Supplementary material

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

Supplementary File S1. Proof of the stationary vector

Let $\boldsymbol{\psi}$ be a stationary vector of $\boldsymbol{q}$ and ψ_{ij}^{n} be the element of the stationary vector corresponding to the state $\{ni, (N-n)j\}$. In the multivariate Moran model with low mutation rates and selection, mutation is only acting 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 lead to equations for the monomorphic and the polymorphic states. In the boundary states, an allele i is either fixed ($n = N$) or absent ($n = 0$, i.e. j is fixed):

$$\psi_i q_{ij}^{N\to N-1} = \psi_{ij}^{N-1} q_{ij}^{N-1\to N} \qquad \psi_j q_{ij}^{0\to 1} = \psi_{ij}^{1} q_{ij}^{1\to 0} \tag{8}$$

Between the polymorphic states, the general condition is valid:

$$\psi_{ij}^{n} q_{ij}^{n\to n+1} = \psi_{ij}^{n+1} q_{ij}^{n+1\to n} \tag{9}$$

Condition (9) can be rewritten in the recursive form:

$$\psi_{ij}^{n+1} = \psi_{ij}^{n} \frac{q_{ij}^{n\to n+1}}{q_{ij}^{n+1\to n}} \tag{10}$$

Equations (8) and (10) can be combined:

$$\psi_i q_{ij}^{N\to N-1} = \psi_{ij}^{n} \frac{q_{ij}^{n\to n+1}}{q_{ij}^{n+1\to n}} \cdots \frac{q_{ij}^{N-2\to N-1}}{q_{ij}^{N-1\to N-2}} q_{ij}^{N-1\to N} = \psi_{ij}^{n} q_{ij}^{n\to n+1} \prod_{r=n+1}^{N-1} \frac{q_{ij}^{r\to r+1}}{q_{ij}^{r\to r-1}} \tag{11}$$

Recognizing that $q_{ij}^{N\to N-1} = \mu_{ij} = \pi_j \rho_{ij}$ and that all the rates inside the product follow expression 2, we can further simplify equation (10) in order to the ψ_{ij}^{n} element of the stationary vector $\boldsymbol{q}$.

$$\psi_{ij}^{n} = \frac{\psi_i \pi_j \rho_{ij}}{q_{ij}^{n\to n+1}} \left(\frac{1+\sigma_j}{1+\sigma_i}\right)^{N-n-1} \tag{12}$$

We can now use equation (12) and make $n = 0$. Because $\psi_{ij}^{0} = \psi_j$ and $q_{ij}^{0\to 1} = \mu_{ji} = \pi_i \rho_{ij}$, we obtain a possible solution for the monomorphic states of the stationary distribution:

$$\frac{\psi_j}{\psi_i} = \frac{\pi_j}{\pi_i} \left(\frac{1+\sigma_j}{1+\sigma_i}\right)^{N-1} \tag{13}$$

And for the polymorphic states, we use equation (12):

$$\psi_{ij}^{n} = \pi_i \pi_j \rho_{ij} (1+\sigma_i)^{n-1} (1+\sigma_j)^{N-n-1} \frac{N}{n(N-n)} \tag{14}$$

The stationary distribution obtained here can be used to obtain the stationary vector of the boundary multivariate Moran model in neutrality. We observe that when $\boldsymbol{\sigma} = \mathbf{0}$, we obtain the solution obtained by Schrempf et al. (2016) for the multivariate Moran model with drift only.

$$\psi_i = \pi_i \qquad \psi_{ij}^{n} = \pi_i \pi_j \rho_{ij} \frac{N}{n(N-n)} \tag{15}$$

References:

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

Supplementary File S2. Proof of the expected number of Moran events per unit of time

In order to have an impression of the consequences 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:

$$d_S(t=1) = -\sum_i \psi_i q_{ii} \tag{16}$$

Where $\boldsymbol{\psi}$ is the stationary vector and q_{ii} the diagonal elements of $\boldsymbol{q}$ (expression (2)). Equation 16 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.

$$d_S(1) = \sum_{i \in \mathcal{A}} \sum_{j \neq i} \psi_i \mu_{ij} + \sum_{ij \in \mathcal{A}^C} \sum_{n=1}^{N-1} \psi_{ij}^n \frac{n(N-n)}{N}(1+\sigma_i+1+\sigma_j) \tag{17}$$

The stationary vector is known from equations (13) and (14). k is the normalization constant defined in 4.

$$d_S(1) = \frac{1}{k} \sum_{i \in \mathcal{A}} \sum_{j \neq i} (1+\sigma_i)^{N-1} \pi_i \rho_{ij} \pi_j + \frac{1}{k} \sum_{ij \in \mathcal{A}^C} \sum_{n=1}^{N-1} \pi_i \rho_{ij} \pi_j [(1+\sigma_i)^n (1+\sigma_j)^{N-n-1} + (1+\sigma_i)^{n-1} (1+\sigma_j)^{N-n}] \tag{18}$$

The expression can be further simplified by observing that the sum in $n = 1, ..., N$ results in doubling every $(1+\sigma_i)^{n-1}(1+\sigma_j)^{N-n}$ element. Therefore, the expected number of events can be simplified to:

$$d_S(1) = \frac{2}{k} \sum_{ij \in \mathcal{A}^C} \sum_{n=1}^{N} \pi_i \rho_{ij} \pi_j (1+\sigma_i)^{n-1} (1+\sigma_j)^{N-n} \tag{19}$$

Supplementary File S3. Proof of the Moran distance in number substitutions

The Moran distance $d_S(t)$ accounts for several events (mutations, drift and selection) and differs from the standard distances $d_S^*(t)$, calculated in terms of the expected number of substitutions. A way to compare them is correcting the Moran distance so it only accounts for substitutions, which can be done by computing the probability of a substitution s.

$$d_S^*(t) = d_S(t)s \tag{20}$$

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.

$$s = \sum_{ij \in \mathcal{A}^P} s_{i \to j} = \sum_{ij} m_{i \to j} \times h_{j|i} \tag{21}$$

$\mathcal{A}^P$ represents all the possible pair-wise permutations without repetition of the K alleles.

1. Solving $m_{i \to j}$

The probability of an event being a mutation is simply the ratio between the rate of mutations and the total rate (i.e the rate of mutations plus the rate of drift and selection). In stationarity, the total rate $r_T = d_S(1)$, the expected number of events of the Moran model defined in equation (19). The rate of a $i \to j$ mutation is the rate of escaping the monomorphic state $\{i\}$.

$$m_{i \to j} = \frac{r_{i \to j}}{r_T} = \frac{\pi_i \pi_j \rho_{ij} (1+\sigma_i)^{N-1}}{2 \sum_{ij \in \mathcal{A}^C} \sum_{n=1}^{N} \pi_i \pi_j \rho_{ij} (1+\sigma_i)^{n-1} (1+\sigma_j)^{N-n}} \tag{22}$$

We can see that $m_{i \to j}$ differs from $m_{j \to i}$ due to the selection coefficient in the numerator.

