Pervasive selection biases inferences of the species tree

Rui Borges[1], Bastien Boussau[2], Gergely J. Szöllősi[3,4,5], and Carolin Kosiol[1,6,*]

[1] Institut für Populationsgenetik, Vetmeduni Vienna, Veterinärplatz 1, 1210 Wien, Austria [2] Univ Lyon, Université Claude Bernard Lyon 1, CNRS UMR 5558, LBBE, F-69100, Villeurbanne, France [3] Department of Biological Physics, Eötvös University, Pázmány P. stny. 1A, H-1117 Budapest, Hungary [4] MTA-ELTE "Lendület" Evolutionary Genomics Research Group, Pázmány P. stny. 1A, H-1117 Budapest, Hungary [5] Evolutionary Systems Research Group, Centre for Ecological Research, Hungarian Academy of Sciences, 8237 Tihany, Hungary [6] Centre for Biological Diversity, University of St Andrews, St Andrews, Fife KY16 9TH, UK

* corresponding author: ck202@st-andrews.ac.uk

Abstract. Despite the importance of natural selection in species' evolutionary history, phylogenetic methods that take into account population-level processes ignore selection. Assuming neutrality is often based on the idea that selection occurs at a minority of loci in the genome and is unlikely to significantly compromise phylogenetic inferences. However, selection might behave more pervasively, as it the case of nearly neutral evolving mutations. Genome-wide processes like GC-bias and some of the variation segregating at the coding regions are known to evolve in the nearly neutral range. As we are now using genome-wide data to estimate species tree, it is just natural to ask whether weak, but pervasive, selection is likely to blur species tree inferences. Here, we employed a polymorphism-aware phylogenetic model, specially tailored for measuring signatures of nucleotide usage biases, to test the impact of nearly neutrally in the substitution process. Analyses with simulated data indicate that while the inferred relationships among species are not significantly compromised, the genetic distances are systematically underestimated, with the deeper nodes suffering more than the younger ones. Such biases have implications for molecular dating. We found signatures of GC-bias considerably affecting the estimated divergence times (up to 21%) of worldwide fruit fly populations. Our findings call for the need to account for nearly neutral forces (or any other form of pervasive selection) when quantifying divergence or dating species evolution.

Keywords: species tree · nearly neutral selection · GC-bias · molecular dating

Introduction

Phylogenetic trees form a basis to address many biological questions and are nowadays used routinely in integrative research. In particular, species trees have many uses, including their primary purpose, disentangling species evolutionary history, and dating species divergence, resolving intricate evolutionary events (e.g., radiations and gene duplications, losses, and transfers) and mapping trait evolution. All in all, species trees have been providing a framework to better understand the divergence process and speciation better.

Modeling species evolution using DNA sequences poses significant challenges. The main one being that different genes or genomic regions narrate different evolutionary histories, leading to discordant gene and species topologies [27]. Incomplete lineage sorting (ILS), i.e., the maintenance of ancestral polymorphisms due to random genetic drift, is a primary cause of such discordance [32]. But other processes were also described, as gene gain or loss, horizontal gene transfer across the species boundaries, and gene flow among diverging populations (reviewed in [38]). Apart from the gene/species tree discordance, difficulties arise when quantifying the divergence. It was reported that calendar times and evolutionary rates can be particularly challenging to untie for deeper evolutionary scales (where the molecular clock hypothesis is violated) [40].

Despite the challenges posed by modeling species evolution, the last two decades have seen an explosion of sophisticated statistical methods for inferring species trees. The multispecies coalescent model (MSC) has arisen as a leading framework for inferring species phylogenies while accounting for ILS and gene tree-species tree conflict. The coalescent traces the genealogical history of a sample of sequences from a population backward in time describing the stochastic process of lineage joining [19,20].

Nevertheless, alternative approaches to the MSC have been proposed, as the polymorphism-aware phylogenetic models (PoMos) [7,8] (Figure 1). PoMos model the allele content of a set of populations over time, thus naturally integrating over all the possible locus histories to directly estimate

the species tree. As a consequence, PoMos naturally account for ILS while avoiding using genealogy samplers that are typically computationally costly. Differently from the MSC, PoMos allow testing hypotheses genome-wide, gathering information from multiple individuals and several populations with high statistical power [36,37]. More importantly, PoMos are versatile, for they can be straightforwardly rebuilt to account for other population forces (e.g., allelic selection [4]). MSC-based methods with selection are notoriously difficult.

