Supplementary Information - Reply to Nakov et al.: Model choice requires biological insight when studying the ancestral habitat of photosynthetic eukaryotes

Patricia Sánchez-Baracaldo, Giorgio Bianchini, John P. Huelsenbeck, John A. Raven, Davide Pisani, and Andrew H. Knoll

27 October, 2017

Table of Contents

Data files

We used data from the Sánchez-Baracaldo et al. [2], which are available on DataDryad [3]. The two datasets analysed consist of a "multistate" dataset, in which taxa character states are defined as *Freshwater* (0), *Marine* (1), or *Brackish* (2), and a "binary" dataset, in which character state for brackish taxa are coded as marine. The "binary" dataset was is a slightly different version presented by Nakov et al. (Nakov, Boyko, Alverson, & Beaulieu, 2017).

We corrected a mistake found in the nexus dataset found in DataDryad, in which the species *Nostoc sp.* PCC 7120 was coded as marine. This is, in fact, a freshwater strain (See Pasteur cyanobacteria collection: https://webext.pasteur.fr/cyanobacteria/). We would like to highlight that all xml scripts ran for the Sanchez-Baracaldo et al. 2017 have the correct character states as shown in SI Appendix Table S2; copy of these files are also found in DataDryad.

Packages

We performed our analyses in R packages (`phytools` v0.6.20 and `corHMM` v1.22) as Nakov et al. [1]. We cannot confirm that the versions of the packages are the same as Nakov et al. [1], since they did not report the version numbers they used.

Analyses

Multistate data

Phytools

For our first set of analyses, we used the `make.simmap` function of `phytools` to perform stochastic character mapping. This function allows three different transition models: an "equal rates" (ER) model, in which all the transitions have the same rate, an "all-rates-different" (ARD) model, in which each transition has a different rate, and a "symmetrical" (SYM) model, in which symmetrical transitions have the same rate. **TABLE 1** summarises these different models. Nakov et al. [1] only implemented ER and ARD models. We have expanded this study by including a SYM model.

Model	Equal rates (ER)	All-rates-different (ARD)	Symmetrical (SYM)
Parameters to be estimated	1	6	3
Rate matrix	0 1 2 0 * 1 1 1 1 * 1 2 1 1 *	0 1 2 0 * 3 5 1 1 * 6 2 2 4 *	0 1 2 0 * 1 2 1 1 * 3 2 2 3 *

Table 1: Evolutionary models implemented in `phytools`. In the rate matrices, each number represents a parameter that is estimated from the data (equal numbers represent parameters that are constrained to be equal), and rows represent the starting state, while columns represent the final state. Cells with asterisks (*) represent parameters that are not estimated from the data. For example, the cell with number 4 in the ARD rate matrix is the rate of going from state 2 (brackish) to state 1 (marine).

In all cases, we set the prior on the root node as `equal` (i.e., we assume, *a priori*, that the last common ancestor at the root had equal probability of living in a freshwater/marine/brackish environment) and we simulated 1000 stochastic maps with an `empirical` rate matrix (this means that a maximum likelihood estimate for the rate matrix is computed once, and then used for all the subsequent stochastic map simulations). The results of these analyses were plotted using functions in the `phytools` package.

In order to assess model fit, we used the log-likelihoods for the rate matrices that are computed inside the `make.simmap` function. Contrary to what the documentation for this function [4] says, this value is never exported outside of the function; thus, to access it, we had to modify the source for the `make.simmap` function (see the make.simmap.R script).

Three nodes are important for our analyses: the root of the tree, the most recent common ancestor (MRCA) of *Gloeomargarita* and Archaeplastida (G+A), and the MRCA of Archaeplastida (A). **TABLE 2** shows the probabilities for the various ancestral habitats of these nodes according to the different models (the complete ancestral state reconstruction plots are available from https://github.com/patrisan/Chloroplast_PNAS).

Node	Model								
	ER			ARD			SYM		
	P_f	P_m	P_b	P_f	P_m	P_b	P_f	P_m	P_b
Root	0.814	0.156	0.030	0.048	0.560	0.392	0.733	0.267	0.000
G+A	0.817	0.178	0.005	0.036	0.927	0.037	0.738	0.262	0.000
A	0.817	0.180	0.003	0.042	0.932	0.026	0.739	0.261	0.000

Table 2: Probabilities for the ancestral habitats for the relevant nodes, reconstructed with different evolutionary models.

These numbers are similar to those reported by Nakov et al. [1] (minor differences may be explained by the incorrect character state of *Nostoc* PCC 7120 in the dataset they used), and show that the results of the ancestral state reconstruction can be radically different under different evolutionary models.

To assess which model best fits the data, we used the log-likelihoods computed by `make.simmap` to calculate AICc (corrected Akaike Information Criterion) and BIC (Bayesian Information Criterion) scores for each model, obtaining the data in TABLE 3.

Model	lnL	k	AICc	w_{AICc}	BIC	w_{BIC}
ER	-85.405	1	172.844	0.000	175.589	0.000
ARD	-67.387	6	147.525	0.803	163.449	0.077
SYM	-72.066	3	150.340	0.197	158.469	0.923

Table 3: Comparison of ER, ARD and SYM models in `phytools`. AICc and BIC identify different models as the best ones, and in any case do not provide unequivocal evidence for accepting any single model. **lnL**: log-likelihood, **k**: number of parameters in the model, w_{AICc}: weights computed using AICc scores, w_{BIC}: weights computed using BIC scores.