2. Solving $h_{i|j}$

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

$$h = \frac{(1+\sigma)^{N-1}}{\sum_{n=0}^{N-1} (1+\sigma)^n} \tag{23}$$

We use the multivariate Moran model, and so we have to extend the denominator of (23) to account for the different possible combinations of two selection coefficients.

$$h_{i|j} = \frac{(1+\sigma_i)^N}{\sum_{n=1}^{N} (1+\sigma_i)^n (1+\sigma_j)^{N-n}} \qquad h_{j|i} = \frac{(1+\sigma_j)^N}{\sum_{n=1}^{N} (1+\sigma_j)^n (1+\sigma_i)^{N-n}} \tag{24}$$

We redefine the denominators in order to make them equal.

$$h_{i|j} = \frac{(1+\sigma_i)^N (1+\sigma_j)}{\sum_{n=1}^{N} (1+\sigma_i)^n (1+\sigma_j)^{N-n+1}} \qquad h_{j|i} = \frac{(1+\sigma_j)^N (1+\sigma_i)}{\sum_{n=1}^{N} (1+\sigma_j)^n (1+\sigma_i)^{N-n+1}} \tag{25}$$

3. Solving s

The probability of a $i \to j$ substitution under the multivariate Moran model with boundary mutations and selection can be computed as:

$$s_{i\to j} = m_{i\to j} \times h_{j|i} = \frac{\pi_i \pi_j \rho_{ij} (1+\sigma_i)^N (1+\sigma_j)^N}{2 \times \sum_{ij\in\mathcal{A}^C} \sum_{n=1}^{N} \pi_i \pi_j \rho_{ij} (1+\sigma_i)^{n-1} (1+\sigma_j)^{N-n} \times \sum_{n=1}^{N} (1+\sigma_j)^n (1+\sigma_i)^{N-n+1}} \tag{26}$$

We can see that $s_{i\to j} = s_{j\to i}$. The probability of an event being an substitution is defined in (21).

$$s = \frac{1}{\sum_{ij\in\mathcal{A}^C} \sum_{n=1}^{N} \pi_i \pi_j \rho_{ij} (1+\sigma_i)^{n-1} (1+\sigma_j)^{N-n}} \sum_{ij\in\mathcal{A}^C} \frac{\pi_i \pi_j \rho_{ij} (1+\sigma_i)^N (1+\sigma_j)^N}{\sum_{n=1}^{N} (1+\sigma_j)^n (1+\sigma_i)^{N-n+1}} \tag{27}$$

The relationship between the Moran distance in events and substitutions can be defined based on equation (20).

$$d_S^*(t) = d_S(t) \frac{1}{\sum_{ij\in\mathcal{A}^C} \sum_{n=1}^{N} \pi_i \pi_j \rho_{ij} (1+\sigma_i)^{n-1} (1+\sigma_j)^{N-n}} \sum_{ij\in\mathcal{A}^C} \frac{\pi_i \pi_j \rho_{ij} (1+\sigma_i)^N (1+\sigma_j)^N}{\sum_{n=1}^{N} (1+\sigma_j)^n (1+\sigma_i)^{N-n+1}} \tag{28}$$

This quantity can be evaluated for neutral regimes: i.e. $\boldsymbol{\sigma} \to (1,1,1,1)$. We obtain the probability of a substitutions under the neutral Moran model and it matches the results of Schrempf et al. (2016):

$$d_S^*(t) = d_S(t) \frac{1}{N^2} \tag{29}$$

References:

Kluth, S., & Baake, E. (2013). The Moran model with selection: Fixation probabilities, ancestral lines, and an alternative particle representation. Theoretical Population Biology, 90, 104-112.

Schrempf, D., Minh, B. Q., De Maio, N., von Haeseler, A., & Kosiol, C. (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 $\boldsymbol{\pi}$, $\boldsymbol{\rho}$, $\boldsymbol{\sigma}$ from the great apes' data were used to calculate the rate matrix $\boldsymbol{Q}$ 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.

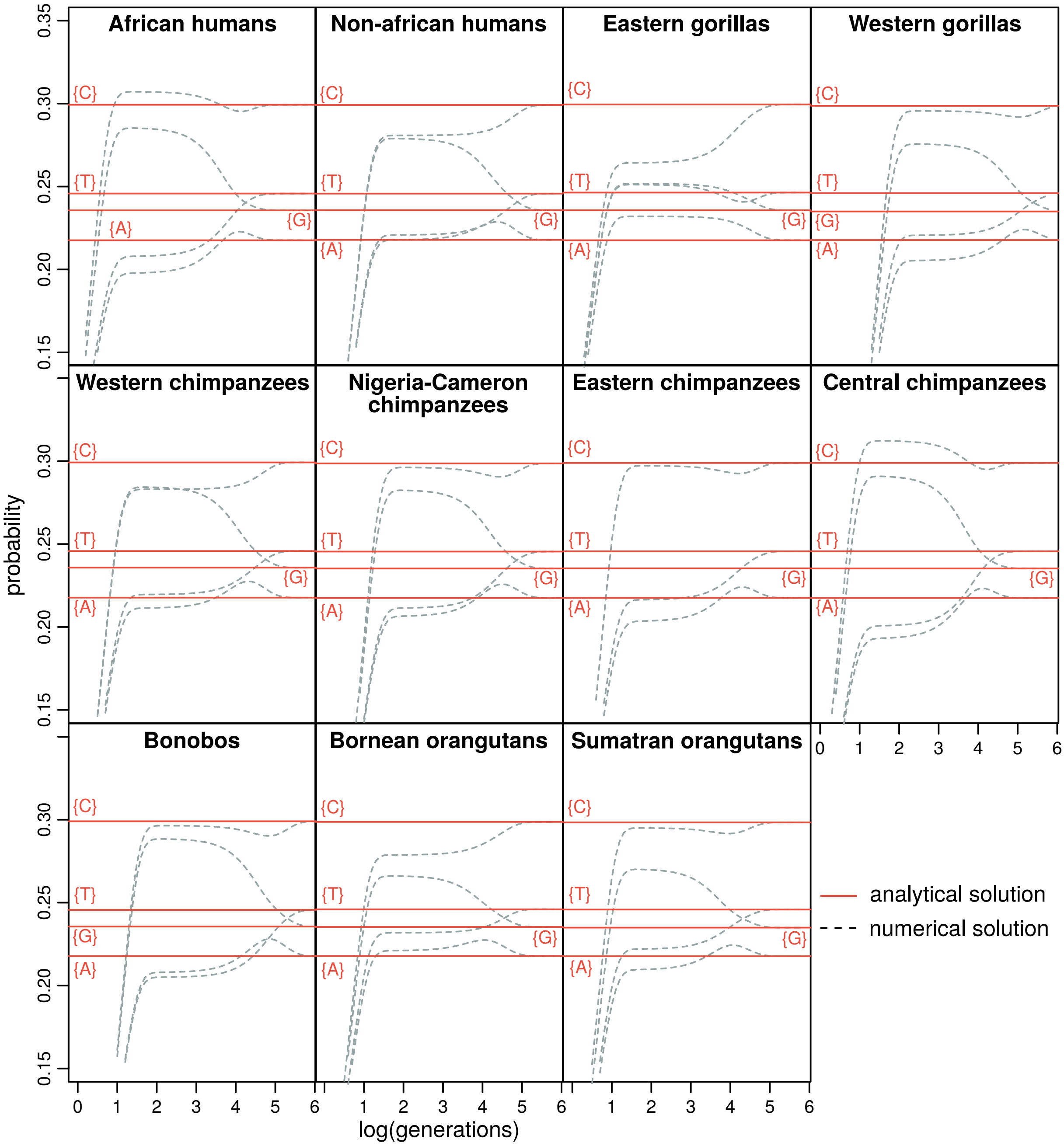


Figure S2: **Validation of the Bayesian algorithms.** Simulation conditions: 1000000 sites, 10 individuals and a simple parameter vector for the Moran model with boundary mutations: $\boldsymbol{\pi} = (0.25, 0.25, 0.25, 0.25)$, $\boldsymbol{\rho} = (0.001, 0.001, 0.001, 0.001, 0.001, 0.001)$. The blue line represents the MCMC moving average whereas the red one represents the true values.

