Theor Appl Genet (2013) 126:1457–1472
DOI 10.1007/s00122-013-2064-2

ORIGINAL PAPER

Contribution of an additive locus to genetic variance when inheritance is multi-factorial with implications on interpretation of GWAS

Daniel Gianola · Frederic Hospital · Etienne Verrier

Received: 29 September 2012 / Accepted: 8 February 2013 / Published online: 19 March 2013

Abstract Although the effects of linkage disequilibrium (LD) on partition of genetic variance have received attention in quantitative genetics, there has been little discussion on how this phenomenon affects attribution of variance to a given locus. This paper reinforces the point that standard metrics used for assessing the contribution of a locus to variance can be misleading when there is linkage LD and that factors such as distribution of effects and of allelic frequencies over loci, or existence of frequency-dependent effects, play a role as well. An apparently new metric is proposed for measuring how much of the variability is contributed by a locus when LD exists. Effects of intervening factors, such as type and extent of LD, number of loci, distribution of effects, and of allelic frequencies over loci, as well as a model for generating frequency-dependent effects, are illustrated via hypothetical simulation scenarios. Implications on the interpretation of genome-wide association studies (GWAS), as typically carried out in human genetics, where single marker regression and the assumption of a sole quantitative trait locus (QTL) are common, are discussed. It is concluded that the standard attributions to variance contributed by a single QTL from a GWAS analysis may be misleading, conceptually and statistically, when a trait is complex and affected by sets of many genes in linkage disequilibrium. Yet another factor to consider in the "missing heritability" saga?.

Communicated by M. Frisch.

D. Gianola (✉)
Department of Animal Sciences,
University of Wisconsin-Madison,
Madison, WI 53706, USA
e-mail: gianola@ansci.wisc.edu

D. Gianola
Department of Animal and Aquacultural Sciences,
Norwegian University of Life Sciences, N-1432 Ås, Norway

F. Hospital
INRA, UMR1313 Génétique animale et biologie intégrative,
78350 Jouy-en-Josas, France

E. Verrier
AgroParisTech, UMR1313 Génétique animale et biologie
intégrative, 75231 Paris 05, 16 Rue Claude Bernard, France

Introduction

Linkage disequilibrium has an impact on the variation of quantitative traits. Early studies include those of Comstock and Robinson (1952); Bulmer (1976), and Avery and Hill (1979), among others. The question of how much a given locus contributes to genetic variability of a trait has resurfaced in the context of genome-wide association studies (e.g., Weir 2008; Manolio et al. 2009; Powell et al. 2011) and in prediction of complex traits via whole-genome marker regression (e.g., Meuwissen et al. 2001; de los Campos et al. 2010; Ober et al. 2012; Heslot et al. 2012). When a single locus affects the trait, an answer to this question can be found in quantitative genetics texts such as Falconer and Mackay (1996). However, in a multi-factorial situation, standard formulae apply provided that genotypes at the loci affecting the target trait have mutually independent distributions, a situation often referred to as one of linkage equilibrium (LE). However, linkage disequilibrium (LD) is the rule, rather than the exception (Hill and Robertson 1968; Sabbati and Risch 2002; Zhao et al. 2005). For example, a significant marker-trait association is based on the premise that this is a reflection of LD between a marker and some unknown "causal" genomic region. Levels of LD are much higher in plants and livestock than

in humans, arguably due to small population sizes, crossing, migration (admixture), and artificial selection keeping alleles affecting the trait or fitness in favorable manners coupled (e.g., Goddard and Hayes 2009), but also creating negative linkage disequilibrium as well (Bulmer 1971).

The objective of this paper is to illustrate and reinforce the point that the standard metrics frequently used for assessing the contribution of a locus to variance are misleading when there is LD. We also argue that factors such as the distribution of effects and of allelic frequencies over loci, or the existence of frequency-dependent effects, can play a role as well and that all these factors need to be considered for interpreting the attribution to variance. This is done using theory and several stylized simulations. The paper flows as follows. Section "Multi-locus setting" introduces notation and a metric proposed for measuring how much of the variability is contributed by a locus. Subsequently, intervening factors, such as type and extent of LD, number of loci, distribution of effects and of allelic frequencies over loci, as well as a model for generating frequency-dependent effects, are discussed. The "Results" section reports several simulated scenarios used to provide quantitative evidence of the extent of over (under) statement of the importance of a locus when LD exists. The paper concludes with a discussion of the implications of the findings of this study on interpretation of genome-wide association studies (GWAS), as typically carried out in animal, human, and plant genetics.

Material and methods

Multi-locus setting

Using the notation of Falconer and Mackay (1996), consider a bi-allelic locus model with genotypes AA, Aa and aa, having effects a, d and $-a$ on some quantitative trait (or latent scale such as liability to disease), respectively. Under Hardy-Weinberg equilibrium and with the allelic frequencies being $\Pr(A) = p$ and $\Pr(a) = 1 - p = q$, the variance generated by the locus (Falconer and Mackay 1996) is $2pq[a + d(q - p)]^2 + (2pqd)^2$, reducing to $2pqa^2$ in the absence of dominance ($d = 0$). In the predecing expression, $a + d(q - p)$ is the average effect of a gene substitution, which is a without dominance. With K additive bi-allelic loci, the genetic value of subject i as

$$u_i = W_{i1}a_1 + W_{i2}a_2 + ... + W_{iK}a_K \quad (1)$$

where W_{ij} is a random indicator variable denoting the genotype of i at locus j, and a_j is the fixed additive effect of such locus, defined as the partial regression of u_i on the number of copies of allele A_j. This distinction is essential, since in whole-genome prediction methods breeders typically treat genotypes as fixed, but the effects a_j as random. As emphasized by Gianola et al. (2009) it is the randomness of the $W's$ that underlies the concept of genetic variance, producing the probability distribution

$$W_{ij}a_j = \begin{cases} -a_j & \text{if } W_{ij} = -1\,(aa);\ \Pr(W_{ij} = -1) = (1-p_j)^2 \\ 0 & \text{if } W_{ij} = 0\,(Aa);\ \Pr(W_{ij} = 0) = 2p_j(1-p_j) \\ a_j & \text{if } W_{ij} = 1\,(AA);\ \Pr(W_{ij} = 1) = p_j^2 \end{cases}$$

Hence, u_i possesses a discrete distribution involving 3^K disjoint events, not all of which are observable in a finite sample, especially, if K is large and some joint frequencies are very small. If $K \to \infty$ and the loci are unlinked the distribution of u_i converges to a Gaussian, as in the infinitesimal model of quantitative genetics (Fisher 1918; Bulmer 1980). If the number of loci is finite, and these are in linkage equilibrium (LE), the additive genetic variance is

$$V_A = \mathrm{Var}(u_i) = \sum_{k=1}^{K} 2p_k(1-p_k)a_k^2 = \sum_{k=1}^{K} V_k, \quad (2)$$

where $V_k = 2p_k(1 - p_k)\ a_k^2$. The fractional contribution of locus j to variance is unambiguous and given by

$$\gamma_j = \frac{V_j}{\sum_{k=1}^{K} V_k};\ j = 1, 2, \ldots, K. \quad (3)$$

If the loci are in LD, the variance decomposition is more involved because the joint distribution of genotypes is no longer trivial, due to the existence of covariances between genotypes at different pairs of loci. In this case

$$\begin{aligned} \mathrm{Var}(u_i) &= 2\sum_{k=1}^{K} p_k(1-p_k)a_k^2 \\ &\quad + 2\sum_{k=1}^{K}\sum_{l=k+1}^{K} Cov(W_{ik}, W_{il})a_k a_l \\ &= 2\sum_{k=1}^{K} p_k(1-p_k)a_k^2 \\ &\quad + 2\sum_{k=1}^{K}\sum_{l=k+1}^{K} \rho_{kl}\sqrt{p_k(1-p_k)p_l(1-p_l)}a_k a_l \\ &= 2\sum_{k=1}^{K} p_k(1-p_k)a_k^2 + 2\sum_{k=1}^{K}\sum_{l=k+1}^{K} 2D_{kl}a_k a_l. \end{aligned} \quad (4)$$

Above, ρ_{kl} is the correlation between genotype codes at loci k and l and D_{kl} is the covariance from gametic disequilibrium between these two loci (e.g., Lewontin 1988). Note that

$$\rho_{kl} = \frac{2D_{kl}}{\sqrt{p_k(1-p_k)p_l(1-p_l)}}.$$

Formula (4) is well known and it appears, for example, in Hill and Robertson (1966); Avery and Hill (1979) and

Lynch and Walsh (1998). An important consequence of LD is that the variability no longer breaks into K components of variance, as opposed to (2). In a path analytic or network contexts, any given locus would be connected to the additive genetic value u via a direct effect and by indirect effects mediated by all other loci with which the focal locus is in LD. This begs the question: how much variance is contributed by a locus, say, k, when LD is prevalent?

