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

ABSTRACT

Several recent studies have reported that multiple genes in the *M. tuberculosis* (Mtb) genome are under positive selection, based on analysis of large collections of Mtb clinical isolates. It is inferred that mutations in these genes are 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. Although the calculations are based on a variety of evolutionary principles, most of these studies have evaluated positive selection at the gene-level. However, positive selection can be evaluated on different scales, including individual codons, and local regions within an open reading frame (clusters of residues). In this paper, we use GenomegaMap to evaluate selection of genes in the Mtb genome at all three levels, and we present evidence that the intermediate level (so-called 'Window model') provides the most credible list of candidate genes under selection. A further advantage of this approach is that it identifies specific regions within an ORF that harbor the selection signal (clusters of SNPs with a local excess of non-synonymous changes), which can be useful for structural and functional analysis. In an analysis of 2,057 Mtb clinical isolates from Moldova, we observe 53 statistically significant genes under selection, including expected drug-resistance genes, as well as *resR* and *phoR* (regulators of antibiotic tolerance, and virulence/stress response, respectively), that have previously been identified to be under selection in Mtb. To evaluate the generality of the results, the analysis was then repeated on a larger, global set of 5,195 clinical isolates from 15 other countries. While many of the same genes were under significant selection in both datasets, the global dataset also revealed positive selection of multiple ESX-1 genes, along with *whiB6* and *phoR*, which have been shown to regulate expression of ESX-1, a type VII secretion system that plays an important role in modulating the host immune response. The selection signals for these ESX-

1 genes lend support to the potential selective advantage of genetic variants affecting ESX-1 function, as adaptations to modern treatment of TB.

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 (Chiner-Oms et al., 2022; Deng et al., 2024; Liu et al., 2022). 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 (Mortimer et al., 2018), and other measures of convergent evolution (Holt et al., 2018) 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. In fact, genes associated with drug resistance, are often found to have among the highest selection signals (Farhat et al., 2019). 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 (Comas et al., 2011).

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 (Nei, 2005). 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 (Nielsen & Yang, 1998) 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 (Wilson & Consortium, 2020). Observed mutations are mapped onto a phylogenetic tree to identify unique evolutionary events, and then the relative number of synonymous and nonsynonymous events is used to compute posterior distributions over the parameters (including ω) under different assumptions (such as whether there is a constant molecular clock, or rates can differ independently in each branch).

PAML (Yang, 2007) is a widely-used tool for evaluating selection in sequence data, which uses maximum-likelihood (ML) to estimate the model parameters. Unfortunately, PAML does not scale-up well, and is difficult to run for collections of thousands of genomes, preventing its use in studies such as (Liu et al., 2022). 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 (Wilson & Consortium, 2020). 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.

83 GenomegaMap uses Markov Chain Monte Carlo (MCMC) methods (described in more detail
84 below) to infer selection parameters for mutations observed in genes. GenomegaMap was in
85 fact applied to analyze >10,000 Mtb genome sequences from the CRyPTIC collection
86 (demonstrating its scalability) and proposed a list of genes under positive selection, which
87 included many drug-resistance-related genes (consistent with expectations).

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

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

111 selection signal (over adjacent codons) by identifying regions within the ORF where there is a
112 cluster of SNPs with a localized over-representation of non-synonymous to synonymous SNPs.

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

117 Comparing the quality of model fits based on total likelihood, Wilson (2020) concluded that the
118 Individual model (identifying genes with individual codons under positive selection) fit the data
119 better than the Window model. However, using a criterion of Pr(ω>1)>0.9 (dN/dS>1 in >90% of
120 MCMC samples), thousands of genes were found to have at least one codon under selection
121 (see Fig 3 in Wilson). While many drug-resistance-related genes were included, when genes
122 were ranked by the proportion of codons under selection, some expected genes, like *gyrA, rpoB,*
123 and *embB*, were found to have ranks in the range of 2000-3000 (out of ~4000 genes). For the
124 purposes of prioritization, it would be desirable to have a method that ranks such genes closer
125 to the top.

126 In this paper, we empirically compare the three different models of selection – at the gene-level,
127 codon-level, and in local windows (gene segments) – by evaluating them on analysis of a set of
128 ~2000 Mtb clinical isolates from Moldova (Yang et al., 2022). While thousands of genes have
129 ω>1 (as a point estimate), determining which genes are statistically significant depends critically
130 on the MCMC sampling (through credible intervals). We show that using tighter criteria for
131 significance than were used in (Wilson & Consortium, 2020) produces a more focused list of
132 candidates for genes under positive selection. While many expected genes that have been
133 shown to be associated with drug resistance appear on both lists, we show that the Window
134 model often better identifies regions within these proteins that are known to be antibiotic
135 binding sites based on X-ray crystal structures. In contrast, we show that the Individual model
136 has a bias toward identifying more codons under selection in large ORFs, suggesting it is more
137 susceptible to noise. Based on these observations, we argue that the Window model in
138 GenomegaMap is the most reliable approach for identifying genes under selection, which

139 results from the smoothing effect achieved by modeling localized variations in ω within regions
140 of a protein (with adjustable borders). As validation, we show that, using the Window model, a
141 similar list of genes under positive selection is derived from an independent set of over 5000
142 Mtb clinical isolates from 15 other countries (from the CRyPTIC collection).

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

146

147 **METHODS**

148 Raw sequencing data (fastq files) for isolates from Moldova were downloaded from SRA using
149 accession numbers provided in (Yang et al., 2022). Sequencing datasets from the CRyPTIC
150 collection were downloaded from Supp Table 3 in (Wilson & Consortium, 2020). While isolates
151 could not be selected based on lineage (data not available), to promote genetic diversity, we
152 performed stratified sampling by selecting approximately 500 isolates each from 15 different
153 countries. TB-profiler (Phelan et al., 2019) was run on the isolates to identify Mtb lineage, drug-
154 resistance genotypes, and estimated depth of coverage (using *M. tuberculosis* H37Rv as a
155 reference). Isolates that had < 40x coverage or were potentially mixed (with more than one
156 sublineage represented, based on characteristic SNPs) were discarded.

157 Next, Pilon (Walker et al., 2014) was used to map read to the H37Rv genome (NC_000962.3_
158 and extract genetic variants (in the form of .vcf files), including SNPs and indels; Pilon also
159 identifies low-coverage and heterogeneous sites. From the .vcf files, the nucleotide sequences
160 of each ORF were extracted from each isolate, and aggregated by ORF (as .fasta files). Codons
161 with low-coverage or heterogeneous nucleotides, or nonsense mutations (stop codons) were
162 converted to gap characters ('---'), which are subsequently disregarded in the calculations by
163 GenomegaMap.

164 GenomegaMap was then run on each ORF using scripts the for Constant, Individual, and
165 Window models (https://github.com/danny-wilson/genomegaMap). For the MCMC sampling

166 for Individual and Window, the trajectory length was 100,000 iterations, saving parameters
167 every 250 steps, and eliminating the first 20% of samples (80 out of 400) as the burn-in phase
168 (to reduce the influence of initial random parameters and ensure the sampler had reached
169 equilibrium). After the MCMC sampling, the value of ω for each codon was summarized over
170 each trajectory by extracting the mean and other statistics (like range and variance), and a 95%
171 credible interval (95%CI) was calculated for the ω value for each codon.

172 129 PPE and PE_PGRS genes were systematically excluded from the analysis, as they harbored
173 apparent mutations that caused many genes in these two families to exhibit selection signals,
174 but these might be artifacts due to uncertainties in sequencing associated with these genes
175 (Marin et al., 2022).

176 Generation of simulated genes for comparison of GenomegaMap models under the assumption
177 of neutral evolution was done as follows. First, a phylogenetic tree based on the 2057 Moldova
178 clinical isolates was constructed using RAxML (Stamatakis, 2014). Then, the least common
179 ancestral sequence (LCA) was computed for each ORF in the Mtb genome (excluding PPE and
180 PE_PGRS genes, and genes known to be involved in drug resistance). The LCA nucleotide at
181 each site in a given gene was derived by using Fitch's algorithm (Yang et al., 2011) to propagate
182 observed nucleotides from leaves (2057 isolates) up to the root. Then, each distinct mutational
183 event (i.e. SNP associated with a particular branch point in the tree) was randomly assigned to a
184 new site within a new gene. This resampling process resulted in 39,564 mutations permuted to
185 new positions among 3,446,666 sites within coding regions of 3877 genes (with the same size
186 distribution as in the Mtb genome), which was then used to generate the 2057 variant
187 sequences for each isolate for each gene.

188

189 **RESULTS**

190 We used GenomegaMap to investigate genes under positive selection in a collection of 2236
191 Mtb clinical isolates from Moldova, representing all reported cases in 2018-2019 (Yang et al.,
192 2022). The aim of the study to identify transmission clusters within the region. The isolates

193 were found to be predominantly members of lineage 2 and lineage 4, and approximately one-
194 half of the isolates were (genotypically) drug-resistant (based on DR markers to one or more
195 antitubercular drugs), with 32.7% MDR-TB (INH+RIF) (lineage and DR distributions are shown in
196 Supp Fig S1). After downloading and preprocessing the raw sequencing data and filtering out
197 low-quality samples, 2057 isolates were selected for analysis.

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

199 With the Window model (where different regions within each ORF are modeled with locally-
200 constant omega values), the MCMC simulation produces a sample of parameter combinations,
201 including a representation of the posterior probability distribution of omega values for each
202 codon. Besides the mean value of omega for each codon, the dispersion of the sample can be
203 used to assess confidence. In (Wilson & Consortium, 2020), the criterion used for significance
204 was if ≥ 90% of the samples had ω >1. This corresponds to a (two-sided) 80% credible interval,
205 which is overly generous, potentially generating many false positive codons and genes under
206 selection (245 genes by this criterion). Instead, we employ a more stringent criterion of ≥ 97.5%
207 of samples having ω >1, equivalent to a (two-sided) 95% credible interval. Using this criterion,
208 53 genes were identified as having at least one codon under positive selection (Table 1). (dN/dS
209 scores and statistical significance for all genes are provided in Supplemental Table T1.)

210

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

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

Rv1093/*glyA1*	Rv2176/*pknL*	Rv2885c/-	Rv3888c/-
Rv1130/*prpD*	Rv2252/-	Rv2934/*ppsD*	Rv3911/*sigM*
			Rv3919c/*gid*

213

214

215 The 53 genes under positive selection in the Moldova dataset (by the Window method) contain
216 many genes known to be involved in drug resistance, including *rpoB* (RNA polymerase, target of
217 rifampicin), *gyrA* (DNA gyrase, target of fluroquinolones, *gidB* (ribosome methyltransferase,
218 affects sensitivity to streptomycin), *pncA* (pyrazinamidase, activator of pyrazinamide), and *embB*
219 (target of ethambutol). These will be discussed in more detail below. *resR* (Rv1830) also
220 exhibits positive selection in this dataset, consistent with the findings in (Liu et al., 2022), where
221 it was shown to be associated with general resilience to antibiotics. *phoR* is a two-component
222 sensor implicated in stress response and virulence (Broset et al., 2015; Parish, 2014) that has
223 been shown to be under positive selection in Mtb using phylodynamics (i.e. increase in non-
224 synonymous mutation rate, coincident with expansion into humans as pathogen) (Chiner-Oms
225 et al., 2019). Other genes under selection that can play a regulatory role and potentially
226 adaptive role include *pknD* (protein Ser/Thr kinase), *devS* (two-component sensor, regulator of
227 hypoxic response (Kumar et al., 2007)), *sigM* (a sigma factor that could influence ESX-4 or PDIM
228 expression (Raman et al., 2006)), and Rv1320 (an adenylate cyclase, which are often involved in
229 signaling/regulation). None of the three models identified *katG* (activator of isoniazid) as
230 positively selected in this collection. Wilson (2020) also noted that *katG* was not found to be
231 under positive selection by the Window model in their larger dataset, but was identified by the
232 Independent model (because the majority of INH-resistance mutations occur in a single codon:
233 Ser315). However, this could also be easily identified by homoplasy analysis (Mortimer et al.,
234 2018), as mutations at this site often occur multiple times independently in clinical populations.

235 The evaluation of genes under positive selection using the Independent model yielded 44
236 significant genes, using the more stringent criterion of at least 1 codon with $\omega>1$ in $\geq 97.5\%$ of
237 the MCMC samples. This list of 44 genes has some overlap with the 53 significant genes by the
238 Window model; there are 19 genes in the intersection, including most of the drug-resistance-

239 related genes (see Venn diagram in Figure 1a). However, there are also many unique genes
240 identified by each of the analysis methods. The Window model also highlighted *ubiA*
241 (arabinogalactan synthesis pathway; harbors resistance mutations to ethambutol) and *folC*
242 (resistance mutations to inhibitors of folate pathway like PAS or sulfamethoxazole), which the
243 Independent model did not.

244 With the Constant model (5000 MCMC iterations), 1924 genes have ω>1, but only 61 are
245 significant (LB of 95%CI > 1). Two-thirds of these are PPE or PE_PGRS genes; after filtering these
246 out, 22 genes remain that exhibit statistically significant positive selection. The only drug-
247 resistance-related genes among these are *pncA, gidB,* and *ubiA*. Other DR genes like *rpoB, gyrA,*
248 *embB,* and *katG* did not reach significance using this model. Many of the significant genes by
249 the Window model have ω below or only marginally above 1 in the Constant model, including
250 many of the genes implicated in drug resistance mentioned above (see Figure 1b). However,
251 the lower bound of the 95% credible interval is less than 1.0 for almost all of the 53 genes
252 significant by the Window model. For example, *embB* has a gene-level ω 1.19, but the 95%CI
253 lower bound is 0.73 (not significant), and is ranked #141. Only 6 genes are significant in both
254 models (including *pncA, gidB,* and *ubiA*). The fact that the Window model captures more drug-
255 resistance-related genes reinforces the utility of focusing the analysis on selected regions within
256 a gene (Window model), whereas the selection signal at the whole-gene level (Constant model)
257 can get diluted by SNPs in other less-relevant parts of the protein.

