The *Mycobacterium tuberculosis* complex pangenome is small and driven by sub-lineage-specific regions of difference

Mahboobeh Behruznia[1,2], Maximillian Marin[3], Maha Farhat[3,4], Jonathan C Thomas[1], Maria Rosa Domingo-Sananes[1], Conor J Meehan[1,5*]

[1] Department of Biosciences, Nottingham Trent University, Nottingham, UK

[2] Institute of Microbiology and Infection, College of Medical and Dental Sciences, University of Birmingham, Birmingham, UK

[3] Department of Biomedical Informatics, Harvard Medical School, Boston, MA 02115, USA

[4] Pulmonary and Critical Care Medicine, Massachusetts General Hospital, Boston, MA 02114, USA

[5] Unit of Mycobacteriology, Institute of Tropical Medicine, Antwerp, Belgium

Corresponding author: conor.meehan@ntu.ac.uk

Authors ORCHID IDs: 0000-0002-3302-7272 (MB), 0000-0002-9108-3328 (MM), 0000-0002-3871-5760 (MF), 0000-0002-1599-9123 (JCT), 0000-0002-3339-8671 (MRDS), 0000-0003-0724-8343 (CJM)

Abstract

The *Mycobacterium tuberculosis* complex (MTBC) is a group of bacteria causing tuberculosis (TB) in humans and animals. Understanding MTBC genetic diversity is crucial for insights into its adaptation and traits related to survival, virulence, and antibiotic resistance. While it is known that within MTBC diversity is characterised by large lineage-specific deletions (regions of difference [RD]), a comprehensive pangenomic analysis incorporating both coding and non-coding regions remains unexplored. We utilised a curated dataset representing various MTBC genomes, including under-represented lineages to quantify the true diversity of the MTBC pangenome. The MTBC was found to have a small, closed pangenome with distinct genomic features and RDs both between lineages (as previously known) and between sub-lineages. The accessory genome was identified to be a product of

genome reduction, showing both lineage-specific and independent deletions. This variation has implications for traits like virulence, drug resistance, and metabolism. The study provides a comprehensive understanding of the MTBC pangenome, highlighting the importance of genome reduction in its evolution and showing that within-lineage genome content diversity is present. The findings underline the significance of genomic variations in determining the pathogenic traits of different MTBC lineages.

Introduction

The *Mycobacterium tuberculosis* complex (MTBC) is a group of closely related bacteria that cause tuberculosis (TB) in humans and other animals. Worldwide, TB is the 13th leading cause of death and is anticipated to once again become the top infectious killer following global efforts to combat the COVID-19 pandemic (WHO, 2023). The MTBC comprises both human-associated lineages (L1 to L9) and animal-associated lineages (La1, La2, La3, *M. microti* and *M. pinnipedii*) (Coscolla et al., 2021, Zwyer et al., 2021). The human-adapted MTBC exhibits a strong phylogeographical population structure, with some lineages occurring globally and others showing a strong geographical restriction (Gagneux, 2018). For example, L2 and L4 are widespread globally, with L2 dominating in Southeast Asia and L4 in Europe (Napier et al., 2020). L1 and L3 occur mainly in regions around the Indian Ocean (Menardo et al., 2021). L5 and L6 are highly restricted to West Africa (de Jong et al., 2010) and L7, L8 and L9 are exclusively found in East Africa (Blouin et al., 2012, Ngabonziza et al., 2020, Coscolla et al., 2021). Animal-associated lineages refer to lineages of the MTBC that primarily infect and cause disease in animals but can also infect humans in close contact with infected animals (Brites et al., 2018). Little is known about the geographic distribution and host range of animal-associated lineages (Zwyer et al., 2021).

Various factors such as human migration, local adaptation to host population, and host immune response are thought to have contributed to the current phylogeographical distribution of human-adapted MTBC members (Darwin and Stanley, 2022, Hershberg et al., 2008). Additionally, different lineages of MTBC are known to differ in virulence, metabolism and transmission potential, which can affect the spread of certain lineages and sub-lineages (Coscolla, 2017, Gagneux et al., 2006). Understanding the genetic diversity within MTBC is

58 essential for gaining insights into how these bacteria adapt to different host populations and
59 environments, and identify traits that may be important for survival, virulence, and antibiotic
60 resistance.

61 MTBC is recognised for its clonal population structure and little genetic diversity between
62 different strains, caused by the absence of horizontal gene transfer, plasmids, and
63 recombination (Godfroid et al., 2018, Chiner-Oms et al., 2022). Single nucleotide
64 polymorphisms (SNPs) and large sequence polymorphisms in the so-called regions of
65 difference (RDs) represent the source of MTBC genetic diversity (Gagneux, 2018). RDs are
66 known to differentiate different MTBC lineages from each other (Gagneux, 2018), but within
67 lineage diversity is thought to be mostly driven by SNPs (Napier et al., 2020). While this has
68 been shown to not be the case in some lineages (Sanoussi et al., 2021), these larger genomic
69 structural differences have not been fully explored. Deletions in some RDs can surpass 10 kb
70 in length and encompass groups of genes with diverse functions, which can be linked to
71 factors like virulence and metabolic diversity (Bespiatykh et al., 2021, Bottai et al., 2020).
72 Additionally, the *IS6110* element has a notable impact on the genetic diversity of the MTBC.
73 Transposition of *IS6110* into various coding or regulatory regions, and genomic deletions
74 through homologous recombination between proximal copies, can also translate into
75 significant genotypic variation at strain level (Alonso et al., 2013, Antoine et al., 2021).

76 The pangenome represents the set of all genes present in a species. It comprises both the
77 core genome, which consists of genes shared by all members of the species, and the
78 accessory genome, which includes genes present in some but not all members (Brockhurst
79 et al., 2019). Previous research has focused on pangenomic investigations of particular MTBC
80 lineages (Baena et al., 2023, Ceres et al., 2022, Sanoussi et al., 2021). A recent study

81 investigated the MTBC pangenome focusing on protein-coding sequences (Silva-Pereira et
82 al., 2024). However, there is a gap in comprehensive pangenomic analyses encompassing
83 both coding and non-coding regions. Despite the name, pangenome analyses tend to not
84 focus on full genome comparisons, but rather on protein-coding sequences. This focus
85 neglects non-coding RNAs and pseudogenes as well as essential elements in intergenic
86 regions, such as promoters, terminators, and regulatory binding sites, despite their
87 demonstrated role in selection and phenotypic impact (Tonkin-Hill et al., 2023). Moreover,
88 the inference drawn from pangenomic analyses is often influenced by the use of short-read
89 datasets. It has been shown that using short-read datasets can artificially inflate the size of
90 the pangenome estimates, leading to incorrect assumptions about evolutionary scenarios
91 (Marin, 2024). Here, we used a diverse collection of closed MTBC genomes including human
92 and animal-associated lineages to develop our understanding of MTBC pangenome and
93 evolution incorporating both coding and non-coding regions. The objectives of this study
94 are: (i) to establish the true size of MTBC pangenome using a highly curated dataset, (ii) to
95 assess specific biological functions that are prevalent in the core and accessory genome and
96 (iii) to assess genetic variation between lineages and sub-lineages that may affect virulence,
97 drug resistance, metabolism and evolutionary characteristics. We find that the MTBC has a
98 small, closed pangenome and genome reduction is the driving evolutionary force that
99 shapes the genomic evolution of the species. We identified distinct genomic features among
100 lineages and sub-lineages, including sub-lineage specific RDs, that can account for variations
101 in virulence, metabolism, and antibiotic resistance.

102

103

104 # Methods

105 **Summary of sequencing data and quality control**

106 Previous studies have shown that closed genomes should be used for best estimation of
107 pangenome size (Marin et al., 2024). From a combination of previously published research
108 and publicly available data, 322 non-duplicated complete and closed (single contig) genome
109 assemblies were selected for having (i) long-read sequencing (Nanopore or PacBio) or (ii)
110 hybrid sequencing with Illumina across the MTBC lineages. Information related to the
111 geographical location, source of isolation, and assembly method can be found in
112 Supplementary File 1. The quality of all assemblies was assessed using BUSCO v5.4.7 (Manni
113 et al., 2021) and only assemblies with ≥ 95% completeness were used for further analysis.
114 The assemblies were set to start on *dnaA* gene using the fixstart command of Circlator v1.5.5
115 (Hunt et al., 2015). Complete circular assemblies were annotated using PGAP v6.4 with
116 default options (Tatusova et al., 2016). This approach is recommended by previous analyses
117 (Marin et al., 2024) as PGAP tends to annotate pseudogenes in cases of pseudogenisation
118 and frameshift, rather than annotating them as new coding sequences. Visualisations of the
119 PGAP annotation were made using the DNA Features Viewer Python library (Zulkower and
120 Rosser, 2020). Lineage classification was performed using TB-profiler v4.4.2 (Phelan et al.,
121 2019).

122 **Nanopore sequencing and assembly**

123 We sequenced lineages and sub-lineages that either had no sequencing data available in
124 public databases or were under-represented in the collection. These isolates belonged to L5
125 (n=2, L5.1 and L5.2), L4.6 (n=3, L4.6.1, L4.6.1.1 and L4.6.1.2), L4.7 (n=3), L9 (n=1), La2 (*M.*
126 *caprae,* n =1) and *M. pinnipedii* (n=1). A list of genomes sequenced in this study and their

127 assembly information can be found in Table S1. Assembled genomes were deposited at the
128 NCBI under the Bioproject PRJNA1085078.

