Evaluating selection at intermediate scales within genes provides robust identification of genes under positive selection in *M. tuberculosis* clinical isolates

Thomas R. Ioerger* and Anthony Shatby
Department of Computer Science and Engineering
Texas A&M University

* Corresponding author (ioerger@cs.tamu.edu)

Keywords: adaptation; selection pressures; drug resistance; genetic variants; clinical isolates

Abstract

Multiple studies have reported genes in the *M. tuberculosis* (Mtb) genome that are under diversifying selection, based on genetic variants among Mtb clinical isolates. These might reflect adaptions to selection pressures associated with modern clinical treatment of TB. Many, but not all, of these genes under selection are related to drug resistance. Most of these studies have evaluated selection at the gene-level. However, positive selection can be evaluated on different scales, including individual sites (codons) and local regions within an ORF. In this paper, we use GenomegaMap, a Bayesian method for calculating selection, to evaluate selection of genes in the Mtb genome at all three levels. We present evidence that the intermediate analysis (windows of codons) provides the most credible list of candidate genes under selection. A further advantage of this approach is that it identifies specific regions within proteins that are under selective pressure, which is useful for structural and functional interpretation. In an analysis of two separate collections of Mtb clinical isolates (from Moldova; and a globally-representative set), we observed 53 and 178 significant genes under selection, with 36% overlap. The lists of genes under selection include many drug-resistance genes, as well as other genes that have previously been reported to be under selection (*resR, phoR*). The specific regions under selection identified within drug-resistance genes are shown to correspond to protein structural features known to be involved in resistance, supporting accuracy of the method. Positive selection in several ESX-1-related genes was also observed, suggesting adaptation to immune pressure.

1. INTRODUCTION

Recent studies examining large collections of genome sequences of *M. tuberculosis* (Mtb) clinical isolates have reported the identification of genes that appear to be under positive selection [1-3]. Traditionally, positive selection is indicated by an excess of non-synonymous over synonymous mutations (in protein-coding regions), although clustering of SNPs (e.g. in non-coding regions), homoplasy [4], and other measures of convergent evolution [5] have also been used in the literature. This is taken to imply on-going evolution and adaptation to modern chemotherapy, including pressures due to antibiotic exposure, as well as adaptive mutations. This is evidenced by the fact that genes associated with drug resistance are often found to have among the highest selection signals [6]. For example, genes directly involved as drug targets or activators (like *rpoB* or *katG*) exhibit selection, as well as genes with indirect effects, like efflux pumps (*tap, mmr, mmpL5*) and regulators (*whiB6, resR, phoR*) that might confer tolerance through changes in cell wall permeability, metabolism, etc. In addition, selection signals sometimes appear in genes due to compensatory mechanisms, where mutations in one gene help reduce the fitness costs of mutations in another [7].

Evaluating selection in ORFs (open reading frames) is usually done within a Bayesian framework, where observed mutations are used to infer posterior probability distributions over mutation rates that would best explain the observations. dN/dS is not just the ratio of raw counts of non-synonymous (NS) to synonymous (S) mutations, but is corrected for a) relative number of possibilities of NS and S changes for each codon (typically twice as many NS as S), and b) for evolutionary time [8]. Positive selection is defined as dN/dS>1. First, a null model is postulated where mutations occur randomly throughout an ORF and accumulate (exponentially) over time. Models such as NY98 [9] use a parameterized codon substitution matrix for rates of substitution between each pair of aligned codons, i and j, that differentiates transition vs transversion rates (using a latent variable κ), and synonymous vs non-synonymous changes (using a latent variable omega, ω), relative to background codon frequencies π_j:

$$Q_{ij} = \begin{cases} 0 \;\; if\ codons\ differ\ by > 1\ nucleotide \\ \pi_j\ for\ synonymous\ transversion \\ \kappa\pi_j\ for\ synonymous\ transitions \\ \omega\pi_j\ for\ non-synonymous\ transversion \\ \omega\kappa\pi_j\ for\ non-synonymous\ transition \end{cases}$$

[Eqn 1]

Over an epoch of evolutionary time *t*, the substitution matrix relaxes by the following equation: $P(t) = e^{Qt}$. The likelihood function for observed allele frequencies, f_i, is assumed to be *Multinomial*(f_i ; *P(t)*). Integrating out *t*, the marginal distribution for allele frequencies among extant sequences can be shown to follow a multinomial-Dirichlet distribution [10]. Observed mutations are mapped onto a phylogenetic tree to identify unique evolutionary events, and then the relative number of synonymous and nonsynonymous events is used to compute posterior distributions over the parameters (including ω) under different assumptions (such as whether there is a constant molecular clock, or rates can differ independently in each branch).

PAML [11] is a widely-used tool for evaluating selection in sequence data, which uses maximum-likelihood (ML) to estimate the model parameters. However, PAML does not scale-up well, and is difficult to run for collections of thousands of genomes, preventing its use in studies such as [1]. The complexity is derived from the requirement of large-scale phylogenetic trees for the computations. Several alternative approaches for evaluating selection have been proposed that are more efficient, enabling their use in analysis of large-scale collections. One such tool is GenomegaMap [10]. GenomegaMap does not require a phylogenetic tree, aggregating SNPs without regard for when the mutations occurred and in which branches, but otherwise relies on a very similar Bayesian framework [Eqn 1] to explain the relative proportion of NS to S mutations in an ORF. GenomegaMap uses Markov Chain Monte Carlo (MCMC) methods (described in more detail below) to infer selection parameters for mutations observed in genes. GenomegaMap was in fact applied to analyze >10,000 Mtb genome sequences from the CRyPTIC collection (demonstrating its scalability) and proposed a list of genes under positive selection, which included many drug-resistance-related genes (consistent with expectations).

As with PAML, GenomegaMap implements several different models for evaluating selection at different levels of granularity. In addition to the gene-level analysis (referred to as the Constant model in GenomegaMap), where all mutations in an ORF are aggregated, selection can also be evaluated at the level of individual codons (referred to as the Individual model). Importantly, these two models have different sensitivities for detecting selection. While analysis of selection in individual codons might seem to have the highest resolution (signal-to-noise ratio), this analysis suffers from the inability to exploit information at multiple nearby sites in a region (such as in an active site, or protein-protein interaction site). On the other hand, while aggregating SNPs at the gene-level can take advantage by combining mutations at multiple sites within a gene, the signal for selection can be diluted by other unrelated mutations occurring in other less-conserved parts of the protein.

An intermediate approach would be to evaluate selection signals in localized regions throughout an ORF. Mutations that can similarly impact the function of a protein often involve multiple residues that cluster together in sequence (and hence in 3D structure), such as residues in an active site or a site of protein-protein interaction. This then produces the appearance of clusters of non-synonymous mutations in multiple alignments of genes under selection. To evaluate selection in such clustered regions, GenomegaMap offers a third mode, called the "Window" model, in which an ORF is divided into several contiguous segments, and an independent ω parameter is used to model selection in each region. This model assumes selection pressure will be constant among codons within each region, but can differ between regions within an ORF. Importantly, the number and boundaries of these regions are themselves parameters of the model, and can be adjusted dynamically to maximize fit to mutation data. The advantage of this Window approach is that it produces a "smoothed" selection signal (over adjacent codons) by identifying regions within the ORF where there is a cluster of SNPs with a localized over-representation of non-synonymous to synonymous SNPs.

GenomegaMap uses MCMC sampling to simulate posterior probability distributions over all model parameters by generating a sequence of parameter combinations from a Markov chain; in the end, the mean value and credible interval for each parameter over the trajectory can be extracted (after removing an initial "burn-in" portion of the trajectory).

112 Comparing the quality of model fits based on total likelihood, Wilson [10] concluded that the
113 Individual model (identifying genes with individual codons under positive selection) fit the data
114 better than the Window model. However, using a criterion of Pr(ω>1)>0.9 (dN/dS>1 in >90% of
115 MCMC samples), thousands of genes were found to have at least one codon under selection
116 (see Fig 3 in Wilson). While many drug-resistance-related genes were included, when genes
117 were ranked by the proportion of codons under selection, some expected genes, like *gyrA, rpoB,*
118 and *embB*, were found to have ranks in the range of 2000-3000 (out of ~4000 genes). For the
119 purposes of prioritization, it would be expected that such genes would be ranked closer to the
120 top.

121 In this paper, we empirically compare the three different models of selection – at the gene-level,
122 codon-level, and in local windows (gene segments) – by evaluating them on analysis of a set of
123 ~2000 Mtb clinical isolates from Moldova [12]. While thousands of genes have ω>1 (as a point
124 estimate), determining which genes are statistically significant depends critically on the MCMC
125 sampling (through credible intervals). We show that using tighter criteria for significance than
126 were used in [10] produces a more focused list of candidates for genes under positive selection.
127 While many expected genes that have been shown to be associated with drug resistance appear
128 on both lists, we show that the Window model often better identifies regions within these
129 proteins that are known to be antibiotic binding sites based on X-ray crystal structures. In
130 contrast, we show that the Individual model has a bias toward identifying more codons under
131 selection in large ORFs, suggesting it is more susceptible to noise. Based on these observations,
132 we argue that the Window model in GenomegaMap is the most reliable approach for
133 identifying genes under selection, which results from the smoothing effect achieved by
134 modeling localized variations in ω within regions of a protein (with adjustable borders). As
135 validation, we show that, a similar list of genes under positive selection is derived from an
136 independent set of over 5,000 Mtb clinical isolates from 15 other countries (from the CRyPTIC
137 collection), also using the Window model.

138 The data on positive selection and genetic variants in both collections of clinical isolates
139 (Moldova and CRyPTIC) is available online (https://orca1.tamu.edu/selection/) as a public
140 resource.

141

142 **2, METHODS**

143 Raw sequencing data (fastq files) for isolates from Moldova were downloaded from SRA using
144 accession numbers provided in [12]. Sequencing datasets from the CRyPTIC collection were
145 downloaded from Supp Table 3 in [10]. While isolates could not be selected based on lineage
146 (data not available), to promote genetic diversity, we performed stratified sampling by selecting
147 approximately 500 isolates each from 15 different countries. TB-profiler [13] was run on the
148 isolates to identify Mtb lineage, drug-resistance genotypes, and estimated depth of coverage
149 (using *M. tuberculosis* H37Rv as a reference). Isolates that had < 40x coverage or were
150 potentially mixed (with more than one sublineage represented, based on characteristic SNPs)
151 were discarded.

152 Next, Pilon [14] was used to map read to the H37Rv genome (NC_000962.3_ and extract genetic
153 variants (in the form of .vcf files), including SNPs and indels; Pilon also identifies low-coverage
154 and heterogeneous sites. From the .vcf files, the nucleotide sequences of each ORF were
155 extracted from each isolate, and aggregated by ORF (as .fasta files). Codons with low-coverage
156 or heterogeneous nucleotides, or nonsense mutations (stop codons) were converted to gap
157 characters ('---'), which are subsequently disregarded in the calculations by GenomegaMap.

158 GenomegaMap was then run on each ORF using scripts the for Constant, Individual, and
159 Window models (https://github.com/danny-wilson/genomegaMap). For the MCMC sampling
160 for Individual and Window, the trajectory length was 100,000 iterations, saving parameters
161 every 250 steps, and eliminating the first 20% of samples as the burn-in phase (to reduce the
162 influence of initial random parameters and ensure the sampler had reached equilibrium). After
163 the MCMC sampling, the value of ω for each codon was summarized over each trajectory by
164 extracting the mean and other statistics (like range and variance), and a 95% credible interval
165 (95%CI) was calculated for the ω value for each codon.

166 129 PPE and PE_PGRS genes were systematically excluded from the analysis, as they harbored
167 apparent mutations that caused many genes in these two families to exhibit selection signals,

168 but these might be artifacts due to uncertainties in sequencing associated with these genes
169 [15].

170 Generation of simulated genes for comparison of GenomegaMap models under the assumption
171 of neutral evolution was done as follows. First, a phylogenetic tree based on the 2057 Moldova
172 clinical isolates was constructed using RAxML [16]. Then, the least common ancestral sequence
173 (LCA) was computed for each ORF in the Mtb genome (excluding PPE and PE_PGRS genes, and
174 genes known to be involved in drug resistance). The LCA nucleotide at each site in a given gene
175 was derived by using Fitch's algorithm [17] to propagate observed nucleotides from leaves
176 (2057 isolates) up to the root. Then, each distinct mutational event (i.e. SNP associated with a
177 particular branch point in the tree) was randomly assigned to a new site within a new gene.
178 This resampling process resulted in 39,564 mutations permuted to new positions among
179 3,446,666 sites within coding regions of 3877 genes (with the same size distribution as in the
180 Mtb genome), which was then used to generate the 2057 variant sequences for each isolate for
181 each gene.

182

183 **3. RESULTS**

184 We used GenomegaMap to investigate genes under positive selection in a collection of 2236
185 Mtb clinical isolates from Moldova, representing all reported cases in 2018-2019 [12]. The aim
186 of the study to identify transmission clusters within the region. The isolates were found to be
187 predominantly members of lineage 2 and lineage 4, and approximately one-half of the isolates
188 were (genotypically) drug-resistant (based on DR markers to one or more antitubercular drugs),
189 with 32.7% MDR-TB (INH+RIF) (lineage and DR distributions are shown in Supp Fig S1). After
190 downloading and preprocessing the raw sequencing data and filtering out low-quality samples,
191 2057 isolates were selected for analysis.

