Supporting Material

Methods

Thinning

Proposal acceptance in Bayesian phylogenetic analysis using MCMC depends on the previous state of the chain, thus a certain amount of inter-generational autocorrelation is expected for all parameters being sampled (1), including tree topologies. Because of auto-correlation, the number of truly independent samples in the posterior is smaller than the total number of sampled generations. Integrated auto-correlated time (IACT) is the generation lag that must be exceeded before significance in temporal autocorrelation breaks down (2). For instance, if the posterior has 20,000 trees (after burn-in) and IACT equals 100, a set of 200 independently sampled trees is obtained by sub-sampling every 100th generation. These 200 trees represent the effective sample size (ESS). Because both distributions have identical ESS's, the sub-sampled set provides an equivalent approximation of the posterior, even though it is 100 times smaller than the original sample size (2). This sub-sampling strategy is called "thinning" (3-4).

Log-likelihood IACT was obtained from MrBayes parameter files (.p) using Tracer v1.3 (5). We used a Perl script (*ThinLog*) to sub-sample parameter files and the resulting distribution was re-analyzed to verify if log-likelihood IACT had been effectively reset to 1 (i.e., no auto-correlation). *ThinLog* was then used to sub-sample tree files (.t) and to combine resulting trees in a single file from which a new consensus tree was computed using MrBayes. The resulting node posterior probabilities were compared to the ones obtained from the original (non-thinned) tree distribution to check for changes in node support across the tree. Thinning reduced the posterior distribution from 28,874 to 312 trees. No nodes with posterior probability above 0.95 were "lost" in the consensus computed from the sub-sampled tree set.

Bayesian ancestral state reconstruction

The statistical framework of maximum-likelihood (6) was used to describe evolution of coloniality in corals. The method is formally defined by an instantaneous rate matrix Q (in this case for 2 states):

$$Q = \begin{bmatrix} \cdots & q_{01} \\ q_{10} & \cdots \end{bmatrix} \quad (1)$$

where q_{01} is the instantaneous rate of change from states 0 (solitary or azooxanthellate, depending on the character considered) to 1 (colonial or zooxanthellate) and q_{10} represents transitions in the opposite direction. The corresponding probabilities are obtained by exponentiating Q after multiplying it by a given branch length. The likelihood of each set of rates is obtained by combining all probabilities across all branches and all possible ancestral states of the phylogeny. The combination of coefficients that maximizes the likelihood function is the maximum-likelihood (ML) solution. One can subsequently attempt to simplify the model by constraining one or more rate coefficients to be equal and comparing the resulting ML values to the ones obtained under the original model via a likelihood-ratio (LR) test. If certain transitions occur very infrequently, one might also attempt to set the corresponding rate coefficients to zero. For more details on the method, see (6-7).

The ML approach has the drawback of not accounting for phylogenetic uncertainty since rates are evaluated on a single tree. Even if ML solutions are computed across a range of trees, there is no unique way of summarizing the corresponding distribution of maximum-likelihoods before performing LR tests. A suitable alternative is applying a Bayesian framework to the problem by sampling trees from a posterior distribution obtained via Markov chain Monte Carlo

(MCMC) and using these topologies to construct a second chain to generate a posterior distribution of rate coefficients. This method can also be used to estimate the Bayesian posterior probability (BPP) of a state on a certain node while factoring phylogenetic uncertainty into this computation. Hence, if the BPP of a node is 0.95, the maximum posterior probability for the reconstruction of an ancestral state at that node will also be 0.95. In other words, the BPP of the state at the node cannot exceed the BPP of the node itself (3).

