



**HAL**  
open science

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

Maxime Godfroid, Charles Coluzzi, Amaury Lambert, Philippe Glaser,  
Eduardo P C Rocha, Guillaume Achaz

► **To cite this version:**

Maxime Godfroid, Charles Coluzzi, Amaury Lambert, Philippe Glaser, Eduardo P C Rocha, et al..  
Evo-Scope: Fully automated assessment of correlated evolution on phylogenetic trees. *Methods in Ecology and Evolution*, 2024, 15 (2), pp.282 - 289. 10.1111/2041-210x.14190 . pasteur-04778723

**HAL Id: pasteur-04778723**

**<https://pasteur.hal.science/pasteur-04778723v1>**

Submitted on 12 Nov 2024

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.






L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial 4.0 International License

## APPLICATION

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

Maxime Godfroid<sup>1</sup>  | Charles Coluzzi<sup>2</sup>  | Amaury Lambert<sup>1,3</sup>  | Philippe Glaser<sup>4</sup>  |  
Eduardo P. C. Rocha<sup>2</sup>  | Guillaume Achaz<sup>1,5</sup>

<sup>1</sup>SMILE Group, Center for Interdisciplinary Research in Biology (CIRB), Collège de France, CNRS, INSERM, Université PSL, Paris, France

<sup>2</sup>Institut Pasteur, Université de Paris Cité, CNRS UMR 3525, Microbial Evolutionary Genomics, Paris, France

<sup>3</sup>Laboratoire de Probabilités, Statistique et Modélisation (LPSM), Sorbonne Université, CNRS UMR 8001, Université de Paris, Paris, France

<sup>4</sup>Institut Pasteur, Université de Paris Cité, CNRS UMR 6047, Ecology and Evolution of Antibiotics Resistance, Paris, France

<sup>5</sup>Éco-anthropologie, Muséum National d'Histoire Naturelle, CNRS UMR 7206, Université de Paris, Paris, France

## Correspondence

Maxime Godfroid

Email: [maxime1godfroid@gmail.com](mailto:maxime1godfroid@gmail.com)

## Present address

Maxime Godfroid, Sciensano, Transversal Activities in Applied Genomics (TAG), J, Wytmanstraat 14, Brussels, 1050, Belgium

## Funding information

INCEPTION, Grant/Award Number: PIA/ANR-16-CONV-0005; Equipe FRM (Fondation pour la Recherche Médicale), Grant/Award Number: EQU201903007835

**Handling Editor:** Steven Kembel

## 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 publicly available penicillin resistance data set of *Streptococcus pneumoniae* and characterize genetic mutations that correlate with antibiotic resistance.
5. 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/u/maximeg/w/evoscope>).

## KEYWORDS

correlated evolution, phylogenetics, workflow

## 1 | INTRODUCTION

The study of correlated evolution between biological traits illustrates the tight interdependence between processes in evolutionary biology. The constant development of new tools (Table S1) to understand these dependencies allows 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 measurement of these correlations is complicated by the phylogenetic relations (i.e. phylogenetic non-independence) between living entities, typically represented by a phylogenetic tree. As such, species belonging to the same taxon may share several residues or biological

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial](https://creativecommons.org/licenses/by-nc/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2023 The Authors. *Methods in Ecology and Evolution* published by John Wiley & Sons Ltd on behalf of British Ecological Society.

traits by vertical inheritance, a process which one usually wants to disentangle from convergence due to other processes, such as natural selection (Achaz & Duthiel, 2021). Therefore, the measurement of correlation between traits must either correct, or explicitly account for phylogenetic structure.

A typical application of the study of correlated evolution is the measure of the correlation between genetic factors and a given phenotype, that is, genetic mapping, generally performed through genome-wide association studies (GWAS). Multiple tools have been devoted to measure such correlation and use different methods to account for phylogenetic relationships (Table S1). While most methods add a correction term to account for the phylogeny, few explicitly rely on the topology of the phylogenetic tree (e.g. treeWAS Collins & Didelot, 2018). Particularly in bacteria, tree-aware GWAS has been critical to detect genetic variants involved in the evolution of antibiotic resistance (Farhat et al., 2019), virulence (Galardini et al., 2020) or niche adaptation (Gori et al., 2020).

