Received: 3 February 2022 | Accepted: 10 June 2022

DOI: 10.1111/2041-210X.13980

Methods in Ecology and Evolution BRITISH ECOLOGICAL SOCIETY

APPLICATION

Polymorphism-aware estimation of species trees and evolutionary forces from genomic sequences with RevBayes

Rui Borges[1] | Bastien Boussau[2] | Sebastian Höhna[3,4] | Ricardo J. Pereira[5] | Carolin Kosiol[1,6]

[1]Institut für Populationsgenetik, Vetmeduni Vienna, Wien, Austria; [2]Université de Lyon, Université Claude Bernard Lyon 1, Villeurbanne, France; [3]GeoBio-Center, Ludwig-Maximilians-Universität München, Munich, Germany; [4]Department of Earth and Environmental Sciences, Paleontology & Geobiology, Ludwig-Maximilians-Universität München, Munich, Germany; [5]Division of Evolutionary Biology, Department of Biology II, Ludwig-Maximilians-Universität München, Martinsried, Germany and [6]Centre for Biological Diversity, University of St Andrews, St Andrews, UK

Correspondence
Carolin Kosiol
Email: ck202@st-andrews.ac.uk

Funding information
Austrian Science Fund, Grant/Award Number: P34524-B; Biotechnology and Biological Sciences Research Council, Grant/Award Number: BB/W000768/1; Deutsche Forschungsgemeinschaft, Grant/Award Number: HO 6201/1-1; Vienna Science and Technology Fund, Grant/Award Number: MA016-061

Handling Editor: Pablo Duchen

Abstract

1. The availability of population genomic data through new sequencing technologies gives unprecedented opportunities for estimating important evolutionary forces such as genetic drift, selection and mutation biases across organisms. Yet, analytical methods that can handle polymorphisms jointly with sequence divergence across species are rare and not easily accessible to empiricists.
2. We implemented polymorphism-aware phylogenetic models (PoMos), an alternative approach for species tree estimation, in the Bayesian phylogenetic software RevBayes. PoMos naturally account for incomplete lineage sorting, which is known to cause difficulties for phylogenetic inference in species radiations, and scale well with genome-wide data. Simultaneously, PoMos can estimate mutation and selection biases.
3. We have applied our methods to resolve the complex phylogenetic relationships of a young radiation of *Chorthippus* grasshoppers, based on coding sequences. In addition to establishing a well-supported species tree, we found a mutation bias favouring AT alleles and selection bias promoting the fixation of GC alleles, the latter consistent with GC-biased gene conversion. The selection bias is two orders of magnitude lower than genetic drift, validating the critical role of nearly neutral evolutionary processes in species radiation.
4. PoMos offer a wide range of models to reconstruct phylogenies and can be easily combined with existing models in RevBayes—for example, relaxed clock and divergence time estimation—offering new insights into the evolutionary processes underlying molecular evolution and, ultimately, species diversification.

KEYWORDS
Bayesian inference, grasshoppers, mutation bias, polymorphism-aware phylogenetic models, RevBayes, selection, species tree

1 | INTRODUCTION

The recent development of sequencing technologies has made it possible to obtain large amounts of genomic data at a low cost, across model and non-model organisms. We now have an unprecedented opportunity to study macroevolutionary and microevolutionary processes at temporal scales where the distinction between species and population is challenging to establish (Leaché & Oaks, 2017). To resolve the intricate phylogenetic relationships between closely related taxa, we need models of species divergence that are able to assess the significance of different evolutionary forces in shaping species' diversity patterns. The polymorphism-aware phylogenetic models (PoMos) are examples of such models (De Maio et al., 2013, 2015). PoMos describe the evolutionary relationship between taxa (i.e. populations or species), in which the frequency of alleles change over time, conditioned on evolutionary forces such as mutational bias, genetic drift and selection (Borges et al., 2019). Thus, PoMos have the ability to leverage population genomic data to estimate parameters of molecular evolution and to jointly provide more accurate estimates of the species tree.

PoMos can be seen as an extension of the traditional phylogenetic substitution models (e.g. JC, HKY and GTR; Hasegawa et al., 1985; Jukes & Cantor, 1969; Tavaré, 1986), in the sense that PoMos additionally consider the existence of polymorphisms. Traditional phylogenetic substitution models assume that a species can be represented using a single genotype and that all differences occur only between (but not within) species. By describing allele changes over time, PoMos operate in a state-space that includes both fixed and polymorphic states. The possibility to model the existence of ancestral polymorphisms is of particular importance as PoMos naturally accounts for incomplete lineage sorting—that is, the persistence of ancestral polymorphisms during speciation—which is known to be a prime cause of phylogenomic conflict in species tree inference (Maddison & Knowles, 2006; Szöllősi et al., 2015).

