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) yields the most credible list of candidate genes under selection (excluding PPE and PE_PGRS genes, which are predicted less reliably due to frequent sequencing errors). 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 173 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 non-synonymous 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 (https://github.com/danny-wilson/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

81 parameters for mutations observed in genes. GenomegaMap was in fact applied to analyze
82 >10,000 Mtb genome sequences from the CRyPTIC collection (demonstrating its scalability) and
83 proposed a list of genes under positive selection, which included many drug-resistance-related
84 genes (consistent with expectations).

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

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

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

114 Comparing the quality of model fits based on total likelihood, Wilson [10] concluded that the
115 codon-level analysis (Independent model, identifying genes with individual codons under
116 positive selection) fit the data better than the Window model. However, using a criterion of
117 Pr(ω>1)>0.9 (dN/dS>1 in >90% of MCMC samples), thousands of genes were found to have at
118 least one codon under selection (see Fig 3 in Wilson). While many drug-resistance-related
119 genes were included, when genes were ranked by the proportion of codons under selection,
120 some expected genes, like *gyrA, rpoB,* and *embB*, were found to have ranks in the range of
121 2000-3000 (out of ~4000 genes). For the purposes of prioritization, it would be expected that
122 such genes would be ranked closer to the top.

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

138 from an independent set of over 5,000 Mtb clinical isolates from 15 other countries (from the
139 CRyPTIC collection), also using the Window model.

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

143

144 **2. METHODS**

145 **2.1. Sources of Genomic Data.** Two sets of Mtb clinical isolates were analyzed in this paper:
146 2,057 isolates from a previous study in Moldova, and 5,195 globally-diverse isolates from the
147 CRyPTIC collection. Raw sequencing data (fastq files) for isolates from Moldova were
148 downloaded from SRA using accession numbers provided in [12]. Sequencing datasets from the
149 CRyPTIC collection were downloaded based on accession numbers obtained from Supplemental
150 Table 3 in [10]. While isolates could not be selected based on lineage (data not available), to
151 promote genetic diversity, we performed stratified sampling by selecting approximately 500
152 isolates each from 15 different countries. TB-profiler [13] was run on the isolates to identify
153 Mtb lineage, drug-resistance genotypes, and estimated depth of coverage (using *M. tuberculosis*
154 H37Rv as a reference). Isolates that had < 40x coverage or were potentially mixed (with more
155 than one sublineage represented, based on characteristic SNPs) were discarded.

156 **2.2. Variant Extraction and Analysis of Positive Selection.** Pilon [14] was used to map read to
157 the H37Rv genome (NC_000962.3) and extract genetic variants (in the form of .vcf files),
158 including SNPs and indels; Pilon also identifies low-coverage and heterogeneous sites. From the
159 .vcf files, the nucleotide sequences of each ORF were extracted from each isolate, and
160 aggregated by ORF (as .fasta files). Codons with low-coverage or heterogeneous nucleotides, or
161 nonsense mutations (stop codons) were converted to gap characters ('---'), which are
162 subsequently disregarded in the calculations by GenomegaMap.

163 GenomegaMap was then run on each ORF using scripts for the gene-level (Constant), codon-
164 level (Independent), and Window models (https://github.com/danny-wilson/genomegaMap).

165 Default parameters were generally used, as defined in the example configuration files (.xml)
166 provided in the GenomegaMap software distribution. The parameters θ and κ in the Nielsen-
167 Yang model [9] of nucleotide evolution (see Eqn 1; θ corresponds to Q_{ij}) are modeled with log-
168 uniform prior distributions, with half-widths of 0.1 and 0.5. The prior distribution for ω in each
169 region is specified by a gamma distribution with parameters size=1 and shape=1. For the
170 Window model, the ω values are modeled by a piecewise continuous function. During the
171 MCMC simulation (Metropolis-Hastings sampling), proposal updates for ω values are generated
172 using a log-uniform distribution with half-width=1. A Dirichlet-Multinomial distribution (derived
173 in [10]) is used as a likelihood function to determine acceptance probabilities for parameter
174 changes based on the fit of the model to the set of alleles observed in each region. Boundaries
175 between regions are updated dynamically by reversible-jump proposals [15] from a geometric
176 distribution with p=0.033 (1-probability that adjacent residues are in the same region), which
177 ultimately controls the sizes of regions. The initial values for the MCMC sampling are set to
178 θ=0.17 and κ=1.0, and π_j was set to equilibrium codon frequencies (0.01639). For the MCMC
179 sampling for the codon-level and Window models, the trajectory length was 100,000 iterations,
180 saving parameters every 250 steps, and eliminating the first 20% of samples as the burn-in
181 phase (to reduce the influence of initial parameters and ensure the sampler had reached
182 equilibrium). After the MCMC sampling, the value of ω for each codon was summarized over
183 each trajectory by extracting the mean and other statistics (like range and variance), and a 95%
184 credible interval (95%CI) was calculated for the ω value for each codon.

185 **2.3. Evaluation of Models on Simulated Gene Sequences.** Generation of simulated genes for
186 comparison of GenomegaMap models under the assumption of neutral evolution was done as
187 follows. First, a phylogenetic tree based on the 2,057 Moldova clinical isolates was constructed
188 using RAxML [16]. Then, the lowest (or most-recent) common ancestral sequence (LCA) was
189 computed for each ORF in the Mtb genome (excluding PPE and PE_PGRS genes, and genes
190 known to be involved in drug resistance – this exclusion was only done for generating the
191 simulated genes). The LCA nucleotide at each site in a given gene was derived by using Fitch's
192 algorithm [17] to propagate observed nucleotides from leaves (2,057 isolates) up to the root.
193 Then, each distinct mutational event (i.e. SNP associated with a particular branch point in the

194 tree) was randomly assigned to a new site within a new gene. This resampling process resulted
195 in 39,564 mutations permuted to new positions among 3,446,666 sites within coding regions of
196 3,877 genes (with the same size distribution as in the Mtb genome), which was then used to
197 generate the 2,057 variant sequences for each isolate for each gene.

198 **2.4. Protein Structure Analysis**. Atomic coordinates of proteins from X-ray crystals were
199 downloaded from the RCSB Protein Data Bank (www.rcsb.org) and visualized in UCSF Chimera
200 (v1.19, https://www.cgl.ucsf.edu/chimera/). The proteins were displayed as ribbon backbone
201 models, using different colors to highlight regions predicted to be under positive selection with
202 the sliding window model in GenomegaMap. Chimera was also used to calculate distances
203 between amino acids, e.g. distances to active site residues.

204 **3. RESULTS**

205 We used the three models in GenomegaMap to identify and compare genes under positive
206 selection in two collections of Mtb clinical isolates. The first was a collection of 2,236 Mtb
207 clinical isolates from Moldova, representing all reported cases in 2018-2019 [12]. The aim of
208 the original study was to identify transmission clusters within the region. The isolates were
209 found to be predominantly members of lineage 2 and lineage 4, and approximately one-half of
210 the isolates were (genotypically) drug-resistant (based on DR markers to one or more
211 antitubercular drugs), with 32.7% MDR-TB (INH+RIF) (lineage and DR distributions are shown in
212 Supplemental Figure S1). After downloading and preprocessing the raw sequencing data and
213 filtering out low-quality samples, 2,057 isolates were selected for analysis.

214 To determine which of the genes identified as under positive selection in the Moldova dataset
215 are generally representative, versus which might be specific to that geographic region, we
216 carried out a similar analysis in a larger collection of 5,195 isolates from 15 other countries.
217 These isolates were also predominantly from lineages L2 and L4, but L1 and L3 were also
218 represented. Approximately one-third of the isolates were drug-resistant, and most of these
219 (33.4%) being MDR (markers for both INH-R and RIF-R, based on TB-profiler). Resistance to
220 EMB and STR was high, as well; fluoroquinolone resistance was lower, at around 10%.
221 (Histograms of the geographic distribution, lineages, and genotypic drug resistance are shown in

222 Supplemental Figures S2 and S3). We applied the same pipeline for genome assembly, quality
223 filtering, and calling of genetic variants as above.

224 **3.1. The Window model in GenomegaMap identifies positive selection in many genes known**
225 **to be involved in drug resistance.**

226 With the Window model (where different regions within each ORF are modeled with locally-
227 constant omega values), the MCMC simulation produces a sample of parameter combinations,
228 including a representation of the posterior probability distribution of omega values for each
229 codon. The mean number of windows per gene was observed to be 12.0 (evaluated on the
230 Moldova dataset), but it scaled-up linearly with gene length: the mean window size was 28.6
231 amino acids. However, even though residues are grouped together in local windows which
232 share a common ω value, the ω estimate for each individual amino acid can be extracted from
233 its assigned window (which can change in each iteration of the MCMC simulation) and averaged
234 independently.

235 Besides the mean value of omega for each codon, the dispersion of the sample can be used to
236 assess confidence. In [10], the criterion used for significance was if ≥ 90% of the samples had ω
237 >1 ('samples' refers to combinations of parameters generated from iterations of MCMC
238 sampling, see Methods). This corresponds to a (two-sided) 80% credible interval, which is
239 overly generous, potentially generating many false positive codons (245 genes in the Moldova
240 set would be significant by this criterion). Instead, we employ a more stringent criterion of ≥
241 97.5% of samples having ω >1, equivalent to a more standard (two-sided) 95% credible interval.

242 Using this criterion, 53 genes were identified as having at least one codon under positive
243 selection in the Moldova dataset (Table 1). (dN/dS scores and statistical significance for all
244 genes are provided in Supplemental Table T1, with indications of strong positive selection using
245 95% confidence criterion, and weak selection using a confidence threshold of 80%, both two-
246 sided.) Significant genes in the PPE and PE_PGRS families were excluded from Table 1, as these
247 genes often harbor base-call errors due to difficulty sequencing (e.g. GC-rich regions, lower
248 coverage) or ambiguities mapping reads to these repetitive families [18, 19], and these

249 artifactual mutations can potentially cause some of these genes to appear to be under positive
250 selection.

251

252 Table 1. 53 genes with statistically significant evidence of positive selection in the Moldova
253 collection of Mtb clinical isolates. using the Window model analysis in GenomegaMap.
254 Asterisks indicate genes also significant in the CRyPTIC collection.

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* *

255

256 The 53 genes under positive selection in the Moldova dataset (by the Window method) contain
257 many genes known to be involved in drug resistance, including *rpoB* (RNA polymerase, target of
258 rifampicin), *gyrA* (DNA gyrase, target of fluroquinolones, *gidB* (ribosome methyltransferase,
259 affects sensitivity to streptomycin), *pncA* (pyrazinamidase, activator of pyrazinamide), and *embB*
260 and *ubiA* (target of ethambutol). *resR* (Rv1830) also exhibits positive selection in this dataset,
261 consistent with the findings in [1], where it was shown to be associated with general resilience
262 to antibiotics. However, although many isolates had the S315T mutation in KatG, associated
263 with isoniazid resistance, it was not identified by Window model (see further discussion below).

264 For comparison, we used the Window model in GenomegaMap to identify genes under positive
265 selection in the larger global dataset (from CRyPTIC), with isolates from 15 other geographically-
266 diverse countries. 173 genes were found to be under positive selection in this global dataset
267 (excluding PPE and PE_PGRS genes) using the same criterion for significance (with at least 1

region with ω>1 for at least 97.5% of the MCMC samples) (see Supplemental Table T2). 19 of the 53 significant genes in the Moldova dataset were also observed in this global collection (marked with asterisks in Table 1), representing a 36% overlap. As expected, this includes many drug-resistance-related genes, like *pncA, gid, rpoB, rpoC, gyrA, embB* and *folC*. An additional gene implicated in drug-resistance that is uniquely significant in the global collection is *aac* (Rv0262c), which is an aminoglycoside acetyltransferase that provides protection against kanamycin. Other notable genes found to be under selection in both datasets include *pks12, phoR,* and *resR*. The global positive-selection gene set also includes some additional genes that have been previously reported in other studies of positive selection in Mtb, such as *whiB6* [6].

3.2. Gene-level analysis of positive selection (Constant Model in GenomegaMap) fails to identify many known drug-resistance-related genes.

With the gene-level analysis (Constant model), 1924 genes have ω>1 in the Moldova dataset, but only 61 are significant (LB of 95%CI > 1). Two-thirds of these are PPE or PE_PGRS genes; after filtering these out, 22 genes remain that exhibit statistically significant positive selection (Supplemental Table T1). The only drug-resistance-related genes among these are *pncA, gidB,* and *ubiA*. Other DR genes like *rpoB, gyrA, embB,* and *katG* did not reach significance using this model. Many of the significant genes by the Window model have ω below or only marginally above 1.0 in the gene-level model, including many of the genes implicated in drug resistance mentioned above (see Figure 1a). However, the lower bound of the 95% credible interval is less than 1.0 for almost all of the 53 genes significant by the Window model. For example, *embB* has a gene-level ω of 1.19, but the 95%CI lower bound is 0.73 (not significant), and is ranked #141. Only 6 genes are significant in both models (including *pncA, gidB,* and *ubiA*). The fact that the Window model captures more drug-resistance-related genes reinforces the utility of focusing the analysis on selected regions within a gene (Window model), whereas the selection signal at the whole-gene level (Constant model) can get diluted by SNPs in other less-relevant parts of the protein.

295 A similar pattern is seen for the gene-level analysis of the CRyPTIC collection (using the Constant
296 model): 1169 genes had ω>1.0, but only 42 are statistically significant (95%CI LB>1.0). Among
297 these, only 3 genes are related to drug-resistance (*pncA, gidB, ethA*), while 4 are PPE or
298 PE_PGRS genes (Supplemental Table T2).

