Title page

Article title: The Effect of Phylogenetic Uncertainty and Imputation on EDGE Scores

Running Head: Effects of Phylogenetic Uncertainty and Imputation on EDGE

Authors: K. Bodie Weedop[1*], Arne Ø. Mooers[2], Caroline M. Tucker[3], and William D. Pearse[1]

[1] Department of Biology & Ecology Center, Utah State University, 5305 Old Main Hill, Logan UT, 84322

[2] Department of Biological Sciences, Simon Fraser University, Burnaby, British Columbia, Canada

[3] Department of Biology, University of North Carolina–Chapel Hill

[*]To whom correspondence should be addressed: K. Bodie Weedop (kbweedop@gmail.com)

Supporting Information

Appendix S1: Effect of Measures of the True, Full Phylogenies

This section contains analyses of a number of metrics from the original phylogenies and their effect on the correlation coefficient of ED values described in body of the manuscript.

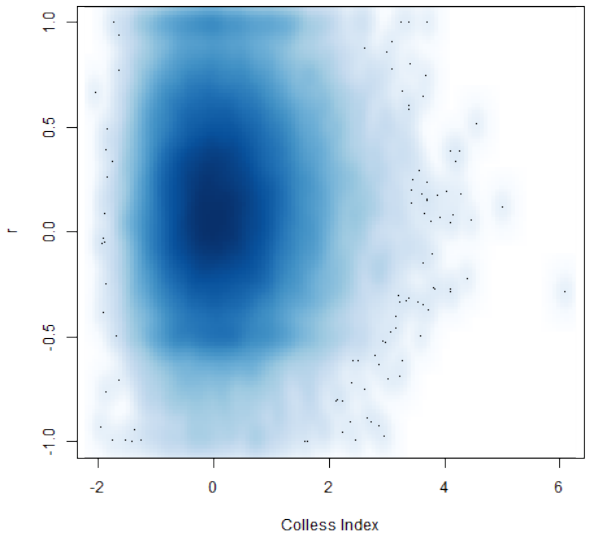

Figure S1.1: Effect of the True Colless Index of Full Phylogeny. Smoothed color density plot of the effect of colless' index (prior to imputation) on the correlation coefficient of ED values . Colless' index is a measure of how imbalanced a tree is using the summed differences of the number of clades in each pair of taxa.

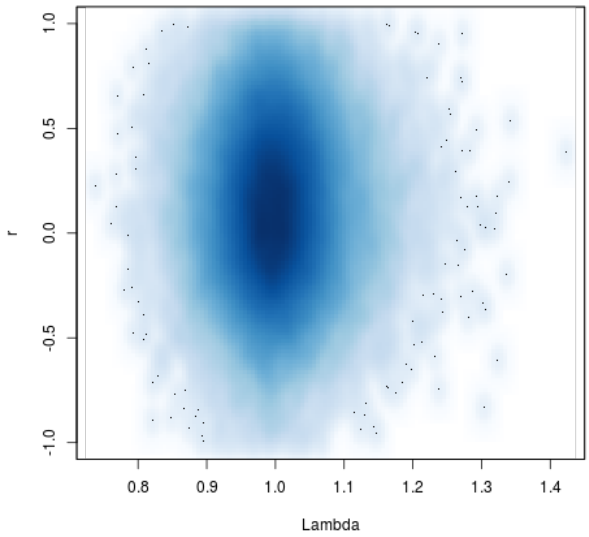

Figure S1.2: Effect of the True Lambda of Full Phylogeny. Smoothed color density plot of estimated speciation rate (λ; prior to imputation) and its effect on the correlation coefficient of ED values.

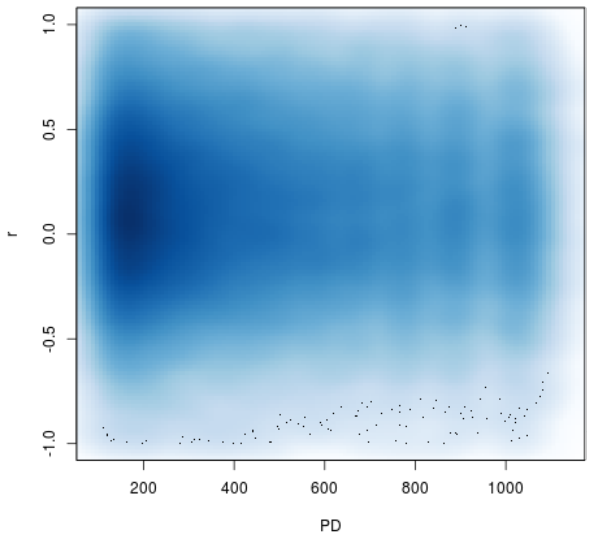

Figure S1.3: Effect of True PD of Full Phylogeny. Smoothed color density plot of the total phylogenetic diversity (prior to imputation) and its effect the correlation coefficient of ED values.