A variant of this Bayesian approach employs reversible-jump Markov chain Monte Carlo (RJ-MCMC) (8). At each step, RJ-MCMC not only proposes new rate coefficient values but it also attempts to modify the number of parameters in the model by constraining some of them to be equal (merging), removing this constraint (splitting), setting one or more rates to zero (reducing) or reversing this move (augmenting). Hence, RJ-MCMC also provides a posterior distribution of models. The significance of the Bayesian belief in a particular model can be formally tested using Bayesian factors which are defined as:

$$BF_{ij} = \frac{P(M_i \mid D)}{P(M_j \mid D)} \times \frac{P(M_j)}{P(M_i)} \qquad (2)$$

where i and j denote two different models, $P(M_n \mid D)$is the posterior probability of model n, given by the proportion of time the chain spent on this model and $P(M_n)$ is its prior probability. The prior probability is calculated as the fraction of the total number of models. For two states, there are 4 possibilities: the one-rate model (i.e. $q_{01} = q_{10}$, transitions in either direction are equally likely), the two-rates model ($q_{01} \neq q_{10}$) and two additional models in which one of the coefficients is set to 0. Given that Bayes factor BF_{ij}, values in the 3-12 range are positive evidence for model i, values above 12 represent strong evidence and values greater than 150 provide very strong evidence in favor of model i. (9)

The RJ-MCMC approach also allows the testing of correlated evolution between binary characters. The assumption is that if characters are evolving independently, the instantaneous rates of transition from one character state to another should be independent of the background state of the other character. Using the same framework described above

$$Q = \begin{array}{c|cccc|} & 0,0 & 0,1 & 1,0 & 1,1 \\ 0,0 & \dots & q_{12} & q_{13} & 0 \\ 0,1 & q_{21} & \dots & 0 & q_{24} \\ 1,0 & q_{31} & 0 & \dots & \mathrm{q}_{34} \\ 1,1 & 0 & q_{42} & q_{43} & \dots \end{array} \tag{3}$$

where the states for both characters were coded using the same convention above (i.e. 0,0 stands for solitary/azooxanthellate, 0,1 for solitary/zooxanthellate, etc.). Under this framework, instantaneous state changes in both characters are not allowed (e.g. q_{14}, the rate that would correspond to a solitary/azooxanthellate to colonial/zooxanthellate transition is set to 0) although these transitions may occur the longer run.

Under the independent model, the probability of going from solitary to colonial should be the same whether it occurs in azooxanthellate or zooxanthellate ancestor-descendant pairs, hence $q_{13} = q_{24}$. Likewise, the rates of loss of symbiosis should be the same whether coloniality is present or not ($q_{21} = q_{43}$). There are 4 such pairs of rates hence the most complex model model is reduced to four parameters

$$Q = \begin{array}{c|cccc|} & 0,0 & 0,1 & 1,0 & 1,1 \\ 0,0 & \ldots & \alpha_1 & \alpha_2 & 0 \\ 0,1 & \beta_1 & \ldots & 0 & \alpha_2 \\ 1,0 & \beta_2 & 0 & \ldots & \alpha_1 \\ 1,1 & 0 & \beta_2 & \beta_1 & \ldots \end{array} \qquad (8)$$

where α is the transition rate between the states 0 and 1 and β is the rate of transition from 1 to 0. The subscript of these rates denotes the character they refer to (1 for symbiosis and 2 for coloniality). Note that the number of parameters in the model can be reduced even further by constraining these pairs to equal each other or setting them to 0. Hence, if $\beta_1 = \beta_2$ or if $\alpha_1 = \beta_2 = 0$, the model is still independent. The total number of possible independent models is 47. If rates within any of these pairs are different or if one of the rates is set to 0, the model becomes dependent. There are 20,001 possible dependent models with 8 rates (see Methods section in Supplemental Material for details on the calculation of number of models).

One can test for correlated evolution of characters by running separate analysis under independent and dependent models and comparing the results. In a likelihood framework, the ML solution would be estimated under the full independent and dependent models and since (8) is a subset of (3), the LR statistic is χ^2 distributed with 4 degrees of freedom (the difference in the number of parameters between the two models). In "discrete" mode, the RJ-MCMC chains are constructed so that the sampling is biased towards the models of choice (independent or dependent) although there is a small probability that they may sample the alternative model. Hence, if there is strong Bayesian belief that the evolution of characters is uncorrelated but the chains are run under the dependent model, the likelihood values sampled from the posterior will be smaller than those obtained by the "independent" run. Significance can be assessed by Bayes factors as the difference between the marginal log-likelihood obtained from each run. Exact computation of marginal log-likelihoods is difficult (8) but they can be unbiasedly estimated by the harmonic mean of the running log-likelihoods(10). Bayes factors are then given by

$$BF_{DI} = 2(H_D - \ H_I) \tag{9}$$

where H_D and H_I are the harmonic means of the log-likelihoods sampled under the dependent and independent models. Harmonic means almost surely converge when the number of generations tends to infinity, but are unstable because of their sensitivity to sporadic sampling of extreme log-likelihood values (10) hence it is advisable to run a number of different chains to assess the robustness of the results (8).