299

300 **3.3. Codon-level analysis (with the Independent model) is biased toward large ORFs.**

301 The evaluation of genes under positive selection using the codon-level (Independent) model
302 yielded 44 significant genes in the Moldova collection, using the more stringent criterion of at
303 least 1 codon with ω>1 in ≥ 97.5% of the MCMC samples. This list of 44 genes has some overlap
304 with the 53 significant genes by the Window model; there are 14 genes in the intersection,
305 including most of the drug-resistance-related genes (see Venn diagram in Figure 1b). However,
306 there are also many unique genes identified by each of the analysis methods. The Window
307 model also highlighted *ubiA* (arabinogalactan synthesis pathway; harbors resistance mutations
308 to ethambutol) and *folC* (resistance mutations to inhibitors of folate pathway like PAS or
309 sulfamethoxazole), which the codon-level analysis did not.

310 One notable aspect of the genes exhibiting significant selection under the codon-level analysis
311 (Independent model) is that it contains many polyketide synthase (PKS) genes. Of the 16 PKS
312 genes in the Mtb H37Rv genome, 9 are significant in the codon-level analysis, along with *nrp*
313 (Rv0101, non-ribosomal peptide synthase, NPRS). These modular (multi-domain) proteins are
314 among the largest genes in the genome, with 10 out of 16 have length over 1000 amino acids
315 (*pks12* is largest, with 4150 aa, and *nrp* is second largest, with 2511 aa). The PKS genes under
316 positive selection are almost all unique calls for the codon-level analysis (see Figure 1b); the
317 only PKS gene that is also detected as significant by the Window model is *pks12*. A comparison
318 of the sizes of the genes under positive selection detected by each model is shown in the Figure
319 2a. Note that the size distribution of the 44 genes significant by Independent (codon-level)
320 model is skewed higher than the 53 genes significant by the Window model, which has a
321 distribution more similar to the overall distribution of ORF sizes in the genome (represented by
322 the histogram on the right). Figure 2b-c shows the number of genes detected by each model in

323 different bin sizes. When normalized by the total number of genes in each bin (Figure 2d-e), the
324 considerable over-representation of genes in the largest bin (1200-4152 aa) by the codon-level
325 analysis is apparent – over 40% of these largest genes contain a one or more codons under
326 positive selection according to the Independent model (Figure 2e). By comparison, the Window
327 model shows much less bias toward larger genes.

328 It is possible that positive selection of PKS genes could reflect biologically-meaningful selection
329 pressures, since PKS genes make glycolipid components of the cell well that often are
330 recognized by and module the immune response in vivo [20-22]. However, the fact that so
331 many PKS genes show signs of positive selection could also reflect a general bias that the codon-
332 level analysis is more likely to identify random codons as under positive selection in larger
333 genes, which could explain why the largest genes tend to be characterized as under selection by
334 this model.

335 To evaluate this hypothesis, we performed a simulation experiment to estimate the number of
336 genes that would be expected to obtain significance by the codon-level analysis or the Window
337 model under a Null hypothesis assuming neutral evolution. Simulated genes were generated
338 that had mutations that occurred at the same rate as the original Moldova dataset, but evolved
339 under the assumption of *neutral evolution*. This was achieved by first computing the lowest-
340 common ancestral (LCA) sequence for each ORF in the Mtb genome (excluding PPE and
341 PE_PGRS genes, and genes known to be involved in drug resistance). The ancestral sequence
342 for each gene was copied for all isolates. Then, the sequences were mutated at random by
343 applying mutations observed in the original genomic data. Each distinct mutational event (in
344 any gene) (defined by a nucleotide substitution associated with a particular branch point in the
345 original phylogenetic tree) was randomly assigned to a new site within a new gene, and the
346 nucleotides at that site were changed to a different nucleotide for the taxa in that clade (isolates
347 in the subtree beneath that branch point). This resampling process resulted in 39,564 distinct
348 nucleotide changes permuted to new positions among 3,446,666 sites within coding regions of
349 3877 genes (with the same size distribution as in the Mtb genome). This process was used to
350 generate a set of "random" sequences for each gene to be used for comparing the methods to
351 evaluate positive selection, where the mutations occur at the same frequency and lineage

352 distribution as in the original clinical isolates, but the localization of each nucleotide substitution
353 is randomized (e.g. breaking-up clustering within ORFs) and the impact on coding (e.g. whether
354 it is synonymous or non-synonymous) is decoupled (based on frame in new codon), hence
355 simulating absence of selection pressure (no filtering based on translated sequence).

356 The 3,877 simulated gene sequences (each with variant sequences for 2,057 isolates) were then
357 analyzed by each of the three GenomegaMap methods. The number of genes detected as
358 significant under this Null model by the codon-level analysis (Independent model) was 66 out of
359 3,877 (~1.7%), whereas only 14 of the simulated genes were identified as under significant
360 positive selection for the Window model (~0.4%) (see Table 2). For comparison, we also ran the
361 gene-level analysis (Constant model) on the same set of simulated genes, which detected 16
362 genes with significant positive selection. Thus, the codon-level analysis would be expected to
363 have a ~4.5x higher false-positive rate (FPR) relative to the Window model. Furthermore, the
364 Independent model again shows a bias toward detecting positive selection among larger genes
365 (see Figure 3). This suggests that the codon-level analysis is more susceptible to false positive
366 calls, due to modeling each codon with its own individual ω parameter, whereas the Window
367 model is more robust because it combines information (on genetic variants) locally over
368 multiple codons (clusters of SNPs) (through use a constant ω parameter for each segment of the
369 ORF, with boundaries that vary over the MCMC simulation).

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

372

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%)

373

374 **3.4. The Window model identifies regions within proteins under positive selection that are**
375 **likely to be functionally important.**

376 The output of the Window model (MCMC trajectory) was further analyzed to identify specific
377 segments (or amino acids) of genes harboring the selection signal. The localization of the
378 selection signals within each ORF can be observed in omega plots, as illustrated in Figure 4
379 (omega plots for *gyrA* and *embB*). These plots show the mean of the posterior distribution of
380 omega (average of the MCMC samples) for each codon (black line). Recall that, in the Window
381 model, even though residues are grouped in contiguous segments of the ORF which share a
382 local ω value, the window boundaries can shift between iterations of the MCMC simulation.
383 Hence each individual residue can have its own value for the ω estimate that can be extracted
384 and averaged over the MCMC trajectory independently, which is why the mean and confidence-
385 interval estimates of ω can be plotted for each amino acid in Figure 4 and subsequent figures.
386 Note that the curve is smoother than the selection signal produced by the codon-level analysis
387 (Independent model). (Omega plots for the corresponding genes based on the Independent
388 model are shown in the Supplemental Figure S4). This smoothing is a consequence of how the
389 omega estimate for one codon effectively combines local information on genetic variants at
390 surrounding codons. Therefore, peaks in the ω signal using the Window model tend to center
391 on clusters of SNPs where there is a local over-abundance of non-synonymous mutations. The
392 upper and lower bounds of the 95% CI are shown as blue dashed lines in the figure. The
393 significant codons are (only) those for which the lower bound of the 95%CI exceeds 1.0, or
394 when ω>1 in 97.5% of the MCMC samples. In contrast, the omega plot of a gene using the
395 codon-level model is much more variable (spiky), because the omega values at one codon is
396 completely independent from the others around it, and there is no sharing of information
397 locally.

398 In the case of GyrA, there are two significant peaks. The predominant peak consists of residues
399 89 to 95 (where omega peaks at 6.7). This contains the two codons most frequently associated

400 with fluoroquinolone (FQ) resistance – A90 and D94 [23]. In fact, although FQ resistance was
401 only ~10% in this population (10.3% for moxifloxacin), A90V was observed in 40 isolates (1.9%)
402 in the Moldova collection, along with D89G/N and D94G/A/N/Y/H. There is also a minor peak at
403 residues 70-74, which is just barely significant (lower bound of 95%CI just barely exceeds 1.0).
404 This is in an adjacent alpha-helix, also in the DNA-binding pocket; A74S been reported in FQ-
405 resistant clinical isolates, but is less frequently observed [24, 25]. Although, some
406 compensatory mutations in *gyrB* have also been reported to be associated with FQ-resistance
407 [26], none of the codons in *gyrB* were statistically significant using the Window model.

408 Regions associated with resistance mutations were also found to be significant in *embB* (Figure
409 4b), which is in the arabinogalactan (AG) synthesis pathway. The three codons most commonly
410 observed to be mutated in ethambutol-resistant mutants– residues 306, 406, and 497 [27] – all
411 exhibit a significant ω value in the Window model. In addition, residues 305 (adjacent to 306)
412 and 317-338 (in between 306 and 406) are highlighted as significant. While no mutation in
413 residue 305 in EmbB has yet been associated with resistance, some sites in the 317-338 window
414 have (such as 319 and 328; see WHO resistance mutation catalog, [28]). In addition, in *ubiA*,
415 residues 171-176 have significant ω values. UbiA is also in the AG pathway, and these residues
416 have been reported in EMB-resistant mutants [29].

417 In the CRyPTIC collection, *aac* (Rv0262c, aminoglycoside acetyltransferase) was identified to be
418 under positive selection. Codons under positive selection an *aac* span 86-94. These residues
419 form a loop that binds CoA (rather than the aminoglycoside), based on crystal structure of the
420 complex of *aac* with kanamycin (PDB: 1M4I; see Supplemental Figure S5).

421 It is notable that *pncA* and *gidB*, two of the genes with the strongest mean selection signals
422 (mean ω=2.93 and 1.94, respectively), show selection in regions throughout the gene (151 and
423 78 codons respectively, see Supplemental Figure S6), which is consistent with the diversity of
424 resistance mutations observed, including loss-of-function [30]. *pncA* is an activator of
425 pyrazinamide, and *gid* is a ribosomal rRNA methyltransferase that natively facilitates binding of
426 streptomycin [31, 32]. Hence any mutations that abrogate function of these two genes could

427 potentially confer resistance [33]. This could explain why GenomegaMap detects a significant
428 signal for positive selection that is diffused across both of these targets.

429

430 **3.4.1. RpoB**

431 Rifampicin resistance mutations are commonly observed in of *rpoB* (beta' chain of RNA
432 polymerase), specifically in the Rifampicin Resistance Determining Region (RRDR), residues 435-
433 452, including the most frequent mutation S450L in RIF-R mutants), which defines the physical
434 binding site for rifampicin. As observed in the omega plot, there are two strong peaks indicating
435 regions of the ORF with significant selection (see Figure 5a). The first peak (residues 435-451),
436 with omega values exceeding 4, clearly corresponds to the RRDR. The second peak consists of
437 residues 488-496. These have been reported to be associated with rifampicin-resistant clinical
438 isolates in China [34]. Since they co-occurred with mutations in the RRDR region of *rpoB*, it was
439 speculated that they could be compensatory. We observe this in the Moldova collection too. In
440 fact, based on the phylogenetic reconstruction (generated using RAxML), there are two clear
441 cases where mutation events in the 488-496 region occurred in lower branches after mutations
442 in the RRDR. For example, 15 isolates with the *rpoB*:I488V were embedded in a cluster of 111
443 isolates with *rpoB*:S450L (see Supplemental Figure S7). In crystal structures of the RNA
444 polymerase, the residues 488-496 are in a parallel beta strand to 426-452 in the RIF-binding
445 pocket (with Ile491 directly adjacent to Ser450; see Figure 5b).

446 In addition to RIF-resistance mutations in *rpoB*, compensatory mutations have been observed in
447 *rpoC*, especially around residues 432-527 [7]. Indeed, *rpoC* is also in the list of 53 genes
448 exhibiting positive selection in the Moldova clinical isolates (using the Window model). As
449 shown in Figure 5c, several regions of codons attain significance, including 431-436, 445-447,
450 482-498, and 502-527, which corresponds well with the region containing compensatory
451 mutations described by Comas et al [7].

452

453 **3.4.2. CinA**

454 CinA was found to be under significant positive selection in the Moldova dataset, and has
455 recently been implicated in INH tolerance [35]. The activity influencing drug sensitivity was
456 localized to the N-terminal pyrophosphatase domain of the two-domain protein. The omega
457 plot (Figure 6a) exhibits a significant peak for residues in the 82-86 range (in the
458 pyrophosphatase domain). Genetic variants observed in this region include: T82I, V83F, and
459 V86A. These residues are in an alpha-helix proximal to Asp80, which projects into the active site
460 and coordinates Mg^{2+} bound to the diphosphate in NADH, in a crystal structure of CinA homolog
461 from *T. thermophilus* (PDB: 4uux) (see Figure 6b). Mutating Asp80 (D80A) abrogates enzyme
462 activity and increases INH sensitivity [35]. In fact, one of the isolates in the Moldova collection
463 had D80E, which likely also affects CinA function. However, none of the isolates with these
464 mutations in CinA were genotypically resistant to INH, ETH, DLM, or PTM (based on drug-
465 resistance markers analyzed by TB-profiler). In contrast, only residue 421 was detected as under
466 positive selection in CinA by the codon-level analysis.

467

468 **3.4.3. PhoR**

469 *phoR,* which has been found to be under positive selection in previous studies of Mtb clinical
470 isolates [36], has among the strongest signals for positive selection in both the Moldova dataset
471 and the CRyPTIC dataset. In the Moldova dataset, the codons in *phoR* with genetic variants
472 under positive selection include: 52-84 and 145-159 (in the periplasmic sensor domain) and
473 188-201 (in the HAMP domain; an alpha-helical domain that facilitates oligomerization by form
474 coiled-coils, often utilized in signal transduction proteins) (see Figure 7).

