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

Mahboobeh Behruznia[1,2], Maximillian Marin[3], Daniel J. Whiley[1,4], Maha Farhat[3,5], Jonathan C Thomas[1], Maria Rosa Domingo-Sananes[1], Conor J Meehan[1,6*]

[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] Medical Technologies Innovation Facility, Nottingham Trent University, Nottingham, UK

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

[6] 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), 0000-0002-3972-5100 (DW), 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 deletions found only in certain lineages (regions of difference [RDs]), 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 full 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 divergent and convergent 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 underlines the significance of genomic variations in determining the pathogenic traits of different MTBC lineages.

Introduction

The *Mycobacterium tuberculosis* complex (MTBC) is the term used for *M. tuberculosis* and all its variants, many of which used to be separate species (Riojas et al., 2018). These bacteria are the primary cause of tuberculosis (TB) in humans and other animals. Worldwide, TB is the 13th leading cause of death and is once again the top infectious killer following global efforts to combat the COVID-19 pandemic (WHO, 2024). The MTBC comprises both human-associated lineages (L1 to L10) and animal-associated lineages (La1, La2, La3, *M. tuberculosis* var *microti* and *M. tuberculosis* var *pinnipedii*) (Coscolla et al., 2021, Zwyer et al., 2021, Guyeux et al., 2024). 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-L10 are exclusively found in Central and East Africa (Blouin et al., 2012, Ngabonziza et al., 2020, Coscolla et al., 2021, Guyeux et al., 2024). 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

58 and transmission potential, which can affect the spread of certain lineages and sub-lineages
59 (Coscolla, 2017, Gagneux et al., 2006). Understanding the genetic diversity within MTBC is
60 essential for gaining insights into how these bacteria adapt to different host populations and
61 environments, and identify traits that may be important for survival, virulence, and antibiotic
62 resistance.

63 MTBC is recognised for its clonal population structure and little genetic diversity between
64 different strains, caused by the absence of horizontal gene transfer and plasmids (Godfroid
65 et al., 2018, Chiner-Oms et al., 2022). Single nucleotide polymorphisms (SNPs) and large
66 sequence polymorphisms (LSPs) represent the source of MTBC genetic diversity (Gagneux,
67 2018). Within the MTBC, LSPs which are deletions in genomes relative to the reference
68 genome H37Rv are referred to as Regions of Difference (RDs) (Bespiatykh et al., 2021). These
69 RDs are known to differentiate MTBC lineages from each other (Gagneux, 2018), but within-
70 lineage diversity is thought to be mostly composed of SNPs (Napier et al., 2020). While some
71 RDs have been described to be differentially present within some lineages (Sanoussi et al.,
72 2021, Bespiatykh et al., 2021), these are thought to be rare and remain unexplored. RDs can
73 surpass 10 kb in length and encompass groups of genes with diverse functions, which can be
74 linked to factors like virulence and metabolic diversity (Bespiatykh et al., 2021, Bottai et al.,
75 2020). Additionally, the *IS6110* element has a notable impact on the genetic diversity of the
76 MTBC. Transposition of *IS6110* into various coding or regulatory regions, and genomic
77 deletions through homologous recombination between proximal copies, can also translate
78 into significant genotypic variation at strain level (Alonso et al., 2013, Antoine et al., 2021).
79 These mechanisms can all contribute to a varying repertoire of genetic features, much of
80 which remains underexplored.

81 The pangenome represents the set of all genes present in a species. It comprises both the
82 core genome, which consists of genes shared by all members of the species, and the
83 accessory genome, which includes genes present in some but not all members. The
84 accessory genome is often split into shell (present in most but not all genomes) and cloud
85 (present in only a few genomes). Pangenomes are further classified as either open (each
86 new genome adds a significant number of new genes to the population's accessory genome)
87 or closed (newly analysed genomes do not significantly increase the accessory genome size)
88 (Brockhurst et al., 2019).

89 Previous research has focused on pangenomic investigations of particular MTBC lineages
90 (Baena et al., 2023, Ceres et al., 2022, Sanoussi et al., 2021) and a recent study investigated
91 pangenome of most MTBC lineages using protein sequences (Silva-Pereira et al., 2024).
92 However, there is a gap in comprehensive pangenomic analyses encompassing both coding
93 and non-coding regions. Despite the name, prokaryotic pangenome analyses tend to not
94 focus on full genome comparisons, but rather on protein-coding sequences. This focus
95 neglects non-coding RNAs and pseudogenes as well as essential elements in intergenic
96 regions, such as promoters, terminators, and regulatory binding sites, despite their
97 demonstrated role in selection and phenotypic impact (Tonkin-Hill et al., 2023). Moreover,
98 the inference drawn from pangenomic analyses is often influenced by short-read datasets. It
99 has been shown that using short-read datasets can artificially inflate the size of the
100 pangenome estimates, leading to incorrect assumptions about evolutionary scenarios
101 (Marin, 2024). Here, we used a diverse collection of closed MTBC genomes including human
102 and animal-associated lineages to develop our understanding of MTBC pangenome and
103 evolution incorporating both coding and non-coding regions. The objectives of this study
104 are: (i) to establish the true size of MTBC pangenome using a highly curated dataset, (ii) to

105 assess specific biological functions that are prevalent in the core and accessory genome and
106 (iii) to assess gene content variation at both lineage and sub-lineage levels. By focusing on
107 sub-lineages, we aim to provide a detailed characterisation of genetic diversity and
108 structural differences that influence MTBC evolution and functional traits long term. We find
109 that the MTBC has a small, closed pangenome and genome reduction via RDs is the driving
110 evolutionary force that shapes the genomic repertoire of the species. We identified distinct
111 genomic features among lineages and sub-lineages, including sub-lineage specific RDs, that
112 may account for variations in virulence, metabolism, and antibiotic resistance.

113 Methods

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

115 Previous studies have shown that closed genomes should be used for the best estimation of
116 pangenome size (Marin et al., 2024). From a combination of previously published research
117 and publicly available data, 328 non-duplicated, complete (single contig) and closed (as
118 listed on NCBI) genome assemblies were selected for having (i) long-read sequencing only
119 (Nanopore or PacBio) or (ii) hybrid sequencing with Illumina across the MTBC lineages.
120 Information related to the geographical location, source of isolation, assembly accessions
121 and quality metrics can be found in Supplementary Table S1. The quality of all assemblies
122 was assessed using BUSCO v5.8.2 (Manni et al., 2021) and only assemblies with ≥ 95%
123 completeness were used for further analysis. For parity, all assemblies were set to start on
124 *dnaA* gene using the fixstart command of Circlator v1.5.5 (Hunt et al., 2015). Complete
125 circular assemblies were annotated using PGAP v6.4 with default options (Tatusova et al.,
126 2016). This approach is recommended by previous analyses (Marin et al., 2024) as PGAP
127 tends to annotate pseudogenes in cases of pseudogenisation and frameshift, rather than

128 annotating them as new coding sequences. Visualisations of the PGAP annotation were
129 made using the DNA Features Viewer Python library (Zulkower and Rosser, 2020). Lineage
130 classification was performed using TB-profiler v6.5 with the TBDB v.72ef6fa database (Phelan
131 et al., 2019) except for *M. tuberculosis* var *microti* (hereafter named *M. microti) and M.*
132 *tuberculosis* var *pinnipedii* (hereafter named *M. pinnipedii)* which are not classified in this
133 database and were assigned using spoligotype using SpolPred2 as implemented in TB-
134 profiler (Napier et al., 2023).

135 **Nanopore sequencing and assembly**

136 We sequenced lineages and sub-lineages that either had no sequencing data available in
137 public databases or were under-represented in the collection. These isolates belonged to L5
138 (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.*
139 *caprae,* n =1) and *M. pinnipedii* (n=1). A list of genomes sequenced in this study and their
140 assembly information can be found in Supplementary Table S1. Raw reads and assembled
141 genomes were deposited in NCBI under the Bioproject PRJNA1085078.

142 DNA extraction suitable for long-read sequencing was performed using the Genomic-tip
143 100/G (Qiagen Inc) as described previously (Ngabonziza et al., 2020). DNA concentration was
144 quantified using a Qubit fluorometer and 1X dsDNA High Sensitivity (HS) assay kit
145 (ThermoFisher Scientific). Nanopore sequencing was performed using the native barcoding
146 kit 24 V14 (SQK-NBD114.24) according to the manufacturer's instruction on the MinION
147 Mk1C platform with R10.4.1 flow cells. Basecalling was undertaken using Guppy v 6.5.7 with
148 the super accurate (SUP) model. Sequence adaptors were trimmed using Porechop v0.2.4
149 with “--end_threshold 95” and “--middle_threshold 85” options
150 (https://github.com/rrwick/Porechop.git). Sequences were assembled using flye v2.9.2 with

151 default options (Kolmogorov et al., 2019) and iteratively polished using Racon v1.5.0 and
152 the "consensus" subcommand of Medaka v1.8.0
153 (https://github.com/nanoporetech/medaka.git). Complete assemblies underwent Circlator,
154 BUSCO and PGAP steps as above.

