Thesis progression

Thesis title: Genomic characterization and pangenome analysis of *Mycobacterium kansasii*

Thesis progression title: Genomic characterization and pangenome analysis of *Mycobacterium kansasii*

Student: Miss Weawwadee Sribunrueang **Student ID:** 685070044-6

Advisor: Asst. Prof. Dr. Auttawit Sirichoat

Date: 18th February 2026

---

1. Introduction

Non-tuberculous mycobacteria (NTM) include all species within the *Mycobacterium* genus, excluding those that cause tuberculosis and leprosy, namely members of the *M. tuberculosis* complex and *M. leprae* or *M. lepromatosis*. These opportunistic bacteria are widespread in the environment, inhabiting water, soil, plants, and animals. NTMs can cause a range of infections throughout the body; pulmonary infections are particularly common (Gu et al., 2023). These infections may manifest lung disease, persistent skin problems, or inflammation of the lymph nodes, and in some cases, can spread to other parts of the body. Currently, over 180 NTM species have been identified, with new species continuously being discovered each year (Parte et al., 2020). This growing diversity has coincided with a global increase in reported NTM-associated infections. This rise is not solely attributable to the expansion of known NTM species, but also reflects improved clinical awareness, a growing population of immunocompromised individuals, and advancements in diagnostic and surveillance technologies (Sood & Parrish, 2017). Despite the potential pathogenicity of many newly described species, fewer than one-third have been consistently associated with clinically significant infections in humans.

Among these, *Mycobacterium kansasii* is one of the most frequently encountered and clinically important NTM species, especially in patients with underlying pulmonary conditions or immunosuppression (Hoefsloot et al., 2013). It was first isolated in 1953 by Buhler and Pollak from respiratory specimens of patients presenting with tuberculosis-like pulmonary symptoms (Buhler & Pollak, 1953). Later named *M. tuberculosis luciflavum* by Middlebrook (1956), and subsequently renamed

M. luciflavum by Manten (1957). The official name *M. kansasii* was proposed by Hauduroy in 1955, reflecting the location of its first isolation in Kansas City, USA (Hauduroy, 1955). *M. kansasii* is a slow-growing opportunistic pathogen known to cause pulmonary infections that resemble tuberculosis, particularly in individuals with compromised immune systems or underlying conditions such as chronic obstructive pulmonary disease (COPD) and various forms of cancer. Primary lung infections caused by *M. kasasii* are generally believed to result from inhalation of aerosolized bacteria present in the environment (Johnston et al., 2017). This species is commonly found in tap water and municipal water systems in urban areas (Vaerewijck et al., 2005).

Initial genetic investigations of *M. kansasii* began in the early 1990s. A foundational study by Ross et al. (1992) used 16S rRNA gene sequencing to identify genetic subgroups. Further studies on clinical and environmental isolates (Taillard et al., 2003) demonstrated significant genetic diversity within the species, leading to the classification of six subtypes (I–VI) based on genotyping approaches (Bakuła et al., 2018). More recent genomic studies have clarified that these subtypes correspond to distinct subspecies: (I) *M. kansasii*, (II) *M. persicum*, (III) *M. pseudokansasii*, (IV) *M. ostraviense*, (V) *M. innocens*, and (VI) *M. attenuatum*. Together with *M. gastri*, these species form the *M. kansasii* complex (MKC) (Jagielski et al., 2019). Notably, the majority of clinical isolates responsible for pulmonary infections across different geographic regions belong to subtype I (*M. kansasii*), underscoring its clinical significance (Chimara et al., 2004).

Whole-genome sequencing (WGS) has become a cornerstone in bacterial genomics. The advent of next-generation sequencing (NGS) technologies has revolutionized the field by dramatically reducing sequencing costs and increasing throughput. Pangenome analysis, a powerful comparative genomics approach, facilitates the identification of both conserved and variable genes across a set of related genomes (Jagielski et al., 2019). A typical pangenome consists of three components: the core genome (shared by all strains), the accessory genome (shared by some but not all strains), and the unique genome (strain-specific genes). These unique genes are often acquired through horizontal gene transfer and mobile genetic elements, and may play important roles in bacterial adaptation and survival (Tettelin et al., 2005).

Previous research has demonstrated the utility of pangenome analysis in characterizing bacterial genomes and exploring intra-species diversity. Many genomic aspects of the MKC that relate to these broad phenotypes are not well elucidated. Genomic analysis of 358 global MKC isolates revealed that recombination, likely via distributive conjugative transfer, drives MKC diversification. Municipal water was identified as a key infection source, with nearly 80% of cases linked to a dominant *M. kansasii* lineage that emerged in the early 1900s. Genes involved in metabolism, secretion, and cell surface remodeling may underlie the pathogen's success and adaptation to human hosts, offering valuable insights into its evolution and epidemiology (Luo et al., 2021). In *M. kansasii* has identified thousands of core genes, along with variable accessory and unique gene sets that differentiate subtypes. For example, subtype I—the most clinically prevalent—contains unique genetic elements such as the espACD operon, which is absent in other subtypes and may contribute to its increased pathogenicity (Tagini et al., 2021). This variation in gene content reflects functional divergence related to host interaction, virulence, and environmental survival. A recent large-scale analysis of gene essentiality in *M. kansasii*, conducted using Himar1 transposon mutagenesis, identified 394 genes required for growth. Of these, 84.8% have essential orthologs in *M. tuberculosis* indicating significant functional overlap. Cross-species genomic comparisons highlighted both conserved and *M. kansasii*-specific essential genes, pointing to possible differences in metabolic or regulatory mechanisms. Additionally, several critical intergenic regions, especially those located upstream of *pe/ppe* gene families were identified, along with essential components of the plasmid pMK1248. These results contribute valuable information for understanding *M. kansasii* pathobiology and may support the development of targeted therapeutics (Keith et al., 2023). Another large-scale study involving 665 MKC strains, sourced from environmental, animal, and human origins, examined the pangenome, mobilome, resistome, virulome, and defense mechanisms of the species. This analysis revealed both shared and unique genomic features, within frequent identification of prophages and diverse defense systems (E. Machado et al., 2024).

