UNDERSTANDING THE LANDSCAPE OF ANTIMICROBIAL RESISTANCE AND VIRULENCE DETERMINANTS IN *Mycobacterium abscessus* THROUGH PAN-GENOME ANALYSIS

DISSERTATION SUBMITTED TO
THAPAR INSTITUTE OF ENGINEERING AND TECHNOLOGY, PATIALA

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

FOR PARTIAL FULFILLMENT OF DEGREE OF
MASTER OF SCIENCE IN BIOTECHNOLOGY

SUBMITTED BY:
SIMRAN GAMBHIR
Reg.no- 302010126

CONDUCTED AT

CSIR-IMTECH

COMPLETED UNDER THE SUPERVISION OF

Dr. ANSHU BHARDWAJ
PRINCIPAL SCIENTIST
BIOINFORMATICS CENTRE, G.N RAMACHANDRAN PROTEIN CENTRE
CSIR-INSTITUTE OF MICROBIAL TECHNOLOGY, SECTOR 39-A, CHANDIGARH

DECLARATION BY STUDENT

I hereby, undertake that the Thesis untitled "UNDERSTANDING THE LANDSCAPE OF ANTIMICROBIAL RESISTANCE AND VIRULENCE DETERMINANTS IN *Mycobacterium abscessus* THROUGH PAN-GENOME ANALYSIS" is prepared by me and that the document reports original work carried out by me under the guidance of Dr. Anshu Bhardwaj. I have made sure that all ideas, expressions that are not a part of my work are properly credited. The work reported in this has not been submitted by me for award of any other degree or diploma.

Simran Gambhir

Date: 15/07/2023

सीएसआईआर – इमटैक
CSIR-IMTECH

सीएसआईआर – सूक्ष्मजीव प्रौद्योगिकी संस्थान
सैक्टर 39-ए, चण्डीगढ़-160 036 (भारत)
CSIR-INSTITUTE OF MICROBIAL TECHNOLOGY
(A CONSTITUENT ESTABLISHMENT OF CSIR)
Sector 39-A, Chandigarh-160 036 (INDIA)

Certificate

This is to certify that the project entitled "Understanding the Landscape of Antimicrobial Resistance and Virulence Determinants in *Mycobacterium abscessus* Through Pan-genome Analysis" has been carried out by Ms. Simran Gambhir as a part of the Dissertation program-2023.

The work is done under my supervision at the CSIR-Institute of Microbial Technology, Sector 39-A, Chandigarh for a period of six months (11th January 2023 to 10th July 2023).

Date: 28th July 2023

Place: Chandigarh

Anshu Bhardwaj

Dr. Anshu Bhardwaj

(Principal Scientist)

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

CERTIFICATE

This is to certify that **Simran Gambhir** (302101026), M.Sc student in the Department of Biotechnology, has completed bonafide work on the thesis entitled "Understanding the Landscape of Antimicrobial Resistance and Virulence Determinants in *Mycobacterium abscessus* through Pan-genome analysis" at CSIR- Institute of Microbial Technology, Chandigarh, India.

Dr. Anshu Bhardwaj
CSIR-IMTECH
July 2023

Dr. Priyankar Dey
TIET
July 2023

ACKNOWLEDGEMENT

I would like to take this opportunity to express my deepest and most sincere gratitude to Dr. Anshu Bhardwaj, Principal Scientist, Bioinformatics Center, G.N. Ramachandran Protein Centre, CSIR-Institute of Microbial Technology, Chandigarh. Her insights and deep understanding of the subject, were essential for carrying out the research. I am especially grateful for the time she took to meet with me regularly to discuss my progress, and for her willingness to help me troubleshoot problems.

I would also like to thank CSIR-IMTECH for the opportunity to have carried out my research at this institute, and for the excellent facilities that were made available to me.

I would also like to express my gratitude to Dr.Rania Assab, for the prior work on NTM. Her work was instrumental in guiding me through the project.

I would like to acknowledge my lab mates for their constant support and encouragement, they were always there to listen to me and helped me to stay motivated.They have made my time in the lab so much more enjoyable.

Additionally, I pay my regards to Dr. MS Reddy, H.O.D, Department of Biotechnology, Thapar Institute of Engineering and Technology and my internal guide Dr.Priyankar Dey, Assistant Professor, Thapar Institute of Engineering and Technology for allowing me to carry out my project here at IMTECH.

Lastly, I would like to express my deepest gratitude to my parents for their love and support throughout my life. They have always been there no matter what, I am forever indebted to my parents for their support.

Thanking You
Simran Gambhir

ABBREVIATIONS

NTM	Non-tuberculous mycobacteria
MTB	*Mycobacterium tuberculosis*
RGM	Rapidly growing mycobacteria
SGM	Slowly growing mycobacteria
MAC	*Mycobacterium avium* complex
MABC	*Mycobacterium abscessus* complex
MAB	*Mycobacterium abscessus*
CF	Cystic fibrosis
MAA	*Mycobacterium abscessus subsp. abscessus*
MAM	*Mycobacterium abscessus subsp. massiliense*
BOL	*Mycobacterium abscessus subsp. bolletii*
ATS	American Thoracic Society
GPL	Glypeptidolipids
TDM	Trehalose dimycolate
RND	Resistance, nodulation and cell division permeases
MmpL	Mycobacterial membrane protein Large
EUCAST	European committee for Antimicrobial susceptibility testing
VF	Virulence Factors
CSV	Comma separated value
GFF	Gene feature format
CD-HIT	Cluster Database at High Identity with Tolerance
MCL	Markov clustering algorithm

SYNOPSIS

Mycobacterium abscessus is an opportunistic pathogen that belongs to the Nontuberculous mycobacteria group. MAB was not identified as a pathogen until 40 years after it was discovered. However, its prevalence has risen over time. MAB primarily affects patients with underlying conditions such as cystic fibrosis, chronic pulmonary obstructive disorder, and bronchiectasis, frequently resulting in a decline in lung function. MAB has been linked to outbreaks of soft tissue infections and infections after cosmetic surgery, in addition to causing pulmonary infections. MAB is resistant to a variety of medications, including macrolides, tetracyclines, beta-lactams, and aminoglycosides. MAB is naturally resistant to anti-TB medications like rifabutin, clofazimine, and bedaquiline.*Mycobacterium abscessus subsp. abscessus* (MAA), *Mycobacterium abscessus subsp. massiliense* (MAM), and *Mycobacterium abscessus subsp. bolletii* (BOL) are the three subspecies of MAB. The resistance phenotypes of these three subspecies to various drugs differ significantly. As a result of different drug susceptibility profiles, targeted methods for delineating subspecies are important. MAB is found in soil, water, and household plumbing systems and can survive inside amoebas. It was long thought that MAB was acquired from the environment. However, efforts to sequence the entire genome have revealed evidence of possible human-to-human transmission via genetically similar clusters known as dominant circulating clones. These DCC have been repeatedly attributed for MAB's nationwide and rare intercontinental spread.

MAB infection incidences are typically underreported, particularly in TB-epidemic countries like India, due to overlapping symptoms such as nodular lesions that mimic TB lung condition and association with other comorbidities such as COPD and CF. Because MAB can acquire genes from other microorganisms, the possibility of developing antimicrobial resistance genes and virulence factors exists, so it is critical to characterize these determinants and map them to their geographical locations.

In this thesis, we use a pan-genome approach to understand the landscape of global core and accessory AMR and VF determinants. We also intend to improve MAB annotation by focusing on AMR and VF determinants. We also aim to understand resistance phenotypes by delving into their basic mechanisms using sequence analysis.

LIST OF TABLES

LIST OF FIGURES

TABLE OF CONTENTS

CHAPTER 1 : INTRODUCTION & REVIEW OF LITERATURE

1.1 NONTUBERCULOUS MYCOBACTERIA: OVERVIEW

The infamous obligate parasites, *Mycobacterium tuberculosis* and *Mycobacterium leprae* have been known to mankind for generations. Mycobacterium was originally the name given to organisms that grew in a mold-like manner. Mycobacteria are taxonomically classified as pertaining to the genus Mycobacterium, which is the sole genus under the Mycobacteriaceae family, in the order Actinomycetales [1]. Mycobacteria are aerobic, slightly curved, rod-shaped bacteria that are characterized by their capacity to synthesize mycolic acids and their resistance to acid-alcohol followed by staining with Ziehl Neelsen stains. Mycobacteria are generally classified into two primary groups, MTB complex which is responsible for causing tuberculosis and Nontuberculous mycobacteria which do not cause TB [2]. NTM were initially classified on the basis of their growth rates into slowly growing mycobacteria (SGM) and rapidly growing mycobacteria (RGM). SGM takes more than 7 days to grow on solid culture media and is further divided into 3 types on the basis of pigment production. Type I is a photochromogen, where colony pigmentation is only visible in the presence of light; Type II is a scotochromogen, where colony pigmentation ranges from yellow to orange despite light; Type III and Type IV are non-pigmented or have just a pale-yellow color regardless of light or darkness. *Mycobacterium avium* complex (MAC), *Mycobacterium kansasii*, *Mycobacterium haemophilum* are some examples of SGM. On the other hand, *Mycobacterium abscessus* complex (MABC) and *Mycobacterium fortuitum* complex fall under the category of RGM (Type IV), and grow in less than 7 days [3]. Many NTM species have been isolated and characterized as a result of advances in microbiological isolation techniques and easily accessible sequencing technology[4]. Of the 170 species recognized, MAC, MABC and *M.kansasii* are the most frequent human pathogens [3]. NTM are widespread in the environment and can be found in soil, water, protozoa, poultry, and humans [6]. Additionally, it has been proposed that NTM's capacity for survival in amoebas and protists allows it to multiply in animal macrophages. NTM are resistant to disinfectants and biocides because they have lipid-rich outer cell walls, which helps them outcompete other organisms that are disinfectant-sensitive. NTM can survive in natural and artificial habitats, such

as pipes, plumbing, and distributions in homes, hospitals, etc., because of the lipid-rich cell wall responsible for surface hydrophobicity [7,8]. In addition to offering defense against disinfectants, the lipid-rich cell wall also operates as a barrier for antibiotics.

Both immunocompetent and immunocompromised patients are adversely affected by NTM. The likelihood of acquiring NTM disease increases in patients with inherited or acquired immunodeficiencies including HIV, organ transplants, or inflammatory conditions [9] .Genetic predisposition to lung diseases such as bronchiectasis and cystic fibrosis (CF) also increase the possibility of patients contacting the NTM pulmonary disease (NTM-PD) [3]. Hermansen and colleagues conducted a nationwide study on half a million mycobacterial cultures to estimate the yearly prevalence of culture-verified NTM disease from 1991 to 2015. They found a bimodal age association of NTM disease incidence in young children (0-4 years) and older people (65–69 years), which suggests that an insufficient immune response may play a role in susceptibility. [11]. Also immunocompetent people have reported an upward trend in the incidence of respiratory and healthcare-associated ailments. In contrast to MTB, reporting of NTM infection is not required, which hinders the understanding of public health affected by NTM. However, western industrialized regions have experienced NTM infection while, with tuberculosis following the downward trend [10]. Furthermore, individuals with no genetic or immune abnormalities, such as postmenopausal women with particular body types, can develop NTM lung disease [9]. Although the pulmonary system is involved in 90% of NTM infections, skin and soft tissue, lymph nodes, and bones are also affected. Central nervous system and disseminated infections, on the other hand, are less frequently observed [3].

The typical treatment regimens for NTM infections involve the use of macrolides such as clarithromycin and azithromycin along with various antibiotics, such as ethambutol and rifampin in the case of SGM and tetracyclines and carbapenems in the case of RGM [3]. Since different NTM species have different drug susceptibility profiles, species identification is crucial before administering a course of treatment.

Treatment of NTM infections is complicated despite advances in molecular identification and microbiological techniques. Limited sensitivity and specificity of symptoms, radiology, which frequently causes tuberculosis diagnosis to be made incorrectly, inherent and acquired resistance to current drugs as well as to first and second-line TB medications, as well as differences in drug susceptibility, temperature tolerance, and growth rates, further complicates the matter [12].

1.2 *Mycobacterium abscessus complex*

Mycobacterium abscessus (MAB) is gram positive, rapidly growing, multi drug resistant NTM species. Patients with chronic respiratory diseases, such as cystic fibrosis,bronchiectasis, and chronic obstructive pulmonary disease (COPD), are more likely to develop MAB infection. These diseases damage the lungs, making them more susceptible to infection by MAB Several incidences have also been reported for soft tissue infections in the non-CF population, with outbreaks due to the use of contaminated surgical instruments [13]. MAB was first isolated in 1953 from knee abscess of a 63 year old woman. When MAB was first isolated, it was thought to be a low-virulent organism, but a 1993 study found that 154 patients with underlying conditions like cystic fibrosis and gastroesophageal conditions had infections that were caused by MAB (82% of the total isolates) [14]. Since it isolation, MAB nomenclature and taxonomy have undergone multiple revisions, beginning with it being designated as a subspecies to *M.chelonae* due to its biochemical identity; in 1992, it was elevated to species status after only 35% relatedness was detected between the two, following a DNA hybridisation experiment.In the following years, scientists discovered *M.massiliense* using partial PCR sequencing of rpoB. Only the *Mycobacterium abscessus* type strain demonstrated 96.0% partial rpoB sequence similarity and 98.0% recA gene sequence similarity. Following that, using rpoB sequencing, scientists discovered a novel species in 2006 that shared 100% 16S gene and 95.6% rpoB gene similarities with the MAB type strain; the species was named *M. bolletii* [15]. Bryant et al. established in 2013 that the *Mycobacterium abscessus complex* has three different subspecies, namely *M. abscessus subsp. abscessus* (MAA), *M. abscessus subsp. bolletii* (BOL), and *M. abscessus subsp. massiliense* (MAM) [16].

MAB has been increasingly isolated from cystic fibrosis patients in North America, Europe, and Asia. MAC and MAB are the most prevalent NTM isolated from CF lung patients.In comparison to all other NTM species, *Mycobacterium abscessus* contributes to nearly 56% of all isolated samples from the lungs of CF patients [17].