PoMos are similar to the multispecies coalescent approach in the sense that both aim at estimating the species tree. However, PoMos describe allele trajectories instead of genealogies, therefore integrating over all possible genealogical histories to directly estimate the species tree (similar to SNAPP; Bryant et al., 2012) and bypassing the computational burden of estimating genealogies. This burden continues to be a major constraint for the coalescent-based species tree methods, as the space of unknown genealogies is large and difficult to explore (Flouri et al., 2020; Ogilvie et al., 2016; Rannala & Yang, 2017). The integration over genealogies in PoMos allows them to scale well with genomic data sampled from many populations and individuals (Schrempf et al., 2016). The neutral PoMos were previously integrated into the maximum likelihood framework in IQ-Tree, primarily for species tree inference (Schrempf et al., 2019). However, an implementation including more recent PoMo models (virtual PoMos and models with selection; Borges et al., 2019; Borges, Boussau, Szöllősi, et al., 2022) and allowing their combination with sophisticated models of molecular dating, phylogeography, trait evolution, branch-heterogeneous processes, all typically used in a Bayesian framework, is still missing.

Furthermore, PoMos permit disentangling the contribution of the various evolutionary forces to the species divergence process that are often confounded in existing methods. For example, selection favouring the fixation of specific alleles can be particularly difficult to model under the coalescent approaches but is quite straightforward under PoMos (Borges et al., 2019). PoMos are also able to estimate possible mutation bias (Borges, Boussau, Szöllősi, et al., 2022), which condition the genetic variability upon which selection can act. As such, PoMos can be helpful in testing the significance of biologically relevant processes responsible for molecular evolution, even in the absence of well-annotated reference genomes, as demonstrated by our application example on grasshopper species. Lastly, PoMos have been shown to be statistically sound; the maximum a posteriori tree is a consistent estimator of the species tree (Borges & Kosiol, 2020).

For all these reasons, PoMos constitute an attractive framework for estimating the species trees and test the role of the different evolutionary forces responsible for molecular evolution across the genome. Here, we implemented these models within the widely used Bayesian programming tool RevBayes (Höhna et al., 2016) and demonstrate their application with population genomic data from a recent species radiation (Nolen et al., 2020).

2 | THE MODEL

PoMos model the evolution of a population of N haploid individuals and K alleles through time, in which changes in the allele content and frequency are both possible. These changes are mediated by evolutionary forces, such as genetic drift, mutation and selection. The PoMo state-space includes fixed (or boundary) states $\{Na_i\}$, in which all N haploid individuals have the same allele $a_i \in \{1, \ldots, K\}$, and polymorphic states $\{na_i, (N-n)a_j\}$, in which two alleles a_i and a_j are present in the population with absolute frequencies n and $N-n$, respectively.

Mutations occur with rate $\mu_{a_i a_j}$ and govern the allele content, only arising in the fixed states:

$$q_{\{Nai\}\to\{(N-1)a_i,1a_j\}} = N\mu_{a_i a_j}. \quad (1)$$

The assumption that mutations can only arise in the fixed states, corresponds to assuming that mutation rates are comparably low, and by the time of a new mutation a fixation of the previous mutation at the same site has already taken place. Low mutations rates are indeed verified for the majority of multicellular eukaryotes (Lynch et al., 2016). Additionally, a reversible mutation model is usually considered. In this case, we break the mutations into a base composition π and exchangeability parameter ρ (i.e. $\mu_{a_i a_j} = \rho_{a_i a_j}\pi_{a_j}$), similar to the definition of substitution rates in the general time-reversible model by Tavaré (1986). Assuming reversible mutations has no biological basis, however, this assumption simplifies obtaining formal quantities such as the stationary distribution, while still allowing to model mutation biases.

Genetic drift is modelled according to the Moran model, in which one individual is chosen to die, and one individual is chosen to reproduce at each time step. Selection acts to favour or disfavour alleles by differentiated fitness: $\phi_{a_i} = 1 + \sigma_{a_i}$, where σ_{a_i} is the selection coefficient of the a_i allele. Together, genetic drift and selection govern the allele frequency changes:

$$q_{\{na_i,(N-n)a_j\}\to\{(n+1)a_i,(N-n-1)a_j\}} = \frac{n(N-n)}{n\phi_{a_i} + (N-n)\phi_{a_j}}\phi_{a_i}. \quad (2)$$

As standard substitution models, PoMos are continuous-time Markov models and are fully characterized by their rate matrices. PoMos express time in Moran generations and the rates in Equations (1) and (2) together define the PoMos rate matrices (Borges, Boussau, Szöllősi, et al., 2022).

3 | PoMos IMPLEMENTED IN RevBayes

Previously, PoMos were implemented in a Maximum Likelihood framework in HyPhy (De Maio et al., 2013) and IQ-Tree (Schrempf et al., 2019). Here, we implemented PoMos in a Bayesian framework in the phylogenetic software package RevBayes (Höhna et al., 2016; Figure 1). The main advantages of our new implementation are extensions of PoMos (described below) and the existing model toolbox of RevBayes which cannot be performed in the Maximum Likelihood framework. RevBayes includes several highly maintained and extensively used routines for phylogeny estimation and hypothesis testing (Höhna et al., 2014, 2016, 2018, 2021). In addition, RevBayes can perform a range of inferences—for example, phylogeny reconstruction, inference of selection, molecular dating, character evolution, ancestral state reconstruction—that can be easily married with PoMos, allowing the development of newly devised models for varied evolutionary applications.

