Unravelling genomic signatures of *Mycobacterium bovis* host specificity through a pangenomic framework

Noah Legall
June 2021

Major Professor: Liliana Salvador

Committee Member: Douda Bensasson
Committee Member: Liang Liu
Committee Member: Justin Bahl

Important information for the Committee:

Instructions for written exam evaluation:
Each committee member has two weeks to review and submit a grade of pass or fail via email to the student and the Graduate Program Administrator (iobgradadmin@UGA.EDU). To pass the written portion and go on to the oral portion, the student must receive no more than one dissenting (failing grade) vote. If a committee member does not provide a grade two weeks after submission of the written exam, the grade will be marked as a pass for that committee member. The written exam takes the form of an NIH grant proposal; if you are not familiar with the IOB written exam requirements, see Appendix B in the most recent graduate student handbook (https://iob.uga.edu/graduate-program/graduate-handbooks/).

Instructions for oral exam: The oral exam will last at least two hours, but not longer than three hours. The student will prepare a presentation of no more than 20 slides that are intended to serve as a framework of the discussion of the proposed research. The student's presentation should last for approximately 20-25 minutes without interruptions, followed by questions from the advisory committee and other faculty present. Questions during the exam will test both general and specific knowledge related to the student's proposed research as described in their presentation and written proposal. A member of the student's committee, other than the advisor, will serve as chair of the exam. The advisor is not allowed to answer questions for the student, and will not participate in the discussion unless granted permission by the exam chair.

Specific Aims

Bovine tuberculosis (bTB) is an agriculturally and economically devastating disease caused by *Mycobacterium bovis*[1]. Studies have shown that *M. bovis* has a wide range of host species with bidirectional transmission between livestock and wildlife, which compromises the success of bTB surveillance and control programs[1–3]. Next Generation Sequencing (NGS) enhances our ability to characterize *M. bovis* transmission by using single nucleotide polymorphisms (SNPs) as input for phylogenetic inference[4,5]. The subsequent phylogenetic clustering of samples is useful for testing hypotheses regarding transmission between geographic locations or host species[6,7]. *M. bovis* genomic signatures other than SNPs, such as sites of positive selection or homologous recombination, have been used to conclude that infection-associated genes are general targets for genomic variations[8–10]. However, researchers still do not fully understand how *M. bovis* genomic variations are associated with *M. bovis* host specificity, or what types of genes are targets when circulating within a host population. Lacking this knowledge will limit the ability of future researchers to exploit features of *M. bovis* evolution as a strategy to mitigate transmission. Thus, there is a *critical need* for tools and methods designed to discover novel genomic variations that potentially influence characteristics of *M. bovis* transmission.

My *long-term goal* is to characterize the relationship between *M. bovis* genomic signatures and transmission within and between host-species populations. The *overall objective* of this proposal, which is one step towards my long-term goal, is to develop a framework for *M. bovis* sequence analysis that will depict the evolutionary mechanisms that are associated with specific host species. To construct the framework, my *central approach* is to infer the pangenome[11–13] (the entire set of conserved or variably present genes) for *M. bovis* datasets through a developed bioinformatics pipeline (*bovpan)*, and use the resulting output for analysis of host-species specific genomic variations. The *rationale* for developing *bovpan* is that incorporating other genomic signatures, rather than SNPs alone, will allow researchers to better characterize associations with host specificity. To attain the overall objectives, the following three specific aims will be pursued:

Aim 1. Develop *bovpan*, a bioinformatics pipeline to analyze *M. bovis* genomic data: Current *M. bovis* genomic data bioinformatic pipelines do not decipher the full extent of genomic variations that are present across *M. bovis* lineages[3,7,14–16]. I will develop a scalable, modular, and reproducible tool that will determine the *M. bovis* pangenome, alongside SNPs. This approach will allow researchers to associate variation data with host specificity.

Aim 2. Determine *M. bovis* host-specific genomic signatures: Previous research based on SNP analysis alone indicates that genomic similarity of *M. bovis* is closely tied to spatial proximity[6]. I will extend this work and compute gene presence/absence clustering, sites of homologous recombination and structural variations, and genes under positive selection to determine genomic signatures that can differentiate host and region associated *M. bovis* lineages.

Aim 3. Unravel *M. bovis* gene interactions that differentiate host-specific lineages: Preliminary data shows that UK *M. bovis* sample accessory genomes from badger and cattle cluster separately. I will analyze SNPs, gene presence/absence, and genetic interactions differentially enriched between badgers and cattle to indicate which genes potentially contribute to *M. bovis* host-pathogen interactions.

After completion of the proposed research, the *expected outcomes* are to use the developed framework to uncover *M. bovis* specific genomic data that can be used to investigate characteristics of *M. bovis* biology, such as genomic signatures and specific genes that enable specificity to a particular host. These outcomes will have a *positive impact* by providing researchers with a framework to unravel *M. bovis* genome variations, while also highlighting the genes and genomic signatures within *M. bovis* that promote host specificity.

Significance

Bovine tuberculosis (bTB) is an agriculturally and economically devastating livestock disease that infects 50 million cattle worldwide and costs farmers $3 billion annually[1]. The disease is caused by the bacterial zoonosis *Mycobacterium bovis,* a gram-positive rod-shaped bacterium that can transmit amongst a wide range of mammalian host species[17]. Close spatial proximity between wild and domestic animals can occur through direct contact (infected individuals and carcasses), or indirect contact (contaminated soil or water resources), which contributes to transmission of *M. bovis* at the wildlife-livestock interface[1]. *M. bovis* is found worldwide with certain regions (United States of America (USA)[7], New Zealand (NZ)[15], and the United Kingdom, (UK))[3] containing wildlife species that transitioned from novel spillover hosts to reservoirs of infection, populations that can maintain disease and also transmit disease to other species[18–20]. Wildlife reservoirs of infection populations can flourish in close contact to livestock species, and consistent contact between the reservoir and livestock can lead to frequent re-introductions of *M. bovis* back into livestock populations[1]. However, only a few species have shown to transition from novel spillover host to reservoir of infection, suggesting that there exist certain factors that limit the ability of *M. bovis* to transmit in specific hosts[21]. Without proper understanding of how these factors contribute to *M. bovis* adaptation to particular host species, researchers will be limited in their ability to exploit *M. bovis* biology to reduce transmission. Therefore, there is a critical need to use new data sources to help characterize transmission within or between species populations.

Next generation sequencing (NGS) holds promise as a way to elucidate transmission dynamics at the wildlife-livestock interface[16]. NGS allows researchers to observe genomic variations that occur through mutational processes during an outbreak. The clustering of pathogens based on their variations can assist in understanding transmission dynamics, and ultimately answer a fundamental question in molecular epidemiology: *who acquired infection from whom* [22]? Characterizing *M. bovis* through the genotyping of collected samples is primarily achieved through Whole-Genome Sequencing (WGS)[23], which proves useful to identify sources of disease outbreaks, tracing chains of transmission, and determining cross-species pathogen migration rates. Most research focusing on the molecular epidemiology of *M. bovis* uses single nucleotide polymorphisms (SNPs)[16] as a good way to determine transmission dynamics, but other informative genomic variations such as homologous recombination (HR), gene presence/absence, or structural variations (SVs) are underutilized in these studies. By continuing this reliance on SNPs to determine aspects about *M. bovis* transmission, researchers might not be able to fully uncover underlying causes associated with *M. bovis* biology[24]. Therefore, being able to characterize the genomic changes that occur alongside transmission will be helpful to find the genomic factors that influence *M. bovis* biology, such as host range.

