DEPARTMENT OF STATISTICS
University of Wisconsin
1300 University Avenue
Madison, WI 53706

TECHNICAL REPORT NO. 1144
July 28, 2008

A Phylogenetic Mixture Model for the Evolution of Gene Expression[1]

Kevin H. Eng
Department of Statistics
Department of Biostatistics and Medical Informatics
University of Wisconsin, Madison

Héctor Corrada Bravo
Department of Computer Sciences
University of Wisconsin, Madison

Grace Wahba
Department of Statistics
Department of Biostatistics and Medical Informatics
University of Wisconsin, Madison

Sündüz Keleş
Department of Statistics
Department of Biostatistics and Medical Informatics
University of Wisconsin, Madison

[1]This work was supported by NSF grants DMS 0604572(GW) and DMS 0804597 (SK); a PhRMA Foundation Research Starter Grant (SK); NIH grants HG03747 (SK) and EY09946(GW) and ONR Grant N0014-06-0095 (GW).

July 28, 2008

A Phylogenetic Mixture Model for the Evolution of Gene Expression

Research Article

Kevin H. Eng[1,3], Héctor Corrada Bravo[2], Grace Wahba[1,2,3], Sündüz Keleş[1,3*]

1. Department of Statistics
2. Department of Computer Sciences
3. Department of Biostatistics and Medical Informatics

*Corresponding Author:
Sündüz Keleş
Department of Statistics
1300 University Ave
University of Wisconsin-Madison
Madison, WI 53706
608-262-2598
608-262-0032 (fax)
keles@stat.wisc.edu

Running Title: Phylogenetic Mixture Model
Keywords: gene expression, mixture model, natural selection, phylogenetic correction.

Abstract

Microarray platforms are used increasingly to make comparative inferences through genome-wide surveys of gene expression. While recent studies focus on describing the evidence for natural selection using estimates of the within and between taxa mutational variances, these methods do not often explicitly or flexibly account for predicted non-independence due to phylogenetic associations between measurements. In the interest of parsing the effects of selection, we introduce a mixture model for the comparative analysis of variation in gene expression across multiple taxa. This class of models isolates the phylogenetic signal from the non-phylogenetic and the heritable signal from the non-heritable while measuring the proper amount of correction. As a result, the mixture model resolves outstanding differences between existing models and relates different ways to estimate the across taxa variance. We investigate by simulation and application the feasibility and utility of estimation of the required parameters. We illustrate the estimation with a gene duplication family data set, discussing previously proposed estimates for testing selection hypotheses and proposing improvements.

Keywords: gene expression, mixture model, natural selection, phylogenetic correction.

1 Introduction

The availability of gene expression data en masse admits a genomic resolution comparative expression experiment which measures many homologous gene transcripts simultaneously across many taxa in the interest of determining which genes are likely to undergo selective forces (Rifkin et al., 2003; Nuzhdin et al., 2004; Whitehead and Crawford, 2006a). Through such an experiment the investigator may determine the relative strengths of neutral drift and natural selection forces on gene expression traits (Fay and Wittkopp, 2007) at the single gene level while isolating whole groups of genes which act together and which might have a common evolutionary history. These investigators propose the use of the variance within and between taxa to determine the strength and form of hypothesized selection forces. The expression of each gene is a single, continuously-valued trait, and, as in the usual comparative experiment, the analysis is potentially obfuscated by the evolutionary dependence common to the taxa.

To account for this dependence, we may examine the structured form of the phylogenetic covariance matrix defined between taxa. The investigator typically considers the evolutionary relationship evidenced by a phylogenetic tree estimated from some other characters, but for model based comparative analyses we wish to translate these trees to covariance matrices. Under the assumption of a Brownian motion process underlying the historical evolution of the trait, we may construct a phylogenetic covariance from a known tree (Felsenstein, 1988). For general phylogenetic covariance matrices, Martins and Housworth (2002) suggested an eigenvector decomposition to identify variance with specific tree shapes. In Corrada Bravo et al. (2008), we developed a new algorithm for estimating a tree and its matching Brownian motion covariance directly from observed continuous-trait data. As opposed to methods like neighbor-joining (Saitou, 1987), this method globally optimizes a projection criterion over all possible tree topologies using proven, efficient methods for combinatorial optimization. For expression from gene duplication families, Gu (2004) and Oakley et al. (2005) both re-parameterize the mutational rates on each branch of a known tree covariance allowing it to better fit the phylogeny information. Of particular note, the addition of an error component allows these covariances to extend to a model for the entire experiment with a single covariance matrix (Ives et al., 2007).

Practically, linear models model both dependence and error by implicitly assuming a covariance structure which decomposes the observed or experimental variance. Such decompositions are especially desirable since they correspond to known structures in the experiment. Lynch (1991) defines a mixed effects model across multiple traits, capturing the phylogenetic structure in a relationship matrix G and covariance between traits as a series of single parameter variance components. While adapting Lynch's model for biological replicate data, Christman et al. (1997) extend a memetic, due to Cheverud et al. (1985), where the trait value (T) is separated into a phylogenetic component (P), a specific value (S) and a random error component (E), namely T=P+S+E. This decomposition leads the authors to conclude that Lynch's model isolates heritable effects (P+S) from noise (E) but fails to separate them from one another (P from S). Housworth et al. (2004) reformulate Lynch's model to address this deficiency by emphasizing a parameter which indirectly estimates the degree of phylogenetic signal in the sample. More recently, Guo et al. (2006) fit three types of Bayesian flavored mixed effects models each parameterizing an increasing amount of phylogenetic

signal, finding that modeling the degree of signal present yields better models.

The importance of determining the amount of phylogenetic signal in a sample cannot be understated. If there is a phylogenetic signal, the comparative analysis ought to find and remove the extra variation. If no signal can be detected, then corrective methods will overzealously bias the final estimates. Permutation tests at the level of tree estimation offer a way of testing for the presence of a signal or not (Blomberg et al., 2003). Pagel (1999) introduced λ as a measure of the strength of the signal and developed a likelihood ratio test for its presence. Similarly, Housworth et al. (2004) adopted h^2 as a measure of the strength of the heritable signal in the data. In both cases, a continuous estimate carries more information than a dichotomous hypothesis test and should it indicate a strong signal, we ought to apply an appropriate phylogenetic correction.

Our goal in this paper is to integrate a framework for studying selection forces into phylogenetic, variance-decomposing models in a gene expression context. With respect to tests of selection, Rifkin et al. (2003) proposed the use of the estimated mean squares to model expected variation between and within taxa. In this framework, evidence of deviation from expectation under neutrality is evidence of the effect of natural selection. Nuzhdin et al. (2004) revised this idea using nested random effects in an ANOVA model and proposing the numerator and denominator of the standard, uncorrected F-ratio to be estimates of the between and within taxa variance. In particular, they give forms of the tests which distinguish between purifying and adaptive selection. Whitehead and Crawford (2006a) continue the use of plain mean square estimates, adding a test for stabilizing selection.

In this article, we present a mixture model for the covariance in order to resolve predecessor models' inability to separate phylogenetic effects from non-phylogenetic ones clearly and to resolve the exclusion of consideration for the structured dependence in the testing frameworks of Nuzhdin et al. (2004) and Whitehead and Crawford (2006a). In such a model, the necessary degree of correction is freely estimated so the investigator may draw inferences on parameters un-confounded by dependence. We discuss the convergence of existing models by demonstrating the relationships between their assumptions on the covariance; the mixture formulation covers a continuum of models set between independent contrasts and the class of phylogenetic mixed effects models. We describe the main assumptions and implications of the model from the practical analysis point of view, illustrating its effect with a simulation study and demonstrating its use in the study of gene duplication families in *Saccharomyces cerevisiae* (Oakley et al., 2005).

2 Methods

2.1 Models for the Structured Variance Among Taxa

We wish to consider existing models for the covariance matrix defined between taxa, so that we may study how they reflect the phylogenetic tree-structured signal and how they measure both the specific variance and the error in the observed data. In an independent contrasts framework (Felsenstein, 1985), suppose V_0 is a tree-structured, phylogenetic matrix (Felsenstein, 1988) representing a true phylogenetic history among T taxa. Since the true structure may be confounded by other processes, we can consider extending this covariance

to the structured form proposed in Pagel (1999) and Freckleton et al. (2002). These studies introduce λ as a measure of the strength of phylogenetic correlation, or as a measure of the "loss of history," which induces a covariance matrix $\mathrm{V}(\lambda)$. In defining $\mathrm{V}(\lambda)$ to be a phylogenetic covariance matrix whose off-diagonals are multiplied by λ, the authors implicitly assume that opposing the phylogenetic structure V_0 is a specific, non-phylogenetic structure Λ_0:

$$\begin{aligned} \mathrm{V}(\lambda) &= [\lambda \mathrm{J}_T + (1-\lambda)\mathrm{I}_T] \circ \mathrm{V}_0 \\ &= \lambda \mathrm{V}_0 + (1-\lambda)\Lambda_0. \end{aligned} \tag{1}$$

Here, J_T is a $T \times T$ matrix of ones, I_T is the identity matrix of the same dimension and $\circ$ is the element-wise (Hadamard) product. We define Λ_0 to be the diagonal matrix with the same main diagonal as V_0 and will assume that $0 \leq \lambda \leq 1$. Even though Pagel (1999) admitted an interpretation for $\lambda > 1$, restricting its range preserves its interpretation as the proportion of signal attributable to phylogeny.

