Quantifying GC-biased gene conversion in great ape genomes using polymorphism-aware models

Rui Borges[1], Gergely Szöllősi[2], and Carolin Kosiol[3*]

1. Institute of Population Genetics, Vetmeduni Vienna, Veterinärplatz 1, 1210 Wien, Austria

2. Department of Biological Physics, ELTE-MTA "Lendulet" Biophysics Research Group, Eötvös University, Pázmány P. stny. 1A, Budapest H-1117, Hungary.

3. Centre for Biological Diversity, University of St Andrews, St Andrews, Fife KY16 9TH, UK

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

Abstract

As multi-individual genome-wide population-scale data is becoming available, more-complex modeling strategies are needed to quantify the 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 proposed 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, which was employed to quantify the patterns of genome-wide GC-biased gene conversion in great apes. Importantly, we showed that great apes have patterns of allelic selection that vary in intensity, a feature that we correlated with the great apes' distinct demographies. We also demonstrate that the AT/GC toggling effect decreases the probability of a substitution, which promotes more polymorphisms in the base composition of great ape genomes. We assessed the impact of CG-bias in molecular analysis and we found that mutation rates and genetic distances are estimated under bias when gBGC is not properly accounted. Our results stress the need for gBGC-aware models in population genetics and phylogenetics.

Keywords: Moran model, boundary mutations, allelic selection, great apes, GC-bias, gBGC

1 Introduction

The field of molecular population genetics is currently been revolutionized by progress in data acquisition. New challenges are emerging as new lines of inquiry posed by increasingly large population-scale sequence data (Casillas and Barbadilla, 2017). Mathematical theory describing population dynamics has been dormant since the early days of population genetics (e.g. Fisher (1930); Wright (1931); Moran (1958); Kimura (1964)), but now that data is 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. 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, the GC alleles are systematically preferred. gBGC is a recombination-associated segregation bias that favors GC-alleles (or strong alleles, hereafter) over AT-alleles (or weak alleles, hereafter) during the repair of mismatches that occur within heteroduplex DNA during meiotic recombination (Marais, 2003). The process of gBGC was studied in several recent publications (e.g. Webster et al. (2006); Escobar et al. (2011); Pessia et al. (2012); Serres-Giardi et al. (2012); Galtier et al. (2018)), but its impact was mainly studied in mammalian genomes (Duret and Galtier, 2009; Romiguier et al., 2010). Apart from some studies in human populations (Katzman et al., 2011; Glémin et al., 2015), 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. The effect of demography on gBGC is still controversial. While theory and some empirical studies advocate a positive relationship between the effective population size and the intensity of gBGC (Nagylaki, 1983; Glémin et al., 2015), Galtier et al. (2018) failed to detect such relationship. 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 questions by revisiting the great ape data (Prado-Martinez et al., 2013) with a Moran model that accounts for allelic selection, which in principle may be able to capture both, episodes of directional selection and gBGC.

The Moran model (Moran, 1958) has a central position describing populations' evolution: 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 (De Maio et al., 2015; 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 we build a Bayesian framework to estimate population parameters (base composition, mutation rates, and selection coefficients) from population data.

Furthermore, our application to great apes shows that most great apes have patterns of GC-bias consistent with gBGC. Our results suggest further that demography has a major role in determining the intensity of gBGC among great apes, as the intensity of allelic selection among the great ape populations significantly correlates with their effective population size. 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.

2 Methods

2.1 The multivariate Moran model with allelic selection

We define the multivariate Moran model with boundary mutations and allelic selection following the terminology proposed by Vogl and Bergman (2015) and Schrempf and Hobolth (2017).

Consider a haploid population of N individuals and a single locus with K alleles: i and j are two possible alleles. The evolution of this population in the course of 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 states can be defined: if all the individuals in a populations have the same allele, the population is monomorphic $\{i\}$, i.e. the N individuals have the allele i; differently, if two alleles are present in the population, the population is polymorphic $\{ni, (N-n)j\}$, meaning that n individuals have the allele i and $(N-n)$ have the allele j.

The trajectory of alleles is defined based on the rate matrix $\boldsymbol{Q}$. Time was accelerated by a factor of N, and therefore instead of describing the Moran dynamics in terms of Moran events (Moran, 1958), we developed a continuous version in which the time is measured in coalescent time.

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

$$q^{\{ni,(N-n)j\}\to\{(n+1)i,(N-n-1)j\}} = q^{\{ni,(N-n)j\}\to\{(n-1)i,(N-n+1)j\}} = \frac{n(N-n)}{N} \qquad (1)$$

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 only occur in the boundary states. The boundary mutations assumption is met if the mutation rates μ_{ij} are small (and N is not too large); more specifically, Schrempf et al. (2016) established that $N\mu_{ij}$ should be lower than 0.1, by comparing the expectations of the diffusion equation with the polymorphic diversity under the Moran model. In fact, most eukaryotes fulfill this condition (Lynch et al., 2016). 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 for 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): $\mu_{ij} = \rho_{ij}\pi_j$, where $\boldsymbol{\rho}$ represents the exchangeabilities between any two alleles and $\boldsymbol{\pi}$ the allele base composition (rate matrix (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.

We modeled allelic selection by defining $K-1$ relative selection coefficients $\boldsymbol{\sigma}$: an arbitrary selection coefficient is fixed to 0. Defining the fitness as the probability that an offspring of allele i is replaced with probability $1+\sigma_i$ (Durrett, 2008), we can formulate the component of allelic selection alongside with drift, and thus among the polymorphic states (rate matrix (2)).

