Mutation rates and adaptive variation among the clinically dominant clusters of *Mycobacterium abscessus*

Nicoletta Commins[1], Mark R. Sullivan[2], Kerry McGowen[2], Evan Koch[1], Eric J. Rubin[2,3], Maha Farhat[1,4] *

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

[2] Department of Immunology and Infectious Diseases, Harvard T. H. Chan School of Public Health, Boston, MA, 02115, USA.

[3] Department of Microbiology, Harvard Medical School, Boston, MA, 02115, USA.

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

* Correspondence: maha_farhat@hms.harvard.edu

Abstract

Mycobacterium abscessus (*Mab*) is a multi-drug resistant pathogen increasingly responsible for severe pulmonary infections. Analysis of whole genome sequences (WGS) of *Mab* demonstrates dense genetic clustering of clinical isolates collected from disparate geographic locations. This has been interpreted as supporting patient-to-patient transmission, but epidemiological studies have contradicted this interpretation. Here we present evidence for a slowing of the *Mab* molecular clock rate coincident with the emergence of phylogenetic clusters. We find that clustered isolates are enriched in mutations affecting DNA repair machinery and have lower spontaneous mutation rates in *in vitro*. We propose that *Mab* adaptation to the host environment through variation in DNA repair genes affects the organism's mutation rate and that this manifests

as phylogenetic clustering. These results inform our understanding of niche switching for facultative pathogens, and challenge the model of transmission as the major mode of dissemination of clinically dominant *Mab* clusters.

Introduction

Mycobacterium abscessus (*Mab*) is an emerging, multi-drug resistant pathogen that is increasingly responsible for opportunistic pulmonary disease, most commonly in patients with underlying structural lung conditions such as cystic fibrosis (CF).[1] *Mab* is divided into three subspecies: *M. a. abscessus*, *M. a. massiliense*, and *M. a. bolletii,* the former two being the most clinically relevant subspecies.[2] Infections with *Mab* are associated with accelerated decline in lung function and can result in severe disseminated disease and/or loss of eligibility for lung transplantation. *Mab* infections are difficult to treat, requiring a prolonged antibiotic regimen with high rates of treatment failure.[3]

Until recently, *Mab* infections were thought to be acquired independently by each patient from environmental reservoirs in soil or water. Large scale analyses of whole genome sequencing (WGS) of Mab clinical isolates have revealed that a large proportion of clinical isolates fall into several phylogenetically characterized clusters, often termed dominant circulating clones.[4] Isolates within these clusters have high core genome similarity that is not explained by point source outbreaks as they were sampled from diverse geographic origins. These observations have been interpreted as supporting widespread recent transmission, either through direct contact or indirectly through fomites. However, a number of epidemiological studies have refuted person-to-person transmission as a major mode of *Mab* dissemination.[5–11] Reconciling these conflicting lines of evidence is important as they dictate public health practice and allocation of resources to protect vulnerable populations from infection. However, it is still unclear what role transmission plays in the emergence of *Mab* clusters.

While widespread transmission is one proposed explanation for the phylogenetic structure observed in *Mab*, it is important to consider other factors in the species' evolutionary history that may contribute. For example, ecological niche switching from environmental sources to a human host may introduce bottlenecks resulting in low diversity and effective population size. Another explanation is a change in mutation rates, either through changes in bacterial generation time or through adaptive genetic variation. A third consideration is sampling bias, where particular *Mab* lineages that cause more frequent, severe and/or persistent disease are more likely to be isolated and sequenced. Consistent with the latter hypothesis, previous work has shown that isolates within clusters are more likely to contain point mutations associated with antibiotic resistance and are associated with greater bacterial burden and lung inflammation in mice.[4]

Interpreting tree structures in the context of population history is aided by phylogenetic dating, which translates genetic distances into time scale. Previous studies[4,9,12,13] have estimated molecular clock rates in *Mab* but these have been limited to subpopulations within clusters. Here we report on molecular clock rate estimation for the deep branches of the *M. a. abscessus* phylogeny and for within-host isolates belonging to clusters. We provide evidence for an evolutionary slow-down coincident with the expansion of clinically dominant clusters and propose that the emergence of dense phylogenetic clustering in *Mab* may result from genetically encoded changes to the mutation rate. These results indicate that transmission is not the only plausible explanation for the dense clustering observed in the *Mab* phylogeny, and that more research is needed to clarify the transmission dynamics of this emerging pathogen.

Results

<u>Phylogenetic analysis confirms the presence of large, dominant clusters (DCs) spanning multiple geographies</u>. We identified 1,629 *M. abscessus* isolates with publicly available whole genome sequencing (WGS) data and associated dates of collection (Table S1). We excluded isolates with

76 likely contamination or low-quality assemblies (Methods). Among the 1,461 isolates passing
77 quality filters, the majority (80.1%) were derived from *in vitro* culture of pulmonary samples. Other
78 samples were derived from a skin swab or biopsy, other body sites, and environmental sources
79 (Table S2). Of the 1,181 pulmonary isolates, 98.3% (1,161/1,181) were collected from patients
80 with cystic fibrosis (CF), while the rest were collected from patients with no documented diagnosis.
81 The dataset represented 483 patients with unique patient identifiers, with 59.6% (288/483) of
82 patients contributing a single isolate, and the remaining contributing between 2 to 24 isolates
83 sampled over time. Isolates originated from nine countries and were collected between 1998-
84 2017, with the exception of the reference strain ATCC 19977 which was collected in 1957 (Figure
85 1B).

86 The majority (91.7%) of samples had ≥98% average nucleotide identity (ANI) with one of
87 three reference genomes (Table S3) representing each of the three Mab subspecies, consistent
88 with the commonly used ANI threshold for bacterial subspecies. Isolates with <98% ANI with any
89 subspecies were excluded from downstream analysis. The resulting dataset overrepresented
90 clinically relevant subspecies of *M. abscessus*, with 93.9% belonging to either *M. a. abscessus*
91 or *M. a. massiliense*. To better resolve true evolutionary relationships among samples, we defined
92 and restricted analysis to the core genome for each subspecies, then further excluded predicted
93 recombination events (Methods). We performed maximum likelihood (ML) phylogenetic analysis
94 for each of the three *Mab* subspecies, using only the first isolate from each patient to focus on
95 between-host diversity. The resulting subspecies trees confirmed the presence of multiple,
96 globally distributed clades with dense phylogenetic clustering. Clusters were defined using a one-
97 dimensional scanning statistic that compared the branch lengths of a clade to the distribution of
98 branch lengths across the entire subspecies tree (Methods). Using this approach, we initially
99 identified 30 clusters. We further excluded clustering due to potential point source outbreaks by
100 limiting to clusters containing isolates from more than one country and a minimum of three isolates

101 for downstream analysis. In total, we confirmed 16 clusters (Figure S2A) including four clusters
102 with > 50 isolates, which we refer to here as dominant clusters (DCs). Isolates within clusters
103 comprised 76.4% of isolates included in the phylogenetic analysis. The median pairwise distance
104 in *M. a. abscessus* DC 1 was 68 SNPs. For comparison, the mean pairwise distance across the
105 subspecies trees were 5537, 4100, and 8034 SNPs for *M. a. abscessus*, *M. a. massiliense*, and
106 *M. a. bolletii*, respectively. All four DCs contained isolates originating from three continents (North
107 America, Europe, and Australia). Clustered isolates also spanned the full range of isolation times,
108 including ATCC 19977, which was placed within *M. a. abscessus* DC 1 despite having been
109 collected at least 40 years prior to any other isolates in the tree (Figure 1A-B).
110
111 Coalescent analysis reveals a faster long-term evolutionary rate than observed in DCs. To aid in
112 transmission inference and to reconstruct the evolutionary history of *Mab*, we performed a
113 coalescent analysis using our SNP alignment. Previous work has reported molecular clock rates
114 within *Mab* clusters, but to our knowledge no study has successfully dated any of the full *Mab*
115 subspecies phylogenies. To evaluate the presence of temporal signal, we performed root-to-tip
116 regression for each subspecies tree but in each case did not observe a positive correlation
117 between sampling date and root-to-tip distance. We reasoned that the high degree of population
118 structure in the *Mab* phylogeny may obscure the genetic-temporal signal. To test this hypothesis,
119 we pruned each subspecies tree until it contained the minimal number of samples required to
120 capture 95% of the genetic diversity in the original tree (Methods). Following pruning of the *M. a.*
121 *abscessus* subspecies tree, root-to-tip regression demonstrated the presence of positive temporal
122 signal (Figure 1C, D). Evidence of temporal signal in *M. a. abscessus* was further strengthened
123 using a permutation test and a Bayesian evaluation of temporal signal (Figure S2C, Table S4).
124 We used a Bayesian implementation of coalescent analysis, BEAST[14], to estimate the
125 molecular clock rate of the pruned *M. a. abscessus* subspecies tree assuming a relaxed clock
126 model, which allows each branch of the tree to have its own clock rate. To explore the effect of