Nakov et al. [1] did not report the log-likelihoods, thus we cannot compare our results with theirs. However, these figures show that the data does not identify a model that clearly over performs the other ones. In fact, AICc and BIC rank the models differently, and model weights computed from neither score give overwhelmingly more credit to one model than the other.

To understand what is going on, we looked at the maximum likelihood estimates for the transition rate matrix that were computed and used by `make.simmap`. TABLE 4 reports these values.

Model	ER			ARD			SYM		
Maximum likelihood estimate of the transition rate matrix		0	1	2	0	1	2	0	1	2
	0	-0.415	0.207	0.207	-0.172	0.172	0.000	-0.467	0.467	0.000
	1	0.207	-0.415	0.207	0.456	-0.744	0.288	0.467	-0.526	0.059
	2	0.207	0.207	-0.415	2.932	3.575	-6.508	0.000	0.059	-0.059

Table 4: Maximum likelihood estimates of the transition rate matrices for the various models.

Nakov et al. [1] did not report these rates, and thus we cannot compare them, but in their letter [5] they recognize that "low frequency states, like the brackish state in this dataset […], require extremely high transition rates away from them to account for their low observed frequency", and this is what emerges from the transition matrix for the ARD model. It is therefore likely that the presence of only two brackish species causes a bias in the maximum likelihood estimate. By constraining the rates in and out of the brackish state to be equal, the SYM model prevents this, while still allowing more flexibility than the ER model.

To take into account the uncertainty in model choice that emerges from AICc and BIC scores, we can do model-averaging analyses. In other words, average the transition rates across the different models, using the weights computed from AICc scores (smoothed AICc, SAICc) or those derived from BIC scores (Bayesian model averaging, BMA). The results of these analyses are shown in TABLE 5.

	Method					
	SAICc			BMA		
Rates		0	1	2	0	1	2
	0	-0.230	0.230	0.000	-0.445	0.445	0.000
	1	0.459	-0.701	0.243	0.466	-0.543	0.076
	2	2.356	2.884	-5.240	0.224	0.328	-0.552

Node	P_f	P_m	P_b	P_f	P_m	P_b
Root	0.098	0.529	0.373	0.610	0.346	0.044
G+A	0.078	0.875	0.047	0.647	0.351	0.002
A	0.093	0.874	0.033	0.650	0.348	0.002

Table 5: Results of the model averaging analyses.

These results (in particular the BMA) show that the data, on its own, does not give a definitive answer to whether any of the ancestors we are interested in inhabited in a freshwater or marine environment. The complete ancestral state reconstruction plots for these model averages are available from https://github.com/patrisan/Chloroplast_PNAS.

To reach a conclusion on the matter, it is thus essential to take into account biological information about the process that we are studying. Few studies have looked at transitions between freshwater ↔ marine habitats, Logares et al. [6] noted that transitions across salinity barriers in the microbial world are likely infrequent. In order to include this insight in our analysis, one could apply a strong prior on low transition rates.

The `make.simmap` function allows us to do so by using an `mcmc` to sample rate matrices. In practice, this means that the program uses a Markov chain to sample rate matrices according to their posterior probability distribution (given the data) and one such sample is taken for each of the requested simulations (in our case, 1000) and used to reconstruct the ancestral states on the tree. The results are then averaged over all of the simulations. The prior distribution on the rate matrix can be specified using a beta distribution. To implement a strong prior on low transition rates, we used a beta distribution with parameters $\alpha = 1$ and $\beta = 400$. The results are shown in TABLE 6.

	Model								
	ER			ARD			SYM		
Mean Rates	0 1 2			0 1 2			0 1 2		
	0	-0.092	0.046	0	-0.042	0.036	0.006	0	-0.067	0.062	0.005
	1	0.046	-0.092	0.046	1	0.031	-0.036	0.005	1	0.062	-0.070	0.008
	2	0.046	0.046	-0.092	2	0.014	0.031	-0.045	2	0.005	0.008	-0.012
Node	P_f	P_m	P_b	P_f	P_m	P_b	P_f	P_m	P_b
Root	0.998	0.002	0.000	0.997	0.003	0.000	0.996	0.004	0.000
G+A	0.997	0.003	0.000	0.998	0.002	0.000	0.994	0.006	0.000
A	0.997	0.003	0.000	0.997	0.003	0.000	0.994	0.006	0.000

Table 6: Results of the analysis when a strong prior on low transition rates is used.

These results show that, when low rates are assumed for transitions across character states, a freshwater habitat for all of the nodes of interest is strongly supported. The complete ancestral state reconstruction plot for the SYM model is shown in FIGURE 1; the other ones are available from https://github.com/patrisan/Chloroplast_PNAS.

