Orthogonal Estimates of Variances for Additive, Dominance, and Epistatic Effects in Populations

Zulma G. Vitezica,*,†,[1] Andrés Legarra,† Miguel A. Toro,‡ and Luis Varona§,**

*Institut National Polytechnique, École Nationale Supérieure Agronomique de Toulouse, Université de Toulouse, and †Institut National de la Recherche Agronomique, UMR 1388 Génétique, Physiologie et Systèmes d'Elevage, F-31326 Castanet-Tolosan, France, ‡Escuela Técnica Superior de Ingenieros Agrónomos, Universidad Politécnica de Madrid, 28040, Spain, §Departamento de Anatomía, Embriología y Genética, Universidad de Zaragoza, and **Instituto Agroalimentario de Aragón, 50013 Zaragoza, Spain

ABSTRACT Genomic prediction methods based on multiple markers have potential to include nonadditive effects in prediction and analysis of complex traits. However, most developments assume a Hardy–Weinberg equilibrium (HWE). Statistical approaches for genomic selection that account for dominance and epistasis in a general context, without assuming HWE (*e.g.*, crosses or homozygous lines), are therefore needed. Our method expands the natural and orthogonal interactions (NOIA) approach, which builds incidence matrices based on genotypic (not allelic) frequencies, to include genome-wide epistasis for an arbitrary number of interacting loci in a genomic evaluation context. This results in an orthogonal partition of the variances, which is not warranted otherwise. We also present the partition of variance as a function of genotypic values and frequencies following Cockerham's orthogonal contrast approach. Then we prove for the first time that, even not in HWE, the multiple-loci NOIA method is equivalent to construct epistatic genomic relationship matrices for higher-order interactions using Hadamard products of additive and dominant genomic orthogonal relationships. A standardization based on the trace of the relationship matrices is, however, needed. We illustrate these results with two simulated F_1 (not in HWE) populations, either in linkage equilibrium (LE), or in linkage disequilibrium (LD) and divergent selection, and pure biological dominant pairwise epistasis. In the LE case, correct and orthogonal estimates of variances were obtained using NOIA genomic relationships but not if relationships were constructed assuming HWE. For the LD simulation, differences were smaller, due to the smaller deviation of the F_1 from HWE. Wrongly assuming HWE to build genomic relationships and estimate variance components yields biased estimates, inflates the total genetic variance, and the estimates are not empirically orthogonal. The NOIA method to build genomic relationships, coupled with the use of Hadamard products for epistatic terms, allows the obtaining of correct estimates in populations either in HWE or not in HWE, and extends to any order of epistatic interactions.

KEYWORDS GenPred; shared data resource; genomic selection; genetic variance components; dominance; epistasis; genomic models; NOIA approach

DOMINANCE and epistasis may play an important role in the genetic determinism of complex traits of interest, such as human health or economic traits in livestock and crops. The existence of interactions within and across loci is supported by classic quantitative genetic studies, QTL mapping, and the wide application of crossbreeding as a breeding strategy. Nowadays, genomics provides tools to understand the effects of the genes and their interactions and to offer new directions for genetic improvement (Mäki-Tanila and Hill 2014). In quantitative genetics, the partition of the variance in statistical components due to additivity, dominance, and epistasis does not reflect the biological (or functional) effect of the genes but it is most useful for prediction, selection, and evolution (Huang and Mackay 2016).

In livestock populations, one of the main reasons why dominance or higher-order interaction terms have not been considered in genetic evaluations is that pedigree relationships are not informative enough. However, genomic selection methods are beginning to demonstrate their potential to include nonadditive effects in evaluation models. Inclusion of dominant or/and epistatic effects in genomic evaluation has been proposed by several authors (Toro and Varona 2010; Su *et al.* 2012; Vitezica *et al.* 2013; Nishio and Satoh 2014; Jiang and Reif 2015). Most epistatic models only consider

doi: https://doi.org/10.1534/genetics.116.199406
Manuscript received December 20, 2016; accepted for publication May 11, 2017; published Early Online May 18, 2017.
Supplemental material is available online at www.genetics.org/lookup/suppl/doi:10.1534/genetics.116.199406/-/DC1.
[1]Corresponding author: Institut National de la Recherche Agronomique, UMR 1388 Génétique, Physiologie et Systèmes d'Elevage, 24 Chemin de Borde Rouge, 31326 Castanet-Tolosan Cedex, France. E-mail: zulma.vitezica@ensat.fr

Table 1 The orthogonal contrast scales ($\boldsymbol{w}'$) for the F_2 population

Genotypes	AABB	AABb	AAbb	AaBB	AaBb	Aabb	aaBB	aaBb	aabb
Frequency (f)	$p_A^2p_B^2$	$p_A^2\ 2p_Bp_b$	$p_A^2p_b^2$	$2p_Ap_ap_B^2$	$4p_Ap_ap_Bp_b$	$2p_Ap_ap_b^2$	$p_a^2p_B^2$	$p_a^2 2p_Bp_b$	$p_a^2p_b^2$
f values	1/16	1/8	1/16	1/8	1/4	1/8	1/16	1/8	1/16
w_1	1	1	1	0	0	0	−1	−1	−1
w_2	−0.5	−0.5	−0.5	0.5	0.5	0.5	−0.5	−0.5	−0.5
w_3	1	0	−1	1	0	−1	1	0	−1
w_4	−0.5	0.5	−0.5	−0.5	0.5	−0.5	−0.5	0.5	−0.5

additive-by-additive epistatic interactions (*e.g.*, Su *et al.* 2012; Jiang and Reif 2015), although dominant-by-dominant and dominant-by-additive interactions may play a major role in heterosis and also in inbreeding and outbreeding depression (Lynch and Walsh 1998, p.223). Moreover, to date these models assume Hardy–Weinberg equilibrium (HWE) (*e.g.*, Vitezica *et al.* 2013). New statistical approaches to genomic selection that account for dominance and epistasis in a general context (*i.e.*, in populations not in HWE, like crosses or inbred populations) are needed both in animal and plant breeding and for QTL association studies.

Genomic evaluation models can fit marker or haplotypic additive genetic effects either explicitly, estimating the effect of each marker (Meuwissen *et al.* 2001), or implicitly through the so-called "genomic" relationship matrix (VanRaden 2008; Goddard 2009; Yang *et al.* 2010), which uses an equivalent model from which the marker effects can be inferred by backsolving. Dominance and higher-order interaction terms can also be modeled using the "genomic" relationship approach. Several approaches exist (Su *et al.* 2012; Muñoz *et al.* 2014; Jiang and Reif 2015) but none has addressed the issue of orthogonality of the model, additive and dominant components, and all possible interactions.

For plant and animal breeders and evolutionary geneticists, a meaningful partition of the variance is such that estimates can be interpreted in the classical terms, as variances of breeding values, dominant deviations, epistatic deviations, and so on (Hill *et al.* 2008). A nonorthogonal partition may lead to the erroneous conclusion that assortative mating and inclusion of dominance and/or epistasis can yield higher genetic gains as opposed to the consideration of additivity and random mating. For instance, Muñoz *et al.* (2014) concluded that dominance accounted for 39% of the total genetic variance when they used a nonorthogonal partition, *vs.* 24% when they used the orthogonal partition in Vitezica *et al.* (2013).

In this study, we develop a general procedure to estimate genomic relationship matrices for interaction terms of any order, expanding the natural and orthogonal interactions (NOIA) approach (Álvarez-Castro and Carlborg 2007) to the scope of the covariances between individuals. We present how to compute epistatic relationships from genotypes. Our results generalize the results of Cockerham (1954) to genomic models (something that had not been proven so far) and to any population, either in HWE or not. In particular, we show that the use of Hadamard products to obtain high-order relationships is correct, something that is frequently used (Su *et al.* 2012; Muñoz *et al.* 2014; Jiang and Reif 2015), but so far unproven for the general case. Two simulated examples, assuming linkage equilibrium (LE) or linkage disequilibrium (LD) and selection, are used to illustrate the approach.

Theory

First, we review the state-of-the-art approaches for orthogonal partitions of genotypic values, including the NOIA model which generalizes these results to any population. Then we extend the NOIA model to genomic prediction and show how this new model translates, for main effects, to the use of genomic-additive and dominant-relationship matrices (built according to the NOIA model) and, for interaction effects, into Hadamard products.

