Traditional phylogenetic models are insensitive to variations in the effective population size

Rui Borges[1*], Ioanna Kotari[1,2], Juraj Bergman[3], Madeline A. Chase[4], Carina F. Mugal[4,5], Carolin Kosiol[6]

1. Institut für Populationsgenetik, Vetmeduni Vienna, 1210, Vienna, Austria
2. Vienna Graduate School of Population Genetics, Vienna, Austria
3. Center for Biodiversity Dynamics in a Changing World (BIOCHANGE) & Section for Ecoinformatics and Biodiversity, Department of Biology, Aarhus University, DK-8000 Aarhus C, Denmark
4. Department of Ecology and Genetics, Uppsala University, Uppsala, Sweden
5. Laboratory of Biometry and Evolutionary Biology, University of Lyon 1, CNRS, UMR 5558, Villeurbanne, France
6. Centre for Biological Diversity, University of St Andrews, St Andrews, Fife, UK

* corresponding author: ruiborges23@gmail.com

Abstract

A substitution represents the emergence and fixation of an allele in a population or species and is the fundamental event from which phylogenetic models of sequence evolution are devised. As a result of the increasing availability of genomic sequences, we can currently leverage interspecific variability while reconstructing the tree of life. Substitutions can thus more realistically be modeled as the product of mutation, selection, and genetic drift. However, it remains unclear whether this additional complexity affects our measures of evolutionary times and rates. This study seeks to answer this question by contrasting the traditional substitution model with a population genetic counterpart by using data from 4385 individuals distributed across 179 populations and representing 17 species of animals, plants, and fungi. We found that when the population genetics dynamic is modeled via the substitution rates, the evolutionary times and rates measured by these models are well correlated, suggesting that the substitution models are able to capture the time and pace of their population counterpart. However, a closer inspection of this result showed that the traditional models largely ignore the effect of population size, even when explicitly accounted for in the substitution rates, while the effects of mutation and selection are well captured. Our results indicate that superimposing population-genetics results on the substitution rates may be efficiently used to study mutation and selection biases, while other data sources (e.g., life history traits or polymorphisms) need to be additionally integrated to infer variations in the population size. We expect our results to help develop mutation-selection phylogenetic models and inspire unified frameworks between large and short evolutionary time scales.

Keywords: models of sequence evolution, Moran model, effective population size, GC-bias, selection and mutation bias

Introduction

Models of sequence evolution have a long tradition in phylogenetics. The JC (Jukes and Cantor, 1969), HKY (Hasegawa et al., 1985), and GTR (Tavaré, 1986) are just a few examples of substitution models that are widely used in phylogenetic applications. These models describe the evolution of nucleotides, codons, or amino acids over time, using the substitution process as their baseline event. A substitution represents the sequence of events by which the allelic content of a specific locus in a population is completely replaced. As a result, substitution models are generally quite vague in terms of the processes they may represent but quite flexible, as they implicitly account for any mutation, selection, or fixation biases leading to the fixation of any new allele.

Substitution models are especially useful when genomic resources are limited, and only one genomic sequence per species is available; as such, one can only aim to characterize interspecific variation. However, due to remarkable advances in sequencing technologies in recent years, we now have access to genomic data across a wide range of species that allow for the characterization of intraspecific diversity, i.e., polymorphisms. This type of variation, which lies at the core of population genetics, has also proven to be highly informative for tree reconstruction (Leaché and Oaks, 2017). A classic example is the contribution of polymorphisms in correcting species trees for the effect of incomplete lineage sorting (i.e., the persistence of polymorphisms over several speciation events) (Maddison and Knowles, 2006).

The attempt to incorporate inter- and intra-specific variation into phylogenetic inference resulted in the development of models that, while describing the substitution process, aim to more generally describe how genetic variation changes over time. To do so, they employ a mechanistic description of well-known population forces such as population size variations, mutation rate, selective pressures, and so forth, typically based on the Moran (Moran, 1958) or Wright-Fisher (Fisher, 1930; Wright, 1931) models or the Kingman's coalescent (Kingman, 1982). Examples of such approaches are the multispecies coalescent (Rannala and Yang, 2003) and the polymorphism-aware phylogenetic models [PoMos; De Maio et al. (2013, 2015)]. Guided by empirical studies, these models are becoming more realistic, accounting for other essential processes such as migration (Flouri et al., 2020) or nucleotide usage biases (Borges et al., 2019b).

As more genomic sequences become available, it is now possible to account for polymorphisms and the processes that govern their dynamic when estimating and dating the tree of life. In this process, many clades and divergence times will undoubtedly be revisited and revised. While it is obvious that different evolutionary models provide different interpretations of evolutionary times and rates, it is still unclear whether population-genetics and phylogenetic narratives are consistent in general. Determining the evolutionary significance of population processes playing a role in sequence evolution will further help us to understand which are the ones that are likely to be neglected and which are already well captured by the traditional approaches.

Here, we used population genomic data from abundantly sequenced biologically diverged taxa across animals, plants, and fungi to compare the traditional substitution model with a population-genetics counterpart that accounts for mutation bias, selection, and genetic drift. Specifically, we compare these models in terms of divergence rates, substitution times, and allele composition. The main goal of this study is to determine the evolutionary contexts where these models express similar time and pace, envisaging possible avenues for integrative approaches at phylogenetic and population-genetics time scales.

Results and Discussion

Modeling the substitution process with two different levels of detail

The substitution process can be modeled at various levels of complexity. In this study, we will concentrate on two different models: a simpler one inspired by phylogenetic models of sequence evolution (which we will refer to as M_1 or substitution model) and a more complex one that takes into account some population-level forces that are likely involved in the substitution process (which we will refer to as M_2 or Moran model). In both cases, the substitution process is modeled as a Markov chain that describes the evolutionary dynamic between two alleles that are present at the same locus in a population or species. A valuable property of both models is that we can easily derive closed-form expressions for the stationary distribution, divergence rate, and substitution time, quantities that we here used to compare these models.

Let us consider a model M_1 with only two states, S and W, and describing the substitution process with only two probabilities: λ_{WS} is the probability of a $W \to S$ substitution and *vice versa* for λ_{SW} (figure 1A). The model M_2 describes the change in the frequencies of S and W over time in a population of size N (figure 1B). In this model, we consider the existence of fixed states, in which all N individuals have the same allele, as well as polymorphic states, in which the alleles S and W coexist at different frequencies. To describe the substitution process in this population, we use a Moran model with boundary mutations and selection.

The probabilities of a mutation from W to S and S to W are represented by μ_{WS} and μ_{SW}, respectively. Mutation rates are assumed to be boundary (or low), biased, and reversible: low, so that by the time a new mutation occurs, the previous one has been fixed [this is a reasonable assumption for most eukaryotes (Lynch et al., 2016)]; biased, because μ_{WS} and μ_{SW} may differ; reversible, because multiple mutations per site are permitted. The Moran model models genetic drift by assuming that one random individual is chosen to reproduce and one random individual is chosen to die at each time step (Moran, 1958). These probabilities depend on the population size N, which defines the strength of genetic drift, and the selection coefficient σ_S, which confers a reproductive advantage to individuals carrying an allele S. σ_S relates to the fitness coefficient via the equation $\phi_S = 1 + \sigma_S$. Four-allelic versions of this model have been previously used for phylogenetic analyses (Borges et al., 2019a; Borges and Kosiol, 2020; Borges et al., 2022).

