

# Supplementary Material

## A detailed description of the PG-BSM

The PG-BSM consists of three components : (1) a model for the evolution of the codon sequence; (2) a model for the evolution of a discrete phenotype; and (3) a model that accounts for the mechanism(s) by which (1) and (2) are associated. The model chosen for the evolution of the phenotype is the proportional rates model, analagous to F81 for DNA (Felsenstein, 1981), characterized by the following rate matrix (in this case assuming  $k = 3$  distinct phenotypes):

$$Q_F = \frac{\lambda}{r_F} \begin{bmatrix} * & \pi_F^2 & \pi_F^3 \\ \pi_F^1 & * & \pi_F^3 \\ \pi_F^1 & \pi_F^2 & * \end{bmatrix} \text{ where } r_F = \sum_{i \neq j} \pi_F^i \pi_F^j \quad (1)$$

Each  $*$  in (1) indicates a value that makes the corresponding row sum to zero. The  $\pi_F^i$  represent the stationary frequencies for the three phenotypic states  $i \in \{1, 2, 3\}$  (the subscript  $F$  is used to indicate that the associated symbol is an element of the model for phenotype evolution). The scaling factor  $r_F$  is included so that the rate constant  $\lambda$  gives the expected number of changes in the discrete phenotypic state per unit branch length. Equation (1) is used not only to compute the probability of the vector of observed phenotypes, but also to generate samples from the distribution of ancestral phenotypes that are required to inform the mechanism(s) of adaptive evolution as described in a separate section below.

The null PG-BSM assumes that some proportion  $\pi_0$  of sites evolved under  $\omega_0 = 0$  over the tree while the remaining sites switched randomly via a covarion-like process between two rate ratios  $\omega_1 < \omega_2$  over time independent of the phenotype. The rate matrix for the covarion-like process can be constructed by expanding the state space to (codon,  $\omega$ ) pairs:

$$Q_{\text{CLM3}} = \frac{1}{c_1} \left[ \begin{array}{c|c} Q(\omega_1) & 0 \\ \hline 0 & Q(\omega_2) \end{array} \right] + \frac{\delta}{c_2} \left[ \begin{array}{c|c} -p_2 I & p_2 I \\ \hline p_1 I & -p_1 I \end{array} \right] \quad (2)$$

Here  $I$  is the identity matrix that is the same size as  $Q(\omega_1)$  and  $Q(\omega_2)$  (e.g.,  $61 \times 61$  for the standard genetic code),  $p_1$  is the average proportion of time sites evolved under  $\omega_1$ , and  $p_2 = 1 - p_1$  the average proportion of time sites evolved under  $\omega_2$ . The vector of stationary frequencies for all possible (codon,  $\omega$ ) pairs is the  $1 \times 122$  vector  $\langle p_1 \boldsymbol{\pi}, p_2 \boldsymbol{\pi} \rangle$ , where  $\boldsymbol{\pi} Q(\omega_1) = \boldsymbol{\pi} Q(\omega_2) = 0$ . The scaling factor  $1/c_1$  is set to make branch lengths equal to the expected number of single nucleotide substitutions per codon:  $c_1 = p_1 r_{\omega_1} + p_2 r_{\omega_2}$ , where  $r_{\omega_k} = \sum_{j \neq i} \pi_i Q_{ij}(\omega_k) \{\ell_1 + 2\ell_2 + 3\ell_3\}$  for  $k \in \{1, 2\}$ . Including the scaling factor  $1/c_2 = 1/(2p_1 p_2)$  permits  $\delta$  to be interpreted as the expected number of switches between  $\omega_1$  and  $\omega_2$  per unit branch length (Jones *et al.*, 2018).

The likelihood function under the null hypothesis that the phenotype and genotype evolved independently is computed as follows (where  $x^h$  is the  $h^{th}$  site pattern in  $X$ ):

$$L_{\text{nul}}(X, \mathbf{F}; \lambda, \boldsymbol{\theta}, \mathbf{t}) = P(\mathbf{F}; \lambda, \mathbf{t}) \prod_{h=1}^n (\pi_0 P_0(x^h; \boldsymbol{\theta}, \mathbf{t}) + (1 - \pi_0) P_{\text{CL}}(x^h; \boldsymbol{\theta}, \mathbf{t})) \quad (3)$$

$P(\mathbf{F}; \lambda, \mathbf{t})$  is the probability of the vector of phenotypes given the rate constant  $\lambda$  and branch lengths  $\mathbf{t}$ ,  $P_0(x^h; \boldsymbol{\theta}, \mathbf{t})$  is the probability of the site pattern  $x^h$  assuming the site evolved under  $\omega_0 = 0$ ,  $P_{\text{CL}}(x^h; \boldsymbol{\theta}, \mathbf{t})$  is the probability of the site pattern assuming it evolved under equation (2), and  $\boldsymbol{\theta} = (\omega_1, \omega_2, p_1, \delta, \kappa)$  is a vector of parameters for sequence evolution. All probabilities are computed using the pruning algorithm (Felsenstein, 1981).

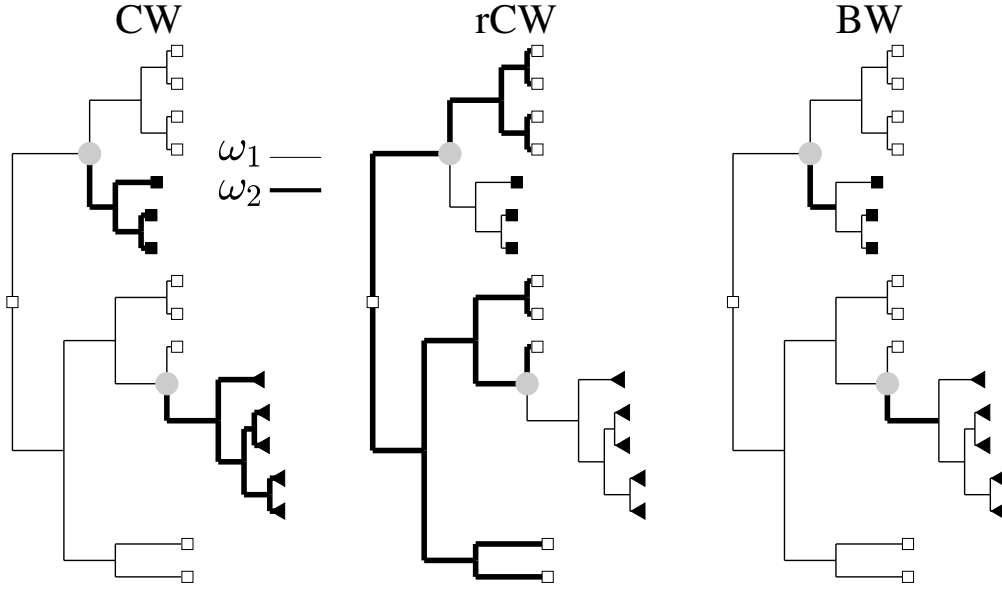


Figure 1: An illustration of the difference between the cladewise (CW and rCW) and branchwise (BW) evolutionary processes assumed under the alternative component of the PG-BSM. Each process accounts for a specific form of heterotachy associated with changes in phenotype. The empty and filled markers at the terminal nodes indicate three phenotypic states.  $\omega_1 < \omega_2$  are  $dN/dS$  rate ratios. The grey disks indicate the nodes at which a change in phenotype occurred. CW sites are assumed to have evolved under  $\omega_1$  prior to a change in phenotype and under  $\omega_2$  after a change. rCW sites are assumed to have evolved under  $\omega_2$  prior to a change in phenotype and under  $\omega_1$  after a change. BW sites are assumed to have evolved under  $\omega_2$  over branches along which a change in phenotype occurred and under  $\omega_1$  everywhere else in the tree. The model assumes a rooted tree because the interpretation of the CW and rCW processes require a particular order of change in rate ratio.

The alternative PG-BSM enforces dependencies between phenotype and genotype evolution at some fraction of sites. Various mechanisms of dependency are amenable to phenomenological representation as distinct modes of heterotachy as illustrated in Figure 1. To infer phenotype-genotype associations of any kind requires knowledge of the ancestral branches over which the phenotype changed or remained

constant. This information is provided by realizations of ancestral phenotypes at the internal nodes of the tree generated using equation (1), each of which is converted to a change map  $\mathbf{z} = (z_1, \dots, z_{2N-2})$  as follows (see Figure 1):

$$z_b = \begin{cases} 0 & \text{if the phenotype is the same at the two ends of branch } b \\ 1 & \text{if the phenotype is different at the two ends of branch } b \end{cases} \quad (4)$$

The likelihood function under the alternative hypothesis can in principle be computed by summing over all possible change maps:

$$L_{\text{alt}}(X, \mathbf{F}; \lambda, \boldsymbol{\theta}, \mathbf{t}) = P(\mathbf{F}; \lambda, \mathbf{t}) \sum_{\mathbf{z}} \left\{ P(\mathbf{z} \mid \mathbf{F}, \lambda, \mathbf{t}) \prod_{h=1}^n g(x^h; \mathbf{z}) \right\} \quad (5)$$

where  $P(\mathbf{z} \mid \mathbf{F}, \lambda, \mathbf{t})$  is the probability of the change map  $\mathbf{z}$  given the terminal states  $\mathbf{F}$ , the rate constant  $\lambda$  and a vector of branch lengths  $\mathbf{t}$ . The mixture probability  $g(x^h; \mathbf{z})$  depends on the particular combination of processes included in the alternate PG-BSM. If the objective was to detect sites consistent with either the cladewise (CW) or branchwise (BW) processes, for example, the mixture probability would be as follows:

$$g(x^h; \mathbf{z}) = \pi_0 P_0(x^h; \boldsymbol{\theta}, \mathbf{t}) + \pi_{\text{CL}} P_{\text{CL}}(x^h; \boldsymbol{\theta}, \mathbf{t}) + \pi_{\text{CW}} P_{\text{CW}}(x^h; \boldsymbol{\theta}, \mathbf{t}, \mathbf{z}) + \pi_{\text{BW}} P_{\text{BW}}(x^h; \boldsymbol{\theta}, \mathbf{t}, \mathbf{z}) \quad (6)$$

Here  $P_{\text{CW}}(x^h; \boldsymbol{\theta}, \mathbf{t}, \mathbf{z})$  and  $P_{\text{BW}}(x^h; \boldsymbol{\theta}, \mathbf{t}, \mathbf{z})$  give the probabilities of the site pattern  $x^h$  assuming the site evolved under the cladewise and branchwise process, respectively, and  $\pi_0 + \pi_{\text{CL}} + \pi_{\text{CW}} + \pi_{\text{BW}} = 1$ .

