Reanalysis of Sanchez-Baracaldo et al. 2017

Teofil Nakov, James Boyko, Andrew Alverson, Jeremy Beaulieu

2017-09-21

Contents

Data files 1
Packages 1
Analyses 1
SIMMAP with three-state coding (freshwater, marine, and brackish) 1
SIMMAP: Equal rates model 1
SIMMAP: All rates different model 2
corHMM with three-state coding 4
SIMMAP with two-state coding (marine vs. freshwater) 5
SIMMAP: Equal rates model 5
SIMMAP: All rates different model 6
corHMM with two-state coding 7
Model-averaging 7

Data files

We use two datasets provided in the data package of Sanches-Baracaldo et al. 2017 available on DataDryad. The first one is an `XML` file used as input to `SIMMAP`, the second one is a `nexus` file. The habitat data from these files were parsed in `bash` before loading into R for ancestral state reconstruction.

There are three differences between the two dataset. Two species coded with "2" corresponding to Brackish habitat in the `XML` file are coded as "1" or Marine in the `nexus` daatset, and one species coded as marine ("1") in the `nexus` dataset is coded as freshwater ("0") in the `XML` dataset.

```
# species              state_in_Nexus state_in_XML
synechococcus_pcc7002        1             2
nostoc_pcc7120               1             0
cyanobium_pcc7001            1             2
```

Packages

To reconstruct ancestral states, we use two packages: `phytools` for stochastic mapping, and `corHMM` for maximum likelihood reconstructions. The package versions are below, see also the `sessionInfo()` output at the bottom of this document.

Analyses

SIMMAP with three-state coding (freshwater, marine, and brackish)

SIMMAP: Equal rates model

We first performed stochastic character mappping (SIMMAP) assuming *equal rates* of transition between marine, brackish, and fresh waters. The prior for the root is set to `equal` to match the methods in the original paper. This means that all states have the same prior probability at the root. We simulated 1000 stochastic maps and summarized the output with functions from the `phytools` package.

The nodes relevant for the reconstruction of the ecology of the primary chloroplast endosymbiosis are the root, the most-recent common ancestor (MRCA) of the cyanobacerium *Gloeomargarita* and Archaeplastida (glaucophytes, red 'algae' and green 'algae' + land plants) and the MRCA of Archaeplastida themselves. The probabilities for different ancestral habitats under the *equal rates* model (summarized from 1000 stochastic maps) are shown in Table 1.

Table 1: Probabilities for Freshwater, Marine, and Brackish ancestry for three nodes on the Cyanobacteria + Archaeplastida phylogeny reconstructed with the equal-rates model.

	P.freshwater.	P.marine.	P.brackish.
Root	0.801	0.170	0.029
MRCA Gloeomargarita + Archaeplastida	0.816	0.179	0.005
MRCA Archaeplastida	0.818	0.180	0.002

These results are not identical to the results from Sanchez-Baracaldo et al. 2017, but nonetheless confirm their major finding: under the assumption that all possible transitions between freshwater, brackish, and marine habitats happen at the same rate, the most likely habitat for the MRCAs of Cyanobacteria and Archaeplastida was freshwater [<0.5 parts per thousand (ppt) salinity].

SIMMAP: All rates different model

It is possible, however, that transitions between these three habitats happen at different rates, perhaps as a result of cellular physiological constraints. Marine and fresh waters differ drastically in many important ways, including concentrations of salts and other ions, osmotic pressure, nutrient regimes, and so on. It is therefore possible that certain types of transitions occur more frequently than others. For example, transitions between brackish (0.5-30 ppt) and marine habitats (> 30 ppt as coded in Sanchez-Baracaldo et al. 2017), might be more frequent over evolutionary time than direct transitions between freshwater and marine environments. It is also possible that such direct marine-to-freshwater (or the reverse) transition are highly unlikely or impossible.

We therefore examined a model that allowed these transitions to vary. In this model, the different transition rates are independently estimated, and importantly, if two evolutionary transitions indeed have similar rates, this model is able to detect that and return similar parameter estimates. We simulated 1000 stochastic character histories using a model with *unequal rates* (all-rates-different), and as before, assumed that the prior state at the root can be either of the states observed at the tips with equal probability.