127 the tree prior on molecular clock estimation, we ran analyses using two different tree priors,
128 assuming either a constant coalescent model, which assumes constant population size over time,
129 or a Bayesian skyline coalescent model, which allows the population size to change over time.
130 While the Bayesian skyline coalescent had a slightly better marginal likelihood, we found no
131 significant difference between the clock rates estimated by the two models (Table 1, Figure 2C).
132 Our estimates for the clock rate in the long, internal branches of the tree were approximately 10-
133 fold faster than nearly all previously reported estimates of the clock rate within *Mab* clusters
134 (Figure 2C). We next attempted to date the full (unpruned) *M. a. abscessus* phylogeny using the
135 faster rate inferred from our coalescent analysis of the pruned tree. These chains suffered from
136 poor convergence and sampled trees with a distorted topology, elongating the short branches
137 and eliminating the densely clustered structure observed in the ML phylogeny (data not shown).
138 We therefore hypothesized that the lack of temporal signal in the full *M. a. abscessus* dataset
139 may result from a difference between the long-term clock rate and the more recent evolutionary
140 rate within phylogenetic clusters.
141
142 Analysis of longitudinally sampled within-host isolates supports a slower clock rate in DCs. To
143 further investigate the differences in clock rate between long-term historical time scales and short-
144 term contemporary time, we used pairs of isolates sampled from the same patient over time to
145 estimate the mutation rate. We focused on estimating a within-host clock rate for *M. a. abscessus*
146 to compare directly with the estimate from our coalescent analysis and because we had the most
147 longitudinal samples for this subspecies. Because polyclonal *Mab* infection is common in CF
148 patients, we first excluded any isolate pairs with a SNP distance > 20. This threshold was selected
149 by comparing the distribution of all possible within-host isolate pairs to the distribution of pairwise
150 SNP distances within *Mab* clusters (Figure 2A, Figure S3). Regressing the SNP distance between
151 isolate pairs on the time between sample pairs yielded a mutation rate of 2.21 SNPs/year [1.06 –
152 3.37 95% CI], or 5.17×10^{-7} SNPs/site/year [2.48×10^{-7} – 7.92×10^{-7} 95% CI]. This value is more

153 comparable to previously reported molecular clock rate estimates within *Mab* clusters[4,9,12,13] than
154 to those estimated in the long branches of the phylogeny (Figure 2C).
155
156 Simulated ancestries support mutation rate slow-down over population size effects as the major
157 driver of phylogenetic clustering. The high core genome similarity of isolates within clusters has
158 been interpreted as supporting rapid global dissemination of *Mab* clones through direct
159 transmission, but may instead relate to population bottlenecks, oversampling, or a relatively slow
160 rate of molecular evolution in human-adapted clones. To test whether sampling from clades with
161 low genetic diversity is sufficient to explain the degree of clustering observed in the ML phylogeny,
162 or whether changes in the molecular clock rate are a better explanation for the observed
163 phylognetic data, we performed coalescent simulations of ancestries with one clustered
164 subpopulation (A) and one unclustered subpopulation (B). Under a range of effective population
165 sizes (N_e) and mutation rates (μ), we simulated ancestries through time with 1000 replicates per
166 simulation. For each replicate we assessed the degree of phylogenetic clustering by measuring
167 d_A/d_B where d_i is the mean pairwise SNP distance within subpopulation i. The total N_e (N_e^T),
168 number of samples, and sampling dates for subpopulations A and B were drawn from the true
169 values for *M. a. abscessus* DC 1 and from all "unclustered" isolates in our dataset, respectively
170 (Figure 3A). We first tested whether a smaller population size for the clustered subpopulation A
171 (N_e^A) could explain the degree of phylogenetic clustering observed. We estimated a range of
172 realistic values for N_e^A using BEAST by assuming a range of possible clock rates (Table S5).
173 These rates spanned a published clock rate estimate for *M. a. abscessus* DC1 to a clock rate that
174 is 50-fold higher, far exceeding the upper 95% highest posterior density (HPD) interval for the
175 published value.[13] Under the the simulated conditions of equal mutation rate between
176 subpopulation A and B, phylogenetic clustering to the extent observed in the clinical isolate
177 phylogeny was not a likely outcome (Figure 3B). We confirmed these finding in simulations where
178 the population sizes N_e^T spanned the 95% confidence intervals under both tree priors shown in

179 Table 1, except in the scenario where $N_e^T >> N_e^A$ (Figure S4). We next tested whether a change
180 in the mutation rate at the root of supopulation A could better explain the observed phylogenetic
181 structure in *Mab*. The simulated ancestries where N_e^A and N_e^B are fixed and the mutation rate
182 slowdown is between 10- and 20-fold better recapitulated the observed phylognetic structure
183 (Figure 3C). This result was robust to changes in N_e^A and N_e^T across the 95% HPD for each
184 parameter (Figure S5). These results indicate that for person-to-person transmission to drive
185 phylogenetic clustering in *Mab*, the total N_e must be very large compared to the N_e of the clusterd
186 subpopulation, an unlikely scenario given the observed genetic data and sampling times. By
187 contrast, a mutation rate change of 10- to 20- fold explains the degree of clustering observed in
188 *Mab* under more realistic population sizes given these data.
189
190 <u>UvrD/Rep family helicases are enriched in DCs in *M. a. abscessus*</u>. To identify genetic variants
191 under positive selection in *Mab* clusters, we searched for evidence of phylogenetic convergence
192 in the core genome of *M. a. abscessus* with regions of recombination excluded. Briefly, we
193 performed ancestral sequence reconstruction for the internal nodes of the subspecies tree and
194 counted the number of independent arisals of each SNP (Methods). Of the 65,202 SNPs identified
195 in *M. a. abscessus*, 34,813 (53.4%) of mutations occurred only once and were excluded. For the
196 remaining SNPs, we ranked each by its enrichment in DCs relative to other regions of the *M. a.*
197 *abscessus* tree (Figure 4A). The most significantly enriched variants in the *M. a. abscessus* DCs
198 included homologs of several known or putative virulence factors as well as a number of cluster-
199 enriched mutations in UvrD-like helicases *MAB_1054*, *MAB_3515c*, and *MAB_3516c* (Table 2,
200 Figure 4A). While the function of these genes is not well characterized in *M. abscessus*, all three
201 share homology with the UvrD/Rep family of helicases. This family of genes is involved in DNA
202 repair and has been associated with growth and persistence *in vivo*.[15,16]
203

204 Isolates in *Mab* DCs are associated with a lower spontaneous mutation rate. To test the
205 association between phylogenetic clustering and spontaneous mutation rate, we selected *M. a.*
206 *abscessus* and *M. a. massiliense* strains that were grouped within DCs or that were not present
207 in those clusters. Each of the strains present in the DCs has the derived (minor) allele at four sites
208 in the UvrD/Rep family genes *MAB_1054*, *MAB_3515c*, and *MAB_3516c*, and the strains outside
209 of DCs possess the ancestral (major) allele at those sites (Table 2, Figure 4A). The strains used
210 are described in Table S6 and their genotypes are described in Table S7. The spontaneous
211 mutation rate was estimated in all strains using a fluctuation assay[17] adapted for use in *M.*
212 *abscessus*. The fluctuation assay counts cells gaining resistance to the antibiotic amikacin during
213 growth in culture without antibiotic. By normalizing the number of mutational events to the
214 population size of the culture, the fluctuation assay estimates the spontaneous mutation rate per
215 generation per the number of bases that can be mutated to confer resistance to amikacin. The
216 strains chosen display similar growth rates (Figure S6A) and very low baseline resistance to
217 amikacin (Table S8) which allows them to be compared using a fluctuation assay. Among *M. a.*
218 *abscessus* strains, those present in DCs with the derived (minor) allele at the four UvrD/Rep loci
219 had lower spontaneous mutation rates than those with the ancestral (major) allele (Figure 4C). A
220 similar trend was observed in *M. a. massiliense* strains (Figure 4D). Further, the average mutation
221 rate was significantly lower in isolates present in DCs across both subspecies (Figure 4E),
222 consistent with the hypothesis that the emergence of dense clustering may be driven in part by
223 changes in mutation rate.
224

225 **Discussion**

226 In this study we have pooled a large global sample of *Mab* genome sequences and
227 examined evolutionary rates and positive selection with the goal of understanding the emergence
228 of *Mab* clusters. We present evidence supporting a slower evolutionary rate within DCs compared
229 to the long internal branches of the phylogeny and further data associating adaptive genetic

230 variation in DNA repair genes with the slower evolutionary rate. We posit that changes to the
231 mutation rate underlie the dense clustering in *Mab*, and that while other effects such as population
232 or transmission bottlenecks and sampling bias may contribute to the low genetic diversity within
233 clusters, they are not sufficient to explain the observed phylogenetic structure.