475

476 **3.5. ESX-1 genes exhibit signs of positive selection**

477 Several components of the ESX-1 complex (a type VII secretion system that is important for
478 virulence [37, 38]) exhibit positive selection in the larger global collection of Mtb clinical
479 isolates, including *eccCa1, eccD1*, and *eccE1*, along with *espA*. In addition, *eccB1, esxA* (ESAT-6),
480 *and espFIJ* all contained segments showing evidence of positive selection by the weaker

481 definition of significance (>90% of MCMC samples had ω>1). Only *espB* and *espK* showed weak
482 significance in the Moldova dataset, possibly because of the smaller number of isolates. The
483 region of *espA* under positive selection includes codons 67-72 (Figure 8a). Non-synonymous
484 genetic variants in this region include: K66R/Q, N67T, R68A, N69S, H70Y, V71L, N72K. These
485 residues lie in an alpha-helix of a predicted 4-helix bundle, proximal to the WXG motif (res. 55-
486 57) (see Figure 8b), which forms a loop that plays a critical role in recognition by the ESX-1 pore
487 complex [39]. As noted above, *whiB6* and *phoR,* which both regulate expression of ESX-1 [40,
488 41], also exhibit positive selection in the global dataset.

489

490 **4. DISCUSSION**

491 Mtb is considered well-adapted to its lifecycle as an obligate human pathogen, and the genome
492 is fairly homogeneous (low diversity) [42]. Strong evidence for purifying selection has been
493 found throughout the genome [43]. Nevertheless, sporadic SNPs and other polymorphisms are
494 often observed in genomes of individual isolates (presumably arising within patients during the
495 course of infection [44]), which are infrequent (unfixed alleles, appearing in only one or a few
496 isolates), reflecting on-going evolution [45]. Some of these mutations might reflect adaptations
497 to pressures related to modern clinical treatment of TB, including chemotherapy, but possibly
498 also effects of co-morbidities (e.g. diabetes, COPD, malnutrition), co-infection (e.g. HIV), which
499 could involve exposure to non-TB antibiotics as well as impacting patient immune status,
500 oxidative environment, etc. Statistical methods for evaluating positive selection are a powerful
501 tool for identifying genes adaptive to these pressures in large genomic databases.

502 There have been multiple studies investigating positive selection among genomes of Mtb clinical
503 isolates, based on principles ranging from phenotype-genotype associations [6], to SNP
504 clustering (e.g. nucleotide diversity, mutation burden) [1, 4], to homoplasy [4], convergent
505 evolution [5], traditional dN/dS analysis [3, 13], and phylodynamics [2]. In addition to genes
506 already known to be involved in drug-resistance, these studies have identified other genes
507 which may play an adaptive role to modern clinical stresses, such as *resR* [1] and *phoR* [36].
508 Most of these studies evaluated selection at the gene-level. However, selection can be analyzed

509 at other scales, including codon-level and clusters of SNPs (intermediate level). The purpose of
510 this study was to evaluate and compare the results of all three levels of analysis.

511 One of the potential problems with gene-level analysis is that the selection signal, which might
512 derive from one or a few residues (e.g. in an active site) could get diluted from mutations in
513 other less-relevant parts of the ORF, when aggregating SNPs over the whole ORF. On the other
514 hand, analyzing each codon independently cannot take advantage of clustering of SNPs, which
515 often occurs in active sites, or sites of protein-protein interaction. We used GenomegaMap as a
516 tool, which implements all 3 analyses within a Bayesian framework, and uses MCMC sampling
517 to make inferences about parameters (omega) and their certainty (credible intervals).

518 We found that the intermediate (or "Window" model) provides the most compelling list of
519 candidate genes. The Window model segments the ORF into multiple contiguous regions, and
520 allows each region to have a different ω parameter, but assumes the codons within a local
521 region all share the same omega value. Many of the genes found to be under positive selection
522 by the Window model are known to be involved in drug resistance, (*rpoB, gyrA, embB, pncA,*
523 *gidB, ubiA, folC*...) which would be expected to be under positive selection in a collection of
524 Mtb clinical isolates from patients receiving chemotherapy, especially in a geographic region
525 with high incidence of drug-resistance (such as Moldova [12]). In addition, on the list of
526 candidates from the Window model, we observed other genes previously reported as targets of
527 positive selection in other studies, such as *resR* [1] and *phoR* [36]. *prpD* is a methylcitrate
528 dehydratase (propionate metabolism pathway) that is regulated by *prpR*, which was identified
529 as associated with multi-drug tolerance [46]. Rv0988 and Rv2082 have both been reported in
530 another study of positive selection in clinical isolates resistant to bedaquiline and clofazimine
531 [47]. While the function of Rv2082 is unknown, Rv0988 is an ABC-transporter implicated in
532 preventing acidification and phagosome maturation [48]. This consistency for "expected"
533 genes under positive selection with previous studies increases confidence that other candidate
534 genes (including those in other pathways which might not be directly related to drug-resistance)
535 might also be genuinely under positive selection and playing an adaptive role.

536 Another advantage of the Window model of selection analysis is that it can identify specific
537 regions (or residues) within an ORF that are under positive selection. In the case of genes
538 commonly associated with drug resistance, the regions under selection were often found to co-
539 occur with known drug-binding regions of the protein. This co-location of selection signals with
540 active sites or structural binding sites for antibiotics lends confidence that the Window model is
541 detecting genuine signals of selective pressure. As implemented in GenomegaMap, this model
542 assumes ω is constant locally but varies among regions. However, the borders between regions
543 also get adjusted dynamically during the MCMC sampling, producing a smoothly varying profile
544 of ω across the ORF. We define significant codons to be those where >97.% of the sampled ω
545 values from the MCMC sample (the lower-bound of the 95% credible interval) exceed 1.0, which
546 is more stringent that the criterion used in the original paper [10]. In cases where protein
547 crystal structures are available, the regions under positive selection often correspond well to
548 parts of the protein involved in interacting with inhibitors (often in active sites), such as in *rpoB*
549 (residues in the RDRR that bind rifampicin) and *gyrA* (residues binding fluoroquinolones), as
550 well as compensatory regions that might reduce the fitness cost of resistance mutations.
551 Another compelling example is the proximity of the residues under selection in *cinA* to the
552 active site of the pyrophosphatase domain, and in *aac*. The insights about which parts of a
553 protein are under selection are helpful in understanding the mechanisms driving it, beyond just
554 labeling a gene as being under positive selection. The correspondence of codons under
555 selection to structural information (e.g. for known drug-binding sites) lends additional
556 confidence in the list of candidates identified by the Window model.

557 Individual codons under selection can also be highlighted by the Independent model in
558 GenomegaMap (where each codon has its own ω parameter). However, we observed that many
559 of the genes on the list of candidates from the codon-level analysis were among the largest
560 genes in the genome, and this size bias suggests the Independent model might be more
561 susceptible to making type I errors. In fact, we used a simulated set of genes with sequences
562 evolved under the assumption of neutral evolution (with the same distribution of ORF sizes and
563 nucleotide content as the clinical isolates from Moldova, but with observed phylogenetic
564 changes permuted randomly to new sites) to show that the codon-level analysis (Independent

565 model) has a roughly 5 times higher estimated false positive rate than the Window model. This
566 is likely due to how the Independent model treats each codon independently, making it more
567 susceptible to errors from random coincidences of SNPs, whereas the Window model is more
568 robust because of how it integrates information over multiple codons in a region (by forcing the
569 codons in a region to share the same ω), and hence takes advantage of clusters of SNPs where
570 there is a local excess of non-synonymous over synonymous SNPs.

571 Based on the analysis of *rpoB* with the Window model, we observed signals of positive selection
572 in the RDRR region (residues 435-452), as well as in an another segment (residues 488-496)
573 that, which, while sequentially discontiguous, packs against the RDRR in 3D space, forming part
574 of the same RIF-binding pocket. This could provide a structural explanation for how mutations
575 at one site could mitigate the fitness cost of mutations at the other. The fact that several cases
576 were observed where the mutations in 488-496 region of *rpoB* came after the mutations in the
577 RDRR (as subsequent evolutionary events) supports the hypothesis that mutations in the 488-
578 496 region might be adaptive and help reduce the fitness cost of the earlier mutations in the
579 RRDR. We also observed *rpoC* to be under positive selection, which is is consistent with the
580 finding of compensatory mutations for RIF resistance in this gene [7].

581 Among drug-resistance-related genes that were *not* detected as significant by the Window
582 model were *katG* and *inhA*, which is unexpected, given the high level of (genotypic) INH
583 resistance in the region (>40%). None of the three models identified *katG* (activator of
584 isoniazid) as positively selected in this collection, despite mutations in *katG* being the most
585 frequently observed in INH-resistant isolates. Wilson [10] also noted that *katG* was not found to
586 be under positive selection by the Window model in their larger dataset, but was identified by
587 the codon-level analysis (Independent model) because the majority of INH-resistance mutations
588 occur in a single codon - Ser315. However, this could also be easily identified by complementary
589 tools such as homoplasy analysis [4], as mutations at this site often occur multiple times
590 independently in clinical populations. *inhA* was probably not detected because mutations occur
591 predominantly in the promoter region upstream of *fabG1*, more frequently than in the coding
592 region, affecting expression levels of *inhA*).

593 Several genes potentially involved *indirectly* in INH resistance were observed to be under
594 positive selection, including CinA and NdhA. CinA is an NADH-metabolizing enzyme whose
595 biological role is not well-understood [35]. Though CinA is non-essential under regular growth
596 conditions (7H9 medium), it was found to be conditionally more essential (decreased Tn
597 insertion counts) in the presence of isoniazid (INH) [49]. It was subsequently shown to be
598 involved in NADH recycling, and to mitigate the impact of INH-NADH adducts [35]. It was also
599 shown to affect MICs for delamanid (DLM) and pretomanid (PTM), which also generate NADH
600 adducts. CinA has two domains: an N-terminal pyrophosphatase and C-terminal deamidase.
601 The signal for positive selection was associated with the pyrophosphatase domain, and the
602 genetic variants observed (in the Moldova collection) where proximal to the active site. The
603 proximity of these mutations to the active site suggest that CinA is under genuine selective
604 pressure in this population, the source of which is still unknown. Another gene observed to be
605 under positive selection that is potentially related to INH resistance indirectly was was *ndhA*.
606 *ndhA* (Rv0392c) is one of two type II NADH dehydrogenases, along with *ndh* (Rv1854c) which
607 can compensate for INH-NADH adduct formation by shifting the intracellular NADH/NAD+ ratio
608 [50]. Although, historically, *ndh* is more often observed to acquire resistance mutations [51],
609 mutations in *ndhA* have also been reported to be associated with INH resistance [52].

610 On the list of 53 genes identified to be under positive selection in the Moldova dataset, we
611 observed several regulatory proteins, including *pknD* and *sigL* (ser-thr protein kinases), *sigM*
612 (sigma factor), Rv1320 (adenylate cyclase), *phoR* and *devS* (two-component sensors).. These
613 regulators could be involved in adaptations to various stresses. For example, *devS* plays a
614 critical role in regulating the response to hypoxia (low oxygen tension, e.g. inside granulomas), a
615 hallmark of the host immune response to TB infection [53-56], and the intracellular oxidative
616 environment could also be influenced by antibiotics (such as isoniazid, which generates
617 oxidative radicals in the process of activation [57]). *whiB6* (under positive selection in the
618 CRyPTIC collecion) is a transcriptional regulator that was found to be associated with multi-
619 drug-resistance in a GWAS study of clinical isolates from Peru [6]. *whiB6* has also been
620 implicated in expression of ESX-1 [40]. Several metabolic genes also show signals of positive
621 selection in this population (*cysN* - sulfate adenyltransferase/kinase; *glyA1* - serine

622 hydroxymethyltransferase; *hemN* - coproporphyrinogen III oxidase; *pyrG* - CTP synthase, *atsD* -
623 arylsulfatase; *fadE17* - acyl-CoA dehydrogenase; *cdsA* - CDP-diglyceride pyrophosphorylase).
624 Metabolic shifts often accompany changes in response to stress [58], though the specific role of
625 these pathways remains to be elucidated.

626 *phoR* had among the strongest signals for positive selection in both datasets (Moldova, and
627 global collection). There is strong evidence for *phoR* being under positive selection in clinical
628 isolates in previous studies [36]. *phoR* is part of a two-component sensor, PhoPR [59], which
629 plays a key role in virulence [60, 61]. *phoR* is membrane-bound sensor coupled to a histidine
630 kinase that phosphorylates *phoP*, which binds DNA and acts as transcription factor. PhoPR
631 function is required for virulence, as disruption mutants are highly attenuated in mice [60, 61].
632 A mutation in *phoP* is one of the few genetic differences between H37Rv and H37Ra, to which
633 the difference in virulence of these strains has been partially attributed [62]. As a part of a two-
634 component sensor, *phoR* regulates many lipid synthesis and cell-wall biogenesis genes [41].
635 PhoPR has been implicated in responses to several types of environmental stress, including
636 acidic pH [60]. Using phylodynamics, Chiner-Oms et al. [36] found that *phoR* exhibited an
637 increase in selection of non-synonymous variants, compared to other animal-adapted members
638 of the MTBC (Mycobacterium Tuberculosis Complex), that coincided roughly with expansion of
639 the pathogen into the human population. This and other evidence suggests mutations in *phoR*
640 might facilitate transmission among humans and might have been a key step in the evolution of
641 Mtb into a human pathogen [63]. The lineage-specific polymorphism that differentiates Mtb
642 *phoR* from the ortholog in *M. bovis*, which is associated with decreased transmission in humans,
643 is G71I [63], which falls in the middle of the sensor domain. Gonzalo-Asensio et al. showed that
644 recombineering this mutation into L2 and L4 strains actually down-regulates expression of the
645 PhoPR regulon. Using the Window model we observed naturally-occurring genetic variants
646 (non-synonymous mutations) in both the sensor domain and HAMP domains of PhoR. The non-
647 synonymous mutations in the sensor domain might cause similar functional changes that are
648 ultimately beneficial and lead to selection in Mtb clinical isolates. Mutations in the HAMP
649 domain will have to be further investigated to determine their impact on PhoPR function and
650 understand how this can lead to adaptation.