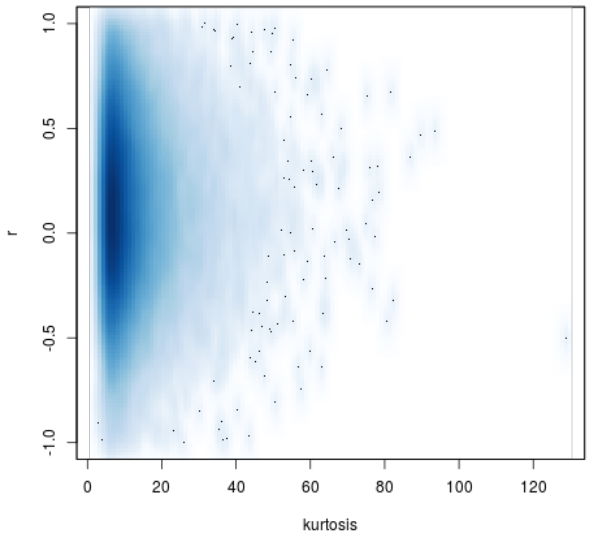

Figure S1.4: Effect of the True Kurtosis of Full Phylogeny. Smoothed color density plot of kurtosis of ED values (prior to imputation) and its effect the correlation coefficient of ED values.

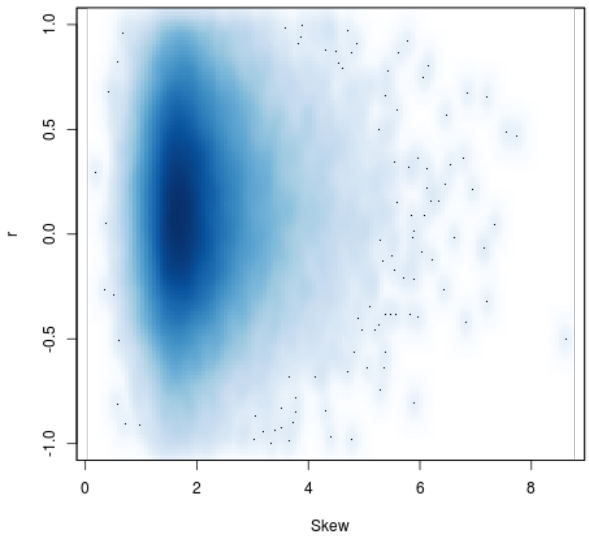

Figure S1.5: Effect of the True Skew of Full Phylogeny. Smoothed color density plot of the skew of ED values (prior to imputation) and its effect the correlation coefficient of ED values.

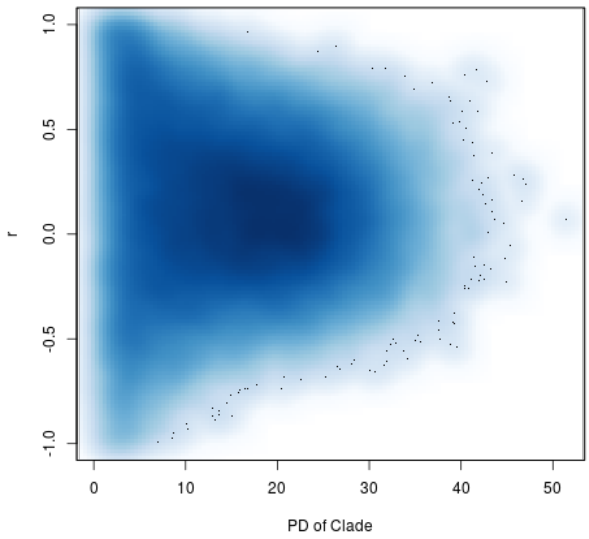

Figure S1.6: Effect of the True PD of The Selected Clade. Smoothed color density plot of the total phylogenetic diversity of the focal clade (prior to imputation) and its effect the correlation coefficient of ED values.

