

¹ A Standardized Effect Size for Evaluating and Comparing the
² Strength of Phylogenetic Signal

³
⁴ **Dean C. Adams^{1,*}, Erica K. Baken^{1,2}, and Michael L. Collyer²**

⁵ 11 May, 2021

⁶ ¹Department of Ecology, Evolution, and Organismal Biology, Iowa State University, Ames, Iowa, USA.

⁷ ²Department of Science, Chatham University, Pittsburgh, Pennsylvania, USA.

⁸ *Correspondence: Dean C. Adams dcadams@iastate.edu

⁹

¹⁰ **Keywords:** comparative analysis, macroevolution

¹¹

¹² **Short Title:** Effect size for phylogenetic signal

¹³

¹⁴ **Research Article:**

words	characters
4221	26597

¹⁵ **Author Contributions:** DCA conceived the original idea for this manuscript, and DCA, EKB, and MLC
¹⁶ collaboratively developed the concept and contributed to all portions of this manuscript. All authors approve
¹⁷ of the final product and are willingly accountable for any portion of the content.

¹⁸

¹⁹ **Data Archiving:** Empirical data used in this paper are available on DRYAD (associated with original
²⁰ articles). R-scripts for simulation tests are found at: <https://github.com/deanadams/PhySigCompareZ>.
²¹ Computer code for implementing the two-sample comparison of effect sizes is found in geomorph (**upon**
²² **article acceptance**): <https://cran.r-project.org/web/packages/geomorph/index.html>

²⁴ **Acknowledgments:** We thank E. Glynne and B. Juarez for comments on early drafts of the manuscript. This
²⁵ work was sponsored in part by National Science Foundation Grants DBI-1902511 (to DCA) and DBI-1902694
²⁶ (to MLC). The authors have no known conflicts of interest.

27 **Abstract**

- 28 1. Macroevolutionary studies frequently characterize the phylogenetic signal in phenotypes, however,
29 analytical tools for comparing the strength of that signal across traits remain largely underdeveloped.
- 30 2. In this paper, we evaluate the efficacy of Pagel's λ to correctly estimate the strength of phylogenetic
31 signal in phenotypic traits across a range of input values. We find that λ behaves as a Bernoulli random
32 variable, where estimates are increasingly skewed at larger and smaller input levels of phylogenetic
33 signal. Further, the precision of λ varies with input signal.
- 34 3. Another measure, Blomberg's K , is more consistent across a range of tree sizes, and exhibits a positive
35 relationship with input levels of phylogenetic signal. However, that relationship is decidedly nonlinear.
36 Thus, neither λ nor K are suitable as effect sizes for measuring the strength of phylogenetic signal, and
37 comparing that signal across datasets.
- 38 4. As an alternative, we propose a standardized effect size based on K , (Z_K), which measures the strength
39 of phylogenetic signal more reliably than does λ , and places that signal on a common scale for statistical
40 comparison. We develop tests based on Z_K to provide a mechanism for formally comparing the strength
41 of phylogenetic signal across datasets, in much the same manner as effect sizes may be used to summarize
42 patterns in quantitative meta-analysis.
- 43 5. Our approach extends the phylogenetic comparative toolkit to address hypotheses that compare the
44 strength of phylogenetic signal between various phenotypic traits, even when those traits are found in
45 different evolutionary lineages or have different units or scales.

46 1 Introduction

47 Investigating macroevolutionary patterns of trait variation requires a phylogenetic perspective, because the
48 shared ancestry of species violates the assumption of independence among trait values that is common for
49 statistical tests (Felsenstein, 1985; Harvey & Pagel, 1991). Accounting for this evolutionary non-independence
50 is the purview of *phylogenetic comparative methods* (PCMs): a suite of analytical tools that condition the
51 data by the phylogenetic relatedness of observations (Grafen, 1989; Martins & Hansen, 1997; Garland & Ives,
52 2000; Rohlf, 2001; O'Meara, Ane, Sanderson, & Wainwright, 2006; Beaulieu, Jhuang, Boettiger, & O'Meara,
53 2012; Adams, 2014b; Adams & Collyer, 2018). PCMs are predicated on the notion that phylogenetic signal –
54 the tendency for closely related species to display similar trait values – is present in cross-species datasets
55 (Felsenstein, 1985; Pagel, 1999; Blomberg, Garland, & Ives, 2003). Indeed, under numerous evolutionary
56 models, phylogenetic signal is expected, as stochastic character change along the hierarchical structure of the
57 tree of life generates trait covariation among taxa (Felsenstein, 1985; Blomberg et al., 2003; Revell, Harmon,
58 & Collar, 2008).

59

60 Several analytical tools have been developed to quantify phylogenetic signal in phenotypic datasets (Gittleman
61 & Kot, 1990; Abouheif, 1999; Pagel, 1999; Blomberg et al., 2003; Klingenberg & Gidaszewski, 2010; Adams,
62 2014a), and their statistical properties – namely type I error rates and statistical power – have been
63 investigated to determine under what conditions phylogenetic signal can be detected (Revell et al., 2008;
64 Revell, 2010; Boettiger, Coop, & Ralph, 2012; Diniz-Filho, Santos, Rangel, & Bini, 2012; Münkemüller et
65 al., 2012; Pavoine & Ricotta, 2012; Adams, 2014a; Molina-Venegas & Rodríguez, 2017). One of the most
66 widely used methods for characterizing phylogenetic signal is Pagel's λ (Pagel, 1999), which transforms the
67 lengths of the internal branches of the phylogeny to improve the fit of data to the phylogeny via maximum
68 likelihood (Pagel, 1999; Freckleton, Harvey, & Pagel, 2002). When incorporated in PGLS, λ serves as a
69 tuning parameter which is optimized via log-likelihood profiling while evaluating the covariation between
70 the dependent and independent variables, given the phylogeny (Pagel, 1999; Freckleton et al., 2002). To
71 infer whether phylogenetic signal differs from no signal or a Brownian motion (BM) model of evolutionary
72 divergence, the observed model fit using $\hat{\lambda}$ may be statistically compared to that using $\lambda = 0$ or $\lambda = 1$ via
73 likelihood ratio tests (Freckleton et al., 2002; Cooper, Jetz, & Freckleton, 2010; Bose, Ramesh, Péliéssier, &
74 Munoz, 2019) or confidence limits (Vandelook et al., 2019).

75

76 Another widely used measure is Blomberg's K (Blomberg et al., 2003), which characterizes phylogenetic

77 signal as the ratio of observed trait variation to the amount of variation expected under Brownian motion.
78 Blomberg's K can be treated as a test statistic by employing a permutation test to generate its sampling
79 distribution (Blomberg et al., 2003; Adams, 2014a) for determining whether significant phylogenetic signal
80 is present in data. Both λ and K seem intuitive to interpret, as a value of 0 for both corresponds to no
81 phylogenetic signal, while a value of 1 corresponds to the amount of phylogenetic signal expected under
82 Brownian motion. Thus, it is tempting to regard both λ and K as descriptive statistics that measure the
83 relative strength of phylogenetic signal, providing an estimate of its magnitude for comparison.