Studying the accessory genome through pangenome analysis is particularly valuable for understanding functional differences among strains, including those associated with virulence, environmental adaptation, and antimicrobial resistance.

While pangenome analysis has been widely applied to various bacterial species providing insights into genome evolution, species differentiation, and functional diversity its application to *M. kansasii,* particularly in isolates derived from human infections in Thailand, remains limited. This study aims to address this gap by characterizing the genomic diversity and structure of *M. kansasii* isolates through comprehensive pangenome analysis, thereby offering new insights into their evolution, pathogenic potential, and adaptability.

2 Objectives

2.1 To investigate the genetic diversity among isolates of the *M. kansasii* isolates.

2.2 To characterize the genome structure of *M. kansasii* using pangenome analysis.

3 Hypothesis

We hypothesize that isolates of the *M. kansasii* exhibit significant genetic diversity. Moreover, pangenome analysis will reveal distinct genomic features and structural variations that contribute to this diversity, enhancing our understanding of the evolutionary and adaptive characteristics of these organisms.

4 Conceptual framework

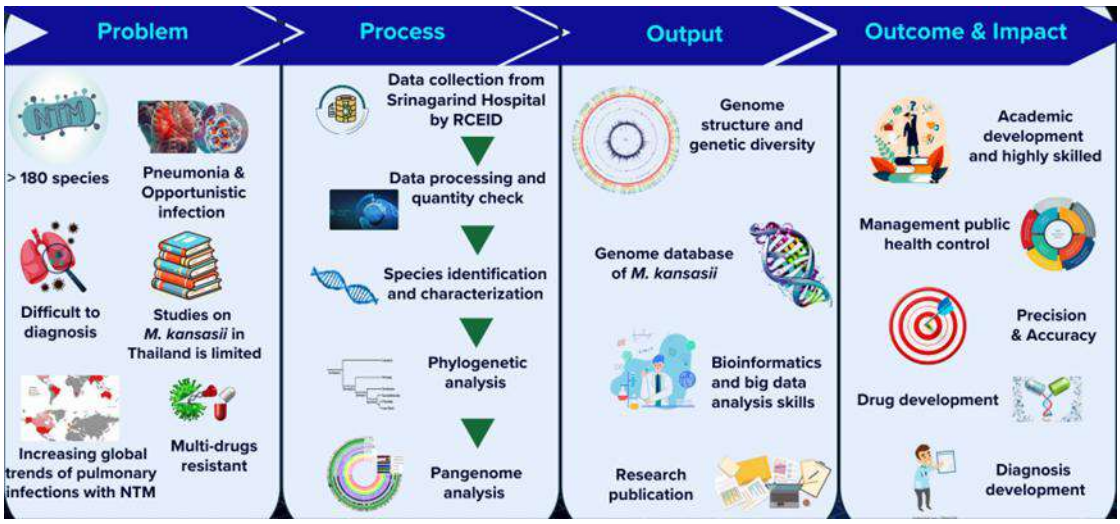

Figure 1 Conceptual framework of the research.

Mycobacterium kansasii, a nontuberculous mycobacterium (NTM), is an opportunistic pathogen causing pulmonary infections. It is difficult to diagnose, genetically diverse, and often resistant to multiple drugs. Global trends show increasing incidence of NTM-related diseases, emphasizing the need for deeper genomic understanding. This study focuses on the genomic characterization and pangenome analysis of *M. kansasii*, which poses diagnostic challenges due to its genetic diversity and increasing drug resistance. By extracting genomic DNA and performing whole genome sequencing (WGS) on clinical and environmental isolates, the study applies pangenome and phylogenetic analyses to investigate genome structure, core and accessory genes, and evolutionary Thailand and global relationships. The resulting outputs include a comprehensive genome database, insights into genetic diversity, and technical skill development in genome analysis. These findings contribute to scientific publications and offer practical impacts such as improved diagnostic tools, drug target identification, public health management strategies, and the advancement of academic and research capacity in microbial genomics (Figure 1).

5 Anticipated outcomes

5.1 The analysis is expected to reveal substantial genetic variability among *M. kansasii* isolates, highlighting their adaptive and dynamic nature.

5.2 Pangenome analysis will provide detailed insights into the genome structure of *M. kansasii*, including the identification of core and accessory genes, as well as genes associated with virulence.