The problem of variance partitioning when the several random factors that affect some response variable are correlated has received much attention in applied statistics, particularly in the context of breaking down variance into hereditary and environmental components for traits such as intelligence tests in humans. Here, considerable debate has focused around the possible existence of a correlation between random genetic and environmental circumstances that is very difficult to take into account in statistical analysis (e.g., Emigh 1977; Goldberger 1977; Kempthorne 1978; Lewontin et al. 1984). For example, Emigh (1977) discussed the partitioning of sums of squares in non-orthogonal analysis of variance settings and suggested a term he called "commonality". For a 2-factor layout this was defined as $C(A, B) = R(A, B) - R(A) - R(B)$, denoting some "joint" contribution of the factors to a sum of squares; $R(A, B)$ is the sum of squares "due to" fitting A and B, and $R(A)$ and $R(B)$ are the sums of squares "due to" fitting either A or B only (Searle 1971). The counterpart of this in a random effects treatment of the factors is clearly $C(A, B) = \sigma^2_{A+B} - \sigma^2_A - \sigma^2_B = 2Cov(A, B)$. Emigh (1977) suggested that

$$\frac{\sigma^2_A + Cov(A,B)}{\sigma^2_{A+B}} = \frac{\sigma^2_A + \frac{C(A,B)}{2}}{\sigma^2_{A+B}}$$

might provide a sensible measure of the total fraction of variance "due to" factor A. Kempthorne (1978) mentioned this metric without discussion and criticized the use of the term "due to", although his arguments were mostly directed to fallacies in causal interpretation rather than to the measure itself.

We adopt this framework here, regroup the covariances resulting from LD and rearrange (4) in the following manner:

$$\mathrm{Var}(u) = \sum_{k=1}^{K} C_k \tag{5}$$

where, for any j ($j = 1, 2, ..., K$)

$$\begin{aligned} C_j = {} & \rho_{j1}\sqrt{p_j(1-p_j)p_1(1-p_1)}a_j a_1 \\ & + \rho_{j2}\sqrt{p_j(1-p_j)p_2(1-p_2)}a_j a_2 + \ldots + 2p_j(1-p_j)a_j^2 \\ & + \ldots + \rho_{jK}\sqrt{p_j(1-p_j)p_K(1-p_K)}a_j a_K \end{aligned} \tag{6}$$

is the net contribution of locus j to additive genetic variance; this follows from inspection of (4). Observe that $C_j = V_j + \Delta_j$ where Δ_j is a disequilibrium term that can be positive or negative, depending on the net effect of the correlations between alleles at locus j with those at different loci and of the additive genetic effects. Further, while $C_1 + C_2 + ... + C_K$ must be positive (because this sum is a variance), any of the C_j can take negative values. Also, observe that if alleles are fixed at locus j ($p_j = 0$ or 1), $C_j = 0$ irrespective of the existence of polymorphisms at other loci, as one would expect.

Letting the variance under LE be

$$\mathrm{Var}_{EQ}(u) = \sum_{k=1}^{K} V_k, \tag{7}$$

one can define the disequilibrium measure

$$D_{\mathrm{diseq}} = \mathrm{Var}(u) - \mathrm{Var}_{EQ}(u) = \sum_{k=1}^{K}(C_k - V_k) = \sum_{k=1}^{K}\Delta_k. \tag{8}$$

The sign of D_{diseq} depends on whether the net contribution of LD to variance is negative or positive, respectively. Also,

$$\frac{D_{\mathrm{diseq}}}{\mathrm{Var}(u)} = 1 - \frac{\mathrm{Var}_{EQ}(u)}{\mathrm{Var}(u)}, \tag{9}$$

expresses the relative contribution of disequilibrium to variance: if LD increases variance relative to the equilibrium situation, this measure is positive; otherwise, it is negative. Further,

$$\lambda_{\mathrm{eq,j}} = \frac{V_j}{\mathrm{Var}(u)}; \quad j = 1, 2, \ldots, K \tag{10}$$

and

$$\lambda_{\mathrm{dis,j}} = \frac{C_j}{\mathrm{Var}(u)} = \frac{V_j + \Delta_j}{\mathrm{Var}(u)}; \quad j = 1, 2, \ldots, K, \tag{11}$$

represent the frational contribution of a locus to variance assuming equilibrium or taking disequilibrium into account in the sense of (6).

As a simple illustration consider a 3-locus model with same allelic frequency p and additive effect a at each locus. Then (4) is

$$\begin{aligned} \mathrm{Var}(u_i) &= 2p(1-p)a^2(3 + \rho_{12} + \rho_{13} + \rho_{23}) \\ &= 6p(1-p)a^2(1+\rho), \end{aligned}$$

where ρ is the average of the three possible correlations. Here, $\mathrm{Var}_{EQ}(u) = 6p(1-p)\ a^2$ and $D_{\mathrm{diseq}} = 6p(1-p)\ a^2\rho$. Further

$$\begin{aligned} V_1 &= 2p(1-p)a^2; \\ C_1 &= [2 + \rho_{12} + \rho_{13}]p(1-p)a^2; \\ \lambda_{\mathrm{eq},1} &= \frac{1}{3(1+\rho)}, \end{aligned}$$

and

$$\lambda_{\text{dis},1} = \frac{2 + \rho_{12} + \rho_{13}}{6(1 + \rho)}.$$

If $\rho_{12} + \rho_{13}$ is replaced by 2ρ (for illustrative purposes), then $\lambda_{\text{dis},1} = 0.33$, and each locus is assessed with an equal relative contribution to variance, whereas $\lambda_{\text{eq},1}$ understates the contribution of the locus to variability if disequilibrium is positive, but makes an overstatement if disequilibrium is negative. Clearly C_j provides a more appealing metric than V_j.

A matrix representation as in Gianola et al. (2009) is more compact. The genetic value (1) can be written as $u_i = \mathbf{w}_i'\mathbf{a}$, where $\mathbf{a} = \{a_j\}$ is a $K \times 1$ column vector containing the additive genetic effects of each of the loci, and $\mathbf{w}_i'$ is a random row vector containing the genotype indicator variables W_{ij}. It follows that

$$\text{Var}(u_i) = \mathbf{a}'\mathbf{M}\mathbf{a}, \tag{12}$$

where $\mathbf{M} = Cov(\mathbf{w}_i, \mathbf{w}_i')$ is a positive-definite matrix of order $K \times K$ having diagonal elements $2p_j(1 - p_j)$ and off-diagonals $\rho_{jl}\sqrt{p_j(1-p_j)p_l(1-p_l)}$. If there is complete pair-wise LE all correlations are null and (12) returns the equilibrium variance

$$\text{Var}_{EQ}(u) = \mathbf{a}'\mathbf{E}\mathbf{a},$$

where $\mathbf{E} = Diag\{2p_j(1 - p_j)\}$. Then

$$D_{\text{diseq}} = \mathbf{a}'(\mathbf{M} - \mathbf{E})\mathbf{a}$$

where $\mathbf{M} - \mathbf{E}$ has null diagonal elements. From (6) it can be noted that $C_j = a_j\mathbf{m}_j'\mathbf{a}$, where $\mathbf{m}_j'$ is the jth row of matrix $\mathbf{M}$; also, observe that

$$\mathbf{M} = 2\mathbf{P}\mathbf{R}\mathbf{P}, \tag{13}$$

where $\mathbf{P} = Diag\left\{\sqrt{p_j(1-p_j)}\right\}$ and $\mathbf{R}$ is a $K \times K$ correlation matrix with off-diagonal elements $r_{kl} = \frac{\rho_{kl}}{2}$, and this is the correlation between alleles at the two loci in question, as in standard LD analysis (Hill and Robertson 1968; Hedrick 1987; Lewontin 1988).

Clearly, the additive genetic variance is defined only if $\mathbf{M}$ is positive-definite and there is a huge number of combinations of mutation, selection, migration, and drift scenarios that can be thought of as candidates for producing a certain covariance structure. A different matter is that of estimating $\mathbf{M}$ from real data. For example, if $\mathbf{R}$ is an LD correlation matrix to be estimated from whole-genome allelic frequencies, standard pairwise methods are bound to produce estimates that will not yield a positive-definite $\mathbf{R}$, producing an invalid estimate of $\mathbf{M}$. Unless care is exercised, taking a naïve estimate of $\mathbf{R}$ could result in an invalid statement of genetic variance with absurd attributions of contributions of individual loci to variance. This point will be retaken later on.

Factors affecting the contribution of a locus to variance

Linkage disequilibrium and number of loci. Expression (6) indicates that, apart from the number of loci, C_j depends on the a effects at all K loci (if most of the effects are either negative or positive, more variance due to disequilibrium would be expected than when their distribution is symmetric), on the distribution of allelic frequencies over loci and on the extent of LD as conveyed by the off-diagonals of $\mathbf{M}$. It is awkward to address the influences of all these factors analytically, but simple simulations serve to provide an idea of the extent to which LD affects the standard measure of an individual locus contribution to variance (V_j) , as well as to compare this with the metric C_j, which we argue provides a more sensible measure.

While conjectures about the "genetic architecture" of quantitative traits are abundant (e.g., Daetwyler at al. 2010), it does not seem unfair to state that the number, effects, mode of gene action, and joint distributions of genotypes or alleles at QTL affecting complex traits remains largely unknown, in spite of a deluge of genomic data. The same holds for evolutionary processes. Hence, simulating genetic systems requires strong and largely untested assumptions about the state of nature and about causation. Any hypothetical evolutionary or breeding scenario will lead to a valid LD structure, but what is the strength of the evidence favoring one scenario over another? At present, this cannot be answered, at least for "complex" traits. For this reason, a purely statistical approach to the study of factors affecting the attribution of variance to a locus is taken here. That is, we examine processes leading to certain net results without arguing from a mechanistic perspective in defense of the statistical setting chosen.