As before, we summarized the probabilities for different ancestral states at the three relevant nodes. We found that the most likely ancestral habitat, for each of the relevant nodes, was now reversed. After we accounted for different rates of transition between states, the probability for the freshwater state went down and probability for marine or brackish ancestry went up (Table 2).

Table 2: Probabilities for Freshwater, Marine, and Brackish ancestry for three nodes on the Cyanobacteria + Archaeplastida phylogeny reconstructed with the all-rates-different model.

	P.freshwater.	P.marine.	P.brackish.
Root	0.063	0.542	0.395
MRCA Gloeomargarita + Archaeplastida	0.038	0.926	0.036
MRCA Archaeplastida	0.047	0.921	0.032

These differences are shown in Figure 1, where the ancestral state reconstructions are plotted side-by-side.

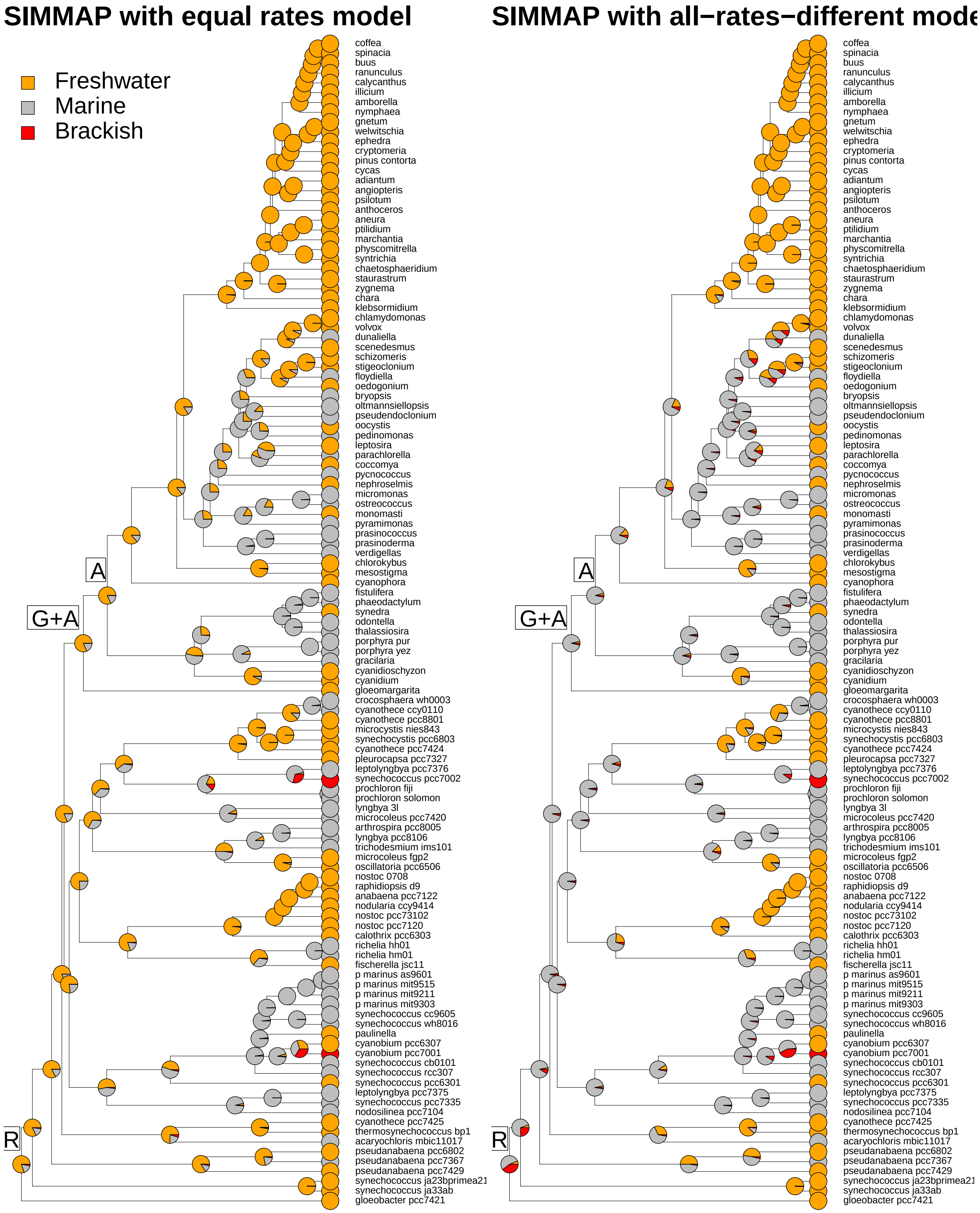

Figure 1: Comparison of ancestral states for habitat (multi-state: marine, freshwater, brackish) reconstructed with equal-rates (ER, left) and all-rates-different (ARD, right) model. The relevant nodes are labeled: R=root, G+A=Gloeomargarita+Archaeplastida, A=Archaeplastida. Each plot is a summary of 1000 stochastic character maps.

corHMM with three-state coding

Next we performed ancestral state reconstructions with a broader set of models under maximum likelihood. We fit the unordered models used above and ordered models in which direct transitions from marine to freshwater, or the reverse, freshwater to marine, are disallowed. In other words, colonization of marine environments by a freshwater lineage has to go through an intermediary brackish state. In both cases (unordere and ordered) we fit models with equal and unequal transition rates and we set the acenstral state reconstruction to provide marginal probabilities.

We compared these models using the Akaike Information Criterion corrected for sample size (AICc) and calculate Akaike weights (relative model likelihoods). We find that models that account for unequal rates are strongly favored while models with rates fixed to be equal (e.g., marine-to-freshwater = freshwater-to-marine) provide a poor fit to the data (Akaike weights close to zero).