Appendix S2: Error Rate in Top Rankings

This section contains plots of mean error rates in ED rankings of the top species overall.

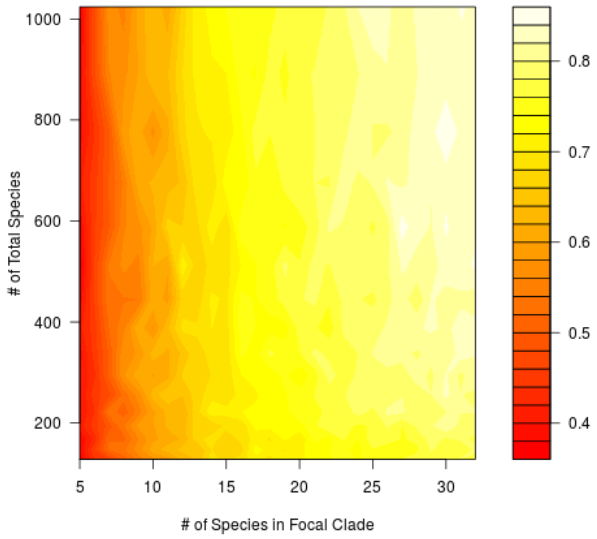

Figure S2.7: Mean error rate in the ranking of top 50 species. Heat map plot demonstrating the average error rate in ED ranking of the top 50 species when imputing a clade of a specified size. The gradient on the right gives the average error rate and the corresponding color.

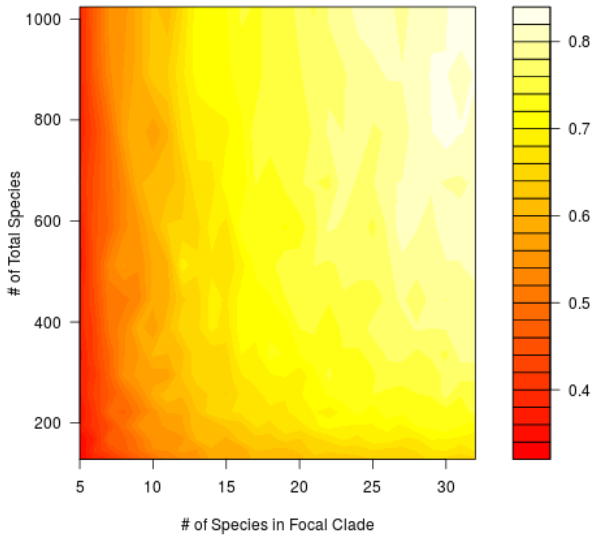

Figure S2.8: Mean error rate in the ranking of top 100 species. Heat map plot demonstrating the average error rate in ED ranking of the top 100 species when imputing a clade of a specified size. The gradient on the right gives the average error rate and the corresponding color.

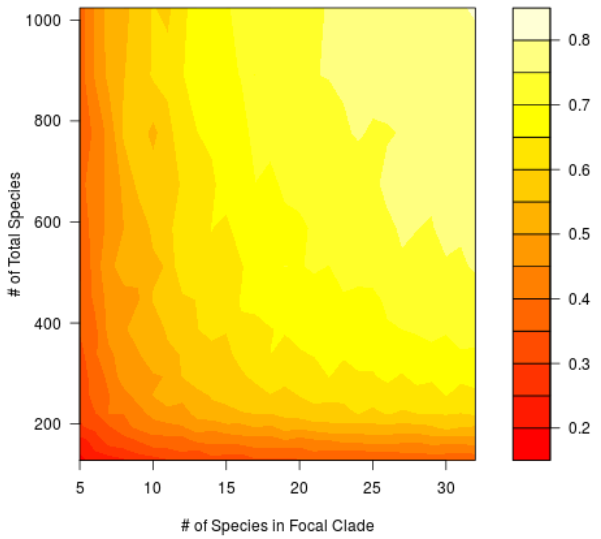

Figure S2.9: Mean error rate in the ranking of top 200 species. Heat map plot demonstrating the average error rate in ED ranking of the top 200 species when imputing a clade of a specified size. The gradient on the right gives the average error rate and the corresponding color.

