Evolution, 53(5), 1999, pp. 1506–1515

HIERARCHICAL COMPARISON OF GENETIC VARIANCE-COVARIANCE MATRICES. I. USING THE FLURY HIERARCHY

PATRICK C. PHILLIPS[1] AND STEVAN J. ARNOLD[2,3]
[1]*Department of Biology, Box 19498, University of Texas at Arlington, Arlington, Texas 76019-0498*
E-mail: pphillips@uta.edu
[2]*Department of Ecology and Evolution, University of Chicago, Chicago, Illinois 60637*

Abstract.—The comparison of additive genetic variance-covariance matrices (*G*-matrices) is an increasingly popular exercise in evolutionary biology because the evolution of the *G*-matrix is central to the issue of persistence of genetic constraints and to the use of dynamic models in an evolutionary time frame. The comparison of *G*-matrices is a nontrivial statistical problem because family structure induces nonindependence among the elements in each matrix. Past solutions to the problem of *G*-matrix comparison have dealt with this problem, with varying success, but have tested a single null hypothesis (matrix equality or matrix dissimilarity). Because matrices can differ in many ways, several hypotheses are of interest in matrix comparisons. Flury (1988) has provided an approach to matrix comparison in which a variety of hypotheses are tested, including the two extreme hypotheses prevalent in the evolutionary literature. The hypotheses are arranged in a hierarchy and involve comparisons of both the principal components (eigenvectors) and eigenvalues of the matrix. We adapt Flury's hierarchy of tests to the problem of comparing *G*-matrices by using randomization testing to account for nonindependence induced by family structure. Software has been developed for carrying out this analysis for both genetic and phenotypic data. The method is illustrated with a garter snake test case.

Key words.—Flury hierarchy, genetic correlation, matrix comparisons, principal components analysis, quantitative genetics, randomization test, statistical resampling.

Received September 17, 1998. Accepted May 6, 1999.

Evolution is an inherently multivariate phenomenon in which the functional and genetic interplay between traits can have a large impact on the both the direction and outcome of evolution (Lande 1988). The relationship between traits is often summarized statistically in the form of a variance-covariance matrix in which the variation for each individual trait is found on the main diagonal and the pattern of covariation among traits is described by the off-diagonal terms. Here, we will focus on the pattern of among-trait covariance from a quantitative genetic point of view, asking specifically to what extent the evolutionary process leads to changes in covariance pattern. This question can be addressed directly by the comparison of quantitative genetic variance-covariance matrices from descendant populations.

Comparison of genetic variance-covariance matrices (*G*-matrices; Lande 1979) is a nontrivial statistical problem and a variety of methods have been proposed as a solution (for a summary, see Roff 1997, p. 101ff). The two leading methods are matrix correlation (which tests the null model of no correlation between two matrices; e.g., Lofsvold 1986; Kohn and Atchley 1988) and maximum-likelihood tests for matrix equality (e.g., Shaw 1991). In this paper we develop and apply a third methodology that has some distinct advantages over previous methods. This new methodology is based on Flury's (1988) model of common principal components, which is built on the observation that matrices can be described and compared by their eigenvalues and eigenvectors (principal components). Unlike other methods, Flury's model provides a hierarchy of tests corresponding to a range of possible relationships among matrices (Fig. 1). This hierarchy provides tests of several other hypotheses besides matrix equality (e.g., matrix proportionality, common principal components). To apply Flury's method to the case of *G*-matrices, we have to account for the nonindependence in the data induced by family structure. We introduce randomization testing as a solution to this problem.

Building the Hierarchy

The purpose of the analysis presented here is to compare the structure of two or more covariance matrices in a hierarchical fashion. The hierarchy of comparisons is built upon the realization that covariance matrices can share more complex relationships between one another than just being equal or unequal (Flury 1988). For example, one matrix might be identical to another except that each element of the matrix is multiplied by a single constant. We would then say that the matrices are proportional. A more precise definition of proportionality is that the matrices share identical eigenvectors (or principal components), but their eigenvalues differ by a proportional constant. This suggests that another relationship between matrices could be that they share principal components in common, but their eigenvalues differ (the common principal component, or CPC, model). In this case, each of the elements of the eigenvectors for each matrix are identical. Similarly, the matrices could share one, two, or up to $p - 2$ (where the matrices have dimension $p \times p$) principal components in common out of the p total possible components. This is the partial principal components (PCPC) model. The PCPC model stops at $p - 2$ components because, as principal components are defined to be orthogonal to one another, if $p - 1$ of the components are known then the final one is already determined, yielding the full CPC model. The hierarchical nature of this set of comparisons can be appreciated by realizing that if two matrices share two principal components in common, then they necessarily share one com-

[3] Present address: Department of Zoology, Oregon State University, Corvallis, Oregon 97331; E-mail: arnolds@bcc.orst.edu.

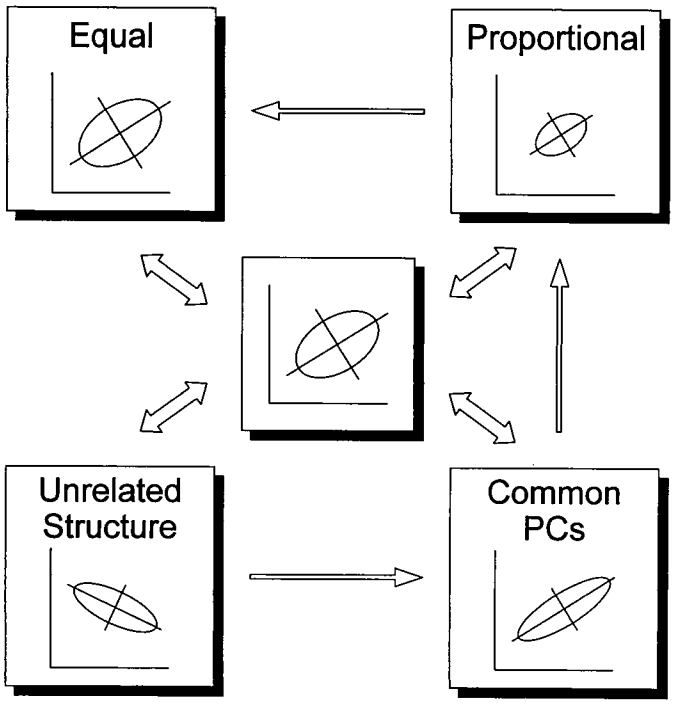

FIG. 1. Diagramatic representation of the Flury hierarchy. The diagram shows the hierarchy of comparisons possible for two covariance matrices, each with two characters. The covariance structure of the data is represented by ellipses with axis orientation representing principal components and with size along each axis representing the eigenvalues. The covariance pattern in the middle is compared to four possible similarity classes (double-headed arrows). The hierarchy of comparisons moves from unrelated structure up through equality (single-headed arrows).

ponent. Similarly, if two matrices are proportional, then they must satisfy the CPC model as well as all of the PCPC models. The hierarchy ends with matrix equality, which can only be true if all other elements of the hierarchy also hold (Figs. 1, 2).