A main difficulty is that of simulating LD settings leading to a positive-definite matrix $\mathbf{M}$. This is essential, as inducing pairwise correlations in a naive manner without ensuring positive-definiteness can produce absurd results, such as a negative genetic variance. This a well-known problem in multivariate analysis of quantitative traits, e.g., Hayes and Hill (1981). Actually, standard estimates of pairwise disequilibrium via the r^2 and D' measures, typically reported as "heat maps" (e.g., Goddard and Hayes 2009; Wu et al. 2011), will rarely lead to a "proper" matrix $\mathbf{M}$.

We examined three scenarios of linkage disequilibrium. In the first one LD was entirely random and this was attained by simulating random correlation matrices (Marsaglia and Olkin 1984) using function *rcorr* in package *ggm* of the *R* software (Marchetti and Drton 2010). Briefly, if $\mathbf{z}_i \sim N(\mathbf{0}, \mathbf{I})$ is a vector containing K independent $N(0, 1)$ variables, then draw K such vectors and form the K^2 matrix $\mathbf{Z}$, whose *ith* column is $\frac{\mathbf{z}_i}{\sqrt{\mathbf{z}_i'\mathbf{z}_i}}$. It follows that using $\mathbf{R} = \mathbf{Z}\mathbf{Z}'$ in (13) yields a matrix whose diagonal elements

are all equal to 1, and its off-diagonals are between −1 and 1. Subsequent to obtaining the correlation matrices, eigenvalues were calculated as a check for positive-definiteness; this was verified in every single instance.

The second scenario aimed to produce positive LD, and the strength eventually attained depended on the number of loci (K) affecting the genetic value and on the value of a single correlation coefficient ρ. A challenge was to ensure that the simulated $\mathbf{R}^+$ was positive definite, and this is a function of the correlation structure, of the strength of the correlation and of K. As an example, consider a situation with $K = 8$ loci and where $\mathbf{R}^+$ has the banded form (e.g., mimicking LD involving 4 "successive" loci, in the sense of physical position in a chromosome)

$$\mathbf{R}^+ = \begin{bmatrix} 1 & \rho & \rho & \rho & 0 & 0 & 0 & 0 \\ \rho & 1 & \rho & \rho & \rho & 0 & 0 & 0 \\ \rho & \rho & 1 & \rho & \rho & \rho & 0 & 0 \\ \rho & \rho & \rho & 1 & \rho & \rho & \rho & 0 \\ 0 & \rho & \rho & \rho & 1 & \rho & \rho & \rho \\ 0 & 0 & \rho & \rho & \rho & 1 & \rho & \rho \\ 0 & 0 & 0 & \rho & \rho & \rho & 1 & \rho \\ 0 & 0 & 0 & 0 & \rho & \rho & \rho & 1 \end{bmatrix}. \quad (14)$$

Here

$$|\mathbf{R}^+| = 4\rho^8 - 32\rho^7 + 91\rho^6 - 104\rho^5 + 25\rho^4 + 32\rho^3 - 18\rho^2 + 1$$

and, as shown in Fig. 1 (left panel), not all values of ρ produce a positive determinant. Even when this condition is satisfied, a ρ that yields a positive determinant does not ensure positive eigenvalues. For instance, $\rho = 0.5$ produces a valid LD as both the determinant and the 8 eigenvalues are all positive. Further, within the 28 off-diagonal elements of $\mathbf{R}^+$, there are 18 that are not 0, giving an average correlation of $\frac{18}{28}\rho$; at $\rho = 0.5$ this average is about 0.32. With $K = 12$, a similar lag-4 banded structure produces a determinant that is non-negative only when the correlation is either weak, or strongly negative (Fig. 1, right panel); at $\rho = 0.3$, the determinant is 1.22×10^{-3} and $\mathbf{R}^+$ has 30 non-zero elements, so the average correlation is $\frac{30}{66}\rho$; with $\rho = 0.3$, the average correlation is only 0.14. This illustrates that, as K increases, this banded correlation structure yields a weaker simulated LD because the proportion of $0's$ in the off-diagonals grows, yet attaining positive-definiteness. Hence, the lag-4 banded structure produces strong positive LD in models with just a few loci, but not when K is large. In short, given K, the positive-definiteness of $\mathbf{R}^+$ depends on ρ and, conversely, at a given ρ, positive definiteness depends on K. In our simulations we used combinations of K and ρ at varying lags, found by trial and error. Once positive definiteness of $\mathbf{R}^+$ was attained, slight additional random disequilibrium was introduced at times by taking as correlation matrix

$$\mathbf{R} = (1 - \alpha)\mathbf{R}^+ + \alpha\mathbf{Z}\mathbf{Z}', \quad (15)$$

where $\mathbf{Z}\mathbf{Z}'$ is a random correlation matrix generated as described earlier and $0 \leq \alpha \leq 1$, with α typically below 0.05 in the trials. Since a weighted average (with positive weights) of two positive-definite matrices is also positive definite, this provided a proper $\mathbf{R}$ for the purpose of this study.

The same approach was followed for negative disequilibrium. Here, the correlation matrix was formed with $\alpha = 0.01$ so that

$$\mathbf{R} = 0.99 \times \mathbf{R}^- + 0.01 \times \mathbf{Z}\mathbf{Z}' \quad (16)$$

where $\mathbf{R}^-$ was a banded matrix similar to (14), employing values of ρ leading to a positive-definite $\mathbf{R}^-$. For example, with $K = 8$ a choice of $\rho = -0.20$ meets the requirements for $\mathbf{R}^-$; here, there are 18 non-zero off-diagonals (out of 36) in either the upper or lower triangles of $\mathbf{R}^-$, producing an average correlation $\frac{18}{28} \times (-0.20) = -0.13$.

Distribution of allelic frequencies and of genetic effects over loci. Since additive genetic variance depends on allelic frequencies, instead of setting arbitrary values of p these were drawn from either uniform $U(0,1)$ or beta, $Beta(\gamma_1, \gamma_2)$, distributions. The latter were J − shaped ($\gamma_1 = 1$, $\gamma_2 = 0.2$), L-shaped ($\gamma_1 = 0.2, \gamma_2 = 1$) or inverted U-shaped ($\gamma_1 = \gamma_2 = 2$).

As noted, the additive effects a are not random variables in the standard quantitative genetics setting (Falconer and Mackay 1996); however, their values were simulated by effecting K draws from either a normal distribution with arbitrarily chosen mean and variance σ^2, or from a double exponential (DE) distribution with the same mean and parameter λ; elicited by setting the variance of this distribution, $2\lambda^2$, equal to σ^2 so that $\lambda = \sqrt{\frac{\sigma^2}{2}}$.

We also examined a situation where the additive effect a depended on the allelic frequency at the locus. Studies on relationships between effects of quantitative trait loci and their frequencies using real data are lacking. One could either adopt a stylized model (e.g., Zhang et al. 2002) that leads to tractable mathematics but without being necessarily relevant to the underlying complexity of the trait(s) in question or adopt a model that does not favor any theory in particular. We adopted the second viewpoint and built an arbitrary relationship where the additive effect at locus j was generated using the function

$$a_j(p_j) = w_j(p_j) + \frac{1}{2}v_{1,j} + \frac{1}{2}v_{2,j} \quad (17)$$

where $v_{1,j} \sim N(\mu_1, \sigma^2)$ and $v_{2,j} \sim N(\mu_2, \sigma^2)$ are two independent normally distributed deviates. Above, $w_j(p_j) = 4 + 4p_j + \sin(15p_j) + \cos(15p_j)$ is a frequency-dependent

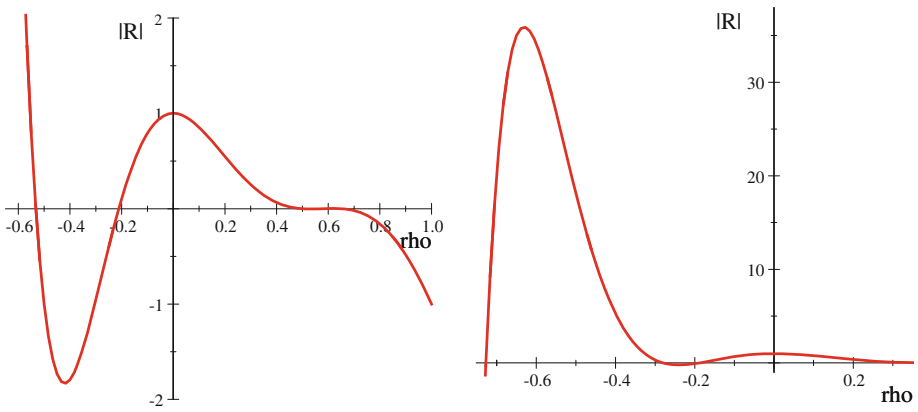

Fig. 1 Determinant of a correlation matrix with a lag-4 banded structure as a function of (rho), the coefficient of correlation and K the number of loci. $K = 8$ (*left panel*). $K = 12$. (*right panel*)

sinusoidal "wave", and the sum of the two normal deviates produces a residual ("away from the wave") having as distribution a 50–50 mixture of normals. Adopting a process that does not favor any specific pet theory about the state of nature (especially if this is far from being firmly established) is common in statistical practice. For example, Newton et al. (2001) used this approach when evaluating a suite of estimators of differential gene expression.