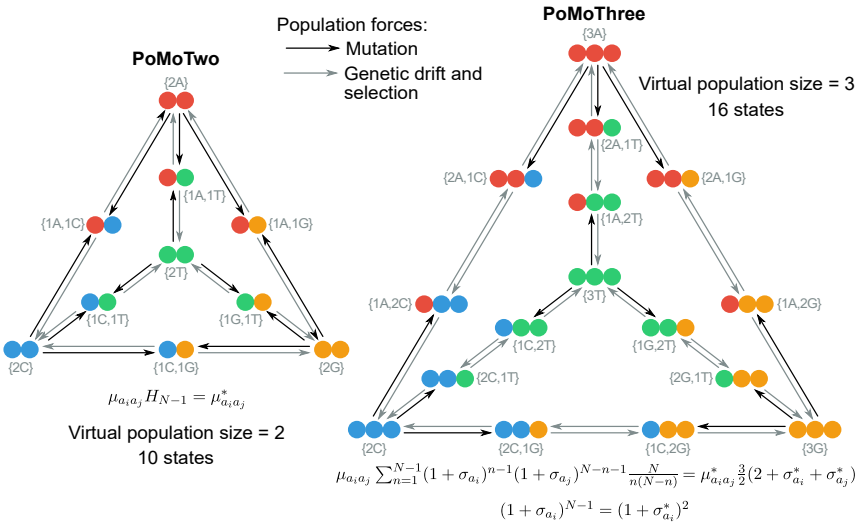

Fig. 1: **State-space and parameter scaling of the virtual PoMos.** The tetrahedrons represent the state-space of the virtual PoMos: the fixed sites $\{Ma_i\}$ are placed in the vertices, while the edges represent the polymorphic states $\{na_i, (N-n)a_j\}$. Black and grey arrows distinguish respectively mutation from genetic drift plus selection events. PoMo is a particular case of the four-variate Moran model with boundary mutations and selection, in which the alleles are the four nucleotide bases). The formulas represent the parameter scaling between the effective and the virtual populations. $\mu_{a_i a_j}$ is the mutation rate from allele a_i to a_j; and σ_{a_i} is the selection coefficient of allele a_i on the effective dynamic of population size N; $\mu^*_{a_i a_j}$ and $\sigma^*_{a_i}$ are the corresponding parameters in the virtual dynamic.

Despite the remarkable effort made to account for forces that can derail the species evolutionary history, models of species evolution generally ignore natural selection. While the literature acknowledges the need for selection-aware models of species evolution, the expectation is that most forms of natural selection are not greatly compromising phylogenetic analysis [9]. Two main arguments justify this expectation. The first argument is that selection often occurs at a minority of loci in the genome, in which case any spurious signals would likely be swamped out by the many neutral loci that are sampled [10]. For example, stabilizing selection (i.e., the most common type of selection acting on protein-coding regions), directional selection and selective sweeps are presumed to mildly lower the overall rate of evolution [26] or moderately violate the molecular clock. The second argument is that the forms of selection that most likely affect the species tree (e.g., selection-driven convergent evolution and balancing selection tend to preserve beneficial alleles at a gene for long periods of time) are thought to be rare [6,9]. As corroboration for these expectations, a recent study found that species-specific positive selection has mild effects on phylogeny estimation across a large range of conditions encountered in empirical data sets [1].

However, selection may act in a more pervasive, though weaker, way. Theoretical expectations predict the existence of an intermediate category between the neutral and the selected modes of evolution, known as nearly neutral. The study of variants with small selection coefficients, particularly slightly deleterious mutations, have been the focus of the so-called nearly neutral theory [30]. This theory provided criteria to define the fate of nearly neutral variants, which depends on both selection and random genetic drift. [31] has suggested that if the relative advantage or disadvantage of a particular allele σ is less than twice the reciprocal of the effective population size N (i.e., $N|\sigma| < 2$) the allele's trajectory is nearly neutral. However, more lenient thresholds were also reasoned (e.g., [29]).

Empirical evidence for nearly neutral evolution has become more substantial within the last few years, due to the possibility of sequencing many genomes and multiple individuals. One such example is GC-biased gene conversion (gBGC), a recombination association mutational bias that favors G and C alleles over A and T alleles [12]. gBGC affects the fixation probability of GC alleles and is best

model as selection [28]. Integrative analysis considering the recombination landscape and nucleotide substitution patterns along the genome provide evidence that gBGC acts in both eukaryotes and bacteria [11,22]. In humans, gBGC was estimated one order of magnitude lower than the reciprocal of the effective population size (1.17×10^{-5}) [13], which suggests that gBGC operates in the nearly neutral range. The same was observed true for other apes where most of their exons evolve under weak gBGC ($N\sigma_{GC} < 1$) [21]. In fruit flies, most amino acid replacements have weak signatures of positive selection, though most of the selective effects are nearly neutral, with around 46% of amino acid replacements exhibiting scaled selection coefficients lower than two, and 84% lower than four [35]. Also, a large number of non-synonymous polymorphisms in functionally important sites of human and bacterial populations were shown to segregate at frequencies of around 1–10% (i.e., much more frequent than variants associate to classic Mendelian diseases), suggesting that ongoing purifying selection on protein-coding regions is mainly slightly deleterious [18,17].

As a substantial part of the genome evolves under nearly neutrally, this form of weak, but pervasive, selection has the potential to impact species' evolutionary history. Here, we present PoMos specially tailored to measure nucleotide usage bias throughout the genome and use them to assess the impact of unaccounted-for near neutrality on the species tree estimation. We show that near neutrality significantly blurs inferences of the species tree by biasing the estimation of the species divergence and to a lesser extend the topology, both having implications for molecular dating.

Results and Discussion

Modelling species evolution with nearly neutral selection

PoMos offer a versatile approach to describe species evolution [24]. PoMos add a new layer of complexity to the standard models of sequence evolution by accounting for population-level processes (such as mutations, genetic drift, and selection) to describe sequence evolution. In specific, PoMo expands the 4×4 state-space to model a population of individuals in which changes in allele content and frequency are both possible. The PoMo state-space includes fixed (or boundary) states $\{Na_i\}$, in which all N individuals have the same allele $a_i \in \{A, C, G, T\}$, and polymorphic states $\{na_i, (N-n)a_j\}$, if two alleles a_i and a_j are present in the population with absolute frequencies n and $N-n$ (Figure 1). This version of PoMo does not consider triallelic sites. A simplification that is generally acceptable for eukaryotes, for which mutations occur at a much lower rate than substitutions by genetic drift [25].

