Mixed Effects Tests of Covariance Structure in Neutral Evolution Inference

Kevin H. Eng* and Sündüz Keleş**

Department of Biostatistics and Medical Informatics

Department of Statistics

University of Wisconsin-Madison

Madison, Wisconsin 53706

U.S.A.

**email:* eng@stat.wisc.edu

***email:* keles@stat.wisc.edu

Summary: Comparative methods study the relationship between phenotypes in the context of expected statistical dependence which arises from common ancestry. Because this phylogenetic dependence is inferred from a tree structured graph and a probability model, the lack of fit of this evolutionary model can signal evidence of adaptive changes driven by selection and can identify traits that warrant further study.

We discuss the use of two goodness-of-fit type hypothesis tests that test for differences in covariance structure. One of these tests is valid for a null hypothesis which allows unknown scaling between error and within-group variances, emphasizing sensitivity to changes in covariance structure. We characterize the power of these tests to detect common alternatives and provide diagnostics to identify low power cases.

Key words: Comparative method; Hypothesis testing; Mixed effects; Natural selection; Neutral drift.

1. Introduction to testing in comparative methods

Phenotypic measurements made across a set of species, strains or populations may be analyzed in a comparative data framework. This analysis must account for an expected dependence structure arising from the common ancestry of the components of a multivariate vector under a phylogenetic evolutionary model (Felsenstein, 1985). Given the covariance implied by such a model, most statistical methods are simply a form of generalized least squares (Grafen, 1989; Martins and Hansen, 1997) regressing one trait on another (Lynch, 1991; Ives et al., 2007).

Recently, it is of interest to test the fit of the expected covariance structure to a trait (Fay and Wittkopp, 2008). Under a Brownian model of evolution (Felsenstein, 1985), continuous traits whose variation may be predicted by a given phylogenetic structure are said to undergo neutral evolution; deviations from neutral expectations are attributed to the influence of selection. That is, by examining the constraints on observed variation (Rifkin et al., 2003; Nuzhdin et al., 2004; Whitehead and Crawford, 2006b), the investigator may infer the functional implication of a trait or its effect on fitness. For traits whose mode of evolution is unknown, for example gene expression traits, this inference may identify a particular model of evolution (Gu, 2004; Oakley et al., 2005).

Most existing tests are based on numeric procedures or asymptotic likelihood ratio tests (LRT). Butler and King (2004) developed an evolution model based on the Ornstein-Uhlenbeck process (Cox and Miller, 1965; Felsenstein, 1988; Hansen, 1997) and proposed a test of a single selection parameter. Allowing phylogenetic signal to attenuate, Freckleton et al. (2002) developed a LRT to detect phylogenetic correlation. Alternatives to large sample assumptions include permutation testing, Blomberg et al. (2003) used permutation testing to identify whether a fitted topology matches the data; and detailed computer simulations to form a reference distribution (Garland et al., 1993).

While parametric methods rely on maximum likelihood estimation, the accuracy of the asymptotic reference distributions are not clear (Hill and Thompson, 1978). More specifically, parametric methods rely on the difficult problem of successfully estimating a matrix-valued covariance parameter (Housworth et al., 2004). Comparative data are most appropriately modeled by a linear mixed model structure (Lynch, 1991; Guo et al., 2007; Felsenstein, 2008), so the usual intuition developed from multivariate methods based on random samples of complete vectors may not apply. Even if these references are appropriate, the parametric models have so far been derived for a specific type of selection model and are only able to detect a single type of alternative.

In this article, we develop hypothesis tests under a goodness-of-fit perspective: we sacrifice power to detect specific alternatives to favor a test able to find a broad classes of differences. Our tests have closed form reference distributions, which allows their implementation free of complex computer simulation, and do not rely on covariance estimation permitting their general use.

In Section 2, we describe the comparative data structure through a mixed effects model. We emphasize two different null hypotheses for testing in Section 3 and Section 4. To characterize specific alternatives, we define a measure of effect size in Section 5. We illustrate the testing framework in Section 6 with both simulated and live data.

2. Mixed effects model for comparative data

In this section, we discuss the statistical model underlying the inference of evolutionary models and develop hypothesis tests under different sets of assumptions. We assume that we have observed a continuous trait for $n_1, n_2, \ldots, n_m$ individuals from m distinct groups with the expectation of a common ancestor. Our goal is to test a given $m \times m$ matrix, $\Sigma = \Sigma(\theta)$, depending on parameters θ, that represents the covariance between these groups.

The structure of $\Sigma(\theta)$ will come from the comparative model applied to a phylogenetic

tree (Felsenstein, 1985). Briefly, the root of the tree graph represents the common ancestor and tips of the tree are the m groups. One imagines that the edges of the graph represent conditionally independent Brownian motion processes modeling the change in the trait over evolutionary time. We give concrete examples of these matrices in Section 5.