The number of possible change maps can be very large depending on the number of taxa and phenotypic states. It is therefore often infeasible to compute equation (5) exactly by summing over all  $\mathbf{z}$ . Instead,  $P(\mathbf{z} \mid \mathbf{F}, \lambda, \mathbf{t})$  is approximated by the relative frequency  $\hat{\pi}_{\mathbf{z}}$  of  $\mathbf{z}$  in a sample of  $10^5$  realizations of ancestral phenotypes. These are generated using the maximum likelihood estimates for  $\lambda$  and  $\mathbf{t}$  and the model for phenotype evolution as described in a separate section below. The summation in equation (5) is then over all unique change maps that appear in the sample using  $\hat{\pi}_{\mathbf{z}}$  in place of  $P(\mathbf{z} \mid \mathbf{F}, \lambda, \mathbf{t})$ . Computation time increases with the number of unique change maps. To reduce computational load, change maps  $\mathbf{z}$  that occurred with a frequency  $< 10^{-3}$  are excluded and the probabilities  $\hat{\pi}_{\mathbf{z}}$  renormalized to sum to one. Note that estimates of  $\lambda$  and  $\mathbf{t}$  are required to generate ancestral phenotypes. An exact but costly approach would be to resample from  $P(\mathbf{z} \mid \mathbf{F}, \lambda, \mathbf{t})$  with each iterative update of  $\lambda$  and  $\mathbf{t}$  inside the optimization function. A less costly approximation is implemented instead as follows. Maximum likelihood estimates  $(\hat{\lambda}, \hat{\mathbf{t}})$  obtained by fitting the null PG-BSM to the data are first used to generate a preliminary sample. The alternate PG-BSM is then fitted to the data using this sample to produce new estimates  $(\hat{\lambda}, \hat{\mathbf{t}})$ . To account for any resulting changes in  $(\hat{\lambda}, \hat{\mathbf{t}})$ , a second sample is generated using the new estimates and the alternate PG-BSM fitted once more to produce the final results.

Rejection of the null hypothesis provides evidence for phenotype-genotype association. Naive empirical Bayes analysis can then be used to identify the most likely category for each site. Let  $c \in \{\text{CW}, \text{rCW}, \text{BW}\}$  index the three categories of phenotype-genotype association possibly included in the alternative model. The posterior probability that a site evolved under the process indicated by category  $c$  is evaluated at the maximum likelihood estimate of the vector of parameters of the alternate PG-BSM as follows:

$$P(c \mid x^h) = \frac{L_{\text{alt}}(x^h \mid c) \hat{\pi}_c}{L_{\text{alt}}(x^h)} \quad (7)$$

To assess the accuracy and power of the *post hoc* analyses of our simulation studies, the mean observed false discovery count and the mean power were computed for each set of  $S$  simulated alignments as follows:

$$\text{FDC}(c) = \frac{1}{S} \sum_{i=1}^S F_i(c), \text{ Power} = \frac{1}{S} \sum_{i=1}^S \frac{D_i(c) - F_i(c)}{n(c)} \quad (8)$$

Here  $F_i(c)$  is the number of false discoveries of category  $c$  sites and  $D_i(c)$  the total number of discoveries of category  $c$  sites, both for the  $i^{\text{th}}$  alignment, and  $n(c)$  is the number of sites in the alignment that were evolved under the process indicated by category  $c$ . The expected false discovery count  $E\{\text{FDC}\} = E\{F_i(c)\}$  can be controlled by setting a posterior threshold that is specific to the alignment under consideration as described in a separate section below. This threshold can change from one data set to another. For the analyses presented in the paper we used  $E\{\text{FDC}\} \in \{1, 2\}$  for each category of sites  $c \in \{\text{CW}, \text{rCW}, \text{BW}\}$  included in the alternate PG-BSM.

To quantify the evidential support for branches over which the phenotype is thought to have changed, the probability of the most frequently sampled change map  $\mathbf{z}^*$  conditioned on the combined data is estimated as follows:

$$\hat{P}(\mathbf{z}^* \mid X, \mathbf{F}) = \frac{L_{\text{alt}}(X, \mathbf{F} \mid \mathbf{z}^*) \hat{\pi}_{\mathbf{z}^*}}{L_{\text{alt}}(X, \mathbf{F})} \quad (9)$$

where  $\hat{\pi}_{\mathbf{z}^*}$  is the frequency of the most commonly sampled change map. Equation (9) is evaluated at the MLE for the alternate PG-BSM. The observed frequency  $\hat{\pi}_{\mathbf{z}^*}$  depends on  $X$  only through branch length estimates. The likelihood  $L_{\text{alt}}(X, \mathbf{F} \mid \mathbf{z}^*)$ , by contrast, also depends on the existence of individual site patterns that match to greater or lesser degree the patterns of phenotype-genotype association indicated by  $\mathbf{z}^*$ . An alignment generated with no phenotype-genotype association will tend to result in  $L_{\text{alt}}(X, \mathbf{F} \mid \mathbf{z}^*)/L_{\text{alt}}(X, \mathbf{F}) \approx 1$  making  $\hat{P}(\mathbf{z}^* \mid X, \mathbf{F}) \approx \hat{\pi}_{\mathbf{z}^*}$ , in concordance with an alignment that contains no information about ancestral phenotypic states. When there is phenotype-genotype association at some sites the ratio of likelihoods will be greater than one depending on the proportion of sites generated under the cladewise, reverse cladewise and branchwise processes and on the extent to which  $\mathbf{z}^*$  matches the change map that generated the data. A good match combined with strong signal will tend to result in  $L_{\text{alt}}(X, \mathbf{F} \mid \mathbf{z}^*)/L_{\text{alt}}(X, \mathbf{F}) > 1$  making  $\hat{P}(\mathbf{z}^* \mid X, \mathbf{F}) > \hat{\pi}_{\mathbf{z}^*}$ .

# Tabulated Simulation Results

Table 1: Simulation 1 Parameter Estimates

| parameter  | generating | C 300            | C 1000           | UC 300           | UC 1000          |
|------------|------------|------------------|------------------|------------------|------------------|
| $\pi_0$    | 0.65       | 0.65/0.65,(0.03) | 0.65/0.65,(0.01) | 0.65/0.65,(0.01) | 0.65/0.65,(0.01) |
| $\omega_1$ | 0.10       | 0.10/0.10,(0.03) | 0.10/0.10,(0.01) | 0.10/0.10,(0.02) | 0.10/0.10,(0.01) |
| $\omega_2$ | 1.50       | 1.51/1.50,(0.25) | 1.57/1.54,(0.18) | 1.56/1.55,(0.26) | 1.51/1.52,(0.14) |
| $p_1$      | 0.80       | 0.78/0.78,(0.03) | 0.83/0.83,(0.02) | 0.80/0.80,(0.03) | 0.79/0.79,(0.02) |
| $\delta$   | 0.20       | 0.19/0.18,(0.07) | 0.21/0.20,(0.04) | 0.20/0.19,(0.06) | 0.20/0.19,(0.03) |
| $\kappa$   | 4.61       | 4.71/4.66,(0.42) | 4.53/4.53,(0.20) | 4.60/4.63,(0.37) | 4.54/4.54,(0.19) |

Table 2: Mean/median, (standard deviation) of select maximum likelihood estimates for Simulation 1. C indicates simulations using the clocked tree, UC simulations using the unclocked tree; 300 and 1000 indicate the number of simulated codon sites. One hundred alignments were generated under each simulation scenario.

Table 3: Simulation 2 Parameter Estimates

| generating model                                | $\hat{\pi}_0$ | $\hat{\omega}_2$ | $\hat{\pi}_{CW}$ | $\hat{\pi}_{BW}$ | rejections |
|---|---------------|------------------|------------------|------------------|------------|
| Scenario 2a ( $\pi_{CW}, \pi_{BW}$ ) = (0%, 0%) |               |                  |                  |                  |            |
| MSmmtDNA 0% DT                                  | 0.59          | 1.10             | 0.01             | 0.00             | 0/50 false |
| MSmmtDNA 6% DT                                  | 0.65          | 1.40             | 0.01             | 0.01             | 0/50 false |
| MSTGdR 0% DT                                    | 0.76          | 2.31             | 0.00             | 0.00             | 0/50 false |
| Scenario 2b ( $\pi_{CW}, \pi_{BW}$ ) = (5%, 0%) |               |                  |                  |                  |            |
| MSmmtDNA 0% DT                                  | 0.55          | 1.11             | 0.07             | 0.01             | 47/50 true |
| MSmmtDNA 6% DT                                  | 0.58          | 1.25             | 0.07             | 0.01             | 46/50 true |
| MSTGdR 0% DT                                    | 0.71          | 1.38             | 0.04             | 0.00             | 42/50 true |
| Scenario 2c ( $\pi_{CW}, \pi_{BW}$ ) = (0%, 5%) |               |                  |                  |                  |            |
| MSmmtDNA 0% DT                                  | 0.59          | 1.10             | 0.01             | 0.09             | 46/50 true |
| MSmmtDNA 6% DT                                  | 0.57          | 2.20             | 0.02             | 0.07             | 38/50 true |
| MSTGdR 0% DT                                    | 0.73          | 1.35             | 0.01             | 0.07             | 50/50 true |
| Scenario 2d ( $\pi_{CW}, \pi_{BW}$ ) = (0%, 0%) |               |                  |                  |                  |            |
| MSmmtDNA 0% DT                                  | 0.60          | 1.27             | 0.01             | 0.01             | 1/50 false |
| MSmmtDNA 6% DT                                  | 0.56          | 2.30             | 0.01             | 0.00             | 1/50 false |
| MSTGdR 0% DT                                    | 0.72          | 1.37             | 0.00             | 0.00             | 2/50 false |
| Scenario 2e ( $\pi_{CW}, \pi_{BW}$ ) = (5%, 5%) |               |                  |                  |                  |            |
| MSmmtDNA 0% DT                                  | 0.55          | 1.10             | 0.06             | 0.09             | 50/50 true |
| MSmmtDNA 6% DT                                  | 0.57          | 1.42             | 0.06             | 0.08             | 50/50 true |
| MSTGdR 0% DT                                    | 0.70          | 1.36             | 0.06             | 0.06             | 50/50 true |

