Supplementary material

Quantifying GC-biased gene conversion in great ape genomes using polymorphism-aware models
Rui Borges¹, Gergely Szöllősi², and Carolin Kosiol¹,³*

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
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 population 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

$$\begin{cases} p(Na_i) = P(Ma_i) \\ \sum_{n=1}^{N-1} p(na_i(N-n)a_i) = \sum_{m=1}^{M-1} p'(ma_i(M-m)a_i) \end{cases}$$ \hspace{1cm} (S1)$$

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

$$\begin{cases} \pi a_i \frac{(1 + \sigma a_i)N-1}{N} = \pi' a_i \frac{(1 + \sigma' a_i)M-1}{M} \\ \frac{\pi a_i \pi a_j \frac{\rho a_i a_j}{N}}{\sum_{n=1}^{N-1} (1 + \sigma a_i)N-1(1 + \sigma a_j)N-n-1} = \frac{\pi' a_i \pi' a_j \frac{\rho a_i a_j}{M}}{\sum_{m=1}^{M-1} (1 + \sigma' a_i)M-1(1 + \sigma' a_j)M-m-1} \end{cases}$$ \hspace{1cm} (S2)$$

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 to conclude that the normalization constants $k$ and $k'$ are equal. Simplifying,

$$\begin{cases} \pi a_i = \pi' a_i \\ (1 + \sigma a_i)N-1 = (1 + \sigma' a_i)M-1 \\ \frac{\rho a_i a_j \frac{\rho a_i a_j}{N}}{\sum_{n=1}^{N-1} (1 + \sigma a_i)N-1(1 + \sigma a_j)N-n-1} = \frac{\rho a_i a_j \frac{\rho a_i a_j}{M}}{\sum_{m=1}^{M-1} (1 + \sigma' a_i)M-1(1 + \sigma' a_j)M-m-1} \end{cases}$$ \hspace{1cm} (S3)$$

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

$$\begin{cases} \pi' a_i = \pi a_i \\ (1 + \sigma' a_i) = (1 + \sigma a_i)N-1 \frac{\sum_{m=1}^{M-1} (1 + \sigma a_i)M-1(1 + \sigma a_j)M-m-1}{\sum_{n=1}^{N-1} (1 + \sigma a_i)N-1(1 + \sigma a_j)N-n-1} \end{cases}$$ \hspace{1cm} (S4)$$

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 A1 depicts the effect of the effective population size on the mutation rates and selection coefficients.

![Figure A1](image_url)

Figure A1. 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 a B. GC-bias case (black dots: mutations rates equal to the neutral scenario and $\sigma_W = 0, \sigma_s = 0.01$).
Appendix B: Proof of the stationary vector

Let \( \psi \) be a stationary vector of \( Q \) with \( \psi_{a,j}^n \) and \( \psi_{a,j}^1 \) 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 only occurring 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 \( a_j \) is either fixed (\( n = N \)) or absent (\( n = 0 \), i.e. \( a_j \) is fixed), for which we may write

\[
\psi_{a,a_j}^N = \psi_{a,a_j}^0 = \psi_{a,a_j}^1, 
\]

while between the polymorphic states, the general condition is valid

\[
\psi_{a,a_j}^{n \rightarrow n+1} = \psi_{a,a_j}^{n+1 \rightarrow n + 1}, 
\]

Condition (S6) can be rewritten in the recursive form

\[
\psi_{a,a_j}^{n+1} = \psi_{a,a_j}^n \frac{\psi_{a,a_j}^{n \rightarrow n+1}}{\psi_{a,a_j}^{n+1 \rightarrow n}}, 
\]

and then combined with equation (S5)

\[
\psi_{a,a_j}^{N \rightarrow N-1} = \psi_{a,a_j}^N \frac{\psi_{a,a_j}^{N-1 \rightarrow N-1}}{\psi_{a,a_j}^{N \rightarrow N-1}} \cdots \frac{\psi_{a,a_j}^{n \rightarrow n+1}}{\psi_{a,a_j}^{n+1 \rightarrow n}} \psi_{a,a_j}^{N-1 \rightarrow N} = \psi_{a,a_j}^N \prod_{r=n+1}^{N-1} \frac{\psi_{a,a_j}^{r \rightarrow r+1}}{\psi_{a,a_j}^{r+1 \rightarrow r}}. 
\]