Orthogonal models of gene effects

The epistatic model proposed by Cockerham (1954), following Fisher (1918), partitioned the genetic variance caused by two genes in LE into orthogonal variance components. For two loci A and B (with alleles A,a and B,b) and allele frequencies equal to $p_A = p_a = p_B = p_b = 0.5$, Cockerham (1954) defined eight contrasts ($\boldsymbol{w}_{l=1\ldots8}$) which are orthogonal with respect to genotypic frequencies (f_{jk} is the joint frequency of genotypes j and k at each locus). The scales $\boldsymbol{w}_l$ (Table 1, see also Kao and Zeng 2002) satisfy two requirements:

$$\sum_{j,k} f_{jk} w_{jkl} = 0 \text{ for all } l,$$

and

$$\sum_{j,k} f_{jk} w_{jkl} w_{jkl'} = 0 \text{ for } l \neq l',$$

where $j(k)$ correspond to the genotype $AA(BB)$, $Aa(Bb)$, and $aa(bb)$ at the locus $A(B)$, and $\boldsymbol{w}_{l=1\ldots8}$ are the contrast scales for the nine genotypes j,k. These conditions allow the orthogonal estimation of genetic effects which are independent to each other in a statistical sense.

Note that $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$ ($\boldsymbol{w}_3$ and $\boldsymbol{w}_4$) are the additive and dominant contrasts for the locus A (for the locus B), and the epistatic components correspond to the relationships among these contrasts. Thus, $\boldsymbol{w}_5 = \boldsymbol{w}_1\boldsymbol{w}_3$, $\boldsymbol{w}_6 = \boldsymbol{w}_1\boldsymbol{w}_4$, $\boldsymbol{w}_7 = \boldsymbol{w}_2\boldsymbol{w}_3$, and $\boldsymbol{w}_8 = \boldsymbol{w}_2\boldsymbol{w}_4$ define the additive-by-additive, additive-by-dominant, dominant-by-additive, and dominant-by-dominant epistatic contrasts. Using the orthogonal contrast

scales and in matrix notation, the genotypic values for an individual assuming two loci j and k is:

$$g_{jk} = [1 \;\; \boldsymbol{w}_1 \; \boldsymbol{w}_2 \; \boldsymbol{w}_3 \; \boldsymbol{w}_4 \; \boldsymbol{w}_5 \; \boldsymbol{w}_6 \; \boldsymbol{w}_7 \; \boldsymbol{w}_8]\boldsymbol{\theta},$$

where the vector $\boldsymbol{\theta}' = [\mu, \alpha_j, d_j, \alpha_k, d_k, (\alpha\alpha), (\alpha d), (d\alpha), (dd)]'$ ($\boldsymbol{\theta}$ was called **E** by Álvarez-Castro and Carlborg 2007) contains the mean and statistical additive (α), dominant (d), and epistatic $[(\alpha\alpha), (\alpha d), (d\alpha), (dd)]$ effects.

Thus, in a general context, the genotypic values are split in additive (or breeding) values, dominance deviations (intralocus interactions), and epistatic deviations (interloci interactions) (Falconer and Mackay 1996). In the classical theory, the genetic effects are defined specifically in reference to a population. The reference population is an "ideal" random mating population where HWE and LE are assumed (Cockerham 1954; Kempthorne 1954). Under these idealized conditions, epistatic variance can be partitioned into orthogonal additive-by-additive, additive-by-dominant, *etc*., variance components.

The orthogonal property is very important and useful. The partition of the genetic effect is directly related to the partition of the genetic variance. The genetic variance can be partitioned into eight independent components and there is no genetic covariance among them. The additive effects contribute to the additive variance, the dominance effects contribute to the dominance variance, and so on. This is a desirable property in modeling because the estimation of one genetic (*i.e.*, additive) effect will not be affected by the presence or absence of other genetic effects in the model (*i.e.*, epistasis). Note that the orthogonal property in Cockerham's model applies only when allelic frequencies are exactly one-half (Zeng *et al.* 2005).

Later on, a model called general 2 allele model (G2A) (Zeng *et al.* 2005) was proposed for genetic effects regardless of the frequencies of the alleles at each locus. This model is a multi-loci, two-allele model which is orthogonal in populations under strict HWE and LE. The G2A model represents the Cockerham's model in a multiple regression context as

$$g = \mu + h_a\alpha + h_d d,$$

where $p = p_A$, $q = (1 - p_A)$, h_a, and h_d are defined as

$$h_a = \begin{cases} (2-2p) \\ (1-2p) \\ -2p \end{cases} \quad \text{for genotypes} \quad \begin{cases} AA \\ Aa \\ aa \end{cases}$$

and

$$h_d = \begin{cases} -2q^2 \\ 2pq \\ -2p^2 \end{cases} \quad \text{for genotypes} \quad \begin{cases} AA \\ Aa \\ aa \end{cases}.$$

Note that h_a and h_d contain the coefficients for the additive and dominant contrasts. This model is identical to the "breeding" or classical parameterization presented in Vitezica *et al.* (2013) for genomic evaluation, where genotypic values are split in additive (or breeding) values (g_A) and dominant deviations (g_D), respectively called u and v in Vitezica *et al.* (2013). Thus,

$$g = \mu + g_A + g_D$$

with $g_A = h_a\alpha$ and $g_D = h_d d$. In this model, note that both h_a and h_d are shifted to have a mean of zero for a population in HWE ($\boldsymbol{f} = [p^2, 2pq, q^2]$). Thus, the mean breeding value is zero in HWE because

$$\sum_j h_{a_j} f_j = (2-2p)p^2 + (1-2p)2pq + (-2p)q^2$$
$$= 2pq(1-p-q) = 0,$$

and the mean of dominant deviations is also zero because

$$\sum_j h_{d_j} f_j = (-2q^2)p^2 + (2pq)2pq + (-2p^2)q^2$$
$$= -4q^2p^2 + 4q^2p^2 = 0.$$

These equations correspond to the first requirement of orthogonality in the Cockerham (1954) model. Since the means of h_a and h_d contrasts are scaled to zero for the population, the effects in the model are all orthogonal in HWE and LE. Thus, the substitution effect (Falconer and Mackay 1996) contributes to the additive variance, the dominance deviation contributes to the dominance variance, *etc*. There is no covariance between the genetic effects due to the orthogonal property of the model. For the allele frequency $p = 0.5$, the Cockerham model is a particular case of the G2A and Vitezica models, thus

$$h_a = \begin{cases} 1 \\ 0 \\ -1 \end{cases} \quad \text{for genotypes} \quad \begin{cases} AA \\ Aa \\ aa \end{cases} \quad \text{and}$$
$$h_d = \begin{cases} -1/2 \\ 1/2 \\ -1/2 \end{cases} \quad \text{for genotypes} \quad \begin{cases} AA \\ Aa \\ aa \end{cases}.$$

Here, the definition of the additive and dominant contrast is the same whether or not other (independently segregating) loci or epistatic effects are fitted in the regression model.

Finally, another orthogonal model called NOIA was proposed by Álvarez-Castro and Carlborg (2007), and whose most relevant feature is that it can deal with departures from HWE while keeping statistical orthogonality. For the NOIA model and assuming the notation by Falconer and Mackay (1996), h_a and h_d in $g = \mu + h_a\alpha + h_d d$ are defined as

$$h_a = \begin{cases} -(-p_{Aa} - 2p_{aa}) \\ -(1 - p_{Aa} - 2p_{aa}) \\ -(2 - p_{Aa} - 2p_{aa}) \end{cases} \quad \text{for genotypes} \quad \begin{cases} AA \\ Aa \\ aa \end{cases} \qquad (1)$$

and

$$h_d = \begin{cases} -\dfrac{2p_{Aa}p_{aa}}{p_{AA}+p_{aa}-(p_{AA}-p_{aa})^2} \\ \dfrac{4p_{AA}p_{aa}}{p_{AA}+p_{aa}-(p_{AA}-p_{aa})^2} \\ -\dfrac{2p_{AA}p_{Aa}}{p_{AA}+p_{aa}-(p_{AA}-p_{aa})^2} \end{cases} \text{for genotypes} \begin{cases} AA \\ Aa\,, \\ aa \end{cases} \tag{2}$$

where p_{AA}, p_{Aa}, and p_{aa} are the genotypic frequencies for the genotypes AA, Aa, and aa for the locus A. It can be shown that these contrasts and their Kronecker products, which model epistasis, fit Cockerham's requirements of orthogonality when assuming LE but even in absence of HWE (Álvarez-Castro and Carlborg 2007). When the denominator is zero (monomorphic markers), the marker is ignored. Note that under the assumption of HWE, G2A and the model in Vitezica *et al.* (2013) are a particular case of the NOIA model when $p_{AA}=p^2$, $p_{Aa}=2pq$, $p_{aa}=q^2$, and the denominator $p_{AA}+p_{aa}-(p_{AA}-p_{aa})^2=2pq$. Expanding the NOIA model, genomic relationship matrices for interaction terms can be obtained for any population, in HWE or not, as is detailed in the next section. The straight relationship between breeding values and α as "regression of value on gene dosage" only holds in HWE (Falconer 1985). Note that α here has a least-squares meaning (regression of value on gene dosage), but when HWE does not hold α is not a substitution effect in the sense of "breeding value" (Falconer 1985). Indeed, it is not clear what a breeding value is in absence of HWE.

