Acta Biologica Szegediensis

Volume 66(2):192-201, 2022
Acta Biologica Szegediensis
http://abs.bibl.u-szeged.hu/index.php/abs
DOI:10.14232/abs.2022.2.192-201

ARTICLE

The pangenome structure of human pathogen *Mycobacterium kansasii*

Saubashya Sur*

Postgraduate Department of Botany, Life Sciences Block, Ramananda College, Bishnupur-722122, West Bengal, India

ABSTRACT The non-tuberculous *Mycobacterium kansasii*, is the causative agent of destructive pulmonary and extrapulmonary infections in immunocompromised persons. Incessant use of multiple antibiotics and lack of effective vaccines did little to combat *M. kansasii* mediated infections. Here, a bioinformatic analysis has been carried out using PanExplorer, to analyze the pangenome aimed at functional characterization of the bacterium, understanding it's pathogenic lifestyle and recognize the factors shaping evolution and variations amongst strains. *M. kansasii* had a large core genome (60.2%), a small (11.9%) dispensable genome and 27.9% strain-specific genes. The core genome of *M. kansasii* had a high concentration of COGs (Cluster of orthologous genes) linked to energy production and conversion, amino acid transport and metabolism, nucleotide transport and metabolism, coenzyme transport and metabolism, and secondary metabolite biosynthesis, transport and metabolism. Interestingly, numerous genes within the core and dispensable genome were associated with pathogenesis and virulence. Noteworthy among them were type VII secretion, ESX, PP and PPE family proteins. Although, *M. kansasii* genomes revealed overall relatedness and conservation, genomic rearrangements caused variability within the strains. The information from this analysis could assist future microbial genomics research on *M. kansasii*, and further studies, e.g., concerning distinctive gene clusters, and evolution.

Acta Biol Szeged 66(2):192-201 (2022)

KEY WORDS
bioinformatics
non-tuberculous mycobacteria
Mycobacterium kansasii
pangenome
pathogenesis

ARTICLE INFORMATION
Submitted
02 November 2022
Accepted
15 February 2023
***Corresponding author**
E-mail: saubashya@gmail.com

Introduction

Non-tuberculous mycobacteria (NTM) are ubiquitously present mycobacteria, responsible for causing opportunistic infections in humans (Ricketts et al. 2014). Most of the NTMs are environmental bacteria (Luo et al. 2021), incorporating over 170 different species, and are associated with skin and pulmonary diseases in humans (Sur and Pal 2021). The incidence of NTM infections is increasing at a fast pace, and there is evidence of acquired infection from environmental sources and human-to-human transmission (Jia et al. 2021). Additionally, there are instances of the zoonotic potential of NTM-mediated infections (Fukano et al. 2021). Anthropogenic activities, immunocompromisation, antibiotic resistance, and lack of effective vaccines have contributed to the surge of NTM infections (Bryant et al. 2016; Sur 2021).

Mycobacterium kansasii is a NTM responsible for destructive pulmonary infections resembling tuberculosis, in individuals suffering from chronic bronchitis, cystic fibrosis, emphysema, and chronic obstructive pulmonary disease (Banks et al. 1983; Ricketts et al. 2014; Luo et al. 2021). *M. kansasii* is also associated with extrapulmonary infections like septic arthritis, and skin and cervical lymph node diseases (Bernard et al. 1999; Guan et al. 2020). *M. kansasii* is commonly found in immunocompromised individuals (e.g., people suffering from HIV) and is increasingly becoming a matter of concern in the USA, South America, Africa, China, and Japan (DeStefano et al. 2018; Luo et al. 2021). Out of the seven subtypes of *M. kansasii,* type I is the major one with a global presence and greater association with human diseases (Guan et al. 2020; Guo et al. 2022). *M. kansasii*-mediated infections can be controlled using a combination of antibiotics like isoniazid, rifampin, and ethambutol; azithromycin and clarithromycin; as well as amikacin, moxifloxacin, and linezolid (DeStefano et al. 2018; Guo et al. 2022). However, usage of multiple antibiotics, prolonged duration of treatment, and increasing drug resistance became limiting factors (DeStefano et al. 2018).

Over the last 20 years, rapid advancement in microbial sequencing technologies catalyzed the exponential growth of bacterial genomes (Zhao et al. 2012). The massive amount of data necessitated the development of efficient computational tools (Perrin and Rocha 2021).

Investigation of the complete genomes of bacterial pathogens and comparative analysis of their various strains offered insights into pathogen biology (Periwal et al. 2015). Moreover, it furnished crucial information regarding the diversity and genomic variability of bacterial pathogens (Kweon et al. 2015). The pangenome approach has been used to explore selection, evolution, conservation, strain-specific virulence, detect functional gene variability, etc. in pathogens (Zhao et al. 2012; Page et al. 2015; Periwal et al. 2015). Pangenome denotes the fusion of the total genetic pool shared by various strains of the species of interest (Periwal et al. 2015). It houses the core and dispensable (accessory) genome (Dereeper et al. 2022). Recent years witnessed some studies on mycobacterial pangenomes (Dumas et al. 2015; Periwal et al. 2015; Zakham et al. 2021) using bioinformatic tools.

This computational study is aimed at elucidating the pangenome architecture of *M. kansasii.* In this context, identification and distribution of the core genome, dispensable genome, and strain-specific genes were performed. This was followed by cluster analysis. Additionally, the distribution of COG (Cluster of Orthologous genes) functional categories (Sen et al. 2008) in *M. kansasii* genomes was studied, to estimate their representation in the genomes. Moreover, synteny between the genomes was assessed to recognize the conservation of gene order between them. The outcome is expected to offer insights into the selection, evolution, and pathogenic lifestyle of *M. kansasii.* Again, it may assist in grasping the role of functional gene clusters in pathogenesis, sensitivity, and resistance. This in turn may benefit future diagnostics and drug development to control *M. kansasii*-mediated infections.

Materials and methods

Retrieval and screening of sequences

The genomes of *M. kansasii* were searched in the NCBI database (https://www.ncbi.nlm.nih.gov/data-hub/genome/). The search revealed the occurrence of 47 genomes of various strains of *M. kansasii* (publicly available as of 15/07/2022). These 47 genomes were subjected to screening, by restricting the assembly status to "complete" or "chromosome" and annotation status to "complete". This quality control step was crucial before pangenome analysis, to filter out poorly/partially annotated and fragmented assembly data since, these low-quality datasets often lead to a fallacious outcome. This was also vindicated by PanExplorer software (Dereeper et al. 2022) since it restricts partially annotated and those with fragmented assemblies. Thus, the resultant dataset consisted of 12 *M. kansasii* genomes. This is represented in Table 1.