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

$$q^{\{ui,(N-u)j\}\to\{vi,(N-v)j\}} = \begin{cases} \mu_{ij} = \rho_{ij}\pi_j & u = N,\ v = N-1 \\ \mu_{ji} = \rho_{ij}\pi_i & u = 0,\ v = 1 \\ \frac{n}{N}(N-n)(1+\sigma_i) & u = n,\ v = n+1,\ 0 < n < N \\ \frac{n}{N}(N-n)(1+\sigma_j) & u = n,\ v = n-1,\ 0 < n < N \\ 0 & |u-v| > 1 \end{cases}, \qquad (2)$$

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 down them to a value small value N while keeping the overall dynamics unchanged. 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.

2.2 The stationary distribution

The stationary distribution of a Markov process can be obtained by computing the vector $\boldsymbol{p}$ satisfying the condition $\boldsymbol{\phi Q} = \mathbf{0}$ (File S1). $\boldsymbol{\phi}$ is the normalized stationary vector and has the solution:

$$\phi_x = \begin{cases} \pi_i(1+\sigma_i)^{N-1}k^{-1} & \text{if } x = \{i\} \\ \pi_i\pi_j\rho_{ij}(1+\sigma_i)^{n-1}(1+\sigma_j)^{N-n-1}\frac{N}{n(N-n)}k^{-1} & \text{if } x = \{ni, (N-n)j\} \end{cases}. \tag{3}$$

k is the normalization constant

$$k = \sum_{i\in\mathcal{A}} \pi_i(1+\sigma_i)^{N-1} + \sum_{ij\in\mathcal{A}^C}\sum_{n=1}^{N-1} \pi_i\pi_j\rho_{ij}(1+\sigma_i)^{n-1}(1+\sigma_j)^{N-n-1}\frac{N}{n(N-n)}, \tag{4}$$

where $\mathcal{A}$ is the alphabet of the K alleles $\{a_1, \ldots, a_K\}$, 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, ..., $a_{K-1}a_K$.

2.3 Expected number of Moran events

From $\boldsymbol{Q}$ and $\boldsymbol{\phi}$, we can compute the expected number of Moran events (mutations, drift and selection). These are the expected state-changes per unit of time for the multivariate Moran model with selection (File S2)

$$d_S(t=1) = d_S = \frac{2}{k}\sum_{ij\in\mathcal{A}^C}\sum_{n=1}^{N} \pi_i\rho_{ij}\pi_j(1+\sigma_i)^{n-1}(1+\sigma_j)^{N-n}. \tag{5}$$

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 $\boldsymbol{\sigma} \to \mathbf{0}$. To compare the Moran distance d_S with the standard models of evolution, we recalculated the Moran distance to only account for substitutions events d_S^*: we corrected d_S by the probability of a mutation and a subsequent fixation under the Moran model (File S3)

$$d_S^* = \frac{2}{k}\sum_{ij\in\mathcal{A}^C} \frac{\pi_i\pi_j\rho_{ij}(1+\sigma_i)^N(1+\sigma_j)^N}{\sum_{n=1}^{N}(1+\sigma_j)^n(1+\sigma_i)^{N-n+1}}. \tag{6}$$

2.4 Bayesian inference with the stationary distribution

We can define a likelihood function for the stationary distribution for a set of S independent sites in N individuals by taking the product of $\boldsymbol{p}$ over the number of monomorphic and polymorphic sites: $\#\{i\}$ and $\#\{ni, (N-n)j\}$, respectively

$$p(\boldsymbol{x}|\boldsymbol{\pi}, \boldsymbol{\rho}, \boldsymbol{\sigma}) = \prod_{s=1}^{S} p(x_s) = k^{-S} \prod_{i \in \mathcal{A}} \left[\pi_i (1+\sigma_i)^{N-1}\right]^{\#\{i\}} \times \prod_{ij \in \mathcal{A}^C} \prod_{n=1}^{N-1} \left[\pi_i \pi_j \rho_{ij} (1+\sigma_i)^{n-1} (1+\sigma_j)^{N-n-1} \frac{N}{n(N-n)}\right]^{\#\{ni,(N-n)j\}}. \quad (7)$$

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 guaranteeing a target acceptance rate of 0.234. We employed the Metropolis-Hastings algorithm (Hastings, 1970) for each conditional posterior in a Markov chain Monte Carlo sequence to obtain random samples from the posterior. The algorithm was coded in the R statistical programing 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). The R script can be assessed in the GitHub branch `pomo-dev/pomo_selection`.

2.5 Polymorphism-aware phylogenetic model

The multivariate Moran model can be also referred as a polymorphism-aware phylogenetic model (PoMo) if we set $k = 4$ alleles (De Maio et al., 2013, 2015; Schrempf et al., 2016), those representing the 4 nucleotide bases. We write $\mathcal{A}$ as the alphabet of the 4 nucleotide bases $\{A, C, G, T\}$ and $\mathcal{A}^C$ as all the possible pairwise combinations of the four nucleotide bases $\{AC, AG, AT, CG, CT, GT\}$. For a population of size N we have $4 + 6(N-1)$ possible states, four of which are monomorphic (Figure 1). Applications and results presented in the following pages were obtained using the 4-variate model.

2.6 Application: great ape population data

The stationary distribution of 4-multivariate model was employed to infer the distribution of allele frequencies, selection coefficients and mutation rates from 4-fold degenerate sites of exome-wide population data from great apes (Prado-Martinez et al., 2013). We used 11 populations with up to 23 individuals, totaling $\sim$ 2.8 million sites per population (Table 1).

