Evo-Scope: Fully automated assessment of correlated evolution on phylogenetic trees

Maxime Godfroid[1], Charles Coluzzi[2], Amaury Lambert[1,3], Philippe Glaser[4], Eduardo P. C. Rocha[2] and Guillaume Achaz[1,5]

Corresponding author: Maxime Godfroid

[1] SMILE group, Center for Interdisciplinary Research in Biology (CIRB), Collège de France, CNRS, INSERM, Université PSL, Paris, France

[2] Institut Pasteur, Université de Paris Cité, CNRS UMR3525, Microbial Evolutionary Genomics, Paris, France

[3] Laboratoire de Probabilités, Statistique et Modélisation (LPSM), Sorbonne Université, CNRS UMR 8001, Université de Paris, Paris, France

[4] Institut Pasteur, Université de Paris Cité, CNRS UMR6047, Ecology and Evolution of Antibiotics Resistance, Paris, France

[5] Éco-anthropologie, Muséum National d'Histoire Naturelle, CNRS UMR 7206, Université de Paris, Paris, France

Abstract

1. Correlated evolution describes how multiple biological traits evolve together. Recently developed methods provide increasingly detailed results of correlated evolution, sometimes at elevated computational costs.

2. Here, we present ***evo-scope***, a fast and fully-automated pipeline with minimal input requirements to compute correlation between discrete traits evolving on a phylogenetic tree.

Notably, we improve two of our previously developed tools that efficiently compute statistics of correlated evolution to characterize the nature, such as synergy or antagonism, and the strength of the interdependence between the traits.

3. Furthermore, we improved the running time and implemented several additional features, such as genetic mapping, Bayesian Markov Chain Monte Carlo estimation, consideration of missing data and phylogenetic uncertainty.

4. As an application, we scan a published penicillin resistance data set of *Streptococcus pneumoniae* and characterize genetic mutations that correlate with antibiotic resistance. The pipeline is accessible both as a self-contained github repository (https://github.com/Maxime5G/EvoScope) and through a graphical galaxy interface (https://galaxy.pasteur.fr/workflows/run?id=44cf9f2340be2b81).

Keywords: phylogenetics, correlated evolution, workflow

Introduction

The study of correlated evolution between biological traits illustrates the tight interdependence between processes in evolutionary biology. The constant development of new tools (listed in Table S1) to understand these dependencies now allow one to predict residues in contact in 3D structures of proteins (Morcos et al., 2011) or to identify protein-protein interactions (Barker & Pagel, 2005). Nevertheless, the measuring of these correlations is complicated by the phylogenetic relations (i.e., phylogenetic inertia) between living entities, typically represented by a phylogenetic tree. As such, species belonging to the same taxon may share several residues or biological traits by vertical descent, a process which one usually wants to disentangle from converge due to other processes, such as natural selection

47 (Achaz & Dutheil, 2021). Therefore, the measurement of correlation between traits must
48 either correct, or explicitly account for phylogenetic structure.
49
50 A typical application of the study of correlated evolution is the measure of the correlation
51 between genetic factors and a given phenotype, i.e., genetic mapping, generally performed
52 through genome-wide association studies (GWAS). Multiple tools have been devoted to
53 measure such correlation, and these use different methods to account for phylogenetic
54 relationships (see Table S1). While most methods add a correction term for this, few rely
55 explicitly on the topology of the phylogenetic tree (e.g., treeWAS (Collins & Didelot, 2018)).
56 Particularly in bacteria, tree-aware GWAS has been critical to detect genetic variants involved
57 in the evolution of antibiotic resistance (Farhat et al., 2019), virulence (Galardini et al., 2020)
58 or niche adaptation (Gori et al., 2020).
59
60 We have previously developed two methods to measure correlated evolution on
61 phylogenetic trees. First, *epics* calculates all possible orderings of mutational events affecting
62 two traits occurring on a phylogenetic tree, and computes efficiently the exact probability to
63 observe a number of co-counts equal or larger than the observed repartition (Behdenna et
64 al., 2016). The second tool, *epocs*, computes the influence of the mutation of a first trait on
65 the mutation rate of a second one by maximum likelihood (Behdenna et al., 2021). Notably,
66 *epocs* assesses different possible interactions between the traits, such as induction or
67 inhibition.
68

Here we describe a novel software, *evo-scope* (EVOlutionary Study of Correlations of Occurrences on a Phylogeny), combining *epics* and *epocs*, to ease the detection of correlated evolution on phylogenetic trees for discrete traits. We further detail new features that we have added in the *evo-scope* implementation. Notable improvements include accelerated run-times (e.g., using sparse matrices in *epics*), consideration of missing data in species traits, analysis of a single trait (i.e., GWAS), the analysis of forests of trees and a Bayesian Markov Chain Monte Carlo (MCMC) estimation.

Presentation of *evo-scope*

The details of the *epics* and *epocs* models are described in Behdenna (2016, 2021). *Evo-scope* provides new versions of these tools with additional features.