We have previously developed two methods to measure correlated evolution on phylogenetic trees. First, *epics* is a heuristic method that calculates all possible orderings of mutational events affecting two traits occurring on a phylogenetic tree, and efficiently computes the probability to observe a number of co-counts equal or larger than the observed repartition (Behdenna et al., 2016). This method is closely related to the concentrated changes test (Maddison, 1990). The second tool, *epocs*, computes an estimate of the influence of the mutation of a first trait on the mutation rate of a second one by maximum likelihood (Behdenna et al., 2022). Notably, *epocs* assesses different possible interactions between the traits, such as induction or inhibition.

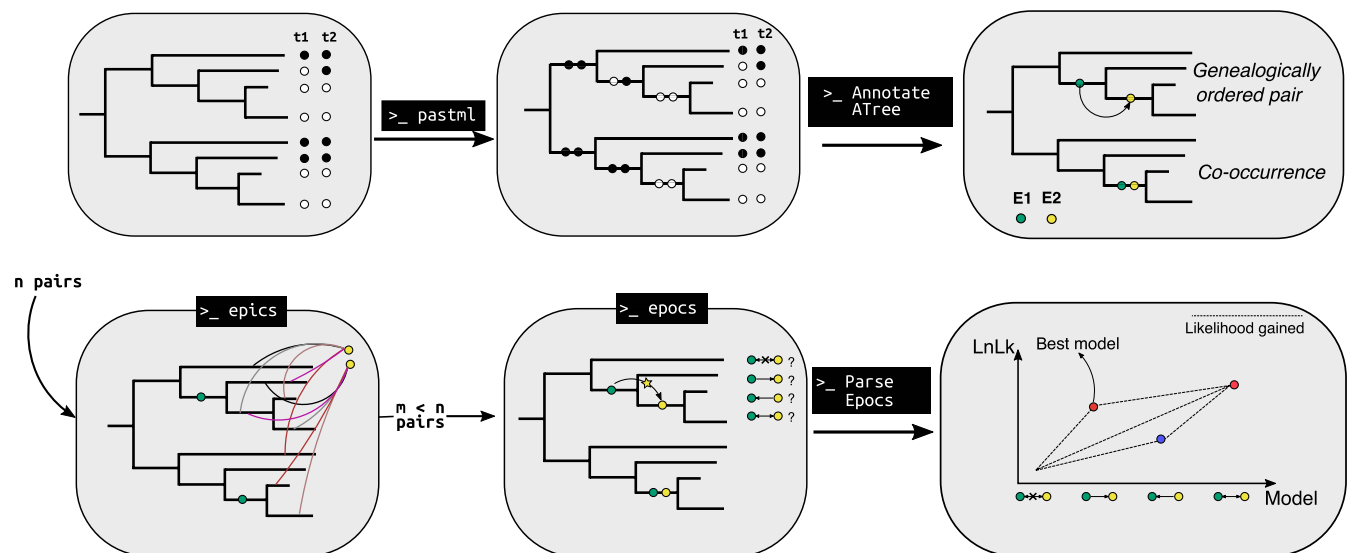
Here, we describe a novel software, *evo-scope* (EVOLUTIONary Study of Correlations of Occurrences on a Phylogeny), combining *epics* and *epocs*, to facilitate the detection of correlated evolution on phylogenetic trees for discrete traits. We further detail new features that we have added in the *evo-scope* 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.

## 2 | DESCRIPTION OF THE EVO-SCOPE PIPELINE

The details of the *epics* and *epocs* models are described in Behdenna et al. (2016, 2022). *Evo-scope* provides new versions of these tools with additional features. The *evo-scope* pipeline consists of five steps: ancestral character reconstruction (ACR), parsing and formatting, running *epics*, running *epocs* and summarizing the results (Figure 1).

### 2.1 | Ancestral character reconstruction

As initial input, a user provides a rooted phylogenetic tree and a file with the values of discrete traits at the tips. *Pastml* reconstructs the trait states on the internal branches of the tree using the JOINT model (Ishikawa et al., 2019; Pupko et al., 2000). Any other ACR tool or algorithm (e.g. Maximum A Posteriori in *pastml*) can be used as



**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 (trait 1 “t1” and trait 2 “t2”) along the tree and an R script formats the reconstruction to display the mutations on the branches. The operation is repeated for  $n$  pairs of traits. *Epics* is then run as a pre-filter to identify significantly associated pairs of evolutionary processes ( $m$ ) among the initial  $n$ . Next, *epocs* maximizes the likelihoods of four evolutionary models, one of independence and three of dependence, on the filtered pairs. An R script then summarizes the results and outputs the best model explaining the repartition of events on the tree. Finally, a figure can be generated for any pair considered to describe the gains of likelihood between nested models.

long as one single state (defined or undefined) is provided on each branch.

