The Phenome-Wide Distribution of Genetic Variance
Author(s): Mark W. Blows, Scott L. Allen, Julie M. Collet, Stephen F. Chenoweth, Katrina McGuigan
Source: *The American Naturalist*, Vol. 186, No. 1 (July 2015), pp. 15-30
Published by: The University of Chicago Press for The American Society of Naturalists
Stable URL: http://www.jstor.org/stable/10.1086/681645
Accessed: 07/10/2015 20:40

vol. 186, no. 1 the american naturalist july 2015

The Phenome-Wide Distribution of Genetic Variance

Mark W. Blows,* Scott L. Allen, Julie M. Collet, Stephen F. Chenoweth, and Katrina McGuigan

School of Biological Sciences, University of Queensland, St Lucia 4072, Australia

Submitted October 28, 2014; Accepted January 22, 2015; Electronically published May 12, 2015

abstract: A general observation emerging from estimates of additive genetic variance in sets of functionally or developmentally related traits is that much of the genetic variance is restricted to few trait combinations as a consequence of genetic covariance among traits. While this biased distribution of genetic variance among functionally related traits is now well documented, how it translates to the broader phenome and therefore any trait combination under selection in a given environment is unknown. We show that 8,750 gene expression traits measured in adult male *Drosophila serrata* exhibit widespread genetic covariance among random sets of five traits, implying that pleiotropy is common. Ultimately, to understand the phenome-wide distribution of genetic variance, very large additive genetic variance-covariance matrices (**G**) are required to be estimated. We draw upon recent advances in matrix theory for completing high-dimensional matrices to estimate the 8,750-trait **G** and show that large numbers of gene expression traits genetically covary as a consequence of a single genetic factor. Using gene ontology term enrichment analysis, we show that the major axis of genetic variance among expression traits successfully identified genetic covariance among genes involved in multiple modes of transcriptional regulation. Our approach provides a practical empirical framework for the genetic analysis of high-dimensional phenome-wide trait sets and for the investigation of the extent of high-dimensional genetic constraint.

Keywords: genetic variance, **G** matrix, pleiotropy, gene expression, matrix completion.

Introduction

Genetic variation is almost invariably found in individual traits (Blows and Hoffmann 2005), but recent multivariate genetic analyses have shown that this focus on individual traits has given a misleading picture of the evolutionary potential of quantitative phenotypes (Walsh and Blows 2009). Genetic variation in combinations of traits is often very low for a substantial proportion of the phenotypic space, as revealed by geometric analyses of the genetic variance-covariance matrix (**G**; Agrawal and Stinchcombe 2009; Hine et al. 2009; Kirkpatrick 2009). The part of the phenotypic space that has very little genetic variance has been called the nearly null genetic subspace (Gomulkiewicz and Houle 2009; Houle and Fierst 2013; Hine et al. 2014), where the term null has been borrowed from the field of linear algebra, meaning the subspace of a covariance matrix that contains no variation. While there are statistical limits on our ability to demonstrate the presence of a true null subspace that contains no variation (Mezey and Houle 2005), the presence of some trait combinations with very low levels of genetic variance appears to be a common property of the distribution of genetic variance (Kirkpatrick 2009; Pitchers et al. 2014), and this nearly null subspace is likely to represent important evolutionary constraints in natural populations (Gomulkiewicz and Houle 2009; Hine et al. 2014).

A direct corollary of the presence of nearly null subspaces is the presence of other trait combinations with relatively high genetic variance, and this too has evolutionary consequences. Artificial selection experiments have shown that such a distribution of genetic variance can result in many of the individual traits under multivariate selection responding in the direction opposite to the selection gradient applied to them (e.g., Hine et al. 2011, 2014). The concentration of most of the genetic variance into only part of the phenotypic space is expected to bias phenotypic evolutionary responses arising from both drift and directional selection toward directions in phenotypic space associated with the most genetic variance (Lande 1979; Hansen and Houle 2008; Walsh and Blows 2009). Empirical studies using natural populations have demonstrated this effect, with responses to selection biased toward trait combinations associated with greater genetic variance (Schluter 1996; Chenoweth et al. 2010).

The concentration of genetic variance into fewer dimensions than the number of phenotypes measured is primarily a consequence of pleiotropy among phenotypes (Lande 1980; Johnson and Barton 2005). To date, investigations of **G** have typically been concerned with small sets of traits ($\leq$10) that are part of the same morphological structure, such as fly wings (Phillips et al. 2001; McGuigan and Blows 2007); have strong biochemical associations, such as insect cuticular hydrocarbons (Blows et al. 2004; Hine et al. 2004; Van Homrigh et al. 2007); or are part of the same gene reg-

* Corresponding author; e-mail: m.blows@uq.edu.au.

Am. Nat. 2015. Vol. 186, pp. 15–30.
DOI: 10.1086/681645

ulatory network (Innocenti and Chenoweth 2013). Strong pleiotropic relationships within such trait sets are perhaps not surprising given their shared development (Cheverud 1984). While these studies of functionally related traits suggest that most genetic variance tends to be confined to few trait dimensions, the implications of this pattern for the broader phenome-wide distribution of genetic variance remains untested. If pleiotropy is extensive, the number of trait dimensions with genetic variance may be much smaller than the number of phenotypes that can be measured on organisms (Walsh and Blows 2009).

The extent of pleiotropy at a phenome-wide scale is unclear. Some evidence from gene knockout studies suggests that pleiotropic effects might typically be highly restricted, corresponding to relatively small variational modules (Wagner and Zhang 2011). In contrast, there are two reasons to suspect that pleiotropy across the phenome might be extensive rather than restricted. The relative rate of mutation in individual traits compared to genome-wide estimates and the strength of stabilizing selection acting on highly heritable quantitative traits are both observations that are difficult to explain without extensive pleiotropy among traits, reducing the number of genetically independent traits (Johnson and Barton 2005). In support of these two inferences, mutational pleiotropy among gene expression traits is widespread, with single putative mutations affecting many traits, even when those traits are considered without regard to known biological function (McGuigan et al. 2014*b*), and mutations are under stronger stabilizing selection when they are pleiotropic (McGuigan et al. 2014*a*). However, empirical studies specifically targeting high-dimensional phenotypes are required to characterize the extent of pleiotropy across the phenome.

To begin to understand the phenome-wide distribution of genetic variance, we need to overcome two difficult challenges. First, characterizing the full set of phenotypes of an organism, termed phenomics, is a problem that has been less advanced than its genetics counterpart, genomics (Houle 2010; Houle et al. 2010). New technologies for the automated measurement of phenotypes are, however, beginning to generate larger trait sets. Gene expression traits are one particular class of traits that provides a useful entry point for phenomic study, with readily available technologies allowing the measurement of thousands of traits. Gene expression traits lie at the interface between genotype and phenotype and might underlie evolutionary diversifications in other phenotypes (Britten and Davidson 1971; King and Wilson 1975; Carroll 2008; Wittkopp and Kalay 2012). While there will clearly be ways in which individuals vary that are not captured by variation in gene expression, expression traits nevertheless represent a broad range of biological functions and have been shown to be associated with responses to selection in the field (e.g., McGraw et al. 2011; Whitehead et al. 2011; Pespeni et al. 2013) and to be genetically correlated with fitness measures under laboratory conditions (e.g., Rest et al. 2013; Runcie and Mukherjee 2013). Gene expression therefore provides perhaps the best opportunity, given current technologies, to explore the distribution of genetic variance in very-high-dimensional phenotypes.

The second challenge to be overcome if we are to understand the phenome-wide distribution of genetic variance is the estimation of high-dimensional **G**. Standard multivariate mixed-model approaches to estimating **G** for n traits typically use restricted maximum likelihood (REML) to fit an unstructured covariance matrix,

$$\mathbf{G} = \begin{bmatrix} \sigma_1^2 & \sigma_{1,2} & \cdots & \sigma_{1,n} \\ & \sigma_2^2 & & \vdots \\ & & \ddots & \vdots \\ & & & \sigma_n^2 \end{bmatrix}, \tag{1}$$

where σ_n^2 is the genetic variance in the nth trait and $\sigma_{i,j}^2$ is the genetic covariance between traits i and j. As n increases, the number of parameters to be estimated increases exponentially by a factor of $n(n+1)/2$, forming the basis of the so-called curse of dimensionality. Given the rapid increase in the number of parameters to be estimated as n increases, it would at first appear to be a daunting task to ask questions concerning the distribution of genetic variance in the hundreds or even thousands of traits necessary to understand what the distribution of genetic variance across the phenome might be. This is primarily because convergence of mixed models with dimensions as few as 10 is often difficult to obtain for reasons that will vary from study to study but are likely to include limited degrees of freedom and nonnormal trait distributions.

Several approaches have recently been advanced that attempt to accommodate large numbers of traits measured within standard quantitative genetic experimental designs (Meyer and Kirkpatrick 2005; Stone and Ayroles 2009; McGraw et al. 2011; Runcie and Mukherjee 2013). The most generalizable of these approaches are likely to be those that result in the estimation of a reduced-rank **G**, where the number of dimensions with genetic variance is constrained to be fewer than the number of traits measured and thus fewer parameters need to be estimated. The Bayesian sparse factor (BSF) approach of Runcie and Mukherjee (2013) in particular promises to be a versatile approach for identifying sparse trait combinations that explain substantial proportions of the observed genetic variance. As a consequence of the Bayesian framework, the implementation of BSF is quite complex and requires a number of simplifying assumptions, the most important of which are that (1) the eigenvectors of the resulting **G** matrix have few traits contributing to them (the sparse assumption) and (2) the residual matrix is also of a reduced rank. Both assumptions are

essentially ways of reducing the number of parameters that need to be estimated by the model.

Quantitative genetics is not alone in the need to estimate and understand the behavior of high-dimensional covariance matrices. Disciplines as diverse as ecology, mathematical physics, signal processing, and finance struggle with the problem of estimating large covariance matrices (Paul and Aue 2014), although in most cases they tend to be sample covariance matrices rather than the more derived variance component matrices that represent **G**. The use of covariance matrices to describe the distribution of variance in high dimensions rests on the additional assumption of multivariate normality (MVN), which also underlies much of quantitative genetic theory and the multivariate response to selection (Lande 1979). Substantial deviation from the MVN assumption can potentially obscure the distribution of variance across trait combinations (see fig. 1 in Mahoney and Drineas 2009 for a graphic illustration). The relative performance of covariance matrices against alternative formulations using approaches such as information theory that do not rely on the MVN assumption for high-dimensional biological applications is currently unknown.