Table 4: Select mean maximum likelihood estimates and omnibus test results for Simulation 2. Note that Scenario 2d used the same alignments as Scenario 2c but with a misspecified vector of phenotypes (i.e.,  $\pi_{BW} = 0$  under Scenario 2d because, although the data was generated with peak shifts, these did not co-occur with changes in phenotype as indicated by the misspecified vector  $\mathbf{F}$ ). % DT indicates the proportion of double or triple substitutions permitted by the generating model.

Table 5: Simulation 2 *post hoc* Analysis

| generating model  | CW FDC | CW Power   | BW FDC | BW Power   | (prior, post) | matches |
|---|--------|------------|--------|------------|---------------|---------|
| Scenario 2b ( $\pi_{\text{CW}}, \pi_{\text{BW}} = (5\%, 0\%)$ ) |        |            |        |            |               |         |
| MSmmtDNA 0% DT  | 0.80   | 0.31(4.72) | 0.86   | -          | (0.73,0.96)   | 48/50   |
| MSmmtDNA 6% DT  | 0.66   | 0.34(5.12) | 1.16   | -          | (0.70,0.96)   | 47/50   |
| MSTGdR 0% DT  | 0.62   | 0.36(5.44) | 1.20   | -          | (0.62,0.88)   | 44/50   |
| Scenario 2c ( $\pi_{\text{CW}}, \pi_{\text{BW}} = (0\%, 5\%)$ ) |        |            |        |            |               |         |
| MSmmtDNA 0% DT  | 1.30   | -          | 1.54   | 0.37(5.60) | (0.76,1.00)   | 50/50   |
| MSmmtDNA 6% DT  | 1.86   | -          | 1.08   | 0.24(3.54) | (0.72,0.98)   | 49/50   |
| MSTGdR 0% DT  | 1.76   | -          | 1.20   | 0.59(8.88) | (0.73,1.00)   | 50/50   |
| Scenario 2e ( $\pi_{\text{CW}}, \pi_{\text{BW}} = (5\%, 5\%)$ ) |        |            |        |            |               |         |
| MSmmtDNA 0% DT  | 1.04   | 0.29(4.38) | 2.12   | 0.28(4.14) | (0.75,1.00)   | 50/50   |
| MSmmtDNA 6% DT  | 1.90   | 0.26(3.90) | 1.76   | 0.34(5.10) | (0.77,1.00)   | 50/50   |
| MSTGdR 0% DT  | 1.44   | 0.46(6.86) | 1.48   | 0.42(6.18) | (0.78,1.00)   | 50/50   |

Table 6: Results of the *post hoc* analysis of Simulation 2 datasets. Column headings: CW FDC - the average number of cladewise sites discovered per alignment that were false; CW Power - average proportion of sites generated under the cladewise processes that were correctly identified (in brackets is the average number of true discoveries per alignment); BW FDC (branchwise false discovery count) and BW power are similarly defined; prior =  $\hat{\pi}_{\mathbf{z}^*}$  - the frequency of the most frequently sampled change map  $\mathbf{z}^*$  averaged over trials; post =  $\hat{P}(\mathbf{z}^* | X, \mathbf{F})$  - the probability of  $\mathbf{z}^*$  conditioned on all of the data averaged over trials; matches - the number of trials for which  $\mathbf{z}^*$  matched the generating change map. % DT indicates the proportion of double or triple substitutions permitted by the generating model.

Table 7: Simulation 3 Parameter Estimates

| Scenario  | $\hat{\pi}_0$ | $\hat{\omega}_2$ | $\hat{\pi}_{\text{CW}}$ | $\hat{\pi}_{\text{BW}}$ | rejections |
|---|---------------|------------------|-------------------------|-------------------------|------------|
| 3a, ( $\pi_{\text{CW}}, \pi_{\text{BW}} = (0\%, 0\%)$ ) | 0.62          | 1.00             | 0.00                    | 0.00                    | 0/50 false |
| 3b, ( $\pi_{\text{CW}}, \pi_{\text{BW}} = (5\%, 0\%)$ ) | 0.59          | 1.05             | 0.02                    | 0.00                    | 15/50 true |
| 3c, ( $\pi_{\text{CW}}, \pi_{\text{BW}} = (0\%, 5\%)$ ) | 0.60          | 2.21             | 0.00                    | 0.06                    | 50/50 true |
| 3d, ( $\pi_{\text{CW}}, \pi_{\text{BW}} = (5\%, 5\%)$ ) | 0.57          | 1.52             | 0.04                    | 0.05                    | 50/50 true |

Table 8: Selected results for Simulation 3 where there were four phenotypes evolved from three discrete changes over the tree.

Table 9: Simulation 3 *post hoc* Analysis

| Scenario | CW FDC | CW Power   | BW FDC | BW Power    | (prior, post) | matches |
|----------|--------|------------|--------|-------------|---------------|---------|
| 3b       | 0.08   | 0.16(2.46) | 0.14   | -           | (0.38,0.80)   | 49/50   |
| 3c       | 0.44   | -          | 0.32   | 0.67(10.18) | (0.50,1.00)   | 50/50   |
| 3d       | 0.68   | 0.36(5.38) | 0.78   | 0.67(10.12) | (0.56,1.00)   | 49/50   |

Table 10: Results of the *post hoc* analysis of Simulation 3 datasets. See caption for Table 6 for definitions of each column heading.

## Simulation 2 Additional Analysis

To determine the impact of the generating scenario on the number of false discoveries in Simulation 2, a generalized linear model was fitted to the false discovery counts of each type for each alignment using as predictor variables: the generating model ( $x_{\text{MS}} = 1, 2$  or  $3$  for MSmmtDNA 0% double-triple mutations, MSmmtDNA 6% double-triple mutations, and MSTGdR 0% double-triple mutations); the presence or absence of sites evolved under the cladewise process ( $x_{\text{CW}} = 1$  or  $0$ ); and the presence or absence of sites evolved under the branchwise process ( $x_{\text{BW}} = 1$  or  $0$ ). The generating model was found to have no significant effect on either of the expected false discovery counts, neither FDC(CW) (p-value = 0.08) nor FDC(BW) (p-value = 0.12). After removing  $x_{\text{MS}}$  from the analysis, the following significant relationships were found:

$$\log \{\text{FDC}(\text{CW})\} = -0.37 + 0.80x_{\text{BW}} \quad (10)$$

$$\log \{\text{FDC}(\text{BW})\} = -0.27 + 0.34x_{\text{CW}} + 0.51x_{\text{BW}} \quad (11)$$

For alignments generated with no phenotype-genotype association (i.e., with  $x_{\text{CW}} = x_{\text{BW}} = 0$ ) the first model predicts an average false discovery count of  $\text{FDC}(\text{CW}) = 0.69$  per alignment for cladewise sites and the second model an average of  $\text{FDC}(\text{BW}) = 0.76$  per alignment for branchwise sites, both slightly under the target  $E\{\text{FDC}\} = 1$ . The first model predicts an average  $\text{FDC}(\text{CW}) = 1.54$  for alignments generated with branchwise sites whether or not cladewise sites were present. The second model predicts  $\text{FDC}(\text{BW}) = 1.07$  for alignments generated with cladewise sites only and  $\text{FDC}(\text{BW}) = 1.27$  for alignments generated with branchwise sites only. The predicted rate is nearly two false detections per alignment when the generating process includes both cladewise and branchwise sites, with  $\text{FDC}(\text{BW}) = 1.78$  per alignment.



# Analysis of mammalian mtDNA

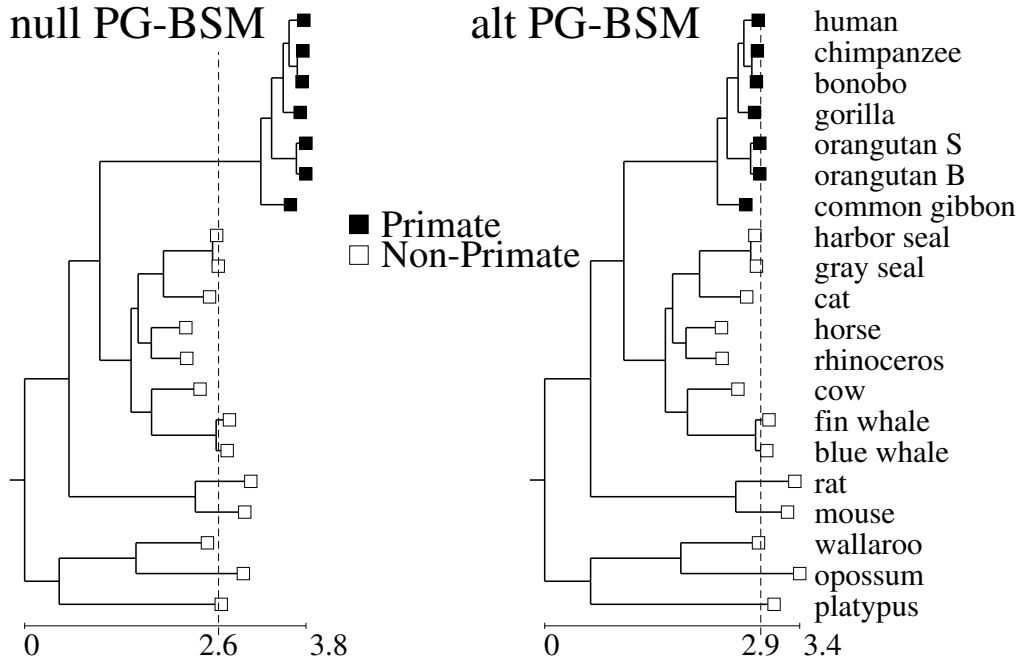


Figure 2: Branch lengths (the expected number of single nucleotide substitutions per codon) estimated by fitting the null and alternate PG-BSM to the alignment of 12 concatenated H-strand mitochondrial DNA sequences (3331 codon sites) from 20 mammalian species distributed in alignment form by the PAML software package (Yang, 2007). Trees are drawn to the same scale. Vertical dashed lines indicate the average tip-to-root distance ( $d_i$ ,  $i = 1, \dots, 20$ ) across terminal nodes. Accounting for non-stationary processes under the alternate model caused the tree to be more consistent with the molecular clock hypothesis, with less variation in the  $d_i$ .