258

259 **Genes significant by Window Model exhibit selection in regions of protein associated with**
260 **drug resistance.**

261 One of the advantages of evaluating genes for positive selection using the Window Model is
262 that it identifies segments of specific amino acids harboring the signal (rather than requiring a
263 constant omega parameter for the entire gene). In the case of genes commonly associated with
264 drug resistance, the regions under selection often co-occur with known drug-binding regions of
265 the protein. This co-location of selection signals with active sites or structural binding sites for

266 antibiotics lends confidence that the Window model is detecting genuine signals of selective
267 pressure.

268 The localization of the selection signals within each ORF can be observed in omega plots, as
269 illustrated in Figure 2 (omega plots for *gyrA* and *embB*). These plots show the mean of the
270 posterior distribution of omega (average of the MCMC samples) for each codon (black line).
271 Note that the curve is smoothed, which is a consequence of the Window model, where the
272 omega value for one codon effectively combines local information on genetic variants at
273 surrounding codons by assuming the codons within discrete regions of the ORF share a common
274 omega value). (The boundaries of the regions get adjusted dynamically during the MCMC
275 sampling.) Therefore, peaks in the selection signal tend to center on clusters of SNPs where
276 there is a local over-abundance of non-synonymous mutations. In contrast, the omega plot of a
277 gene using the Independent model is much more variable (spiky), because the omega values at
278 one codon is completely independent from the other around it, and there is no sharing of
279 information locally. (Omega plots for *gyrA* and *embB* based on the Independent model, which
280 are much less smooth, are shown in the Supplemental Figure S2).

281 Importantly, although the omega value might exceed 1.0 in multiple regions of the ORF, this
282 does not necessarily mean it is significant. To distinguish genuine selection signals from random
283 fluctuations, we can utilize the dispersion in the MCMC samples to assess certainty. The 95%
284 credible interval (95%CI) of the MCMC samples represents the range of possible values of
285 omega for each codon. The upper and lower bounds of the 95% CI are shown as blue dashed
286 lines in the figure. The significant codons are (only) those for which the lower bound of the
287 95%CI exceeds 1.0, or when $\omega>1$ in 97.5% of the MCMC samples.

288 In the case of GyrA, there are two significant peaks. The predominant peak consists of residues
289 89 to 95 (where omega peaks at 6.7). This contains the two codons most frequently associated
290 with fluoroquinolone (FQ) resistance – A90 and D94 (Maruri et al., 2012). In fact, although FQ
291 resistance was only ~10% in this population (10.3% for moxifloxacin), A90V was observed in 40
292 isolates in the Moldova collection, along with D89G/N and D94G/A/N/Y/H. The region also
293 included S95T, which is known to be a lineage-specific marker not associated with drug

294 resistance (Sreevatsan et al., 1997). The reason it was included was likely due to: a) proximity to
295 D94, and b) a single isolate had an additional non-synonymous mutation, S95A. There is also a
296 minor peak at residues 70-74, which is just barely significant (lower bound of 95%CI just barely
297 exceeds 1.0). This is in an adjacent alpha-helix, also in the DNA-binding pocket; A74S been
298 reported in FQ-resistant clinical isolates, but is less frequently observed (Shi et al., 2006). In
299 one study, A74S was only observed in mutants with D94G, so it might be compensatory (Sun et
300 al., 2008). Although, some mutations in *gyrB* have also been reported to be associated with FQ-
301 resistance, especially in the region around Arg482 (which is proximal to where fluoroquinolones
302 bind in the 3D structures of the GyrAB complex (Blower et al., 2016), and some segments of
303 *gyrB* have ω>1 by the Window model, none of these is statistically significant.

304 A similar pattern of selection is observed in *embB* (Figure 2b) in the arabinogalactan (AG)
305 synthesis pathway. The three codons most commonly observed to be mutated in ethambutol-
306 resistant mutants– residues 306, 406, and 497 (Sun et al., 2018) – all exhibit a significant ω
307 value. In addition, residues 305 (adjacent to 306) and 317-338 (in between 306 and 406) are
308 highlighted as significant. In addition, in *ubiA*, residues 171-176 have significant ω values. UbiA
309 is also in the AG pathway, and these residues have been reported in EMB-resistant mutants
310 (Nan et al., 2024).

311 It is notable that *pncA* and *gidB*, two of the genes with the strongest selection signals (mean
312 ω=2.93 and 1.94, respectively), show selection in regions throughout the gene (151 and 78
313 codons respectively, see Supp Fig S3), which is consistent with the diversity of resistance
314 mutations observed, including loss-of-function (Yadon et al., 2017). *pncA* is an activator of
315 pyrazinamide, and *gid* is a ribosomal rRNA methyltransferase that natively facilitates binding of
316 streptomycin (Spies et al., 2011; Wong et al., 2013). Hence any mutations that abrogate
317 function of these two genes could confer resistance (Gomes et al., 2021). This could explain
318 why GenomegaMap detects a significant signal for positive selection that is diffused across both
319 of these targets.

320

321 **RpoB**

322 Rifampicin resistance mutations are commonly observed in of *rpoB* (beta' chain of RNA
323 polymerase), specifically in the Rifampicin Resistance Determining Region (RRDR), residues 435-
324 452, including the most frequent mutation S450L in RIF-R mutants), which lines the physical
325 binding site for rifampicin. As observed in the omega plot, there are two strong peaks indicating
326 regions of the ORF with significant selection (see Figure 3a). The first peak (residues 435-451),
327 with omega values exceeding 4, clearly corresponds to the RRDR. The second peak consists of
328 residues 488-496. These have been reported to be associated with rifampicin-resistant clinical
329 isolates in China (Li et al., 2021). Since they co-occurred with mutations in the RRDR region of
330 *rpoB*, it was speculated that they could be compensatory. We observe this in the Moldova
331 collection too. In(Xu et al., 2017) fact, based on the phylogenetic reconstruction (generated
332 using RAxML), there are two clear cases where mutation events in the 488-496 region occurred
333 in lower branches after mutations in the RRDR. For example, 15 isolates with the *rpoB*:I488V
334 were embedded in a cluster of 111 isolates with *rpoB*:S450L. This supports the hypothesis that
335 mutations in the 488-496 region might help reduce the fitness cost of the earlier mutations in
336 the RRDR. In crystal structures of the RNA polymerase, the residues 488-496 are in a parallel
337 beta strand to 426-452 in the RIF-binding pocket (with Ile491 directly adjacent to Ser450; see
338 Figure 3b), which could provide a structural explanation for how mutations at one site could
339 mitigate the fitness cost of mutations at the other.

340 In addition to RIF-resistance mutations in *rpoB*, compensatory mutations have been observed in
341 *rpoC*, especially around residues 432-527 (Comas et al., 2011). Indeed, *rpoC* is also in the list of
342 53 genes exhibiting positive selection in the Moldova clinical isolates (using the Window model).
343 As shown in Figure 3c, several regions of codons attain significance, including 431-436, 445-447,
344 482-498, and 502-527, which corresponds well with the region described by Comas et al (2011).

345

346 **CinA**

347 CinA has recently been implicated in INH tolerance through TnSeq studies, where it was found
348 to be non-essential under regular growth conditions (7H9 medium), but conditionally more
349 essential (decreased Tn insertion counts) in the presence of isoniazid (INH) (Xu et al., 2017). It

350 was subsequently shown to be involved in NADH recycling, and to mitigate the impact of INH-
351 NADH adducts (Kreutzfeldt et al., 2022). It was also shown to affect MICs for delamanid (DLM)
352 and pretomanid (PTM), which also generate NADH adducts. CinA has two domains: an N-
353 terminal pyrophosphatase and C-terminal deamidase. The activity influencing drug sensitivity
354 was localized to the pyrophosphatase domain. The omega plot (Figure 4a) exhibits a significant
355 peak for residues in the 82-86 range (in the pyrophosphatase domain). Genetic variants
356 observed in this region include: T82I, V83F, and V86A. These residues are in an alpha-helix
357 proximal to Asp80, which projects into the active site and coordinates Mg^{2+} bound to the
358 diphosphate in NADH, in a crystal structure of CinA homolog from *T. thermophilus* (PDB: 4uux)
359 (see Figure 4b). Mutating Asp80 (D80A) abrogates enzyme activity and increases INH sensitivity
360 (Kreutzfeldt et al., 2022). In fact, one of the isolates in the Moldova collection had D80E, which
361 likely also affects CinA function. None of the isolates with these mutations in CinA were
362 genotypically resistant to INH, ETH, DLM, or PTM (based on drug-resistance markers analyzed by
363 TB-profiler). Nonetheless, the proximity of these mutations to the active site suggest that CinA
364 might be under genuine selective pressure in this population, the source of which is still
365 unknown. In contrast, only residue 421 was detected as under positive selection in CinA by the
366 Independent model.

367

368 **PhoR**

369 *phoR* has among the strongest signals for positive selection in the Moldova dataset. *phoR* is
370 part of a two-component sensor, PhoPR (Parish, 2014), which plays a key role in virulence
371 (Martin et al., 2006; Singh et al., 2023). *phoR* is membrane-bound sensor coupled to a histidine
372 kinase that phosphorylates *phoP*, which binds DNA and acts as transcription factor. There is
373 strong evidence for *phoR* being under positive selection in clinical isolates jn previous studies
374 (Chiner-Oms et al., 2019). and mutations in *phoR* have been hypothesized to influence
375 transmissibility and might have been a key step in the evolution of Mtb into a human pathogen
376 (Gonzalo-Asensio et al., 2014).

377 In the Moldova dataset, the codons in *phoR* with genetic variants under positive selection
378 include: 52-84 and 145-159 (in the periplasmic sensor domain) and 188-201 (in the HAMP
379 domain; an alpha-helical domain that facilitates oligomerization by form coiled-coils, often
380 utilized in signal transduction proteins) (see Figure 5). The polymorphism in *phoR* that
381 differentiates Mtb *phoR* from the ortholog in *M. bovis*, which is associated with decreased
382 transmission in humans, is G71I (Gonzalo-Asensio et al., 2014), which falls in the middle of the
383 sensor domain. Gonzalo-Asensio et al. showed that recombineering this mutation into L2 and
384 L4 strains actually down-regulates expression of the PhoPR regulon. Thus, the analysis of
385 positive selection using the Window model in GenomegaMap suggests that other non-
386 synonymous mutations in the sensor domain might cause similar functional changes that are
387 ultimately beneficial and lead to selection in Mtb clinical isolates. Mutations in the HAMP
388 domain will have to be further investigated to determine their impact on PhoPR function and
389 understand how this can lead to adaptation.

390

391 **The Independent Model is biased toward large ORFs.**

392 One notable aspect of the genes exhibiting significant selection under the Independent Model is
393 that it contains many polyketide synthase (PKS) genes. Of the 16 PKS genes in the Mtb H37Rv
394 genome, 9 are significant in the Independent model, along with *nrp* (Rv0101, non-ribosomal
395 peptide synthase, NPRS). These modular (multi-domain) proteins are among the largest genes
396 in the genome, with 10 out of 16 have length over 1000 amino acids (*pks12* is largest, with 4150
397 aa, and NRPS is second largest, with 2511 aa). The PKS genes under positive selection are
398 almost all unique calls for the Independent Model (see Figure 1a); the only PKS gene that is also
399 detected as significant by the Window Model is *pks12*. A comparison of the sizes of the genes
400 under positive selection detected by each model is shown in the Figure 6a. Note that the size
401 distribution of the 44 genes significant by Independent Model is skewed higher than the 53
402 genes significant by the Window, which has a distribution more similar to the overall
403 distribution of ORF sizes in the genome (represented by the histogram). Figure 6b-c shows the
404 number of genes detected by each model in different bin sizes. When normalized by the total

405 number of genes in each bin (Figure 6d-e), the considerable over-representation of genes in the
406 largest bin (1200-4152 aa) by the Independent Model is apparent – over 40% of these largest
407 genes contain a one or more codons under positive selection according to the Independent
408 Model (Figure 6e). By comparison, the Window Model shows much less bias toward larger
409 genes.

410 It is possible that positive selection of PKS genes could reflect biologically-meaningful selection
411 pressures, since PKS genes make glycolipid components of the cell well that often are
412 recognized by and module the immune response in vivo (Constant et al., 2002; Matsunaga et
413 al., 2004; Sirakova et al., 2001). However, the fact that so many PKS genes show signs of
414 positive selection could also reflect a general bias that the Independent Model is more likely to
415 (spuriously) identify individual codons under positive selection in larger genes, which could
416 explain why the largest genes tend to be characterized as under selection.

417 To evaluate this hypothesis, we performed a simulation experiment to estimate the number of
418 genes that would be expected to obtain significance by the Independent or the Window model
419 under a Null model (approximating neutral evolution). Simulated genes were generated that
420 had mutations that occurred at the same rate as the original Moldova dataset, but evolved
421 under the assumption of *neutral evolution*. The was achieved by first computing the least-
422 common ancestral (LCA) sequence for each ORF in the Mtb genome (excluding PPE and
423 PE_PGRS genes, and genes known to be involved in drug resistance). Then, each distinct
424 mutational event (i.e. SNP associated with a particular branch point in the tree) was randomly
425 assigned to a new site within a new gene. This resampling process resulted in 39,564 distinct
426 nucleotide changes permuted to new positions among 3,446,666 sites within coding regions of
427 3877 genes (with the same size distribution as in the Mtb genome), which was then used to
428 generate the 2057 variant sequences for each isolate for each gene. This process decouples the
429 localization of each nucleotide substitution (e.g. clustering within ORFs) and coding impact
430 (whether it is synonymous or non-synonymous).