Here, we bring together a series of recent theoretical results establishing a number of properties of large Hermitian matrices completed from smaller submatrices contained within them. We use these theoretical results to develop a framework for the estimation of a large **G** matrix ($n = 8{,}750$) of gene expression traits measured in male *Drosophila serrata*. Beginning with small matrices ($n = 5$) of a size typical of many evolutionary quantitative genetic studies, we show that genetic covariance is common among random sets of gene expression traits. Building from this low-dimensional base, we demonstrate how small submatrices can be used to complete a higher-dimensional **G** while estimating just a fraction of the elements of the larger matrix directly from the data. Having validated the method, we use the approach to estimate the 8,750-trait **G** matrix (from $n = 50$ submatrices) and present evidence for widespread genetic covariance among a very large number of expression traits. Finally, we employ gene ontology (GO) enrichment analysis and show that the widespread genetic covariance uncovered by the analysis of the 8,750-trait **G** was associated with genes involved in multiple modes of transcriptional regulation, supplying a plausible mechanism for widespread pleiotropy among gene expression traits.

Methods

We employed two complementary analytical strategies in this study to investigate the phenome-wide distribution of genetic variance. First, a large number of **G** matrices were estimated using standard methodologies for small (5- and 50-trait) sets of randomly chosen gene expression traits to determine the extent of genetic covariance among these random sets sampled from the entire set of 8,780 traits. Second, these **G** matrices were used to determine the utility of a matrix completion approach for approximating very-large-dimensional **G** matrices (here, a **G** matrix of up to 8,750 traits). We begin by outlining the problem of matrix completion and then detail the specific approach we have adapted for use in estimating very-large-dimensional **G**.

The Problem of Matrix Completion

Real symmetric matrices like **G** are a subclass of Hermitian matrices, with the latter broader category allowing for complex entries. If the *i*th rows and *j*th columns of a symmetric matrix are sampled so that $i = j$, the resulting smaller matrix is called a principal submatrix. If all n traits are allocated to a subset k in this fashion, **G** can be represented as

$$\mathbf{G} = \begin{bmatrix} \mathbf{B}_{1,1} & \mathbf{B}_{1,2} & \cdots & \mathbf{B}_{1,m} \\ & \mathbf{B}_{2,2} & & \vdots \\ & & \ddots & \vdots \\ & & & \mathbf{B}_{m,m} \end{bmatrix}, \tag{2}$$

where the $m = n/k$ $\mathbf{B}_{m,m}$ are principal submatrices positioned along the diagonal. The off-diagonal submatrices each contain the set of k^2 genetic covariances between the traits contained in any two principal submatrices. If the off-diagonal blocks remain unestimated from the data, equation (2) becomes

$$\mathbf{K} = \begin{bmatrix} \mathbf{B}_{1,1} & 0 & \cdots & 0 \\ & \mathbf{B}_{2,2} & & \vdots \\ & & \ddots & 0 \\ & & & \mathbf{B}_{m,m} \end{bmatrix}, \tag{3}$$

where **K** is an example of a block diagonal partial Hermitian matrix (Tian 2010). Clearly, the number of individual elements to be estimated in **K** is far less than in **G**. For example, for $n = 8{,}750$ and $k = 50$, as in our empirical analysis below, **K** contains 223,125 unique elements, compared to 3.83×10^7 unique elements in **G**. Therefore, only 0.6% of the unique elements of **G** are contained in **K**. Such an enormous difference in the number of elements to be estimated provides a strong incentive to find ways to complete **G** when only **K** is estimated from the data (Candes and Recht 2009; Tian 2010).

The problem of how to estimate the missing entries of a matrix from the fewer entries that are known is called the matrix completion problem, and many of the potential solutions to this problem rest on the assumption that the completed matrix is of low rank (Candes and Recht 2009). Since we expect **G** matrices to be of rank $r < n$ as a consequence of pleiotropy (Johnson and Barton 2005) and current estimates

of **G** from multivariate mixed-model analyses are consistent with the vast majority of genetic variance being confined to a subspace of dimension $< n$ (Hine et al. 2009; Kirkpatrick 2009), such an assumption seems reasonable in a quantitative genetic context. As a consequence of the vast reduction in information in equation (3) compared to in equation (2), however, $r \ll n$ for the completed matrix and the completed matrix will necessarily be an approximation of the true covariance matrix in equation (2). One goal of this article is to determine whether a **G** of very high dimension completed using estimates for only a fraction of the elements contained in equation (3) can further our understanding of the distribution of genetic variance across the phenome.

*Completing **G** from a Set of Principal Submatrices*

Matrices of the form of equation (3) can potentially be completed using only the information provided by the principal submatrices because of the regular nature of the block sampling (Candes and Recht 2009). In particular, Bourin and Lee (2012) established that for $m = 2$ principal matrices $\mathbf{B}_{1,1}$ and $\mathbf{B}_{2,2}$, each of size k, the off-diagonal matrix $\mathbf{B}_{1,2}$ could be estimated under specific conditions if it is assumed that the completed matrix of size $n = 2k$ was of rank $r = k$. Expanding to $m > 2$ blocks, Bourin and Lee (2013, theorem 2.1, corollary 2.7) showed that a completed matrix of size n and rank k ($\mathbf{G}_n^k$) can be estimated from m principal submatrices as

$$\mathbf{G}_n^k = \begin{bmatrix} \mathbf{B}_{1,1}^{1/2} & 0 & \cdots & 0 \\ \mathbf{B}_{2,2}^{1/2} & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ \mathbf{B}_{m,m}^{1/2} & 0 & \cdots & 0 \end{bmatrix} \begin{bmatrix} \mathbf{B}_{1,1}^{1/2} & \mathbf{B}_{2,2}^{1/2} & \cdots & \mathbf{B}_{m,m}^{1/2} \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 \end{bmatrix}, \quad (4)$$

where 0 elements fill the two symmetrical matrices on the right-hand side of equation (4) to each be size n. Although the algebraic proof is given in Bourin and Lee (2013), here equation (4) is expressed in the notation used in Bourin and Lee (2012), which is more amenable to empirical usage.

The $\mathbf{B}_{m,m}^{1/2}$ in equation (4) are calculated as

$$\mathbf{B}_{m,m}^{1/2} = \mathbf{U}\Phi\mathbf{U}^{-1},$$

where **U** contains the eigenvectors of $\mathbf{B}_{m,m}$ as columns and Φ is a diagonal matrix containing the square root of the eigenvalues of $\mathbf{B}_{m,m}$. Equation (4) can be understood as substituting the geometric mean of any pair of principal submatrices for the associated off-diagonal submatrix of genetic covariances for that pair. Therefore, any (unknown) genetic covariance between two traits that is not estimated directly from the data within one of the principal submatrices is approximated from the (known) genetic covariances of those two traits, with the other $k - 1$ traits contained in their respective principal submatrices.

Note that the diagonal of the completed matrix $\mathbf{G}_n^k$ is the same as **K** and therefore the traces of $\mathbf{G}_n^k$ and **K** are the same. The sum of the k nonzero eigenvalues of $\mathbf{G}_n^k$ is therefore equal to the trace of **K**. This means that the magnitudes of the eigenvalues of $\mathbf{G}_n^k$ in isolation are of little use in interpreting how much genetic variance is captured by a particular phenome-wide trait combination. Rather, it is the comparison of the size of the eigenvalues from a null distribution, determined from a $\mathbf{G}_n^k$ where the covariance among traits is absent or due only to sampling covariance (see below), that enables a test for the presence of widespread genetic covariance among high-dimensional trait sets.

The one remaining complication that needs to be addressed before we can use equation (4) to complete $\mathbf{G}_n^k$ from **K** concerns the estimation of the large set of m principal submatrices. All principal submatrices need to be positive semidefinite, but empirical estimates of **G** tend to be negative definite as a consequence of sampling (Hill and Thompson 1978). Estimation methods that constrain the estimates of $\mathbf{B}_{m,m}$ to be positive semidefinite are therefore required. Two general approaches to guarantee positive semidefinite matrices are available, differing in their utility depending on the magnitude of k.

First, if k is chosen to be small enough, the m principal submatrices of **G** can each be estimated in separate standard multivariate mixed models. One readily available method to constrain a REML estimate of a **G** matrix to be positive semidefinite is by applying a factor-analytic covariance structure of full rank for the appropriate random effect within a mixed model (Meyer and Kirkpatrick 2005; Hine and Blows 2006). We apply this approach using $k = 5$ (such matrices are referred to as $\mathbf{G}_5$ below), where k was chosen as a compromise between the desire to include as many traits as possible and model convergence. Notably, of the published estimates of **G** surveyed by Pitchers et al. (2014), an average of five traits were included in any particular analysis.

A second approach that can be applied to matrices with larger values of k is to estimate **G** from a series of bivariate models and manually arrange the resulting covariances into a symmetrical matrix. This approach has been used to estimate single **G** where $n > 5$ (e.g., Mezey and Houle 2005; Leinonen et al. 2011), a practical solution to the estimation of large covariance matrices that is also used in other disciplines such as finance (Higham 2002). Once in symmetrical form, this $k \times k$ matrix can then be subjected to a bending or shrinkage procedure (Hayes and Hill 1981; Higham 2002; Meyer and Kirkpatrick 2010) to result in a positive semidefinite matrix. Here, we used matrices formed from bivariate mixed-model analyses of all possible pairwise combinations of 50 traits ($k = 50$, where such matrices are referred to as $\mathbf{G}_{50}$ below). We chose $k = 50$ as it was an order of magnitude larger than the $k = 5$ used for

the mixed-model approach and because we wished to establish the behavior of the Bourin and Lee (2012, 2013) approach when the number traits in each block exceeded the number of inbred lines (30) in the experiment. We applied the shrinkage estimator of Higham (2002) to ensure that each matrix was positive semidefinite, which first required the $\mathbf{G}_{50}$ covariance matrices to be transformed to correlation matrices. Higham's shrinkage estimator was applied using SAS IML code given in Wicklin (2013).

Null Distributions for the Eigenvalues of $\mathbf{G}_n^k$

From random matrix theory, it is known that the spectral distribution of a Hermitian matrix with random entries will follow the Marchenko-Pastur (MP) distribution (Bai and Silverstein 2010) and the leading eigenvalues will follow Tracy-Widom (TW) distributions (Tracy and Widom 2009). Therefore, we can expect that the leading eigenvalues of $\mathbf{G}_n^k$ will be inflated to some extent by this random process. As $\mathbf{G}$ is derived from a mixed linear model and is not a standard sample covariance matrix, it is not straightforward to establish the MP distribution in this case (see Martin 2014 for a discussion). Similarly, it was not possible to test the leading eigenvalues of $\mathbf{G}_n^k$ against the TW distribution, as the centering and scaling constants required to define the TW distribution for variance-component-based covariance matrices are unknown (Blows and McGuigan 2015).

We therefore took two alternative approaches to directly estimate the extent of the bias generated by random processes. First, we assumed no covariance among any of the randomly combined traits within a set, taking the estimated genetic variances for each individual trait (the diagonal) from the observed data and setting all off-diagonal elements (covariances) to zero. This null model has been used previously to explore the extent of bias imposed on phenotypic evolution by the genetic covariance among traits (Agrawal and Stinchcombe 2009) and has the useful property that eigenanalyses of these matrices will return eigenvalues corresponding to the genetic variance in individual expression traits in descending order of individual trait genetic variances.