Here, we developed two new PoMo models defined on a virtual population: the virtual PoMos. The idea resumes to mimicking a population dynamic that unfolds on the effective population N (estimated at around $10^4 - 10^6$ in multicellular eukaryotes [25]), using a virtual population of smaller size M. M defines a lighter and more efficient state-space. By matching the expected diversity (i.e., the proportion of fixed and polymorphic sites) in both the effective and the virtual populations, one can obtain scaling laws for the mutation rates and selection coefficients [5]. These are intuitive for the neutral case, in which the mutation rates are scaled by the ratio of harmonic numbers of the virtual and effective population sizes (Figure 1 and Supporting Text S1).

Using the smallest virtual population sizes, we defined PoMoTwo (M=2) and PoMoThree (M=3). PoMoTwo has only one polymorphic state per allelic pair and is used to describe species evolution under neutrality (Figure 1). PoMoThree includes two polymorphic states per allelic pair and additionally allows one to infer allelic selection (at least two polymorphic states are necessary to make the selection coefficients identifiable; Figure 1). PoMoThree, in particular, allows quantifying signatures of nearly neutral selection and their impact on species evolution.

We assessed the fit of the virtual PoMos mimicking effective population dynamics by evaluating the absolute error between the fixed states' probability as estimated by the virtual and effective PoMos. The probabilities of the fixed states, differently from the probabilities of the polymorphic ones, can be directly compared among the PoMo models. We checked several scenarios by varying the effective population size, the strength of selection and mutation, and the population's diversity. An average error smaller than 0.5% (Figure S1) was observed for all the scenarios tested, which indicates that virtual PoMos are good approximations of the effective dynamic.

Besides, we compared the computational efficiency of the virtual PoMos to the standard general time-reversible model (GTR, [39]), as implemented in RevBayes [16]. We observed that PoMoTwo requires, on average, four times more cpu hours than the GTR model, whereas, PoMoThree uses nine times more (e.g., data sets of 1 million sites and 50 populations require 42 hours to complete a MCMC chain of 40 000 generations on a standard desktop and without parallelization; Figure S2). Hence, the virtual PoMos can efficiently handle extensive genomic data sets, while being particularly suited to test the impact of weak, but widespread, nearly neutral evolution in species tree estimation.

Ignored nearly neutral evolution affects species tree topology estimation

The impact of unaccounted-for nearly neutral selection on the reconstructed species tree can only be truly known by assessing the correctness of both the topology and the branch lengths estimation. Such aspects can be easily tested using simulated data on a known species tree. We simulated phylogenetic data sets with signatures of nearly neutral, but pervasive, nucleotide usage biases as observed in populations of great apes and fruit flies [4,33,2]. In specific, we mimicked patterns of GC-bias with scaled selection coefficients of around 1.4 for the G and C nucleotides, and 0.0 for A and T nucleotides.

We simulated the A data set with expected heterozygosity of 0.018, which quite well describes the fruit flies' diversity [2]. A second data set, i.e., the B data set, was also created by setting the mutation rates one order of magnitude lower, thus having less segregating polymorphisms. The expected heterozygosity for the B data set is 0.0014, representing the diversity of great ape populations [33].

Regarding the species evolutionary history, we considered four evolving populations, including two closely related populations, a third one recently diverging from the other two, and a fourth one included as an outgroup. We set the evolutionary distances matching an expected divergence per site of 0.025 and 0.23 [33,2] in the B and A data sets, respectively. We produced simulated alignments of 100 to 100 000 sites and 100 repetitions to further assess the methods' precision.

The levels of polymorphism and divergence in the simulated data sets are intended to produce situations of ILS. Directional selection is known to accelerate the fixation process, and fewer polymorphisms are expected to segregate in the population, overall contributing to a reduction of ILS [15]. Hence, phylogenetic inference on data sets generated with selection should produce better estimates of the species tree topology. We investigated the impact of nearly neutral evolution on the topology, by comparing the maximum *a posteriori* (MAP) as estimated by the virtual PoMos and the general time-reversible (GTR, [39]) model of sequence evolution. The MAP topology has sound statistical properties, namely being a consistent estimator of the species topology under Bayesian PoMos. [3].

We evaluated the accuracy of each method via the percentage of correct topologies per data set and alignment size (Figure 2A). Our results show that the virtual PoMos perform better than the GTR in estimating the correct topology in smaller sequence alignments. Such a feature is particularly obvious in polymorphism-rich data sets, in which PoMos reach an accuracy of around 95% for alignments with just 1000 sites, while the GTR method can only reach 60% (A data set panel of Figure 2A). Although the GTR model's accuracy is generally lower than that of PoMos, we observed that it increases with the alignment size, showing that even standard sequence evolution methods can recover the correct topology with genome-scale data. PoMoThree is slightly more accurate than PoMoTwo, but both methods perform equally well as the size of the alignments grows. This suggests that neutral models may already provide fairly good estimates of the species tree topology.