Viewing the evolution of G-matrices within the context of the Flury hierarchy greatly expands the set of evolutionary questions that can be addressed in matrix-comparison studies. For example, most previous studies of G-matrix evolution have focused on whether the G-matrices are unchanged (equal) through time (e.g., Lofsvold 1986; Shaw 1991). Although this question is of central importance to the issue of reconstructing historical patterns of selection (Lande 1979; Arnold 1988; Lofsvold 1988; Turelli 1988), the evolution of G can be interesting in its own right. The structure of G can have an important influence on the direction of evolutionary change (Lande 1979; Felsenstein 1988; Zeng 1988; Schluter 1996) and may reflect the underlying pattern of developmental interactions and associations (Cheverud 1984; Maynard Smith et al. 1985). Here, we present tools that allow this underlying structure to be investigated. We will first provide the statistical context for evaluating the Flury hierarchy for quantitative genetic data and then demonstrate its application using a worked example. A companion paper (Arnold and Phillips 1999) uses these methods to more fully investigate G-matrix evolution within two populations of the garter snake, *Thamnophis elegans*.

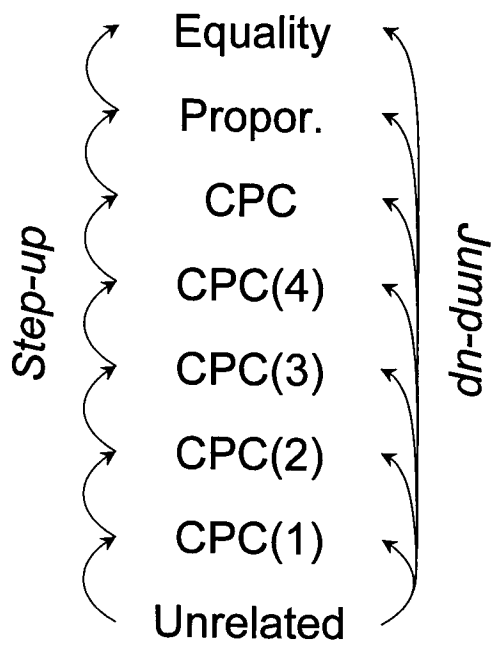

FIG. 2. A comparison of the different approaches to locating a set of matrices on the Flury hierarchy. This example assumes 6 × 6 matrices. In the step-up approach, each level of the hierarchy is evaluated in turn and the next step up is predicated on the previous lower step, which is used as the current null hypothesis. In the jump-up approach, each level of the hierarchy is tested against the lowest level (unrelated structure), yielding a single test for the appropriate level. This approach is based on the fact that for the current level of the hierarchy to be valid (e.g., CPC), all the lower levels of the hierarchy must also be true (e.g., CPC[4] and below). The step-up approach has the advantage that it shows the entire decomposition of the model (Flury 1988), whereas the jump-up approach has the advantage that it is simpler to interpret and ultimately involves reporting a single statistical test.

METHODS

Finding One's Place in the Hierarchy

Comparison Statistic.—The first step in comparing two or more matrices is creating a metric or statistic by which the comparison can be evaluated. A solution based on maximum-likelihood methods has been known for some time for the case of matrix equality (Anderson 1958). Basically, in this case each separate matrix is compared to the average of all of the matrices. The more different each matrix is from the average, the less likely it is that the matrices are equal to one another (the exact value of this likelihood depends on the underlying distribution of the traits—usually assumed to be multivariate normal). Flury (1987, 1988) greatly expanded on this approach by adding other levels of similarity to the comparisons (Fig. 1), although the overall approach remains the same. For each hypothesis in the hierarchy, a new set of matrices (based on the sample matrices) is constructed that are constrained so that the hypothesis in question is true. The relative degree of difference between the sample and constrained matrices determines the likelihood that that particular hypothesis is true. Because there are a number of different hypotheses in the hierarchy, the relative likelihood of any given hypothesis must be evaluated against the other hypotheses in the hierarchy. It is important to note that this

process does not consist of separately calculating the eigenvalues and eigenvectors and then comparing them in some way. Instead, the eigenstructure of the matrices is estimated simultaneously under the constrained hypothesis (Flury 1988).

Step-up Approach.—The various models in the hierarchy can be tested in several ways. First, and perhaps most logically, the hierarchy can be built in a stepwise fashion starting with no relation between the matrices (unrelated structure). From this one can go to CPC(1), then to CPC(2), etc., through CPC, proportionality, and equality (Figs. 1, 2). The likelihood that a particular model is valid is then tested against the next lowest model in the hierarchy (Fig. 2), the next step being taken if the current test in the hierarchy is nonsignificant. Flury (1988) has shown that this approach leads to a decomposition of the log-likelihood ratio used to test each step of the hierarchy, although it is currently unknown if the tests in this decomposition are completely independent of one another (Flury 1988, p. 151), because sums of chi-squares that are known to be stochastically dependent sometimes appear to be independent (Flury 1986).

We call this method the "step-up approach." This is the statistical-testing framework outlined in detail by Flury (1988). A difficulty with this approach is that if the lower model is not the correct model, then the significance tests involving this model will not have any real meaning (Flury 1988). This can sometimes make interpretation of where in the hierarchy the matrices belong difficult, because one is forced to decide whether a marginally nonsignificant test means that one should move up in the hierarchy. The following approach simplifies interpretation, but loses the elegance of the decomposition provided by the step-up approach.

Jump-up Approach.—An alternative testing procedure is that any model hypothesis can be tested against any other model that is lower in the hierarchy than the first hypothesis. For example, equality could also be tested against CPC (skipping proportionality) or directly against unrelated structure for a more traditional "equal-or-not" type test (this latter test is in fact equivalent to the standard test of matrix equality discussed above). The test of each level of the hierarchy against unrelated structure is probably the most logical from a hypothesis-testing point of view, which is how most biologists will probably approach these problems. Because each level of the hierarchy is inclusive of the lower levels, one can start at the bottom of the hierarchy and test each successive level against the hypothesis of unrelated structure until a significant deviation is encountered (Fig. 2). The move up the hierarchy should be stopped at this point. The statistic for the jump-up level is simply the sum of the likelihoods (χ^2-values) and degrees of freedom for the underlying steps. We call this method the "jump-up approach" because one jumps all the way from the bottom of the hierarchy to the model in question. One caveat to this approach is that the significance tests at each level are not independent, because the tests are not orthogonal. However, if the tests are build from the bottom up with a strong stopping rule (i.e., not going above the first significant difference rather than simply looking for highest level in the model with a nonsignificant *P*-value), then this nonindependence should not be too problematic, because it is the *P*-values at higher levels of the hierarchy that are dependent on those below.