Calculation of prior distributions of models

Under the most complex model, all rates are estimated separately, hence $q_{12} \neq q_{13} \neq q_{21} \neq q_{24} \neq q_{31} \neq q_{34} \neq q_{42} \neq q_{43}$. However, as explained above, one can reduce the number of parameters in the model by constraining some of the rates to be equal. In other words, these rates are placed in the same rate category. The most complex model has 8 rates and 8 categories. A convenient notation would be (0, 1, 2, 3, 4, 5, 6, 7) where each algorism corresponds to a category and their position in the set corresponds to the rates arranged as (q_{12}, q_{13}, q_{21}, q_{24}, q_{31}, q_{34}, q_{42}, q_{43}). If any 2 rates are constrained to be equal, there are 8 rates and 7 categories. The models ($\underline{0}$, $\underline{0}$, 1, 2, 3, 4, 5, 6) and (1, $\underline{0}$, $\underline{0}$, 2, 3, 4, $\underline{0}$, $\underline{0}$) are 2 possibilities (we designate the first category as 0 instead of 1 in accordance with the notation used in Bayes Traits). In the first model $q_{12} = q_{13}$ (hence the first two rates are assigned to category 0) and $q_{13} = q_{21} = q_{42} = q_{43}$ in the second. The total number of models is given by the Stirling number of the second kind (8), which counts the way one can arrange *n* elements in *c* partitions:

$$S_2(n,c) = \frac{1}{c!}\sum_{i=0}^{c-1}(-1)^{\mathrm{i}}\binom{c}{i}(\mathrm{c}-\mathrm{i})^{\mathrm{n}} \qquad (2)$$

The total number of models is given by n$^{\text{th}}$ Bell number

$$B_n = \sum_{i=1}^{c} S_2(n,i) \qquad (3)$$

which is 4,140 for 8 rates. The full set of Stirling numbers make up the prior distribution of models and for 8 rates we have S_2(8,1) = 1, S_2 (8,2) = 127, S_2 (8,3) = 966, S_2 (8,4) = 1,701, S_2 (8,5) = 1,050, S_2(8,6) = 266, S_2(8,7) = 28, and S_2 (8,8) = 1 (8).

Because categories can be also set to zero, the total number of models is much larger. For instance, (Z, 0, 1, 2, 3, 4, 5, 6) would be a 8 rate/8 parameter model in which rate q_{12} was set to zero (the use of Z follows the notation adopted in Bayes Traits to prevent confusion with the algorism 0, which designates a non-Z category).

We can build 8 additional 8 category models by constraining each of the other categories to zero, i.e. (0, Z, 1, 2, 3, 4, 5, 6), (0, 1, Z, 2, 3, 4, 5, 6), etc. We can only assign one category at a time to zero without reducing the total number of categories. For instance, if we assign 2 categories to zero, we now have a 7 category model, since these 2 rates now to belong to the "Z" category. The total number of models for *n* rates and *c* categories is then given by the "expanded" Stirling number (the name was introduced in (8))

$$E_2(n,c) = S_2(n,c)(1+c) \qquad (4)$$

and the total number of models is given by the summation

$$\sum_{i=1}^{c} E_2(n,i) = B_{n+1} \qquad (5)$$

which is 21,147 in the case of 8 rates. The value reported in (8) is actually 21,146. This is because if the one-category model is set to zero, i.e (Z, Z, Z, Z, Z, Z, Z, Z), this means no character evolution, so this model should not be considered part of the prior. Hence, the *M* number of non-trivial models for *n* rates is