The product can be further simplified by recognizing that for \( r = N - 1 \), \( \psi_{a,a_j}^{N-1 \rightarrow N-1} = \mu_{a,a_j} = \pi_a \rho_{a,a_j} \), while for \( r < N - 1 \), the rates inside the product are just the combined rate of drift and selection (according to the rate matrix in (2)). We can now rewrite equation (S7) in order to the \( \psi_{a,a_j}^n \) element of the stationary vector of \( Q \)

\[
\psi_{a,a_j}^n = \frac{\psi_{a,a_j}^1 \pi_a \rho_{a,a_j}}{1 + \sigma_{a_j}} \left( \frac{1 + \sigma_{a_j}}{1 + \sigma_{a_j}} \right)^{N-n-1}. 
\]

Because \( \psi_{a,a_j}^0 = \psi_{a} \) and \( \psi_{a,a_j}^0 = \mu_{a} = \pi_a \rho_{a,a_j} \), we obtain a possible solution for the monomorphic states of the stationary distribution by making \( n = 0 \) in equation (S9)

\[
\psi_a = \frac{\pi_a \sigma_{a_j}}{1 + \sigma_{a_j}}. 
\]

The stationary solution for the polymorphic states can be obtained from equation (S9) by noting that \( \psi_{a} = \pi_a \sigma_{a_j}^{-1} \) and \( \psi_{a,a_j}^{n \rightarrow n+1} = \frac{n(N-n)}{N} (1 + \sigma_{a_j}) \)

\[
\psi_{a,a_j}^n = \pi_a \pi_a \rho_{a,a_j} (1 + \sigma_{a_j})^{-1} (1 + \sigma_{a_j})^{N-n-1} \frac{N}{n(N-n)} \cdot 
\]

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

\[
\psi_a = \pi_a, \quad \psi_{a,a_j} = \pi_a \pi_a \rho_{a,a_j} \frac{N}{n(N-n)}. 
\]
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

\[ d_{MS}(t = 1) = -\sum_{u} \psi_{u} q_{uu}, \quad (S13) \]

Where \( \psi \) is the stationary vector and \( q_{uu} \) the diagonal elements of \( Q \). Equation (S13) can be solved by observing that a monomorphic state can only be escaped by mutation, while a polymorphic state can only be escaped by selection and drift

\[ d_{MS} = \sum_{a_i \in A} \sum_{a_j \in \mathcal{A}^c} \psi_{a_i} \mu_{a_i a_j} + \sum_{a_i a_j \in \mathcal{A}^c} \sum_{n=1}^{N-1} \psi_{a_i a_j}^n \frac{n(N-n)}{N} (1 + \sigma_{a_i} + 1 + \sigma_{a_j}) \quad (S14) \]

The stationary vector is known from equations (S10) and (S11)

\[ d_{MS} = \frac{1}{k} \sum_{a_i \in A} \sum_{j \neq i} (1 + \sigma_{a_i})^{N-1} \tau_{a_i a_j} \rho_{a_i a_j} \tau_{a_j} + \frac{1}{k} \sum_{a_i a_j \in \mathcal{A}^c} \sum_{n=1}^{N-1} \tau_{a_i a_j} \rho_{a_i a_j} \tau_{a_j} [(1 + \sigma_{a_i})^n (1 + \sigma_{a_i})^{N-n-1} + (1 + \sigma_{a_i})^{n-1}(1 + \sigma_{a_i})^{N-n}] \quad (S15) \]

where \( k \) is the normalization constant defined in equation (4). Expression (S15) can be further simplified by observing that the sum in \( n \) results in doubling every \( (1 + \sigma_{a_i})^{n-1}(1 + \sigma_{a_i})^{N-n} \) element. Therefore, the expected number of events can be simplified to

\[ d_{MS} = \frac{2}{k} \sum_{a_i a_j \in \mathcal{A}^c} \sum_{n=1}^{N} \tau_{a_i a_j} \rho_{a_i a_j} \tau_{a_i} (1 + \sigma_{a_i})^{n-1}(1 + \sigma_{a_i})^{N-n} \quad . \quad (S16) \]
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 $d'_{MS}$.

$$d'_{MS} = d_{MS} \times s \quad ,$$

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

$$s = \sum_{a_i \rightarrow a_j} s_{a_i \rightarrow a_j} = \sum_{a_i} m_{a_i \rightarrow a_j} \times h_{a_i} \quad ,$$

where $A^p$ represents all the possible pair-wise permutations without repetition of $K$ alleles.