Model-Building Approach.—A final approach is the one advocated by Flury (1988), in which the overall best-fitting model is taken. In this "model-building approach," the choice criterion is not statistical significance, but rather is how well a particular hierarchical model fits based on how much information is available to fit that model. Flury (1988, p. 151) suggests that the Akaike information criterion (AIC; Akaike 1973) can be used for this decision-making. The AIC balances the goodness of fit of a particular model (the log-likelihood statistic in this case) against the number of parameters used to fit the model. Models with more parameters tend to fit better out of necessity, so the best model in this scheme is chosen using a "penalized log likelihood," which is a simple difference between the likelihood and the number of parameters (Flury 1988, p. 152). The AIC is akin to the reciprocal of this function, so that the minimum AIC value represents the best fitting model.

Although Flury (1988) recommends this method over other possibilities, there are two difficulties in applying the AIC to quantitative genetic matrices. First, there is no statistical testing in this method, so it is impossible to address the question of whether one model fits the data significantly better than another model. Second, proper calculation of the AIC is dependent on the number of parameters in a model (i.e., a combination of sample size, matrix dimension, and the number of populations being compared), which cannot be precisely estimated in quantitative genetic designs. Estimating the appropriate value for the number of parameters is difficult in quantitative genetics experiments because the accuracy of the covariance components that are used to estimate the genetic variance-covariance matrices depend in a complex way on both the number of families and the numbers of individuals per family. None of these values can be substituted into the parametric version of the Flury hierarchy to yield sensible results (see below).

Choosing an Approach.—The correct approach to use in building the hierarchy is currently unclear and it is likely that there will be no single best approach for every situation. All three approaches usually yield the same qualitative answer, but not always (as we shall see, they can sometimes be several steps away from one another on the hierarchy). Often the biological "reality" of the situation can be gleaned by looking at how well the matrices constructed using the constrained model match the actual matrices from the populations. The model-building approach will serve best when the primary goal is parameter estimation, whereas the step-up approach highlights which point in the hierarchy causes the overall greatest effect. The jump-up approach is most applicable when a test of a single hypothesis is desired. We tend to rely on the jump-up approach because it provides the most straightforward way of understanding and interpreting the results.

Other Hierarchies.—Flury also presents a slightly alternative hierarchy called the common space model in which the focus is on a shared spherical subspace among some of the components. Although this is an important question in many aspects of multivariate statistics (e.g., in repeated measures analysis; Winer et al. 1991), it probably is not generally

applicable to quantitative genetic variation, and so is ignored here (see also Krzanowski 1979).

Statistical Tests

Randomization Test.—The problem with testing quantitative genetic data using this hierarchy is that the significance tests constructed by Flury (1988) rely on the likelihood statistic to be chi-square distributed, which in turn requires both multivariate normality and that the degrees of freedom under the null hypothesis is known. The former can be a problem for any data, but it is especially problematic for genetic data where we usually have little information about the distribution of breeding values. The latter problem is uniquely severe in variance component estimation procedures because it is not clear what the appropriate degrees of freedom for the comparison should be. For additive genetic variances and covariances, the appropriate number of degrees of freedom is related to the number of families or sires in the breeding design. However, the number of individuals per family also contributes to the error structure of the covariance estimates. In practice, the distributions do appear to be chi-square distributed, but with the inappropriate degrees of freedom (see below).

One solution to these problems is to use a statistical resampling approach. We have devised a randomization procedure in which families are randomly reassigned to populations before the comparisons are made (Phillips 1998a; see also Roff 1997, p. 106). The randomized populations satisfy the null hypothesis of similarity at each level of the hierarchy and can therefore be used as a null distribution against which the actual comparison statistic can be tested. The comparison statistic for the actual populations are then compared to the distribution of randomized populations to estimate the probability of obtaining a statistic that large just by chance. The appropriate population means are subtracted off each trait value before the randomization procedure so that the among-population variance does not confound the comparisons of within-population covariance structure.

An alternative testing procedure is to use a bootstrapping approach (B. Flury, pers. comm. 1993) in which each population is repeatedly compared to itself to build a null distribution (see also Zhang and Boos 1992, 1993; Paulsen 1996; Goodnight and Schwartz 1997). We have implemented such a procedure (Phillips 1998b), which gives results comparable to the randomization tests, but report only the randomization results here because of the appropriateness of the test and the greater ease of interpretation. However, bootstrapping is used for setting errors on the genetic estimates.

Singular Matrices.—Because of the nature of the maximum-likelihood matrix comparison statistics, the methods devised by Flury (1988) will not work on singular or nonpositive definite matrices (i.e., matrices with zero or negative eigenvalues). This is usually not a problem for product-moment-based matrices in which the matrices are guaranteed to be positive-definite. Matrices constructed using variance component estimates are known to often have negative eigenvalues, so this can be a real problem for quantitative genetic matrices (Hill and Thompson 1978). When negative eigenvalues occur during the randomization procedure, the most straightforward approach is to eliminate the offending matrices from the randomization sample. This elimination undoubtedly introduces some bias into the statistical test because it is likely that the most extreme matrices are those that are eliminated, although the degree of bias has yet to be determined. A more serious problem occurs when the initial population estimates result in singular matrices, thus precluding analysis. We use a "matrix-bending" procedure (Hayes and Hill 1981; Kirkpatrick et al. 1990) in which the eigenvalues of the matrix are adjusted just enough to eliminate the negative eigenvalues. All subsequent matrices in the resampling procedure are bent to the same degree. This bending creates another potential source of bias, although comparison with parametric approaches that do not require bending yield very similar results (see below). Maximum-likelihood estimators of the quantitative genetic parameters themselves may ultimately prove useful for overcoming these difficulties (Shaw 1987, 1991, 1992).

