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 1
Equal rates model 1
All rates different model 2
corHMM with three-state coding 4
SIMMAP with two-state coding 5
SIMMAP: Equal rates model 5
SIMMAP: All rates different model 5
corHMM with two-state coding 6
Model-averaging 6

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.

```
synechococcus_pcc7002 1 2 NO
nostoc_pcc7120 1 0 NO
cyanobium_pcc7001 1 2 NO
```

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

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) were as follows:

Table 1: Probabilities for Freshwater, Marine, and Brackish ancestry for three nodes on the Cyanobacteria + Archaeplastida phylogeny

	P.freshwater.	P.marine.	P.brackish.
Root	0.7	0.3	0
MRCA Gloeomargarita + Archaeplastida	0.6	0.4	0
MRCA Archaeplastida	0.6	0.4	0

These results confirm the finding of Sanchez-Baracaldo et al. 2017. 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 are fresh waters [<0.5 parts per thousand (ppt) salinity].

All rates different model

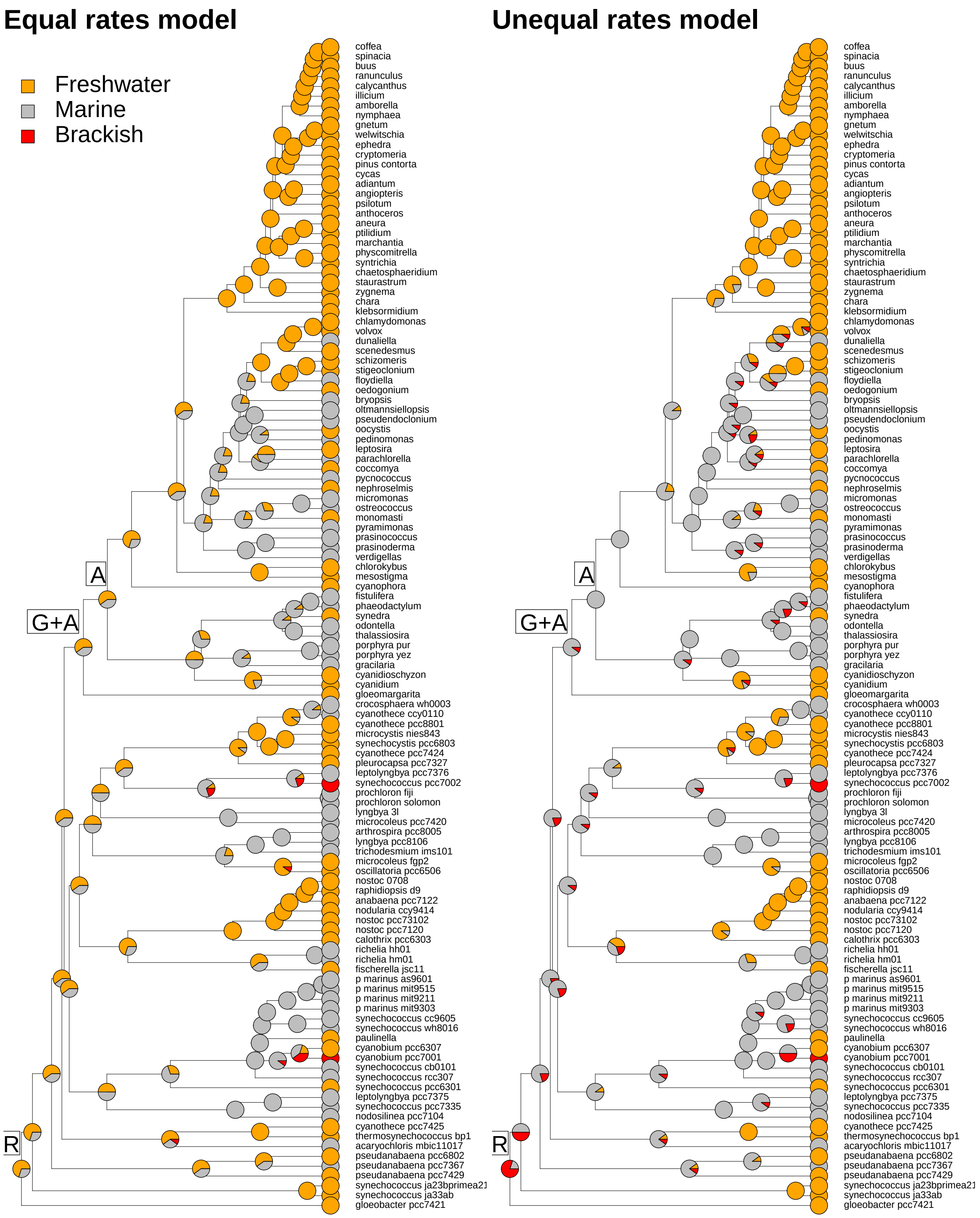
It is possible, however, that transitions between these three habitats happen at different rates, perhaps as a result of cellular physiological constraints. Marine and fresh waters, after all, differ drastically in many important ways, including concentrations of salts and other ions, osmotic pressure, 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 models that allowed these transitions to vary. In these models, the different transition rates are independently estimated, and importantly, if two evolutionary transitions indeed have similar rates, these models are able to detect that and return similar parameter estimates. We simulated 1000 stochastic character histories using a model with *unequal rates*, 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 completely reversed. The probability for the fresh water state went down and probability for marine or brackish ancestry went up after we accounted for different rates of transition between states.