192 **3.1. Comparison of genes under positive selection by 3 models with GenomegaMap.**

193 With the Window model (where different regions within each ORF are modeled with locally-
194 constant omega values), the MCMC simulation produces a sample of parameter combinations,

195 including a representation of the posterior probability distribution of omega values for each
196 codon. Besides the mean value of omega for each codon, the dispersion of the sample can be
197 used to assess confidence. In [10], the criterion used for significance was if ≥ 90% of the
198 samples had ω >1. This corresponds to a (two-sided) 80% credible interval, which is overly
199 generous, potentially generating many false positive codons and genes under selection (245
200 genes by this criterion). Instead, we employ a more stringent criterion of ≥ 97.5% of samples
201 having ω >1, equivalent to a more standard (two-sided) 95% credible interval. Using this
202 criterion, 53 genes were identified as having at least one codon under positive selection (Table
203 1). (dN/dS scores and statistical significance for all genes are provided in Supplemental Table
204 T1.)

205

206 Table 1. 53 genes with statistically significant evidence of positive selection in the Moldova
207 collection of Mtb clinical isolates. using the Window model analysis in GenomegaMap.

Rv0006/*gyrA*	Rv1286/*cysN*	Rv2258c/-	Rv2940c/*mas*
Rv0093c/-	Rv1288/-	Rv2388c/*hemN*	Rv3109/*moaA1*
Rv0229c/-	Rv1320c/-	Rv2395/-	Rv3132c/*devS*
Rv0275c/-	Rv1699/*pyrG*	Rv2407/-	Rv3299c/*atsB*
Rv0392c/*ndhA*	Rv1830/*resR*	Rv2447c/*folC*	Rv3452/*cut4*
Rv0667/*rpoB*	Rv1899c/*lppD*	Rv2544/*lppB*	Rv3695/-
Rv0668/*rpoC*	Rv1901/*cinA*	Rv2652c/-	Rv3736/-
Rv0758/*phoR*	Rv1934c/*fadE17*	Rv2779c/-	Rv3795/*embB*
Rv0931c/*pknD*	Rv2043c/*pncA*	Rv2825c/-	Rv3806c/*ubiA*
Rv0969/*ctpV*	Rv2048c/*pks12*	Rv2874/*dipZ*	Rv3822/-
Rv0988/-	Rv2082/-	Rv2881c/*cdsA*	Rv3828c/-
Rv1093/*glyA1*	Rv2176/*pknL*	Rv2885c/-	Rv3888c/-
Rv1130/*prpD*	Rv2252/-	Rv2934/*ppsD*	Rv3911/*sigM*
			Rv3919c/*gid*

208

209

210 The 53 genes under positive selection in the Moldova dataset (by the Window method) contain
211 many genes known to be involved in drug resistance, including *rpoB* (RNA polymerase, target of
212 rifampicin), *gyrA* (DNA gyrase, target of fluroquinolones, *gidB* (ribosome methyltransferase,
213 affects sensitivity to streptomycin), *pncA* (pyrazinamidase, activator of pyrazinamide), and *embB*

214 (target of ethambutol). These will be discussed in more detail below. *resR* (Rv1830) also
215 exhibits positive selection in this dataset, consistent with the findings in [1], where it was shown
216 to be associated with general resilience to antibiotics. *phoR* is a two-component sensor
217 implicated in stress response and virulence [18, 19] that has been shown to be under positive
218 selection in Mtb using phylodynamics (i.e. increase in non-synonymous mutation rate,
219 coincident with expansion into humans as pathogen) [20]. Other genes under selection that can
220 play a regulatory role and potentially adaptive role include *pknD* (protein Ser/Thr kinase), *devS*
221 (two-component sensor, regulator of hypoxic response [21]), *sigM* (a sigma factor that could
222 influence ESX-4 or PDIM expression [22]), and Rv1320 (an adenylate cyclase, which are often
223 involved in signaling/regulation). None of the three models identified *katG* (activator of
224 isoniazid) as positively selected in this collection, despite mutations in *katG* being the most
225 frequently observed in INH-resistant isolates. Wilson [10] also noted that *katG* was not found to
226 be under positive selection by the Window model in their larger dataset, but was identified by
227 the Independent model (because the majority of INH-resistance mutations occur in a single
228 codon: Ser315). However, this could also be easily identified by homoplasy analysis [4], as
229 mutations at this site often occur multiple times independently in clinical populations.

230 The evaluation of genes under positive selection using the Independent model yielded 44
231 significant genes, using the more stringent criterion of at least 1 codon with $\omega>1$ in ≥ 97.5% of
232 the MCMC samples. This list of 44 genes has some overlap with the 53 significant genes by the
233 Window model; there are 19 genes in the intersection, including most of the drug-resistance-
234 related genes (see Venn diagram in Figure 1a). However, there are also many unique genes
235 identified by each of the analysis methods. The Window model also highlighted *ubiA*
236 (arabinogalactan synthesis pathway; harbors resistance mutations to ethambutol) and *folC*
237 (resistance mutations to inhibitors of folate pathway like PAS or sulfamethoxazole), which the
238 Independent model did not.

239 With the Constant model (5000 MCMC iterations), 1924 genes have $\omega>1$, but only 61 are
240 significant (LB of 95%CI > 1). Two-thirds of these are PPE or PE_PGRS genes; after filtering these
241 out, 22 genes remain that exhibit statistically significant positive selection. The only drug-
242 resistance-related genes among these are *pncA, gidB,* and *ubiA*. Other DR genes like *rpoB, gyrA,*

243 *embB,* and *katG* did not reach significance using this model. Many of the significant genes by
244 the Window model have ω below or only marginally above 1 in the Constant model, including
245 many of the genes implicated in drug resistance mentioned above (see Figure 1b). However,
246 the lower bound of the 95% credible interval is less than 1.0 for almost all of the 53 genes
247 significant by the Window model. For example, *embB* has a gene-level ω 1.19, but the 95%CI
248 lower bound is 0.73 (not significant), and is ranked #141. Only 6 genes are significant in both
249 models (including *pncA, gidB,* and *ubiA*). The fact that the Window model captures more drug-
250 resistance-related genes reinforces the utility of focusing the analysis on selected regions within
251 a gene (Window model), whereas the selection signal at the whole-gene level (Constant model)
252 can get diluted by SNPs in other less-relevant parts of the protein.

253

254 **3.2. Genes significant by Window model exhibit selection in regions of protein associated with**
255 **drug resistance.**

256 One of the advantages of evaluating genes for positive selection using the Window model is that
257 it identifies segments of specific amino acids harboring the signal (rather than requiring a
258 constant omega parameter for the entire gene). In the case of genes commonly associated with
259 drug resistance, the regions under selection often co-occur with known drug-binding regions of
260 the protein. This co-location of selection signals with active sites or structural binding sites for
261 antibiotics lends confidence that the Window model is detecting genuine signals of selective
262 pressure.

263 The localization of the selection signals within each ORF can be observed in omega plots, as
264 illustrated in Figure 2 (omega plots for *gyrA* and *embB*). These plots show the mean of the
265 posterior distribution of omega (average of the MCMC samples) for each codon (black line).
266 Note that the curve is smoothed, which is a consequence of the Window model, where the
267 omega value for one codon effectively combines local information on genetic variants at
268 surrounding codons by assuming the codons within discrete regions of the ORF share a common
269 omega value). (The boundaries of the regions get adjusted dynamically during the MCMC
270 sampling.) Therefore, peaks in the selection signal tend to center on clusters of SNPs where

271 there is a local over-abundance of non-synonymous mutations. In contrast, the omega plot of a
272 gene using the Independent model is much more variable (spiky), because the omega values at
273 one codon is completely independent from the other around it, and there is no sharing of
274 information locally. (Omega plots for *gyrA* and *embB* based on the Independent model, which
275 are much less smooth, are shown in the Supplemental Figure S2).

276 Importantly, although the omega value might exceed 1.0 in multiple regions of the ORF, this
277 does not necessarily mean it is significant. To distinguish genuine selection signals from random
278 fluctuations, we can utilize the dispersion in the MCMC samples to assess certainty. The 95%
279 credible interval (95%CI) of the MCMC samples represents the range of possible values of
280 omega for each codon. The upper and lower bounds of the 95% CI are shown as blue dashed
281 lines in the figure. The significant codons are (only) those for which the lower bound of the
282 95%CI exceeds 1.0, or when $\omega>1$ in 97.5% of the MCMC samples.

283 In the case of GyrA, there are two significant peaks. The predominant peak consists of residues
284 89 to 95 (where omega peaks at 6.7). This contains the two codons most frequently associated
285 with fluoroquinolone (FQ) resistance – A90 and D94 [23]. In fact, although FQ resistance was
286 only ~10% in this population (10.3% for moxifloxacin), A90V was observed in 40 isolates in the
287 Moldova collection, along with D89G/N and D94G/A/N/Y/H. The region also included S95T,
288 which is known to be a lineage-specific marker not associated with drug resistance [24]. The
289 reason it was included was likely due to: a) proximity to D94, and b) a single isolate had an
290 additional non-synonymous mutation, S95A. There is also a minor peak at residues 70-74,
291 which is just barely significant (lower bound of 95%CI just barely exceeds 1.0). This is in an
292 adjacent alpha-helix, also in the DNA-binding pocket; A74S been reported in FQ-resistant clinical
293 isolates, but is less frequently observed [25]. In one study, A74S was only observed in mutants
294 with D94G, so it might be compensatory [26]. Although, some mutations in *gyrB* have also been
295 reported to be associated with FQ-resistance, especially in the region around Arg482 (which is
296 proximal to where fluoroquinolones bind in the 3D structures of the GyrAB complex [27], and
297 some segments of *gyrB* have $\omega>1$ by the Window model, none of these is statistically significant.

298 A similar pattern of selection is observed in *embB* (Figure 2b) in the arabinogalactan (AG)
299 synthesis pathway. The three codons most commonly observed to be mutated in ethambutol-
300 resistant mutants– residues 306, 406, and 497 [28] – all exhibit a significant ω value. In
301 addition, residues 305 (adjacent to 306) and 317-338 (in between 306 and 406) are highlighted
302 as significant. In addition, in *ubiA*, residues 171-176 have significant ω values. UbiA is also in
303 the AG pathway, and these residues have been reported in EMB-resistant mutants [29].

304 It is notable that *pncA* and *gidB*, two of the genes with the strongest selection signals (mean
305 ω=2.93 and 1.94, respectively), show selection in regions throughout the gene (151 and 78
306 codons respectively, see Supp Fig S3), which is consistent with the diversity of resistance
307 mutations observed, including loss-of-function [30]. *pncA* is an activator of pyrazinamide, and
308 *gid* is a ribosomal rRNA methyltransferase that natively facilitates binding of streptomycin [31,
309 32]. Hence any mutations that abrogate function of these two genes could confer resistance
310 [33]. This could explain why GenomegaMap detects a significant signal for positive selection
311 that is diffused across both of these targets.

312

313 **3.2.1. RpoB**

314 Rifampicin resistance mutations are commonly observed in of *rpoB* (beta' chain of RNA
315 polymerase), specifically in the Rifampicin Resistance Determining Region (RRDR), residues 435-
316 452, including the most frequent mutation S450L in RIF-R mutants), which lines the physical
317 binding site for rifampicin. As observed in the omega plot, there are two strong peaks indicating
318 regions of the ORF with significant selection (see Figure 3a). The first peak (residues 435-451),
319 with omega values exceeding 4, clearly corresponds to the RRDR. The second peak consists of
320 residues 488-496. These have been reported to be associated with rifampicin-resistant clinical
321 isolates in China [34]. Since they co-occurred with mutations in the RRDR region of *rpoB*, it was
322 speculated that they could be compensatory. We observe this in the Moldova collection too. In
323 fact, based on the phylogenetic reconstruction (generated using RAxML), there are two clear
324 cases where mutation events in the 488-496 region occurred in lower branches after mutations
325 in the RRDR. For example, 15 isolates with the *rpoB*:I488V were embedded in a cluster of 111

326 isolates with *rpoB*:S450L. This supports the hypothesis that mutations in the 488-496 region
327 might help reduce the fitness cost of the earlier mutations in the RRDR. In crystal structures of
328 the RNA polymerase, the residues 488-496 are in a parallel beta strand to 426-452 in the RIF-
329 binding pocket (with Ile491 directly adjacent to Ser450; see Figure 3b), which could provide a
330 structural explanation for how mutations at one site could mitigate the fitness cost of mutations
331 at the other.

332 In addition to RIF-resistance mutations in *rpoB*, compensatory mutations have been observed in
333 *rpoC*, especially around residues 432-527 [7]. Indeed, *rpoC* is also in the list of 53 genes
334 exhibiting positive selection in the Moldova clinical isolates (using the Window model). As
335 shown in Figure 3c, several regions of codons attain significance, including 431-436, 445-447,
336 482-498, and 502-527, which corresponds well with the region described by Comas et al [7].