Tests on Individuals and Family Means.—Some of the sampling difficulties with covariance components can be overcome if these components are approximated using calculations of the variances and covariances from the population samples directly. For example, the genetic covariance between two traits can be estimated by the covariance among the family means (Via 1984), whereas the phenotypic covariance can be estimated by the covariance among individuals, while ignoring family structure. The advantage of these methods is that the statistical properties of these estimates are better understood, allowing direct tests of significance. Further, as long as every trait is measured on every individual/family, then the covariance matrices constructed from these estimates are guaranteed to be positive-definite. (They can still be singular, but this is unlikely unless some of the traits are simple functions of the other traits in the matrix.) Flury's (1988) methods and tests are designed for these types of estimates, and so in this case the hierarchy can be tested directly using parametric methods. The disadvantage of these estimates is that they are biased, sometimes substantially so (Roff and Preziosi 1994). The family-mean covariance is biased by the within-family covariance (Via 1984), whereas the among-individual covariance is biased by the among-family covariance. The individual and family-mean analysis is implemented in the program *CPC* (Phillips 1998c). The hierarchy for phenotypic data can also be tested using matrix manipulation packages (Klingenberg et al. 1996) or structural equation modeling routines (Dolan 1996). It seems prudent to analyze both the variance component and the family-mean estimates separately, which in turn yield nonparametric and parametric tests of the hierarchy. The former estimates have the advantage of being unbiased and not as subject to distributional assumptions, but have not been rigorously justified in this context, whereas the latter estimates are more precise and based on well-established methodology, but are subject to bias. Each approach has its own set of strengths and weaknesses, and therefore any similarities or differences in the results of these two approaches should provide separate insights into the actual structure of the matrices.

Calculating these statistical tests is computationally intensive, but not prohibitive. The time for analysis scales similarly to matrix inversion, or slightly less than the square of

TABLE 1. Genetic variance-covariance matrices (± SE) for female offspring from the inland population of *Thamnophis elegans*. Data from the unconstrained model (i.e., the normal analysis) are on the top line, whereas data constrained to fit the common principal component (CPC) model are on the second line. A description of the data and characters can be found in Arnold and Phillips (1999).

	VENT	SUB	MID	ILAB	SLAB	POST
VENT	8.17 ± 1.70 8.59 ± 2.00	3.78 ± 1.67 3.52 ± 1.67	0.09 ± 0.29 0.17 ± 0.33	0.65 ± 0.21 0.78 ± 0.33	0.07 ± 0.18 −0.06 ± 0.27	0.12 ± 0.20 0.18 ± 0.23
SUB		8.16 ± 1.73 7.60 ± 1.77	0.28 ± 0.40 0.15 ± 0.35	0.24 ± 0.31 0.25 ± 0.34	0.23 ± 0.21 0.14 ± 0.26	0.17 ± 0.23 0.21 ± 0.22
MID			0.27 ± 0.10 0.39 ± 0.13	0.08 ± 0.05 0.09 ± 0.07	0.05 ± 0.06 0.05 ± 0.07	−0.10 ± 0.05 −0.07 ± 0.06
ILAB				0.03 ± 0.05 0.24 ± 0.15	−0.02 ± 0.04 −0.03 ± 0.08	0.03 ± 0.05 0.07 ± 0.06
SLAB					0.02 ± 0.03 0.22 ± 0.12	0.02 ± 0.03 0.02 ± 0.04
POST						0.09 ± 0.05 0.27 ± 0.11

size of the matrices and the number of populations. Single runs of the *CPC* program on modern computer hardware take from several seconds for small matrices to four hours for a 40 × 40 matrix. Randomization tests multiply these values by the number of resampling runs conducted. Each randomization statistic reported here is the result of about a half day of computing time on a 200Mz PentiumPro computer.

EXAMPLE

To illustrate the principles laid out above, we will analyze a sample dataset in detail. The example is based on a quantitative genetic analysis of genetic covariation among scalation traits in the garter snake, *Thamnophis elegans*. A full description of the dataset is available in Arnold and Phillips (1999). The data consist of parent-offspring regression estimates of additive genetic variance and covariance among six traits for two snake populations, one from a coastal and the other from an inland site in northern California. There were 102 litters from the coastal population, 156 litters from the inland population, and about 10 offspring in each litter. The *G*-matrix estimates for female offspring from these populations are given in Tables 1 and 2. See Arnold and Phillips (1999) for an analysis of the phenotypic and environmental matrices, as well as for comparisons between sexes.

Simple inspection of the matrices (Tables 1, 2) shows that change in variance and covariance from one population to another are not uniform across characters. It is equally evident that patterns of similarity and difference cannot be obtained by mere element-by-element comparisons. However, using the hierarchical approach outlined above reveals a great deal of shared underlying structure across populations. Here, the hierarchy is built by beginning at the hypothesis of one shared principal component and ends at common principal component structure for the jump-up approach (Table 3). Results are consistent for both the randomization and parametric tests. The step-up and model-building approaches (Table 4) are somewhat less consistent, however. The randomization analysis supports the CPC result, whereas the parametric results suggest CPC(4), the partial common principal component model with four of the six possible components shared in common. Note that the difference here is between the jump-up and step-up approaches, not between the randomization and parametric results, because the parametric results themselves are different across the two methods. These two

TABLE 2. Genetic variance-covariance matrices (± SE) for female offspring from the coastal population of *Thamnophis elegans*. Data from the unconstrained model (i.e., the normal analysis) are on the top line, whereas data constrained to fit the common principal component (CPC) model are on the second line. A description of the data and characters can be found in Arnold and Phillips (1999).

	VENT	SUB	MID	ILAB	SLAB	POST
VENT	6.88 ± 1.97 5.36 ± 1.81	0.19 ± 1.66 0.43 ± 1.31	−0.04 ± 0.25 0.10 ± 0.22	1.15 ± 0.47 0.56 ± 0.34	0.30 ± 0.35 −0.09 ± 0.20	−0.04 ± 0.20 0.07 ± 0.17
SUB		7.80 ± 2.66 5.17 ± 1.58	−0.02 ± 0.30 0.04 ± 0.26	0.03 ± 0.48 −0.07 ± 0.35	0.12 ± 0.35 0.11 ± 0.21	−0.11 ± 0.25 0.11 ± 0.19
MID			0.01 ± 0.03 0.45 ± 0.17	0.04 ± 0.06 −0.03 ± 0.11	0.11 ± 0.06 0.07 ± 0.12	0.03 ± 0.04 0.10 ± 0.09
ILAB				0.36 ± 0.14 0.58 ± 0.30	0.47 ± 0.11 0.22 ± 0.14	0.00 ± 0.06 0.03 ± 0.10
SLAB					0.37 ± 0.12 0.56 ± 0.18	0.05 ± 0.06 0.05 ± 0.10
POST						0.10 ± 0.05 0.48 ± 0.16

TABLE 3. The Flury hierarchy for the comparison of genetic matrices for coastal and inland females using the jump-up procedure. At each step in the hierarchy a hypothesis is tested against the hypothesis at the bottom of the hierarchy, viz. Unrelated. Two estimates of the genetic matrix are compared: one based the variance-covariance of litter means (with parametric evaluation of sampling properties) and one based on offspring-mother regressions (with randomization evaluation of sampling properties).