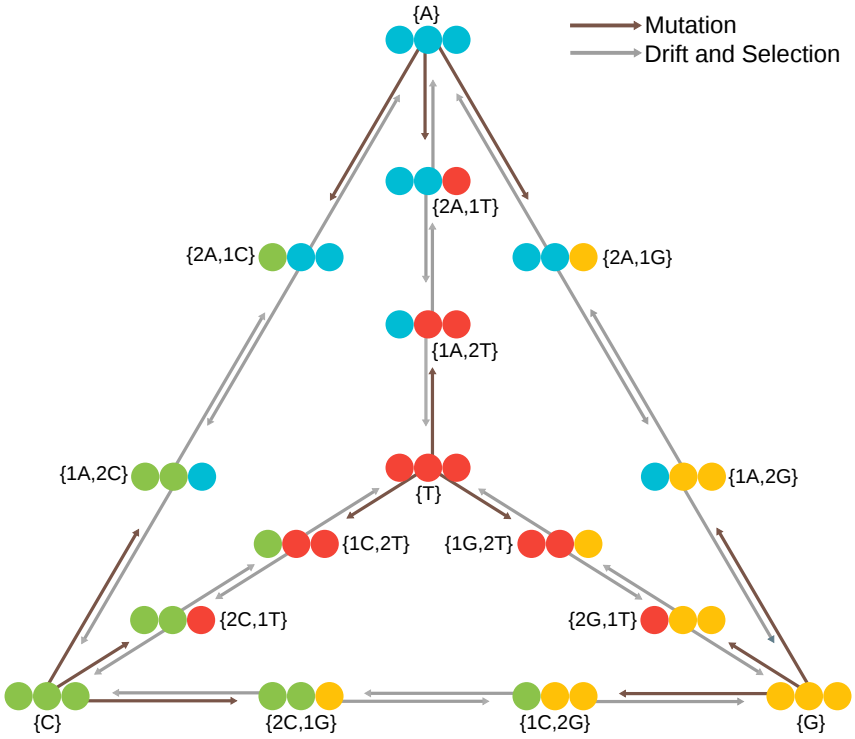

Figure 1: **PoMo state-space using $N = 3$.** The 4 alleles represent the four nucleotide bases. Brow and grey arrows indicate mutations, and genetic drift and selection, respectively. Monomorphic or boundary states $\{i\}$ are represented in the tetrahedron's vertices, while the polymorphic states $\{ni, (N-n)j\}$ are represented in its edges. Monomorphic states interact with polymorphic states via mutation, but a polymorphic can only reach a monomorphic state via drift or selection. Between polymorphic states only drift and selection events occur.

Data preparation follows the pipeline described in De Maio et al. (2015). The allelic counts of all 11 primate subspecies are available in the GitHub branch `pomo-dev/pomo_selection`. Estimates of the Watterson's θ genetic diversity is below 0.003 for all the studied populations (Schrempf et al., 2016), validating the boundary mutations assumption of 0.1.

3 Results

3.1 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 $\boldsymbol{Q}t$ for large enough t. We confirmed that the numerical solution converges to the analytical solution (Figure S1).

We validated the Bayesian algorithm for estimating population parameters from the stationary distribution by performing simulations (Table S1 and Figures S2-S5). Our algorithms efficiently recover the true population parameters from simulated allele counts. We tested the algorithms for different number of sites (10^3, 10^6 and 10^9) and state-spaces ($N = 5$, 10 and

50). While the number of sites does not increase the computation time substantially and is not being a limiting factor for genome-wide analysis, 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.

3.2 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 multivariate Moran model (M_M) and the model with allelic selection (M_S). Using the predictive stationary distribution and the observed allele counts, we computed the Bayes' factors favoring the more complex model M_S (i.e. log BF > 0 favors the model with allelic selection) for all populations. It is clear that M_S fits the data considerably better for most of the studied great apes (log BF > 100, Table 1). The only exception is the Eastern gorillas population, for each a lower log BF was obtained (log BF $= 5.497$, Table 1).

Table 1: **Evidence of allelic selection among the great ape populations.** The number of individuals and the number 4-fold degenerate sites per population are indicated by I and S, respectively. The log Bayes' factors (log BF) were calculated as the sum over the product of the allele counts and the posterior predictive probabilities under the Moran model with boundary mutations (M_M) and allelic selection (M_S). BF favor the model with allelic selection when higher than 1.

Population	I	S	$\log p(\boldsymbol{x}\|M_M)$	$\log p(\boldsymbol{x}\|M_S)$	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

We have also corroborated our Bayes' factors by inspecting the fit of the predictive distribution of alleles of M_M and M_S with the allele counts (Figure S6A-K). The allele counts for the polymorphic states are not symmetric, generally one allele if preferred and so are the polymorphic states that have it in higher proportions. As expected, we observed that M_S better reproduces the skewed distribution of allele counts among great apes.

We further explored the parameter estimates under M_S to know how strong and variable are the patterns of allelic selection among great apes. We analyzed the posterior distribution of the relative selection coefficients of C, G, and T (σ_A was set to 0). 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 the 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 Bayes' factor found for evidence of allelic selection in the Eastern gorilla population.

