TYPE I ERROR RATES FOR TESTING GENETIC DRIFT WITH PHENOTYPIC COVARIANCE MATRICES: A SIMULATION STUDY

Miguel Próa,1 Paul O’Higgins,1 and Leandro R. Monteiro1,2

1Centre for Anatomical and Human Sciences, The Hull York Medical School, The University of York, Heslington, York, YO10 5DD, United Kingdom
2Laboratório de Ciências Ambientais, Universidade Estadual do Norte Fluminense, Av. Alberto Lamego 2000, cep 28013-620, Campos dos Goytacazes, RJ, Brazil

Received November 22, 2010
Accepted July 3, 2012

Studies of evolutionary divergence using quantitative genetic methods are centered on the additive genetic variance–covariance matrix (G) of correlated traits. However, estimating G properly requires large samples and complicated experimental designs. Multivariate tests for neutral evolution commonly replace average G by the pooled phenotypic within-group variance–covariance matrix (W) for evolutionary inferences, but this approach has been criticized due to the lack of exact proportionality between genetic and phenotypic matrices. In this study, we examined the consequence, in terms of type I error rates, of replacing average G by W in a test of neutral evolution that measures the regression slope between among-population variances and within-population eigenvalues (the Ackermann and Cheverud [AC] test) using a simulation approach to generate random observations under genetic drift. Our results indicate that the type I error rates for the genetic drift test are acceptable when using W instead of average G when the matrix correlation between the ancestral G and P is higher than 0.6, the average character heritability is above 0.7, and the matrices share principal components. For less-similar G and P matrices, the type I error rates would still be acceptable if the ratio between the number of generations since divergence and the effective population size ($t/N_e$) is smaller than 0.01 (large populations that diverged recently). When G is not known in real data, a simulation approach to estimate expected slopes for the AC test under genetic drift is discussed.

KEY WORDS: Cheverud’s conjecture, divergence rates, G-matrix, genetic drift test, multivariate evolution.

Quantitative genetic methods provide inferences of evolutionary processes via the study of evolutionary divergence patterns and their relationship to intrapopulation adult variation (Lande 1979; Ackermann and Cheverud 2002, 2004; Marroig and Cheverud 2004; Monteiro and Gomes-Jr 2005; Perez and Monteiro 2009). The connection between neutral microevolutionary processes and macroevolutionary patterns is centered around the additive genetic variance–covariance matrix (G) (Lande 1980; Arnold et al. 2001; Jones et al. 2003; Bégin and Roff 2004), which is thought to determine both the response to selection and the pattern of neutral divergence, at least among populations over a small time scale (Lande 1980; Felsenstein 1988; Zeng 1988).

The expected pattern of phenotypic divergence among populations caused by random genetic drift in correlated traits can be used as a null hypothesis to test for neutral evolution (Lande 1979, 1980). The sampling distribution of the change in trait means in one generation ($\Delta \bar{z}$) has a mean of 0 and variance–covariance matrix $G/N_e$, the genetic covariance matrix in a population divided by the effective population size (Lande 1979). If the average phenotype of a population $a$ is represented by a column vector $\bar{z}_a$ of
polygenic traits with additive genetic and environmental components following multivariate normal distributions (Lande 1980), the probability distribution $\Phi$ after $t$ generations will be

$$
\Phi(z_a, t) \sim N(z_0, G(t/N_e)),
$$

which is a normal distribution with a mean equal to that of the initial population and variance–covariance matrix $G(t/N_e)$ (Lande 1979). If a number of populations are evolving independently (i.e., without gene flow), the expected among-population polygenic variance–covariance matrix $(B)$ is a function of the genetic covariance matrix $(G)$, effective population size $(N_e)$, and the number of generations $(t)$:

$$
B = G(t/N_e).
$$

As a result, the comparison of among-population $(B$ phenotypic) and within-group $(G$ genetic) variance–covariance matrices can be used as a means to determine whether genetic drift as a null model explains the pattern of divergence observed (Lofsvold 1986, 1988; Roff et al. 1999; Ackermann and Cheverud 2002; Bégin and Roff 2004).

Because phenotypic covariances are much easier to estimate than their genetic counterparts, replacing average $G$ with the pooled phenotypic within-group covariance matrix $(W)$, provided that the phenotypic covariance matrices for diverging populations remain similar, has been a widely used approach to study the evolutionary mechanisms of divergence (Ackermann and Cheverud 2002, 2004; Marroig and Cheverud 2004; Perez and Monteiro 2009). Cheverud (1988) investigated the relationship between genetic and phenotypic correlation matrices using data taken from the literature and concluded that phenotypic correlations were reasonable estimates (and generally proportional, although perhaps not in a strict mathematical sense) of the respective genetic correlations. A second conclusion from these data was that phenotypic covariances $W$ estimated with large samples might approach $G$ more accurately than genetic covariances estimated from small effective sample sizes, at least for morphometric data (Cheverud 1988; Revell et al. 2010). A number of meta-analyses from literature reviews and empirical results have to some degree corroborated Cheverud’s findings (Roff 1995; 1996; Koots and Gibson 1996; Roff et al. 1999; Waitt and Levin 1998). Nonetheless, this approach has been criticized on several grounds (Willis et al. 1991), but mostly because $W$ is not mathematically proportional (i.e., having a constant ratio) to average $G$. Apart from the issue of similarity and proportionality between matrices, more specific consideration of the actual consequences of using $W$ as a surrogate of average $G$ in empirical studies (Bégin and Roff 2004; Klingenberg et al. 2010) should prove fruitful and one such aspect, the impact in terms of type I error rates, is the focus of the present study.