Table 1. Dataset of studied *Mycobacterium kansasii* genomes

No.	Name of the microorganism	GenBank assembly ID
1	*M. kansasii* 1MK	GCA_002085625.1
2	*M. kansasii* 4MK	GCA_002085645.1
3	*M. kansasii* 6MK	GCA_002085775.1
4	*M. kansasii* 10MK	GCA_002085795.1
5	*M. kansasii* 11MK	GCA_002085815.1
6	*M. kansasii* 9MK	GCA_002085835.1
7	*M. kansasii* FDAARGOS_1534	GCA_020341475.1
8	*M. kansasii* FDAARGOS_1615	GCA_021183685.1
9	*M. kansasii* FDAARGOS_1614	GCA_021183705.1
10	*M. kansasii* FDAARGOS_1616	GCA_021183865.1
11	*M. kansasii* Kuro-I	GCA_014701265.1
12	*M. kansasii* ATCC12478	GCA_000157895.2

Software

The PanExplorer software (Dereeper et al. 2022) was used to explore the pangenome of *M. kansasii.*

Data processing

The Genbank assembly identifiers of the 12 genomes of *M. kansasii* were entered, and the pangenome analysis pipeline PGAP (Zhao et al. 2012) was selected for data processing. Next, Genbank assemblies were scanned by PanExplorer for compatibility. Here, *M. kansasii* Kuro-I (Genbank assembly ID: GCA_014701265.1) was discarded by PanExplorer owing to incompatibility. Hence, the final dataset consisted of 11 genomes of *M. kansasii.* The minimum percentage identity for BLAST was set at 80%.

Investigation of core and dispensable genomes

Post data processing, the distribution of genes comprising the core genome (genes incident in all the strains) and dispensable genome (genes forming a part of the accessory genome) were investigated (Dereeper et al. 2022). Strain-specific genes from each *M. kansasii* genome were also identified. The interactive presence/absence matrix was explored to inspect the genes in clusters from a subset of strains. The intersections within the dispensable genome were determined. Additionally, the size of various intersections along with their degree of abundance was evaluated.

Analysis of COG functional categories

RPSblast (Tatusov et al. 2000) was utilized by PanExplorer, to probe into the allotment and degree of representation of COG functional categories in 11 *M. kansasii* genomes.

Exploratory analysis

The core and strain-specific genes from 11 *M. kansasii* genomes were portrayed in a Circos, to explore similari-

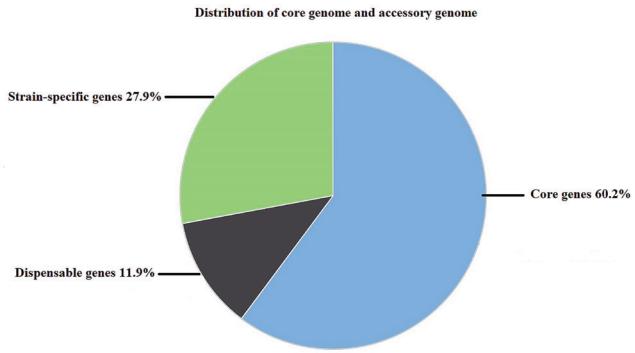

Figure 1. Distribution of core genome and dispensable (accessory) genome of *M. kansasii*. The pie-chart depicts the percentage of core genes, dispensable genes and strain-specific genes.

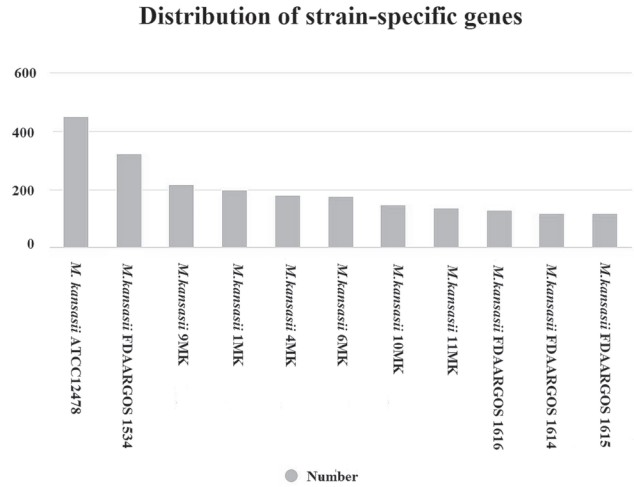

Figure 2. Distribution of strain-specific genes in *M. kansasii*.

ties and differences (Krzywinski et al. 2009). The genes were colored based on their COG categories. Synteny between the genomes was evaluated by selecting three genomes at a time and hive plots (Krzywinski et al. 2012) were generated to visualize rearrangements.

Results

General features

In this work, a comparative genome analysis was carried out to describe the pangenome of *M. kansasii*. The data pool consisted of 11 complete and fully annotated genomes. These strains derived from NCBI were human pathogens isolated from different countries. The genome sizes of these 11 *M. kansasii* strains ranged from 6.4-6.7 MB. The total number of genes ranged between 5669 to 5901, with a high average GC content of 66%. The number of protein-coding genes ranged from 5327 to 5483. *M. kansasii* FDAARGOS 1534 had the largest genome (5901 genes) while *M. kansasii* FDAARGOS 1614 was the smallest (5669 genes).

The pan, core, and dispensable genome

The pangenome of 11 complete genomes of *M. kansasii* strains consisted of 7852 genes (Fig. 1). It portrayed the overall genetic content of the strains. Out of these, 4728 genes (60.2%) formed the core genome and 933 genes (11.9%) formed the dispensable genome. The number of strain-specific genes was 2191 (27.9%).

There were 1241 hypothetical protein genes within the core genome. The rest 3487 genes within the core genome were associated with essential functions, crucial for the survival, metabolism, and pathogenic lifestyle of *M. kansasii*. These included genes linked to different types of transcription, translation, energy metabolism, lipid metabolism, carbohydrate and amino acid transport, DNA repair, recombination, cell division, motility, cellular secretion, posttranslational modification, pathogenesis and virulence. There were substantial numbers of ABC transporters, transcriptional regulators, TetR family transcriptional regulators, type VII secretion system proteins, universal stress proteins, VapC toxin family PIN domain ribonucleases, WXG100 family type VII secretion targets, 2Fe-2S-binding proteins, acetyl-CoA acetyltransferases, acyl-CoA dehydrogenases, alpha/beta hydrolases, cytochrome P450s, enoyl-CoA hydratases, ESX secretion-associated proteins, GntR family transcriptional regulators, LLM class F420-dependent oxidoreductases, MFS transporters, oxidoreductases, PE family proteins, polyketide synthases, PPE family proteins, and SAM-dependent methyltransferase short-chain dehydrogenase within the core genome. Thus, most of these genes are related to housekeeping activities.

