Received: 1 June 2021 | Accepted: 12 October 2021

DOI: 10.1111/2041-210X.13749

Methods in Ecology and Evolution
BRITISH ECOLOGICAL SOCIETY

RESEARCH ARTICLE

A standardized effect size for evaluating and comparing the strength of phylogenetic signal

Michael L. Collyer[1] | Erica K. Baken[1,2] | Dean C. Adams[2]

[1]Department of Science, Chatham University, Pittsburgh, PA, USA

[2]Department of Ecology, Evolution, and Organismal Biology, Iowa State University, Ames, IA, USA

Correspondence
Michael L. Collyer
Email: m.collyer@chatham.edu

Handling Editor: José Miguel Ponciano

Abstract

1. Macroevolutionary studies frequently characterize the phylogenetic signal in phenotypes; however, analytical tools for comparing the strength of that signal across traits remain largely underdeveloped.
2. We developed a non-parametric, permutation test for the log-likelihood of an evolutionary model, plus a standardized statistic, Z, from this test, which can be considered a phylogenetic signal effect size. This statistic can be used in two-sample tests to compare the strength of phylogenetic signal for multiple traits.
3. We performed simulation experiments that revealed that Z had a linear association with Pagel's λ, which could be predicted by tree size, plus could be quickly interpreted as a hypothesis for phylogenetic signal based on a standard normal distribution. We additionally found that the permutation test had greater statistical power for detecting phylogenetic signal than parametric likelihood ratio tests, especially for small trees.
4. The analytical framework we present extends the phylogenetic comparative methods toolkit, allowing for statistical comparison of phylogenetic signal in multiple traits. Future studies could also consider this framework for the comparison of different evolutionary models, especially in light of different null processes.

KEYWORDS

comparative analysis, macroevolution, RRPP

1 | INTRODUCTION

The shared evolutionary history of closely related species often implies the trait similarity among them, a pattern referred to as phylogenetic signal (Blomberg et al., 2003; Felsenstein, 1985; Pagel, 1999). Many phylogenetic comparative methods (PCMs) seek to address the non-independence of species' traits (Felsenstein, 1985; Harvey & Pagel, 1991) in their analytic framing, by conditioning data on the phylogenetic relatedness among observations (Adams, 2014b; Adams & Collyer, 2018; Beaulieu et al., 2012; Garland & Ives, 2000; Grafen, 1989; Martins & Hansen, 1997; O'Meara et al., 2006; Rohlf, 2001). Indeed, under numerous evolutionary models, phylogenetic signal is expected, as stochastic character change along the hierarchical structure of the tree of life generates trait covariation among taxa (Blomberg et al., 2003; Felsenstein, 1985; Revell et al., 2008). Quantifying and comparing phylogenetic signal among traits, however, remains quite challenging.

Several analytical tools have been developed to quantify phylogenetic signal in phenotypic datasets (Abouheif, 1999; Adams, 2014a; Blomberg et al., 2003; Gittleman & Kot, 1990; Klingenberg & Gidaszewski, 2010; Pagel, 1999), and their statistical properties namely type I error rates and statistical power have been investigated to determine under what conditions phylogenetic signal can be detected (Adams, 2014a; Boettiger et al., 2012; Diniz-Filho et al., 2012; Molina-Venegas & Rodríguez, 2017; Münkemüller et al., 2012; Pavoine & Ricotta, 2013; Revell, 2010; Revell et al., 2008). One of the most widely used methods for characterizing phylogenetic signal is Pagel's λ (Pagel, 1999), which transforms the lengths (by compression) of the internal branches of the phylogeny, while leaving the tips unaffected, to improve the fit of data to the phylogeny via maximum likelihood (Freckleton et al., 2002; Pagel, 1999). To infer whether phylogenetic signal differs from no signal or a Brownian motion (BM) model of evolutionary divergence, the observed model fit using $\widehat{\lambda}$ may be statistically compared with that using $\lambda = 0$ or $\lambda = 1$ via likelihood ratio tests (Bose et al., 2019; Cooper et al., 2010; Freckleton et al., 2002) or confidence limits (Vandelook et al., 2019).

Another widely used measure is Blomberg's K (Blomberg et al., 2003), which characterizes phylogenetic signal as the ratio of observed trait variation to the amount of variation expected under BM. Blomberg's K can be treated as a test statistic by using a permutation test to generate its sampling distribution (Adams, 2014a; Blomberg et al., 2003) for determining whether significant phylogenetic signal is present in data. Both λ and K seem intuitive to interpret, as a value of 0 for both corresponds to no phylogenetic signal, and a value of 1 corresponds to the amount of phylogenetic signal expected under BM. Thus, it is tempting to regard both λ and K as descriptive statistics (and effect sizes, Münkemüller et al., 2012) that measure the relative strength of phylogenetic signal, providing an estimate of its magnitude for comparison.

The potential appeal of Pagel's λ and Blomberg's K as effect sizes is that they provide a basis for interpreting weak versus strong phylogenetic signal; that is, small versus large values of $\widehat{\lambda}$ or K, respectively, in a comparative sense (De Meester et al., 2019; Pintanel et al., 2019; Su et al., 2019). They are also important statistics in hypothesis tests. The optimized value of lambda, $\widehat{\lambda}$, is the location where the log-likelihood is maximized, and is, therefore, compelling for finding the maximum phylogenetic signal in the data, which can be deemed significant by rejecting the null hypothesis of $\lambda = 0$ in a likelihood ratio test. Although Pagel's λ has an upper bound of 1, Blomberg's K can measure phylogenetic signal that is greater than expected under BM, as it has no upper bound. Blomberg's K or more specifically, the GLS estimation of variance that is a part of its calculation can serve as a test statistic in a permutation test that randomizes tip data in random permutations. However, K is quite sensitive to tree size, exhibits high type II error rates for intermediate strength of phylogenetic signal, has higher type I error rates than likelihood ratio tests based on $\widehat{\lambda}$, and exhibits greater uncertainty for strong phylogenetic signals (whereas $\widehat{\lambda}$ has greater uncertainty at intermediate phylogenetic signal strength: Münkemüller et al., 2012). Both of these statistics offer good support as test statistics for determining whether phylogenetic signal exists in a trait, but they are limited for comparing phylogenetic signals between traits.

Here, we present an alternative, standardized effect size calculation, which can be used in hypothesis tests to compare phylogenetic signals for different traits, and which is based on a normalized distribution of random log-likelihoods of a phylogenetic model, generated from a model of phylogenetic independence. Much like a likelihood ratio test for Pagel's λ, this non-parametric approach can assess the significance of the observed phylogenetic signal, but unlike the parametric test, the standardized location of the observed likelihood can be used as an effect size, which can be statistically compared with similarly calculated effect sizes to consider hypotheses regarding the relative strengths of phylogenetic signal for multiple traits. We use simulation experiments to compare this standardized effect size to $\widehat{\lambda}$ and K and demonstrate its utility with an empirical example. Comprehensively we illustrate that this standardized effect size provides an additional necessary tool to the phylogenetic comparative toolkit.

2 | CONCEPTUAL DEVELOPMENT

A hypothesis test for phylogenetic signal involves calculating the variance among taxa trait values, conditioned on phylogenetic covariances (evolutionary rates) and comparing this variance to a variance that assumes phylogenetic independence. This can be appreciated by the GLS log-likelihood equation for a BM phylogenetic model of a univariate trait (Blomberg et al., 2003; Freckleton, 2012; Freckleton et al., 2002; Garland & Ives, 2000):

$$\log\mathscr{L}(\sigma|\mathbf{V}) = -\frac{1}{2}\log|2\pi\mathbf{V}| - \frac{1}{2}\left[(\mathbf{y}-E(\mathbf{y}))\mathbf{V}^{-1}(\mathbf{y}-E(\mathbf{y}))\right], \tag{1}$$

where, $\mathbf{y}$ is a vector of N trait values, $\mathbf{y} - E(\mathbf{y})$ is a vector of phylogenetic residuals, $\mathbf{V}$ is an $N \times N$ phylogenetic covariance matrix, equal to $\sigma^2\mathbf{C}$, and $|\mathbf{V}|$ represents its determinant. The $N \times N$ covariance matrix, $\mathbf{C}$, is a matrix of phylogenetic variances along the diagonal, and covariances that are proportional to or exactly the covariances from a BM model of evolution (Revell et al., 2008). The variance (evolutionary rate), σ^2, is calculated as $\sigma^2 = N^{-1}(\mathbf{y}-E(\mathbf{y}))^T\mathbf{C}^{-1}(\mathbf{y}-E(\mathbf{y}))$, where ^{T} represents vector transposition. N^{-1} is used for the maximum likelihood estimate of σ^2; $(N-1)^{-1}$ is used in place of N for the restricted maximum likelihood (REML) estimator (Freckleton, 2012). The expected value (tree root) is computed as $E(\mathbf{y}) = (1^T\mathbf{C}^{-1}1)^{-1}1^T\mathbf{C}^{-1}\mathbf{y}$, where 1 is an $N \times 1$ vector of 1s. Because $\mathbf{V} = \sigma^2\mathbf{C}$, Equation 1 can be expanded, that is,

$$\log\mathscr{L}(\sigma|\mathbf{V}) = -\frac{N}{2}\log(2\pi) - \frac{N}{2}\log\sigma^2 - \frac{1}{2}\log|\mathbf{C}| - \frac{1}{2}\left[(\mathbf{y}-E(\mathbf{y}))\mathbf{V}^{-1}(\mathbf{y}-E(\mathbf{y}))\right]. \tag{2}$$

This expansion helps to elucidate the portions of the log-likelihood equation that are constant when comparing traits. In Equation 2, $\frac{N}{2}\log(2\pi)$ is a constant and $(\mathbf{y}-E(\mathbf{y}))\mathbf{V}^{-1}(\mathbf{y}-E(\mathbf{y})) = N$ for any traits and any trees, for a BM model of evolution. We can, therefore, update Equation 2:

$$\begin{aligned}\log\mathscr{L}(\sigma|\mathbf{V}) &= -\frac{N}{2}\log(2\pi) - \frac{N}{2}\log\sigma^2 - \frac{1}{2}\log|\mathbf{C}| - \frac{1}{2}\left[(\mathbf{y}-E(\mathbf{y}))\mathbf{V}^{-1}(\mathbf{y}-E(\mathbf{y}))\right] \\ &= -\frac{N}{2}\left[\log(2\pi) + \log\sigma^2 + \frac{1}{N}\log|\mathbf{C}| + \frac{1}{N}\left[(\mathbf{y}-E(\mathbf{y}))\mathbf{V}^{-1}(\mathbf{y}-E(\mathbf{y}))\right]\right] \\ &= -\frac{N}{2}\left[\log(2\pi) + \log\sigma^2 + \frac{1}{N}\log|\mathbf{C}| + \frac{1}{N}N\right] \\ &= -\frac{N}{2}\left[\log\sigma^2 + \frac{1}{N}\log|\mathbf{C}| + \log(2\pi) + 1\right].\end{aligned}$$

Thus, Equation 2 can be simplified:

$$\log \mathscr{L}(\sigma|\mathbf{V}) = -\frac{N}{2}\left[\log\sigma^2 + \frac{1}{N}\log|\mathbf{C}| + C\right], \tag{3}$$

where C is a constant for all of the parts in Equation 1 that would not be changed by changing $\mathbf{C}$. Equation 3 helps one appreciate that in a likelihood ratio test to compare estimates of evolutionary rates (e.g. $\sigma^2_{\lambda=0}$ and $\sigma^2_{\hat{\lambda}}$), N and C are the same in the two likelihood calculations; the only parts that change are $\mathbf{C}$ (as a proportional change in covariances, based on $\hat{\lambda}$) and the values of σ^2, as a result of $\mathbf{C}$. Therefore, a likelihood ratio test is a direct comparison of evolutionary rates.

