# Multivariate Analysis of Genotype–Phenotype Association

**Philipp Mitteroecker,*[,1] James M. Cheverud,† and Mihaela Pavlicev‡**

*Department of Theoretical Biology, University of Vienna, A-1090 Vienna, Austria, †Department of Biology, Loyola University of Chicago, Chicago, Illinois 60660, and ‡Department of Pediatrics, Cincinnati Children's Hospital Medical Centre, Cincinnati, Ohio 45229

**ABSTRACT** With the advent of modern imaging and measurement technology, complex phenotypes are increasingly represented by large numbers of measurements, which may not bear biological meaning one by one. For such multivariate phenotypes, studying the pairwise associations between all measurements and all alleles is highly inefficient and prevents insight into the genetic pattern underlying the observed phenotypes. We present a new method for identifying patterns of allelic variation (genetic latent variables) that are maximally associated—in terms of effect size—with patterns of phenotypic variation (phenotypic latent variables). This multivariate genotype–phenotype mapping (MGP) separates phenotypic features under strong genetic control from less genetically determined features and thus permits an analysis of the multivariate structure of genotype–phenotype association, including its dimensionality and the clustering of genetic and phenotypic variables within this association. Different variants of MGP maximize different measures of genotype–phenotype association: genetic effect, genetic variance, or heritability. In an application to a mouse sample, scored for 353 SNPs and 11 phenotypic traits, the first dimension of genetic and phenotypic latent variables accounted for $>70\%$ of genetic variation present in all 11 measurements; 43% of variation in this phenotypic pattern was explained by the corresponding genetic latent variable. The first three dimensions together sufficed to account for almost 90% of genetic variation in the measurements and for all the interpretable genotype–phenotype association. Each dimension can be tested as a whole against the hypothesis of no association, thereby reducing the number of statistical tests from 7766 to 3—the maximal number of meaningful independent tests. Important alleles can be selected based on their effect size (additive or nonadditive effect on the phenotypic latent variable). This low dimensionality of the genotype–phenotype map has important consequences for gene identification and may shed light on the evolvability of organisms.

**KEYWORDS** genetic mapping; genotype–phenotype map; mouse; multivariate analysis; partial least squares

STUDIES of genotype–phenotype association are central to several branches of contemporary biology and biomedicine, but they suffer from serious conceptual and statistical problems. Most of these studies consist of a vast number of pairwise comparisons between single genetic loci and single phenotypic variables, typically leading—among other reasons—to very low fractions of phenotypic variance explained by genetic effects ["missing heritability" (Manolio *et al.* 2009; Eichler *et al.* 2010)]. *Post hoc* corrections for multiple testing can lead to a dramatic loss of statistical power and in fact violate standard rules of statistical inference. Biologically more important, most phenotypes are not determined by single alleles, but by the joint effects, both additive and nonadditive, of a number of alleles at multiple loci. With the advent of modern imaging and measurement technology, complex phenotypes, such as the vertebrate brain or cranium, often are represented by large numbers of variables. This further complicates the study of genotype–phenotype association by tremendously increasing the number of pairwise comparisons between genetic loci and phenotypic variables, which may not be meaningful traits *per se* [for instance, in geometric morphometrics, voxel-based image analysis, and many behavioral studies (Bookstein 1991; Ashburner and Friston 2000; Mitteroecker and Gunz 2009; Houle *et al.* 2010)]. The genotype–phenotype associations we actually

doi: 10.1534/genetics.115.181339
Manuscript received July 29, 2015; accepted for publication February 15, 2016; published Early Online February 18, 2016.
Supplemental material is available online at www.genetics.org/lookup/suppl/doi:10.1534/genetics.115.181339/-/DC1.

[1] Corresponding author: Department of Theoretical Biology, University of Vienna, Althanstrasse 14, A-1090 Vienna, Austria. E-mail: philipp.mitteroecker@univie.ac.at

seek are between certain allele combinations from multiple loci and certain combinations of phenotypic variables that bear biological interpretation. The number of such pairs of "latent" allele combinations and phenotypes that underlie the observed genotype–phenotype association depends on the genetic-developmental system under study, but typically is less than the number of assessed loci and phenotypic variables (Hallgrimsson and Lieberman 2008; Martinez-Abadias *et al.* 2012).

Several methods have been suggested for such a multivariate mapping, including multiple and multivariate regression (Haley and Knott 1992; Jansen 1993; Hackett *et al.* 2001; de Los Campos *et al.* 2013), principal component regression (Wang and Abbott 2008), low-rank regression models (Zhu *et al.* 2014), partial least-squares regression (Bjørnstad *et al.* 2004; Bowman 2013), and canonical correlation analysis (Leamy *et al.* 1999; Ferreira and Purcell 2009). We present a multivariate analytic strategy—which we term *multivariate genotype–phenotype mapping* (MGP)—that embraces and relates all of these methods and that circumvents several of the problems resulting from pairwise univariate mapping and from the multivariate analysis of the loci separately from the phenotypes. Our approach does not primarily aim for the detection and location of single loci segregating with a given phenotypic trait. Instead, we present an approach that identifies patterns of allelic variation that are maximally associated—in terms of effect size—with patterns of phenotypic variation. In this way, we gain insight into the multivariate structure of genotype–phenotype association, including its dimensionality and the clustering of genetic and phenotypic variables within this association—the genetic-developmental properties determining the evolvability of organisms (Wagner and Altenberg 1996; Hendrikse *et al.* 2007; Mitteroecker 2009; Pavlicev and Hansen 2011).

## The Principle of Multivariate Genotype–Phenotype Mapping

Let there be $p$ genetic loci and $q$ phenotypic measurements scored for $n$ specimens. Instead of assessing each of the $pq$ pairwise genotype–phenotype associations, we seek a genetic effect—composed of the additive and nonadditive effects of multiple alleles—onto a phenotypic trait that is a composite of multiple measured phenotypic variables. As these genetic and phenotypic features are not directly measured, but perhaps present in the data, we refer to them as genetic and phenotypic "latent variables," $\mathrm{LV_G}$ and $\mathrm{LV_P}$ (Figure 1). The molecular, physiological, and developmental processes that underlie the genotype–phenotype relationship and that constitute the latent variables likely are complex nonlinear processes. In a first approximation, however, we consider the latent variables as linear combinations of the alleles and phenotypic measurements, respectively.

How do we identify these latent variables? This problem can be viewed in a dual way. First, it can be assumed that the effect of a genetic latent variable on a phenotypic latent variable is stronger than the effect of any single locus on any single phenotypic variable. We may thus seek genetic and phenotypic latent variables (linear combinations) with *maximal genotype–phenotype association*. In addition, there might be further pairs or "dimensions" of genetic and phenotypic latent variables with maximal associations, mutually independent across the dimensions, that together account for the observed genotype–phenotype association. The classic measures of genotype–phenotype association in the quantitative genetic literature are (i) genetic effect (average, additive, and dominance effect); (ii) genetic variance (the phenotypic variance accounted for by genetic effects); and (iii) heritability, the ratio of genetic to total phenotypic variance. Accordingly, we may compute latent variables that maximize one of these measures of genotype–phenotype association, depending on the scientific question and the information content in the data.

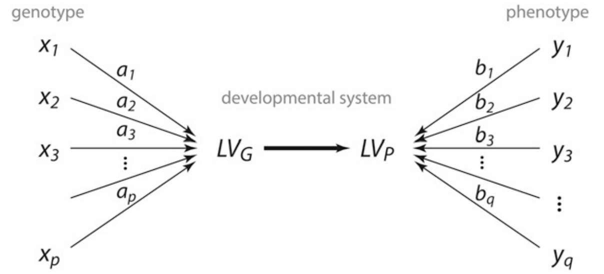


**Figure 1** Path diagram illustrating the principle of multivariate genotype–phenotype mapping. For $p$ loci $\mathbf{x}_1, \ldots, \mathbf{x}_p$ and $q$ phenotypes $\mathbf{y}_1, \ldots, \mathbf{y}_q$, the genotype–phenotype map acts via a genetic latent variable ($\mathrm{LV_G}$)—a joint effect of multiple loci—on a phenotypic latent variable ($\mathrm{LV_P}$), which is a combination of multiple measured phenotypic variables. The effects of the loci on the phenotype are denoted by the coefficients $a_1, a_2, \ldots, a_p$, and the composition of the phenotypic latent variable is denoted by the coefficients $b_1, b_2, \ldots, b_q$. These latent variables and their differential effects are properties of the genetic-developmental system of the studied organisms. Multivariate genotype–phenotype mapping seeks latent variables with a maximal genotype–phenotype association in the given sample (thick arrow).

A second, equivalent, way to view the problem is that of a search for simple patterns underlying the observed pairwise genotype–phenotype associations. Technically, we seek low-rank (*i.e.*, "simple") matrices that approximate the $p \times q$ matrix of pairwise associations. A powerful standard technique in multivariate statistics for this purpose is singular value decomposition (SVD). For a $p \times q$ matrix of pairwise genotype–phenotype associations, SVD finds pairs of singular vectors (one of length $p$, one of length $q$) of which the outer product best approximates the matrix in a least-squares sense. The two singular vectors can be interpreted as the genetic and phenotypic effects of the corresponding latent variables (the coefficients $a_i, b_i$ in Figure 1) that best approximate the observed genotype–phenotype associations. When connecting the two views, effect maximization and pattern search, the question arises: For which kind of matrices do the singular vectors (as simple patterns) lead to latent variables that maximize the above measures of genotype–phenotype association?

In *Materials and Methods* we demonstrate how to identify these latent genetic and phenotypic variables by an SVD of the appropriate association matrix. A more detailed derivation, including proofs, is given in *Appendix A*. In an application to a classic mouse sample, scored for 353 genetic markers and 11 phenotypic variables, we demonstrate the effectiveness of multivariate genotype–phenotype mapping. We show that a single dimension of latent variables already suffices to account for >70% of genetic variance present in the 11 traits. The first three dimensions together account for almost 90% of genetic variation and capture all the interpretable genotype–phenotype association in the data. Each dimension can be tested as a whole against the hypothesis of no association by a permutation approach, which reduces the number of significance tests from 7766 to 3. We discuss the consequences of this low dimensionality of the genotype–phenotype map for gene identification and for understanding the evolvability of organisms.

## Materials and Methods

### *Maximizing genotype–phenotype association*

Let each allele at each locus be represented by a vector $\mathbf{x}_i$ that contains the additive genotype scores (0, 1, or 2) for all $n$ subjects. Additional vectors may represent interactions of alleles at one locus (dominance scores, 0 or 1). See below for the implementation of epistasis, the interaction of alleles at different loci. We seek the combined effect of the alleles $(\mathbf{x}_1, \ldots, \mathbf{x}_p) = \mathbf{X}$ on a phenotype composed of the measured variables $(\mathbf{y}_1, \ldots, \mathbf{y}_q) = \mathbf{Y}$. For notational convenience, let the columns of $\mathbf{X}$ and $\mathbf{Y}$ be mean centered. Let the effects of the alleles on the phenotype be denoted by the $p \times 1$ vector $\mathbf{a} = (a_1, a_2, \ldots, a_p)'$ and the weightings of the measured variables that determine the phenotype by the $q \times 1$ vector $\mathbf{b} = (b_1, b_2, \ldots, b_q)'$. Then the genetic and phenotypic latent variables are given by the linear combinations $a_1\mathbf{x}_1 + a_2\mathbf{x}_2 + \ldots + a_p\mathbf{x}_p = \mathbf{Xa}$ and $b_1\mathbf{y}_1 + b_2\mathbf{y}_2 + \ldots + b_q\mathbf{y}_q = \mathbf{Yb}$ (Figure 1). Since, by definition, the latent variables have a stronger genotype–phenotype association than any single variable, the coefficient vectors $\mathbf{a}, \mathbf{b}$ are chosen to maximize the association between the corresponding latent variables: (i) genetic effect, (ii) genetic variance, or (iii) heritability (see *Appendix A*, *Univariate genotype–phenotype association*). In addition to this pair of latent variables, there are further pairs of latent variables with effects $\mathbf{a}_i$ and $\mathbf{b}_i$, independent of the previous ones, that together account for the observed genotype–phenotype association.

For any real $p \times q$ matrix, SVD yields a first pair of real singular vectors $\mathbf{u}_1, \mathbf{v}_1$, both of unit length, and a real singular value $\lambda_1$. The "left" singular vector $\mathbf{u}_1$ is of dimension $p \times 1$, and the "right" vector $\mathbf{v}_1$ is of dimension $q \times 1$. The outer product $\mathbf{u}_1\mathbf{v}_1'$, scaled by $\lambda_1$, is the rank-1 matrix that best approximates the matrix in a least-squares sense. There is also a second pair of singular vectors $\mathbf{u}_2, \mathbf{v}_2$, orthogonal to the first singular vectors, which are associated with a second singular value $\lambda_2$. Together, the two pairs of singular vectors give the best rank-2 approximation $\lambda_1\mathbf{u}_1\mathbf{v}'_1 + \lambda_2\mathbf{u}_2\mathbf{v}'_2$ of the matrix, and so forth for further dimensions. SVD yields optimal low-rank approximations of the original matrix by *maximizing* the singular values, that is, the contribution of the corresponding rank-1 matrix to the matrix approximation. The summed squared singular values equal the summed squared elements of the approximated matrix. The number of relevant dimensions (the rank of the approximation) can thus be determined by the squared singular values, expressed as a fraction of the total squared singular values.

For a matrix of pairwise genotype–phenotype associations, the singular vectors may serve as weightings for the genetic and phenotypic variables to compute the latent variables $\mathbf{Xu}_i$ and $\mathbf{Yv}_i$. The question is, For which kind of association matrices are the singular vectors $\mathbf{u}_i, \mathbf{v}_i$ the vectors $\mathbf{a}_i, \mathbf{b}_i$ maximizing (i) genetic effect, (ii) genetic variance, and (iii) heritability, respectively?

***Genetic effect (i):*** The *additive genetic effect* of one allele substitution is half the difference between the homozygote mean phenotypes, and the *dominance effect* is the deviation of the heterozygote mean phenotype from the midpoint of the homozygote mean phenotypes. By contrast, the *average effect* of an allele substitution is the average difference between offspring that get this allele and random offspring (Falconer and Mackay 1996; Roff 1997). In a sample of measured individuals, the average effect can be estimated by the regression slope of the phenotype on the additive genotype scores, whereas additive and dominance effects can be estimated by the two multiple-regression coefficients of the phenotype on both the additive and dominance genotype scores (see *Appendix A*, *Univariate genotype–phenotype association*).

For multivariate genotypes and phenotypes, maximizing the effect of the genetic latent variable on the phenotypic latent variable translates into finding vectors $\mathbf{a}$ and $\mathbf{b}$ that maximize the slope

$$\mathrm{Cov}(\mathbf{Xa}, \mathbf{Yb})/\mathrm{Var}(\mathbf{Xa}) \tag{1}$$

of the regression of the phenotype $\mathbf{Yb}$ on the allele combination $\mathbf{Xa}$. Under the constraint $\mathbf{a}'\mathbf{a} = \mathbf{b}'\mathbf{b} = 1$, this regression slope is maximized by the first pair of singular vectors $\mathbf{a} = \mathbf{u}_1, \mathbf{b} = \mathbf{v}_1$ of the matrix of regression coefficients

$$\mathbf{F} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y}. \tag{2}$$

For detailed derivations and proofs see *Appendix A*, *Multivariate genotype–phenotype association* and *Maximizing genotype–phenotype association via SVD*. See *Appendix B*, *Computational properties* for a discussion of the computational difficulties arising from the matrix inversion if $p > n$.

If the genetic variables $\mathbf{X}$ comprise the $p$ additive genotype scores only, maximizing the regression slope (1) maximizes the average effect, and the linear combination $\mathbf{Xa}$ can be interpreted as breeding values (sum of average effects). The $p$ elements of $\mathbf{a}$ are the partial average effects of the corresponding alleles on the phenotype $\mathbf{Yb}$ (average effects conditional on the other alleles); they are proportional to the regression coefficients of $\mathbf{Yb}$ on $\mathbf{X}$. The maximal slope (maximal average effect) associated with the linear combinations is given by the singular value $\lambda_1$.