Equivalent genomic model with epistasis

Using the NOIA model, the additive values of a set of individuals are, for multiple loci, $\boldsymbol{g}_A = \boldsymbol{H}_a\boldsymbol{\alpha}$, with $\boldsymbol{H}_a$ containing one column per locus coded as in Equation 1 for an individual. Also, the dominant value of an individual can be parameterized as in Equation 2 and $\boldsymbol{g}_D = \boldsymbol{H}_d\boldsymbol{d}$ for a set of individuals.

A linear model including additive, dominant, and higher-order interaction terms can be written as:

$$\boldsymbol{y} = 1\mu + \boldsymbol{g} = 1\mu + \boldsymbol{g}_A + \boldsymbol{g}_D + \sum_{i=A,D}\sum_{\substack{j=A,D \\ i\geq j}} \boldsymbol{g}_{ij} + \sum_{i=A,D}\sum_{j=A,D}\sum_{\substack{k=A,D \\ i\geq j\geq k}} \boldsymbol{g}_{ijk} + \ldots + \boldsymbol{e},$$

where $\boldsymbol{y}$ is the vector of phenotypic records, μ is the population mean, $\boldsymbol{g}$ is the total genotypic value, $\boldsymbol{g}_A$ is the additive value of the individual (breeding value if the population is in HWE), $\boldsymbol{g}_D$ is the dominant value, $\boldsymbol{g}_{ij}$ is the second-order epistatic value, $\boldsymbol{g}_{ijk}$ is the third order epistatic value, and so on.

The additive genomic (co)variance relationship matrix will be computed from Equation 1 as:

$$\mathrm{Cov}(\boldsymbol{g}_A) = \frac{\boldsymbol{H}_a\boldsymbol{H}_a'}{tr(\boldsymbol{H}_a\boldsymbol{H}_a')/n}\sigma_A^2 = \boldsymbol{G}_A\sigma_A^2,$$

where $\boldsymbol{H}_a$ is a matrix with n rows (number of individuals) and m columns (number of markers) containing "additive" coefficients (Equation 1), as

$$\boldsymbol{H}_a = \begin{pmatrix} \boldsymbol{h}_{a_i} \\ \vdots \\ \boldsymbol{h}_{a_n} \end{pmatrix},$$

where $\boldsymbol{h}_{a_i}$ is a row vector for the ith individual with m columns. For individual 1, the vector $\boldsymbol{h}_{a_1}$ is equal to $(h_{a_{11}},\ldots,h_{a_{1m}})$, and the element $h_{a_{1j}}$ for the $j=1,\ldots,m$ marker is equal to

$$h_{a_{1j}} = \begin{cases} -(-p_{Aa}-2p_{aa}) \\ -(1-p_{Aa}-2p_{aa}) \\ -(2-p_{Aa}-2p_{aa}) \end{cases} \text{for genotypes} \begin{cases} AA \\ Aa\,. \\ aa \end{cases}$$

On the other hand, $tr(\boldsymbol{H}_a\boldsymbol{H}_a')$ corresponds to the expected variance of $\boldsymbol{H}_a\boldsymbol{\epsilon}_a$ with $\mathrm{Var}(\boldsymbol{\epsilon}_a)=\boldsymbol{I}$ (Searle 1982). In other words, $tr(\boldsymbol{H}_a\boldsymbol{H}_a')$ standardizes the cross-product matrix $\boldsymbol{H}_a\boldsymbol{H}_a'$ to a variance of 1. In a Hardy–Weinberg population, $tr(\boldsymbol{H}_a\boldsymbol{H}_a')$ is equal to the heterozygosity of the markers $2\sum p_iq_i$ (Vitezica *et al.* 2013). Overall, $\boldsymbol{G}_A$ in HWE is similar to the VanRaden (2008) genomic-relationship matrix.

For the dominance, the genomic (co)variance matrix will be:

$$\mathrm{Cov}(\boldsymbol{g}_D) = \frac{\boldsymbol{H}_d\boldsymbol{H}_d'}{tr(\boldsymbol{H}_d\boldsymbol{H}_d')/n}\sigma_D^2 = \boldsymbol{G}_D\sigma_D^2,$$

where the matrix $\boldsymbol{H}_d$ is

$$\boldsymbol{H}_d = \begin{pmatrix} \boldsymbol{h}_{d_i} \\ \vdots \\ \boldsymbol{h}_{d_n} \end{pmatrix},$$

and the elements of the vector $\boldsymbol{h}_{d_i} = (h_{d_{i1}},\ldots,h_{d_{im}})$ for the ith individual are equal to

$$h_{d_{ij}} = \begin{bmatrix} -\dfrac{2p_{Aa}p_{aa}}{p_{AA}+p_{aa}-(p_{AA}-p_{aa})^2} \\ \dfrac{4p_{AA}p_{aa}}{p_{AA}+p_{aa}-(p_{AA}-p_{aa})^2} \\ -\dfrac{2p_{AA}p_{Aa}}{p_{AA}+p_{aa}-(p_{AA}-p_{aa})^2} \end{bmatrix} \text{for genotypes} \begin{cases} AA \\ Aa\,. \\ aa \end{cases}$$

The $tr(\boldsymbol{H}_d\boldsymbol{H}_d')$ standardizes the cross-product matrix $\boldsymbol{H}_d\boldsymbol{H}_d'$ to a variance of 1. Again, for Hardy–Weinberg populations, $tr(\boldsymbol{H}_d\boldsymbol{H}_d') = 4\sum(p_iq_i)^2$ as expected (Vitezica *et al.* 2013).

Álvarez-Castro and Carlborg (2007) proved that the coefficients of the incidence matrix for second-order epistatic effects between two loci can be computed as the Kronecker products of the respective incidence matrices for single locus effects, that is

$$\boldsymbol{h}_{ad_{ij}} = \boldsymbol{h}_{a_i} \otimes \boldsymbol{h}_{d_j},$$

and subsequently (*e.g.*, $\boldsymbol{h}_{add_{ijk}} = \boldsymbol{h}_{a_i} \otimes \boldsymbol{h}_{d_j} \otimes \boldsymbol{h}_{d_k}$) for third- and higher-order epistatic effects. This results in orthogonality of epistatic effects. For instance, consider individual k with two loci:

$$g_k = h_{a_1}\alpha_1 + h_{a_2}\alpha_2 + h_{d_1}d_1 + h_{d_2}d_2 + h_{a_1}\otimes h_{a_2}(\alpha\alpha)_{12} + h_{a_1}\otimes h_{d_2}(\alpha d)_{12} + h_{d_1}\otimes h_{a_2}(d\alpha)_{12} + h_{d_1}\otimes h_{d_2}(dd)_{12}.$$

In fact, the Kronecker product above reorders the effects as follows:

$$g_k = h_{a_1}\alpha_1 + h_{d_1}d_1 + h_{a_2}\alpha_2 + h_{a_1}\otimes h_{a_2}(\alpha\alpha)_{12} + h_{d_1}\otimes h_{a_2}(d\alpha)_{12} + h_{d_2}d_2 + h_{a_1}\otimes h_{d_2}(\alpha d)_{12} + h_{d_1}\otimes h_{d_2}(dd)_{12}.$$

Following this idea, for the interactions, such as additive-by-dominant interactions, the matrix $\boldsymbol{H}_{ad}$ can be written using Kronecker products of each row of the preceding matrices as

$$\boldsymbol{H}_{ad} = \begin{pmatrix} \boldsymbol{h}_{a_i}\otimes\boldsymbol{h}_{d_i} \\ \boldsymbol{h}_{a_{i+1}}\otimes\boldsymbol{h}_{d_{i+1}} \\ \cdots \\ \boldsymbol{h}_{a_n}\otimes\boldsymbol{h}_{d_n} \end{pmatrix}.$$

For instance, for individual 1, the incidence matrix of epistatic effects is:

$$\boldsymbol{h}_{ad_1} = \boldsymbol{h}_{a_1}\otimes\boldsymbol{h}_{d_1}.$$

For instance, the second element of $\boldsymbol{h}_{ad_1}$ contains the additive-by-dominant interaction for the respective loci 1 and 2. For individual 2, this is $\boldsymbol{h}_{ad_2} = \boldsymbol{h}_{a_2}\otimes\boldsymbol{h}_{d_2}$, and so on. The matrix $\boldsymbol{H}_{ad}$ has as many columns as marker interactions, and as many rows as individuals. This matrix is of a very large size (*e.g.*, for a 50K SNP chip and 1000 individuals the matrix contains $1000\times 50,000^2$ elements).