Table 2: Probabilities for Freshwater, Marine, and Brackish ancestry for three nodes on the Cyanobacteria + Archaeplastida phylogeny

	P.freshwater.	P.marine.	P.brackish.
Root	0	0.2	0.8
MRCA Gloeomargarita + Archaeplastida	0	0.9	0.1
MRCA Archaeplastida	0	1.0	0.0



Equal rates model

Unequal rates model

Figure 1: Comparison of ancestral states 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.

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 transitions from marine to freshwater, or the reverse, have to pass through an intermediary brackish state. In both cases we have 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. 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

Probabilities for ancestral states with different models:

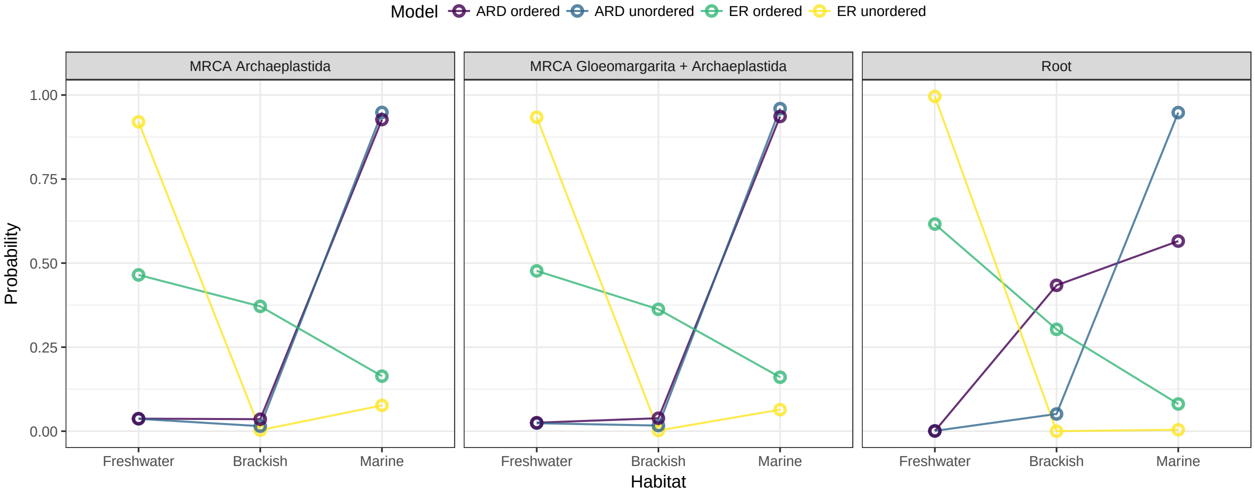


Figure 2: Probabilities for Freshwater, Marine, and Brackish ancestry for three nodes on the Cyanobacteria + Archaeplastida phylogeny using different models of habitat transitions. Models allowing transition rates to vary support marine ancestry, while freshwater anestry is more likely only under the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

The maximum likelihood estimates of the rates are show that the marine-to-freshwater transition rate is 0.367 per lineage per billion years, whereas the reverse, freshwater-to-marine rate is much lower at 0.16 events per lineage per billion years (all-rates-different, unordered model). 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 the all-rates-different model estimated them independently.

