A Standardized Effect Size for Evaluating the Strength of Phylogenetic Signal, and Why Lambda is not Appropriate

Abstract

Macroevolutionary studies frequently characterize the phylogenetic signal in phenotypes, and wish to compare the strength of that signal across traits. However, analytical tools for such comparisons have largely remained underdeveloped. In this study, we evaluated the efficacy of one commonly used parameter (Pagel's λ) to estimate the strength of phylogenetic signal in phenotypic traits, and evaluate the degree to which λ correctly identifies known levels of phylogenetic signal. We find that the precision of λ in estimating actual levels of phylogenetic signal is often inaccurate, and that biological interpretations of the strength of phylogenetic signal based on λ are therefore compromised. We then propose a standardized effect size based on κ (Z_κ), which measures the strength of phylogenetic signal, and places it on a common scale for statistical comparison. Tests based on Z_κ provide a mechanism for formally comparing the strength of phylogenetic signal across datasets, in much the same manner as effect sizes may be used to summarize patterns in quantitative meta-analysis. Our approach extends the phylogenetic comparative toolkit to address hypotheses that compare the strength of phylogenetic signal between various phenotypic traits, even when those traits are found in different evolutionary lineages or have different units or scales.

Introduction

Investigating macroevolutionary patterns of trait variation requires a phylogenetic perspective, because the shared ancestry among species violates an assumption of independence among trait values that is common for statistical tests (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 trends in the data on the phylogenetic relatedness of observations (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).

Several analytical tools have been developed to quantify phylogenetic signal in phenotypic datasets, including measures of serial independence ($\mathbf{C}$: Abouheif 1999), autocorrelation estimates (I: Gittleman and Kot 1990), statistical ratios of trait variation relative to what is expected given the phylogeny (κ: Blomberg et al. 2003; Adams 2014a), and scaling parameters used in maximum likelihood fitting of the data to the phylogeny (λ: Pagel 1999), among others (e.g., Klingenberg and Gidaszewski 2010). The statistical properties of these methods – namely type I error rates and power – have also been investigated to determine when phylogenetic signal can be detected and under what conditions (e.g., Münkemüller et al. 2012; Pavoine and Ricotta 2012; Diniz-Filho et al. 2012; Adams 2014a; Molina-Venegas and Rodríguez 2017; see also Revell et al. 2008; Revell 2010). One of the most widely used methods for characterizing phylogenetic signal in macroevolutionary studies is Pagel's λ (Pagel 1999). The parameter (λ) transforms the lengths of the internal branches of the phylogeny to improve the fit of data to the phylogeny via maximum likelihood (Pagel 1999; Freckleton et al. 2002). Pagel's λ ranges from $0 \rightarrow 1$, with larger values signifying a greater dependence of observed trait variation on the phylogeny. Pagel's λ also has the appeal that it may be included in phylogenetic generalized least-squares regression (PGLS) to account for the degree of phylogenetic signal in comparative analyses (see Freckleton et al. 2002).

In PGLS analysis, λ is a parameter that is tuned via log-likelihood profiling for its estimation. After obtaining an optimized λ value, $\hat{\lambda}$, likelihood ratio tests can be performed to compare observed model fits using $\hat{\lambda}$ to those obtained when $\lambda = 0$ or $\lambda = 1$ to infer whether phylogenetic signal differs from no signal or a Brownian motion (BM) model of evolutionary divergence, respectively (Freckleton et al. 2002; Cooper et al. 2010; Bose et al. 2019). Confidence limits on $\hat{\lambda}$ can be also estimated based on percentiles of a χ^2 distribution and where these values are mapped onto a log-likelihood profile (*sensu* Freckleton et al. 2002). With respect to confidence limits, one can ascertain whether $\lambda = 0$ or $\lambda = 1$ is contained within an interval of $\hat{\lambda}$, as an analog to likelihood ratio tests (see, e.g., Vandelook et al. 2019). It is, therefore, tempting to regard $\hat{\lambda}$ as a descriptive statistic that measures the relative strength of phylogenetic signal, for inferring the extent to which shared evolutionary history has influenced trait covariation among taxa, is common. The appeal of $\hat{\lambda}$ as a descriptive statistic is that it serves as the basis for interpreting "weak" versus "strong" phylogenetic signal; i.e., small versus large values of $\hat{\lambda}$, respectively, in a comparative sense (e.g., De Meester et al. 2019; Pintanel et al. 2019; Su et al. 2019). For example, qualitative comparisons of $\hat{\lambda}$ for multiple traits on the same phylogenetic tree provide a basis for inferring whether the strength of phylogenetic signal is greater in one trait as compared to another (e.g., Liu et al. 2019; Bai et al. 2019). Indeed, statements regarding the strength of phylogenetic signal based on $\hat{\lambda}$ are rather common in the evolutionary literature. For instance, of the 204 papers published in 2019 that estimated and reported Pagel's λ (found from a literature survey we conducted in Google.scholar), 40% explicitly interpreted the strength of phylogenetic signal for at least one phenotypic trait. Further, because nearly half of the 1,572 $\hat{\lambda}$ values reported were near 0 or 1 (Figure 1) where the biological interpretation of λ is known ($\lambda = 0$ represents no phylogenetic signal, while $\lambda = 1$ is phylogenetic signal as expected under BM), this percentage is even higher.

[insert Figure 1 here]

It seems intuitive to interpret the strength of phylogenetic signal based on the value of $\hat{\lambda}$, as λ is a parameter on a bounded scale ($0 \rightarrow 1$) for which interpretation of its extremal points are understood. However, equating values of $\hat{\lambda}$ directly to the strength of phylogenetic signal presumes two important statistical properties that have not been fully explored. First, it presumes that values of $\hat{\lambda}$ can be precisely estimated, as biological inferences regarding the strength of phylogenetic signal depend on high accuracy in its estimation. (In order to be precise, the log-likelihood profile would have to be peaked rather than flat.) Therefore, understanding the precision in estimating $\hat{\lambda}$ is paramount. One study (Boettiger et al. 2012) found that estimates of Pagel's λ displayed less variation (i.e., greater precision) when data were simulated on a large phylogeny ($N = 281$)

as compared to a small one ($N = 13$). From this observation it was concluded that insufficient data (i.e., low number of species) was the underlying cause of the increased variation across parameter estimates (Boettiger et al. 2012). Indeed, such a pattern is common with statistical estimators, as summary statistics and parameters are often more precise at greater sample sizes (Cohen 1988). However, this conclusion could imply that the shape of log-likelihood profiles are assumed consistent with species number. Such an assumption and the importance of tree shape have to date not been verified. Thus, despite widespread use of Pagel's (1999) λ in macroevolutionary studies, at present, we lack a general understanding of the precision with which $\hat{\lambda}$ can estimate levels of phylogenetic signal in phenotypic datasets.

Second, while estimates of λ are within a bounded scale ($0 \rightarrow 1$), this does not *de-facto* imply that the estimated values, $\hat{\lambda}$, of this parameter correspond to the actual strength of the underlying input signal in the data. For this to be the case, $\hat{\lambda}$ must be a statistical effect size. Effect sizes are a measure of the magnitude of a statistical effect in data, represented on a common scale (Glass 1976; Cohen 1988). Effect sizes have widespread use in many areas of the quantiative sciences, as they represent measures that may be readily summarized across datasets as in meta-analyses (Glass 1976; Hedges and Olkin 1985; Arnqvist and Wooster 1995), or compared among datasets (e.g., Adams and Collyer 2016, 2019a). Unfortunately, not all model parameters and descriptive statistics are effect sizes, directly, and thus many summary measures must first be converted to statistics with standardized units (i.e., conversion to an effect size) for meaningful comparison (see Rosenthal 1994). As a consequence, it follows that only if $\hat{\lambda}$ is a statistical effect size can comparisons of estimates across datasets be interpretable. However, the calculation and statistical behavior of $\hat{\lambda}$ as an effect size has not yet been explored.