337

338 **3.2.2. CinA**

339 CinA has recently been implicated in INH tolerance through TnSeq studies, where it was found
340 to be non-essential under regular growth conditions (7H9 medium), but conditionally more
341 essential (decreased Tn insertion counts) in the presence of isoniazid (INH) [35]. It was
342 subsequently shown to be involved in NADH recycling, and to mitigate the impact of INH-NADH
343 adducts [36]. It was also shown to affect MICs for delamanid (DLM) and pretomanid (PTM),
344 which also generate NADH adducts. CinA has two domains: an N-terminal pyrophosphatase
345 and C-terminal deamidase. The activity influencing drug sensitivity was localized to the
346 pyrophosphatase domain. The omega plot (Figure 4a) exhibits a significant peak for residues in
347 the 82-86 range (in the pyrophosphatase domain). Genetic variants observed in this region
348 include: T82I, V83F, and V86A. These residues are in an alpha-helix proximal to Asp80, which
349 projects into the active site and coordinates Mg^{2+} bound to the diphosphate in NADH, in a
350 crystal structure of CinA homolog from *T. thermophilus* (PDB: 4uux) (see Figure 4b). Mutating
351 Asp80 (D80A) abrogates enzyme activity and increases INH sensitivity [36]. In fact, one of the
352 isolates in the Moldova collection had D80E, which likely also affects CinA function. None of the
353 isolates with these mutations in CinA were genotypically resistant to INH, ETH, DLM, or PTM

354 (based on drug-resistance markers analyzed by TB-profiler). Nonetheless, the proximity of
355 these mutations to the active site suggest that CinA is under genuine selective pressure in this
356 population, the source of which is still unknown. In contrast, only residue 421 was detected as
357 under positive selection in CinA by the Independent model.

358

359 **3.2.3. PhoR**

360 *phoR* has among the strongest signals for positive selection in the Moldova dataset. *phoR* is
361 part of a two-component sensor, PhoPR [18], which plays a key role in virulence [37, 38]. *phoR*
362 is membrane-bound sensor coupled to a histidine kinase that phosphorylates *phoP*, which binds
363 DNA and acts as transcription factor. There is strong evidence for *phoR* being under positive
364 selection in clinical isolates in previous studies [20]. and mutations in *phoR* have been
365 hypothesized to influence transmissibility and might have been a key step in the evolution of
366 Mtb into a human pathogen [39].

367 In the Moldova dataset, the codons in *phoR* with genetic variants under positive selection
368 include: 52-84 and 145-159 (in the periplasmic sensor domain) and 188-201 (in the HAMP
369 domain; an alpha-helical domain that facilitates oligomerization by form coiled-coils, often
370 utilized in signal transduction proteins) (see Figure 5). The polymorphism in *phoR* that
371 differentiates Mtb *phoR* from the ortholog in *M. bovis*, which is associated with decreased
372 transmission in humans, is G71I [39], which falls in the middle of the sensor domain. Gonzalo-
373 Asensio et al. showed that recombineering this mutation into L2 and L4 strains actually down-
374 regulates expression of the PhoPR regulon. Thus, the analysis of positive selection using the
375 Window model in GenomegaMap suggests that other non-synonymous mutations in the sensor
376 domain might cause similar functional changes that are ultimately beneficial and lead to
377 selection in Mtb clinical isolates. Mutations in the HAMP domain will have to be further
378 investigated to determine their impact on PhoPR function and understand how this can lead to
379 adaptation.

380

381 **3.3. The Independent model is biased toward large ORFs.**

382 One notable aspect of the genes exhibiting significant selection under the Independent model is
383 that it contains many polyketide synthase (PKS) genes. Of the 16 PKS genes in the Mtb H37Rv
384 genome, 9 are significant in the Independent model, along with *nrp* (Rv0101, non-ribosomal
385 peptide synthase, NPRS). These modular (multi-domain) proteins are among the largest genes
386 in the genome, with 10 out of 16 have length over 1000 amino acids (*pks12* is largest, with 4150
387 aa, and NRPS is second largest, with 2511 aa). The PKS genes under positive selection are
388 almost all unique calls for the Independent omdel (see Figure 1a); the only PKS gene that is also
389 detected as significant by the Window model is *pks12*. A comparison of the sizes of the genes
390 under positive selection detected by each model is shown in the Figure 6a. Note that the size
391 distribution of the 44 genes significant by Independent model is skewed higher than the 53
392 genes significant by the Window, which has a distribution more similar to the overall
393 distribution of ORF sizes in the genome (represented by the histogram). Figure 6b-c shows the
394 number of genes detected by each model in different bin sizes. When normalized by the total
395 number of genes in each bin (Figure 6d-e), the considerable over-representation of genes in the
396 largest bin (1200-4152 aa) by the Independent model is apparent – over 40% of these largest
397 genes contain a one or more codons under positive selection according to the Independent
398 model (Figure 6e). By comparison, the Window model shows much less bias toward larger
399 genes.

400 It is possible that positive selection of PKS genes could reflect biologically-meaningful selection
401 pressures, since PKS genes make glycolipid components of the cell well that often are
402 recognized by and module the immune response in vivo [40-42]. However, the fact that so
403 many PKS genes show signs of positive selection could also reflect a general bias that the
404 Independent model is more likely to identify random codons as under positive selection in
405 larger genes, which could explain why the largest genes tend to be characterized as under
406 selection by this model.

407 To evaluate this hypothesis, we performed a simulation experiment to estimate the number of
408 genes that would be expected to obtain significance by the Independent or the Window model

409 under a Null model (approximating neutral evolution). Simulated genes were generated that
410 had mutations that occurred at the same rate as the original Moldova dataset, but evolved
411 under the assumption of *neutral evolution*. This was achieved by first computing the least-
412 common ancestral (LCA) sequence for each ORF in the Mtb genome (excluding PPE and
413 PE_PGRS genes, and genes known to be involved in drug resistance). Then, each distinct
414 mutational event (i.e. SNP associated with a particular branch point in the tree) was randomly
415 assigned to a new site within a new gene. This resampling process resulted in 39,564 distinct
416 nucleotide changes permuted to new positions among 3,446,666 sites within coding regions of
417 3877 genes (with the same size distribution as in the Mtb genome), which was then used to
418 generate the 2057 variant sequences for each isolate for each gene. This process decouples the
419 localization of each nucleotide substitution (e.g. clustering within ORFs) and coding impact
420 (whether it is synonymous or non-synonymous).

421 The 3877 simulated gene sequences were then re-analyzed by each of the three GenomegaMap
422 methods. The number of genes detected as significant under this Null model by the
423 Independent model was 66 out of 3877 (~1.7%), whereas only 14 of the simulated genes were
424 identified as under significant positive selection for the Window model (~0.4%) (see Table 2).
425 For comparison, we also ran the Constant model on the same set of simulated genes, which
426 detected 16 genes with significant positive selection. Thus, the Independent model would be
427 expected to have a ~4.5x higher false-positive rate (FPR) relative to the Window model.
428 Furthermore, the Independent model again shows a bias toward detecting positive selection
429 among larger genes (see Figure 7). This suggests that the Independent model is more
430 susceptible to false positive calls, due to modeling each codon with its own individual ω
431 parameter, whereas the Window model is <u>more robust</u> because it combines information (on
432 genetic variants) locally over multiple codons (clusters of SNPs) (through use a constant ω
433 parameter for each segment of the ORF, with boundaries that vary over the MCMC simulation).

434 Table 2. Evaluation of positive selection on 3877 simulated genes (under a Null
435 hypothesis of neutral evolution).

GenomegaMap model	level of selection analysis	number of significant genes under positive selection (estimated FPR)
Constant	individual codons	16/3877 (0.41%)
Window	groups of codons	14/3877 (0.36%)
Independent	whole gene	66/3877 (1.7%)

436

437

438 **3.4. Validation by analysis of an independent collection of Mtb isolates from other countries.**

439 To evaluate whether the genes identified as under positive selection in the Moldova dataset are
440 generally representative, or might be specific to that geographic region, we carried out a similar
441 analysis in a larger collection of 5,195 isolates from 15 other countries. These isolates were also
442 predominantly from lineages L2 and L4, but L1 and L3 were also represented. Approximately
443 one-third of the isolates were drug-resistant, and most of these (33.4%) being MDR (markers for
444 both INH-R and RIF-R, based on TB-profiler). Resistance to EMB and STR was high, as well;
445 fluoroquinolone resistance was lower, at around 10%. (Histograms of the geographic
446 distribution, lineages, and genotypic drug resistance are shown in Supplemental Figures S4 and
447 S5). We applied the same pipeline for genome assembly, quality filtering, and calling of genetic
448 variants as above.

449 Using the Window model in GenomegaMap, 178 genes were found to be under positive
450 selection in this global dataset, using the same criterion for significance (with at least 1 region
451 with ω>1 for at least 97.5% of the MCMC samples) (see Supplemental Table T2). 19 of the 53
452 significant genes in the Moldova dataset were also observed in this global collection,
453 representing a 36% overlap. Predictably, this includes many drug-resistance-related genes, like
454 *pncA, gid, rpoB, rpoC, gyrA, embB* and *folC*. An additional gene implicated in drug-resistance
455 that is uniquely significant in the global collection is *aac* (Rv0262c, aminoglycoside
456 acetyltransferase). Codons under positive selection an *aac* span 86-94. These residues form a
457 loop that binds CoA (rather than the aminoglycoside), based on crystal structure of the complex

458 of *aac* with kanamycin (PDB: 1M4I; see Supplemental Figure S6). Other notable genes found to
459 be under selection in both datasets include *pks12, phoR,* and *resR*. The global positive-selection
460 gene set also includes some additional genes that have been previously reported in other
461 studies of positive selection in Mtb. For example, *whiB6* is a transcriptional regulator that was
462 found to be associated with multi-drug-resistance in a GWAS study of clinical isolates from Peru
463 [6]. *whiB6* has also been implicated in expression of ESX-1 (see below) [43].

464

465 **3.5. ESX-1 genes exhibit signs of positive selection**

466 Several components of the ESX-1 complex exhibit positive selection in the larger global
467 collection of Mtb clinical isolates, including *eccCa1, eccD1*, and *eccE1*, along with *espA*, one of
468 several proteins secreted by ESX-1. In addition, *eccB1, esxA* (ESAT-6)*, and espFIJ* all contained
469 segments showing evidence of positive selection by the weaker definition of significance (>90%
470 of MCMC samples had ω>1). Only *espB* and *espK* showed weak significance in the Moldova
471 dataset, possibly because of the smaller number of isolates. ESX-1 is one of five type-VII
472 secretion systems (T7Ss) in Mtb [44]. It plays an important role in interacting with the host
473 immune system, including phagosome-maturation arrest, host cell membrane lysis, inhibition of
474 autophagy [45], and up-regulation of host-cell genes and cytokine signaling [46-48], and is
475 required for virulence in animal models of infection [49]. The ESX-1 locus (Rv3864-Rv3883c)
476 contains members of the transmembrane pore complex (*eccA1-E1*), the *mycP1* protease, and
477 several secreted effectors (*espBEFGHIJKL, esxAB, pe35, ppe68*). Interestingly *espACD* are in a
478 separate locus (Rv3514c-Rv3616c). Specifically, *espA* binds to the ESX-1 pore complex and
479 facilitates secretion of EsxA and EsxB [50]. The region of *espA* under positive selection includes
480 codons 67-72 (Figure 8a). Non-synonymous genetic variants in this region include: K66R/Q,
481 N67T, R68A, N69S, H70Y, V71L, N72K. These residues lie in an alpha-helix of a predicted 4-helix
482 bundle, proximal to the WXG motif (res. 55-57) (see Figure 8b), which forms a loop that plays a
483 critical role in recognition by the ESX-1 pore complex [50]. As noted above, *whiB6*, and *phoR,*
484 which both regulate expression of ESX-1 [43, 51], also exhibit positive selection in the global
485 dataset.

486

487 **4. DISCUSSION**

488 Mtb is considered well-adapted to its lifecycle as an obligate human pathogen, and the genome
489 is fairly homogeneous (low diversity) [52]. Strong evidence for purifying selection has been
490 found throughout the genome [53]. Nevertheless, sporadic SNPs and other polymorphisms are
491 often observed in genomes of individual isolates (presumably arising within patients during the
492 course of infection [54]), which are infrequent (unfixed alleles, appearing in only one or a few
493 isolates), reflecting on-going evolution [55]. Some of these mutations might reflect adaptations
494 to pressures related to modern clinical treatment of TB, including chemotherapy, but also
495 effects of co-morbidities (e.g. diabetes, COPD, malnutrition), co-infection (e.g. HIV), which could
496 involve exposure to non-TB antibiotics as well as impacting patient immune status, oxidative
497 environment, etc. Statistical methods for evaluating positive selection are a powerful tool for
498 identifying genes adaptive to these pressures in large genomic databases.

499 There have been multiple studies investigating positive selection among genomes of Mtb clinical
500 isolates, based on principles ranging from phenotype-genotype associations [6], to SNP
501 clustering (e.g. nucleotide diversity, mutation burden) [1, 4], to homoplasy [4] and, more
502 broadly, convergent evolution [5], to traditional dN/dS analysis [3, 13], to phylodynamics [2]. In
503 addition to genes already known to be involved in drug-resistance, they have identified other
504 genes which may play an adaptive role to modern clinical stresses, such as *resR* [1] and *phoR*
505 [20]. Most of these studies evaluated selection at the gene-level. However, selection can be
506 analyzed at other scales, including codon-level and clusters of SNPs (intermediate level). The
507 purpose of this study was to evaluate and compare the results of all three levels of analysis.

508 One of the potential problems with gene-level analysis is that the selection signal, which might
509 derive from one or a few residues (e.g. in an active site) could get diluted from mutations in
510 other less-relevant parts of the ORF, when aggregating SNPs over the whole ORF. On the other
511 hand, analyzing each codon independently cannot take advantage of clustering of SNPs, which
512 often occurs in active sites, or sites of protein-protein interaction. We used GenomegaMap as a

513 tool, which implements all 3 analyses within a Bayesian framework, and uses MCMC sampling
514 to make inferences about parameters (omega) and their certainty (credible intervals).

