Estimates of Phylogenetic Signal Based on Lambda are Often Inaccurate

Keywords: Pagel's lambda, phylogenetic signal

Short Title: Inaccuracies in Pagel's Lambda

Abstract

{conclusion holds: interpreting the regression is not appreciably different (in terms of slopes and f values)}

Introduction

Investigating macroevolutionary patterns of trait variation requires a phylogenetic perspective, because the shared ancestry among species generates statistical non-independence (Felsenstein 1985; Harvey and Pagel 1991). Accounting for this evolutionary non-independence is the purview of *phylogenetic comparative methods* (PCMs); a suite of analytical tools that condition the data on the phylogeny through the course of statisical evaluations of phenotypic trends (e.g., Grafen 1989; Garland and Ives 2000; Rohlf 2001; Butler and King 2004). The past several decades have witnessed a rapid expansion in the development of PCMs to address an ever-growing set of macroevolutionary hypotheses (Martins and Hansen 1997; O'Meara et al. 2006; Revell and Harmon 2008; Beaulieu et al. 2012; Adams 2014b,a; Adams and Collyer 2018). These methods are predicated on the notion that phylogenetic signal – the tendancy for closely related species to display similar trait values – is present in cross-species datasets (Felsenstein 1985; Pagel 1999; Blomberg et al. 2003). Indeed, under numerous evolutionary models, phylogenetic signal is to be expected, as stochastic character change along the hierarchical structure of the tree of life generates trait covaration among related taxa (see Felsenstein 1985; Blomberg et al. 2003; Revell et al. 2008).

For many macroevolutionary analyses, it is often of interest to quantify the degree to which phylogenetic signal is displayed in continuous traits. Several analytical tools have been developed for this purpose (e.g., Gittleman and Kot 1990; Abouheif 1999; Pagel 1999; Blomberg et al. 2003; Klingenberg and Gidaszewski 2010; Adams 2014a), which differ primarily in how they characterize the phylogenetic dependency of trait variation among taxa. One commonly used statistical measure, *Kappa* (K), expresses the strength of phylogenetic signal as the ratio of observed trait variation to the trait variation conditioned on the phylogeny; scaled by what is expected under Brownian motion given the phylogeny's size and shape (Blomberg et al. 2003; also Adams 2014a). Another approach, Pagel's λ (Pagel 1999), uses maximum likelihood to fit the data to the phylogeny under some model of evolutionary change (typically Brownian motion). The inclusion of a scaling parameter, λ, transforms the lengths of the internal branches of the phylogeny to improve the fit, and this parameter describes the degree of phylogenetic signal in the dataset (Pagel 1999; Freckleton et al. 2002). Pagel's λ also has the advantage that it may be included when estimating the association of traits in a phylogenetic context, meaning that one may account for the degree of phylogenetic signal while conducting phylogenetic regression or ANOVA (see Freckleton et al. 2002).

Several studies have investigated the statistical properties of methods for estimating phylogenetic signal under various conditions (e.g., Munkemuller et al. 2012; Pavoine and Ricotta 2012; Diniz-Filho et al. 2012; Molina-Venegas and Rodriguez 2017; see also Revell et al. 2008; Revell 2010). These have largely focused on the ability of methods to detect the presence of phylogenetic signal (i.e., type I and type II error rates) under complex models of evolutionary change, across a range of phylogeny sizes, and with varying degrees of phylogenetic uncertainty or unresolved topologies. In terms of parameter estimation, Revell (2010) found that regression parameters were accurately estimated when λ was included during phylogenetic regression, and Munkmuller et al. (2012) demonstrated that estimates of phylogenetic signal obtained using various measures generally increased when input levels of phylogenetic signal were stronger. However, the precision of those estimates could not be determined, because the input levels of phylogenetic signal were simulated via a scaling factor (w) that was not directly comparable to the measures of phylogenetic signal being compared (see Munkemuller et al. 2012). One study (Boettiger et al. 2012) found that estimates of Pagel's λ displayed less variation when data were simulated on a large phylogeny ($N = 281$) as compared to a small one ($N = 13$), concluding that insufficient data (i.e., the number of species) was the underlying cause of the lack of precision in parameter estimation. However, this conclusion assumes that the precision of parameter estimation remains constant across the range of values ($\lambda = 0 \rightarrow 1$); an assumption that to date, has not been verified. Thus, despite widespread use of Pagel's (1999) λ in macroevolutionary studies, at present, we still lack a general understanding of the precision with which λ can estimate levels of phylogenetic signal in phenotypic datasets.