In this proposal, I will discuss the development of the tool, *bovpan*, which will be used for the inference of gene presence/absence variation and SNPs from *M. bovis* sequence data. Using *bovpan,* I plan to investigate the genomic factors that contribute to *M. bovis* host specificity through **(i)** determining genomic signatures (such as accessory gene presence/absence, HR, SVs, and positive selection) that act as potential markers for *M. bovis* host specific lineages and **(ii)** investigating the specific genes and gene interactions, through SNP detection and gene presence/absence, found between two spatially close host species with indications of host specificity. The *expected outcomes* of this proposal will be (i) *bovpan,* which will enable users to determine the *M. bovis* pangenome for downstream analysis, (ii) genomic signatures that associate with *M. bovis* host specificity and, (iii) a list of genes and gene interactions, that have the ability to discriminate host species-associated *M. bovis* samples. These results will have a *positive impact* because understanding the genomic processes that contribute to *M. bovis* host adaptation can be useful in leveraging *M. bovis* biology to craft control strategies, which will alleviate the economic pressure bTB presents.

Innovation

This proposal will determine potential genomic signatures associated with host specificity in *M. bovis*, which has not been fully investigated in the *M. bovis* literature. To briefly summarize the perceived innovations, this proposal seeks to achieve the following advancements related to *M. bovis* adaptation research:

- Development of a new bioinformatics pipeline, *bovpan,* that can infer the *M. bovis* pangenome given sequence read data (Aim 1)

- Explore patterns of genome evolution that can be associated with specific host species and geographical regions through analysis of accessory genome clustering, homologous recombination, structural variation, and positive selection (Aim 2)

- Identify genes and gene interactions enriched in *M. bovis* isolated from specific hosts through statistical analysis (Aim 3)

Approach

Aim 1: Develop *bovpan*, a bioinformatics pipeline to analyze *M. bovis* genomic data:

Introduction: Current pipelines to investigate *M. bovis* evolution target SNPs as a measure of genomic variability[14,25]. Detected SNPs are then used to effectively investigate the spatial and temporal dynamics of *M. bovis* transmission through phylogenetic techniques. However, SNPs are only a subset of the total amount of genomic variation that is present within *M. bovis* samples*,* and certain strains of the bacteria possess genome characteristics that vary amongst other *M. bovis* samples[8]. Exclusively using SNPs for analyzing *M. bovis* samples might present challenges if other genomic variations are better for characterizing features of *M. bovis*. This challenge can be overcome by utilizing the same sequence data for SNP detection by inferring the pangenome, the entire set of conserved or variably present genes amongst a group of genomic samples. Therefore, the objective of this aim is to enhance *M. bovis* genome analysis by developing *bovpan,* a bioinformatics pipeline that will determine the *M. bovis* pangenome and consolidate the data for downstream analysis.

To achieve this, my approach will be to first implement *bovpan* using bioinformatics tools useful for pangenome inference[26], and then demonstrate *bovpan's* utility through analysis of real-world data. *Bovpan* will be implemented through a scalable, reproducible, and modular workflow management technology, Nextflow[27], to simplify the development and testing process. *Bovpan* will then be compared against conventional pipelines used to analyze *M. bovis* for accuracy in determining host range from *M. bovis* sequences. Comparing the predicted host species with the true value assigned before the analysis will produce the true positive rate (TPR) and the false positive rate (FPR), metrics that are useful to create a receiver operating characteristic curve to visually conclude *bovpan's* utility as a genome analysis pipeline. The rationale for developing *bovpan* is that this tool will assist in crafting datasets that can be leveraged to do further downstream analysis to link *M. bovis* evolution with phenotype. Upon the completion of Aim 1, it is my expectation that *bovpan* will be fully developed as a proper tool to generate genomic data from *M. bovis* sequence reads.

Justification and Feasibility: The growing use of WGS in epidemiological work necessitated the development of bioinformatic pipelines that could be used to infer SNPs in pathogenic bacteria outbreaks [28]. Typically, these pipelines take bacterial WGS as input and SNPs are then calculated as output. General pipelines to locate SNPs in microbial genomes include RedDog[25], Split K-mer Analysis [29], and GATK[30]. For *M. bovis* specific SNP detection, in-house pipelines are often developed for research projects [31], but there are limited *M. bovis* SNP pipelines available

for general use, with USDA vSNP[14] and the pipeline developed by Crispell et al., 2017[32] being major examples. All-in-one pipelines to gather data on a wider extent of bacterial genomic variations are also present, such as Bactopia[33] and TORMES[34]. However, there is not an *M. bovis* specific pipeline that can detect the full genomic variations present in *M. bovis* samples. Processing the sequence data also requires computationally intensive tools that can be non-intuitive to end users, difficult to edit because of unstable software design, and challenging to maintain in different computational environments. In an effort to alleviate some of these issues, sequence analysis pipelines such as RseqFlow[35], PRADA[36], and Galaxy[37], have been developed to simplify the process. Some of these pipelines are open-source, but often require the user to install certain dependencies separately, which may lead to errors in downstream analysis. *In-silico* workflow management systems such as Nextflow or Snakemake[38] make it simpler to develop sustainable, reproducible, and modular pipelines.

To understand the extent of genomic variations present in *M. bovis* samples, I developed in python a prototype pangenome analysis pipeline and analyzed 350 random *M. bovis* isolates downloaded from publicly available genome repositories. The genomes were reconstructed using a bacterial genome assembly software SPAdes[39], and the assembled contigs were aligned to the *M. bovis* reference genome NC_002945v4[17] to filter for low quality assemblies. If less than 90% of the contigs were mapped to the reference genome, the assembly was expunged from the analysis. A final dataset of 232 genomes was analyzed after filtering. Genome annotation software Prokka[40] and pangenome inference software Roary[41] were used to distinguish the core genome from the accessory genome. For the *M. bovis* core genome, 3784 genes (35% of the identified genes in the pangenome) were inferred, while 374 genes (4%) were variable in their presence in the genome. The total number of genes that were unique in presence/absence across samples represent 61% of the total genome. In the accessory genome, 3986 inferred genes (57%) were unknown in function, and 3,036 genes (43%) were annotated. This work showed how a pangenome pipeline could provide extra information that could be useful in *M. bovis* WGS studies.