Pagel's λ is a scaling parameter by which the covariances (i.e. the off-diagonals) of $\mathbf{C}$ are multiplied (Pagel, 1999). A value of 0 changes all covariances to 0 (a star phylogeny or phylogenetic independence), and a value of 1 does not change the covariances from those expected by a BM model of evolution. (If $\lambda = 0$ and the tree is ultrametric, $\mathbf{C}$ is a diagonal matrix proportional to an $N \times N$ identity matrix, $\mathbf{I}$. It is convenient in this case to refer to $\mathbf{I}$ rather than $\mathbf{C}$ as a model of evolutionary independence because the lengths of branches in a star phylogeny all equal are inconsequential in estimation of the trait variance.) The value of λ that minimizes $\log\sigma^2 + \frac{1}{N}\log|\mathbf{C}|$ maximizes the log-likelihood. This value can be found for the interval between 0 and 1, yielding the optimized value, $\hat{\lambda}$. In a likelihood ratio test, the log-likelihood at $\hat{\lambda}$ can be compared with the log-likelihood found at $\lambda = 0$; rejection of the null hypothesis indicates significant phylogenetic signal. Likelihood ratio tests could also be used to explicitly test $\hat{\lambda}$ against a model of pure BM ($\lambda = 1$).

By contrast, Blomberg's K finds $(\mathbf{y} - E(\mathbf{y}))$ and calculates variance (mean-squared error) from these residuals two different ways: $MSE_0 = (N-1)^{-1}(\mathbf{y}-E(\mathbf{y}))^T(\mathbf{y} - E(\mathbf{y}))$ and $MSE = (N-1)^{-1}(\mathbf{y}-E(\mathbf{y}))^T\mathbf{C}^{-1}(\mathbf{y} - E(\mathbf{y}))$, where $\mathbf{C}$ is typically the untransformed covariance matrix based on a BM model of evolutionary divergence ($\lambda = 1$). The only difference between MSE and σ^2 in Equation 2 is the use of the REML estimator ($N - 1$ degrees of freedom) for MSE. MSE_0, however, ignores phylogenetic covariances in its estimation (does not correct for phylogenetic relatedness). Blomberg's K is the ratio, $\frac{MSE_0}{MSE}$, divided by its expectation under BM for a given phylogeny; that is, $K = \text{observed}\left(\frac{MSE_0}{MSE}\right)/\text{expected}\left(\frac{MSE_0}{MSE}\right)$ (Blomberg et al., 2003). This equation could be equivalently calculated (Revell et al., 2008), as

$$K = \frac{(\mathbf{y}-E(\mathbf{y}))^T(\mathbf{y} - E(\mathbf{y}))}{(\mathbf{y}-E(\mathbf{y}))^T\mathbf{C}^{-1}(\mathbf{y} - E(\mathbf{y}))} \Big/ \frac{trace(\mathbf{C}_{BM}) - N(1^T\mathbf{C}_{BM}^{-1}1)^{-1}}{N-1}, \tag{4}$$

where *trace* is the sum of diagonal elements, and we use the subscript, BM, to indicate this is an untransformed ($\lambda = 1$) version of $\mathbf{C}$. Typically, $\mathbf{C}_{BM}$ is also used in the calculation of $(\mathbf{y}-E(\mathbf{y}))^T\mathbf{C}^{-1}(\mathbf{y} - E(\mathbf{y}))$, but this need not be the case, as least for considering K as a test statistic rather than a descriptive statistic. K will tend toward 0 if there is no phylogenetic signal, and tend toward or exceed 1 if there is. Whereas a likelihood ratio test can be used for Pagel's λ, a permutation test (which randomizes the trait data across the tips of the phylogeny) is used to generate random distributions of MSE (e.g. Blomberg et al., 2003) or K (e.g. Adams, 2014a). A p-value is found as the percentile of the observed statistic in its sampling distribution. Because a permutation test and likelihood ratio test are non-parametric and parametric solutions for different test statistics, respectively, it might not be surprising that they could produce different results with respect to the same null hypothesis of no phylogenetic signal. However, it is because of the potential difference in λ values used in calculation of the test statistics more so than the statistic or method used that different results are possible. With the same λ used to calculate $\mathbf{C}$, and thus, MSE, the two tests should produce similar results. This can be appreciated by considering the process that generates variation in the permutation test.

Blomberg et al. (2003) proposed that $\mathbf{y} - E(\mathbf{y})$ could be replaced in Equation 3 by $\mathbf{y} - \mu$, where μ is the ordinary least squares (OLS) mean of $\mathbf{y}$. A permutation test that randomizes tip data performed with this altered K statistic or MSE produces a distribution of values that are perfectly rank correlated because the only random element recalculated in each permutation is $(\mathbf{y}-E(\mathbf{y}))^T\mathbf{C}^{-1}(\mathbf{y} - E(\mathbf{y}))$. (All other portions of the K calculation would be constant.) It can be appreciated why Blomberg et al. (2003) suggested MSE as a statistic, and asserted that using $\mathbf{C}$ that is transformed (e.g. by $\hat{\lambda}$) would mean having greater statistical power to detect phylogenetic signal. It can be seen from Equations 3 and 4 that for the same $\mathbf{C}$,

$$\frac{(\mathbf{y}-E(\mathbf{y}))^T\mathbf{C}^{-1}(\mathbf{y} - E(\mathbf{y}))}{N-1} \propto \left[\log\hat{\sigma}^2 + \frac{1}{N}\log|\mathbf{C}|\right]. \tag{5}$$

If one optimizes λ via maximum likelihood, uses this value to transform $\mathbf{C}$, and performs a permutation test on K, using MSE as the test statistic, then a test on $\hat{\lambda}$ and a test on K are commensurate. Furthermore, such a permutation test can be considered a non-parametric alternative to a likelihood ratio test. (We provide additional empirical detail in Appendix 1 of the Supporting Information that confirms rank correlation.)

Randomizing tip data is a simplified form of randomization of residuals in a permutation procedure (RRPP). RRPP works best if residuals are the most appropriate exchangeable units under the null hypothesis (Adams & Collyer, 2018; Commenges, 2003). RRPP is a process that randomizes null model residuals and adds them to null model fitted values in every random permutation to create random pseudodata used to fit alternative models. If the null hypothesis is phylogenetic independence, a star phylogeny is assumed, $\mathbf{C} \propto \mathbf{I}$, $E(\mathbf{y}) = \mu$, the OLS mean, $\mathbf{y} - E(\mathbf{y})$ are the OLS residuals, and random outcomes of $\mathbf{y}^* = E(\mathbf{y}) + (\mathbf{y}-E(\mathbf{y}))^*$, where $*$ indicates randomization, are the pseudodata produced in each permutation. If $E(\mathbf{y}) = \mu$, then randomizing residuals is the same as randomizing tip data (the root and data mean are the same). This process preserves first- and second-moment exchangeability; that is, the OLS mean and variance of the trait are constant across random permutations. (If phylogenetic independence is not assumed, RRPP still functions the same, but has a GLS solution with second-moment exchangeability only.) It is important to understand that when the portions of the log-likelihood expression for phylogenetic independence (summarized as C in Equation 3) are held constant, p-values found from the sampling distributions of either component of Equation 5, the log-likelihood, or the likelihood ratio statistic, $-2\left(\log\mathscr{L}_{\mathbf{I}} - \log\mathscr{L}^{*}_{\mathbf{C}}\right)$, will be exactly the same (as $\log\mathscr{L}_{\mathbf{I}}$ is also constant in every random permutation). Therefore, a permutation test with RRPP is a non-parametric application of the likelihood ratio statistic in a hypothesis test (Potter, 2005), which does not rely on a mixture of parametric probability distributions as a proxy of the true sampling distribution (Molenberghs & Verbeke, 2007; Self & Liang, 1987). However, it remains to be seen if a permutation test, using the log-likelihood statistic and RRPP, is as reliable as a parametric likelihood ratio test.

Assuming a comparable tree transformation, Pagel's λ and Blomberg's K can be considered two phylogenetic signal effect sizes (Münkemüller et al., 2012), but a test of phylogenetic signal, as demonstrated above, is more explicitly an assessment of the rarity of the observed $\log\sigma^2$ in a hypothetical distribution of $\log\sigma^2$, if $\lambda = 0$ is the null model process. This process can be applied with RRPP and a sampling distribution of either σ^2, $-\frac{N}{2}\log\sigma^2$, or $\log\mathscr{L}_{\mathbf{C}}$ can be generated, and all would provide the same p-values for the same RRPP permutations (see Appendix 1 of the Supporting Information for an example of this outcome). However, sampling distributions from permutation tests do not need to be a means to an end, a tool to merely find a p-value. The location of the observed statistic in its sampling distribution can also be considered an effect size (Adams & Collyer, 2016, 2018, 2019; Collyer et al., 2015). From, Equation 3, either the non-constant portion of the log-likelihood equation, $-\left(\log\sigma^2 + \frac{1}{N}\log|\mathbf{C}|\right)$, or the log-likelihood itself, are good statistics for estimating effect sizes (Equation 3 demonstrates that the log-likelihood is a linear transformation of $-\left(\log\sigma^2 + \frac{1}{N}\log|\mathbf{C}|\right)$, so effect sizes estimated from the RRPP distributions of either will be the same. We will henceforth refer to the random forms of $\log\mathscr{L}_{\mathbf{C}}$, for simplicity.) The location of the observed value in a standardized distribution of random $\log\mathscr{L}_{\mathbf{C}}$ outcomes, based on the appropriate null hypothesis of phylogenetic independence, provides a standardized (statistical) effect size that can be used in comparative tests (Adams & Collyer, 2016, 2018, 2019; Collyer et al., 2015).

Letting $\theta = f\left(\log\mathscr{L}_{\mathbf{C}}\right)$, where f represents a normalizing function (if random $\log\mathscr{L}_{\mathbf{C}}$ are not sufficiently normally distributed), the standardized effect size of phylogenetic signal for a trait is estimated as,

$$Z = \frac{\theta_{obs} - \hat{\mu}_{\theta}}{\hat{\sigma}_{\theta}}, \qquad (6)$$

where $\hat{\mu}_{\theta}$ is the mean of the sampling distribution and $\hat{\sigma}_{\theta}$ is the standard error (the standard deviation of the sampling distribution, not the trait). (The ^ indicates these values are estimated, based on the number of random permutations used, which is probably fewer than the finite but large possible number of all permutations.) As a standard deviate, we would expect correspondence between a p-value estimated from the location of Z in a standard normal distribution and the percentile of $\log\mathscr{L}_{\mathbf{C}}$ in its sampling distribution. Therefore, it is obvious that, e.g. $Z = 2.5$ means significant phylogenetic signal and $Z = 0.7$ means phylogenetic signal that is not significant, based on a significance level of $\alpha = 0.05$. More importantly, because sampling distributions are approximately normal, two effect sizes can be compared in a hypothesis test, by finding the two-sample test statistic,

$$Z_{12} = \frac{\left(\theta_{1_{\text{obs}}} - \hat{\mu}_{\theta_1}\right) - \left(\theta_{2_{\text{obs}}} - \hat{\mu}_{\theta_2}\right)}{\sqrt{\hat{\sigma}^2_{\theta_1} + \hat{\sigma}^2_{\theta_2}}}. \qquad (7)$$

The Z_{12} statistic can be assumed to follow a standard normal distribution, meaning a p-value can be obtained for a null hypothesis test that the phylogenetic signals for two separate traits are the same. There is no explicit expectation that the traits have to come from the same phylogeny, but the scope for comparison of traits is something that can only be considered by examining the behavior of these effect sizes for varied tree sizes and phylogenetic signal strength.