515 We found that the intermediate (or "Window" model) provides the most compelling list of
516 candidate genes. The Window model segments the ORF into multiple contiguous regions, and
517 allows each region to have a different ω parameter, but assumes the codons within a local
518 region all share the same omega value. Many of the genes found to be under positive selection
519 by the Window model are known to be involved in drug resistance, (*rpoB, gyrA, embB, pncA,*
520 *gidB, ubiA, folC*...) which would be expected to be under positive selection in a collection of
521 Mtb clinical isolates from patients receiving chemotherapy, especially in a geographic region
522 with high incidence of drug-resistance (such as Moldova [12]). In addition, on the list of
523 candidates from the Window model, we observed other genes previously reported as targets of
524 positive selection in other studies, such as *resR* [1] and *phoR* [20]. *prpD* is a methylcitrate
525 dehydratase (propionate metabolism pathway) that is regulated by *prpR*, which was identified
526 as associated with multi-drug tolerance [56]. Rv0988 and Rv2082 have both been reported in
527 another study of positive selection in clinical isolates resistant to bedaquiline and clofazimine
528 [57]. While the function of Rv2082 is unknown, Rv0988 is an ABC-transporter implicated in
529 preventing acidification and phagosome maturation [58]. This consistency for "expected"
530 genes under positive selection with previous studies increases confidence that other candidate
531 genes (including those in other pathways which might not be directly related to drug-resistance)
532 might also be genuinely under positive selection and playing an adaptive role.

533 Another advantage of the Window model of selection analysis is that it can identify specific
534 regions within an ORF that are under positive selection. As implemented in GenomegaMap,
535 this model assumes ω is constant locally but varies among regions. However, the borders
536 between regions also get adjusted dynamically during the MCMC sampling, producing a
537 smoothly varying omega signal across the ORF. We define significant codons to be those where
538 >97.% of the sampled ω values from the MCMC sample (the lower-bound of the 95% credible
539 interval) exceed 1.0, which is more stringent that the criterion used in the original paper [10]. In
540 cases where protein crystal structures are available, the regions under positive selection often
541 correspond well to parts of the protein involved in interacting with inhibitors, such as in *rpoB*

542 (residues in the RDRR that bind rifampicin) and *gyrA* (residues binding fluoroquinolones).
543 Another compelling example is the proximity of the residues under selection in *cinA* to the
544 active site (of the pyrophosphatase domain). The insights about which parts of a protein are
545 under selection are helpful in understanding the mechanisms driving it, beyond just labeling a
546 gene as being under positive selection. The correspondence of codons under selection to
547 structural information (e.g. for known drug-binding sites) lends additional confidence in the list
548 of candidates identified by the Window model.

549 Individual codons under selection can also be highlighted by the Independent model in
550 GenomegaMap (where each codon has its own ω parameter). However, we observed that many
551 of the genes on the list of candidates from the Independent model were among the largest
552 genes in the genome, and this size bias suggests the Independent model might be more
553 susceptible to making type I errors. In fact, we used a simulated set of genes with sequences
554 evolved under the assumption of neutral evolution (with the same distribution of ORF sizes and
555 nucleotide content as the clinical isolates from Moldova, but with observed phylogenetic
556 changes permuted randomly to new sites) to show that the Independent model has a roughly 5
557 times higher estimated false positive rate than the Window model. This is likely due to how the
558 Independent model treats each codon independently, making it more susceptible to errors from
559 random coincidences of SNPs, whereas the Window model is more robust because of how it
560 integrates information over multiple codons in a region (by assuming the codons in a region
561 share the same ω), and hence takes advantage of clusters of SNPs where there is a local excess
562 of non-synonymous over synonymous SNPs.

563 On the list of 53 genes identified to be under positive selection in the Moldova dataset, we
564 observed several regulatory proteins, including *pknD* and *sigL* (ser-thr protein kinases), *sigM*
565 (sigma factor), Rv1320 (adenylate cyclase), and *devS* (two-component sensor). As exemplified
566 by *devS*, these regulators could be involved in adaptations to various stresses. For example,
567 *devS* plays a critical role in regulating the response to hypoxia (low oxygen tension, e.g. inside
568 granulomas), a hallmark of the host immune response to TB infection [21, 59-61], and the
569 intracellular oxidative environment could also be influenced by antibiotics (such as isoniazid,
570 which generates oxidative radicals in the process of activation [62]). Several metabolic genes

571 also show signals of positive selection in this population (*cysN* - sulfate
572 adenyltransferase/kinase; *glyA1* - serine hydroxymethyltransferase; *hemN* - coproporphyrinogen
573 III oxidase; *pyrG* - CTP synthase, *atsD* - arylsulfatase; *fadE17* - acyl-CoA dehydrogenase; *cdsA* -
574 CDP-diglyceride pyrophosphorylase). Metabolic shifts often accompany changes in response to
575 stress [63], though the specific role of these pathways remains to be elucidated.

576 Genes classically associated with INH resistance were not detected as significant by the Window
577 model (*katG* and *inhA*). *katG* was probably not detected because mutations of only a single
578 codon, S315; and *inhA* was probably not detected because mutations occur predominantly in
579 the promoter region upstream of *fabG1*, more frequently than in the coding region, affecting
580 expression levels of *inhA*). However, several genes that are potentially indirectly related to INH
581 resistance were observed as under positive selection, including *ndhA*. *ndhA* (Rv0392c) is one of
582 two type II NADH dehydrogenases, along with *ndh* (Rv1854c) which can compensate for INH-
583 NADH adduct formation by shifting the intracellular NADH/NAD+ ratio [64]. Although,
584 historically, *ndh* is more often observed to acquire resistance mutations [65], mutations in *ndhA*
585 have also been reported to be associated with INH resistance [66].

586 In order to determine whether the genes under selection might be specific to Moldova, we
587 applied the same analysis of positive selection to a larger collection of Mtb clinical isolates from
588 15 other countries. Many of the drug-resistance genes, along with genes such as *resR* and *phoR*
589 were also detected as significant in the global dataset. There were approximately 3 times as
590 many genes (178) with statistically significant positive selection as in the Moldova dataset (53).
591 This is possibly due in part to the increased incidence of unfixed mutations (often unique SNPs
592 observed only in single or a few isolates), the probability of which increases with size of the
593 collection (number of genomes, barring duplications due to outbreak clusters) (or possibly due
594 to the increased diversity among isolates selected from 15 different countries). However, in
595 theory, dN/dS analysis should factor-out differences in overall mutation rate, and instead
596 identify genes where there is a relative excess in the rate of non-synonymous over synonymous
597 SNPs, which is not necessarily a consequence of aggregating SNPs from a larger collection of
598 genomes.

599 *phoR* has among the strongest signals for positive selection in both datasets (Moldova, and
600 global collection). PhoPR function is required for virulence, as disruption mutants are highly
601 attenuated in mice [37, 38]. A mutation in *phoP* is one of the few genetic differences between
602 H37Rv and H37Ra, to which the difference in virulence of these strains has been partially
603 attributed [67]. As a part of a two-component sensor, *phoR* regulates many lipid synthesis and
604 cell-wall biogenesis genes [51]. PhoPR has been implicated in responses to several types of
605 environmental stress, including acidic pH [37]. We observed multiple codons in the periplasmic
606 sensor domain to be under positive selection (using the Window model). Using phylodynamics,
607 Chiner-Oms et al. [20] found that *phoR* exhibited an increase in selection of non-synonymous
608 variants, compared to other animal-adapted members of the MTBC (Mycobacterium
609 Tuberculosis Complex), that coincided roughly with expansion of the pathogen into the human
610 population. This and other evidence suggests mutations in *phoR* might facilitate transmission
611 among humans [39].

612

613 Among the genes under positive selection in the larger global collection, we observed several
614 genes from the ESX-1 cluster. In particular, *eccCa1, eccD1, eccE1* were under positive selection,
615 along with *espA*. Based on analogy to a cryo-EM structure of the ESX-5 complex [68], these
616 three subunits interact to form the cytoplasmic face of the transmembrane pore. *espA* is
617 encoded in a different operon (*espACD*), but is thought to bind to the pore complex and
618 facilitate secretion of substrates [46]. ESX-1 plays a crucial role in virulence, secreting a number
619 of effectors (including *esxA* and *esxB*, along with other *esx* and PPE proteins), and can disrupt
620 host-cell membranes and modulate cytokine expression [48, 69]. Thus, we speculate that the
621 genetic variants observed could play a role in fine-tuning the host immune response. ESX
622 components have been identified in other studies of positive selection. For example, *esxW*
623 (secreted via ESX-5) was identified to harbor convergent mutations in different Mtb lineages in a
624 collection of isolates from Viet Nam, which was used to help explain the increased transmission
625 of the Beijing strain in the geographic region [70]. *esxW* was not under significant positive
626 selection in our global collection, and the only *esx* protein that was under significant positive

627 selection was *esxH*, a member of ESX-3. ESX-1 proteins *eccD1* and *eccE1* were also observed to
628 be under positive selection in the analysis of over 51,000 clinical isolates by Liu et al. [1].

629 Increased selection of genetic variants in the ESX-1 system could be the result of adaptation to
630 modern TB treatment paradigms (e.g. shortening of disease course through chemotherapy,
631 better clinical control of inflammation) that might benefit from adjustments in host-pathogen
632 interactions (e.g. enhancement or suppression of immune response). It is interesting to note
633 that PhoPR has been shown to regulate expression of ESX-1 and *esxA* (ESAT-6) secretion [71].
634 Similarly, *whiB6*, which is also on the list of genes under significant positive selection in the
635 global collection, also regulates expression of ESX-1 (demonstrated in *M. marinum*; [43]). The
636 selection signals for these two regulators lend additional support to the potential adaptive role
637 of genetic variants affecting ESX-1 function in this global sample of clinical isolates.

638 One of the limitations of using positive selection to identify genes potentially involved in
639 adaptation is that it is restricted to SNPs in coding regions. Liu et al. [1] also looked for clusters
640 of SNPs (mutational burden) in intergenic regions, especially those associated with regulatory
641 regions (e.g. promotors). In general, dN/dS analysis focuses on SNPs, and thus does not directly
642 take indels into account. However, indels can play a critical role in activating/deactivating
643 proteins; for example, there is a 7-bp insertion that breaks *pks15/*1 into two ORFs and
644 eliminates the synthesis of PGL (phenolglycolipid) found in more virulent strains of Mtb [41].
645 Also, because GenomegaMap does not explicitly use a phylogenetic tree (which is key to its
646 efficiency/scalability over methods like PAML), it is insensitive to homoplasy. However, this can
647 be easily circumvented by using other tools, such as SNPPar [72], which builds a rapid,
648 approximate phylogenetic tree and can then be used to identify homoplasic sites (such as S315
649 in KatG, the primary resistance mutation for INH).

650 In summary, GenomegaMap is an effective tool for evaluating genes under positive selection
651 that scales up well to analyzing large genomic databases. Furthermore, our results suggest that
652 evaluating selection at an intermediate scale (using the "Window" model) identifies the most
653 credible list of candidate genes under selection, compared to gene-level analysis (the
654 "Constant" model) and codon-level analysis (the "Independent" model). The Window model

takes advantage of clusters of SNPs with a local increased rate of non-synonymous over synonymous mutations, which amplifies signals of selection over the other two models. Some of the genes under selection are likely responding to antibiotic treatment (anti-TB chemotherapy), whereas other genes exhibiting positive selection might be reflecting adaptations to other unknown pressures related to the clinical environment, and represent avenues for further investigation into TB biology and potential new drug targets.

Supplemental Data

Supplemental Table T1. Selection scores and statistical significance for all ORFs in the Mtb H37Rv genome for the Moldova dataset (2,057 clinical isolates). Results for 3 models in GenomegaMap are provided: Constant model (one omega parameter per gene), Window model (omega values shared by codons locally within regions), and Independent model (separate omega parameter for each codon). For the Window and Independent model, the number of significant codons ("strongly selected codons") are given, using a 95% credible interval as a significance criterion (ω>1 for >97.5% of the MCMC samples). (The number of weakly selected codons using 80%CI is given too.) The genes are sorted by the "max_omega_LB" which is the peak height of the lower blue curve in the omega plots (lower bound of the 95%CI). The 53 significant genes by the Window model (excluding PPE/PGRS genes) are marked in yellow.

Supplemental Table T2. Selection scores and statistical significance for all ORFs in the Mtb H37Rv genome for the global Mtb genome collection (5,195 clinical isolates from 15 countries). Results for the Window model in GenomegaMap are provided: The number of significant codons ("strongly selected codons") are given, using a 95% credible interval as a significance criterion (ω>1 for >97.5% of the MCMC samples). (The number of weakly selected codons using 80%CI is given too.) The genes are sorted by the "max_omega_LB" which is the peak height of the lower blue curve in the omega plots (lower bound of the 95%CI). The 178 significant genes by the Window model (excluding PPE/PGRS genes) are marked in yellow.

682 **Funding Statement**

683 Funding for this work was provided in part by NIH grant AI143575 (TRI).

684

685 **Acknowledgements**

686 Supercomputing time was provided by the High Performance Research Computing (HPRC)
687 center at Texas A&M University.

688

689 **CRediT authorship contribution statement**

690 TRI - Conceptualization, Data Curation, Formal Analysis, Writing - original manuscript; AS -
691 Visualization, Writing – review and editing.