1.3 THE SUBSPECIES DILEMMA

As mentioned before, MABC comprises 3 subspecies, MAA, BOL and MAM, however the molecular identification of MABC is a challenging and complicated process to date. Precise

molecular identification of NTM organisms is expected to facilitate better patient management and treatment outcomes [18]. During the 1950s, when MAB was first isolated, it was thought to be identical to M.chelonae, another RGM which inhabits amphibians [19]. After 40 years, MAB has been recognised as a different species thanks to DNA-DNA hybridisation. However, there is still some ambiguity because several laboratory tests still report MAB as part of the *Mycobacterium abscessus*/*M.chelonae* complex. Making this distinction and accurate species delineation is critical because *M.chelonae* rarely causes lung infection and lacks the erythromycin methylase (erm) gene, which is responsible for resistance to macrolides such as clarithromycin and amikacin in MAB. Although 16S rRNA sequencing is the FDA-approved technique for NTM evaluation in the United States, it cannot distinguish between the two species. Target genes, such as the rpoB gene, must be sequenced to determine the two species[18]. The three subspecies of MABC further complicates the matter of taxonomic classification and antibiotic prescription.In 2011, some scientists proposed combining two mycobacterial subspecies, MAM and BOL, into a single subspecies known as *M. abscessus subsp. bolletii*. However, in 2013, studies using whole-genome sequencing to compare the genetic makeup of these subspecies discovered that they were distinct from one another and should be classified as three separate subspecies belonging to the MABC *(Mycobacterium abscessus* complex) group. The three subspecies show 100% sequence identity according to complete 16S rRNA gene sequence analysis. However, based on partial sequencing of the rpoB gene, these three subspecies vary by more than 3.5% of their base pairs, a distinction seen only between isolates from different species.

Moreover, MAM as compared to the two other subspecies has a large deletion erm(41) gene which makes the bacterial subspecies susceptible to macrolides.On the other hand, a 14-day exposure to the functional gene in MAA and BOL confers resistance to macrolides. However, laboratories have been slow to adopt a simple procedure that can aid in the improvement of MAB, MIC testing accuracy. There is no single, agreed-upon method for identifying new mycobacterial species, and the techniques are constantly evolving.The field of NTM taxonomy is. in flux , and many laboratories are not using the most recent methods for identifying these bacteria. This can result in a wrong diagnosis and unsuccessful therapy [18].

1.4 TRANSMISSION AND PATHOGENESIS

MAB, like other NTMs, can be found in a wide range of environments, including soil, water sources, hospital water supplies, and household plumbing such as sinks and showerheads. MAB can survive in harsh environments due to its impenetrable, lipid-rich cell wall and ability to form a biofilm. A long held belief was that MAB is acquired from the environment; however, this belief was challenged when a study on 1080 samples from 517 patients revealed that human-human transmission is possible from fomites and aerosoles. They also discovered that three dominant circulating clones (DCC) MAA DCC1 & DCC2 and MAM DCC 3, accounted for 74% of the isolates. These genetically similar clusters have repeatedly been observed as the cause of outbreaks in single hospital settings and nationwide .Another study in Australian cystic fibrosis patients found MAB in 6.7% (22/328) of CF patients' respiratory specimens, and whole genome sequencing revealed genomically related clusters in 3 patients, indicating probable transmission. These studies shifted the strategies used to control MAB infection in CF centers from standard sterilization to patient segregation with air changed independently 15 times an hour to control transmission[20]. A recent study conducted in a tertiary hospital in Japan also suggests that genetically closely related isolated non-CF patients (less than 21 SNP distance between a cluster found in the UK and Japan) can be transmitted without any contact between the affected populations [24]. Studies also show that these genetically related dominant circulating clones emerged when CF patients' life expectancy increased, resulting in more susceptible individuals for MAB to grow. Even though MAB primarily affects the lungs of patients with CF, resulting in a decline, lung function studies show that non-CF patients are a secondary vector responsible for MAB global transmission.

One of the important points to discuss here is the unique mode of MAB for acquiring genes from other organisms, distributive conjugal transfer(DCT). Horizontal gene transfer is one of the major forces driving evolution other than mutation and transposition. Conjugation in *M. smegmatis* can result in large amounts of DNA being transferred between strains, resulting in a wide range of genomic variation. This variation is random and can occur anywhere in the genome. DCT involves the transfer of multiple, large segments of DNA to the recipients. A study published in 2016, clearly demonstrates that MAB undergoes genomic rearrangements, just as *M.smegmatis*.

NTMs are dangerous to people with life-threatening conditions such as cystic fibrosis, AIDS, bronchiectasis, and gastroesophageal reflux disease (GERD). In the last decade, the prevalence of NTM, a common pathogen in the CF population, has increased from 3% to 22.6%. Cystic fibrosis is an autosomal recessive disorder in which most deaths are caused by CF lung infection caused by impaired clearance of thick mucus secretion, which promotes bacterial colonization and progressive decline in lung function. MAB has been reported in more significant numbers in CF centers worldwide than other NTM subspecies[22, 23]. Along with being reported in causing infections in patients with compromised immune systems, central nervous system (CNS) infections, manifesting as cerebral abscesses, have also been observed.Bacteria can enter the body through wounds in the skin, like cuts, scrapes, or ulcers. They can also get in through contaminated soil or water. Most infections in healthy people occur due to a penetrating injury. This type of injury breaks the skin, allowing bacteria into the body. Soft tissue infections can also occur due to naturally occurring puncture wounds, contaminated fluid injections, or surgical procedures[25].

MAB has been linked to worsening lung function in people with CF. Medical professionals are concerned about the risk of MABC infections in hospitals, especially in patients with lung transplants or cystic fibrosis. This is because MABC infections can be difficult to treat and can lead to serious complications. Additionally, there have been cases of nosocomial MABC infections, which means that the infection was acquired in the hospital. Along with genotypic heterogeneity among subspecies, MABC also showcases phenotypic heterogeneity which is an essential factor that confers virulence to MAB[26]. NTM including MAB shows two types of phenotypic morphology which is differentiated on the basis of presence and absence of Glycopeptidolipids (GPL) on their cell wall. MAB shows smooth and rough colony morphologies, where smooth strains produce a high amount of GPL, whereas the rough strains have trehalose dimycolate and a low amount of GPL.The wild-type smooth strains with abundant GPL, tend to mask surface molecules, thus helping in evading the innate immune response. TDM in rough colonies is associated with cording phenotype, which helps MAB in inducing an immune response[25].

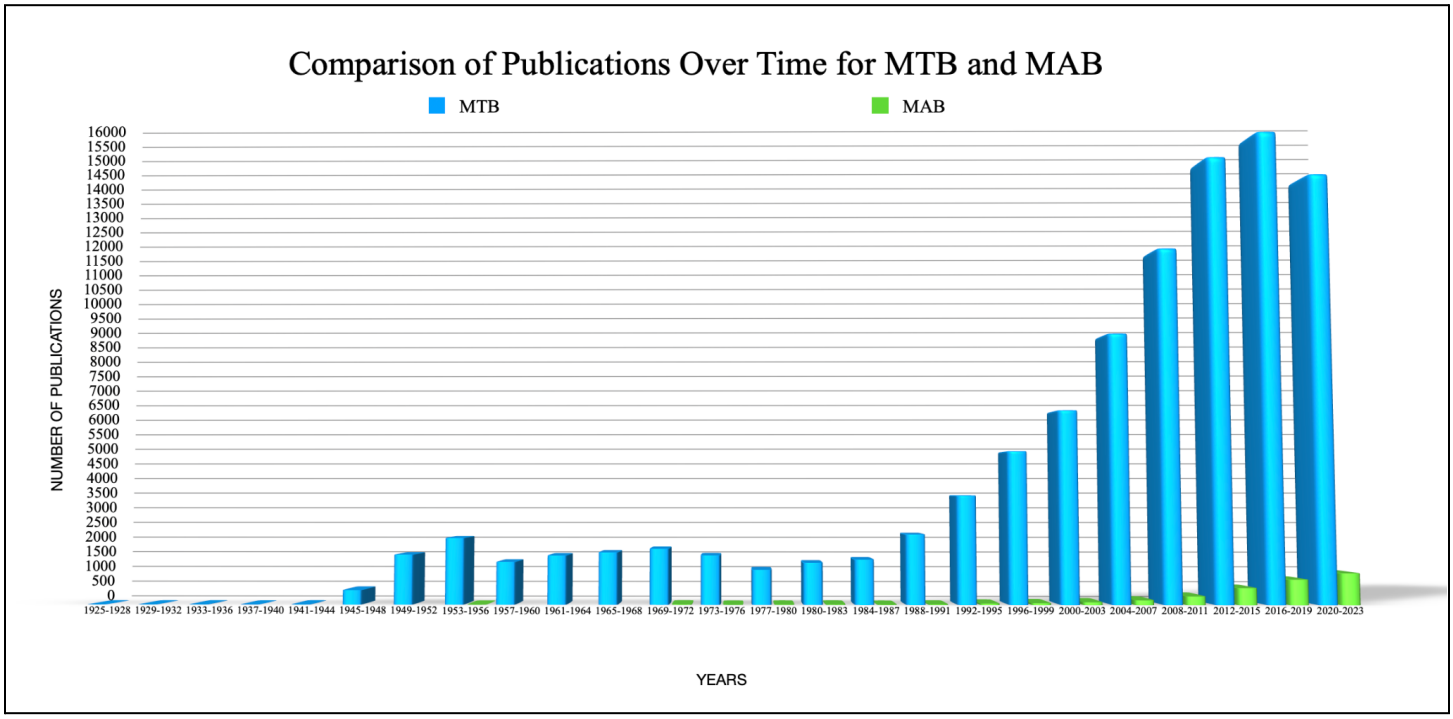

Fig. 1: Comparison of number publications from PUBMED over time for MTB and MAB

Fig 1 represents the number of publications for MTB and MAB over a number of years.

Even though the first article published dates for both organisms differ (1925 for MTB and 1953 for MAB), the graph shows a significant difference in the number of publications over time.This may be because of several reasons such as dilemma in species identification for MAB, misdiagnosis of MAB with TB which leads to underreporting of MAB infections. Apart from this, documenting the infection cases due to MAB is not mandatory as with TB.

1.5 THE PAN-GENOME APPROACH

Tettelin et al. published a paper in Proceedings of the National Academy of Sciences made a major contribution to our understanding of the pan-genome concept. The authors of this paper sequenced the genomes of six different strains of *Streptococcus agalactiae*, a bacterium that can cause serious infections in newborns. The word “pan” in pan-genome is derived from a Greek word meaning “whole”, pan-genome is the full repertoire of genes of a bacterial species and can be classified into “core” ,which includes the genes present in all the strains, and “accessory” which represents dispensable, partially shared and genes unique to some or a specific strains.

Tettelin et al. clearly demonstrates the importance of the pan-genome concept in describing genetic complexity and variability among the bacterial trains. They argued that studying a bacterial species based on its reference strain can be misleading because it does not account for its genetic diversity.The pan-genome concept can help us understand how bacterial species evolve in response to different environments and stressors. By studying the pan-genome, we can identify genes essential for adaptation to different environments and track how these genes have changed over time.

Choo et al. published the first pan-genome of MAB in 2014 after sequencing 12 genomes from a hospital setting in Malaysia. Their findings clearly show that MAB has an open pan-genome, which means that the size of the pan-genome grows as the number of genomes sequenced increases. The open pan-genome suggests that MAB can acquire new genes and continuously evolve.

As discussed previously, MABCcomprises three subspecies, however the majority of infections have been caused by MAA [3,5,6]. So, currently in our study we are focusing on MAA for pan-genome analysis.

The objective of our study is to understand and improve our knowledge on drug resistance and pathogenesis in MAB. In this study, we aim to landscape the genes responsible for antimicrobial resistance and pathogenesis in MAB, and understand the genetic basis of acquisition of traits responsible for pathogenesis. We also intend to annotate the bacterial genomes by specifically focusing on AMR and VF determinants.

1.6 GENOMIC TOOLS AND RESOURCES FOR PAN-GENOME ANALYSIS

1.6.1 BACTERIAL AND VIRAL BIOINFORMATICS RESOURCE CENTRE (BV-BRC)

BV-BRC which stands for Bacterial and Viral Bioinformatics Resource Center was used to retrieve complete and whole genome sequences of MAB.

BV-BRC was formed from the integration of three BRC (Bioinformatics Resource Center) sources listed below:

- PAThosystems Resource Integration Center (PATRIC)
- Influenza Research Database (IRD)
- The Virus Pathogen Database and Analysis Resource (ViPR)

BV-BRC offers bacterial (both pathogenic and non-pathogenic) and viral genomic data along with other essential biological metadata. BV-BRC also offers various inbuilt tools for visualizations, analysis and command line interface with numerous switches to retrieve the required data. Several metadata attributes such as geographic location, BioSample Id, BioProject Id, information about host are also provided by BV-BRC. In addition to this BV-BRC also provides test results for Antimicrobial susceptibility testing for bacterial species, curated from NCBI and publications[34].

1.6.2 QUAST: GENOME QUALITY ASSESSMENT

QUAST was used for genome quality analysis, on the basis of certain metrics that will be discussed in the material and methods section. Quast is a tool available for genome quality assessment, it evaluates a number of parameters and also provides informative visualizations. In comparison to other tools such as Plantagora and Gage can evaluate assembly quality without reference genome. Quast collectively uses already available tools such as Plantagora, GlimmerHMM, GeneMark.hmm, along with extending the quality control with its own metrics. Although the quality control metrics can be computed without the reference genome, Quast utilizes Nucmer aligner for aligning assemblies to their reference genomes. Along with calculating standard metrics such as number of contigs, length, length of largest contigs, GC%,

Quast also computes NGA50, the number of miss assemblies as well as number of unaligned contigs. In all, Quast provides a comprehensive quality evaluation of genomic assemblies [35].

1.6.3 PROKKA: RAPID PROKARYOTIC GENOME ANNOTATION

For structural and functional annotation of genomes, Prokka was used. It produces high quality genome annotation and accomplishes this by combining other software tools and employing many processing cores. This enables Prokka to annotate genomes in a relatively less time that other methods would require. Prokka combines a suite of different softwares for prediction of coordinates of different genomic features:

- Prodigal for prediction of coding sequences
- RNAmmer for identifying rRNA gene,
- Aragorn for predicting tRNA genes,
- SignalP for identification signal leader peptides
- Infernal for non-coding RNA.

After identification of coordinates of coding sequences, Prokka performs functional annotation by following a hierarchical approach to predict the putative gene products. BLAST+ is used with default e-value of 1e-6 to look for matches in Uniprot and Refseq for the particular genus and species as per the user's requirement . If no match is found in the aforementioned databases, the results are returned as hypothetical protein [37].Prokka accepts fasta files for annotation and gives 12 different files for further analysis and integration into different software tools.

1.6.4 ROARY: PROKARYOTIC PAN-GENOME CONSTRUCTION

Roary, a command line software tool for pan-genome construction specifically for prokaryotes was used. The coding regions of the input sequences are extracted and converted to protein sequences to identify orthologous groups of genes. Incomplete arrangements are eliminated. The protein sequences are clustered using a combination of CD-HIT and BLASTP.