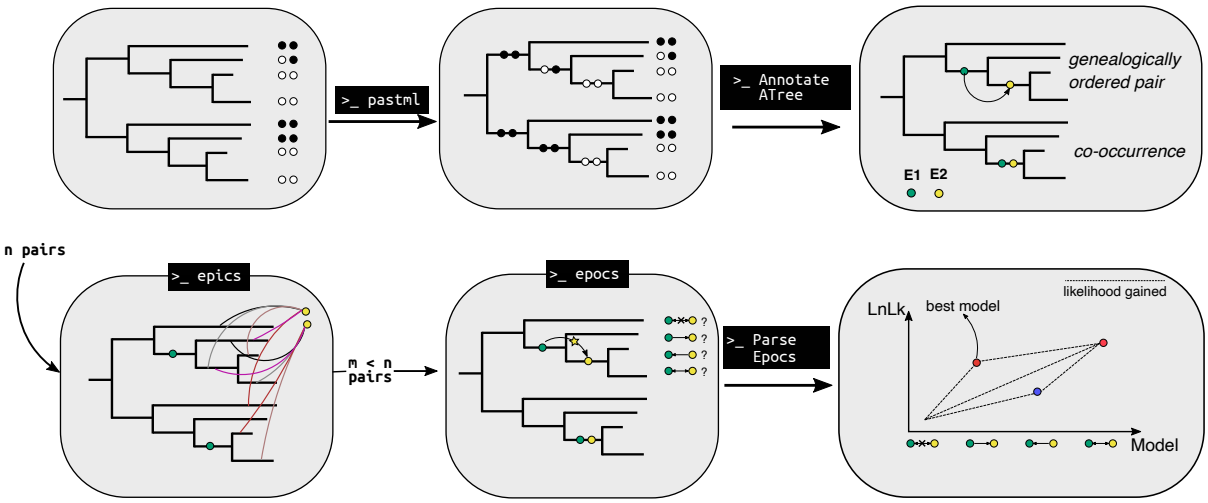


Figure 1. Schematic representation of the *evo-scope* pipeline (see Methods). The user supplies a rooted phylogenetic tree with the values of *n* discrete traits at the tips. Pastml reconstructs the evolution of each trait along the tree and an R script formats the reconstruction to display the mutations on the branches. *Epics* is run as a pre-filter to identify significantly associated pairs of evolutionary processes (*m*) among the initial *n*. Next, *epocs*

86 maximizes the likelihoods of four evolutionary models, one of independence and three of
87 dependence, on the selected pairs. Finally, an R script summarizes the results and outputs
88 the best model explaining the repartition of events on the tree. Additionally, a figure can be
89 generated for any pair considered to describe the gains of likelihood between nested
90 models.
91

92 Description of the pipeline

93 The *evo-scope* pipeline consists of five steps: ancestral character reconstruction, parsing and
94 formatting, running *epics*, running *epocs* and summarizing the results (Fig 1).

95 1. Ancestral Character Reconstruction (ACR)

96 As initial input, a user provides a rooted phylogenetic tree and a file with the values of discrete
97 traits at the tips. Pastml reconstructs the trait states on the internal branches of the tree using
98 the JOINT model (Ishikawa et al., 2019; Pupko et al., 2000). Any other ACR tool or algorithm
99 (e.g., Maximum A Posteriori in *pastml*) can be used as long as one single state (defined or
100 undefined) is provided on each branch.

101 2. ACR parsing and tree reformatting

102 *Evo-scope* reconstructs all mutational events on the branches based on the ancestral
103 reconstruction. A mutational event is defined as any change of a trait value between two
104 consecutive nodes of the tree. At most one mutation per branch per trait is allowed: either a
105 change occurred between the parent and daughter node or none.

106 3. *Epics* as a pre-filter

107 *Evo-scope* then runs *epics* to discover by default pairs of traits that co-occur frequently on
108 the branches of the trees (Behdenna et al., 2016). Other scenarios for the pairs of traits

109 placement are available, such as genealogically ordered pairs. At this step, *epics* is used as a
110 pre-filter as it runs very fast and it only detects statistical association between the mutational
111 events on the tree. Importantly, *epics* does not allow to predict an evolutionary scenario. As
112 such, additional in-depth analysis is required to better describe the interactions between the
113 mutational events. However, for very large data sets, *epics* might be the only option available
114 as running time increase substantially for the maximum likelihood-based tool *epocs*. The
115 output of this step is the list of all pair of events that are significantly associated on the tree
116 for a given p-value.

117 4. *Epocs* for the model discovery

118 For all significantly associated traits, *epocs* maximizes the likelihood of multiple scenarios of
119 induction to determine which of the trait influenced the occurrence of the other one. The
120 multiple scenarios can contain from two to eight parameters and describe the
121 interrelationship between the occurrence of two events on the tree. The parameters are
122 divided into natural occurrence rates ($\mu_{\{i,j\}}$, $\upsilon_{\{i,j\}}$) and excited occurrence rates ($\mu^*_{\{i,j\}}$, $\upsilon^*_{\{i,j\}}$)
123 for each of the trait (Fig 3A). The ratio between the excited occurrence rates and the natural
124 occurrence rates defines the induction.

