

# Quantifying GC-Biased Gene Conversion in Great Ape Genomes Using Polymorphism-Aware Models

Rui Borges,\* Gergely J. Szöllösi,<sup>†</sup> and Carolin Kosiol\*<sup>\*,†,1</sup>

\*Institut für Populationsgenetik, Vetmeduni Vienna, 1210 Wien, Austria, <sup>†</sup>Department of Biological Physics, MTA-ELTE “Lendulet” Evolutionary Genomics Research Group, Eötvös University, Pázmány P. stny. 1A, Budapest H-1117, Hungary, and <sup>‡</sup>Centre for Biological Diversity, School of Biology, University of St Andrews, Fife KY16 9TH, UK

ORCID IDs: 0000-0002-5905-3778 (R.B.); 0000-0002-8556-845X (G.J.S.); 0000-0002-3219-6648 (C.K.)

**ABSTRACT** As multi-individual population-scale data become available, more complex modeling strategies are needed to quantify genome-wide patterns of nucleotide usage and associated mechanisms of evolution. Recently, the multivariate neutral Moran model was proposed. However, it was shown insufficient to explain the distribution of alleles in great apes. Here, we propose a new model that includes allelic selection. Our theoretical results constitute the basis of a new Bayesian framework to estimate mutation rates and selection coefficients from population data. We apply the new framework to a great ape dataset, where we found patterns of allelic selection that match those of genome-wide GC-biased gene conversion (gBGC). In particular, we show that great apes have patterns of allelic selection that vary in intensity—a feature that we correlated with great apes’ distinct demographies. We also demonstrate that the AT/GC toggling effect decreases the probability of a substitution, promoting more polymorphisms in the base composition of great ape genomes. We further assess the impact of GC-bias in molecular analysis, and find that mutation rates and genetic distances are estimated under bias when gBGC is not properly accounted for. Our results contribute to the discussion on the tempo and mode of gBGC evolution, while stressing the need for gBGC-aware models in population genetics and phylogenetics.

**KEYWORDS** Moran model; boundary mutations; allelic selection; great apes; GC-bias; gBGC

**T**HE field of molecular population genetics is currently being revolutionized by progress in data acquisition. New challenges are emerging as new lines of inquiry are posed by increasingly large population-scale sequence data (Casillas and Barbadilla 2017). Mathematical theory describing population dynamics was developed before molecular sequences were available (e.g., Fisher 1930; Wright 1931; Moran 1958; Kimura 1964); now that ample data are available to perform statistical inference, many models have been revisited. Recently, the multivariate Moran model with boundary mutations was developed and applied to exome-wide allele frequency data from great ape populations

(Schrempf and Hobolth 2017). However, drift and mutation are not fully sufficient to explain the observed allele counts (Schrempf and Hobolth 2017). It was hypothesized that other forces, such as directional selection and GC-biased gene conversion (gBGC), may also play a role in shaping the distribution of alleles in great apes.

Directional selection and gBGC have different causes but similar signatures: under directional selection, the advantageous allele increases as a consequence of differences in survival and reproduction among different phenotypes; under gBGC, GC alleles are systematically preferred. gBGC is a recombination-associated segregation bias that favors GC-alleles (hereafter, strong alleles) over AT-alleles (hereafter, weak alleles) during the repair of mismatches that occur within heteroduplex DNA during meiotic recombination (Marais 2003). gBGC has been studied in several taxa including mammals (Duret and Galtier 2009; Romiguier *et al.* 2010; Lartillot 2013), birds (Webster *et al.* 2006; Weber *et al.* 2014; Smeds *et al.* 2016; Corcoran *et al.* 2017), reptiles (Figuet *et al.* 2015), plants (Muyle *et al.* 2011; Serres-Giardi *et al.* 2012; Clément *et al.* 2017; Liu *et al.* 2018), and fungi (Pessia *et al.*

Copyright © 2019 by the Genetics Society of America

doi: <https://doi.org/10.1534/genetics.119.302074>

Manuscript received September 13, 2018; accepted for publication May 20, 2019; published Early Online May 30, 2019.

Available freely online through the author-supported open access option.

Supplemental material available at FigShare: <https://doi.org/10.6084/m9.figshare.8180960>.

<sup>1</sup>Corresponding author: Centre for Biological Diversity, University of St Andrews, Greenside Place, Dyers Brae House, St Andrews KY16 9TH, United Kingdom. E-mail: [ck202@st-andrews.ac.uk](mailto:ck202@st-andrews.ac.uk)

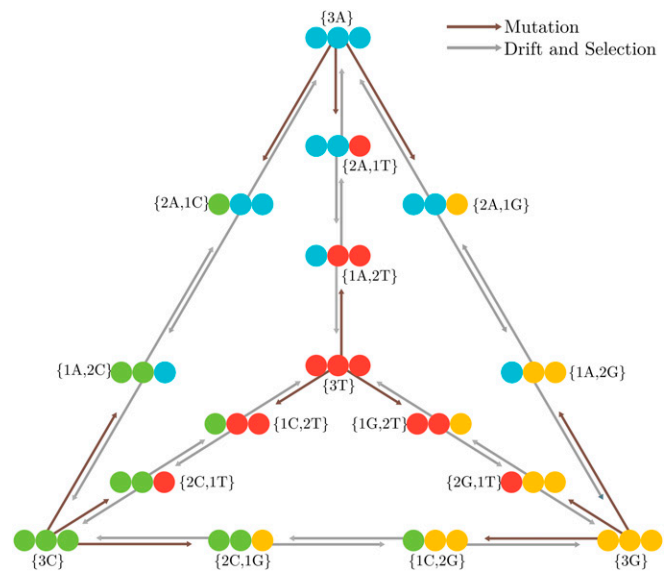
2012; Lesecque *et al.* 2013; Liu *et al.* 2018). However, apart from some studies in human populations (Katzman *et al.* 2011; Glémin *et al.* 2015; Pouyet *et al.* 2018), a population-level perspective of the intensity and diversity of patterns of gBGC among closely related populations is still lacking.

Several questions remain open regarding the tempo and mode of gBGC evolution. For example, the effect of demography on gBGC is still controversial. While theoretical results and studies in mammals and birds advocate a positive relationship between the effective population size and the intensity of gBGC (Nagylaki 1983; Romiguier *et al.* 2010; Weber *et al.* 2014), Galtier *et al.* (2018) failed to detect such relationship between animal phyla. These results open the question as to which extent demography shapes the intensity of gBGC in closely vs. distantly related species/populations. Another aspect that is not completely understood is the impact of GC-bias on the base composition of genomes (Phillips *et al.* 2004; Romiguier *et al.* 2013). In particular, the individual and joint effect of gBGC and mutations shaping the substitution process remains elusive. Here, we address these two questions by revisiting great ape data (Prado-Martinez *et al.* 2013) with a Moran model that accounts for allelic selection.

The Moran model (Moran 1958) has a central position in describing the evolution of a population in that it models the dynamics of allele frequency changes in a finite haploid population. Recently, an approximate solution for the multivariate Moran model with boundary mutations (*i.e.*, low mutation rates) was derived (Schrempf and Hobolth 2017). In particular, the stationary distribution was shown useful to infer population parameters from allele frequency data (Schrempf *et al.* 2016; Schrempf and Hobolth 2017). Here, we present the Moran model with boundary mutations and allelic selection, derive the stationary distribution, and build a Bayesian framework to estimate population parameters. While De Maio *et al.* (2013) had previously proposed a Moran model with allelic selection, we introduce further assumptions on the mutation scheme that permit us to mechanistically describe the relative importance and impact of the population processes mediating the base composition of genomes and expected divergence.

Other approaches making use of allele frequency data to estimate mutation rates and selection coefficients have been proposed in the literature. Glémin *et al.* (2015) proposed a method to quantify gBGC from the derived allele frequency spectra that incorporates polarization errors, takes spatial heterogeneity into account, and jointly estimates mutation bias. The number of derived alleles is modeled by a Poisson distribution on the mutation rates among weak, strong, and neutral alleles (Muyle *et al.* 2011). Our approach differs from that of Glémin *et al.* (2015) as it does not require polarized data or need to account for polarization errors. In addition, our method makes use of the information given by the fixed sites—information that is usually discarded by other methods (Glémin *et al.* 2015 included).

Furthermore, our application to great apes shows that most great apes have patterns of allelic selection consistent with



**Figure 1** PoMo state-space using  $N = 3$ . The four alleles represent the four nucleotide bases. Black and gray arrows indicate mutations, and genetic drift plus selection, respectively. Monomorphic or boundary states  $\{Na_j\}$  are represented in the tetrahedron's vertices, while the polymorphic states  $\{na_i, (N - n)a_j\}$  are represented in its edges. Monomorphic states interact with polymorphic states via mutation, but a polymorphic state can reach a monomorphic state only via drift or selection. Between polymorphic states, only drift and selection events occur.

gBGC. Our results suggest further that demography plays a major role in determining the intensity of gBGC among great apes, as the intensity of the obtained selection coefficients correlates significantly with the effective population size of great apes. We also show that not accounting for GC-bias may considerably distort the reconstructed evolutionary process, as mutation and substitution rates are estimated under bias.

## Methods

### The multivariate Moran model with allelic selection

The modeling framework defined in this work builds on the model described by Schrempf *et al.* (2016), which, according to some proposed terminology (Vogl and Bergman 2015; Schrempf and Hobolth 2017), can be addressed as the multivariate Moran model with boundary mutations (hereafter, *MM*). Here, we describe the *MM* and allelic selection (hereafter, *MS*). The multivariate Moran model can be also referred to as a polymorphism-aware phylogenetic model (PoMo) if we consider the four-variate case (De Maio *et al.* 2013, 2015; Schrempf *et al.* 2016), *i.e.*, representing the four nucleotide bases (Figure 1).

Consider a haploid population of  $N$  individuals and a single locus with  $K$  alleles:  $a_i$  and  $a_j$  are two possible alleles. The evolution of this population over time is described by a continuous-time Markov chain with a discrete character-space defined by  $N$  and  $K$ , each of which represents a specific assortment of alleles. Two types of state can be defined: if all the individuals in a population have the same allele, the

population is monomorphic  $\{Na_i\}$ , i.e., the  $N$  individuals have the allele  $a_i$ ; differently, if two alleles are present in the population, the population is polymorphic  $\{na_i, (N-n)a_j\}$ , meaning that  $n$  individuals have the allele  $a_i$  and  $(N-n)$  have the allele  $a_j$ .  $n/N$  is therefore the frequency of allele  $a_i$  in the population.

Allele trajectories are given by the rate matrix  $\mathbf{Q}$ . Time is accelerated by a factor of  $N$ , and, therefore, instead of describing Moran dynamics in terms of Moran events (Moran 1958), we developed a continuous version in which time is measured as the coalescent in generation time (in units of  $N$ ).

