Title of Thesis

Characterisation, discrimination and phylogenetic study of clinically significant members of the *Mycobacterium abscessus/chelonae* taxonomic complex.

Number of volumes: Volume 1 of 1

Name: Anne Barrett

Qualification: Doctor of Philosophy

School/Institute: Chemical Engineering and Advanced Materials (CEAM)

Submitted: July 2022

Abstract

Rapidly growing mycobacteria are rare in clinical samples or as significant human pathogens. Members of the *Mycobacterium abscessus*/*chalone* complex are notable exceptions, implicated as colonisers and potential pathogens in a variety of clinical situations.

Cystic fibrosis patients, with reduced lung function, can be colonised with organisms from this complex. An apparent poorer prognosis, post lung transplant, associated with *Mycobacterium abscessus* colonisation, potentially precluding lung transplantation, complicates this.

At the start of this project reliable discrimination between these species was problematic. Given the clinical importance this project sought to compare the genomes of *M. abscessus* and *M. chelonae*, improve identification and explain differences in pathogenicity.

Initially there were many whole genomes for *Mycobacterium abscessus* but none for *Mycobacterium chelonae* so a whole genome sequence for a clinical strain confidently assigned as *M. chelonae* (HPA 006) was obtained using the Ion Torrent (IT) and MinION (ONT) platforms. These data were used to identify regions of difference between these two species for identification and to explain putative variations in virulence and antibiotic resistance.

Average Nucleotide Analysis (ANI) confirmed *Mycobacterium abscessus* and *Mycobacterium chelonae* complex as separate species.

A pangenome study using PPanGGOLiN identified 655 genes present in *M. abscessus* (ATCC 19977) but absent from *M. chelonae* strains. AntiSMASH analysis highlighted differences in the overall number of Biosynthetic Gene Clusters (BGC) present in each species and identified BGCs present in one species but absent in the other.

Of the 655 differential genes identified the largest category was hypothetical, including proteins with domains of unknown function. Despite annotation other genes proved challenging to correlate with known mechanisms of resistance or virulence.

There were many genomic elements, present in both *M. chelonae* HPA 006 and *M. abscessus* (ATCC 19977), which show more difference in sequence than average. These differences, suggesting selection and a modified function, may account for the difference in virulence between the strains

Acknowledgments

I would like to extend my sincere thanks to Professor Jarka Glassey (Chemical Engineering Education, Newcastle University) for her support throughout this study, and in particular, her understanding of the challenges presented by the COVID 19 pandemic. Thank you also to Dr Katarina Novakovic (Senior Lecturer in Chemical Engineering, Newcastle University).

Thank you to my colleagues in Public Health England who supported my studies and to Jonathan Turner, Interim Head of Scientific and Technical Services in particular, whose support professionally and personally was freely given and gratefully received. Mr. Neil Bentley, Jonathon's predecessor, was likewise an encouraging and wise mentor.

I would like to acknowledge the technical assistance and advice offered by Dr Jeni Devi and Dr Nick Allenby from Demuris at Newcastle University. Thank you for your patience Jeni, it was truly appreciated. Thank you Nick for the kind use of the laboratory facilities.

I would like to offer grateful thanks to Alan Ward, Professor Emeritus, Newcastle University, who led me through this thesis project. Without his help, expertise, guidance and feedback, I would not have come this far. And to Dr John Magee, who, in his role as Clinical Services Director in Public Health England supported this study, and for his mentorship and unwavering encouragement over the course of my career in public health microbiology.

Dedicated to my parents, James and May Barrett, in loving memory.

Table of Contents

List of Figures

List of Tables

List of Abbreviations and Bioinformatic Terms

Programme	Function	Installation
Anaconda	A free and open-source distribution of the programming languages Python and R	https://repo.anaconda.com/archive/Anaconda3-2021.05-Linux-x86_64.sh
ANI	Average Nucleotide Identity	
BLAST	Basic Local Alignment Search Tool	https://blast.ncbi.nlm.nih.gov
Canu	Fork of the Celera Assembler, designed for high-noise single-molecule sequencing such as Oxford Nanopore MinION. It is a hierarchical assembly pipeline	https://github.com/marbl/canu/releases
CASSIS	Cluster Assignment by Islands of Sites. Bacterial biosynthetic gene cluster boundary prediction using ClusterAssignment.	Integrated into antiSMASH
Conda	An open-source package and environment management system which makes it easy to install/update packages and create/load environments. It is included in Anaconda	

Diamond	Program for finding homologs of protein and DNA sequences in reference databases	Releases · bbuchfink/diamond · GitHub
EC	Enzyme commission numbers. A numerical classification scheme for enzymes, based on the chemical reactions they catalyse.	
eggNOG	(evolutionary genealogy of genes: Non supervised Orthologous Groups) Database of orthologous proteins and functional annotations at multiple taxonomical levels hosted by EMBL (European Molecular Biology Laboratory)	http://eggnog5.embl.de/#/app/home
Flye	Long read assembler De novo assembler for single molecule sequencing reads using repeat graphs	https://github.com/fenderglass/Flye https://github.com/fenderglass/Flye/blob/flye/docs/USAGE.md#quickusage
Geneious	Genomic Sequence Alignment and Assembly software package.	Geneious Prime 2022.0.1 http://www.geneious.com/
Interpro2GO	Integrated resource of protein families, domains and sites which are combined from a number	https://www.ebi.ac.uk/GOA/InterPro2GO

	of different protein signature databases	
KEGG	Kyoto Encyclopedia of Genes and Genomes	https://www.genome.jp/kegg/
Megan 6	Metagenome Analyzer	https://software-ab.informatik.uni-tuebingen.de/download/megan6/welcome.html
MiBiG	Minimum Information about a Biosynthetic Gene Cluster. A specification which provides a robust community standard for annotations and metadata on biosynthetic gene clusters and their molecular products.	MIBiG: Minimum Information about a Biosynthetic Gene cluster (secondarymetabolites.org)
MIRA 5	Mimicking Intelligent Read Assembly A multi-pass DNA sequence data assembler/mapper for whole genome and EST/RNASeq projects. MIRA assembles reads from Ion Torrent	GitHub - bachev/mira: MIRA sequence assembler
MUMmer	Rapid entire genome alignment system	NUCmer:github.com/gmarcais/mummer
NCBI PGAP	NCBI Prokaryotic Genome Annotation Pipeline	https://github.com/ncbi/pgap

PPanGGOLiN	Software suite used to create and manipulate prokaryotic pangenomes from a set of either genomic DNA sequences or provided genome annotations. Employed to build a partitioned pangenome graph from microbial genomes	https://github.com/labgem/PPanGGOLiN
proovread	A hybrid error correction tool, it used short read data to correct long reads. Used on the *M. chelonae* HPA 006 MinION reads (long reads)	https://anaconda.org/imperial-college-research-computing/proovread Imperial College research computing
Pyani	Program that calculates average nucleotide identity (ANI) and related measures for whole genome comparisons, and renders graphical summary output.	https://github.com/widdowquinn/pyani https://github.com/widdowquinn/pyani/blob/master/README_v_0_2_x.md
SANDPUMA	Ensemble algorithm, based on newly trained versions of all high-performing algorithms, which significantly outperforms in dividual methods.	Integrated into antiSMASH
SEED	Annotation environment Database infrastructure	https://www.theseed.org/wiki/Home_of_the_SEED

	applied in the comparative genomics in MEGAN software.	
SPAdes	St Petersburg genome Assembler Used in both single-cell and standard (multicell) assembly	https://github.com/ablab/spades
Ubuntu	Linux distribution based on Debian and composed mostly of free and open-source software.	Get Ubuntu | Download | Ubuntu
Unicycler	A pipeline running a series of programs; the input data for one program produces output that is passed as input data to the next program in the pipeline:	https://bioconda.github.io/recipes/unicycler/README.html

List of Chemicals used in the project

All chemicals and reagents used in this project were of analytical grade and purchased at the highest purity available. Purities, suppliers, state (solid (S), liquid (L) or gas (G)) and abbreviations (used along the text) of the compounds used are listed below

Chemical	Purity	Supplier	State
Cetyltrimethylammonium bromide (CTAB)		VWR International Ltd	L
Chloroform	>99%	Sigma Aldrich	L
Chloroform/ isoamyl alcohol	>99.9% (chloroform), ≥99.0% (isoamyl alcohol)	Fisher Scientific	L
Ethanol	70%	VWR International Ltd	L
Isopropanol	99.5%	Fisher Scientific UK Ltd	L
Lysozyme		Thermo Fisher Scientific	S
Proteinase K solution		Merck Life Science Limited	L
RNase A		VWR International Ltd	L
Sodium Chloride (NaCl)		Fisher Scientific UK Ltd	S
Sodium dodecyl sulphate (SDS)	≥97%	Sigma Aldrich	S
Tris-EDTA (TE buffer)		Thermo Fisher Scientific	L
Tris HCL		VWR International Ltd	L

List of papers being prepared for publication

Complete genome sequence of a clinical isolate of *Mycobacterium chelonae*

The application of ANI to the classification of organisms assigned to the *Mycobacterium abscessus* clade

Chapter 1. Introduction, Literature Review and Aims

1.1 Genus *Mycobacterium*

Species assigned to the genus *Mycobacterium*, the only genus in the family *Mycobacteriaceae* have been isolated from many diverse environments. *Mycobacteria* all lie within the genus *Mycobacterium* and are aerobic, non-motile, mycolic acid containing bacteria that are characteristically acid-fast and contain DNA with a G+C content of 62–72 mol%. *Mycobacteria* are phenotypically classified into slow growing or rapidly growing species, and are pigmented or non-pigmented. At the time of writing, the List of Prokaryotic names with standing in Nomenclature (LPSN) (Euzeby, 1997) included 260 species (209 validly published with a correct name and 14 synonyms) within the genus *Mycobacterium*. The majority of these species are free living in soil and water, and many have been isolated from environmental, plant and animal sources (World Health Organisation, 2004; Magee & Ward, 2012). Though the principle human pathogens are *M. leprae* and organisms assigned to the *M. tuberculosis* complex, it is well established that members of other mycobacterial species, variously described as atypical mycobacteria; mycobacteria other than tuberculosis, or non-tuberculous mycobacteria may be opportunistically responsible for disease in humans (Magee & Ward, 2012). Everyday contact between humans and environmental mycobacteria may result in little or no adverse effect; transient colonisation or; in some individuals, may lead to a pathogenic process (Falkinham, 1996; Primm *et al*., 2004). Modern developments in clinical approaches to the treatment of clinical, genetic and immunological disorders including the use of immune-suppressive therapies have, paradoxically, led to an increased recognition of the dangers of opportunistic infection by environmental organisms, including mycobacteria (Olivier 1998; Wallace *et al*., 1998; Phillips & von Reyn, 2001).

Classically, mycobacteria are separated into two broad taxonomic groups based on the rate of growth of a controlled inoculum on solid media. Those species termed slowly-growing require greater than 7 days for the appearance of colonial growth, visible to the naked eye, from a dilute inoculum. Conversely, rapidly-growing species require less than 7 days under the same conditions (Wayne & Kubica, 1986). This phenotypic distinction is supported by phylogenies based on gene sequence studies (Goodfellow & Magee, 1998, Nouioui *et al.,* 2018). Those species which are potential opportunistic infectors and therefore clinically relevant, more

frequently occur in the slowly growing group (Goodfellow & Magee, 1998; Magee & Ward, 2012). Despite this, some rapidly growing species may occasionally be of clinical significance, (Armstrong & Parrish, 2021; Tortoli *et al.,* 2017). Most notably those assigned to the *Mycobacterium abscessus/chelonae* clade (Magee & Ward, 2012). This clade includes *M. abscessus* subsp. *abscessus*, *M. abscessus* subsp. *bolletii* (Adékambi *et al*., 2006b)*, M. abscessus* subsp *massiliense* (Adékambi *et al*., 2004)*, M. chelonae* subsp *chelonae, M. chelonae* subsp *bovis* (Kim *et al*., 2017) and *M. chelonae* subsp gwanakae (Kim *et al*., 2018) *M. immunogenum (*Wilson *et al*., 2001), *M. salmoniphilum* (Whipps *et al*., 2007)*, M. franklinii (*Nogueira *et al*., 2015a*), M. stephanolepidis* (Fukano *et al*., 2017a) and *M. saopaulense* (Bergey *et al*., 1923; Leao *et al*., 2011).

Given their predilection for growth temperatures lower than that of the body temperature of humans, it is not surprising that *M. abscessus* (*sensu lato*) and *M. chelonae* have been implicated as causal agents in infections of wounds due to skin surface trauma (Arnold *et al*., 2012), rather than as systemic pathogens. However, other areas of the human body where lower natural temperatures may prevail can provide a suitable environment. One such area is the upper respiratory tract. As these mycobacterial species may be present in water systems (including potable water) they may be transient colonisers of the respiratory tract (Primm *et al*., 2004; World Health Organisation, 2004). In an immune-competent individual such colonisation is usually temporary, but when there is impairment of respiratory function then colonisation may become more established, difficult to eradicate and lead to a pathogenic effect.

The *M. chelonae* clade as described by Magee and Ward (2012) has had a tangled taxonomic history. *M. abscessus* and *M. chelonae* were initially thought to be synonymous but were described as separate species by Kusunoki & Ezaki, (1992). More recently, several organisms which were initially validly named as independent species have been reclassified as subspecies of *M. abscessus* (Leao *et al*., 2011, Minias *et al.,* 2020). These shifts in classification have further confused the relationships between species and their resultant pathogenicity/virulence. However, taxonomic studies have largely been based upon phenotypic characters or short sequences of housekeeping genes, notably sequences of the 16S rRNA and *rpo*B genes.

The taxonomic relatedness of these species has not been studied in depth by reference to the whole genome. Equally, the association of taxonomic variance with pathogenicity or virulence has not been explored at the genomic level. Previously, study of these factors was hindered by the lack of whole genome sequences for *Mycobacterium chelonae*, however, the number of deposited nucleotide sequences in the National Centre for Biotechnology (NCBI) database has increased significantly during the course of this project, which began in 2013 with practical work completed in 2019

1.1.1 The Mycobacterium tuberculosis complex

The existence of microorganisms and their role in disease was unknown until the early 18th century when several scientists postulated the possibility of unseen living organisms. Arguably, the science of microbiology began with the pioneering Dutch microscopist Antonie van Leeuwenhoek (1632-1723). He is often referred to as the "father of microbiology" since he constructed the first compound microscope and used it to describe the "animalcules" which he was then able to detect (Corliss, 1975). The cell culture techniques developed by Robert Koch (1843-1910) enabled an increasing clinical interest in microbiology. However, it was Koch himself who became interested in tuberculosis and postulated that the disease was caused by a microorganism (Koch, 1882).

Initial interest in mycobacteria was, unsurprisingly, driven by clinical need and thus focussed primarily on *Mycobacterium tuberculosis* and on *Mycobacterium leprae*. Although the existence of bacteria and specifically these two species may not have been known until the 18th century, the medical conditions later ascribed to them, namely leprosy and tuberculosis have been recognised since the beginning of recorded history. Reference to tuberculosis in early literature uses terms descriptive of the clinical appearance of the condition; for example, consumption, phthisis and the white plague. The term tuberculosis dates from the 19th century in reference to the tubercules found to be present in the lungs of deceased sufferers (from Modern Latin, tuberculum "small swelling, lump"). In contrast, leprosy has no such synonyms, although to avoid concerns in relatives and carers the term Hansen's disease was often used in reference to G.H.A. Hansen (Irgens, 2002) who discovered the causative organism.

The use of tandem repeat sequences as genetic markers suggest that tuberculosis can be traced back some 40,000 years into history, coincident with the beginning of the migration of "modern man" from the African continent (Wirth *et al*., 2008). Subsequently, tuberculosis most probably spread to other humans and to domesticated animals such as goats and cattle as trade routes developed. A genomic study of modern isolates suggests that *M. tuberculosis* attained its worldwide distribution following human migration out of Africa during the Pleistocene epoch (Comas *et al*., 2013). A study of three 1,000 year-old mycobacterial genomes from Peruvian human skeletons (Bos *et al*., 2014) revealed that a member of the *M. tuberculosis* complex caused human disease before contact with this group could have occurred. The ancient strains were shown to be distinct from known human-adapted strains and were related most closely to those adapted to seals and sea lions, supporting the theory sea lions and seals could have carried the disease across the Atlantic to South America. Hunters are likely to have come into contact with the disease via this source. These transmission routes are somewhat speculative but variants of the bovine strain of tuberculosis (*Mycobacterium bovis*) occurring in goats (*Mycobacterium caprae*) and seals (*Mycobacterium pinnipedii*) are well recognised (Magee & Ward, 2012).

The past three decades has seen steady advances in the laboratory diagnosis of tuberculosis, including, automated liquid culture, polymerase chain reaction (PCR) applied directly to clinical samples and positive cultures, random access platforms which utilise PCR to detect drug resistance directly from samples or positive cultures and the use of Whole Genome Sequencing (WGS) to identify relatedness between strains. All of these have improved patient pathways by reducing the time taken to culture the organism and provide susceptibility profiles, enhance contract tracing and allow meaningful interventions which can reduce transmission events and control spread (Cheng *et al*., 2004; Chihota *et al*., 2010; Hamdi *et al*., 2020; Park *et al*, 2022).

The World Health Organisation (WHO) reported that globally, TB incidence is falling at around 2% per year and between 2015 and 2020 the cumulative reduction was 11%, excellent progress toward the End TB Strategy milestone of 20% reduction between 2015 and 2020. The COVID 19 pandemic has had a detrimental effect on this strategy with the most obvious impact being a large global drop in the number of new diagnoses, which fell from 7.1 million in 2019

to 5.8 million in 2020, an 18% decline back to the level of 2012. Provisional data up to June 2021 show ongoing shortfalls (World Health Organisation; Global Tuberculosis report 2021). Of the five key priorities of the United Kingdom Health Security Agency (UKHSA) TB Action Plan for England 2021-2026, the first is the recovery of services affected by the COVID 19 pandemic so that the expected increase in undetected and unreported cases of active disease and latent infection can be addressed (UKHSA TB Action Plan for England 2021-2026)

This focus on *M. tuberculosis* and *M. leprae* led to other species, as they were increasingly detected, being collectively labelled as "non-tuberculous mycobacteria" NTM; other terms have since come into use e.g., environmental, opportunist or atypical mycobacteria. However, although a convenient shorthand, these terms can be misleading since there is considerable variation in the pathogenic potential as well as the clinical manifestation of these species.

1.1.2 Non-tuberculous mycobacteria

As noted earlier the *Mycobacterium tuberculosis* complex is a term used to refer to organisms classified as *M. tuberculosis* and to those which are known to cause tuberculosis in man or animals (e.g., *Mycobacterium bovis*) or are genetically indistinguishable except by the most exacting of analyses (e.g., *Mycobacterium africanum*) (Magee & Ward 2012). There are a number of terms which are used to label those mycobacterial species which are not assigned to the *Mycobacterium tuberculosis* complex. The most common of these terms is non-tuberculous mycobacteria (NTM). However, the terms "atypical" or "opportunist" will often be noted as will "mycobacteria other than tuberculous" or MOTT. Although NTM may be the most accurate of these descriptive terms, the other terms have some understandable validity. The term "atypical" highlights differences in phenotypic properties often shown by these species (from those of *M. tuberculosis*). Alternatively, "opportunist" refers to the non-obligate pathogenicity of these species, again in contrast to *M. tuberculosis*. However, some species have no known occurrence as pathogens which belies their labelling as opportunist infectors. The weakness of these terms in identifying these species lies in the fact that some of those labelled as non-tuberculous (i.e. NTM) may cause pulmonary infections which can be difficult to distinguish from tuberculosis on primary presentation (Johnson & Odell, 2014). Nevertheless, a surprising number of these species can cause disease if certain conditions are

met. In a sense they are niche pathogens i.e. causing infections when opportunity and occurrence coincide.

NTM are ubiquitous environmental organisms mostly found in soil and water which cause lung, lymph node, joint and catheter-related and disseminated infections in susceptible individuals. NTM are also implicated in progressive inflammatory lung damage, a condition which is termed NTM pulmonary disease (NTM-PD) (Haworth *et al*., 2017). Despite this worldwide distribution NTM are geographically heterogeneous. Two studies that attempted to study this geographic diversity were carried out by Marras & Daley in 2002 and by Martín-Casabona *et al*., in 2004. The latter reported NTM isolation over a period of three decades up to 1996, but included laboratories from Europe, Turkey, Iran and Brazil only, additionally sampling processes were not consistent over the time period. Both studies confirmed that *Mycobacterium avium* complex (MAC) isolates predominate worldwide, whilst others, *Mycobacterium malmoense*, *Mycobacterium xenopi* and *Mycobacterium kansasii* demonstrated characteristic geographic occurrence. *M. malmoense* being more commonly found in Northern Europe, *M. xenopi* in Canada and the United Kingdom and *M. kansasii* in midwestern and southwestern states of USA.

In 2013 Hoefsloot *et al*., published the results of a collaborative study in which global partners in the NTM-Network European Trials Group (NET) framework (a branch of the Tuberculosis Network European Trials Group (TB-NET) were invited to provide data on the total number of patients from whom NTM were isolated from pulmonary samples in their hospital, regional or reference laboratory in the year 2008. Species identification results and details of the identification methods used were submitted and only those partners who had NTM isolates in excess of 30 could contribute to ensure sufficient experience and interpretability of results was demonstrated. One isolate per species, per patient was included in the analysis. In total, species identification data was received on 20,182 patients, from 62 laboratories in 30 countries across 6 continents. In all 91 different NTM species were isolated.

The species distribution among NTM isolates from pulmonary specimens in 2008 was found to differ by continent and by country within these continents, therefore the frequency and indicators of pulmonary NTM disease in each geographical location may be influenced by these differences in species distribution. The six most frequently isolated NTM were *M. avium*

complex (9421 isolates; 47%), *M. gordonae* (2170 isolates; 11%), *M. xenopi* (1605 isolates; 8%), *M. fortuitum complex* (1322 isolates; 7%), *M. abscessus* (664 isolates; 3%) and *M. kansasii* (720 isolates; 4%). These six species accounted for 80% of all mycobacteria identified (Table 1, Figures 1 and 2).

The most frequently encountered rapidly growing species were *M. abscessus* and *M. fortuitum* and again, these species demonstrated clear geographical differences. Rapidly growing species were found to be more prevalent in centres in East Asia. Overall, they made up 27% of all NTM isolates, compared to 17.9%, 16% and 14% from collaborating centres in North America, South America and Europe respectively. Examining data from within countries in East Asia further differences could be discerned. In Tokyo, rapidly growing species accounted for 6.6% of all isolates, in contrast to collaborating centres in Taiwan and S. Korea where the rates were 50% and 28.7% respectively. In S. Korea *M. abscessus* was the second most frequently isolated NTM after isolates belonging to the *M. avium* complex. The data confirmed previous findings that *M. malmoense* is encountered most frequently in northern Europe.

Table 1. Distribution of respiratory non tuberculous mycobacteria (NTM) isolates. Data are presented as n or n (%), where n is the number of patients or isolates. MAC: *Mycobacterium avium* complex isolates and (percentage of all non tuberculous mycobacteria). Reproduced with permission from Hoefsloot *et al*., 2013.

Region	Number of participating Laboratories	Number of Patients from whom NTM were isolated	Number and (percentage of) isolates identified as members of MAC *
Europe	43	6803	2500 (36.9)
North America	4	4913	2553 (52.0)
South America	3	393	123 (31.3)
Australia (Queensland)	1	453	322 (71.1)
Asia	3	1974	1062 (53.8)
South Africa	2	5646	2849 (50.5)
Total	56	20182	9421 (46.7)

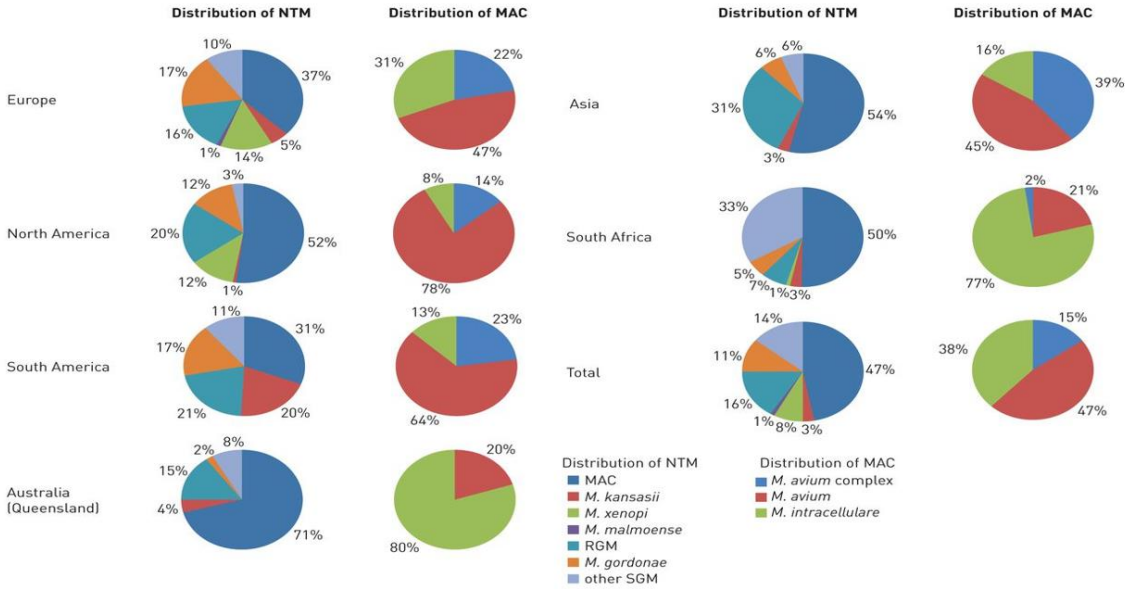


Figure 1. NTM-Network European Trials Group (NET) Study: Distribution of respiratory nontuberculous mycobacteria (NTM) isolates, isolated from 62 collaborating laboratories across 6 continents. Reproduced with permission from Hoefsloot *et al*., 2013.

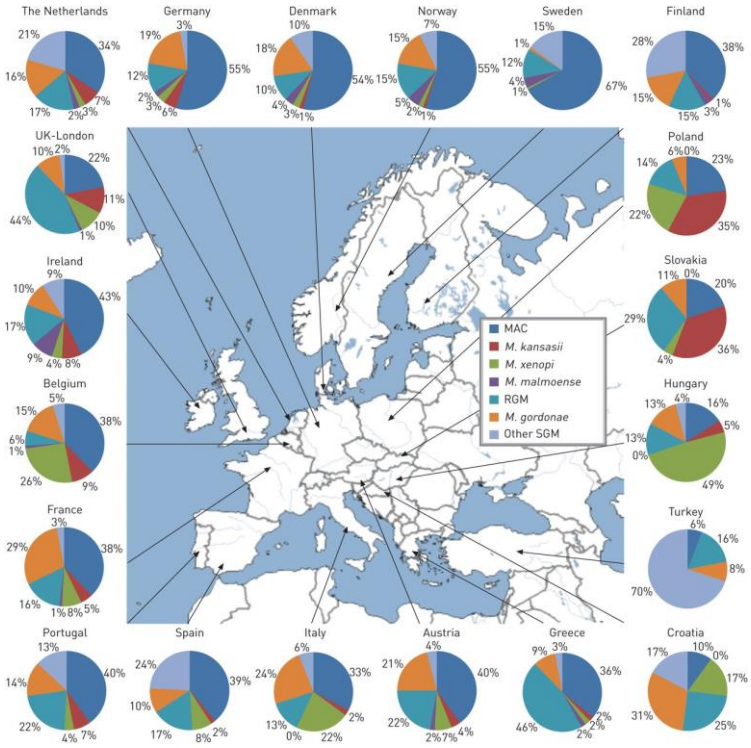


Figure 2. NTM-Network European Trials Group (NET) Study: Distribution of NTM by species isolated from pulmonary samples in 2008 in 20 European countries. Reproduced with permission from Hoefsloot *et al.,* 2013.

NTM are divided into two broad categories, slow and rapidly growing species with the most common non tuberculous mycobacteria implicated in human infections belonging to the slow-growing category (Wallace *et al*., 1990; Campbell & Jenkins, 1995). Several recent studies report that the incidence of infection with NTM is increasing worldwide (Simons *et al*., 2011; Koh *et al*., 2013; Chou *et al*., 2014; Ratnatunga *et al*., 2020; Stout *et al.,* 2016) with the rate of infection in England, Wales and Northern Ireland more than doubling between 1996 and 2006. (Shah *et al*., 2012). The same advances in culture and molecular techniques which drove improvement in the laboratory diagnosis of tuberculosis have also enhanced the recovery of NTM from clinical samples, additionally, improved recognition by clinicians of the relationship between these organisms and the diseases they produce is also a factor.

Perhaps the most widely reported of those NTM implicated in disease are organisms assigned to the *Mycobacterium avium* complex (MAC). Originally named for its occurrence in birds (Magee & Ward, 2012), it was later established that the species had several variants and close antigenic relationships with other species, notably with *M. intracellulare* (which led to the occasional use of the term MAI complex). It was accepted that *M. avium* was not an homogenous species (Stanford & Grange, 1974). However, separation of the component variants was exacting, and this led to the term *M. avium* complex becoming commonly used. Strains assigned to the *M. avium* complex have been found to be common in the environment and thus unsurprisingly detected as commensal or occasionally causing disease in humans and animals (Gordin & Horsburgh, 2015). MAC strains have been detected in various clinical settings, in respiratory infections in the elderly and most notably in childhood cervical lymphadenitis (Thavagnanam *et al*., 2006). These limited clinical interactions with humans suggest that immune defence mechanisms are effective in normal situations. The human immunodeficiency virus (HIV) epidemic and associated progression to acquired immune-deficiency syndrome (AIDS) triggered an escalation in systemic MAC infections, due to the concomitant severe reduction in immune capability. Quite rapidly an organism with a history of limited impact in human disease became a dangerous pathogen to those suffering from HIV/AIDS. Although treatment of the infections was attempted, results were poor, and the impact of MAC was only alleviated when HAART (highly active anti-retrovirus therapy) was introduced. This therapy breakthrough resulted in immune response capability returning to

normal levels, and the incidence of disseminated MAC was dramatically reduced (Moore & Chaisson, 1999).

Mycobacterium chimaera is an organism closely related to *Mycobacterium intracellulare* and one of those species separated from the "MAI complex" by modern taxonomic methods. Although now known to be common in the environment, and particularly in water, it was previously not encountered as a human pathogen. In 2014, six cases of severe infection due to *Mycobacterium chimaera*, were reported in cardiac surgery patients in Zurich (Sax *et al*., 2015). These patients had undergone heart surgery with cardiac by-pass. It was established that these infections resulted from mycobacteria-laden biofilms formed within the heater/cooler units of the by-pass apparatus with acquisition by patients as a result of bioaerosols emitted from the water systems of these units. The inherent characteristics which allow this species to selectivity colonise such equipment are yet to be elucidated; but it is now clear that *Mycobacterium chimaera* is an emerging opportunistic threat to patients undergoing coronary bypass surgery and open-heart procedures requiring extracorporeal devices (Walker *et al*., 2017). Public Health England first published Guidance for healthcare professionals on the infection control and clinical aspects of *Mycobacterium chimaera* infection associated with cardiopulmonary bypass in June 2015. The most recent update to this guidance indicates that, as of 7 October 2022, there were 49 cases of *Mycobacterium chimaera* infection following surgery on cardiopulmonary bypass, of which 34 have died. The median interval between surgery and diagnosis is 24 months but ranges from less than 1 month to 154 months (UKHSA, 2022)

The opportunity to make the step from the environment to a human infection may sometimes be a function of geographic constraints. *Mycobacterium kansasii*, for example, appears to be most commonly encountered in association with watercourses (Kaustova *et al*., 1981; Fisheder *et al*., 1991). In contrast, infections with *Mycobacterium malmoense* are predominantly encountered in more northerly geographic locations. Named for its first recognition as a cause of infections in the city of Malmo, Sweden; *M. malmoense* has been isolated from patients in the North of England and in Scotland but less often in more southerly regions (Connolly *et al*., 1985; France *et al*., 1987). Both of these species will cause lung infections in elderly patients, especially those with underlying respiratory conditions.

Notwithstanding the examples above, systemic diseases associated with NTM are uncommon. These mycobacterial species are more likely to be encountered in superficial lesions (Wallace *et al*., 1990). *Mycobacterium marinum* for example is a fish pathogen found in temperate climates (Clark & Shepard, 1963). However, it can cause cutaneous granulomas on the extremes of arms (e.g., forearms, wrists, fingers) and legs (knees, feet, toes). These infections are associated with swimming pools and fish tanks and are often referred to as "swimming pool" or "fish tank" granulomas. The lesions can spontaneously heal, albeit slowly, but may require chemotherapy or even surgical debridement (Norden & Linell, 1951; Magee & Ward, 2012). Sequence analysis of the 16S rRNA and 16S-23S rRNA genes shows *Mycobacterium marinum* to be closely related to *Mycobacterium ulcerans* (Roth *et al*., 1998). However, this species will cause far more serious progressive and malignant cutaneous ulceration of the lower limbs which can lead to muscle and tendon damage and possibly deformity. *M. ulcerans* is found in several tropical regions (Boisvert, 1977) and will sometimes be referred to by one of several geographic epithets e.g., Buruli ulcer (a region of sub-Saharan Africa), Bairnsdale ulcer (a region of the Australian state of Victoria).

Aside from the fact that any microorganism contaminating an area of superficial damage may be unwelcome, few of the rapidly growing mycobacterial species have been implicated as pathogens in humans. There are however some exceptions. *Mycobacterium cosmeticum* for example has been implicated as a cause of granulomatous lesions on the fingers of individuals using nail salons (Cooksey *et al*., 2004). Somewhat similarly, *Mycobacterium mageritense* has been found as a cause of furunculosis associated with footbaths in nail salons (Gira *et al*., 2004). *Mycobacterium novocastrense*, initially thought to be a variant of *M. marinum*, was isolated from a slowly spreading skin granulation on the hand of a child (Shojaei *et al*., 1997). However, the only rapidly growing mycobacterial species encountered as systemic pathogens are those assigned to the *Mycobacterium abscessus/chelonae* clade; most specifically *M. abscessus* and its known variants.

1.2 Mycobacterial Systematics

1.2.1 Genus description

Before continuing it is necessary to give a brief evaluation of recent taxonomic changes which were considered for this genus. It has been proposed that the genus *Mycobacterium* be divided into five different genera based on phylogenetic data. The results have proposed the separation of the various species into five major clades (Gupta, Lo & Son 2018). The authors proposed that the genus *Mycobacterium* would contain members of the *Mycobacterium tuberculosis* complex, while most of the nontuberculous mycobacterial species would be classified in the following four genera: *Mycobacteroides*, *Mycolicibacter*, *Mycolicibacterium*, and *Mycolicibacillus*. The proposed division of the single genus *Mycobacterium* by Gupta and colleagues was undoubtedly supported by the criteria used for establishing novel species using genomic comparisons. However, it has created debate in the wider community, where changing the taxonomic designation of so many mycobacterial species of clinical importance could be problematic from a laboratory diagnostic and patient care standpoint. Amalio Telenti, stated 20 years ago: "clinical meaningfulness should be the key to taxonomic precision" (Telenti 1998). Others active in the field of mycobacteriology have noted their reservations to these proposed changes. (Tortoli *et al.,* 2019).
For the purposes of this thesis, I will continue to use the designation *Mycobacterium*.

The type genus *Mycobacterium* is the sole genus within the taxonomic Family *Mycobacteriaceae* which is of the Order *Corynebacteriales* and the Phylum *Actinobacteria* (Ludwig *et al*., 2012). The name *Mycobacterium* is derived from the Latin root "myco" which may mean fungus-like (some species do show filamentous and branched forms) but also "waxy" (which reflects their high lipid content). Mycobacteria have a high Guanine and Cytosine content (57-73%). Their cell walls contain a high content of phospholipids and are rich in chloroform soluble waxes, including mycolic acids with distinctive long branched chains (60-90 carbon atoms). On pyrolysis cells release fatty acid esters with 22-26 carbon atoms (Magee & Ward, 2012). Species encompassed by this genus characteristically display "acid-fastness" at some stage of growth. This particular characteristic classically refers to the retention of a dye (basic fuchsin) in stained preparations of the organisms despite attempted decolourisation with strong acid. This trait may also be shown by some other members of the order *Corynebacteriales* albeit to a lesser degree. However, mycobacteria will often display a

unique resistance to decolourisation by a mixture of both acid and alcohol (usually methyl alcohol is used in this context).

1.2.2 16S rRNA gene

The 16S rRNA gene is universal in bacteria, having a high degree of within-species conservation resulting from the critical importance of the gene to cell function (Woese, 1987; Clarridge, 2004). The stability of this gene in one species and difference from the equivalent gene in another species led to 16S rRNA gene sequence analysis being used as a mainstay of bacterial genus and species description (Garrity & Holt 2001)..

16S rRNA gene sequence analysis data has been used with confidence in the differentiation of genera. Stackebrandt *et al*., (1997) and later Zhi *et al*., (2009) proposed a definition of the family *Mycobacteriaceae* based on 16S rRNA signature nucleotides at positions 128:233 (G-C), 250 (U), 316:337 (C-G), 418:425 (C-G), 586:755 (U-G), 599:639 (U-G), 662:743 (C-G), 987:1218 (G-C), 1000:1040 (A-U) and 1026:1035 (U-G). The genus currently consists of over 260 species and subspecies.

In 2017, Beye *et al*., assessed the accuracy of 16S rRNA nucleotide sequence similarity for the classification of new mycobacterium isolates at species level, studying the pairwise identity values of the 16S rRNA gene sequence for 131 validly published Mycobacterium species. They observed that 16S rRNA gene sequence similarity thresholds for delineating bacterial species were valid for only 0.76% of 131 Mycobacterium species studied (a single species *Mycobacterium poriferae).* Additionally, as the study covered over 70% of the Mycobacterium species currently validly described, they propose that the 95% and 98.65% thresholds are not suitable for this genus and should at best be used as indicators, not as a reference standard, for classifying new Mycobacterium species.

This reinforces earlier studies which also demonstrated that 16S rRNA gene sequence analysis is not always sufficiently discriminatory at the species level and that this approach is dependent upon very high precision in determining detailed nucleotide sequence data (Mende *et al*., 2013; Varghese *et al*., 2015)

1.2.3 Species assignment in mycobacteria

The first attempts to classify *Mycobacterium* species, in the 1950s, resulted in a division into four groups based on growth rate and pigmentation (Timpe & Runyon, 1954; Runyon, 1958, 1959). Groups I, II, and III were composed of slow-growing strains. Group I was composed of non-chromogenic species. Conversely groups II and III contained species which display notable pigmentation either when subjected to light stimulation (photo-chromogenic) or without such a stimulus (scoto-chromogenic). Group IV consisted of rapidly growing species. A more comprehensive collection of phenotypic characters was introduced in a series of numerical taxonomy studies, many of which were under the auspices of the International Working Group on Mycobacterial Taxonomy (IWGMT) (Table 2). From the late 1960s into the 1990s this *ad hoc* grouping of scientists applied numerical taxonomic principles to both slow- and rapid-growing mycobacteria, bringing some much-needed organisation to the classification of this genus (Wayne, 2000).

The phenotypic characters utilised in numerical taxonomic studies were valuable in allowing easier recognition of some species on primary isolation. However, it quickly became apparent that several important species were not homogenous, and that some "species" were a catch-all name for ill-defined isolates. For example, the species *M. avium* and *M. intracellulare*, each with a separate pathogenesis, were often grouped together as the *M. avium/intracellulare* complex. To overcome this problem, it was necessary to move from analysis of cell reactions as a mirror of genetic make-up, to analysis of elements of the genome itself. This approach was the basis of the final IWGMT study which focussed on problems among slowly-growing species (Wayne *et al*., 1996). This study used two methods at the molecular level; sequence analysis of the 16S rRNA gene and DNA: DNA pairing experiments, alongside the determination of pheno- and chemo-taxonomic characteristics. Termed polyphasic taxonomy (after Colwell, 1970) this approach recognises that both genotypic and phenotypic data are required for the accurate delineation of a bacterial species.

Table 2. Phenotypic characters used in differentiating members of the *Mycobacterium abscessus*/*chelonae* clade.

	M. abscessus	*M. abscessus subsp. bolletii*	*M. abscessus subsp. massiliense*	*M. chelonae*	*M. franklinii*	*M. immunogenum*	*M. salmoniphilum*	*M. stephanolepidis*	*M saopaulense*
Colony type	Smooth or rough	Smooth or rough	Intermediate between rough and smooth	Smooth/rough Moist/shiny	Not Available	Rough	Smooth/ shiny	Smooth or rough (mainly rough)	Smooth
Pigment/ colour	None/ White/grey	None	None	None/buff	None/	None/ Off-white	None/ cream	None/White	None/ Not Available
Temperature Range (^{0}C)	28^{0}C-37^{0}C	24^{0}C -37^{0}C Optimum 30^{0}C	24^{0}C -37^{0}C Optimum 30^{0}C	22^{0}C -40^{0}C	25^{0}C -37^{0}C Optimum 28^{0}C	30^{0}C -35^{0}C Optimum 30^{0}C	20^{0}C -30	15^{0}C -35^{0}C Optimum 20^{0}C	
Citrate utilisation	-	-	-	+	+/-	-	Not Available	Not Available	+
5% w/v NaCl	+	-	+	-	+	-	Not Available	-	+
MacConkey agar	+	Not Available	+	+	Not Available	+	+	Not Available	Not Available

Analyses of 16S rRNA gene sequences have proven to be an effective tool in the determination and definition of bacterial species. Many novel *Mycobacterium* species have been characterised and validly described using 16S rRNA gene sequence analysis with associated phenotypic data (Shojaei *et al*., 1997, 2000). 16S rRNA gene sequences from many bacterial species, including mycobacteria have been deposited in international databases, thus providing a resource to aid workers in the classification of bacterial strains. Although an effective technique in most instances, in some closely related species the 16S rRNA gene may be too conserved to allow differentiation at the species level (Magee & Ward, 2012). To resolve such problems DNA: DNA pairing has been applied (Stackebrandt & Goebel, 1994). This technique examines the degree of re-association of single stranded DNA from one organism with that of another to which it is being compared. The degree of relatedness is a factor of the degree of hybridisation; organisms with 70% or more DNA similarity are considered to be of the same species consistent with having at least 96% 16S rRNA gene sequence identity (Stackebrandt & Goebel, 1994). For example, several DNA pairing techniques were used by Kusunoki and Ezaki (1992) in studying the so-called "*M. fortuitum* complex" and led to the proposal to re-instate *M. peregrinum* as an independent species. However, though highly discriminatory, DNA pairing techniques are technically demanding, time consuming and sensitive to experimental conditions. Given that each strain to be studied must be compared individually with strains of each possible related species, the technique is also expensive to perform and as a result now rarely utilised. This has led to the analysis of sequence data from genes less conserved than the 16S rRNA gene. The description of *Mycobacterium conceptionense* for example, was largely based on analysis of a partial *rpo*B gene sequence (Adékambi *et al*., 2006a). In some cases, small variations in several genes have been used to define a new species as with *Mycobacterium paraseoulenese* (Lee *et al*., 2010). However, dependence on small base-pair variances in relatively short, single gene sequences may be misleading and this led to the idea that data from multiple sequences should be concatenated (i.e. aligned and compared in a linear fashion), to add rigor to the species definition (Devulder *et al*., 2005). This approach was used by van Ingen *et al*., (2009) in their description of *Mycobacterium vulneris* when they analysed concatenated sequences from the 16S rRNA, *hsp65* and *rpoB* genes.

Whether based on phenotypic or genomic data, the relationships of individual species to one another can be diagrammatically represented in phylogenetic trees. In principle, such trees show the degree of similarity (or dissimilarity) between representative strains of the organisms being studied. This may allow variances at the genus or species level to be apparent, aiding their definition. Although phylogenetic trees can be constructed from any comprehensive dataset, 16S rRNA gene sequences are the basis of many comparative analyses of bacteria since they are available for the great majority of species.

A limitation of the published 16S rRNA gene sequence databases is that deposited sequences of some species are occasionally too short to provide reliable comparison with those of related species. The analysis of mycobacterial species relationships prepared for Bergey's Manual of Systematic Bacteriology used, wherever possible, sequence data from original species descriptions. However, where these sequences were seen to be of short length, they were replaced by sequence data of higher quality derived from subsequent studies (Magee & Ward, 2012). The resulting phylogenetic tree gathered species into clades *i.e.*, groups based on a single branch within the tree. Standard two-dimensional trees may occasionally mislead since they do not allow the perspective of depth.

Thus, species apparently closely aligned in two-dimensions may in fact be more distantly related when the sequence analyses are presented in three dimensions. However, the analyses carried out by Magee and Ward (2012) and shown as Figures 3 and 4, clearly demonstrate that within the 16S rRNA gene sequence phylogeny determined for rapidly-growing mycobacteria, there was a distinct clade centred on *M. chelonae* and comprising, in addition to this species, *M. abscessus*, *M. bolletii, M. immunogenum, M. massiliense* and *M. salmoniphilum*.

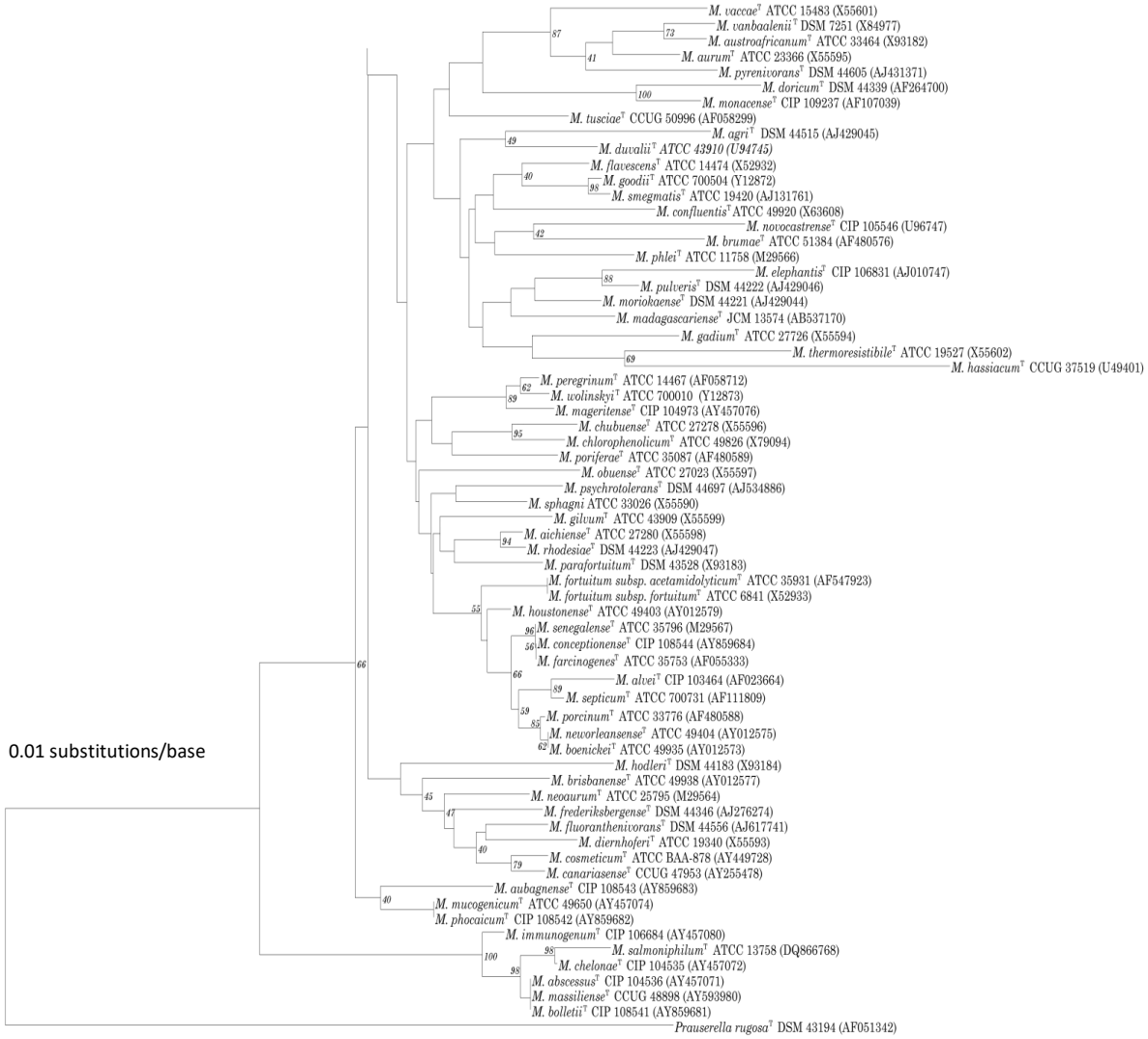


Figure 3. Phylogenetic tree constructed from 16S rRNA gene sequence data of rapidly growing mycobacterial species.

Generated using Jukes and Cantor (1969) similarities, the neighbour-joining algorithm (Saitou & Nei, 1987) using SeaView (Gouy *et al*., 2010) and Dendroscope v2.74 (Huson *et al*., 2007). Bootstrap values from 1000 iteration. From Magee and Ward (2012).

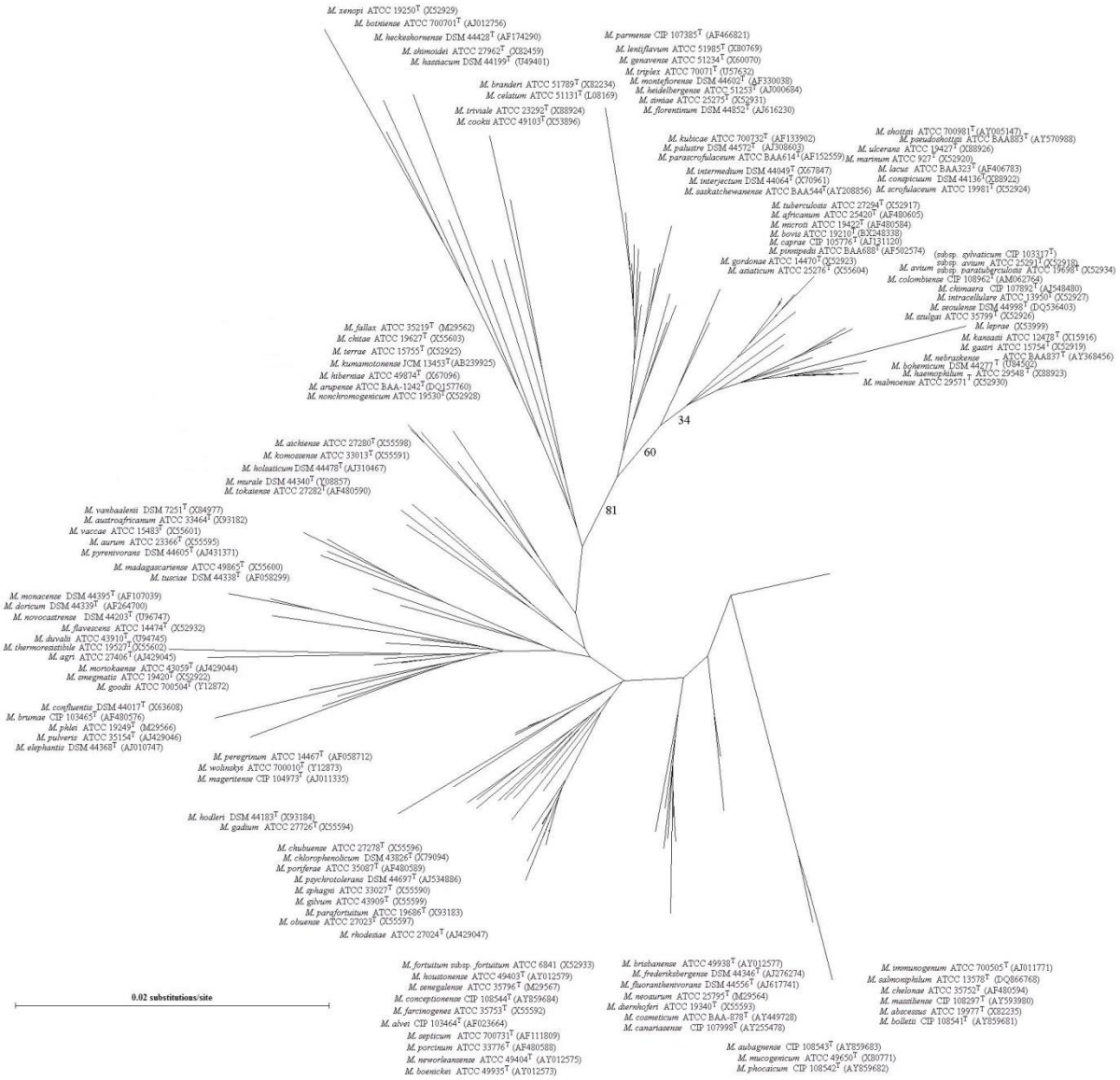


Figure 4. Phylogenetic tree constructed from 16S rRNA gene sequence data of rapidly growing mycobacterial species shown as a radial tree.

Generated using Jukes and Cantor (1969) similarities, the neighbour-joining algorithm (Saitou & Nei, 1987) using SeaView (Gouy *et al*., 2010) and Dendroscope v2.74 (Huson *et al*., 2007). Bootstrap values from 1000 iteration encompassing all mycobacterial species. From Magee and Ward (2012).

The cladistic relationship of species associated with *M. chelonae* was established by Magee & Ward (2012) by an analysis of 16S rRNA gene sequence data. The results of this analysis were realised in a three-dimensional dendrogram. There are several species and variants which have been subsequently described but, in each case, the close relationship to *M. chelonae* seen in 16S rRNA gene sequence analysis has been noted. Thus, the continued use of the term "*M. chelonae* clade" to encompass these organisms seems reasonable.

Progressively, next generation sequencing platforms have made whole genome sequencing affordable and more accessible to researchers. While the phenotypic characteristics of species may remain valuable, descriptions of both existing and newly detected microorganisms may increasingly be dependent upon whole genome signatures (Thompson *et al*., 2011; 2013). This will allow taxonomic trees to be based upon the evolutionary information contained in the whole genome sequences, so called super-trees (Daubin *et al*., 2001). However, it should be noted that phylogenetic trees based on whole genome sequences often showed close agreement with those based on the 16S rRNA gene (Thompson *et al*., 2013). Nevertheless, whole genome sequencing is already finding applications in outbreak management and the surveillance of pathogens (Deurenberg *et al*., 2017; Lalor *et al*., 2018).

1.3 *M. abscessus*, *M. chelonae* and Other Members of the *M. chelonae* Clade

1.3.1 Phenotypic characteristics

Microscopically, cells are rod shaped and Gram positive. They are acid-alcohol-fast in young cultures but in older cultures non-acid-fast forms may be seen. Even within individual species there is variation in cell form, from short to long rods (1.0 – 6.0 µm) and from narrow to thick (0.2 – 0.6 µm) with coccoid forms also possible.

Colonial morphology is equally variable. All species are considered non-chromogenic in the strict sense, that is none produce yellow/orange pigmentation either with or without exposure to light (Magee & Ward, 2012). Some strains may show white or grey colony colouration and *M. salmoniphilum* may show cream colouration in older cultures. Classically growth/no growth on MacConkey agar and on media with sodium chloride have been used in

primary separation of species, but these characteristics are unreliable. There are very few other phenotypic characteristics of value in distinguishing species within this clade.

Earlier comparative studies of phenotypic characters were confined to *M. abscessus* and *M. chelonae* along with the invalidly named species *"M. borstelense".* These studies held that *M. abscessus* could be distinguished from *M. chelonae* (and "*M. borstelense",* now seen as a species variant of *M. chelonae*). Furthermore, it was suggested that *M. abscessus* and *M. chelonae* should be considered as subspecies (Kubica *et al*., 1972). This view was supported by subsequent studies which named the subspecies as *M. chelonae* subsp. *abscessus* and *M. chelonae* subsp. *chelonae* (Silcox *et al*., 1981). This conclusion was also supported by DNA homology studies (Baess, 1982). However, at least one of the strains used in the latter study was not the Type strain (Goodfellow & Magee, 1998). Further DNA pairing studies by Lévy-Frébault *et al*., (1986) led to the proposal by Kusonoki and Ezaki (1992) to re-establish *M. abscessus* as a distinct species.

1.3.2 Molecular description

Phenotypic characters may give a guide to the identity of strains provisionally assigned to the *M. abscessus/chelonae* clade, but they are neither reliable nor accurate in conclusive species determination. Phylogenetic trees based on 16S rRNA gene sequence analyses show good separation between most clades and species represented within the rapidly-growing mycobacteria (Figures 3 and 4; Magee & Ward, 2012). However, although these analyses clearly demonstrate a distinct *M. abscessus/chelonae* clade, separation of the individual species encompassed within the clade is uncertain when based on 16S rRNA gene sequence data alone. Although 16S rRNA gene sequence data will allow the separation of *M. immunogenum* and *M. salmoniphilum* from *M. abscessus* and *M. chelonae,* DNA homology studies were necessary to firmly establish the division of *M. abscessus* from *M. chelonae* (Wilson *et al*., 2001; Whipps *et al*., 2007).

The relationship of *M. bolletii* and *M. massiliense* to *M. abscessus* introduces even more complexity. *M. bolletii* was described by Adékambi *et al*., (2006b) based on *rpoB* gene sequence data, despite 16S rRNA gene sequence data showing 100% similarity to *M. abscessus*. Similarly, *M. massiliense* was proposed by Adékambi *et al*., (2004; 2006c) on the

basis of small numbers of nucleotide variances from *M. abscessus* (in *hsp65, sodA, recA, rpoB* and ITS gene sequence analyses), although again showing 100% similarity to that species in 16S rRNA gene sequence studies. The close similarity of *M. bolletii* and *M. massiliense* led to a proposal (Leao *et al*., 2009) for the union of these two species under one epithet and the recognition of two subspecies *M. abscessus* subsp. *abscessus* and *M. abscessus* subsp. *massiliense*. However, the epithet "*bolletii*" was held to have priority over "*massiliense*" and thus Leao *et al*., (2011) proposed the correct name of the subspecies to be *Mycobacterium abscessus* subsp. *bolletii*. In an attempt to dispel the confusion surrounding the species and subspecies assigned to *M. abscessus* Tortoli *et al*., (2016) assessed the functionality of the *erm* gene. This gene they considered to be functional in "*M. bolletii*" but non-functional (although present) in "*M. massiliense*". Since the *erm* gene confers inducible resistance to macrolides (which are often used in therapeutic regimes) the authors considered the variance to be significant. Additional studies, including genome to genome distance assessment, led these workers to propose that *M. abscessus* subsp. *massiliense* be re-instated alongside *M. abscessus* subsp. *abscessus* and *M. abscessus* subsp. *bolletii*. Thus, there are presently three subspecies of *M. abscessus*.

1.3.3 The clinical significance of rapidly growing mycobacteria

There are currently over 260 validly described species and subspecies within the genus *Mycobacterium*. With the exception of the obligate pathogens *Mycobacterium leprae*; species assigned to the *M. tuberculosis* complex and possibly *M. ulcerans* (since an environmental habitat for this species is unknown), mycobacteria are either non-pathogenic to humans or opportunistic pathogens. The great majority of those species which show a propensity for opportunistic pathogenicity are slow-growing species (Goodfellow & Magee, 1998). In contrast, the majority of rapidly growing species are environmental and have not been found in a pathogenic role. Some have been incriminated as pathogens of animals, for example *M. farcinogenes* and *M. senegalense*, which are believed to be the cause of bovine farcy, a glanders like disease in African cattle which is of considerable economic significance (Chamoiseau, 1979). Some species have been detected in human clinical samples but not thought to be contributing to a clinical syndrome, for example *M. insubricum* was isolated from the sputum of five patients but was not thought to be significant despite each patient having a pre-existing respiratory condition (Tortoli *et al*., 2009). However, some species may

cause significant infections but be of rare or unusual occurrence, for example *M. setense* which was isolated from an infected soft tissue wound of the foot, resulting from the patient stepping on a nail (Lamy *et al*., 2008). Though clearly environmental in occurrence, some species may impact opportunistically on human health due to lifestyle or behavioural changes. For example, *M. cosmeticum* was named for its causal relationship to furunculosis resulting from cosmetic nail treatments; a condition in which *M. mageritense* has also been incriminated (Cooksey *et al*., 2004; Domenech *et al*., 1997; Wintrop *et al*., 2002).

Those rapidly growing mycobacteria which are consistently detected as opportunistic pathogens of humans most usually fall into or are related to either the *M. fortuitum* or *M. abscessus/chelonae* clades as defined by Magee & Ward (2012). Within the *M. fortuitum* clade, *M. fortuitum* subsp. *fortuitum* may cause injection site abscesses, *M. conceptionense* was isolated from tissue associated with post-traumatic osteitis, *M. houstonense* was isolated initially from a face wound, *M. neworleansense* from a scalp wound and *M. septicum* from a catheter tip (Adékambi *et al*., 2006a; Schinsky *et al*., 2004; Schinsky *et al*., 2000; Magee & Ward, 2012). Isolates of *Mycobacterium mucogenicum* were initially labelled as *M. chelonae*-like organisms (MCLO) before being shown to be clearly distinguishable by 16S rRNA gene sequence data and established as a separate species (Springer *et al*., 1995). Although originally associated with nosocomial outbreaks in peritoneal dialysis patients, isolates of this species have been recovered from wound infections including catheter related sepsis (Wallace *et al*., 1993). However, some of those species assigned to the *M. abscessus/chelonae* clade have become even more significant in their clinical occurrence.

1.3.4 The clinical significance of the M. abscessus/chelonae clade

The *M. abscessus/chelonae* clade currently contains 7 species: *M. abscessus*, *M. chelonae*, *M. immunogenum*, *M. salmoniphilum*, *M. franklinii*, *M. saopaulense* and *M. stephanolepidis*. *M. abscessus* is further divided into 3 subspecies, as is *M. chelonae*. Each of the species assigned to this clade is associated with water sources (even potable water) and several share with other rapidly growing species an occasional implication in wound infections. *M. salmoniphilum* has not been associated with human infections but is nevertheless a water-borne species which is a pathogen of salmonid fish (Whipps *et al*., 2007). Gene sequence analyses show a close relationship between *M. salmoniphilum* and other members of the *M.*

abscessus/chelonae clade (Figures 3, 4). *M. immunogenum*, although initially isolated from bronchoscopy wash water has been specifically linked to hypersensitivity pneumonitis associated with metalworking fluid (Wilson *et al*., 2001; Shelton *et al*., 1999). This species has also been identified as a cause of keratitis following laser eye surgery (Sampaio *et al*., 2006).

M. abscessus was originally isolated from synovium of the knee. Subsequently, it has been found as a pathogen in wound and soft tissue infections (Kusunoki & Ezaki, 1992). *M. chelonae* has also been found in wound and soft tissue infections and as a cause of cervical adenitis (Magee & Ward, 2012). Both *M. bolletii* and *M. massiliense* were initially isolated from sputum samples (Adékambi *et al*., 2004; 2006b). The lower optimum growth temperature and highly aerobic metabolism of species within this clade make them ill-fitted for deep-seated systemic infections. However, where temperature is more variable and redox potential is high, they may be encountered as colonisers. Bearing in mind the occurrence of these organisms in water and therefore in water distribution systems (Le Dantec *et al*., 2002; Thomson *et al*., 2013) it is perhaps unsurprising that they may transiently occur in the upper respiratory tract, due to contact with water droplets. Unfortunately, when this occurs in individuals with pre-existing respiratory conditions, these species may become more enduring colonisers. In some circumstances, such colonisation can lead to an exacerbation of the underlying disease process. It is in regard to this potential infectivity that *M. abscessus* and *M. chelonae* have emerged as significant species in some individuals particularly those suffering from suffering from cystic fibrosis.

As described earlier *M. abscessus* is now more correctly *M. abscessus* subsp. *abscessus*. "*M. massiliense"* is considered by some taxonomists to be synonymous with "*M. bolletii"* (more correctly named *M. abscessus* subsp. *bolletii* (Leao *et al*., 2011)). However, "*M. massiliense*" is now also named as a subspecies of *M. abscessus* – *M. abscessus* subsp. *massiliense*. Prior to the description of these new subspecies and their re-classification, *M. abscessus* (*sensu lato*) was considered to have a more serious impact on the prognosis for individuals suffering from cystic fibrosis than *M. chelonae* (Arnold *et al*., 2012). As noted earlier, it may be difficult, accurately, to separate *M. abscessus* from *M. chelonae*. It is now unclear whether strains previously assigned to either *M. abscessus* or *M. chelonae* were in fact representatives of one

of these subspecies. It is equally unclear whether these taxonomic niceties have a bearing on clinical colonisation or pathogenicity.

1.3.5 The M. abscessus/chelonae clade and cystic fibrosis

Cystic Fibrosis is a genetic disorder which is due to a mutation in the gene responsible for trans-membrane conductance of chloride and sodium ions, the cystic fibrosis transmembrane conductance regulator (CFTR). It is inherited in an autosomal recessive manner. It is particularly prevalent in people of European/Caucasian heritage. The disease affects secretory cells and results in the build-up of thick viscous secretions, notably in the lungs. In severe forms, this condition causes progressive disability and may result in early death. Respiratory problems are common and are exacerbated by frequent lung infections (Arnold *et al*., 2012). In more severe cases lung damage may be progressive and continued survival becomes dependent upon lung transplantation. There are several factors which have a bearing on suitability for transplantation but infection with *M. abscessus (sensu lato*) has been linked to a poor post-transplant prognosis and infection with this organism can become a disqualifying factor for the procedure. In contrast, *M. chelonae* infections are not held to be of similar significance (Sanguinetti *et al*., 2001; Yakrus *et al*., 2001). This variation in pathogenicity and clinical significance belies the close taxonomic relatedness of these species.
It has been estimated that the incidence of NTM in CF patients has increased from 3.3% to 22.6% over the last 20 years, with an associated increase in morbidity and mortality (Salsgiver *et al*., 2016: Martiniano *et al*., 2019: Daniel-Wayman *et al*., 2019).

As previously discussed, in many areas of the world *M. abscessus* is second only to the *M. avium* complex as a clinically important NTM isolate. Additionally, after a great deal of debate it is now accepted that *M. abscessus* contains three distinct subspecies, *M. abscessus subsp*. a*bscessus*, *M. abscessus* subsp. *massiliense*, and *M. abscessus* subsp*. bolletii* (Tortoli *et al*., 2016). The majority of clinical isolates are *M. abscessus subsp*. a*bscessus and M. abscessus* subsp*. massiliense*. *M. abscessus* subsp*. bolletii* is rare in East Asia, with a lower incidence also reported in the USA, it does, however, appear to be more prevalent in Europe (Kim *et al*., 2008; Koh *et al*., 2014). The determination by Koh *et al*., (2014) that the *erm*(41) gene is functional in *M. abscessus* strains but is not functional in the *M. massiliense* variant emphasises the value of accurate taxonomic speciation of *M. abscessus.*

The majority of clinical isolates consist of *M. abscessus subsp*. a*bscessus* and *M. abscessus* subsp. *massiliense*; *M. abscessus* subsp*. bolletii* is rare in the US and East Asia, but appears more frequent in Europe (Koh *et al*., 2014). Subspecies distinction is clinically relevant in that the presence of a functional erythromycin ribosomal methylase gene, *erm*(41), is associated with delayed or failed response to macrolide combination therapy (Koh *et al*., 2011), despite *in vitro* susceptibility to clarithromycin at 3 days. Nash *et al*., (2009) were first to demonstrate that *erm*(41), present in *M. abscessus* (but not *M. chelonae*), conferred inducible resistance (through methylation of 23S ribosomal RNA) in a majority of *M. abscessus subsp*. a*bscessus* isolates but was mutated in a smaller number of susceptible strains (Nash *et al*., 2009). The gene is also functional in most *M. abscessus* subsp*. bolletii*, conferring inducible resistance unless mutated, while deletions/truncation present in *M. abscessus* subsp. *massiliense* renders it susceptible (Nash *et al*., 2009). High-level resistance (not inducible) can also occur in all 3 subspecies due to single point mutations in the 23S rRNA gene. To detect inducible macrolide resistance, MIC determination requires incubation for up to 14 days (CLSI, 2011) and can be predicted by *erm*(41) gene sequencing (Brown-Elliott *et al*., 2015; Kim *et al*., 2016).

A multinational WGS study of more than one thousand *M. abscessus* strains from over 500 cystic fibrosis patients revealed that recent spread of 3 dominant circulating clones of *M. abscessus* (2 *M. abscessus subsp*. a*bscessus* and 1 *M. abscessus* subsp. *massiliense*) has occurred within the cystic fibrosis community across three continents, namely the US, Europe and Australia (Bryant *et al*., 2016). It is postulated that *M. abscessus* is possibly spreading between patients indirectly via fomites or desiccation resistant aerosols. The isolate clusters were noted to correlate most closely with chronic infection, poorer clinical outcomes and showed high rates of resistance to amikacin and/or macrolide antibiotics. Increased phagocytic uptake and survival in the macrophages was noted in those isolates which clustered, factors which could clearly enhance the pathogenic potential of the organism. The possibility that *M. abscessus* is evolving to become a true pulmonary pathogen cannot be ruled out (Bryant *et al*., 2016).

The *M. abscessus* colonisation of lung alveoli begins with smooth strains producing glycopeptidolipids and a biofilm, whilst in the invasive infection, “rough” mutants are responsible for the production of trehalose dimycolate, with cord formation as a consequence

of this. Using a *M. abscessus* infected CF zebrafish model Bernut *et al*., (2014), were able to demonstrate the crucial role of cording in the *in vivo* physiopathology of *M. abscessus* infection and the role of cording as a mechanism to evade the host immune response. *M. abscessus* is also intrinsically resistant to many drugs but recently alternative strategies have been investigated. Jerry *et al*., 2022 reported the successful compassionate intravenous administration of two types of mycobacteriophage (BPsD33HTH_HRM10 and D29_HRMGD40) to a male patient with treatment refractory *M. abscessus* subsp *abscessus.* The phages had been specifically engineered to improve their capacity to lyse the *M. abscessus* infecting the patient*,* subsequent to the organism having been sequenced to facilitate this.

1.4 Chemotherapy of Mycobacterial Infections

The initial focus of attention in anti-mycobacterial drug discovery and development was in response to the pressure to find treatments for classical tuberculosis. This led to a series of developments throughout the 20th Century which arrived eventually at the use of combination therapies based particularly upon rifampicin and isoniazid. Issues of patient compliance and of clinical expertise have led to problems of drug resistance. This has been exacerbated by the inter-relationship of tuberculosis with HIV infection. Nevertheless, for the most part classical tuberculosis is eminently treatable but may require extended periods of treatment.

There were a number of false starts and disappointments as this development process proceeded. For example, the focus on Streptomycin in the 1950s, which illustrated the failure of monotherapy and of the emergence of toxicity and resistance.

Monotherapy, especially over protracted periods, was likely to lead to the selection of resistant mutants and thus to treatment failure. The addition of a second drug, p-amino salicylic acid (PAS) to streptomycin in a combination therapy was aimed at overcoming the development of resistance (MRC, 1952). Isonicotinylhydrazide (isoniazid; INAH) was similarly used in combination with PAS, streptomycin or both (Crofton, J.W. 1959). Thiacetazone, despite its toxicity issues is a low-cost agent and as a result was used in various combinations with the other available agents in treatment trials in Africa and India throughout the 1960's and 1970's (Tuberculosis Chemotherapy Centre, Madras, 1966; East African/British MRC, 1973). These drugs, used in various combinations were the mainstay of treatment regimens

for tuberculosis until *circa* 1970 when more effective and better tolerated agents, notably rifamycins, became available.

The efficacy of combination therapies was further developed with the introduction of pyrazinamide following several successful trials (Hong Kong Chest Service/MRC, 1981). The mode of action of this agent remains unconfirmed but is known to relate to its conversion within the bacterial cell to pyrazinoic acid. The killing action of the drug only occurs in acidic conditions (pH 5-6) which pertain intracellularly. Pyrazinamide is used as part of a drug combination in the early months of anti-tuberculosis therapy where it has aided in the reduction of overall treatment periods, it is not used as a monotherapy (British Thoracic Society, 1984). The drug has no activity against other mycobacteria including *M. bovis, M. leprae* and non-tuberculous mycobacteria, a characteristic which has some value in species discrimination. As with rifampicin, pyrazinamide is on the WHO list of essential medicines (see above).

It had now been established that successful treatment of *Mycobacterium tuberculosis* infection needed a multi-drug combination allied to prolonged periods of drug administration.

Emergence of resistance to isoniazid and rifampicin is seen as particularly problematic since these two agents are the key elements in successful treatment of tuberculosis (Migliori *et al*., 2008). Strains exhibiting resistance to both isoniazid and rifampicin (irrespective of other drug susceptibilities) are said to cause Multi-Drug Resistant Tuberculosis (MDR-TB). The spread of such strains is now world-wide, albeit with notable regional variations in incidence (Zignol *et al*., 2006). Treatment of MDR-TB disease is based on the administration of 4 or sometimes 5 drugs chosen from a categorisation of drug classes based on efficacy and toxicity. The therapy in these cases would be over an extended period. (WHO, 1993). In 2006 the World Health Organization and the US Centers for Disease Control and Prevention Treatment drew attention to what was termed XDR-TB (Extensively Drug Resistant TB). These organisations described a severe form of disease caused by strains of *Mycobacterium tuberculosis* which were resistant not only to INH and RIF but also to at least three of the six classes of second-line anti-TB drugs (fluoroquinolones, aminoglycosides, polypeptides, thioamides, cycloserine and para-aminosalicylic acid). Although XDR-TB is relatively uncommon it was clear that the problem of drug resistance had intensified (Shah *et al*., 2007).