155 **Pangenome analysis (protein coding genes)**

156 Panaroo v 1.5.1 inferred the MTBC protein-coding pangenome (Tonkin-Hill et al., 2020) with
157 and without the "--merge_paralogs" option to investigate gene duplication (paralogy).
158 Additionally, "--clean-mode strict" and "--core_threshold 0.95" were applied to reduce
159 potential contamination by removing genes found at low frequencies (present in fewer than
160 5% of genomes) and to define core alignment gene set for the subsequent phylogenetics,
161 respectively. Ceres et al. (2022) showed that genes could be over-split by annotation
162 software, creating additional accessory genes, which has been confirmed elsewhere (Marin
163 et al., 2024). To counteract this we followed the approach recommended in Ceres et al.
164 (2022) by merging genes that map to the same gene in a reference genome. The reference
165 pangenome created by Panaroo was compared to the H37Rv genome (standard reference
166 genome for the MTBC) using BLASTn with a similarity e-value cutoff of 1×10^{-10} (Zhang et al.,
167 2000). The BLASTn result was then filtered to only include queries with ≥90% identity and
168 ≥75% of the H37Rv gene length covered using in-house Python scripts
169 (https://github.com/conmeehan/pathophy). A core genome alignment was constructed
170 using the built-in MAFFT aligner included in Panaroo (Katoh et al., 2002). The core
171 phylogeny was created using RAxML-NG v1.2.2 with a GTR substitution model (Kozlov et al.,
172 2019). Tree visualisations were created using the Treeio package v1.20.0 (Wang et al., 2020).

173 **Pangenome graph construction**

174 The above pangenome analysis only compares coding gene content and omits all non-coding
175 genes and intra-gene genetic content. To examine the total pangenome, including both
176 coding and non-coding regions, all genomes were aligned into a graph using Pangraph v0.7.2
177 (Noll et al., 2023). A minimum genome region block size of 250bp was selected to look only
178 at large polymorphisms. As above, an additional check was undertaken on blocks labelled as
179 absent, to ensure they truly are not present in the given genome, and not just mis-labelled
180 as absent by the Pangraph pipeline. BLASTn was used to compare the consensus sequence
181 for that region and the genome it is predicted to be absent from. Any BLASTn results with an
182 e-value less than $1e^{-30}$ and additionally covering 50% of the region length and a 50%
183 identical match were re-labelled as present. Genomic regions were classified as a core region
184 if present in all genomes or an accessory region if present in some but not all genomes. The
185 DNA sequences from genomic blocks present in at least one sub-lineage but completely
186 absent in others were extracted to look for long-term evolution patterns in the pangenome.

187 Known Regions of Differences (RDs) were identified using RDscan (Bespiatykh et al., 2021).
188 Structural variants with Rv (H37Rv gene) loci overlapping previously annotated RDs were
189 assigned the same RD names. Novel structural variants that did not align with any previously
190 reported RDs were classified as new RDs if they are present in H37Rv, or as a new LSP if
191 absent. Accessory regions were further classified as a deletion if present in over 50% of the
192 sub-lineages or an insertion/duplication if present in less than 50% of sub-lineages. This
193 approach provides a reference-free perspective for investigating gene content differences
194 and labels relevant LSPs as reductive (deletions) or increasing (insertions/duplications).

195 **Functional genome analysis**

196 The reference pangenome created by Panaroo was translated using transeq v5.0.0 (Rice et
197 al., 2000) to obtain the protein sequences. COG functional categories were assigned using
198 eggNOG Mapper v5.0.2 using default options (Huerta-Cepas et al., 2017). Genes that did not
199 return a match within the eggNOG database were included in the "not found" category.
200 Functional domain annotation was performed with InterProScan v5.63-95 (Jones et al.,
201 2014). For the gene association analysis, the accessory genome (genes present in less than
202 95% of genomes) was used as input for Coinfinder v1.2.1 (Whelan et al., 2020). The STRING
203 database (https://string-db.org/) was used to infer interaction networks of genes. Predicted
204 functional partners with a high confidence interaction score (≥0.7) by STRING v12.0 were
205 considered in this study.

206 **Statistical analysis**

207 Statistical analysis was conducted using R studio v2024.04.2+764 (R Core Team, 2023). Due
208 to the use of multiple public genomes created using different sequencing technologies, we
209 looked for significant associations between accessory genome size and sequencing
210 technology. For comparative purposes and to ensure sufficient sample size per group, all
211 types of Illumina sequencing (e.g., MiSeq, HiSeq, etc) were combined, resulting in
212 comparisons between different long-read technologies, with or without Illumina hybrid
213 sequencing. Any sequencing technology group with fewer than three genomes was
214 excluded. The Shapiro-Wilk test was used to test for normality and then a Kruskal-Wallis test
215 was used to look for associations between accessory genome count (the number of missing
216 genes in a genome) and sequencing platform(s). A Dunn's test was then used to look for
217 pairwise significant differences with a Bonferroni correction for multiple comparisons. The

218 same process was undertaken on pseudogene counts, based on the GFF files output from
219 PGAP.

220 Principal component analysis (PCA) based on gene presence and absence data from
221 Panaroo-derived accessory genome contents and Pangraph-derived accessory region
222 contents was performed using the prcomp package in R. The fviz_contrib function in
223 Factoextra (v1.0.7) was used to identify the contribution of important variables from the PCA
224 results (Kassambara and Mundt, 2020).

225 To assess the openness and pairwise uniqueness of the pangenome, micropan was used to
226 estimate Heaps law, genome fluidity and rarefaction curves with 100 permutations (Snipen
227 and Liland, 2015). An alpha of 1 or lower in Heap's law indicates an open genome; higher
228 than 1 indicates a closed genome.

229

230 Results

231 **Overview of the genomic dataset**

232 We combined 339 whole-genome sequences covering all known MTBC lineages except L10
233 for which no closed genome is available (Guyeux et al., 2024). This included 328 genomes
234 published previously, as well as 11 genomes sequenced here from sub-lineages with few or
235 no representatives in the public dataset to better cover the known phylogeny of MTBC
236 (Supplementary Table S1). These include the first closed genomes for L5.2, L9, La2 (formerly
237 *M. caprae*) and *M. pinnipedii*. All genomes were sequenced to a depth greater than 30x
238 coverage, as required for >99% sequence recovery (Sanderson et al., 2024).

239 Except for Antarctica, our collection includes genomes from every continent, and different
240 host species including humans, cattle, voles and BCG vaccine strains (Supplementary Figure
241 S1-A). Africa has the highest diversity of MTBC genomes of all continents, which is consistent
242 with the fact that the continent is known to have the most diverse MTBC strains
243 (Supplementary Figure S1-B) (Gagneux, 2018). Most L1 genomes have been collected from
244 East Africa, South and Southeast Asia, where this lineage is geographically common. L3
245 genomes were mostly isolated from India, while L2 and L4 genomes have been collected
246 from different continents, representing the global distribution of these lineages. L5 and L6
247 genomes were exclusively found in West Africa, where they contribute significantly to the
248 overall burden of TB in this region (Coscolla et al., 2021).

249 Across the total collection, genomes contained on average 3,950 protein-coding genes (CDS)
250 (range 3,579-4,018), of which 76 were predicted to be pseudogenes (range 64-93)
251 (Supplementary Table 1). No significant association was found between the sequencing
252 platform and pangenome size (Kruskal-Wallis test: p-value < 0.001). Pairwise comparisons

253 also found no significant differences in pangenome size (Dunn's test: p-value < 0.001).
254 However, a less stringent threshold (p-value < 0.05) indicated a difference in pangenome size
255 between Nanopore/Illumina hybrid assemblies and PacBio SMRT (no hybrid assembly), with
256 Nanopore hybrid assemblies showing lower accessory gene counts. No difference in
257 pseudogene count was observed based on sequencing technology, suggesting this was not
258 impacted either.

259 **Pangenome characteristics and functional diversity**

260 Panaroo estimated a protein-coding pangenome (with merged paralogs) consisting of 4,118
261 genes, including 3,638 core genes (≥99% of genomes) and 480 accessory genes including
262 180 soft core (≥95% and <99% of genomes), 182 shell (≥15% and <95% of genomes) and 118
263 cloud (<15% of genomes) genes. After mapping all the genes in the pangenome to H37Rv to
264 look for over-splitting of annotated genes, we identified 86 genes (2% of total genes) that
265 were misclassified as accessory genes, including 9 out of 180 soft core, 12 out of 182 shell
266 and 65 out of 118 cloud genes (Table 1 and Supplementary Table S2). This approach reduced
267 the pangenome size from 4,118 to 4,032 genes and the accessory genome from 480 to 394
268 genes. Incorporating duplicated genes (i.e. not merging paralogs in Panaroo) increased the
269 pangenome size to 4,427 genes, including 3,635 core and 792 accessory genes (195 soft
270 core, 200 shell and 397 cloud). After combing genes based on mapping to H37Rv to reduce
271 annotation splitting bias, the (including paralogs) pangenome size decreased to 4,321 genes,
272 including 3,627 core genes and the accessory genome was reduced to 694 (187 soft core,
273 186 shell and 321 cloud) (Table 1 and Supplementary Table S2).