129 DNA extraction suitable for PacBio SMRT sequencing was performed using the Genomic-tip
130 100/G (Qiagen Inc) as described previously (Ngabonziza et al., 2020). DNA was enriched in
131 samples with low DNA concentration (1-60 ng/µl) using AMPureXP Beads (SPRIselect,
132 Beckman Coulter) in a 1:1 ratio. Samples were incubated for 10 minutes at room
133 temperature followed by pelleting the beads on a magnetic stand. The supernatant was
134 discarded, and DNA was eluted from the bead using 10-20 µl of nuclease-free water
135 (ThermoFisher Scientific). DNA concentration was quantified using a Qubit fluorometer and
136 1X dsDNA High Sensitivity (HS) assay kit (ThermoFisher Scientific). Nanopore sequencing was
137 performed using the native barcoding kit 24 V14 (SQK-NBD114.24) according to the
138 manufacturer's instruction on the MinION Mk1C platform with R10.4.1 flow cells. Sequence
139 adaptors were trimmed using Porechop v0.2.4 with "--end_threshold 95" and "--
140 middle_threshold 85" options (https://github.com/rrwick/Porechop.git). Sequences were
141 assembled using flye v2.9.2 with default options (Kolmogorov et al., 2019) and iteratively
142 polished using Racon v1.5.0 (Vaser et al., 2017) and the "consensus" subcommand of
143 Medaka v1.8.0 (https://github.com/nanoporetech/medaka.git). Complete assemblies
144 underwent Circlator, BUSCO and PGAP steps as above.

145 **Pangenome analysis**

146 Panaroo was used to characterise the MTBC protein-coding pangenome (Tonkin-Hill et al.,
147 2020) using the "--merge_paralogs", "--clean-mode strict" and "--core_threshold 0.95"
148 options. Ceres et al. (2022) showed that genes impacted by pseudogenisation could emerge
149 as accessory genes. To find split genes due to pseudogenisation, the reference pangenome

150 created by Panaroo was compared to the H37Rv genome using BLASTn with a similarity e-
151 value cutoff of 1×10^{-10} (Zhang et al., 2000). The BLASTn result was then filtered to only
152 include queries with ≥90% identity and ≥75% of the H37Rv gene length covered using in-
153 house Python scripts (Conor, could you please add the script to your GitHub page?). A core
154 genome alignment was constructed using the built-in MAFFT aligner included in Panaroo
155 (Katoh et al., 2002). The core phylogeny was created using RAxML-NG v1.2.0 with a GTR
156 substitution model (Kozlov et al., 2019). Tree visualisations were created using the Treeio
157 package v1.20.0 (Wang et al., 2020).

158 **Pangenome graph construction**

159 To look at total pangenome content (both coding and non-coding), multiple whole genomes
160 were aligned into a graph using Pangraph v0.7.2 (Noll et al., 2023). The DNA sequences from
161 gene blocks present in at least one sub-lineage but completely absent in others were
162 extracted. These sequences were subjected to BLAST against the H37Rv genome, along with
163 RvD1 and TbD1 genes to identify their corresponding Rv loci within the H37Rv genome and
164 determine sub-lineage-specific regions of difference.

165 **Functional genome analysis**

166 The reference pangenome created by Panaroo was translated using transeq v5.0.0 (Rice et
167 al., 2000) to obtain the protein sequences. COG functional categories were assigned using
168 eggNOG Mapper v5.0.2 using default options (Huerta-Cepas et al., 2017). Genes that did not
169 return a match within the eggNOG database were included in the not found category.
170 Functional domain annotation was performed with InterProScan v5.63-95 (Jones et al.,
171 2014). For the gene association analysis, the accessory genome (genes present in less than
172 95% of genomes) was used as input for Coinfinder v1.2.1 (Whelan et al., 2020). STRING

database (https://string-db.org/) was used to infer interaction networks of genes. Predicted functional partners with a high confidence interaction score (≥0.7) by STRING v12.0 were considered in this study.

Statistical analysis

Statistical analysis was conducted using R studio v2023.12.0+369 (R Core Team, 2023). Principal component analysis (PCA) based on gene presence and absence data from Panaroo and Pangraph derived accessory genome contents was performed using the prcomp package in R. The fviz_contrib function in Factoextra (v1.0.7) was used to identify the contribution of important variables from the PCA results (Kassambara and Mundt, 2020).

184 Results

185 **Overview of the genomic dataset**

186 We combined a total of 335 whole-genome sequences covering all known MTBC lineages. This
187 included 324 genomes published previously, as well as 11 genomes from sub-lineages with few or no
188 representatives in our collection to better cover the known phylogeny of MTBC (Supplementary Table
189 S1). Except Antarctica, our collection includes genomes from every continent, and different host
190 species including humans, cattle, voles and BCG vaccine strains (Figure 1-A). Africa represents the
191 highest diversity of MTBC genomes among other continents, which is consistent with the fact that
192 the continent is known to have the most diverse MTBC strains (Figure 1-B) (Gagneux, 2018). Most L1
193 genomes have been collected from East Africa, South and Southeast Asia, where this lineage is
194 geographically common. L3 genomes were mostly isolated from India, while L2 and L4 genomes have
195 been collected from different continents, representing the global distribution of these lineages. L5
196 and L6 genomes were exclusively found in West Africa, where they contribute significantly to the
197 overall burden of TB in this region (Coscolla et al., 2021).

198 **Pangenome characteristics and functional diversity**

199 Panaroo estimated a protein-coding pangenome consisting of 4,252 genes, including 3,639
200 core genes (present in more than 99% of genomes) and 613 accessory genes including 258
201 soft core, 218 shell and 137 cloud genes. After mapping all the genes in the pangenome to
202 H37Rv, we identified 127 genes that were misclassified as accessory genes, including 25 out
203 of 258 soft core, 24 out of 218 shell and 78 out of 137 cloud genes (Supplementary Table
204 S2). This approach increased the core genome size from 3,639 to 3,746 and reduced the
205 accessory genome from 613 to 506. This shift occurred because 107 genes out of 127 genes
206 detected as false accessories were re-classified as part of the core genome. To examine
207 genomic diversity among lineages, we investigated variations in the size of the accessory

208 genome among different lineages. This analysis revealed significant differences in accessory
209 genome size among different lineages (Kruskal-Wallis test, p= 2.2e-16). Post-hoc Dunn tests
210 with Bonferroni correction identified significant pairwise differences between lineages
211 (Supplementary Table S3). L6, La1 and *M. microti* have significantly smaller accessory
212 genomes compared to others (Figure 2, Supplementary Table S3). L2 has a significantly
213 smaller accessory genome in comparison with L1, L3, and L4.

214 Functional annotation of core and accessory genes showed that genes of unknown function were the
215 largest functional group in both (26.8 % and 12.0% respectively). The core genome of MTBC encodes
216 a high proportion of genes involved in lipid transport and metabolism (category I) and cellular
217 transport and secretion systems (category N), indicating the importance of these biological functions
218 in defining the species. Functional breakdown of the accessory genome also revealed variations in
219 genes involved in DNA replication, repair and mobile elements (category L) and lipid transport and
220 metabolism (category I) across the species. A breakdown of the functional composition of the core
221 and accessory genome is shown in Figure 3.

222 **Population structure analysis**

223 The population structure was assessed using SNPs in the core genome and PCA plots were
224 constructed based on presence-absence patterns in the accessory genome determined with Panaroo
225 and Pangraph (Figure 4). In the PCAs, genomes group together based on their lineage, indicating
226 variations in accessory genome are driven mainly by lineage structure (Figure 4-B and 4-C). The first
227 principal component (PC1) highlights the contrast between L2 from other members of MTBC. In both
228 PCA plots, L2 formed a distinct group, defined primarily by RD105 and RD207 deletions. RD105
229 serves as a classic marker for all L2 genomes and RD207 deletion is found in all L2.2 sub-lineages; the
230 single L2.1 genome is separated from the L2.2 cluster due to the absence of RD207. The second
231 principal component (PC2) demonstrates the difference between modern human-associated lineages
232 (L2, L3 and L4) from other lineages. Important variables driving this separation are the TbD1 gene

233 deletion in modern lineages, and gene deletions associated with RD11 (prophage phiRv2) in L6, L7,
234 La1 and La3 genomes and some L1 sub-lineages. L1 genomes exhibit considerable genetic diversity
235 primarily due to RD3 (related to prophage phiRv1) and RD11. The latter was identified in all L1.1 sub-
236 lineages and certain L1.2 sub-lineages (L1.2.1, L1.2.2, and L1.2.2.2). The positioning of the single L8
237 genome, marked with a positive sign at the centre of PCAs, implies that L8, serving as the ancestor of
238 the MTBC, exhibits similarities in the accessory genome with other lineages.