234 Previous studies have restricted molecular clock rate estimation to *Mab* clusters, most
235 likely because of the lack of temporal signal at the subspecies-level. By pruning the *Mab* tree
236 while preserving the majority of the subspecies diversity, we were able to measure a molecular
237 clock rate for *Mab* spanning the time scale from the common ancestor of the subspecies to the
238 present. Estimates of evolutionary rates are recognized to depend on the time scale of
239 measurement. In such examples, the clock rate usually decreases with increasing time depth.[19]
240 Here we observe the opposite effect where the clock rate increases with time depth. The pruned
241 *Mab* tree most likely reflects an evolutionary process taking place predominantly in environmental
242 reservoirs, in contrast to phylogenetic clusters where terminal branches reflect evolution within
243 the human lung environment. We also note that our measurement of within-host mutation rate in
244 *Mab* was independent of phylogenetic inference and the full temporal signal in our sample.
245 Nevertheless, the measured rate within-host was comparable to the clock rate measured
246 previously using coalescent analysis within clusters (Figure 2C).

247 The molecular clock rate is influenced by both the spontaneous mutation rate and the
248 generation time of an organism. Changes in generation time could result from physiological
249 constraints of the human lung environment. For example, differences in nutrient availability may
250 change the growth rate in the lung compared to those living in environmental reservoirs. Longer
251 generation times for bacteria growing in the lung would result in a decrease in the molecular clock
252 rate. Ruis et al. reported differences in mutational signatures along internal branches of the
253 subspecies phylogeny compared to internal branches of DCs, supporting a change in the
254 mutational process within some clades of the tree.[13] The authors attributed these changes to
255 differences in mutagen exposure across niches. Generation times and spontaneous mutation

256 rates can also be modified by genetic changes, for example to DNA replication or repair
257 machinery. The observation that some lineages have adapted to the host environment[4] raises the
258 possibility that genetic changes that underlie or associate with this adaptation affect mutation rate,
259 either directly by affecting DNA replication or repair machinery, or indirectly by affecting the growth
260 rate.

261 We scanned the core genomes for variants enriched in the DCs in *M. a. abscessus*, the
262 subspecies for which we have the most WGS data and for which there are the highest number of
263 clusters. Notably, we observed several cluster-enriched mutations in UvrD-like helicases. UvrD is
264 a superfamily 1 helicase that is involved in DNA repair and is widely conserved in gram-negative
265 bacteria. UvrD1 in *M. tuberculosis* plays an important role both in nucleotide excision repair (NER)
266 and in pathogenesis and persistence because its DNA repair activity confers tolerance to reactive
267 oxygen intermediates (ROI) and reactive nitrogen intermediates (RNI)[15]. Mutations that promote
268 the activity of UvrD1 and its homologs are likely advantageous in the context of lung infection and
269 may have the additional consequence of decreasing the spontaneous mutation rate. We
270 additionally observed mutations in UvrD-like helicases occurring along the cluster-defining branch
271 of the large *M. a. massiliense* cluster, further supporting the role of the NER pathway in *Mab*
272 pathogenesis.

273 Other factors may contribute to the low genetic diversity within clusters. Clinical and
274 experimental data have shown that clustered *Mab* isolates are more likely to cause chronic
275 disease, have point mutations associated with antibiotic resistance, have increased intracellular
276 survival in macrophages, and cause greater bacterial burden and inflammation in mice.[4] Mab
277 genotypes that cause more frequent, persistent, or severe disease are more likely to be sampled
278 and sequenced, and dense sampling of closely related isolates may contribute to the appearance
279 of extreme clustering in the *Mab* phylogeny. Similarly, population bottlenecks due to transmission
280 or other factors may purge diversity and result in the observation of closely related isolates. While
281 epidemiological evidence does not currently support transmission as a major driver of *Mab*

282 dissemination, isolated examples of putative recent transmission events have been described,
283 which may contribute to the observation of dense phylogenetic clustering independent of
284 bottlenecks. However, the results of our ancestry simulations suggest that population size or
285 sampling effects alone do not sufficiently account for the observed phylogenetic structure in *Mab*
286 and instead favor mutation rate changes as a more likely explanation.

287 Our study has several limitations. Our dataset overrepresented clinical isolates from the
288 UK where routine whole genome sequencing is part of *Mab* surveillance. Oversampling of some
289 lineages and geographies may mean that we have not adequately captured the global population
290 structure and that we may have focused on large phylogenetic clusters because of their
291 geographic distribution and not because of their clinical significance. While our study argues
292 against transmission as the main driver of *Mab* dissemination and the emergence of DCs, we do
293 not examine the geographic distribution of Mab within clusters. It is possible that future work
294 considering this data can highlight mechanisms of parallel evolution that explain why clustered
295 *Mab* isolates are detected in distant geographic locations. While our results support a faster
296 molecular clock rate in the long internal branches of the phylogeny compared to clustered
297 lineages, we are unable to distinguish between rate differences due to the spontaneous mutation
298 rate from generation time effects. Differences in growth conditions across environments is difficult
299 to quantify as *Mab* is ubiquitous in both natural and built environments, but nutrient availability
300 likely differs greatly between natural reservoirs such as soil and water compared to man-made
301 reservoirs such as showerheads. A healthy lung supports relatively little bacterial growth, but
302 inflammation increases nutrient availability by promoting mucus production and vascular leakage
303 into the airway.[20] Because there is little publicly available sequencing data from environmental
304 isolates, we were also unable to compare genotypes or mutation rates in environmental samples
305 compared to clinical isolates. Our analysis of variants associated with phylogenetic clustering is
306 also limited in its ability to identify true causal genotypes or those that are enriched in clusters due
307 to hitchhiking to other adaptive variants.

308 In summary, we present new evidence for changes in the molecular clock rate that
309 contribute to the dense phylogenetic clustering observed in *Mab*. These results argue against
310 person-to-person transmission as the primary driver of clustering in the *Mab* phylogeny.
311 Continuing genomic surveillance is needed to understand how *Mab* may be adapting to human
312 infection and to characterize the risk of transmission to at-risk patients. We argue further that the
313 phylogeographic structure of *Mab* is influenced by its complex ecological history as an emerging
314 facultative pathogen that should be considered carefully in future analysis of *Mab* genomic data.
315 Future work should extend these analyses by examining differences in mutation rates and
316 phenotypes between clinical isolates and environmental strains for evidence of host adaptation.
317

318 **Acknowledgements**

319 We thank members of the Farhat lab for feedback and discussion. We are grateful to Luca Freschi
320 for advice on processing WGS data, calling variants, and phylogenetic inference; Maximillian
321 Marin for input on calculating mappability pileup; and Roger Vargas for support in homoplasy
322 counting. We also thank Wilder Wohns for discussions on implementing ancestry simulations with
323 mutation rate changes through time. M.R.S. is a Merck Fellow of the Damon Runyon Cancer
324 Research Foundation, DRG-2415-20. Computational resources and support were provided by the
325 Orchestra High Performance Compute Cluster at Harvard Medical School, which is funded by the
326 NIH (NCRR 1S10RR028832-01).
327

328 **Author Contributions**

329 Conceptualization, N.C. and M.F.; Methodology – Computational analysis, N.C., E.K., and M.F.;
330 Methodology – Experimental studies, M.R.S., K.M.; Investigation – Computational analysis, N.C.;
331 Investigation – Experimental studies, M.R.S., K.M.; Data Curation, N.C.; Writing – Original Draft,
332 N.C. and M.R.S.; Writing – Review & Editing, N.C., M.R.S., E.K., M.F.; Supervision, E.J.R., M.F.
333

Declaration of Interests

The authors declare no competing interests.

360 **Tables**

361

362 **Table 1**. Coalescent analysis results and model comparison

Model	**Marginal Log Likelihood (SD)**	**UCLD Rate (SNPs/site/year)**	**UCLD St. Dev.**
Coalescent constant population	-7144301.64 (21.42)	2.82×10^{-6} [1.33×10^{-6}, 4.83×10^{-6}]	0.182 [0.135, 0.234]
Bayesian skyline coalescent	-7144118.76 (19.93)	2.63×10^{-6} [1.16×10^{-6} , 4.0×10^{-6}]	0.181 [0.138, 0.232]