In this study, we evaluate the precision of Pagel's λ to estimate known levels of phylogenetic signal in phenotypic data. First we use computer simulations across differing numbers of species, differently shaped phylogenies, and differing input levels of phylogenetic signal, to explore the degree to which λ correctly identifies known levels of phylogenetic signal, and under what circumstances. For comparison, we also evaluate estimates of $Kappa$ and its empirically-derived effect size (Z_K) for this purpose. Additionally, we use simulations to determine how the inclusion of λ in phylogenetic regression and ANOVA (i.e., PGLS) affects parameter estimation, and whether levels of phylogenetic signal estimated in a PGLS framework are accurate. We then survey the recent macroevolutionary literature for published papers containing estimates of λ from empirical datasets, and compare these empirical estimates to patterns gleaned from our computer simulations. In general we find that while PGLS parameters (e.g., β) are accurately estimated with the inclusion of phylogenetic signal, estimates of λ are not. Further, we find that estimates of λ vary widely for a given input value of phylogenetic signal, and that this variation is not constant across the range of input signal, with decreased precision when phylogenetic signal is of intermediate strength. Alternatively, variation across effect sizes (Z_K) obtained from $Kappa$ is far more uniform across the range of input values, and estimates of phylogenetic signal with this measure increase in a uniform manner with increasing phylogenetic signal. We conclude that estimates of phylogenetic signal using Pagel's λ are often innacurate, and thus interpreting strength of phylogenetic signal in phenotypic datasets based on this measure should be treated with caution. By contrast, effect sizes obtained from $Kappa$ hold promise for characterizing phylogenetic signal, and comparing the strength of phylogenetic signal across datasets.

Materials and Methods

Simulations

We conducted a series of computer simulations to evaluate the ability of Pagel's (1999) λ to estimate known levels of phylogenetic signal. Our primary simulations were based on pure-birth phylogenies; however, we also evaluated patterns on both balanced and pectinate trees to determine the extent to which tree shape affected our findings (see Supplemental Material). Our first set of simulations evaluated the extent to which values of λ estimated from the data corresponded with actual input levels of phylogenetic signal. For these simulations we generated 50 pure-birth phylogenies at each of six different tree sizes, ranging from 32 to 1024 taxa ($n = 2^5 - 2^{10}$). Next, we rescaled the simulated phylogenies by multiplying the internal branches by λ. We used a set of values that encompassed the entire range of λ: ($\lambda_{in} = 0.0 - 1.0$; in 21 intervals of 0.05 units), resulting in 1050 scaled phylogenies at each level of species richness (n). Continuous traits were then simulated on each phylogeny under a Brownian motion model of evolution to obtain datasets with known and differing levels of phylogenetic signal. These varied from datasets with no phylogenetic signal (when $\lambda_{in} = 0$) to datsets with phylogenetic signal corresponding to what was expected under Brownian motion (when $\lambda_{in} = 1$).