We model the observations as an $n \times 1$ vector $Y = \mu 1_n + Zb + e$, where μ is a common mean parameter, Z is an incidence matrix identifying group membership which satisfies $Z1_m = 1_n$ and $Z^T Z = diag(n_1, \ldots, n_m)$. The random effect $b \sim \mathcal{N}(0, \sigma^2\Sigma(\theta))$ captures variation across groups and $e \sim \mathcal{N}(0, \sigma^2 I_n)$ is the measurement error random effect for known $n \times n$ covariance I_n.

Marginally, Y is normally distributed with $\mathrm{E}(Y) = \mu 1_n$ and $\mathrm{Var}(Y) = \sigma^2 Z\Sigma(\theta)Z^T + \sigma^2 I_n$. The likelihood for Y is

$$\begin{aligned} L(\mu, \sigma^2, \theta; Y) =& (2\pi\sigma^2)^{-\frac{n}{2}} \left| Z\Sigma(\theta)Z^T + I_n \right|^{-\frac{1}{2}} \times \\ & \times \exp\left\{ -\frac{1}{2\sigma^2} (Y - \mu 1_n)^T \left(Z\Sigma(\theta)Z^T + I_n \right)^{-1} (Y - \mu 1_n) \right\}, \end{aligned}$$

where $|\cdot|$ denotes the determinant. We denote the group means as $M = (Z^T Z)^{-1} Z^T Y$ and the residual sum of squares orthogonal to the group subspace as $R^T R = Y^T \left(I_n - Z(Z^T Z)^{-1} Z^T \right) Y$. It can be shown that these are independent and have the following distributions (Seber and Lee, 2003)

$$\begin{aligned} M &\sim \mathcal{N}\left(\mu 1_m, \sigma^2\Sigma(\theta) + \sigma^2\Omega \right), \\ R^T R &\sim \sigma^2 (n - m) \chi^2_{n-m}, \end{aligned}$$

where we write $\Omega = (Z^T Z)^{-1}$.

We write M as M_n to emphasize dependence on the number of observations and, under balanced replicates, we note that the form of the Ω matrix is $\Omega_n = \frac{m}{n} I_m$. By the usual properties of normal random variables, we may write M_n as $M_n = M_\infty + \sigma \Omega_n^{1/2} e_n$ where $M_\infty \sim \mathcal{N}(\mu 1_m, \sigma^2\Sigma(\theta))$, and $e_n \sim \mathcal{N}(0, I_m)$. That is, as $n \to \infty$, $M_n \to_d M_\infty$. So, even

with a large number of replicates, we have effectively one vector of data with which to make inferences about $\Sigma(\theta)$.

3. Point null hypothesis

We wish to test $H_0 : \Sigma(\theta) = \Sigma_0$ where Σ_0 is a known covariance matrix. Under the null hypothesis, a statistic whose exact distribution is known is

$$F_n = \frac{n-m}{m-1} \frac{M^T K^T \left(K(\Sigma_0 + \Omega)K^T\right)^{-1} KM}{Y^T(I_n - Z(Z^TZ)^{-1}Z^T)Y},$$

where K is a $(m-1) \times m$ matrix satisfying $K1_m = 0$. Under $H_0 : \Sigma(\theta) = \Sigma_0$, it follows that

$$KM \sim \mathcal{N}\left(0, \sigma^2 K(\Sigma_0 + \Omega)K^T\right),$$
$$\frac{M^T K^T \left(K(\Sigma_0 + \Omega)K^T\right)^{-1} KM}{m-1} \sim \sigma^2 \chi^2_{m-1},$$

and that, $F_n \sim F_{m-1,n-m}$ under H_0. Writing the αth quantile of $F_{m-1,n-m}$ as $F^{\alpha}_{m-1,n-m}$, a size α, one-sided test may be constructed by considering a test with rejection region $\{F_n > F^{1-\alpha}_{m-1,n-m}\}$ or $\{F_n < F^{\alpha}_{m-1,n-m}\}$ and a two-sided test from considering the complement of the interval $\{F^{\alpha/2}_{m-1,n-m} < F_n < F^{1-\alpha/2}_{m-1,n-m}\}$.

4. Compound null hypothesis

The first test assumes that the error variance $\sigma^2 > 0$ is constant for both the variation in group means $\Sigma(\theta)$ and the error portion Ω. Allowing for scale differences leads to the compound null hypothesis: $H_0 : \Sigma(\theta) = \theta\Sigma_0$ where $\theta > 0$ is a relative rate between the group and error variances. The F test is not expected to work under this null hypothesis as it does not account for the second scale θ.

We may use the following observation to construct a statistic that is insensitive to θ under this compound null hypothesis. Suppose the $p \times 1$ vector X is normally distributed with mean zero and variance $\vartheta\Omega$, where $\vartheta > 0$. Statistics of the form $Q(A_1, A_2) = X^T A_1 X / X^T A_2 X$,

$A_2 \neq 0$, are ancillary to ϑ:

$$\begin{aligned}\Pr(X^T A_1 X / X^T A_2 X \leqslant t) &= \Pr\left(\frac{X^T (A_1 - tA_2) X}{\vartheta} \leqslant 0\right) \\ &= \Pr\left(Z^T \Omega^{1/2} (A_1 - tA_2) \Omega^{1/2} Z \leqslant 0\right),\end{aligned}$$

where Z has a standard multivariate normal distribution. The distribution of this type of multivariate quadratic form has been studied (Wilson and Hilferty, 1931; Imhof, 1961; Gupta and Nagar, 2000).