Even though both models allow for the study of the substitution process, they do so in different ways. There are, however, some expectations that can be drawn in order to make them comparable. We assume that the composition of fixed (or monomorphic) states S and W to be the same in both models. This is equivalent to the common assumption made in phylogenetic models, in which each genomic position is assumed fixed in the species/population. In such a situation, we should expect the frequencies of the fixed states 0 and N in the M_2 model to match those of the states W and S in M_1, as both represent fixed sites. If s is the stationary distribution of these models, we expected that

$$s_W^{M_1} = \frac{s_0^{M_2}}{s_0^{M_2} + s_N^{M_2}} \quad \text{and} \quad s_S^{M_1} = \frac{s_N^{M_2}}{s_0^{M_2} + s_N^{M_2}} \ , \tag{1}$$

where the denominator normalizes the stationary frequencies of the fixed states in M_2, so they sum up to one (the M_2 also accounts for polymorphic states; text S1). By substituting the stationary frequencies in equation (1), we obtain that the substitutions rates can be expressed by

$$\lambda_{SW} \propto \mu_{SW} \quad \text{and} \quad \lambda_{WS} \propto \mu_{WS}(1+\sigma_S)^{N-1} \ . \tag{2}$$

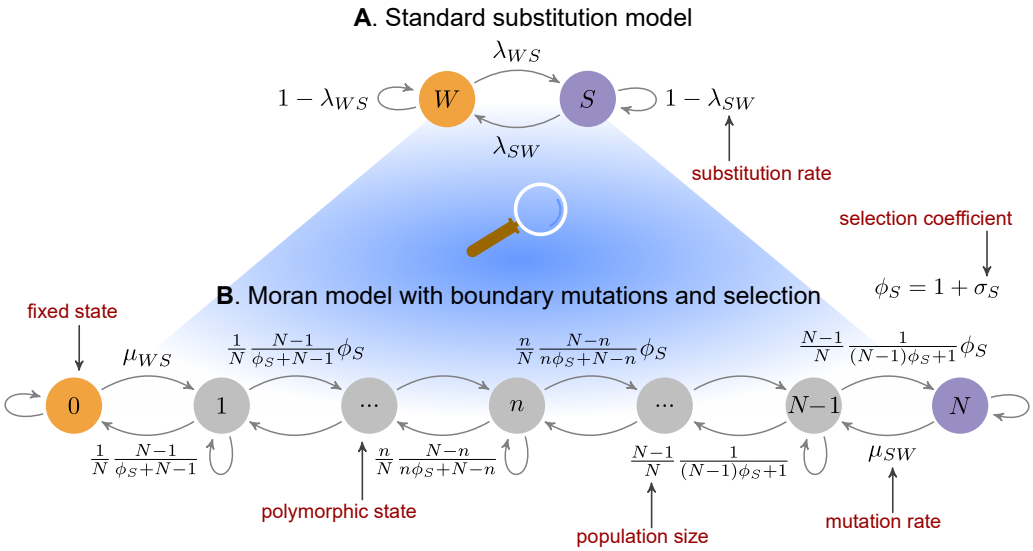

Figure 1: **Two possible models for the substitution process (A)** A simple substitution model inspired by the phylogenetic models of sequence evolution. The two alleles W and S interchange each other with substitution probabilities λ_{WS} and λ_{SW}. **(B)** A Moran model with boundary mutations and selection. The states of this model represent the frequency of the S allele in a population of N individuals: 0 and N represent fixed states, while the remaining ones are polymorphic. The probabilities of a mutation from W to S and S to W are respectively denoted by μ_{WS} and μ_{SW}. Under the Moran model, the probability of choosing one of the alleles to die and another to reproduce is $\frac{n}{N}\frac{N-n}{N}$. This ratio is, in this scheme, modified to account for directional selection, where σ_S is the selection coefficient of the S allele; we assume $\sigma_W = 0$. For the sake of clarity, the probabilities of remaining in the same state were not depicted.

If we further assume that there is no selection ($\sigma_S = 0$), we obtain that the substitution rates are equal to the mutation rates, which is a well-known expectation of the neutral theory of evolution (Kimura, 1968).

The identity depicted in equation (2) is yet another example of the use of population genetics-informed parameters in phylogenetic models. Similar strategies have been used by several authors, such as Halpern and Bruno (1998); Robinson et al. (2003) in developing the first mutation-selection codon models or Lartillot (2013) to study the dynamic of GC-biased gene conversion (gBGC) across mammals. In the following sections, we used identity (2) to redefine the substitution rates of model M_1, making it thus dependent on the same evolutionary processes that govern the M_2 dynamic.

The evolutionary dynamic between strong and weak nucleotides

To make our model comparisons more biologically insightful, we used population parameters estimated from empirical data. We concentrate on the dynamic of strong (S: G or C) and weak (W: A or T) nucleotides. There were two reasons for selecting such a system. First, it is hypothesized that the composition of strong and weak nucleotides is regulated throughout the genome by two antagonistic processes: while mutations are generally found to be biased towards weak nucleotides (Keightley et al., 2009; Lynch, 2010b), another molecular process, known as gBGC, acts in the opposite direction. gBGC is a recombination-associated process that induces a transmission distortion towards strong nucleotides at WS heterozygote sites, which ultimately promotes the fixation of strong alleles (Duret and Galtier, 2009). Second, this nucleotide dynamic is thought to act on a genome-wide scale and has been observed across a wide range of taxa (Pessia et al., 2012; Galtier et al., 2018). gBGC is therefore exceptionally well suited for testing the effects of the population-genetics dynamic at larger time scales.

Under the weak-strong allelic system, μ_{WS} and μ_{SW} are the mutation rates from weak to strong and

strong to weak alleles, respectively. gBGC is typically modeled equivalent to directional selection (Nagylaki, 1983). Thus, σ_S is the selection coefficient of strong nucleotides or the GC-bias preference. These parameters, along with the effective population size N, were estimated using allele counts from various genomic positions, assuming that a sample of M individuals was drawn from a population of N individuals with a mutation-selection dynamic close to or at the steady-state (we assumed the stationary frequencies of model M_2; text S1). The inferences were conducted using a Bayesian framework, which estimates the population parameters described above using empirically obtained mutation rates. We created a program called BESS (Bayesian Estimator of the Site-frequency Spectrum) that performs the aforementioned steps. More information on BESS and the inferences can be found in the Material and Methods section.

Our data set includes 179 populations from six different taxa, including great apes, mice, flycatchers, fruit flies, Arabidopsis, and baker's yeast (tables 1 and S1). These taxa were chosen because multi-individual and genome-wide data are available and are particularly well-characterized in terms of their intraspecific variation. An average of 145 million genomic sites (ranging from 150 thousand to 26 million sites) were used for the inferences, with approximately 26 (ranging from 2 to 243) individuals sampled per population. It was our intention to cover a wide range of population dynamics with variable population sizes, mutation rates, and GC preferences. As figure 2 shows, all of the population parameters vary significantly within and between taxa. To better assess the magnitude of these forces, the mutation rates and the selection coefficient are scaled with the population size.

Table 1: **Population genomic data used in this study.**

Taxa	**Species and no. of populations**	**Type of genomic sites and source**	**Mutation rate and source**
Great apes	*Pongo pygmaeus* (1), *P. abelii* (1), *Gorilla gorilla* (3), *Pan troglodytes* (4), *P. paniscus* (1) and *Homo sapiens* (1)	Four-fold degenerate sites (Bergman and Schierup, 2021)	25×10^{-9} (Nachman and Crowell, 2000)
Mice	*Mus musculus* (8) and *Mus spretus* (1)	Whole genome (Harr et al., 2016)	118.0×10^{-9} (Lynch, 2010a)
Flycatchers	*Ficedula albicollis* (1), *F. hypoleuca* (1), *F. parva* (1), *F. albicilla*, *F. speculigera* and *F. semitorquata* (1)	Four-fold-degenerate sites (Burri et al., 2015; Chase and Mugal, 2022)	4.6×10^{-9} (Smeds et al., 2016)
Fruit flies	*Drosophila melanogaster* (30)	whole genome (Hervas et al., 2017)	3.5×10^{-9} (Keightley et al., 2009)
Arabidopsis	*Arabidopsis thaliana* (37)	coding regions (Initiative, 2000)	7.0×10^{-9} (Ossowski et al., 2010)
Baker's yeast	*Saccharomyces cerevisiae* (86)	whole genome (Peter et al., 2018)	0.3×10^{-9} (Lynch et al., 2008)