Using the simulated phylogenies we estimated the phylogenetic signal in each dataset using Pagel's λ. Next, for the 1050 datasets at each level of species richness (n), we assessed the relationship between λ_{in} and λ_{est} using regression. Additionally, the precision of λ_{est} was approximated by observing the variation of λ_{est} obtained at each input level of phylogenetic signal (λ_{in}). Finally, for comparison we estimated the degree of phylogenetic signal in each dataset using $Kappa$ (Blomberg et al. 2003; Adams 2014a), and calculated its effect size (Z_K: sensu Adams and Collyer 2016, 2019a) for each dataset. Here, the effect size was estimated as: $Z_K = \frac{K_{obs} - \mu_K}{\sigma_K}$, where K_{obs} was the observed $Kappa$, and μ_K and σ_K were the mean and standard deviation of the empirical sampling distribution of values obtained from the permutation distribution (for discussion see: Adams and Collyer 2019b). Because $Kappa$ and Z_K differ in scale from λ, both measures were linearly normalized to a standard uniform distribution ($0 \rightarrow 1$), and the variation in the set of normalized $Kappa$ and Z_K values at each input level of phylogenetic signal (λ_{in}) were estimated for comparison.

Our second set of simulations evaluated the extent to which values of λ estimated in PGLS regression and ANOVA (i.e., $Y \sim X$) corresponded with actual input levels of phylogenetic signal in the response variable. As before we generated 50 pure-birth phylogenies at each of six levels of species richness ($n = 2^5 - 2^{10}$), and rescaled each with a set of input λ values ($\lambda_{in} = 0.0 - 1.0$; in 21 intervals of 0.05 units). Next, an independent variable X was simulated on each phylogeny under a Brownian motion model of evolution (for PGLS regression). For phylogenetic ANOVA, random groups (X) were obtained by simulating a discrete (binary) character on each phylogeny. Next, the dependent variable was simulated in such a manner as to contain a known relationship with X plus random error containing phylogenetic signal. This was accomplished as: $Y = \beta X + \epsilon$. Here, the association between Y and X was modeled using a range of values: $\beta = (0.0, 0.25, 0.5, 0.75, 1.0)$, and the residual error was modeled to contain phylogenetic signal simulated under a Brownian motion model of evolution: $\epsilon = \mathcal{N}(\mu = 0, \sigma = \mathbf{C})$: (see Revell 2010 for a similar simulation design). For each dataset, the fit of the phylogenetic regression was estimated using maximum likelihood, and parameter estimates (β_{est} and λ_{est}) were obtained and evaluated as above. All analyses were performed in R v3.6.0 (R Core Team 2019) using the packages `geiger` (Harmon et al. 2008), `caper` (Orme et al. 2013), `phytools` (Revell 2012), and `geomorph` (Adams and Otárola-Castillo 2013; Adams et al. 2020). R-scripts are found in the Supplemental Material.

Literature Survey

To determine how Pagel's λ is commonly utilized in empirical studies, we conducted a literature survey. From Google.scholar we obtained a list of all papers published in 2019 that used λ; resulting in 341 studies. For each study, we extracted all λ_{est}, the size of the phylogeny (n), and noted whether authors reported confidence intervals or performed significance tests assessing difference of λ_{est} from either zero or one. We also noted whether biological interpretations based on λ_{est} were made, and for studies that reported more than one λ_{est}, we also noted whether these were compared in some manner, and whether such comparisons were accompanied with statistical tests between λ_{est}.

Results

Simulations

lambda: (Fig 1) not good predictor of input physignal (regression), particularly at low sample sizes. Also, similar to early results of Boettiger (but greatly expanded) range and variation of lambda very large at low N, and generally decreases with larger N. However, the precision was not uniform across the range of lambda-in. Variation in lambda-est was greatest at intermediate levels of physig, and descreased at lower and higher values. Thus, the overall dispersion across the range of input values resembled an asymmetric lens (**NAME?**), with less variation at lambda-in close to 1, greatest variation in middle, and less (but substantial) variation near lambda-in of zero. Notably, even at very large species richness (n=1024) the range of lam-est was quite large: e.g., l-in 0.5 has range of 45% of all possible values. Thus, lambda-est is not very precise in terms of estimating the true l-in parameter, meaning that using lambda to biologically interpret the 'strength' of phylogenetic signal in a dataset, or compare between datasets, is fraught with difficulty.

