Polymorphism-aware species trees with advanced mutation models, bootstrap and rate heterogeneity

Dominik Schrempf[1,2,4], Bui Quang Minh[3,4], Arndt von Haeseler[4,5], and Carolin Kosiol [*2,6]

[1]Department of Biological Physics, Eötvös Loránd University, Budapest, Hungary
[2]Centre for Biological Diversity, University of St Andrews, United Kingdom
[3]Ecology and Evolution, Research School of Biology, Australian National University, Australia
[4]Center for Integrative Bioinformatics Vienna, Max F. Perutz Laboratories, University of Vienna, Medical University of Vienna, Austria
[5]Bioinformatics and Computational Biology, Faculty of Computer Science, University of Vienna, Austria
[6]Institut für Populationsgenetik, Vetmeduni Vienna, Austria

December 1, 2018

Abstract

Molecular phylogenetics has neglected polymorphisms within present and ancestral populations for a long time. Recently, multispecies coalescent based methods have increased in popularity, however, their application is limited to a small number of species and individuals. We introduced a polymorphism-aware phylogenetic model (PoMo), which overcomes this limitation and scales well with the increasing amount of sequence data while accounting for present and ancestral polymorphisms. PoMo circumvents handling of gene trees and directly infers species trees from allele frequency data. Here, we extend the PoMo implementation in IQ-TREE and integrate search for the statistically best-fit mutation model, the ability to infer mutation rate variation across sites, and assessment of branch support values. We exemplify an analysis of a hundred species with ten haploid individuals each, showing that PoMo can perform inference on large data sets. While PoMo is more accurate than standard substitution models applied to concatenated alignments, it is almost as fast. We also provide `bmm-simulate`, a software package that allows simulation of sequences evolving under PoMo. The new options consolidate the value of PoMo for phylogenetic analyses with population data.

Keywords: Incomplete Lineage Sorting, Species Tree, Phylogenetics, Polymorphism-aware Phylogenetic Model, Boundary Mutation Model.

*Corresponding author. E-mail: ck202@st-adnrews.ac.uk.

1 Introduction

Molecular phylogenetics seeks to estimate evolutionary relationships of species depicted as species trees by modeling the change and development of hereditary sequences. Established methods (e.g., Tavaré, 1986; Yang, 2006) that neglect molecular variation on the population level suffer from incongruencies between genomic regions which may arise due to incomplete lineage sorting (ILS, Knowles, 2009; Maddison, 1997). The effect of ILS becomes large if the considered species are closely related or if internal branches of the species tree are short when measured in number of generations divided by the effective population size Pamilo and Nei (1988). Short branches, especially when appearing in caterpillar-like topologies may lead to statistical inconsistency Degnan (2013); Degnan and Rosenberg (2009) of approaches such as concatenation Gadagkar *et al.* (2005), where sequences for different genes of the same individual are joined to form one overall alignment.

Consequently, development of phylogenetic methods has increasingly focused on explicit modeling of population genetic effects Leaché and Oaks (2017). Most methods employ the multispecies coalescent model Rannala and Yang (2003) to reconcile gene trees, i.e., the evolutionary histories of genes, with the species tree. The species tree is either jointly estimated with the gene trees (e.g., Drummond and Rambaut, 2007; Heled and Drummond, 2010; Liu, 2008) or reconstructed from previously estimated gene trees (e.g., Liu *et al.*, 2010; Mirarab *et al.*, 2014).

We have recently proposed a polymorphism-aware phylogenetic model (PoMo, De Maio *et al.*, 2015; Schrempf *et al.*, 2016) for estimating species trees from genome-wide data for up to dozens of species with multiple individuals each, which bypasses the computational burden of estimating gene trees. PoMo can be viewed as an extension of classical substitution models which additionally considers polymorphisms. Present as well as ancestral polymorphisms are described along a species tree by separating mutation and genetic drift in a population genetic framework that we have termed multivariate boundary mutation model Schrempf and Hobolth (2017); Vogl and Clemente (2012). The substitution model (e.g., HKY model, Hasegawa *et al.*, 1985) extended by the multivariate boundary mutation model is referred to as mutation model, because frequency changes are considered separately. PoMo improves estimation of mutation parameters and species trees compared to established methods such as concatenation when ILS has blurred the phylogenetic signal. Extensive evaluations showed good performance of PoMo against other methods De Maio *et al.* (2015); Schrempf *et al.* (2016). Here we extend our implementation into the maximum likelihood software IQ-TREE Nguyen *et al.* (2015), referred to as IQ-TREE-PoMo, which includes several new features as detailed in the next section.

2 New Approaches

Big data. In contrast to the multispecies coalescent model, PoMo scales much better with the number of analyzed species. The performance of IQ-TREE PoMo was tested by applying it to simulated multiple sequence alignments (MSA) of a length of up to 1 million nucleotides, and 100 species with ten sampled individuals each corresponding to gene trees with 1000 leaves (see Materials and Methods).

Simulator. We developed and released a software package to generate sequences evolved under various boundary mutation models (bmms), in particular PoMos, called `bmm-simulate` (`https://github.com/pomo-dev/bmm-simulate`). Given a species tree, the simulator directly generates polymorphic sequences using the discrete multivariate boundary mutation model with mutation rate heterogeneity (see below). This simulation tool complements the coalescent simulators available which we have used in this and in previous studies and allows exact assessment of inference accuracy under PoMo.