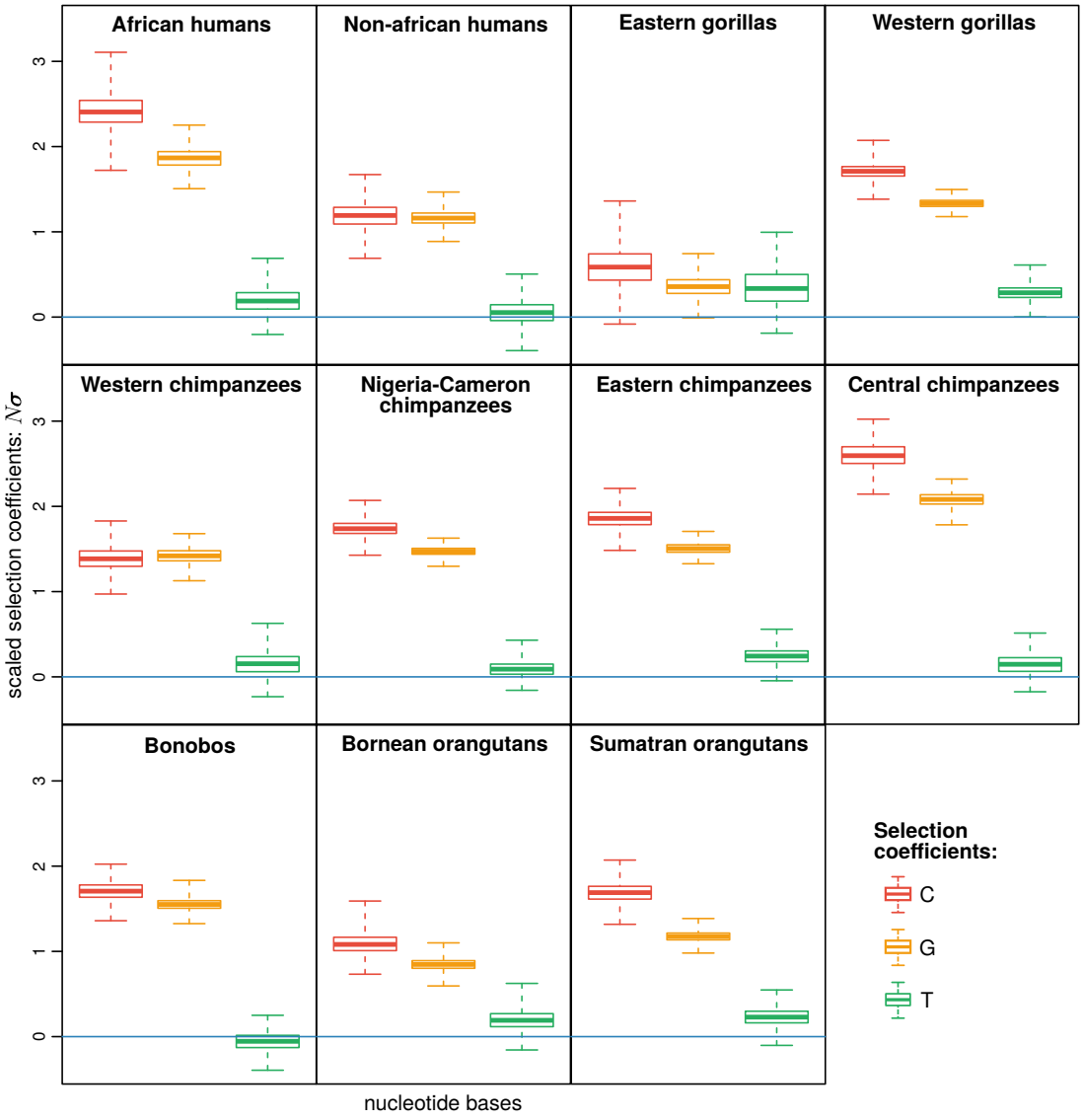

Figure 2: **Scaled allelic selection coefficients for the great apes 4-fold degenerate synonymous sites.** The boxplots represent the posterior distribution of the C, G and T scaled selection coefficients (σ_A was set to 0); the estimates were obtained using the 4-variate Moran model. The line in blue represents $\sigma_A = 0$. Table S2 summarizes the average scaled selection coefficients for each great ape 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 ($\sigma_C + \sigma_G$) with the chromosome size and recombination rate in the non-African human population (Figure S7), for which this data is particularly well characterized (Jensen-Seaman, 2004). We found a

significant positive correlation between the GC-bias and recombination rate (Spearman's ρ = 0.57, p-value = 0.006), but a negative correlation with the chromosome length (Spearman's ρ = -0.52, p-value = 0.014).

Although the patterns of selection among great apes are similar qualitatively, they differ quantitatively. For example, the Central chimpanzees have patterns of GC-bias around 2.08/2.60 (σ_C/σ_G, Table S2 and Figure 2), while the closely related population of Western chimpanzees shows less strong patterns (around 1.38/1.42). Likewise, the GC-bias content in African and non-African human populations contrasts: 2.41/1.86 and 1.19/1.16, respectively. These results show that the patterns of allelic selection greatly vary among great apes, even among closely related populations.

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; Philippe et al., 2011): the AT/GC toggling effect. Congruently with this expectations, we observed that mutation rates from strong to weak alleles are higher (by a factor of 3.05 in average), but rather similar between alleles of the same type (around 1.02 in average; supplementary table S2), while the selection coefficients, as shown, have a clear pattern of GC-bias in most of the great ape populations.