651 In order to determine whether the genes under selection might be specific to Moldova, we
652 applied the same analysis of positive selection to a larger collection of Mtb clinical isolates from
653 15 other countries. Many of the drug-resistance genes, along with genes such as *resR* and *phoR*
654 were also detected as significant in the global dataset. There were approximately 3 times as
655 many genes (173) with statistically significant positive selection as in the Moldova dataset (53).
656 This is possibly due in part to the increased incidence of unfixed mutations (often unique SNPs
657 observed in only one or a few isolates), the probability of which increases with size of the
658 collection (number of genomes, barring duplications due to outbreak clusters), or possibly due
659 to the increased diversity among isolates selected from 15 different countries. However, in
660 theory, dN/dS analysis should factor-out differences in overall mutation rate, and instead
661 identify genes where there is a *relative* excess in the rate of non-synonymous over synonymous
662 SNPs, which is not necessarily a consequence of aggregating SNPs from a larger collection of
663 genomes.

664 Among the genes under positive selection in the larger global collection, we observed several
665 genes from the ESX-1 cluster. In particular, *eccCa1, eccD1, eccE1* were under positive selection,
666 along with *espA*. ESX-1 is one of five type-VII secretion systems (T7Ss) in Mtb [37]. It plays an
667 important role in interacting with the host immune system, including phagosome-maturation
668 arrest, host cell membrane lysis, inhibition of autophagy [64], and up-regulation of host-cell
669 genes and cytokine signaling [65-67], and is required for virulence in animal models of infection
670 [38]. ESX-1 secretes a number of effectors (including *esxA* and *esxB*, along with other *esx* and
671 PPE proteins), and can disrupt host-cell membranes and modulate cytokine expression [67, 68].
672 The ESX-1 locus (Rv3864-Rv3883c) contains members of the transmembrane pore complex
673 (*eccA1-E1*), the *mycP1* protease, and several secreted effectors (*espBEFGHIJKL, esxAB, pe35,*
674 *ppe68*). Based on analogy to a cryo-EM structure of the ESX-5 complex from *M. xenopi* [69],
675 *eccCa1, eccD1,* and *eccE1* are predicted to interact and form the cytoplasmic face of the
676 transmembrane pore. Interestingly *espACD* are in a separate locus (Rv3514c-Rv3616c) from the
677 main ESX-1 cluster. Specifically, *espA* binds to the ESX-1 pore complex and facilitates secretion
678 of EsxA and EsxB [39, 65]. Thus, we speculate that the genetic variants observed in ESX-1 genes
679 could play a role in fine-tuning the host immune response. ESX components have been

680 identified in other studies of positive selection. For example, *esxW* (secreted via ESX-5) was
681 identified to harbor convergent mutations in different Mtb lineages in a collection of isolates
682 from Viet Nam, which was proposed to help explain the increased transmission of the Beijing
683 strain in the geographic region [70]. *esxW* was not under significant positive selection in our
684 global collection, and the only *esx* protein that was under significant positive selection was
685 *esxH*, a member of ESX-3. ESX-1 proteins *eccD1* and *eccE1* were also observed to be under
686 positive selection in the analysis of over 51,000 clinical isolates by Liu et al. [1].

687 Increased selection of genetic variants in the ESX-1 system could be the result of adaptation to
688 modern TB treatment paradigms (e.g. shortening of disease course through chemotherapy,
689 better clinical control of inflammation) that might benefit from adjustments in host-pathogen
690 interactions (e.g. enhancement or suppression of immune response). It is interesting to note
691 that PhoPR has been shown to regulate expression of ESX-1 and *esxA* (ESAT-6) secretion [71].
692 Similarly, *whiB6*, which is also on the list of genes under significant positive selection in the
693 global collection, also regulates expression of ESX-1 (demonstrated in *M. marinum*; [40]). The
694 selection signals for these two regulators lend additional support to the potential adaptive role
695 of genetic variants affecting ESX-1 function in this global sample of clinical isolates.

696

697 **5. LIMITATIONS**

698 One of the limitations of GenomegaMap is that it does not take into account phylogenetic
699 information when analyzing residues for positive selection, and hence is not sensitive to
700 mutations that are selected multiple times independently (homoplasy). This was done for
701 efficiency, to avoid have to build phylogenetic trees, which does not scale-up well for thousands
702 of genomes. While most regions under positive selection exhibit a cluster of SNPs enriched for
703 non-synonymous over synonymous mutations, there are some cases where mutations can be
704 focused primarily on a single site, such as for the KatG:S315T mutation, which GenomegaMap
705 had trouble detecting. However, GenomegaMap can easily be complemented with other tools,
706 such as for homoplasy analysis (e.g. SNPPar [72]). Pepperell [4] and others have demonstrated

707 that homoplasy analysis can effectively identify residues and genes under positive selection,
708 including those associated with drug resistance among Mtb clinical isolates.

709 A related limitation is that GenomegaMap is restricted to identifying positive selection in coding
710 regions, because of its reliance on interpreting mutations in terms of non-synonymous versus
711 synonymous substitutions (i.e. to evaluate dN/dS). However, methods like homoplasy analysis
712 and SNP burden tests that focus on evolution or clustering (and do not rely on translation) can
713 again be used to complement GenomegaMap to identify signs of positive selection in non-
714 coding (or intergenic) regions. For example, Liu et al [1] used a SNP burden test to assess
715 positive selection in promoter regions upstream of ORFs, suggesting that the mutations might
716 produce their phenotypes through altering gene expression levels.

717 While we found that the Window model of GenomegaMap worked better than the gene-level
718 or codon-level analysis for identifying regions under selection, we did not systematically explore
719 the effect of window size. Using default parameters, amino acids were grouped into windows of
720 approximately 28 amino acids. This produced a smoothing effect on omega values (averaged
721 over the MCMC trajectory) because it allowed neighboring residues to share the same local
722 omega parameter (depending on the location of boundaries). It is possible that by varying the
723 parameter *p* of the geometric distribution in the MCMC simulation, average window sizes could
724 be increased or decreased, which could have influenced the sensitivity for detecting regions
725 under positive selection (by over- or under-smoothing). For example, it could be that enforcing
726 smaller window sizes might allow increased resolution by focusing more precisely on clusters of
727 residues sporting an excess of non-synonymous mutations, undiluted by mutations in
728 surrounding residues.

729 The implementation of the Window model in GenomegaMap implicitly assumes that residues
730 under positive selection are clustered in sequentially contiguous regions (segments of an ORF).
731 However, it is possible that discontiguous residues fold together in 3D space to be functional,
732 such as in some active sites. Actually, these are sometimes recognized as *separate* windows
733 under positive selection, as we illustrated with the case of *rpoB*, though finding a way to
734 combine them could increase statistical sensitivity, rather than analyzing the regions

735 independently. A variety of extensions to the method could be imagined, such as higher-level
736 clustering of residues with elevated ω estimates, but the challenge would be revising the
737 Bayesian model to calculate the statistical significance for these clusters in a rigorous way.

738 Another limitation of GenomegaMap is that it does not take into account indels
739 (insertions/deletions), only nucleotide substitutions. However, indels can play a critical role in
740 activating/deactivating proteins, in some cases producing phenotypes. For example, Indels are
741 commonly observed as a mechanism of resistance in non-essential drug-activator genes, like
742 *pncA* [30], and have also been reported in genes related to PDIM synthesis (*ppsA-E, mas*), which
743 have been found to produce a slight growth advantage in laboratory settings [73]. There is a 7-
744 bp insertion that divides *pks15/*1 into two ORFs and eliminates the synthesis of PGL
745 (phenolglycolipid) found in more virulent strains of Mtb (e.g. W-Beijing clade in lineage 2) [21].
746 Some indels (including large-scale deletions spanning >10kb and tens of genes) can serve as
747 lineage-specific markers [74, 75]. Indels have recently been found to be a driver of variation
748 among PPE and PE_PGRS genes (relying on long-read sequencing for accuracy) [76]. Association
749 of indels with specific phenotypes, such as drug resistance, can be evaluated through statistical
750 methods such GWAS (Genome-Wide Association testing), when aligned and encoded properly
751 as a genotype [77].

752 A final limitation is that GenomegaMap does not work as robustly on genes in the PPE and
753 PE_PGRS families because difficulty sequencing these regions [18] can generate incorrect or
754 uncertain base calls, which can inflate the apparent mutation rate in such genes. This can be
755 related to higher GC-content, which causes problem for Illumina sequencing technologies and
756 often leads to lower coverage and increased base-call ambiguity [19]. We observed that many
757 PPE and PGRS genes were identified as being under positive selection by GenomegaMap,
758 though some of these might be artifacts (false positives). These genes are mostly secreted,
759 structural proteins (non-enzymes) [78], and almost all are non-essential in-vitro (e.g. knockout
760 by transposon mutagenesis has no identifiable phenotype, [79]). Few genes in these families
761 have been found to have identifiable functions, with a few notable exceptions, such as PPE51
762 [80, 81]. However, some of these genes have been implicated in pathogenesis via modulation
763 of host immune cells [82, 83], for example, PGRS33 [84]. In fact, several PPE and PGRS genes

764 have been highlighted in another study of positive selection (also based on dN/dS) among Mtb
765 clinical isolates in China [3]. Therefore, we provide selection data (from GenomegaMap) on
766 these genes in the Supplemental Tables for completeness.

767

768 **6. CONCLUSION**

769 In summary, GenomegaMap is an effective tool for evaluating genes under positive selection
770 that scales up well to analyzing large genomic databases. Furthermore, our results suggest that
771 evaluating selection at an intermediate scale (using the "Window" model) identifies the most
772 credible list of candidate genes under selection, compared to gene-level analysis (the Constant
773 model) and codon-level analysis (the Independent model). The Window model takes advantage
774 of clusters of SNPs with a local increased rate of non-synonymous over synonymous mutations,
775 which amplifies signals of selection over the other two models. Some of the genes under
776 selection are likely responding to antibiotic treatment (anti-TB chemotherapy), whereas other
777 genes exhibiting positive selection might be reflecting adaptations to other unknown pressures
778 related to the clinical environment and represent avenues for further investigation into TB
779 biology and potential new drug targets.

780

781 **Supplemental Data**

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

791 bound of the 95%CI). The 53 significant genes by the Window model (excluding PPE/PGRS
792 genes) are marked in yellow.

793 **Supplemental Table T2.** Selection scores and statistical significance for all ORFs in the Mtb
794 H37Rv genome for the global Mtb genome collection (5,195 clinical isolates from 15 countries).
795 Results for the Window model in GenomegaMap are provided: The number of significant
796 codons ("strongly selected codons") are given, using a 95% credible interval as a significance
797 criterion (ω>1 for >97.5% of the MCMC samples). (The number of weakly selected codons using
798 80%CI is given too.) The genes are sorted by the "max_omega_LB" which is the peak height of
799 the lower blue curve in the omega plots (lower bound of the 95%CI). The 178 significant genes
800 by the Window model (excluding PPE/PGRS genes) are marked in yellow.

801

802 **Funding Statement**

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

804

805 **Acknowledgements**

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

808

809 **CRediT authorship contribution statement**

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

812

813 **Declaration of Competing Interests**

814 The authors declare no competing interests.

REFERENCES

1. Liu, Q., et al., *Tuberculosis treatment failure associated with evolution of antibiotic resilience.* Science, 2022. **378**(6624): p. 1111-1118.
2. Chiner-Oms, A., et al., *Gene evolutionary trajectories in Mycobacterium tuberculosis reveal temporal signs of selection.* Proc Natl Acad Sci U S A, 2022. **119**(17): p. e2113600119.
3. Deng, L., et al., *Identification of positively selected genes in Mycobacterium tuberculosis from southern Xinjiang Uygur autonomous region of China.* Front Microbiol, 2024. **15**: p. 1290227.
4. Mortimer, T.D., A.M. Weber, and C.S. Pepperell, *Signatures of Selection at Drug Resistance Loci in Mycobacterium tuberculosis.* mSystems, 2018. **3**(1).
5. Farhat, M.R., et al., *Genomic analysis identifies targets of convergent positive selection in 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. Wilson, D.J. and G. McVean, *Estimating diversifying selection and functional constraint in the presence of recombination.* Genetics, 2006. **172**(3): p. 1411-25.
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. Marin, M., et al., *Benchmarking the empirical accuracy of short-read sequencing across the M. tuberculosis genome.* Bioinformatics, 2022. **38**(7): p. 1781-1787.