## 2.2 | ACR parsing and tree reformatting

*Evo-scope* then reconstructs the minimum set of mutational events compatible with the joint ML reconstruction of node states on the tree. A mutational event is defined as any change of a trait value between two consecutive nodes of the tree. At most one mutation per branch per trait is allowed: either a change occurred between the parent and daughter node or none.

## 2.3 | *Epics* as a pre-filter

*Evo-scope* then runs *epics* to discover (by default) pairs of traits that co-occur frequently on the branches of the trees (Behdenna et al., 2016). *Epics* calculates all possible orderings of mutational events affecting two traits occurring on a phylogenetic tree, and computes efficiently the exact probability to observe a number of co-counts equal or larger than the observed repartition. Other scenarios for the pairs of traits placement are available, such as genealogically ordered pairs. Of particular importance is that the calculation of the p-value depends on the reconstruction of the ancestral traits. At this step, *epics* is used as a pre-filter as it runs very fast and it only detects statistical association between the mutational events on the tree. Importantly, *epics* does not allow to predict an evolutionary scenario. As such, additional in-depth analysis is required to better describe the interactions between the mutational events. However, for very large datasets, *epics* might be the only option available as running time increase substantially for the maximum likelihood-based tool *epocs*. The output of this step is the list of every event pairs that are significantly associated on the tree for a selected p-value threshold.

## 2.4 | *Epocs* for the model discovery

For all significantly associated traits, *epocs* maximizes the likelihood of multiple scenarios of induction to determine which of the trait influenced the occurrence of the other one. The multiple scenarios can contain from two to eight parameters and describe the interrelationship between the occurrence of two events on the tree. The parameters are divided into natural occurrence rates ( $\mu_{(ij)}$ ,  $v_{(ij)}$ ) and excited occurrence rates ( $\mu_{(ij)}^*$ ,  $v_{(ij)}^*$ ) for each of the trait (Figure 3a). The ratio between the excited occurrence rates and the natural occurrence rates defines the induction (hereafter termed “lambda”).

In the current version, the model exploration is restricted to three models of correlation compared to a model of independence. The three models of correlation describe either an induction of the

first trait on the second one, vice-versa or a co-induction of both events. Scenarios of co-induction are modelled as independent of a third variable.

## 2.5 | Summarizing the results

Finally, *evo-scope* extracts the likelihoods and parameter values inferred for each of the models. To select the best model for each pair, the program performs likelihood ratio tests (LRT) between nested models (Behdenna et al., 2022). Next, it compares the calculated LRT value to a  $\chi^2$  distribution whose degrees of freedom is the difference in the number of parameters between models. Finally, based on significant LRTs, *evo-scope* selects the “best” model by a trade-off measure which maximizes the LRT value and minimizes the number of model parameters. *Evo-scope* then allows the user to select a pair and plot the likelihood gains between nested models.

## 3 | ADDITIONAL NOVEL FEATURES TO *EPICS* AND *EPOCS*

### 3.1 | Analysis on multiple independent trees

The user can provide a list of independent trees (non-overlapping taxa), to either *epics* or *epocs*, with the same type of events. This feature is particularly useful whenever the same character can be analysed across unrelated clades. In particular, *epics* and *epocs* will collect the information of all the trees and generate a global result by integrating the information across all trees. To do so, *epics* computes an aggregated p-value convoluted across all trees, while *epocs* transmits the likelihood calculation across all trees.

### 3.2 | Forest of trees

To account for phylogenetic uncertainty, the user can provide a forest of trees, such as generated by BEAST or MrBayes (Ronquist et al., 2012; Suchard et al., 2018), to either *epics* or *epocs*, and the results are ranked to extract the mode and 95% credibility intervals. In *epics*, the p-values are ranked by the tree likelihoods, while in *epocs* the results are ranked by the product of the *epocs* likelihood and the tree likelihood.