274 Heap's Law analysis resulted in an alpha of 2 for the merged paralog dataset and 1.548 for
275 the unmerged dataset. This indicates a closed or near closed pangenome based on both

276 datasets. Genome fluidity was estimated to be 0.0085 for the merged paralog dataset and
277 0.0088 for the unmerged paralog dataset. This indicates any two randomly selected
278 genomes only differ by 0.85-0.88% of gene content. (Supplementary Figure S2).

279 All subsequent analyses based on Panaroo was undertaken using the merged paralog
280 dataset, unless otherwise stated.

Method	Pangenome statistics	Raw estimates	Corrected estimates
Panaroo Merged paralogs	Total Genes	4,118	4,032
	Core genes	3,638	3,627
	Accessory genes:	480	394
Panaroo Including paralogs	Total Genes	4,427	4,321
	Core genes:	3,635	3,627
	Accessory genes:	792	694

281 **Table 1: Pangenome (Panaroo) size estimations with and without corrections and merged**
282 **paralogs.** Corrected estimates are derived from filtering out false accessory genes through
283 merging of genes with strong hits to the same gene in H37Rv (≥90% identity and ≥75% gene
284 length coverage).

285 To examine genomic diversity among lineages, we investigated variations in the size of the
286 accessory genome among different lineages. This analysis revealed significant differences in
287 accessory genome size among different lineages (Kruskal-Wallis test, p= 2.2e-16). Post-hoc
288 Dunn tests with Bonferroni correction identified significant pairwise differences between
289 lineages (Supplementary Table S3). L6, La1 and *M. microti* have significantly smaller number
290 of accessory genes per strain compared to others (Figure 1, Supplementary Table S3); L2 has
291 a significantly smaller accessory genome in comparison with L1, L3, and L4.

292 Functional annotation of core and accessory genes showed that genes of unknown function
293 were the largest functional group in both, with over 20% (Figure 2). The core genome of

294 MTBC encodes a high proportion of genes involved in lipid transport and metabolism
295 (category I), transcription (category K) and cellular transport and secretion systems (category
296 N), indicating the importance of these biological functions in defining the species. Functional
297 breakdown of the accessory genome also revealed variations in genes involved in DNA
298 replication, repair and mobile elements (category L) and lipid transport and metabolism
299 (category I) across the species. The functional composition of core and accessory genes with
300 the inclusion of paralogs is shown in Supplementary Figure S3.

301 Pangraph revealed a similar distribution of core and accessory regions (Table 2). In total,
302 1,338 regions were identified, with 1,015 (75.8%) present in all genomes. The accessory
303 regions were distributed as follows: 129 (9.6%) in the soft core, 140 (10.4%) in the shell and
304 54 (4%) in the cloud. Regions ranged in length from 250bp (the set lower limit) to 54.6kb
305 (mean: 3.2kb, median: 1kb). Some regions (2.3% of total pangenome regions) were found to
306 be mislabelled as absent in certain genomes, which were corrected using BLAST. This
307 resulted in small corrections to the pangenome category distributions, primarily with an
308 increase to the core region and reduction in the cloud accessory region count (Table 2),
309 similar to the pangenome corrections above. Blocks that had these spurious hits discovered
310 by BLASTn ranged in size from 254bp to 5.4kb (mean: 815bp, median:584bp), suggesting this
311 is an issue mainly impacting smaller Pangraph blocks.

312 After corrections, accessory blocks made up ~23% of the pangenome. These accessory
313 blocks ranged in size from 250bp to 9.8kb (mean: 993bp; median 629bp). This suggests most
314 accessory blocks each encompass only one or a few genes each, with most larger regions
315 (i.e. all above 10kb) being common to all the MTBC strains.

316 Heaps law based on the Pangraph block distribution resulted in an alpha of 2 and a genome
317 fluidity of 0.041 (i.e. 4.1% uniqueness between any two genomes). This indicates a closed
318 genome, similar to the Panaroo-based analyses, but with more unique blocks per genome
319 when both coding and non-coding regions are considered compared to coding sequences
320 alone (Supplementary Figure 2).

Pangenome partition	Raw estimates	Corrected estimates
Total Regions	1,338	1,338
Core Regions	1,015	1,040
Soft core	129	124
Shell	140	143
Cloud	54	31

321 Table 2: **Pangenome graph (Pangraph) size estimations with and without corrections**.
322 Corrected estimates are derived from filtering out false labelling of absent regions using
323 BLASTn.

324 **Population structure analysis**

325 The population structure was assessed using SNPs in the core genome (Figure 3A) and PCA
326 plots were constructed based on presence-absence patterns in the accessory genome
327 determined with Panaroo (includes coding genes only) and the accessory regions derived
328 from Pangraph (includes both coding and non-coding genomic regions) (Figure 3B-D;
329 Supplementary figure S4). In the PCAs, genomes group together based on their lineage,
330 indicating variations in accessory genome are mainly structured by lineage (Figure 3B-D).
331 The first principal component (PC1) highlights the contrast between L2 from other members
332 of MTBC. In all PCA plots, L2 formed a distinct group, defined primarily by the important
333 variables RD105 and RD207 deletions. RD105 serves as a classic marker for all L2 genomes
334 and RD207 deletion is found in all L2.2 sub-lineages; the single L2.1 genome is separated

335 from the L2.2 cluster due to the absence of RD207. The second principal component (PC2)
336 demonstrates the difference between modern human-associated lineages (L2, L3 and L4)
337 from other lineages. Important variables driving this separation are the TbD1 gene deletion
338 in modern lineages, and gene deletions associated with RD11 (prophage phiRv2) in L6, L7,
339 La1 and La3 genomes and some L1 sub-lineages. L1 genomes exhibit considerable genetic
340 diversity primarily due to RD3 (related to prophage phiRv1) and RD11. The latter was
341 identified in all L1.1 sub-lineages and certain L1.2 sub-lineages (L1.2.1, L1.2.2, and L1.2.2.2).
342 The positioning of the single L8 genome, (positive sign, Figure 3B-D) sits separate from all
343 other lineages, likely due to being the most basal known lineage (Ngabonziza et al., 2020).

344 **The accessory genome of MTBC sub-lineages**

345 To counteract the potentially misleading effect of mis-annotations of some genomes on
346 accessory genome analyses, we looked for larger patterns of genomic differences in the
347 MTBC, shared by multiple genomes. The accessory genome (Panaroo) and accessory region
348 (Pangraph) patterns outlined above were filtered to look only at those genes and larger
349 genomic regions present in all strains of one or more sub-lineages and completely absent in
350 others (i.e. sub-lineage-specific genes and LSPs).

351 Among the 394 filtered accessory genes, 111 were identified as specific to certain lineages
352 and sub-lineages (Supplementary Figure S5-A, Supplementary Table S4). Four of these genes
353 were absent from La3 but present in at least two copies in all other lineages: Rv3117 (cysA3)
354 and Rv0815c (cysA2), involved in the formation of thiosulfate, and Rv3118 (sseC1) and
355 Rv0814c (sseC2), involved in sulphur metabolism. Panaroo failed to identify partial gene
356 deletions caused by RD9, RD10, RD13, RD105, RD207, and RD702. For example, RD702 is
357 present (i.e. deleted) in all L6 and L9 genomes and covers a 641 bp region (Figure 4). RD702

358 affects less than a quarter of the *bglS* gene (470 bp from 2,076 bp) and *mymT* gene (126 bp
359 from 162 bp gene) in the 5' to 3' direction as well as a regulatory RNA (ncRv0186c) in the 3'
360 to 5' direction. The Pangraph approach identified most partial gene deletion and non-coding
361 regions of the DNA that were impacted by genomic deletion (relative to H37Rv). However,
362 the graph approach failed to identify four structural variants, including a gene deletion
363 corresponding to RD8 (Rv3618) in L6 and the animal-associated genomes.

364 Gene association analysis using coinfinder confirmed that most accessory genes (as
365 determined by Panaroo) co-occur in pairs or groups in larger multi-gene RDs (as determined
366 by Pangraph) (Supplementary Figure S5). Thus, the small accessory genome is contained
367 within an even smaller number of RDs. Coincident genomic blocks that were associated with
368 lineages were RD8, RD9, RD10, and RD178, whereas those that were distributed across the
369 phylogeny belonged to RD3 and RD11.