6. Materials and methods

6.1 Experimental design

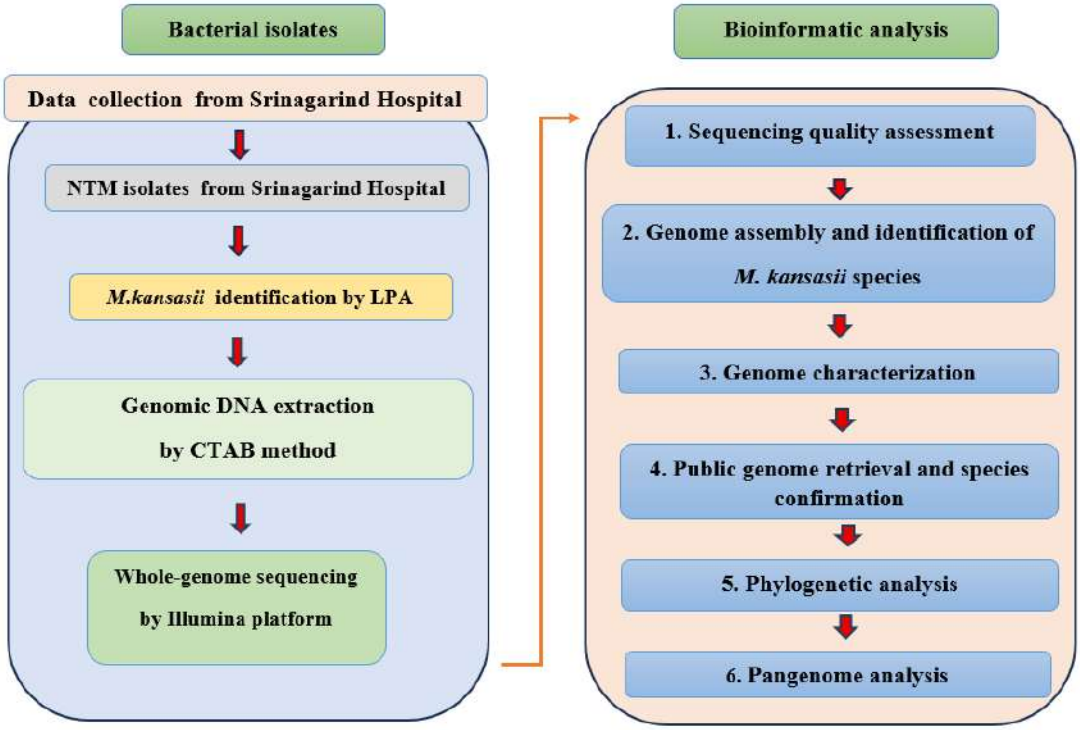

Figure 2 Experimental design of the research

6.2 Bacterial isolates

The *Mycobacterium kansasii* isolates in this study, including mycobacterial cultures, initial isolation, genomic DNA extraction, and whole-genome sequencing, were kindly provided by the staff at the Research and Diagnostic Center for Emerging Infectious Disease (RCEID), Khon Kaen University, Khon Kaen, Thailand.

Six clinical isolates of *M. kansasii* were obtained from the culture collection at the Clinical Microbiology Laboratory, Srinagarind Hospital, Khon Kaen University, Thailand. These isolates originated from patients diagnosed with nontuberculous mycobacterial (NTM) infections and were recovered from various clinical specimens, including sputum, pleural fluid, and lymph node tissues, collected between 2012 and 2016. The initial diagnosis followed the published criteria of the American Thoracic Society and Infectious Diseases Society of America (ATS/IDSA). Molecular identification was initially performed using the line probe assay (LPA; GenoType

Mycobacterium CM VER 2.0, Hain Lifesciences, Nehren, Germany) according to the manufacturer's instructions. For subsequent analyses, the isolates were subcultured on Lowenstein–Jensen medium and incubated at 37 °C more than seven days (Newstead et al., 2025).

Genomic DNA was extracted using the cetyl-trimethyl-ammonium bromide–sodium chloride (CTAB) method described by Larsen et al. (2007). High-quality genomic DNA was subsequently submitted to NovogeneAIT (Singapore) for whole-genome sequencing using the Illumina HiSeq 2500 platform. Library preparation followed the manufacturer's protocol and paired-end sequencing (2 × 150 bp) were performed. The resulting FASTQ files will be processed for downstream bioinformatic analyses.

6.3 Bioinformatic analysis

6.3.1 Sequencing quality assessment

The quality of the raw sequencing reads will be evaluated using FastQC v0.12.1 (Andrews, 2010). Reads shorter than 75 bp and those containing adapter sequences will be removed using Trimmomatic v0.39 (Bolger et al., 2014) with the following parameters: LEADING:3, TRAILING:3, SLIDINGWINDOW:4:15, and MINLEN:75. High-quality, filtered reads will be retained for all subsequent analyses.