[insert Figure 1 here]

By contrast, (FIg 2) Kappa. Much smaller range and variation as compared with lambda, but with larger l-in, variation in kappa does increase. However, this is not the case with Z-K. Z-K increases predictably (linearly, progressively. . .) with L-in, and importantly variation in Z-K remains nearly constant across the range of input values. this means that Z-K is equally precise in estimating the strength of phylogenetic signal, and may be more adept at facilitating useful statements on teh relative strength of phylogenetic signal within and across datasets.

NOTE: Fig 2 is 4 panels for N=128: Lambda, Kappa, Z-K, Z-k-normalized.

[insert Figure 2 here]

Literature Survey

We found 182 manuscripts from 2019 that estimated and reported Pagel's lambda values using PGLS methods. These papers averaged 8.527 lambda values, ranging from a single lambda estimate up to 71 estimated lambdas. Almost exactly half of the published lambda estimates were either below 0.05 (25.32%) or above 0.9 (24.74%; Figure 3). 73.32% of the published lambdas were estimated using phylogenies with fewer than 200 tips, and 348 lambda estimates (8.57% of all published estimates) came from phylogenies with fewer than 30 tips.

[insert Figure 3 here]

Many of the reviewed manuscripts liberally interpreted the magnitude of the estimate lambda, using phrases such as “strong” or “weak” phylogentic signal when statistically, all that was clear was a difference between the estimated lambda and 0 or 1 respectively. We estimated that about 20.49% of the manuscripts revealed some sort of biological interpretation of the magnitude of estimated phylogenetic signal that overreached the statistical findings. We also identified seven manuscripts as having inappropriately interpreted differences in lambda values, indicating that some traits had stronger or weaker signal than other traits without the appropriate statistical tests.

As is evidenced by macroevolutionary papers published in 2019 papers, Pagel’s lambda estimation methods are often misused and over-interpretted. Despite the urging of Boettiger and colleagues to publish confidence intervals with all lambda parameter estimates, only 18% of papers published in 2019 do so.

Results

Discussion

1: summary paragraph

2: expand on Lambda.. lambda innacurate, not precise, level of precision varies with input physig (worse in mid-range). Patterns worse with PGLS, though beta still estimated properly. Conclusion, lambda not overly useful.

3: By contrast, effect size Z-K useful, equally precise across range of values. Can be used to characterize the strength of physignal, and because robust to input levels, etc. may be used to compare across datasets.

Somewhere, recognize that this is somewhat ‘backwards’ from prior recommendations where Kappa had somewhat lower performance in terms of type I and type II error (which?? I forget). However, recall that those studies did not examine the precision of the estimates. Nor was Z-k included, because it was not yet invented. So Use of Z-k should make good sense here.

Closing paragraph.

More discussion paragraphs

References

Abouheif, E. 1999. A method for testing the assumption of phylogenetic independence in comparative data. Evolutionary Ecology Research 1:895–909.

Adams, D. C. 2014a. A generalized Kappa statistic for estimating phylogenetic signal from shape and other high-dimensional dultivariate data. Systematic Biology 63:685–697.

Adams, D. C. 2014b. A method for assessing phylogenetic least squares models for shape and other high-dimensional multivariate data. Evolution 68:2675–2688.

Adams, D. C., and M. L. Collyer. 2019a. Comparing the strength of modular signal, and evaluating alternative modular hypotheses, using covariance ratio effect sizes with morphometric data. Evolution 73:2352–2367.

Adams, D. C., and M. L. Collyer. 2016. On the comparison of the strength of morphological integration across morphometric datasets. Evolution 70:2623–2631.