125 In the current version, the model exploration is restricted to three models of correlation
126 compared to a model of independence. The three models of correlation describe either an
127 induction of the first trait on the second one, vice-versa or a co-induction of both events.

128 5. Summarizing the results

129 Finally, *evo-scope* extracts the likelihoods and parameter values inferred for each of the
130 models. To select the best model for each pair, the program performs likelihood ratio tests
131 (LRT) between nested models (Behdenna et al., 2021). Next, it compares the calculated LRT

132 value to a χ^2 distribution whose degrees of freedom is the difference in the number of
133 parameters between models. Finally, based on significant LRTs, *evo-scope* selects the "best"
134 model by a trade-off measure which maximizes the LRT value and minimizes the number of
135 model parameters. *Evo-scope* then allows the user to select a pair and plot the likelihood
136 gains between nested models.
137
138 **Additional novel features to *epics* and *epocs***
139 1. Analysis on multiple independent trees
140 The user can provide a list of independent trees (non-overlapping taxa), to either *epics* or
141 *epocs*, with the same type of events. This feature is particularly useful whenever the same
142 character can be analyzed across unrelated clades. In particular, *epics* and *epocs* will collect
143 the information of all the trees and generate a global result by integrating the information
144 across all trees. To do so, *epics* computes an aggregated p-value convoluted across all trees,
145 while *epocs* transmits the likelihood calculation across all trees.
146 2. Forest of trees
147 To account for phylogenetic uncertainty, the user can provide a forest of trees, such as
148 generated by BEAST or MrBayes (Ronquist et al., 2012; Suchard et al., 2018), to either *epics*
149 or *epocs*, and the results are ranked to extract the mode and 95% credibility intervals. In
150 *epics*, the p-values are ranked by the tree likelihoods, while in *epocs* the results are ranked
151 by the product of the *epocs* likelihood and the tree likelihood.
152 3. MCMC exploration of likelihood surface
153 In the standard *epocs*, the likelihood is maximized considering all possible arrangements,
154 with some restrictions on the number of co-occurrences due to computational complexity. In

155 the MCMC case, the program is allowed to explore freely any possible arrangement of the
156 co-occurrences. This feature is implemented in a third executable *epocs_mcmc*. In this
157 module, the user supplies a tree with events to analyze and a standard Metropolis-Hastings
158 algorithm will explore the parameter distribution across the number of replicates selected by
159 the user (Hastings, 1970). Notable behaviors of the MCMC chain determine the relevancy of
160 each parameter in the model. For example, a flat uniform distribution of the posterior is likely
161 to reflect that this parameter is not pertinent in the model. In addition, the algorithm explores
162 the order in the co-occurrences of the traits. *Epocs_mcmc* outputs a tab-delimited file
163 compatible with Tracer (Rambaut et al., 2018).

164 4. Missing data

165 The presence of uncharacterized traits at the tips can be taken into account in each of the
166 tools above. For *epics*, the branches where traits are unknown are masked, whereas for *epocs*
167 and *epocs_mcmc*, branches containing an unknown trait value and those below are excluded
168 in the likelihood calculation.

169 5. GWAS-like procedure

170 *Epics* and *epocs* have now the possibility of selecting either one character or two to test in
171 the procedures. Whenever the user selects only one character, the analysis is similar to a
172 GWAS, where any other character is tested against the one selected.

173

174 **Analysis of an antibiotic resistance dataset**

175 As test data, we retrieved 603 genomes of *Streptococcus pneumoniae* (Chewapreecha et al.,
176 2014; Croucher et al., 2013, 2015; Lees et al., 2018). In this data set, the resistance to
177 penicillin is known and encoded as a binary variable (R for resistant strains, S for susceptible

178 strains). Also, 198,248 single-nucleotide polymorphisms (SNPs) have been called in the
179 genomes. For the post-hoc MCMC analysis, we retrieved a significant association from *evo-*
180 *scope* and ran *epocs_mcmc* with a chain length of 10^7 steps and sampling every 50 steps
181 on this pair.
182