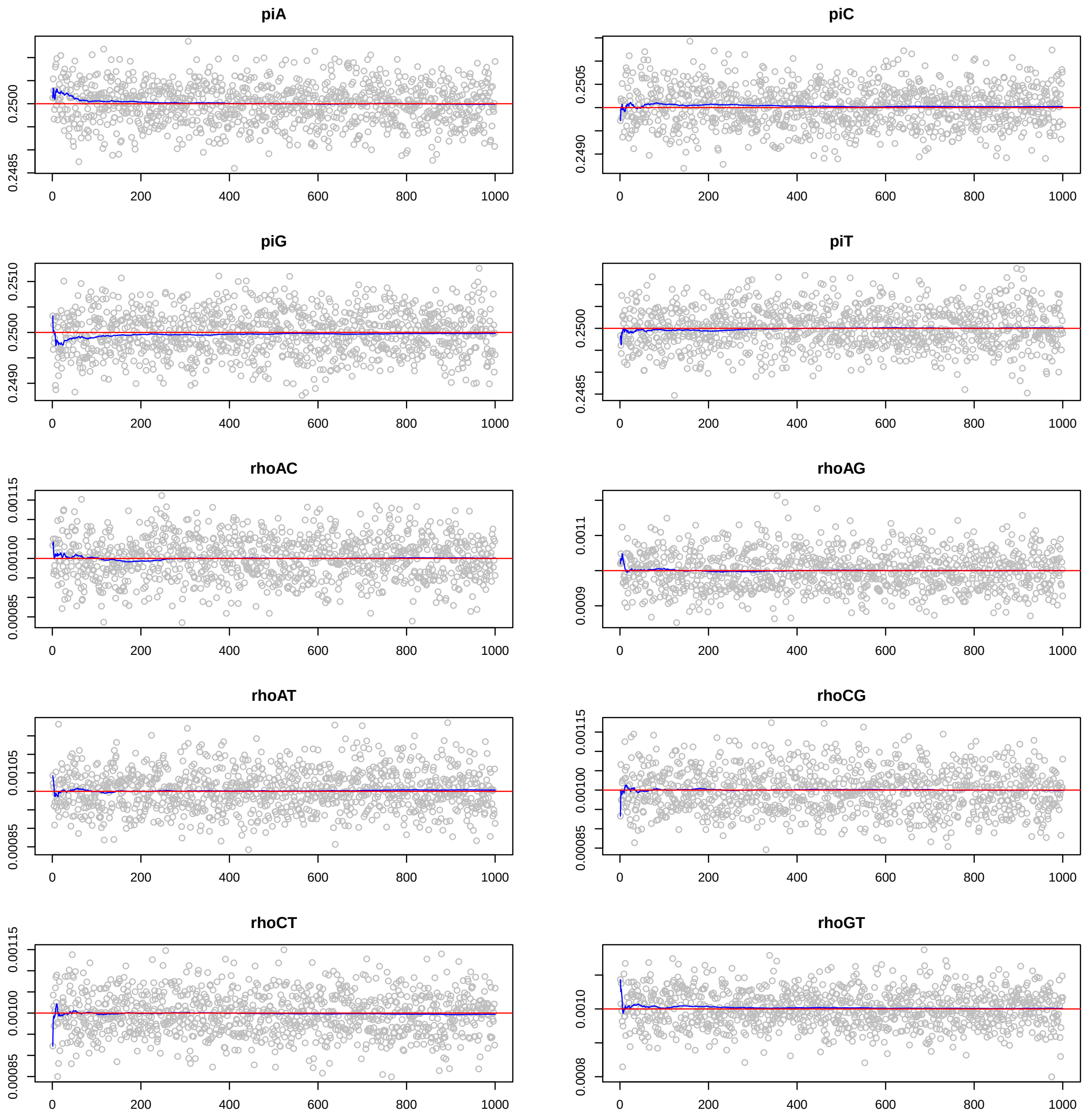


Figure S3: **Validation of the Bayesian algorithms.** Simulation conditions: 1000000 sites, 10 individuals and a complex parameter vector for the Moran model with boundary mutations: $\boldsymbol{\pi} = (0.10, 0.20, 0.30, 0.40)$, $\boldsymbol{\rho} = (0.003, 0.006, 0.009, 0.011, 0.014, 0.017)$. The blue line represents the MCMC moving average whereas the red one represents the true values.

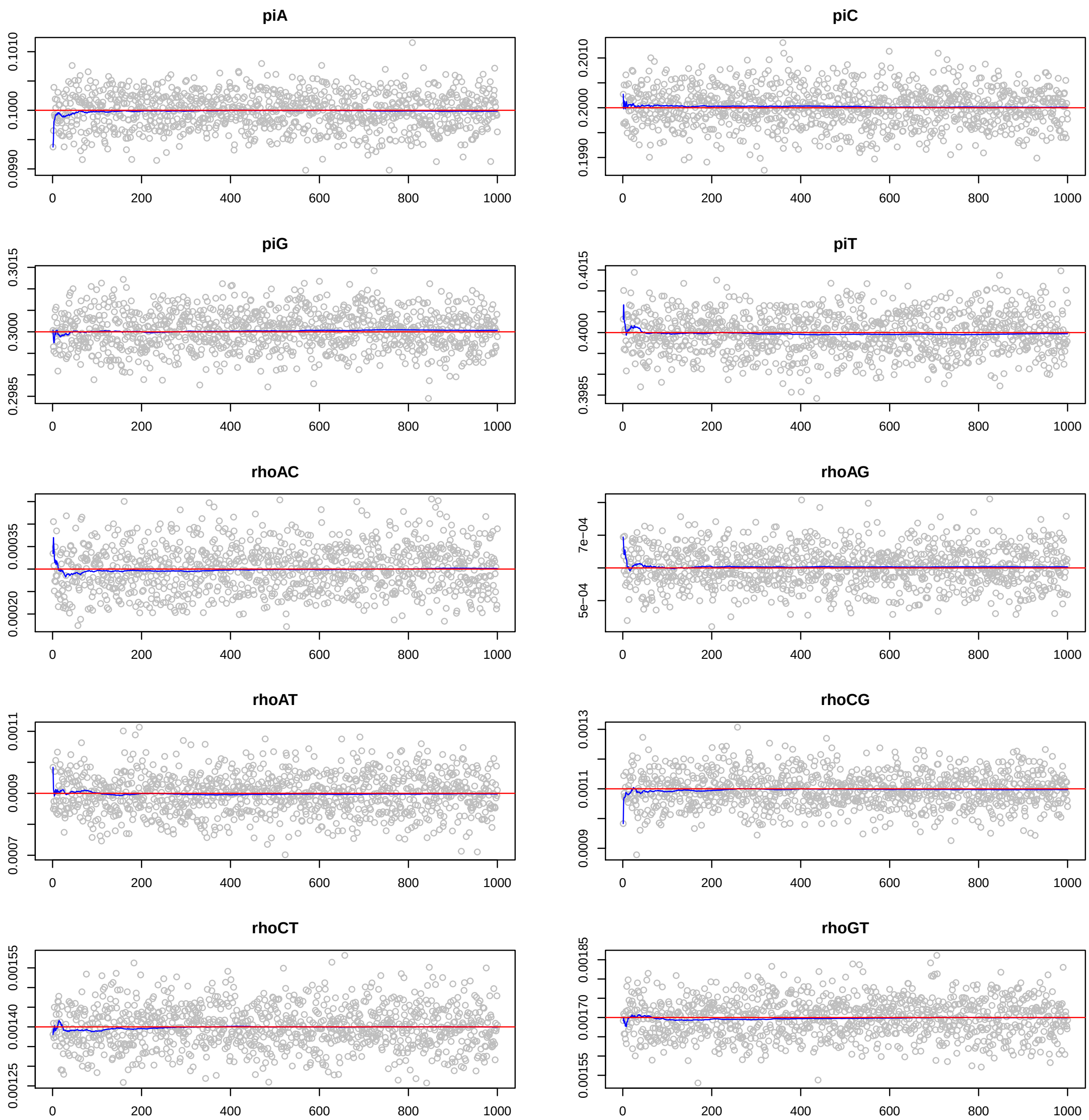


Figure S4: **Validation of the Bayesian algorithms.** Simulation conditions: 1000000 sites, 10 individuals and a simple parameter vector for the Moran model with allelic selection: $\boldsymbol{\pi} = (0.25, 0.25, 0.25, 0.25)$, $\boldsymbol{\rho} = (0.001, 0.001, 0.001, 0.001, 0.001, 0.001)$, $\boldsymbol{\sigma} = (1.00, 1.00, 1.00, 1.00)$. The blue line represents the MCMC moving average whereas the red one represents the true values.