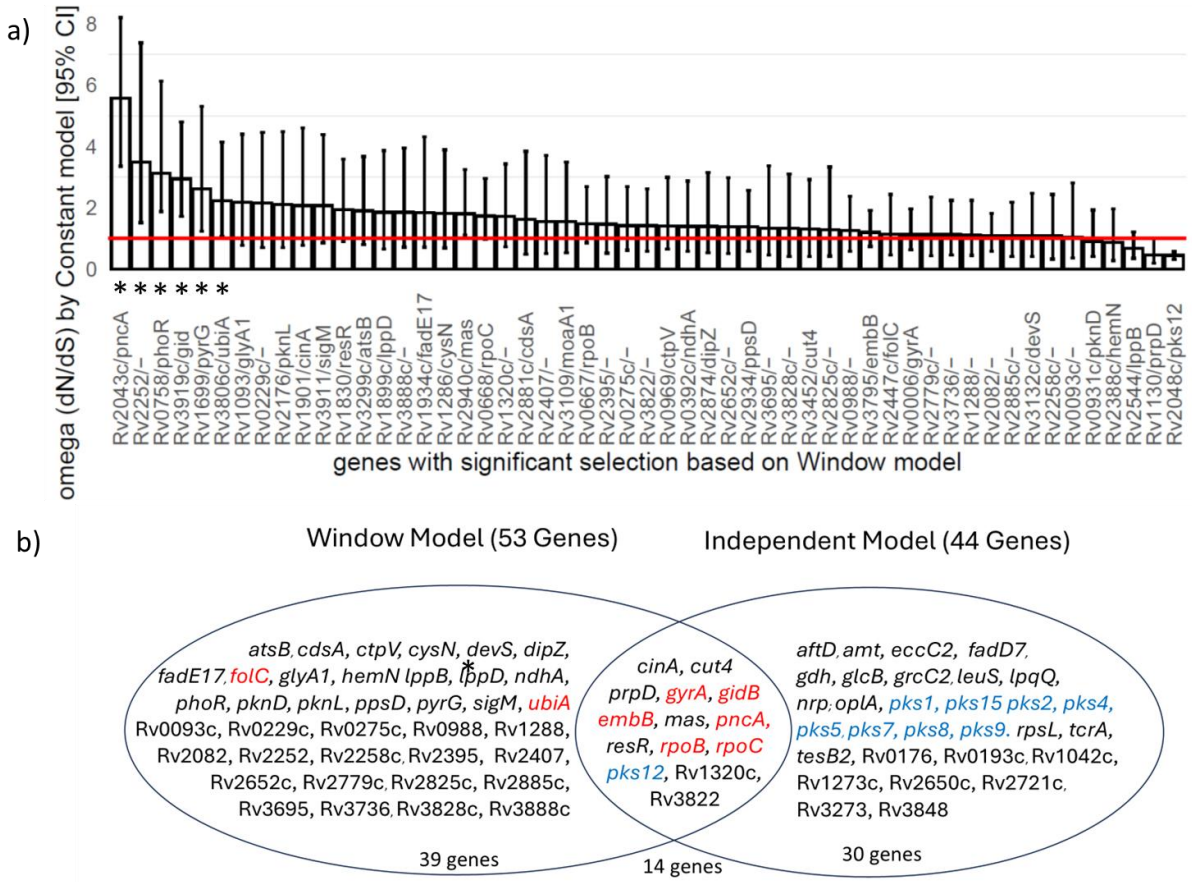
857 19. Modlin, S.J., et al., *Exact mapping of Illumina blind spots in the Mycobacterium*
858 *tuberculosis genome reveals platform-wide and workflow-specific biases.* Microb
859 Genom, 2021. **7**(3).
860 20. Sirakova, T.D., et al., *The Mycobacterium tuberculosis pks2 gene encodes the synthase for*
861 *the hepta- and octamethyl-branched fatty acids required for sulfolipid synthesis.* J Biol
862 Chem, 2001. **276**(20): p. 16833-9.
863 21. Constant, P., et al., *Role of the pks15/1 gene in the biosynthesis of phenolglycolipids in*
864 *the Mycobacterium tuberculosis complex. Evidence that all strains synthesize*
865 *glycosylated p-hydroxybenzoic methyl esters and that strains devoid of phenolglycolipids*
866 *harbor a frameshift mutation in the pks15/1 gene.* J Biol Chem, 2002. **277**(41): p. 38148-
867 58.
868 22. Matsunaga, I., et al., *Mycobacterium tuberculosis pks12 produces a novel polyketide*
869 *presented by CD1c to T cells.* J Exp Med, 2004. **200**(12): p. 1559-69.
870 23. Maruri, F., et al., *A systematic review of gyrase mutations associated with*
871 *fluoroquinolone-resistant Mycobacterium tuberculosis and a proposed gyrase numbering*
872 *system.* J Antimicrob Chemother, 2012. **67**(4): p. 819-31.
873 24. Shi, R., et al., *Emergence of ofloxacin resistance in Mycobacterium tuberculosis clinical*
874 *isolates from China as determined by gyrA mutation analysis using denaturing high-*
875 *pressure liquid chromatography and DNA sequencing.* J Clin Microbiol, 2006. **44**(12): p.
876 4566-8.
877 25. Sun, Z., et al., *Comparison of gyrA gene mutations between laboratory-selected*
878 *ofloxacin-resistant Mycobacterium tuberculosis strains and clinical isolates.* Int J
879 Antimicrob Agents, 2008. **31**(2): p. 115-21.
880 26. Blower, T.R., et al., *Crystal structure and stability of gyrase-fluoroquinolone cleaved*
881 *complexes from Mycobacterium tuberculosis.* Proc Natl Acad Sci U S A, 2016. **113**(7): p.
882 1706-13.
883 27. Sun, Q., et al., *Mutations within embCAB Are Associated with Variable Level of*
884 *Ethambutol Resistance in Mycobacterium tuberculosis Isolates from China.* Antimicrob
885 Agents Chemother, 2018. **62**(1).
886 28. Walker, T.M., et al., *The 2021 WHO catalogue of Mycobacterium tuberculosis complex*
887 *mutations associated with drug resistance: A genotypic analysis.* Lancet Microbe, 2022.
888 **3**(4): p. e265-e273.
889 29. Nan, X.T., et al., *Different Contributions of embB and ubiA Mutations to Variable Level of*
890 *Ethambutol Resistance in Mycobacterium tuberculosis Isolates.* Infect Drug Resist, 2024.
891 **17**: p. 3125-3132.
892 30. Yadon, A.N., et al., *A comprehensive characterization of PncA polymorphisms that confer*
893 *resistance to pyrazinamide.* Nat Commun, 2017. **8**(1): p. 588.
894 31. Spies, F.S., et al., *Streptomycin resistance and lineage-specific polymorphisms in*
895 *Mycobacterium tuberculosis gidB gene.* J Clin Microbiol, 2011. **49**(7): p. 2625-30.
896 32. Wong, S.Y., et al., *Functional role of methylation of G518 of the 16S rRNA 530 loop by*
897 *GidB in Mycobacterium tuberculosis.* Antimicrob Agents Chemother, 2013. **57**(12): p.
898 6311-8.

899 33. Gomes, L.C., et al., *Whole genome sequencing reveals large deletions and other loss of*
900 *function mutations in Mycobacterium tuberculosis drug resistance genes.* Microb
901 Genom, 2021. **7**(12).
902 34. Li, M.C., et al., *rpoB Mutations and Effects on Rifampin Resistance in Mycobacterium*
903 *tuberculosis.* Infect Drug Resist, 2021. **14**: p. 4119-4128.
904 35. Kreutzfeldt, K.M., et al., *CinA mediates multidrug tolerance in Mycobacterium*
905 *tuberculosis.* Nat Commun, 2022. **13**(1): p. 2203.
906 36. Chiner-Oms, A., et al., *Genomic determinants of speciation and spread of the*
907 *Mycobacterium tuberculosis complex.* Sci Adv, 2019. **5**(6): p. eaaw3307.
908 37. Bitter, W., et al., *Systematic genetic nomenclature for type VII secretion systems.* PLoS
909 Pathog, 2009. **5**(10): p. e1000507.
910 38. Tiwari, S., et al., *Infect and Inject: How Mycobacterium tuberculosis Exploits Its Major*
911 *Virulence-Associated Type VII Secretion System, ESX-1.* Microbiol Spectr, 2019. **7**(3).
912 39. Chen, J.M., et al., *Phenotypic profiling of Mycobacterium tuberculosis EspA point*
913 *mutants reveals that blockage of ESAT-6 and CFP-10 secretion in vitro does not always*
914 *correlate with attenuation of virulence.* J Bacteriol, 2013. **195**(24): p. 5421-30.
915 40. Bosserman, R.E., et al., *WhiB6 regulation of ESX-1 gene expression is controlled by a*
916 *negative feedback loop in Mycobacterium marinum.* Proc Natl Acad Sci U S A, 2017.
917 **114**(50): p. E10772-E10781.
918 41. Walters, S.B., et al., *The Mycobacterium tuberculosis PhoPR two-component system*
919 *regulates genes essential for virulence and complex lipid biosynthesis.* Mol Microbiol,
920 2006. **60**(2): p. 312-30.
921 42. Coscolla, M. and S. Gagneux, *Consequences of genomic diversity in Mycobacterium*
922 *tuberculosis.* Semin Immunol, 2014. **26**(6): p. 431-44.
923 43. Pepperell, C.S., et al., *The role of selection in shaping diversity of natural M. tuberculosis*
924 *populations.* PLoS Pathog, 2013. **9**(8): p. e1003543.
925 44. Trauner, A., et al., *The within-host population dynamics of Mycobacterium tuberculosis*
926 *vary with treatment efficacy.* Genome Biol, 2017. **18**(1): p. 71.
927 45. Goig, G.A., et al., *Ecology, global diversity and evolutionary mechanisms in the*
928 *Mycobacterium tuberculosis complex.* Nat Rev Microbiol, 2025.
929 46. Hicks, N.D., et al., *Clinically prevalent mutations in Mycobacterium tuberculosis alter*
930 *propionate metabolism and mediate multidrug tolerance.* Nat Microbiol, 2018. **3**(9): p.
931 1032-1042.
932 47. Ismail, N., et al., *Emergence of Canonical and Noncanonical Genomic Variants following*
933 *In Vitro Exposure of Clinical Mycobacterium tuberculosis Strains to Bedaquiline or*
934 *Clofazimine.* Antimicrob Agents Chemother, 2023. **67**(4): p. e0136822.
935 48. Pethe, K., et al., *Isolation of Mycobacterium tuberculosis mutants defective in the arrest*
936 *of phagosome maturation.* Proc Natl Acad Sci U S A, 2004. **101**(37): p. 13642-7.
937 49. Xu, W., et al., *Chemical Genetic Interaction Profiling Reveals Determinants of Intrinsic*
938 *Antibiotic Resistance in Mycobacterium tuberculosis.* Antimicrob Agents Chemother,
939 2017. **61**(12).
940 50. Vilcheze, C., et al., *Altered NADH/NAD+ ratio mediates coresistance to isoniazid and*
941 *ethionamide in mycobacteria.* Antimicrob Agents Chemother, 2005. **49**(2): p. 708-20.

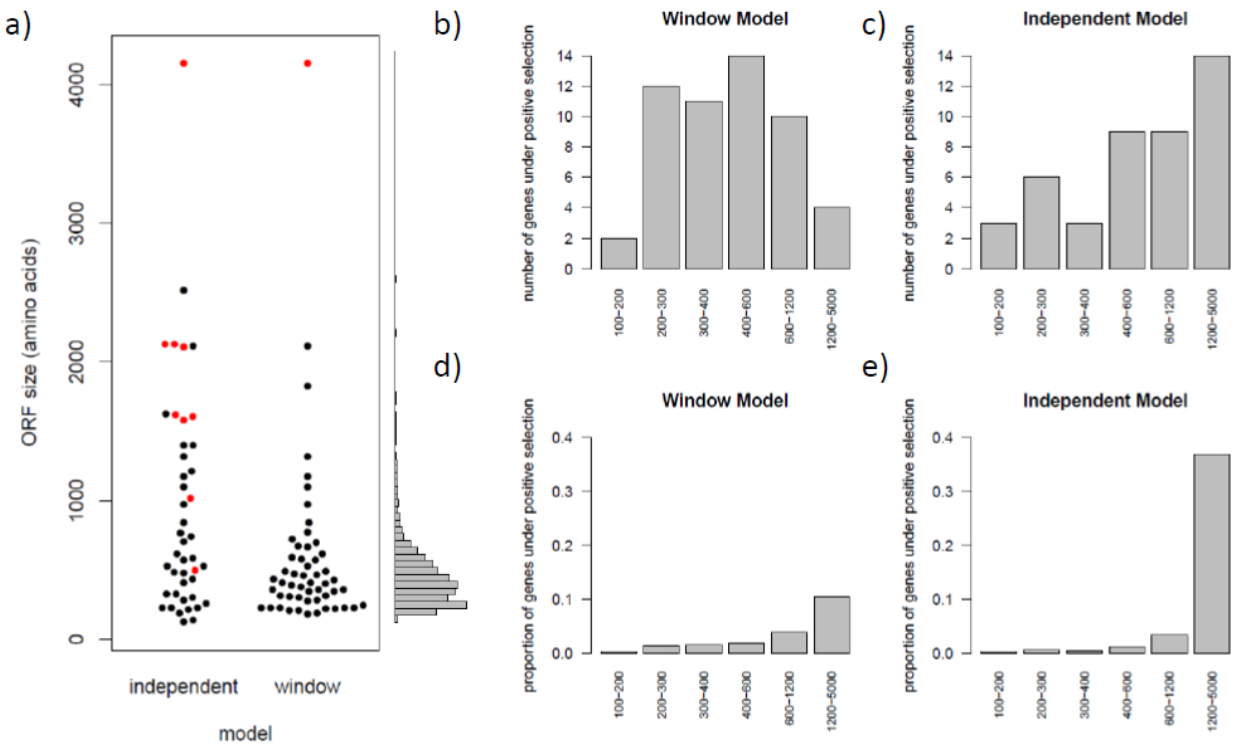
942 51. Hazbon, M.H., et al., *Population genetics study of isoniazid resistance mutations and*
943 *evolution of multidrug-resistant Mycobacterium tuberculosis.* Antimicrob Agents
944 Chemother, 2006. **50**(8): p. 2640-9.
945 52. Furio, V., et al., *An evolutionary functional genomics approach identifies novel candidate*
946 *regions involved in isoniazid resistance in Mycobacterium tuberculosis.* Commun Biol,
947 2021. **4**(1): p. 1322.
948 53. Park, H.D., et al., *Rv3133c/dosR is a transcription factor that mediates the hypoxic*
949 *response of Mycobacterium tuberculosis.* Mol Microbiol, 2003. **48**(3): p. 833-43.
950 54. Rustad, T.R., et al., *The enduring hypoxic response of Mycobacterium tuberculosis.* PLoS
951 One, 2008. **3**(1): p. e1502.
952 55. Sherman, D.R., et al., *Regulation of the Mycobacterium tuberculosis hypoxic response*
953 *gene encoding alpha -crystallin.* Proc Natl Acad Sci U S A, 2001. **98**(13): p. 7534-9.
954 56. Kumar, A., et al., *Mycobacterium tuberculosis DosS is a redox sensor and DosT is a*
955 *hypoxia sensor.* Proc Natl Acad Sci U S A, 2007. **104**(28): p. 11568-73.
956 57. Sherman, D.R., et al., *Compensatory ahpC gene expression in isoniazid-resistant*
957 *Mycobacterium tuberculosis.* Science, 1996. **272**(5268): p. 1641-3.
958 58. Rizvi, A., et al., *Rewiring of Metabolic Network in Mycobacterium tuberculosis During*
959 *Adaptation to Different Stresses.* Front Microbiol, 2019. **10**: p. 2417.
960 59. Parish, T., *Two-Component Regulatory Systems of Mycobacteria.* Microbiol Spectr, 2014.
961 **2**(1): p. MGM2-0010-2013.
962 60. Singh, P.R., et al., *Dual functioning by the PhoR sensor is a key determinant to*
963 *Mycobacterium tuberculosis virulence.* PLoS Genet, 2023. **19**(12): p. e1011070.
964 61. Martin, C., et al., *The live Mycobacterium tuberculosis phoP mutant strain is more*
965 *attenuated than BCG and confers protective immunity against tuberculosis in mice and*
966 *guinea pigs.* Vaccine, 2006. **24**(17): p. 3408-19.
967 62. Lee, J.S., et al., *Mutation in the transcriptional regulator PhoP contributes to avirulence*
968 *of Mycobacterium tuberculosis H37Ra strain.* Cell Host Microbe, 2008. **3**(2): p. 97-103.
969 63. Gonzalo-Asensio, J., et al., *Evolutionary history of tuberculosis shaped by conserved*
970 *mutations in the PhoPR virulence regulator.* Proc Natl Acad Sci U S A, 2014. **111**(31): p.
971 11491-6.
972 64. Romagnoli, A., et al., *ESX-1 dependent impairment of autophagic flux by Mycobacterium*
973 *tuberculosis in human dendritic cells.* Autophagy, 2012. **8**(9): p. 1357-70.
974 65. Champion, P.A., et al., *ESX-1 secreted virulence factors are recognized by multiple*
975 *cytosolic AAA ATPases in pathogenic mycobacteria.* Mol Microbiol, 2009. **73**(5): p. 950-
976 62.
977 66. Chirakos, A.E., et al., *Conserved ESX-1 Substrates EspE and EspF Are Virulence Factors*
978 *That Regulate Gene Expression.* Infect Immun, 2020. **88**(12).
979 67. Conrad, W.H., et al., *Mycobacterial ESX-1 secretion system mediates host cell lysis*
980 *through bacterium contact-dependent gross membrane disruptions.* Proc Natl Acad Sci U
981 S A, 2017. **114**(6): p. 1371-1376.
982 68. Zheng, W., et al., *Single-cell analysis reveals Mycobacterium tuberculosis ESX-1-mediated*
983 *accumulation of permissive macrophages in infected mouse lungs.* Sci Adv, 2025. **11**(3):
984 p. eadq8158.