CD-HIT is a program that iteratively pre-clusters the sequences multiple times, each time improving the accuracy of the results by using the results of the previous clustering. BLASTP is a programme that compares protein sequences and finds similar ones.The resulting clusters are

merged and refined using conserved gene neighborhood information.Roary takes Prokka GFF files and generates multiple output files for further analysis and visualization[38].
Aside from that, roary provides a variety of switches such as minimum percentage identity and MCL inflation value to fine-tune results based on the user's analysis.

1.6.5 RESISTANCE GENE IDENTIFIER (RGI)

RGI is a software tool for prediction of resistance genes available through galaxy wrapper, and command line interface which can be obtained through Conda or Docker. Apart from this RGI can be accessed through RGI portal via CARD website. RGI utilizes CARD (COMPREHENSIVE ANTIBIOTIC RESISTANCE DATABASE) as a database to look for antibiotic resistance genes on the basis of homology and SNP model. Protein homolog model is used for detection of AMR genes based on the sequence homology with already curated sequences present in the database based on certain blastP threshold. Protein variant model works same as protein homolog model, but also looks for through a database of curated set of mutations. RNA variant models are used to look for acquired mutations but in ribosomal RNA genes. RGI classifies the obtained blast hits as Perfect, which includes exact matches to the curated set of sequences in th CARD database, Strict matches which includes matches of unknown variants of predicated AMR genes. Besides perfect and strict matches, loose matches usually look for matches below the decided cut-offs, and results in identification of unfamiliar AMR genes [39].

1.6.6 AMRFINDERPLUS

AMRFINDERPLUS is an NCBI software tool for identifying AMR determinants using nucleotide and protein fasta sequences. Most AMR prediction tools use Hidden Markov models or require databases to find homology between the input sequences and the database. For classifying AMR genes, HMM employs a hierarchical approach. AMRFINDERPLUS makes use of NCBI's own curated Reference database as well as curated HMM models found in the Reference HMM catalog. The curators regulate and review the database, collaborating with CARD and domain experts[40].

AMRFINDERPLUS not only aids in the search for AMR genes and point mutations, but it also includes genes that are activated in response to stress, such as biocide and heat resistance. Porins have been added to the list of virulence factors[41].
AMRFINDERPLUS can predict AMR determinants using both protein and nucleotide fasta files. However, a GFF file can be submitted alongside the sequence fasta files to obtain the coordinates of the predicated fasta sequence.

1.6.7 VFDB - VIRULENCE FACTOR DATABASE

VFDB serves as a reference, comprehensive platform for bacterial virulence factors that are responsible for pathogenesis. VFDB comprises an extensive collection of virulence factor from 32 genera of bacteria pathogens as per the latest update. The hierarchical architecture of categories and subcategories for classification of virulence factors provides an easy interface for the user to have a compressive information of a particular bacterial pathogen. Apart from that, the section for comparative pathogenesis allows the user to make a comparative assessment of the virulence factors present in the same genera among different species.

1.7 SUBSYSTEM ANALYSIS

A subsystem is a collection of functional roles that are interconnected. Annotators determine which functional roles should be combined to form a subsystem. Subsystems can represent a wide range of biological entities, including metabolic pathways, complexes, and protein classes.
The SEED project was started in the year 2003, by Fellowship of Interpretation of Genomes(FIG), in order to develop a technology for accurate and reliable annotation with the goal to annotate 1000 genomes.
The goal of the Project to Annotate 1000 Genomes is to create a collection of carefully defined and curated subsystems. This is accomplished by having a group of experts annotate the same subsystem across a variety of genomes. This enables annotators to gain a thorough understanding of the variation found in each subsystem, as well as the relevant literature[43].

1.8 SYNTENY ANALYSIS

The study of gene order conservation across genomes is known as synteny analysis. It is a very useful technique for understanding genome evolution and discovering functionally linked genes. Synteny analysis can be used to detect orthology, find conserved syntenic blocks, analyze genome evolution, and predict gene function, among other things. Conserved syntenic blocks are genomic areas that have been preserved in gene order across species. These blocks frequently contain genes involved in the same biological process. Synteny analysis can be used to investigate how genomes evolve through time. It is feasible to identify portions of the genome that have been rearranged or duplicated by comparing the synteny of various genomes. We used a software tool Sibelia to perform synteny for the longest complete genomes between different countries. Sibelia also allows visualization of syntenic blocks in a circular plot through Circos.

1.9 GENE ONTOLOGY MAPPING

With the advent of genomics, biology has undergone a revolution in recent years. This has resulted in massive amounts of data being gathered, posing new challenges and opportunities for biologists.One of the difficulties is that this data is frequently stored in different formats and standards, making integration and analysis complex. Another challenge is that data can be very complex and heterogeneous, making understanding the relationships between other pieces of data complex. This necessitates the creation of new data integration methods and tools.

Ontologies are one approach to data integration. Ontologies are formal vocabularies that describe concepts and relationships within a specific domain. They can represent various data types in a biological database and map data between databases. The Gene Ontology project (GO) began in 1998 as a collaboration between various databases based on model organisms and has since expanded and integrated the world's significant data collection repositories, facilitating communication among scientists[33] GO brings together three non-overlapping categories of molecular biology, namely biological processes (BP) which describe the larger-scale activities that take place in a cell or organism. For example, the BP "cell cycle" describes the process by which a cell divides to produce two new cells. Molecular function (MF) describes the specific

activities that are carried out by individual gene products and cellular components (CC) describe the structures within a cell where gene products are located or act.

Directed acyclic graphs (DAGs) are commonly used to represent ontologies. DAGs are a graph with no cycles, which means you can't start at any node and follow the edges back to the same node. As a result, they are a valuable tool for organizing and managing biological knowledge.

Here, GO terms for reported AMR genes were mapped with InterPro. InterPro is a database that provides information about the function of proteins. It does this by combining data from various protein signature databases, which are collections of protein sequences that have been classified into families, domains, and functional sites. InterPro can be used to characterize new protein sequences by identifying the features that they share with known proteins. It is a curated resource that combines data from various protein signature databases.

CHAPTER 2 : MATERIALS AND METHODS

2.1 DATA RETRIEVAL FROM BV-BRC

BV-BRC was used to retrieve genome sequences for MAA. The steps for the retrieval are listed below:

- Different keywords such as "*Mycobacteriodes abscessus" , "Abscessus" "Mycobacterium abscessus" and "Mycobacteriodes abscessus* OR *Mycobacterium abscessus"* were used to retrieve the total number of genomes from BV-BRC.
- CSV files were downloaded for the above mentioned query keywords.
- The downloaded files were processed using AWK for the following criteria

➢ Organism name: " *Mycobacteriodes abscessus subsp. abscessus*" OR "*Mycobacterium abscessus subsp.abscessus*"
➢ Genome status: WGS OR Complete
➢ Genome quality : Good
➢ Host species: *Homo sapiens*

Commands used to filter the files according to the criteria mentioned above:

- awk -F"," '$2 ~ /*Mycobacteriodes abscessus subsp. abscessus*/ || $2 ~/*Mycobacterium abscessus subsp. abscessus*/ {print $2}' bvbrc_2153.csv | wc -l
- awk -F"," ' $15 ~/Complete/ || $15~/WGS/ || $15=="" {print $2 "\t" $15}' status.csv | wc -l
- awk -F"," '($36~/Good/ || $36 == "") {print}' quality.csv| wc -l
- awk -F"," '$75 !~ "seakrait" {print}' host.csvl| wc -l

Empty entries for host species in the file were looked in NCBI to confirm for the host species as *Homo sapiens.* After sorting out the list according to the criteria, BV-BRC CLI (command line interface) was used to retrieve the genomes CLI tools also known as p3-scripts allow access to data, different services provided by BV-BRC.

An in-house PERL script was used to download genome sequences using BV-BRC CLI.

Perl script used as follows:

```
open(FH,"genome2153.txt") || die "file not found";
$count=0;
while($ln=<FH>)
{
        chop($ln);
        $ln_new=$ln.".fasta";
        print $ln_new; print "\n";$count++;
        `p3-genome-fasta $ln > $ln_new`;
}
print "\n",$count,"\n";
```

2.2 GENOME QUALITY ANALYSIS

Genome quality was assessed using the genomes that were finally retrieved from BV-BRC. Quast takes fasta nucleotide files as input sequences. Along with the fasta nucleotide files, we also provided a reference genome of MAB, ATCC 19977 for computing the metrics. The -r switch is used to specify the reference genome.

Command used to run Quast

```
ls *.fasta | parallel --verbose "quast.py -r 19977_6con.fasta  {} -o {.}_out_q1"
```

The transposed report file was used to select the genome that followed the EUCAST(European Committee on Antimicrobial Susceptibility Testing) quality control guidelines for whole genome sequencing [36].

Quality Control Metric	Explanation
Total number of contigs	Less than 100 contigs for organisms with 5-6 Mb genome size
Number of contigs> 500bp	Its corresponds to the total number of contigs have length more than 500 bp, the number should well correspond to total number of contigs
NG50	More than 30000 bp is preferred.

Table 1: Quality controls metrics for WGS by EUCAST

2.3 STRUCTURAL AND FUNCTIONAL ANNOTATION

Prokka accepts fasta files for annotation and gives 12 different files for further analysis and integration into different software tools for further analysis. Prokka provides options for mentioning the kingdom and genus of the bacterial pathogen under analysis. Thus we have mentioned Mycobacterium as the genus.

Command used to run prokka:

```
ls *.fasta | parallel --verbose "prokka  --kingdom Bacteria  --genus Mycobacterium {} --prefix {.}_outr"
```

2.4 PAN-GENOME CONSTRUCTION

Roary accepts GFF files produced by Prokka. Roary was run at default minimum percentage identity of 95% for all against all Blastp.

Command used to run Roary:

```
roary -f roary_output *.gff
```

2.5 IDENTIFICATION AND ANNOTATION OF ANTIMICROBIAL RESISTANCE GENES

2.5.1 RGI

Nucleotide fasta files were used as input for the prediction of AMR genes at default parameters. We did not opt for the loose matches. RGI produces results in txt format as well as outputs JSON files for producing heatmaps.

The following script was used as obtain AMR genes from RGI

```
#!/usr/bin/bash
for f in /home/bic/rgi/sample_input2/*.fasta
do
        base=$(basename "$f" .fasta)
        rgi main --input_sequence "/home/bic/rgi/sample_input2/$base.fasta" --output_file
"/home/bic/rgi/847_output/${base}" --local --clean
done
~
```

2.5.2 AMRFINDERPLUS

AMRFINDERPLUS can predict AMR determinants using both protein and nucleotide fasta files. GFF and protein fasta files from Prokka were used as input query files as AMRFINDERPLUS.

The following script was used as obtain AMR genes from AMRFINDERPLUS

```
#!/usr/bin/bash

protein_files=(/home/bic/test_finder+/final_run/faa_files/*.faa)
for protein_file in "${protein_files[@]}"
do
        protein_basename=$(basename "$protein_file")
        gff_file="/home/bic/test_finder+/final_run/gff_files/${protein_basename%.*}.gff"
        output_file="/home/bic/test_finder+/final_run/${protein_basename%.*}_output.txt"
        amrfinder -p "$protein_file" -g "$gff_file" --annotation_format prokka > "$output_file"
done
```

2.5.3 CARD

Aside from AMR gene prediction using the tools above, we also tried to find homology using sequence alignment by downloading the CARD database and performing a nucleotide BLAST search with the following parameters.

- percentage identity : 95%
- query coverage: 95%
- word size: 48
- e-value: 1e-10

Command used for making nucleotide blast from the different models provided by CARD

```
makeblastdb -in card.fasta -dbtype nucl  -title card_db -out card_db
```

Script used to perform nucleotide blast

```
#!/usr/bin/bash
for f in /home/bic/Updated_exc/Main_Folder_893/card_analysis/ffn_files/*.ffn
do
        filename=$(basename "$f".ffn)
        db=/home/bic/Updated_exc/Main_Folder_893/card_analysis/card_db

output_file=/home/bic/Updated_exc/Main_Folder_893/card_analysis/output_card/${filename}
_blastn.txt

        blastn -query "$f" -db "$db" -evalue 1e-10  -word_size 48  -qcov_hsp_perc 95
-perc_identity 95   -out "$output_file"  -outfmt 7

done
```

2.6 IDENTIFICATION AND ANNOTATION OF VIRULENCE FACTORS

Full dataset of virulence factors(nucleotide fasta files) was downloaded from VFDB. Nucleotide fasta files(.ffn) files from Prokka were used as input query files.

Commands used to make database of the full dataset of VFDB nucleotide sequences

```
makeblastdb -in VFDB.setb.fasta -dbtype nucl -title vfdb -out vfdb
```

Script used to perform nucleotide blast for ffn files

```
#!/usr/bin/bash
for f in /home/bic/Updated_exc/Main_Folder_893/vfdb_blast/ffn_files/*.ffn
do
        filename=$(basename "$f".ffn)
        db=/home/bic/Updated_exc/Main_Folder_893/vfdb_blast/vfdb

output_file=/home/bic/Updated_exc/Main_Folder_893/vfdb_blast/output_setB/${filename}_b
lastn.txt

        blastn -query "$f" -db "$db" -evalue 1e-10  -word_size 48  -qcov_hsp_perc 95
-perc_identity 95   -out "$output_file"  -outfmt 7

done
```

2.7 SUBSYSTEM ANALYSIS

Pubmed Seed Viewer was to visualize the subsystem category distribution particularly for MAB and MTB.

2.8 SYNTENY ANALYSIS

Sibelia a command line software, was used to perform synteny analysis. Nucleotide fasta files of complete genome representing a different country were used as input files for Sibelia.

Following command was used for synteny analysis

```
Sibelia -s loose 1185650.4.fasta 1185650.1046  -o output_uscan
```

Following command was used for visualization of synteny plot

```
perl "/home/bic/Circos/circos-0.69-9/bin/circos" -conf
"/home/bic/Sibelia/build/output_spain_us/circos/circos.conf"
```

2.9 GO MAPPING

Protein fasta files from the reported AMR genes were retrieved from NCBI and were looked through InterPro database to retrieve the specific GO terms for all the three categories as described previously.

CHAPTER 3 : RESULTS AND DISCUSSION

3.1 GENOMES RETRIEVED FROM BV-BRC

Different number of genomes were obtained as shown in the table below after using the four different keywords as mentioned previously. After sorting out the data with the aforementioned commands on the basis of decided criteria, the union of 4 keywords resulted in 848 genomes, while the three keywords "*abscessus*","*Mycobacteriodes abscessus* OR *Mycobacterium abscessus"* and "*Mycobacteriodes abscessus*" resulted in 72 extra genomes. Thus in total 920 genomes were taken to the next step for analysis.