For each gene, $g = 1, \ldots, G$, suppose we measure $N = RT$ many individuals from $t = 1, \ldots, T$ many taxa with $r = 1, \ldots, R$ balanced replicates. That is, for each observation vector $Y_{gr} = (Y_{gr1}, \ldots, Y_{grT})'$, we measure the homologous transcript once in each taxa and assume that $\mathrm{E}(Y_{gr}) = \mu_g$. Freckleton et al.'s model (2002) assumes that this vector has variance

$$\begin{aligned} Var(Y_{gr}) &= \tau^2 V(\lambda) \\ &= \tau^2 \lambda \mathrm{V}_0 + \tau^2 (1-\lambda)\Lambda_0. \end{aligned} \tag{2}$$

Freckleton et al. (2002) did not make explicit use of this additive relationship between λ and V_0, but this sum makes clear that hidden inside $\mathrm{V}(\lambda)$ is a decomposition of the observed signal into a phylogenetic and non-phylogenetic component. While this partition is desirable, it lacks a sense of error due to measurements or within the taxa in the sample.

When the taxa span species, the within taxa error is typically assumed to be negligible relative to the across taxa error. Ives et al. (2007), however, observe that this error component may include error due to populations and measurement. For gene expression microarray studies, the sampling error induced having only a few samples may be of some concern. Further, one might argue for the interpretation of the within taxa error as a rate of mutation (Guo et al., 2006; Rifkin et al., 2003; Whitehead and Crawford, 2006a).

An alternative way to model the covariance, a linear mixed model (Lynch, 1991; Housworth et al., 2004; Guo et al., 2006), induces a covariance by assuming that there ought to be a random component for the heritable portion of the signal b_{gr} independent of the component for error e_{gr}. As Christman et al. (1997) observed, this decomposition only isolates the error component from the signal of interest. But, if we substitute the form of Pagel's

matrix $\mathrm{V}(\lambda)$ for the variance of b_{gr}, we obtain the following decomposition.

$$\begin{aligned} Var(Y_{gr}) &= Var(b_{gr}) + Var(e_{gr}) && (3)\\ &= \tau^2 V(\lambda) + \sigma^2 \mathrm{I}_T && (4)\\ &= \tau^2 \lambda \mathrm{V}_0 + \tau^2(1-\lambda)\Lambda_0 + \sigma^2 \mathrm{I}_T && (5)\\ &= \frac{\tau^2\lambda}{\tau^2+\sigma^2}(\tau^2+\sigma^2)\mathrm{V}_0 + \frac{\tau^2(1-\lambda)}{\tau^2+\sigma^2}(\tau^2+\sigma^2)\Lambda_0 + \frac{\sigma^2}{\tau^2+\sigma^2}(\tau^2+\sigma^2)\mathrm{I}_T && (6)\\ &= p_1(\kappa \mathrm{V}_0) + p_2(\kappa\Lambda_0) + p_3(\kappa \mathrm{I}_T), && (7) \end{aligned}$$

where the first part is a phylogenetic component, the second a non-phylogenetic component and the third an error component for $p_1 + p_2 + p_3 = 1$. This implies that utilizing Pagel's construction (1999) of $\mathrm{V}(\lambda)$ results in a model which delivers precisely Christman et al.'s (1997) desired decomposition into a phylogenetic component (P), a taxon specific one (S) and a random error part (E). Here, λ controls the relative strength of P versus S. This form of the variance naturally suggests the mixture model as the combination of these two forms (one which models only P versus S and another which separates only P+S from E).

2.2 A Phylogenetic Mixture Model

Suppose that the gene expression measurement $Y_{gr} = \mu_g + Z_{gr}$ may be decomposed into a fixed mean μ_g and a random part Z_{gr} which follows a phylogenetic mixture distribution. For some probability densities, $f_{V_0}(z), f_{\Lambda_0}(z), f_{I_T}(z)$, the mixture density is given by

$$f_{mix}(Z_{gr}) = p_1 f_{V_0}(Z_{gr}) + p_2 f_{\Lambda_0}(Z_{gr}) + p_3 f_{I_T}(Z_{gr}), \tag{8}$$

where $p_1 + p_2 + p_3 = 1$. That is, the mixture model is defined by the weighted average of the three densities. From the arguments in Felsenstein (1973) it is natural to assume that a pure phylogenetic history component follows a Brownian motion process and thus we assume that $f_{V_0}(z)$ is a normal density. Likewise, the non-phylogenetic component, $f_{\Lambda_0}(z)$, may be assumed normal and, for tractability, one tends to assume error distributions, $f_{I_T}(z)$, are also normal. We assume, therefore, that

$$\mathcal{N}(0, \kappa \mathrm{V}_0), \mathcal{N}(0, \kappa\Lambda_0), \mathcal{N}(0, \kappa \mathrm{I}_T) \tag{9}$$

are the distributions of the independent components. It is important to note that the marginal density $f_{mix}(z)$ is not a normal density, but since each of the components are mean zero and normal, the mixture may be similar to the multivariate normal marginal density of a mixed effects model (Lynch, 1991; Martins and Hansen, 1997; Guo et al., 2006). Further, we believe that this mixture model is a reasonable generating model both for the data structure we described in section 2.1 and for the gene family case study we will present in section 3.3.

In a mixture model, the mixing proportions p_1, p_2, p_3, and the scaling constant κ are commonly estimated by the application of an Expectation Maximization (EM) algorithm (Dempster et al., 1977). Having done so, we can retrieve the original parameters by transforming the estimates into $\lambda, \sigma^2, \tau^2$. The required steps are laid out in Appendix A. Since μ_g is a fixed effect separate from the mixture, its estimation is trivial.

These component distributions represent particular archetypical scenarios. If the data show phylogenetic signal (a particular type of non-independence) then we believe that they come from the $f_{V_0}(z)$ component. If the data were independent but not identically distributed (each has is own specific variance) then $f_{\Lambda_0}(z)$ is the correct model. If the taxa were truly independent, identically distributed noise then $f_{I_T}(z)$ takes precedence. Mixing proportions p_1, p_2, p_3 represent the relative strengths or relative probability of each component.

In estimating the parameters from this mixture model, we obtain estimates of the decomposed, "marginal" covariance whose parameters have important connotations:

$$Var(Y_{gr}) = \tau^2 \lambda \mathrm{V}_0 + \tau^2(1-\lambda)\Lambda_0 + \sigma^2 I_T. \tag{10}$$

If λ is the strength of phylogenetic signal in the data, τ^2 scales the variance of the signal and is commonly called the taxa specific variance. Likewise the error variance is σ^2. Since V_0 is scaled by it, λ represents a degree of correction to the observed covariance. If $\lambda = 1$ then the correction is strong whereas if $\lambda = 0$ there is no correction for dependence.

Recalling that our goal is to model the correlated signal in the data, a different type of analysis would fit each component model separately and use a likelihood ratio test to choose between them (this is the original proposal in Freckleton et al. (2002)). But, the authors noted in their empirical survey of the available studies that comparative studies tend to lack the sample size that generates sufficient power to make clear decisions. While Martins and Hansen (1997) charged that these models may generally lack the power to make the decision between a phylogenetic and non-phylogenetic analysis, Freckleton et al.'s (2002) counterargument holds: because we propose to estimate λ rather than to test for $\lambda = 1$ or $\lambda = 0$, in this model, λ admits a continuum of possible corrections and controls the type of analysis by attempting to find the optimal amount of correction. Even if we do not reach optimal power, we get as good a guess as the data allows with respect to the degree of correction. Similarly, the permutation test in Blomberg et al. (2003), the model selection procedure in Oakley et al. (2005) or the Bayesian model selection problem in Guo et al. (2006) all choose only one among many candidate models.

Per Freckleton et al. (2002)'s argument, this desirable property derives from the transformative interpretation of $\mathrm{V}(\lambda)$. In practice, we do not have to choose between extremes since they may be chosen for us: one can observe, from the form of the covariances, that when $\lambda = 1$ our variance and the variance of the original class of mixed effects models agree and when $\lambda = 0$ we have a properly scaled form of an uncorrected, non-phylogenetic analysis.

$$\begin{array}{lll} \lambda = 1 & \tau^2 \mathrm{V}_0 + \sigma^2 I_T, & \text{(Phylogenetic)} \\ 0 < \lambda < 1 & \tau^2 \mathrm{V}(\lambda) + \sigma^2 I_T, & \\ \lambda = 0 & \tau^2 \Lambda_0 + \sigma^2 I_T. & \text{(Non-phylogenetic)} \end{array} \tag{11}$$

An alternative measurement for the strength of signal was proposed in Housworth et al. (2004) to reparameterize the model by Lynch (1991). The phylogenetic heritability, h^2, is defined in our model as the proportion, $h^2 = \frac{\tau^2}{\tau^2+\sigma^2}$, interpreting σ^2 as Lynch's or Housworth's nonheritable variance or Christman et al.'s (1997) error variance. It follows that the mixture variance can be written as,

$$Var(Y_{gr}) \propto h^2 \left[\lambda \mathrm{V}_0 + (1-\lambda)\Lambda_0\right] + (1-h^2)\mathrm{I}_T, \tag{12}$$

where $\propto$ indicates "proportional to" the total variation, $\tau^2 + \sigma^2$. By their observation, when $h^2 = 1$, Housworth et al.'s (2004) mixed effects model is identical to Felsenstein's (1985) phylogenetic independent contrasts method (i.e., a model where there is only phylogenetic signal). In our model, h^2 is close to 1 when σ^2 is close to zero. So, as σ^2 decreases our marginal covariance becomes increasingly close to the covariance assumed in the independent contrasts method (V_0) provided λ is large; h^2 remains a measure of the strength of the heritable variance part versus the non-heritable part.