### 3.3 | MCMC exploration of likelihood surface

In the standard *epocs*, the likelihood is maximized considering all possible arrangements, with some restrictions on the number of co-occurrences due to computational complexity. In the MCMC case,

the program is allowed to explore freely any possible arrangement of the co-occurrences. This feature is implemented in a third executable *epocs\_mcmc*. In this module, the user supplies a tree with events to analyse and a standard Metropolis-Hastings algorithm will explore the parameter distribution across the number of replicates selected by the user (Hastings, 1970). Notable behaviours of the MCMC chain determine the relevancy of each parameter in the model. For example, a flat uniform distribution of the posterior is likely to reflect that this parameter is not pertinent in the model. In addition, the algorithm explores the order in the co-occurrences of the traits. *Epocs\_mcmc* outputs a tab-delimited file compatible with Tracer (Rambaut et al., 2018).

### 3.4 | Missing data

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

### 3.5 | GWAS-like procedure

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

## 4 | ANALYSIS OF AN ANTIBIOTIC RESISTANCE DATASET

As test data, we retrieved 603 genomes of *Streptococcus pneumoniae* (Chewapreecha et al., 2014; Croucher et al., 2013, 2015; Lees et al., 2018). In this dataset, the resistance to penicillin is known and encoded as a binary variable (R for resistant strains, S for susceptible strains). Also, 198,248 single-nucleotide polymorphisms (SNPs) have been called in the genomes. For the post-hoc MCMC analysis, we retrieved a significant association from *evo-scope* and ran *epocs\_mcmc* with a chain length of  $10^7$  steps and sampling every 50 steps on this pair.

## 5 | RESULTS

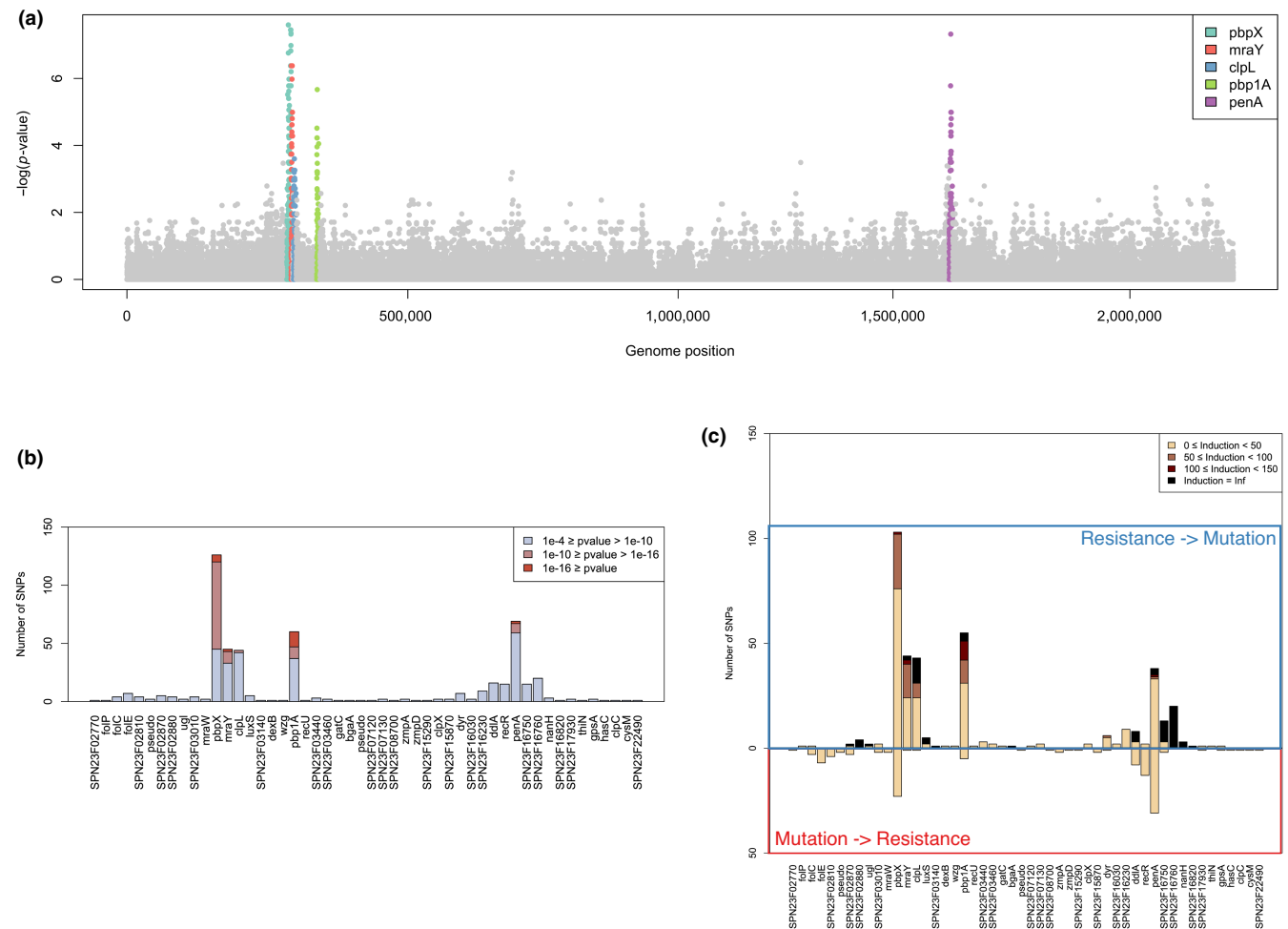
To illustrate the applicability of the present pipeline, we evaluated *evo-scope* by performing a GWAS-like analysis to a dataset of 603 genomes of *S. pneumoniae* (Chewapreecha et al., 2014; Croucher et al., 2013, 2015; Lees et al., 2018). As a positive