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 transitions from brackish to marine (q=100, hitting the upper bound in `corHMM::rayDisc`) and from brackish to freshwater (q=13.7) unrealistically high. This is because the brackish state in this dataset is very rare (2 out of 119 taxa).

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	0.369	NA
brackish-to-freshwater	0.198	0.998	5.073	13.738
freshwater-to-marine	0.198	NA	0.160	NA
brackish-to-marine	0.198	0.998	1.764	100.000
freshwater-to-brackish	0.198	0.998	0.000	0.182
marine-to-brackish	0.198	0.998	0.301	4.937

SIMMAP with two-state coding

SIMMAP: Equal rates model

Table 5: Probabilities for Freshwater and Marine ancestry for three nodes on the Cyanobacteria + Archaeplastida phylogeny

	P.freshwater.	P.marine.
Root	0.3	0.7
MRCA Gloeomargarita + Archaeplastida	0.3	0.7
MRCA Archaeplastida	0.4	0.6

SIMMAP: All rates different model

Table 6: Probabilities for Freshwater and Marine ancestry for three nodes on the Cyanobacteria + Archaeplastida phylogeny

	P.freshwater.	P.marine.
Root	0	1
MRCA Gloeomargarita + Archaeplastida	0	1
MRCA Archaeplastida	0	1

corHMM with two-state coding

Table 7: Comparison of unordered models 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:

Table 8: Transition rate estimates between marine and freshwater habitats under different models.