If both additive and dominance scores are included in $\mathbf{X}$, the sum of both additive and dominance effects is maximized and the linear combination can be interpreted as genotypic values. The elements of $\mathbf{a}$, which is now of dimension $2p \times 1$, would then correspond to the partial additive and dominance effects on the phenotype $\mathbf{Yb}$, and the singular value is the sum of additive and dominance effects.

***Genetic variance (ii):*** For a population in Hardy–Weinberg equilibrium, the additive genetic variance can be expressed as the product of the variance of the additive genotype scores and the squared average effect (*cf*. Equation A3 in *Appendix A*). Maximizing genetic variance for multivariate genotypes and phenotypes thus is achieved by maximizing $\mathrm{Cov}(\mathbf{Xa}, \mathbf{Yb})^2/\mathrm{Var}(\mathbf{Xa})$ over the vectors $\mathbf{a}$ and $\mathbf{b}$. The latent variables with maximal genetic variance are given by the vectors $\mathbf{a} = (\mathbf{X}'\mathbf{X})^{-1/2}\mathbf{u}_1$ and $\mathbf{b} = \mathbf{v}_1$, where $\mathbf{u}_1$ and $\mathbf{v}_1$ are the first left and right singular vectors of the matrix

$$\mathbf{G} = (\mathbf{X}'\mathbf{X})^{-1/2}\mathbf{X}'\mathbf{Y}. \quad (3)$$

If the genetic variables comprise the additive scores only, this approach maximizes the additive genetic variance, given by $\lambda_1^2/(n-1)$, whereas if both additive and dominance scores are included, the total genetic variance (additive plus dominance variance) is maximized. This approach is computationally equivalent to reduced-rank regression (Izenman 1975; Aldrin 2000).

***Heritability (iii):*** Heritability can be expressed as the squared correlation coefficient of the phenotype and the genotype scores (*Appendix A*, *Univariate genotype–phenotype association*). Hence, maximizing heritability in the multivariate context amounts to maximizing the squared correlation $\mathrm{Cor}(\mathbf{Xa}, \mathbf{Yb})^2$, which is achieved by the vectors $\mathbf{a} = (\mathbf{X}'\mathbf{X})^{-1/2}\mathbf{u}_1$ and $\mathbf{b} = (\mathbf{Y}'\mathbf{Y})^{-1/2}\mathbf{v}_1$, where $\mathbf{u}_1$ and $\mathbf{v}_1$ are the left and right singular vectors of the matrix

$$\mathbf{H} = (\mathbf{X}'\mathbf{X})^{-1/2}\mathbf{X}'\mathbf{Y}(\mathbf{Y}'\mathbf{Y})^{-1/2}. \quad (4)$$

Including only the additive genotype scores in $\mathbf{X}$ maximizes the narrow-sense heritability $h^2$, whereas including both additive and dominance scores maximizes the broad-sense heritability $H^2$ resulting from additive and dominance variance. These maximal heritabilities equal the squared singular value $\lambda_1^2$. This approach is equivalent to canonical correlation analysis (*e.g.*, Mardia *et al.* 1979).

***Covariance (iv):*** Bjørnstad *et al.* (2004) and Mehmood *et al.* (2011) applied partial least-squares analysis (PLS) to identify genetic and phenotypic latent variables. PLS maximizes the covariance between the two linear combinations $\mathrm{Cov}(\mathbf{Xa}, \mathbf{Yb})$. The unit vectors $\mathbf{a}, \mathbf{b}$ maximizing this covariance can be computed as the first pair of singular vectors $\mathbf{u}_1, \mathbf{b}_1$ of the cross-covariance matrix $\mathbf{X}'\mathbf{Y}$. The maximal covariance is given by $\lambda_1/(n-1)$. The covariance between a phenotypic variable and a genetic variable has no correspondence in the classic genetic framework. However, this approach has convenient computational properties as it requires no matrix inverse (see *Appendix B*, *Computational properties*). The scaled genetic coefficients $\lambda_i\mathbf{a}_i/(n-1)$ are equal to the covariances between the corresponding locus and the phenotypic latent variable, without conditioning on the other loci as in approaches i–iii. See *Appendix A*, *Partial least-squares analysis* for more details.

### Properties of the four approaches

For each of the four approaches (Table 1), further dimensions (pairs of latent variables $\mathbf{a}_i, \mathbf{b}_i$) can be extracted by the subsequent pairs of singular vectors of the corresponding association matrix. The second dimension consists of a new allele combination (genetic latent variable) and a new phenotype (phenotypic latent variable) that are independent of the ones from the first dimension and have the second largest association, and similarly for further dimensions. The maximal number of dimensions is $\min(p, q, n-1)$. However, the notion of "independence" differs among the four approaches. In approaches i and iv—the maximizations of genetic effect and covariance—the genetic coefficient vectors are orthogonal ($\mathbf{a}_i'\mathbf{a}_j = 0$ for $i \neq j$ and 1 for $i = j$), whereas in approaches ii and iii—the maximizations of genetic variance and heritability—the genetic latent variables are uncorrelated ($\mathbf{a}_i'\mathbf{X}'\mathbf{X}\mathbf{a}_j = 0$ for $i \neq j$ and 1 for $i = j$). The phenotypic effects are orthogonal in approaches i, ii, and iv, whereas the phenotypic latent variables are uncorrelated in approach iii.

In the first three approaches, the corresponding association is maximized conditional on all other linear combinations of alleles, including the other latent variables. The matrix of regression coefficients of the phenotypic latent variables $\mathbf{YB}$ on the genetic latent variables $\mathbf{XA}$ thus is diagonal, where the matrix $\mathbf{A}$ contains the vectors $\mathbf{a}_i$ and $\mathbf{B}$ the vectors $\mathbf{b}_i$. In words, the prediction of the $i$th phenotypic latent variable by the $i$th genetic latent variable is not improved by adding any other latent variable or any linear combination of them as predictor. In this sense, the effect of the allele combination $\mathbf{Xa}_i$ on the phenotype $\mathbf{Yb}_i$ is "independent" of that of any other allele combination $\mathbf{Xa}_j$. Furthermore, in approaches ii, iii, and iv the genetic latent variables are correlated only with the corresponding phenotypic latent variable but not with any of the other latent variables: $\mathrm{Cov}(\mathbf{Xa}_i, \mathbf{Yb}_j) = 0$ for $i \neq j$ (see *Appendix A*, *Maximizing genotype–phenotype association via SVD* for proofs).

The constraints on the mutual independence (orthogonality or uncorrelatedness) of the genetic coefficient vectors $\mathbf{a}_i$

**Table 1 The four different approaches and the quantities they maximize, as well as the statistical methods to which they relate**

| Approach | Maximization | Related methods |
|---|---|---|
| i | Genetic effect | |
| ii | Genetic variance | Reduced-rank regression |
| iii | Heritability | Canonical correlation analysis |
| iv | Covariance | Partial least-squares analysis |

and the phenotypic coefficient vectors $\mathbf{b}_i$ are unlikely to reflect biological relationships and can complicate the reification of multiple latent variables as biological factors (see, *e.g.*, Cheverud 2007; Bookstein 2014). Supplemental Material, Figure S1 illustrates these properties by applications to simple simulated data sets. Unless only a single dominant dimension (pair of latent variables) is present in the data or all the important latent variables can clearly be interpreted, multiple dimensions should be interpreted jointly, as spanning a genetic subspace that is maximally associated with a phenotypic subspace. The summed squared singular values of these dimensions indicate their joint genotype–phenotype association.

Maximizing genetic effect (phenotypic change per unit genetic change) in approach i imposes a constraint on the norms of $\mathbf{a}$ and $\mathbf{b}$; that is, it implies a concept of "total" genetic and phenotypic effect. The vectors are computed to have a norm (summed squared elements) of 1, which is common in statistics but not an obvious choice in genetics (see *Appendix B*, *Geometric properties* for further discussion). By contrast, maximizing genetic variance and heritability in approaches ii and iii requires no constraint on the norm of $\mathbf{a}$; the explained phenotypic variance is unaffected by the norm of $\mathbf{a}$. (The SVD scales $\mathbf{a}$ so that the latent variable $\mathbf{Xa}$ has unit variance, but this choice has no effect on the maximal genetic variance.) In approaches ii and iii, the singular values and the right singular vectors are even unchanged under all linear transformations of the genetic variables $\mathbf{X}$ (see *Appendix B*, *Geometric properties* for a proof and Figure S1 for a demonstration). This implies that the maximal genetic variances (singular values) in approach ii as well as the phenotypes that show these maximal variances (right singular vectors) do not depend on the variances and covariances of the genetic variables (as they are scale dependent), that is, on genetic variance and linkage disequilibrium. They depend on the *relative* differences between the genotype scores only. The same applies to approach iii, but the heritabilities are even invariant to affine transformation of the phenotypic variables. To understand these properties intuitively, consider that the regression slope (genetic effect) depends on the variance of the predictor variable (the genotype scores), but the explained variance (genetic variance, heritability) does not. The genetic coefficients (left singular vectors) are affected by linear transformations of $\mathbf{X}$ because they are the multiple-regression coefficients of the phenotypic latent variable on the loci.

Approaches i, ii, and iv require meaningful phenotypic variances and covariances that are commensurate across the variables (because $\mathbf{b}$ is constrained to a norm of 1). This can be the case in morphometrics and chemometrics, but rarely in the behavioral sciences and psychometrics (see also Mitteroecker and Huttegger 2009; Huttegger and Mitteroecker 2011). If correlations, but not variances, are interpretable, for example when the variables have incommensurate scales, approaches i, ii, and iv may be applied after standardizing the variables separately to unit variance or unit mean (Hansen and Houle 2008). However, linear combinations of variance-standardized variables may be difficult to interpret; multivariate genotype–phenotype mapping is most powerful in contexts such as geometric morphometrics and image analysis, where the variables do not require standardization and interpretations are always in terms of linear combinations of measurements. When neither variances nor correlations are interpretable for the phenotypes, only approach iii—the maximization of heritability—can be applied.

### *Epistasis*

The interactions of alleles at different loci are not considered explicitly in the above descriptions. Epistatic effects contribute to the estimates as far as they contribute to the average or additive allelic effects (Cheverud and Routman 1995, 1996). However, it is possible to explicitly include epistatic effects by appending a design matrix to $\mathbf{X}$ that represents pairwise or higher-order allele interactions. When adding such interaction terms, approach i maximizes the sum of additive, dominance, and epistatic effects; approach ii maximizes genetic variance resulting from all three kinds of genetic effects (additive plus interaction variance); and approach iii maximizes the corresponding heritability.

As noted earlier, the allele combinations $\mathbf{Xa}_i$ have no mutual interactions in their effects on the corresponding phenotypic latent variable and in this sense show no epistasis. But this property should not be overinterpreted; it is primarily a convenient mathematical property, which holds true for all data sets.

### *Statistical significance*

The presented approaches are exploratory techniques for identifying multivariate patterns underlying the observed genotype–phenotype relationships. As long as no specific hypotheses about the effect of particular alleles have been formulated, no allele-specific hypothesis tests are appropriate. But one can test each pattern (pair of genetic and phenotypic latent variables) as a whole against the hypothesis of no genotype–phenotype association by a permutation test (*e.g.*, Churchill and Doerge 1994; Good 2000): Randomly reassign the phenotype vector of each individual to another individual in the data set and recompute the singular value. The *P*-value is given by the fraction of such permutations that yield an equal or larger singular value than the original one. A structured pedigree in an experimental cross may require a restricted or hierarchical permutation approach. Confidence

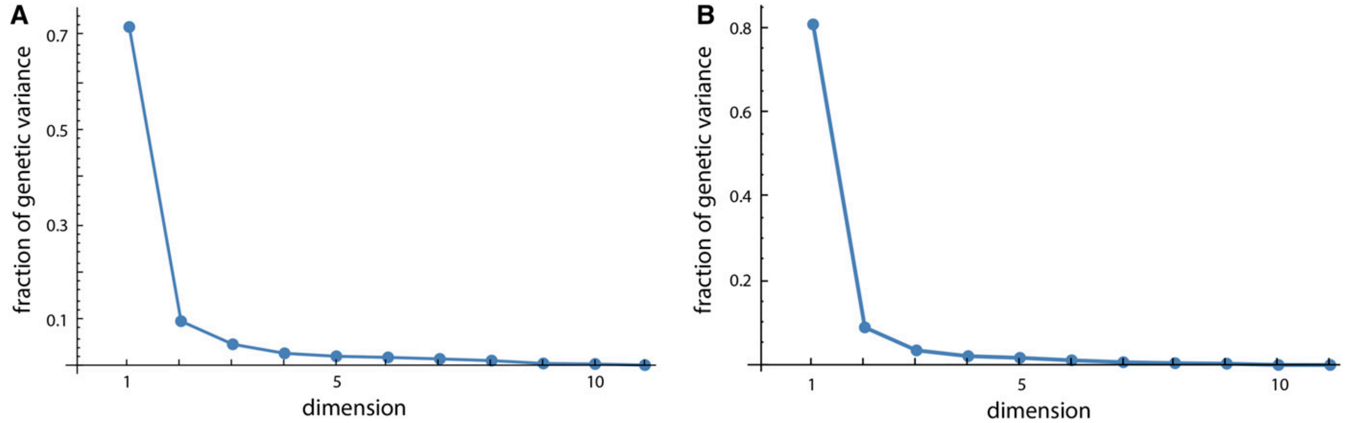


**Figure 2** (A) Genetic variances of the 11 latent variables, expressed as fractions of total genetic variance summed over all 11 dimensions, resulting from approach ii applied to the 353 loci and the 11 phenotypic measurements. The genetic variance comprises both additive and dominance variance. These are the 11 squared singular values of the matrix **G** in Equation 3, divided by their sum of squares. (B) Fractions of genetic variance for the 11 latent variables of the analysis of chromosome 6 only.

intervals of the genetic and phenotypic coefficients can be estimated by a bootstrap approach (*e.g.*, Visscher *et al.* 1996). These intervals can be used to compare the computational stability of coefficients across alleles or across dimensions. For the reasons mentioned earlier, they should not be used for inferences about allele-specific statistical significance.

### *Data and analysis*

We demonstrate multivariate genotype–phenotype mapping by an application to a classic mouse sample, comprising 1581 specimens of the $F_2$ and $F_3$ generations of an intercross of inbred LG/J and SM/J mice, obtained from The Jackson Laboratory (for details see Cheverud *et al.* 1996; Norgard *et al.* 2008). Each individual was genotyped at 384 polymorphic SNPs, of which we used the 353 SNPs on the 19 autosomal chromosomes. Every locus was represented by one variable for the additive genotype scores and one variable for the dominance scores. Total body weight (WTN) and the weights of the reproductive fat pad (FP), the heart (HT), the kidney (KD), the spleen (SP), and the liver (LV), as well as tail length (TL), were recorded for each individual. In addition, the lengths of the right set of long bones (humerus, ulna, femur, and tibia) were measured. Measurements were corrected for the effects of sex, litter size, and age at necropsy and standardized to unit variance. All work on mice was performed using a protocol approved by the Institutional Animal Care and Use Committee of the Washington University School of Medicine (St. Louis). These data may not be representative of modern high-throughput standards, but they allowed us to test our approach on a well-explored sample with easily interpretable variables.

We analyzed these data by approach ii—the maximization of total genetic variance—because genetic variance is the focus of most breeding studies and evolutionary models. The computation of the inverse square root matrix was based on a generalized inverse, using the first 80 principal components of the genetic variables, which accounted for 76% of total variation (see *Appendix B*, *Computational properties*). To check for overfitting the data (see *Discussion* and *Appendix B*, *Computational properties*), we used a leave-one-out cross-validation: Each individual's phenotype was predicted by the genetic and phenotypic coefficient vectors that were computed using only the other $n - 1$ individuals. The explained variance was then computed from these predictions.

### *Data availability*

All data used in the analysis are available at the DRYAD repository: DOI:10.5061/dryad.fc55k. All computations and visualizations were performed in Mathematica 9.0 (Wolfram Research Inc.).