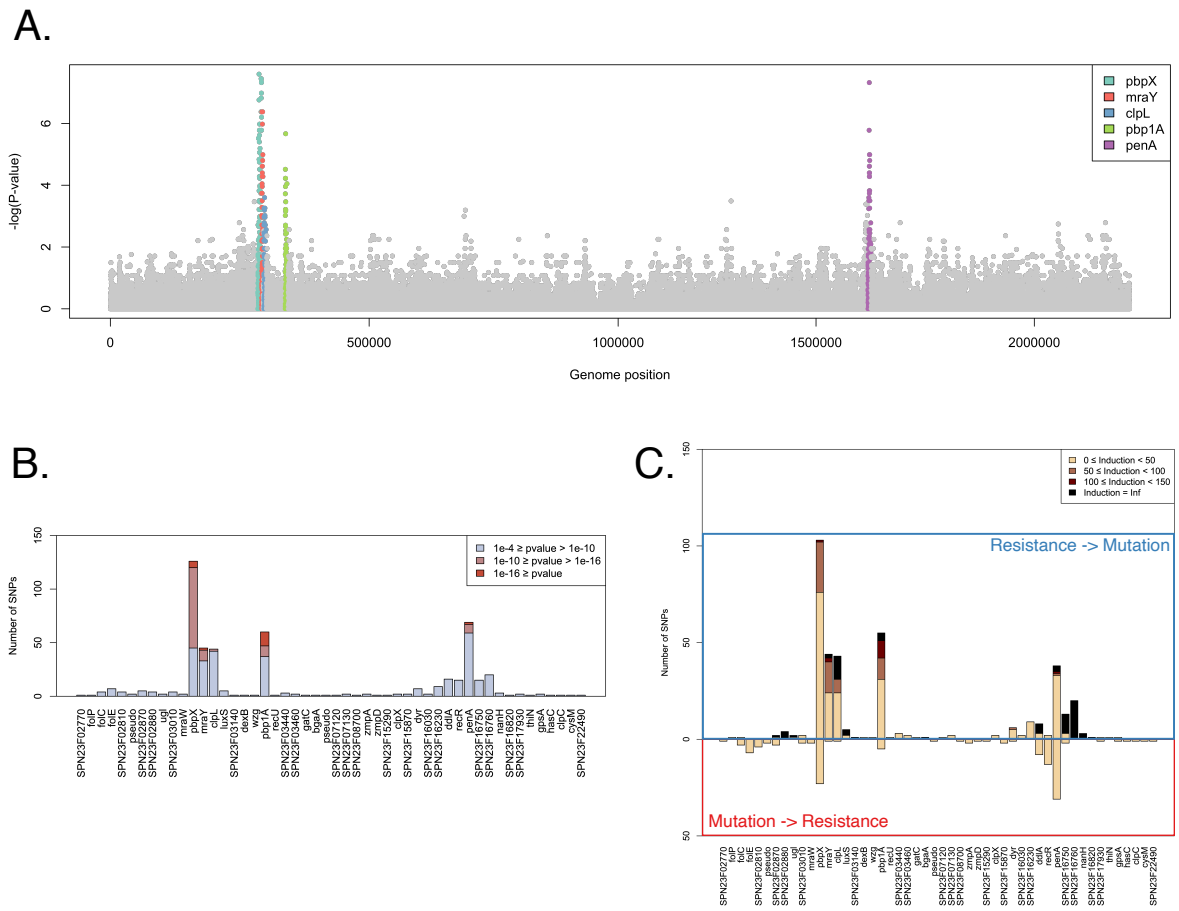
183 **Availability**

184 The source code of *evo-scope*, *epics* and *epocs* are available on github under the GNU
185 General Public License v3: https://github.com/Maxime5G/EvoScope, and pastml is available
186 either on bioconda (Grüning et al., 2018) or on github:
187 https://github.com/evolbioinfo/pastml. We also provide a galaxy access to a "push-button"
188 pipeline at the Pasteur institute server:
189 https://galaxy.pasteur.fr/workflows/run?id=44cf9f2340be2b81.
190

191 **Results**

192 To illustrate the applicability of the present pipeline, we evaluated *evo-scope* by performing
193 a GWAS-like analysis to a data set of 603 genomes of *S. pneumoniae* (Chewapreecha et al.,
194 2014; Croucher et al., 2013, 2015; Lees et al., 2018). As a positive control, *evo-scope*
195 detected genetic variants correlated to penicillin resistance (following (Chewapreecha et al.,
196 2014)). In addition, *evo-scope* inferred the induction values for the significant associations to
197 determine the strength and the direction of the correlation between the occurrence of
198 antibiotic resistance and the SNPs. Of the 82,829 SNPs present in at least eight genomes, we
199 report a total of 1,629 SNPs (1.97%) associated with the resistance by *epics* (Fig. 2A) – of
200 which 543 (33.3% of the significant SNPs) were further inferred to be correlated with the

resistance by *epocs*, with a P-value < 10^{-4} (Fig. 2B,C). In the end, *evo-scope* found that nearly half of the correlated SNPs (n=255, 47%) are located in only three genes, *penA*, *pbpX* and *pbp1A*, in accordance with previous results, as mutations in these three genes are the major drivers of penicillin resistance in *S. pneumoniae* (Chewapreecha et al., 2014) (Fig. 2B,C). In addition, we detected 45 significant SNPs in *mraY*, a gene involved in cell wall biogenesis (Chewapreecha et al., 2014), and 44 SNPs in *clpL*, reported to be related to penicillin susceptibility and compensation of the antibiotic resistance fitness cost (Hakenbeck et al., 2012; Tran et al., 2011). In our study, most of the inferred inductions revealed that the resistance phenotype occurred first in the tree, with the genetic mutations occurring very quickly afterwards (i.e., induction values > 100 – Fig 2C).