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

Model-averaging

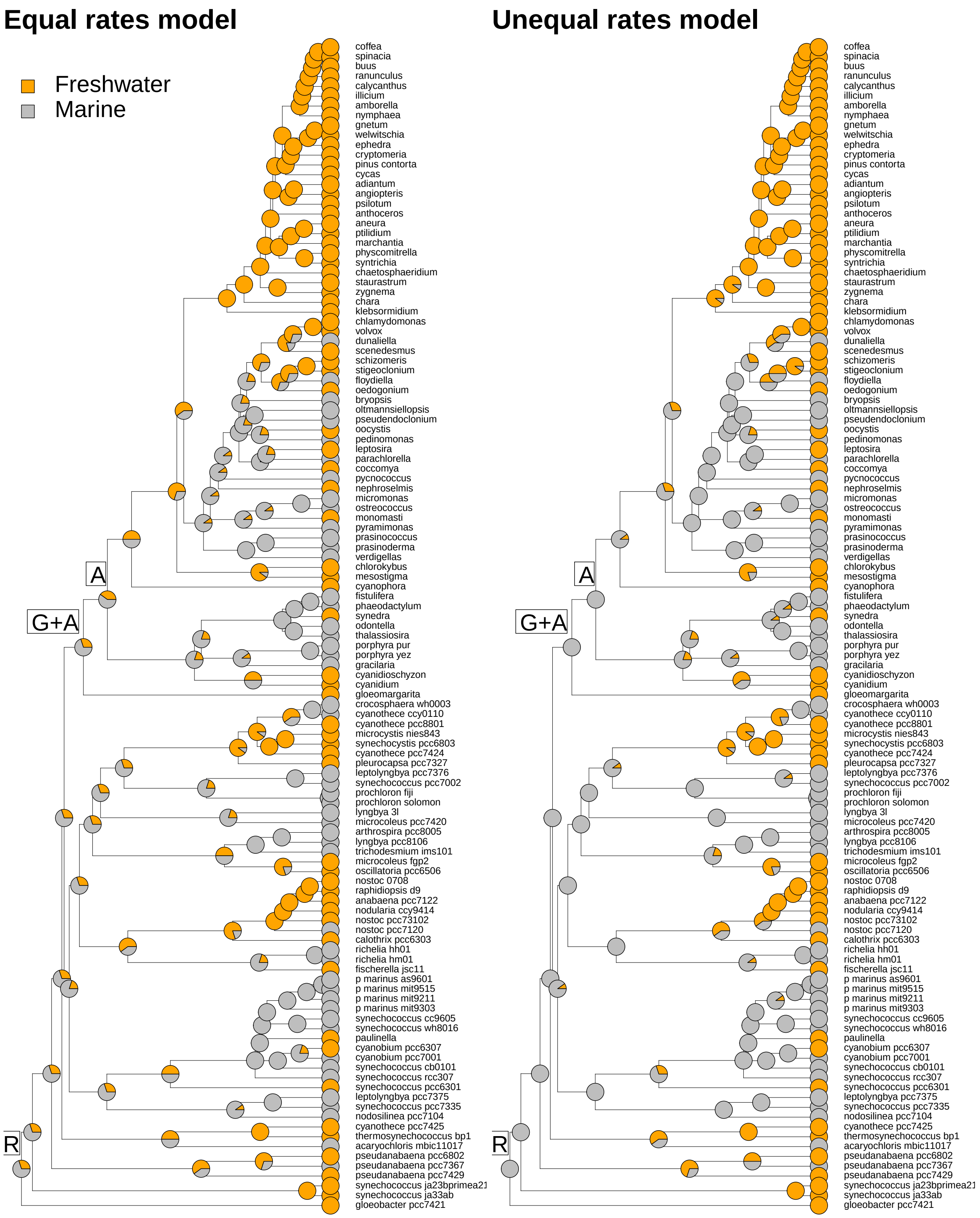


Figure 3: Comparison of ancestral states reconstructed with an equal rates (ER, left) and all-rates-different (ARD, right) model. The relevant nodes are labeled: R=root, G+A=Gloeomargarita+Archaeplastida, A=Archaeplastida. Each plot is a summary of 1000 stochastic characte maps.