Hierarchy	Litter-mean estimate (parametric) df	χ^2	P	Regression estimate (randomization) P
Equality	21	81.66	<0.0001	0.0005
Proportionality	20	73.77	<0.0001	0.0003
Full CPC	15	20.95	0.1384	0.3949
CPC(4)	14	14.65	0.4020	0.3719
CPC(3)	12	14.12	0.2931	0.3790
CPC(2)	9	7.10	0.6264	0.5107
CPC(1)	5	5.10	0.4035	0.3950
Unrelated				

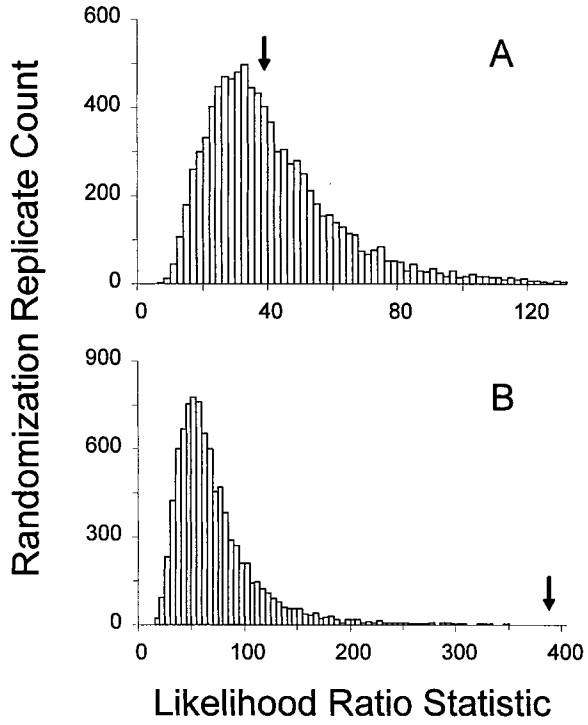

FIG. 3. Null distributions for tests of the Flury hierarchy using a randomization test for the comparison of genetic matrices for females from coastal and inland populations of *Thamnophis elegans*. (A) Test of the common principal component model. (B) Test of the proportionality model. *P*-values (Table 3) are calculated as the fraction of randomization replicates that exceed the likelihoods for the actual comparisons (arrows). The histograms show results from 10,000 randomization runs.

approaches can yield differences when one step in the hierarchy yields a large change (step-up) while the sum over all the effects is relatively small (jump-up). The multiple decisions that must be made while using the step-up procedure can make it difficult to interpret, but with the jump-up procedure there is a danger that an overall pattern of similarity might swamp more subtle patterns of difference when summing over all differences (see a similar discussion in Roff 1997, p. 111). Neither method is foolproof. We will use the CPC result because it is based on the covariance components and provides the simplest statistical interpretation. However, the model-building perspective suggests that parameter estimates based on CPC(4) will probably be better. Nevertheless, the qualitative result of overall similarity in matrix structure is the same no matter which approach is used.

The *P*-values from the randomization results are calculated from 10,000 resampling runs, yielding distributions of the form shown in Figure 3. These distributions have a chi-square shape (indeed they are almost normal), but the magnitude of the values on the x-axis do not match the expectation based on parametric calculations (Flury 1988). Thus, barring further theoretical developments, the randomization approach appears necessary for the proper analysis of the covariance component-based matrices. Again, there is a satisfying consistency between the parametric and randomization results in both in the level of the hierarchy chosen and in the approximate *P*-values calculated (Table 3).

TABLE 4. The Flury hierarchy for the comparison of genetic matrices for coastal and inland females using the step-up procedure. At each step in the hierarchy the hypothesis labeled "upper" is tested against the hypothesis on the step below, "lower." Two estimates of the genetic matrix are compared: one based the variance-covariance of litter means (with parametric evaluation of sampling properties) and one based on offspring-mother regressions (with randomization evaluation of sampling properties). The best solution under the model-building approach is indicated by the minimum value of Akaike information criterion (AIC) under the parametric estimates (CPC[4] in this case).

Hierarchy Upper	Lower	Litter-mean estimate (parametric) df	χ^2	P	AIC	Regression estimate (randomization) P
Equality	Proportionality	1	7.88	0.0050	81.6	0.2029
Proportionality	Full CPC	5	52.82	<0.0001	75.8	0.0002
Full CPC	CPC(4)	1	6.30	0.0121	33.0	0.5083
CPC(4)	CPC(3)	2	0.54	0.7648	28.7	0.5177
CPC(3)	CPC(2)	3	7.02	0.0714	32.1	0.5223
CPC(2)	CPC(1)	4	2.00	0.7357	31.1	0.5062
CPC(1)	Unrelated	5	5.10	0.4035	37.1	0.3992
Unrelated	—				42.0	
Total		21	81.66			

The adequacy of the CPC model for fitting the data is demonstrated by the correspondence between the estimated matrices and the matrices generated under the constrained hypothesis (Tables 1, 2). The fit is somewhat better for the inland population (Table 1) than for the coastal population (Table 2), probably because a high correlation between ILAB and SLAB required the coastal population to be bent before analysis. Element by element, there is a fairly good correspondence, given the errors in the estimates. The size of the errors do raise the concern that part of the pattern of shared similarity could be caused by a lack of power (see below), despite the relatively large number of litters sampled in both populations. A more detailed interpretation of the pattern of correlation among the traits is given in Arnold and Phillips (1999).

