

# A METHOD FOR ASSESSING PHYLOGENETIC LEAST SQUARES MODELS FOR SHAPE AND OTHER HIGH-DIMENSIONAL MULTIVARIATE DATA

Dean C. Adams<sup>1,2,3</sup>

<sup>1</sup>*Department of Ecology, Evolution, and Organismal Biology, Iowa State University, Ames, Iowa 50011*

<sup>2</sup>*Department of Statistics, Iowa State University, Ames, Iowa 50011*

<sup>3</sup>*E-mail: dcadams@iastate.edu*

Received April 22, 2014

Accepted May 28, 2014

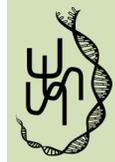
Studies of evolutionary correlations commonly use phylogenetic regression (i.e., independent contrasts and phylogenetic generalized least squares) to assess trait covariation in a phylogenetic context. However, while this approach is appropriate for evaluating trends in one or a few traits, it is incapable of assessing patterns in highly multivariate data, as the large number of variables relative to sample size prohibits parametric test statistics from being computed. This poses serious limitations for comparative biologists, who must either simplify how they quantify phenotypic traits, or alter the biological hypotheses they wish to examine. In this article, I propose a new statistical procedure for performing ANOVA and regression models in a phylogenetic context that can accommodate high-dimensional datasets. The approach is derived from the statistical equivalency between parametric methods using covariance matrices and methods based on distance matrices. Using simulations under Brownian motion, I show that the method displays appropriate Type I error rates and statistical power, whereas standard parametric procedures have decreasing power as data dimensionality increases. As such, the new procedure provides a useful means of assessing trait covariation across a set of taxa related by a phylogeny, enabling macroevolutionary biologists to test hypotheses of adaptation, and phenotypic change in high-dimensional datasets.

**KEY WORDS:** Geometric morphometrics, macroevolution, morphological evolution, phylogenetic comparative method.

In comparative biology it is generally accepted that the evaluation of evolutionary correlations among traits requires a phylogenetic perspective. Because of shared evolutionary history, species in a phylogeny are not independent (Felsenstein 1985; Harvey and Pagel 1991; Felsenstein 2004; Revell et al. 2008; Pennell and Harmon 2013). This leads to phenotypic similarities among taxa, whose expected covariance may be described in proportion to their shared ancestry under a specified model of trait evolution (Felsenstein 1985; Grafen 1989; Martins and Hansen 1997; Rohlf 2001). As a consequence, statistical procedures that assess patterns of covariation among traits from cross-species data must take the evolutionary relationships among taxa into consideration during the analysis, so that the lack of independence among

species may be properly accounted for (Revell 2010). In these circumstances, when evolutionary correlations among traits are identified via phylogenetic comparative methods, such correlations provide important evidence of adaptation and inform on the evolutionary causes of phenotypic change (Garland et al. 1992; Rüber and Adams 2001; Pennell and Harmon 2013), as the potential effect of shared evolutionary history may be ruled out.

Over the past several decades, numerous analytical approaches have been developed to account for phylogenetic nonindependence while assessing patterns of trait covariation (Cheverud et al. 1985; Felsenstein 1985; Grafen 1989; Martins and Hansen 1997; Diniz-Filho et al. 1998; Garland and Ives 2000; Guill et al. 2003). Of the approaches available, phylogenetic



regression (Felsenstein 1985; Grafen 1989) is most commonly used for assessing evolutionary associations in a comparative framework. Phylogenetic regression, which can be implemented either using independent contrasts (PIC: Felsenstein 1985) or generalized least squares (PGLS: Grafen 1989; Martins and Hansen 1997), evaluates trait associations as described by linear models, where one trait is treated as the dependent (response) variable, and one or more additional traits are treated as explanatory (independent) variables (Grafen 1989; Rohlf 2001; Pennell and Harmon 2013; described below). Most empirical studies of evolutionary covariation examine patterns between one or a few traits. Some classic examples include characterizing the evolutionary correlations between body mass and home range size in mammals (Garland et al. 1992), evaluating the correlation between life history traits and growth-related variables in maples (Ackerly and Donoghue 1998), assessing the relationship between limb measurements and functional performance in lizards (Losos 1990), and evaluating the association between genome size and developmental rate in salamanders (Sessions and Larson 1987).

Phylogenetic regression can also be used to evaluate evolutionary patterns in multivariate datasets. In this case, the response variable may be a set of measured traits treated simultaneously, or may be a complex multidimensional trait like shape (for instance, as quantified from geometric morphometric methods: Bookstein 1991; Adams et al. 2013). Indeed, several studies have investigated multivariate phenotypic evolution using phylogenetic regression using shape as data (e.g., Rüber and Adams 2001; McPeck et al. 2008; Monteiro and Nogueira 2011; Blankers et al. 2012; Klingenberg and Marugán-Lobón 2013; Outomuro et al. 2013a,b; Piras et al. 2013; for applications to other multivariate data types see: Lawing and Polly 2011). However, one underappreciated challenge for these approaches when evaluating highly multivariate datasets is that as the number of trait dimensions equals or exceeds the number of species in the phylogeny ( $p \geq N$ ), parametric methods such as multivariate regression and MANOVA cannot be used to assess significance. The reason for this is that when  $p \geq N$ , the covariance matrices obtained from these models are singular. Thus, parametric test statistics that require inverting a covariance matrix or finding its determinant (Wilks' lambda, Pillai's trace, etc.) cannot be computed, and the matrix computations for obtaining the likelihood of such models cannot be completed (see Adams 2014a). Furthermore, for a given number of species, it may be expected that the ability to detect evolutionary patterns in such data will decrease as trait dimensionality increases (a property that is demonstrated below).

Thus, while standard procedures for implementing phylogenetic regression can accommodate multivariate datasets, their utility is compromised by the complexity of the phenotypic data under consideration. For this reason, numerous authors have reduced the dimensionality of their data by representing it using

a smaller set of principal component axes before implementing phylogenetic analyses (Bergmann et al. 2009; Nogueira et al. 2009; Monteiro and Nogueira 2011; Brusatte et al. 2012; see recommendations in Monteiro 2013). However, this procedure is not optimal, because one cannot assume that patterns of evolutionary covariation between multivariate phenotypes and other factors align with the major axes of variation that summarize the phenotypic traits (Monteiro 2013). Therefore, for high-dimensional phenotypic datasets, an alternative procedure is required.

In this article, I propose a new statistical procedure for evaluating linear models in a phylogenetic context, under a Brownian motion model of evolution, for high-dimensional multivariate datasets like shape. The approach is derived from the statistical equivalency between methods based on covariance matrices and those based on distance matrices. I show that for univariate data, the distance-based approach (*D*-PGLS) provides numerically identical estimates of evolutionary patterns to those obtained from standard implementations of phylogenetic regression, demonstrating the statistical equivalency between the two. I then demonstrate that the approach displays acceptable Type I error and high statistical power for detecting evolutionary correlations between an independent variable and high-dimensional multivariate traits like shape. Further, I show that for multivariate data, standard implementations of phylogenetic regression exhibit decreasing statistical power with increasing dimensionality of the response variable. Thus, these methods are hampered in their ability to identify evolutionary associations in a phylogenetic context for high-dimensional phenotypic data. Finally, I present a biological example demonstrating the utility of the new approach for high-dimensional datasets. Computer code written in R for implementing the procedure is also provided.

## *Phylogenetic Regression via Generalized Least Squares*

One approach for assessing the evolutionary covariation between variables in a phylogenetic context is phylogenetic regression, or phylogenetic generalized least squares (PGLS: Grafen 1989; Martins and Hansen 1997). For regression models, PGLS is mathematically equivalent to phylogenetically independent contrasts (PIC: Garland and Ives 2000; Rohlf 2001; Blomberg et al. 2012). However, phylogenetic regression can more easily accommodate alternative statistical designs such as ANOVA and more complicated factorial designs (Pennell and Harmon 2013), and can be used with different evolutionary models (see below). With phylogenetic regression, phylogenetic relationships are taken into account while parameterizing the linear model  $\mathbf{Y} = \mathbf{XB} + \epsilon$ , which is then statistically evaluated. Model parameterization is accomplished by incorporating the expected covariance due to the

phylogeny into the residual error,  $\varepsilon$  (Martins and Hansen 1997). Under a Brownian motion model of evolution, the lack of independence due to shared ancestry is described by the phylogenetic covariance matrix ( $\mathbf{C}$ ), which is an  $N \times N$  matrix whose diagonal elements contain the phylogenetic distance from the root of the tree to each of the  $N$  species tips, and the off-diagonals contain the phylogenetic distances from the root of the tree to the most recent common ancestor for each pair of species (Garland and Ives 2000; Rohlf 2001). In some cases, suitable branch-length transformations of  $\mathbf{C}$  can be used to represent alternative evolutionary models, such as Ornstein–Uhlenbeck processes or an ACDC model of evolutionary change (see Butler et al. 2000; Blomberg et al. 2003).