Strong to weak base mutations were found to be more common than weak to strong base mutations, which is consistent with other studies (Keightley et al., 2009; Lynch, 2010b) (figure 2A–B). We also observed that the selection coefficients of the S allele varied greatly between taxa and populations, being quite pronounced in great apes, mice, and flycatchers but almost nonexistent in fruit flies (figure 2C). Our results are supported by studies showing that the composition of weak and strong bases is driven by gBGC in great apes (Borges et al., 2019a), mice (Galtier, 2021), flycatchers (Bolívar et al., 2016), Arabidopsis (Günther et al., 2013) and baker's yeast (Lesecque et al., 2013), but not in fruit flies (Robinson et al., 2014). However, our results also showed that selection in favor of strong bases varies considerably among taxa and correlates negatively with the effective population size ($\rho = 0.79$ p–value<0.001; figure S3). Since the recombination landscape is remarkably conserved between closely-related populations and species, we attribute this variation to nonequilibrium conditions related to fluctuating demography or gBGC (we note that our model assumes that both processes are constant) (Brandvain and Wright, 2016; Müller et al., 2022). In any case, the estimated GC preferences appear to operate in the nearly neutral range for all taxa (i.e., $N\sigma_S < 1$), as observed in other studies (Glémin et al., 2015; Bolívar et al., 2019).

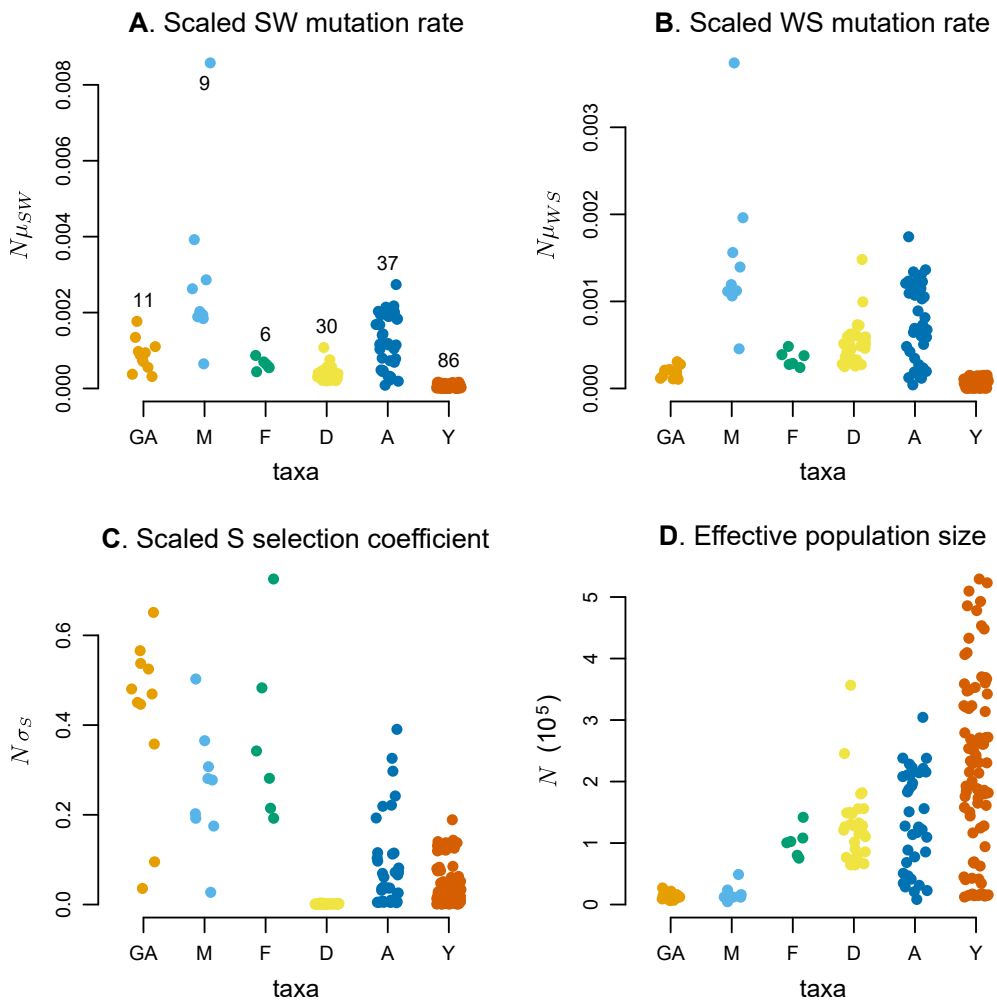

Figure 2: **Population estimates of the forces at work in the weak vs. strong allele dynamic.** To better assess the magnitude of these forces, the mutation rates and selection coefficient were scaled with population size. The parameter estimates were obtained with BESS (see Material and Methods for more details) using allelic counts from population genomic data. The numbers depicted in panel A represent the number of populations per taxa. Legend: great apes (GA), mice (M), flycatchers (F), fruit flies (D), arabidopsis (A), and baker's yeast (Y).

The population size varies significantly among taxa, ranging from around 4 500 individuals in great apes to about 530 000 in yeasts (figure 2D). Other studies have estimated far larger population sizes for fruit flies and yeast in the order of millions of individuals [e.g., Hodgins-Davis et al. (2015)]. These differences may be attributed to different model assumptions. Concretely, we account for reversible and biased mutations as well as fixed sites, unlike many estimators that assume the infinite sites model and rely solely on polymorphic sites. Fixed sites are by far the most common type among the genomic positions sampled, and we expect them to dominate our estimates of mutation rates, selection coefficients, and effective population sizes. We further note that our estimates of mutation rates, selection coefficients, and population sizes vary independently of the different types of genomic positions used to perform the inferences (whole genome vs. coding sequences; table 1).

Comparing the pace and time of traditional and population-genetics models

Divergence rates express the speed at which genetic differences arise in a given lineage over time. Evolutionary times are usually relative to the process of evolution and are posteriorly translated into calendar times using fossil calibrations or generation times. Together, these quantities represent an important

aspect of evolutionary studies and are extensively used in comparative genomics to establish organismal relationships and estimate divergence times. We compared the models M_1 and M_2 in a rate and time dimension by making use of their expected divergence rate and substitution time.

The divergence rate in the substitution model M_1 simply expresses the expected number of substitutions per unit of time. Differently, in the Moran model M_2, it represents the expected number of Moran events (including both mutations and frequency shifts) per unit of time. These quantities can be obtained formally (text S1) if the populations are at or near the steady-state:

$$d_{M_1} = 2\frac{\lambda_{WS}\lambda_{SW}}{\lambda_{SW}+\lambda_{WS}} \quad \text{and} \quad d_{M_2} = 2\mu_{WS}\mu_{SW}\frac{1}{k}\sum_{n=0}^{N-1}(1+\sigma_S)^n \ , \tag{3}$$

where k is the normalization constant [equation S8 in text S1]. We used the substitution time from a weak to a strong base to establish a time dimension comparison between the models. Despite representing the same quantity, these are expressed differently by the two models. The weak to strong base substitution time in model M_1 represents a single step from a W to a S state, while in the M_2 model is the time it takes to reach the boundary state N from the boundary state 0, which explicitly includes the time to a mutation plus the fixation time. Both models have closed-form expressions for the substitution time:

$$\begin{aligned} t_{M_1} = \frac{1}{\lambda_{WS}} \quad \text{and} \quad t_{M_2} =& \frac{1}{\mu_{SW}}\left(1+\sum_{n=1}^{N-1}\frac{1}{(1+\sigma_S)^n}\right) \\ &+ N\sum_{n=1}^{N-1}\frac{n+(N-n)(1+\sigma_S)}{n(N-n)}\frac{(1+\sigma_S)^{N-n}-1}{(1+\sigma_S)^{N-n+1}-(1+\sigma_S)^{N-n}} \ . \end{aligned} \tag{4}$$

We further replaced the substitution rates λ_{WS} and λ_{SW} using the identity (2), so these rates and times are comparable and explicitly dependent on the same population processes. Because the population genetics model operates in a larger state-space that includes polymorphic states, substitution times are expectedly higher than in the substitution model. Differently, for the divergence rate, we expect no trend. Nevertheless, we are less interested in the absolute values of these quantities but rather in how much they correlate with each other. Phylogenetic distances are usually normalized, so one substitution is expected per unit of time.