Drift is defined by the neutral Moran model: the transition rates of the allelic frequency shifts depend only on the allele frequency, and are therefore equal regardless of allele increases or decreases in the population (Durrett 2008).

$$q^{\{na_i, (N-n)a_j\} \rightarrow \{(n+1)a_i, (N-n-1)a_j\}} = q^{\{na_i, (N-n)a_j\} \rightarrow \{(n-1)a_i, (N-n+1)a_j\}} = \frac{n(N-n)}{N}. \quad (1)$$

We accommodated mutation and selection in the framework of the neutral Moran model by assuming them to be decoupled (Baake and Bialowons 2008; Etheridge *et al.* 2010).

Mutation is incorporated based on a boundary mutation model, in which mutations occur only in the boundary states.

resents the exchangeabilities between any two alleles, and  $\boldsymbol{\pi}$  the allele base composition (Equation 2). Here, we restricted ourselves to the GTR, as this model simplifies obtaining formal results (Burden and Tang 2016). Because  $\boldsymbol{\pi}$  has  $K-1$  free parameters, and  $\boldsymbol{\rho}$  includes the exchangeabilities for all the possible pairwise combinations of  $K$  alleles, we ended up having  $K(K+1)/2-1$  free parameters in the GTR-based boundary mutation model.

Until now, we have essentially described the model proposed by Schrepf *et al.* (2016); this work extends this model by including allelic selection. We modeled allelic selection by defining  $K-1$  relative selection coefficients  $\boldsymbol{\sigma}$ : the selection coefficient of an arbitrary allele (A in our experiments) is fixed to 0. The selection coefficients defined this way guarantee that our multi-allelic model behaves neutrally only under the condition that all the selection coefficients are the same and equal to 0. Defining the fitness as the probability that an offspring of allele  $a_i$  is replaced with probability  $1 + \sigma_{a_i}$  (Durrett 2008), we can formulate the component of allelic selection alongside with drift, and thus among the polymorphic states (Equation 2).

Altogether, the instantaneous rate matrix  $\mathbf{Q}$  of the multivariate Moran model with boundary mutations and allelic selection can be defined as

$$q^{\{ua_i, (N-u)a_j\} \rightarrow \{va_i, (N-v)a_j\}} = \begin{cases} \mu_{a_i a_j} = \rho_{a_i a_j} \pi_{a_j} & u = N, v = N-1 \\ \mu_{a_j a_i} = \rho_{a_i a_j} \pi_{a_i} & u = 0, v = 1 \\ \frac{n}{N}(N-n)(1 + \sigma_{a_i}) & u = n, v = n+1, 0 < n < N \\ \frac{n}{N}(N-n)(1 + \sigma_{a_j}) & u = n, v = n-1, 0 < n < N \\ 0 & |u-v| > 1 \end{cases}, \quad (2)$$

The boundary mutations assumption is met if the mutation rates  $\mu_{a_i a_j}$  are small (and  $N$  is not too large). More specifically, by comparing the expectations of the diffusion equation with the polymorphic diversity under the Moran model, Schrepf *et al.* (2016) established that  $N\mu_{a_i a_j}$  should be lower than 0.1. In fact, most eukaryotes fulfill this condition [see Lynch *et al.* (2016) for a review of mutation rates]. Another assumption of our boundary mutation model is that the polymorphic states can only be biallelic. However, this assumption is not a significant constraint as tri-or-more allelic sites are rare in sequences with low mutation rates.

We employed the strategy used by Burden and Tang (2016), and separated our model into a time-reversible and a flux part. We wrote the mutation rates as the entries of a specific mutation model, the general time-reversible model (GTR) (Tavaré 1986):  $\mu_{a_i a_j} = \rho_{a_i a_j} \pi_{a_j} = \rho_{a_j a_i} \pi_{a_i}$ , where  $\boldsymbol{\rho}$  rep-

resents where  $u$  and  $v$  represent a frequency change in the allele counts (though  $N$  remains constant). The diagonal elements are defined by the mathematical requirement such that the respective row sum is 0.

As the parameters of the population size, mutation rate, and selection coefficients are confined, it is possible to scale them down to a small value of  $N$  while keeping the overall dynamics unchanged (Appendix A). The virtual population size  $N$  becomes a parameter describing the number of bins the allele frequencies can fall into. As a result, we can think of  $N$  either as a population size or a discretization scheme.

### The stationary distribution

The stationary distribution of a Markov process can be obtained by computing the vector  $\boldsymbol{\psi}$  satisfying the condition

$$\psi_x = \begin{cases} \pi_{a_i}(1 + \sigma_{a_i})^{N-1}k^{-1} & \text{if } x = \{Na_i\} \\ \pi_{a_i}\pi_{a_j}\rho_{a_i a_j}(1 + \sigma_{a_i})^{n-1}(1 + \sigma_{a_j})^{N-n-1}\frac{N}{n(N-n)}k^{-1} & \text{if } x = \{na_i, (N-n)a_j\} \end{cases} \quad (3)$$

$\psi\mathbf{Q} = \mathbf{0}$  (Appendix B).  $\psi$  is the normalized stationary vector, and has the solution

$k$  is the normalization constant

$$k = \sum_{a_i \in \mathcal{A}} \pi_{a_i}(1 + \sigma_{a_i})^{N-1} + \sum_{a_i, a_j \in \mathcal{A}^c} \sum_{n=1}^{N-1} \pi_{a_i}\pi_{a_j}\rho_{a_i a_j}(1 + \sigma_{a_i})^{n-1}(1 + \sigma_{a_j})^{N-n-1}\frac{N}{n(N-n)}, \quad (4)$$

where  $\mathcal{A}$  is the alphabet of the  $K$  alleles  $\{a_1, \dots, a_K\}$ , representing the monomorphic states, and  $\mathcal{A}^c$  all the possible pairwise combinations of  $\mathcal{A}$  representing the  $K(K-1)/2$  types of polymorphic states  $\{a_1a_2, a_1a_3, \dots, a_{K-1}a_K\}$ . For example, for the four-multivariate case, we write  $\mathcal{A}$  as the alphabet of the four nucleotide bases  $\{A, C, G, T\}$  and  $\mathcal{A}^c$  as all the possible pairwise combinations of the four nucleotide bases  $\{AC, AG, AT, CG, CT, GT\}$ .

events  $d_{MS}^*$ : we corrected  $d_{MS}$  by the probability of a mutation and subsequent fixation under the Moran model (Appendix D)

$$d_{MS}^* = \frac{2}{k} \sum_{a_i, a_j \in \mathcal{A}^c} \frac{\pi_{a_i}\pi_{a_j}\rho_{a_i a_j}(1 + \sigma_{a_i})^N(1 + \sigma_{a_j})^N}{\sum_{n=1}^N (1 + \sigma_{a_j})^n(1 + \sigma_{a_i})^{N-n+1}}. \quad (6)$$

### Bayesian inference with stationary distribution

We can define a likelihood function based on the stationary distribution for a set of  $S$  independent sites in  $N$  individuals by taking the product of  $\psi_x$  over counts of monomorphic and polymorphic sites  $c(x)$ , thus:

We employed a Bayesian approach: we define the prior distributions independently, a Dirichlet prior for  $\boldsymbol{\pi}$  and an exponential prior for  $\boldsymbol{\rho}$  and  $\boldsymbol{\sigma}$ ; a Dirichlet and multiplier proposals were set for the aforementioned parameters with tuning parameters

$$p(\mathbf{c}|\boldsymbol{\pi}, \boldsymbol{\rho}, \boldsymbol{\sigma}) = \prod_x \psi_x^{c(x)} = k^{-S} \prod_{a_i \in \mathcal{A}} [\pi_{a_i}(1 + \sigma_{a_i})^{N-1}]^{c(\{Na_i\})} \times \prod_{a_i, a_j \in \mathcal{A}^c} \prod_{n=1}^{N-1} \left[ \pi_{a_i}\pi_{a_j}\rho_{a_i a_j}(1 + \sigma_{a_i})^{n-1}(1 + \sigma_{a_j})^{N-n-1}\frac{N}{n(N-n)} \right]^{c(\{na_i, (N-n)a_j\})}. \quad (7)$$

For a population of size  $N$ , we have  $4 + 6(N-1)$  possible states, four of which are monomorphic (Figure 1).

### Expected number of Moran events

From  $\mathbf{Q}$  and  $\psi$ , we can compute the expected number of Moran events (mutations, drift, and selection) or the expected divergence per unit of time (in generations) under the  $MS$  model (Appendix C):

$$d_{MS} = \frac{2}{k} \sum_{a_i, a_j \in \mathcal{A}^c} \sum_{n=1}^N \pi_{a_i}\rho_{a_i a_j}\pi_{a_j}(1 + \sigma_{a_i})^{n-1}(1 + \sigma_{a_j})^{N-n}. \quad (5)$$

The quantity (5) can also be interpreted as the overall rate of the model. The expected number of Moran events for the neutral model can be easily calculated by letting  $\boldsymbol{\sigma} \rightarrow \mathbf{0}$ . To compare the Moran distance  $d_{MS}$  with the standard models of evolution, we recalculated the Moran distance to account only for substitutions

guaranteeing a target acceptance rate of 0.234 (Roberts *et al.* 1997). We employed the Metropolis-Hastings algorithm (Hastings 1970) for each conditional posterior in a Markov chain Monte Carlo (MCMC) sequence to obtain random samples from the posterior. The algorithm was coded in the R statistical programming language (R Core Team 2015): the packages MCMCpack and expm were integrated in our code to obtain samples from the Dirichlet density and to compute the matrix exponential, respectively (Martin *et al.* 2011; Goulet *et al.* 2017).

### Application: great ape population data

The stationary distribution of the four-multivariate model was employed to infer the distribution of allele frequencies, selection coefficients, and mutation rates from fourfold degenerate sites of exome-wide population data from great apes (Prado-Martinez *et al.* 2013). We used 11 populations with up to

**Table 1 Evidence of allelic selection among the great ape populations.**

Population	<i>N</i>	<i>S</i>	$\log p(\mathbf{c} MM)$	$\log p(\mathbf{c} MS)$	$\log BF$
African humans	6	2827135	-3941390.98	-3940993.95	397
Non-African humans	12	2826956	-3940071.64	-3939858.12	213
Eastern gorillas	6	2823830	-3917375.00	-3917370.00	5
Western gorillas	54	2813092	-3955462.98	-3954663.09	799
Western chimpanzees	10	2823911	-3935188.83	-3934928.50	260
Nigeria-Cameroon chimpanzees	20	2825739	-3980386.43	-3979429.05	957
Eastern chimpanzees	12	2822976	-3961202.57	-3960561.15	641
Central chimpanzees	8	2822685	-3958674.29	-3957704.55	969
Bonobos	26	2824240	-3948520.55	-3947835.54	685
Bornean orangutans	10	2824768	-3952527.89	-3952358.67	169
Sumatran orangutans	10	2824618	-3973247.40	-3972725.44	521

The number of haploid individuals and the number of fourfold degenerate sites per population are indicated by *N* and *S*, respectively. The log Bayes factors ( $\log BF$ ) were calculated as the sum over the product of the allele counts  $\mathbf{c}$  and the posterior predictive probabilities under the Moran model with boundary mutations (*MM*) and allelic selection (*MS*). *BF* favors the model with allelic selection when  $>1$ .