Research Design:

Objective 1-1: Implementation of *M. bovis* pangenome pipeline, *bovpan*

Input for the pipeline will be *M. bovis* pair end sequence reads generated from Illumina sequencing, due to the increased accuracy that paired read data provides[5]. SNP calling will be conducted following the protocol established in a previously published study[31]. Reads will be mapped against the *M. bovis* AF2122/97 genome (GenBank Accession NC_002945.4) using the Burrows Wheeler Aligner tool[42]. SNP Calling will be jointly calculated with SAMtools[43] and Genome Alignment Toolkit (GATK) software, and concordant SNPs calculated between the two methods will be kept. A minimum site coverage of 20 reads and allele frequency of 75% or higher was necessary for SNPs to be considered accurate. Otherwise, the base would be considered a missing call, with SNP sites having an excess of 10% missing calls leading to removal. SNPs falling in highly repetitive mycobacterial *pe* and *ppe* genes will be removed from the final alignment. Pangenome inference will be determined first by taking the reads and then by applying a de Bruijn *de novo* assembly algorithm implemented in SPAdes, due to its efficiency in working with short read data (Figure 1a). Assembly quality will be analyzed by calculating assembly metrics through the tool QUAST[44]. The resulting scaffolds are then annotated for gene features through the use of the tool Prokka, which is specifically designed for prokaryotic genome annotation (Figure 1a). Lastly, Roary will be used to construct clusters of orthologous genes (COGs) based on annotations present in each *M. bovis* assembly (Figure 1b). If a COG is made up of orthologous genes present in virtually every sample, then that COG will be defined as part of the core genome. Otherwise, the COG will have variable presence of

the gene between the samples, indicating that COG as a part of the accessory genome (Figure 1b).

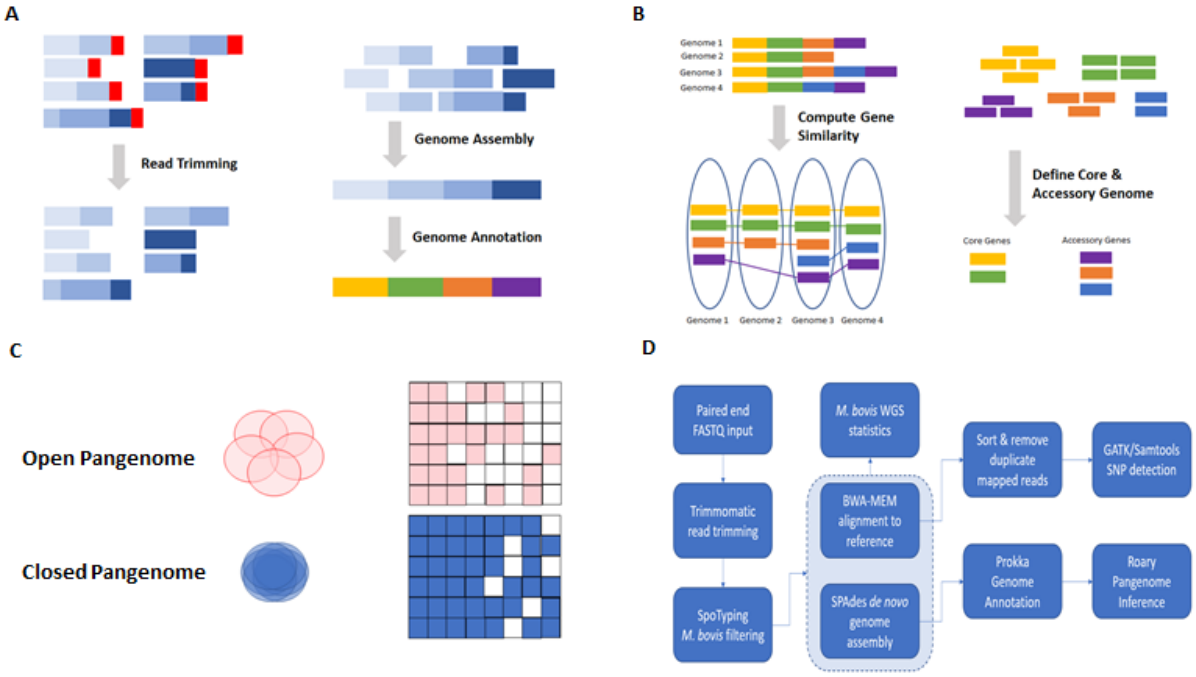


Figure 1. *The pangenome is defined as the entire genomic repertoire associated with a group of individuals from the same species, or closely related species.* ***A*** *To properly define a pangenome, sequence reads are first processed through read trimming to improve low quality ends (red) and remove adapters. The trimmed reads are next overlapped together to reconstruct the full M. bovis genome, and gene regions (non-blue) are identified through annotation.* ***B*** *Clusters of orthologous genes (COGs) are next computed by grouping genes that are present between genomes. If a COG possesses a gene that is present in virtually every sample, this gene is labelled as a core gene, and an accessory gene otherwise.* ***C*** *Pangenomes are defined as either open (the addition of more samples increases the gene set size; pink) or closed (The addition of more samples doesn't greatly influence the gene set size; blue).* ***D*** *The layout of bovpan, which takes in M. bovis sequence reads and infers both the pangenome and SNPs.*

To avoid issues of scalability, reproducibility, and portability when other users want to utilize this pipeline for their own analyses, the entire pipeline described above will be constructed using Nextflow. By defining modular sections of the pipeline that run independently from each other, the pipeline has the ability to be refactored without causing instability in the underlying software. Nextflow will also assist in alleviating issues of dependencies through its integrated Conda support[45], which allows for each section of code to have its own dependencies.

Objective 1-2: Compare pangenome and SNP only approach to identify host species

The development of the pipeline must be complemented with a demonstration in order to showcase *bovpan's* viability as a pipeline. To demonstrate usefulness of the pipeline, I will compute the *M. bovis* pangenome from samples with host-associated metadata, and use the genomic data to predict a sample's host species. The United Kingdom (UK) has collected a

large amount of sequence data associated with two host species implicated in the transmission of *M. bovis,* the Eurasian badger and cattle[3]. The true classification of isolates would be known, but in order to test the accuracy in predicting the UK host from the pipeline, I will first train 80% of data on a classification model based on (i) gene presence/absence and SNP data or (ii) SNP data alone. Logistic regression and linear support vector machine classification will be the two methods used to perform classification. The remaining 20% of the data will have their host species predicted, with comparison to the actual host species resulting in the calculation of the true positive rate and also the false positive rate. Both the TPR and FPR are useful in creating the Receiver Operating Characteristic (ROC) curve, which can be useful in visually determining if a method is more accurate compared to other methods. The area under the curve for a particular method will give an indication of how accurate that model is in classifying the samples. To create the ROC curve, the process of training and testing the models will require multiple iterations where each iteration randomizes which samples are used as training and testing data.

Expected Outcomes:

My expectation for this aim is to successfully develop *bovpan* using available bioinformatics software to extract pangenomic data given *M. bovis* paired end sequencing data. I also expect to showcase *bovpan* as an effective tool to generate genomic data for other analyses. Completion of this aim will lead to researchers having the ability to tie *M. bovis* characteristics to genomic factors, and provide a needed step towards understanding genomic signatures of *M. bovis* host specificity.