Table 3: Comparison of ordered and unordered models with equal or unequal transition rates for the multi-state coding of habitat (freshwater, marine, brackihs). Models allowing rates to vary provide much better fit to the data.

Model	lnL	AICc	delta_AICc	AICc_w
ER.unord	-84.650	171.334	28.303	0.000
ER.ord	-101.296	204.626	61.594	0.000
ARD.unord	-67.032	146.814	3.783	0.131
ARD.ord	-67.340	143.032	0.000	0.869

As above, we look at the probabilities for marine, freshwater or brackish ancestry under different models (Figure 2).

The maximum likelihood estimates of the rates show that the marine-to-freshwater transition rate is 0.367 per billion years, whereas the reverse, freshwater-to-marine rate is much lower at 0.16 events per billion years (unordered, all-rates-different model, Table 4). If the history of transitions between marine and freshwaters on this phylogeny was consistent with the equal rates model, we would expect these two parameter estimates to be much closer even though they are estimated independently with the all-rates-different model.

The best model over all was a model with different parameters for each transition ordered in a way that disallows direct marine-to-freshwater or freshwater-to-marine shifts (i.e., freshwater <-> brackish <-> marine; Table 3). However, the optimization of this model results with poor parameter estimates, with unrealistically high transitions from brackish to marine (hitting the upper bound in `corHMM::rayDisc`) and from brackish to freshwater habitats. This is because the brackish state in this dataset is very rare (2 out of 119 taxa). This implies low persistence of this character state and requires extremely high transition rates out of it.

Table 4: Transition rate estimates between marine, freshwater, and brackish habitats under different models.

Param	ER.unord	ER.ord	ARD.unord	ARD.ord
marine-to-freshwater	0.198	NA	0.365	NA
brackish-to-freshwater	0.198	0.998	5.160	13.723
freshwater-to-marine	0.198	NA	0.160	NA
brackish-to-marine	0.198	0.998	1.701	100.000
freshwater-to-brackish	0.198	0.998	0.000	0.182
marine-to-brackish	0.198	0.998	0.302	4.942