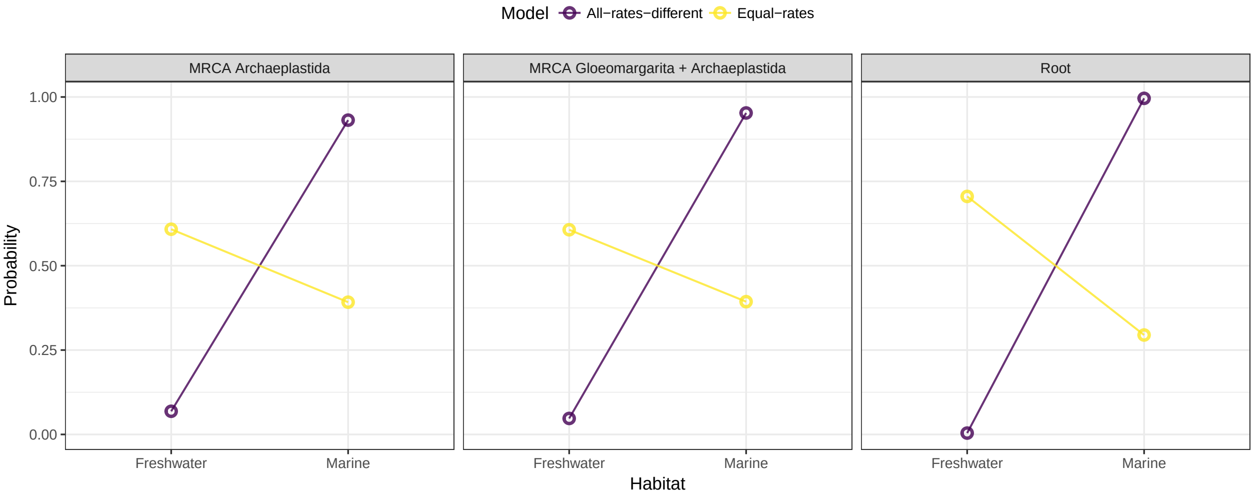
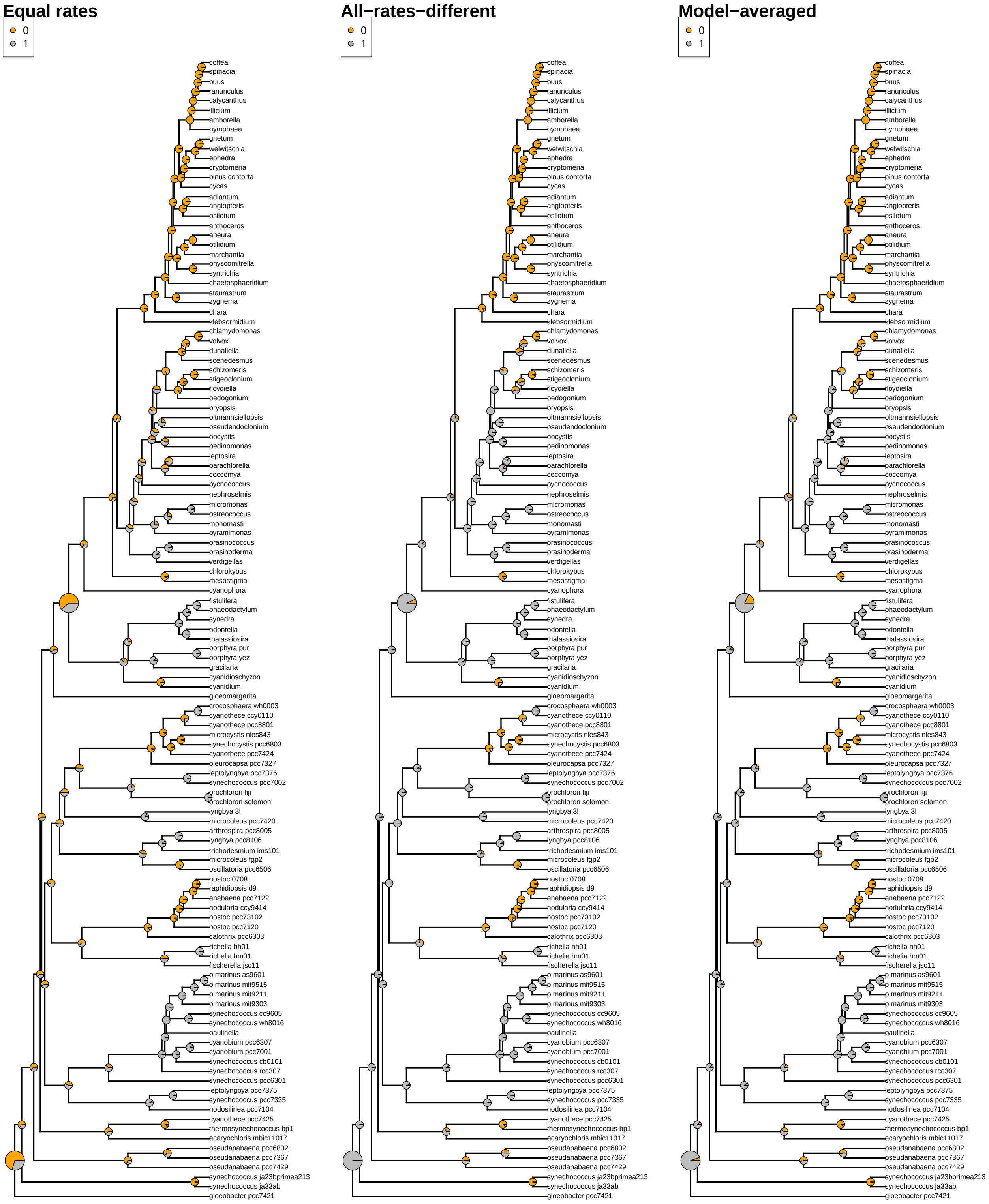


Figure 4: Probabilities for Freshwater and Marine ancestry for three nodes on the Cyanobacteria + Archaeplastida phylogeny using different models of habitat transitions. Models allowing transition rates to vary support marine ancestry, while freshwater anestry is more likely only under the unordered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



Equal rates

All−rates−different

Model−averaged