Potential Problems and Alternative Strategies:

Two potential problems can complicate the implementation of *bovpan.* For instance, sequence data is not guaranteed to be highly accurate when used as input to the pipeline, with sequencing reads typically having decreasing quality the further the read length progresses. This can lead to downstream steps of *bovpan* being biased by the error-prone data. To account for this potential problem, automatic read trimming will be employed, requiring that the reads be processed using the read trimming software Trimmomatic[46] (Figure 1a,d). This ensures that the sequence reads can be used in upcoming pipeline steps without introducing error-prone data. Another challenge of working with *M. bovis* sequence data is the chance that the reads are mislabeled as *M. bovis*, when actually the sequences are from a genomically similar species such as *Mycobacterium tuberculosis* that is 99.95% similar to *M. bovis*[17]. Spoligotyping, the characterization of mycobacterial specific Direct Repeat “spacer” regions, is an effective way of differentiating *M. bovis* from distantly related bacterial species as well as closely related *Mycobacterium tuberculosis* complex species[47] (Figure 1d). SpoTyping[48] will be utilized to define spoligotyping patterns directly from the trimmed reads, where the pattern of the spacer regions will be used to classify samples as being *M. bovis.*

Aim 2: Determine *M. bovis* host-specific genomic signatures:

Introduction: Analysis of *M. bovis* NGS typically employed SNP detection as a method to first define genomic variation and then associate that variation with hypotheses on pathogen transmission[16]. Recent work by researchers Patané et al.[8] and Zimpel et al.[9] have expanded the use of *M. bovis* NGS by incorporating genomic signatures to analyses to understand patterns of genomic variation of unique *M. bovis* strains. These analyses were limited in the amount of samples that were utilized, resulting in researchers being unable to study how genomic variations differed between different host-associated lineages. Host specific patterns of evolution may provide clues to how *M. bovis* adapts to particular host species. Identifying these variations as potential genomic signatures coinciding with host specificity can be useful as an

indicator that *M. bovis* is adapting to a particular host species. The objective of this aim is to investigate *M. bovis* genomic signatures (including gene presence/absence, HR, SVs, and positive selection) that associate with host range. To obtain this objective, my approach will be to utilize existing bioinformatic software alongside statistical metrics for association testing to relate pangenomic data with host species. The metric used to assess if association is present is the contingency coefficient, which takes a value of '0' when two random variables (in this case a genomic variation vs. a particular species) are independent of each other, while a value of '1' indicates that the variables are perfectly associated. The rationale for detecting genomic signatures that can distinguish host-associated *M. bovis* lineages is that this information can be used to explore if *M. bovis* is developing specificity in a new host population. After completing Aim 2, it is my expectation that I will have a list of genomic signatures that coincide with *M. bovis's* host range.

Justification and Feasibility: *M. bovis* is considered to be a generalist pathogen based on the large host range that the pathogen can infect[6]. This is further supported by the pattern of geographic clustering that is observed after phylogenetic inference of *M. bovis* SNPs[6]. While evidence for host specificity is lacking through phylogenetic analysis, other research has shown that miniscule genomic changes in *M. bovis* strains lead to variability in the pathogen's virulence and transmission characteristics[21]. Specific investigations into how evolutionary mechanisms are influenced within different host species populations are still an understudied focus of *M. bovis* biology. Whole genome analysis of *M. bovis* isolates from hosts using phylogenomic methods is increasing, but only a few researchers have investigated genotype/phenotype relationships. *M. bovis* whole genome analysis was used to tie genomic variations with virulence phenotypes through the comparison of 38 *M. bovis* sequences to understand genomic

A

Summarization of Host-Associated *M. bovis* data

	Species Count
New Zealand	
BOVINE-NZ	192
CERVINE-NZ	4
FERRET-NZ	38
PORCINE-NZ	7
POSSUM-NZ	54
STOAT-NZ	1
United Kingdom	
BADGER-UK	193
BOVINE-UK	159
USA	
BOVINE-Michigan	12
ELK-Michigan	5
WTD-Michigan	117

B

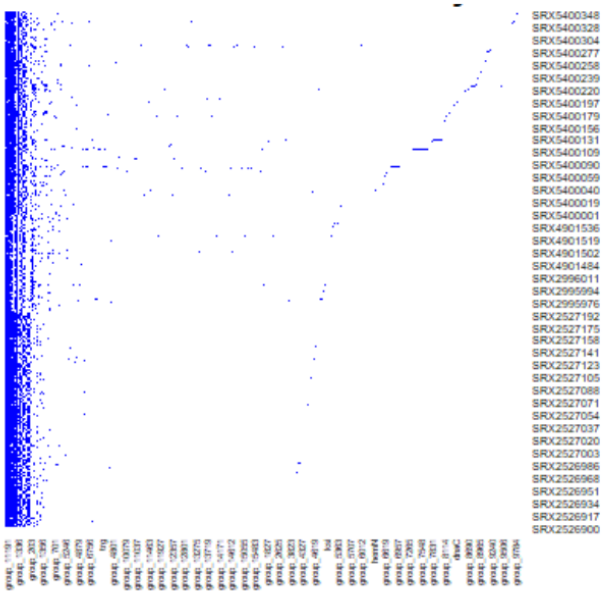


Figure 2. ***A*** *The amount of host-associated M. bovis samples that are represented in the dataset, stratified by different countries.* ***B*** *The gene presence/absence matrix representing the accessory genome of the preliminary 700 data samples. Dark blue navy color indicates a sample contains a particular gene.*

characteristics such as regions under positive selection and recombination sites [8]. Other research focused on using comparative genomics of *M. bovis* strains to investigate the genomic characteristics of an unique Brazilian strain *SP38* [9], while population structure analysis was conducted on isolates from Uruguay[10]. Although there is potential to use comparative methods to investigate the basis of *M. bovis* host specificity, an analysis of the genomic signatures of host-associated *M. bovis* has not been examined.

In order to understand the genomic factors that associate with the host range of *M. bovis,* I compiled a dataset of host-associated *M. bovis* samples from an online sequence repository. The dataset is composed of three rich publicly available datasets (782 isolates) of *M. bovis* samples collected from both livestock and wildlife (11 host-species represented) from the UK, United States of America, and New Zealand (Figure 2a). The isolates were then processed through a precursor pipeline that will inspire *bovpan,* but has generated *de novo* assemblies of the samples. Genome assembly statistics were quantified for each sample through the use of QUAST, and samples that produced statistics that were not biologically supported by *M. bovis* properties were next filtered out. This led to a dataset of 700 isolates for pangenome inference (Figure 2b) and preliminary clustering of the *M. bovis* accessory genome through Principal Component Analysis, PCA[49]. The results provide support that geographic proximity between hosts plays a role in shaping genomic evolution of *M. bovis*, but instances of host specific adaptation might still be occurring, suggested by the apparent separation of cattle associated and badger associated *M. bovis* samples from the UK.

Research Design

Objective 2-1: Cluster analysis of *M. bovis* accessory genome through dimensionality reduction

To investigate the variability associated with the accessory genome of *M. bovis,* I used the same data that was generated from the preliminary analysis. After inferring the pangenome of the dataset, I will obtain a data matrix that describes COGs for each *M. bovis* sample. For proper visualization, PCA will be utilized in order to preserve intrinsic characteristics of the dataset while simultaneously having the ability to interpret the variation of the accessory genome. After this dimension reduction, I will next use unsupervised clustering through k-means[50] to distinguish accessory genomes that are more similar based on their distance from each other[51,52]. Once cluster assignments are made, I will use association testing to validate associations between the country of origin or host species origin to decipher if evolution is based more on geography than species. The contingency coefficient value that is higher between the geographic and host data will provide an indication if the clustering is based more on geographic proximity.