Second, we replicated the entire set of mixed-model analyses that generated the sets of $\mathbf{G}_5$ and $\mathbf{G}_{50}$ matrices and subsequently the construction of $\mathbf{G}_n^k$ using a randomly generated data set with the observations for each gene expression trait represented as standard normal deviates and structured to have the same number of lines (30) and replicates per line (two) such that the same models could be fit. We then discarded the diagonals of these random $\mathbf{G}_5$ and $\mathbf{G}_{50}$ matrices and substituted the observed expression trait genetic variances (the diagonal) into these matrices. In this way, the total genetic variance remained the same in the null data and in the observed data, but the covariances differed, with the covariances in these null $\mathbf{G}$ matrices reflecting the sampling of random associations between traits. Substituting the observed genetic variances for the randomly generated variances does not change the fundamental nature of the null matrix, which will be negative definite using either diagonal as a consequence of sampling error. This is a more stringent null model, capturing the sampling error associated with estimation of covariances among traits, and allows the size of the eigenvalues of the null model to be directly compared to the observed data.

Gene Expression Data for Drosophila serrata *Males*

We measured gene expression traits in males from a set of 30 inbred lines derived from a collection of inseminated females from a natural population of *Drosophila serrata*, with 15 subsequent generations of brother-sister inbreeding of the offspring of those wild-caught females (Allen et al. 2013). Expression of 11,604 genes was measured using a custom-made NimbleGen microarray designed from a *D. serrata* expression library (Frentiu et al. 2009). Full details can be found in McGuigan et al. (2014*b*), and the data are available via the National Center for Biotechnology Information Gene Expression Omnibus accession GSE45801. Briefly, five probes, each of which was represented twice on each array, targeted each gene. Mixed-model analyses of the average $\log_{10}$ expression of the two replicates per probe were implemented per gene using a model where probe was fit as a fixed effect and among- and within-line variance were fit as random effects. Expression for each gene was standardized to a mean of 0 and standard deviation of 1 before analysis, negating the need for fitting gene as a fixed effect. Using log-likelihood ratio tests (with 1 degree of freedom, comparing a model with and without an among-line effect, and using the *P* values from those analyses to set a 5% false discovery rate (FDR) correction; Benjamini and Hochberg 2000; Storey and Tibshirani 2003), McGuigan et al. (2014*b*) determined that there was significant (at FDR) genetic (among-line) variance in 8,782 individual expression traits, with an average heritability of 0.41.

We randomly discarded two of the heritable expression traits, resulting in $n = 8{,}780$ traits (or $n = 8{,}750$ traits when $k = 50$ below), which were randomly assigned to $k = 5$ trait sets and subjected to multivariate mixed-model analyses to estimate the $m = 1{,}756$ $\mathbf{G}_5$ principal submatrices. The mixed model, fitted in SAS, version 9.3 (SAS Institute 2011), took the form

$$\mathbf{Y} = \boldsymbol{\mu} + \mathbf{X}\boldsymbol{b}_i + \mathbf{Z}_l\boldsymbol{\delta}_l + \mathbf{Z}_r\boldsymbol{\delta}_r + \boldsymbol{\varepsilon} \quad (5)$$

where $\mathbf{X}$ is a design matrix for the fixed effects of replicate probe per gene; $\boldsymbol{\varepsilon}$ is a diagonal matrix containing the residual (among probe mean) variances for each trait; $\mathbf{Z}_l$ and $\mathbf{Z}_r$

are design matrices for the line and replicate within-line random effects, respectively; and $\boldsymbol{\delta}_l$ and $\boldsymbol{\delta}_r$ are the covariance matrices for these effects. We fit two different types of covariance structures in these analyses. The within-line variance ($\boldsymbol{\delta}_r$) was modeled as an unstructured covariance matrix. To ensure positive semidefinite $\mathbf{B}_{m,m}$ was obtained, the among-line variance ($\boldsymbol{\delta}_l$) was modeled using a factor-analytic structure, $\boldsymbol{\delta}_l = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\mathrm{T}}$, where $\boldsymbol{\Lambda}$ is a lower triangular matrix of factor loadings and superscript T indicates transpose. We fit a full-rank factor-analytic model (see McGuigan and Blows 2010 for SAS code), with the number of dimensions equal to $k = 5$.

To estimate the $\mathbf{G}_{50}$ principal submatrices, we allocated all 50 traits contained in 10 of the $\mathbf{G}_5$ submatrices to one of 175 50-trait sets. The allocation of the traits contained in 10 complete sets of $\mathbf{G}_5$ matrices to a single $\mathbf{G}_{50}$ matrix allowed us to directly compare the estimated $\mathbf{G}_{50}$ to the $\mathbf{G}_{50}^5$ completed from the $\mathbf{G}_5$ principal submatrices using the Bourin and Lee (2012, 2013) approach. Specifically, we could empirically determine whether completing a matrix using only a fraction of the elements could successfully capture information on the genetic variance in the larger trait set. We therefore had to discard 30 traits (corresponding to six $\mathbf{G}_5$) from this analysis, resulting in 175 50-trait sets for estimation of $\mathbf{G}_{50}$ matrices. The 214,375 bivariate genetic covariances required to construct these 175 matrices were each estimated in a bivariate mixed model using model (5) but where the $\boldsymbol{\delta}_l$ were modeled with an unconstrained covariance structure and the matrices were subsequently transformed to be positive semidefinite, as described above.

Results

The Distribution of Genetic Variance in $\mathbf{G}_5$ *and* $\mathbf{G}_{50}$

We begin by describing the patterns of genetic variation in the $\mathbf{G}_5$ and $\mathbf{G}_{50}$ matrices (fig. 1), where each element has been estimated directly from the data. The $\mathbf{G}_5$ matrices, estimated in a multivariate mixed model using REML, are typical of those found in many quantitative genetic studies. Across the 1,756 $\mathbf{G}_5$ matrices, the average pattern of decay of the eigenvalues (the spectral distribution) was similar to that established for small sets of functionally related traits, illustrating the unevenness of the distribution of genetic variance across trait combinations (Kirkpatrick 2009; Walsh and Blows 2009). A high proportion of the genetic variation was associated with the first eigenvector, $\boldsymbol{g}_{\max}$ (median = 55.6%), with relatively little overlap of the eigenvalues of $\boldsymbol{g}_{\max}$ and $\boldsymbol{g}_2$ in the proportion of genetic variance explained (fig. 1*A*). The last eigenvectors ($\boldsymbol{g}_4$ and $\boldsymbol{g}_5$) accounted for very little genetic variance, typically for less than 10% of the total genetic variance in the trait set. As demonstrated by Hine et al. (2014) through the application of artificial directional selection, trait combinations with low levels of genetic variance are likely to represent a genetic nearly null subspace in which evolutionary responses to selection are inconsistent.

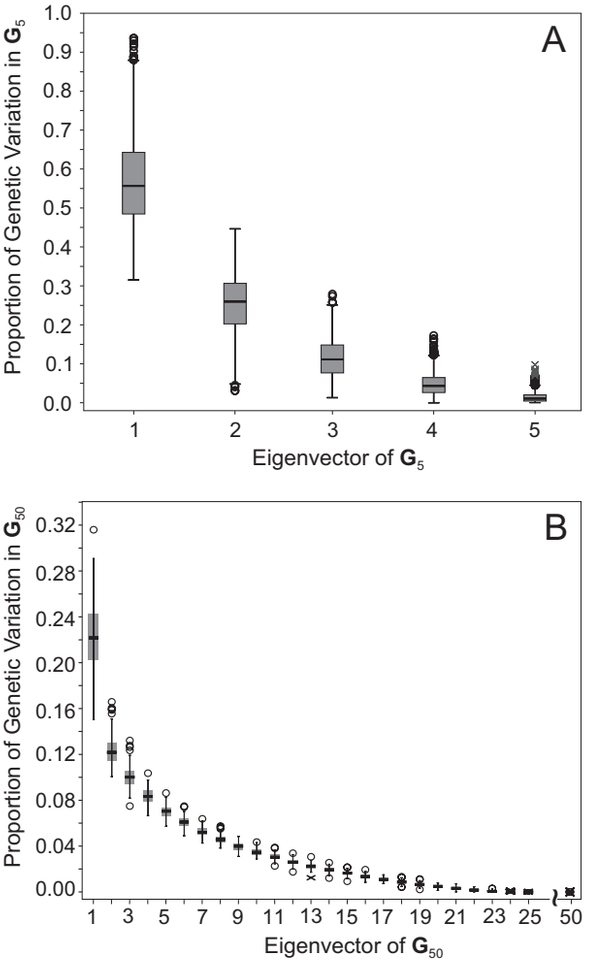

Figure 1: The distribution of eigenvalues (as a proportion of the total variance within the matrix) across 1,756 $\mathbf{G}_5$ principal submatrices (*A*) and 175 $\mathbf{G}_{50}$ principal submatrices (*B*). Boxes represent the range between the first and third quartile, with the median shown as a black band within the box. The whiskers represent the full range of the data, with values falling between 1.5 and 3 times the interquartile range (IQR) from the end of the box indicated by circles and extreme cases (>3 IQR from the end of the box) shown as crosses. In *B*, to improve the visibility of patterns among higher eigenvectors, eigenvectors $\boldsymbol{g}_{26}$–$\boldsymbol{g}_{49}$ are not plotted. Eigenvalues of these eigenvectors were not distinguishable from zero, as shown for $\boldsymbol{g}_{25}$ and $\boldsymbol{g}_{50}$.

The relatively high variance in $\boldsymbol{g}_{\max}$ in the $\mathbf{G}_5$ matrices may not only be a consequence of real genetic covariance but might also be contributed to by sampling variance in this trait combination. To address this issue, we compared the level of genetic variance in the observed $\boldsymbol{g}_{\max}$ with the

major axis of variation in the two null models of **G**, both of which have the same per-trait genetic variance (the diagonal) as the observed data, but in which we assumed no genetic covariance or we explicitly estimated the genetic covariances generated among traits through random sampling. The genetic variance of the observed $\boldsymbol{g}_{\max}$ was always greater than the genetic variance in any individual trait, as captured by the diagonal null model (fig. 2*A*). As expected, the random covariance null was a more stringent null than the diagonal null, but the observed $\boldsymbol{g}_{\max}$ eigenvalue was also typically greater than $\boldsymbol{g}_{\max}$ of the corresponding random covariance null model, with only 152 of the 1,756 $\mathbf{G}_5$ having a $\boldsymbol{g}_{\max}$ eigenvalue less than the null model (fig. 2*B*). A paired *t*-test indicated that significantly more genetic variance was captured in the eigenvalues of $\boldsymbol{g}_{\max}$ from the observed data than in the random covariance null model (excluding one extreme null model value, $t_{1754} = 15.789$, $P < .001$). The eigenvalue for the second eigenvector ($\boldsymbol{g}_2$) was significantly smaller in the observed data than in the null model ($t_{1755} = -6.642$, $P < .001$), which reflects the fact that total genetic variance is constrained to be the same in the observed and null data; when more genetic variance is contained in $\boldsymbol{g}_{\max}$, it is necessarily the case that less genetic variance is available to be allocated to the remaining eigenvalues.