The wave $w_j(p_j)$ is illustrated in the left panels of Fig. 2; the figure used 2,000 draws from a uniform $U(0, 1)$ distribution of frequencies (top panel), or from a *Beta*(2, 2) distribution (inverted *U-shaped*) in the bottom panel. The corresponding genetic values are in the right panels: sinusoidal genetic values are seen more clearly when allelic frequencies are distributed uniformly over the 2,000 loci. The density of the distribution of allelic substitution effects is

$$f(a) = \int g(a|p)h(p)dp, \tag{18}$$

where $g(a|p)$ is the density of the conditional distribution of the genetic values given the allelic frequency, and $h(p)$ is the density of the allelic frequency distribution. Given the allelic frequencies, the only random term in (17) is $\frac{1}{2}v_{1,j} + \frac{1}{2}v_{2,j}$. Thus, $g(a|p_j)$ is a mixture of normals, with expected value

$$E\left(a_j(p_j)|p_j\right) = \frac{\mu_1 + \mu_2}{2} + w_j(p_j),$$

and variance $\frac{\sigma^2}{2}$. Moments or the density $f(a)$ cannot be written in closed form and, to illustrate, this density was estimated non-parametrically assuming $\mu_1 = 4$, $\mu_2 = -4$ and $U(0, 1)$ or *Beta*(2, 2) as allelic frequency distributions. For this purpose, 20,000 draws were obtained from these distributions, with (17) evaluated to produce samples from the marginal distribution of genetic values a. The distributions were bimodal and the distribution of allelic frequencies did not make a difference (results not shown).

Results

Many arbitrarily chosen settings were investigated; salient ones serving purposes of the study are reported.

3-locus model

The model had 3 loci and positive linkage disequilibrium inducing the correlation matrix

$$\mathbf{R}^{+} = \begin{bmatrix} 1 & 0.8 & 0.6 \\ 0.8 & 1 & 0.8 \\ 0.6 & 0.8 & 1 \end{bmatrix}$$

This matrix is positive-definite. Assuming that the three loci had allelic frequencies $0.5, 0.5 + \Delta$ and $0.5 + 2\Delta$, where $-0.25 < \Delta < 0.25$, and the same additive effect a, one can write using (5) $Var(u) = C_1 + C_2 + C_3$, where

$$C_1 = \left[2 \times 0.5^2 + 0.8 \times 0.5\sqrt{\left(0.25 - \Delta^2\right)} + 0.6 \times 0.5\sqrt{\left(0.25 - 4\Delta^2\right)}\right]a^2,$$

$$C_2 = \left[0.8 \times 0.5\sqrt{\left(0.25 - \Delta^2\right)} + 2\left(0.25 - \Delta^2\right) + 0.8\sqrt{\left(0.25 - \Delta^2\right)\left(0.25 - 4\Delta^2\right)}\right]a^2,$$

and

$$C_3 = \left[0.6 \times 0.5\sqrt{\left(0.25 - 4\Delta^2\right)} + 0.8\sqrt{\left(0.25 - 4\Delta\right)^2\left(0.25 - \Delta^2\right)} + 2\left(0.25 - 4\Delta^2\right)\right]a^2.$$

Likewise,

$$V_1 = 2 \times 0.5^2 a^2;\ V_2 = 2\left(0.25 - \Delta^2\right)a^2;\ V_3 = 2\left(0.25 - 4\Delta^2\right)a^2.$$

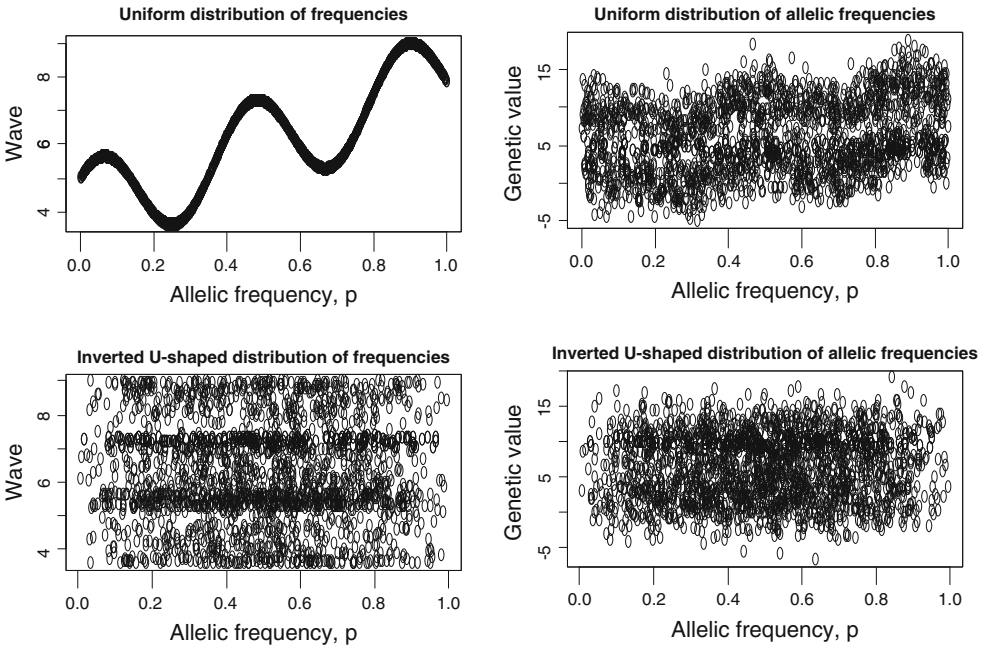

Fig. 2 Sinusoidal wave (2,000 points in the plot) used in creating frequency-dependent genetic effects for two distributions of allelic frequencies: *U*(0, 1) in the *top left panel* and *Beta*(2, 2) in the *bottom left panel*. Corresponding genetic values are in the *right-side panels*

Then

$$\lambda_{\mathrm{eq},j} = \frac{V_j}{C_1 + C_2 + C_3}; j = 1, 2, 3,$$

and

$$\lambda_{\mathrm{dis},j} = \frac{C_j}{C_1 + C_2 + C_3}; j = 1, 2, 3.$$

The relative contributions $\lambda_{\mathrm{eq},j}$ and $\lambda_{\mathrm{dis},j}$ of the three loci to variance were plotted against Δ, as shown in Fig. 3 (left panel). The picture was clear: because LD was positive and strong, the standard formula based on V_j produced a severe understatement of the contribution of any of the three loci to genetic variability. For example, in the case of locus 3, its maximum contribution, as deemed by V_j, is attained when $\Delta = 0\,(p = 0.5)$, at nearly 20 % of the variance (dotted green line). However, this locus makes a contribution of at most 30–31 % of the total genetic variance at frequencies near $p = 0.35$ when indirect contributions stemming from LD (as conveyed by C_j) are taken into account. Importantly, note that while equilibrium formulae suggest that locus 1 is the most important contributor to variance at most allelic frequencies (dotted black line), this is not so when both direct and indirect effects of a locus are brought into the picture. For example, the relative importance of loci 1 and 2 crisscross and locus 2 (solid red line) is the main contributor to variance at intermediate frequencies, but no so at other values of p.

Consider a negative disequilibrium case, with correlation structure

$$\mathbf{R} = \begin{bmatrix} 1 & -0.7 & -0.3 \\ -0.7 & 1 & -0.2 \\ -0.3 & -0.2 & 1 \end{bmatrix},$$

with the three loci having the same additive effect. The eigenvalues are {1.708, 1.140, 0.152} and V_1,V_2 and V_3 are as before. Now

$$C_1 = \left[2 \times 0.5^2 - 0.7 \times 0.5\sqrt{(0.25 - \Delta^2)} - 0.3 \times 0.5\sqrt{(0.25 - 4\Delta^2)}\right] a^2,$$

$$C_2 = \left[-0.7 \times 0.5\sqrt{(0.25 - \Delta^2)} + 2(0.25 - \Delta^2) - 0.2\sqrt{(0.25 - \Delta^2)(0.25 - 4\Delta^2)}\right] a^2,$$

and

$$C_3 = \left[-0.3 \times 0.5\sqrt{(0.25 - 4\Delta^2)} - 0.2\sqrt{(0.25 - 4\Delta)^2(0.25 - \Delta^2)} + 2(0.25 - 4\Delta^2)\right] a^2.$$

Figure 3 (right panel) depicts the relative importance of these three loci in terms of contribution to variance. The equilibrium formulae now overstate the relative importance of loci 1 and 3, but slightly understate the contribution of locus 2 to variance. In this setting, negative disequilibrium results in negative contributions of locus 3 to variance at allelic frequencies that are approximately larger than 0.72

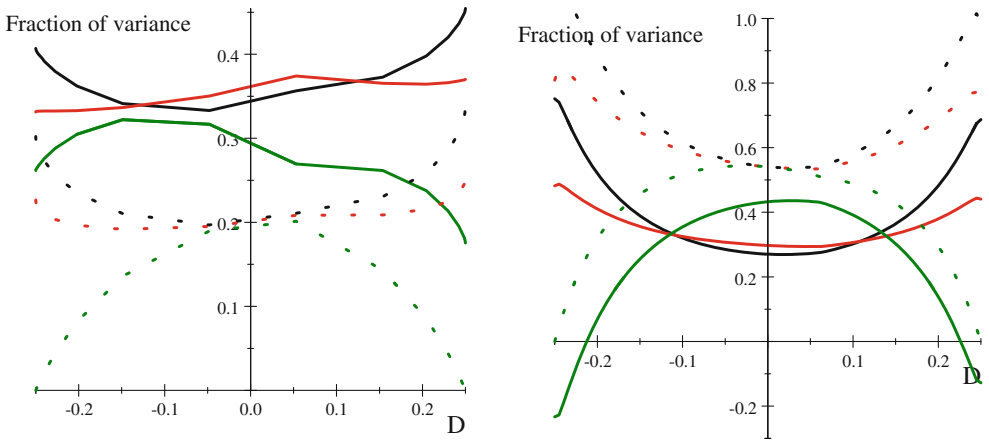

Fig. 3 Relative contribution to variance of three loci under positive (*left panel*) or negative (*right panel*) LD; *dotted lines* give the contributions as deemed by equilibrium formulae. Locus 1 *black*, Locus 2 *red*, Locus 3 *green*. *D* departure of allelic frequency from 0.5 (color figure online)

or smaller than about 0.28. The effect of negative disequilibrium on total variance results in a re-ranking of loci.