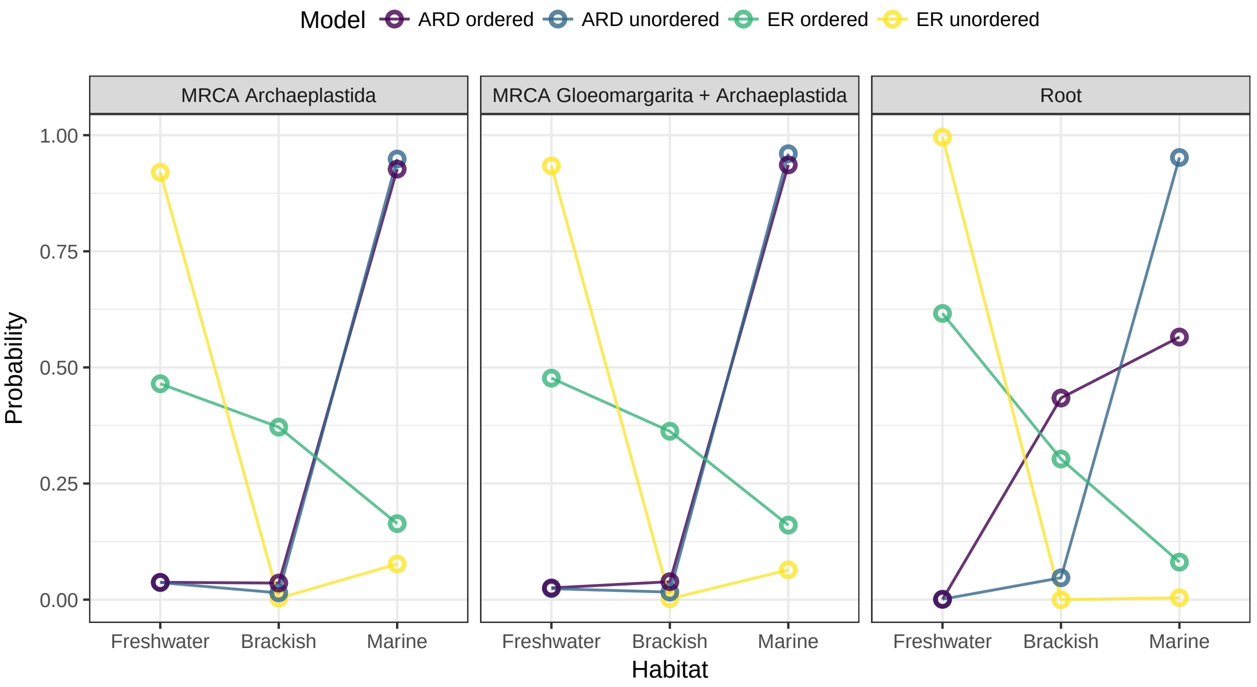

Figure 2: Probabilities for freshwater, marine, and brackish ancestry for three nodes on the Cyanobacteria + Archaeplastida phylogeny using different models of habitat transitions. Models allowing transition rates to vary provide better fit to the data and support marine ancestry, while freshwater anestry is more likely only under the unordered equal-rates model.

SIMMAP with two-state coding (marine vs. freshwater)

To bypass this optimization issue, we can look at binary coding for habitat: marine vs. freshwater. The second dataset available in the original data package had the brackish taxa scored as marine. Next, we use this dataset to reconstruct ancestral states.

SIMMAP: Equal rates model

As before, we simulate and summarise 1000 stochastic character histories under the equal-rates and all-rates-different models. The probabilities of ancestral habitat are shown in Tables 5 and 6. With binary scoring for habitat, the probability for freshwater ancestry is not very high, even under the equal-rates model, and goes further down when reconstructing with the all-rates-different model.

Table 5: Probabilities for Freshwater and Marine ancestry for three nodes on the Cyanobacteria + Archaeplastida phylogeny reconstructed with the equal-rates model.

	P.freshwater.	P.marine.
Root	0.571	0.429
MRCA Gloeomargarita + Archaeplastida	0.560	0.440
MRCA Archaeplastida	0.555	0.445

SIMMAP: All rates different model

Table 6: Probabilities for Freshwater and Marine ancestry for three nodes on the Cyanobacteria + Archaeplastida phylogeny reconstructed with the all-rates-different model.

	P.freshwater.	P.marine.
Root	0.187	0.813
MRCA Gloeomargarita + Archaeplastida	0.108	0.892
MRCA Archaeplastida	0.125	0.875

The reconstructions across the entire tree clearly show differences between model and uncertainty in ancestral states (Figure 3.)

corHMM with two-state coding

As before, we fit the same models using maximum likelihood and look at the relative probabilities of models and the transition rate estimates. The all-rates-different model is favored (Table 7) and shows strong asymmetry in transition rates (Table 8).