SEARCH KEYWORD	TOTAL No. OF GENOMES	No. OF GENOMES AFTER ACCORDING TO THE CRITERIA
Mycobacteroides abscessus	2083	920
Abscessus	2153	920
Mycobacterium abscessus	1961	848
Mycobacteroides abscessus OR *Mycobacterium abscessus*	2172	920

Table 2 :Number of genomes from BV-BRC from different keywords

Table 2 represents total number of genomes obtained from BV-BRC with different keyword search and the number of genomes after sorting out according to the criteria mentioned in section 2.1 of Materials and Methods

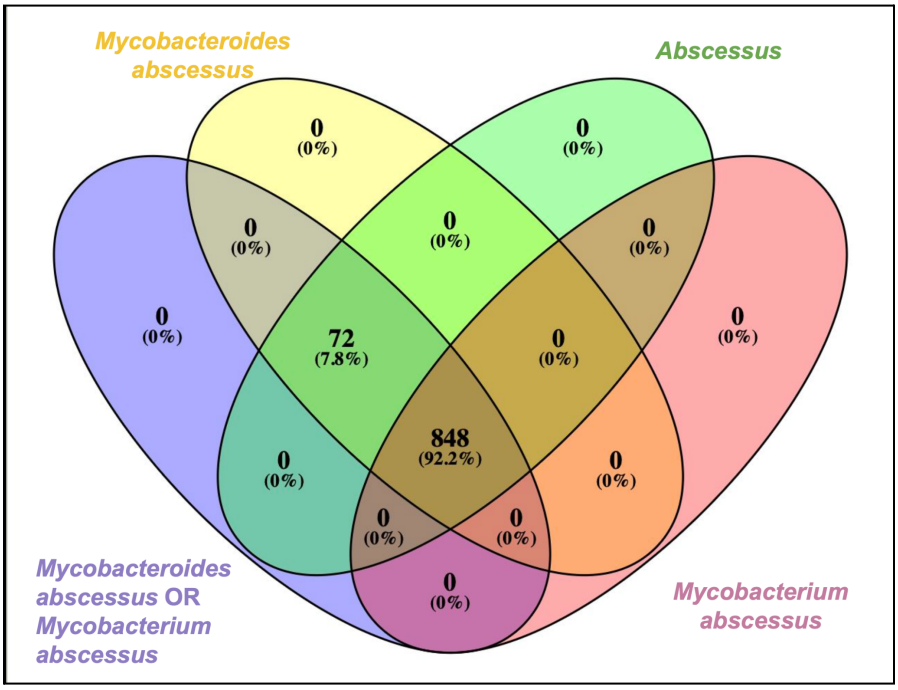

Fig. 2: Common and unique genomes obtained from the set of 4 keywords

However, out of 920 genomes, 281 genomes had no species data mentioned, so they were individually mapped back to NCBI through their BioSample Ids as provided by BV-BRC.

From 281, only 10 genomes had no information about host species as human. The data was either missing or the isolation source was listed as an environment.

Hence, the total number of genomes after sorting out the data and verifying the host species was 910. These fasta files of the 910 genomes were used as an input for the next step that is evaluation of genome quality. Information about these 10 genomes is shared below in Table 2

GENOME ID	BIOSAMPLE ID	GENOME STATUS	HOST SPECIES
1185650.1090	SAMN22365134	complete	missing
1185650.1091	SAMN22365133	complete	missing
1185650.1092	SAMN22365132	complete	missing
1185650.560	SAMEA2275842	WGS	environmental
1185650.566	SAMEA2259657	WGS	environmental
1185650.807	SAMEA2275827	WGS	missing
1185650.808	SAMEA2275830	WGS	environmental
1185650.809	SAMEA2275826	WGS	environmental
1185650.810	SAMEA2275828	WGS	environmental
1185650.811	SAMEA2275829	WGS	environmental

Table 3 : Genomes eliminated due to the absence of host species data

3.2 GENOME QUALITY EVALUATION

After analyzing the nucleotide fasta files, according the criteria for good genomes as provided by WGS sequence guidelines by EUCAST, 17 genomes were eliminated from the total list of 910 genomes, as they had either more than 100 contigs or NG50 less than 30000 base pairs.Thus the number of genomes left after evaluation of genome quality assessment is 893.

GENOME ID	TOTAL NUMBER OF CONTIGS	CONTIGS WITH LENGTH MORE THAN 500bp	NG50
1185650.1006	102	85	181588
1185650.1014	718	367	60645
1185650.1020	102	99	99028
1185650.1024	572	565	14200
1185650.1027	128	121	109321
1185650.1029	132	112	110809
1185650.1030	745	557	29655
1185650.1035	196	171	68984
1185650.130	261	160	246682
1185650.230	283	40	586089
1185650.319	256	150	335388
1185650.375	191	109	374934
1185650.387	145	96	381412
1185650.436	229	132	221127
1185650.597	126	122	81780
1185650.813	130	96	419849
1185650.996	173	56	345171

Table 4 : Genomes Excluded Based on Failure to meet EUCAST Genome Quality Threshold

The table above represents 17 MAA strains that were eliminated because of the failure to meet the threshold as per EUCAST guidelines. Each of the 17 genomes, have contigs more than 100, while genome 1185650.1024 and 1185650.1030 have NG50 less than 30000.

3.3 ANNOTATION FROM PROKKA

With nucleotide fasta files as input, and specifying genus as Mycobacterium, structural and functional annotation was performed with Prokka. Following files were obtained from Prokka for each 893 genome files.

FILE FORMAT	DESCRIPTION
.gff	GFF stands for gene feature format. It is the principal file which contains annotation , coordinates as well as sequence.
.gbk	Standard genbank file for .gff format
.faa	Protein fasta file of input sequence
.fna	Nucleotide fasta file of input sequence
.ffn	Nucleotide fasta file of transcripts by Prokka
.sqn	Sequin file for submission to Genbank after correction for required attributes
.fsa	Same as .fna file along with sequin tags
.tbl	Feature file produced for creating .sqn file
.err	Files containing unacceptable annotations
.log	Contains all the information of annotation file during Prokka's run
.txt	Contains statistics related to predicted annotation
.tsv	Tab-separated files of features predicated by Prokka

Table 5: Different files provided by Prokka for each genome

Out of the above files mentioned above, .gff, .ffn, .fna & .faa files were further used for input and analysis for other tools.

3.4 PAN-GENOME ANALYSIS

A total of 893 genomes were used for pan-gnome construction by Roary at the default settings i.e 95% minimum percentage identity. The pan-genome of MAB had a total of 39841 clusters, with 3525(8.84% of total) genes located in core, 480 (1.20% of total) in soft shell, 1638(4.11% of total) in shell and 34198 (85.83% of total) in cloud. The above data clearly indicates that MAB has an open pan-genome which corroborates with observations made by Choo et al. in 2014.
An open pan-genome increases in size as the number of genomes sequenced increases.

Roary Summary File

Core genes	(99% <= strains <= 100%)	3525
Soft core	(95% <= strains < 99%)	480
Shell	(15% <= strains < 95%)	1638
Cloud	(0% <= strains < 15%)	34198
Total	(0% <= strains <= 100%)	39814

Table 6: Number of clusters in core, soft core, shell and cloud (as provided by roary)

Table 6 and Figure 3 shows the distribution of clusters in core, soft core, shell and cloud, with core accounting for only 9% of the whole pan-genome, whereas clouds accounts for almost 86% of the genome. This shows that a vast majority of genes are strain-specific, whereas only a small number of genes are common to the 893 strains of MAA. These figures indicate that accessory and strain-specific genes should be prioritized in order to understand species diversity.

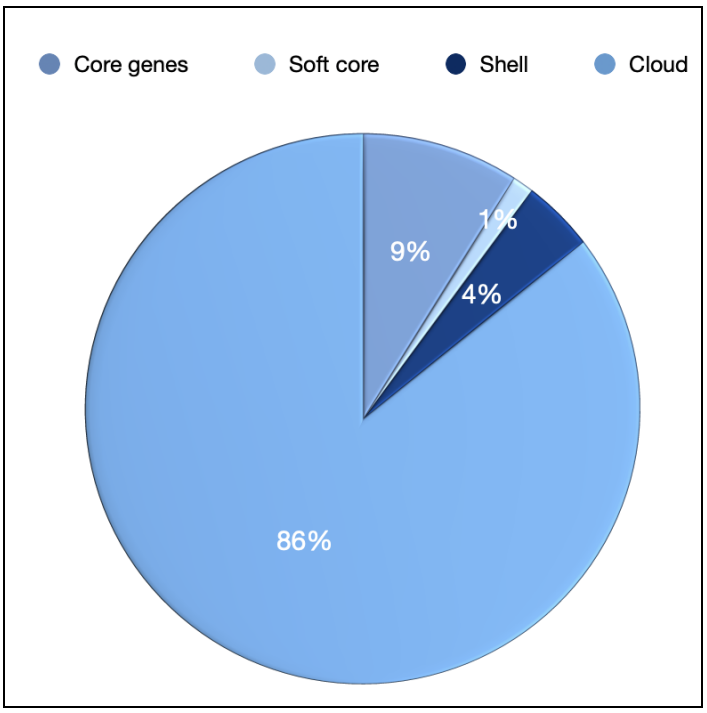

Fig. 3 : Diagrammatic representation of distribution of core, soft core, shell and cloud clusters

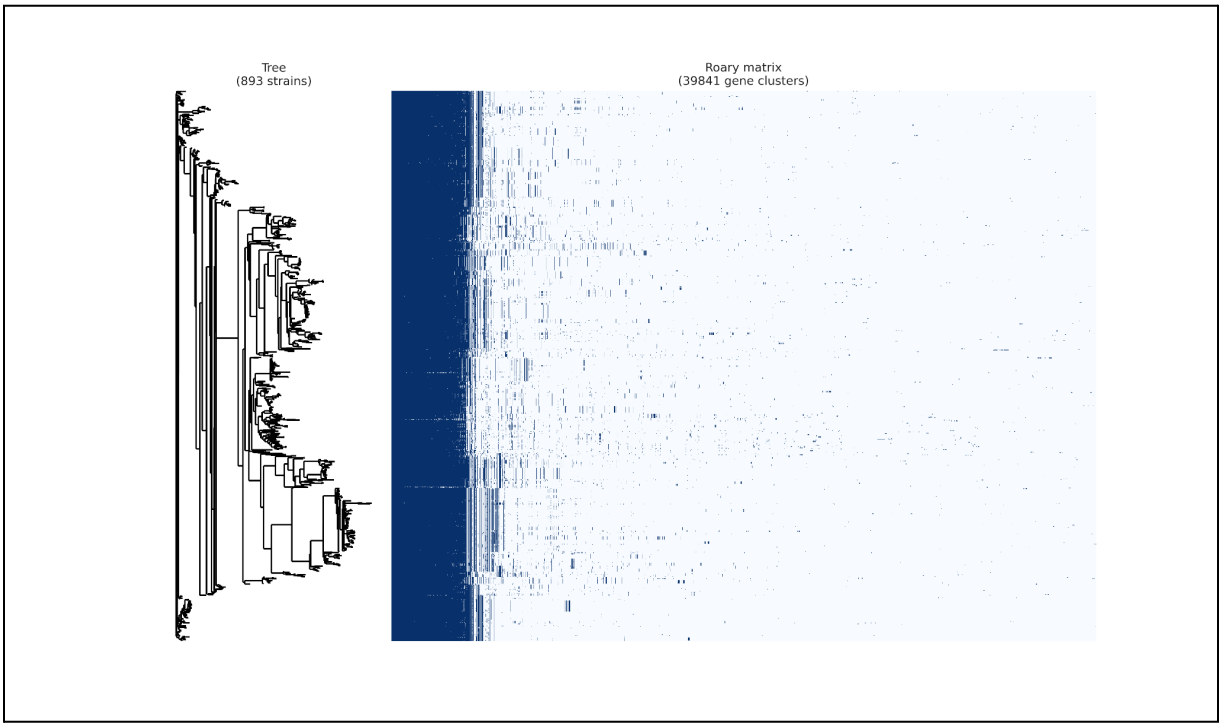

Fig. 4: Matrix representation of core gene clusters in pan-genome of 893 strains of MAA