Several loci

The setting reported had $K = 12$ loci and random, positive or negative LD, with $|\rho| = 0.40$. Since only a few loci are involved, the forms of the distribution of additive genetic effects and of frequencies over loci are unimportant. However, the proportion of loci with either positive or negative a effects does matter. For instance, if 50 % of these effects are positive and 50 % are negative, there is little build up of disequilibrium variance, as they cancel with each other. On the other hand, if most of the effects are either positive or negative, the relative contribution of D_{diseq} becomes patent. Here, we drew a effects from the normal distribution $N(2, 9^2)$, where about 59 % of the values are expected to be positive.

The $\mathbf{R}^+$ and $\mathbf{R}^-$ matrices used in the positive and negative LD settings, respectively, had a lag-6 type of structure, e.g., $\mathbf{R}^+[1, 6] = 0.40$ and $\mathbf{R}^+[1, 7] = 0$, with $\mathbf{R}^-[1, 2] = -0.40$ and $\mathbf{R}^-[1, 10] = 0$; these are positive-definite matrices. These matrices were blended with a random correlation matrix, as indicated in (15) and (16).

The random LD setting produced $\text{Var}(u) = 59.19$, $\text{Var}_{EQ}(u) = 62.81$ and $D_{\text{diseq}} = 59.19 - 62.81 = -3.62$ so that some mild random negative disequilibrium was created, with $\frac{D_{\text{diseq}}}{\text{Var}(u)} = -0.06$. When LD was positive, $\text{Var}(u) = 86.77$, $\text{Var}_{EQ}(u) = 62.81$ as before, and $D_{\text{diseq}} = 23.96$, with $\frac{D_{\text{diseq}}}{\text{Var}(u)} = 0.28$. For negative LD, $\text{Var}(u) = 58.69$, $\text{Var}_{EQ}(u) = 62.81$ and $D_{\text{diseq}} = -4.12$, with this parameter equivalent to -7 % of the variance. Estimates of the slopes of the regression of $\lambda_{\text{dis},j}$ on $\lambda_{\text{eq},j}$ were calculated. With positive LD, the standard formula for the relative contribution of a locus to variance understated the actual contribution (in the sense used in this paper) by about 14 % ($b = 1.16$), as measured by the slope of the regression; when LD was negative, there was no overstatement observed because the simulation did not generate sizable disequilibrium variance. Since LD levels in animals and plants subject to selection are typically stronger than those examined here, these results probably represent lower bounds for the relative understatement or overstatement of the contribution of a locus to variance, provided that the distribution of effects is not symmetric.

Many loci: frequency-independent effects

We report on a setting with $K = 80$ loci and with combinations of distributions of allelic frequencies (uniform or beta), additive effects (normal or double exponential) and type of LD (random, positive or negative). We used either a $N(3, 9^2)$ distribution of additive effects, yielding about 63 % positive realizations, or a double exponential process with mean 3 and parameter $\lambda = \sqrt{\frac{9}{2}}$, producing a variance equal to 9. In the latter case, about 90 % of the realizations were expected to be positive.

For positive LD, a correlation of 0.23 between pairs of "contiguous" loci with a lag of 8 in the banded correlation structure yielded a positive-definite $\mathbf{R}^+$; because the average of its off-diagonal elements was only 0.039, thus resulting in weak overall LD, this matrix was used in lieu of $\mathbf{R}$ ($\alpha = 0$). Otherwise, the disequilibrium contribution to variance would have been essentially zero. For negative LD, it was found that a lag-4 correlation of -0.16 produced a positive-definite $\mathbf{R}^-$, translating into an overall average correlation of -0.01; $\mathbf{R}^-$ was used instead of $\mathbf{R}$ as well. Although these two settings produced little disequilibrium variance (see Table 1), they can be construed as providing a lower bound for the effects of LD on variance partitioning.

Table 1 Additive genetic variance, Var(u); equilibrium additive variance, $\mathrm{Var}_{EQ}(u)$, and relative contribution of disequilibrium, D_{diseq}, to genetic variance at random (R), positive (+) and negative (−) disequilibrium under normal $N(3, 9^2)$ or double exponential (DE) distribution of effects; the latter with mean 3 and variance 9

	Effects⟹	N	N	N	DE	DE	DE
Frequencies ⇓	Disequilibrium⟹	R	+	−	R	+	−
Uniform	Var (u)	307	385	277	455	616	385
	$\mathrm{Var}_{EQ}(u)$	305	305	305	442	442	442
	$100\frac{D_{\mathrm{diseq}}}{\mathrm{Var}(u)}$	0.19	20.9	−10.2	2.9	28.2	−14.8
Inverted U	Var (u)	358	460	332	521	700	438
	Var $_{EQ}(u)$	364	364	364	503	503	503
	$100\frac{D_{\mathrm{diseq}}}{\mathrm{Var}(u)}$	−1.6	20.9	−9.5	3.3	28.1	−15.0
J	Var (u)	132	154	127	167	210	158
	Var $_{EQ}(u)$	135	135	135	170	170	170
	$100\frac{D_{\mathrm{diseq}}}{\mathrm{Var}(u)}$	−2.7	12.0	−6.28	−1.4	19.1	−11.2
L	Var (u)	158	180	149	194	231	171
	Var $_{EQ}(u)$	155	155	155	186	186	186
	$100\frac{D_{\mathrm{diseq}}}{\mathrm{Var}(u)}$	1.8	13.5	−4.5	4.1	19.8	−8.6

The number of loci is 80 and allelic frequency distributions are uniform, $U(0, 1)$; inverted U-shaped, $Beta(2, 2)$; J = shaped, $Beta(1, 0.20)$, and L-shaped, $Beta(0.20, 1)$

As shown in Table 1, for random LD the contribution of D_{diseq} to variance ranged from −1.6 to 4.1 %; when it was positive, this parameter accounted for 12 to 28 % of the variability, and the disequilibrium parameter was relatively larger for the DE than for the normal distribution because of a larger proportion of positive realizations. With negative LD, parameter D_{diseq} represented from about −15 to −4.5 % of the variability; again the disequilibrium contribution was stronger for the DE distribution of effects. Overall, the uniform and the inverted-U distributions of allelic frequencies tended to produce stronger disequilibrium than the other two beta distributions, this being due to the fact that frequencies near 0.5 are more rare under the J- and L-shaped beta processes.

Plots of (11) versus (10), as percentages, when the distribution of effects was normal or DE were made both for positive and negative LD. Plots for the double exponential distribution are in Fig. 4 and 5 for positive and negative LD, respectively. The slopes of the regression of (11) on (10) were calculated for each case. When LD was positive, the contribution of a locus to variability was understated (the slope "b" ranged between 1.08 and 1.23), and more markedly so when allelic frequencies were uniform or inverted U, because these settings produced stronger LD. When LD was negative, the equilibrium expressions overstated the "importance" of a locus, with b ranging from 0.91 to 0.99. As expected, when LD was random scatterplots (not shown) did not reveal departures from the 45° angle line, although the slopes were slightly below 1 and slightly larger than 1 for the DE situation; the effect of the distribution of allelic frequencies on the slopes was nil.

Many loci: frequency-dependent effects

The setting had $K = 80$ loci, the same disequilibrium structure as in the preceding section, a uniform distribution of allelic frequencies and additive genetic effects were generated as in (17) with $v_{1,j} \sim N(-1, 2.25^2)$ and $v_{2,j} \sim N(2, 2.25^2)$. For the DE

$$a_j(p_j) = 4 + 4p_j + \sin(15p_j) + \cos(15p_j) + \frac{1}{2}v'_{i,j} + \frac{1}{2}v'_{2,j}, \tag{19}$$

where $v'_{1,j} \sim DE(-1, 2.25^2)$ and $v'_{2,j} \sim DE(2, 2.25^2)$, so that $\lambda = 2.25\sqrt{\frac{1}{2}}$. Results are shown in Fig. 6, with similar qualitative results as for the frequency-independent situation: ignoring positive (negative) LD results in an understatement (overstatement) of the contribution of an individual locus to variability. Results obtained when using either J or L-shaped distribution of allelic frequencies led to the same conclusions, but with milder effects of LD on attribution of variance to a locus.