To assess how both models measure evolutionary rates and times, we correlated the divergence rates and substitution times between the two models using the population parameters inferred with BESS and the identities (3) and (4). We observed that the divergence rates are strongly and significantly correlated (correlation coefficient $\rho_B = 0.795$, $p_0 > 13.815$; figure 3A and S6A). ρ_B is corrected for the heterogeneous number of populations per taxa; significant correlations are obtained for $p_0 > 2.944$ (see Material and Methods for more details on the bootstrap analyses). However, divergence rates under the M_2 model vary significantly more across populations than under the M_1 model. Similarly, we found that the substitution times measured by the two models are significantly and strongly correlated ($\rho_B = 0.955$, $p_0 > 13.815$; figure 3B and S6B) and that the M_2 model captures considerably more variation at substitution times.

To better understand the lower variability in the M_2 estimates, we conducted a sensitivity analysis. More specifically, we investigated the effect of different population forces on the estimates of divergence rates and substitution times under the two models. To do that, we perturbed the estimates of each population parameter at a time by $\pm 10\%$ and calculated the relative difference between the perturbed and unperturbed rates and times. The expectation being that the higher the relative difference, the more sensitive the model is to variations in the population parameter that was perturbed. Figure 4 shows the relative difference in divergence rates and substitution times for each population parameter and model.

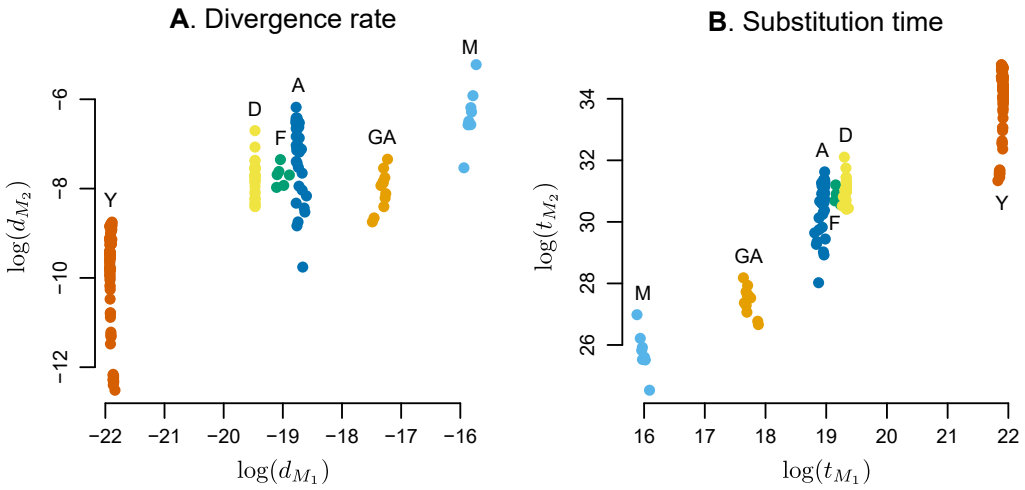

Figure 3: **Comparing divergence rates and substitution times between traditional substitution and the Moran model** Comparison between the **(A)** log divergence rates and **(B)** log substitution times in the standard substitution model M_1 and the Moran model with boundary mutations and selection M_2. Legend: great apes (GA), mice (M), flycatchers (F), fruit flies (D), arabidopsis (A), and baker's yeast (Y).

The sensitivity analyses showed that the impact of mutation rates is overall consistent between the two models: $5 - 15\%$ for the case of the divergence rate and around 20% for the substitution time (figure 4A–B and S4), suggesting that their impact is likely to affect inference in both models. This result is in part expected, as it expresses a known expectation from the neutral theory of molecular evolution [i.e., the correspondence between mutation and substitution rates (Kimura, 1968)].

The GC-bias rate, on the other hand, has a less noticeable effect: $0 - 6.3\%$ for the case of the divergence rate and 0.0–8.4% for the substitution time; figures 4C–D). An intriguing feature is that the impact of selection on the divergence rate and substitution is on average greater in the M_1 model than in the M_2 model (figures 4C–D). This finding suggests that using population genetics-informed substitution rates may slightly overestimate the contribution of nucleotide usage biases to the overall evolutionary process. Our findings also showed that the impact of selection is clearly more pronounced in taxa where GC-bias is stronger, such as the great apes, mice, and the flycatchers, and had almost no effect in taxa where the GC-bias is weak, such as baker's yeast, or almost inexistent, as in fruit flies. This suggests that the substitution models may only be able to capture the net effects of selection. This result is expected by the nearly neutral theory, in which the substitution rates might be accelerated or decelerated depending on the population-scaled selection coefficient (Ohta, 1973). In fact, if we expand the selection term in its Maclaurin series and further assume that σ_S is small, we obtain that

$$(1 + \sigma_S)^N \approx 1 + N\sigma_S \ , \tag{5}$$

from which we can recapture the scaled selection coefficient. We further note that these results on selection must be generalized with care since here, we only investigated a particular case of directional selection acting on GC alleles. As shown by our results (figure 2C) and other studies [e.g., Galtier et al. (2018); Bolívar et al. (2019); Borges et al. (2022)], although ubiquitous, gBGC is likely to act in the nearly neutral range. Thus, despite our results showing selection has a minor effect on substitution times and rates, one should not underestimate the effect of other selection forces in the substitution process. For example, purifying selection acting on coding regions is known to be both pervasive and strong (Cvijović et al., 2018).

The effective population size follows a completely different pattern than that of the selection coefficients