Combining all traits from 10 of our five-trait sets, we used a pairwise mixed-modeling approach to estimate 175 $\mathbf{G}_{50}$ matrices (fig. 1*B*). This approach of building a larger matrix from pairwise analyses is frequently employed to deal with the computational difficulty of estimating relatively large covariance matrices. Such matrices will typically be negative definite, and we subjected our matrices to a shrinkage procedure to transform them to semipositive definite matrices. The $\mathbf{G}_{50}$ exhibited a similar spectral distribution to the $\mathbf{G}_5$, with the eigenvalue of $\boldsymbol{g}_{\max}$ much larger than and exhibiting relatively little overlap with the eigenvalues of subsequent eigenvectors (fig. 1*B*). Although the pattern of decay was similar, the proportion of genetic variation accounted for by $\boldsymbol{g}_{\max}$ was lower in the $\mathbf{G}_{50}$ than in the $\mathbf{G}_5$ matrices—a median of 22.1% compared to 55.6%. This difference in relative magnitude can at least in part be attributed to differences in the estimation of these matrices, particularly the transformation of $\mathbf{G}_{50}$ to a correlation matrix as part of the procedure for generating a semipositive definite matrix. Correlation matrices typically have a major eigenvector that explains a lower proportion of total variance than does the major eigenvector of a covariance matrix. When the same shrinkage approach was applied to the $\mathbf{G}_5$ matrices, the median proportion of genetic variance attributable to $\boldsymbol{g}_{\max}$ declined from 55.6% (fig. 1) to 33.6%.

Because of the limited number of inbred lines (30) in the experiment, eigenvalues for $\mathbf{G}_{50}$ eigenvectors after $\boldsymbol{g}_{29}$ are unlikely to contain any biological information. The eigenvalues of $\mathbf{G}_{50}$ were consistently estimated close to zero from $\boldsymbol{g}_{24}$ on (fig. 1*B*), consistent with the presence of a substantial nearly null subspace in these matrices. As with $\mathbf{G}_5$, comparison to the distribution of the random covariance null model **G** again provided strong support for widespread covariance among random sets of gene expression traits, with the eigenvalue of the observed $\boldsymbol{g}_{\max}$ significantly greater than in the null model ($t_{174} = 12.658$, $P < .001$).

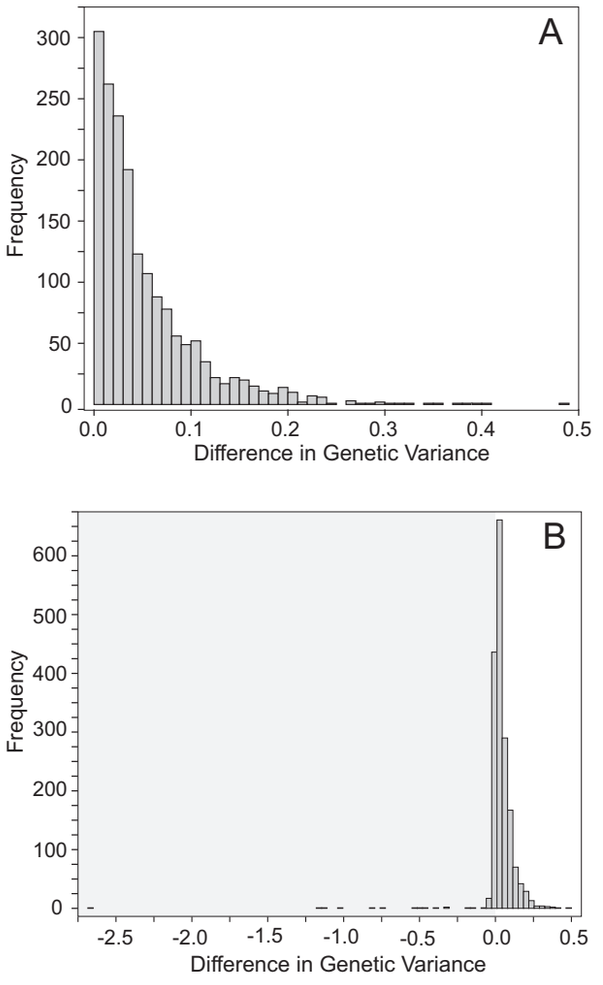

Figure 2: The difference between the eigenvalue of $\boldsymbol{g}_{\max}$ observed in the data and the $\boldsymbol{g}_{\max}$ eigenvalue of the diagonal null (*A*) and the $\boldsymbol{g}_{\max}$ eigenvalue of the random covariance null (*B*). The observed $\boldsymbol{g}_{\max}$ was always greater than $\boldsymbol{g}_{\max}$ of the diagonal null (only positive values in *A*). In *B*, the shaded background indicates values below zero (i.e., where the random covariance null $\boldsymbol{g}_{\max}$ was associated with greater variance than the observed $\boldsymbol{g}_{\max}$). In *B*, one extreme difference value (−26) occurred due to an extreme null model, and this matrix pair is not plotted and was not included in analyses comparing the observed and null distributions.

Prediction of $\mathbf{G}_{50}$ from the $\mathbf{G}_{50}^{5}$ Approximation

One goal of this article was to explore whether high-dimensional **G** matrices could be completed from only a very small fraction of all the unique elements of the full matrix to advance our understanding of the distribution of genetic variance. We first tested our approach using matrices of relatively small dimension ($n = 50$) that could be both estimated in their entirety from the data and completed from a small fraction of estimated elements. We compared 175 $\mathbf{G}_{50}$ matrices, which had all 1,275 elements estimated through pairwise covariance analyses, with their corresponding $\mathbf{G}_{50}^{5}$ matrices that we completed from the 150 elements directly estimated in the 10 $\mathbf{G}_{5}$ principal submatrices. Remembering that for the $\mathbf{G}_{50}^{5}$ matrices, completed from principal submatrices with $k = 5$, only the first $k = 5$ eigenvectors will have nonzero eigenvalues, it is therefore only this subspace of $\mathbf{G}_{50}^{5}$ that we can compare to $\mathbf{G}_{50}$.

To determine whether $\mathbf{G}_{50}$ and $\mathbf{G}_{50}^{5}$ captured the same information, we projected each of the first five eigenvectors of $\mathbf{G}_{50}^{5}$ through $\mathbf{G}_{50}$ (using $\mathbf{e}_i^{\mathrm{T}}\mathbf{G}_{50}\mathbf{e}_i$, where $\mathbf{e}_i$ is the ith eigenvector of $\mathbf{G}_{50}^{5}$ scaled to unit length and superscript T indicates matrix transpose) to determine how much of the genetic variance in $\mathbf{G}_{50}$ had been captured by $\mathbf{G}_{50}^{5}$. Given that $\mathbf{G}_{50}^{5}$ has a maximum of five dimensions, we compared the sum of the variances from the projection of the five $\mathbf{G}_{50}^{5}$ eigenvectors to the sum of the first five eigenvalues of $\mathbf{G}_{50}$, which represents the maximal level of genetic variance contained in five dimensions in $\mathbf{G}_{50}$. For the 175 replicate pairs of $\mathbf{G}_{50}$ and $\mathbf{G}_{50}^{5}$, on average, 40% of the genetic variance associated with the five-dimensional subspace of $\mathbf{G}_{50}$ was recovered by the five eigenvectors from $\mathbf{G}_{50}^{5}$. Notably, the ordered projection of the five $\mathbf{G}_{50}^{5}$ eigenvectors recovered genetic variance in the same decreasing order as the observed eigenvalues of the fully estimated $\mathbf{G}_{50}$ (fig. 3), indicating that the order of eigenvectors in $\mathbf{G}_{50}^{5}$ successfully predicted the order in $\mathbf{G}_{50}$. We again compared the observed genetic variance to the random covariance null model, projecting the five $\mathbf{G}_{50}^{5}$ eigenvectors through the null $\mathbf{G}_{50}$. For all five eigenvectors of $\mathbf{G}_{50}^{5}$, the genetic variance in the observed $\mathbf{G}_{50}$ was significantly greater than the genetic variance in the null model $\mathbf{G}_{50}$ (paired Wilcoxon tests with 174 degrees of freedom, $P < .002$; fig. 3). Therefore, estimating only the 150 elements contained in the 10 $\mathbf{G}_{5}$ principal submatrices rather than all 1,275 unique elements in $\mathbf{G}_{50}$ still resulted in the completed $\mathbf{G}_{50}^{5}$ matrices capturing a considerable portion of the shared genetic covariance among the 50 traits in each set.

*Completion of Extreme High-Dimensional **G** from Principal Submatrices*

The spectral distribution of the 1,756 $\mathbf{G}_{5}$ and 175 $\mathbf{G}_{50}$ exhibited the exponential decline that is typically observed for **G** matrices of morphological traits (Kirkpatrick 2009). Recent theory has established that all principal submatrices sampled from a larger Hermitian matrix will have almost the same spectral distribution if k is sufficiently large (Chatterjee and Ledoux 2009; Meckes and Meckes 2011). The consistency among the spectral distributions of the principal submatrices is therefore an expected consequence of the random sampling of such submatrices. Importantly, what is not apparent from the independent analyses of small random subsets of traits, many of which revealed genetic covariance among traits, is whether each submatrix is capturing a different variational module (sensu Wagner et al. 2007) or whether all submatrices are sampling the same large variational module. These different scenarios would reflect very different spectral distributions of larger-dimensional **G**. In general, there appears to be no established relationship between the spectral distribution of a larger covariance matrix and the spectral distributions of its principal submatrices (Chatterjee and Ledoux 2009). We can therefore not infer that the genetic covariance among a much larger set of traits will have a similar decay in the eigenvalues as observed for these smaller spaces (fig. 1). Therefore, to determine the properties of the distribution of genetic variance across a larger number of traits, it is necessary to estimate the spectral distribution of higher-dimensional **G** completed from different ran-

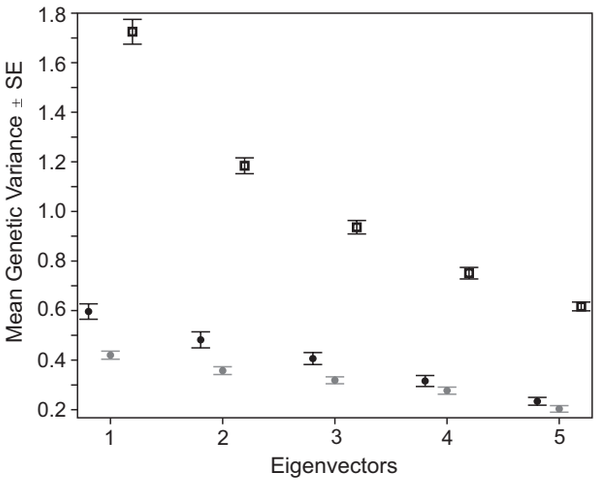

Figure 3: Distribution of genetic variance in $\mathbf{G}_{50}$ and $\mathbf{G}_{50}^{5}$. The genetic variance in each of the first five eigenvectors of $\mathbf{G}_{50}$ (open squares) is plotted on the original covariance scale (without applying the matrix shrinkage method to ensure a semipositive definite matrix). The variance $\mathbf{G}_{50}$ associated with each of the first five (and only positive) eigenvectors of $\mathbf{G}_{50}^{5}$ was determined by projecting each of these eigenvectors through $\mathbf{G}_{50}$ (black circles) and compared with the projection of these same eigenvectors through the random covariance null **G** estimated for 50 traits (gray circles).

dom samples of the data, and we implement this using both $\mathbf{G}_5$ and $\mathbf{G}_{50}$ principal submatrices.