1. **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 (S16). The rate of a $a_i \rightarrow a_j$ mutation is the rate of escaping the monomorphic state $\{N_{a_i}\}$, from which we can write

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

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

2. **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

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

As we are using the multivariate Moran model, we have to extend the denominator of (S20) to account for the different possible combinations of two selection coefficients. In particular, we may have

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

We further redefine the denominators in order to make them equal

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

3. **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

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

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 (S18)

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

The relationship between the Moran distance in events and substitutions can be defined based on equation (S17),

$$d_{MS}^* = d_{MS} \frac{1}{N^2} \quad .$$

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):

$$d_{MS}^* = \frac{1}{N^2} \quad .$$
Literature Cited
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 Research 25: 1215–1228.
Jensen-Seaman, M. I., 2004 Comparative Recombination Rates in the Rat, Mouse, and Human Genomes. Genome Research 14: 528–538.
Kluth, S. and E. Baake, 2013 The Moran model with selection: Fixation probabilities, ancestral lines, and an alternative particle representation. Theoretical Population Biology 90: 104–112.
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. Journal of Theoretical Biology 407: 362–370.
Figure S1 Numerical validation of the stationarity vector. Estimated vectors of $\pi$, $\rho$, $\sigma$ from the great apes’ data were used to calculate the rate matrix $Q$ and the probabilities for the state space at several time points (time in generations). The initial probabilities were set uniformly as $\frac{1}{4+6(N-1)}$, i.e. the number of states. For sake of clarity only the monomorphic states $\{i\}$ are represented.
Figure S2 Validation of the Bayesian algorithms. Trace plot depicting the convergence of the MCMC runs (grey dots and blue lines) to the true parameter values (red lines). Simulation conditions: 1,000,000 sites, 10 individuals and a simple parameter vector for the Moran model with boundary mutations: $\pi = (0.25, 0.25, 0.25, 0.25)$, $\rho = (0.001, 0.001, 0.001, 0.001, 0.001)$. The blue line represents the MCMC moving average whereas the red one represents the true values. The codes used to performed these simulations is available in GitHub: https://github.com/pomo-dev/pomo_selection.
**Figure S3 Validation of the Bayesian algorithms.** Trace plot depicting the convergence of the MCMC runs (grey dots and blue lines) to the true parameter values (red lines). Simulation conditions: 100 000 sites, 10 individuals and a complex parameter vector for the Moran model with boundary mutations: \( \pi = (0.10, 0.20, 0.30, 0.40) \), \( \rho = (0.0003, 0.0006, 0.0009, 0.0011, 0.0014, 0.0017) \). The blue line represents the MCMC moving average whereas the red one represents the true values. The codes used to performed these simulations is available in GitHub: [https://github.com/pomo-dev/pomo_selection](https://github.com/pomo-dev/pomo_selection).
Figure S4 Validation of the Bayesian algorithms. Trace plot depicting the convergence of the MCMC runs (grey dots and blue lines) to the true parameter values (red lines). Simulation conditions: 1000000 sites, 10 individuals and a simple parameter vector for the Moran model with allelic selection: $\pi = (0.25, 0.25, 0.25, 0.25)$, $\rho = (0.001, 0.001, 0.001, 0.001, 0.001, 0.001)$, $\sigma = (1.00, 1.00, 1.00, 1.00)$. The blue line represents the MCMC moving average whereas the red one represents the true values. The codes used to performed these simulations is available in GitHub: https://github.com/pomo-dev/pomo_selection.
Figure S5 Validation of the Bayesian algorithms. Trace plot depicting the convergence of the MCMC runs (grey dots and blue lines) to the true parameter values (red lines). Simulation conditions: 1000000 sites, 10 individuals and a complex parameter vector for the Moran model with allelic selection: \( \pi = (0.10, 0.20, 0.30, 0.40) \), \( \rho = (0.0003, 0.0006, 0.0009, 0.0011, 0.0014, 0.0017) \), \( \sigma = (1.00, 1.01, 1.02, 1.03) \). The blue line represents the MCMC moving average whereas the red one represents the true values. The codes used to performed these simulations is available in GitHub: https://github.com/pomo-dev/pomo_selection.
Figure S6 Prediction of the site-frequency spectrum in great ape populations. The gray points represent the observed counts and the vertical lines the posterior predictive distribution of the stationary distribution under the 4-variate Moran model: boundary mutation model (blue) and allelic selection model (red).
C. Eastern gorillas

D. Western gorillas
E. Western chimpanzees

F. Nigeria-Cameroon chimpanzees
G. Eastern chimpanzees

H. Central chimpanzees
K. Sumatran orangutans
Figure S7 GC-bias vs. recombination rate and chromosome length in non-African humans. The scaled selection coefficients were estimated based on the posterior average. We scaled or selection coefficients according to equation (S4). Recombination rates were estimated by comparing the genetic distance (cM) between markers to the physical (Mb) as described in (Jensen-Seaman (2004)) and based on the human Iceland pedigree map.