As a prortional random variable bounded by 0 and 1, we might expect that $\hat{\lambda}$ follows A Bernoulli distribution; i.e., branch lengths in a tree are scaled proportionally to the probability that data arise from a BM process. As such, we would also expect in simulation experiment that given a known λ value used to generate random data on a tree: the mean of an empirical sampling distribution of $\hat{\lambda}$ would approximately equal λ; the standard error of $\hat{\lambda}$ would be largest at intermediate values of λ, and predictable over the range of λ with respect to treesize; the distribution of $\hat{\lambda}$ would be symmetric at intermediate values of λ and more skewed toward values of 0 or 1; and that the distribution of $\hat{\lambda}$ will be more platykurtic at intermediate values of λ, becoming more leptokurtic toward 0 and 1. These properties seemed to be superficiallyreasonable from the simulations perfomed by Münkemüller et al.(2012, see Fig. 2). Their study simulated strength of phylogenetic signal as a varied weighted-average of data simulated on trees with $\lambda = 0$ and $\lambda = 1$, rather than from prescribed

values of λ. Nonetheless, their results revealed that as a random variable, $\hat{\lambda}$ resembled a Bernoulli random variable based superficially on statistical moments (mean, variance, skewness, and kurtosis; see their Fig. 2) with respect to strength of phylogenetic signal, for a given tree size. However, for comparisons among tree sizes for a given strength of phylogenetic signal, median values of $\hat{\lambda}$ were slightly, positively associated with the number of species in a tree, especially at weaker strength of signals, suggesting $\hat{\lambda}$ might not be an accurate statistic across a range of tree sizes. In order for $\hat{\lambda}$ to be considered a reliable statistic as a measure of phylogenetic signal, as a minimum, it should be accurate and precise, properties that could be measured from simulation experiments for known values of λ.

In this study, we evaluate the precision of $\hat{\lambda}$ as an optimized parameter, used as a statistic, for estimating known levels of phylogenetic signal in phenotypic data (data simulated with known λ). We use computer simulations with differing numbers of species, differently shaped phylogenies, and differing input levels of λ, to explore the degree to which $\hat{\lambda}$ correctly identifies known levels of phylogenetic signal, λ, and under which circumstances. We find that estimates of $\hat{\lambda}$ vary widely for a given input value of λ, and that the precision of $\hat{\lambda}$ is not constant across its range. Rather, there is decreased precision when input levels of phylogenetic signal are of intermediate strength. Additionally, the same values of $\hat{\lambda}$ may be obtained from datasets containing vastly different input levels of λ. Thus, $\hat{\lambda}$ is not a reliable indicator of the strength of phylogenetic signal in phenotypic data. We then describe a standardized effect size for measuring the strength of phylogenetic signal in phenotypic datasets and apply the concept to two common measures of phylogenetic signal: $\hat{\lambda}$ and κ. Through simulations we find that the precision of effect sizes based on $\hat{\lambda}$ (Z_λ) are less reliable than that those based on κ (Z_κ), implying that Z_κ is a more robust effect size measure. We also propose a two-sample test statistic that may be used to compare the strength of phylogenetic signal among datasets, and provide an empirical example to demonstrate its use. We conclude that estimates of phylogenetic signal using Pagel's λ (as a descriptive statistic, not as a PGLS parameter) are often inaccurate, and thus interpreting strength of phylogenetic signal in phenotypic datasets based on this measure is compromised. By contrast, effect sizes obtained from κ hold promise for characterizing phylogenetic signal, and for comparing the strength of phylogenetic signal across datasets.

Methods and Results

The Precision of $\hat{\lambda}$ is Variable

We conducted a series of computer simulations to evaluate the precision of $\hat{\lambda}$ based on varied input of Pagel's λ to generate data. Our primary simulations were based on pure-birth phylogenies; however, we also evaluated patterns on both balanced and pectinate trees to determine whether tree shape affected our findings (see Supporting Information). First we generated 50 pure-birth phylogenies at each of six different tree sizes, ranging from 32 to 1024 taxa ($n = 2^5 - 2^{10}$). For each simulation run, we rescaled the simulated phylogenies by multiplying the internal branches by λ_{in}, using 21 intervals of 0.05 units across its range ($\lambda_{in} = 0.0 \rightarrow 1.0$), resulting in 1050 scaled phylogenies at each level of species richness (n). Continuous traits were then simulated on each phylogeny under a BM model of evolution to obtain datasets with differing levels of phylogenetic signal, that ranged from no phylogenetic signal (when $\lambda_{in} = 0$), to phylogenetic signal reflecting a BM model of evolution (when $\lambda_{in} = 1$). For each dataset we then estimated phylogenetic signal ($\hat{\lambda}$), and calculated its variance ($\sigma^2_{\hat{\lambda}}$) at each input level of phylogenetic signal and level of species richness as an estimate of precision. Random data samples were generated from a normal distribution, i.e., $\mathcal{N}(\mu = 0, \sigma = \sigma^2\mathbf{C})$. We verified that the level of σ^2 had no effect on $\hat{\lambda}$ or $\sigma^2_{\hat{\lambda}}$, but the results we present used $\sigma^2 = 1$.

We also evaluated the precision of $\hat{\lambda}$ when estimated in PGLS regression and ANOVA (i.e., $Y \sim X$). For these simulations, $Y = \beta X + \epsilon$ was used to generate data with varied phylogenetic signal and variable associations (*sensu* Revell 2010). In the case of PGLS regression, X was a variable simulated under a BM model of evolution, constrained by rescaling phylogeny according to λ, and in the case of ANOVA, X was a balanced binary variable (values of 0 and 1), randomly ordered. The strength of variable associations ranged as $\beta = 0.0, 0.25, 0.5, 0.75$ and 1.0. Residuals, ϵ, were simulated under a BM model of evolution, constrained by rescaling phylogeny according to λ. Simulated independent variables or residuals were generated from a normal distribution, i.e., $\mathcal{N}(\mu = 0, \sigma = \sigma^2\mathbf{C})$, with the phylogenetic covariance matrix, $\mathbf{C}$, implicitly varied for each tree by the level of λ. The fit of the phylogenetic regression was estimated using maximum likelihood, and parameter estimates ($\hat{\lambda}$ and $\hat{\beta}$) were obtained. We then calculated precision estimates ($\sigma^2_{\hat{\lambda}}$) at each input level of phylogenetic signal and level of species richness. We verified that the level of σ^2 had no effect on $\sigma^2_{\hat{\lambda}}$ but did influence the precision of $\hat{\beta}$ estimated from the linear model (precision increased with smaller ϵ, as expected). The results we present used $\sigma^2 = 1$.

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` 3.2.1 (Adams and Otárola-Castillo 2013; Adams et al. 2020). R-scripts are found in the Supporting Information.