PoMo rate matrices include selection by modelling allelic fitness (argument phi). Selection can also be helpful to model nucleotide usage biases that are not caused by selection but behave like it (Nagylaki, 1983). This is the case of GC-biased gene conversion, a mutational bias present in many living organisms (Galtier et al., 2009), that prefers GC alleles over AT during recombination. Neutral dynamics can simply be defined by assigning the fitness coefficients vector to 1.

PoMo functions in RevBayes permit modelling mutational biases. Mutations can be assumed non-reversible (via the argument mu) or reversible (functions including the prefix Reversible). Reversible mutations assume that the mutation rate between any two alleles is the product of an exchangeability and a base composition term (arguments rho and pi). Another advantage of having PoMos into a Bayesian framework is that empirically measured mutation rates, population sizes or nucleotide usage biases can be accounted for during inference via informative priors. Additionally, we implemented virtual PoMos, which were recently introduced (Borges, Boussau, Szöllősi, et al., 2022): fnReversiblePoMoTwo4N and fnReversiblePoMoThree4N. Virtual PoMos mimic a population dynamic

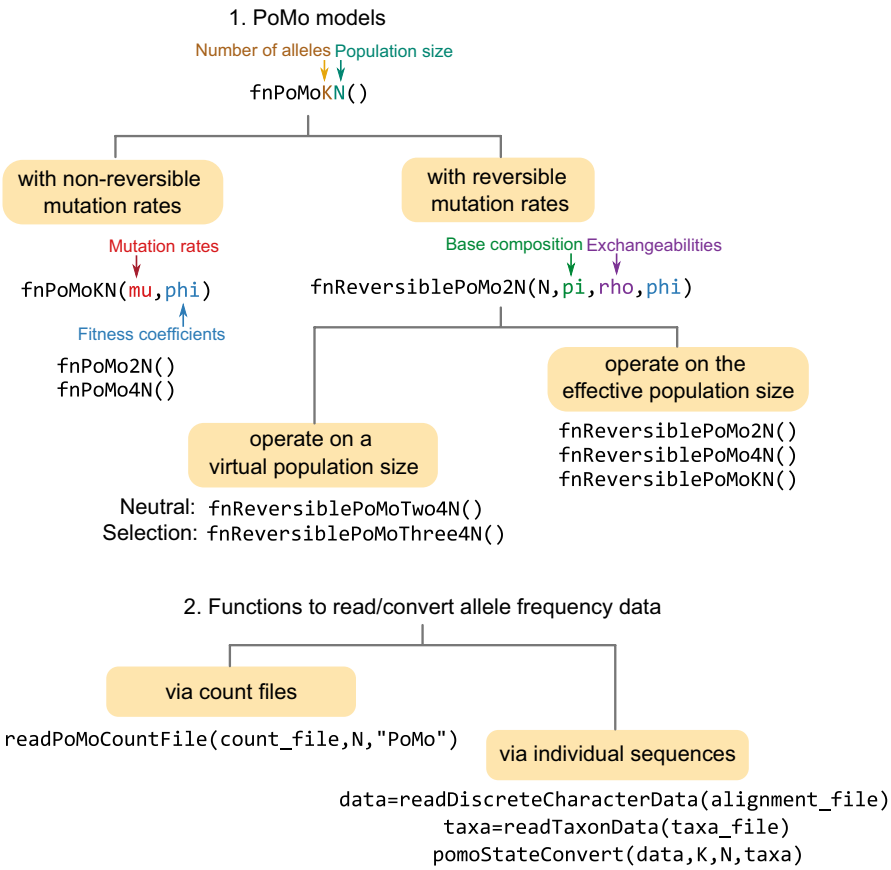

FIGURE 1 An overview of the PoMos functions implemented in RevBayes (Höhna et al., 2016).

that unfolds in the effective population using a much smaller virtual population size, making it computationally more efficient. While inferences are performed in the virtual population, the mutation rates and selection coefficients are rescaled to real dynamic using theoretically obtained scaling laws (Borges, Boussau, Szöllősi, et al., 2022).

Apart from the standard PoMos, which usually operate with the four nucleotide bases, we have also added PoMo rate matrices to be used in genetic systems that might have any number of variants (i.e. fnPoMoKN). In addition, we included specific models for the biallelic case, as this is typically used in population genetic applications (e.g. fnReversiblePoMo2N).

Furthermore, we have included functions to read allele frequency data and correct for sampling biases. Genetic diversity is usually undersampled because sampled fixed sites might not necessarily be fixed in the original population. To correct for such biases, we implemented a variation of the weighted method proposed by Schrempf et al. (2016), which employs a binomial sampling at the tips of the tree. This method is integrated into the functions that read allele frequency data, either from count files or from FASTA alignments (i.e. convertCountFileToNaturalNumbers and convertFastaFileToNaturalNumbers, respectively).

4 | IMPLEMENTATION, AVAILABILITY AND VALIDATION

The RevBayes source code is freely available at https://github.com/revbayes/revbayes. A detailed tutorial about Bayesian phylogenetic inference with PoMos is available at https://revbayes.github.io/tutorials/pomos/.