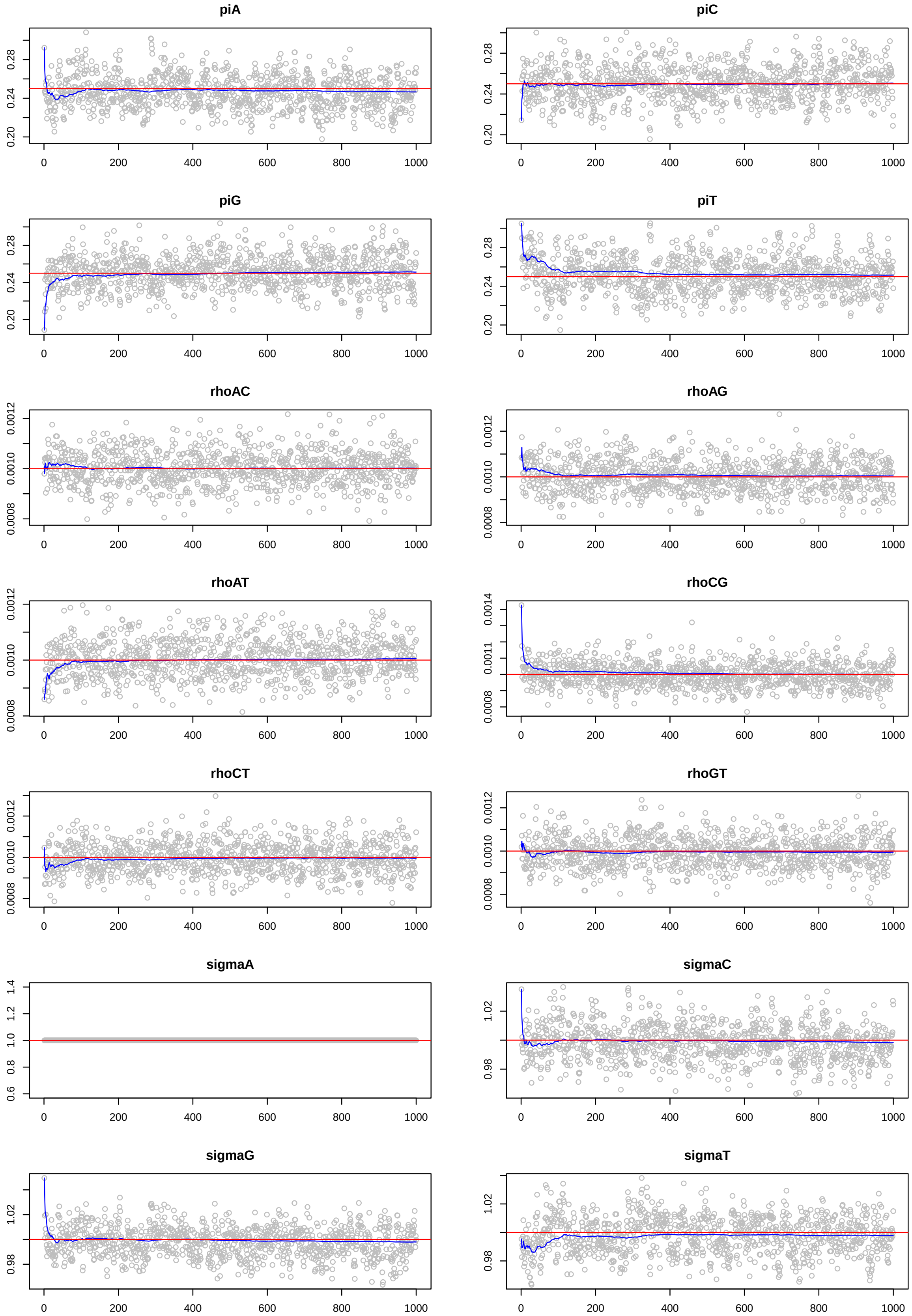


Figure S5: **Validation of the Bayesian algorithms.** Simulation conditions: 1000000 sites, 10 individuals and a complex parameter vector for the Moran model with allelic selection: $\boldsymbol{\pi} = (0.10, 0.20, 0.30, 0.40)$, $\boldsymbol{\rho} = (0.003, 0.006, 0.009, 0.011, 0.014, 0.017)$, $\boldsymbol{\sigma} = (1.00, 1.01, 1.02, 1.03)$. The blue line represents the MCMC moving average whereas the red one represents the true values.