239 **The accessory genome of MTBC**

240 Among the 506 filtered accessory genes from Panaroo output, 107 were identified as being specific
241 to certain lineages and sub-lineages (Figure 5-A, Supplementary Table S4). Similarly, we identified 82
242 lineage and sub-lineage-specific genomic regions ranging from 270 bp to 9.8kb, encompassing a total
243 of 116 genes and 4 non-coding RNA using the Pangraph approach (Figure 5-B, Supplementary Table
244 S4 and Supplementary File S1). Our analysis revealed both previously known RDs and new deletions
245 in the accessory genome, suggesting that it originates from genome deletions. Lineage and sub-
246 lineage-specific genomic regions were consistently identified using both methods, including TbD1,
247 RD9, RD10, RD105, RD303, RD312, RD316, RD702, RD711, RD713, RD750, ND1, RDoryx, and N-
248 RD25bov/cap (Figure 6).

249 Panaroo failed to identify partial gene deletions caused by RD9, RD10, RD13, RD105, RD207,
250 and RD702. For example, RD702 is present in all L6 and L9 genomes and covers a 641 bp
251 region (Figure 7). RD702 affects less than a quarter of the *bglS* gene (470 bp from 2,076 bp)
252 and *mymT* gene (126 bp from 162 bp gene) in the 5’ to 3’ direction as well as a regulatory
253 RNA (ncRv0186c) in the 3’ to 5’ direction. Panaroo did not identify the 470 bp deletion
254 downstream of the *bglS* and this gene was classified as being in the core genome. These
255 results could be explained by the fact that Panaroo uses a relaxed alignment threshold along
256 with neighbourhood gene information to cluster gene families (Tonkin-Hill et al., 2020). This

257 can lead to the grouping of partially deleted genes with those that are complete. The
258 Pangraph alignment approach identified partial gene deletion and non-coding regions of the
259 DNA that were impacted by genomic deletion. However, the graph approach failed to
260 identify four structural variants, including a gene deletion corresponding to RD8 (Rv3618) in
261 L6 and animal-associated genomes. Despite these limitations, both methods yielded
262 consistent results indicating that the accessory genome of MTBC is small and is acquired
263 vertically from a common ancestor within the lineage. Gene association analysis using
264 coinfinder confirmed that most accessory genes (as determined by Panaroo) appear in pairs
265 or groups in larger multi-gene RDs (as determined by Pangraph) (Supplementary Figure S1).
266 Thus, the small accessory genome is contained within an even smaller number of RDs.
267 Coincident gene blocks that were associated with lineages were RD8, RD9, RD10, and
268 RD178, whereas those that were distributed across the phylogeny belonged to RD3 and
269 RD11.

Discussion

MTBC accessory genome is small and is acquired vertically

The MTBC pangenome size has been reported to vary across studies. In a recent investigation conducted by Marin (2024), the accessory genome ranged from 314 to 2951 genes, as a function of the assembly quality, software and specific parameters used. Genome sequencing using short-read technology has some limitations that can result in incomplete assemblies, particularly in regions with repetitive sequences that are difficult to sequence (Meehan et al., 2019). About 10% of the coding sequences of the MTBC genome are composed of repetitive polymorphic sequences known as the PE (Proline-Glutamic acid) and PPE (Proline-Proline-Glutamic acid) families, implicated in host-pathogen interactions and virulence (Cole et al., 1998). Incomplete assemblies can impact the accurate detection of genes and may lead to artificial inflation of the estimated pangenome size (Ceres et al., 2022, Marin et al., 2024). We used a curated dataset to address these issues and estimate the true size of the MTBC pangenome. We showed that MTBC has a compact accessory genome that is acquired vertically. This finding is consistent with recent research, which found a small accessory genome with a clonal genetic structure in MTBC (Baena et al., 2023, Ceres et al., 2022, Silva-Pereira et al., 2024). Using a BLASTn analysis approach, we identified core genes that were designated as accessory and further reduced the accessory genome from 613 to 506. For example, indel mutations observed in the *lppA* gene across 128 genomes resulted in the misclassification of this gene as an accessory (labelled as group_1700). Notably, over 50% of the cloud genes (genes present in less than 15% of strains) were found to be core genes. These genes mapped to the reference genes using the BLAST approach and appeared to be fragmented genes caused by misassemblies.

293 **Attributes of the accessory genome**

294 We employed a pangenomic approach to analyse the genetic diversity within the MTBC,
295 aiming to understand its evolution and functional potential. In addition to inferring a gene-
296 based pangenome with Panaroo, we also analysed variation in non-coding regions by
297 inferring a pangenome graph with Pangraph. Examining non-coding regions is particularly
298 important for species with limited genetic diversity, as they play a vital role in controlling
299 gene expression, stress response and host adaptation (Arnvig et al., 2014). Our study
300 demonstrated that, although both methods produced reliable outcomes, the whole-genome
301 alignment approach could capture higher genetic variations, especially those involving large
302 structural variants (>200 bp). Moreover, this method identified non-coding regions of the
303 genome that were affected by genomic deletions. Results from both methodologies revealed
304 that the accessory genome of MTBC is a product of gene deletions, that can be classified
305 into lineage-specific deletions and lineage-independent deletions, with the former category
306 being the more common. Coinfinder results also showed that coincident genes in the
307 accessory genome either belonged to particular lineages or were dispersed throughout the
308 phylogeny. Region of difference, which are large deletions often encompassing multiple
309 genes, can act as distinctive markers not only for lineages, as demonstrated previously, but
310 also for sub-lineages. These warrant further investigation to understand the biological
311 consequences of such deletions on traits like virulence, drug resistance, and metabolism,
312 along with SNP only analyses. On the other hand, we observed several genomic deletions
313 that appear to be lineage-independent and may indicate gene instability (Brosch et al.,
314 2002). This includes RD3 and RD11 corresponding to the two prophages phiRv1 and phiRv2.
315 One drawback of our study was that many of the lineages were poorly represented,
316 especially in the L5, L6, L9 and all animal-associated lineages. The production of more closed

317 genomes from the geographic regions where these are abundant (e.g. East and West Africa)
318 will provide further information about the RD distributions within the MTBC and better
319 understanding of the driving evolutionary forces.

320 **Differential traits between lineages and sub-lineages**

321 The size of the accessory genome content varied significantly among different lineages. L2
322 genomes exhibited a high degree of clonality and a smaller accessory genome. L2 strains are
323 characterised by a large number of *IS6110* copies, possibly due to the RD207, which affects
324 the CRISPR-Cas type III-A system, responsible for adaptive immunity against mobile genetic
325 elements (Alonso et al., 2013). While RD207 encompasses seven genes (Rv2814c-Rv2820c)
326 (Bespiatykh et al., 2021), our analysis revealed that only two genes, Rv2819c and Rv2820c,
327 are specifically absent in L2 and can serve as a genetic marker for this lineage. Other RDs
328 affecting all L2 genomes are RD3 (9.2 kb), and RD105 (3.4 kb). The latter is exclusively
329 present in all L2 genomes and is caused by the insertion of *IS6110* (Alonso et al., 2013).
330 Previous research proposed a potential role of RD105 in conferring drug resistance in L2.2
331 strains (Qin et al., 2019). RD207 (7.4 kb), linked to the insertion of the *IS6110* element, is
332 also present in all L2 genomes except for L2.1. L3 genomes exhibited significantly less
333 genomic deletion in comparison with L2, and RDs affecting all L3 genomes were RD316 (1.2
334 kb) and RD750 (765 bp). The genes affected by these RDs are involved in lipid metabolism
335 (category I) and secondary metabolite biosynthesis and transport respectively (category Q),
336 which may contribute to metabolic variations in L3 strains. L4 exhibits substantial genetic
337 diversity primarily due to variations in the presence and absence of specific RDs at the sub-
338 lineage level. For example, RD145 (4.7 kb) and RD182 (6.5 kb) were detected in L4.1.2.1
339 genomes isolated from Africa, Asia, Europe and South America. However, none of the

340 affected genes appears advantageous or essential for MTBC growth and survival and this
341 sub-lineage is successful worldwide (https://tbdr.lshtm.ac.uk/sra/browse). Other L4 sub-
342 lineage-specific RDs detected in this study with no known advantage are RD115 in L4.3.4,
343 RD727 in L4.6.1, RD219 in L4.8, RvD1 in L4.8 and L4.9. RD4 in La1, L5.3 Del in L5.3, and
344 RD750 in L3, RD147c in L1.

345 Nonetheless, some genomic deletions can affect the growth advantage gene group in
346 various lineages. Disruption of these genes has been shown to provide a fitness advantage
347 for the *in vitro* growth of H37Rv by analysis of saturated Himar 1 transposon libraries
348 (DeJesus et al., 2017). For example, RD174 specific to L4.3.4 intersects with RD743 specific
349 to L5, and both deletions affect two genes (Rv1995 and Rv1996) including *cmtR*
350 transcriptional regulator. The latter regulates bacterial stress response to metal deficiency
351 and is important during persistence in phagocyte (Flores-Valdez et al., 2020). While the
352 absence of these genes may provide a fitness advantage for *in vitro* growth, the impact of
353 this deletion is likely to be different in *in vivo* conditions. The low prevalence of L4.3.4 strains
354 in the TB Profiler database may suggest either the lower fitness of this sub-lineage, or its
355 presence in under-sampled geographic areas. RDs impacting the growth advantage gene
356 group were prevalent in L5 genomes (L5.3 Del, RD743 and RD711). For instance, RD711
357 affects the Rv1335 gene, which is annotated as a sulphur carrier protein involved in cysteine
358 biosynthesis by PGAP. This may imply that the genome reduction observed in L5 has evolved
359 to depend on host factors for essential nutrients, consequently making these strains difficult
360 to grow in a culture medium (Sanoussi et al., 2017).