Under this compound null hypothesis, $Var(M_\infty) = \sigma^2 \theta \Sigma_0$, so we expect the distribution of statistics $Q(A_1, A_2)$ constructed from KM to not depend on rates σ^2 or θ. We re-purpose the numerator of F_n to choose A_1 in the following statistic.

Letting $A_1 = \left(K(\Sigma_0 + \Omega)K^T\right)^{-1}$ and $A_2 = I_{m-1}$, we have

$$Q_n = \frac{M^T K^T \left(K(\Sigma_0 + \Omega)K^T\right)^{-1} KM}{M^T K^T K M}.$$

We may construct one and two sided tests from Q_n using the following lemma. We consider the limiting distribution of Q_n in the sense that as n increases, $M_n \to_d M_\infty$.

LEMMA 1: *The limiting distribution of Q_n is*

$$Pr\left(Q \leqslant t\right) = Pr\left(\left[\sum_{j=1}^{m-1} d_j W_j\right]^{-1} \leqslant t\right),$$

where d_i are eigenvalues of $K\Sigma_0 K^T$ and $W = (W_1, \ldots, W_{m-1})$ are Symmetric Dirichlet (1/2) where $\sum_{j=1}^{m-1} W_j = 1$.

Proof. See appendix.

This Q_n test is similar, in spirit, to the K statistic proposed in Blomberg et al. (2003) which compares the ratios of observed neutral model and independent model mean squared errors to their expected values. However, inference about their statistic is based on a simulated reference distribution.

Subsequently, we investigate the small sample properties of our tests by simulation. We

may observe immediately that as $\theta \to 0$, we approach cases where $\Sigma(\theta)$ is unidentifiable from zero. We might expect, then, that the distribution of group means might deviate from M_∞ when θ is small and the test's validity will suffer. These cases correspond to low heritability traits which are of decreasing importance; in practice, we may estimate the heritability under the assumed comparative model. We investigate the validity of the test under these challenging settings in the simulation study section.

5. Characterizing effect size

We have generally adopted a goodness of fit type philosophy by constructing our tests to be valid under their specified null hypotheses without respect to a specified alternative. While we expect these tests to be powered for a broad set of alternatives to avoid the problem of guessing a particular one, it is useful to know what kind of covariance differences can be detected.

A practical measure of difference between the covariance of normally distributed vectors may be considered by computing the power of the uniformly most powerful (UMP) test using the Neymann-Pearson Lemma (Neyman and Pearson, 1933). Throughout the article, we use this UMP test power to measure the difference between two matrices. Writing $g(m; \Sigma)$ to be the density of M_∞ given covariance Σ, the UMP test of $H_0 : \Sigma = \Sigma_0$ versus $H_1 : \Sigma = \Sigma_1$ is based on

$$T = \begin{cases} 1 & g(M; \Sigma_0) \geqslant \lambda g(M; \Sigma_1) \\ 0 & otherwise, \end{cases}$$

for some critical value λ. This power is generally evaluated by simulation. Further, we introduce a family of tree matrices to aid our study.

5.1 *Pectinate tree structures*

Comparative inference assumes that under a Gaussian evolution model (Felsenstein, 1985; Martins and Hansen, 1997), the covariance between groups, Σ_0, can be mapped to a tree structured graph where the main diagonal of Σ_0 is proportional to the total graph length from root to tips (left to right) and the off-diagonals are proportional to the path length from the root to the most recent common ancestor (internal node) between two tips (Felsenstein, 2004).

[Figure 1 about here.]

One sub-class of tree structured matrices of arbitrary dimension may be generated by the following recursive formula, where J_m is the $m \times m$ matrix of ones:

$$\Psi_{m+1} = \begin{pmatrix} \frac{m-1}{m}\Psi_m + \frac{1}{m}J_m & 0 \\ 0 & 1 \end{pmatrix},$$

and $\Psi_1 = I_1$. For concreteness, the matrices Ψ_3 through Ψ_5 are

$$\Psi_3 = \begin{pmatrix} 1 & .5 & 0 \\ .5 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \Psi_4 = \begin{pmatrix} 1 & .67 & .33 & 0 \\ .67 & 1 & .33 & 0 \\ .33 & .33 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \Psi_5 = \begin{pmatrix} 1 & .75 & .50 & .25 & 0 \\ .75 & 1 & .50 & .25 & 0 \\ .50 & .50 & 1 & .25 & 0 \\ .25 & .25 & .25 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix},$$

and correspond to the nested graphs in Fig. 1. Operationally, the recursion adds a branch leading from the root node (furthest left), adds a new root node and a branch of leading to a new leaf and re-scales the entire graph. These trees are sometimes called pectinate topologies, referring to their shape which resembles the parallel teeth of a comb.