Adams, D. C., and M. L. Collyer. 2018. Phylogenetic anova: Group-clade aggregation, biological challenges, and a refined permutation procedure. Evolution 72:1204–1215.

Adams, D. C., and M. L. Collyer. 2019b. Phylogenetic comparative methods and the evolution of multivariate phenotypes. Annual Review of Ecology, Evolution, and Systematics 50:405–425.

Adams, D. C., M. L. Collyer, and A. Kaliontzopoulou. 2020. Geomorph: Software for geometric morphometric analyses. R package version 3.2.1.

Adams, D. C., and E. Otárola-Castillo. 2013. Geomorph: An r package for the collection and analysis of geometric morphometric shape data. Methods in Ecology and Evolution 4:393–399.

Beaulieu, J. M., D. C. Jhwueng, C. Boettiger, and B. C. O'Meara. 2012. Modeling stabilizing selection: Expanding the ornstein-uhlenbeck model of adaptive evolution. Evolution 66:2369–2383.

Blomberg, S. P., T. Garland, and A. R. Ives. 2003. Testing for phylogenetic signal in comparative data: Behavioral traits are more labile. Evolution 57:717–745.

Boettiger, C., G. Coop, and P. Ralph. 2012. Is your phylogeny informative? Measuring the power of comparative methods. Evolution 67:2240–2251.

Butler, M. A., and A. A. King. 2004. Phylogenetic comparative analysis: A modeling approach for adaptive evolution. American Naturalist 164:683–695.

Diniz-Filho, J. A. F., T. Santos, T. F. Rangel, and L. M. Bini. 2012. A comparison of metrics for estimating phylogenetic signal under alternative evolutionary models. Genetics and Molecular Biology 35:673–679.

Felsenstein, J. 1985. Phylogenies and the comparative method. American Naturalist 125:1–15.

Freckleton, R. P., P. H. Harvey, and M. Pagel. 2002. Phylogenetic analysis and comparative data: A test and review of evidence. American Naturalist 160:712–726.

Garland, T. J., and A. R. Ives. 2000. Using the past to predict the present: Confidence intervals for regression equations in phylogenetic comparative methods. American Naturalist 155:346–364.

Gittleman, J. L., and M. Kot. 1990. Adaptation: Statistics and a null model for estimating phylogenetic effects. Systematic Zoology 39:227–241.

Grafen, A. 1989. The phylogenetic regression. Philosophical Transactions of the Royal Society of London B, Biological Sciences 326:119–157.

Harmon, L. J., J. T. Weir, C. D. Brock, R. E. Glor, and W. Challenger. 2008. GEIGER: Investigating evolutionary radiations. Bioinformatics 24:129–131.

Harvey, P. H., and M. D. Pagel. 1991. The comparative method in evolutionary biology. Oxford University Press, Oxford.

Klingenberg, C. P., and N. A. Gidaszewski. 2010. Testing and quantifying phylogenetic signals and homoplasy in morphometric data. Systematic biology 59:245–261.

Martins, E. P., and T. F. Hansen. 1997. Phylogenies and the comparative method: A general approach to incorporating phylogenetic information into the analysis of interspecific data. American Naturalist 149:646–667.

Molina-Venegas, R., and M. A. Rodriguez. 2017. Revisiting phylogenetic signal; strong or negligible impacts of polytomies and branch length information? BMC evolutionary biology 17:53.

Munkemuller, T., S. Lavergne, B. Bzeznik, S. Dray, T. Jombart, K. Schiffers, and W. Thuiller. 2012. How to measure and test phylogenetic signal. Methods in Ecology and Evolution 3:743–756.

O'Meara, B. C., C. Ane, M. J. Sanderson, and P. C. Wainwright. 2006. Testing for different rates of continuous trait evolution using likelihood. Evolution 60:922–933.