212 Figure 2. **Application of *evo-scope* to a dataset of *S. pneumoniae*.** (A) Manhattan plot of the
213 P-values inferred by *epics*. In color are shown genes of interest with multiple SNPs of low p-
214 values. Notably, the three highest peaks correspond to *pbpX*, *pbp1A* and *penA*. (B) Barplot
215 showing the p-value inferred by *epocs*, grouped by gene IDs. (C) Barplots showing the
216 inductions inferred by *epocs* (*i.e.*, the lambda in (Behdenna et al., 2021)). Top row represents
217 the inductions Resistance -> Mutation, bottom row represents the other way round. For each
218 statistical association between the resistance and one of the 543 significant SNP, we retrieved
219 the best model using LRT. Most notably, the three genes of interest *pbpX*, *penA* and *pbp1A*
220 display low P-values and strong inductions.
221
222 Additionally, we include a post-hoc MCMC analysis with *epocs_mcmc* to demonstrate the
223 usefulness of this addition in the *evo-scope* toolbox. For this purpose, we tested it on a
224 significant association from *evo-scope*, which indicated an induction resistance -> mutation.
225 (Table 1 and Fig. 3B).
226
227
228
229
230
231
232
233
234

235 Table 1. Parameter estimates of the MCMC run for the selected pair.

Parameter	*ML estimates*	*Mean of the bayesian estimates*	*Mode of the posterior distribution*	*95% Credible interval*
μ1	46.263	41.03	38.81	[25.39 – 57.51]
μ1*	292.22	543.25	503.8	[214.42 – 934.32]
μ2	6.8197	14.66	12.79	[4.41 – 25.73]
μ2*	377.08	338.24	317.8	[148.78 – 543.70]
υ1	41.474	46.23	41.76	[20.36 – 73.91]
υ1*	?	475.22	68.12	[0.33 – 37.86]
υ2	0	4.67	0.5724	[1.4E-5 – 13.94]
υ2*	0	294.33	41.59	[3.6E-5 – 41.23]

236 Parameter naming follows the convention from (Behdenna et al., 2021). Non-starred rates
237 represent natural occurrence rates, whereas starred rates represent excited occurrence rates
238 (see also Fig 3A for a summary graph). ML estimates are the values of each parameter
239 selected with the free-for-all model in *epocs* (model "8"). Mean of the Bayesian estimates
240 and mode of the posterior distribution are, respectively, the mean and mode values
241 calculated the MCMC chain. The last column shows the 95% credible intervals of the posterior
242 distribution.