370 The filtered Pangraph results identified 82 lineage and sub-lineage-specific genomic regions
371 ranging from 270 bp to 9.8kb, encompassing a total of 116 genes and 4 non-coding RNA
372 (Supplementary Table S4). Many known lineage and sub-lineage-specific genomic regions
373 were consistently identified using both methods (Panaroo and Pangraph) and found to be
374 differentially present in two or more genomes, including TbD1, RD9, RD10, RD13, RD105,
375 RD303, RD312, RD316, RD702, RD711, RD713, RD750, ND1, RDoryx, and N-RD25bov/cap
376 (Figure 4).

377 Sub-lineage-specific accessory regions were classified as either RDs (deletions relative to
378 H37Rv; present in most sub-lineages but absent in less than 50%) or LSPs
379 (insertions/duplications; present in less than 50%). All sub-lineage-specific accessory regions
380 were found to be RDs, i.e. present in H37Rv (L4.9 in Figure 4) and absent in only a minority

381 of sub-lineages. Additional analysis of paralogous blocks, those present in all genomes but in
382 multiple copies in some, did not reveal any new sub-lineage-specific accessory regions.
383 Overall, these findings point towards deletion-driven evolution of the MTBC.

384 Discussion

385 **The MTBC accessory genome is small and prone to inflation due to mis-annotation**

386 The MTBC pangenome size has been reported to vary across studies. In a recent
387 investigation conducted by Marin (2024), the accessory genome ranged from 314 to 2951
388 genes as a function of the assembly quality, software and specific parameters used. Genome
389 sequencing using short-read technology has some limitations that can result in incomplete
390 assemblies, particularly in regions with repetitive sequences that are difficult to sequence
391 (Meehan et al., 2019). About 10% of the coding sequences of the MTBC genome are
392 composed of repetitive polymorphic sequences known as the PE (Proline-Glutamic acid) and
393 PPE (Proline-Proline-Glutamic acid) families, implicated in host-pathogen interactions and
394 virulence (Cole et al., 1998). Incomplete assemblies can impact the accurate detection of
395 genes and may lead to artificial inflation of the estimated pangenome size (Ceres et al.,
396 2022, Marin et al., 2024). We used a curated dataset to address these issues and estimate
397 the true size of the MTBC pangenome. We showed that the MTBC has a compact accessory
398 genome, totalling 10% of the total pangenome size (16% when including paralogs). This
399 finding is consistent with recent research, which found a small accessory genome with a
400 clonal genetic structure in the MTBC (Baena et al., 2023, Ceres et al., 2022, Silva-Pereira et
401 al., 2024). Using a BLASTn analysis approach, we identified core genes that were designated
402 as accessory and further reduced the accessory genome from 480 to 394 genes, a reduction
403 of almost 20%. For example, two genes labelled as *lppA* and group_1645, both annotated as

404 lipoproteins, were initially classified as accessory genes. However, these two genes exhibited
405 98% query coverage and 94% identity to the same gene in H37Rv, leading to their
406 classification as a single core gene. Notably, 50% of the cloud genes (genes present in less
407 than 15% of strains) were found to be core genes after this correction. These genes mapped
408 to the reference genes using the BLAST approach and appeared to be fragmented genes
409 caused by over-splitting in the PGAP annotation step.

410 The graph-based approach employed by Pangraph, encompassing all coding and non-coding
411 regions, found a slightly larger accessory genomic content, with 22% of the total pangenome
412 lying within accessory regions, after correction for spurious hits using BLASTn. This indicates
413 that there is more variation to be uncovered when both coding and non-coding portions of
414 the genome are assessed. However, correction of the Pangraph present/absent region
415 patterns was required, showing caution and additional scrutiny is needed when assessing
416 accessory region counts, especially for smaller blocks.

417 Identification of RDs was also likely impacted by the pangenome construction pipelines.
418 Many partial RDs known to occur in specific lineages and sub-lineages were not identified by
419 Panaroo. These results could be explained by the fact that Panaroo uses a relaxed alignment
420 threshold along with neighbourhood gene information to cluster gene families (Tonkin-Hill et
421 al., 2020). This can lead to the grouping of partially deleted genes with those that are
422 complete. Pangraph appeared to overcome many of these issues as it uses the whole
423 genome information as opposed to coding genes only. Despite these smaller limitations,
424 both methods yielded consistent results indicating that the accessory genomic content
425 (including coding and non-coding regions) of the MTBC is small, with much of the accessory
426 genome contents shared between multiple strains (i.e. in the soft core and shell

427 compartments). However, this does indicate that moving towards analysis of all genome
428 regions, such as pangenome graphs, is required to identify the true accessory regions of
429 difference within a population, especially if that population has a near closed pangenome.

430 **The MTBC accessory genomic diversity is driven by reductive evolution**

431 We employed a pangenomic approach to analyse the genetic diversity within the MTBC,
432 aiming to understand its evolution and functional potential. In addition to inferring a gene-
433 based pangenome with Panaroo, we also analysed variation in non-coding regions by
434 inferring a pangenome graph with Pangraph. Examining non-coding regions is particularly
435 important for species with limited genetic diversity, as they play a vital role in controlling
436 gene expression, stress response and host adaptation (Arnvig et al., 2014). Our study
437 demonstrated that, although both methods produced reliable outcomes, the whole-genome
438 alignment approach is needed to capture all genetic variations, especially those involving
439 large structural variants (>200 bp), as much variation appeared to be outside of protein
440 coding genes. Moreover, this method identified non-coding regions of the genome that were
441 affected by genomic variation. Results from both methodologies revealed that the accessory
442 genomic content of MTBC is a product of reductive evolution, that can be classified into
443 divergent deletions arising from a single event (RDs) and convergent deletions, where the
444 same deletion arises multiple times from independent events, with the former category
445 being the more common. Coinfinder results also showed that coincident genes in the
446 accessory genome primarily belonged to particular lineages.

447 However, a limitation of this analysis is that syntenic changes (those at the same places in
448 different genomes) were not accounted for. Therefore, insertions or deletions which do not
449 radically change the overall copy number but change the genome architecture could still

occur and contribute to the overall evolutionary trajectory of the species. A more in-depth structural variation analysis accounting for such potential rearrangements is needed to assess the full impact of large insertions and deletions on the *M. tuberculosis* pangenome.

Regions of difference (RDs), which are large deletions often encompassing multiple genes, can act as distinctive markers not only for lineages, as demonstrated previously, but also for sub-lineages. These warrant further investigation to understand the biological consequences of such deletions on traits like virulence, drug resistance, and metabolism, along with SNP only analyses. On the other hand, we observed several genomic deletions that appear to be due to convergent evolution and may indicate gene instability (Brosch et al., 2002). This includes RD3 and RD11 corresponding to the two prophages phiRv1 and phiRv2. One drawback of our study was that many of the lineages were poorly represented, especially in the L5, L6, L9 and all animal-associated lineages. The production of more closed genomes from the geographic regions where these are abundant (e.g. East and West Africa) will provide further information about the RD distributions within the MTBC and better understanding of the driving evolutionary forces.

All analysis undertaken here points towards genomic deletion as the main evolutionary factor creating the non-SNP-based variation within the MTBC. Reductive evolution is a driving factor of the other two *Mycobacterium* obligate pathogens, *M. ulcerans* and *M. leprae* (Stinear et al., 2007). The *M. tuberculosis* genome has also been suggested to be a product of genome reduction as its genome is much smaller than its sister species' (Orgeur et al., 2024). Our findings fit within this context, indicating that within-MTBC long-term evolution also tends to be reductive, with all discovered (sub-)lineage RDs being deletions relative to the complex as a whole. Much of this reduction appears to occur within the

473 animal-associated lineages and the closely related human-associated lineages L5/L6/L9
474 (figure 4). This is not to say duplication is not also an important factor in driving evolution
475 within this group. It has previously been shown that duplication of genes implicated in drug
476 resistance and virulence genes (e.g. PE/PPE genes) are common within the accessory
477 genome, albeit in a lineage-independent manner (Espinoza et al., 2025). While accounting
478 for duplicated genes did increase the accessory genome size here, especially the cloud
479 genome (Table 1), this did not translate into significantly different patterns between sub-
480 lineages. More closed genomes of closely related strains are required to look for duplication-
481 driven evolution in shorter timescales, especially if this is associated with changes in
482 virulence.