Advanced models. Heterogeneity of rates across sites is a consequence of varying mutation and fixation rates. If rate heterogeneity is not taken into account, sequence distance is underestimated and phylogenetic analyses may suffer from long branch attraction artifacts (Yang, 2006, p.19).

Therefore, we developed new theory for PoMo to incorporate Γ distributed mutation rate heterogeneity across sites Yang (1994). The new theory is necessary as only the mutation rates are heterogeneous across sites, while drift rates stay constant for our model (see Materials and Methods). A simulation study with `bmm-simulate` showed high accuracy when estimating the Γ shape parameter and other parameters on trees with twelve species.

Furthermore, species tree search can be performed with fixed parameters. Manual specification of parameters is useful when detailed knowledge is available or when likelihoods are compared between different analyses. In practice, we often do not have total control over the number of individuals in the population that are sampled and sequenced (e.g., sequences are retrieved from existing databases, re-sequencing runs failed or are simply too costly even for as few as 10 individuals). We therefore present two strategies, called weighted binomial and weighted hypergeometric sampling, to initialize the likelihoods of the PoMo states at the tips of the trees, and to account for this missing data problem.

ModelFinder. Model selection, which includes finding the best-fit mutation model as well as a model of mutation rate heterogeneity across sites,

is crucial when performing molecular phylogenetic analyses. The improved implementation of PoMo allows for highly flexible polymorphism-aware phylogenetic analyses in that the most suitable evolutionary model for the data at hand can be automatically determined using statistical model search with ModelFinder Kalyaanamoorthy *et al.* (2017). ModelFinder is a model selection framework in IQ-TREE which has now been made available for IQ-TREE-PoMo. The Bayesian Schwarz (1978) or the Akaike (1973) information criterion are employed to determine the best-fit model.

Bootstrapping. IQ-TREE-PoMo is fast enough to allow for bootstrapping Efron (1979); Felsenstein (1985) on large data sets. Evaluation of branch support values using the branch-wise approximate likelihood ratio test (SH-aLRT, Guindon *et al.*, 2010) as well as standard nonparameteric bootstrap Efron (1979) and ultrafast bootstrap (UFBoot2, Hoang *et al.*, 2018) is now possible. The ability of bootstrapping was tested on a data set of great ape genomes (see Material and Methods, Prado-Martinez *et al.*, 2013).

3 Results

Big data. The phylogenetic analysis of the simulated MSAs of 100 species with ten sampled individuals each shows that consideration of heterozygosity improves estimates considerably compared to the standard concatenation approach. Branch score distance (Fig. 1), as well as Robinson-Foulds distance (Fig. S1) strongly decrease, especially when sufficient data is available. Remarkably, IQ-TREE-PoMo already outperforms concatenation for as little as ten genes (10 000 sites) and exhibits branch score distances around ten times more accurate when 1000 genes (one million sites) are available. Further, the progression of branch score distance for the concatenation method with increasing amount of data is not monotonically decreasing. The average run time (wall-clock time) of IQ-TREE-PoMo for 1000 genes which corresponds to a sequence length of one million sites is 12.3 ± 0.7 hours on a 2.6 GHz processor with 16 physical cores (Intel® Xeon® CPU E5-2650 v2 @ 2.60GHz) . The concatenation method is approximately six times faster (2 ± 0.25 hours).

Advanced models We assess the accuracy of estimating rate heterogeneity by focussing on the shape parameter α of the Γ distribution. The relative error (difference between true and estimated value in percent) of the shape parameter α for ten replicate analyses is usually within two percent (see Fig. 2 for $\alpha = 0.3$, 1.0, and 5.0, respectively) but can be higher for more extreme α values (≈ 20 and ≈ 400 percent for $\alpha = 0.1$, and 10, respectively; Fig. S2; – S4). The relative errors in terms of branch score distance and

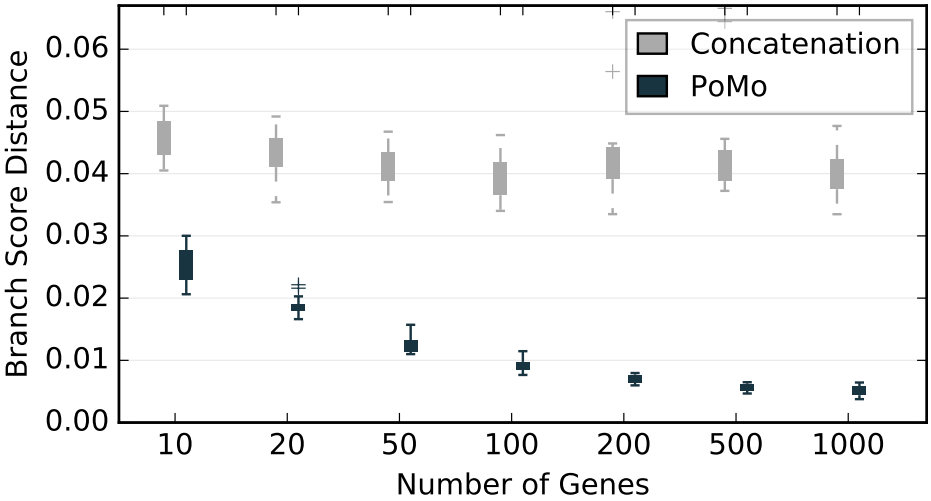

Figure 1: Branch score distance of concatenation approach and IQ-TREE-PoMo with $N = 10$ and weighted binomial sampling for Yule trees with 100 species and ten individuals each. The tree height measured in coalescent units is $6N_e$, where N_e is the effective population size. The HKY model was used for both inference methods. The heterozygosity is $\theta_W = 0.005$ per site. Each gene spans 1000 sites. The error bars are standard deviations of ten replicate analyses.

the transition to transversion ratio parameter of the HKY mutation model κ are mostly below 0.5 and one percent, respectively.