6.3.2 Genome assembly and species identification of *M. kansasii*

High-quality reads will be *de novo* assembled using Unicycler v0.5.1 (Wick et al., 2017). Assembly quality will be improved using RagTag v2.1.0 (Alonge et al., 2022) and polished with Pilon v1.24 (Walker et al., 2014). Assembly metrics will be assessed using QUAST v5.3.0 (Gurevich et al., 2013), and genome completeness and contamination will be evaluated using CheckM v1.2.3 (Parks et al., 2015). Only genomes with ≥95% completeness and ≤5% contamination will be included in subsequent analyses.

Species identification will be performed using NTM Profiler v0.6.1 (Phelan et al., 2019), the Type Strains Genome Sever (TYGS) (Meier-Kolthoff & Göker, 2019) (available on https://tygs.dsmz.de/), Pathogenwatch (available on https://pathogen.watch/), and GTDB-Tk (Chaumeil et al., 2020). Pairwise average

nucleotide identity (ANI) will be calculated using FastANI (Jain et al., 2018) for accurate taxonomic classification. ANI-based clustering and heatmap visualization will be generated using ANIclustermap v1.4.0 (Shimoyama, 2022).

6.3.3 Genome characterization

The assembled genomes will be submitted to the Bacterial and Viral Bioinformatics Resource Center (BV-BRC) for a comprehensive genome analysis pipeline (Olson et al., 2023). Genome annotation will be conducted using the RASTtk pipeline to assess protein subsystem distribution. Virulence factors (VFs) will be screened using ABRicate v1.0.1 (Seemann, 2020) with default parameters and the Virulence Factor Database (VFDB) (Chen et al., 2016).

6.3.4 Public genome retrieval and species confirmation

Publicly available genome assemblies will be retrieved from the National Center for Biotechnology Information (NCBI) database (https://www.ncbi.nlm.nih.gov/datasets/genome/) (accessed November 2025). All retrieved genomes will be evaluated using QUAST and CheckM. Species confirmation will be performed using NTM-Profiler, TYGS, Pathogenwatch, GTDB-Tk, and FastANI as described above. Genomes confirmed as *M. kansasii* genomes will be combined with the six clinical isolates from this study for subsequent analyses.

6.3.5 Core-SNP phylogeny

To complement the gene-based analysis, core-genome alignment and single-nucleotide polymorphism (SNP) calling will be performed using Parsnp v1.2 within the Harvest Suite (Treangen et al., 2014). *De novo* assemblies will be aligned to the *M. kansasii* reference genome ATCC 12478 (accession no. CP006835.1). Recombination detection and refinement of the core genome will be conducted using the PhiPack backend with default parameters to ensure high-quality variant calling. The final alignment will be converted to FASTA format using HarvestTools. A maximum-likelihood phylogenetic tree will be constructed using IQ-TREE v2.4.0 (Nguyen et al., 2015) with the optimal nucleotide substitution model and 1,000 bootstrap replicates. Phylogenetic visualization will be performed using iTOL v7 (Letunic & Bork, 2007).

6.3.6 Pangenome analysis and functional annotation

All *M. kansasii* genomes will be re-annotated using Prokka v1.14.6 (Seemann, 2014). Pangenome analysis will be conducted using Roary v3.12.0 (Page et al., 2015) with default parameters. Roary outputs, including the gene presence–absence matrix and summary statistics, will be generated. Genes will be categorized into core (present in all genomes), soft-core (present in most but not all genomes), accessory (present in several genomes), or unique (strain-specific) categories. Visualization of pangenome results will be performed using the roary_plots.py script and the ggplot2 package in R program.

Functional annotation of core, accessory, and unique gene sets will be performed using the Clusters of Orthologous Groups (COGs) classification system via the Pan-Explorer platform (Dereeper et al., 2022), facilitating the comparative analysis of metabolic pathways and functional roles across isolates.

7. Results

Characteristics of the studied isolates

Six *M. kansasii* clinical isolates were identified using a comprehensive bioinformatics framework comprising NTM-Profiler, TYGS, Pathogenwatch, GTDB-Tk, and FastANI (Table 1). Most isolates (83.33%, 5/6) were obtained from female patients, with a mean age of 63.5 years (range, 49–88 years). The isolates originated from diverse clinical specimens, including sputum, pleural fluid, and lymph node tissues (Table 1).

Gene prediction and functional annotation

The six *M. kansasii* isolates exhibited a mean genome size of 6.49 Mbp (Table 1). The number of predicted coding sequences (CDSs) ranged from 5,997 to 6,167, with an average of 6,104.5 CDSs per genome. All genomes displayed a GC content of approximately 66.15%. Each isolate contained three rRNA genes, while the number of tRNA genes ranged from 49 to 51.

Sequence types (STs) were assigned using the PubMLST database. Among the six *M. kansasii* isolates analyzed, 5 were successfully classified into previously defined STs, while one isolate could not be assigned to any existing profile, suggesting the presence of potentially novel genetic diversity within the local population. ST38 was the predominant lineage, accounting for five of the six isolates (83.33%), whereas one isolate (16.67%) remained unassigned.

Functional categorization revealed that the majority of genes were associated with metabolic processes, followed by protein processing, stress response, energy production, DNA-related functions, cellular processes, and RNA processing. This functional distribution highlights the metabolic versatility and adaptive capacity of *M. kansasii* in diverse host and environmental niches (Fig. 3).