69. Beckham, K.S.H., et al., *Structure of the mycobacterial ESX-5 type VII secretion system pore complex.* Sci Adv, 2021. **7**(26).
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).
73. Domenech, P. and M.B. Reed, *Rapid and spontaneous loss of phthiocerol dimycocerosate (PDIM) from Mycobacterium tuberculosis grown in vitro: implications for virulence studies.* Microbiology (Reading), 2009. **155**(Pt 11): p. 3532-3543.
74. Baena, A., et al., *Large genomic deletions delineate Mycobacterium tuberculosis L4 sublineages in South American countries.* PLoS One, 2023. **18**(5): p. e0285417.
75. Alland, D., et al., *Role of large sequence polymorphisms (LSPs) in generating genomic diversity among clinical isolates of Mycobacterium tuberculosis and the utility of LSPs in phylogenetic analysis.* J Clin Microbiol, 2007. **45**(1): p. 39-46.
76. Gomez-Gonzalez, P.J., et al., *Functional genetic variation in pe/ppe genes contributes to diversity in Mycobacterium tuberculosis lineages and potential interactions with the human host.* Front Microbiol, 2023. **14**: p. 1244319.
77. Oppong, Y.E.A., et al., *Genome-wide analysis of Mycobacterium tuberculosis polymorphisms reveals lineage-specific associations with drug resistance.* BMC Genomics, 2019. **20**(1): p. 252.
78. Sampson, S.L., *Mycobacterial PE/PPE proteins at the host-pathogen interface.* Clin Dev Immunol, 2011. **2011**: p. 497203.
79. DeJesus, M.A., et al., *Comprehensive Essentiality Analysis of the Mycobacterium tuberculosis Genome via Saturating Transposon Mutagenesis.* mBio, 2017. **8**(1).
80. Babu Sait, M.R., et al., *PPE51 mediates uptake of trehalose across the mycomembrane of Mycobacterium tuberculosis.* Sci Rep, 2022. **12**(1): p. 2097.
81. Dechow, S.J., et al., *ppe51 Variants Enable Growth of Mycobacterium tuberculosis at Acidic pH by Selectively Promoting Glycerol Uptake.* J Bacteriol, 2022. **204**(11): p. e0021222.
82. De Maio, F., et al., *PE_PGRS proteins of Mycobacterium tuberculosis: A specialized molecular task force at the forefront of host-pathogen interaction.* Virulence, 2020. **11**(1): p. 898-915.
83. Koleske, B., et al., *In vivo profiling of the PE/PPE proteins of Mycobacterium tuberculosis reveals diverse contributions to virulence.* Front Microbiol, 2025. **16**: p. 1634229.
84. Talarico, S., et al., *Association of Mycobacterium tuberculosis PE PGRS33 polymorphism with clinical and epidemiological characteristics.* Tuberculosis (Edinb), 2007. **87**(4): p. 338-46.

Figure 1. Comparison of genes under significant positive selection in Moldova dataset using 3 models. a) Barchart showing the omega estimates by the gene-level analysis (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. b) Venn diagram shows overlap of 53 significant genes from Window model with 44 genes from codon-level analysis (Independent model). Drug-resistance genes are marked red; polyketide synthases are marked blue.



1039 **Figure 2.** The codon-level analysis (Independent model) is biased toward larger genes. a)
1040 Beeswarm plots showing that the significant genes under positive selection by the Independent
1041 model are skewed larger than those from the Window model, whereas the Window model
1042 better matches the overall ORF size distribution (histogram to right). Polyketide synthase genes
1043 are colored red. b) and c) Number of genes under positive selection, divided into size bins
1044 (amino acids). The Independent model shows an increasing trend toward the rightmost bin
1045 (ORFs with 1200-5000 amino acids). d) and e) Proportion of genes under positive selection
1046 (counts of significant genes normalized by the overall number of genes in each size bin).

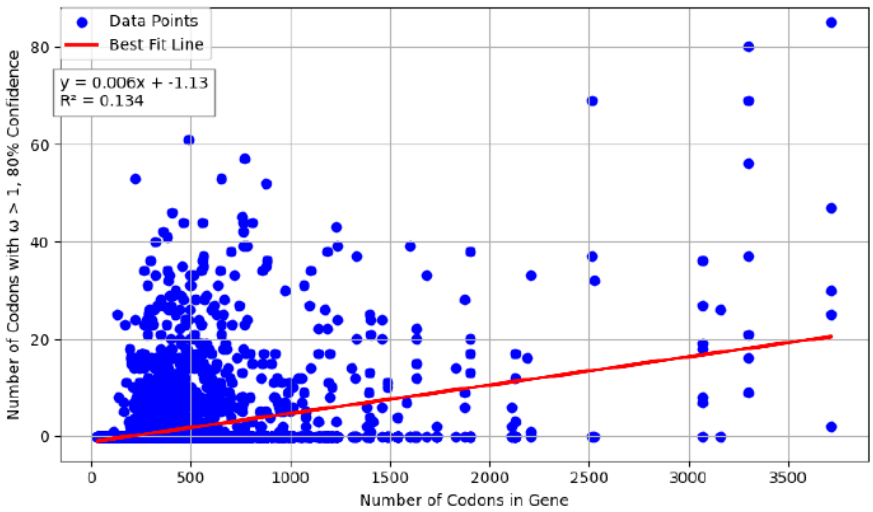


1047

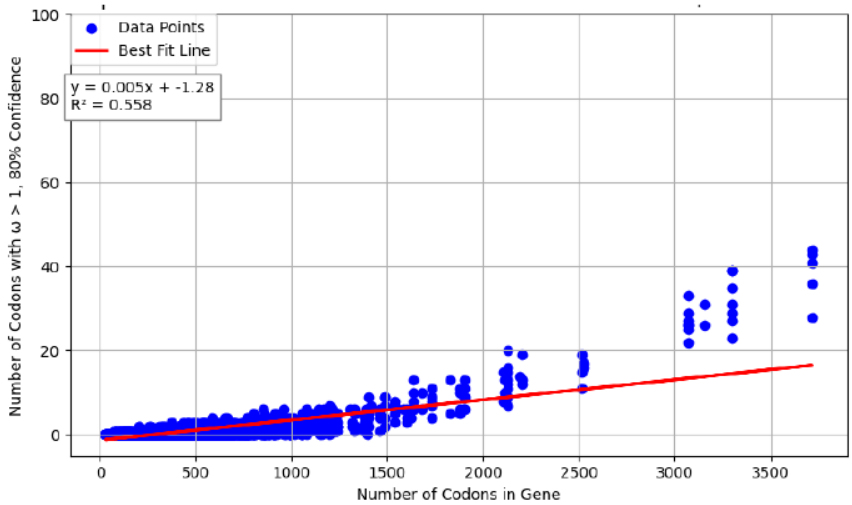
1048

Figure 3. 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 (codon-level analysis). The Independent model exhibits a strong correlation with ORF size ($r^2 = 0.558$) compared to the Window model ($r^2 = 0.134$).