## Results

Figure 2A shows the genetic variances for the 11 dimensions of latent variables in the style of a scree plot. The first dimension strikingly dominates the genotype–phenotype map by accounting for 72% of the total genetic variance summed over all 11 phenotypic variables (Table 2). The corresponding phenotypic latent variable represented overall limb length (high positive coefficients for all long bones; Figure 3A). More than 43% of the variance in this phenotypic latent variable was accounted for by the corresponding genetic latent variable (Table 2). The additive and dominance effects of the loci are represented by the genetic coefficients shown in Figure 4. They indicate strong additive effects on limb length mainly on chromosomes 1, 2, 3, 6, 8, 9, 13, and 17.

The second phenotypic latent variable reflected body and organ weight (Figure 3B), with additive and dominance effects mainly on chromosomes 2, 3, 6, 9, 10, 11, 12, 13, and 19 (Figure S2). Compared to dimension 1, the explained variance of body/organ weight was relatively small (7% of phenotypic variance). This pattern differs from the third phenotypic latent variable, which contrasted distal *vs.*

**Table 2 The first three dimensions of latent variables resulting from approach ii, the maximization of total genetic variance**

| Dimension | Phenotypic variance | Genetic variance (%) | Fraction of genetic variance | *P*-value |
|---|---|---|---|---|
| 1 | 3.35 | 1.43 (0.43) | 0.72 | <0.001 |
| 2 | 3.01 | 0.20 (0.07) | 0.10 | 0.063 |
| 3 | 0.39 | 0.10 (0.25) | 0.05 | <0.001 |

Shown are the variance of the phenotypic latent variable (linear combination of standardized phenotypic variables), the variance of this phenotypic latent variable explained by the genetic latent variable (*i.e.*, genetic variance), the genetic variance as a fraction of total genetic variance of all traits (plotted in Figure 2), and the *P*-value for the test of this dimension against the hypothesis of complete independence between genetic and phenotypic latent variables.

proximal long bone length. This trait varied little in the studied population but was under stronger genetic control than body weight (25% explained phenotypic variance). The strongest additive genetic effects were located on chromosomes 1 and 11 (Figure S3). These three dimensions together accounted for 87% of the genetic variance present in the 11 variables of this sample. The subsequent dimensions accounted for small portions of genetic variance only (<3%) and had no obvious interpretation.

Figure 5 shows a plot of the three scaled phenotypic coefficient vectors, constituting a phenotype space in which the phenotypic variables cluster according to their genetic structure. For approach ii applied here, the squared length of the vectors approximates the genetic variance of the corresponding phenotypic variable, and the cosine of the angle between the vectors approximates their genetic correlation. The long bones clustered together to the exclusion of the weight measurements, indicating a shared genetic basis. Humerus and femur as well as tibia und ulna showed particularly strong genetic correlations. Tail length was more closely correlated with the weight measurements than with the long bone lengths.

In a second analysis, we performed a separate and more detailed study of chromosome 6, which showed strong genetic effects. As genetic predictors we used the additive and dominance genotype scores for each of the 22 screened SNPs as well as for 89 loci imputed every 1 cM. This allows for the implementation of interval mapping (Lander and Botstein 1989) in multivariate genotype–phenotype mapping. These 222 genetic variables and the 11 phenotypic variables were analyzed again with approach ii, the maximization of genetic variance, resulting in two interpretable dimensions (Figure 2B and Figure 6). Bootstrap confidence intervals are shown for both the genetic and phenotypic effects. The computation of the matrix inverse was based on the first eight principal components of the genetic variables (89% of variance). The first dimension again represented limb length and accounted for even 85% of total genetic variance within the 11 phenotypic variables ($P < 0.001$). The additive genetic effects had two distinguished peaks, one at ~60–70 Mb and one at ~140 Mb. Both peaks were associated with small dominance effects. These results correspond well to the two loci affecting long bone length identified by Norgard *et al.* (2008); they were estimated at 85 and 144 Mb on chromosome 6 by applying traditional methods to the same data. The second dimension represented body and organ weight and accounted for 8% of total genetic variance ($P = 0.090$). Additive and dominance effects had three peaks, one close to the centromere, one at ~60–70 Mb, and one at the end of the chromosome. For the latter region, additive and dominance effects were of opposite sign. These three locations are in accordance with the loci identified by Vaughn *et al.* (1999) and Fawcett *et al.* (2008) for body and organ weight in the $F_2$ and $F_3$ generations of the same cross. The confidence intervals of both genetic and phenotypic coefficients for dimension 2 were considerably wider than that for dimension 1. Note that the confidence intervals should not be used for statistical inference in this exploratory context, only for the comparison of computational stability within the sample.

Since the $F_3$ population consisted of 200 sets of full sibs, we accounted for the genetic similarity between individuals in a separate analysis. We constructed an $n \times n$ matrix that represents genetic similarity between individuals (we tested both the expected similarity based on relatedness and the actual similarity based on the SNP data) and implemented this matrix in the estimation by a generalized least-squares approach (see *Appendix B*, *Generalized least squares*). This had only a limited effect on the results and we thus presented just the ordinary least-squares solutions here. We also repeated the analyses with different numbers of principal components (PCs) used for the matrix inversion. The genetic coefficients were stable against small changes of the numbers of PCs; the phenotypic coefficients and the shape of the scree plot were stable even over a very wide range of PCs. Taking the cube root of the weight measurements before variance standardization had basically no effect on the results. We checked whether outliers could drive some of the results, but found no evidence in scatter plots of genetic *vs.* phenotypic latent variables. We also applied the other three maximization approaches to the data, which basically resulted in the same first two pairs of phenotypic and genetic latent variables with similar explained variances (see *Appendix C* for a brief presentation of these results).

## Discussion

Many traits are affected by numerous alleles with small or intermediate effects, which are difficult to detect by mapping each locus separately. Loci selected by separately computed *P*-values often account for low fractions of phenotypic variance and provide an incomplete picture of the genotype–phenotype map (Manolio *et al.* 2009; Eichler *et al.* 2010). "Whole-genome prediction" methods, which are based on all scored loci, have proved more effective in explaining phenotypic variation of a trait (Yang *et al.* 2010; de Los Campos *et al.*

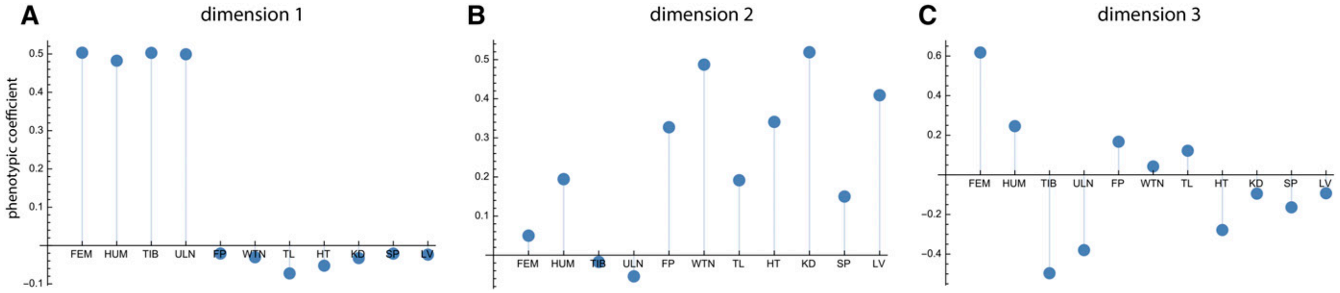


**Figure 3** Phenotypic coefficients for the first three dimensions (A–C) of the analysis of all 19 chromosomes. These are the vectors $\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3$ resulting from approach ii and represent the composition of the phenotypic latent variables. The phenotypic measurements are the lengths of the femur, the humerus, the tibia, and the ulna (FEM, HUM, TIB, ULN); weight of the fat pad (FP); body weight (WTN); tail length (TL); and the weights of the heart, the kidneys, the spleen, and the liver (HT, KD, SP, LV).

2013). However, many complex phenotypes cannot be adequately represented by a single variable but require multiple measurements. For such multivariate traits, we presented a "whole-genome, whole-phenotype prediction" method that identifies the genetically determined traits and their associated allelic effects in a single step. This avoids an inefficient decomposition (such as principal component analysis) of the phenotypic variables separately from the genetic variables (Cheverud 2007); instead our method provides a decomposition of the genotype–phenotype map itself.

Multivariate genotype–phenotype mapping separates phenotypic features under strong genetic determination from features with less genetic control, whereas traditional mapping of complex traits typically lumps different phenotypic features with different heritabilities. For example, we found that overall limb length in our mouse sample is both highly variable and highly heritable (43% explained phenotypic variance); a second feature, distal *vs.* proximal limb bone length, is much less variable but still shows an explained variance of 25%. All other aspects of long bone variation, *e.g.*, forelimb *vs.* hindlimb length, show very little genetic variation. Mapping each of the four limb bones separately thus leads to estimates of explained variance that are averages across all these features, some of which have high heritability and some of which have basically none. It thus misses the actual signal: the traits (latent variables) under strong genetic control. Multivariate genotype–phenotype mapping can tell one *where* to look for the (missing) heritability in the phenotype and shows the allelic pattern associated with this phenotype.

For the same mouse population, Norgard *et al.* (2008) estimated the heritability of long bone length to be ~0.9, of which we could explain almost half by dimension 1 of the multivariate mapping. The lack of the remaining heritability is due likely to the limited number of SNPs and incomplete linkage disequilibrium between causal variants and genotyped SNPs (Yang *et al.* 2010).

Estimates of explained variance based on all scored loci tend to be too high because of massive overfitting and, hence, may not reflect actual prediction accuracy (Makowsky *et al.* 2011; Gianola *et al.* 2014). After applying leave-one-out cross-validation, dimension 1 (limb length) still showed an explained variance of 0.36 and dimension 3 of 0.17. The reduction of the genetic variables to the first 80 PCs together with the identification of the relevant predictor variable (the genetic latent variable) thus prevented severe overfitting. Our estimates include both additive and dominance effects, but most of the explained variance was due to additive gene effects (fractions of phenotypic variance explained by additive effects were 0.40, 0.05, and 0.23 for the three dimensions).

The variance of body/organ weight (second latent variable) that was explained by the genetic latent variable was relatively small (7% of phenotypic variance), which is somewhat surprising since the two parental strains were selected for small and large body size, respectively. Apart from substantial environmental variance, this may result from the considerable sex interactions identified by Fawcett *et al.* (2008), which we did not include in our analysis. Note also that dimension 2 covers the genetic effects on body/organ weight, *independent* of the effects on limb length, so some of the QTL effects on weight might have been captured by a more general size factor with the limb length. Higher estimates of explained variance of body weight in earlier genome-wide studies likely resulted from overfitting the genotype–phenotype relationship. For example, Kramer *et al.* (1998) explained 47% of phenotypic variance in body weight by a multiple regression on the additive and dominance scores of all scored SNPs. We could reproduce this result with the present sample, but when applying a leave-one-out cross-validation, this fraction dropped to ~3%. This severe overfitting by the multiple regression is not surprising, given that we found only a single allele combination to be considerably (and presumably causally) associated with body/organ weight. The 705 remaining combinations of genotype scores inflated the "explained variance" by random associations with body weight (note that we had additive and dominance scores for 353 loci and hence 706 independent linear combinations of scores).

Multivariate genotype–phenotype mapping is primarily an exploratory method for investigating the multivariate structure of genotype–phenotype association. However, it can provide crucial information for gene identification. In our mouse

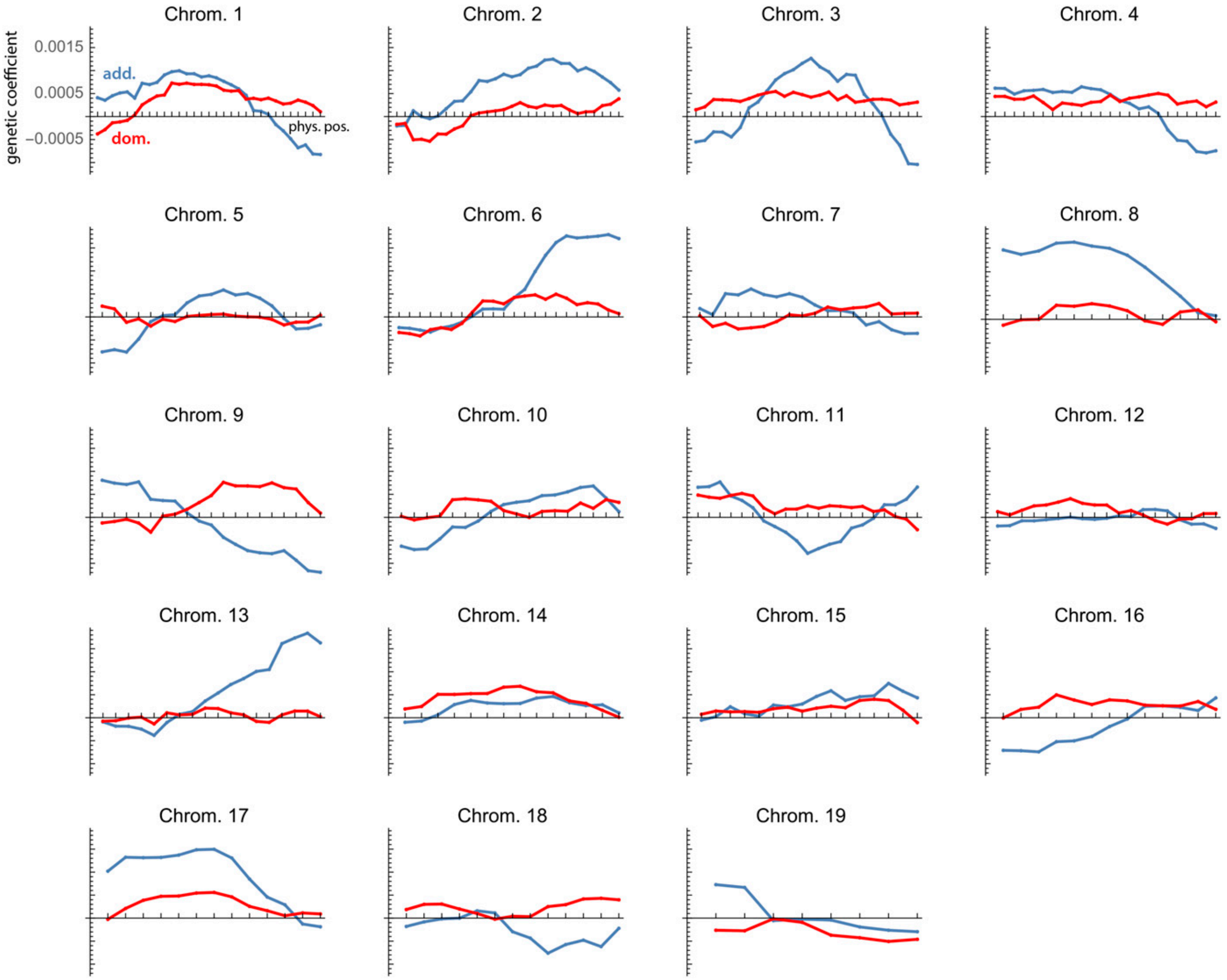


**Figure 4** Genetic coefficients for the first dimension of the analysis of all 19 chromosomes. They are the elements of the vector $\mathbf{a}_1$ from approach ii and represent the partial additive and dominance effects (blue and red lines) of all 353 loci on the corresponding phenotypic latent variable (limb length; *cf*. Figure 3).

data, for instance, we found three independent dimensions of genotype–phenotype association, each of which could be tested as a whole against the null hypothesis of no association. In fact, this is the number of independent statistically meaningful tests that can be made. The performance of all pairwise tests between genetic and phenotypic variables (7766 for our data) is a misuse of significance testing in an entirely exploratory context (McCloskey and Zilik 2009; Bookstein 2014; Mitteroecker 2015) that does not guarantee repeatable results (Morgan *et al.* 2007). Of course, the biological meaning of a hypothesis about the complete lack of genotype–phenotype association remains doubtful nonetheless. Dimensions 1 and 3 were highly statistically significant as a whole in our data; dimension 2 was convincing as a pattern but not as clearly significant as the other dimensions because of the large fraction of environmental variance ($P = 0.063$; Table 2). Identification of important alleles should be based on the effect sizes (the genetic coefficients), unless more specific prior hypotheses about gene effects existed. The fourth and all subsequent dimensions did not differ significantly from a random association ($P > 0.30$).