2.3 Covariance Transformations

In addition to an interpretation as a decomposition of the variance, we give the following transformative re-interpretation of Freckleton et al.'s (2002) model. They originally call λ a measure of how closely the trait follows the Brownian motion process, so that for λ small, the Brownian motion assumption may be violated. We argue that every value of λ corresponds to a particular covariance matrix which could have been generated by a particular Brownian motion history. If true, the utility in λ comes from indexing which process is the best fit for the expression trait data.

If we consider the tree implied by the covariance matrix, $V(\lambda)$, since λ shrinks the off diagonal entries, it shrinks the ancestral branches of the tree corresponding to V_0 while fixing the total branch length. When $\lambda = 0$, every off diagonal is exactly zero and this covariance matrix corresponds to the independent "star tree" configuration with total branch lengths the same as the total branch lengths in V_0. When $\lambda = 1$, the covariance is identically V_0. For $0 < \lambda < 1$, the covariance corresponds to some tree with ancestral branches shrunk towards the root, assuming a Brownian motion process. Thus, λ helps our model find, for the expression trait, the best covariance corresponding to a Brownian motion tree between V_0 and the star tree Λ_0. A graphical representation appears in Figure 1.

Here there are trees for the phylogenetic ($\lambda > 0$) and non-phylogenetic ($\lambda = 0$) cases. The dotted segments represent variation attributable to σ^2, the independent variation common to all taxa, accordingly they all have the same length. τ^2 controls variation in the branched part of the diagram, and, even in the non-phylogenetic case, it isolates the independent variation specific to each taxa. If we think about the position of the ancestor nodes as mobile, λ controls the relative size of the branch between the last split and all ancestors versus the splits higher up the tree.

[Figure 1 about here.]

This is an adaptation of the figure from Housworth et al. (2004) who wrote that h^2 separates two components as the phylogenetic mixed model "envisions extant taxon phenotypes to be the result of a linear combination of gradually accumulated evolutionary changes occurring along a true species phylogeny and short-lived evolutionary changes ... occurring in each taxon independently and not passed on between ancestor and descendent taxa" (pp 85).

Since V_0 is assumed estimated from the sequence data, of issue is how the tree implied by the expression process relates to the phylogenetic tree. Since the parametrization by λ only changes the branch lengths and not the topology of the tree, we might infer that the differences between the expression tree and the phylogeny depend on the comparisons of the

rates of trait mutation along different parts of the tree and their relation to the observed sequence divergence. That is, if we know that the tree defines a short nearest-ancestor-to-leaf evolutionary time, but we observe strongly independent expression signal between taxa, we might conclude that the rate on those branches, σ^2, is very large. In particular, σ^2 and τ^2 are interpretable as estimates of the gene-specific constraints inherent in some function which relates variation within taxa and between taxa as discussed in Whitehead and Crawford (2006b). This idea will be highlighted in the next section.

We give some consideration to the form of V_0 which is generally assumed to be a known, given, tree-structured matrix estimated from separate characters. While we generally focus on the Brownian motion interpretation for this model (see Felsenstein (1988) on the link between trees and Brownian motion covariances), within the greater framework of statistical linear models, it is relatively simple to extend this formulation to Lynch's (1991) relationship matrix G, a functional covariance suggested in Martins and Hansen's (1997) PGLS method, or indeed any method that relies on placing phylogenetic structure into a covariance term. In each of these cases, the form and estimation of the mixture model remains the same; one only needs to change the given V_0. It is a little less clear what it means to shrink the off diagonals of the matrix in these cases.

2.4 Testing Selection Hypotheses

Following Rifkin et al. (2003), Nuzhdin et al. (2004) and Whitehead and Crawford (2006a), we consider the evidence in favor of natural selection forces characterized by the variance within and between taxa. The exact definition is situational: the first article uses estimates of the expected mean squared error for the variance within taxa and the mutational variance scaled by time for the variance between taxa. The second uses the variance of a nesting factor (species) and the nested factor (line). The last uses the variance among the population means and the variance within populations. The variance models follow in standard variance component notation,

$$\begin{aligned} Var(Y_{gr}) &= \sigma^2_{\text{error}}\mathrm{I}_T && \text{(Rifkin et al. 2003)} && (13)\\ &= \sigma^2_{\text{species}}\mathrm{I}_T + \sigma^2_{\text{line (species)}}\mathrm{I}_T && \text{(Nuzhdin et al. 2004)} && (14)\\ &= \sigma^2_{\text{pop.}}\mathrm{I}_T + \sigma^2_{\text{individual (pop.)}}\mathrm{I}_T && \text{(Whitehead and Crawford 2006)} && (15) \end{aligned}$$

In all cases, the intuitive interpretation is that the estimate of between taxa variance is the numerator and the within taxa variance is the denominator of the ANOVA F-test of interest. We call this mean square estimation for variance based (mutational variance/rate) testing. In noting that none of the ANOVA based methods take into account V_0 or Λ_0 (they all use some form of I_T), we wish to emphasize that none of the three techniques accounts for phylogeny in the estimation of the mutational variances. (Rifkin et al.'s method (2003) uses estimated distances in the computation of expected mean squares, but this is not the same.) In order to apply the ANOVA estimates, we must believe that the data are independent and identically distributed across taxa.

Since each method tries to derive the variance component under neutral drift, the hypothesis that a gene's expression divergence is explainable by neutral drift alone makes a

natural null hypothesis. Then, deviations from the null hypothesis are deviations from neutral drift and alternative hypotheses indicate evidence for natural selection. Each method has a different interpretation for the evidence of a particular type of natural selection. For their mutation-drift test, Rifkin et al. (2003) only observe that rejected hypotheses show evidence of selection. Nuzhdin et al. (2004) identify genes with both variance estimates low as undergoing stabilizing selection; genes with low F-ratios may be undergoing balancing selection; genes with large F-ratios may undergo adaptive divergence. Whitehead and Crawford (2006a) add the constraint that genes undergoing adaptive divergence ought to favor a particular direction, i.e., correlate with an additional environmental covariate. (Whitehead and Crawford (2006a) actually propose an "Inverse-F" test, but note that this is the same as considering the lower tail of the usual F-test. The term may be an anachronism, without the inverse form, one would need two sets of statistical tables to conduct two tailed tests. It is sufficient to consider a "two tailed" F-test.)

In our phylogenetic corrective parametrization, recall that the marginal variance is given by (10). Then, τ^2 is the analog of the gene specific between taxa variance ("numerator of the F-ratio") and σ^2 is the analog of the gene specific within taxa variance ("denominator of the F-ratio"). Since σ^2 is interpreted as the rate of mutation in the expression trait (Rifkin et al., 2003; Nuzhdin et al., 2004; Whitehead and Crawford, 2006a), the relative sizes of τ^2 and σ^2 imply different evolutionary scenarios. When $\tau^2 = \sigma^2$, the signal is consistent with a Brownian motion process evolving along the given tree, representing the neutral drift null hypothesis. If we conclude that $\tau^2 < \sigma^2$, the observed dependence is consistent with slower than expected variation (less expression divergence) suggesting that the gene may be undergoing balancing selection. Similarly, $\tau^2 > \sigma^2$ favors directional selection since the observed divergence is larger than expectation (we relax the requirement that the residuals must also show correlation with environmental covariates, i.e., a that they show a particular direction as well). If τ^2 and σ^2 are both "small" then we conclude that there is evidence of purifying or stabilizing selection.

In the following sections, we attempt to demonstrate that the naive mean square estimate does not behave exactly as required in the mixture framework. Whitehead and Crawford (2006b) observed that the ANOVA method tries to approximate the function that relates within and between taxa variance. Such a function depends on gene-specific constraints and divergence times, but a standard ANOVA estimate ignores the latter.

3 Results

3.1 The Need for Corrections

We construct the following simulation study to illustrate the cost of failing to correct a phylogenetic signal and the effect on the statistical evolutionary hypotheses posited above.

Suppose we have the following phylogeny structure encoded in a covariance matrix:

$$V_0 = \begin{pmatrix} 5 & 4 & 3 & 1 & 0 \\ 4 & 7 & 3 & 1 & 0 \\ 3 & 3 & 7 & 1 & 0 \\ 1 & 1 & 1 & 5 & 0 \\ 0 & 0 & 0 & 0 & 8 \end{pmatrix}.$$

[Figure 2 about here.]

Under the mixture model proposed above, we define the selection hypotheses in the table below and draw a simulation dataset. We construct an artificial array of 350 genes (50 per hypothesis) and an artificial experiment where each gene is measured 500 times (5 taxa in V_0 above, 100 individuals). λ is generated by sampling a Uniform(0,1) random variable once for each of the 350 genes.

Hypothesis	τ^2	σ^2	Number of Genes	Plot Color
Neutral Drift	1.00	1.00	50	black
Balancing Selection, Weak	1.00	5.00	50	light red
Balancing Selection, Strong	1.00	10.00	50	dark red
Directional Selection, Weak	5.00	1.00	50	light blue
Directional Selection, Strong	10.00	1.00	50	dark blue
Stabilizing Selection, Weak	0.10	0.10	50	light green
Stabilizing Selection, Strong	0.05	0.05	50	dark green
			350 genes total	

The mean square approach generates estimates from the following nested ANOVA table. The proper reference for these estimates is the F-distribution with 4 and 95 degrees of freedom.