1.5 Treatment of Non-Tuberculous Mycobacterial Diseases

The widespread environmental occurrence of non-tuberculous mycobacteria means that the detection of strains in clinical specimens may often be considered insignificant, representing transient colonisation or commensalism. Nevertheless, these organisms can cause both superficial and systemic infective syndromes, particularly if there are pre-existing clinical conditions (Campbell & Jenkins 1995; Adjemian *et al.,* 2014; Faverio *et al*., 2021). Treatment of such infections focused initially on the slowly growing mycobacterial species notably organisms assigned to the *M. avium* complex and *M. kansasii*, *M. malmoense*, *M. xenopi*. These species were more frequently encountered, and systemic infections potentially confused with those due to *M. tuberculosis*. A number of treatment trials established that although often recalcitrant, these infections were most likely to respond to proven anti-tuberculosis drugs, notably a combination of ethambutol and rifampicin, albeit over much longer treatment periods (Wallace *et al*., 1990; BTS research committee, 1994). In the early 1990's the macrolides azithromycin and clarithromycin were investigated both as monotherapy and as adjuncts to ethambutol and rifampicin combination therapy (Wallace *et al*., 1994; Dautzenberg *et al*., 1995; Griffith *et al*., 1996). In the same period a new rifamycin semi-synthetic derivative, rifabutin, became available and was included in trials for the treatment of non-tuberculous mycobacteria (Sullam *et al*., 1994).

Rapidly growing mycobacterial species are most likely to present as skin surface infections, when surgical debridement and topical therapies may be used (De Groote & Huitt, 2006; Gonzalez-Santiago & Drage, 2015). They are also a rare but treatable cause of prosthetic joint infections (Henry *et al.,* 2016). Although transient systemic colonisation is not uncommon, this can lead to significant problems in immune-compromised patients, especially where there is underlying tissue damage such as those suffering from fibrocystic disease. In this respect *Mycobacterium abscessus* is particularly significant and is one of the most drug resistant of the rapidly growing mycobacterial species. In contrast to slowly growing nontuberculous species, these rapid-growers are resistant to conventional anti-tuberculosis regimens. The American Thoracic Society recommend a long period (at least 1 year) of a combination treatment regimen including macrocyclic lactones (macrolides; clarithromycin or azithromycin), aminoglycosides (amikacin), and β-lactams (cefoxitin or imipenem) for *M.*

abscessus infections (Griffith *et al*., 2007). However, Pasipanodya *et al*., (2017) showed that the effectiveness of such regimes was still limited with curative rates of 34-54% for newly diagnosed disease involving *M. abscessus*.

The British Thoracic Society (BTS) first published the Guideline on the 'Management of opportunistic mycobacterial infections' in the year 2000 (Subcommittee OT, 2000). The last two decades have heralded enormous improvements in the understanding of the epidemiology, diagnostic microbiology and management of non-tuberculous mycobacterial-pulmonary disease (NTM-PD). Although the incidence of NTM-PD is increasing, advances in culture techniques, access to direct molecular testing of samples, improved clinician awareness and advances in mycobacterial taxonomy are all considered responsible for this increase. We have simply become better at isolating and identifying organisms which were always present but missed due to ineffective diagnostic techniques. Diagnosis and treatment of non-tuberculous mycobacterial disease requires the combination of clinical, radiographic and microbiology data.

Updated guidelines published in 2017 (Haworth *et al.,* 2017) and drawing on the American Thoracic Society statement (Griffith *et al.,* 2007) define the clinical and microbiological criteria for diagnosing non-tuberculous mycobacterial lung disease. Clinically, nodular or cavitary opacities and or multifocal bronchiectasis must be noted on CT scan and there must be appropriate exclusion of other diagnoses; both criteria must be satisfied.
Microbiological diagnostic criteria for NTM-PD, unlike tuberculosis, stipulate that a patient must have two or more positive sputum samples, or one positive bronchial wash/lavage or compatible histopathological findings with one positive culture of the same NTM species.

In 2020 The American Thoracic Society (ATS), European Respiratory Society (ERS), European Society of Clinical Microbiology and Infectious Diseases (ESCMID), and Infectious Diseases Society of America (IDSA) jointly sponsored the development of a guideline to update the treatment recommendations for nontuberculous mycobacterial pulmonary disease (NTM-PD) in adults (Griffith *et al.,* 2007). Daley *et al.,* 2020 carried out a systematic review of treatment of NTM-PD in adults (without HIV infection or cystic fibrosis) which centred around each of twenty two PICO (Population, Intervention, Comparator, Outcome) questions. Recommendations were formulated and graded using the GRADE approach (Grading of

Recommendations Assessment, Development, and Evaluation). In total, thirty-one evidence-based recommendations regarding treatment of NTM-PD were proposed, covering laboratory isolation, identification and susceptibility methods and proposed treatment regimens based on the species isolated and the extent of pulmonary disease. Where a patient meets the diagnostic criteria described previously and which have not changed, it is recommended that treatment is initiated with no period of watchful waiting. Treatment regimens should be based upon susceptibility testing of isolates over empiric therapy for patients with disease caused by MAC and *M. kansasii*. In the case of *M. xenopi* disease however, the committee members felt that there was insufficient evidence to make a recommendation for or against susceptibility-based treatment. *M. abscessus* pulmonary disease susceptibility-based treatment for macrolides and amikacin over empiric therapy was recommended. For macrolides, a 14-day incubation and/or sequencing of the *erm* (41) gene should be performed to evaluate for potential inducible macrolide resistance

Accurate identification and speciation of NTM is important as it can predict the clinical relevance of an isolate (van Ingen *et al*., 2009) and equally importantly, direct the treatment regimen. Molecular identification is better and can be achieved using line probe assays or genetic sequencing. Line probe assays have been widely used over the past two decades as they were easy for diagnostic centres to implement, however, they lack discriminatory power which can lead to misidentification of isolates (van Ingen *et al*., 2010; Tortoli *et al*., 2010). Genetic sequencing provides a more robust identification, often to sub species level. Several target genes have been described, e.g., 16S rRNA, hsp65, and rpoB. 16S rRNA gene sequencing alone offers limited discriminatory power, particularly for the *M. chelonae*/*abscessus* group (van Ingen *et al*., 2010). The hsp65 and rpoB genes are more discriminatory and using all or a combination of these gene targets can offer a higher level of sub-species discrimination particularly for *M. abscessus* (Zelazny *et al*., 2009).

Given the very clear advantages offered by molecular testing of NTM isolates in directing treatment and patient outcomes, it is not surprising that whole genome sequencing (WGS) has the potential to improve things further. A large prospective study carried out in 2018 (Quan *et* al., 2018) demonstrated that WGS predicted species and drug susceptibility data with great accuracy, achieving very high agreement with existing diagnostic tests for both

mycobacterial species identification (96.0%) and MTBC first-line drug resistance detection, 99.3% versus Line Probe Assay and 99.2% versus phenotypic testing.

Resistance may be due to one or more of several mechanisms. Intrinsic resistance may be derived from low permeability of the *M. abscessus* cell envelope, as well as to drug export systems. The mycobacterial transcriptional regulator *whi*B7 has been shown to contribute to intrinsic strain resistance; activating its own expression and many drug resistance genes when stimulated by exposure to antibiotics (Burian *et al*., 2012). In *M. abscessus whi*B7 is induced by exposure to the ribosome-targeting antibiotics erythromycin, clarithromycin, amikacin, tetracycline, and spectinomycin. However, deletion of the *whi*B7 reverses this effect and allows susceptibility to each of these agents (Hurst-Hess *et al*., 2017). These authors also showed that *whi*B7 induces the gene eis2 which is a factor in the resistance of *M. abscessus* to amikacin. Recent research has shown that the expression of numerous enzymes which either modify the drug-target or the drug itself, is a significant factor in mycobacterial resistance to several classes of antibiotics. Luthra *et al*., (2018) drew attention to an erythromycin ribosome methyltransferase, two aminoglycoside acetyltransferases, an aminoglycoside phosphotransferase, a rifamycin ADP-ribosyl-transferase, a β-lactamase and a monooxygenase.

1.5.1 Macrolides – Clarithromycin/Azithromycin

Rapidly growing mycobacteria, including organisms of the *M. abscessus/chelonae* clade have been shown to have some susceptibility to the macrolides azithromycin and clarithromycin. The current American Thoracic Society (ATS), European Respiratory Society (ERS), European Society of Clinical Microbiology and Infectious Diseases (ESCMID), and Infectious Diseases Society of America (IDSA) guidelines for treatment of NTM-PD recommend the use of one of these agents as part of treatment regimens for these organisms (Davey *et al*., 2020). Clarithromycin has previously been more commonly used despite some evidence that azithromycin would be a better choice (as noted below). Current recommendations recommend azithromycin be used in patients with newly diagnosed macrolide-susceptible MAC pulmonary disease Increasingly, resistance to macrolides has been detected, notably inducible resistance due to the erythromycin ribosomal methylase gene *erm*(41) (Nash *et al*., 2009). The gene, modifies the binding site for macrolide antibiotics, resulting in the inducible

macrolide resistance (Koh *et al*., 2014). Bastian *et al*., (2011) looked for *erm*(41) in the three variants of *M. abscessus* (i.e. *M. abscessus*, *M. bolletii* and *M. massiliense* - identified on a molecular basis). Additionally, mutations in *rrl* (23S rRNA gene) known to confer acquired clarithromycin resistance were looked for. The *erm*(41) gene was detected in all strains but in "*M. massiliense"* there were two deletions in the gene and these strains were clarithromycin susceptible - excepting those with *rrl* mutations which were clarithromycin resistant. Koh *et al*., (2014) demonstrated that the *erm*(41) gene is functional in *M. abscessus* (*sensu stricto*) but is non-functional in "*M massiliense*". Thus, this latter species/subspecies does not display inducible macrolide resistance. Treatment success rates with macrolide-based antibiotic treatment are thought to be much higher in patients with "*M. massiliense*" infections than in those infected with *M. abscessus*.

Precise speciation of *M. abscessus* complex strains is important not just on taxonomic grounds but in predicting antibiotic susceptibilities and subsequent patient outcome. This conclusion was emphasized by Choi *et al*., (2012) who also found that clarithromycin induces *erm*(41) to a significantly greater extent than azithromycin. The phenotypic study by Park *et al*., (2017) showed that sustained culture conversion after macrolide treatment was more common in patients with "*M. massiliense*" infections than those with *M. abscessus* infections. It also agreed that azithromycin rather than clarithromycin was a predictor of sustained culture conversion.

Kim *et al*., (2016) describe a case of pulmonary disease caused by a strain of *M. abscessus* subsp. *abscessus* showing clarithromycin susceptibility despite presence of the *erm*(41) gene. The susceptibility was due to a non-functional *erm*(41) allele which carried a C-to-T mutation at position 19 in the gene. This, they conclude, strengthens the case for genetic assessment of strains to guide therapy. This concept was further developed by Guo *et al*., (2018) using clarithromycin susceptibility genotyping of *M. abscessus* strains. Although the clinical results were somewhat predictable (patients infected with clarithromycin-sensitive and -resistant *M. abscessus* genotypes differed significantly in clarithromycin-based combination treatment outcomes) it is important to note that these outcomes were predicted from genotyping.

1.5.2. Aminoglycosides

Although not involved in primary treatment of *M. tuberculosis*, the 2-deoxystreptamine aminoglycosides (kanamycin, amikacin, gentamicin and tobramycin) have found usage in the treatment of multidrug-resistant strains and in nontuberculous mycobacterial infection (Sander & Böttger, 1999). These agents target the rRNA operon leading to inhibition of protein synthesis. Prammananan *et al*., (1998) showed that an adenine to guanine substitution within the 16S rRNA gene of *M. abscessus* is responsible for a high level of resistance to aminoglycosides. Despite this there are recommendations to combine clarithromycin with an aminoglycoside and another injectable (such as imipenem) although results are variable (Griffith *et al*., 2007)

1.5.3. β-Lactams

Mycobacterium abscessus is highly resistant to most β-lactam antibiotics. Such resistance can potentially be overcome by the concomitant use of the β-lactamase inhibitor clavulanic acid. However, Soroka *et al*., (2014) established the presence of a clavulanate-insensitive broad-spectrum β-lactamase in *M. abscessus* which limits the *in-vivo* efficacy of these agents. Despite this imipenem and cefoxitin have been found to have moderate activity *in-vitro* against *M. abscessus* and these agents have been used to treat infections with this organism. Furthermore, it is believed that L, D-transpeptidases, (which are involved in the synthesis of cell wall peptidoglycan), are inhibited by the carbapenem class of β-lactams (Kumar *et al*., 2017a). Recent studies have demonstrated that inhibition of these enzymes determines their activity against *Mycobacterium tuberculosis* Kumar *et al*., (2017b). Kumar *et al*., (2017a) demonstrated that two L, D-transpeptidases in *M. abscessus*, namely, LdtMab1 and LdtMab2 were inhibited by the carbapenem and cephalosporin, but not penicillin, subclasses of β-lactams. These authors state that, contrary to the commonly held belief that combination therapy with β-lactams is redundant, doripenem and cefdinir exhibit synergy against both pan-susceptible *M. abscessus* and clinical isolates shown to be resistant to most antibiotics. This suggests that dual β-lactam therapy has potential for the treatment of *M. abscessus*.

Whilst studies of individual β-lactam antibiotics have failed to show significant efficacy against *M. abscessus* when combined with β-lactamase inhibitors (BLI), β-lactam antibiotics may still prove valuable (Story-Roller *et al*., 2018). Recently, two novel carbapenem-BLI combinations have been developed. These are meropenem-vaborbactam, which was recently FDA-approved for use against gram-negative organisms, and imipenem-relebactam, which is currently in phase II clinical trials (Zhanel *et al*., 2018). There are no published studies assessing such combinations against *M. abscessus*, but there is some potential for clinical use and further studies are needed.

1.5.4. Fluoroquinolones

Fluoroquinolones are broad spectrum antibiotics developed from a 4-Quinolone basic molecule with the addition of a fluorine atom (Andersson & MacGowan, 2003). They are active against most strains assigned to the *Mycobacterium tuberculosis* complex and some nontuberculous mycobacterial species (Jacobs, 1999). This class of chemotherapeutic agents includes ciprofloxacin, ofloxacin, sparfloxacin and moxifloxacin (Zhanel *et al*., 2006). Resistance to these agents is often due to mutations in the DNA-gyrase complex (Aubry *et al*., 2006). However, *Mycobacterium abscessus* displays resistance to fluoroquinolones not always explained by such mutations suggesting that other mechanisms are involved (de Moura *et al*., 2012; Kim *et al*., 2016).

1.5.5 Rifamycins

Although of considerable value in the treatment of *M. tuberculosis* and some slowly growing mycobacterial species, rifamycins show no impact on *M. abscessus*. Respiratory arsenate reductase (Arr), which catalyzes ADP-ribosylation of rifamycins, is one mechanism conferring resistance to these agents. Genes encoding for Arr enzymes are widely distributed in the genomes of bacteria (Baysarowich *et al*., 2008). These authors analyzed three representative Arr enzymes from bacterial sources and showed them to have drug resistance capacity *in-vitro* and *in-vivo*. They found that the 3D structure of an orthologue from *Mycobacterium smegmatis* revealed structural homology with ADP-ribosyl-transferases strengthening the view that Arr enzyme activity is of particular significance in rifampicin drug resistance.

Rominski *et al*., (2017) postulated that intrinsic resistance to rifampicin in *M. abscessus* was mediated by the ADP-ribosyltransferase (Arr_*Mab*), encoded by the protein *MAB_0591*.

However, working with *M. smegmatis* Combrink *et al*., (2007) showed that the antimicrobial activity of rifampicin was significantly increased by a series of 3-morpholino rifamycins in which the C25 acetate group was replaced by a carbamate group. These workers also suggest that relatively large groups attached to the rifamycin core *via* a C25 carbamate linkage will prevent inactivation *via* ribosylation of the C23 alcohol as catalyzed by the endogenous rifampin ADP-ribosyl transferase of *M. smegmatis*. Such studies hold out the hope that modifications of the rifamycin core may protect against enzymic resistance whilst retaining anti-mycobacterial activity. In a variation of this approach, Kanglemycin, a newly described antibiotic has been shown to bind bacterial RNA polymerase at the rifampicin-binding pocket while maintaining potency against rifampicin resistant strains containing RNA polymerases (Mosaei *et al*., 2018).

1.5.6.Bedaquiline

Bedaquiline (a diarylquinoline) is a new anti-tuberculosis agent which acts by blocking ATP synthesis (Koul *et al*., 2007). Use of this agent in clinical practice is restricted due to its potential as a treatment of last resort for multidrug-resistant tuberculosis (MDR-TB) and extremely resistant (XDR-TB) strains of *M. tuberculosis* (World Health Organization, 2014). However, in-vitro studies offer the possibility of bedaquiline usage as a salvage therapy for clinically critical *M. abscessus* infections (Philley *et al*., 2015). There are reasons for caution since Pang *et al*., (2017) found a significant subset of *M. abscessus* strains to show in-vitro resistance to bedaquiline (66/381). As with *M. tuberculosis* strains, resistance to bedaquiline in *M. abscessus* may involve mutations in the *atpE* gene (Koul *et al*., 2007). This leads to the view that use of the agent for the treatment of *M. abscessus* infections may be discouraged since it may impact on the treatment of MDR/XDR-TB.

The discovery of new drug targets leading to the development of new treatment agents for *M. abscessus* infections is urgently needed. Improved knowledge of the mechanisms of *M. abscessus* drug resistance could improve drug selection and may promote development of novel antimicrobials.

1.6 Project Aims

The aim of this research is to further the understanding of the clinical relevance of the *Mycobacterium abscessus*/*chelonae* clade through sequencing studies, with particular emphasis on the discovery of putative variations in virulence and antibiotic resistance which could explain the differences in clinical outcomes.

This is achieved through the following objectives:

- Extract, amplify and obtain a WGS of a clinical isolate of *Mycobacterium chelonae*
- Compare this, *M. chelonae* HPA 006 genome sequence with *Mycobacterium abscessus* (CIP 104536T = ATCC 19977T)
- Identify genomic regions which explain putative variations in virulence and antibiotic resistance between the representatives of these species
- Clarify the taxonomy of the organisms assigned to the *Mycobacterium abscessus*/*chelonae* clade

Chapter 2. Whole Genome Sequencing and Assembly of *M. chelonae* HPA 006

2.1 Genomic Sequencing: History, Technologies and Applications

2.1.1 Introduction

Sequencing allows the order and identity of bases in a fragment of DNA to be determined. The determination of part or whole genomes from animals, plants and prokaryote organisms continues to be a scientific research target. The information gained from The Human Genome Project (HGP) has helped in understanding genetic variation, given insights into genetic defects and helped identify the genetic causes of predisposition to disease (Collins and McKusick 2001., Collins, 2003). There are several projects which are aimed specifically at understanding the connection between human genetic variation and health, including the HapMap project and its successor the 1000 Genomes Project (The International HapMap Consortium 2005; The International HapMap3 Consortium, 2010; The 1000 Genomes Project Consortium, 2012).

It is hoped that research into the genomics of various crop plants will lead to improved yields, thus benefitting food production worldwide (Feuillet *et al*., 2011). The cloning of a potato late blight–resistance gene using RenSeq (resistance (R) gene sequence capture) and SMRT sequencing (single molecule real time sequencing) allows rapid cloning of multiple R genes for engineering pathogen-resistant crops, in this case, a potato resistant to *Phytophthora infestans*, one of the main pathogens in the agricultural sector (Witek *et al*., 2016).

The determination of the whole genome sequences of some domesticated animals may prove to be of similar value (Elsik *et al*., 2009; Canavez *et al*., 2012). The capability to map DNA sequences was also seen as valuable in more arcane sciences such as the systematics of prokaryotes, notably bacteria. Such research can also be of great value, since it can allow clearer elucidation of species, and thus aid the understanding of bacterial pathogenicity. The first whole genome sequence of a bacterium to be described was that

of *Haemophilus influenzae* (Fleischmann *et al*., 1995). This was followed by the publication of many more sequences, each intended to lead to better detection methods of significant microorganisms, more understanding of virulence markers and better targeted vaccination and chemotherapy (Fraser *et al*., 2002).

Improvements in sequencing chemistry, platforms and downstream bioinformatics applications used to analyse results, have ensured that sequencing of DNA can take place on a larger scale and at a lower cost, both of which are necessary for these technologies to have meaningful applications in the clinical and biological fields (Schwarze *et al*., 2019). However, the term "lower cost" is an important distinction from low cost. There has been a great deal of expectation surrounding the $1000 genome, however, a recent study (Schwarze *et al*,. 2019) suggests that this aspiration likely could only be achieved when the consumable costs alone are assessed, with insufficient consideration of the true costs of the full sequencing process. The group studied the cost of genome sequencing for a cancer case (tumour and germline sample) and a rare disease trio case (three samples). Costs were £6840.85 (£3420 per genome) for cancer with sequencing being 76% of these costs and £7050 for the three rare disease samples (£2350 per genome), with sequencing being 79% of these costs.

Sequencing was carried out on a medium throughput Illumina HiSeq 4000 in a laboratory that processes around 400 samples per year. The authors acknowledge that they could not compare their costs with an alternative higher throughput sequencing platform, however, the level of detail presented in their costings could be used as a benchmark for future comparisons. They conclude that high throughput services which could be delivered from a national scale facility combined with meaningful reductions in laboratory consumable costs would have the greatest impact on reducing costs (a personal observation is that such mega labs supported the community COVID testing during the recent pandemic and therefore proof of concept has been shown).

2.1.2 Introduction of DNA sequencing

Starting with the introduction of the Sanger chain-terminating method over forty years ago, and followed by the steady development of new approaches, sequencing has created

a paradigm shift across the field of biological science, enabling sweeping advances in the way in which we understand disease.

For almost thirty years, until the first of the next-generation technologies became widely available (Roche 454; Margulies *et al*., 2005), DNA sequence determination was dominated by the methods developed by Sanger and colleagues (Sanger, *et al*., 1977). Though modified in various ways to improve practicality and speed, the basic principles of the Sanger methodology remain essentially unchanged.

The Sanger method utilises specific di-deoxynucleotides that are used to disrupt the DNA synthesis reaction, each has a base-specific radioactive isotope label. This means that after gel electrophoresis, the DNA sequences of the sample can be determined according to the position of the electrophoretic band. This method is routinely known as chain termination sequencing or the dideoxy sequencing method

The first step is to extract DNA from the organism under study to give multiple fragments of DNA. Each of these fragments can be multiplied (cloned) many times. As will be apparent, massively parallel sequencing is necessary to achieve accurate and effective definition of each fragment. A key element of the Sanger approach is an *in vivo* cloning stage (of DNA fragments) carried out within host bacterial cells. In this case the vector which allows the cloning is a plasmid from a rapidly growing bacterium, very often a strain of *Escherichia coli* (Cohen and Chang, 1973; Stoker, *et al*., 1982). Plasmids are self-replicating within a bacterial cell, producing a stable and characteristic number of copies. Some bacteria produce high numbers of plasmids which make them ideal cloning vectors. Thus, a DNA segment inserted into a plasmid which is then replaced within the bacterial cell will produce large numbers of cloned copies.

The next step is to determine the sequence of each single-strand fragment and in the Sanger method this is done by chain termination. The basis of DNA replication in the laboratory is the addition of deoxynucleotides (dNTPs; deoxynucleotide tri-phosphates) to the 3’OH (3-prime) end-group of the developing chain, catalysed by DNA polymerase. The Sanger method is based on termination of these developing chains using di-deoxynucleotides (ddNTPs). These molecules lack a 3’OH group; thus, when a ddNTP is

added to a developing chain it terminates the reaction since it cannot receive any more dNTPs.

It is necessary to know at least some part (however short) of the genome sequence. This information may be laboriously obtained by chemical and/or radiographic analysis. With this knowledge, the first primer set can be designed. In each set of 4 reaction tubes, primer, polymerase and dNTPs of all 4 nucleotides (Adenine, Cytosine, Guanine and Thymine) are present. Also present in each tube is one specific di-deoxynucleotide, (e.g., dd-Adenosine tri-phosphate; ddATP). In this example replication would proceed normally with base after base (A, T, G, C) being added to the developing chain. The polymerase supports the addition of one molecule of dATP as its turn arrives, until a ddATP is randomly added, at which point extension of that fragment would cease. So, in the tube in this example, there will be DNA strands of varying lengths but all ending in A. In the other 3 tubes, there will be an equivalent process involved, producing lots of DNA strands, again of varying length but all ending in T, G or C.

Now the contents of all 4 tubes can be run out on a gel, in 4 lanes side by side. The smallest fragment will be that which has built the fewest number of nucleotides before being stopped by a di-deoxynucleotide. This fragment will run furthest along the gel, and this will be the start of the sequence (e.g., A). The next smallest fragment (reaching the next furthest position on the gel) will indicate the second base in the sequence (e.g., T) and so on until the sequence of this fragment is complete. Now the last group of nucleotides making up this fragment can be used to design a primer to extend the chain through another sequence, and this is repeated. If several short sequences are known, then this process can be simultaneously repeated with several primers. Multiple extracted fragments in multiple groups of 4 tubes will result in many single strand sequences. These must now be manually aligned using their classical coupling arrangements with a complimentary strand to form double stranded DNA (i.e., C with G, and A with T). Thus, fragment by fragment, the whole genome sequence is gradually elucidated.

Clearly, the manual process is highly laborious and determination of whole genomes a time-consuming process. Despite automation of the Sanger method in the late 1980s (referred to as First Generation sequencing), sequencing of large genomes remained a

costly process requiring a high investment in equipment to achieve reasonable timescales for data generation. This limitation led to the development of large-scale sequencing centres such as the Wellcome Trust Sanger Institute in Cambridge. Additionally, centralised facilities such as the European Bioinformatics Institute provided data analysis support and aided in the collation of the resulting information (http://www.ebi.ac.uk).

Genomes sequenced using the Sanger method included many of medical and economic importance notably, in 2001, the human genome. Where genome sequences became available, scientific breakthroughs followed. However, time and cost meant that only a limited number of entire genomes became available and the published catalogue of whole genomes increased only slowly. Without the resources of the larger centres, smaller units could only focus on individual genes and access to sequencing at the whole genome level was not available to most scientists. The automated Sanger methodology is often referred to as first generation sequencing (which is based on electrophoretic separation of chain-termination products produced in individual sequencing reactions).

In 1987 the first commercial platform from Applied Biosystems, Inc. (ABI) utilising fluorescent DNA sequencing technology became available (one fluor per nucleotide reaction). The platform had been created in Leroy Hood's laboratory at the California Institute of Technology (Smith *et al*., 1986). The move away from radiolabeled dATP to fluorescently labeled primers automated the result analysis, making this process much less laborious and also reduced the errors which can accompany manual interpretation.

2.2. Massively Parallel Sequencing Technologies

The sequencing technologies which appeared from 2005 onwards are referred to as next generation sequencing (NGS) and third generation sequencing technologies (TGST) (Figure 5 Mardis, 2011). The fact that these technologies are performing hundreds of thousands (or hundreds of millions) of sequencing reactions at the same time has led to the term massively parallel sequencing, a term considered by many to be a much more accurate reflection of what takes place. Additionally, each technology is further differentiated by the sequencing read length produced. NGS technologies provide short read lengths (whilst third generation technologies are noted for their longer read lengths.

To more easily differentiate the technologies, I will continue to use the terms NGS and TGST.

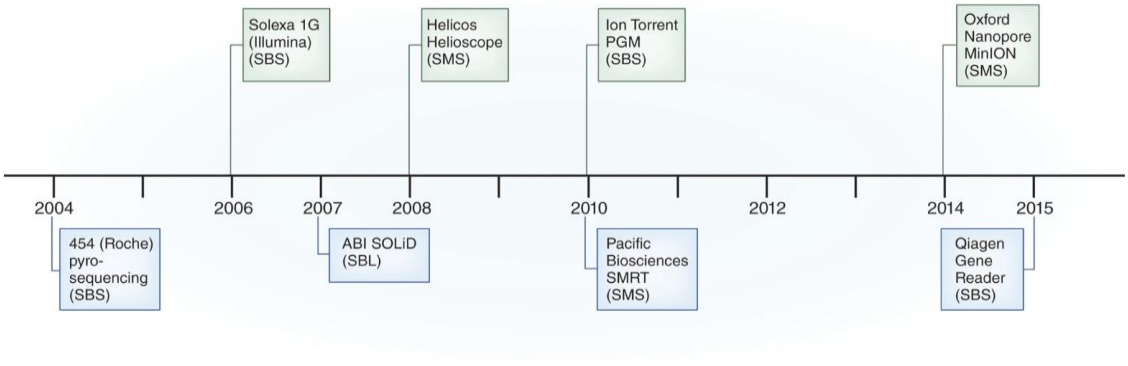


Figure 5. Timeline of Next Generation Sequencing instruments introduced from 2005-2015 (Reproduced from Mardis, 2011).

2.2.1 Next generation sequencing technology

The first NGS platforms performed sequencing by synthesis (SBS). SBS strategies may be classified as either single molecule–based or combination based (involving the sequencing of multiple identical copies of a DNA molecule, which, as previously stated, are typically amplified together on a solid support e.g., a bead). The first of these to be commercially available was the Roche/454 platform (Margulies *et al*., 2005), but this was quickly followed by other commercial systems such as ABI (Illumina), SOLiD (Applied Biosystems), and Ion Torrent (Life Technologies, now ThermoFisher Scientific).

Sequencing by ligation does not use a DNA polymerase to create a second strand, instead, a known probe sequence that is bound to a fluorophore hybridises to a DNA fragment which is then ligated to an adjacent oligonucleotide for imaging. The emission spectrum of the fluorophore is indicative of the base identity complementary to specific positions within the probe (Goodwin *et al*., 2016).

There are several steps that are key to successful sequencing using NGS technologies,

- Extraction and purification of the DNA to produce the DNA template

The most common methods for extracting good quality DNA are chemical, physical or enzymatic

- Fragmentation of the DNA

Samples of the purified DNA are sheared into short fragments. This can be achieved mechanically using sonication, shearing via nebulisation or enzymatically (Knierim *et al*., 2011).

- Library preparation and PCR amplification of target sequence

One of the main differences in NGS compared to the Sanger method is the way in which the sequencing library is constructed. DNA ligase is used to covalently add synthetic DNA adapters to the end of each fragment (Mardis 2013; Head *et al*., 2014). The adapter sequences are specifically designed to interact with the chosen NGS platform and they are immobilised onto a solid surface This can be the surface of a flow cell, as found in Illumina platforms or a bead, as found in the Ion Torrent platform.

There are three techniques used to amplify DNA fragments prior to sequencing, two are PCR based, namely, emulsion PCR and bridge amplification.. The third amplification technology, DNA nanoball generation, solves this potential problem because copies generated are all made from the original DNA template.

Emulsion PCR was developed by Jonathan Rothberg at the 454 Life Sciences Corporation (Margulies *et al*., 2005). The basic premise of emulsion PCR is the dilution of template DNA molecules in water droplets in a water-in-oil emulsion. The best results are obtained if each droplet contains a single template DNA molecule and functions as a micro-PCR reactor (Kanagal-Shamanna, 2016). Platforms which utilise emulsion PCR are 454 Pyrosequencing (Roche 454; Margulies *et al*., 2005), Ion Torrent (Thermofisher Scientific) and Qiagen GeneReader (Qiagen).

Bridge amplification takes place on a flow cell coated by two types of oligonucleotides complementary to the two adapter oligonucleotides attached to the DNA template strand. Essentially, bridge amplification permits the formation of dense groups of amplified fragments. This allows a fluorescent signal to be detected every time a single

dNTP is added sequentially as sequencing-by-synthesis proceeds. Over time, the number of groups being read grows.

DNA nanoball generation uses rolling circle replication to amplify small fragments of genomic DNA into DNA nanoballs. Fluorescent nucleotides bind to complementary nucleotides and are then polymerised to anchor sequences bound to known sequences on the DNA template. The base order is determined via the fluorescence of the bound nucleotides. After purchasing Complete Genomics, the Beijing Genomics Institute (BGI) refined DNA nanoball sequencing to sequence nucleotide samples on their own platform (Huang *et al.,* 2017).

The Ion-Torrent Personal Genome Machine (PGM), one of two sequencing platforms used in this project, is an example of a next generation sequencing system. It was first introduced commercially in 2011 (Rothberg *et al*., 2011). Ion Torrent uses semi-conductor chip technology containing millions of tiny micro-wells located under a sensing pixelated layer similar to the complementary metal oxide semiconductor (CMOS) light sensor chip (Quail *et al*., 2012). The CMOS sensor has been modified and paired with an Ion Sensitive Field Effect Transistor (ISFET) sensor to sense chemical changes instead of changes in light. It therefore does not detect the identity of incorporated bases by using chemical luminescence dyes during the sequencing process, instead, the PGM utilises the fact that the addition of a dNTP to a DNA polymer results in the release of a hydrogen ion. The resulting pH change from the released hydrogen ion is detected using semiconductors which are capable of directly translating chemical signals into digital information. (Rothberg *et al*., 2011). Many millions of such changes can be measured simultaneously to determine the sequence of each fragment. Each of the four nucleotides (dNTPs) is added iteratively onto a massively parallel semiconductor-sensing device to ensure only one dNTP will be responsible for the electrical signal.

Target DNA is fragmented, ligated to adapters, and the adaptor-ligated libraries undergo clonal amplification which results in the molecule being bound onto a solid surface in the form of a bead. The template DNA bearing beads are then PCR amplified to create a set of identical clones. A magnetic bead-based enrichment process selects template-carrying

beads This enrichment process serves several purposes; it increases the amount of template DNA; selects the molecules which have adapters successfully ligated to them; facilitates incorporation of oligonucleotide sequences for the attachment of the library to the beads and allows the addition of indices for multiplexing technique (Rothberg *et al*., 2011).

During the sequencing run, each of the four nucleotide bases is introduced sequentially. A nucleotide complementary to the base on the template is incorporated into the growing genome strand by DNA polymerase, signal processing software measures incorporation and filters out low-accuracy readings (Rothberg *et al*., 2011).

Finally, per-base quality values are predicted using an adaptation of the Phred method (Ewing & Green, 1998) which quantifies the concordance between the phasing model predictions and the observed signal. These *ab initio* scores track closely with post-alignment derived quality scores, and are used to trim back low-quality sequence from the 3' end of a read (Rothberg *et al*., 2011).

Figure 6 provides a representation of the Ion Torrent sequencing semiconductor, illustrating the technology. of the wafer, die and chip packaging as well as the sensor, well and chip architecture.

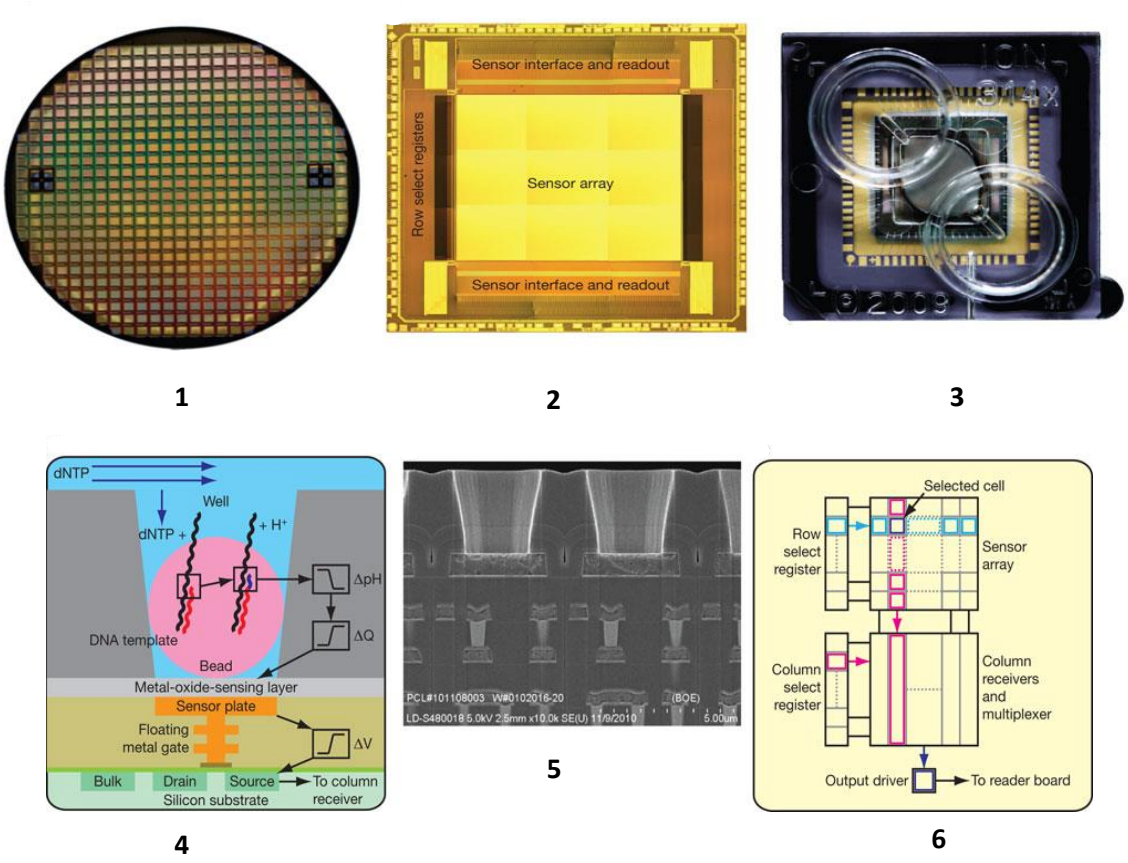


Figure 6. Ion Torrent sequencing semiconductor: technology. wafer, die and chip packaging (1-3); sensor, well and chip architecture (4-6).

1. Fabricated CMOS 8" wafer containing approximately 200 individual functional ion sensor die.
2. Underlying functional electronic elements and sensors board,
3. Image of die within ceramic package with visible moulded fluidic lid which allows addition of sequencing reagents.
4. A schematic representation of the technology behind semiconductor sequencing with DNA template releasing hydrogen ions
resulting in a pH change in the well. This signal is transformed into potential voltage which sensed by the under lying sensor and electronics,
5. Electron micrograph revealing the alignment of the wells over the ISFET metal sensor plate and the underlying electronic layers
6. Schematic diagram for the sensor detection workflow in two-dimensional array.
Reproduced from Rothberg *et al*., 2011

2.2.2 Third generation sequencing technology (TGST)

Whereas NGS platforms proceed in a stepwise fashion producing output from amplified library fragments, TGST are capable of sequencing single molecules (single molecule sequencing, SMS) without the requirement for DNA amplification and it is this feature which sets them apart from the technologies that went before.

SMS technology was first developed by Stephen Quake and colleagues (Braslavsky *et al.,* 2003; Harris *et al.,* 2008) and later commercialised by Helicos in 2008 (Helicos BioSciences Corporation, Cambridge MA, USA). Despite being the first system of its kind to allow sequencing of DNA without the need for amplification, this platform is no longer available. Currently SMS platforms are available from Oxford Nanopore Technologies (ONT) (Oxford Nanopore Technologies, Oxford, UK) and Pacific Biosciences (Pacific Biosciences of California, Inc.). The latter are noted for the ability to produce reads up to and over 10Kb in length, an attribute which are beneficial in *de novo* genome assembly (van Dijk *et al.,* 2014). There has been a great deal of anticipation over the last 6-8 years as to what the nanopore system would offer sequencing.

The second sequencing platform used in this project is the MinION. The MinION was the first commercial third generation sequencer which was nanopore based to be released. The portability of the MinION, (a USB device which is the size of a mobile phone) coupled with the fact that it generates long read sequence data and has a fast run time makes it an accessible option for smaller groups to employ. MinION technology have been used alone to generate *de novo* bacterial reference sequences (Loman *et al*., 2015). MinION sequencers have also been deployed in the field, notably during the Ebola outbreak in W Africa (Loman *et al*., 2016).

The technology has not been without issues, Tyler *et al.*, 2018 evaluated the MinION for microbial whole genome sequencing application, reviewing flow cell quality, sequencing yield, mapping, sequencing depth and *de novo* assembly (amongst other parameters). They determined two key things, changes to the MinION software and a simplifying of the ID protocols did improve overall performance. A recent review by Kerkhof, 2021 suggests that MinION had made major improvements in the five years up to 2021 (Kerkhof, 2021).

MinION has also been shown to be a useful adjunct to help generate scaffolds to map Illumina reads. Combining the long read length with the accuracy and depth of the short reads generated by the Illumina technology (Madoui *et al.,* 2015).

2.2.3 Long read lengths versus short read lengths and the impact of this on sequencing results

Different sequencing platforms generate different read lengths. Both short-read sequencing and long-read sequencing have positive and negative aspects to the results they generate, which can also be influenced by the research goal.

Platforms that carry out sequencing by synthesis or ligation (using DNA polymerase or ligase enzymes) to extend many DNA strands in parallel, will produce short reads, generally from 75 to 700 base pairs in length depending upon the platform. Short-read sequencing is a low-cost, accurate method that is supported by numerous analysis tools and workflows, however, for all but the least complex and smallest genomes e.g., viral genomes, one must align the reads generated by SBS and SBL to a reference genome before assembly can be attempted. This contrasts with assembling the genome by using shared sequence overlaps, as is done for Sanger sequencing reads. It was easy to predict therefore that the relative ease with which sequencing can be undertaken has necessarily created an expanding series of bioinformatics/computational tools to support the analysis of end results.

Sequencing technologies which produce longer read lengths are based on single molecule sequencing (SMS) and can read longer lengths of the DNA, RNA, gene or protein being studied, generally, between 5,000 and 30,000 base pairs. This helps tackle the main issue presented by short-read sequencing, spanning repeat regions. Since they sequence a single molecule, it also helps to exclude amplification bias and generates enough length to overlap one sequence with another, resulting in improved sequence assembly, but typically has a higher error rate.

2.3 Genome Assembly and Annotation: Tools and Challenges

2.3.1 Background

Once the sequencing process is complete the next challenge is assembly and annotation of the proteins and genes. Assembly is the process of putting the large number of short DNA sequences generated during sequencing back together again to recreate the original chromosome(s) from which the DNA originated. In terms of bacterial genome sequencing, the goal is a complete contiguous DNA sequence of the bacterial chromosome.

There are two types of genome assembly, *de novo*, which refers to assembly from scratch without the help of a reference genome and a reference-based alignment (or mapping to a reference genome), the latter being a digital nucleic acid sequence held in a database as an exemplar of the gene set of an individual organism within a species. There are several databases with such deposited sequences available, e.g., National Centre for Biotechnology Information (NCBI). NCBI does not generate these sequences but rather is a repository for the sequences which are deposited by individual researchers and sequencing centres. The assemblies are made publicly available by being submitted to NCBI via GenBank (The National Institutes of Health, genetic sequence database) or to another member of the International Nucleotide Sequence Database Collaboration (INSDC).

Neither NGS nor TGST has the capability to sequence the genomes of prokaryote and eukaryote organisms in one complete string. This, therefore, poses the problem of how the reads generated from either technology can be assembled to produce a complete and accurate genome assembly.

The factors which can impact upon microbial genome assembly are:

- Nucleic acid extraction, *de novo* sequencing requires high quality nucleic acid. For NGS technologies purity and structural integrity of extracted nucleic acid are the key parameters

- Sufficient data has been generated from sequencing, generally, the number of total nucleotides in the generated sequences should be at least 60 times the number of nucleotides in the genome under study (Dominguez Del Angel *et al*., 2018)
- Read length generated, short read lengths whilst accurate can be fragmented. Whilst coverage may help mitigate this, if a repetitive region is longer than a read, coverage on its own will not compensate for this, resulting in gaps in the assembly produced. Paired reads (two reads which have been generated from a single DNA fragment and which are separated by a known distance) can span these gaps but again, they must also produce read lengths longer than the gap (Schatz *et al*., 2010). Longer reads generated from TGST
- GC-content. Inhomogenous GC content can cause a problem for Illumina sequencing, resulting in low coverage in these regions (Chen *et al*., 2013)

2.3.2 Software and bioinformatics tools for data analysis and assembly.

Genome assembly is a computational process which utilises a series of algorithms to take the reads derived from the target genome during sequencing and, rather like a jigsaw puzzle, put them back together in the correct order based on their overlap information. This should then provide an accurate representation of that original genome. Software tools, referred to as assemblers are employed to achieve this.

- De novo assemblers which assemble without the use of reference genomes e.g., SPAdes (Bankevich *et al*., 2012), SGA (String Graph Assembler) (Simpson & Durbin, 2012), MEGAHIT (Li *et al*., 2015), Velvet (Zerbino & Birney, 2008), Canu (Koren *et al*., 2017) and Flye (Kolmogorov *et al*., 2019)

- Reference guided assemblers which assemble by mapping sequences to reference genomes

These algorithms will assemble the short reads into longer contigs. Contig is derived from the word contiguous and represents a set of DNA sequences (reads) that overlap in a way that provides a contiguous representation of a region in the genome. There are three main classes of assembly algorithms:

- Overlap-layout-consenses (OLC)

This assembly method takes all the reads and locates overlaps between them, it then creates a consensus sequence from the aligned overlapping read

- de Bruijn graph (DBG) (Aardenne-Ehrenfest and de Bruijn, 1951)

This method breaks the sequencing reads down further into so called k-mers (with length *k*). The k-mers overlap by k-1, the next k-mer. A de Bruijn graph is then built using all of the *k*-mers

- Repeat graph (Kolmogorov *et al*., 2019)

Published long-read assemblers follow an OLC approach that involves raw read mapping, error correction, pre-assembly, consensus build-up and consensus polishing (Athanasopoulou *et al*., 2022) or a repeat graph approach. Examples are Flye (Kolmogorov *et al*., 2019; Canu (Koren *et al*., 2017) or miniasm (Li, 2016) and Racon (Vaser *et al*., *2017).*

TGST generate longer read lengths which have the advantage of being able to span repetitive regions of the genome, resolving one of the challenges raised by short read technologies, however they are prone to error. It is possible to overcome the shortfalls of both technologies by carrying out a hybrid assembly, in which long reads can be used to scaffold contigs generated by short read technologies. This in turn helps to resolve regions of the assembly graph where short reads have failed (Chen *et al*., 2020). SPAdes (Bankevich *et al*., 2012) and Unicycler (Wick *et al.,* 2017) are examples of assemblers which have been used in hybrid assemblies.

2.3.3 Annotation

Annotation of assembled bacterial genomes can be challenging (Salzberg, 2019). A range of tools has been developed and published enabling even individual researchers to carry out whole-genome analyses on their sequences. However, these tools can be complex requiring an optimised computational strategy and sufficient data processing resources to achieve good results. Annotation of prokaryotic sequences can be separated into structural and functional annotation. Structural annotation elucidates the exact location of different elements in a genome**,** such as open reading frames (ORFs), coding sequences

(CDS), repeats and start/stop codons. Functional annotation compares similarity to other known genes or proteins to assess the function of the gene. Combining both will promote a greater degree of accuracy in the final annotated product.

There are a range of annotation pipelines which can be accessed, notably NCBI's Prokaryotic Genome Annotation Pipeline (PGAP). This pipeline was developed in 2001 and has undergone regular updates to ensure the quality of the results it produces (Tatusova *et al*., 2016; Haft *et al*., 2018 and Li *et al*., 2021). It is designed to annotate bacterial and archaeal genomes by combining *ab initio* gene prediction algorithms with homology-based methods, where the molecular function of a protein can be inferred by analysing the similarity that exists due to common evolutionary ancestry among different organisms. Most genomes in the NCBI Genbank databases have been annotated with PGAP.

Genome annotation, which starts with the unannotated FASTA file can be broken down into three areas, the nucleotide, the protein and process levels, analogous to where, what and how? Where are genes located in the genome, what function do they carry out and how do they function. At the nucleotide level, finding where genes are situated in the genome is the most important outcome. This is usually achieved by a combination of *ab initio* gene prediction and the comparison of sequences with deposited sequence databases such GenBank available through the National Centre for Biotechnology (NCBI) or mapping to a reference genome.

Protein level annotation will allow researchers to assign a function to gene products. Functional annotation pipelines, assign functional roles to coding sequences inferred in the gene prediction process. There are three parallel routes by which the definition of functions can be achieved. The first is through protein domains and motifs using, for example, NCBI PGAP (Prokaryotic Genome Annotation Pipeline) or InterPro (Mitchell *et al*., 2015). The second using orthology data, an example being the KEGG Orthology database (Kyoto Encyclopedia of Genes and Genomes) (Kanehisa *et al.*, 2016) and finally the third by protein homology using for example NCBI-nr, a comprehensive database of

non-identical protein sequences compiled by the NCBI or DIAMOND and MEGAN (Huson *et al.,* 2016 and Bagci *et al*., 2021) and UniProt (UniProt Consortium, 2021).
Other examples of annotation pipelines are listed below.

- MicrobeAnnotator (Ruiz-Perez *et al*., 2021) is an annotation pipeline for microbial genomes that combines results from several reference protein databases
- Prokka (Seemann, 2014) a prokaryotic annotation pipeline which assumes preassembled genomic DNA sequences in FASTA format. Whilst complete sequences without gaps would be ideal, the pipeline will accommodate scaffold sequences produced by *de* novo assemblers
- RAST (Rapid Annotations using Subsystem Technology) (Aziz *et al*., 2008), A web-based annotation pipeline for bacterial and archaeal genomes
- DRAM (Distilled and Refined Annotation of Metabolism) (Shaffer *et al*., 2020). This pipeline is designed for annotating bacterial genomes

Genome assembly is much more accessible in terms of cost, access to next generation sequencing platforms and genome assemblers, however, annotation of the generated contigs can, as previously stated, present challenges. Large, fragmented genomes can be difficult to annotate using automated annotation pipelines and errors in the assembly of thousands of contigs may generate subsequent errors in annotation which can be propagated across species (Salzberg, 2019). Additionally, reference databases such as RefSeq (Tatusova *et al*., 2016) and GenBank can have incomplete sequence deposits (further compounded by the fact that most new genomes are in draft form) making them more prone to inaccuracy (Lu & Salzberg, 2018). Despite these issues, locating genes in bacteria is relatively straightforward as bacterial genomes are approximately 90% protein-coding with short spacer DNA regions between every pair of genes. Once computational gene finders detect which of the six possible reading frames contains the protein, they can exploit this to produce accurate results. We can at least then be confident that we have the amino acid sequences correct despite not being able to assign a definitive function to them.

2.4 *Variance in Prognosis*

The variance in prognosis between CF patients colonised with *M. abscessus* (sensu lato) and *M. chelonae* can have a significant effect on clinical decision making in such individuals. Understanding of the cause of this variance might be gained from understanding the genetic makeup of these species. However, there is considerable taxonomic uncertainty among species within this clade. *M. abscessus* subspecies include *M. bolletii* and *M. massiliense*; *M. chelonae.* To these can be added other related members in the *M. chelonae* clade, *M. franklinii*, *M. immunogenum*, *M. salmoniphilum*, *M. stephanolepidis* and *M. saopaulense.*

Therefore, more detailed understanding of the clinical relevance of the *Mycobacterium abscessus/chelonae* clade through sequencing studies was sought in this research with particular emphasis on the discovery of putative variations in virulence and antibiotic resistance, with the goal of explaining the differences in clinical outcomes.

2.5 Materials and Methods

2.5.1 Strains

A key element of this project is the comparison of the genomic sequence of *M. chelonae* with the sequences of other members of the *M. chelonae* clade, notably *M. abscessus* variants. However, at the start of the project in January 2013, a whole genome sequence for *Mycobacterium chelonae* had not been deposited. The Type strain (ATCC 35752; NCTC 946) was originally deposited in 1923, isolated from a turtle tubercle, and is not representative of clinically important strains.

It was therefore decided that this study would use a clinical isolate retrieved from the Regional Centre for Mycobacteriology (RCM) culture collection and rigorously characterised in an earlier study (Arnold *et al*., 2012) using *rpo*B gene sequence analysis, Matrix-assisted laser desorption/ionisation Time of Flight Mass Spectrometry (MALDI-ToF MS) and DNA-strip technology targeting a region of the ITS gene (Genotype® Mycobacterium AS, Hain Lifescience, Nehren, Germany). Six strains were selected based on viability and culture purity and are listed in Table 3.

Table 3. Representative strains from the 100 strain collection.

Strain Number	Study Designation
M. abscessus (30)	HPA 001
M. abscessus (44)	HPA 002
M. abscessus (47)	HPA 003
M. chelonae (19)	HPA 004
M. chelonae (20)	HPA 005
M. chelonae (32)	HPA 006

Safety statement

None of the mycobacterial species studied in this project are Hazard Group 3 pathogens as recorded by the Advisory Committee for Dangerous Pathogens (ACDP). Nevertheless, all work with live organisms was carried out in the Containment Level 3 Laboratory of the Regional Mycobacterial Reference Centre located in Freeman Hospital, Newcastle upon Tyne. These facilities are approved by the Health and Safety Executive (HSE) and subject to regular internal audit by Public Health England (PHE). The strains were stored within this secure environment and all work was carried out inside a Class I microbiological safety cabinet (MSC). All materials for disposal were autoclaved before discard. The hazard control protocols of the Reference Laboratory were adhered to in all respects

2.5.2 DNA extraction

Six strains (3 x *M. abscessus* and 3 x *M. chelonae*) with comprehensive multiple identification data were selected and subjected to the extraction procedure below:

Protocol for the isolation of high molecular weight genomic DNA from mycobacteria for Ion Torrent.

Two 10 µl loopfuls of mycobacterial growth were transferred into a microcentrifuge tube containing 400ml of 1 x Tris/EDTA buffer (TE buffer). The cells were killed by heating for 20 minutes at 80^0C and cooled at room temperature. To the growth was added 50 µl of 10mg/mL lysozyme, this was vortexed and incubated, while shaking, for at least one hour at 37^0C. 75 µl of 10% of SDS/proteinase K solution (Sodium dodecylsulphate/Proteinase K) was added and again vortexed and incubated for 10 minutes at 65^0C. 100 µl of 5M NaCl and 100ml of pre warmed CTAB/NaCl solution (Cetyltrimethylammonium bromide/NaCl; pre-warmed to 65^0C) was added and the mixture vortexed until the liquid became white ("milky appearance"). This was then incubated for 10 minutes at 65^0C. 75ml of chloroform/isoamyl alcohol was now added and vortexed for at least 10 seconds. The mixture was now centrifuged for 8 minutes at 11,000g.

The aqueous phase at the top was transferred to a fresh microcentrifuge tube, by pipetting small aliquots of 18 µl at a time. To this was added 0.6 volume (450 µl) of isopropanol. The tube was manually moved slowly upside down to precipitate the nucleic acids and an estimate made of the amount of 1 x TE buffer in which the DNA should be re-dissolved at the completion of the process. The mixture was now placed at -20^0C for 30 minutes and then centrifuged for 15 minutes at *ca*. 11,000g. The supernatant was discarded leaving approximately 20 µl above the pellet. To this 1ml of cold 70% ethanol was added, and the tube inverted several times to wash the precipitate. This was then centrifuged for 5 minutes at *ca*. 11,000g and the supernatant discarded to leave about 20 µl above the pellet. After a further centrifugation for 1 minute at *ca*. 11,000g the remaining supernatant was removed using a pipette. The pellet was allowed to dry for approximately 15 minutes at room temperature ensuring all ethanol had evaporated. Finally, the pellet was dissolved in the volume of TE buffer previously estimated.

The resulting DNA was assessed using the Nanodrop 2000C UV-VIS Spectrophotometer. Absorbance measurements made on a spectrophotometer, including the NanoDrop Spectrophotometer, include the absorbance of all molecules in the sample that absorb at the wavelength of interest. Since nucleotides, RNA, ssDNA, and dsDNA all absorb at 260 nm, they will contribute to the total absorbance of the sample. Thus, to ensure accurate

results when using a NanoDrop Spectrophotometer, nucleic acid samples require purification prior to measurement.

The ratio of absorbance at 260 nm and 280 nm is used to assess the purity of DNA and RNA. A ratio of ~1.8 is generally accepted as pure for DNA (Glasel, 1995). If the ratio is appreciably lower, it may indicate the presence of protein, phenol or other contaminants that absorb strongly at or near 280 nm. The aim was to obtain a DNA sample for sequencing with a purity of approximately 1.80 at (260/280 nm)

Protocol for the isolation of high molecular weight genomic DNA from mycobacteria for Nanopore sequencing

The salting out method adapted from Kieser *et al.,* (2000) was employed to isolate high molecular weight genomic DNA from *M. chelonae* HPA 006, as it is particularly useful for long read sequencing such as PacBio and Nanopore.

The biomass from 2 loopfuls (~10 µl) of biomass was resuspended in salt-EDTA-Tris buffer (SET - 75mM NaCl, 25mM EDTA pH8, 20mM Tris-HCl pH7.5) with 100µl lysozyme (50mg ml^{-1}) and incubated at 37°C until viscous (1-3h). Then 140µl proteinase K (20mg ml^{-1}) and 600µl 10% SDS were added, mixed gently by inversion and incubated at 55° for 2 hours with gentle mixing. Two ml of 5M NaCl was added, mixed by inversion and cooled to room temperature then 5ml of chloroform was added and mixed gently for 30 minutes at 20°C. The layers were separated by centrifugation at 4500g for 15 min and the top layer transferred carefully to a clean tube using a wide bore pipette tip and placed on ice. After 2 minutes 0.6 volume of ice-cold isopropanol was added, mixed gently by inversion and incubated on ice to precipitate the DNA. DNA was spooled onto a glass rod, rinsed in 5ml of 70% ethanol and air dried. DNA was dissolved in TE buffer by pipetting gently through a wide bore tip.

DNA was made up to 200 µl and 10 µl RNase A (10 mg ml^{-1}) added and incubated at 37°C for 1 hour and DNA recovered by isopropanol precipitation as above and dissolved in TE buffer.

2.5.3 Ion Torrent sequencing

Access to an Ion Torrent next generation sequencing platform was by courtesy of Professor Anil Wipat, Professor of Integrative Bioinformatics, School of Computing Science, University of Newcastle upon Tyne and Dr. Wendy Smith. Ion Torrent Systems was a commercial company (http://www.iontorrent.com/) now part of ThermoFisher (2023) accessed in 2014.

DNA was sheared into fragments (no smaller than 1.5kb), end repaired and ligated to Ion Torrent adaptors using an NEB library preparation kit. Purified template sequences were mixed with beads and emulsified with oil to form microdroplets containing a single bead with primers and a single template DNA molecule. The template DNA was then amplified by emulsion PCR and denatured to single strands to hybridise to bead-bound primers and generate a multi-template DNA bead. The beads were extracted and amplified beads enriched on a glycerol gradient, with unamplified beads sedimenting to the bottom. Sequencing was performed on an Ion Personal Genome Machine (Pennisi, 2010; Rusk, 2011). A micro-well plate was prepared with a single micro-bead in each well, the 316D chip had a total of 6,348,326 addressable wells. During sequencing the micro-well layer is flooded with a solution containing sequencing reagents and a single dNTP base, if the next base from the primer is complementary this base is incorporated by the DNA polymerase and H^+ ions released and detected by the ISFET (ion-sensitive, field-effect transistor) which measures the ion concentrations in a solution. The detected pH change sends a signal to a computer database where that reaction, in that well, with that particular dNTP, is recorded, using Ion Torrent server software (Ion Torrent Suite version 2.2). In homopolymeric regions, with runs of more than one of a single base, multiple dNTPs are incorporated and H^+ ions released, giving a higher signal.

The micro-well plate is now washed to remove the previous base and the process repeated with the next base. Each run takes about an 1 hour and 100-200 nucleotides are sequenced in each well in a massively parallel sequencing run. Data output was as fastq sequence reads (Cock *et al*., 2010).

2.5.4 Nanopore long read sequencing

The MinION long read sequencing was performed in the laboratories of Demuris at Newcastle University, courtesy of Dr. Nick Allenby with Dr Jeni Devi. Sequencing was performed on a MinION handheld sequencing device (Figure 7) with R9.1 flow cells and SQK-MAP005 library preparation, by ligation with barcodes, and was part of a 12 genome sequencing run. Data output was collected with MinKNOW and cloud-based base-calling using Metrichor to generate fast5 format data files. Fastq sequence reads were extracted with poretools (Loman & Quinlan, 2014).

Figure 7. Oxford Nanopore Technologies: MinION portable genome sequencer

2.5.5 Assembly

Fastq reads from Ion Torrent were assembled individually and together, hybrid assembly, using several assemblers, using a strategy described in Wick *et al*., (2021) in which the contigs produced by different assembly methods are resolved by alignment and consensus.

Ion Torrent Fastq reads were assembled with SPAdes (Bankevich *et al*., 2012) and MIRA (Chevreux, 2005), nanopore reads were assembled with Canu (Koren *et al.,* 2017) and Flye (Kolmogorov *et al*., 2019).

Hybrid assembly of nanopore long reads and Ion Torrent reads was performed with Unicycler (Wick *et al*., 2017).

Read data, contig assemblies and alignments were stored in Geneious (https://www.geneious.com) which was used to convert between formats for input to

other programs. Alignments using MAFFT (Katoh & Standley, 2013), mapping, BLAST (Altschul *et al.*, 1990) and MIRA assembly of short reads, mapped to assembled contigs, was performed in Geneious using Geneious R9.1 and Geneious Prime 2021.

2.5.6 Annotation

Genome annotation was performed with PGAP (Tatusova *et al*., 2016) and Diamond + Megan (Metagenome Analyser) (Bagci *et. al.* 2021) with the workflow illustrated in Figure 8.

All of the reads generated for *M. chelonae* HPA 006 were aligned in DIAMOND against the NCBI-nr protein database. DIAMOND produces a DAA file which has three blocks of data:

- Reference protein header lines
- All aligned reads and
- All alignments.

Following this MEGANISATION was performed. The MEGANIZER tool indexes all of the reads and alignments and bins the reads into taxonomic and functional classes. The computed classifications, together with indexes of all reads, are attached to the bottom of the DAA file producing a "meganized" DAA file, which can be opened in MEGAN.

MEGAN supports a number of different classifications. Taxonomic classification is performed using the NCBI taxonomy browser and the Genome Taxonomy Database (GTDB) (Parks *et al*., 2020). Functional classification is currently performed using Enzyme Commission database (EC) (Barrett, 1992), (Powell *et al*., 2012), InterPro (Mitchell *et al*., 2015) or SEED (Overbeek *et al*.,2013). Functional classification using KEGG (Kyoto Encyclopedia of Genes and Genomes)(Kanehisa & Goto, 2000) is also available in the Ultimate Edition of MEGAN 6.

The main taxonomy viewer in MEGAN provides an overview of the assignment of all sequences to "nodes" in the NCBI taxonomy database. Each node is represented by a circle whose area is relative to the number of reads assigned to that node. The system allows an interactive expansion and collapse of nodes allowing the user to see a greater or lesser degree of detail.

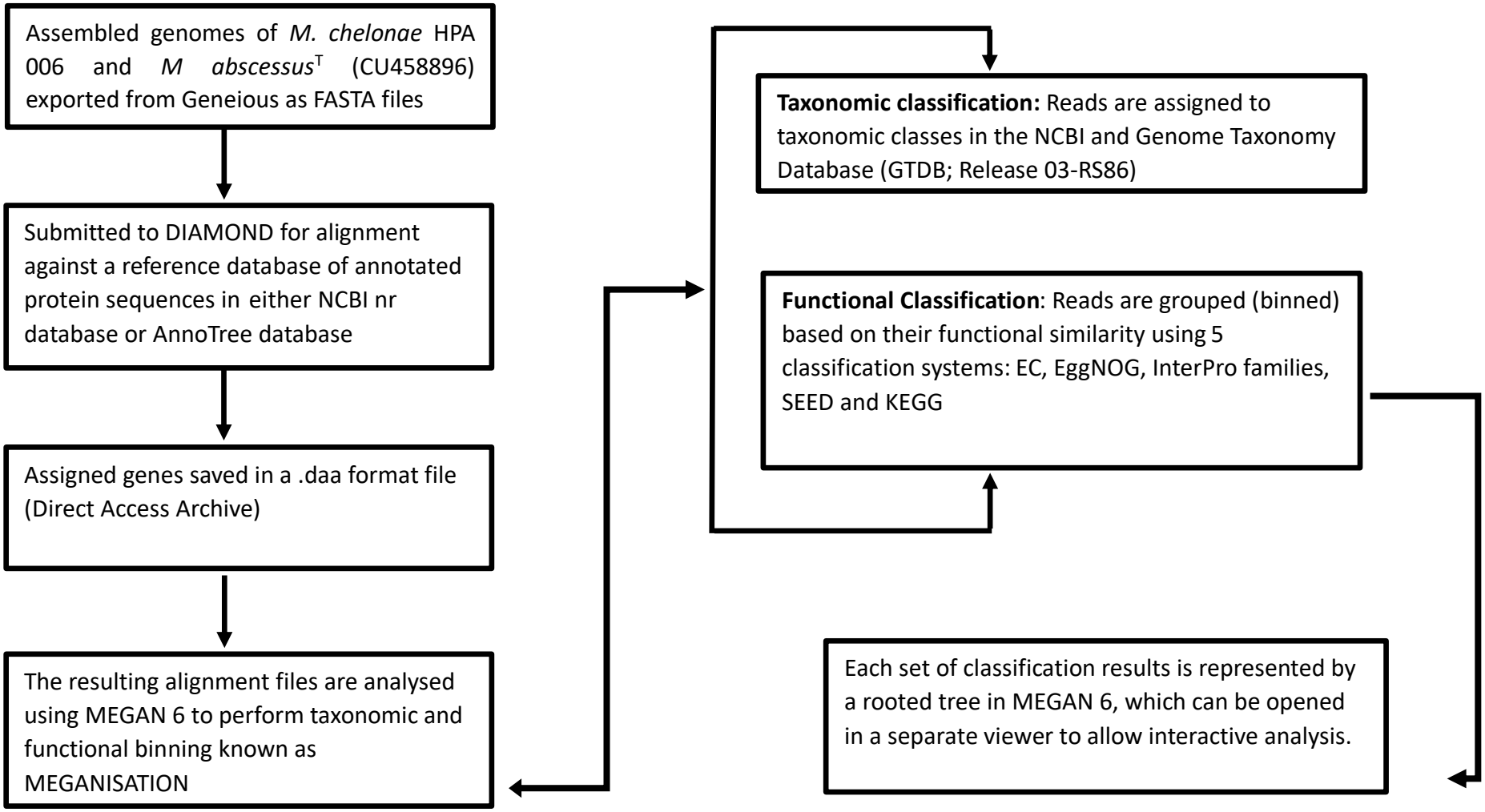


Figure 8. *M. chelonae* HPA 006 genome annotation workflow carried out using PGAP and Diamond and MEGAN 6.

2.6 Results

2.6.1 Selection of representative strain

The 6 strains (3 x *M. abscessus* and 3 x *M. chelonae*) processed through the DNA extraction procedure were assessed for DNA content and volume by spectrophotometry (NanoDrop200C UV-VIS spectrophotometer). The results are presented in Table 4 and from these strain HPA 006 was selected for further analysis and sequence determination using Ion Torrent and the analysis software.

Table 4. Results of DNA quantification using NanoDrop™ 2000.

Sample	Strain ID	DNA Conc (ng/μl)	A260 nm	A280 nm	260/ 280 nm	260/ 230 nm	Volume (μl)	Total DNA (μg)
M. abscessus (30)	HPA 001	8.8	0.176	0.085	2.07	0.86	800	7.0
M. abscessus (44)	HPA 002	10.2	0.203	0.111	1.82	1.27	800	8.1
M. abscessus (47)	HPA 003	2.3	0.046	0.009	4.83	0.58	800	1.8
M. chelonae (19)	HPA 004	4.9	0.099	0.063	1.57	0.86	800	3.9
M. chelonae (20)	HPA 005	3.0	0.061	0.026	2.32	0.63	800	2.4
M. chelonae (32)	HPA 006	5.2	0.105	0.070	1.49	0.78	800	4.2

The 260/280 ratio gives an indication of how pure the sample is from contaminating protein. The optimal 260/280 ratio for DNA is 1.8. The 260/230 ratio reflects how pure the sample is from salts and other contaminants which can absorb at 230 nm e.g., EDTA, phenol and guanidine hydrochloride. Pure nucleic acid samples have a 260/230 ratio of 2 or above.

Table 4 shows that HPA 004, HPA 005 and HPA 006 as *M. chelonae* strains, potentially fulfil these criteria. Whilst HPA 004 and HPA 005 exhibited a greater level of DNA purity than HPA 006 none demonstrated the optimal level at the 260/230 nm ratio. Additionally on solid culture HPA 006 gave a pure growth on subculture, whereas the other two strains showed a degree of bacterial contamination. HPA 006 had a satisfactory level of DNA and an acceptable level of contamination with proteins and other organic compounds.

2.7 Whole Genome Sequencing

2.7.1 Ion Torrent Sequence analysis

The processing steps and proposed workflow utilised by Ion Torrent are summarised in Figure 9. The Ion Torrent server writes the sequence reads to Fastq (Cock *et al*., 2010) and Standard Flowgram Format (SFF) files, binary files containing the raw data from the sequencer, including the base calls, sequencing quality scores, and metadata such as the Flowgrams and signal processing outputs.

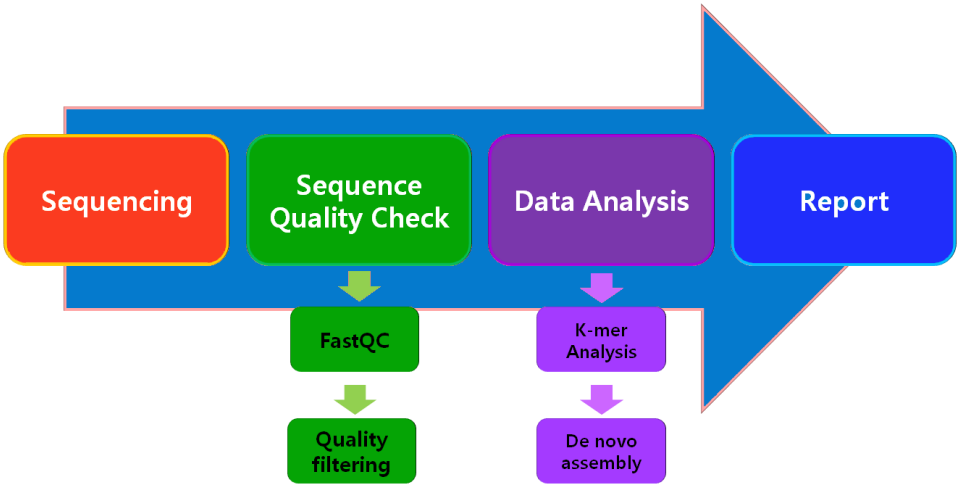


Figure 9. Ion Torrent Whole genome sequence analysis workflow for *M. chelonae* HPA 006.

2.7.2 Sequence quality check

The Ion Torrent server software (Ion Torrent Suite version 2.2) reported the total number of reads as 2,539,246 with a mean length of 209 giving a total of 530 Mbp, with 460 Mbp >Q20. The read length histogram is shown in Figure 10, indicating some very short reads.

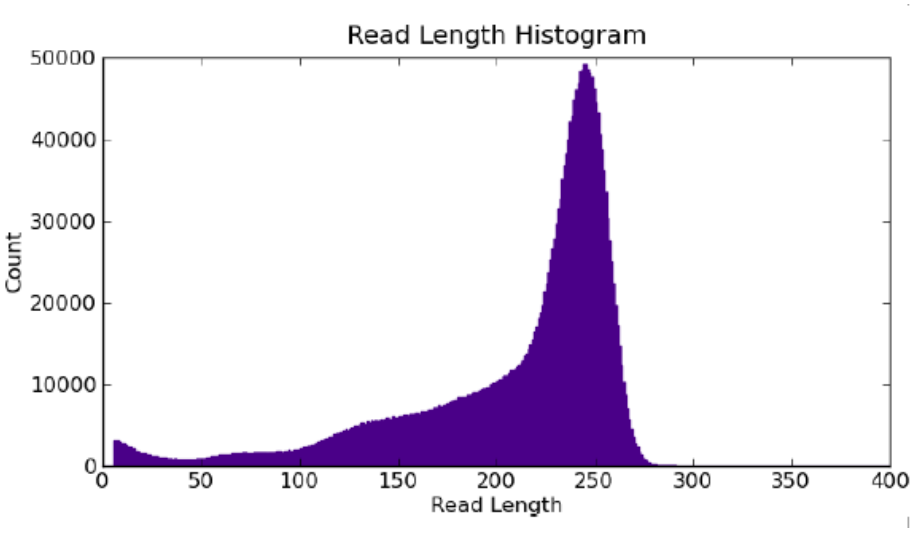


Figure 10. Read length histogram for *M. chelonae* HPA 006 Ion Torrent whole genome sequencing run.

The *Mycobacterium chelonae* genome is about 5Mb in size, so this gives 100x coverage. The 316 semi-conductor chip (present in the Personal Genome Machine (PGM) which captures chemical information from DNA sequencing, translating it into digital information or base calls) had a total of 6,348,326 addressable wells with 77% of the wells loaded with Ion Sphere particles (Figure 11).

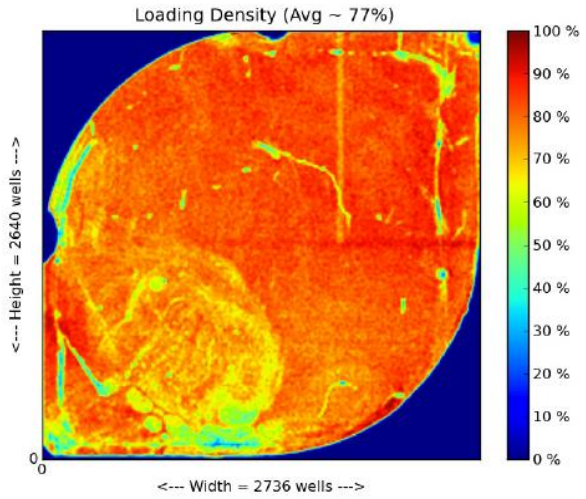


Figure 11. Ion-Torrent PGM 316 Ion-chip sequencing density plot analysis for *M. chelonae* HPA 006 sequencing run.