431 The 3877 simulated gene sequences were then re-analyzed by each of the three GenomegaMap
432 methods. The number of genes detected as significant under this Null model by the

433 Independent Model was 66 out of 3877 (~1.7%), whereas only 14 of the simulated genes were
434 identified as under significant positive selection for the Window Model (~0.4%) (see Table 2).
435 For comparison, we also ran the Constant model on the same set of simulated genes, which
436 detected 16 genes with significant positive selection. Thus, the Independent model would be
437 expected to have a ~4.5x higher false-positive rate (FPR) relative to the Window model.
438 Furthermore, the Independent Modal again shows a bias toward detecting positive selection
439 among larger genes (see Figure 7). This suggests that the Independent Model is more
440 susceptible to false positive calls, due to modeling each codon with its own individual ω
441 parameter, whereas the Window Model is more robust because it combines information (on
442 genetic variants) locally over multiple codons (clusters of SNPs) (through use a constant ω
443 parameter for each segment of the ORF, with boundaries that vary over the MCMC simulation).

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

GenomegaMap model	**number of significant genes under positive selection (estimated FPR)**
Constant	16/3877 (0.41%)
Window	14/3877 (0.36%)
Independent	66/3877 (1.7%)

446

447

448 **Validation through analysis of an independent collection of Mtb isolates from other countries.**

449 To evaluate whether the genes identified as under positive selection in the Moldova dataset are
450 generally representative, or might be specific to that geographic region, we carried out a similar
451 analysis in a larger collection of 5,195 isolates from 15 other countries. These isolates were also
452 predominantly from lineages L2 and L4, but L1 and L3 were also represented. Approximately
453 one-third of the isolates were drug-resistant, and most of these (33.4%) being MDR (markers for
454 both INH-R and RIF-R, based on TB-profiler). Resistance to EMB and STR was high, as well;

455 fluoroquinolone resistance was lower, at around 10%. (Histograms of the geographic
456 distribution, lineages, and genotypic drug resistance are shown in Supplemental Figures S4 and
457 S5). We applied the same pipeline for genome assembly, quality filtering, and calling of genetic
458 variants as above.

459 Using the Window model in GenomegaMap, 178 genes were found to be under positive
460 selection in this global dataset, using the same criterion for significance (with at least 1 region
461 with ω>1 for at least 97.5% of the MCMC samples) (see Supplemental Table T2). 19 of the 53
462 significant genes in the Moldova dataset were also observed in this global collection,
463 representing a 36% overlap. Predictably, this includes many drug-resistance-related genes, like
464 *pncA, gid, rpoB, rpoC, gyrA, embB* and *folC*. An additional gene implicated in drug-resistance
465 that is uniquely significant in the global collection is *aac* (Rv0262c, aminoglycoside
466 acetyltransferase). Codons under positive selection an *aac* span 86-94. These residues form a
467 loop that binds CoA (rather than the aminoglycoside), based on crystal structure of the complex
468 of *aac* with kanamycin (PDB: 1M4I; see Supplemental Figure S6). Other notable genes found to
469 be under selection in both datasets include *pks12, phoR,* and *resR*. The global positive-selection
470 geneset also includes some additional genes that have been previously reported in other
471 studies of positive selection in Mtb. For example, *whiB6* is a transcriptional regulator that was
472 found to be associated with multi-drug-resistance in a GWAS study of clinical isolates from Peru
473 (Farhat et al., 2019). *whiB6* has also been implicated in expression of ESX-1 (see below)
474 (Bosserman et al., 2017).

475

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

477 Several components of the ESX-1 complex exhibit positive selection in the larger global
478 collection of Mtb clinical isolates, including *eccCa1, eccD1*, and *eccE1*, along with *espA*, one of
479 several proteins secreted by ESX-1. In addition, *eccB1, esxA* (ESAT-6)*, and espFIJ* all contained
480 segments showing evidence of positive selection by the weaker definition of significance (>90%
481 of MCMC samples had ω>1). Only *espB* and *espK* showed weak significance in the Moldova
482 dataset, possibly because of the smaller number of isolates. ESX-1 is one of five type-VII

483 secretion systems (T7Ss) in Mtb (Bitter et al., 2009). It plays an important role in interacting
484 with the host immune system, including phagosome-maturation arrest, host cell membrane
485 lysis, inhibition of autophagy (Romagnoli et al., 2012), and up-regulation of host-cell genes and
486 cytokine signaling (Champion et al., 2009; Chirakos et al., 2020; Conrad et al., 2017), and is
487 required for virulence in animal models of infection (Tiwari et al., 2019). The ESX-1 locus
488 (Rv3864-Rv3883c) contains members of the transmembrane pore complex (*eccA1-E1*), the
489 *mycP1* protease, and several secreted effectors (*espBEFGHIJKL, esxAB, pe35, ppe68*).
490 Interestingly *espACD* are in a separate locus (Rv3514c-Rv3616c). Specifically, *espA* binds to the
491 ESX-1 pore complex and facilitates secretion of EsxA and EsxB (Chen et al., 2013). The region of
492 *espA* under positive selection includes codons 67-72 (Figure 8a). Non-synonymous genetic
493 variants in this region include: K66R/Q, N67T, R68A, N69S, H70Y, V71L, N72K. These residues lie
494 in an alpha-helix of a predicted 4-helix bundle, proximal to the WXG motif (res. 55-57) (see
495 Figure 8b), which forms a loop that plays a critical role in recognition by the ESX-1 pore complex
496 (Chen et al., 2013). As noted above, *whiB6*, and *phoR,* which both regulate expression of ESX-1
497 (Bosserman et al., 2017; Walters et al., 2006), also exhibit positive selection in the global
498 dataset.

499

500 **DISCUSSION**

501 Mtb is considered well-adapted to its lifecycle as an obligate human pathogen, and the genome
502 is fairly homogeneous (low diversity) (Coscolla & Gagneux, 2014). Strong evidence for purifying
503 selection has been found throughout the genome (Pepperell et al., 2013). Nevertheless,
504 sporadic SNPs and other polymorphisms are often observed in genomes of individual isolates
505 (presumably arising within patients during the course of infection (Trauner et al., 2017)), which
506 are infrequent (unfixed alleles, appearing in only one or a few isolates), reflecting on-going
507 evolution (Goig et al., 2025). Some of these mutations might reflect adaptations to pressures
508 related to modern clinical treatment of TB, including chemotherapy, but also effects of co-
509 morbidities (e.g. diabetes, COPD, malnutrition), co-infection (e.g. HIV), which could involve
510 exposure to non-TB antibiotics as well as impacting patient immune status, oxidative

511 environment, etc. Statistical methods for evaluating positive selection are a powerful tool for
512 identifying genes adaptive to these pressures in large genomic databases.

513 There have been multiple studies investigating positive selection among genomes of Mtb clinical
514 isolates, based on principles ranging from phenotype-genotype associations (Farhat et al.,
515 2019), to SNP clustering (e.g. nucleotide diversity, mutation burden) (Liu et al., 2022; Mortimer
516 et al., 2018), to homoplasy (or convergent evolution) (Holt et al., 2018; Mortimer et al., 2018),
517 to traditional dN/dS analysis (Deng et al., 2024; Phelan et al., 2019), to phylodynamics (Chiner-
518 Oms et al., 2022). In addition to genes already known to be involved in drug-resistance, they
519 have identified other genes which may play an adaptive role to modern clinical stresses, such as
520 *resR* (Liu et al., 2022) and *phoR* (Chiner-Oms et al., 2019). Most of these studies evaluated
521 selection at the gene-level. However, selection can be analyzed at other scales, including
522 codon-level and clusters of SNPs (intermediate level). The purpose of this study was to evaluate
523 and compare the results of all three levels of analysis.

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

531 We found that the intermediate (or "Window" model) provides the most compelling list of
532 candidate genes. The Window model segments the ORF into multiple contiguous regions, and
533 allows each region to have a different ω parameter, but assumes the codons within a local
534 region all share the same omega value. Many of the genes found to be under positive selection
535 by the Window model are known to be involved in drug resistance, (*rpoB, gyrA, embB, pncA,*
536 *gidB, ubiA, folC*...) which would be expected to be under positive selection in a collection of
537 Mtb clinical isolates from patients receiving chemotherapy, especially in a geographic region
538 with high incidence of drug-resistance (such as Moldova (Yang et al., 2022)). In addition, on the

539 list of candidates from the Window model, we observed other genes previously reported as
540 targets of positive selection in other studies, such as *resR* (Liu et al., 2022) and *phoR* (Chiner-
541 Oms et al., 2019). *prpD* is a methylcitrate dehydratase (propionate metabolism pathway) that is
542 regulated by *prpR*, which was identified as associated with multi-drug tolerance (Hicks et al.,
543 2018). Rv0988 and Rv2082 have both been reported in another study of positive selection in
544 clinical isolates resistant to bedaquiline and clofazimine (Ismail et al., 2023). While the function
545 of Rv2082 is unknown, Rv0988 is an ABC-transporter implicated in preventing acidification and
546 phagosome maturation (Pethe et al., 2004). This consistency for "expected" genes under
547 positive selection with previous studies increases confidence that other candidates with less-
548 well understood role might also be genuinely under positive selection.

549 Another advantage of the Window model of selection analysis is that it can identify specific
550 regions within an ORF that are under positive selection. As implemented in GenomegaMap,
551 this model assumes ω is constant locally but varies among regions. However, the borders
552 between regions also get adjusted dynamically during the MCMC sampling, producing a
553 smoothly varying omega signal across the ORF. We define significant codons to be those where
554 >97.% of the sampled ω values from the MCMC sample (the lower-bound of the 95% credible
555 interval) exceed 1.0, which is more stringent that the criterion used in the original paper (Wilson
556 & Consortium, 2020). In cases where protein crystal structures are available, the regions under
557 positive selection often correspond well to parts of the protein involved in interacting with
558 inhibitors, such as in *rpoB* (residues in the RDRR that bind rifampicin) and *gyrA* (residues binding
559 fluoroquinolones). Another compelling example is the proximity of the residues under selection
560 in *cinA* to the active site (of the pyrophosphatase domain). The insights about which parts of a
561 protein are under selection are helpful in understanding the mechanisms driving it, beyond just
562 labeling a gene as being under positive selection. The correspondence of codons under
563 selection to structural information (e.g. for known drug-binding sites) lends additional
564 confidence in the list of candidates identified by the Window model.

565 Individual codons under selection can also be highlighted by the Independent model in
566 GenomegaMap (where each codon has its own ω parameter). However, we observed that many
567 of the genes on the list of candidates from the Independent model were among the largest

568 genes in the genome, and this size bias suggests the Independent model might be more
569 susceptible to making type I errors. In fact, we used a simulated set of genes with sequences
570 evolved under the assumption of neutral evolution (with the same distribution of ORF sizes and
571 nucleotide content as the clinical isolates from Moldova, but with observed phylogenetic
572 changes permuted randomly to new sites) to show that the Independent model has a roughly 5
573 times higher estimated false positive rate than the Window model. This is likely due to how the
574 Independent model treats each codon independently, making it more susceptible to errors from
575 random coincidences of SNPs, whereas the Window model is more robust because of how it
576 integrates information over multiple codons in a region (by assuming the codons in a region
577 share the same ω), and hence takes advantage of clusters of SNPs where there is a local excess
578 of non-synonymous over synonymous SNPs.

579 On the list of 53 genes identified to be under positive selection in the Moldova dataset, we
580 observed several regulatory proteins, including *pknD* and *sigL* (ser-thr protein kinases), *sigM*
581 (sigma factor), Rv1320 (adenylate cyclase), and *devS* (two-component sensor). As exemplified
582 by *devS*, these regulators could be involved in adaptations to various stresses. For example,
583 *devS* plays a critical role in regulating the response to hypoxia (low oxygen tension, e.g. inside
584 granulomas), a hallmark of the host immune response to TB infection (Kumar et al., 2007; Park
585 et al., 2003; Rustad et al., 2008; Sherman et al., 2001), and the intracellular oxidative
586 environment could also be influenced by antibiotics (such as isoniazid, which generates
587 oxidative radicals in the process of activation (Sherman et al., 1996)). Several metabolic genes
588 also show signals of positive selection in this population (*cysN* - sulfate
589 adenyltransferase/kinase; *glyA1* - serine hydroxymethyltransferase; *hemN* - coproporphyrinogen
590 III oxidase; *pyrG* - CTP synthase, *atsD* - arylsulfatase; *fadE17* - acyl-CoA dehydrogenase; *cdsA* -
591 CDP-diglyceride pyrophosphorylase). Metabolic shifts often accompany changes in response to
592 stress (Rizvi et al., 2019), though the specific role of these pathways remains to be elucidated.

593 Genes classically associated with INH resistance were not detected as significant by the Window
594 model (*katG* and *inhA*). *katG* was probably not detected because mutations of only a single
595 codon, S315; and *inhA* was probably not detected because mutations occur predominantly in
596 the promoter region upstream of *fabG1*, more frequently than in the coding region, affecting

597 expression levels of *inhA*). However, several genes that are potentially indirectly related to INH
598 resistance were observed as under positive selection, including *ndhA*. *ndhA* (Rv0392c) is one of
599 two type II NADH dehydrogenases, along with *ndh* (Rv1854c) which can compensate for INH-
600 NADH adduct formation by shifting the intracellular NADH/NAD+ ratio (Vilcheze et al., 2005).
601 Although, historically, *ndh* is more often observed to acquire resistance mutations (Hazbon et
602 al., 2006), mutations in *ndhA* have also been reported to be associated with INH resistance
603 (Furio et al., 2021).

604 In order to determine whether the genes under selection might be specific to Moldova, we
605 applied the same analysis of positive selection to a larger collection of Mtb clinical isolates from
606 15 other countries. Many of the drug-resistance genes, along with genes such as *resR* and *phoR*
607 were also detected as significant in the global dataset. There were approximately 3 times as
608 many genes (178) with statistically significant positive selection as in the Moldova dataset (53).
609 This is possibly due in part to the increased incidence of unfixed mutations (often unique SNPs
610 observed only in single or a few isolates), the probability of which increases with size of the
611 collection (number of genomes, barring duplications due to outbreak clusters) (or possibly due
612 to the increased diversity among isolates selected from 15 different countries). However, in
613 theory, dN/dS analysis should factor-out differences in overall mutation rate, and instead
614 identify genes where there is a relative excess in the rate of non-synonymous over synonymous
615 SNPs, which is not necessarily a consequence of aggregating SNPs from a larger collection of
616 genomes.