Results. We found that the precision of λ_{est} varied widely across simulation conditions. Predictably, precision improved as the number of species increased (Figure 2). This confirmed earlier findings of Boettiger et al. (2012), and adhered to parametric statistical theory. However, in many cases the set of λ_{est} spanned nearly the entire range of possible values (e.g., $n = 32$; $\lambda_{in} = 0.5$: $\lambda_{est} = 0.0 \rightarrow 0.985$), revealing that estimates of λ were not a reliable indicator of input phylogenetic signal. Importantly, the precision of λ_{est} was not uniform across all levels of phylogenetic signal. The worst precision was observed at intermediate levels of phylogenetic signal ($\lambda_{in} \approx 0.5$), while precision improved as input levels approached the extremes of λ's range (i.e., $\lambda_{in} \rightarrow 0$ & $\lambda_{in} \rightarrow 1$). Thus, estimates of λ were least reflective of the true input signal at intermediate values. Additionally, even at large levels of species richness, we found that the range of λ_{est} still encompassed a substantial portion of possible values (e.g., $n = 512$; $\lambda_{in} = 0.5$: $\lambda_{est} = 0.32 \rightarrow 0.68$). Likewise, the same λ_{est} could be obtained from datasets containing vastly different input levels of phylogenetic signal (e.g., $n = 512$; $\lambda_{est} = 0.5$; $\lambda_{in} = 0.25 \rightarrow 0.65$). These findings were particularly unsettling when considered in light of our literature survey. Over one quarter of the λ estimates published in empirical studies (421 of 1,572) were between $\lambda = 0.25$ and $\lambda = 0.75$ (Figure 1). This range reflected the region that our simulations identified as being the least reliable in terms of accurately characterizing levels of phylogenetic signal, yet 30% of these mid-range empirical estimates were explicitly interpreted in terms of the strength of phylogenetic signal that they represented (i.e., weak, intermediate, strong phylogenetic signal).

When phylogenetic signal was investigated on balanced or pectinate trees, patterns in the precision of λ were largely the same, with decreased precision at intermediate levels of phylogenetic signal (Supporting Information). Likewise, when λ was co-estimated with regression parameters in PGLS regression and ANOVA, the results of our simulations were quite similar. Regression parameters (β) were accurately estimated, confirming earlier findings of Revell (2010) (see Supporting Information). However, estimates of phylogenetic signal (λ) were less precise (Figure 3; see also Supporting Information), and the spread of λ_{est} was similar to that observed when λ was estimated for only the dependent variable, as in Figure 2. Taken together, these findings reveal that λ_{est} does not precisely characterize known levels of phylogenetic signal in phenotypic datasets, and that biological interpretations of the strength of phylogenetic signal based on λ may be highly inaccurate.

[insert Figure 2 here]

[insert Figure 3 here]

A Standardized Effect Size for Phylogenetic Signal

The results above demonstate that λ is not a reliable estimate of the phylogenetic signal in phenotypic data. As such, biological interpretations of the strength of phylogenetic signal, and comparisons of the magnitude of such effects across datasets are severely compromised when based on this parameter. As an alternative, we propose that summary estimates of phylogenetic signal be converted to effect sizes for interpretation and comparison of the relative strength of phylogenetic signal in phenotypic datasets. Statistically, a standardized effect size may be found as:

$$Z_\theta = \frac{\theta_{obs} - E(\theta)}{\sigma_\theta} \tag{1}$$

where θ_{obs} is the observed test statistic, $E(\theta)$ is its expected value under the null hypothesis, and σ_θ is its standard error (Glass 1976; Cohen 1988; Rosenthal 1994). Z_θ expresses the magnitude of the effect in θ_{obs} by transforming the original test statistic to its standard normal deviate (Glass 1976; Kelley and Preacher 2012). Typically, θ_{obs} and σ_θ are estimated from the data, while $E(\theta)$ is obtained from the distribution of θ derived from parametric theory. However, recent advances in resampling theory (Collyer et al. 2015; Adams and Collyer 2016, 2019a; Collyer and Adams 2018) have shown that $E(\theta)$ and σ_θ may also be obtained from an empirical sampling distribution of θ obtained from permutation procedures.

Adams and Collyer (2019b) suggested that the strength of phylogenetic signal could be represented as an effect size based on the κ statistic and its empirical sampling distribution from permutation. Here we formalize that suggestion, resulting in an effect size of:

$$Z_\kappa = \frac{log(\kappa_{obs}) - \hat{\mu}_{log(\kappa)}}{\hat{\sigma}_{log(\kappa)}}, \tag{2}$$

where κ_{obs} is the observed phylogenetic signal, and $\hat{\mu}_\kappa$ and $\hat{\sigma}_\kappa$ are the mean and standard deviation of the empirical sampling distribution of $log(\kappa)$ obtained via permutation. Note that the logarithm was used because κ takes only positive values ($0 \rightarrow \infty$) and its sampling distribution is log-normally distributed (for a similar transformation when calculating multivariate effect sizes see: Appendix 1 of Collyer and Adams 2018).

An effect size based on λ is a little more challenging because, as a statistic, it does not have a sampling distribution on which to measure standard error. Confidence intervals for λ are not generated from standard error calculations, but rather for the values of lambda that intersect the log-likehihood profile for corresponding percentiles of the χ^2 distribution used to compare the compare the putative appropriate model to a null model with $\lambda = 0$ [citation needed, think Boettiger has a paper on this]. The simplest similar effect size is

$$|Z_\lambda| = \sqrt{\chi^2_{\hat{\lambda}}} \tag{3}$$

where, $\hat{\lambda}$ is the maximized likelihood value of λ and $\chi^2_{\hat{\lambda}}$ is the likelihood ratio statistic for the value. Hypothetically, this equation would be better represented as $Z_\lambda = d\sqrt{\chi^2_{\hat{\lambda}}}$, where d is a binary value $(-1, 1)$ to indicate a direction based on whether $\hat{\lambda}$ is below or above a critical value of λ, for a quantile from a χ^2 distribution at a probability of 0.5. The critical value would correspond to a value of $Z = 0$; i.e., the strength of phylogenetic signal that might be observed at the 50th percentile in a distribution of random outcomes, generated from a model where $\lambda = 0$. However, in some preliminary simulations, we found that mapping $\chi^2_{\hat{\lambda}}$ values this way did not produce effect sizes that were symmetrical about $Z = 0$, as the mapping is not linear and, especially for small trees, the log-likelihood profiles can be rather flat, making estimation of $\hat{\lambda}$ rather imprecise (as we elaborate more below.) We will heneforth use equation 3, recognizing that large values of Z would likely be positive, anyway, based on the reference model for the likelihood ratio test.

In this case, λ_{obs} and $\hat{\sigma}_\lambda$ are empirically derived using maximum likelihood, as permutation approaches have not been developed for evaluating λ. Note also that under the null hypothesis, no phylogenetic signal is expected (Freckleton et al. 2002), and thus $E(\lambda) = 0$ under this condition.