For complex phenotypes measured by multiple variables, multivariate genotype–phenotype mapping should precede any other mapping technique to identify the number of independent dimensions of genotype–phenotype association. In particular, this applies to variables that do not bear biological meaning one by one, such as in modern morphometrics and image analysis, but also to gene expression profiles and similar “big data.” Multivariate genotype–phenotype mapping identifies the phenotypes (linear combinations of measurements) under strong genetic control that are worth considering for further genetic analysis. In addition to the allelic effects estimated by multivariate mapping, other measures of allelic effects or LOD scores can be computed by more classic methods for the identified phenotypic latent variables.

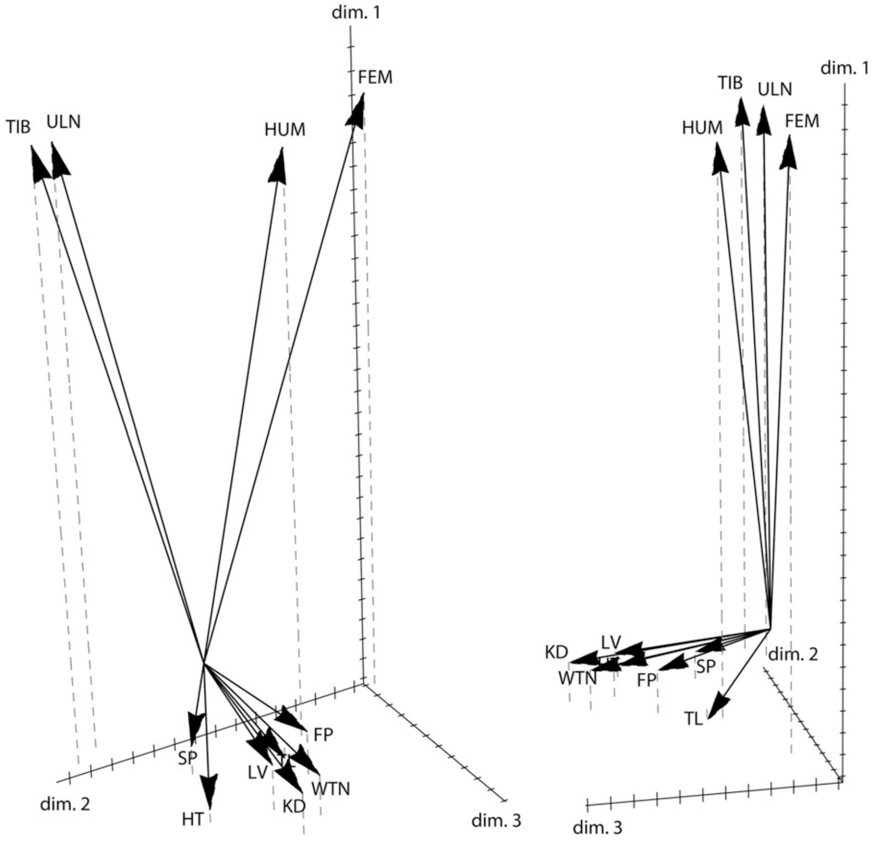


**Figure 5** Two different projections of the three-dimensional phenotype space resulting from the first three scaled phenotypic coefficient vectors ($\lambda_1\mathbf{b}_1, \lambda_2\mathbf{b}_2, \lambda_3\mathbf{b}_3$). For approach ii applied here—the maximization of genetic variance—the squared length of the vectors approximates the genetic variance of the corresponding phenotypic variable, and the cosine of the angle between the vectors approximates their genetic correlation. Clustering of phenotypic variables in this diagram thus indicates shared genetic control.

We presented four different variants of multivariate genotype–phenotype mapping, which maximize different measures of association: (i) genetic effect, (ii) genetic variance, (iii) heritability, and (iv) the covariance between genetic and phenotypic latent variables (Table 1). The choice among them depends on the scientific question and the kind of phenotypic variables. While the genetic effect may be of interest in certain medical studies, additive genetic variance is central to many evolutionary studies and breeding experiments. Maximizing heritability tends to be the most unstable approach because it maximizes genetic variance and minimizes environmental variance at the same time. Approach iv, the maximization of covariance, has no correspondence in quantitative genetics but it is computationally simple and avoids severe overfitting without prior variable reduction (Martens and Naes 1989). The genetic coefficients in approaches i–iii represent partial effects (*i.e.*, effects conditional on the other loci), whereas the coefficients in approach iv do not depend on the other loci in this way. Approach iv thus offers an alternative to the other approaches when computational simplicity is preferred over interpretability or when partial coefficients should be avoided.

In addition to purely additive or average gene effects, nonadditive effects can be incorporated into the analysis by adding variables representing dominance or epistasis (pairwise or higher-order interaction terms) to the genetic predictors. Accordingly, the genetic latent variables can be interpreted either as breeding values or as genotypic values. Multivariate genotype–phenotype mapping can be applied to crosses of inbred strains as well as to natural populations. Genetic similarity and common ancestry can be accounted for by generalized least-squares variants (see *Appendix B*, *Generalized least squares*). In addition to genetic variables, covariates such as environmental variables can be included in the predictors as well. In approaches i–iii, the resulting coefficients of the gene effects are then conditional on these covariates. The presented methods make no distributional assumptions and do not require linear relationships between genetic and phenotypic latent variables or covariates. However, only if all relationships are linear (and, hence, the variables jointly normally distributed), uncorrelatedness implies actual independence. The interpretation of the singular values of **F**, **G**, and **H** as genetic effect, genetic variance, and heritability, respectively, is exact only for randomly mating populations in Hardy–Weinberg equilibrium. The more a population deviates from equilibrium, the more the singular values may deviate from these genetic quantities.

For a single phenotypic variable only, approaches i–iii lead to the same genetic latent variables and the same genetic coefficients, which are the regression coefficients of the phenotype on the loci. The corresponding linear combination of loci maximizes all three measures of genotype–phenotype association: genetic effect, genetic variance, and heritability. In the classic genetic literature, this is also known as the “selection index” (Smith 1936; Hazel 1943). The three approaches can thus be construed as three different generalizations of multiple

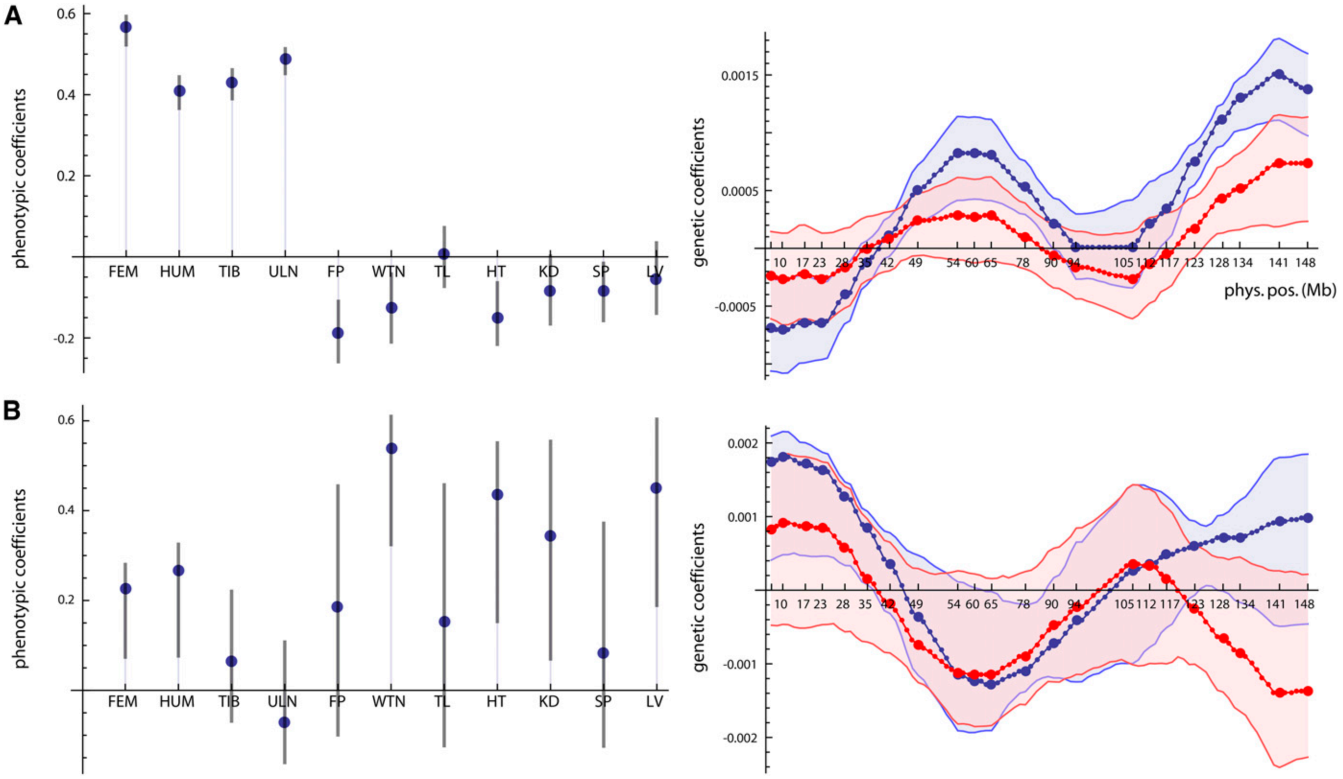


**Figure 6** (A and B) Phenotypic and genetic coefficients with 90% confidence intervals of the first (A) and the second (B) pair of latent variables for the separate analysis of chromosome 6. The first phenotypic latent variable reflected limb length and the second one body and organ weight. Additive and dominance effects are represented by blue and red lines, respectively. The 22 screened loci are represented by large symbols and the 89 imputed loci by small symbols.

regression to many phenotypic traits; the genetic coefficients (the elements of the vectors $\mathbf{a}_i$) in all three approaches equal the multiple-regression coefficients of the corresponding $i$th phenotypic latent variable on the loci.

This property allows one to rotate the phenotypic latent variables to increase their biological interpretability, as in exploratory factor analysis, and to estimate the corresponding genetic effects. For the mouse data, the first 3 dimensions of phenotypic latent variables constituted a 3-dimensional subspace of the 11-dimensional phenotype space, which contained almost all of the genetic variation in the data. The first latent variable was overall limb length, and the third one was a contrast between distal and proximal limb bone lengths. Hence, femur and humerus, as well as tibia and ulna, were highly correlated and clustered in Figure 5. Alternative latent variables would thus be proximal limb length (*femur* + *humerus*) and distal limb length (*tibia* + *ulna*). The corresponding genetic latent variables can be computed by multiple regression of the new latent variable on the loci.

In most modern genetic data sets $p$ clearly exceeds $n$, which challenges least-squares methods such as approaches i–iii. In *Appendix B, Computational properties* we show how they can be computed based on generalized inverses or matrix regularizations. Approach iv—the partial least-squares analysis—does not require the inversion of a matrix and can also be applied to collinear genetic variables and when $p > n$. Clearly, there is much room for improvement, such as an implementation in a Bayesian framework and the application of other penalized or BLUE/BLUP methods (*e.g.*, Meuwissen *et al.* 2001; Lopes and West 2004; de Los Campos *et al.* 2013; Zhu *et al.* 2014). The use of information measures, such as the application of PLS to Kullback–Leibler divergences by Bowman (2013), may allow for the application of the presented approaches to a wide range of heterogeneous variables.

In our 11-dimensional phenotype space, only 3 dimensions had considerable genetic variation, but the majority of genetic variation was even concentrated in a single dimension. In this sense, the genotype–phenotype map in this population is of surprisingly "low dimension" (even if the metaphor of dimensionality does not uniquely translate into an integer or a real number). In a preliminary analysis, we found similar results for the $F_9$ and $F_{10}$ generations of the same mouse cross. At least in part, this low-dimensional genotype–phenotype map resulted from the intercross of two inbred populations. It remains to be investigated to what degree it is also characteristic of outbred populations. Current studies of phenotypic and genetic variance–covariance patterns provide inconsistent results in this regard (*e.g.*, Kirkpatrick and Lofsvold 1992; Mezey and Houle 2005; Hine and Blows 2006; Pavlicev *et al.* 2009). Hallgrimsson and Lieberman (2008) speculated that a low-dimensional pattern of phenotypic variation is a general phenomenon that results from the "funneling" of the vast amount of genetic variation by a few central developmental pathways and morphogenetic processes. This would massively bias and constrain a population's phenotypic response to natural or artificial selection and generate a broad heterogeneity of genetic responses within a single selection scenario. If such funneling processes exist, multivariate genotype–phenotype mapping can help to identify these central pathways.

## Acknowledgments

We are very grateful to Fred L. Bookstein, Hans Nemeschkal, and Günter Wagner, as well as to three anonymous reviewers and William Valdar for very helpful comments and discussion. M.P. acknowledges support from the March of Dimes Prematurity Research Center Ohio Collaborative (grant 22-FY14-470).

## Literature Cited

Aldrin, M., 2000 Multivariate prediction using softly shrunk reduced-rank regression. Am. Stat. 54(1): 29–34.

Ashburner, J., and K. J. Friston, 2000 Voxel-based morphometry—the methods. Neuroimage 11: 805–821.

Bjørnstad, A., F. Westad, and H. Martens, 2004 Analysis of genetic marker-phenotype relationships by jack-knifed partial least squares regression (plsr). Hereditas 141(2): 149–165.

Bookstein, F., 1991 *Morphometric Tools for Landmark Data: Geometry and Biology*. Cambridge University Press, Cambridge, UK/London/New York.

Bookstein, F. L., 2014 *Reasoning and Measuring: Numerical Inferences in the Sciences*. Cambridge University Press, Cambridge, UK.

Bowman, C. E., 2013 Discovering pharmacogenetic latent structure features using divergences. *J Pharmacogenomics Pharmacoproteomics*, 4(1): 1000e134.

Cheverud, J. M., 2007 The dangers of diagonalization. J. Evol. Biol. 20(1): 15–16.

Cheverud, J. M., and E. J. Routman, 1995 Epistasis and its contribution to genetic variance components. Genetics 139: 1455–1461.

Cheverud, J. M., and E. J. Routman, 1996 Epistasis as a source of increased additive genetic variance at population bottlenecks. Evolution 50(3): 1042–1051.

Cheverud, J. M., E. J. Routman, F. A. Duarte, B. van Swinderen, K. Cothran *et al.*, 1996 Quantitative trait loci for murine growth. Genetics 142: 1305–1319.

Churchill, G. A., and R. W. Doerge, 1994 Empirical threshold values for quantitative trait mapping. Genetics 138: 963–971.

de Los Campos, G., J. M. Hickey, R. Pong-Wong, H. D. Daetwyler, and M. P. L. Calus, 2013 Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics 193: 327–345.

Eichler, E. E., J. Flint, G. Gibson, A. Kong, S. M. Leal *et al.*, 2010 Missing heritability and strategies for finding the underlying causes of complex disease. Nat. Rev. Genet. 11(6): 446–450.

Falconer, D. S., and T. F. C. Mackay, 1996 *Introduction to Quantitative Genetics*. Longman, Essex, UK.

Fawcett, G. L., C. C. Roseman, J. P. Jarvis, B. Wang, J. B. Wolf *et al.*, 2008 Genetic architecture of adiposity and organ weight using combined generation QTL analysis. Obesity 16(8): 1861–1868.

Ferreira, M. A. R., and S. M. Purcell, 2009 A multivariate test of association. Bioinformatics 25(1): 132–133.

Gianola, D., K. A. Weigel, N. Krämer, A. Stella, and C.-C. Schön, 2014 Enhancing genome-enabled prediction by bagging genomic BLUP. PLoS ONE 9(4): e91693.

Good, P., 2000 *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses* (Springer Series in Statistics, Ed. 2). Springer-Verlag, New York.

Hackett, C. A., R. C. Meyer, and W. T. Thomas, 2001 Multi-trait QTL mapping in barley using multivariate regression. Genet. Res. 77(1): 95–106.

Haley, C. S., and S. A. Knott, 1992 A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. Heredity 69(4): 315–324.

Hallgrimsson, B., and D. E. Lieberman, 2008 Mouse models and the evolutionary developmental biology of the skull. Integr. Comp. Biol. 48: 373–384.

Hansen, T., C. Pélabon, and D. Houle, 2011 Heritability is not evolvability. Evol. Biol. 38: 258–277.