Quantitative genetic theory predicts phenotypic covariances within a single population $(P)$ to be the sum of the genetic covariance $(G)$ and the environmental covariances $(E)$, $P = G + E$ (Falconer and Mackay 1996). A part-whole correlation is expected between phenotypic and genetic covariances; therefore, phenotypic covariances can be considered an estimate of genetic covariances with added error due to environmental covariances, even if not mathematically proportional.

Most of the discussion on the surrogacy of average $G$ by $W$ revolves around the similarities and differences between phenotypic and genetic covariances in single populations or from literature reviews, and the differences in empirical comparative results obtained when using one kind of estimate or the other. The latter are rare, due to the difficulty in estimating genetic parameters for a large number of species at the same time (Bégin and Roff 2004). Considering that Lande’s (1979, 1980) model expects the among-population covariance matrix $B$ to be proportional to the average $G$ when genetic drift is the sole evolutionary mechanism, for the purpose of evolutionary divergence tests of neutral evolution, the relevant discussion is not whether $G$ and $P$ are exactly proportional in single populations, but whether using the phenotypic pooled within-group covariance matrix $W$ instead of the average $G$ will add enough error (caused by the environmental covariances) to lead into erroneous conclusions. The tests that have been used in the comparison of among-species phenotypic covariances and genetic covariances (Lofsvold 1988; Ackermann and Cheverud 2002, 2004) do not test for exact proportionality between $B$ and average $G$, but for similarity in different matrix features, such as the correlation of principal components and the distribution of eigenvalues. The expectation of proportionality rests on a number of assumptions (Lande 1979) that are probably violated in most natural populations (Lofsvold 1988), for example, through the lack of large effective population sizes (Lofsvold 1988), or because of differences in the starting times of lineages (Revell 2007). Furthermore, error in the estimation of the average $G$ might lead to unpredictable deviations from the expectation. Lofsvold (1988) has suggested that the acceptance of genetic drift as a null hypothesis will be more robust to the breaking of the model’s assumptions than the rejection (so type I error rates are of more concern than the power), and in real studies it might be hard to determine the actual cause of rejection, natural selection being one of the possible explanations. One might expect that a consequence of using pooled within-group phenotypic instead of genetic covariances would be to increase the probability of rejecting (type I error rate) a true null hypothesis of genetic drift.

In this study, we examined the consequences of using pooled within-group phenotypic instead of average genetic covariance matrices in the Ackermann and Cheverud (2002) test of genetic drift (referred to as the AC test from here on) in terms of type I error rates using a simulation of phenotypic evolution in
diverging populations. We identified the most relevant parameters and discusses a number of recommendations.

Material and Methods

SIMULATION MODELS: GENERAL DESCRIPTION

The simulations were performed using the quantitative genetic theory from Lande (1979, 1980). Starting from an ancestral population with genetic covariance matrix $G$ and mean vector $\bar{z}$, a number (15 or 30) of descendant population mean vectors $\bar{z}_a$ were generated using the $t$-fold convolution equation (1) for a range of $t/N_e$ ratios ($0.000001–100$ in increments of $1$ in $\log_{10}$ scale). This approach is equivalent to a random walk in multivariate space where each descendant population is evolving at a rate equivalent to $G/N_e$. Instead of generating the intermediate phenotypes for each step (generation) of the random walk, the convolution allows for a direct generation of the end points with the same results and in a computationally efficient way.

The descendant populations from the ancestral distribution $N[\bar{z}_a, G]$ were sampled $n$ times (sample sizes $10–100$, in increments of $10$) according to the multivariate normal distribution $N[\bar{z}_a, P]$, for each population. The first step in the simulations required an ancestral genetic variance–covariance matrix ($G$) to generate species means and the second step required a phenotypic within-population variance–covariance matrix ($P$) to generate individual specimens for each population. The same $P$ was used for all populations (the pooled within-group phenotypic covariance matrix $W$ is an estimate of the original $P$). Different simulation models were used, either generating random $G$ and $P$ matrices as starting parameters (fully stochastic), or using predetermined matrices obtained from real datasets. The fully stochastic set of simulations required the generation of random positive definite covariance matrices (where all eigenvalues are $>0$) that could be used as parameters in the generation of random multivariate normal numbers representing individuals sampled, as described below.

Throughout the paper, we used correlations of lower triangular covariance matrices excluding diagonals (variances) as one measure of structural similarity (alongside Common Principal Components—CPCs, Phillips and Arnold 1999). Note that permutations of the matrix elements were not used for testing significance of correlations. This procedure is not indicated for testing similarity in covariance matrices if variables have differences in scale (Cheverud and Marroig 2007), but was the most appropriate choice in our simulations. This is because the algorithm for the generation of random positive definite matrices (see details below) yielded matrices where covariances were small relative to variances. As a result, even when two covariance matrices were independently generated, they presented high positive matrix correlations when the diagonal was included (or when using comparison methods such as random skewers), because the variances and covariances would systematically form two groups of values in the matrix scatterplot. Only when diagonals were excluded, was the expected correlation for independent matrices 0. This particular structure in the random matrices (high variances, small covariances) is a consequence of generating positive definite matrices, because matrices with high covariances relative to variances are likely to be nonpositive definite. Therefore, the most accurate description of matrix similarities in our simulations was derived from matrix correlations, using the lower triangular elements, excluding the diagonals. This is equivalent to comparing correlation matrices derived from the covariance matrix, as the information regarding variances is disregarded. In real datasets, there would be no justification to exclude the variances from the structural comparisons, as differences in scale of variances and covariances are a relevant part of the structure.