As critical part of a robust Bayesian workflow, we validated our implementations of PoMos in RevBayes using a simulation-based calibration approach (Talts et al., 2018). This method assesses the soundness of a posterior sampler by testing the expectation that if data are generated according to a model, then the Bayesian posterior inferences with respect to that same model are calibrated by construction. In practical terms, if one draws S parameter values from their prior distributions and subsequently simulates S datasets using these parameter values, the inferred α% credible intervals will contain the true parameter value in $S\alpha$ of times. This simulation-based calibration approach has advantage over other validation techniques because the correct coverage of credible intervals is only achieved if the likelihood function, the Markov chain Monte Carlo (MCMC) sampler and the simulation function are implemented correctly.

We simulated 1000 alignments of four species and 1000 sites, considering a population dynamic including three virtual individuals, four alleles (A, C, G and T) and general mutation (nine free parameters in total: three base composition parameters π and six exchange abilities ρ) and selection (three fitness coefficients; ϕ_A was set to 1) schemes. We calculated the coverage probabilities for each parameter and compared them with the credibility interval size. Our analyses show a 1:1 correlation (i.e. identity line) between the coverage and credible interval size, indicating that our implementations are statistically and computationally sound (Figure 2). The simulated datasets and the RevBayes scripts and output files used in simulation-based calibration analyses are all provided as additional material (see Data availability statement for further details).

5 | APPLICATION TO GRASSHOPPERS TRANSCRIPTOME DATA

We studied the recent radiation of five grasshopper species using transcriptome data (Nolen et al., 2020) and PoMos in RevBayes. This dataset includes nine grasshopper populations with 5–15 sampled individuals per population (Figure 3). FASTA alignments from 1895 protein-coding genes were generated by calling the most common base at each position so that each individual grasshopper is represented by one allele. In the final alignment, we have only included the third codon position and sites for which at least four individuals were observed per population. A total of 2,744,646 sites were obtained, but for computational reasons, only a sample of 1 million randomly selected sites was used for the phylogenetic inferences. This represents a manageable number of sites that are usually easily reached in standard empirical studies.

Phylogenetic inferences were carried out with the virtual PoMoThree model, which assumes a reversible mutation scheme. As we aimed to test the dynamic where AT alleles (or weak alleles, W) are preferred by mutational bias, and GC alleles (or strong alleles, S) are favoured by GC-biased gene conversion, we defined the allele composition, exchangeabilities and fitness coefficients in the following manner. The allele composition is generally defined as $\pi = [\pi_A, \pi_C, \pi_G, \pi_T]$, which in terms of strong and weak alleles becomes $\pi = [\pi_W, \pi_S, \pi_S, \pi_W]$. The allele exchangeabilities are generally defined as $\rho = [\rho_{AC}, \rho_{AG}, \rho_{AT}, \rho_{CG}, \rho_{CT}, \rho_{GT}]$, but we have only considered two exchangeability classes, one between weak and strong alleles, and another one between alleles of the same type: that is, $\rho = [\rho_{WS}, \rho_{WS}, \rho_{WW}, \rho_{SS}, \rho_{WS}, \rho_{WS}]$, where $\rho_{WW} = \rho_{SS}$. The fitness coefficients are generally defined as $\phi = [\phi_A, \phi_C, \phi_G, \phi_T]$, but because we are only interested in modelling GC-bias, we simplified it to $\phi = [1.0, \phi_S, \phi_S, 1.0]$, where $\sigma_S = \phi_S - 1$ is the selection coefficient favouring the strong alleles.

In total, four parameters were estimated: π_W, ρ_{WS}, $\rho_{SS} = \rho_{WW}$ and ϕ_S. We set a Beta prior on π_W and an exponential prior on the exchangeabilities ρ_{WS} and ρ_{SS}. A reversible jump mixture was set on ϕ_S, including a point mass at 1.0 and a gamma prior for all the positive values, both with probability 0.5. This means that the GC-bias rate can take on the constant value 0.0 (suggesting neutrality) or be drawn from the base distribution gamma (indicating selection). A special reversible-jump MCMC algorithm (for details, see Freyman & Höhna, 2018) is then used to infer whether GC-bias was active or not. We assumed a strict molecular clock rate drawn from an exponential prior with an arbitrary root age that we fixed to 1.0. A uniform time tree prior was set for the phylogeny. We ran the MCMC with two chains for 300,000 generations. The Bayesian inferences in RevBayes with the grasshopper's sequence alignment