Neglected nearly neutral evolution distorts the measured divergence

Every phylogenetic method aims at providing a realistic description of the divergence process by which populations become species. While more simplistic approaches exist, including the standard substitution models, PoMos account for the convoluted interplay between elemental forces of species evolution. A general expectation is that more elaborate descriptions of the evolutionary process should provide more reliable species tree estimates. We investigated the impact of nearly neutral selection on the branch lengths by comparing the genetic distances as estimated by the virtual PoMos and the standard GTR model. We measured these methods' accuracy by calculating the absolute error between the true and the estimated populations' pair-wise distances. This strategy has the advantage of permitting

comparing distances, even if the underlying topologies are not exactly matching. Normalized distances were used for this comparison, as the GTR and PoMos operate in different evolutionary units: number of substitutions and Moran events, respectively.

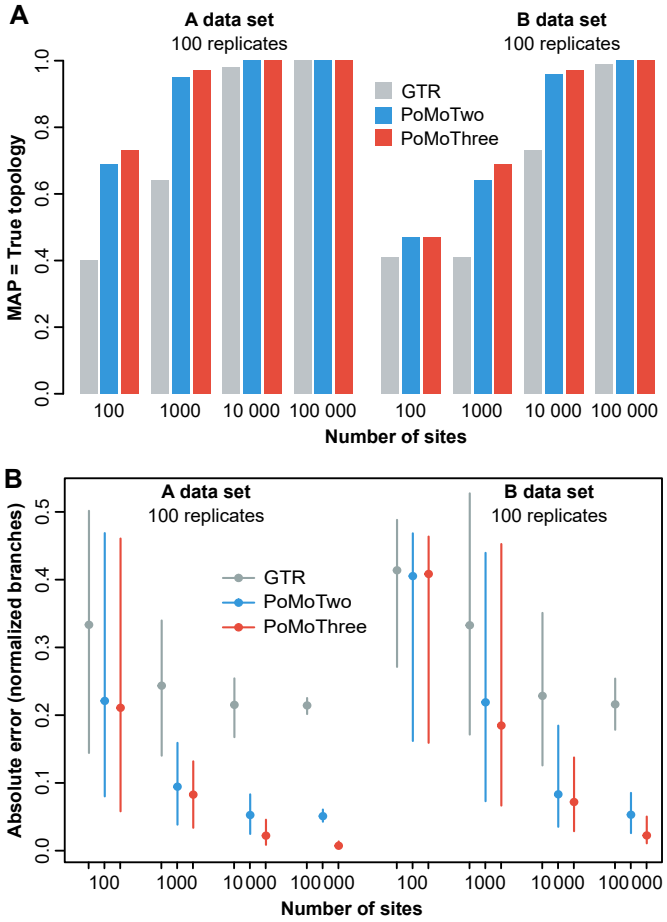

Fig. 2: **Estimating the species tree under nearly neutral selection.** Simulated alignments with different lengths (x-axis) and levels of polymorphism (A and B data sets) were fed to three methods of species tree inference: GTR, PoMoTwo, and PoMoThree, the latter accounting for genome-wide signatures of selection. We investigated the accuracy of these methods regarding the topology **(A)** and the branch lengths **(B)**. The accuracy of the topology estimates was evaluated by the proportion of correctly estimated topologies among the simulated replicates (100 per scenario). The accuracy of the branch lengths was measured by the absolute error between the estimated pair-wise distances among the evolving populations. The branch lengths were normalized, as the three methods employ different evolutionary units. We applied these measures of accuracy to the maximum *a posteriori* tree.

Our results show that the error in the branch length estimations decreases as the size of the alignments increases (Figure 2B). This decay is much more pronounced for the PoMo models than the GTR, showing that polymorphisms significantly improve the divergence process's quantification. Similarly, the polymorphism-rich data set A produces more accurate branch lengths than the B data set, for comparable alignment sizes (Figure 2B). However, PoMoThree is the only method that systematically reduces its error as the number of sites increases. PoMoTwo plateaus with an error of 5% and the GTR of 20% (2B). These results provide evidence that genetic distances neglecting nearly neutral selection provide biased estimators of the divergence process.

To investigate the source of this bias further, we tested whether the placement of the estimated divergence varies across evolutionary scales in nearly neutral regimes by comparing the node heights of the virtual PoMos. We observed that PoMoTwo produces on average smaller node heights than PoMoThree; however, this bias is not uniform across nodes (Figure3A). Our results show that the height of the deeper nodes is more underestimated than that of the shallow ones (the ratio of node heights decreases from the shallow to the deeper nodes; Figure 3A and S3). Such a result suggests that weak selection signatures can mislead neutral models to estimate species histories with more recent divergence events. As expected the rate of evolution is challenging to untie for deeper evolutionary