Mitochondrial genes encode proteins involved in basic metabolic processes and are therefore thought to be functionally constrained most of the time. Signatures of adaptation in mammalian mtDNA have nevertheless been detected. For example, Pupko and Galtier (2002) detected significant differences in the replacement rate at some sites in mitochondrial genes in a clade of 7 simian primates compared to a clade of 27 other mammalian lineages. The authors interpreted this as evidence of adaptation and speculated that it might have been due to changes in metabolic requirements related to an increase in the size of the neocortex among primates compared to the other mammals included in their analysis (Pupko and Galtier, 2002). Here we present an analysis of a similar alignment of 12 mammalian mtDNA genes 3331 codons in length taken from 20 mammalian species (Yang, 2007). We used primate/non-primate as a binary phenotype to facilitate comparison with Pupko and Galtier (2002), but consider this to be a proxy for an unobserved and unknown phenotype. The analysis presented here is included because it illustrates the potential impact of accounting for non-stationary clade-wise and branch-wise processes on branch-length estimates (i.e., by making the tree more clock-like), and provides a case study that suggests that the

PG-BSM can be robust to confounding by branch-length effects.

Four models of increasing complexity were fitted to the alignment: (i) the model M3( $k = 2$ ) (Yang *et al.*, 2000), which assumes that some proportion  $p_1$  of sites evolved under a smaller  $\omega_1$  and the remaining sites under a larger  $\omega_2$  across the tree without heterotachy; (ii) the covarion-like model CLM3( $k = 2$ ), which assumes that sites evolved under  $\omega_1$  a proportion  $p_1$  of the time but switched randomly between  $\omega_1$  and  $\omega_2$  at a rate of  $\delta$  switches per single nucleotide substitution; (iii) the null PG-BSM, which combines CLM3( $k = 2$ ) with a category of sites that evolved under  $\omega_0 = 0$ ; and (iv) the alternate PG-BSM, which adds to the null model two categories of sites for the cladewise and branchwise processes. Note that CLM3( $k = 2$ ) is fitted to the alignment  $X$  alone, whereas the PG-BSM is fitted to  $X$  together with the vector of phenotypes  $\mathbf{F}$ . We therefore report the log-likelihood of the PG-BSM as the sum of two values:

$$\begin{aligned} \ln\{\mathcal{L}_{\text{null}}(X, \mathbf{F}; \lambda, \boldsymbol{\theta}, \mathbf{t})\} &= \ln\{P(\mathbf{F}; \lambda, \mathbf{t})\} + \ln\left\{\prod_{h=1}^n (\pi_0 P_0(x^h; \boldsymbol{\theta}, \mathbf{t}) + (1 - \pi_0) P_{\text{CL}}(x^h; \boldsymbol{\theta}, \mathbf{t}))\right\} \\ &= \text{LL}_{\mathbf{F}}(\mathbf{F}; \lambda, \mathbf{t}) + \text{LL}_X(X; \boldsymbol{\theta}, \mathbf{t}) \end{aligned}$$

Similarly:

$$\text{LL}_{\text{CL}}(X; \boldsymbol{\theta}, \mathbf{t}) = \ln\left\{\prod_{h=1}^n P_{\text{CL}}(x^h; \boldsymbol{\theta}, \mathbf{t})\right\}$$

where  $\boldsymbol{\theta} = \langle \omega_1, \omega_2, p_1, \delta, \kappa, \rangle$ . The log-likelihood ratio  $\text{LLR} = 2\{\text{LL}_X(X; \boldsymbol{\theta}, \mathbf{t}) - \text{LL}_{\text{CL}}(X; \boldsymbol{\theta}, \mathbf{t})\}$  evaluated at the MLEs for the two models provides a means to test the significance of accounting for sites evolving under  $\omega_0 = 0$ . Note that, whereas  $\mathbf{t}$  is estimated from  $(X, \mathbf{F})$  under the null PG-BSM, it is estimated from  $X$  alone under CLM3( $k = 2$ ). Hence the LLR is only approximate. It is likely to be a close approximation however, since the MLE for  $\mathbf{t}$  is determined primarily by the information in  $X$ .

| Model           | $\text{LL}_X$ | $\text{LL}_{\mathbf{F}}$ | $\hat{\pi}_0$ | $\hat{\omega}_1$ | $\hat{\omega}_2$ | $\hat{p}_1$ | $\hat{\delta}$ | $\hat{\pi}_{\text{CW}}$ | $\hat{\pi}_{\text{BW}}$ |
|-----------------|---------------|--------------------------|---------------|------------------|------------------|-------------|----------------|-------------------------|-------------------------|
| M3( $k = 2$ )   | -89,162       | -                        | -             | 0.01             | 0.15             | 0.71        | -              | -                       | -                       |
| CLM3( $k = 2$ ) | -88,880       | -                        | -             | 0.00             | 0.21             | 0.77        | 0.06           | -                       | -                       |
| null PG-BSM     | -88,719       | -4                       | 0.47          | 0.02             | 0.35             | 0.76        | 0.12           | -                       | -                       |
| alt PG-BSM      | -88,681       | -4                       | 0.48          | 0.02             | 0.31             | 0.70        | 0.13           | 0.06                    | 0.04                    |

Table 11: Results of the analysis of the mammalian mtDNA.

Accounting for heterotachy resulted in a large improvement in fit, as the M3 vs CLM3 contrast produced  $\text{LLR} = 2(89,162 - 88,880) = 564$  on one parameter ( $\delta$ ) (Table 11). Including a category for sites that evolved under  $\omega_0 = 0$  also resulted in a large improvement in fit, with  $\text{LLR} = 2(88,880 - 88,719) = 322$  on one parameter ( $\pi_0$ ) for the CLM3 vs null PG-BSM contrast. Accounting for phenotype-genotype association resulted in a smaller but also highly significant improvement, with  $\text{LLR} = 2(88,719 - 88,681) =$

76 on two parameters ( $\pi_{\text{CW}}$  and  $\pi_{\text{BW}}$ ). Trees with branch lengths estimated under the null and alternate PG-BSM are shown in Figure 2. It is interesting that accounting for non-stationary cladewise and branchwise processes under the alternate PG-BSM reduced the length of the branch leading to the primate clade from 2.2 single nucleotide substitutions per codon under the null model to 1.3 under the alternate model. This was accompanied by an increase in the average of the distances  $d_1, \dots, d_{20}$  from each terminal node to the root from 2.6 to 2.9. The combined effect was a reduction in the variance of the  $d_i$ . Accounting for the non-stationary cladewise and branchwise processes therefore caused the tree to be more consistent with the molecular clock hypothesis.

The alternate PG-BSM inferred a fraction of sites ( $\hat{\pi}_{\text{CW}} = 0.06, \hat{\pi}_{\text{BW}} = 0.04$ ) to be associated with the change from non-primate to primate. The *post hoc* analysis identified a total of 17 sites (11 cladewise and 6 branchwise) that likely underwent changes in their site-specific fitness landscapes along the branch leading to the primate clade. Substitution patterns for these sites are shown in Figure 3. Cladewise sites with posteriors  $0.77 \leq P(\text{CW}) \leq 0.87$  (where the upper bound 0.87 corresponds to the maximum value of  $P(\text{CW})$  and the lower bound was determined by setting  $E\{\text{FDC}(\text{CW})\} = 2$ ) exhibit two or more amino acids among primates and typically a single amino acid among non-primates, consistent with relaxation of selection pressure among the primate clade. Branchwise sites with posteriors  $0.67 \leq P(\text{BW}) \leq 0.75$  (where the upper bound 0.75 corresponds to the maximum value of  $P(\text{BW})$  and the lower bound was determined by setting  $E\{\text{FDC}(\text{BW})\} = 2$ ) exhibit a single amino acid among primates and a single different amino acid among non-primates consistent with peak shifts along the branch leading to the primate clade.