Virulence factors profiling

Virulence gene screening using VFDB revealed a largely conserved repertoire of genes involved in key pathogenic mechanisms. These included genes associated with cell wall biosynthesis and lipid metabolism (*aftB, aftD, embC, pks1, pks15, ppsC–E*), ESX secretion systems (*esxA, esxB, esxH, esxV, eccA3, eccB3, eccC3, eccD3, eccE3*), stress response and persistence (*ahpC, relA, ideR, sodA*), and intracellular survival (*devR, devS*). Genes implicated in host cell adhesion (*hbhA, fbpA–D*), host membrane damage (*plcA, plcB, plcC*), iron acquisition (*mbtB, mbtC, mbtI*), vitamin biosynthesis (*panC, panD*), and PE family proteins (*PE5, PE19*) were also consistently detected across the isolates. Minor variations were observed within the ESX-associated gene cluster, with *espG3* absent in five isolates and *esxG* absent in one isolate (Fig. 4).

Table.1 General genome features of the 6 *Mycobacterium kansasii* isolates.

Feature/protein coding genes	**Isolate**					
	MKA80499	MKA80686	MKA80709	MKA82243	MKA82252	MKA82253
Source	Sputum	Sputum	Sputum	Pleural fluid	Lymph node (TISSUE)	Sputum
Collection date (d/m/y)	20/2/2013	11/3/2013	12/2/2013	17/9/2016	19/9/2016	19/9/2016
Genome size (bp)	6,505,392	6,425,212	6,416,927	6,538,639	6,533,506	6,536,265
GC content(%)	66.09	66.2	66.2	66.14	66.14	66.13
No. of contigs	24	28	28	26	25	27
Contig N50 (bp)	6,476,335	6,384,716	6,374,337	6,513,204	6,508,363	6,509,982
Contig L50	1	1	1	1	1	1
No. of CDSs	6,159	5,997	6,003	6,142	6,167	6,159
No. of subsystems	310	309	309	309	310	309
tRNAs	50	50	51	49	49	49
rRNA	3	3	3	3	3	3

Notes: d/m/y, date/month/year; bp, base pairs; CDSs, coding DNA sequences.

The N50 length, which is defined as the shortest sequence length at 50 % of the genome.

The L50 count, which is defined as the smallest number of contigs whose length sum produces N50.

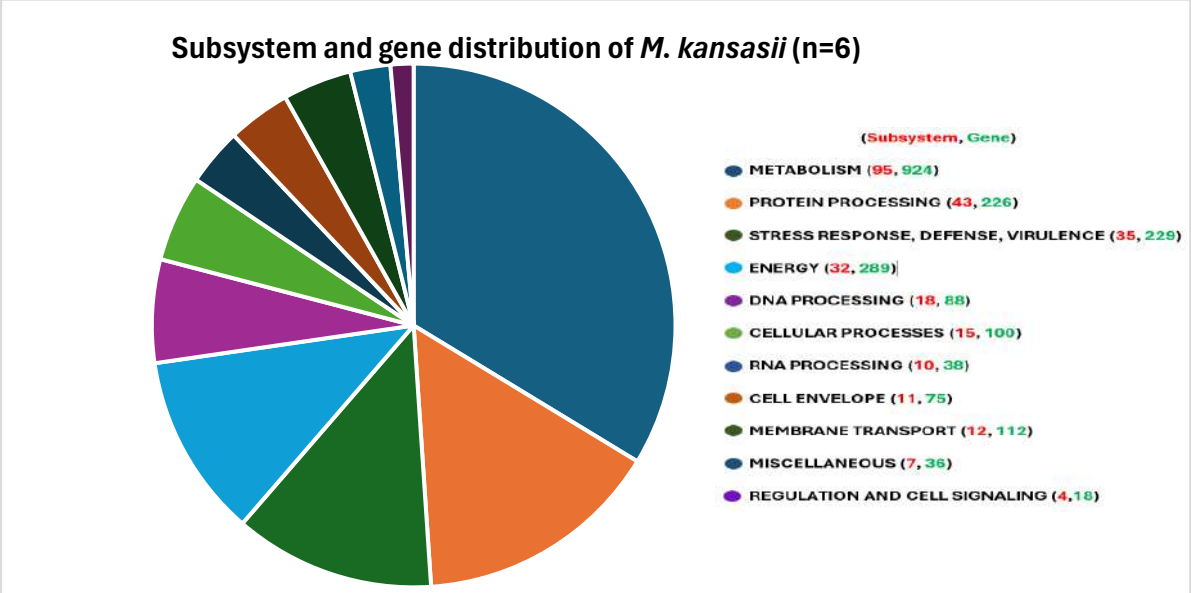

Figure 3. Functional subsystem annotation of six *Mycobacterium kansasii* isolates. Subsystem categories and gene distributions were derived from genome annotations generated using the RAST toolkit through the Bacterial and Viral Bioinformatics Resource Center platform.