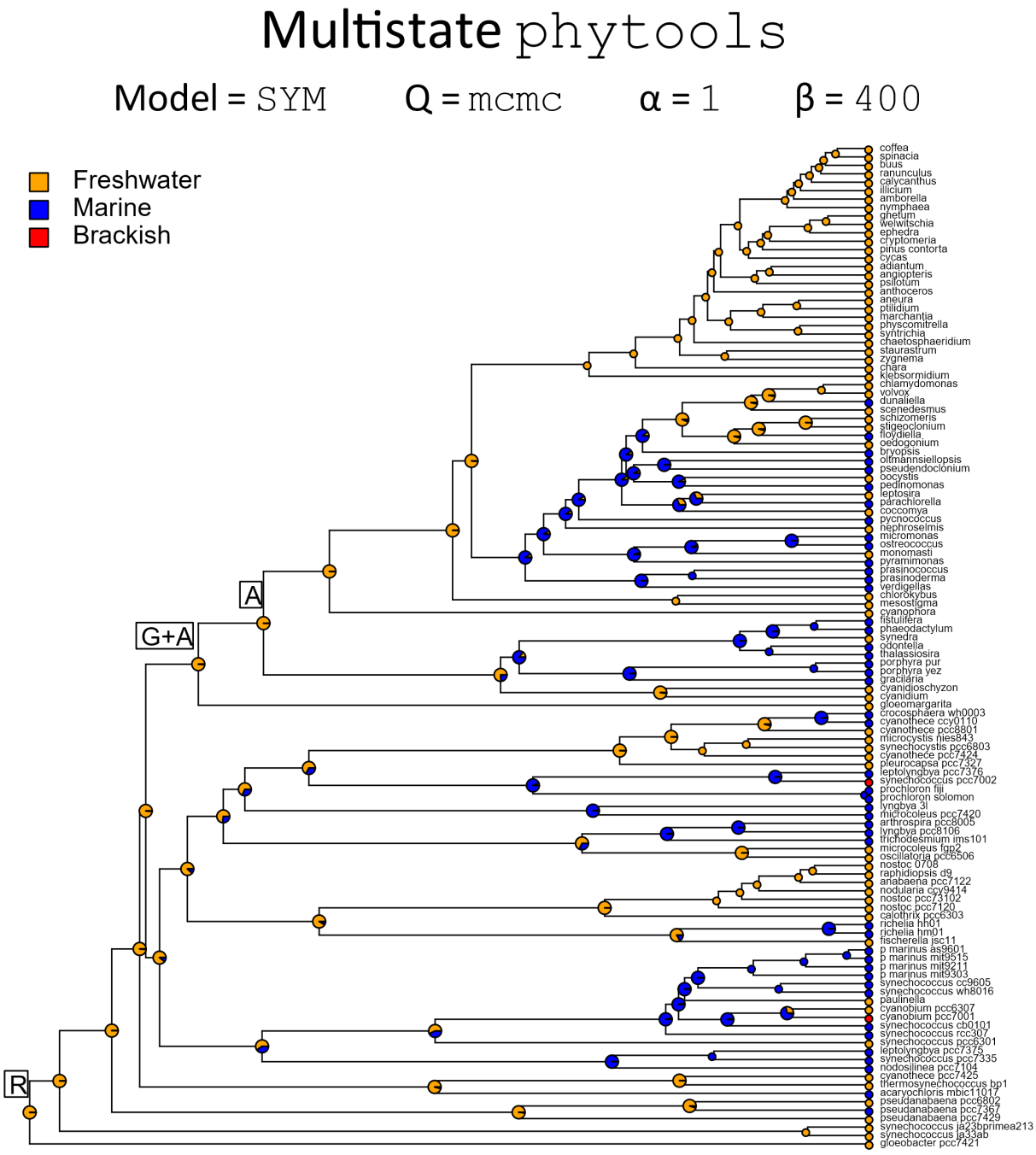

***Figure 1**:* Ancestral state reconstruction for the SYM model with a strong prior on low transition rates (beta distribution with parameters α = 1 and β = 400).

It could be argued that our choice of prior biases the analysis (i.e., it makes the freshwater ancestry the only viable hypothesis, regardless of the data). To show that this is not the case, we did a Bayes factor analysis, comparing the posterior odds of a freshwater state for the root and the MRCAs of Archaeplastida and *Gloeomargarita* + Archaeplastida to their prior odds.

To do this analysis, we need to sample the prior probability distribution of clocklike trees for the 119-taxa dataset. We thus built 1000 ultrametric trees using a gamma prior on the age of the root node (with mean 2500Mya and standard deviation 100Mya) and a flat Dirichlet prior on branch length proportions. Furthermore, we constrained the trees to have the three nodes we are interested in:

- We constrained the Archaeplastida to be monophyletic
- We constrained *Gloeomargarita* to be the sister taxon to the Archaeplastida
- We constrained *Gloeobacter* to be the sister taxon to all the other 118 species in the tree

For each of the 1000 trees, we then ran an analysis using the `make.simmap` function with the strong prior on low transition rates (doing one simulation per tree). Finally, for each interesting node we counted in what proportion of the 1000 stochastic maps it was in a freshwater state, and used that proportion as the prior probability of the node being freshwater.

The results of this analysis are shown in TABLE 7.

	Model					
	ER		ARD		SYM	
Node	p_f	B_f	p_f	B_f	p_f	B_f
Root	0,741	174,414	0,805	80,503	0,739	87,942
G+A	0,989	3,696	0,965	18,098	0,990	1,673
A	0,994	2,006	0,980	6,782	0,992	1,336

Table 7: Results of the Bayes factor analysis. p_f: prior probability of the node being freshwater, B_f: Bayes factor (here computed as the ratio of posterior to prior odds: $B_f = \frac{P_f}{1-P_f} \cdot \frac{1-p_f}{p_f}$).

This table shows that the prior on low transition rates does add a certain bias towards the freshwater origin; in fact, the contribution of the data (i.e., the correct topology and branch lengths) to our conclusions, if we stand by Kass and Raftery's modification[1] [7] to the Bayes factor interpretation table developed by Jeffreys [8], can be summarised like this:

- For the ER model, the data provides *barely worth mentioning* evidence that the MRCA of Archaeplastida inhabited freshwater habitats, *positive* evidence that the MRCA of Archaeplastida and *Gloeomargarita* lived in a freshwater environment, and *very strong* evidence that the root node inhabited freshwater habitats.
- For the ARD model, the data provides *positive* evidence that the MRCA of Archaeplastida and the MRCA of Archaeplastida and *Gloeomargarita* lived in a freshwater habitat and *strong* evidence that the root node inhabited a freshwater habitat[2].
- For the SYM model, the data contributes *barely worth mentioning* evidence that the MRCA of Archaeplastida and *Gloeomargarita* and the MRCA of Archaeplastida inhabited freshwater habitats, and *strong* evidence that the root node inhabited freshwater habitats.