Discussion

Our study discussed factors affecting the partition of additive variance for a quantitative trait into locus-specific

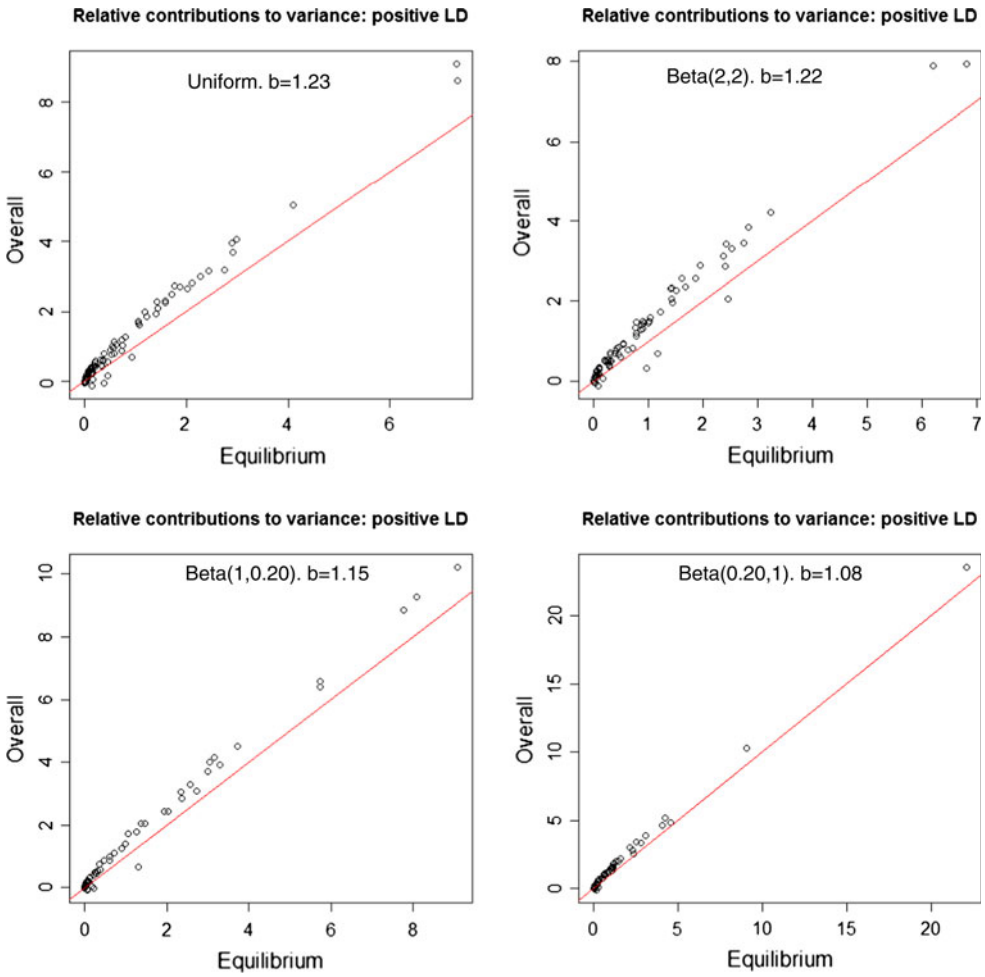

Fig. 4 Plots of relative contributions to variance considering (*y*-axis) and ignoring (*x*-axis) positive LD. The number of loci is 80; the distribution of effects is double exponential. Allele frequencies are uniform or Beta(γ_1, γ_2). Slope of the regression of λ_{dis} on λ_{eq} represented as *b*

components, when linkage disequilibrium exists. Factors included the extent and type (random, positive or negative) of LD, the distribution of allelic frequencies (e.g., *J*-shaped or *L*-shaped) and the distribution of additive effects over loci. As one would expect, the attribution of variance to a given locus is overstated by the usual equilibrium formula when LD is predominantly negative, and understated when LD is positive. Linkage disequilibrium was created randomly or by producing banded correlation structures over pairs of contiguous loci. The settings were arbitrary and used primarily to provide a proof of concept, as opposed to proposing structural models for the analysis of correlation matrices stemming from gametic disequilibrium. An alternative would have been to simulate some evolutionary or selective process leading to a predictable LD structure. A difficulty is that there is a huge number of possible scenarios reflecting population size, drift, selection, mutation and demographic structure, and any choice of setting would have been no less arbitrary than the approach followed here. A related issue is the technical difficulty of retrieveing a positive-definite estimate of **R** from highly dimensional genomic data. In a nutshell, the point of this paper was to llustrate the effects of net negative or positive disequilibrium on variance attribution, without reference to why such structure arose.

There may be finer ways of studying the effect of LD on genetic variability. Here, we created LD statistically via the positive-definite matrix **R** (**R**$^+$ or **R**$^-$, depending on whether LD was positive or negative). For example, LD can be modeled in terms of some latent variable that is linear on effects after suitable transformation, e.g., as inTurelli and Barton (1990); Hospital (1992) and Barton (2000). In a randomly picked gamete the latent variable could be expressed as

$$l_{ij} = \mu + c_i + A_{ij}^{[p]} + A_{ij}^{[m]},$$
$$l_{i'j'} = \mu + c_{i'} + B_{i'j'}^{[p]} + B_{i'j'}^{[m]},$$

where c_i is a random effect due to chromosome i and $A_{ij}^{[p]}$, and $A_{ij}^{[m]}$ represent effects due to paternal (p) and maternal (m) origin of alleles at randomly chosen locus j (A, say), and same for $B_{i'j'}^{[p]}$ and $B_{i'j'}^{[m]}$. One could assume, for example

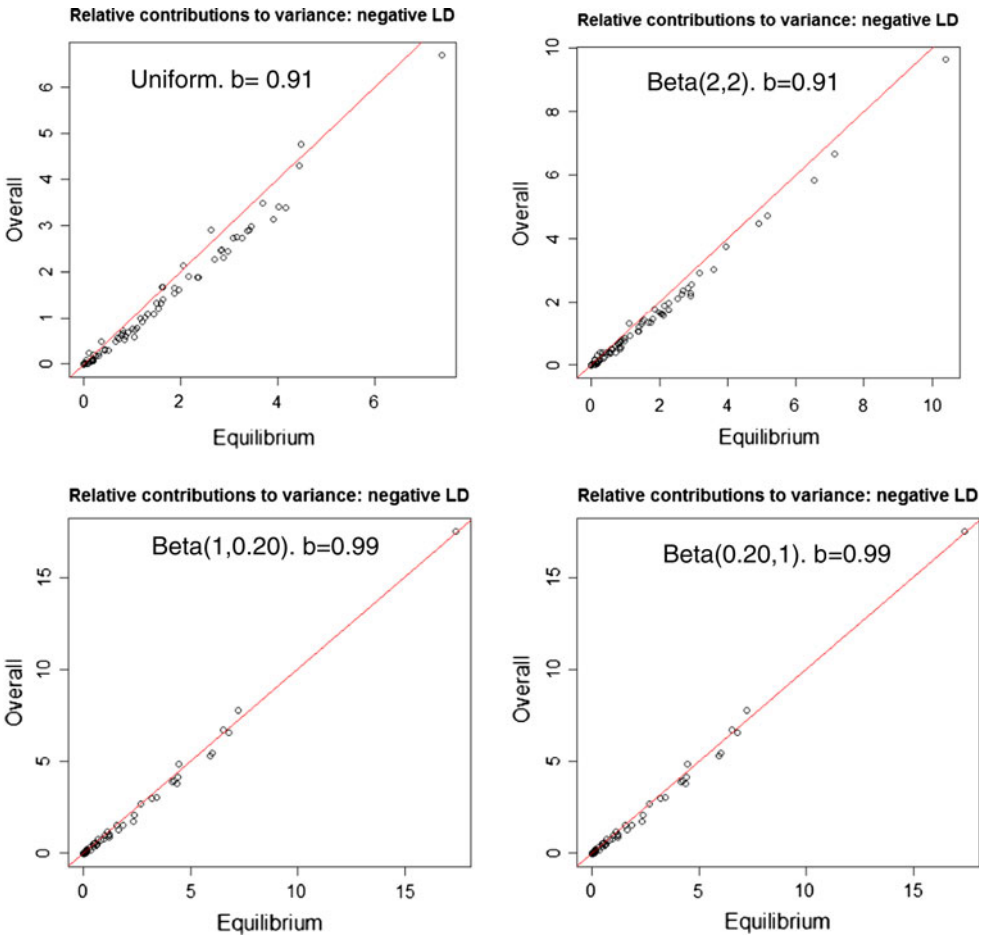

Fig. 5 Plots of relative contributions to variance considering (*y*-axis) and ignoring (*x*-axis) negative LD. The number of loci is 80; the distribution of effects is double exponential. Allele frequencies are uniform or Beta(γ_1, γ_2). Slope of the regression of λ_{dis} on λ_{eq} represented as b

$$Cov(l_{ij}, l_{i'j'}) = Cov(c_i, c_{i'}) + Cov\left(A_{ij}^{[p]}, B_{i'j'}^{[p]}\right) + Cov\left(A_{ij}^{[p]}, B_{i'j'}^{[m]}\right) + Cov\left(A_{ij}^{[m]}, B_{i'j'}^{[p]}\right) + Cov\left(A_{ij}^{[m]}, B_{i'j'}^{[m]}\right),$$

with

$$Cov(c_i, c_{i'}) = \begin{cases} \sigma_c^2 & \text{if } i = i' \\ \rho_c \sigma_c^2 & \text{if } i \neq i' \end{cases},$$

$$Cov\left(A_{ij}^{[p]}, B_{i'j'}^{[p]}\right) = \begin{cases} \sigma_p^2 & \text{if } i = i' \\ \rho_p \sigma_p^2 & \text{if } i \neq i' \end{cases},$$

$$Cov\left(A_{ij}^{[m]}, B_{i'j'}^{[m]}\right) = \begin{cases} c\sigma_m^2 & \text{if } i = i' \\ \rho_m \sigma_m^2 & \text{if } i \neq i' \end{cases},$$

and

$$Cov\left(A_{ij}^{[p]}, B_{i'j'}^{[m]}\right) = Cov\left(A_{ij}^{[m]}, B_{i'j'}^{[p]}\right) = \begin{cases} \sigma_{pm,w} & \text{if } i = i' \\ \sigma_{pm,b} & \text{if } i \neq i' \end{cases}.$$

where σ_{pm} is a covariance and w and b represent "within" and "between" chromosomes. This random effects model is indexed by four variance components, σ_c^2 (variance due to chromosomes); σ_p^2 (variance among alleles of paternal origin); σ_m^2 (variance among alleles of maternal origin) and σ_{pm}, covariance due to one of the alleles being of paternal (maternal) origin and the other having maternal (paternal) origin. In addition, three among-chromosome correlations arise: ρ_c, ρ_p, and ρ_m. The issue of how these parameters ought to be estimated remains to be addressed. We note that Barton (2000) did not provide a solution to this problem, although he casted it in a contingency table framework (log-linear models).