617 *phoR* has among the strongest signals for positive selection in both datasets (Moldova, and
618 global collection). PhoPR function is required for virulence, as disruption mutants are highly
619 attenuated in mice (Martin et al., 2006; Singh et al., 2023). A mutation in *phoP* is one of the few
620 genetic differences between H37Rv and H37Ra, to which the difference in virulence of these
621 strains has been partially attributed (Lee et al., 2008). As a part of a two-component sensor,
622 *phoR* regulates many lipid synthesis and cell-wall biogenesis genes (Walters et al., 2006). PhoPR
623 has been implicated in responses to several types of environmental stress, including acidic pH
624 (Singh et al., 2023). We observed multiple codons in the periplasmic sensor domain to be under
625 positive selection (using the Window model). Using phylodynamics, Chiner-Oms et al. (2019)

626 found that *phoR* exhibited an increase in selection of non-synonymous variants, compared to
627 other animal-adapted members of the MTBC (Mycobacterium Tuberculosis Complex), that
628 coincided roughly with expansion of the pathogen into the human population. This and other
629 evidence suggests mutations in *phoR* might facilitate transmission among humans (Gonzalo-
630 Asensio et al., 2014).

631

632 Among the genes under positive selection in the larger global collection, we observed several
633 genes from the ESX-1 cluster. In particular, *eccCa1, eccD1, eccE1* were under positive selection,
634 along with *espA*. Based on analogy to a cryo-EM structure of the ESX-5 complex (Beckham et
635 al., 2021), these three subunits interact to form the cytoplasmic face of the transmembrane
636 pore. *espA* is encoded in a different operon (*espACD*), but is thought to bind to the pore
637 complex and facilitate secretion of substrates (Champion et al., 2009). ESX-1 plays a crucial role
638 in virulence, secreting a number of effectors (including *esxA* and *esxB*, along with other *esx* and
639 PPE proteins), and can disrupt host-cell membranes and modulate cytokine expression (Conrad
640 et al., 2017; Zheng et al., 2025). Thus, we speculate that the genetic variants observed could
641 play a role in fine-tuning the host immune response. ESX components have been identified in
642 other studies of positive selection. For example, *esxW* (secreted via ESX-5) was identified to
643 harbor convergent mutations in different Mtb lineages in a collection of isolates from Viet Nam,
644 which was used to help explain the increased transmission of the Beijing strain in the
645 geographic region (Holt et al., 2018). *esxW* was not under significant positive selection in our
646 global collection, and the only *esx* protein that was under significant positive selection was
647 *esxH*, a member of ESX-3. ESX-1 proteins *eccD1* and *eccE1* were also observed to be under
648 positive selection in the analysis of over 51,000 clinical isolates by Liu et al. (2022).

649 Increased selection of genetic variants in the ESX-1 system could be the result of adaptation to
650 modern TB treatment paradigms (e.g. shortening of disease course through chemotherapy,
651 better clinical control of inflammation) that might benefit from adjustments in host-pathogen
652 interactions (e.g. enhancement or suppression of immune response). It is interesting to note
653 that PhoPR has been shown to regulate expression of ESX-1 and *esxA* (ESAT-6) secretion (Frigui

654 et al., 2008). Similarly, *whiB6*, which is also on the list of genes under significant positive
655 selection in the global collection, also regulates expression of ESX-1 (demonstrated in *M.*
656 *marinum*; (Bosserman et al., 2017)). The selection signals for these two regulators lend
657 additional support to the potential adaptive role of genetic variants affecting ESX-1 function in
658 this global sample of clinical isolates.

659 One of the limitations of using positive selection to identify genes potentially involved in
660 adaptation is that it is restricted to SNPs in coding regions. Liu et al. (2022) also looked for
661 clusters of SNPs (mutational burden) in intergenic regions, especially those associated with
662 regulatory regions (e.g. promotors). In general, dN/dS analysis focuses on SNPs, and thus does
663 not directly take indels into account. However, indels can play a critical role in
664 activating/deactivating proteins; for example, there is a 7-bp insertion that breaks *pks15/*1 into
665 two ORFs and eliminates the synthesis of PGL (phenolglycolipid) found in more virulent strains
666 of Mtb (Constant et al., 2002). Also, because GenomegaMap does not explicitly use a
667 phylogenetic tree (which is key to its efficiency/scalability over methods like PAML), it is
668 insensitive to homoplasy. However, this can be easily circumvented by using other tools, such
669 as SNPPar (Edwards et al., 2021), which builds a rapid, approximate phylogenetic tree and can
670 then be used to identify homoplasic sites (such as S315 in KatG, the primary resistance mutation
671 for INH).

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

683

684 **Figure Legends**

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

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

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

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

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

708 **Figure 6.** The Independent model is biased toward larger genes. a) Beeswarm plots showing
709 that the significant genes under positive selection by the Independent model are skewed larger

710 than those from the Window model, whereas the Window model better matches the overall
711 ORF size distribution (histogram to right). Polyketide synthase genes are colored red. b) and c)
712 Number of genes under positive selection, divided into size bins (amino acids). The
713 Independent model shows an increasing trend toward the rightmost bin (ORFs with 1200-5000
714 amino acids). d) and e) Proportion of genes under positive selection (counts of significant genes
715 normalized by the overall number of genes in each size bin).

716 **Figure 7.** Size bias in the Simulated gene set (neutral evolution). Linear regressions of the
717 number of significant codons found against ORF among genes with weakly significant positive
718 selection (ω>1 for >90% of the MCMC samples) using a) the Window model, b) and the
719 Independent model. The Independent model exhibits a strong correlation with ORF size ($r^2 =$
720 0.558) compared to the Window model ($r^2 = 0.134$).

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

725

726 **Supplemental Figures**

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

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

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

732 **Figure S4.** Countries represented in global collection of 5,195 Mtb clinical isolates.

733 **Figure S5.** Lineages and drug-resistance genotypes represented in global collection of 5,195 Mtb
734 clinical isolates.

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

739

740 **Supplemental Data**

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

751 **Supplemental Table T2.** Selection scores and statistical significance for all ORFs in the Mtb
752 H37Rv genome for the global Mtb genome collection (5,195 clinical isolates from 15 countries).
753 Results for the Window model in GenomegaMap are provided: The number of significant
754 codons ("strongly selected codons") are given, using a 95% credible interval as a significance
755 criterion (ω>1 for >97.5% of the MCMC samples). (The number of weakly selected codons using
756 80%CI is given too.) The genes are sorted by the "max_omega_LB" which is the peak height of
757 the lower blue curve in the omega plots (lower bound of the 95%CI). The 178 significant genes
758 by the Window model (excluding PPE/PGRS genes) are marked in yellow.

759

760 **Acknowledgements**

761 Funding for this work was provided in part by NIH grant AI143575 (TRI). Supercomputing time
762 was provided by the High Performance Research Computing (HPRC) at Texas A&M University.

763

764 **REFERENCES**

765 Beckham, K. S. H., Ritter, C., Chojnowski, G., Ziemianowicz, D. S., Mullapudi, E., Rettel, M.,
766 Savitski, M. M., Mortensen, S. A., Kosinski, J., & Wilmanns, M. (2021). Structure of the
767 mycobacterial ESX-5 type VII secretion system pore complex. *Sci Adv*, *7*(26).
768 https://doi.org/10.1126/sciadv.abg9923
769 Bitter, W., Houben, E. N., Bottai, D., Brodin, P., Brown, E. J., Cox, J. S., Derbyshire, K., Fortune, S.
770 M., Gao, L. Y., Liu, J., Gey van Pittius, N. C., Pym, A. S., Rubin, E. J., Sherman, D. R., Cole, S.
771 T., & Brosch, R. (2009). Systematic genetic nomenclature for type VII secretion systems.
772 *PLoS Pathog*, *5*(10), e1000507. https://doi.org/10.1371/journal.ppat.1000507
773 Blower, T. R., Williamson, B. H., Kerns, R. J., & Berger, J. M. (2016). Crystal structure and stability
774 of gyrase-fluoroquinolone cleaved complexes from Mycobacterium tuberculosis. *Proc*
775 *Natl Acad Sci U S A*, *113*(7), 1706-1713. https://doi.org/10.1073/pnas.1525047113
776 Bosserman, R. E., Nguyen, T. T., Sanchez, K. G., Chirakos, A. E., Ferrell, M. J., Thompson, C. R.,
777 Champion, M. M., Abramovitch, R. B., & Champion, P. A. (2017). WhiB6 regulation of
778 ESX-1 gene expression is controlled by a negative feedback loop in Mycobacterium
779 marinum. *Proc Natl Acad Sci U S A*, *114*(50), E10772-E10781.
780 https://doi.org/10.1073/pnas.1710167114
781 Broset, E., Martin, C., & Gonzalo-Asensio, J. (2015). Evolutionary landscape of the
782 Mycobacterium tuberculosis complex from the viewpoint of PhoPR: implications for
783 virulence regulation and application to vaccine development. *mBio*, *6*(5), e01289-01215.
784 https://doi.org/10.1128/mBio.01289-15
785 Champion, P. A., Champion, M. M., Manzanillo, P., & Cox, J. S. (2009). ESX-1 secreted virulence
786 factors are recognized by multiple cytosolic AAA ATPases in pathogenic mycobacteria.
787 *Mol Microbiol*, *73*(5), 950-962. https://doi.org/10.1111/j.1365-2958.2009.06821.x
788 Chen, J. M., Zhang, M., Rybniker, J., Basterra, L., Dhar, N., Tischler, A. D., Pojer, F., & Cole, S. T.
789 (2013). Phenotypic profiling of Mycobacterium tuberculosis EspA point mutants reveals
790 that blockage of ESAT-6 and CFP-10 secretion in vitro does not always correlate with
791 attenuation of virulence. *J Bacteriol*, *195*(24), 5421-5430.
792 https://doi.org/10.1128/JB.00967-13
793 Chiner-Oms, A., Lopez, M. G., Moreno-Molina, M., Furio, V., & Comas, I. (2022). Gene
794 evolutionary trajectories in Mycobacterium tuberculosis reveal temporal signs of
795 selection. *Proc Natl Acad Sci U S A*, *119*(17), e2113600119.
796 https://doi.org/10.1073/pnas.2113600119
797 Chiner-Oms, A., Sanchez-Buso, L., Corander, J., Gagneux, S., Harris, S. R., Young, D., Gonzalez-
798 Candelas, F., & Comas, I. (2019). Genomic determinants of speciation and spread of the
799 Mycobacterium tuberculosis complex. *Sci Adv*, *5*(6), eaaw3307.
800 https://doi.org/10.1126/sciadv.aaw3307

801 Chirakos, A. E., Nicholson, K. R., Huffman, A., & Champion, P. A. (2020). Conserved ESX-1
802 Substrates EspE and EspF Are Virulence Factors That Regulate Gene Expression. *Infect*
803 *Immun*, *88*(12). https://doi.org/10.1128/IAI.00289-20
804 Comas, I., Borrell, S., Roetzer, A., Rose, G., Malla, B., Kato-Maeda, M., Galagan, J., Niemann, S.,
805 & Gagneux, S. (2011). Whole-genome sequencing of rifampicin-resistant Mycobacterium
806 tuberculosis strains identifies compensatory mutations in RNA polymerase genes. *Nat*
807 *Genet*, *44*(1), 106-110. https://doi.org/10.1038/ng.1038
808 Conrad, W. H., Osman, M. M., Shanahan, J. K., Chu, F., Takaki, K. K., Cameron, J., Hopkinson-
809 Woolley, D., Brosch, R., & Ramakrishnan, L. (2017). Mycobacterial ESX-1 secretion system
810 mediates host cell lysis through bacterium contact-dependent gross membrane
811 disruptions. *Proc Natl Acad Sci U S A*, *114*(6), 1371-1376.
812 https://doi.org/10.1073/pnas.1620133114
813 Constant, P., Perez, E., Malaga, W., Laneelle, M. A., Saurel, O., Daffe, M., & Guilhot, C. (2002).
814 Role of the pks15/1 gene in the biosynthesis of phenolglycolipids in the Mycobacterium
815 tuberculosis complex. Evidence that all strains synthesize glycosylated p-hydroxybenzoic
816 methyl esters and that strains devoid of phenolglycolipids harbor a frameshift mutation
817 in the pks15/1 gene. *J Biol Chem*, *277*(41), 38148-38158.
818 https://doi.org/10.1074/jbc.M206538200
819 Coscolla, M., & Gagneux, S. (2014). Consequences of genomic diversity in Mycobacterium
820 tuberculosis. *Semin Immunol*, *26*(6), 431-444.
821 https://doi.org/10.1016/j.smim.2014.09.012
822 Deng, L., Wang, Q., Liu, H., Jiang, Y., Xu, M., Xiang, Y., Yang, T., Yang, S., Yan, D., Li, M., Zhao, L.,
823 Zhao, X., Wan, K., He, G., Mijiti, X., & Li, G. (2024). Identification of positively selected
824 genes in Mycobacterium tuberculosis from southern Xinjiang Uygur autonomous region
825 of China. *Front Microbiol*, *15*, 1290227. https://doi.org/10.3389/fmicb.2024.1290227
826 Edwards, D. J., Duchene, S., Pope, B., & Holt, K. E. (2021). SNPPar: identifying convergent
827 evolution and other homoplasies from microbial whole-genome alignments. *Microb*
828 *Genom*, *7*(12). https://doi.org/10.1099/mgen.0.000694
829 Farhat, M. R., Freschi, L., Calderon, R., Ioerger, T., Snyder, M., Meehan, C. J., de Jong, B., Rigouts,
830 L., Sloutsky, A., Kaur, D., Sunyaev, S., van Soolingen, D., Shendure, J., Sacchettini, J., &
831 Murray, M. (2019). GWAS for quantitative resistance phenotypes in Mycobacterium
832 tuberculosis reveals resistance genes and regulatory regions. *Nat Commun*, *10*(1), 2128.
833 https://doi.org/10.1038/s41467-019-10110-6
834 Frigui, W., Bottai, D., Majlessi, L., Monot, M., Josselin, E., Brodin, P., Garnier, T., Gicquel, B.,
835 Martin, C., Leclerc, C., Cole, S. T., & Brosch, R. (2008). Control of M. tuberculosis ESAT-6
836 secretion and specific T cell recognition by PhoP. *PLoS Pathog*, *4*(2), e33.
837 https://doi.org/10.1371/journal.ppat.0040033
838 Furio, V., Moreno-Molina, M., Chiner-Oms, A., Villamayor, L. M., Torres-Puente, M., & Comas, I.
839 (2021). An evolutionary functional genomics approach identifies novel candidate regions
840 involved in isoniazid resistance in Mycobacterium tuberculosis. *Commun Biol*, *4*(1), 1322.
841 https://doi.org/10.1038/s42003-021-02846-z
842 Goig, G. A., Windels, E. M., Loiseau, C., Stritt, C., Biru, L., Borrell, S., Brites, D., & Gagneux, S.
843 (2025). Ecology, global diversity and evolutionary mechanisms in the Mycobacterium
844 tuberculosis complex. *Nat Rev Microbiol*. https://doi.org/10.1038/s41579-025-01159-w