| 0.87 ≥ P(CW) ≥ 0.77 |   |   |   |   |   |   |   |   |   | 0.75 ≥ P(BW) ≥ 0.67 |   |   |   |   |   |   |               |
|---------------------|---|---|---|---|---|---|---|---|---|---------------------|---|---|---|---|---|---|---------------|
| Y                   | A | T | A | V | D | S | S | I | A | I                   | A | P | T | H | S | N | human         |
| H                   | A | T | A | V | D | S | S | V | A | I                   | A | P | T | H | S | N | chimpanzee    |
| H                   | A | T | A | V | D | S | S | V | A | T                   | A | P | T | H | S | N | bonobo        |
| Y                   | A | T | A | V | D | S | S | I | A | T                   | A | P | T | H | S | N | gorilla       |
| H                   | T | Q | I | V | S | T | S | V | A | A                   | A | P | T | H | S | N | orangutan S   |
| H                   | T | S | V | T | S | T | G | I | A | A                   | A | P | T | H | S | N | orangutan B   |
| Y                   | T | T | V | A | N | T | S | I | M | T                   | A | P | T | H | S | N | common gibbon |
| R                   | S | A | I | I | K | K | A | F | A | V                   | I | S | C | C | E | G | harbor seal   |
| R                   | S | A | I | I | K | K | A | F | A | V                   | I | S | C | C | E | G | gray seal     |
| R                   | S | A | I | I | K | K | A | F | A | V                   | I | S | C | C | E | G | cat           |
| R                   | S | A | I | I | K | K | A | F | S | V                   | I | S | C | C | E | G | horse         |
| R                   | S | A | I | I | K | K | A | F | S | V                   | I | S | C | C | E | G | rhinoceros    |
| R                   | S | A | I | V | K | K | A | F | S | V                   | I | S | C | C | E | G | cow           |
| R                   | S | A | I | I | K | K | A | F | S | V                   | I | S | C | C | E | G | fin whale     |
| R                   | S | A | I | I | K | K | A | F | S | V                   | I | S | C | C | E | G | blue whale    |
| R                   | S | A | I | I | K | K | A | F | S | V                   | I | S | C | C | E | G | rat           |
| R                   | S | A | I | I | K | K | A | F | S | V                   | I | S | C | C | E | G | mouse         |
| R                   | S | A | I | I | K | K | A | L | A | V                   | I | S | C | C | E | G | wallaroo      |
| R                   | S | A | I | I | K | K | A | L | A | V                   | I | S | C | C | E | G | opossum       |
| R                   | S | A | I | I | K | K | A | F | S | V                   | I | S | C | C | E | G | platypus      |
| 2                   | 2 | 3 | 3 | 3 | 3 | 3 | 2 | 2 | 3 | 2                   | I | S | C | C | E | G |               |

Figure 3: Patterns for sites in the mammalian mtDNA alignment inferred to have evolved in association with a change in the primate/non-primate binary character state after application of false discovery count control with  $E\{FDC\} = 2$ . Rows designate taxa in the same order as they appear in the tree shown in Figure 2. Columns designate sites. The horizontal line separates primates from non-primates. The vertical line separates CW from BW sites. Sites of each type are arranged in order of decreasing posterior probability. The value below each CW site pattern indicates the number of different amino acids among primates at that site. Colors indicate physicochemical properties: small nonpolar (orange), hydrophobic (green), polar (magenta), negatively charged (red), positively charged (blue).

The YN-BSM A and YN-BSM B were also fitted to the alignment using the long branch leading to the primate clade as the foreground. The YN-BSM B is similar to YN-BSM A but estimates the three rate ratios  $\omega_0$ ,  $\omega_1$ , and  $\omega_2$  freely and uses  $M3(k = 2)$  as the null model (Yang and Nielsen, 2002). The YN-BSM B is not a test for  $\omega_2 > 1$ , but only for a change to  $\omega_2$  at some sites along the designated foreground branch. The YN-BSM A detected no evidence of positive selection along the branch leading to the primate clade ( $LLR = 2.62$  compared to a critical value of 3.84 for a 5% test assuming  $LLR \sim \chi_1^2$ , Table 12). Model B detected strong evidence ( $LLR = 54.44$  compared to a critical value of 5.99 for a 5% test assuming  $LLR \sim \chi_2^2$ ) for an elevation in the rate ratio at some sites along the same branch ( $\hat{\omega}_2 = 0.98, \hat{p}_2 = 0.15$ ). The PG-BSM provided the best fit, as the AIC was 177,456 for the PG-BSM, 178,358 for the YN-BSM B and 181,777 for the YN-BSM A.

|          | YN-BSM A LLR = 2.62 < 3.84 |             |             | YN-BSM B LLR = 54.44 > 5.99 |             |             |
|----------|----------------------------|-------------|-------------|-----------------------------|-------------|-------------|
| category | proportion                 | BG $\omega$ | FG $\omega$ | proportion                  | BG $\omega$ | FG $\omega$ |
| 0        | 0.89                       | 0.03        | 0.03        | 0.70                        | 0.01        | 0.01        |
| 1        | 0.07                       | 1.00        | 1.00        | 0.28                        | 0.15        | 0.15        |
| 2a       | 0.04                       | 0.03        | 1.79        | 0.01                        | 0.01        | 0.98        |
| 2b       | 0.00                       | 1.00        | 1.79        | 0.01                        | 0.15        | 0.98        |

Table 12: Results of the fit of alternate YN-BSM A and B to the mammalian mtDNA data. LLR is the log-likelihood ratio for the model contrast.

Figure 4 shows 13 sites inferred by YN-BSM B to have undergone an elevation in rate ratio to  $\hat{\omega}_2 = 0.98$  on the foreground branch after false discovery count control was applied to the Bayes empirical Bayes posteriors with  $E\{FDC\} = 2$ . Three of these sites have amino acid patterns that match the primate/non-primate phenotype (marked by filled circles in Figure 4). All three are among the 6 sites inferred by the PG-BSM to have undergone the branchwise process. None of the sites are consistent with the cladewise process. However, seven of them are consistent with the reverse cladewise process, exhibiting one amino acid among primates and several among non-primates (marked by filled triangles in Figure 4). Similar site patterns were identified by Pupko and Galtier (2002) in their data, and were interpreted as evidence of functional shifts. Motivated by the presence of site patterns of this kind, the alternate PG-BSM accounting for the reverse cladewise process alone was fitted to the alignment. No evidence for the reverse cladewise process was detected ( $\hat{\pi}_{\text{rcw}} = 0.00$ ) despite the fact that the alignment includes a fair number of site patterns that are apparently consistent with this process (e.g., there are 828 sites patterns with one amino acid among primates and 2 or more among the 13 non-primates).

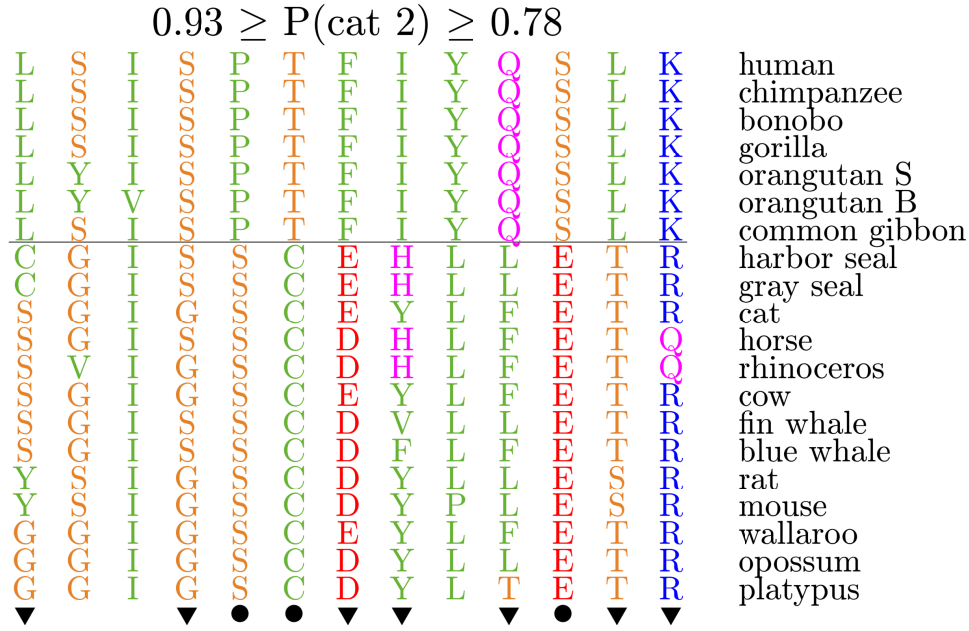


Figure 4: Patterns for sites in the mammalian mtDNA alignment inferred to be in category 2a or 2b by the YN-BSM B after application of false discovery count control with  $E\{FDC\} = 2$ . Rows designate taxa in the same order as they appear in the tree shown in Figure 2. Columns designate sites. The horizontal line separates primates from non-primates. Sites are arranged in order of decreasing posterior probability estimated using YB-BSM B. Filled triangles and circles mark site patterns most consistent with the reverse cladewise and branchwise process, respectively. Colors indicate physicochemical properties: small nonpolar (orange), hydrophobic (green), polar (magenta), negatively charged (red), positively charged (blue).

The fact that the PG-BSM did not detect evidence for the reverse cladewise process suggests that the covarion-like component of the null model provides a measure of protection against branch-length effects (although possibly at the expense of statistical power, cf. Scenario 3b  $(\pi_{CW}, \pi_{BW}) = (5\%, 0\%)$ ). A site that evolved on a static fitness landscape over the tree in Figure 2 is likely to exhibit a single amino acid among the primates because the branches within the primate clade are very short. The same site is likely to exhibit considerable amino acid diversity among the remaining terminal nodes because the non-primates consist of three clades each with relatively long terminal branches. It follows that the site-pattern distribution implied by evolution on static fitness landscapes is similar to that implied by the mechanism whereby sites undergo an increase in the stringency of selection along the branch leading to the primate clade. Heterotachous site patterns  $x^h$  in the alignment consistent with the reverse cladewise process therefore tended to be approximately equally likely under the covarion-like process. Inclusion of the covarion-like model CLM3( $k = 2$ ) in the null PG-BSM consequently resulted in no evidence for the reverse cladewise process, possibly preventing a type I error.