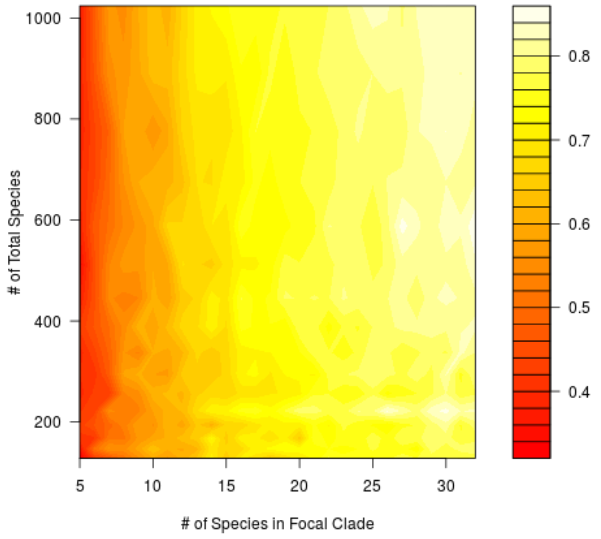

Figure S2.10: Mean error rate in the ranking of top 5% of species. Heat map plot demonstrating the average error rate in ED ranking of the top 5% of all species in the phylogeny when imputing a clade of a specified size. The gradient on the right gives the average error rate and the corresponding color.

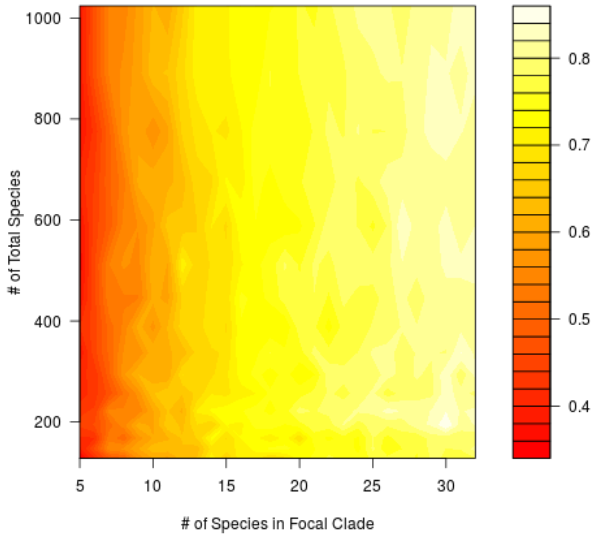

Figure S2.11: Mean error rate in the ranking of top 10% of species. Heat map plot demonstrating the average error rate in ED ranking of the top 10% of all species in the phylogeny when imputing a clade of a specified size. The gradient on the right gives the average error rate and the corresponding color.

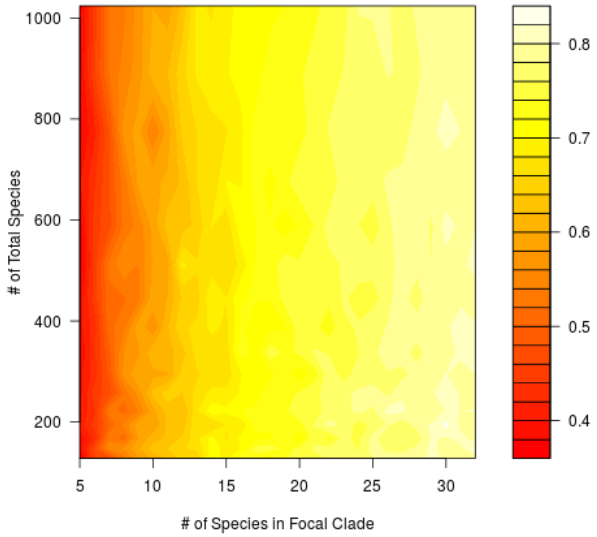

Figure S2.12: Mean error rate in the ranking of top 20% of species. Heat map plot demonstrating the average error rate in ED ranking of the top 20% of all species in the phylogeny when imputing a clade of a specified size. The gradient on the right gives the average error rate and the corresponding color.

Appendix S3: Ranking Error When Using Average ED Value

This section contains a plot of the mean ranking error when species in the focal clades were assigned the average ED value of the clade.