SIMULATION MODEL 1

In this set of simulations, we used real matrices rather than randomly generated ones. The matrices for the main simulations were obtained from a honey bee (Apis mellifera) wing shape dataset with 16 shape variables (partial warps), modified (used only landmarks 11–20) from Monteiro et al. (2002), and a gastropod shell shape dataset (Physa heterostropha) with 14 shape variables (DeWitt 1996, 1998). The average heritability of bee wing variables (calculated as $I(IoG)/(IoP)^{-1} I m^{-1}$, where $I$ is a row vector of ones, $I$ is the identity matrix, $o$ is a Hadamard [element-wise] product, and $m$ is the number of variables) was 0.217, and the effective sample size, given that 21 bee colonies were used, was 4.6. The effective sample size was calculated as the product of heritability and the number of families, as described in Cheverud (1988). For the shell shape dataset, the average heritability was 0.607, and the effective sample size was 11.5 (19 families were used). These two datasets present important differences in the structural similarity of $G$ and $P$. For the bee wing data, the matrix correlation between $G$ and $P$ was 0.804, whereas for the shell dataset the correlation was 0.442. Although the average heritability was smaller for the bee wing data, their genetic and phenotypic matrices were more similar than in the shell dataset. This is not unexpected, as these average heritabilities do not measure matrix similarity, only the relative magnitudes of the genetic and phenotypic variances. A comparison of these genetic and phenotypic covariance matrices via CPC indicated that the bee dataset matrices shared the full set of principal components (full CPC model supported by the jump-up approach) and the shell dataset matrices shared no principal components (Unrelated model supported).

SIMULATION MODEL 2

In this model, $P$ and $G$ were exactly proportional and differed only by a scalar multiplication. $G$ was defined first as a
random positive definite covariance matrix using the eigenvector method from Marsaglia and Olkin (1984) and Joe (2006). The eigenvector method first generates random eigenvalues \((\lambda_1, \ldots, \lambda_m)\) from a uniform distribution (the diagonal matrix \(\mathbf{L}\)). A lower bound of eigenvalues was set to 1 and an eigenvalue ratio (between upper and lower bound) set to 10. The algorithm then generates a random orthogonal matrix of eigenvectors \(\mathbf{Q}\) (via QR-decomposition) and constructs the genetic covariance matrix \(\mathbf{G}\) as \(\mathbf{QLQ}^T\). The phenotypic covariance matrix was defined by the scalar multiplication \(\mathbf{P} = k\mathbf{G}\), where \(k\) is a uniform random number from 1 to 10. This approach generates a random uniform distribution of covariance matrices in the space of positive definite covariance matrices (Joe 2006). In this set of simulations, as in all other fully stochastic models, \(\mathbf{G}\) and \(\mathbf{P}\) had 15 dimensions.

**SIMULATION MODEL 3**

In this model, \(\mathbf{G}\) and \(\mathbf{P}\) are defined as independent random positive definite covariance matrices using the uniform correlation matrix method (Joe, 2006), where a random correlation matrix \(\mathbf{R}\) is first generated from a uniform distribution of partial correlation coefficients. The variances are generated separately as a diagonal matrix \(\mathbf{S} = \text{diag}(\sigma_1^2, \ldots, \sigma_m^2)\) with elements obtained from a uniform distribution ranging from 1 to 10. The random covariance matrices are constructed as SRS. \(\mathbf{G}\) and \(\mathbf{P}\) were independently derived in this model, with the restriction that the variances in \(\mathbf{P}\) are always larger than the respective variances in \(\mathbf{G}\). To achieve this, the variances (diagonal) of \(\mathbf{P}\) were random multiples of the respective variances in \(\mathbf{G}\). This procedure ensures that the variances of \(\mathbf{P}\) were always larger, but \(\mathbf{P}\) and \(\mathbf{G}\) were independent. A series of 1000 simulations using this model yielded a distribution of \(\mathbf{G}\) and \(\mathbf{P}\) matrices correlations with a 95% confidence interval (CI) (using 2.5 and 97.5 quantiles) of \(-0.194\) to \(0.201\), and a median of \(0.0004\). A further structural comparison of model 3 matrices was performed by CPC of 100 simulated \(\mathbf{G}\) \& \(\mathbf{P}\) pairs.

We compared estimated covariance matrices after generating 300 random observations from a multivariate normal distribution with a mean vector of zeros and random \(\mathbf{G}\) and \(\mathbf{P}\) (defined as above) as parametric covariance matrices. Because the model fitting in CPC depends on sample sizes, we have maintained a standard \(n = 300\) for all other comparisons as well. The results indicated the Unrelated model (no shared principal components) in all comparisons. Although the covariance structure is independently generated, all matrices generated by this method have variances on a much larger scale than the covariances. Therefore, there is some structural similarity because all matrices have clearly two groups of elements (covariances and variances), and the variances are always much larger than the covariances. This model is not biologically reasonable because \(\mathbf{G}\) and \(\mathbf{P}\) independence is unlikely (even if not proportional) due to a part-whole relationship. The model is included as a control, as the opposite to the mathematical proportionality of simulation model 2, allowing for a check that the simulations behaved as expected at extremes of \(\mathbf{G}\) and \(\mathbf{P}\) similarity and independence.