Equation 3 implicitly assumes that the compared traits evolve independently, which might be an illogical assumption for traits measured on species from the same phylogeny. In such cases, there are two options worth considering. First, one could generalize the log-likelihood equation (Equation 2) for multivariate data (see, Revell & Harmon, 2008) and consider the relative strength of multivariate phylogenetic signal with respect to the univariate signals. This is not necessarily a simple generalization, if one allows $\hat{\lambda}$ to vary among traits (requiring $p\left(p-1\right)/2$ covariance matrix estimations in the log-likelihood for p traits; see Appendix 2 in the Supporting Information for further details). However, one could compare multivariate Z-scores between models that assume a common $\hat{\lambda}$ or allow $\hat{\lambda}$ to vary among traits, as a test of evolutionary independence of traits; much like one can compare models with common or separate evolutionary rates among traits (see, Adams, 2013). (We provide further details for this future research consideration in Appendix 2 of the Supporting Information.) Second, one could compare the relative strength of phylogenetic signal between principal components of a multivariate data set. With this option, the principal components would be independent, but one would have to reconcile principal component loadings with test results to determine whether suites of traits have different phylogenetic signals.

Multivariate considerations are expansive and exceed the scope of this paper. However, RRPP is a process that generates sampling distributions of log-likelihoods in a consistent manner, irrespective of the number of traits. Research questions that require multivariate analysis should have tractable solutions, provided log-likelihoods can be estimated (residual covariance matrices are not singular). Regarding single traits, we perform simulation experiments to determine type I error rates, correspondence between hypothesis test outcomes, statistical power, and the relationship between effect size and simulated phylogenetic strength, below. However, we first provide a simple example to help illustrate the purpose of this type of analysis.

2.1 | Illustrative example

As an illustrative example, we simulated two traits on a phylogeny ($N = 60$), one with moderate phylogenetic signal and one with stronger phylogenetic signal. (Simulation details are explained in the next section.) For one variable, X, $\widehat{\lambda}_X = 0.36$, and for the other, Y, $\widehat{\lambda}_Y = 0.77$. We performed RRPP to recalculate the GLS log-likelihoods (using a covariance matrix for a tree transformed by $\widehat{\lambda}$ for each variable), with 10,000 random permutations, each. These distributions were normalized (with a Box-Cox transformation) and standardized (Figure 1), yielding Z scores of 2.21 and 6.33 standard deviations, respectively, each of which was significant at $\alpha = 0.05$ ($p = 0.0146$ for X and $p = 0.0001$ for Y). Performing parametric likelihood ratio tests with a null model of $\lambda = 0$ yielded slightly different results $\chi^2 = 2.07$, $p = 0.0750$, and $\chi^2 = 25.89$, $p < 0.0001$, for X and Y, respectively. The difference, as we show below, is likely due to the limited statistical power (type II error) of the parametric likelihood ratio test. We performed a two-sample z-test to compare the phylogenetic signal effect sizes; $|Z_{XY}| = 2.92$, $p = 0.0018$, indicating that the phylogenetic signal in Y was significantly larger than in X.

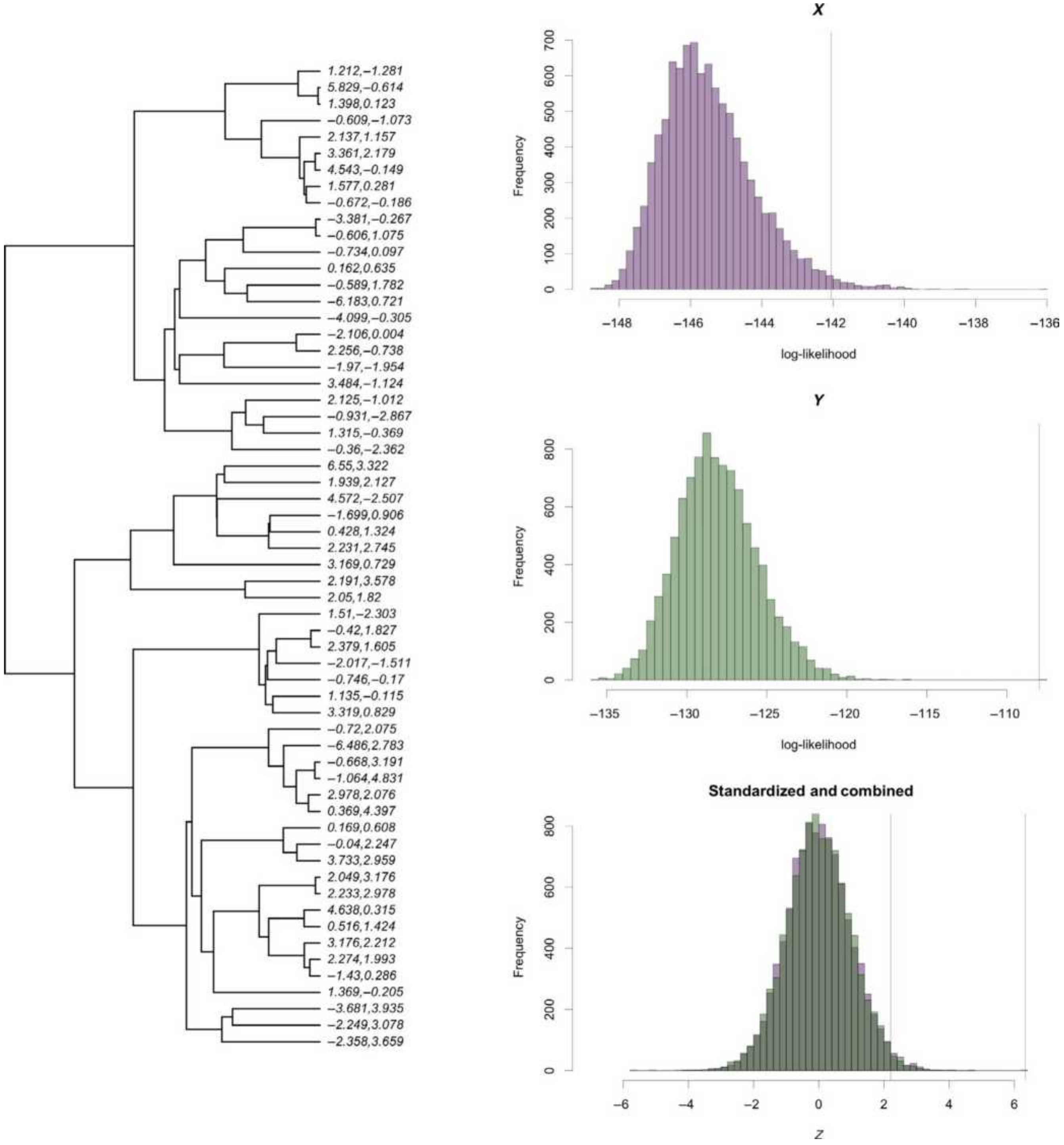

FIGURE 1 Plot of phylogenetic tree with x, y values, plus frequency histograms for the RRPP log-likelihood values for two variables, X and Y. Vertical lines indicate observed values. In the last panel, histograms have been combined for standardized values

Comparatively, $K_X = 0.96$ and $K_Y = 1.02$, which like $\widehat{\lambda}$ is not as useful for determining a significant difference between phylogenetic signals.

3 | SIMULATION METHODS AND RESULTS

We examined the behavior of RRPP-based likelihood ratio tests and standardized log-likelihood effect size with simulation experiments that varied the strength of phylogenetic signal, for various sized pure-birth phylogenetic trees. In our simulation experiments, we sought to examine the statistical power of likelihood ratio tests with RRPP compared to parametric tests, and the relationships between effect size, Pagel's λ, and Blomberg's K.

3.1 | Simulation methods

Simulating data from a model with known phylogenetic signal is challenging, as the process requires an a priori definition of phylogenetic signal, and there is no guarantee that the process will produce data that are similar to the intended effect. It is possible to simulate data with an intended λ, as this requires merely simulating a tree, rescaling the tree, and simulating BM data on the rescaled tree (see, e.g. Adams, 2014a; Molina-Venegas & Rodríguez, 2017). Alternatively, a weighted average of data simulated with BM and without BM could be used, which Münkemüller et al. (2012) described as the (simulated) BM strength. However, with either approach, there is no guarantee that $\widehat{\lambda}$ will resemble λ, especially for small trees (see, e.g. Figure 2 of Münkemüller et al., 2012). Furthermore, there is no easily conceivable way to simulate data from a model with known K. For previous studies that sought to evaluate statistical properties (type I or type II errors, accuracy, and precision), defining λ or a weight of BM strength, as simulated, was sufficient for calculating summary statistics over many simulation runs with the same input value. However, we were more interested in understanding the association of simulated phylogenetic signal strength and the effect size estimated from the log-likelihood of an evolutionary model, over a continuum from $\lambda = 0$ to 1.

Initial trials to simulate data (sensu Adams, 2014a; Molina-Venegas & Rodríguez, 2017) from a uniform distribution of λ revealed that, especially with smaller trees, distributions of $\widehat{\lambda}$ tended to be skewed toward 0 or 1, despite uniform sampling of λ. This was consistent with the research of Münkemüller et al. (2012). (see, e.g. their Figure 2 and their Table 2, which indicates skewing of $\widehat{\lambda}$ toward 0 or 1 for small trees.) Therefore, we used an algorithm to first simulate λ from a uniform distribution, and then simulate data that produced $\widehat{\lambda}$ within 5% of λ to assure that there was an approximately uniform distribution of $\widehat{\lambda}$ throughout the simulation runs.

We simulated 5,000 pure-birth, ultrametric trees (with a branching rate of 0.05) for each of $2^{5:9}$ sized trees (25,000 trees total). All trees were created with the function, `pbtree`, from the `phytools R` package (Revell, 2010). For each tree, we randomly sampled λ from a uniform distribution (minimum of 0, maximum of 0.99), scaled the tree branch lengths by λ, and simulated random BM data on the transformed tree, using the `sim.char` function of the `geiger R` package (Harmon et al., 2008). Subsequently, we found the maximum likelihood estimate, $\widehat{\lambda}$ (see code in Appendix 3 of the Supporting Information) from the data generated. We used an upper limit of $\lambda = 0.99$ because like Cooper et al. (2016), we observed a rare but discernible trend for data simulated with $\lambda = 1$ to not fit as well with a BM model of evolutionary divergence as alternative models, such as Ornstein Uhlenbeck models (Lande, 1976). By using a cut-off of 0.99, instances of $\widehat{\lambda} = 1$ were still frequent, but anomalies from simulating non-BM data were largely mitigated. For every tree we simulated, we repeatedly simulated data until we found $\widehat{\lambda}$ within a 5% interval of λ, and then retained the data for analysis.

For every simulated tree and its corresponding data, we performed a parametric likelihood ratio test, with $\lambda = 1$ (untransformed) and $\lambda = \widehat{\lambda}$ (transformed) adjustments of **C**. We also used RRPP to generate distributions of 1,000 random log-likelihoods for each tree:data combination, and for both untransformed and transformed **C** matrices, from which the percentile of the observed statistic was used to estimate a p-value. The parametric likelihood ratio test performed for $\lambda = \widehat{\lambda}$ and the permutation test performed for $\lambda = 1$ (null $\lambda = 0$ in both cases) correspond exactly with the tests typically performed for Pagel's λ and Blomberg's K, respectively. We verified p-values estimated this way were the exact same as using the distribution of random MSE, more typically used for a test of Blomberg's K.