The dispensable genome too had a significant share of hypothetical protein genes (520 out of 933). The remaining 413 genes within the dispensable genome were associated with myriad functions. Several genes related to pathogenesis and virulence are accommodated within the dispensable genome. Fig. 2. illustrated the distribution pattern of strain-specific genes in *M. kansasii* genomes. *M. kansasii* ATCC12478 had the highest number of strain-specific genes (451) followed by *M. kansasii* FDAARGOS 1534 (325) and *M. kansasii* 9MK (217). *M. kansasii* 1MK, *M. kansasii* 4MK, *M. kansasii* 6MK, *M. kansasii* 10MK, and *M. kansasii* 11 MK had 199, 179, 175, 146, and 135 strain-specific genes respectively. On the other hand, *M. kansasii* FDAARGOS 1614 (118) and *M. kansasii* FDAARGOS 1615 (118) had the lowest number of strain-specific genes. Fig. 3. portrayed the variation of gene clusters orders based on hierarchical clustering. The outcome of pangenome

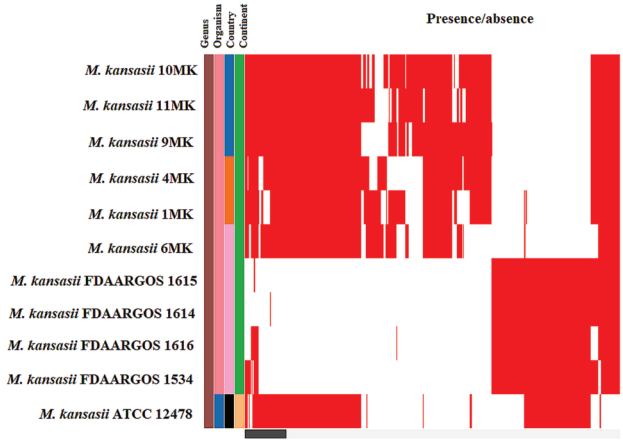

Figure 3. Snapshot of the presence/absence matrix showing how gene clusters have been ordered based on hierarchical clustering. White and red colours indicate absence and presence, respectively.

analysis indicated that the *M. kansasii* pangenome had a large core genome and a small dispensable genome. This implied a closed pangenome, which added a smaller number of gene families while new genomes were integrated during pangenome analysis.

Fig. 4. demonstrated the abundance of intersections within the dispensable genome of *M. kansasii*. It revealed that there are 9 intersections. The largest intersection incorporated genes shared between 10 genomes, while the smallest one consisted of genes shared between 2 genomes. It is evident from Fig. 4. that moderate number of genes were shared between the *M. kansasii* strains. The *M. kansasii* strains 4MK, 10MK, 11MK, 6MK and 9MK seemed to be strongly related owing to their incidence in the top three intersections. Likewise, the *M. kansasii* strains FDAARGOS 1614, 1615, 1616 and 1534 were distinctly related due to their occurrence in two large sized intersections. On the other hand, *M. kansasii* 1MK and *M. kansasii* ATCC12478 showed some variation from the other strains due to their presence in one of the largest intersections.

Evaluation of COG functional categories

The evaluation of the functional distribution of COGs within the pangenome components divulged some interesting features. The four COG categories were: 1) Information and storage processing (sub categories J, A, K, L, and B); 2) Cellular processes and signaling (sub categories D, Y, V, T, N, M, Z, W, U, and O); 3) Metabolism (sub categories C, G, E, F, H, I, P, Q); 4) Poorly characterized (sub categories R, and S). It was noticed that COGs associated with metabolism, was the most abundant category followed by poorly characterized in *M. kansasii* genomes. While COGs associated with cellular processes and signaling had a modest presence, those related to information and storage processing were least represented. It was also found that relative abundance of the distribution of COG categories, was comparable among the *M. kansasii* strains.

A substantial number of genes within the core (1138) and dispensable (519) were not assigned to any of the COG categories. The genes related to metabolism in the core and dispensable genome were 58.16% and 17.47% respectively. A considerable proportion of the genes within COG functional category metabolism, belonged to the functional subcategories viz. energy production and conversion (C), amino acid transport and metabolism (E), nucleotide transport and metabolism (F), coenzyme transport and metabolism (H) and secondary metabolite biosynthesis, transport and metabolism (Q). Among the poorly characterized COG functional category, subcategories general functional prediction (R) and function unknown (S) had a noteworthy representation in the core and dispensable genome. In fact, the latter had 19.18% genes in the poorly characterized COG functional category. Apart from these, the core genome consisted of genes in COG functional categories cellular process and signaling and information and storage processing. The functional subcategories with fair share of genes included, defense mechanisms (V), cell wall/membrane/

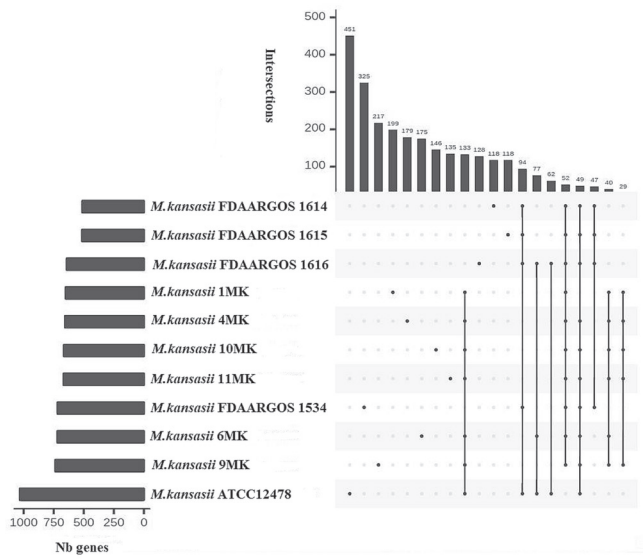

Figure 4. Upset diagram illustrates the degree of abundant intersections within the dispensable genome of *M. kansasii*. The top barplot depicts the incidence of each intersection and the bottom plot highlights the intersections. While each column resembled an intersection between *M. kansasii* strains, the bar charts display the size of the intersection. The rows indicate possible intersections and, the filled-in cells correspond to the particular *M. kansasii* strain involved in the intersection. Nb indicates number of genes.