Source	df
Taxa	5-1 = 4
Individual (Taxa)	100-5 = 95
Error	499-99 = 400

The resulting "idealized" data is analyzed in Figure 3 which plots the logged values of τ^2 and σ^2 under the scenarios tabled above. The seven versions of the variance based hypotheses are color coded: the black points represent a neutral drift null scenario, the two shades of blue are genes undergoing strong and weak directional selection; two shades of red, balancing selection and two shades of green, stabilizing selection. Two grey lines indicate the 0.05 two-sided thresholds for the F-ratio test. Points above the upper threshold are declared to show evidence of directional selection. For the sake of argument, we omit the condition that genes undergoing directional selection need to correlate with an environmental covariate. Points below the lower threshold show evidence of balancing selection. We do not implement the corresponding stabilizing selection test.

[Figure 3 about here.]

The top two panels illustrate the same data when V_0 captures the true correlation between taxa. The true τ^2 and σ^2 is the same for every gene in the same group, so the spread of points represents sampling variability (and to some extent the effect of λ). The plot on the left (3a) shows the ANOVA estimates and on the right (3b) shows the estimates from the mixture model. Intuitively, both variance estimation procedures partition the total observed variance into within and between taxa parts. Since we assume the mixture model is the true generating model, we can see that the ANOVA estimates tend to over estimate σ^2 and make up for the excess by increasing the variance in the estimate of τ^2. In a joint bias-variance tradeoff, the ANOVA estimate trades low variance in the σ^2 estimate for bias and higher overall error in the τ^2 estimate.

The left hand plots (panels 3a and 3c) employ the mean square estimates for the between and within taxa variances. In plot 3a, since all the groups of genes in each class of hypotheses are centered about the identity line, it is clear that choosing genes using their F-ratio is not specific for the directional alternative or the balancing alternative; some genes from each group fall above or below the corresponding threshold. This pattern holds even in the lower left plot (3c) where we assume the data really are independent and identically distributed.

In addition to problems with the hypothesis tests, the estimates of σ^2 and τ^2 appear over-estimated when we do not account for the dependence structure. In Figure 3a and 3c, the neutral drift cluster and the stabilizing selection clusters (which are and should be centered on the identity line), appear biased much farther up the identity line than they should.

Contrast these observations with the estimates from with the mixture model (Panel 3b). The plot shows what we would ideally like to see: all the genes clearly separate based on the true values of their parameters. The effects are clearly separated implying that there are a sufficient number of replicates to identify all the effects. Note that this scenario represents artificially ideal conditions: a large number of observations, good separation, each gene class has the same true parameter. The point is that the mean square estimates do not behave as expected, even under this optimal setting. In practice, we might expect each gene to have a different set of parameters (τ^2 and σ^2) and the groups to overlap significantly. Furthermore, the proportion of genes undergoing natural selection may alter the plot significantly. That is, we do not expect to see nicely separated groups in practice, the actual form depends on the proportion of genes under each type of selection and their relative strengths.

We noted previously that Housworth et al. (2004) observe that the small number of replicates available in comparative experiments may not reach the statistical power necessary to make strong inferences. For gene expression data, we observe elsewhere (Eng et al., 2008) that a per gene analysis may also suffer from low power, but propose that clustered analyses may use similar genes to generate additional power. That is, if we believe that several genes act in concert and are willing to draw selection inferences on a whole group (i.e., a each member of a group undergoes the same selection force versus a single gene under a unique force) then we may employ genes as identical replicates in order to increase the power of the test.

3.2 Calibration problem

As we discussed in the methods section, the mixture model relies on a V_0 matrix that captures the phylogenetic relationship between the taxa. Additionally, tests of selection hypotheses require a model that preserves the relationship between τ^2 and σ^2. Since estimates of the phylogenetic tree are typically obtained from sequence information (or similar independent sources), there is no reason to believe that it is of the appropriate scale for expression level data. If the given covariance structure is scaled too small then estimates of τ^2 will be artificially large; likewise if the given covariance is too large, τ^2 will be too small.

One simple correction is to add an additional scale factor to the mixture. Suppose instead of $(\mathrm{V}_0, \Lambda_0)$, the true generating covariance had components $\mathrm{V}_a = a\mathrm{V}_0$ and $\Lambda_a = a\Lambda_0$ for $a > 0$.

$$\begin{aligned} Var(Y_{gr}) &= p_1(\kappa \mathrm{V}_a) + p_2(\kappa \Lambda_a) + p_3(\kappa \mathrm{I}_T) \\ &= p_1(\kappa a \mathrm{V}_0) + p_2(\kappa a \Lambda_0) + p_3(\kappa \mathrm{I}_T) \\ &= p_1(\kappa_1 \mathrm{V}_0) + p_2(\kappa_1 \Lambda_0) + p_3(\kappa_2 \mathrm{I}_T) \end{aligned} \tag{16}$$

Keeping in mind that the mixture-based EM algorithm will estimate the scales of the individual components (κ_1, κ_2), we can obtain an identifiable estimate for all the parameters, separating the effect of a and τ^2. Details of this EM algorithm are available in Appendix B.

We need to emphasize that the mixture assumption, made in section 2.2, is necessary to obtain an identifiable estimate of a. Had we assumed a marginal mixed effects model (Lynch, 1991; Martins and Hansen, 1997; Guo et al., 2006) with the same variance (Equation 5), the scale parameter and the variance would only be estimable as $a\tau^2$. Practically, the investigator would have to assume some value of a in order to conduct selection tests, but this would create an uncorrectable bias in the testing framework.

Figure 4 summarizes the scaling problem and this correction using the same set of seven hypotheses from the previous section. For these plots, we assume that we know V_0, the same as in the last section, but that data are generated from scaled versions of V_0. In the left panels (4a and 4d), the true $\mathrm{V}_{1/100} = \mathrm{V}_0/100$ is 100 times smaller than the given V_0. In the middle (4b and 4e), the scale is correct. In the right panels (4c and 4f), the true $\mathrm{V}_{100} = 100\mathrm{V}_0$ is 100 times larger than the given V_0.

The top row of Figure 4 demonstrates the effect of the wrong sized covariance by fitting the mixture with V_0 given. Notice that estimates are drawn uniformly downwards in panel 4a but pushed upwards in panel 4c. For reference panel 4b is the same plot from Figure 3b. Fewer points appear in the latter plot since estimates may be unobtainable when this scaling is too far off. The bottom row of Figure 4 shows the effect of estimating nuisance scale a for large and small true values.

[Figure 4 about here.]

It makes sense that the procedure fails for a very small, since this case corresponds to the scenario where the heritable component is weak, i.e., there is very little signal. At present, this case can be identified by observing an unusually large proportion of genes for which $\lambda = 0$ since very small a forces λ to shrink even if the signal is present.

Oakley et al. (2005) anticipated this problem in constructing their various model types: in one model the known covariance structure enters in assuming a is a unit rate, in another, different scalings of the branch lengths correspond to a different value for a. It is useful at this point to observe that $\lambda = 0$ corresponds to Oakley's class of "non-phylogenetic" models and $\lambda = 1$ to the "pure-phylogenetic" models. Also, the need for a common scale appears in the methodological notes in Rifkin et al. (2003) as an estimate of the common mutational variance. These interpretations lead to the idea that if a is a common mutational rate, it is natural to think of $\tau^2 > 1$ or $\sigma^2 > 1$ indicating a higher mutational rate between and within taxa versus a rate common to all genes (likewise $\tau^2 < 1$ and $\sigma^2 < 1$ indicate slower mutational rates).

3.3 *Saccharomyces* Data Example

The application of phylogenetic techniques to gene duplication families in the yeast *Saccharomyces cerevisiae* supposes that individual genes' sequences are linked to a common ancestor sequence, the target of a duplication event, and that the expression level of the descendent sequences is itself a trait subject to evolutionary forces (Gu, 2004; Oakley et al., 2005). In such an analysis, the members of these families constitute the taxa of interest and the models developed in Gu (2004) and Oakley et al. (2005) test how well a sequence derived covariance matrix matches the predicted history of the expression trait. Since there is good reason to expect a phylogenetic structure between the genes, we will re-analyze the data set presented in Oakley et al. (2005) to illustrate the mixture model.

Using Gu's procedure (2004) for searching the proteome to identify 10 large gene families (between 7 and 18 genes each), Oakley et al. (2005) process expression arrays from 19 experiments from the Stanford Microarray Database (`http://genome-www5.stanford.edu`) and compute maximum likelihood phylogenetic trees for each family. Each experiment represents a different experimental condition, so we may draw inferences about the evidence of selection under particular conditions. There are 19 experiments each of which contains some of the 10 gene families for a total of 169 family specific measurements. Each experiment is a separate dataset where g corresponds to a gene family, t a single transcript in the gene family and r an array in the experiment. The data are analyzed are available from the supplementary materials from Oakley et al. (2005) (`http://www.lifesci.ucsb.edu/eemb/labs/oakley/pubs/MBE2005data/`).

First consider the application of an ANOVA model which assumes that the residuals from its fit will be independent and identically distributed (iid). For each gene family in Table 1, the maximum residual correlation between all pairs of taxa over all replicates in all experiments demonstrates that the residuals are frequently not independent (8 out of 10 families have correlation greater than 0.50 in at least one pair of taxa) and Levene's test for the homogeneity of variances rejects the identically distributed assumption for 6 of the 10 families. These observations reinforce the need for an adjustment to account for the violation of the iid assumptions.