K results and standardized log-likelihood effect sizes were plotted against $\widehat{\lambda}$, with points scaled and hued in association with $\widehat{\lambda}$ to visualize patterns. In such plots, points were colored if significant, based on the RRPP permutation test, or gray if not significant. We anticipated that $Z = 1.96$ should correspond to the null hypothesis rejection limit for a one-tailed test, a line we superimposed into plots to visually determine the consistency of effect sizes and hypothesis test results. (We expected colored points to lie above this line and gray points to lie beneath, if effect sizes reflected hypothesis test outcomes.)

Because we had p-values from both parametric and permutation tests, we could create 2 × 2 tables of hypothesis test outcome correspondence, to assess the consistency of parametric and non-parametric tests. These tables report both the consistencies (parametric tests and RRPP tests found same result) and two types of inconsistencies: the parametric test finds a significant result but RRPP does not, or RRPP finds a significant result, but the parametric test does not. The inconsistent results were labeled in plots, along with type I error rates (calculated from the frequency of occasions that for $\widehat{\lambda} \approx 0$, a significant result was observed). Because we had 5,000 $\widehat{\lambda}$ values, approximately uniform in distribution, we were able to estimate statistical power curves as a moving proportion of null hypothesis rejections across the landscape of $\widehat{\lambda}$ values. We estimated the proportion of rejections for four test types: parametric/untransformed, parametric/transformed, RRPP/untransformed, and RRPP/transformed. Proportions were estimated by culling data by intervals of $\widehat{\lambda}$ and producing a vector of 0s (did not reject null

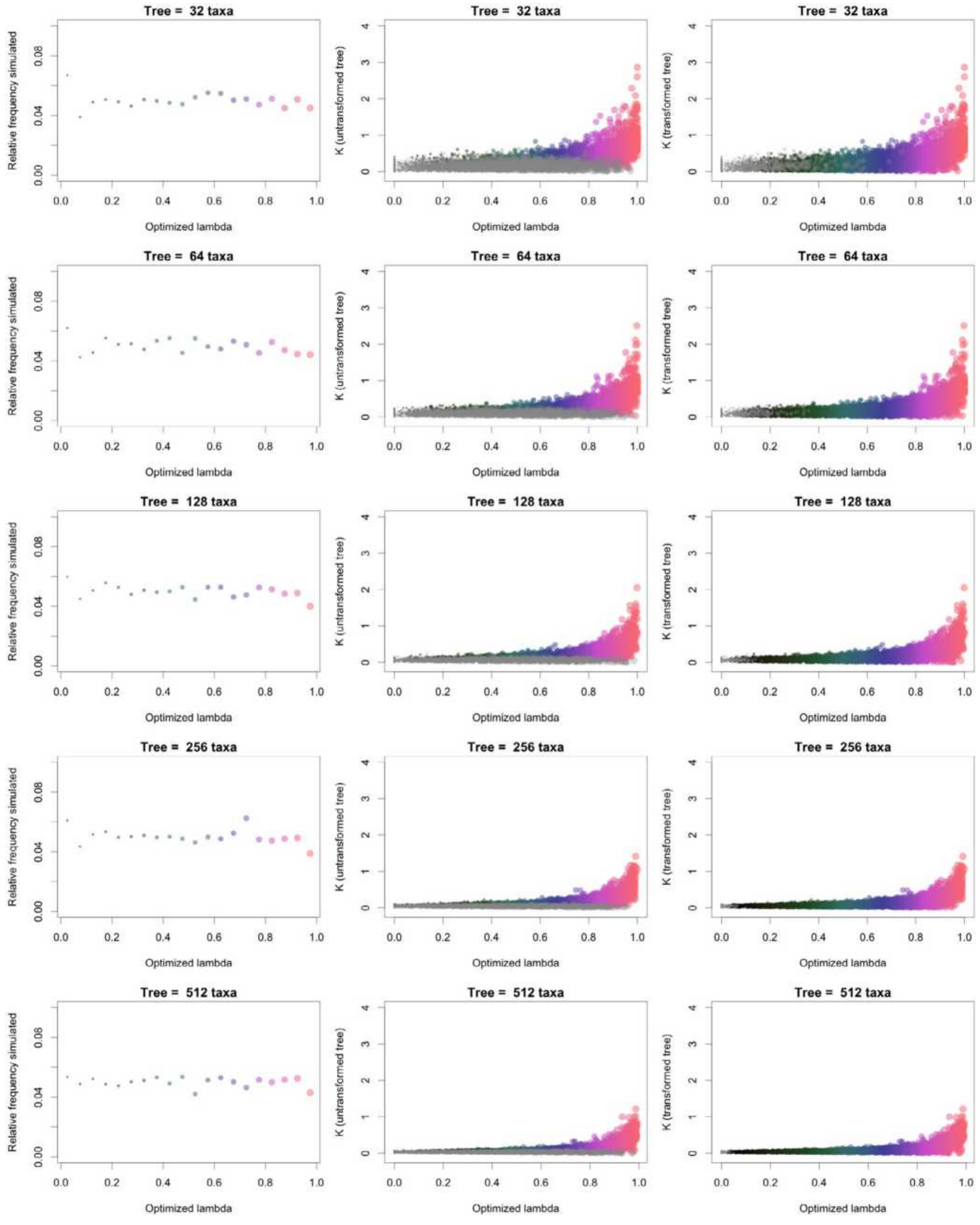

FIGURE 2 Plots of simulations including relative frequencies of $\hat{\lambda}$ generated (left column), K tested with $\lambda = 1$ (middle column), and K tested with $\lambda = \hat{\lambda}$ (right column). Rows separate results by tree size. Points corresponding to non-significant results from permutation tests are colored gray; significant results are scaled, colored and hued according to the magnitude of $\hat{\lambda}$, as in the left column

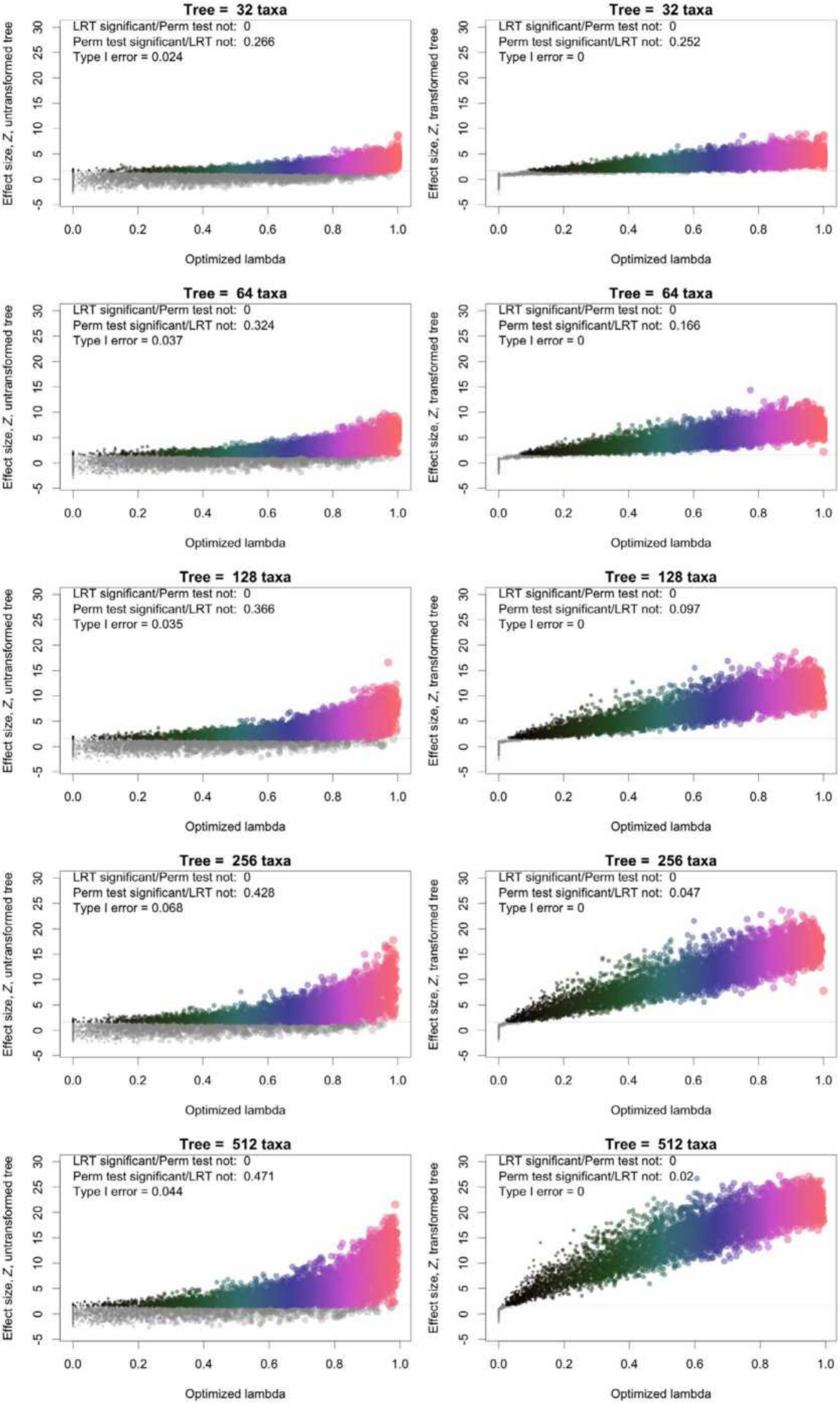
Tree = 32 taxa
LRT significant/Perm test not: 0
Perm test significant/LRT not: 0.266
Type I error = 0.024
Effect size, Z, untransformed tree
-5 0 5 10 15 20 25 30
0.0 0.2 0.4 0.6 0.8 1.0
Optimized lambda
Tree = 32 taxa
LRT significant/Perm test not: 0
Perm test significant/LRT not: 0.252
Type I error = 0
Effect size, Z, transformed tree
-5 0 5 10 15 20 25 30
0.0 0.2 0.4 0.6 0.8 1.0
Optimized lambda
Tree = 64 taxa
LRT significant/Perm test not: 0
Perm test significant/LRT not: 0.324
Type I error = 0.037
Effect size, Z, untransformed tree
-5 0 5 10 15 20 25 30
0.0 0.2 0.4 0.6 0.8 1.0
Optimized lambda
Tree = 64 taxa
LRT significant/Perm test not: 0
Perm test significant/LRT not: 0.166
Type I error = 0
Effect size, Z, transformed tree
-5 0 5 10 15 20 25 30
0.0 0.2 0.4 0.6 0.8 1.0
Optimized lambda
Tree = 128 taxa
LRT significant/Perm test not: 0
Perm test significant/LRT not: 0.366
Type I error = 0.035
Effect size, Z, untransformed tree
-5 0 5 10 15 20 25 30
0.0 0.2 0.4 0.6 0.8 1.0
Optimized lambda
Tree = 128 taxa
LRT significant/Perm test not: 0
Perm test significant/LRT not: 0.097
Type I error = 0
Effect size, Z, transformed tree
-5 0 5 10 15 20 25 30
0.0 0.2 0.4 0.6 0.8 1.0
Optimized lambda
Tree = 256 taxa
LRT significant/Perm test not: 0
Perm test significant/LRT not: 0.428
Type I error = 0.068
Effect size, Z, untransformed tree
-5 0 5 10 15 20 25 30
0.0 0.2 0.4 0.6 0.8 1.0
Optimized lambda
Tree = 256 taxa
LRT significant/Perm test not: 0
Perm test significant/LRT not: 0.047
Type I error = 0
Effect size, Z, transformed tree
-5 0 5 10 15 20 25 30
0.0 0.2 0.4 0.6 0.8 1.0
Optimized lambda
Tree = 512 taxa
LRT significant/Perm test not: 0
Perm test significant/LRT not: 0.471
Type I error = 0.044
Effect size, Z, untransformed tree
-5 0 5 10 15 20 25 30
0.0 0.2 0.4 0.6 0.8 1.0
Optimized lambda
Tree = 512 taxa
LRT significant/Perm test not: 0
Perm test significant/LRT not: 0.02
Type I error = 0
Effect size, Z, transformed tree
-5 0 5 10 15 20 25 30
0.0 0.2 0.4 0.6 0.8 1.0
Optimized lambda