**SIMULATION MODEL 4**

Simulation model 4 was designed to generate correlated \(\mathbf{G}\) and \(\mathbf{P}\) matrices, but without a CPC structure. To achieve this, we have used the quantitative genetic relation \(\mathbf{P} = \mathbf{G} + \mathbf{E}\). In these models, \(\mathbf{G}\) and \(\mathbf{E}\) were defined first and independently, \(\mathbf{P}\) was then defined as a random matrix with expected value \(\mathbf{G}\) and a random perturbation \(\mathbf{E}\) (Marsaglia and Olkin 1984). \(\mathbf{E}\) and \(\mathbf{G}\) were generated by the uniform correlation matrix method described in simulation model 3, where \(\mathbf{G}\) has a range of variances between 1 and \(\sigma_{\text{maxG}}^2\) (we used a maximum of 10), and \(\mathbf{E}\) has a range of variances between 0 and \(\sigma_{\text{maxE}}^2\), where \(\sigma_{\text{maxG}}^2\) and \(\sigma_{\text{maxE}}^2\) are the function parameters determining the upper limits of the range of variances in \(\mathbf{G}\) and \(\mathbf{E}\), respectively. The expected value of the average heritability of the variables in the simulations is the ratio \((\sigma_{\text{maxG}}^2 - 1)/(\sigma_{\text{maxG}}^2 + \sigma_{\text{maxE}}^2)\). This method generated correlated \(\mathbf{P}\) and \(\mathbf{G}\), but without a common PC structure. This pattern is ensured because the variables with larger variances in \(\mathbf{E}\) will be generally different than the variables with larger variances in \(\mathbf{G}\) so that \(\mathbf{P}\) is less likely to inherit principal components from \(\mathbf{G}\) (H. Joe, pers. comm.). Of course, as the variances in \(\mathbf{E}\) become smaller than variances in \(\mathbf{G}\) (\(\sigma_{\text{maxE}}^2 << \sigma_{\text{maxG}}^2\)), CPCs between \(\mathbf{P}\) and \(\mathbf{G}\) appear. The distribution of matrix correlations from 1000 model 4 simulations (using \(\sigma_{\text{maxG}}^2 = 10\) and \(\sigma_{\text{maxE}}^2 = 9\) for a similar range of variances) presented a 95% CI of 0.560–0.897, and a median of 0.776. To check for common eigenstructure, we performed a CPC analysis of 100 model 4 simulations of \(\mathbf{G}\) and \(\mathbf{P}\). The simulations showed strong support for the Unrelated model (no CPCs) in 65% of the cases, using the jump-up approach. The remaining simulations supported 1 (26%) or 2 (9%) CPCs. In simulation model 4, the perturbation of expected value \(\mathbf{G}\) by \(\mathbf{E}\) included random rotations of its eigenstructure, even if matrix correlations were high.

**SIMULATION MODEL 5**

In simulation model 5, \(\mathbf{G}\) was defined as a random positive definite covariance matrix using the eigenvector method from Marsaglia and Olkin (1984) and Joe (2006), but where \(\sigma_{\text{maxG}}^2\) is the max/min eigenvalue ratio (this parameter will have a different interpretation than in model 4, but the average heritabilities expected are exactly the same in models 4 and 5). \(\mathbf{P}\) was defined as the sum of \(\mathbf{G} + \mathbf{E}\), where \(\mathbf{E}\) was generated by the uniform correlation method, with a variance range of 0 to \(\sigma_{\text{maxE}}^2\). In this model, \(\mathbf{P}\) readily inherits the principal component structure of \(\mathbf{G}\), even when \(\sigma_{\text{maxE}}^2 \sim \sigma_{\text{maxG}}^2\). The matrix correlation in these simulated matrices (with \(\sigma_{\text{maxG}}^2 = 10\), and \(\sigma_{\text{maxE}}^2 = 9\) was smaller than in model 4 (matrix correlation distribution 95% CI = 0.061–0.462, median = 0.274), but the
Figure 1. Simulation of genetic drift in four populations. The means of each population were evolved from an ancestral multivariate normal distribution with mean $\mu = (0,0)$ and covariance matrix $\Sigma_G$. Each population was randomly sampled 30 times using the respective average and covariance matrix $\Sigma_P$. Left panels correspond to simulation model 4, where $\Sigma_P$ and $\Sigma_G$ are correlated, but do not share principal components. Right panels correspond to simulation model 5, where $\Sigma_P$ and $\Sigma_G$ share principal components but have low correlation. The ancestral genetic covariance matrix is depicted as a dashed ellipse. The population phenotypic covariance matrices are depicted as solid ellipses. Filled circles correspond to population means and open circles correspond to individual observations.

CPC analysis shows strong support for a shared latent structure, where 13% of the simulations supported the Unrelated model (0 CPCs), and 60% of the simulations supported models with 3 or more common PCs. The perturbation caused by $E$ generates random differences between $\Sigma_P$ and $\Sigma_G$, but not a random rotation of the eigenstructure of $\Sigma_G$ (when $\sigma_{\text{max}}^2 > \sigma_{\text{max}}^2 E$). This pattern is caused by a lambda ratio ($\sigma_{\text{max}}^2 > \sigma_{\text{max}}^2 E$) of 10 or larger, which will produce $\Sigma_G$ matrices with sharp elliptical contours (noticeable principal components), ensuring that the principal component structure of $\Sigma_G$ is inherited by $\Sigma_P$, even when $\sigma_{\text{max}}^2 \sim \sigma_{\text{max}}^2 E$ (H. Joe, pers. comm.).

An illustrative bivariate example of the typical main differences between simulation models 4 and 5 is depicted in Figure 1. We simulated for each model and $t/N_e$, four populations descending from an ancestor (0,0) with a random genetic covariance matrix (shown as dashed lines in Fig. 1) and a random phenotypic matrix. The same phenotypic covariance matrices were used to generate 30 observations in each population and these are depicted as distinct clusters around each descendant. In simulation model 4, the matrix $\Sigma_P$ is a random rotation of $\Sigma_G$, whereas in simulation model 5, the main axes of $\Sigma_G$ are preserved in $\Sigma_P$.