Objective 2-2: Associate regions affected by and SVs with host range

HR could play a role in mycobacterial adaptation, as shown in research from Patané et al.[8]. Additionally, SVs for *M. bovis* were identified in five worldwide lineages, but the relationship between these SVs and host range was unable to be elucidated due to the small amount of sequence data[8]. To investigate if there exists heterogeneity in the amount of HR and SVs present in host-associated *M. bovis*, I will develop a whole genome alignment from the *M. bovis* sequences that is variable depending on the detected SNPs that were inferred through the pipeline. To detect HR, a maximum likelihood sliding window approach implemented through Gubbins[53] will be utilized to distinguish specific regions in each sample that exhibit some evidence for HR. Mapped read data output from *bovpan* will be used as input for Wham[54], which detects SV from short read data. I will again use contingency coefficients to see which association of host data is most correlated to the HR/SV events between isolates. Regions that

are highly associated with the host data and *M. bovis* genes that overlap with these regions will be recorded.

Objective 2-3: Compute positively selected genes from different host-associated population

To investigate if certain genes under positive selection can help differentiate *M. bovis* residing in different hosts, I will first create host-associated sub-alignments from the full *M. bovis* genome alignment. For each sub-alignment, the whole genome alignment will be further split into individual gene alignments based on coordinates from the *M. bovis* reference genome. IQTree[55] will be used to create individual gene trees for each gene alignment. The gene alignments and gene trees will be used as input for PAML[56], which will calculate dN/dS, a metric that can be used as an indicator of selective pressure acting on a protein coding gene. Genes that are determined to be evolving under positive selection will be recorded.

Expected Outcomes

My expectation for this aim is to determine the genomic signatures that associate with *M. bovis* host specificity. Cluster analysis of *M. bovis* pangenome, association testing of HR and SVs, and genes under positive selection will all be used to determine associations between host-associated *M. bovis* lineages and genomic variations. Completion of this aim will unravel the genomic factors that might be related to *M. bovis* adaptation in particular hosts.

Potential Problems and Alternative Strategies:

A potential problem that could arise is the limitations of dimension reduction techniques like PCA. The method is designed to capture patterns in the data based on present variation. This variation however is not guaranteed to be based on biological differences, but may be due to technical differences when preparing the dataset. This is labelled as a batch effect and can lead to error-prone interpretations of the resulting clustering. In order to limit the effects that technical variation might introduce into the analysis, principal component axes will be investigated to see if variation is based on technical aspects (such as choice of sequencing instrument). If there is an indication this is influencing the results, then the axis will not be used in the analysis. Eventually, the first two axes that are devoid of any technical bias will be the axes used for subsequent analysis.

Aim 3: Unravel *M. bovis* gene interactions that differentiate host-specific lineages:

Introduction: One predominant example of a genetic locus in *M. bovis* that appears to be essential for host pathogen interactions is the ESX-1 Type VII Secretion System[57]. Recent research into characteristics of *M. bovis* adaptation also demonstrated a role for ESX-1[58], potentially making the group of genes interesting candidates for understanding the role they play in *M. bovis* host range. However, an analysis of the ESX-1 genes and their relationship with host specificity has not been conducted despite the large amount of host-associated sequence data present in online repositories. Additionally, the processes that influence ESX-1 gene interactions when *M. bovis* circulates within a particular species is still understudied. The objective of this aim is to investigate ESX-1 genes and gene interactions that exist between two host species and could potentially act as indicators for host adaptation. To obtain the objective of this aim, my approach will be to generate genomic variation data from a large amount of *M. bovis* samples using the *bovpan* pipeline that was detailed in Aim 1. To limit the effects that species from different geographic regions might introduce, *M. bovis* sequence data associated with UK badgers and cattle will be exclusively used for this analysis. Following the computation of SNP and gene/presence absence data, I will proceed to perform a microbial genome wide association study (GWAS) and a genome wide epistatic study (GWES), alongside a pangenome GWAS (panGWAS)[59–61]. I will use these results to filter a gene presence/absence coincidence

network to find relevant genetic interactions[62]. For important genes and gene interactions that are predicted through this analysis, their correlation with membership to the ESX-1 pathway will be investigated. The rationale behind this aim is that if ESX-1 genes are shown to be implicated in *M. bovis* host specificity, then it will characterize a previously established mechanism for virulence as also important for host range. Upon the completion of Aim 3, it is my expectation that I will elucidate the role that ESX-1 genes have in promoting the adaptation of *M. bovis* to a particular host species.

Justification and Feasibility: ESX-1 type VII secretion systems (T7S system) are clusters of genes whose combined function provides mechanisms for mycobacterial virulence and host pathogen interactions[57]. Full or partial deletion of genes from the ESX-1 T7S system have led to mycobacterial pathogens becoming attenuated for successful infection, with the current *M. tuberculosis* vaccine, *M. bovis* BCG, being constructed through deletion of ESX-1[63]. Knockout experiments of ESX-1 components through random transposon insertion predominantly showed a limited ability of *M. bovis* to survive while being phagocytosed by both mammalian macrophages and soil dwelling amoeba predators[58], suggesting that ESX-1 has an ability to provide *M. bovis* mechanisms to survive in different types of hosts. However, the specific genes of ESX-1 that are implicated in the adaptation of *M. bovis* to a particular host are not well understood. Additionally, epistasis interactions are also important for understanding mycobacterial phenotype, with multiple studies focused mostly on *M. tuberculosis* epistatic interactions being helpful in promoting antibiotic resistance phenotypes[64]. The extent to which ESX-1 genes work in tandem with other mycobacterial genes to promote host specificity is ignored in genomic analyses of *M. bovis,* but understanding these interactions will lead researchers to better understand mechanisms behind *M. bovis* host specificity.

In order to study the effect that ESX-1 genes have on *M. bovis* adapting to a new host, sequence data from cattle and badger from the UK will be utilized. Preliminary clustering of the *M. bovis* accessory genome through PCA provided support that instances of host specific adaptation might be occurring, suggested by the apparent separation of cattle associated and badger associated *M. bovis* samples from the UK. These genomic data are only a small subset of the total sequence data available for comparative analysis, with 598 *M. bovis* isolates available from Allen et al.[65] and 619 *M. bovis* isolates from Akhmetova et al.[66].

Research Design

Objective 3-1: Conduct GWAS and GWES on *M. bovis* from the two different host species

To detect the SNPs that work independently and in tandem to help distinguish between *M. bovis* circulating in both badgers and cattle, I will use GWAS and GWES techniques. While GWAS analysis is a common technique in finding variants that can properly segregate sequence samples, these methods are not optimized to handle the wide variation presented by bacterial accessory genes and horizontal gene transfer. To circumvent this shortcoming, I will use a genome wide association study implemented through sequence element enrichment analysis (SEER)[67,68] to detect genome wide variations that can distinguish the host range phenotype of *M. bovis* samples. To find SNPs that show evidence of coevolution for specific host range, SpyderPick[61] will be used to compute pairwise SNPs that are enriched between cattle and badgers. To validate if the SNPs found have a correlation with membership in the ESX-1 T7S system, I will use Fisher Exact test as described in Butler et al.[58].