845 Gomes, L. C., Campino, S., Marinho, C. R. F., Clark, T. G., & Phelan, J. E. (2021). Whole genome
846 sequencing reveals large deletions and other loss of function mutations in
847 Mycobacterium tuberculosis drug resistance genes. *Microb Genom*, *7*(12).
848 https://doi.org/10.1099/mgen.0.000724
849 Gonzalo-Asensio, J., Malaga, W., Pawlik, A., Astarie-Dequeker, C., Passemar, C., Moreau, F., Laval,
850 F., Daffe, M., Martin, C., Brosch, R., & Guilhot, C. (2014). Evolutionary history of
851 tuberculosis shaped by conserved mutations in the PhoPR virulence regulator. *Proc Natl*
852 *Acad Sci U S A*, *111*(31), 11491-11496. https://doi.org/10.1073/pnas.1406693111
853 Hazbon, M. H., Brimacombe, M., Bobadilla del Valle, M., Cavatore, M., Guerrero, M. I., Varma-
854 Basil, M., Billman-Jacobe, H., Lavender, C., Fyfe, J., Garcia-Garcia, L., Leon, C. I., Bose, M.,
855 Chaves, F., Murray, M., Eisenach, K. D., Sifuentes-Osornio, J., Cave, M. D., Ponce de Leon,
856 A., & Alland, D. (2006). Population genetics study of isoniazid resistance mutations and
857 evolution of multidrug-resistant Mycobacterium tuberculosis. *Antimicrob Agents*
858 *Chemother*, *50*(8), 2640-2649. https://doi.org/10.1128/AAC.00112-06
859 Hicks, N. D., Yang, J., Zhang, X., Zhao, B., Grad, Y. H., Liu, L., Ou, X., Chang, Z., Xia, H., Zhou, Y.,
860 Wang, S., Dong, J., Sun, L., Zhu, Y., Zhao, Y., Jin, Q., & Fortune, S. M. (2018). Clinically
861 prevalent mutations in Mycobacterium tuberculosis alter propionate metabolism and
862 mediate multidrug tolerance. *Nat Microbiol*, *3*(9), 1032-1042.
863 https://doi.org/10.1038/s41564-018-0218-3
864 Holt, K. E., McAdam, P., Thai, P. V. K., Thuong, N. T. T., Ha, D. T. M., Lan, N. N., Lan, N. H., Nhu, N.
865 T. Q., Hai, H. T., Ha, V. T. N., Thwaites, G., Edwards, D. J., Nath, A. P., Pham, K., Ascher, D.
866 B., Farrar, J., Khor, C. C., Teo, Y. Y., Inouye, M., . . . Dunstan, S. J. (2018). Frequent
867 transmission of the Mycobacterium tuberculosis Beijing lineage and positive selection
868 for the EsxW Beijing variant in Vietnam. *Nat Genet*, *50*(6), 849-856.
869 https://doi.org/10.1038/s41588-018-0117-9
870 Ismail, N., Dippenaar, A., Warren, R. M., Peters, R. P. H., & Omar, S. V. (2023). Emergence of
871 Canonical and Noncanonical Genomic Variants following In Vitro Exposure of Clinical
872 Mycobacterium tuberculosis Strains to Bedaquiline or Clofazimine. *Antimicrob Agents*
873 *Chemother*, *67*(4), e0136822. https://doi.org/10.1128/aac.01368-22
874 Kreutzfeldt, K. M., Jansen, R. S., Hartman, T. E., Gouzy, A., Wang, R., Krieger, I. V., Zimmerman,
875 M. D., Gengenbacher, M., Sarathy, J. P., Xie, M., Dartois, V., Sacchettini, J. C., Rhee, K. Y.,
876 Schnappinger, D., & Ehrt, S. (2022). CinA mediates multidrug tolerance in
877 Mycobacterium tuberculosis. *Nat Commun*, *13*(1), 2203.
878 https://doi.org/10.1038/s41467-022-29832-1
879 Kumar, A., Toledo, J. C., Patel, R. P., Lancaster, J. R., Jr., & Steyn, A. J. (2007). Mycobacterium
880 tuberculosis DosS is a redox sensor and DosT is a hypoxia sensor. *Proc Natl Acad Sci U S*
881 *A*, *104*(28), 11568-11573. https://doi.org/10.1073/pnas.0705054104
882 Lee, J. S., Krause, R., Schreiber, J., Mollenkopf, H. J., Kowall, J., Stein, R., Jeon, B. Y., Kwak, J. Y.,
883 Song, M. K., Patron, J. P., Jorg, S., Roh, K., Cho, S. N., & Kaufmann, S. H. (2008). Mutation
884 in the transcriptional regulator PhoP contributes to avirulence of Mycobacterium
885 tuberculosis H37Ra strain. *Cell Host Microbe*, *3*(2), 97-103.
886 https://doi.org/10.1016/j.chom.2008.01.002
887 Li, M. C., Lu, J., Lu, Y., Xiao, T. Y., Liu, H. C., Lin, S. Q., Xu, D., Li, G. L., Zhao, X. Q., Liu, Z. G., Zhao,
888 L. L., & Wan, K. L. (2021). rpoB Mutations and Effects on Rifampin Resistance in

889 Mycobacterium tuberculosis. *Infect Drug Resist*, *14*, 4119-4128.
890 https://doi.org/10.2147/IDR.S333433
891 Liu, Q., Zhu, J., Dulberger, C. L., Stanley, S., Wilson, S., Chung, E. S., Wang, X., Culviner, P., Liu, Y.
892 J., Hicks, N. D., Babunovic, G. H., Giffen, S. R., Aldridge, B. B., Garner, E. C., Rubin, E. J.,
893 Chao, M. C., & Fortune, S. M. (2022). Tuberculosis treatment failure associated with
894 evolution of antibiotic resilience. *Science*, *378*(6624), 1111-1118.
895 https://doi.org/10.1126/science.abq2787
896 Marin, M., Vargas, R., Harris, M., Jeffrey, B., Epperson, L. E., Durbin, D., Strong, M., Salfinger, M.,
897 Iqbal, Z., Akhundova, I., Vashakidze, S., Crudu, V., Rosenthal, A., & Farhat, M. R. (2022).
898 Benchmarking the empirical accuracy of short-read sequencing across the M.
899 tuberculosis genome. *Bioinformatics*, *38*(7), 1781-1787.
900 https://doi.org/10.1093/bioinformatics/btac023
901 Martin, C., Williams, A., Hernandez-Pando, R., Cardona, P. J., Gormley, E., Bordat, Y., Soto, C. Y.,
902 Clark, S. O., Hatch, G. J., Aguilar, D., Ausina, V., & Gicquel, B. (2006). The live
903 Mycobacterium tuberculosis phoP mutant strain is more attenuated than BCG and
904 confers protective immunity against tuberculosis in mice and guinea pigs. *Vaccine*,
905 *24*(17), 3408-3419. https://doi.org/10.1016/j.vaccine.2006.03.017
906 Maruri, F., Sterling, T. R., Kaiga, A. W., Blackman, A., van der Heijden, Y. F., Mayer, C., Cambau, E.,
907 & Aubry, A. (2012). A systematic review of gyrase mutations associated with
908 fluoroquinolone-resistant Mycobacterium tuberculosis and a proposed gyrase
909 numbering system. *J Antimicrob Chemother*, *67*(4), 819-831.
910 https://doi.org/10.1093/jac/dkr566
911 Matsunaga, I., Bhatt, A., Young, D. C., Cheng, T. Y., Eyles, S. J., Besra, G. S., Briken, V., Porcelli, S.
912 A., Costello, C. E., Jacobs, W. R., Jr., & Moody, D. B. (2004). Mycobacterium tuberculosis
913 pks12 produces a novel polyketide presented by CD1c to T cells. *J Exp Med*, *200*(12),
914 1559-1569. https://doi.org/10.1084/jem.20041429
915 Mortimer, T. D., Weber, A. M., & Pepperell, C. S. (2018). Signatures of Selection at Drug
916 Resistance Loci in Mycobacterium tuberculosis. *mSystems*, *3*(1).
917 https://doi.org/10.1128/mSystems.00108-17
918 Nan, X. T., Li, M. C., Xiao, T. Y., Liu, H. C., Lin, S. Q., Wang, W., Qian, C., Hang, H., Li, G. L., Zhao, X.
919 Q., Wan, K. L., & Zhao, L. L. (2024). Different Contributions of embB and ubiA Mutations
920 to Variable Level of Ethambutol Resistance in Mycobacterium tuberculosis Isolates.
921 *Infect Drug Resist*, *17*, 3125-3132. https://doi.org/10.2147/IDR.S466371
922 Nei, M. (2005). Selectionism and neutralism in molecular evolution. *Mol Biol Evol*, *22*(12), 2318-
923 2342. https://doi.org/10.1093/molbev/msi242
924 Nielsen, R., & Yang, Z. (1998). Likelihood models for detecting positively selected amino acid
925 sites and applications to the HIV-1 envelope gene. *Genetics*, *148*(3), 929-936.
926 https://doi.org/10.1093/genetics/148.3.929
927 Parish, T. (2014). Two-Component Regulatory Systems of Mycobacteria. *Microbiol Spectr*, *2*(1),
928 MGM2-0010-2013. https://doi.org/10.1128/microbiolspec.MGM2-0010-2013
929 Park, H. D., Guinn, K. M., Harrell, M. I., Liao, R., Voskuil, M. I., Tompa, M., Schoolnik, G. K., &
930 Sherman, D. R. (2003). Rv3133c/dosR is a transcription factor that mediates the hypoxic
931 response of Mycobacterium tuberculosis. *Mol Microbiol*, *48*(3), 833-843.
932 https://doi.org/10.1046/j.1365-2958.2003.03474.x

933 Pepperell, C. S., Casto, A. M., Kitchen, A., Granka, J. M., Cornejo, O. E., Holmes, E. C., Birren, B.,
934 Galagan, J., & Feldman, M. W. (2013). The role of selection in shaping diversity of natural
935 M. tuberculosis populations. *PLoS Pathog*, *9*(8), e1003543.
936 https://doi.org/10.1371/journal.ppat.1003543
937 Pethe, K., Swenson, D. L., Alonso, S., Anderson, J., Wang, C., & Russell, D. G. (2004). Isolation of
938 Mycobacterium tuberculosis mutants defective in the arrest of phagosome maturation.
939 *Proc Natl Acad Sci U S A*, *101*(37), 13642-13647.
940 https://doi.org/10.1073/pnas.0401657101
941 Phelan, J. E., O'Sullivan, D. M., Machado, D., Ramos, J., Oppong, Y. E. A., Campino, S., O'Grady, J.,
942 McNerney, R., Hibberd, M. L., Viveiros, M., Huggett, J. F., & Clark, T. G. (2019). Integrating
943 informatics tools and portable sequencing technology for rapid detection of resistance
944 to anti-tuberculous drugs. *Genome Med*, *11*(1), 41. https://doi.org/10.1186/s13073-019-
945 0650-x
946 Raman, S., Puyang, X., Cheng, T. Y., Young, D. C., Moody, D. B., & Husson, R. N. (2006).
947 Mycobacterium tuberculosis SigM positively regulates Esx secreted protein and
948 nonribosomal peptide synthetase genes and down regulates virulence-associated
949 surface lipid synthesis. *J Bacteriol*, *188*(24), 8460-8468.
950 https://doi.org/10.1128/JB.01212-06
951 Rizvi, A., Shankar, A., Chatterjee, A., More, T. H., Bose, T., Dutta, A., Balakrishnan, K., Madugulla,
952 L., Rapole, S., Mande, S. S., Banerjee, S., & Mande, S. C. (2019). Rewiring of Metabolic
953 Network in Mycobacterium tuberculosis During Adaptation to Different Stresses. *Front*
954 *Microbiol*, *10*, 2417. https://doi.org/10.3389/fmicb.2019.02417
955 Romagnoli, A., Etna, M. P., Giacomini, E., Pardini, M., Remoli, M. E., Corazzari, M., Falasca, L.,
956 Goletti, D., Gafa, V., Simeone, R., Delogu, G., Piacentini, M., Brosch, R., Fimia, G. M., &
957 Coccia, E. M. (2012). ESX-1 dependent impairment of autophagic flux by Mycobacterium
958 tuberculosis in human dendritic cells. *Autophagy*, *8*(9), 1357-1370.
959 https://doi.org/10.4161/auto.20881
960 Rustad, T. R., Harrell, M. I., Liao, R., & Sherman, D. R. (2008). The enduring hypoxic response of
961 Mycobacterium tuberculosis. *PLoS One*, *3*(1), e1502.
962 https://doi.org/10.1371/journal.pone.0001502
963 Sherman, D. R., Mdluli, K., Hickey, M. J., Arain, T. M., Morris, S. L., Barry, C. E., 3rd, & Stover, C. K.
964 (1996). Compensatory ahpC gene expression in isoniazid-resistant Mycobacterium
965 tuberculosis. *Science*, *272*(5268), 1641-1643.
966 https://doi.org/10.1126/science.272.5268.1641
967 Sherman, D. R., Voskuil, M., Schnappinger, D., Liao, R., Harrell, M. I., & Schoolnik, G. K. (2001).
968 Regulation of the Mycobacterium tuberculosis hypoxic response gene encoding alpha -
969 crystallin. *Proc Natl Acad Sci U S A*, *98*(13), 7534-7539.
970 https://doi.org/10.1073/pnas.121172498
971 Shi, R., Zhang, J., Li, C., Kazumi, Y., & Sugawara, I. (2006). Emergence of ofloxacin resistance in
972 Mycobacterium tuberculosis clinical isolates from China as determined by gyrA mutation
973 analysis using denaturing high-pressure liquid chromatography and DNA sequencing. *J*
974 *Clin Microbiol*, *44*(12), 4566-4568. https://doi.org/10.1128/JCM.01916-06
975 Singh, P. R., Goar, H., Paul, P., Mehta, K., Bamniya, B., Vijjamarri, A. K., Bansal, R., Khan, H.,
976 Karthikeyan, S., & Sarkar, D. (2023). Dual functioning by the PhoR sensor is a key