To evaluate the utility of Z_κ and Z_λ we calculated both effect sizes for the simulated datasets generated above, and summarized the precision of each using its variance ($\sigma^2_{Z_\kappa}$ and $\sigma^2_{Z_\lambda}$, Figure 4: additional results in the Supporting Information). Here two things are evident. First, estimates of Z_κ linearly track the input

phylogenetic signal whereas estimates of Z_λ do not (Figure 4A, B). Thus, actual changes in the strength of phylogenetic signal are reflected more evenly in the corresponding values of the effect size Z_κ. Second, the precision of Z_κ is considerably more stable as compared with Z_λ. This may be seen by calculating the coefficients of variation for the set of precision estimates (i.e., $\sigma^2_{Z_\kappa}$ and $\sigma^2_{Z_\lambda}$) across input levels of phylogenetic signal. Coefficients of variation in the precision of Z_κ were up to an order of magnitude smaller than for Z_λ (Figure 4C), implying that estimates of the strength of phylogenetic signal were more reliable and robust when using Z_κ.

[insert Figure 4 here]

Statistical Comparisons of Phylogenetic Signal

Once the magnitude of phylogenetic signal is characterized using Z_κ, one may wish to compare such measures across datasets, to determine whether the strength of phylogenetic signal is significantly greater in one phenotypic trait than another. As with other effect sizes derived from permutation distributions (e.g., Adams and Collyer 2016, 2019a), a two-sample test statistic may be calculated as:

$$\hat{Z}_{12} = \frac{|(\kappa_1 - \hat{\mu}_{\kappa_1}) - (\kappa_2 - \hat{\mu}_{\kappa_2})|}{\sqrt{\hat{\sigma}^2_{\kappa_1} + \hat{\sigma}^2_{\kappa_2}}} = \frac{|Z_{\kappa_1} - Z_{\kappa_2}|}{\sqrt{2}} \tag{4}$$

where κ_1, κ_2, $\hat{\mu}_{\kappa_1}$, $\hat{\mu}_{\kappa_2}$, $\hat{\sigma}_{\kappa_1}$, and $\hat{\sigma}_{\kappa_2}$ are as defined above for equation 2. The right side of the equation illustrates that if Z_κ has already been calculated for two sampling distributions as in equation 2, the sampling distributions have unit variance for each of the Z_κ statistics. Estimates of significance of $\hat{Z}_{12}$ may be obtained from a standard normal distribution. Typically, $\hat{Z}_{12}$ is considered a two-tailed test, however directional (one-tailed) tests may be specified should the empirical situation require it (see Adams and Collyer 2016, 2019a).

Empirical Example

To demonstrate the utility of $\hat{Z}_{12}$ we quantified and compared the strength of phylogenetic signal of two phenotypic traits across species of plethodontid salamander. The data were part of a series of studies

examining macroevolutionary trends in phenotypic diversification in this group (Baken and Adams 2019; Baken et al. 2020). Our dataset contained surface area to volume ratios (SA:V) and relative body width ($\frac{BW}{SVL}$) for 305 species (Figure 5A). For SA:V, 11 linear body measurements from 2,781 individuals were taken, from which estimates of the surface area and volume of the head, body, and tail were calculated and subsequently combined to arrive at the SA:V for each indiviual (for mathematical details see Baken et al. 2020). Species means were then obtained. Likewise, body size (SVL) and body width (BW) measurements were taken from 3,371 individuals, and species means of relative body width ($\frac{BW}{SVL}$) were calculated (data from Baken and Adams 2019). A time-dated molecular phylogeny for the group (Bonett and Blair 2017) was then pruned to match the species in the phenotypic dataset. The phylogenetic signal in each trait was then characterized using κ, which was converted to its effect size (Z_κ) using `geomorph` 3.2.1 (Adams and Otárola-Castillo 2013; Adams et al. 2020). Finally, the strength of phylogenetic signal was compared across traits using $\hat{Z}_{12}$ as described above (**to be incorporated in `geomorph` upon manuscript acceptance**).

Results. Both SA:V and relative body width displayed significant phylogenetic signal ($\kappa_{SA:V} = 0.7608$; $P = 0.001$; $\kappa_{BW/SVL} = 0.2515$; $P = 0.001$). For both phenotypic traits, κ_{obs} differed markedly from their corresponding permutation distributions, which were found to overlap almost perfectly (Figure 5B). However, while both traits displayed significant phylogenetic signal, there was nearly a four-fold difference in the magnitude of their effect sizes, with SA:V displaying the greater phylogenetic signal (Figure 5C). Using the two-sample test statistic above, this difference was found to be highly significant ($\hat{Z}_{12} = 4.13$; $P = 0.000036$). Thus it may be concluded that SA:V displays significantly stronger phylogenetic signal than does relative body width, and that shared evolutionary history has strongly influenced trait covariation among taxa for SA:V. Biologically, this observation corresponds with the fact that tropical species – which form a monophyletic group within plethodontids – display greater variation in SA:V which covaries with disparity in their climatic niches (Baken et al. 2020). We hypothesize that because of this macroevolutionary association, strong phylogenetic signal in SA:V is observed.

Discussion

It is common in comparative evolutionary studies to characterize the phylogenetic signal in phenotypic traits to determine the extent to which shared evolutionary history has generated trait covariation among taxa. However, while numerous analytical approaches may be used to quantify phylogenetic signal (e.g., Abouheif 1999; Gittleman and Kot 1990; Pagel 1999; Blomberg et al. 2003; Adams 2014a), methods that explicitly

measure the strength of phylogenetic signal, or facilitate comparisons among datasets, have remained underdeveloped. In this study, we evaluated the precision of one common measure, Pagel's λ, and explored its efficacy for characterizing the strength of phylogenetic signal in phenotypic data. Using computer simulations, we found that the precision of λ increased with increasing sample sizes; a pattern noted previously (Boettiger et al. 2012), and one that conformed with parametric statistical theory (Cohen 1988). However, we also found that vastly different λ estimates could be obtained from data containing the same level of phylogenetic signal, and that similar λ estimates may be obtained from data containing differing levels of phylogenetic signal. Further, the precision of λ varied with the strength of phylogenetic signal, where lower precision was observed when in data whose phylogenetic signal was of intemediate strength. From these findings we conclude that λ is not a reliable indicator of the observed strength of phylogenetic signal in phenoytpic datasets, and that biological interpretations of the strength of signal based on this parameter may innacurately characterize such effects.

As an alternative, we described a standardized effect size (Z) for assessing the strength of phylogenetic signal. Z expresses the magnitude of phylogenetic signal as a standard normal deviate, which is easily interpretable as the strength of phylogenetic signal relative to the mean. We applied this concept to both λ and κ, and found that Z_κ was a better estimate of the strength of phylogenetic signal in phenotypic data. First, Z_κ was more precise than Z_λ, and precision was more consistent across the range of input levels of phylogenetic signal. Additionally, values of Z_κ more accurately tracked known changes in the magnitude of phylogenetic signal, as demonstrated by the linear relationship between Z_κ and λ_{in}. Thus, Z_κ holds promise as a measure of the relative strength of phylogenetic signal that reflects the magnitude of this effect in phenotypic data. We therefore recommend that future studies interested in the strength of phylogenetic signal incorporate Z_κ as a statistical measure of this effect.