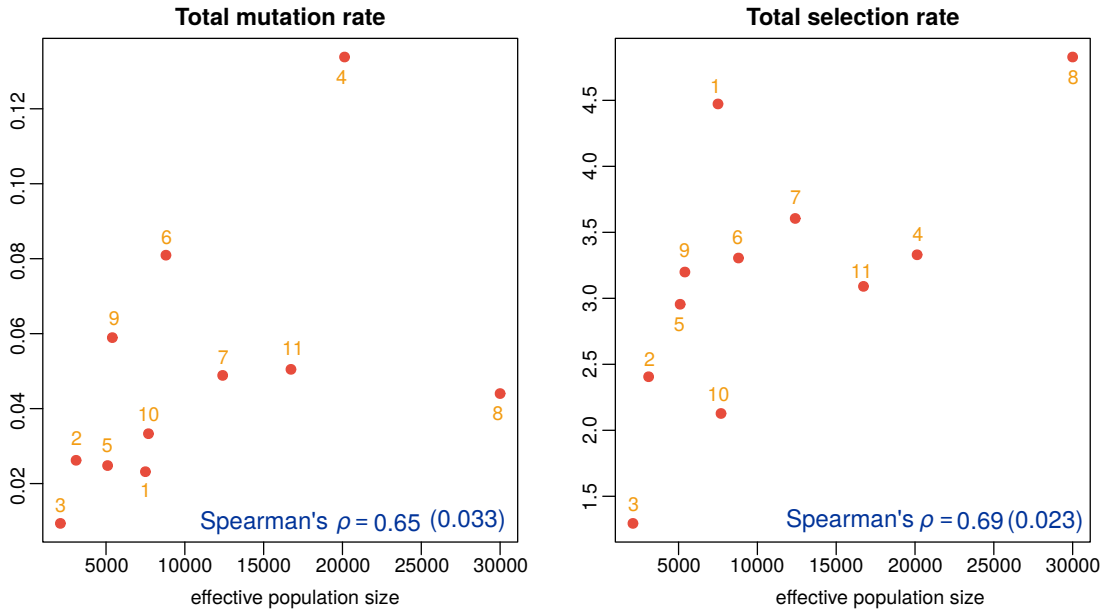

Figure 3: **Correlating N_e and the total rate of mutation and selection in great apes.** Great ape populations are numbered: 1. African human, 2. Non-African human, 3. Eastern gorilla, 4. Western gorilla, 5. Western chimpanzee, 6. Nigeria-Cameroon chimpanzee, 7. Eastern chimpanzee, 8. Central chimpanzee, 9. Bonobo, 10. Bornean orangutan and 11. Sumatran orangutan. Estimates of N_e were taken from Prado-Martinez et al. (2013) and Tenesa et al. (2007).

3.3 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 and significant correlations between the total mutation and selection rates and the effective population size were obtained (Figure 3): Spearman's correlation coefficient of 0.65 (p-value = 0.033) and 0.69 (p-value = 0.023), respectively.

This result shows that N_e plays an important role in determining the intensity of mutations and selection. In particular, it becomes clear that the different patterns of GC-bias found among great apes are, in part, due 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.

Table 2: **Expected number of substitutions per unit of time**. The expected number of substitutions for the multivariate Moran model with boundary mutations d_M^* and allelic selection d_S^* were calculated based on the posterior distributions of the model parameters and equation (6). The relative difference between these distances was calculated as the ratio between the average number of events between the two models (d_S^*/d_M^*) and was used to assess how dissimilar these distances are.

Population	$d_M^* \times 10^3$	$d_S^* \times 10^3$	d_S^*/d_M^*
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

3.4 Comparing the expected number of substitutions in great apes

We calculated the expected number of substitutions under $\mathrm{M_M}$ and $\mathrm{M_S}$ to evaluate the impact of allelic selection (in particular, GC-bias) in the evolutionary process. With formula (6), we calculated d_M^* and d_S^* using the posterior estimates of the respective model parameters. We

observe that for most of the great ape populations, the expected number of substitutions is lower when allelic selection is accounted (Table 2). Eastern gorillas are an exception, and the opposite pattern was observed. We also calculated the relative difference between the expected number of substitutions in both models (i.e. d_S/d_M) and we obtained minor (-0.26% in Bornean orangutans) to major (-17.8% in bonobos) relative differences; 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.