27 diploid individuals, totaling  $\sim 2.8$  million sites per population (Table 1). Data preparation follows the pipeline described in De Maio *et al.* (2015). Estimates of the Watterson's  $\theta$  genetic diversity is  $<0.003$  for all the studied populations (Schrepf *et al.* 2016), validating the boundary mutations assumption of 0.1.

#### Data availability

Data and R scripts necessary to confirm the findings of this article are available on GitHub ([https://github.com/pomo-dev/pomo\\_selection](https://github.com/pomo-dev/pomo_selection)). Supplemental material available at FigShare: <https://doi.org/10.6084/m9.figshare.8180960>.

## Results

### Simulations and algorithm validation

To validate the analytical solution for the stationary distribution of the multivariate Moran model, we compare it to the numerical solution obtained by calculating the probability matrix of  $Q^t$  for large enough *t*. We confirmed that the numerical solution converges to the analytical solution (Supplemental Material, Figure S1).

We validated the Bayesian algorithm for estimating population parameters from the stationary distribution by performing simulations (Figures S2–S5 and Table S2). Our algorithm efficiently recovers the true population parameters from simulated allele counts. We tested the algorithms for different numbers of sites ( $10^3$ ,  $10^6$ , and  $10^9$ ) and state-spaces ( $N = 5, 10$ , and  $50$ ). The number of sites does not increase the computational time substantially and is not a limiting factor for genome-wide analysis. In contrast, the size of the state-space influences the computational time. For larger state-spaces, *N*, more iterations are needed to obtain converged, independent, and mixed MCMC chains during the posterior estimation.

### Patterns of allelic selection in great apes

To test the role of allelic selection defining the distribution of alleles in the great apes, we compared the neutral model (*MM*) and the model with allelic selection (*MS*). Using the predictive stationary distribution and the observed allele counts, we

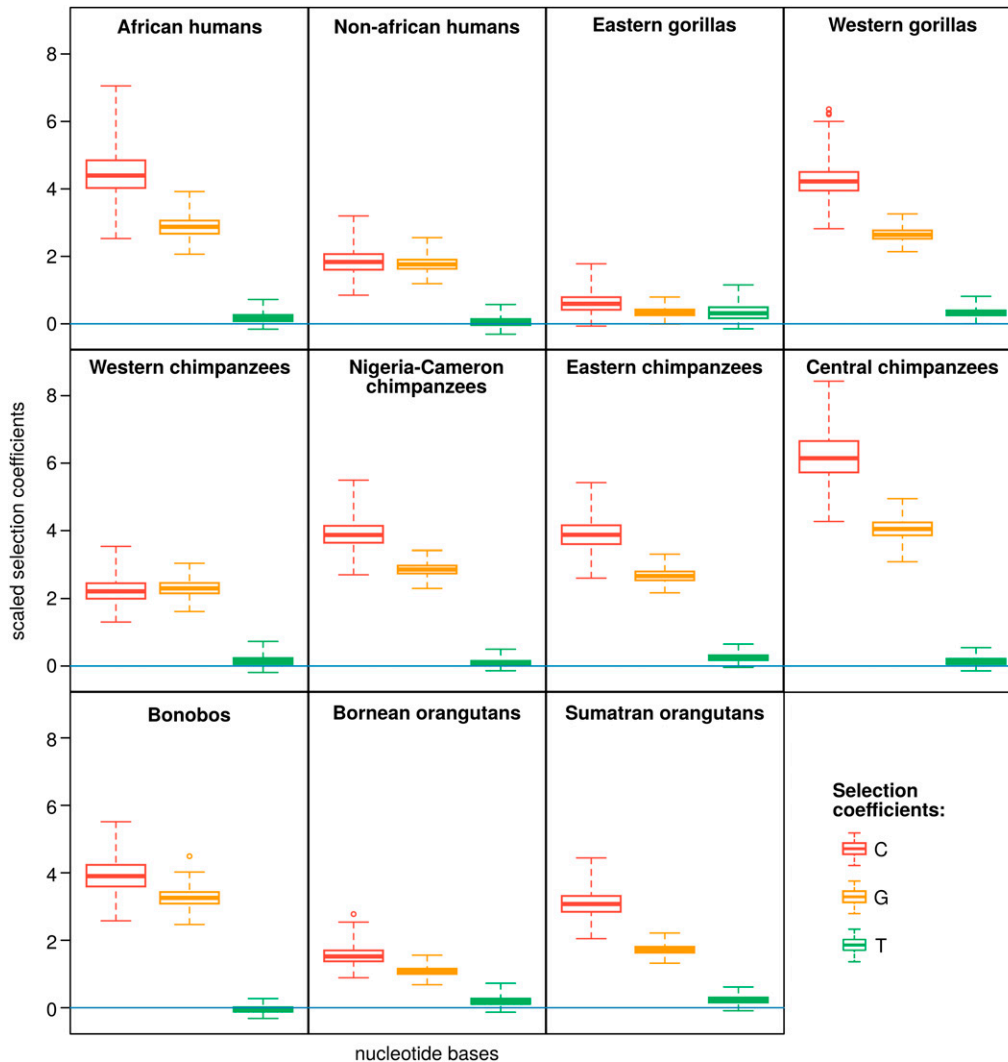
computed the Bayes factors (*BF*) favoring the more complex model *MS* (*i.e.*,  $\log BF > 0$  favors the model with allelic selection) for all populations. It is clear that *MS* fits the data considerably better for most of the studied great apes ( $\log BF > 100$ , Table 1). The only exception is the Eastern gorilla population, for which a lower  $\log BF$  was obtained ( $\log BF = 5.497$ , Table 1).

We have also corroborated our *BF* by inspecting the fit of the predictive distribution of *MM* and *MS* with the allele counts (Figure S6, A–K). The allele counts for the polymorphic states are not symmetrical; generally, one allele is preferred, and thus also the polymorphic states that have it in higher proportions. As expected, we observed that *MS* better reproduces the skewed distribution of allele counts among great apes.

We further investigated the patterns of allelic selection in great apes by analyzing the posterior distribution of the relative selection coefficients of C, G, and T ( $\sigma_A$  was set to 0) under *MS*. A general pattern of allelic selection is observed in great apes: the selection coefficients of C and G are similar (meaning that their posterior distributions largely overlap), but different from the selection coefficient of T, which, in turn, overlaps 0 (approximately equal to the selection coefficient of A) (Figure 2). The only exception is Eastern gorillas, for which the selection coefficients are all only slightly higher than 0 and rather similar (Figure 2). This result corroborates the relatively low *BF* found for evidence of allelic selection in the Eastern gorilla population.

We further explored this result in order to check if the patterns of GC-bias found among great apes can be associated with gBGC. We correlate the GC-bias per chromosome (by averaging the scaled  $\sigma_C$  and  $\sigma_G$ ) with the chromosome size and recombination rate in the non-African human population (Figure S7), for which these data are particularly well characterized (Jensen-Seaman 2004). We found a significant positive correlation between the GC-bias and recombination rate (Spearman's  $\rho = 0.57$ ,  $P = 0.006$ ), but a negative correlation with the chromosome length (Spearman's  $\rho = -0.52$ ,  $P = 0.014$ ).

Although the patterns of selection among great apes are similar qualitatively, they differ quantitatively. For example,



**Figure 2** Scaled allelic selection coefficients for the great apes fourfold degenerate synonymous sites. The boxplots represent the posterior distribution of the C, G, and T scaled selection coefficients ( $\sigma_A$  was set to 0); the estimates were obtained using the four-variate Moran model. The line in blue represents  $\sigma_A = 0$ . We scaled or selection coefficients according to Equation 11 in Appendix A:  $(1 + \sigma)^{N-1}$ . Note that, if  $\sigma$  is small enough, the scaling  $(N - 1)\sigma$  is valid, as  $(1 + \sigma)^{N-1} \approx 1 + (N - 1)\sigma$ . Table S3 summarizes the average unscaled selection coefficients for each great ape population.

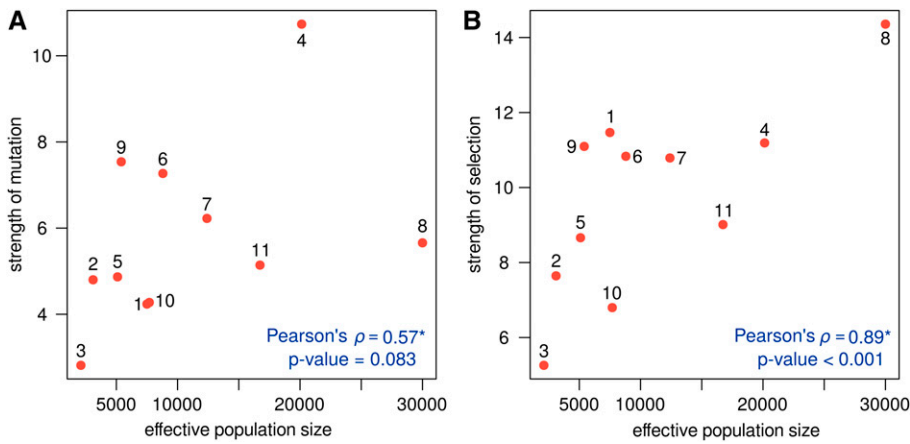
the Central chimpanzees have patterns of GC-bias  $\sim 6.17/4.05$  ( $\sigma_C/\sigma_G$ , Table S3 and Figure 2; scaled according to Equation 11), while the closely related population of Western chimpanzees shows less strong patterns ( $\sim 2.21/2.30$ ). Likewise, the GC-bias content in African and non-African human populations contrasts: 4.47/2.86 and 1.83/1.76, respectively. These results show that the patterns of allelic selection vary greatly among great apes, even among closely related populations.

It has been hypothesized that gBGC is a compensation mechanism for the mutational bias that exists in favor of the weak alleles, A and T (Duret and Galtier 2009; Philippe *et al.* 2011)—the AT/GC toggling effect. We observed that mutation rates from strong to weak alleles are more frequent (by a factor of 2.80; 3.26 if the stationary frequencies are accounted for), while no mutational bias was found between alleles of the same type (1.02; 0.98 if the stationary frequencies are accounted for; Table S3). As the estimated selection coefficients have a clear pattern of GC-bias in most great apes, we can conclude that our anal-

yses are congruent with the expectations of the AT/GC toggling effect.