Based on the effect size Z_κ, we then proposed a two-sample test, which provides means of determining whether the strength of phylogenetic signal is greater in one phenotypic trait as compared to another, via a hypothesis test. Prior studies have summarized patterns of variation in phylogenetic signal across datasets using summary test values, such as κ (e.g., Blomberg et al. 2003). However, κ does not scale linearly with input levels of phylogenetic signal, and its variance increases (i.e., precision decreases) with increasing strength of phylogenetic signal (Münkemüller et al. 2012; Diniz-Filho et al. 2012: see also Supporting Information). Thus, κ should not be considered an effect size that measures the strength of phylogenetic signal on a common scale. By contrast, standardizing κ (Z_κ, via equation 2) alleviates these concerns, and facilitates formal statistical comparisons of the strength of signal across datasets. Thus when viewed from this

perspective, the approach developed here aligns well with other statistical approaches such as meta-analysis (sensu Hedges and Olkin 1985; Glass 1976; Arnqvist and Wooster 1995), where summary statistics across datasets are converted to standardized effect sizes for subsequent "higher order" statistical summaries or comparisons. As such, our approach enables evolutionary biologists to quantitatively examine the relative strength of phylogenetic signal across a wide range of phenotypic traits, and thus opens the door for future discoveries that inform on how phenotypic diversity accumulates in macroevolutionary time across the tree of life.

One important advantage of the approach advocated here is that the resulting effect sizes (Z_κ) are dimensionless, as the units of measurement cancel out during the calculation of Z (Sokal and Rohlf 2012). Thus, Z_κ represents the strength of phylogenetic signal on a common and comparable scale – measured in standard deviations – regardless of the initial units and original scale of the phenotypic variables under investigation. This means that the strength of phylogenetic signal may be compared across datsets for continuous phenotypic traits measured in different units and scale, because those units have been standardized through their conversion to Z_κ. For example, our approach could be utilized to determine whether the strength of phylogenetic signal (say, in response to ecological differentiation) is stronger in morphological traits (linear traits: mm), physiological traits (metabolic rate: $\frac{O^2}{min}$), or behavioral traits (aggression: $\frac{\#displays}{second}$). In fact, our empirical example provided such a comparison, as SA:V is represented in mm^{-1} while relative body size is a unitless ratio ($\frac{BW}{SVL}$). Additionally, our method is capable of comparing the strength of phylogenetic signal in traits of different dimensionality, as estimates of phylogenetic signal using κ have been generalized for multivariate data (κ_{mult}: see Adams 2014a). Furthermore, tests based on $\hat{Z}_{12}$ may be utilized for comparing the strength of phylogenetic signal among datasets containing a different number of species, and even for phenotypes obtained from species in different lineages, because their phylogenetic non-independence and observed variation are taken into account in the generation of the empirical sampling distribution via permutation.

This study is not the first to compare λ and κ for their ability as statistics to measure phylogenetic signal. Our results for λ and κ values are consistent with those found in the simulations performed by Münkemüller et al. (2012), but that study investigated type I error rates and statistical power, finding that λ performed better in both regards, irrespective of species number in trees. Although not the central focus of their study, the same tendency for variable λ and consistent κ at intermediate phylogenetic signal strengths was observed (see Fig. 2, Münkemüller et al. 2012). Recent work by Molina-Venegas and Rodríguez (2017) found that κ but not λ tended to inflate the estimate of phylogenetic signal, leading to moderate type I and type II

biases, if polytomic chronograms were used. Their work more thoroughly addressed previous observations of inflated κ for incompletely resolved phylogenetic trees (Davies et al. 2012; Münkemüller et al. 2012). An interesting question is whether an inflated κ value leads to an inflated Z_κ or does a tendency of a particular tree to inflate estimates of κ also inflate the values in random permutations of a test, in which case Z_κ is robust to polytomies? We repeated the analyses in Figure 4, adjusting trees to have 50% collapsed nodes, per the technique of Molina-Venegas and Rodríguez (2017), and found results were consistent (see Supporting Information). This confirms that any tendency of incompletely resolved trees to inflate κ as a descriptive statistic does not inflate Z_κ as an effect size. Furthermore, because comparison of effect sizes in a test is a comparison of locations of observed values in their sampling distributions, which would shift concomitantly because of this tendency, the Z_{12} test statistic in equation 4 appears to be robust in spite of unresolved trees.

Phylogenetic signal can be thought of as both an attribute to be measured in the data and a parameter that can be tuned to account for the phylogenetic non-independence among observations, for analysis of the data. As such, λ is appealing, as a statistic that potentially fulfills both roles. However, the inability to estimate phylogenetic signal with λ for data simulated with known phylogenetic signal is troublesome, and we recommend evolutionary biologists refrain from viewing it as a statistic to describe the amount of phylogenetic signal in the data. Interestingly, κ – when standardized to an effect size Z_κ – is a better statistic for measuring the amount of phylogenetic signal in data simulated with respect to known levels of λ. Although λ might be viewed as an important parameter for modifying the the conditional estimation of linear model coefficients with respect to phylogeny, it is neither a statistic that has meaningful comparative value as a measure of phylogenetic signal nor a statistic that lends itself well to reliable calculation of a test statistic. By contrast, κ has been shown here to be a reliable statistic, but only when standardized by the mean and standard deviation of its empirical sampling distribution (i.e., when converted to the effect size, Z_κ). Because one has control over the number of permutations used in analysis, one can be assured with many permutations that the empirical sampling distribution is representative of true probability distributions (Adams and Collyer 2018). With low coefficients of variation for Z_κ (Figure 4), it is difficult to imagine that a hypothesis test can improve equation 4 for efficiently comparing phylogenetic signal for different traits, different trees, or a combination of both.

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.

Arnqvist, G., and D. Wooster. 1995. Meta-analysis: Synthesizing research findings in ecology and evolution. Trends in Ecology and Evolution 10:236–240.

Bai, K., S. Lv, S. Ning, D. Zeng, Y. Guo, and B. Wang. 2019. Leaf nutrient concentrations associated with phylogeny, leaf habit and soil chemistry in tropical karst seasonal rainforest tree species. Plant and Soil 434:305–326.

Baken, E. K., and D. C. Adams. 2019. Macroevolution of arboreality in salamanders. Ecology and Evolution 9:7005–7016.

Baken, E. K., L. E. Mellenthin, and D. C. Adams. 2020. Macroevolution of desiccation-related morphology in plethodontid salamanders as inferred from a novel surface area to volume ratio estimation approach. Evolution 74:476–486.

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.

Bonett, R. M., and A. L. Blair. 2017. Evidence for complex life cycle constraints on salamander body form diversification. Proceedings of the National Academy of Sciences, U.S.A. 114:9936–9941.

Bose, R., B. R. Ramesh, R. Pélissier, and F. Munoz. 2019. Phylogenetic diversity in the western ghats biodiversity hotspot reflects environmental filtering and past niche diversification of trees. Journal of Biogeography 46:145–157.

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

Cohen, J. 1988. Statistical power analysis for the behavioral sciences. Routledge.

Collyer, M. L., and D. C. Adams. 2018. RRPP: An r package for fitting linear models to high-dimensional data using residual randomization. Methods in Ecology and Evolution 9:1772–1779.

Collyer, M. L., D. J. Sekora, and D. C. Adams. 2015. A method for analysis of phenotypic change for phenotypes described by high-dimensional data. Heredity 115:357–365.

Cooper, N., W. Jetz, and R. P. Freckleton. 2010. Phylogenetic comparative approaches for studying niche conservatism. Journal of Evolutionary Biology 23:2529–2539.

Davies, T. J., N. J. Kraft, N. Salamin, and E. M. Wolkovich. 2012. Incompletely resolved phylogenetic trees inflate estimates of phylogenetic conservatism. Ecology 93:242–247. Wiley Online Library.

De Meester, G., K. Huyghe, and R. Van Damme. 2019. Brain size, ecology and sociality: A reptilian perspective. Biological Journal of the Linnean Society 126:381–391.