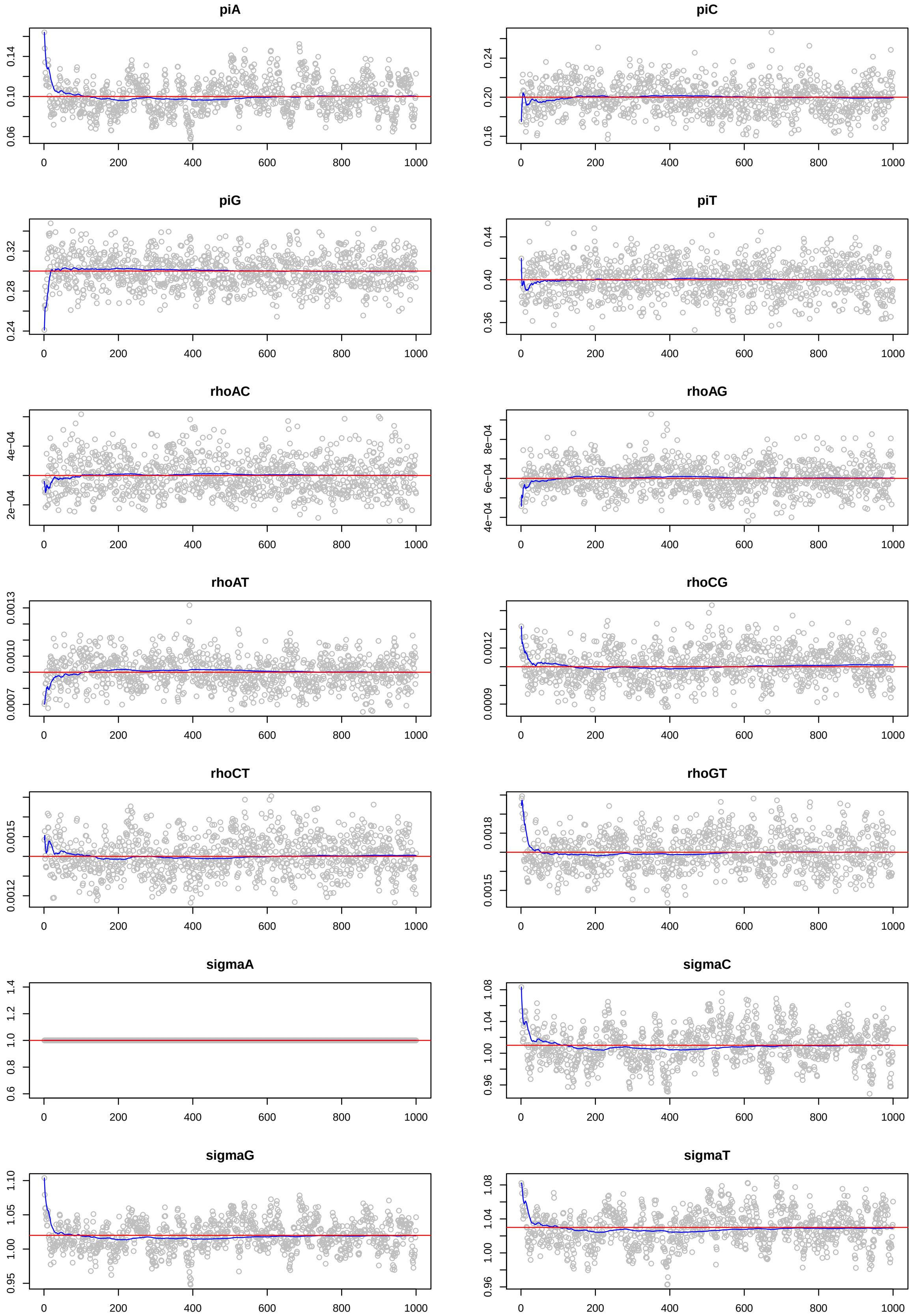
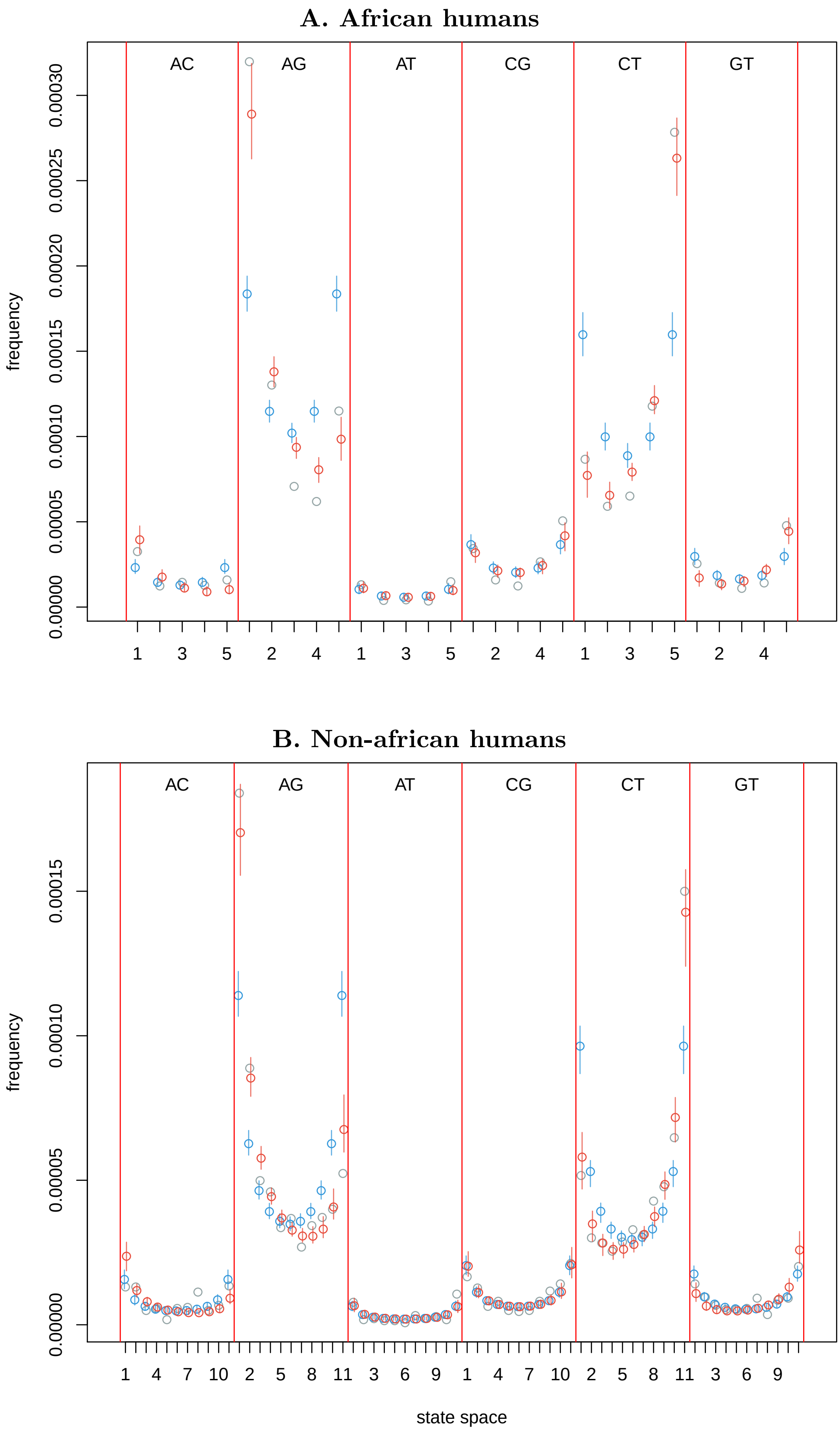
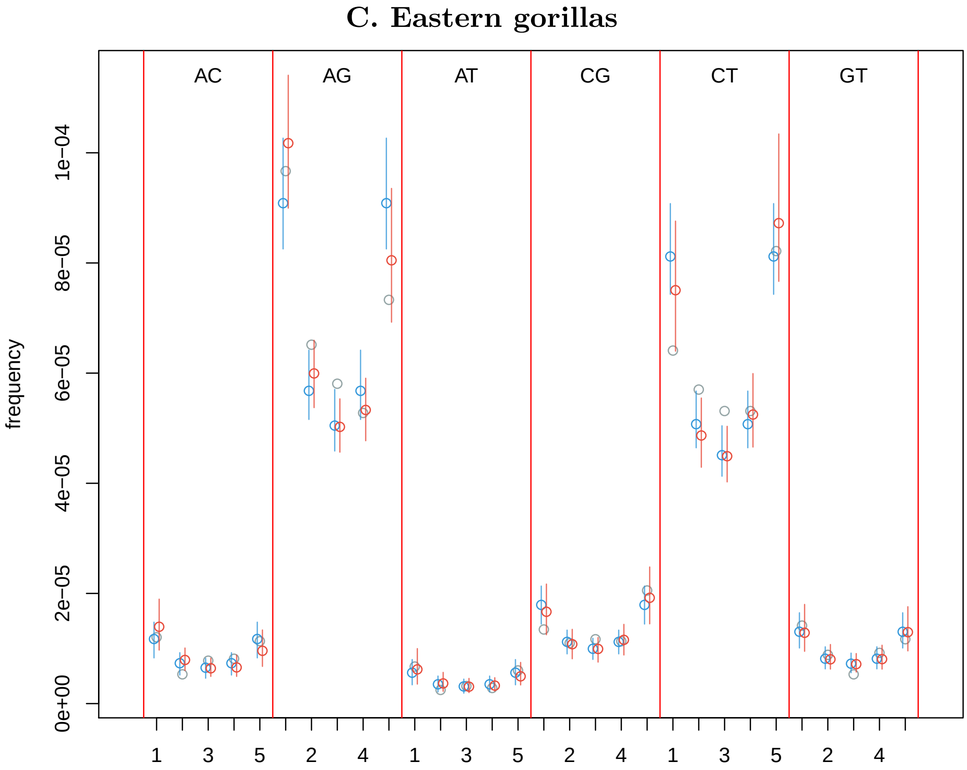
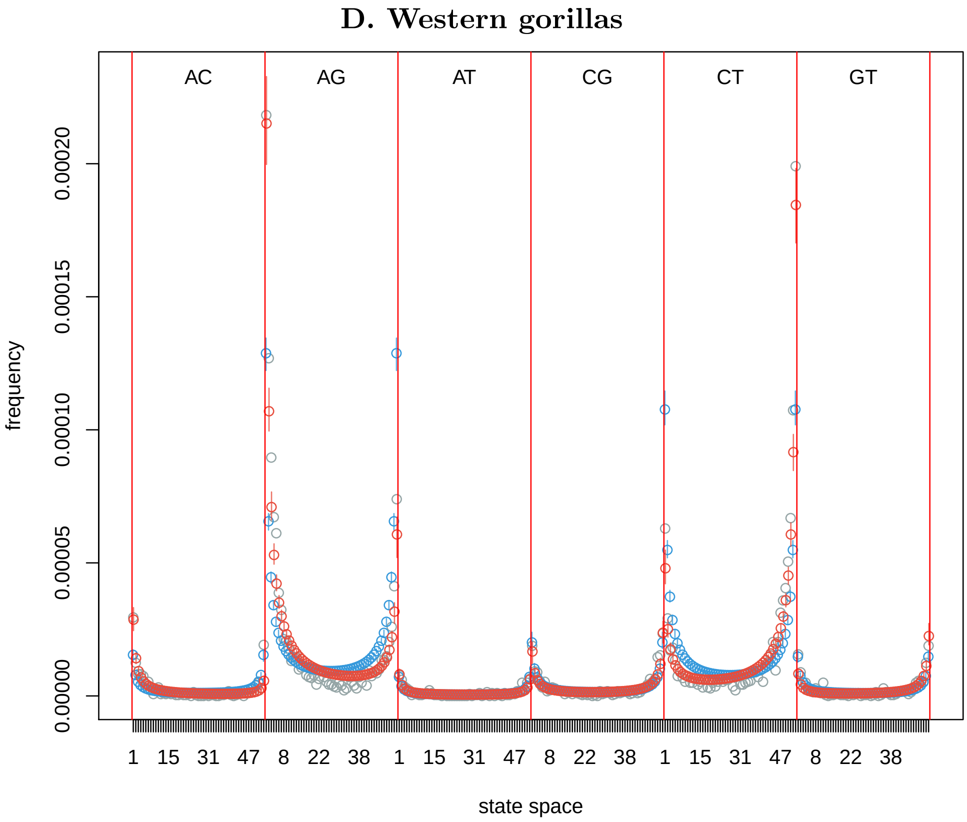