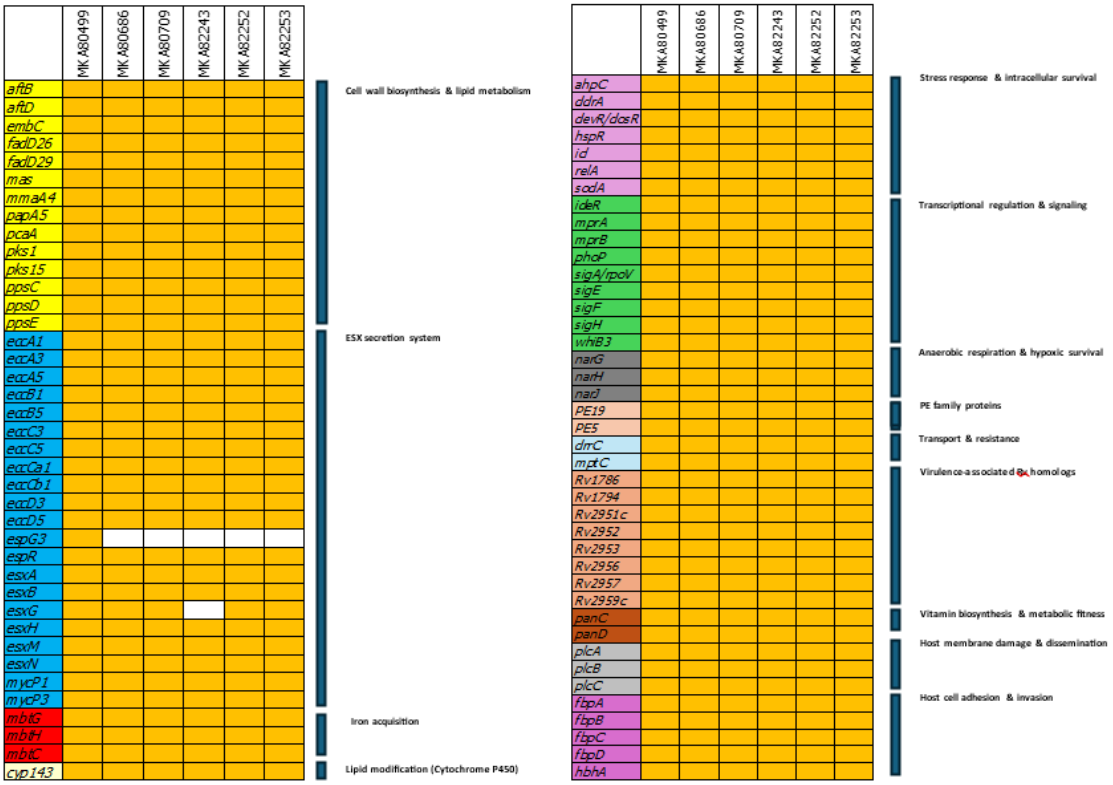

Figure 4. Distribution of virulence factors across 6 *Mycobacterium kansasii* clinical isolates. Gene presence determined using the Virulence Factor Database via ABRicate. Yellow indicates gene pre sence; white indicates gene absence.

8. Conclusion

In this study, six clinical *M. kansasii* isolates from Thailand were comprehensively characterized using multiple bioinformatics classification tools, confirming their taxonomic identity and genomic consistency. The isolates exhibited highly conserved genomic features, including genome size, GC content, and core gene composition, reflecting genomic stability within the species. Multilocus sequence typing revealed that ST38 was the predominant lineage, while the presence of unassigned sequence types suggests previously unrecognized genetic diversity within the local population. Functional annotation demonstrated that genes related to metabolism and cellular processes were highly represented, highlighting the metabolic versatility and adaptive potential of *M. kansasii*. Virulence profiling further showed a largely conserved repertoire of genes involved in cell wall biosynthesis, secretion systems, stress response, and intracellular survival, with only minor variations observed among ESX-associated genes. Overall, these findings provide valuable genomic insights into Thai clinical *M. kansasii* isolates and contribute to a better understanding of their population structure, functional capacity, and pathogenic potential.

9 Research plan

Table 2 Research plan of the research.

Activities	2025		2026				2027	
	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2
1. Course work								
2. Literature reviews and planning								
3. Proposal examination								
4. Data collection of *M. kansasii*								
5. Sequencing quality assessment								
6. Genome assembly and identification of *M. kansasii* species								
7. Genome characterization								
8. Public genome retrieval and species confirmation								
9. Phylogenetic analysis								
10. Pangenome analysis								
11. Manuscript and preparation and submission								
12. Thesis preparation and examination								

Blue color ( ) indicates finished work, orange ( ) indicates ongoing work, and grey color ( ) indicates future work. This semester, I registered for 15 credits for the thesis course.