Table 2. *Mycobacterium kansasii* genome datasets showing rearrangements

No.	Genome datasets
1	*M. kansasii* 10MK, FDAARGOS 1534, 1614
2	*M. kansasii* 10MK, FDAARGOS 1614, 1616
3	*M. kansasii* 11MK, FDAARGOS 1534, 1614
4	*M. kansasii* 11MK, FDAARGOS 1614, 1615
5	*M. kansasii* 11MK, FDAARGOS 1615, 1616
6	*M. kansasii* 1MK, ATCC12478, FDAARGOS 1534
7	*M. kansasii* 1MK, FDAARGOS 1534, 1614
8	*M. kansasii* 1MK, FDAARGOS 1614, 1615
9	*M. kansasii* 1MK, FDAARGOS 1615, 1616
10	*M. kansasii* 4MK, ATCC12478, FDAARGOS 1534
11	*M. kansasii* 4MK, FDAARGOS 1534, 1614
12	*M. kansasii* 4MK, FDAARGOS 1614, 1615
13	*M. kansasii* 4MK, FDAARGOS 1615, 1616
14	*M. kansasii* 6MK, ATCC12478, FDAARGOS 1534
15	*M. kansasii* 6MK, FDAARGOS 1614, 1615
16	*M. kansasii* 6MK, FDAARGOS 1615, 1616
17	*M. kansasii* 9MK, FDAARGOS 1534, 1614
18	*M. kansasii* 9MK, FDAARGOS 1614, 1615
19	*M. kansasii* 9MK, FDAARGOS 1615, 1616
20	*M. kansasii* ATCC12478, FDAARGOS 1534, 1614
21	*M. kansasii* ATCC12478, FDAARGOS 1614, 1615
22	*M. kansasii* ATCC12478, FDAARGOS 1615, 1616
23	*M. kansasii* FDAARGOS 1534, 1614, 1615
24	*M. kansasii* FDAARGOS 1534, 1615, 1616
25	*M. kansasii* FDAARGOS 1614, 1615, 1616

envelope biogenesis (M), signal transduction mechanisms (T), cell motility (N) and post translational modification, protein turnover, chaperone (O), translation, ribosomal structure, and biogenesis (J), transcription (K) and replication, recombination and repair (L).

At the strain-specific level, all *M. kansasii* strains had abundant genes in the COG subcategory general function prediction (R), while COG subcategories viz., intracellular trafficking, secretion and vesicular transport (U) and cell cycle control, cell division, chromosome partitioning (D) had the least. Besides, other COG subcategories having modest representation were amino acid transport and metabolism (E), inorganic ion transport metabolism (P) and energy production and conversion (C). COG functional subcategories lipid transport and metabolism (I), signal transduction mechanism (T), posttranslational modification, protein turnover, chaperones (O), transcription (K) and function unknown (S) too had some presence in the strain-specific genes. Furthermore, it was observed that *M. kansasii* FDAARGOS 1534 had 29.23% of strain-specific genes in COGs, whereas in *M. kansasii* FDAARGOS 1616 it was 16.40%. In other strains it ranged between 16.94% to 25.13%.

Physical maps and synteny of M. kansasii genomes

Fig. 5 (a-c). represented the physical map of the genes from *M. kansasii* genomes. It was observed that there is very little difference between the strains in terms of genome size. *M. kansasii* 6MK had large number of genes that are not associated with COGs. That *M. kansasii* genomes show variation, especially in the core genes and strain-specific genes at the COGs level were visible from Fig. 5 (a-c).

Supplementary figures (Suppl. Fig. 1-8) displayed the hive plots of whole genome comparisons between pre-selected *M. kansasii* genome datasets. On the whole, absence of gaps in the direction of the center of all the hive plots revealed synteny between *M. kansasii* genomes. However, several hive plots depicted inverted regions between the genomes. Additionally, it is clear from these figures (Suppl. Fig. 1-8) that genome rearrangements occurred in the genome datasets shown in Table 2. The links between the genomes are mediated by the core genes, since they form connections between genomes. Each node corresponded to a core gene between the dataset of three genomes. Links colored with a color gradient estimated genomic rearrangements. These core genes formed locally colinear blocks (LCBs) between pairs of aligned genomes that are connected.

Discussion

M. kansasii mediated diseases have been increasing steadily in different parts of the world. The clinical significance of these infections often resulted in severity and even fatality in immunocompromised patients (Taillard et al. 2003). That strains isolated from geographically diverse regions are homogeneous, indicated the dominance of an infectious clone (Zhang et al. 2004). The relatedness between the *M. kansasii* strains coupled with clinical symptoms quite similar to *M. tuberculosis* infections, created significant challenges in understanding the pathogen (Griffith et al. 2007). Thus, a pangenome study was undertaken to examine the genetic repertoire of the strains, recognize the variation between them and determine the factors influencing their lifestyle. In the present study, a comprehensive comparative analysis of 11 *M. kansasii* genomes were carried out, facilitating the determination of core genome, dispensable genome, strain-specific genes, exploration of gene clusters and synteny.

Identical genomic GC content among the *M. kansasii* strains, indicated uniform taxonomy. Despite almost similar genome size, the *M. kansasii* strains demonstrated small variation in the number of genes and protein-coding genes. This was probably due to the events of gene gain