Using phylogenetic regression on a single dependent ( $\mathbf{Y}$ ) and independent ( $\mathbf{X}$ ) variable, parameter estimates describing the relationship between the two variables while accounting for phylogeny may be found as

$$\hat{\beta} = (\mathbf{X}^t \mathbf{C}^{-1} \mathbf{X})^{-1} \mathbf{X}^t \mathbf{C}^{-1} \mathbf{Y}. \quad (1)$$

Typically, a column of ones is appended to  $\mathbf{X}$  to include the intercept (Rohlf 2001). Predicted values from the model ( $\hat{\mathbf{Y}}$ ) are then estimated, and used to determine the residual sums-of-squares (Martins and Hansen 1997):

$$\hat{\mathbf{Y}} = \mathbf{X} \hat{\beta} = \mathbf{X} (\mathbf{X}^t \mathbf{C}^{-1} \mathbf{X})^{-1} \mathbf{X}^t \mathbf{C}^{-1} \mathbf{Y}, \quad (2)$$

$$\mathbf{SS}_{\text{Resid}} = (\mathbf{Y} - \hat{\mathbf{Y}})^t \mathbf{C}^{-1} (\mathbf{Y} - \hat{\mathbf{Y}}). \quad (3)$$

Next, the total sums of squares ( $\mathbf{SS}_{\text{Tot}}$ ) is obtained by repeating equations 1–3 on a design matrix  $\mathbf{X}$  containing only the column of ones. Variation explained by the regression model is then estimated as  $\mathbf{SS}_{\text{Tot}} - \mathbf{SS}_{\text{Resid}}$ , from which mean squares,  $F$ -ratios, and  $R^2$  may be obtained and evaluated. For more complicated statistical designs, multiple terms may be included in  $\mathbf{X}$ , and the statistical relationship between each  $\mathbf{X}$ -variable and  $\mathbf{Y}$  may be determined by repeating equations 1–3 above in a sequential fashion, each time removing a term from the model to produce a new reduced model  $\mathbf{X}_{\text{red}}$  (see Rencher and Christensen 2012). Finally, for multivariate  $\mathbf{Y}$  data, the analytical procedure described here is identical, but in this case sums-of-squares and cross-product matrices (SSCP) are obtained rather than univariate sums-of-squares (Rencher and Christensen 2012).

## *Phylogenetic Regression and the Problem of High-Dimensional Data*

Phylogenetic regression provides a flexible analytical tool for assessing the degree of evolutionary association between variables while accounting for phylogeny. Statistically, phylogenetic regression is a parametric method that summarizes information in

$\mathbf{Y}$  based on variances and covariances. As such it is considered an R-mode technique (sensu Legendre and Legendre 1998; see also Krzanowski 1993; Rencher and Christensen 2012). However, one complication with R-mode techniques is that for a given sample size, as the number of trait dimensions ( $p$ ) of the response data increases, the ability to detect patterns in the data decreases (Rao 1966), because the number of parameters to be estimated increases with the number of dependent variables. This phenomenon is known as Rao's paradox, and its effects are well-appreciated for standard multivariate R-mode methods such as Hotelling's  $T^2$  and MANOVA (Healy 1969; Olson 1974; Stevens 1980). Importantly, Rao's paradox should also apply to phylogenetic regression, because it uses the same algebraic machinery to obtain variances and parameter estimates as is implemented in regression and ANOVA (equations 1–3 above). Thus it is expected that the statistical power to detect patterns in a phylogenetic context using standard implementations of phylogenetic regression will decrease as the dimensionality of the dependent variables ( $\mathbf{Y}$ ) increases. This relationship is demonstrated empirically below. Furthermore, when the number of trait dimensions equals or exceeds the number of species in the phylogeny ( $p \geq N$ ), significance testing using standard approaches to phylogenetic regression cannot be completed, because the covariance matrices obtained from the model are singular, and thus parametric test statistics that require inverting a covariance matrix cannot be obtained.

Unfortunately, the implications of these statistical shortcomings have not been widely appreciated by comparative biologists, though they pose severe limitations on the assessment of multivariate data in a phylogenetic context. Specifically, because of Rao's paradox, phylogenetic assessments of phenotypic data will be compromised by the power of phylogenetic regression to detect patterns in these data, because their statistical power is a function of sample size ( $N$ ) relative to trait dimensionality ( $p$ ). Thus, it will generally be more difficult to detect evolutionary patterns in phenotypes that are characterized multivariately (like shape), than it will be to identify patterns in single, univariate traits such as body size. Further, in the extreme case where only a few species exist in a clade, evolutionary biologists will be limited to examining single traits, or must simplify their data through dimension reduction approaches (e.g., principal components) so that the statistical analyses themselves may be completed (see discussion in Monteiro 2013). The undesirable consequence is that methodological limitations are restricting the scope of phenotypic traits that are able to be examined with standard procedures. In essence, comparative biologists are unable to fully test hypotheses of adaptation in complex phenotypic traits because they are constrained by the analytical tools currently at their disposal for evaluating patterns in such data. Therefore, to assess evolutionary patterns in high-dimensional phenotypic traits using

a phylogenetic perspective, an alternative analytical framework is required.

In classical multivariate analysis, some of the challenges of analyzing high-dimensional data may be alleviated by using procedures based on the matrix of pairwise distances among specimens, rather than methods that use covariance matrices (Krzanowski 1993; Rencher and Christensen 2012). Such distance-based (Q-mode) methods provide a complementary (dual) view of multivariate data, as they summarize the same information represented by covariance matrices but do so in a different manner (Krzanowski 1993; Legendre and Legendre 1998). Further, for Euclidean datasets there exists a statistical equivalency between distance-based and covariance-based approaches for many procedures such that empirical results derived from both methods are numerically identical. This equivalency has been demonstrated analytically for ordination methods (principal components analysis versus principal coordinate analysis: Gower 1966), sums-of-squares from linear models (ANOVA and regression versus permutational-MANOVA: Anderson 2001; McArdle and Anderson 2001), methods for estimating rates of phenotypic evolution in a phylogenetic context (Adams 2014a), and methods for estimating phylogenetic signal (Adams 2014b). Additionally, distance-based techniques have a critical advantage over covariance-based approaches when used on high-dimensional data, such that even if the number of trait dimensions exceeds the number of observations ( $p \geq N$ ), statistical summaries from Q-mode methods may still be computed (see Anderson 2001; McArdle and Anderson 2001; Zapala and Schork 2006). Thus, for high-dimensional datasets, distance-based procedures provide a practical and useful means of assessing statistical patterns.

## A Q-Mode Phylogenetic Regression Method for High-Dimensional Data

Here I propose a distance-based (Q-mode) procedure for evaluating linear models for high-dimensional data in a phylogenetic context (hereafter: *D*-PGLS). The method assumes that trait variation accumulates over time following a Brownian motion model of evolution (Felsenstein 1973, 1988). Under Brownian motion, phenotypic changes are assumed to be independent from time step to time step, and variation increases proportionally with time ( $\sigma^2 t$ ). For multivariate data, each trait dimension may differ in the rate at which variation accumulates, and changes may also be correlated across trait dimensions (Felsenstein 1988, 2004). For the procedure developed here, both the phenotypic data ( $\mathbf{Y}$ ) and the design matrix ( $\mathbf{X}$ ) are transformed by the phylogeny, which are then used to obtain matrices that capture variation in  $\mathbf{Y}$  relative to  $\mathbf{X}$  as an among-specimen distance matrix. The sums of squares explained by the model is then estimated from the

distances among species in the phylogeny-transformed multivariate space, and factors in the multivariate model are evaluated statistically using permutation.

To implement the *D*-PGLS procedure, the phylogenetic transformation matrix  $\mathbf{E}$  is first obtained (Garland and Ives 2000). This matrix is found from the eigenvectors ( $\mathbf{U}$ ) and eigenvalues ( $\mathbf{W}$ ) of the phylogenetic covariance matrix ( $\mathbf{C}$ ) obtained from  $\mathbf{C} = \mathbf{U}\mathbf{W}\mathbf{U}^{-1}$ . From these, the phylogenetic transformation matrix is constructed as

$$\mathbf{E} = (\mathbf{U}\mathbf{W}^{1/2}\mathbf{U}^t)^{-1}. \quad (4)$$

Next, the independent variables,  $\mathbf{X}$  (including a column of ones), and the dependent variables,  $\mathbf{Y}$ , are projected onto  $\mathbf{E}$ , thereby transforming them by the phylogeny such that the resulting data no longer contain phylogenetic covariances:

$$\mathbf{X}_{phy} = \mathbf{E}\mathbf{X}, \quad (5)$$

$$\mathbf{Y}_{phy} = \mathbf{E}\mathbf{Y}. \quad (6)$$

Additionally, a  $N \times 1$  column vector of ones is transformed as  $\mathbf{1}_{phy} = \mathbf{E}\mathbf{1}$ . From the transformed data, predicted values ( $\hat{\mathbf{Y}}_X$ ) from the linear model:  $\mathbf{Y}_{phy} \sim \mathbf{X}_{phy}$  are obtained. Predicted values from  $\mathbf{Y}_{phy} \sim \mathbf{1}_{phy}$  are also estimated ( $\hat{\mathbf{Y}}_1$ ). Note that unlike standard linear models, the values in the rows of  $\hat{\mathbf{Y}}_1$  are not identical. This is because the phylogenetic relationships among taxa have been taken into account, thereby adjusting the expected value for each species with respect to its relationship to the other species in the phylogeny.