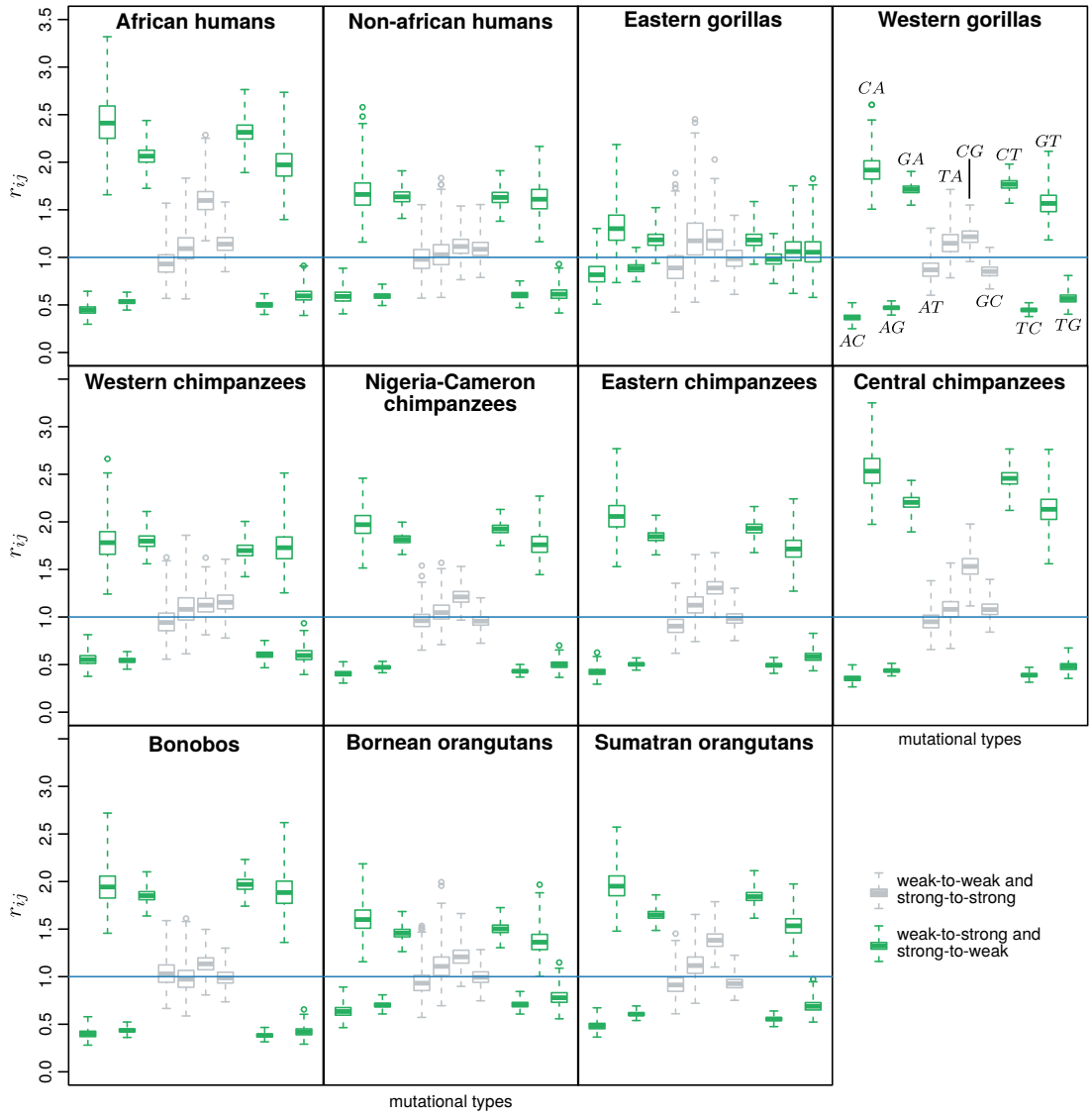

Figure 4: **Relative difference in the mutation rates estimated under the neutral and non-neutral Moran model.** r_{ij} represents the ratio between the mutation from allele i to allele j in the model with allelic selection and the model with boundary mutations: $r_{ij} = \mu^S_{ij}/\mu^M_{ij}$. The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

We complement this result by comparing the posterior distribution of the mutations rates in M_M and M_S. Because we wanted to identify the mutational types that may be differently estimated between these models, we calculated the relative difference between the mutation

rate from allele i to allele j under the models, respectively: $r_{ij} = \mu^S_{ij}/\mu^M_{ij}$. If $r_{ij} > 1$ for a certain mutation rate ij, then this mutation rate is being underestimated in $\mathrm{M_M}$ when compared to $\mathrm{M_S}$ (and *vice versa* if $r_{ij} < 1$); if $r_{ij} \approx 1$ 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 non-differentially 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 $\mathrm{M_M}$. 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 population parameters. Eastern gorillas behave differently by not showing significant differences between the estimated mutations rates (all r_{ij} overlap 1, Figure 4).

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

4.1 Variable patterns gBGC among great apes

A genome-wide application in 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 non-human 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 (around 2.8 million sites in our case) are summarized to perform inferences.

The patterns of GC-bias we have 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 (because at least one crossover per chromosome is necessary for proper segregation during meiosis, the crossover rate (in cM per Mb) will be higher for small than for large chromosomes (Farré et al., 2013)). We have performed these analyses in non-African Humans, for which this data is available; however, we are confident that the patterns of GC-bias found in great apes are due to gBGC.

In agreement with previous studies in mammals and humans (Spencer et al., 2006; Lartillot, 2013; Capra et al., 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 ranges between 1.49 ± 0.53, consistent with the nearly-neutral scenario (Ohta and Gillespie, 1996; Vogl and Bergman, 2015). These estimates are in congruence with other 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 to predict scaled conversion coefficients lower than 1 in all apes. Notice that the latter works employed the Wright-Fisher model. As the rate of genetic drift is twice as fast in the Moran model, we expect to estimate twice as hight selection coefficients with our approach.

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 (Galtier et al., 2009; Duret and Galtier, 2009; Philippe et al., 2011). Congruently with this expectations, we observed that mutation rates from strong to weak alleles are higher but rather similar between alleles of the same type. Interestingly, this symmetric manner by which mutations and selection are acting in great apes leads, as we have demonstrated, the number of substitutions to decrease in average, which suggests that the AT/GC toggling may actually increase the population variability by promoting more polymorphic sites.

Glémin et al. (2015) hypothesized that differences in GC-bias intensity among human populations were due to 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 the patterns of gBGC can rapidly change due to demography, even among closely related populations. In fact, most of the studied populations are known to have diverged less than 0.5 million years ago (Prado-Martinez et al., 2013).

Here, we showed that GC-bias determines the genome-wide base composition of genomes in a factor proportional to $(1+\sigma_{C/G})^{N-1}$ (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, we are convinced that demography has a major role determining the base composition of great apes genomes. Intriguingly, Galtier et al. (2018) have not found this correlation at the species level in animals. This is most likely happening because genome-wide recombination rate, length of gene conversion tracts and repair biases should significantly vary across species, but not so much across related populations, which explains why the correlation between the intensity of gBGC and N_e was found in great ape populations, but not more generally in animals.