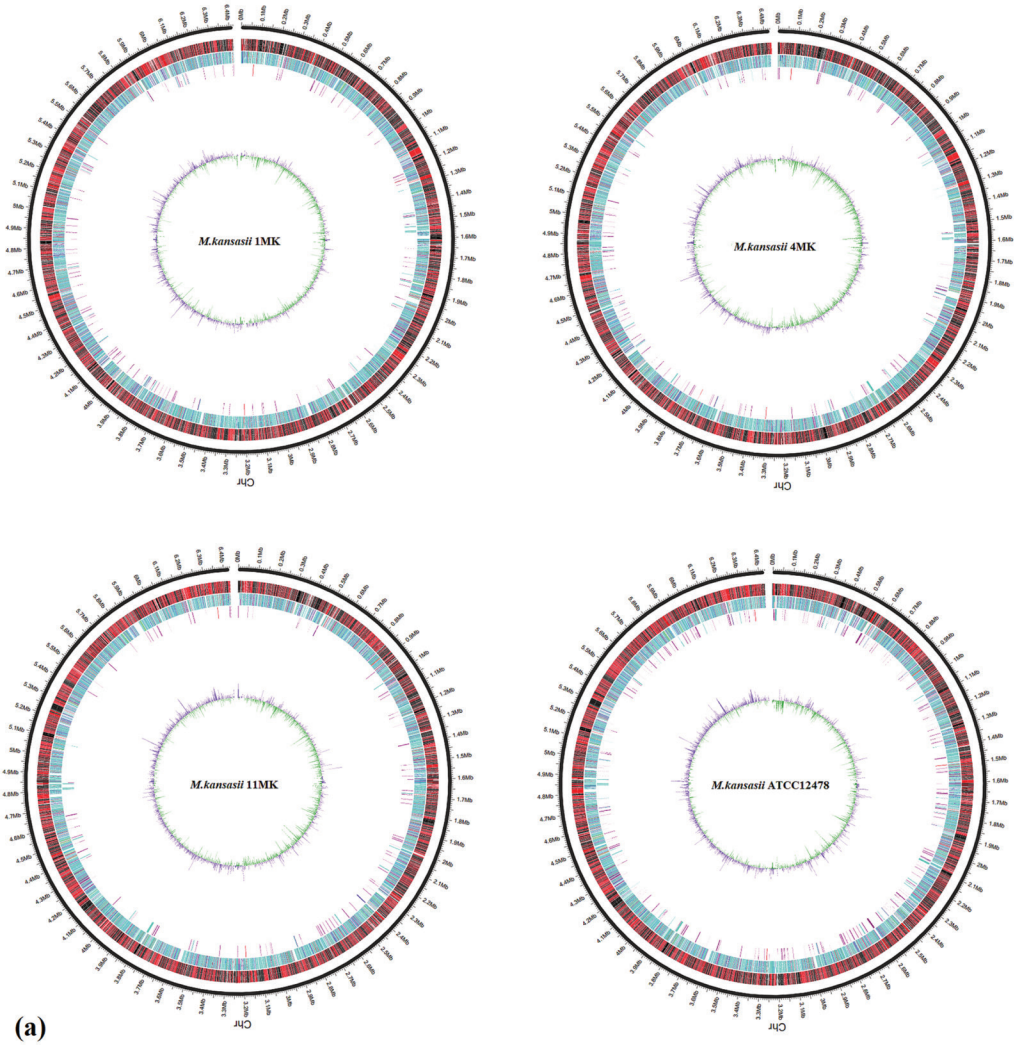

Figure 5. Physical map of the core and strain-specific genes from (a) *M. kansasii* strains 1MK, 4MK, 11MK and ATCC12478; (b) *M. kansasii* strains 6MK, 9MK, FDAARGOS 1534 and FDAARGOS 1614; (c) *M. kansasii* strains 10MK, FDAARGOS 1615, FDAARGOS 1616. The core and strain-specific genes are coloured based on their COG categories. Green, red, sky blue and blue denote the COG categories viz. Information and storage processing, Metabolism, Cellular processing and signaling and Poorly characterized, respectively. Genes not associated with COGs are coloured purple. The tracks from outside to inside indicate genes in the i) forward and reverse strand ii) core genes iii) strain-specific genes and iv) GC deviation from the average. In the latter, the ones in green indicate GC percentage higher than the average while those in blue have lower GC percentage than the average. Genes in forward strand are coloured red while those in reverse black.

or loss. This variation corresponded to the pangenome size of 7852 genes comprising of the core, dispensable and strain-specific genes. Nevertheless, the core genome of 4728 genes included the set of vital genes present across all the 11 strains. Interestingly, the core genome is quite large consisting of 60.2% of the pangenome. This is in line with the observations for bacterial genomes (Tettelin and Medini 2020). These core genes are conserved across the 11 strains of *M. kansasii* and serve as an identity for the same. The relatively small size of the dispensable genome specified low variability on the basis of genome size. The closed pangenome of *M. kansasii* implied that, these strains had lesser possibility to gain newer genetic material from other sources. Therefore, the pangenome embedded a more homogeneous set of gene content. High proportion of hypothetical protein genes with no functional validation, in the core and dispensable genome signified unexplored potential of these strains. These could be earmarked for a combination of experimental and bioinformatic analysis in future, especially in the context of comprehending their role in pathogen physiology, infection, host survival and as prospective therapeutic targets. The number of strain-specific genes corresponded to the genome size. These genes contributed to the genomic differences and