Variation explained by the model is then found from the trace of the outer-product of predicted values:

$$SS_X = tr \left( (\hat{\mathbf{Y}}_X - \hat{\mathbf{Y}}_1) (\hat{\mathbf{Y}}_X - \hat{\mathbf{Y}}_1)^t \right). \quad (7)$$

This  $N \times N$  outer-product matrix is known as the Gower-centered distance matrix, and represents the pairwise relationships among species in the high-dimensional data space (see Gower 1966; McArdle and Anderson 2001). If the design matrix  $\mathbf{X}$  contains more than one factor, predicted values are obtained from the terms of  $\mathbf{X}_{phy}$  in a sequential fashion to obtain the contributions of each factor to the model sums-of-squares (see Anderson 2001). Residual values are obtained from the full model ( $\mathbf{Y}_{phy} \sim \mathbf{X}_{phy}$ ) and used to estimate the  $SS_{resid}$ . Sums-of-squares are then used to obtain F-ratios and  $R^2$  values for all terms in the model. Next, the significance of each term in the model is assessed using permutation, in which the phenotypic data ( $\mathbf{Y}$ ) are permuted across the tips of the phylogeny, and the above procedure is repeated. Finally, to account for phylogenetic uncertainty the procedure described above can be repeated on an empirical distribution of alternative phylogenetic hypotheses to obtain a distribution of test

statistics, from which confidence intervals can be derived (see, e.g., de Villemereuil et al. 2012).

It is important to recognize that for univariate data, the *D-PGLS* procedure outlined in equations 4–7 yields statistical estimates that are numerically identical to those obtained from parametric implementations of phylogenetic regression. A demonstration of this property is found in Appendix A. Therefore, for univariate data, the distance—covariance equivalency between phylogenetic regression and *D-PGLS* has been preserved, since estimates of SS, F, and  $R^2$  for these models are the same. However, when used on multivariate data, *D-PGLS* is not restricted by the number of trait dimensions ( $p$ ), and in fact may be used even when the number of trait dimensions exceeds the number of species. Additionally, from a computational perspective the approach proposed here is more efficient than standard procedures, as calculations based on the among species distance matrix require considerably less effort than those based on the variable covariance matrix (whose size increases dramatically as the number of trait dimensions increases). This is particularly important for phylogenetic assessments of traits such as shape, where the number of trait dimensions frequently exceeds the number of taxa in a phylogeny (e.g., McPeck et al. 2008; Klingenberg and Gidaszewski 2010; Adams 2014a). As such, *D-PGLS* alleviates the analytical challenges found when using phylogenetic regression with high-dimensional data. Computer code written in R for implementing the approach is found in Appendix B, and may be found in the R library *geomorph* (Adams et al. 2014).

## Statistical Performance

To evaluate the statistical performance of *D-PGLS*, I executed a series of computer simulations. Simulations were designed to examine the statistical association between a multivariate response variable  $\mathbf{Y}$  and a single continuous variable  $\mathbf{X}$ . The multivariate data used in these simulations were generated under a Brownian motion model of evolution using two different patterns of error covariance: isotropic error and nonisotropic error. Initial simulations were conducted on a balanced phylogeny containing 32 species. Simulations under a wider set of conditions are found in the Supporting Information.

For each simulation, the number of trait dimensions for  $\mathbf{Y}$  was first selected ( $p = 2, 10, 15, 20, 30$ ). Next, input covariance matrices ( $\mathbf{S}$ ) of dimension  $(p+1) \times (p+1)$  were constructed, and used to generate the phenotypic data (the additional dimension was included to simulate the data for the  $\mathbf{X}$  variable). For simulations under an isotropic error model, the diagonal elements of  $\mathbf{S}$  were all identical ( $\sigma^2 = 1.0$ ). The off-diagonal elements of  $\mathbf{S}$  were then varied depending upon simulation conditions. Simulations evaluating type I error rates assumed no relationship between  $\mathbf{X}$  and

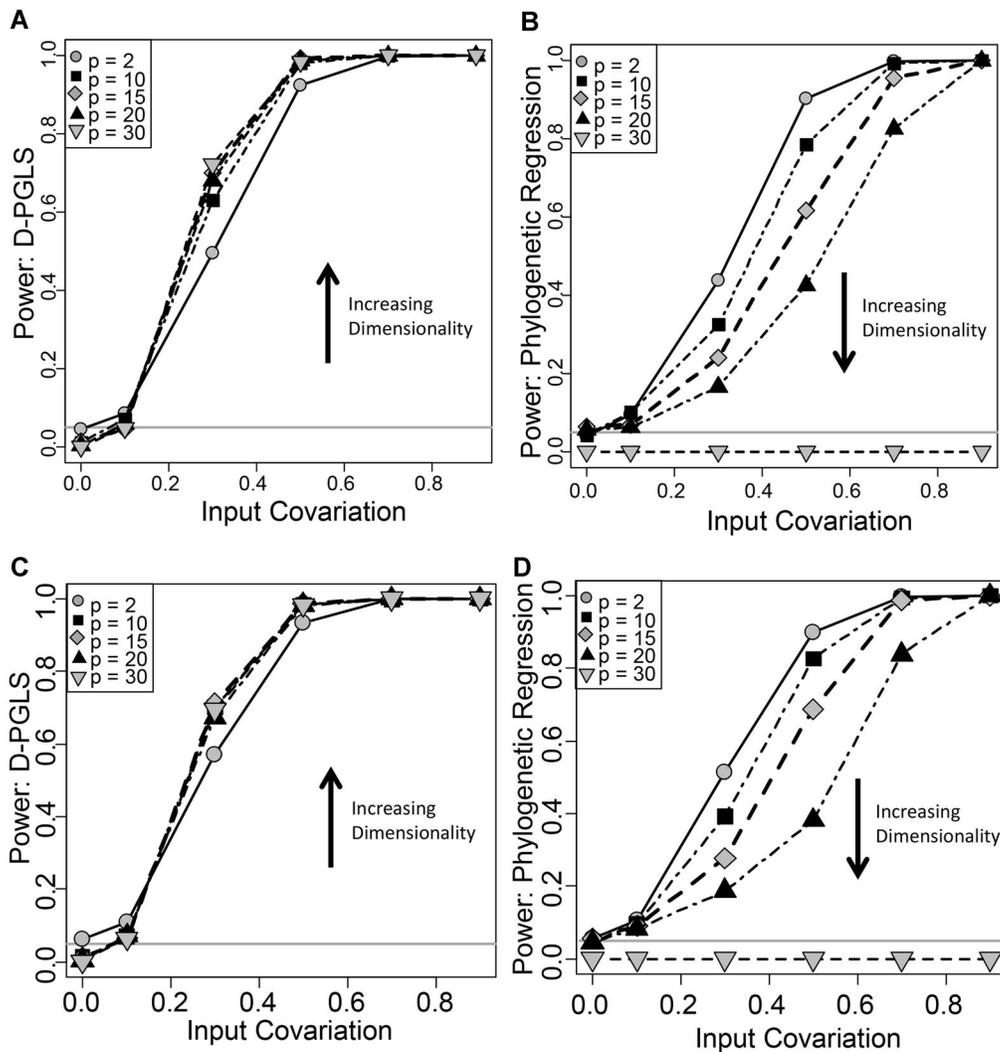
$\mathbf{Y}$ , and thus used no initial covariation between traits ( $\sigma_{1,2} = 0$ ). By contrast, simulations evaluating statistical power assumed a positive relationship between  $\mathbf{X}$  and  $\mathbf{Y}$ , and thus used positive initial levels of covariation between traits. The degree of covariation between  $\mathbf{X}$  and  $\mathbf{Y}$  varied depending upon the desired strength of the  $\mathbf{Y} \sim \mathbf{X}$  relationship ( $\sigma_{1,2} = 0.1, 0.3, 0.5, 0.7, 0.9$ ). For simulations under a model of nonisotropic error, the diagonal elements of  $\mathbf{S}$  were drawn from a normal distribution ( $\mu = 1; \text{std} = 0.02$ ) for each dimension, and the off diagonal elements of  $\mathbf{S}$  were drawn from a normal distribution whose mean followed those used in the isotropic simulations ( $\mu = \sigma_{1,2}; \text{std} = 0.02$ ). Values were drawn repeatedly from these distributions until  $\mathbf{S}$  satisfied all mathematical conditions of a covariance matrix.