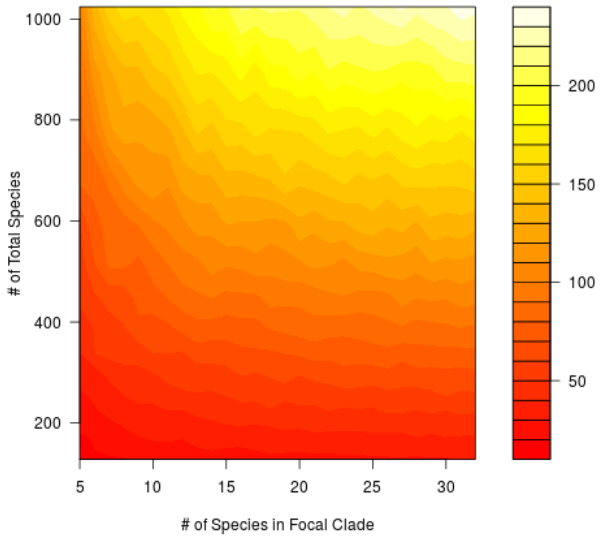

Figure S3.13: Mean Ranking Error of Species Assigned Average ED. Heat map plot demonstrating the average ranking error when species in the focal clade were assigned the average ED value of the clade. The gradient on the right demonstrates average number of positions within the full ranking that focal clade species shifted from their true rank.

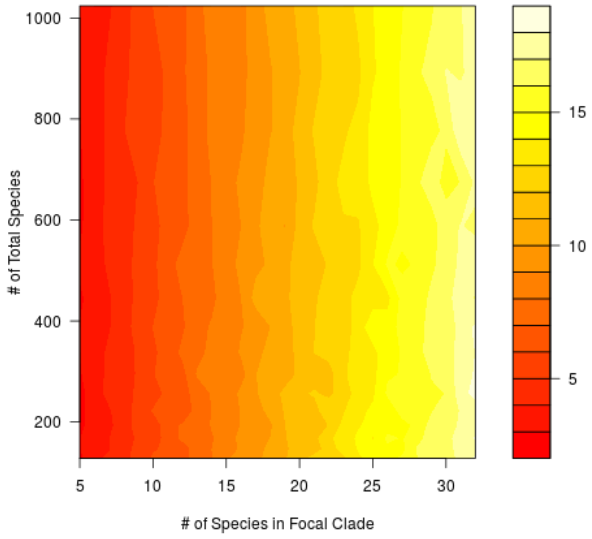

Figure S3.14: Mean Ranking Error of Non-imputed Species. Heat map plot demonstrating the average ranking error of non-imputed species when species in the focal clade were imputed. The gradient on the right demonstrates average number of positions within the full ranking that focal clade species shifted from their true rank.

Appendix S4: Figures and tables of results when using a small amount (d=0.5) of extinction

This section contains all plots and tables that are associated with the alternative simulations which were performed incorporating past extinction. The underlying analyses of all plots and graphs here are identical to those that can be seen in the main text and preceding Supporting Information appendices.

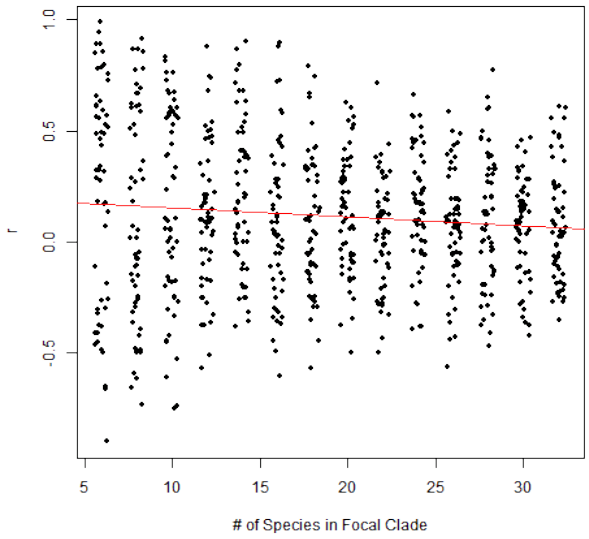

Figure 1: Figure S4.15: The correlation between species' imputed and true ED scores plotted as a function of the number of species imputed (focal clade size from all sizes of phylogenies used (n = 256, ..., 1024)). Each data point represents the correlation between ED values within the focal clades where imputation has occurred, comparing species' true ED values with their imputed ED values. This plot, and the statistical analysis of it in table S4.1, show limited support for an association between true and imputed ED values.