Hansen, T. F., and D. Houle, 2008 Measuring and comparing evolvability and constraint in multivariate characters. J. Evol. Biol. 21(5): 1201–1219.

Hayes, B. J., P. M. Visscher, and M. E. Goddard, 2009 Increased accuracy of artificial selection by using the realized relationship matrix. Genet. Res. 91(1): 47–60.

Hazel, L. N., 1943 The genetic basis for constructing selection indexes. Genetics 28: 467–490.

Hendrikse, J. L., T. E. Parsons, and B. Hallgrímsson, 2007 Evolvability as the proper focus of evolutionary developmental biology. Evol. Dev. 9(4): 393–401.

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

Houle, D., D. R. Govindaraju, and S. Omholt, 2010 Phenomics: the next challenge. Nat. Rev. Genet. 11(12): 855–866.

Huttegger, S., and P. Mitteroecker, 2011 Invariance and meaningfulness in phenotype spaces. Evol. Biol. 38: 335–352.

Izenman, A. J., 1975 Reduced-rank regression for the multivariate linear model. J. Multivariate Anal. 5: 248–264.

Jansen, R. C., 1993 Interval mapping of multiple quantitative trait loci. Genetics 135: 205–211.

Kirkpatrick, M., and D. Lofsvold, 1992 Measuring selection and constraint in the evolution of growth. Evolution 46(4): 954–971.

Kramer, M. G., T. T. Vaughn, L. S. Pletscher, K. King-Ellison, E. Adams *et al.*, 1998 Genetic variation in body weight gain and composition in the intercross of large (lg/j) and small (sm/j) inbred strains of mice. Genet. Mol. Biol. 21(2): 211–218.

Lander, E. S., and D. Botstein, 1989 Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics 121: 185–199.

Leamy, L., E. J. Routman, and J. M. Cheverud, 1999 Quantitative trait loci for early- and late-developing skull characteristics in mice: a test of the genetic independence model of morphological integration. Am. Nat. 153(2): 201–214.

Lopes, H. F., and M. West, 2004 Bayesian model assessment in factor analysis. Stat. Sin. 14(1): 41–67.

Makowsky, R., N. M. Pajewski, Y. C. Klimentidis, A. I. Vazquez, C. W. Duarte *et al.*, 2011 Beyond missing heritability: prediction of complex traits. PLoS Genet. 7(4): e1002051.

Manolio, T. A., F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff *et al.*, 2009 Finding the missing heritability of complex diseases. Nature 461(7265): 747–753.

Mardia, K. V., J. T. Kent, and J. M. Bibby, 1979 *Multivariate Analysis*. Academic Press, London.

Martens, H., and T. Naes, 1989 *Multivariate Calibration*. John Wiley & Sons, Chichester, UK.

Martinez-Abadias, N., P. Mitteroecker, T. E. Parsons, M. Esparza, T. Sjovold *et al.*, 2012 The developmental basis of quantitative craniofacial variation in humans and mice. Evol. Biol. 39(4): 554–567.

McCloskey, D., and S. Zilik, 2009 The unreasonable ineffectiveness of Fisherian "tests" in biology, and especially in medicine. Biol. Theory 4(1): 44–53.

Mehmood, T., H. Martens, S. Saebø, J. Warringer, and L. Snipen, 2011 Mining for genotype-phenotype relations in Saccharomyces using partial least squares. BMC Bioinformatics 12: 318.

Meuwissen, T. H., B. J. Hayes, and M. E. Goddard, 2001 Prediction of total genetic value using genome-wide dense marker maps. Genetics 157: 1819–1829.

Mezey, J. G., and D. Houle, 2005 The dimensionality of genetic variation for wing shape in Drosophila melanogaster. Evolution 59(5): 1027–1038.

Mitteroecker, P., 2009 The developmental basis of variational modularity: insights from quantitative genetics, morphometrics, and developmental biology. Evol. Biol. 36: 377–385.

Mitteroecker, P., 2015 Systems mapping has potential to overcome inherent problems of genetic mapping: comment on "mapping complex traits as a dynamic system" by L. Sun and R. Wu. Phys. Life Rev. 13: 190–191.

Mitteroecker, P., and P. Gunz, 2009 Advances in geometric morphometrics. Evol. Biol. 36: 235–247.

Mitteroecker, P., and S. Huttegger, 2009 The concept of morphospaces in evolutionary and developmental biology: mathematics and metaphors. Biol. Theory 4(1): 54–67.

Morgan, T. M., H. M. Krumholz, R. P. Lifton, and J. A. Spertus, 2007 Nonvalidation of reported genetic risk factors for acute coronary syndrome in a large-scale replication study. JAMA 297(14): 1551–1561.

Norgard, E. A., C. C. Roseman, G. L. Fawcett, M. Pavlicev, C. D. Morgan *et al.*, 2008 Identification of quantitative trait loci affecting murine long bone length in a two-generation intercross of lg/j and sm/j mice. J. Bone Miner. Res. 23(6): 887–895.

Pavlicev, M., and T. Hansen, 2011 Genotype-phenotype maps maximizing evolvability: modularity revisited. Evol. Biol. 38(4): 371–389.

Pavlicev, M., J. M. Cheverud, and G. P. Wagner, 2009 Measuring morphological integration using eigenvalue variance. Evol. Biol. 36: 157–170.

Roff, D. A., 1997 *Evolutionary Quantitative Genetics*. Chapman & Hall, New York.

Sampson, P. D., A. P. Streissguth, H. M. Barr, and F. L. Bookstein, 1989 Neurobehavioral effects of prenatal alcohol: Part ii. Partial least squares analysis. Neurotoxicol. Teratol. 11(5): 477–491.

Smith, H. F., 1936 A discriminant function for plant selection. Ann. Eugen. 7(3): 240–250.

Tibshirani, R., 1996 Regression shrinkage and selection via the lasso. J. R. Stat. Soc. B 58(1): 267–288.

Vaughn, T. T., L. S. Pletscher, A. Peripato, K. King-Ellison, E. Adams *et al.*, 1999 Mapping quantitative trait loci for murine growth: a closer look at genetic architecture. Genet. Res. 74(3): 313–322.

Visscher, P. M., R. Thompson, and C. S. Haley, 1996 Confidence intervals in QTL mapping by bootstrapping. Genetics 143: 1013–1020.

Wagner, G. P., and L. Altenberg, 1996 Complex adaptations and the evolution of evolvability. Evolution 50(3): 967–976.

Wang, K., and D. Abbott, 2008 A principal components regression approach to multilocus genetic association studies. Genet. Epidemiol. 32(2): 108–118.

Yang, J., B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders *et al.*, 2010 Common SNPs explain a large proportion of the heritability for human height. Nat. Genet. 42(7): 565–569.

Zhu, H., Z. Khondker, Z. Lu, and J. G. Ibrahim Alzheimer's Disease Neuroimaging Initiative, 2014 Bayesian generalized low rank regression models for neuroimaging phenotypes and genetic markers. J. Am. Stat. Assoc. 109(507): 977–990.

*Communicating editor: W. Valdar*

# Appendices

## Appendix A: Derivation and Proofs

### *Univariate genotype–phenotype association*

To show how—for an idealized, randomly mating population—the singular value decompositions of the matrices $\mathbf{F}$, $\mathbf{G}$, and $\mathbf{H}$ introduced in (2), (3), and (4), respectively, lead to latent variables with maximal genotype–phenotype association, we first review the classic measures of association in the simple case of one diploid locus with two alleles affecting one quantitative phenotypic trait. For a sample of $n$ specimens, let the $n \times 1$ vector $\mathbf{x}$ contain the additive genotype scores (1 for heterozygotes and 0 or 2 for the two possible homozygotes) and the $n \times 1$ vector $\mathbf{y}$ contain the corresponding phenotypes. Consider further the $n \times 1$ vector $\mathbf{d}$ containing the dominance scores (0 for homozygotes and 1 for heterozygotes) and the $n \times 2$ matrix $\mathbf{Z} = (\mathbf{x}, \mathbf{d})$. For notational convenience, let $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{d}$ be mean centered so that $\mathbf{x}'\mathbf{1} = \mathbf{y}'\mathbf{1} = \mathbf{d}'\mathbf{1} = 0$, where the superscript $'$ indicates the transpose operation and $\mathbf{1}$ a vector of 1's.

In a sample of measured individuals, the additive and dominance effects, $a$ and $d$, are numerically identical to the regression coefficients of the multiple regression of the phenotype $\mathbf{y}$ on both $\mathbf{x}$ and $\mathbf{d}$ :

$$(a, d)' = \left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{y}. \tag{A1}$$

By contrast, the average effect, $\alpha$, of an allele substitution equals the bivariate regression slope of the phenotype $\mathbf{y}$ on the additive genotype scores $\mathbf{x}$ :

$$\alpha = \mathrm{Cov}(\mathbf{x}, \mathbf{y})/\mathrm{Var}(\mathbf{x}) = \mathbf{x}'\mathbf{y}/\mathbf{x}'\mathbf{x}. \tag{A2}$$

The sum of the average effects of all alleles constitutes the breeding value of an individual.

The additive genetic variance, $V_{\mathrm{A}}$, of $\mathbf{y}$ owing to the variance in $\mathbf{x}$ equals $2p_1p_2\alpha^2$ (Falconer and Mackay 1996; Roff 1997). For a population in Hardy–Weinberg equilibrium, $2p_1p_2 = \mathrm{Var}(\mathbf{x})$, and thus the additive genetic variance can be expressed as

$$\begin{aligned} V_{\mathrm{A}} &= \mathrm{Var}(\mathbf{x})\mathrm{Cov}(\mathbf{x}, \mathbf{y})^2/\mathrm{Var}(\mathbf{x})^2 \\ &= \mathrm{Cov}(\mathbf{x}, \mathbf{y})^2/\mathrm{Var}(\mathbf{x}) \\ &= \mathrm{Cov}(\mathbf{x}_{\mathrm{s}}, \mathbf{y})^2, \end{aligned} \tag{A3}$$

where $\mathbf{x}_{\mathrm{s}} = \mathbf{x}/\mathrm{Var}(\mathbf{x})^{1/2}$ is $\mathbf{x}$ standardized to unit variance. The dominance variance, $V_{\mathrm{D}}$, is equal to $(2p_1p_2d)^2 = \mathrm{Var}(\mathbf{x})^2d^2$.

The additive genetic variance of $\mathbf{y}$ due to $\mathbf{x}$, expressed as a fraction of the total variance of $\mathbf{y}$, is the narrow-sense heritability $h^2$ of $\mathbf{y}$ resulting from variation in the studied locus. It can be expressed as the squared correlation between the locus and the phenotype,

$$\begin{aligned} h^2 &= V_{\mathrm{A}}/\mathrm{Var}(\mathbf{y}) \\ &= \mathrm{Cor}(\mathbf{x}, \mathbf{y})^2 \\ &= \mathrm{Cov}(\mathbf{x}_{\mathrm{s}}, \mathbf{y}_{\mathrm{s}})^2, \end{aligned} \tag{A4}$$

where $\mathbf{y}_{\mathrm{s}}$ is the phenotype standardized to unit variance. The broad-sense heritability $H^2$ is the sum of additive and dominance variance as a fraction of phenotypic variance, which equals the squared multiple correlation coefficient resulting from the regression of $\mathbf{y}$ on $\mathbf{Z}$.

All these parameters represent different aspects of the genotype–phenotype relationship. While $a$ and $d$ are properties of the genotype alone, $\alpha$ and $V_{\mathrm{A}}$ are population properties that represent the potential to respond to natural or artificial selection. In contrast to $a$ and $d$, the average effect $\alpha$ depends on the allele frequencies $p_1$ and $p_2 = 1 - p_1$ in a population: $\alpha = a + d(p_2 - p_1)$. The heritability depends on both genetic and nongenetic variation in the population (see also Hansen *et al.* 2011).

### *Multivariate genotype–phenotype association*

For multiple loci and multiple phenotypic variables, we seek the combined effect of the alleles $(\mathbf{x}_1, \ldots, \mathbf{x}_p) = \mathbf{X}$ on a phenotype composed of the measured variables $(\mathbf{y}_1, \ldots, \mathbf{y}_q) = \mathbf{Y}$. In a cross of two inbred lines, each locus is represented by one variable for the additive genotype scores and one for the dominance scores. In natural populations, where each locus can have multiple alleles, each allele at each locus is represented by a separate variable containing the number of this allele (0, 1, 2) at the locus and one variable for the dominance scores. In the following notation, the variables $\mathbf{x}_{\mathrm{i}}$, $\mathbf{y}_{\mathrm{i}}$ are again assumed to be mean centered. Let the effects of the alleles on the phenotype be denoted by the $p \times 1$ vector $\mathbf{a} = (a_1, a_2, \ldots, a_p)'$ and the weightings of the measured variables that determine the phenotype by the $q \times 1$ vector $\mathbf{b} = (b_1, b_2, \ldots, b_q)'$. Then the two latent variables are given by the linear combinations $\mathbf{Xa}$ and $\mathbf{Yb}$ (Figure 1). The coefficient vectors $\mathbf{a}$, $\mathbf{b}$ are chosen to maximize the association between the corresponding latent variables,

$$\max_{\mathbf{a},\mathbf{b}} A(\mathbf{Xa}, \mathbf{Yb}),$$

where $A$ represents one of the above association functions (genetic effect, genetic variance, heritability) between the genetic and phenotypic latent variables. In addition to this pair of latent variables, there might be further pairs of latent variables with effects $\mathbf{a}_i$ and $\mathbf{b}_i$, independent of the previous ones, that together account for the observed genotype–phenotype association.

### *Maximizing genotype–phenotype association via SVD*

The association functions (A1)–(A4) extend naturally from a single locus and a single trait to a linear combination of loci and a linear combination of phenotypic variables.

***Genetic effect (i):*** Under the constraint $\mathbf{a}'\mathbf{a} = \mathbf{b}'\mathbf{b} = 1$, the regression slope $\mathrm{Cov}(\mathbf{Xa}, \mathbf{Yb})/\mathrm{Var}(\mathbf{Xa})$ of $\mathbf{Yb}$ on $\mathbf{Xa}$, conditional on all other linear combinations of $\mathbf{X}$, is maximized by the first pair of singular vectors $\mathbf{a} = \mathbf{u}_1, \mathbf{b} = \mathbf{v}_1$ of the matrix of multiple-regression coefficients $\mathbf{F} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$. If $\mathbf{X}$ contains the $p$ additive scores only, the singular value $\lambda_1$ represents the average effect associated with the allele combination and the phenotype specified by the singular vectors. Whereas if $\mathbf{X}$ comprises both additive and dominance scores ($2p$ in total), the singular value is the sum of additive and dominance effects.

To prove this, consider the matrix of regression coefficients for the linear combinations $\mathbf{XA}$ and $\mathbf{YB}$, where $\mathbf{A}$ and $\mathbf{B}$ are orthonormal matrices containing the vectors $\mathbf{a}_i$ and $\mathbf{b}_i$, respectively:

$$\begin{aligned}\Big((\mathbf{XA})'\mathbf{XA}\Big)^{-1}(\mathbf{XA})'\mathbf{YB} &= \mathbf{A}'\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{YB} \\ &= \mathbf{A}'\mathbf{FB}.\end{aligned} \tag{A5}$$

(Note that this equation holds only under the constraint that the vectors $\mathbf{a}_i$ are mutually orthogonal so that the matrix $\mathbf{A}$ is orthonormal and $\mathbf{A}' = \mathbf{A}^{-1}$.) The right part of Equation A5 is a classic singular value problem (*e.g.*, Mardia *et al.* 1979). If $\mathbf{A}$ is equal to the matrix $\mathbf{U}$ of left singular vectors of $\mathbf{F}$ and $\mathbf{B}$ is equal to the matrix $\mathbf{V}$ of right singular vectors, then $\mathbf{A}'\mathbf{FB} = \boldsymbol{\Lambda}$ is the diagonal matrix of singular values of $\mathbf{F}$. These singular values are equal the regression slopes of the phenotypic latent variables on the corresponding genetic latent variables. The pair of singular vectors associated with the largest singular value determines the pair of genetic and phenotypic latent variables with maximal regression slope (*i.e.*, maximal genetic effect), $\lambda_1$.