From each initial covariance matrix  $\mathbf{S}$ , 1000 phenotypic datasets were obtained by evolving multidimensional traits along the phylogeny following a Brownian motion model of evolution. Then, the evolutionary relationship between the multivariate response variable ( $\mathbf{Y}$ ) and the dependent variable ( $\mathbf{X}$ ) for each dataset was statistically evaluated using *D-PGLS*, as well as with standard phylogenetic regression methods (e.g., PIC). The proportion of significant results (out of 1000) was then treated as an estimate of the Type I error (when  $\sigma_{1,2} = 0$ ) or statistical power (when  $\sigma_{1,2} > 0$ ) of both approaches.

Simulations were also performed across a wider set of conditions to evaluate the robustness of *D-PGLS*. These simulations examined the effect of the number of taxa in the phylogeny ( $N = 16, 32, 64, 128$ ), as well as the effect of randomly generated phylogenies on statistical performance ( $N = 16, 32, 64, 128$ ). Further, simulations were executed to investigate the effect of random noise on the statistical performance of *D-PGLS*. Here, some trait dimensions of the response variable ( $\mathbf{Y}$ ) were obtained using the  $\mathbf{Y} \sim \mathbf{X}$  relationship as described above, while other dimensions of the response variable contained random noise. Additional implementation details and results from all simulations are found in the Supporting Information.

## RESULTS

For all simulations, *D-PGLS* displayed slightly conservative Type I error rates, ranging between the nominal  $\alpha = 0.05$  and  $\alpha = 0.002$ . For both isotropic and nonisotropic conditions, Type I error became more conservative as traits of higher dimensionality were examined (Figs. 1 A, C). With respect to statistical power, the power of tests using *D-PGLS* increased rapidly as the degree of covariation between  $\mathbf{X}$  and  $\mathbf{Y}$  increased. Statistical power also increased as the number of trait dimensions ( $p$ ) increased (Figs. 1 A, C; Supporting Information). This finding runs counter to Rao's paradox, and implies that for the same number of taxa, *D-PGLS* is capable of detecting evolutionary trends more easily in high-dimensional phenotypic traits than in data represented by fewer dimensions. This pattern is due to the fact that as additional dimensions of



**Figure 1.** Simulation results evaluating the Type I error and statistical power of hypothesis testing procedures evaluating evolutionary covariation between a multivariate response variable ( $Y$ ) and a continuous predictor variable ( $X$ ) for the proposed approach  $D$ -PGLS (A, C) and standard phylogenetic regression (B, D). Data were simulated under a Brownian motion model of evolution on randomly generated phylogenies containing 32 species under an isotropic model of input covariance (panels A & B), and a nonisotropic model of input covariance (panels C & D). Curves for increasing numbers of trait dimensions are shown.

data are included the distances between specimens can only increase as the total dispersion among specimens increases. Thus, any covariation between the independent and dependent variables will be more easily detected, especially if the additional dimensions covary with  $X$ . A similar finding was obtained with  $Q$ -mode methods for comparing evolutionary rates for multivariate traits on phylogenies (Adams 2014a), and for estimating phylogenetic signal in multivariate data (Adams 2014b). Further, the power of tests using  $D$ -PGLS did not decrease as dimensions of random noise were added to the response variable  $Y$  (Supporting Information). Thus, unlike parametric approaches, the distance-based procedure is insensitive to the addition of variables containing unstructured variation. Together, these findings reveal that  $D$ -PGLS

is capable of detecting significant associations between variables, even when the response variable is highly dimensional.

In contrast to results using  $D$ -PGLS, and as expected, the parametric version of phylogenetic regression displayed decreasing power as trait dimensionality ( $p$ ) increased (Fig. 1 B, D). Further, for traits of high-dimensionality, the power of these approaches was 0.0, as the statistical significance from standard multivariate test statistics (e.g., Pillai's trace) was not able to be computed. Similar findings were revealed for the broader set of simulations examined in the Supporting Information. Overall these results demonstrate that for high-dimensional datasets, standard phylogenetic comparative approaches such as implementation with phylogenetic regression suffer from Rao's paradox,

implying that the ability of these methods to detect evolutionary associations are compromised as trait dimensionality increases. Further, in cases where  $p$  approaches  $N$ , statistical assessment of evolutionary patterns cannot be evaluated with phylogenetic regression, because the matrix computations for assessing significance with these approaches are singular. By contrast, the Q-mode approach developed here ( $D$ -PGLS) does not suffer from these shortcomings, and displays increasing power as trait dimensionality increases. As such,  $D$ -PGLS provides a useful means of detecting patterns of evolutionary association in a phylogenetic context for high-dimensional phenotypic datasets.

## A Biological Example

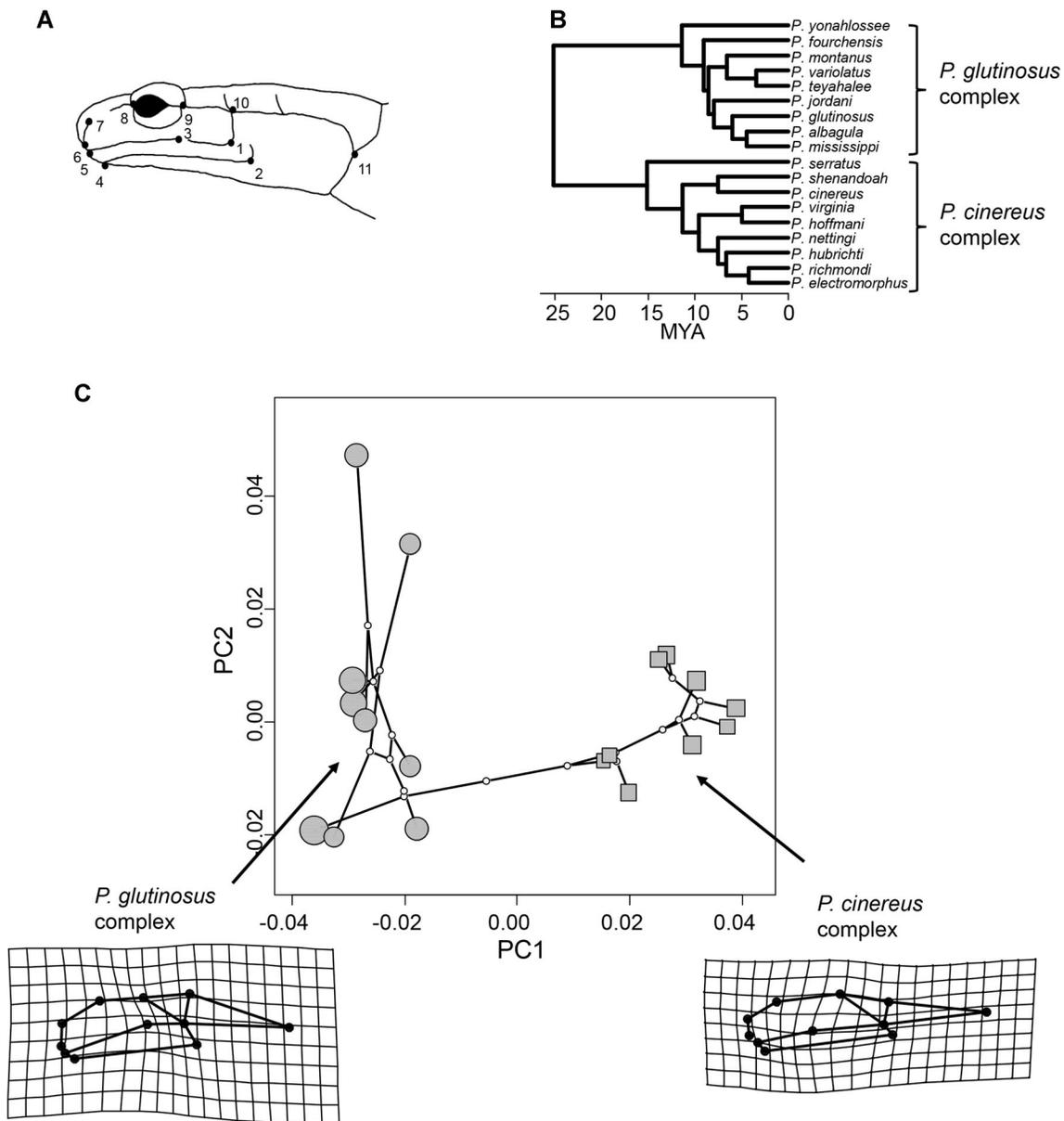
To illustrate the utility of the approach described above, I evaluated the degree of evolutionary association between body size and head shape (i.e., allometry) in a phylogenetic context in *Plethodon* salamanders. Allometry is the degree of covariation between size and shape (Gould 1966; Mosimann 1970), and thus describes the consequences of size changes on patterns of shape variation (Mitteroecker et al. 2013; Voje et al. 2014). Comparisons of allometric trajectories among species are frequently used to identify evolutionary changes in the size-shape relationship across taxa (e.g., Bookstein et al. 2003; Mitteroecker et al. 2004; Adams and Nistri 2010; Piras et al. 2010; Gunz 2012). Additionally, evolutionary allometry may be inferred by characterizing the relationship between size and shape among species in a phylogenetic context (e.g., Monteiro and Nogueira 2011; Cardini and Polly 2013; Klingenberg and Marugán-Lobón 2013; Outomuro et al. 2013b). In *Plethodon*, body size varies widely among species (Highton 1995; Adams and Church 2008, 2011), with species in some lineages displaying relatively small body sizes (e.g., *P. cinereus* species complex), while species in other lineages exhibit larger body sizes (e.g., *P. glutinosus* species complex). In many ecological communities, body size also appears to be a key trait in determining species coexistence in the group (Adams 2007). Similarly, head shape varies within and among species, and considerable evidence suggests that competitive interactions among species may drive morphological evolution (e.g., Adams 2010). Head shape also displays a strong genetic component (Adams 2011), thereby enabling microevolution via selection. Therefore, because both head shape and body size vary widely in the genus, it is of interest to determine whether the two traits coevolve among species allometrically.