A. $\sigma$ versus recombination rate

B. $\sigma$ versus chromosome length
Figure S8 Comparison of gBGC estimates between Glémin et al. (2015) and our method. Estimates of gBGC rate coefficients were obtained using Glémin et al. (2015) and the method proposed here (\( B \) in and \( \sigma_S \), respectively). We used a human data set available as supplementary material in Glémin et al. (2015). The alleles counts correspond to 51 regions of 1 million sites from the human chromosome 1. We adapted our 4-variate model to only account for these two types of alleles: \( S \) and \( W \) stand for strong and weak alleles, respectively. The code needed to reproduce these analyses can be found in https://github.com/pomo-dev/pomo_selection. The correlation tests were FDR corrected for multiple testing.
### Table S1: Summary of great apes’ count data.

| Population                      | Monomorphic | Polymorphic |
|---------------------------------|-------------|-------------|
|                                 | A           | C           | G           | T           | AC          | AG          | AT          | CG          | CT          | GT          |
| African humans                  | 615878      | 845533      | 666393      | 694569      | 249         | 1972        | 112         | 395         | 1716        | 318         |
| Non-african humans              | 615966      | 845701      | 666455      | 694604      | 245         | 1780        | 101         | 321         | 1508        | 275         |
| Eastern gorillas                | 615513      | 844903      | 666554      | 694492      | 126         | 977         | 60          | 192         | 874         | 139         |
| Western gorillas                | 612727      | 839959      | 661019      | 691995      | 389         | 3238        | 180         | 506         | 2707        | 372         |
| Bonobos                         | 615217      | 844431      | 665216      | 693948      | 280         | 2422        | 134         | 408         | 1904        | 280         |
| Nigeria-Cameroon chimpanzees    | 614656      | 843715      | 664742      | 693631      | 484         | 3888        | 227         | 632         | 3294        | 470         |
| Eastern chimpanzees             | 614403      | 843458      | 664327      | 693337      | 377         | 3317        | 203         | 552         | 2593        | 409         |
| Central chimpanzees             | 614292      | 843310      | 664183      | 693270      | 440         | 3369        | 214         | 521         | 2626        | 460         |
| Western chimpanzees             | 615125      | 844759      | 665810      | 693893      | 246         | 1876        | 107         | 335         | 1499        | 261         |
| Sumatran orangutans             | 615164      | 842548      | 663805      | 694268      | 414         | 3980        | 181         | 598         | 3192        | 468         |
| Bornean orangutans              | 615573      | 843480      | 664912      | 694666      | 333         | 2655        | 135         | 420         | 2274        | 320         |
**Table S2 Simulation schemes.** Simulation schemes used to validate the Bayesian algorithms for estimating the model parameters under the multivariate Moran model with mutation (M schemes) and mutation plus selection (S schemes). $\sigma_A$ is set to 1.

| Scheme | $\pi_A$ | $\pi_C$ | $\pi_G$ | $\pi_T$ | $\rho_{AC}$ | $\rho_{AG}$ | $\rho_{AT}$ | $\rho_{CG}$ | $\rho_{CT}$ | $\rho_{GT}$ | $\sigma_A$ | $\sigma_C$ | $\sigma_G$ | $\sigma_T$ |
|--------|--------|--------|--------|--------|-------------|-------------|-------------|-------------|-------------|-------------|--------|--------|--------|--------|
| M1     | 0.25   | 0.25   | 0.25   | 0.25   | 0.001      | 0.001       | 0.001       | 0.001       | 0.001       | -           | -      | -      | -      | -      |
| M2     | 0.22   | 0.30   | 0.23   | 0.25   | 0.00028    | 0.00300     | 0.00016     | 0.00036     | 0.00172     | 0.00033     | -      | -      | -      | -      |
| S1     | 0.25   | 0.25   | 0.25   | 0.25   | 0.001      | 0.001       | 0.001       | 0.001       | 0.001       | 1           | 1      | 1      | 1      | 1      |
| S2     | 0.22   | 0.30   | 0.23   | 0.25   | 0.00028    | 0.00300     | 0.00016     | 0.00036     | 0.00172     | 0.00033     | 1      | 1.030  | 1.024  | 1.004  |
Table S3 Great apes mutation rates and selection coefficients. Scaled mutation rates and selection coefficients estimated for the great apes populations using the multivariate Moran model with boundary mutations and allelic selection.