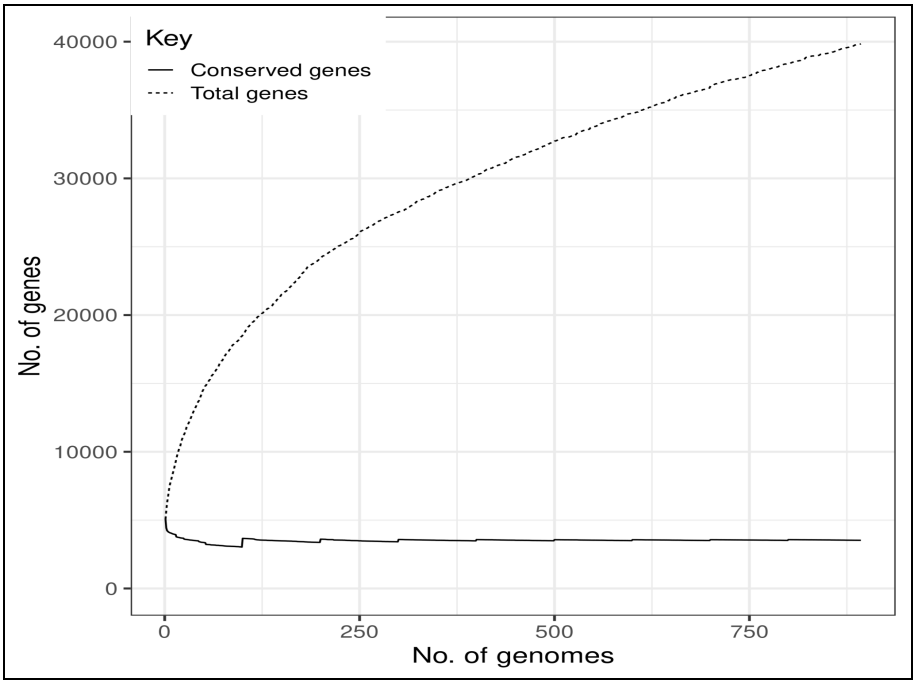

Fig 5: Conserved vs total genes of MAA

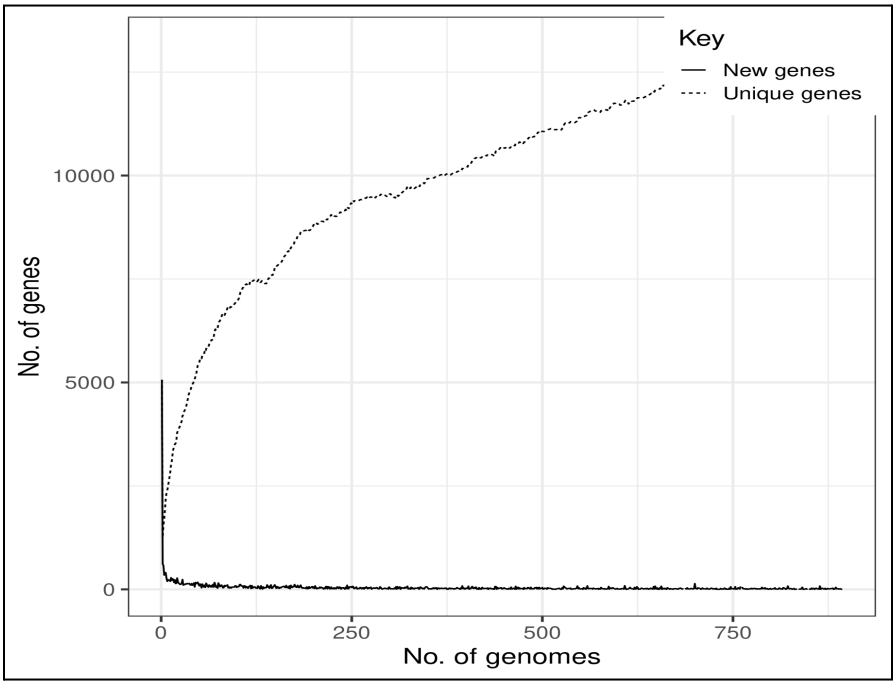

Fig 6: New vs unique genes of MAA

3.5 IDENTIFICATION AND MAPPING OF AMR GENES

3.5.1 RESULTS FROM RGI

S.No	ARO	AMR GENE FAMILY	DRUG CLASS	AMR MECHANISM	No. OF ISOLATES WITH THE PARTICULAR AMR GENE
1	3007180	class A *Mycobacterium abscessus* beta-lactamase	cephalosporin; penam; penem	antibiotic inactivation	893
2	3002956	vanY gene in vanB cluster	glycopeptide antibiotic	antibiotic target alteration	892
3	3003515	Mycobacteroides chelonae 16S rRNA mutation conferring resistance to kanamycin A	aminoglycoside antibiotic	antibiotic target alteration	48
4	3003518	*Mycobacteroides chelonae* 16S rRNA mutation conferring resistance to neomycin	aminoglycoside antibiotic	antibiotic target alteration	48
5	3003516	*Mycobacteroides chelonae* 16S rRNA mutation conferring resistance to tobramycin	aminoglycoside antibiotic	antibiotic target alteration	48

6	3003514	*Mycobacteroides chelonae* 16S rRNA mutation conferring resistance to amikacin	aminoglycoside antibiotic	antibiotic target alteration	48
7	3003240	*Mycobacteroides abscessus* 16S rRNA mutation conferring resistance to gentamicin	aminoglycoside antibiotic	antibiotic target alteration	48
8	3003239	*Mycobacteroides abscessus* 16S rRNA mutation conferring resistance to amikacin	aminoglycoside antibiotic	antibiotic target alteration	48
9	3003238	*Mycobacteroides abscessus* 16S rRNA mutation conferring resistance to neomycin	aminoglycoside antibiotic	antibiotic target alteration	48
10	3003236	*Mycobacteroides abscessus* 16S rRNA mutation conferring resistance to kanamycin	aminoglycoside antibiotic	antibiotic target alteration	48
11	3003237	*Mycobacteroides abscessus* 16S rRNA mutation conferring resistance to	aminoglycoside antibiotic	antibiotic target alteration	48

		tobramycin			
12	3004163	*Mycobacteroides abscessus* 23S rRNA with mutation conferring resistance to clarithromycin	macrolide antibiotic	antibiotic target alteration	889
13	3000603	Erm(41)	macrolide antibiotic; lincosamide antibiotic; streptogramin antibiotic; streptogramin A antibiotic; streptogramin B	antibiotic target alteration	7
14	3000250	ErmC	macrolide antibiotic; lincosamide antibiotic; streptogramin antibiotic; streptogramin A antibiotic; streptogramin B antibiotic	antibiotic target alteration	1
15	3003295	*Mycobacterium tuberculosis* gyrA conferring resistance to fluoroquinolones	fluoroquinolone antibiotic	antibiotic target alteration	5

Table 7: AMR Genes predicted by RGI on the basis of homology and SNP models

The AMR genes predicted by RGI are listed in table 7, column 1 to 4 gives the description about the particular AMR gene in terms of its class and AMR mechanism. Column 5, represents the

numbers of isolates that contain the particular gene, This number was obtained after mapping the output of RGI to the gene presence and absence file of Roary.

This number corresponds to the gene category i.e core, soft core, shell and cloud genes. As clearly represented in the above table, that beta lactamase gene responsible for resistance against cephalosporin and the penems class of antibiotics is located in the core. VanY gene which is responsible for resistance to vancomycin is also located in the core, however resistance to vancomycin is not so far reported in MAB. Aminoglycosides such as amikacin, kanamycin, gentamicin are present in cloud according to the results by RGI. Antibiotic target alteration by erm(41) is the prominent mechanism towards macrolide resistance in MAB, however it is observed to be present in only a few genomes.

3.5.3 RESULTS FROM AMRFINDERPLUS

S.No	ACCESSION	SEQUENCE NAME	No. OF ISOLATES WITH THE PARTICULAR AMR GENE
1	WP_063844243.1	Cmx/CmrA family chloramphenicol efflux MFS transporter	887
2	WP_005091054.1	MAB family class A beta-lactamase	893
3	WP_012377648.1	APH(3'') family aminoglycoside O-phosphotransferase	888
4	WP_063842202.1	NAD(+)--rifampin ADP-ribosyltransferase	890
5	WP_122630832.1	subclass B3 metallo-beta-lactamase	15
6	WP_063842998.1	BLMA family bleomycin binding protein	6
7	WP_001003263.1	23S rRNA (adenine(2058)-N(6))-methyltransferase Erm(C)	1
8	WP_000027050.1	TEM family class A beta-lactamase	1

Table 8: AMR genes predicted by AMRFINDERPLUS

AMRFINDERPLUS utilizes NCBI's curated reference database to find homology between the input query sequence provided by the user. Along with that it also employs HMM models for the prediction of AMR genes. Table 8 lists the AMR genes predicted by AMRFINDERPLUS, whereas column 1-2 lists the accessions and sequence names of the predicted AMR gene. The last column of the table represents the number of isolates, obtained after mapping to the roary gene presence absence file.

3.5.3 RESULTS FROM CARD DATABASE

S.No	ARO	GENE	GENE DESCRIPTION	No. OF ISOLATES WITH THE PARTICULAR AMR GENE
1	3000250	ermC	rRNA adenine N-6-methyltransferase	1
2	3007180	*Mycobacterium abscessus* beta-lactamase	Beta-lactamase	910

Table 9: BLAST results for .ffn files from Prokka vs CARD database for AMR genes

The above table represents the results for sequence alignment between CARD and the .ffn files from Prokka. Only beta-lactamase and erm(C) give BLAST hits. The beta-lactamase gene gives consistent results as in RGI and AMRFINDERPLUS and maps to the core. Whereas the erm(C) is strain-specific.

3.5.4 RESULTS FOR AMR DETERMINANTS FROM LITERARY SOURCES

S.No	GENE	DESCRIPTION	No. OF ISOLATES WITH THE PARTICULAR AMR GENE
1	MAB_0006	DNA gyrase B	892
2	MAB_0163c	Probable phosphotransferase	892
3	MAB_0173	Prenyltransferase family protein UbiA	891
4	MAB_0180	Polyketide synthase PSK13	888
5	MAB_0192c	Probable oxidoreductase	892
6	MAB_0408c	Probable bifunctional membrane-associated penicillin binding protein	892
7	MAB_0540	hypothetical protein	893
8	MAB_0591	Probable rifampin ADP-ribosyl transferase	886
9	MAB_0856c	Putative transcriptional regulator, Tetr family	892
10	MAB_0937c	Putative membrane protein, MmpL family	892
11	MAB_0945	Putative drug resistance drug resistance transporter, EmrB/QacA family	890
12	MAB_1359c	Putative ABC	893

		transporter, ATP binding protein	
13	MAB_1409c	Multidrug efflux transporter Tap	892
14	MAB_1448	ATP synthase C chain AtpE	892
15	MAB_1472c	Putative nicotinamidase/ pyrazinamidase	892
16	MAB_1689	Probable daunorubicin resistance ABC transporter ATP binding domain	890
17	MAB_1846	Putative ABC transporter ATP-binding protein	890
18	MAB_1858	Probable ABC transporter antibiotic transport ATP-binding protein	878
19	MAB_1859	Probable ABC transporter antibiotic transport integral membrane protein	888
20	MAB_1860	Probable ABC transporter antibiotic transport integral membrane protein	888
21	MAB_1877c	3-oxoacyl-[acyl-carrier-protein] synthase 1 KasA	890
22	MAB_2108	Probable undecaprenyl-diphosphate (Bacitracin resistance protein)	892

23	MAB_2208c	Hypothetical protein	892
24	MAB_2299c	Tetr regulator	892
25	MAB_2415C	Conserved hypothetical protein (penicillinase repressor)	892
26	MAB_2644C	Tryptophan synthase, beta subunit Tr	893
27	MAB_2685	Putative transcriptional regulator, Tetr Family	890
28	MAB_2705c	Isoleucyl-tRNA synthetase IleS	893
29	MAB_2722c	Enoyl-(acyl-carrier-protein)reductase (NADH)	893
30	MAB_2780c	Putative transporter	893
31	MAB_2875	Beta- lactamase precursor	893
32	MAB_3080	Putative antibiotic biosynthesis monooxygenase	892
33	MAB_3637c	Putative aminoglycoside phosphotransferase	892
34	MAB_4124	Conserved hypothetical protein (GNAT acetyltransferase?)	890
36	MAB_4237c	Putative amino acid ABC transporter, ATP-binding protein	893
37	MAB_4283c	Conserved hypothetical protein(isoniazid-ind	892

		ucible gene protein INiA)	
38	MAB_4382c	Putative membrane protein MmpL5	596
39	MAB_4383c	Putative membrane protein MmpS5	596
40	MAB_4384	Tetr Regulator	596
41	MAB_4482	Putative phosphotransferase	878
42	MAB_4659	Conserved hypothetical protein(phosphotransferase?)	893
43	MAB_4910c	Putative aminoglycoside phosphotransferase	892
44	MAB_0951	Putative aminoglycoside phosphotransferase	891
45	MAB_3508c	Putative transcriptional regulator	890

Table 10 : AMR genes from literary sources and mapped to pan-genome of 893 MAA strains

Table 10 represents the AMR genes that were mined from different literature sources for MAB and mapped to gene presence absence files provided by the roary.

The table clearly shows that AMR genes that encode for efflux pumps, transporters and enzymes responsible for resistance against tetracyclines and aminoglycosides are present in core . However, only a fraction of these genes were predicted by RGI, AMRFINDERPLUS and CARD. This clearly indicates that already available tools do not suffice for the identification and annotation of AMR genes.This is a concerning finding because it implies that these tools may overlook several AMR genes.

3.6 IDENTIFICATION AND MAPPING FOR VIRULENCE FACTORS

S.No	UNIQUE IDENTIFIER	DESCRIPTION	VF CLASS	No. OF ISOLATES WITH THE PARTICULAR VF
1	VFG022350(gb|YP_001700923)	(erp) exported repetitive protein precursor Erp	Erp (VF0312) - Others (VFC0346)	890
2	VFG022358(gb|YP_001704810)	(hbhA) iron-regulated heparin binding hemagglutinin hbhA	HbhA (VF0313) - Adherence (VFC0001)	892
3	VFG022495(gb|YP_001702285)	(stf0) Conserved hypothetical protein (sulfotransferase?)	Sulfolipid-1 biosynthesis and transport (VF0803)	888
4	VFG022517(gb|YP_001700930)	(fbpA) Secreted antigen 85-a FbpA (mycolyl transferase 85A)	Antigen 85 (VF0311) - Adherence (VFC0001)	884
5	VFG022539(gb|YP_001705126)	(hspX) Heat shock protein HspX (alpha-crystallin homolog)	HspX (VF0301) - Stress survival (VFC0282)	892
6	VFG022545(gb|YP_001704951)	(pknG) Probable serine/threonine-protein kinase	Protein kinase G (VF0805) - Immune modulation (VFC0258)	893
7	VFG022553(gb|YP_001701633)	(lpqH) Hypothetical lipoprotein lpqH precursor	19-kD protein (VF0806) - Immune modulation (VFC0258)	890
8	VFG022554(gb|YP_001703114)	(lpqH) Hypothetical lipoprotein LpqH	19-kD protein (VF0806) -	734

		precursor	Immune modulation (VFC0258)	
9	VFG022561(gb|YP_001704851)	(eis) acetyltransferase	Enhanced intracellular survival protein (VF0807) - Antimicrobial activity/Competitive advantage (VFC0325)	890
10	VFG022581(gb|YP_001704822)	(icl) Isocitrate lyase Icl (isocitrate)	Isocitrate lyase (VF0253) - Others (VFC0346)]	892
11	VFG022609(gb|YP_001701294)	(panC) Probable pantoate--beta-alanine ligase (PanC)	PanC/PanD (VF0319) - Nutritional/Metabolic factor (VFC0272)	893
12	VFG022617(gb|YP_001701295)	(panD) Probable aspartate 1-decarboxylase precursor	PanC/PanD (VF0319) - Nutritional/Metabolic factor (VFC0272)	893
13	VFG022625(gb|YP_001701440)	(purC) Probable phosphoribosylaminoimidazole-succinocarboxamide synthase PurC	Purine synthesis (VF0811) - Nutritional/Metabolic factor (VFC0272)	891
14	VFG022633(gb|YP_001704733)	(proC) Probable pyrroline-5-carboxylate reductase ProC	Proline synthesis (VF0812) - Nutritional/Metabolic factor (VFC0272)	892
15	VFG022641(gb|YP_001702706)	(trpD) Anthranilate phosphoribosyltransfera	Tryptophan synthesis	891

		se TrpD	(VF0813) - Nutritional/Met abolic factor (VFC0272)	
16	VFG022648(gb|YP_001704023)	(leuD) 3-isopropylmalate dehydratase small subunit	Leucine synthesis (VF0814) - Nutritional/Met abolic factor (VFC0272)	893
17	VFG022656(gb|YP_001702174)	(lysA) Probable diaminopimelate decarboxylase	Lysine synthesis (VF0815) - Nutritional/Met abolic factor (VFC0272)	892
18	VFG022664(gb|YP_001702670)	(glnA1) Probable glutamine synthetase, type I (GlnA1)	Glutamine synthesis (VF0816) - Nutritional/Met abolic factor (VFC0272)	891
19	VFG022673(gb|YP_001704322)	(mgtC) Possible Mg2+ transport P-type ATPase C MgtC	MgtC (VF0289) - Nutritional/Met abolic factor (VFC0272)	892
20	VFG022680(gb|YP_001703761)	(ideR) Iron-dependent repressor and activator IdeR	IdeR (VF0300) - Regulation (VFC0301)	892
21	VFG022688(gb|YP_001704827)	(mbtH) MbtH-like protein	GPL locus (VF0841) - Immune modulation (VFC0258)	881
22	VFG022695(gb|YP_001702984)	(mbtG) Lysine-N-oxygenase MbtG (L-lysine 6-monooxygenase)	Mycobactin (VF0299) - Nutritional/Met abolic factor	892