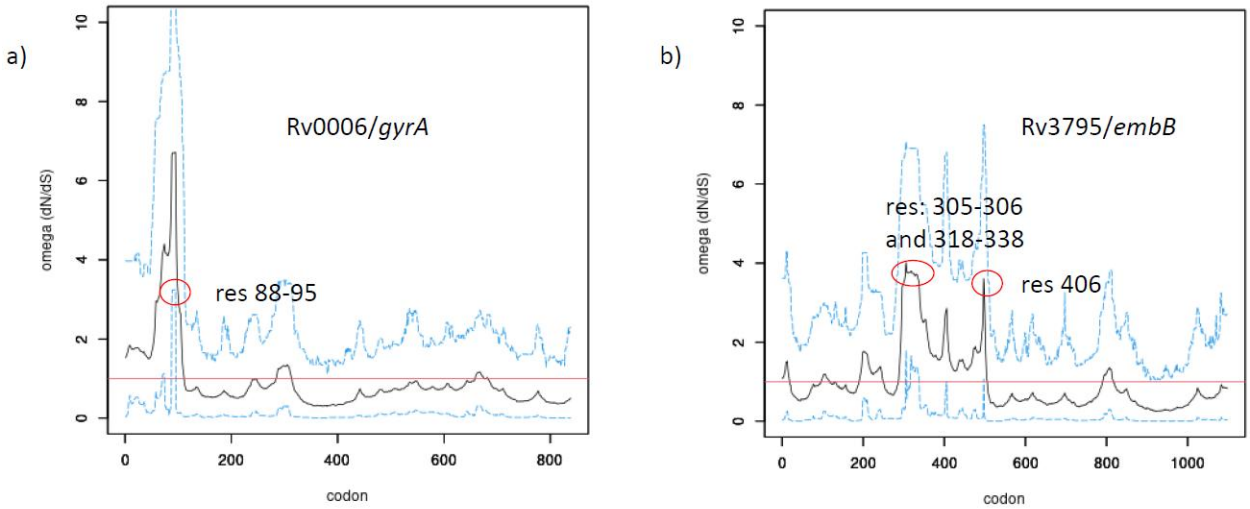
a) Window model



b) Independent model



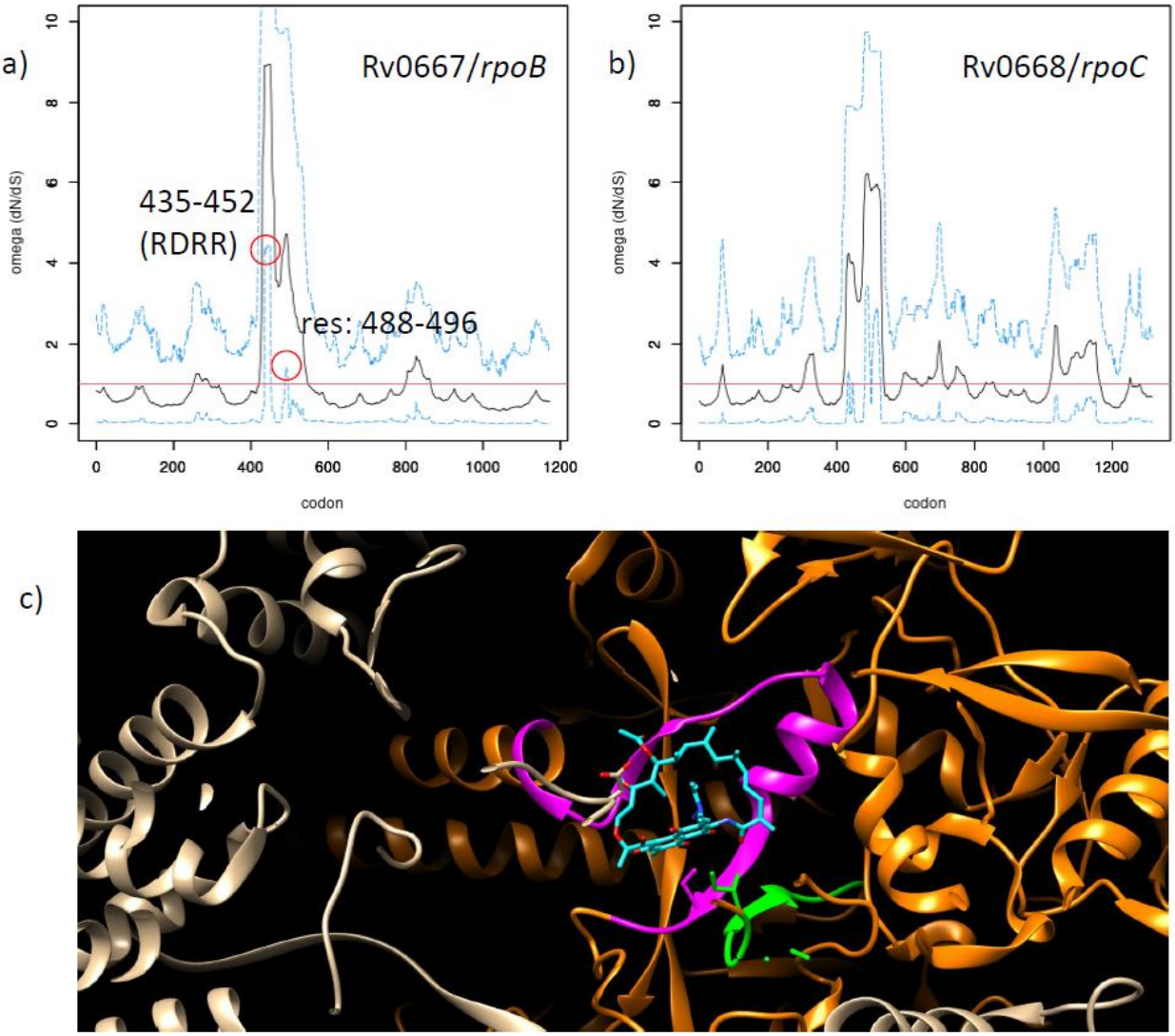
1057 **Figure 4**. Example omega plots for *gyrA* and *embB*, showing the dN/dS for each codon
1058 estimated by the Window model. The black curve shows the mean omega value for each codon
1059 averaged over the MCMC samples; the blue lines show the upper- and lower-bounds of the 95%
1060 credible interval. Codons under significant positive selection are those where the lower blue
1061 line peaks above 1.0.



1062

1063

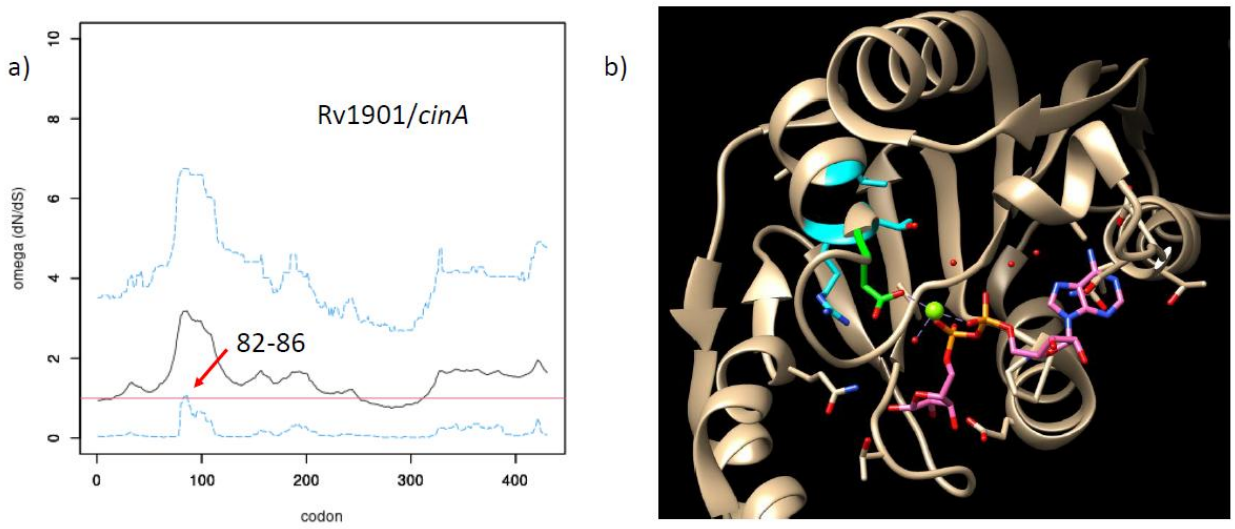
1064 **Figure 5.** Analysis of positive selection in RpoB and RpoC. a) Omega plot for *rpoB*. b) omega
1065 plot for *rpoC*. c) X-ray crystal structure of complex of RNA polymerase (*rpoABC*) (PDB: 5uhc)
1066 with rifampicin (cyan), showing that 488-496 (green) forms a parallel beta-strand to 426-452
1067 (purple); both are regions under positive selection.



1068

1069

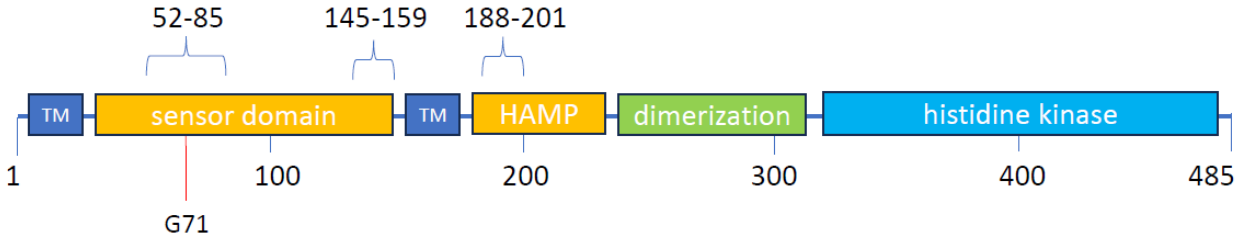
1070 **Figure 6.** Analysis positive selection in CinA. a) Omega plot for *cinA*, showing selection of
1071 residue 82-86 (where lower blue line peaks above 1.0). b) Crystal structure of CinA from *T.*
1072 *thermophilus* (N-terminal domain) (PDB: 4uux); green: Asp80 (proximal to NADH in active site,
1073 purple), blue residues: codons under positive selection.



1074

1075

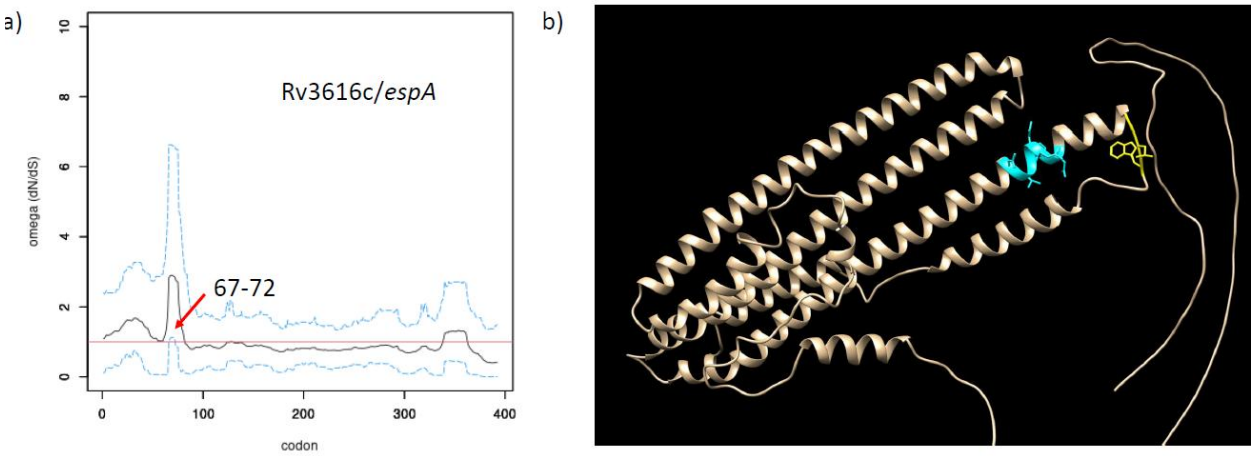
1076 **Figure 7.** Regions under positive selection in *phoR*, mapped onto domain structure. 'TM' refers
1077 to transmembrane domains. The residue (G71) that differentiates the *Mtb* and *M. bovis*
1078 orthologs is indicated.



1079

1080

1081 **Figure 8.** Analysis of positive selection for EspA. a) Omega plot for *espA,* showing significant
1082 selection for codons 67-72. b) AlphaFold model of EspA. Residues under positive selection (67-
1083 72) are colored blue. The WXG motif (residues 55-57), involved in recognition by the ESX-1 pore
1084 complex, are colored blue.



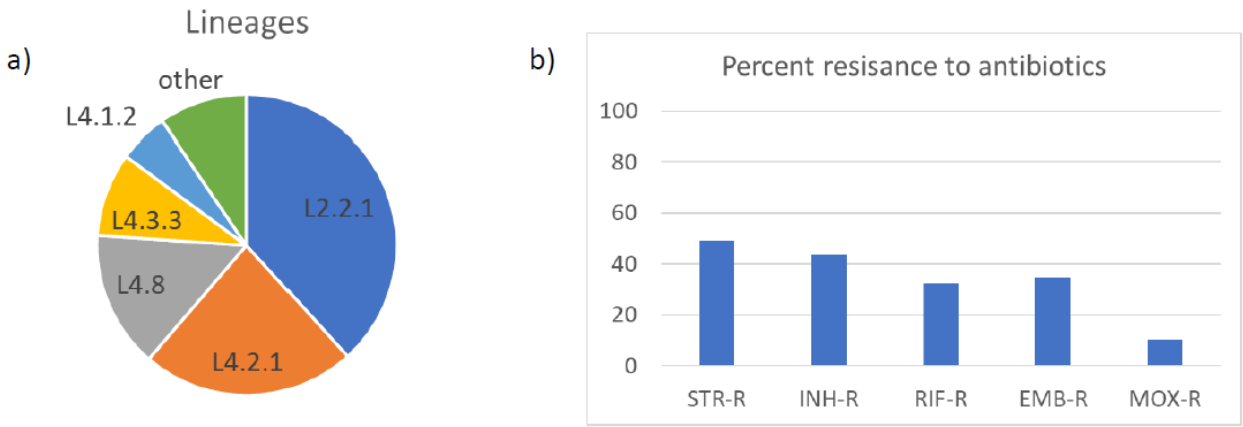
1085

1086

1087

1088 **Supplemental Figure S1.** Lineage distribution and drug-resistance profile for 2,057 Mtb clinical
1089 isolates from Moldova (as determined by TB-profiler).

1090 .

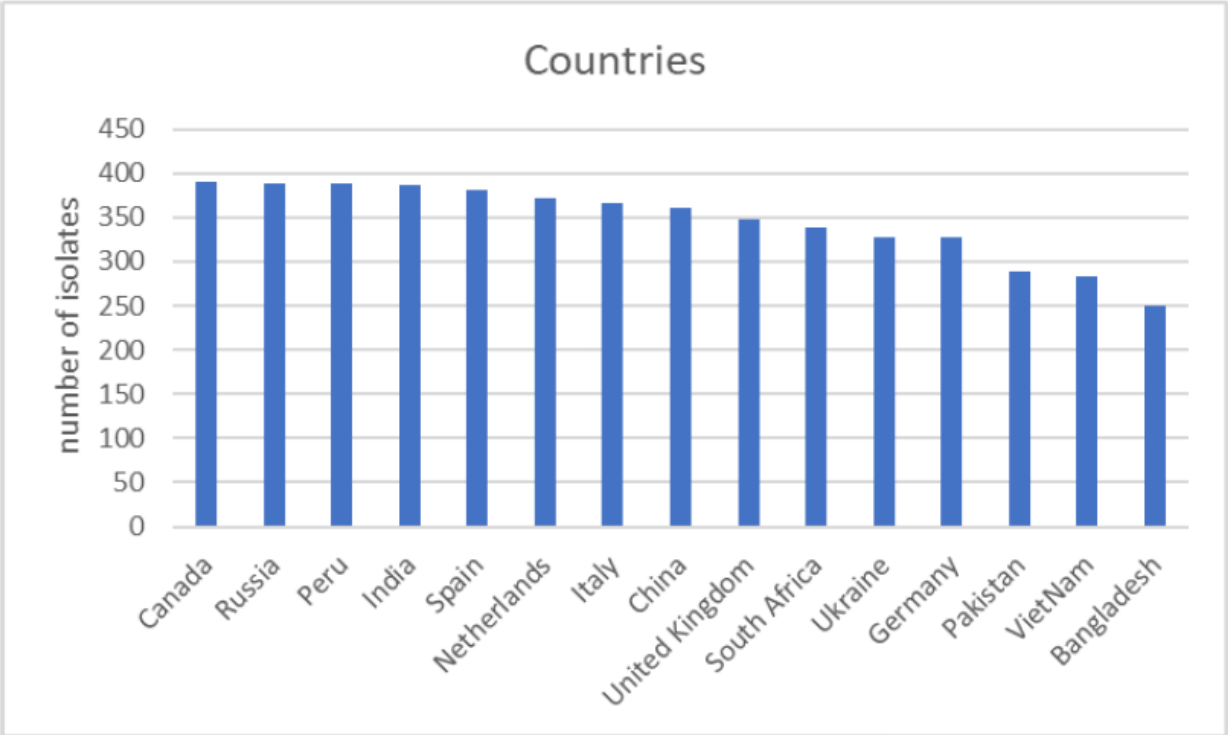


1091

1092

1093

1094 **Supplemental Figure S2.** Countries represented in global collection of 5,195 Mtb clinical isolates
1095 (from CRyPTIC collection).