In practice, how much a given locus contributes to genetic variability is an important question, one that has become central in the explosion of genome-wide association studies, or GWAS (e.g., Manolio et al. 2009; Stranger et al. 2011). Zuk et al. (2012) argue that the issue may be irrelevant with regard to the relevance of a gene in biology or medicine. As indicated by our study, the answer to our question is not as straightforward as suggested by quantitative genetics texts such as Falconer and Mackay (1996), even in a single locus model. Under multi-factorial inheritance, additional complications are introduced by the fact that genotypes at the intervening loci are correlated due to LD and by how allelic frequencies and effects are

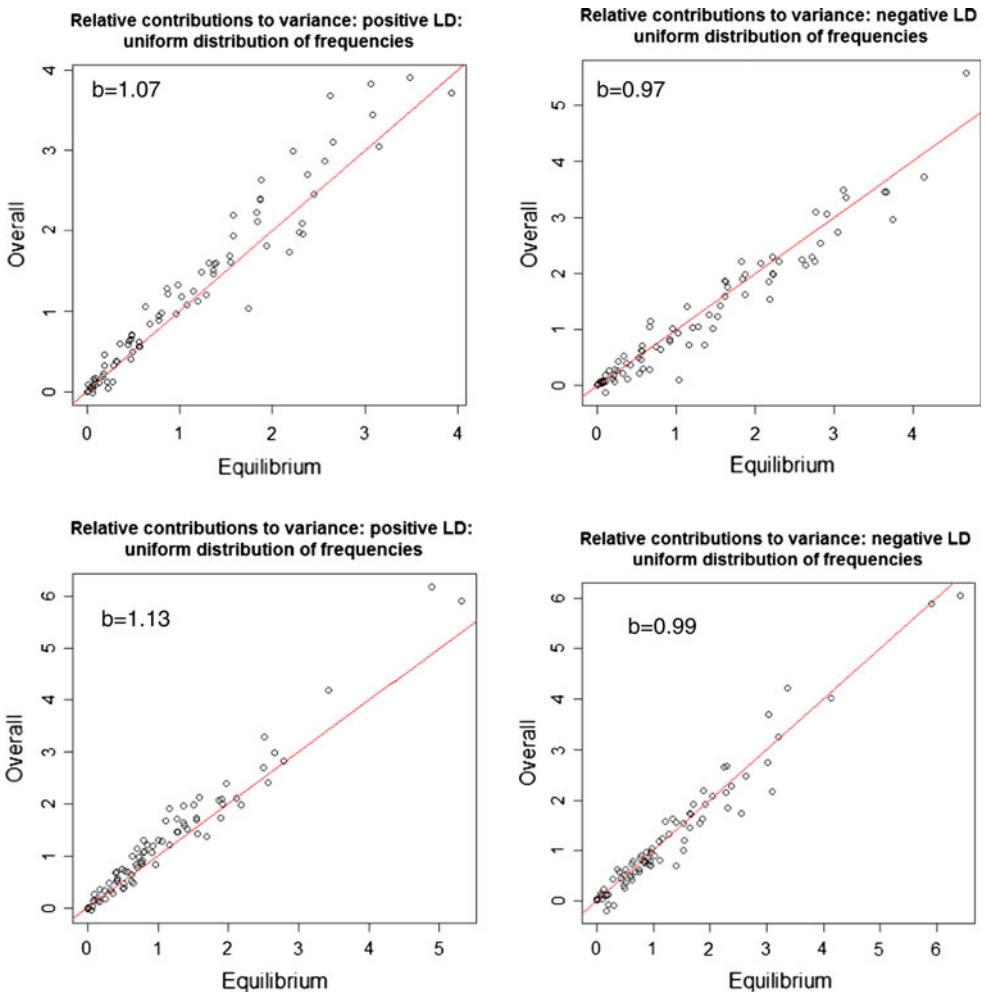

Fig. 6 Plots of relative contributions to variance considering (*y*-axis) and ignoring (*x*-axis) LD. The number of loci is 80; distribution of effects is frequency-dependent (see text), with normal (*top*) or double-exponential (*bottom*) residuals. Slope of the regression of λ_{dis} on λ_{eq} represented as *b*

distributed over loci (Sabbati and Risch 2002; Zhao et al. 2005). We proposed parameter C_j for measuring direct and indirect (through LD) contributions of a locus to genetic variance. It is clear from the form of C_j that how the contribution to variance evolves over time depends not only on forces affecting locus *j*, but on all other loci with which *j* is correlated. We note that C_j in (6) can be inferred from data, e.g., using a "plug-in" method. For example, if a whole-genome additive Bayesian model is fitted to data, the $a's$ can be estimated from the mean of their posterior distributions, and estimation of allelic frequencies is straightforward. The main difficulty is that of obtaining estimates of LD parameters leading to a positive-definite LD structure. As noted above, the estimates obtained from pair-wise statistics do not lead to positive-definiteness and ignore parametric bounds that are hard to establish with high-dimensional SNP data (Svetlana Miller and Henner Simianer, personal communication).

Next we discuss the connection between LD and the single-marker approach typically used in GWAS context and some related estimation issues. The advent of genome-wide markers, such as single nucleotide polymorphisms, has produced thousands of GWAS, where a main objective is that of relating variation for some disease-connected phenotype to variation of marker genotypes. A prototypical GWAS uses naive single-marker regression models, typically linear if the trait is quantitative, or logistic (or probit) if the response is a discrete response. Then, using stringent significance levels a few markers are retained, and sometimes validated in meta-analysis. Stranger et al. (2011) reviewed many such studies and discussed difficulties posed when the traits are suspected to be multi-factorial. This is certainly the case for most economically important characters in plants and animals, and arguably for many diseases in livestock and humans. In connection with a study of rheumatoid arthritis in humans (RA), Stranger et al. (2011) stated:

> "...On the basis of their ORs [odds ratios] and allele frequencies, we can calculate the proportion of phenotypic variance explained in RA for each SNP under a liability threshold model (Falconer and Mackay 1996), and these can be assumed to sum to the total percentage of variance explained by validated RA risk alleles."

Details on how this was done are lacking in their paper, but it is not always obvious how variance components from some generalized linear model (especially if all effects in the explanatory structure are fixed!) translate into variance in some observed scale. Examples are provided by Kathiresan et al. (2008) and Speliotes et al. (2010), who carried out GWAS studies for cholesterol and body mass index, respectively, and found that 18 and 32 loci explained 2–4 and 5–6 % of the variation of the respective traits. These reports are not explicit on how such estimates were arrived at, but it is probable that this was done via single marker regression. In such an approach, the model relates the centered phenotype of subject i (y_i) to the number of copies of a given allele (x_i) at some marker via the relationship $y_i = x_i\beta + e_i$, where β is the allelic substitution effect (corresponding to a in the notation of this paper), and $e_i \sim (0, \sigma^2)$ is a residual with variance σ^2; $i = 1, 2, ..., N$. In an ideal situation $\beta = a$, so that the marker would correspond to a quantitative trait locus (QTL), and suppose this is the only locus affecting the trait. If the regression is estimated by ordinary least-squares, the estimate of additive variance attributed to the locus, assuming that one knows the allelic frequencies without error, is

$$\widehat{V} = 2p(1-p)\widehat{a}^2.$$

This provides an upwardly biased estimator of the variance "due" to the locus, since

$$E\left(\widehat{V}|x\right) = 2p(1-p)\left[a^2 + \frac{\sigma^2}{\sum x_i^2}\right].$$

Thus, even when the model is "true", the standard assessment exaggerates the contribution of the locus to variability, unless the sample is very large.