ISP loading density is the percentage of chip wells that contain an Ion Sphere Particle (ISP; templated and non-templated, or live and failed ISPs). This percentage value considers

only the potentially addressable wells and is a result of the software well classification step. The ISP density image is a colour image of the Ion Chip that shows the percentage of loading across the physical surface. Red indicates adequate ISP loading density, yellow indicates less-than adequate loading density, and blue indicates an absence of loaded beads.

91% of the 4,898,267 loaded wells were live (had amplified fragments indicated by the red colour in Figure 11) and 52,693 were test fragments leaving 4,386,622 library reads. It is important to get the input concentration correct to ensure that there are sufficient amplicons to provide enough data, however, if the concentration is too high more than one amplicon can end up bound to the Ion Sphere Particle (ISP), giving a polyclonal ISP which cannot be used and is discarded from the final data set. In this sequencing run 1,110,372 reads were filtered as polyclonal and therefore discarded. 144 identified as primer-dimer, i.e. reads where no or only a very short sequencing insert is present and reads that, after P1 adapter trimming, have a trimmed length of <25 bases are considered primer dimers and 736,860 filtered as low quality reads leaving 2,539,246. Low quality reads can occur for several reasons but most commonly are associated with inaccurate flow-calls, which introduce insertion/deletion (indel) errors at a raw rate of 2.84% (1.38% after quality clipping) (Bragg *et al* 2013). Filters are also applied to remove reads with low signal quality, and reads trimmed to less than 25 bases

The average Q17 read length for test fragments TF-A and TF-D was 94 and 92 of the 97 bp length test fragments. The Ion Torrent sequencing platform uses test fragments which are slightly less than 100 bp in length. They are spiked into the experimental sample before the sequencing chip is loaded. This sequencing kit used 2 test fragments, A and D. These Q17 quality results (corresponding to 1 base error allowed per 50 bases) indicate a high level of accuracy in base calling.

Analysis of the Fastq data file:

(http://www.bioinformatics.babraham.ac.uk/projects/FastQC) shown in Figure 12 indicates that sequences were generally of high quality up to position 149 before dropping in quality after 250 bases. The quality of Ion Torrent sequencing reads can drop off at around 250 bases due to the accumulation of errors in the reads (Forth & Hoper, 2019).

The technology utilises a pH meter to detect the release of hydrogen ions during DNA synthesis. However, as the length of the DNA sequence increases, the pH change becomes smaller and harder to detect accurately. This can lead to errors in base calling and a decrease in read quality (Bragg *et al*., 2013). The quality of base calls on most platforms is known to degrade as the sequencing run progresses, so it is common to see base calls falling below a quality score of 28 towards the end of a read (Forth & Hoper, 2019).

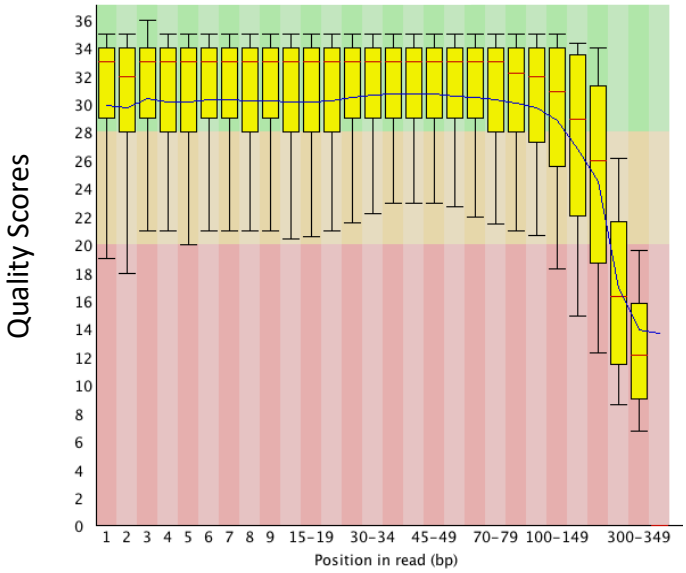


Figure 12. Quality scores versus read position for *M. chelonae* HPA 006 Ion Torrent whole genome sequencing run. Per Base sequence quality showing mean and standard deviation of sequencing quality for each position in all reads of the data set.

Initially the Ion Torrent reads were left untrimmed as the MIRA assembler has an option to trim these, specifically recommending the input of untrimmed reads (Chevreux, 2005).

The distribution of GC content looks sensible *i.e*., overall, sequence peaks at around 63% with at least 160,000 reads and is consistent with what would be expected for the genus *Mycobacterium*. The GC content of microorganisms is a hugely variable attribute. As early as 1962 it was noted that in bacteria, the GC content can range from less than 25% to higher than 75% (Sueoka, 1962). More recent studies demonstrate that the GC content of bacterial genomes can be as low as 13% as seen in *Zinderia insecticola* (McCutcheon & Moran, 2011). The GC content range for the genus Mycobacterium is 61 to 71%, with the

average being in the region of 65.6% (Lévy-Frébault & Portaels, 1992) The theoretical distribution of GC content in a normal random library is expected to be a roughly normal distribution where the central peak corresponds to the overall GC content of the underlying genome. The modal GC content is calculated from the observed data and used to build a reference distribution. An unusually shaped distribution could indicate a contaminated library or the presence of a biased subset. The GC content of the mycobacterial genome is 65.6%, the theoretical distribution and GC count per read for this sequencing run for *M. chelonae* HPA 006 (Figure 13) gives confidence that the data for *M chelonae* HPA006 represents that of a member of the genus *Mycobacterium.*

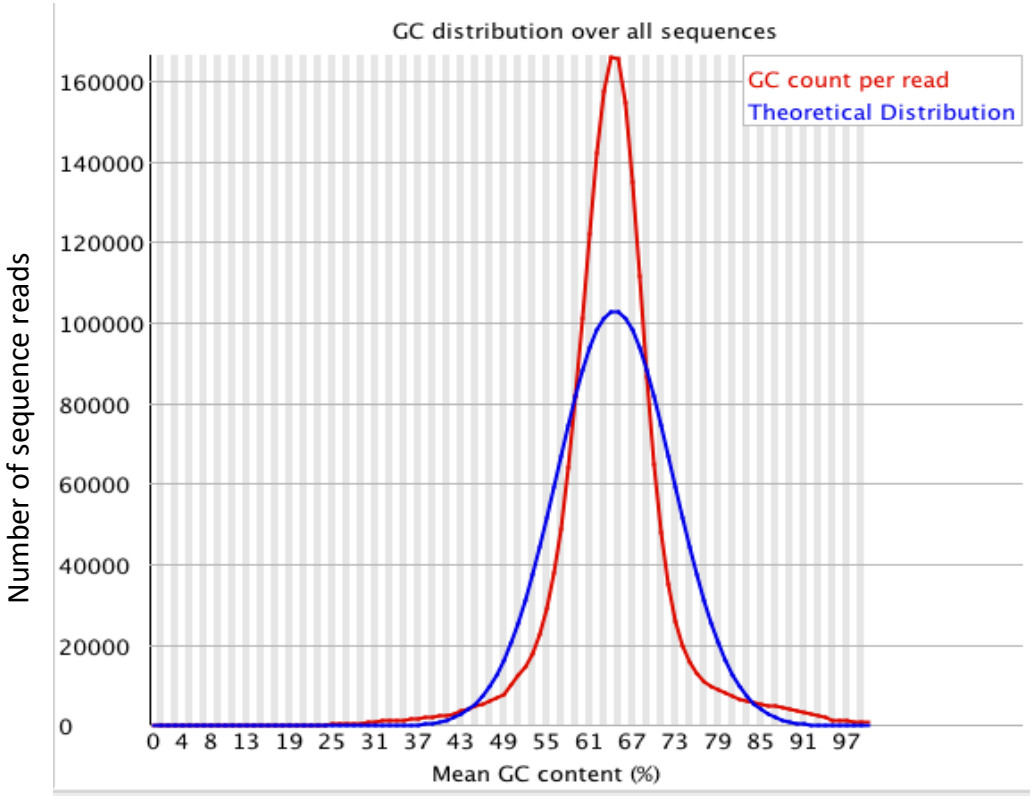


Figure 13. GC count per read analysis of *M. chelonae* HPA 006 Ion Torrent sequencing reads.

The observed GC count per read distribution (red line) is compared with the Theoretical Distribution that would result from a Poisson process with the same mean (blue line). For *M. chelonae* HPA006 the correspondence between the expected and observed distributions is excellent.

2.8 Data Analysis

2.8.1. K-mer analysis

FastQC does a simple k-mer (5-mers) analysis (Figure 14). There were no overrepresented sequences but a high peak of repetitive K-mer sequences were encountered from position 300 bp to 349 bp. Specific k-mers occur frequently when the same sequence is present, when these occur at the ends of reads, it is likely to be adapter sequences which will be trimmed by MIRA in this project. In Figure 14 they are visible in positions 1 through to 9 and again at the end around base 200-250.

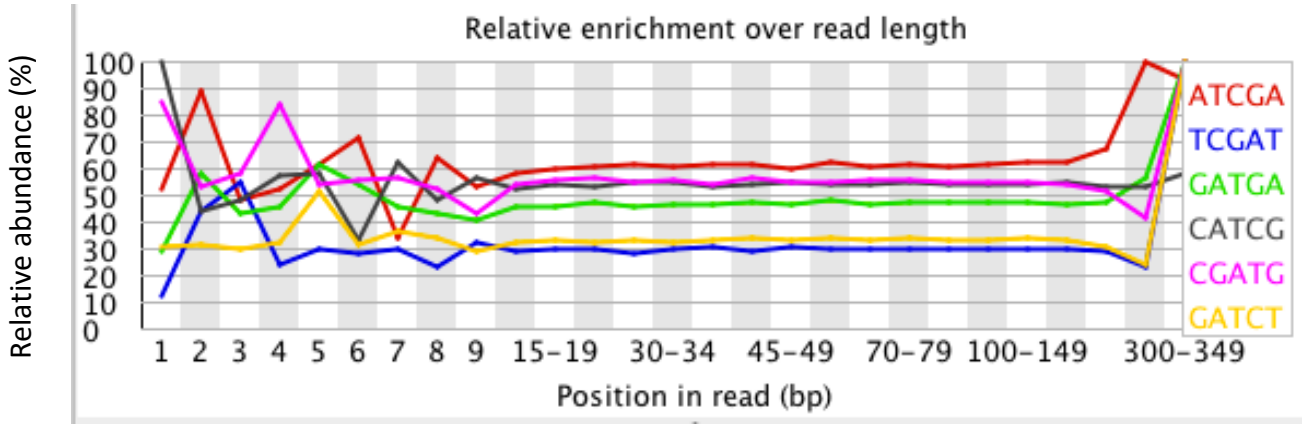


Figure 14. FastQC k-mer analysis of *M. chelonae* HPA 006 Ion Torrent sequencing reads, indicating the presence of repetitive sequences amongst the reads located at 300-349bp.

2.9 Mapping to Reference Genome

The analysis of the *Mycobacterium abscessus* genomes (Ripoll *et al.,* 2009; Sassi & Drancourt, 2014) shows that the *M. abscessus* genome is collinear with *M. smegmatis* with no major rearrangements so it can be expected that the *M. chelonae* genome should map reasonably well onto *M. abscessus*. Mapping the *M. chelonae* HPA 006 reads, using Geneious software (Biomatters Ltd) onto the NC_010397 *Mycobacterium abscessus* genome (Ripoll *et al*., 2009) demonstrates the alignment of nearly 2 million reads onto a 6Mb genome of *M. abscessus*. The reads map with high similarity and reasonably complete coverage, but with some gaps. Figure 15 illustrates how similar *M. chelonae* is to *M. abscessus*.

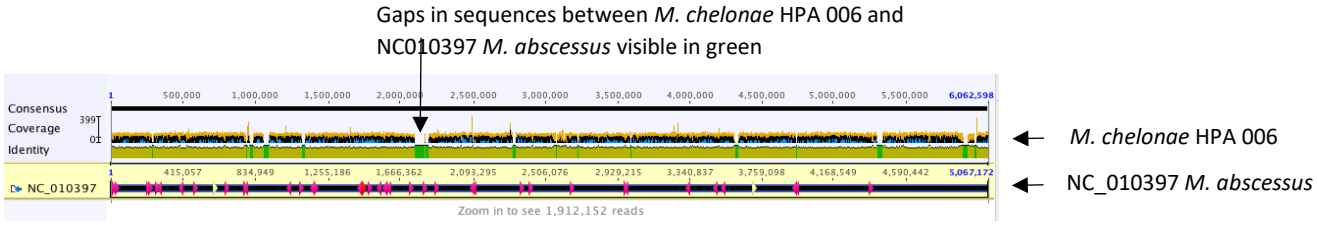


Figure 15. Ion Torrent reads from *M. chelonae* HPA 006 mapped onto *M. abscessus* (CIP 104536T = ATCC 19977T) using Geneious software.

Zooming in shows this at the individual gene level (Figure 16).

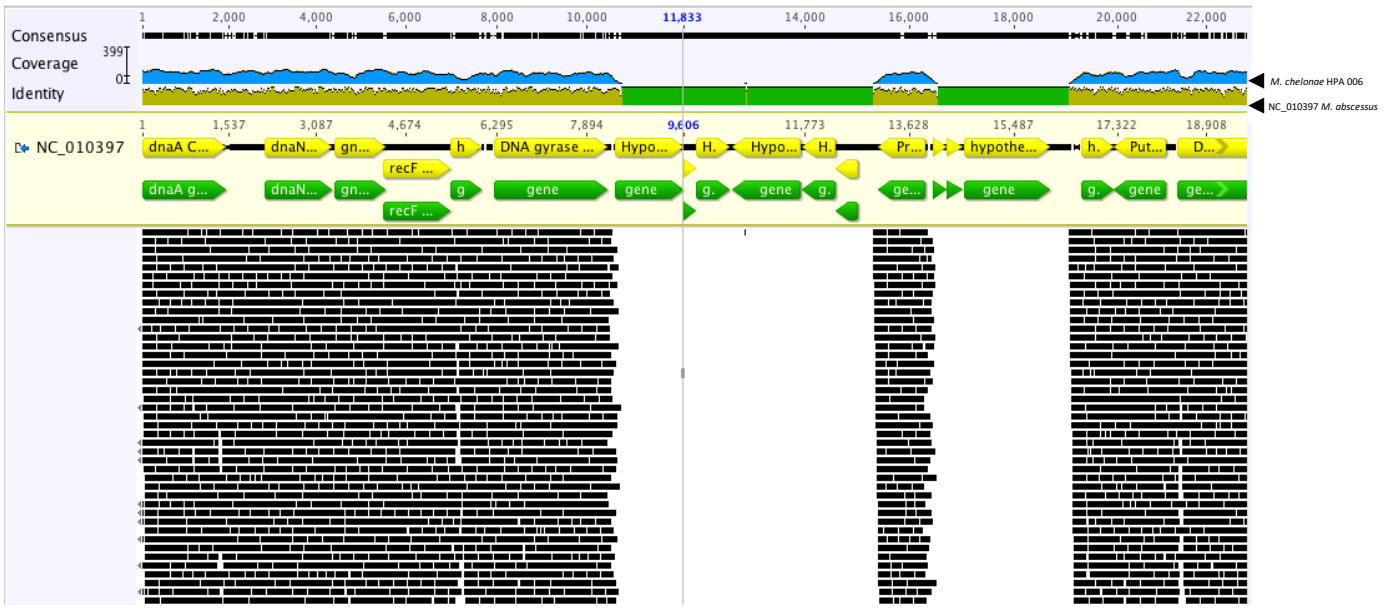


Figure 16. *M. chelonae* HPA 006 FastQ reads mapped onto the Mab_0001 – Mab_0019 region of the *M. abscessus* (CIP 104536T = ATCC 19977T) genome.

The dnaA gene, a protein that activates initiation of DNA replication in bacteria, is the start of the genome. The first block of genes, present in *M. abscessus* but for which no reads from *M. chelonae* match, Mab0007-Mab0012c are all hypothetical protein coding regions, as are Mab0014-Mab0016. Missing genes, annotated as hypothetical was a recurring theme, which means that their function could not be classified.

Mab0013cis a probable arylamine n-acetyl transferase. The retention of the arylamine N-acetyl transferase in this deleted region suggests it is being actively evolutionarily

conserved, however as it is present in both *M. chelonae* and *M. abscessus* it does nothing to explain the difference in clinical outcome between these two species.

Aromatic amines are highly toxic and arylamine n-acetyl transferase, found for example in human liver may detoxify xenobiotics, including, for example isoniazid, a front line antituberculosis drug. Arylamine n-acetyl transferase is found in *Mycobacterium tuberculosis* and other mycobacteria, including non-pathogens such as *M. smegmatis*. They may also have a role in lipid metabolism (Bhakta *et al*., 2004) and loss in *M. tuberculosis* results in cell wall defects.

The sequence of the DNA region Mab0007-Mab0012c in *M. abscessus* with blastn in Genbank only finds hits with *M. abscessus*, *M. massiliense* and *M. bolletii.* However, against the *Mycobacteriaceae* (you must select organisms before doing a nucleotide blast against the WGS database) it finds hits to many mycobacteria, including *M. tuberculosis* and the contigs submitted for another project sequencing a putative *M. chelonae* 1518 (Institute for Genome Sciences, "Whole genome sequencing of 14 antibiotic resistant nontuberculous mycobacteria used in preclinical compound testing at Colorado State University").

Annotated protein sequences from WGS projects are present in the general nr database and can be searched with blastp. Blastp of the Mab0007 protein sequence shows sequence similarities to a generic hypothetical protein in mycobacteria but also similarity to hypothetical proteins in other actinomycetes such as nocardia and rhodococci (Figure 17).

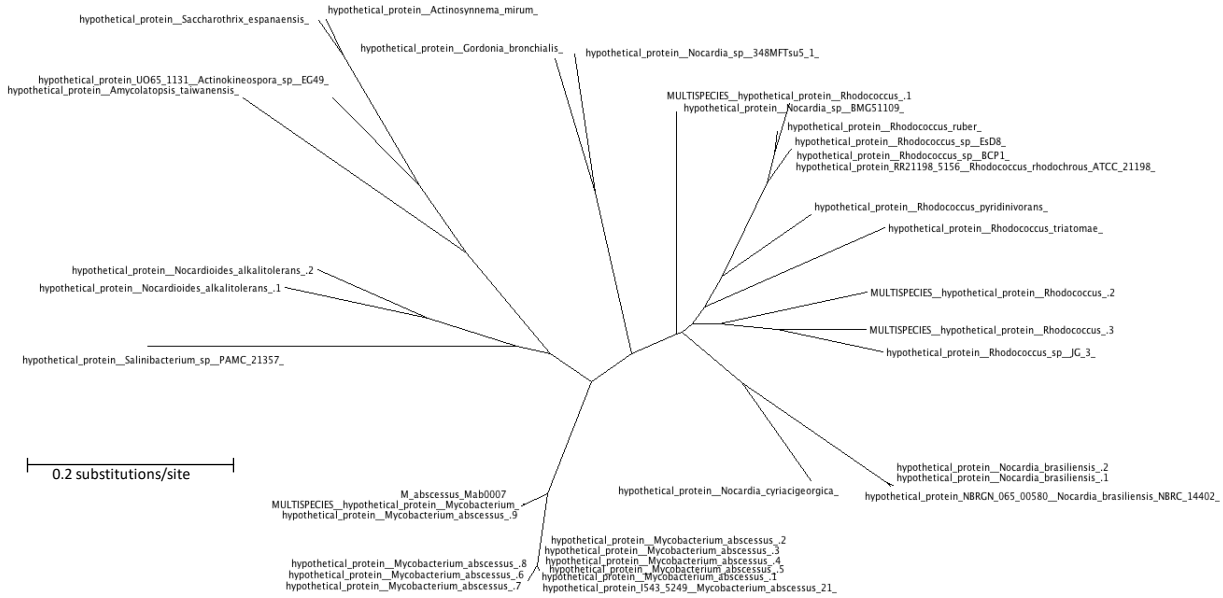


Figure 17. Minimum evolution pairwise tree for *M. abscessus* (CIP 104536T = ATCC 19977T) genome DNA region Mab007 and related protein sequences (Genbank).

Zooming in to view individual bases and reads (Figure 18) shows the coverage and identity, as well as highlighting read errors. However, it is clear that the errors, even the high possible sequence error at the end of RZ4P3:277:631 (reversed), are diluted out by the sequence read coverage.

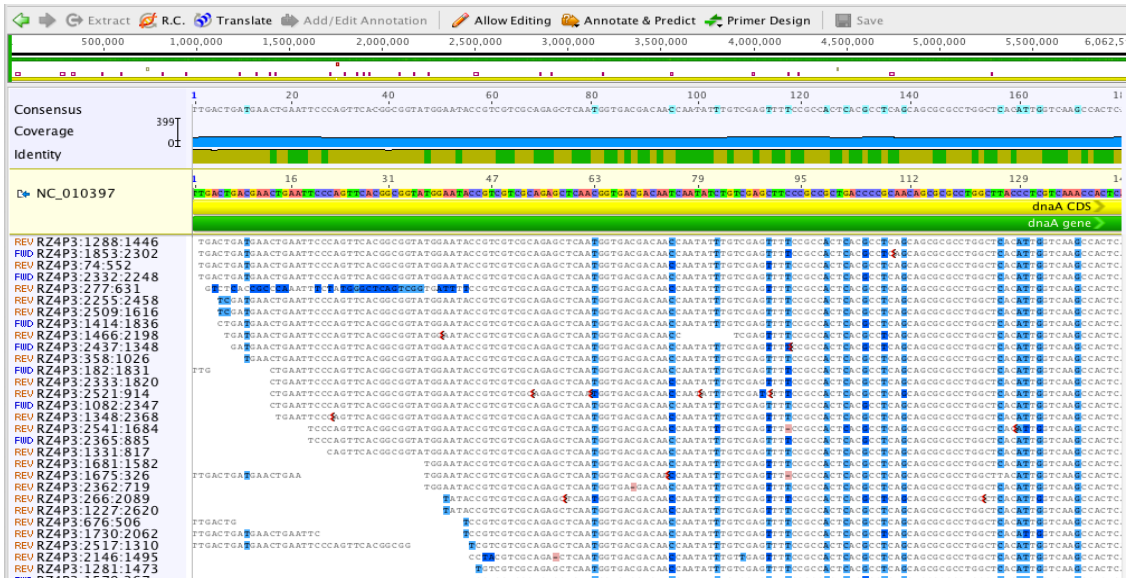


Figure 18. *M. chelonae* HPA 006 sequence reads mapped to the 5' end of the dnaA gene in *M. abscessus*^{T}.

The mapping leaves multiple small gaps along the genome where *M. chelonae* has no matching sequence. However, gaps are not uncommon. The sequence assembler MIRA is very good at trimming poor quality ends and these sporadic errors were corrected by the consensus assembly. An initial scan of the genes present in these regions of difference shows that they are dominated by hypothetical proteins. There are also numerous transcriptional regulators, ABC transporters and lipid associated genes. Some of the regions of difference are linked to phage associated sequences including one of the largest regions of difference (Figure 19).

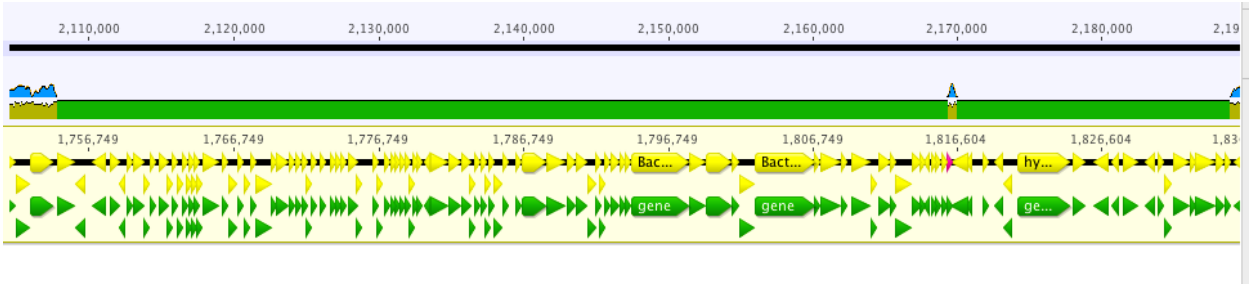


Figure 19. Bacteriophage associated region of difference.

This region is one of the likely sources of variation between strains, as mycobacteria, due to cell wall structure, do not typically exchange a lot of genetic material.

There are a few antibiotic resistance-associated genes: 3 beta-lactamase; 1 streptomycin resistance gene; 1 chloramphenicol resistance protein and 1 chloramphenicol acetyl transferase; and 1 glyoxalase/Bleomycin resistance protein. Only 1 MCE and 1 PE/PPE associated protein are in the regions of difference and there is 1 putative surface layer protein present in *M. abscessus* and missing from *M. chelonae*.

The ribosomal rRNA genes are often absent from whole genomes which are not finished. One problem is assembly, with repetitive sequences, but there is only 1 rRNA cluster in *M. abscessus,* and mapping reads does not depend upon assembly. The coverage of this rRNA cluster from the reads for *M. chelonae* is low, suggesting there is a bias in obtaining sequence coverage. In many genomes (available in the databases) sequenced by short read sequencing (Illumina, Ion Torrent etc.) there is no ribosomal RNA sequence (no 16s rRNA) because sequence coverage is often lower and assembly methods for short reads

cope poorly with longer repeats found in these regions. Figure 20 illustrates that there is less coverage of ribosomal RNA

Figure 20. Ion Torrent Read coverage of the ribosomal RNA gene cluster of *M. chelonae* HPA 006.

2.10 Nanopore Sequence Analysis

Nanopore sequence analysis, 1/12th of a flow cell, yielded 24 Mb of sequence data in 8,377 reads, the largest read was 43 Kb. The read length distribution is dominated by small reads, but the minimum size is >200 bp and the most bases, >10Mb, are in reads between 1kb and 5kb and 22Mb in reads larger than 1kb (Figure 21).

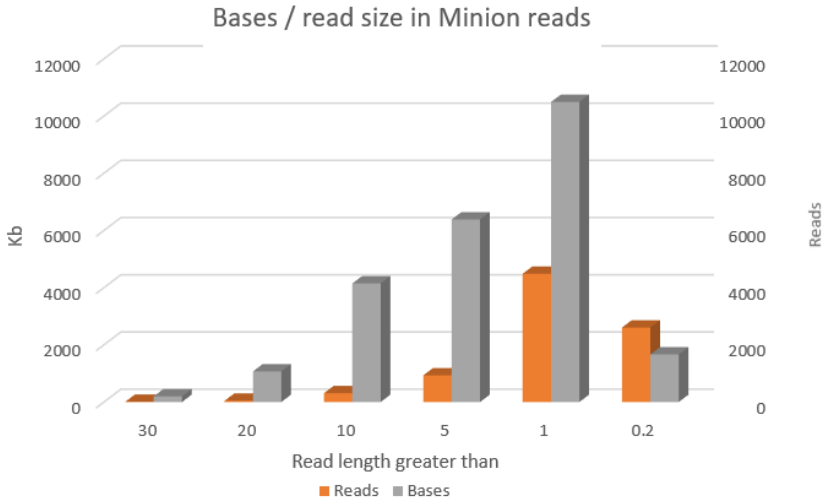


Figure 21. *M. chelonae* HPA 006 sequence reads and read bases histogram generated from sequencing carried out on the MinION.

The read coverage at 24 Mb is only 4-5x coverage for the expected *M. chelonae* genome size, but long reads are required to cover the repeat regions over which short reads fail. The long reads map poorly to *Mycobacterium abscessus* (CIP 104536^T = ATCC 19977^T as long reads span regions of difference between *M. chelonae* and *M. abscessus*. Mapping the read sequence data, split into 200Kb fragments, mapped to the *M. abscessus* genome gives a more valid comparison between the long and short read data (Figure 22). The main difference is the coverage – the maximum coverage for the MinION fragment coverage (Figure 22a) is 17 compared to the maximum for Ion Torrent of 371 (Figure 22b). In Figure 22 and Figure 15 there are too many Ion Torrent reads mapped to the *M. abscessus* genome for Geneious to display, see zoomed in displays (Figures 16 and 20) to get a visual impression of Ion Torrent coverage.

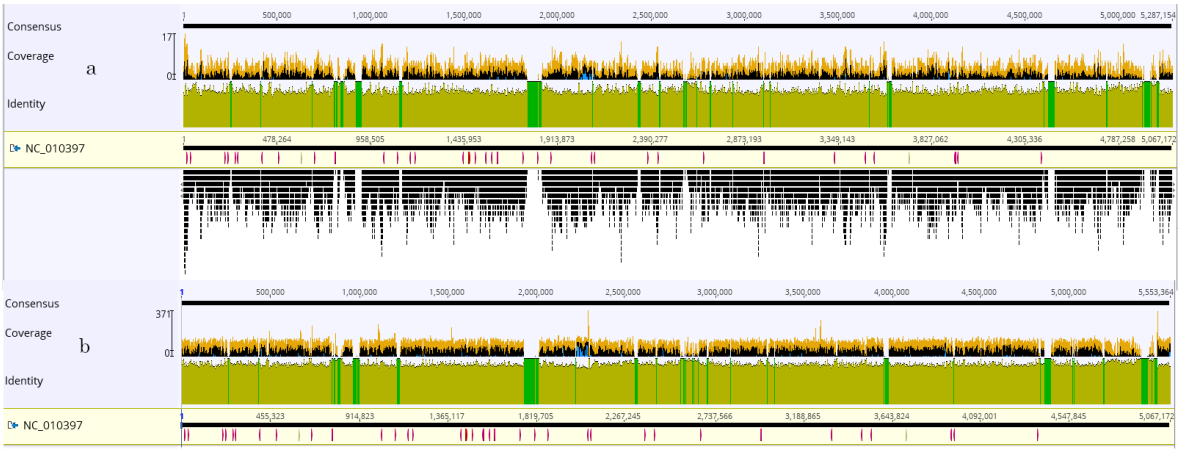


Figure 22. **a**. Mapping of *M. chelonae* HPA 006 MinION long read fragments and **b**. Mapping of *M. chelonae* HPA006 Ion Torrent short read fragments to *M. abscessus* (CIP 104536T = ATCC 19977T).

However, the long reads do cover repeat regions, for example a number of long reads span the rRNA operon (Figure 23)

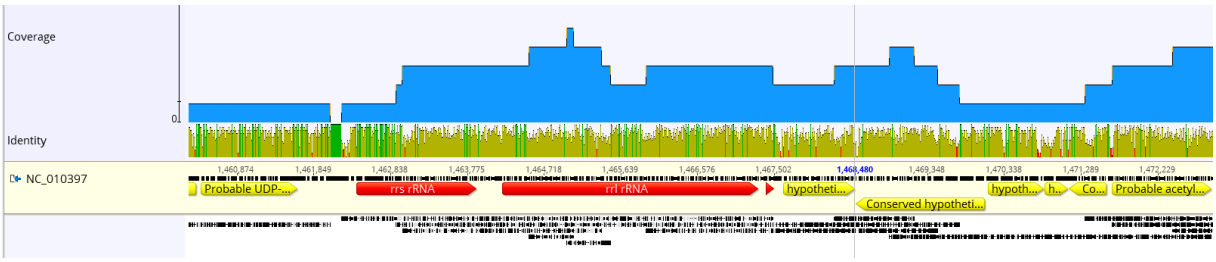


Figure 23. MinION read coverage of the ribosomal RNA gene cluster in *M. chelonae* HPA 006.

2.11 Assembly

The rapid development in sequencing technologies and the massive amounts of data generated has meant rapid changes in both the technologies and the software to analyse. The results here are based upon relatively early adoption of the technologies (2016) but the results presented are based upon analysis using the latest available version of the software in 2019. In 2016 the Nanopore technology (pore design and library preparation and base calling) software resulted in high error rates. Subsequent improvements in base calling software and the sequencing technology have given much lower error rates. The improvements in software

facilitated better quality sequence data by re-base calling the original raw, 2016 data in 2019.

2.11.1 Canu assembly

The Canu assembler was one of the first assemblers available for nanopore long reads, an overlap (OLC) assembler derived from the original Celera Assembler (Myers *et al*., 2000; Miller *et al*., 2008). Canu assembly includes three stages: correction, trimming, and assembly, as well as the final assembled contigs the corrected reads, corrected by consensus from overlapping reads, can be saved. Canu created 127 contigs and 5,759 corrected reads.

2.11.2 Flye assembly

The Flye assembler uses a repeat graph (Kolmogorov *et al*., 2019) strategy to assemble long read data and was developed in response to the availability of long read data from PacBio and Nanopore. Assembly of the 8,377 nanopore reads with Flye generated 85 contigs.

2.11.3 SPAdes assembly

SPAdes (St. Petersburg Genome Assembler, Bankevich *et al*., 2012) is a genome assembly algorithm that was originally developed for de novo assembly of genome sequencing data produced for cultivated microbial isolates and single celled genomic DNA sequencing. It works with Ion Torrent, PacBio, Oxford Nanopore, and Illumina paired-end, mate-pairs and single reads. However, it is not thought to be suitable for large genome projects, for example, those studying mammalian sized genomes. SPAdes assembled the Ion Torrent reads to 246 contigs.

2.11.4 MIRA assembly

MIRA 5 (Mimicking Intelligent Read Assembly) is a whole genome shotgun and EST (Expressed Sequence Tag) sequence assembler which maps the sequencing reads gained form Sanger, 454 pyro sequencing, Solexa (Illumina) and Ion Torrent platforms into contiguous sequences or contigs (Chevreux, 2005).

It is an iterative assembler that works over several passes, extending reads and therefore contigs based on extra information gained by overlapping read pairs in contigs and the corrections made by the automated editor. MIRA 5 is recommended for genome de-novo assemblies or mapping projects for haploid organisms up to 20 to 40 mega bases. It has an Ion Torrent option and assembles the Ion Torrent sequencing reads in this project to 518 contigs.

2.11.5 Unicycler hybrid assembly

Unicycler uses SPAdes to assemble the short reads and then uses information from the SPAdes assembly to identify bridges and align short reads and long reads, in a semi-global alignment (end gap free alignment) taking advantage of the fact that the short and long reads come from the same sample, so, unlike alignment of sequence data from different samples implemented in most long read alignment tools, there should be no structural rearrangements in the read sets (Wick *et al*., 2017). The Unicycler pipeline is not quite optimal as the SPAdes assembler implementation called is optimised for Illumina data, so the SPAdes assembly utilising the Ion Torrent option may give improved assembly for later consensus assembly.

Unicycler assembles the 2,539,246 Ion Torrent reads and 8,377 nanopore long reads into 17 contigs.

2.12 Consensus Assembly

The Unicycler assembly of 17 contigs totalled 5,203,609 bases, close to an expected genome size of about 5 Mb (Figure 24).

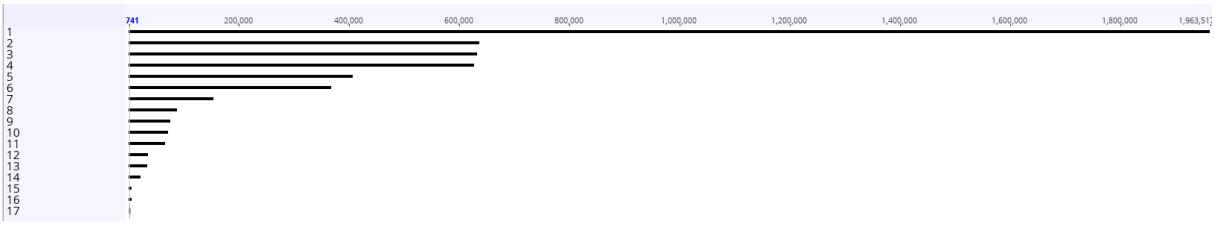


Figure 24. Size in base pairs of Unicycler assembled contigs from long and short read sequences generated for *M. chelonae* HPA 006.

The Unicycler contigs were validated by mapping short reads, short read contigs, long read contigs and corrected long reads to each contig using the Geneious mapper (Figure 25). In this context, other mappers such as BWA_MEM (Li, 2013), did not offer significant advantages and were more time consuming to execute and visualise the results.

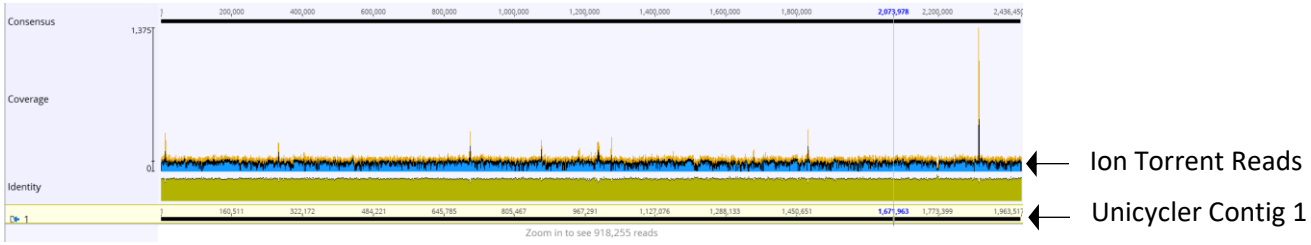


Figure 25. *M. chelonae* HPA006 sequencing reads mapped against Unicycler Contig 1 using the Geneious mapper.

Zooming in to the rRNA operon (Figure 26) and mapping the Ion Torrent reads to the junction of the 16S gene and the next gene, *murA*, in the contig, annotated with PGAP, shows the contiguity of the assembly.

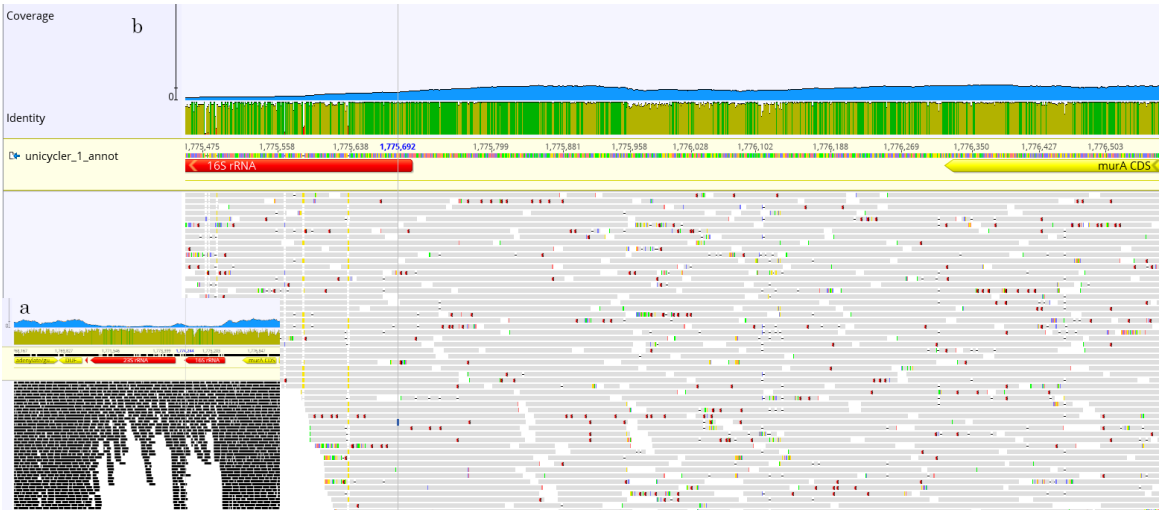


Figure 26. Mapping of Ion Torrent reads to the rRNA operon in Unicycler contig 1 is shown in inset **(a)** and zoomed into the transition from 16S to the adjacent gene (murA), Unicycler contig 1 annotated by PGAP**(b).**

Nevertheless, the mapping shows some differences between the reads and the assembly (Figure 27).

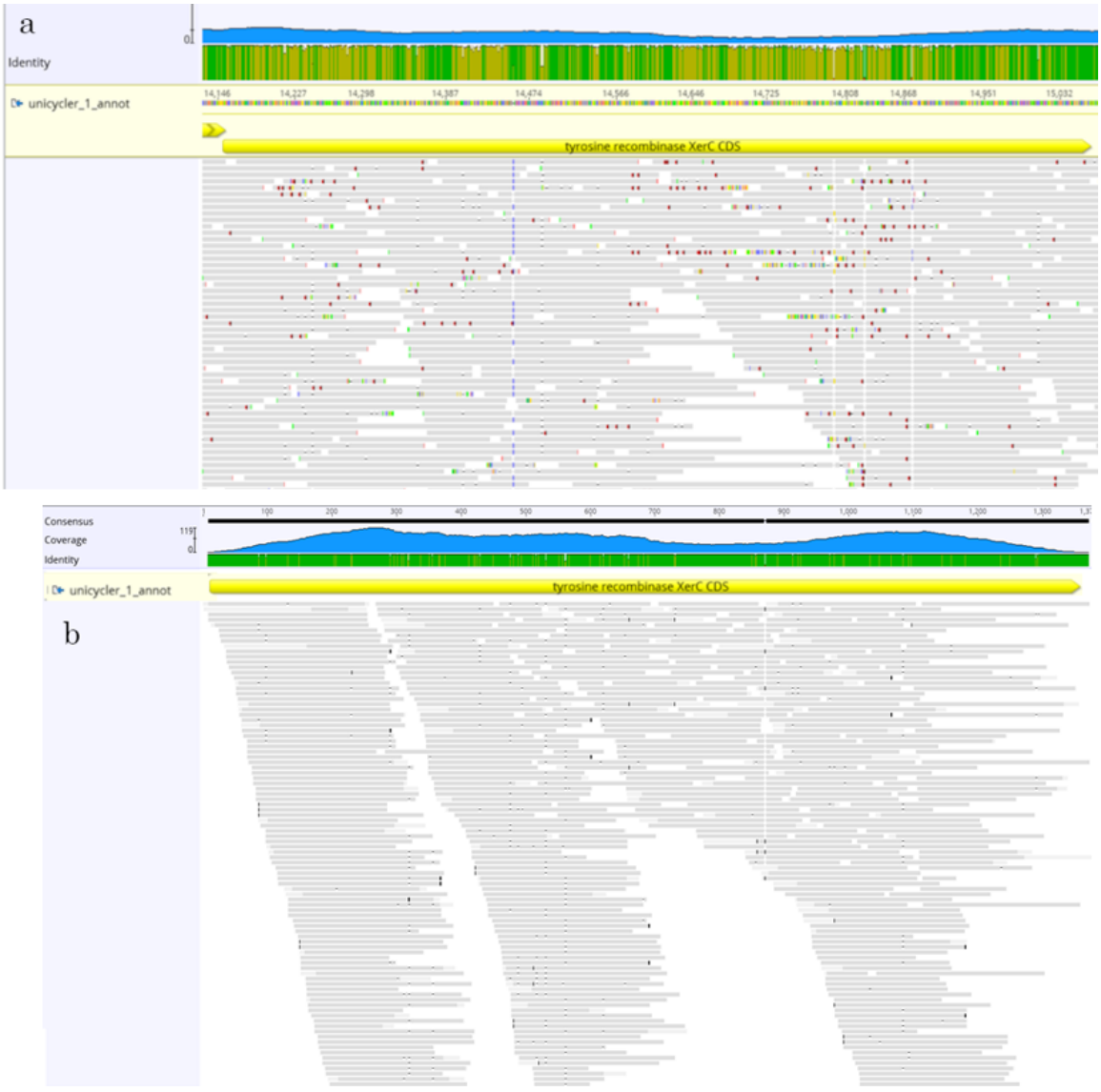


Figure 27. **a.** Mapping of *M. chelonae* HPA 006 Ion Torrent reads to a tyrosine recombinase in Unicycler contig 1 and **b.** Assembly of the mapped reads by MIRA 5 validating the assembly with high accuracy Ion Torrent reads.

MIRA does not produce the fewest or the longest contigs, but it assembles short reads very carefully, MIRA 5 has an option, recommended for long read assembly error correction with short read data, to assemble short reads to a reference genome. After scaffolding the Unicycler contigs this option will be applied to error correct the final assembly, so there is no need to error correct reads or contigs individually. However, the option remains, during scaffolding, to check the accuracy of data e.g., when joining contigs with low coverage and/or error prone long reads. In the case of the tyrosine recombinase region of Unicycler contig 1 440 Ion Torrent reads map to the sequence, MIRA 5 uses 436

reads to re-assemble the tyrosine recombinase sequence and the overlapping ends, from reads captured by the sequence bait, which will include reads with partial overlaps to the ends of the sequence. Figure 27 illustrates that merely mapping the reads to a reference is not an assembly as there are numerous mismatches visible at the ends of reads. Comparing this with the MIRA assembly (b) we see a clear demonstration that it trims the reads, aligning them accurately, resulting in just one correction from the Unicycler assembly

After MIRA assembly there is one correction, an indel error, a C corrected to CC (Figure 28).

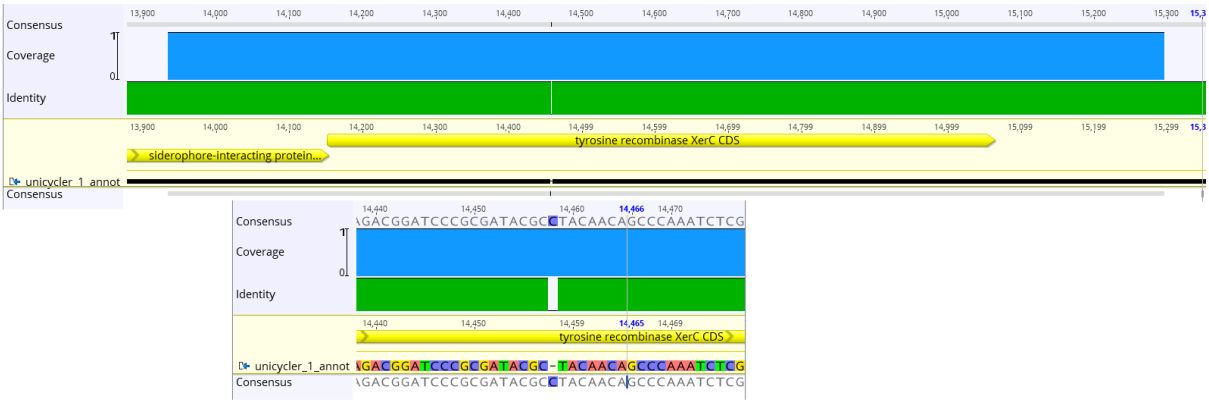


Figure 28. MIRA 5 assembled *M. chelonae* HPA 006 reads for tyrosine recombinase mapped back to Unicycler contig 1.

The end of Unicycler contig 1 reveals the reason for termination of the assembly, namely low coverage of Ion Torrent reads (Figure 29).

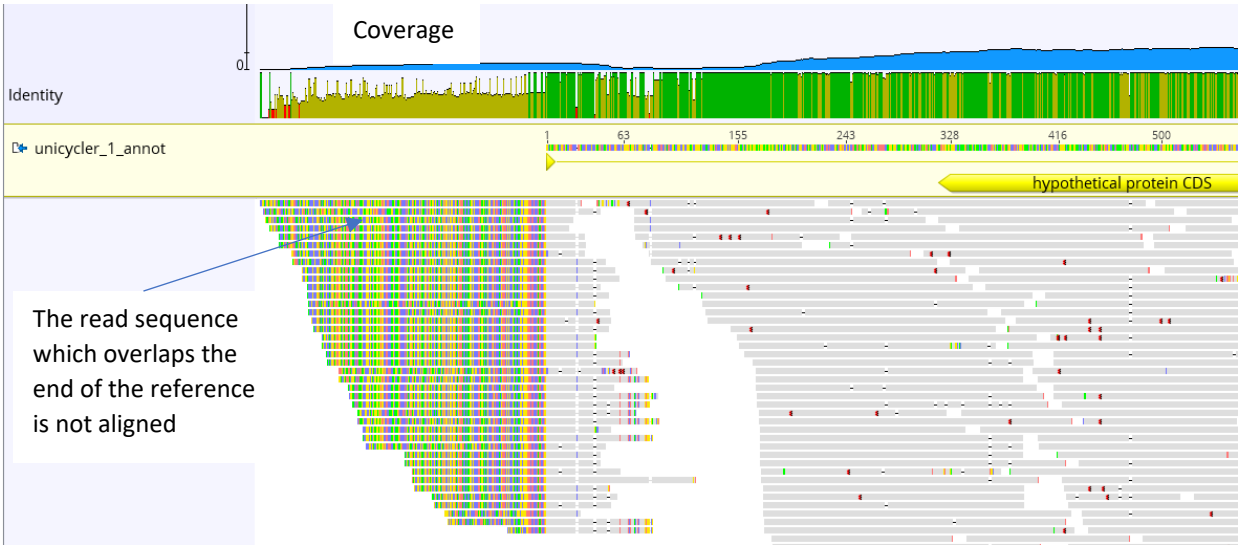


Figure 29. Mapping of *M. chelonae* HPA 006 Ion Torrent reads to the start of Unicycler contig 1.

Similar results occur for each of Unicycler contigs 2 through to 9. Contig 10 is a relatively short contig, which makes it easier to illustrate in these figures

Sets of assembled contigs and reads, namely, MinION reads, Ion Torrent reads, Canu corrected reads, Canu assembled contigs, Flye assembled contigs, Unicycler assembled contigs, MIRA assembled contigs and SPAdes assembled contigs were collected and stored in a folder in Geneious Prime. A sequence, such as those overlapping ends seen in Figure 30 can be searched for in these data sets using BLAST (Basic Local Alignment Search Tool)

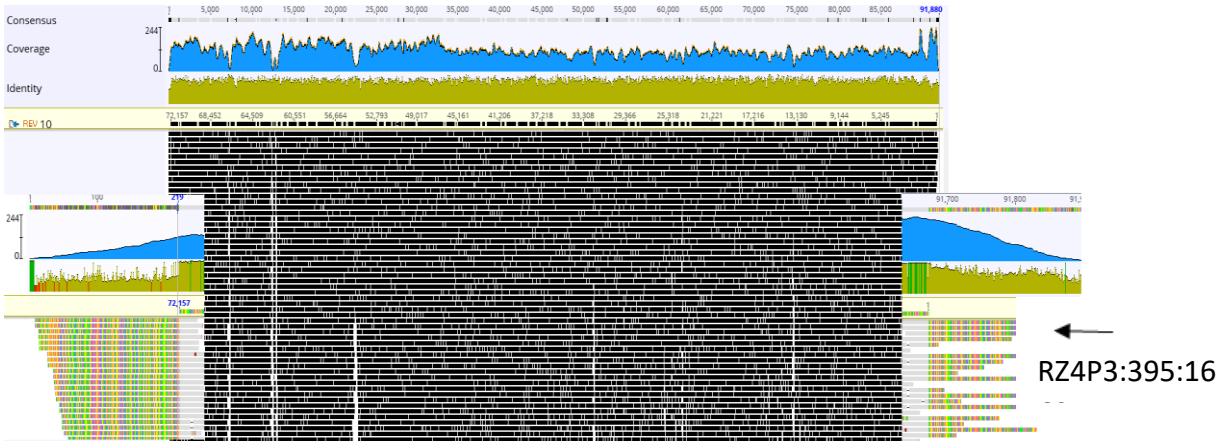


Figure 30. Unicycler contig 10 (reversed) illustrating the ends of *M. chelonae* HPA 006 Ion Torrent read coverage.

Figure 30 illustrates that there are reads overlapping both ends of the contig, suggesting that there is read data which can extend the contigs. The left end is very clear, with numerous reads which will align, the other end is poorer, but only in the sense that the coverage is somewhat lower. Coverage or depth of sequencing equates to the number of sequencing reads uniquely mapped to a reference genome and which therefore cover a known part of the genome. In an ideal situation, these reads would be equally distributed across the reference genome, providing uniform coverage. In actuality, coverage is often uneven. There are a variety of reasons for this, e.g., homologous regions which have similar sequences and the fact that the genome itself is complex with noncoding DNA and repetitive sequences. This can make it difficult to align the sequencing reads to the precise start and end of a genomic element.

Figure 31 further illustrates that when the first Ion Torrent read at the end of contig 10, RZ4P3:395:1669r, is blasted against the MinION reads stored in Geneious Prime, four blast hits (denoted by the symbol ▶) are achieved. The fourth read has a blast hit at the end of a60414f6 and it actually aligns with Unicycler contig 10 i.e., it supports the contig already identified. However, the three other MinION reads, 9f6a86d1, 2f8990ed and b79b3520 also contain this sequence in the middle and hence must overlap, and therefore extend the sequence.

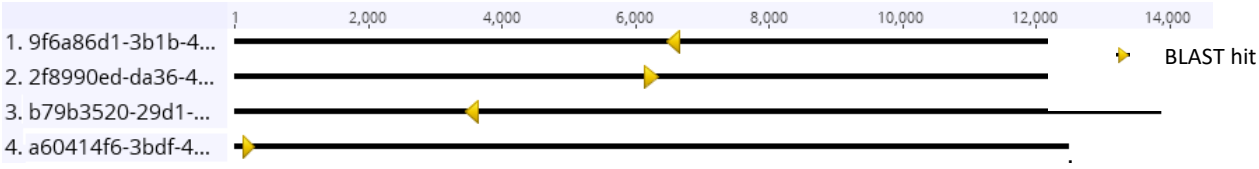


Figure 31. BLAST of Ion Torrent read RZ4P3:395:1669r against all MinION reads.

Blast of these three reads against a database of contigs identify hits, at the end of Unicycler contig 2. Blast against a database of MIRA contigs hits contigs c46 and c84. Assembly of these reads and contigs shows the 3 MinION reads overlapping the ends of both MIRA contigs c46 and c86 and Unicycler contigs 10 and 2 (Figures 32, 33 and 34).

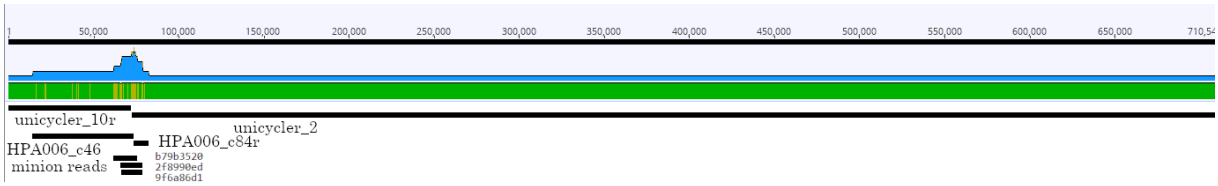


Figure 32. Proposed junction between Unicycler 10 and Unicycler 2 contigs in the *M. chelonae* HPA 006 genome assembly.

This hypothesis then allows collection of Ion Torrent reads from both sides, correction of the MinION reads with Ion Torrent reads, and assembly walking (in which a starting sequence is used as bait, the reads are assembled with MIRA 5 and the new extended sequence is used as bait for the next iteration). Iterating these steps generates a sequence bridging this Unicycler assembly gap.

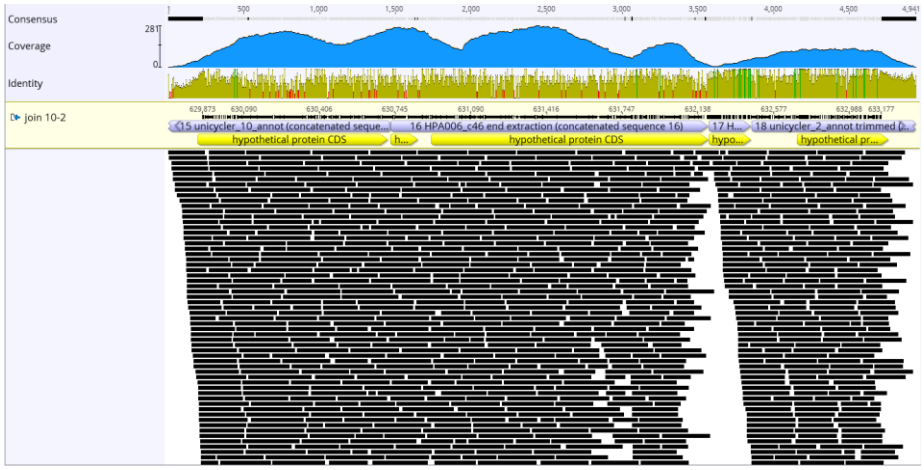


Figure 33. Mapping of *M. chelonae* HPA 006 Ion Torrent reads to the scaffold join for Unicycler contigs 10 and 2.

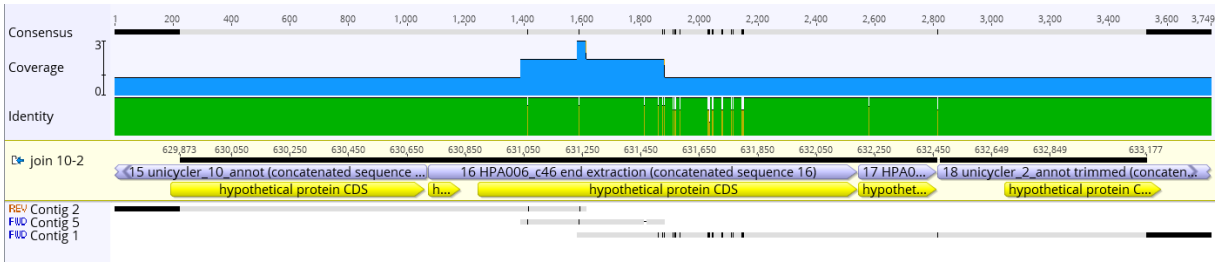


Figure 34. Alignment of *M. chelonae* HPA 006 MIRA 5 contigs to the proposed Unicycler 10: Unicycler 2 scaffold join.

The MIRA 5 assembled contigs map to the assembly of the multiple assembly fragments with 99.4% identity. The generation of multiple contigs, by MIRA, from the mapped reads, reflects the uneven read coverage which is also partly explained by repeat sequences in other parts of the genome (MIRA 5 assembles other contigs which do not map here (data not shown). Applying this strategy to all the Unicycler contigs generates a hypothesis for scaffolding the Unicycler contigs and provides the data to generate a consensus genome from all assembly strategies (Figure 35).

This genome was corrected by MIRA 5 assembling all Ion Torrent reads to the final scaffold as reference genome to give a final corrected genome. This genome was annotated by PGAP when the PGAP annotation pipeline became available for local installation which allowed the joins between assembly fragments to be validated against protein annotations.

The *M. chelonae* HPA 006 genome was aligned with *M. abscessus*^{T} (Figure 36) against *M. abscessus*^{T} (NC_010397 = ref sequence for CIP 104536^{T} = ATCC 19977^{T} = CU458896).

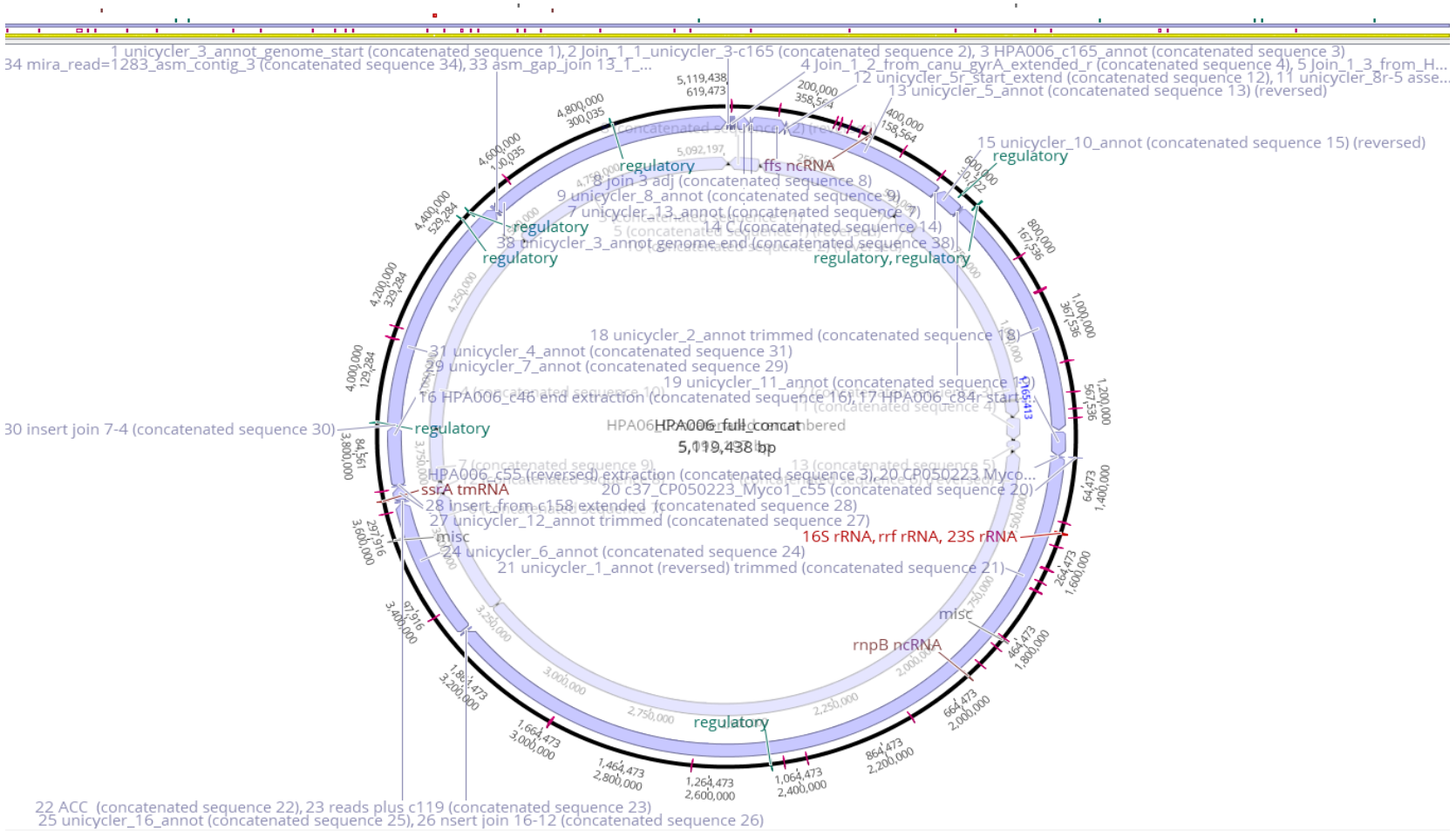


Figure 35. Final scaffolded genome assembly of *M. chelonae* HPA 006.

Inner circle initial assembly of the Unicycler contigs into order, orientation and circularisation, outer circle adjusted scaffold, assembly of multiple fragments from all assembly strategies.

Figure 36 (a) illustrates the Ion Torrent reads of *M. chelonae* HPA 006 mapped onto *M. abscessus*^{T} (NC_010397). An image first illustrated in Figure 15.

The *de novo* assembled genome for *M. chelonae* HPA006 aligned against *M. abscessus*^{T} (NC_010397) is illustrated in Figure 36 (b). This gives a more detailed and a more accurate base to base comparison. It shows not only the major regions of difference obvious in the Ion Torrent mapping but many small differences, particularly exemplified by the 61.4% overall similarity compared to the 82% identity of regions present in both strains. The 82% identity further emphasises that these two strains are different species.

Figure 36(c) compares the *M. chelonae* HPA006 genome to *M. chelonae* M77, one of the most closely similar *M. chelonae* sequences deposited since the start of the project. Despite the high sequence similarity, >99%, there are many regions of difference.
This could reflect rapid acquisition and loss of DNA sequence and means that looking for significant differences between *M. chelonae* and *M. abscessus* will be difficult given the high variation within *M. chelonae* and *M abscessus*.

The first gap between the strains in Figure 36 (c) is a phage insert of genes which are most similar to homologous genes in *M. abscessus* strains, but located in different parts of their genomes, the same gap is seen in the *M. chelonae* HPA 006 – *M. abscessus* alignment. The differences in sequence between *M. chelonae* HPA 006 and *M. chelonae* M77 are almost invariably supported by high confidence sequence data in the *M. chelonae* HPA 006 assembly.

There is one example given in the text here, a phage insert, present in *M. chelonae* HPA 006 but absent in *M. chelonae* M77, one shared with *M. abscessus* strains. Compared with Figure 19 where the biggest region of difference of *M. chelonae* HPA 006 from M. abscessus^{T} is a phage insert in *M. abscessus* not present in *M. chelonae* HPA 006.

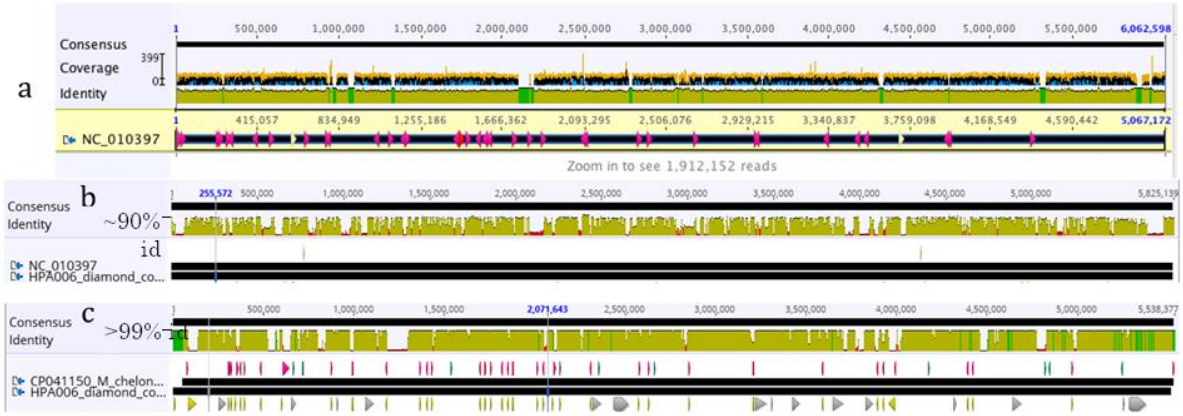


Figure 36. **(a)**. Ion Torrent reads mapped to NC-010397 from Figure 13. (**b)**. *M. chelonae* HPA006 genome aligned against *M. abscessus*^{T} NC-010397 with MAFFT and **(c)**. *M. chelonae* HPA006 genome aligned to *M. chelonae* M77 CP041150 with MAFFT.

Note: NC_010397 is the RefSeq (Reference Sequence) for ATCC 19977T = CIP 104536T = ID CU458896

Chapter 3. Taxonomy of the *Mycobacterium abscessus/chelonae* Clade

3.1 Taxonomic Overview

Fundamentally species are defined by their phenotypic behaviour and interaction, and evolutionary history with populations of similar organisms. "A species is a group of organisms which share a genetic heritage and are able to interbreed and create offspring that are also fertile" (Dictionary of Biology - https://biologydictionary.net/species/ accessed July, 2022). Such a biological definition poses significant problems for prokaryotes. Nevertheless, in the genomes of bacteria we have that evolutionary history and a record of those interactions.

The genetic basis for the delineation of bacterial species, estimating their genetic relatedness (Jain *et al*., 2018), has for many years been the paper by Wayne *et al*., (1987) on DNA: DNA re-association. The criteria proposed by Wayne *et al*., (1987) and Stackebrandt *et al*., (2002) were based on DNA:DNA re-association, and sequencing data from the 16S rRNA gene has underpinned the determination of species boundaries (Konstantinidis *et al*., 2006). However, DNA: DNA pairing has problems with reproducibility and workability (Stackebrandt *et al*., 2002). In contrast to DNA pairing, 16S rRNA gene sequencing is rapid, portable, and unambiguous, but many studies have demonstrated that it is not always discriminatory at the species level taxa (Varghese *et al.*, 2015; Mende *et al*., 2013) and that this approach is dependent upon high precision in determining nucleotide sequence data. Nevertheless, 16S rRNA sequence data has supported for a rapid expansion of the numbers of validly described species. These issues were addressed by Stackebrandt *et al*., (2002) who concluded that species description based on a single gene seemed ill-advised.

The application of second and third generation sequencing has resulted in the availability of genomic information on a diverse range of prokaryotic species, which is deposited and therefore accessible on public databases. When bacteria evolve, most of the functional gene transmission takes place vertically, but prokaryotic genomes are also subject to horizontal gene transfer and this can result in an extensive exchange of functional genomic DNA (Pritchard *et al.*, 2016). If the intention of taxonomy is to provide a classification system that is functional, predictable and reproducible, then utilising these deposited data with a new approach can give a better correlation between taxonomy and phenotype, for example, in

clinical prediction of mechanisms of infection, strategies for therapeutic intervention and case outcomes based upon accurate identification.

3.2 Taxonomic Tools

3.2.1 Average nucleotide identity

ANI (Average Nucleotide Identity) is a calculation of the relationship of all nucleotides within a genome to those of the genome of a purported related species, it is an *in silico* method with results which could best represent those achievable using DNA:DNA pairing (Richter and Rossello-Mora, 2009). The aim is to find all the nucleotides of one genome which have a matching nucleotide in the comparator genome. These matches are termed "best hits" and allow a numerical relationship to be derived as a percentage value. ANI can be calculated using both best hits (one-way ANI) and reciprocal best hits (two-way ANI) between two genomic datasets. Two-way ANI (i.e., comparing A with B and B with A) reduces the likelihood of miscalculation due to variation in the alignment of individual nucleotides within a genome.

Typically, the ANI values between genomes of the same species are above 95% which is seen as comparable to the 70% DNA: DNA re-association standard (Goris *et al*., 2007). The problem is that these calculations were difficult when limited to a small number of genomes, and even more complex when multiple strains were to be assessed against an equally large set of comparator strains. This situation can be overcome with the availability of specifically designed software packages e.g., Pyani. Pyani is a Python3 module that provides support for calculating ANI for multiple whole genome comparisons (Pritchard *et al.*, 2016).

3.2.2 Pangenome analysis

The concept of species has undergone considerable change in the years since Wayne *et al*., proposed DNA: DNA hybridisation as a cornerstone of the definition. This has particularly occurred due to the availability of faster and less expensive genomic analysis. Baumdicker *et al*., (2012) postulated that whilst individual bacterial cells have compact genomes, at the level of a species population a higher number of genes exist. This distributed genome of a bacterial species can be assessed using pan-genomics.

A single analysis and description of the genome of a strain of a species may be unique but nevertheless contain variations from other genomic analyses of strains of the same species. That is, the full genomes of several strains of the same species may contain variations. Thus, the sequence of a single genome does not reflect the entire genetic variability of a bacterial species (Costa *et al*., 2020). Variations may include deletions, insertions or substitutions or even loss of a more significant area of the genome, perhaps encompassing a whole gene. Taken in isolation the presence of such regions of difference may challenge the species definition of a particular strain of an organism.

In a pan-genomic analysis, all available deposited genomes (full or partial) are comparatively analysed. In principle, the pangenome is derived by analysis of multiple copies of the available genomes as deposited in a gene bank. The resulting pangenome is made up of the core genome, dispensable or accessory genome and singleton or cloud genes (i.e., species-specific genes). Within the genomes of organisms assigned to a broad "species" group, the core genome is the set of genes present in all the strains. These genes will have vital roles for strain survival. Accessory or dispensable genes are those which are present in only a fraction of the genomes. Cloud genes are present in only one or a few genomes. Accessory and cloud genes may have been acquired by horizontal gene transfer (Treangen & Rocha, 2011) or by extensive mutation of a pre-existing gene.

These genes may have become established by conferring an evolutionary benefit such as drug resistance or virulence. In what is termed a "closed" pangenome, a more dominant core genome (and thus limited accessory genome) will result in a highly stable pangenome (e.g., *Bacillus anthracis*; Park *et al*., 2019). Pathogenic and symbiotic species tend to have closed pangenomes, having evolved away, in part or whole, from an environmental existence. In contrast, an "open" pangenome will have a greater composition of accessory genetic material and will continue evolving with the acquisition of more genes. Species such as *Escherichia coli* are reported to have open pangenomes (Park *et al*., 2019). Mobile genetic elements and hypervariable regions are found in many genomes and these may form part of the accessory genetic material.

Analogous genes code for similar functions but are otherwise unrelated; by contrast, homologous genes are derived from a common ancestor. Pangenomic analysis seeks out homologous genes. Homologous genes may be orthologous or paralogous. Orthologous genes

are related by vertical descent from a common ancestor and may encode proteins with the same function in species that have diverged from one another. Paralogous genes have evolved by duplication (with some copy variation) and may code for proteins that show similar but not identical functions. However, whilst orthologous genes are likely to be highly conserved, paralogous genes may have mutated after duplication resulting in a functional change.

A pan-genomic analysis presents all the gene variability of a group of organisms. The set of genes shared among all organisms as well as species- or strain-specific genes are also a source of extremely useful information. All of these data allow an improvement of time and technology in different areas of biology and bioinformatics.

For example, traditional methods of vaccine development require cultivation of large amounts of the target microbe followed by laborious experimentation to determine inhibitory manipulations to validate the vaccine target. In contrast a comparison of several examples of the genomes of a target species can allow identification of common essential proteins which may be potential vaccine targets. The approach is termed reverse vaccinology method and was first applied to the serogroup B meningococci (Pizza *et al*., 2000). The method has advantages over classic approaches of vaccine development because it is less laborious, less costly, and more accurate in choosing a gene target. Several reverse vaccinology studies have now used pan-genomics to determine the main targets for vaccine development (Seib *et al*., 2012).

The host-pathogen interaction can be evaluated at the genomic level through genes that are responsible for processes such as adhesion, invasion, and toxin production. Therefore, a pan-genome analysis helps to define which virulence genes are shared among all pathogenic species, as well as which genes are specific to one isolate. This has direct implications for understanding the evolution of pathogenic species. Pan-genomic analysis has been increasingly used to assist in the taxonomic classification of microorganisms (Caputo *et al*., 2019), to determine a set of molecular markers for phylogenomic analysis (Velsko *et al*., 2019) and in the analysis of multiple pathogenic isolates of *Streptococcus agalactiae* (Tettelin *et al*., 2005).

The concept of what constitutes a species is being challenged, largely due the possibilities opened by sequencing of whole genomes. Hitherto the criteria proposed by Wayne *et al*., (1987) and Stackebrandt *et al*., (2002) based on DNA:DNA re-association and sequencing data

from the 16S rRNA gene respectively, have largely underpinned this determination of species boundaries for the last several decades. Recently, Bobay & Ochman, (2017) proposed that bacterial strains should be classified as the same species only if they show an intra-group rate of allele exchange (gene flow) which is greater than the rate between that group and any other strains. The concept of a distributed genome hypothesis, proposed by Baumdicker *et al.,*(2012) supports this.