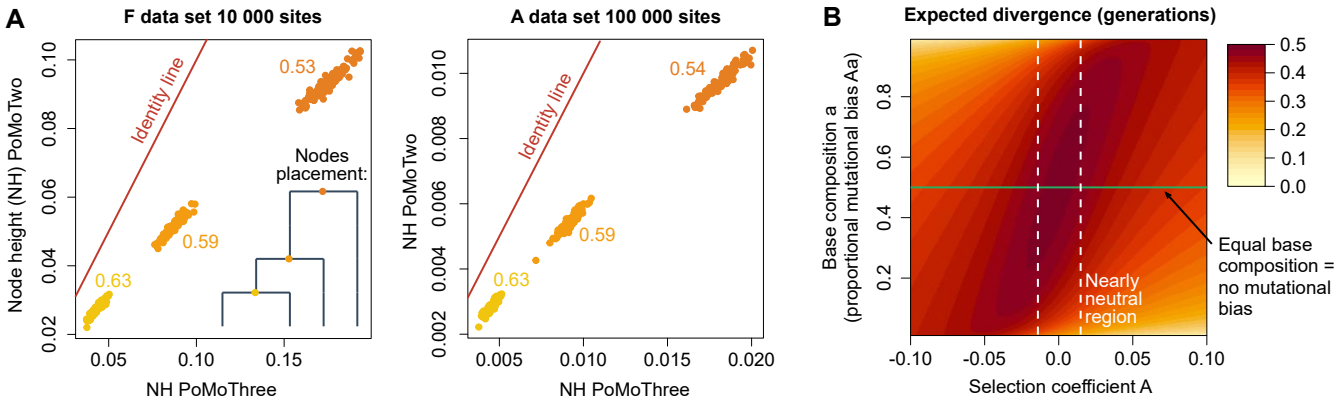

Fig. 3: **Estimating divergence in the presence of nearly neutral selection. (A)** We investigated the association between the node heights as estimated by the virtual PoMos Two and Three, the latter accounting for allelic selection. The phylogenetic tree depicts the three nodes' placement compared in the true phylogeny (the node heights double as we approach the root). Values accompanying each cloud of points denote the ratio of node heights as predicted by PoMoTwo and PoMoThree, respectively; these ratios are always smaller than 1.0 for all the other simulated scenarios (Figure S3). Each cloud represents 100 replicated alignments. **(B)** The expected divergence was calculated for different regimes of selection and mutation. For the biallelic case, the base composition π is proportional to the mutational bias: $\frac{\mu_{Aa}}{\mu_{aA}} = \frac{\pi_a}{\pi_A}$, where μ is the mutation rate. If the base composition is equal, there is no mutational bias (Supporting Text S2).

scales [40]: this can be due to likely violations of the molecular clock hypothesis, but also due to pervasive weak selection as shown by our results.

While our simulations intend to represent realistic scenarios, one might want to know what is the overall impact of selection (not necessarily nearly neutral) on the branch lengths. An advantage of PoMos is that some quantities can be formally obtained; the expected divergence per generation in diverse mutation-selection regimes is one of such cases (Figure 3B and Supporting Text S3).

We observe that the expected divergence is generally higher for the neutral case, and tends to decrease as selection becomes more intense (Figure 3B; such trend is not crucially affected by the effective population size as shown in Figure S4). Directional selection fades alleles that could potentially be drifting in the population to fixation, overall reducing the expected frequency changes and hence decreasing the expected divergence. Consequently, the same patterns of diversity lead selection-aware models to return longer branch lengths than the neutral models, corroborating the results obtained in figure 3A. Furthermore, the magnitude of biases on the estimated branch lengths is strongly determined by the underlying population dynamic. For instance, the expected divergence is generally higher when selection counteracts mutation (rarer alleles are favored and *vice versa*, as it is the case witH gBGC) than when selection and the mutational bias act concordantly (rarer alleles are disfavoured and *vice versa*; 3B).

Molecular dating with fruit fly populations

The geographic origins of the globally distributed fruit fly (*Drosophila melanogaster*) are still not fully understood. This is certainly not due to the lack of genomic data, as hundreds sequenced genomes are available for the fruit flies (e.g., [14]), but mostly due to the methods of species tree estimation not scaling with multiple populations and genome-wide data. We employed the virtual PoMos to estimate and date the evolutionary history of 31 fruit fly populations. We used 966 sequences (Table S1) from 1 million distantly located genomic sites, acknowledging the sites independence assumption. We also assumed a global clock for molecular dating analyses, which is generally valid for short time scales. We set an uniform prior of 60 000 $\pm$ 15 000 years dating the African population expansion, as is suggested in the field's literature (e.g., [23]).

The phylogenies estimated by the virtual PoMos are agreeing for most of the major clades (Figures 4, S6 and S7), overall supporting the same phylogeographic history for the fruit fly populations. As expected, the fruit fly populations have flourished in the southeast of the African continent. Next, the species spread to the north, establishing in western equatorial Africa. From there, fruit flies colonized the western equatorial and the northern regions of Africa, with the northern wave leading to the spread of fruit flies worldwide. This phylogeography is not captured by the GTR model (Figure S5), suggesting high levels of ILS among fruit flies.