363

364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398

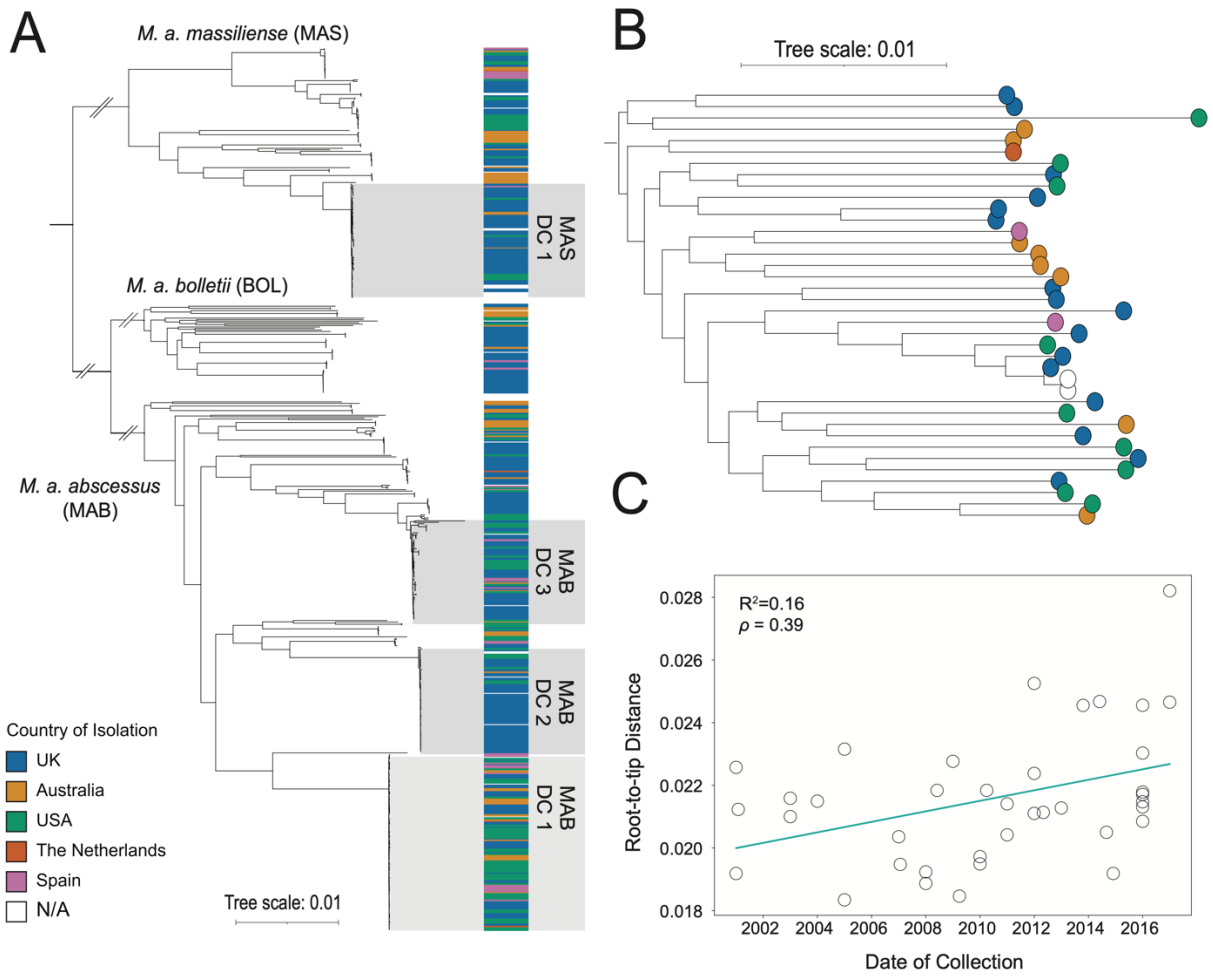
Table 2. Summary of top 10 nonsynonymous variants enriched in DCs of *M. a. abscessus*

Gene	Gene Symbol	Description	Comments	Enrichment p-value	Number of clustered taxa containing variant	Number of unclustered taxa containing variant
MAB_3334c	*gatB* [c]	Aspartyl/glutamyl-tRNA (Asn/Gln) amidotransferase subunit B [c,d]	Essential for growth and survival *in vivo* in *M. tuberculosis* [17]	1.03×10^{-65}	240	1
MAB_3244		Mercury resistance transport protein/cytochrome c biogenesis protein [c]		1.72×10^{64}	189	3
MAB_3480		Putative short chain dehydrogenase/ Reductase [c]		7.23×10^{-64}	232	13
MAB_4141[a]		PE/PPE family protein [c]		1.58×10^{-62}	189	2
MAB_3242	*fni* [c]	Isopentenyl-diphosphate delta-isomerase [c,d]		8.07×10^{-62}	189	3
MAB_3481	*fadE* [c]	Probable acyl-CoA dehydrogenase [c]	FadE mutations in *Mtb* result in decreased growth and virulence in macrophages and in mice[18]	2.10×10^{-61}	232	15
MAB_4057c	*mshA* [c]	D-inositol 3-phosphate glycosyltransferase [d]	Catalyzes biosynthesis of mycothiol, an important compound in resisting oxidative stress and detoxifying harmful agents such as antibiotics[19]	2.78×10^{-60}	189	2
MAB_3581c		Hypothetical protein		6.09×10^{-60}	169	3
MAB_1054 [b]	*pcrA*[c] / *uvrD1*[d]	ATP-dependent DNA helicase [c,d]	Involved in nucelotide excision repair (NER), mismatch repair (MMR), homologous recombination (HR), and rolling circle replication[15]	2.72×10^{-58}	216	7
MAB_3515c	*rep* [c,d]	ATP-dependent DNA helicase [c,d]	Involved in NER, MMR, HR and rolling circle replication[15]	8.42×10^{-57}	189	7

450 Table 2 Legend:
451 [a] Multiple variants in *MAB_4141* have been collapsed and the lowest p-value is shown. See
452 Supplementary Material for full list of homoplasic SNPs ranked by their enrichment in clusters.
453 [b] Gram-negative bacteria possess UvrD and a closely related homolog, Rep, while gram-positive
454 bacteria possess a single homolog called PcrA.[18]
455 [c] Mycobrowser annotation
456 [d] prokka annotation
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478

479 **Figures**

480



481

482 **Figure 1. Dominant clusters of *M. abscessus* and the presence of temporal signal in**
483 **internal branches.** A) Species-wide phylogeny of *M. abscessus* representing one isolate per patient to avoid within-patient clustering of samples. Each subspecies tree was constructed
484 independently using a subspecies specific reference genome. The four largest phylogenetic clusters are highlighted. B) Histogram showing the distribution of collection dates of all
485 clustered isolates v. all unclustered isolates in *M. a. abscessus*. C) Pruned *M. a. abscessus* phylogeny representing 95% of the original subspecies tree diversity. C) Root-to-tip
486 regression showing evidence of positive temporal signal in the pruned tree shown in (B).