FIGURE 3 Plots of Z score simulations for untransformed $\mathbf{C}$ (left columns) and transformed $\mathbf{C}$ (right column). Points are scaled and colored as in Figure 2. Inconsitencies (from a 2×2 correspondence table) between parametric likelihood ratio tests (LRT) and RRPP permutation tests are noted, as well as type I error rates for $\hat{\lambda} = 0$. The dashed line at $Z = 1.96$ indicates expected separation of significant and non-significant results, based on a level of significance of $\alpha = 0.05$

hypothesis) and 1s (rejected the null hypothesis), for each test type. The means of these vectors were the proportion of tests for which the null hypothesis was rejected. We found that an interval length of 0.1 assured more than 500 values for all interior points ($\hat{\lambda} = 0.1{:}\,0.9$), and produced rather smooth curves that were qualitatively as informative as any curves produced with a greater number of intervals.

Additional functions and code for simulation experiments can be found in Appendix 3 of the Supporting Information. Several support functions from the `RRPP R` package (Collyer & Adams, 2018) were used to create functions to estimate log-likelihoods and effect sizes from RRPP distributions. In some initial simulations, we also considered balanced and pectinate trees. We found no qualitative differences and simulations could only produce new sets of data on the same tree, so we did not consider them further.

3.2 | Simulation results

Figure 2 shows $\hat{\lambda}$ and K results from simulations, and Figure 3 shows Z score results, with corresponding points scaled and hued the same, based on the value of $\hat{\lambda}$ in the first column of Figure 2. The relative frequencies of $\hat{\lambda}$ suggested the simulations produced approximately uniformly distributed phylogenetic signals. A rejection of the null hypothesis of no phylogenetic signal (significant phylogenetic signal) resulted in points that were colored, with hue changing as $\hat{\lambda}$ increased; non-significant values were gray in color. These figures allow for visual clarification of various attributes acquired from the simulation runs, such that patterns are easy to interpret. Statistical power curves are shown in Figure 4.

Regarding Pagel's λ and Blomberg's K, our simulation results tracked the results of Münkemüller et al. (2012) in one particular way (Figure 2). It was possible to simulate larger K values for smaller trees, but within any tree size, K tended to be less than 1 except for the largest simulated $\hat{\lambda}$ values. Despite this trend, the hypothesis test results using alternative transformations of the $\mathbf{C}$ matrix for estimation of MSE as a test statistic for K revealed profound differences. In Figure 2, significant or non-significant K values can be found for any $\hat{\lambda}$, if $\lambda = 1$ is forced in the test statistic, which is the common way this test is performed (middle column). The simple act of using $\lambda = \hat{\lambda}$ and MSE as a test statistic (not K) alleviated this concern, and was consistent with the assertion of Blomberg et al. (2003) that doing so increases statistical power (Figure 4). Forcing $\lambda = 1$ for hypothesis tests of K also elevated type I error rates, but they were still close to the nominal $\alpha = 0.05$ level.

Issues with λ forced to be equal to 1 were also revealed by using a standardized effect size based on the location of $\mathscr{L}_{\mathbf{C}}$ in its RRPP-generated sampling distribution. Significant and non-significant results spanned the entire range of $\hat{\lambda}$ (Figure 3). These results are not surprising, as they do not seek to maximize likelihood, but help to confirm that the permutation test with K, using MSE as a test statistic, is flawed (since MSE might not be minimized via a best fit of the tree to the comparative data). Furthermore, using Z as an effect size if λ is forced to equal 1 makes little sense because of its curvilinear association with $\hat{\lambda}$ (Figure 3). However, for cases where $\lambda = \hat{\lambda}$, both the permutation test on the log-likelihood statistic as well as the Z score from the RRPP sampling distribution, as an effect size, had several desirable attributes.

First, the permutation test for the log-likelihood of the evolutionary model had greater statistical power than the parametric likelihood ratio test (Figure 4). A statistical power advantage was greatest for smaller trees, and the power curves of the two methods tended to converge with larger trees. The cases of inconsistent results from the 2×2 hypothesis test correspondence tables (Figure 3) were always due to the permutation test finding significant results when the parametric likelihood ratio test did not, but the rate of inconsistencies decreased with increased tree size. (By contrast, if $\lambda = 1$ is forced, the rate of inconsistencies increased with tree size.) A likelihood ratio statistic only asymptotically follows a χ^2 distribution, as $N \to \infty$ (Wilks, 1938), so it is not surprising that a parametric likelihood ratio test would have larger type II error rates with small tree sizes. Furthermore, the asymptotic null distribution for a one-sided likelihood ratio statistic, in which null hypotheses are at the limits of the constrained parameter space ($\lambda = 0$ or $\lambda = 1$), is a mixture of two χ^2 distributions (Molenberghs & Verbeke, 2007; Verbeke & Molenberghs, 2003). Generally, an unconstrained χ^2 statistic is reported, but a p-value is considered to be 1/2 of the classical χ^2 approximation, when mixture proportions are equal. Therefore, a tendency toward high statistical power might be expected for traits from large trees with likelihood ratio tests. Nonetheless, the statistical power was as good or better with permutation tests in our results, irrespective of tree size. In addition to having greater statistical power, the RRPP sampling distribution allows the standard deviate of the observed log-likelihood to be used as an effect size (Z), which also has nice attributes.

Second, Z based on a maximum likelihood estimate of λ has a linear association with $\hat{\lambda}$. The slope of this linear association increases with tree size, unfortunately, as it is not possible to disentangle a goodness of fit (σ^2) from the size of a tree. Thus, one might consider comparing effect sizes for traits from two vastly different trees with caution. The range of simulated Z also increased with phylogenetic signal strength and tree size (Figure 3). This result can be explained by the fact that for a large value of $\hat{\lambda}$, also for a large tree, the breadth of possible σ^{2*} values in RRPP permutations increases, so it is also possible to have a larger span of possible Z values.

Third, with $\alpha = 0.05$, there was a clear demarcation of Z above a value of 1.96 corresponding to significant hypothesis test outcomes,

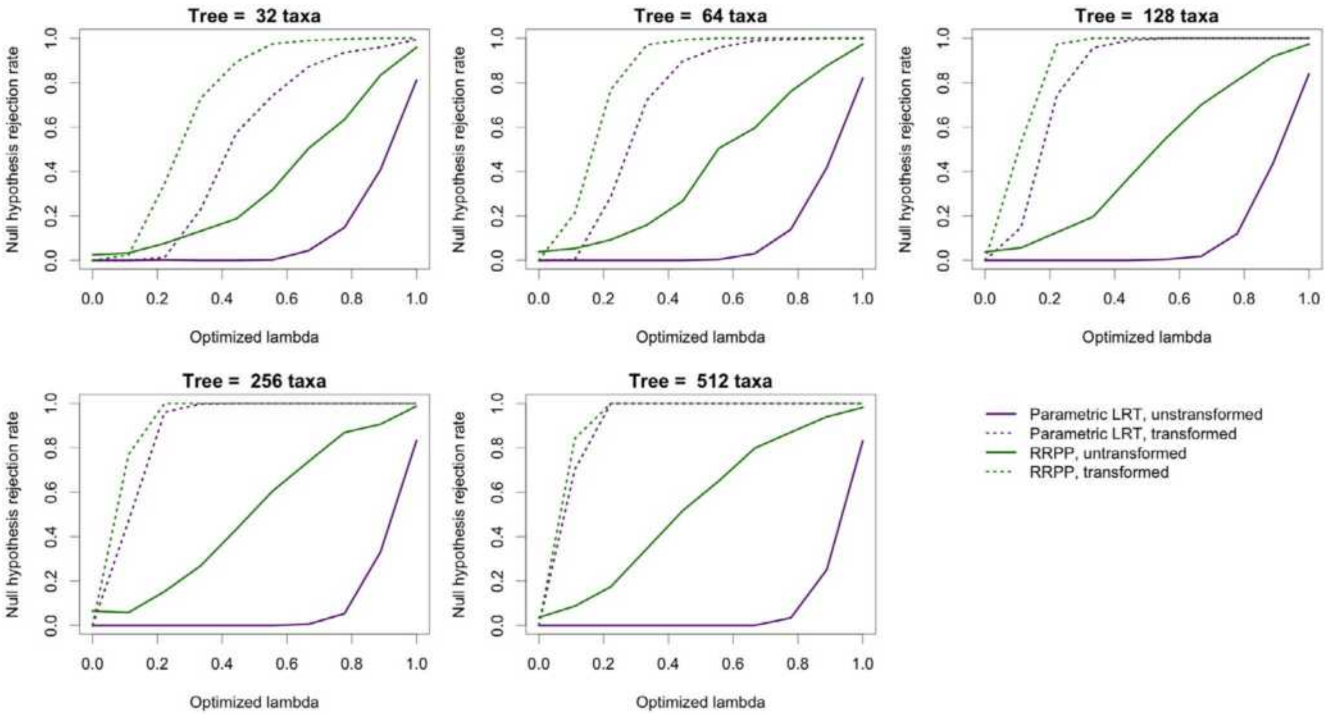

FIGURE 4 Statistical power curves for indicated methods: parametric likelihood ratio test (LRT), and RRPP. Whether the **C** matrix is transformed is noted

especially for tests with $\lambda = \hat{\lambda}$ (Figure 3). This is helpful, as an effect size of say, $Z = 2.5$ reported from an empirical study, indicates significant phylogenetic signal. We found sampling distributions to be consistently normally distributed (see Figure 1 as an example), especially for larger trees. Results in Figure 3 (second column) had a distinction of significant results consistently for $Z > 1.96$. A reliance on the normal distribution of RRPP sampling distributions means that two-sample Z statistics are also reliable, and quick interpretation of $Z = 25$ to, e.g. $Z = 6$, for two traits from the same tree, indicates which has greater phylogenetic signal.

Ideally, there would have been no relationship between Z and tree size, but such an expectation would be unwarranted, as phylogenetic signal is inherently related to the largeness of the phylogeny. However, we determined that there was a precise relationship between tree size and the slope of Z with respect to $\hat{\lambda}$. The slopes of the lines in Figure 3 fit (nearly perfectly) the relationship, $\log\left(\frac{Z}{\hat{\lambda}}\right) = \frac{1}{2}\log N$. Thus, the expected value of Z, given N and $\hat{\lambda}$ is $E(Z|N, \hat{\lambda}) = \exp\left[\frac{1}{2}\log N + \log\hat{\lambda}\right]$. One can calculate $Z - E(Z|N, \hat{\lambda})$ for the traits (see Figure 5) that are compared to ascertain if Z is larger (more positive) or lesser (more negative) than expected, given the tree size and optimized value of λ. This adjustment could be seen at best as a tool to help understand the multifarious nature of phylogenetic signal, rather than fix Z for comparative tests. For example, when comparing traits from two different trees, more positive values of $Z - E(Z|N, \hat{\lambda})$ might be considered stronger phylogenetic signal, if $\hat{\lambda}$ are comparable.