Table 7: Comparison of unordered models for the evolution of habitat (binary: marine vs. freshwater) with equal or unequal transition rates. Models allowing rates to vary provide much better fit to the data.

Model	lnL	AICc	delta_AICc	AICc_w
ER.unord	-62.745	127.524	0.876	0.392
ARD.unord	-61.272	126.648	0.000	0.608

Probabilities for ancestral states with different models for the three relevant nodes show again that freshwater ancestry is likely only under the unsupported equal-rates model (Figure 4, Table 7). The better-fitting all-rates-different model strongly supports marine ancestry for the MRCA of cyanobacteria and the MRCA of archaeplastids (Figure 4).

Table 8: Transition rate estimates between marine and freshwater habitats under models with equal and unequal transition rates.

Param	ER.unord	ARD.unord
marine-to-freshwater	0.503	0.594
freshwater-to-marine	0.503	0.230

Although the all-rates-different model is favored (AICc weight = 0.61), the equal-rates model is plausible (AICc weight = 0.39). To account for this uncertainty in model choice, we can average the transition rates across the two models. We do this by taking the average of the corresponding rates across models weighted by the relative likelihoods (Akaike weights) of the models. Model-averaging reduced the difference between the rates, due to influence from the equal-rates model, however, the asymmetry is still strong: marine-to-freshwater = 0.55 vs. freshwater-to-marine=0.33. We then reconstructed ancestral states using these model-averaged transition rates and found again stronger support for marine ancestry for the MRCA of cyanobacteria and archaeplastids (Figure 5).