By studying the pangenome of the species within the *M. abscessus*/*chelonae* clade this study hopes to identify genes in *M. abscessus* subsp *abscessus* which, if they are responsible for conferring a greater degree of virulence, pathogenicity or resistance to antimicrobial agents, have the potential to explain the differences in the clinical outcome of infection with this organism versus *M. chelonae* in patients with cystic fibrosis.

3.3 Materials and Methods

3.3.1 Average nucleotide analysis with pyani

There are multiple tools for ANI, some specifically for taxonomy, ANI Calculator at EZBioCloud (Joon *et al*., 2017), the genome-to-genome calculator (GGDC) at DSMZ (Meier-Kolthoff *et al*., 2022) or the ANI calculator at the Kostas Lab (Goris *et al*., 2007). However, Pyani (Pritchard *et al*., 2016) offers a comprehensive package to calculate ANI between thousands of genomes, both complete genomes and multi-contig WGS.

Pyani was run in ubuntu 18.04 in Windows Subsystem for Linux (WSL2) on a Dell XPS 15 9500 laptop with 10th generation core i7 processor, 32Gb RAM and 1Tb SSD hard disk or an ASUS ROG Strix scar 17 similarly configured, or a native Ubuntu 18.04 installation on a Dell HP with 4th generation core i7 and 32Gb RAM.

It was installed within an Anaconda base environment set up for Bioconda and requires MUMmer (NUCmer: github.com/gmarcais/mummer) to produce the ANIm output for each species.

R (R Core Team, 2013) was installed with RStudio (RStudio Team, 2020) for windows 10.

Pyani has a convenient Python script to retrieve whole genome sequence data from the NCBI website based on the NCBI taxid taxonomy. This allowed rapid retrieval of all the genomes of

M. abscessus, M. chelonae, M. franklinii, M. immunogenum, M. salmoniphilum, M. saopaulense and M. stephanolepidis for all initial exploratory data analysis. (see Supplementary Data File (S1) for a list of genomes). Additionally, as part of the output, it prepares graphical summaries of the data, which simplified that exploratory data analysis phase.

The command-line interface to Pyani uses subcommands. These separate the individual steps of an analysis into distinct actions. The following processes are common to any of the analyses which were carried out in this study. The version of pyani used was https://github.com/widdowquinn/pyani/blob/master/README_v_0_2_x.md

- Download required genomes. Pyani requires each genome to be individually presented in a FASTA file format. The FASTA files were placed in a Pyani subdirectory created in UBUNTU 18.04.
- Create a database to hold genome data and analysis results
- Perform ANI analysis
- Each analysis produced an output file, which was moved back into Windows
- The percentage identity file for each Pyani analyses was analysed further using R which has software designed for integrated data manipulation
- Generate species hypotheses (classify genomes) using the analysis results

Output is as tab-separated plain text format tables describing:

- alignment coverage
- total alignment lengths
- similarity errors
- percentage identity (ANIm)

The percentage identity (ANIm) was used for further analysis in R (R Core Team, 2013), the R environment is a collection of software facilities designed for integrated data manipulation, calculation, and graphical display. It assembles and runs on a wide variety of UNIX platforms, in addition to Windows and MacOS. The Comprehensive R Archive Network (CRAN) is available at The Comprehensive R Archive Network. R studio which is the graphical user interface (GUI) for R can be downloaded at http://www.rstudio.com/ide

The tsv file for each dataset was used to generate cluster heatmaps using the R package Heatmaply (Galili *et al.*, 2018). A tooltip display allows the user to visualise values whilst hovering over cells. The zoom capability allows closer inspection of specific areas in the heatmap and there is the additional advantage of interactive relationship with other R packages such as ggplot2 (Wickham, 2009), plotly (Sievert *et al*., 2016), dendextend (Galili, 2015).

Interactivity includes a tooltip display of values when hovering over cells, as well as the ability to zoom in to specific sections of the figure from the data matrix, the side dendrograms, or annotated labels. Thanks to the synergistic relationship between heatmaply and other R packages, the user can, if required, have control over the statistical and visual aspects of the heatmap layout. The heatmaply package is available under the GPL-2 Open Source license and is freely available from: http://cran.r-project. org/package/heatmaply.

3.3.2 R analysis

R code for ANI analysis

#load library for heat map and library for plot3d

>library(heatmaply)

>library(rgl)

#read in tab delimited output from pyani – file in R's default working dir

>Mab_rpoB_clus_sim <- read.table("Mab_rpoB_clus0_ANIb_percentage_identity.tab", header = TRUE, sep = "\t", row.name = 1)

#plot a histogram of similarities in the matrix with 100 bins

> hist(as.matrix(Mab_rpoB_clus_sim), breaks = 100)

#do the heatmap – revC to get the diagonal top left to bottom right – use 256 viridis colors and reverse so v similar is dark

> heatmaply(Mab_rpoB_clus_sim, revC = TRUE, col = rev(viridis(256)))

#get a distance matrix for plotting

> Mab_rpoB_clus_dist <- 1 - as.matrix(Mab_rpoB_clus_sim)

#calculate 3d positions by principal co-ordinates analysis

> Mab_rpoB_clus_points <- cmdscale(Mab_rpoB_clus_dist, k = 3)

#plot the points as blue spheres

> plot3d(Mab_rpoB_clus_points, type = "s", col = "blue", size = 1)

#add text labels positioned starting at centre of point and text ½ size

> text3d(Mab_rpoB_clus_points, text = rownames(Mab_rpoB_clus_points), adj = c(0,0), cex = 0.5)

#or enable add labels by pointing and right-clicking – left click quits, quits if move plot (re-run to add more)

>identify3d(Mab_rpoB_clus_points, labels = rownames(Mab_rpoB_clus_points), plot = TRUE, buttons = c("right", "left"))

3.3.3 Pangenome with PPanGGOLiN

PPanGGOLiN (Partitioned PanGenome Graph Of Linked Neighbors) presents the gene repertoire and its variations in a graph format. Each node represents an homologous gene family and each edge is indicative of a relation of genetic contiguity (Gautreau, *et al*., 2020).

One possible drawback of this approach could be the fact that polymorphisms in genes are ignored, and variations in intragenic regions and introns are overlooked. However, prokaryotic genomes have small intragenic regions and have almost no introns; therefore, this approach has merit (Koonan & Wolf, 2008).

In comparison to other software packages PPanGGOLiN can produce more cloud gene partitions but for this analysis the focus is on persistent and shell genes. PPanGGOLiN is easy to install and run and will accept FASTA and multi-FASTA genome files and annotate all the genomes consistently using Prodigal (Hyatt *et al*., 2010).

PPanGGOLiN Materials and Methods

PPanGGOLiN was installed on Ubuntu 20.04 running under wsl 2 on Windows 10 or 11 with Anaconda and Mamba. Anaconda was installed as conda version 4.1.10 from

https://repo.anaconda.com/archive/Anaconda3-2021.05-Linux-x86_64.sh and updated to 4.1.11 during the project. Conda channels defaults, R, BioConda and conda-forge were installed into a base environment. Mamba, a faster replacement for the conda core written in C++ is recommended for installation of PPanGGOLiN and was installed in the base environment with the command:

$ conda install -n base -c conda-forge mamba

$ mamba update -n base mamba

to version 0.7.3

Then PPanGGOLiN was installed, in its own environment within the base environment:

(base) anne@DESKTOP-17LEH4R:~$ mamba create -n pangenome PPanGGOLiN.

(base) anne@ DESKTOP-17LEH4R:~$ conda activate pangenome

(pangenome) anne@ DESKTOP-17LEH4R:~$ ppanggolin --version

ppanggolin 1.1.136

3.3.4 Genomes

Genome sequence data was retrieved from NCBI for *Mycobacterium* (*Mycobacteroides*) *abscessus* subsp. *abscessus, bolletii* and *massiliense*, *M. chelonae*, *M. franklinii*, *M. salmoniphilum* and *M. immunogenum.* The reference rpoB gene sequences from the type strains (except *M. chelonae*, where the type strain is not representative and rpoB from *M. chelonae* HPA 006 was utilised) were used to identify the genomes of members of the genus using blast. There was only one whole genome for *M. stephanolepidis* and four for *M. saopaulense*. This procedure retrieved genomes for this taxonomic group even if not identified in the NCBI tax id and were added to the genome sequence for *M. chelonae* HPA 006.

The rpoB sequences, retrieved by blast, were aligned with MAFFT (Katoh & Standley, 2013) in Geneious (Geneious Prime: Geneious Prime 2022.0.1 http://www.geneious.com/) and the phylogenetic tree viewed in Dendroscope (Huson *et al*., 2007) and the source data for rpoB data in the clade corresponding to each species identified.

All the genomes were retrieved from complete genomes or as multi-contig WGS, as full genbank (gb) or .gbff format files and read into geneious into a folder for each species. The names in geneious were adjusted to – genbank id_species_name_strain - joined with underlines (no spaces or punctuation).

For analysis in PPanGGOLiN files were selected in Geneious and exported to an analysis folder in FASTA format, supplying PPanGGOLiN with input files in FASTA format ensured that all genomes were annotated in the same way. The input to PPanGGOLiN is a file with a label to be used to identify each genome followed by the path to the corresponding FASTA file

e.g., CP007220_M_chelonae_CCUG47445 CP007220_M_chelonae_CCUG47445.FASTA

This file was generated using the ls command with output redirected to a file

(pangenome) anne@ DESKTOP-17LEH4R:~/M_chel_pan: $ ls > M_chel_pan_lst

The file names were copied in block select mode in Textpad, the .FASTA extension deleted and pasted with a tab separator, to the start of each line to generate the PPanGGOLiN input.

PPanGGOLiN was run as:

(pangenome) anne@ DESKTOP-17LEH4R:~/M_chel_pan: $ ppanggolin panrgp --cpu 8 --FASTA M_chel_pan_lst which generates an output folder with a unique name e.g., ppanggolin_output_DATE2021-11-03_HOUR07.58.28_PID24979.

Instructions for PPanGGOLiN are in the wiki at

https://github.com/labgem/PPanGGOLiN/wiki/Introduction (last accessed July 2021)

3.4 Results

3.4.1 ANI analysis

With the description of *Mycobacterium salmoniphilum* in 2007 (Whipps *et al*., 2007) the group of mycobacterial species referred to as the *M. abscessus/chelonae* clade were composed of *Mycobacterium chelonae* (Bergey *et al*., 1923), *Mycobacterium abscessus* (Kusunoki & Ezaki, 1992), *Mycobacterium immunogenum* (Wilson *et al*., 2001), *Mycobacterium massiliense* (Adékambi *et al*., 2004), and *Mycobacterium bolletii* (Adékambi *et al*., 2006b). To these have been added *Mycobacterium franklinii* (Nogueira *et al*., 2015a), *Mycobacterium saopaulense* (Nogueira *et al*., 2015b) and *Mycobacterium stephanolepidis* (Fukano *et al*., 2017a). The

taxonomic relationships of *Mycobacterium abscessus* variants (*M. abscessus, M. bolletii* and *M. massiliense*) was rationalised (Tortoli *et al*., 2016) to create subspecies, *M. abscessus* subsp. *abscessus*, *M. abscessus* subsp. *bolletii* and *M. abscessus* subsp. *massiliense* though there is still debate over these taxonomic designations.

3.4.2 Analysis of species in the M. abscessus/chelonae clade

A total of 471 representative strains of the *M. abscessus*/*chelonae* clade were downloaded as FASTA files from the NCBI database using the Pyani script. This provided a cohort of samples, representative of the clade, for ANIm analysis (Figure 37).

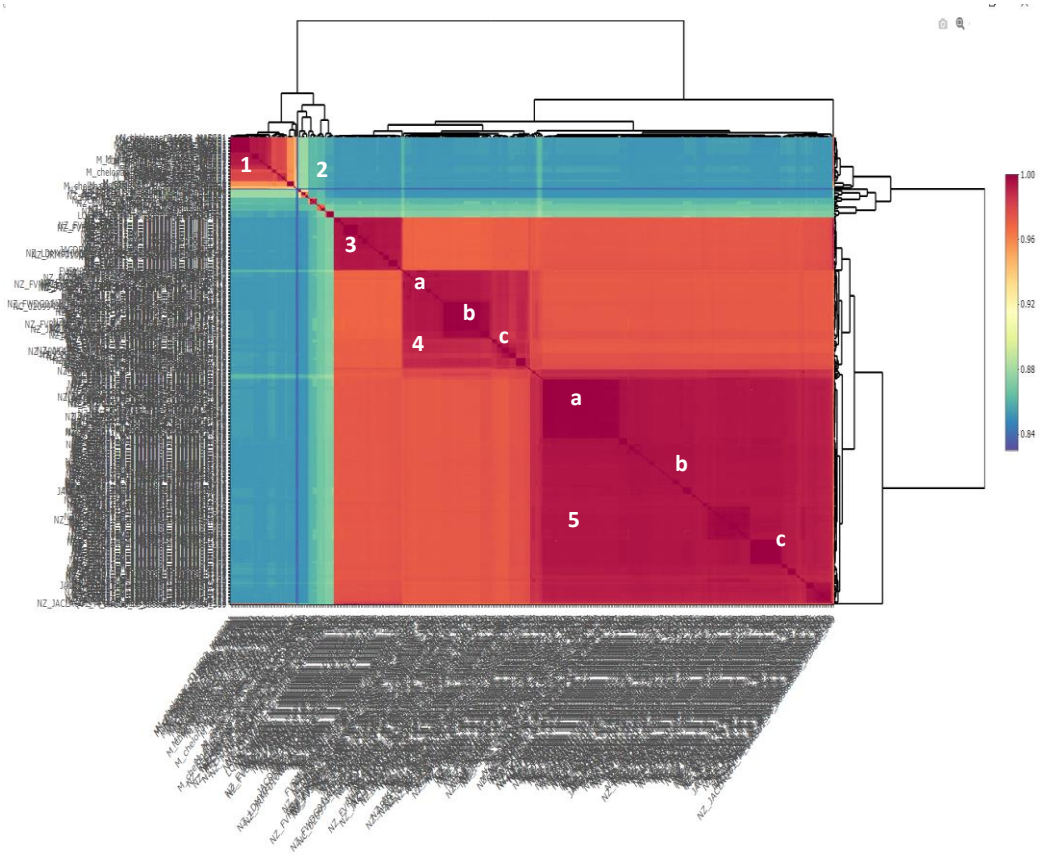


Figure 37. Heatmap of average nucleotide identity (ANIm) for 1. *M. chelonae* (including *M. chelonae* HPA 006) and *M. stephanolepidis* 2. *M. immunogenum, M. salmoniphilum, M. franklinii, M. saopaulense* 3. *M. abscessus* subsp. *bolletii* 4. *M. abscessus* subsp. *massiliense* 5. *M. abscessus* subsp. *abscessus*. a, b and c represent putative subclusters of strains within the main clusters.

It is clear that the number of strains analysed makes this representation of the data difficult to read within the constraints of an A4 page but the blocks of strains corresponding to individual species have been labelled and are analysed below. There is also a PDF of the original pyani generated (percentage identity file) output as a Supplementary File (S2), with the position of *M. chelonae* HPA 006 highlighted.

Cluster 1, *M. chelonae* strains, all show a ≥ 98% average nucleotide identity with the exception of strain NCTC 946 (nominally the originally deposited type strain) about 82%* similarity. *M. stephanolepidis* has average ANI values comparable to the other *M. chelonae* strains of about 95% while the sequence data for multiple culture collection samples of the type strain are distinct from the rest of the species at about 96%. The sequence for NCTC 946 matches *M. phlei*, by 16S, rpoB and ANI – presumably *M. phlei* NCTC 8156, which is an *M. chelonae*.

* There are 470 individual ANI values which show some variation.

Cluster 2. All of the representatives of *M. franklinii*, *M. immunogenum*, *M. salmoniphilum* and *M. saopaulense* have clustered in area 2, all of the clusters are distinct and all strains display a ≥98% average nucleotide identity within their separate clusters. A strain, *M. chelonae* S00154 is also observed in this cluster and is most closely related to *M. saopaulense*, though, like *M. stephanolepidis*, at the border-line of new species identity.

Cluster 3. Strains of *M. abscessus* subsp. *bolletii* which have clustered in area 3. There are several groups of strains observed, all strains display a ≥98% average nucleotide identity, these groups represent clusters of nearly identical strains. There is a strain *M. abscessus* subsp. *abscessus* (JACDRH01) which is clearly a member of this species.

Cluster 4 predominantly includes strains which have been identified as *M. abscessus* subsp. *massiliense* which have clustered here. Although there are some groups with high similarity discerned in the data all strains display a ≥98% average nucleotide identity to one another, and any subgroups comprise clusters of very high identity. There are more strains which seem to have been misidentified, 12 strains described as *M. abscessus* subsp. *bolletii* and 1 *M. abscessus* subsp. *abscessus.* At least 3 strains, designated as *M. abscessus* subsp. *bolletii* in NCBI were submitted as *M. abscessus* subsp. *massiliense*.

Although these strain identifications are not taking place in a routine clinical setting it is clear that few of these strains appear to be misidentified.

A 3D principal coordinates plot of the ANI similarities (percentage similarities as rendered by the ANI output converted to distances from a median point) for representative strains from each of the clusters (there are too many data points and too many labels in a full analysis) shows the same pattern. At this resolution, scaled to the distances between species, the subspecies in *M. abscessus* do not separate (Figure 38).

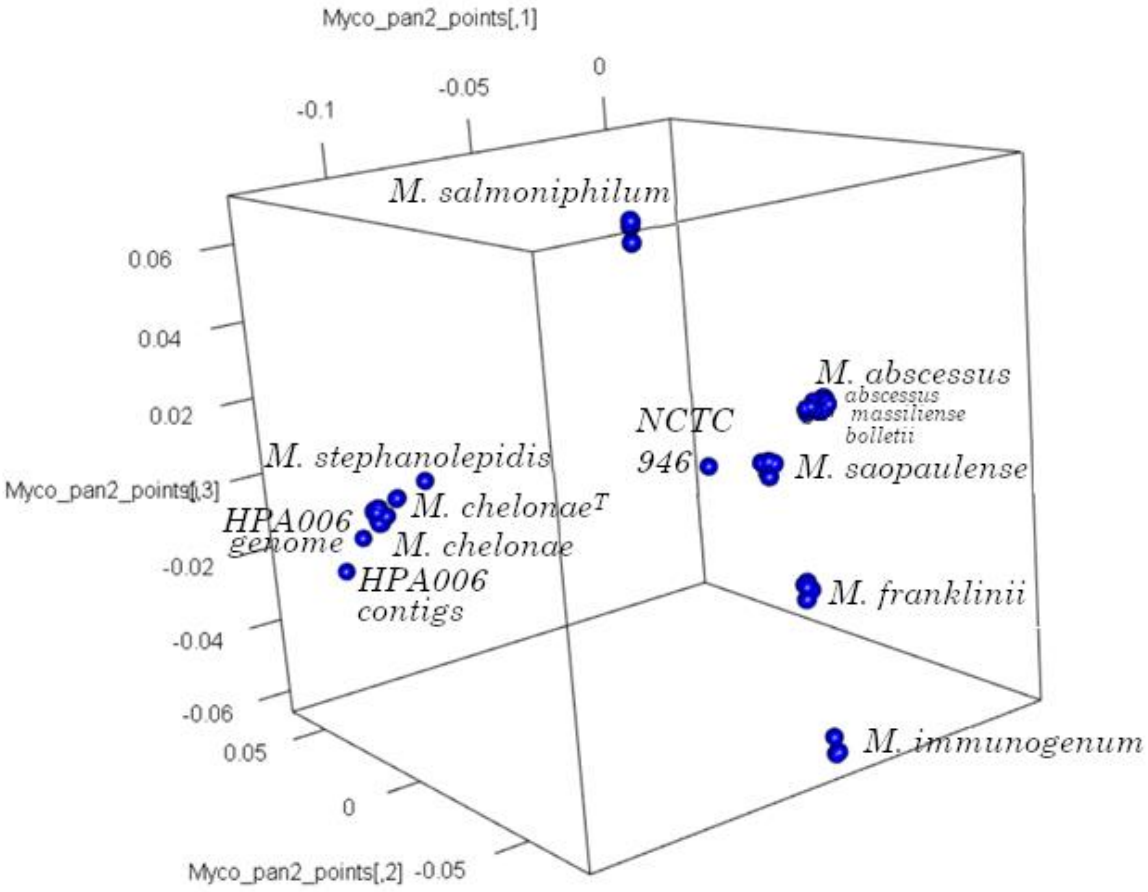


Figure 38. 3D ordination plot of Average Nucleotide Identity (ANI) analysis of the representative strains of the *M. abscessus chelonae* taxonomic clade.

A plot of the histogram for all the ANI similarities shows two peaks at close to 99% which corresponds to distances between members of the same species while a second peak at about 97.5% may correspond to distances between strains that are members of different subspecies. The basis of the 70% DNA:DNA re-association, as a guide to the separation of species, proposed by Wayne *et al*., (1987) was that this figure corresponded to a minimum between

DNA:DNA re-association data available to them, perhaps capturing a discontinuity in organism similarities at the natural species boundary. For the *M. abscessus* species there is a clear discontinuity in the data which may represent the subspecies separation (Figure 39). Similarities between members of different species are represented by the peaks at around 0.85% similarity.

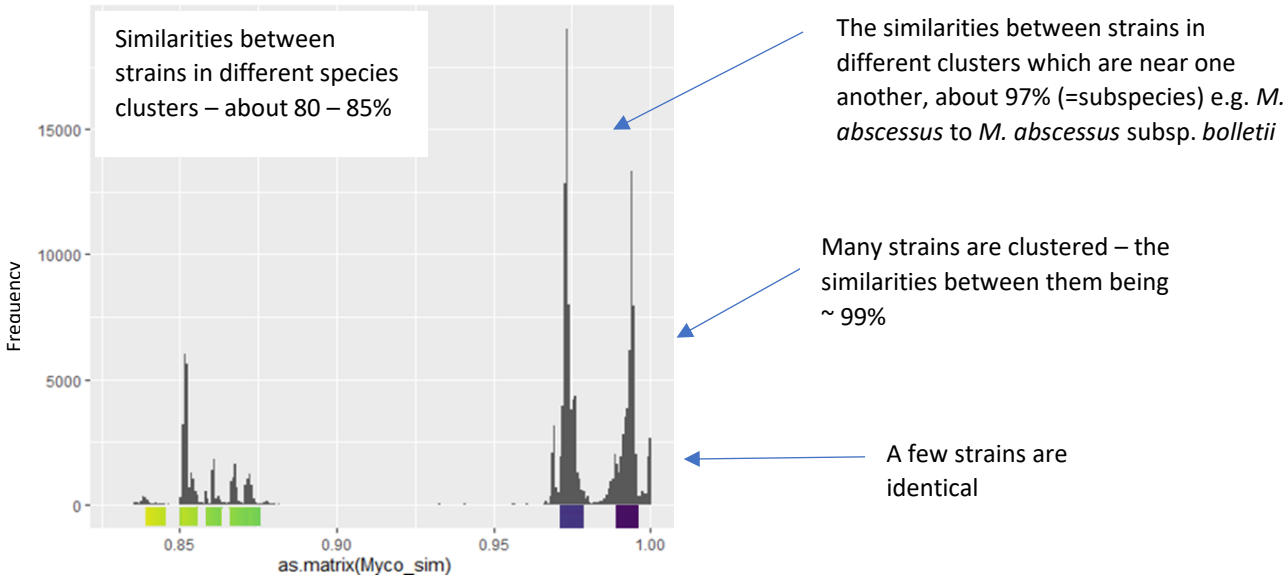


Figure 39. Histogram of ANI similarities from all vs all *M. abscessus*/*M. chelonae* strains.

3.4.3 Analysis of M. abscessus and subspecies

The ANI analysis of *M. abscessus* strains is shown in Figure 40

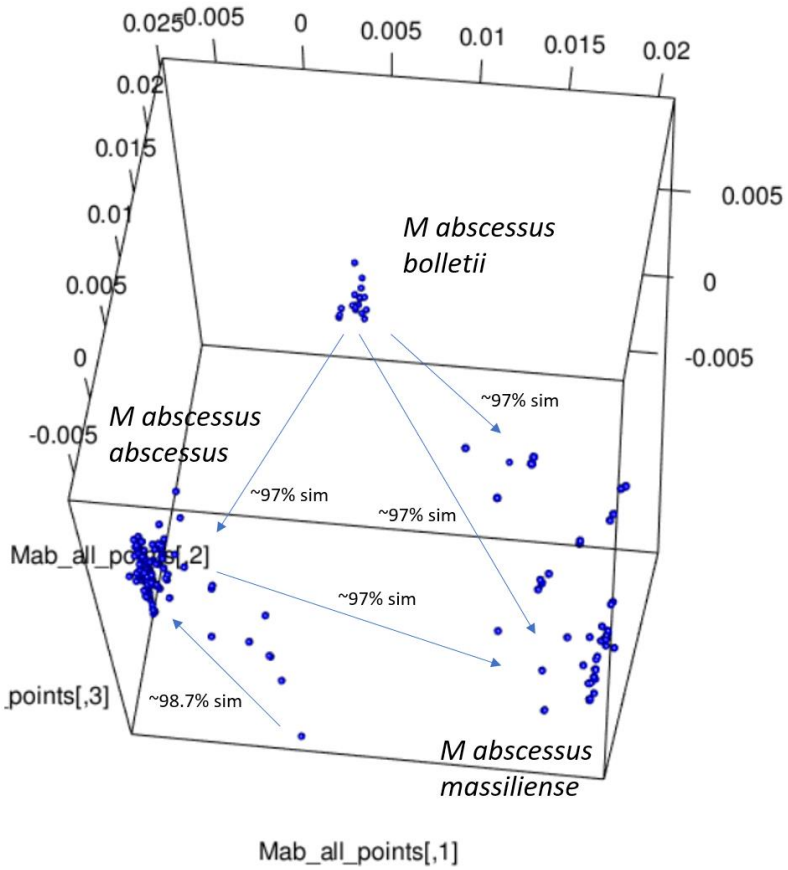


Figure 40. Average Nucleotide Identity analysis of *M. abscessus* showing the separation into the 3 common subspecies present: *M. abscessus* subsp. *abscessus*, *M. abscessus* subsp. *bolletii* and *M. abscessus* subsp*. massiliense.*

The plot in Figure 39 can be visualised in the ANI analysis of *M. abscessus* and its subspecies in Figure 40. A clear stable *M. abscessus* subsp. *bolletii* cluster is evident here, but a core subspecies for *M abscessus* subsp. *abscessus* with a short trail of strains emerging at 45^{o} (inside the red circle), to the strain at the bottom middle is evident. This could perhaps represent the act of evolving as there is a continuum of variation and no discontinuity or gap to that last point, which may be the nucleus for a new subspecies, if it finds a niche in which it is successful. Similarly, the *M. abscessus* subsp *massiliense* cluster is showing a lot of variation with the group circled in blue, potentially a new subcluster. It would be interesting to try and correlate this diversity of the *M. abscessus* subsp *massiliense* cluster based on the source of the strain as either clinical or environmental.

3.4.4 Analysis of M. chelonae and subspecies

The analysis of *M. chelonae* strains was carried out omitting the sequence for NCTC 946^{T} as it is identical to *M. phlei.* Additionally, it was established that the other deposited type strains of *M. chelonae*, namely, ATCC 35752, ATCC 35752.1, ATCC 35752.2, CCUG 47445 and DSM 43804 were duplicates of one another. Of these, only ATCC 35752 was included.

M. chelonae S00154 (JACHLF01) which, although deposited in NCBI as a chelonae species, was found to be derived from a strain of *M. immunogenum.* The chelonae analysis is shown in Figure 41.

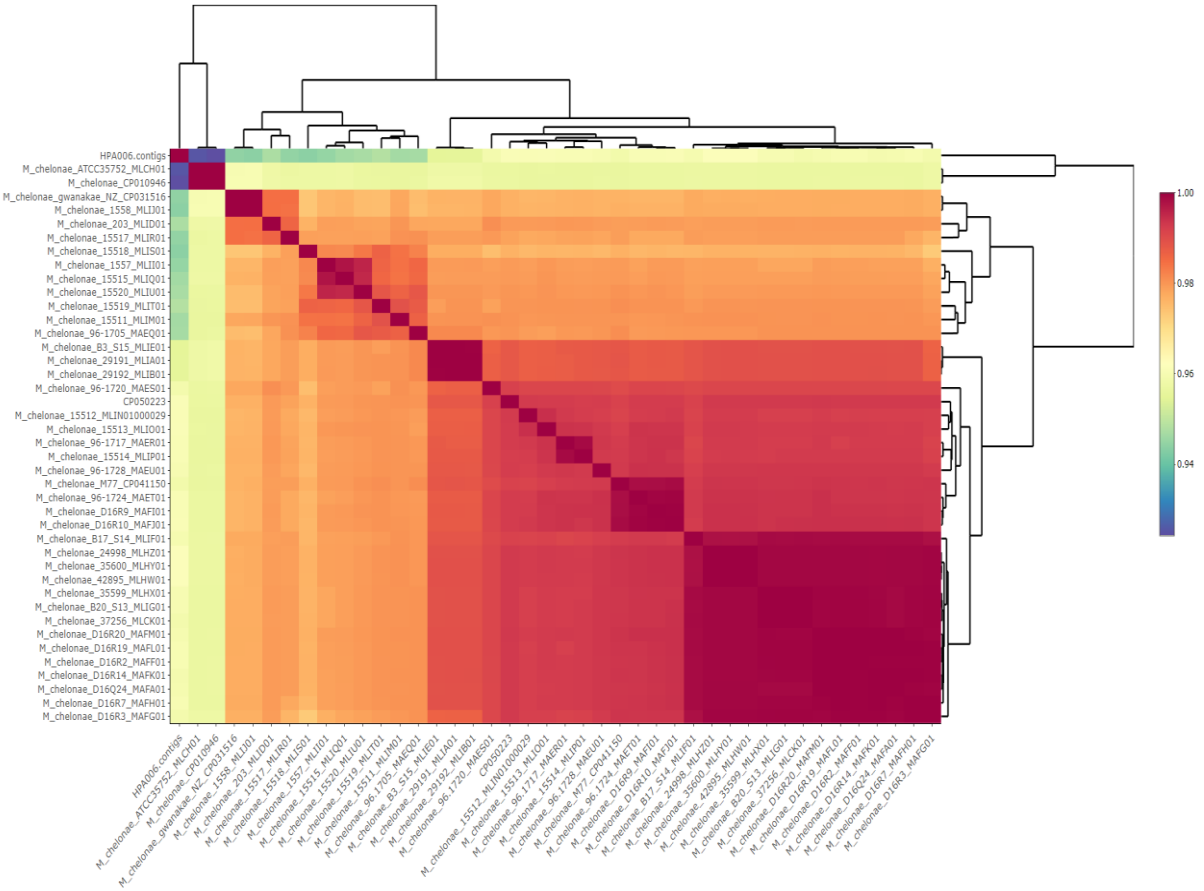


Figure 41. *M. chelonae* strains ANIm percentage identity Heatmap and Dendrogram minus aberrant strain data and with multiple Type strain data inclusions removed.

The relationship between *M. chelonae* strains is analysed further in the PPanGGOLiN analysis (Section 3.5.2).

3.4.5 PPanGGOLiN analysis

The 1526 genomes downloaded as members of the *M. abscessus/chelonae* clade, based on rpoB analysis, required too much RAM to analyse with PPanGGOLiN (Bazin *et al*., 2020; Gautreau *et al*., 2020) in a 32 Gb RAM linux computer, so the genomes were reduced to 870 *M. abscessus* subsp. *abscessus*, 119 *M. abscessus* subsp. *bolletii*, 362 *M. abscessus* subsp. *massiliense*, 53 *M. chelonae*, 12 *M. franklinii*, 7 *M. salmoniphilum* and 12 *M. immunogenum*. Genomes which were identical or highly similar and multi-contig whole genome sequences (WGS) with large numbers of contigs were reduced and *M. stephanolepidis* and *M. saopaulense* were left out.

Analysis of 1435 genomes was successful and generated a tab separated matrix of locus tags for each genome with each row containing the homologous genes from each genome, calculated from the annotated genes, using Prodigal (Hyatt *et al*., 2010). The genomes were submitted, ordered to cluster the *M. abscessus*/*M. chelonae* clade species together, so that the columns in the matrix were ordered by species. This tab separated matrix was read into excel and the number and percentage of each gene in each species and subspecies calculated (Figure 42).

	Gene	ATCC_19977 CDSs	Non-unique Gene name	Annotation	No. isolates	No. sequences	Avg sequences per isolate	M franklinii	M franklinii %	M abscessus abscessus	M abscessus abscessus %	M abscessus bolletii	M abscessus bolletii %	M abscessus massiliense	M abscessus massiliense %	M chelonae	M chelonae %	M salmoniphilum	M salmoniphilum %	M immunogenum	M immunogenum %
3	NZ_FSJM01	CDS_0001	shell		163	163	1	0	0	140	16.09195	1	0.840336	15	4.143646	4	7.54717	0	0	0	0
4	NZ_FRYO01	CDS_0002	shell		166	166	1	0	0	143	16.43678	1	0.840336	16	4.41989	4	7.54717	0	0	0	0
5	NZ_FSMX01	CDS_0003	shell		168	168	1	0	0	144	16.55172	1	0.840336	16	4.41989	4	7.54717	0	0	0	0
6	NZ_CAACXQ	CDS_0004	shell		101	102	1.01	0	0	99	11.37931	0	0	1	0.276243	0	0	0	0	0	0
7	NZ_FSFP01_	CDS_0005	shell		119	121	1.02	0	0	110	12.64368	0	0	8	2.209945	0	0	0	0	0	0
8	NZ_FWDC01	CDS_0006	cloud		10	10	1	0	0	10	1.149425	0	0	0	0	0	0	0	0	0	0
9	FVBY01_M_	CDS_0007	shell		100	101	1.01	0	0	99	11.37931	0	0	1	0.276243	0	0	0	0	0	0
10	FVBY01_M_	CDS_0008	shell		100	101	1.01	0	0	99	11.37931	0	0	1	0.276243	0	0	0	0	0	0
11	NZ_FVYG01	CDS_0009	cloud		11	11	1	0	0	10	1.149425	0	0	1	0.276243	0	0	0	0	0	0
12	NZ_FWDC01	CDS_0010	cloud		11	11	1	0	0	10	1.149425	0	0	1	0.276243	0	0	0	0	0	0
13	NZ_FVXF01	CDS_0011	cloud		30	30	1	0	0	29	3.333333	0	0	1	0.276243	0	0	0	0	0	0

Figure 42. Calculation of the number and percent of each gene present in the genomes of *M. franklinii, M. abscessus* subsp*. abscessus, M. abscessus* subsp*. bolletii, M. abscessus* subsp*. massiliense, M. chelonae, M. salmoniphilum, M. immunogenum*.

Table 5. PPanGGOLiN gene classification results for genomes of *M. franklinii, M. abscessus* subsp*. abscessus, M. abscessus* subsp*. bolletii, M. abscessus* subsp*. massiliense, M. chelonae, M. salmoniphilum, M. immunogenum*.

Gene class identified	Number of genes identified
Cloud genes	115,249
Shell genes	3157
Persistent	4231
Core genes	3251

There were an enormous number of cloud genes. PPanGGOLiN, compared to other pangenome programs, like Panaroo (Tonkin-Hill *et al*., 2020), does classify many more cloud genes. Otherwise, the number of core genes (genes present in > 99% of *Mycobacterial* genomes) at 3251 looks reasonable for genomes with about 5,000 genes (Segerman, 2012; Pear**c**e *et al*., 2020; Lyu *et al*., 2021)

There were 75 genomes retrieved from Genbank, identified by rpoB phylogeny, as members of the *M chelonae* clade, including *M. phlei* NCTC 8151 but not including *M. chelonae*^{T} NCTC 946. The other sequence data for the type strain such as ATCC 35752 are included, and correspond to the NCTC 8151 sequence. The data was read into R from the presence_absence.Rtab tab separated data file generated from the PPanGGOLiN analysis of 75 *M. chelonae* genomes. The Euclidean distance was calculated with dist() and plotted as an interactive heatmap with heatmaply() (Figure 43)

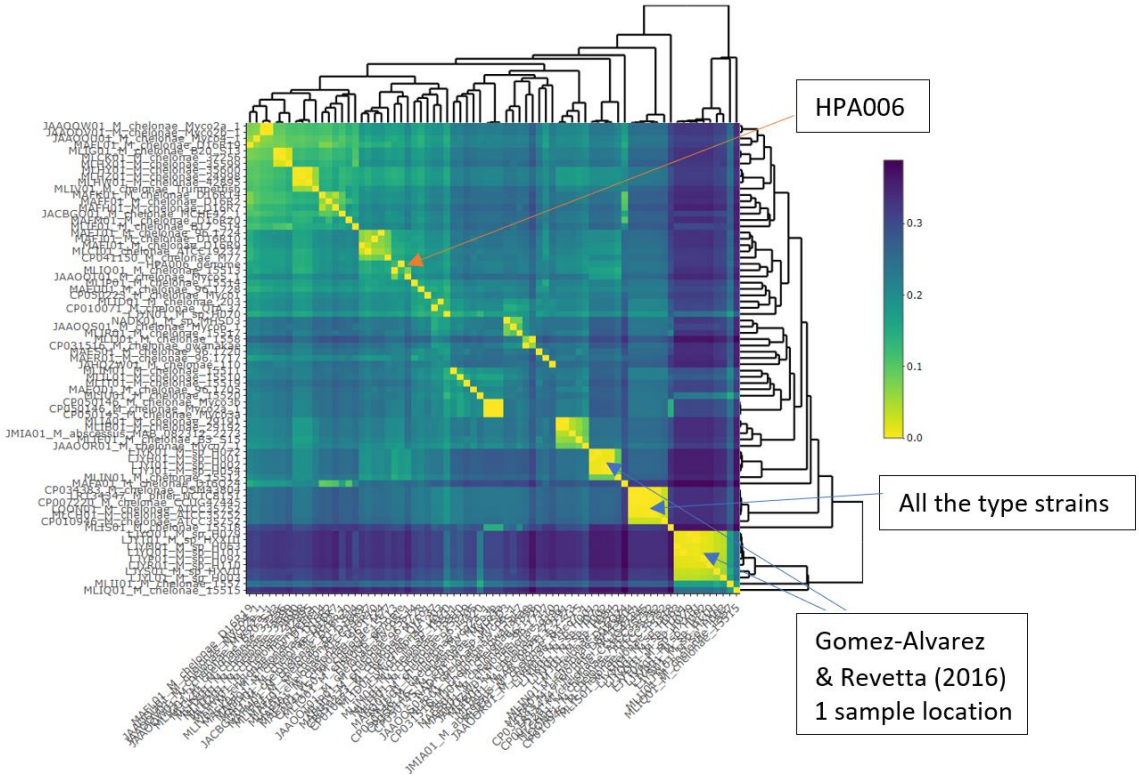


Figure 43. Heatmap *of M. chelonae* strains based on the presence/absence of homologous genes.

M. chelonae HPA 006 clusters with *M. chelonae* M77. The yellow blocks (labelled Gomez-Alvarez & Revetta, 2016) are numerous strains which were isolated at the same time from one source and so essentially amount to two strains represented by multiple identical genomes. There are multiple sequences for the type strain from different sequencing projects and culture collections, the NCTC strains are not included here. These results, however, confirm again that the type strain differs from most clinical isolates of *M. chelonae*.

The same data plotted as a 3D ordination using hclust_method using the option “average” (UPGMA, unweighted pair group with arithmetic mean method) in R software is shown in Figure 44.

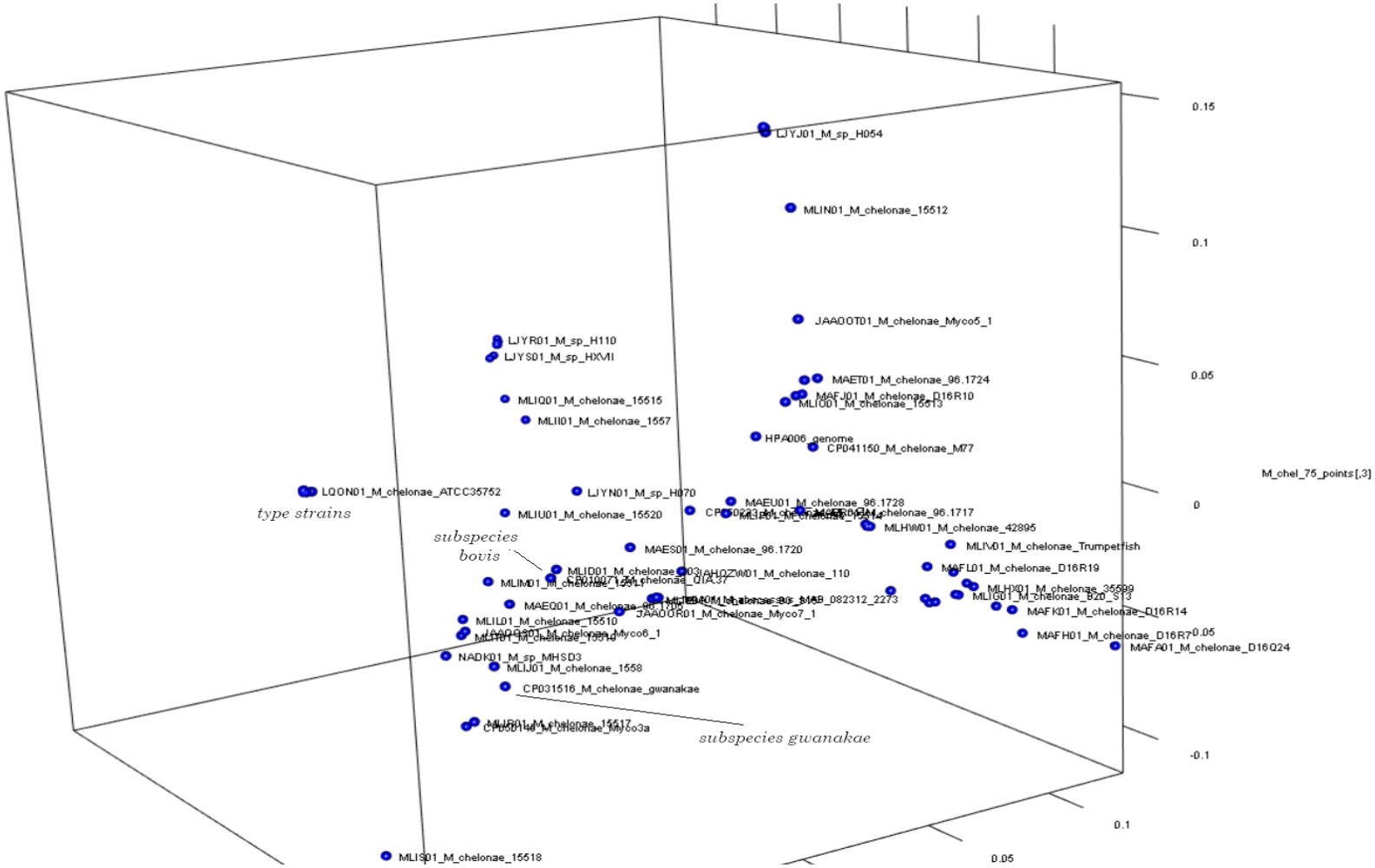


Figure 44. 3D plot of *M. chelonae* strains based upon the UPGMA distance calculated from the presence/absence of genes.

Despite the congruence seen in the ANI data and the heatmaps (see Figure. 43), Figure 44 suggests that the species shows a greater level of heterogeneity and that while all strains fall within a defined area of relationship compatible with a species definition there are some strains which are at the edge of the species envelope. The plot also suggests that *M. gwanakae* and *M. chelonae* subsp. *bovis* (on this evidence) are not sufficiently variant to be considered subspecies.

Figure 45 confirms the picture seen in Figure 44, demonstrating that *M. chelonae* is potentially still evolving.

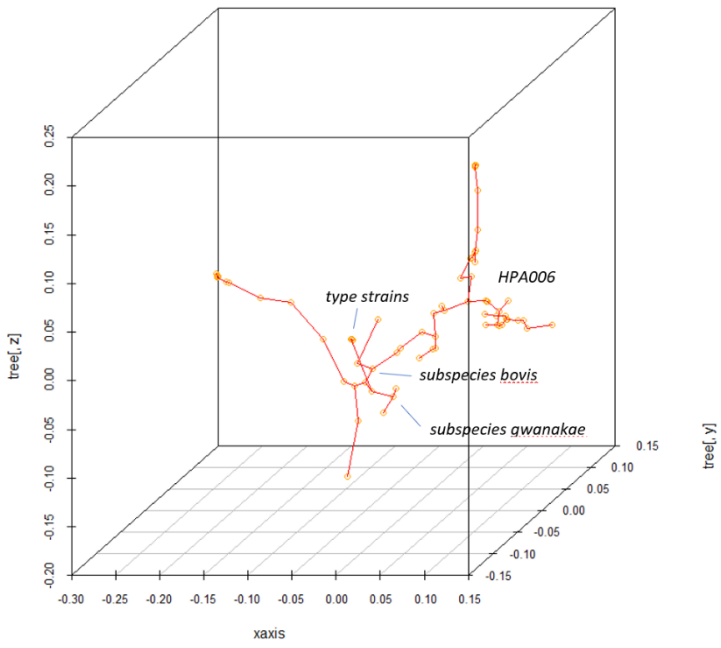


Figure 45. Minimum spanning tree of *M. chelonae* strains based upon UPGMA distance calculated from the presence/absence of genes.

Figure 45 confirms the picture seen in Figure 44, demonstrating that *M. chelonae* is potentially still evolving.

3.5 Analysis of Each of the Members of the *M. abscessus*/*chelonae* Clade

3.5.1 Mycobacterium chelonae clade

With the description of *Mycobacterium salmoniphilum* in 2007 (Whipps *et al*.,) the group of mycobacterial species referred to as the *M. abscessus*/*chelonae* clade were composed of *Mycobacterium chelonae* (Bergey *et al*., 1923), *Mycobacterium abscessus* (Kusunoki & Ezaki, 1992), *Mycobacterium immunogenum* (Wilson *et al*., 2001), *Mycobacterium massiliense* (Adékambi *et al*., 2004), and *Mycobacterium bolletii* (Adékambi *et al*., 2006b). To these have been added *Mycobacterium franklinii* (Nogueira *et al*., 2015a) and *Mycobacterium saopaulense* (Nogueira *et al*., 2015b). The taxonomic relationships of *Mycobacterium abscessus* variants (*M. abscessus, M. bolletii* and *M. massiliense*) were rationalised by Tortoli *et al*., (2016). To these species has been further added *Mycobacterium stephanolepidis* (Fukano *et al*., 2017).

3.5.2 Mycobacterium chelonae

This species was originally isolated from a sea turtle (Freidman, unpublished) and was suggested as a vaccine strain. However, the source of the isolate which was subsequently validly described was a tortoise (Bergey *et al*., 1923). Both of these organisms are members of the taxonomic order *Testudinae* but tortoises have the family name of *Chelonidae* giving rise to the epithet *chelonae*. *Mycobacterium chelonae* is an environmental species found in water and soil. Nevertheless, it may cause cervical adenitis, corneal infections, prosthetic valve endocarditis and wound infections. Phenotypically, the species is rapid-growing (3-4 days) producing non-chromogenic or buff-coloured colonies of rod-shaped organisms which are strongly acid-fast in young cultures. Growth temperatures vary from 22-40^{o}C, but the optimum lies between 33^{o} and 35^{o} (Magee & Ward, 2012). The Type strain is ATCC 35752 but many culture collections hold Type strain representatives. The rather ambiguous phenotypic features have led to several new species being noted as closely resembling *M. chelonae*.

Notably, the relationship of this species to *M. abscessus* (*sensu lato*) has in the past been a source of some confusion (Magee and Ward, 2012).

3.5.3 Mycobacterium abscessus subspecies analysis

Found in soil and originally isolated from the synovium of a knee and a cause of wound infections. A potential water-borne coloniser of vulnerable patients notably those with fibrocystic disease. The original species was validly named by Kusunoki and Ezaki (1992). However, Leao *et al*., (2011) based on analysis of rpoB and hsp65 sequence data, proposed *M. abscessus*, *M. massiliense* and *M. bolletii* to be a single species (*M. abscessus*) with two subspecies (*M. abscessus* subsp*. abscessus* and *M. abscessus* subsp. *bolletii*). Furthermore, those strains formerly representing *M. massiliense* were reclassified as *M. abscessus* subsp. *bolletii* (Leao *et al*., 2011). However, whole genomic sequencing of clinical isolates supported the distinctiveness of strains classified as "massiliense", thus suggesting that there are indeed three taxonomic groups within *M. abscessus* (Tettelin *et al*., 2014). This led to the proposal by Tortoli *et al*., (2016) to describe 3 subspecies of *Mycobacterium abscessus*, namely *M. abscessus* subsp. *abscessus*, *M. abscessus* subsp. *bolletii* and *M. abscessus* subsp. *massiliense* with the Type strains ATCC 19977T, CCUG 50184T and CCUG 48898T respectively. In arriving at these taxonomic conclusions Tortoli *et al.,* (2016) used ANI, genome to genome distance and single nucleotide polymorphism analysis as distinguishing characteristics. In a taxonomic oddity, Tortoli *et al*., (2016) erroneously described *M. abscessus subsp. massiliense* as a comb. nov., this was corrected to *M. abscessus subsp. massiliense* subsp. nov. when the valid publication was eventually officially recognised (Oren & Garrity, 2017).

The current study assessed 307 whole genome sequences of *M. abscessus* strains downloaded from the NCBI GenBank data base by ANI and 807 by PPanGGOLiN. Each of these strains had been assigned to a subspecies of *M. abscessus* by the depositing scientists. Once again, the software program Pyani was used to determine the Average Nucleotide Identities (ANI) for the collected strains.

The ANI data of the *M. abscessus* subsp *abscessus, M. abscessus* subsp *massiliense* and *M. abscessus* subsp *bolletii* strains studied show all 3 variants to be very closely related but very

distinct from all the other species and, in particular, from *M. chelonae*. However, if 95% ANI is accepted as an appropriate cut off for species and 95-98% as the range for the description of subspecies (Figure 39), these subspecies hold up well.

The impetus of this project was to consider whether there were genetic factors evident which would explain the apparent difference in virulence between strains of *M. abscessus* and *M. chelonae*. However, other members of the clade, when they were also implicated in disease were consigned to the vague classification "mycobacterium chelonae like organisms" (Wallace *et al*., 1993). Advances in identification techniques, including but not limited to WGS have identified that these species are distinct. A short analysis of the findings for each of these strains is detailed below.

3.5.4 Mycobacterium stephanolepidis analysis

The description of this species was based on 7 isolates from 5 thread-sail filefish (Fukano *et al*., 2017a). These isolates were subjected to sequence analyses of the 16S rRNA, *rpo*B, *hsp*65, recA and sodA genes. Although the 16S rRNA sequence showed 99.9% similarity to *M. chelonae*, a phylogenetic tree based on concatenation of these 5 genes separated *M. stephanolepidis* from *M. chelonae* and indicated the nearest related species to be *M. salmoniphilum*. The Type strain is JCM 31611.

However, only one of these isolates resulted in a deposited WGS. The genome description (Fukano *et al*., 2017b) records 93.56% ANI between *M. stephanolepidis* and *M. chelonae* CCUG 47445^{T}. The type strain of *M. chelonae* is not representative of the species and *M. stephanolepidis* clearly falls within the radius of the *M. chelonae* diversity

3.5.5 Mycobacterium immunogenum analysis

This species was validly named by Wilson *et al*., (2001) following an extensive assessment of unassigned strains putatively linked to the *M. abscessus/chelonae* clade. The Type strain being listed as ATCC 700505T.

This species has been implicated in hypersensitivity pneumonitis associated with metalworking fluid (Kreiss and Cox-Ganser 1997; Shelton *et al*., 1999). Also identified as a cause of keratitis following laser *in situ* keratomileusis (Sampaio *et al*., 2006).

Restriction enzyme analysis, 16S rRNA sequencing studies and DNA: DNA hybridisation (Wayne *et al*., (1987) suggested that this was a distinct species. An unexpected finding was that, despite the phenotypic similarity of *M. immunogenum* to *M*. *chelonae* and *M*. *abscessus*, the species possesses two copies of the rRNA operon, whereas *M*. *chelonae* and *M*. *abscessus* have only one.

In the current study all 17 of the available deposited whole genome sequences were included in a pangenomic assessment. The results of which supports the separation of *M. immunogenum* from the other species studied.

The outlier (*M. immunogenum*_CD11-6) is an unpublished strain, with only a partial sequence deposited. This sequence (NZ_LQYE01) lacks housekeeping genes which removes the possibility of further Blast analysis. Thus, the identity of this strain remains unclear, but it seems unlikely to be compatible with *M. immunogenum*. The remaining 16 strains show an ANI relationship of greater than 99.9%.

3.5.6 Mycobacterium salmoniphilum analysis

Isolated from viscera of salmonid fish – originally described by Ross (1960) the species was not cited in the *Approved List of Bacterial Names* (Skerman *et al*., 1980) but was revived and validly named by Whipps *et al*., (2007). The Type strain is CCUG 60883T. Whipps *et al*., (2007) used sequence analysis of the SSU (16S) rRNA gene, *hsp65*, *rpoB* and ITS regions to show that isolates were phylogenetically distinct. However, the authors also noted the diversity of isolates assigned to *M. salmoniphilum* with a resulting nervousness regarding deposited genetic sequences.

In this study all 7 whole genome sequences deposited in NCBI were assessed. *M. salmoniphilum* strains are shown to be separated from other strains of the *M. abscessus/chelonae* clade in the 3D plot. However, some ANI values between strains fall around 94-95%. The conclusion would be to agree with Whipps *et al*., (2007) that there is

diversity within this species. Despite this the species is distinct from other members of the *M. abscessus/chelonae* clade with the closest relative being *M. stephanolepidis*.

3.5.7 Mycobacterium franklinii analysis

Originally described on the basis of phenotypic data and analysis of concatenated sequences of the 16S rRNA, ITS, *hsp*65, *rpo*B housekeeping genes (Nogueira *et al*., 2015a). The Type strain is DSM 45524T. This was supported by a DNA:DNA relatedness study of the Type strain versus the Type strains of *M. abscessus* subsp. *abscessus*, *M. chelonae*, *M. immunogenum* and *M. salmoniphilum*.

In the current study 12 deposited whole genome sequences were downloaded and ANI data determined. The results support the conclusion that this organism is distinct from the other members of the *M. abscessus/chelonae* clade. However, the species may not be homogenous since there appear to be several clusters although all fall within the 96% similarity range.

3.5.8 Mycobacterium saopaulense analysis

Five isolates were studied by Nogueira *et al*., (2015b). These were originally obtained from 2 patients with corneal infections following LASIK surgery, 1 patient with a cervical abscess and 2 from laboratory kept zebrafish. Although initially thought to be variants of *Mycobacterium chelonae*, analysis of concatenated sequences of the 16S rRNA, *hsp*65 and *rpo*B genes showed *Myobacterium saopaulense* to be a distinct species. This was confirmed by DNA: DNA hybridisation against all of the other species listed as members of a *M. abscessus/chelonae* clade. The Type strain of *M. saopaulense* is CCUG 66554T.

For this study only 4 whole genome sequences of *M. saopaulense* were available. The ANI data demonstrates *M. saopaulense* to be separated from other members of the *M. abscessus/chelonae* clade.

Chapter 4. Resistance and Virulence Factors in the *Mycobacterium abscessus chelonae* Clade and Comparative Genomic Analysis of *M. abscessus* (ATCC 19977) and the Clinical Isolate *M. chelonae HPA 006*

4.1 Antimicrobial Resistance: Background

Many of the antibiotics we currently use are natural products, produced by microbes themselves. They are generated by one microbe as a way of competing with other microbes occupying the same ecological niche. Screening soil samples to look for these natural products led to the discovery of most of the antibiotics we currently use. Penicillins and cephalosporins are derived from fungi and antibiotics such as streptomycin from different strains of *Streptomyces*. A natural progression was the modification of these natural products to make semi-synthetic compounds which allowed the production of second and third generation antibiotics such as β lactams from the penicillins and cephalosporins as well as totally synthetic flouroquinolone compounds such as ciprofloxacin in the 1990's (Walsh, 2000).

4.1.1 Mechanisms of antibiotic resistance in bacteria

Antimicrobial resistance is a long-standing phenomenon that occurs as a result of an organism's interactions with its habitat. Because the majority of antimicrobial substances are naturally occurring chemicals, co-existing bacteria have evolved methods to counteract their effects in order to survive. As a result, these organisms are described as being "intrinsically" resistant to one or more antimicrobial agents (Munita and Cesar, 2016). Resistance can also be transferred between bacteria, so called acquired resistance. Genetic elements which contain resistance genes are picked up in a process of horizontal gene transfer (Lee, 2019). There is a third type of resistance, adaptive resistance. There is a wide range of opinion on the definition of adaptive resistance but Fernandez and Hancock, 2012 define it as "*a temporary increase in the ability of a bacterium to survive an antibiotic insult due to alterations in gene and or protein expression as a result of exposure to an environmental trigger*". Environmental situations which can initiate this include, ion density, pH, nutrient levels and exposure to a non lethal dose of antibiotics. One area that has sparked interest is that of biofilms. Yeasts or bacteria that form biofilms (cells growing on a surface and enclosed in an exopolysaccharide

matrix) are protected from antimicrobial agents (Szomolay *et al*., 2005), which raises the possibility that this is a form of adaptive resistance. Biofilms are comprised mostly of water and bacteria have no problem diffusing into them. Stewart, 2002, noted a reduction of antibiotic mobility by a factor of two to three in biofilms, however, this was not enough to explain the level of resistance of the aggregated bacteria to killing. It is apparent that mobility is not the only factor required for optimum penetration of the biofilm. If an antibiotic is inactivated or becomes bound as it moves within the biofilm, then its delivery deeper into the biofilm will be reduced (Stewart, 2002). A model biofilm formed by wild-type *Klebsiella pneumoniae* in a study carried out by Anderl *et al*., was shown to resist killing by ampicillin and ciprofloxacin (Anderl *et al*., 2000). Four potential explanations which explain biofilm resistance mechanisms have been put forward (Stewart, 2002). The first is that the antibiotic penetrates too slowly or incompletely, second, a concentration gradient of a metabolic substrate within the biofilm encourages areas of slow growing bacteria or completely inhibits growth. The third postulates an adaptive stress response and finally a small percentage of cells become persister cells, cells which become highly protected within the film. There is evidence that whilst these cells may constitute only a very small percentage of the biofilm, they enter into an almost spore like state which helps them evade killing by antibiotics (Lewis, 2001., Stewart and Costerton, 2001)

In summary resistance in bacteria arises either by mutation or acquisition followed by selection. Resistance is spread vertically, by the dissemination of resistant clones, e.g., the spread of methicillin resistant *Staphylococcus aureus* (MRSA) or horizontally by gene transfer. Examples of the latter are

- transduction mediated by bacteriophages
- conjugation, the direct transfer of genetic material directly by cell to cell contact via a pilus (the conjugative element).
- transformation, the direct uptake and incorporation of DNA by a competent cell from its surroundings via cell membranes.

4.1.2 Evading the effects of antibiotics

Antimicrobial agents can be classified according to the mechanisms which they employ to act on microbial cells (Reygaert, 2018), namely:

- inhibition of cell wall synthesis e.g., b -Lactam, carbapenem and cephalosporin classes of antibiotic
- depolarisation of the cell membrane e.g., lipopeptide class of antibiotic
- inhibition of protein synthesis e.g., aminoglycosides, tetracycline and macrolide classes of antibiotic
- inhibition of nucleic acid synthesis e.g., quinolone and fluoroquinolone classes of antibiotic
- inhibition of the microbial cell metabolic pathways e.g., sulfonamide and trimethoprim

Bacteria have several mechanisms which they employ to evade the mechanisms of antibiotic action.

- inactivate the antibiotic by enzymatic degradation e.g., β -lactamase which destroys the active component (the β-lactam ring) of penicillins. Extended spectrum β -lactamase producing bacteria (ESBL) are a major concern. These bacteria are able to hydrolyse extended spectrum cephalosporin, rendering ceftazidime, ceftriaxone, cefotaxime and oxyimino-monobactam ineffective (Bush and Jacoby, 2010)
- prevent the antibiotic from reaching its intended target, an example being the *mecA* gene, which encodes an alternative penicillin-binding protein, PBP 2a, causing high level resistance to methicillin in *Staphylococcus aureus* is such an example (Beceiro *et al*., 2013)
- reduce the susceptibility of the target by making it difficult for the antibiotic to gain entry into the cell, by modifying porin channels bacteria can restrict the influx of β-lactam and fluoroquinolone antibiotics
- increase the number of efflux pumps, effectively reducing the effectiveness of the antibiotic by removing it from the cell. Reygaert (2018) described five main families of efflux pumps in bacteria. They are classified according to their structure and the energy source used: the resistance-nodulation-cell division (RND) family: the ATP-binding cassette (ABC) family, the small multidrug resistance (SMR) family, the multidrug and toxic compound extrusion (MATE) family and the major facilitator superfamily (MFS)

Gram positive bacteria essentially have two resistance mechanisms. Enzymatic degradation of the antibiotic (for example by β-lactamases) or decreasing the affinity and susceptibility of the

penicillin-binding protein by either acquisition of exogenous DNA or by changes in the native PBP genes (Munita *et al*., 2015; Berger-Bächi, 2002), in effect two of the four mechanisms described above. Gram negative bacteria, however, may utilise all four in one capacity or another.

The United Kingdom's five-year national action plan, "Tackling antimicrobial resistance 2019–2024" was published in January 2019 (Tackling antimicrobial resistance 2019 to 2024 (publishing.service.gov.uk)). The plan's ultimate goal is to ensure progress toward the United Kingdom's 20-year vision on AMR, which would result in resistance being both contained and controlled and which focuses on three critical strategies to combat AMR: minimising the need for antimicrobials and unintended exposure to them; optimising antimicrobial usage; and investing in innovation, supply, and access.

4.1.3 Antibiotic resistance in the M. abscessus/chelonae clade

As discussed in Chapter 1 (section 1.5) resistance may be due to one or more of several mechanisms.

Clinically acquired drug resistance is usually the result of mutations in the target genes for the specific antibiotic being used as a therapeutic agent. Clinically acquired macrolide resistance in mycobacteria is conferred by mutation in the 23S rRNA gene. Meier *et al.,* (1996) examined isolates of *Mycobacterium avium* cultured from the blood of 38 patients before and after treatment with clarithromycin. They observed point mutations in the peptidyltransferase region of the 23S rRNA gene in 100% of the 74 resistant relapse blood isolates, however, no mutations were observed in the 69 susceptible isolates which were isolated prior to the start of treatment. Additionally, multiple mutations were observed in isolates from 23 of the 38 patients. In total 63 mutations were identified and 95% involved adenine at base pair 2058 (*Escherichia coli* numbering). Resistance due to this mutation develops infrequently during the treatment of infections due to *M. abscessus* or *M. chelonae* (Wallace *et al.,* 1996)

Intrinsic resistance may be derived from low permeability of the *M. abscessus/chelonae* cell envelope, as well as to numerous drug export systems and *M. abscessus* exhibits intrinsic resistance to numerous antibiotics of different classes.

There are also multiple enzymatic mechanisms which can confer resistance as Luthra *et al*., (2018) noted

- an erythromycin ribosome methyltransferase which acts to lower the binding affinity of macrolides. This happens when the organism is exposed to macrolide antibiotics, classic inducible resistance

- Acyl and phosphate groups can be added to sites on the aminoglycoside molecule by enzymes in *M. abscessus* preventing the antibiotic binding to its ribosomal target

- There are two putative acetyltransferases, *AAC*(2') and *eis2*, which can be active not only on different aminoglycosides but also within different regions of the aminoglycoside. It is noteworthy that two genes present in the *M. abscessus* genome, MAB_4124 (*eis1*) and MAB_4532c (*eis2*), have been demonstrated to show homology to *eis* from *M. tuberculosis* (Rv2416c) and *A. variabilis* (Ava_4977), respectively. *eis* is an effector which when released into a host cell, negatively affects the host immune response. Antibiotic testing has demonstrated a role for *eis2* in intrinsic aminoglycoside resistance in *M. abscessus*. It was also noted that as capreomycin also targets the bacterial ribosome, it too, is vulnerable to *eis2*

- Environmental bacteria have been shown to inactivate rifampicin by glycosylation, and phosphorylation of the molecule and to decompose it using monooxygenation. Given that members of the *M. abscessus*/*chelonae* clade occupy an environmental niche in which they will be exposed to actinomycetes (noted for their rifamycin inactivation mechanisms) it is feasible that members of this clade also have rifamycin resistance genes, such as a rifamycin glycosyltransferase

Whilst the use of macrolides (azithromycin and clarithromycin) has had some success they are susceptible, as described above, to inducible resistance caused by the erythromycin ribosomal methylase gene *erm(*41) (Nash *et* al, 2009). Moreover, mutations in the *rrl* gene encoding the 23S rRNA peptidyl transferase are also associated with acquired macrolide resistance (Choi *et al*., 2017). Of particular interest, is the determination by Koh *et al*., (2014) that the *erm*(41)

gene is functional in *M. abscessus* subsp*. abscessus* strains, but *M. abscessus* subsp. *massiliense* has a truncated *erm*(41) which renders it intrinsically susceptible.

M. abscessus subsp. *bolletii* has the T28 polymorphism and may develop resistance to macrolides during therapy (Nessar *et al*., 2012; Kim *et al*., 2016; Pavan *et al*., 2017). This emphasises the value of accurate taxonomic speciation of *M. abscessus.*

In *M. abscessus* the mycobacterial transcriptional regulator *whi*B7 is induced by exposure to the ribosome-targeting antibiotics erythromycin, clarithromycin, amikacin, tetracycline, and spectinomycin. However, deletion of the *whi*B7 reverses this effect and allows susceptibility to each of these agents (Hurst-Hess *et al*., 2017). Hurst *et al*., (2017) demonstrated that *whi*B7 specifically induces the gene *eis2* which is a factor in higher levels of intrinsic resistance to amikacin in *M. abscessus*.

Strains assigned to the *M. abscessus/chelonae* clade are highly resistant to most β lactam antibiotics due to the production of broad spectrum β lactamases. However, Kumar *et al*., (2017b) stated a belief that combination therapy using a β lactam antibiotic and a β lactamase inhibitor has potential for the treatment of *M. abscessus*. This hypothesis was further explored by Story-Roller *et al.,* (2018) and Zhanel *et al*., (2018) and has led to clinical trials of such combination treatments. In the same vein intrinsic resistance to the rifamycin group of antibiotics may be overcome by modification of the rifamycin core (Combrink *et al*., 2007) an approach utilised by Mosaei *et al.,* (2018) in the newly described antibiotic kanglemycin.
These data emphasise the importance of new drug developments and drug targets in the treatment of infections caused by organisms in this clade.

In a recent paper Victoria *et al*., (2021) reviewed developments of respiratory disease with *M. abscessus* as the etiological agent. Taxonomic developments, incidence, resistance profile and virulence determinants are amongst the areas evaluated, reflecting the increasing importance of *M. abscessus* as a respiratory pathogen.
Table 6 reproduced from Victoria *et al.,* (2021) details the antibiotic family, resistance mechanism, enzyme/gene implicated and its location.

Table 6. Mechanisms responsible for antimicrobial resistance in *Mycobacterium abscessus* Reproduced from Victoria *et al*., (2021).

ANTIMICROBIAL	MECHANISMS OF RESISTANCE	ENZYME/GENE	LOCATION
Aminoglycosides	Target modifying enzymes	Aminoglycoside2- N –acetyltransferase and aminoglycosies phosphotransferases[a]	
Deoxystreptamine Aminoglycosides	Acquired resistance by point gene mutations	*rrs* gene encoding 16S rRNA protein	Mutations include T1406A, Cl409T, A1408G, and G1491T.[b]
Beta Lactams	Antibiotic degrading enzymes	β-lactamase encoding genes.	Class A β-lactamase Bla_Mab (MAB_2875)
Macrolides	Target modifying enzymes Acquired resistance by point gene mutations	Functioning erythromycin ribosome methylase *erm(41)* gene rrl gene encoding 23S rRNA transferase	Reversion to susceptibility: 274 bp deletion at positions 159- 432 and T28C point mutation.[C] Point mutations at positions 2058 and 2059
Fluoroquinolones	Polymorphism in target genes	Nucleotide variation at the Quinolone resistance determining region in the DNA gyrase- GyrA – GyrB[d]	Ala-90 gyrA gene Arg-516 and Asp-533 gyrB gene
Tetracyclines	Enzymatic inactivation	Flavin- adenine- dinucleotide (FAD)- inactivating monooxygenase (MabTetX)	

[a]*Twelve putative aminoglycoside phosphotransferases are encoded within the MABC genome, which could contribute to resistance to this group of antibiotics (Nessar et al., 2012; Luthra et al., 2018).*
[b]*Mutations associated with aminoglycoside resistance (Ananta et al., 2018).*
[c]*These mechanisms are associated with reversion to clarithromycin susceptibility (Nie et al., 2014; Zhu et al., 2015). A T28C point mutation (thymidine to cytosine polymorphism at the position 28) results in tryptophan to arginine amino acid change at codon 10, rendering a non-functional erm 41 gene (Pavan et al., 2017), C28 polymorphism is related to susceptibility (Luthra et al., 2018).*
[d]*Quinolone resistance is associated with gyrA and gyrB mutations, a previous study showed all resistant isolates encoded the same amino acids in the quinolone resistance determining region (Kim et al., 2016).*

4.2 Bacterial Virulence Factors

A microbe's ability to evade host defence mechanisms, grow and persist are key to its pathogenicity. Overcoming these mechanisms is achieved by various methods, which can be split into four broad categories:

- factors that allow the organism to attach to cells in the host e.g., adhesins. Different species have developed different solutions, but the typical mechanism is a fimbria or pilus. These are important virulence factors in circumstances where the organism has to attach to a mucosal layer such as the respiratory or urinary tract (Zanin *et al*., 2016). Saprophytic bacteria may also possess them; however many pathogenic bacteria have an array of different adhesins which they can call upon at various stages during the infective process, a key factor for virulence (Klemm & Schembri, 2000)
- mechanisms to escape phagocytosis such as capsules. This polysaccharide layer lies outside the cell envelope is not easily destroyed by the host cell making it an important virulence factor. It is present in both gram positive and gram negative organisms. The polysaccharide capsule of *Streptococcus pneumoniae* is the pathogen's main virulence factor allowing it to evade phagocytosis (Moscoso & García, 2009). *Haemophilus influenza, Klebsiella pneumonia* are also examples of capsulated pathogens

- enzymatic factors are also employed by pathogenic bacteria. These cause damage to host tissues and include hyaluronidase (which breaks down hyaluronic acid present in connective tissue), lipases, haemolysins (which break down, amongst other cells, red blood cells) and DNases (active against DNA). *Staphylococcus aureus*, for example, produces hyaluronidase, lipase and coagulase. The latter, which converts fibrinogen to fibrin, causing clots is also used to differentiate between types of staphylococci in the clinical/laboratory setting. In general coagulase negative staphylococci are considered to be opportunistic pathogens, though it is worthwhile bearing in mind that not all *Staphylococcus aureus* isolates produce coagulase (Vandenesch *et al*., 1993)
- production of toxins, categorised as exotoxins and endotoxins.

  Endotoxin is a complex lipopolysaccharide and a major component of the outer membrane of the gram negative bacterial cell wall. It is the lipid A component of the lipopolysaccharide which is toxic. Endotoxins are not released outside of the bacterial cell and act by inducing fever and in high enough concentrations they can cause septic shock

  Exotoxins are produced by both gram positive and gram negative bacteria. In contrast to endotoxins they are actively secreted by the organisms. There are three types:

  - superantigens (Type I toxins) which stimulate the production of large amounts of T cells, which in turn increase the production of cytokines. *Streptococcus pyogenes* and *Staphylococcus aureus* produce superantigen toxins in the form of streptococcal pyogenic exotoxin and toxic shock syndrome toxin respectively

  - membrane disrupting toxins (Type II toxins) as the name implies damage host cells by disrupting the structural integrity of the plasma membrane. This is achieved by one of two methods. Either by formation of protein channels into the plasma membrane (*S. aureus*) or by disruption of the phospholipid layer (*C. perfringens*)

  - A-B toxin, a polypeptide comprising two components. A (active portion) alters the function of host cells by inhibiting protein synthesis, and B, the binding component, facilitates the attachment of the toxin to the receptors of the host cells. Perhaps the

two best known examples are tetanus toxin produced by *Clostridium tetani and botulinum toxin which is secreted by Clostridium botulinum*

4.3 Mycobacterial Virulence Factors

The study of the virulence factors associated with the genus *Mycobacterium* has focussed on tuberculosis and its etiological agent *M. tuberculosis*.

M. tuberculosis is an obligate pathogen, transmitted via the inhalation of aerosolised droplets containing the bacterium. It most often infects the lungs causing pulmonary disease but it can also cause extra pulmonary disease when dissemination to the central nervous system or lymph nodes takes place (Yang *et al*., 2004). Infection in the lungs is primarily of the type 2 pneumocytes, polymorphonuclear neutrophils and alveolar macrophages (González-Cano *et al*., 2010). Following an immune response by the host, multicellular granulomas are formed with the purpose of limiting the spread of the organism, thus, infection with *M. tuberculosis* does not necessarily result in immediate progression to disease, as this interaction between the host immune system and pathogen can result in a period of latency (Getahun *et al*., 2015). Less than 10% of infected individuals will progress to developing the disease (Silva Miranda *et al*., 2012). This containment by the immune system, however, can be disrupted by a weakened immune system or an immunocompromised host, leading to systemic dissemination of the organism and active disease (Madacki *et al*., 2019).

Mycobacteria have become adept at evading the host immune system and secreting several virulence factors, all of which enable it to survive and persist inside the host.
M. tuberculosis does not produce a single dominant factor which supports virulence, but instead has a blend of virulence factors and host responses.