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.

Glass, G. V. 1976. Primary, secondary, and meta-analysis of research. Educational Researcher 5:3–8.

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.

Hedges, L. V., and I. Olkin. 1985. Statistical methods for meta-analysis. Elsevier.

Kelley, K., and K. J. Preacher. 2012. On effect size. Psychological Methods 17:137–152.

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

Liu, H., C. P. Osborne, D. Yin, R. P. Freckleton, G. Jiang, and M. Liu. 2019. Phylogeny and ecological processes influence grass coexistence at different spatial scales within the steppe biome. Oecologia 191:25–38.

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. Rodríguez. 2017. Revisiting phylogenetic signal; strong or negligible impacts

of polytomies and branch length information? BMC evolutionary biology 17:53.

Münkemüller, 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. Freckleton, G. Thomas, T. Petzoldt, S. 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.

Pintanel, P., M. Tejedo, S. R. Ron, G. A. Llorente, and A. Merino-Viteri. 2019. Elevational and microclimatic drivers of thermal tolerance in andean pristimantis frogs. Journal of Biogeography 46:1664–1675.

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.

Rosenthal, R. 1994. The handbook of research synthesis. Pp. 231–244 *in* L. V. Cooper H Hedges, ed. Russell Sage Foundation.

Sokal, R. R., and F. J. Rohlf. 2012. Biometry. 4th ed. W.H. Freeman & Co., San Francisco.

505 Su, G., S. Villéger, and S. Brosse. 2019. Morphological diversity of freshwater fishes differs between realms,
506 but morphologically extreme species are widespread. Global ecology and biogeography 28:211–221.

507 Vandelook, F., S. Janssens, P. Gijbels, E. Fischer, W. Van den Ende, O. Honnay, and S. Abrahamczyk. 2019.
508 Nectar traits differ between pollination syndromes in balsaminaceae. Annals of Botany 124:269–279.

Figure Legends

Figure 1. Frequency distribution of λ estimates published in 2019. The majority of these values were close to 0 or 1, and from phylogenies with fewer than 200 taxa.

Figure 2. Precision of Pagel's λ across known levels of input phylogenetic signal (λ_{in}) on phylogenies of various sizes. As phylogenies increase in size, variation in λ_{in} decreases; however the precision is not constant across the range of input levels ($\lambda_{in} : 0 \rightarrow 1$), and is highest at intermediate levels of phylogenetic signal.

Figure 3. Precision of Pagel's λ when incorporated in phylogenetic regression ($Y \sim X$), across known levels of input phylogenetic signal (λ_{in}) on phylogenies of various sizes. As phylogenies increase in size, variation in λ_{in} decreases; however the precision is not constant across the range of input levels ($\lambda_{in} : 0 \rightarrow 1$), and is highest at intermediate levels of phylogenetic signal.

Figure 4. Variation in effect size estimates of phylogenetic signal across input levels of phylogenetic signal. (A) Estimates Z_λ for data simulated on phylogenies with 32 taxa ($n = 32$), (B) Estimates of Z_κ for data simulated on phylogenies with 32 taxa ($n = 32$), (C) Coefficients of variation of precision estimates of Z_λ and Z_κ across input levels of phylogenetic signal, estimated on phylogenies containing differing numbers of species.

Figure 5. (A) Linear measures for relative body size, and regions of the body used to estimate surface area to volume (SA:V) ratios. (B) Permutation distributions of phylogenetic signal for SA:V and $\frac{BW}{SVL}$, with observed values shown as vertical bars. (C) Effect sizes (Z_κ) for SA:V and $\frac{BW}{SVL}$, with their 95% confidence intervals (CI not standardized by $\sqrt{(}n)$).

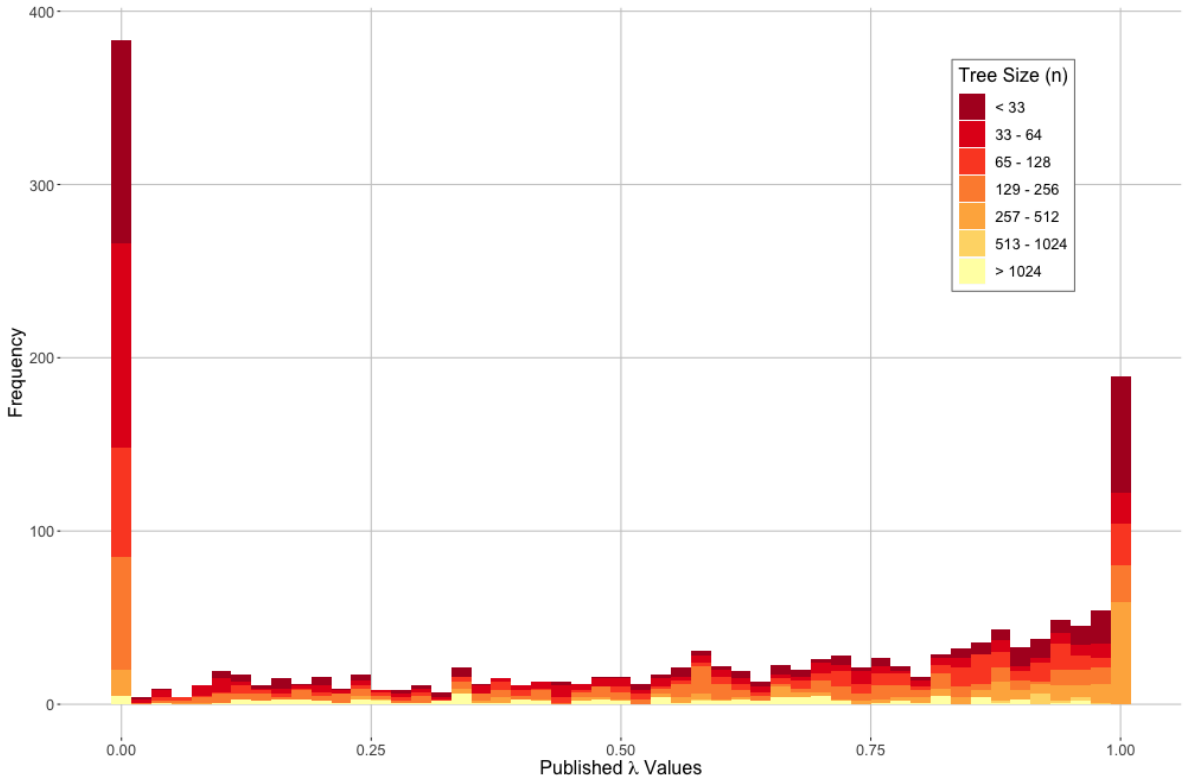

Figure 1. Frequency distribution of λ estimates published in 2019. The majority of these values were close to 0 or 1, and from phylogenies with fewer than 200 taxa.