These results are thus in line with the idea that more analyses are likely needed to be able to give a definite answer to whether these three ancestors lived in a freshwater or marine environment.

[1] The table is reported here for the reader's convenience:

B_f	Strength of evidence
1 to 3	Not worth more than a bare mention
3 to 20	Positive
20 to 150	Strong
> 150	Very strong

[2] These analyses are still running, thus the results presented here for the ARD model are based on a reduced sample of 200 trees.

CorHMM

We also replicated and extended the analyses done by Nakov et al. [1] using the `corHMM` R package. This package allows users to modify the matrices identifying what parameters of the transition model should be estimated, and thus makes it possible to implement many different models. Nakov et al., [1] used this to analyse the ER and ARD models, as well as "ordered" versions of these two models (i.e., models in which a taxa transitioning from a freshwater state to a marine state has to go through a brackish state first). We extended their analyses by considering the SYM model, the ordered SYM model, and other models that we refer to as "some-rates-different" (SRD). A summary of the models that we analysed, with the respective rate matrices, is in TABLE 8.

Model	Equal rates (ER)	All-rates-different (ARD)	Symmetrical (SYM)
Parameters to be estimated	1	6	3
Rate matrix	0 1 2 0 * 1 1 1 1 * 1 2 1 1 *	0 1 2 0 * 3 5 1 1 * 6 2 2 4 *	0 1 2 0 * 1 2 1 1 * 3 2 2 3 *
Model	**Ordered equal rates (oER)**	**Ordered all-rates-different (oARD)**	**Ordered symmetrical (oSYM)**
Parameters to be estimated	1	4	2
Rate matrix	0 1 2 0 * * 1 1 * * 1 2 1 1 *	0 1 2 0 * * 3 1 * * 4 2 1 2 *	0 1 2 0 * * 1 1 * * 2 2 1 2 *
Model	**Some-rates-different 1 (SRD1)**	**Some-rates-different 2 (SRD2)**	**Some-rates-different 3 (SRD3)**
Parameters to be estimated	5	3	2
Rate matrix	0 1 2 0 * 1 4 1 1 * 5 2 2 3 *	0 1 2 0 * 1 3 1 1 * 3 2 2 2 *	0 1 2 0 * 1 2 1 1 * 2 2 2 2 *
Model	**Ordered some-rates-different 4 (oSRD4)**		
Parameters to be estimated	2		
Rate matrix	0 1 2 0 * * 2 1 * * 2 2 1 1 *		

Table 8: Evolutionary models that were tested in `corHMM`. In the rate matrices, each number represents a parameter that is estimated from the data (equal numbers represent parameters that are constrained to be equal), and rows represent the starting state, while columns represent the final state. Cells with asterisks (*) represent parameters that are not estimated from the data. For example, the cell with number 4 in the ARD rate matrix is the rate of going from state 2 (brackish) to state 1 (marine).

We ran the analyses with these models using the `rayDISC` function of the `corHMM` package, obtaining the maximum likelihood estimates for the rate matrices and the probabilities for the interesting nodes that are displayed in **TABLE 9**.

Model	Equal rates (ER)			All-rates-different (ARD)			Symmetrical (SYM)		
Rate matrix	0	1	2	0	1	2	0	1	2
0	-0.396	0.198	0.198	-0.161	0.161	0.000	-0.452	0.452	0.000
1	0.198	-0.396	0.198	0.362	-0.666	0.304	0.452	-0.516	0.065
2	0.198	0.198	-0.396	5.241	1.642	-6.883	0.000	0.065	-0.065
Node	**P_f**	**P_m**	**P_b**	**P_f**	**P_m**	**P_b**	**P_f**	**P_m**	**P_b**
Root	0.996	0.004	0.000	0.001	0.956	0.043	0.954	0,046	0,000
G+A	0.934	0.064	0.002	0.024	0.960	0.016	0.835	0,165	0,000
A	0.920	0.077	0.003	0.037	0.949	0.014	0.814	0,186	0,000
Model	**Ordered equal rates (oER)**			**Ordered all-rates-different (oARD)**			**Ordered symmetrical (oSYM)**		
Rate matrix	0	1	2	0	1	2	0	1	2
0	-0.998	0.000	0.998	-0.253	0.000	0.253	-0.474	0.000	0.474
1	0.000	-0.998	0.998	0.000	-1.254	1.254	0.000	-17.47	17.47
2	0.998	0.998	-1.997	14.85	17.62	-32.47	0.474	17.47	-17.94
Node	**P_f**	**P_m**	**P_b**	**P_f**	**P_m**	**P_b**	**P_f**	**P_m**	**P_b**
Root	0.616	0.081	0.303	0.000	0.788	0.212	1.000	0.000	0.000
G+A	0.477	0.161	0.363	0.019	0.954	0.027	0.986	0.006	0.008
A	0.465	0.164	0.371	0.029	0.951	0.021	0.977	0.010	0.013
Model	**Some-rates-different 1 (SRD1)**			**Some-rates-different 2 (SRD2)**			**Some-rates-different 3 (SRD3)**		
Rate matrix	0	1	2	0	1	2	0	1	2
0	-0.180	0.180	0.000	-0.524	0.414	0.110	-0.498	0.474	0.024
1	0.180	-0.606	0.426	0.414	-0.524	0.110	0.474	-0.498	0.024
2	9.057	0.000	-9.057	2.767	2.767	-5.533	0.024	0.024	-0.049
Node	**P_f**	**P_m**	**P_b**	**P_f**	**P_m**	**P_b**	**P_f**	**P_m**	**P_b**
Root	0.001	0.998	0.001	0.568	0.133	0.299	0.789	0.211	0.000
G+A	0.023	0.967	0.010	0.625	0.337	0.038	0.665	0.335	0.000
A	0.041	0.952	0.007	0.628	0.344	0.027	0.658	0.342	0.000
Model	**Ordered some-rates-different 4 (oSRD4)**								
Rate matrix	0	1	2						
0	-0.952	0.000	0.952						
1	0.000	-0.952	0.952						
2	25.713	25.713	-51.426						
Node	**P_f**	**P_m**	**P_b**						
Root	0.556	0.147	0.297						
G+A	0.629	0.353	0.018						
A	0.626	0.355	0.019						