To evaluate patterns of evolutionary allometry in *Plethodon* I used head shape data from 691 adult salamanders from 18 species (data from Maerz et al. 2006; Adams et al. 2007; Arif et al. 2007; Myers and Adams 2008; Adams 2010; Deitloff et al. 2013). For all individuals, 11 landmarks were digitized from left-lateral images

of the side of each head (Fig. 2A), and geometric morphometric methods were used to generate a set of variables representing head shape (Bookstein 1991; Adams et al. 2013). Specifically, variation in the jaw position relative to the skull was mathematically standardized among specimens (Adams 1999), and a generalized Procrustes analysis was then performed to align specimens to a common coordinate system and remove variation in their position, orientation, and size (Rohlf and Slice 1990). This resulted in a set of shape variables (Procrustes tangent coordinates) for all specimens, from which the mean head shape was estimated for all 18 species. Typical adult body sizes for all species were obtained from Adams and Church (2008, 2011).

With these data, I examined trends of evolutionary allometry by assessing the multivariate regression between head shape and body size in a phylogenetic context using  $D$ -PGLS. A time-calibrated multigene molecular phylogeny for the genus (Wiens et al. 2006; Fig. 2B) was pruned to match the species in the dataset, and provided an estimate of the evolutionary relationships among taxa. For comparison I also performed a nonphylogenetic evaluation of evolutionary allometry using standard regression based on Procrustes distances (sensu Goodall 1991). In addition, because the species studied here were members of two clades within *Plethodon*, I compared head shape and body size between them using Procrustes ANOVA and ANOVA, respectively. Finally, phylogenetic patterns of head shape evolution were visualized using a phylomorphospace approach (sensu Sidlauskas 2008), and thin-plate spline deformation grids of representative specimens were generated to facilitate description of the observed shape changes. All analyses were performed in R 3.0.2 (R Development Core Team 2014) using routines in the library geomorph (Adams and Otárola-Castillo 2013; Adams et al. 2014).

**Results:** Standard multivariate regression revealed a significant relationship between head shape and body size ( $F = 9.481$ ;  $P = 0.001$ ;  $R^2 = 0.372$ ), suggesting that evolutionary allometry was present in *Plethodon*. However, when this relationship was examined in a phylogenetic context, the association between head shape and body size was no longer significant, with body size explaining little of the variation in head shape ( $F = 1.885$ ;  $P = 0.188$ ;  $R^2 = 0.105$ ). Thus, in contrast to the ahistorical analysis, the phylogenetic analysis using  $D$ -PGLS strongly implied that evolutionary allometry was not present in the group. This apparent discrepancy between the approaches was resolved by examining patterns of head shape variation in phylomorphospace (Fig. 2C). Here it was found that the species examined in this study formed two distinct clusters, which perfectly corresponded with two monophyletic lineages (species complexes) within the genus. Specifically, species in the *P. glutinosus* complex were found to display heads that were stouter and relatively more compressed in the antero-posterior direction (Fig. 2C), whereas species in the *P. cinereus* complex exhibited heads that were more elongated



**Figure 2.** (A) Positions of 11 anatomical landmarks used to quantify head shape in *Plethodon* salamanders (image from [Adams et al. 2007]). (B) Fossil-calibrated molecular phylogeny displaying the estimated phylogenetic relationships among the species of *Plethodon* examined here. (C) Plot of phylomorphospace viewed as the first two principal component axes of tangent space. Thin-plate spline deformation grids representing exemplar individuals from each species complex are shown (magnified by 3 $\times$ ).

anterio-posteriorly, and were slightly compressed in the dorsal-ventral direction. Further, both lineages also differed in body size, where species in the *P. glutinosus* clade were larger than species in the *P. cinereus* clade (Fig. 2C). Indeed, statistical analyses confirmed that the two clades differed significantly in both head shape ( $F = 16.384$ ;  $P = 0.001$ ;  $R^2 = 0.506$ ) and body size ( $F = 35.612$ ;  $P < 0.0001$ ;  $R^2 = 0.69$ ), suggesting that patterns of variation in these phenotypic traits was distinct between lineages. Finally, a comparison of  $R^2$  values between phylogenetic regression models revealed that considerably more variation in head shape

was explained by a model describing lineage-specific differences ( $R^2 = 0.506$ ) than by a model describing evolutionary allometry ( $R^2 = 0.105$ ), implying that clade-level differences provided a better overall explanation of the observed trends in shape variation when taking phylogeny into account.

In conclusion, when all patterns were taken into consideration, it was clear that the apparent evolutionary allometry observed in *Plethodon* was the result of two lineages within the genus evolving differences in both body size and head shape, as opposed to the alternative explanation in which more subtle changes

in both traits cooccurred from species to species. Thus, this example highlights the importance of incorporating phylogeny directly into cross-species evaluations of evolutionary allometry. Without a phylogenetic perspective, the hypothesized mechanism responsible for patterns of size-shape covariation would have been misidentified, and incorrectly attributed to allometric trends related to size. However, when phylogenetic nonindependence was included in the multivariate regression, this explanation became untenable. Instead, the observed patterns appear due to the fact that two sublineages exist within the group, and these lineages have evolved differences in both body size and head shape. As such, it is lineage-specific differences and not evolutionary allometric trends that best explain patterns in head shape variation at the macroevolutionary level. Only when phylogeny was taken into consideration could the correct pattern be revealed (see also Garland et al. 1992).

## Discussion

A critical component in the study of adaptation is to identify evolutionary correlations between phenotypic traits in species related by a phylogeny. To evaluate such hypotheses, phylogenetic regression is typically used, but this method is incapable of assessing evolutionary patterns in highly multivariate datasets like shape, as the large number of parameters to be estimated prohibits the analyses from being analytically completed. In this article, I developed a distance-based procedure (*D*-PGLS) for evaluating ANOVA and regression models for high-dimensional datasets in a phylogenetic framework under a Brownian motion model of evolution. I showed that when used on univariate data, the approach yields statistical estimates that are numerically identical to those obtained from phylogenetic regression, thereby demonstrating the equivalency between phylogenetic regression and *D*-PGLS for such datasets. I also illustrated that the approach exhibits appropriate type I error and high statistical power for detecting evolutionary patterns in high-dimensional datasets. By contrast, I found that standard implementations of phylogenetic regression have decreasing power as trait dimensionality increases. Together, these findings have several important implications for the evaluation of evolutionary patterns in high-dimensional data and how they may be identified phylogenetically.

First, the results presented here imply that parametric implementations of phylogenetic comparative approaches, such as independent contrasts and PGLS, may not be the most appropriate procedures for assessing patterns in high-dimensional phenotypic datasets, because the ability of these procedures for detecting patterns in multivariate data are analytically compromised by trait dimensionality. Recently, it was suggested that comparative analyses of multivariate data like shape may be accomplished simply by converting the data to their phylogenetic independent

contrasts and using the appropriate statistical procedure to assess the given hypothesis (see Klingenberg and Marugán-Lobón 2013). Although this recommendation may be acceptable for large datasets where the number of species examined greatly outpaces the dimensionality of the phenotypic data (see e.g., Blankers et al. 2012; Klingenberg and Marugán-Lobón 2013), the results presented here show that employing parametric implementations of the standard phylogenetic comparative toolkit ubiquitously to all datasets (including high-dimensional data), may not be sufficient, as these methods lose statistical power as trait dimensionality increases. Thus, as comparative biologists quantify phenotypic data by more complex and comprehensive methods, standard phylogenetic regression approaches are not guaranteed to identify patterns that may be present (see Fig. 1). Further, when the number of trait dimensions equals or exceeds the number of species in the phylogeny, the parametric approach to phylogenetic regression has zero statistical power as its algebra cannot be completed, thereby limiting the datasets that can be examined with this method. Indeed, the biological example presented here represents such a scenario, as the number of phenotypic variables describing salamander head shape exceeded the number of species used in the study. By contrast, evaluating evolutionary patterns in such data via the distance-based approach developed here has no such data restrictions. Therefore, as comparative biologists continue to characterize phenotypic attributes in increasingly more complex ways (from univariate measures to sets of a few traits to highly multivariate phenotypes), so too must they alter their analytical procedures; moving from standard implementations of phylogenetic regression that evaluate multivariate test statistics such as Wilks' lambda and Pillai's trace to the distance-based procedure proposed here (*D*-PGLS).