487

488

489

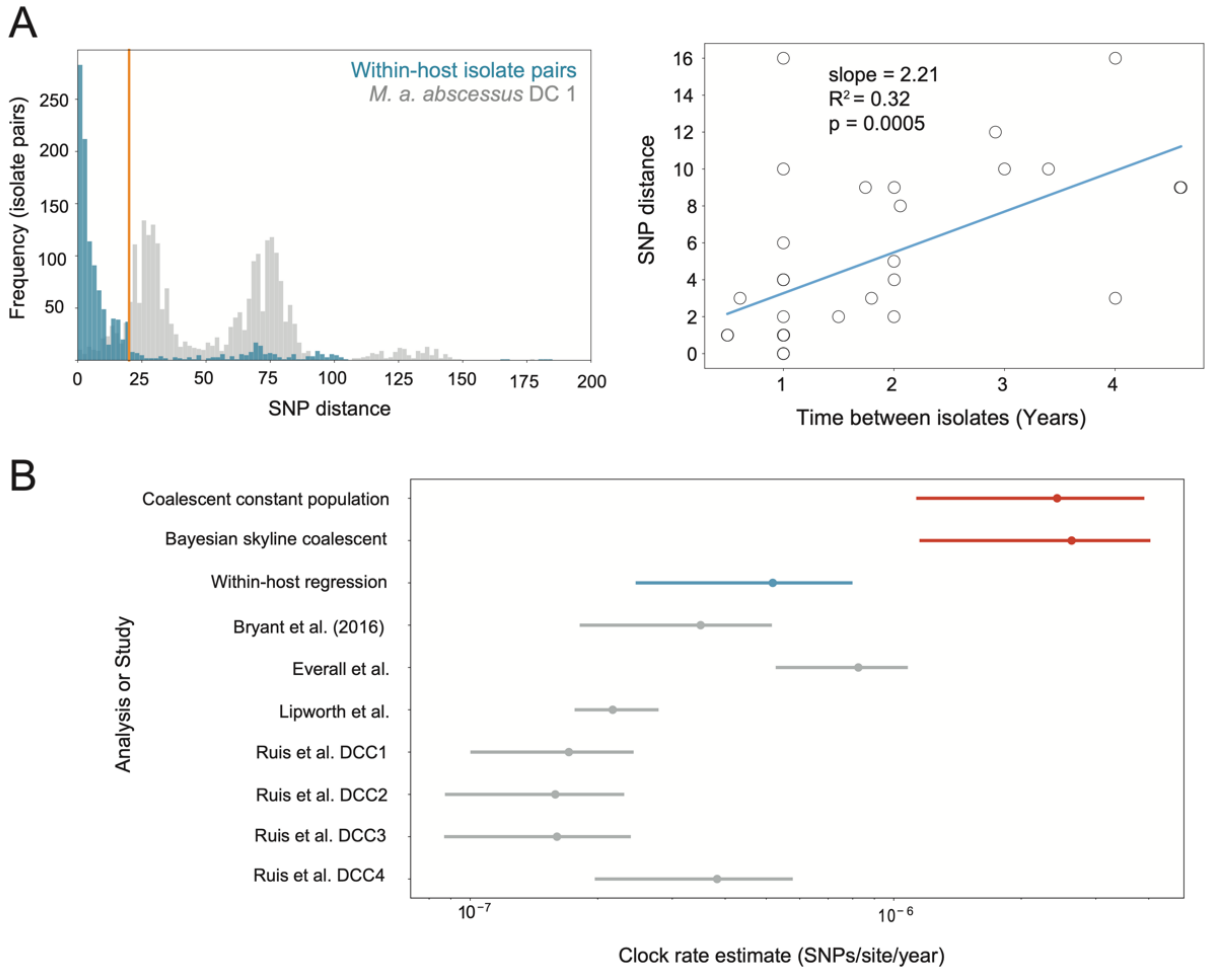


Figure 2. Longitudinally sampled within-host isolates and coalescent analysis support a slower clock rate within DCs. A) Clock rate estimation using isolate pairs sampled from the same patient over time. Left: distribution of SNP distances between all possible within-host isolate pairs compared to SNP distances between all possible isolate pairs within *M. a. abscessus* DC 1. The orange line represents the threshold of 20 SNPs used to exclude pairs representing independent infections. Right: regression of SNP distance on the time between isolates for each isolate pair. The slope of the regression line was 2.51 SNPs/year [1.24 – 3.78 95% CI], or roughly 5.93×10^{-7} SNPs/site/year [2.92×10^{-7} – 8.9×10^{-7} 95% CI]. B) UCLD clock rate estimates for the three models used in this study (two coalescent models shown in red and one regression model shown in blue) compared to mutation rate estimates from the within-host regression (blue) and published clock rates (grey).

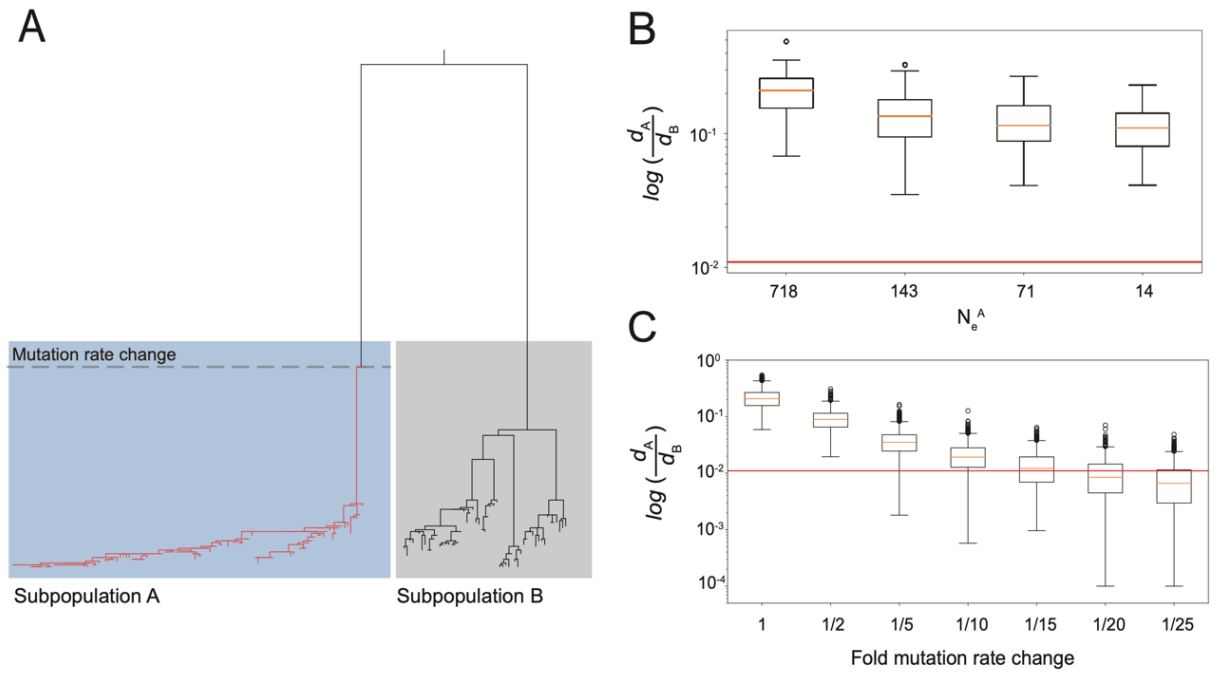
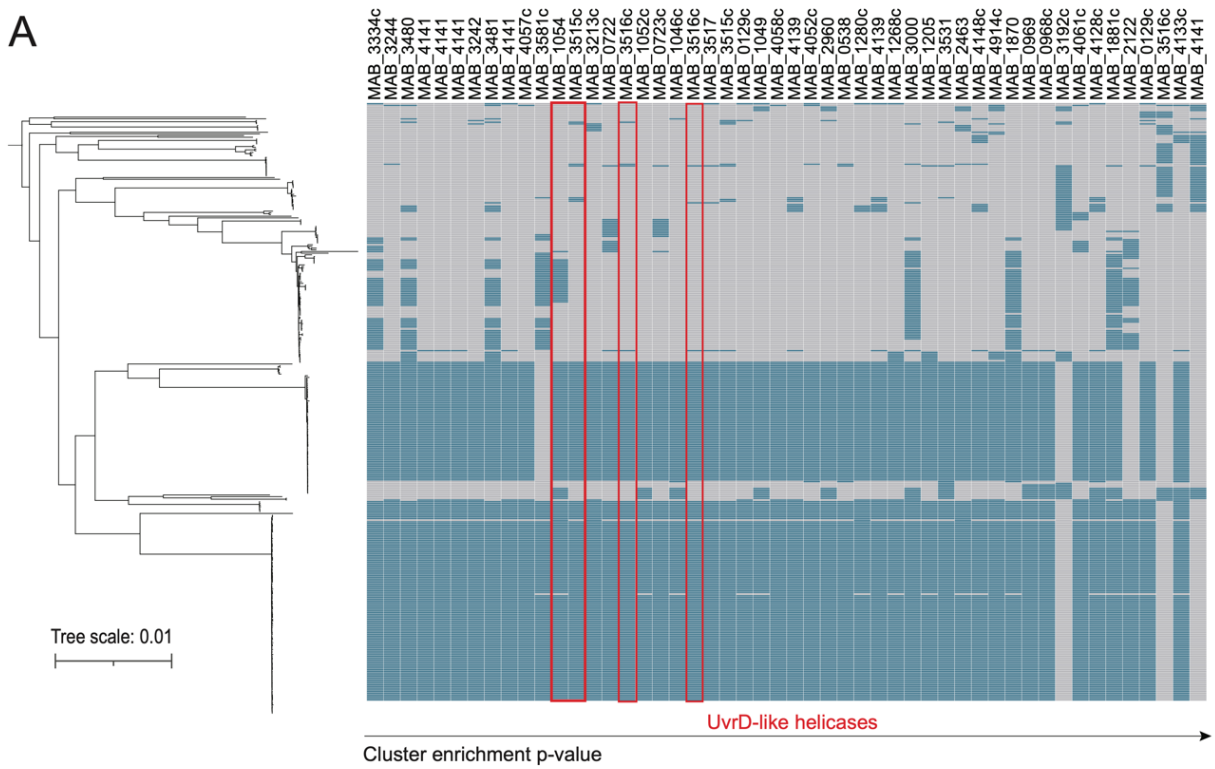
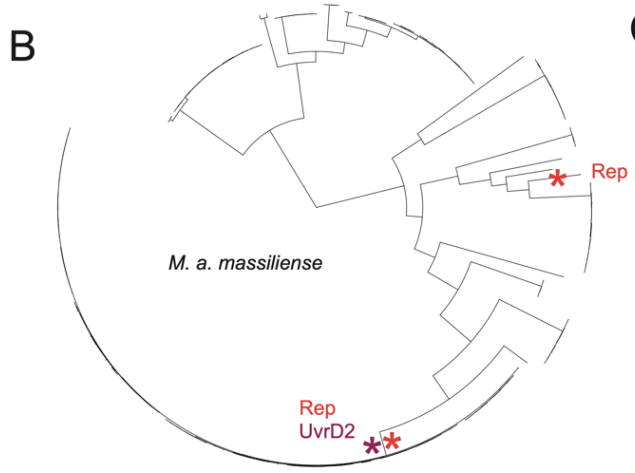
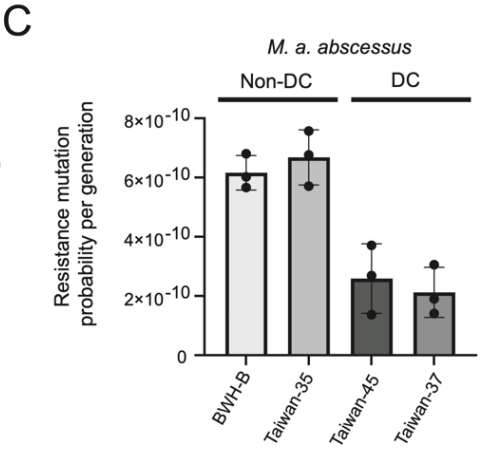
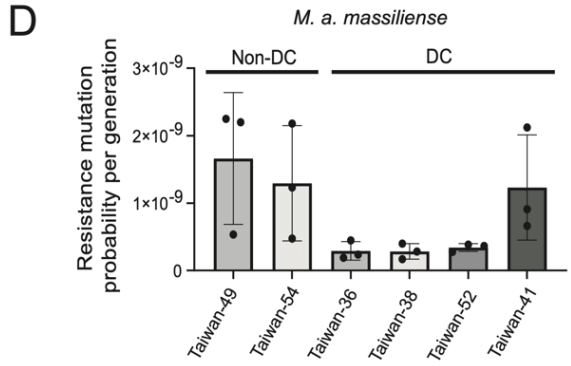