4 | EMPIRICAL EXAMPLE

To demonstrate the utility of Z_{12}, we compared Z for two ecologically-relevant traits in plethodontid salamanders (Figure 6): surface area to volume ratios $\left(\frac{SA}{V}\right)$ and relative (to snout to vent length) body width $\left(\frac{BW}{SVL}\right)$ (Baken & Adams, 2019; Baken et al., 2020). For this example, surface area to volume ratios and relative body width measures were obtained from individuals of 305 species, from which species means were obtained (Baken & Adams, 2019; Baken et al., 2020). A time-dated molecular phylogeny for the group (Bonett & Blair, 2017) was pruned to match the species in the phenotypic dataset. The phylogenetic signal effect size in each trait was obtained from 10,000 RRPP permutations, using functions described in Appendix 3 of the Supporting Information. The absolute value of the two-sample effect size (Equation 5) was calculated, as we had no a priori expectation of direction in the hypothesis test; i.e. it was treated as a two-tailed hypothesis test.

Although both traits contained significant phylogenetic signal ($Z_{\frac{BW}{SVL}} = 16.17; p = 0.0001$ and $Z_{\frac{SA}{V}} = 21.20; p = 0.0001$), a test based on Z_{12} revealed that the degree of phylogenetic signal was significantly stronger in $\frac{SA}{V}$ ($|Z_{12}| = 7.10$; $p < 0.0001$: Figure 5). Biologically, this observation may be interpreted by the fact that the tropical species which form a monophyletic group within plethodontids display greater variation in $\frac{SA}{V}$, which covaries with disparity in their climatic niches (Baken et al., 2020). Thus, greater phylogenetic signal in $\frac{SA}{V}$ is to be expected. Coincidentally,

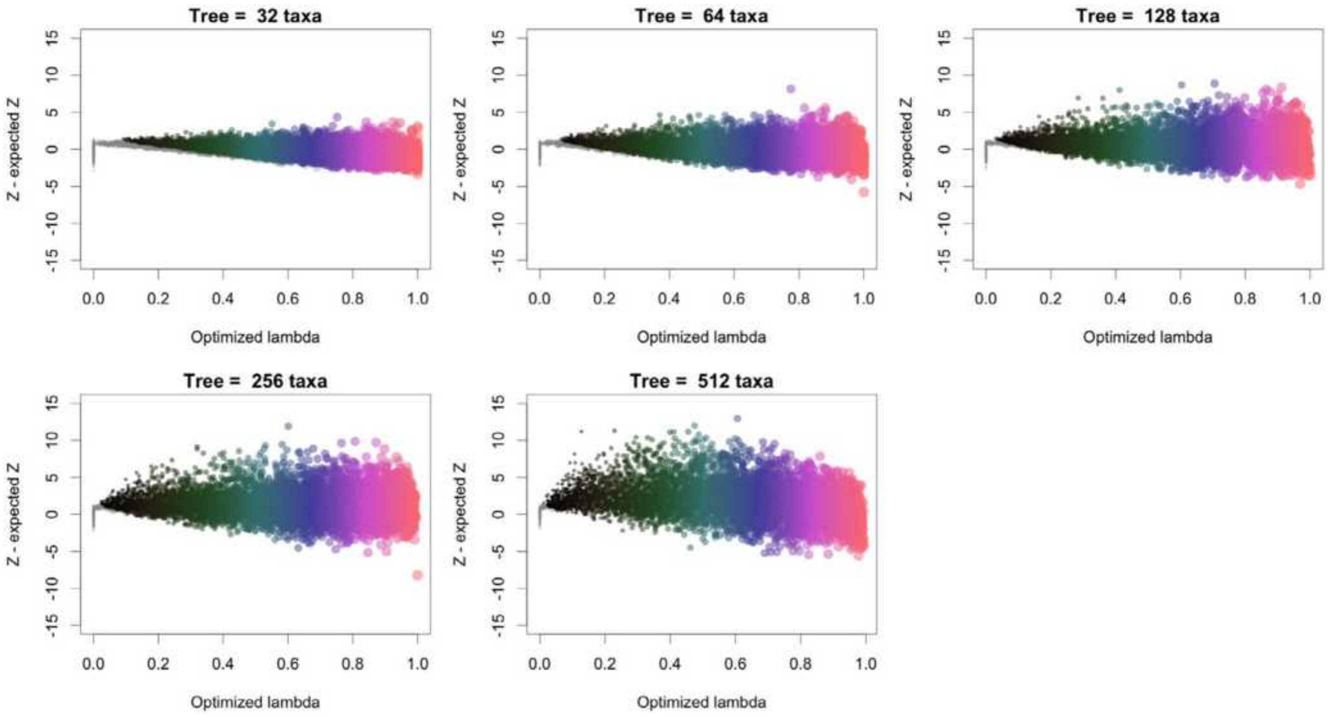

FIGURE 5 Z scores from Figure 3 (right column) after subtracting the expected value of Z, based on $\hat{\lambda}$ and N

$\hat{\lambda}$ was 0.76 and 0.91, and K was 0.25 and 0.76, for $\frac{BW}{SVL}$ and $\frac{SA}{V}$, respectively.

5 | DISCUSSION

To be able to ask if traits differ in their amount of phylogenetic signal, resolving how to best measure phylogenetic signal is essential. In this study, we considered the two most common measures of phylogenetic signal, and our simulation results did not dispute any issues that were already known about these measures. For example, the precision to estimate $\hat{\lambda}$ is tree-dependent, with more taxa-rich trees required for better precision (Boettiger et al., 2012; Münkemüller et al., 2012). K does not scale linearly with increased phylogenetic signal strength, and its variance is positively associated with phylogenetic signal strength (Diniz-Filho et al., 2012; Münkemüller et al., 2012). Our simulation results confirmed these attributes. These issues make the comparison of phylogenetic signals challenging, even if only qualitatively comparing $\hat{\lambda}$ or K between traits, for the same phylogeny. That there has been no statistical test only makes inference more speculative.

In this study, we made three important advances for the comparison of phylogenetic signals among different traits. First, we demonstrated that a permutation-based procedure (RRPP) using the log-likelihood as a statistic is not only reliable but performs better than a parametric test, especially for smaller trees. Second, we demonstrated that if the RRPP-log-likelihood permutation test is used, a test of $\hat{\lambda}$ and K are the same, provided that $\mathbf{C}$ is transformed by $\hat{\lambda}$ in the calculation of the GLS variance that is at the heart of the calculation of either statistic. Indeed, Blomberg et al. (2003) introduced K as a statistic that had an associated permutation test, based on a distribution of MSE, not K, noting that statistical power would be higher if MSE was calculated from a transformed version of $\mathbf{C}$ that resulted in better fit of the tree to the data. Because $-MSE$ and $\mathcal{L}_{\mathbf{C}}$ are perfectly rank-order correlated for the same set of RRPP permutations, viewing $\hat{\lambda}$ and K as statistics that have different hypothesis test outcomes is not necessary. Previous simulation studies have found differences between them, but did so by relying on adjudication of $\hat{\lambda}$ by a parametric likelihood ratio test ($\mathbf{C}$ transformed by $\hat{\lambda}$), and K by a permutation test with no transformation of $\mathbf{C}$ ($\lambda = 1$) (see, e.g. Molina-Venegas & Rodríguez, 2017; Münkemüller et al., 2012). Our work reveals that these differences were the result of the incommensurate transformation step, and not in test statistic performance, per se. Third, having demonstrated that a test of phylogenetic signal is a test of the rarity of the observed $\mathcal{L}_{\mathbf{C}}$ in a distribution of random outcomes, generated by a null model of phylogenetic independence, we can measure phylogenetic signal an alternative way: as the standardized location of the observed $\mathcal{L}_{\mathbf{C}}$ in the RRPP-generated distribution of random values (i.e. as an effect size). This alternative makes it possible to perform hypothesis tests for the comparison of the strength of phylogenetic signal across traits.

This third advance is important but perhaps unsettling. The convenience of $\hat{\lambda}$ or K is that a value of 0 should mean data devoid of phylogenetic signal, and a value of 1 should mean data have a phylogenetic signal that matches a BM model of evolutionary divergence. By contrast, a Z-score is a value measured in standard deviations that indicates a location in a normal distribution relative to expectation (mean), given a null model of phylogenetic

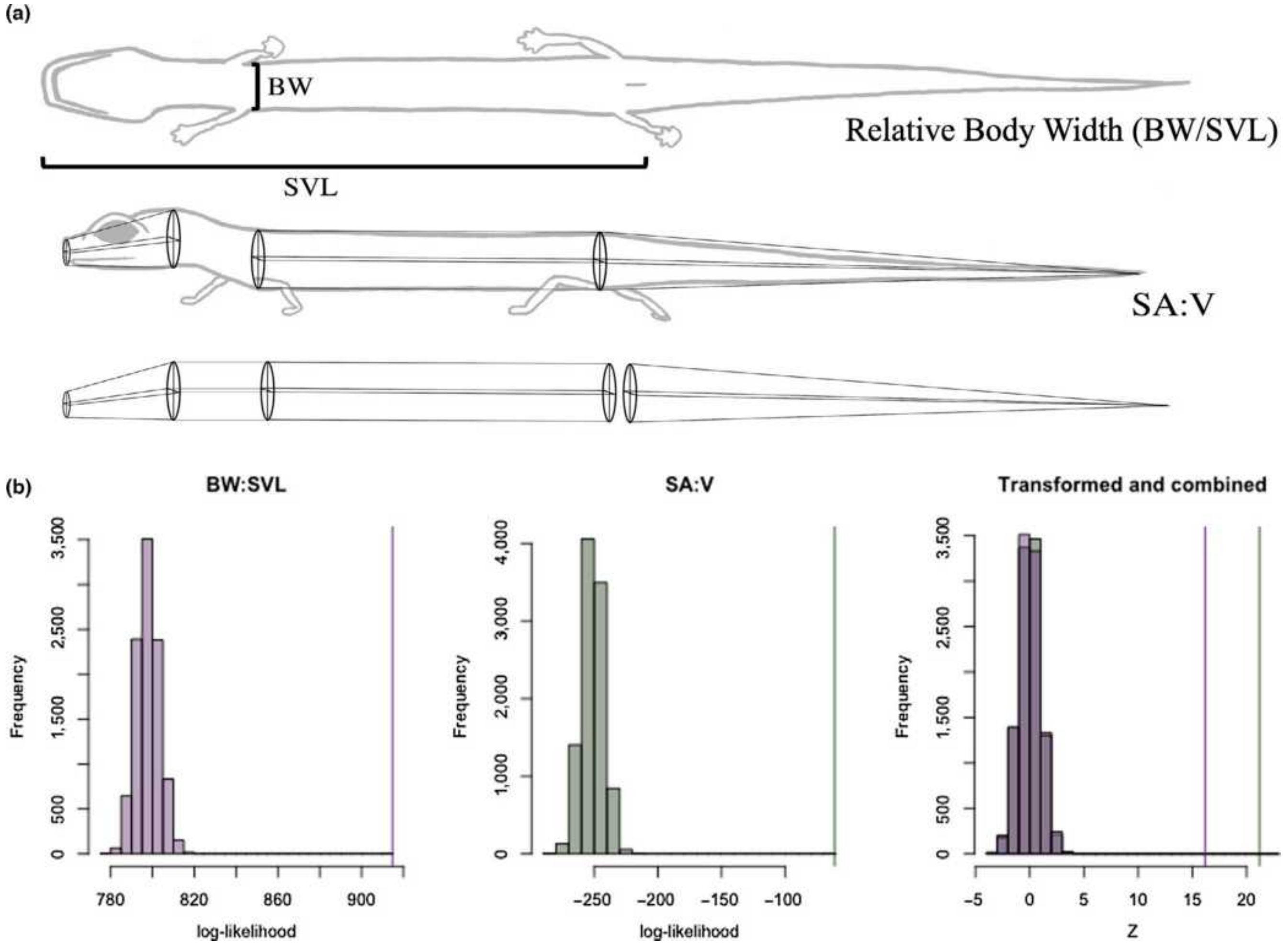

FIGURE 6 A Description of traits compared; B comparison of traits via standardized effect sizes (shown as locations in standardized sampling distributions, after transforming sampling distributions of log-likelihoods)