Objective 3-2: Investigate accessory genes associated with a particular host niche

Next, I will conduct an association study between accessory genes presence/absence and host species. The gene presence/absence data will be converted into binary ‘1’ (presence) and ‘0’ (absence). This data matrix will comprise genes on the columns, and isolate samples on the rows. In order to find out which genes within the accessory genome are enriched in a

particular host species, Scoary[60] will use a panGWAS approach to detect genes that can distinguish between particular host species. The top 20 accessory genes that are found in *M. bovis* from a particular host will be correlated with membership into the ESX-1 T7S System through a Fisher's Exact test. The results of this panGWAS will be further validated by association metrics and mutual information independence test to see if similar genes are enriched despite the method of determining association[69].

These enriched genes will next be analyzed to see if co-occurrence between the enriched genes are present. Coinfinder[62] will be used to create a gene co-occurrence network, and genes that are not enriched or not directly connected to enriched genes based on the Scoary analysis will be filtered out. For genes that are a part of the ESX-1 pathway, the co-occurrence network will be analyzed for differences in centrality metrics between ESX-1 pathway genes, general virulence associated genes, and non-pathogenic associated genes. Performing a Mann Whitney U Test will be able to detect differences in the distributions of the centrality metrics in a non-parametric approach. If the ESX-1 genes have significantly different centrality measured compared to other genes, then this will be an indication that the ESX-1 pathway has connections with adapting to the two particular species.

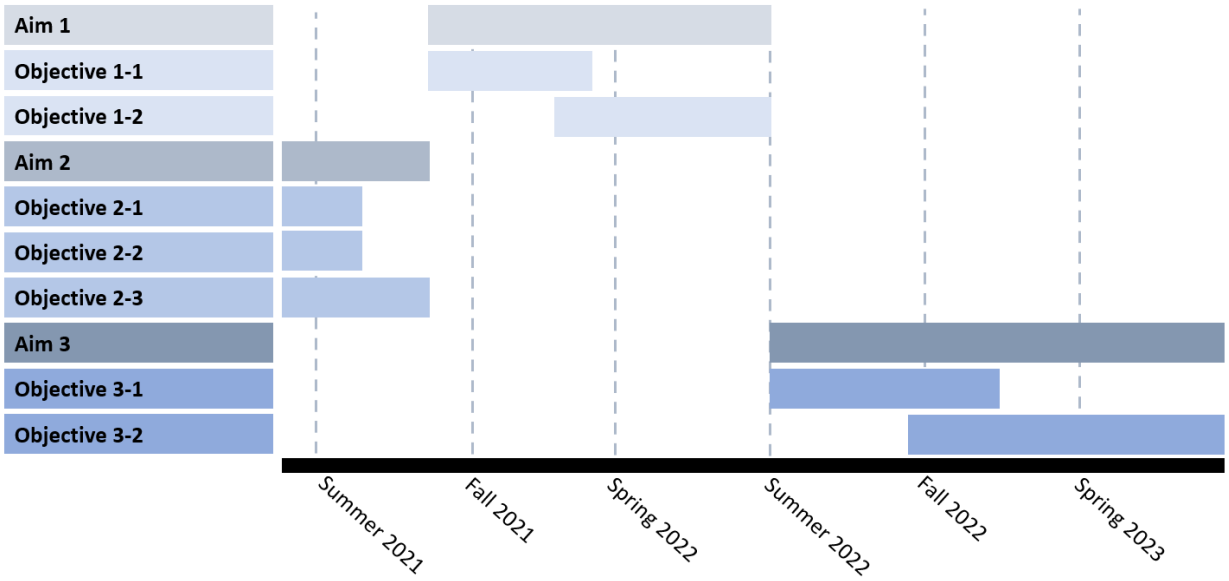
Expected Outcomes

My expectation for this aim is to determine the specific genes and gene interactions within the ESX-1 gene cluster that contribute to *M. bovis* host specificity in UK badgers and cattle. SNPs will be calculated and used to find genes with enriched amounts of mutations between badger and cattle, with interacting SNPs being computed through the use of GWES analysis. Gene presence/absence will be used to find both enriched genes, and the patterns of pairwise gene presence/absence will be used to compute networks of co-occurring genes. Completion of this aim will lead to determination of evolutionary processes that interact with *M. bovis* mechanisms to promote host specificity.

Potential Problems and Alternative Strategies

A potential problem that can arise through this analysis could be that mutational changes in ESX-1 genes and the genes they interact with are not enriched in either cattle or badgers, indicating that the genes are likely not responsible for *M. bovis* developing specificity within a host species. In the case that this might occur, I will shift from studying only the ESX-1 genes, and instead focus on other key genes/groups of genes that are enriched in cattle or badger. Since other groups of genes (such as *pe and ppe* genes that are unique to *M. bovis,* alongside *MCE4* cholesterol transport system genes) were similarly shown to be just as important for *M. bovis* phagocytosis survival as ESX-1 genes, this is a good indication that focusing on these other genes will assist in characterizing host specific variations in the absence of evidence from ESX-1 genes[58].

Timeline


Aim 1
Objective 1-1
Objective 1-2
Aim 2
Objective 2-1
Objective 2-2
Objective 2-3
Aim 3
Objective 3-1
Objective 3-2
Summer 2021
Fall 2021
Spring 2022
Summer 2022
Fall 2022
Spring 2023

Work Cited:

1. Palmer, M. V. *Mycobacterium bovis*: Characteristics of Wildlife Reservoir Hosts. *Transbound. Emerg. Dis.* **60**, 1–13 (2013).
2. Palmer, M. V., *et al*. *Mycobacterium bovis* : A Model Pathogen at the Interface of Livestock, Wildlife, and Humans. *Vet. Med. Int.*, 1–17 (2012).
3. Crispell, J. *et al.* Combining genomics and epidemiology to analyse bi-directional transmission of *Mycobacterium bovis* in a multi-host system. *eLife* **8**, e45833 (2019).
4. Yang, Z. & Rannala, B. Molecular phylogenetics: principles and practice. *Nat. Rev. Genet.* **13**, 303–314 (2012).
5. Olson, N. D. *et al.* Best practices for evaluating single nucleotide variant calling methods for microbial genomics. *Front. Genet.* **6**, (2015).
6. Zimpel, C. K. *et al.* Global Distribution and Evolution of *Mycobacterium bovis* Lineages. *Front. Microbiol.* **11**, (2020).
7. Salvador, L. C. M. *et al.* Disease management at the wildlife-livestock interface: Using whole-genome sequencing to study the role of elk in *Mycobacterium bovis* transmission in Michigan, USA. *Mol. Ecol.* **28**, 2192–2205 (2019).
8. Patané, J. S. L. *et al.* Patterns and Processes of *Mycobacterium bovis* Evolution Revealed by Phylogenomic Analyses. *Genome Biol. Evol.* **9**, 521–535 (2017).
9. Zimpel, C. K. *et al.* Complete Genome Sequencing of *Mycobacterium bovis* SP38 and Comparative Genomics of *Mycobacterium bovis* and *M. tuberculosis* Strains. *Front. Microbiol.* **8**, (2017).
10. Lasserre, M. *et al.* Whole genome sequencing of the monomorphic pathogen *Mycobacterium bovis* reveals local differentiation of cattle clinical isolates. *BMC Genomics* **19**, (2018).
11. Rouli, L. *et al*. The bacterial pangenome as a new tool for analysing pathogenic bacteria. *New Microbes New Infect.* **7**, 72–85 (2015).