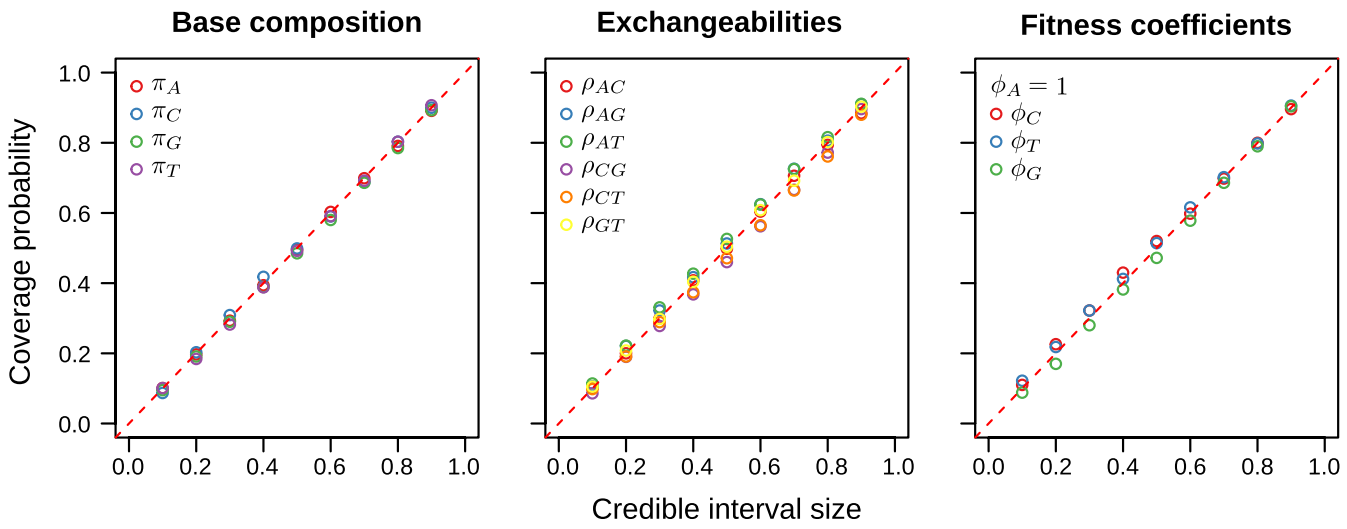

FIGURE 2 Simulation-based calibration of PoMos in RevBayes (Höhna et al., 2016). the plots were obtained using 1000 simulated datasets of multiple sequence alignments of 4 species and 1000 sites. We simulated a population dynamic including three virtual individuals, four alleles (*A*, *C*, *G* and *T*) and general mutation (nine free parameters in total: Three πs and six ρs) and selection (three fitness coefficients; ϕ_A was set to 1) schemes. The CI probability is the size of the credible interval. The coverage probability corresponds to the proportion of credible intervals that included the true parameter value. The expectation from mathematical theory is that coverage and credible interval size (CI probability) show a 1:1 correlation.

Chorthippus brunneus *Chorthippus mollis*

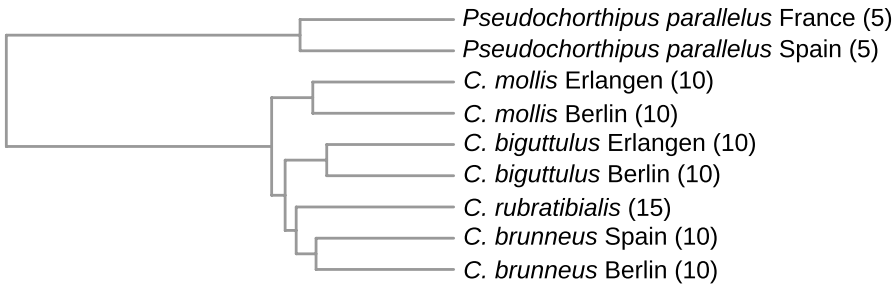

FIGURE 3 Grasshopper's evolutionary history. Application example of the Bayesian PoMos in RevBayes to transcriptome data from nine grasshopper populations of the genus *Chorthippus sp.* (Nolen et al., 2020). Phylogenetic inferences were conducted using one million randomly sampled sites from the third codon position of 1895 protein-coding genes together with the virtual PoMo Three model. The maximum a posteriori tree is depicted; all the branches had a posterior clade probability of 1.00 (not shown). The number of sampled individuals per population are indicated inside parenthesis. Photos of *Chorthippus brunneus* (credits to Jörg Hempel) and *Chorthippus mollis* (credits to G.-U. Tolkiehn) grasshoppers taken from Wikipedia under the creative commons attribution-share alike 3.0 licence.

(one million sites and nine populations) took approximately 29 hr in an iMac desktop with a 3.7 GHz Intel Core i5 processor and 32GB memory using four parallel processes. Convergence was assessed via the effective sample size parameter, which was higher than 500 for the continuous parameters and tree splits (Fabreti & Höhna, 2022). The multiple sequence alignment, the RevBayes script and output files used in the grasshopper's phylogenetic analyses are provided as additional material (see Data availability statement for further details).

The posterior distributions of the mutation rate parameters and selection coefficient, despite all showing clear peaks (Figure 4), cannot be directly interpreted as they are relative to a virtual population size of three individuals. To obtain sensible estimates of these evolutionary forces, we need to scale them in terms of the grasshoppers'