Computation of this cross-product matrix across individuals $\boldsymbol{H}_{ad}\boldsymbol{H}_{ad}'$ can take a considerable time because the matrix $\boldsymbol{H}_{ad}$ is very large, *e.g.*, $(1000\times 50,000^2)^2$ operations. However, there is an algebraic shortcut that allows easy computation of $\boldsymbol{H}_{ad}\boldsymbol{H}_{ad}'$and the rest of the epistatic matrices, even for third and higher orders. Using Searle (1982), p. 265:

$$\boldsymbol{H}_{ad}\boldsymbol{H}_{ad}' = \begin{pmatrix} \boldsymbol{h}_{a_1}\otimes\boldsymbol{h}_{d_1} \\ \boldsymbol{h}_{a_2}\otimes\boldsymbol{h}_{d_2} \\ \cdots \\ \boldsymbol{h}_{a_n}\otimes\boldsymbol{h}_{d_n} \end{pmatrix}\begin{pmatrix} \boldsymbol{h}_{a_1}\otimes\boldsymbol{h}_{d_1} \\ \boldsymbol{h}_{a_2}\otimes\boldsymbol{h}_{d_2} \\ \cdots \\ \boldsymbol{h}_{a_n}\otimes\boldsymbol{h}_{d_n} \end{pmatrix}' = \begin{pmatrix} \boldsymbol{h}_{a_1}\otimes\boldsymbol{h}_{d_1} \\ \boldsymbol{h}_{a_2}\otimes\boldsymbol{h}_{d_2} \\ \cdots \\ \boldsymbol{h}_{a_n}\otimes\boldsymbol{h}_{d_n} \end{pmatrix}\begin{pmatrix} \boldsymbol{h}_{a_1}'\otimes\boldsymbol{h}_{d_1}' & \boldsymbol{h}_{a_2}'\otimes\boldsymbol{h}_{d_2}' & \cdots & \boldsymbol{h}_{a_n}'\otimes\boldsymbol{h}_{d_n}' \end{pmatrix}.$$

And this product is

$$\boldsymbol{H}_{ad}\boldsymbol{H}_{ad}' = \begin{pmatrix} \boldsymbol{h}_{a_1}\otimes\boldsymbol{h}_{d_1}\boldsymbol{h}_{a_1}'\otimes\boldsymbol{h}_{d_1}' & \boldsymbol{h}_{a_1}\otimes\boldsymbol{h}_{d_1}\boldsymbol{h}_{a_2}'\otimes\boldsymbol{h}_{d_2}' & \cdots & \boldsymbol{h}_{a_1}\otimes\boldsymbol{h}_{d_1}\boldsymbol{h}_{a_n}'\otimes\boldsymbol{h}_{d_n}' \\ & \boldsymbol{h}_{a_2}\otimes\boldsymbol{h}_{d_2}\boldsymbol{h}_{a_2}'\otimes\boldsymbol{h}_{d_2}' & \cdots & \boldsymbol{h}_{a_2}\otimes\boldsymbol{h}_{d_2}\boldsymbol{h}_{a_n}'\otimes\boldsymbol{h}_{d_n}' \\ & & \cdots & \cdots \\ \text{symm} & & & \boldsymbol{h}_{a_n}\otimes\boldsymbol{h}_{d_n}\boldsymbol{h}_{a_n}'\otimes\boldsymbol{h}_{d_n}' \end{pmatrix}.$$

Thus, the i,j element of $\boldsymbol{H}_{ad}\boldsymbol{H}_{ad}'$ has the form $\boldsymbol{h}_{a_i}\otimes\boldsymbol{h}_{d_i}\boldsymbol{h}_{a_j}'\otimes\boldsymbol{h}_{d_j}'$. In Searle (1982), p.265, we have that $(\boldsymbol{A}\otimes\boldsymbol{B})(\boldsymbol{C}\otimes\boldsymbol{D}) = (\boldsymbol{AC}\otimes\boldsymbol{BD})$. Applying this, $\boldsymbol{h}_{a_i}\otimes\boldsymbol{h}_{d_i}\boldsymbol{h}_{a_j}'\otimes\boldsymbol{h}_{d_j}'$ has the form $\boldsymbol{h}_{a_i}\boldsymbol{h}_{a_j}'\otimes\boldsymbol{h}_{d_i}\boldsymbol{h}_{d_j}'$. However, each of the products $\boldsymbol{h}_{a_i}\boldsymbol{h}_{a_j}'$, $\boldsymbol{h}_{d_i}\boldsymbol{h}_{d_j}'$ is scalar because $\boldsymbol{h}$ are row vectors. Thus, we have proven that, for instance, the cross-product matrix of the additive-by-dominant interaction can be put as the direct product of the *unscaled* cross-product matrices of additive and dominance:

$$\boldsymbol{H}_{ad}\boldsymbol{H}_{ad}' = \boldsymbol{H}_a\boldsymbol{H}_a'\odot\boldsymbol{H}_d\boldsymbol{H}_d'.$$

However, to standardize and get meaningful relationships we need to divide by the trace:

$$\boldsymbol{G}_{AD} = \frac{(\boldsymbol{H}_a\boldsymbol{H}_a'\odot\boldsymbol{H}_d\boldsymbol{H}_d')}{tr(\boldsymbol{H}_a\boldsymbol{H}_a'\odot\boldsymbol{H}_d\boldsymbol{H}_d')/n}.$$

But $tr(\boldsymbol{H}_a\boldsymbol{H}_a'\odot\boldsymbol{H}_d\boldsymbol{H}_d')/n \neq [tr(\boldsymbol{H}_a\boldsymbol{H}_a')/n][tr(\boldsymbol{H}_d\boldsymbol{H}_d')/n]$ and thus $\boldsymbol{G}_{AD}\neq\boldsymbol{G}_A\odot\boldsymbol{G}_D$ unless all elements in the diagonals of the $\boldsymbol{G}$ matrices equal 1.

Assume that we know $\boldsymbol{G}_A$, $\boldsymbol{G}_D$, and the traces $tr(\boldsymbol{H}_a\boldsymbol{H}_a')$ and $tr(\boldsymbol{H}_d\boldsymbol{H}_d')$. Then

$$\boldsymbol{H}_{ad}\boldsymbol{H}_{ad}' = \boldsymbol{H}_a\boldsymbol{H}_a'\odot\boldsymbol{H}_d\boldsymbol{H}_d' = \frac{1}{n}\,tr(\boldsymbol{H}_a\boldsymbol{H}_a')\,\boldsymbol{G}_A\odot\frac{1}{n}\,tr(\boldsymbol{H}_d\boldsymbol{H}_d')\,\boldsymbol{G}_D = \frac{1}{n}\,tr(\boldsymbol{H}_a\boldsymbol{H}_a')\,\frac{1}{n}\,tr(\boldsymbol{H}_d\boldsymbol{H}_d')\boldsymbol{G}_A\odot\boldsymbol{G}_D$$

and

$$tr(\boldsymbol{H}_{ad}\boldsymbol{H}_{ad}') = \frac{1}{n}\,tr(\boldsymbol{H}_a\boldsymbol{H}_a')\,\frac{1}{n}\,tr(\boldsymbol{H}_d\boldsymbol{H}_d')tr(\boldsymbol{G}_A\odot\boldsymbol{G}_D)$$

$$\frac{1}{n}tr(\boldsymbol{H}_{ad}\boldsymbol{H}_{ad}') = \frac{1}{n}\,tr(\boldsymbol{H}_a\boldsymbol{H}_a')\,\frac{1}{n}\,tr(\boldsymbol{H}_d\boldsymbol{H}_d')\,tr(\boldsymbol{G}_A\odot\boldsymbol{G}_D)\,\frac{1}{n}.$$