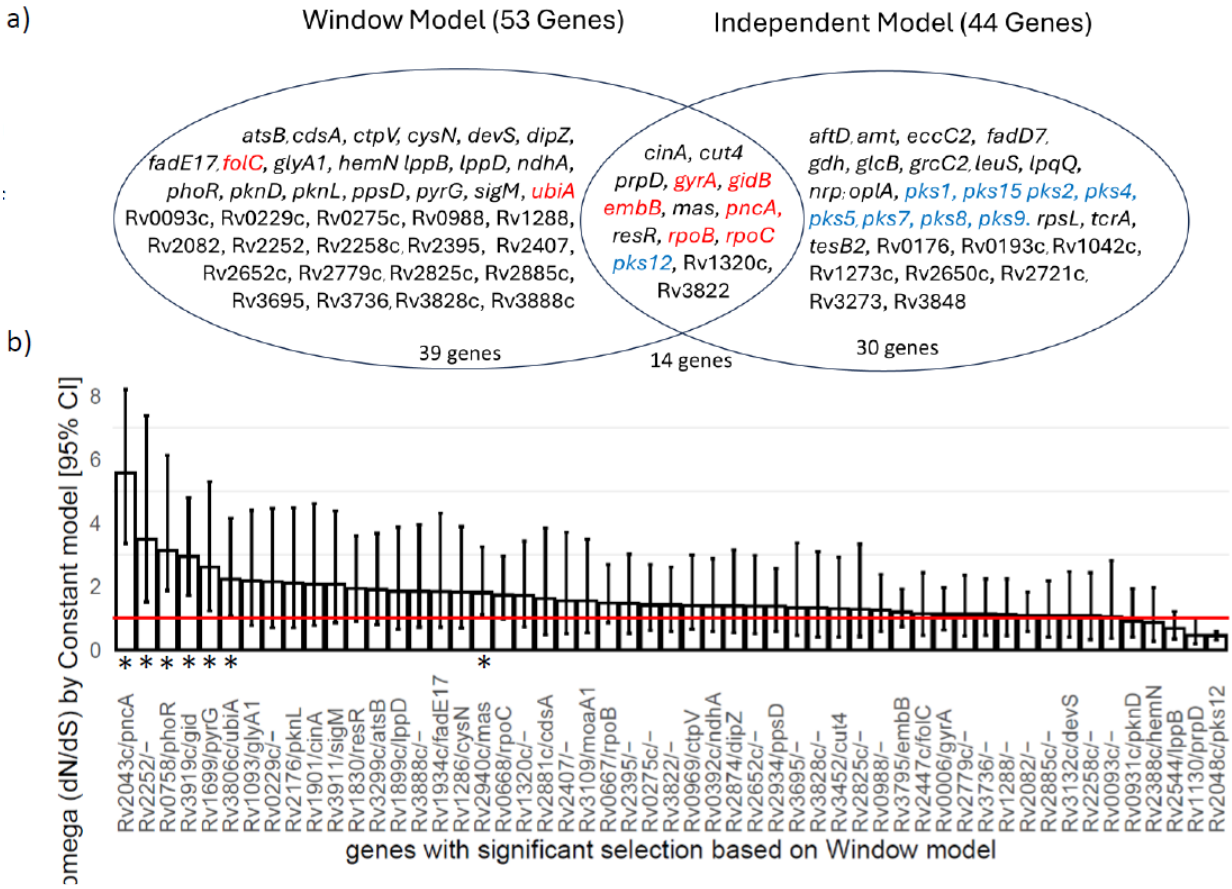
977 determinant to Mycobacterium tuberculosis virulence. *PLoS Genet*, *19*(12), e1011070.
978 https://doi.org/10.1371/journal.pgen.1011070
979 Sirakova, T. D., Thirumala, A. K., Dubey, V. S., Sprecher, H., & Kolattukudy, P. E. (2001). The
980 Mycobacterium tuberculosis pks2 gene encodes the synthase for the hepta- and
981 octamethyl-branched fatty acids required for sulfolipid synthesis. *J Biol Chem*, *276*(20),
982 16833-16839. https://doi.org/10.1074/jbc.M011468200
983 Spies, F. S., Ribeiro, A. W., Ramos, D. F., Ribeiro, M. O., Martin, A., Palomino, J. C., Rossetti, M. L.,
984 da Silva, P. E., & Zaha, A. (2011). Streptomycin resistance and lineage-specific
985 polymorphisms in Mycobacterium tuberculosis gidB gene. *J Clin Microbiol*, *49*(7), 2625-
986 2630. https://doi.org/10.1128/JCM.00168-11
987 Sreevatsan, S., Pan, X., Stockbauer, K. E., Connell, N. D., Kreiswirth, B. N., Whittam, T. S., &
988 Musser, J. M. (1997). Restricted structural gene polymorphism in the Mycobacterium
989 tuberculosis complex indicates evolutionarily recent global dissemination. *Proc Natl*
990 *Acad Sci U S A*, *94*(18), 9869-9874. https://doi.org/10.1073/pnas.94.18.9869
991 Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of
992 large phylogenies. *Bioinformatics*, *30*(9), 1312-1313.
993 https://doi.org/10.1093/bioinformatics/btu033
994 Sun, Q., Xiao, T. Y., Liu, H. C., Zhao, X. Q., Liu, Z. G., Li, Y. N., Zeng, H., Zhao, L. L., & Wan, K. L.
995 (2018). Mutations within embCAB Are Associated with Variable Level of Ethambutol
996 Resistance in Mycobacterium tuberculosis Isolates from China. *Antimicrob Agents*
997 *Chemother*, *62*(1). https://doi.org/10.1128/AAC.01279-17
998 Sun, Z., Zhang, J., Zhang, X., Wang, S., Zhang, Y., & Li, C. (2008). Comparison of gyrA gene
999 mutations between laboratory-selected ofloxacin-resistant Mycobacterium tuberculosis
1000 strains and clinical isolates. *Int J Antimicrob Agents*, *31*(2), 115-121.
1001 https://doi.org/10.1016/j.ijantimicag.2007.10.014
1002 Tiwari, S., Casey, R., Goulding, C. W., Hingley-Wilson, S., & Jacobs, W. R., Jr. (2019). Infect and
1003 Inject: How Mycobacterium tuberculosis Exploits Its Major Virulence-Associated Type VII
1004 Secretion System, ESX-1. *Microbiol Spectr*, *7*(3).
1005 https://doi.org/10.1128/microbiolspec.BAI-0024-2019
1006 Trauner, A., Liu, Q., Via, L. E., Liu, X., Ruan, X., Liang, L., Shi, H., Chen, Y., Wang, Z., Liang, R.,
1007 Zhang, W., Wei, W., Gao, J., Sun, G., Brites, D., England, K., Zhang, G., Gagneux, S., Barry,
1008 C. E., 3rd, & Gao, Q. (2017). The within-host population dynamics of Mycobacterium
1009 tuberculosis vary with treatment efficacy. *Genome Biol*, *18*(1), 71.
1010 https://doi.org/10.1186/s13059-017-1196-0
1011 Vilcheze, C., Weisbrod, T. R., Chen, B., Kremer, L., Hazbon, M. H., Wang, F., Alland, D.,
1012 Sacchettini, J. C., & Jacobs, W. R., Jr. (2005). Altered NADH/NAD+ ratio mediates
1013 coresistance to isoniazid and ethionamide in mycobacteria. *Antimicrob Agents*
1014 *Chemother*, *49*(2), 708-720. https://doi.org/10.1128/AAC.49.2.708-720.2005
1015 Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C. A., Zeng, Q.,
1016 Wortman, J., Young, S. K., & Earl, A. M. (2014). Pilon: an integrated tool for
1017 comprehensive microbial variant detection and genome assembly improvement. *PLoS*
1018 *One*, *9*(11), e112963. https://doi.org/10.1371/journal.pone.0112963
1019 Walters, S. B., Dubnau, E., Kolesnikova, I., Laval, F., Daffe, M., & Smith, I. (2006). The
1020 Mycobacterium tuberculosis PhoPR two-component system regulates genes essential for

1021 virulence and complex lipid biosynthesis. *Mol Microbiol*, *60*(2), 312-330.
1022 https://doi.org/10.1111/j.1365-2958.2006.05102.x
1023 Wilson, D. J., & Consortium, C. R. (2020). GenomegaMap: Within-Species Genome-Wide dN/dS
1024 Estimation from over 10,000 Genomes. *Mol Biol Evol*, *37*(8), 2450-2460.
1025 https://doi.org/10.1093/molbev/msaa069
1026 Wong, S. Y., Javid, B., Addepalli, B., Piszczek, G., Strader, M. B., Limbach, P. A., & Barry, C. E., 3rd.
1027 (2013). Functional role of methylation of G518 of the 16S rRNA 530 loop by GidB in
1028 Mycobacterium tuberculosis. *Antimicrob Agents Chemother*, *57*(12), 6311-6318.
1029 https://doi.org/10.1128/AAC.00905-13
1030 Xu, W., DeJesus, M. A., Rucker, N., Engelhart, C. A., Wright, M. G., Healy, C., Lin, K., Wang, R.,
1031 Park, S. W., Ioerger, T. R., Schnappinger, D., & Ehrt, S. (2017). Chemical Genetic
1032 Interaction Profiling Reveals Determinants of Intrinsic Antibiotic Resistance in
1033 Mycobacterium tuberculosis. *Antimicrob Agents Chemother*, *61*(12).
1034 https://doi.org/10.1128/AAC.01334-17
1035 Yadon, A. N., Maharaj, K., Adamson, J. H., Lai, Y. P., Sacchettini, J. C., Ioerger, T. R., Rubin, E. J., &
1036 Pym, A. S. (2017). A comprehensive characterization of PncA polymorphisms that confer
1037 resistance to pyrazinamide. *Nat Commun*, *8*(1), 588. https://doi.org/10.1038/s41467-
1038 017-00721-2
1039 Yang, C., Sobkowiak, B., Naidu, V., Codreanu, A., Ciobanu, N., Gunasekera, K. S., Chitwood, M.
1040 H., Alexandru, S., Bivol, S., Russi, M., Havumaki, J., Cudahy, P., Fosburgh, H., Allender, C.
1041 J., Centner, H., Engelthaler, D. M., Menzies, N. A., Warren, J. L., Crudu, V., . . . Cohen, T.
1042 (2022). Phylogeography and transmission of M. tuberculosis in Moldova: A prospective
1043 genomic analysis. *PLoS Med*, *19*(2), e1003933.
1044 https://doi.org/10.1371/journal.pmed.1003933
1045 Yang, J., Li, J., Dong, L., & Grunewald, S. (2011). Analysis on the reconstruction accuracy of the
1046 Fitch method for inferring ancestral states. *BMC Bioinformatics*, *12*, 18.
1047 https://doi.org/10.1186/1471-2105-12-18
1048 Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol*, *24*(8),
1049 1586-1591. https://doi.org/10.1093/molbev/msm088
1050 Zheng, W., Borja, M., Dorman, L. C., Liu, J., Zhou, A., Seng, A., Arjyal, R., Sunshine, S., Nalyvayko,
1051 A., Pisco, A. O., Rosenberg, O. S., Neff, N., & Zha, B. S. (2025). Single-cell analysis reveals
1052 Mycobacterium tuberculosis ESX-1-mediated accumulation of permissive macrophages
1053 in infected mouse lungs. *Sci Adv*, *11*(3), eadq8158.
1054 https://doi.org/10.1126/sciadv.adq8158

1055

1056

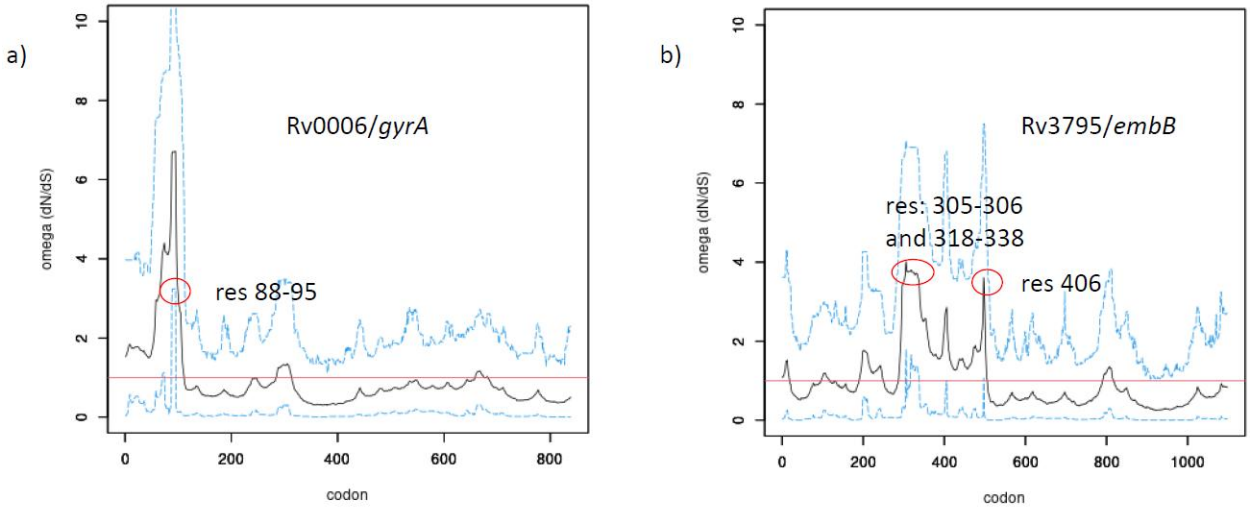
1057 Figure 1.