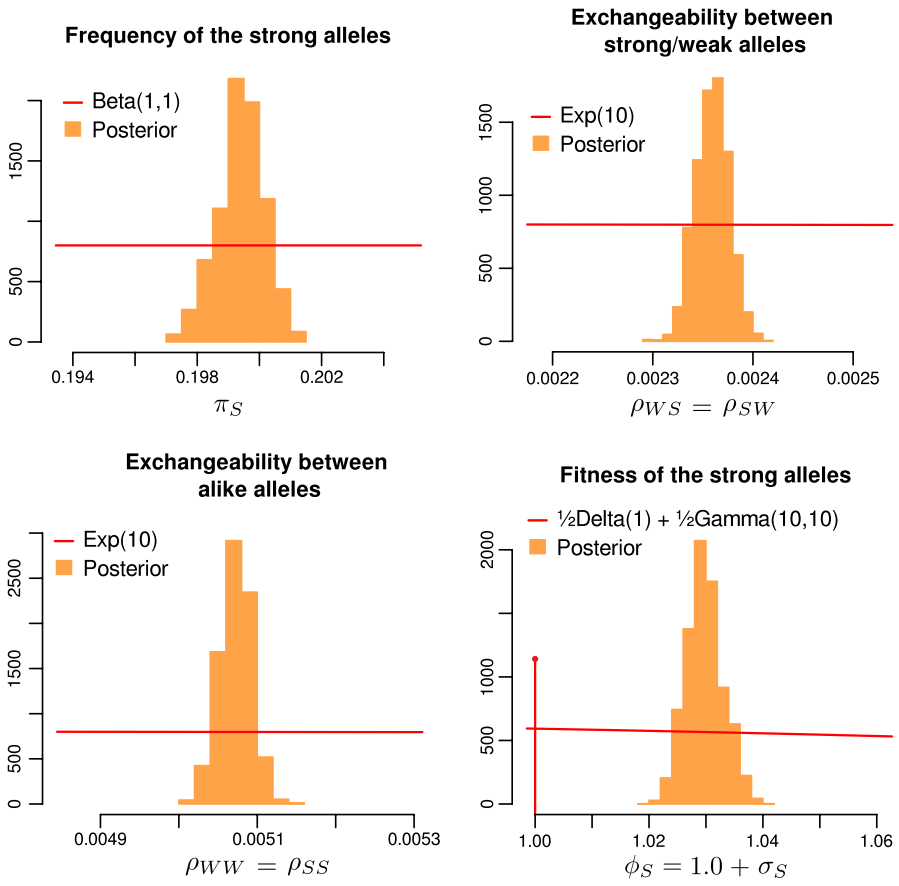

FIGURE 4 Posterior and prior distributions of the mutation biases and selection parameters at the grasshoppers' third codon position. Delta indicates the mass point distribution. The prior densities are rescaled only to facilitating comparisons between the prior and posterior distributions and should not be read on the *y*-axis.

TABLE 1 Posterior mean estimates and 95% credible intervals of the molecular dynamic parameters between weak (AT) and strong (GC) alleles at the grasshoppers' third codon position. Note that $2\pi_S + 2\pi_W = 1$, $\rho_{SS} = \rho_{WW}$ and $\rho_{SW} = \rho_{WS}$. The estimates of the mutations rates and selection coefficients were rescaled for a population of 850,000 individuals based on the theoretical results of Borges, Boussau, Szöllősi, et al. (2022)

Evolutionary force	Parameter	Posterior mean	95% credible interval
Mutation bias towards strong alleles	$\mu_{WS} = \rho_{WS}\pi_S$	2.63×10^{-10}	$[2.60, 2.67]\times10^{-10}$
Mutation bias towards weak alleles	$\mu_{SW} = \rho_{WS}\pi_W$	5.67×10^{-10}	$[5.62, 5.71]\times10^{-10}$
Mutation bias between strong alleles	$\mu_{SS} = \rho_{SS}\pi_S$	3.76×10^{-10}	$[3.73, 3.79]\times10^{-10}$
Mutation bias between weak alleles	$\mu_{WW} = \rho_{SS}\pi_W$	1.75×10^{-10}	$[1.72, 1.77]\times10^{-10}$
Selection favouring GC alleles	$\sigma_S = \phi_S - 1$	6.89×10^{-8}	$[5.71, 8.19]\times10^{-8}$

effective population size. While empirical estimates of the genetic diversity (Watterson's $\theta = 0.012$, according to Nolen et al., 2020) are available for these species, we lack direct measurements of the mutation rate that would allow us to estimate their effective population size via $\theta = 4N\mu$. We have instead used the fruit fly mutation rate (3.5×10^{-9}, according to Keightley et al., 2014) and obtained an effective population size of approximately 850,000 individuals. We note that this estimate must be considered with care, as mutation rates can greatly vary across phylogenetic scales.

Using the estimated population size and the scaling laws in Borges, Boussau, Szöllősi, et al. (2022), we calculated the relative mutation rates and GC-bias for the grasshoppers' transcriptomic data (Table 1). Despite the four mutation types being around the same order of magnitude, we found evidence for mutational bias, with mutations from strong to weak alleles being more frequent than mutations from weak to strong alleles: $\mu_{SW} = 5.67\times10^{-10}$ and $\mu_{WS} = 3.76\times10^{-10}$. Mutations between alleles of the same type are the least frequent: $\mu_{WW} = 2.64\times10^{-10}$ and $\mu_{SS} = 1.75\times10^{-10}$. The GC-bias rate is two orders of magnitude higher than the mutation rates but two orders of magnitude lower than the reciprocal of the population size: $\sigma_S = 6.89\times10^{-8}$ and $N\sigma_S = 0.059$. This indicates that GC-selection (or possibly GC-biased gene conversion) operates in a nearly neutral range. Nonetheless, the posterior of ϕ_S never included the point mass 1.0 (log Bayes Factor = 3.68), indicating that selection in favour of GC alleles at the third codon position is very significant despite acting in the nearly neutral range.