control, *evo-scope* detected genetic variants correlated to penicillin resistance (following Chewapreecha et al., 2014). In addition, *evo-scope* inferred the induction values for the significant associations to determine the strength and the direction of the correlation between the occurrence of antibiotic resistance and the SNPs. Of the 82,829 SNPs present in at least eight genomes, we report a total of 1629 SNPs (1.97%) associated with the resistance by *epics* (Figure 2a)—of which 543 (33.3% of the significant SNPs) were further inferred to be correlated with the resistance by *epocs*, with a  $p$ -value  $<10^{-4}$  (Figure 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; Figure 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 afterward (i.e. induction values  $>100$ —Figure 2c).

Additionally, we include a post-hoc MCMC analysis with *epocs\_mcmc* to demonstrate the usefulness of this addition in the *evo-scope* toolbox. For this purpose, we tested it on a significant association from *evo-scope*, which indicated an induction resistance  $\rightarrow$  mutation (Table 1 and Figure 3b).

Parameter naming follows the convention from (Behdenna et al., 2022). Non-starred rates represent natural occurrence rates, whereas starred rates represent excited occurrence rates (see also Figure 3a for a summary graph). ML estimates are the values of each parameter selected with the free-for-all model in *epocs* (model “8”, see Behdenna et al., 2022). Mean of the Bayesian estimates and mode of the posterior distribution are, respectively, the mean and mode values calculated by MCMC chain. The last column shows the 95% credible intervals of the posterior distribution.

After running *epocs\_mcmc*, we see that  $\mu_1$  is clearly defined, with a narrow distribution (Figure 3c). Interestingly, even though the preferred model by ML is the induction Resistance  $\rightarrow$  Mutation,  $\mu_1^*$  (demonstrating an induction mutation  $\rightarrow$  resistance) is still somewhat relevant to the model, but the confidence interval is quite wide (Figure 3e). Importantly,  $\mu_2$  is close to zero and  $\mu_2^*$  is clearly defined with a narrow distribution (Figure 3d,f). We conclude from these observations that the selected mutation in *pbpX* mostly occurs after the resistance, therefore indicating an induction resistance  $\rightarrow$  mutation. In addition, the order of the co-occurrences is inferred in the same direction when estimated by ML and furthermore, the  $v$  parameter posterior distributions are centred around the parameter values inferred by ML (Figure S1). It is important to mention that



**FIGURE 2** Application of *evo-scope* to a dataset of *Streptococcus pneumoniae*. (a) Manhattan plot of the *p*-values inferred by *epics*. In colour are shown genes of interest with multiple SNPs of low *p*-values. Notably, the three highest peaks correspond to *pbpX*, *pbp1A* and *penA*. (b) Barplot showing the *p*-value inferred by *epics*, grouped by gene IDs. (c) Barplots showing the inductions inferred by *epics* (i.e. the lambda in Behdenna et al., 2022). Top row represents the inductions resistance  $\rightarrow$  mutation, bottom row represents the other way round. For each statistical association between the resistance and one of the 543 significant single-nucleotide polymorphism, we retrieved the best model using likelihood ratio tests. Most notably, the three genes of interest *pbpX*, *penA* and *pbp1A* display low *p*-values and strong inductions.