Discussion

The analysis of genetic variance-covariance structure using a hierarchical approach can reveal shared similarities across matrices that go well beyond the simple question of matrix equality (Fig. 1, Table 3). Given the central role that the *G*-matrix can play in understanding evolutionary processes (Lande 1988), an appropriate analysis of *G* itself should contribute to our understanding of multivariate evolution. The *G*-matrix must evolve on some sort of continuum. On short evolutionary time scales, we would expect *G* to show equality across populations, whereas on longer time scales, *G* must necessarily diverge as large-scale morphological evolution occurs (Lofsvold 1986). A hierarchical analysis of *G*-matrix structure allows for a finer-scale analysis of how this divergence is generated. Would we expect divergence in *G* to follow the steps down the hierarchy? Clearly, as two populations diverge, equality of *G* will be lost at some point, but it is not clear at what time scale this will occur. Also, it is not necessary that populations should maintain common matrix structure as they diverge, so the presence of shared structure provides an evolutionary and/or genetic signal that warrants further study. Under restricted circumstances, selection can maintain common structure throughout the process of divergence (Phillips, unpubl. data), whereas genetic drift is expected to generate proportional matrices (Lande 1979), although there is likely to be a large amount of variance around this average (P. C. Phillips, M. C. Whitlock, and K. Fowler, unpubl. data). As more studies begin to explore the evolution of *G*-matrix structure, questions relating to the time scale of divergence and the way that structure is maintained can begin to be addressed (e.g., Pfrender 1998).

Further, the inherent structure of *G* is interesting in its own right. For instance, it is possible that multivariate divergence in population means might follow the direction determined by the principal component structure of *G* (Lande 1979; Zeng 1988; Schluter 1996). It has also been argued that the structure of *G* can reflect underlying developmental processes and the nature of morphological integration (Cheverud 1982, 1984). As a practical matter, statistical techniques that rely on covariance structure such as distance metrics (Krzanowski 1996), canonical variates (Neuenschwander and Flury 1995), and discriminant functional analysis (Flury and Schmid 1992) can be affected by divergence in covariance matrices among populations, yet the preserved common structure can be used as a source of additional information.

A number of studies have demonstrated the utility of the Flury hierarchy in phenotypic analyses (Airoldi and Flury 1988; Klingenberg and Zimmermann 1992; Klingenberg and Spence 1993; Klingenberg et al. 1996; Steppan 1997). In particular, Steppan (1997) has examined the difficult question of how phenotypic covariance structure evolves in a phylogenetic context. This approach will be crucial to understanding the pattern and timescale of *G* divergence. The methods presented here will need to be combined with a more rigorous comparative approach (Felsenstein 1985; Hansen and Martins 1996) before this question can be adequately addressed. Klingenberg et al. (1996; see also Flury and Neuenschwander 1985; Neuenschwander 1991) have extended the CPC approach to even more complexly patterned matrices. For example, in studies of development (e.g., Cheverud et al. 1983; Atchley 1987) and genotype-by-environment interactions (Via 1984; Via and Lande 1985; Gomulkiewicz and Kirkpatrick 1992) one is concerned with both correlations among traits within a single age or environment and correlations across ages or environments. This increase in trait number greatly enlarges the *G*-matrix and generates natural partitions of within and between age/environment relationships within the matrix. We might expect the structure of the within-class covariances to be different from the between-class covariances, but these separate structures can in turn share similarities with one another (Klingenberg et al. 1996). Expanding the current analysis for quantitative genetic data to include these sorts of models is in principle straightforward, although in practice it will take a fair degree of computational work.

Covariances versus Correlation

Measuring the association between traits using covariances provides a natural scale for evolutionary analysis (Lande 1979). However, it is well known that the results from principal component analysis are strongly influenced by measurement scale. Traits measured on larger scales tend to have larger variances and therefore tend to cluster by themselves in the analysis. A common solution to this scaling problem is to standardize the variables so that the analysis is conducted on the correlation structure instead of the covariance structure (e.g., Kohn and Atchley 1988). Unfortunately, although they are related to one another, covariance structure and correlation structure can be substantially different. Indeed, it is not difficult to construct two genetic variance-covariance matrices that yield the same genetic correlations yet diverge substantially in their response to selection (K. Spitze, pers. comm. 1995).

There is nothing to prevent the application of the Flury hierarchy to correlation matrices, although the hypothesis of proportionality does not really apply in this case. Also, the sampling properties of correlations are quite different from covariances, so care should be exercised when using the parametric approaches. The problem with changes in scale is that principal component grouping of the traits can differ from analysis to analysis. In particular, the order of the principal components is strongly dependent on the total variance of

the traits being analyzed. Although not discussed by Flury (1988), the common principal component hypotheses need not be constrained to follow any particular ordering rule. Principal components have no inherent order, although they are traditionally ranked according to the magnitude of their associated eigenvalues. There is no biological basis for believing that the principal components with the largest eigenvalues are more likely to have conserved structure than principal components with smaller eigenvalues (although those with larger eigenvalues will tend to be better estimated). Therefore, the scaling problem can be eliminated to some extent by formulating hypotheses of similarity on biological grounds rather than on the more arbitrary rankings based on scale. For example, there may be functional or selective reasons to expect a particular set of traits to maintain their covariance relationships, even if the total amount of variance explained by those traits is lower than other traits (indeed this may often be the case). A complication of changing order is that the principal components with larger eigenvalues are likely to have significantly smaller errors associated with them, so that the lower-order principal components might be found to be similar simply on the basis of measurement error. The full implications of PC ordering have yet to be worked out. Because of the role they play in evolutionary analysis, we advocate testing the covariances themselves on their natural scale, with careful attention being paid to the nature of the hypotheses under study.

Null Hypotheses and Statistical Power

One potentially confusing aspect of the way in which we have described the Flury hierarchy is that although we have presented building the hierarchy from the bottom up, statistical power in the analysis actually comes from the top down. Beginning at the top of the hierarchy with equality, each successive step down the hierarchy involves adding more parameters to fit the matrices (e.g., proportionality requires one additional parameter per population, the proportionality constant, in addition to the maximum-likelihood estimate for the best pooled matrix from equality). Statistical testing is based on evaluating whether adding an additional parameter improves the fit of the model, and therefore the overall null hypothesis in these comparisons is matrix equality. Experiments with insufficient power will therefore tend to find matrix similarity when in fact there is none. Although power flows from the top down, our hypothesis-testing framework is from the bottom up, which is the appropriate framework given the way that matrix similarity is actually built. A cynical view of this approach would suggest that the location in the hierarchy is determined by a balance between Type I and Type II statistical errors; the point in which one has sufficient power to reject a hypothesis of certain types of similarity (e.g., equality) while not being able to reject more complex measures of similarity (e.g., CPC). Although there is undoubtedly some truth to this idea, in practice, comparing matrices in this hierarchical fashion appears to reflect the true underlying similarities between matrices, and it is actually not too difficult to show that a set of matrices do not share any common structure (e.g., Steppan 1997). Power analysis of the Flury hierarchy remains an unsolved problem at this point, especially for quantitative genetic data, but one that should not prove to be difficult to address (for a power analysis of individual correlation coefficients, see Phillips 1998d).