Table 9: Probabilities for the ancestral habitats for the relevant nodes and maximum likelihood estimates of the transition rate matrices, reconstructed with different evolutionary models.

As was the case with the results of the phytools analysis, the numbers in this table are similar to those reported by Nakov et al. [1] and show that different models result in very different reconstructions.

As we did before, in order to assess which model best fits the data, we compared the various models using the AICc and BIC compute from the log-likelihoods reported by the `rayDISC` function, obtaining the values of TABLE 10.

Model	lnL	k	AICc	w_{AICc}	BIC	w_{BIC}
ER	-84.650	1	171.334	0.000	174.079	0.000
ARD	-67.032	6	146.815	0.079	162.739	0.002
SYM	-71.479	3	149.166	0.024	157.295	0.023
oER	-101.296	1	204.626	0.000	207.371	0.000
oARD	-67.671	4	143.693	0.377	154.458	0.096
oSYM	-96.218	3	198.645	0.000	206.774	0.000
SRD1	-67.130	5	144.792	0.218	158.156	0.015
SRD2	-70.216	3	146.640	0.086	154.769	0.082
SRD3	-72.476	2	149.055	0.026	154.509	0.094
oSRD4	-70.484	2	145.071	0.189	150.525	0.688

Table 10: Comparison the models that we tested in `corHMM`. AICc and BIC identify different models as the best ones, and in any case do not provide unequivocal evidence for accepting any single model.

Again, these figures show that the data does not identify a model that clearly over performs the other ones: in fact, AICc and BIC rank the models differently, and model weights computed from neither score give overwhelmingly more credit to one model than to the other. These results are different than what was described by Nakov et al. [1], because they did not take into account all of the models that we considered (and thus reported that, amongst the models that they used, oARD was favoured).

To take into account the different models, we performed again two model averaging analyses, using weights derived from AICc (SAICc) or from BIC (BMA). The results of these analyses are shown in TABLE 11.

	Method					
	SAICc			BMA		
Rates	0 1 2 0 -0,397 0,111 0,286 1 0,127 -0,909 0,782 2 13,09 11,89 -24,98			0 1 2 0 -0,783 0,092 0,691 1 0,092 -0,888 0,795 2 19,48 19,61 -39,09		
Node	P_f	P_m	P_b	P_f	P_m	P_b
Root	0,008	0,782	0,210	0,417	0,254	0,329
G+A	0,072	0,903	0,024	0,522	0,459	0,019
A	0,097	0,883	0,021	0,526	0,454	0,020

Table 11: Results of the model averaging analyses.

As before, the BMA particularly shows that the data, on its own, does not give a definitive answer to whether any of the ancestors of our interest lived in a freshwater or marine environment. The complete ancestral state reconstruction plots for these model averages are available from https://github.com/patrisan/Chloroplast_PNAS.

Unfortunately, the `rayDISC` function does not allow the users to set a prior on the transition rates, and thus we could not implement in the analysis the biological insight that transition rates should be low.

Binary data

Phytools

The "binary" dataset, in which brackish species are coded as marine, should help address the bias in the transition rates resulting from the presence of only two brackish species. As Nakov et al [1] did, we thus used this dataset to perform the same analyses that were performed on the multistate dataset.

We first analysed the data using the `make.simmap` function of the `phytools` package. For binary data, the SYM model is identical to the ER model, and thus we only tested the ER and ARD models.

As before, we set the prior on the root node as `equal` and we simulated 1000 stochastic maps with an `empirical` rate matrix. The results of these analyses are summarised in TABLE 12.

Model	Equal Rates (ER)		All-rates-different (ARD)	
Maximum likelihood estimate of the rate matrix	0 / 1		0 / 1	
	0: -0.507	0.507	0: -0.178	0.178
	1: 0.507	-0.507	1: 0.576	-0.576
Node	**P_f**	**P_m**	**P_f**	**P_m**
Root	0.620	0.380	0.087	0.913
G+A	0.595	0.405	0.050	0.950
A	0.607	0.393	0.059	0.941

Table 12: Results of the `phytools` analysis on the binary dataset.

The probabilities in this table are comparable to those obtained by Nakov et al. [1], and show that different models yield different conclusions also for the binary dataset (we cannot compare the transition rates, as Nakov et al. [1] did not report them).

As we did before, we analysed model fit using the AIC and BIC: the results are shown in TABLE 13.

Model	lnL	k	AICc	w_{AICc}	BIC	w_{BIC}
ER	-61.386	1	124.807	0.300	127.552	0.625
ARD	-59.506	2	123.115	0.700	128.570	0.375

Table 13: Results of the model fit analysis performed on the binary dataset.