$$M_n = B_{n+1} - 1 \tag{6}$$

Let us now consider the 4 parameter *Q* matrix that corresponds to the independent model, described in (3). Because 4 pairs are constrained to the same rate, the total number of models is $M_4 = 51$. However, models where $\alpha_1 = \beta_1 = \mathrm{Z}$ or $\alpha_2 = \beta_2 = \mathrm{Z}$ are also trivial, because under these models, no state transitions ever occur for the constrained characters. Using the notation described above and arranging the rates as (α_1, β_1, α_2, β_2) these models are (Z, Z, 0, 0), (0, 0, Z, Z), (Z, Z, 0, 1), etc. If we constrain all the rates of a character 1 to zero, we are left with two unconstrained rates for character 2 and vice-versa. So the total number of *N* non-trivial models for a 4 rate matrix is

$$N_{(4,2)} = M_4 - 2M_2 \tag{8}$$

or 43 models. This formula can be generalized to

$$N_{(f,r)} = M_f - 2M_{f-r} \tag{9}$$

Where *f* is the number of free parameters and *r* is the number of parameters that need to be constrained in order to make the model trivial. In the case of the 8-parameter model in (1):

$$N_{(8,4)} = M_8 - 2M_4 \tag{10}$$

Hence, there are 21,044 non-trivial models under the prior distribution, but since 43 of these models are *independent*, the number of non-trivial *dependent* models under the prior is given by

$$D = N_{(8,4)} - N_{(4,2)} \tag{10}$$

which is 21,044 - 43 = 21,001.

Hypotheses testing

One can use this framework to compute the prior probabilities of models used in the Bayes factors

$$BF_{ij} = \frac{P(M_i \mid D)}{P(M_j \mid D)} \times \frac{P(M_j)}{P(M_i)} \tag{11}$$

Bayes factors can be used to address very specific hypotheses. For instance, one may wish to verify if models in which one rate is set to Z are sampled from the posterior more frequently than expected under the prior. The number I of non-trivial independent models for 8 rates in which one specific rate was set to Z is given by

$$I = M_3 - M_2 - M_1 \tag{13}$$

or 14 – 4 – 1 = 9. The difference with respect to equation (8) arises because of the following. Let's assume that the rate α_1 in equation (7) is set to Z. Since one category has been assigned to Z, there are 3 free parameters in the model. There are M_2 trivial models in which $\alpha_1 = \beta_1$ (i.e. when all the rates in character 1 are set to Z). However, there's just one non-trivial independent model when $\alpha_2 = \beta_2$. This is because α_1 is already constrained to Z, so this leaves only one other possible model, $\beta_1 \neq$ Z.

Following the same reasoning, under the 8-parameter prior, the total number of non-trivial dependent models with one specific rate assigned to Z is given by

$$D = M_7 - M_4 - M_3 - I \tag{14}$$

or 4139 – 51 – 14 – 9 = 4,065. This corresponds to approximately 20% of all the 20,001 non-trivial dependent models under the prior.

One may also be interested in testing if two specific rates are placed in the same (non-Z) category more often than the prior expectation. Under the independent prior, when 2 rates are constrained to the same category, we are left with 2 free parameters. If the constrained rates refer to the same character (i.e. $\alpha_1 = \beta_1$ or $\alpha_2 = \beta_2$), the number of independent models is $M_2 = 4$. However, if the rates constrained not to be Z refer to different models (i.e. $\alpha_1 = \alpha_2$, $\alpha_1 = \beta_2$, $\alpha_2 = \beta_1$ or $\beta_1 = \beta_2$) then the number of non-trivial independent models (I) is $B_3 = 5$ (recall from equation 5 that the summation over all c categories of the expanded Stirling number $E_2(n,c)$ is B_{n+1}). This is because even if all free rates are assigned to Z, the model will not be trivial since one rate from each character was constrained to be non-Z. Likewise, under the dependent model, if we assign two rates to the same category, we are left with 6 free categories. If these rates refer to transitions occurring in different characters, all non-trivial dependent models are given by

$$D = B_7 - I = B_7 - B_3 \tag{15}$$

or 877 – 5 = 872. These models make up 4.36% of the prior.