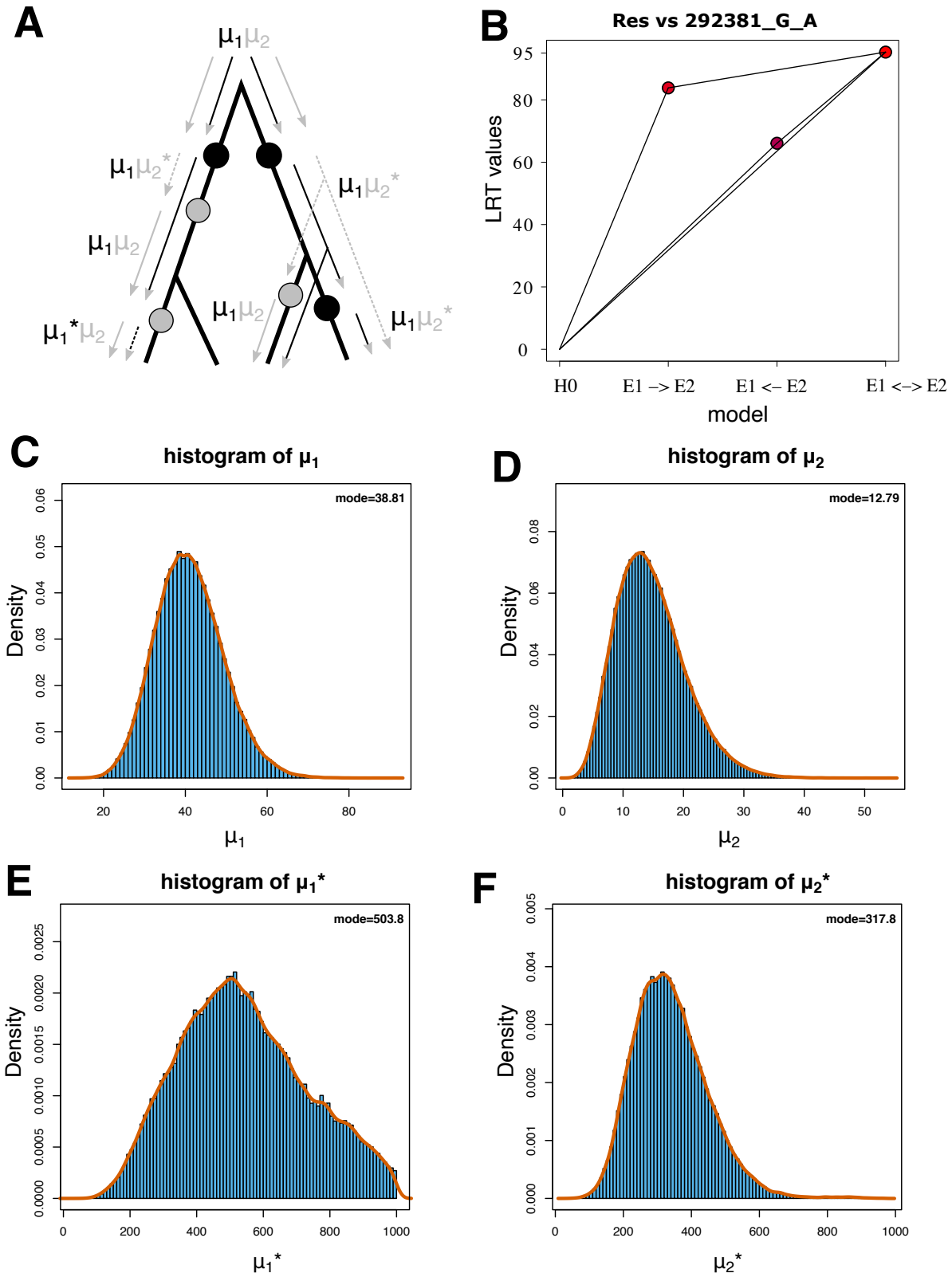


Figure 3. Likelihood ratios and output from an *epocs_mcmc* run on one significant association between the resistance phenotype and a SNP in *pbpX*. (A) Summary graph of parameter

models and transitions upon evolutionary events on a mock tree, adapted from (Behdenna et al., 2021). Non-starred rates are natural rates of occurrence, starred rates are excited rates of occurrence. An occurrence of the first event in the pair activates the excited rates of the second event in the pair. Once the second event occurs, the excitation is consumed. (B) Likelihood ratios between the four models tested by *evo-scope*, following the convention from (Behdenna et al., 2021) . Lines connect nested models. The comparison providing the best likelihood gain while minimizing the number of parameters is the comparison between the model of independence and the model of induction E1->E2 (i.e., Resistance -> Mutation). (C-F) Histograms of parameter values taken from the MCMC chain. Scales are parameter-based in order to best display the shape of the parameter distribution.

After running *epocs_mcmc*, we see that $\mu 1$ is clearly defined, with a narrow distribution (Fig 3C). Interestingly, even though the preferred model by ML is the induction Resistance -> Mutation, $\mu 1^*$ (demonstrating an induction Mutation -> Resistance) is still somewhat relevant to the model, but the confidence interval is quite wide (Fig 3E). Importantly, $\mu 2$ is close to zero and $\mu 2^*$ is clearly defined with a narrow distribution (Fig 3D,F). We conclude from these observations that the selected mutation in *pbpX* mostly occurs after the resistance, therefore indicating an induction Resistance -> Mutation. In addition, the order of the co-occurrences are inferred in the same direction as ML and further, the $\boldsymbol{\upsilon}$ parameter posterior distributions are centered around the parameter values inferred by ML (Fig S1).

Discussion

Inferring correlations between evolutionary events has multiple useful applications. However, fast and easily interpretable tools are still lacking. Here, we provide a self-contained and fully automated pipeline to detect correlated evolution, with minimal input required from the user. *Evo-scope* takes as input any number of discrete traits and a rooted phylogenetic tree and outputs a model of correlated evolution best explaining the repartition of the events on the tree. The flexibility of *evo-scope* allows to perform different types of analyses, such as GWAS as exemplified in this paper. Here, we applied *evo-scope* on a well-described dataset of antibiotic resistance and SNPs in *S. pneumoniae* to 1) test the accuracy of the pipeline and 2) to add on the literature possible inductions either from the mutations to the resistance or in the other way round. We showed that *evo-scope* is able to retrieve consistent results with the literature in a timely manner.

The Induction values inferred on the significantly associated pairs revealed that most of the mutations associated with the resistance occurred after the resistance phenotype. Two effects may contribute to explain this result. Methodologically, the ACR step might introduce biases regarding the real occurrence timing of both mutations and resistance as the algorithm tries to minimize the number of steps to reconstruct a trait, while antibiotic resistance and resistance-conferring mutations are known to be subject to homoplasies (see for example (Coolen et al., 2021)).