REFERENCES

Alonge, M., Lebeigle, L., Kirsche, M., Jenike, K., Ou, S., Aganezov, S., Wang, X., Lippman, Z. B., Schatz, M. C., & Soyk, S. (2022). Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. *Genome Biology*, *23*(1), 258. https://doi.org/10.1186/s13059-022-02823-7

Andrews, S. (2010). *FastQC: A Quality Control Tool for High Throughput Sequence Data*.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics (Oxford, England)*, *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Buhler, V. B., & Pollak, A. (1953). Human infection with atypical acid-fast organisms; report of two cases with pathologic findings. *American Journal of Clinical Pathology*, *23*(4), 363–374. https://doi.org/10.1093/ajcp/23.4.363

Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P., & Parks, D. H. (2020). GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics*, *36*(6), 1925–1927. https://doi.org/10.1093/bioinformatics/btz848

Chen, L., Zheng, D., Liu, B., Yang, J., & Jin, Q. (2016). VFDB 2016: hierarchical and refined dataset for big data analysis--10 years on. *Nucleic Acids Research*, *44*(D1), D694-7. https://doi.org/10.1093/nar/gkv1239

Chimara, E., Giampaglia, C. M. S., Martins, M. C., Telles, M. A. da S., Ueki, S. Y. M., & Ferrazoli, L. (2004). Molecular characterization of Mycobacterium kansasii isolates in the State of São Paulo between 1995-1998. *Memorias Do Instituto Oswaldo Cruz*, *99*(7), 739–743. https://doi.org/10.1590/s0074-02762004000700013

Dereeper, A., Summo, M., & Meyer, D. F. (2022). PanExplorer: a web-based tool for exploratory analysis and visualization of bacterial pan-genomes. *Bioinformatics*, *38*(18), 4412–4414. https://doi.org/10.1093/bioinformatics/btac504

Gu, Y., Nie, W., Huang, H., & Yu, X. (2023). Non-tuberculous mycobacterial

disease: progress and advances in the development of novel candidate and repurposed drugs. *Frontiers in Cellular and Infection Microbiology*, *13*, 1243457. https://doi.org/10.3389/fcimb.2023.1243457

Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics (Oxford, England)*, *29*(8), 1072–1075. https://doi.org/10.1093/bioinformatics/btt086

Hauduroy, P. (1955). *Derniers aspects du monde des mycobacteries*.

Hoefsloot, W., van Ingen, J., Andrejak, C., Ängeby, K., Bauriaud, R., Bemer, P., Beylis, N., Boeree, M. J., Cacho, J., Chihota, V., Chimara, E., Churchyard, G., Cias, R., Daza, R., Daley, C. L., Dekhuijzen, P. N. R., Domingo, D., Drobniewski, F., Esteban, J., … Wagner, D. (2013). The geographic diversity of nontuberculous mycobacteria isolated from pulmonary samples: an NTM-NET collaborative study. *European Respiratory Journal*, *42*(6), 1604–1613. https://doi.org/10.1183/09031936.00149212

Jagielski, T., Borówka, P., Bakuła, Z., Lach, J., Marciniak, B., Brzostek, A., Dziadek, J., Dziurzyński, M., Pennings, L., van Ingen, J., Žolnir-Dovč, M., & Strapagiel, D. (2019). Genomic Insights Into the Mycobacterium kansasii Complex: An Update. *Frontiers in Microbiology*, *10*, 2918. https://doi.org/10.3389/fmicb.2019.02918

Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T., & Aluru, S. (2018). High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nature Communications*, *9*(1), 5114. https://doi.org/10.1038/s41467-018-07641-9

Johnston, J. C., Chiang, L., & Elwood, K. (2017). Mycobacterium kansasii. *Microbiology Spectrum*, *5*(1). https://doi.org/10.1128/microbiolspec.TNMI7-0011-2016

Keith, L., Niklas, J., & N., Q. L. E. (2023). Comprehensive essentiality analysis of the Mycobacterium kansasii genome by saturation transposon mutagenesis and deep sequencing. *MBio*, *14*(4), e00573-23. https://doi.org/10.1128/mbio.00573-23

Letunic, I., & Bork, P. (2007). Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics (Oxford, England)*, *23*(1), 127–128. https://doi.org/10.1093/bioinformatics/btl529

Luo, T., Xu, P., Zhang, Y., Porter, J. L., Ghanem, M., Liu, Q., Jiang, Y., Li, J., Miao, Q., Hu, B., Howden, B. P., Fyfe, J. A. M., Globan, M., He, W., He, P., Wang, Y., Liu, H., Takiff, H. E., Zhao, Y., … Gao, Q. (2021). Population genomics provides insights into the evolution and adaptation to humans of the waterborne pathogen Mycobacterium kansasii. *Nature Communications*, *12*(1), 2491. https://doi.org/10.1038/s41467-021-22760-6

Machado, E., Vasconcellos, S., Gomes, L., Catanho, M., Ramos, J., de Carvalho, L., Goldenberg, T., Redner, P., Caldas, P., Campos, C., Dalcolmo, M., Lourenço, M. C., Lasunskaia, E., Mussi, V., Spinassé, L., Vinhas, S., Rigouts, L., Cogneau, S., de Rijk, P., … Suffys, P. (2024). Phylogenomic and genomic analysis reveals unique and shared genetic signatures of Mycobacterium kansasii complex species. *Microbial Genomics*, *10*(7). https://doi.org/10.1099/mgen.0.001266

Manten, A. (1957). Antimicrobial susceptibility and some other properties of photochromogenic Mycobacteria associated with pulmonary disease. *Antonie van Leeuwenhoek*, *23*(3–4), 357–363. https://doi.org/10.1007/BF02545888

Meier-Kolthoff, J. P., & Göker, M. (2019). TYGS is an automated high-throughput platform for state-of-the-art genome-based taxonomy. *Nature Communications*, *10*(1), 2182. https://doi.org/10.1038/s41467-019-10210-3