**GENETIC DRIFT TEST**

Genetic drift as a neutral model for phenotypic divergence was tested by comparing the among-population covariance matrix ($\Sigma_B$) and the within-population phenotypic covariance matrix ($\Sigma_W$, as a surrogate of the average $\Sigma_G$) for the simulated data using the method of Ackermann and Cheverud (2002, 2004). This involved extracting the eigenvectors ($\mathbf{M}$) and eigenvalues ($\mathbf{m}$) of $\Sigma_W$, and projecting each population phenotypic vector of means $\bar{z}$ on $\mathbf{M}$, $Y = \bar{z} \mathbf{M}$. The vector of means for each population was the one estimated from the simulated samples, not the parametric means generated from the ancestral $\Sigma_G$ and ancestral vector of means. Finally, we calculated the variances for each column of $Y$ and performed a regression of the variances of $Y$ on $\mathbf{m}$

$$\ln(Y_i) = \ln(t/N_e) + \beta \ln(m_i)$$

(3)
Testing with a $t$-test whether the slope of the regression ($\beta$) is different from 1 indicates whether the pattern is compatible with genetic drift. The null hypothesis of genetic drift is rejected if the slope deviates significantly from 1 (Ackermann and Cheverud 2002).

For each combination of parameters (ancestral $G$, $P$, $t/N_e$ ratio, sample size, number of descendant populations, $\sigma^2_{\text{max}}/\sigma^2_{\text{maxG}}$) in different models, we simulated 1000 datasets to estimate type I error rates. In the simulated datasets, the only mechanism producing phenotypic divergence among the descendant populations was genetic drift. When using a significance level of $\alpha = 0.05$, we expect that a true null hypothesis has a 5% chance of being rejected (a type I error). If the use of phenotypic covariances as proxies for genetic ones in the genetic drift test does increase the type I error rates, we expect to find that, using a significance level of 5%, the null model of genetic drift will be rejected in more than 5% of the simulated samples.

All the simulations and analyses were run in the R environment (R Development Core Team, 2010) using functions from the packages MASS (Venables and Ripley 2002), clusterGeneration (Qiu and Joe 2009), and vegan (Oksanen et al. 2010). The R code (commented) used for the simulations is available as Supporting information.

**Results**

For the simulation using the bee wing shape data (genetic and phenotypic covariance matrices) as starting parameters, the type I error rate decreased with increasing sample sizes for small $t/N_e$ ratios (between 0.01 and 0.000001) irrespective of the number of populations (15, 30) used (Figs. 2A, S1A). The error rate increased for larger sample sizes when $t/N_e \geq 0.1$. The correlation between $G$ and $W$ remained stable over simulations for all $t/N_e$, with a median matrix correlation of 0.788, and a 95% CI (based on 0.025 and 0.975 quantiles) from 0.741 to 0.830. The matrix correlation for the ancestral (original) $P$ and $G$ was 0.804.

For the simulation using the shell shape data as starting parameters, the type I error rates remain at acceptable levels for small sample sizes above 20 in $t/N_e$ ratios equal to or below 0.001, and both numbers of populations (15, 30) (Figs. 2B, S1B). For the simulation with $t/N_e = 0.01$, the error rates remain increase with sample size. This is a slightly worse result than in the simulations with bee wing parameters, because in the latter, the simulation with $t/N_e = 0.01$ yielded acceptable error rates (Fig. 2A, B). The correlation between average $G$ and $W$ also remained stable over simulations using the shell dataset for all $t/N_e$, with a median matrix correlation of 0.441, and a 95% CI (based on 0.025 and 0.975 quantiles) from 0.405 to 0.479. The matrix correlation for the ancestral (original) $P$ and $G$ was 0.442.

In the simulation model 2, where $G$ and $P$ differed only by a random constant (Fig. 2C), the resulting pattern showed slight fluctuations around the expected type I error rate (0.05) for any value of $t/N_e$. This result was observed for sample sizes above 40 individuals per population regardless of the number of populations (15 or 30; Figs. 2C, S1C).

The simulation model 3, where $P$ and $G$ were generated independently (Figs. 2D, S1D), presented acceptable type I error rates only for $t/N_e$ ratios equal to or below 0.00001, regardless of the number of populations. The simulations with $t/N_e > 0.001$ all presented type I error rates above 0.8 and are not shown in the Figure. Because in this model $G$ and $P$ have independent covariances, the test would be expected to show significant deviations from the unity slope for any combination of parameters. This suggests that the power of the test must be small for such values of $t/N_e$.

Simulation models 4 and 5 were designed to generate $G$ and $P$ correlated matrices, where $P = G + E$. In simulation model 4, the random matrix $E$ adds variation to the genetic covariances and variances, including a random rotation of the eigenstructure when $P$ is calculated, even if the range of variances in $E$ ($\sigma^2_{\text{max}}$) is the same or a bit smaller than the range of variances in $G$ ($\sigma^2_{\text{maxG}}$). In simulation model 5, the $E$ matrix only causes differences in the principal components of $G$ and $P$ when $\sigma^2_{\text{max}} > \sigma^2_{\text{maxG}}$. The first set of analyses was performed using the same range of variances in $G$ and $E$ for both models. The simulation model 4 presented acceptable error rates for sample sizes larger than 20 regardless of $t/N_e$ ratio and number of populations. The simulation model 5 presented acceptable type I error rates only for $t/N_e$ ratios equal to or below 0.001, regardless of the number of populations (Figs. 2E, F, S1E, S1F).