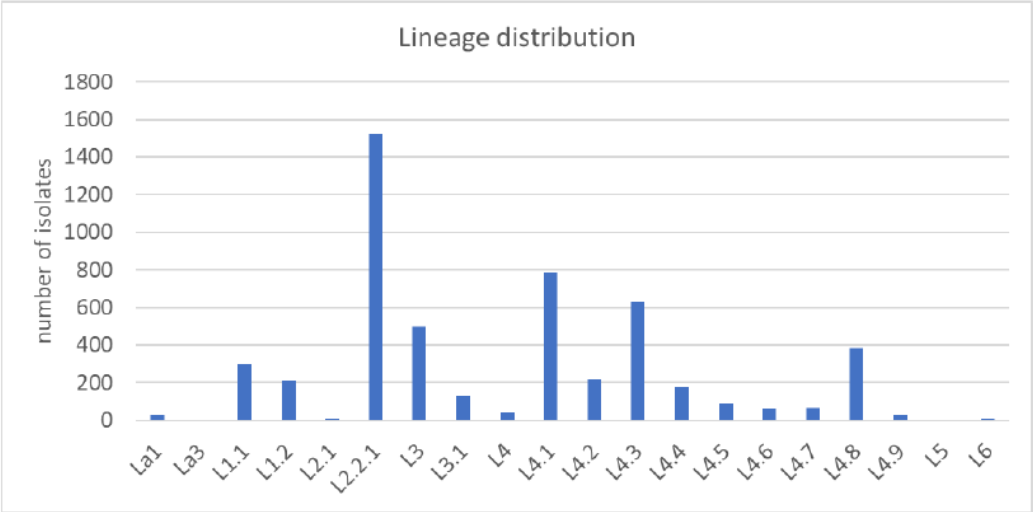


1096

1097

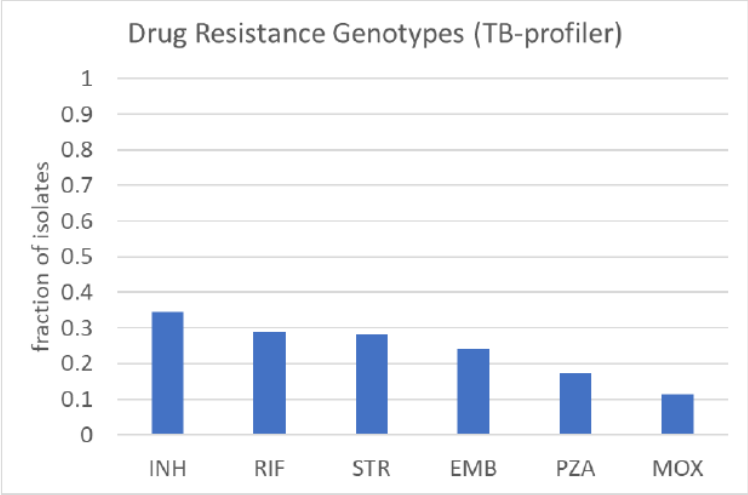
1098 **Supplemental Figure S3.** Lineages and drug-resistance genotypes represented in global
1099 collection of 5,195 Mtb clinical isolates.

a)



1100

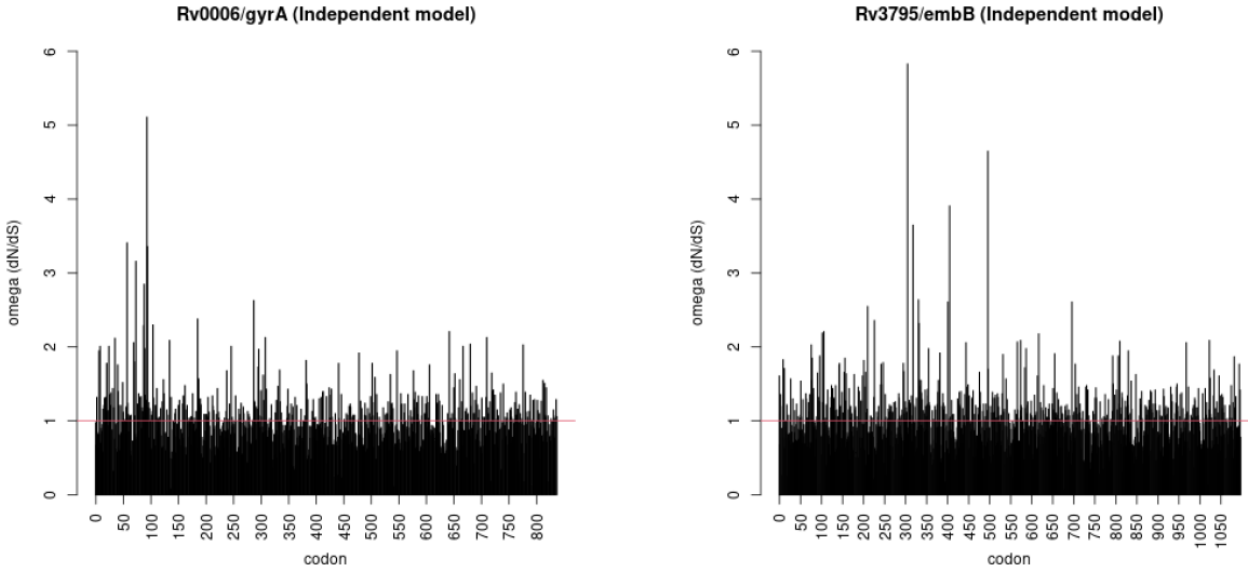
b)



1101 b)

1102

1103 **Supplemental Figure S4.** Omega plots for *gyrA* and *embB* based on the Independent model
1104 (codon-level analysis).

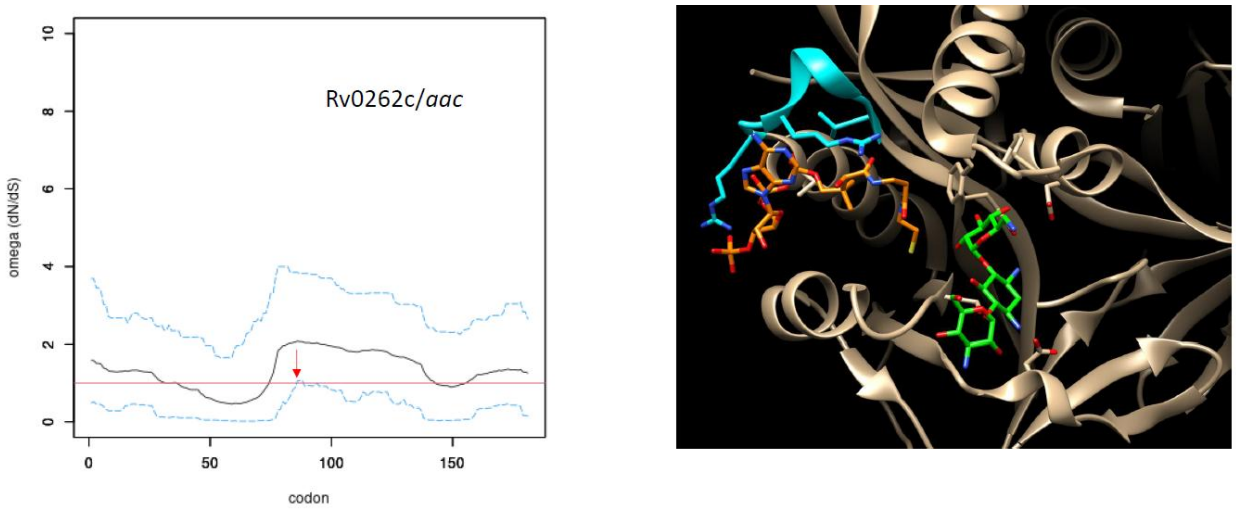


1105

1106

1107

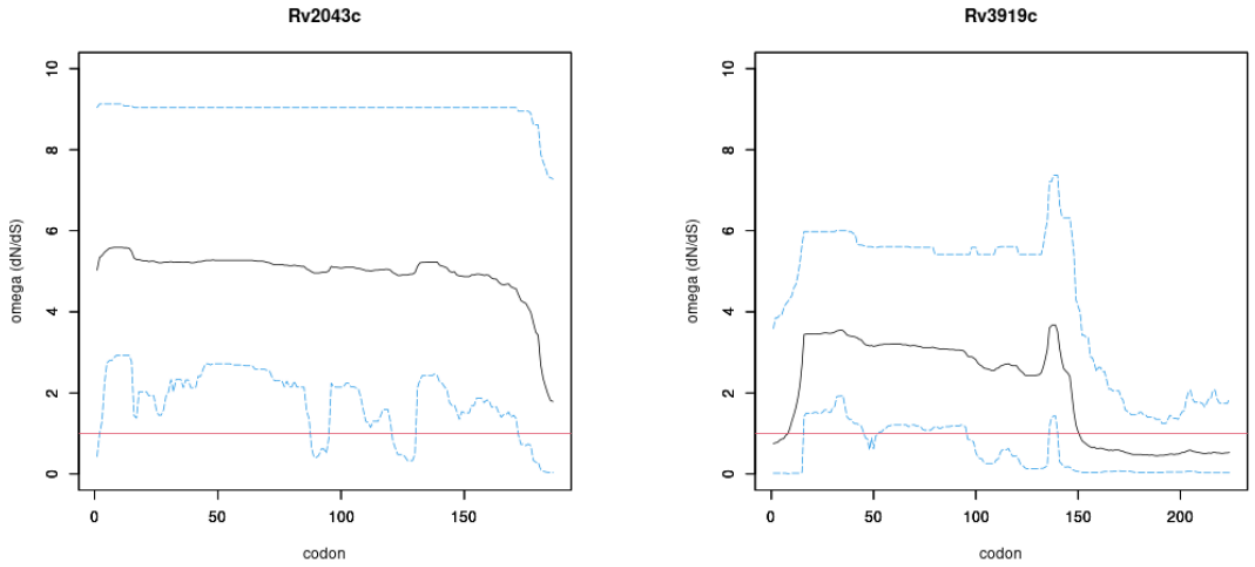
1108 **Supplemental Figure S5**. Analysis of positive selection for *aac* (Rv0262c, aminoglycoside 2'-N-
1109 acetyltransferase). a) Omega plot for *aac*: b) Genetic variants in *aac* mapped onto crystal
1110 structure (PDB: 1M4I) of complex of Mtb *aac* with CoA (orange) and kanamycin (green);
1111 residues under selection (86-94) are marked in blue.



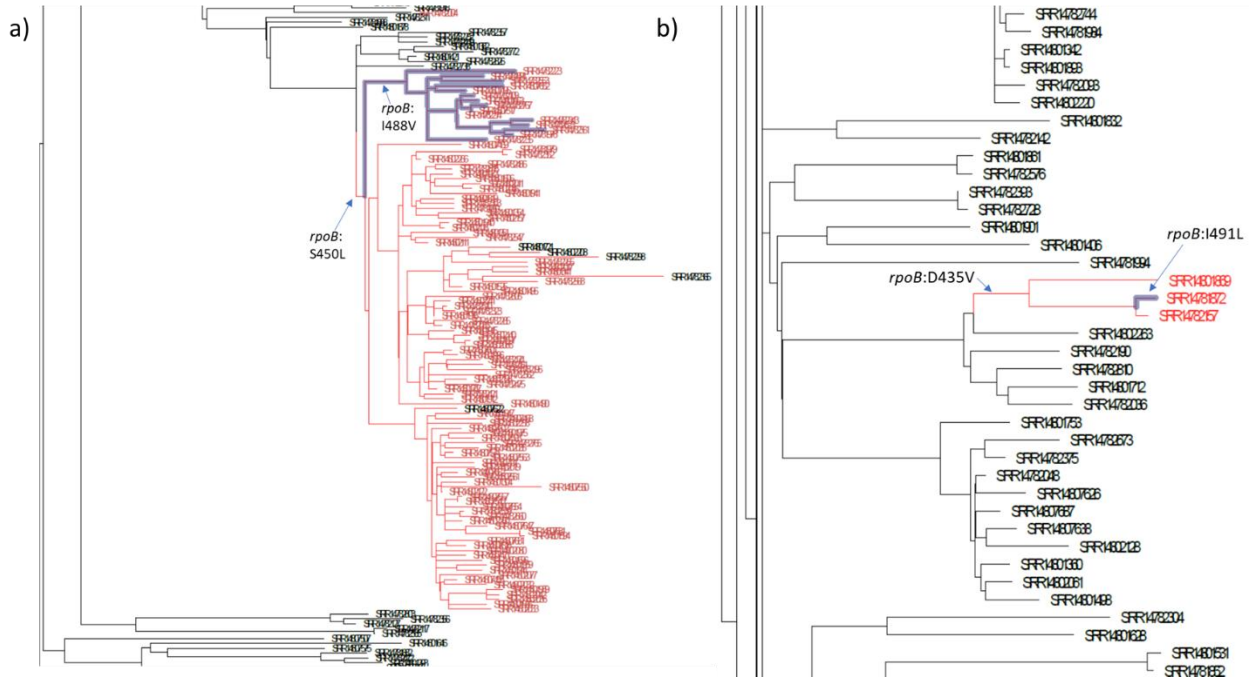
1112

1113

1114 **Supplemental Figure S6.** Omega plots for *pncA* (Rv2043c) and *gidB* (Rv3919c) based on the
1115 Window model, showing a high fraction of the coding regions under positive selection (lower
1116 blue line > 1).



1117

1118

Supplemental Figure S7. Portions of the phylogeny of Moldova isolates showing the order of acquisition of RpoB mutations (the leaves are SRA accession numbers of isolates). a) The *rpoB*:I488V mutation in the highlighted clade of 15 isolates occurred after the *rpoB*:S450L mutation (111 isolates, taxa highlighted red), within the L2.2.1 lineage. b) The *rpoB*:I491L mutation in a single isolate (highlighted branch) occurred after the *rpoB*:D435V mutation within a clade of 3 isolates (red) in the L4.3.3 lineage.