Regarding the topology, our phylogenetic inferences show very well-supported branches (posterior probabilities all equal to 1.00;

Figure 3), even in clades that previous methods were unable to resolve using more sites (Nolen et al., 2020). Notably, the relative divergence among taxa is significantly larger than previously reported (Nolen et al., 2020), likely because with PoMos we can account for mutation and nucleotide usage bias, which lead other methods to underestimate genetic distances (Borges, Boussau, Szöllősi, et al., 2022).

Overall, these results suggest that a dynamic including mutational bias favouring AT alleles and selection bias promoting the fixation of GC alleles governs the codon usage patterns of the *Chorthippus* grasshoppers protein-coding sequences. Despite being reported here for the first time for the case of grasshoppers, these patterns have been observed in other taxa (e.g. Sueoka & Kawanishi, 2000), suggesting that gene GC-biased gene conversion is a general molecular process throughout the tree of life.

6 | CONCLUSION

We implemented PoMos into the widely used phylogenetic Bayesian software RevBayes. Our implementation allows researchers to use the latest species tree inference models and combine state-of-the-art methods in phylogenetics to build robust inferences or hypotheses testing. PoMos are able to perform accurate species tree inference while jointly estimating parameters of the molecular evolutionary dynamic. Together, PoMos offer more complete species tree inference methods that can be useful to characterize both species and molecular evolution. We anticipate that our methods will be particularly interesting to researchers studying genomic datasets including multiple species, populations and individuals, which are increasingly being produced but still lack appropriate analytical methods. Furthermore, RevBayes is open-source and multiplatform; thus, it can easily be used on a variety of systems and settings for both academic research and teaching.

AUTHOR CONTRIBUTIONS

Rui Borges and Carolin Kosiol conceived the project; Rui Borges, Bastien Boussau and Sebastian Höhna implemented the models in RevBayes; Rui Borges and Ricardo J. Pereira performed the application example; Rui Borges and Carolin Kosiol led the writing of the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

ACKNOWLEDGEMENTS

This work was funded by the Vienna Science and Technology Fund (WWTF) [MA16-061] and partially supported by the Austrian Science Fund (FWF) [P34524-B] and Biotechnology and Biological Sciences Research Council (BBSRC) [BB/W000768/1]. This work was supported by the Deutsche Forschungsgemeinschaft (DFG) Emmy Noether-Program (Award HO 6201/1-1 to S.H.).

CONFLICT OF INTEREST

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

The data used and produced by the simulation-based calibration analyses of PoMos in RevBayes and the phylogenetic analyses conducted on the *Chorthippus* sp. grasshoppers (Nolen et al., 2020) are available from Zenodo https://doi.org/10.5281/zenodo.6592395 (Borges, Boussau, Höhna, et al., 2022).

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1111/2041-210X.13980.

ORCID

Rui Borges https://orcid.org/0000-0002-5905-3778
Bastien Boussau https://orcid.org/0000-0003-0776-4460
Sebastian Höhna https://orcid.org/0000-0001-6519-6292
Ricardo J. Pereira https://orcid.org/0000-0002-8076-4822
Carolin Kosiol https://orcid.org/0000-0002-3219-6648

REFERENCES

Borges, R., Boussau, B., Höhna, S., Pereira, R., and Kosiol, C. (2022). Supplemental material: Polymorphism-aware estimation of species trees and evolutionary forces from genomic sequences with RevBayes. https://doi.org/10.5281/zenodo.6592395

Borges, R., Boussau, B., Szöllősi, G. J., & Kosiol, C. (2022). Nucleotide usage biases distort inferences of the species tree. *Genome biology and evolution*, *14*(1), evab290.

Borges, R., & Kosiol, C. (2020). Consistency and identifiability of the polymorphism-aware phylogenetic models. *Journal of Theoretical Biology*, *486*, 110074.

Borges, R., Szöllősi, G. J., & Kosiol, C. (2019). Quantifying GC-biased gene conversion in great ape genomes using polymorphism-aware models. *Genetics*, *212*(4), 1321–1336.

Bryant, D., Bouckaert, R., Felsenstein, J., Rosenberg, N. A., & RoyChoudhury, A. (2012). Inferring species trees directly from biallelic genetic markers: Bypassing gene trees in a full coalescent analysis. *Molecular Biology and Evolution*, *29*(8), 1917–1932.

De Maio, N., Schlötterer, C., & Kosiol, C. (2013). Linking great apes genome evolution across time scales using polymorphism-aware phylogenetic models. *Molecular Biology and Evolution*, *30*(10), 2249–2262.

De Maio, N., Schrempf, D., & Kosiol, C. (2015). PoMo: An allele frequency-based approach for species tree estimation. *Systematic Biology*, *64*(6), 1018–1031.

Fabreti, L. G., & Höhna, S. (2022). Convergence assessment for bayesian phylogenetic analysis using mcmc simulation. *Methods in Ecology and Evolution*, *13*(1), 77–90.