Furthermore, we compared our method with that of Glémin *et al.* (2015), by considering only two alleles [the strong (S) and weak (W) alleles] using human allele counts from the first human chromosome, divided into 51 regions of 1 million sites (data taken from Glémin *et al.* 2015). We compared estimates of the gBGC rate coefficient as predicted by our model and that of Glémin *et al.* (2015) ( $\sigma_S$  and  $B$ , respectively), and observed that they are negatively correlated (Spearman's  $\rho = -0.37$ ,  $p$ -value = 0.012). Interestingly,  $B$  correlates significantly with our estimates of  $\mu_{WS}$  (the mutation rate of weak to strong alleles;  $\rho = 0.50$ ,  $p$ -value = 0.001). We have further checked the influence of the fixed sites in our estimates of gBGC, and, as expected, we observed that  $\sigma_S$  correlates positively with the percentage of monomorphic sites ( $\rho = 0.36$ ,  $p$ -value = 0.012); intriguingly,  $B$  is negatively correlated ( $\rho = -0.46$ ,  $p$ -value = 0.001). Scatter plots of the mentioned correlation tests can all be found in Figure S8.





**Figure 3** Correlating  $N_e$  with the strength of (A) mutation and (B) selection in great apes. The strength of mutation and selection were obtained by summing up the scaled mutation rates and selection coefficients, respectively. Great ape populations are numbered: 1. African humans, 2. Non-African humans, 3. Eastern gorillas, 4. Western gorillas, 5. Western chimpanzees, 6. Nigeria-Cameroon chimpanzees, 7. Eastern chimpanzees, 8. Central chimpanzees, 9. Bonobos, 10. Bornean orangutans, and 11. Sumatran orangutans. Estimates of  $N_e$  were taken from Prado-Martinez *et al.* (2013) and Tenesa *et al.* (2007). \*Correlation coefficients calculated using independent contrasts and correcting for the effect of the great apes phylogeny [as predicted in Prado-Martinez *et al.* (2013)].

### $N_e$ and the total rate of mutation and selection in great apes

It is widely known that the intensity of mutation and selection reflect population demography. To check whether the estimated mutation and selection coefficients among great ape populations may be explained by demography, we tested the correlation between the total rate of mutation and selection and  $N_e$  (obtained from Tenesa *et al.* (2007); Prado-Martinez *et al.* (2013)). Positive correlations between the total mutation and selection rates and the effective population size were obtained (Figure 3): Pearson's correlation coefficient of 0.57 ( $P = 0.089$ ) and 0.89 ( $P < 0.001$ ), respectively. These correlations were obtained using independent contrasts (Felsenstein 1985) accounting for the great apes phylogeny as predicted in Prado-Martinez *et al.* (2013).

This result shows that  $N_e$  plays an important role in determining the intensity of selection. In particular, it becomes clear that the different patterns of GC-bias found among great apes are due, in part, to different demographies. For example, Central chimpanzees have the highest GC-bias among the studied great apes, and they are indeed the population that was estimated with the largest  $N_e$  (30,000; Prado-Martinez *et al.* 2013). Eastern gorillas showed the opposite pattern; this population had no evidence of GC-bias (with very homogeneous selection coefficients) and congruently Prado-Martinez *et al.* (2013) estimated its  $N_e$  as only 2000—the lowest of the studied populations.

### Comparing the expected number of substitutions in great apes

We calculated the expected number of substitutions under *MM* and *MS* to evaluate the impact of allelic selection (in particular, GC-bias) in the evolutionary process. With Equation 6, we calculated  $d_{MM}^*$  and  $d_{MS}^*$  using the posterior estimates of the respective model parameters. We observe that, for most great ape populations, the expected number of substitutions is lower when allelic selection is accounted for (Table 2); Eastern gorillas are an exception, and the opposite pattern was observed. We also calculated the ratio between

the expected number of substitutions in both models (*i.e.*,  $d_{MS}^*/d_{MM}^*$ ), and we obtained minor (99.8% in Bornean orangutans) to major (82.1% in bonobos) deviations; the average difference is  $-7.3\%$  (Table 2). These results suggest that not accounting for GC-bias may distort the reconstructed evolutionary process by overestimating the expected number of substitutions.

We complement this result by comparing the posterior distribution of the mutations rates in *MM* and *MS*. Because we wanted to identify the mutational types that may be estimated differently between these models, we calculated the relative difference between the mutation rate from allele  $a_i$  to allele  $a_j$  using the following ratio:  $r_{a_i a_j} = \mu_{a_i a_j}^{MS} / \mu_{a_i a_j}^{MM}$ . If  $r_{a_i a_j} > 1$  for a certain mutation rate  $a_i a_j$ , then this mutation rate is being underestimated in *MM* when compared with *MS* (and vice versa if  $r_{a_i a_j} < 1$ ); if  $r_{a_i a_j} \approx 1$  the mutation rates are equally estimated in both models.

We observed a systematic bias among great apes. While weak-to-weak and strong-to-strong mutation rates are generally nondeferentially estimated in both models (most of their  $r$  overlap 1, Figure 4) the strong-to-weak and weak-to-strong mutation rates are generally biased in *MM*. In particular, we obtained that weak-to-strong mutation rates are augmented, while mutations rates from strong-to-weak alleles are deprecated (Figure 4), which suggests that not accounting for GC-bias may bias the estimation of mutation rates. Eastern gorillas behave differently by not showing significant differences between the estimated mutations rates (all  $r_{a_i a_j}$  overlap 1, Figure 4).

### Discussion

In this work, we built on the multivariate Moran model with boundary mutations and allelic selection to explain the population processes shaping the observed distribution of alleles. We obtained new formulae to characterize this model. In particular, we derived the stationary distribution and the rate of the process. In addition, we built a Bayesian framework to estimate population parameters (mutation rates and selection

**Table 2 Expected number of substitutions per unit of time**

Population	$d_{MM}^* \times 10^3$	$d_{MS}^* \times 10^3$	$d_{MS}^*/d_{MM}^*$
African humans	0.123	0.120	0.978
Non-African humans	0.041	0.039	0.954
Eastern gorillas	0.061	0.064	1.045
Western gorillas	0.011	0.009	0.845
Western chimpanzees	0.054	0.052	0.956
Nigeria-Cameroon chimpanzees	0.045	0.038	0.858
Eastern chimpanzees	0.073	0.066	0.910
Central chimpanzees	0.130	0.114	0.873
Bonobos	0.019	0.016	0.821
Bornean orangutans	0.077	0.077	0.998
Sumatran orangutans	0.111	0.106	0.959

The expected number of substitutions for the four-variate Moran model with boundary mutations  $d_{MM}^*$  and allelic selection  $d_{MS}^*$  were calculated based on the posterior distributions of the model parameters and Equation 6. The relative difference between the average number of events between the two models ( $d_{MS}^*/d_{MM}^*$ ) was used to assess how dissimilar these distances are.

coefficients) from population data. This work accomplishes tasks set by Schrepf and Hobolth (2017), who observed derivations from neutrality without having a model in place to enlighten the causes.

### Variable patterns of gBGC among great apes

A genome-wide application to the great apes provides important insight into the strength and magnitude of GC-bias patterns and also the impact of gBGC in the evolutionary process. To our knowledge, this is the first work giving a population perspective of the patterns of GC-bias in nonhuman populations.

Here, we focus on GC-bias because it is a genome-wide effect. Mathematically speaking, it is difficult to disentangle gBGC from directional selection: they may have different biological explanations, but represent the exact same process modeling-wise (*i.e.*, one allele is preferred over the others). Therefore, existing signatures of directional selection are most likely canceling out, when several site-histories ( $\sim 2.8$  million sites in our case) are summarized to perform inferences.

In agreement with previous studies in mammals and humans (Spencer *et al.* 2006; Capra *et al.* 2013; Lartillot 2013; Lachance and Tishkoff 2014; Glémin *et al.* 2015), we found that gBGC is weak on average. Indeed, among great apes, the effect of GC-bias is  $2.75 \pm 1.27$  (value obtained by averaging scaled  $\sigma_C$  and  $\sigma_G$ ), consistent with the nearly neutral scenario (Ohta and Gillespie 1996; Vogl and Bergman 2015). Other studies provided estimates of the scaled conversion coefficient in coding regions: Lynch (2010) estimated  $4N_e s$  as 0.82 in humans and Lartillot (2013) adopted a phylogenetic approach that predicted scaled conversion coefficients  $< 1$  in all apes. Our estimates are comparatively higher; however, our methods and those of Lynch (2010) and Lartillot (2013) have different underlying assumptions. In particular, our method employs the Moran model, which has a rate of genetic drift twice as fast as the Wright-Fisher model. Therefore, we expect to estimate selection coefficients that are twice as high as those in the studies cited.

We found no quantitative agreement between our estimates of the gBGC rate coefficient and those derived from the method of Glémin *et al.* (2015). In addition, we found that our model attributes to mutation what Glémin *et al.* (2015) attributes to gBGC. This might be a consequence the use of monomorphic sites in our method. Indeed, unlike those of Glémin *et al.* (2015), our estimates of gBGC correlate positively with the percentage of fixed sites. In general, the gBGC rate coefficient should promote greater fixation by boosting the purging of polymorphic sites (at least for low mutation rates, as those observed in humans). On the other hand, Glémin *et al.* (2015) also considered a varying GC-content, which may explain why their estimates of gBGC do not correlate with the percentage of fixed sites. We have preliminary evidence showing that monomorphic sites can significantly impact estimates of population parameters. Nevertheless, a more comprehensive model accounting for both fixed states and variable GC-content would be necessary to disentangle their relative contribution to explaining allele counts.

The patterns of GC-bias we found in great apes are in concordance with the well-known process of gBGC. As expected, we observed that the larger the recombination rate or the lower the chromosome length, the higher the GC-effect. Evidently, recombination promotes gBGC; however, a negative association between gBGC and chromosome size is expected [in most organisms, small chromosomes undergo more recombination per unit of physical distance than large chromosomes (Kaback *et al.* 1989)]. We performed these analyses in non-African humans, for which these data are available; however, we are confident that the patterns of GC-bias found in great apes are due to gBGC.