The cladewise process was intended to detect sites that might have undergone a reduction in functional

importance along branches over which the phenotype changed, a process not necessarily associated with adaptation. However, many of the cladewise site patterns identified in the real mammalian mtDNA suggest not only a reduction in the stringency of selection among primates but also a peak shift along the branch leading to the primate clade. The first cladewise site pattern in Figure 3, for example, consists of one amino acid (arginine) among non-primates and two amino acids among primates. A peak shift is suggested by the fact that the amino acids among primates are both different than arginine (histidine and tyrosine). Most of the cladewise sites in Figure 3 have similar patterns. It might therefore be the case that sites that underwent a reduction in the stringency of selection in combination with a peak shift are more readily detected by the cladewise component of the alternate PG-BSM than sites that underwent a reduction in stringency alone. This provides an alternative explanation for the low power to detect cladewise sites in Simulation Scenario 3b ( $\pi_{\text{CW}}, \pi_{\text{BW}} = (5\%, 0\%)$ ), where cladewise sites were generated by reducing the stringency of selection alone. It should be pointed out however that the alternate PG-BSM does not provide a formal test for the co-occurrence of a reduction in the stringency of selection with a peak shift because both processes are expressed in the model by a phenomenological switch to the larger rate ratio  $\omega_2$ . That co-occurrence can be identified informally by appealing to contextual information demonstrates that it might be formally detected by the PG-BSM via inclusion of measures to discriminate between amino acids (cf., Gu, 2006). We leave this task for future efforts.

## Analysis of Phytochrome A&CF

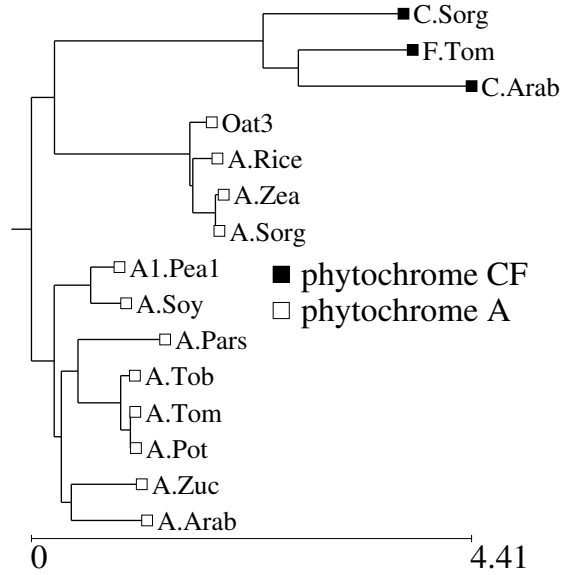


Figure 5: Branch lengths (the expected number of single nucleotide substitutions per codon) estimated by fitting the alternate PG-BSM to the phytochrome A&CF alignment.

Phytochrome is a plant photoreceptor associated with the regulation of developmental processes. Most seed plants contain variants phyA, phyB and phyC, but some also include variants phyD, phyE and phyF that apparently arose following duplication of phyB (Mathews, 2010). We conducted an analysis of an alignment of 15 angiosperm phytochrome sequences 1072 codons in length (Figure 5). Previous analyses of the same data performed as a test case for the YN-BSM partitioned the sequences as phyA versus phyCF (Yang and Nielsen, 2002; Zhang *et al.*, 2005). We used the same partition and conducted an analysis to compare the YN-BSM A and YN-BSM B with the PG-BSM.

No phenotype-genotype association was detected by the PG-BSM. However, both YN-BSM A and YN-BSM B rejected their respective nulls. YN-BSM A detected evidence of positive selection at some sites along the branch leading to the phyCF clade (i.e., category 2 sites with  $\hat{\omega}_2 = 30.17, \hat{p}_2 = 0.12$ ) and identified 29 sites with Bayes empirical Bayes posterior probabilities  $0.95 < P(\text{cat } 2) < 1.00$ , all of which remained after false discovery count control was applied with  $E\{\text{FDC}\} = 1$  (Figure 6). YN-BSM B detected evidence for an elevation in the rate ratio at some sites along the same branch ( $\hat{\omega}_2 = 5.56, \hat{p}_2 = 0.05$ ). Akaike's Information Criterion was 57,510 for the null PG-BSM, 57,700 for the YN BSM B and 58,255 for the YN BSM A. Thus the null PG-BSM provided the best fit of the three models.



| Model           | $LL_X$  | $LL_F$ | $\hat{\pi}_0$ | $\hat{\omega}_1$ | $\hat{\omega}_2$ | $\hat{p}_1$ | $\hat{\delta}$ | $\hat{\pi}_{CW}$ | $\hat{\pi}_{BW}$ |
|-----------------|---------|--------|---------------|------------------|------------------|-------------|----------------|------------------|------------------|
| M3( $k = 2$ )   | -28,818 | -      | -             | 0.03             | 0.22             | 0.55        | -              | -                | -                |
| CLM3( $k = 2$ ) | -28,739 | -      | -             | 0.01             | 0.30             | 0.62        | 0.13           | -                | -                |
| null PG-BSM     | -28,708 | -4     | 0.22          | 0.04             | 0.40             | 0.68        | 0.20           | -                | -                |
| alt PG-BSM      | -28,708 | -4     | 0.22          | 0.04             | 0.40             | 0.68        | 0.20           | 0.00             | 0.00             |

Table 13: Results of the analysis of the phyA&CF data.

|          | YN-BSM A LLR = 22.67 > 3.84 |             |             | YN-BSM B LLR = 23.27 > 5.99 |             |             |
|----------|-----------------------------|-------------|-------------|-----------------------------|-------------|-------------|
| category | proportion                  | BG $\omega$ | FG $\omega$ | proportion                  | BG $\omega$ | FG $\omega$ |
| 0        | 0.81                        | 0.09        | 0.09        | 0.55                        | 0.03        | 0.03        |
| 1        | 0.07                        | 1.00        | 1.00        | 0.40                        | 0.23        | 0.23        |
| 2a       | 0.11                        | 0.09        | 30.17       | 0.03                        | 0.03        | 5.56        |
| 2b       | 0.01                        | 1.00        | 30.17       | 0.02                        | 0.23        | 5.56        |

Table 14: Results of the fit of alternate YN-BSM A and B to the phyA&CF data. LLR is the log-likelihood ratio for the model contrast.

| scenario | % variant | % CW | % BW | PG-BSM | YN-BSM A |
|----------|-----------|------|------|--------|----------|
| 4a       | 40        | 5    | 5    | 50/50  | 50/50    |
| 4b       | 70        | 5    | 5    | 42/50  | 50/50    |
| 4c       | 40        | 0    | 0    | 0/50   | 11/50    |
| 4d       | 70        | 0    | 0    | 0/50   | 31/50    |

Table 15: Counts of the number of times the null was rejected for omnibus tests applied to Simulation 4 alignments. % variant gives the approximate proportion of site patterns in the simulated alignments that exhibited some degree of heterotachy. The columns PG-BSM and YN-BSM A show the number of trials out of 50 for which the relevant model rejected the null hypothesis. Rejections are true for scenarios 4a and 4b and false for scenarios 4c and 4d.

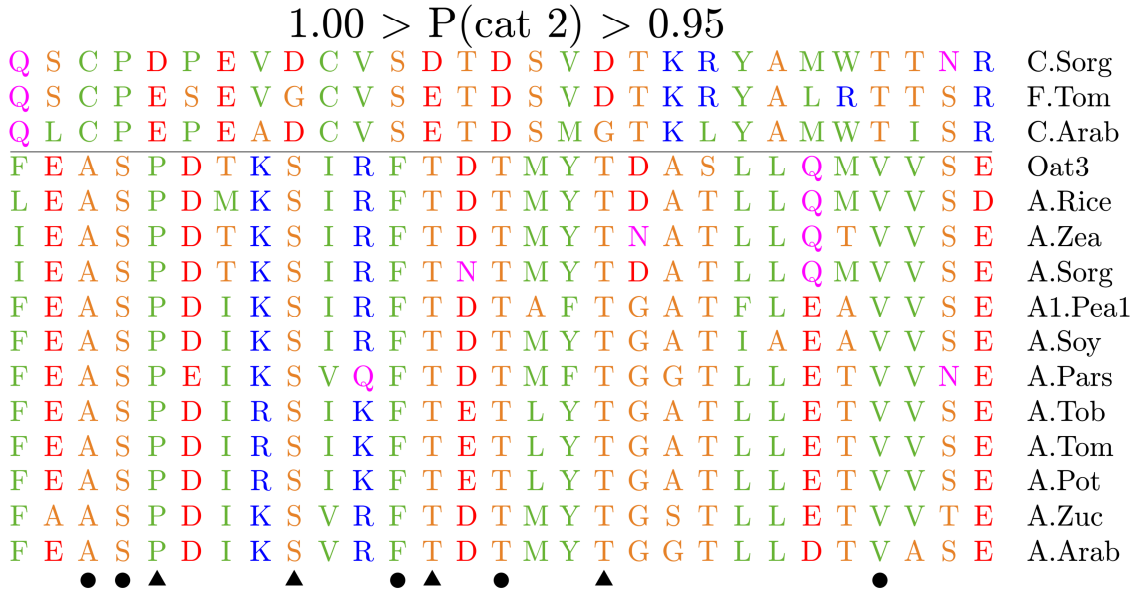


Figure 6: Patterns for sites in the phytochrome alignment inferred to be in category 2a or 2b by the YN-BSM A after application of FDC control with  $E\{\text{FDC}\} = 1$ . Rows designate taxa in the same order as they appear in the tree shown in Figure 5. Columns designate sites. The horizontal line separates phyA from phyCF. Sites are arranged in order of decreasing posterior probabilities estimated using YB-BSM A. Filled triangles and circles mark site patterns most consistent with the CW and BW processes, respectively. Colors indicate physicochemical properties: small nonpolar (orange), hydrophobic (green), polar (magenta), negatively charged (red), positively charged (blue).