We cannot compare these results with Nakov et al. [1] because they did not perform this analysis, but the table nevertheless shows that, once again, there is no clearly favoured model, and AIC and BIC rank the two models differently.

Model averaging analyses using SAICc and BMA are not conclusive either, as shown in tabella.

	Method			
	SAICc		BMA	
Rates	0 / 1		0 / 1	
	0: -0,277	0,277	0: -0,384	0,384
	1: 0,556	-0,556	1: 0,533	-0,533
Node	**P_f**	**P_m**	**P_f**	**P_m**
Root	0,205	0,795	0,410	0,590
G+A	0,136	0,864	0,332	0,668
A	0,156	0,844	0,323	0,677

Table 14: Model averaging analyses performed on the binary dataset using `phytools`.

It is therefore apparent that (as was the case with the multistate data) the binary data is not enough to confidently choose a single model or a single conclusion and it is still necessary to use biological information.

As before, we implemented the biological insight that transition rates should be low [6] by repeating the analysis with a strong prior on the transition rates (again, a beta distribution with parameters $\alpha = 1$ and $\beta = 400$). The results are shown in TABLE 15.

Model	Equal Rates (ER)		All-rates-different (ARD)	
Mean rates	0: -0.049, 0.049 1: 0.049, -0.049		0: -0.042, 0.042 1: 0.015, -0.015	
Node	P_f	P_m	P_f	P_m
Root	0.999	0.001	0.984	0.016
G+A	0.993	0.007	0.982	0.018
A	0.994	0.006	0.982	0.018

Table 15: Results of the analysis when a strong prior on low transition rates is used.

These results show that with low transition rates, a freshwater origin for all of the three nodes is favoured. The complete ancestral state reconstruction for the ARD model is shown in FIGURE 2, the one for the ER model is available from https://github.com/patrisan/Chloroplast_PNAS.

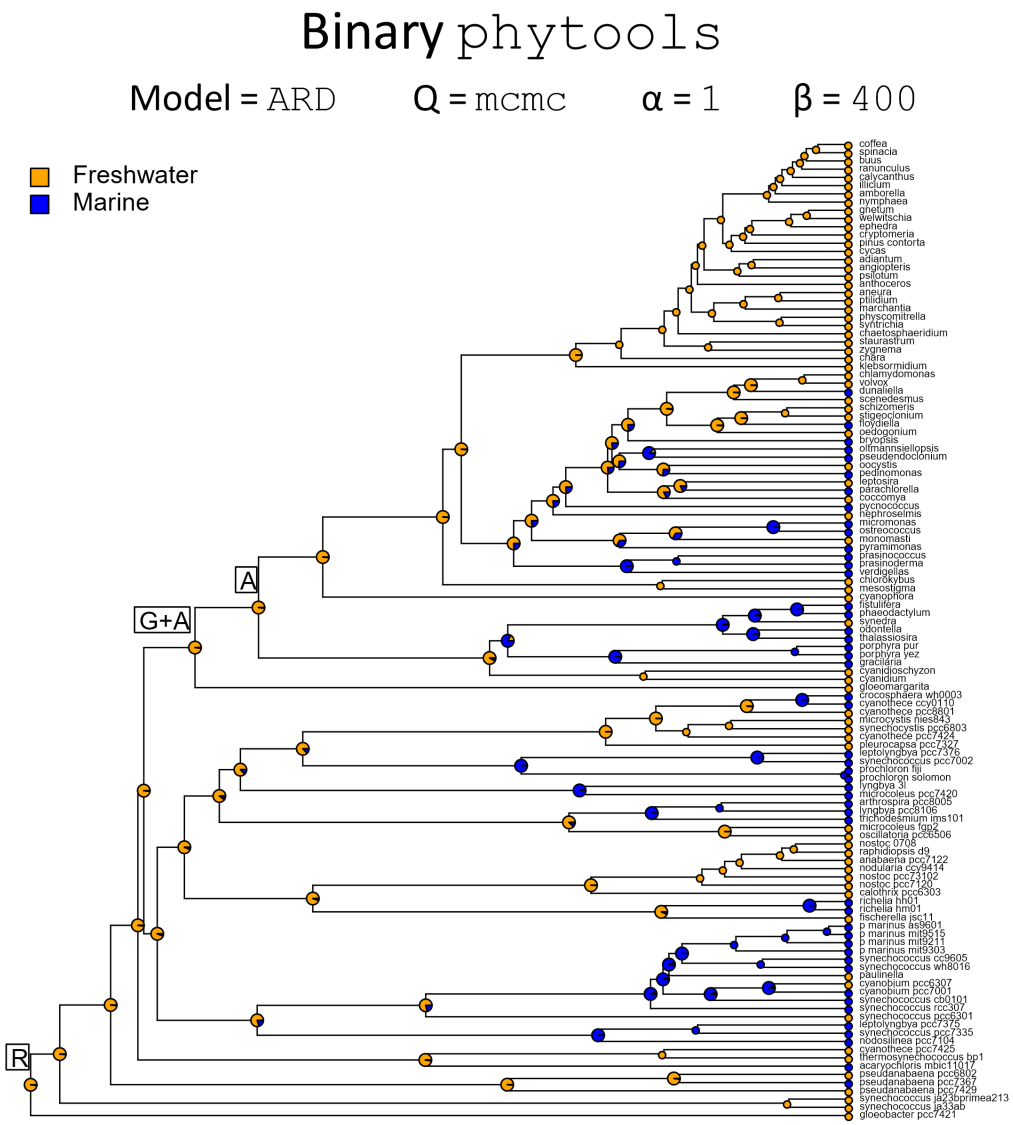

Figure 2: Ancestral state reconstruction for the ARD model with a strong prior on low transition rates.