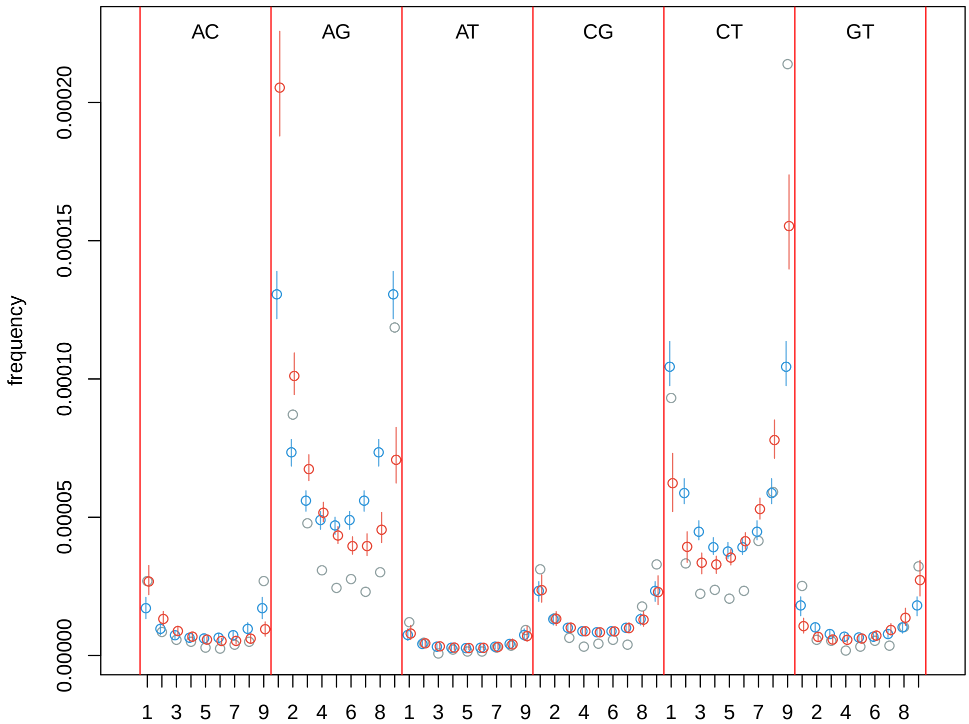
Figure S6: **Prediction of the site-frequency spectrum in great ape populations.** The gray points represent the observed counts and the vertical lines the posterior predictive distribution of the stationary distribution under the 4-variate Moran model: boundary mutation model (blue) and allelic selection model (red).