361 L6 and animal-associated lineages accumulated a higher number of deletions including RD7,
362 RD8, RD9 and RD10. Comparative analysis of the accessory genome size in L6, La1, and *M.*

363 *microti* revealed significant differences compared to other lineages. The identified genomic
364 regions in L6, L9, and animal-associated lineages impact genes associated with virulence,
365 potentially explaining their reduced virulence in human hosts. RD7 contains the mammalian
366 cell entry (*mce3*) operon implicated in the invasion of host cells, while RD9 includes genes
367 related to biliverdin reduction, potentially aiding in protection against oxidative stress inside
368 macrophages (Szklarczyk et al., 2023). RD10 impacts genes related to lipid storage crucial for
369 MTBC transition to a dormant state (Deb et al., 2009). RD1BCG and RD1mic affect genes
370 encoding the ESX-1 secretion system, crucial for virulence and phagosomal escape, with
371 RD1BCG being deleted during BCG vaccine development (Szklarczyk et al., 2023). Overall,
372 the findings of this study suggest that, despite having a compact accessory genome, it can
373 account for certain biological distinctions among MTBC. Previous research has highlighted
374 the significance of random genetic drift in shaping the evolutionary dynamics of the species
375 (Hershberg et al., 2008). Gene loss, driven by genetic drift, is likely to be a key contributor to
376 the observed genetic diversity within the MTBC.

377 **A need for more high-quality genomes**

378 While this dataset represents high-quality MTBC genome sequences that cover most of its
379 known diversity, it suffers from sampling bias as some lineages are over-represented
380 whereas others have a few representatives. The over-representation of L2 and L4 in the
381 dataset is primarily attributed to their association with Bioprojects analysing TB outbreaks or
382 originating from countries with extensive sampling, highlighting a bias in sequencing efforts
383 towards more prevalent or historically studied lineages. Many sub-Saharan African countries
384 do not have a representative sample in our dataset, and this may explain the under-
385 representation of L5 and L6 in the dataset. These lineages contribute significantly to the

386 overall burden of TB in this region (Coscolla et al., 2021). Additionally, we found that certain
387 RDs that were previously assumed to be lineage-specific were identified in other lineages
388 (Supplementary table S4). The observed variation in RD distribution emphasizes the need for
389 an expanded genomic dataset to provide a more comprehensive understanding of their
390 prevalence, specificity, and implications in the evolution of the MTBC. Future sequencing
391 efforts should target high-burden TB countries, ensuring a comprehensive representation of
392 the population structure to enhance our understanding of MTBC diversity and evolution.

393 **Conclusion**

394 In conclusion, this study reveals several significant findings. Firstly, the MTBC has a small,
395 closed pangenome, predominantly driven by genome reduction. This observation underlines
396 the importance of genomic deletions in the evolution of the MTBC. The study also highlights
397 distinct genomic features across MTBC, including sublineage-specific RDs, which may play a
398 crucial role in variations in virulence, metabolism, and antibiotic resistance. The analysis
399 further reveals that the accessory genome of MTBC is a product of gene deletions, which can
400 be classified into lineage-specific and independent deletions. Finally, the study underscores
401 the impact of non-coding regions and large structural variations on MTBC's genetic diversity,
402 offering a new perspective on MTBC pangenomics.

403 **Acknowledgements and funding**

404 We would like to thank all the staff at the BCCM/ITM collection for the growth of strains for
405 sequencing in this study. We thank Alasdair Hubbard for assistance with Nanopore
406 sequencing.

407 C.J.M. and M.B. are supported by the Academy of Medical Sciences (AMS), the Wellcome
408 Trust, the Government Department of Business, Energy and Industrial Strategy (BEIS), the

British Heart Foundation and Diabetes UK and the Global Challenges Research Fund (GCRF) via a Springboard grant [SBF006\1090].

Author contributions

M.B. contributed to conceptualization, methodology, formal analysis, investigation, data curation, writing -original draft, writing – reviewing & editing, and visualization.

M.M. contributed to methodology, formal analysis, resources, data curation and writing – reviewing & editing.

M.F. contributed to methodology, data curation, writing – reviewing & editing and supervision.

J.C.T contributed to methodology, resources, and writing – reviewing & editing.

M.R.D.S. contributed to methodology, validation, formal analysis, investigation, writing – reviewing & editing and visualisation.

C.J.M. contributed to conceptualization, methodology, formal analysis, investigation, writing -original draft, writing – reviewing & editing, supervision, project administration and funding acquisition.

References

ALONSO, H., SAMPER, S., MARTÍN, C. & OTAL, I. 2013. Mapping IS6110 in high-copy number Mycobacterium tuberculosis strains shows specific insertion points in the Beijing genotype. *BMC Genomics,* 14**,** 422.

ANTOINE, R., GAUDIN, C. & HARTKOORN, R. C. 2021. Intragenic Distribution of IS6110 in Clinical Mycobacterium tuberculosis Strains: Bioinformatic Evidence for Gene Disruption Leading to Underdiagnosed Antibiotic Resistance. *Microbiol Spectr,* 9**,** e0001921.

ARNVIG, K. B., CORTES, T. & YOUNG, D. B. 2014. Noncoding RNA in Mycobacteria. *Microbiology Spectrum,* 2**,** 10.1128/microbiolspec.mgm2-0029-2013.

BAENA, A., CABARCAS, F., OCAMPO, J. C., BARRERA, L. F. & ALZATE, J. F. 2023. Large genomic deletions delineate Mycobacterium tuberculosis L4 sublineages in South American countries. *PLOS ONE,* 18**,** e0285417.

BESPIATYKH, D., BESPYATYKH, J., MOKROUSOV, I. & SHITIKOV, E. 2021. A Comprehensive Map of Mycobacterium tuberculosis Complex Regions of Difference. *mSphere,* 6**,** e0053521.

BLOUIN, Y., HAUCK, Y., SOLER, C., FABRE, M., VONG, R., DEHAN, C., CAZAJOUS, G., MASSOURE, P.-L., KRAEMER, P., JENKINS, A., GARNOTEL, E., POURCEL, C. & VERGNAUD, G. 2012. Significance of the Identification in the Horn of Africa of an Exceptionally Deep Branching Mycobacterium tuberculosis Clade. *PLOS ONE,* 7**,** e52841.

BOTTAI, D., FRIGUI, W., SAYES, F., DI LUCA, M., SPADONI, D., PAWLIK, A., ZOPPO, M., ORGEUR, M., KHANNA, V., HARDY, D., MANGENOT, S., BARBE, V., MEDIGUE, C., MA, L., BOUCHIER, C., TAVANTI, A., LARROUY-MAUMUS, G. & BROSCH, R. 2020. TbD1 deletion as a driver of the evolutionary success of modern epidemic *Mycobacterium tuberculosis* lineages. *Nature Communications,* 11**,** 684.

BRITES, D., LOISEAU, C., MENARDO, F., BORRELL, S., BONIOTTI, M. B., WARREN, R., DIPPENAAR, A., PARSONS, S. D. C., BEISEL, C., BEHR, M. A., FYFE, J. A., COSCOLLA, M. & GAGNEUX, S. 2018. A New Phylogenetic Framework for the Animal-Adapted Mycobacterium tuberculosis Complex. *Front Microbiol,* 9**,** 2820.

BROCKHURST, M. A., HARRISON, E., HALL, J. P. J., RICHARDS, T., MCNALLY, A. & MACLEAN, C. 2019. The Ecology and Evolution of Pangenomes. *Curr Biol,* 29**,** R1094-r1103.

BROSCH, R., GORDON, S. V., MARMIESSE, M., BRODIN, P., BUCHRIESER, C., EIGLMEIER, K., GARNIER, T., GUTIERREZ, C., HEWINSON, G., KREMER, K., PARSONS, L. M., PYM, A. S., SAMPER, S., VAN SOOLINGEN, D. & COLE, S. T. 2002. A new evolutionary scenario for the *Mycobacterium tuberculosis* complex. *Proceedings of the National Academy of Sciences,* 99**,** 3684-3689.

CERES, K. M., STANHOPE, M. J. & GRÖHN, Y. T. 2022. A critical evaluation of Mycobacterium bovis pangenomics, with reference to its utility in outbreak investigation. *Microb Genom,* 8.

CHINER-OMS, Á., LÓPEZ, M. G., MORENO-MOLINA, M., FURIÓ, V. & COMAS, I. 2022. Gene evolutionary trajectories in *Mycobacterium tuberculosis* reveal temporal signs of selection. *Proceedings of the National Academy of Sciences,* 119**,** e2113600119.