As has been done for the multistate data, we also performed a Bayes factor analysis, whose results are shown in TABLE 16.

	Model			
	ER		ARD	
Node	**p_f**	**B_f**	**p_f**	**B_f**
Root	0,686	457,268	0,981	1,191
G+A	0,977	3,340	0,998	0,109
A	0,985	2,523	0,998	0,109

Table 16: Results of the Bayes factor analysis. p_f: prior probability of the node being freshwater, B_f: Bayes factor.

These figures show that, if a strong prior on low transition rates is assumed, the data does in fact support the freshwater origin of the nodes of interest for the ER model (as it provides *very strong* evidence that the root node lived in a freshwater habitat, *positive* evidence that the MRCA of *Gloeomargarita* and Archaeplastida was a freshwater taxon, and *barely worth mentioning* evidence that the ancestor of Archaeplastida was a freshwater species), but not for the ARD models (it appears that the data provides *barely worth mentioning* evidence that the root node was a freshwater taxon, and *positive* evidence that

the other two nodes lived in marine habitats, as the posterior freshwater probability is lower than the prior).

Once again, these results highlight that further studies are necessary to definitively tackle this problem.

CorHMM

Finally, we repeated the analysis of Nakov et al. [1] on the binary dataset using the `corHMM` package. TABLE 17 shows the results of this analysis.

Model	Equal Rates (ER)		All-rates-different (ARD)	
Maximum likelihood estimate of the rate matrix	0: -0,466 / 0,466 1: 0,466 / -0,466		0: -0,160 / 0,160 1: 0,575 / -0,575	
Node	**P_f**	**P_m**	**P_f**	**P_m**
Root	0.772	0.228	0.001	0.999
G+A	0.650	0.350	0.025	0.975
A	0.647	0.353	0.037	0.963

Table 17: Results of the `corHMM` analysis on the binary dataset.

These results are similar to those obtained using `phytools` and those reported by Nakov et al. [1], and show once again that different models provide very different results.

To assess which model best fits the data, we performed a likelihood analysis (TABLE 18).

Model	lnL	k	AICc	w_{AICc}	BIC	w_{BIC}
ER	-61.348	1	124.731	0.212	127.476	0.510
ARD	-59.000	2	122.103	0.788	127.558	0.490

Table 18: Results of the model fit analysis performed on the binary dataset.

This analysis shows one more time that the data alone does not clearly favour any one of the two models.

Lastly, we performed SAICc and BMA model averaging analyses in `corHMM`, which provided the numbers shown in TABLE 19.

	Method			
	SAICc		BMA	
Rates	0: -0,225 / 0,225 1: 0,552 / -0,552		0: -0,316 / 0,316 1: 0,520 / -0,520	
Node	**P_f**	**P_m**	**P_f**	**P_m**
Root	0,005	0,995	0,059	0,941
G+A	0,049	0,951	0,150	0,850
A	0,070	0,930	0,178	0,822

Table 19: Model averaging analyses performed on the binary dataset using `corHMM`.

The results in this case lean a bit towards a marine origin of the root node, but are still not conclusive.

Conclusions

Nakov et al. [1] have shown that using different models in reconstructing the habitat of the ancestors of chloroplast can yield fundamentally different results. In their analysis [1] they suggest that the data mostly supports models that result in a marine ancestry for the three nodes of relevance (the root node, the MRCA of Archaeplastida and *Gloeomargarita* and the MRCA of Archaeplastida).

Extending their analyses while still using the same framework they used, however, we have shown that the data does not support a single model. Most importantly without considering biological information, no reliable conclusion can be made from the statistical analyses alone.

While not many studies have looked into transition rates between freshwater ↔ marine taxa across the tree of life, it has been noted that transitions across salinity barriers are likely rare [6]. By incorporating this biological insight in the models that we used, our analyses show that when low transition rates are assumed, a clear freshwater origin for the three nodes emerges.

Understanding the evolution of ecological preferences across photosynthetic eukaryotes will require more research. It can be argued that by looking at one single character this could be an oversimplification, and further studies should expand these analyses taking into several traits describing the mechanisms behind the biosynthetic pathways responsible for 'salt tolerance', and a Bayesian approach to accurately assess the results.

References

[1] T. Nakov, J. Boyko, A. Alverson e J. Beaulieu, «Ecology of Primary Endosymbiosis,» 2017. [Online]. Available: https://github.com/teofiln/Ecology-of-Primary-Endosymbiosis.

[2] S.-B. P, R. JA, P. D e K. AH, «Early photosynthetic eukaryotes inhabited low-salinity habitats,» *Proc Natl Acad Sci USA,* n. 114, pp. E7737-7745, 2017.

[3] S.-B. P, R. JA, P. D e K. AH, «Data from: Early photosynthetic eukaryotes inhabited low-salinity habitats,» 17 08 2017. [Online]. Available: http://datadryad.org/resource/doi:10.5061/dryad.421p2.

[4] L. Revell, «Simulate stochastic character maps on a phylogenetic tree or trees,» [Online]. Available: http://www.phytools.org/static.help/make.simmap.html. [Consultato il giorno 10 11 2017].

[5] N. T, B. JB, A. AJ e B. JM, «Models with unequal transition rates favor marine origins of cyanobacteria and photosynthetic eukaryotes,» *Proc Natl Acad Sci USA,* 2017.