84

85 The appeal of Pagel's λ and Blomberg's K as descriptive statistics is that they provide a basis for
86 interpreting "weak" versus "strong" phylogenetic signal; i.e., small versus large values of $\hat{\lambda}$ or K , respectively,
87 in a comparative sense (De Meester, Huyghe, & Van Damme, 2019; Pintanel, Tejedo, Ron, Llorente, &
88 Merino-Viteri, 2019; Su, Villéger, & Brosse, 2019). Nonetheless, an important question that has yet to be
89 considered is whether such comparisons are analytically appropriate, and whether these statistics are, or
90 can be, converted to effect sizes for comparative analyses across datasets. To be statistics representing
91 phylogenetic signal, they should have reliable distributional properties, which could be revealed with
92 simulation experiments. For instance, as a proportional random variable bounded by 0 and 1, we might
93 expect that $\hat{\lambda}$ is a random variable that follows the distribution of a Bernoulli probability parameter (Forbes,
94 Evans, Hastings, & Peacock, 2011); i.e., branch lengths in a tree are scaled proportionally to the probability
95 that data arise from a BM process. Given a known λ value used to generate random data on a tree, we
96 would also expect that the mean of an empirical sampling distribution of $\hat{\lambda}$ would approximately equal λ ; the
97 dispersion of $\hat{\lambda}$ would be largest at intermediate values of λ , $\hat{\lambda}$ would be predictable over the range of λ with
98 respect to tree size; the distribution of $\hat{\lambda}$ would be symmetric at intermediate values of λ and more skewed
99 toward values of 0 or 1; and that the distribution of $\hat{\lambda}$ would be more platykurtic at intermediate values of
100 λ , becoming more leptokurtic toward 0 and 1 (Forbes et al., 2011). Prior work (Münkemüller et al., 2012)
101 seems to support some of these conjectures, based superficially on statistical moments for a given tree size
102 (mean, variance, skewness, and kurtosis; see Fig. 2 of ref. (Münkemüller et al., 2012)). However, because
103 the "strength of Brownian motion" was simulated as a varied weighted-average of data simulated on trees
104 with $\lambda = 0$ and $\lambda = 1$ and not as prescribed values of λ (Münkemüller et al., 2012), interpretation of these
105 patterns is challenging.

106

107 By contrast, for Blomberg's K , which is positively unbounded, we might expect that for any λ used to generate
108 data, estimates of K might be a random variable that follows a normal distribution, with values distributed

109 symmetrically (Forbes et al., 2011). This attribute seemed less reasonable based on the simulations performed
110 by Münkemüller et al. (2012), which suggested that distributions were positively skewed and that Blomberg's
111 K might not behave as a statistic that follows a normal distribution. However, because their simulations used a
112 weighted combination of simulated phylogenetic signal strengths, strong inferences are not possible (and distri-
113 butional attributes were not the intended result of their simulations). Thus, for both Pagel's λ or Blomberg's K ,
114 evaluation of statistical moments across a range of λ used to generate data would be valuable for adjudicating
115 the reliability of these statistics as effect sizes. Furthermore, the expected values of these statistics appear to
116 vary with tree size (Münkemüller et al., 2012), making comparisons across studies challenging. Therefore,
117 transformation of these statistics into Z -scores would allow evaluation of the efficacy of each statistic to yield
118 effect sizes that could be used for comparisons of the strength of phylogenetic signal across traits and lineages.

119

120 Here we use simulation experiments to compare the distributional attributes of $\hat{\lambda}$ and K , plus their effect
121 sizes (Z -scores), across a range of tree size and phylogenetic signal strength. We find that estimates of $\hat{\lambda}$
122 are increasingly skewed at larger and smaller input levels of phylogenetic signal and at smaller tree sizes,
123 vary widely for a given input value of λ , and that the precision of $\hat{\lambda}$ is not constant across its range. By
124 contrast, estimates of K are more consistent across tree sizes, and are normally distributed across the range
125 of input levels of λ , making K a more reliable statistic. However, the relationship between K and input
126 levels of phylogenetic signal is decidedly nonlinear. Thus, neither λ nor K are suitable as effect sizes for
127 measuring the strength of phylogenetic signal. As an alternative, we propose an effect size based on K , (Z_K),
128 which provides consistent estimates of the strength of phylogenetic signal across tree sizes and signal strength.
129 Further, because Z_K places that phylogenetic signal on a common scale, it facilitates statistical comparisons
130 of the relative strength of phylogenetic signal across datasets. We propose a two-sample statistic (\hat{Z}_{12}) to
131 accomplish this task, and show that it displays appropriate levels of type I error and model misspecification.
132 An empirical example is then provided to illustrate its use.

133 2 Simulation methods and results

134 Methods for characterizing phylogenetic signal were evaluated by computer simulation. Briefly, simulations
135 were conducted by generating pure-birth phylogenies at each of six different tree sizes ($n = 2^5, 2^6, \dots, 2^{10}$),
136 and with differing levels of phylogenetic signal ($\lambda = 0.0, 0.5, \dots, 1.0$). We generated 50 random trees for each
137 intersection of tree size and λ . For each λ within each tree size, continuous traits were then simulated on
138 each phylogeny under a BM model of evolution. For each set of 50 trees we measured the mean values of $\hat{\lambda}$
139 and K , their standard deviation, and calculated the Shapiro-Wilk W statistic as a departure from normality

¹⁴⁰ (symmetry). For the latter, a value of 1.0 indicates normally distributed values, while departures from 1.0
¹⁴¹ indicate skewness. Simulations were then repeated for both balanced and pectinate trees, which yielded
¹⁴² qualitatively similar results (see Supporting Information). Trees containing polytomies, and an evaluation
¹⁴³ of $\hat{\lambda}$ from models of linear regression and phylogenetic ANOVA, were also investigated, and results were
¹⁴⁴ qualitatively similar to those reported above (see Supporting Information).

¹⁴⁵

¹⁴⁶ 2.1 Lambda (λ) estimates of phylogenetic signal are inaccurate

¹⁴⁷ Computer simulations reveal that for $\hat{\lambda}$, the distributional expectations of a Bernoulli variable were mostly
¹⁴⁸ upheld. First, the mean value of $\hat{\lambda}$ increases as λ increases. Second, the precision in estimating λ varies
¹⁴⁹ across the range of input values, as the standard deviation of $\hat{\lambda}$ is largest at intermediate values of λ and
¹⁵⁰ smallest at extreme values (Fig. 1 red line). Third, the distributions of $\hat{\lambda}$ tend toward normal distributions at
¹⁵¹ intermediate levels of λ but become increasingly skewed at more extreme values of λ (Fig. 1 blue line). For
¹⁵² small tree sizes, it is also clear that distributions are more platykurtic at intermediate values of $\hat{\lambda}$. However,
¹⁵³ the mean value of $\hat{\lambda}$ is negatively-biased (particularly for small tree sizes but also consistently across most of
¹⁵⁴ its range; Fig. 1 black line) and standard deviations of $\hat{\lambda}$ are negatively associated with tree size. For trees
¹⁵⁵ of 128 species or less, $\hat{\lambda}$ are quite variable, except for cases when λ is near or equal to 1. Taken together
¹⁵⁶ these results reveal that $\hat{\lambda}$ is a biased statistic that inconsistently estimates phylogenetic signal, both across
¹⁵⁷ tree sizes and across the range of input values. Additional simulations (Supporting Information) reveal that
¹⁵⁸ incorporating $\hat{\lambda}$ in PGLS ANOVA and regression does not adversely affect the statistical properties of PGLS
¹⁵⁹ parameter estimation or model evaluation (type I error, power, bias in coefficients). Thus, it is reasonable to
¹⁶⁰ incorporate $\hat{\lambda}$ in PGLS as a parameter for tuning the degree of phylogenetic signal in the dependent variables
¹⁶¹ during the analysis. However, the statistical properties shown in Fig. 1 demonstrate that λ is unsuitable as
¹⁶² an effect size for measuring the strength of phylogenetic signal in data, and thus λ should not be used for
¹⁶³ comparing phylogenetic signal across datasets.