We first divided the 8,750 traits into 10 independent sets, each comprised of 175 $\mathbf{G}_5$ principal submatrices, and then completed the 10 estimates of $\mathbf{G}^5_{875}$, repeating the same process for the random covariance null matrices. In general, the 10 replicates of these $\mathbf{G}^5_{875}$ matrices all behaved in a very similar way (fig. 4*A*). The eigenvalues for the five eigenvectors of $\mathbf{G}^5_{875}$, which are the only ones that can be nonzero (as $k = 5$), indicated that the completed matrices of the 875-dimensional **G** had successfully captured a single dimension that consistently had greater genetic variance than under the null model (fig. 4*A*). Because of the very limited number of dimensions ($k = 5$) that can have genetic variance, combined with the constraint that all genetic variance in the 875 traits is forced into these five dimensions, the remaining four eigenvectors had lower genetic variance than the null model and hence are of little value for understanding the biology of these traits.

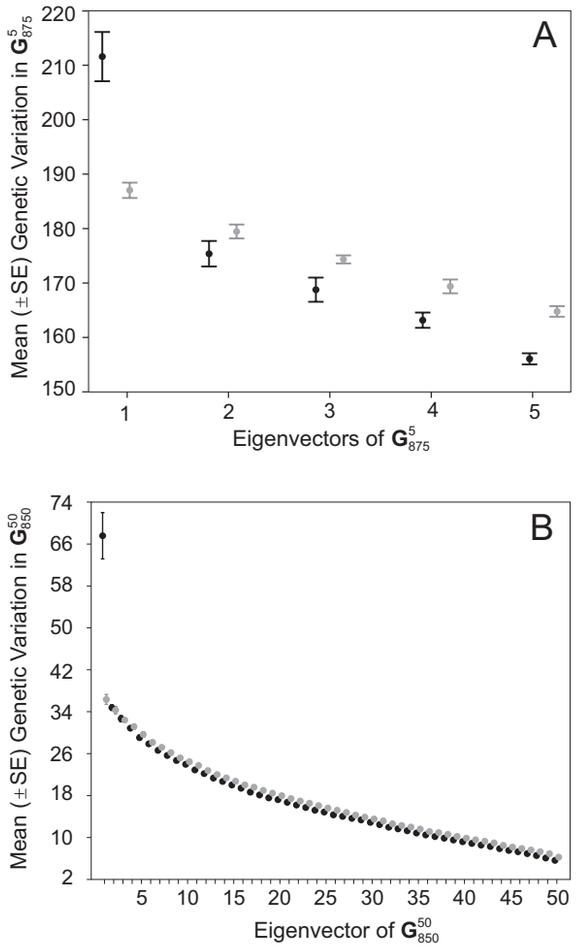

Figure 4: Mean (±SE) of the genetic variance partitioned to each eigenvector over 10 replicate random trait sets for $\mathbf{G}^5_{875}$ (*A*) and $\mathbf{G}^{50}_{850}$ (*B*) for the matrices estimated from the observed data (black circles) and from the null model of random trait covariances (gray circles).

We repeated this approach using the 10 independent sets of 850 traits, each comprising 17 $\mathbf{G}_{50}$ principal submatrices, which allowed many more dimensions ($k = 50$) to have genetic variance in the completed $\mathbf{G}^{50}_{850}$ matrices. Once again, the first dimension captured genetic variance that was well above the null model (fig. 4*B*; paired *t*-test: $t_9 = 12.977$, $P < .001$). The subsequent two dimensions also had mean eigenvalues above the null but overlapped with the null distribution substantially (fig. 4*B*). Finally, we used all 175 $\mathbf{G}_{50}$ principal submatrices to complete one single **G** for most of the gene expression traits (8,750 of 8,780) with significant individual heritability. This $\mathbf{G}^{50}_{8750}$ returned very similar results of a large $\boldsymbol{g}_{\max}$ eigenvalue but little evidence of subsequent dimensions. Here, we have only one observed and one null model matrix, but the leading eigenvector $\boldsymbol{g}_{\max}$ of $\mathbf{G}^{50}_{8750}$ had an eigenvalue of 640.9 (7.3% of the trace of all 8,750 traits), which was well in excess of the leading eigenvalue of 227.6 for the random covariance null estimate of $\mathbf{G}^{50}_{8750}$.

To this point, we have interpreted only the eigenvalues of **G** for evidence of genetic covariance among traits. However, the distribution of contributions of expression traits to the eigenvectors also provides insight into the among-trait relationships (McGuigan et al. 2014*b*). In the absence of covariance among traits, eigenvectors will have very large (approaching +1 or −1 for normalized vectors) contributions from a single trait and very low contributions (approaching 0) from the remaining traits. In the presence of positive covariance among traits, traits will load in the same direction (either positive or negative), whereas under negative covariance, traits will load in opposing directions. For any given eigenvector, direction is arbitrary such that all positively covarying traits might all have positive or all have negative loadings on the eigenvector; it is only the direction of contribution of traits relative to other traits within their set that is informative.

For the $\mathbf{G}_{50}$ principal submatrices, we determined whether there was any evidence of bias in the direction of covariance among traits (e.g., whether traits typically positively covary). Because of the arbitrary nature of direction within each of the 175 matrices, we took the sum of positive and negative loadings to infer bias. On average, $\boldsymbol{g}_{\max}$ of the 175 $\mathbf{G}_{50}$ principal submatrices had 12 more positive than negative loadings (fig. 5*A*). The random covariance null model also exhibited a bias toward more positive loadings (an average of three more positive loadings per principal submatrix) but to a lesser extent than in the observed data (fig. 5*A*; paired *t*-test: $t_{174} = 11.254$, $P < .001$).

In the $\mathbf{G}^{50}_{8750}$ matrix, because there was only one matrix and thus only one eigenvector, we can directly consider the trait

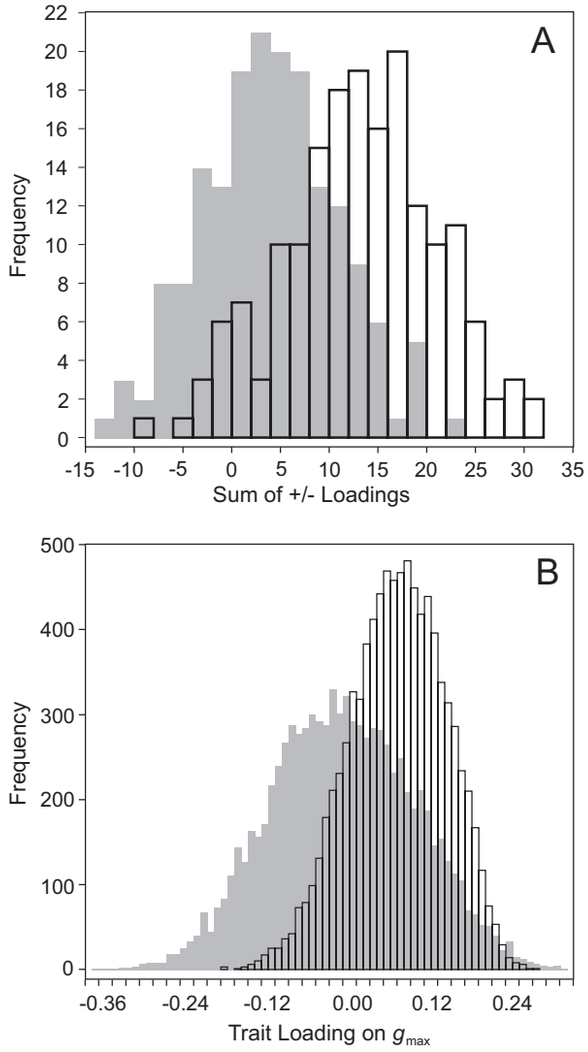

Figure 5: The distribution of expression trait contributions to $\boldsymbol{g}_{max}$. *A*, The distribution of the sum of the positive and negative contributions of individual traits to each of the 175 $\mathbf{G}_{50}$ of the observed and random covariance null (gray-shaded) data. For each matrix, we assigned all negative trait loadings a value of −1 and all positive loadings a value of +1 and summed these across all 50 traits. *B*, The distribution of individual trait contributions to the completed $\mathbf{G}^{50}_{8750}$ $\boldsymbol{g}_{max}$ for observed and null (gray-shaded) data.

loadings. The observed distribution of the eigenvector coefficients for $\boldsymbol{g}_{max}$ (i.e., contributions of each of the 8,750 individual gene expression traits to this statistically significant eigenvector) displayed a substantial skew in one direction (mean = 0.008048) compared to the null eigenvector (mean = 0.000035; fig. 5*B*), consistent with the suggestion from the analysis of the $\mathbf{G}_{50}$ that covariances were consistently directionally biased. This pattern of a shift in mean contribution away from zero provides further support for the interpretation of significant covariance among traits in this dimension of trait space. Furthermore, the directional bias in trait loadings indicates that this variational module is associated with coordinated regulation of expression in the same direction (up/down) of the many genes involved.

Gene Ontology Analysis

Having established the existence of widespread genetic covariance among the large number of expression traits that contribute to $\boldsymbol{g}_{max}$, more detailed questions concerning the mechanistic nature of pleiotropy among expression traits can be considered. To determine the underlying nature of the genetic covariance identified in $\boldsymbol{g}_{max}$, we performed gene ontology (Ashburner et al. 2000) term enrichment analysis using PANTHER (Mi et al. 2013), as implemented on the Gene Ontology Consortium website (http://geneontology.org/), on the 100 genes with the largest absolute loading on $\boldsymbol{g}_{max}$ (high group) and on the 100 genes with the lowest absolute loading on $\boldsymbol{g}_{max}$ (low group). Gene ontology terms were inferred for *Drosophila serrata* based on DNA sequence BLAST to *Drosophila melanogaster* genes under the assumption that hits between the two species are potential orthologs (Waterhouse et al. 2013). Standalone BLAST, version 2.2.27+ (Altschul et al. 1990), was used to perform tBLASTx (default settings) on the *D. melanogaster* chromosome, coding, gene, transcript, and pseudogene sequences obtained from Flybase (St. Pierre et al. 2014). *Drosophila serrata* genes with an e value less than 0.1 were considered potential orthologs and assigned the *D. melanogaster* GO terms. In the high group, 90 of the 100 genes were classed as potential *D. melanogaster* orthologs and 86 contained GO terms. For the low group, 90 of the original 100 genes were classified as potential orthologs; GO terms were available for 89 of these.