[6] R. Logares, J. Bråte, S. Bertilsson, J. Clasen, K. Shalchian-Tabrizi e K. Rengefors, «Infrequent marine–freshwater transitions in the microbial world,» *Trends Microbiol,* n. 17, pp. 414-422, 2009.

[7] R. Kass e A. Raftery, «Bayes Factors,» *Journal of the American Statistical Association,* vol. 90, n. 430, pp. 773-795, 1995.

[8] H. Jeffreys, The Theory of Probability, Oxford, 1961, p. 432.

Scripts

This section describes the scripts and data files that were used to produce the results that have been presented. All the files can be download from https://github.com/patrisan/Chloroplast_PNAS.

sp_tree.tre

Calibrated molecular clock tree that was used as the tree underlying the ancestral state reconstruction.

Matrix_states.csv

Data file containing the habitats for the 119 taxa, for both the binary and multistate coding.

make.simmap.R

This script is used to override the `make.simmap` function of the `phytools` package and allow it to report the log-maximum likelihood of the selected model.

plotSimmap.r

This script runs a simmap analysis on both the binary data and the multistate data, for a given model. The output consists of two PDF files with the ancestral state reconstruction plots and of the rate matrices, log-maximum likelihoods and probabilities for the three key nodes, which are printed to the standard output. It should be invoked this way:

```
Rscript plotSimmap.r sp_tree.tre Matrix_states.csv <model> <seed>
```

We used 20170921 as seed (in order to match what has been done by Nakov et al. [1].

plotSimmapBinarySAICc.r
plotSimmapBinaryBMA.r
plotSimmapMultistateSAICc.r
plotSimmapMultistateBMA.r

These scripts run the model-averaged simmap analyses. The output consists of a pdf file with the ancestral state reconstruction plot and of the rate matrix and probabilities for the three key nodes, which are printed to the standard output. They should be invoked this way:

```
Rscript <script name> sp_tree.tre Matrix_states.csv <seed>
```

Again, we used 20170921 as seed. Note that the transition rates are hard-coded in the scripts.

plotSimmap_MCMC.r

This script is used to run simmap analyses with the transition rate matrix sampled from a posterior distribution with a specified prior, for a given model. The output consists of two PDF files with the ancestral state reconstruction plots, of the average transition rate matrices and of the probabilities for the three key nodes, which are printed to the standard output. It should be invoked this way:

```
Rscript plotSimmap_MCMC.r sp_tree.tre Matrix_states.csv <model> <seed> <alpha> <beta>
```

Where `alpha` and `beta` are the parameters of the beta distribution that is used as prior for the transition rates. The seed that we used was still 20170921.

plotCorHMM.r

This script runs the `corHMM` analyses with both datasets and all models. The output consists of 11 pdf files with ancestral state reconstructions and of the rate matrices, log-maximum likelihoods and probabilities for the three key nodes, which are printed to the standard output. It should be invoked this way:

```
Rscript plotCorHMM.r sp_tree.tre Matrix_states.csv <seed>
```

We still used 20170921 as seed.

plotCorHMM_MA.r

These scripts run the model-averaged `corHMM` analyses. The output consists of 4 pdf files with the ancestral state reconstruction plots and of the rate matrices and probabilities for the three key nodes, which are printed to the standard output. It should be invoked this way:

```
Rscript plotCorHMM_MA.r sp_tree.tre Matrix_states.csv <seed>
```

Once more, we used 20170921 as seed. Note that the transition rates are hard-coded in the scripts.

Archaeplastida.txt

This file contains the Archaeplastida species that have been included in the analysis.

OtherSpecies.txt

This file contains the species that are not Archaeplastida that were included in the analysis (except for *Gloeomargarita* and *Gloeobacter*).

generateTopologies.r

This script is used to generate random topologies. It should be called like this:

```
Rscript generateTopologies.r <number of trees> <output file> <seed>
```

Again, we used 20170921 as seed.

Infer.Runtime.dll
InferNetLicense_2.6.pdf

Infer.NET library, that is used in the following script to get the Dirichlet and Gamma distributions, and relative licence.

SetBranchLengths.cs

Source code of the program that assigns branch lengths to the topologies created by the generateTopologies.r script. After having been appropriately compiled (referencing the included Infer.Runtime assembly), the program should be called like this:

```
SetBranchLengths.exe <tree file> <root age mean> <root age std deviation> <root age
                     multiplier> <output file>
```

To match what had been done in the original paper by Sánchez-Baracaldo et al. [1], we used a root age mean of 2500, root age standard deviation of 100 and root age multiplier of 0.001.

prior_distrib_binary.r
prior_distrib_multistate.r

These scripts are used to run the simmap analysis on the prior distribution of trees. Each script performs the analysis on 100 random trees, once per tree, and outputs the resulting stochastic maps to a file named respectively `binary.maps` or `multistate.maps`. To speed up the analysis, 10 instances of each of these scripts were run in parallel for each model, and the resulting maps files were concatenated and then read using the next script. The scripts should be invoked this way:

```
Rscript <script name> sp_tree.tre Matrix_states.csv <model> <seed> <alpha> <beta>
```

We used a different seed each time we ran the scripts; the seeds were of the form `2017abc`, where `a` was 1 for ER models, 2 for ARD models and 3 for SYM models, `b` was 2 for the binary dataset and 3 for the multistate dataset and `c` was a number between 0 and 9, depending on the run.

GetProbs.cs

This script is used to read the simmap files originated by the previous scripts and print to the standard output the probabilities for the “interesting” nodes. After having been appropriately compiled, it should be invoked this way:

```
GetProbs.exe <maps file>
```