Model-averaging

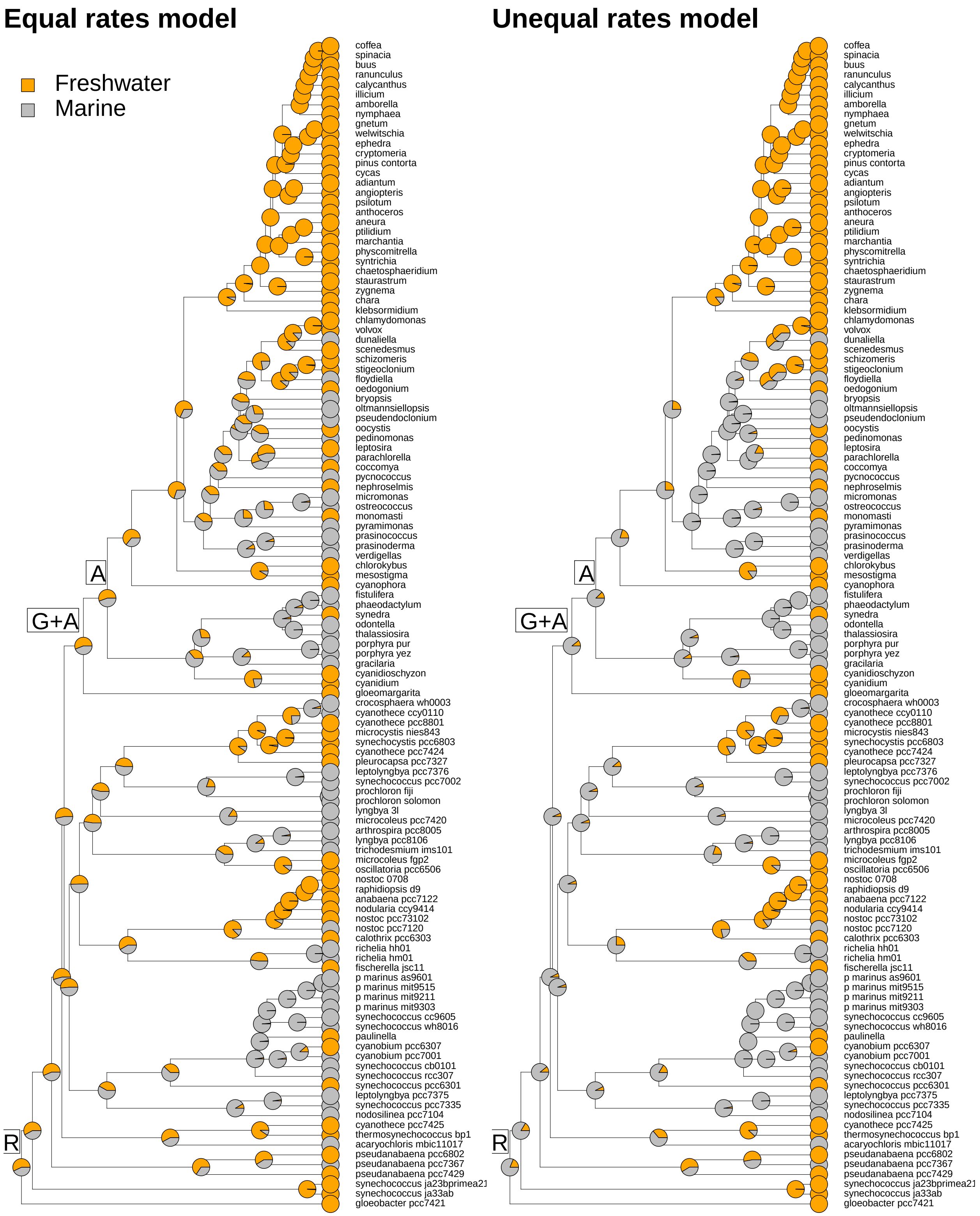

Figure 3: Comparison of ancestral states for habitat (binary: marine vs. freshwater) reconstructed with an equal rates (ER, left) and all-rates-different (ARD, right) model. The relevant nodes are labeled: R=root, G+A=Gloeomargarita+Archaeplastida, A=Archaeplastida. Each plot is a summary of 1000 stochastic characte maps.

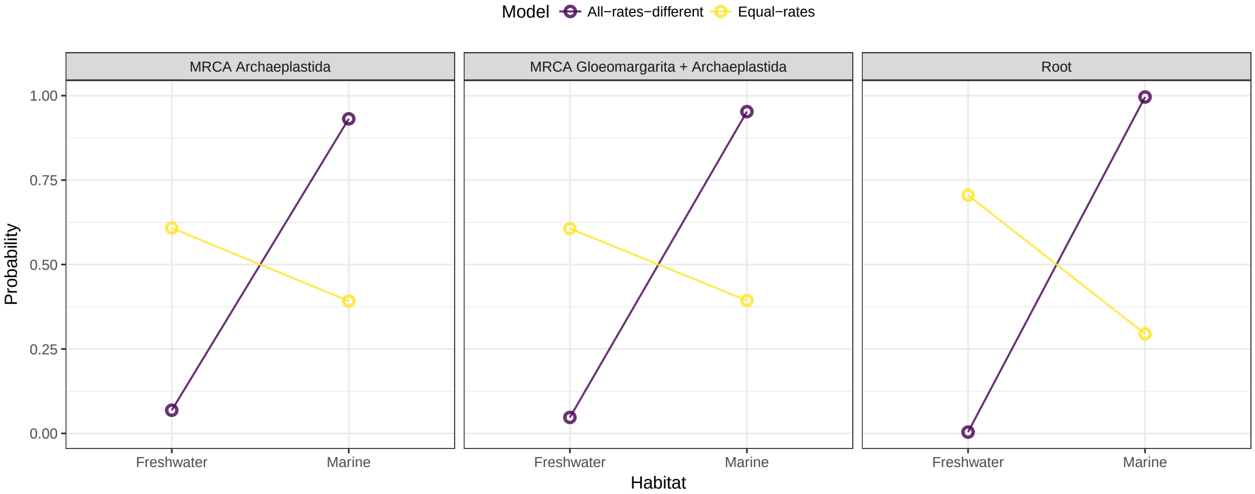

Figure 4: Probabilities for Freshwater and Marine ancestry for three nodes on the Cyanobacteria + Archaeplastida phylogeny using different models of habitat transitions and binary coding for habitat. Models allowing transition rates to vary support marine ancestry, while freshwater ancestry is more likely only under the unordered equal-rates model.