The heterozygosity θ determines the level of polymorphism present in the species. We tested the accuracy of PoMo with increasing heterozygosity because for high θs the model assumption of having boundary mutations only is violated. We observe low errors in branch score distance for heterozygosity values up to $\theta = 0.1$ in analyses of species trees with twelve species with ten individuals each and a tree height of $3N_e$ (Fig. 3).

For the case of uneven sampling of individuals of the populations, we tested two different strategies to account for the missing data in the poorly sampled populations. We found that weighted binomial sampling outperforms weighted hypergeometric sampling (see Materials and Methods and Supplement S3 for details on these strategies).

ModelFinder and bootstrapping. Model selection and bootstrapping were tested on data of 12 great ape species (see Material and Methods). For PoMo, the GTR Tavaré (1986) mutation model with Γ rate heterogeneity was determined to be the best fitting model. The inferred Γ shape parameter is 1.26. The estimated phylogeny (Fig. 4) confirms previous results Schrempf *et al.* (2016). UFBoot2 and SH-aLRT both report 100 percent support for all branches.

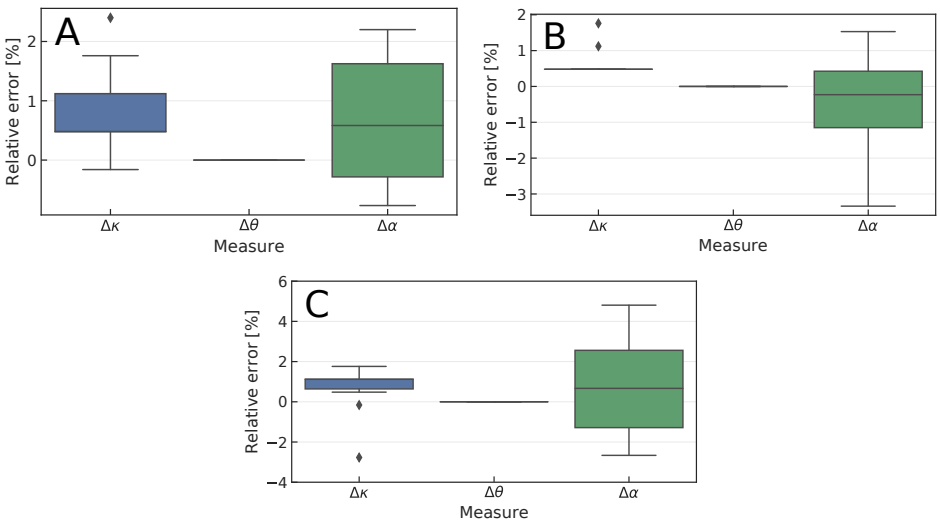

Figure 2: Relative errors of the transition to transversion ratio $\Delta\kappa$, the heterozygosity $\Delta\theta$, and the shape parameter of the Γ distributed mutation rate heterogeneity $\Delta\alpha$. The true shape parameter is $\alpha = 0.3$ (A), $\alpha = 1.0$ (B), and $\alpha = 5.0$ (C), respectively.

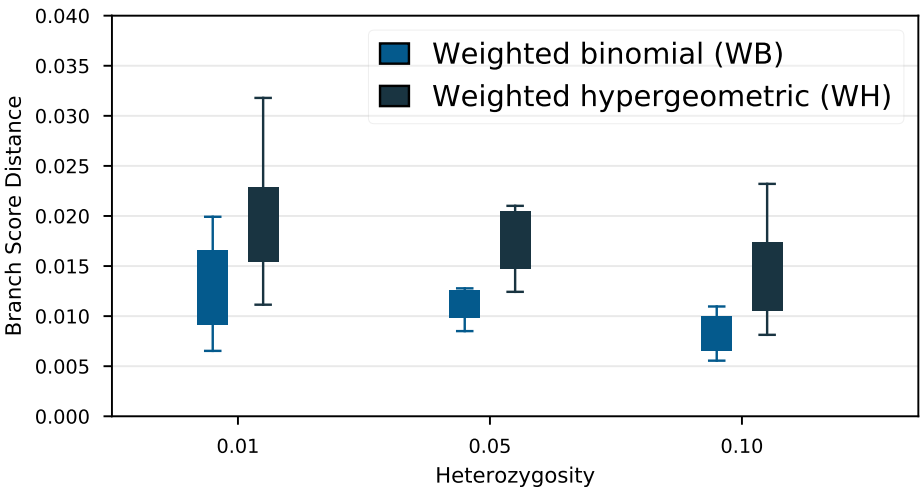

Figure 3: Branch score distance of weighted binomial and weighted hypergeometric sampling for Yule trees of height $3N_e$ with twelve species and ten individuals each. Heterozygosity varies between 0.01 and 0.1.