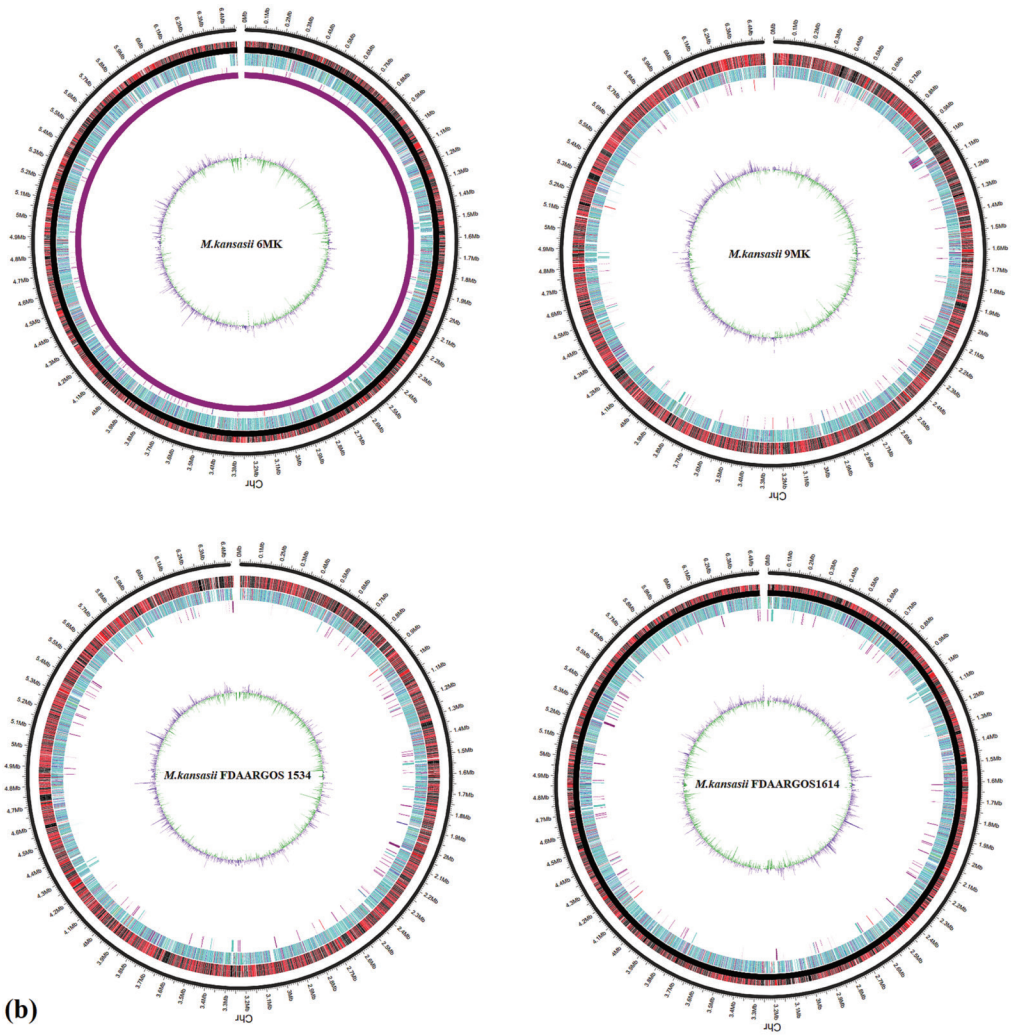

Figure 5. Continued.

complexity among the strains. Again, this variability may be attributed to the diversity and geographical location of the strains.

The presence of ABC transporters and major facilitator superfamily transporters (MFS) in the core genome emphasized their potential in nutrient uptake (Gustaw et al. 2021). Existence of large number of core genome genes associated with the type VII secretion system, WxG 100 family type VII secretion targets, virulence associated toxin C family (VapC), ESX secretion associated proteins, PE and PPE family proteins pointed out to the success of *M. kansasii* as a pathogen (Luo et al. 2021). These pathogenic genes bestowed identity to the human pathogen *M. kansasii.* Multiple lines of evidence have described their role in the manifestation of clinical infections and virulence in humans (Griffin et al. 2012; Houben et al. 2014; Gröschel et al. 2016). High proportion of 2Fe-2S binding proteins within the core genome, in all probability assisted the bacteria to overcome iron deficit at the time of infection by maintaining homeostasis and energy activity (Luo et al. 2021). Substantial incidence of TetR family transcriptional regulators within the core genome, signified the antibiotic resistance potential of *M. kansasii.* Moreover, the existence of numerous cytochrome P450 monooxygenases substantiated their importance in facilitating the organism's metabolism (Sur 2021) for pathogenesis. Interestingly, the occurrence of similar type of pathogenic and virulence genes in the core and dispensable genome reflected panvirulence distribution. The presence of these genes portrayed the pathogenic potential of the bacteria.

Analysis of the COGs underlined the functionalities of *M. kansasii* pangenomic components. The outcome revealed that *M. kansasii* core genes had a high metabolic

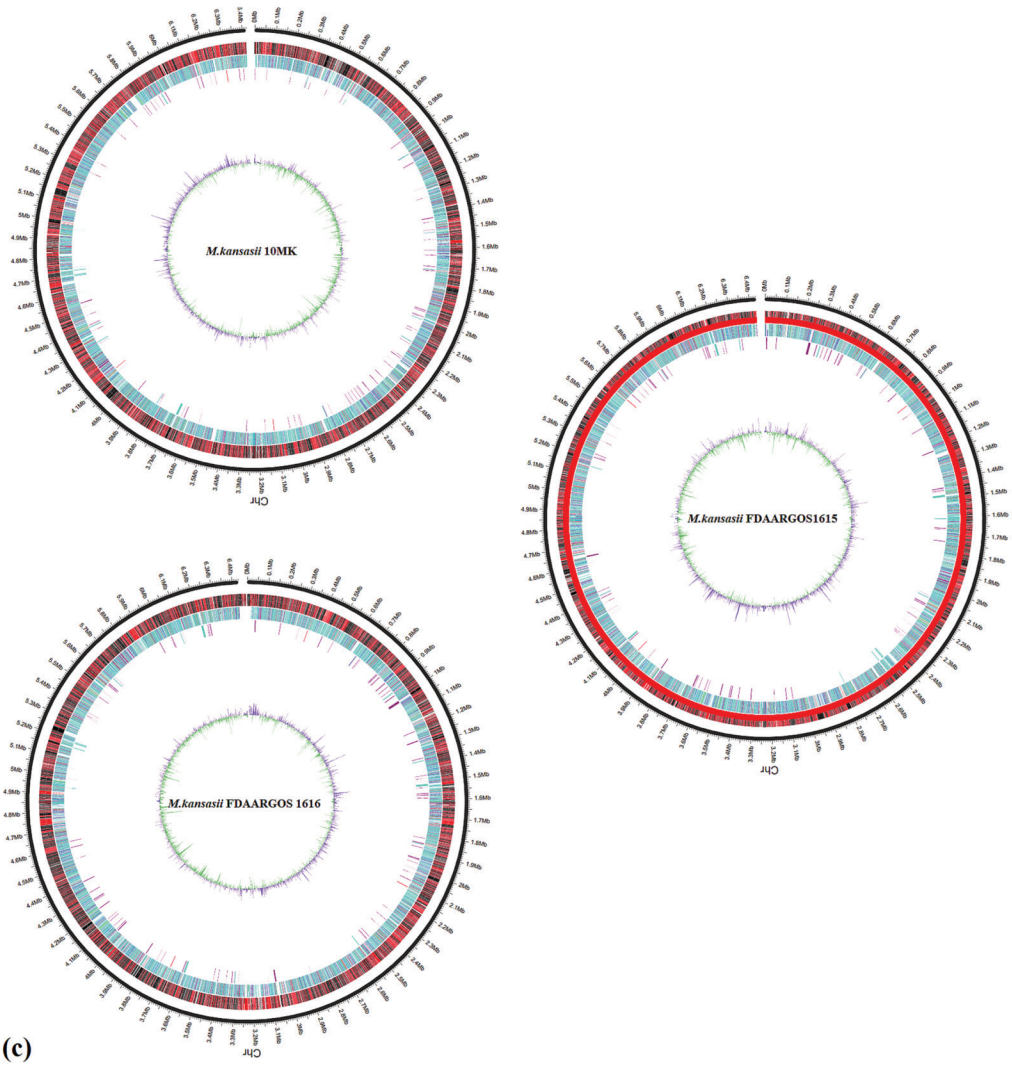

Figure 5. Continued.