	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	-0.0724	0.5205	-0.14	0.8894
Size of Focal Clade	-0.0042	0.0030	-1.43	0.1531
Size of Phylogeny	0.0010	-0.01	0.9888	
PD	0.0001	0.0007	0.18	0.8544
Estimated speciation rate	0.4050	0.7068	0.57	0.5669
Colless' Index	-0.0000	0.0000	-0.92	0.3577
Skew	-0.0681	0.0675	-1.01	0.3133
Kurtosis	0.0083	0.0063	1.32	0.1878
Depth of Imputed Clade	0.0001	0.0020	0.06	0.9503

Table 1: Table S4.1: Statistical model of the potential drivers of the correlation between imputed and true ED values when incorporating low (d=0.5) amounts of extinction. Results of a multiple regression fitted to the data shown in figure , showing a relatively poor correlation between imputed and true ED scores ($F_{761,8} = 1.49$, $r^2 = 0.005$, $p < 0.157$). Given the extremely low predictive power, just as in the case of the pure birth model, of this statistical model we are reticent to make strong claims about drivers of the correlation between imputed and observed ED.

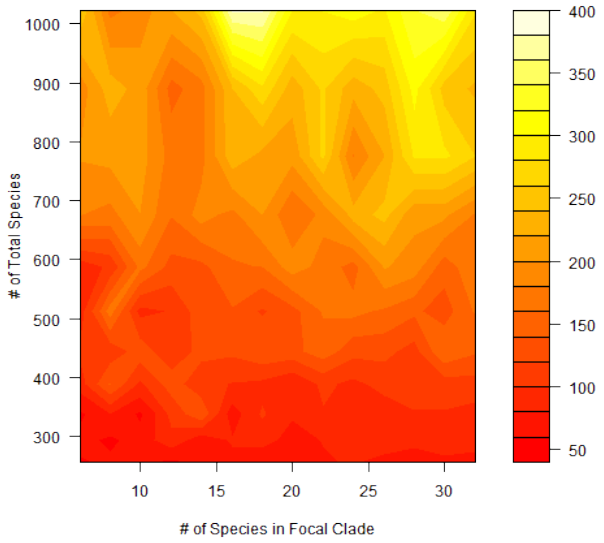

Figure 2: Figure S4.16: Mean ranking error of imputed species when incorporating low (d=0.5) amounts of extinction. An interpolated heat-map of the mean ranking error of imputed species as a function of the total number of species in the phylogeny (vertical axis) and number of species in the focal (imputed) clade (horizontal axis). Table S4.2 gives statistical support for the trend of increased error in larger phylogenies and imputed clades.

	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	-1.6783	0.3832	-4.38	0.0000
Size of focal (imputed) clade	0.0779	0.0091	8.60	0.0000
Size of phylogeny	0.5243	0.0144	36.31	0.0000

Table 2: Table S4.2: Statistical model of the effect of clade and phylogeny size on ranking error when incorporating low (d=0.5) amounts of extinction. Model of the raw data underlying figure S4.16, regressing the ranking error of imputed species against the number of species in the imputed clade and the entire phylogeny ($F_{767,2} = 696.3$, $r^2 = 0.6439$, $p < 0.0001$). As can be seen in figure S4.16, the average ranking error is positively correlated with the size of the clade being imputed and the entire phylogeny. Square-root transformations have been applied to both ranking error and size of phylogeny.

Appendix S5: Figures and tables of results when using a large amount (d=0.95) of extinction

This section contains all plots and tables that are associated with the alternative simulations which were performed incorporating past extinction. The underlying analyses of all plots and graphs here are identical to those that can be seen in the main text and preceding Supporting Information appendices.