4.3.1 Surviving phagocytosis

Bacterial pathogens, once inside macrophages are eliminated by phago-lysosomal fusion, with the macrophage removing waste material through exocytosis. The fact that *M. tuberculosis* can remain dormant within these cells for extended periods indicates that it has developed

the means to overcome this host defence. As a result, the pathogenesis of *M. tuberculosis* is more complex, with no single virulence factor predominating.

Sassetti *et al.*, (2003) sought to determine the molecular mechanisms which *M. tuberculosis* employed to infect and persist inside the human host. Using a murine model of infection and by mutating every non-essential gene of *M. tuberculosis*, the group were able to identify 194 genes which were specifically needed for *in vivo* mycobacterial growth. It was also determined that a large proportion of these genes were unique to mycobacteria, indicating that the approach taken by this species is novel. A second study carried out by Zhang *et al.*, (2013) which employed deep sequencing found an additional 400 genes which were essential for *in vivo* survival (Madacki *et al*., 2019)

A similar study carried out by Rengarajan *et al.,* (2005) but looking at macrophages, specifically the *M. tuberculosis* genes which were required to survive inside macrophages by screening for transposon mutants which failed to grow within primary macrophages. In total they identified 126 genes needed to survive inside macrophages. Additionally, they found that the majority of genes required for survival are expressed rather than regulated by macrophages.

4.3.2 Type VII secretion systems

Type VII secretion systems (ESX systems) play a key role in *M. tuberculosis* and other non tuberculous mycobacteria surviving in the host by enabling the transport of selected protein substrates across the cell envelope (Rivera-Calzada *et al*., 2021). Several gene clusters that encode proteins have been identified in mycobacteria, confirming that they play an important role in pathogenesis. Furthermore, the secretion systems in mycobacteria appear to be unique, a fact explained by their cell wall structure which is hydrophobic and impermeable being rich in mycolic acids. Abdallah *et al*., (2007) proposed that this secretion system should be called a Type VII secretion system. *M. tuberculosis* has 5 Type VII secretion systems: ESX-1, ESX-2, ESX-3, ESX-4 and ESX-5. All five show a similarity in gene content but each has a different function, this information is summarised in Table 7 below

Table 7. Type VII secretion systems (ESX systems) present in *M. tuberculosis* and other Non-Tuberculous Mycobacteria.

Type VII Secretion	**Organism**	**Substrate produced**	**Function**
ESX-1	*M. tuberculosis*	ESAT-6, CFP-10, EspB, EspC, EspD, EspE, EspF, EspG, EspH, EspI, EspJ, EspK, EspL, PE-PGRS family proteins	**Virulence factor.** ESAT-6 and CFP-10 are potent T cell antigens and both play a crucial role in the virulence of *M. tuberculosis*
ESX-2	*M. tuberculosis*	PPE proteins	Permeabilisation of the phagosomal membrane
ESX-3	*M. tuberculosis*	EsxG, EsxH, EsxR, EsxS, EsxT	Required for siderophore mediated iron acquisition and for zinc uptake.
ESX-4	*M. tuberculosis*	PE proteins	Required for export of CpnT and surface accessibility of tuberculosis necrotizing toxin (TNT)
(ESX-5)	*M. tuberculosis* and other *Actinobacteria*	EsxW, EsxX, EsxY and EsxZ proteins; PE and PPE proteins; LpqH; LipY;	Cell wall synthesis and maintenance
(ESX-5)	*M. xenopi*	EccB5, EccC5, EccD5 and EccE5	Transport of substrates across the outer membrane of the mycobacterial cell

4.4 *Mycobacterium abscessus* Virulence Factors

The review by Victoria *et al*., (2021) has comprehensively detailed the recent developments which have taken place with respect to *M. abscessus* respiratory infections. These authors described a number of key points related to virulence. These are reproduced below and collated in Table 8. Whilst these have not been studied as part of this project, in describing the broad range of resistance and virulence mechanisms employed by this organism, they offer an insight into the clinical challenges faced in the treatment of infections caused by *M. abscessus*.

Table 8. Review of immune response and virulence factors for *Mycobacterium abscessus* antimicrobial resistance. Taken from Victoria *et al*., (2021)

Immune Response	Outcome
(i) Type I interferons (IFN-I) such us IFNβ and IFNα play a critical role during MABC infection (Ruangkiattikul *et al*., 2019)	
(ii) Activation of toll-like receptors (TLR), specifically TLR2 and TLR4 results in MyD88/TRIF/IRF3 dependent IFN-I induction (Ruangkiattikul *et al*., 2019 Peignier and Parker, 2021)	Data suggest that persistence of *M. abscessus* infections in CF patients could be explained by a limited IFN-I response. Interferon gamma (IFN-γ), a type II IFN is also required to control *M. abscessus* infection, there have been several reports of disseminated disease in patients with defects in the IFN-γ pathway (Rottman *et al*., 2007)
(iii) IFN-I production in infected macrophages activates inducible nitric oxide synthase (NOS2) and nitric oxide (NO) production which can kill or induce dormancy. IFN-I plays a key role to induce NO production and intensify the ability of macrophages to clear MABC	

infection (Ruangkiattikul *et al*., 2019)	
(i)*M. abscessus* induces a strong TLR2 mediated TNFa response (Tumour Necrosis Factor), a key inflammatory cytokine that mediates mycobacterial killing, host defence and granuloma formation (Bernut *et al*., 2016; Kim *et al*., 2017). (ii) TNF/IL8 signalling pathway activates macrophage bactericidal activity, restrict extracellular growth, and increase neutrophil recruitment and mobilization which is required for granuloma formation (Bernut *et al*., 2016). (iii) highly virulent isolates stimulate TNF secretion by macrophages (Medjahed *et al*., 2019)	Impairment of this TNF/IL8 signalling pathway correlates with disseminated disease and lethal infection (Bernut *et al*., 2016; Bernut *et al*., 2019). However, excessive levels of TNF-α can lead to detrimental effects in the host secondary to tissue damage (Kim *et al*., 2017).
Virulence Factors	
(i)M. abscessus from clinical isolates have shown a **rough vs smooth colony morphology** and is able to shift between these forms (Medjahed *et al*., 2019)	A zebrafish experimental model has been recently used for the study of *M. abscessus* virulence, using this model, researchers observed that the rough morphology forms serpentine cords and large bacterial clumps, the formation of large cords allows *M. abscessus* to escape the immune system as the extremely large size of cords might prevent *M. abscessus* from being internalized as they are larger in size than macrophages (Bernut *et al*., 2014; Bernut *et al*., 2017)

	This also promotes spread to other tissues and extracellular replication that results in abscess formation and tissue damage (Bernut *et al*., 2014). The rough morphology is also associated with increased apoptosis leading to increases in extracellular bacteria and promotion of cord formation (Bernut *et al*., 2014) Rough morphology induces higher levels of TNF-I (Ruangkiattikul *et al*., 2019) Smooth morphology induces lower levels of IFN-I which favours persistence (Ruangkiattikul *et al*., 2019).
CFTR (CF transmembrane conductance regulator) defects, such as those seen in CF, has been associated with impaired NADPH oxidase production.	This leads to increase intracellular growth and reduced neutrophil chemotaxis, compromising granuloma integrity (Bernut *et al*., 2019)
(i)The absence of glycopeptidolipid (GLP) has been associated with rough colony morphology. (ii) Defects in the mmpL4b gene are associated with the loss of GLP, leading to conversion to a rough phenotype showing morphological plasticity (Nessar *et al*., 2011; Bernut *et al*., 2016)	The GLP loss unmask lipoproteins that produce a strong inflammatory response (Nessar *et al*., 2011).
Intra-macrophage survival of the smooth morphology of *M. abscessus*	Due to phago-lysosomal fusion block and resistance to apoptosis, the phagosome shows membrane disruption at early stages of infection leading to phagosomecytosol communication and phagosomal escape allowing extracellular replication (Bernut *et al*., 2017)

	Phagosomal escape is independent of ESX-1 mechanism as *M. abscessus* only has two ESX gene clusters (ESX-3 and ESX-4) which differs from *M. tuberculosis* (Kim *et al*., 2017). Cord formation is a unique and new immune evasion mechanism in *M. abscessus* infection
The three subspecies of *M. abscessus* are separated based on multi-locus sequencing of housekeeping genes (Harris and Kenna, 2014) *M. abscessus* has a single ribosomal RNA operon	Unlike other rapidly growing mycobacteria, *M. abscessus* has a single ribosomal RNA operon, making the phenotypic expression of single mutation more likely.
The reference strain of MABC (ATCC19977) includes an 81- Kb full-length prophage, five insertion elements, and a 23 Kb- mercury resistance plasmid, (Cortes *et al*., 2010; Medjahed *et al.*, 2019).	Associated with infection in young patients with CF The resistant plasmid is highly similar to an episome present in *M. marinum*, suggesting that these species may have exchanged the plasmid (Medjahed *et al*., 2019).
M. abscessus contains unique genes not present in other mycobacteria that appear to have been acquired by horizontal gene transfer from different species such as pseudomonas species and streptomyces species (Howard, 2013; Medjahed *et al*., 2019)	These shared genes are thought to contribute to the pathogenesis of Pseudomonas sp. and MABC-PD (*M. abscessus* pulmonary disease) facilitating respiratory tract colonization in CF patients (Nessar *et al*., 2011).

4.5 Secondary metabolite biosynthetic gene clusters

Microbial secondary metabolites are low molecular weight products of secondary metabolism. They are typically produced during the stationary and/or late growth phase and have varied chemical structures and biological functions. They are not essential for growth of the organism (which led to the designation secondary metabolites) but do play a key role in the interaction of the microbe with its environment (Sanchez & Demain, 2011).

Gokulan *et al.,* (2014) have covered the principal synthetic pathways of secondary metabolite production in bacteria e.g., β-lactam, oligosaccharide, shikimate (Shikimic acid), polyketide and non-ribosomal pathways in their review in the Encyclopedia of Food Microbiology. This review illustrates that these metabolites have various biological functions related to antimicrobial substances, toxins, anti-cancer agents, pesticides and more. Sharrar *et al., (*2020) also examined understudied phylogenetic groups with great biosynthetic potential, reviewing the most common Biosynthetic Gene Cluster types and their possible functions.

antiSMASH (**anti**biotics and **S**econdary **M**etabolites **A**nalysis **Sh**ell) (Medema *et al*., 2011; Blin *et al.,* 2021) is a fully automated pipeline which searches the genomes of bacteria and fungi for secondary metabolite biosynthetic gene clusters (BGCs). antiSMASH was initially developed in a collaborative project between Tübingen University (Tilmann Weber, Kai Blin), Groningen University (Eriko Takano, Rainer Breitling, Marnix Medema) and UCSF (Michael Fischbach). Currently, antiSMASH development is coordinated at Wageningen University and the Novo Nordisk Foundation Center for Bio-sustainability/ Technical University of Denmark.

Glycopeptidolipids (Ripoll *et al*., 2007) play a major role in creating the mycobacterial cell surface which is exposed to the host cell. They are often strong antigens but also modulate the immune system response. Many mycobacterial BGCs produce glycopeptidolipids.

The sequence reads generated for the genomes of *M. abscessus* (ATCC 19977) and the clinical isolate *M. chelonae HPA 006* from the PPanGGOLiN gene classification study, were aligned against the AnnoTree protein database (Mendler *et al*., 2019) using DIAMOND (Buchfink *et al*., 2015) eggNOG. (evolutionary genealogy of genes: Non supervised Orthologous Groups) Powell

et al., 2012), InterPro (Mitchell *et al.,* 2015) SEED (Overbeek *et al*., 2013). and KEGG (Kanehisa & Goto, 2000) classification systems.

4.6 Materials and Methods PPanGGOLiN

4.6.1. PPanGGOLiN

(see Chapter 3 Section 3.3.3)

4.6.2 Installation of Diamond + Megan

A binary executable for Windows was downloaded, to a folder Diamond, from the Github releases page at https://github.com/bbuchfink/diamond and the Visual C++ redistributable package was installed from https://www.microsoft.com/en-us/download/details.aspx?id=48145 the official Microsoft Download Center. The nr protein database from ncbi was downloaded from ftp://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/nr.gz together with prot.accession2taxid.gz and nodes.dmp from ftp://ftp.ncbi.nlm.nih.gov/pub/taxonomy/accession2taxid/prot.accession2taxid.gz and ftp://ftp.ncbi.nlm.nih.gov/pub/taxonomy/taxdmp.zip.

A Diamond index was created in Powershell:
PS C:\programs\Diamond>./diamond makedb --in nr.gz -d nr --taxonmap prot.accession2taxid.gz --taxonnodes nodes.dmp

Megan6 was downloaded to a folder Megan from the Megan6 download page at Tuebingen from the link MEGAN_Community_windows-x64_6_23_2.exe and the mapping file megan-map-Feb2022.db.zip which maps NCBI-nr accessions to taxonomic and functional classes (NCBI, GTDB, EC, eggNOG, InterPro2GO, SEED). Instructions for analysis are detailed in Bagci *et al*., (2021).

The Annotree database (Mendler *et al*., 2019) provides an alternative protein database for gene assignment by Diamond and Megan (Gautam *et al*., 2022). The Diamond index is created with annotree.FASTA.gz downloaded from the Megan-Annotree download page and Megan is run with the mapping file megan-mapping-annotree-June-2021.db.zip.

4.6.3 AntiSMASH The antibiotics and secondary metabolites analysis SHell

AntiSMASH analysis (Medema *et al*., 2011; Blin *et al.,* 2021) is available as a web service at https://antismash.secondarymetabolites.org/#!/start and data is returned through a browser interface and as annotated sequence files and links to databases.
The sequence data for both strains were submitted to the database, to identify the potential function of biosynthetic gene clusters.

Previously, the discovery of natural compounds produced by microorganisms which could have potential as new antimicrobial agents, was a protracted process of extraction, chemical isolation and purification from the natural source (Blin *et al.,*2021). The accessibility of whole genome sequencing means that this process can now be enhanced by the subsequent mining of genome and metagenome data to determine biosynthetic pathways for these potential antimicrobial products (Ziemert *et al*., 2016). Software tools to support these process have included packages such as PRISM (Skinnider *et al*., 2020) and TOUCAN (Almeida *et al*., 2020). antiSMASH (Medema *et al*., 2011; Blin *et al.,* 2021)

The secondary metabolism of bacteria and fungi constitutes a rich source of bioactive compounds which play a key role in the interaction of the microbe with its environment (Sanchez & Demain, 2011). The genes encoding the biosynthetic pathway responsible for the production of any given secondary metabolite are often spatially clustered together at a certain position on the chromosome (Skinnider *et al*., 2020); this collection of genes is referred to as a secondary metabolite biosynthesis gene cluster (BGC). Therefore, locating their gene clusters simplifies the detection of secondary metabolite biosynthesis pathways.

Based on profile hidden Markov models (HMM) of genes that are specific for certain types of gene clusters, antiSMASH is able to accurately identify the gene clusters encoding secondary metabolites of all known broad chemical classes. antiSMASH not only detects the gene clusters, but also offers detailed sequence analysis .

The workflow of an antiSMASH analysis is illustrated in Figure 46. Cluster detection is achieved using CASSIS (Cluster Assignment by Islands of Sites), generic analyses is carried out using Minimum Information about a Biosynthetic Gene cluster, (MiBiG) a specification which

provides a robust community standard for annotations and metadata on biosynthetic gene clusters and their molecular products. Specific analyses using modules which annotate domains with polyketide synthase (PKS) and non-ribosomal peptide synthase (NRPS) related functions are also carried out. Output data is returned through a browser interface and as annotated sequence files and links to databases.

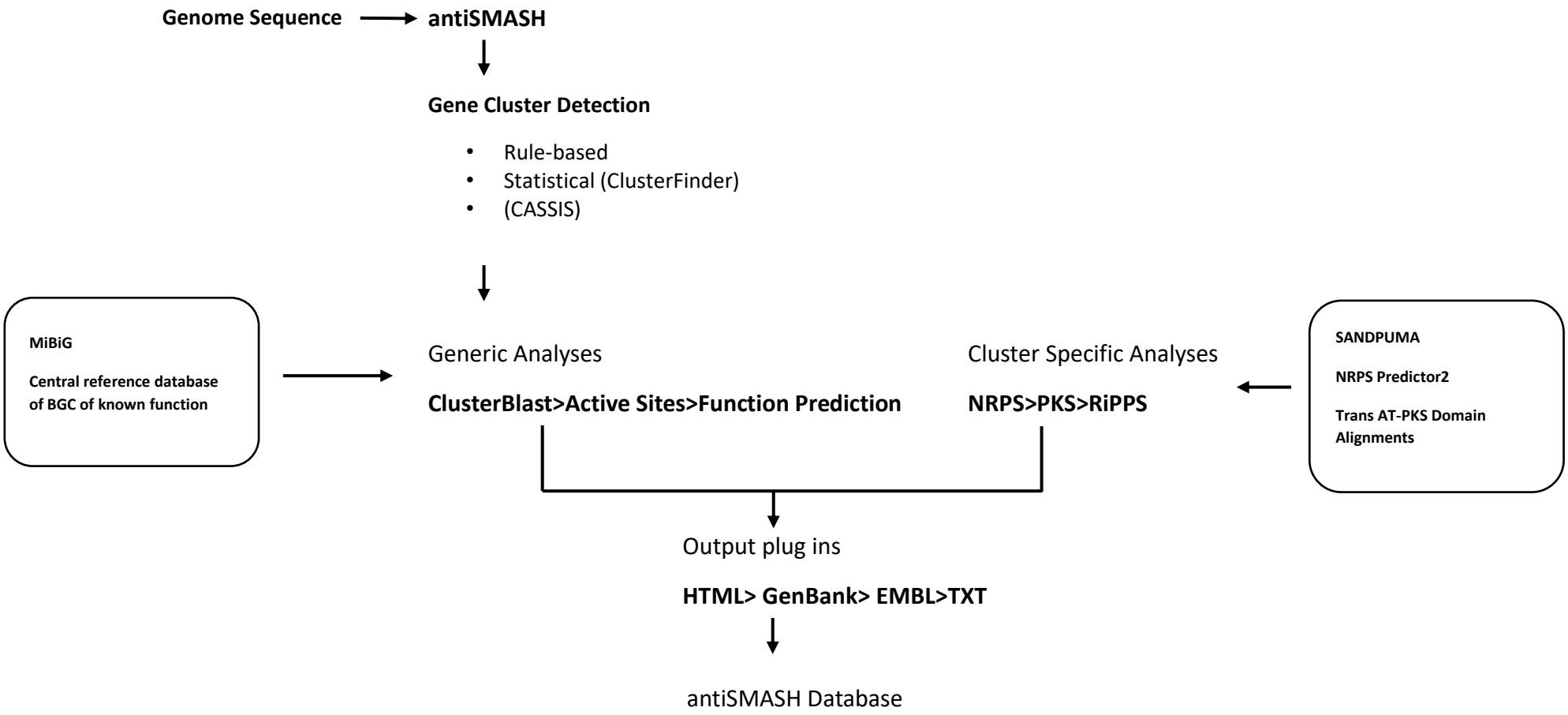


Figure 46. Workflow utilised by antiSMASH for the analysis of bacterial and fungal genomes. Illustrating the cluster detection using CASSIS, generic analyses using MiBiG and specific analyses using modules which annotate domains with polykeytide synthase (PKS) and non- ribosomal peptide synthase (NRPS) related functions. Output data is returned through a browser interface and as annotated sequence files and links to databases.

4.7 Results

4.7.1 Comparison of the genomes of M. abscessus (ATCC 19977) and the clinical isolate M. chelonae HPA 006

Genome size is reflective of lifestyle, with genome size reducing as an organism specialises in a pathogenic lifestyle compared to one in which they are managing the more diverse challenges required to survive in the environment. e.g., compare *Mycobacterium ulcerans* with *Mycobacterium marinum* (Stinear *et al*., 2007; Tan *et. al.,* 2020) Further studies by Murray *et al*., in 2020 showed that a reduced genome size correlated with a pathogenic lifestyle in bacteria. The Tuberculosis Database (TBDB) is an integrated database providing access to mycobacterial genomic data and resources, particularly where it pertains to tuberculosis. Currently it has genome sequence data and annotations for 28 different *M. tuberculosis* strains and related bacteria which includes information on genome size (Reddy *et al.,* 2009; TB Genomes Database (bu.edu)

Table 9 shows that the genome size of *Mycobacterium abscessus* is intermediate between that of the obligate pathogen *Mycobacterium tuberculosis*, and the environmental species *Mycobacterium smegmatis*, but there is no appreciable difference in the genome sizes of *Mycobacterium abscessus* and *Mycobacterium chelonae*.

Table 9. Genome size of the *M. chelonae* HPA 006 study strain and three mycobacterial species representative of an obligate pathogen *M. tuberculosis*, an opportunistic pathogen *M. abscessus*^{T} and an environmental strain *M. smegmatis*.

Mycobacterial species	**Genome size in base pairs**
M. tuberculosis H37R	4,411,709
M. abscessus ^{T}	5,067,192
M. smegmatis ATCC 19420	6,983,767
M. chelonae (HPA 006) study strain	5,116,005

4.7.2 Comparison of the functional classification of genes in M. abscessus (ATCC 19977) and the clinical isolate M. chelonae HPA 006

Many genes code for the synthesis of enzymes whose actions are the basis for cell function. There are 5 gene classifications available to analyse in Diamond + Megan (Bagci *et. al.* 2021)

- Interpro2GO (https://www.ebi.ac.uk/GOA/InterPro2GO Camon *et. al.* 2005; Gene Ontology Consortium. 2021)
- eggNOG (http://eggnog5.embl.de/#/app/home Huerta-Cepas *et. al.* 2019)
- SEED (https://www.theseed.org/wiki/Home_of_the_SEED Overbeek *et al.*, 2005)
- KEGG (https://www.genome.jp/kegg/ Kanehisa. 2000)
- EC Enzyme commission numbers. (https://web.archive.org/web/20180910045839/http://www.sbcs.qmul.ac.uk/iubmb/enzyme/) accessed 25/05/2022)

The genomes of *M. chelonae* HPA 006 and *M. abscessus*^{T} (CU458896) were exported from Geneious as FASTA files and submitted to Diamond for assignment of genes in either the nr database or the AnnoTree database using the long read option. The assigned genes were saved in a .daa format file. Megan 6 was used to classify the assigned genes by taxonomy and functional classification using Interpro2GO, EggNOG, SEED, KEGG and EC. The KEGG classification was only freely available with the AnnoTree gene assignments. The number of genes classified for *M. chelonae* HPA 006 is shown in Table 10.

Table 10. Number of *M. chelonae* HPA 006 genes assigned in each of the different classification systems applied in MEGAN 6.

System	Total number of genes assigned	Percentage coverage of the *M chelonae* HPA 006 genome achieved
Interpro2GO	2,461,418	52.8%
EggNOG	739,311	16.3%
SEED	1,481,104	31.7%
KEGG	1,412,255	30.31%
EC	1,469,160	31.5%

Interpro2GO classified most genes from *M. chelonae* HPA 006 but the Gene Ontology (GO) hierarchical levels mix specific categories and are very general e.g., flagellar motility and

Metabolic process are categories at the same level. At level 1 there is a category Carbohydrate binding but it isn't a subcategory of Carbohydrate metabolism. Carbohydrate metabolic process is at level 2 under Metabolic process. So, the GO annotations do not produce such a useful high-level summary of gene categories.

The proposal to classify the activities of enzymes hierarchically, in what became the Enzyme Commission nomenclature, was published in 1961 and subsequently updated, now only on-line (https://web.archive.org/web/20180910045839/ and http://www.sbcs.qmul.ac.uk/iubmb/enzyme/ accessed 25/05/2022). Assigning genes in the whole genome to the 7 top level classes and 63 subclasses provides an enzymic profile of the strain. However, the assignment of genes e.g., to the category transferases, which differ as a percent of enzymes between *M. chelonae* HPA 006 and *M. abscessus*^{T}, does not address function.

Uniprot (The Universal Protein Resource) is a comprehensive resource for protein sequence and annotation data and function. Each protein field name has a link to the EC data for that entry. . Diamond + Megan (Bagci *et al.,* ,2021) annotates genes with their EC classification.

The KEGG classification (Kanehisa, 2000) encompasses the enzyme classification in that it assigns proteins to the reactions that they catalyse within a metabolic pathway (and they are labelled with EC numbers). This classification is much more directly linked to function than classification by enzyme type. The top level categories are shown in Table 11.

Table 11. Top level categories assigned by KEGG database (Kyoto Encyclopaedia of Genes and Genomes).

K1001100: Metabolism
K2000011: Genetic Information Processing
K2000016: Environmental Information Processing
K2000020: Cellular Processes
K2000025: Organismal Systems
K2000034: Human Diseases

The last two top level categories (organismal systems and human diseases) are directed more toward human and eukaryote systems and therefore not relevant to this project. The first four categories which are directed at both prokaryote and eukaryote organisms, could, when examined in greater detail potentially explain putative variations in virulence and antibiotic resistance between *M. chelonae* and *M. abscessus.* Each of these top level categories is further broken down into sub categories as described in Table 12.

Table 12. KEGG level 2 classification headings.

Top level KEGG classification	Classification sub-headings
Metabolism	Carbohydrate metabolism
	Transcription
	Lipid metabolism
	Energy metabolism
	Nucleotide metabolism
	Amino acid metabolism
	Metabolism of other amino acids
	Membrane transport
	Glycan biosynthesis and metabolism
	Metabolism of co-factors and vitamins
	Biosynthesis of polyketides and terpenoids
	Biosynthesis of other secondary metabolites
	Xenobiotics biodegradation and metabolism
Genetic Information	Transcription
	Translation
	Folding, sorting, degradation
	Replication and repair
Environmental Information	Membrane transport
	Signal transduction
	Signalling molecules/Interaction
Cellular processes	Transport and catabolism
	Cell motility
	Cell growth and death

It was also useful to compare the data for *M. tuberculosis* as a representative human pathogen and *M. smegmatis* as an environmental strain from the same genus as *M. chelonae* and *M. abscessus*, and finally *E. coli I* and *S. coelicolor* as well studied model organisms. Each KEGG category for each organism was created and a bar chart generated (Figure 47).

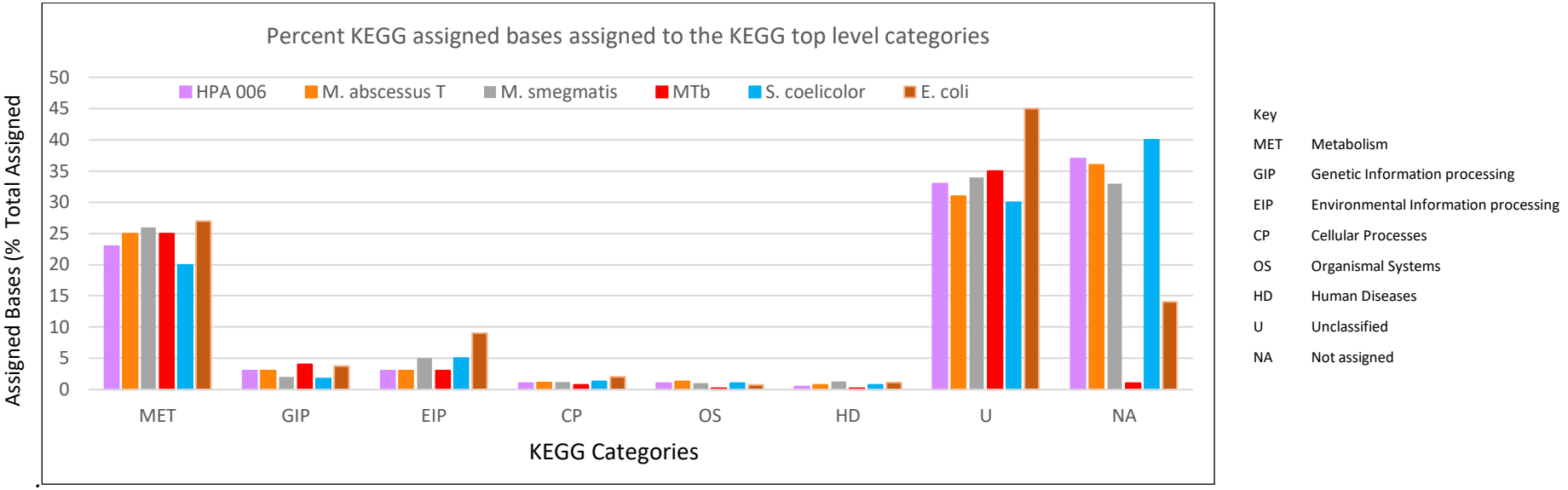
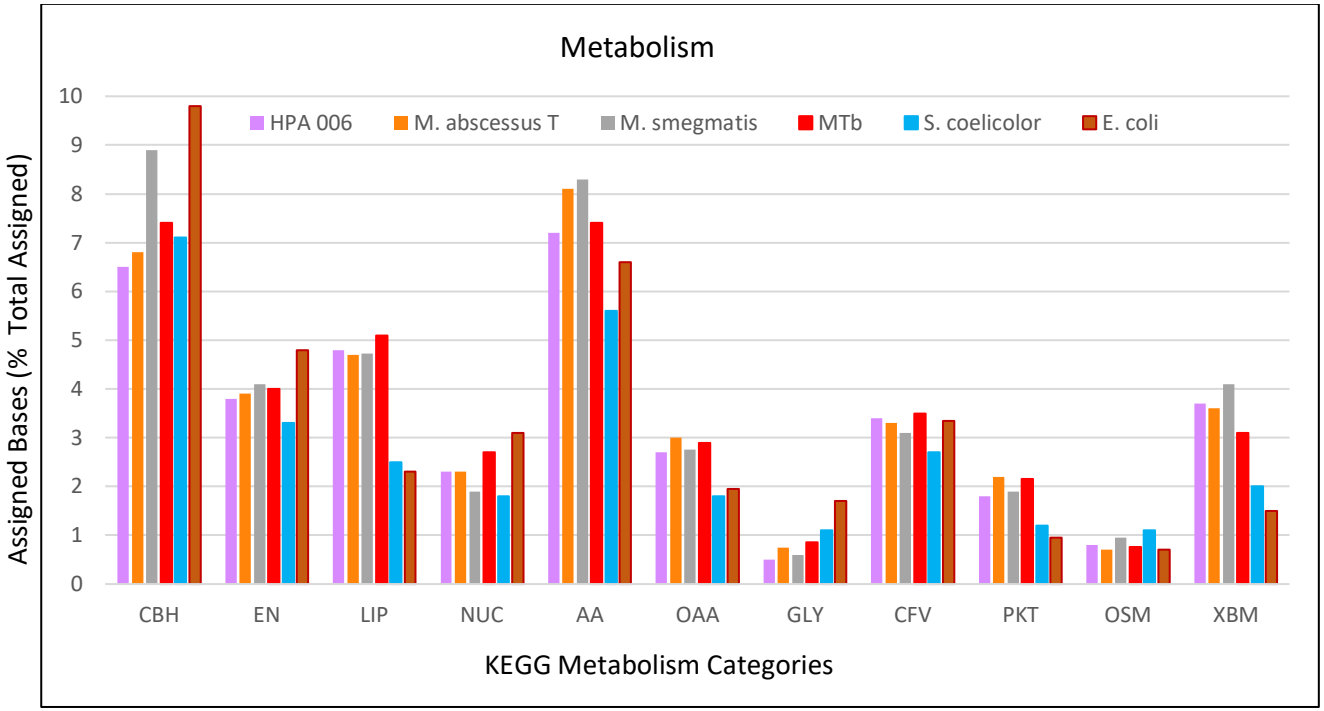


Figure 47. Assignment of genes to top level KEGG categories for the human pathogen *M. tuberculosis* H37Rv, the type strain of *M. abscessus*, the clinical isolate of *M. chelonae* HPA 006, the environmental strains *M. smegmatis* and *S. coelicolor* A3(2) and the model organism *E. coli*

In this comparison of genes assigned to the top level categories of KEGG (expressed as a percentage of the genome size) *M. tuberculosis* and *M. abscessus* do show a slight increase in genes assigned by AnnoTree to metabolism compared to *M. chelonae* HPA 006. There are two interesting categories not present in other classifications, Genetic Information Processing and Environmental Information Processing. As may be expected *M. smegmatis* (and *S. coelicolor*) is, relatively, higher in Environmental Information Processing than *M. tuberculosis. M. abscessus* and *M. chelonae* HPA 006 as opportunistic pathogens are also lower than those more environmental strains, but they are the same as one another. (The *E. coli* data is distorted relative to the other strains because it is so well studied that far fewer genes are unassigned).

Figure 48 illustrates the level 2 categories for Metabolism for all the strains. These high-level summaries of gene categories, including those in the other top level categories, show little difference between *M. chelonae* HPA 006 and *M. abscessus* or even other *Mycobacteria* including *M. tuberculosis*.



Key

CBH	Carbohydrate metabolism
EN	Energy metabolism
LIP	Lipid metabolism
AA	Amino Acid metabolism
OAA	Metabolism of other amino acids
GLY	Glycan biosynthesis and metabolism
CFV	Metabolism of cofactors and vitamins
PKT	Biosynthesis of polyketides and terpenoids
OSM	Biosynthesis of other secondary metabolites
XBM	Xenobiotics degradation and metabolism

Figure 48. Assignment of genes to KEGG categories under Metabolism for the human pathogen *M. tuberculosis* H37Rv, the type strain of *M. abscessus*, the clinical isolate of *M. chelonae* HPA 006, the environmental strains *M. smegmatis* and *S. coelicolor* A3(2) and the model organism *E. coli*.

The most interesting category of genes which differ between *M. chelonae* HPA 006 and *M. abscessus*^{T} is secondary metabolism (Figure 48) – Biosynthesis of Polyketides and Terpenoids.

Polyketides are biosynthesised by sequential condensation of simple two-carbon acetate units derived from malonyl-CoA with an acyl starter. These reactions are catalysed by polyketide synthases (PKSs), a family of multi domain enzymes or enzyme complexes that produce polyketides, a class of secondary metabolites which are present in bacteria, fungi and plants. Polyketide synthases are classified into three types (I, II and III) on the basis of their domain structures and subunit organisations (Shimuzu *et al*., 2017).

These complex secondary metabolites act iteratively or are linked through directed substrate transfer into modular groups (Herbst *et al*., 2018). They are one of the largest groups of natural secondary metabolic products and have been shown to play an important role in the life cycles of organisms which produce them, notably serving as chemical defence agents (Pfeifer & Khosla, 2001; Funabashi *et al*., 2008 ; Zeng *et al*., 2012). Type III polyketide synthase genomic clusters have been associated with diverse bio-functionalities such as conferring antibiotic resistance in Streptomyces species (Funabashi *et al*., 2008), the development of decay resistant outer coatings, exines, in dormant Azotobacter cells (Funa *et al*., 2006) , and the survival capability of mycobacteria in anaerobic biofilms (Anand *et al*., 2015). A large number are also biologically active with concomitant pharmaceutical potential (Singh *et al*., 2018).

When the complete genome sequence of *M. tuberculosis* became available in 1998 it showed an unexpectedly high number of open reading frames encoding proteins with homology to polyketide synthases (Cole *et al.,* 1998). A study carried out in 2005 by Jenke-Kodama *et al.,* further indicated that PKS and fatty acid synthases (FAS) have passed through a joint evolution process, in which modular PKS have a central position. Mega-synthetic polyketide synthases alongside fatty acid synthases, have been shown to generate a range of polyketide lipids that coat the *M. tuberculosis* cell envelope, enhancing the organism's virulence (Trivedi *et al*., 2005).

The majority of the PKS encoding genes of *M. tuberculosis* have been linked to specific biosynthetic pathways required to produce unique lipids or glycolipid conjugates that are

critical for virulence and host pathogen interactions (Quadri, 2014). The availability of comparative published genomes and their analyses has demonstrated the conservation of several biosynthetic PKS genes across various mycobacterial species (Kneitz & Dandekar, 2006). Many are clustered in biosynthetic operons and are uniquely conserved in genomes of pathogenic mycobacteria (Parvez *et al*., 2018).

The *M. marinum* genome has 34 open reading frames (ORF) homologous to PKS genes, with four being Type III (Stinear *et al*., 2008). Type III polyketide synthases have shown potential to biosynthesise structurally diverse and distinctive metabolites, such as spirolaxine, isolated from *Sporotrichum laxum* ATCC 15155, which exhibits many biological activities including promising anti-*Helicobacter pylori* properties (Sun *et al*., 2016) and germicidin synthase from *Streptomyces coelicolor* (Chemler *et al*., 2012), germicidin is known to act as an autoregulatory inhibitor of spore germination.

A study carried out by Parvez *et al.,* 2018, analysed the functional characteristics of two novel Type III PKSs, namely, MMAR_2470 and MMAR_2474, in *M. marinum*. MMAR_2470 and MMAR_2474 belong to a unique PKS genomic cluster found exclusively in pathogenic mycobacterial species. The MMAR_2747 protein was shown to play a key role in the survival of mycobacteria in stationary biofilms.

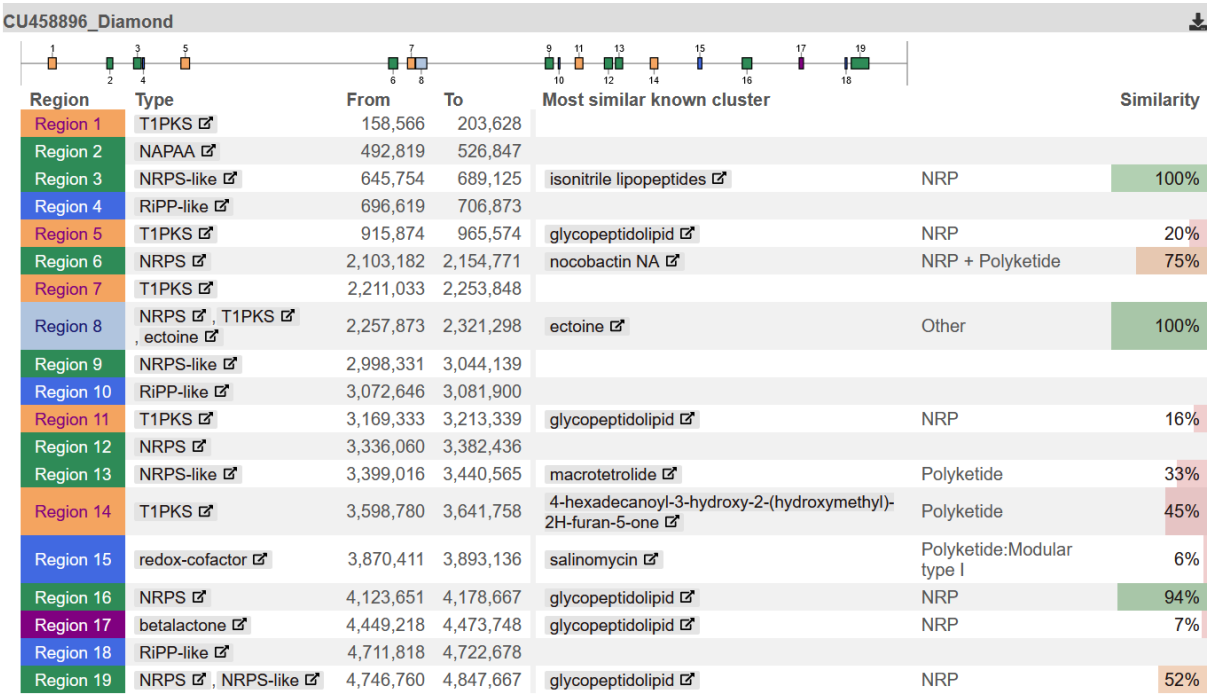
4.7.3 AntiSMASH analysis of M. chelonae HPA 006 and M. abscessus^T

AntiSMASH analysis (Medema *et al*., 2011; Blin *et al.,* 2021) shows 19 Biosynthetic Gene Clusters (BGCs) for the *M. abscessus* ^{T} (ATCC 19977) but 15 BGCs in *M. chelonae* HPA 006 (Figures 49, 50 and 51) Figure 51 shows 5 BGCs in *M. abscessus* not present in *M. chelonae* and 2 BGCs present in *M. chelonae* HPA 006 and not present in *M. abscessus*. At least 5 BGCs are linked to glycopeptidolipids which are important in evading the immune response.

A study which examined the natural biosynthetic gene products of the actinomycetes (Doroghazi & Metcalfe, 2013) concluded that these bacteria have a large number of natural product gene clusters. They noted that there were very similar repeated domains in the most common biosynthetic machinery, namely polyketide synthases (PKSs) and nonribosomal peptide synthetases (NRPSs) making comparative genomics of these areas challenging.

Additionally, the study found that conservation of these natural gene products within Mycobacteria (Streptomyces and Frankia) indicated key roles for these products within the genera.

For many of the genera examined, the most conserved secondary metabolite clusters were siderophores (NRPS products). Within the mycobacteria, many of the PKS gene clusters were well conserved in large phylogenetic groups. This was in contrast to NRPS clusters which were either unique or shared with a close relative. One exception to this is the gene cluster for mycobactin synthesis, a characterized siderophore, which was found in all strains of mycobacteria examined except *M. leprae*. The *M. abscessus*^{T} strain analysed in this study had approximately 5,000 genes and 15-17 BGCs identified comprising NRPS and PKS classes.



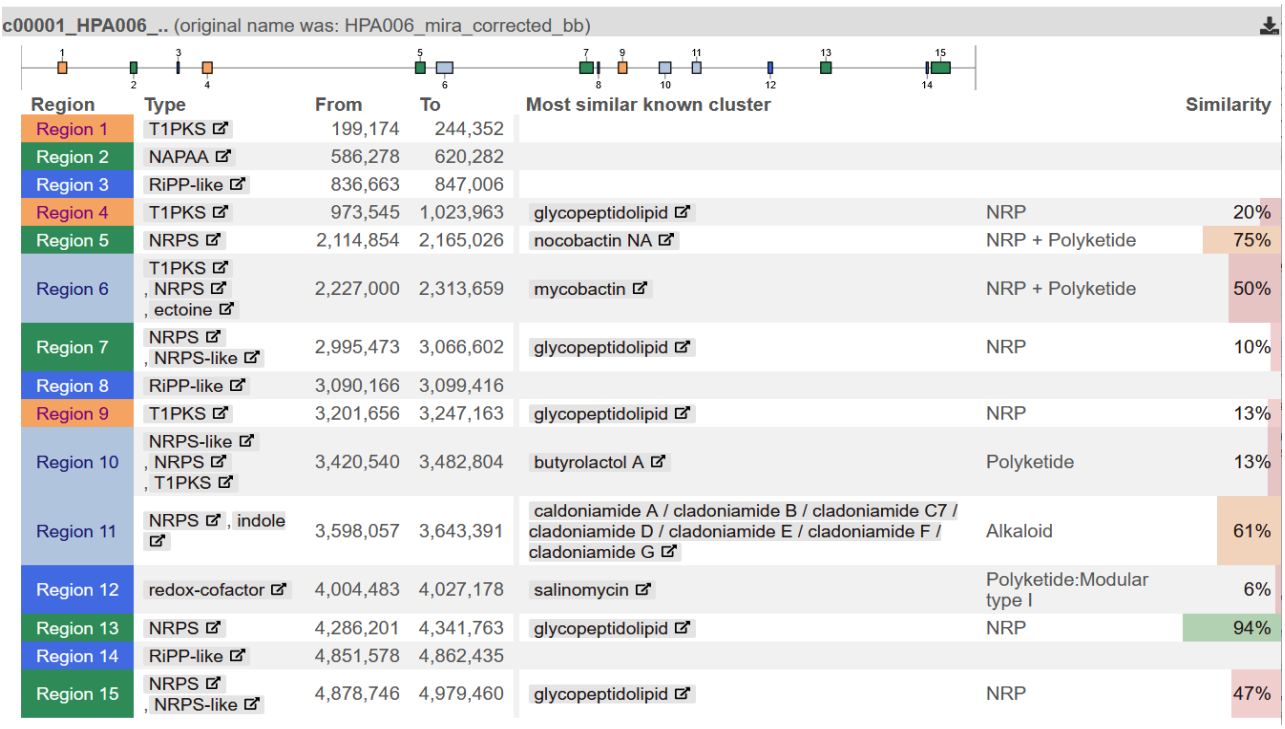
Region	Type	From	To	Most similar known cluster		Similarity
Region 1	T1PKS	158,566	203,628			
Region 2	NAPAA	492,819	526,847			
Region 3	NRPS-like	645,754	689,125	isonitrile lipopeptides	NRP	100%
Region 4	RiPP-like	696,619	706,873			
Region 5	T1PKS	915,874	965,574	glycopeptidolipid	NRP	20%
Region 6	NRPS	2,103,182	2,154,771	nocobactin NA	NRP + Polyketide	75%
Region 7	T1PKS	2,211,033	2,253,848			
Region 8	NRPS , T1PKS , ectoine	2,257,873	2,321,298	ectoine	Other	100%
Region 9	NRPS-like	2,998,331	3,044,139			
Region 10	RiPP-like	3,072,646	3,081,900			
Region 11	T1PKS	3,169,333	3,213,339	glycopeptidolipid	NRP	16%
Region 12	NRPS	3,336,060	3,382,436			
Region 13	NRPS-like	3,399,016	3,440,565	macrotetrolide	Polyketide	33%
Region 14	T1PKS	3,598,780	3,641,758	4-hexadecanoyl-3-hydroxy-2-(hydroxymethyl)-2H-furan-5-one	Polyketide	45%
Region 15	redox-cofactor	3,870,411	3,893,136	salinomycin	Polyketide:Modular type I	6%
Region 16	NRPS	4,123,651	4,178,667	glycopeptidolipid	NRP	94%
Region 17	betalactone	4,449,218	4,473,748	glycopeptidolipid	NRP	7%
Region 18	RiPP-like	4,711,818	4,722,678			
Region 19	NRPS , NRPS-like	4,746,760	4,847,667	glycopeptidolipid	NRP	52%

Figure 49. antiSMASH analysis results for *M. abscessus*^{T}.

Each region is colour coded e.g., Region 3 in this figure denotes an NRPS like region, the location in nucleotides (645,754 to 689,125) and that it shows 100% similarity with a known isonitrile lipopeptide cluster.

antiSMASH gives an overview of all the output results in a single page, showing all the detected biosynthetic gene clusters (regions) with their type classifications and nucleotide positions. Gene cluster types are signified by specific colors with green denoting regulatory genes, blue,

transport related genes, orange, additional biosynthetic genes, burgundy, core biosynthetic genes and grey, “other” genes.



c00001_HPA006_.. (original name was: HPA006_mira_corrected_bb)

Region	Type	From	To	Most similar known cluster		Similarity
Region 1	T1PKS	199,174	244,352			
Region 2	NAPAA	586,278	620,282			
Region 3	RiPP-like	836,663	847,006			
Region 4	T1PKS	973,545	1,023,963	glycopeptidolipid	NRP	20%
Region 5	NRPS	2,114,854	2,165,026	nocobactin NA	NRP + Polyketide	75%
Region 6	T1PKS, NRPS, ectoine	2,227,000	2,313,659	mycobactin	NRP + Polyketide	50%
Region 7	NRPS, NRPS-like	2,995,473	3,066,602	glycopeptidolipid	NRP	10%
Region 8	RiPP-like	3,090,166	3,099,416			
Region 9	T1PKS	3,201,656	3,247,163	glycopeptidolipid	NRP	13%
Region 10	NRPS-like, NRPS, T1PKS	3,420,540	3,482,804	butyrolactol A	Polyketide	13%
Region 11	NRPS, indole	3,598,057	3,643,391	caldoniamide A / cladoniamide B / cladoniamide C7 / cladoniamide D / cladoniamide E / cladoniamide F / cladoniamide G	Alkaloid	61%
Region 12	redox-cofactor	4,004,483	4,027,178	salinomycin	Polyketide:Modular type I	6%
Region 13	NRPS	4,286,201	4,341,763	glycopeptidolipid	NRP	94%
Region 14	RiPP-like	4,851,578	4,862,435			
Region 15	NRPS, NRPS-like	4,878,746	4,979,460	glycopeptidolipid	NRP	47%

Figure 50. antiSMASH analysis results for *M. chelonae* HPA 006.

In this figure Region 13 denotes an NRPS like region, the location in nucleotides (4,286,578 to 4,341,763) showing 94% similarity with a known glycopeptidolipid cluster.

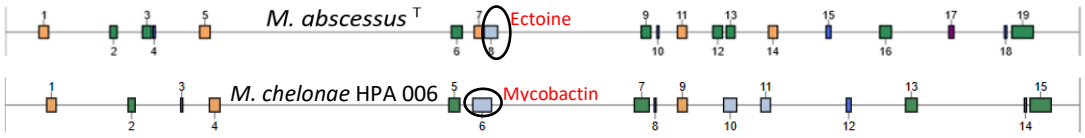


Figure 51. Comparison of Biosynthetic Gene Clusters (BGC) in *M. abscessus*^{T} and *M. chelonae* HPA 006.

Each of the colour coded regions indicate a type of BGC, as determined by antiSMASH. The presence of the same sized block in the same colour in a corresponding position in the genome (given the 82% identity established earlier between matching regions of *M. chelonae* HPA 006 and *M. abscessus*^{T}) implies a strong likelihood that these BGCs correspond to one another. Table 13 also illustrates this and highlights some differences e.g., Region 6 (*M. chelonae* HPA 006) and Region 8 (*M. abscessus*^{T}), appear to be corresponding regions (colour coding matches (Figure 51)) which should have similar functions. However, Region 6 in *M. chelonae* HPA 006

is a BGC which codes for mycobactin, whilst Region 8 (*M. abscessus*^{T}) is a BGC coding for ectoine. Region 6 in *M. chelonae* HPA 006, however, shows only a 50% similarity to the to the most similar known cluster, implying that these differences could be due to issues with detecting BGCs, and do not represent differences between the strains as described in Figure 52

The presence and absence of BGCs detected by antiSMASH is summarised in Table 13 and discussed in section 4.7.4

Table 13. AntiSMASH BGC regions present in *M. abscessus*^{T} and *M. chelonae* HPA 006 with red highlighting those present only in *M. abscessus*^{T} and blue highlighting those present only in *M. chelonae* HPA 006.

BGC Region	Description	*M. abscessus*	HPA 006 (Corresponding BGC Region)
1	T1PKS	*M. abscessus*	HPA 006 (1)
2	NAPAA	*M. abscessus*	HPA 006 (2)
3	Isonitrile lipopeptide	*M. abscessus*	
4	RiPP	*M. abscessus*	HPA 006 (3)
5	Glycopeptidolipid	*M. abscessus*	HPA 006 (4)
6	Nocobactin	*M. abscessus*	HPA 006 (5)
7	T1PKS	*M. abscessus*	
8	Ectoine	*M. abscessus*	HPA 006 (6) Mycobactin
9	NRPS-like	*M. abscessus*	HPA 006 Glycopeptidolipid (7)
10	RiPP	*M. abscessus*	HPA 006 (8)
11	Glycopeptidolipid	*M. abscessus*	HPA 006 (9)
12	NRPS	*M. abscessus*	
13	Macrotetrolide	*M. abscessus*	
14	4-hexadecanoyl etc	*M. abscessus*	
10	Butyrolactol A		HPA 006 (10)
11	Caldonamide		HPA 006 (11)
15	Salinomycin	*M. abscessus*	HPA 006 (12)
16	Glycopeptidolipid	*M. abscessus*	HPA 006 (13)
17	Betalactone	*M. abscessus*	
18	RiPP	*M. abscessus*	HPA 006 (14)
19	Glycopeptidolipid	*M. abscessus*	HPA 006 (15)

4.7.4 Isonitrile lipopeptide BGC

The first BGC of interest is the isonitrile lipopeptide present in *M. abscessus*^{T} but absent in *M. chelonae* HPA 006 (Figure 52). This BGC is 100% identical (antiSMASH calculation) to the

reference BGC (MiBIG BGC0001627) from *M. tuberculosis* for isonitrile lipopeptides (Harris *et al*., 2017) whose role may be in metal transport. The genes responsible for isonitrile peptide synthesis are significantly upregulated in biofilm formation by *M. abscessus* in synthetic cystic fibrosis medium (Belardinelli *et al.,* 2021). Two NRPS-encoding gene clusters (*mbt* and *Rv0096-0101*) have been identified from the genome of *Mycobacterium tuberculosis*, *mbt* has been characterized to biosynthesize mycobactin, a virulence-related iron siderophore (Harris *et al.,* 2017). *Rv0096*-0101 has been shown, to be essential for virulence and is present in pathogenic mycobacteria but not non-pathogens like *M. smegmatis* (Wang, *et al.,* 2007; Harris *et al.,* 2017). It has been shown to be responsible for the biosynthesis of a family of isonitrile lipopeptides probably involved in metal transport with zinc being the most likely candidate (Harris *et al.,* 2017).

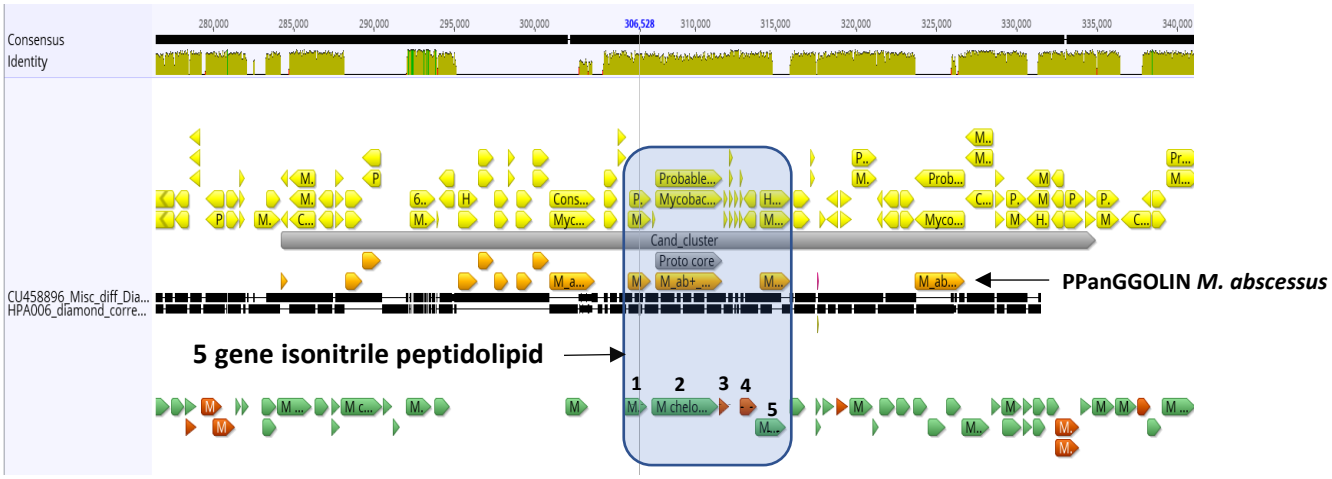


Figure 52. Isonitrile lipopeptide BGC in *M. abscessus*^{T} aligned with *M. chelonae* HPA 006.

Figure 52 shows that although antiSMASH does not identify an isonitrile lipopeptide BGC at this location in *M. chelonae* HPA 006 (denoted in green), a significant part of the BGC is actually present and the core gene is identical as identified in PPanGGOLiN.

There are 5 genes essential for isonitrile lipopeptide in *M. tuberculosis* (Figure 53), they correspond to MAB_0659 – MAB_0663 although PPanGGOLiN identifies 2 of these genes, MAB_0661 and MAB_0663 as missing, this is because, although the genes are present in *M. chelonae* HPA 006, they are low similarity. Genes HPA 006_000827 – HPA 006_000831 annotated as TauD/TfdA family dioxygenase, FcoT family thioesterase, AMP-binding protein, acyl carrier protein, AMP-binding protein align with MAB_0659 – MAB_0663 but only at 74.2% across the BGC and only 71.5% for the major NRPS (Non ribosomal peptide synthase).

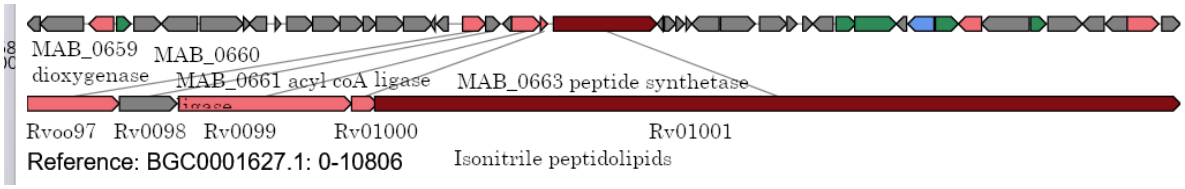


Figure 53. *M. abscessus*^{T} query for isonitrile lipopeptide BGC vs MiBiG BGC0001627 in *M. tuberculosis.*

Pangenome software is primarily designed for determining the pangenome of a species, so, in applying it to the *M. abscessus*/*M. chelonae* clade many proteins are about 82% nucleotide similarity (and higher protein similarity) and identified as homologous, but some genes are too distant. Similarly, in antiSMASH, the genes in *M. chelonae* HPA 006 (and other *M. chelonae*) are too dissimilar to identify as similar to BGC0001627, or as any putative BGC. Detecting homology in metagenome data (Diamond + Megan) and pan-genome data (PPanGGOLiN) are different problems. This explains the differences seen between comparisons performed by antiSMASH (Table 13) and PPanGGOLiN (Table 14) and the failure of antiSMASH to identify isonitrile lipopeptide BGC in *M. chelonae* HPA 006

Table 14. PPanGGOLiN analysis results for the MAB_0659 – MAB_0663 gene region.

	ATCC_19977 CDSs	Non-unique Gene name	No. sequences	M franklinii %	M abscessus abscessus %	M abscessus massiliense %	M chelonae %	M salmoniphilum %	M immunogenum %
	CDS_4795	persistent	1482	100	100	100	100	98	100	100
	CDS_4796	persistent	1485	100	100	100	100	98	100	100
	CDS_4797	persistent	1389	100	100	100	100	0	0	100
	CDS_4798	persistent	1481	100	100	100	100	98	100	100
	CDS_4799	persistent	1359	0	100	100	100	0	0	0

This BGC appears to be present in all species in the *M. abscessus*/*M. chelonae* clade, with the NRPS (MAB_0663) most dissimilar.

4.7.5 Mycobactin

A cluster of secondary metabolite BGCs are identified in *M. abscessus* (regions 6 nocobactin, 7 and 8 ectoine) matched by only 2 BGCs in *M. chelonae* (regions 5 nocobactin and 6 ectoine/mycobactin). Nocobactin is a lipid soluble iron binding compound similar to mycobactin, isolated from Nocardia. The genes in this cluster are annotated as mycobactin in *M. abscessus*^{T} CU458896 (Figure 54).

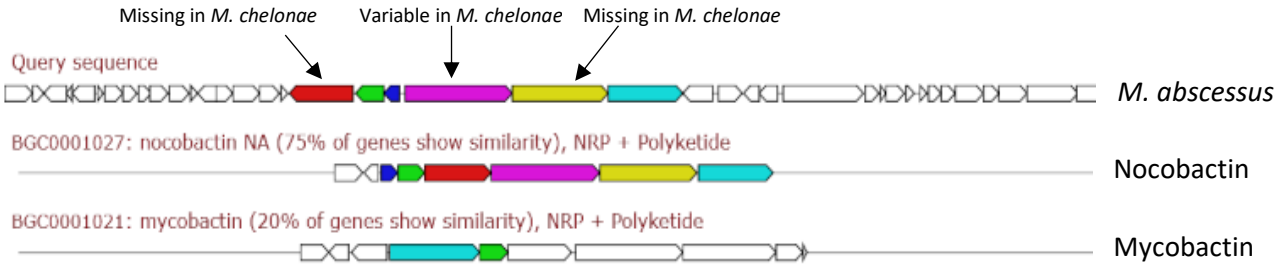


Figure 54. Comparison of the *M. abscessus*^{T} Region 6 BGC with MiBiG nocobactin and mycobactin gene clusters.

MAB_2119c and MAB_2123 are flagged as missing in all *M. chelonae* in PPanGGOLiN and others as somewhat variable across the strains. Comparing with these missing genes the *M. chelonae* HPA 006 cluster looks more like mycobactin. But alignment of *M. chelonae* HPA 006 and *M abscessus*^{T} (Figure 55) shows this alignment does not match the PPanGGOLiN analysis.

The MAB_2119c and MAB_2123 genes are flagged as missing in all *M. chelonae* strains included in the PPanGGOLiN analysis. Carrying out a comparison against these missing genes the *M. chelonae* HPA006 cluster looks more like mycobactin. But alignment of *M. chelonae* HPA006 and *M abscessus*^{T} (Figure 55 and Figure 56) shows this alignment does not match the PPanGGOLiN analysis.

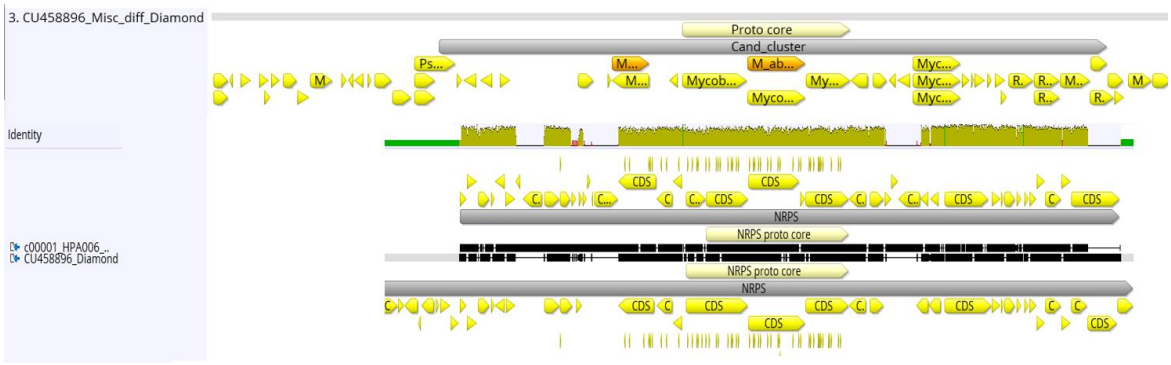


Figure 55. Alignment of region 6 in *M. abscessus*^{T} CU458896 with *M. chelonae* HPA 006.

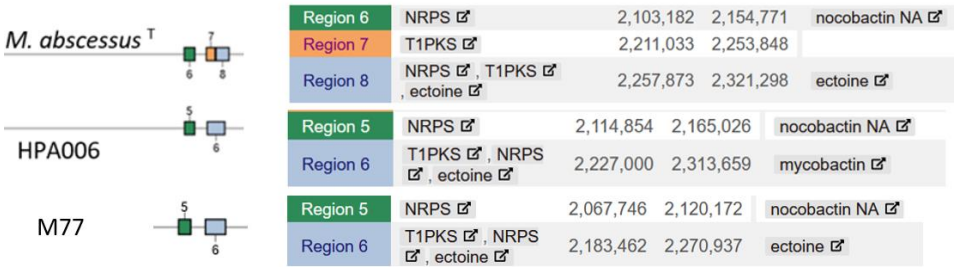


Figure 56. AntiSMASH description of region 6 in *M. abscessus*^{T} and the corresponding BGC regions in *M. chelonae* HPA 006 and *M. chelonae* M77.

Aligning the sequences which cover regions 5 and 6 in *M. chelonae* strains (HPA 006 and M77) and regions 6, 7 and 8 in *M. abscessus*^{T}, with *M. chelonae* HPA 006 annotated with Diamond + Megan, there is a region of mismatch just following the nocobactin protocluster in which *M. chelonae* HPA 006 has 2 genes more similar to *M. abscessus* than *M. chelonae*, these seem to be a regulatory protein and methionine synthase (metH)

M77 has insertion sequences, which are not in *M. chelonae* HPA 006, and only in some other *M. chelonae* strains (and as a complication M77 has an assembly gap and a small DUF3329 protein, not present in other *M. chelonae* strains). But *M. chelonae* HPA 006 and other *M. chelonae* do align contiguously across that gap. Figures 57 and 58 illustrate the variability in sequence, not only between species but within species.

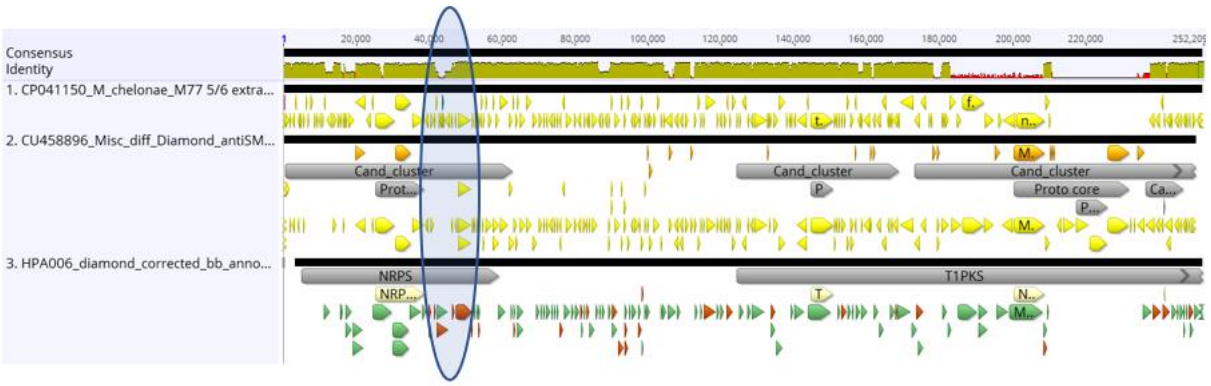


Figure 57. ProgressiveMauve alignment (displayed in Geneious) of regions 6, 7, 8 in *M. abscessus*^{T} and the corresponding regions in *M. chelonae* HPA 006 and *M. chelonae* M77.

Figure 58. NCBI Basic Local Alignment Search Tool (BLAST) analysis of *M. chelonae* M77 sequence against the NCBI nr-protein database showing that an IS3 insertion sequence is present in only some strains of *M chelonae.*

M. abscessus^{T} has 3 regions annotated by antiSMASH (6, 7 and 8) as nocobactin, a type 1 PKS cluster (T1PKS) and the BGC for ectoine. *M. chelonae* HPA 006 has only 2 regions (5 and 6), annotated as nocobactin and mycobactin. However, the same 2 regions in M77 are annotated as nocobactin and ectoine. And alignment of the sequences covering these regions shows that *M. chelonae* HPA 006 and M77 region 6 extends over regions 7 and 8 in *M. abscessus*^{T}. i.e. 2 regions in *M. abscessus*^{T} have been joined as 1 in the *M. chelonae* HPA 006 and M77 antiSMASH analysis.

The candidate cluster T1PKS (region 7 in *M. abscessus*^{T} and the start of region 6 in *M. chelonae* HPA 006) begins with 6 genes in the cobalamin biosynthesis pathway which are conserved

across *M. abscessus* and *M. chelonae*. This is followed by 2 pyridoxime-5-phosphate (PPOX) F420 oxidoreductases and an fsxA-like membrane protein (exclusion of T7 phage from E coli) only 1 of which is annotated in *M. abscessus*^{T} (with 3 short deletions compared to *M. chelonae*). Three genes in *M. abscessus*^{T} are then conserved across the strains but with 3 extra genes present in *M. chelonae,* between genes 2 and 3, with partial sequence matches in *M. abscessus*^{T} (being deleted in *M. abscessus*^{T}?). Then a Type 1 PKS gene, the basis for BCG region 7 in *M. abscessus*^{T} is conserved. After a run of genes, which seem to vary more in annotation than sequence, a CocE/NonD family hydrolase, which seems to be a dipeptide peptidase in some distantly related-organisms, is conserved followed by a tryptophan rich sensory protein, present in *M. abscessus*^{T} but absent in the *M chelonae*.

Then a run of 11 Type VII secretion associated proteins (Table 15) are conserved in *M. chelonae* and *M. abscessus*^{T} (Figure 59). The first, annotated as Type VII secretion protein EccE, is not annotated in *M. abscessus*^{T} but the DNA sequence aligns, but only at 73% similarity, and in PPanGGOLiN, its annotated protein in *M. abscessus*^{T} at that location is absent from *M. chelonae*.

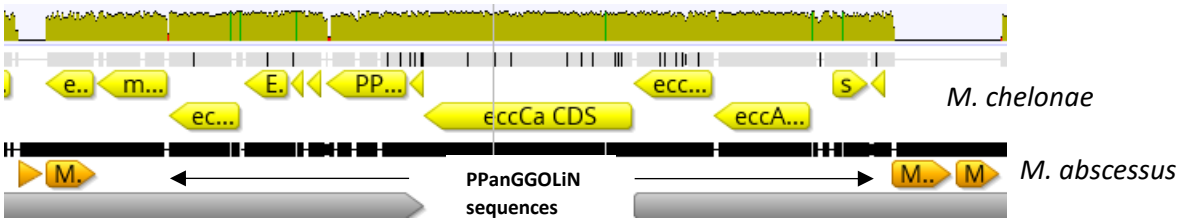


Figure 59. Type VII secretion genes in *M. chelonae* HPA 006 compared with the *M. abscessus*^{T} sequence alignment and the presence/absence of genes in PPanGGOLiN . The PPanGGOLiN sequences are denoted by the orange triangles.

Table 15. Type VII secretion system genes

Type VII Secretion	**Finding**
Type VII secretion protein EccE	This gene is not annotated in *M. abscessus*^{T} CU458896, however the sequence is 73% identical to the same sequence in *M. chelonae* HPA 006
Type VII serine protease mycosin MycP Type VII integral membrane protein EccD ESX secretion-associated protein EspG WXG100 Type VII secretion target	These genes are not annotated in *in M. abscessus*^{T} CU458896, however the sequence is 94% identical to the same sequence in *M. chelonae* HPA 006
Type VII secretion EsxS (ESAT-6 locus esx3) PPE family protein PE family protein	These genes are not annotated in *M. abscessus*^{T} CU458896, however the sequence is 85% identical to the same sequence in *M. chelonae* HPA 006
Type VII secretion protein EccCa Type VII secretion protein EccB Type VII secretion AAA-ATPase EccA	These genes are annotated as FtsK *in M. abscessus*^{T} CU458896, however the sequence is 86% identical to the same sequence in *M. chelonae* HPA 006

These genes (Table 15) probably have a role in virulence (Bunduc *et al.,* 2021), but they are present in both *M. abscessus* and *M. chelonae* so do not account for differences in the strains. The gene present in *M. abscessus* but absent in *M. chelonae* next to eccE is a regulatory protein, the pair of genes after eccA are a flavin-dependent oxidoreductase and LysR regulatory protein.

Although the sequence at the start of *M. abscessus*^{T} region 8 (Figure 54) is shown as mismatched with *M. chelonae* HPA 006, the sequences are annotated similarly (global alignments of long sequences such as whole genome sequences can show regions of misalignment in sequence alignment programmes such as MAFFT (Lee *et al*., 2021)). Extraction of these misaligned sequences and realignment with MAFFT gives a better alignment (Figure 60).

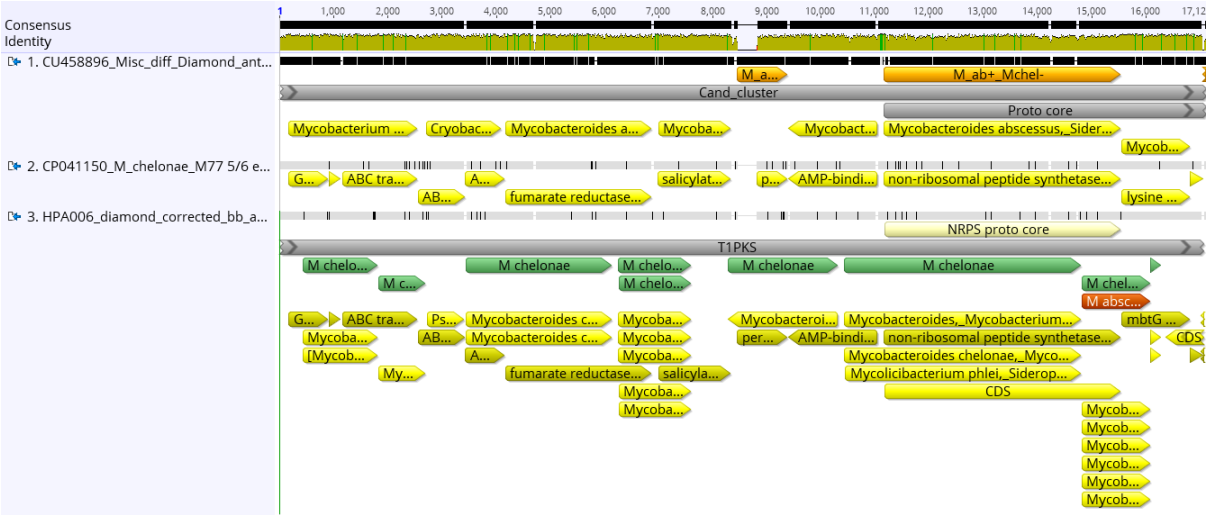


Figure 60. MAFFT alignment of misaligned sequences seen previously in Figure 52. *M. chelonae* HPA 006 annotated with Diamond + Megan.

This region has a lower similarity (78%) between *M. chelonae* HPA 006 and *M. abscessus*^{T} (*M. chelonae* HPA 006 and M77 are >99% identical) compared to a typical value of 85%. The NRPS gene highlighted as present in *M. abscessus* but missing in *M. chelonae* by PPanGGOLiN is, in fact, present in *M. chelonae* but at lower similarity (74%). Pan-genome software is designed to determine the pan-genome within a species, so we are stretching the software using it to compare across all the *M. abscessus* clade species. In metagenomes the detection of homologous genes may be based upon a threshold of about 40% but homology in PPanGGOLiN looks as though it may be based upon a threshold more like 75% (the actual comparison is made between protein sequences), which is probably appropriate within a species. So, the protein coding genes annotated as present in *M. abscessus* but absent in *M. chelonae* actually represent present/absent or significantly different.