1058

1059

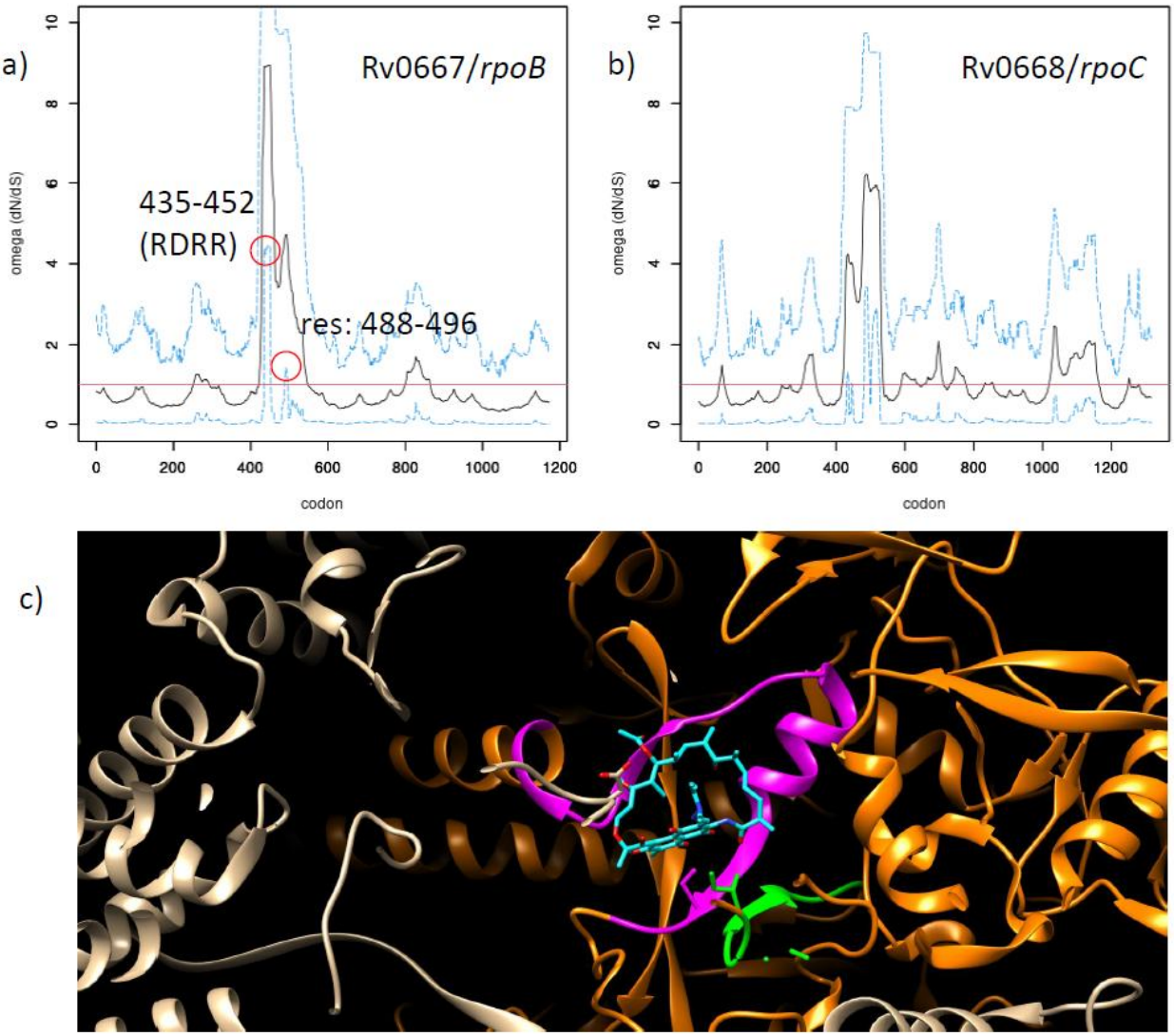
1060 Figure 2.



1061

1062

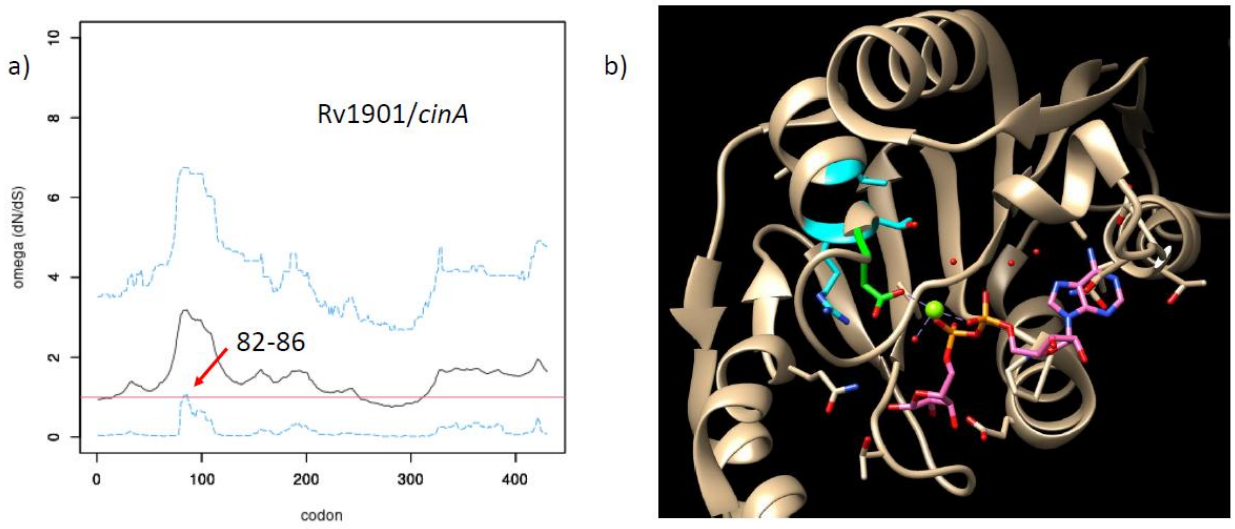
1063 Figure 3.



1064

1065

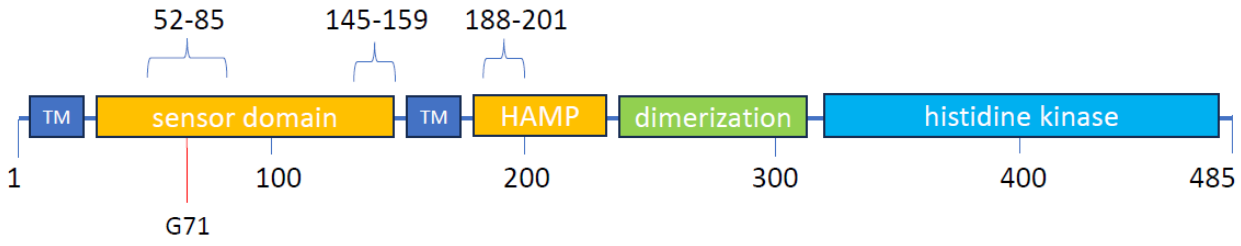
1066 Figure 4.



1067

1068

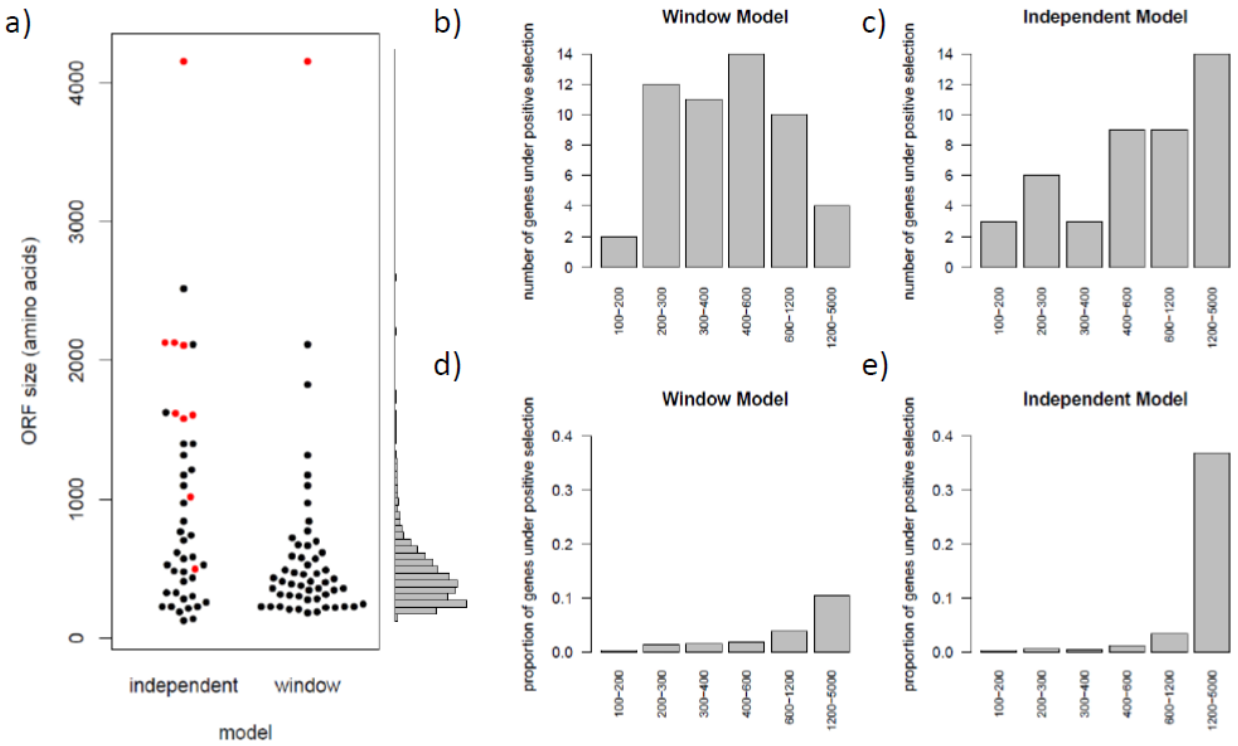
1069 Figure 5.



1070

1071

1072 Figure 6.

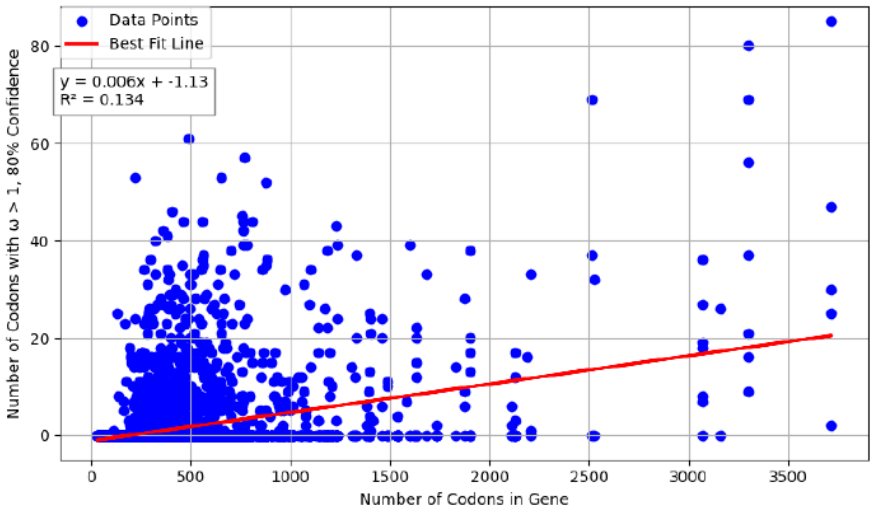


1073

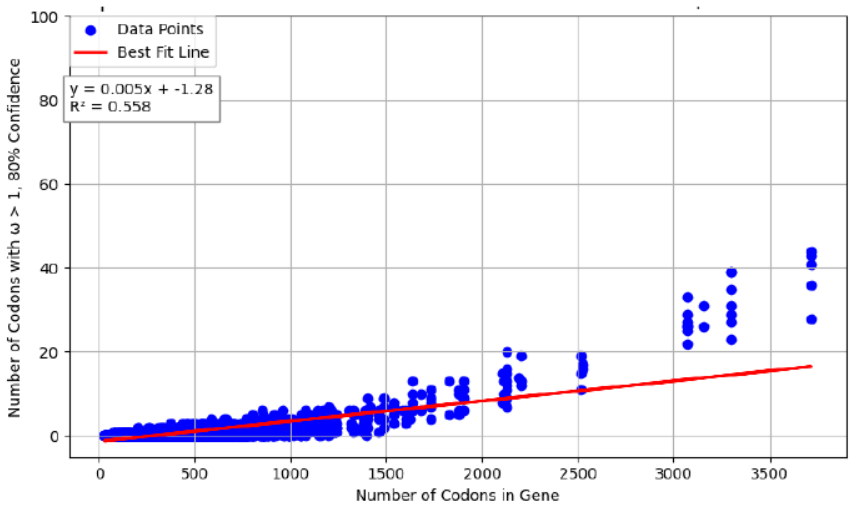
1074

Figure 7.