¹⁶⁴ 2.2 Kappa (K) estimates of phylogenetic signal are more reliable

¹⁶⁵ Simulation results demonstrate that K displays better statistical properties. First, as expected, mean values
¹⁶⁶ of K increase with increasing signal (λ) irrespective of tree size, albeit non-linearly (Fig. 2 black line). Second,
¹⁶⁷ the standard deviation of K is consistent across tree sizes (Fig. 2 red line), and while it increases with λ ,
¹⁶⁸ it is always less than the mean (low coefficient of variation). This finding is perhaps unsurprising, as K is
¹⁶⁹ lower-bounded by 0, and is never large for small values of λ . Importantly, K is normally distributed across

170 the range of input λ ; a consistent pattern regardless of tree size (Fig. 2 blue line). This differs from results
 171 of (Münkemüller et al., 2012), where the skewing appears to be due to combining random values generated
 172 independently, rather than being a property of K itself. Overall, these findings reveal that while K is more
 173 reliable as an estimate of phylogenetic signal, the non-linear scaling with input signal implies that it should
 174 not be considered an effect size that measures the strength of phylogenetic signal on a common scale for
 175 comparison across datasets.

176 2.3 Effect sizes from K (Z_K) better characterize phylogenetic signal

177 To measure the strength of phylogenetic signal on a common scale, we propose effect sizes (Z-scores) for both
 178 λ and K . Statistically, a standardized effect size may be found as:

$$179 Z_\theta = \frac{\theta_{obs} - E(\theta)}{\sigma_\theta} \quad (1)$$

180 where θ_{obs} is the observed test statistic, $E(\theta)$ is its expected value under the null hypothesis, and σ_θ is its
 181 standard error (Glass, 1976; Cohen, 1988; Rosenthal, 1994). Typically, θ_{obs} and σ_θ are estimated from the
 182 data, while $E(\theta)$ is obtained from the distribution of θ derived from parametric theory. However, recent
 183 advances in resampling theory (Collyer, Sekora, & Adams, 2015; Adams & Collyer, 2016, 2019a; Collyer &
 184 Adams, 2018) have shown that $E(\theta)$ and σ_θ may also be obtained from an empirical sampling distribution of
 185 θ simulated from permutation procedures.

186 Formalizing the suggestion of Adams and Collyer (Adams & Collyer, 2019b), an effect size for K may be
 187 found as:

$$188 Z_K = \frac{K_{obs} - \hat{\mu}_K}{\hat{\sigma}_K}, \quad (2)$$

189 where K_{obs} is the observed phylogenetic signal, and $\hat{\mu}_K$ and $\hat{\sigma}_K$ are the mean and standard deviation of the
 190 empirical sampling distribution of K obtained via permutation. The empirical sampling distribution of K
 191 can be first transformed via a Box-Cox transformation to better adhere to the assumption of normality.

192 For λ , deriving an effect size is more challenging, as λ does not have a sampling distribution from which the
 193 standard error and confidence intervals may be obtained, and estimates from the Hessian matrix from PGLS
 194 are unreliable (Boettiger et al., 2012). Confidence intervals are therefore generated for the values of λ that
 195 intersect the log-likelihood profile for corresponding percentiles of the χ^2 distribution used to compare the
 196 putative model to a null model with $\lambda = 0$ (Orme et al., 2013). Thus, an effect size for λ may be found as:

$$|Z_\lambda| = \sqrt{\chi_{\hat{\lambda}}^2} \quad (3)$$

197 where, $\hat{\lambda}$ is the maximized likelihood value of λ and $\chi_{\hat{\lambda}}^2$ is the likelihood ratio statistic for the value.

198

199 Simulations reveal that both Z_λ and Z_K are associated with input phylogenetic signal (λ), indicating that
 200 both statistics capture the observed signal (Fig. 3). However, effect sizes from $\hat{\lambda}$ made little sense, as they
 201 are more strongly associated with tree size than they are with the actual phylogenetic signal in the data (Fig.
 202 3). By contrast, Z_K is much more consistent across tree sizes, and increases more linearly with increasing
 203 levels of phylogenetic signal. Additionally, Z_K exhibits a much stronger association with phylogenetic signal
 204 strength as compared to tree size (Fig. 3), and its standard deviation is more consistent, implying similar
 205 levels of precision across the range of input signal (Supporting Information). Thus, Z_K is a more reliable
 206 measure of the strength of phylogenetic signal, and may be used to compare levels of phylogenetic signal
 207 across datasets.

208 2.4 A test statistic (\hat{Z}_{12}) allows meaningful comparisons across datasets

209 To statistically compare the strength of phylogenetic signal across datasets we propose a two-sample test
 210 statistic (\hat{Z}_{12}). Based on statistical theory, a two-sample test statistic may be calculated as:

$$\hat{Z}_{12} = \frac{|(K_1 - \hat{\mu}_{K_1}) - (K_2 - \hat{\mu}_{K_2})|}{\sqrt{\hat{\sigma}_{K_1}^2 + \hat{\sigma}_{K_2}^2}} \quad (4)$$

211 where K_1 , K_2 , $\hat{\mu}_{K_1}$, $\hat{\mu}_{K_2}$, $\hat{\sigma}_{K_1}$, and $\hat{\sigma}_{K_2}$ are as defined above. Estimates of significance of \hat{Z}_{12} may be
 212 obtained from a standard normal distribution, or permutation. Typically, \hat{Z}_{12} is considered a two-tailed test,
 213 however directional (one-tailed) tests may be specified should the empirical situation require it (Adams &

214 Collyer, 2016, 2019a).

215

216 We evaluated the type I error and false discovery rates of tests based on \hat{Z}_{12} using a procedure similar to that
217 above. Simulations were performed using pure-birth trees of different sizes ($n = 2^5, 2^6, \dots, 2^{10}$), and with
218 differing levels of phylogenetic signal ($\lambda = 0.0, 0.5, \dots, 1.0$). For each combination of n and λ , a total of 100
219 traits were simulated on pure-birth trees (50 traits per tree) under a Brownian motion model of evolution.
220 Next, the phylogenetic signal between traits was compared in pairwise fashion for all combinations of traits,
221 using \hat{Z}_{12} . The proportion of significant results provided an estimate of the type I error or false discovery
222 rate. Specifically, type I error was evaluated when traits were simulated using no input phylogenetic signal
223 (i.e., $\lambda = 0$). Likewise, false discovery rates were evaluated for data simulated with some known, but equal,
224 level of phylogenetic signal (i.e., $\lambda > 0$).

225

226 Tests revealed that across simulation conditions, the average type I error of \hat{Z}_{12} was approximately 0.05
227 (Fig. 4). Likewise, average false discovery rates were also low, and were at or below the nominal 5% level
228 (Fig. 4). Importantly, the statistical performance of \hat{Z}_{12} appears unaffected by phylogeny size (Supporting
229 Information). Overall, these results reveal that tests based on \hat{Z}_{12} have acceptable type I error and false
230 discovery rates. Therefore, \hat{Z}_{12} is an appropriate statistic for comparing the degree of phylogenetic signal
231 across traits.

232 3 Empirical example

233 To demonstrate the utility of \hat{Z}_{12} , we compared Z_K for two ecologically-relevant traits in plethodontid
234 salamander (Fig. 5): surface area to volume ratios (SA:V) and relative body width ($\frac{BW}{SVL}$) (Baken
235 & Adams, 2019; Baken, Mellenthin, & Adams, 2020). For this example, surface area to volume
236 ratios (SA:V) and relative body width ($\frac{BW}{SVL}$) measures were obtained from individuals of 305 species,
237 from which species means were obtained (Baken & Adams, 2019; Baken et al., 2020). A time-dated
238 molecular phylogeny for the group (Bonett & Blair, 2017) was pruned to match the species in the
239 phenotypic dataset. The phylogenetic signal in each trait was then characterized using K , which was
240 converted to its effect size (Z_K) using **geomorph** 3.3.1 (Adams & Otárola-Castillo, 2013; Adams, Collyer, &
241 Kaliantzopoulou, 2020), and routines by the authors (**to be incorporated in geomorph upon acceptance**).