***Genetic variance (ii):*** The variance in the phenotype $\mathbf{Yb}$ explained by the allele combination $\mathbf{Xa}$ is given by $\mathrm{Cov}(\mathbf{Xa}, \mathbf{Yb})^2/\mathrm{Var}(\mathbf{Xa})$ (compare Equation A3). It is maximized by the vectors $\mathbf{a} = (\mathbf{X}'\mathbf{X})^{-1/2}\mathbf{u}_1$ and $\mathbf{b} = \mathbf{v}_1$, where $\mathbf{u}_1$ and $\mathbf{v}_1$ are the first left and right singular vectors of $\mathbf{G} = (\mathbf{X}'\mathbf{X})^{-1/2}\mathbf{X}'\mathbf{Y}$. The matrix $\mathbf{G}$ is equal to the covariance matrix between $\mathbf{Y}$ and $\mathbf{X}$ after $\mathbf{X}$ is transformed to a spherical distribution. If $\mathbf{X}$ contains the additive scores only, $\lambda_1^2/(n-1)$ is the additive genetic variance associated with the corresponding allele combination and phenotype, whereas if $\mathbf{X}$ contains both additive and dominance scores, it represents the total genetic variance.

To show this, express the maximization of $\mathrm{Cov}(\mathbf{Xa}, \mathbf{Yb})^2/\mathrm{Var}(\mathbf{Xa})$ as the maximization of $\mathrm{Cov}(\mathbf{Xa}, \mathbf{Yb})^2 = (\mathbf{a}'\mathbf{X}'\mathbf{Yb}/(n-1))^2$ subject to $\mathrm{Var}(\mathbf{Xa}) = \mathbf{a}'\mathbf{X}'\mathbf{Xa}/(n-1) = 1$. Write $\alpha = (\mathbf{X}'\mathbf{X})^{1/2}\mathbf{a}$; then the maximization is of $(\alpha'(\mathbf{X}'\mathbf{X})^{-1/2}\mathbf{X}'\mathbf{Yb}/(n-1)^2)$, subject to $\alpha'\alpha/(n-1) = 1$. This is well known to be solved by the first pair of singular vectors $\mathbf{u}_1, \mathbf{v}_1$ of $(\mathbf{X}'\mathbf{X})^{-1/2}\mathbf{X}'\mathbf{Y}$, where $\mathbf{a} = (\mathbf{X}'\mathbf{X})^{-1/2}\mathbf{u}_1$ and $\mathbf{b} = \mathbf{v}_1$, with the maximal variance equal to $\lambda_1^2/(n-1)$ (*cf.* Mardia *et al.* 1979; p. 284).

***Heritability (iii):*** The squared correlation coefficient $\mathrm{Cor}(\mathbf{Xa}, \mathbf{Yb})^2$ is maximized by the vectors $\mathbf{a} = (\mathbf{X}'\mathbf{X})^{-1/2}\mathbf{u}_1$ and $\mathbf{b} = (\mathbf{Y}'\mathbf{Y})^{-1/2}\mathbf{v}_1$, where $\mathbf{u}_1$ and $\mathbf{v}_1$ are the left and right singular vectors of the matrix $\mathbf{H} = (\mathbf{X}'\mathbf{X})^{-1/2}\mathbf{X}'\mathbf{Y}(\mathbf{Y}'\mathbf{Y})^{-1/2}$, the covariance matrix between the spherized $\mathbf{Y}$ and spherized $\mathbf{X}$. Depending on whether $\mathbf{X}$ contains the additive scores or both additive and dominance scores, the squared singular value $\lambda_1^2$ (squared canonical correlation) is the narrow-sense or broad-sense heritability, respectively. The proof follows by extending the one of approach ii; it equals the standard proof of canonical correlation analysis (*e.g.*, in Mardia *et al.* 1979).

For each of the three approaches, further dimensions (pairs of latent variables) can be extracted by the subsequent pairs of singular vectors of the corresponding association matrix:

$$\text{(i) } \mathbf{a}_\mathrm{i} = \mathbf{u}_\mathrm{i}, \mathbf{b}_\mathrm{i} = \mathbf{v}_\mathrm{i};$$

$$\text{(ii) } \mathbf{a}_\mathrm{i} = \left(\mathbf{X}'\mathbf{X}\right)^{-1/2}\mathbf{u}_\mathrm{i}, \mathbf{b}_\mathrm{i} = \mathbf{v}_\mathrm{i};$$

$$\text{(iii) } \mathbf{a}_\mathrm{i} = \left(\mathbf{X}'\mathbf{X}\right)^{-1/2}\mathbf{u}_\mathrm{i}, \mathbf{b}_\mathrm{i} = \left(\mathbf{Y}'\mathbf{Y}\right)^{-1/2}\mathbf{v}_\mathrm{i}.$$

In approach i, the genetic effects (the vectors $\mathbf{a}_i$) are orthogonal because they are the singular vectors of a real matrix. In the maximizations of (ii) genetic variance and (iii) heritability, by contrast, the genetic latent variables are uncorrelated. This can be shown by expressing the covariance matrix of the genetic latent variables as $\mathbf{U}'(\mathbf{X}'\mathbf{X})^{-1/2}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1/2}\mathbf{U} = \mathbf{U}'\mathbf{U} = \mathbf{I}$, where the orthonormal matrix $\mathbf{U}$ contains the singular vectors $\mathbf{u}_i$, and $\mathbf{I}$ is the identity matrix. The phenotypic effects are orthogonal in approaches i and ii, whereas the phenotypic latent variables are uncorrelated in approach iii.

In all three approaches, the corresponding association is maximized conditional on all other linear combinations of alleles, including the other dimensions $\mathbf{a}_i$. This implies that $b_{\mathbf{Xa}_i,\mathbf{Yb}_i} = b_{\mathbf{Xa}_i,\mathbf{Yb}_i.\mathbf{Xc}}$, where $b_{\mathbf{Xa}_i,\mathbf{Yb}_i}$ is the regression slope of the $i$th phenotypic latent variable on the $i$th genetic latent variable and $b_{\mathbf{Xa}_i,\mathbf{Yb}_i.\mathbf{Xc}}$ is the regression slope conditional on another linear combination $\mathbf{Xc}$, for any vector $\mathbf{c} \neq \mathbf{a}_i$. The matrix of regression coefficients of the phenotypic latent variables $\mathbf{YB}$ on the genetic latent variables $\mathbf{XA}$ thus is diagonal.

The proof of these properties for approach i follows from the fact that the matrix of regression coefficients for the latent variables, $\mathbf{XA}$ and $\mathbf{YB}$, can be written as $\mathbf{A}'\mathbf{FB} = \mathbf{\Lambda}$ (see Equation A5), where $\mathbf{\Lambda}$ is the diagonal matrix of singular values of $\mathbf{F}$. For approach ii, $\mathbf{A} = (\mathbf{X}'\mathbf{X})^{-1/2}\mathbf{U}$ and $\mathbf{B} = \mathbf{V}$, where $\mathbf{U}$ and $\mathbf{V}$ are the orthonormal matrices of left and right singular vectors of $\mathbf{G}$. The matrix of regression coefficients for the latent variables can thus be written as

$$\left(\mathbf{U}'\left(\mathbf{X}'\mathbf{X}\right)^{-1/2}\mathbf{X}'\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1/2}\mathbf{U}\right)^{-1}\mathbf{U}'\left(\mathbf{X}'\mathbf{X}\right)^{-1/2}\mathbf{X}'\mathbf{YV} = \mathbf{U}'\left(\mathbf{X}'\mathbf{X}\right)^{-1/2}\mathbf{X}'\mathbf{YV} = \mathbf{\Lambda}, \qquad \text{(A6)}$$

where the diagonal matrix $\mathbf{\Lambda}$ now contains the singular values of $\mathbf{G}$. The proof for approach iii can be constructed similarly. Furthermore, in approaches ii and iii the genetic latent variables are correlated only with the corresponding phenotypic latent variable but not with any of the other latent variables: $\mathrm{Cov}(\mathbf{Xa}_i, \mathbf{Yb}_j) = 0$ for $i \neq j$. This can be shown by expressing the cross-covariance matrix of the latent variables as $\mathbf{U}'(\mathbf{X}'\mathbf{X})^{-1/2}\mathbf{X}'\mathbf{YV}$ in approach ii and as $\mathbf{U}'(\mathbf{X}'\mathbf{X})^{-1/2}\mathbf{X}'\mathbf{Y}(\mathbf{Y}'\mathbf{Y})^{-1/2}\mathbf{V}$ in approach iii, which both are diagonal.

### *Partial least-squares analysis*

The fourth approach, PLS, maximizes the covariance between the two linear combinations $\mathrm{Cov}(\mathbf{Xa}, \mathbf{Yb})$. The unit vectors $\mathbf{a}, \mathbf{b}$ maximizing this covariance can be computed as the first pair of singular vectors $\mathbf{u}_1, \mathbf{v}_1$ of the cross-covariance matrix $\mathbf{X}'\mathbf{Y}$. The proof of this classic singular value problem is given, *e.g.*, in Mardia *et al.* (1979); see also Sampson *et al.* (1989). Subsequent pairs of singular vectors yield further genetic and phenotypic dimensions $\mathbf{a}_i, \mathbf{b}_i$ that are mutually orthogonal. Furthermore, $\mathrm{Cov}(\mathbf{Xa}_i, \mathbf{Yb}_j) = 0$ for $i \neq j$ and $\lambda_i/(n-1)$ for $i = j$, where $\lambda_i$ is the $i$th singular value of $\mathbf{X}'\mathbf{Y}$.

The scaled genetic coefficients (the elements of the scaled singular vectors $\lambda_i\mathbf{a}_i/(n-1)$ are equal to the covariances between the corresponding locus and the phenotypic latent variable, without conditioning on the other loci as in approaches i–iii. When the genetic variables are standardized to unit variance through division by their standard deviation, the squared scaled genetic coefficients equal the explained variance of the phenotypic latent variable owing to the corresponding locus considered separately, *i.e.*, without conditioning on the other loci (compare Equation A3).

## Appendix B: Properties

### *Geometric properties*

The maximizations of genetic effect and of covariance in approaches i and iv require a constraint on the length of $\mathbf{a}$ and $\mathbf{b}$. Because they are the singular vectors $\mathbf{u}_i, \mathbf{v}_i$, they are computed to have a 2-norm of 1. When the variables can be equipped with a meaningful Euclidean metric, this constraint translates into an interpretable notion of total genetic and phenotypic effects. For the genetic variables, this choice of constraint is not particularly obvious as it implies that an allele with an effect of 1 is equivalent in magnitude to two alleles with effects of $\sqrt{1/2}$ each. It may seem more intuitive that two alleles with effects of $1/2$ are equivalent to a single allele with an effect of 1. This latter choice would impose a constraint on the sum of the absolute values of the elements of $\mathbf{a}$ (the 1-norm), not on the sum of the squared elements. Regression approaches with constraints other than the 2-norm have been proposed [*e.g.*, the Lasso technique (Tibshirani 1996)], but this generalization goes beyond the scope of the present article.

The invariance to the length of $\mathbf{a}$ in approach ii can also be seen from Equation A3, which expresses the additive genetic variance of a single trait as its squared covariance with $\mathbf{x}_s$, the additive genotype scores scaled to unit variance. The scale of $\mathbf{x}$ is removed through dividing by its standard deviation. In the multivariate context, the maximal genetic variance, $\mathrm{Cov}(\mathbf{Xa}, \mathbf{Yb})^2/\mathrm{Var}(\mathbf{Xa})$, can be found by maximizing $\mathrm{Cov}(\mathbf{X}_S\mathbf{a}, \mathbf{Yb})^2$, where $\mathbf{X}_S$ is the matrix of genetic variables transformed so that every linear combination has unit variance: $\mathrm{Var}(\mathbf{X}_S\mathbf{c}) = 1$ for any unit vector $\mathbf{c}$. This transformation is achieved by multiplying $\mathbf{X}$ by the inverse square root of its covariance matrix: $\mathbf{X}_S = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1/2}$. The vectors $\mathbf{a}$ and $\mathbf{b}$ can thus be found by the singular vectors $\mathbf{u}_1, \mathbf{v}_1$ of $(\mathbf{X}'\mathbf{X})^{-1/2}\mathbf{X}'\mathbf{Y}$, after $\mathbf{u}_1$ has been transformed back into the original coordinate system. The singular values of this matrix (the genetic variances) as well as the right singular vectors determining the phenotypes are invariant to affine transformations of $\mathbf{X}$. This can be shown when considering that the right singular vectors and squared singular values of $\mathbf{G}$ are the eigenvectors and eigenvalues of $\mathbf{G}'\mathbf{G}$ (*cf.* Equation B1). Let $\mathbf{X}^* = \mathbf{XT}$ be a linear transformation of $\mathbf{X}$, where $\mathbf{T}$ is a full-rank $p \times p$ matrix, and $\mathbf{G}^* = ((\mathbf{X}^*)'\mathbf{X}^*)^{-1/2}(\mathbf{X}^*)'\mathbf{Y}$ :

$$\begin{aligned}(\mathbf{G}*)'\mathbf{G}^* &= \mathbf{Y}'\mathbf{XT}((\mathbf{XT})'\mathbf{XT})^{-1/2}((\mathbf{XT})'\mathbf{XT})^{-1/2}(\mathbf{XT})'\mathbf{Y}\\ &= \mathbf{Y}'\mathbf{XT}\left(\mathbf{T}'\mathbf{X}'\mathbf{XT}\right)^{-1}\mathbf{T}'\mathbf{X}'\mathbf{Y}\\ &= \mathbf{Y}'\mathbf{XTT}^{-1}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\left(\mathbf{T}'\right)^{-1}\mathbf{T}'\mathbf{X}'\mathbf{Y}\\ &= \mathbf{Y}'\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y}\\ &= \mathbf{G}'\mathbf{G}.\end{aligned}$$

It follows from this property that the maximal genetic variances (singular values of **G**) as well as the phenotypes that show these maximal genetic variances (right singular vectors) remain unchanged by linear transformations of the genetic variables; thus they also do not depend on the variances and covariances of the genetic variables, that is, on genetic variance and linkage disequilibrium. The same property holds for approach iii. Approaches i and iv are not invariant to transformations of **X**, implying that the genetic variances and covariances need to be interpretable for computing the maximal genetic effects. Only approach iii is invariant to affine transformation of the phenotypic variables **Y**; all other approaches require meaningful phenotypic variances and covariances as well as commensurate units.

### *Computational properties*

For typical genetic data, approaches i–iii are not computable by the presented least-squares methods because of collinearities between loci or because $p > n$. The covariance matrix $\mathbf{X}'\mathbf{X}$ is singular and its inverse or inverse square root cannot be computed without the use of a pseudoinverse or of regularization techniques. The Moore–Penrose pseudoinverse of a matrix **M** is $\mathbf{M}^+ = \mathbf{Q}\mathbf{\Lambda}^+\mathbf{Q}'$, where **Q** is the matrix of eigenvectors of **M** and $\mathbf{\Lambda}^+$ is a diagonal matrix with the reciprocal of the $m$ largest nonzero eigenvalues in the diagonal. The remaining $p - m$ eigenvalues are set to 0. This is equivalent to reducing the data to the first $m$ principal components for the inversion, discarding all subsequent principal components with small or zero variance. The partial coefficients resulting from such an approach are not conditional on all other variables, but only on the major patterns of multivariate variation (discarding rare alleles). In the simplest form of Tikhonov regularization, also referred to as ridge regression, $(\mathbf{X}'\mathbf{X})^{-1}$ is replaced by $(\mathbf{X}'\mathbf{X} + \gamma\mathbf{I})^{-1}$, where **I** is the identity matrix and $\gamma$ is a positive real. The larger $\gamma$ is, the more are components with low variance downweighted in the matrix inverse; *i.e.*, rare alleles or less variable allele combinations are downweighted relative to more variable alleles or allele combinations.