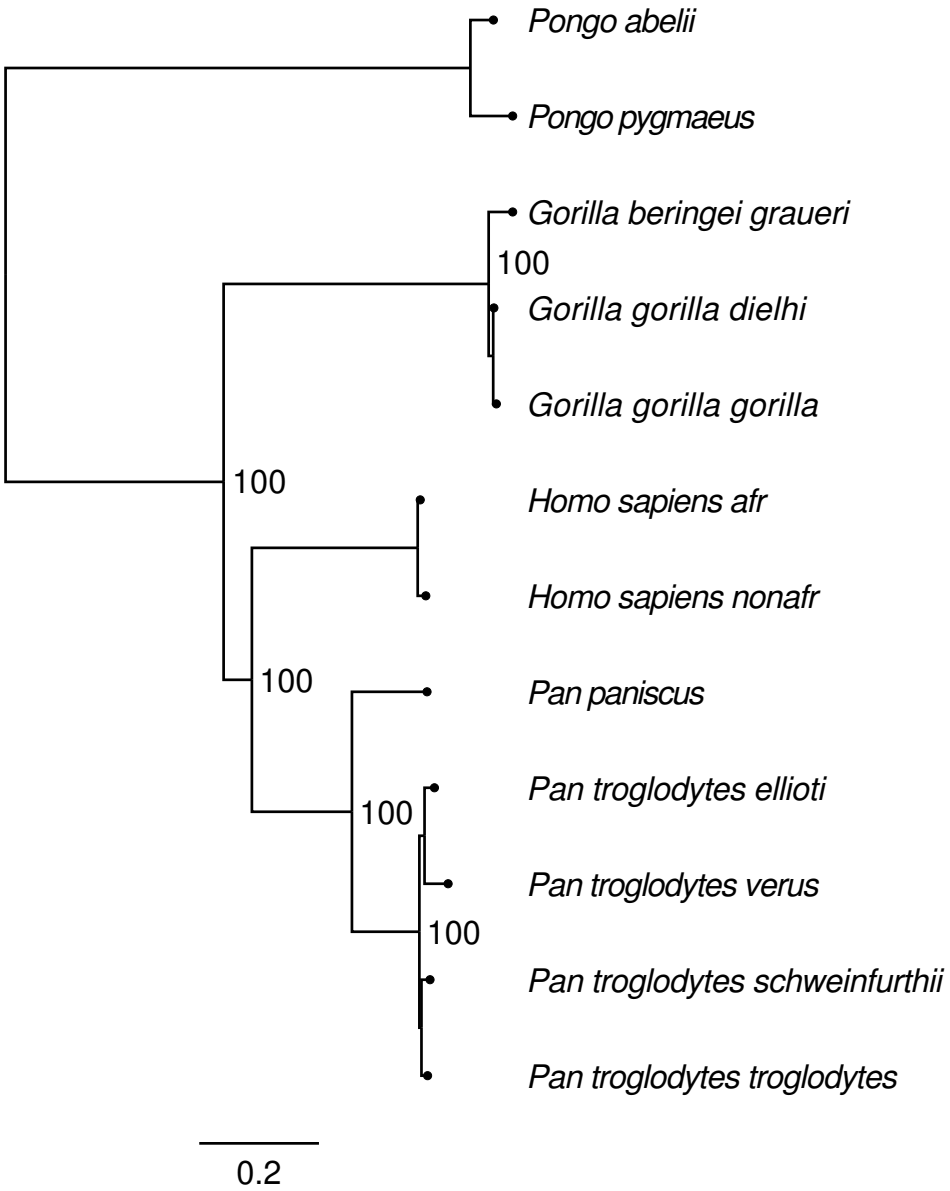

Figure 4: Phylogeny inferred from primate data Prado-Martinez *et al.* (2013). Both, UFBoot2 Hoang *et al.* (2018) and SH-aLRT Guindon *et al.* (2010) branch support tests evaluated to hundred percent support values.

4 Discussion

The analysis of data from 100 species underlines that IQ-TREE-PoMo is fast and scales well with increasing number of species. The chosen species tree exhibits significant amount of ILS, and hence, it is expected that our method which accounts for polymorphisms has higher accuracy than concatenation. In contrast to the concatenation method, which exhibits again statistical inconsistency (Fig. 1) as previously reported by Degnan and colleagues (Degnan, 2013; Degnan and Rosenberg, 2009), the branch score distance of IQ-TREE-PoMo continuously decreases when more data is analyzed (Fig. 1).

Further, it is now possible to infer parameters of more advanced sequence evolution models with IQ-TREE-PoMo. When modeling mutation rate heterogeneity, the shape parameter α of the Γ distribution is recovered with relative errors below two percent for $0.3 \leq \alpha \leq 5.0$. The inferred value of $\alpha = 1.26$ for the great ape data set lies well in this range. We observe a slight overestimation of the shape parameter, when a significant amount of sites evolves extremely slowly and is nearly invariant ($\alpha \leq 0.1$, Fig. S2). Similarly and as expected, when the mutation rate distribution is homogeneous ($\alpha \geq 10.0$), the variability of estimated shape parameters is higher (Fig. S4).

We decided not to implement the invariant sites model. The reason is that polymorphisms are the consequence of strictly positive mutation rates. When mutation rates are zero, all states but the boundary states have a stationary frequency of zero. This leads to problems because (1) observed polymorphic sites have zero likelihood and cannot be assigned to any model state and (2) numerical problems are encountered when calculating the transition probabilities. Although a separate model could be used to calculate the likelihood conditioned on the mutation rates being zero, we strongly believe that, if the data exhibits nearly invariant sites, those are sufficiently covered by Γ distributed mutation rate heterogeneity with low shape parameters.