Exploring the simulations with a larger range of parameters, we found that the ratio of upper limits of environmental and genetic variance ranges ($\sigma^2_{\text{max}}/\sigma^2_{\text{maxG}}$) also influences the type I error rates of the test. One unexpected result was that in simulation model 4, as $\sigma^2_{\text{max}}$ gets smaller than $\sigma^2_{\text{maxG}}$, the type I error rates increase. We performed the simulations again, with fixed sample sizes (100), number of groups (15), and $t/N_e$ (10) to assess the influence of $\sigma^2_{\text{max}}/\sigma^2_{\text{maxG}}$ on the slope of the test (Fig. 3). In the right panel of Figure 3, using simulation model 5 (where $P$ readily inherits the eigenvectors of $G$), the value of $\sigma^2_{\text{max}}/\sigma^2_{\text{maxG}}$ gets smaller, the slope of the test converges to 1, as expected under genetic drift. On the other hand, in the simulation model 4 (left panel of Fig. 3), the expected value of the slope under simulation of drift is 1 only when $\sigma^2_{\text{max}} \sim \sigma^2_{\text{maxG}}$. As the ratio of variance ranges get smaller, the expected slope converges to approximately 1.3, and this pattern explains why the type I error rates increase when $\sigma^2_{\text{max}}$ gets smaller than $\sigma^2_{\text{maxG}}$. The simulations using the real matrices (model 1) and the same parameters described above had expected slopes of 1.3 (bees) and 0.8 (Physa shells).
Figure 2. Type I error rates for the simulated analyses with varying sample sizes and $t/N_e$ ratios (drift intensities). The legends and line types indicate the value of $t/N_e$ used (only when differences among lines are noticeable). The dashed horizontal straight line indicates the expected type I error rate of 0.05. All simulations in this figure were performed with 15 populations. (A) Error rates for the bee wing dataset. (B) Error rates for the shell dataset. (C) Stochastic simulations (model 2) where $G$ was random and $P$ was exactly proportional to it $P = kG$ (multiplication by a random scalar $k$ drawn from a uniform distribution between 1 and 10). (D) Stochastic simulation (model 3) where both $G$ and $P$ were random and completely independent. (E) Stochastic simulation (model 4) where $G$ and $P$ were correlated ($P = G + E$), but did not share a common latent structure ($G$ and $E$ with the same range of variances). (F) Stochastic simulation (model 5) where $G$ and $P$ were correlated and shared a common latent structure ($G$ and $E$ with the same range of variances). See text for model details.

Slopes larger than 1 might be obtained when the variance among population averages projected on the first eigenvectors of $W$ is larger than the corresponding eigenvalues, whereas slopes smaller than 1 are the result of less among-population variation than predicted by the eigenvalues of the first PCs of $W$.

Considering that, for simulation model 5, smaller variance range ratios lead to the expected slope under genetic drift, we explored the combination of simulation parameters that would lead to acceptable type I error rates on the AC test (Table 1). When we decrease the $\sigma^2_{maxE}/\sigma^2_{maxG}$ ratio, the correlations...
Figure 3. Slopes (β) of the Ackermann and Cheverud test in relation to the ratio of upper bounds of environmental ($\sigma^2_{\text{maxE}}$) and genetic ($\sigma^2_{\text{maxG}}$) variances in the simulations for models 4 (Sim4) and 5 (Sim5), using $t/N_e = 10$, 15 dimensions in $G$, 15 populations, and 100 observations per population. Genetic variances ranged between 1 and $\sigma^2_{\text{maxG}} = 10$ and the environmental variances ranged between 0 and $\sigma^2_{\text{maxE}} = 1$–15. The solid lines show the expected (mean) value for the slope over 1000 simulations, whereas the dashed lines indicate the upper and lower limits of 95% CIs. The dotted line indicates the unity slope, which is the theoretical expectation under genetic drift.

Table 1. Type I error rates for the genetic drift test using simulation model 5 (1000 repetitions), with 15 variables, 15 groups, and 50 individuals per group ($\alpha = 0.05$), with varying $t/N_e$. $\sigma^2_{\text{maxE}}/\sigma^2_{\text{maxG}}$ is the ratio of the upper bounds of variances in the environmental and genetic matrices (see text). CI-$h^2$ is the 95% CI for the average heritability in each set of simulations, CI-MatCor is the 95% CI for $G \times P$ matrix correlations in each set of simulations, fCPC is the percentage of significant full CPC models for $G$ and $P$ in 100 simulations, CICPCs are the 95% CI (percentiles) for the number of common principal components for $G$ and $P$ in 100 simulations.

<table>
<thead>
<tr>
<th>$\sigma^2_{\text{maxE}}/\sigma^2_{\text{maxG}}$</th>
<th>CI-$h^2$ (95% CI)</th>
<th>CI-MatCor (95% CI)</th>
<th>fCPC</th>
<th>CICPCs (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0.86 – 0.94</td>
<td>0.808 – 0.970</td>
<td>100</td>
<td>14</td>
</tr>
<tr>
<td>0.2</td>
<td>0.76 – 0.88</td>
<td>0.543 – 0.891</td>
<td>100</td>
<td>14</td>
</tr>
<tr>
<td>0.3</td>
<td>0.70 – 0.84</td>
<td>0.385 – 0.812</td>
<td>79</td>
<td>7–14</td>
</tr>
<tr>
<td>0.4</td>
<td>0.64 – 0.80</td>
<td>0.277 – 0.729</td>
<td>67</td>
<td>4–14</td>
</tr>
</tbody>
</table>

between $P$ and $G$ increase, as well as the number of CPCs. If $\sigma^2_{\text{maxE}}$ is around 20% of $\sigma^2_{\text{maxG}}$, the matrix correlations observed are not particularly high, as compared to real $P$ and $G$ matrices estimated with large sample sizes, but they do share a common eigenstructure, and for any value of $t/N_e$, the type I error rates approach acceptable values. Performing the same simulations with more variables ($n = 30$), the same results are obtained with larger within-population sample sizes ($n > 100$) (results not shown). It is evident from these results that the combination of parameters yielding acceptable type I error rates is sensitive to the models under which the starting matrices were generated.