			(VFC0272)	
23	VFG022723(gb|YP_001702855)	(mbtC) Polyketide synthetase MbtC (polyketide synthase)	Mycobactin (VF0299) - Nutritional/Metabolic factor (VFC0272)	892
24	VFG022737(gb|YP_001702982)	Bifunctional enzyme MbtA: salicyl-AMP ligase (SAL-AMP ligase) + salicyl-S-ArCP synthetase	Mycobactin (VF0299) - Nutritional/Metabolic factor (VFC0272)	892
25	VFG022751(gb|YP_001702980)	(mbtI) Isochorismate synthase MbtI	Nutritional/Metabolic factor (VFC0272)	892
26	VFG022774(gb|YP_001703856)	(mbtK) Lysine N-acetyltransferase MbtK	Mycobactin (VF0299) - Nutritional/Metabolic factor (VFC0272)	889
27	VFG022781(gb|YP_001702997)	(irtA) Iron-regulated transporter IrtA	ABC transporter (VF0818) - Nutritional/Metabolic factor (VFC0272)	893
28	VFG022829(gb|YP_001705132)	ahpC) Putative alkyl hydroperoxidase C	AhpC (VF0306) - Stress survival (VFC0282)	892
29	VFG022845(gb|YP_001700872)	(sodA) Probable superoxide dismutase (Mn)	odA (VF0304) - Stress survival (VFC0282)	886
30	VFG022875(gb|YP_001702102)	(sigE) Probable alternative RNA polymerase sigma factor	SigE (VF0295) - Regulation (VFC0301)	891
31	VFG022897(gb|Y	(whiB3) Putative	WhiB3	892

	P_001704454)	transcriptional regulator, WhiB family	(VF0288) - Regulation (VFC0301)	
32	VFG022919(gb|YP_001701702)	(prrA) Probable transcriptional regulatory protein PrrA	PrrA/B (VF0823) - Regulation (VFC0301)	44
33	VFG022935(gb|YP_001701820)	(mprA) Mycobacterial persistence regulator MrpA, (two component response transcriptional regulatory protein)	MprAB (VF0298) - Regulation (VFC0301	893
34	VFG022943(gb|YP_001701821)	(mprB) Probable two component sensor kinase MprB	MprAB (VF0298) - Regulation (VFC0301)	893
35	VFG022951(gb|YP_001704619)	(devR/dosR) Probable transcriptional regulator, LuxR family	DevRS (VF0317) - Regulation (VFC0301)	1
36	VFG022958(gb|YP_001704618)	(devS) Probable histidine kinase response regulator	DevRS (VF0317) - Regulation (VFC0301)	891
37	VFG022965(gb|YP_001703609)	(relA) Probable GTP pyrophosphokinase RelA (ATP:GTP 3'-pyro phosphotransferase) (PPGPP synthetase I) ((P)PPGPP synthetase) (GTP diphosphokinase)	RelA (VF0287) - Regulation (VFC0301)	893
38	VFG023071(gb|YP_001704879)	(mce4B) Hypothetical MCE-family protein	Mce4 (VF0827) - Immune modulation (VFC0258)	892
39	VFG023078(gb|YP_001704878)	(mce4C) Hypothetical	Mce4	892

		MCE-family protein	(VF0827) - Immune modulation (VFC0258)	
40	VFG023109(gb|YP_001701753)	(mce5B) Putative MCE family protein	Mce5 (VF0828) - Others (VFC0346	809
41	VFG023126(gb|YP_001702435)	(mce6B) Putative Mce family protein	Mce6 (VF0829) - Others (VFC0346	890
42	VFG023128(gb|YP_001702436)	(mce6C) Putative Mce family protein	Mce6 (VF0829) - Others (VFC0346	890
43	VFG023130(gb|YP_001702437)	mce6D) Putative Mce family protein	Mce6 (VF0829) - Others (VFC0346	890
44	VFG023132(gb|YP_001702438)	(mce6E) Putative Mce family protein	Mce6 (VF0829) - Others (VFC0346	890
45	VFG023311(gb|YP_001702969)	(eccA3) Type VII secretion system protein EccA3	ESX-3 (VF0521) - Effector delivery system (VFC0086)	892
46	VFG023327(gb|YP_001702968)	(eccB3) Type VII secretion system protein EccB3	[ESX-3 (VF0521) - Effector delivery system (VFC0086)	892
47	VFG023335(gb|YP_001702967)	(eccC3) Type VII secretion system protein EccC3	ESX-3 (VF0521) - Effector delivery system	853

			(VFC0086)	
47	VFG023367(gb|YP_001702962)	(espG3) Type VII secretion system protein EspG3	ESX-3 (VF0521) - Effector delivery system (VFC0086)	811
48	VFG023391(gb|YP_001702959)	(essE3) Type VII secretion system transmembrane protein EccE3	ESX-3 (VF0521) - Effector delivery system (VFC0086)	892
50	VFG023418(gb|YP_001704484)	(eccC4) Putative FtsK/SpoIIIE family protein	SX-4 (T7SS) (VF0836) - Effector delivery system (VFC0086)	893
51	VFG023432(gb|YP_001704486)	(mycP4) Putative protease	ESX-4 (T7SS) (VF0836) - Effector delivery system (VFC0086)	886
52	VFG023546(gb|YP_001703132)	(secA2) accessory Sec system translocase SecA2	Accessory secretion factor (VF0808) - Others (VFC0346)	892

Table 11 : Virulence factors identified from VFDB and mapped to Pan-genome of 893 MAA strains

The table above lists the results from sequence alignment between VFDB and .ffn files. Out of the 52 VF mapped, 49 belongs to the core, while only one factor is located in the shell and two virulence factors are in the cloud. The results clearly indicate that the majority of the VF are common to most of the strains.

3.7 SUBSYSTEM ANALYSIS

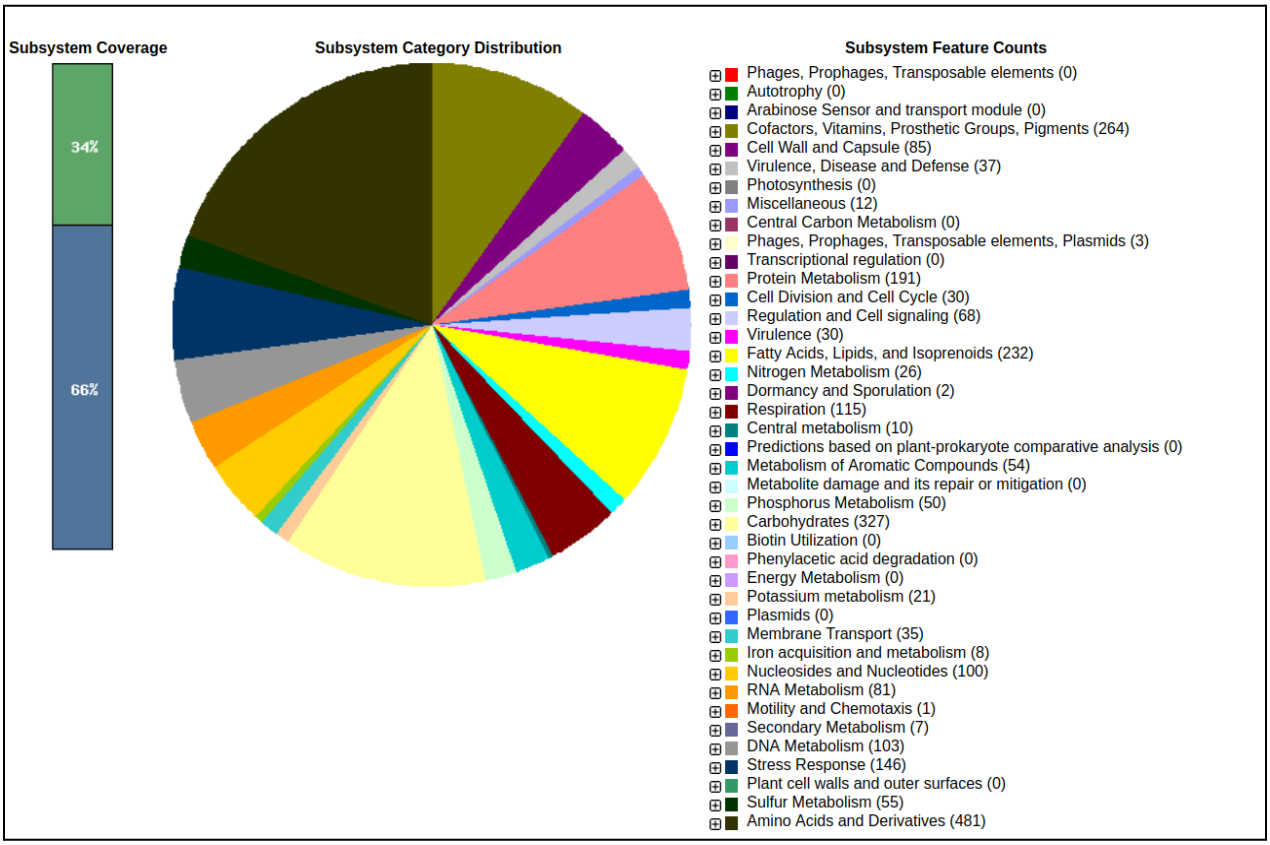

Fig.7 : Diagrammatic representation of subsystem category distribution of *Mycobacterium abscessus* 36809.5

The above figure represents the subsystem category distribution in Mycobacterium abscessus 36809.5. Subsystems are the collection of functional roles that are connected to each other.

These subsystems are manually curated by annotators and are provided through the Seed Viewer platforms for numerous organisms. The bar on the left side shows the subsystem coverage which accounts for only 35% here.

The pie chart depicts the number of subsystem feature counts, here, the category amino acid and Derivatives accounts for most of the subsystem feature counts followed by cofactors, vitamins, prosthetic groups and pigments.

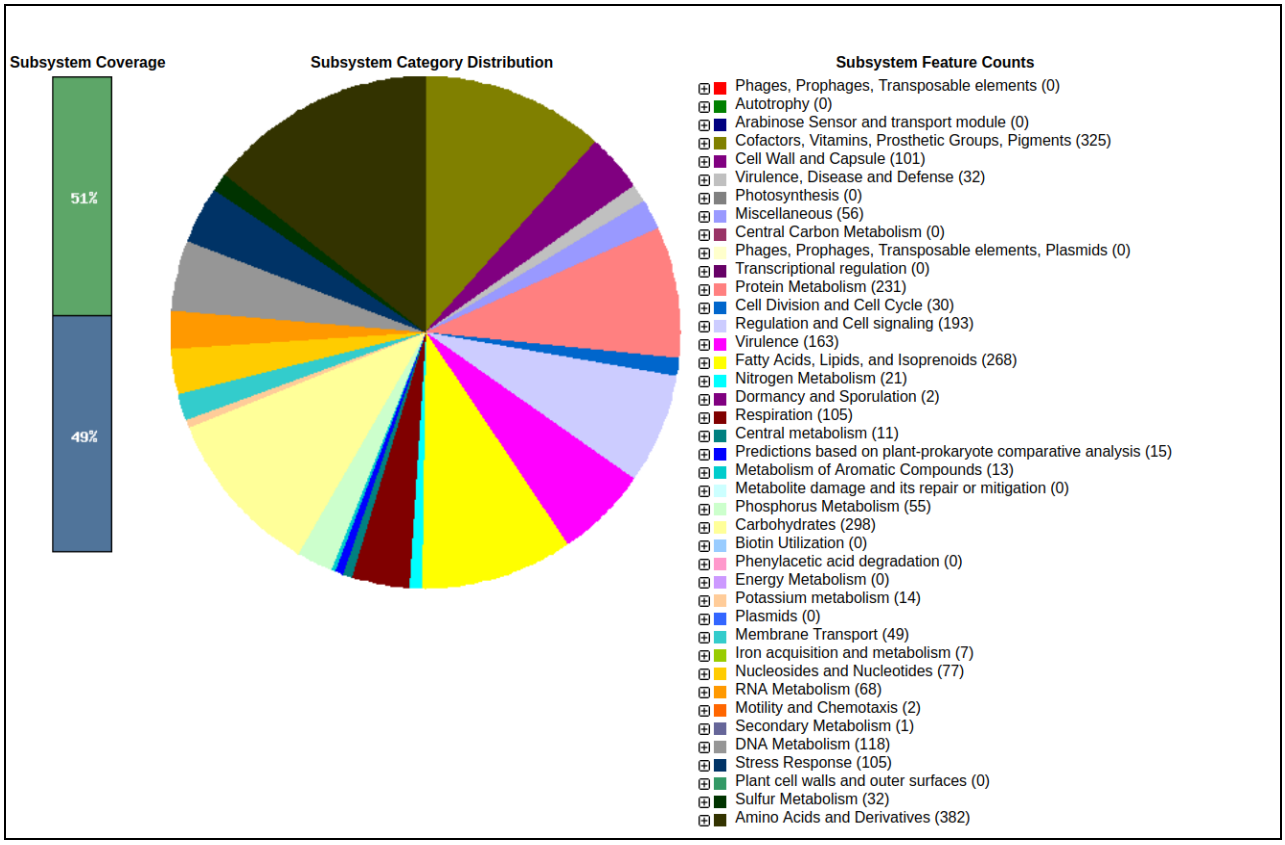

Fig.8: Diagrammatic representation of subsystem category distribution of *Mycobacterium tuberculosis* H37Rv

The above figure represents the distribution of the subsystem category in *Mycobacterium tuberculosis* H37Rv. Here, our objective was to do a comparative analysis between the subsystem category distribution between MAB and MTB. As previously observed in MAB, the subsystem category of amino acids and derivatives accounts for most of the feature counts in MTB. The major difference in terms of number of feature counts, is in the category of regulation and cell signaling.