COLE, S. T., BROSCH, R., PARKHILL, J., GARNIER, T., CHURCHER, C., HARRIS, D., GORDON, S. V., EIGLMEIER, K., GAS, S., BARRY, C. E., TEKAIA, F., BADCOCK, K., BASHAM, D., BROWN,

472 D., CHILLINGWORTH, T., CONNOR, R., DAVIES, R., DEVLIN, K., FELTWELL, T., GENTLES,
473 S., HAMLIN, N., HOLROYD, S., HORNSBY, T., JAGELS, K., KROGH, A., MCLEAN, J.,
474 MOULE, S., MURPHY, L., OLIVER, K., OSBORNE, J., QUAIL, M. A., RAJANDREAM, M. A.,
475 ROGERS, J., RUTTER, S., SEEGER, K., SKELTON, J., SQUARES, R., SQUARES, S.,
476 SULSTON, J. E., TAYLOR, K., WHITEHEAD, S. & BARRELL, B. G. 1998. Deciphering the
477 biology of Mycobacterium tuberculosis from the complete genome sequence.
478 *Nature,* 393**,** 537-544.
479 COSCOLLA, M. 2017. Biological and Epidemiological Consequences of MTBC Diversity. *Adv*
480 *Exp Med Biol,* 1019**,** 95-116.
481 COSCOLLA, M., GAGNEUX, S., MENARDO, F., LOISEAU, C., RUIZ-RODRIGUEZ, P., BORRELL, S.,
482 OTCHERE, I. D., ASANTE-POKU, A., ASARE, P., SÁNCHEZ-BUSÓ, L., GEHRE, F.,
483 SANOUSSI, C. N., ANTONIO, M., AFFOLABI, D., FYFE, J., BECKERT, P., NIEMANN, S.,
484 ALABI, A. S., GROBUSCH, M. P., KOBBE, R., PARKHILL, J., BEISEL, C., FENNER, L.,
485 BÖTTGER, E. C., MEEHAN, C. J., HARRIS, S. R., DE JONG, B. C., YEBOAH-MANU, D. &
486 BRITES, D. 2021. Phylogenomics of Mycobacterium africanum reveals a new lineage
487 and a complex evolutionary history. *Microb Genom,* 7.
488 DARWIN, K. H. & STANLEY, S. A. 2022. The aldehyde hypothesis: metabolic intermediates as
489 antimicrobial effectors. *Open Biol,* 12**,** 220010.
490 DE JONG, B. C., ANTONIO, M. & GAGNEUX, S. 2010. Mycobacterium africanum—Review of
491 an Important Cause of Human Tuberculosis in West Africa. *PLOS Neglected Tropical*
492 *Diseases,* 4**,** e744.
493 DEB, C., LEE, C.-M., DUBEY, V. S., DANIEL, J., ABOMOELAK, B., SIRAKOVA, T. D., PAWAR, S.,
494 ROGERS, L. & KOLATTUKUDY, P. E. 2009. A Novel In Vitro Multiple-Stress Dormancy
495 Model for Mycobacterium tuberculosis Generates a Lipid-Loaded, Drug-Tolerant,
496 Dormant Pathogen. *PLOS ONE,* 4**,** e6077.
497 DEJESUS, M. A., GERRICK, E. R., XU, W., PARK, S. W., LONG, J. E., BOUTTE, C. C., RUBIN, E. J.,
498 SCHNAPPINGER, D., EHRT, S., FORTUNE, S. M., SASSETTI, C. M. & IOERGER, T. R. 2017.
499 Comprehensive Essentiality Analysis of the Mycobacterium tuberculosis Genome via
500 Saturating Transposon Mutagenesis. *mBio,* 8.
501 FLORES-VALDEZ, M. A., ACEVES-SÁNCHEZ, M. J., PETERSON, E. J. R., BALIGA, N., BRAVO-
502 MADRIGAL, J., DE LA CRUZ-VILLEGAS, M., ARES, M. A., BORN, S., VOSKUIL, M., PÉREZ-
503 PADILLA, N. A., BURCIAGA-FLORES, M., CAMACHO-VILLEGAS, T. A. & ESPINOZA-
504 JORGE, M. G. 2020. Transcriptional portrait of M. bovis BCG during biofilm
505 production shows genes differentially expressed during intercellular aggregation and
506 substrate attachment. *Sci Rep,* 10**,** 12578.
507 GAGNEUX, S. 2018. Ecology and evolution of Mycobacterium tuberculosis. *Nat Rev*
508 *Microbiol,* 16**,** 202-213.
509 GAGNEUX, S., DERIEMER, K., VAN, T., KATO-MAEDA, M., DE JONG, B. C., NARAYANAN, S.,
510 NICOL, M., NIEMANN, S., KREMER, K., GUTIERREZ, M. C., HILTY, M., HOPEWELL, P. C.
511 & SMALL, P. M. 2006. Variable host-pathogen compatibility in *Mycobacterium*
512 *tuberculosis*. *Proceedings of the National Academy of Sciences,* 103**,** 2869-2873.
513 GODFROID, M., DAGAN, T. & KUPCZOK, A. 2018. Recombination Signal in Mycobacterium
514 tuberculosis Stems from Reference-guided Assemblies and Alignment Artefacts.
515 *Genome Biol Evol,* 10**,** 1920-1926.
516 HERSHBERG, R., LIPATOV, M., SMALL, P. M., SHEFFER, H., NIEMANN, S., HOMOLKA, S.,
517 ROACH, J. C., KREMER, K., PETROV, D. A., FELDMAN, M. W. & GAGNEUX, S. 2008. High

Functional Diversity in Mycobacterium tuberculosis Driven by Genetic Drift and Human Demography. *PLOS Biology,* 6**,** e311.

HUERTA-CEPAS, J., FORSLUND, K., COELHO, L. P., SZKLARCZYK, D., JENSEN, L. J., VON MERING, C. & BORK, P. 2017. Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. *Mol Biol Evol,* 34**,** 2115-2122.

HUNT, M., SILVA, N. D., OTTO, T. D., PARKHILL, J., KEANE, J. A. & HARRIS, S. R. 2015. Circlator: automated circularization of genome assemblies using long sequencing reads. *Genome Biology,* 16**,** 294.

JONES, P., BINNS, D., CHANG, H. Y., FRASER, M., LI, W., MCANULLA, C., MCWILLIAM, H., MASLEN, J., MITCHELL, A., NUKA, G., PESSEAT, S., QUINN, A. F., SANGRADOR-VEGAS, A., SCHEREMETJEW, M., YONG, S. Y., LOPEZ, R. & HUNTER, S. 2014. InterProScan 5: genome-scale protein function classification. *Bioinformatics,* 30**,** 1236-40.

KASSAMBARA, A. & MUNDT, F. Extract and Visualize the Results of Multivariate Data Analyses [R package factoextra version 1.0.7]. 2020.

KATOH, K., MISAWA, K., KUMA, K. & MIYATA, T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res,* 30**,** 3059-66.

KOLMOGOROV, M., YUAN, J., LIN, Y. & PEVZNER, P. A. 2019. Assembly of long, error-prone reads using repeat graphs. *Nature Biotechnology,* 37**,** 540-546.

KOZLOV, A. M., DARRIBA, D., FLOURI, T., MOREL, B. & STAMATAKIS, A. 2019. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics,* 35**,** 4453-4455.

MANNI, M., BERKELEY, M. R., SEPPEY, M., SIMÃO, F. A. & ZDOBNOV, E. M. 2021. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Mol Biol Evol,* 38**,** 4647-4654.

MARIN, M. 2024. Analysis of the limited *Mycobacterium tuberculosis* pangenome reveals potential pitfalls of pan-genome analysis approaches

MEEHAN, C. J., GOIG, G. A., KOHL, T. A., VERBOVEN, L., DIPPENAAR, A., EZEWUDO, M., FARHAT, M. R., GUTHRIE, J. L., LAUKENS, K., MIOTTO, P., OFORI-ANYINAM, B., DREYER, V., SUPPLY, P., SURESH, A., UTPATEL, C., VAN SOOLINGEN, D., ZHOU, Y., ASHTON, P. M., BRITES, D., CABIBBE, A. M., DE JONG, B. C., DE VOS, M., MENARDO, F., GAGNEUX, S., GAO, Q., HEUPINK, T. H., LIU, Q., LOISEAU, C., RIGOUTS, L., RODWELL, T. C., TAGLIANI, E., WALKER, T. M., WARREN, R. M., ZHAO, Y., ZIGNOL, M., SCHITO, M., GARDY, J., CIRILLO, D. M., NIEMANN, S., COMAS, I. & VAN RIE, A. 2019. Whole genome sequencing of Mycobacterium tuberculosis: current standards and open issues. *Nature Reviews Microbiology,* 17**,** 533-545.

MENARDO, F., RUTAIHWA, L. K., ZWYER, M., BORRELL, S., COMAS, I., CONCEIÇÃO, E. C., COSCOLLA, M., COX, H., JOLOBA, M., DOU, H. Y., FELDMANN, J., FENNER, L., FYFE, J., GAO, Q., GARCÍA DE VIEDMA, D., GARCIA-BASTEIRO, A. L., GYGLI, S. M., HELLA, J., HIZA, H., JUGHELI, L., KAMWELA, L., KATO-MAEDA, M., LIU, Q., LEY, S. D., LOISEAU, C., MAHASIRIMONGKOL, S., MALLA, B., PALITTAPONGARNPIM, P., RAKOTOSAMIMANANA, N., RASOLOFO, V., REINHARD, M., REITHER, K., SASAMALO, M., SILVA DUARTE, R., SOLA, C., SUFFYS, P., BATISTA LIMA, K. V., YEBOAH-MANU, D., BEISEL, C., BRITES, D. & GAGNEUX, S. 2021. Local adaptation in populations of Mycobacterium tuberculosis endemic to the Indian Ocean Rim. *F1000Res,* 10**,** 60.