independence. As a measure of the degree of phylogenetic signal, Z, might feel less intuitively comfortable. However, this discomfort is perhaps predicated on one's definition of phylogenetic signal. For example, in a tree with 128 taxa (Figure 3), for a value of $\hat{\lambda} = 0.5$, Z might range from 2 to 17. If $\hat{\lambda}$ is the definition of phylogenetic signal strength, the range of effect sizes suggest that Z is not a good effect size to consider. Conversely, an effect size $Z = 5$ might be found to have $\hat{\lambda}$ range from 0.1 to 1.0. That is, if Z is the measure of phylogenetic signal, a corresponding $\hat{\lambda}$ indicates the tree transformation that best reveals the phylogenetic signal, not the amount of phylogenetic signal. However, the appeal of Z is that it allows a statistical comparison of the phylogenetic signals from multiple traits, but those traits might also have quite different values of $\hat{\lambda}$ or K. The best analysis is probably one that statistically compares Z but also reports both $\hat{\lambda}$ and K, as these two statistics have important meaning: the optimized branch-length transformation and a ratio that expresses the relative amount of BM contribution to the GLS variance estimate, respectively. That is, one can report Z, $\hat{\lambda}$, K, and one p-value, and not have to view phylogenetic signal statistics as a means to an end for different statistical tests.

It is common for researchers to report weak but significant phylogenetic signal when K is considerably less than 1 but the null hypothesis test is rejected. We also demonstrated with our simulations that it is possible to find significant phylogenetic signal when $\hat{\lambda}$ is small compared to a non-significant result when λ is forced to be equal to 1 (compare plots between left and right columns of Figure 3). Our work demonstrates that it is not helpful to declare weak but significant phylogenetic signal (especially if not simultaneously reporting strong but not significant phylogenetic signal by increasing λ), based on $\hat{\lambda}$ or K values. However, we feel it is more appropriate to declare $Z = 2$ as weak but significant, compared with say, $Z = 15$, which is strong and significant. Phylogenetic signal strength can be viewed as measure of rarity to generate such a strong signal by chance, which Z describes well. Although $\hat{\lambda}$ and K are useful statistics, their ability to discern strong versus weak phylogenetic signal is questionable. Only Z, which is a statistical effect size, affords this statistical interpretation. However, an interpretation of phylogenetic strength still cannot be made independent of phylogeny size.

One less desirable outcome of our simulations is that Z (more precisely the slope between Z and $\hat{\lambda}$) was positively associated with N, the number of taxa represented in a tree. We were able to demonstrate that the slope between Z and $\hat{\lambda}$ is predicted by N, such that one could find an expected value of Z, given N and $\hat{\lambda}$. When comparing the same or multiple traits between trees, this added step might help to better elucidate differences between a two-sample test of phylogenetic signal, especially if Z_{12} is significant, but it is not clear if the test result is because of differences in phylogenetic signal strength or tree size. This would not be a panacea, as it would also involve using $\hat{\lambda}$, which could vary between traits, but it is a tool that might assist inferences made about differences in phylogenetic signal involving multiple trees, or different Z scores also involving different $\hat{\lambda}$ transformations.

Although using Z scores for comparative analysis offers new opportunities, it also presents new challenges. Chief among the challenges that will have to be addressed is how to generalize the Z score as an effect size for multivariate data, especially if the number of variables precludes calculating log-likelihood. We see three possible approaches. First, like the generalization of the K statistic for multivariate data (Adams, 2014a), it might be possible to use the trace of the evolutionary rate matrix, rather than the matrix determinant, which would not be variable-limited (for example, $\log\sigma^2 + \frac{1}{N}\log|\mathbf{C}|$ could be generalized by taking either the trace or determinant of $\mathbf{R}$, the multivariate generalization of σ^2). Research demonstrating the adequacy of this approach would be needed, and certainly, the random outcomes could not be called log-likelihoods, but if the sampling distributions of log-likelihoods and modified statistics using traces were commensurate for comparable sets of variables, and yielded similar Z scores, then using an alternative generalization would be possible. Second, one could use a penalized-log-likelihood based on a regularization of near-singular or singular $\mathbf{R}$ matrices (Clavel et al., 2019). Because this approach assures a $\mathbf{R}$ matrix that is positive-definite and invertable, it also assures that $\log\mathscr{L}_{\mathbf{C}}$ can be estimated in every random permutation. Whether, the distribution of random $\log\mathscr{L}_{\mathbf{C}}$ obtained from RRPP, followed by regularization in each permutation, yields appropriate sampling distributions would remain to be seen. The statistical properties have been adjudicated using a penalized-likelihood framework for evaluating Wilks' Λ, with RRPP (Clavel & Morlon, 2020), so there is promise that this framework would also work for calculating multivariate Z scores.

The third potential solution is to use phylogenetically aligned component analysis (PACA; Collyer & Adams, 2021) as a dimension reduction method. The perils of data reduction before likelihood estimation have been clearly demonstrated (Uyeda et al., 2015), but this was for cases where the data reduction method (principal component analysis, PCA) did not find components specifically aligned to phylogenetic signal. PACA specifically aligns components to phylogenetic signal, such that greater phylogenetic signal rather than variance is predominantly found in the first few components. It might be possible to use a subset of data dimensions that contain most or all phylogenetic signal to estimate pseudo-likelihoods. Again, it might not be sufficient to refer to a statistic calculated this way as model likelihood, but if random outcomes across many permutations produce a sampling distribution that yields similar Z values in fewer dimensions, it might be trusted for estimating Z for highly multivariate data.

Regardless of these three possible solutions, another consideration is whether different variables could have different $\hat{\lambda}$ in the estimation of log-likelihoods; that is, can it be assumed traits evolve independently? In Appendix 2 of the Supporting Information, we outline a method for calculating log-likelihoods for multivariate data, both assuming common and independent λ for traits. Model selection could be used to compare these two likelihoods to ascertain if traits evolve independently, and if so, the two-sample Z test described here could be used to determine which traits have greater phylogenetic signal. However, appropriate optimization methods for multiple λ should be rigorously researched, in addition to the statistical properties of different likelihood estimators, before solutions for multivariate traits are eagerly embraced.

Regardless of future challenges, the ability to estimate an effect size that can be used for hypothesis tests to compare phylogenetic signal in multiple traits, as a tool, is a boon for the PCM toolkit. We feel that measuring phylogenetic signal is more nuanced than using a single statistic, but adding Z to the suite of statistics used can help decipher between weak and strong phylogenetic signals, rather than misinterpreting values of $\hat{\lambda}$ or K. Our scope of investigation concerned BM models of evolutionary divergence and one transformation parameter, Pagel's λ. Pagel's λ is generally considered to be most associated with phylogenetic signal, but one could also consider using RRPP with additional transformation parameters, including δ and κ (Pagel, 1999). Because the transformation of the $\mathbf{C}$ matrix is an a priori step and this transformation is retained through random permutations, it would be easy to extend the RRPP-log-likelihood computations to additional $\mathbf{C}$ matrix transformations. Furthermore, RRPP could be used with alternative models of evolution (e.g. multi-rate Brownian, early burst, OU, AC/DC), recognizing that the simplifications we made from Equations 1 to 3 would not be the same. Random versions of $\mathbf{V}$ in Equation 1 would have to be calculated in each RRPP permutation, accounting appropriately for parameters that are fixed or variable in each permutation. Insomuch as phylogenetic signal effect size (using Z) is a measure of the fit of tree to comparative data for a BM model of evolution, any similar approach can be considered a model effect size for an alternative evolutionary model. Thus, there could be some appeal with using the RRPP-log-likelihood effect size as a model selection criterion, especially because multiple models could be compared, not just assuming a null model of phylogenetic independence, but other null models as well. (For example, a single rate BM model could serve as null model for various multi-rate alternative models. Comparison of Z among the models, using both phylogenetic independence and

BM as different null models, could be valuable for inferring the best evolutionary model for trait data.)

In answering certain evolutionary questions, such as comparing the strength of phylogenetic signal, traditional parametric approaches offer challenges. First, the parameter space of λ is bounded, and thus, a mixture of χ^2 distributions is required as a proxy for a sampling distribution (Molenberghs & Verbeke, 2007; Self & Liang, 1987; Verbeke & Molenberghs, 2003). Second, χ^2 distributions are asymptotically appropriate for likelihood ratio statistics for very large sample sizes (Wilks, 1938), a situation rarely afforded when working with phylogenies. The permutation test we presented is not constrained to use a parametric probability distribution as a proxy, and is additionally capable of providing effect sizes, which are comparable across datasets to evaluate comparative hypotheses. Prior work (Adams & Collyer, 2018) has shown that empirical sampling distributions generated from RRPP match nearly perfectly the parametric *F*-distributions typically used in ANOVA, when data are simulated to match the assumptions of ANOVA. Based on our work here, one might speculate that RRPP-generated sampling distributions are better proxies for statistics without appropriate parametric sampling distributions and converge on parametric distributions in cases where sampling distribution solutions are tractable. When viewed from this perspective, permutation methods such as RRPP should not be considered mere analytical band-aids to be used for ill-conditioned datasets, or scenarios where standard tests are not applicable. Rather, they are equivalent to parametric procedures for standard biostatistical problems and can supersede them in cases where parametric methods are not applicable. Thus, our perspective is that this work helps to continue to pave the way for advancement of PCMs as sets of tools that take advantage of the computational power of modern computers rather than force evolutionary biology questions into limited traditional statistics applications.

ACKNOWLEDGEMENTS

We thank Elizabeth Glynne and Bryan Juarez for comments on early drafts of the manuscript, and Simone Blomberg and Joe Felsenstein for their helpful comments on the penultimate draft. This work was sponsored in part by National Science Foundation Awards DBI-1902694 (to M.L.C.) and DBI-1902511 (to D.C.A.). National Science Foundation (DBI-1902511, DBI-1902694).

CONFLICTS OF INTEREST

The authors have no conficts of interest to declare.

AUTHORS' CONTRIBUTIONS

D.C.A. conceived the original idea for an effect size measure of phylogenetic signal; M.L.C. developed the log-likelihood permutation procedure that was implemented and D.C.A., E.K.B., and M.L.C. collaboratively developed the concept and contributed to all portions of this manuscript. All authors approve of the final product and are willingly accountable for any portion of the content.

DATA AVAILABILITY STATEMENT

Empirical data used in this paper are available on Dryad (associated with original article, Baken et al., 2020), https://doi.org/10.5061/dryad.59zw3r23m. R-scripts for simulated data tests are provided as Supporting Information.