3.8 SYNTENY ANALYSIS

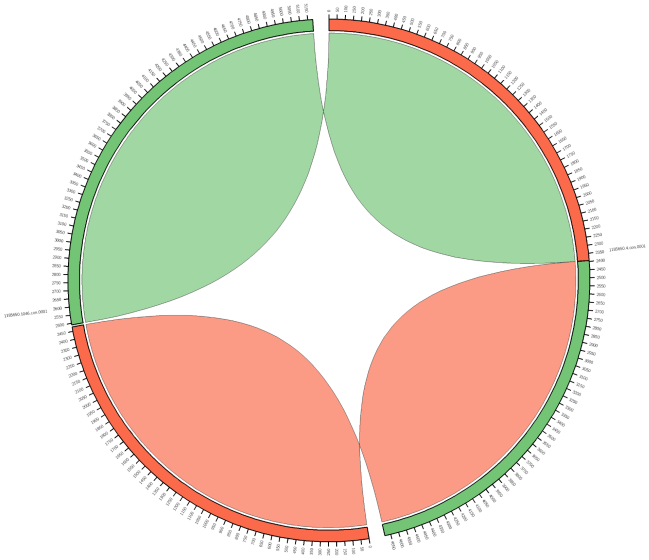

Fig. 9 : Synteny plot for USA vs Canada

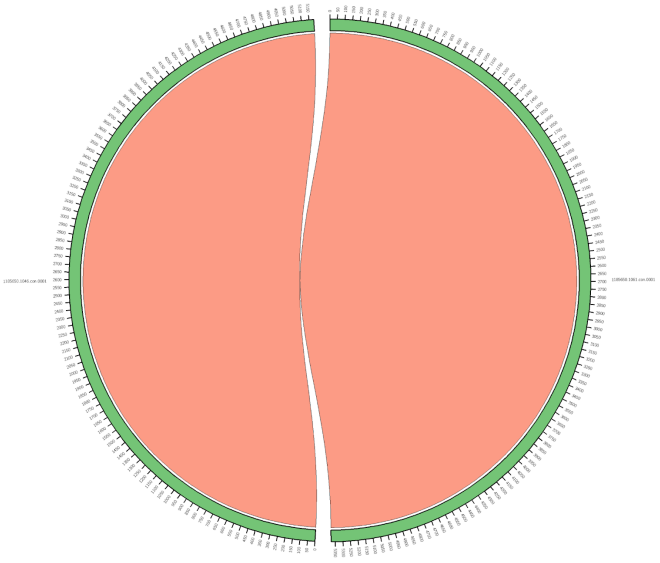

Fig. 10: Synteny plot for Canada vs Australia

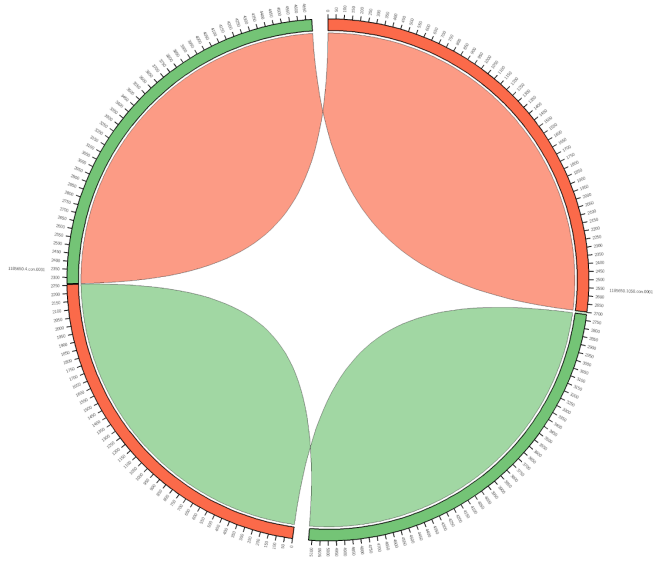

Fig. 11 Synteny plot for Spain vs USA

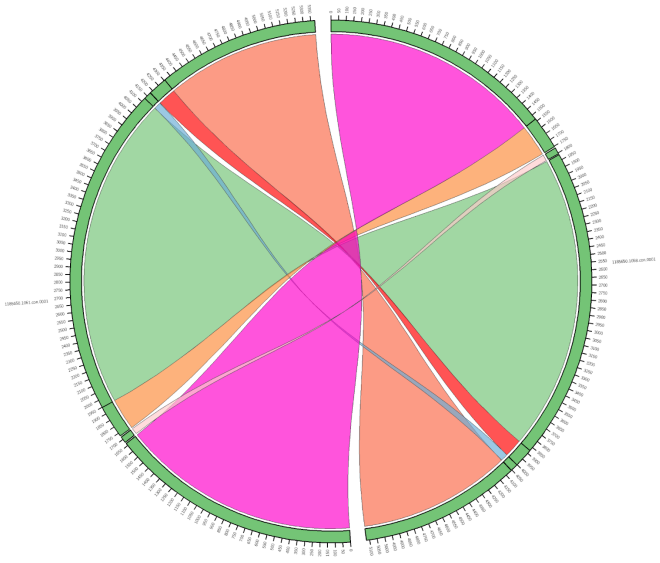

Fig. 12 : Synteny plot for Spain vs Australia

Figure 9, 10, 11, 12 represent the plots generated after synteny analysis for complete genomes of different countries. In figure 9 and 11 the strands are reversed within the same syntenic block. On the other hand, figure 12 shows syntenic regions that have been conserved between the strains of MAB for Spain and Australia.

3.9 GO MAPPING

S.No	GENE	DESCRIPTION	GO:BIOLOGICAL PROCESS	GO: MOLECULAR FUNCTION	GO: CELLULAR COMPONENT
1	MAB_0006	DNA gyrase B	DNA topological change (GO:0006265)	DNA topoisomerase type II (double strand cut, ATP-hydrolyzing) activity (GO:0003918)	chromosome (GO:0005694
				ATP binding (GO:0005524	
				DNA binding (GO:0003677	
2	MAB_0163c	Probable phosphotransferase	response to antibiotic (GO:0046677	ATP binding (GO:0005524)	
				phosphotransferase activity, alcohol group as acceptor(GO:0016773	
3	MAB_0173	Prenyltransferase family protein UbiA		transferase activity, transferring alkyl or aryl (other than methyl) groups (GO:0016765)	membrane (GO:0016020)
4	MAB_0180	Polyketide synthase PSK13	biosynthetic process (GO:0009058)	acyltransferase activity (GO:0016746)	

			fatty acid biosynthetic process (GO:0006633)	phosphopantetheine binding (GO:0031177)	
				transferase activity (GO:0016740)	
				3-oxoacyl-[acyl-carrier-protein] synthase activity(GO:0004315)	
5	MAB_0192c	Probable oxidoreductase		flavin adenine dinucleotide binding (GO:0050660)	membrane (GO:0016020)
				oxidoreductase activity, acting on the CH-OH group of donors, oxygen as acceptor (GO:0016899)	
				D-arabinono-1,4-lactone oxidase activity (GO:0003885	
				FAD binding (GO:0071949)	
6	MAB_0408c	Probable bifunctional membrane-associated penicillin binding protein		penicillin binding (GO:0008658)	
7	MAB_0540	hypothetical protein			

8	MAB_0591	Probable rifampin ADP-ribosyl			
9	MAB_0856c	Putative transcriptional regulator, Tetr family	negative regulation of DNA-templated transcription(GO:0045892)	DNA binding (GO:0003677)	
			response to antibiotic (GO:0046677)		
10	MAB_0937c	Putative membrane protein, MmpL family			membrane (GO:0016020)
11	MAB_0945	Putative drug resistance drug resistance transporter, EmrB/QacA family	transmembrane transport (GO:0055085	transmembrane transporter activity (GO:0022857)	membrane (GO:0016020)
12	MAB_1359c	Putative ABC transporter, ATP binding protein		ATP hydrolysis activity (GO:0016887)	
				ATP binding (GO:0005524)	
13	MAB_1409c	Multidrug efflux transporter Tap	transmembrane transport (GO:0055085)	transmembrane transporter activity (GO:0022857)	membrane (GO:0016020)
14	MAB_1448	ATP synthase C chain AtpE	proton motive force-driven ATP synthesis (GO:0015986)	proton transmembrane transporter activity (GO:0015078)	proton-transporting ATP synthase complex, coupling factor F(o) (GO:0045263)
			proton transmembrane		proton-transporting

			transport (GO:1902600)		two-sector ATPase complex, proton-transporting domain (GO:0033177)
`15	MAB_1472c	Putative nicotinamidase/ pyrazinamidase			
16	MAB_1689	Probable daunorubicin resistance ABC transporter ATP binding domain		ATP hydrolysis activity (GO:0016887)	
				ATP binding (GO:0005524)	
17	MAB_1846	Probable daunorubicin resistance ABC transporter ATP binding domain		ATP hydrolysis activity (GO:0016887)	
				ATP binding (GO:0005524)	
18	MAB_1858	Probable ABC transporter antibiotic transport ATP-binding protein		ATP binding (GO:0005524)	
				ATP hydrolysis activity (GO:0016887	
19	MAB_1859	Probable ABC transporter antibiotic transport integral membrane protein			

20	MAB_1860	Probable ABC transporter antibiotic transport integral membrane protein			
21	MAB_1877c	3-oxoacyl-[acyl-carrier-protein] synthase 1 KasA		acyltransferase activity (GO:0016746)	
22	MAB_2108	Probable undecaprenyl-diphosphate (Bacitracin resistance protein)	dephosphorylation (GO:0016311)	undecaprenyl-diphosphatase activity (GO:0050380)	membrane (GO:0016020
23	MAB_2208c	Hypothetical protein	positive regulation of DNA-templated transcription(GO:0045893	bacterial-type RNA polymerase core enzyme binding(GO:0001000)	
24	MAB_2299	Tetr regulator		DNA binding (GO:0003677)	
25	MAB_2415c	Conserved hypothetical protein (penicillinase repressor)	negative regulation of DNA-templated transcription(GO:0045892)	DNA binding (GO:0003677)	
26	MAB_2644C	Tryptophan synthase, beta subunit Tr	tryptophan biosynthetic process (GO:0000162)	tryptophan synthase activity (GO:0004834)	
			tryptophan metabolic process (GO:0006568)		
27	MAB_2685	Putative transcriptional regulator, Tetr Family	response to antibiotic (GO:0046677)	DNA binding (GO:0003677)	

			negative regulation of DNA-templated transcription(GO:0045892)		
28	MAB_2705c	Isoleucyl-tRNA synthetase IleS	tRNA aminoacylation for protein translation (GO:0006418)	aminoacyl-tRNA ligase activity (GO:0004812)	
			isoleucyl-tRNA aminoacylation (GO:0006428)	tRNA binding (GO:0000049)	
				ATP binding (GO:0005524)	
				nucleotide binding (GO:0000166)	
				isoleucine-tRNA ligase activity (GO:0004822)	
				aminoacyl-tRNA editing activity (GO:0002161	
29	MAB_2722c	Enoyl-(acyl-carrier-protein)reductase (NADH)	fatty acid biosynthetic process (GO:0006633)	enoyl-[acyl-carrier-protein] reductase (NADH) activity(GO:0004318)	
30	MAB_2780c	Putative transporter	transmembrane transport (GO:0055085)	transmembrane transporter activity (GO:0022857)	
31	MAB_2875	Beta- lactamase precursor	beta-lactam antibiotic catabolic	beta-lactamase activity	

			process (GO:0030655	(GO:0008800	
			response to antibiotic (GO:0046677)		
			antibiotic catabolic process (GO:0017001)		
32	MAB_3080	Putative antibiotic biosynthesis monooxygenase			
33	MAB_3637c	Putative aminoglycoside phosphotransferase			
34	MAB_4124	Conserved hypothetical protein (GNAT acetyltransferase?)		acyltransferase activity, transferring groups other than amino-acyl groups (GO:0016747)	
35	MAB_4237c	Putative amino acid ABC transporter, ATP-binding protein	amino acid transmembrane transport (GO:0003333)	ATP hydrolysis activity (GO:0016887)	
				ATP binding (GO:0005524)	
				ABC-type amino acid transporter activity (GO:0015424)	
36	MAB_4283c	Conserved			

		hypothetical protein(isoniazid-inducible gene protein INiA)			
37	MAB_4382c	Putative membrane protein MmpL5			membrane (GO:0016020)
38	MAB_4383c	Putative membrane protein MmpS5			
39	MAB_4384	Tetr Regulator		DNA binding (GO:0003677)	
40	MAB_4482	Putative phosphotransferase			
41	MAB_4659	Conserved hypothetical protein(phosphotransferase?)			
42	MAB_4910c	Putative aminoglycoside phosphotransferase			
43	MAB_0951	Putative aminoglycoside phosphotransferase			
44	MAB_3508c	Putative transcriptional regulator	regulation of DNA-templated transcription (GO:0006355)		
45	MAB_0019	DNA gyrase (subunit A) GyrA (DNA topoisomerase)	DNA topological change (GO:0006265)	DNA topoisomerase type II (double strand cut, ATP-hydrolyzi	chromosome (GO:0005694)

				ng) activity (GO:0003918)	
			DNA metabolic process (GO:0006259)	ATP binding (GO:0005524)	
				DNA topoisomerase activity (GO:0003916)	
				DNA binding (GO:0003677)	
46	MAB_0185c	Probable arabinosyltransferase B	Actinobacterium-type cell wall biogenesis (GO:0071766)	arabinosyltransferase activity (GO:0052636)	
47	MAB_0189c	Probable arabinosyltransferase C	Actinobacterium-type cell wall biogenesis (GO:0071766	arabinosyltransferase activity (GO:0052636)	
48	MAB_1134c	Probable membrane protein, MmpL			membrane (GO:0016020)
49	MAB_1496c	Putative FAD-binding monooxygenase		FAD binding (GO:0071949)	
50	MAB_1497c	Putative regulatory protein, TetR family		DNA binding (GO:0003677)	
51	MAB_1560	Probable ABC transporter (macrolide-transport) ATP-binding protein	negative regulation of translational elongation (GO:0045900)	ATP binding (GO:0005524)	
				ATP hydrolysis activity (GO:0016887)	

52	MAB_1703	Probable aminotransferase		transaminase activity (GO:0008483)	
				pyridoxal phosphate binding (GO:0030170	
53	MAB_1935	Putative drug resistance transporter	transmembrane transport (GO:0055085	transmembrane transporter activity (GO:0022857)	
54	MAB_2301	Putative membrane protein, mmpL			membrane (GO:0016020)
55	MAB_2319c	Probable lysyl-tRNA synthetase 2 LysX	lysyl-tRNA aminoacylation (GO:0006430)	nucleic acid binding (GO:0003676)	cytoplasm (GO:0005737
			tRNA aminoacylation for protein translation (GO:0006418	ATP binding (GO:0005524)	
				lysine-tRNA ligase activity (GO:0004824)	
				nucleotide binding (GO:0000166)	
				aminoacyl-tRNA ligase activity (GO:0004812)	
56	MAB_2355c	Putative ABC transporter ATP-binding protein		ATP binding (GO:0005524)	
				ATP hydrolysis	

				activity (GO:0016887)	
57	MAB_2358	Conserved hypothetical protein			
58	MAB_2643c	Tryptophan synthase, alpha subunit (TrpA)	tryptophan metabolic process (GO:0006568)	tryptophan synthase activity (GO:0004834)	
59	MAB_2958	Putative transmembrane-transport protein	transmembrane transport (GO:0055085)	transmembrane transporter activity (GO:0022857)	
60	MAB_2989	Probable chloramphenicol acetyltransferase		chloramphenicol O-acetyltransferase activity (GO:0008811)	
61	MAB_3499c	Conserved hypothetical protein	proteolysis (GO:0006508)	protein binding (GO:0005515)	
			protein catabolic process (GO:0030163)	ATP-dependent peptidase activity (GO:0004176)	
				serine-type endopeptidase activity (GO:0004252)	
				ATP binding (GO:0005524)	
62	MAB_3851c	30S ribosomal protein S12	translation (GO:0006412)	structural constituent of ribosome (GO:0003735)	small ribosomal subunit (GO:0015935)
					ribosome

					(GO:0005840)
63	MAB_4117c	Putative membrane protein, MmpS family			
64	MAB_4508	Putative membrane protein, MmpL			membrane (GO:0016020)
65	MAB_4532c	Conserved hypothetical protein			

Table 12: GO terms obtained from InterPro for reported AMR genes

Out of 71 AMR reported genes, 65 AMR genes that map to the core of the pan-genome of 893 MAA strains were mapped to their respective GO terms, in all three categories , i.e Biological process, Molecular functions and Cellular components via InterPRO.