The high group of genes was significantly enriched for 31 GO terms, including a number of processes related to the regulation of gene expression (table 1). In contrast, the low group of genes was enriched for only two GO terms, and these were not directly related to transcriptional regulation. The construction of the major axis of genetic variance through the matrix completion approach was therefore successful in identifying a mechanistically related set of genes that share a functional relationship with transcriptional regulation, which in turn has the potential to affect a large number of genes, as was shown to be the case from the distribution of loadings in $\boldsymbol{g}_{max}$ (fig. 5*B*).

Discussion

The distribution of genetic variance among quantitative traits is fundamental to our understanding of how genetic variation is maintained and how traits will evolve in nat-

Table 1: Gene enrichment analysis of gene ontology (GO) terms for the high and low groups where the Bonferroni corrected *P* value was less than .05

Group	Type	Term	Background	Sample	Expected	*P* value
High	BP	Cellular metabolic process (GO:0044237)	3262	39	20.6000	.0004
High	BP	**Positive regulation of transcription initiation from RNA polymerase II promoter (GO:0060261)**	2	2	.0126	.0024
High	BP	**Positive regulation of DNA-templated transcription, initiation (GO:2000144)**	2	2	.0126	.0024
High	BP	Trachea morphogenesis (GO:0060439)	18	3	.1140	.0067
High	BP	**Negative regulation of mRNA splicing, via spliceosome (GO:0048025)**	4	2	.0253	.0096
High	BP	Gene expression (GO:0010467)	886	15	5.5900	.0128
High	BP	Mitotic spindle organization (GO:0007052)	214	7	1.3500	.0133
High	BP	Regulation of protein complex assembly (GO:0043254)	55	4	.3470	.0134
High	BP	Positive regulation of protein complex assembly (GO:0031334)	23	3	.1450	.0138
High	BP	**Regulation of transcription initiation from RNA polymerase II promoter (GO:0060260)**	5	2	.0316	.0150
High	BP	**Regulation of DNA-templated transcription, initiation (GO:2000142)**	5	2	.0316	.0150
High	BP	Trachea development (GO:0060438)	24	3	.1520	.0156
High	BP	**Negative regulation of mRNA metabolic process (GO:1903312)**	6	2	.0379	.0214
High	BP	**Negative regulation of mRNA processing (GO:0050686)**	6	2	.0379	.0214
High	BP	Spindle organization (GO:0007051)	244	7	1.5400	.0287
High	BP	**Negative regulation of RNA splicing (GO:0033119)**	7	2	.0442	.0291
High	MF	**TFIIF-class transcription factor binding (GO:0001096)**	2	2	.0126	.0010
High	MF	**RNA polymerase II basal transcription factor binding (GO:0001091)**	7	2	.0442	.0122
High	MF	**Basal RNA polymerase II transcription machinery binding (GO:0001099)**	8	2	.0505	.0159
High	MF	**Basal transcription machinery binding (GO:0001098)**	8	2	.0505	.0159
High	MF	**snRNA binding (GO:0017069)**	10	2	.0631	.0246
High	CC	**Transcription factor TFIIF complex (GO:0005674)**	3	2	.0189	.0033
High	CC	Macromolecular complex (GO:0032991)	2552	30	16.1000	.0056
High	CC	Cytosolic ribosome (GO:0022626)	94	5	.5930	.0065
High	CC	Endosome membrane (GO:0010008)	25	3	.1580	.0107
High	CC	Cytoplasmic part (GO:0044444)	1910	24	12.1000	.0110
High	CC	Ribonucleoprotein complex (GO:0030529)	468	10	2.9500	.0142
High	CC	Organelle part (GO:0044422)	2349	27	14.8000	.0183
High	CC	Endosomal part (GO:0044440)	36	3	.2270	.0305
High	CC	Cytosolic part (GO:0044445)	135	5	.8520	.0325
High	CC	Intracellular organelle part (GO:0044446)	2315	26	14.6000	.0332
Low	MF	Transferase activity (GO:0016740)	1276	19	8.3400	.0155
Low	CC	Cell surface (GO:0009986)	64	4	.4180	.0199

Note: Type refers to the GO term categories biological process (BP), molecular function (MF), and cellular component (CC). Background = the number of genes containing the GO term in *Drosophila melanogaster*. Sample = the number of genes with the GO term in our sample of 100 genes. Expected = the number of genes expected in our sample by chance. GO terms in boldface are those directly related to transcription.

ural populations. If all parts of the phenotypic space that comprise an organism have levels of genetic variation that are commonly found for individual traits, it is very difficult to reconcile the maintenance of such levels of genetic variation in the presence of selection (Johnson and Barton 2005; Zhang and Hill 2005). Taking a geometric view of the distribution of genetic variance and selection may potentially alleviate the severity of the conflict between the

simultaneous presence of strong selection and observed levels of genetic variance in natural populations (Walsh and Blows 2009). However, beyond the patterns found among small sets of functionally related traits, we have very little understanding of how many phenotypic dimensions of an organism might typically exhibit appreciable levels of genetic variance (Mezey and Houle 2005; Hine and Blows 2006; Kirkpatrick 2009; Hine et al. 2014).

There were three features of the distribution of genetic variance that our genetic analyses of 8,780 gene expression traits have highlighted. First, genetic covariance was common among expression traits that were randomly assigned to small trait sets without regard to biological function. Second, the matrix completion approximation of the high-dimensional **G** consistently uncovered genetic covariance among a very large number of expression traits, covariance that was well above random expectations. Finally, the single high-dimensional genetic factor identified in $\mathbf{G}^{50}_{8750}$ suggested a common up/down regulation pattern across a very large number of genes, an interpretation that was supported by the GO term enrichment for transcriptional regulation. We discuss each of these features of the distribution of genetic variance in turn below.

Genetic Covariance among Small Random Sets of Gene Expression Traits

There is a long tradition in evolutionary biology of highlighting the potential importance of pleiotropy among traits (Fisher 1930). The ubiquitous nature of correlated responses to selection between pairs of traits (Bohren et al. 1966) is perhaps the most widespread demonstration that even seemingly disparate traits can sometimes share a genetic basis. More recently, direct evidence for genetic covariance between functionally related multivariate sets of quantitative traits such as wing measures or cuticular hydrocarbons is well established (McGuigan and Blows 2007; Hine et al. 2009; Kirkpatrick 2009). However, given the likelihood of such traits sharing common developmental pathways, covariance among traits comprising related morphological structures or linked by shared chemistry is not particularly surprising. More broadly, life-history traits have also been found to be frequently genetically correlated (e.g., Houle 1991; Garant et al. 2008), an observation that can be explained by postulating that major fitness components will compete for the same finite resources (Lande 1982; Van Noordwijk and Dejong 1986; Riska 1989).

Our analysis of both the 1,756 random sets of five and the 175 random sets of 50 gene expression traits revealed that at least one dimension ($\boldsymbol{g}_{\max}$) was typically associated with greater variance than in any individual gene expression trait and was greater than expected through random associations among traits. This common pattern demonstrates that genetic covariance is widespread across the transcriptome. Assignment of traits was made completely randomly, without any a priori information on gene function, and we therefore only expect to detect genetic covariance if covariance among gene expression traits is prevalent, in contrast to analyses that focus on sets of traits known to be functionally or developmentally related. The genetic covariance detected here could be a consequence of either physical linkage of loci affecting different gene expression traits or pleiotropy (Lande 1980). Analyzing mutation accumulation lines, which differ from one another by relatively few mutations, McGuigan et al. (2014*b*) reported extensive covariance among five-trait sets of these same gene expression traits. Because of the few genetic differences among lines and the limited opportunity for selection, this covariance is more likely to be caused by pleiotropy than by linkage. Probability analyses of the mutational pleiotropy suggested that variational modules (sensu Wagner et al. 2007) spanned at least 70 traits on average, although there was likely to be a considerable range of module size around this figure that is yet to be quantified (McGuigan et al. 2014*b*). Therefore, the mutational pleiotropy among expression traits is likely to underlie much of the widespread standing genetic covariance among random sets of expression traits that we observed in these lines derived from a natural population.

The Distribution of Genetic Variance in High Dimensions

As highlighted in several recent publications (Houle 2010; Houle et al. 2010), if we are to extend our understanding of the development and evolution of phenotypes, we need to resolve the challenge of both quantifying high-dimensional phenotypes and analyzing such data. Here, we have addressed the second issue, extending our understanding of the phenome-wide distribution of genetic variance by estimating very large covariance matrices under a quantitative genetic framework. The completion of **G** from a small fraction of the elements, contained in small principal submatrices, was successful in capturing a substantial proportion of the genetic variance. Specifically, completing matrices of 50 dimensions using principal submatrices of five dimensions recovered 40% of the maximum genetic variance associated with five dimensions in the fully estimated **G**. Furthermore, the approximation represented by the completed matrix was able to distinguish between a large number of dimensions that had little or no genetic variance and the 10 or so dimensions that contained the vast majority of the genetic variance in these 50-dimensional spaces (fig. 1*B*).

For much higher dimensions, such as that recovered in the $\mathbf{G}^{50}_{850}$ and $\mathbf{G}^{5}_{875}$ matrices, it was not feasible to determine exactly what part of the subspace was recovered since such large matrices are not readily estimable from the data in

the way that the $\mathbf{G}_{50}$ matrices were. Instead, the use of the null distribution allowed us to demonstrate that in these extreme high-dimensional cases, at least one dimension recovered a substantial amount of genetic variance, well in excess of that expected through random sampling of traits. Subtracting the level of genetic variance observed in the random covariance null $\boldsymbol{g}_{\max}$ vector, the observed $\boldsymbol{g}_{\max}$ vectors from the 10 replicate $\mathbf{G}^{50}_{850}$ and $\mathbf{G}^{5}_{875}$ matrices explained between 3% and 4% of the genetic variance, respectively. Given the nature of completed matrices established from the comparison between $\mathbf{G}^{5}_{50}$ and $\mathbf{G}_{50}$, we expect that the leading eigenvector of the real 850- and 875-trait matrices accounts for a much larger proportion of the total genetic variance in the trait set. Nonetheless, the similarity of order of decay of the real $\mathbf{G}_{50}$ eigenvalues and the genetic variance in the projected $\mathbf{G}^{5}_{50}$ eigenvectors suggest that the major axis that is recovered from the matrix completion approach will capture biological information about the shared genetic basis of a very large number of traits.