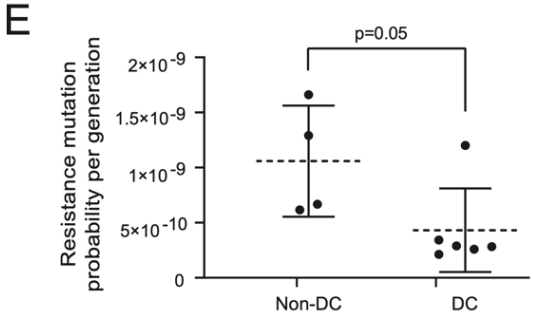
Figure 3. Simulated ancestries support a mutation rate slow-down coincident with DC emergence. A) Example neighbor joining (NJ) tree from the simulation shown in B. The number of samples and sampling dates for subpopulations A and B are drawn from real data from *M. a. abscessus* cluster 1 and from all unclustered samples, respectively. The dotted line crosses the common ancestor node for subpopulation A, and all branches descended from this node (shown here in red) are subject to the change in mutation rate. B) Boxplot showing the degree of phylogenetic clustering in subpopulation A relative to subpopulation B over a range of effective population sizes of subpopulation A (N_e^A). The simulation assumes that N_e^T=8000. C) Boxplot showing the degree of phylogenetic clustering in subpopulation A relative to subpopulation B over a range of mutation rate changes. The simulation assumes that N_e^T=8000 and N_e^A=718. In both A) and B) the degree of clustering is estimated with metric $\log(\frac{d_A}{d_B})$ where d_i is the mean pairwise SNP distance among all sample pairs in subpopulation i. The red line represents the clustering metric estimated using the observed ML phylogeny.


A
MAB_3334c
MAB_3244
MAB_3480
MAB_4141
MAB_4141
MAB_4141
MAB_3242
MAB_3481
MAB_4141
MAB_4057c
MAB_3581c
MAB_1054
MAB_3515c
MAB_3213c
MAB_0722
MAB_3516c
MAB_1052c
MAB_0723c
MAB_1046c
MAB_3516c
MAB_3517
MAB_3515c
MAB_0129c
MAB_1049
MAB_4058c
MAB_4139
MAB_4052c
MAB_2960
MAB_0538
MAB_1280c
MAB_4139
MAB_1268c
MAB_3000
MAB_1205
MAB_3531
MAB_2463
MAB_4148c
MAB_4914c
MAB_1870
MAB_0969
MAB_0968c
MAB_3192c
MAB_4061c
MAB_4128c
MAB_1881c
MAB_2122
MAB_0129c
MAB_3516c
MAB_4133c
MAB_4141
Tree scale: 0.01
UvrD-like helicases
Cluster enrichment p-value


B
Rep
M. a. massiliense
Rep
UvrD2


C
M. a. abscessus
Non-DC
DC
Resistance mutation probability per generation
8×10-10
6×10-10
4×10-10
2×10-10
0
BWH-B
Taiwan-35
Taiwan-45
Taiwan-37


D
M. a. massiliense
Non-DC
DC
Resistance mutation probability per generation
3×10-9
2×10-9
1×10-9
0
Taiwan-49
Taiwan-54
Taiwan-36
Taiwan-38
Taiwan-52
Taiwan-41


E
p=0.05
Resistance mutation probability per generation
2×10-9
1.5×10-9
1×10-9
5×10-10
0
Non-DC
DC

515

516 **Figure 4. Dominant clusters are associated with a decrease in the spontaneous**
517 **mutation rate *in vitro***. A) Subspecies phylogeny for *M. a. abscessus* (left) and corresponding heatmap representing the presence of the derived allele (blue) or ancestral
518 allele (grey) for each nonsynonymous variant. Variants are ranked in ascending order by their p-value derived from Fisher's tests for enrichment in the three largest clusters. The top 50
519 hits are shown. The genes where each variant is located are denoted. B) Subspecies phylogeny for *M. a. massiliense*. The branches where mutations have occurred in homologs
520 of Rep and UvrD2 are shown with asterisks. Probability of acquiring an amikacin resistance mutation per generation in C) *M. a. abscessus* strains, D) *M. a. massiliense* strains, or D) in
521 all strains tested that are part of a dominant cluster or not. Mean +/- SD is displayed. In C-D, n=3 biological replicates. In E, each point represents the average mutation probability for a
522 single strain. p-value derived from two-tailed, unpaired t-test.

523

524 **Methods**

525 Data Sources and Description: WGS data were curated from published studies[4,7,10,12,21–23] and
526 additional data were downloaded from the NCBI Sequence Read Archive (SRA). A search of
527 "abscessus" was conducted on October 4, 2019 using the filters "DNA" and "Illumina". Sequences
528 were selected that had an associated date of collection with the year at minimum. Collection dates
529 were obtained either from the literature, SRA metadata table, or provided by the authors. A
530 summary of all samples used in this study is provided in Table S1. Raw sequencing reads were
531 downloaded using the NCBI SRA toolkit[24]. Our procedure for downloading and processing
532 sequencing reads is available as a snakemake pipeline as well as all custom scripts are available
533 at https://github.com/nicolettacommins/mab_mutation_rates_2022.

534

535 Quality control of Illumina reads: Reads were trimmed and filtered with trimmomatic[25]. To detect
536 contaminated samples, the mean GC content of the reads were compared to a modelled normal
537 distribution of GC content using fastqc[26]. If the sum of the deviations from the normal distribution
538 represented more than 30% of the reads, the sequencing run was excluded from further analysis.

539

540 Subspecies assignment and genome assembly: Multiple sequencing runs corresponding to one
541 biological sample were combined into one file. Reads were then assembled *de novo* using

542 SPAdes[27]. Assemblies were compared to a reference sequence for each of the three subspecies
543 (Table S3) and a whole-genome based average nucleotide identity (gANI) score was calculated
544 using fastANI[28]. The resulting alignments were assigned to a subspecies based on having gANI
545 of at least 98% with exactly one reference strain. Isolates that did not have at least 98% ANI with
546 any reference strain were excluded. No isolates matched more than one reference genome. After
547 assigning isolates to a subspecies, reads were mapped to the corresponding reference genome
548 using BWA MEM[29]. After alignment, isolates with <80% coverage of the reference genome at a
549 depth of at least 20x or with missing data at >30% of sites were excluded.
550
551 Variant calling and SNP filtering: Variants were called using Pilon[30]. We excluded ambiguous calls
552 as wells as indels and MNVs. We also excluded calls with low coverage using a minimum depth
553 of either 10% of the mean coverage, or 5, whichever was greater (default settings for Pilon –
554 mindepth). We used bcftools[31] to exclude calls with mapping quality or base quality scores < 20.
555
556 Core genome inference: Unlike the professional mycobacterial pathogens *M. tuberculosis* or *M.*
557 *leprae*, which have small accessory genomes, *M. abscessus* has a large, open accessory
558 genome.[32] To define the core genome for phylogenetic analysis and to exclude loci that are highly
559 divergent from the reference sequence, we masked all loci for downstream analysis where read
560 depth was <20x in more than 5% of the isolates in our sample set. The remaining loci were
561 compared to an estimate of the core genome calculated by Roary[33], and the two methods yielded
562 nearly identical results. We further excluded regions with average mapping quality scores or
563 average based quality scores <20. Finally, we used GenMap[34] to calculate mappability scores
564 across each reference genome as defined in reference[35]. We then calculated a mappability pileup
565 score for each base position as described in reference[36] and excluded loci where the mappability
566 pileup score was <95% from downstream analysis. In total, we excluded 14.4%, 13.5%, and