While correlating the strength of selection with N_e, we obtained a correlation coefficient (0.69) that suggests that other processes may be determining the strength of allelic selection: we can refer two likely reasons. First, the effect of recent demographic effects. We have considered a fixed population size and stationarity, which are good assumptions to recover long-standing population processes, but may not capture the more-recent demographic events and therefore, their impact on GC-bias. Second, variations in s due to species-specific recombination landscapes may also contribute to different GC-bias. Indeed, variation in the karyotypes (number and length of chromosomes) and the short-life and self-destructive nature of recombination hotspots are known to contribute to generating different patterns of GC-bias among species (Duret and Galtier, 2009; Lesecque et al., 2014). For the particular case of great apes, changes in the karyotype should not be a major aspect, as it is very conserved among primates: humans have 46 diploid chromosomes whereas the other great apes (orangutans, gorillas, and chimps) have 48. However, it is known that humans and chimpanzees, and even human populations share few recombination hotspots (Auton et al., 2012; Lesecque et al., 2014), which may explicate differences in the great apes' recombination landscapes and, ultimately, why the intensity of allelic selection cannot be completely explained by the effects of demography.

Knowing to what extent variations in N_e or s determine the base composition of genomes will require further studies. In particular, determining s experimentally, ideally in different populations, would help to assess the real impact of gBGC and how variable it is among species/populations. If as for the mutation rate, we could assume that s vary 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.

4.2 gBGC calls for caution in molecular and phylogenetic analyses

The effects of gBGC in the molecular analysis have been extensively described 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. In particular, we observed that mutations rates from weak-to-strong and strong-to-weak alleles are systematically over and underestimated, respectively. Biased estimators are not necessarily worthless (particularly when the bias is known) for parameter estimation. Being able to describe the distribution of alleles with fewer parameters is in principle a good aspect modeling-wise. However, we have strong evidence that the model with allelic selection much better fits the observed allele counts for all the studied great ape populations.

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.

Recently, a nucleotide substitution process that accounts for gBGC was proposed by Lartillot (2013). In this model, the scaled conversion coefficient is used to correct the substitution rates in a similar fashion as we have done to calculate the expected number of substitutions for the Moran distance (i.e. assessing the relative fixation probabilities under GC-bias, File S3). Therefore, we expect to obtain similar results with this nucleotide substitution model and our model: the only differences being that our model accounts for polymorphic sites and is based on the Moran model (while in Lartillot (2013) populations follow the Wright-Fisher model).

5 Conclusion

Despite the widespread evidence of gBGC in several taxa, several questions remain open regarding the role of gBGC determining the base composition of genomes. In this work, we quantify the patterns of gBGC in great apes while contributing to the discussion of the tempo and mode of gBGC evolution in vertebrate genomes.

Our Moran model adds a significant contribution to the endeavor of estimating population parameters from multi-individual, genome-wide population-scale data. Our model was used to estimate genome-wide signature of gBGC, but it can also be more generally employed to estimate patterns of nucleotide usage and associated mechanisms of evolution. Importantly, our analysis showed that gBGC may significantly distort estimates of population parameters and genetic distances, stressing that gBGC-aware models should be used when employing molecular phylogenetics and population genetics analyses.

Here, we have not performed phylogenetic inference, but previous applications of the Moran model to phylogenetic problems (De Maio et al., 2015; Schrempf et al., 2016) show that it can be done. Therefore, a necessary future work would be testing the effect of gBGC in phylogeny reconstruction, in particular, determining how much of its signal can be accounted for increasing the accuracy of tree estimation both on the topology and branch lengths.

Acknowledgements

This work has been funded by the Vienna Science and Technology Fund (WWTF) through project MA16-061. We thank Dominik Schrempf for helpful comments on the manuscript.

References

Auton, A., Fledel-Alon, A., Pfeifer, S., Venn, O., Segurel, L., Street, T., Leffler, E. M., Bowden, R., Aneas, I., Broxholme, J., Humburg, P., Iqbal, Z., Lunter, G., Maller, J., Hernandez, R. D., Melton, C., Venkat, A., Nobrega, M. A., Bontrop, R., Myers, S., Donnelly, P., Przeworski, M., and McVean, G. (2012). A Fine-Scale Chimpanzee Genetic Map from Population Sequencing. *Science*, 336(6078):193–198.

Baake, E. and Bialowons, R. (2008). Ancestral processes with selection: Branching and Moran models. *Banach Center Publications*, 80:33–52.

Burden, C. J. and Tang, Y. (2016). An approximate stationary solution for multi-allele neutral diffusion with low mutation rates. *Theoretical Population Biology*, 112:22–32.

Capra, J. A., Hubisz, M. J., Kostka, D., Pollard, K. S., and Siepel, A. (2013). A Model-Based Analysis of GC-Biased Gene Conversion in the Human and Chimpanzee Genomes. *PLoS Genetics*, 9(8):e1003684.

Casillas, S. and Barbadilla, A. (2017). Molecular population genetics. *Genetics*, 205(3):1003–1035.

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(1):285–311.

Durrett, R. (2008). *Probability Models for DNA Sequence Evolution*. Probability and its Applications. Springer New York, New York, NY.

Escobar, J. S., Glémin, S., and Galtier, N. (2011). GC-Biased Gene Conversion Impacts Ribosomal DNA Evolution in Vertebrates, Angiosperms, and Other Eukaryotes. *Molecular Biology and Evolution*, 28(9):2561–2575.

Etheridge, A. M., Griffiths, R. C., and Taylor, J. E. (2010). A coalescent dual process in a Moran model with genic selection, and the lambda coalescent limit. *Theoretical Population Biology*, 78(2):77–92.