692

693 **Declaration of Competing Interests**

694 The authors declare no competing interests.

695

696 **REFERENCES**

697 1. Liu, Q., et al., *Tuberculosis treatment failure associated with evolution of antibiotic*
698 *resilience.* Science, 2022. **378**(6624): p. 1111-1118.
699 2. Chiner-Oms, A., et al., *Gene evolutionary trajectories in Mycobacterium tuberculosis*
700 *reveal temporal signs of selection.* Proc Natl Acad Sci U S A, 2022. **119**(17): p.
701 e2113600119.
702 3. Deng, L., et al., *Identification of positively selected genes in Mycobacterium tuberculosis*
703 *from southern Xinjiang Uygur autonomous region of China.* Front Microbiol, 2024. **15**: p.
704 1290227.
705 4. Mortimer, T.D., A.M. Weber, and C.S. Pepperell, *Signatures of Selection at Drug*
706 *Resistance Loci in Mycobacterium tuberculosis.* mSystems, 2018. **3**(1).
707 5. Farhat, M.R., et al., *Genomic analysis identifies targets of convergent positive selection in*
708 *drug-resistant Mycobacterium tuberculosis.* Nat Genet, 2013. **45**(10): p. 1183-9.

6. Farhat, M.R., et al., *GWAS for quantitative resistance phenotypes in Mycobacterium tuberculosis reveals resistance genes and regulatory regions.* Nat Commun, 2019. **10**(1): p. 2128.
7. Comas, I., et al., *Whole-genome sequencing of rifampicin-resistant Mycobacterium tuberculosis strains identifies compensatory mutations in RNA polymerase genes.* Nat Genet, 2011. **44**(1): p. 106-10.
8. Nei, M., *Selectionism and neutralism in molecular evolution.* Mol Biol Evol, 2005. **22**(12): p. 2318-42.
9. Nielsen, R. and Z. Yang, *Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene.* Genetics, 1998. **148**(3): p. 929-36.
10. Wilson, D.J. and C.R. Consortium, *GenomegaMap: Within-Species Genome-Wide dN/dS Estimation from over 10,000 Genomes.* Mol Biol Evol, 2020. **37**(8): p. 2450-2460.
11. Yang, Z., *PAML 4: phylogenetic analysis by maximum likelihood.* Mol Biol Evol, 2007. **24**(8): p. 1586-91.
12. Yang, C., et al., *Phylogeography and transmission of M. tuberculosis in Moldova: A prospective genomic analysis.* PLoS Med, 2022. **19**(2): p. e1003933.
13. Phelan, J.E., et al., *Integrating informatics tools and portable sequencing technology for rapid detection of resistance to anti-tuberculous drugs.* Genome Med, 2019. **11**(1): p. 41.
14. Walker, B.J., et al., *Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement.* PLoS One, 2014. **9**(11): p. e112963.
15. Marin, M., et al., *Benchmarking the empirical accuracy of short-read sequencing across the M. tuberculosis genome.* Bioinformatics, 2022. **38**(7): p. 1781-1787.
16. Stamatakis, A., *RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies.* Bioinformatics, 2014. **30**(9): p. 1312-3.
17. Yang, J., et al., *Analysis on the reconstruction accuracy of the Fitch method for inferring ancestral states.* BMC Bioinformatics, 2011. **12**: p. 18.
18. Parish, T., *Two-Component Regulatory Systems of Mycobacteria.* Microbiol Spectr, 2014. **2**(1): p. MGM2-0010-2013.
19. Broset, E., C. Martin, and J. Gonzalo-Asensio, *Evolutionary landscape of the Mycobacterium tuberculosis complex from the viewpoint of PhoPR: implications for virulence regulation and application to vaccine development.* mBio, 2015. **6**(5): p. e01289-15.
20. Chiner-Oms, A., et al., *Genomic determinants of speciation and spread of the Mycobacterium tuberculosis complex.* Sci Adv, 2019. **5**(6): p. eaaw3307.
21. Kumar, A., et al., *Mycobacterium tuberculosis DosS is a redox sensor and DosT is a hypoxia sensor.* Proc Natl Acad Sci U S A, 2007. **104**(28): p. 11568-73.
22. Raman, S., et al., *Mycobacterium tuberculosis SigM positively regulates Esx secreted protein and nonribosomal peptide synthetase genes and down regulates virulence-associated surface lipid synthesis.* J Bacteriol, 2006. **188**(24): p. 8460-8.
23. Maruri, F., et al., *A systematic review of gyrase mutations associated with fluoroquinolone-resistant Mycobacterium tuberculosis and a proposed gyrase numbering system.* J Antimicrob Chemother, 2012. **67**(4): p. 819-31.

751 24. Sreevatsan, S., et al., *Restricted structural gene polymorphism in the Mycobacterium*
752 *tuberculosis complex indicates evolutionarily recent global dissemination.* Proc Natl Acad
753 Sci U S A, 1997. **94**(18): p. 9869-74.
754 25. Shi, R., et al., *Emergence of ofloxacin resistance in Mycobacterium tuberculosis clinical*
755 *isolates from China as determined by gyrA mutation analysis using denaturing high-*
756 *pressure liquid chromatography and DNA sequencing.* J Clin Microbiol, 2006. **44**(12): p.
757 4566-8.
758 26. Sun, Z., et al., *Comparison of gyrA gene mutations between laboratory-selected*
759 *ofloxacin-resistant Mycobacterium tuberculosis strains and clinical isolates.* Int J
760 Antimicrob Agents, 2008. **31**(2): p. 115-21.
761 27. Blower, T.R., et al., *Crystal structure and stability of gyrase-fluoroquinolone cleaved*
762 *complexes from Mycobacterium tuberculosis.* Proc Natl Acad Sci U S A, 2016. **113**(7): p.
763 1706-13.
764 28. Sun, Q., et al., *Mutations within embCAB Are Associated with Variable Level of*
765 *Ethambutol Resistance in Mycobacterium tuberculosis Isolates from China.* Antimicrob
766 Agents Chemother, 2018. **62**(1).
767 29. Nan, X.T., et al., *Different Contributions of embB and ubiA Mutations to Variable Level of*
768 *Ethambutol Resistance in Mycobacterium tuberculosis Isolates.* Infect Drug Resist, 2024.
769 **17**: p. 3125-3132.
770 30. Yadon, A.N., et al., *A comprehensive characterization of PncA polymorphisms that confer*
771 *resistance to pyrazinamide.* Nat Commun, 2017. **8**(1): p. 588.
772 31. Spies, F.S., et al., *Streptomycin resistance and lineage-specific polymorphisms in*
773 *Mycobacterium tuberculosis gidB gene.* J Clin Microbiol, 2011. **49**(7): p. 2625-30.
774 32. Wong, S.Y., et al., *Functional role of methylation of G518 of the 16S rRNA 530 loop by*
775 *GidB in Mycobacterium tuberculosis.* Antimicrob Agents Chemother, 2013. **57**(12): p.
776 6311-8.
777 33. Gomes, L.C., et al., *Whole genome sequencing reveals large deletions and other loss of*
778 *function mutations in Mycobacterium tuberculosis drug resistance genes.* Microb
779 Genom, 2021. **7**(12).
780 34. Li, M.C., et al., *rpoB Mutations and Effects on Rifampin Resistance in Mycobacterium*
781 *tuberculosis.* Infect Drug Resist, 2021. **14**: p. 4119-4128.
782 35. Xu, W., et al., *Chemical Genetic Interaction Profiling Reveals Determinants of Intrinsic*
783 *Antibiotic Resistance in Mycobacterium tuberculosis.* Antimicrob Agents Chemother,
784 2017. **61**(12).
785 36. Kreutzfeldt, K.M., et al., *CinA mediates multidrug tolerance in Mycobacterium*
786 *tuberculosis.* Nat Commun, 2022. **13**(1): p. 2203.
787 37. Singh, P.R., et al., *Dual functioning by the PhoR sensor is a key determinant to*
788 *Mycobacterium tuberculosis virulence.* PLoS Genet, 2023. **19**(12): p. e1011070.
789 38. Martin, C., et al., *The live Mycobacterium tuberculosis phoP mutant strain is more*
790 *attenuated than BCG and confers protective immunity against tuberculosis in mice and*
791 *guinea pigs.* Vaccine, 2006. **24**(17): p. 3408-19.
792 39. Gonzalo-Asensio, J., et al., *Evolutionary history of tuberculosis shaped by conserved*
793 *mutations in the PhoPR virulence regulator.* Proc Natl Acad Sci U S A, 2014. **111**(31): p.
794 11491-6.

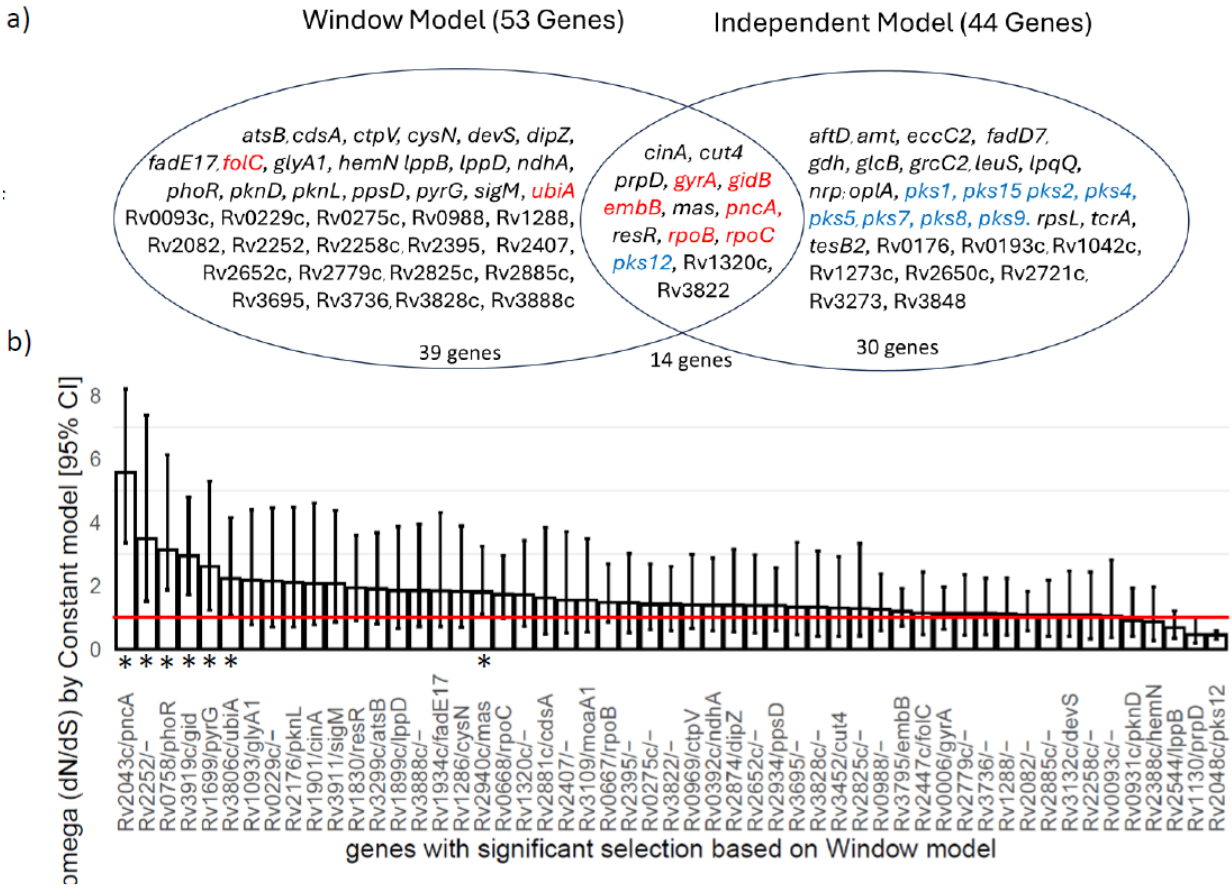
40. Sirakova, T.D., et al., *The Mycobacterium tuberculosis pks2 gene encodes the synthase for the hepta- and octamethyl-branched fatty acids required for sulfolipid synthesis.* J Biol Chem, 2001. **276**(20): p. 16833-9.
41. Constant, P., et al., *Role of the pks15/1 gene in the biosynthesis of phenolglycolipids in the Mycobacterium tuberculosis complex. Evidence that all strains synthesize glycosylated p-hydroxybenzoic methyl esters and that strains devoid of phenolglycolipids harbor a frameshift mutation in the pks15/1 gene.* J Biol Chem, 2002. **277**(41): p. 38148-58.
42. Matsunaga, I., et al., *Mycobacterium tuberculosis pks12 produces a novel polyketide presented by CD1c to T cells.* J Exp Med, 2004. **200**(12): p. 1559-69.
43. Bosserman, R.E., et al., *WhiB6 regulation of ESX-1 gene expression is controlled by a negative feedback loop in Mycobacterium marinum.* Proc Natl Acad Sci U S A, 2017. **114**(50): p. E10772-E10781.
44. Bitter, W., et al., *Systematic genetic nomenclature for type VII secretion systems.* PLoS Pathog, 2009. **5**(10): p. e1000507.
45. Romagnoli, A., et al., *ESX-1 dependent impairment of autophagic flux by Mycobacterium tuberculosis in human dendritic cells.* Autophagy, 2012. **8**(9): p. 1357-70.
46. Champion, P.A., et al., *ESX-1 secreted virulence factors are recognized by multiple cytosolic AAA ATPases in pathogenic mycobacteria.* Mol Microbiol, 2009. **73**(5): p. 950-62.
47. Chirakos, A.E., et al., *Conserved ESX-1 Substrates EspE and EspF Are Virulence Factors That Regulate Gene Expression.* Infect Immun, 2020. **88**(12).
48. Conrad, W.H., et al., *Mycobacterial ESX-1 secretion system mediates host cell lysis through bacterium contact-dependent gross membrane disruptions.* Proc Natl Acad Sci U S A, 2017. **114**(6): p. 1371-1376.
49. Tiwari, S., et al., *Infect and Inject: How Mycobacterium tuberculosis Exploits Its Major Virulence-Associated Type VII Secretion System, ESX-1.* Microbiol Spectr, 2019. **7**(3).
50. Chen, J.M., et al., *Phenotypic profiling of Mycobacterium tuberculosis EspA point mutants reveals that blockage of ESAT-6 and CFP-10 secretion in vitro does not always correlate with attenuation of virulence.* J Bacteriol, 2013. **195**(24): p. 5421-30.
51. Walters, S.B., et al., *The Mycobacterium tuberculosis PhoPR two-component system regulates genes essential for virulence and complex lipid biosynthesis.* Mol Microbiol, 2006. **60**(2): p. 312-30.
52. Coscolla, M. and S. Gagneux, *Consequences of genomic diversity in Mycobacterium tuberculosis.* Semin Immunol, 2014. **26**(6): p. 431-44.
53. Pepperell, C.S., et al., *The role of selection in shaping diversity of natural M. tuberculosis populations.* PLoS Pathog, 2013. **9**(8): p. e1003543.
54. Trauner, A., et al., *The within-host population dynamics of Mycobacterium tuberculosis vary with treatment efficacy.* Genome Biol, 2017. **18**(1): p. 71.
55. Goig, G.A., et al., *Ecology, global diversity and evolutionary mechanisms in the Mycobacterium tuberculosis complex.* Nat Rev Microbiol, 2025.
56. Hicks, N.D., et al., *Clinically prevalent mutations in Mycobacterium tuberculosis alter propionate metabolism and mediate multidrug tolerance.* Nat Microbiol, 2018. **3**(9): p. 1032-1042.