567 15.1% of the *M. a. abscessus*, *M. a. massiliense*, and *M. a. bolletii* reference genome length,
568 respectively.
569
570 Phylogenetic tree inference: A full-length sequence alignment was created using a custom script,
571 masking loci excluded by our filtering criteria described above. Gubbins[37] was used to predict and
572 mask regions of variation produced by recombination from the full-length alignment. A
573 recombination-free SNP alignment generated by Gubbins was then used to build a phylogeny for
574 each subspecies using IQ-TREE[38], using the ModelFinder Plus (-mfp) option and 1000 bootstrap
575 replicates. Clusters were defined as previously described[4] using a one-dimensional scanning
576 statistic to identify clades with a distribution of branch lengths that is shorter than the distribution
577 of branch lengths in the rest of the tree[39].
578
579 Coalescent analysis: Temporal signal was assessed using root-to-tip regression. In both *M. a.*
580 *abscessus* and *M. a. massiliense* subspecies trees and all three major clusters, a slightly negative
581 slope was observed, indicated lack of temporal signal. We recognized that the high degree of
582 clustering and short time span over which the clustered isolates were sampled may confound the
583 genetic temporal signal. To address this, we pruned the trees down to taxa that preserved 95%
584 of the relative tree length (RTL) of the original trees using Treemmer[40]. A threshold of 95%
585 eliminated most of the dense clustering, but preserved the long branches of the original tree
586 (Figure 2B, Figure S2B). Using the pruned tree, we observed evidence of temporal signal in the
587 MAB subspecies using root-to-tip regression. We generated a SNP alignment including only the
588 taxa present in the pruned *M. a. abscessus* tree and used this as input for coalescent analysis
589 using BEAST2. ModelTest-NG[41] was used to select the site model GTR+4. MegaX[42] was used to
590 perform a maximum likelihood test to determine that a relaxed clock model is best suited for our
591 data. This choice was confirmed using the nested sampling algorithm to compare marginal
592 likelihoods for a strict v. relaxed clock model. We then used nested sampling to compare various

593 tree priors. We using Bayesian Evaluation of Temporal Signal (BETS) to confirm the presence of
594 temporal signal by comparing marginal likelihoods when the tip dates were used to those when
595 all samples were assumed to be isochronous (Table S4). Based on marginal likelihood values,
596 we ran both a coalescent constant population model and a Bayesian skyline coalescent model,
597 assuming a relaxed clock and including the sampling dates for both models. We used the ML
598 phylogeny of the pruned dataset as a starting tree. For each model we ran two independent chains
599 for at least 100 million states to ensure convergence.
600
601 Phylogenetic dating: To attempt to date the full MAB phylogeny, we ran both coalescent constant
602 population and Bayesian skyline coalescent models assuming a relaxed clock without sampling
603 dates. We set the mean UCLD clock rate to the values estimated from our coalescent analysis of
604 the pruned tree. We ran each chain for at least 100 million states.
605
606 Estimation of mutation rate within-host: We identified 90 patients with two or more *M. a. abscessus*
607 isolates. Because coinfection with multiple *M. abscessus* clones is common in cystic fibrosis
608 patients, we first sought to exclude any isolate pairs that represented distinct clonal infections.
609 We selected a threshold of 20 SNPs by comparing the distribution of all possible within-host
610 isolate pairs to the distribution of pairwise SNP distances within *Mab* clusters. We further excluded
611 any isolates where >20% of the genome is predicted to be rearranged by recombination. For the
612 remaining isolate pairs we regressed the pairwise SNP distance between the first and last isolates
613 sampled from each patient on the time between the collection of each sample. The slope of the
614 regression line was used as an estimate of the number of SNPs accumulated over time.
615
616 Ancestry simulations: We used msprime[43] to simulate ancestries to model two subpopulations:
617 (1) a “clustered” subpopulation (A) and (2) an “unclustered” subpopulation (B). We assumed a
618 constant population size and initially set the total effective population size (N_e) to 8000, the N_e

619 estimated in our coalescent analysis of the pruned *Mab* tree. We chose a range of realistic N_e for
620 subpopulation A, N_e^A by running a separate coalescent analysis of *M. a. abscessus* cluster 1 using
621 BEAST, assuming a constant population size, relaxed clock, and a range of UCLD clock rates
622 spanning 1.7×10^{-7} SNPs/site/year, as previously reported for this cluster in ref[13], to 8.5×10^{-6}
623 SNPs/site/year, a value well above the 95% HPD interval for our estimate of the clock rate in the
624 pruned tree (Table S5). We assumed the N_e of subpopulation B to be: $N_e^B = N_e^{total} - N_e^A$. The number
625 of samples and sampling dates supplied to the model for subpopulations A and B are drawn from
626 the true values for *Mab* cluster 1 and from all "unclustered" isolates in the tree, respectively. We
627 assumed an initial mutation rate of 1×10^{-9} mutations/generation with a sequence length of 5Mb
628 and a divergence time of 30,000 generations between subpopulations A and B. We also assumed
629 the nucleotide subsitution model estimated using BEAST.

630 We first simulated a scenario where the mutation rate remains constant, varying N_e^A over
631 the range shown in Table S5. For each simulated ancestry we overlaid mutation events, supplying
632 the model with a starting substitution rate of 1×10^{-9} mutations/generation, a nucleotide substitution
633 model estimated using BEAST, and a sequence length of 5Mb. We then extracted the simulated
634 genotypes and constructed a neighbor joining (NJ) tree. The NJ tree was used to exclude any
635 simulations where the two subpopulations did not exhibit reciprocal monophyly. We used the
636 extracted genotypes to estimate the degree of clustering in each tree by comparing the mean
637 pairwise SNP distance in subpopulation A (d_A) to the mean pairwise SNP distance in
638 subpopulation B (d_B). For each simulated condition we performed 1000 replicates. To simulate
639 ancestries with a mutation rate change, we repeated the above procedure conservatively
640 assuming $N_e^A = 718$ (Table S5). We modeled a per-generation mutation rate change starting at
641 the common ancestor node for subpopulation A and ranging from 1-fold (no change) to 1/25-fold
642 spanning a range of reasonable rate changes based on the range of clock rate estimates
643 observed in this study and reported in the literature (Figure 2B). For all simulated conditions, we
644 the simulation over a range of values for N_e^{total} spanning the 95% HPD intervals for the N_e

645 estimates using the Treemmer tree under both the constant coalescent and coalescent skyline
646 tree priors.
647
648 Scan for variants enriched in host-adapted lineages: To search for variants associated with host-
649 adapted MAB lineages, we first inferred the number of independent arisals in the MAB phylogeny
650 for each SNP using SNPPar[44]. Briefly, SNPPar uses TreeTime to perform an ancestral sequence
651 reconstruction at each node, then infers mutation events arising on each branch of the tree. We
652 quantified the number of independent mutation events for each SNP. We considered only
653 mutation events where the derived call was the minor allele and the ancestral call was the major
654 allele. Because of the extreme phylogenetic clustering in *Mab*, we defined the major and minor
655 alleles using the pruned alignment generated by Treemmer. For all homoplasic SNPs (occurring
656 two or more times independently in the tree), we used Fisher's exact test to quantify the
657 enrichment of that variant within phylogenetic clusters. We conducted two separate analyses: a)
658 focusing on the three largest clusters in the *M. a. abscessus* tree (described in the main text), and
659 b) considering all clusters identified by this approach. In each case we obtained similar results.
660 For both analyses, full lists of variants ranked by their enrichment scores are provided in
661 Supplementary Data. We also used SNPPar to identify all mutation events occurring on the
662 branch ancestral to *M. a. massiliense* cluster 1.
663
664 *M. abscessus* culture: *M. abscessus* clinical isolates used for fluctuation assays were isolated
665 from patients in Taiwan and in Boston, MA, USA. All *M. abscessus* strains were cultured in
666 Middlebrook 7H9 broth (271310, BD Diagnostics, Franklin Lakes, NJ, USA) with 0.2% (v/v)
667 glycerol, 0.05% (v/v) Tween-80 (P1754, Sigma-Aldrich, St. Louis, MO, USA), and 10% (v/v)
668 OADC (90000-614, VWR, Radnor, PA, USA). Proliferation rates were determined by inoculating
669 5 mL media with 500,000 colony forming units (cfu), then measuring OD600 at 24 and 48 hours

670 post-inoculation. Proliferation rate was calculated as follows: proliferation rate in doublings per
671 day = $\log_2$(OD600 at 48 hours / OD600 at 24 hours).
672
673 Fluctuation Assay: Fluctuation assays were performed as described in reference[17] with
674 adaptations for *M. abscessus*. *M. abscessus* strains were grown to saturation, then diluted to an
675 OD600 < 0.01 and grown overnight until cultures reached an OD600 = 0.6-0.8. These cultures
676 were used for fluctuation assays as well as determination of baseline amikacin sensitivity.
677 Baseline amikacin sensitivity was tested by spreading dilutions of 10 million, 100 million, or 1
678 billion cfu on agar 7H10 (262710, BD Diagnostics) + OADC (90000-614, VWR) + 300 µg/mL
679 amikacin sulfate (A2324, Sigma-Aldrich) plates. For fluctuation assays, 5000 cfu were plated in 1
680 mL media in each of 21 wells of a 24-well culture plate (10861-558, VWR), and initial cultures
681 were plated for cfu on 7H10 + OADC to confirm number of cells in initial culture. 24-well plates
682 were incubated shaking at 150 rpm for 6 days at 37°C. 18 wells of each strain were plated onto
683 6-well plates containing 5 mL 7H10 agar + OADC + 300 µg/mL amikacin sulfate in each well.
684 Amikacin was used as the antibiotic selection for the assay because all tested strains had low
685 baseline resistance (Table S8), which is required for the validity of the fluctuation assay. The
686 remaining 3 wells from each strain were diluted and plated on 7H10 + OADC plates to enumerate
687 average final cfu. Colonies on all agar plates were counted after 5 days incubation at 37°C.
688 Mutation rate per generation per total number of bases that can be mutated to confer amikacin
689 resistance was calculated using the Shinyflan application for the flan R package[17] using the
690 Maximum Likelihood estimation method, assumption of unknown fitness effects of mutations, the
691 Exponential (LD model) distribution of mutant lifetime, and a Winsor parameter of 1024. One
692 strain (BWH-E) was excluded from analysis due to highly variable results across replicates (Figure
693 S6B).
694
695

696 **References**

697 1. Adjemian, J., Olivier, K. N. & Prevots, D. R. Epidemiology of Pulmonary Nontuberculous
698 Mycobacterial Sputum Positivity in Patients with Cystic Fibrosis in the United States, 2010-
699 2014. *Ann. Am. Thorac. Soc.* **15**, 817–826 (2018).