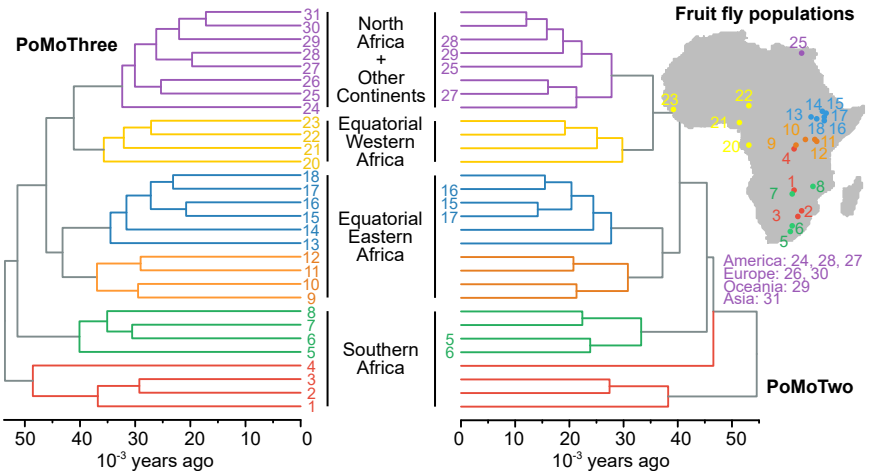

Fig. 4: **Dating the evolutionary history of 31 worldwide fruit fly populations.** We dated the evolutionary history of 31 worldwide populations with 4 to 205 individuals (966 sequences in total, Table S1) of *Drosophila melanogaster* distributed across the globe (data taken from [14]). The phylogenetic analyses were employed in RevBayes [16], using alignments of 1 million distantly located genomic sites and the PoMoTwo and PoMoThree models of evolution, the latter being selection-aware. The few populations named at the tips of the PoMoTwo tree indicate disagreement between the two models. The maximum *a posteriori* tree is depicted in both cases. Estimated node ages can be found in Figures S5-S7.

The time inferences are less consistent between the virtual PoMos, with PoMoThree providing, as expected by our formal results and simulations, more ancient node ages (around 21%; Figures 4, S6 and S7). Our divergence times are also comparatively more ancient than other studies. For example, some studies suggest a split of the African lineage at about 16 000 years ago [23], while our analyses estimated it around 32 000 years ago. While unaccounted signatures of selection can underestimate the divergence times, we also believe that the non-monophyly of the non-African fruit fly populations (e.g., the Egyptian population (number 25 in Figure 4) branches within the European clade) can cause such volatility in the estimated divergence times among studies.

An additional advantage of using selection-aware methods is that we can quantify selection coefficients. We measured the fruit flies' GC-bias rate at around 5.35×10^{-7} per site per generation (Figure S8B), one order of magnitude lower than their effective population size and matching the expectations of gBGC. It remains unclear whether fruit flies have gBGC [34]. While it is difficult to think of any other process than gBGC favoring G and C alleles genome-wide, further studies associating these signatures with the recombination landscape are still needed.

Conclusions

Methods of species tree estimation have become an essential tool in studies of molecular evolution. The models presented here represent an important extension of existing approaches by allowing selection and, therefore, more realistic descriptions of the divergence process shaping sequence evolution. We have shown evidence that unaccounted-for low magnitude, but pervasive, selection underestimates genetic distances, with the older nodes being more affected than the young ones. Therefore, selection-aware models of species evolution are crucial if the phylogenetic analyses aim to quantify divergence or dating species evolution.

Materials and Methods

Performance of the virtual PoMos

The error of virtual PoMos was calculated using the `expokit` package in R language to exponentiate the rate matrix of the real process (i.e., with larger N). We tested several scenarios including varying the effective population size, elapsed time, and the strength of selection and mutation (more details in Supporting Text S1). The virtual PoMos are implemented in RevBayes and can be employed using the functions fnPoMoTwo and fnPoMoThree, which can be found in the devPoMo branch of the RevBayes GitHub project.

Simulations

To test our algorithm, we simulated data using a four-variate Moran model with mutation and allelic selection. The simulated data was based on a phylogenetic tree of four populations. The relative genetic distances were considered as such `(((pop4:0.25,pop3:0.25):0.25,pop2:0.5):0.5,pop1:1);` so that the expected divergence was 0.23 and 0.025 substitutions per site for B and A data set respectively. The mutation and selection rates were set after [4]. In particular, we considered a base frequencies of $\pi = (0.37, 0.12, 0.15, 0.36)$ for the A, C, G and T nucleotides, the exchangeabilities as $N\rho = (0.005, 0.04, 0.001, 0.01, 0.04, 0.005)$ (and one order of magnitude higher for the A data set; exchangeabilities in the following order AC, AG, AT, CG, CT and GT) and the selection coefficients of A, C, G and T as such $N\sigma = (0.0, 1.4, 1.4, 0.0)$. We simulated two data sets, A and B, by mimicking fast and slowly evolving populations with an expected heterozygosity (0.018 and 0.0014) as observed in *Drosophila simulans* populations and among great apes, respectively [33,2]. We produced alignments of 100 to 100 000 sites for each data set and 100 repetitions.

Application to fruit fly populations

Genome-wide data from 31 world-wide distributed populations of *Drosophila melanogaster* with 4 to 205 individuals (966 individuals) were extracted from the PopFly database [14]. One million sites were used to perform molecular dating analyses in RevBayes (data sets S1-S3). We run the GTR and the Virtual PoMos considering a global clock model and an uniform prior of 60 000 $\pm$ 15 000 dating the African population expansion, following [23] (more details are provided in Supporting Text S3).

Acknowledgements

The molecular dating analyses of the fruit fly data set have been achieved using the Vienna Scientific Cluster (VSC). This work was supported by the Vienna Science and Technology Fund (WWTF) [MA16-061]. GJSz received funding from the European Research Council under the European Union's Horizon 2020 research and innovation program under grant agreement no. 714774 and the grant GINOP-2.3.2.-15-2016-00057.