483 **The virulence potential of sub-lineages is differentially impacted by their accessory**
484 **genomic content**

485 The size of the accessory genome content varied significantly among different lineages. L2
486 genomes exhibited a high degree of clonality and a smaller accessory genome. L2 strains are
487 characterised by a large number of *IS6110* copies, possibly due to the RD207, which affects
488 the CRISPR-Cas type III-A system, responsible for adaptive immunity against mobile genetic
489 elements (Alonso et al., 2013). While RD207 encompasses seven genes (Rv2814c-Rv2820c)
490 (Bespiatykh et al., 2021), our analysis revealed that only two genes, Rv2819c and Rv2820c,
491 are specifically absent in L2 and can serve as a genetic marker for this lineage. Other RDs
492 affecting all L2 genomes are RD3 (9.2 kb), and RD105 (3.4 kb). The latter is exclusively
493 present in all L2 genomes and previous research proposed a potential role of RD105 in
494 conferring drug resistance in L2.2 strains (Qin et al., 2019). RD207 (7.4 kb), caused by the
495 recombination of two *IS6110* elements, is also present in all L2 genomes except for L2.1. L3

496 genomes exhibited significantly less genomic deletion in comparison with L2, and RDs
497 affecting all L3 genomes were RD316 (1.2 kb) and RD750 (765 bp). The genes affected by
498 these RDs are involved in lipid metabolism (category I) and secondary metabolite
499 biosynthesis and transport respectively (category Q), which may contribute to metabolic
500 variations in L3 strains. L4 exhibits substantial genetic diversity primarily due to variations in
501 the presence and absence of specific RDs at the sub-lineage level. For example, RD145 (4.7
502 kb) and RD182 (6.5 kb) were detected in L4.1.2.1 genomes isolated from Africa, Asia, Europe
503 and South America. However, none of the affected genes appears advantageous or essential
504 for MTBC growth and survival and this sub-lineage is successful worldwide
505 (https://tbdr.lshtm.ac.uk/sra/browse). Other L4 sub-lineage-specific RDs detected in this
506 study with no known advantage are RD115 in L4.3.4, RD727 in L4.6.1, RD219 in L4.8, RvD1 in
507 L4.8 and L4.9. RD4 in La1, L5.3 Del in L5.3, and RD750 in L3, RD147c in L1.

508 Nonetheless, some genomic deletions can affect the genes which confer growth advantages
509 in various lineages. Disruption of these accessory genes has been shown to provide a fitness
510 advantage for the *in vitro* growth of H37Rv by analysis of saturated Himar 1 transposon
511 libraries (DeJesus et al., 2017). For example, RD174 specific to L4.3.4 intersects with RD743
512 specific to L5, and both deletions affect two genes (Rv1995 and Rv1996) including *cmtR*
513 transcriptional regulator. The latter regulates bacterial stress response to metal deficiency
514 and is important during persistence in phagocyte (Flores-Valdez et al., 2020). While the
515 absence of these genes may not be necessary for *in vitro* growth (DeJesus et al., 2017) the
516 impact of this deletion may be different in *in vivo* conditions. The low prevalence of L4.3.4
517 strains in the TB Profiler database may suggest either the lower fitness of this sub-lineage, or
518 its presence in under-sampled geographic areas. RDs impacting the growth-advantage gene
519 group were prevalent in L5 genomes (L5.3 Del, RD743 and RD711). For instance, RD711

520 affects the Rv1335 gene, which is annotated as a sulphur carrier protein involved in cysteine
521 biosynthesis by PGAP. This may imply that the genome reduction observed in L5 has led to
522 dependance on host factors for essential nutrients, consequently making these strains
523 difficult to grow in a culture medium (Sanoussi et al., 2017).

524 L6 and animal-associated lineages accumulated a higher number of deletions including RD7,
525 RD8, RD9 and RD10. Comparative analysis of the accessory genome size in L6, La1, and *M.*
526 *microti* revealed significant differences compared to other lineages. The identified genomic
527 regions in L6, L9, and animal-associated lineages impact genes associated with virulence,
528 potentially explaining their reduced virulence in human hosts. RD7 contains the mammalian
529 cell entry (*mce3*) operon implicated in the invasion of host cells, while RD9 includes genes
530 related to biliverdin reduction, potentially aiding in protection against oxidative stress inside
531 macrophages (Szklarczyk et al., 2023). RD10 impacts genes related to lipid storage crucial for
532 MTBC transition to a dormant state (Deb et al., 2009). RD1BCG and RD1mic affect genes
533 encoding the ESX-1 secretion system, crucial for virulence and phagosomal escape, with
534 RD1BCG being deleted during BCG vaccine development (Szklarczyk et al., 2023). Overall,
535 the findings of this study suggest that, despite having a compact accessory genome, it can
536 account for certain biological distinctions among MTBC. While genetic drift has been
537 proposed as a key driver of diversity in MTBC (Hershberg et al., 2008), recent research has
538 argued that it does not fully explain the observed low genetic diversity in MTBC. Instead,
539 purifying selection may be the dominant force, removing harmful mutations and maintaining
540 genetic uniformity (Stritt and Gagneux, 2023).

541 **A need for more high-quality genomes and annotations**

542 While this dataset represents high-quality MTBC genome sequences that cover most of its
543 known diversity, it suffers from sampling bias as some lineages are over-represented
544 whereas others have a few representatives. The over-representation of L2 and L4 in the
545 dataset is primarily attributed to their association with Bioprojects analysing TB outbreaks or
546 originating from countries with extensive sampling, highlighting a bias in sequencing efforts
547 towards more prevalent or historically studied lineages. Many sub-Saharan African countries
548 do not have a representative sample in our dataset, and this may explain the under-
549 representation of L5 and L6 in the dataset. These lineages contribute significantly to the
550 overall burden of TB in West Africa (Coscolla et al., 2021). Additionally, we found that certain
551 RDs that were previously assumed to be specific to certain lineages were identified in other
552 lineages (Supplementary table S4). The observed variation in RD distribution emphasizes the
553 need for an expanded genomic dataset to provide a more comprehensive understanding of
554 their prevalence, specificity, and implications in the evolution of the MTBC. Future
555 sequencing efforts should target high-burden TB countries, ensuring a comprehensive
556 representation of the population structure to enhance our understanding of MTBC diversity
557 and evolution. Closed genomes for closely related strains are also required to better
558 understand short-term evolutionary patterns and the impact upon virulence within
559 circulating strains. Additionally, investigating human transcriptomic data could further
560 improve our understanding of the MTBC host adaptation, geographic distribution and
561 pathogenicity.

562 Alongside these genomes, highly accurate annotation of the genes within is also needed.
563 The availability of the H37Rv genome, which has undergone many rounds of intense
564 reannotation (Modlin et al., 2021), allowed for merging of genes that were over-split by
565 PGAP. However, this meant only genes present in H37Rv could be re-merged where needed,

potentially leaving other genes mis-annotated in the accessory genome. High quality, and high-throughout, annotation of other genomes will help rectify this in future expanded studies.

Conclusion

In conclusion, this study reveals that the MTBC has a small, near-closed pangenome, predominantly driven by genome reduction. This observation underlines the importance of genomic deletions in the evolution of the MTBC. We found many distinct genomic features across the MTBC, including sublineage-specific RDs, which likely play a crucial role in variations in virulence, metabolism, and antibiotic resistance. Finally, the study underscores the impact of non-coding regions and large structural variations on the MTBC's genetic diversity, offering a new perspective on MTBC pangenomics.

Acknowledgements and funding

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

C.J.M., D.J.W. and M.B. are supported by the Academy of Medical Sciences (AMS), the Wellcome 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.

D.W. contributed to methodology, formal analysis, data curation, 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,