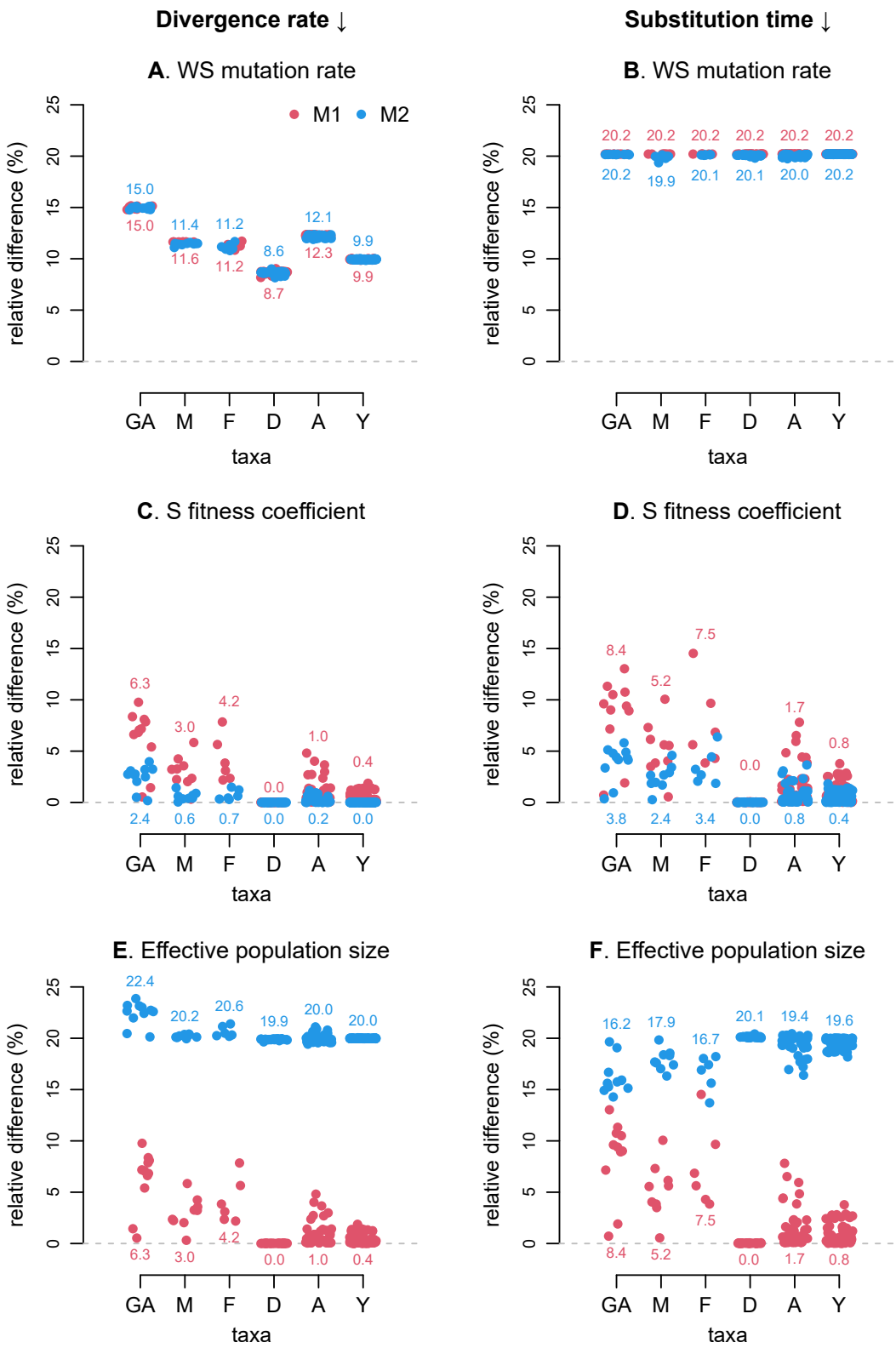

Figure 4: **Sensitivity analyses for the divergence rate and substitution time (A, C and E)** Sensitivity analyses for the divergence rate. The plot relative to the impact of SW mutation rate can be found in figure S4. **(B, D, and F)** Sensitivity analyses for the substitution time. Because we are only considering the WS substitution time, the μ_{WS} has no measurable effect [equation (4)].The relative difference expresses the effect of changing each population parameter by $\pm 10\%$ on the divergence rates. The numbers represent the averaged relative divergence across all populations. Legend: great apes (GA), mice (M), flycatchers (F), fruit flies (D), arabidopsis (A), and baker's yeast (Y).

and mutation rates (figure 4E–F). Changes in population size may explain approximately 19.9 – 22.4% of variation in substitution rates in the M_2 model, but only approximately 0.0 – 6.3% percent in the M_1 model (figure 4E). Similarly, the effective population size is the driving force explaining the main differences in the estimated substitution times between models: while it accounts for 16.2 – 20.1% of its variation in the M_2 model, it only accounts for 0.0 – 8.4% in the M_1 model (figure 4F). These results suggest that the effect of the effective population size, or genetic drift, is largely ignored in the substitution model. We also observed that populations with stronger GC-bias have a more pronounced effect of population size on the substitution rates estimated by the model M_1. This is likely because the population size influences the substitution rates only through the selection coefficient [as shown in equation (2)].

We conclude that while evolutionary rates and times correlate well between the traditional substitution model and its population-genetics equivalent, the former considerably underappreciates genetic drift, despite being explicitly modeled into the substitution rates, while being able to capture the effect of mutation and selection. This result is in agreement with studies that superimposed population genetic formulae into substitution rates. While these studies have been able to quite successfully estimate mutation and selection bias [e.g., Lartillot (2013); Latrille and Lartillot (2022)], inferring variations in the population size has been more challenging. To circumvent this limitation, proxies for population size, such as life history traits (Romiguier et al., 2014; Ellegren and Galtier, 2016) or polymorphic data are incorporated (Brevet and Lartillot, 2021), or further assumptions about the demographic history along the phylogeny are made (e.g., Latrille et al. (2021) assumes an autocorrelated geometric Brownian process to model variations in the effective population size). These studies show that the patterns of nucleotide substitution are not *per se* enough to estimate the effective population.

The impact of unseen variation on the substitution process

When we established the equality (2) we assumed that the expectation for the observed fixed sites is the same in the two models. This is not usually the case, as in phylogenetics, one usually assumes that the observed fixed sites are fixed in the population/species, thus overlooking the underlying mechanisms generating polymorphisms. We properly accounted for this bias in our inferences with BESS by using hypergeometric sampling on the observed counts.

A well-known consequence of sampling alleles from a population of N individuals is that one always assesses a subset of the existing diversity. This bias occurs primarily because low-frequency polymorphisms are more likely to be sampled as monomorphic. More formally, the probability c_M of sampling M S alleles from the population is

$$c_M = s_N + \overbrace{\sum_{n=M}^{N-1} \frac{\binom{n}{M}}{\binom{N}{n}} s_n}^{f} , \tag{6}$$

where s_n is the expected stationary frequency of a population having n S alleles and $(N-n)$ W alleles. Equation (6) quantifies the contribution of true fixed and polymorphic sites to the frequency of observed fixed sites of type S. The quantity f represents the excess of fixed sites and is the proportion of polymorphic sites that are likely to be sampled as monomorphic in a sample of M individuals. As M approaches N, f becomes less important; however, this excess is also affected by the mutation-selection dynamic via the stationary quantity s_n. As a result, the relative contribution of the dynamic and sampling effects to the excess of sampled fixed sites is not immediately clear.

Because the excess of observed fixed sites comes from sampling polymorphisms, we correlated the excess

of fixed sites with the proportion of polymorphic sites in the data. We found an unexpectedly weak and nonsignificant correlation ($\rho_B = -0.177$, $p_0 = 0.792$; figure 5A and S7A). We also correlated the excess of fixed sites with sample size because it is expected that the greater the number of individuals sampled, the smaller the excess of monomorphic sites. Again, we found a weak and insignificant correlation ($\rho_B = 0.069$, $p_0 = 0.347$; figure 5B and S7B). We also observed that the relation between these variables varies considerably across taxa; however, no clear trends are observed when correlations are conducted per taxa: some correlations are positive, others are negative, and not all of them are significant (table S2). These analyses suggest that the population dynamic is the main driver of sampling bias.

To quantify the relative contribution of the dynamic and the sample size to the excess of fixed sites, we again employed a sensitivity analysis. These analyses were only performed with the model M_2 because the excess of fixed sites cannot be assessed with model M_1, which disregards polymorphisms. The mutation rates and the effective population size were found to be the most influential in determining the excess of fixed observed sites, accounting for approximately 3.6 – 16.3% (figure 5C and S5) and 22.1 – 28.3% (figure 5D), respectively. The selection coefficient and the sample size both had a residual impact of approximately 0.0 – 5.2% (figure 5E) and 1.4 – 3.7% respectively (figure 5F).

It is thus clear that the sample size has a more negligible effect on the sampling biases than the population dynamic, and we conclude that when the dynamic is accounted for, it may not only evidently help correct for biases in the allele composition that may distort parameter estimation, but also result in better assessments of evolutionary times and rates. The latter aspect is greatly corroborated by the strong correlations obtained for the divergence rates and the substitution time (figure 3). We emphasize that the number of individuals sampled is only one aspect of our sampling process. Another factor is the number of genomic sites, which we did not address as gathering a larger number of sites is not currently a sampling challenge. We further note that great apes, mice, and flycatchers show similar patterns in regards to the impact of selection and population size, despite having considerable differences in the analyzed sites (9, 26, and 1 million, respectively; table S1).

Thus far, we have been using a population-genetic description for the substitution rates, which is represented by equation (2). However, in most phylogenetic applications, we have no information about the underlying population dynamic or genomic data allowing us to infer it. As we know now that the population dynamic is fundamental to accurately determine the proportion of fixed sites, we test how do evolutinary rates and times relate when we completely ignore it and directly rely on the proportion of observed fixed sites. In order to represent such a situation, we re-estimated the substitution rates directly from the fixed observed sites. This results in notoriously weaker and, for the case of divergence rate, insignificant correlations: $\rho_B = 0.172$ ($p_0 = 1.691$; 6A and S8A) and $\rho_B = 0.468$ ($p_0 > 13.815$; 6B and S8A).