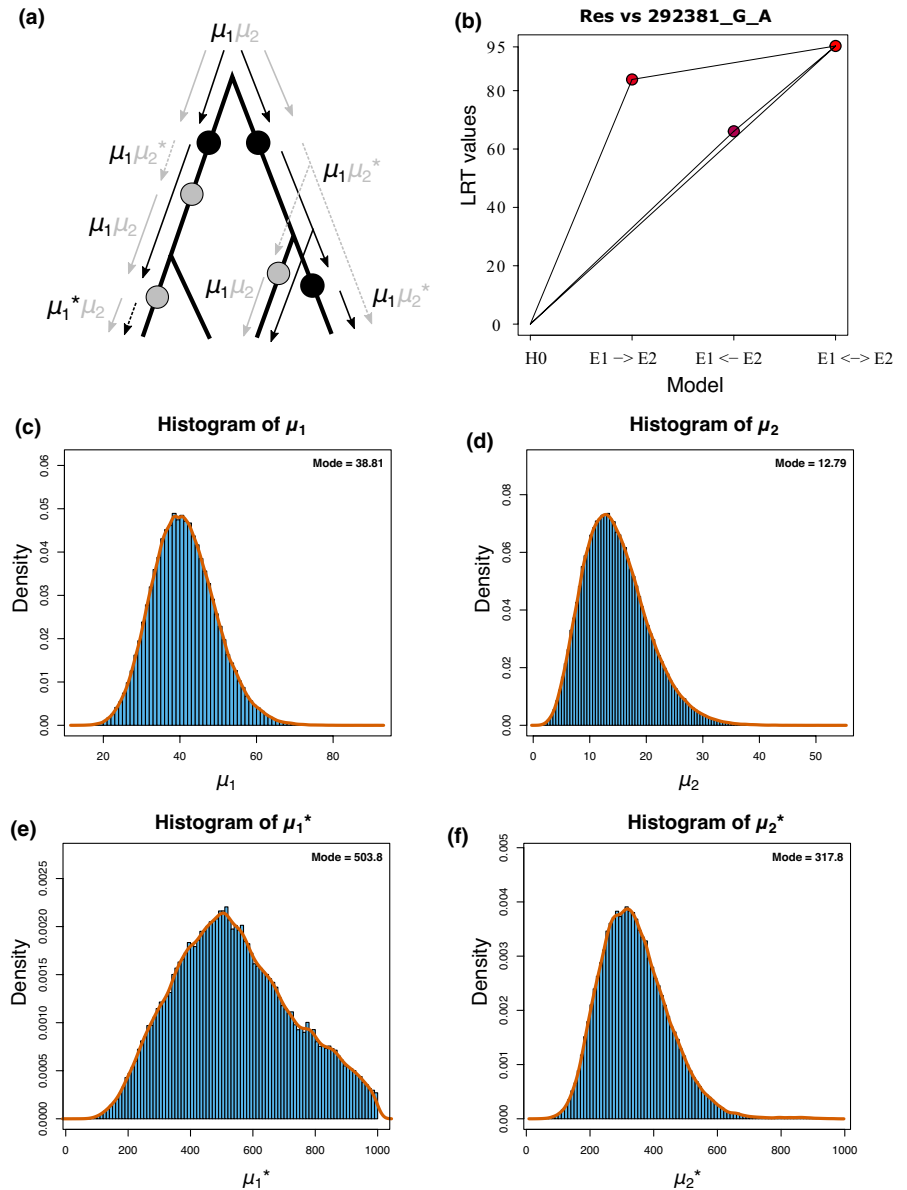
Parameter	ML estimates	Mean of the Bayesian estimates	Mode of the posterior distribution	95% credible interval
$\mu_1$	46.263	41.03	38.81	[25.39–57.51]
$\mu_1^*$	292.22	543.25	503.8	[214.42–934.32]
$\mu_2$	6.8197	14.66	12.79	[4.41–25.73]
$\mu_2^*$	377.08	338.24	317.8	[148.78–543.70]
$\nu_1$	41.474	46.23	41.76	[20.36–73.91]
$\nu_1^*$	?	475.22	68.12	[0.33–37.86]
$\nu_2$	0	4.67	0.5724	[1.4E-5–13.94]
$\nu_2^*$	0	294.33	41.59	[3.6E-5–41.23]