The results of approaches i–iii depend on the number of selected components or on $\gamma$ and require a careful decision. Typically, after adding the first few components that cover the relevant signals, adding further components has little effect, until the number of components becomes too large and the increasing noise leads to unstable results. Adding further components may still increase the explained phenotypic variance, but this is due to overfitting; the genetic coefficients may not be interpretable. Exploring different numbers of components underlying the pseudoinverse (or different values of $\gamma$) thus often leads to a range of stable components (or a stable range of $\gamma$) that lead to similar and equally interpretable results. A cross-validation approach can help to find the optimal number of principal components or the optimal $\gamma$ for the data set. Approach iv—the partial least-squares analysis—involves no matrix inverse and can also be computed for collinear loci and if $p > n$, overfitting is less a problem in this approach (Martens and Naes 1989). It is thus useful to compare the results of approaches i–iii to that of approach iv. Bayesian approaches and numerous other penalized methods offer promising alternatives to the presented least-squares methods (Meuwissen *et al.* 2001; Lopes and West 2004; de Los Campos *et al.* 2013; Zhu *et al.* 2014).

If the number of genetic variables $p$ or the number of phenotypic variables $q$ is very large, the singular value decomposition of the association matrix can be computationally demanding. Here one can make use of the property that the left singular vectors $\mathbf{u}_i$ of a matrix **M** are equal to the eigenvectors of $\mathbf{MM}'$ and the right singular vectors $\mathbf{v}_i$ are the eigenvectors of $\mathbf{M}'\mathbf{M}$. Thus, if $p \ll q$ or $q \ll p$, one can compute either $\mathbf{u}_i$ or $\mathbf{v}_i$ as the eigenvectors of the smaller matrix product. Since $\mathbf{M}'\mathbf{u}_i = \lambda_i\mathbf{v}_i$ and $\mathbf{Mv}_i = \lambda_i\mathbf{u}_i$, the other singular vectors can be obtained by premultiplication with **M**. In approach ii, for instance, the vectors $\mathbf{b}_i$ are given by the eigenvectors of

$$\mathbf{G}'\mathbf{G} = \mathbf{Y}'\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y}. \tag{B1}$$

Note that the eigenvalues of (B1) are equal to the squared singular values of **G**.

If both $p$ and $q$ are very large and if only the first few dimensions need to be computed, the singular vectors can be computed more effectively via an iterative approach. Start with any $p \times 1$ vector $\mathbf{u}_1$ and estimate $\mathbf{v}_1$ as $\mathbf{M}'\mathbf{u}_1$, scaled to unit vector length. In the next step, $\mathbf{u}_1$ is estimated as $\mathbf{Mv}_1$, again scaled to unit vector length. These steps are repeated until convergence, which is usually reached fast. The singular value equals $\lambda_1 = \left\|\mathbf{M}'\mathbf{u}_1\right\| = \|\mathbf{Mv}_1\|$. To compute the next pair of singular vectors $\mathbf{u}_2, \mathbf{v}_2$, let $\mathbf{M}^{(1)} = \mathbf{M} - \lambda_1\mathbf{u}_1\mathbf{v}_1'$ and repeat the iterative approach with $\mathbf{M}^{(1)}$ instead of **M**, and similarly for subsequent dimensions.

### *Generalized least squares*

In a sample with a family structure, the presented least-squares estimates are unbiased but the standard errors of the parameters may be inflated. This can be addressed by generalized least squares. Let the $n \times n$ matrix $\mathbf{\Omega}$ contain measures of expected or realized genetic relatedness between pairs of individuals (*e.g.*, Hayes *et al.* 2009). Then the maximization in the four approaches is based on the following matrices:

$$\text{(i)}\left(\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{Y},$$

$$\text{(ii)}\left(\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X}\right)^{-1/2}\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{Y},$$

$$\text{(iii)}\left(\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X}\right)^{-1/2}\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{Y}\left(\mathbf{Y}'\boldsymbol{\Omega}^{-1}\mathbf{Y}\right)^{-1/2},$$

$$\text{(iv)}\,\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{Y}.$$

### *The genetic variance–covariance structure*

Because in approach ii—the maximization of genetic variance—the vectors $\mathbf{b}_i$ constitute an orthonormal basis of the phenotype space with associated genetic variances $\lambda_i^2/(n-1)$, the genetic variance–covariance matrix $\mathbf{S}_\mathrm{G}$ of the measured phenotypic variables can be computed as $\mathbf{B}\boldsymbol{\Lambda}^2\mathbf{B}'/(n-1)$, where $\mathbf{B}$ contains the phenotypic coefficient vectors $\mathbf{b}_i$ (the left singular vectors of $\mathbf{G}$), and $\boldsymbol{\Lambda}^2$ is a diagonal matrix of the squared singular values of $\mathbf{G}$. Because of (B1), the genetic variance–covariance matrix can also be computed directly as

$$\mathbf{S}_\mathrm{G} = \mathbf{Y}'\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y}\Big/(n-1). \tag{B2}$$

This can also be seen when considering the phenotypic predictions from the genetic variables, $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$, of which $\mathbf{S}_\mathrm{G}$ is the variance–covariance matrix. Note that when $\mathbf{X}$ contains the additive genotype scores only, (B2) is the additive genetic variance–covariance matrix, whereas if $\mathbf{X}$ contains additive and dominance scores, (B2) is the total genetic covariance matrix.

## Appendix C: Alternative Analyses

Here we present the results of the other three approaches applied to the same mouse data as in the main text. All three approaches lead to similar scree plots as in approach ii: The first dimension clearly dominates the genotype–phenotype relationship (Figure C1). In all approaches, the first dimension represents limb length (Figure C2) and, hence, also the genetic coefficients are highly consistent (not shown).

Approach iii, the maximization of heritability, is most unstable and requires a dimension reduction of the phenotypic variables. In this analysis we used the first four principal components of the 11 measurements. By contrast, the partial least-squares analysis in approach iv does not require variable reduction or regularization for the genetic and phenotypic variables.

The second dimension largely reflects body and organ weight in all four approaches (Figure C2), and also the genetic coefficients are very similar (not shown). The third dimension differs among the approaches. Only in approach iv the third dimension reflects a contrast between distal and proximal long bones as in approach ii.

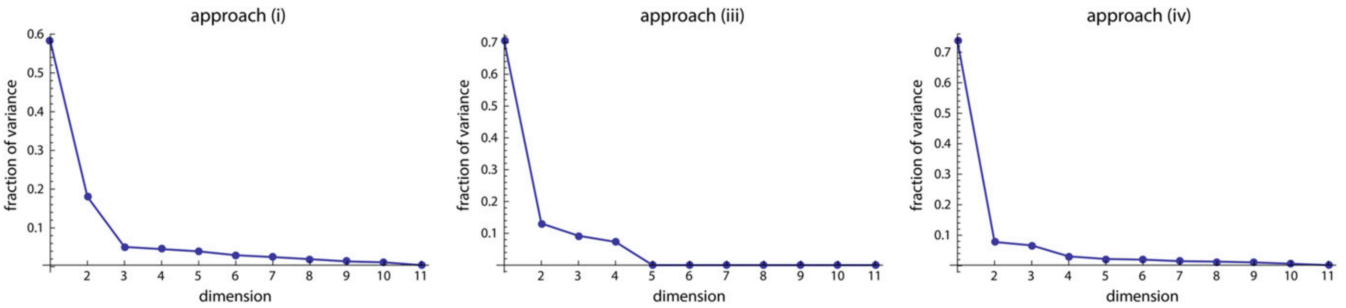


**Figure C1** Scree plots resulting from approaches i, iii, and iv, which maximize genetic effect, genetic variance, and the covariance between genetic and phenotypic latent variables, respectively.

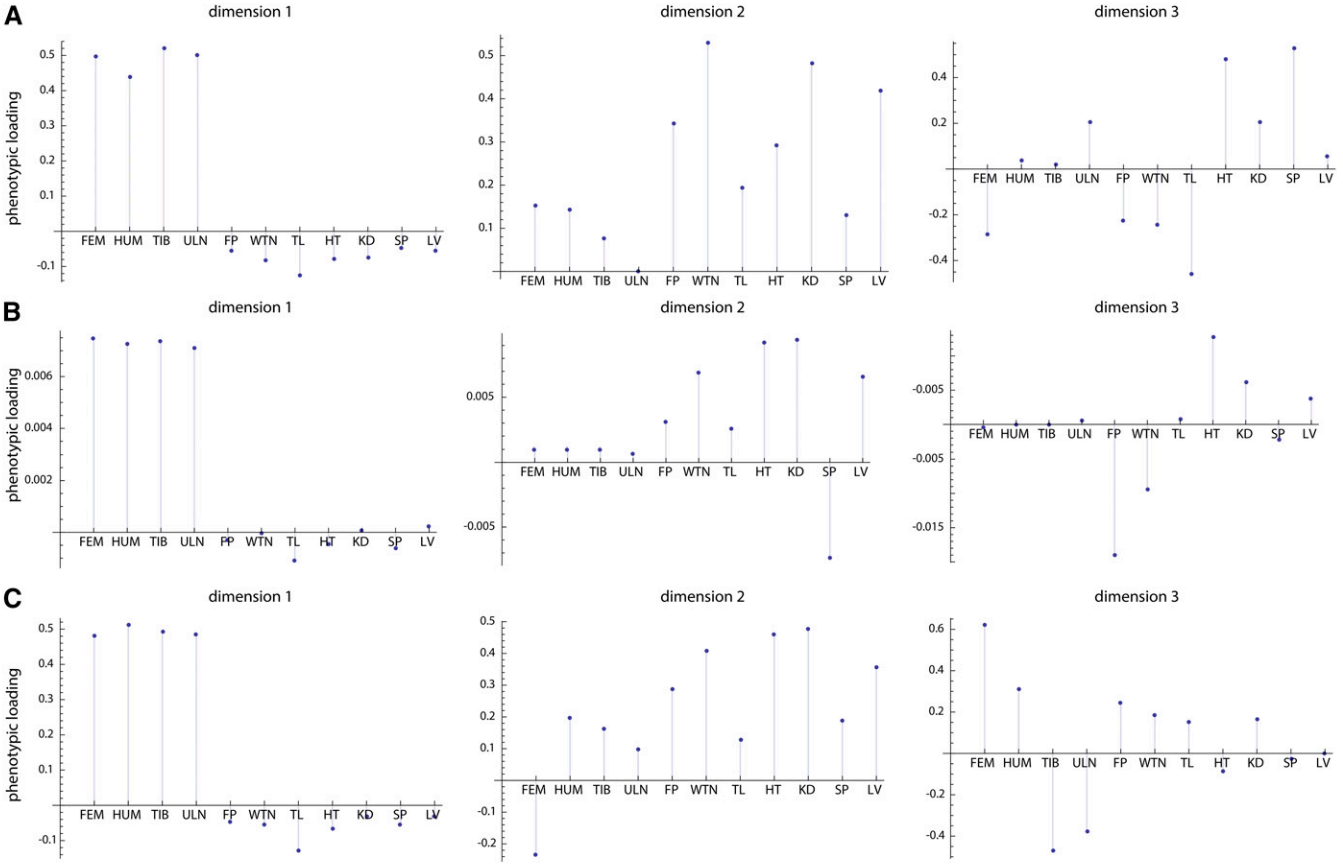


**Figure C2** Phenotypic coefficients resulting from approaches i, iii, and iv applied to the 353 loci on the 19 chromosomes. The phenotypic measurements are the lengths of the femur, the humerus, the tibia, and the ulna (FEM, HUM, TIB, ULN); weight of the fat pad (FP); body weight (WTN); tail length (TL); and the weights of the heart, the kidneys, the spleen, and the liver (HT, KD, SP, LV).