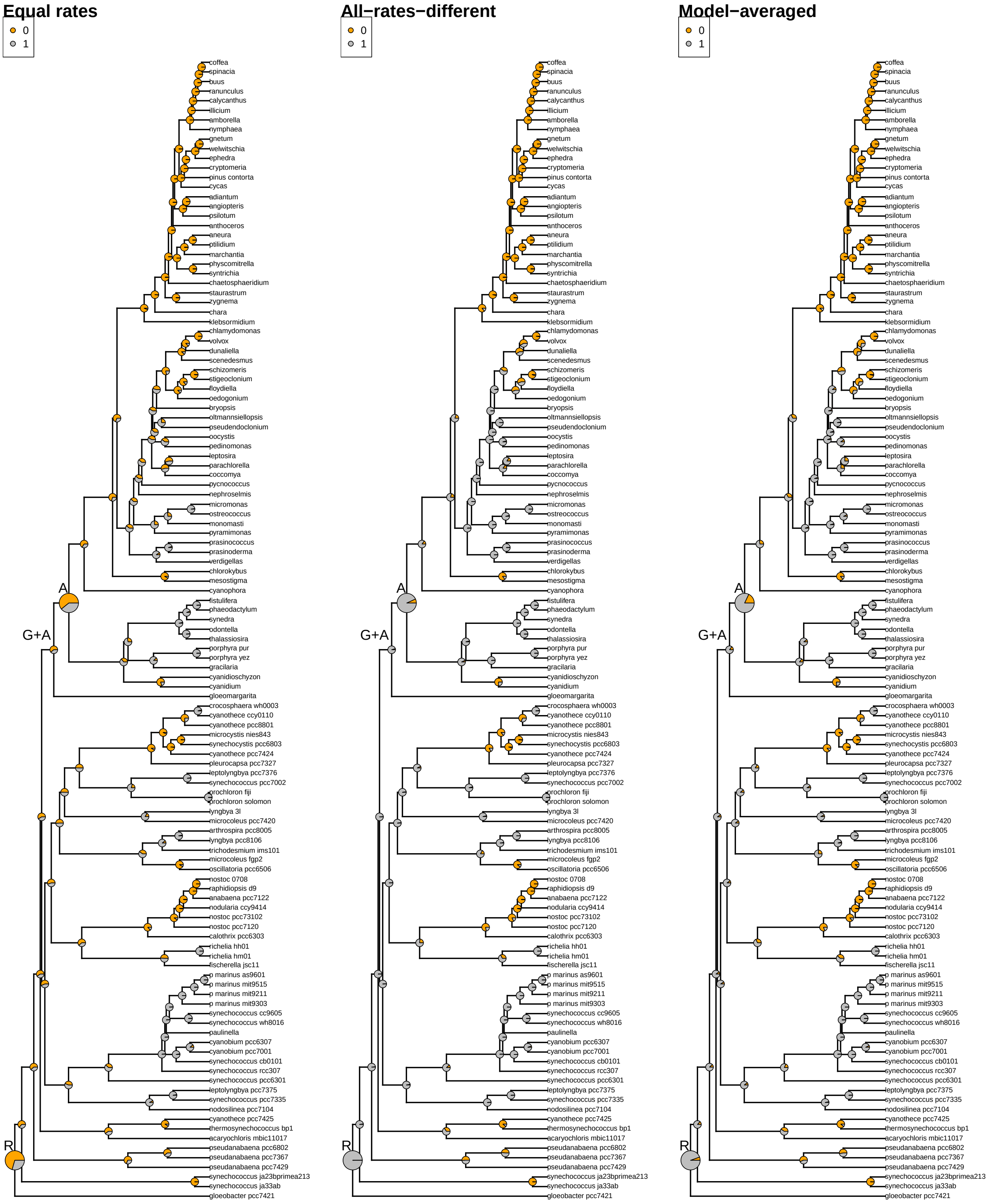

Figure 5: Ancestral state reconstruction of habitat [binary: marine (grey) or freshwater (orange)] for the Cyanobacteria + Archaeplastida phylogeny under an equal rates model (left), unequal rates model (middle) and with model-averaged transition rates (right). Accounting for the asymmetry in transition rates (unequal and model-averaged reconstruction) clearly favors a marine, rather than freshwater, ancestry for the MRCA of cyanobacteria and archaeplastids.