As Mallo (2017) reported increased errors of the estimates of PoMo for high heterozygosity values up to $\theta = 0.05$, we tested the robustness of our approach in this respect. We find that the accuracy of IQ-TREE-PoMo with respect to BSD increases with higher heterozygosity up to a value of $\theta = 0.1$, which is well above the observed value in primates ($\theta \sim 0.001$, Prado-Martinez *et al.*, 2013) and most other organisms (except, Lynch *et al.*, 2016).

We recommend to use weighted binomial sampling instead of weighted hypergeometric sampling. Although weighted hypergeometric sampling has the advantage of retaining the heterozygosity between the observed data and the model, questions arise when the number of samples is lower than the number of PoMo frequency bins. In practice, the latter is often the cause

of missing data. Correspondingly, IQ-TREE-PoMo uses weighted binomial sampling with $N = 9$ as default, a setting that has repeatedly been observed to be the most accurate and stable one Schrempf *et al.* (2016).

As more advanced mutation models are now available, the choice of the statistically most adequate mutation model is of major importance. ModelFinder does not require user input and is fast. The likelihood improvement when accounting for mutation rate heterogeneity is also tested for and, consequently, the most appropriate mutation rate heterogeneity model is reported and automatically used.

For the great ape data set, the best-fit model for PoMo coincides with results from concatenation methods and from the phylogenetic analysis in the original paper Prado-Martinez *et al.* (2013). In the latter case, not only the model choice coincides, but also the topology, which is a further confirmation of the validity of PoMo.

Additionally, assessment of branch support values is now possible with normal bootstrap, UFBoot2 and likelihood ration based tests (SH-aLRT). Branch support values for the primate data are high because 2.8 million sites are considered. This is not surprising and emphasizes that the results are robust. Especially for smaller data sets, this feature will be highly useful.

Overall, we provide important extensions to further establish the use of polymorphism-aware models in phylogenetics. In the future, we would like to implement probability distribution free rate category models Kalyaanamoorthy *et al.* (2017). We have repeatedly discussed allowing non-reversible mutation models to be used with PoMo since IQ-TREE offers a whole set of non-reversible mutation models consistent with heterogeneous mutation rates (Lie-Markov models, Woodhams *et al.*, 2015). IQ-TREE does not only provide an extensive set of mutation models or allows usage of new model selection methods such as ModelFinder, but also comes with other benefits such as the ability to resume analyses (checkpointing), or carefully designed parallelization techniques. In conclusion, PoMo has evolved to be a mature, well-tested and flexible method to perform phylogenetic inference with population data.

5 Materials and Methods

PoMo. A detailed description of PoMo and the discrete multivariate boundary mutation model can be found in Schrempf *et al.* (2016) and Schrempf and Hobolth (2017), respectively. PoMo is a time-continuous Markov process modeling sequence evolution along a species tree. Sites are assumed to be independent (composite likelihood, free recombination). At each site, not only evolution of a single character $a \in \mathcal{A}$ (e.g., $\mathcal{A} = \{A, C, G, T\}$, $|\mathcal{A}| = 4$) of the reference genome is considered but rather the evolution of the collective characters of a population of genomes.

Actual populations are big in size and direct treatment is not feasible. For neutral evolution, effective population size is confined with mutation rates. It is possible to scale down the effective population size to a small value while scaling up mutation rates such that the overall dynamics remain unchanged. We take advantage of this property by choosing a rather small number of collected characters N of the multivariate boundary mutation model. Consequently, the parameter N should be interpreted as a discretization parameter (not to be confused with effective population size) describing the number of bins that allele frequencies can fall into.

The rates of frequency shifts, are determined using the time-continuous Moran process Moran (1958). PoMo assumes that drift removes variation fast, and mutations are disallowed when more than one allele is present in the population. Hence, the population can either be monomorphic for an allele $a \in \mathcal{A}$ or polymorphic for two alleles $a, b \in \mathcal{A}$ with counts i and $(N-i)$, respectively. The monomorphic and polymorphic states of the multivariate boundary mutation model are denoted by $\{a\}$ and $\{ia|(N-i)b\}$, $a \neq b$, respectively. Disallowing mutations when the population is polymorphic is a good approximation as long as the heterozygosity is below a value of 0.1, a requirement that is readily satisfied in most cases.

The transition rate matrix

$$\boldsymbol{Q} = \boldsymbol{Q}_M + \boldsymbol{Q}_D, \tag{1}$$

of dimension $|\mathcal{A}| + \binom{|\mathcal{A}|}{2}(N-1)$ is composed by mutation (M) and genetic drift (D). The only off-diagonal, non-zero entries of $\boldsymbol{Q}_M$ are mutations away from monomorphic states

$$(\boldsymbol{Q}_M)_{\{a\} \to \{(N-1)a|1b\}} = q_{ab}, \tag{2}$$

where the q_{ab} are mutation rates defined by an underlying mutation model such as the HKY mutation model Hasegawa *et al.* (1985), or the GTR mutation model Tavaré (1986). Drift rates are non-zero for neighboring states only $(1 \leq i \leq N-1)$

$$(\boldsymbol{Q}_D)_{\{ia|(N-i)b\} \to \{(i\pm1)a|(N-i\mp1)b\}} = \frac{i(N-i)}{N}. \tag{3}$$

Diagonal elements of $\boldsymbol{Q}_M$ and $\boldsymbol{Q}_D$ are set such that all row sums are zero.