Farré, M., Micheletti, D., and Ruiz-Herrera, A. (2013). Recombination Rates and Genomic

Shuffling in Human and ChimpanzeeA New Twist in the Chromosomal Speciation Theory. *Molecular Biology and Evolution*, 30(4):853–864.

Fisher, R. a. (1930). The Genetical Theory of Natural Selection. *Genetics*, 154:272.

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

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(5):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.

Goulet, V., Dutang, C., Maechler, M., Firth, D., Shapira, M., and Stadelmann, M. (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(1):97–109.

Hobolth, A., Dutheil, J. Y., Hawks, J., Schierup, M. H., and Mailund, T. (2011). Incomplete lineage sorting patterns among human, chimpanzee, and orangutan suggest recent orangutan speciation and widespread selection. *Genome Research*, 21(3):349–356.

Jensen-Seaman, M. I. (2004). Comparative Recombination Rates in the Rat, Mouse, and Human Genomes. *Genome Research*, 14(4):528–538.

Katzman, S., Capra, J. A., Haussler, D., and Pollard, K. S. (2011). Ongoing GC-Biased Evolution Is Widespread in the Human Genome and Enriched Near Recombination Hot Spots. *Genome Biology and Evolution*, 3:614–626.

Kimura, M. (1964). Diffusion Models in Population Genetics. *Journal of Applied Probability*, 1(2):177.

Lachance, J. and Tishkoff, S. A. (2014). Biased Gene Conversion Skews Allele Frequencies in Human Populations, Increasing the Disease Burden of Recessive Alleles. *The American Journal of Human Genetics*, 95(4):408–420.

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(3):489–502.

Lesecque, Y., Glémin, S., Lartillot, N., Mouchiroud, D., and Duret, L. (2014). The Red Queen Model of Recombination Hotspots Evolution in the Light of Archaic and Modern Human Genomes. *PLoS Genetics*, 10(11):e1004790.

Lynch, M. (2010). Rate, molecular spectrum, and consequences of human mutation. *Proceedings of the National Academy of Sciences of the United States of America*, 107(3):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.

Marais, G. (2003). Biased gene conversion: Implications for genome and sex evolution. *Trends in Genetics*, 19(6):330–338.

Martin, A. D., Quinn, K. M., and Park, J. H. (2011). MCMCpack: Markov chain monte carlo in R. *Journal of Statistical Software*, 42(9):22.

McCormack, J. E., Faircloth, B. C., Crawford, N. G., Gowaty, P. A., Brumfield, R. T., and Glenn, T. C. (2012). Ultraconserved elements are novel phylogenomic markers that resolve placental mammal phylogeny when combined with species-tree analysis. *Genome Research*, 22(4):746–754.

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

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. and Gillespie, J. (1996). Development of Neutral and Nearly Neutral Theories. *Theoretical population biology*, 49(2):128–42.

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(7):675–682.

Philippe, H., Brinkmann, H., Lavrov, D. V., Littlewood, D. T. J., Manuel, M., Wörheide, G., and Baurain, D. (2011). Resolving Difficult Phylogenetic Questions: Why More Sequences Are Not Enough. *PLoS Biology*, 9(3):e1000602.

Phillips, M. J., Delsuc, F., and Penny, D. (2004). Genome-Scale Phylogeny and the Detection of Systematic Biases. *Molecular Biology and Evolution*, 21(7):1455–1458.

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

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Romiguier, J., Ranwez, V., Delsuc, F., Galtier, N., and Douzery, E. J. (2013). Less Is More in Mammalian Phylogenomics: AT-Rich Genes Minimize Tree Conflicts and Unravel the Root of Placental Mammals. *Molecular Biology and Evolution*, 30(9):2134–2144.

Romiguier, J., Ranwez, V., Douzery, E. J., and Galtier, N. (2010). Contrasting GC-content dynamics across 33 mammalian genomes: Relationship with life-history traits and chromosome sizes. *Genome Research*, 20(8):1001–1009.

Romiguier, J. and Roux, C. (2017). Analytical Biases Associated with GC-Content in Molecular Evolution. *Frontiers in Genetics*, 8:16.

Schrempf, D. and Hobolth, A. (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. *Theoretical Population Biology*, 114:88–94.

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

Serres-Giardi, L., Belkhir, K., David, J., and Glémin, S. (2012). Patterns and Evolution of Nucleotide Landscapes in Seed Plants. *The Plant Cell*, 24(4):1379–1397.

Spencer, C. C. A., Deloukas, P., Hunt, S., Mullikin, J., Myers, S., Silverman, B., Donnelly, P., Bentley, D., and McVean, G. (2006). The Influence of Recombination on Human Genetic Diversity. *PLoS Genetics*, 2(9):e148.

Tavaré, S. (1986). Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on Mathematics in the Life Sciences*, 17(2):57–86.

Tenesa, A., Navarro, P., Hayes, B. J., Duffy, D. L., Clarke, G. M., Goddard, M. E., and Visscher, P. M. (2007). Recent human effective population size estimated from linkage disequilibrium. *Genome Research*, 17(4):520–526.

Vogl, C. and Bergman, J. (2015). Inference of directional selection and mutation parameters assuming equilibrium. *Theoretical Population Biology*, 106:71–82.

Webster, M. T., Axelsson, E., and Ellegren, H. (2006). Strong Regional Biases in Nucleotide Substitution in the Chicken Genome. *Molecular Biology and Evolution*, 23(6):1203–1216.

Wright, S. (1931). Evolution in Mendelian Populations. *Genetics*, 16(2):97–159.