ORCID

Michael L. Collyer https://orcid.org/0000-0003-0238-2201
Erica K. Baken https://orcid.org/0000-0003-2972-7900
Dean C. Adams https://orcid.org/0000-0001-9172-7894

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. (2013). Comparing evolutionary rates for different phenotypic traits on a phylogeny using likelihood. *Systematic Biology*, *62*(2), 181 192. https://doi.org/10.1093/sysbio/sys083

Adams, D. C. (2014a). A generalized K statistic for estimating phylogenetic signal from shape and other high-dimensional multivariate data. *Systematic Biology*, *63*(5), 685 697. https://doi.org/10.1093/sysbio/syu030

Adams, D. C. (2014b). A method for assessing phylogenetic least squares models for shape and other high-dimensional multivariate data. *Evolution*, *68*, 2675 2688. https://doi.org/10.1111/evo.12463

Adams, D. C., & Collyer, M. L. (2016). On the comparison of the strength of morphological integration across morphometric datasets. *Evolution*, *70*(11), 2623 2631. https://doi.org/10.1111/evo.13045

Adams, D. C., & Collyer, M. L. (2018). Phylogenetic ANOVA: Group-clade aggregation, biological challenges, and a refined permutation procedure. *Evolution*, *72*(6), 1204 1215. https://doi.org/10.1111/evo.13492

Adams, D. C., & Collyer, M. L. (2019). Comparing the strength of modular signal, and evaluating alternative modular hypotheses, using covariance ratio effect sizes with morphometric data. *Evolution*, *73*(12), 2352 2367. https://doi.org/10.1111/evo.13867

Baken, E. K., & Adams, D. C. (2019). Macroevolution of arboreality in salamanders. *Ecology and Evolution*, *9*(12), 7005 7016. https://doi.org/10.1002/ece3.5267

Baken, E. K., Mellenthin, L. E., & Adams, D. C. (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. https://doi.org/10.1111/evo.13898

Beaulieu, J. M., Jhwueng, D. C., Boettiger, C., & O'Meara, B. C. (2012). Modeling stabilizing selection: Expanding the ornstein-uhlenbeck model of adaptive evolution. *Evolution*, *66*(8), 2369 2383. https://doi.org/10.1111/j.1558-5646.2012.01619.x

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

Boettiger, C., Coop, G., & Ralph, P. (2012). Is your phylogeny informative? Measuring the power of comparative methods. *Evolution*, *66*(7), 2240 2251. https://doi.org/10.1111/j.1558-5646.2011.01574.x

Bonett, R. M., & Blair, A. L. (2017). Evidence for complex life cycle constraints on salamander body form diversification. *Proceedings of the National Academy of Sciences of the United States of America*, *114*, 9936 9941. https://doi.org/10.1073/pnas.1703877114

Bose, R., Ramesh, B. R., Pélissier, R., & Munoz, F. (2019). Phylogenetic diversity in the Western Ghats biodiversity hotspot reflects environmental filtering and past niche diversification of trees. *Journal of Biogeography*, *46*(1), 145 157. https://doi.org/10.1111/jbi.13464

Clavel, J., Aristide, L., & Morlon, H. (2019). A penalized likelihood framework for high-dimensional phylogenetic comparative methods and an application to new-world monkeys brain evolution. *Systematic Biology*, *68*(1), 93 116. https://doi.org/10.1093/sysbio/syy045

Clavel, J., & Morlon, H. (2020). Reliable phylogenetic regressions for multivariate comparative data: Illustration with the MANOVA and application to the effect of diet on mandible morphology in phyllostomid bats. *Systematic Biology*, *69*(5), 927 943. https://doi.org/10.1093/sysbio/syaa010

Collyer, M. L., & Adams, D. C. (2018). RRPP: An r package for fitting linear models to high-dimensional data using residual randomization. *Methods in Ecology and Evolution*, *9*(7), 1772 1779. https://doi.org/10.1111/2041-210x.13029

Collyer, M. L., & Adams, D. C. (2021). Phylogenetically aligned component analysis. *Methods in Ecology and Evolution*, *12*(2), 359 372. https://doi.org/10.1111/2041-210X.13515

Collyer, M. L., Sekora, D. J., & Adams, D. C. (2015). A method for analysis of phenotypic change for phenotypes described by high-dimensional data. *Heredity*, *115*(4), 357 365. https://doi.org/10.1038/hdy.2014.75

Commenges, D. (2003). Transformations which preserve exchangeability and application to permutation tests. *Journal of Nonparametric Statistics*, *15*, 171 185. https://doi.org/10.1080/1048525031000089310

Cooper, N., Jetz, W., & Freckleton, R. P. (2010). Phylogenetic comparative approaches for studying niche conservatism. *Journal of Evolutionary Biology*, *23*(12), 2529 2539.

Cooper, N., Thomas, G. H., Venditti, C., Meade, A., & Freckleton, R. P. (2016). A cautionary note on the use of ornstein uhlenbeck models in macroevolutionary studies. *Biological Journal of the Linnean Society*, *118*(1), 64 77. https://doi.org/10.1111/bij.12701

De Meester, G., Huyghe, K., & Van Damme, R. (2019). Brain size, ecology and sociality: A reptilian perspective. *Biological Journal of the Linnean Society*, *126*(3), 381 391. https://doi.org/10.1093/biolinnean/bly206

Diniz-Filho, J. A. F., Santos, T., Rangel, T. F., & Bini, L. M. (2012). A comparison of metrics for estimating phylogenetic signal under alternative evolutionary models. *Genetics and Molecular Biology*, *35*(3), 673 679. https://doi.org/10.1590/S1415-47572012005000053

Felsenstein, J. (1985). Phylogenies and the comparative method. *The American Naturalist*, *125*(1), 1 15. https://doi.org/10.1086/284325

Freckleton, R. P. (2012). Fast likelihood calculations for comparative analyses. *Methods in Ecology and Evolution*, *3*(5), 940 947. https://doi.org/10.1111/j.2041-210X.2012.00220.x

Freckleton, R. P., Harvey, P. H., & Pagel, M. (2002). Phylogenetic analysis and comparative data: A test and review of evidence. *The American Naturalist*, *160*, 712 726. https://doi.org/10.1086/343873

Garland, T. Jr, & Ives, A. R. (2000). Using the past to predict the present: Confidence intervals for regression equations in phylogenetic comparative methods. *The American Naturalist*, *155*, 346 364. https://doi.org/10.1086/303327

Gittleman, J. L., & Kot, M. (1990). Adaptation: Statistics and a null model for estimating phylogenetic effects. *Systematic Zoology*, *39*(3), 227. https://doi.org/10.2307/2992183

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

Harmon, L. J., Weir, J. T., Brock, C. D., Glor, R. E., & Challenger, W. (2008). GEIGER: Investigating evolutionary radiations. *Bioinformatics*, *24*, 129 131. https://doi.org/10.1093/bioinformatics/btm538

Harvey, P. H., & Pagel, M. D. (1991). *The comparative method in evolutionary biology. 239*, Oxford University Press.

Klingenberg, C. P., & Gidaszewski, N. A. (2010). Testing and quantifying phylogenetic signals and homoplasy in morphometric data. *Systematic Biology*, *59*(3), 245 261. https://doi.org/10.1093/sysbio/syp106

Lande, R. (1976). Natural selection and random genetic drift in phenotypic evolution. *Evolution*, 314 334. https://doi.org/10.1111/j.1558-5646.1976.tb00911.x

Martins, E. P., & Hansen, T. F. (1997). Phylogenies and the comparative method: A general approach to incorporating phylogenetic information into the analysis of interspecific data. *The American Naturalist*, *149*, 646 667. https://doi.org/10.1086/286013

Molenberghs, G., & Verbeke, G. (2007). Likelihood ratio, score, and Wald tests in a constrained parameter space. *The American Statistician*, *61*(1), 22 27. https://doi.org/10.1198/000313007X171322

Molina-Venegas, R., & Rodríguez, M. (2017). Revisiting phylogenetic signal; strong or negligible impacts of polytomies and branch length information? *BMC Evolutionary Biology*, *17*(1). https://doi.org/10.1186/s12862-017-0898-y

Münkemüller, T., Lavergne, S., Bzeznik, B., Dray, S., Jombart, T., Schiffers, K., & Thuiller, W. (2012). How to measure and test phylogenetic signal. *Methods in Ecology and Evolution*, *3*, 743 756. https://doi.org/10.1111/j.2041-210X.2012.00196.x

O'Meara, B. C., Ané, C., Sanderson, M. J., & Wainwright, P. C. (2006). Testing for different rates of continuous trait evolution using likelihood. *Evolution*, *60*, 922 933.

Pagel, M. D. (1999). Inferring the historical patterns of biological evolution. *Nature*, *401*, 877 884. https://doi.org/10.1038/44766

Pavoine, S., & Ricotta, C. (2013). Testing for phylogenetic signal in biological traits: The ubiquity of cross-product statistics. *Evolution*, *67*(3), 828 840. https://doi.org/10.1111/j.1558-5646.2012.01823.x

Pintanel, P., Tejedo, M., Ron, S. R., Llorente, G. A., & Merino-Viteri, A. (2019). Elevational and microclimatic drivers of thermal tolerance in Andean *Pristimantis* frogs. *Journal of Biogeography*, *46*(8), 1664 1675. https://doi.org/10.1111/jbi.13596

Potter, D. M. (2005). A permutation test for inference in logistic regression with small-and moderate-sized data sets. *Statistics in Medicine*, *24*(5), 693 708. https://doi.org/10.1002/sim.1931

Revell, L. J. (2010). Phylogenetic signal and linear regression on species data. *Methods in Ecology and Evolution*, *1*, 319 329. https://doi.org/10.1111/j.2041-210X.2010.00044.x

Revell, L. J., & Harmon, L. J. (2008). Testing quantitative genetic hypotheses about the evolutionary rate matrix for continuous characters. *Evolutionary Ecology Research*, *10*, 311 331.

Revell, L. J., Harmon, L. J., & Collar, D. C. (2008). Phylogenetic signal, evolutionary process, and rate. *Systematic Biology*, *57*(4), 591 601. https://doi.org/10.1080/10635150802302427

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

Self, S. G., & Liang, K.-Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, *82*(398), 605 610. https://doi.org/10.1080/01621459.1987.10478472

Su, G., Villéger, S., & Brosse, S. (2019). Morphological diversity of freshwater fishes differs between realms, but morphologically extreme species are widespread. *Global Ecology and Biogeography*, *28*(2), 211 221. https://doi.org/10.1111/geb.12843

Uyeda, J. C., Caetano, D. S., & Pennell, M. W. (2015). Comparative analysis of principal components can be misleading. *Systematic Biology*, *64*(4), 677 689. https://doi.org/10.1093/sysbio/syv019

Vandelook, F., Janssens, S., Gijbels, P., Fischer, E., Van den Ende, W., Honnay, O., & Abrahamczyk, S. (2019). Nectar traits differ between pollination syndromes in balsaminaceae. *Annals of Botany*, *124*(2), 269 279. https://doi.org/10.1093/aob/mcz072

Verbeke, G., & Molenberghs, G. (2003). The use of score tests for inference on variance components. *Biometrics*, *59*(2), 254 262. https://doi.org/10.1111/1541-0420.00032

Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, *9*(1), 60 62. https://doi.org/10.1214/aoms/1177732360

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher s website.

How to cite this article: Collyer, M. L., Baken, E. K., & Adams, D. C. (2021). A standardized effect size for evaluating and comparing the strength of phylogenetic signal. *Methods in Ecology and Evolution*, 00, 1 16. https://doi.org/10.1111/2041-210X.13749