The antiSMASH results for these regions (Figure 56) indicate nocobactin, a T1PKS and ectoine as the BGCs for *M. abscessus*^{T} but this low similarity mismatched NRPS follows the nocobactin BGC and a type 1 PKS gene, just before the type VII secretion genes. Following that is a putative salicylate synthase and a salicylate-mycobactin ligase and a partial match to mycobactin

(Figure 61f). In *M. abscessus*^{T} this is followed by an unknown BGC, absent from *M. chelonae*, between the partial mycobactin BGC and the ectoine BGCs (Figure 61).

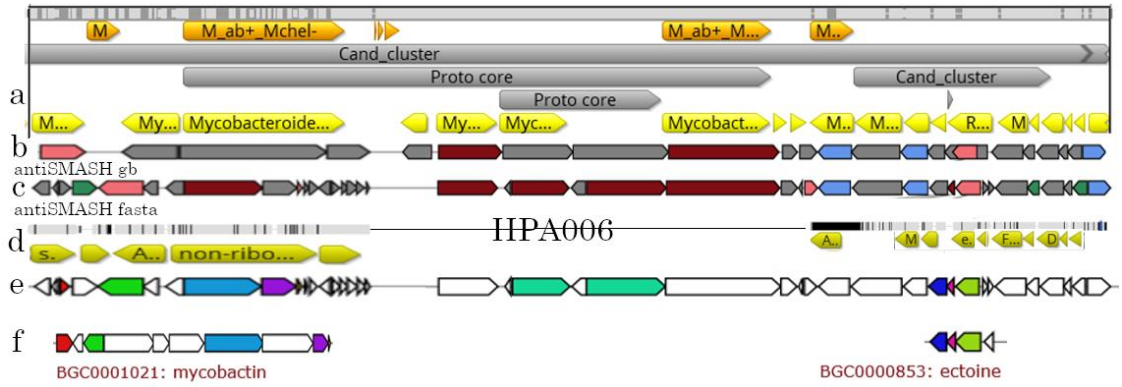


Figure 61. **a.** *M. abscessus*^{T} CU458896 annotation **b.** antiSMASH on CU458896 annotated genbank file **c.** antiSMASH on CU458896 FASTA (annotation by antiSMASH) **d.** alignment of *M. chelonae* HPA 006 to *M. abscessus*^{T} showing deletion. **e.** *M. abscessus*^{T} query sequence, from FASTA submission, displayed in matches to KnownClusterBlast. **f.** Matches to mycobactin and ectoine in KnownClusterBlast.

This combined secondary metabolite region of *M. chelonae* HPA 006 antiSMASH regions 5/6 and *M. abscessus*^{T} regions 6/7/8 begins with a recombinase and ends with a long run (missing region) of genes present in *M. abscessus* but absent in *M. chelonae* as shown in Figure 62.

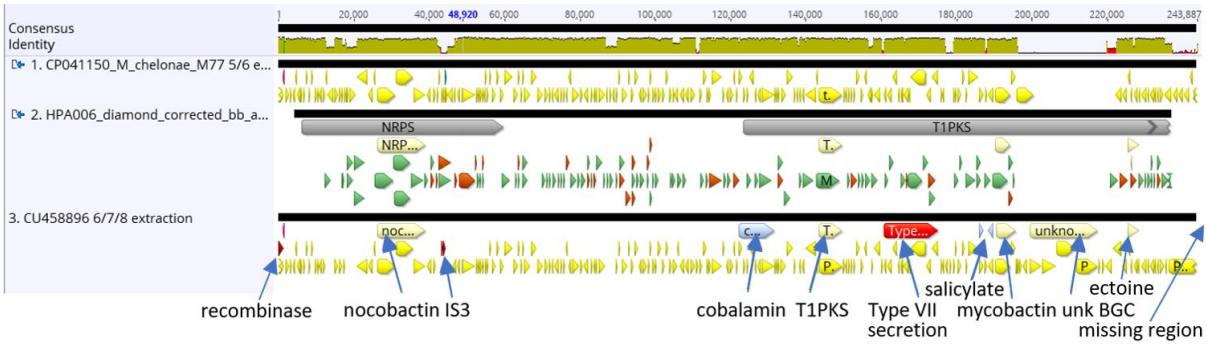


Figure 62. MAFFT alignment of *M. chelonae* strains (M77 and HPA 006) and *M. abscessus*^{T}, identifying BGCs in antiSMASH regions 5,6 in *M. chelonae* and 6,7,8 in *M. abscessus*^{T}.

Genes classified as present in *M. abscessus* and absent in *M. chelonae* by PPanGGOLiN, begin 4 genes downstream of the ectoine BGC, beginning with the DoxX gene (Figure 63). These genes are further collated in Table 16.

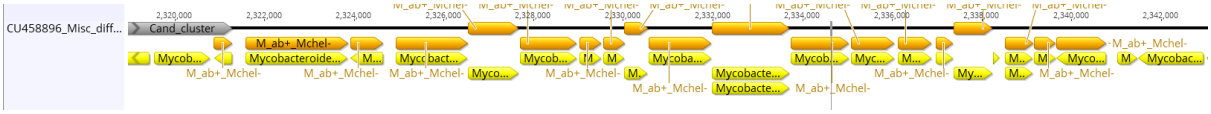


Figure 63. Genes classified as present in *M. abscessus* and absent in *M. chelonae* by PPanGGOLiN which begin 4 genes downstream of the ectoine BGC, beginning with the *Dox*X gene.

Table 16. *M. abscessus* specific genes in the DoxX cluster

PPanGGOLiN	**MLCG01**	**CU458896**	**Description**
CDS_4108	BKG55_20425	MAB_2274c	DoxX
CDS_4109	BKG55_20430	MAB_2275	**MMPL family transporter** (Ferrell *et al*., 2022)
CDS_4110	BKG55_20435	MAB_2276c	Winged helix-turn-helix domain - regulator
CDS_4111	BKG55_20440	MAB_2277	Aldehyde dehydrogenase
CDS_4112	BKG55_20445	MAB_2278	Acyl-CoA/acyl-ACP dehydrogenase
CDS_4113	BKG55_20450	MAB_2279	Aminotransferase class III-fold
CDS_4114	BKG55_20455	MAB_2280	SRPBCC
CDS_4115	BKG55_20460	MAB_2281	Hypothetical
CDS_4116	BKG55_20465	MAB_2282	**Flavin reductase** – required by MAB_2284
CDS_4117	BKG55_20470	MAB_2283	FAD-dependent monooxygenase
CDS_4118	BKG55_20475	MAB_2284	**Tryptophan 7-halogenase** (Ye *et al*., 2005)
CDS_4119	BKG55_20480	MAB_2285	Cation:proton antiporter - Na+/H+ antiporter
CDS_4120	BKG55_20485	MAB_2286	DUF2156 - O-seryl-isobutylhydroxylamine esterase
CDS_4121	BKG55_20490	MAB_2287	HD domain - metal dependent P-hydrolase
CDS_4122	BKG55_20495	MAB_2288	Hypothetical
CDS_4123	BKG55_20500	MAB_2289	Alpha/beta hydrolase
CDS_4124	BKG55_20505	MAB_2290	TetR/AcrR transcriptional regulator
CDS_4125	BKG55_20510	MAB_2291	SRPBCC
CDS_4126	BKG55_20515	MAB_2292c	Hypothetical

This cluster of *M. abscessus* and *M. immunogenum* specific genes codes for an MmpL family transporter (Table 16), these proteins are important in virulence and resistance in *Mycobacteria* and are over-represented in *M. abscessus* (Ferrell *et al*., 2022). They transport lipids, glycolipids, glycopeptidolipids and siderophores to the periplasmic space where they

contribute to the functional integrity of the outer cell surface which forms a barrier to the immune system response and chemotherapeutic agents and modifies host-pathogen interactions. They have been implicated in the virulence of *M. abscessus* and resistance to isoniazid, clofazimine and bedaquiline (see Ferrell *et al.*, 2022). There is no MmpL-associated regulatory protein MmpS nearby. There is also a flavin reductase and tryptophan-7 halogenase, required for the regio-specific halogenation of tryptophan which can act as the substrate to incorporate a halogen into halogenated natural products (Ye *et al.*, 2005).

In *M. chelonae* (HPA 006/*M. chelonae* M77) these 19 genes are replaced with 24 different genes. These genes, the *M. chelonae* DoxX cluster, are mostly present, as a cluster, elsewhere in the *M. abscessus*^{T} genome (Figure 64) in the reverse orientation.

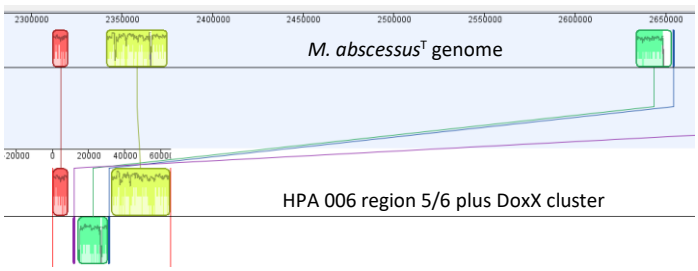


Figure 64. ProgressiveMauve alignment of *M. chelonae* region 5/6 to *M. abscessus*^{T}.

The corresponding 19 genes in the *M. abscessus*^{T} genome are absent from the *M. chelonae* genome. There are also genes which are missing from *M. chelonae* HPA 006 in the alignment (Figure 61, between the partial mycobactin BGC and the ectoine BGC, but present in *M. abscessus*^{T}. They are not flagged as present/absent, *M.* abscessus/M. *chelonae*, by PPanGGOLiN because these genes are present elsewhere in the *M. chelonae* HPA 006 genome (Figure 65).

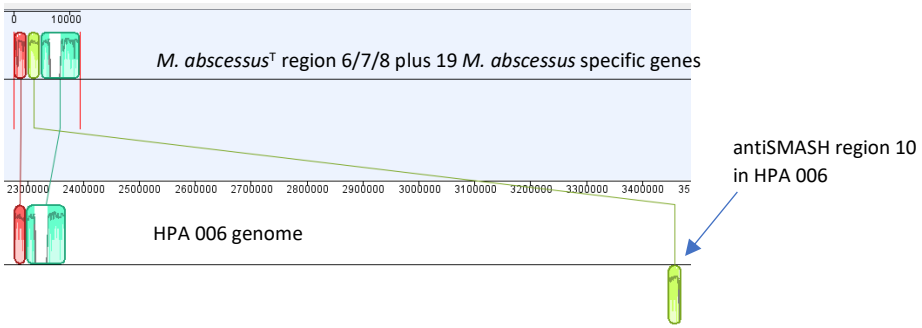


Figure 65. ProgressiveMauve alignment of *M. abscessus*^{T} region 6/7/8, plus 19 *M. abscessus* specific genes, to the *M. chelonae* genome.

M. abscessus genes(denoted by) are missing from region 5/6 in *M. chelonae* HPA 006 but present elsewhere in the genome. *M. abscessus* genes (19) (denoted by) absent from *M. chelonae* HPA 006 and replaced by 24 *M. abscessus* genes from elsewhere in the genome (Figure 64).

They correspond to antiSMASH region 10 in *M. chelonae* HPA 006 (Figure 66), which appears as an additional BGC compared to *M. abscessus*^{T} (Figure 49 and 50).

Region 10	NRPS-like, NRPS, T1PKS	3,420,540	3,482,804	butyrolactol A	Polyketide 13%

Figure 66. antiSMASH description of *M. chelonae* HPA 006 region 10.

However, it is clear that this BGC corresponds to the unknown BGC (Figure 62) between the partial mycobactin and ectoine BGCs in *M. abscessus*^{T}. I.e. *M. chelonae* HPA 006 BGC at region 10 is not missing from *M. abscessus* but relocated, the corresponding BGC is part of the complex antiSMASH *M. abscessus* region 6/7/8.

4.7.6 PE and PPE Family Proteins

PE/PPE family proteins are substrates for the type VII ESX export system and may be small molecule–selective channels analogous to outer membrane porins, which allow *M. tuberculosis* to take up nutrients while maintaining an otherwise impermeable barrier. There are over 168 PE/PPE (named for the proline (P) and glutamate (E) in the N-terminal) in the *M. tuberculosis* genome and they are widely thought to be important for virulence (Qinglan *et al*., 2020; Qian *et al*., 2020). *M. tuberculosis* has 5 five ESX gene clusters, *M. abscessus* only has two ESX gene clusters ESX-3 and ESX-4 (Kim *et al*., 2017). There are 8 PPE family proteins

annotated in *M. abscessus*^{T} (CU458896) by Diamond + Megan and 11 in *M. chelonae* HPA 006. There is some difficulty annotating them in *M. tuberculosis*.

PPE 1

MAB_0046 – MAB_0047 (Figure 67) are annotated as (probable) PE and PPE, they are identified by PPanGGOLiN as present in *M. abscessus* and *M. chelonae* (HPA 006_00093, HPA 006_00094) though they sit next to a region of difference MAB_0050c – MAB_0055c (Figure 64) though with only 76.3% identity, significantly lower than most homologous genes compared between *M. abscessus*^{T} and *M. chelonae* HPA 006. The next 2 genes MAB_0048 and MAB_0049 are shared EsxS type VII secretion and type VII secretion target genes at much higher similarity (89.4%). There are three patterns of similarity observed, genes with average similarity around 82%, genes with lower similarity, which PPanGGOLiN may not detect when homology is ~75% similarity or less and conserved genes with ~90% similarity. Both higher and lower similarity may reflect selection pressures, either for conservation of function or change in a specific function.

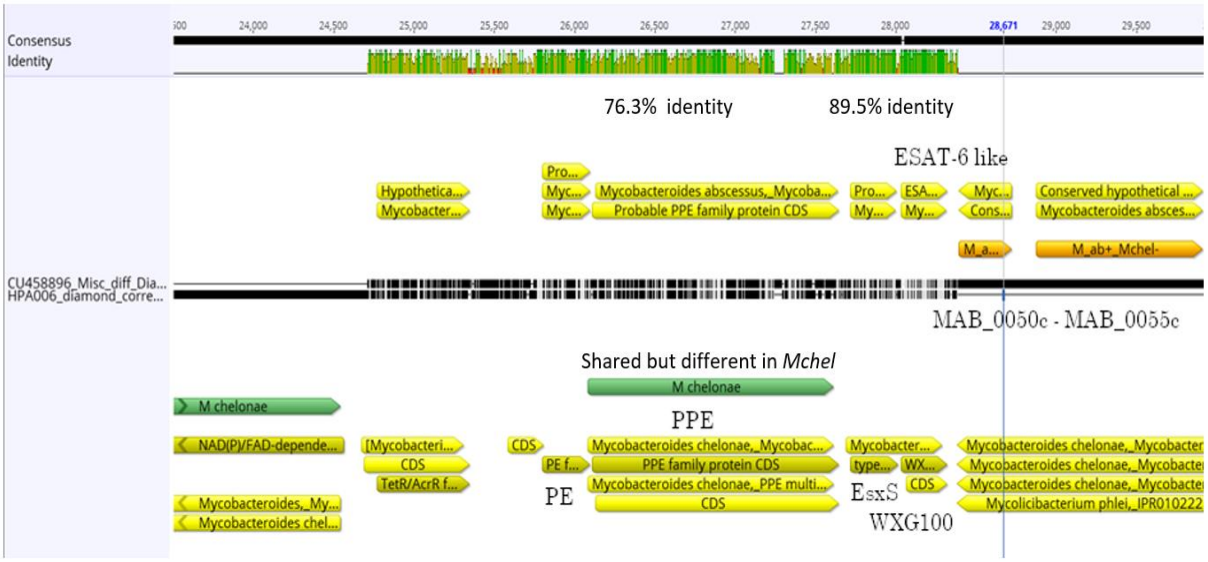


Figure 67. PE1 - PE/PPE genes shared by *M. abscessus*^{T} and *M. chelonae* HPA 006, MAB_0046 – MAB_0049 (PE, PPE – 76.3% and EsxS, WXG100 – 89.5%) ➡ denotes accepted gene annotations ➡ denotes other alternative gene matches.

Immediately preceding the MAB_0050c – MAB_0055c region of difference (this region of difference, present in *M. abscessus*^{T} and absent in *M. chelonae* HPA 006, contains three

conserved hypothetical genes, a cytidine amidase, a cyanate hydratase and a LysR regulator) is a set of PE/PPE and type VII secretion genes. They are annotated by PPanGGOLiN as present in both *M. abscessus* and *M. chelonae* but the PE and PPE genes are much less similar than the associated type VII secretion genes (annotated as EsxS/WXG100 in *M. chelonae* but ESAT-6 like in *M.* abscessus, despite sharing ~90% identity which may imply a different functional role for this type VII secretion system. This difference in PE/PPE between the strains is consistent with its presence in a general region of difference. The genes are preceded by an FAD dependent dehydrogenase present in *M. chelonae* but absent in *M. abscessus* and followed by the MAB_0050c – MAB_0055c region of difference. The LysR regulator (MAB_0055c) has been shown to play a role in the survival of *M. abscessus* in mice and a CRISPR based transcription knockdown shows it is involved in increased resistance to rifabutin (Nguyen *et al*., 2023).

There are eight PPE genes shared between *M. abscessus*^{T} and *M. chelonae* HPA 006 and three found only in *M. chelonae* HPA 006 (Table 17). The three in *M. chelonae* HPA 006 are not shared, at that location, by *M. abscessus* but the absence of PPE genes present in *M. abscessus* but absent in *M. chelonae* HPA 006 might be down to annotation.

Table 17. PE/PPE genes in *M. abscessus* ^{T} and *M. chelonae* HPA 006

PPE gene	***M. abscessus* (MAB)**	***M. chelonae* HPA006**	**Type VII secretion**
PPE 1	0046 – 0047	000093 – 000094	EsxS WXG100
PPE 2	0148c – 0149c	000184 – 000185	EspG
PPE 3		000225 PPE	
PPE 4		000773 – 000774	EspG
		000782 IS256	TNT EspG eccA mycP eccD eccCa
			eccD eccB eccE
PPE 5	0664 – 0665	000832 – 000833	EsxS WXG100
PPE 6	0809c PPE	000946 PPE	
PPE 7	0664 – 0665	000832 – 000833	EsxS WXG100
PPE 8	2230c – 2231c	002255 – 002256	eccE mycP eccD EspG WXG100
			EsxS (PPE PE) eccCa eccB eccA
PPE 9	4141 PE-PPE domain	004314	8 genes annotated MCE
PPE 10	4783	004931	

MAB_0809c PPE 6

The two genes annotated as PPE only are conserved within a long sequence of hypothetical proteins (Figure 68), presumably acquired by recombination around the, also conserved, 4 tRNAs. There is one gene annotated as a putative bacteriophage protein. This insert of hypothetical proteins is present in about 80% of *M. abscessus* subsp. *abscessus* and *bolletii* but only the occasional *M. abscessus* subsp. *massiliense* and none of the other species. At the end of this insert is the conserved (78% identity) PPEs immediately adjacent to a sequence present in all *M. abscessus* but absent in *M. chelonae* these include a putative choline oxidoreductase and a carnitine hydratase and are present in almost all *M. abscessus* subsp. *abscessus* and *bolletii*, 75% of *M. abscessus* subsp. *massiliense* and *M. immunogenum*.

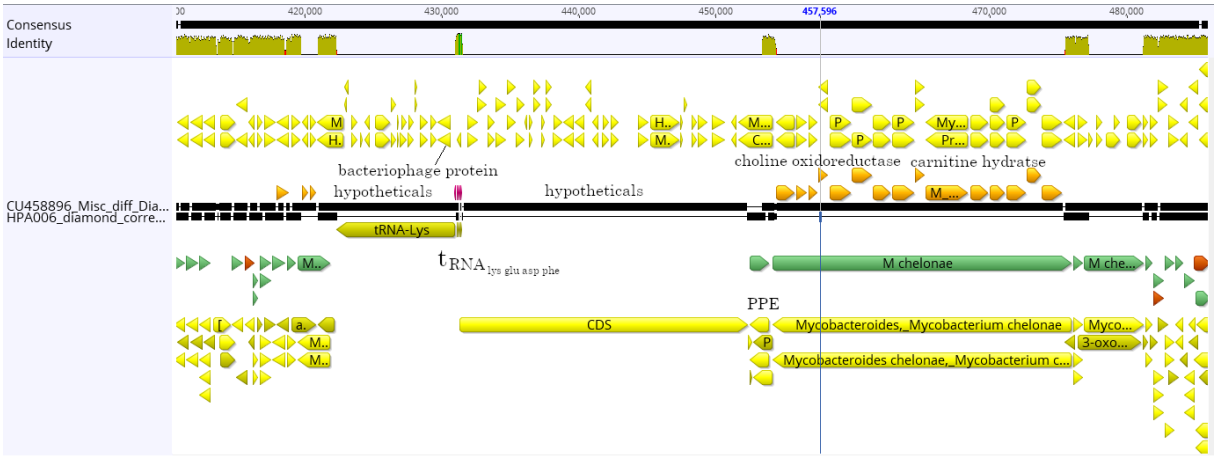


Figure 68. Conservation of a PPE (MAB_0809c) in a region of insertion in *M. abscessus*^{T}.

MAB_2230c PPE 8

The entire MAB_2230c PPE 8 cluster (eccE mycP eccD EspG WXG100 EsxS PPE PE eccCa eccB eccA) is embedded in the complex secondary metabolite BGCs for ectoine and mycobactin (see section 4.7.5) overall the cluster of genes, with the PPE and PE in its midst, shows 83% pairwise identity, but the PPE gene is only 76.8% identical (Figure 69).

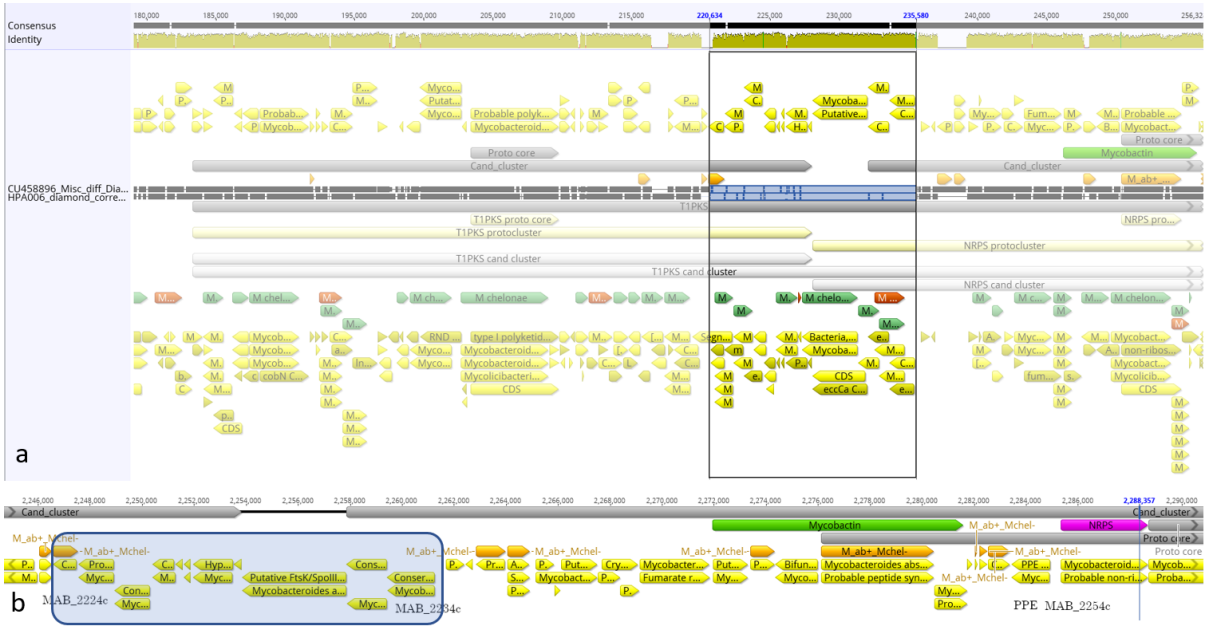


Figure 69. PPE 8 **a.** progressiveMauve alignment of *M. abscessus*^{T} and *M. chelonae* HPA 006 **b.** *M. abscessus*^{T} CU458896 annotated.

MAB_4141 PPE 9

This region of 3 genes, identified by PPanGGOLiN analysis as differing between *M. abscessus* and *M. chelonae* shows only 20% pairwise identity and only 40% identity between the PE-PPE domain proteins annotated on both *M. abscessus*^{T} and *M. chelonae* HPA 006 (Figure 70). Both proteins match a C-terminal domain present in PE and PPE proteins and are shortened compared to blastp hits outside the *M. abscessus*/*M. chelonae* clade.

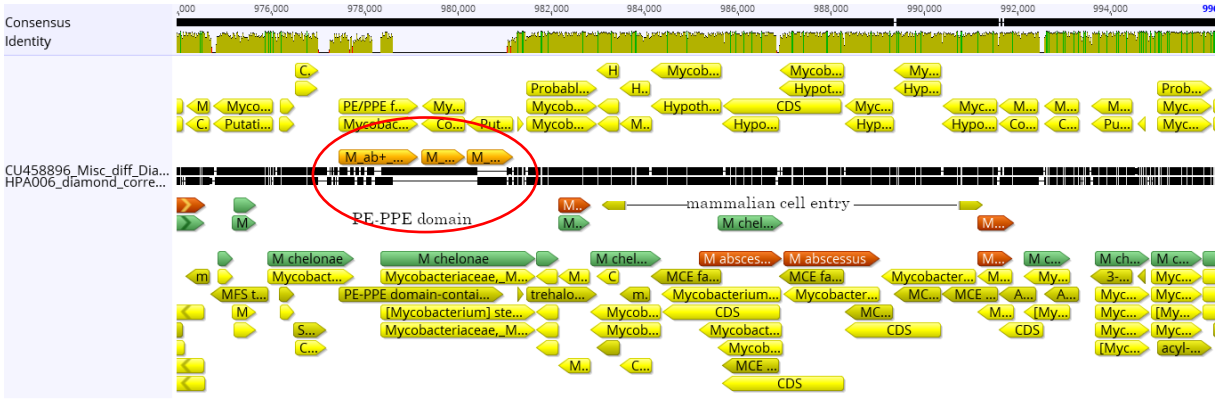


Figure 70. Illustrating variable identity in the PE-PPE region between the MAB_4141 gene and the homologous gene in *M. chelonae* HPA 006.

4.7.7 Daptomycin Resistance

Daptomycin (Baltz, 2009) is a calcium-dependent lipopeptide antibiotics which interacts with phosphatidylglycerol (PG) and other lipids such as undecaprenyl-linked cell wall precursors (Grein *et al*., 2018), compromising cell membrane integrity and leading to inhibition of cell wall synthesis. Resistance is linked to membrane structure and phospholipid composition (Wan-Ting *et al*., 2021) and overlaps with resistance to cationic antimicrobial peptides such as defensins (Montoya-Rosales *et al*., 2017).

Two genes in *M. abscessus*^{T} MAB_0123 and MAB_0573 are annotated as linked to daptomycin resistance, both are also present in *M. chelonae* HPA 006 at 83% nt similarity and conserved in all the *M. abscessus* clade species and strains.

Resistance to cationic antimicrobial peptides, resistance to acidic conditions and daptomycin (Montoya-Rosales *et al*., 2017) has been linked to lysyl-phosphatidylglycerol in the outer membrane which can change the surface charge of the cell. MprF, in e.g., *Staphylococcus aureus*, and lysX in *M. tuberculosis* code for a lysyl aminoacyl phosphatidylglycerol synthase (aaPG). An aaPg transfers an aa from its cognate tRNA to PG. The N-terminal domain is transmembrane and acts a flippase to translocate aaPG from inner to outer membrane, the C-terminal domain is in the cytosol and catalyses the formation of the aaPG. Expression levels of LysX (*rv1640c) in M. tuberculosis* strains can be correlated with virulence (Montoya-Rosales *et al.,* 2017). A lysX homologue is annotated in *M. abscessus*^{T} and *M. chelonae* HPA *006* but not as daptomycin resistant*,* MAB_2319c and HPA 006_002335 (81% id) and is present in all the *M. abscessus* clade species analysed in PPanGGOLiN.

A specific aaPG synthease (LysX2 – *rv1610*) has been identified in *M. tuberculosis* and is found in pathogens and not non-pathogens (Boldrin *et al*., 2022). It is found in *M. abscessus*^{T} (MAB_2639c) and *M. chelonae* HPA 006 (HPA 006_002643 – 91% id) by blastp of LysX2 and is present in all the *M. abscessus* clade strains for all species.

This is typical of the pattern, *M. abscessus* and *M. chelonae* have evolved significantly, most genes show only about 82-85% nucleotide identity, and there is considerable insertion e.g., of bacteriophage genes, and gene deletion/change (655 genes deleted or significantly different

out of 5012 = 13%), yet most genes that can be linked to resistance and virulence are conserved.

It is clear that many resistance/virulence mechanisms are not the result, simply, of the possession of a specific gene but also expression levels and single nucleotide polymorphisms. On that basis it is interesting that genes linked to resistance/virulence can fall into three classes: those that are strongly conserved (~90%), those showing the same similarity as the majority of genes (82-85%) and those showing greater variation (<78%)

4.7.8 Aminoglycoside Resistance

Victoria *et al*., (2021) identify aminoglycoside 2'-N-acetyltransferase and phosphotransferases as primary resistance mechanisms to aminoglycoside antibiotics. Deletion of the *eis2* gene increased susceptibility to kanamycin B, amikacin, hygromycin B and capreomycin, but did not affect susceptibility to tobramycin, dibekacin, arbekacin, gentamicin C, isepamicin, kanamycin A, apramycin and streptomycin.

Aminoglycoside 2'-N-acetyltransferase is MAB_4395c and the *eis2* gene is MAB_4532c. Both these genes are conserved in *M. chelonae* HPA 006 (Figures 71 and 72) and *M. chelonae* (PPanGGOLiN). MAB_4395 Aminoglycoside 2'-N-acetyltransferase --> HPA 006_004561 GNAT acetyltransferase

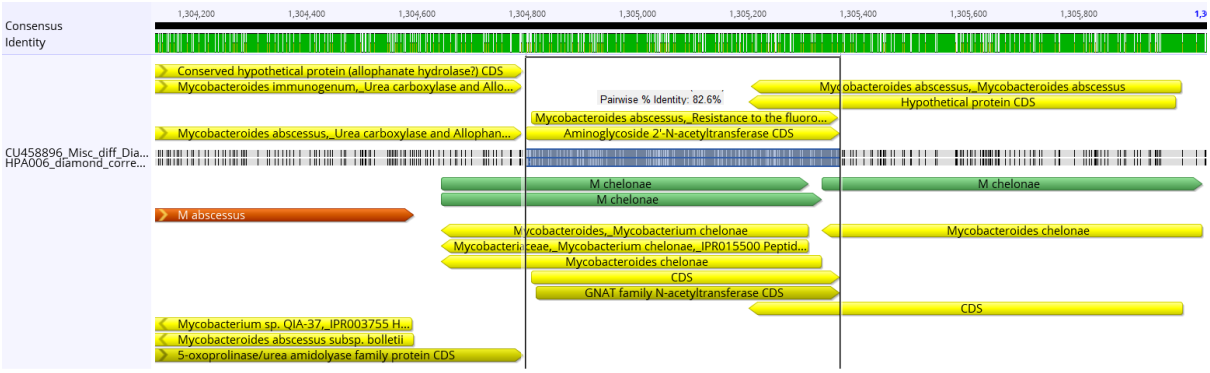


Figure 71. Aminoglycoside 2'-N-acetyltransferase (MAB_4395) in *M. abscessus*^{T} and *M. chelonae* HPA 006.

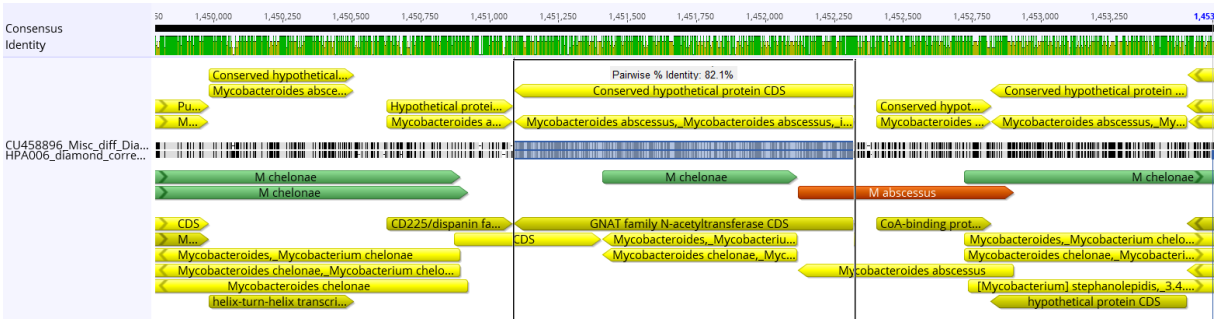


Figure 72. Aminoglycoside 2'-N-acetyltransferase (MAB_4532c) conserved in both *M. abscessus*^{T} and *M. chelonae* HPA 006.

MAB_0327, MAB_0951, MAB_3637c and MAB_4910c are aminoglycoside phosphotransferases (Nessar *et al*., 2012) except for MAB_0951, which is also annotated as a rifampin FADP-ribosyl transferase all these genes are conserved, in the same genomic context, in both *M. abscessus*^{T} and *M. chelonae* HPA 006, and present in both species (PPanGGOLiN). MAB_0951 is absent from *M. chelonae* HPA 006 and the PPanGGOLiN analyses shows it is one of 4 genes present in all *M. abscessus* strains but absent in all *M. chelonae* (Figure 73).

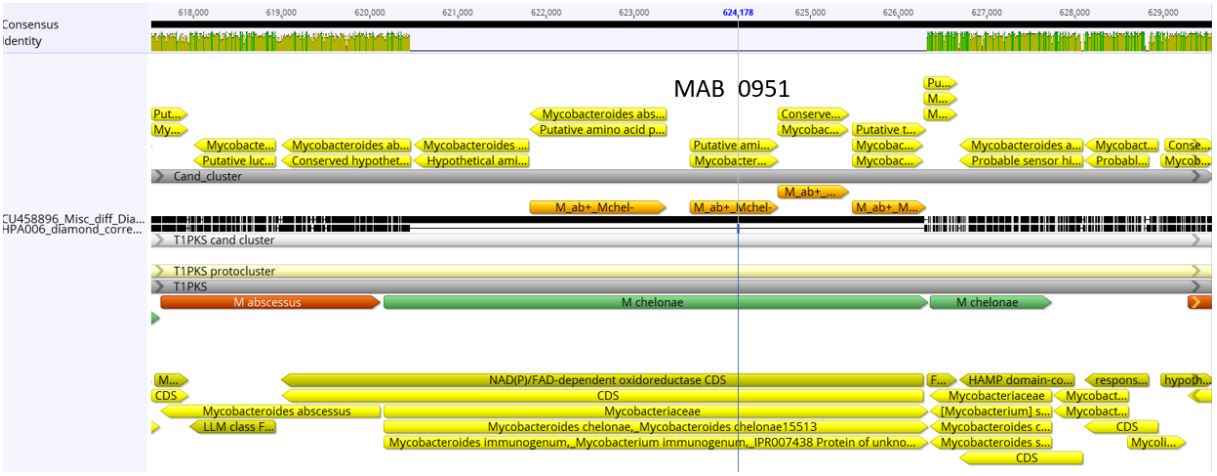


Figure 73. MAB_0951 gene present in *M. abscessus*^{T} but absent from *M. chelonae* HPA 006.

Similarly the beta-lactamase Bla_{mab} MAB_2875 is present in both species (Figure 74).

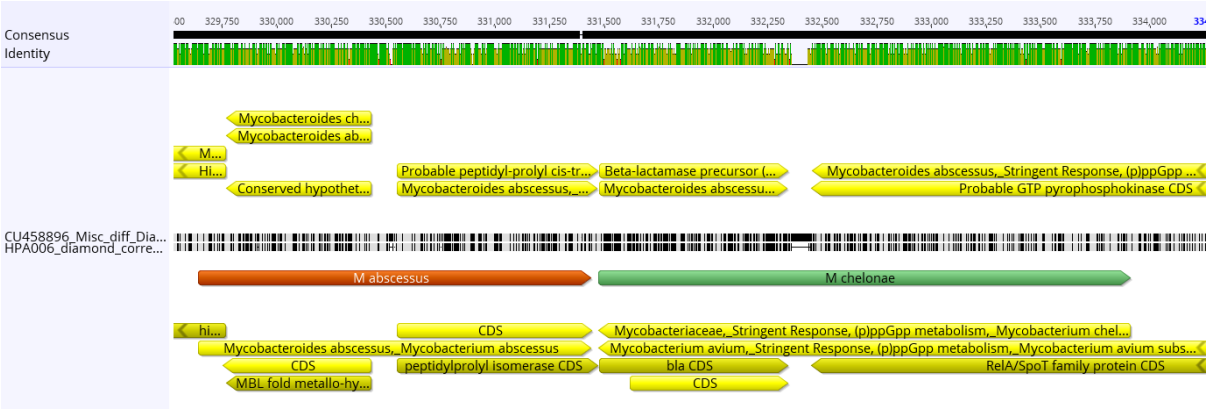


Figure 74. The Bla_{mab} beta-lactamase gene present in *M. abscessus*^{T} and *M. chelonae* HPA 006, and all other *M. chelonae* genomes.

4.7.9 *whi*B

MAB_3508c, resembles (75%) of the *M. tuberculosis whi*B7 gene, and is present in all strains in the *M. abscessus/chelonae* clade. Exposure to sub-inhibitory levels of tetracycline upregulates *whi*B7 and a cluster of regulated genes. Deletion of *whi*B7 *M. tuberculosis* makes it hypersensitive to multiple antibiotics (Morris *et al*., 2005). Deletion of MAB_3508c makes *M. abscessus* highly susceptible to erythromycin, clarithromycin, streptomycin, spectinomycin, amikacin and tetracycline (Hurst-Hess *et al*., 2017). There is 90% similarity in the sequence identity between *M. abscessus*^{T} and *M. chelonae* HPA 006. Seven *whi*B-like genes in *M. tuberculosis* are similarly induced by antibiotics and stress conditions (Geiman *et al*., 2006). Recent studies on tigecycline resistance in *M. abscessus* uncovered the involvement of the MAB_3508c gene which encodes the transcriptional activator *whi*B7. The deletion of *whi*B7 has been observed to cause a 4-fold decrease in the minimum inhibitory concentration of tigecycline (Ng and Ngeow, 2022).

There are 9 *whi*B annotated genes in *M. chelonae* HPA 006 and 7 *whi*B in *M. abscessus*^{T}.
The *whi*B **HPA 006_004489** (Figure 75) is missing from *M. abscessus*^{T}, replaced by MAB_4302 (hypothetical) and MAB_4303 (α-galactosidase).

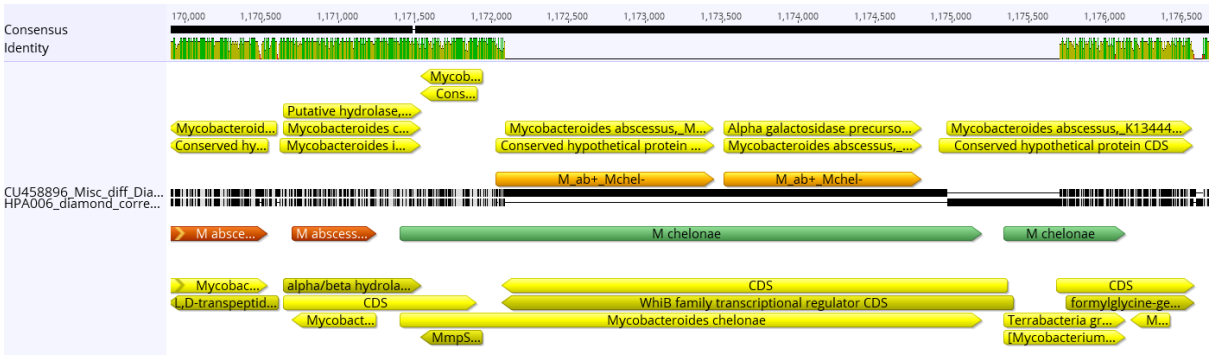


Figure 75. *M. chelonae* HPA 006 *whi*B annotated gene region HPA 006_004489.

HPA 006_004910 (Figure 76) is annotated as *whi*B in *M. chelonae* HPA 006 but it is not annotated in *M. abscessus*^{T} (CU458896). The intergene region between genes MAB_4762 and MAB_4763c is annotated by Diamond + Megan as a hypothetical but is 80% similar to the *whi*B gene annotated in *M. chelonae* HPA 006.

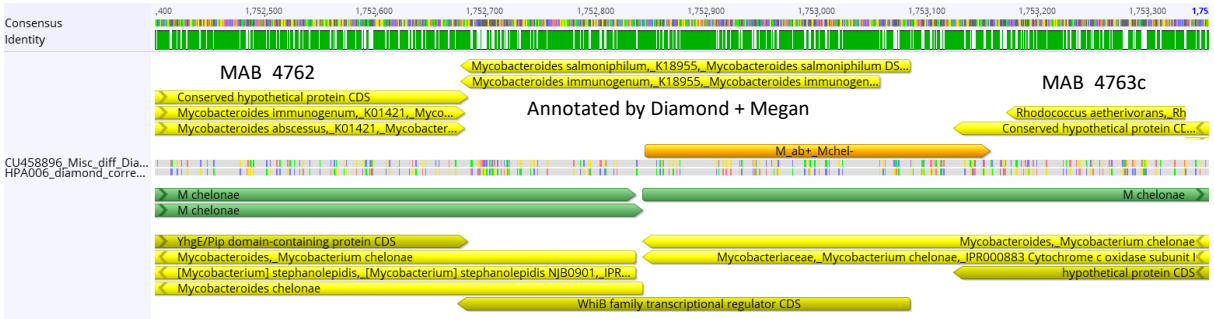


Figure 76. *M. chelonae* HPA 006 *whi*B annotated gene region HPA006_004910.

MAB_

MAB_1756 is in a long region absent from *M. chelonae* HPA 006 beginning with MAB_1723, a phage integrase and multiple hypothetical and phage-related proteins including another integrase, finishing at MAB_1834. In the middle is a conserved $tRNA_{met}$ (MAB_t5028). It is preceded by a shorter region, present in *M. chelonae* HPA 006 but absent in *M. abscessus* containing an IS3 transposase (HPA 006_001828).

The proteins in the region MAB_1723 – MAB_1834 are sporadically present, almost exclusively in *M. abscessus* and subspecies, at low levels (10/20/30% to absent).

MAB_4343c is at the start of a long region absent from *M. chelonae* HPA 006 beginning with MAB_4342c – MAB_4375 with no phage or integrase-like proteins

CDS_1173 – CDS_1208 in the PPanGGOLiN analysis in *M. abscessus*^{T}, corresponding to MAB_4343c – MAB_4375 are shell proteins present in about 60% of *M. abscessus* subsp. *abscessus* and a few in *M. abscessus* subsp. *bolletii* (25%) and only the odd protein in the odd strain in *M. abscessus* subsp. *massiliense*. Alsarraf *et al.*, 2022, demonstrated that whilst MAB_4324c is not essential for in vitro growth of *M. abscessus*, overexpression of the protein enhanced the uptake and survival of *M. abscessus* in THP-1 macrophages.

MAB_3446 (Figure 77) is in a region of poor similarity, it is not annotated by Prodigal (Hyatt *et al.*, 2010) within the PPanGGOLiN analysis as MAB_3445 and MAB_3447 are annotated in MLCG01 as CDS_0272 and CDS_0273.

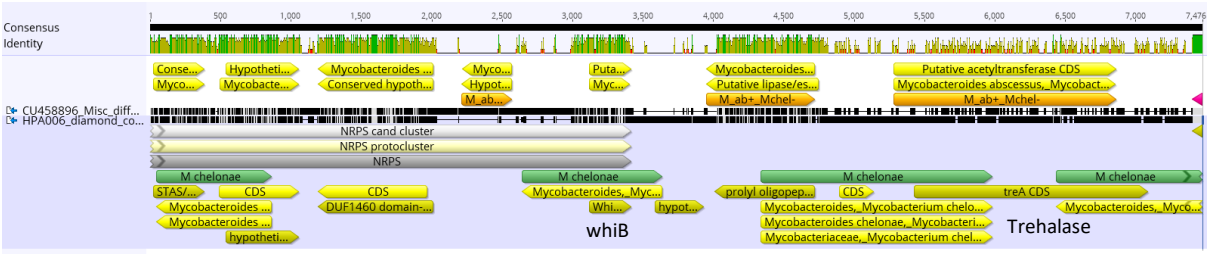


Figure 77. MAB_3446 gene (*whi*B) conserved within a region of variation between *M. abscessus* and *M. chelonae.*

Table 18. provides a summary of six additional MAB *whi*B regions and their corresponding status in *M. chelonae* HPA 006

Table 18. *whi*B regions present in *M. abscessus* T and their corresponding status in *M. chelonae* HPA 006

MAB *whi*B region	Status in *M. chelonae* HPA 006
MAB_3508c	Not annotated as *whi*B but the sequence shares 90.5% sequence identity to *whi*B HPA 006_003669
MAB_4715c MAB_4716c	This inter-spacer region shares 79.9% sequence identity with *whi*B HPA 006_004858
MAB_0409	This region shares 92% sequence identity with HPA 006_000450
MAB_3539	This region shares 96% sequence identity with HPA 006_003700
MAB_3606c	This region shares 96% sequence identity with HPA 006_003700
MAB_3726	This region shares 91% sequence identity with HPA 006_003860

The *whi*B-like genes at MAB_1756 and MAB_4343c are candidates for differential resistance between *M. abscessus* strains as well as *M. chelonae* (Burian *et al*., 2012). There is a *whi*B present in *M. chelonae* HPA 006 but absent from *M. abscessus*^{T} (HPA 006_004489**).**

4.7.10 Gorzynski Mutants

Gorzynski *et al*., (2021) generated a transposon knock out library of *M. abscessus*^{T} (ATCC 19977) and screened it against minimal inhibitory (MIC) or bactericidal (BC) concentrations of the antibiotics amikacin, clarithromycin, or cefoxitin. Knock out strains showing an increased resistance (BC) or decreased susceptibility (MIC) were sequenced to identify the gene location of the transposon. Selected genes with changed resistance from Table 3 in Gorzynski *et al*., (2021) are listed in Table 19 below

Table 19. Presence/absence of gene knockouts changing resistance to amikacin, clarithromycin, or cefoxitin in Gorzynski *et al*., (2021)

Gene	Function	MIC/BC	MIC/BC (%)	Gene Identification in HPA 006	Status in members of *M. abscessus/chelonae* Clade
MAB_0734	MspA porin	BC	82.0	HPA 006_000907	Present in all
MAB_0937c	MmpL	MIC	86.7	HPA 006_001027	Present in all
MAB_1137c	MmpL	MIC	0.0	-	*M. abscessus* 42%
MAB_1170	TauE *	MIC	81.8	HPA 006_001250	Present in all
MAB_1171c	Hypothetical	MIC	86.5	HPA 006_001251	Present in all
MAB_1839	GGDEF †	MIC	0.0	-	*M. abscessus* subsps
MAB_2421c	Hypothetical	BC	68.8	HPA 006_002438	*M. abscessus* subsps
MAB_2435	Mo transport	MIC	76.0	HPA 006_002452	Present in all
MAB_2787c	Mo transport	MIC	0.0	-	NOT *M. chelonae/M. salmoniphilum*
MAB_3384c	Mo transport	MIC	84.7	HPA 006_003464	Present in all
MAB_3465	SO_4 transport	MIC	0.0		*M. abscessus* subsps
MAB_4036	Hypothetical	MIC	72.1	HPA 006_004204	only 87% *M. ab abscessus*
					57% *M. salmoniphilum*
					NOT *M. immunogenum*
MAB_4117c	MmpS	BC	86.5	HPA 006_004290	Present in all
MAB_4116c	MmpL	-	85.0	HPA 006_004289	Present in all
MAB_4237c	GlnQ	MIC	86.1	HPA 006_004408	Present in all
MAB_4691c	NRPS antiSMASH 19 ‡	MIC	73.5	HPA 006_004833-5	Present in all
MAB_4915c	Hypothetical	BC	62.9	HPA 006_004971	*M. abscessus* subsps

Key : *TauE sulfite exporter

† GGDEF Diguanylate-cyclase (cyclic di-GMP)

‡ antiSMASH region 19 glycopeptidolipid

bold genes not present in *M. chelonae* but in all *M. abscessus* clade species

4.7.11 PPanGGOLiN Analysis

Compared to other software PPanGGOLiN can produce more cloud gene partitions than other software but for this analysis the focus is on persistent and shell genes. PPanGGOLiN is easy to install and run, and will accept FASTA and multi-FASTA genome files and annotate all the genomes consistently, using Prodigal (Hyatt *et al*., 2010).

The 1526 genomes downloaded as members of the *M. abscessus* clade required too much RAM to analyse with PPanGGOLiN (Bazin *et al*., 2020; Gautreau *et al*., 2020) in a 32 Gb RAM linux computer, so the genomes were reduced to 870 *M. abscessus* subsp. *abscessus*, 119 *M. abscessus* subsp. *bolletii*, 362 *M. abscessus* subsp. *massiliense*, 53 *M. chelonae*, 12 *M. franklinii*, 7 *M. salmoniphilum* and 12 *M. immunogenum*. Genomes which were identical or highly similar and multi-contig whole genome sequences (WGS) with large numbers of contigs were reduced.

Analysis of 1435 genomes was successful and generated a tab separated matrix of locus tags for each genome with each row containing the homologous genes from each genome, calculated from the annotated genes. The genomes were submitted, ordered to cluster the *M. abscessus* clade species together, so that the columns in the matrix were ordered by species. This tab separated matrix was read into Excel and the number and percentage of each gene in each species and subspecies calculated (Figure 78).

	Gene	ATCC_19977 CDSs	Non-unique Gene name	Annotation	No. isolates	No. sequences	Avg sequences per isolate	M franklinii	M franklinii %	M abscessus abscessus	M abscessus abscessus %	M abscessus bolletii	M abscessus bolletii %	M abscessus massiliense	M abscessus massiliense %	M chelonae	M chelonae %	M salmoniphilum	M salmoniphilum %	M immunogenum	M immunogenum %
2																					
3	NZ_FSJM01	CDS_0001	shell		163	163	1	0	0	140	16.09195	1	0.840336	15	4.143646	4	7.54717	0	0	0	0
4	NZ_FRYO01	CDS_0002	shell		166	166	1	0	0	143	16.43678	1	0.840336	16	4.41989	4	7.54717	0	0	0	0
5	NZ_FSMX01	CDS_0003	shell		168	168	1	0	0	144	16.55172	1	0.840336	16	4.41989	4	7.54717	0	0	0	0
6	NZ_CAACXQ	CDS_0004	shell		101	102	1.01	0	0	99	11.37931	0	0	1	0.276243	0	0	0	0	0	0
7	NZ_FSFP01_	CDS_0005	shell		119	121	1.02	0	0	110	12.64368	0	0	8	2.209945	0	0	0	0	0	0
8	NZ_FWDC01	CDS_0006	cloud		10	10	1	0	0	10	1.149425	0	0	0	0	0	0	0	0	0	0
9	FVBY01_M_	CDS_0007	shell		100	101	1.01	0	0	99	11.37931	0	0	1	0.276243	0	0	0	0	0	0
10	FVBY01_M_	CDS_0008	shell		100	101	1.01	0	0	99	11.37931	0	0	1	0.276243	0	0	0	0	0	0
11	NZ_FVYG01	CDS_0009	cloud		11	11	1	0	0	10	1.149425	0	0	1	0.276243	0	0	0	0	0	0
12	NZ_FWDC01	CDS_0010	cloud		11	11	1	0	0	10	1.149425	0	0	1	0.276243	0	0	0	0	0	0
13	NZ_FVXF01	CDS_0011	cloud		30	30	1	0	0	29	3.333333	0	0	1	0.276243	0	0	0	0	0	0

Figure 78. Calculation of the number and percent of each gene in genomes of *M. franklinii, M. abscessus* subsp*. abscessus, M. abscessus* subsp*. bolletii, M. abscessus* subsp*. massiliense, M. chelonae, M. salmoniphilum* and *M. immunogenum.*

The matrix was sorted on the percentage of each gene in *M. abscessus* subsp. *abscessus.* Those genes contained in more than 95% of *M. abscessus* subsp. *abscessus* strains were sorted on their percentage contained in *M. chelonae*. The genes present in more than 95% of *M.*

abscessus subsp. *abscessus* and absent in *M. chelonae* were selected. The list of genes is in Supplementary file (S3) "Ppanggolin_diff_Mab_Mchel.xlsx"

The type strain of *M. abscessus* was sequenced in 2007 (CU458896) but was re-sequenced in 2016 (MLCG01) this later, multi-contig sequence and assembly was chosen to include in the pangenome analysis. This genome was annotated and the genes assigned locus tags within PPanGGOLiN. These locus tags were used to label each gene row but most of the literature uses the locus tags (MAB_xxxx) from the CU458896 genome. The genes annotated in PPanGGOLiN, for each genome, are identified in a tab separated file (tsv) containing the PPanGGOLiN locus tag, the start and end position, and the containing contig. The tsv file for MLCG01 was used to determine the genes in the MLCG01 WGS contigs and these were mapped to the CU458896 genome to determine the corresponding MAB locus tags for all the genes present in *M. abscessus* but absent in *M. chelonae*. These genes were added, as an annotation, to the *M. abscessus* CU458896 genome using Geneious software (Biomatters Ltd).

The descriptions, in the annotations, of the 655 differential genes identified were sorted and duplicates merged to generate a list of gene types e.g., hypothetical, transcriptional regulator. The number of genes matching each type was counted in Excel using COUNTIFS(). Most genes had a unique annotation (Figure 79) but the largest category was hypothetical and proteins with domains of unknown function.

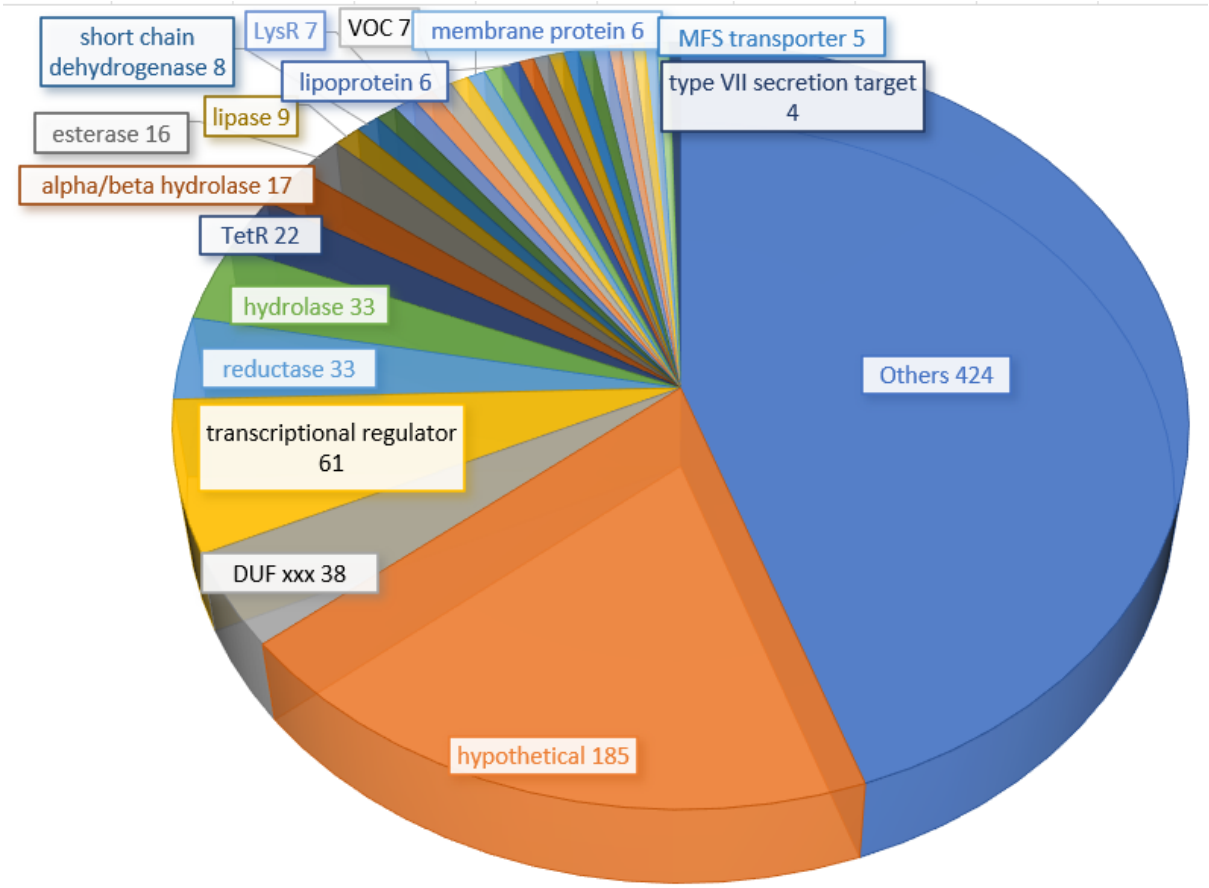


Figure 79. Categories of genes present in *M. abscessus* but absent in *M. chelonae* from the gene annotations.

Many of the gene descriptions are solely generic so transcriptional regulators may be defined as TetR or LysR but the genes they regulate are usually unknown. Often the transcriptional regulator will be clustered on the genome with an adjacent gene, also absent from *M. chelonae*, frequently hypothetical.

The reference gene sequences from PPanGGOLiN were submitted to Diamond + Megan and the SEED classification determined. Given the significant evolutionary difference between *M. abscessus* and *M. chelonae* the differences between these two species are not likely to be restricted to differences in antibiotic resistance and virulence. The SEED classification shows the major category, consistent with the gene annotations, is unassigned. Nevertheless, the assigned categories are informative.

An example of a difference, probably not relevant to resistance or virulence is Galactose Utilisation, midway down the SEED classification figure (Figure 80). Near the top, biosynthesis of arabinogalactans, which may change the cell wall and surface structure may affect uptake

of antibiotics and immunomodulation, influencing survival and virulence (Faller *et al.,* 2004) . Similarly: outer membrane porins (de Moura *et al*., 2021); biofilm formation (Szomolay *et al*., 2005; Dokic *et al*., 2021); putrescine utilisation and polyamine metabolism (El-Halfawy & Valvano, 2014; Tkachenko *et al.,* 2012); folate biosynthesis (Morgan *et al.,* 2018); heme acquisition (Choby & Skaar, 2016) and siderophores (in several categories) (Khasheii *et al*., 2021; Ribeiro & Simões, 2019); glycerolipid metabolism (Yu *et al*., 2019); cholesterol catabolism (Abuhammad, 2017); tRNA aminoacylation (Fields & Roy, 2018); para-aminosalicylic acid resistance; and protection against reactive oxygen species (Ma *et al*., 2016); all seem categories with potential links to antibiotic resistance mechanisms or improved survival in the host environment, which are present in these differential genes.

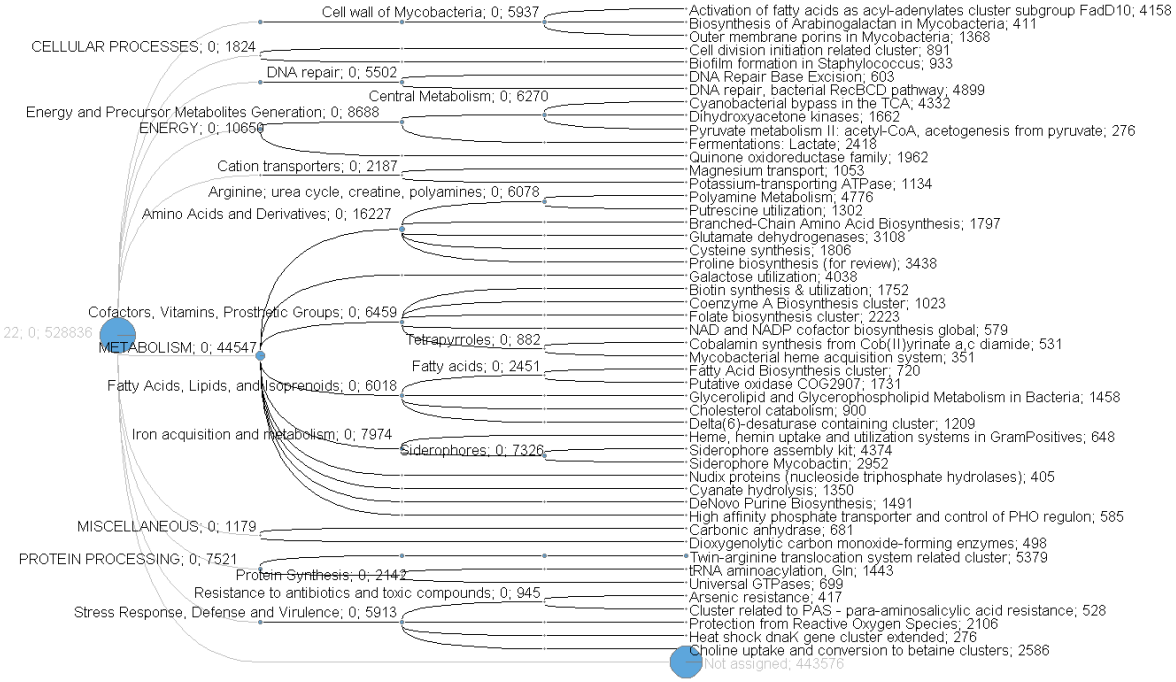


Figure 80. SEED classification of genes present in *M. abscessus* genomes and absent in *M. chelonae* genomes.

Chapter 5. Conclusions and Potential for Future Studies

Systematists will always strive for taxonomic exactitude when defining a species or an individual organism, but in many clinical situations the precise identification of a bacterium causing an infection is less important than its antibiogram. This is understandable since the predominant imperative for clinicians is the health and well-being of their patient. Thus, the antibiotic susceptibility of a bacterial strain will always take precedence over any taxonomic niceties. The clinical objective is to affect a remission of symptoms and wherever possible a cure. This may not always be true; frequent occurrences of a particular infecting organism in certain definable situations can be significant (Chiappini *et al*., 2021). Thus, epidemiological markers may be needed to follow and, if possible, intercede in transmission events. Similarly, chemotherapeutic markers may come to define a strain. There are several examples of this, such as MRSA (methicillin resistant *Staphylococcus aureus*); and MDR-TB (multi-drug resistant tuberculosis). However, accurate speciation can have implications in clinical prognosis. The long-term clinical management of progressive fibrocystic disease may be profoundly affected by species assigned to the so-called *Mycobacterium abscessus*/*chelonae* clade. Differentiation of *Mycobacterium abscessus* from *Mycobacterium chelonae* can have considerable significance in these circumstances. Empirical clinical experience suggests that when lung damage has reached the point where transplantation is the only recourse, the ensuing prognosis is poorer if the patient is colonised with a variant of *M. abscessus* rather than *M. chelonae* (Orens *et al.,* 2006).

The initial objective of this study was to select a clinical isolate of *Mycobacterium chelonae* and to derive a whole genome sequence (WGS). The sequencing of *M. chelonae* HPA006 with early adoption sequencing technologies (both Ion Torrent and Nanopore) meant higher errors and, in the case of Nanopore MinION, lower sequence coverage than would have been achieved currently. Subsequently, access to these sequencing resources was limited. Assembly of a complete genome was challenging and depended on the ongoing improvements in software. Nevertheless, subsequent ANI and pangenome analyses have given confidence that the *M. chelonae* HPA006 genome is robust and representative of clinically significant *M. chelonae* isolates.

There have been several recent publications which have assessed genome assemblers both for short and long read assemblies. A study carried out by Wick and Holt, 2021 reported on the benchmarking of current popular long-read assemblers (which included two used in this study, namely Canu and Flye, the remaining assemblers being Miniasm/Minipolish, NextDenovo/NextPolish and Raven) on various prokaryotic genomes. The authors simulated 500 long-read datasets to reflect various genomic features such as repeat length as well as several key sequencing parameters such as sequencing depth and read length.

Concluding that each of the different assemblers had pros and cons and that no single assembler emerged as the perfect choice for prokaryote genome long-read assembly. Flye was found to be reliable and was the best performing assembler at low read depths making the fewest large-scale sequence errors. Canu was found to be the most configurable with hundreds of adjustable parameters. Canu also created corrected and trimmed reads in its pipeline which had low error rates and which were, for the best part, free of adapters and chimeric sequences. As such Canu can considered as both an assembler and a long-read correction tool as used in this study.

Neubert *et al*., 2021 also assessed assembly strategies for short-read, long-read, and hybrid assemblers in genome studies of *Francisella tularensis*. They assessed the correctness, contiguity and overall completeness of the assemblies. Short-read sequences of *F. tularensis* were generated by MiSeq, HiSeq, and Ion Torrent sequencing technologies, these were then evaluated with eight de novo assembly tools focussing on short-read sequencing data; ABySS, A5-miseq, IDBA, MaSuRCA, MIRA, SGA, SPAdes, Tadpole and VelvetOptimiser. A5-miseq did not apply to Ion Torrent data. Neubert *et al*., concluded that optimal results were obtained by using assembly platforms which are adapted to the characteristics of the sequence platform producing the reads, and that the MIRA assembler performed best on Ion Torrent data. SPAdes gave good contiguity of the assembled *F. tularensis* genome with very few assembly errors. The study also assessed hybrid assemblies for each combination of long reads (PacBio, ONT) and short reads (HiSeq, MiSeq, Ion Torrent) using Canu/Pilon, Flye/Pilon, SPAdes, and Unicycler. SPAdes and Unicycler were found to be less prone to sequence errors such as mismatches, and indels.

Overall, in the choice of assembly tools for this study (Flye, Canu, Spades, MIRA and Unicycler) were appropriate choices.

A recent review by Kerkhof 2021 acknowledged that since its introduction in 2014 the MinION has made major improvements in both read quantity and accuracy. Oxford Nanopore Technologies have specifically addressed and improved sequencing chemistry, pore design and algorithms for base calling. Additionally, there is now a quantifiable and predictive signal from the MinION with respect to target molecule abundance and a simplified graphical user interface-based pathways for data analysis (Kerkhof, 2021).

The central aim of this study was to consider if the poorer prognosis for cystic fibrosis sufferers colonised with *M. abscessus* (in contrast to those colonised with *M. chelonae*) could be explained by identifying differences in the annotated genomes of these two species which could support this putative variation in virulence.

Additionally, these studies aimed to clarify the confused picture presented by the species covered by the umbrella term *Mycobacterium abscessus*/*chelonae* clade. Since it is now clear that there are three subspecies of *M. abscessus* (*M. abscessus* subsp. *abscessus*, *M. abscessus* subsp. *bolletii* and *M. abscessus* subsp. *massiliense),* there may be variation in the clinical impact of each of these. It is also apparent that several species have been validly named on the basis of phenotypic similarity to *M. chelonae* but differentiated due to somewhat limited variations in housekeeping genes.

The whole genome sequence of *M. chelonae* HPA 006 was compared to the deposited sequence of the *M. abscessus* Type strain (ATCC 19977^{T}) using a series of functional annotation and taxonomic tools. Genome annotation was performed with PGAP (Tatusova *et al*., 2016) and Diamond + Megan (Metagenome Analyser) (Bagci *et. al.,* 2021). Megan 6 was used to classify the assigned genes by taxonomy and functional classification using Interpro2GO, EggNOG, SEED, KEGG and EC.

The annotated sequence data for both the *M. abscessus* Type strain (ATCC 19977^{T}) and *M. chelonae* HPA 006 strains was also submitted to the antiSMASH database (Medema *et al.,*

2011; Blin *et al.,* 2021), to identify the gene clusters encoding secondary metabolites of all known broad chemical classes.

The differences seen in the KEGG (Kyoto Encyclopedia of Genes and Genomes, (Kanehisa, 2000)) profile of *M. abscessus*^{T} and *M. chelonae* HPA006 were minimal, there was also a high level of unassigned genes. Indeed, even in those genes which were assigned a KEGG category, the category was assigned in a very general way, a feature which made it difficult to ascribe such differences as the potential source of enhanced virulence or increased resistance in *M. abscessus* compared to *M. chelonae*.

The most interesting category of genes which differed between and *M. abscessus*^{T} and *M. chelonae* HPA006 were those associated with secondary metabolism. The study identified 19 BGCs in *M. abscessus* ^{T} and 15 BGCs in *M. chelonae* HPA006. 5 BGCs in *M.* abscessus^{T} were not present in *M. chelonae* HPA 006, and 2 BGCs present in HPA006 not present in *M. abscessus*^{T}. Secondary metabolite gene clusters are part of a mobile pan-genome in the species of the *M. abscessus*/*M. chelonae* clade. The antiSMASH BGCs are, intentionally, much larger than the core BGC (to ensure capturing all relevant genes even in poorly identified BGCs) and seem associated with regions of variation and regions associated with virulence and resistance. However, differences seem to be variation more than presence/absence and shuffling of genomic regions makes comparison difficult.

The *whi*B-like genes at MAB_1756 and MAB_4343c are candidates for differential resistance between *M. abscessus* strains as well as *M. chelonae* (Burian *et al*., 2012). There is also a *whi*B present in *M. chelonae* HPA 006 but absent from *M. abscessus*^{T} (HPA 006_004489).

A putative isonitrile lipopeptide is present in both *M. abscessus* and *M. chelonae* strains, but with some significant differences in sequence.

Of the 16 genes identified in Gorzynski *et al*., (2021) affecting resistance to the clinically significant antibiotics amikacin, clarithromycin, and cefoxitin 7 were identified as present in all *M. abscessus* but absent in *M. chelonae* (2 on the basis of low similarity), shown in **bold** in Table 19, and are candidates for contributing to the increased resistance of *M. abscessus*.

MAB_1137c and *MAB_4036 are not present in all M. abscessus strains and might be responsible for some variation in resistance between M. abscessus strains and the subspecies (M. abscessus subsp. bolletii and M. abscessus subsp. massiliense).*

In summary, the pangenome study using PPanGGOLiN (Gautreau *et al*., 2020) identified 655 genes present in *M. abscessus* (ATCC 19977) but absent from *M. chelonae* strains. antiSMASH analysis highlighted differences in the overall number of Biosynthetic Gene Clusters (BGC) present in each species and identified BGCs present in one species but absent in the other.
Of the 655 differential genes identified the largest category was hypothetical, including proteins with domains of unknown function. Despite annotation other genes proved challenging to correlate with known mechanisms of resistance or virulence.

Clarification of the taxonomy of the organisms assigned to the *M. abscessus*/*chelonae* clade was carried out by performing and Average Nucleotide Analysis (ANI) with Pyani (Pritchard *et al.*, 2016). It was also an opportunity to examine the relatedness of the species in the *M. chelonae* clade and the *M. abscessus* clade

ANI values of > 95% are commonly used as a tool to measure prokaryotic species delineation (Goris *et al*., 2007; Richter and Rosselló-Móra, 2009), furthermore, if supported by other phenotypic and epidemiological values between 95 and 98% reasonably infer subspecies status (Turenne, 2019). To date ANI values cannot be used as a genomic standard for prokaryotic genus delineation, and more than 60% of interspecies ANI values within a genus are around 68–72% (Qin *et al*., 2014).

The *M. chelonae* strains studied form a relatively compact cluster of closely related strains (see Figures 44 and 45). Nevertheless, the type strain (ATCC 35752; NCTC 946), isolated from turtle tubercle, is well recognised as not representative of the majority of isolates assigned to this species, probably a prime candidate as the nucleus of a distinct subspecies. There are also some strains which appear even more distinct than the type strain but there is no clear break in the chain of relatedness, nor clustering of strains to justify grouping them together.

Two subspecies have been described, *M. chelonae* subsp. *gwanakae* (Kim *et* al., 2018) and *M. chelonae* subsp. *bovis* (Kim *et al*., 2017) for which there is no whole genome sequence deposited. Based on limited sequence data, *M. chelonae* subsp. *bovis* may be related to the strains indicated in the Minimum spanning tree (Figure 45). Neither the ANI data nor the whole genome relationships based on presence/absence of genes (see Figure 45 would support these 2 subspecies designations and they are not designated as such in the proposed new genus *Mycobacteroides* (Gupta *et al*., 2018).

The ANI data of the *M. abscessus* subsp *abscessus, M. abscessus* subsp *massiliense* and *M. abscessus* subsp *bolletii* strains studied show all 3 variants to be very closely related but very distinct from all the other species and, in particular, from *M. chelonae*. However, if 95% ANI is accepted as an appropriate cut off for species and 95-98% as the range for the description of subspecies (Figure 39), this delineation into separate subspecies holds up well.

This study has demonstrated that separation of the species listed as part of the *M. abscessus/chelonae* clade may be difficult using phenotypic methods and indeed with more commonly applied genomic analyses such as 16S rRNA gene comparisons. However, the ANI analyses demonstrate that all species listed as part of the *M. abscessus*/*chelonae* clade are indeed separate species. There is evidence however that *M. stephanolepidis* aligns better with *M. salmoniphilum*.

ANI analyses of *M. abscessus* "variants" has allowed a supportive contribution to the taxonomy of the subspecies of this organism. These taxonomic issues are undoubtedly linked to an improved understanding of virulence and drug resistance. These studies have demonstrated that average nucleotide identities (ANI) derived from deposited whole genome sequences and analysed using R and RStudio allow a more definitive classification of these species

Standard empirical identification based on phenotypic characters and rapid assessment of some genetic markers via PCR techniques, is currently supported by sequencing of housekeeping genes (notably 16S rRNA sequence data). These approaches need to be more actively supported by emergent genetic techniques, particularly when species or strain identities are of greater significance. These techniques may be complex at present but these

studies suggest that new software and easier access to dedicated IT facilities can provide this new approach to determining species designation. Such approaches also raise new questions. Does the empirical belief that *M. abscessus* (sensu lato) suggest a poorer prognosis in fibrocystic disease hold true for all the subspecies now identified? Similarly, in this context, does this subspecies variation have any bearing on clinical significance? The ability to reliably classify these subspecies needs to be allied to research on clinical occurrence and outcome. To determine whether there is such variation a progressive investigation with taxonomic characterisation of strains allied to clinical findings would be necessary.