242

243 While both traits contained significant phylogenetic signal, tests based on \hat{Z}_{12} revealed that the degree of

²⁴⁴ phylogenetic signal was significantly stronger in SA:V ($\hat{Z}_{12} = 16.51$; $P < 0.00001$: Fig. 5). Biologically, this
²⁴⁵ observation may be interpreted by the fact that the tropical species – which form a monophyletic group
²⁴⁶ within plethodontids – display greater variation in SA:V, which covaries with disparity in their climatic niches
²⁴⁷ (Baken et al., 2020). Thus, greater phylogenetic signal in SA:V is to be expected.

²⁴⁸ 4 Discussion

²⁴⁹ It is common in comparative evolutionary studies to characterize the phylogenetic signal in phenotypic traits
²⁵⁰ to determine the extent to which shared evolutionary history has generated trait covariation among taxa.
²⁵¹ However, while numerous analytical approaches may be used to quantify phylogenetic signal (Gittleman &
²⁵² Kot, 1990; e.g., Abouheif, 1999; Pagel, 1999; Blomberg et al., 2003; Adams, 2014a), methods that explicitly
²⁵³ measure the strength of phylogenetic signal, or facilitate comparisons among datasets, have remained
²⁵⁴ underdeveloped. We evaluated the precision of one common measure, Pagel's λ , and explored its efficacy
²⁵⁵ for characterizing the strength of phylogenetic signal in phenotypic data. Using computer simulations, we
²⁵⁶ found that λ behaves as a Bernoulli random variable, with estimates that are increasingly skewed at larger
²⁵⁷ and smaller input levels of phylogenetic signal. Further, the precision of λ in estimating actual levels of
²⁵⁸ phylogenetic signal varies with both tree size (see also ref. Boettiger et al. (2012)) and input levels of
²⁵⁹ phylogenetic signal. From these findings we conclude that λ is not a reliable indicator of the observed
²⁶⁰ strength of phylogenetic signal in phenotypic datasets, and should not be used as an effect size for comparing
²⁶¹ the degree of phylogenetic signal between datasets.

²⁶²

²⁶³ As an alternative, we described a standardized effect size (Z) for assessing the strength of phylogenetic
²⁶⁴ signal. Z expresses the magnitude of phylogenetic signal as a standard normal deviate, which is easily
²⁶⁵ interpretable as the strength of phylogenetic signal relative to the mean. We applied this concept to both
²⁶⁶ λ and K , and found that Z_K was a better estimate of the strength of phylogenetic signal in phenotypic
²⁶⁷ data. First, values of Z_K more accurately tracked known changes in the magnitude of phylogenetic signal,
²⁶⁸ as demonstrated by the near linear relationship between Z_K and input signal. Additionally, the precision
²⁶⁹ of Z_K was more consistent across the range of input levels of phylogenetic signal (Fig S1; Supporting
²⁷⁰ Information). Thus, Z_K is a more reliable measure of the relative strength of phylogenetic signal, and
²⁷¹ places that effect on a common and comparable scale. We therefore recommend that future studies in-
²⁷² terested in evaluating the strength of phylogenetic signal incorporate Z_K as a statistical measure of this effect.

²⁷³

274 Next we proposed a two-sample test (\hat{Z}_{12}), which provides a formal statistical procedure for determining
275 whether the strength of phylogenetic signal is greater in one phenotypic trait as compared to another. Prior
276 studies have summarized patterns of variation in phylogenetic signal across datasets using summary test
277 values, such as K (e.g., Blomberg et al., 2003). However, because K does not scale linearly with input
278 levels of phylogenetic signal (Fig. 2), and its variance increases with increasing strength of phylogenetic
279 signal (Diniz-Filho et al., 2012; Münkemüller et al., 2012), it should not be considered an effect size
280 that measures the strength of phylogenetic signal on a common scale. By contrast, standardizing K
281 to Z_K via equation 2 alleviates these concerns, and facilitates formal statistical comparisons of the
282 strength of signal across datasets. Thus when viewed from this perspective, the approach developed
283 here aligns well with other statistical approaches such as meta-analysis (Glass, 1976; sensu Hedges
284 & Olkin, 1985; Arnqvist & Wooster, 1995), where summary statistics across datasets are converted
285 to standardized effect sizes for subsequent “higher order” statistical summaries or comparisons. As
286 such, our approach enables evolutionary biologists to quantitatively examine the relative strength of
287 phylogenetic signal across a wide range of phenotypic traits, and thus opens the door for future dis-
288 coveries that inform on how phenotypic diversity accumulates in macroevolutionary time across the tree of life.

289

290 One important advantage of the approach advocated here is that the resulting effect sizes (Z_K) are
291 dimensionless, as the units of measurement cancel out during the calculation of Z (Sokal & Rohlf, 2012).
292 Thus, Z_K represents the strength of phylogenetic signal on a common and comparable scale – measured
293 in standard deviations – regardless of the initial units and original scale of the phenotypic variables under
294 investigation. This means that the strength of phylogenetic signal may be compared across datasets for
295 continuous phenotypic traits measured in different units and scale, because those units have been standardized
296 through their conversion to Z_K . For example, our approach could be utilized to determine whether the
297 strength of phylogenetic signal (say, in response to ecological differentiation) is stronger in morphological
298 traits (linear traits: mm), physiological traits (metabolic rate: $\frac{O^2}{min}$), or behavioral traits (aggression:
299 $\frac{\#displays}{second}$). In fact, our empirical example provided just such a comparison, as SA:V is represented in mm^{-1}
300 while relative body size is a unitless ratio ($\frac{BW}{SVL}$). Additionally, our method is capable of comparing the
301 strength of phylogenetic signal in traits of different dimensionality, as estimates of phylogenetic signal using K
302 have been generalized for multivariate data (κ_{mult} : see Adams, 2014a). Furthermore, tests based on \hat{Z}_{12} may
303 be utilized for comparing the strength of phylogenetic signal among datasets containing a different number of
304 variables, and even for phenotypes obtained from species in different lineages, because their phylogenetic
305 non-independence and observed variation are taken into account in the generation of the empirical sampling

306 distribution via permutation.

307

308 This study is not the first to compare λ and K for their ability as statistics to measure phylogenetic signal.
309 Our results for λ and K values are consistent with those found in the simulations performed by Münkemüller
310 et al. (2012), but that study investigated type I error rates and statistical power, finding that λ performed
311 better in both regards, irrespective of species number in trees. Although not the central focus of their study,
312 the same tendency for variable λ and consistent K at intermediate phylogenetic signal strengths was observed
313 (Fig. 2 of ref. (Münkemüller et al., 2012)). Recent work by Molina-Venegas and Rodríguez (2017) found that
314 K but not λ tended to inflate the estimate of phylogenetic signal, leading to moderate type I and type II
315 biases, if polytomic chronograms were used. Their work more thoroughly addressed previous observations
316 of inflated K for incompletely resolved phylogenetic trees (Davies, Kraft, Salamin, & Wolkovich, 2012;
317 Münkemüller et al., 2012). An interesting question is whether an inflated K value leads to an inflated Z_K or
318 does a tendency of a particular tree to inflate estimates of K also inflate the values in random permutations
319 of a test, in which case Z_K is robust to polytomies? We repeated the analyses in Figs. 1 & 2, adjusting trees
320 to have 20% collapsed nodes, per the technique of Molina-Venegas and Rodríguez (2017), and found results
321 were consistent (Supporting Information). This confirms that any tendency of incompletely resolved trees to
322 inflate K as a descriptive statistic does not inflate Z_K as an effect size. Furthermore, because comparison of
323 effect sizes in a test is a comparison of locations of observed values in their sampling distributions, which
324 would shift concomitantly because of this tendency, the Z_{12} test statistic in equation 4 appears to be robust
325 in spite of unresolved trees.