**TABLE 1** Parameter estimates of the Markov Chain Monte Carlo run for the selected pair.

even though the induction seem to indicate the occurrence of resistance first, correlation does not mean causation, and further investigation should be carried to determine the causal relationships. In

our previous papers, we extensively described the outcome of the algorithms using real-world examples and simulations (Behdenna et al., 2016, 2022).

**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. (2022). 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., 2022). 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 Markov Chain Monte Carlo chain. Scales are parameter-based in order to best display the shape of the parameter distribution.



## 6 | 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, where 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, mechanisms exist that compensate the cost of such antibiotic resistance, for example the evolution of a secondary mutation in the genome. For *S. pneumoniae*, bacteria possessing multiple mutations in the penicillin binding proteins exhibit a phenotype of resistance and compensation (Orio et al., 2011). Using *evo-scope*, we showed that many significantly

associated pairs of mutations concur with these observations. Furthermore, the patterns of mutation acquisitions and inductions could shed light on evolutionary trajectories of the acquisition of antibiotic resistance in *S. pneumoniae*.

In conclusion, we developed a fully automated pipeline, with minimal input required from the user. Our approach, by considering explicitly the phylogenetic component, allows to detect correlated evolution between discrete traits of any species by explicitly taking into account the underlying structure. We show that our pipeline can also be applied to tree-aware GWAS analyses in bacteria, expanding previous results in the field linking genetic variants to penicillin resistance in *S. pneumoniae*.

## AUTHOR CONTRIBUTIONS

Guillaume Achaz, Maxime Godfroid, Amaury Lambert, Eduardo P. C. Rocha, Philippe Glaser and Charles Coluzzi conceived the ideas and designed the methodology. Maxime Godfroid collected the data. Guillaume Achaz, Maxime Godfroid analysed the data. Maxime Godfroid led the writing of the manuscript. Guillaume Achaz, Amaury Lambert, Philippe Glaser and Eduardo P. C. Rocha obtained the funding for this project. All authors contributed critically to the drafts and gave final approval for publication.

## ACKNOWLEDGEMENTS

This work used the computation and storage services (TARS cluster) provided by the IT department at Institut Pasteur, Paris.

## FUNDING INFORMATION

This work was supported by INCEPTION (PIA/ANR-16-CONV-0005) and Equipe FRM (Fondation pour la Recherche Médicale) EQU201903007835.

## CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

## PEER REVIEW

The peer review history for this article is available at <https://www.webofscience.com/api/gateway/wos/peer-review/10.1111/2041-210X.14190>.

## DATA AVAILABILITY STATEMENT

The source code of *evo-scope*, *epics* and *epocs* are available on github under the GNU General Public Licence v3: <https://github.com/Maxime5G/EvoScope>. Pastml is available either on bioconda (Grüning et al., 2018) or on github: <https://github.com/evolbioinf/pastml>. We also provide a galaxy access to all tools and a “push-button” workflow at the Pasteur institute server: <https://galaxy.pasteur.fr/u/maximeg/w/evoscope>.