The appropriateness of the randomization procedure also needs more critical evaluation, and this analysis is currently being undertaken. Nevertheless, it is very encouraging that the parametric tests, which are based on a number of unlikely assumptions, and the randomization tests, which do not make these assumptions but need more statistical validation, usually give the same qualitative answer. As discussed above, each type of test will have its own power properties, and Arnold and Phillips (1999) provide an example of when the increased power of the parametric tests may yield slightly more discrimination among matrices. However, power can not be the sole criterion for statistical adequacy because with sufficient power it is likely that any two sample matrices will be found to differ in many respects. Ultimately, the reasonableness of the results depends on an inspection of the matrices themselves, their principal component structure, and the results from the Flury hierarchy under the constrained hypothesis.

Other Techniques for Comparing Matrices

Flury's (1988) hierarchical approach to matrix comparisons provides reconciliation for two opposing positions that have developed within the field of quantitative genetics. Matrix correlation was used in the first quantitative comparisons of G-matrices (Lofsvold 1986; Kohn and Atchley 1988; Cowley and Atchley 1992). Objections were raised to this approach on two grounds (Shaw 1991, 1992). First, the sampling distribution of the matrix correlation is complicated by the fact that the same set of families is used to estimate each element in the G-matrix. This family structure in the data induces sampling correlations between matrix elements. Proponents of matrix correlation tried to circumvent this problem by permutation tests in which the labels on the rows and columns of one G-matrix are shuffled, the correlation is calculated, and so on. Opponents of matrix correlation argued that such permutation tests were inappropriate because the labels correspond to traits, rather than to a set of commensurate identities that could be shuffled at random (e.g., a set of population names). Second, matrix correlation usually tests the null hypothesis that the correlation is zero (r_M = 0). Opponents have argued on evolutionary grounds that the appropriate null hypothesis is equality of G-matrices, which is related to the hypothesis $r_M = 1$. Maximum-likelihood ratio tests were put forward as a direct way to test for equality of G-matrices (Shaw 1987, 1991). Using Flury's (1988) perspective, we can see these two camps at opposite ends of a hierarchy of tests (Cowley and Atchley 1992). For example, in the snake data presented here, the correlation between the matrices is extremely high (r_M = 0.94), yet the matrices are not equal to one another. Information about the shared structure of the matrices is lost using both approaches.

A variety of interesting tests reside in the middle of the hierarchy. In an illuminating discussion of the history of matrix comparisons, Flury (1988) points out that many tests are extrapolations from the univariate to the multivariate case.

In the univariate case there are only two outcomes: either the variances are the same or they are different. In the multivariate case, however, a whole range of possibilities exists between equality and difference. In particular, the set of comparisons involving principal components has no counterpart in the single trait case. In this sense the Flury hierarchy provides a truly multivariate perspective on the problem of comparing variance-covariance matrices. The power of the hierarchical approach is that it enables us to move beyond the simple issue of whether the *G*-matrices differ and to ask in what ways they differ.

How can we measure the degree of difference between two matrices? There seems to be no universal answer to this question. No known metric is a silver bullet. Matrix correlation, likelihood ratios, or total χ^2 each give a single answer to the question of how much matrices differ. The problem with all of these single answers is that they cannot reflect the diversity of ways in which pairs of matrices can differ. Graphical approaches (Campbell 1981; Arnold 1992) are plagued by this same difficulty. The *G*-matrix is multivariate and so too is the difference between any pair of *G*-matrices.

Acknowledgments

We wish to thank B. Flury for software and many helpful comments and suggestions as to how to best implement his methods. Two anonymous reviewers provided a number of very useful suggestions that we have freely incorporated. This work was supported by National Science Foundation grants BSR-8111489, BSR-8918581, and BSR-9119588 to SJA and BIR-9612469 and DBI-9722921 to PCP.

Literature Cited

Airoldi, J.-P., and B. K. Flury. 1988. An application of common principal component analysis to cranial morphometry of *Microtus californicus* and *M. ochrogaster* (Mammalia, Rodentia). J. Zool. Lond. 216:21–36.

Akaike, H. 1973. Information theory and an extension of the maximum-likelihood principle. Pp. 267–281 *in* B. N. Petrov and F. Csaki, ed. Second international symposium on information theory. Akademiai Kiado, Budapest.

Anderson, T. W. 1958. An introduction to multivariate statistical analysis. Wiley, New York.

Arnold, S. J. 1988. Quantitative genetics and selection in natural populations: microevolution of vertebral numbers in the garter snake *Thamnophis elegans*. Pp. 619-636 *in* B. S. Weir, E. J. Eisen, M. M. Goodman, and G. Namkoong, eds. Proceedings of the second international conference on quantitative genetics. Sinauer, Sunderland, MA.

———. 1992. Constraints on phenotypic evolution. Am. Nat. 140: S85–S107.

Arnold, S. J., and P. C. Phillips. 1999. Hierarchical comparison of genetic variance-covariance matrices. II. Coastal-inland divergence in the garter snake, *Thamnophis elegans*. Evolution 53: 1516–1527.

Atchley, W. R. 1987. Developmental quantitative genetics and the evolution of ontogenies. Evolution 41:316–330.

Campbell, N. A. 1981. Graphical comparison of covariance matrices. Aust. J. Sci. 23:21–37.

Cheverud, J. M. 1982. Phenotypic, genetic, and environmental morphological integration in the cranium. Evolution 36:499–516.

———. 1984. Quantitative genetics and developmental constraints on evolution by selection. J. Theor. Biol. 110:155–171.

Cheverud, J. M., J. J. Rutledge, and W. R. Atchley. 1983. Quantitative genetics of development: genetic correlations among age-specific trait values and the evolution of ontogeny. Evolution 37:895–905.

Cowley, D. E., and W. R. Atchley. 1992. Comparison of quantitative genetic parameters. Evolution 46:1965–1967.

Dolan, C. 1996. Principal component analysis using LISREL 8. Struc. Eq. Model. 3:307–322.

Felsenstein, J. 1985. Phylogenies and the comparative method. Am. Nat. 125:1–15.

———. 1988. Phylogenies and quantitative genetics. Annu. Rev. Ecol. Syst. 19:445–471.

Flury, B. 1986. On sums of random variable and independence. Am. Stat. 40:214–215.

———. 1987. A hierarchy of relationships between covariance matrices. Pp. 31–43 *in* A. K. Gupta, ed. Advances in multivariate statistical analysis. Reidel, Boston, MA.

———. 1988. Common principal components and related multivariate models. Wiley, New York.