Thus

$$\boldsymbol{G}_{AD} = \frac{\boldsymbol{H}_{ad}\boldsymbol{H}_{ad}'}{tr(\boldsymbol{H}_{ad}\boldsymbol{H}_{ad}')/n} = \frac{\frac{1}{n}\,tr(\boldsymbol{H}_a\boldsymbol{H}_a')\,\frac{1}{n}\,tr(\boldsymbol{H}_d\boldsymbol{H}_d')\boldsymbol{G}_A\odot\boldsymbol{G}_D}{\frac{1}{n}\,tr(\boldsymbol{H}_a\boldsymbol{H}_a')\,\frac{1}{n}\,tr(\boldsymbol{H}_d\boldsymbol{H}_d')\,tr(\boldsymbol{G}_A\odot\boldsymbol{G}_D)\,\frac{1}{n}} = \frac{\boldsymbol{G}_A\odot\boldsymbol{G}_D}{tr(\boldsymbol{G}_A\odot\boldsymbol{G}_D)/n}$$

which results in

$$\boldsymbol{G}_{AD} = \frac{\boldsymbol{H}_{ad}\boldsymbol{H}_{ad}'}{tr(\boldsymbol{H}_{ad}\boldsymbol{H}_{ad}')/n} = \frac{\boldsymbol{G}_A\odot\boldsymbol{G}_D}{tr(\boldsymbol{G}_A\odot\boldsymbol{G}_D)/n},$$

which is very simple to do because the element i,j of $\boldsymbol{G}_{AD}$ is the product of the elements i,j of the respective matrices $\boldsymbol{G}_A$ and $\boldsymbol{G}_D$, scaled by $tr(\boldsymbol{G}_A\odot\boldsymbol{G}_D)/n$. Thus, this is the Hadamard product of both matrices divided by the average of its diagonal. Therefore, the covariance matrix can be calculated as:

$$\mathrm{Cov}(\boldsymbol{g}_{AD}) = \frac{\boldsymbol{H}_{ad}\boldsymbol{H}_{ad}'}{tr(\boldsymbol{H}_{ad}\boldsymbol{H}_{ad}')/n}\sigma^2_{AD} = \frac{\boldsymbol{G}_A\odot\boldsymbol{G}_D}{tr(\boldsymbol{G}_A\odot\boldsymbol{G}_D)/n}\sigma^2_{AD} = \boldsymbol{G}_{AD}\sigma^2_{AD}.$$

Accordingly, $\boldsymbol{G}_{AA} = \boldsymbol{G}_A\odot\boldsymbol{G}_A/[tr(\boldsymbol{G}_A\odot\boldsymbol{G}_A)/n]$, and, *e.g.*, $\boldsymbol{G}_{AAD} = \boldsymbol{G}_A\odot\boldsymbol{G}_A\odot\boldsymbol{G}_D/[tr(\boldsymbol{G}_A\odot\boldsymbol{G}_A\odot\boldsymbol{G}_D)/n]$. The normalization factor based on the traces was already used by Xu (2013) but

Table 2 Genotypic effects at two loci under epistasis

	Genotypes at locus 2		
Genotypes at locus 1	BB	Bb	bb
AA	0	0	0
Aa	0	$(dd)_{12}$	0
aa	0	0	0

several authors ignore it (Muñoz *et al.* 2014). Here we presented the reasoning for pairwise interactions but it extends to third and higher order interactions.

We have thus shown how to proceed to an orthogonal decomposition of the variances in any population in HWE or not, assuming LE. We have also shown that orthogonal high-order (epistatic) genomic-relationship matrices involved the Hadamard products of additive and dominant genomic-relationship matrices.

Genetic variances in the presence of two-locus epistasis and LE

In the simulation part of this article we simulate and estimate two-loci epistasis. Here we present the decomposition of the expected variance in such a situation, to recover the expected parameters from the simulation. We will use properties of Cockerham's (1954) idea of orthogonal weighted partitioning in genetic systems, applied to populations not necessarily in HWE, although assuming LE. The expression $\boldsymbol{y} = \boldsymbol{Wb}$ relates genotypic values with statistical effects in $\boldsymbol{b}$, and $\boldsymbol{W}$ is an incidence matrix including an overall mean and all contrasts. Orthogonality of $\boldsymbol{W}$ is warranted in LE if it is constructed by the NOIA system.

For instance, in a two-loci (j, k) case, $\boldsymbol{y}$ has nine values, ordered as AABB, AaBB, aaBB. . .aabb. $\boldsymbol{W} = \boldsymbol{W}_j \otimes \boldsymbol{W}_k$, where $\boldsymbol{W}_k = (\,1 \quad h_a \quad h_d\,)$ is a 3×3 matrix for locus k with alleles A/a. In $\boldsymbol{W_k}$, the first column contains 1, and the second and third columns are equal to Equations 1 and 2.

The orthogonal system has the two properties described before: $\sum_{j,k} f_{jk} \boldsymbol{w}_{jkl} = 0$ and $\sum_{j,k} f_{jk} \boldsymbol{w}_{jkl} \boldsymbol{w}_{jkl'} = 0$ for $l \neq l'$. These properties allow a straightforward estimation of statistical effects and variance components. The statistical effects can be obtained from weighted linear regression:

$$\boldsymbol{b} = (\boldsymbol{W}'\boldsymbol{D}\boldsymbol{W})^{-1}\boldsymbol{W}'\boldsymbol{D}\boldsymbol{y},$$

where $\boldsymbol{D}$ is a diagonal matrix (for two loci, a 9×9 matrix) with genotypic frequencies (weights) in the diagonal. This approach extends to any order of interaction. The total variance can be obtained from $\boldsymbol{y}'\boldsymbol{D}\boldsymbol{y} = \boldsymbol{b}'\boldsymbol{W}'\boldsymbol{D}\boldsymbol{W}\boldsymbol{b}$. All the variance components can be obtained from

$$\boldsymbol{V} = \boldsymbol{B}'\boldsymbol{W}'\boldsymbol{D}\boldsymbol{W}\boldsymbol{B},$$

where $\boldsymbol{B}$ is a diagonal matrix that contains the effects in $\boldsymbol{b}$ in its diagonal. Matrix $\boldsymbol{V}$ contains, in its diagonal, the variance components associated to each contrast; in the case of two loci, these are nine. Because the system is orthogonal, the out-of-diagonal terms in $\boldsymbol{V}$ contain 0. Because the Kronecker product reorders effects in the linear system, for the two-loci case we have:

$$V_A = V_{[2,2]} + V_{[4,4]},$$

$$V_D = V_{[3,3]} + V_{[7,7]},$$

$$V_{AA} = V_{[5,5]},$$

$$V_{AD} = V_{[6,6]} + V_{[8,8]},$$

$$V_{DD} = V_{[9,9]}.$$

Table 3 Proportion of loci not in HWE and LD (r^2) in the population simulated under divergent selection and LD

	Line 1	Line 2	F_1
Proportion of loci not in HWE			
Markers	0.05	0.04	0.11
QTL	0.04	0.05	0.13
LD			
r^2 at <1 cM	0.43	0.47	0.40
r^2 at 4–5 cM	0.14	0.14	0.10

For the F_1 case that will be simulated, genotypic frequencies are as follows. Consider two loci k, A/a, and j, B/b. The allelic frequencies in the parental populations 1 and 2 are p_A^1 and p_A^2 and p_B^1 and p_B^2. The F_1 population has genotypic frequencies for locus A $(p_A^1 p_A^2,\ p_A^1 q_A^2 + q_A^1 p_A^2,\ q_A^1 q_A^2)$ for the three genotypes (AA,Aa,aa) and similarly for locus B. In the F_1 population, each locus has the following genotypic frequencies:

$$\boldsymbol{D}_A = \begin{pmatrix} p_A^1 p_A^2 & 0 & 0 \\ 0 & p_A^1 q_A^2 + q_A^1 p_A^2 & 0 \\ 0 & 0 & q_A^1 q_A^2 \end{pmatrix},$$

$$\boldsymbol{D}_B = \begin{pmatrix} p_B^1 p_B^2 & 0 & 0 \\ 0 & p_B^1 q_B^2 + q_B^1 p_B^2 & 0 \\ 0 & 0 & q_B^1 q_B^2 \end{pmatrix}.$$

The frequencies of the nine genotypes at the two loci are, assuming LE, the Kronecker product of each locus' frequencies: $\boldsymbol{D} = \boldsymbol{D}_B \otimes \boldsymbol{D}_A$.

For instance, let $\boldsymbol{y}' = (\,0 \quad 0 \quad 0 \quad 0 \quad 1 \quad 0 \quad 0 \quad 0 \quad 0\,)$. Only the double heterozygote has a functional value different from 0. Let $p_A^1 = 0.2$, $p_B^1 = 0.7$, $p_A^2 = 0.7$, and $p_B^2 = 0.3$. This gives $\text{diag}(\boldsymbol{V}) = (\,0.129 \quad 0.003 \quad 0.076 \quad 0.000 \quad 0.000 \quad 0.0000.094 \quad 0.003 \quad 0.055)$, corresponding to the nine sums of squares attributed to the overall mean μ, α_1, d_1, α_2, $(\alpha\alpha)_{12}$, $(d\alpha)_{12}$, d_2, $(\alpha d)_{12}$, and $(dd)_{12}$. From this, we get $V_A = 0.00349$, $V_D = 0.1694$, $V_{AA} = 0$, $V_{AD} = 0.00253$, and $V_{DD} = 0.05486$. For n pairs of epistatic loci, the procedure is repeated n times and the n variances are added. The procedure is generalizable to any number of interactions. An R code for this example is provided in Supplemental Material, File S1.