Comparing the results for the completed $\mathbf{G}^{5}_{875}$ and $\mathbf{G}^{50}_{850}$ matrices, increasing the number of elements directly estimated via the principal submatrices from 0.7% to 6.0% increased the level of genetic variance recovered by only 1% (3% vs. 4%), which suggests little benefit from the increased computational effort. However, with the increase in the number of elements estimated from the data in the $\mathbf{G}^{50}_{850}$ matrices, there was the suggestion that at least a further two dimensions of $\mathbf{G}^{50}_{850}$ might have been associated with genetic variance, and these dimensions might have been recovered and available for interpretation if the experiment had more power. The relative merits of using smaller principal submatrices each estimated within a single multivariate mixed model to be positive semidefinite, compared to using much larger negative definite principal submatrices manually constructed from an enormous number of bivariate estimates of covariance, deserves further consideration in future work.

Finally, the distribution of eigenvalues uncovered by the extreme high-dimension $\mathbf{G}^{50}_{850}$ and $\mathbf{G}^{50}_{8750}$ completed matrices displayed spectral distributions consistent with the behavior of spiked covariance models of high-dimensional covariance matrices (Paul and Aue 2014). The term spiked refers to a small number of dimensions that have large eigenvalues, while the vast majority of dimensions have eigenvalues equal in value to some arbitrary small number. Under spiked covariance models, the eigenvalues are subject to a phase transition behavior in relation to their detectability from the null Marcenko-Pasteur distribution. Given the ratio between the number of dimensions and the number of observations in a data set, eigenvalues below a certain magnitude are very unlikely to be distinguishable from the null model. Such behavior of eigenvalues has been shown in a population genetics context for the detection of genetic structure among populations from a large number of markers (Patterson et al. 2006). In our case, we suspect that the detection of a single dimension of genetic variance in our high-dimensional completed matrices is likely to be influenced by a similar phase transition. The presence of a single significant dimension should therefore not be interpreted as evidence for a sole underlying variation module for gene expression but is rather likely to be a consequence of the structure of the data set we have used. Whether this is a common characteristic of high-dimensional **G** estimated from other experimental designs and trait types is a question for further exploration.

The Extent of Genetic Covariance among Gene Expression Traits

While there are numerous examples of the pleiotropic effects of single genes, the extent of pleiotropy is a contentious issue (Wagner and Zhang 2011; Hill and Zhang 2012; Paaby and Rockman 2013). Gene knockout studies suggest that only a modest number of traits may be affected by each gene (Wang et al. 2010; Wagner and Zhang 2011), while the frequency of mutational covariance among traits indicates that pleiotropy of naturally occurring mutations might be widespread (McGuigan et al. 2014*b*). The frequency of covariance among small sets of random expression traits found here suggests that there is likely to be some underlying factor(s) that might affect many such traits at once. We have shown, using the high-dimensional matrix completion approach of Bourin and Lee (2013), that the high frequency of covariance among random sets of expression traits is underpinned by at least one genetic factor that affects a very large number of these traits. Several previous studies of gene expression profiles, taking various approaches such as coexpression (Denver et al. 2005) or eQTL (West et al. 2007) mapping and the Bayesian sparse factor approach of Runcie and Mukherjee (2013), have similarly revealed that the expression of very many genes might covary.

Consideration of the eigenvector loadings suggested that genetic covariances were strongly biased in one direction. At the genomic level, a pattern where a substantial number of expression traits positively covary would suggest the involvement of genes with regulatory function. Our GO term analysis on the 100 genes with the strongest association with $\boldsymbol{g}_{\max}$ of $\mathbf{G}^{50}_{8750}$ illuminated an appreciable number of terms linked to regulatory functions, including mRNA processing (Le Hir et al. 2003), transcription initiation (Shilatifard et al. 2003), and transcription factors that underlie gene regulatory networks (Erwin and Davidson 2009). A recent study constructing transcription factor protein interaction networks in *Drosophila* has indicated that many hundreds of transcription factor proteins display bivariate interactions, with up to 63% of 647 known or putative transcription factors forming a single protein interaction net-

work (Rhee et al. 2014). The genetic variance captured by $\boldsymbol{g}_{\max}$ represents the possible genetic control of such widespread protein networks.

As in our study, where the major axis of genetic variation in gene expression of an outbred population of *Drosophila serrata* was for up/down regulation of many genes, Runcie and Mukherjee (2013) in their genetic analysis of *Drosophila melanogaster* gene expression also identified a major axis of genetic variation whereby a substantial proportion of traits changed expression in the same direction (their factor 2). Importantly, Runcie and Mukherjee (2013) found that this was one of only two factors that were genetically correlated with a measure of competitive fitness. Studies in a variety of taxa have suggested a general pattern across many genes of stabilizing selection on expression levels (Denver et al. 2005; Rifkin et al. 2005; Bedford and Hartl 2009; Warnefors and Eyre-Walker 2012), and we have recently demonstrated that stabilizing selection is intensified in the presence of pleiotropic effects across traits (McGuigan et al. 2014*a*). This emergent picture of general transcriptome-wide mechanisms of gene regulation across many traits also provides a potential basis for the ubiquitous presence of correlated responses in both evolutionary experiments and applied animal and plant breeding.

Conclusion

Establishing the phenome-wide distribution of genetic variance is a key component of attempting to understand the maintenance of genetic variance from an empirical perspective, and the approach we have developed here is just a first step in this regard. If mutation-selection balance is a leading cause of the maintenance of genetic variance, then it will be important to determine why some trait combinations display so much genetic variance when others do not. Is strong selection responsible for the depletion of genetic variance in the nearly null space or are these phenotypic combinations simply not prone to the effects of new mutations? On the one hand, two recent mutation-accumulation studies showed that mutational pleiotropy among functionally related traits (Houle and Fierst 2013) and random sets of gene expression traits (McGuigan et al. 2014*b*) is common but that some trait combinations exhibit more mutational variance than others, at least within the relatively short time frames of such experiments. On the other hand, stronger stabilizing selection against mutations with greater pleiotropic effects has been demonstrated for random sets of gene expression traits (McGuigan et al. 2014*a*). Disentangling the relative contribution of low mutation variance and strong selection in generating nearly null subspaces will require the simultaneous assessment of phenome-wide patterns of mutational variance and selection.

Acknowledgments

We thank E. Hine for discussions and comments on a previous draft. This work was funded by the Australian Research Council.

Literature Cited

Agrawal, A. F., and J. R. Stinchcombe. 2009. How much do genetic covariances alter the rate of adaptation? Proceedings of the Royal Society B: Biological Sciences 276:1183–1191.

Allen, S. L., R. Bonduriansky, and S. F. Chenoweth. 2013. The genomic distribution of sex-biased genes in *Drosophila serrata*: x-chromosome demasculinization, feminization, and hyperexpression in both sexes. Genome Biology and Evolution 5:1986–1994.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. Journal of Molecular Biology 215:403–410.

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, et al. 2000. Gene ontology: tool for the unification of biology. Nature Genetics 25:25–29.

Bai, Z., and J. W. Silverstein. 2010. Spectral analysis of large dimensional random matrices. Springer Series in Statistics. Springer, New York.

Bedford, T., and D. L. Hartl. 2009. Optimization of gene expression by natural selection. Proceedings of the National Academy of Sciences of the USA 106:1133–1138.

Benjamini, Y., and Y. Hochberg. 2000. On the adaptive control of the false discovery fate in multiple testing with independent statistics. Journal of Educational and Behavioral Statistics 25:60–83.

Blows, M. W., S. F. Chenoweth, and E. Hine. 2004. Orientation of the genetic variance-covariance matrix and the fitness surface for multiple male sexually selected traits. American Naturalist 163:E329–E340.

Blows, M. W., and A. A. Hoffmann. 2005. A reassessment of genetic limits to evolutionary change. Ecology 86:1371–1384.

Blows, M. W., and K. McGuigan. 2015. The distribution of genetic variance across phenotypic space and the response to selection. Molecular Ecology 24:2056–2072.

Bohren, B. B., W. G. Hill, and A. Robertson. 1966. Some observations on asymmetrical correlated responses to selection. Genetical Research 7:44–57.

Bourin, J. C., and E. Y. Lee. 2012. Unitary orbits of Hermitian operators with convex or concave functions. Bulletin of the London Mathematical Society 44:1085–1102.

———. 2013. Decomposition and partial trace of positive matrices with Hermitian blocks. International Journal of Mathematics 24: 1350010.

Britten, R. J., and E. H. Davidson. 1971. Repetitive and non-repetitive DNA sequences and a speculation on origins of evolutionary novelty. Quarterly Review of Biology 46:111–138.

Candes, E. J., and B. Recht. 2009. Exact matrix completion via convex optimization. Foundations of Computational Mathematics 9:717–772.

Carroll, S. B. 2008. Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. Cell 134:25–36.

Chatterjee, S., and M. Ledoux. 2009. An observation about submatrices. Electronic Communications in Probability 14:495–500.

Chenoweth, S. F., H. D. Rundle, and M. W. Blows. 2010. The contribution of selection and genetic constraints to phenotypic divergence. American Naturalist 175:186–196.

Cheverud, J. M. 1984. Quantitative genetics and developmental constraints on evolution by selection. Journal of Theoretical Biology 10:155–171.

Denver, D. R., K. Morris, J. T. Streelman, S. K. Kim, M. Lynch, and W. K. Thomas. 2005. The transcriptional consequences of mutation and natural selection in *Caenorhabditis elegans*. Nature Genetics 37:544–548.

Erwin, D. H., and E. H. Davidson. 2009. The evolution of hierarchical gene regulatory networks. Nature Reviews Genetics 10:141–148.

Fisher, R. A. 1930. The genetical theory of natural selection. Clarendon, Oxford.

Frentiu, F. D., M. Adamski, E. A. McGraw, M. W. Blows, and S. F. Chenoweth. 2009. An expressed sequence tag (EST) library for *Drosophila serrata*, a model system for sexual selection and climatic adaptation studies. BMC Genomics 10:40.

Garant, D., J. D. Hadfield, L. E. B. Kruuk, and B. C. Sheldon. 2008. Stability of genetic variance and covariance for reproductive characters in the face of climate change in a wild bird population. Molecular Ecology 17:179–188.

Gomulkiewicz, R., and D. Houle. 2009. Demographic and genetic constraints on evolution. American Naturalist 174:E218–E229.

Hansen, T. F., and D. Houle. 2008. Measuring and comparing evolvability and constraint in multivariate characters. Journal of Evolutionary Biology 21:1201–1219.

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

Higham, N. J. 2002. Computing the nearest correlation matrix: a problem from finance. IMA Journal of Numerical Analysis 22:329–343.

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

Hill, W. G., and X. S. Zhang. 2012. Assessing pleiotropy and its evolutionary consequences: pleiotropy is not necessarily limited, nor need it hinder the evolution of complexity. Nature Reviews Genetics 13:295–295.