646 D., CHILLINGWORTH, T., CONNOR, R., DAVIES, R., DEVLIN, K., FELTWELL, T., GENTLES,
647 S., HAMLIN, N., HOLROYD, S., HORNSBY, T., JAGELS, K., KROGH, A., MCLEAN, J.,
648 MOULE, S., MURPHY, L., OLIVER, K., OSBORNE, J., QUAIL, M. A., RAJANDREAM, M. A.,
649 ROGERS, J., RUTTER, S., SEEGER, K., SKELTON, J., SQUARES, R., SQUARES, S.,
650 SULSTON, J. E., TAYLOR, K., WHITEHEAD, S. & BARRELL, B. G. 1998. Deciphering the
651 biology of Mycobacterium tuberculosis from the complete genome sequence.
652 *Nature,* 393**,** 537-544.
653 COSCOLLA, M. 2017. Biological and Epidemiological Consequences of MTBC Diversity. *Adv*
654 *Exp Med Biol,* 1019**,** 95-116.
655 COSCOLLA, M., GAGNEUX, S., MENARDO, F., LOISEAU, C., RUIZ-RODRIGUEZ, P., BORRELL, S.,
656 OTCHERE, I. D., ASANTE-POKU, A., ASARE, P., SÁNCHEZ-BUSÓ, L., GEHRE, F.,
657 SANOUSSI, C. N., ANTONIO, M., AFFOLABI, D., FYFE, J., BECKERT, P., NIEMANN, S.,
658 ALABI, A. S., GROBUSCH, M. P., KOBBE, R., PARKHILL, J., BEISEL, C., FENNER, L.,
659 BÖTTGER, E. C., MEEHAN, C. J., HARRIS, S. R., DE JONG, B. C., YEBOAH-MANU, D. &
660 BRITES, D. 2021. Phylogenomics of Mycobacterium africanum reveals a new lineage
661 and a complex evolutionary history. *Microb Genom,* 7.
662 DARWIN, K. H. & STANLEY, S. A. 2022. The aldehyde hypothesis: metabolic intermediates as
663 antimicrobial effectors. *Open Biol,* 12**,** 220010.
664 DE JONG, B. C., ANTONIO, M. & GAGNEUX, S. 2010. Mycobacterium africanum—Review of
665 an Important Cause of Human Tuberculosis in West Africa. *PLOS Neglected Tropical*
666 *Diseases,* 4**,** e744.
667 DEB, C., LEE, C.-M., DUBEY, V. S., DANIEL, J., ABOMOELAK, B., SIRAKOVA, T. D., PAWAR, S.,
668 ROGERS, L. & KOLATTUKUDY, P. E. 2009. A Novel In Vitro Multiple-Stress Dormancy
669 Model for Mycobacterium tuberculosis Generates a Lipid-Loaded, Drug-Tolerant,
670 Dormant Pathogen. *PLOS ONE,* 4**,** e6077.
671 DEJESUS, M. A., GERRICK, E. R., XU, W., PARK, S. W., LONG, J. E., BOUTTE, C. C., RUBIN, E. J.,
672 SCHNAPPINGER, D., EHRT, S., FORTUNE, S. M., SASSETTI, C. M. & IOERGER, T. R. 2017.
673 Comprehensive Essentiality Analysis of the Mycobacterium tuberculosis Genome via
674 Saturating Transposon Mutagenesis. *mBio,* 8.
675 ESPINOZA, M. E., SWING, A. M., ELGHRAOUI, A., MODLIN, S. J. & VALAFAR, F. 2025. Interred
676 mechanisms of resistance and host immune evasion revealed through network-
677 connectivity analysis of M. tuberculosis complex graph pangenome. *mSytems,* 10
678 FLORES-VALDEZ, M. A., ACEVES-SÁNCHEZ, M. J., PETERSON, E. J. R., BALIGA, N., BRAVO-
679 MADRIGAL, J., DE LA CRUZ-VILLEGAS, M., ARES, M. A., BORN, S., VOSKUIL, M., PÉREZ-
680 PADILLA, N. A., BURCIAGA-FLORES, M., CAMACHO-VILLEGAS, T. A. & ESPINOZA-
681 JORGE, M. G. 2020. Transcriptional portrait of M. bovis BCG during biofilm
682 production shows genes differentially expressed during intercellular aggregation and
683 substrate attachment. *Sci Rep,* 10**,** 12578.
684 GAGNEUX, S. 2018. Ecology and evolution of Mycobacterium tuberculosis. *Nat Rev*
685 *Microbiol,* 16**,** 202-213.
686 GAGNEUX, S., DERIEMER, K., VAN, T., KATO-MAEDA, M., DE JONG, B. C., NARAYANAN, S.,
687 NICOL, M., NIEMANN, S., KREMER, K., GUTIERREZ, M. C., HILTY, M., HOPEWELL, P. C.
688 & SMALL, P. M. 2006. Variable host-pathogen compatibility in *Mycobacterium*
689 *tuberculosis*. *Proceedings of the National Academy of Sciences,* 103**,** 2869-2873.
690 GODFROID, M., DAGAN, T. & KUPCZOK, A. 2018. Recombination Signal in Mycobacterium
691 tuberculosis Stems from Reference-guided Assemblies and Alignment Artefacts.
692 *Genome Biol Evol,* 10**,** 1920-1926.

GUYEUX, C., SENELLE, G., LE MEUR, A., SUPPLY, P., GAUDIN, C., PHELAN, J., CLARK, T. G., RIGOUTS, L., DE JONG, B., SOLA, C. & REFRÉGIER, G. 2024. Newly Identified *Mycobacterium africanum* Lineage 10, Central Africa. *Emerging Infectious Disease journal,* 30**,** 560.

HERSHBERG, R., LIPATOV, M., SMALL, P. M., SHEFFER, H., NIEMANN, S., HOMOLKA, S., 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.

MODLIN, S. J., ELGHRAOUI, A., GUNASEKARAN, D., ZLOTNICKI, A. M., DILLON, N. A., DHILLON, N., KUO, N., ROBINHOLD, C., CHAN, C. K., BAUGHN, A. D. & VALAFAR, F. 2021. Structure-Aware Mycobacterium tuberculosis Functional Annotation Uncloaks Resistance, Metabolic, and Virulence Genes. *mSystems,* 6**,** e00673-21.

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.

NAPIER, G., COUVIN, D., REFRÉGIER, G., GUYEUX, C., MEEHAN, C. J., SOLA, C., CAMPINO, S., PHELAN, J. & CLARK, T. G. 2023. Comparison of in silico predicted Mycobacterium tuberculosis spoligotypes and lineages from whole genome sequencing data. *Scientific Reports,* 13**,** 11368.

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.

ORGEUR, M., SOUS, C., MADACKI, J. & BROSCH, R. 2024. Evolution and emergence of Mycobacterium tuberculosis. *FEMS Microbiology Reviews,* 48.

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.

RIOJAS, M. A., MCGOUGH, K. J., RIDER-RIOJAS, C. J., RASTOGI, N. & HAZBÓN, M. H. 2018. Phylogenomic analysis of the species of the Mycobacterium tuberculosis complex demonstrates that Mycobacterium africanum, Mycobacterium bovis, Mycobacterium caprae, Mycobacterium microti and Mycobacterium pinnipedii are later heterotypic synonyms of Mycobacterium tuberculosis. *Int J Syst Evol Microbiol,* 68**,** 324-332.

SANDERSON, N. D., HOPKINS, K. M. V., COLPUS, M., PARKER, M., LIPWORTH, S., CROOK, D. & STOESSER, N. 2024. Evaluation of the accuracy of bacterial genome reconstruction with Oxford Nanopore R10.4.1 long-read-only sequencing. *Microb Genom,* 10.

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.

SNIPEN, L. & LILAND, K. H. 2015. micropan: an R-package for microbial pan-genomics. *BMC Bioinformatics,* 16**,** 79.

STINEAR, T. P., SEEMANN, T., PIDOT, S., FRIGUI, W., REYSSET, G., GARNIER, T., MEURICE, G., SIMON, D., BOUCHIER, C., MA, L., TICHIT, M., PORTER, J. L., RYAN, J., JOHNSON, P. D., DAVIES, J. K., JENKIN, G. A., SMALL, P. L., JONES, L. M., TEKAIA, F., LAVAL, F., DAFFÉ, M., PARKHILL, J. & COLE, S. T. 2007. Reductive evolution and niche adaptation inferred from the genome of *Mycobacterium ulcerans*, the causative agent of Buruli ulcer. *Genome Res,* 17**,** 192-200.

STRITT, C. & GAGNEUX, S. 2023. How do monomorphic bacteria evolve? The Mycobacterium tuberculosis complex and the awkward population genetics of extreme clonality. *Peer Community Journal,* 3.

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.

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. 2024. *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