References

1. R. H. Adams, D. R. Schield, D. C. Card, and T. A. Castoe. Assessing the Impacts of Positive Selection on Coalescent-Based Species Tree Estimation and Species Delimitation. *Systematic Biology*, 67(6):1076–1090, nov 2018.
2. D. J. Begun, A. K. Holloway, K. Stevens, L. W. Hillier, Y.-P. Poh, M. W. Hahn, P. M. Nista, C. D. Jones, A. D. Kern, C. N. Dewey, L. Pachter, E. Myers, and C. H. Langley. Population Genomics: Whole-Genome Analysis of Polymorphism and Divergence in Drosophila simulans. *PLoS Biology*, 5(11):e310, nov 2007.
3. R. Borges and C. Kosiol. Consistency and identifiability of the polymorphism-aware phylogenetic models. *Journal of Theoretical Biology*, 486:110074, feb 2020.
4. R. Borges, G. J. Szöllősi, and C. Kosiol. Quantifying GC-Biased Gene Conversion in Great Ape Genomes Using Polymorphism-Aware Models. *Genetics*, 212(4):1321–1336, aug 2019.
5. R. Borges, G. J. Szöllősi, and C. Kosiol. Quantifying GC-Biased Gene Conversion in Great Ape Genomes Using Polymorphism-Aware Models. *Genetics*, 212(4):1321–1336, aug 2019.
6. T. A. Castoe, A. P. J. de Koning, H.-M. Kim, W. Gu, B. P. Noonan, G. Naylor, Z. J. Jiang, C. L. Parkinson, and D. D. Pollock. Evidence for an ancient adaptive episode of convergent molecular evolution. *Proceedings of the National Academy of Sciences*, 106(22):8986–8991, jun 2009.
7. N. De Maio, C. Schlötterer, and C. Kosiol. Linking great apes genome evolution across time scales using polymorphism-aware phylogenetic models. *Molecular Biology and Evolution*, 30(10):2249–2262, 2013.
8. N. De Maio, D. Schrempf, and C. Kosiol. PoMo: An allele frequency-based approach for species tree estimation. *Systematic Biology*, 64(6):1018–1031, nov 2015.
9. S. V. Edwards. Natural selection and phylogenetic analysis. *Proceedings of the National Academy of Sciences*, 106(22):8799–8800, jun 2009.
10. S. V. Edwards, Z. Xi, A. Janke, B. C. Faircloth, J. E. McCormack, T. C. Glenn, B. Zhong, S. Wu, E. M. Lemmon, A. R. Lemmon, A. D. Leaché, L. Liu, and C. C. Davis. Implementing and testing the multispecies coalescent model: A valuable paradigm for phylogenomics. *Molecular Phylogenetics and Evolution*, 94:447–462, jan 2016.
11. N. Galtier, L. Duret, S. Glémin, and V. Ranwez. GC-biased gene conversion promotes the fixation of deleterious amino acid changes in primates, 2009.