326

327 Phylogenetic signal can be thought of as both an attribute to be measured in the data and a parameter
328 that can be tuned to account for the phylogenetic non-independence among observations, for analysis of
329 the data. As such, λ is appealing, as a statistic that potentially fulfills both roles. However, the inability
330 to estimate phylogenetic signal with λ for data simulated with known phylogenetic signal is troublesome,
331 and we recommend evolutionary biologists refrain from viewing it as a statistic to describe the amount of
332 phylogenetic signal in the data. Interestingly, K – when standardized to an effect size Z_K – is a better
333 statistic for measuring the amount of phylogenetic signal in data simulated with respect to known levels of
334 λ . Although λ might be viewed as an important parameter for modifying the the conditional estimation of
335 linear model coefficients with respect to phylogeny, it is neither a statistic that has meaningful comparative
336 value as a measure of phylogenetic signal nor a statistic that lends itself well to reliable calculation of a test
337 statistic. By contrast, K has been shown here to be a reliable statistic, but only when standardized by the

338 mean and standard deviation of its empirical sampling distribution (i.e., when converted to the effect size,
339 Z_K). Because one has control over the number of permutations used in analysis, one can be assured with
340 many permutations that the empirical sampling distribution is representative of true probability distributions
341 (Adams & Collyer, 2018). Given the greater consistency in estimates of Z_K across tree sizes and input signal,
342 it is difficult to imagine a hypothesis test that can improve equation 4 for efficiently comparing phylogenetic
343 signal for different traits, different trees, or a combination of both.

344 **References**

- 345 Abouheif, E. (1999). A method for testing the assumption of phylogenetic independence in comparative
346 data. *Evolutionary Ecology Research*, 1, 895–909. Journal Article.
- 347 Adams, D. C. (2014a). A generalized Kappa statistic for estimating phylogenetic signal from shape and
348 other high-dimensional multivariate data. *Systematic Biology*, 63, 685–697.
- 349 Adams, D. C. (2014b). A method for assessing phylogenetic least squares models for shape and other
350 high-dimensional multivariate data. *Evolution*, 68, 2675–2688.
- 351 Adams, D. C., & Collyer, M. L. (2016). On the comparison of the strength of morphological integration
352 across morphometric datasets. *Evolution*, 70, 2623–2631. Journal Article.
- 353 Adams, D. C., & Collyer, M. L. (2018). Phylogenetic ANOVA: Group-clade aggregation, biological
354 challenges, and a refined permutation procedure. *Evolution*, 72(6), 1204–1215.
- 355 Adams, D. C., & Collyer, M. L. (2019a). Comparing the strength of modular signal, and evaluating
356 alternative modular hypotheses, using covariance ratio effect sizes with morphometric data. *Evolution*,
357 73(12), 2352–2367.
- 358 Adams, D. C., & Collyer, M. L. (2019b). Phylogenetic comparative methods and the evolution of
359 multivariate phenotypes. *Annual Review of Ecology, Evolution, and Systematics*, 50, 405–425.
- 360 Adams, D. C., Collyer, M. L., & Kaliontzopoulou, A. (2020). Geomorph: Software for geometric
361 morphometric analyses. R package version 3.3.1. Retrieved from <https://cran.r-project.org/package=geomorph>
- 363 Adams, D. C., & Otárola-Castillo, E. (2013). Geomorph: An r package for the collection and analysis of
364 geometric morphometric shape data. *Methods in Ecology and Evolution*, 4, 393–399.
- 365 Arnqvist, G., & Wooster, D. (1995). Meta-analysis: Synthesizing research findings in ecology and
366 evolution. *Trends in Ecology and Evolution*, 10, 236–240.
- 367 Baken, E. K., & Adams, D. C. (2019). Macroevolution of arboreality in salamanders. *Ecology and
368 Evolution*, 9(12), 7005–7016.
- 369 Baken, E. K., Mellenthin, L. E., & Adams, D. C. (2020). Macroevolution of desiccation-related morphology
370 in plethodontid salamanders as inferred from a novel surface area to volume ratio estimation approach.

- 371 *Evolution*, 74, 476–486.
- 372 Beaulieu, J. M., Jhwueng, D. C., Boettiger, C., & O'Meara, B. C. (2012). Modeling stabilizing selection:
373 Expanding the ornstein-uhlenbeck model of adaptive evolution. *Evolution*, 66, 2369–2383. Journal
374 Article.
- 375 Blomberg, S. P., Garland, T., & Ives, A. R. (2003). Testing for phylogenetic signal in comparative data:
376 Behavioral traits are more labile. *Evolution*, 57, 717–745.
- 377 Boettiger, C., Coop, G., & Ralph, P. (2012). Is your phylogeny informative? Measuring the power of
378 comparative methods. *Evolution*, 67, 2240–2251. Journal Article.
- 379 Bonett, R. M., & Blair, A. L. (2017). Evidence for complex life cycle constraints on salamander body
380 form diversification. *Proceedings of the National Academy of Sciences, U.S.A.*, 114, 9936–9941.
381 doi:10.1073/pnas.1703877114
- 382 Bose, R., Ramesh, B. R., Péllissier, R., & Munoz, F. (2019). Phylogenetic diversity in the western ghats
383 biodiversity hotspot reflects environmental filtering and past niche diversification of trees. *Journal of
384 Biogeography*, 46(1), 145–157.
- 385 Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Routledge.
- 386 Collyer, M. L., & Adams, D. C. (2018). RRPP: An r package for fitting linear models to high-dimensional
387 data using residual randomization. *Methods in Ecology and Evolution*, 9, 1772–1779.
- 388 Collyer, M. L., Sekora, D. J., & Adams, D. C. (2015). A method for analysis of phenotypic change for
389 phenotypes described by high-dimensional data. *Heredity*, 115, 357–365.
- 390 Cooper, N., Jetz, W., & Freckleton, R. P. (2010). Phylogenetic comparative approaches for studying
391 niche conservatism. *Journal of Evolutionary Biology*, 23(12), 2529–2539.
- 392 Davies, T. J., Kraft, N. J., Salamin, N., & Wolkovich, E. M. (2012). Incompletely resolved phylogenetic
393 trees inflate estimates of phylogenetic conservatism. *Ecology*, 93(2), 242–247.
- 394 De Meester, G., Huyghe, K., & Van Damme, R. (2019). Brain size, ecology and sociality: A reptilian
395 perspective. *Biological Journal of the Linnean Society*, 126(3), 381–391.
- 396 Diniz-Filho, J. A. F., Santos, T., Rangel, T. F., & Bini, L. M. (2012). A comparison of metrics for
397 estimating phylogenetic signal under alternative evolutionary models. *Genetics and Molecular Biology*,
398 35(3), 673–679.

- 399 Felsenstein, J. (1985). Phylogenies and the comparative method. *American Naturalist*, 125(1), 1–15.
- 400 Journal Article.
- 401 Forbes, C., Evans, M., Hastings, N., & Peacock, B. (2011). *Statistical distributions*. John Wiley & Sons.
- 402 Freckleton, R. P., Harvey, P. H., & Pagel, M. (2002). Phylogenetic analysis and comparative data: A test
403 and review of evidence. *American Naturalist*, 160, 712–726.
- 404 Garland, T. Jr., & Ives, A. R. (2000). Using the past to predict the present: Confidence intervals for
405 regression equations in phylogenetic comparative methods. *American Naturalist*, 155, 346–364.
- 406 Gittleman, J. L., & Kot, M. (1990). Adaptation: Statistics and a null model for estimating phylogenetic
407 effects. *Systematic Zoology*, 39(3), 227–241.
- 408 Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher*, 5, 3–8.
- 409 Grafen, A. (1989). The phylogenetic regression. *Philosophical Transactions of the Royal Society of
410 London B, Biological Sciences*, 326, 119–157.
- 411 Harvey, P. H., & Pagel, M. D. (1991). *The comparative method in evolutionary biology*. book, Oxford:
412 Oxford University Press.
- 413 Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Elsevier.
- 414 Klingenberg, C. P., & Gidaszewski, N. A. (2010). Testing and quantifying phylogenetic signals and
415 homoplasy in morphometric data. *Systematic Biology*, 59(3), 245–261.
- 416 Martins, E. P., & Hansen, T. F. (1997). Phylogenies and the comparative method: A general approach
417 to incorporating phylogenetic information into the analysis of interspecific data. *American Naturalist*,
418 149, 646–667.
- 419 Molina-Venegas, R., & Rodríguez, M. A. (2017). Revisiting phylogenetic signal; strong or negligible
420 impacts of polytomies and branch length information? *BMC Evolutionary Biology*, 17(1), 53.
- 421 Münkemüller, T., Lavergne, S., Bzeznik, B., Dray, S., Jombart, T., Schiffers, K., & Thuiller, W. (2012).
422 How to measure and test phylogenetic signal. *Methods in Ecology and Evolution*, 3, 743–756. Journal
423 Article.
- 424 O'Meara, B. C., Ane, C., Sanderson, M. J., & Wainwright, P. C. (2006). Testing for different rates of
425 continuous trait evolution using likelihood. *Evolution*, 60, 922–933. Journal Article.