CHAPTER 4 : CONCLUSION

In this study, we used a pan-genome approach to map the antimicrobial resistance (AMR) and virulence determinants of MAA.The pan-genome of a bacterial species is the complete set of genes that can be found in any strain of that species. It includes genes that are shared by all strains, as well as genes that are only found in certain strains. Our findings from the pan-genome of 893 MAA strains support the open pan-genome concept, which states that the pan-genome contains a high proportion of strain-specific genes. This suggests that MAA strains have a high level of genetic diversity, which may contribute to their ability to adapt to different environments and resist antimicrobial treatment.

We also tried to annotate AMR and virulence determinants through already available tools such as RGI, AMRFINDERPLUS, CARD and VFDB and also through literature reported determinants. Our findings indicate that, in addition to existing tools, it is critical to review the literature when annotating AMR and virulence determinants.

Efflux pumps, transporters and enzymes responsible for conferring resistance against antibiotics tetracyclines, aminoglycosides are present in core. Almost all virulence factors which are responsible for immune modulation, metabolism, adherence and the factors which provide a competitive advantage are present in core.

Finally, we looked at the GO terms linked to AMR determinants in the core pan-genome. We discovered that these genes are involved in major cellular processes, such as the breakdown and synthesis of fatty acids and amino acids, translation, and transcription and translation regulation. This implies that these genes are involved in a variety of cellular functions, which may contribute to MAA's ability to resist antimicrobial treatment and cause disease.

In addition to the findings presented above, our investigation is still ongoing. Our research aims to gain a complete understanding of MAA's core AMR genes and VF by mapping their distribution across geographical isolation sites. Additionally, we aim to elucidate the fundamental mechanisms underlying these resistant phenotypes, while also dedicating attention to the phenomenon of horizontal gene transfer among different MAA strains.This detailed examination of transmission and pathogenesis will considerably advance our understanding MAA's virulence and antibiotic resistance mechanisms.

Thesis_ver4

ORIGINALITY REPORT

2%
SIMILARITY INDEX

PRIMARY SOURCES

1	card.mcmaster.ca Internet	98 words —	1%
2	www.mdpi.com Internet	20 words —	< 1%
3	pubmed.ncbi.nlm.nih.gov Internet	17 words —	< 1%
4	journals.lww.com Internet	15 words —	< 1%
5	researchmgt.monash.edu Internet	15 words —	< 1%
6	www.sgm.ac.uk Internet	15 words —	< 1%
7	effectors.org Internet	14 words —	< 1%
8	uspto.report Internet	13 words —	< 1%
9	www.ncbi.nlm.nih.gov Internet	12 words —	< 1%

Anshu

EXCLUDE QUOTES	OFF	EXCLUDE SOURCES	OFF
EXCLUDE BIBLIOGRAPHY	ON	EXCLUDE MATCHES	< 10 WORDS

Anshu Bhardwaj
17/7/2023

REFERENCES

1. Rastogi N, Legrand E, Sola C. The mycobacteria: an introduction to nomenclature and pathogenesis. Rev Sci Tech. 2001;20(1):21-54.
2. Lee MR, Sheng WH, Hung CC, Yu CJ, Lee LN, Hsueh PR. Mycobacterium abscessus Complex Infections in Humans. Emerg Infect Dis. 2015;21(9):1638-46.
3. Koh WJ. Nontuberculous Mycobacteria-Overview. Microbiol Spectr. 2017;5(1).
4. Velayati AA, Farnia P, Mozafari M, Malekshahian D, Seif S, Rahideh S, et al. Molecular epidemiology of nontuberculous mycobacteria isolates from clinical and environmental sources of a metropolitan city. PLoS One. 2014;9(12):e114428.
5. Falkinham JO, 3rd. Environmental sources of nontuberculous mycobacteria. Clin Chest Med. 2015;36(1):35-41.
6. Primm TP, Lucero CA, Falkinham JO, 3rd. Health impacts of environmental mycobacteria. Clin Microbiol Rev. 2004;17(1):98-106
7. Boeck L, Burbaud S, Skwark M, Pearson WH, Sangen J, Wuest AW, et al. Mycobacterium abscessus pathogenesis identified by phenogenomic analyses. Nat Microbiol. 2022;7(9):1431-41.
8. Brown-Elliott BA, Wallace RJ, Jr. Clinical and taxonomic status of pathogenic nonpigmented or late-pigmenting rapidly growing mycobacteria. Clin Microbiol Rev. 2002;15(4):716-46
9. Baldwin SL, Larsen SE, Ordway D, Cassell G, Color RN. The complexities and challenges of preventing and treating nontuberculous mycobacterial diseases. PLoS Negl Trop Dis. 2019;13(2):e0007083.
10. Faria S, Joao I, Jordao L. General Overview on Nontuberculous Mycobacteria, Biofilms, and Human Infection. J Pathog. 2015;2015:809014.
11. Hermansen TS, Ravn P, Svensson E, Lillebaek T. Nontuberculous mycobacteria in Denmark, incidence and clinical importance during the last quarter-century. Sci Rep. 2017;7(1):6696.
12. van Ingen J. Diagnosis of nontuberculous mycobacterial infections. Semin Respir Crit Care Med. 2013;34(1):103-9

13. Petrini B. Mycobacterium abscessus: an emerging rapid-growing potential pathogen. APMIS. 2006;114(5):319-28.
14. Griffith DE, Girard WM, Wallace RJ, Jr. Clinical features of pulmonary disease caused by rapidly growing mycobacteria. An analysis of 154 patients. Am Rev Respir Dis. 1993;147(5):1271-8.
15. Adekambi T, Berger P, Raoult D, Drancourt M. rpoB gene sequence-based characterization of emerging non-tuberculous mycobacteria with descriptions of Mycobacterium bolletii sp. nov., Mycobacterium phocaicum sp. nov. and Mycobacterium aubagnense sp. nov. Int J Syst Evol Microbiol. 2006;56(Pt 1):133-43.
16. Bryant JM, Grogono DM, Greaves D, Foweraker J, Roddick I, Inns T, et al. Whole-genome sequencing to identify transmission of Mycobacterium abscessus between patients with cystic fibrosis: a retrospective cohort study. Lancet. 2013;381(9877):1551-60.
17. Sapriel G, Konjek J, Orgeur M, Bouri L, Frezal L, Roux AL, et al. Genome-wide mosaicism within Mycobacterium abscessus: evolutionary and epidemiological implications. Bmc Genomics. 2016;17.
18. Griffith DE, Brown-Elliott BA, Benwill JL, Wallace RJ. Mycobacterium abscessus "Pleased to Meet You, Hope You Guess My Name...". Ann Am Thorac Soc. 2015;12(3):436-9.
19. Lopeman RC, Harrison J, Desai M, Cox JAG. Mycobacterium abscessus: Environmental Bacterium Turned Clinical Nightmare. Microorganisms. 2019;7(3)
20. Degiacomi G, Sammartino JC, Chiarelli LR, Riabova O, Makarov V, Pasca MR. Mycobacterium abscessus, an Emerging and Worrisome Pathogen among Cystic Fibrosis Patients. Int J Mol Sci. 2019;20(23).
21. Fujiwara K, Yoshida M, Murase Y, Aono A, Furuuchi K, Tanaka Y, et al. Potential Cross-Transmission of Mycobacterium abscessus among Non-Cystic Fibrosis Patients at a Tertiary Hospital in Japan. Microbiology Spectrum. 2022;10(3).
22. Viviani L, Harrison MJ, Zolin A, Haworth CS, Floto RA. Epidemiology of nontuberculous mycobacteria (NTM) amongst individuals with cystic fibrosis (CF). J Cyst Fibros. 2016;15(5):619-23.

23. Pennelly KP, Ojano-Dirain C, Yang QP, Liu L, Lu L, Progulske-Fox A, et al. Biofilm Formation by Mycobacterium abscessus in a Lung Cavity. Am J Resp Crit Care. 2016;193(6):692-3.
24. Yan J, Kevat A, Martinez E, Teese N, Johnson K, Ranganathan S, et al. Investigating transmission of Mycobacterium abscessus amongst children in an Australian cystic fibrosis center. J Cyst Fibros. 2020;19(2):219-24.
25. To K, Cao R, Yegiazaryan A, Owens J, Venketaraman V. General Overview of Nontuberculous Mycobacteria Opportunistic Pathogens: Mycobacterium avium and Mycobacterium abscessus. J Clin Med. 2020;9(8).
26. Johansen MD, Herrmann JL, Kremer L. Non-tuberculous mycobacteria and the rise of Mycobacterium abscessus. Nat Rev Microbiol. 2020;18(7):392-407.
27. Griffith DE, Aksamit T, Brown-Elliott BA, Catanzaro A, Daley C, Gordin F, et al. An official ATS/IDSA statement: diagnosis, treatment, and prevention of nontuberculous mycobacterial diseases. Am J Respir Crit Care Med. 2007;175(4):367-416
28. Nessar R, Cambau E, Reyrat JM, Murray A, Gicquel B. Mycobacterium abscessus: a new antibiotic nightmare. J Antimicrob Chemother. 2012;67(4):810-8.
29. Rindi L. Efflux Pump Inhibitors Against Nontuberculous Mycobacteria. Int J Mol Sci. 2020;21(12).
30. Hurst-Hess K, Rudra P, Ghosh P. Mycobacterium abscessus WhiB7 Regulates a Species-Specific Repertoire of Genes To Confer Extreme Antibiotic Resistance. Antimicrob Agents Chemother. 2017;61(11).
31. Wallace RJ, Meier A, Brown BA, Zhang YS, Sander P, Onyi GO, et al. Genetic basis for clarithromycin resistance among isolates of Mycobacterium chelonae and Mycobacterium abscessus. Antimicrob Agents Ch. 1996;40(7):1676-81.
32. Richard M, Gutierrez AV, Kremera L. Dissecting erm(41)-Mediated Macrolide-Inducible Resistance in Mycobacterium abscessus. Antimicrob Agents Ch. 2020;64(2).
33. Nessar R, Reyrat JM, Murray A, Gicquel B. Genetic analysis of new 16S rRNA mutations conferring aminoglycoside resistance in Mycobacterium abscessus. J Antimicrob Chemoth. 2011;66(8):1719-24.

34. Olson RD, Assaf R, Brettin T, Conrad N, Cucinell C, Davis JJ, et al. Introducing the Bacterial and Viral Bioinformatics Resource Center (BV-BRC): a resource combining PATRIC, IRD and ViPR. Nucleic Acids Res. 2022.
35. Gurevich A, Saveliev V, Vyadhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 2013;29(8):1072-5.
36. Ellington MJ, Ekelund O, Aarestrup FM, Canton R, Doumith M, Giske C, et al. The role of whole genome sequencing in antimicrobial susceptibility testing of bacteria: report from the EUCAST Subcommittee. Clin Microbiol Infec. 2017;23(1):2-22.
37. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014;30(14):2068-9.
38. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG, et al. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics. 2015;31(22):3691-3.
39. McArthur AG, Waglechner N, Nizam F, Yan A, Azad MA, Baylay AJ, et al. The Comprehensive Antibiotic Resistance Database. Antimicrob Agents Ch. 2013;57(7):3348-57.
40. Feldgarden M, Brover V, Haft DH, Prasad AB, Slotta DJ, Tolstoy I, et al. Validating the AMRFinder Tool and Resistance Gene Database by Using Antimicrobial Resistance Genotype-Phenotype Correlations in a Collection of Isolates (vol 63, e00483-19, 2019). Antimicrob Agents Ch. 2020;64(4).
41. Feldgarden M, Brover V, Gonzalez-Escalona N, Frye JG, Haendiges J, Haft DH, et al. AMRFinderPlus and the Reference Gene Catalog facilitate examination of the genomic links among antimicrobial resistance, stress response, and virulence. Sci Rep-Uk. 2021;11(1).
42. Overbeek R, Begley T, Butler RM, Choudhuri JV, Chuang HY, Cohoon M, et al. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. Nucleic Acids Res. 2005;33(17):5691-702.
43. Alcock BP, Raphenya AR, Lau TTY, Tsang KK, Bouchard M, Edalatmand A, et al. CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. Nucleic Acids Res. 2020;48(D1):D517-D25.
44. Chen LH, Yang J, Yu J, Ya ZJ, Sun LL, Shen Y, et al. VFDB: a reference database for bacterial virulence factors. Nucleic Acids Res. 2005;33:D325-D8.

45. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, et al. Circos: An information aesthetic for comparative genomics. Genome Res. 2009;19(9):1639-45.