[Table 1 about here.]

The same 10 gene families appear in Table 2 which lists the number of experiments in which the gene family was measured, the number of these experiments which show some

evidence of phylogenetic signal ($\lambda > .5$) and the number of experiments which may have significant balancing or directional selection tests. The last two are defined by the same reference F distribution for the ANOVA estimates. Note that this is not a proper test, the reference distribution for the "corrected F" estimate is presently unclear so the table serves as a heuristic for comparing the mixture model and ANOVA estimates.

[Table 2 about here.]

The plot in Figure 5 shows the ANOVA estimates and the mixture model corrected estimates plotted on log scale (τ^2 is between taxa and σ^2 is within taxa). As in the simulation plots, points about the identity line favor neutral expectations and points significantly distant from the line favor selection hypotheses. The ANOVA estimates appear to have a strong trend where τ^2 is smaller than expected, reflecting the tendency of the ANOVA estimate to favor σ^2 at the cost of shrinking τ^2 to zero if necessary (we saw this same pattern in Figure 3). The mixture estimates are more in line with neutral expectations. We have highlighted extreme points (ones which have large or small τ^2/σ^2 ratios when properly corrected. In the left hand plot, notice that the pattern is fairly random suggesting that the raw ANOVA estimates will lead to incorrect inferences.

[Figure 5 about here.]

Both Gu (2004) and Oakley et al. (2005) proposed models for gene family data similar to our decomposition into phylogenetic (V_0), independent (Λ_0) and iid (I_T) hypotheses, but these models differ from ours in that they attempt to choose a single best fitting hypothesis. To underscore the difference between the testing frameworks consider Table 3, where we list the experiments which showed some evidence of a particular type of selection. Keeping in mind that the test is not formal, note that the inferences possible under mutational rate testing (with τ^2 and σ^2) highlight types of selection while Oakley's relative rate testing picks a best fitting model. The choice between the "pure-phylogenetic distance" and "non-phylogenetic distance" models is analogous to the choice between $\lambda = 0$ and $\lambda = 1$ assuming $\sigma^2 = 0$.

[Table 3 about here.]

Concordant with the finding in Oakley et al. (2005) that most families have a "non-phylogenetic" model in different experimental conditions (117 of 152), a large proportion of experiments corrected with the mixture model show weak phylogenetic signal, $\lambda < .5$ (115 of 169). This raises some questions about how to interpret the results since Whitehead and Crawford (2006a) only defined selection scenarios for τ^2 and σ^2 supposing that $\lambda = 1$. We do find, however that the Hexose Transport gene family appears to show strong phylogenetic signal in 8 of 14 experiments versus 12 of 14 in Oakley et al.'s analysis (2005). This family is also strongly represented in the balancing selection list (10 of 14 experiments).

The equivalence with Oakley et al.'s models (2005) may give us some idea about how to interpret the results taking λ into account. Under strong signal ($\lambda = 1$), the pure-phylogenetic model in Oakley et al. (2005) agrees with the $\tau^2 = \sigma^2$ neutral drift hypothesis, while the $\lambda = 0, \tau^2 = \sigma^2$ case agrees with the non-phylogenetic model. Of the latter they give

the possible evolutionary implication: "genetic distances since last gene duplication predict change in expression, consistent with an initial coupling during evolution of expression and coding sequence" (pp47). We might hypothesize that this aligned evolution holds for genes which have small λ values and further that, since τ^2 is the rate associated with V_0, it reflects the relative variance of the expression phenotype over the rate of divergence in coding sequence while σ^2 reflects the relative rate of variation after the effect of this coupling fades in evolutionary time.

4 Discussion

The use of gene expression microarrays for studying heritable variation on a genome-wide scale yields a fine-resolution look at the interplay between natural selection and neutral drift. The ability for investigators to discern more subtle effects is severely limited by the ability of the models to identify and remove extraneous non-heritable noise and confounding non-phylogenetic signal. For these comparative gene expression studies, we have presented a mixture model which attempts to address precisely this fault. In adapting existing models of the phylogenetic variance for use in gene expression data, our model's primary innovation is the use of Pagel's $V(\lambda)$ matrix (1999) to decompose the heritable signal into phylogenetic and non-phylogenetic components. This model readily parameterizes between and within taxa variances and therefore provides a framework for studying evolutionary hypotheses defined by their mutational variance. We illustrated the drawbacks of estimating relative sizes of within and among taxa variances without a phylogenetic correction and demonstrated via simulation that an uncorrected analysis can lead to incorrect identification of the strongest drift/selection hypothesis. Further this model may be implemented for the analysis of data using independent contrasts, phylogenetic generalized least squares type methods as well as for data using likelihood based models. We close this section with a discussion of a few outstanding points.

4.1 Signal Detection

Mechanically, the EM solution for the mixture formulation attempts to "classify" each observed replicate according to the given tree structure V_0, so we can rely on statistical intuition from classification problems. If V_0 is too similar to a diagonal matrix, it is too similar to Λ_0; likewise, if it is nearly the identity matrix I_T, none of the three can be easily separated. These cases are easily identified by inspection, but, as a rule of thumb, we suggest looking carefully at V_0 if all genes have estimates of λ at about $1/2$ since this is a sign from the estimation algorithm that something may be wrong. Similarly, if all genes have an estimate of λ near 0 then the chosen V_0 may be too small and the real effect of any shared history may be negligible.

4.2 Measurements of phylogenetic signal

Freckleton et al. (2002) and Housworth et al. (2004) disagree on the interpretation of λ and h^2. Through the decomposition due to Christman et al. (1997) and the representation

of V(λ) in this paper, we can begin to sort out their differences. Arguing by algebra, the following are equivalent parameterizations for the partitioned variance of the trait (T).

$$\begin{array}{ccccccc} \text{Var P} & + & \text{Var S} & + & \text{Var E} & & \text{(Christman et al., 1997)} \\ & & h^2\text{V} & + & (1-h^2)\text{I}_T & & \text{(Housworth et al., 2004)} \\ & & \tau^2\text{V}(\lambda) & + & n/a & & \text{(Freckleton et al., 2002)} \\ p_1\kappa\text{V}_0 & + & p_2\kappa\Lambda_0 & + & p_3\kappa\text{I}_T & & \text{(Mixture Model)} \end{array}$$

Note that in Freckleton et al.'s (2002) description, V(λ) is evaluated without decomposition. In Housworth et al.'s (2004) model, V(λ) is denoted V since it has no parameterized equivalent, and we have omitted a scale factor of $\tau^2 + \sigma^2$ for clarity. So, Housworth et al.'s (2004) claim to the equivalence of h^2 and λ is understandably incorrect; Freckleton et al.'s (2002) parametrization only partitions Var P and Var S implicitly (so implicitly that the authors argue that the utility of V(λ) derives from a transforming and not partitioning interpretation, but this transformation is, in fact, also a partition), and the representation in our model makes it explicit. However, as we observed before, Freckleton et al.'s (2002) claim that λ measures something different than h^2 is preserved even if the variance can be partitioned since it breaks the variance not into (P+S) and E parts, but separates P from S. Thus, in our model λ tells us about the relative strength of the tree component, and $1 - h^2$ tells us about the relative strength of the error.

We suggest λ as a measure of the phylogenetic signal and the need for correction in the model. Housworth et al. (2004) offers h^2 the phylogenetic heritability as the same measure but they refer to subtly different things. From above, λ separates the phylogenetic from the non-phylogenetic (Var P from Var S) and h^2 separates the heritable (Var P + Var S from Var E). If the heritable signal of the observed data is weak then λ is less important (in the calibration problem, we saw empirically that λ approaches zero as the heritable signal weakens). If the signal is strong then h^2 is not so discriminating and the importance of a proper correction and therefore λ rises. It may also be observed that h^2 is a one-to-one transformation of the ratio τ^2/σ^2 which means that h^2 has selection inference implications.

4.3 Interpreting non-phylogenetic signal

We wish to follow up the earlier conjecture about the relationship between λ, σ^2 and τ^2 when their estimates favor the non-phylogenetic model. In the style of Gu (2004), we may write the covariance matrix in equation (10) explicitly for a fixed V_0. Suppose the ith, jth entry of V_0 is v_{ij} then the marginal covariance has the following form for components y_i:

$$Var(y_i) = \tau^2 v_{ii} + \sigma^2, \tag{17}$$

$$Cov(y_i, y_j) = \tau^2 \lambda v_{ij}. \tag{18}$$

Since the non-phylogenetic model is characterized by $Cov(y_i, y_j) = 0$ for every off diagonal entry, there are two scenarios: either $\lambda = 0$ or σ^2 is very large. Large σ^2 implies some combination of rapid rate of expression evolution or a long evolutionary time since the last split in the tree. When we compute the calibration factor a, we make σ^2 and τ^2 comparable absolute rates of evolution, so non-phylogenetic settings are characterized by $\lambda = 0$ alone.

Oakley et al. (2005) had given two reasons for the over performance of non-phylogenetic models and the under performance of the phylogenetic models. First, if the family has little influence over physiological functions, noise might dominate the across taxa signal. Because our decomposition allows the separation of noise from non-phylogenetic signal, genes families for which noise is strong appear below the identity line in the right panel of Figure 5. Since λ measures the support for phylogenetic models, we may determine which of these families show small phylogenetic signal and which show evidence of selection. This is equivalently the idea that an un-calibrated σ^2 is very large or that the heritable signal is weak versus the non-heritable signal. Recalling Figure 1, the non-phylogenetic scenario suggested by Oakley et al. (2005) can be explained by lengthening the dotted segments (σ^2) until they dominate the shape of a truly phylogenetic ($\lambda = 1$) tree.