These correlations suggests that when the substitution rates are estimated directly from the fixed sites, they might not express the tempo and pace of the population process. This is not necessarily an aspect of concern, as one might be indeed only interested in studying the long-term effect of population forces. For example, Mugal et al. (2020) showed that the contribution of polymorphisms to sequence divergence becomes negligible for the comparison of distantly related species. However, we expect that the substitution models perform sub-optimally when the mutation-selection process is particularly crucial to establishing the evolutionary relationships and divergence times, which should be the case for closely-related taxa. Indeed, we recently employed the GTR model and PoMos to infer the evolutionary history of thirty fruit fly populations, and, unlike PoMos, the GTR model failed to reconstruct some well-established phylogeographic events of the emergence and worldwide colonization of fruit flies (Borges et al., 2022). We thus conclude that modeling the population dynamic, even if imperfectly, becomes necessary at shorter time

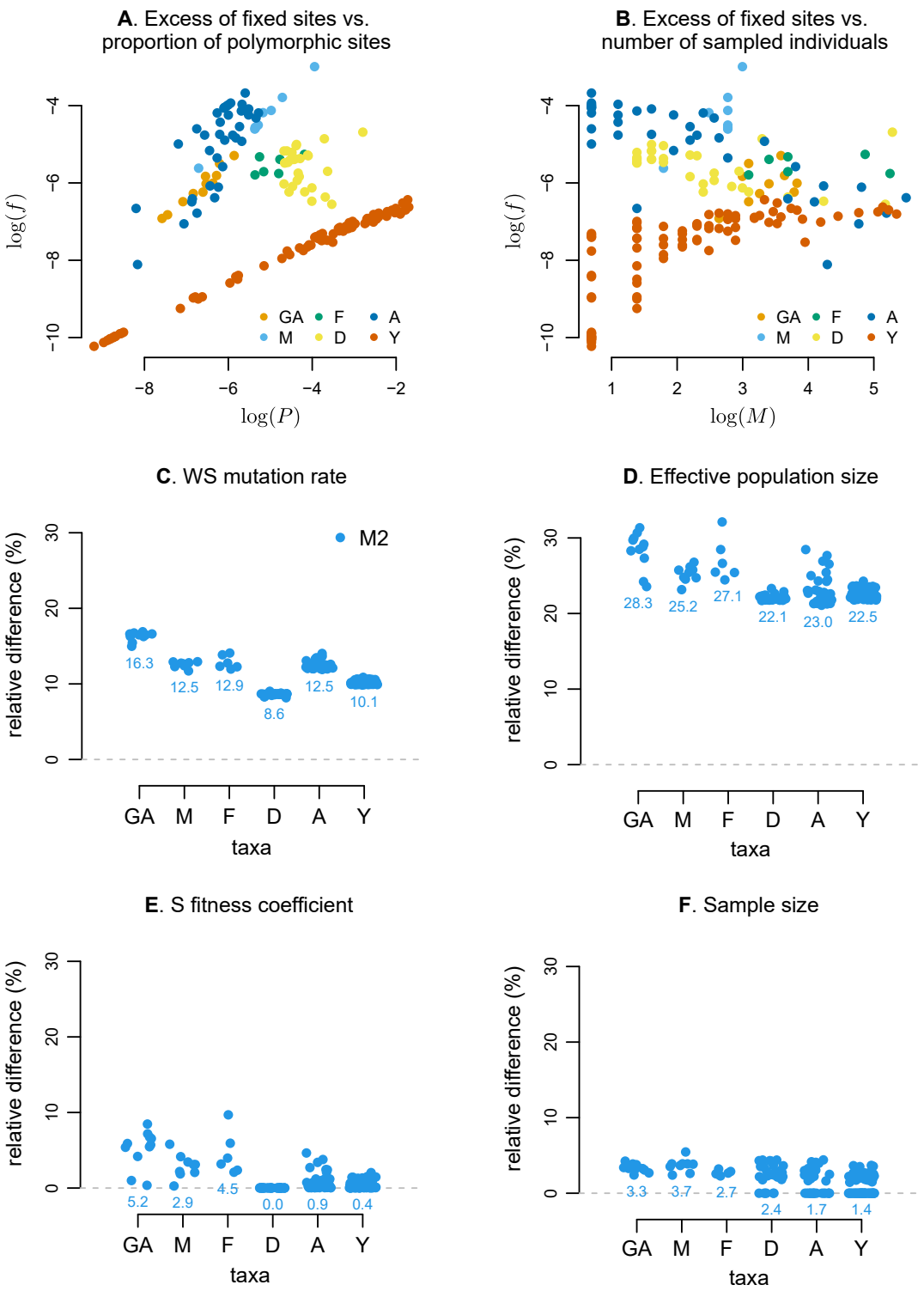

Figure 5: **Assessing the relative contribution of the population dynamic and sampling of individuals to the excess of observed fixed sites (A–B)** Comparison between the log excess of fixed sites (f) and proportion of polymorphism (P) and sample size (M; for diploid organisms, M represent the number of haploid individuals). **(C–F)** Sensitivity analyses for the different population forces. The relative difference expresses the effect of changing each population parameter by ±10% on the excess of observed sites [equation (4)]. The plot relative to the impact of SW mutation rate can be found in figure S5. The numbers represent the average relative divergence across all populations. Legend: great apes (GA), mice (M), flycatchers (F), fruit flies (D), arabidopsis (A), and baker's yeast (Y).

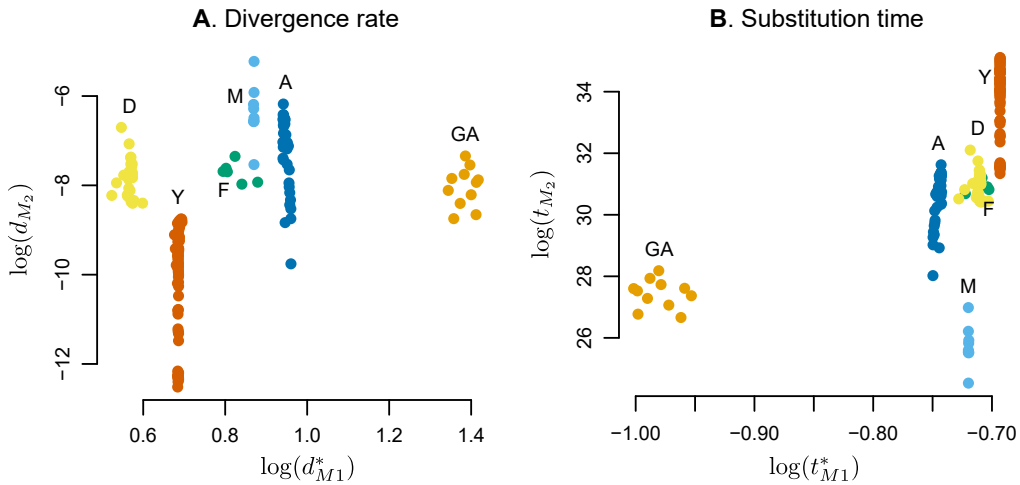

Figure 6: **Comparing rates and times between population-informed and population-naïve substitution models** The asterisk marks the models for which the substitution rates were directly estimated from the fixed sites. Legend: great apes (GA), mice (M), flycatchers (F), fruit flies (D), arabidopsis (A), and baker's yeast (Y).

scales. This observation is in correspondence with the results obtained by several studies on tree discordance due to random drift of ancestral polymorphisms (or incomplete lineage sorting) [e.g., Maddison and Knowles (2006)]. Also, in such a case, traditional substitution models, being polymorphism-unaware, are likely to infer wrong topologies, again showing that the substitution process is not enough to recover the true evolutionary relationships.