Flury, B. D., and B. E. Neuenschwander. 1985. Principal component models for patterned covariance matrices, with applications to canonical correlation analysis of several sets of variables. Pp. 179–206 *in* W. J. Krzanowski, ed. Recent advances in descriptive multivariate analysis. Oxford Univ. Press, Oxford.

Flury, B. D., and M. J. Schmid. 1992. Quadratic discriminant functions with constraints on the covariance matrices some asymptotic results. J. Multivar. Anal. 40:244–261.

Gomulkiewicz, R., and M. Kirkpatrick. 1992. Quantitative genetics and the evolution of reaction norms. Evolution 46:390–411.

Goodnight, C. J., and J. M. Schwartz. 1997. A bootstrap comparison of genetic covariance matrices. Biometrics 53:1026–1039.

Hansen, T. F., and E. P. Martins. 1996. Translating between microevolutionary process and macroevolutionary patterns: the correlation structure of interspecific data. Evolution 50:1404–1417.

Hayes, J. F., and W. G. Hill. 1981. Modification of estimates of parameters in the construction of genetic selection indices ("bending"). Biometrics 37:483–493.

Hill, W. G., and R. Thompson. 1978. Probabilities of non-positive definite between-group or genetic covariance matrices. Biometrics 34:429–439.

Kirkpatrick, M., D. Lofsvold, and M. Bulmer. 1990. Analysis of the inheritance, selection, and evolution of growth trajectories. Genetics 124:979–993.

Klingenberg, C. P., and J. R. Spence. 1993. Heterochrony and allometry: lessons from the water strider gene *Limnoporus*. Evolution 47:1834–1853.

Klingenberg, C. P., and M. Zimmermann. 1992. Static, ontogenetic, and evolutionary allometry: a multivariate comparison in nine species of water striders. Am. Nat. 140:601–620.

Klingenberg, C. P., B. E. Neuenschwander, and B. D. Flury. 1996. Ontogeny and individual variation: analysis of patterned covariance matrices with common principal components. Syst. Biol. 45:135–150.

Kohn, L. A. P., and W. R. Atchley. 1988. How similar are genetic correlation structures? Data from mice and rats. Evolution 42: 467–481.

Krzanowski, W. J. 1979. Between-groups comparison of principal components. J. Am. Stat. Assoc. 74:703–707.

———. 1996. Rao's distance between normal populations that have common principal components. Biometrics 52:1467–1471.

Lande, R. 1979. Quantitative genetic analysis of multivariate evolution, applied to brain:body size allometry. Evolution 33:402–416.

———. 1988. Quantitative genetics and evolutionary theory. Pp. 71–84 *in* B. S. Weir, E. J. Eisen, M. M. Goodman, and G. Namkoong, ed. Proceeding of the second international conference on quantitative genetics. Sinauer, Sunderland, MA.

Lofsvold, D. 1986. Quantitative genetics of morphological differentiation in *Peromyscus*. I. Test of the homogeneity of genetic covariance structure among species and subspecies. Evolution 40:559–573.

———. 1988. Quantitative genetics of morphological differentiation in *Peromyscus*. II. Analysis of selection and drift. Evolution 42:54–67.

Maynard Smith, J., R. Burian, S. Kauffman, P. Albrech, J. Campbell, B. Goodwin, R. Lande, D. Raup, and L. Wolpert. 1985. Developmental constraints and evolution. Q. Rev. Biol. 60:265–287.

Neuenschwander, B. 1991. Common principal components for dependent random vectors. Ph.D. diss. University of Bern, Bern, Switzerland.

Neuenschwander, B. E., and B. D. Flury. 1995. Common canonical variates. Biometrika 82:553–560.

Paulsen, S. M. 1996. Quantitative genetics of wing color pattern in the buckeye butterfly (*Precis coenia* and *Precis evarete*): evidence against the constancy of *G*. Evolution 50:1585–1597.

Pfrender, M. E. 1998. Evolutionary dynamics of molecular and quantitative genetic variation in ephemeral pond populations of *Daphnia pulex*. Ph.D. diss. University of Oregon, Eugene, OR.

Phillips, P. C. 1998a. CPCrand: randomization test of the CPC hierarchy. Univ. of Texas at Arlington. Software available at www.uta.edu/biology/phillips/software.

———. 1998b. H2boot: bootstrap estimates and tests of quantitative genetic data. Univ. of Texas at Arlington. Software available at www.uta.edu/biology/phillips/software.

———. 1998c. CPC: common principal components analysis. Univ. of Texas at Arlington. Software available at www.uta.edu/biology/phillips/software.

———. 1998d. Designing experiments to maximize the power of detecting correlations. Evolution 52:251–255.

Roff, D. A. 1997. Evolutionary quantitative genetics. Chapman and Hall, New York.

Roff, D. A., and R. Preziosi. 1994. The estimation of the genetic correlation: the use of the jackknife. Heredity 73:544–548.

Schluter, D. 1996. Adaptive radiation along genetic lines of least resistance. Evolution 50:1766–1774.

Shaw, R. G. 1987. Maximum-likelihood approaches applied to quantitative genetics of natural populations. Evolution 41:812–826.

———. 1991. The comparison of quantitative genetic parameters between populations. Evolution 45:143–151.

———. 1992. Comparison of quantitative genetic parameters: reply to Cowley and Atchley. Evolution 46:1967–1969.

Steppan, S. J. 1997. Phylogenetic analysis of covariance structure. I. Contrasting results from matrix correlations and common principal component analyses. Evolution 51:571–586.

Turelli, M. 1988. Phenotypic evolution, constant covariances, and the maintenance of additive genetic variance. Evolution 42: 1342–1347.

Via, S. 1984. The quantitative genetics of polyphagy in an insect herbivore. II. Genetic correlations in larval performance within and among host plants. Evolution 38:896–905.

Via, S., and R. Lande. 1985. Genotype-environment interaction and the evolution of phenotypic plasticity. Evolution 39:505–522.

Winer, B. J., D. R. Brown, and K. M. Michels. 1991. Statistical Principles in Experimental Design. McGraw Hill, New York.

Zeng, Z.-B. 1988. Long-term correlated response, interpopulation covariation, and interspecific allometry. Evolution 42:363–374.

Zhang, J., and D. D. Boos. 1992. Bootstrap critical values for testing homogeneity of covariance matrices. J. Am. Stat. Assoc. 87:425–429.

———. 1993. Testing hypothesis about covariance matrices using bootstrap methods. Commun. Statist.-Theory Meth. 22:723–739.

Corresponding Editor: L. Leamy