Orme, D., R. P. Freckleton, G. H. Thomas, T. Petzoldt, S. A. Fritz, and N. Isaac. 2013. CAPER: Comparative analyses of phylogenetics and evolution in r. Methods in Ecology and Evolution 3:145–151.

Pagel, M. D. 1999. Inferring the historical patterns of biological evolution. Nature 401:877–884.

Pavoine, S., and C. Ricotta. 2012. Testing for phylogenetic signal in biological traits: The ubiquity of cross-product statistics. Evolution: International Journal of Organic Evolution 67:828–840.

R Core Team. 2019. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Revell, L. J. 2010. Phylogenetic signal and linear regression on species data. Methods in Ecology and Evolution 1:319–329.

Revell, L. J. 2012. Phytools: An r package for phylogenetic comparative biology (and other things). Methods in Ecology and Evolution 3:217–223.

Revell, L. J., and L. J. Harmon. 2008. Testing quantitative genetic hypotheses about the evolutionary rate matrix for continuous characters. Evolutionary Ecology Research 10:311–331.

Revell, L. J., L. J. Harmon, and D. C. Collar. 2008. Phylogenetic signal, evolutionary process, and rate. Systematic Biology 57:591–601.

Rohlf, F. J. 2001. Comparative methods for the analysis of continuous variables: Geometric interpretations. Evolution 55:2143–2160.

Figure Legends

Figure 1. Accuracy of Pagel's lambda estimations across known lambda inputs on various tree sizes. As trees increase in size, the estimates more closely resemble the input lambdas, however considerable and concerning variation is apparent in trees smaller than those with 200 tips.

Figure 2. Estimated ANOVA slopes under PGLS. Across tree sizes, the mean estimated slope matches the input slope, and as trees increase in size, the variance around this mean estimate decreases. However, for trees with fewer than 200 tips, the error around the estimated slope is considerable, where these analyses frequently estimate slopes in the opposite direction of the known pattern.

Figure 3. Frequency of estimated lambda values published in manuscripts in 2019. The majority of these values were close to 0 or 1, and from phylogenies with fewer than 200 taxa.