Second, the distances involved in tree building might have been poor estimates for the true correlation among genes. This problem can be addressed using the phylogenetic minded approach in Corrada Bravo et al. (2008) to attempt to estimate the best fitting tree for the given expression trait. One imagines that if that tree appears very similar to the sequence based tree then there is strong evidence that λ is not zero.

4.4 Extensions

Since this model only considers decompositions of the variance, we can augment it with the application of standard statistical linear model theory to accommodate much more complicated experiments. In time course expression experiments, this form of linear model modeled the correlation over time (Eng et al., 2008), approximating gene associations by clustering similar genes together. For comparison, Oakley et al. (2005) corrected for correlated adjacent time points by using the first order differences, while Gu (2004) found the effect negligible. It is not unbelievable that more complex factors like expression under various conditions/treatments across taxa will be of interest and the model we have presented may serve as a useful component in that analysis.

These models anticipate the use of microarray platforms to make general inferences about the strength of evidence for natural selection forces. As described previously, mutational variability/mean square type selection inference relies on estimating τ^2 and σ^2. We have paid less attention to the frameworks put forward in Gu (2004) and Oakley et al. (2005) which define natural selection on the basis of a likely history parameterized in the form of V_0. This highlights the transformative role of λ, τ^2 and σ^2 in the sense that they also find a tree-structured covariance matrix consistent with observed data.

As an alternative to the Brownian model, Freckleton et al. (2002) considered the interpretation that $\lambda = 0$ corresponds to a rapid or instantaneous response to selection. When one considers the Ornstein Uhlenbeck process (Butler and King, 2004) in a setting where selection towards some optimum trait value overpowers the random perturbation one should see a similar independent, diagonal covariance. (In fact, one can show that the covariance approaches the identity matrix, not Λ_0, in the limit Smith et al. (2008)). Thus we could consider that, when $\lambda = 0$, selection hypotheses refer to a very rapid reversion to the fitness optima (completely non-phylogenetic). We intend to return to this alternative class of models in a later article.

The mixture model is ideal for the gene duplication family data structure because arrays

represent independent replications of the same experiment. In a more general experiment, we imagine similar models that structure dependence among all observations may be developed.

One point we have left ambiguous is the proper estimation of nuisance parameter a which we believe is common to all genes but we estimate once for each gene. This is clearly inefficient and we intend a future statistical paper on its estimation across genes.

Acknowledgements

We thank Carol Lee (UW-Madison) and Anthony Ives (UW-Madison) for their feedback on earlier drafts of this article and Andrew Whitehead (Louisiana State University) for discussion on his experiments. This work was supported by NSF grants DMS 0604572(GW) and DMS 0804597 (SK); a PhRMA Foundation Research Starter Grant (SK); NIH grants HG03747 (SK) and EY09946(GW) and ONR Grant N0014-06-0095 (GW).

References

Blomberg, S. P., T. G. Garland, and A. R. Ives. 2003. Testing for Phylogenetic Signal in Comparative Data: Behavioral Traits are More Labile. Evolution **57**:717–745.

Butler, M. A., and A. A. King. 2004. Phylogenetic Comparative Analysis: A modeling approach for adaptive evolution. The American Naturalist **164**:683–695.

Cheverud, J. M., M. M. Dow, and W. Leutenegger. 1985. The Quantitative assessment of phylogenetic constraints in comparative analyses: sexual dimorphism in body weight among primates. Evolution **39**:1335–1351.

Christman, M. C., R. W. Jernigan, and D. Culver. 1997. A comparison of two models for estimating phylogenetic effect on trait variation. Evolution **51**:262–266.

Corrada Bravo, H., K. H. Eng, S. Keleş, G. Wahba, and S. Wright. 2008. Estimating tree-structured covariance matrices via mixed integer programming with an application to phylogenetic analysis of gene expression. Technical Report 1142, University Of Wisconsin-Madison, Department of Statistics.

Dempster, A. P., N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society **39**:185–197.

Eng, K. H., S. Keleş, and G. Wahba. 2008. A Linear Mixed Effects Clustering Model for Multi-Species Time Course Gene Expression Data. Technical Report 1143, University Of Wisconsin-Madison, Department of Statistics.

Fay, J. C., and P. J. Wittkopp. 2007. Evaluating the role of natural selection in the evolution of gene regulation. Heredity Advance publication May 23,2007.

Felsenstein, J. 1973. Maximum-Likelihood Estimation of Evolutionary Trees from Continuous Characters. American Journal of Human Genetics **25**:471–492.

———. 1985. Phylogenies and the comparative method. The American Naturalist **125**:1–15.

———. 1988. Phylogenies and the Quantitative Characters. Annual Review of Ecological Systems **19**:447–471.

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

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

Guo, H., R. E. Weiss, X. Gu, and M. Suchard. 2006. Time squared: repeated measures on phylogenies. Molecular Biology and Evolution Advance access: November 1, 2006.

Housworth, E. A., E. P. Martins, and M. Lynch. 2004. The Phylogenetic Mixed Model. The American Naturalist **163**:84–96.

Ives, A. R., P. E. Midford, and T. Garland. 2007. Within-Species Variation and Measurment Error in Phylogenetic Comparative Methods. Systematic Biology **56**:252–270.

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

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

Martins, E. P., and E. A. Housworth. 2002. Phylogeny shape and the phylogenetic comparative method. Systematic Biology **51**:873–880.

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

Oakley, T. H., Z. Gu, E. Abouheif, N. H. Patel, and W. H. Li. 2005. Comparative methods for the analysis of gene-expression evolution: an example using yeast functional genomic data. Molecular Biology and Evolution **22**:40–50. `http://www.lifesci.ucsb.edu/eemb/labs/oakley/pubs/MBE2005data/`.

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

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

Saitou, N. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular Biology and Evolution **4**:406–425.

Smith, S. D., C. Ané, and D. A. Baum. 2008. The role of pollinator shifts in the floral diversification of *Iochroma* (Solanaceae). Evolution **62**:793–806.

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

———. 2006b. Variation within and among species in gene expression: raw material for evolution. Molecular Biology **15**:1197–1211.

Appendix A

EM Algorithm for $\mathbf{V}_0$ and Λ_0 known for a single gene. Let C_r be the class variable for replicate $r = 1, \ldots, R$, taking values 1,2 and 3 for components $\mathcal{N}(\mu, \kappa \mathrm{V}_0)$, $\mathcal{N}(\mu, \kappa\Lambda_0)$, and $\mathcal{N}(\mu, \kappa \mathrm{I}_T)$ respectively. Suppose T is the rank of V_0. The algorithm stops when the log-likelihood increases by less than 0.001. Since the estimate μ does not depend on C_r, its estimate is $\hat{\mu} = \frac{1}{R}\sum_{r=1}^{R}(Z'Z)^{-1}Z'Y_r$ for design matrix Z (i.e., for balanced replicates these are the group means).

1. E-Step

$$\begin{aligned}\hat{C}_{ri} &= E(C_r = i | Y_r, \hat{\mu}, \hat{\kappa}^{(t)}, \hat{p}_i^{(t)}), \\ &= \frac{P\left(Y_r - \hat{\mu} | C_r = i, \hat{\kappa}^{(t)}\right)\hat{p}_i^{(t)}}{\sum_{i'=1}^{3} P\left(Y_r - \hat{\mu} | C_r = i', \hat{\kappa}^{(t)}\right)\hat{p}_{i'}^{(t)}}.\end{aligned}$$

2. M-step

$$\begin{aligned}\hat{p}_i^{(t+1)} &= \frac{\sum_r \hat{C}_{ri}}{R}, \\ \hat{\kappa}^{(t+1)} &= \frac{1}{RT}\sum_{r=1}^{R}\left[\hat{C}_{r1}(Y_r - \hat{\mu})'\mathrm{V}_0^{-1}(Y_r - \hat{\mu}) + \hat{C}_{r2}(Y_r - \hat{\mu})'\Lambda_0^{-1}(Y_r - \hat{\mu}) + \hat{C}_{r3}(Y_r - \hat{\mu})'(Y_r - \hat{\mu})\right].\end{aligned}$$

The final estimates from this algorithm are converted back to the original parametrization:

$$\begin{aligned}\hat{\tau}^2 &= \hat{\kappa}(1 - \hat{p}_3), \\ \hat{\sigma}^2 &= \hat{\kappa}(\hat{p}_3), \\ \hat{\lambda} &= \frac{\hat{p}_1}{1 - \hat{p}_3}.\end{aligned}$$

Appendix B

EM Algorithm for $\mathbf{V}_0$ and Λ_0 known up to scale a for a single gene. As in appendix A, except suppose that V_0 is known up to scale constant a. The components of interest are $\mathcal{N}(\mu, \kappa_1 \mathrm{V}_0)$, $\mathcal{N}(\mu, \kappa_1\Lambda_0)$, and $\mathcal{N}(\mu, \kappa_2 \mathrm{I}_T)$.