Matrices of dimension 3 and higher of this type correspond to tree graphs with maximal distance 2 between leaves (Ives and Garland, 2010):

$$\begin{aligned}\|\Psi_m\|_{Ives} &= \max_{i\neq j}\left\{(\Psi_m)_{ii} + (\Psi_m)_{jj} - 2(\Psi_m)_{ij}\right\}, \\ &= (\Psi_m)_{11} + (\Psi_m)_{mm} - 2(\Psi_m)_{1m}, \\ &= 2.\end{aligned}$$

As a function of the dimension m, the greatest correlation will always appear between the first and second leaves and increases with m. For $m > 1$, it can be shown that the maximum correlation is $(\Psi_m)_{12} = \frac{m-2}{m-1}$. This means that Ψ_m matrices of larger dimension reflect an increasing amount of covariation, and in the limit, are singular. While the minimum eigenvalue remains positive, for $m > 1$, the fact that it is exactly $\frac{1}{m-1}$ may aid numerical studies. We use this class of tree structured matrices to study the properties of the F and Q tests in the next section.

6. Illustrations

6.1 *Validity and sensitivity to heritability by artificial data*

In this section, we verify that tests based on F_n are valid for $H_0 : \Sigma(\theta) = \Sigma_0$ and have power for $\theta \neq 1$. If the known, tree structured matrix Σ_0 cannot be rejected, we say that the trait follows a neutral evolution model. We also consider the validity of the Q_n test, showing that it is valid for the compound null hypothesis specified earlier. To draw an inference about the non-neutrality of a trait, we assume that this tree-covariance Σ_0 is known and we wish to test for observable deviations from the structure.

We first study the pectinate tree matrix $\Sigma_0 = \Psi_{30}$, varying θ between 0 and 4. In Fig. 2, we show the size and power for the two-sided test based on F_n under $H_0 : \Sigma(\theta) = \Sigma_0$ versus $H_1 : \Sigma(\theta) = \theta\Sigma_0$. We see that this F-test is properly sized under H_0 and has good power to detect alternatives which differ from Σ_0 by a scale constant.

In the right panel, we have transformed θ to heritability scale, $h^2 = \frac{\theta}{1+\theta}$, to reflect an evolutionary interpretation (Lynch, 1991). The estimated sizes for a two-sided Q test show that it is insensitive to most reasonable values of h^2. As we noted previously, we expect small heritability values to lead to an identifiability problem as $\theta\Sigma_0$ approaches the zero matrix. In general, comparative analysis tends to focus on traits which are believed to be heritable, so we are less concerned about an invalid test when heritability is low.

[Figure 2 about here.]

6.2 *Tree structured alternative power studies by artificial data*

We study two alternative hypotheses which are based on evolutionary models. First, Butler and King (2004) suggested a non-linear transformation of the covariance matrix Σ_0 based on an Ornstein-Uhlenbeck (OU) process with evolutionary selection parameter $\alpha \geqslant 0$. We denote this OU model covariance as Σ_α with entries:

$$(\Sigma_\alpha)_{ij} = \frac{1}{2\alpha}\left(1 - e^{-2\alpha(\Sigma_0)_{ij}}\right)\left(e^{-2\alpha d_{ij}}\right),$$

where $d_{ij} = (\Sigma_0)_{ii} + (\Sigma_0)_{jj} - 2(\Sigma_0)_{ij}$. One recovers Σ_0 as $\alpha \to 0$, reflecting a hypothesis of no selection. As α grows large, Σ_α becomes approximately diagonal, that is it is close to $\frac{1}{2\alpha} I_m$, before approaching a stationary, degenerate distribution.

Second, Pagel (1999) proposed that the linear scaling of the off-diagonal entries of Σ models loss of phylogenetic signal (PS). We write this matrix as Σ_λ for $0 \leqslant \lambda \leqslant 1$ with entries:

$$(\Sigma_\lambda)_{ij} = \begin{cases} \lambda\Sigma_{ij} & i \neq j \\ \Sigma_{ii} & i = j \end{cases}$$

so that when $\lambda = 0$, $\Sigma_{\lambda=0}$ has no covariance between groups representing completely independent evolution. When $\lambda = 1$, $\Sigma_{\lambda=1} = \Sigma_0$ is the null hypothesis covariance.

In Fig. 3, we use the $\Sigma_0 = \Psi_{30}$ covariance, drawing θ uniformly in $(0.5, 2)$ to reflect the unknown scale and estimate the power for the two-sided test based on Q to detect non-linear

and linear alternatives Σ_α and Σ_λ. Comparing power curves from three values of n, reflecting the number of balanced replicates, we see rapid convergence to the limiting power curve.

We have plotted the region of possible powers accessible by the UMP test which knows the true alternative covariance. For this Σ_0, the UMP test upper bound reflects that even under the best cases, the power to detect these types of selection is only moderate. The difference between the UMP bound and the Q test's power is attributable to the cost of looking at many possible alternatives and controlling for unknown heritability. We emphasize that it performs as well as it can be expected to given our criteria for an accessible, valid and broadly applicable test.

[Figure 3 about here.]