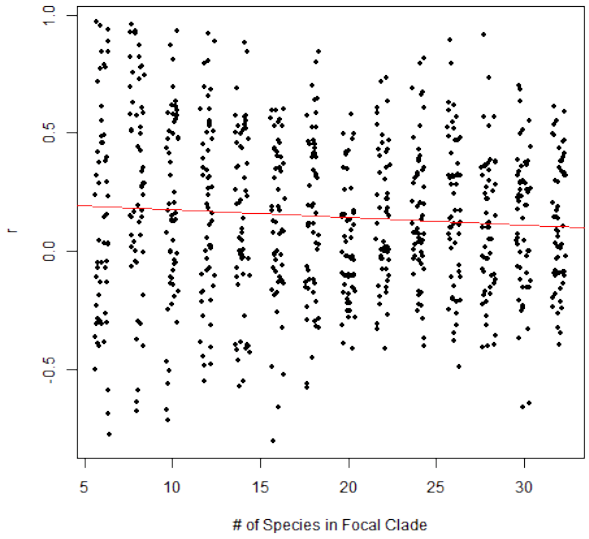

Figure 3: Figure S5.17: The correlation between species' imputed and true ED scores plotted as a function of the number of species imputed (focal clade size from all sizes of phylogenies used (n = 256, ..., 1024)). Each data point represents the correlation between ED values within the focal clades where imputation has occurred, comparing species' true ED values with their imputed ED values. This plot, and the statistical analysis of it in table S5.3, show limited support for an association between true and imputed ED values.

	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	0.5230	0.1825	2.87	0.0043
Size of Focal Clade	-0.0065	0.0023	-2.77	0.0057
Size of Phylogeny	0.0002	0.0003	0.71	0.4786
PD	-0.0001	0.0001	-1.17	0.2429
Estimated speciation rate	-0.7027	0.4634	-1.52	0.1298
Colless' Index	0.0000	0.0000	0.56	0.5746
Skew	-0.0200	0.0254	-0.79	0.4307
Kurtosis	0.0014	0.0015	0.99	0.3237
Depth of Imputed Clade	0.0014	0.0008	1.85	0.0649

Table 3: Table S5.3: Statistical model of the potential drivers of the correlation between imputed and true ED values when incorporating high (d=0.95) amounts of extinction. Results of a multiple regression fitted to the data shown in figure S5.17, showing a relatively poor correlation between imputed and true ED scores ($F_{761,8} = 1.53$, $r^2 = 0.006$, $p < 0.142$). Given the extremely low predictive power, just as in the case of the pure birth model, of this statistical model we are reticent to make strong claims about drivers of the correlation between imputed and observed ED.

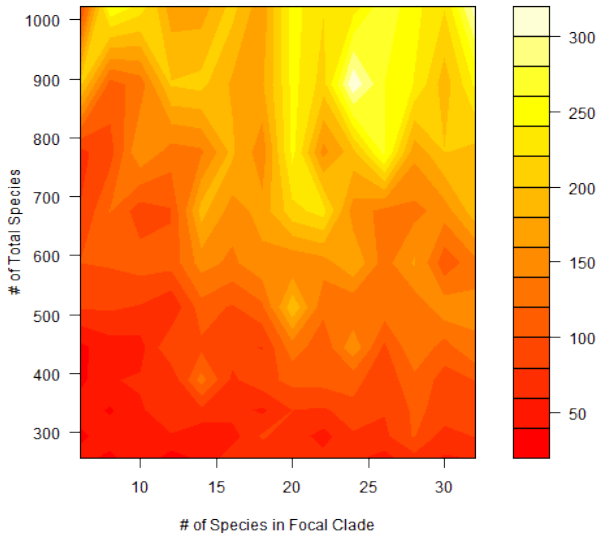

Figure 4: Figure S5.18: Mean ranking error of imputed species when incorporating high (d=0.95) amounts of extinction. An interpolated heat-map of the mean ranking error of imputed species as a function of the total number of species in the phylogeny (vertical axis) and number of species in the focal (imputed) clade (horizontal axis). Table S5.4 gives statistical support for the trend of increased error in larger phylogenies and imputed clades.

	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	-2.5074	0.4739	-5.29	0.0000
Size of focal (imputed) clade	0.1408	0.0112	12.58	0.0000
Size of phylogeny	0.4498	0.0179	25.19	0.0000

Table 4: Table S5.4: Statistical model of the effect of clade and phylogeny size on ranking error when incorporating low (d=0.95) amounts of extinction. Model of the raw data underlying figure S5.18, regressing the ranking error of imputed species against the number of species in the imputed clade and the entire phylogeny ($F_{767,2} = 396.4$, $r^2 = 0.507$, $p < 0.0001$). As can be seen in figure S5.18, the average ranking error is positively correlated with the size of the clade being imputed and the entire phylogeny. Square-root transformations have been applied to both ranking error and size of phylogeny.