If the rates refer to the same character, the number of non-trivial dependent models is given by

$$D = B_7 - M_2 - I = B_7 - M_2 - M_2 \tag{16}$$

and 877 – 30 – 30 = 817. The first term refers to the total number of models with 6 rates. The second term refers to the M_2 trivial models, obtained when all of the unconstrained character's rates are assigned to the Z category. The proportion of such prior models is slightly smaller in this case (4.08%).

References

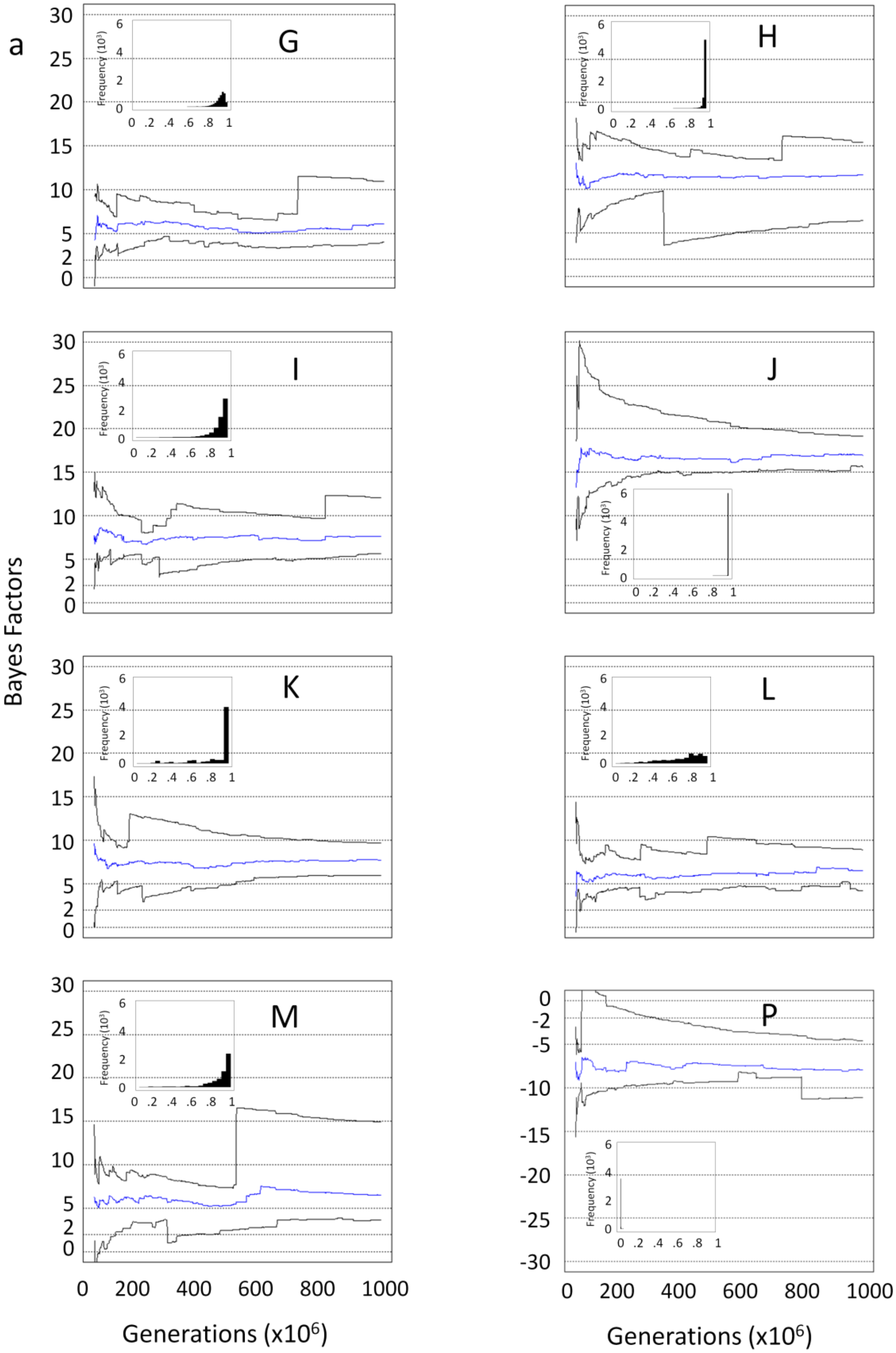
1. Geyer CJ (1992) Practical Markov Chain Monte Carlo. *Statistical Science* 7(4):473-483.
2. Drummond AJ, Nicholls GK, Rodrigo AG, & Solomon W (2002) Estimating Mutation Parameters, Population History and Genealogy Simultaneously From Temporally Spaced Sequence Data. *Genetics* 161(3):1307-1320.
3. Pagel M, Meade A, & Barker D (2004) Bayesian Estimation of Ancestral Character States on Phylogenies. *Syst. Biol.* 53(5):673–684.
4. Buschbom J & Barker D (2006) Evolutionary History of Vegetative Reproduction in *Porpidia s.l.* (Lichen-Forming Ascomycota). *Syst. Biol.* 55(3):471-484.
5. Rambaut A & Drummond AJ (2003) TracerEdinburgh, 1.3.