NAPIER, G., CAMPINO, S., MERID, Y., ABEBE, M., WOLDEAMANUEL, Y., ASEFFA, A., HIBBERD, M. L., PHELAN, J. & CLARK, T. G. 2020. Robust barcoding and identification of Mycobacterium tuberculosis lineages for epidemiological and clinical studies. *Genome Medicine,* 12**,** 114.

NGABONZIZA, J. C. S., LOISEAU, C., MARCEAU, M., JOUET, A., MENARDO, F., TZFADIA, O., ANTOINE, R., NIYIGENA, E. B., MULDERS, W., FISSETTE, K., DIELS, M., GAUDIN, C., DUTHOY, S., SSENGOOBA, W., ANDRÉ, E., KASWA, M. K., HABIMANA, Y. M., BRITES, D., AFFOLABI, D., MAZARATI, J. B., DE JONG, B. C., RIGOUTS, L., GAGNEUX, S., MEEHAN, C. J. & SUPPLY, P. 2020. A sister lineage of the Mycobacterium tuberculosis complex discovered in the African Great Lakes region. *Nature Communications,* 11**,** 2917.

NOLL, N., MOLARI, M., SHAW, L. P. & NEHER, R. A. 2023. PanGraph: scalable bacterial pan-genome graph construction. *Microbial Genomics,* 9.

PHELAN, J. E., O'SULLIVAN, D. M., MACHADO, D., RAMOS, J., OPPONG, Y. E. A., CAMPINO, S., O'GRADY, J., MCNERNEY, R., HIBBERD, M. L., VIVEIROS, M., HUGGETT, J. F. & CLARK, T. G. 2019. Integrating informatics tools and portable sequencing technology for rapid detection of resistance to anti-tuberculous drugs. *Genome Medicine,* 11**,** 41.

QIN, L., WANG, J., LU, J., YANG, H., ZHENG, R., LIU, Z., HUANG, X., FENG, Y., HU, Z. & GE, B. 2019. A deletion in the RD105 region confers resistance to multiple drugs in Mycobacterium tuberculosis. *BMC Biol,* 17**,** 7.

R CORE TEAM 2023. R: A language and environment for statistical computing. R Foundation for Statistical Computing. *R Core Team.* Vienna, Austria.

RICE, P., LONGDEN, I. & BLEASBY, A. 2000. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet,* 16**,** 276-7.

SANOUSSI, C. N., APOS, DIRA, COSCOLLA, M., OFORI-ANYINAM, B., OTCHERE, I. D., ANTONIO, M., NIEMANN, S., PARKHILL, J., HARRIS, S., YEBOAH-MANU, D., GAGNEUX, S., RIGOUTS, L., AFFOLABI, D., DE JONG, B. C. & MEEHAN, C. J. 2021. Mycobacterium tuberculosis complex lineage 5 exhibits high levels of within-lineage genomic diversity and differing gene content compared to the type strain H37Rv. *Microbial Genomics,* 7.

SANOUSSI, C. N. D., AFFOLABI, D., RIGOUTS, L., ANAGONOU, S. & DE JONG, B. 2017. Genotypic characterization directly applied to sputum improves the detection of Mycobacterium africanum West African 1, under-represented in positive cultures. *PLOS Neglected Tropical Diseases,* 11**,** e0005900.

SILVA-PEREIRA, T. T., SOLER-CAMARGO, N. C. & GUIMARÃES, A. M. S. 2024. Diversification of gene content in the Mycobacterium tuberculosis complex is determined by phylogenetic and ecological signatures. *Microbiology Spectrum,* 0**,** e02289-23.

SZKLARCZYK, D., KIRSCH, R., KOUTROULI, M., NASTOU, K., MEHRYARY, F., HACHILIF, R., GABLE, A. L., FANG, T., DONCHEVA, N. T., PYYSALO, S., BORK, P., JENSEN, L. J. & VON MERING, C. 2023. The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Res,* 51**,** D638-d646.

TATUSOVA, T., DICUCCIO, M., BADRETDIN, A., CHETVERNIN, V., NAWROCKI, E. P., ZASLAVSKY, L., LOMSADZE, A., PRUITT, K. D., BORODOVSKY, M. & OSTELL, J. 2016. NCBI prokaryotic genome annotation pipeline. *Nucleic Acids Res,* 44**,** 6614-24.

TONKIN-HILL, G., GLADSTONE, R. A., PÖNTINEN, A. K., ARREDONDO-ALONSO, S., BENTLEY, S. D. & CORANDER, J. 2023. Robust analysis of prokaryotic pangenome gene gain and loss rates with Panstripe. *Genome Res,* 33**,** 129-140.

TONKIN-HILL, G., MACALASDAIR, N., RUIS, C., WEIMANN, A., HORESH, G., LEES, J. A., GLADSTONE, R. A., LO, S., BEAUDOIN, C., FLOTO, R. A., FROST, S. D. W., CORANDER, J., BENTLEY, S. D. & PARKHILL, J. 2020. Producing polished prokaryotic pangenomes with the Panaroo pipeline. *Genome Biology,* 21**,** 180.

VASER, R., SOVIĆ, I., NAGARAJAN, N. & ŠIKIĆ, M. 2017. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res,* 27**,** 737-746.

WANG, L. G., LAM, T. T., XU, S., DAI, Z., ZHOU, L., FENG, T., GUO, P., DUNN, C. W., JONES, B. R., BRADLEY, T., ZHU, H., GUAN, Y., JIANG, Y. & YU, G. 2020. Treeio: An R Package for Phylogenetic Tree Input and Output with Richly Annotated and Associated Data. *Mol Biol Evol,* 37**,** 599-603.

WHELAN, F. J., RUSILOWICZ, M. & MCINERNEY, J. O. 2020. Coinfinder: detecting significant associations and dissociations in pangenomes. *Microb Genom,* 6.

WHO. 2023. *Global tuberculosis report 2023* [Online]. World Health Organization. Available: https://www.who.int/teams/global-tuberculosis-programme/tb-reports/global-tuberculosis-report-2023 [Accessed].

ZHANG, Z., SCHWARTZ, S., WAGNER, L. & MILLER, W. 2000. A greedy algorithm for aligning DNA sequences. *J Comput Biol,* 7**,** 203-14.

ZULKOWER, V. & ROSSER, S. 2020. DNA Features Viewer: a sequence annotation formatting and plotting library for Python. *Bioinformatics,* 36**,** 4350-4352.

ZWYER, M., ÇAVUSOGLU, C., GHIELMETTI, G., PACCIARINI, M. L., SCALTRITI, E., VAN SOOLINGEN, D., DÖTSCH, A., REINHARD, M., GAGNEUX, S. & BRITES, D. 2021. A new nomenclature for the livestock-associated Mycobacterium tuberculosis complex based on phylogenomics. *Open Res Eur,* 1**,** 100.