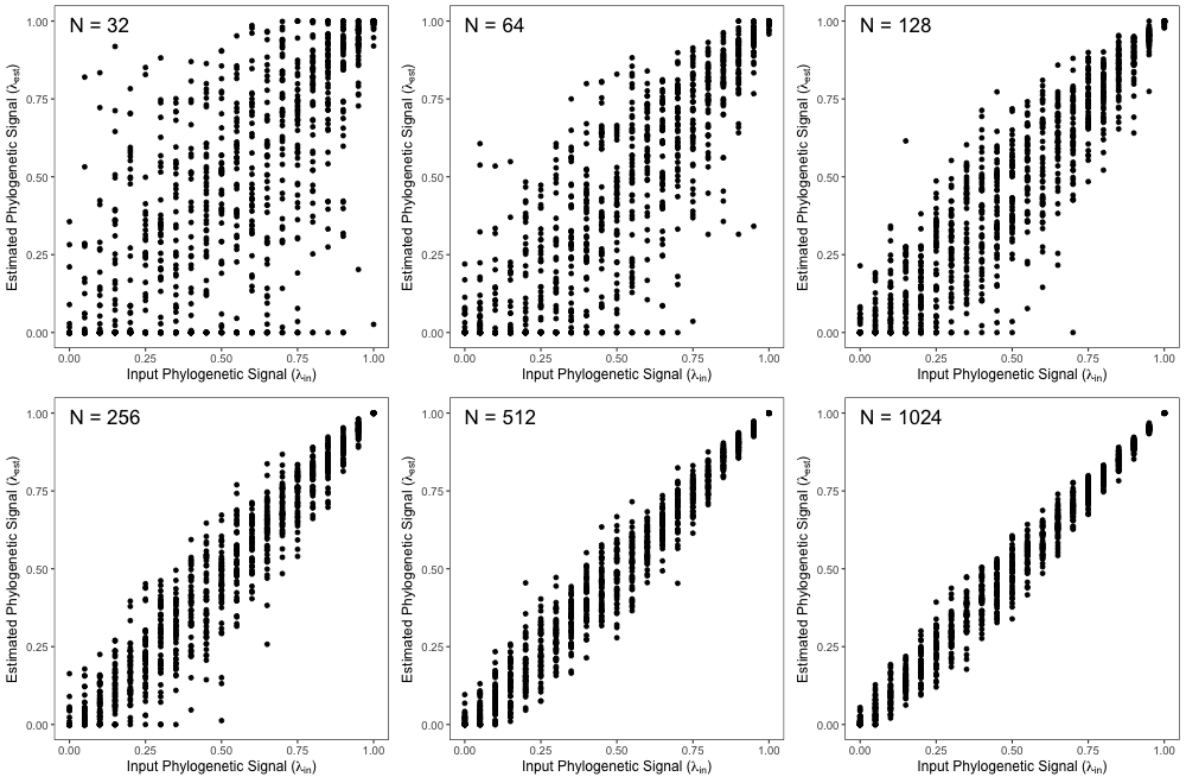

Figure 2. Precision of Pagel's λ across known levels of input phylogenetic signal (λ_{in}) on phylogenies of various sizes. As phylogenies increase in size, variation in λ_{in} decreases; however the precision is not constant across the range of input levels ($\lambda_{in} : 0 \rightarrow 1$), and is highest at intermediate levels of phylogenetic signal.

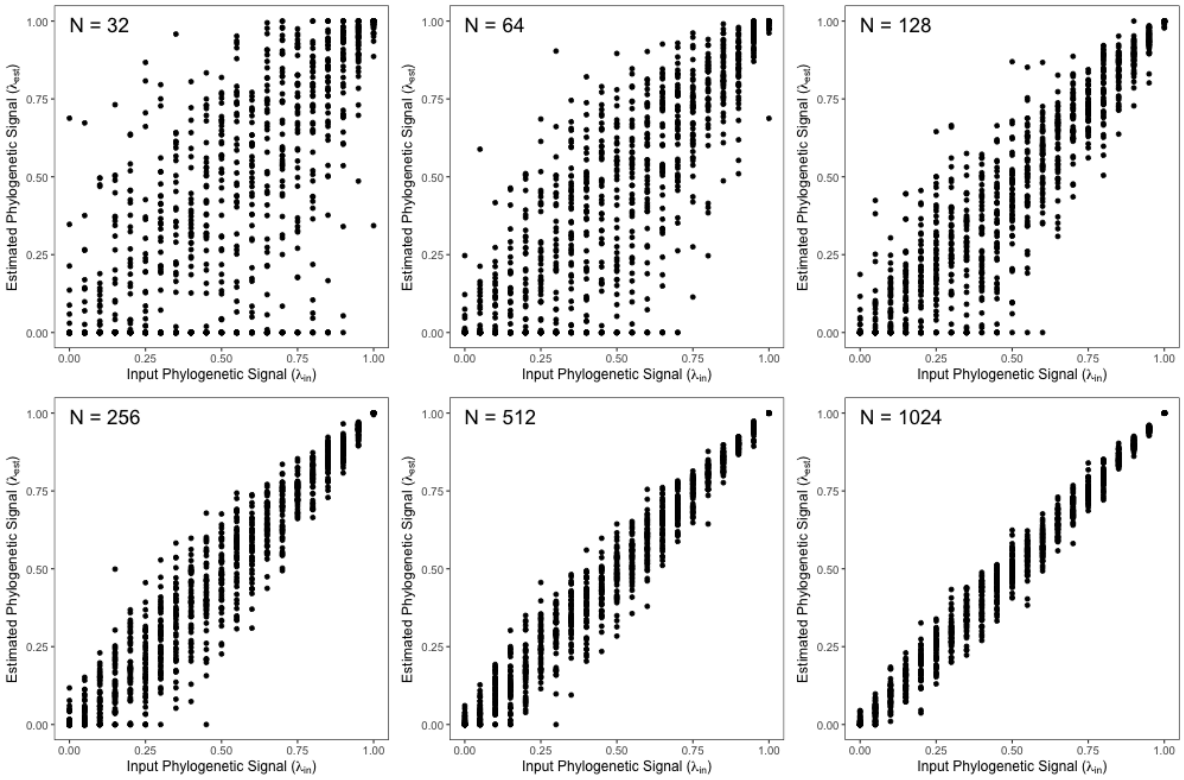

Figure 3. Precision of Pagel's λ when incorporated in phylogenetic regression ($Y \sim X$), across known levels of input phylogenetic signal (λ_{in}) on phylogenies of various sizes. As phylogenies increase in size, variation in λ_{in} decreases; however the precision is not constant across the range of input levels ($\lambda_{in} : 0 \rightarrow 1$), and is highest at intermediate levels of phylogenetic signal.