6.3 *Gene expression families*

Gene expression duplication families are sets of genes which have evolved from a common ancestral sequence (Thornton and DeSalle, 2000). Because these families have biological roles, geneticists might infer how proteins develop their specific function by studying evolutionary process underlying their family history (Fay and Wittkopp, 2008).

Oakley et al. (2005) estimate the frequency of non-neutral evolutionary models in 10 yeast gene families under a variety of experimental conditions by model selection (using Akaike's Information Criterion). Out of the possible variance models, they find 35/152 cases where the selected model shows dependence and 117/152 cases where groups (gene transcripts) appear independent. Of these 35 cases showing covariance, only 3 confirm the neutral expectation implied by the history of sequence evolution. The most common model (94/152 cases) has independent but heteroskedastic across-group variances. The authors conclude that the primary evolutionary process may be the "erasure of historical signal during the rapid evolution of gene expression."

From our synthetic example, our F test should be sensitive to loss of heritability in exactly

this sense. In contrast to the findings in Oakley et al. (2005), we observe 99/152 neutral (not significant) tests. This may be symptomatic of low power, however 23 of the 53 significant tests are rejected in the upper tail suggesting that, in half the cases, more variation was observed than was expected.

6.4 *Inbred mouse morphological trait study*

Inbred strains of laboratory mice have positive dependence arising from their common origin (Beck et al., 2000; Frazer et al., 2007; Yang et al., 2007) which needs to be accounted for when making inferences on phenotypes measured across strains (Kang et al., 2008). Analogous to quantitative and neutral marker comparisons (Whitlock, 2008), it is also of interest to know whether a given trait displays the same phenotypic divergence across strains as would be expected under a neutral evolution model (Kimura, 1964). Should these populations' phenotypic plasticity be constrained, the investigator infers that a selective force has influenced the phenotype's variation (Whitehead and Crawford, 2006a).

We performed an exploratory analysis of morphological phenotypes from $m = 35$ inbred strains measured on 644 mice collected by Bret Payseur (University of Wisconsin-Madison) and Christopher Vinyard (Northeastern Ohio Universities Colleges of Medicine and Pharmacy). We used mouse hapmap genotypes to construct an identity by state distance matrix and used a neighbor joining algorithm to construct Σ_0. Following the suggestion in Leinonen et al. (2007) to use this kind of test to screen traits for candidates for further study, we examine the p-values of the Q test in table 1.

We observe that the traits with significant tests at level 0.05 tend to be in male mice. Despite being not statistically significant, a number of alluringly small p-values may signal a general lack of power to detect differences in this large covariance matrix. Across both sexes, each trait's test points to rejection in the downward direction suggesting that we observe less covariance than expected. We also tabulate phylogenetic generalized least squares (Grafen,

1989) type estimates of h^2 that suggest that measurement error is not dominating the signal. This implies that had these tests been significant, we would have inferred a form of stabilizing selection. We must also point out that some traits are highly correlated (e.g., body length is the sum of tail and crown to rump lengths) so the common trends are not wholly unexpected. From this analysis, we hypothesize that gross characters like body length might undergo selection which restricts their variation and that smaller skeletal measurements may as well.

[Table 1 about here.]

We have used the Q test here because the scale associated with Ω reflects pure environmental error: no genetic variation is expected within fully inbred individuals. Further, we expected morphological traits to be highly heritable. So, the Q test has focused on scale free changes in dependence.

7. Discussion

We have presented statistics for testing a phenotype for deviations from the predictions of a neutral evolution model. Importantly, these tests have closed form reference distributions and scale invariance properties. They follow a goodness-of-fit approach, looking for any deviation from neutrality. In contrast, selection and estimation methods (Freckleton et al., 2002; Butler and King, 2004; Gu, 2004; Oakley et al., 2005) are able to characterize specific types of models but are unable to identify whether the true alternative lies outside of the class of models under consideration.

Additional complications like missing data might be accommodated by adapting the mixed effects model and adjusting the distribution of the group means accordingly.

Assuming that we observe balanced replicates, we have not explicitly examined the conditions under which $M_n \rightarrow_d M_\infty$. Except for distinctly pathological cases, we imagine that

moderate amounts of imbalance are acceptable given that these group means will quickly converge.

The neutral inference is not complete without a sense of direction, having rejected a neutral model one naturally asks what kind of selection may be present and how to rule out environmental causes which look like heritable variation.

While the pectinate topologies have been described previously, to our knowledge, this is the first formal statement of some of their properties in the context of comparative methods. Besides their ability to scale to a size m, one of their attractive points is that they induce extreme correlation as they grow, contrasting the Ornstein-Uhlenbeck models which erase dependence as a function of their selection parameter.

Acknowledgements

The authors thank the Bret Payseur and Chris Vinyard for permitting use of their data. Conversations with Bret Payseur and Cécile Ané significantly improved our understanding of the details of the Brownian models. This work was supported by NIH grants LM007359, HL083806, and HG003747.