## Data Generation using MSmmtDNA and MSTGdR

Simulations under the MS framework were conducted as follows. First, parameters of the mutation process  $M$ , including position-specific nucleotide frequencies and the transition/transversion rate ratio, were estimated from a real alignment consisting of 12 concatenated H-strand mitochondrial DNA sequences (3331 codon sites) from 20 mammalian species as distributed in alignment form by the PAML software package (Yang, 2007). Most alignments were simulated without fixation of double-triple mutations by setting  $(\alpha, \beta) = (0, 0)$ . In cases where alignments were generated with fixation of double-triple mutations, we set  $(\alpha, \beta) = (0.0371, 0.0030)$  so that  $\approx 6\%$  of all mutations would be double ( $\approx 5.8\%$ ) or triple ( $\approx 0.2\%$ ). Recent studies suggest that double-triple mutations comprise between 1% and 3% of all mutations (Keightley *et al.*, 2009; Schrider *et al.*, 2014; De Maio *et al.*, 2013; Harris and Nielsen, 2014). We used a larger value (i.e., 6%) to investigate the impact of double-triple substitutions on model power and accuracy when left unaccounted for by the fitted model. Next, vectors of amino acid fitness coefficients were drawn for each site using either MSmmtDNA (as described in Jones *et al.*, 2018) or MSTGdR (i.e., drawn with replacement from the set of 3598 vectors estimated from real mammalian mtDNA using swMutSel with a Dirichlet-based penalty and with  $\sigma = 0.1$  as described in Tamuri *et al.*, 2014). These were scaled and converted to site-specific substitution rate matrices as follows:

1. A scaling factor  $\sigma^h \sim 0.001 + (0.01 - 0.001) \times B$  was drawn to determine the stringency of selection at the site, where  $B \in [0, 1]$  is a beta random variable with shape parameters  $u, v > 0$ . Values of  $\sigma^h \in [0.001, 0.01]$  closer to the upper bound correspond to greater selection stringency, whereas values closer to the lower bound correspond to a balance between selection and drift that typically results in shifting balance (Jones *et al.*, 2017). Parameters  $u$  and  $v$  for the beta distribution were chosen to make the distributions of scaled selection coefficients  $s_{ij}$  drawn under MSmmtDNA match those reported by Tamuri *et al.* (2012) as closely as possible (Jones *et al.*, 2018) (i.e.,  $u = 0.08$  and  $v = 0.02$ ).
2. A vector  $\mathbf{f}^h$  of fitness coefficients for the 60 codons for the mammalian mitochondrial genetic code was then constructed from the amino acid fitnesses with the assumption that synonymous codons are equally fit. This vector was scaled to make its standard deviation equal to  $\sigma^h$ .
3.  $\mathbf{f}^h = \langle f_1^h, \dots, f_{60}^h \rangle$  for the 60 codons was converted into a matrix  $W^h$  of fixation probabilities computed from scaled selection coefficients  $s_{ij}^h = N_e(f_j^h - f_i^h)$  assuming an effective population size of  $N_e = 1000$  and a ploidy of one for mtDNA:

$$W_{ij}^h \propto \begin{cases} 1 & \text{if } s_{ij}^h = 0 \\ \frac{2s_{ij}^h}{1 - \exp(-2s_{ij}^h)} & \text{otherwise} \end{cases} \quad (12)$$

The corresponding site-specific rate matrix  $A^h = M \circ W^h$  was then constructed by taking the element-wise product of the matrix of mutation rates  $M$  and the matrix of fixation probabilities  $W^h$ .

The expected number of single nucleotide substitutions per codon per unit branch length at each site was then computed as follows:

$$r^h = \sum_{i \neq j} \pi_i A^h(i, j) \{\ell_1 + 2\ell_2 + 3\ell_3\} \quad (13)$$

where the indicator  $\ell_s$  is one if  $i$  and  $j$  differ by  $s \in \{1, 2, 3\}$  nucleotides and zero otherwise. When generating an alignment with no phenotype-genotype association, all rate matrices were divided by  $\bar{r} = (1/n) \sum_{h=1}^n r^h$  to make branch lengths equal to the expected number of single nucleotide substitutions per codon.

A CW shift was implemented by reducing the stringency of selection to  $\sigma^h = 0.0001$  at a subset of sites. Such shifts were made to occur at the ancestral node of the branch over which the phenotype changed and were made to persist along all descendant branches. This was intended to mimic an increase in the replacement rate among a subset of sites over a clade; rCW shifts were similarly implemented but with the stringency of selection increased to  $\sigma^h = 0.01$ . A BW shift was implemented by drawing new vectors of fitness coefficients for a subset of sites to mimic peak shifts. These new vectors were scaled to increase the stringency of selection to  $\sigma^h = 0.01$  and were applied starting at the ancestral node of the branch over which the phenotype changed and along all descendant branches. When CW, rCW and/or BW shifts were implemented, all rate matrices were scaled by dividing by  $\bar{r}_a = (1/n) \sum_{h=1}^n r^h$  with site-specific rates averaged over branches:

$$r^h = \frac{\sum_{b=1}^{2N-2} r^h(b) \mathbf{t}(b)}{\sum_{b=1}^{2N-2} \mathbf{t}(b)} \quad (14)$$

Here  $r^h(b)$  is the rate for the site along branch  $b$  computed using (13) and  $\mathbf{t}(b)$  is the length of that branch.

## Generating Ancestral Phenotypes

Ancestral phenotypic states were sampled using the marginal approach described in Yang (2006) pp 121. Consider the root node of the tree where the state  $x_r$  is unknown. The probability of  $x_r$  conditioned on the vector of phenotypes  $\mathbf{F}$  at the terminal nodes of the tree can be expressed using Bayes' theorem as follows (omitting the rate parameter  $\lambda$  and the vector of branch lengths  $\mathbf{t}$  for brevity):

$$P(x_r = x \mid \mathbf{F}) = \frac{P(\mathbf{F} \mid x_r = x)P(x_r = x)}{P(\mathbf{F})} = \frac{P(\mathbf{F} \mid x_r = x)\pi_F^x}{\sum_{x=1}^3 P(\mathbf{F} \mid x_r = x)\pi_F^x} \quad (15)$$

Here we assume three discrete phenotypes. Note that the conditional probabilities  $P(\mathbf{F} \mid x_r = x)$  are readily computed using the pruning algorithm (Felsenstein, 1981). The vector

$$\langle P(x_r = 1 \mid \mathbf{F}), P(x_r = 2 \mid \mathbf{F}), P(x_r = 3 \mid \mathbf{F}) \rangle$$

can be used to draw a realization of the state at the root node of the tree. This is used in turn to compute a vector of marginal probabilities for the two nodes that descend from the root node. Thus the algorithm moves inductively from the root to the terminal nodes of the tree.

Consider the  $i^{th}$  internal node of the tree. A realization  $x_a$  for the parent of the  $i^{th}$  node will already have been drawn. The conditional probability of the state at the  $i^{th}$  node can be computed as follows:

$$P(x_i = x \mid x_a, \mathbf{F}_i) = \frac{P(\mathbf{F}_i \mid x_i = x)P(x_i = x \mid x_a)}{P(\mathbf{F}_i)} = \frac{P(\mathbf{F}_i \mid x_i = x)P_i(x_a, x)}{\sum_{x=1}^3 P(\mathbf{F}_i \mid x_i = x)P_i(x_a, x)} \quad (16)$$

Here  $\mathbf{F}_i$  denotes the vector of phenotypes at terminal nodes that descend from the  $i^{th}$  node.  $P_i(x_a, x)$  is the element of the transition probability matrix  $P_i = \exp(t_i Q_F)$  corresponding to the  $x_a \rightarrow x$  change of state, where  $t_i$  is the length of the branch connecting the  $i^{th}$  node to its parent. The vector

$$\langle P(x_i = 1 \mid x_a, \mathbf{F}_i), P(x_i = 2 \mid x_a, \mathbf{F}_i), P(x_i = 3 \mid x_a, \mathbf{F}_i) \rangle$$

can be used to draw a realization of the state at the  $i^{th}$  node. The algorithm continues in this way until all internal nodes have been assigned a phenotypic state. The resulting vector of states then constitutes one realization of the evolution of the phenotype conditioned on the values  $\mathbf{F}$  at the terminal nodes of the tree.

## Computing Scaling Constants for the PG-BSM

All rate matrices included in the PG-BSM were constructed as follows for  $k \in \{0, 1, 2\}$ :

$$Q(\omega_k) = M \circ (\ell_S + \omega_k \ell_N) / r_{\omega_k} \quad (17)$$

The value of the common scaling constant was specified so that estimated branch lengths would give the expected number of single nucleotide substitutions per codon. The scaling constant for any individual rate matrix depends only on  $\omega_k$  and can be computed as follows:

$$r_{\omega_k} = \sum_{i \neq j} \pi_i M_{ij} (\ell_S(i, j) + \omega_k \ell_N(i, j)) \{\ell_1 + 2\ell_2 + 3\ell_3\} \quad (18)$$

If  $\pi_0$  is the proportion of sites that evolved under  $Q(\omega_0 = 0)$ ,  $\pi_{CL} = 1 - \pi_0$  the proportion of sites that evolved covarion-like with random switching between  $\omega_1 < \omega_2$ , and  $p_1$  the proportion of time a covarion-like site is expected to spend evolving under  $\omega_1$ , then the scaling constant for the null PG-BSM is:

$$r = \pi_0 r_{\omega_0} + \pi_{CL} r_{CL} = \pi_0 r_{\omega_0} + (1 - \pi_0)(p_1 r_{\omega_1} + (1 - p_1) r_{\omega_2}) \quad (19)$$

Under the alternate PG-BSM accounting for cladewise (CW) and branchwise (BW) sites, the scaling constant is:

$$r = \pi_0 r_{\omega_0} + (1 - \pi_0 - \pi_{CW} - \pi_{BW}) r_{CL} + \pi_{CW} r_{CW} + \pi_{BW} r_{BW} \quad (20)$$

Values for  $r_{CW}$  and  $r_{BW}$  can be calculated using an average that accounts for both the rate ratio at a site on any particular branch weighted by the length of that branch and for the distribution of ancestral histories (i.e., via change maps). For example, the scaling factor for the BW process is:

$$r_{BW} = \sum_{\mathbf{z}} \hat{\pi}_{\mathbf{z}} \frac{\sum_{b=1}^{2N-2} r(z_b) \mathbf{t}(b)}{\sum_{b=1}^{2N-2} \mathbf{t}(b)} \quad (21)$$

where  $r(z_b) = r_{\omega_1}$  if  $z_b = 0$  and  $r_{\omega_2}$  if  $z_b = 1$ . Scaling factors  $r_{CW}$  and  $r_{rCW}$  are similarly computed by taking into account branches over which the site was evolved under  $\omega_1$  or  $\omega_2$ .