Figures

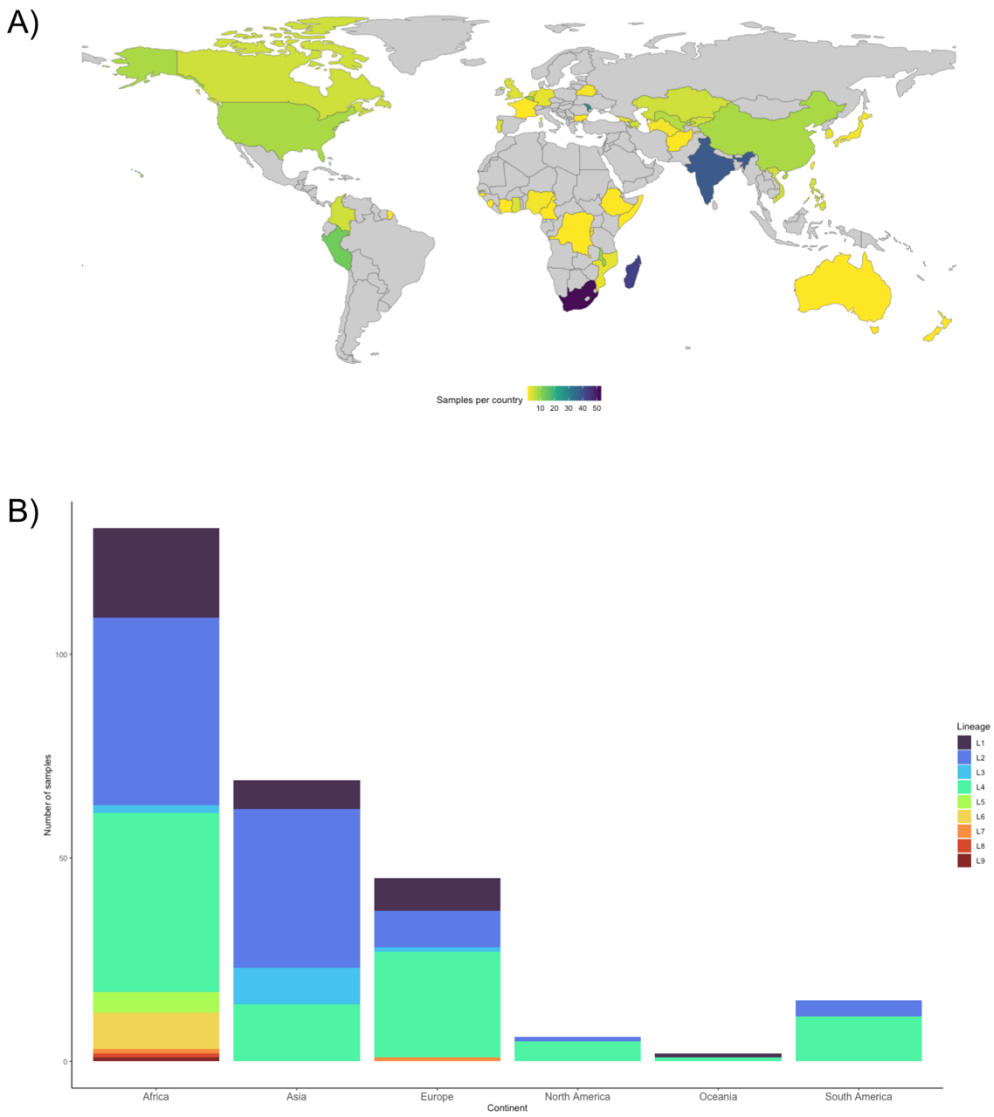
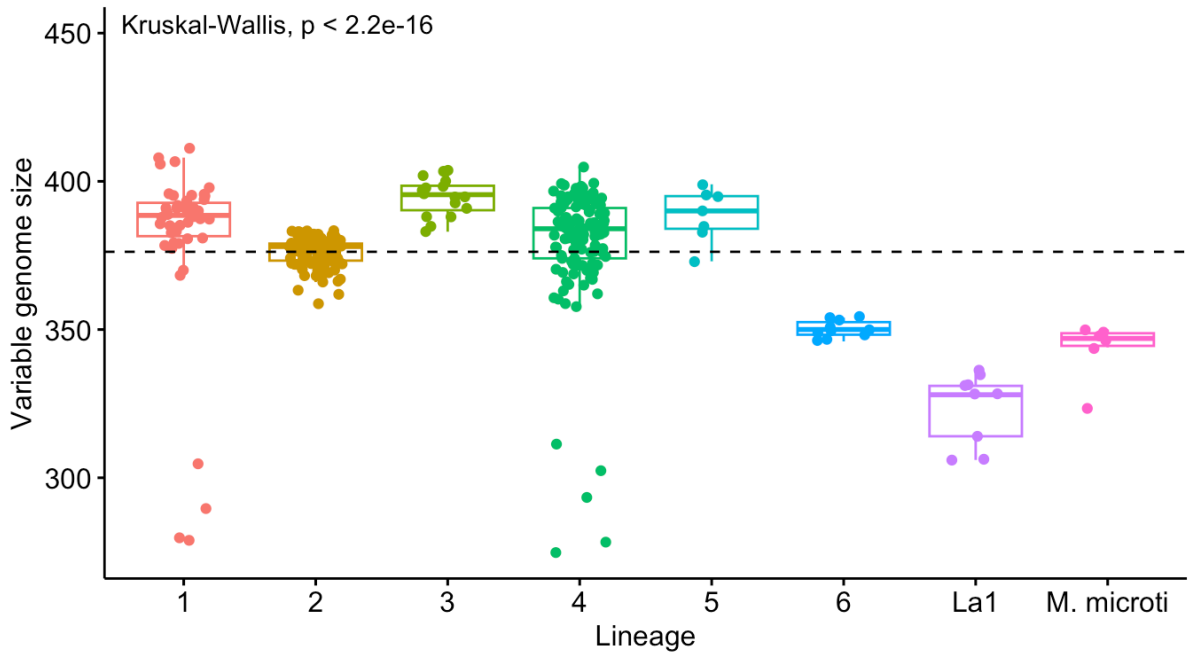


Figure 1: The geography of the dataset. (A) sample collection distribution by country; (B) the number of genomes of each lineage included from each continent.



640

641 **Figure 2: Boxplot showing the distribution of accessory genes within each lineage.** Lineages with a

642 small number of genomes were excluded from the analysis. L6, La1 and *M. microti* have significantly

643 smaller accessory genomes compared to other lineages.

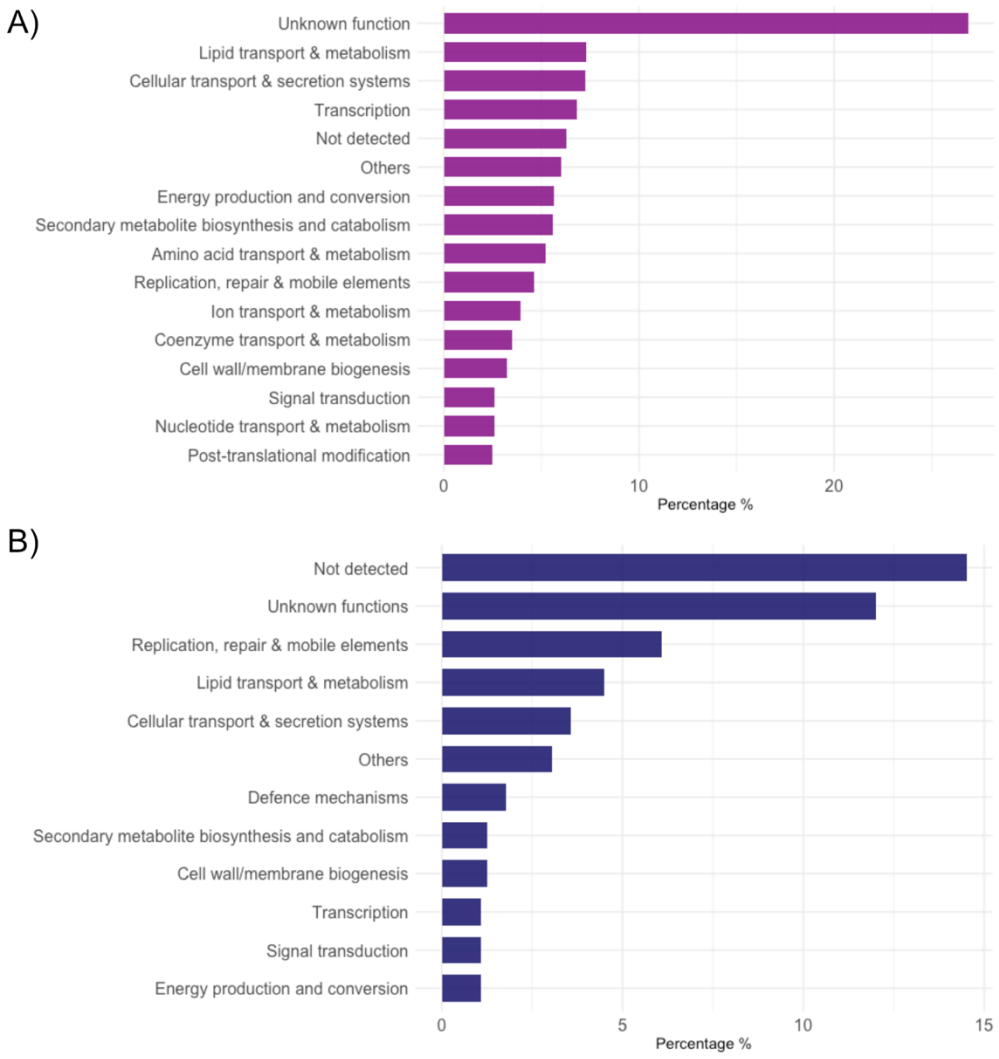
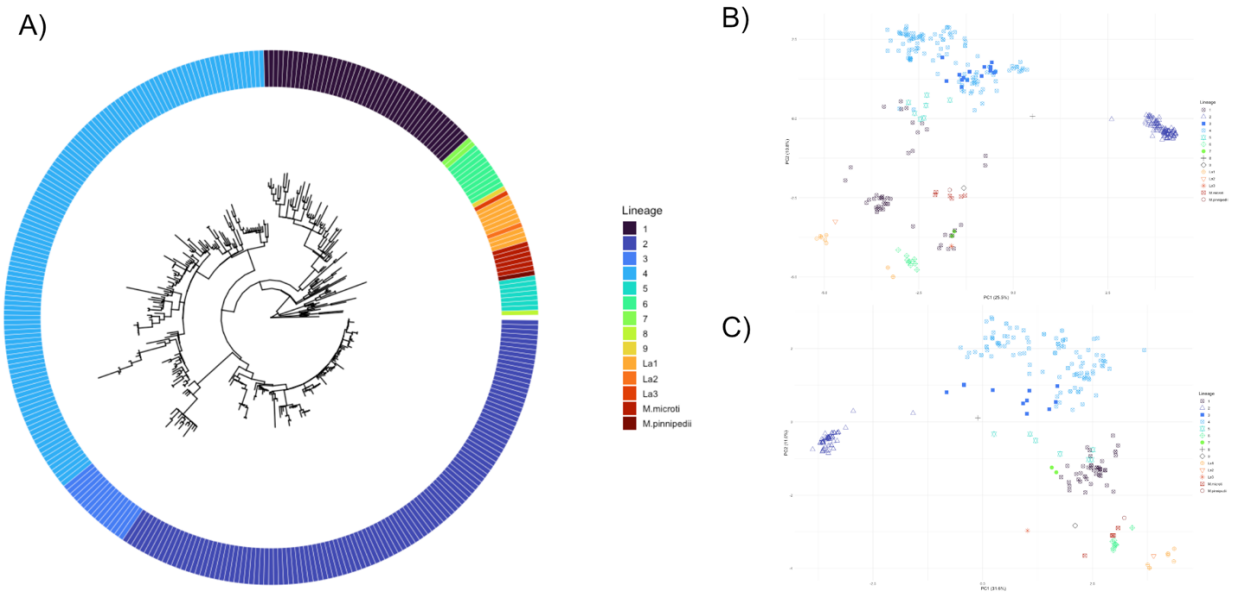