Allied to the taxonomic discussions is the finding that there are differences in gene content between *M. abscessus* (and variants) and *M. chelonae*. Do these gene variations explain subsequent variations in virulence? Additionally, a further question would be, do these genes confer enhanced resistance to chemotherapeutics? Side by side with this is the possibility that they suggest potential targets for development of new antibiotics. In an era of gene therapy is it possible to reduce/limit the virulence of the infecting organism, not by eradication but by inhibiting the gene expression. Once again, these questions cannot easily be answered without marrying detailed strain assessment and subsequent clinical occurrence and outcome. In terms of future studies there are many potential areas of supplementary research.

Perhaps most valuable would be further characterisation of those members of the Pro-Glu and Pro-Pro-Glu (PPE) family of proteins which may have significance in organism resistance and virulence. Equally important would be utilising these approaches alongside a dedicated clinical interface.

The rapid development of DNA sequencing and the ease, relative cheapness and portability of nanopore sequencing suggest that identification of these strains in a clinical setting by nanopore sequencing would be fast and accurate. It is also not dependent upon prior identification of the causative organism and would identify unexpected, and unknown, organisms in clinical samples generically. There could also be consideration of a nanopore based sequencing service using the MinION platform to provide a respiratory metagenomic service to identify other respiratory targets regularly implicated in causing infection in cystic fibrosis patients.

References

Aardenne-Ehrenfest, van, T., & Bruijn, de, N. G. (1951). Circuits and trees in oriented linear graphs. Simon Stevin : Wis- en Natuurkundig Tijdschrift. Volume 28: 203-217.

Abdallah, A.M., Gey van Pittius, N.C., Champion, P.A., Cox, J., Luirink, J., Vandenbroucke-Grauls, C.M., Appelmelk B.J. and Bitter W. (2007). Type VII secretion--mycobacteria show the way. Nature reviews. Microbiology. *5* (11): 883–891.

Abuhammad, A. (2017). Cholesterol metabolism: a potential therapeutic target in Mycobacteria. British journal of pharmacology. *174* (14): 2194–2208.

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

Adékambi, T., Stein, A., Carvajal, J., Raoult, D. and Drancourt, M. (2006a). Description of *Mycobacterium conceptionense* sp. nov., a *Mycobacterium fortuitum* group organism isolated from a posttraumatic osteitis inflammation. J Clin Microbiol *44*: 1268-1273.

Adékambi, T., Berger, P., Raoult, D. and Drancourt, M. (2006b). *rpoB* gene sequence-based characterisation 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 *56*: 133-143.

Adékambi, T., Gaubert, M.R., Greub, G., Gevaudan, M.J., La Scola, B., Raoult, D. and Drancourt, M. (2006c). *In*: List of new names and new combinations previously effectively, but not validly, published. Validation List no. 111. Int J Syst Evol Microbiol *56*: 2025-2027.

Adjemian, J., Olivier, K.N. and Prevots, R. (2104). Non-tuberculous Mycobacteria among Patients with Cystic Fibrosis in the United States Screening Practices and Environmental Risk. Am J Respir Crit Care Med. *190* (5): 581-586.

Akeson, M., Branton, D., Kasianowicz, J.J., Brandin, E. and Deamer, D.W. (1999). Microsecond time-scale discrimination among polycytidylic acid, polyadenylic acid and polyuridylic acid as homopolymers or as segments within single RNA molecules. Biophys J. *77*(6):3227-33.

Alsarraf, H. M. A. B., Ung, K. L., Johansen, M. D., Dimon, J., Olieric, V., Kremer, L. and Blaise, M. (2022). Biochemical, structural, and functional studies reveal that MAB_4324c from *Mycobacterium abscessus* is an active tandem repeat N-acetyltransferase. FEBS letters. *596*, (12):1516–1532.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990). Basic local alignment search tool. J Mol Biol. *215* (3):403-410.

Anand, A., Verma, P., Singh, A. K., Kaushik, S., Pandey, R., Shi, C., Kaur, H., Chawla, M., Elechalawar, C. K., Kumar, D., Yang, Y., Bhavesh, N. S., Banerjee, R., Dash, D., Singh, A., Natarajan, V. T., Ojha, A. K., Aldrich, C. C. and Gokhale, R. S. (2015). Polyketide Quinones Are Alternate Intermediate Electron Carriers during Mycobacterial Respiration in Oxygen-Deficient Niches. Molecular cell. *60* (4): 637–650.

Ananta, P., Kham-Ngam, I., Chetchotisakd, P., Chaimanee, P., Reechaipichitkul, W., Namwat, W., Lulitanond, V. and Faksri, K. (2018). Analysis of drug-susceptibility patterns and gene sequences associated with clarithromycin and amikacin resistance in serial Mycobacterium abscessus isolates from clinical specimens from Northeast Thailand. PloS one. *13* (11): e0208053.

Anderl, J. N., Franklin, M.J. and Stewart, P.S. (2000). Role of antibiotic penetration limitation in *Klebsiella pneumoniae* biofilm resistance to ampicillin and ciprofloxacin. Antimicrobial agents and chemotherapy. *44* (7): 1818–1824.

Andersson, M.I. and MacGowan, A.P. (2003). Development of the quinolones. J Antimicrob Chemother. *51* (Suppl. S1): 1–11.

Armstrong, D.T. and Parrish, N. (2021). Current Updates on Mycobacterial Taxonomy, 2018 to 2019. J Clin Microbiol. *59* (7).

Arnold, C., Barrett, A., Cross, L. and Magee, J.G. (2012). The use of *rpo*B sequence analysis in the differentiation of *Mycobacterium abscessus* and *Mycobacterium chelonae*: a critical judgement in cystic fibrosis. Clin Microbiol Inf *18*(5): E131-E133.

Arumugam, K., Bagci, C., Bessarab, I., Beier, S., Buchfink, B., Górska, A., Qiu, G., Huson, D.H. and Williams, R.B.H. (2019). Annotated bacterial chromosomes from frame-shift-corrected long-read metagenomic data. Microbiome. *7* : 61.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M. and Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. *25* (1):25-29.

Athanasopoulou, K., Boti, M.A., Adamopoulos, P.G., Skourou, P.C. and Scorilas, A. (2022) Third-Generation Sequencing: The Spearhead towards the Radical Transformation of Modern Genomics. Life. *12*, 30.

Aubry, A., Veziris, N., Cambau, E., Truffot-Pernot, C., Jarlier, V. and Fisher, L.M. (2006). Novel gyrase mutations in quinolone-resistant and hypersusceptible clinical isolates of *Mycobacterium tuberculosis*: Functional analysis of mutant enzymes. Antimicrob agents Chemother *50* (1): 104-112.

Ayikpoe, R., Govindarajan, V. and Latham, J.A. (2019). Occurrence, function, and biosynthesis of mycofactocin. Appl Microbiol Biotechnol. *103***,** 2903–2912

Aziz, R.K., Bartels, D., Best, A.A., DeJongh, M., Disz, T., Edwards, R.A., Formsma, K., Gerdes, S., Glass, E.M., Kubal, M., Meyer, F., Olsen, G.J., Olson, R., Osterman, A.L., Overbeek, R.A., McNeil, L.K., Paarmann, D., Paczian, T., Parrello, B., Pusch, G.D., Reich, C., Stevens, R., Vassieva, O., Vonstein, V., Wilke, A. and Zagnitko, O. (2008). The RAST Server: Rapid Annotations using Subsystems Technology. BMC Genomics. *9*: 75.

Babaki Zadeh Karbalaei, M., Soleimanpour, S. and Rezaee, S.A. (2017). Antigen 85 complex as a powerful Mycobacterium tuberculosis immunogene: Biology, immune-pathogenicity, applications in diagnosis, and vaccine design. Microbial Pathogenesis. *112*: 20-29.

Baess, I. (1982). Deoxyribonucleic acid relatedness among species of rapidly-growing mycobacteria. Acta Pathol Microbiol Scand. *90*: 371-375.

Bagci, C., Patz, S. and Huson, D.H. (2021). DIAMOND+MEGAN: fast and easy taxonomic and functional analysis of short and long microbiome sequences. Current Protocols. *1*: e59

Baltz, R.H. (2009) Daptomycin: mechanisms of action and resistance, and biosynthetic engineering. Current Opinion in Chemical Biology. *13* (2):144-151.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G. Alekseyev, M. A. and Pevzner, P. A. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. Journal of computational biology : a journal of computational molecular cell biology. *19* (5): 455–477.

Barrett, A. J. (1992). Enzyme nomenclature. recommendations 1992.European Journal of Biochemistry. *232* (1).

Bastian S., Veziris, N., Roux, A-L., Brossier, F., Gaillard, J-L., Jarlier, V. and Cambau, E. (2011). Assessment of clarithromycin susceptibility in strains belonging to the *Mycobacterium abscessus* group by *erm*(41) and *rrl* sequencing. Antimicrob Agents Chemother *55*: 775-781.

Baumdicker, F., Hess, W.R. and Pfaffelhuber, P. (2012). The infinitely many genes model for the distributed genome of bacteria. Genome Biol Evol *4*: 443-456.

Baysarowich, J., Koteva, K., Hughes, D.W., Ejim, L., Griffiths, E., Zhang, K., Junop, M. and Wright G.D. (2008). Rifamycin antibiotic resistance by ADP-ribosylation: Structure and diversity of Arr PNAS *105*: 4886-4891.

Bazin, A., Gautreau, G., Médigue, C., Vallenet, D. and Calteau, A. (2020). panRGP: a pangenome-based method to predict genomic islands and explore their diversity. Bioinformatics. *36* (Suppl 2):i651-i658.

Beceiro, A., Tomás, M. and Bou, G. (2013). Antimicrobial resistance and virulence: a successful or deleterious association in the bacterial world?. Clinical microbiology reviews. *26* (2): 185–230.

Beckham, K.S.H., Ciccarelli, L., Bunduc, C.M., Mertens, H.D.T., Ummels, R., Lugmayr, W., Mayr, J., Rettel, M., Savitski, M.M., Svergun, D.I., Bitter, W., Wilmanns, M., Marlovits, T.C., Parret, A.H.A. and Houben, E.N.G. (2017). Structure of the mycobacterial ESX-5 type VII secretion system membrane complex by single-particle analysis. *Nature Microbiology*. *2* (6): 17047.

Belardinelli, J.M., Li, W., Avanzi, C., Angala, S.K., Lian, E., Wiersma, C.J., Palčeková, Z., Martin, K.H., Angala, B., Moura, de V., Kerns, C., Jones, V., Gonzalez-Juarrero, M., Davidson, R.M., Nick, J.A., Borlee, B.R. and Jackson, M. (2021). Unique Features of *Mycobacterium abscessus* Biofilms Formed in Synthetic Cystic Fibrosis Medium. Frontiers in Microbiology. *12*: 743126.

Berger-Bächi, B. (2002). Resistance mechanisms of Gram-positive bacteria. International Journal of Medical Microbiology. *292* (1): 27-35.

Bergey, D.H., Harrison, F.C., Breed, R.S., Hammer, B.W. and Huntoon, F.M. (1923). Bergey's Manual of Determinative Bacteriology, 1st edn. Williams & Wilkins, Baltimore.

Bernut, A., Dupont, C., Ogryzko, N.V., Neyret, A., Herrmann, J.-L., Floto, R.A., Renshaw, S.A. and Kremer, L. (2019). CFTR Protects against Mycobacterium abscessus Infection by Fine-Tuning Host Oxidative Defences. Cell Reports. *26*: 1828–1840.

Bernut A., Herrmann, J-L., Kissa, K., Dubremetz, J-F., Gaillard, J-L., Lutfalla, G. and Kremer, L. (2014). Mycobacterium abscessus cording prevents phagocytosis and promotes abscess formation. Proc Natl Acad Sci U S A. *111* (10): E943–E952.

Bernut, A., Herrmann, J-L., Ordway, D. and Kremer, L. (2017). The Diverse Cellular and Animal Models to Decipher the Physiopathological Traits of Mycobacterium abscessus infection. Frontiers in cellular and infection microbiology. *7*:100.

Bernut, A., Nguyen-Chi, M., Halloum, I., Herrmann, J-L., Lutfalla, G. and Kremer, L. (2016). Mycobacterium abscessus-induced granuloma formation is strictly dependent on TNF signalling and neutrophil trafficking. PLoS Pathog. *12* (11): e1005986

Bhakta, S., Besra, G. S., Upton, A. M., Parish, T., Sholto-Douglas-Vernon, C., Gibson, K. J., Knutton, S., Gordon, S., DaSilva, R. P., Anderton, M. C. and Sim, E. (2004). Arylamine N-acetyltransferase is required for synthesis of mycolic acids and complex lipids in Mycobacterium bovis BCG and represents a novel drug target. The Journal of Experimental Medicine. 199 (9), 1191–1199.

Basic Local Alignment Search Tool (BLAST) https://blast.ncbi.nlm.nih.gov

Blin, K., Shaw, S., Kloosterman, A.M., Charlop-Powers, Z., van Weezel, G.P., Medema, M.H. and Tilmann, W. (2021). antiSMASH 6.0: improving cluster detection and comparison capabilities. Nucleic Acids. *49*: (W1):W29-W35.

Bobay, L.M. and Ochman, H. (2017). Biological species are universal across Life's domains. Genome Biol Evol. *9* (3): 491–501.

Boisvert, H. (1977). L'ulcere cutane a *Mycobacterium ulcerans* au Cameroun. II. Étude bacteriologique. Bull Soc Pathol Exot *70*: 125-131.

Bolotin, E and Hershberg, R. (2015). Gene Loss Dominates As a Source of Genetic Variation within Clonal Pathogenic Bacterial Species. Genome Biol Evol. *7* (8):2173–2187.

Bolotin, E and Hershberg, R. (2017). Horizontally Acquired Genes Are Often Shared between Closely Related Bacterial Species. Front Microbiol. *8*:1536.

Bos, K. I., Harkins, K. M., Herbig, A., Coscolla, M., Weber, N., Comas, I., Forrest, S. A., Bryant, J. M., Harris, S.R., Schuenemann, V. J., Campbell, T. J., Majander, K., Wilbur, A. K., Guichon, R.A., Wolfe Steadman, D. L., Collins, Cook D., Niemann, S., Behr, M. A., Zumarraga, M., Bastida, R., Huson D., Nieselt, K., Young, D., Parkhill, J., Buikstra, J. E., Gagneux, S., Stone, A. C. and Krause J. Pre-Columbian mycobacterial genomes reveal seals as a source of New World human tuberculosis. (2014). Nature. *514* (7523): 494–497.

Bragg, L. M., Stone, G., Butler, M. K., Hugenholtz, P. and Tyson, G. W.. (2013). Shining a light on dark sequencing: characterising errors in Ion Torrent PGM data. PLoS computational biology. *9* (4).

Braslavsky, I., Hébert, B., Kartalov, E.P. and Quake, S.R. (2003). Sequence information can be obtained from single DNA molecules. Proceedings of the National Academy of Sciences of the United States of America. *100*: 3960 - 3964.

British Thoracic Society. (1984). A controlled trial of 6 months chemotherapy in pulmonary tuberculosis, final report: results during the 36 months after the end of chemotherapy and beyond. Br J Dis Chest *78* (4): 330-336.

Note: in 1982 The British Thoracic Association merged with the Thoracic Society to form the British Thoracic Society.

British Thoracic Society. (1994). *Mycobacterium kansasii* pulmonary infection: a prospective study of the results of 9 months of treatment with rifampicin and ethambutol. Thorax *49*: 442-445.

Brodin, P., Eiglmeier, K., Marmiesse, M., Billault, A., Garnier, T., Niemann, S., Cole, S.T. and Brosch, R. (2002). Bacterial Artificial Chromosome-Based Comparative Genomic Analysis Identifies *Mycobacterium microti* as a Natural ESAT-6 Deletion Mutant. *Infection and Immunity. 70* (10): 5568-5578.

Brown-Elliott, B. A., Vasireddy, S., Vasireddy, R., Iakhiaeva, E., Howard, S.T., Nash, K., Parodi, N., Strong, A., Gee, M., Smith, T. and Wallace, R.J. Jr. (2015). Utility of sequencing the *erm*(41) gene in isolates of Mycobacterium abscessus subsp. abscessus with low and intermediate clarithromycin MICs. Journal of Clinical Microbiology. *53*(4): 1211–1215.

Bryant, J.M., Grogono, D.M., Rodriguez-Rincon, D., Everall, I., Brown, K.P., Moreno, P., Verma, D., Hill, E., Drijkoningen J. (72 further authors who collaborated in this study). (2016). Emergence and spread of a human-transmissible multidrug-resistant nontuberculous mycobacterium. Science. *354* (6313): 751-757.

Buchfink, B., Xie, C. and Huson, D. (2015). Fast and sensitive protein alignment using DIAMOND. Nat Methods. *12* : 59-60.

Bunduc, C. M., Fahrenkamp, D., Wald, J., Ummels, R., Bitter, W., Houben, E. and Marlovits, T. C. (2021). Structure and dynamics of a mycobacterial type VII secretion system. Nature. *593* (7859), 445–448.

Bush, K. and Jacoby, G.A. (2010). Updated functional classification of beta-lactamases. Antimicrobial agents and chemotherapy. *54* (3): 969–976.

Camon, E., Barrell, D., Dimmer, E., Lee, V., Magrane, M., Maslen, J., Binns, D. and Apweiler, R. An evaluation of GO annotation retrieval for BioCreAtIvE and GOA. (2005). *BMC Bioinformatics*. *6*: (Suppl 1): S17.

Campbell, I.A. and Jenkins, P.A. (1995). Opportunist mycobacterial infections. In: Brewis, R.A.L., Corrin, B. Geddes, D.M. Gibson, G.J., eds. Respiratory medicine. London: W.B. Saunders.

Canavez F.C., Luche, D.D., Stothard, P., Leite, K. R. M., Sousacanavez, J. M., Plastow, G., Meidanis, J., Souza, M. A., Feijao, P., Moore, S. S. and Camara-Lopes, L. H. (2012). Genome sequence and assembly of Bos Indicus. J Heredity. *103* (3): 342-348.

Canessa, C.M., A.M. Merillat and B.C. Rossier. 1994. Membrane topology of the epithelial sodium channel in intact cells. Am J Physiol 267 (6): 1682-90.

CANU Long read assembly pipeline: https://github.com/marbl/canu/releases

Caputo, A., Fournier, P.E. and Raoult, D. (2019). Genome and pan-genome analysis to classify

emerging bacteria. Biology Direct *14* (1): 5.

Chamoiseau, G. (1979) Etiology of farcy in African bovines: nomenclature of the causal organisms *Mycobacterium farcinogenes* Chamoiseau and *Mycobacterium senegalense* (Chamoiseau) comb. nov. Int. J. Syst. Bacteriol. *29*:407-410.

Chan, A.P., Sutton, G., DePew, J., Krishnakumar, R., Choi, Y., Huang ,X.Z., Beck, E., Harkins, D.M., Kim, M., Lesho, E.P., Nikolich, M.P. and Fouts, D.E. (2015). A novel method of consensus pan-chromosome assembly and large-scale comparative analysis reveal the highly flexible pan-genome of Acinetobacter baumannii. Genome biology. *16* (1):143.

Chand, M., Lamagni, T., Kranzer, K., Hedge, J., Moore, G., Parks, S., Collins, S., Del Ojo Elias, C., Ahmed, N., Brown, T., Smith, E. G., Hoffman, P., Kirwan, P., Mason, B., Smith-Palmer, A., Veal, P., Lalor, M. K., Bennett, A., Walker, J., Yeap, A., Martin, A.I.C., Dolan, G., Bhatt, S., Skingsley, A., Charlett, A., Pearce, D., Russell, K., Kendall, S., Klein, A. A., Robins, S., Schelenz, S., Newsholme, W., Thomas, S., Collyns, T., Davies, E., McMenamin, J., Doherty, L., Peto, T E.A., Crook, D., Zambon, M. and Phin, N. (2017). Insidious Risk of Severe *Mycobacterium chimaera* Infection in Cardiac Surgery Patients. Clinical infectious diseases : an official publication of the Infectious Diseases Society of America. *64* (3), 335–342.

Chaumeil, P.A., Mussig, A.J., Hugenholtz, P. and Parks, D.H. (2019). GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. Bioinformatics. *36* (6):1925–19277.

Chemler, J. A., Buchholz, T. J., Geders, T. W., Akey, D. L., Rath, C. M., Chlipala, G. E., Smith, J. L. and Sherman, D. H. (2012). Biochemical and structural characterization of germicidin synthase: analysis of a type III polyketide synthase that employs acyl-ACP as a starter unit donor. Journal of the American Chemical Society. *134* (17): 7359–7366.

Chen, Y,C., Liu, T., Yu, C-H, Chiang, T-Y. and Hwang, C-C. (2013). Effects of GC bias in Next Generation Sequence Data on De Novo Genome Assembly. PLoS ONE. *8* (4):e62856

Chen, Z., Erickson, D.E. and Meng, J. (2020). Benchmarking hybrid assembly approaches for genomic analyses of bacterial pathogens using Illumina and Oxford Nanopore sequencing. BMC Genomics. *21*:631

Cheng, V.C., Yam, W.C., Hung, I.F., Woo, P.C., Lau, S.K., Tang, B.S. and Yuen, K.Y. (2004). Clinical evaluation of the polymerase chain reaction for the rapid diagnosis of tuberculosis. Journal of Clinical Pathology. *57* (3): 281–285.

Chevreux, B (2005) MIRA: An Automated Genome and EST Assembler. Ph.D. Thesis. The Ruprecht-Karls-University. http://sourceforge.net/projects/MIRA-assembler.

Chiappini, E., Santamaria, F., Marseglia, G.L., Marchisio, P., Galli, L., Cutrera, R., de Martino, M., Antonini, S., Becherucci, P., Biasci, P., Bortone, B., Bottero, S., Caldarelli, V., Cardinale, F., Gattinara, G.C., Ciarcià, M., Ciofi, D., D'Elios, S., Di Mauro, G., Doria, M., Indinnimeo, L., Lo Vecchio, A., Macrì, F., Mattina, R., Miniello, V.L., Del Giudice, M.M., Morbin, G., Motisi, M.A., Novelli, A., Palamara, A.T., Panatta, M.L., Pasinato, A., Peroni, D., Perruccio, K., Piacentini, G., Pifferi, M., Pignataro, L., Sitzia, E., Tersigni, C., Torretta, S., Trambusti, I., Trippella, G., Valentini, D., Valentini, S., Varricchio, A., Verga, M.C., Vicini, C., Zecca, M. and Villani, A. (2021) Prevention of recurrent respiratory infections : Inter-society Consensus. Italian Journal of Paediatrics. *47* (1): 211.

Chihota, V.N., Grant, A.D., Fielding, K., Ndibongo, B., van Zyl, A., Muirhead, D. and Churchyard, G.J. (2010). Liquid vs. solid culture for tuberculosis: performance and cost in a resource-constrained setting. Int J Tuberc Lung Dis. *14* (8): 1024–1031.

Choby, J. E. and Skaar, E. P. (2016). Heme Synthesis and Acquisition in Bacterial Pathogens. Journal of Molecular Biology. *428* (17): 3408–3428.

Choi, G-E., Shin, S.J., Won, C-J., Min, K-N., Oh, T., Hahn, M.Y., Lee, K., Daley, C.L., Kim, S., Jeong, B., Jeon, H-K. and Koh, W-J. (2012). Macrolide treatment for *Mycobacterium abscessus* and *Mycobacterium massiliense* infection and inducible resistance. Am J Respir Crit Care Med. *186:* 917–925.

Choi, H., Kim, S.Y., Lee, H., Jhun, B.W., Park, H.Y., Jeon, K., Kim, D.H., Huh, H.J., Ki, C.S., Lee, N.Y., Lee, S.H., Shin, S.J., Daley, C.L. and Koh, W.J. (2017). Clinical characteristics and treatment outcomes of patients with macrolide-resistant Mycobacterium massiliense lung disease. Antimicrobial Agents and Chemotherapy. *61* (2): e02189.

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

Chung, Y.J. and Saier, Jr M.H. (2002). Overexpression of the Escherichia coli sugE gene confers resistance to a narrow range of quaternary ammonium compounds. J Bacteriol. *184*(9):2543-2545.

Ciaramella, A., Martino, A., Cicconi, R., Colizzi, V. and Fraziano, M. (2000). Mycobacterial 19-kDa lipoprotein mediates Mycobacterium tuberculosis-induced apoptosis in monocytes/macrophages at early stages of infection. Cell Death Differ. *7* (12):1270-1272.

Clark, H.F. and Shepard, C.C. (1963). Effect of Environmental Temperatures on Infection with *Mycobacterium Marinum* (Balnei) of Mice and a Number of Poikilothermic Species. Journal of Bacteriology *86*: 1057 - 1069.

Clark, J., Wu, H-C., Jayasinghe, L., Patel, A., Reid, S. and Bayley, H. (2009). Continuous base identification for single molecule nanopore DNA sequencing. Nature Nanotechnology *4*: 265-270.

Clarridge, J.E. (2004. Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases. Clin. Microbiol. Rev. *17* (4): 840-862.

CLSI. Clinical and Laboratory Standards Institute. (2011). Performance Standards for Antimicrobial Susceptibility Testing. 21st Informational Supplement. CLSI Document M100-S21.

Cobbett L. (1918). An acid-fast bacillus obtained from a pustular eruption. British Medical Journal. *2* (3007), 158–159.

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

Cohen, S.N. and Chang, A.C.Y. (1973). Re-circularisation and autonomous replication of a sheared R-factor DNA segment in *Escherichia coli* transformants. Nat. Acad. Sc of USA *70*: 1293-1297.

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

Collins, F.S. and McKusick, V.A. (2001). Implications of the Human Genome Project for medical science. JAMA *285*:540–544.

Collins, F.S., Green, E.D., Guttmacher, A.E. and Guyer, M.S. (2003). A vision for the future of genomics research. Nature *422*: 835-847.

Collins, R.E. and Higgs, P.G. (2012). Testing the infinitely many genes model for the evolution of the bacterial core genome and pangenome. Mol Biol Evol. *29* (11):3413–3425.

Colwell, R.R. (1970). Polyphasic taxonomy of bacteria. In: Culture Collections of Microorganisms (edited by H. Iizuka & T. Hasegawa). University of Tokyo Press, Tokyo pp. 421-436.

Comas, I., Coscolla, M., Luo, T., Borrell ,S., Holt, K. E., Kato-Maeda, M., Parkhill, J., Malla, B., Berg, S., Thwaites, G., Yeboah-Manu, D., Bothamley, G., Mei, J., Wei, L., Bentley, S., Harris, S. R., Niemann, S., Diel, R., Aseffa, A., Gao, Q., Young, D. and Gagneux, S. (2013). Out-of-Africa

migration and Neolithic co-expansion of Mycobacterium tuberculosis with modern humans. Nature genetics. *45*(10):1176–1182.

Combrink, K.D., Denton, D.A., Harran, S., Ma, Z., Chapo, K., Yan, D., Bonventre, E., Roche, E.D., Doyle, T.B., Robertson, G.T. and Lynch, A.S. 2007. New C25 carbamate rifamycin derivatives are resistant to inactivation by ADP-ribosyl transferases. Bioorganic & Medicinal Chemistry Letters *17*: 522–526.

Connolly, M.J., Magee, J.G., Hendrick, D.J., and Jenkins, P.A. (1985). *Mycobacterium malmoense* in the North-east of England. Tubercle *66* (3): 211-217.

Cooksey, R.C., de Waard, J.H., Yakrus, M.A., Rivera, I., Chopite, M., Toney, S.R., Morlock, G.P. and Butler, W.R. (2004). *Mycobacterium cosmeticum* sp.nov., a novel rapidly growing species isolated from a cosmetic infection and from a nail salon. Int. J. Syst. Evol. Microbiol. *54*: 2385-2391.

Corliss, J.O. (1975). Three Centuries of Protozoology: A Brief Tribute to its Founding Father, A. van Leeuwenhoek of Delft. J Protozoology. *22 (1)*: 3-7.

Cortes, M.A., Nessar, R. and Singh, A.K. (2010). Laboratory maintenance of Mycobacterium abscessus. Current Protocols in Microbiology. Chapter 10. https://doi.org/10.1002/9780471729259.mc10d01s18

Costa, S.S., Guimarães, L.C., Silva, A., Soares, S.C. and Baraúna, R.A. (2020). First steps in the analysis of prokaryotic pan-genomes. Bioinformatics and Biology Insights *14*: 1-9.

Crofton, J.W. (1959). Chemotherapy of pulmonary tuberculosis. Br Med J 1: 1610-1614.

Daley, C. L., Iaccarino, J. M., Lange, C., Cambau, E., Wallace, R. J., Jr, Andrejak, C., Böttger, E. C., Brozek, J., Griffith, D. E., Guglielmetti, L., Huitt, G. A., Knight, S. L., Leitman, P., Marras, T. K., Olivier, K. N., Santin, M., Stout, J. E., Tortoli, E., van Ingen, J., Wagner, D. and Winthrop, K. L. (2020). Treatment of nontuberculous mycobacterial pulmonary disease: An official

ATS/ERS/ESCMID/IDSA clinical practice guideline. The European Respiratory Journal. *56* (1): 2000535.

Daniel-Wayman, S., Abate, G., Barber, D.L., Bermudez, L.E., Coler, R.N., Cynamon, M.H., Daley, C.L., Davidson, R.M., Dick, T., Floto, R.A., Henkle, E., Holland, S.M., Jackson, M., Lee, R.E., Nuermberger, E. L., Olivier, K.N., Ordway, D. J., Prevots, D. R., Sacchettini, J. C., Salfinger, M., Sassetti, C. M., Sizemore, C. F., Winthrop, K.L. and Zelazny, A. M. (2019). Advancing Translational Science for Pulmonary Nontuberculous Mycobacterial Infections. A Road Map for Research. Am. J. Respir. Crit. Care Med. *199*: 947–951.

Daubin, V., Gouy, M., and Perrière, G. (2001). Bacterial molecular phylogeny using supertree approach. Genome Inform *12*:155–164.

Dautzenberg, B., Piperno, D., Diot, P., Truffot-Pernot, C., Chauvin, J-P. and the clarithromycin study group of France. (1995). Clarithromycin in the treatment of *Mycobacterium avium* lung infections in patients without AIDS. Chest *107*: 1035-40.

Davies J.C., Alton, E.W.F.W. and Bush, A. (2007). Cystic Fibrosis. BMJ *335*: 1255-1259.

Davies, P.D.O. (1981). Drug-resistant tuberculosis. J R Soc Med *94* (6): 261-263.

Davis, J.J., Boisvert, S., Brettin, T., Kenyon, R.W., Mao, C., Olson, R., Overbeek, R., Santerre, J., Shukla, M., Wattam, A.R., Will, R., Xia, F. and Stevens, R. (2016). Antimicrobial Resistance Prediction in PATRIC and RAST. Sci Rep. *6*:27930.

DeBlois, R.W. and Bean, C.P. (1970). Counting and sizing of submicron particles by the resistive pulse technique. Rev Sci Instrum *41*: 909-916.

De Groote, M.A. and G. Huitt. (2006). Infections due to rapidly growing mycobacteria. Clin Infect Dis. *42* (12): 1756-1763.

de Moura V. C., da Silva, M.G., Gomes, K.M., Coelho, F.S., Sampaio, J.L., Mello, F.C., Lourenco, M.C., Amorim, Ede L. and Duarte, R.S. (2012). Phenotypic and molecular characterisation of

quinolone resistance in *Mycobacterium abscessus* subsp. *bolletii* recovered from postsurgical infections. J Med Microbiol. *61* (1): 115-125.

de Moura, V. C. N., Verma, D., Everall, I., Brown, K. P., Belardinelli, J. M., Shanley, C., Stapleton, M., Parkhill, J., Floto, R. A., Ordway, D. J. and Jackson, M. (2021). Increased Virulence of Outer Membrane Porin Mutants of Mycobacterium abscessus. Frontiers in microbiology. *12*, 706207.

Deamer, D.W. and Akeson, M. (2000). Nanopores and nucleic acids: prospects for ultrarapid sequencing. Trends Biotechnol. *18*: 147-51.

Deamer, D.W. and Branton, D. (2002). Characterization of nucleic acids by nanopore analysis. Acc Chem Res. *35*: 817-825.

Dedrick, R. M., Aull, H.G., Jacobs-Sera, D., Garlena, R.A., Russell, D.A., Smith, B.E., Mahalingam, V., Abad, L., Gauthier, C. and Hatfull, G.F. (2021) The Prophage and Plasmid Mobilome as a Likely Driver of Mycobacterium abscessus Diversity. mBio. *12* (2) e03441-20

Deurenberga, R.H., Bathoorna, E., Chlebowicza, M.A., Coutoa, N., Ferdousa, M., García-Cobosa, S., Kooistra-Smida, A.M.D., Raangsa, E.C., Rosemaa, S., Velooa, A.C.M., Zhouc, K., Friedricha, A.W. and Rossen, J.W.A. (2017). Application of next generation sequencing in clinical microbiology and infection prevention. J Biotec. *243*: 16-24.

Devulder, G., Perouse de Montclos, M. and Flandrois, J.P. (2005). A multigene approach to phylogenetic analysis using the genus *Mycobacterium* as a model. Int J Syst Evol Microbiol. *55*: 293-302.

Diamond: Releases · bbuchfink/diamond · GitHub

Dokic, A., Peterson, E., Arrieta-Ortiz, M.L., Pan, M., Di Maio, A., Baliga, N. and Bhatt, A. (2021). *Mycobacterium abscessus* biofilms produce an extracellular matrix and have a distinct mycolic acid profile. The Cell Surface. *7*:100051.

Domenech, P., Jiménex, M.S., Menendez, M.C., Bull, T.J., Samper, S., Manrique, A. and Garcia, M.J. (1997). *Mycobacterium mageritense* sp. nov. Int. J. Syst. Bacteriol. *47*: 535-540

Dominguez Del Angel, V., Hjerde, E., Sterck, L., Capella-Gutierrez, S., Notredame, C., Vinnere Pettersson, O., Amselem, J., Bouri, L., Bocs, S., Klopp, C., Gibrat, J.F., Vlasova, A., Leskosek, B. L., Soler, L., Binzer-Panchal, M. and Lantz, H. 2018. Ten steps to get started in Genome Assembly and Annotation. F1000Research. *7* :ELIXIR-148.

Doroghazi, J.R. and Metcalf, W.W. (2013). Comparative genomics of actinomycetes with a focus on natural product biosynthetic genes. BMC Genomics. 14: 611.

Dumas, E., Boritsch, E.C., Vandenbogaert, M., Rodríguez de la Vega, R.C., Thiberge, J.M., Caro, V., Gaillard, J.L., Heym, B., Girard-Misguich, F., Brosch, R. and Sapriel, G. (2016). Mycobacterial Pan-Genome Analysis Suggests Important Role of Plasmids in the Radiation of Type VII Secretion Systems. Genome Biol Evol. *8* (2): 387-402.

East African/British Medical Research Council. (1973). Isoniazid with thiacetazone (thioacetazone) in the treatment of pulmonary tuberculosis in East Africa – third report of fifth investigation. Tubercle *54*: 169-179.

El-Halfawy, O. M. and Valvano, M. A. (2014). Putrescine reduces antibiotic-induced oxidative stress as a mechanism of modulation of antibiotic resistance in *Burkholderia cenocepacia*. Antimicrobial agents and chemotherapy. *58* (7): 4162–4171. https://doi.org/10.1128/AAC.02649-14

Elsik, C.G., Tellam, R.L. and Worley, K.C. (2009). The genome sequence of taurine cattle: a window to ruminant biology and evolution. Science. *324*: 522-528.

Euzeby, J.P. (1997). List of bacterial names with standing in nomenclature: A folder available on the internet. Int J Syst Bacteriol. *47*:590–592.
https://lpsn.dsmz.de/genus/mycobacterium

Ewing, B. and Green, P. (1998). Base-calling of automated sequencer traces using Phred. II. Error probabilities. Genome research. *8*(3): 186-194.

Falkinham, J. O. III, (1996). Epidemiology of infection by nontuberculous mycobacteria. Clin Microbiol Rev *9*: 177-215.

Faller, M., Niederweis, M. and Schulz, G.E. (2004). The Structure of a Mycobacterial Outer-Membrane Channel. Science. *303*: 1189 - 1192.

Faverio, P., De Giacomi, F., Bodini, B. D., Stainer, A., Fumagalli, A., Bini, F., Luppi, F. and Aliberti, S. (2021). Nontuberculous mycobacterial pulmonary disease: an integrated approach beyond antibiotics. ERJ Open Research. *7* (2): 00574-2020.

Federhen, S. 2012. The NCBI Taxonomy database. Nucleic Acids Res. *40* (Database issue): D136-43.

Fernández, L. and Hancock, R.E. (2012). Adaptive and mutational resistance: role of porins and efflux pumps in drug resistance. Clin Microbiol Rev. *25*:661–681

Ferrell, K.C., Johansen M. D. , Triccas, J. A. and Counoupas, C. (2022). Virulence Mechanisms of *Mycobacterium abscessus*: Current Knowledge and Implications for Vaccine Design. Frontiers in Microbiology. *13*: 842017.

Feuillet, C., Leach J.E., Rogers J., Schnable, P.S. and Eversole, K. (2011). Crop genome sequencing: Lessons and rationales. Trends Plant Sci *16***:** 77-88.

Fields, R. N. and Roy, H. (2018). Deciphering the tRNA-dependent lipid aminoacylation systems in bacteria: Novel components and structural advances. RNA Biology. *15* (4-5): 480–491

Fisheder, R., Schulze-Röbbecke, R. and Weber, A. (1991). Occurrence of mycobacteria in drinking water samples. Zentralbl Hyg Umweltmed. *192*: 154-158.

Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., Bult, C.J., Tomb, J.F., Dougherty, B.A., Merrick, J.M., McKenny, K., Sutton, G., Fitzhugh, W., Fields,

C., Gocayne, J.D., Scott, J., Shirley, R., Liu, L.I., Glodek, A., Kelley, J.M., Weidman, J.F., Phillips, C.A., Spriggs, T., Hedblom, E., Cotton, M.D., Utterback, T.R., Hanna, M.C., Nguyen, D.T., Saudek, D.M., Brandon, R.C., Fine, L.D., Fritchman, J.L., Fuhrmann, J.L., Geoghagen, N.S.M., Gnehm, C.L., McDonald, L.A., Small, K.V., Fraser, C.M., Smith , H.O. and Venter, J.C. (1995). Whole genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science. *269* (5223): 496-512.

Flye: De novo assembler for single molecule sequencing reads using repeat graphs https://github.com/fenderglass/Flye

Fologea, D., Gershow, M., Ledden, B., McNabb, D.S., Golovchenko, J.A. and Li, J. (2005). Detecting single stranded DNA with a solid state nanopore. Nano lett *5* (10): 1905-1909.

Forth, L.F. and Höper, D. (2019). Highly efficient library preparation for Ion Torrent sequencing using Y-adapters. BioTechniques *67*: 229-237

France, A.J., McLeod, D.T., Calder, M.A. and Seaton, A. (1987). *Mycobacterium malmoense* infections in Scotland: an increasing problem. Thorax *42*(8): 593-595.

Fraser, C.M., Eisen, J.A., Nelson, K.E., Paulsen, I.T. and Salzberg, S.L. (2002). The value of complete microbial genome sequencing. J Bacteriol. *184* (23): 6403-6405.

Fukano, H., Wada, S., Kurata, O., Katayama, K., Fujiwara, N. and Hoshino, Y. (2017a). *Mycobacterium stephanolepidis* sp. nov., a rapidly growing species related to Mycobacterium chelonae, isolated from marine teleost fish, Stephanolepis cirrhifer. Int J Syst Evol Microbiol. *67*:2811–2817.

Fukano H., Yoshida, M., Katayama, Y., Omatsu, T., Mizutani, T., Kurata, O., Wada, S. and Hoshino, Y. (2017b). Complete genome sequence of *Mycobacterium stephanolepidis*. Genome Announce *5*: e00810-17. https://doi.org/10.1128/genomeA.00810-17.

Funa, N., Ozawa, H., Hirata, A. and Horinouchi, S. (2006). Phenolic lipid synthesis by type III polyketide synthases is essential for cyst formation in *Azotobacter vinelandii*. Proceedings of the National Academy of Sciences. *103*: 6356–6361

Funabashi, M., Funa, N. and Horinouchi, S. (2008). Phenolic Lipids Synthesized by Type III Polyketide Synthase Confer Penicillin Resistance on *Streptomyces griseus*. Journal of Biological Chemistry. *283* (20): 13983–13991

Fussle, R., Bhakdi, S., Sziegoleit, A., Tranum-Jensen, J., Kranz, T. and Weflensiek, H.L. (1981). On the mechanism of membrane damage by *S. aureus* a-toxin. J Cell Biol. *91*: 83-94.

Galili, T. (2015). dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering. Bioinformatics. *31*: 3718–372.

Galili, T., O'Callaghan, A., Sidi, J. and Sievert, C. (2018). heatmaply: an R package for creating interactive cluster heatmaps for online publishing Bioinformatics. *34* (9): 1600–1602.

Garaj, S., Hubbard, W., Reina, A., Kong, J., Branton, D. and Golovchenko, J.A. (2010). Graphene as a sub-nanometer trans-electrode membrane. Nature. *467* (7312): 190-193.

Garrity, G.M. and Holt, J.G. (2001). The road map to the manual. In: Bergey's Manual of Systematic bacteriology, 2nd edn, vol. 1 (edited by Boone and Castenholz). Springer, New York, pp. 119-155.

Gautam, A., Felderhoff, H., Bagci, C. and Huson, D.H. (2022). Using AnnoTree to Get More Assignments, Faster, in DIAMOND+MEGAN Microbiome Analysis. mSystems. *7* (1):e0140821.

Gautreau, G., Bazin, A., Gachet, M., Planel, R., Burlot, L., Dubois, M., Perrin, A., Médigue, C., Calteau, A., Cruveiller, S., Matias, C., Ambroise, C., Rocha, E.P.C. and Vallenet, D. (2020). PPanGGOLiN: Depicting microbial diversity via a partitioned pangenome graph. PLoS Comput Biol. *16* (3):e1007732.
Erratum in: PLoS Comput Biol. 2021 Dec 10;17(12):e1009687.

Geiman, D.E., Raghunand, T.R., Agarwal, N. and Bishai, W.R. (2006). Differential Gene Expression in Response to Exposure to Antimycobacterial Agents and Other Stress Conditions among Seven Mycobacterium tuberculosis *whi*B-Like Genes. *Antimicrobial Agents and Chemotherapy. 50*: 2836 - 2841.

Geneious Prime: Geneious Prime 2022.0.1 http://www.geneious.com/

Gene Ontology Consortium. (2021). The Gene Ontology resource: enriching a GOld mine. Nucleic Acids Research. *49* (D1): D325–D334.

Getahun, H., Matteelli, A., Chaisson, R.E. and Raviglione, M. (2015). Latent Mycobacterium tuberculosis Infection. N Engl J Med. *372*: 2127-2135.

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

Gira, A.K., Reisenauer, A.H., Hammock, L., Nadiminti, U., Macy, J.T., Reeves, A., Burnett, C., Yakrus, M.A., Toney, S., Jensen, B.J., Blumberg, H.M., Caughman, S.W. and Nolte, F.S. (2004). Furunculosis due to *Mycobacterium mageritense* associated with footbaths at a nail salon. J Clin Microbiol. *42*: 1813-1817.

GitHub - bbuchfink/diamond: Accelerated BLAST compatible local sequence aligner.

Glasel, J.A. (1995). Validity of nucleic acid monitored by 260 nm/280 nm absorbance ratios. BioTechniques. *18*:62-63

Glenn, T.C. (2011). Field guide to next-generation DNA sequencers. Molecular Ecology Resources. *11*: 759-769.

Gokulan, K., Sangeeta, K. and Cerniglia, C. (2014). Metabolic Pathways: Production of Secondary Metabolites of Bacteria. Encyclopedia of Food Microbiology. vol 2. Elsevier Ltd, Academic Press. Pages 561–569.

González-Cano, P., Mondragon-Flores, R., Sánchez-Torres, L.E., González-Pozos, S., Silva-Miranda, M., Monroy-Ostria, A., Estrada-Parra, S. and Estrada-García, I. (2010). Mycobacterium tuberculosis H37Rv induces ectosome release in human polymorphonuclear neutrophils. *Tuberculosis*. *90* (2): 125-34.

Gonzalez-Santiago, T. M. and Drage, L. A. (2015). Nontuberculous Mycobacteria: Skin and Soft Tissue Infections. Dermatologic clinics. *33* (3): 563–577.

Goodfellow, M. and Magee, J.G. (1998). Taxonomy of mycobacteria. In: Mycobacteria, vol. 1: Basic Aspects (edited by P. Gangadharam and Jenkins P.A.) Chapman & Hall, New York pp. 1-71.

Goodwin, S., McPherson, J. D. and McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. Nature reviews. Genetics. *17* (6): 333–351.

Gordin, F.M. and Horsburgh, C.R. Jr. (2015). *Mycobacterium avium* complex. In: Mandell, Douglas and Bennett's, Principles and Practice of Infectious Diseases (Bennett, Dolin & Blaser eds.) Elsevier.

Goris, J., Konstantinidis, K.T., Klappenbach, J.A., Coenye, T., Vandamme, P. and Tiedje, J.M. (2007). DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. International Journal of Systematic and Evolutionary Microbiology. *57* (1): 81–91

Gorzynski, M., Week, T., Jaramillo, T., Dzalamidze, E. and Danelishvili, L. (2021). *Mycobacterium abscessus* Genetic Determinants Associated with the Intrinsic Resistance to Antibiotics. Microorganisms. *9* (12): 2527.

Gouaux, E. (1998). a Hemolysin from *Staphylococcus aureus*: an archetype of b-barrel, channel-forming toxins. J Struct Biol. *121*: 110-22.

Gouy, M., Guindon, S. and Gascuel, O. (2010). SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol Biol Evol. *27*: 221-224.

Graham, M.D. (2003). The Coulter principle: Foundation of an industry. J Assoc Lab Automation. *8*: 72-81.

Grein, F., Müller, A., Scherer, K.M., Liu, X., Ludwig, K.C., Klöckner, A., Strach, M., Sahl, H.G., Kubitscheck, U. and Schneider, T. (2020). Ca^{2+}-Daptomycin targets cell wall biosynthesis by forming a tripartite complex with undecaprenyl-coupled intermediates and membrane lipids. Nature communications. *11*(1):1455.

Griffith, D.E., Brown, B.A., Girard, W.M., Murphy, D.T. and Wallace, R.J. (1996). Azithromycin activity against *Mycobacterium avium* complex lung disease in HIV negative patients. Clin Infect Dis. *23*: 983-989.

Griffith, D.E., Aksamit, T., Brown-Elliott, B.A., Catanzaro, A., Daley, C., Gordin, F., Holland, S.M., Horsburgh, R., Huitt, G., Iademarco, M.F., Iseman, M., Olivier, K., Ruoss, S., von Reyn, C.F., Wallace, R.J. and Winthrop, K. (2007). An official ATS/IDSA statement: diagnosis, treatment, and prevention of nontuberculous mycobacterial diseases. Am J Respir Crit Care Med. *175*: 367– 416.

Guo, Q., Chu, H., Ye, M., Zhang, Z., Li, B., Yang, S., Ma, W. and Yu, F. (2018). The clarithromycin susceptibility genotype affects the treatment outcome of patients with *Mycobacterium abscessus* lung disease. Antimicrob Agents Chemother. *62*: e02360-17.

Gupta R.S., Lo, B. and Son, J. (2018). Phylogenomics and comparative genomic studies robustly support division of the genus Mycobacterium into an emended genus Mycobacterium and four novel genera. Front Microbiol. *9*:67.

Haft, D. H., DiCuccio, M., Badretdin, A., Brover, V., Chetvernin, V., O'Neill, K., Li, W., Chitsaz, F., Derbyshire, M. K., Gonzales, N. R., Gwadz, M., Lu, F., Marchler, G. H., Song, J. S., Thanki, N., Yamashita, R. A., Zheng, C., Thibaud-Nissen, F., Geer, L. Y., Marchler-Bauer, A. and Pruitt, K. D.

(2018). RefSeq: an update on prokaryotic genome annotation and curation. Nucleic Acids Research. *46* (D1): 851-860.

Harris, T.D., Buzby, P.R., Babcock, H., Beer, E., Bowers, J., Braslavsky, I., Causey, M., Colonell, J., Dimeo, J., Efcavitch, J.W., Giladi, E., Gill, J., Healy, J., Jarosz, M., Lapen, D., Moulton, K., Quake, S.R., Steinmann, K., Thayer, E., Tyurina, A., Ward, R., Weiss, H. and Xie, Z. (2008). Single-molecule DNA sequencing of a viral genome. Science. *320* (5872):106-109.

Harris, K.A. and Kenna. D. **(**2014). Mycobacterium abscessus infection in cystic fibrosis: Molecular typing and clinical outcomes. Journal of Medical Microbiology. *63*:1241-1246

Harris, N.C., Sato, M., Herman, N.A., Twigg, F., Cai, W., Liu, J., Zhu, X., Downey, J., Khalaf, R., Martin, J., Koshino, H. and Zhang, W. (2017). Biosynthesis of isonitrile lipopeptides by conserved non ribosomal peptide synthetase gene clusters in Actinobacteria. Proc Natl Acad Sci U S A. *114* (27):7025-7030.

Haworth, C. S., Banks, J., Capstick, T., Fisher, A. J., Gorsuch, T., Laurenson, I. F., Leitch, A., Loebinger, M.R., Milburn, H. J., Nightingale, M., Ormerod, P., Shingadia, D., Smith, D., Whitehead, N., Wilson, R. and Floto, R. A. (2017). British Thoracic Society guidelines for the management of non-tuberculous mycobacterial pulmonary disease (NTM-PD). Thorax. 72 (Suppl 2): ii1–ii64.

Hamdi, A., Fida, M., Deml, S.M., Saleh, Abu O. and Wengenack, N.L. 2020. Utility of *Mycobacterium tuberculosis* PCR in ruling out active disease and impact on isolation requirements in a low prevalence setting. J Clin Tuberc Other Mycobact Dis. *21*: 100181

Hayden, E.C. (2012). Nanopore genome sequencer makes its debut. Nature, News doi:10.1038/nature.2012.10051.

Head, S. R., Komori, H. K., LaMere, S. A., Whisenant, T., Van Nieuwerburgh, F., Salomon, D. R. and Ordoukhania, P. (2014). Library construction for next-generation sequencing: Overviews and challenges. Biotechniques. *56* (2): 61.

Henry, M.W., Miller, A.O., Kahn, B., Windsor, R.E and Brause, B.D. (2016) Prosthetic joint infections secondary to rapidly growing mycobacteria: Two case reports and a review of the literature. Infectious Diseases. *48* (6): 453-460.

Herbst, D. A., Townsend, C. A. and Maier, T. (2018). The architectures of iterative type I PKS and FAS. Natural Product Reports. *35* (10): 1046–1069.

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

Irons, J.L., Hodge-Hanson, K.M. and Downs, D.M. (2020). RidA Proteins Protect against Metabolic Damage by Reactive Intermediates. *Microbiology and Molecular Biology Reviews. 84*, Issue 3: e00024-20.

Hong Kong Chest Service, Medical Research Council. (1981). Controlled trial of four thrice weekly regimens and a daily regimen given for 6 months for pulmonary tuberculosis. *Lancet*. *1* (8213): 171–174.

Howard, S.T. (2013) Recent Progress Towards Understanding Genetic Variation in the *Mycobacterium abscessus* complex. Tuberculosis. *93*: S15-S20

Huang, J., Liang, X., Xuan, Y., Geng, C., Li, Y., Lu, H., Qu, S., Mei, X., Chen, H., Yu, T., Sun, N., Rao, J., Wang, J., Zhang, W., Chen Y., Liao, S., Jiang, H., Liu, X., Yang, Z., Mu, F. and Gao, S.

(2017). A reference human genome dataset of the BGISEQ-500 sequencer. Gigascience. *6* (5):1-9.

Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S.K., Cook, H., Mende, D.R., Letunic, I., Rattei, T., Jensen, L.J., von Mering, C. and Bork, P. (2019). eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic acids research. *47* (D1), D309–D314.

Hug, L.A., Baker, B.J., Anantharaman, K., Brown, C.T., Probst, A.J., Castelle, C.J., Butterfield, C.N., Hernsdorf, A.W., Amano, Y., Ise, K., Suzuki, Y., Dudek, N., Relman, D.A., Finstad, K.M., Amundson, R., Thomas, B.C. and Banfield, J.F. (2016). A new view of the tree of life. Nature Microbiology. *1*: 16048.

Hurst, K., Rudra, P. and Ghosh, P. (2017). Mycobacterium abscessus *whi*B7 Regulates a Species-Specific Repertoire of Genes To Confer Extreme Antibiotic Resistance. Antimicrobial Agents and Chemotherapy. *61*.

Hurst-Hess, K., Rudra, P. and Ghosh, P. (2017). *Mycobacterium abscessus whi*B7 regulates a species-specific repertoire of genes to confer extreme antibiotic resistance. Antimicrobial Agents and Chemotherapy. *61* (11): e01347-17.

Huson, D.H., Richter, D.C., Rausch, C., Dezulian, T., Franz, M. and Rupp, R. (2007). Dendroscope: An interactive viewer for large phylogenetic trees. BMC Bioinformatics. *8*:460.

Huson, D.H., Beier S, Flade, I., Górska, A., El-Hadidi, M., Mitra, S., Ruscheweyh, H-J. and Tappu, R. (2016). MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data. PloS Computational Biology. *12* (6): e1004957

Hyatt, D., Chen, G-L., LoCascio, P.F., Land, M.L., Larimer, F.W. and Hauser, L.J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. *11*:119.

Irgens, L.M. (2002). The discovery of the leprosy bacillus. Tidsskrift for den Norske laegeforening. *122* (7): 708–9.

Isom, G.L., Davies, N.J., Chong, Z-S., Bryant, J.A., Jamshad, M., Sharif, M., Cunningham, A.F., Knowles, T.J., Chng, S-S., Cole, J.A. and Henderson, I.R. (2017). 'MCE domain proteins: conserved inner membrane lipid-binding proteins required for outer membrane homeostasis. *Scientific Reports*. *7* (1) 8608.

Jackson, M., Stevens, C.M., Zhang, L., Zgurskaya, H.I. and Niederweis, M. (2021). Transporters Involved in the Biogenesis and Functionalization of the Mycobacterial Cell Envelope. Chem Rev. *121*(9):5124-5157.

Jacobs, M.R. (1999). Activity of quinolones against mycobacteria. Drugs *58* (2): 19-22:

Jain, C., Rodriguez-R, L.M., Phillippy A.M., Konstantinidis, T. and Aluru, S. (2018). High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nature Communications. *9* **:** 5114.

Jerry, A. N,. Dedrick, R.M., Gray, A. L., Vladar, E. K., Smith, B. E., Freeman, K.G., Malcolm, K. C., Epperson, L. E., Hasan, N. A., Hendrix, J., Callahan, K., Walton, K., Vestal, B., Wheeler, E., Rysavy, N. M., Poch, K., Caceres, S., Lovell, V. K., Hisert, K. B., de Moura, V. C., Chatterjee, D., De, P., Weakly, N., Martiniano, S. L., Lynch, D. A., Daley, C. L., Strong, M. l, Jia, F., Hatfull, G. F. and Davidson, R. M. 2022. Host and pathogen response to bacteriophage engineered against Mycobacterium abscessus lung infection. Cell. *185*: 1860–1874.

Johansen, M.D., Herrmann, J.L. and Kremer, L. (2020). Non-tuberculous mycobacteria and the rise of Mycobacterium abscessus. Nat Rev Microbiol. *18*: 392-407.

Johnson, M.M. and Odell, J.A. (2014). Nontuberculous mycobacterial pulmonary infections. J Thorac Dis *6*(3): 210-220.

Jukes, T.H. and Cantor, C. (1969). Evolution of protein molecules. In: mammalian protein metabolism (edited by H.N. Munro). Academic Press, New York pp. 21-132.

Kanagal-Shamanna, R. Emulsion PCR: Techniques and Applications. (2016). Methods Mol Biol. *1392*:33-42.

Kanehisa, M.; "Post-genome Informatics", Oxford University Press (2000)

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. and Tanabe, M. (2016). KEGG as a reference resource for gene and protein annotation. Nucleic Acids Research. *44*, (D1) 457-462.

Kasianowicz, J.J., Brandin, E., Branton, D. and Deamer, D.W. (1996). Characterization of individual polynucleotide molecules using a membrane channel. Proc Natl Acad Sci USA. *93*: 13770–13773.

Katoh, K. and Standley, D.M. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol Biol Evol. 30* (4): 772–780. https://doi.org/10.1093/molbev/mst010

Kaustova, J., Olsovsky, Z., Kubin, M., Zatloukal, O., Pelikan, M. and Hradil, V. (1981). Endemic occurrence of *Mycobacterium kansasii* in water supply systems. J Hyg Epidemiol Microbiol Immunol. *25*: 24-30.

Kautsar, S.A., Blin, K., Shaw, S., Navarro-Muñoz, J.C., Terlouw, B.R., van der Hooft, J.J.J., van Santen, J.A., Tracanna, V., Suarez Duran, H.G., Andreu, V.P., Selem-Mojica, N., Alanjary, M., Robinson, S.L., Lund, G., Epstein, S.C., Sisto, A.C., Charkoudian, L.K., Collemare, J., Linington, R.G., Weber, T. and Medema, M.H. 2020. MIBiG 2.0: a repository for biosynthetic gene clusters of known function. Nucleic Acids Research. *48* (D1): D454–D458.

Kawano, R., Schibel, A.E., Cauley, C. and White, H.S. (2009). Controlling the translocation of single-stranded DNA through alpha-hemolysin ion channels using viscosity. Langmuir *25*:1233-7.

Kerkhof, L. J. (2021). Is Oxford Nanopore sequencing ready for analyzing complex microbiomes? FEMS Microbiology Ecology. *97* (3), fiab001.

Kieser, T., Bibb, M.J., Buttner, M.J., Chater, K.F. and Hopwood, D.A. (2000). Practical *Streptomyces* genetics. Norwich: John Innes Foundation.

Kim, B-J., Kim, B-R., Jeong, J., Lim, J-H., Park, S.H., Lee, S-H., Kim, C.K., Kook, Y-H. and Kim, B-J. (2018). A description of *Mycobacterium chelonae* subsp. *gwanakae* subsp. nov., a rapidly growing mycobacterium with a smooth colony phenotype due to glycopeptidolipids. Int J Syst Evol Microbiol. *68:*.3772-3780.

Kim, B-J., Kim, G-N., Kim, B-R., Jeon, C-O., Jeong, J., Lee, S.H., Lim, J-H., Lee, S-H., Kim ,C.K., Kook, Y-H. and Kim, B-J. (2017). Description of *Mycobacterium chelonae* subsp. *bovis* subsp. nov., isolated from cattle (Bos taurus coreanae), emended description of *Mycobacterium chelonae* and creation of *Mycobacterium chelonae* subsp. *chelonae* subsp. nov. Int J Syst Evol Microbiol. *67*:3882-3887.

Kim, H. Y., Kook, Y., Yun, Y. J., Park, C. G., Lee, N. Y., Shim, T. S., Kim, B. J. and Kook, Y. H. (2008). Proportions of Mycobacterium massiliense and Mycobacterium bolletii strains among Korean Mycobacterium chelonae-Mycobacterium abscessus group isolates. Journal of Clinical Microbiology. *46* (10): 3384–3390.

Kim, J., Sung, H., Park, J.S., Choi, S.H., Shim, T.S. and Kim, M.N. (2016). Subspecies distribution and macrolide and fluoroquinolone resistance genetics of Mycobacterium abscessus in Korea. Int J Tuberc Lung Dis. *20* (1): 109-114.

Kim, S-Y., Shin, S.J., Jeong, B-H. and Koh, W-J. (2016). Successful antibiotic treatment of pulmonary disease caused by *Mycobacterium abscessus* subsp. *abscessus* with C-to-T mutation at position 19 in *erm*(41) gene: case report. BMC Infectious Diseases. *16*:207.

Kim, Y. S., Yang, C.S., Nguyen, L.T., Kim, J.K., Jin, H.S., Choe, J., Kim, S.Y., Lee, H-M., Jung, M., Kim, J-M., Kim, M. H., Jo, E-K. and Jang, J-C. (2017). Mycobacterium abscessus ESX-3 plays an

important role in host inflammatory and pathological responses during infection. Microbes and Infection. *19* (1): 5-17

Khasheii, B., Mahmoodi, P. and Mohammadzadeh, A. (2021). Siderophores: Importance in bacterial pathogenesis and applications in medicine and industry. Microbiological research. *250*, 126790.

Klemm, P. and Schembri, M.A. (2000). Bacterial adhesins: function and structure. International Journal of Medical Microbiology. *290* (1): 27-35.

Kneitz, S. and Dandekar, T. (2006). Pathogenomics: Genome Analysis of Pathogenic Microbes. Wiley-VCH Verlag GmbH & Co.

Knierim, E., Lucke, B., Schwarz, J.M., Schuelke, M. and Seelow, D. (2011). Systematic Comparison of Three Methods for Fragmentation of Long-Range PCR Products for Next Generation Sequencing. PLoS ONE *6*(11): e28240.

Koch, R. (1882). Die Aetiologie der Tuberculose. Berliner Klin Wochenschr *19*: 221-238.

Koh, W.J., Chang, B., Jeong, B.H., Jeon, K., Kim, S.U., Lee, N.M, Ki, C.S. and Kwon, O.J. (2013). Increasing Recovery of Nontuberculous Mycobacteria from Respiratory Specimens over a 10-year period in a Tertiary Referral Hospital in South Korea. Tuberculosis and Respiratory Diseases. *75*(5):199-204.

Koh, W-J., Jeon, K., Lee, N.Y., Kim, B-J, Kook, Y-H., Lee, S-H., Park, Y.K., Kim, C.K., Shin, S.J., Huitt, G.A., Daley, C.L. and Kwon, O.J. (2011). Clinical significance of differentiation of Mycobacterium massiliense from Mycobacterium abscessus. American Journal of Respiratory and Critical Care Medicine. *183* (3): 405-410.

Koh, W-J., Stout J. E. and Yew W.W. (2014). Advances in the management of pulmonary disease due to Mycobacterium abscessus complex. Int J Tuberc Lung Dis. *18 (10)*: 1141–1148.

Kolmogorov, M., Yuan, J., Lin, Y. and Pevzner, P.A. (2019). Assembly of long, error-prone reads using repeat graphs. Nature Biotechnology. *37* (5): 540–546.

Konstantinidis, K., Ramette, A. and Tiedje, J. (2006). The bacterial species definition in the genomic era. Philosophical transactions of the Royal Society of London. Series B, Biological sciences. *361*: 1929-40.

Koonin, E.V. and Wolf, Y.I. (2008). Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. Nucleic Acids Res. *36* (21):6688–6719.

Koren., S, Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H. and Phillipy, A.M. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome research. *27* 5:722-236

Koul, A., Dendouga, N., Vergauwen, K., Molenberghs, B., Vranckx, L., Willebrords, R., Ristic, Z., Lill, H., Dorange, I., Guillemont, J., Bald, D. and Andries, K. (2007). Diarylquinolines target subunit c of mycobacterial ATP synthase. Nat Chem Biol. *3*:323–324.

Kreiss, K. and Cox-Ganser, J. (1997). Metalworking fluid-associated hypersensitivity pneumonitis: a workshop summary. Am J Ind Med. *32*: 423-432.

Krishnamoorthy, G., Kaiser, P., Lozza, L., Hahnke, K., Mollenkopf, H.J. and Kaufmann, S. (2019). Mycofactocin Is Associated with Ethanol Metabolism in Mycobacteria. Molecular Biology and Physiology. *10* (3): e00190-19.

Kubica, G. P., Baess, I., Gordon, R.E., Jenkins, P.A, Kwapinski, J.B.G., McDurmont, C., Pattyn, S.R., Saito, H., Silcox, V., Stanford, J.L., Takeya, K. and Tsukamura, M. (1972). A cooperative numerical analysis of the rapidly growing mycobacteria. J. Gen. Microbiol. *73*:55-70.

Kumar, P., Chauhan, V., Silva, J.R.A., Lameira, J., d'Andrea, F.B., Li, S-G., Ginell, S.L., Freundlich, J.S., Alves, C.N., Bailey, S., Cohen, K.A. and Lamichhane, G. (2017a). *Mycobacterium abscessus* L, D-transpeptidases are susceptible to inactivation by carbapenems and cephalosporins but not penicillins. Antimicrob Agents Chemother. *61*: e00866-17.

Kumar, P., Kaushik, A., Lloyd, E.P., Li, S.G., Mattoo, R., Ammerman, N.C., Bell, D.T., Perryman, A.L., Zandi, T.A., Ekins, S., Ginell, S.L, Townsend, C.A., Freundlich, J.S. and Lamichhane, G. (2017b). Non-classical transpeptidases yield insight into new antibacterials. Nat Chem Biol. *13*:54–61.

Kuo, C.J., Ptak, C.P., Hsieh, C.L., Akey, B.L. and Chang, Y.F. (2013). Elastin, a novel extracellular matrix protein adhering to mycobacterial antigen 85 complex. The Journal of Biological Chemistry. *288* (6): 3886–3896.

Kusunoki, S. and Ezaki, T. (1992). Proposal of *Mycobacterium peregrinum* sp. nov., nom. rev., and elevation of *Mycobacterium chelonae* subsp. *abscessus* (Kubica *et al*) to species status: *Mycobacterium abscessus* comb. nov. Int. J. Syst. Bacteriol. *42*:240-245.

Lalor M.K., Casali, N., Walker, T.M., Anderson, L.F., Davidson, J.A., Ratna, N., Mullarkey, C., Gent, M., Foster, K., Brown, T., Magee, J., Barrett, A., Crook, D.W., Drobniewski, F., Thomas H. L., and Abubakar, I. (2018). The use of whole genome sequencing in cluster investigation of a multidrug-resistant tuberculosis outbreak. Eur Respir J. *51*: 1702313.

Lamy, B., Marchandin, H., Hamitouche, K. and Laurent, F. (2008). *Mycobacterium setense* sp. nov., a *Mycobacterium fortuitum*-group organism isolated from a patient with soft tissue infection and osteitis. Int. J. Syst. Evol. Microbiol. *58*:486-490.

Lapierre, P. and Gogarten, J.P. (2009). Estimating the size of the bacterial pan-genome. Trends Genet. *25* (3): 107–110.

Le Dantec, C., Duguet, J-P., Montiel, A., Dumoutier, N., Dubrou, S. and Vincent, V. (2002). Occurrence of mycobacteria in water treatment lines and in water distribution systems. Appl Environ Microbiol. *68* (11): 5318-5325.

Leao, S. C., Tortoli, E., Viana-Niero, C., Ueki, Sy.-M., Batista Lima, K.V., Lopes, M.L., Yubero, J., Menendez, M.C. and Garcia, M.J. (2009). Characterisation 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. *47*: 2691-2698.

Leao, S. C., Tortoli, E., Euzéby, G.P. and Garcia, M.J. (2011). Proposal that *Mycobacterium massiliense* and *Mycobacterium bolletii* be united and reclassified as *Mycobacterium abscessus* subsp. *bolletii*, conv. nov., designation of *Mycobacterium abscessus* subsp. *abscessus* subsp. nov. and emended description of *Mycobacterium abscessus.* Int. J. Syst. Evol. Microbiol. *61*:2311-2313.

Lee, J.H. (2019). Perspectives towards antibiotic resistance: from molecules to population. J Microbiol. *57* 181–184.

Lee, H.K., Lee, S-A., Lee, I-K., Yu, H-K, Park, Y-G., Hyun, J-W., Kim, K., Kook, Y-H. and Kim, B-J. (2010). *Mycobacterium paraseoulense* sp. nov., a slowly growing, scotochromogenic species related genetically to *Mycobacterium seoulenese*. Int J Syst Evol Microbiol. *59*: 2803-2808.

Lee, C.Y., Yen, H.Y., Zhong, A.W. and Gao, H. (2021). Resolving misalignment interference for NGS-based clinical diagnostics. Human Genetics. *140*: 477–492.

Lévy-Frébault, V., Grimont, F., Grimont, P.A. and David, H.L. (1986). Deoxyribonucleic acid relatedness study of the *Mycobacterium fortuitum-Mycobacterium chelonae* complex. Int J Syst Bacteriol. *36*: 456-460.

Lévy-Frébault, V. V. and Portaels, F. (1992). Proposed minimal standards for the genus Mycobacterium and for description of new slowly growing Mycobacterium species. International journal of systematic bacteriology. *42* (2), 315–323.

Lewis, K. (2001). Riddle of biofilm resistance. Antimicrobial agents and chemotherapy. *45* (4): 999-1007.

Li, D., Liu, C.M., Luo, R., Sadakane, K. and Lam, T.W. (2015). MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics*. *31* (10):1674-1676.

Li, W., O'Neill, K. R., Haft, D. H., DiCuccio, M., Chetvernin, V., Badretdin, A., Coulouris, G., Chitsaz, F., Derbyshire, M. K., Durkin, A. S., Gonzales, N. R., Gwadz, M., Lanczycki, C. J., Song, J. S., Thanki, N., Wang, J., Yamashita, R. A., Yang, M., Zheng, C., Marchler-Bauer, A. and Thibaud-Nissen, F. (2021). RefSeq: expanding the Prokaryotic Genome Annotation Pipeline reach with protein family model curation. Nucleic Acids Research. *49* (D1): 1020-1028.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *Genomics arXiv:*1303-3997 v2.

Li, H. (2016). Minimap and miniasm: Fast mapping and de novo assembly for noisy long sequences. Bioinformatics. *32*: 2103–2110.

Li, J., Stein, D., McMullan, C., Branton, D., Aziz, M.J. and Golovchenko, J.A. (2001). Ion-beam sculpting at nanometre length scales. Nature. 412: 166-169.

Liu, W-T., Chen, E-Z., Yang, L., Peng, C., Wang, Q., Xu, Z. and Chen, D-Q. (2021). Emerging resistance mechanisms for 4 types of common anti-MRSA antibiotics in Staphylococcus aureus: A comprehensive review. Microbial Pathogenesis *156*: 104915.

Loman, N.J. and Quinlan, A.R. (2014). Poretools: a toolkit for analyzing nanopore sequence data. *Bioinformatics. 30* (23):3399–3401.

Loman, N.J. and Watson, M. (2015). Successful test launch for nanopore sequencing. Nature Methods. Vol 12, No 4: 303-304.

Loman, N.J., Constantinidou, C., Chan, J.Z.M., Halachev, M., Sergeant, M., Penn, C.W., Robinson, E.R. and Pallen, M.J. (2012). High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity. Nature Reviews Microbiology. 10: 599-606.

Loman, N.J., Quick, J. and Simpson, J.T. (2015). A complete bacterial genome assembled de novo using only nanopore sequencing data. Nature Methods. Vol 12, No 8: 733-736

LPSN/DSMZ https://www.bacterio.net

Lu, J. and Salzberg, S. L. (2018). Removing contaminants from databases of draft genomes. PLoS computational biology. 14 (6) : e1006277.

Ludwig, W., Euzéby, J. and Whitman, W.B. (2012). Taxonomic outline of the phylum *Actinobacteria*. In Bergey's Manual of Systematic Bacteriology, Volume 5, The Actinobacteria, Part A. (edited by M. Goodfellow, P. Kämpfer, H-J. Busse, M.E. Trujillo, K. Suzuki, W. Ludwig and W.B. Whitman). Springer, New York, pp.29-31.

Luthra S., Rominski, A. and Sander, P. (2018). The role of antibiotic-target-modifying and antibiotic-modifying enzymes in *Mycobacterium abscessus* drug resistance. Front Microbiol. *9*:2179.

Lyu, N., Feng, Y., Pan, Y., Huang, H., Liu, Y., Xue, C., Zhu, B. and Hu, Y. (2021) Genomic Characterization of *Salmonella enterica* Isolates From Retail Meat in Beijing, China. Frontiers in Microbiology. *12* (636332).

Ma, L.N., Mi, H.F., Xue, Y.X., Wang, D. and Zhao, X.L. (2016). The mechanism of ROS regulation of antibiotic resistance and antimicrobial lethality. Yi chuan = Hereditas. *38* (10), 902–909.

Madacki, J., Mas Fiol, G. and Brosch, R. (2019). Update on the virulence factors of the obligate pathogen Mycobacterium tuberculosis and related tuberculosis-causing mycobacteria. Infection Genetics and Evolution. *72*: 67-77.

Madoui, M-A., Engelen, S., Cruaud, C., Belser, C., Bertrand, L., Alberti, A., Lemainque, A., Wincker, P. and Aury, J.M. (2015). Genome assembly using Nanopore guided long and error free DNA reads. BMC Genomics. *16*: 327.

Magee, J.G. and Ward, A.C. (2012). *Mycobacteriaceae*. In Bergey's Manual of Systematic Bacteriology, Volume 5, The Actinobacteria, Part A. (edited by M. Goodfellow, P. Kämpfer, H-

J. Busse, M.E. Trujillo, K. Suzuki, W. Ludwig and W.B. Whitman). Springer, New York, pp. 312-375.

Manrao, E., Derrington, I., Pavlenok, M., Niederweis, M. and Gundlach, J. (2011). Nucleotide discrimination with DNA immobilized in the MspA nanopore. PloS ONE *6:10*; e25723.

Marcais, G., Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics. *27*: 764-770.

Mardis E.R. (2013). Next-generation sequencing platforms. Annu Rev Anal Chem (Palo Alto Calif). *6*:287-303.