- 426 Orme, D., Freckleton, R. P., Thomas, G. H., Petzoldt, T., Fritz, S. A., & Isaac, N. (2013). CAPER:
427 Comparative analyses of phylogenetics and evolution in r. *Methods in Ecology and Evolution*, 3,
428 145–151.
- 429 Pagel, M. D. (1999). Inferring the historical patterns of biological evolution. *Nature*, 401, 877–884.
430 Journal Article.
- 431 Pavoine, S., & Ricotta, C. (2012). Testing for phylogenetic signal in biological traits: The ubiquity of
432 cross-product statistics. *Evolution: International Journal of Organic Evolution*, 67(3), 828–840.
- 433 Pintanel, P., Tejedo, M., Ron, S. R., Llorente, G. A., & Merino-Viteri, A. (2019). Elevational and
434 microclimatic drivers of thermal tolerance in andean pristimantis frogs. *Journal of Biogeography*,
435 46(8), 1664–1675.
- 436 Revell, L. J. (2010). Phylogenetic signal and linear regression on species data. *Methods in Ecology and
437 Evolution*, 1, 319–329. Journal Article.
- 438 Revell, L. J., Harmon, L. J., & Collar, D. C. (2008). Phylogenetic signal, evolutionary process, and rate.
439 *Systematic Biology*, 57, 591–601. Journal Article.
- 440 Rohlf, F. J. (2001). Comparative methods for the analysis of continuous variables: Geometric interpreta-
441 tions. *Evolution*, 55, 2143–2160. Journal Article.
- 442 Rosenthal, R. (1994). The handbook of research synthesis. In L. V. Cooper H Hedges (Ed.) (pp.
443 231–244). Russell Sage Foundation.
- 444 Sokal, R. R., & Rohlf, F. J. (2012). *Biometry* (4th ed.). San Francisco: W.H. Freeman & Co.
- 445 Su, G., Villéger, S., & Brosse, S. (2019). Morphological diversity of freshwater fishes differs between
446 realms, but morphologically extreme species are widespread. *Global Ecology and Biogeography*, 28(2),
447 211–221.
- 448 Vandelooy, F., Janssens, S., Gijbels, P., Fischer, E., Van den Ende, W., Honnay, O., & Abrahamczyk,
449 S. (2019). Nectar traits differ between pollination syndromes in balsaminaceae. *Annals of Botany*,
450 124(2), 269–279.

451 **5 Figures**

452 Figure 1. Response of Pagel's λ to increasing strength of Brownian motion. Gray line signifies the 1:1
453 line where the input value matches the estimate $\hat{\lambda}$. At each input level, the dark black line represents the
454 empirically derived expected value (mean) of $\hat{\lambda}$, the red line is the standard deviation of $\hat{\lambda}$, and the blue
455 line is Shapiro Wilks statistic of $\hat{\lambda}$ ($W = 1.0$ signifies normality, $W < 1.0$ represent skewed distributions).

456

457

458 Figure 2. Response of Blomberg's K to increasing strength of Brownian motion. At each input level, the
459 black line represents the empirically derived expected value (mean) of K , the red line is the standard
460 deviation of K , and the blue line is Shapiro Wilks statistic of K ($W = 1.0$ signifies normality, $W < 1.0$
461 represent skewed distributions).

462

463 Figure 3. Response of effect sizes Z_λ and Z_K to increasing strength of Brownian motion. Means from
464 simulation runs are shown for comparative ease. Individual values from each simulation run are available
465 in Supporting Information.

466

467 Figure 4. Type I error rates and model misspecification for \hat{Z}_{12} . Type I error is found on the far-left of the
468 plot ($\lambda = 0$). The black line signifies the average type I error and false discovery rates across simulations
469 of different tree sizes.