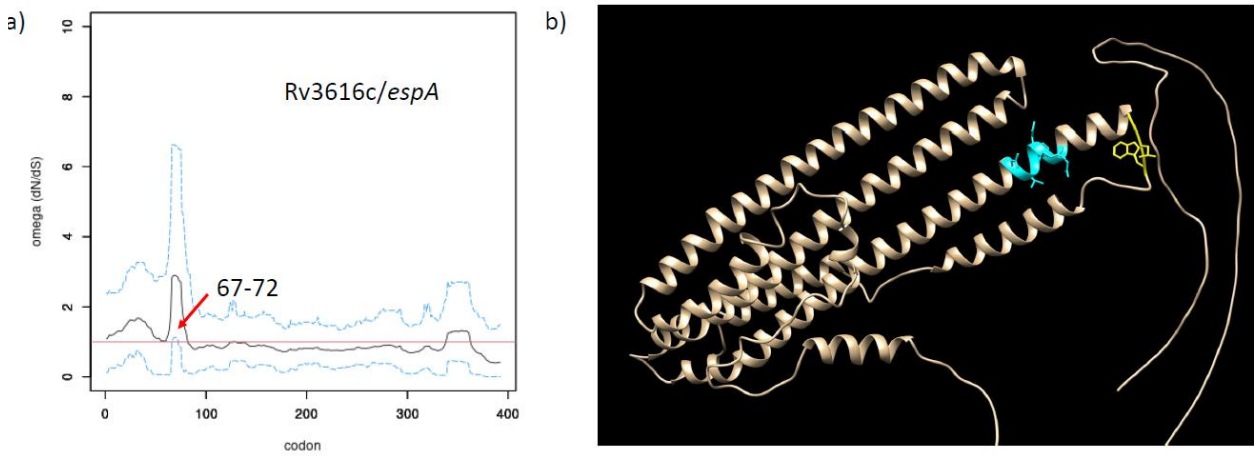
a) Window model



b) Independent model



1079 Figure 8.

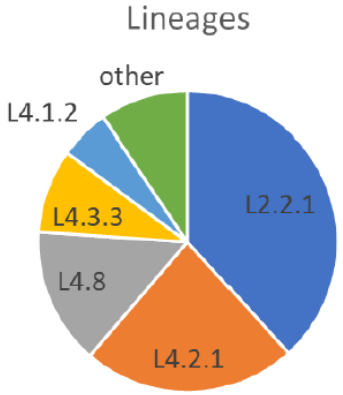


1080

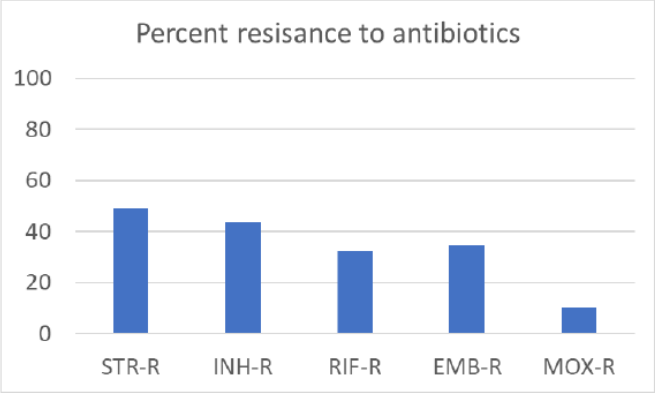
1081

1082 Supplemental Figure S1.

a)



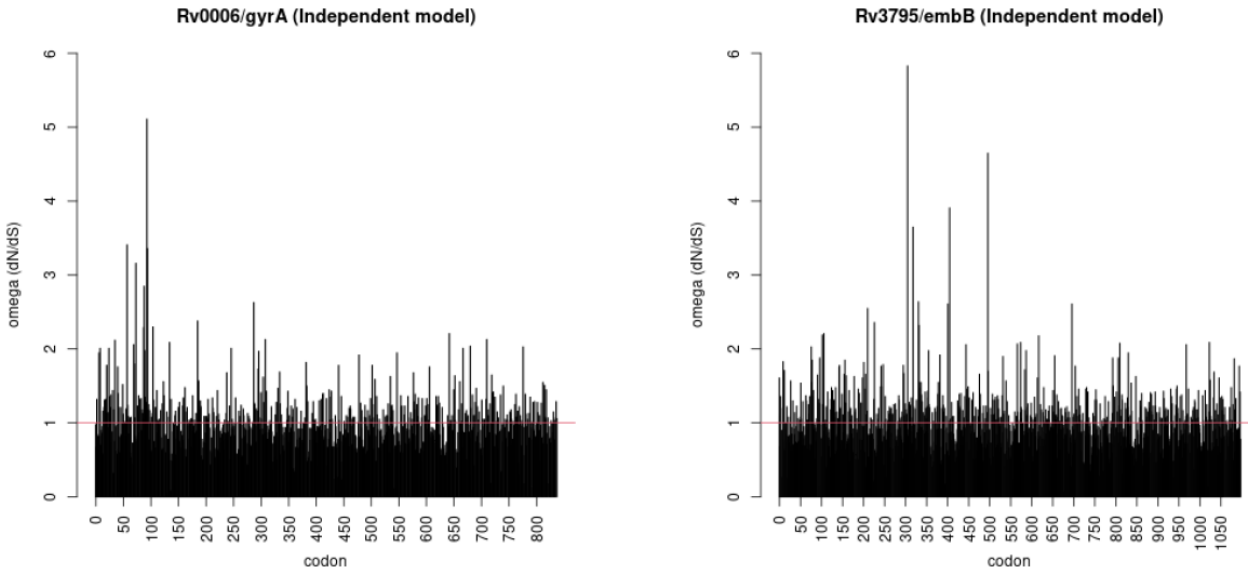
b)



1083

1084

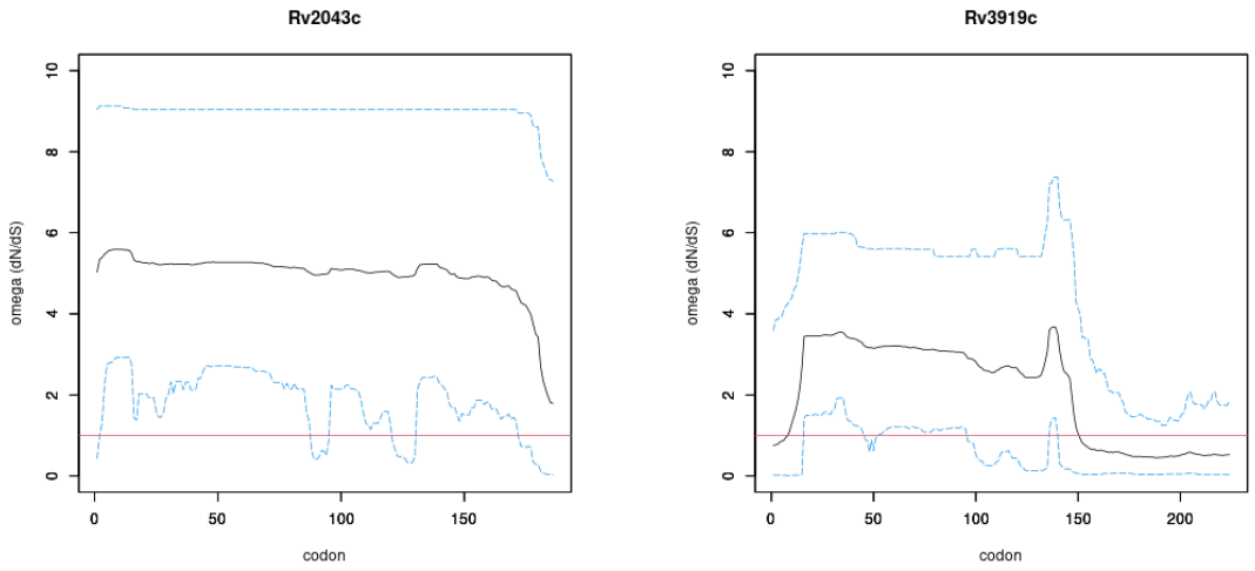
1085 Supplemental Figure S2.



1086

1087

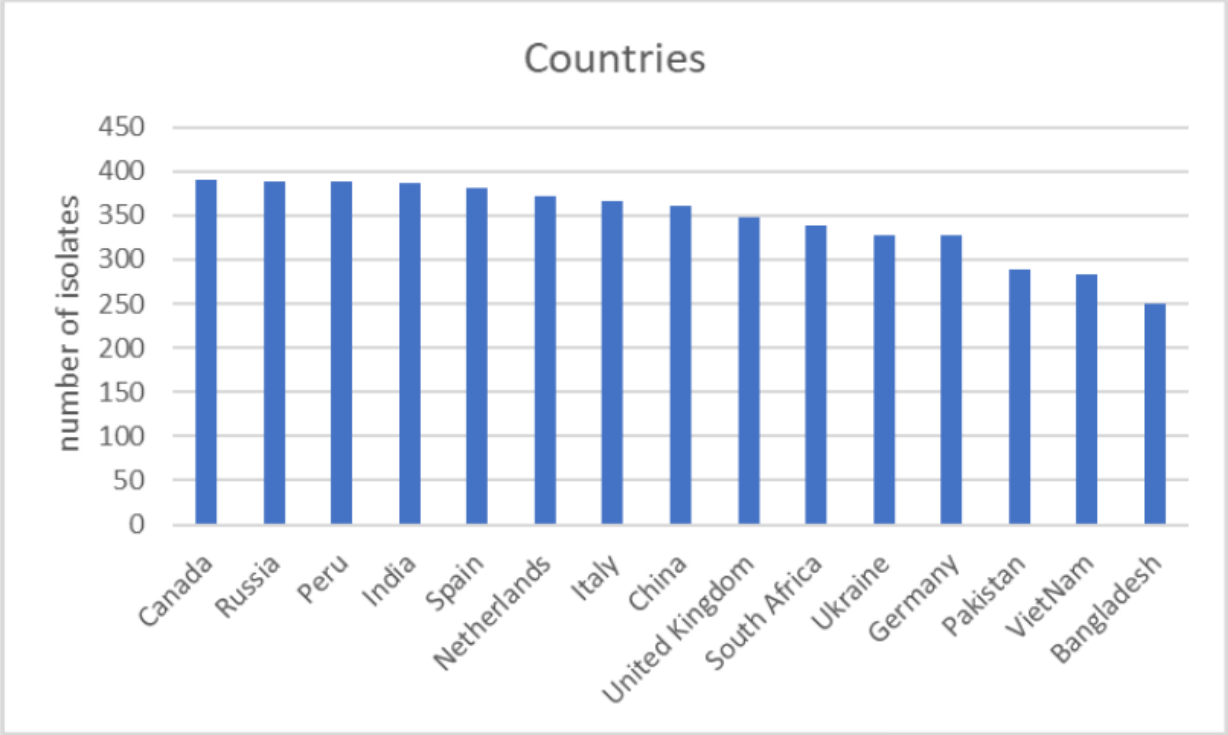
1088 Supplemental Figure S3.



1089

1090

1091 Supplemental Figure S4.

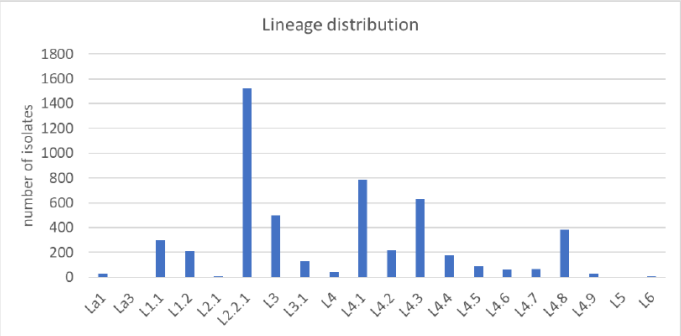


1092

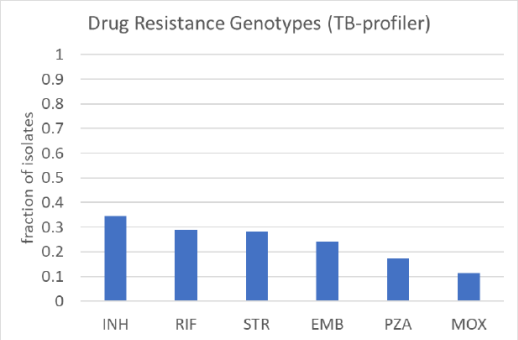
1093

1094 Supplemental Figure S5.

a)



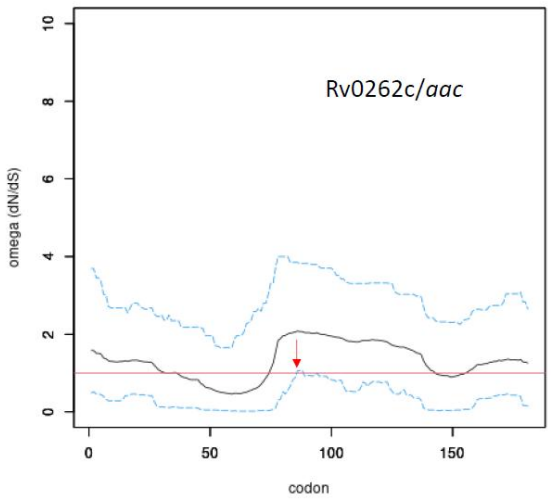
b)

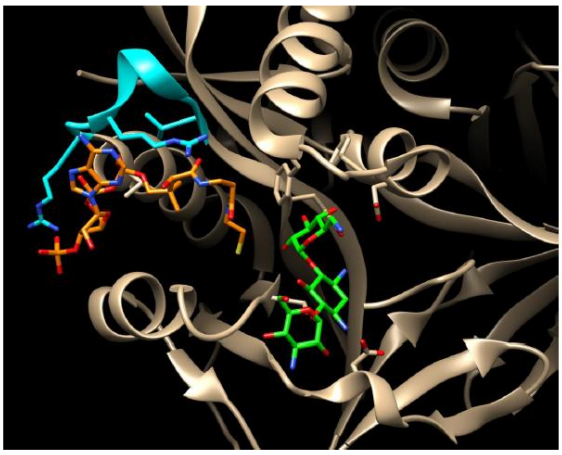


1095

1096

1097 Supplemental Figure S6.





1098

1099