Big data. For the simulation of large data sets, we employed a pipeline that follows Schrempf *et al.* (2016) First, ten species trees with 100 leaves were randomly generated under the Yule birth model Yule (1925). Each of the ten species trees is referred to as one of ten replicates. The height of the species trees measured in number of generations was 6 times the effective population size which is assumed constant. The Yule birth rate

was set such that the expected number of species for the given height is 100. Second, for each replicate, 1000 gene trees were simulated under the multispecies coalescent model. SimPhy Mallo *et al.* (2016) was used for these steps. Finally, for each gene tree, DNA sequences with 1000 sites were generated with Seq-Gen Rambaut and Grass (1997) under the HKY mutation model Hasegawa *et al.* (1985). The transition to transversion ratio was $\kappa = 6.25$, the stationary nucleotide frequencies were $\pi_A = 0.3$, $\pi_C = 0.2$, $\pi_G = 0.2$, and $\pi_T = 0.3$. The simulated sequences had a heterozygosity of 0.005 which is approximately four times the value observed in primates (e.g., Prado-Martinez *et al.*, 2013).

Estimation of the original species trees with 100 leaves was performed with IQ-TREE using either PoMo or a standard concatenation method. The HKY model was used for both methods. The discretization parameter of the multivariate boundary mutation model was $N = 10$ and weighted binomial sampling was used. Command lines for simulation and analysis are in Section S1. The accuracy of the inferences was measured by comparing the estimated to the original species trees. Trees were normalized to a tree height of 1.0. The branch score distance (BSD, Kuhner and Felsenstein, 1994) and the Robinson-Foulds distance Robinson and Foulds (1981) were used.

High heterozygosity. For the assessment of the accuracy of tree inference for different heterozygosity values (Fig. 3) the same simulation pipeline was used. In contrast to above, the height of the species trees was 3 times the effective population size, the number of taxa was twelve, and the heterozygosity values was 0.01, 0.05, and 0.1, respectively.

Simulator. Although the above pipeline corresponds to the way how methods based on the multispecies coalescent model perform inference, estimates of PoMo are accurate (Fig. 3 and Schrempf *et al.*, 2016). When considering mutation rate heterogeneity, we have to keep in mind that the mutation rates of the multivariate boundary mutation model are not independent of the effective population size. A mutational event $\{a\} \rightarrow \{(N-1)a|1b\}$, $a, b \in \mathcal{A}$, $a \neq b$, corresponds to a mutation in the real population with subsequent frequency shift to a value of $1/N$. Of course, the probability of such an event depends on the effective population size of the real population. As a result, higher mutation rates in the multivariate boundary mutation model, associated with a higher probability to be in a polymorphic state when species split, lead to a higher probability of ILS.

In contrast, the probability of ILS derived by the multispecies coalescent model does not explicitly depend on the mutation rate. For example, a rooted, three-taxon tree has a probability of ILS proportional to $e^{-c\Delta t/N_e}$ (e.g., Nei, 1987) , where c is a constant, Δt is the branch length

of the internal branch measured in number of generations, and N_e is the effective population size. These considerations urged us to develop a multivariate boundary mutation model simulator for direct assessment of the accuracy of mutation rate heterogeneity inference.

Advanced models. Heterogeneity in evolutionary rates across sites may be modeled with a parametric distribution such as the Γ distribution Yang (1994). For the multivariate boundary mutation model, complications arise because the transition rate matrix contains contributions from mutations $\boldsymbol{Q}_M$ as well as frequency shifts due to random genetic drift $\boldsymbol{Q}_D$. A general scaling of the transition rate matrix, like it is usually done when using Γ rate heterogeneity with substitution models, would erroneously (de-)accelerate both processes. This difficulty was overcome by employing the concept of mixture models.

In brief, K mutation rate categories with rate modifiers r_k and corresponding probabilities $1/K$ of a site belonging to a category are defined. The total likelihood of a mutation model accounting for mutation rate heterogeneity M_h and species tree T given data D is

$$\mathcal{L}(M_h, T|D) = \Pr(D|M_h, T) = \sum_{k=1}^{K} \frac{1}{K} \Pr(D|\boldsymbol{Q}_k, T), \tag{4}$$

where

$$\boldsymbol{Q}_k = r_k \boldsymbol{Q}_M + \boldsymbol{Q}_D. \tag{5}$$

Hereby, the mutation rates modifiers r_k are the means of the mutation rate categories. The latter is similar to the classical treatment using Γ rate heterogeneity, however, this setup allows modeling of mutation rate heterogeneity with any free mutation rate categories. We chose to use a parametrized Γ distribution because it is most used. Naturally, the usual efficiency of Γ rate heterogeneous models is not retained, because the transition matrices for different mutation rate categories are intrinsically different. That is, an analysis with k categories requires eigendecomposition of k transition rate matrices.