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

474

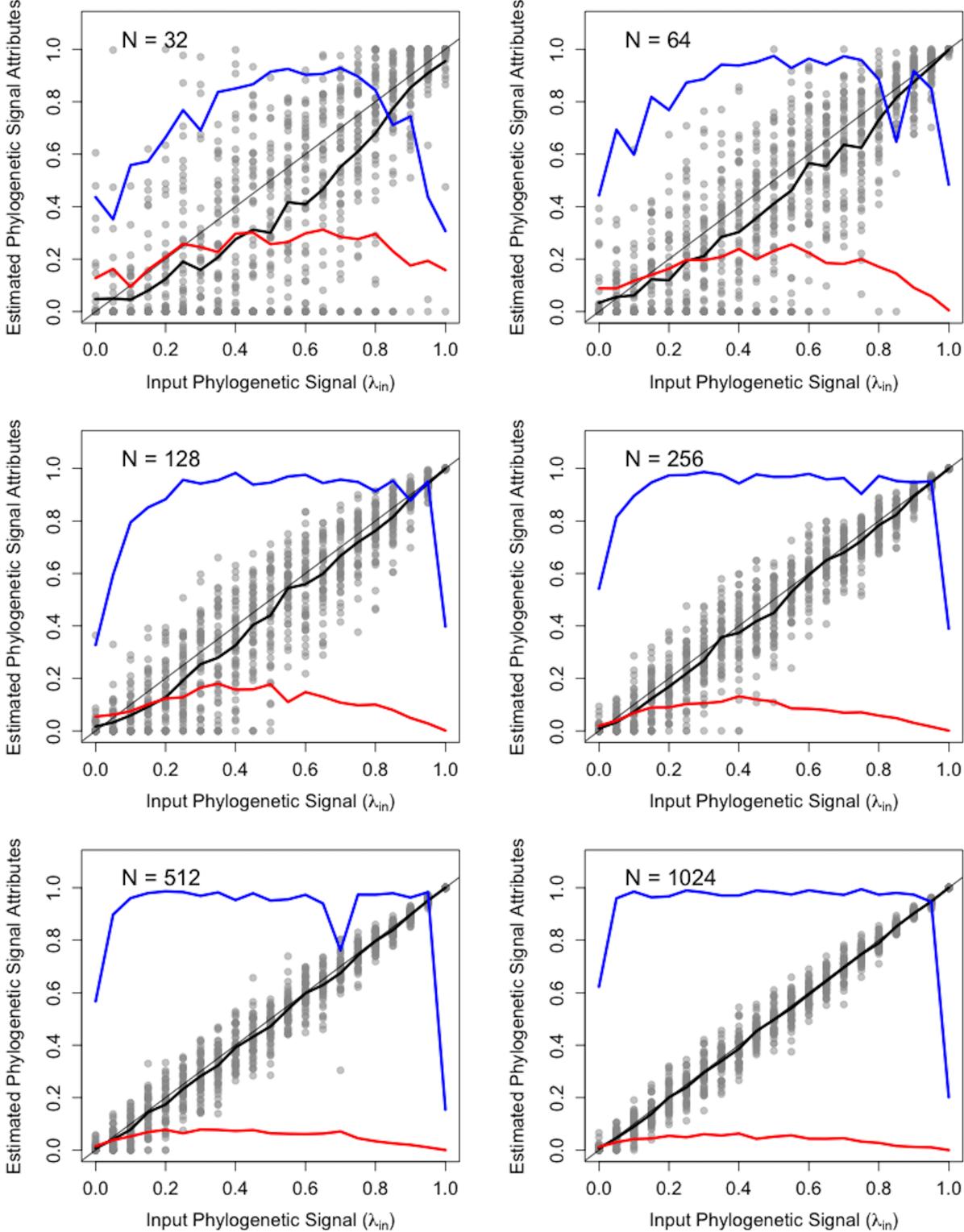


Figure 1: Response of Pagel's λ to increasing strength of Brownian motion. Gray line signifies the 1:1 line where the input value matches the estimate $\hat{\lambda}$. At each input level, the dark black line represents the empirically derived expected value (mean) of $\hat{\lambda}$, the red line is the standard deviation of $\hat{\lambda}$, and the blue line is Shapiro Wilks statistic of $\hat{\lambda}$ ($W = 1.0$ signifies normality, $W < 1.0$ represent skewed distributions).

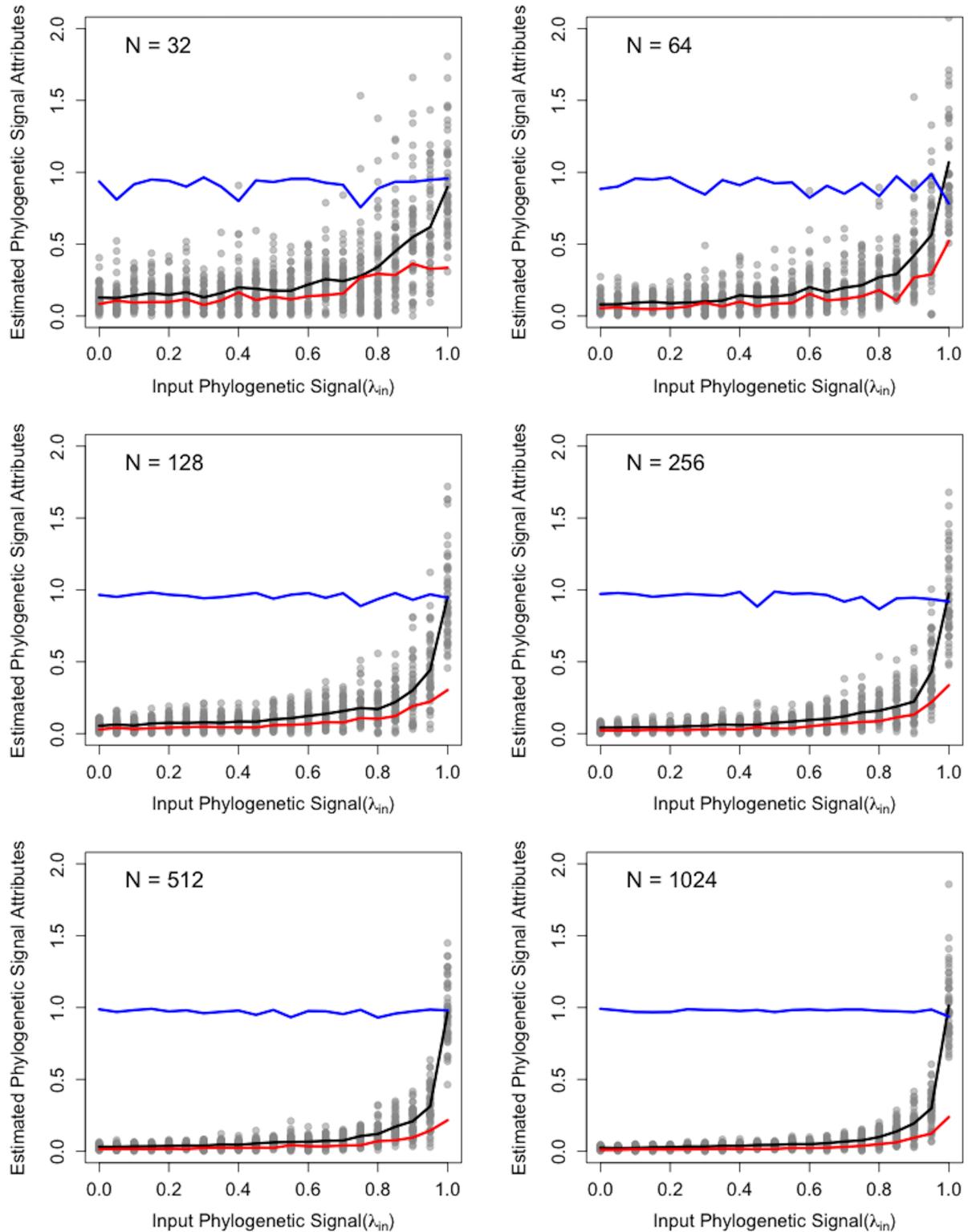


Figure 2: Response of Blomberg's K to increasing strength of Brownian motion. At each input level, the black line represents the empirically derived expected value (mean) of K , the red line is the standard deviation of K , and the blue line is Shapiro Wilks statistic of $*K*$ ($W = 1.0$ signifies normality, $W < 1.0$ represent skewed distributions).

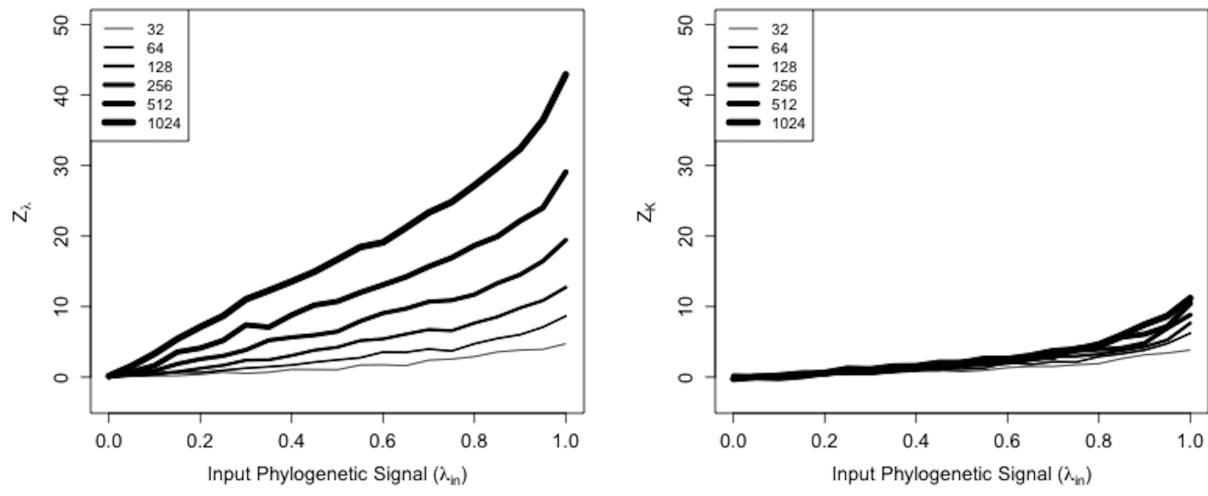


Figure 3: Response of effect sizes Z_λ and Z_K to increasing strength of Brownian motion. Means from simulation runs are shown for comparative ease. Individual values from each simulation run are available in Supporting Information.