Table 4 Features of the genomic relationship matrices constructed assuming HWE or not (NOIA) in the F_1 population in LE

	HWE		NOIA	
Component	Mean	Mean (diagonal)	Mean	Mean (diagonal)
Additive	0	0.59	0	1
Dominant	0.29	0.93	0	1
Additive-by-additive	0	0.35	0	1
Additive-by-dominant	0	0.55	0	1
Dominant-by-dominant	0.08	0.87	0	1

Table 5 Estimates of variance components assuming HWE or not (NOIA) using a complete model with two-locus epistasis in LE scenario

Variance components	HWE	HWE corrected[a]	NOIA	True
Additive	12.04 (2.15)	7.10	15.43 (1.66)	13.01
Dominant	37.90 (4.12)	24.26	19.01 (1.99)	18.49
Additive-by-additive	1.34 (1.41)	0.46	0.72 (0.79)	1.27
Additive-by-dominant	3.26 (1.72)	1.79	2.59 (0.88)	3.38
Dominant-by-dominant	3.32 (1.92)	2.59	0.63 (0.60)	2.20
Residual	1.37 (0.87)	1.37	1.62 (1.03)	0
Total	59	38	40	38

Posterior SD values shown in parenthesis.
[a] Variance components were corrected as in Legarra (2016).

Example

Simulated data

We simulated an F_1 population deviating strongly from HWE and two scenarios: (1) QTL in LE and directly observed, and (2) QTL in LD and estimation via markers.

In the LE scenario, we simulated a quantitative trait with pure biological (or functional) dominant-by-dominant effects (see Table 2) of pairs of QTL loci ($n/2$ exclusive pairs from n loci), such that the double heterozygote has a value $(dd)_{ij}$ sampled from a Bernouilli distribution $(-1,1)$ and all other genotypes have value 0. It is important to recall that, when allelic frequencies differ from 0.5 and across loci, this dominant-by-dominant epistatic interaction generates statistical additive, dominance, and the three kinds of epistatic variances. Thus, we consider this as an extreme case for illustration purposes. Expected true variance components were obtained as explained in the previous section from the simulated effects and allelic frequencies.

Total genetic values were calculated for 1000 individuals by adding the genotypic values of 200 pairs of unlinked interacting loci from 400 biallelic QTL. Each individual has two alleles that each come from a different population, say line 1 and line 2, with different allelic frequencies both drawn from a U-shaped Beta (0.5,0.5) distribution. Allelic frequencies were simulated as negatively correlated (−0.5) across the populations to mimic divergent selection. The proportion of loci not in HWE (P-value < 0.001) turned out to be 0.94. The heritability was equal to 1 and the total genetic (and phenotypic) variance was equal to 39. No residual/environmental variance was simulated.

In the LD scenario, from a historical common population in drift- mutation equilibrium, two divergent lines were divergently selected based on phenotype for 10 generations, followed by an F_1 cross. The simulation included 30 chromosomes of 1 M with, potentially, 5000 markers and 100 QTL each. We used QMSim (Sargolzaei and Schenkel 2009) and our own software. A quantitative trait with heritability equal to 0.5 was simulated in the creation and selection of lines 1 and 2, using the same biological dominant-by-dominant mechanism as above. Lines included 20 males and 20 females per generation, with a number of offspring of 10 per couple. Thus, selection was quite strong. To generate the F_1, 80 males and 80 females were drawn from the last generations of line 1 and 2 to generate 10 offspring per couple. Therefore, the data set consisted of 800 crossed individuals. In the F_1, the heritability was equal to one and no residual variance was simulated. The mean phenotypic values were equal to 6.69, −1.87, and 3.20 for lines 1 and 2 in generation 10 and in F_1, respectively, for a total genetic variance of 29 in the F_1. The pairs of epistatic loci and the functional effects of their interactions were generated at random at the beginning of the simulation and did not change with time.

The number of segregating loci in the F_1 was 10,225 SNPs and 1432 QTL. The overall deviation from HWE was measured as the proportion of loci not in HWE (P-value < 0.001) on 200 individuals from line 1, line 2, and F_1 (Table 3). LD was monitored from QMSim output in generation 10 for line 1 and 2, and in the F_1 animals (Table 3). Levels and decay of LD (r^2) with genetic map distance are in agreement with the values simulated by Schopp *et al.* (2017) for a scenario of "ancestral long-range LD" in plant breeding. Expected pseudotrue variance components were obtained (incorrectly assuming LE) from the simulated effects and allelic frequencies in the F_1 as explained in the previous section.

Variance components were estimated for the two scenarios (LE and LD) with two different models, *i.e.*, assuming HWE (model HWE) or not (model NOIA). The general model was:

$$\boldsymbol{y} = 1\mu + \boldsymbol{g}_A + \boldsymbol{g}_D + \boldsymbol{g}_{AA} + \boldsymbol{g}_{AD} + \boldsymbol{g}_{DD} + \boldsymbol{e}$$

$$\mathrm{Var}(\boldsymbol{g}_A) = \boldsymbol{G}_A\sigma^2_A; \quad \mathrm{Var}(\boldsymbol{g}_D) = \boldsymbol{G}_D\sigma^2_D;$$

$$\mathrm{Var}(\boldsymbol{g}_{AA}) = \boldsymbol{G}_{AA}\sigma^2_{AA}; \quad \mathrm{Var}(\boldsymbol{g}_{AD}) = \boldsymbol{G}_{AD}\sigma^2_{AD};$$
$$\mathrm{Var}(\boldsymbol{g}_{DD}) = \boldsymbol{G}_{DD}\sigma^2_{DD}; \quad \mathrm{Var}(\boldsymbol{e}) = \boldsymbol{I}\sigma^2_e;$$

where the covariance matrices $\boldsymbol{G}_A$ and $\boldsymbol{G}_D$ (involved in the construction of $\boldsymbol{G}_{AA}$, $\boldsymbol{G}_{AD}$, and $\boldsymbol{G}_{DD}$) were constructed either assuming HWE as in Vitezica *et al.* (2013) or using the NOIA approach as in Equations 1 and 2. The genotypes at the QTL were directly used to construct the matrices.

Genetic (σ^2_A, σ^2_D, σ^2_{AA}, σ^2_{AD}, and σ^2_{DD}) variances were estimated by Bayesian methods using Gibbs sampling using the

Table 6 Estimates of variance components assuming HWE or not (NOIA) using a complete model with two-locus epistasis in LD scenario

Variance components	HWE	HWE corrected[a]	NOIA	Pseudotrue[b]
Additive	9.92 (2.67)	9.16	10.25 (2.78)	8.54
Dominant	9.52 (3.19)	8.35	8.47 (2.81)	8.62
Additive-by-additive	5.52 (3.68)	4.68	8.89 (3.79)	2.60
Additive-by-dominant	4.71 (3.55)	3.89	0.77 (0.73)	6.35
Dominant-by-dominant	3.78 (3.12)	3.03	1.22 (1.09)	3.02
Residual	1.21 (1.21)	1.21	1.45 (1.51)	0
Total	35	29	31	29

Posterior SD values shown in parenthesis.
[a] Variance components were corrected as in Legarra (2016).
[b] Variances if loci were in LE.

software gibbs2f90 (Misztal *et al.* 2002), available at http://nce.ads.uga.edu/wiki/. After estimation, variances estimated assuming HWE equilibrium were transformed to a proper scale, multiplying the estimates by $D_k = \overline{\text{diag}(\boldsymbol{K})} - \overline{\boldsymbol{K}}$ (Legarra 2016), where $\boldsymbol{K}$ is a relationship matrix (*e.g.*, $\boldsymbol{G}_A$ or $\boldsymbol{G}_{AD}$). This is needed because, in absence of HWE, these relationship matrices do not have a mean of 1 in the diagonal and 0 overall; see Legarra (2016) for a detailed explanation. This was not needed for the NOIA approach as in all cases $D_K \approx 1$. Finally, to ascertain if variance component estimates were empirically orthogonal, we computed their correlation in the posterior distribution.

Data availability

Programs and data are available at http://genoweb.toulouse.inra.fr/~zvitezic/simuepi_genetics.