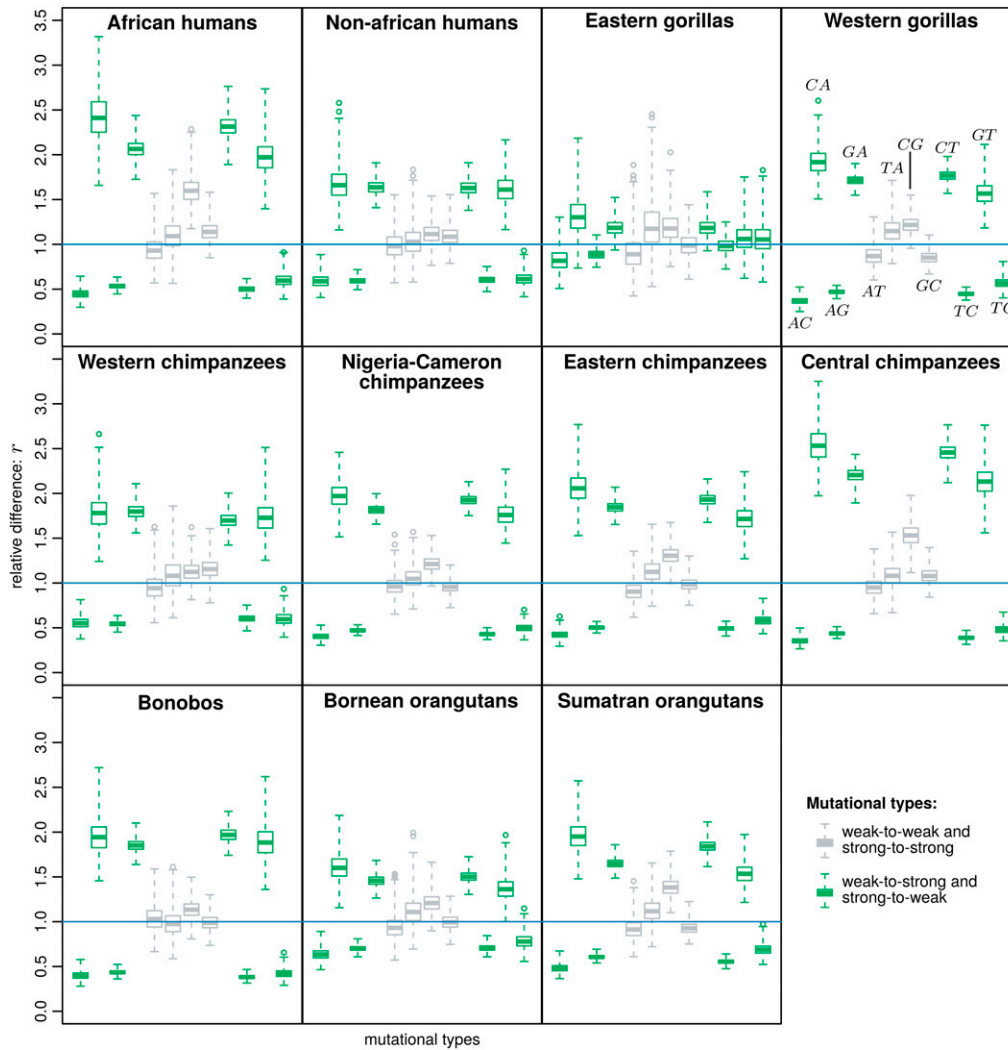
It has been hypothesized that GC-bias is a compensation mechanism for the mutational bias that exists in favor of the weak alleles, A and T (Duret and Galtier 2009; Galtier *et al.* 2009; Philippe *et al.* 2011). Congruent with this expectation, we observed that mutation rates from strong to weak alleles are higher, but rather similar between alleles of the same type. Interestingly, as we have demonstrated, this symmetric manner by which mutations and selection act in great apes leads the number of substitutions to decrease on average. This suggests that AT/GC toggling may increase population variability by promoting more polymorphic sites; however, further studies would be necessary to clarify this prediction.

### Intensity of gBGC and demography in great apes

Glémin *et al.* (2015) hypothesized that differences in GC-bias intensity among human populations were due to the effects of demography. We also advance that demography regulates the intensity of gBGC in great apes. We obtained a positive correlation between the total rate of selection and  $N_e$  in great apes. An important conclusion of our study is that patterns of gBGC can change rapidly due to demography, even among closely related populations. In fact, most of the studied populations are known to have diverged  $< 0.5$  MYA (Prado-Martinez *et al.* 2013).

Here, we showed that GC-bias determines the genome-wide base composition of genomes in a factor proportional to





**Figure 4** Relative difference in the mutation rates estimated under the neutral and non-neutral Moran model.  $r_{a_i a_j}$  represents the ratio between the mutation from allele  $a_i$  to allele  $a_j$  in the model with allelic selection and the model with boundary mutations:  $r_{a_i a_j} = \mu_{a_i a_j}^{MS} / \mu_{a_i a_j}^{MM}$ . The 12 mutational types are indicated in the western gorillas plot: all of the plots follow this arrangement.

$(1 + \sigma_{C/G})^{N_e - 1}$  [or  $(1 + s)^{N_e - 1}$  in the true dynamic]. Therefore, by either changing  $N_e$  or  $s$ , we are able to change the AT/GC composition of genomes. Because we were able to correlate  $N_e$  with the intensity of allelic selection (Pearson's  $\rho = 0.89$ ), we are convinced that demography plays a major role determining the base composition of great ape genomes. Studies using life history traits (*i.e.*, body size) in mammals (Romiguier *et al.* 2010) and ancestral reconstructions of the effective population size in birds (Weber *et al.* 2014) also advocated for correlations between  $N_e$  and GC-content [although not as strong as that found here;  $\rho \sim 0.30 - 0.55$  in Weber *et al.* (2014)].

In contrast, Galtier *et al.* (2018) did not find this correlation in a data set covering 31 species of distinct metazoa phyla (including vertebrates, insects, molluscs, crustaceans, echinoderms, tunicates, annelids, nematodes, nemertians, and cnidarians). This is most likely happening because aspects of the recombination landscape, such as genome-wide recombination rate, length of gene conversion tracts, and repair biases, may also affect the intensity of gBGC (Duret and Galtier 2009; Lesecque *et al.* 2014; Galtier *et al.* 2018). As the recombination landscape varies significantly across

species, but not so much across closely related populations (*e.g.*, the karyotype is very conserved among great apes, with humans having 46 diploid chromosomes whereas other great apes having 48), we expected stronger correlations between the intensity of gBGC and demography.

Knowing to what extent variations in  $N_e$  or  $s$  determine the base composition of genomes will require further study. In particular, determining  $s$  experimentally in different populations/species would help assess the real impact of gBGC. If we assume that  $s$  varies slightly among closely related populations/species, then we might attribute different intensities of GC-bias almost solely to demographic effects, which simplifies the task of accommodating gBGC in population models.

#### ***gBGC calls for caution in molecular and phylogenetic analyses***

The effects of gBGC in molecular analysis have been described extensively in the literature [reviewed in Romiguier and Roux (2017)]. We complement these results by showing how GC-bias affects the base composition of genomes, and how the mutation rates and genetic distances may be biased if

GC-bias is not properly accounted for. In particular, we observed that mutation rates from weak-to-strong and strong-to-weak alleles are systematically over and underestimated, respectively.

The idea that gBGC may distort the reconstructed evolutionary process comes mainly from phylogenetic studies. For example, it is hypothesized that gBGC may promote substitution saturation (Romiguier and Roux 2017). We have shown that the number of substitutions may be significantly overestimated if we do not account for GC-bias, meaning that gBGC may indeed promote branch saturation. Based on this and other gBGC-related complications [e.g., GC-bias promotes incomplete lineage sorting (Hobolth *et al.* 2011)], some authors advocate that only GC-poor markers should be used for phylogenetic analysis (McCormack *et al.* 2012; Romiguier *et al.* 2013). Contradicting this approach, our results show that we may gain more inferential power if GC-bias is accounted for when estimating evolutionary distances.

Here, we have not performed phylogenetic inference, but previous applications of the Moran model to phylogenetic problems (*i.e.*, PoMo) (De Maio *et al.* 2015; Schrempf *et al.* 2016) show that it can be done. Therefore, a necessary future work would be to test the effect of allelic selection (or, more specifically, GC-bias) in phylogeny reconstruction; in particular, it would be of major interest to determine how much of its signal can account for the increase in accuracy of tree estimation.

Recently, a nucleotide substitution process that accounts for gBGC was proposed by Lartillot (2013). In this latter model, a scaled conversion coefficient is used to correct substitution rates in a manner similar to that used here to calculate the expected number of substitutions for the Moran distance (*i.e.*, assessing the relative fixation probabilities under GC-bias, File S3). Therefore, these models should perform similarly, with the exception that PoMo should be able to disentangle the contribution of selection and mutation to the observed diversity, as it additionally accounts for polymorphic sites.

## Conclusions

Despite widespread evidence of gBGC in several taxa, several questions remain open regarding the role of gBGC in determining the base composition of genomes. In this work, we provide a mechanistic model and theoretical results that allow quantification of patterns of gBGC in nonhuman closely related populations, providing a new layer of understanding of the tempo and mode of gBGC evolution in vertebrate genomes.

In addition, our multivariate Moran model with allelic selection makes a significant contribution to the endeavor of estimating population parameters from multi-individual population-scale data. Importantly, our analysis showed that gBGC may significantly distort estimates of population parameters and genetic distances, highlighting that gBGC-aware models should be used when employing molecular phyloge-

netics and population genetics analyses. We stress that, although our application to great apes show evidence of GC-bias, our framework can be employed more generally to estimate patterns of nucleotide usage and associated mechanisms of evolution.

## Acknowledgments

We thank Dominik Schrempf, Claus Vogl, and Sylvain Glémin for helpful discussions, and the three anonymous reviewers for comments improving the manuscript. This work was funded by the Vienna Science and Technology Fund (WWTF) through project MA16-061. G.J.S. received funding from the European Research Council under the European Unions Horizon 2020 research and innovation program under grant agreement no. 714774.

## Literature Cited

- Baake, E., and R. Bialowons, 2008 Ancestral processes with selection: branching and Moran models. *Banach Center Publications* 80: 33–52. <https://doi.org/10.4064/bc80-0-2>
- Burden, C. J., and Y. Tang, 2016 An approximate stationary solution for multi-allele neutral diffusion with low mutation rates. *Theor. Popul. Biol.* 112: 22–32. <https://doi.org/10.1016/j.tpb.2016.07.005>
- Capra, J. A., M. J. Hubisz, D. Kostka, K. S. Pollard, and A. Siepel, 2013 A model-based analysis of GC-biased gene conversion in the human and chimpanzee genomes. *PLoS Genet.* 9: e1003684. <https://doi.org/10.1371/journal.pgen.1003684>
- Casillas, S., and A. Barbadilla, 2017 Molecular population genetics. *Genetics* 205: 1003–1035. <https://doi.org/10.1534/genetics.116.196493>
- Clément, Y., G. Sarah, Y. Holtz, F. Homa, S. Pointet *et al.*, 2017 Evolutionary forces affecting synonymous variations in plant genomes. *PLoS Genet.* 13: e1006799. <https://doi.org/10.1371/journal.pgen.1006799>
- Corcoran, P., T. I. Gossmann, H. J. Barton, J. Great Tit HapMap Consortium, Slate *et al.*, 2017 Determinants of the efficacy of natural selection on coding and noncoding variability in two passerine species. *Genome Biol. Evol.* 9: 2987–3007.
- De Maio, N., C. Schlötterer, and C. Kosiol, 2013 Linking great apes genome evolution across time scales using polymorphism-aware phylogenetic models. *Mol. Biol. Evol.* 30: 2249–2262. <https://doi.org/10.1093/molbev/mst131>
- De Maio, N., D. Schrempf, and C. Kosiol, 2015 PoMo: an allele frequency-based approach for species tree estimation. *Syst. Biol.* 64: 1018–1031. <https://doi.org/10.1093/sysbio/syv048>
- Duret, L., and N. Galtier, 2009 Biased gene conversion and the evolution of mammalian genomic landscapes. *Annu. Rev. Genomics Hum. Genet.* 10: 285–311. <https://doi.org/10.1146/annurev-genom-082908-150001>
- Durrett, R., 2008 *Probability Models for DNA Sequence Evolution* (Probability and its Applications Series, Vol. 34). Springer, New York. <https://doi.org/10.1007/978-0-387-78168-6>
- Etheridge, A. M., R. C. Griffiths, and J. E. Taylor, 2010 A coalescent dual process in a Moran model with genic selection, and the lambda coalescent limit. *Theor. Popul. Biol.* 78: 77–92. <https://doi.org/10.1016/j.tpb.2010.05.004>
- Felsenstein, J., 1985 Phylogenies and the comparative method. *Am. Nat.* 125: 1–15. <https://doi.org/10.1086/284325>
- Figuet, E., M. Ballenghien, J. Romiguier, and N. Galtier, 2015 Biased gene conversion and GC-content evolution in