1. E-Step

$$\begin{aligned}\hat{C}_{ri} &= E(C_r = i|Y_r, \hat{\mu}, \hat{\kappa_1}^{(t)}, \hat{\kappa_2}^{(t)}, \hat{p}_i^{(t)}), \\ &= \frac{P\left(Y_r - \hat{\mu}|C_r = i, \hat{\kappa_1}^{(t)}\hat{\kappa_2}^{(t)}\right)\hat{p}_i^{(t)}}{\sum_{i'=1}^{3} P\left(Y_r - \hat{\mu}|C_r = i', \hat{\kappa_1}^{(t)}, \hat{\kappa_2}^{(t)}\right)\hat{p}_{i'}^{(t)}}.\end{aligned}$$

2. M-step

$$\begin{aligned}\hat{p}_i^{(t+1)} &= \frac{\sum_r \hat{C}_{ri}}{R}, \\ \hat{\kappa}_1^{(t+1)} &= \frac{\frac{1}{T}\sum_{r=1}^{R}\left[\hat{C}_{r1}(Y_r - \hat{\mu})'\mathrm{V}_0^{-1}(Y_r - \hat{\mu}) + \hat{C}_{r2}(Y_r - \hat{\mu})'\Lambda_0^{-1}(Y_r - \hat{\mu})\right]}{\sum_{r=1}^{R}\hat{C}_{r1} + \hat{C}_{r2}}, \\ \hat{\kappa}_2^{(t+1)} &= \frac{\frac{1}{T}\sum_{r=1}^{R}\left[\hat{C}_{r3}(Y_r - \hat{\mu})'(Y_r - \hat{\mu})\right]}{\sum_{r=1}^{R}\hat{C}_{r3}}.\end{aligned}$$

Again, the final estimates from this algorithm are converted back to the original parametrization:

$$\begin{aligned}\hat{\tau}^2 &= \hat{\kappa_2}(1 - \hat{p}_3), \\ \hat{\sigma}^2 &= \hat{\kappa_2}(\hat{p}_3), \\ \hat{\lambda} &= \frac{\hat{p}_1}{1 - \hat{p}_3}, \\ \hat{a} &= \frac{\hat{\kappa}_1}{\hat{\kappa}_2}.\end{aligned}$$

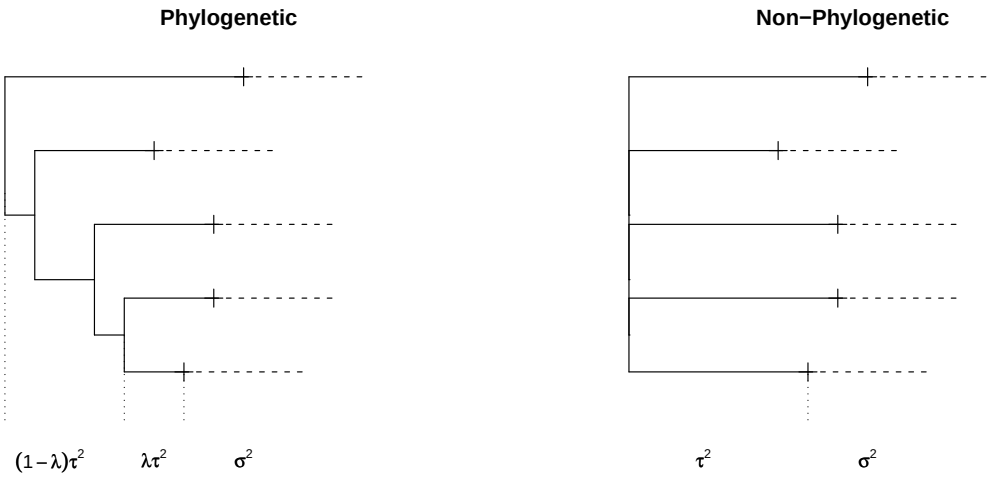

Figure 1: *Rate decompositions on the tree.* The variance decomposition may be interpreted as partitions of the taxa specific branch of the phylogenetic tree. Under phylogenetic $\lambda > 0$ and non-phylogenetic $\lambda = 0$ scenarios we show the decomposition of the rate of mutation on the bottom branch of this tree. The dotted segments correspond to a common proportion of variation attributable to σ^2.

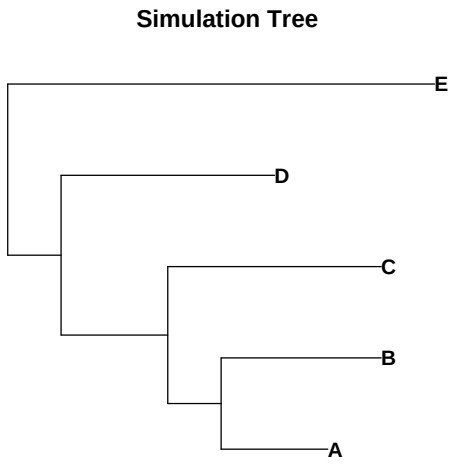

Figure 2: *Example tree for simulation study.*

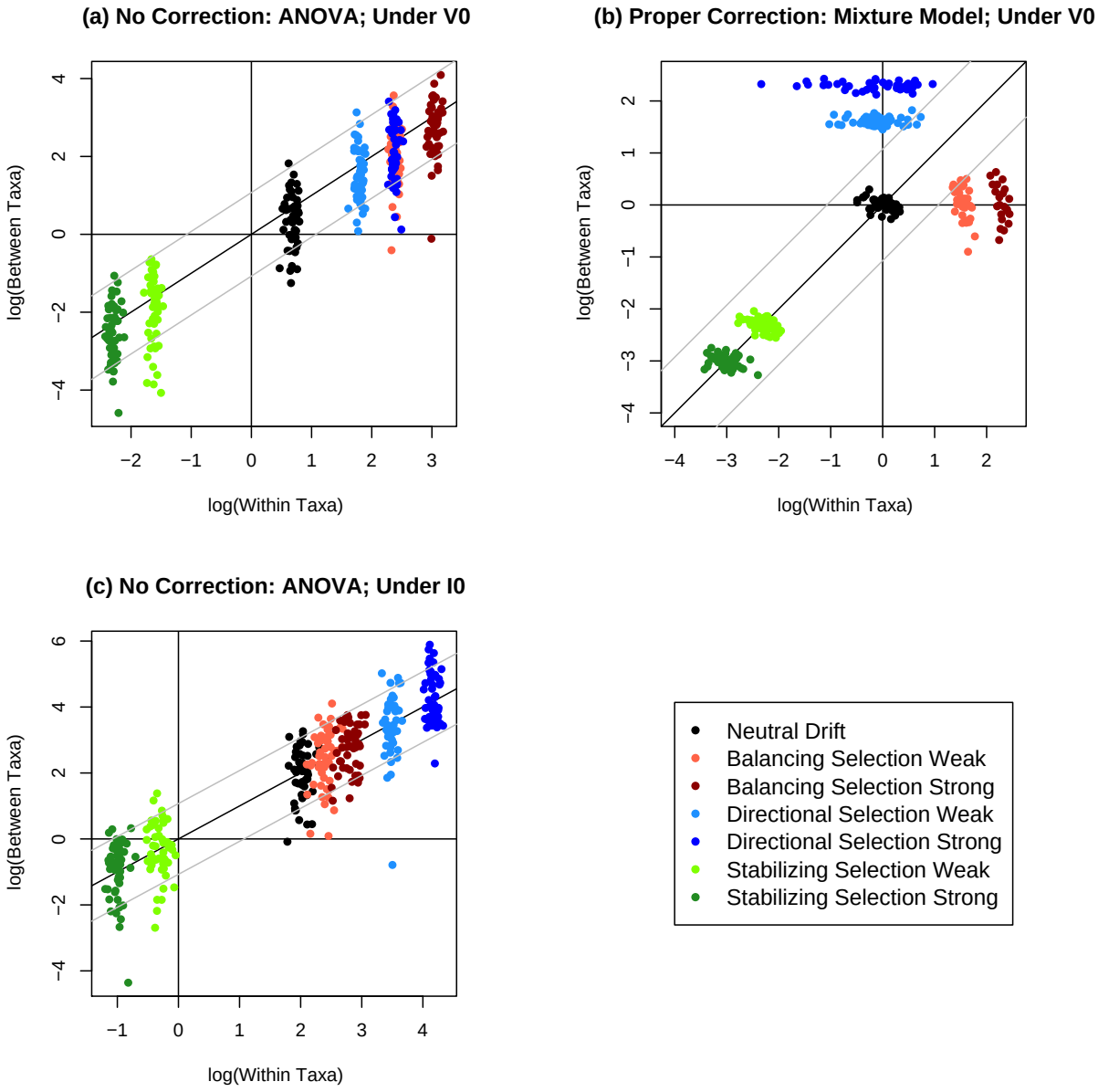

Figure 3: *Simulation Example.* Simulation under ideal settings for selection hypotheses defined in the text with a tree-structured covariance (V_0) and a non-phylogenetic covariance (I_0) shows that the ANOVA estimators do not discriminate between the hypotheses. The grey lines identify tests of selection: the corresponding two-sided F-test thresholds at $\alpha = 0.05$ for $F_{4,95}$. Panel (a) shows the ANOVA estimates of data generated under V_0; Panel (b) shows the mixture model estimates under V_0 and Panel (c) illustrates that the ANOVA estimates are inflated even under independent identically distributed characters (the primary ANOVA assumption).