12. Tettelin, H. *et al.* Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial “pan-genome”. *Proc. Natl. Acad. Sci.* **102**, 13950–13955 (2005).
13. Medini, D. *et al*. The Pangenome: A Data-Driven Discovery in Biology. in *The Pangenome: Diversity, Dynamics and Evolution of Genomes* (eds. Tettelin, H. & Medini, D.) 3–20 (Springer International Publishing, 2020).
14. *USDA-VS/vSNP*. (USDA-VS, 2020).
15. Crispell, J. *et al.* Using whole genome sequencing to investigate transmission in a multi-host system: bovine tuberculosis in New Zealand. *BMC Genomics* **18**, 180 (2017).
16. Biek, R. *et al.* Whole Genome Sequencing Reveals Local Transmission Patterns of *Mycobacterium bovis* in Sympatric Cattle and Badger Populations. *PLOS Pathog.* **8**, e1003008 (2012).
17. Garnier, T. *et al.* The complete genome sequence of *Mycobacterium bovis*. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 7877–7882 (2003).
18. Viana, M. *et al.* Assembling evidence for identifying reservoirs of infection. *Trends Ecol. Evol.* **29**, 270–279 (2014).
19. Haydon, D. T. *et al*. Identifying reservoirs of infection: a conceptual and practical challenge. *Emerg. Infect. Dis.* **8**, 1468–1473 (2002).
20. Hallmaier-Wacker, L. K., Munster, V. J. & Knauf, S. Disease reservoirs: from conceptual frameworks to applicable criteria. *Emerg. Microbes Infect.* **6**, e79 (2017).
21. Allen, A. R. One bacillus to rule them all? - Investigating broad range host adaptation in *Mycobacterium bovis*. *Infect. Genet. Evol. J. Mol. Epidemiol. Evol. Genet. Infect. Dis.* **53**, 68–76 (2017).
22. Volz, E. M. & Frost, S. D. W. Inferring the Source of Transmission with Phylogenetic Data. *PLOS Comput. Biol.* **9**, e1003397 (2013).
23. Fitak, R. R. *et al.* The Expectations and Challenges of Wildlife Disease Research in the Era

of Genomics: Forecasting with a Horizon Scan-like Exercise. *J. Hered.* **110**, 261–274 (2019).

24. Guimaraes, A. M. S. & Zimpel, C. K. *Mycobacterium bovis*: From Genotyping to Genome Sequencing. *Microorganisms* **8**, (2020).
25. Holt, K. *katholt/RedDog*. (2021).
26. Vernikos, G. S. A Review of Pangenome Tools and Recent Studies. in *The Pangenome: Diversity, Dynamics and Evolution of Genomes* (eds. Tettelin, H. & Medini, D.) 89–112 (Springer International Publishing, 2020).
27. Di Tommaso, P. *et al.* Nextflow enables reproducible computational workflows. *Nat. Biotechnol.* **35**, 316–319 (2017).
28. Leipzig, J. A review of bioinformatic pipeline frameworks. *Brief. Bioinform.* **18**, 530–536 (2017).
29. Harris, S. R. SKA: Split Kmer Analysis Toolkit for Bacterial Genomic Epidemiology. *bioRxiv* 453142 (2018)
30. Van der Auwera, G. A. *et al.* From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinforma.* **43**, 11.10.1-11.10.33 (2013).
31. Conceição, M. L. *et al.* Phylogenomic Perspective on a Unique Mycobacterium bovis Clade Dominating Bovine Tuberculosis Infections among Cattle and Buffalos in Northern Brazil. *Sci. Rep.* **10**, 1747 (2020).
32. Crispell, J. *et al. Mycobacterium bovis* genomics reveals transmission of infection between cattle and deer in Ireland. *Microb. Genomics* **6**, (2020).
33. Petit, R. A. & Read, T. D. Bactopia: a Flexible Pipeline for Complete Analysis of Bacterial Genomes. *mSystems* **5**, (2020).
34. Quijada, N. M. *et al*. TORMES: an automated pipeline for whole bacterial genome analysis. *Bioinformatics* **35**, 4207–4212 (2019).

35. Wang, Y. *et al.* RseqFlow: workflows for RNA-Seq data analysis. *Bioinformatics* **27**, 2598–2600 (2011).
36. Torres-García, W. *et al.* PRADA: pipeline for RNA sequencing data analysis. *Bioinforma. Oxf. Engl.* **30**, 2224–2226 (2014).
37. Jalili, V. *et al.* The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2020 update. *Nucleic Acids Res.* **48**, W395–W402 (2020).
38. Köster, J. & Rahmann, S. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics* **28**, 2520–2522 (2012).
39. Bankevich, A. *et al.* SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).
40. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).
41. Page, A. J. *et al.* Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* **31**, 3691–3693 (2015).
42. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
43. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
44. Gurevich, A. *et al*. QUAST: quality assessment tool for genome assemblies. *Bioinforma. Oxf. Engl.* **29**, 1072–1075 (2013).
45. Grüning, B. *et al.* Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nat. Methods* **15**, 475–476 (2018).
46. Bolger, A. *et al*. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
47. Bauer, J. *et al*. Usefulness of Spoligotyping To Discriminate IS6110 Low-Copy-Number *Mycobacterium tuberculosis* Complex Strains Cultured in Denmark. *J. Clin. Microbiol.* **37**,

2602–2606 (1999).

48. Xia, E., Teo, Y.-Y. & Ong, R. T.-H. SpoTyping: fast and accurate in silico Mycobacterium spoligotyping from sequence reads. *Genome Med.* **8**, 19 (2016).
49. Jolliffe, I. T. & Cadima, J. Principal component analysis: a review and recent developments. *Philos. Trans. R. Soc. Math. Phys. Eng. Sci.* **374**, 20150202 (2016).
50. Steinley, D. & Brusco, M. J. Choosing the number of clusters in K-means clustering. *Psychol. Methods* **16**, 285–297 (2011).
51. Freschi, L. *et al.* The *Pseudomonas aeruginosa* Pan-Genome Provides New Insights on Its Population Structure, Horizontal Gene Transfer, and Pathogenicity. *Genome Biol. Evol.* **11**, 109–120 (2019).
52. Jeukens, J. *et al.* A Pan-Genomic Approach to Understand the Basis of Host Adaptation in Achromobacter. *Genome Biol. Evol.* **9**, 1030–1046 (2017).
53. Croucher, N. J. *et al.* Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Res.* **43**, e15–e15 (2015).
54. Kronenberg, Z. N. *et al.* Wham: Identifying Structural Variants of Biological Consequence. *PLOS Comput. Biol.* **11**, e1004572 (2015).
55. Minh, B. Q. *et al.* IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
56. Yang, Z. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).
57. Gröschel, M. I. *et al*. ESX secretion systems: mycobacterial evolution to counter host immunity. *Nat. Rev. Microbiol.* **14**, 677–691 (2016).
58. Butler, R. E. *et al. Mycobacterium bovis* uses the ESX-1 Type VII secretion system to escape predation by the soil-dwelling amoeba Dictyostelium discoideum. *ISME J.* **14**, 919–930 (2020).
59. San, J. E. *et al.* Current Affairs of Microbial Genome-Wide Association Studies: Approaches,

Bottlenecks and Analytical Pitfalls. *Front. Microbiol.* **10**, (2020).