57. Ismail, N., et al., *Emergence of Canonical and Noncanonical Genomic Variants following In Vitro Exposure of Clinical Mycobacterium tuberculosis Strains to Bedaquiline or Clofazimine.* Antimicrob Agents Chemother, 2023. **67**(4): p. e0136822.
58. Pethe, K., et al., *Isolation of Mycobacterium tuberculosis mutants defective in the arrest of phagosome maturation.* Proc Natl Acad Sci U S A, 2004. **101**(37): p. 13642-7.
59. Park, H.D., et al., *Rv3133c/dosR is a transcription factor that mediates the hypoxic response of Mycobacterium tuberculosis.* Mol Microbiol, 2003. **48**(3): p. 833-43.
60. Rustad, T.R., et al., *The enduring hypoxic response of Mycobacterium tuberculosis.* PLoS One, 2008. **3**(1): p. e1502.
61. Sherman, D.R., et al., *Regulation of the Mycobacterium tuberculosis hypoxic response gene encoding alpha -crystallin.* Proc Natl Acad Sci U S A, 2001. **98**(13): p. 7534-9.
62. Sherman, D.R., et al., *Compensatory ahpC gene expression in isoniazid-resistant Mycobacterium tuberculosis.* Science, 1996. **272**(5268): p. 1641-3.
63. Rizvi, A., et al., *Rewiring of Metabolic Network in Mycobacterium tuberculosis During Adaptation to Different Stresses.* Front Microbiol, 2019. **10**: p. 2417.
64. Vilcheze, C., et al., *Altered NADH/NAD+ ratio mediates coresistance to isoniazid and ethionamide in mycobacteria.* Antimicrob Agents Chemother, 2005. **49**(2): p. 708-20.
65. Hazbon, M.H., et al., *Population genetics study of isoniazid resistance mutations and evolution of multidrug-resistant Mycobacterium tuberculosis.* Antimicrob Agents Chemother, 2006. **50**(8): p. 2640-9.
66. Furio, V., et al., *An evolutionary functional genomics approach identifies novel candidate regions involved in isoniazid resistance in Mycobacterium tuberculosis.* Commun Biol, 2021. **4**(1): p. 1322.
67. Lee, J.S., et al., *Mutation in the transcriptional regulator PhoP contributes to avirulence of Mycobacterium tuberculosis H37Ra strain.* Cell Host Microbe, 2008. **3**(2): p. 97-103.
68. Beckham, K.S.H., et al., *Structure of the mycobacterial ESX-5 type VII secretion system pore complex.* Sci Adv, 2021. **7**(26).
69. Zheng, W., et al., *Single-cell analysis reveals Mycobacterium tuberculosis ESX-1-mediated accumulation of permissive macrophages in infected mouse lungs.* Sci Adv, 2025. **11**(3): p. eadq8158.
70. Holt, K.E., et al., *Frequent transmission of the Mycobacterium tuberculosis Beijing lineage and positive selection for the EsxW Beijing variant in Vietnam.* Nat Genet, 2018. **50**(6): p. 849-856.
71. Frigui, W., et al., *Control of M. tuberculosis ESAT-6 secretion and specific T cell recognition by PhoP.* PLoS Pathog, 2008. **4**(2): p. e33.
72. Edwards, D.J., et al., *SNPPar: identifying convergent evolution and other homoplasies from microbial whole-genome alignments.* Microb Genom, 2021. **7**(12).

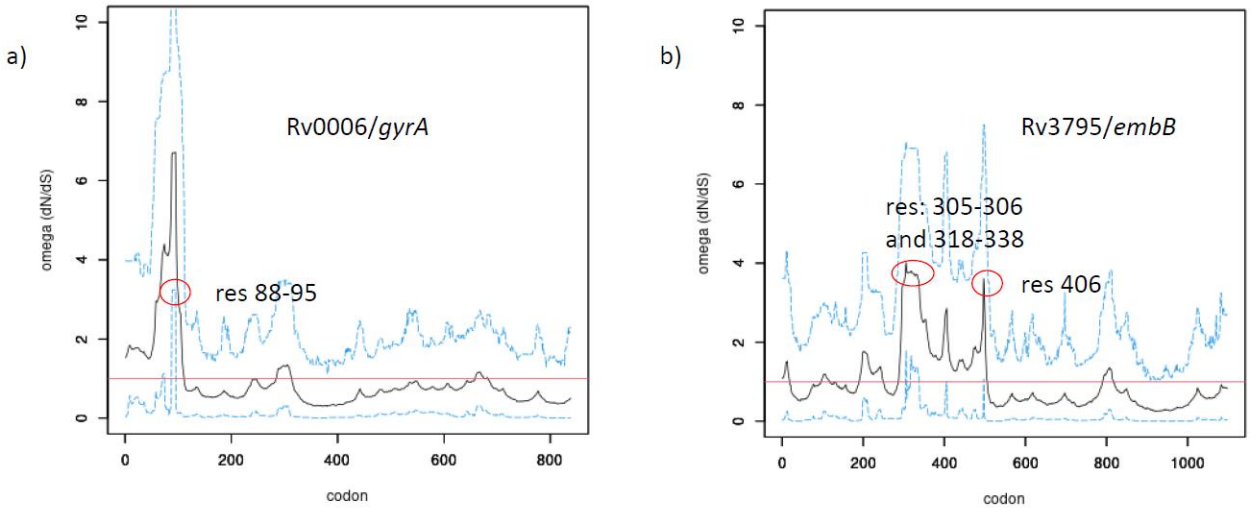
878

879

Figure 1. Comparison of genes under significant positive selection in Moldova dataset using 3 models. a) Venn diagram shows overlap of 53 significant genes from Window model with 44 genes from Independent model. Drug-resistance genes are marked red; polyketide synthases are marked blue. b) Barchart showing the omega estimates by the Constant model for the 53 significant genes based on the Window model. Error bars represent 95% credible intervals for omega based on MCMC sampling. Genes that would be significant based on the Constant model (where omega lower bound>1.0, red line) are marked with an asterisk.



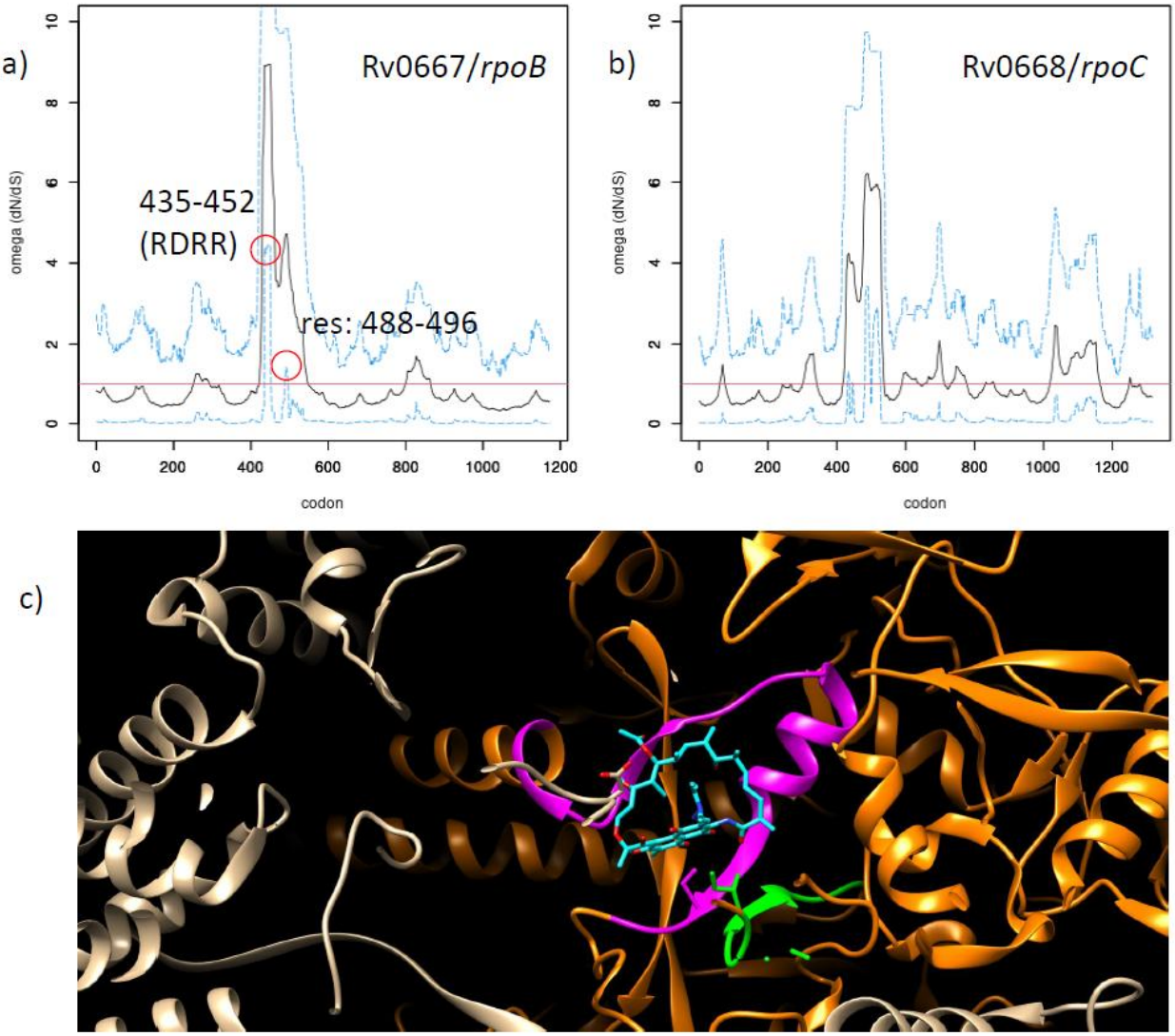
889 **Figure 2**. Example omega plots for gyrA and embB, showing the dN/dS for each codon
890 estimated by the Window model. The black curve shows the mean omega value for each codon
891 averaged over the MCMC samples; the blue lines show the upper- and lower-bounds of the 95%
892 credible interval. Codons under significant positive selection are those where the lower blue
893 line peaks above 1.0.



894

895

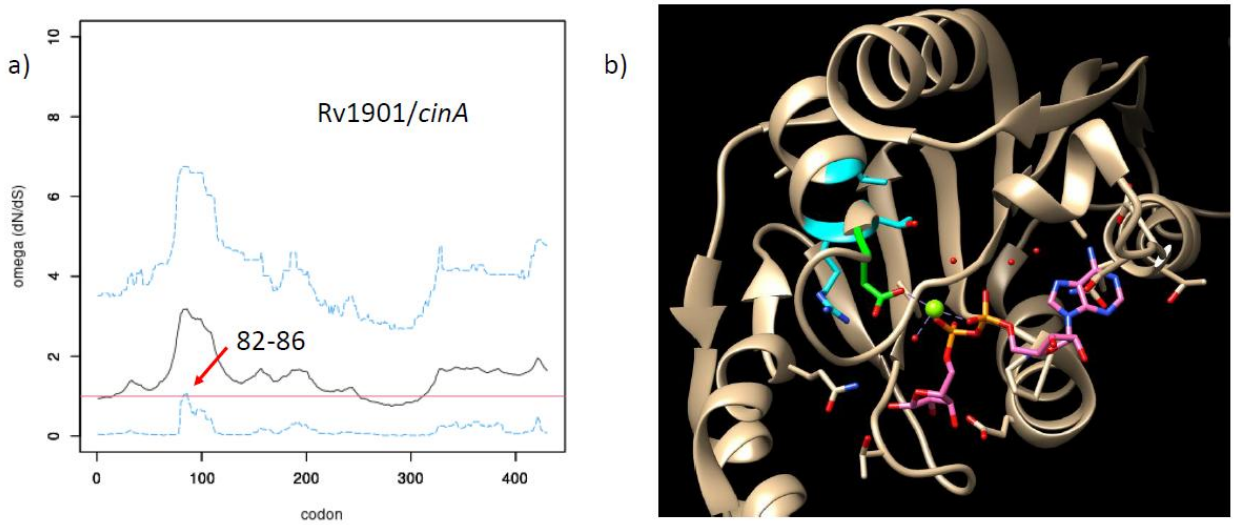
896 **Figure 3.** Analysis of positive selection in RpoB and RpoC. a) Omega plot for *rpoB*. b) omega
897 plot for *rpoC*. c) X-ray crystal structure of complex of RNA polymerase (*rpoABC*) (PDB: 5uhc)
898 with rifampicin (cyan), showing that 488-496 (green) forms a parallel beta-strand to 426-452
899 (purple); both are regions under positive selection.