Second, the findings presented here demonstrate that the comparative biologist need not reduce the dimensionality of the phenotypic data to a few principal component axes to perform phylogenetic assessments of evolutionary correlations (e.g., Bergmann et al. 2009; Nogueira et al. 2009; Monteiro and Nogueira 2011; Brusatte et al. 2012). Indeed, the method developed here exhibits higher statistical power as trait dimensionality increases (Fig. 1), implying that more data can be leveraged in a meaningful way to evaluate evolutionary patterns. This empirical finding is in direct contrast to the suggestion that dimension reduction is a required step for evaluating patterns in high-dimensional data like shape when using phylogenetic comparative methods (Monteiro 2013). While this remains the case for covariance-based procedures that require the estimation of many parameters, by moving from the parametric implementations of PGLS to the distance-based *D*-PGLS procedure there is no need to simplify one's phenotypic data to force it into the standard analytical paradigm (for a related conclusion on dimension reduction see Bookstein 2013). Thus, when viewed in this manner,

*D*-PGLS provides a useful complement to existing phylogenetic comparative approaches, that enables the evaluation of hypotheses of adaptation and phenotypic change for high-dimensional phenotypic data in a manner analogous to that provided by PIC and PGLS for univariate traits.

One current limitation with the approach developed here is that the method is only formulated under a Brownian motion model of evolution. Thus, evolutionary patterns derived from other evolutionary processes, such as an Ornstein–Uhlenbeck model (OU: Hansen and Martins 1996; Butler and King 2004) cannot be investigated. Recently, multivariate extensions of the OU models have been developed for assessing phenotypic trends in a phylogenetic context (Bartoszek et al. 2012; see also Butler and King 2004). However, like phylogenetic regression, this method requires the estimation of a large number of parameters that increases greatly with trait dimensionality (see discussion in Monteiro 2013). As such, the multivariate OU model will suffer the same limitations when used on highly multivariate data as does standard phylogenetic regression, and thus in practice it will be limited in utility to those scenarios where one evaluates a few number of traits across a phylogeny containing many species. Future work is required to derive an OU equivalent of *D*-PGLS to fill this void. With such an approach, evolutionary biologists could then compare the evolution of highly multivariate phenotypes as described by both Brownian motion and Ornstein–Uhlenbeck processes, providing the multivariate equivalent of evolutionary model selection approaches currently used for univariate traits (e.g., Butler and King 2004; Collar et al. 2010; Harmon et al. 2010; Beaulieu et al. 2012).

Despite this, the distance-based phylogenetic regression procedure proposed here (*D*-PGLS) provides an important quantitative merging of methods that characterize phenotypes using high-dimensional data (morphometrics) with those that evaluate interspecific patterns in a phylogenetic context. Recent years have seen an increased attention on the empirical intersection between morphometrics and phylogenetic comparative biology, and how data from the former may be used in the analytics of the latter (e.g., Adams et al. 2009; Klingenberg and Gidaszewski 2010; Adams et al. 2011; Klingenberg and Marugán-Lobón 2013; Monteiro 2013; Polly et al. 2013; Adams and Felice 2014; Adams 2014a,b). While in some cases the standard tools of the phylogenetic comparative toolkit may be sufficient for evaluating patterns in morphometric data (e.g., Sidlauskas 2008; Monteiro and Nogueira 2010; Klingenberg and Marugán-Lobón 2013; Piras et al. 2013), in other instances the characteristics of this high-dimensional data type dictate that alternative analytical procedures are sometimes required (see Klingenberg and Gidaszewski 2010; Adams 2014a,b). The analysis of evolutionary correlations between high-dimensional phenotypic data and other factors represents one such

case, as the number of parameters that require estimation using standard approaches quickly becomes prohibitive. However, by deriving a Q-mode alternative that leverages the statistical equivalency between covariance-based and distance-based approaches, comparative biologists may circumvent this conundrum, and may evaluate evolutionary trends in high-dimensional data while accounting for phylogenetic nonindependence.

## ACKNOWLEDGMENTS

I thank M. Collyer, G. Hunt, A. Kaliontzopoulou, and E. Sherratt for comments and discussion. This work was sponsored in part by NSF grants DEB-1257827 and DEB-111884.

## DATA ARCHIVING

The doi for my data is 10.5061/dryad.36df0.

## LITERATURE CITED

- Ackerly, D. D., and M. J. Donoghue. 1998. Leaf size, sapling allometry, and Corner's rules: phylogeny and correlated evolution in maples (*Acer*). *Am. Nat.* 152:767–791.
- Adams, D. C. 1999. Methods for shape analysis of landmark data from articulated structures. *Evol. Ecol. Res.* 1:959–970.
- . 2007. Organization of *Plethodon* salamander communities: guild-based community assembly. *Ecology* 88:1292–1299.
- . 2010. Parallel evolution of character displacement driven by competitive selection in terrestrial salamanders. *BMC Evol. Biol.* 10:1–10.
- . 2011. Quantitative genetics and evolution of head shape in *Plethodon* salamanders. *Evol. Biol.* 38:278–286.
- . 2014a. Quantifying and comparing phylogenetic evolutionary rates for shape and other high-dimensional phenotypic data. *Syst. Biol.* 63:166–177.
- . 2014b. A generalized K statistic for estimating phylogenetic signal from shape and other high-dimensional multivariate data. *Syst. Biol.* 63. doi: 10.1093/sysbio/syu030.
- Adams, D. C., and J. O. Church. 2008. Amphibians do not follow Bergmann's rule. *Evolution* 62:413–420.
- . 2011. The evolution of large-scale body size clines in *Plethodon*: evidence of heat-balance or species-specific artifact? *Ecography* 34:1067–1075.
- Adams, D. C., and R. Felice. 2014. Assessing phylogenetic morphological integration and trait covariation in morphometric data using evolutionary covariance matrices. *PLoS ONE* 9:e94335.
- Adams, D. C., and A. Nistri. 2010. Ontogenetic convergence and evolution of foot morphology in European cave salamanders (Family: Plethodontidae). *BMC Evol. Biol.* 10:1–10.
- Adams, D. C., and E. Otárola-Castillo. 2013. geomorph: an R package for the collection and analysis of geometric morphometric shape data. *Methods Ecol. Evol.* 4:393–399.
- Adams, D. C., M. E. West, and M. L. Collyer. 2007. Location-specific sympatric morphological divergence as a possible response to species interactions in West Virginia *Plethodon* salamander communities. *J. Anim. Ecol.* 76:289–295.
- Adams, D. C., C. M. Berns, K. H. Kozak, and J. J. Wiens. 2009. Are rates of species diversification correlated with rates of morphological evolution? *Proc. Royal Soc. B* 276:2729–2738.
- Adams, D. C., A. Cardini, L. R. Monteiro, P. O'Higgins, and F. J. Rohlf. 2011. Morphometrics and phylogenetics: principal components of shape from