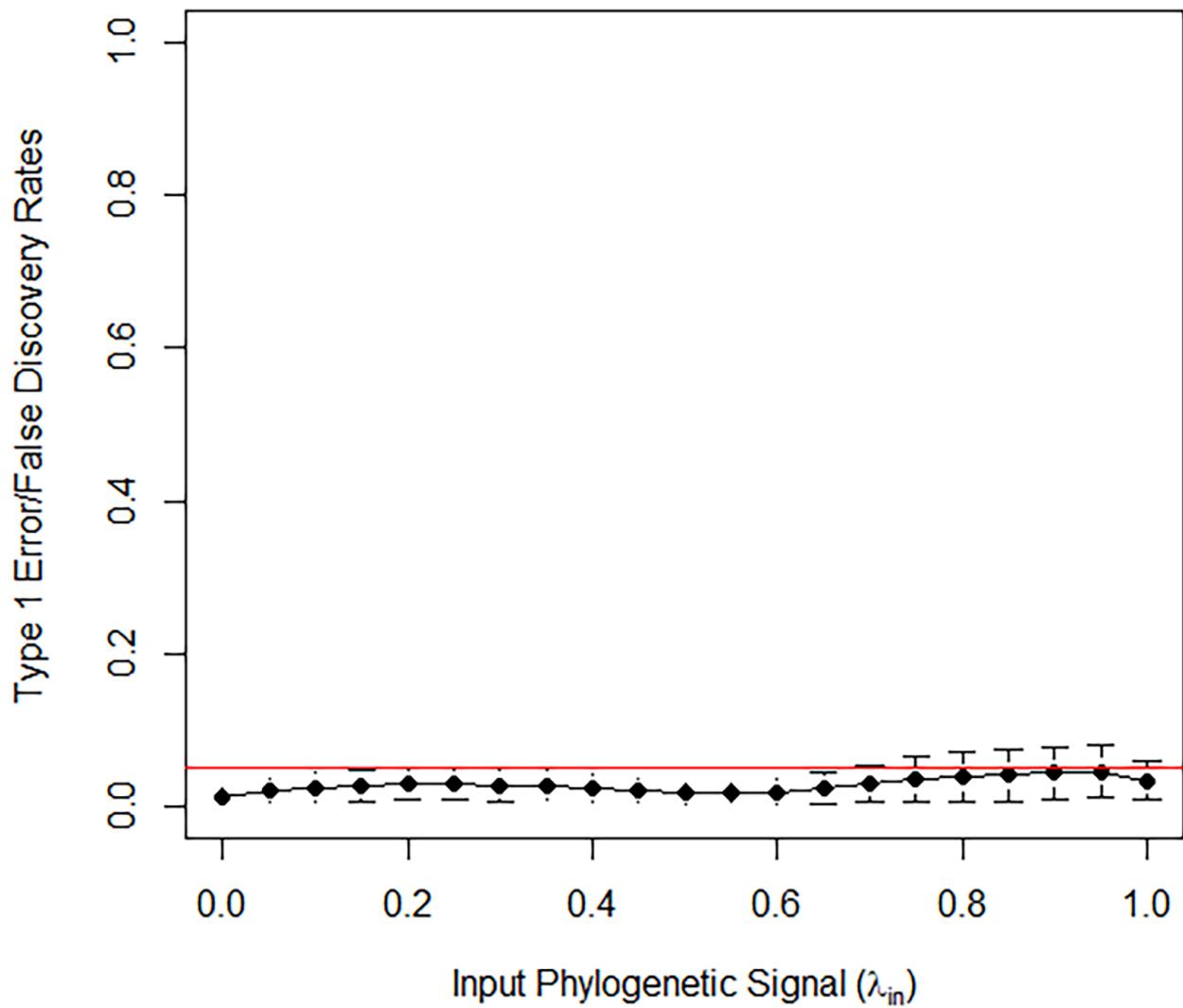


Figure 4: Type I error rates and model misspecification for \hat{Z}_{12} . Type I error is found on the far-left of the plot ($\lambda = 0$). The black line signifies the average type I error and false discovery rates across simulations of different tree sizes.

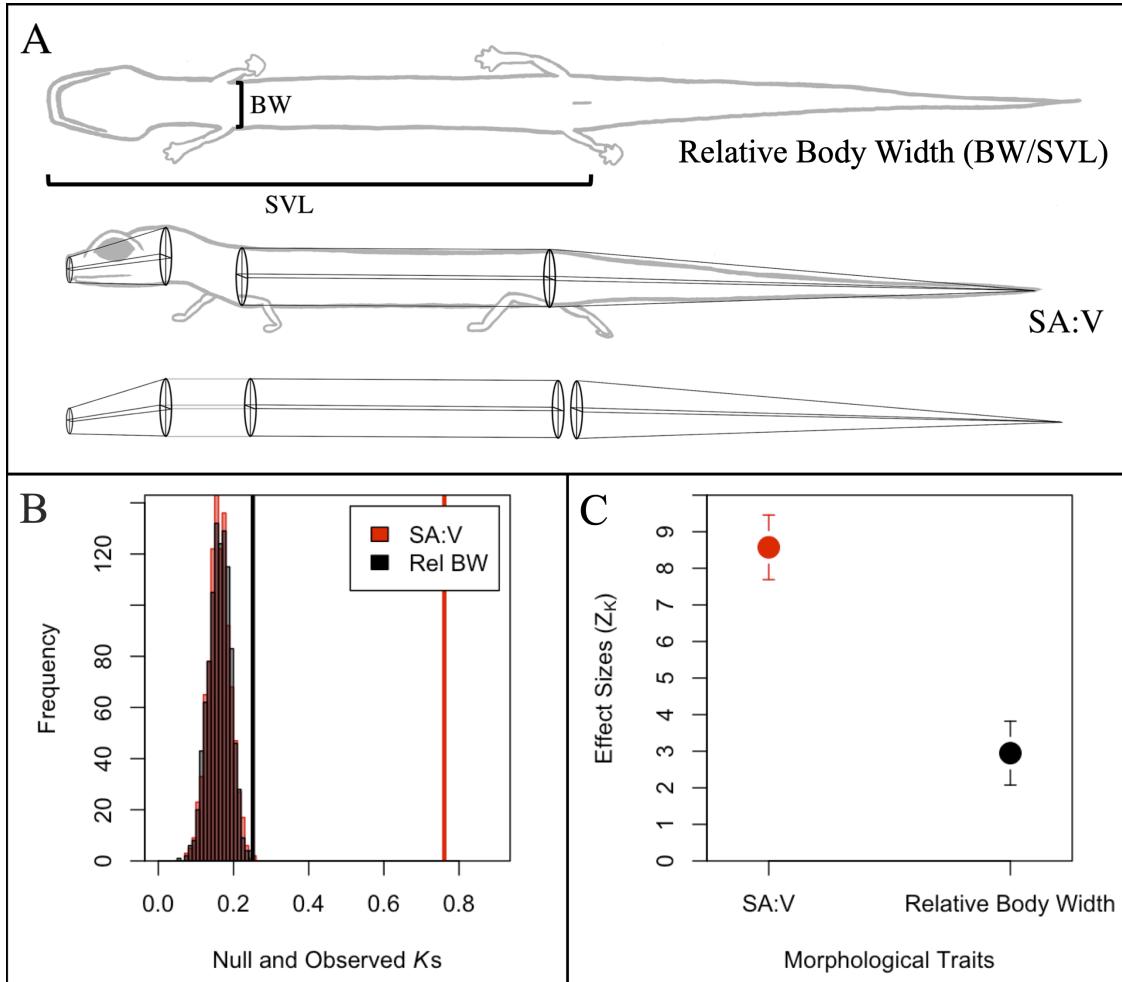


Figure 5: (A) Linear measures for relative body size, and regions of the body used to estimate surface area to volume (SA:V) ratios. (B) Permutation distributions of phylogenetic signal for SA:V and $\frac{BW}{SVL}$, with observed values shown as vertical bars. (C) Effect sizes (Z_K) for SA:V and $\frac{BW}{SVL}$,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