Results

Table 4 shows the statistics of the different relationship matrices for the F_1 population in LE. The diagonal elements do not have an average diagonal of 1, as the divisors $2\sum p_i q_i$, $4\sum p_i^2 q_i^2$, and so on in Vitezica *et al.* (2013) are conceived for populations in HWE. In addition, in absence of HWE, the mean of the matrix is not zero, in particular for the dominant matrix. Both phenomena affect the interpretation of variance components (Legarra 2016). However, NOIA is free of this problem [although the mean of $\boldsymbol{G}_{AA}$ and $\boldsymbol{G}_{DD}$ is actually not zero, it is very small (<0.005)]. Similar results were observed in the scenario with LD. Whereas divergence in the LE scenario was very strong (correlation of allelic frequencies across lines 1 and 2 equal to −0.5), here divergence was much lower, even after 10 generations of divergent selection (correlation of allelic frequencies across populations of 0.70, both for markers and QTL). The number of loci not in HWE in the F_1 was 11%, which is not too high. In the LE simulation, the number of loci not in HWE was 94%.

In the scenario with LE, estimates of variance components for the two different models (HWE and NOIA) are shown in Table 5. Most of the variance goes to nonepistatic effects as expected (Falconer and Mackay 1996; Hill *et al.* 2008; Huang and Mackay 2016). However, the HWE and NOIA models radically differ in their estimates. The NOIA model retrieves the simulated variance components (last column called "true" in Table 5). The HWE model (which incorrectly assumes HWE) gives a completely different partitioning of

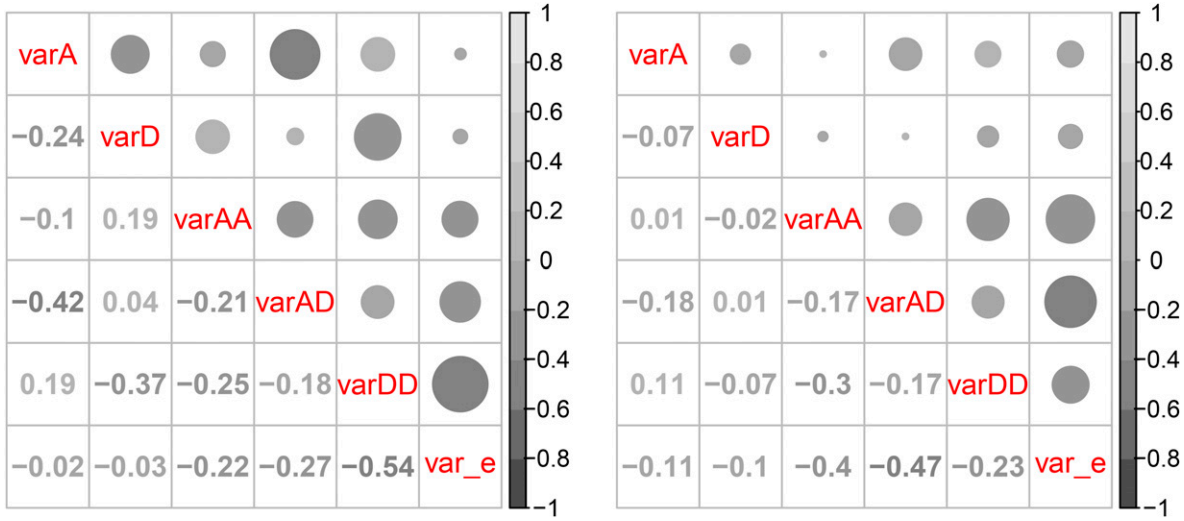

Figure 1 *A-posteriori* correlations among variance component estimates (var_e is the residual variance) in LE scenario. Left: assuming HWE. Right: using the orthogonal NOIA model. Size of the ● is proportional to absolute magnitude of the correlation.

the variance, yielding an estimate of total genetic variance of 58 whereas the simulated variance is 39. After correction as in Legarra (2016), the sum of genetic variances in the HWE model yields a correct estimate of total genetic variance. However, the partitioning in the different components is still incorrect and quite different from NOIA.

Estimates of variance components fitting, exclusively, additive relationships were computed for the LE scenario. The NOIA model is orthogonal; the estimate (a value of 15) is the same as the estimate with the complete model with all direct and epistatic components (Table 5). When only additive effects are fit, NOIA assigns dominant and epistatic components to the residual. The HWE model is *not* orthogonal; the estimate fitting additive relationships only (a value of 25) is different from the one obtained fitting all components (a value of 12 in Table 5). When only additive effects are estimated and the correction D_k is used in the HWE model, results are similar to NOIA values.

In the LD scenario, the variance component estimates are similar in both models, assuming HWE or NOIA (Table 6), although the NOIA model obtains a correct estimate of total genetic variance even with LD, while the HWE model overestimates it by 20%. In Table 6, the simulated variance components are pseudotrue variance components because they do not account for LD. Note that LD among loci has an effect on estimations that cannot be ignored. Differences between the HWE and NOIA models are less clear than in the LE scenario, nonetheless because divergence of lines is much lower, as well as the deviations from HW in the F_1. The sum of nonadditive variances is higher than the additive variance, and, in particular, the additive-by-additive variance is quite high. This could be due to the selection of linked epistatic pairs of loci, something that the expected variances under LE cannot consider. Empirically, neither the HWE nor NOIA models were orthogonal under LD. Both models estimated different values when only additive relationships were fit than when the model fit all variance components (estimates of 17 *vs.* 9, respectively).

Lastly, in Figure 1 we show the *a-posteriori* correlations across estimates for the LE scenario. The HWE analysis shows a correlation of -0.24 between additive and dominant variance components, which shows difficulty in disentangling both components using this model. This correlation is reduced to -0.07 in NOIA. High-order terms are more difficult to estimate than linear ones. The *a-posteriori* correlations between epistatic variance components were not always the lowest for NOIA. Nevertheless, the mean absolute correlation among all (additive, dominant, and epistatic) variance components is lower for NOIA (0.11) than for HWE (0.22).

Discussion

In this work, we propose using the NOIA model, which was conceived for QTL studies, to account for epistasis and partition the variance in a genomic evaluation context. The main advantage of the NOIA model is that the assumption of HWE is relaxed, so that the model applies equally well to populations in HWE or other populations not in HWE, such as F_1 (*e.g.*, corn, pigs), three-way crosses, or backcrosses, all of which are quite common in agriculture and livestock. In this genomic evaluation framework, the NOIA model can be used to define appropriate genomic-relationship matrices for the additive, dominant, and epistatic genetic effects for each individual. This was initially shown by Varona *et al.* (2014), a work that we generalize and complete here.

We have also shown that using Hadamard products of relationship matrices is equivalent to the direct estimation of loci-based epistatic effects for all possible levels of interactions, including dominance, something that had not been proven before. Jiang and Reif (2015) make an asymptotic proof for pairwise additive interactions, whereas here we provide a complete proof including dominance and for any order of interactions. The Hadamard product relies on the orthogonal property of the model because no covariance exists between main genetic effects (*i.e.*, additive and epistatic effects). However, the derivation of the epistatic relationship matrices using the Hadamard product depends on the assumption of noninbreeding and random mating, or, in other words, HWE (Cockerham 1954). Henderson (1985) suggested using the Hadamard product of the (pedigree-based) additive and dominant relationship matrices to obtain the epistatic relationship matrices. Henderson's approach was extended to the genomic framework by Xu (2013) for an F_2 design. For instance, Xu *et al.* (2014) predicted hybrid performance in a rice F_2 population. Recently, the Hadamard product of additive and dominant genomic-relationship matrices was used to construct epistatic relationship matrices in pigs (Su *et al.* 2012), pine trees (Muñoz *et al.* 2014), and wheat and maize (Jiang and Reif 2015), often without standardization of the resulting matrices or incorrectly assuming HWE. In absence of HWE, and to get meaningful estimates of variances that sum to the phenotypic variance, a further standardization (Legarra 2016) may be needed—this has not been considered by other authors.

In simulating an F_1 population, we observed that even if there is only biological epistasis (dominant-by-dominant epistasis), the statistical additive and dominant effects capture an important part of the variation. This is as expected because epistasis contributes to additive and dominant genetic variances (Hill *et al.* 2008; Mäki-Tanila and Hill 2014). In LE, the difference between the HWE and NOIA model estimates can be large. Both yield different partitions of the genetic variance and, worse, HWE is neither correctly scaled nor orthogonal. Thus, inclusion of more genetic effects in analysis assuming HWE can dramatically change estimates. In some studies, the nonorthogonality can be seen because introducing new variance components dramatically changes previous estimates, something that does not occur for orthogonal models. Our approach based on the orthogonal NOIA model allows correct partitioning of the genetic variance, even in Hardy–Weinberg disequilibrium, through

the correct use of Hadamard products to obtain epistatic relationships.