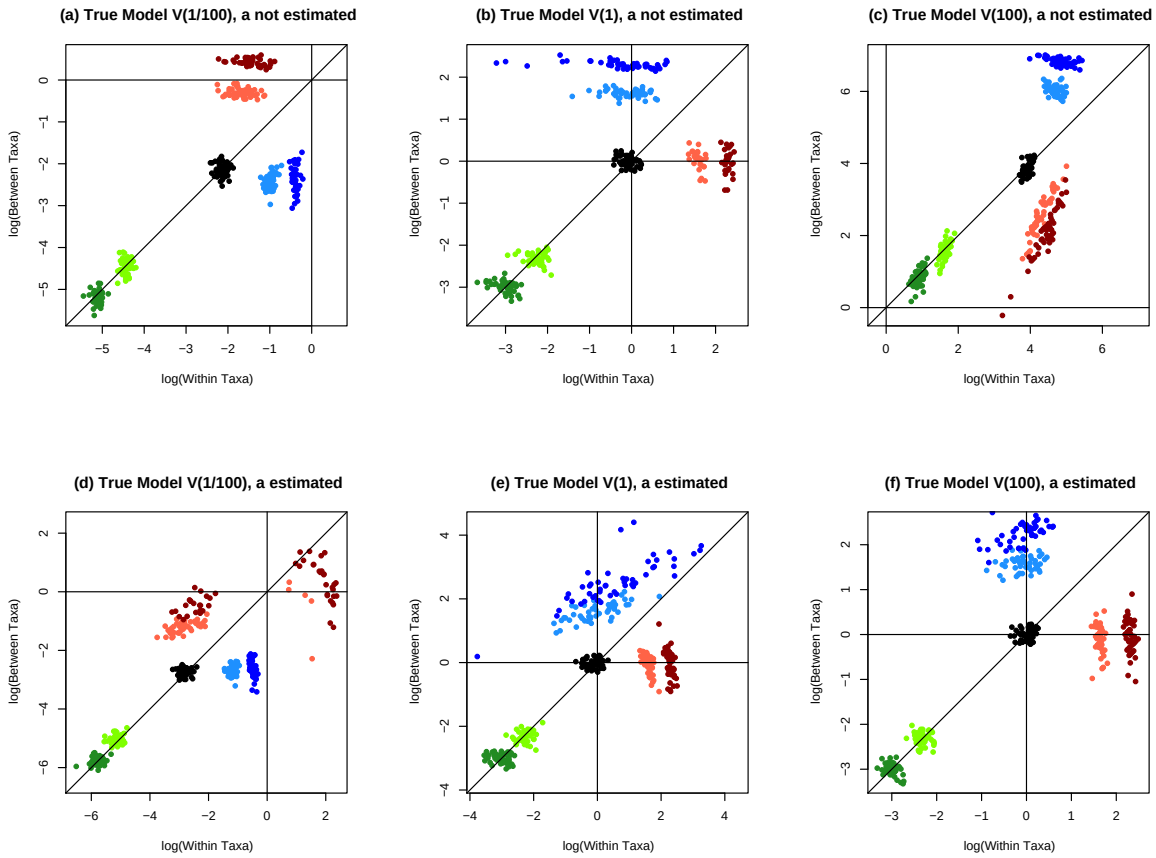

Figure 4: *Calibration Problem/Solution.* Estimation of σ^2 and τ^2 is sensitive to misspecifying the scale of the phylogenetic covariance matrix, a. When a is not accounted for (Panels a, b, c), estimates are shrunk for a small (Panel a). When a is big, estimates are too big (Panel c). The same ideal pattern as in Figure 3 appears in the top center (Panel b). Simultaneously estimating a fixes the problem (Panels e, f) for all but the smallest case (Panel d).

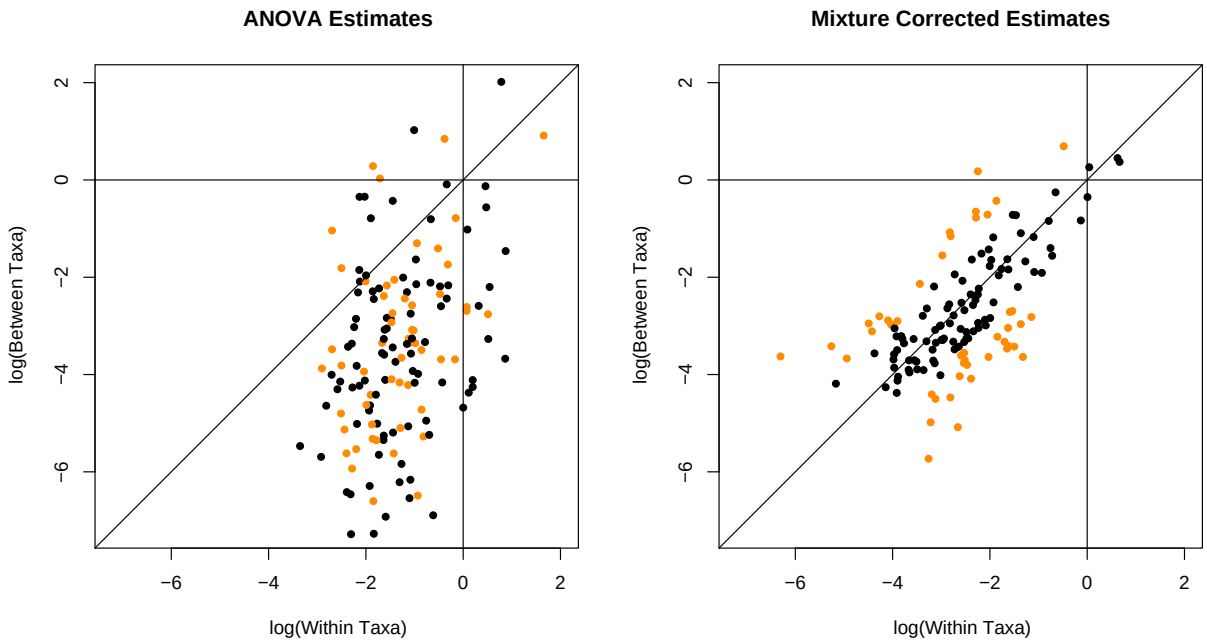

Figure 5: *ANOVA and mixture model estimates for data from Oakley et al. (2005).* Uncorrected ANOVA estimates show a marked trend towards small between taxa variances (τ^2) while corrected estimates fit more neutral expectation. The ANOVA estimates show the same low variance pattern in the σ^2 estimate as in Figure 3. Three extreme points are omitted from the right hand plot to make the scales comparable. Points which appear extreme when properly corrected are highlighted in both plots.

Gene Family	Max residual correlation	Levene's Test (p-value)
ABC Transporters	0.40	0.0093
ADP Ribosylation	0.56	0.0751
Alpha Glucosidases	0.71	0.1139
DUP	0.84	0.3998
GTP Binding	0.51	<0.0001
HSP DnaK	0.78	<0.0001
Hexose Transport	0.93	<0.0001
Kinases	0.43	0.0001
Permeases	0.60	<0.0001
Putative Helicases	0.75	0.0626

Table 1: *Diagnostics for ANOVA residuals.* The ANOVA method for estimating the mutational variances assumes that the residuals will be independent and identically distributed. The maximum residual correlation between pairs of taxa over all replicates in all experiments demonstrates that the residuals are frequently not independent (8 out of 10 have correlation greater than 0.50) and Levene's test for the homogeneity of variances shows that the identically distributed assumption holds for only 4 of the 10 families.

Gene Family (no. taxa)	No. Experiments	$\lambda > .5$	Balancing Selection Test	Directional Selection Test
ABC Transporters (8)	17	4	0	3
ADP Ribosylation (7)	17	7	0	2
Alpha Glucosidases (6)	19	4	0	1
DUP (10)	13	8	1	5
GTP Binding (11)	17	7	2	2
HSP DnaK (10)	16	1	2	4
Hexose Transport (18)	14	8	10	0
Kinases (7)	16	8	2	4
Permeases (17)	12	5	5	3
Putative Helicases (11)	11	2	3	0

Table 2: *Yeast Gene Family Data*. Gene family data analyzed under the mixture model. A small number of experiments show strong phylogenetic signal ($\lambda > .5$), while the number of experiments with ratios τ^2/σ^2 large (directional) or small (balancing) are tabulated above.

Gene Family	Experiments with favoring Balancing Selection
DUP	H2O2
GTP Binding	AminoAcidStarvation, CellCycleElu
HSP DnaK	CellCycle15, CellCycleAlpha
Hexose Transport	CellCycle15, CellCycleAlpha, CellCycleElu, DTT2, AminoAcidStarvation, Menadione, Diamide, NitrogenDeletion, Sorbitol, YPD
Kinases	DiauxicShift, Sporulation
Permeases	CellCycleElu, CellCycle28, CarbonChange, Sorbitol, Menadione
Putative Helicases	AminoAcidStarvation, CarbonChange, Menadione

Gene Family	Experiments with favoring Directional Selection*
ABC Transporters	CellCycle15, CellCycle28, Diamide
ADP Ribosylation	CellCycleAlpha, DiauxicShift
Alpha Glucosidases	YPD
DUP	DTT2, Sorbitol, Menadione, Zinc, CellCycle28
GTP Binding	DTT2, H2O2
HSP DnaK	DiauxicShift, DTT2, NitrogenDeletion, Sorbitol
Kinases	CellCycle15, DTT, H2O2, HeatShock2
Permeases	DTT2, DiauxicShift, Zinc

Table 3: *Lists of experiments used in Oakley et al. (2005) with selection evidence.* These tables contain the lists of experiments which showed evidence of selection for the listed gene family. *The directional selection test uses only the ratio of τ^2 and σ^2, it does not check for correlation with an environmental covariate.