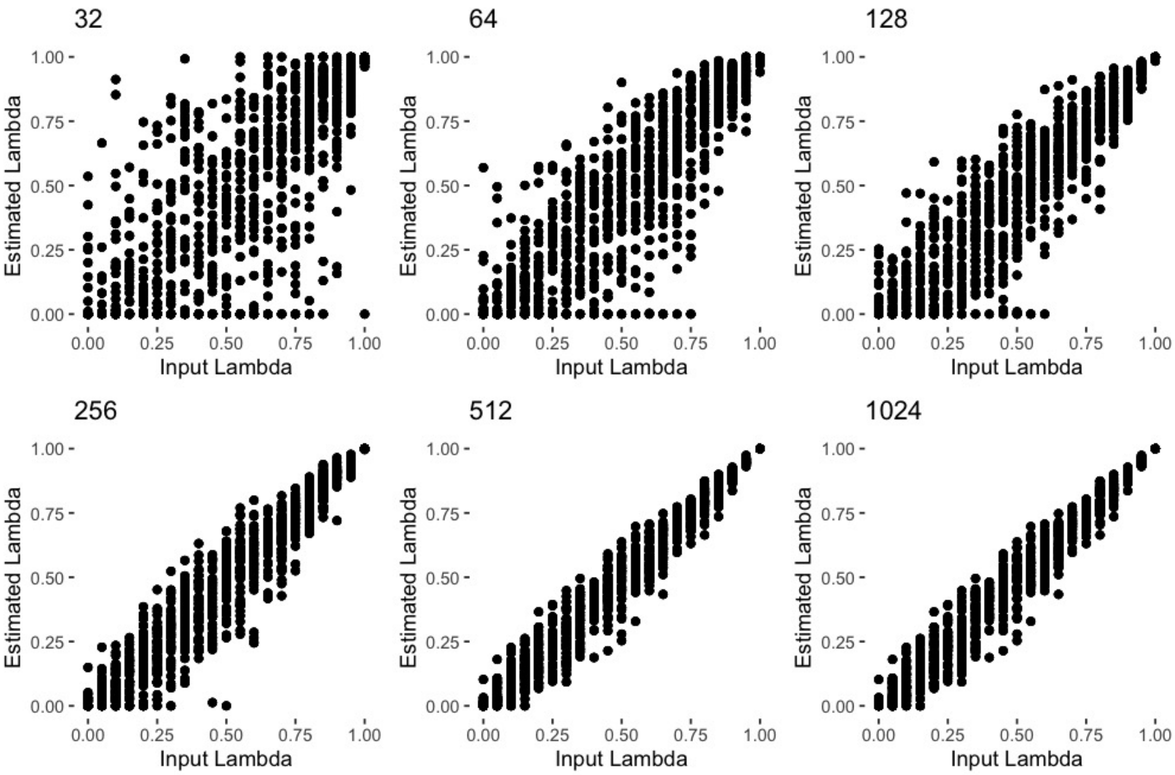

Figure 1. Accuracy of Pagel's lambda estimations across known lambda inputs on various tree sizes. As trees increase in size, the estimates more closely resemble the input lambdas, however considerable and concerning variation is apparent in trees smaller than those with 200 tips.