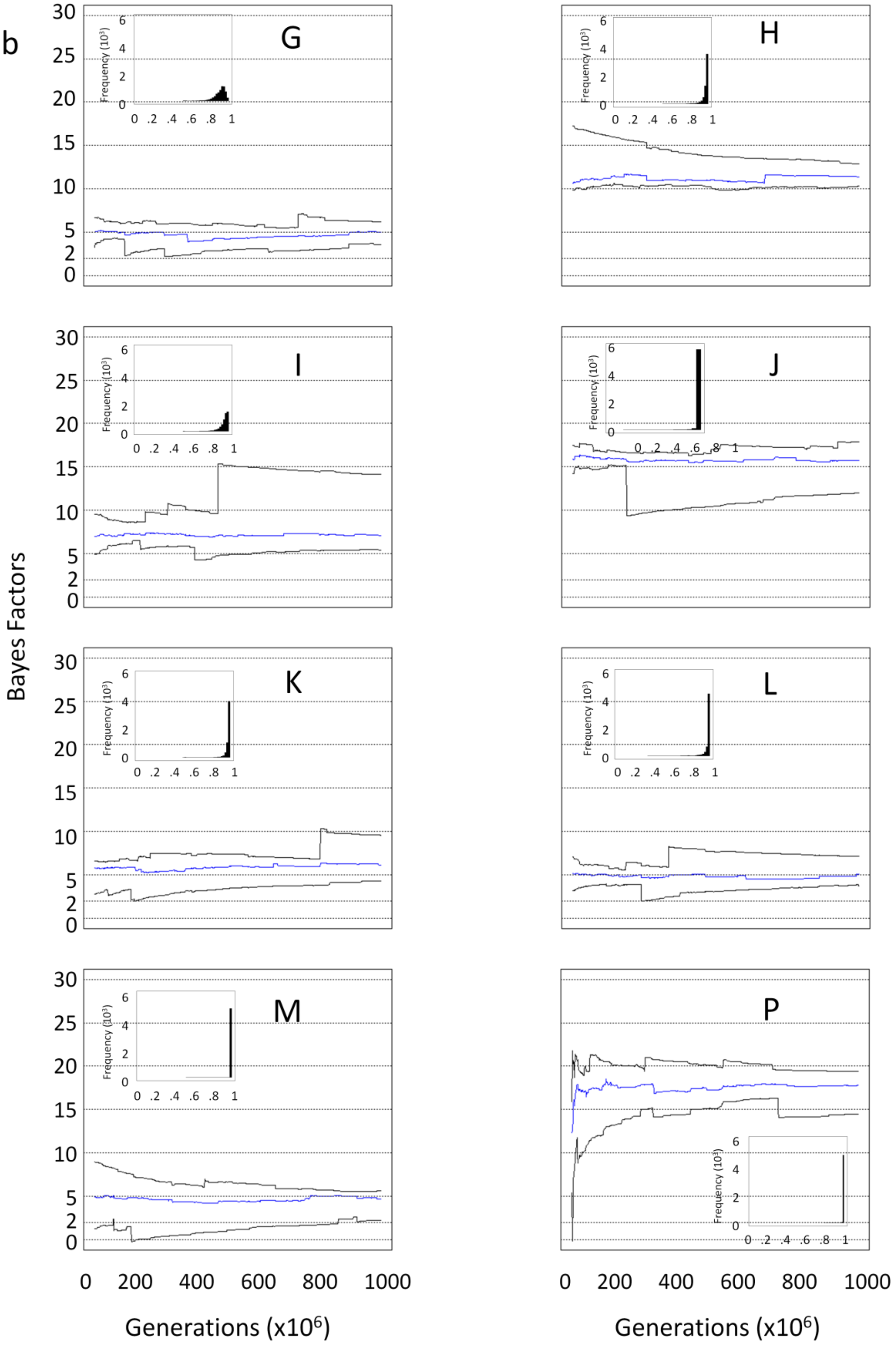
6. Pagel M (1999) The maximum likelihood approach to reconstructing ancestral character states of discrete characters on phylogenies. *Syst. Biol.* 48:612-622.
7. Pagel M (1994) Detecting Correlated Evolution on Phylogenies: A General Method for the Comparative Analysis of Discrete Characters. *Proceedings: Biological Sciences* 255(1342):37-45.
8. Pagel M & Meade A (2006) Bayesian Analysis of Correlated Evolution of Discrete Characters by Reversible-Jump Markov Chain Monte Carlo. *Am Nat* 167(6):808-825.
9. Raftery AE (1996) Hypothesis testing and model selection. *Markov chain Monte Carlo in practice,* eds Gilks WR, Richardson S, & Spiegelhalter DJ (Chapman & Hall, London), pp 163-188.
10. Newton MA & Raftery AE (1994) Approximate Bayesian Inference with the Weighted Likelihood Bootstrap. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 56(1):3-48.