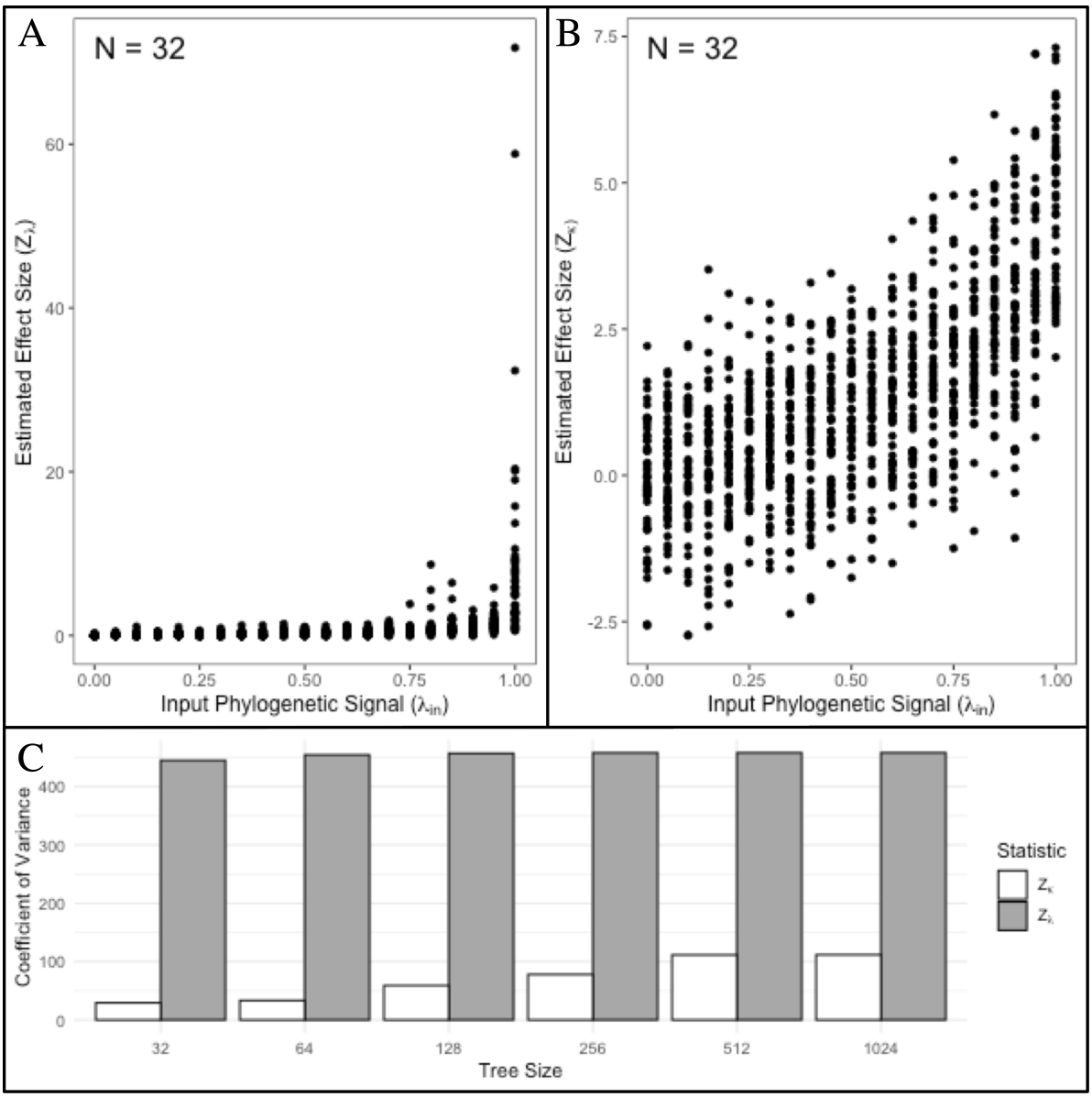

Figure 4. Variation in effect size estimates of phylogenetic signal across input levels of phylogenetic signal. (A) Estimates Z_λ for data simulated on phylogenies with 32 taxa ($n = 32$), (B) Estimates of Z_κ for data simulated on phylogenies with 32 taxa ($n = 32$), (C) Coefficients of variation of precision estimates of Z_λ and Z_κ across input levels of phylogenetic signal, estimated on phylogenies containing differing numbers of species.

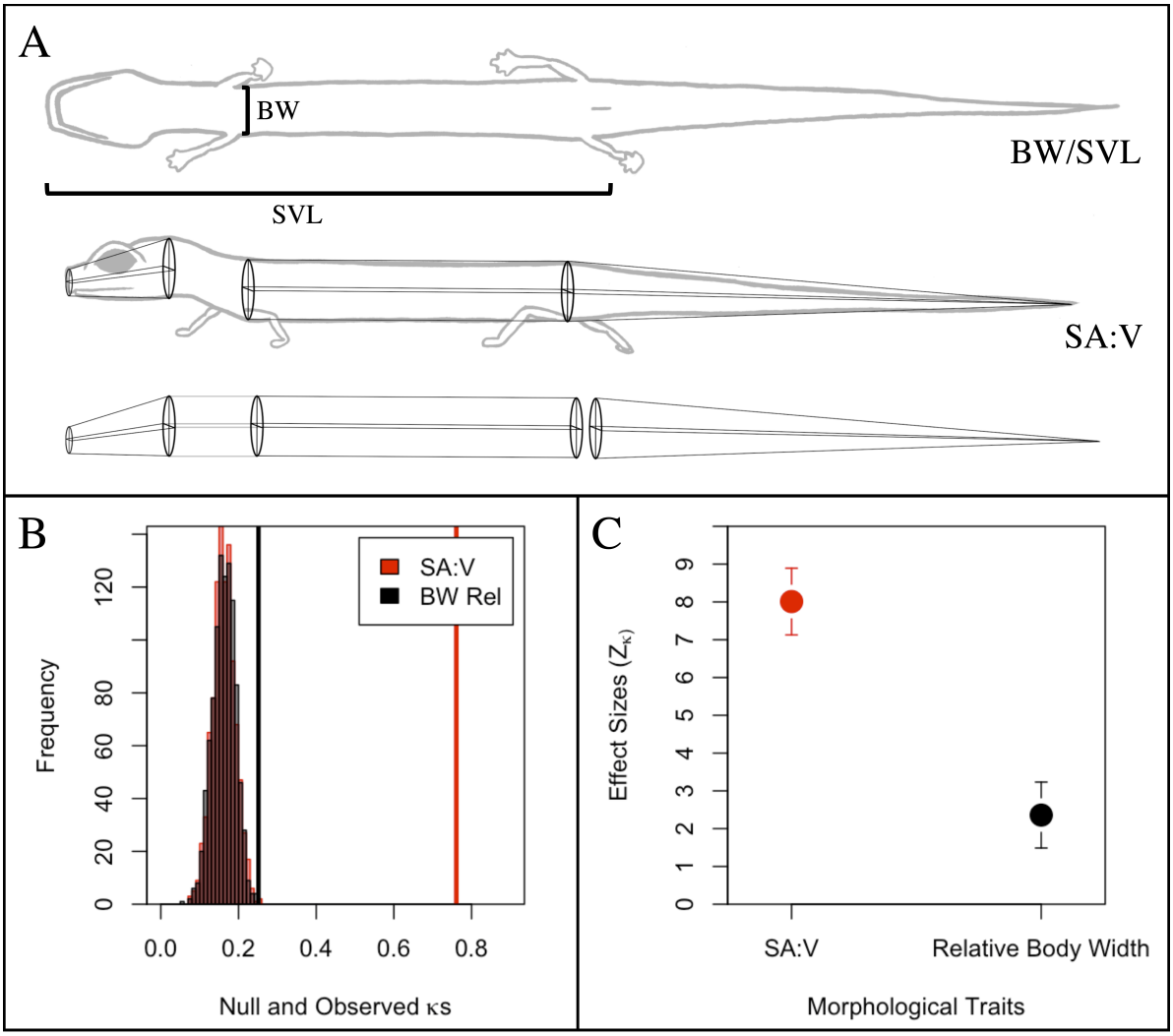

Figure 5. (A) Linear measures for relative body size, and regions of the body used to estimate surface area to volume (SA:V) ratios. (B) Permutation distributions of phylogenetic signal for SA:V and $\frac{BW}{SVL}$, with observed values shown as vertical bars. (C) Effect sizes (Z_κ) for SA:V and $\frac{BW}{SVL}$, with their 95% confidence intervals (CI not standardized by $\sqrt{(n)}$).