The accuracy in estimating the shape parameter of the Γ distributed mutation rate heterogeneity was assayed using `bmm-simulate`. Species trees were randomly sampled from a Yule process. The tree height measured in average number of substitutions per site was $h = 0.01$. The speciation rate λ was set such that $n = 12$ species are present on average. That is (e.g., Kendall, 1949)

$$\lambda = \frac{1}{h} \left(\sum_{i=1}^{n} \frac{1}{i} - 1 \right). \tag{6}$$

For each sampled species tree, sequences were simulated under the discrete multivariate boundary mutation model. The discretization parameter N was set to 10, and the heterozygosity to 0.0025. Similar to the big data analysis above, the HKY mutation model was used with $\kappa = 6.25$ and stationary nucleotide frequencies $\pi_A = 0.3$, $\pi_C = 0.2$, $\pi_G = 0.2$, and $\pi_T = 0.3$. Four Γ mutation rate categories were used and one million sites were simulated. Ten replicate analyses with different, randomly sampled species trees were performed. The shape parameter α of the Γ distributed mutation rate heterogeneity was set to 0.1, 0.3, 0.5, 1.0, 5.0, and 10.0. Command lines for simulation and analysis with rate heterogeneity are in Section S2.

Sampling. Interpretation of data is mostly predetermined when using phylogenetic substitution models, because observed states can directly be mapped to internal states of the used substitution model. Special handling is required when encountering an unknown site. Then, the likelihood of all states is set to 1.0, and all model states have the same probability of leading to the observed data.

Equivalently, we would like to assign the likelihoods of all multivariate boundary model states at a specific site in the alignment and leaf of the tree. If the number of samples coincides with the discretization parameter N, and if not more than two alleles are present, the observed data is a multivariate boundary mutation model state. Like above with substitution models, the likelihood of this state can be set to 1.0, and the likelihood of all other states to 0.0.

In practice, we often do not have total control over the number of individuals sampled. For example, sequences are retrieved from existing databases, or the high sequencing costs set constraints on the number of individuals. We therefore designed various strategies to assign the likelihoods of the multivariate boundary mutation model states Schrempf *et al.* (2016) if the number of samples does not match N. The easiest one is to randomly draw N alleles with replacement from the data. We refer to this methods as *sampled*. We can also assign likelihoods in an inverse manner, similar to handling unknown sites with substitution models. A specific multivariate boundary mutation model state is assumed at the considered leaf and the probability of observing the data given this state is calculated. We can sample with or without replacement and calculate this probability for all multivariate boundary mutation model states. We have termed these ways of assigning likelihoods *weighted binomial* and *weighted hypergeometric* sampling. A detailed description is given in Section S3.

Bootstrapping using real data. We revisited the great ape data set of six species subdivided into 12 populations with up to 23 individuals each Prado-Martinez *et al.* (2013). The exome-wide data set includes roughly

2.8 million, 4-fold degenerate sites. Data preparation is described in De Maio *et al.* (2015), the counts file is available on `https://github.com/pomo-dev/data`.

For the analysis we used ModelFinder and assessed branch support with 1000 bootstraps. We tested estimations with and without Γ mutation rate heterogeneity with four discrete categories combined with UFBoot2 and SH-aLRT. The discretization parameter was $N = 9$. The used command line was

```
iqtree -s hg18-all.cf -alrt 1000 -bb 1000
```

6 Acknowledgments

This work was funded by the Vienna Science and Technology Fund (WWTF) through project MA16-061. DS was supported by the Austrian Science Fund [FWF-P24551, I-2805-B29] and received funding from the European Research Council under the European Union's Horizon 2020 research and innovation programme under grant agreement no. 741774. The computational results presented have been achieved in part using the Vienna Scientific Cluster (VSC) and the St Andrews Bioinformatics Unit (StABU) which is funded by a Wellcome Trust ISSF award (grant 105621/Z/14/Z).

References

Akaike, H. 1973. Information theory and an extension of the maximum likelihood principle. In B. N. Petrov and F. Csaki, editors, *Proceedings of the 2nd International Symposium on Information Theory, Tsahkadsor, Armenia, USSR, September 2-8, 1971*, pages 267–281, Budapest. Akadémiai Kiadó.

De Maio, N., Schrempf, D., and Kosiol, C. 2015. PoMo: An Allele Frequency-Based Approach for Species Tree Estimation. *Systematic Biology*, 64(6): 1018–1031.

Degnan, J. H. 2013. Anomalous Unrooted Gene Trees. *Systematic Biology*, 62(4): 574–590.

Degnan, J. H. and Rosenberg, N. A. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends in Ecology & Evolution*, 24(6): 332–340.

Drummond, A. J. and Rambaut, A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology*, 7: 214.

Efron, B. 1979. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7(1): 1–26.

Felsenstein, J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, 39(4): 783–791.

Gadagkar, S. R., Rosenberg, M. S., and Kumar, S. 2005. Inferring species phylogenies from multiple genes: Concatenated sequence tree versus consensus gene tree. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*, 304B(1): 64–74.

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. 2010. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Systematic Biology*, 59(3): 307–321.

Hasegawa, M., Kishino, H., and Yano, T.-a. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution*, 22(2): 160–174.

Heled, J. and Drummond, A. J. 2010. Bayesian Inference of Species Trees from Multilocus Data. *Molecular Biology and Evolution*, 27(3): 570–580.

Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q., and Le, S. V. 2018. UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Molecular Biology and Evolution*, 35(2): 518–522.

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A., and Jermiin, L. S. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature Methods*.

Kendall, D. G. 1949. Stochastic Processes and Population Growth. *Journal of the Royal Statistical Society. Series B (Methological)*, 11(2): 230–282.