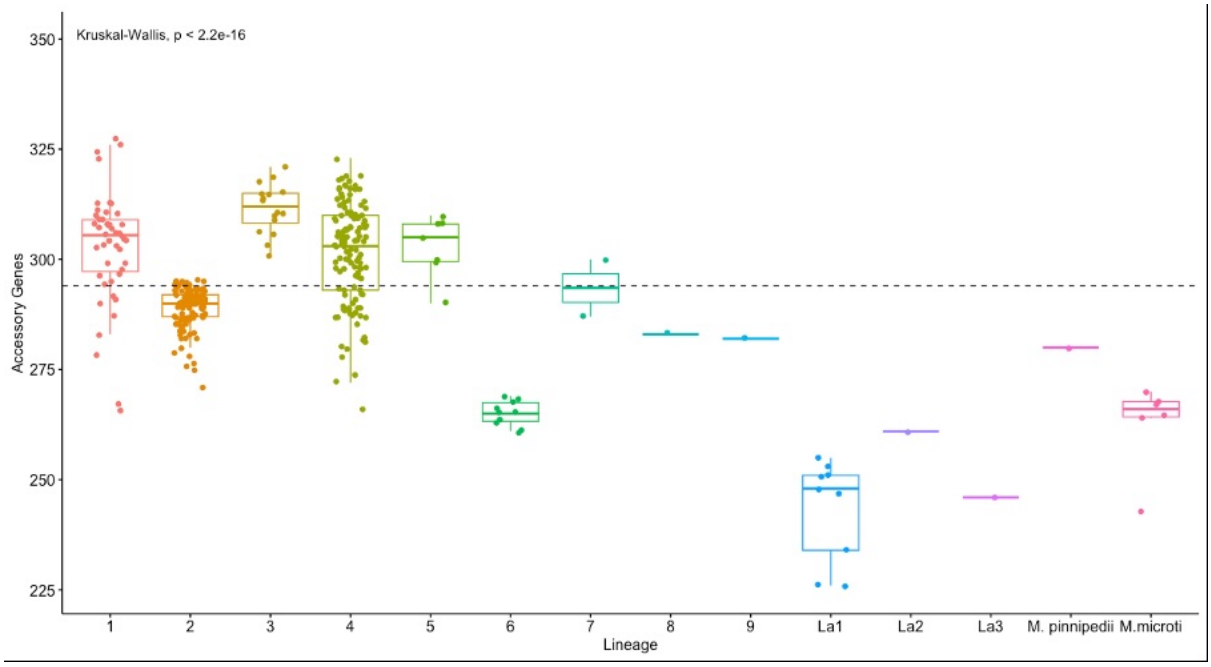
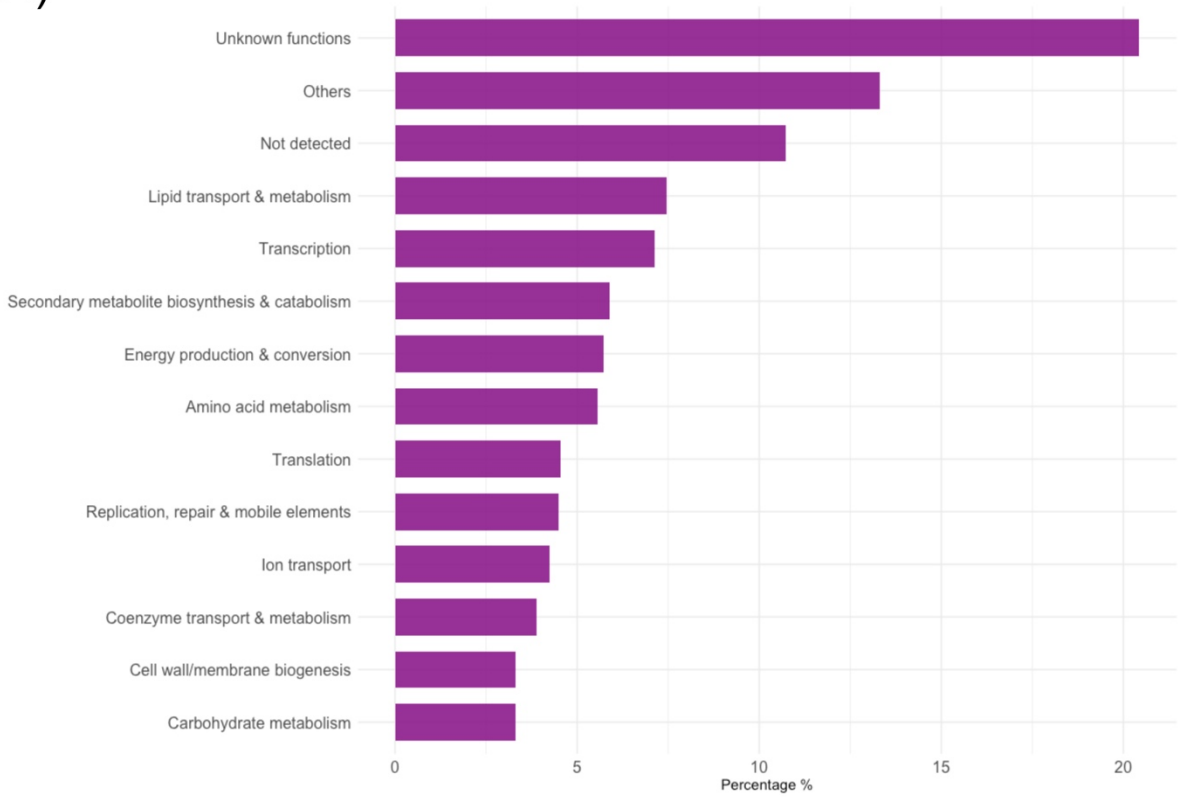
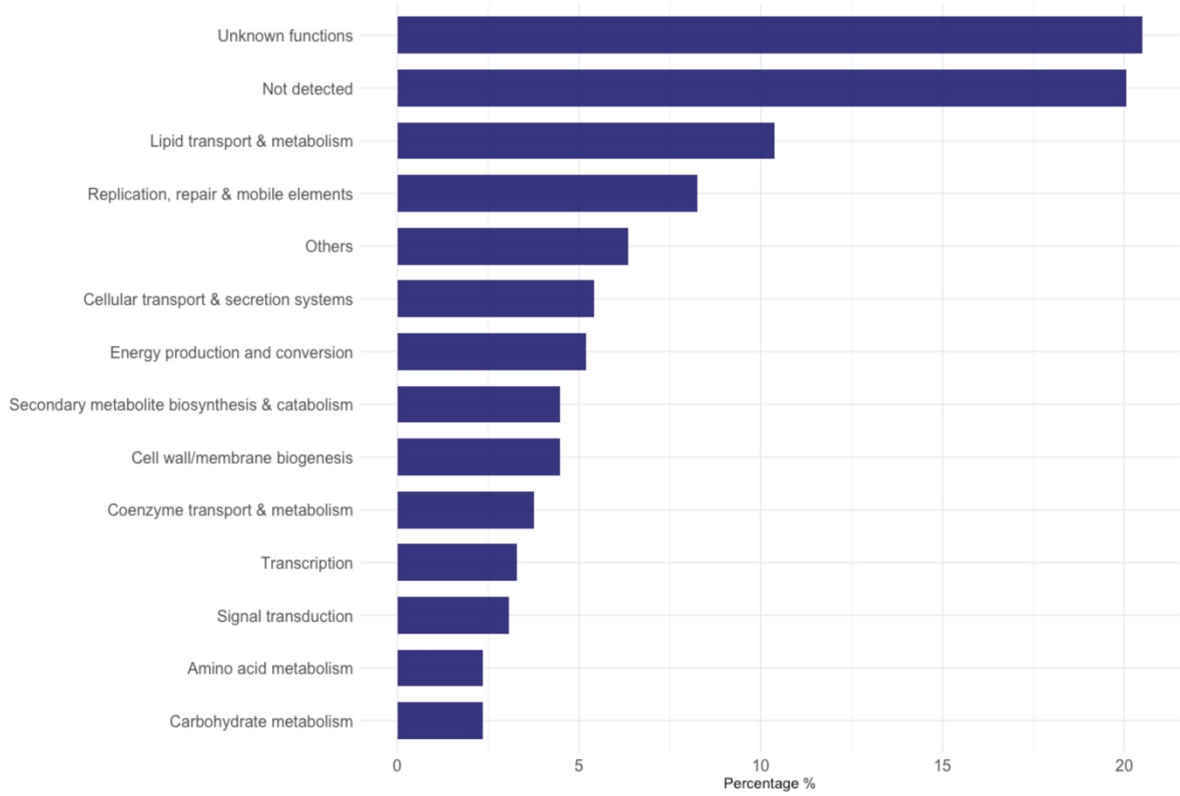


Figure 1: Boxplot showing the distribution of accessory genes within each lineage. Lineages with a small number of genomes were excluded from the statistical analysis. L6, La1 and *M. microti* have significantly smaller accessory genomes compared to other lineages.

A)


Unknown functions
Others
Not detected
Lipid transport & metabolism
Transcription
Secondary metabolite biosynthesis & catabolism
Energy production & conversion
Amino acid metabolism
Translation
Replication, repair & mobile elements
Ion transport
Coenzyme transport & metabolism
Cell wall/membrane biogenesis
Carbohydrate metabolism
0
5
10
15
20
Percentage %

B)


Unknown functions
Not detected
Lipid transport & metabolism
Replication, repair & mobile elements
Others
Cellular transport & secretion systems
Energy production and conversion
Secondary metabolite biosynthesis & catabolism
Cell wall/membrane biogenesis
Coenzyme transport & metabolism
Transcription
Signal transduction
Amino acid metabolism
Carbohydrate metabolism
0
5
10
15
20
Percentage %