- the coding sequences of reptiles and vertebrates. *Genome Biol. Evol.* 7: 240–250. <https://doi.org/10.1093/gbe/evu277>
- Fisher, R. a., 1930 The genetical theory of natural selection. *Genetics* 154: 272.
- Galtier, N., L. Duret, S. Glémin, and V. Ranwez, 2009 GC-biased gene conversion promotes the fixation of deleterious amino acid changes in primates. *Trends Genet.* 25:1–5. (erratum: *Trends Genet.* 25: 287)
- Galtier, N., C. Roux, M. Rousselle, J. Romiguier, E. Figuet *et al.*, 2018 Codon usage bias in animals: disentangling the effects of natural selection, effective population size, and GC-biased gene conversion. *Mol. Biol. Evol.* 35: 1092–1103. <https://doi.org/10.1093/molbev/msy015>
- Glémin, S., P. F. Arndt, P. W. Messer, D. Petrov, N. Galtier *et al.*, 2015 Quantification of GC-biased gene conversion in the human genome. *Genome Res.* 25: 1215–1228. <https://doi.org/10.1101/gr.185488.114>
- Goulet, V., C. Dutang, M. Maechler, D. Firth, M. Shapira *et al.*, 2017 expm: matrix exponential, log, 'etc'. R package version 0.999–2.
- Hastings, W. K., 1970 Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 57: 97–109. <https://doi.org/10.1093/biomet/57.1.97>
- Hobolth, A., J. Y. Dutheil, J. Hawks, M. H. Schierup, and T. Mailund, 2011 Incomplete lineage sorting patterns among human, chimpanzee, and orangutan suggest recent orangutan speciation and widespread selection. *Genome Res.* 21: 349–356. <https://doi.org/10.1101/gr.114751.110>
- Jensen-Seaman, M. I., 2004 Comparative recombination rates in the rat, mouse, and human genomes. *Genome Res.* 14: 528–538. <https://doi.org/10.1101/gr.1970304>
- Kaback, D. B., H. Y. Steensma, and P. de Jonge, 1989 Enhanced meiotic recombination on the smallest chromosome of *Saccharomyces cerevisiae*. *Proc. Natl. Acad. Sci. USA* 86: 3694–3698. <https://doi.org/10.1073/pnas.86.10.3694>
- Katzman, S., J. A. Capra, D. Haussler, and K. S. Pollard, 2011 Ongoing GC-biased evolution is widespread in the human genome and enriched near recombination hot spots. *Genome Biol. Evol.* 3: 614–626. <https://doi.org/10.1093/gbe/evr058>
- Kimura, M., 1964 Diffusion models in population genetics. *J. Appl. Probab.* 1: 177–232. <https://doi.org/10.2307/3211856>
- Kluth, S., and E. Baake, 2013 The Moran model with selection: fixation probabilities, ancestral lines, and an alternative particle representation. *Theor. Popul. Biol.* 90: 104–112. <https://doi.org/10.1016/j.tpb.2013.09.009>
- Lachance, J., and S. A. Tishkoff, 2014 Biased gene conversion skews allele frequencies in human populations, increasing the disease burden of recessive alleles. *Am. J. Hum. Genet.* 95: 408–420. <https://doi.org/10.1016/j.ajhg.2014.09.008>
- Lartillot, N., 2013 Phylogenetic patterns of GC-biased gene conversion in placental mammals and the evolutionary dynamics of recombination landscapes. *Mol. Biol. Evol.* 30: 489–502. <https://doi.org/10.1093/molbev/mss239>
- Lesecque, Y., D. Mouchiroud, and L. Duret, 2013 GC-biased gene conversion in yeast is specifically associated with crossovers: molecular mechanisms and evolutionary significance. *Mol. Biol. Evol.* 30: 1409–1419. <https://doi.org/10.1093/molbev/mst056>
- Lesecque, Y., S. Glémin, N. Lartillot, D. Mouchiroud, and L. Duret, 2014 The red queen model of recombination hotspots evolution in the light of archaic and modern human genomes. *PLoS Genet.* 10: e1004790. <https://doi.org/10.1371/journal.pgen.1004790>
- Liu, H., J. Huang, X. Sun, J. Li, Y. Hu *et al.*, 2018 Tetrad analysis in plants and fungi finds large differences in gene conversion rates but no GC bias. *Nat. Ecol. Evol.* 2: 164–173. <https://doi.org/10.1038/s41559-017-0372-7>
- Lynch, M., 2010 Rate, molecular spectrum, and consequences of human mutation. *Proc. Natl. Acad. Sci. USA* 107: 961–968. <https://doi.org/10.1073/pnas.0912629107>
- Lynch, M., M. S. Ackerman, J.-F. Gout, H. Long, W. Sung *et al.*, 2016 Genetic drift, selection and the evolution of the mutation rate. *Nat. Rev. Genet.* 17: 704–714. <https://doi.org/10.1038/nrg.2016.104>
- Marais, G., 2003 Biased gene conversion: implications for genome and sex evolution. *Trends Genet.* 19: 330–338. [https://doi.org/10.1016/S0168-9525\(03\)00116-1](https://doi.org/10.1016/S0168-9525(03)00116-1)
- Martin, A. D., K. M. Quinn, and J. H. Park, 2011 MCMCpack: Markov chain Monte Carlo in R. *J. Stat. Softw.* 42: 22. <https://doi.org/10.18637/jss.v042.i09>
- McCormack, J. E., B. C. Faircloth, N. G. Crawford, P. A. Gowaty, R. T. Brumfield *et al.*, 2012 Ultraconserved elements are novel phylogenomic markers that resolve placental mammal phylogeny when combined with species-tree analysis. *Genome Res.* 22: 746–754. <https://doi.org/10.1101/gr.125864.111>
- Moran, P., 1958 Random processes in genetics. *Math. Proc. Camb. Philos. Soc.* 54: 60. <https://doi.org/10.1017/S0305004100033193>
- Muyle, A., L. Serres-Giardi, A. Ressayre, J. Escobar, and S. Glémin, 2011 GC-biased gene conversion and selection affect GC content in the *Oryza* genus (rice). *Mol. Biol. Evol.* 28: 2695–2706. <https://doi.org/10.1093/molbev/msr104>
- Nagylaki, T., 1983 Evolution of a finite population under gene conversion. *Proc. Natl. Acad. Sci. USA* 80: 6278–6281. <https://doi.org/10.1073/pnas.80.20.6278>
- Ohta, T., and J. Gillespie, 1996 Development of neutral and nearly neutral theories. *Theor. Popul. Biol.* 49: 128–142. <https://doi.org/10.1006/tpbi.1996.0007>
- Pessia, E., A. Popa, S. Mousset, C. Rezvoy, L. Duret *et al.*, 2012 Evidence for widespread GC-biased gene conversion in eukaryotes. *Genome Biol. Evol.* 4: 675–682. <https://doi.org/10.1093/gbe/evs052>
- Philippe, H., H. Brinkmann, D. V. Lavrov, D. T. J. Littlewood, M. Manuel *et al.*, 2011 Resolving difficult phylogenetic questions: why more sequences are not enough. *PLoS Biol.* 9: e1000602. <https://doi.org/10.1371/journal.pbio.1000602>
- Phillips, M. J., F. Delsuc, and D. Penny, 2004 Genome-scale phylogeny and the detection of systematic biases. *Mol. Biol. Evol.* 21: 1455–1458. <https://doi.org/10.1093/molbev/msh137>
- Pouyet, F., S. Aeschbacher, A. Thiéry, and L. Excoffier, 2018 Background selection and biased gene conversion affect more than 95% of the human genome and bias demographic inferences. *eLife* 7: pii: e36317. <https://doi.org/10.7554/eLife.36317>
- Prado-Martinez, J., P. H. Sudmant, J. M. Kidd, H. Li, J. L. Kelley *et al.*, 2013 Great ape genetic diversity and population history. *Nature* 499: 471–475. <https://doi.org/10.1038/nature12228>
- R Core Team, 2015 *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna.
- Roberts, G. O., A. Gelman, and W. R. Gilks, 1997 Weak convergence and optimal scaling of random walk Metropolis algorithms. *Ann. Appl. Probab.* 7: 110–120.
- Romiguier, J., and C. Roux, 2017 Analytical biases associated with GC-content in molecular evolution. *Front. Genet.* 8: 16. <https://doi.org/10.3389/fgene.2017.00016>
- Romiguier, J., V. Ranwez, E. J. P. Douzery, and N. Galtier, 2010 Contrasting GC-content dynamics across 33 mammalian genomes: relationship with life-history traits and chromosome sizes. *Genome Res.* 20: 1001–1009. <https://doi.org/10.1101/gr.104372.109>
- Romiguier, J., V. Ranwez, F. Delsuc, N. Galtier, and E. J. Douzery, 2013 Less is more in mammalian phylogenomics: AT-rich genes minimize tree conflicts and unravel the root of placental