capability, owing to high proportion of genes linked to energy production and conversion, amino acid transport and metabolism, nucleotide transport and metabolism, coenzyme transport and metabolism, and secondary metabolite biosynthesis, transport and metabolism. The upregulation of various enzymes secreted by these metabolic activities, are crucial for adaptation and survival of *M. kansasii* within the host. High metabolic capacity is advantageous, in serving as a carbon source for *M. kansasii* to enhance its ability to replicate within the human macrophage ecosystem also (Billig et al. 2017). Moreover, high proportion of genes associated with secondary metabolite biosynthesis indicated the impact of various molecules in regulating metabolism and morphology of the pathogen (Luo et al. 2021). That the dispensable genome housed 19.18% genes within the poorly characterized COG functional category, signified capability to explore them for novel functionalities. Again, presence of 17.47% genes within the dispensable genome related to COGs associated with metabolism emphasized capability of this mycobacteria to overcome challenges of the host ecosystem. It is noteworthy that the strain-specific genes had greater distribution of genes in the general function prediction (R) subcategory followed by those linked with metabolic activities. These are vital for maintaining energy activities. In general, the distribution and proportion of COG functional categories within the core, dispensable and strain-specific genes aided *M. kansasii* in regulating gene expression to adapt itself to the environment.

Synteny analysis of *M. kansasii* genomes pointed out overall conservation and signified the role of many core genes in shaping the connections between genomes.

Despite this, genomic rearrangements between several *M. kansasii* genomes, implied the role of translocation, inversion, deletions etc. in variability amongst them.

Conclusions

In this work, PanExplorer was used to investigate the pangenome of *M. kansasii* and comprehend the lifestyle of the pathogen. Despite the proliferation of large scale sequencing of bacterial genomes, the sequence quality had been a headache for the bioinformaticians and microbiologists at large. This is especially true for the draft genomes. The good thing about PanExplorer is that, it filtered out poor quality draft and incompatible genomes at the start of the analysis. Although it resulted in a smaller dataset, it enhanced the quality and precision of the analysis by restricting sequencing errors commonly associated with partially annotated/assembled and draft genomes. The approach uncovered functionally relevant gene clusters, characteristics and conserved regions that could serve as a basis for further studies. To this end, it was observed that the core genome represented a substantial part of the pangenome and contained numerous genes associated with pathogenicity and virulence. The dispensable genome albeit small still housed many genes related to pathogenesis. These information could be utilized for targeted drug discovery and vaccine development. High and diverse metabolic capacity allowed *M. kansasii* to adapt and survive within the human ecosystem and maintain it's pathogenic lifestyle. Although *M. kansasii* strains displayed relatedness and are known to have shared evolutionary history, genomic rearrangement events and geographical locations shaped the variability amongst them. The findings from this work could aid in controlling *M. kansasii*-mediated infections.

Acknowledgments

The author acknowledges the support of Ramananda College, India for providing infrastructure to carry out the work. The author also thank anonymous reviewers for constructive comments.

References

Banks J, Hunter AM, Campbell IA, Jenkins PA, Smith AP (1983) Pulmonary infection with *Mycobacterium kansasii* in Wales, 1970-9: review of treatment and response. Thorax 38:271-274.

Bernard L, Vincent V, Lortholary O, Raskine L, Vettier C, Colaitis D, Mechali D, Bricaire F, Bouvet E, Bani Sadr F, Lalande V, Perronne C (1999) *Mycobacterium kansasii* septic arthritis: French retrospective study of 5 years and review. Clin Infect Dis 29:1455-1460.

Billig S, Schneefeld M, Huber C, Grassl GA, Eisenreich W, Bange FC (2017) Lactate oxidation facilitates growth of *Mycobacterium tuberculosis* in human macrophages. Sci Rep 7:6484.

Bryant JM, Grogono DM, Rodriguez-Rincon D, Everall I, Brown KP, Moreno P, Verma D, Hill E, Drijkoningen J, Gilligan P, Esther CR, Noone PG, Giddings O, Bell SC, Thomson R, Wainwright CE, Coulter C, Pandey S, Wood ME, Stockwell RE, Ramsay KA, Sherrard LJ, Kidd TJ, Jabbour N, Johnson GR, Knibbs LD, Morawska L, Sly PD, Jones A, Bilton D, Laurenson I, Ruddy M, Bourke S, Bowler IC, Chapman SJ, Clayton A, Cullen M, Daniels T, Dempsey O, Denton M, Desai M, Drew RJ, Edenborough F, Evans J, Folb J, Humphrey H, Isalska B, Jensen-Fangel S, Jönsson B, Jones AM, Katzenstein TL, Lillebaek T, MacGregor G, Mayell S, Millar M, Modha D, Nash EF, O'Brien C, O'Brien D, Ohri C, Pao CS, Peckham D, Perrin F, Perry A, Pressler T, Prtak L, Qvist T, Robb A, Rodgers H, Schaffer K, Shafi N, van Ingen J, Walshaw M, Watson D, West N, Whitehouse J, Haworth CS, Harris SR, Ordway D, Parkhill J, Floto RA (2016) Emergence and spread of a human-transmissible multidrug-resistant nontuberculous mycobacterium. Science 354:751-757.

Dereeper A, Summo M, Meyer DF (2022) PanExplorer: a web-based tool for exploratory analysis and visualization of bacterial pan-genomes. Bioinformatics https://doi.org/10.1093/bioinformatics/btac504.

DeStefano MS, Shoen CM, Cynamon MH (2018) Therapy for *Mycobacterium kansasii* infection: Beyond 2018. Front Microbiol 9:2271.

Dumas E, Christina Boritsch E, Vandenbogaert M, Rodríguez de la Vega RC, Thiberge JM, Caro V, Gaillard JL, Heym B, Girard-Misguich F, Brosch R, Sapriel G (2016) Mycobacterial pan-genome analysis suggests important role of plasmids in the radiation of Type VII secretion systems. Genome Biol Evol 8:387-402.

Fukano H, Terazono T, Hirabayashi A, Yoshida M, Suzuki M, Wada S, Ishii N, Hoshino Y (2021) Human pathogenic *Mycobacterium kansasii* (former subtype I) with zoonotic potential isolated from a diseased indoor pet cat, Japan. Emerg Microbes Infect 10:220-222.

Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA (2009) Circos: an information aesthetic for comparative genomics. Genome Res 19:1639-1645.

Krzywinski M, Birol I, Jones SJ, Marra MA (2012) Hive plots--rational approach to visualizing networks. Brief Bioinform 13:627-44.

Kweon O, Kim SJ, Blom J, Kim SK, Kim BS, Baek DH, Park

SI, Sutherland JB, Cerniglia CE (2015) Comparative functional pan-genome analyses to build connections between genomic dynamics and phenotypic evolution in polycyclic aromatic hydrocarbon metabolism in the genus *Mycobacterium*. BMC Evol Biol 15:21.

Griffin JE, Pandey AK, Gilmore SA, Mizrahi V, McKinney JD, Bertozzi CR, Sassetti CM (2012) Cholesterol catabolism by *Mycobacterium tuberculosis* requires transcriptional and metabolic adaptations. Chem Biol 19:218-27.

Griffith DE, Aksamit T, Brown-Elliott BA, Catanzaro A, Daley C, Gordin F, Holland SM, Horsburgh R, Huitt G, Iademarco MF, Iseman M, Olivier K, Ruoss S, von Reyn CF, Wallace RJ, Jr, Winthrop K, ATS Mycobacterial Diseases Subcommittee, American Thoracic Society, Infectious Disease Society of America (2007) An official ATS/IDSA statement: diagnosis, treatment, and prevention of nontuberculous mycobacterial diseases. Am J Respir Crit Care Med 175:367-416.

Gröschel MI, Sayes F, Simeone R, Majlessi L, Brosch R (2016) ESX secretion systems: mycobacterial evolution to counter host immunity. Nat Rev Microbiol 14:677-691.

Guan Q, Ummels R, Ben-Rached F, Alzahid Y, Amini MS, Adroub SA, van Ingen J, Bitter W, Abdallah AM, Pain A (2020) Comparative genomic and transcriptomic analyses of *Mycobacterium kansasii* subtypes provide new insights into their pathogenicity and taxonomy. Front Cell Infect Microbiol 10:122.

Guo Y, Cao Y, Liu H, Yang J, Wang W, Wang B, Li M, Yu F (2022) Clinical and microbiological characteristics of *Mycobacterium kansasii* pulmonary infections in China. Microbiol Spectr 10:e0147521.

Gustaw K, Koper P, Polak-Berecka M, Rachwał K, Skrzypczak K, Waśko A (2021) Genome and pangenome analysis of *Lactobacillus hilgardii* FLUB-a new strain isolated from mead. Int J Mol Sci 22:3780.

Houben EN, Korotkov KV, Bitter W (2014) Take five - Type VII secretion systems of Mycobacteria. Biochim Biophys Acta 1843:1707-1716.

Jia X, Yang L, Li C, Xu Y, Yang Q, Chen F (2021) Combining comparative genomic analysis with machine learning reveals some promising diagnostic markers to identify five common pathogenic non-tuberculous mycobacteria. Microb Biotechnol 14:1539-1549.

Luo T, Xu P, Zhang Y, Porter JL, Ghanem M, Liu Q, Jiang Y, Li J, Miao Q, Hu B, Howden BP, Fyfe JAM, Globan M, He W, He P, Wang Y, Liu H, Takiff HE, Zhao Y, Chen X, Pan Q, Behr MA, Stinear TP, Gao Q (2021) Population genomics provides insights into the evolution and adaptation to humans of the waterborne pathogen *Mycobacterium kansasii*. Nat Commun 12:2491.

Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MT, Fookes M, Falush D, Keane JA, Parkhill J (2015) Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics 31:3691-3693.

Periwal V, Patowary A, Vellarikkal SK, Gupta A, Singh M, Mittal A, Jeyapaul S, Chauhan RK, Singh AV, Singh PK, Garg P, Katoch VM, Katoch K, Chauhan DS, Sivasubbu S, Scaria V (2015) Comparative whole-genome analysis of clinical isolates reveals characteristic architecture of *Mycobacterium tuberculosis* pangenome. PLoS One 10:e0122979.

Perrin A, Rocha EPC (2021) PanACoTA: a modular tool for massive microbial comparative genomics. NAR Genom Bioinform 3:lqaa106.

Ricketts WM, O'Shaughnessy T, van Ingen J (2014) Human-to-human transmission of *Mycobacterium kansasii* or victims of a shared source? Eur Respir J 44:1085-1087.

Sen A, Sur S, Bothra AK, Benson DR, Normand P, Tisa LS (2008) The implication of lifestyle on codon usage patterns and predicted highly expressed genes for three *Frankia* genomes. Antonie Van Leeuwenhoek 93:335-346.

Sur S, Pal B (2021) Comprehensive review of *Mycobacterium ulcerans* and Buruli ulcer from a bioinformatics perspective – what have we learnt? Acta Biol Szeged 65(2):233-45.

Sur S (2021) Understanding the nature and dynamics of *Mycobacterium ulcerans* cytochrome P450 monooxygenases (CYPs) – a bioinformatics approach. Acta Biol Szeged 65(1):93-103.

Taillard C, Greub G, Weber R, Pfyffer GE, Bodmer T, Zimmerli S, Frei R, Bassetti S, Rohner P, Piffaretti JC, Bernasconi E, Bille J, Telenti A, Prod'hom G (2003) Clinical implications of *Mycobacterium kansasii* species heterogeneity: Swiss national survey. J Clin Microbiol 41:1240-1244.

Tatusov RL, Galperin MY, Natale DA, Koonin EV (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res 28:33-36.

Tettelin H, Medini D (Eds) (2020) The Pangenome. Diversity, Dynamics and Evolution of Genomes. Springer Cham, Switzerland.

Zakham F, Sironen T, Vapalahti O, Kant R (2021) Pan and core genome analysis of 183 *Mycobacterium tuberculosis* strains revealed a high inter-species diversity among the human adapted strains. Antibiotics 10:500.

Zhang Y, Mann LB, Wilson RW, Brown-Elliott BA, Vincent V, Iinuma Y, Wallace RJ Jr (2004) Molecular analysis of *Mycobacterium kansasii* isolates from the United States. J Clin Microbiol 42:119-25.

Zhao Y, Wu J, Yang J, Sun S, Xiao J, Yu J (2012) PGAP: pan-genomes analysis pipeline. Bioinformatics 28:416-418.