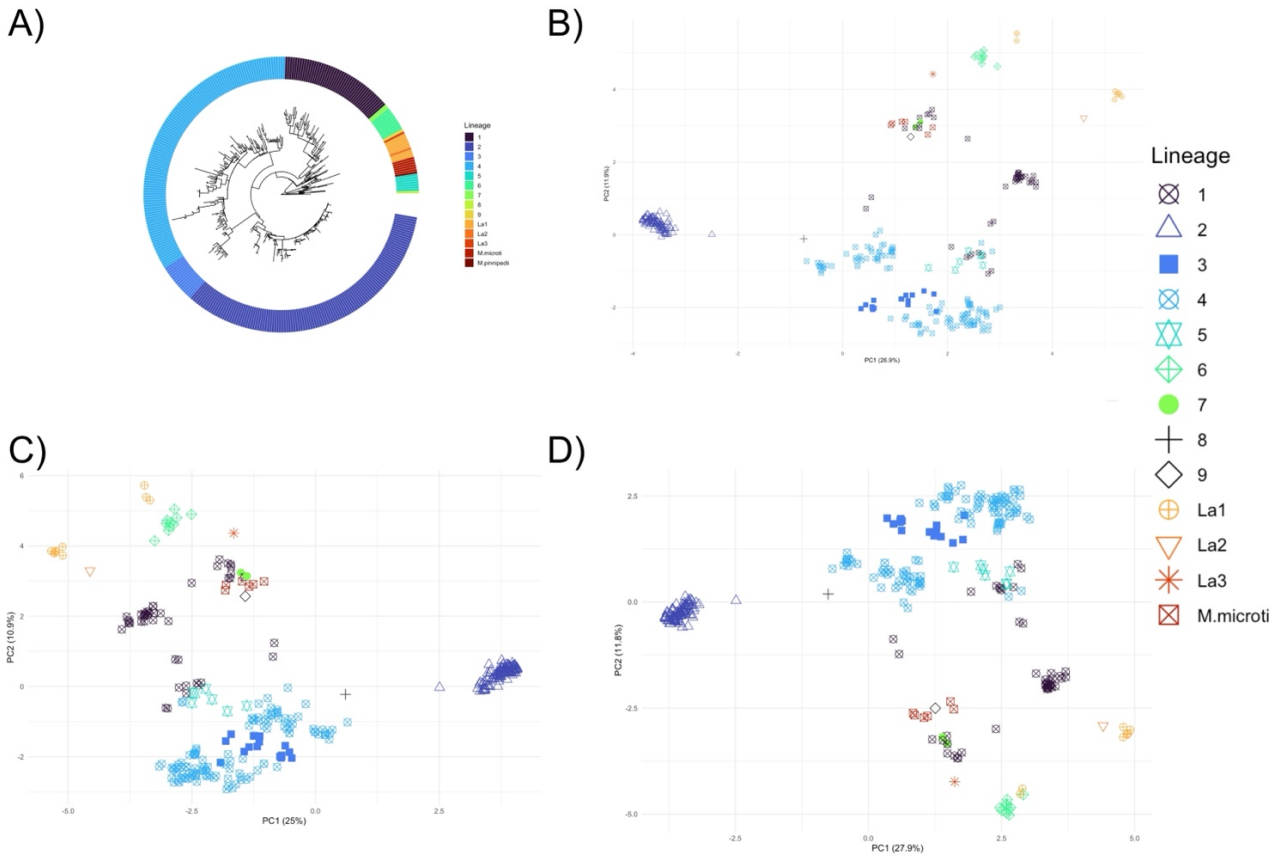
849

850 **Figure 2: Analysis of the functional components in the core (A) and accessory (B) genome using**

851 **EggNOG mapper and InterProScan.** Core and accessory genes are derived from Panaroo with

852 merged paralogs. For without merged paralogs see Supplementary Figure S3.

853



854

855 **Figure 3: (A) phylogenetic tree based on MTBC core genome. PCA based on the accessory genome data from (B) Panaroo (merged paralogs), (C) Panaroo**

856 **(no merged paralogs) and (D) accessory regions data from Pangraph.**

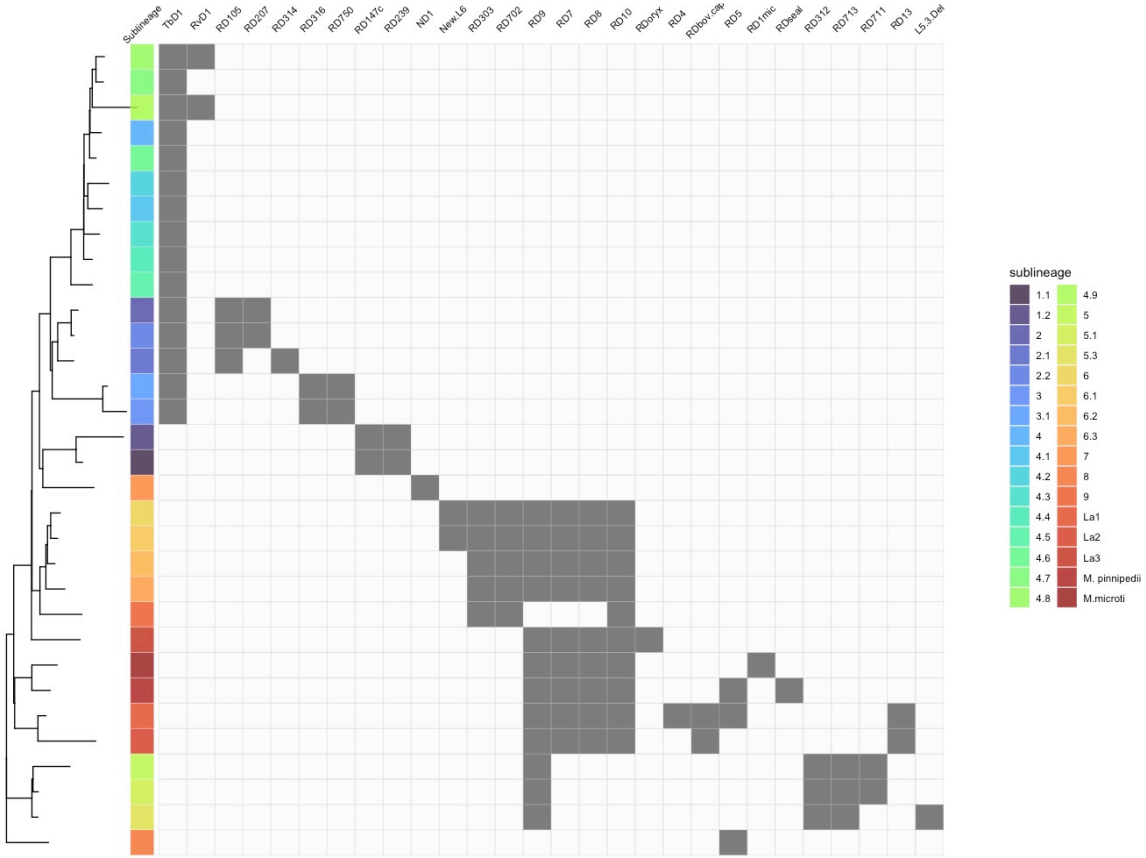


Figure 4: Sub-lineage specific regions of differences (RDs) identified using pangenome-based approaches. Sub-lineages are shown on the Y-axis coloured as per the legend. H37Rv sits within L4.9 at the top of the Y-axis. RDs (structural variants present in all members of one or more sub-lineages and absent in all members of one or more other sub-lineages) are listed on the X-axis, grouped by their pattern of presence/absence. Only RDs detected by both the Panaroo

861 and Pangraph-based approaches are shown. Grey boxes indicate that the region is absent from that sub-lineage. Only RDs that are due to divergent

862 evolution and found present or absent in 2 or more genomes are shown here.

863

864 **Supplementary file legends**

865 **Supplementary Table S1.** An overview of the genome dataset, both publicly acquired data
866 and those strains sequenced within this study including assembly accessions and online
867 locations, where relevant. BUSCO completeness information for each genome is also
868 provided, as is the CDS and pseudogene count as predicted by PGAP.

869 **Supplementary Table S2.** A list of all gene groups which were combined to reduce over-
870 splitting of pangenome due to annotation errors. Each gene groups name is listed alongside
871 its original classification (Core, Sort-core, Cloud, Shell) and its new classification after
872 merging with others after annotation correction. This is shown for both the pangenome with
873 and without merged paralog setting enabled.

874 **Supplementary Table S3.** Pairwise comparisons of accessory genome size between lineages.
875 A * indicates a significant difference between these lineages. Comparisons on both merged
876 and unmerged paralog datasets are shown.

877 **Supplementary Table S4.** A list of all RDs (known and new) along with the genes contained
878 in each and the lineages the regions are absent from.

879 **Supplementary Figure S1: The geography of the dataset.** (A) sample collection distribution by
880 country; (B) the number of genomes of each lineage included from each continent.

881 **Supplementary Figure S2: Pangenome openness assessment using Heap's Law analysis, genome**
882 **fluidity and rarefaction curve for A) Panaroo with merged paralogs; B) Panaroo with unmerged**
883 **paralogs and C) Pangraph blocks.**

884 **Supplementary Figure S3: Analysis of the functional components in the core (A) and accessory (B)**
885 **genome using EggNOG mapper and InterProScan.** Core and accessory genes are derived from
886 Panaroo without merged paralogs.

887 **Supplementary Figure S4: MTBC accessory genome distribution based on (A) Panaroo (merged**
888 **paralogs) and (B) Pangraph.** The MTBC phylogenetic tree is shown on the left beside a coloured bar
889 indicating the sub-lineage of each tip genome. The accessory genes/regions are indicated by columns
890 in the heatmap with a blue box if present in that strain's genome.

891 **Supplementary Figure S5: Coinfinder gene association heatmap.** Coinfinder reveals gene
892 association patterns within the accessory genome of the MTBC. Accessory genes, listed on the
893 X-axis, cluster into groups (shown here by varying colours of the blocks).

894 **Supplementary Figure S6: MTBC accessory genome distribution based on (A) Panaroo (merged**
895 **paralogs) and (B) Pangraph.** The MTBC phylogenetic tree is shown on the left beside a coloured bar
896 indicating the sub-lineage of each tip genome. The accessory genes/regions are indicated by columns
897 in the heatmap with a blue box if present in that strain's genome.

898

899

900

901