- Maerz, J. C., E. M. Myers, and D. C. Adams. 2006. Trophic polymorphism in a terrestrial salamander. *Evol. Ecol. Res.* 8:23–35.
- Martins, E. P., and T. F. Hansen. 1997. Phylogenies and the comparative method: a general approach to incorporating phylogenetic information into the analysis of interspecific data. *Am. Nat.* 149:646–667.
- McArdle, B. H., and M. J. Anderson. 2001. Fitting multivariate models to community data: a comment on distance-based redundancy analysis. *Ecology* 82:290–297.
- McPeck, M. A., L. Shen, J. Z. Torrey, and H. Farid. 2008. The tempo and mode of three-dimensional morphological evolution in male reproductive structures. *Am. Nat.* 171:E158–E178.
- Mitteroecker, P., P. Gunz, M. Bernhard, K. Schaefer, and F. L. Bookstein. 2004. Comparison of cranial ontogenetic trajectories among great apes and humans. *J. Hum. Evol.* 46:679–698.
- Mitteroecker, P., P. Gunz, S. Windhager, and K. Schaefer. 2013. A brief review of shape, form, and allometry in geometric morphometrics, with applications to human facial morphology. *Hystrix* 24:59–66.
- Monteiro, L. R. 2013. Morphometrics and the comparative method: studying the evolution of biological shape. *Hystrix* 24:25–32.
- Monteiro, L. R., and M. R. Nogueira. 2010. Adaptive radiations, ecological specialization, and the evolutionary integration of complex morphological structures. *Evolution* 64:724–744.
- . 2011. Evolutionary patterns and processes in the radiation of phyllostomid bats. *BMC Evol. Biol.* 11:1–23.
- Mosimann, J. E. 1970. Size allometry: Size and shape variables with characterizations of the lognormal and generalized gamma distributions. *J. Am. Stat. Assoc.* 65:930–945.
- Myers, E. M., and D. C. Adams. 2008. Morphology is decoupled from interspecific competition in *Plethodon* salamanders in the Shenandoah Mountains. *Herpetologica* 64:281–289.
- Nogueira, M. R., A. L. Peracchi, and L. R. Monteiro. 2009. Morphological correlates of bite force and diet in the skull and mandible of phyllostomid bats. *Func. Ecol.* 23:715–723.
- Olson, C. L. 1974. Comparative robustness of six tests in multivariate analysis of variance. *J. Am. Stat. Assoc.* 69:894–908.
- Outomuro, D., D. C. Adams, and F. Johansson. 2013a. Evolution of wing shape in ornamented-winged damselflies. *Evol. Biol.* 40:300–309.
- . 2013b. Wing shape allometry and aerodynamics in calopterygid damselflies: a comparative approach. *BMC Evol. Biol.* 13:1–11.
- Pennell, M. W., and L. J. Harmon. 2013. An integrative view of phylogenetic comparative methods: connections to population genetics, community ecology, and paleobiology. *Ann. N. Y. Acad. Sci.* 1289:90–105.
- Piras, P., P. Colangelo, D. C. Adams, A. Buscalioni, J. Cubo, T. Kotsakis, C. Meloro, and P. Raia. 2010. The Gavialis-Tomistoma debate: the contribution of skull ontogenetic allometry and growth trajectories to the study of crocodylian relationships. *Evol. Develop.* 12:568–579.
- Piras, P., L. Maiorino, L. Teresi, C. Meloro, F. Lucci, T. Kotsakis, and P. Raia. 2013. Bite of the cats: relationships between functional integration and mechanical performance as revealed by mandible geometry. *Syst. Biol.* 62:878–900.
- Polly, P. D., A. M. Lawing, A.-C. Fabre, and A. Goswami. 2013. Phylogenetic principal components analysis and geometric morphometrics. *Hystrix* 24:1–9.
- R Development Core Team. 2014. R: a language and environment for statistical computing. Version 3.0.2. R Foundation for Statistical Computing, Vienna, Austria. Available at <http://cran.R-project.org>.
- Rao, C. R. 1966. Covariance adjustment and related problems in multivariate analysis. Pp. 87–103 in P. R. Krishnaiah, ed. *Multivariate analysis*. Academic Press, New York.
- Rencher, A. C., and W. F. Christensen. 2012. *Methods of multivariate analysis*. John Wiley & Sons, Hoboken, NJ.
- Revell, L. J. 2010. Phylogenetic signal and linear regression on species data. *Methods Ecol. Evol.* 1:319–329.
- Revell, L. J., L. J. Harmon, and D. C. Collar. 2008. Phylogenetic signal, evolutionary process, and rate. *Syst. Biol.* 57:591–601.
- Rohlf, F. J. 2001. Comparative methods for the analysis of continuous variables: geometric interpretations. *Evolution* 55:2143–2160.
- Rohlf, F. J., and D. E. Slice. 1990. Extensions of the Procrustes method for the optimal superimposition of landmarks. *Syst. Zool.* 39:40–59.
- Rüber, L., and D. C. Adams. 2001. Evolutionary convergence of body shape and trophic morphology in cichlids from Lake Tanganyika. *J. Evol. Biol.* 14:325–332.
- Sessions, S. K., and A. Larson. 1987. Developmental correlates of genome size in plethodontid salamanders and their implications for genome evolution. *Evolution* 41:1239–1251.
- Sidlauskas, B. 2008. Continuous and arrested morphological diversification in sister clades of characiform fishes: a phylomorphospace approach. *Evolution* 62:3135–3156.
- Stevens, J. P. 1980. Power of the multivariate analysis of variance tests. *Psych. Bull.* 88:728–737.
- Voje, K. L., T. F. Hansen, C. K. Egset, G. H. Bolstad, and C. Pelabon. 2014. Allometric constraints and the evolution of allometry. *Evolution* 68:866–885.
- Wiens, J. J., T. N. Engstrom, and P. T. Chippendale. 2006. Rapid diversification, incomplete isolation, and the ‘speciation clock’ in North American salamanders (genus: *Plethodon*): testing the hybrid swarm hypothesis of rapid radiation. *Evolution* 60:2585–2603.
- Zapala, M. A., and N. J. Schork. 2006. Multivariate regression analysis of distance matrices for testing associations between gene expression patterns and relative variables. *Proc. Natl. Acad. Sci. USA* 103:19430–19435.

Associate Editor: P. David Polly

## Appendix A:

### Worked Example Demonstrating the Equivalency of Phylogenetic Regression and D-PGLS for Univariate Data

In this example, five hypothetical species are related by the following phylogenetic tree and have the following phenotypic values for **X** and **Y**:

For these data, the phylogenetic independent contrasts are found as:

$$PIC_y = \begin{bmatrix} 0.2294 \\ 0.5773 \\ -1.4042 \\ 1.4142 \end{bmatrix} \quad PIC_x = \begin{bmatrix} -1.4708 \\ -1.1547 \\ -0.7071 \\ -0.7071 \end{bmatrix}$$

A linear regression of  $pic_y$  on  $pic_x$  yields the following statistical summary values:  $SS = 0.2369$ ;  $F = 0.1713$ ;  $R^2 = 0.054$ , which are identical to those found using phylogenetic generalized least squares (PGLS).

To estimate this relationship via distances, the phylogenetic covariance matrix (**C**) representing these evolutionary

relationships under Brownian motion is first obtained as:

$$\mathbf{C} = \begin{matrix} & \begin{matrix} A & B & C & D & E \end{matrix} \\ \begin{matrix} A \\ B \\ C \\ D \\ E \end{matrix} & \begin{bmatrix} 3 & 0 & 0 & 0 & 0 \\ 0 & 3 & 2 & 1 & 1 \\ 0 & 2 & 3 & 1 & 1 \\ 0 & 1 & 1 & 3 & 2 \\ 0 & 1 & 1 & 2 & 3 \end{bmatrix} \end{matrix} \quad (8)$$

Next, the phylogenetic transformation matrix (Garland and Ives 2000) is obtained through the components of an eigen-decomposition of the phylogenetic covariance matrix:  $\mathbf{C} = \mathbf{U}\mathbf{W}\mathbf{U}^{-1}$ . For this example,  $\mathbf{U}$  and  $\mathbf{W}$  are:

$$\mathbf{U} = \begin{bmatrix} 0.0 & 0.0 & 1 & 0.0000 & 0.0000 \\ -0.5 & 0.5 & 0 & 0.0000 & -0.7071 \\ -0.5 & 0.5 & 0 & 0.0000 & 0.7071 \\ -0.5 & -0.5 & 0 & -0.7071 & 0.0000 \\ -0.5 & -0.5 & 0 & -0.7071 & 0.0000 \end{bmatrix}$$

$$\mathbf{W} = \begin{bmatrix} 7 & 0 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 & 0 \\ 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \quad (9)$$

The eigenvectors and eigenvalues of  $\mathbf{C}$  are then used to construct the phylogenetic transform matrix, as  $\mathbf{E} = (\mathbf{U}\mathbf{W}^{1/2}\mathbf{U}^t)^{-1}$ :

$$\mathbf{E} = \begin{bmatrix} 0.5773 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.7388 & -0.2612 & -0.0498 & -0.0498 \\ 0.0000 & -0.2612 & 0.7388 & -0.0498 & -0.0498 \\ 0.0000 & -0.0498 & -0.0498 & 0.7388 & -0.2612 \\ 0.0000 & -0.0498 & -0.0498 & -0.2612 & 0.7388 \end{bmatrix}$$

Using this matrix, both  $\mathbf{Y}$  and  $\mathbf{X}$  (with a column of ones to estimate the intercept) are transformed by the phylogeny:

$$\mathbf{Y}_{transf} = \begin{bmatrix} 2.3094 \\ 0.6116 \\ 2.6116 \\ 2.0342 \\ 0.0342 \end{bmatrix} \quad \mathbf{X}_{transf} = \begin{bmatrix} 0.5775 & 0.5773 \\ 0.3779 & 0.2455 \\ 0.3779 & 1.2455 \\ 0.3779 & 1.4002 \\ 0.3779 & 2.4002 \end{bmatrix}$$

Predicted values are then obtained from the columns of  $\mathbf{X}_{trans}$  in a sequential manner ( $X_1, X_{1\&2}, X_{1-3}$ , etc.), using a linear model ( $\mathbf{Y} \sim \mathbf{X}$ ):

$$\mathbf{Y}_{pred.1} = \begin{bmatrix} 2.1271 \\ 1.3925 \\ 1.3925 \\ 1.3925 \\ 1.3925 \end{bmatrix} \quad \mathbf{Y}_{pred.x} = \begin{bmatrix} 2.3594 \\ 1.5783 \\ 1.3235 \\ 1.2841 \\ 1.029 \end{bmatrix}$$

Variation explained by the model ( $SS_X$ ) is found from the trace of the outer product of predicted values (see text for explanation):

$$SS_X = tr(\widehat{\mathbf{Y}}\widehat{\mathbf{Y}}^t) = 0.2369$$

Finally, residual values are estimated from the model in a similar manner, from which  $SS_{res}$  is obtained. From  $SS_X$  and  $SS_{res}$  one obtains  $F = 0.1713$  and  $R^2 = 0.054$ ; which are identical to the values obtained using phylogenetic regression as shown above.