References

N. Balakrishnan and V. B. Nevzorov. *A Primer on Statistical Distributions*. Wiley, 2003. ISBN 0-471-42798-5.

J.A. Beck, S. Lloyd, M. Hafezparast, M. Lennon-Pierce, J.T. Eppig, M.F.W. Festing, and E.M.C. Fisher. Genealogies of mouse inbred strains. *Nature genetics*, 24(1):23–25, 2000.

S. P. Blomberg, T. G. Garland, and A. R. Ives. Testing for phylogenetic signal in comparative data: Behavioral traits are more labile. *Evolution*, 57(4):717–745, 2003.

M. A. Butler and A. A. King. Phylogenetic comparative analysis: A modeling approach for adaptive evolution. *The American Naturalist*, 164(6):683–695, 2004.

D. R. Cox and H. D. Miller. *The Theory of Stochastic Processes.* Chapman and Hall, 1965.

J. C. Fay and P. J. Wittkopp. Evaluating the role of natural selection in the evolution of gene regulation. *Heredity*, 100:191–199, 2008.

J. Felsenstein. Phylogenies and the comparative method. *The American Naturalist*, 125(1): 1–15, 1985.

J. Felsenstein. Phylogenies and quantitative characters. *Annual Review of Ecological Systems*, 19:447–471, 1988.

J. Felsenstein. *Inferring phylogenies.* Sinauer Associates, 2004. ISBN 9780878931774.

J. Felsenstein. Comparative methods with sampling error and within-species variation: Contrasts revisited and revised. *The American Naturalist*, 171(6):713–725, 2008.

K.A. Frazer, E. Eskin, H.M. Kang, M.A. Bogue, D.A. Hinds, E.J. Beilharz, R.V. Gupta, J. Montgomery, M.M. Morenzoni, and G.B. Nilsen. A sequence-based variation map of 8.27 million SNPs in inbred mouse strains. *Nature*, 448(7157):1050–1053, 2007. ISSN 0028-0836.

R. P. Freckleton, P. H. Harvey, and M. Pagel. Phylogenetic analysis and comparative data: a test and review of evidence. *The American Naturalist*, 160(6):712–726, 2002.

T. Garland, Jr., A. W. Dickerman, C. M. Janis, and J. A. Jones. Phylogenetic aanalysis of covariance by computer simulation. *Systematic Biology*, 42(3):265–292, 1993.

A. Grafen. The phylogenetic regression. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 326(1233):119–157, 1989. ISSN 0080-4622.

X. Gu. Statistical framework for phylogenomic analysis of gene family expression profiles. *Genetics*, 167:531–542, 2004.

H. Guo, R. E. Weiss, X. Gu, and M.A. Suchard. Time squared: repeated measures on phylogenies. *Molecular Biology and Evolution*, 24(2):352–362, 2007.

A. K. Gupta and D. K. Nagar. *Matrix Variate Distributions.* Chapman & Hall, 2000.

T. F. Hansen. Stabilizing selection and the comparative analysis of adaptation. *Evolution*, 51(5):1341–1351, 1997.

W. G. Hill and R. Thompson. Probabilities of non-positive definite between-group or genetic covariance matrices. *Biometrics*, 34(3):429–439, 1978.

E. A. Housworth, E. P. Martins, and M. Lynch. The phylogenetic mixed model. *The American Naturalist*, 163(1):84–96, 2004.

JP Imhof. Computing the distribution of quadratic forms in normal variables. *Biometrika*, 48(3/4):419–426, 1961. ISSN 0006-3444.

A. R. Ives and T. Garland. Phylogenetic logistic regression for binary dependent variables. *Systematic Biology*, 59(1):9–26, 2010.

A. R. Ives, P. E. Midford, and T. Garland, Jr. Within-species variation and measurment error in phylogenetic comparative methods. *Systematic Biology*, 56(2):252–270, 2007.

H. M. Kang, N. A. Zaitlen, C. M. Wade, A. Kirby, D. Heckerman, M. J. Daly, and E. Eskin. Efficient control of population structure in model organism association mapping. *Genetics*, 178:1709–1723, 2008.

M. Kimura. Diffusion models in population genetics. *Journal of Applied Probability*, 1: 177–232, 1964.

T. Leinonen, R. B. O'Hara, J. M Cano, and J. Merilä. Comparative studies of quantitative trait and neutral marker divergence. *Journal of Evolutionary Biology*, 21:1–17, 2007.

M. Lynch. Methods for the analysis of comparative data in evolutionary biology. *Evolution*, 45(5):1065–1080, 1991.

E. P. Martins and T. F. Hansen. Phylogenies and the comparative method: a general approach to incorporating phylogenetic information into the analysis of interspecific data. *The American Naturalist*, 149(4):646–667, 1997.

J. Neyman and E.S. Pearson. On the problem of the most efficient tests of statistical

hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231(694-706):289, 1933.

S. V. Nuzhdin, M. L. Wayne, K.L. Harmon, and L. M. McIntyre. Common pattern of evolution of gene expression level and protein sequence in *drosophila*. *Molecular Biology and Evolution*, 21(7):1308–1317, 2004.