- mammals. *Mol. Biol. Evol.* 30: 2134–2144. <https://doi.org/10.1093/molbev/mst116>
- Schrempf, D., and A. Hobolth, 2017 An alternative derivation of the stationary distribution of the multivariate neutral WrightFisher model for low mutation rates with a view to mutation rate estimation from site frequency data. *Theor. Popul. Biol.* 114: 88–94. <https://doi.org/10.1016/j.tpb.2016.12.001>
- Schrempf, D., B. Q. Minh, N. De Maio, A. von Haeseler, and C. Kosiol, 2016 Reversible polymorphism-aware phylogenetic models and their application to tree inference. *J. Theor. Biol.* 407: 362–370. <https://doi.org/10.1016/j.jtbi.2016.07.042>
- Serres-Giardi, L., K. Belkhir, J. David, and S. Glémin, 2012 Patterns and evolution of nucleotide landscapes in seed plants. *Plant Cell* 24: 1379–1397. <https://doi.org/10.1105/tpc.111.093674>
- Smeds, L., C. F. Mugal, A. Qvarnström, and H. Ellegren, 2016 High-resolution mapping of crossover and non-crossover recombination events by whole-genome Re-sequencing of an avian pedigree. *PLoS Genet.* 12: e1006044. <https://doi.org/10.1371/journal.pgen.1006044>
- Spencer, C. C. A., P. Deloukas, S. Hunt, J. Mullikin, S. Myers *et al.*, 2006 The influence of recombination on human genetic diversity. *PLoS Genet.* 2: e148. <https://doi.org/10.1371/journal.pgen.0020148>
- Tavaré, S., 1986 Some probabilistic and statistical problems in the analysis of DNA sequences. *Lect. Math Life Sci.* 17: 57–86.
- Tenesa, A., P. Navarro, B. J. Hayes, D. L. Duffy, G. M. Clarke *et al.*, 2007 Recent human effective population size estimated from linkage disequilibrium. *Genome Res.* 17: 520–526. <https://doi.org/10.1101/gr.6023607>
- Vogl, C., and J. Bergman, 2015 Inference of directional selection and mutation parameters assuming equilibrium. *Theor. Popul. Biol.* 106: 71–82. <https://doi.org/10.1016/j.tpb.2015.10.003>
- Weber, C. C., B. Boussau, J. Romiguier, E. D. Jarvis, and H. Ellegren, 2014 Evidence for GC-biased gene conversion as a driver of between-lineage differences in avian base composition. *Genome Biol.* 15: 549. <https://doi.org/10.1186/s13059-014-0549-1>
- Webster, M. T., E. Axelsson, and H. Ellegren, 2006 Strong regional biases in nucleotide substitution in the chicken genome. *Mol. Biol. Evol.* 23: 1203–1216. <https://doi.org/10.1093/molbev/msk008>
- Wright, S., 1931 Evolution in Mendelian populations. *Genetics* 16: 97–159.

*Communicating editor: M. Beaumont*

## Appendix A

### Virtual Population Size

Consider two populations,  $A$  and  $A'$ , with different population size,  $N$  and  $M$ , respectively. We want to mimic the dynamics of population  $A$ , relying on the population parameters of a population  $A'$  of different size (larger or smaller). Both populations have the same number of monomorphic states (equaling the number of alleles  $K$ ) and so we assume them equally frequent in both populations. The number of polymorphic states differs: there are  $K(N - 1)$  polymorphic states in population  $A$ , while  $A'$  has  $K(M - 1)$ . Because we cannot make polymorphic states equivalent, we assume that the sum of polymorphic states for each pairwise comparison of the  $K$  alleles should be equal in both populations. These conditions can be written in the following system of equations

$$\begin{cases} P_{\{Na_i\}} = P'_{\{Ma_i\}} \\ \sum_{n=1}^{N-1} P_{\{na_i, (N-n)a_j\}} = \sum_{m=1}^{M-1} P'_{\{ma_i, (M-m)a_j\}} \end{cases} \quad (8)$$

As we have derived an estimator of the site frequency spectrum, we can write this conditions for the multivariate Moran model with boundary mutations and selection as

$$\begin{cases} \pi_{a_i}(1 + \sigma_{a_i})^{N-1} \frac{1}{k} = \pi'_{a_i}(1 + \sigma')^{M-1} \frac{1}{k'} \\ \pi_{a_i} \pi_{a_j} \frac{\rho_{a_i a_j}}{k} \sum_{n=1}^{N-1} (1 + \sigma_{a_i})^{n-1} (1 + \sigma_{a_j})^{N-n-1} \frac{N}{n(N-n)} = \pi'_{a_i} \pi'_{a_j} \frac{\rho'_{a_i a_j}}{k'} \sum_{m=1}^{M-1} (1 + \sigma'_{a_i})^{m-1} (1 + \sigma'_{a_j})^{M-m-1} \frac{M}{m(M-m)} \end{cases} \quad (9)$$

This system has  $K + K(K - 1)/2$  conditions and  $2K - 2 + K(K - 1)/2$  parameters and therefore cannot be solved. However, we know that the entries of  $\pi$  are constrained in  $[0, 1]$  and should sum up to 1 in both populations, therefore we make the additional assumption that  $\pi_{a_i} = \pi'_{a_i}$ . In addition, and by definition, the reference allele  $a_i^*$  is considered to evolve neutrally in both systems, which permits the conclusion that the normalization constants  $k$  and  $k'$  are equal. Simplifying,

$$\begin{cases} \pi_{a_i} = \pi'_{a_i} \\ (1 + \sigma_{a_i})^{N-1} = (1 + \sigma'_{a_i})^{M-1} \\ \rho_{a_i a_j} \sum_{n=1}^{N-1} (1 + \sigma_{a_i})^{n-1} (1 + \sigma_{a_j})^{N-n-1} \frac{N}{n(N-n)} = \rho'_{a_i a_j} \sum_{m=1}^{M-1} (1 + \sigma'_{a_i})^{m-1} (1 + \sigma'_{a_j})^{M-m-1} \frac{M}{m(M-m)} \end{cases} \quad (10)$$

we obtain that the population parameters of population  $A'$  can be expressed in terms of the parameters of population  $A$

$$\begin{cases} \pi'_{a_i} = \pi_{a_i} \\ (1 + \sigma'_{a_i}) = (1 + \sigma_{a_i})^{\frac{N-1}{M-1}} \\ \rho'_{a_i a_j} = \rho_{a_i a_j} \frac{\sum_{n=1}^{N-1} (1 + \sigma_{a_i})^{n-1} (1 + \sigma_{a_j})^{N-n-1} \frac{N}{n(N-n)}}{\sum_{m=1}^{M-1} (1 + \sigma_{a_i})^{\frac{N-1}{M-1}(m-1)} (1 + \sigma_{a_j})^{\frac{N-1}{M-1}(M-m-1)} \frac{M}{m(M-m)}} \end{cases} \quad (11)$$

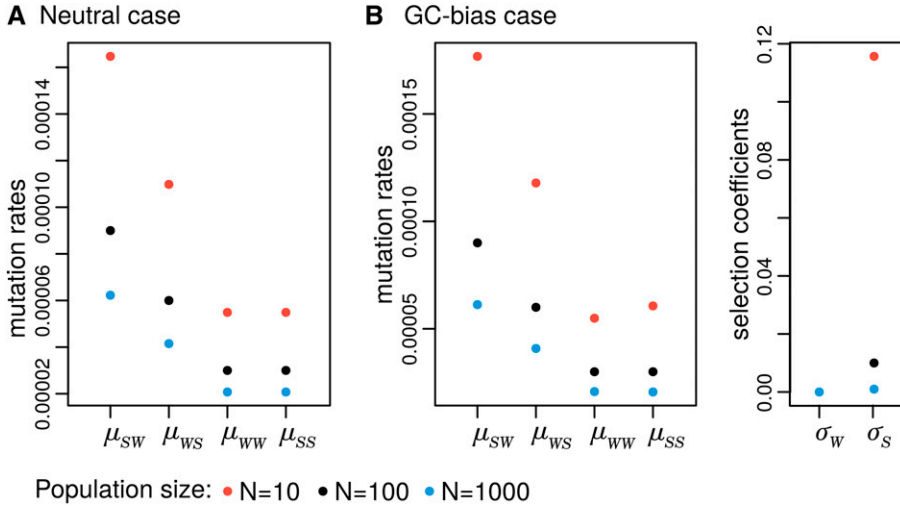
This expression looks tedious, but the neutral case ( $\sigma_{a_i} = 0$ ) can be very intuitive. In this scenario, mutation rates of populations  $A$  and  $A'$  change by a factor that is simply the ratio of two harmonic numbers, each of which determined by the population size of the respective population. Intuitively, if  $N > M$ , then  $\rho'_{a_i a_j} > \rho_{a_i a_j}$ , meaning that, in order to compensate the smaller number of individuals,  $M$  (i.e. stronger effect of genetic drift), mutation rates are augmented in population  $A'$ . Figure 5 depicts the effect of the effective population size on the mutation rates and selection coefficients.

## Appendix B

### Proof of the Stationary Vector

Let  $\psi$  be a stationary vector of  $Q$ , with  $\psi_{a_i a_j}^n$  and  $\psi_i$  being the elements of the stationary vector corresponding to the states  $\{na_i, (N - n)a_j\}$  and  $\{Na_i\}$ , respectively. In the multivariate Moran model with low mutation rates and selection, mutation is





**Figure 5** Population parameters transformation for different population sizes. We considered the simple case of two alleles: W stands for the weak alleles A and T, and S stands for the strong alleles C and G. Model parameters were set to represent (A) a neutral case (black dots:  $\mu_{SW} = 9 \times 10^{-5}$ ,  $\mu_{WS} = 6 \times 10^{-5}$  and  $\mu_{WW} = \mu_{SS} = 3 \times 10^{-5}$ ) and (B) a GC-bias case (black dots: mutations rates equal to the neutral case and  $\sigma_W = 0$ ,  $\sigma = 0.01$ ).

occurring only in the boundary states, permitting the monomorphic states to communicate with the polymorphic states, while drift and selection are both acting among the polymorphic states. The detailed balance conditions can be defined and lead to equations for the monomorphic and the polymorphic states. In the boundary states, an allele  $a_i$  is either fixed ( $n = N$ ) or absent ( $n = 0$ , i.e.  $a_j$  is fixed), for which we may write

$$\psi_i q_{a_i a_j}^{N \rightarrow N-1} = \psi_{a_i a_j}^{N-1} q_{a_i a_j}^{N-1 \rightarrow N} \quad \psi_j q_{a_i a_j}^{0 \rightarrow 1} = \psi_{a_i a_j}^1 q_{a_i a_j}^{1 \rightarrow 0}, \quad (12)$$

while, between the polymorphic states, the general condition is valid

$$\psi_{a_i a_j}^n q_{a_i a_j}^{n \rightarrow n+1} = \psi_{a_i a_j}^{n+1} q_{a_i a_j}^{n+1 \rightarrow n}. \quad (13)$$

Condition (13) can be rewritten in the recursive form

$$\psi_{a_i a_j}^{n+1} = \psi_{a_i a_j}^n \frac{q_{a_i a_j}^{n \rightarrow n+1}}{q_{a_i a_j}^{n+1 \rightarrow n}} \quad (14)$$

and then combined with Equation 12

$$\psi_i q_{a_i a_j}^{N \rightarrow N-1} = \psi_{a_i a_j}^n \frac{q_{a_i a_j}^{n \rightarrow n+1}}{q_{a_i a_j}^{n+1 \rightarrow n}} \cdots \frac{q_{a_i a_j}^{N-2 \rightarrow N-1}}{q_{a_i a_j}^{N-1 \rightarrow N-2}} q_{a_i a_j}^{N-1 \rightarrow N} = \psi_{a_i a_j}^n q_{a_i a_j}^{n \rightarrow n+1} \prod_{r=n+1}^{N-1} \frac{q_{a_i a_j}^{r \rightarrow r+1}}{q_{a_i a_j}^{r+1 \rightarrow r}}. \quad (15)$$