Nguyen, L.-T., Schmidt, H. A., von Haeseler, A., & Minh, B. Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*, *32*(1), 268–274. https://doi.org/10.1093/molbev/msu300

Olson, R. D., Assaf, R., Brettin, T., Conrad, N., Cucinell, C., Davis, J. J., Dempsey, D. M., Dickerman, A., Dietrich, E. M., Kenyon, R. W., Kuscuoglu, M., Lefkowitz, E. J., Lu, J., Machi, D., Macken, C., Mao, C., Niewiadomska, A., Nguyen, M., Olsen, G. J., … Stevens, R. L. (2023). Introducing the Bacterial and Viral Bioinformatics Resource Center (BV-BRC): a resource combining PATRIC, IRD and ViPR. *Nucleic Acids Research*, *51*(D1), D678–D689. https://doi.org/10.1093/nar/gkac1003

Page, A. J., Cummins, C. A., Hunt, M., Wong, V. K., Reuter, S., Holden, M. T. G., Fookes, M., Falush, D., Keane, J. A., & Parkhill, J. (2015). Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics (Oxford, England)*,

31(22), 3691–3693. https://doi.org/10.1093/bioinformatics/btv421

Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., & Tyson, G. W. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research*, *25*(7), 1043–1055. https://doi.org/10.1101/gr.186072.114

Parte, A. C., Sardà Carbasse, J., Meier-Kolthoff, J. P., Reimer, L. C., & Göker, M. (2020). List of Prokaryotic names with Standing in Nomenclature (LPSN) moves to the DSMZ. *International Journal of Systematic and Evolutionary Microbiology*, *70*(11), 5607–5612. https://doi.org/10.1099/ijsem.0.004332

Phelan, J., O'Sullivan, D., Machado, D., Ramos, J., Oppong, Y., Campino, S., O'Grady, J., Mcnerney, R., Hibberd, M., Viveiros, M., Huggett, J., & Clark, T. (2019). Integrating informatics tools and portable sequencing technology for rapid detection of resistance to anti-tuberculous drugs. *Genome Medicine*, *11*. https://doi.org/10.1186/s13073-019-0650-x

Ross, B. C., Jackson, K., Yang, M., Sievers, A., & Dwyer, B. (1992). Identification of a genetically distinct subspecies of Mycobacterium kansasii. *Journal of Clinical Microbiology*, *30*(11), 2930–2933. https://doi.org/10.1128/jcm.30.11.2930-2933.1992

Runyon, E. H. (1959). Anonymous mycobacteria in pulmonary disease. *The Medical Clinics of North America*, *43*(1), 273–290. https://doi.org/10.1016/s0025-7125(16)34193-1

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, *30*(14), 2068–2069. https://doi.org/10.1093/bioinformatics/btu153

Seemann, T. (2020). *ABRicate: Mass screening of contigs for antimicrobial and virulence genes*. https://github.com/tseemann/abricate

Shimoyama, Y. (2022). *ANIclustermap: A tool for drawing ANI clustermap between all-vs-all microbial genomes*. GitHub (Moshi4). https://github.com/moshi4/ANIclustermap

Sood, G., & Parrish, N. (2017). Outbreaks of nontuberculous mycobacteria. *Current Opinion in Infectious Diseases*, *30*(4), 404–409. https://doi.org/10.1097/QCO.0000000000000386

Tagini, F., Pillonel, T., Bertelli, C., Jaton, K., & Greub, G. (2021). Pathogenic

Determinants of the Mycobacterium kansasii Complex: An Unsuspected Role for Distributive Conjugal Transfer. *Microorganisms*, *9*(2). https://doi.org/10.3390/microorganisms9020348

Tettelin, H., Masignani, V., Cieslewicz, M. J., Donati, C., Medini, D., Ward, N. L., Angiuoli, S. V, Crabtree, J., Jones, A. L., Durkin, A. S., Deboy, R. T., Davidsen, T. M., Mora, M., Scarselli, M., Margarit y Ros, I., Peterson, J. D., Hauser, C. R., Sundaram, J. P., Nelson, W. C., … Fraser, C. M. (2005). Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". *Proceedings of the National Academy of Sciences of the United States of America*, *102*(39), 13950–13955. https://doi.org/10.1073/pnas.0506758102

Treangen, T. J., Ondov, B. D., Koren, S., & Phillippy, A. M. (2014). The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. *Genome Biology*, *15*(11), 524. https://doi.org/10.1186/s13059-014-0524-x

Vaerewijck, M. J. M., Huys, G., Palomino, J. C., Swings, J., & Portaels, F. (2005). Mycobacteria in drinking water distribution systems: ecology and significance for human health. *FEMS Microbiology Reviews*, *29*(5), 911–934. https://doi.org/10.1016/j.femsre.2005.02.001

Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C. A., Zeng, Q., Wortman, J., Young, S. K., & Earl, A. M. (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PloS One*, *9*(11), e112963. https://doi.org/10.1371/journal.pone.0112963

Wick, R. R., Judd, L. M., Gorrie, C. L., & Holt, K. E. (2017). Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Computational Biology*, *13*(6), e1005595. https://doi.org/10.1371/journal.pcbi.1005595