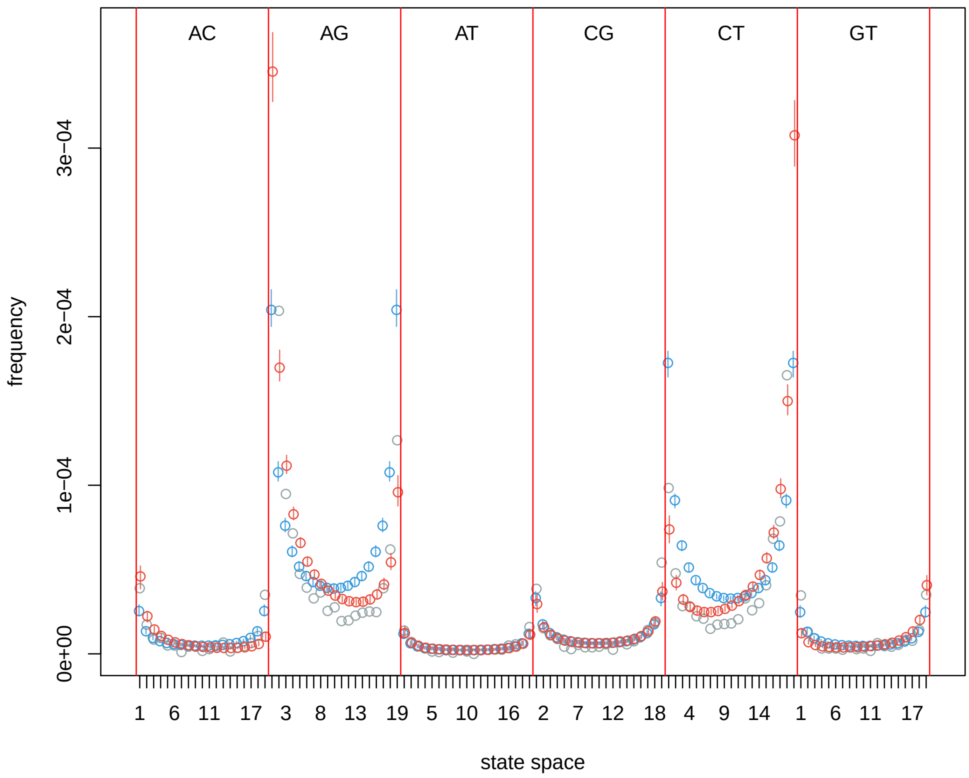

C. Eastern gorillas
AC
AG
AT
CG
CT
GT
frequency
0e+00
2e−05
4e−05
6e−05
8e−05
1e−04
1
3
5
2
4


D. Western gorillas
AC
AG
AT
CG
CT
GT
frequency
0.00000
0.00005
0.00010
0.00015
0.00020
1
15
31
47
8
22
38
state space

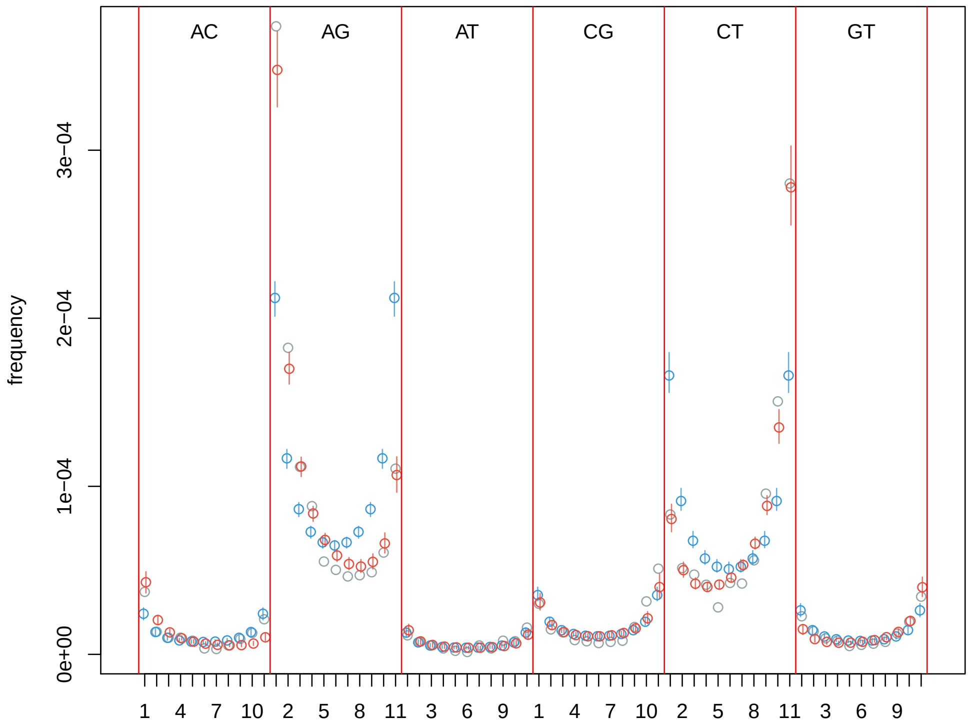
E. Western chimpanzees



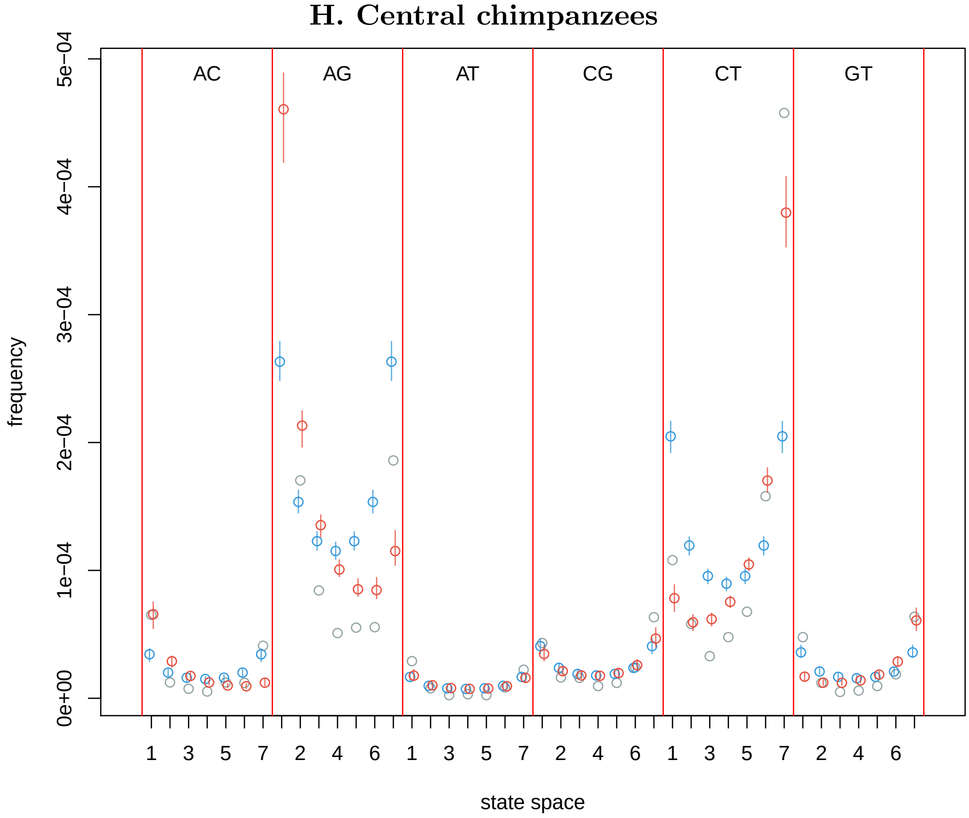
F. Nigeria-Cameroon chimpanzees



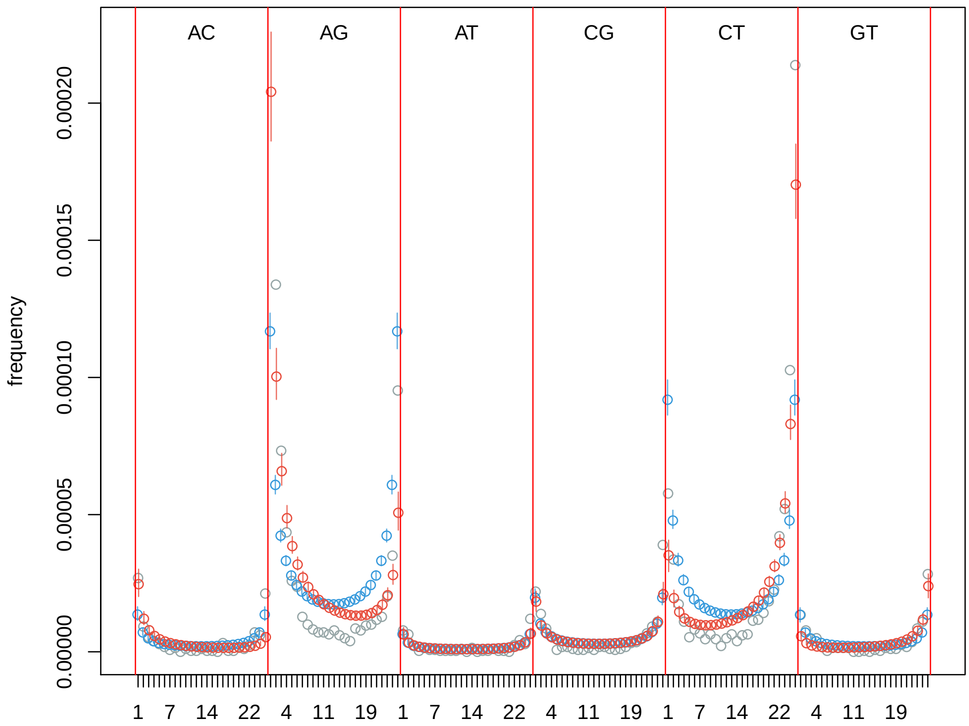
G. Eastern chimpanzees



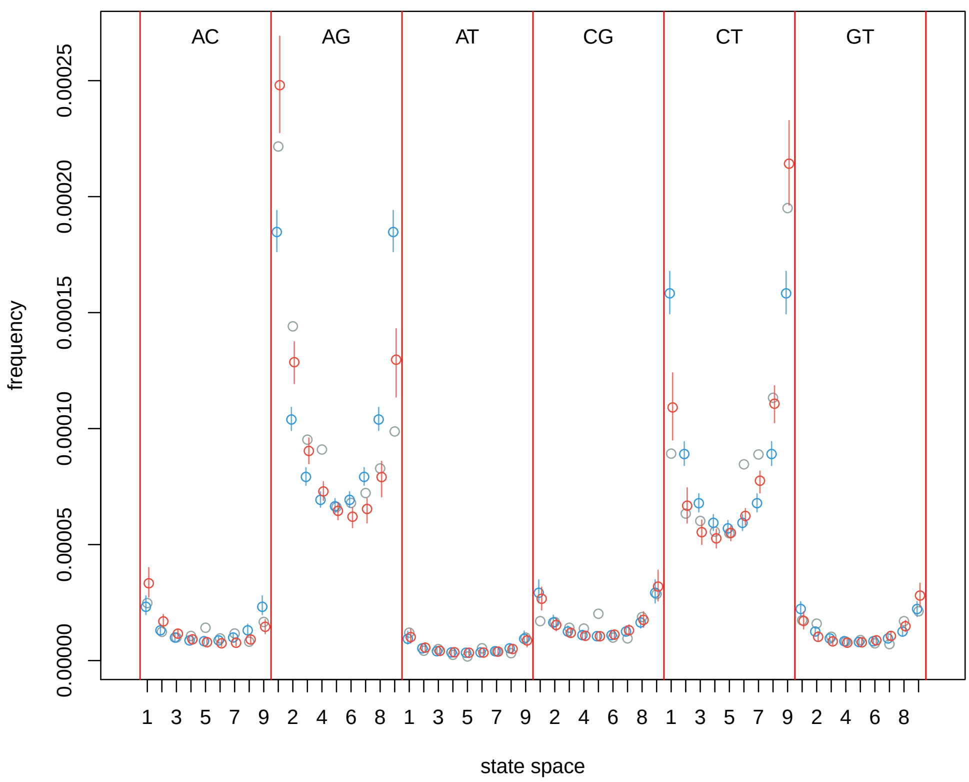
H. Central chimpanzees



I. Bonobos



J. Bornean orangutans



K. Sumatran orangutans

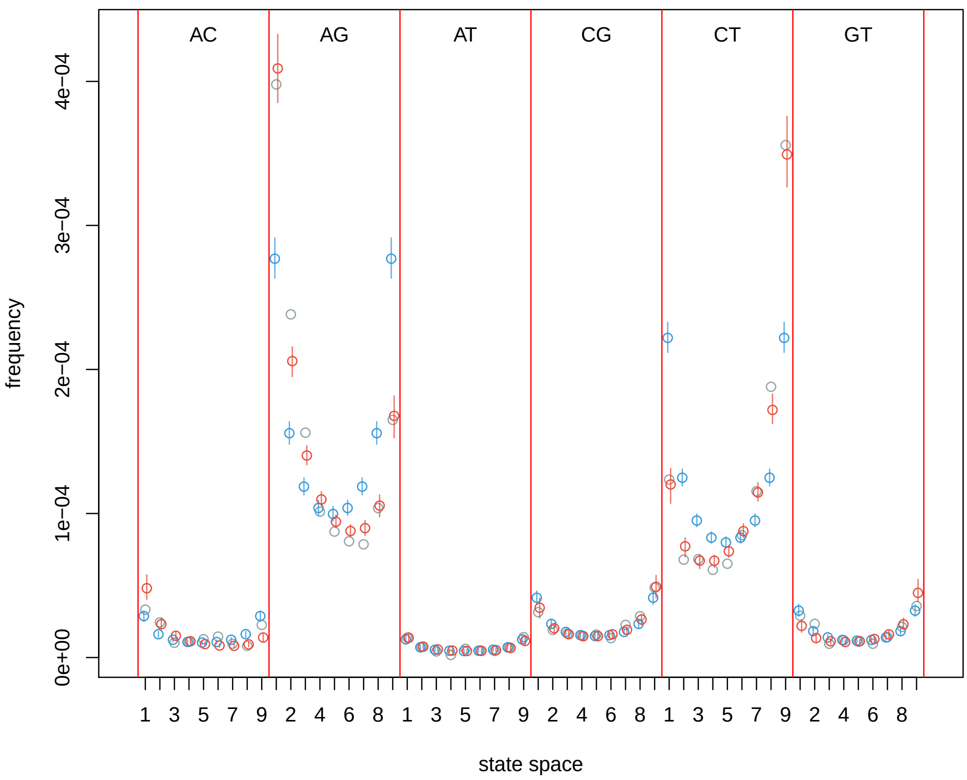
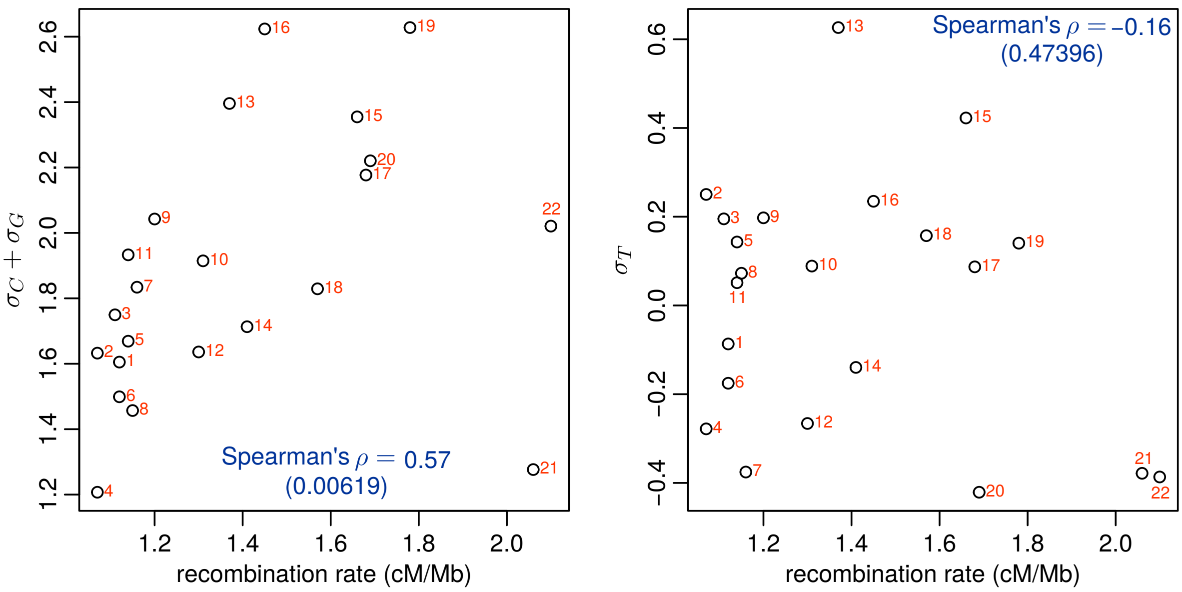


Figure S7: **GC-bias *vs.* recombination rate and chromosome length in non-African humans.** The scaled selection coefficients were estimated based on the posterior average. Recombination rates were estimated by comparing the genetic distance (cM) between markers to the physical (Mb) as described in (Jensen-Seaman (2004)) and based on the human Iceland pedigree map.

A. *σ versus* recombination rate



B. *σ versus* chromossome length

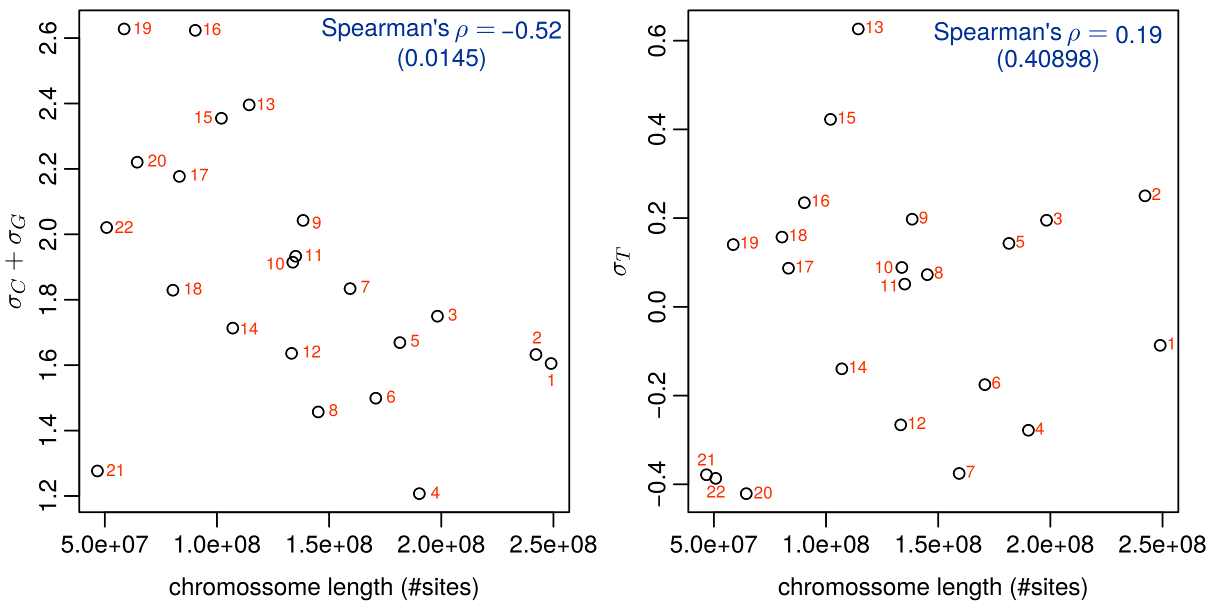