The product can be further simplified by recognizing that, for  $r = N - 1$ ,  $q_{a_i a_j}^{N \rightarrow N-1} = \mu_{a_i a_j} = \pi_{a_j} \rho_{a_i a_j}$ , while for  $r < N - 1$ , the rates inside the product are just the combined rate of drift and selection [according to expression (2)]. We can now rewrite Equation 14 in order to the  $\psi_{a_i a_j}^n$  element of the stationary vector of  $\mathbf{Q}$

$$\psi_{a_i a_j}^n = \frac{\psi_i \pi_{a_j} \rho_{a_i a_j}}{q_{a_i a_j}^{n \rightarrow n+1}} \left( \frac{1 + \sigma_{a_j}}{1 + \sigma_{a_i}} \right)^{N-n-1}. \quad (16)$$

Because  $\psi_{a_i a_j}^0 = \psi_j$  and  $q_{a_i a_j}^{0 \rightarrow 1} = \mu_{ji} = \pi_{a_i} \rho_{a_i a_j}$ , we obtain a possible solution for the monomorphic states of the stationary distribution by making  $n = 0$  in Equation 16

$$\frac{\psi_j}{\psi_i} = \frac{\pi_{a_j}}{\pi_{a_i}} \left( \frac{1 + \sigma_{a_j}}{1 + \sigma_{a_i}} \right)^{N-1}. \quad (17)$$

The stationary solution for the polymorphic states can be obtained from Equation 16 by noting that  $\psi_i = \pi_{a_i} \sigma_{a_i}^{N-1}$  and  $q_{a_i a_j}^{n \rightarrow n+1} = \frac{n(N-n)}{N} (1 + \sigma_{a_i})$



$$\psi_{a_i a_j}^n = \pi_{a_i} \pi_{a_j} \rho_{a_i a_j} (1 + \sigma_{a_i})^{n-1} (1 + \sigma_{a_j})^{N-n-1} \frac{N}{n(N-n)}. \quad (18)$$

The stationary distribution obtained here can be related with the stationary vector of the neutral boundary multivariate Moran model. We observe that, when  $\sigma = 0$ , we obtain the solution computed by Schrempf *et al.* (2016) for the multivariate Moran model with drift only

$$\psi_i = \pi_{a_i} \quad \psi_{a_i a_j}^n = \pi_{a_i} \pi_{a_j} \rho_{a_i a_j} \frac{N}{n(N-n)}. \quad (19)$$

## Appendix C

### Proof of the Expected Number of Moran Events per Unit of Time

To assess the impact of allelic selection in branch length estimation (or the total rate of the process), we computed the expected number of events per unit of time for the multivariate Moran model with selection

$$d_{MS}(t=1) = - \sum_u \psi_u q_{uu}, \quad (20)$$

Where  $\psi$  is the stationary vector and  $q_{uu}$  the diagonal elements of  $Q$ . Equation 20 can be solved by observing that a monomorphic state can be escaped only by mutation, while a polymorphic state can be escaped only by selection and drift

$$d_{MS} = \sum_{a_i \in \mathcal{A}} \sum_{j \neq i} \psi_i \mu_{a_i a_j} + \sum_{a_i a_j \in \mathcal{A}^c} \sum_{n=1}^{N-1} \psi_{a_i a_j}^n \frac{n(N-n)}{N} (1 + \sigma_{a_i} + 1 + \sigma_{a_j}). \quad (21)$$

The stationary vector is known from Equations 17 and 18

$$d_{MS} = \frac{1}{k} \sum_{a_i \in \mathcal{A}} \sum_{j \neq i} (1 + \sigma_{a_i})^{N-1} \pi_{a_i} \rho_{a_i a_j} \pi_{a_j} + \frac{1}{k} \sum_{a_i a_j \in \mathcal{A}^c} \sum_{n=1}^{N-1} \pi_{a_i} \rho_{a_i a_j} \pi_{a_j} \left[ (1 + \sigma_{a_i})^n (1 + \sigma_{a_j})^{N-n-1} + (1 + \sigma_{a_i})^{n-1} (1 + \sigma_{a_j})^{N-n} \right], \quad (22)$$

where  $k$  is the normalization constant defined in Equation 4. Expression (22) can be further simplified by observing that the sum in  $n$  results in doubling every  $(1 + \sigma_{a_i})^{n-1} (1 + \sigma_{a_j})^{N-n}$  element. Therefore, the expected number of events can be simplified to

$$d_{MS} = \frac{2}{k} \sum_{a_i a_j \in \mathcal{A}^c} \sum_{n=1}^N \pi_{a_i} \rho_{a_i a_j} \pi_{a_j} (1 + \sigma_{a_i})^{n-1} (1 + \sigma_{a_j})^{N-n}. \quad (23)$$

## Appendix D

### Proof of the Moran Distance in Number Substitutions

The Moran distance  $d_{MS}$  accounts for several events (mutation, drift, and selection), and differs from the standard evolutionary distances because they are calculated in terms of the expected number of substitutions  $d_{MS}^*$ .

$$d_{MS}^* = d_{MS} \times s, \quad (24)$$

where  $s$  is the probability of a substitution.  $s$  can be calculated multiplying the probability  $m$  of an event being a mutation, by the probability  $h$  of that mutation getting fixed in the population

$$s = \sum_{a_i a_j \in \mathcal{A}^p} s_{a_i \rightarrow a_j} = \sum_{a_i a_j} m_{a_i \rightarrow a_j} \times h_{j|i}, \quad (25)$$

where  $\mathcal{A}^p$  represents all the possible pairwise permutations without repetition of  $K$  alleles.

### Solving $m_{a_i \rightarrow a_j}$

The probability of an event being a mutation is simply the ratio between the rate of mutation and the total rate (*i.e.*, the rate of mutation plus the rate of drift and selection). In stationarity, we know that the total rate  $r_T = d_{MS}(1)$  is simply the expected number of events of the Moran model and follows Equation 23. The rate of a  $a_i \rightarrow a_j$  mutation is the rate of escaping the monomorphic state  $\{Na_i\}$ , from which we can write

$$m_{a_i \rightarrow a_j} = \frac{r_{i \rightarrow j}}{r_T} = \frac{\pi_{a_i} \pi_{a_j} \rho_{a_i a_j} (1 + \sigma_{a_i})^{N-1}}{2 \sum_{a_i, a_j \in \mathcal{A}^c} \sum_{n=1}^N \pi_{a_i} \pi_{a_j} \rho_{a_i a_j} (1 + \sigma_{a_i})^{n-1} (1 + \sigma_{a_j})^{N-n}}. \quad (26)$$

We can see that  $m_{a_i \rightarrow a_j}$  differs from  $m_{a_j \rightarrow a_i}$  only due to the selection coefficient in the numerator.

### Solving $h_{a_i|a_j}$

According to Kluth and Baake (2013), the fixation probability of an allele with fitness  $1 + \sigma$  is for the Moran model

$$h = \frac{(1 + \sigma)^{N-1}}{\sum_{n=0}^{N-1} (1 + \sigma)^n}. \quad (27)$$

As we are using the multivariate Moran model, we have to extend the denominator of (27) to account for the different possible combinations of two selection coefficients. In particular, we may have

$$h_{a_i|a_j} = \frac{(1 + \sigma_{a_i})^N}{\sum_{n=1}^N (1 + \sigma_{a_i})^n (1 + \sigma_{a_j})^{N-n}} \quad \text{and} \quad h_{a_j|a_i} = \frac{(1 + \sigma_{a_j})^N}{\sum_{n=1}^N (1 + \sigma_{a_j})^n (1 + \sigma_{a_i})^{N-n}}. \quad (28)$$

We further redefine the denominators in order to make them equal

$$h_{a_i|a_j} = \frac{(1 + \sigma_{a_i})^N (1 + \sigma_{a_j})}{\sum_{n=1}^N (1 + \sigma_{a_i})^n (1 + \sigma_{a_j})^{N-n+1}} \quad \text{and} \quad h_{a_j|a_i} = \frac{(1 + \sigma_{a_j})^N (1 + \sigma_{a_i})}{\sum_{n=1}^N (1 + \sigma_{a_j})^n (1 + \sigma_{a_i})^{N-n+1}}. \quad (29)$$

### Solving $s$

The probability of a  $a_i \rightarrow a_j$  substitution under the multivariate Moran model with boundary mutations and selection can be computed as

$$s_{a_i \rightarrow a_j} = m_{a_i \rightarrow a_j} \times h_{a_j|a_i} = \frac{\pi_{a_i} \pi_{a_j} \rho_{a_i a_j} (1 + \sigma_{a_i})^N (1 + \sigma_{a_j})^N}{2 \times \sum_{a_i, a_j \in \mathcal{A}^c} \sum_{n=1}^N \pi_{a_i} \pi_{a_j} \rho_{a_i a_j} (1 + \sigma_{a_i})^{n-1} (1 + \sigma_{a_j})^{N-n} \times \sum_{n=1}^N (1 + \sigma_{a_j})^n (1 + \sigma_{a_i})^{N-n+1}}. \quad (30)$$

We see that  $s_{a_i \rightarrow a_j} = s_{a_j \rightarrow a_i}$ , which is an expected consequence of stationarity. We can now generalize  $s_{a_i \rightarrow a_j}$  for all the substitution types by using Equation 25

$$s = \frac{1}{\sum_{a_i, a_j \in \mathcal{A}^c} \sum_{n=1}^N \pi_{a_i} \pi_{a_j} \rho_{a_i a_j} (1 + \sigma_{a_i})^{n-1} (1 + \sigma_{a_j})^{N-n}} \sum_{a_i, a_j \in \mathcal{A}^c} \frac{\pi_{a_i} \pi_{a_j} \rho_{a_i a_j} (1 + \sigma_{a_i})^N (1 + \sigma_{a_j})^N}{\sum_{n=1}^N (1 + \sigma_{a_j})^n (1 + \sigma_{a_i})^{N-n+1}}. \quad (31)$$

The relationship between the Moran distance in events and substitutions can be defined based on Equation 24,

$$d_{MS}^* = d_{MS} \frac{1}{\sum_{a_i, a_j \in \mathcal{A}^c} \sum_{n=1}^N \pi_{a_i} \pi_{a_j} \rho_{a_i a_j} (1 + \sigma_{a_i})^{n-1} (1 + \sigma_{a_j})^{N-n}} \sum_{a_i, a_j \in \mathcal{A}^c} \frac{\pi_{a_i} \pi_{a_j} \rho_{a_i a_j} (1 + \sigma_{a_i})^N (1 + \sigma_{a_j})^N}{\sum_{n=1}^N (1 + \sigma_{a_j})^n (1 + \sigma_{a_i})^{N-n+1}}. \quad (32)$$

This quantity can be evaluated for neutral regimes: *i.e.*  $\sigma \rightarrow (0, 0, 0, 0)$ . We obtain the probability of a substitutions under the neutral Moran model, and it matches the result computed by Schrempf *et al.* (2016):

$$d_{MS}^* = d_{MS} \frac{1}{N^2} \quad (33)$$