Discussion

Testing diversification by genetic drift is a useful starting point in the study of evolutionary variation (Lynch 1990; Ackermann and Cheverud 2004; Weaver et al. 2007; Perez and Monteiro 2009). Cheverud’s (1988) suggestion that genetic covariance matrices could be safely replaced by phenotypic matrices for evolutionary inferences was greeted with scepticism, and “Cheverud’s conjecture” (Roff 1995) has been tested and discussed in a number of papers (e.g., Roff 1995, 1996; Koots and Gibson, 1996; Waitt and Levin 1998; Roff et al. 1999; Begon and Roff 2004; Hadfield et al. 2006; Kruuk et al. 2007), usually by comparing the similarity of genetic and phenotypic covariances, seldom by checking the influence of matrix differences in the results of tests. Thus, the evidence gathered has been equivocal and the most relevant studies (large reviews of data) indicate a general agreement with Cheverud (1988), but also recommend caution in the interpretations of results because matrix comparisons among isolated populations using genetic or phenotypic covariances might
Our results indicate that the type I error of Ackermann and Cheverud’s (2002, 2004) test of proportionality between \( \mathbf{B} \) and \( \mathbf{W} \) is influenced mainly by the structural similarity between the ancestral \( \mathbf{G} \) and \( \mathbf{P} \), the ratio of variance ranges (approximated by the average heritability), and the ratio of time and effective population size \( t/t_N \). If the parametric genetic and phenotypic covariance matrices are exactly proportional, as in the simulation model 2, the type I error rates are acceptable for any \( t/t_N \) ratio (as expected). On the other extreme (simulation model 3), where \( \mathbf{G} \) and \( \mathbf{P} \) were generated with an unrealistic minimum of structural similarity, the type I error rate is unacceptable for most values of \( t/t_N \).

The simulations showed that even if the ancestral \( \mathbf{G} \) and \( \mathbf{P} \) are not proportional but do share a large number of principal components, have an average heritability around 0.5, and matrix correlation above 0.7 over all variables (as in our simulation model 5), acceptable type I error rates will be obtained for any \( t/t_N \) ratio. When \( \mathbf{G} \) and \( \mathbf{P} \) do not share principal components but are highly correlated (\( r > 0.7 \)) and have average heritabilities approaching 0.5, the type I error rates should be acceptable for any \( t/t_N \) ratio (as in our simulation model 4). Average heritabilities different from 0.5 will bias the expectation of the slope in the AC test due to concentration of variation among projections of population averages in the first eigenvectors of \( \mathbf{W} \). In these cases, type I error rates will still be acceptable for \( t/t_N < 0.01 \).

The combination of parameters laid out is not an unrealistic expectation. The literature indicates that considerable agreement between genetic and phenotypic correlations is often found and that the correlations between \( \mathbf{G} \) and \( \mathbf{P} \) are usually above 0.6 for morphological data when effective sample sizes are large (Cheverud 1988; Koots and Gibson, 1996; Roff 1996; Waitt and Levin 1998; Begin and Roff 2004; Kruuk et al. 2007; Oliveira et al. 2009).

In a study where only phenotypic data are available, it might be complicated or impossible to determine whether the relationship between the ancestral \( \mathbf{G} \) and \( \mathbf{P} \) fits into the assumptions outlined above. These parameter values can, nevertheless, be used as guidelines for comparisons among populations as indirect evidence of ancestral \( \mathbf{G} \) and \( \mathbf{P} \) similarity (Oliveira et al. 2009), or one might use the Monte Carlo simulation approach described below to estimate a CI for the slope of the AC test under drift.

Our example datasets (simulation model 1) seem to behave in a similar way to simulation model 4 for extremes of low and high \( \sigma^2_{\text{max}}/\sigma^2_{\text{max}} \). The expected slope for the simulations using the bee matrices was 1.3, the similarity of \( \mathbf{G} \) and \( \mathbf{P} \) was high, they did share principal components, but the average heritability was low (it should have been higher than 0.6 to fit the model 4 more closely). On the other hand, the simulations with shell matrices had an expected slope of 0.8. \( \mathbf{G} \) and \( \mathbf{P} \) similarity was low, they did not share principal components, but the average heritability was high (should have been lower than 0.3). Such results would be observed if model 4 was changed to calculate \( \mathbf{P} = k(\mathbf{G} + \mathbf{E}) \), so that the average heritability would be decreased or increased by the scalar \( k \) without influence in the correlation or shared structure between \( \mathbf{P} \) and \( \mathbf{G} \). These results suggest that \( \mathbf{G} \) and \( \mathbf{P} \) are related in complex ways, which can hardly be reduced to scalar comparisons without considerable loss of information. If some information about \( \mathbf{G} \) and \( \mathbf{P} \) is available, one might use this simulation approach to estimate the expected slope of the AC test and use this expectation in the test of the real data (instead of the theoretical unity slope). For example, in the bee wing analyses, we could have used a slope of 1.3 as parameter in the \( t \)-test of the AC tests and the type I error rates would be acceptable for any value of \( t/t_N \). Alternatively, the 95% CI for the expected AC test slope under genetic drift simulations ranged from 1.1 to 1.5, and an observed slope could be compared with this interval for evidence of departure from the neutral expectation. When genetic data are not available, it might be possible to use the between-population covariance matrix (\( \mathbf{B} \)), estimated from phylogenetic independent contrasts if possible (Revell 2007) and the within-population phenotypic covariance matrix \( \mathbf{W} \) as proxies for the ancestral \( \mathbf{G} \) and \( \mathbf{P} \), respectively, in the simulations to estimate the expected slope under drift. A simulation function provided as Supporting information (simulationAC-slope.R) will calculate a mean estimate and a 95% CI for AC test slopes under genetic drift for any ancestral \( \mathbf{G} \) and \( \mathbf{P} \). Observed slopes can be compared to the CI or the mean estimate can replace the parameter slope = 1 in the ordinary \( t \)-test.