T. H. Oakley, Z. Gu, E. Abouheif, N. H. Patel, and W. H. Li. Comparative methods for the analysis of gene-expression evolution: an example using yeast functional genomic data. *Molecular Biology and Evolution*, 22(1):40–50, 2005.

M. Pagel. Inferring the historical patterns of biological evolution. *Nature*, 401:877–884, 1999.

S. A. Rifkin, J. Kim, and K. P. White. Evolution of gene expression in the *drosophila melanogaster* subgroup. *Nature Genetics*, 33:138–144, 2003.

G. A. F. Seber and A. J. Lee. *Linear Regression Analysis*. Wiley, 2 edition, 2003.

J. W. Thornton and R. DeSalle. Gene family evolution and homology: Genomics meets phylogenetics. *Annual Review of Genomics and Human Genetics*, 1:41–73, 2000.

A. Whitehead and D. L. Crawford. Variation within and among species in gene expression: raw material for evolution. *Molecular Biology*, 15:1197–1211, 2006a.

A. Whitehead and D. L. Crawford. Neutral and adaptive variation in gene expression. *Proceedings of the National Academy of Sciences*, 103(14):5425–5430, 2006b.

M.C. Whitlock. Evolutionary inference from q_{ST}. *Molecular Ecology*, 17(8):1885–1896, 2008. ISSN 1365-294X.

E.B. Wilson and M.M. Hilferty. The distribution of chi-square. *proceedings of the National Academy of Sciences of the United States of America*, 17(12):684, 1931.

H. Yang, T.A. Bell, G.A. Churchill, and F.P.M. de Villena. On the subspecific origin of the laboratory mouse. *Nature genetics*, 39(9):1100–1107, 2007. ISSN 1061-4036.

Appendix

Proof of lemma 1 Under $H_0 : \Sigma(\theta) = \theta\Sigma_0$, let $X = KM_\infty \sim \mathcal{N}(0, \sigma^2\theta K\Sigma_0 K^T)$ and write $A = K\Sigma_0 K^T$ and $p = m - 1$. The positive definite symmetric A has eigensystem decomposition $A = TDT^T$ for orthogonal T and diagonal D where $T^T X = D^{1/2}Z$ for p dimensional multivariate normal Z.

$$\begin{aligned}
\Pr\left(Q_\infty \leqslant t\right) &= \Pr\left(\frac{X^T A^{-1} X}{X^T X} \leqslant t\right) \\
&= \Pr\left(X^T\left[A^{-1} - tI_p\right]X \leqslant 0\right) \\
&= \Pr\left(X^T\left[TD^{-1}T^T - tI_p\right]X \leqslant 0\right) \\
&= \Pr\left(Z^T\left[I_p - tD\right]Z \leqslant 0\right) \\
&= \Pr\left(\frac{1}{\sum_{j=1}^p d_j \frac{Z_j^2}{\sum_{i=1}^p Z_i^2}} \leqslant t\right) \\
&= \Pr\left(\frac{1}{\sum_{j=1}^p d_j W_j} \leqslant t\right)
\end{aligned}$$

Where the last equality follows by observing that the individual Z_j^2 are independent χ_1^2 random variables, their sum is gamma distributed with parameters shape $1/2$ and scale p. From Balakrishnan and Nevzorov (2003), the joint distribution of $W_j = Z_j^2 / \sum_{i=1}^p Z_i^2$ is symmetric dirichlet with parameter $1/2$. For clarity, we emphasize that $\sum_{j=1}^p W_j = 1$.

Reference quantiles from this non-standard distribution can be obtained quickly by numerical integration using R function `gtools::rdirichlet`.

Figure 1. *Nesting among pectinate tree topologies.* The three-leaf tree with solid edges represents the covariance matrix Ψ_3. The recursion operation generating Ψ_4 adds the dashed lines to the solid graph. Likewise the dotted lines are added when generating Ψ_5. The tree is re-scaled at each iteration, so no scales are shown.

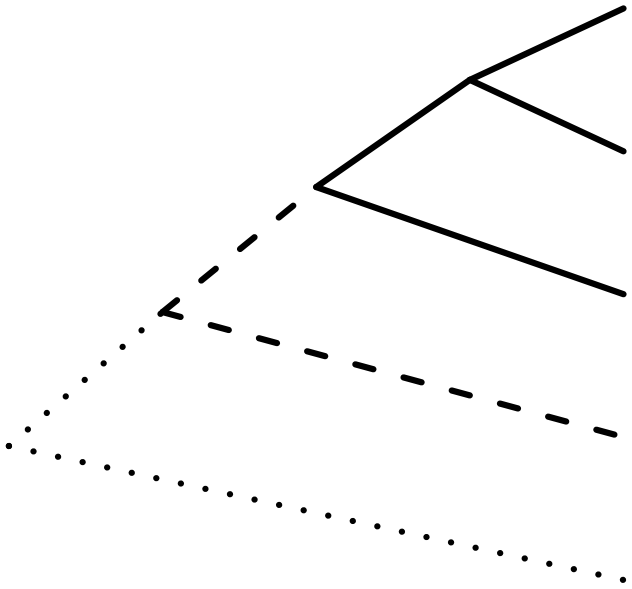

Figure 2. *Relative rates hypotheses.* (left) The F test is valid at $\theta = 1$ (marked on the axis) and has power for $\theta \neq 1$. (right) The Q test considers all values of θ (on the heritability scale $h^2 = \theta/(1+\theta)$) to be part of the null hypothesis. The target size is 0.05. Parameter values considered are indicated by the rug.