Table S1: **Simulation schemes.** Simulation schemes used to validate the Bayesian algorithms for estimating the model parameters under the multivariate Moran model with mutation (M schemes) and mutation plus selection (S schemes). σ_A is set to 1.

Scheme	π_A	π_C	π_G	π_T	ρ_{AC}	ρ_{AG}	ρ_{AT}	ρ_{CG}	ρ_{CT}	ρ_{GT}	σ_A	σ_C	σ_G	σ_T
M1	0.25	0.25	0.25	0.25	0.001	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	0.001	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 S2: **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	μ_{AC}	μ_{CA}	μ_{AG}	μ_{GA}
African humans	0.000237	0.000934	0.002248	0.008030
Non-african humans	0.000464	0.000957	0.003411	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	μ_{AT}	μ_{TA}	μ_{CG}	μ_{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	μ_{CT}	μ_{TC}	μ_{GT}	μ_{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	σ_C	σ_G	σ_T
African humans	2.415	1.864	0.193
Non-african humans	1.192	1.161	0.053
Eastern gorillas	0.592	0.357	0.346
Western gorillas	1.711	1.334	0.285
Bonobos	1.705	1.551	-0.057
Nigeria-Cameroon chimpanzees	1.741	1.473	0.091
Eastern chimpanzees	1.858	1.506	0.241
Central chimpanzees	2.600	2.083	0.145
Western chimpanzees	1.385	1.420	0.150
Sumatran orangutans	1.687	1.176	0.228
Bornean orangutans	1.087	0.846	0.194