### Appendix B:

#### Computer Code for R

```

#The function below performs distance-based phylogenetic least-
squares analysis. The method may
#be used to assess regression, ANOVA, and other linear mod-
els in a phylogenetic context.
#The approach is particularly useful for high-dimensional
data where standard (parametric)
#phylogenetic generalized least squares analyses, or the anal-
ysis of phylogenetically
#independent contrasts, cannot be performed. The function
below obtains the sums of squares
#(SS) for all factors in the linear model, and statistically
evaluates them via permutation.
D.pgls<-function(f1,phy,iter = 999){
  library(ape)
  data = NULL
  form.in<-formula(f1)
  Terms<-terms(form.in,keep.order = TRUE)
  Y<-as.matrix(eval(form.in[[2]],parent.frame()))
  N<-length(phy$tip.label)
  p<-ncol(Y)
  if(is.null(rownames(Y))){
    stop("No species names with Y-data.")}
  if(length(match(rownames(Y), phy$tip.label))!= N)
    stop("Data matrix missing some taxa present on the tree.")
  if(length(match(phy$tip.label,rownames(Y)))!= N)

```

```

stop("Tree missing some taxa in the data matrix.")
C<-vcv.phylo(phy); C<-C[rownames(Y),rownames(Y)]
eigC <- eigen(C)
D.mat<-solve(eigC$eigenvalues)
diag(sqrt(eigC$eigenvalues))%*% t(eigC$eigenvalues)
Y.new<-D.mat%*% (Y)
ones.new<-D.mat%*%(array(1,N))
pred.l <- predict(lm(Y.new~ones.new-1))
dat<-model.frame(form.in,data)
df<-df.tmp<-SS.tmp<-SS.obs<-F<-array()
for (i in 1:ncol(attr(Terms, "factors"))){
  mod.mat<-model.matrix(Terms[1:i],data = dat)
  x.new<-D.mat%*%mod.mat
  pred.y<-predict(lm(Y.new~x.new-1))
  G<-(pred.y-pred.l)%*%t(pred.y-pred.l)
  SS.tmp[i]<-sum(diag(G))
  ifelse(i == 1, SS.obs[i]<-SS.tmp[i], SS.obs[i]<-
(SS.tmp[i]-SS.tmp[i-1]))
  df.tmp[i]<-ifelse(ncol(mod.mat) == 1,1,(ncol(mod.mat)-
1))
  ifelse(i == 1, df[i]<-df.tmp[i], df[i]<-(df.tmp[i]-df.tmp[i-
1]))
}
MS<-SS.obs/df
mod.mat<-model.matrix(Terms)
x.new<-D.mat%*%mod.mat
y.res<-residuals(lm(Y.new~x.new-1))
SS.r.res<-sum(diag(y.res%*%t(y.res)))
df.res<-nrow(Y)-1-sum(df)
MS.res<-SS.res/df.res
Rsqr<-SS.obs/(sum(SS.obs)+SS.res)
F<-MS/MS.res

F.r<-P.val<-array(1,dim = length(SS.obs))
for(i in 1:iter){
  SS.tmp<-SS.r<-array()
  Y.r<-as.matrix(Y[sample(nrow(Y)),i])
  row.names(Y.r)<-row.names(Y)
  Y.r.new<-D.mat%*% (Y.r)
  pred.l.r<- predict(lm(Y.r.new~ones.new-1))
  for (ii in 1:ncol(attr(Terms, "factors"))){
    mod.mat<-model.matrix(Terms[1:ii])
    x.new<-D.mat%*%mod.mat
    pred.y.r<-predict(lm(Y.r.new~x.new-1))
    G.r<-(pred.y.r-pred.l.r)%*%t(pred.y.r-pred.l.r)
    SS.tmp[ii]<-sum(diag(G.r))
    ifelse(ii == 1, SS.r[ii]<-SS.tmp[ii], SS.r[ii]<-(SS.tmp[ii]-
SS.tmp[ii-1]))
  }
  MS.r<-SS.r/df
  mod.mat<-model.matrix(Terms)
  x.new<-D.mat%*%mod.mat
  y.res.r<-residuals(lm(Y.r.new~x.new-1))
  SS.r.res<-sum(diag(y.res.r%*%t(y.res.r)))
  MS.r.res<-SS.r.res/df.res
  F.r<-MS.r/MS.r.res
  P.val<-ifelse(F.r>= F, P.val+1,P.val)
}
P.val<-P.val/(iter+1)
anova.tab<-cbind(df,SS.obs,MS,F,P.val,Rsq)
anova.tab<-rbind(anova.tab,c(df.res,SS.res,MS.res,NA,
NA,NA))
rownames(anova.tab)<-c(colnames(attr(Terms, "factors")),
"Residual")
return(anova.tab)
}

```

## Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's website:

**Figure S1.** Statistical power curves for tests evaluating evolutionary covariation using *D*-PGLS when data are simulated on balanced phylogenies using isotropic error.

**Figure S2.** Statistical power curves for tests evaluating evolutionary covariation using *D*-PGLS when data are simulated on balanced phylogenies using nonisotropic error.

**Figure S3.** Statistical power curves for tests evaluating evolutionary covariation using *D*-PGLS when data are simulated on random phylogenies using isotropic error.

**Figure S4.** Statistical power curves for tests evaluating evolutionary covariation using *D*-PGLS when data are simulated on random phylogenies using nonisotropic error.

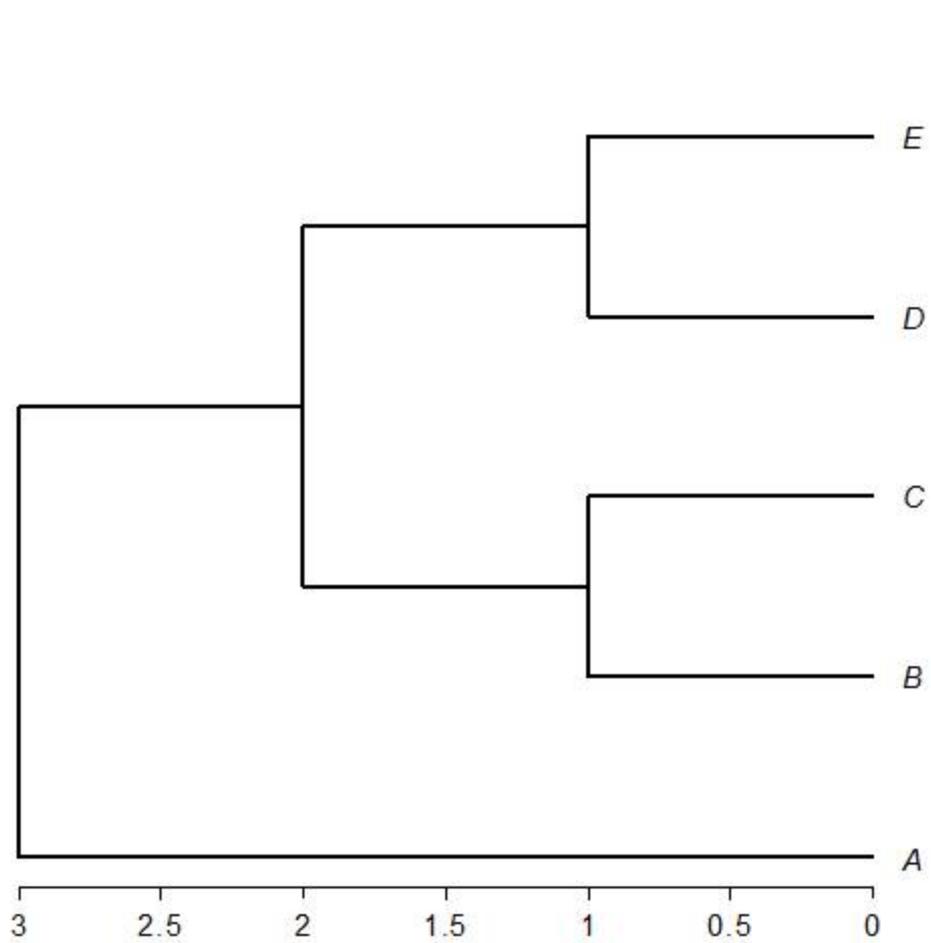
**Figure S5.** Statistical power curves for tests evaluating evolutionary covariation using phylogenetic regression when data are simulated on balanced phylogenies using isotropic error.

**Figure S6.** Statistical power curves for tests evaluating evolutionary covariation using phylogenetic regression when data are simulated on balanced phylogenies using nonisotropic error.

**Figure S7.** Statistical power curves for tests evaluating evolutionary covariation using phylogenetic regression when data are simulated on random phylogenies using isotropic error.

**Figure S8.** Statistical power curves for tests evaluating evolutionary covariation using phylogenetic regression when data are simulated on random phylogenies using nonisotropic error.

**Figure S9.** Statistical power curves for tests evaluating evolutionary covariation using *D*-PGLS.



Y	X
2	5
4	4
5	3
3	2
4	1