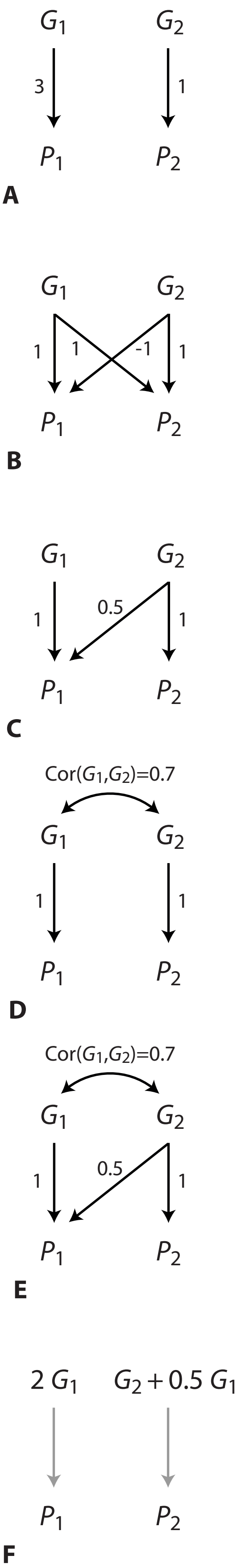
GENETICS

**Supporting Information**
www.genetics.org/lookup/suppl/doi:10.1534/genetics.115.181339/-/DC1

# Multivariate Analysis of Genotype–Phenotype Association

**Philipp Mitteroecker, James M. Cheverud, and Mihaela Pavlicev**

DOI: 10.1534/genetics.115.181339



| | approach (I) | approach (II) | approach (III) | approach (IV) |
|---|---|---|---|---|
| A | $\mathbf{a}_1$=(1, 0)<br>$\mathbf{a}_2$=(0, 1)<br>$\mathbf{b}_1$=(1, 0)<br>$\mathbf{b}_2$=(0, 1)<br>$\lambda_1$=3 $\lambda_2$=1 | $\mathbf{a}_1$=(0.10, 0)<br>$\mathbf{a}_2$=(0, 0.10)<br>$\mathbf{b}_1$=(1, 0)<br>$\mathbf{b}_2$=(0, 1)<br>$\lambda_1^2/(n\text{-}1)$=9 $\lambda_2^2/(n\text{-}1)$=1 | $\mathbf{a}_1$=(0.10, 0)<br>$\mathbf{a}_2$=(0, 0.10)<br>$\mathbf{b}_1$=(0.03, 0)<br>$\mathbf{b}_2$=(0, 0,10)<br>$\lambda_1^2$=1 $\lambda_2^2$=1 | $\mathbf{a}_1$=(1, 0)<br>$\mathbf{a}_2$=(0, 1)<br>$\mathbf{b}_1$=(1, 0)<br>$\mathbf{b}_2$=(0, 1)<br>$\lambda_1/(n\text{-}1)$=3 $\lambda_2/(n\text{-}1)$=1 |
| B | $\mathbf{a}_1$=(1, 0)<br>$\mathbf{a}_2$=(0, 1)<br>$\mathbf{b}_1$=(0.71, 0.71)<br>$\mathbf{b}_2$=(-0.71, 0.71)<br>$\lambda_1$=1.41 $\lambda_2$=1.41 | $\mathbf{a}_1$=(0.10, 0)<br>$\mathbf{a}_2$=(0, 0.10)<br>$\mathbf{b}_1$=(0.71, 0.71)<br>$\mathbf{b}_2$=(-0.71, 0.71)<br>$\lambda_1^2/(n\text{-}1)$=2 $\lambda_2^2/(n\text{-}1)$=2 | $\mathbf{a}_1$=(0.10, 0)<br>$\mathbf{a}_2$=(0, 0.10)<br>$\mathbf{b}_1$=(0.05, 0.05)<br>$\mathbf{b}_2$=(-0.05, 0.05)<br>$\lambda_1^2$=1 $\lambda_2^2$=1 | $\mathbf{a}_1$=(1, 0)<br>$\mathbf{a}_2$=(0, 1)<br>$\mathbf{b}_1$=(0.71, 0.71)<br>$\mathbf{b}_2$=(-0.71, 0.71)<br>$\lambda_1/(n\text{-}1)$=1.41 $\lambda_2/(n\text{-}1)$=1.41 |
| C | $\mathbf{a}_1$=(0.62, 0.79)<br>$\mathbf{a}_2$=(0.79, -0.62)<br>$\mathbf{b}_1$=(0.79, 0.62)<br>$\mathbf{b}_2$=(0.62,-0.79)<br>$\lambda_1$=1.28 $\lambda_2$=0.78 | $\mathbf{a}_1$=(0.62, 0.79)<br>$\mathbf{a}_2$=(0.79, -0.62)<br>$\mathbf{b}_1$=(0.79, 0.62)<br>$\mathbf{b}_2$=(0.62,-0.79)<br>$\lambda_1^2/(n\text{-}1)$=1.64<br>$\lambda_2^2/(n\text{-}1)$=0.61 | $\mathbf{a}_1$=(0.10, 0)<br>$\mathbf{a}_2$=(0, 0.10)<br>$\mathbf{b}_1$=(-0.10, 0.05)<br>$\mathbf{b}_2$=(0, 0.10)<br>$\lambda_1^2$=1 $\lambda_2^2$=1 | $\mathbf{a}_1$=(0.62, 0.79)<br>$\mathbf{a}_2$=(0.79, -0.62)<br>$\mathbf{b}_1$=(0.79, 0.62)<br>$\mathbf{b}_2$=(0.62,-0.79)<br>$\lambda_1^2/(n\text{-}1)$=1.28 $\lambda_2^2/(n\text{-}1)$=0.78 |
| D | $\mathbf{a}_1$=(1, 0)<br>$\mathbf{a}_2$=(0, 1)<br>$\mathbf{b}_1$=(1, 0)<br>$\mathbf{b}_2$=(0, 1)<br>$\lambda_1$=1 $\lambda_2$=1 | $\mathbf{a}_1$=(0.05, 0.05)<br>$\mathbf{a}_2$=(-0.13, 0.13)<br>$\mathbf{b}_1$=(0.71, 0.71)<br>$\mathbf{b}_2$=(-0.71, 0.71)<br>$\lambda_1^2/(n\text{-}1)$=1.71<br>$\lambda_2^2/(n\text{-}1)$=0.29 | $\mathbf{a}_1$=(-0.13, 0.05)<br>$\mathbf{a}_2$=(-0.05, 0.13)<br>$\mathbf{b}_1$=(-0.13, 0.05)<br>$\mathbf{b}_2$=(-0.05, 0.13)<br>$\lambda_1^2$=1 $\lambda_2^2$=1 | $\mathbf{a}_1$=(0.71, 0.71)<br>$\mathbf{a}_2$=(-0.71, 0.71)<br>$\mathbf{b}_1$=(0.71, 0.71)<br>$\mathbf{b}_2$=(-0.71, 0.71)<br>$\lambda_1^2/(n\text{-}1)$=1.71 $\lambda_2^2/(n\text{-}1)$=0.29 |
| E | $\mathbf{a}_1$=(0.62, 0.79)<br>$\mathbf{a}_2$=(0.79, -0.62)<br>$\mathbf{b}_1$=(0.79, 0.62)<br>$\mathbf{b}_2$=(0.62,-0.79)<br>$\lambda_1$=1.28 $\lambda_2$=0.78 | $\mathbf{a}_1$=(0.05, 0.06)<br>$\mathbf{a}_2$=(-0.12, 0.13)<br>$\mathbf{b}_1$=(0.83, 0.56)<br>$\mathbf{b}_2$=(-0.56,0.83)<br>$\lambda_1^2/(n\text{-}1)$=2.78<br>$\lambda_2^2/(n\text{-}1)$=0.18 | $\mathbf{a}_1$=(-0.05, 0.13)<br>$\mathbf{a}_2$=(-0.13, 0.05)<br>$\mathbf{b}_1$=(-0.05, 0.16)<br>$\mathbf{b}_2$=(-0.13, 0.12)<br>$\lambda_1^2$=1 $\lambda_2^2$=1 | $\mathbf{a}_1$=(0.70, 0.72)<br>$\mathbf{a}_2$=(-0.72, 0.70)<br>$\mathbf{b}_1$=(0.83, 0.56)<br>$\mathbf{b}_2$=(-0.56, 0.83)<br>$\lambda_1^2/(n\text{-}1)$=2.18 $\lambda_2^2/(n\text{-}1)$=0.23 |
| F | $\mathbf{a}_1$=(-0.98, 0.18)<br>$\mathbf{a}_2$=(0.18, 0.98)<br>$\mathbf{b}_1$=(-0.96, 0.28)<br>$\mathbf{b}_2$=(0.28, 0.96)<br>$\lambda_1$=1.54 $\lambda_2$=0.98 | $\mathbf{a}_1$=(0.05, 0.00)<br>$\mathbf{a}_2$=(-0.03, 0.10)<br>$\mathbf{b}_1$=(1, 0)<br>$\mathbf{b}_2$=(0, 1)<br>$\lambda_1^2/(n\text{-}1)$=9 $\lambda_2^2/(n\text{-}1)$=1 | $\mathbf{a}_1$=(0.05, 0)<br>$\mathbf{a}_2$=(-0.03, 0.10)<br>$\mathbf{b}_1$=(0.03, 0)<br>$\mathbf{b}_2$=(0, 0,10)<br>$\lambda_1^2$=1 $\lambda_2^2$=1 | $\mathbf{a}_1$=(0.97, 0.25)<br>$\mathbf{a}_2$=(-0.25, 0.97)<br>$\mathbf{b}_1$=(0.99, 0.04)<br>$\mathbf{b}_2$=(-0.04, 0.99)<br>$\lambda_1/(n\text{-}1)$=6.20 $\lambda_2/(n\text{-}1)$=1.00 |

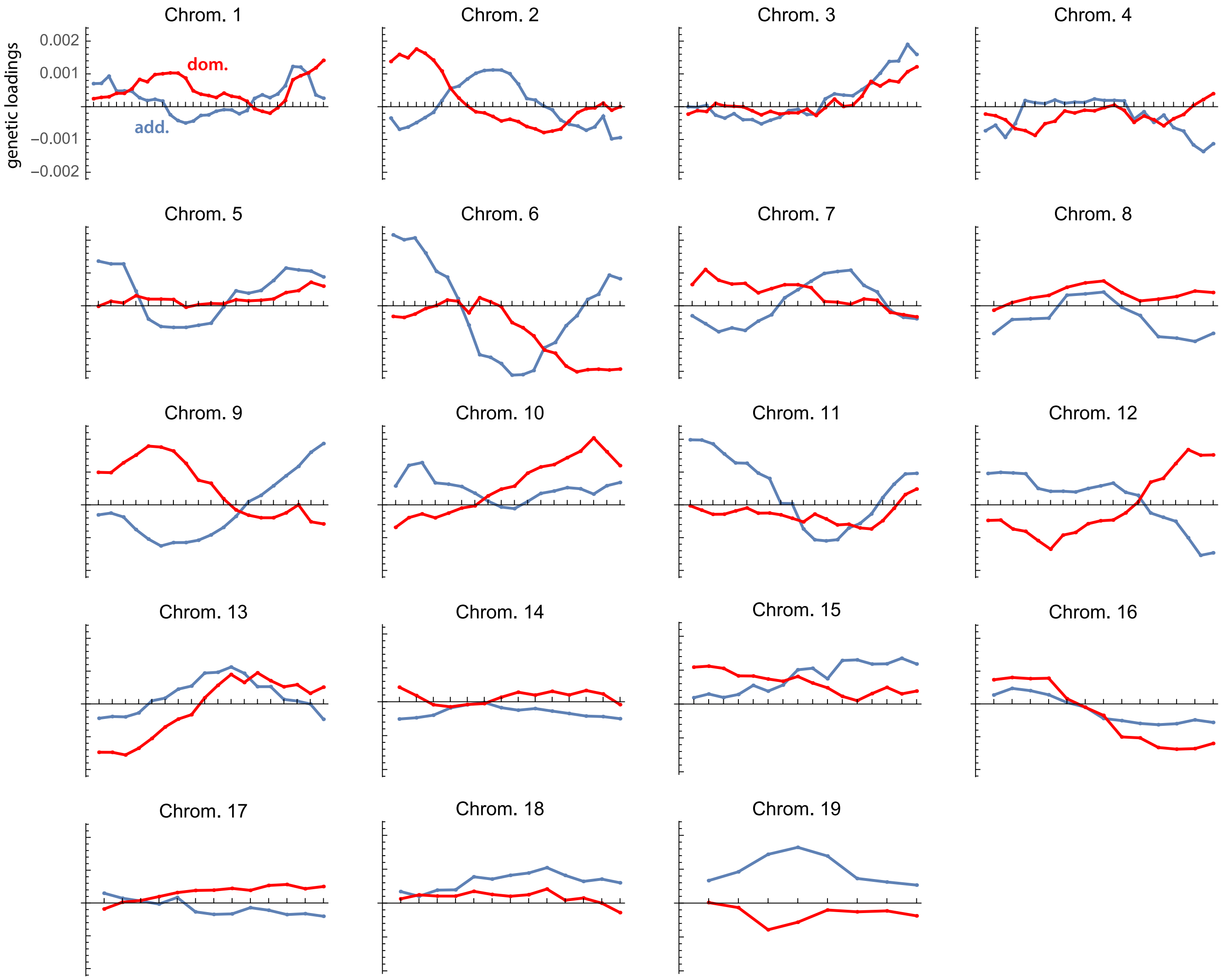
## Figure S1

Application of the four approaches to simulated data. Each dataset consists of two genetic variables (random variables scaled to unit variance), $G_1, G_2$, and two phenotypic variables, $P_1, P_2$, that are a linear function of the genetic variables without adding any noise. (A) The genetic variables are completely uncorrelated and affect one phenotypic variable each, where the effect of $G_1$ is three times as large as the effect of $G_2$. (B) The genetic variables are uncorrelated but have pleiotropic effects on both phenotypic variables. The effects of $G_1$ are orthogonal to that of $G_2$ (the vectors of genetic effects are (1,1) and (-1,1), respectively). (C) The genetic variables are uncorrelated and their effects are non-orthogonal: (1,0) versus (0.5,1). (D) The genetic variables have a correlation of 0.7 and equally affect the phenotypic variables. (E) The genetic variables are correlated and have non-orthogonal effects.

In (A) and (B), where the genetic variables are uncorrelated and have orthogonal effects, all four approaches recover the structure: the vectors $\mathbf{a}_i, \mathbf{b}_i$ correspond to the path coefficients in the models. In (C), where the genetic effects are non-orthogonal, approaches (I), (II), and (IV) lead to the same vectors $\mathbf{a}_i, \mathbf{b}_i$, which deviate from the path coefficients. The first dimension is a "common factor", representing the joint effect of both loci on both phenotypic variables, whereas the second dimension is a "contrast". The vectors resulting from approach (III) are more difficult to interpret. These results are similar to those of (E).

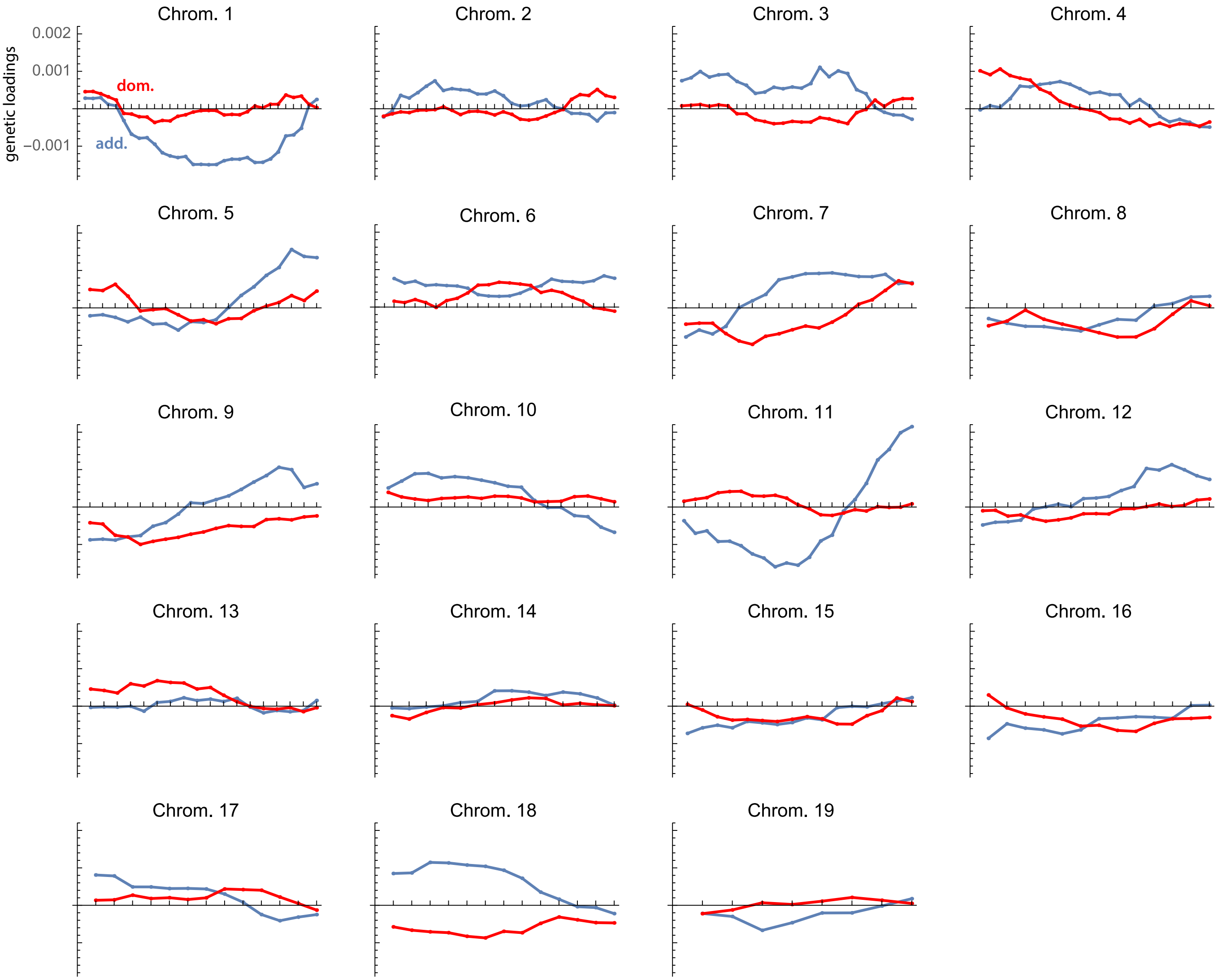
In (D) the genetic variables are correlated und have orthogonal effects. Approach (I) recovers the path coefficients, because $\mathbf{a}_1$ and $\mathbf{a}_2$ contain the genetic effects *independent* of the other loci/alleles. By contrast, approach (II) maximizes genetic variance, not genetic effect, and so the first dimension captures the joint effect of the two correlated loci on both phenotypic variables, which has almost six times as much genetic variance as the contrast between the loci and 1.7 times as much genetic variance as each locus considered separately. For these data, approach (IV) leads to the same results.

In (F) the same data is used as in (A) except that the genetic variables are linearly transformed: $G_1$ is multiple by 2, and $G_1/2$ is added to $G_2$, thus inducing a correlation between the two transformed genetic variables. The results of approaches (I) and (IV) differ between (A) and (F) because the genetic variances are changed by the transformation. For approaches (II) and (III), the vectors $\mathbf{b}_1, \mathbf{b}_2$ as well as the singular values $\lambda_1, \lambda_2$ are not affected by the linear transformation (see A.2.1).


Chrom. 1
Chrom. 2
Chrom. 3
Chrom. 4
genetic loadings
0.002
0.001
−0.001
−0.002
dom.
add.
Chrom. 5
Chrom. 6
Chrom. 7
Chrom. 8
Chrom. 9
Chrom. 10
Chrom. 11
Chrom. 12
Chrom. 13
Chrom. 14
Chrom. 15
Chrom. 16
Chrom. 17
Chrom. 18
Chrom. 19

## Figure S2

Genetic coefficients for the second dimension of the analysis of all 19 chromosomes. They are the elements of the vector $\mathbf{a}_2$ from approach (II) – the maximization of genetic variance – and represent the partial additive and dominance effects (blue and red lines) of all 353 loci on the corresponding phenotypic latent variable (Figure 3b).


Chrom. 1
Chrom. 2
Chrom. 3
Chrom. 4
genetic loadings
0.002
0.001
−0.001
dom.
add.
Chrom. 5
Chrom. 6
Chrom. 7
Chrom. 8
Chrom. 9
Chrom. 10
Chrom. 11
Chrom. 12
Chrom. 13
Chrom. 14
Chrom. 15
Chrom. 16
Chrom. 17
Chrom. 18
Chrom. 19

## Figure S3

Genetic coefficients for the third dimension of the analysis of all 19 chromosomes. They are the elements of the vector $\mathbf{a}_3$ from approach (II) and represent the partial additive and dominance effects (blue and red lines) of all 353 loci on the corresponding phenotypic latent variable (Figure 3c).