900

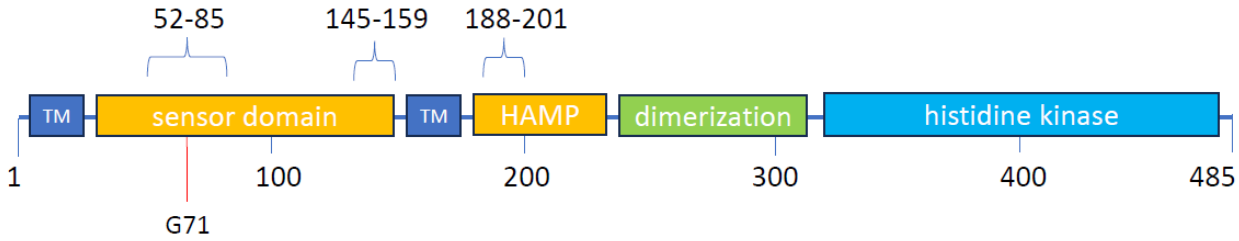
901

902 **Figure 4.** Analysis positive selection in CinA. a) Omega plot for *cinA*, showing selection of
903 residue 82-86 (where lower blue line peaks above 1.0). b) Crystal structure of CinA from *T.*
904 *thermophilus* (N-terminal domain) (PDB: 4uux); green: Asp80 (proximal to NADH in active site,
905 purple), blue residues: codons under positive selection.



906
907

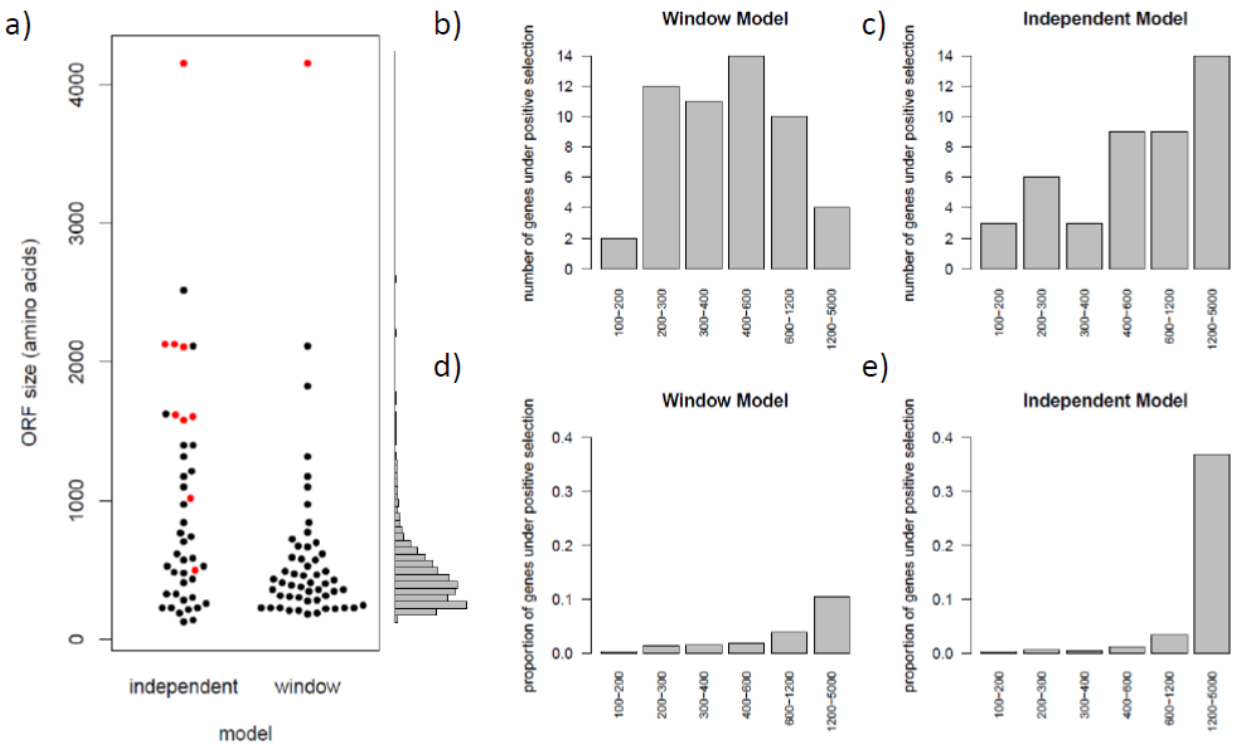
908 **Figure 5.** Regions under positive selection in *phoR*, mapped onto domain structure. 'TM' refers
909 to transmembrane domains. The residue (G71) that differentiates the *Mtb* and *M. bovis*
910 orthologs is indicated.



911

912

913 **Figure 6.** The Independent model is biased toward larger genes. a) Beeswarm plots showing
914 that the significant genes under positive selection by the Independent model are skewed larger
915 than those from the Window model, whereas the Window model better matches the overall
916 ORF size distribution (histogram to right). Polyketide synthase genes are colored red. b) and c)
917 Number of genes under positive selection, divided into size bins (amino acids). The
918 Independent model shows an increasing trend toward the rightmost bin (ORFs with 1200-5000
919 amino acids). d) and e) Proportion of genes under positive selection (counts of significant genes
920 normalized by the overall number of genes in each size bin).

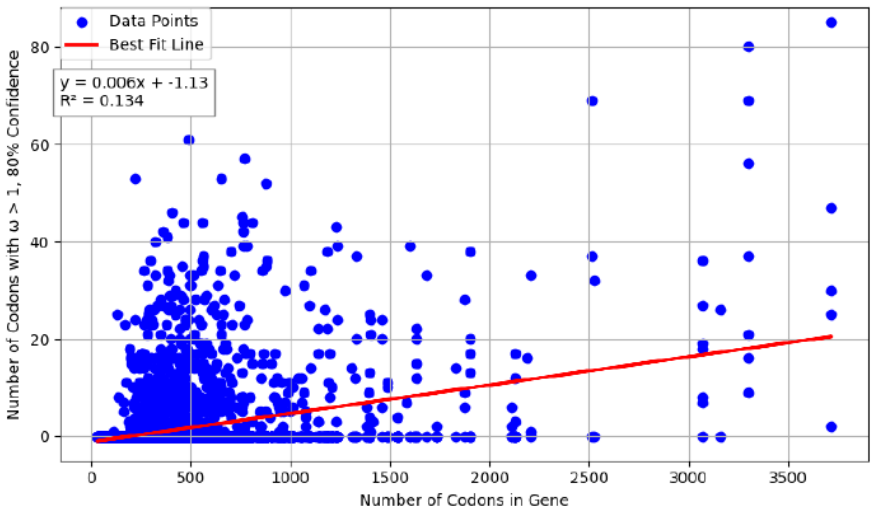


921

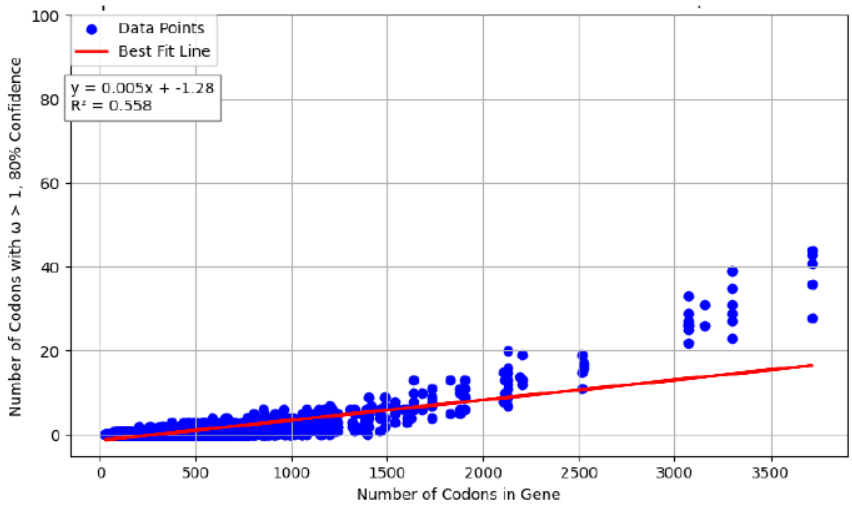
922

Figure 7. Size bias in the Simulated gene set (neutral evolution). Linear regressions of the number of significant codons found against ORF among genes with weakly significant positive selection (ω>1 for >90% of the MCMC samples) using a) the Window model, b) and the Independent model. The Independent model exhibits a strong correlation with ORF size ($r^2 = 0.558$) compared to the Window model ($r^2 = 0.134$).

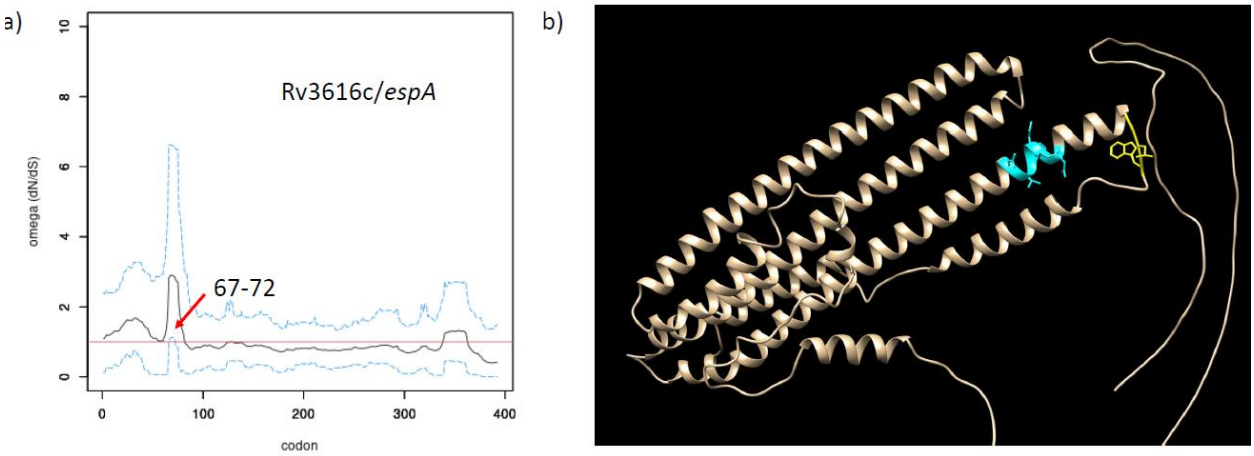
a) Window model



b) Independent model



931 **Figure 8.** Analysis of positive selection for EspA. a) Omega plot for *espA,* showing significant
932 selection for codons 67-72. b) AlphaFold model of EspA. Residues under positive selection (67-
933 72) are colored blue. The WXG motif (residues 55-57), involved in recognition by the ESX-1 pore
934 complex, are colored blue.



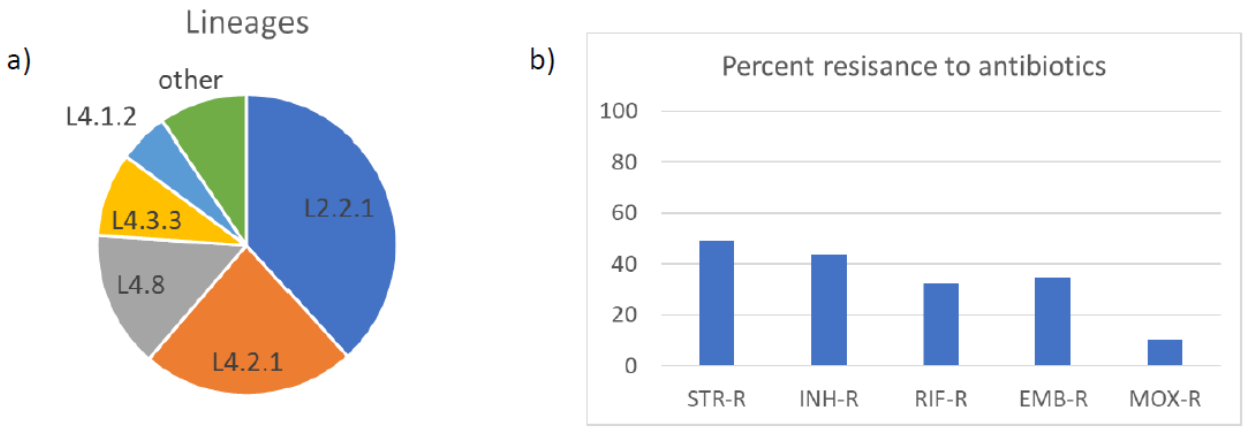
935

936

937 **Supplemental Figure S1.** Lineage distribution and drug-resistance profile for 2,057 Mtb clinical

938 isolates from Moldova (as determined by TB-profiler).