Hine, E., and M. W. Blows. 2006. Determining the effective dimensionality of the genetic variance-covariance matrix. Genetics 173:1135–1144.

Hine, E., S. F. Chenoweth, and M. W. Blows. 2004. Multivariate quantitative genetics and the lek paradox: genetic variance in male sexually selected traits of *Drosophila serrata* under field conditions. Evolution 58:2754–2762.

Hine, E., S. F. Chenoweth, H. D. Rundle, and M. W. Blows. 2009. Characterizing the evolution of genetic variance using genetic covariance tensors. Philosophical Transactions of the Royal Society B: Biological Sciences 364:1567–1578.

Hine, E., K. McGuigan, and M. W. Blows. 2011. Natural selection stops the evolution of male attractiveness. Proceedings of the National Academy of Sciences of the USA 108:3659–3664.

———. 2014. Evolutionary constraints in high-dimensional trait sets. American Naturalist 184:119–131.

Houle, D. 1991. Genetic covariance of fitness correlates: what genetic correlations are made of and why it matters. Evolution 45:630–648.

———. 2010. Numbering the hairs on our heads: the shared challenge and promise of phenomics. Proceedings of the National Academy of Sciences of the USA 107:1793–1799.

Houle, D., and J. Fierst. 2013. Properties of spontaneous mutational variance and covariance for wing size and shape in *Drosophila melanogaster*. Evolution 67:1116–1130.

Houle, D., D. R. Govindaraju, and S. Omholt. 2010. Phenomics: the next challenge. Nature Reviews Genetics 11:855–866.

Innocenti, P., and S. F. Chenoweth. 2013. Interspecific divergence of transcription networks along lines of genetic variance in *Drosophila*: dimensionality, evolvability, and constraint. Molecular Biology and Evolution 30:1358–1367.

Johnson, T., and N. Barton. 2005. Theoretical models of selection and mutation on quantitative traits. Philosophical Transactions of the Royal Society B: Biological Sciences 360:1411–1425.

King, M. C., and A. C. Wilson. 1975. Evolution at two levels in humans and chimpanzees. Science 188:107–116.

Kirkpatrick, M. 2009. Patterns of quantitative genetic variation in multiple dimensions. Genetica 136:271–284.

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

———. 1980. The genetic covariances between characters maintained by pleiotropic mutations. Genetics 94:203–215.

———. 1982. A quantitative genetic theory of life history evolution. Ecology 63:607–615.

Le Hir, H., A. Nott, and M. J. Moore. 2003. How introns influence and enhance eukaryotic gene expression. Trends in Biochemical Sciences 28:215–220.

Leinonen, T., J. M. Cano, and J. Merila. 2011. Genetics of body shape and armour variation in threespine sticklebacks. Journal of Evolutionary Biology 24:206–218.

Mahoney, M. W., and P. Drineas. 2009. CUR matrix decompositions for improved data analysis. Proceedings of the National Academy of Sciences of the USA 106:697–702.

Martin, G. 2014. Fisher's geometrical model emerges as a property of complex integrated phenotypic networks. Genetics 197:237–255.

McGraw, E. A., Y. H. Ye, B. Foley, S. F. Chenoweth, M. Higgie, E. Hine, and M. W. Blows. 2011. High dimensional variance partitioning reveals the modular genetic basis of adaptive divergence in gene expression during reproductive character displacement. Evolution 65:3126–3137.

McGuigan, K., and M. W. Blows. 2007. The phenotypic and genetic covariance structure of Drosophilid wings. Evolution 61:902–911.

———. 2010. Evolvability of individual traits in a multivariate context: partitioning the additive genetic variance into common and specific components. Evolution 64:1899–1911.

McGuigan, K., J. M. Collet, S. L. Allen, S. F. Chenoweth, and M. W. Blows. 2014*a*. Pleiotropic mutations are subject to strong stabilizing selection. Genetics 197:1051–1062.

McGuigan, K., J. M. Collet, E. A. McGraw, Y. X. H. Ye, S. L. Allen, S. F. Chenoweth, and M. W. Blows. 2014*b*. The nature and extent of mutational pleiotropy in gene expression of male *Drosophila serrata*. Genetics 196:911–921.

Meckes, E. S., and M. W. Meckes. 2011. Another observation about operator compressions. Proceedings of the American Mathematical Society 139:1433–1439.

Meyer, K., and M. Kirkpatrick. 2005. Restricted maximum likelihood estimation of genetic principal components and smoothed covariance matrices. Genetics Selection Evolution 37:1–30.

———. 2010. Better estimates of genetic covariance matrices by "bending" using penalized maximum likelihood. Genetics 185:1097–1110.

Mezey, J. G., and D. Houle. 2005. The dimensionality of genetic variation for wing shape in *Drosophila melanogaster*. Evolution 59:1027–1038.

Mi, H. Y., A. Muruganujan, and P. D. Thomas. 2013. PANTHER in 2013: modeling the evolution of gene function, and other gene at-

tributes, in the context of phylogenetic trees. Nucleic Acids Research 41:D377–D386.
Paaby, A. B., and M. V. Rockman. 2013. The many faces of pleiotropy. Trends in Genetics 29:66–73.
Patterson, N., A. L. Price, and D. Reich. 2006. Population structure and eigenanalysis. PLoS Genetics 2:e190.
Paul, D., and A. Aue. 2014. Random matrix theory in statistics: a review. Journal of Statistical Planning and Inference 150:1–29.
Pespeni, M. H., B. T. Barney, and S. R. Palumbi. 2013. Differences in the regulation of growth and biomineralization genes revealed through long-term common-garden acclimation and experimental genomics in the purple sea urchin. Evolution 67:1901–1914.
Phillips, P. C., M. C. Whitlock, and K. Fowler. 2001. Inbreeding changes the shape of the genetic covariance matrix in *Drosophila melanogaster*. Genetics 158:1137–1145.
Pitchers, W., J. B. Wolf, T. Tregenza, J. Hunt, and I. Dworkin. 2014. Evolutionary rates for multivariate traits: the role of selection and genetic variation. Philosophical Transactions of the Royal Society B: Biological Sciences 369:20130252.
Rest, J. S., C. M. Morales, J. B. Waldron, D. A. Opulente, J. Fisher, S. Moon, K. Bullaughey, et al. 2013. Nonlinear fitness consequences of variation in expression level of a eukaryotic gene. Molecular Biology and Evolution 30:448–456.
Rhee, D. Y., D.-Y. Cho, B. Zhai, M. Slattery, L. Ma, J. Mintseris, C. Y. Wong, et al. 2014. Transcription factor networks in *Drosophila melanogaster*. Cell Reports 8:2031–2043.
Rifkin, S. A., D. Houle, J. Kim, and K. P. White. 2005. A mutation accumulation assay reveals a broad capacity for rapid evolution of gene expression. Nature 438:220–223.
Riska, B. 1989. Composite traits, selection response, and evolution. Evolution 43:1172–1191.
Runcie, D. E., and S. Mukherjee. 2013. Dissecting high-dimensional phenotypes with Bayesian sparse factor analysis of genetic covariance matrices. Genetics 194:753–767.
SAS Institute. 2011. SAS. Version 9.3. SAS Institute, Cary, NC.
Schluter, D. 1996. Adaptive radiation along genetic lines of least resistance. Evolution 50:1766–1774.
Shilatifard, A., R. C. Conaway, and J. W. Conaway. 2003. The RNA polymerase II elongation complex. Annual Review of Biochemistry 72:693–715.
Stone, E. A., and J. F. Ayroles. 2009. Modulated modularity clustering as an exploratory tool for functional genomic inference. PLoS Genetics 5:e1000479.
Storey, J. D., and R. Tibshirani. 2003. Statistical significance for genomewide studies. Proceedings of the National Academy of Sciences of the USA 100:9440–9445.
St. Pierre, S. E., L. Ponting, R. Stefancsik, P. McQuilton, and FlyBase Consortium. 2014. FlyBase 102—advanced approaches to interrogating FlyBase. Nucleic Acids Research 42:D780–D788.
Tian, Y. G. 2010. Completing block Hermitian matrices with maximal and minimal ranks and inertias. Electronic Journal of Linear Algebra 21:124–141.
Tracy, C. A., and H. Widom. 2009. The distributions of random matrix theory and their applications. Pages 753–765 *in* V. Sidoravicius, ed. New trends in mathematical physics. Springer, Dordrecht.
Van Homrigh, A., M. Higgie, K. McGuigan, and M. W. Blows. 2007. The depletion of genetic variance by sexual selection. Current Biology 17:528–532.
Van Noordwijk, A. J., and G. Dejong. 1986. Acquisition and allocation of resources: their influence on variation in life-history tactics. American Naturalist 128:137–142.
Wagner, G. P., M. Pavlicev, and J. M. Cheverud. 2007. The road to modularity. Nature Reviews Genetics 8:921–931.
Wagner, G. P., and J. Z. Zhang. 2011. The pleiotropic structure of the genotype-phenotype map: the evolvability of complex organisms. Nature Reviews Genetics 12:204–213.
Walsh, B., and M. W. Blows. 2009. Abundant genetic variation + strong selection = multivariate genetic constraints: a geometric view of adaptation. Annual Review of Ecology, Evolution, and Systematics 40:41–59.
Wang, Z., B. Y. Liao, and J. Z. Zhang. 2010. Genomic patterns of pleiotropy and the evolution of complexity. Proceedings of the National Academy of Sciences of the USA 107:18034–18039.
Warnefors, M., and A. Eyre-Walker. 2012. A selection index for gene expression evolution and its application to the divergence between humans and chimpanzees. PLoS ONE 7:e34935.
Waterhouse, R. M., F. Tegenfeldt, J. Li, E. M. Zdobnov, and E. V. Kriventseva. 2013. OrthoDB: a hierarchical catalog of animal, fungal and bacterial orthologs. Nucleic Acids Research 41:D358–D365.
West, M. A. L., K. Kim, D. J. Kliebenstein, H. van Leeuwen, R. W. Michelmore, R. W. Doerge, and D. A. S. Clair. 2007. Global eQTL mapping reveals the complex genetic architecture of transcript-level variation in *Arabidopsis*. Genetics 175:1441–1450.
Whitehead, A., J. L. Roach, S. J. Zhang, and F. Galvez. 2011. Genomic mechanisms of evolved physiological plasticity in killifish distributed along an environmental salinity gradient. Proceedings of the National Academy of Sciences of the USA 108:6193–6198.
Wicklin, R. 2013. Simulating data with SAS. SAS Institute, Cary, NC.
Wittkopp, P. J., and G. Kalay. 2012. Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. Nature Reviews Genetics 13:59–69.
Zhang, X. S., and W. G. Hill. 2005. Genetic variability under mutation selection balance. Trends in Ecology and Evolution 20:468–470.

Associate Editor: Christina M. Caruso
Editor: Troy Day