## False Discovery Count Control

The algorithm described by Newton *et al.* (2004) was applied to naive empirical Bayesian posteriors to control the false discovery count (FDC) for our *post hoc* analyses. We aimed to identify as many sites with evidence of phenotype-genotype association as possible while controlling the number of false discoveries to some specified value. The procedure was as follows, here described for BW sites:

1.  $p^h = 1 - \text{P}(\text{BW} \mid x^h)$  (cf. equation 7) approximates the conditional probability that assigning the  $h^{\text{th}}$  site to the BW category is a type I error.
2. Let  $p^1, \dots, p^n$  be a list of these probabilities for all  $n$  sites.
3. For any specified bound  $\kappa$ , assign a site to the BW category if  $p^h \leq \kappa$ .
4. By this rule, the expected number of false discoveries given the data is (Newton *et al.*, 2004):

$$\text{E}\{\text{FDC}\} = \sum_{h=1}^n p^h \ell(p^h \leq \kappa) \quad (22)$$

where  $\ell(p^h \leq \kappa)$  is 1 if  $p^h \leq \kappa$  and zero otherwise.

5. To control the false discovery count to be no more than  $k \in \{1, 2\}$ ,  $\kappa$  can be set to the largest value for which  $\text{E}\{\text{FDC}\} \leq k$ . Note that this expectation is across data sets, and is only approximate because it depends on how well the fitted model matches the data-generating process (Newton *et al.*, 2004).

## Dealing with Underflow

Likelihood functions are typically optimized in log-space. For example, the log-likelihood for the alternate component of PG-BSM is:

$$\ln \{L_{\text{alt}}(X, \mathbf{F}; \lambda, \boldsymbol{\theta}, \mathbf{t})\} = \ln \{P(\mathbf{F}; \lambda, \mathbf{t})\} + \ln \left\{ \sum_{\mathbf{z}} \hat{\pi}_{\mathbf{z}} \prod_{h=1}^n g(x^h; \mathbf{z}) \right\} \quad (23)$$

It is convenient to express the second addend of (23) as follows:

$$\ln \left\{ \sum_{\mathbf{z}} \hat{\pi}_{\mathbf{z}} \prod_{h=1}^n g(x^h; \mathbf{z}) \right\} = \ln \left\{ \sum_{\mathbf{z}} \hat{\pi}_{\mathbf{z}} \exp(\ell_{\mathbf{z}}) \right\} \quad (24)$$

$$\text{where } \ell_{\mathbf{z}} = \ln \left\{ \prod_{h=1}^n g(x^h; \mathbf{z}) \right\} \quad (25)$$

A problem arises when the probability  $\exp(\ell_{\mathbf{z}})$  is too small to be represented digitally, an issue commonly referred to as underflow. This can be mitigated by a simple transformation:

$$\ln \left\{ \sum_{\mathbf{z}} \hat{\pi}_{\mathbf{z}} \prod_{h=1}^n g(x^h; \mathbf{z}) \right\} = \max\{\ell_{\mathbf{z}}\} + \ln \left\{ \sum_{\mathbf{z}} \hat{\pi}_{\mathbf{z}} \exp(\ell_{\mathbf{z}} - \max\{\ell_{\mathbf{z}}\}) \right\} \quad (26)$$

Since  $\forall \mathbf{z} \ell_{\mathbf{z}} < 0$ , the transformation  $\ell_{\mathbf{z}} - \max\{\ell_{\mathbf{z}}\}$  moves the log-probabilities to the right toward zero (i.e., toward larger values), making underflow less likely.

The transformation method can also be used to avoid underflow in equation (9). Let  $\mathbf{z}^*$  represent the change map that maximizes  $\ell_{\mathbf{z}}$  (i.e. maximizes the likelihood of the alignment). The natural log of (9) at  $\mathbf{z}^*$  is:

$$\ln \{P(\mathbf{z}^* | X, \mathbf{F})\} = \ln \{L_{\text{alt}}(X, \mathbf{F} | \mathbf{z}^*)\} + \ln \{\hat{\pi}_{\mathbf{z}^*}\} - \ln \{L_{\text{alt}}(X, \mathbf{F})\} \quad (27)$$

Applying the transformation:

$$\ln \{L_{\text{alt}}(X, \mathbf{F})\} = \ln \{P(\mathbf{F})\} + \ell_{\mathbf{z}^*} + \ln \left\{ \sum_{\mathbf{z}} \hat{\pi}_{\mathbf{z}} \exp(\ell_{\mathbf{z}} - \ell_{\mathbf{z}^*}) \right\} \quad (28)$$

$$= \ln \{L_{\text{alt}}(X, \mathbf{F} | \mathbf{z}^*)\} + \ln \left\{ \sum_{\mathbf{z}} \hat{\pi}_{\mathbf{z}} \exp(\ell_{\mathbf{z}} - \ell_{\mathbf{z}^*}) \right\} \quad (29)$$

Substituting into (27) gives:

$$P(\mathbf{z}^* | X, \mathbf{F}) = \frac{\hat{\pi}_{\mathbf{z}^*}}{\hat{\pi}_{\mathbf{z}^*} + \sum_{\mathbf{z} \neq \mathbf{z}^*} \hat{\pi}_{\mathbf{z}} \exp(\ell_{\mathbf{z}} - \ell_{\mathbf{z}^*})} \quad (30)$$

Equation (30) demonstrates that the strength of the evidence for the ancestral reconstruction corresponding to  $\mathbf{z}^*$  is a function of the relative frequency of the most frequently sampled change map  $\hat{\pi}_{\mathbf{z}^*}$  and the differences  $\ell_{\mathbf{z}} - \ell_{\mathbf{z}^*}$  in log-likelihoods of the alignment under the various other  $\mathbf{z}$ . When evidence for the PG processes dictated by  $\mathbf{z}^*$  is strong, it will be the case  $\forall \mathbf{z} \neq \mathbf{z}^*$  that  $\exp\{\ell_{\mathbf{z}} - \ell_{\mathbf{z}^*}\} \approx 0$  making  $P(\mathbf{z}^* | X, \mathbf{F}) \approx 1$ .



# References

- De Maio, N., Holmes, I., Schlötterer, C., and Kosiol, C. 2013. Estimating empirical codon hidden Markov models. *Mol. Biol. Evol.*, 30: 725–736.
- Felsenstein, J. J. 1981. Evolutionary trees from dna sequences: a maximum likelihood approach. *J. Mol. Evol.*, 17: 368–376.
- Gu, X. 2006. A simple statistical model for estimating type-ii (cluster-specific) functional divergence of protein sequences. *Mol. Biol. Evol.*, 23: 1937–1945.
- Harris, K. and Nielsen, R. 2014. Error-prone polymerase activity causes multinucleotide mutations in humans. *Genome Research*, 9: 1445–1554.
- Jones, C. T., Youssef, N., Susko, E., and Bielawski, J. P. 2017. Shifting balance on a static mutation-selection landscape: a novel scenario of positive selection. *Mol. Biol. Evol.*, 34: 391–407.
- Jones, C. T., Youssef, N., Susko, E., and Bielawski, J. P. 2018. Phenomenological load on model parameters can lead to false biological conclusions. *Mol. Biol. Evol.*, 35: 1473–1488.
- Keightley, P., Trivedi, U., Thomson, M., Oliver, F., Kumar, S., and Blaxter, M. 2009. Analysis of the genome sequences of three *Drosophila melanogaster* spontaneous mutation accumulation lines. *Genet. Res.*, 19: 1195–1201.
- Mathews, S. 2010. Evolutionary studies illuminate the structural-functional model of plant phytochromes. *The Plant Cell*, 22: 4–16.
- Newton, M. A., Noueiry, A., Sarkar, D., and Ahlquist, P. 2004. Detecting differential gene expression with a semiparametric hierarchical mixture model. *Biostatistics*, 5: 155–176.
- Pupko, T. and Galtier, N. 2002. A covarion-based method for detecting molecular adaptation: application to the evolution of primate mitochondrial genomes. *Proc. R. Soc. Lond.*, 269: 1313–1316.
- Schrider, D., Hourmozdi, J., and Hahn, M. 2014. Pervasive multinucleotide mutational events in eukaryotes. *Curr. Biol.*, 21: 1051–1054.
- Tamuri, A. U., dos Reis, M., and Goldstein, R. A. 2012. Estimating the distribution of selection coefficients from phylogenetic data using sitewise mutation-selection models. *Genetics*, 190: 1101–1115.
- Tamuri, A. U., Goldman, N., and dos Reis, M. 2014. A penalized-likelihood method to estimate the distribution of selection coefficients from phylogenetic data. *Genetics*, 197: 257–271.

- Yang, Z. H. 2006. *Computational Molecular Evolution*. Oxford University Press, Oxford.
- Yang, Z. H. 2007. PAML4: Phylogentic analysis by maximum likelihood. *Mol. Biol. Evol.*, 24: 1586–1591.
- Yang, Z. H. and Nielsen, R. 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol. Biol. Evol.*, 19: 908–917.
- Yang, Z. H., Nielsen, R., Goldman, N., and Pedersen, A. M. K. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics*, 155: 431–449.
- Zhang, J., Nielsen, R., and Yang, Z. H. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol. Biol. Evol.*, 22: 2472–2479.