Conclusion

In this study, we demonstrated that the evolutionary rates and times correlate well between the traditional substitution model and a population-genetics equivalent. However, despite most mutation and selection biases being well captured, the effect of drift is largely underappreciated even if the population size is explicitly accounted for in the substitution rates. Our results have important implications for the development of mutation-selection phylogenetic models. In particular, they show that reducing the evolutionary process to the substitution event is still sufficient for studying the involvement of mutation and selection on genome base composition; however, population size variations can only be adequately accounted for if intraspecific variation or other co-variates are appropriately leveraged.

Material and Methods

Models for the substitution process

We present and contrast two alternative substitution models. The derivations of the mathematical results presented on these pages are all thoroughly described in the text S1.

Genomic data

We gathered genomic data from six different biological taxa: great apes, mice, flycatchers, fruit flies, Arabidopsis, and baker's yeast, totaling 4385 individuals, 179 populations, and 17 different species. The

type and source of genomic data used in this study are summarized in table 1. The number of individuals and sites for each population is given in table S1, and the sampled site frequency spectra for each population are provided as supplementary material. The software package VCFtools (Danecek et al., 2011) was used to extract information about individual genotypes from the VCF and GVCF files. For each genotype, the four nucleotide bases were recoded posteriorly to W (weak base, A or T) and S (strong base, G or C), and the site-frequency spectra of each population were calculated by counting the absolute frequency of S alleles at various genomic positions.

Bayesian inferences with population genomic data

The stationary distribution of the Moran model with boundary mutations and selection [equations S6 and S7 of the text S1] represents the distribution of allele frequencies in a population and therefore is the population site-frequency spectrum. To model the distribution of the allele content in M allelic counts sampled from a population of N individuals, we used the stationary quantity in conjunction with a hypergeometric distribution. This model was then used to infer the mutation rates, selection coefficient, and population size of a given biallelic system using a Bayesian framework. We set the Markov Chain Monte Carlo algorithm using a Metropolis-Hastings step based on normal proposals. We used a gamma prior for the selection coefficient and a discrete gamma distribution for the population size. The average mutation rate is assumed to be known. The Bayesian estimator of the Site-frequency Spectrum (BESS) was implemented in C++ and is deposited on GitHub at `https://github.com/mrborges23/BESS`. At the same location, we provide a small tutorial explaining how to perform inferences with BESS. BESS takes as input allelic counts in a sample of genomic sites and empirically obtained mutation rates. The program also includes an estimator for the neutral Moran model with biased mutations [based on equation (S16)] that is several orders of magnitude faster than the one with selection.

For our genomic data, we have used the sampled counts for the weak-strong alleles, along with the mutation rates provided in table 1. To estimate the population parameters, we used BESS in two steps: first, we used the neutral model to estimate the population size; second, we used the model with selection, employing the posterior distribution of the population size as priors. The parameter estimates for the 179 populations are provided as supplementary material.

Sensitivity analyses

We conducted a sensitivity analysis to quantify the effect of different population forces on the measures of rate, time, and allele composition under the two models. In particular, we varied each population parameter θ by 10% and calculated the relative difference between the perturbed and unperturbed estimate g of the divergence rate, substitution time, or excess of fixed sites. The relative difference was calculated via the formula

$$RD = \frac{|g(\theta + 0.1\theta) - g(\theta - 0.1\theta)|}{g(\theta)} \times 100\% \ . \tag{7}$$

Statistical analyses

The statistical analyses were performed using R version 4.1.2 (R Core Team, 2021). Because the number of populations varies across taxa and we wanted to avoid biasing our correlations by well-sampled taxa (e.g., yeasts are represented by 82 populations), we computed the correlation coefficients ρ_B between any two variables using the bootstrap technique (Li and Zharkikh, 1994). The correlation coefficient distribution was obtained by randomly sampling one variable value per taxon at each draw. We used

1000 bootstraps; the resulting distributions are in figures S6–S8. We then calculated the quantity

$$p_0 = \left| \log \frac{p(\rho_B > 0)}{p(\rho_B < 0)} \right| , \tag{8}$$

which is zero when the distribution is symmetric around zero. We set $p_0 > 2.944$ as the significance threshold, which expresses the situation where ρ_B is greater or less than 0 with a probability of 0.95. Despite correcting for sampling bias, we have not performed a correction for the phylogenetic effects because we intended to also include the impact of the evolutionary relationships among the taxa. All correlations were performed with log-transformed variables.

Acknowledgements

This research was funded by the Austrian Science Fund (FWF) [P34524-B] awarded to RB.

Authors' contributions

RB and CK conceived the idea. RB implemented the models, analyzed the data, and led the writing of the manuscript. IK, JB, MC, CFM, and CK collected data and contributed to the data analysis. All authors contributed critically to the drafts and gave their final approval for publication.

Conflict of Interest statement

The authors declare no conflicts of interest. RB and CK are review/associate editors of Frontiers in Molecular Biosciences, and CFM is a member of the PCI Evolutionary Biology recommenders.

Data availability statement

The data used and produced in this manuscript are available at `https://vetcloud.vetmeduni.ac.at/nextcloud/s/JNMRXkKAn65ScHy`.

References

Bergman, J. and Schierup, M. H. (2021). Population dynamics of gc-changing mutations in humans and great apes. *Genetics*, 218.

Bolívar, P., Guéguen, L., Duret, L., Ellegren, H., and Mugal, C. F. (2019). Gc-biased gene conversion conceals the prediction of the nearly neutral theory in avian genomes. *Genome Biology*, 20:5.

Bolívar, P., Mugal, C. F., Nater, A., and Ellegren, H. (2016). Recombination rate variation modulates gene sequence evolution mainly via gc-biased gene conversion, not hill–robertson interference, in an avian system. *Molecular Biology and Evolution*, 33:216–227.

Borges, R., Boussau, B., Szöllősi, G. J., and Kosiol, C. (2022). Nucleotide usage biases distort inferences of the species tree. *Genome Biology and Evolution*, 14.

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

Borges, R., Szöllősi, G. J., and Kosiol, C. (2019a). Quantifying GC-Biased Gene Conversion in Great Ape Genomes Using Polymorphism-Aware Models. *Genetics*, 212(4):1321–1336.

Borges, R., Szöllősi, G., and Kosiol, C. (2019b). Quantifying gc-biased gene conversion in great ape genomes using polymorphism-aware models. *Genetics*, page genetics.302074.2019.

Brandvain, Y. and Wright, S. I. (2016). The limits of natural selection in a nonequilibrium world. *Trends in Genetics*, 32:201–210.

Brevet, M. and Lartillot, N. (2021). Reconstructing the history of variation in effective population size along phylogenies. *Genome Biology and Evolution*, 13.

Burri, R., Nater, A., Kawakami, T., Mugal, C. F., Olason, P. I., Smeds, L., Suh, A., Dutoit, L., Bureš, S., Garamszegi, L. Z., Hogner, S., Moreno, J., Qvarnström, A., Ružić, M., Sæther, S.-A., Sætre, G.-P., Török, J., and Ellegren, H. (2015). Linked selection and recombination rate variation drive the evolution of the genomic landscape of differentiation across the speciation continuum of ¡i¿ficedula¡/i¿ flycatchers. *Genome Research*, 25:1656–1665.

Chase, M. A. and Mugal, C. F. (2022). The role of recombination dynamics in shaping signatures of direct and indirect selection across the ficedula flycatcher genome. *bioRxiv*.

Cvijović, I., Good, B. H., and Desai, M. M. (2018). The effect of strong purifying selection on genetic diversity. *Genetics*, 209:1235–1278.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., and Durbin, R. (2011). The variant call format and vcftools. *Bioinformatics*, 27:2156–2158.

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

Duret, L. and Galtier, N. (2009). Biased gene conversion and the evolution of mammalian genomic landscapes. *Annual Review of Genomics and Human Genetics*, 10:285–311.

Ellegren, H. and Galtier, N. (2016). Determinants of genetic diversity. *Nature Reviews Genetics*, 17:422–433.

Fisher, R. (1930). The genetical theory of natural selection. *Genetics*, 154:272.

