RH: SOWHAT?

Appendix 1

RAxML results and the BFGS optimization

We compared the results of SOWH tests performed using the same data, model, and sample size, but different likelihood software (GARLI and RAxML) for six datasets (Table 2). For the Buckley dataset we performed this comparison using two different models for each likelihood software tool. Under BFGS optimization (the RAxML default), the choice of likelihood software had an effect on the outcome for two datasets.

In order to understand the underlying cause of the different outcomes when using RAxML and GARLI, we compared the null distributions and test statistics generated by each of the SOWH tests performed using GARLI and RAxML (Appendix Table 1). We did not observe differences in the calculated test statistics on the real data between RAxML and GARLI that would change the outcome of the test, suggesting that this component of the analysis is not responsible for the discrepancies. Initial results instead indicated that the differences in significance between programs are due to irregularities in RAxML performance on the simulated data used to generate the null distribution. For every SOWH test performed, the range of the null distribution generated using RAxML was larger than the range generated using GARLI. This is because RAxML occasionally returned large δ values in both tails of the distribution relative to those returned by GARLI. The presence of these large values in the left tail of the null distribution signifies that for certain simulated datasets RAxML returns a much lower likelihood score using an unconstrained search than using a constrained search. This indicated that RAxML was not identifying the most likely tree in the unconstrained searches on simulated data, as the constrained

topology is within the treespace of the unconstrained search. Following consultation with the authors of RAxML, the problem was identified as a failure of the BFGS optimization of model parameters. Rerunning the tests using RAxML with the optimization suppressed resulted in universally smaller ranges of null distribution (Table 3) and the outcomes of the tests run under RAxML were the same as those under GARLI.

Table 1: **Effects of Likelihood Software on Null Distributions**

Dataset	Model	ML Software	P-value	Test Stat.	Null Distribution					
					0%	25%	50%	75%	100%	Range
Buckley	GTR+I+Γ	GARLI	0.018*	4.703	-0.147	0.102	0.831	1.775	8.022	8.168
Buckley	GTR+I+Γ	RAxML	0.025*	4.875	-55.879	<0.001	0.010	1.118	73.743	129.622
Buckley	GTR+I+Γ	RAxML (No BFGS)	0.010*	4.873	-1.076	0.000	0.534	1.489	9.530	10.605
Buckley	GTR+Γ	GARLI	0.118	4.888	-0.860	0.165	1.348	3.254	17.192	18.052
Buckley	GTR+Γ	RAxML	0.241	4.961	-21.136	0.214	2.385	4.787	28.437	49.573
Buckley	GTR+Γ	RAxML (No BFGS)	0.252	4.871	-1.581	0.499	2.566	4.873	23.381	24.962
Dunn	GTR+I+Γ	GARLI	<0.01**	21.796	-0.023	<0.01	0.279	1.101	4.499	4.522
Dunn	GTR+I+Γ	RAxML	<0.01**	17.810	-4.319	<0.01	0.767	2.808	8.592	12.911
Dunn	GTR+I+Γ	RAxML (No BFGS)	<0.01**	16.427	-6.292	0.015	0.707	2.062	5.276	11.569
Edwards	GTR+I+Γ	GARLI	<0.01**	9.684	-0.001	<0.01	<0.01	<0.01	0.008	0.009
Edwards	GTR+I+Γ	RAxML	<0.01**	9.732	-0.039	-0.003	-0.002	<0.01	0.019	0.058
Edwards	GTR+I+Γ	RAxML (No BFGS)	<0.01**	9.731	-0.033	-0.004	-0.002	<0.01	0.009	0.041
Liu	GTR+I+Γ	GARLI	<0.01**	335.653	-0.001	<0.01	<0.01	0.298	4.009	4.010
Liu	GTR+I+Γ	RAxML	<0.01**	335.620	-0.041	-0.001	0.002	0.094	4.018	4.059
Liu	GTR+I+Γ	RAxML (No BFGS)	<0.01**	335.628	-1.805	0.000	0.002	0.304	4.831	6.637
Sullivan	GTR+I+Γ	GARLI	<0.01**	8.828	<0.01	0.320	1.354	2.720	7.434	7.434
Sullivan	GTR+I+Γ	RAxML	0.290	8.413	-560.827	0.612	2.139	10.575	581.588	1,142.414
Sullivan	GTR+I+Γ	RAxML (No BFGS)	<0.01**	8.466	-1.308	0.155	0.730	1.957	7.049	8.357
Wang	GTR+Γ	GARLI	<0.005**	12.542	-0.002	0.000	0.200	1.169	8.442	8.444
Wang	GTR+Γ	RAxML	0.026*	13.018	-652.832	0.012	0.269	1.515	780.203	1,433.035
Wang	GTR+Γ	RAxML (No BFGS)	<0.005**	13.026	-2.802	0.005	0.190	1.262	10.337	13.139

Notes: A SOWH test was performed using GARLI and RAxML for each dataset. RAxML initially resulted in a larger total range of the null distribution for all datasets. Large values in the left tail of the distribution, as seen in the Sullivan and Wang analyses, indicate failure of RAxML to find the most likely topology. Following consultation with the authors of RAxML, a SOWH test was performed using RAxML with the BFGS routine suppressed for each dataset, which resulted in a smaller range in all cases and no extreme values in the tails. The outcomes of the tests between GARLI and RAxML with the BFGS routine suppressed are equivalent. Test statistics varied between SOWH tests, though this variance alone would not have had an effect on the final outcome.