12. N. Galtier, G. Piganeau, D. Mouchiroud, and L. Duret. GC-content evolution in mammalian genomes: the biased gene conversion hypothesis. *Genetics*, 159(2):907–11, oct 2001.
13. S. Glémin. Surprising Fitness Consequences of GC-Biased Gene Conversion: I. Mutation Load and Inbreeding Depression. *Genetics*, 185(3):939–959, jul 2010.
14. S. Hervas, E. Sanz, S. Casillas, J. E. Pool, and A. Barbadilla. PopFly: the Drosophila population genomics browser. *Bioinformatics*, 33(17):2779–2780, sep 2017.
15. A. Hobolth, J. Y. Dutheil, J. Hawks, M. H. Schierup, and T. Mailund. Incomplete lineage sorting patterns among human, chimpanzee, and orangutan suggest recent orangutan speciation and widespread selection. *Genome Research*, 21(3):349–356, mar 2011.
16. S. Höhna, M. J. Landis, T. A. Heath, B. Boussau, N. Lartillot, B. R. Moore, J. P. Huelsenbeck, and F. Ronquist. RevBayes: Bayesian Phylogenetic Inference Using Graphical Models and an Interactive Model-Specification Language. *Systematic Biology*, 65(4):726–736, jul 2016.
17. A. L. Hughes. Evidence for Abundant Slightly Deleterious Polymorphisms in Bacterial Populations. *Genetics*, 169(2):533–538, feb 2005.
18. A. L. Hughes, B. Packer, R. Welch, A. W. Bergen, S. J. Chanock, and M. Yeager. Widespread purifying selection at polymorphic sites in human protein-coding loci. *Proceedings of the National Academy of Sciences*, 100(26):15754–15757, dec 2003.
19. J. Kingman. The coalescent. *Stochastic Processes and their Applications*, 13(3):235–248, sep 1982.
20. J. F. C. Kingman. On the genealogy of large populations. *Journal of Applied Probability*, 19(A):27–43, jul 1982.
21. N. Lartillot. Phylogenetic Patterns of GC-Biased Gene Conversion in Placental Mammals and the Evolutionary Dynamics of Recombination Landscapes. *Molecular Biology and Evolution*, 30(3):489–502, mar 2013.
22. F. Lassalle, S. Périan, T. Bataillon, X. Nesme, L. Duret, and V. Daubin. GC-Content Evolution in Bacterial Genomes: The Biased Gene Conversion Hypothesis Expands. *PLOS Genetics*, 11(2):e1004941, feb 2015.
23. S. J. Laurent, A. Werzner, L. Excoffier, and W. Stephan. Approximate Bayesian Analysis of Drosophila melanogaster Polymorphism Data Reveals a Recent Colonization of Southeast Asia. *Molecular Biology and Evolution*, 28(7):2041–2051, jul 2011.
24. A. D. Leaché and J. R. Oaks. The Utility of Single Nucleotide Polymorphism (SNP) Data in Phylogenetics. *Annual Review of Ecology, Evolution, and Systematics*, 48(1):69–84, nov 2017.
25. M. Lynch, M. S. Ackerman, J.-F. Gout, H. Long, W. Sung, W. K. Thomas, and P. L. Foster. Genetic drift, selection and the evolution of the mutation rate. *Nature Reviews Genetics*, 17(11):704–714, nov 2016.
26. K. M. *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge, 1983.
27. W. P. Maddison and L. L. Knowles. Inferring Phylogeny Despite Incomplete Lineage Sorting. *Systematic Biology*, 55(1):21–30, feb 2006.
28. T. Nagylaki. 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, 1983.
29. M. Nei. Selectionism and Neutralism in Molecular Evolution. *Molecular Biology and Evolution*, 22(12):2318–2342, aug 2005.
30. T. OHTA. Slightly Deleterious Mutant Substitutions in Evolution. *Nature*, 246(5428):96–98, nov 1973.
31. T. Ohta. Near-neutrality in evolution of genes and gene regulation. *Proceedings of the National Academy of Sciences*, 99(25):16134–16137, dec 2002.
32. D. A. Pollard, V. N. Iyer, A. M. Moses, and M. B. Eisen. Widespread Discordance of Gene Trees with Species Tree in Drosophila: Evidence for Incomplete Lineage Sorting. *PLoS Genetics*, 2(10):e173, 2006.
33. J. Prado-Martinez, P. H. Sudmant, J. M. Kidd, H. Li, J. L. Kelley, B. Lorente-Galdos, K. R. Veeramah, A. E. Woerner, T. D. O'Connor, G. Santpere, A. Cagan, C. Theunert, F. Casals, H. Laayouni, K. Munch, A. Hobolth, A. E. Halager, M. Malig, J. Hernandez-Rodriguez, I. Hernando-Herraez, K. Prüfer, M. Pybus, L. Johnstone, M. Lachmann, C. Alkan, D. Twigg, N. Petit, C. Baker, F. Hormozdiari, M. Fernandez-Callejo, M. Dabad, M. L. Wilson, L. Stevison, C. Camprubí, T. Carvalho, A. Ruiz-Herrera, L. Vives, M. Mele, T. Abello, I. Kondova, R. E. Bontrop, A. Pusey, F. Lankester, J. A. Kiyang, R. A. Bergl, E. Lonsdorf, S. Myers, M. Ventura, P. Gagneux, D. Comas, H. Siegismund, J. Blanc, L. Agueda-Calpena, M. Gut, L. Fulton, S. A. Tishkoff, J. C. Mullikin, R. K. Wilson, I. G. Gut, M. K. Gonder, O. A. Ryder, B. H. Hahn, A. Navarro, J. M. Akey, J. Bertranpetit, D. Reich, T. Mailund, M. H. Schierup, C. Hvilsom, A. M. Andrés, J. D. Wall, C. D. Bustamante, M. F. Hammer, E. E. Eichler, and T. Marques-Bonet. Great ape genetic diversity and population history. *Nature*, 499(7459):471–475, jul 2013.
34. M. C. Robinson, E. A. Stone, and N. D. Singh. Population Genomic Analysis Reveals No Evidence for GC-Biased Gene Conversion in Drosophila melanogaster. *Molecular Biology and Evolution*, 31(2):425–433, feb 2014.
35. S. A. Sawyer, J. Parsch, Z. Zhang, and D. L. Hartl. Prevalence of positive selection among nearly neutral amino acid replacements in Drosophila. *Proceedings of the National Academy of Sciences*, 104(16):6504–6510, apr 2007.
36. D. Schrempf, B. Q. Minh, N. De Maio, A. von Haeseler, and C. Kosiol. Reversible polymorphism-aware phylogenetic models and their application to tree inference. *Journal of Theoretical Biology*, 407:362–370, 2016.
37. D. Schrempf, B. Q. Minh, A. von Haeseler, and C. Kosiol. Polymorphism-Aware Species Trees with Advanced Mutation Models, Bootstrap, and Rate Heterogeneity. *Molecular Biology and Evolution*, 36(6):1294–1301, jun 2019.
38. G. J. Szöllősi, E. Tannier, V. Daubin, and B. Boussau. The Inference of Gene Trees with Species Trees. *Systematic Biology*, 64(1):e42–e62, jan 2015.
39. S. Tavaré. Some Probabilistic and Statistical Problems in the Analysis of DNA Sequences. *Lectures on Mathematics in the Life Sciences*, 17:57–86, 1986.
40. B. Xu and Z. Yang. Challenges in Species Tree Estimation Under the Multispecies Coalescent Model. *Genetics*, 204(4):1353–1368, dec 2016.