Figure legends

Fig. S1. Running Bayes Factors for (a) symbiosis and (b) coloniality. Median is represented by the blue line. Posterior probability distribution in of symbiotic and colonial states are in the insets. Note that in the case of node P, the probability symbiosis is close to zero. Hence, Bayes factors in favor of symbiosis are negative (note difference in scale of the y axis).

Fig.S1


a
Bayes Factors
G
H
I
J
K
L
M
P
30
25
20
15
10
5
2
0
-2
-5
-10
-15
-20
-25
-30
Frequency (10³)
6
4
2
0
0 .2 .4 .6 .8 1
0 200 400 600 800 1000
Generations (x10⁶)


b
G
H
I
J
K
L
M
P
Bayes Factors
30
25
20
15
10
5
2
0
Frequency (10³)
6
4
2
0
0 .2 .4 .6 .8 1
0
200
400
600
800
1000
Generations (x10⁶)

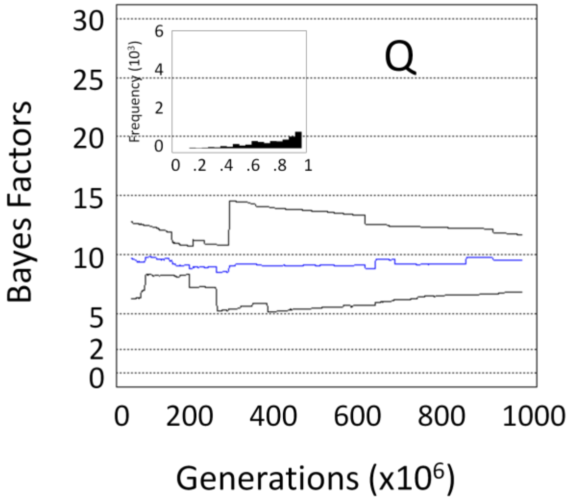

Q
Bayes Factors
30
25
20
15
10
5
2
0
0
200
400
600
800
1000
Generations (x10^6)
Frequency (10^3)
6
4
2
0
0
.2
.4
.6
.8
1

Fig. S1 (Cont.)