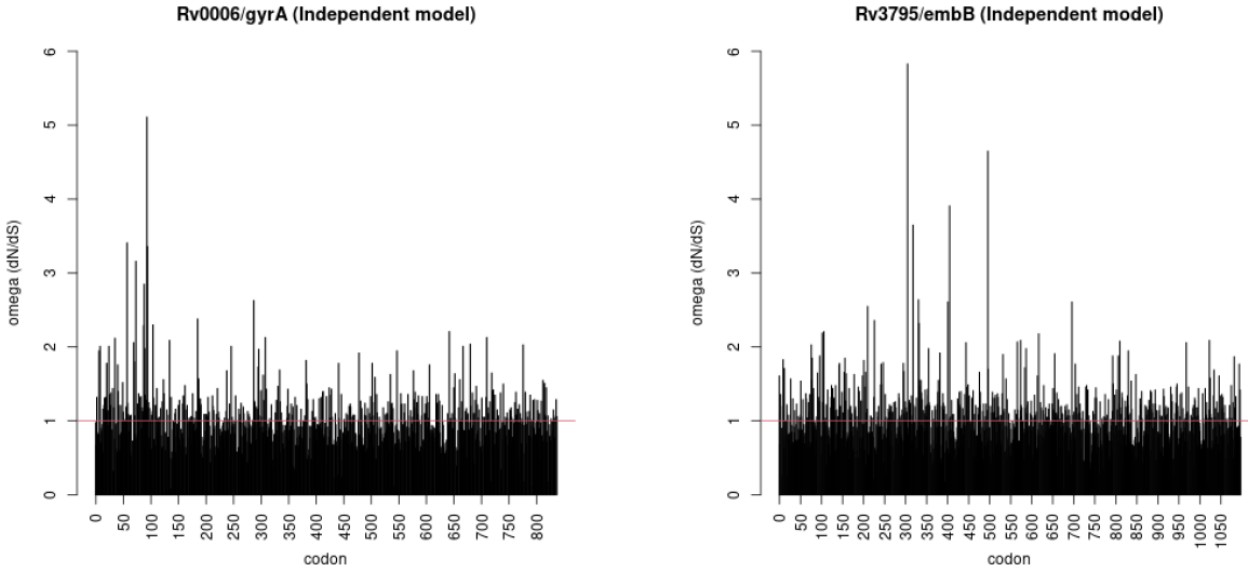
939 .



940

941

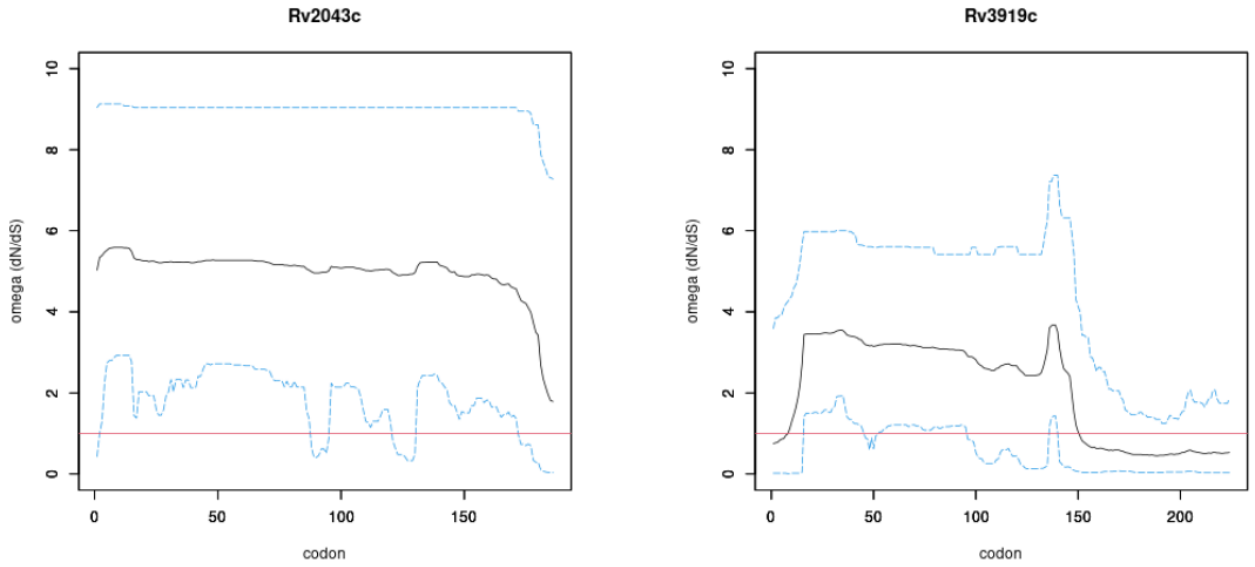
942 **Supplemental Figure S2.** Omega plots for *gyrA* and *embB* based on the Independent model.



943

944

945 **Supplemental Figure S3.** Omega plots for *pncA* (Rv2043c) and *gidB* (Rv3919c) based on the
946 Window model, showing a high fraction of the coding regions under positive selection (lower
947 blue line > 1).

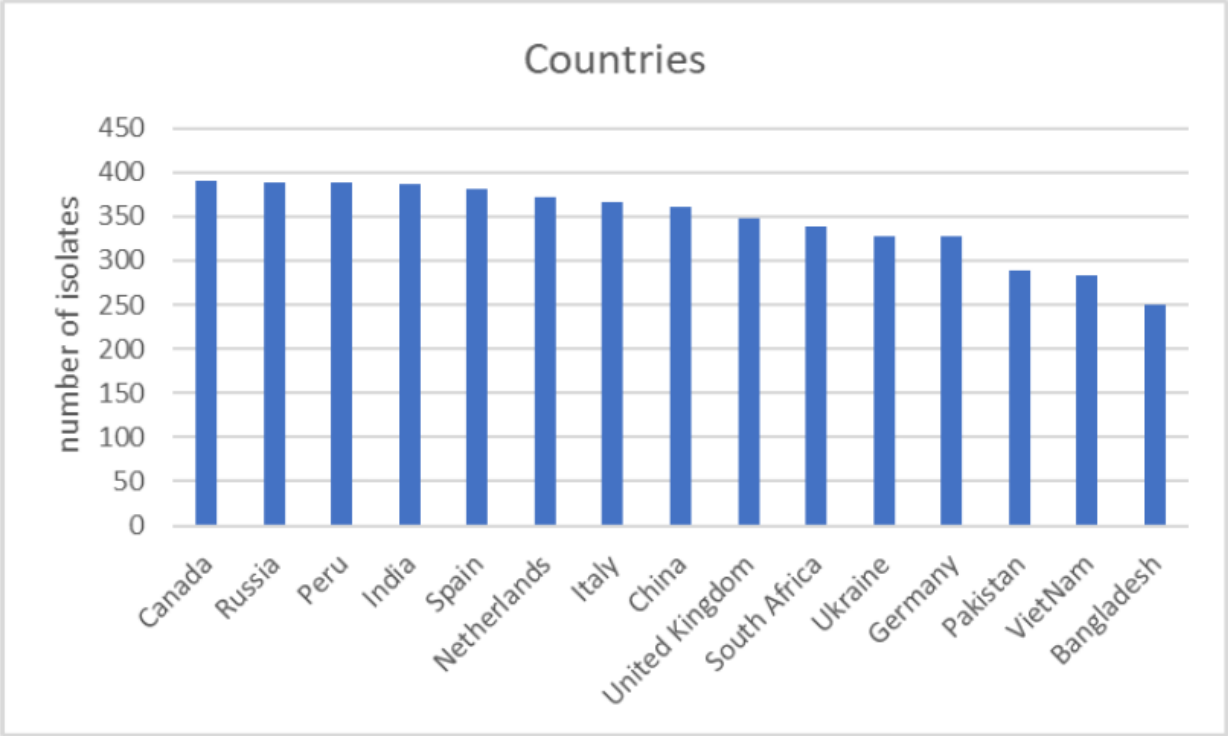


948

949

950 **Supplemental Figure S4.** Countries represented in global collection of 5,195 Mtb clinical

951 isolates.



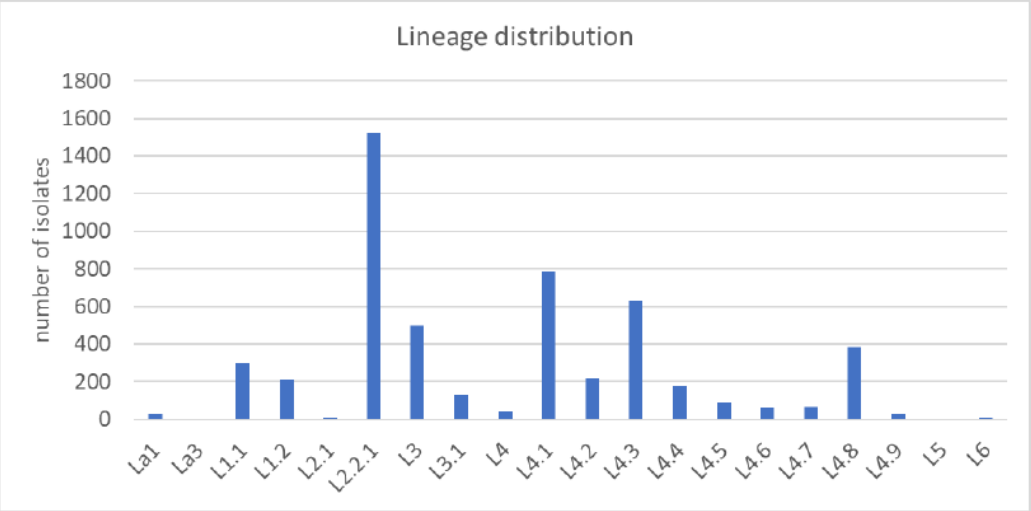
952

953

954 **Supplemental Figure S5.** Lineages and drug-resistance genotypes represented in global

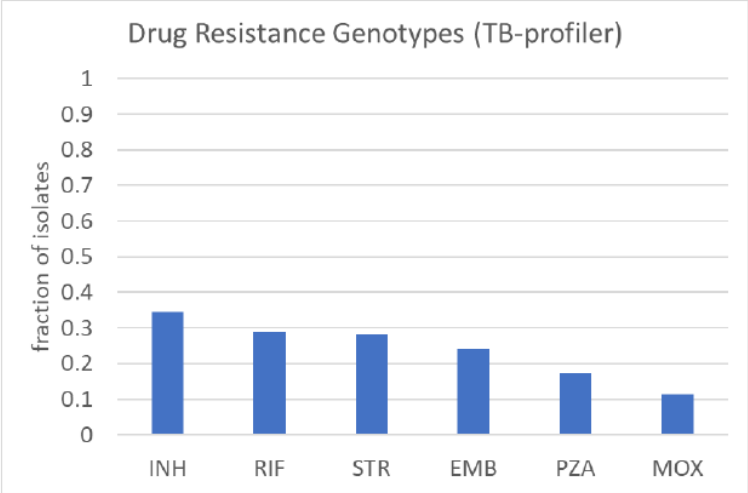
955 collection of 5,195 Mtb clinical isolates.

a)



956

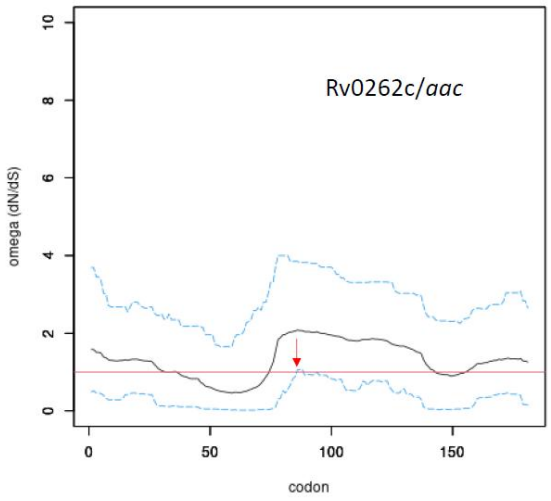
b)

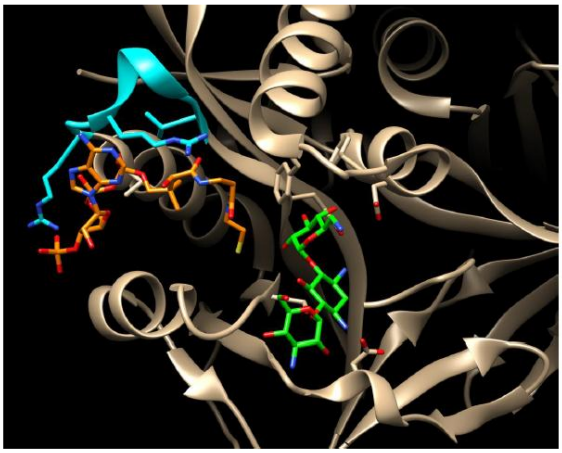


957 b)

958

959 **Supplemental Figure S6**. Analysis of positive selection for *aac* (Rv0262c, aminoglycoside 2'-N-
960 acetyltransferase). a) Omega plot for *aac*: b) Genetic variants in *aac* mapped onto crystal
961 structure (PDB: 1M4I) of complex of Mtb *aac* with CoA (orange) and kanamycin (green);
962 residues under selection (86-94) are marked in blue.





963
964
965