Flouri, T., Jiao, X., Rannala, B., & Yang, Z. (2020). A Bayesian implementation of the multispecies coalescent model with introgression for phylogenomic analysis. *Molecular Biology and Evolution*, *37*(4), 1211–1223.

Freyman, W. A., & Höhna, S. (2018). Cladogenetic and anagenetic models of chromosome number evolution: A Bayesian model averaging approach. *Systematic Biology*, *67*(2), 1995–1215.

Galtier, N., Duret, L., Glémin, S., & Ranwez, V. (2009). GC-biased gene conversion promotes the fixation of deleterious amino acid changes in primates. *Trends in Genetics*, *25*(1), 1–5.

Hasegawa, M., Kishino, H., & Yano, T. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial {DNA}. *Journal of Molecular Evolution*, *22*, 160–174.

Höhna, S., Coghill, L. M., Mount, G. G., Thomson, R. C., & Brown, J. M. (2018). P3: Phylogenetic posterior prediction in RevBayes. *Molecular Biology and Evolution*, *35*(4), 1028–1034.

Höhna, S., Heath, T. A., Boussau, B., Landis, M. J., Ronquist, F., & Huelsenbeck, J. P. (2014). Probabilistic graphical model representation in phylogenetics. *Systematic Biology*, *63*(5), 753–771.

Höhna, S., Landis, M. J., Heath, T. A., Boussau, B., Lartillot, N., Moore, B. R., Huelsenbeck, J. P., & Ronquist, F. (2016). RevBayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. *Systematic Biology*, *65*(4), 726–736.

Höhna, S., Landis, M. J., & Huelsenbeck, J. P. (2021). Parallel power posterior analyses for fast computation of marginal likelihoods in phylogenetics. *PeerJ*, *9*, e12438.

Jukes, T. H., & Cantor, C. R. (1969). Evolution of protein molecules. In H. N. Munro (Ed.), *Mammalian protein metabolism* (pp. 21–132). Academic Press.

Keightley, P. D., Ness, R. W., Halligan, D. L., & Haddrill, P. R. (2014). Estimation of the spontaneous mutation rate per nucleotide site in a Drosophila melanogaster full-sib family. *Genetics*, *196*(1), 313–320.

Leaché, A. D., & Oaks, J. R. (2017). The utility of single nucleotide polymorphism (SNP) data in phylogenetics. *Annual Review of Ecology, Evolution, and Systematics*, *48*(1), 69–84.

Lynch, M., Ackerman, M. S., Gout, J.-F., Long, H., Sung, W., Thomas, W. K., & Foster, P. L. (2016). Genetic drift, selection and the evolution of the mutation rate. *Nature Reviews Genetics*, *17*(11), 704–714.

Maddison, W. P., & Knowles, L. L. (2006). Inferring phylogeny despite incomplete lineage sorting. *Systematic Biology*, *55*(1), 21–30.

Nagylaki, T. (1983). Evolution of a finite population under gene conversion. *Proceedings of the National Academy of Sciences of the United States of America*, *80*(20), 6278–6281.

Nolen, Z. J., Yildirim, B., Irisarri, I., Liu, S., Groot Crego, C., Amby, D. B., Mayer, F., Gilbert, M. T. P., & Pereira, R. J. (2020). Historical isolation facilitates species radiation by sexual selection: Insights from Chorthippus grasshoppers. *Molecular Ecology*, *29*(24), 4985–5002.

Ogilvie, H. A., Heled, J., Xie, D., & Drummond, A. J. (2016). Computational performance and statistical accuracy of *BEAST and comparisons with other methods. *Systematic Biology*, *65*(3), 381–396.

Rannala, B., & Yang, Z. (2017). Efficient Bayesian species tree inference under the multispecies coalescent. *Systematic Biology*, *66*(5), 823–842.

Schrempf, D., Minh, B. Q., De Maio, N., von Haeseler, A., & Kosiol, C. (2016). Reversible polymorphism-aware phylogenetic models and their application to tree inference. *Journal of Theoretical Biology*, *407*, 362–370.

Schrempf, D., Minh, B. Q., von Haeseler, A., & Kosiol, C. (2019). Polymorphism-aware species trees with advanced mutation models, bootstrap, and rate heterogeneity. *Molecular Biology and Evolution*, *36*(6), 1294–1301.

Sueoka, N., & Kawanishi, Y. (2000). DNA G+C content of the third codon position and codon usage biases of human genes. *Gene*, *261*(1), 53–62.

Szöllősi, G. J., Tannier, E., Daubin, V., & Boussau, B. (2015). The inference of gene trees with species trees. *Systematic Biology*, *64*(1), e42–e62.

Talts, S., Betancourt, M., Simpson, D., Vehtari, A., & Gelman, A. (2018). Validating bayesian inference algorithms with simulation-based calibration. *arXiv* [online]. http://arxiv.org/abs/1804.06788

Tavaré, S. (1986). Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on Mathematics in the Life Sciences*, *17*, 57–86.

How to cite this article: Borges, R., Boussau, B., Höhna, S., Pereira, R. J., & Kosiol, C. (2022). Polymorphism-aware estimation of species trees and evolutionary forces from genomic sequences with RevBayes. *Methods in Ecology and Evolution*, *00*, 1–8. https://doi.org/10.1111/2041-210X.13980