Flouri, T., Jiao, X., Rannala, B., and Yang, Z. (2020). A Bayesian Implementation of the Multispecies Coalescent Model with Introgression for Phylogenomic Analysis. *Molecular Biology and Evolution*, 37(4):1211–1223.

Galtier, N. (2021). Fine-scale quantification of gc-biased gene conversion intensity in mammals. *Peer Community Journal*, 1:e17.

Galtier, N., Roux, C., Rousselle, M., Romiguier, J., Figuet, E., Glémin, S., Bierne, N., and Duret, L. (2018). Codon usage bias in animals: Disentangling the effects of natural selection, effective population size, and gc-biased gene conversion. *Molecular Biology and Evolution*, 35:1092–1103.

Glémin, S., Arndt, P. F., Messer, P. W., Petrov, D., Galtier, N., and Duret, L. (2015). Quantification of GC-biased gene conversion in the human genome. *Genome Research*, 25(8):1215–1228.

Günther, T., Lampei, C., and Schmid, K. J. (2013). Mutational bias and gene conversion affect the intraspecific nitrogen stoichiometry of the arabidopsis thaliana transcriptome. *Molecular Biology and Evolution*, 30:561–568.

Halpern, A. L. and Bruno, W. J. (1998). Evolutionary distances for protein-coding sequences: modeling site- specific residue frequencies. *Molecular Biology and Evolution*, 15(7):910–917.

Harr, B., Karakoc, E., Neme, R., Teschke, M., Pfeifle, C., Željka Pezer, Babiker, H., Linnenbrink, M., Montero, I., Scavetta, R., Abai, M. R., Molins, M. P., Schlegel, M., Ulrich, R. G., Altmüller, J., Franitza, M., Büntge, A., Künzel, S., and Tautz, D. (2016). Genomic resources for wild populations of the house mouse, mus musculus and its close relative mus spretus. *Scientific Data*, 3:160075.

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

Hervas, S., Sanz, E., Casillas, S., Pool, J. E., and Barbadilla, A. (2017). Popfly: the drosophila population genomics browser. *Bioinformatics*, 33:2779–2780.

Hodgins-Davis, A., Rice, D. P., and Townsend, J. P. (2015). Gene expression evolves under a house-of-cards model of stabilizing selection. *Molecular Biology and Evolution*, 32:2130–2140.

Initiative, A. G. (2000). Analysis of the genome sequence of the flowering plant arabidopsis thaliana. *Nature*, 408:796–815.

Jukes, T. H. and Cantor, C. R. (1969). Evolution of protein molecules. In Munro, H., editor, *Mammalian Protein Metabolism*, pages 21–132. Elsevier, New York.

Keightley, P. D., Trivedi, U., Thomson, M., Oliver, F., Kumar, S., and Blaxter, M. L. (2009). Analysis of the genome sequences of three ¡i¿drosophila melanogaster¡/i¿ spontaneous mutation accumulation lines. *Genome Research*, 19:1195–1201.

Kimura, M. (1968). Evolutionary Rate at the Molecular Level. *Nature*, 217(5129):624–626.

Kingman, J. F. C. (1982). On the genealogy of large populations. *Journal of Applied Probability*, 19(A):27–43.

Lartillot, N. (2013). Phylogenetic patterns of gc-biased gene conversion in placental mammals and the evolutionary dynamics of recombination landscapes. *Molecular Biology and Evolution*, 30:489–502.

Latrille, T., Lanore, V., and Lartillot, N. (2021). Inferring long-term effective population size with mutation–selection models. *Molecular Biology and Evolution*, 38:4573–4587.

Latrille, T. and Lartillot, N. (2022). An improved codon modeling approach for accurate estimation of the mutation bias. *Molecular Biology and Evolution*, 39.

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.

Lesecque, Y., Mouchiroud, D., and Duret, L. (2013). Gc-biased gene conversion in yeast is specifically associated with crossovers: Molecular mechanisms and evolutionary significance. *Molecular Biology and Evolution*, 30:1409–1419.

Li, W.-H. and Zharkikh, A. (1994). What is the bootstrap technique? *Systematic Biology*, 43:424–430.

Lynch, M. (2010a). Evolution of the mutation rate. *Trends in Genetics*, 26:345–352.

Lynch, M. (2010b). Rate, molecular spectrum, and consequences of human mutation. *Proceedings of the National Academy of Sciences*, 107:961–968.

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.

Lynch, M., Sung, W., Morris, K., Coffey, N., Landry, C. R., Dopman, E. B., Dickinson, W. J., Okamoto, K., Kulkarni, S., Hartl, D. L., and Thomas, W. K. (2008). A genome-wide view of the spectrum of spontaneous mutations in yeast. *Proceedings of the National Academy of Sciences*, 105:9272–9277.

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

Moran, P. A. P. (1958). Random processes in genetics. *Mathematical Proceedings of the Cambridge Philosophical Society*, 54(1):60.

Mugal, C. F., Kutschera, V. E., Botero-Castro, F., Wolf, J. B. W., and Kaj, I. (2020). Polymorphism data assist estimation of the nonsynonymous over synonymous fixation rate ratio ω for closely related species. *Molecular Biology and Evolution*, 37:260–279.

Müller, R., Kaj, I., and Mugal, C. F. (2022). A nearly neutral model of molecular signatures of natural selection after change in population size. *Genome Biology and Evolution*, 14.

Nachman, M. W. and Crowell, S. L. (2000). Estimate of the mutation rate per nucleotide in humans. *Genetics*, 156:297–304.

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.

Ohta, T. (1973). Slightly deleterious mutant substitutions in evolution. *Nature*, 246:96–98.

Ossowski, S., Schneeberger, K., Lucas-Lledó, J. I., Warthmann, N., Clark, R. M., Shaw, R. G., Weigel, D., and Lynch, M. (2010). The rate and molecular spectrum of spontaneous mutations in ¡i¿arabidopsis thaliana¡/i¿. *Science*, 327:92–94.

Pessia, E., Popa, A., Mousset, S., Rezvoy, C., Duret, L., and Marais, G. A. B. (2012). Evidence for widespread gc-biased gene conversion in eukaryotes. *Genome Biology and Evolution*, 4:675–682.

Peter, J., Chiara, M. D., Friedrich, A., Yue, J.-X., Pflieger, D., Bergström, A., Sigwalt, A., Barre, B., Freel, K., Llored, A., Cruaud, C., Labadie, K., Aury, J.-M., Istace, B., Lebrigand, K., Barbry, P., Engelen, S., Lemainque, A., Wincker, P., Liti, G., and Schacherer, J. (2018). Genome evolution across 1,011 saccharomyces cerevisiae isolates. *Nature*, 556:339–344.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rannala, B. and Yang, Z. (2003). Bayes estimation of species divergence times and ancestral population sizes using dna sequences from multiple loci. *Genetics*, 164:1645–1656.

Robinson, D., Jones, D., Kishino, H., Goldman, N., and Thorne, J. (2003). Protein Evolution with Dependence Among Codons Due to Tertiary Structure. *Molecular Biology and Evolution*, 20(10):1692–1704.

Robinson, M. C., Stone, E. A., and Singh, N. D. (2014). Population genomic analysis reveals no evidence for gc-biased gene conversion in drosophila melanogaster. *Molecular Biology and Evolution*, 31:425–433.

Romiguier, J., Gayral, P., Ballenghien, M., Bernard, A., Cahais, V., Chenuil, A., Chiari, Y., Dernat, R., Duret, L., Faivre, N., Loire, E., Lourenco, J. M., Nabholz, B., Roux, C., Tsagkogeorga, G., Weber, A. A.-T., Weinert, L. A., Belkhir, K., Bierne, N., Glémin, S., and Galtier, N. (2014). Comparative population genomics in animals uncovers the determinants of genetic diversity. *Nature*, 515:261–263.

Smeds, L., Qvarnström, A., and Ellegren, H. (2016). Direct estimate of the rate of germline mutation in a bird. *Genome Research*, 26:1211–1218.

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

Wright, S. (1931). Evolution in mendelian populations. *Genetics*, 16:97–159.