700 2. Lee, M.-R. *et al.* Mycobacterium abscessus Complex Infections in Humans. *Emerg. Infect.*
701 *Dis.* **21**, 1638–1646 (2015).

702 3. Nessar, R., Cambau, E., Reyrat, J. M., Murray, A. & Gicquel, B. Mycobacterium abscessus:
703 a new antibiotic nightmare. *J. Antimicrob. Chemother.* **67**, 810–818 (2012).

704 4. Bryant, J. M. *et al.* Emergence and spread of a human-transmissible multidrug-resistant
705 nontuberculous mycobacterium. *Science* **354**, 751–757 (2016).

706 5. Bange, F. C., Brown, B. A., Smaczny, C., Wallace, R. J., Jr & Böttger, E. C. Lack of
707 transmission of mycobacterium abscessus among patients with cystic fibrosis attending a
708 single clinic. *Clin. Infect. Dis.* **32**, 1648–1650 (2001).

709 6. Olivier, K. N. *et al.* Nontuberculous mycobacteria. I: multicenter prevalence study in cystic
710 fibrosis. *Am. J. Respir. Crit. Care Med.* **167**, 828–834 (2003).

711 7. Doyle, R. M. *et al.* Cross-transmission is not the source of new Mycobacterium abscessus
712 infections in a multi-centre cohort of cystic fibrosis patients. *Clin. Infect. Dis.* (2019)
713 doi:10.1093/cid/ciz526.

714 8. Sermet-Gaudelus, I. *et al.* Mycobacterium abscessus and children with cystic fibrosis. *Emerg.*
715 *Infect. Dis.* **9**, 1587–1591 (2003).

716 9. Lipworth, S. *et al.* Epidemiology of Mycobacterium abscessus in England: an observational
717 study. *The Lancet Microbe* (2021) doi:10.1016/S2666-5247(21)00128-2.

718 10. Harris, K. A. *et al.* Whole-genome sequencing and epidemiological analysis do not provide
719 evidence for cross-transmission of mycobacterium abscessus in a cohort of pediatric cystic
720 fibrosis patients. *Clin. Infect. Dis.* **60**, 1007–1016 (2015).

721 11. Tortoli, E. *et al.* Mycobacterium abscessus in patients with cystic fibrosis: low impact of inter-
722 human transmission in Italy. *Eur. Respir. J.* **50**, (2017).
723 12. Everall, I. *et al.* Genomic epidemiology of a national outbreak of post-surgical Mycobacterium
724 abscessus wound infections in Brazil. *Microb Genom* **3**, e000111 (2017).
725 13. Ruis, C. *et al.* Dissemination of Mycobacterium abscessus via global transmission networks.
726 *Nature Microbiology* 1–10 (2021).
727 14. Bouckaert, R. *et al.* BEAST 2.5: An advanced software platform for Bayesian evolutionary
728 analysis. *PLoS Comput. Biol.* **15**, e1006650 (2019).
729 15. Houghton, J. *et al.* Important role for Mycobacterium tuberculosis UvrD1 in pathogenesis and
730 persistence apart from its function in nucleotide excision repair. *J. Bacteriol.* **194**, 2916–2923
731 (2012).
732 16. Sinha, K. M., Unciuleac, M.-C., Glickman, M. S. & Shuman, S. AdnAB: a new DSB-resecting
733 motor-nuclease from mycobacteria. *Genes Dev.* **23**, 1423–1437 (2009).
734 17. Krašovec, R. *et al.* Measuring Microbial Mutation Rates with the Fluctuation Assay. *J. Vis.*
735 *Exp.* (2019) doi:10.3791/60406.
736 18. Curti, E., Smerdon, S. J. & Davis, E. O. Characterization of the helicase activity and substrate
737 specificity of Mycobacterium tuberculosis UvrD. *J. Bacteriol.* **189**, 1542–1555 (2007).
738 19. Ho, S. Y. W., Shapiro, B., Phillips, M. J., Cooper, A. & Drummond, A. J. Evidence for time
739 dependency of molecular rate estimates. *Syst. Biol.* **56**, 515–522 (2007).
740 20. Huffnagle, G. B., Dickson, R. P. & Lukacs, N. W. The respiratory tract microbiome and lung
741 inflammation: a two-way street. *Mucosal Immunol.* **10**, 299–306 (2017).
742 21. Bryant, J. M. *et al.* Whole-genome sequencing to identify transmission of Mycobacterium
743 abscessus between patients with cystic fibrosis: a retrospective cohort study. *Lancet* **381**,
744 1551–1560 (2013).
745 22. Hasan, N. A. *et al.* Population Genomics of Nontuberculous Mycobacteria Recovered from
746 United States Cystic Fibrosis Patients. *bioRxiv* 663559 (2019) doi:10.1101/663559.

747 23. Shaw, L. P. *et al.* Children with cystic fibrosis are infected with multiple subpopulations of
748 Mycobacterium abscessus with different antimicrobial resistance profiles. *Clin. Infect. Dis.*
749 (2019) doi:10.1093/cid/ciz069.
750 24. Download : Software : Sequence Read Archive : NCBI/NLM/NIH.
751 https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software.
752 25. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence
753 data. *Bioinformatics* **30**, 2114–2120 (2014).
754 26. Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence
755 Data. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
756 27. Prjibelski, A., Antipov, D., Meleshko, D., Lapidus, A. & Korobeynikov, A. Using SPAdes De
757 Novo Assembler. *Curr. Protoc. Bioinformatics* **70**, e102 (2020).
758 28. Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T. & Aluru, S. High throughput
759 ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat. Commun.*
760 **9**, 5114 (2018).
761 29. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.
762 *arXiv [q-bio.GN]* (2013).
763 30. Walker, B. J. *et al.* Pilon: an integrated tool for comprehensive microbial variant detection
764 and genome assembly improvement. *PLoS One* **9**, e112963 (2014).
765 31. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and
766 population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–
767 2993 (2011).
768 32. Choo, S. W. *et al.* Genomic reconnaissance of clinical isolates of emerging human pathogen
769 Mycobacterium abscessus reveals high evolutionary potential. *Sci. Rep.* **4**, 4061 (2014).
770 33. Page, A. J. *et al.* Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics*
771 **31**, 3691–3693 (2015).

772 34. Pockrandt, C., Alzamel, M., Iliopoulos, C. S. & Reinert, K. GenMap: ultra-fast computation of
773 genome mappability. *Bioinformatics* **36**, 3687–3692 (2020).
774 35. Derrien, T. *et al.* Fast computation and applications of genome mappability. *PLoS One* **7**,
775 e30377 (2012).
776 36. Marin, M. *et al.* Benchmarking the empirical accuracy of short-read sequencing across the
777 M. tuberculosis genome. *Bioinformatics* (2022) doi:10.1093/bioinformatics/btac023.
778 37. Croucher, N. J. *et al.* Rapid phylogenetic analysis of large samples of recombinant bacterial
779 whole genome sequences using Gubbins. *Nucleic Acids Res.* **43**, e15 (2015).
780 38. Minh, B. Q. *et al.* IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference
781 in the Genomic Era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
782 39. Harris, S. TreeGubbins. (2016).
783 40. Menardo, F. *et al.* Treemmer: a tool to reduce large phylogenetic datasets with minimal loss
784 of diversity. *BMC Bioinformatics* **19**, 164 (2018).
785 41. Darriba, D. *et al.* ModelTest-NG: A New and Scalable Tool for the Selection of DNA and
786 Protein Evolutionary Models. *Mol. Biol. Evol.* **37**, 291–294 (2020).
787 42. Kumar, S., Stecher, G., Li, M., Knyaz, C. & Tamura, K. MEGA X: Molecular Evolutionary
788 Genetics Analysis across Computing Platforms. *Mol. Biol. Evol.* **35**, 1547–1549 (2018).
789 43. Kelleher, J. & Lohse, K. Coalescent Simulation with msprime. in *Statistical Population*
790 *Genomics* (ed. Dutheil, J. Y.) 191–230 (Springer US, 2020).
791 44. Edwards, D. J., Duchêne, S., Pope, B. & Holt, K. E. SNPPar: identifying convergent evolution
792 and other homoplasies from microbial whole-genome alignments. *bioRxiv*
793 2020.07.08.194480 (2020) doi:10.1101/2020.07.08.194480.