## ORCID

Maxime Godfroid <https://orcid.org/0000-0001-9037-4265>

Charles Coluzzi <https://orcid.org/0000-0003-2238-0836>

Amaury Lambert <https://orcid.org/0000-0002-7248-9955>

Philippe Glaser <https://orcid.org/0000-0003-4156-2782>

Eduardo P. C. Rocha <https://orcid.org/0000-0001-7704-822X>

## REFERENCES

- Achaz, G., & Duthel, 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. (2022). A minimal yet flexible likelihood framework to assess correlated evolution. *Systematic Biology*, 71(4), 823–838. <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 associated with beta-lactam resistance within pneumococcal mosaic genes. *PLoS Genetics*, 10(8), e1004547. <https://doi.org/10.1371/journal.pgen.1004547>
- Collins, C., & Didelot, X. (2018). A phylogenetic method to perform genome-wide association studies in microbes that accounts for population structure and recombination. *PLoS Computational Biology*, 14(2), e1005958. <https://doi.org/10.1371/journal.pcbi.1005958>
- Coolen, J. P. M., den Drijver, E. P. M., Verweij, J. J., Schildkraut, J. A., Neveling, K., Melchers, W. J. G., Kolwijck, E., Wertheim, H. F. L., Kluytmans, J. A. J. W., & Huynen, M. A. Y. (2021). Genome-wide analysis in *Escherichia coli* unravels a high level of genetic homoplasy associated with cefotaxime resistance. *Microbial Genomics*, 7(4), 000556. <https://doi.org/10.1099/mgen.0.000556>
- Croucher, N. J., Finkelstein, J. A., Pelton, S. I., Mitchell, P. K., Lee, G. M., Parkhill, J., Bentley, S. D., Hanage, W. P., & Lipsitch, M. (2013). Population genomics of post-vaccine changes in pneumococcal epidemiology. *Nature Genetics*, 45(6), 656–663. <https://doi.org/10.1038/ng.2625>
- Croucher, N. J., Finkelstein, J. A., Pelton, S. I., Parkhill, J., Bentley, S. D., Lipsitch, M., & Hanage, W. P. (2015). Population genomic datasets describing the post-vaccine evolutionary epidemiology of *Streptococcus pneumoniae*. *Scientific Data*, 2(1), Article 1. <https://doi.org/10.1038/sdata.2015.58>
- Farhat, M. R., Freschi, L., Calderon, R., loerger, T., Snyder, M., Meehan, C. J., de Jong, B., Rigouts, L., Sloutsky, A., Kaur, D., Sunyaev, S., van Sooling, D., Shendure, J., 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, A. C. J., & Parkhill, J. (2021). Genomic epidemiology of *Escherichia coli* O157:H7 in cattle. *Microbial Genomics*, 7(4), 000556. <https://doi.org/10.1099/mgen.0.000556>



- 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), 1–17. <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., Denapate, D., & Maurer, P. (2012). Molecular mechanisms of  $\beta$ -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>
- Maddison, W. P. (1990). A method for testing the correlated evolution of two binary characters: Are gains or losses concentrated on certain branches of a phylogenetic tree? *Evolution*, 44(3), 539–557. <https://doi.org/10.2307/2409434>
- 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 of the United States of America*, 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  $\beta$ -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>
- Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D. L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M. A., & Huelsenbeck, J. P. (2012). MrBayes 3.2: Efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology*, 61(3), 539–542. <https://doi.org/10.1093/sysbio/sys029>
- Suchard, M. A., Lemey, P., Baele, G., Ayres, D. L., Drummond, A. J., & Rambaut, A. (2018). Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evolution*, 4(1), 1–5. <https://doi.org/10.1093/ve/vey016>
- Tran, T. D.-H., Kwon, H.-Y., Kim, E.-H., Kim, K.-W., Briles, D. E., Pyo, S., & Rhee, D.-K. (2011). Decrease in penicillin susceptibility due to heat shock protein ClpL in *Streptococcus pneumoniae*. *Antimicrobial Agents and Chemotherapy*, 55(6), 2714–2728. <https://doi.org/10.1128/AAC.01383-10>

## SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

**Figure S1.** Likelihood ratios and output from `epocs_mcmc`, with the state-dependent parameters, 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. (2022). 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 eight models tested by `scoop`. Lines connect nested models, where plain lines show significant LRT and dotted lines show non-significant LRT. The black square shows the LRTs for models without state-dependence and the light grey square shows the LRTs for the models with state-dependence. 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) with state-dependence, (C–F) histogram of parameter values taken from the MCMC chain for the state-dependence parameters, (G) order of mutations in the co-occurrences in the branches of the tree. We observed that the MCMC algorithm selected preferentially the direction E1 → E2 (i.e. Resistance → Mutation) between 65.5% (co-occurrence 15) and 70.8% (co-occurrence 7) of the time.

**Table S1.** List of bioinformatics tools performing Genome-Wide Association Studies.

**How to cite this article:** Godfroid, M., Coluzzi, C., Lambert, A., Glaser, P., Rocha, E. P. C., & Achaz, G. (2024). Evo-Scope: Fully automated assessment of correlated evolution on phylogenetic trees. *Methods in Ecology and Evolution*, 15, 282–289. <https://doi.org/10.1111/2041-210X.14190>