Within-population sample sizes influence the type I error rates, but they need to be considered in conjunction with the number of populations and the dimensionality of the matrices. For our fully stochastic simulations, all matrices had 15 dimensions and most acceptable type I errors were observed for within-population samples larger than 40. The number of populations used had a slight but negligible effect.

It is possible that sampling error in the estimation of \( \mathbf{G} \) might lead to a similar pattern of type I errors as when average \( \mathbf{G} \) is replaced by \( \mathbf{W} \), because the parametric and estimated \( \mathbf{G} \) matrices are not likely to be exactly proportional as well. It is not clear whether sampling error in the estimation of \( \mathbf{G} \) is comparable to the environmental covariance matrix \( \mathbf{E} \), but a part of Cheverud’s conjecture was that \( \mathbf{W} \) could be a more reliable estimate of the parametric \( \mathbf{G} \) than a genetic matrix estimated from a small effective sample size (Cheverud 1988), and phenotypic correlation estimates are often within the CIs of genetic correlations (Koots and Gibson 1996; Roff 1996). The instability of covariance matrix and factor estimation for small sample sizes is well known in multivariate statistics (MacCallum et al. 1999; Krzanowski 2000).
2000), and genetic covariance matrices can be particularly demanding with respect to samples sizes (Cheverud 1988). Patterns caused by sampling error in the estimation of genetic covariance matrices, such as biases on eigenvalues, are well known (Meyer and Kirkpatrick 2010) and a considerably large statistical literature is devoted to such topics. As long as the sampling error can be considered independent from the parametric $G$, the simulation function provided as Supporting information can be adjusted to address specific concerns regarding the error in the estimation of $G$.

In some of the simulations, particularly model 1 (with predetermined matrices) and the fully stochastic simulation where $G$ and $P$ where random and completely independent (model 2), a trend was observed where for higher values of $t/N_e$, the type I error rates increase with within-population sample size (see Fig. 2A, B, D). This counterintuitive result was also observed in simulation model 4, when $\sigma^2_{\text{max}}$ is smaller than $\sigma^2_{\text{max}}$ (Fig. S2). Considering that, depending on this ratio, the differences between $G$ and $P$ caused the expected value of the slope of the AC test to be larger than 1 (as shown in Fig. 3 due to more variation among populations than predicted by the eigenvalues of $W$), the type I error rates increase with sample sizes because the CIs become narrower (there is an expected increase in power) and a larger percentage of simulated tests will show significant results. The type I error converges to a value that depends on the magnitude of deviation of the expected AC test slope from 1 and the size of the CI. For smaller $t/N_e$ ratios, there is a reduction in the contribution of the $G$ matrix to among-population variation (it will be proportional to $t/N_e, G$). Because the simulations calculate among-group variation using averages estimated from the $n$ observations generated by $P$ at each population (and not the parametric means generated by $G$), when $t/N_e$ decreases, most among-population variation is generated and predicted by $W$, and the expected slope of the AC test is 1. This also explains the effect in reverse, when $\sigma^2_{\text{max}} > \sigma^2_{\text{max}}$, causing among-population variation to be smaller than the eigenvalues of $W$ and the expected slope of the AC test to be $< 1$ (Fig. 3).

In summary, replacing $G$ with $W$ when testing the null hypothesis of divergence by genetic drift is not likely to increase the type I error rates of the AC test, unless the ancestral $G$ and $W$ are structurally dissimilar (mathematical proportionality is not a required condition), the $t/N_e$ ratio is large, and sample sizes are small ($< 40$ per group). A Monte Carlo simulation approach might be used to estimate the expected slope of the AC test under drift, taking into account the structural differences between $G$ and $P$. A number of other methods have been proposed, which were used to compare among-population and within-population covariance matrices (Lofsvold 1988; Bégin and Roff 2001; Revell et al. 2007). Not all of these alternative methods will have their type 1 error rates increased when average $G$ is replaced by $W$. Type II errors are also a possibility, as genetic drift is the alternative hypothesis in some tests (Revell et al. 2007). The simulation function provided as Supporting information can be modified to account for other methods of matrix comparison. Alternatively, model-based approaches (Butler and King 2004) should provide reliable and possibly more informative tests of evolutionary processes and scenarios. The simulation approaches, particularly the more sophisticated individual-based models (Revell 2007), should prove useful in further analyses, comparing methods and testing evolutionary quantitative genetics models.

ACKNOWLEDGMENTS

The authors would like to thank T. DeWitt for allowing the use of the Physa shape data as simulation parameters. Previous versions of the manuscript were greatly improved by comments from G. Marroig and anonymous reviewers. MP was funded by the Fundação para a Ciência e a Tecnologia (Portugal), through the Ph.D. Programme in Computational Biology, Instituto Gulbenkian de Ciência (Portugal). LRM is funded by the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq, Brazil) and the Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ). The authors declare no conflict of interest.

LITERATURE CITED


Q3


Associate Editor: G. Marroig
Supporting Information

The following supplementary material is available for this article:

Figure S1. Type I error rates for the simulated analyses with varying sample sizes and $t/N_e$ ratios (drift intensities).

Figure S2. Type I error rates for the simulated analyses under model 4 with varying $\sigma_{2\text{max}G}/\sigma_{2\text{max}E}$, sample sizes, and $t/N_e$ ratios (drift intensities).

Supporting Information may be found in the online version of this article.

Please note: Wiley-Blackwell is not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.
Queries

Q1 Author: Please provide E-mail address for the corresponding author.

Q2 Author: “Oliveira et al. 2009” has not been included in the Reference List, please supply full publication details.

Q3 Author: “Oliveiera et al.” has not been cited in the text. Please indicate where it should be cited; or delete from the Reference List. Also provide year of publication in the same reference.