In bacteria, genomic changes that are necessary to evolve antibiotic resistance are often a burden for the bacteria, imposing a fitness cost (Andersson & Hughes, 2010). However,

292 mechanisms exist that compensate the cost of such antibiotic resistance, such as the evolution
293 of a secondary mutation in the genome. For *S. pneumoniae*, bacteria exhibiting multiple
294 mutations in the penicillin binding proteins exhibit a phenotype of resistance and
295 compensation (Orio et al., 2011). Using *evo-scope*, we showed that many significantly
296 associated pairs of mutations concur with these observations. Furthermore, the patterns of
297 mutation acquisitions and inductions could shed light on evolutionary trajectories of the
298 acquisition of antibiotic resistance in *S. pneumoniae*.
299
300 In conclusion, we developed a fully-automated pipeline, with minimal input required from the
301 user. Our approach, by considering explicitly the phylogenetic component, allows to detect
302 correlated evolution between discrete traits of any species by explicitly taking into account
303 the underlying structure. We show that our pipeline can also be applied to tree-aware GWAS
304 analyses in bacteria, expanding previous results in the field linking genetic variants to
305 penicillin resistance in *S. pneumoniae*.

306 ## Acknowledgements
307 This work used the computation and storage services (TARS cluster) provided by the IT
308 department at Institut Pasteur, Paris.
309 ## Funding
310 This work was supported by INCEPTION (PIA/ANR-16-CONV-0005) and Equipe FRM
311 (Fondation pour la Recherche Médicale) EQU201903007835.

312 ## Author contribution
313 GA, MG, AL, ER, PG, CC conceived the ideas and designed methodology; MG collected the
314 data; GA, MG analyzed the data; MG led the writing of the manuscript. GA, AL, PG, ER

obtained the funding for this project. All authors contributed critically to the drafts and gave final approval for publication.

Conflicts of interest: None

References

Achaz, G., & Dutheil, J. (2021). Correlated evolution: Models and methods. *ArXiv:2103.11809 [q-Bio]*. https://doi.org/10.48550/arXiv.2103.11809

Andersson, D. I., & Hughes, D. (2010). Antibiotic resistance and its cost: Is it possible to reverse resistance? *Nature Reviews Microbiology*, *8*(4), 260–271. https://doi.org/10.1038/nrmicro2319

Barker, D., & Pagel, M. (2005). Predicting Functional Gene Links from Phylogenetic-Statistical Analyses of Whole Genomes. *PLOS Computational Biology*, *1*(1), e3. https://doi.org/10.1371/journal.pcbi.0010003

Behdenna, A., Godfroid, M., Petot, P., Pothier, J., Lambert, A., & Achaz, G. (2021). A minimal yet flexible likelihood framework to assess correlated evolution. *Systematic Biology*, syab092. https://doi.org/10.1093/sysbio/syab092

Behdenna, A., Pothier, J., Abby, S. S., Lambert, A., & Achaz, G. (2016). Testing for Independence between Evolutionary Processes. *Systematic Biology*, *65*(5), 812–823. https://doi.org/10.1093/sysbio/syw004

Chewapreecha, C., Marttinen, P., Croucher, N. J., Salter, S. J., Harris, S. R., Mather, A. E., Hanage, W. P., Goldblatt, D., Nosten, F. H., Turner, C., Turner, P., Bentley, S. D., & Parkhill, J. (2014). Comprehensive Identification of Single Nucleotide Polymorphisms

336 Associated with Beta-lactam Resistance within Pneumococcal Mosaic Genes. *PLOS*
337 *Genetics*, *10*(8), e1004547. https://doi.org/10.1371/journal.pgen.1004547

338 Collins, C., & Didelot, X. (2018). A phylogenetic method to perform genome-wide
339 association studies in microbes that accounts for population structure and
340 recombination. *PLOS Computational Biology*, *14*(2), e1005958.
341 https://doi.org/10.1371/journal.pcbi.1005958

342 Coolen, J. P. M., den Drijver, E. P. M., Verweij, J. J., Schildkraut, J. A., Neveling, K.,
343 Melchers, W. J. G., Kolwijck, E., Wertheim, H. F. L., Kluytmans, J. A. J. W., &
344 Huynen, M. A. Y. 2021. (2021). Genome-wide analysis in Escherichia coli unravels a
345 high level of genetic homoplasy associated with cefotaxime resistance. *Microbial*
346 *Genomics*, *7*(4), 000556. https://doi.org/10.1099/mgen.0.000556

347 Croucher, N. J., Finkelstein, J. A., Pelton, S. I., Mitchell, P. K., Lee, G. M., Parkhill, J.,
348 Bentley, S. D., Hanage, W. P., & Lipsitch, M. (2013). Population genomics of post-
349 vaccine changes in pneumococcal epidemiology. *Nature Genetics*, *45*(6), 656–663.
350 https://doi.org/10.1038/ng.2625

351 Croucher, N. J., Finkelstein, J. A., Pelton, S. I., Parkhill, J., Bentley, S. D., Lipsitch, M., &
352 Hanage, W. P. (2015). Population genomic datasets describing the post-vaccine
353 evolutionary epidemiology of Streptococcus pneumoniae. *Scientific Data*, *2*(1),
354 Article 1. https://doi.org/10.1038/sdata.2015.58