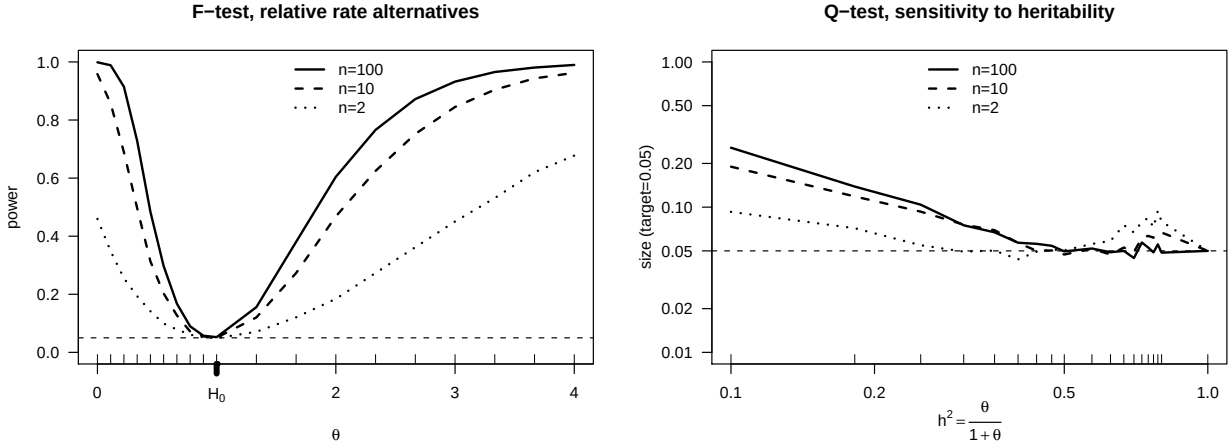

Figure 3. *Power for Q tests versus* Σ_α *and* Σ_λ *type alternatives.* The grey shaded area indicates the limits of detection using a theoretical UMP test. The null hypothesis is H_0 : $\Sigma(\theta) = \theta\Psi_{30}$.

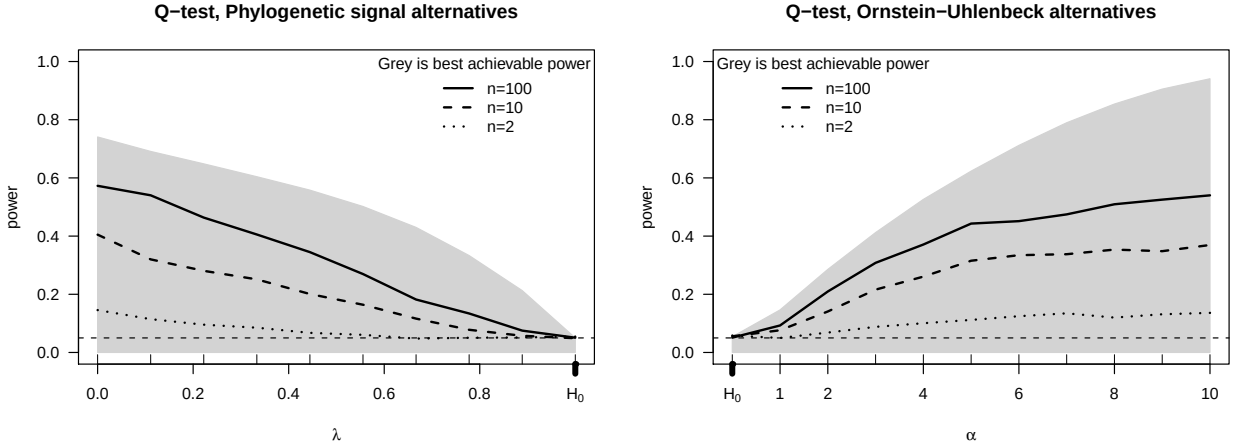

Table 1
Two-sided Q test p-values and heritability estimates for mouse morphological traits. Traits with non-neutral covariance structure ($p < 0.05$) are highlighted.

	Female		Male	
	h^2	p	h^2	p
Scapula Length	0.77	*0.0470*	0.74	*0.0440*
Crown Rump Length	0.77	0.0502	0.76	*0.0370*
Tarsus Length	0.82	0.0848	0.76	*0.0260*
Body Length	0.80	0.1022	0.78	*0.0458*
Hind Length	0.76	0.1098	0.74	0.0768
Leg Length	0.78	0.1690	0.76	0.0640
Tail Length	0.79	0.1992	0.78	0.0646
Thigh Length	0.77	0.2236	0.76	0.0912
Arm Length	0.77	0.2552	0.70	0.0952