60. Brynildsrud, O., Bohlin, J., Scheffer, L. & Eldholm, V. Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. *Genome Biol.* **17**, 238 (2016).
61. Pensar, J. *et al.* Genome-wide epistasis and co-selection study using mutual information. *Nucleic Acids Res.* **47**, e112–e112 (2019).
62. Whelan, F. J. *et al.* Coinfinder: detecting significant associations and dissociations in pangenomes. *Microb. Genomics* **6**, e000338 (2020).
63. Matsuo, K. & Yasutomi, Y. *Mycobacterium bovis Bacille Calmette-Guérin* as a Vaccine Vector for Global Infectious Disease Control. *Tuberc. Res. Treat.*, 574591 (2011).
64. Borrell, S. *et al.* Epistasis between antibiotic resistance mutations drives the evolution of extensively drug-resistant tuberculosis. *Evol. Med. Public Health*, 65–74 (2013).
65. Allen, A. *et al.* Genome epidemiology of *Mycobacterium bovis* infection in contemporaneous, sympatric badger and cattle populations in Northern Ireland. *Access Microbiol.* **1**, 385.
66. Akhmetova, A. *et al.* Genomic epidemiology of *Mycobacterium bovis* infection in sympatric badger and cattle populations in Northern Ireland. *bioRxiv* 435101 (2021)
67. Lees, J. A. *et al*. pyseer: a comprehensive tool for microbial pangenome-wide association studies. *Bioinformatics* **34**, 4310–4312 (2018).
68. Lees, J. A. *et al.* Sequence element enrichment analysis to determine the genetic basis of bacterial phenotypes. *Nat. Commun.* **7**, 12797 (2016).
69. Kavvas, E. S. *et al.* Machine learning and structural analysis of *Mycobacterium tuberculosis* pan-genome identifies genetic signatures of antibiotic resistance. *Nat. Commun.* **9**, 4306 (2018).

APPENDIX

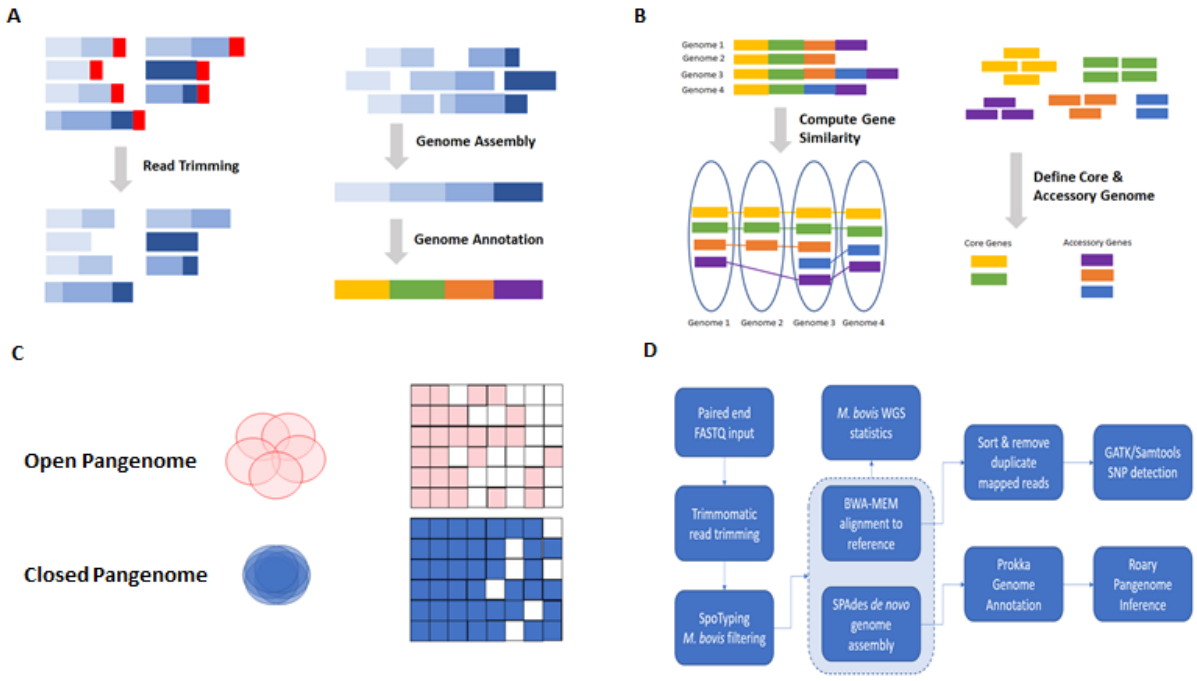


Figure 1. *The pangenome is defined as the entire genomic repertoire associated with a group of individuals from the same species, or closely related species.* ***A*** *To properly define a pangenome, sequence reads are first processed through a read trimming process to improve low quality ends (red) and remove adapters. The trimmed reads are next overlapped together to reconstruct the full M. bovis genome, and gene regions (non-blue) are identified through annotation.* ***B*** *Clusters of orthologous genes (COGs) are next computed by grouping genes that are present between genomes. If a COG possesses a gene that is present in virtually every sample, this gene is labelled as a core gene, and an accessory gene otherwise.* ***C*** *Pangenomes are defined as either open (the addition of more samples increases the gene set size; pink) or closed (The addition of more samples doesn't greatly influence the gene set size; blue).* ***D*** *The layout of bovpan, which takes in M. bovis sequence reads and infers both the pangenome and SNPs.*

APPENDIX

A

Summarization of Host-Associated *M. bovis* data	
	Species Count
New Zealand	
BOVINE-NZ	192
CERVINE-NZ	4
FERRET-NZ	38
PORCINE-NZ	7
POSSUM-NZ	54
STOAT-NZ	1
United Kingdom	
BADGER-UK	193
BOVINE-UK	159
USA	
BOVINE-Michigan	12
ELK-Michigan	5
WTD-Michigan	117

B

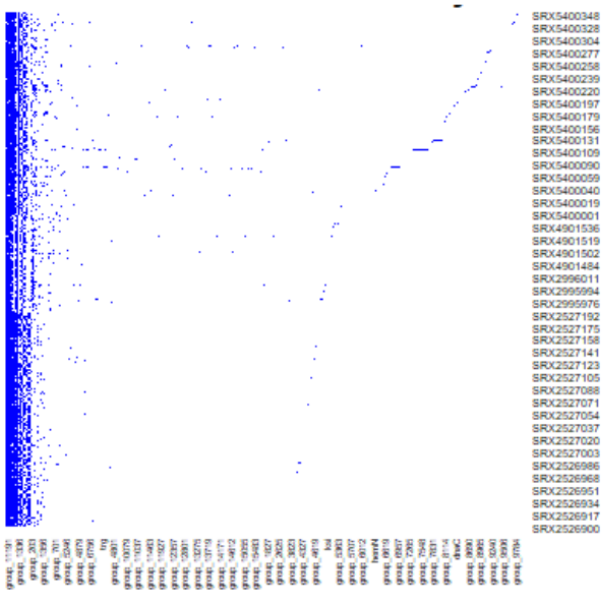


*Figure 2. **A** The amount of host-associated M. bovis samples that are represented in the dataset, stratified by different countries. **B** The gene presence/absence matrix representing the accessory genome of the preliminary 700 data samples. Dark blue navy color indicates a sample contains a particular gene.*