355 Farhat, M. R., Freschi, L., Calderon, R., Ioerger, T., Snyder, M., Meehan, C. J., Jong, B. de,
356 Rigouts, L., Sloutsky, A., Kaur, D., Sunyaev, S., Soolingen, D. van, Shendure, J.,
357 Sacchettini, J., & Murray, M. (2019). GWAS for quantitative resistance phenotypes in

Mycobacterium tuberculosis reveals resistance genes and regulatory regions. *Nature Communications*, *10*(1), 2128. https://doi.org/10.1038/s41467-019-10110-6

Galardini, M., Clermont, O., Baron, A., Busby, B., Dion, S., Schubert, S., Beltrao, P., & Denamur, E. (2020). Major role of iron uptake systems in the intrinsic extra-intestinal virulence of the genus Escherichia revealed by a genome-wide association study. *PLOS Genetics*, *16*(10), e1009065. https://doi.org/10.1371/journal.pgen.1009065

Gori, A., Harrison, O. B., Mlia, E., Nishihara, Y., Chan, J. M., Msefula, J., Mallewa, M., Dube, Q., Swarthout, T. D., Nobbs, A. H., Maiden, M. C. J., French, N., & Heyderman, R. S. (2020). Pan-GWAS of Streptococcus agalactiae Highlights Lineage-Specific Genes Associated with Virulence and Niche Adaptation. *MBio*, *11*(3). https://doi.org/10.1128/mBio.00728-20

Grüning, B., Dale, R., Sjödin, A., Chapman, B. A., Rowe, J., Tomkins-Tinch, C. H., Valieris, R., & Köster, J. (2018). Bioconda: Sustainable and comprehensive software distribution for the life sciences. *Nature Methods*, *15*(7), Article 7. https://doi.org/10.1038/s41592-018-0046-7

Hakenbeck, R., Brückner, R., Denapaite, D., & Maurer, P. (2012). Molecular mechanisms of β-lactam resistance in Streptococcus pneumoniae. *Future Microbiology*, *7*(3), 395–410. https://doi.org/10.2217/fmb.12.2

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, *57*(1), 97–109. https://doi.org/10.1093/biomet/57.1.97

Ishikawa, S. A., Zhukova, A., Iwasaki, W., & Gascuel, O. (2019). A Fast Likelihood Method to Reconstruct and Visualize Ancestral Scenarios. *Molecular Biology and Evolution*, *36*(9), 2069–2085. https://doi.org/10.1093/molbev/msz131

Lees, J. A., Galardini, M., Bentley, S. D., Weiser, J. N., & Corander, J. (2018). pyseer: A comprehensive tool for microbial pangenome-wide association studies. *Bioinformatics*, *34*(24), 4310–4312. https://doi.org/10.1093/bioinformatics/bty539

Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D. S., Sander, C., Zecchina, R., Onuchic, J. N., Hwa, T., & Weigt, M. (2011). Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences*, *108*(49), E1293–E1301. https://doi.org/10.1073/pnas.1111471108

Orio, A. G. A., Piñas, G. E., Cortes, P. R., Cian, M. B., & Echenique, J. (2011). Compensatory Evolution of pbp Mutations Restores the Fitness Cost Imposed by **β**-Lactam Resistance in Streptococcus pneumoniae. *PLOS Pathogens*, *7*(2), e1002000. https://doi.org/10.1371/journal.ppat.1002000

Pupko, T., Pe, I., Shamir, R., & Graur, D. (2000). A Fast Algorithm for Joint Reconstruction of Ancestral Amino Acid Sequences. *Molecular Biology and Evolution*, *17*(6), 890–896. https://doi.org/10.1093/oxfordjournals.molbev.a026369

Rambaut, A., Drummond, A. J., Xie, D., Baele, G., & Suchard, M. A. (2018). Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. *Systematic Biology*, *67*(5), 901–904. https://doi.org/10.1093/sysbio/syy032

399 Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D. L., Darling, A., Höhna, S., Larget, B.,
400 Liu, L., Suchard, M. A., & Huelsenbeck, J. P. (2012). MrBayes 3.2: Efficient Bayesian
401 Phylogenetic Inference and Model Choice Across a Large Model Space. *Systematic*
402 *Biology*, *61*(3), 539–542. https://doi.org/10.1093/sysbio/sys029

403 Suchard, M. A., Lemey, P., Baele, G., Ayres, D. L., Drummond, A. J., & Rambaut, A. (2018).
404 Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus*
405 *Evolution*, *4*(1). https://doi.org/10.1093/ve/vey016

406 Tran, T. D.-H., Kwon, H.-Y., Kim, E.-H., Kim, K.-W., Briles, D. E., Pyo, S., & Rhee, D.-K.
407 (2011). Decrease in Penicillin Susceptibility Due to Heat Shock Protein ClpL in
408 Streptococcus pneumoniae. *Antimicrobial Agents and Chemotherapy*, *55*(6), 2714–
409 2728. https://doi.org/10.1128/AAC.01383-10

410