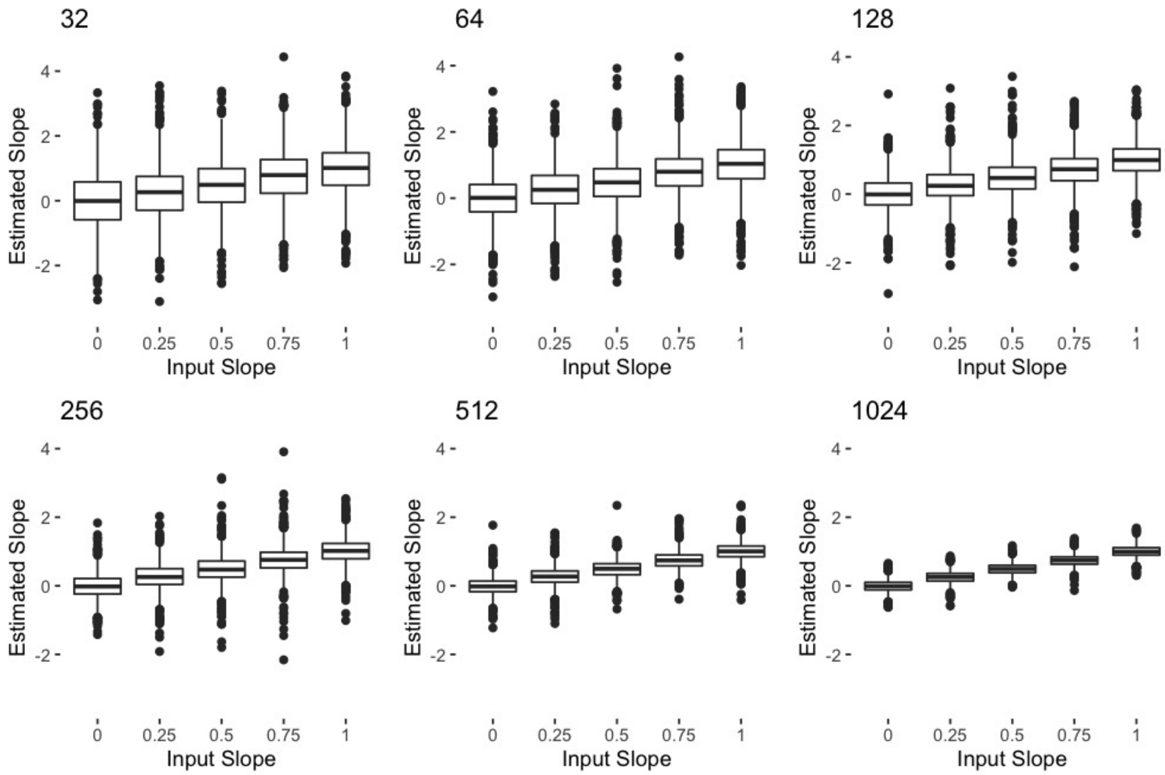

Figure 2. Estimated ANOVA slopes under PGLS. Across tree sizes, the mean estimated slope matches the input slope, and as trees increase in size, the variance around this mean estimate decreases. However, for trees with fewer than 200 tips, the error around the estimated slope is considerable, where these analyses frequently estimate slopes in the opposite direction of the known pattern.

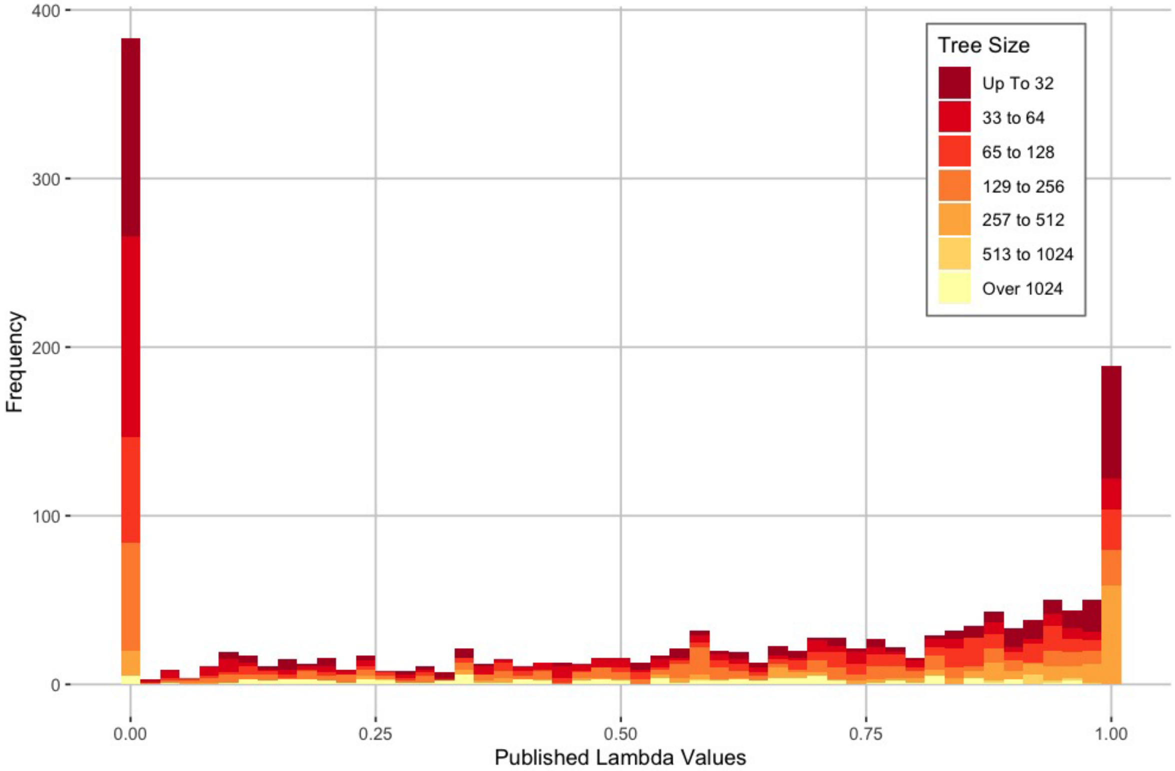

Figure 3. Frequency of estimated lambda values published in manuscripts in 2019. The majority of these values were close to 0 or 1, and from phylogenies with fewer than 200 taxa.