Although the NOIA model removes the assumption of HWE, another factor that generates nonindependence between loci and lack of orthogonality is LD. The LD affects the partition into additive, dominance, and epistatic components, such that an orthogonal partition is not possible (Hill and Mäki-Tanila 2015). In fact, LD may introduce genetic covariances between different genetic effects and complicates the definition of genetic effects and the partition of the genetic variance, in particular in presence of epistasis (Zeng *et al.* 2005). In addition, there are suspicions of favorable epistatic combinations of linked loci, leading to outbreeding depression and recombination loss (Dickerson 1969; Lynch 1991). Still, in our simulation in LD, the NOIA model returned the correct total genetic variance, while the model assuming HWE overestimated the total genetic variance by 20%. In F_1 and other crosses, the NOIA model is a more realistic model able to deal with the HWE assumption even if estimates may be biased due to lack of orthogonality of the effects due to LD. Note that a tight linkage is needed to yield substantial LD in outbred populations (Hill and Mäki-Tanila 2015). For reliable parameter estimation, it is important to work with the most "reliable" statistical model before claiming high importance of the epistatic variance.

It is tempting to fit high-order epistasis terms given the relative ease of computing the direct products. There are two words of caution: First, the products $\boldsymbol{G}\odot\boldsymbol{G}\odot\boldsymbol{G}\ldots$ tend quickly to the identity matrix, in which case there is no hope of distinguishing genetic components from residual environment. This can already be observed in Table 5 and Table 6: only the first components are estimated with some accuracy, and in Figure 1 epistatic terms have *a-posteriori* high correlations with residual variance. Second, variance decomposition in terms of additive, dominant, and epistatic terms does not indicate the relative importance of additive and nonadditive gene actions (Falconer and Mackay 1996; Hill *et al.* 2008; Huang and Mackay 2016).

We have not undertaken the change of reference operation included in the NOIA framework, which allows to transform genetic effects in a reference system (say, statistical effects) into another one (say, functional genetic effects). In principle, this is feasible by a multiple loci extension of the operations in Álvarez-Castro and Carlborg (2007). This should be a line of future research.

Conclusions

Our approach to construct NOIA genomic relationships is flexible and general for populations not in HWE, such as in inbred lines or F_1 crosses of plants and animals, but also for populations in HWE. Genetic effects in NOIA are orthogonal in LE regardless of the genotypic frequencies in the population. Further, we have proven that any genetic modeling with interactions generalizes to Hadamard products of additive and dominant relationships, something that had not been proven before for the general case; but with special care for standardization, something that is not always done. The underlying orthogonal NOIA model is useful on the partition of additive and nonadditive genetic variation and is able to recover the total genetic variance. Assuming HWE when it is not present can yield biased and nonorthogonal partitions of the genetic variance. For reliable parameter estimation, it is important to work with the most reliable statistical model such as the NOIA model which can deal with lack of HWE.

Acknowledgments

We are grateful to Mehdi Sargolzaei for a modified version of QMSim and to members of the EpiSel project for their helpful and constructive comments. Discussions with Simon Teyssèdre (RAGT, Rodez, France) are greatly acknowledged. This work has been financed by the Institut National de la Recherche Agronomique Sélection Génomique metaprogram, project EpiSel (Z.V., A.L.), as well as the project CGL2016-80155 of the Spanish Ministerio de Economía y Competitividad (L.V.). The project was partly supported by Toulouse Midi-Pyrénées bioinformatics platform.

Literature Cited

Álvarez-Castro, J. M., and Ö. Carlborg, 2007 A unified model for functional and statistical epistasis and its application in quantitative trait loci analysis. Genetics 176: 1151–1167.

Cockerham, C. C., 1954 An extension of the concept of partitioning hereditary variance for analysis of covariances among relatives when epistasis is present. Genetics 39: 859–882.

Dickerson, G. E., 1969 Techniques for research in quantitative animal genetics, pp. 36–79 in *Techniques and Procedures in Animal Production Research*. American Society of Animal Science, New York.

Falconer, D. S., 1985 A note on Fisher's 'average effect' and 'average excess'. Genet. Res. 46: 337–347.

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

Fisher, R. A., 1918 The correlation between relatives on the supposition of mendelian inheritance. Trans. R. Soc. Edinb. 52: 399–433.

Goddard, M., 2009 Genomic selection: prediction of accuracy and maximisation of long term response. Genetica 136: 245–257.

Henderson, C. R., 1985 Best linear unbiased prediction of nonadditive genetic merits. J. Anim. Sci. 60: 111–117.

Hill, W. G., and A. Mäki-Tanila, 2015 Expected influence of linkage disequilibrium on genetic variance caused by dominance and epistasis on quantitative traits. J. Anim. Breed. Genet. 132: 176–186.

Hill, W. G., M. E. Goddard, and P. M. Visscher, 2008 Data and theory point to mainly additive genetic variance for complex traits. PLoS Genet. 4: e1000008.

Huang, W., and T. F. C. Mackay, 2016 The genetic architecture of quantitative traits cannot be inferred from variance component analysis. PLoS Genet. 12: e1006421.

Jiang, Y., and J. C. Reif, 2015 Modeling epistasis in genomic selection. Genetics 201: 759–768.

Kao, C. H., and Z. B. Zeng, 2002 Modeling epistasis of quantitative trait loci using Cockerham's model. Genetics 160: 1243–1261.

Kempthorne, O., 1954 The correlation between relatives in a random mating population. Proc. R. Soc. Lond. B Biol. Sci. 143: 103–113.

Legarra, A., 2016 Comparing estimates of genetic variance across different relationship models. Theor. Popul. Biol. 107: 26–30.

Lynch, M., 1991 The genetic interpretation of inbreeding depression and outbreeding depression. Evolution 45: 622–629.

Lynch, M., and B. Walsh, 1998 *Genetics and Analysis of Quantitative Traits*. Sinauer associates, Sunderland, MA.

Mäki-Tanila, A., and W. G. Hill, 2014 Influence of gene interaction on complex trait variation with multilocus models. Genetics 198: 355–367.

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

Misztal, I., S. Tsuruta, T. Strabel, B. Auvray, T. Druet *et al.*, 2002 BLUPF90 and related programs (BGF90). Proceedings from the 7th World Congress on Genetics Applied to Livestock Production, Communication No. 28–07, Montpellier, France.

Muñoz, P. R., M. F. R. Resende, S. A. Gezan, M. D. Vilela Resende, G. de los Campos *et al.*, 2014 Unraveling additive from nonadditive effects using genomic relationship matrices. Genetics 198: 1759–1768.

Nishio, M., and M. Satoh, 2014 Including dominance effects in the genomic BLUP method for genomic evaluation. PLoS One 9: e85792.

Sargolzaei, M., and F. S. Schenkel, 2009 QMSim: a large-scale genome simulator for livestock. Bioinformatics 25: 680–681.

Schopp, P., D. Müller, F. Technow, and A. E. Melchinger, 2017 Accuracy of genomic prediction in synthetic populations depending on the number of parents, relatedness, and ancestral linkage disequilibrium. Genetics 205: 441–454.

Searle, S. R., 1982 *Matrix Algebra Useful for Statistics*. John Wiley & Sons, New York.

Su, G., O. F. Christensen, T. Ostersen, M. Henryon, and M. S. Lund, 2012 Estimating additive and non-additive genetic variances and predicting genetic merits using genome-wide dense single nucleotide polymorphism markers. PLoS One 7: e45293.

Toro, M. A., and L. Varona, 2010 A note on mate allocation for dominance handling in genomic selection. Genet. Sel. Evol. 42: 33.

VanRaden, P. M., 2008 Efficient methods to compute genomic predictions. J. Dairy Sci. 91: 4414–4423.

Varona, L., Z. G. Vitezica, S. Munilla, and A. Legarra, 2014 A general approach for calculation of genomic relationship matrices for eEpistatic effects. Proceedings from the 10th World Congress on Genetics Applied to Livestock Production, Vancouver, Canada, pp. 11–22.

Vitezica, Z. G., L. Varona, and A. Legarra, 2013 On the additive and dominant variance and covariance of individuals within the genomic selection scope. Genetics 195: 1223–1230.

Xu, S., 2013 Mapping quantitative trait loci by controlling polygenic background effects. Genetics 195: 1209–1222.

Xu, S., D. Zhu, and Q. Zhang, 2014 Predicting hybrid performance in rice using genomic best linear unbiased prediction. Proc. Natl. Acad. Sci. USA 111: 12456–12461.

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: 565–569.

Zeng, Z. B., T. Wang, and W. Zou, 2005 Modeling quantitative trait loci and interpretation of models. Genetics 169: 1711–1725.

Communicating editor: M. P. L. Calus