Knowles, L. L. 2009. Estimating Species Trees: Methods of Phylogenetic Analysis When There Is Incongruence across Genes. *Systematic Biology*, 58(5): 463–467.

Kuhner, M. K. and Felsenstein, J. 1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Molecular Biology and Evolution*, 11(3): 459–468.

Leaché, A. D. and 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.

Liu, L. 2008. BEST: Bayesian estimation of species trees under the coalescent model. *Bioinformatics*, 24(21): 2542–2543.

Liu, L., Yu, L., and Edwards, S. V. 2010. A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. *BMC Evolutionary Biology*, 10(1): 302.

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

Maddison, W. P. 1997. Gene Trees in Species Trees. *Systematic Biology*, 46(3): 523–536.

Mallo, D. 2017. *Evaluation of phylogenomic methods for species tree estimation*. Ph.D. thesis, Universidad de Vigo.

Mallo, D., De Oliveira Martins, L., and Posada, D. 2016. SimPhy: Phylogenomic Simulation of Gene, Locus, and Species Trees. *Systematic Biology*, 65(2): 334–344.

Mirarab, S., Reaz, R., Bayzid, M. S., Zimmermann, T., Swenson, M. S., and Warnow, T. 2014. ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics*, 30(17): i541–i548.

Moran, P. a. P. 1958. Random processes in genetics. *Mathematical Proceedings of the Cambridge Philosophical Society*, 54(01): 60–71.

Nei, M. 1987. *Molecular Evolutionary Genetics*. Columbia University Press.

Nguyen, L.-T., Schmidt, H. A., von Haeseler, A., and Minh, B. Q. 2015. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular Biology and Evolution*, 32(1): 268–274.

Pamilo, P. and Nei, M. 1988. Relationships between gene trees and species trees. *Molecular Biology and Evolution*, 5(5): 568–583.

Prado-Martinez, J., Sudmant, P. H., Kidd, J. M., Li, H., Kelley, J. L., Lorente-Galdos, B., Veeramah, K. R., Woerner, A. E., O'Connor, T. D., Santpere, G., Cagan, A., Theunert, C., Casals, F., Laayouni, H., Munch, K., Hobolth, A., Halager, A. E., Malig, M., Hernandez-Rodriguez, J., Hernando-Herraez, I., Prüfer, K., Pybus, M., Johnstone, L., Lachmann, M., Alkan, C., Twigg, D., Petit, N., Baker, C., Hormozdiari, F., Fernandez-Callejo, M., Dabad, M., Wilson, M. L., Stevison, L., Camprubí, C., Carvalho, T., Ruiz-Herrera, A., Vives, L., Mele, M., Abello, T., Kondova, I., Bontrop, R. E., Pusey, A., Lankester, F., Kiyang, J. A., Bergl, R. A., Lonsdorf, E., Myers, S., Ventura, M., Gagneux, P., Comas, D., Siegismund, H., Blanc, J., Agueda-Calpena, L., Gut, M., Fulton, L., Tishkoff, S. A., Mullikin, J. C., Wilson, R. K., Gut, I. G., Gonder, M. K., Ryder, O. A., Hahn, B. H., Navarro, A., Akey, J. M., Bertranpetit, J., Reich, D., Mailund, T., Schierup, M. H., Hvilsom, C., Andrés, A. M., Wall, J. D., Bustamante, C. D., Hammer, M. F., Eichler, E. E., and Marques-Bonet, T. 2013. Great ape genetic diversity and population history. *Nature*, 499(7459): 471–475.

Rambaut, A. and Grass, N. C. 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Computer Applications in the Biosciences : CABIOS*, 13(3): 235–238.

Rannala, B. and Yang, Z. 2003. Bayes Estimation of Species Divergence Times and Ancestral Population Sizes Using DNA Sequences From Multiple Loci. *Genetics*, 164(4): 1645–1656.

Robinson, D. F. and Foulds, L. R. 1981. Comparison of phylogenetic trees. *Mathematical Biosciences*, 53(1-2): 131–147.

Schrempf, D. and Hobolth, A. 2017. An alternative derivation of the stationary distribution of the multivariate neutral Wright–Fisher model for low mutation rates with a view to mutation rate estimation from site frequency data. *Theoretical Population Biology*, 114: 88–94.

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

Schwarz, G. E. 1978. Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2): 461–464.

Tavaré, S. 1986. Some Probabilistic and Statistical Problems in the Analysis of DNA Sequences. *Lectures on Mathematics in the Life Sciences*, 17: 57–86.

Vogl, C. and Clemente, F. 2012. The allele-frequency spectrum in a decoupled Moran model with mutation, drift, and directional selection, assuming small mutation rates. *Theoretical Population Biology*, 81(3): 197–209.

Woodhams, M. D., Fernández-Sánchez, J., and Sumner, J. G. 2015. A New Hierarchy of Phylogenetic Models Consistent with Heterogeneous Substitution Rates. *Systematic Biology*, 64(4): 638–650.

Yang, Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *Journal of Molecular Evolution*, 39(3): 306–314.

Yang, Z. 2006. *Computational molecular evolution*, volume 284. Oxford University Press Oxford.

Yule, G. U. 1925. A Mathematical Theory of Evolution, Based on the Conclusions of Dr. J. C. Willis, F.R.S. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 213(402-410): 21–87.