| Population                        | $\mu_{AC}$ | $\mu_{CA}$ | $\mu_{AG}$ | $\mu_{GA}$ |
|-----------------------------------|------------|------------|------------|------------|
| African humans                    | 0.000237   | 0.000934   | 0.002248   | 0.008030   |
| Non-african humans                | 0.00464    | 0.00957    | 0.03411    | 0.008700   |
| Eastern gorillas                  | 0.000220   | 0.000257   | 0.001850   | 0.002280   |
| Western gorillas                  | 0.001383   | 0.005259   | 0.014739   | 0.049773   |
| Bonobos                           | 0.000617   | 0.002196   | 0.005840   | 0.022977   |
| Nigeria-Cameroon chimpanzees     | 0.000896   | 0.003185   | 0.008399   | 0.029962   |
| Eastern chimpanzees              | 0.000516   | 0.001829   | 0.005393   | 0.018297   |
| Central chimpanzees              | 0.000391   | 0.002039   | 0.003698   | 0.017267   |
| Western chimpanzees              | 0.000391   | 0.000913   | 0.002932   | 0.008944   |
| Sumatran orangutans               | 0.000573   | 0.001699   | 0.006940   | 0.017486   |
| Bornean orangutans                | 0.000604   | 0.001116   | 0.005353   | 0.010287   |

| Population                        | $\mu_{AT}$ | $\mu_{TA}$ | $\mu_{CG}$ | $\mu_{GC}$ |
|-----------------------------------|------------|------------|------------|------------|
| African humans                    | 0.000224   | 0.000233   | 0.000982   | 0.000888   |
| Non-african humans                | 0.000318   | 0.000296   | 0.000838   | 0.001036   |
| Eastern gorillas                  | 0.000114   | 0.000134   | 0.000352   | 0.000373   |
| Western gorillas                  | 0.001508   | 0.001765   | 0.004350   | 0.003859   |
| Bonobos                           | 0.000761   | 0.000639   | 0.001869   | 0.002065   |
| Nigeria-Cameroon chimpanzees     | 0.000998   | 0.000964   | 0.002558   | 0.002566   |
| Eastern chimpanzees              | 0.000594   | 0.000655   | 0.001700   | 0.001626   |
| Central chimpanzees              | 0.000511   | 0.000514   | 0.001457   | 0.001304   |
| Western chimpanzees              | 0.000289   | 0.000293   | 0.000788   | 0.001028   |
| Sumatran orangutans               | 0.000475   | 0.000515   | 0.001739   | 0.001477   |
| Bornean orangutans                | 0.000359   | 0.000378   | 0.001065   | 0.001108   |

| Population                        | $\mu_{CT}$ | $\mu_{TC}$ | $\mu_{GT}$ | $\mu_{TG}$ |
|-----------------------------------|------------|------------|------------|------------|
| African humans                    | 0.006177   | 0.001625   | 0.001235   | 0.000360   |
| Non-african humans                | 0.005777   | 0.002607   | 0.001324   | 0.000484   |
| Eastern gorillas                  | 0.001609   | 0.001619   | 0.000291   | 0.000277   |
| Western gorillas                  | 0.033763   | 0.010382   | 0.005220   | 0.001810   |
| Bonobos                           | 0.015136   | 0.003569   | 0.002698   | 0.000576   |
| Nigeria-Cameroon chimpanzees     | 0.021183   | 0.005752   | 0.003521   | 0.000954   |
| Eastern chimpanzees              | 0.011799   | 0.003667   | 0.002102   | 0.000683   |
| Central chimpanzees              | 0.011798   | 0.002272   | 0.002279   | 0.000491   |
| Western chimpanzees              | 0.005329   | 0.002309   | 0.001194   | 0.000397   |
| Sumatran orangutans               | 0.012329   | 0.004508   | 0.001913   | 0.000825   |
| Bornean orangutans                | 0.007159   | 0.004084   | 0.001160   | 0.000636   |

| Population                        | $1 + \sigma_C$ | $1 + \sigma_G$ | $1 + \sigma_T$ |
|-----------------------------------|----------------|----------------|----------------|
| African humans                    | 1.403          | 1.311          | 1.032          |
| Non-african humans                | 1.099          | 1.097          | 1.004          |
| Eastern gorillas                  | 1.099          | 1.060          | 1.058          |
| Western gorillas                  | 1.032          | 1.025          | 1.005          |
| Bonobos                           | 1.066          | 1.060          | 0.998          |
| Nigeria-Cameroon chimpanzees     | 1.087          | 1.074          | 1.005          |
| Eastern chimpanzees              | 1.155          | 1.125          | 1.020          |
| Central chimpanzees              | 1.325          | 1.260          | 1.018          |
| Western chimpanzees              | 1.139          | 1.142          | 1.015          |
| Sumatran orangutans               | 1.169          | 1.118          | 1.023          |
| Bornean orangutans                | 1.109          | 1.085          | 1.019          |