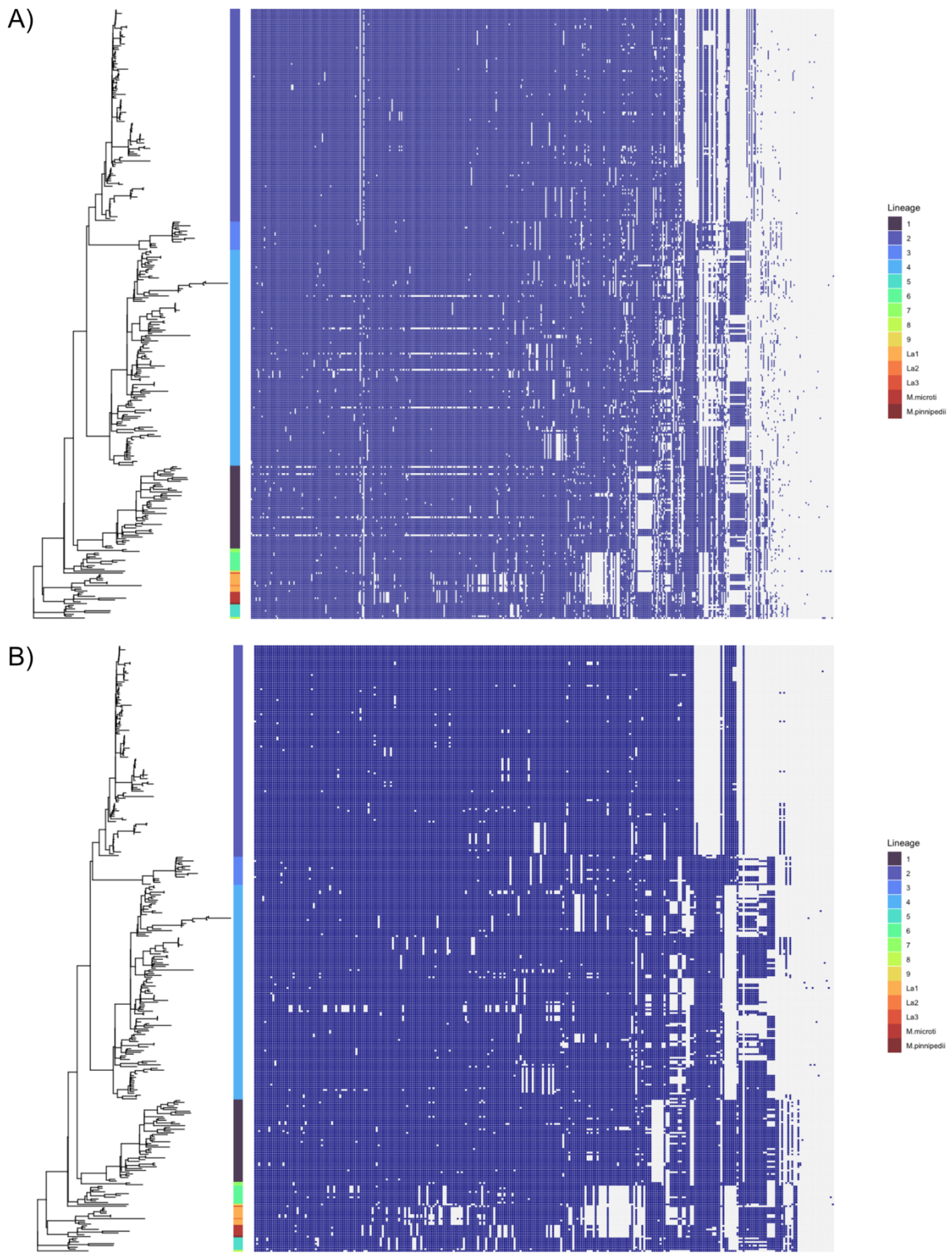
Figure 3: Analysis of the functional components in the core (A) and accessory (B) genome using EggNOG mapper and InterProScan.



647

648 **Figure 4: (A) phylogenetic tree based on MTBC core genome. PCA based on the accessory genome**

649 **data from Panaroo (B) and pangraph (C).**



650

651 **Figure 5: MTBC Accessory genome identified by Panaroo (A) and Pangraph (B).**

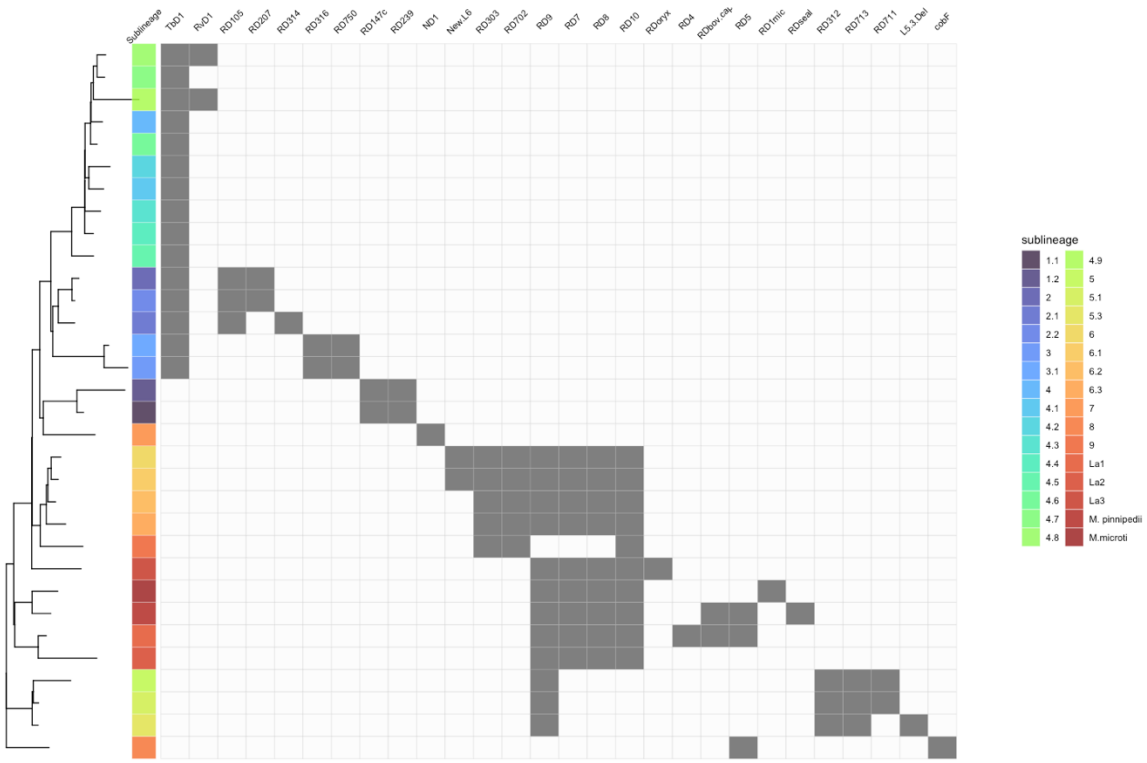
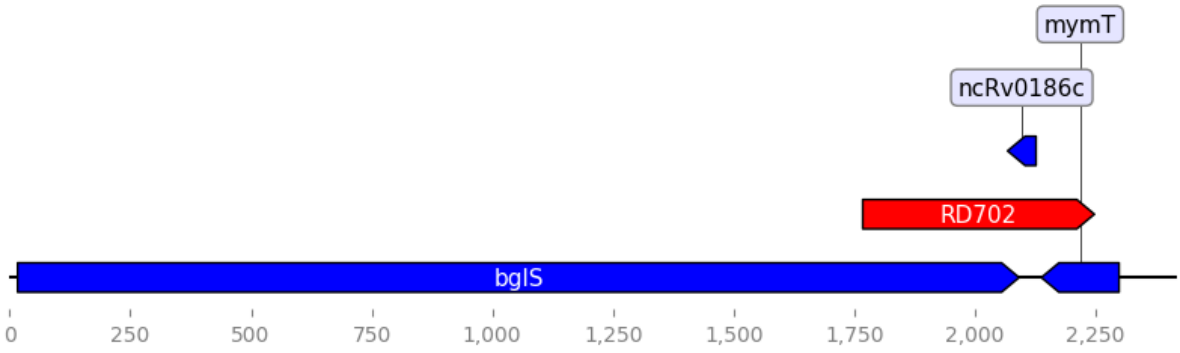


Figure 6: Sub-lineage specific regions of differences (RDs) identified using pangenome-based and genome alignment-based approaches. Sub-lineages are shown on the Y-axis coloured as per the legend and clustered based on the RD presence/absence patterns. RDs are listed on the X-axis, grouped by their pattern of presence/absence.



659 **Figure 7: Visual presentation of deletion caused by RD702.** The truncated *bglS* gene was classified as

660 a core gene with Panaroo but has specific deletions within, which can be problematic for

661 pangenomic analyses.

662