Mardis E. R. (2017). DNA sequencing technologies: 2006–2016. Nature Protocols. *12* (2): 213-218.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y. J., Chen, Z., Dewell, S.B., Du, L., Fierro, J.M., Gomes, X. V., Godwin, B.C., He, W., Helgesen, S., Ho, C.H., Irzyk, G.P., Jando, S.C., Alenquer, M.L., Jarvie, T.P., Jirage, K. B., Kim, J.B., Knight, J.R., Lanza, J.R., Leamon, J.H., Lefkowitz, S. M., Lei, M., Li, J., Lohman, K. L., Lu, H., Makhijani, V.B., McDade, K.E., McKenna, M.P., Myers, E.W., Nickerson, E., Nobile, J.R., Plant, R., Puc, B.P., Ronan, M.T., Roth, G.T., Sarkis, G.J., Simons, J.F., Simpson, J.W., Srinivasan, M., Tartaro, K. R., Tomasz, A., Vogt, K. A., Volkmer, G.A., Wang, S.H., Wang, Y., Weiner, M.P., Yu, P., Begley, R.F. and Rothberg, J.M. (2005). Genome sequencing in microfabricated high-density picolitre reactors. Nature *437*(7057): 376-80.

Marras, T.K. and Daley, C.L. (2002). Epidemiology of human pulmonary infection with nontuberculous mycobacteria. Clin Chest Med. *23*(3): 553-567.

Martín-Casabona, N., Bahrmand, A.R., Bennedsen, J., Thomsen, V. Ø., Curcio, M.L., Fauville-Dufaux, M., Feldman, K., Havelková, M., Katila, M.L. Köksalan, K., Pereira, M.D., Rodrigues, F., Pfyffer, G.E., Portaels, F., Urgell, J.R., Rüsch-Gerdes, S., Tortoli, E., Vincent, V. and Watt, B. (2004). Non-tuberculous mycobacteria: patterns of isolation. A multi-country retrospective

survey. The international journal of tuberculosis and lung disease: the official journal of the International Union against Tuberculosis and Lung Disease. *8* (10)*:*1186-93.

Martiniano, S.L., Nick, J.A., and Daley, C.L. (2019). Nontuberculous Mycobacterial Infections in Cystic Fibrosis. Thorac. Surg. Clin. *29*: 95–108.

Masignani, V., Pizza, M. and Moxon, E.R. (2019). The Development of a Vaccine Against Meningococcus B Using Reverse Vaccinology. Fronthopen Immunol. *10*: 751.

Maxam, A M. and Gilbert, W. (1977). A new method for sequencing DNA. *Proc. Natl. Acad. Sci.* U.S.A. 74 (2): 560–4.

McCutcheon, J. P. and Moran, N. A. (2011). Extreme genome reduction in symbiotic bacteria. Nature reviews. Microbiology, 10 (1), 13–26.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M. and DePristo, M.A. (2010). The genome analysis toolkit: a MapReduce framework for analysing next-generation DNA sequencing data. Genome Res. *9*: 1297-1303.

Medema, M.H., Blin, K., Cimermancic, P., de Jager, V., Zakrzewski, P., Fischbach, M.A., Weber, T., Breitling, R. and Takano, E. (2011). antiSMASH: Rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters. Nucleic Acids Research. *39* (Web Server issue): W339-46.

Medha., Sharma, S. and Sharma, M. (2021). Proline-Glutamate/Proline-Proline-Glutamate (PE/PPE) proteins of Mycobacterium tuberculosis: The multifaceted immune-modulators. Acta Tropica. *222*: 106035.

Medical Research Council. (1952). The prevention of streptomycin resistance by combined chemotherapy. Br Med J. *1*: 1157-1162.

Medjahed, H., Gaillard, J.L. and Reyrat, J.M. (2019). Mycobacterium abscessus: a new player in the mycobacterial field. Trends in microbiology. *18* (3):117-123.

Megan: MEtaGenome Analyzer. Please also see Huson *et al*., 2016. https://software-ab.informatik.uni-tuebingen.de/download/megan6/welcome.html

Meier, A., Heifets, L., Wallace, R.J. Jr., Zhang, Y., Brown, B.A., Sander, P. and Böttger, E.C. (1996). Molecular mechanisms of clarithromycin resistance in Mycobacterium avium: observation of multiple 23S rDNA mutations in a clonal population. J. Infect. Dis. *174*:354–360.

Meier-Kolthoff, J.P., Sardà Carbasse, J., Peinado-Olarte, R.L. and Göker, M. (2022). TYGS and LPSN: a database tandem for fast and reliable genome-based classification and nomenclature of prokaryotes. Nucleic Acid Res. *50* (D1): D801–D807.

Meller A., Nivon, L., Brandin, E., Golovchenko, J. and Branton, D. (2000). Rapid nanopore discrimination between single polynucleotide molecules. Proc Natl Acad Sci USA. *97*:1079-84.

Mende, D. R., Sunagawa, S., Zeller, G and Bork, P. (2013). Accurate and universal delineation of prokaryotic species. Nat. Methods. *10* :881.

Mendler, K., Chen, H., Parks, D.H., Lobb, B., Hug, L.A. and Doxey, A.C. (2019). AnnoTree: visualization and exploration of a functionally annotated microbial tree of life. Nucleic Acids Res. *47* (9): 4442–4448.

Merriman, B., Rothberg, J., Aguinaldo, K. A ., Alanjary, M., Allen M., Altun, G., Andersen, M., Andruzzi, L., Angulo, M., Atehortua-Khalsa, K., Balineni, A., Ball, J., Barbee, K., Beasley, E., Beauchemin, M., Bee, G., Belhachemi, D., Bennett, R., Benson, S. and Zou, R. (2012). Progress in Ion Torrent semiconductor chip based sequencing. Electrophoresis. Electrophoresis. 3*3*. 3397-3417.

Migliori, G.B., Matteelli, A., Cirillo, D. and Pai, M. (2008). Diagnosis of multidrug-resistant tuberculosis and extensively drug-resistant tuberculosis: Current standards and challenges. Can J Infect Dis Med Microbiol. *19*(2): 169–172.

Mikheyev, A.S. and Tin, M.M.Y. (2014). A first look at the Oxford Nanopore MinION sequencer. Molecular Ecology Resources. *14*: 1097-1102.

Miller, J. R., Delcher, A.L., Koren, S., Venter, E., Walenz, B.P., Brownley, A., Johnson, J., Li, K., Mobarry, C. and Sutton, G. (2008). Aggressive assembly of pyrosequencing reads with mates. Bioinformatics *24* (24):2818–2824.

Minato, Y., Thiede, J.M., Kordus, S.L., McKlveen, E.J., Turman, B.J. and Baughn, A.D. (2015). Mycobacterium tuberculosis folate metabolism and the mechanistic basis for para-aminosalicylic acid susceptibility and resistance. Antimicrob agents and chemotherapy. *59* (9): 5097–5106.

Minias, A., Żukowska, L., Lach, J., Jagielski, T., Strapagiel, D., Kim, S.Y., Koh, W.J., Adam, H., Bittner, R., Truden, S., Žolnir-Dovč, M. and Dziadek, J. (2020). Subspecies-specific sequence detection for differentiation of *Mycobacterium abscessus* complex. Scientific reports. *10* (1): 16415.

Mitchell, N. and Howorka, S. 2008. Chemical tags facilitate the sensing of individual DNA strands with nanopores. Angew Chem Int Ed Engl. *47*:5565-8.

Mitchell, A., Chang, H-Y., Daugherty, L., Fraser, M., Hunter, S., Lopez, R., McAnulla, C., McMenamin, C., Nuka, G., Pesseat, S., Sangrador-vegas, A., Scheremetjew, M., Rato, C., Yong, S-Y., Bateman, A., Punta, M., Attwood, T.K., Sigrist, C.J.A., Redaschi, N., Rivoire, C., Xenarios, I., Kahn, D., Guyot, D., Bork, P., Letunic, I., Gough, J., Oates, M., Haft, D., Huang, H., Natale, D.A., Wu, C.H., Orengo, C., Sillitoe, I., Mi, H., Thomas, P.D. and Finn, R.D. (2015). The InterPro protein families database: The classification resource after 15 years. Nucleic Acids Research *43*: D213-D221.

Moldovan, M.A. and Gelfand, M.S. (2018). Pangenomic Definition of Prokaryotic Species and the Phylogenetic Structure of Prochlorococcus spp. Frontiers in Microbiology. *9*:428.

Montoya-Rosales, A., Provvedi, R., Torres-Juarez, F., Enciso-Moreno, J.A., Hernandez-Pando, R., Manganelli, R. and Rivas-Santiago, B. (2017). lysX gene is differentially expressed among *Mycobacterium tuberculosis* strains with different levels of virulence. Tuberculosis. *106*: 106-117.

Morgan, J., Smith, M., Mc Auley, M. T. and Enrique Salcedo-Sora, J. (2018). Disrupting folate metabolism reduces the capacity of bacteria in exponential growth to develop persisters to antibiotics. Microbiology (Reading, England). *164* (11), 1432–1445.

Morris, R.P., Nguyen, L., Gatfield, J., Visconti, K., Nguyen, K., Schnappinger, D., Ehrt, S., Liu, Y., Heifets, L., Pieters, J., Schoolnik, G. and Thompson, C.J. (2005). Ancestral antibiotic resistance in Mycobacterium tuberculosis. Proc Natl Acad Sci USA. *102* (34): 12200–12205.

Mosaei, H., Molodtsov, V., Kepplinger, B., Harbottle, J., Moon, C.W., Rose, E.J., Ceccaroni, L., Shin, Y., Morton-Laing, S., Marrs, E.C.L., Wills, C., Clegg, W., Yuzenkova, Y., Perry, J.D., Bacon, J., Errington, J., Allenby, N.E.E., Hall, M.J., Murakami, K.S. and Zenkin, N. (2018). Mode of Action of Kanglemycin A, an Ansamycin Natural Product that Is Active against Rifampicin-Resistant Mycobacterium tuberculosis. Molecular Cell. *72*: 263–274.

Moscoso, M., and García, E. (2009). Transcriptional Regulation of the Capsular Polysaccharide Biosynthesis Locus of Streptococcus Pneumoniae: a Bioinformatic Analysis. DNA Res. *16*: 177–186

Mukherjee, R. and Chatterji, D. (2012). Glycopeptidolipids: immuno-modulators in greasy mycobacterial cell envelope. IUBMB Life. *64* (3):215-25.

Munita, J.M., Bayer, A.S and Arias, C.A. (2015). Evolving Resistance Among Gram-positive Pathogens. *Clinical Infectious Diseases*. *61* (2): S48–S57.

Munita, J.M. and Cesar, A.A. (2016). Mechanisms of Antibiotic Resistance. Microbiology Spectrum. *4* (2).

Murray, G. G. R., Charlesworth, J., Miller, E. L., Casey, M. J., Lloyd, C. T., Gottschalk, M., Tucker, A. W. D., Welch, J. J., & Weinert, L. A. (2021). Genome Reduction Is Associated with Bacterial Pathogenicity Across Different Scales of Temporal and Ecological Divergence. Molecular biology and evolution. *38* (4), 1570–1579.

Myers, E. W., Sutton, G.G., Delcher, A.L., Dew, I.M., Fasulo, D.P., Flanigan, M.J., Kravitz, S.A. Mobarry, C.M., Reinert, K.H., Remington, K.A., Anson, E.L., Bolanos, R.A., Chou, H.H., Jordan, C.M., Halpern, A.L., Lonardi, S., Beasley, E.M., Brandon, R.C., Chen, L., Dunn, P.J. and Venter, J.C. (2000). A whole-genome assembly of Drosophila. Science. *287* (5461): 2196–2204.https://genome.cshlp.org/content/27/5/722.long - ref-45

Nadal Jimenez, P., Koch, G., Thompson, J.A., Xavier, K.B., Cool, R.H. and Quax, W.J. (2012). The multiple signalling systems regulating virulence in Pseudomonas aeruginosa. Microbiology and Molecular Biology Reviews MMBR. *76* (1): 46–65.

Nakane, J.J., Akeson, M. and Marziali, A. (2003). Nanopore sensors for nucleic acid analysis. J. Phys. Condens. Matter. *15*: R1365–R1393.

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

NCBI Prokaryotic Genome Annotation Pipeline (PGAP):
https://github.com/ncbi/pgap

Nessar, R., Cambau, E., Reyrat, J.M, Murray, A. and Gicquel, B. (2012). Mycobacterium abscessus: A New Antibiotic Nightmare. J Antimicrob Chemother. *67* (4): 810-818.

Nessar, R., Reyrat, J-M., Davidson, L.B. and Byrd, T.F. (2011). 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. 157 (4): 1187–1195.

Neubert, K., Zuchantke, E., Leidenfrost, R.M., Wünschiers, R., Grützke, J., Malorny, B., Brendebach, H., Al Dahouk, S., Homeier, T., Hotzel, H., Reinert, K., Tomaso, H. and Busch, A. (2021). Testing assembly strategies of *Francisella tularensis* genomes to infer an evolutionary conservation analysis of genomic structures. BMC Genomics. *22*: 822.

Ng, H.F. and Ngeow, Y.F. (2022). Genetic Determinants of Tigecycline Resistance in *Mycobacteroides abscessus*. Antibiotics. 11: 572.

Nguyen, T. Q., Heo, B. E., Park, Y., Jeon, S., Choudhary, A., Moon, C. and Jang, J. (2023). CRISPR Interference-Based Inhibition of MAB_0055c Expression Alters Drug Sensitivity in *Mycobacterium abscessus*. Microbiology Spectrum. *11* (3): e0063123.

Nie, W., Duan, H., Huang, H., Lu, Y., Bi, D. and Chu, N. (2014). Species Identification of Mycobacterium Abscessus Subsp. Abscessus and Mycobacterium Abscessus Subsp. Bolletii Using RpoB and hsp65, and Susceptibility Testing to Eight Antibiotics. Int. J. Infect. Dis. *25*: 170–174.

Nogueira, L.C., Simmon, K.E., Chimara, E., Cnockaert, M., Palomino, J.C., Martin, A., Vandamme, P., Brown-Elliott, B.A., Wallace, R.J. and Leao, S.C. (2015a). *Mycobacterium franklinii* sp. nov., a species closely related to members of the Mycobacterium chelonae-Mycobacterium abscessus group. Int J Syst Evol Microbiol. *65* (7):2148-2153.

Nogueira C.L., Whipps, C.M., Matsumoto, C.K., Chimara, E., Droz, S., Tortoli, E., de Freitas, D., Cnockaert, M., Palomino, J.C., Martin, A., Vandamme, P. and Leão, S.C. (2015b). Mycobacterium saopaulense sp. nov., a rapidly growing Mycobacterium closely related to members of the Mycobacterium chelonae – Mycobacterium abscessus group. Int J Syst Evol Microbiol. *65* (12): 4403-4409.

Norden, A. and Linell, F. (1951). A new type of pathogenic Mycobacterium. Nature. *168*: 826.

Nouioui, I., Carro, L., García-López, M., Meier-Kolthoff, J.P., Woyke, T., Kyrpides, N.C., Pukall, R., Klenk, H-P., Goodfellow, M. and Göker, M. (2018). Genome-Based Taxonomic Classification of the Phylum Actinobacteria. Frontiers in Microbiology. *Vol 9*, Article 2007.

O'Brien, R.J., Geiter, L.J. and Snider, D.E. Jr. (1987). The epidemiology of non-tuberculous mycobacterial disease in the United States. Am Rev Respir Dis. *135*: 1007-14.

Olivier, K. N. (1998). Nontuberculous mycobacterial pulmonary disease. Curr. Opin. Pulm. Med. *4*:148-153.

Oren A. and Garrity, G.M. (2017). Notification that new names of prokaryotes, new combinations and new taxonomic opinions have appeared in volume 66, part 11, of the IJSEM.

Orens, J. B., Estenne, M., Arcasoy, S., Conte, J. V., Corris, P., Egan, J. J., Egan, T., Keshavjee, S., Knoop, C., Kotloff, R., Martinez, F. J., Nathan, S., Palmer, S., Patterson, A., Singer, L., Snell, G., Studer, S., Vachiery, J. L., Glanville, A. R. and Pulmonary Scientific Council of the International Society for Heart and Lung Transplantation. (2006). International guidelines for the selection of lung transplant candidates: 2006 update--a consensus report from the Pulmonary Scientific Council of the International Society for Heart and Lung Transplantation. The Journal of heart and lung transplantation : the official publication of the International Society for Heart Transplantation. *25* (7): 745–755.

Overbeek, R., Begley, T., Butler, R.M., Choudhuri, J.V., Chuang, H.Y., Cohoon, M., de Crécy-Lagard, V., Diaz, N., Disz, T., Edwards, R., Fonstein, M., Frank, E.D., Gerdes, S., Glass, E.M., Goesmann, A., Hanson, A., Iwata-Reuyl, D., Jensen, R., Jamshidi, N., Krause, L., Kubal, M., Larsen, N., Linke, B., McHardy, A.C. Meyer, F., Neuweger, H., Olsen, G., Olson, R., Osterman, A., Portnoy, V., Pusch, G.D., Rodionov, D.A., Rückert, C., Steiner, J., Stevens, R., Thiele, I., Vassieva, O., Ye, Y., Zagnitko, O. and Vonstein, V. (2005). The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. Nucleic Acids Research. *33* (17): 5691-702.

Overbeek, R., Olson, R., Pusch, G.D., Olsen, G.J., Davis, J.J., Disz, T., Edwards, R.A., Gerdes, S., Parrello, B., Shukla, M., Vonstein, V., Wattam, A.R., Xia, F. and Stevens, R. (2013). The SEED

and the rapid annotation of microbial genomes using subsystems technology (RAST). Nucleic Acids Research. *42* (1): D206-214.

Ozsolak, F., (2012). Third-generation sequencing techniques and applications to drug discovery. Expert Opin Drug Discov. *7* (3): 231-243.

Pace, N. R. (2009). Mapping the tree of life: progress and prospects. Microbiol. Mol. Biol. Rev. *73*: 565–576.

Pang Y., Zheng, H., Tan, Y., Song, Y. and Zhao, Y. (2017). In vitro activity of bedaquiline against nontuberculous mycobacteria in China. Antimicrob Agents Chemother. *61*: e02627-16.

PPanGGOLIN: https://github.com/labgem/PPanGGOLiN

Park, J., Cho, J., Lee, C-H., Han, S.K. and Yim, J-J. (2017). Progression and treatment outcomes of lung disease caused by *Mycobacterium abscessus* and *Mycobacterium Massiliense*. Clin Infect Dis. *64*: 301-308.

Park, M., Lalvani, A., Satta, G. and Kon, O.M. (2022). Evaluating the clinical impact of routine whole genome sequencing in tuberculosis treatment decisions and the issue of isoniazid mono-resistance. BMC Infect Dis. *22:* 349.

Park, S. C., Lee, K., Kim, Y.O., Won, S. and Chun, J. (2019). Large-Scale Genomics Reveals the Genetic Characteristics of Seven Species and Importance of Phylogenetic Distance for Estimating Pan-Genome Size. Frontiers in Microbiology. *10*: 834.

Parks, D. H., Chuvochina, M., Chaumeil, P.A., Rinke, C., Mussig, A.J. and Hugenholtz, P. (2020). A complete domain-to-species taxonomy for bacteria and archaea. Nature Biotechnology. *38* (9): 1079–1086.

Parra, J., Marcoux, J., Poncin, I., Canaan, S., Herrmann, J.L., Nigou, J., Burlet-Schiltz, O. and Rivière, M. (2017). Scrutiny of Mycobacterium tuberculosis 19 kDa antigen proteoforms

provides new insights in the lipoglycoprotein biogenesis paradigm. Scientific reports. *7*: 43682. https://doi.org/10.1038/srep43682

Parte, A. C., Sardà Carbasse, J., Meier-Kolthoff, J.P., Reimer, L.C. and Göker, M. (2020). List of Prokaryotic names with Standing in Nomenclature (LPSN) moves to the DSMZ. Int. J. Syst. Evol. Microbiol. *70* (11): 5607–5612.

Parvez, A., Giri, S., Giri, G. R., Kumari, M., Bisht, R. and Saxena, P. (2018). Novel Type III Polyketide Synthases Biosynthesize Methylated Polyketides in *Mycobacterium marinum*. Scientific Reports. *8*, 6529

Pasipanodya, J.G., Ogbonna, D., Ferro, B.E., Magombedze, G., Srivastava, S., Deshpande, D. and Gumbo, T. (2017). Systematic review and meta-analyses of the effect of chemotherapy on pulmonary *Mycobacterium abscessus* outcomes and disease recurrence. Antimicrob Agents Chemother. *61*: e01206-17.

Pavan, F., Chimara, E., Leite, C. Q .F., Arbeit, R.D., Sato, D.N. and de Carvalho, N. F. G. (2017). Genetic Correlates of Clarithromycin Susceptibility Among Isolates of the Mycobacterium Abscessus Group and the Potential Clinical Applicability of a PCR-based Analysis of *erm(*41). J. Antimicrob. Chemother. *73* (4): 862–866.

Pearce, M. E., Chattaway, M. A., Grant, K. and Maiden, M. C. J. (2020). A proposed core genome scheme for analyses of the Salmonella genus. Genomics. *112* (1): 371-378.

Peignier, A. and Parker, D. (2021). Impact of Type I Interferons on Susceptibility to Bacterial Pathogens. Trends in Microbiology. *29* (9): 823–835.

Pennisi, E. (2010). Semi-conductors inspire new sequencing technologies. Science. *327*: 1190.

Pesci, E.C., Milbank, J. B. J., Pearson, J.P., McKnight, S.L., Kende, A., Greenberg, E.P. and Iglewski, B. H. (1999). Quinolone signalling in the cell-to-cell communication system of Pseudomonas aeruginosa. *Proceedings of the National Academy of Sciences of the United States of America. 96* (20): 11229-34.

Pfeifer, B. A. and Khosla, C. (2001). Biosynthesis of polyketides in heterologous hosts. Microbiology and molecular biology reviews. *65* (1), 106–118.

Phillips, M. S. and von Reyn, C.F. (2001). Nosocomial infections due to nontuberculous mycobacteria. Clin. Infec. Dis. *33*:1363-1374.

Philley J.V., Wallace, R.J. Jr, Benwill, J. L., Taskar, V., Brown-Elliott, B.A., Thakkar, F., Aksamit, T.R. and Griffith, D.E. (2015). Preliminary results of bedaquiline as salvage therapy for patients with nontuberculous mycobacterial lung disease. Chest. *148*:499 –506.

Pizza, M., Scarlato, V., Masignani, V., Giuliani, M.M., Aricò, B., Comanducci, M., Jennings, G.T., Baldi, L., Bartolini, E., Capecchi, B., Galeotti, C.L., Luzzi, E., Manetti, R., Marchetti, E., Mora, M., Nuti, S., Ratti, G., Santini, L., Savino, S., Scarselli, M., Storni, E., Zuo, P., Broeker, M., Hundt, E., Knapp, B., Blair, E., Mason, T., Tettelin, H., Hood, D. W., Jeffries, A.C., Saunders, N.J., Granoff, D.N., Venter, J.C., Moxon, E.R., Grandi, G. and Rappuoli, R. (2000). Identification of vaccine candidates against serogroup B meningococcus by whole-genome sequencing. Science (New York, N.Y.) *287* (5459): 1816–1820.

Powell, S., Szklarczyk, D., Trachana, K., Roth, A., Kuhn, M., Muller, J., Arnold, R., Rattei, T., Letunic, I., Doerks, T., Jensen, L.J., von Mering, C. and Bork, P. (2012). eggNOG v3.0: Orthologous groups covering1133 organisms at 41 different taxonomic ranges. Nucleic Acids Research. *40*: 284–289.

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

Primm, T.P., Lucero, C.A. and Falkinham, J.O. III. (2004). Health Impacts of Environmental Mycobacteria. Clin Microbiol Rev. *17*: 98-106.

Pritchard, L., Glover, R.H., Humphris, S., Elphinstone, J.G. and Toth, I.K. (2016). Genomics and taxonomy in diagnostics for food security: soft-rotting enterobacterial plant pathogens. *Anal. Methods.* ***8*****:** 12-24.

proovread: Hybrid error correction tool: Imperial College research computing https://anaconda.org/imperial-college-research-computing/proovread

Pyani: https://github.com/widdowquinn/pyani

Qian J., Chen, R., Wang, H. and Zhang, X. (2020). Role of the PE/PPE Family in Host–Pathogen Interactions and Prospects for Anti-Tuberculosis Vaccine and Diagnostic Tool Design. Frontiers in Cellular and Infection Microbiology. *10*: Article 594288.

Qin, Q-L., Xie, B-B., Zhang, X-Y., Chen, X-L., Zhou, B-C., Zhou, J., Oren, A. and Zhanga, Y-Z. (2014). A Proposed Genus Boundary for the Prokaryotes Based on Genomic Insights. Journal of Bacteriology. *196* (12): 2210 –2215.

Qinglan, W., Boshoff, H.M., Harrison, J.R., Ray, P.C., Green, R., Wyatt, P.G. and Barry, C.E. III. (2020). PE/PPE proteins mediate nutrient transport across the outer membrane of *Mycobacterium tuberculosis*. Science. *167* (6482): 1147-1151.

Quadri, L. E. (2014). Biosynthesis of mycobacterial lipids by polyketide synthases and beyond. Critical reviews in biochemistry and molecular biology. *49* (3): 179–211.

Quail, M.A., Smith, M., Coupland, P., Otto, T.D. Harris**,** S.R., Connor, T.R., Bertoni**,** A., Swerdlow, H.P. and Guet, Y. (2012). A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics *13*: 341-353.

Quan, T. P., Bawa, Z., Foster, D., Walker, T., Del Ojo Elias, C., Rathod, P., MMM Informatics Group, Iqbal, Z., Bradley, P., Mowbray, J., Walker, A. S., Crook, D. W., Wyllie, D. H., Peto, T. E. A. and Smith, E. G. (2018). Evaluation of Whole-Genome Sequencing for Mycobacterial Species Identification and Drug Susceptibility Testing in a Clinical Setting: a Large-Scale Prospective

Assessment of Performance against Line Probe Assays and Phenotyping. Journal of Clinical Microbiology. *56* (2): e01480-17.

Queval, C.J., Song, O.R., Carralot, J.P., Saliou, J.M., Bongiovanni, A., Deloison, G., Deboosère, N., Jouny, S., Iantomasi, R., Delorme, V., Debrie, A.S., Park, S.J. Gouveia, J.C., Tomavo, S., Brosch, R., Yoshimura, A., Yeramian, E. and Brodin, P. (2017). Mycobacterium tuberculosis Controls Phagosomal Acidification by Targeting CISH-Mediated Signaling. Cell Reports. *20* (13): 3188–3198.

Quick, J., Loman, N.J., Duraffour, S., Simpson, J., Severi, E., Cowley, L. et al. (2016). Real-time, portable genome sequencing for Ebola surveillance. Nature. *530* (7589): 228–232.
(please note this paper has contributions from a further 97 collaborators)

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/

RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA. http://www.rstudio.com/

Ratnatunga, C. N., Lutzky, V. P., Kupz, A., Doolan, D.L., Reid, D.W., Field, M., Bell, S.C., Thomson, R.M. and Miles, J.J. (2020). The Rise of Non-Tuberculosis Mycobacterial Lung Disease. Frontiers in Immunology. *11*: 303.

Reddy, T. B., Riley, R., Wymore, F., Montgomery, P., DeCaprio, D., Engels, R., Gellesch, M., Hubble, J., Jen, D., Jin, H., Koehrsen, M., Larson, L., Mao, M., Nitzberg, M., Sisk, P., Stolte, C., Weiner, B., White, J., Zachariah, Z. K., Sherlock, G., Galagan, J.E., Ball, C.A. and Schoolnik, G. K. (2009). TB database: an integrated platform for tuberculosis research. Nucleic Acids Research. *37* : 499–508. https://doi.org/10.1093/nar/gkn652

Rengarajan, J., Bloom, B.R. and Rubin, E.J. (2005). Genome-wide requirements for Mycobacterium tuberculosis adaptation and survival in macrophages. *Proc Natl Acad Sci U S A*. *102* (23):8327-8332.

Reygaert, W.C. (2018). An overview of the antimicrobial resistance mechanisms of bacteria. AIMS microbiology. *4* (3): 482–501.

Rhee, M. and Burns, M.A. (2006). Nanopore sequencing technology: research trends and applications. Trends in Biotechnology. *24* (*12*): 580-586.

Ribeiro, M. and Simões, M. (2019). Siderophores: A Novel Approach to Fight Antimicrobial Resistance. In: Arora, D., Sharma, C., Jaglan, S. and Lichtfouse, E. (eds) Pharmaceuticals from Microbes. Environmental Chemistry for a Sustainable World. Vol 28. Springer, Cham.

Richard, M., Gutiérrez, A.V. and Kremer, L. (2020). Dissecting *erm*(41)-mediated macrolide-inducible resistance in mycobacterium abscessus. Antimicrobial Agents and Chemotherapy. *64* (2): e01879-19.

Richter, M., and Rosselló-Móra, R. (2009). Shifting the genomic gold standard for the prokaryotic species definition. Proc Natl Acad Sci USA. *106*, (45): 19126-19131.

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

Ripoll, F., Pasek, S., Schenowitz, C., Dossat, C., Barbe, V., Rottman, M., Macheras, E., Heym, B., Herrmann, J.L., Daffé, M., Brosch, R., Risler, J-L. and Gaillard, J-L. (2009). Non Mycobacterial Virulence Genes in the Genome of the Emerging Pathogen *Mycobacterium abscessus*. PLoS ONE *4*: e5660.

Roberts, H.E., Lopopolo, M., Pagnamenta, A.T., Sharma, E., Parkes, D., Lorne Lonie, L., Freeman, C., Knight, S.J.L., Lunter, G., Dreau, H., Lockstone , H., Taylor, J.C., Schuh, A., Bowden, R. and Buck, D. (2021). Short and long-read genome sequencing methodologies for somatic variant detection; genomic analysis of a patient with diffuse large B-cell lymphoma. Sci Rep. *11*: 6408.

Rominski, A., Roditscheff, A., Selchow, P., Böttger, E.C. and Sander, P. (2017). Intrinsic rifamycin resistance of *Mycobacterium abscessus* is mediated by ADP-ribosyltransferase MAB_0591. J Antimicrob Chemother. *72*: 376–384.

Ross, A.J. (1960). *Mycobacterium salmoniphilum* sp. nov. from salmonid fishes. Am Rev Respir Dis. *81*: 241-250.

Roth, A., Fisher, M., Hamid, M.E., Michalke, S., Ludwig, W. and Mauch, H. (1998). Differentiation of phylogenetically related slowly growing mycobacteria based on 16S-23S rRNA gene internal transcribed spacer sequences. J Clin Microbiol. *36*: 139-147.

Rothberg, J.M., Wolfgang Hinz, W., Rearick, T.M., Schultz, J., Mileski, W., Davey, M., Leamon, J.H., Johnson, K., Milgrew, M.J., Edwards, M., Hoon, J., Simons, J.F., Marran, D., Myers, J.W., Davidson, J.F., Branting, A., Nobile, J.R., Puc, B.P., Light, D., Clark, T.A., Huber, M., Branciforte, J.T., Stoner, I. B., Cawley, S.E., Lyons, M., Fu, Y., Homer, N., Sedova, M., Miao, X., Reed, B., Sabina, J., Feierstein, E., Schorn, M., Alanjary, M., Dimalanta, E., Dressman, D., Kasinskas, R., Sokolsky, T., Fidanza, J.A., Namsaraev, E., McKernan, K.J., Williams, A., Roth, G.T. and Bustillo, J. (2011). An integrated semiconductor device enabling non-optical genome sequencing. Nature. *475*: 348–352.

Rottman, M., Catherinot, E., Hochedez, P., Emile, J-F., Casanova, J-L., Gaillard, J.L. and Soudais, C. (2007). Importance of T Cells, Gamma Interferon, and Tumor Necrosis Factor in Immune Control of the Rapid Grower Mycobacterium abscessus in C57BL/6 Mice Am. Soc. for Microbiol. *75* (12): 5898-5907.

Ruangkiattikul, N., Rys, D., Abdissa, K., Rohde, M., Semmler, T., Tegtmeyer, P.K., Kalinke, U., Schwarz, C., Lewin, A. and Goethe, R. (2019). Type I interferon induced by TLR2-TLR4-MyD88-TRIF-IRF3 controls Mycobacterium abscessus subsp. abscessus persistence in murine macrophages via nitric oxide. International journal of medical microbiology : IJMM. *309* (5): 307–318.

Ruiz-Perez, C.A., Conrad, R.E. and Konstantinidis, K.T. (2021). MicrobeAnnotator: a user-friendly, comprehensive functional annotation pipeline for microbial genomes. BMC Bioinformatics. *22*: 11

Runyon, E.H. (1958). Mycobacteria encountered in clinical laboratories. Leprosy Briefs *9*: 21.

Runyon, E.H. (1959). Anonymous mycobacteria in pulmonary disease. Med Clin North Am. *43*: 273-290.

Rusk, N. (2011). Torrents of sequence. Nat Meth. *8* (10): 44.

Sadaka, C., Ellsworth, E., Hansen, P.R., Ewin, R., Damborg, P. and Watts, J.L. (2018). Review on Abyssomicins: Inhibitors of the Chorismate Pathway and Folate Biosynthesis. *Molecules*. *23*: 1371. https://doi.org/10.3390/molecules23061371

Saitou, N. and Nei, M. (1987). The neighbour-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. *4*: 406-425.

Salsgiver, E.L., Fink, A.K., Knapp, E.A., LiPuma, J.J., Olivier, K.N., Marshall, B.C. and Saiman, L. (2016). Changing Epidemiology of the Respiratory Bacteriology of Patients with Cystic Fibrosis. Chest. *149*: 390–400.

Salzberg, S.L. (2019) .Next-generation genome annotation: we still struggle to get it right. Genome Biology. *20*: 92

Sampaio, J. L., Junior, D.N., de Freitas, D., Hofling-Lima, A.L., Miyashiro, K., Alberto, F.L. and Leao, S.C. (2006). An outbreak of keratitis caused by *Mycobacterium immunogenum.* J. Clin. Microbiol. *44*: 3201-3207.

Sanchez, S. and Demain, A.L. (2011). Secondary Metabolites in Comprehensive Biotechnology (Academic Press Second Edition), Pages 155-167.

Sander P. and Böttger, E.C. (1999). Mycobacteria: genetics of resistance and implications for treatment. Chemotherapy. *45*: 95-108.

Sanger, F., Nicklen, S. and Coulson, A.R. (1977). DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA. *74* (12): 5463-5467.

Sanguinetti, M., Ardito, F., Fiscarelli, E., La Sarda, M., D'Argenio, P., Ricciotti, G. and Fadda, G. (2001). Fatal pulmonary infection due to multidrug resistant *Mycobacterium abscessus* in a patient with cystic fibrosis. J Clin Microbiol. *39*: 816-819.

Sassi, M, and Drancourt, M. (2014). Genome analysis reveals three genomospecies in *Mycobacterium abscessus*. BMC Genomics. *15*: 359.

Sassetti, C. M. and Rubin, E.J. (2013). Genetic requirements for mycobacterial survival during infection. Proc Natl Acad Sci U S A. *100* (22):12989-12994.

Sax, H., Bloemberg, G., Hasse, B., Sommerstein, R., Kohler, P., Achermann, Y., Rössle, M., Falk, V., Kuster, S.P. Böttger, E.C. and Weber, R. (2015). Prolonged outbreak of *Mycobacterium chimaera* infection after open-chest heart surgery. Clin Infect Dis. *61*: 67-75.

Schatz, M.C., Delcher, A.L. and Salzberg, S.L. (2010). Assembly of large genomes using second-generation sequencing. Genome Res. *20*(9):1165-1173.

Schein, S.J., Colombini, M. and Finkelstein, A. (1976). Reconstitution in planar lipid bilayers of a voltage-dependent anion-selective channel obtained from paramecium mitochondria. J Membr Biol. *30* (2): 99-120.

Schinsky, M.F., McNeil, M.M., Whitney, A.M., Steigerwalt, A.G., Lasker, B.A., Floyd, M.M., Hogg, G.G., Brenner, D.J. and Brown, J.M. (2000). *Mycobacterium septicum* sp. nov., a new rapidly growing species associated with catheter-related bacteraemia. Int. J. Syst. Evol. Microbiol. *50*: 575-581.

Schinsky, M.F., Morey, R.E., Steigerwalt, A.G., Douglas, M.P., Wilson, R.W., Floyd, M.M., Butler, W.R., Daneshvar, M.I., Brown-Elliott, B.A., Wallace, R.J. Jr, McNeil, M.M., Brenner D.J. and Brown, J.M. (2004). Taxonomic variation in the *Mycobacterium fortuitum* third biovariant complex: description of *Mycobacterium boenickei* sp. nov., *Mycobacterium houstonense* sp. nov., *Mycobacterium neworleansense* sp. nov. and *Mycobacterium brisbanense* sp. nov. and recognition of *Mycobacterium porcinum* from human clinical isolates. Int. J. Syst. Evol. Microbiol. *54*: 1653-1667.

Schorey, J.S. and Sweet, L. (2008). The mycobacterial glycopeptidolipids: structure, function, and their role in pathogenesis. Glycobiology. *18* (11):832-41.

Schröder, K. H., and Juhlin, I.M. (1977). Mycobacterium malmoense sp. nov. *International Journal of Systematic and Evolutionary Microbiology. 27*: 241-246.

Schwarze K., Buchanan, J., Fermont, J.M., Dreau, H., Tilley, M.W., Taylor, J.M., Antoniou, P., Knight, S.J., Camps, C., Pentony, M.M., Kvikstad, E.M., Harris, S., Popitsch, N., Pagnamenta, A.T., Schuh, A., Taylor, J.C. and Wordsworth, S. (2019). The complete costs of genome sequencing: a microcosting study in cancer and rare diseases from a single center in the United Kingdom. Genetics in Medicine. *22* (1).

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. Bioinformatics (Oxford, England). *30* (14), 2068-2069.

Segerman, B. (2012). The genetic integrity of bacterial species: the core genome and the accessory genome, two different species. Frontiers in Cellular and Infection Microbiology. *2* (116)

Seib, K.L., Zhao, X. and Rappuoli, R. (2012). Developing vaccines in the era of genomics: a decade of reverse vaccinology. Clin Microbiol Infect. *18*:109-116.

Shaffer, M., Borton, M. A., McGivern, B. B., Zayed, A. A., La Rosa, S. L., Solden, L. M., Liu, P., Narrowe, A. B., Rodríguez-Ramos, J., Bolduc, B. and Wrighton, K. C. (2020). DRAM for distilling

microbial metabolism to automate the curation of microbiome function. Nucleic Acids Research. *48*: (16), 8883-8900.

Shah, N.M., Davidson, J.A., Anderson, L.F., Lalor, M.K., Kim, J., Thomas, H.L., Lipman, M. and Abubakar, I. (2016). Pulmonary Mycobacterium avium-intracellulare is the main driver of the rise in non-tuberculous mycobacteria incidence in England, Wales and Northern Ireland, 2007–2012. BMC Infectious Diseases. *16*:195.

Shah N.S., Wright, A., Bai, G.H., Barrera, L., Boulahbal, F., Martín-Casabona, N., Drobniewski, F., Gilpin, C., Havelková, M., Lepe, R., Lumb, R., Metchock, B., Portaels, F., Rodrigues, M.F., Rüsch-Gerdes, S., Van Deun, A., Vincent, V., Laserson, K., Wells, C. and Cegielski, J.P. (2007). Worldwide emergence of extensively drug-resistant tuberculosis. Emerg Infect Dis. *13* (3): 380-7.]

Sharrar, A. M., Crits-Christoph, A., Méheust, R., Diamond, S., Starr, E. P. and Banfield, J.F. (2020). Bacterial Secondary Metabolite Biosynthetic Potential in Soil Varies with Phylum, Depth, and Vegetation Type. mBio 11: e00416-20.

Shelton, B.G., Flanders, W.D. and Morris, G.K. (1999). *Mycobacterium* sp. as a possible cause of hypersensitivity pneumonitis in machine workers. Emerg. Infec. Dis. *5*: 270-273.

Shendure, J. and Ji, H. (2008). Next-generation DNA sequencing. Nature Biotechnology. *26*: 1135-1145.

Shimizu, Y., Ogata, H. and Goto, S. (2017). Type III Polyketide Synthases: Functional Classification and Phylogenomics. ChemBioChem. *18* (1): 50-65.

Shojaei, H., Goodfellow, M., Magee, J.G., Freeman, R., Gould, F.K. and Brignall, C.J. (1997). *Mycobacterium novocastrense* sp. nov., a rapidly growing photochromogenic *Mycobacterium*. Int J Syst Bacteriol. *47*: 1205-1207.

Shojaei, H., Magee, J.G., Freeman, R., Yates, M., Horadagoda, N.U. and Goodfellow, M. (2000). *Mycobacterium elephantis* sp. nov., a rapidly growing non-chromogenic *Mycobacterium* isolated from an elephant. J Syst Evol Microbiol. *50*: 1817-1820.

Sievert, C., Parmer, C., Hocking, T., Chamberlain, S.A., Ram, K., Corvellec M. and Despouy, P. (2020). "Create Interactive Web Graphics via 'plotly.js' [R package plotly version 4.9.2.1]. https://CRAN.R-project.org /package=plotly

Silcox, V.A., Good, R,C, and Floyd, M.M. (1981). Identification of clinically significant *Mycobacterium fortuitum* complex isolates. J Clin Microbiol. *14*: 686-691.

Silva Miranda, M., Breiman, A., Allain, S., Deknuyd, F. and Altare, F. (2012). The tuberculous granuloma: an unsuccessful host defence mechanism providing a safety shelter for the bacteria? Clinical & Developmental Immunology. Volume 2012: Article ID 139127.

Simons, S., van Ingen, J., Hsueh, P.R., Van Hung, N., Dekhuijzen, P.N.R., Boeree, M.J. and van Soolingen, D. (2011). Nontuberculous Mycobacteria in Respiratory Tract Infections, Eastern Asia. Emerging Infectious Diseases. *17*(3):343-349.

Simpson, J.T. and Durbin, R. (2012). Efficient de novo assembly of large genomes using compressed data structures. Genome Research. *22* (3): 549–556.

Singh, S. and S. Bhatia. 2021. Quorum Sensing Inhibitors: Curbing Pathogenic Infections through Inhibition of Bacterial Communication. Iran J Pharm Res. *20* (2):486-514.

Singh, H. B., Gupta, V. G. and Jogaiah, S. (Eds.). (2018). New and future developments in microbial biotechnology and bioengineering: Microbial genes biochemistry and applications. Elsevier.

Skerman, V.B.D., McGowan, V. and Sneath, P.H.A. (1980). Approved lists of bacterial names. Int J Syst Bacteriol. *30*: 225-420.

Smith, L.M., Sanders, J.Z., Kaiser, R.J., Hughes, P., Dodd, C., Connell, C.R., Heiner, C., Kent, S.B. and Hood, L.E. (1986). Fluorescence detection in automated DNA sequence analysis. Nature *321*:674–79.

Smyth, A. and Walters, S. (2003). Prophylactic anti-staphylococcal antibiotics for cystic fibrosis. Cochrane Database Syst Rev. *3*: CD001912.

Sondén, B., Kocíncová, D., Deshayes, C., Euphrasie, D., Rhayat, L., Laval, F., Frehel, C., Daffé, M., Etienne, G. and Reyrat, J.M. (2005). Gap, a mycobacterial specific integral membrane protein, is required for glycolipid transport to the cell surface. Molecular Microbiology. *58* (2):426-40.

Song, L., Hobaugh, M.R., Shustak, C., Cheley, S., Bayley, H. and Gouaux, J.E. (1996). Structure of staphylococcal a hemolysin, a heptameric transmembrane pore. Science. *274*: 1859-66.

Soroka, D., Dubée, V., Soulier-Escrihuela, O., Cuinet, G., Hugonnet, J.E., Gutmann, L., Mainardi, J-L. and Arthur, M. (2014). Characterization of broad-spectrum *Mycobacterium abscessus* class A β-lactamase. J Antimicrob Chemother. *69*: 691–696.

SPAdes Genome Assembler: https://github.com/ablab/spades
(Please also see Bankevich, A *et al* 2012)

Springer, B., Böttger, E.C., Kirschner, P. and Wallace, R.J. (1995). Phylogeny of the *Mycobacterium chelonae*-like organism based on partial sequencing of the 16S ribosomal RNA gene and proposal of *Mycobacterium mucogenicum* sp. nov. Int. J. Syst. Bacteriol. *45*: 262-267.

Sreevatsan, S., Stockbauer, K.E., Pan, X., Kreiswirth, B.N., Moghazeh, S.L., Jacobs, W.R. Jr, Telenti, A. and Musser, J.M. (1997). Ethambutol resistance in Mycobacterium tuberculosis: critical role of embB mutations. Antimicrob Agents Chemother. *41* (8):1677-81.

Stackebrandt, E. and Goebel, B.M. (1994). Taxonomic note; A place for DNA:DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology. Int J System Bacteriol. *44*: 846-849.

Stackebrandt, E., Rainey, F.E. and Ward-Rainey, N.L. (1997). Proposal for a new hierarchic classification system, *Actinobacteria* classis nov. Int J Syst Evol Microbiol. *47*: 479-491.

Stackebrandt, E., Frederiksen, W., Garrity, G.M., Grimont, P.A.D., Kämpfer, P., Maiden, M.C.J., Nesme, X., Rosselló-Mora, R., Swings, J., Trüper, H.G., Vauterin, L., Ward, A.C. and Whitman, W.B. (2002). Report of the ad hoc committee for the re-evaluation of the species definition in bacteriology. Int J Syst Evol Micro. *52*: 1043-1047.

Stanford, J.L. and Grange, J.M. (1974). The meaning and structure of species as applied to mycobacteria. Tubercle. *55*: 143-152.

Stewart, P. S., and Costerton, J.W. (2001). Antibiotic resistance of bacteria in biofilms. Lancet. *358* :135 -138

Stewart, P.S. (2002). Mechanisms of antibiotic resistance in bacterial biofilms. Int. J. Med. Microbiol. *292*: 107-113

Stinear, T.P., Seemann, T., Pidot, S.J., Frigui, W., Reysset, G., Garnier, T., Meurice, G., Simon, D., Bouchier, C., Ma, L., Tichit, M., Porter, J.L., Ryan, J., Johnson, P.D.R., Davies, J.K., Jenkin, G.A., Small, P.L.C., Jones, L.M., Tekaia, F., Laval, F., Daffe, M., Parkhill, J. and Cole, S.T. (2007). Reductive evolution and niche adaptation inferred from the genome of Mycobacterium ulcerans, the causative agent of Buruli ulcer. *Genome Research. 17* (2): 192-200.

Stoker, N. G., Fairweather, N.F. and Spratt, B.G. (1982). Versatile low-copy-number plasmid vectors for cloning in *Escherichia coli*. Gene. *18*: 335–341.

Storm, A.J., Chen, J.H., Ling, X.S., Zandbergen, H.W. and Dekker, C. (2003). Fabrication of solid-state nanopores with single-nanometre precision. Nature Materials. *2*: 537-540.

Story-Roller, E., Maggioncalda, E.C., Cohen, K.A. and Lamichhane, G. (2018). *Mycobacterium abscessus* and β-Lactams: Emerging Insights and Potential Opportunities. Front Microbiol. *9*: 2273.

Stout, J. E., Koh, W.J. and Yew, W.W. (2016). Update on Pulmonary Disease Due to non-Tuberculous Mycobacteria. Int. J. Infect. Dis. *45*: 123–134.

Sueoka, N. (1962). On The Genetic Basis of Variation and Heterogeneity of DNA Base Composition. Proceedings of the National Academy of Sciences of the United States of America. *48*: 582-592.

Sullam, P.M., Gordin, F.M., Wynne, B.A. and the rifabutin treatment group. (1994). Efficacy of rifabutin in the treatment of disseminated infection due to *Mycobacterium avium* complex. Clin Infect Dis *19*: 84-86.

Sun, L., Wang, S., Zhang, S., Yu, D., Qin, Y., Huang, H., Wang, W. and Zhan, J. (2016). Identification of a type III polyketide synthase involved in the biosynthesis of spirolaxine. Applied Microbiol Biotechnol. *100* (16): 7103-7113.

Szomolay, B., Klapper, I., Dockery, J. and Stewart, P.S. (2005). Adaptive responses to antimicrobial agents in biofilms. Environmental Microbiology. *7* (8): 1186–1191.

Ta, P., Buchmeier, N., Newton, G.L., Rawat, M. and Fahey, R.C. (2011). Organic hydroperoxide resistance protein and ergothioneine compensate for loss of mycothiol in Mycobacterium smegmatis mutants. Journal of Bacteriology. *193* (8): 1981–1990.

Tan, L.T-H., Raghunath, P., Ming, L.C. and Law, J.W-F. (2020) Mycobacterium ulcerans and Mycobacterium marinum: Pathogenesis, Diagnosis and Treatment. Prog Microbes Mol Biol. *3* (1): a0000114.

Tatusova, T., DiCuccio, M., Badretdin, A., Chetvernin, V., Nawrocki, E.P., Zaslavsky, L., Lomsadze, A., Pruitt, K.D., Borodovsky, M. and Ostell, J. (2016). NCBI prokaryotic genome annotation pipeline. Nucleic acids research. *44* (14): 6614–6624.

Telenti A. (1998). More on “what’s in a name …” – pragmatism in mycobacterial taxonomy. Int J Tuberc Lung Dis. *2*: 182–183.

Tettelin, H., Masignani, V., Cieslewicz, M.J., Donati, C., Medini, D., Ward, N.L., Angiuoli, S.V., Crabtree, J., Jones, A.L., Durkin, A.S., Deboy, R.T., Davidsen, T.M., Mora, M., Scarselli, M., Margarit y Ros, I., Peterson, J.D., Hauser, C.R., Sundaram, J.P., Nelson, W.C. Madupu, R., Brinkac, L.M., Dodson, R.J. Rosovitz, M.J., Sullivan, S.A., Daugherty, S.C., Haft, D.H., Selengut, J., Gwinn, M.L., Zhou, L., Zafar, N., Khouri, H., Radune, D., Dimitrov, G., Watkins, K., O'Connor, K.J., Smith, S., Utterback, T.R., White, O., Rubens, C.E., Grandi, G., Madoff, L.C., Kasper, D.L., Telford, J.L., Wessels, M.R., Rappuoli, R. and Fraser, C.M. (2005). Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". Proc Natl Acad Sci U S A. *102* (39):13950-5.

Tettelin, H., Riley, D., Cattuto, C. and Medini, D. (2008). Comparative genomics: the bacterial pan-genome. Current Opinions in Microbiology. *12*:472-477.

Tettelin, H., Davidson, R.M., Agrawal, S., Aitken, M.L., Shallom, S., Hasan, N.A., Strong, M., Nogueira de Moura, V.C., De Groote, M.A., Duarte, R.S., Hine, E., Parankush, S., Su, Q., Daugherty, S.C., Fraser, C.M., Brown-Elliott, B.A., Wallace, R.J., Holland, S.M., Sampaio, E.P., Olivier, K.N., Jackson, M. and Zelazny, A.M. (2014). High-level relatedness among Mycobacterium abscessus subsp. massiliense strains from widely separated outbreaks. Emerg Infect Dis. *20*: 364–371.

Thavagnanam, S., McLoughlin, L.M., Hill, C. and Jackson, P.T. (2006). Atypical Mycobacterial Infections in Children: The Case for Early Diagnosis. Ulster Med J. *75*: 192-194.

The International HapMap Consortium: A haplotype map of the human genome. (2005). Nature. *437*:1299–1320.

The International HapMap3 Consortium: Integrating common and rare genetic variation in diverse human populations. (2010). Nature. *467*:52–58.

The Pew Charitable Trusts. 2019-11-22. "How We Work".

The 1000 Genomes Project Consortium: An integrated map of genetic variation from 1,092 human genomes. (2012). Nature. *491*:56–65.

Thermofisher. (2023).
https://www.thermofisher.com/uk/en/home/life-science/sequencing/next-generation-sequencing/ion-torrent-next-generation-sequencing-technology.html

Thompson, C.C., Vieira, N.M., Vicente, A. and Thompson, F. (2011). Towards a genome based taxonomy of Mycoplasmas. Infect Genet Evol. *11*:1798–1804.

Thompson, C.C., Emmel, V.E., Fonseca, E.L., Marin, M.A. and Vicente, A.C.P. (2013). Streptococcal taxonomy based on genome sequence analyses. F1000Research *67*:1–9.

Thomson, R., Tolson, C., Sidjabat, H., Huygens, F. and Hargreaves, M. (2013). *Mycobacterium abscessus* isolated from municipal water – a potential source of human infection. BMC Infectious Diseases. *13*: 241.

Timpe, A. and Runyon, E.H. (1954). The relationship of “atypical” acid-fast bacteria to human disease; a preliminary report. J Lab Clin Med. *44*: 202-209.

Tkachenko, A. G., Akhova, A. V., Shumkov, M. S. and Nesterova, L. Y. (2012). Polyamines reduce oxidative stress in *Escherichia coli* cells exposed to bactericidal antibiotics. Research in Microbiology. *163*(2), 83–91.

Tonkin-Hill, G., MacAlasdair, N., Ruis, C., Weimann, A., Horesh, G., Lees, J.A., Gladstone, R.A., Lo, S., Beaudoin, C., Floto, R.A., Frost, S.D.W., Corander, J., Bentley, S.D. and Parkhill, J. (2020). Producing polished prokaryotic pangenomes with the Panaroo pipeline. *Genome Biology*. *21* (1): 180.

Tortoli. E., Baruzzo, S., Heijdra, Y., Klenk, H.P., Lauria, S., Mariottini, A. and van Ingen, J. (2009). *Mycobacterium insubricum* sp. nov. Int J Syst Evol Microbiol. *59*: 1518-1523.

Tortoli E., Brown-Elliott, B.A., Chalmers, J.D., Cirillo, D.M., Daley, C.L., Emler, S., Floto, R.A., Garcia, M.J., Hoefsloot, W., Koh, W.J., Lange, C., Loebinger, M., Maurer, F.P., Morimoto, K., Niemann, S., Richter, E., Turenne, C.Y., Vasireddy, R., Vasireddy, S., Wagner, D., Wallace, R.J. Jr, Wengenack, N. and van Ingen, J. (2019). Same meat, different gravy: ignore the new names of mycobacteria. Eur Respir J. *54*:1900795.

Tortoli, E., Fedrizzi, T., Meehan, C.J., Trovato, A., Grottola, A., Giacobazzi, E., Serpini, G.F., Tagliazucchi, S., Fabio, A., Bettua, C., Bertorelli, R., Frascaro, F., De Sanctis, V., Pecorari, M., Jousson, O., Segata N. and Cirillo, D.M. (2017). The new phylogeny of the genus Mycobacterium: The old and the news. Infection, Genetics and Evolution *56*: 19–25.

Tortoli, E., Kohl, T.A., Brown-Elliott, B.A., Trovato, A., Leao, S.C., Garcia, M.J., Vasireddy, S., Turenne, C.Y., Griffith, D.E., Philley, J.V., Baldan, R., Campana, S., Cariani, L., Colombo, C., Taccetti, G., Teri, A., Niemann, S., Wallace, R.J. Jr. and Cirillo, D.M. (2016). 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. *66*: 4471-4479.

Tortoli, E., Pecorari, M., Fabio, G., Messinò, M. and Fabio, A. (2010). Commercial DNA probes for mycobacteria incorrectly identify a number of less frequently encountered species. Journal of Clinical Microbiology. *48* (1): 307–310.

Tran, T.T., Munita J.M. and Arias, C.A. (2015). Mechanisms of drug resistance: daptomycin resistance. Ann N Y Acad Sci. *1354*:32-53.

Treangen, T.J. and Rocha, E.P.C. (2011). Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. PLoS genetics. *7* (1): e1001284.

Trivedi, O. A., Arora, P., Vats, A., Ansari, M.Z., Tickoo, R., Sridharan, V., Mohanty, D. and Gokhale, R. S. (2005). Dissecting the Mechanism and Assembly of a Complex Virulence Mycobacterial Lipid. Molecular Cell. *17* (5): 631-643.

Tyler, A. D., Mataseje, L., Urfano, C.J., Schmidt, L., Antonation, K.S., Mulvey, M.R. and Corbett, C.R. (2018). Evaluation of Oxford Nanopore's MinION Sequencing Device for Microbial Whole Genome Sequencing Applications Scientific reports. *8* (1),10931.

Tuberculosis Chemotherapy Centre, Madras. (1966). Isoniazid plus thiacetazone compared with two regimens of Isoniazid plus PAS in the domiciliary treatment of pulmonary tuberculosis in South Indian patients. Bull WHO. *34*: 483-515.

Tufariello, J.M., Chapman, J.R., Kerantzas, C.A., Wong, K.W., Vilchèze, C., Jones, C.M., Cole, L.E., Tinaztepe, E., Thompson, V., Fenyö, D., Niederweis, M., Ueberheide, B., Philips, J.A. and Jacobs, W.R. Jr. (2016). Separable roles for Mycobacterium tuberculosis ESX-3 effectors in iron acquisition and virulence. Proc Natl Acad Sci USA. *113*(3): E348–E357.

UKHSA TB Action Plan for England 2021-2026. (2021). PHE publications gateway number: GOV-7952.
https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/998158/TB_Action_Plan_2021_to_2026.pdf

UKHSA *Mycobacterium chimaera*: Infections linked to heater cooler units – GOV.UK (www.gov.uk)

Unicycler: Hybrid assembly pipeline for bacterial genomes
https://bioconda.github.io/recipes/Unicycler/README.html

UniProt: The universal protein knowledgebase in 2021. (2021). Nucleic Acids Research. Oxford University Press. 49: 480-489.

van Dijk, E., Auger, H., Jaszczyszyn, Y. and Thermes, C. (2014). Ten years of next-generation sequencing technology, Trends in Genet. *30* (9): 418-426.

van Ingen, J., Bendien, S. A., de Lange, W. C., Hoefsloot, W., Dekhuijzen, P. N., Boeree, M. J. and van Soolingen, D. (2009). Clinical relevance of non-tuberculous mycobacteria isolated in

the Nijmegen-Arnhem region, The Netherlands. Thorax. *64* (6): 502–506. https://doi.org/10.1136/thx.2008.110957

van Ingen, J., Boeree, M.J., Klosters, K., Wieland, A., Tortoli, E., Dekhuijzen, P.N.R. and van Sooligen, D. (2009). Proposal to elevate *Mycobacterium avium* complex ITS sequevar MAC-Q to *Mycobacterium vulneris* sp. nov. Int J Syst Evol Microbiol. *59*: 2277-2282.

van Ingen, J., de Zwaan, R., Enaimi, M., Dekhuijzen, P. N., Boeree, M. J. and van Soolingen, D. (2010). Re-analysis of 178 previously unidentifiable Mycobacterium isolates in the Netherlands in 1999-2007 Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases. *16* (9): 1470–1474.

Vandenesch, F., Bes, M., Lebeau, C., Greenland, T., Brun, Y. and Etienne, J. (1993). Coagulase-negative Staphylococcus aureus. Lancet (London, England). *342* (8877): 995–996.

Varghese, N. J., Mukherjee, S., Ivanova, N., Konstantinidis, K.T., Mavrommatis, K., Kyrpides, N.C. and Pati, A. (2015). Microbial species delineation using whole genome sequences. Nucleic acids research. *43* (14): 6761–6771.

Vaser, R., Sovic, I., Nagarajan, N. and Sikic, M. Fast and accurate de novo genome assembly from long uncorrected reads. (2017). Genome Res. *27*: 737-746

Velsko, I.M., Perez, M.S. and Richards, V.P. (2019). Resolving phylogenetic relationships for Streptococcus mitis and Streptococcus oralis through core- and pan-genome analyses. Genome Biol Evol *11*:1077-1087.

Victoria, L., Gupta, A., Gomez, J.L. and Robledo, J. (2021). Mycobacterium abscessus complex: A Review of Recent Developments in an Emerging Pathogen. Front. Cell. Infect. Microbiol. *11*:659997.

Voskuil, M.I., Bartek, I.L., Viscontia, K. and Schoolnik, G.K. (2011). The response of Mycobacterium tuberculosis to reactive oxygen and nitrogen species. Frontiers in Microbiology. *2*: 105.

Wainwright, B.J., Scambler, P.J., Schmidtke, J., Watson, E.A., Law, H-Y., Farrall, M., Cooke, H.J., Eiberg, H. and Williamson, R. (1985). Localisation of cystic fibrosis locus to human chromosome 7cen-q22. Nature. *318*: 384-385.

Walker, J., Moore, G., Collins, S., Parks, S., Garvey, M.I., Lamagni, T., Smith, G., Dawkin, L., Goldenberg, S. and Chand, M. (2017). Microbiological problems and biofilms associated with *Mycobacterium chimaera* in heater-cooler units used for cardiopulmonary bypass. J Hosp Infect. *96*: 209-220.

Wallace, R.J. Jr., Swenson, J.M., Silcox, V.A., Good, R.C., Tischen, J.A. and Stone, M.S. (1983). Spectrum of disease due to rapidly growing mycobacteria. Rev Inf Dis. *5*: 657-679.

Wallace, R.J. Jr., O'Brien, R., Glassroth, J., Raleigh, J. and Dutt, A. (1990). Diagnosis and treatment of disease caused by nontuberculous mycobacteria. NTM statement, American Thoracic Society.

Wallace, R.J., Silcox, V.A., Tsukamura, M., Brown, B.A., Kilburn, J.O., Butler, W.R. and Onyi, G. (1993). Clinical significance, biochemical features, and susceptibility patterns of sporadic isolates of the *Mycobacterium chelonae*-like organism. J. Clin. Microbiol. *31*: 3231-3239.

Wallace, R.J. Jr., Brown, B.A., Griffith, D.E., Girard, W.M., Murphy, D.T., Onyi, G.O., Steingrube, V.A and Mazurek, G.H. (1994). Initial clarithromycin monotherapy for *Mycobacterium avium-intracellulare* complex lung disease. Am R Respir Crit Care Med. *149*: 1335-1341.

Wallace, R.J., Jr., Meier, A., Brown, B.A., Zhang, Y., Sander, P., Onyi, G.O. and Böttger, E.C.. (1996). Genetic basis for clarithromycin resistance among isolates of Mycobacterium chelonae and Mycobacterium abscessus. Antimicrob. Agents Chemother. *40*:1676–168.

Wallace, R. J., Jr., Brown, B.A. and Griffith, D.E. (1998). Nosocomial outbreaks/pseudo-outbreaks caused by nontuberculous mycobacteria. Annu. Rev. Microbiol. *52*: 453-490.

Wallace, R.J. (2001). *Mycobacterium immunogenum* sp. nov., a novel species related to *Mycobacterium abscessus* and associated with clinical disease, pseudo-outbreaks and contaminated metalworking fluids: an international cooperative study on mycobacterial taxonomy. Int. J. Syst. Evol. Microbiol. *51*: 1751–1764.

Walsh C. (2000). Molecular mechanisms that confer antibacterial drug resistance. Nature. *406* (6797):775-781.

Wang, F., Langley, R., Gulten, G., Wang, L. and Sacchettini, J.C. (2007). Identification of a Type III Thioesterase Reveals the Function of an Operon Crucial for Mtb Virulence. Chemistry & Biology. *14*, (5): 543-551.

Ward, A.C. and Kim, W. (2015). MinIONTM: New, Long Read, Portable Nucleic Acid Sequencing Device. J Bacteriology and Virology. *45* (4): 285 **–** 303.

Wayne, L.G. (2000). A slow ramble in the acid-fast lane: the coming of age of mycobacterial taxonomy. In: Applied microbial Systematics (edited by Priest & Goodfellow). Springer, New York, pp. 389-420.

Wayne, L.G., Brenner, D.J., Colwell, R.R., Grimont, P.A.D., Kandler, P., Krichevsky, M.I., Moore, L.H., Moore, W.E.C., Murray, R.G.E., Stackebrandt, E., Starr, M.P. and Trüper, H.G. (1987). Report of the *ad hoc* committee on reconciliation of approaches to bacterial systematics. Int J Syst Bact. *37*: 463-464.

Wayne, L.G., Good, R.C., Böttger, E.C., Butler, R., Dorsch, M., Ezaki, T., Gross, W., Jonas, V., Kilburn, J., Kirschner, P., Krichevsky, M.J., Ridell, M., Shinnick, T.M., Springer, B., Stackebrandt, E., Tárnok, I., Tasaka, H., Vincent, V., Warren, N.G., Knott, C.A. and Johnson, R. (1996). Semantide- and chemotaxonomy-based analyses of some problematic phenotypic clusters of slowly-growing mycobacteria, a cooperative study of the International Working Group on Mycobacterial taxonomy. Int J Syst Bacteriol. *46*: 280-297.

Wayne, L.G. and Kubica, G. (1986). Genus, Mycobacterium. In Bergey's Manual of Systematic Bacteriology, vol. 2 (edited by Sneath, Mair, Sharpe & Holt). Williams & Wilkins, Baltimore, pp. 1436-1457.

Wick, R.R. and Holt, K.E. (2021). Benchmarking of long-read assemblers for prokaryote whole genome sequencing. F1000Research. *8*: 2138

Wickham, H. (2009). ggplot2: Elegant Graphics for Data Analysis. https://books.google.co.uk/books?id=XgFkDAAAQBAJ

Whipps, C.M., Butler, W.R., Pourahmad, F., Watral, V.G. and Kent, M.L. (2007). Molecular systematics support the revival of Mycobacterium salmoniphilum (ex. Ross 1960) sp. nov., nom. rev., a species closely related to Mycobacterium chelonae. Int. J. Syst. Evol. Microbiol. *57*: 2525–2531.

Wick, R.R., Judd, L.M., Cerdeira, L.T., Hawkey, J., Méric, G., Vezina, B., Wyres, K.L. and Holt, K.E. (2021). Trycycler: consensus long-read assemblies for bacterial genomes. Genome Biology. *22*: 266.

Wick, R.R., Judd, L.M., Gorrie, C.L. and Holt, K.E. (2017). Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. PLoS Comput Biol. *13*(6):e1005595.

Wilson, R.W., Steingrube, V.E., Bottger, E.C., Springer, B., Brown-Elliott, B.A., Vincent, V., Jost, K.C., Zhang, Y.S., Garcia, M.J., Chiu, S.H., Onyi, G.O., Rossmoore, H., Nash, D.R. and Wallace, R.J. Jr. (2001). Mycobacterium immunogenum sp. nov., a novel species related to Mycobacterium abscessus and associated with clinical disease, pseudo-outbreaks and contaminated metalworking fluids: an international cooperative study on mycobacterial taxonomy. Int J Syst Evol Microbiol. *51*: 1751–1764.

Wirth, T., Hildebrand, F., Allix-Béguec, C., Wölbeling, F., Kubica, T., Kremer, K., van Soolingen, D., Rüsch-Gerdes, S., Locht, C., Brisse, S., Meyer, A., Supply, P. and Niemann, S. (2008). Origin,

Spread and Demography of the Mycobacterium tuberculosis Complex. PLoS Pathogens. *4* (issue 9) e1000160.

Wiryaman, T. and Toor, N. (2022). Recent advances in the structural biology of encapsulin bacterial nano compartments. J Struct Biol X. *6*: 100062.

Witek K., Jupe, F., Witek, A.I., Baker, D., Clark, M.D. and Jones, J.D.G. (2016). Accelerated cloning of a potato late blight–resistance gene using RenSeq and SMRT sequencing. Nature Biotechnology. *34* (6).

Withrop, K.L., Abrams, M., Yakrus, M.A., Schwartz, I., Ely, J., Gillies, D and Vugia, D.J. (2002). An outbreak of mycobacterial furunculosis associated with footbaths at a nail salon. N. Engl. J. Med. *346*: 1366-1371.

Woese, C.R. (1987). Bacterial evolution. Microbiol. Rev. *51*: 221-271.

Wong, D., Bach, H., Sun, J., Hmama, Z. and Av-Gay, Y. (2011). Mycobacterium tuberculosis protein tyrosine phosphatase (PtpA) excludes host vacuolar-H+-ATPase to inhibit phagosome acidification. Proc Natl Acad Sci USA. *108* (48):19371–19376.

Wood, D.M. and Smyth, A.R. (2006). Antibiotic strategies for eradicating *Pseudomonas aeruginosa* in people with cystic fibrosis. Cochrane Database Syst Rev *1*: CD004197.

World Health Organization. (1993). Treatment of tuberculosis. Guidelines for National Programmes. World Health Organization, Geneva, Switzerland.
Treatment of tuberculosis (who.int)

World Health Organisation. (2004). Pathogenic Mycobacteria in Water: A guide to Public Health Consequences, Monitoring and Management (edited by S Pedley, J Bartram, G Rees, A Dufour and J Cotruvo). IWA Publishing, London, UK.
https://apps.who.int/iris/rest/bitstreams/471465/retrieve

World Health Organization. (2014). Companion handbook to the WHO guidelines for the programmatic management of drug-resistant tuberculosis. World Health Organization, Geneva, Switzerland.
Companion handbook to the WHO guidelines for the programmatic management of drug-resistant tuberculosis

World Health Organization. (2021). Global Tuberculosis Report. World Health Organization, Geneva, Switzerland. Global Tuberculosis Programme (who.int)

World Health Organization. (2021). Antimicrobial Resistance. World Health Organization, Geneva, Switzerland. Antimicrobial resistance (who.int)

Yakrus, M.A., Hernandez, S.M., Floyd, M.M., Sikes, D., Butler, W.R. and Metchock, B. (2001). Comparison of methods for identification of *Mycobacterium abscessus* and *M. chelonae* isolates. J Clin Microbiol. *39*: 4103-4110.

Yang, Z., Kong, Y., Wilson, F., Foxman, B., Fowler, A.H., Marrs, C.F., Cave, M.D. and Bates, J.H. (2004). Identification of Risk Factors for Extrapulmonary Tuberculosis. *Clinical Infectious Diseases*. *38* (2):199–205.

Yeh, E., Garneau, S. and Walsh, C.T. (2005). Robust in vitro activity of RebF and RebH, a two-component reductase/halogenase, generating 7-chlorotryptophan during rebeccamycin biosynthesis. Proceedings of the National Academy of Sciences of the United States of America. *102* (11): 3960–3965.

Yoon, S.H., Ha, S.M., Kwon, S., Lim, J., Kim, Y., Seo, H. and Chun, J. (2017). Introducing EzBioCloud: A Taxonomically United Database of 16S rRNA Gene Sequences and Whole-Genome Assemblies. IJSEM. *67*: 1613-1617.

Yu, W-B., Pan, Q. and Ye, B-C. (2019). Glucose-Induced Cyclic Lipopeptides Resistance in Bacteria via ATP Maintenance through Enhanced Glycolysis. iScience. *21*: 135-144.

Zanin, M., Baviskar, P., Webster, R. and Webby, R. (2016). The Interaction between Respiratory Pathogens and Mucus. Cell Host & Microbe. *19* (2): 159–168.

Zelazny, A. M., Root, J. M., Shea, Y. R., Colombo, R. E., Shamputa, I. C., Stock, F., Conlan, S., McNulty, S., Brown-Elliott, B. A., Wallace, R. J., Olivier, K. N., Holland, S. M. and Sampaio, E. P. (2009). Cohort study of molecular identification and typing of *Mycobacterium abscessus*, *Mycobacterium massiliense*, and *Mycobacterium bolletii*. Journal of Clinical Microbiology. *47* (7): 1985–1995.

Zeng, J., Decker, R. and Zhan, J. (2012). Biochemical Characterization of a Type III Polyketide Biosynthetic Gene Cluster from *Streptomyces toxytricini* . Appl Biochem Biotechnol. *166*, 1020–1033.

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

Zhanel G.G., Fontaine, S., Adam, H., Schurek, K., Mayer, M., Noreddin, A.M., Gin, A.S., Rubinstein, E. and Hoban, D.J. (2006). A review of new fluoroquinolones: Focus on their use in respiratory tract infections. Treat Respir Med. *5* (6): 437–65.

Zhanel, G.G., Lawrence, C.K., Adam, H., Schweizer, F., Zelenitsky, S., Zhanel, M., Lagacé-Wiens, P.R.S., Walkty, A., Denisuik, A., Golden, A., Gin, A.S., Hoban, D.J., Lynch, J.P. 3rd and Karlowsky, J.A. (2018). Imipenem-relebactam and meropenem-vaborbactam: two novel carbapenem- β-lactamase inhibitor combinations. Drugs. *78*: 65–98.

Zhang, D., Iyer, L.M. and Aravind, L. (2011). A novel immunity system for bacterial nucleic acid degrading toxins and its recruitment in various eukaryotic and DNA viral systems. Nucleic Acids Res. *39* (11):4532-4552.

Zhang, Y.J., Reddy, M.C., Ioerger, T.R., Rothchild, A.C., Dartois, V., Schuster, B.M., Trauner, A., Wallis, D., Galaviz, S., Huttenhower, C., Sacchettini, J.C., Behar, S.M. and Rubin, E.J. (2013). Tryptophan biosynthesis protects mycobacteria from CD4 T-cell mediated killing. Cell. *155* (6):1296-1308.

Zhi, X-Y., Li, W-J. and Stackebrandt, E. (2009). An update of the structure and 16S rRNA gene sequence-based definition of higher ranks of the class *Actinobacteria*, with the proposal of two new suborders and four new families and emended descriptions of the existing higher taxa. Int J Syst Evol Microbiol. *59*: 589-608.

Zhu, Y.C., Mitchell, K.K., Nazarian, E.J., Escuyer, V.E. and Musser, K.A. (2015). Rapid Prediction of Inducible Clarithromycin Resistance in Mycobacterium Abscessus. Mol. Cell Probes. *29* (6), 514–516.

Zignol, M., Hosseini, M.S., Wright, A., Weezenbeek, C.L., Nunn, P., Watt, C.J. Williams, B.G. and Dye, C. (2006). Global incidence of multidrug-resistant tuberculosis. J Infect Dis. *194* (4): 479-485.

Supplementary data

Table 20. Supplementary Files

Supplementary File ID	Supplementary File Title	Attached file
S1. List of genomes used	List of genomes used.txt	 S1. List of genomes used.txt
S2 Pyani generated ANIm analysis	Pyani generated ANIm Percentage Identity File.pdf	 S2. Pyani generated ANIm Percentage Ider
S3. List of genes present in *M. abscessus* and absent in *M. chelonae*	Ppanggolin_diff_Mab_Mchel.xlsx	 S3. Ppanggolin_diff_Mab_