A related issue in GWAS via single marker least-squares is the interpretation of the proportion of variance accounted for by regression, or R^2. Typically, this is assessed (assume all variables have been "centered", so that $\sum x_i = 0$) as

$$R^2 = \frac{\widehat{\beta}^2\left(\sum x_i^2\right)}{\sum y_i^2}.$$

In a strict sense, this is the proportion of the total sum of squares that is accounted for by the fitted line. Unfortunately, R^2 is often interpreted as a "proportion of variance", but this is not correct as variance is generated only by random factors in a linear model: fixed effects do not contribute to variance (Henderson 1953; Searle 1971). Hence, R^2 does not possess an interpretation in a strict variance components setting. On the other hand $h^2 = 2p(1-p)\beta^2/Var(y)$ represents the proportion of phenotypic variance due to the additive effect of the locus, assuming $\beta = a$. This is heritability in a narrow sense in a model where genotypes are random but their effects on the trait are fixed, contrary to the regression model where both the observed genotypes and the effects are fixed entities. Now, under Hardy-Weinberg equilibrium assumptions $E\left(x_i^2\right) = 2p(1-p)$ (e.g., Gianola et al. 2009), so that

$$E\left(R^2\right) = E_x\left[E\left(R^2|x\right)\right] \approx \frac{2p(1-p)\beta^2 N + \sigma^2}{2p(1-p)\beta^2 N + N\sigma^2}. \quad (20)$$

Using the definition of heritability in (20) and taking $\sigma^2 = (1-h^2)\mathrm{Var}(y)$ produces

$$E\left(R^2\right) \approx \frac{(N-1)h^2 + 1}{N} \approx h^2 \quad (21)$$

only if N is large and, accepting that something that is treated as fixed becomes suddenly random, an approach termed at least once by Thompson (1979) as "schizophrenic".

QTLs are elusive but an optimistic view is that one or more markers may be in linkage disequilibrium with a "causal" variant, therefore serving as a proxy for this QTL. This induces a well-known bias (e.g., Beavis 1998; Xu 2003 and Weir 2008) that is not corrected by an increase in sample size. To illustrate, suppose that the unobserved QTL has genotypes (additive effects) $QQ(a), Qq(0)$ and $qq(-a)$; then, the regression of the genetic value (G) on the number of copies of Q is a. We observe a neutral marker with genotypes MM, Mm and mm, with this marker being in LD with the QTL. The marker-based regression (φ) of the genetic value on the number of copies of M can be shown to be

$$\varphi = \frac{E(G|MM) - E(G|mm)}{2} = (1-\tau)a,$$

where

$$\tau = \frac{\Pr(Qq|MM) + \Pr(Qq|mm)}{2} + [\Pr(qq|MM) + \Pr(QQ|mm)]$$

The regression φ is equal to a only if τ is 0, and this would happen only if the marker is the QTL. Hence, the true effect of the QTL on the quantitative trait is estimated with a downward bias. If the estimate of the marker-based regression is $\widehat{\beta}$, the variance attributed to the locus is now deemed to be

$$V_{\text{marked}} = 2p_m(1-p_m)\widehat{\beta}^2,$$

where p_m is the frequency of marker allele m. Now, since $E\left(\widehat{\beta}\right) = (1-\tau)a$

$$E(V_{\text{marked}}|p_m) = 2p_m(1-p_m)\left[(1-\tau)^2a^2 + V_{\widehat{\beta}}\right],$$

where $V_{\widehat{\beta}} = \left(\sum x_i^2\right)^{-1}\sigma^2$ is the variance of the least-squares estimator of β. However, the frequency of allele m is not the frequency of allele q, that is, $p_m = p + \delta$. If sample size is very large so that $V_{\widehat{\beta}}$ is close to 0,

$$E(V_{\text{marked}}|p_m) = 2(p+\delta)(1-p-\delta)(1-\tau)^2a^2.$$

It is seen that, even when sample sizes are very large, two sources of bias remain, one due to the fact that the regression is estimated downwardly and the second associated with the fact that the allelic frequencies at the marker and QTL loci differ by δ.

The problem is much more complicated if many QTLs affect the trait and if a battery of markers is engaged in the expedition of searching for a QTL, even under the (naive) assumption of pure additivity. Suppose that one fits p markers and that sample size (N) is large enough to produce unique least-squares estimates of each regression on a marker. The regression model fitted is

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{e} = [\,\mathbf{x}_1 \quad \mathbf{x}_2 \quad . \quad . \quad . \quad \mathbf{x}_p\,]\begin{bmatrix}\beta_1\\ \beta_2\\ .\\ .\\ .\\ \beta_p\end{bmatrix} + \mathbf{e},$$

where $\mathbf{x}_i$ is an $N \times 1$ column vector linking the effect of marker i to the phenotype. Assume that there are two epistatic QTL, so that the "true" model for the trait is

$$\mathbf{y} = [\,\mathbf{q}_1 \quad \mathbf{q}_2 \quad \mathbf{q}_{12}\,]\begin{bmatrix}\alpha_1\\ \alpha_2\\ \alpha_{12}\end{bmatrix} + \mathbf{e} = \mathbf{Q}\alpha + \mathbf{e},$$

where $\mathbf{q}_1, \mathbf{q}_2$ and $\mathbf{q}_{12}$ are unknown incidence vectors linking the additive effects α_1, α_2 and the additive × additive effect α_{12} to the phenotypes. If marker effects are estimated by ordinary least-squares, the expected value of the estimator is

$$E\left(\widehat{\beta}\right) = \begin{bmatrix}\mathbf{x}_1'\mathbf{x}_1 & \mathbf{x}_1'\mathbf{x}_2 & . & . & . & \mathbf{x}_1'\mathbf{x}_p\\ . & \mathbf{x}_2'\mathbf{x}_2 & . & . & . & \mathbf{x}_2'\mathbf{x}_p\\ . & . & . & . & . & .\\ . & . & . & . & . & .\\ . & . & . & . & . & .\\ \text{symmetric} & . & . & . & . & \mathbf{x}_p'\mathbf{x}_p\end{bmatrix}^{-1} \times \begin{bmatrix}\mathbf{x}_1'(\mathbf{q}_1\alpha_1 + \mathbf{q}_2\alpha_2 + \mathbf{q}_{12}\alpha_{12})\\ \mathbf{x}_2'(\mathbf{q}_1\alpha_1 + \mathbf{q}_2\alpha_2 + \mathbf{q}_{12}\alpha_{12})\\ .\\ .\\ .\\ \mathbf{x}_p'(\mathbf{q}_1\alpha_1 + \mathbf{q}_2\alpha_2 + \mathbf{q}_{12}\alpha_{12})\end{bmatrix}.$$

It is seen that the bias of the estimator is extraordinarily complex. It is affected by *all* LD relationships among markers (note that $\mathbf{x}_i'\mathbf{x}_j$ is the sum over individuals of products of genotype codes for markers i and j, interpretable as a sample covariance if markers have been centered and standardized), and the bias is conveyed by the inverse matrix in the preceding expression. The bias of the estimator is also affected by *all* LD relationships between *all* markers and *all* unknown QTLs (and by their joint distribution of QTL genotypes over loci represented by the Hadamard vector product $\mathbf{q}_{12}$) affecting the quantitative trait, as well as by their "true" effects ($\alpha's$) on the trait. For the special case of single marker regression, the expected value of the estimator reduces to

$$E\left(\widetilde{\beta}_i\right) = \frac{\mathbf{x}_i'(\mathbf{q}_1\alpha_1 + \mathbf{q}_2\alpha_2 + \mathbf{q}_{12}\alpha_{12})}{\mathbf{x}_i'\mathbf{x}_i},$$

and note that as sample sizes goes to ∞, the probability limit of $\widetilde{\beta}_i$ is

$$\begin{aligned}&\text{plim}_{N\to\infty}\frac{\mathbf{x}_i'(\mathbf{q}_1\alpha_1 + \mathbf{q}_2\alpha_2 + \mathbf{q}_{12}\alpha_{12})}{\mathbf{x}_i'\mathbf{x}_i}\\ &\quad = \delta_{i,1}\alpha_1 + \delta_{i,2}\alpha_2 + \delta_{i,12}\alpha_{12},\end{aligned}$$

where, for example, $\delta_{i,1}$ is the regression of marker i genotype codes on QTL *1* genotype codes. This shows that even when the marker is the QTL $\left(\delta_{i,1} = 1\right)$, the regression remains biased unless all other δ-coefficients are zero.

It seems that all energies in GWAS seem to center on the problem of multiple testing, as opposed to criticism of what is clearly an inadequate model for analysis of complex traits. What is the effect of the bias discussed above on the attribution of variance stemming from a standard GWAS? The expected value of the estimator of residual variance from single marker regression (assuming centered data) is

$$E\left(\widetilde{\sigma}_e^2\right) = \frac{E(\mathbf{y}'\mathbf{y}) - \mathbf{x}_i'\mathbf{x}_i E\left(\widetilde{\beta}_i^2\right)}{N-1},$$

where

$$E(\mathbf{y}'\mathbf{y}) = \alpha'\mathbf{Q}'\mathbf{Q}\alpha + N\sigma_e^2,$$

and

$$E\left(\widetilde{\beta}_i^2\right) = \left(\frac{\mathbf{x}_i'\mathbf{Q}\alpha}{\mathbf{x}_i'\mathbf{x}_i}\right)^2 + Var\left(\widetilde{\beta}_i\right) = \frac{\alpha'\mathbf{Q}'\mathbf{x}_i\mathbf{x}_i'\alpha}{\left(\mathbf{x}_i'\mathbf{x}_i\right)^2} + \frac{\sigma_e^2}{\mathbf{x}_i'\mathbf{x}_i}.$$

Hence,

$$\begin{aligned}E\left(\widetilde{\sigma}_e^2\right) &= \frac{\alpha'\mathbf{Q}'\mathbf{Q}\alpha + N\sigma_e^2 - \left[\frac{\alpha'\mathbf{Q}'\mathbf{x}_i\mathbf{x}_i'\mathbf{Q}\alpha}{\mathbf{x}_i'\mathbf{x}_i} + \sigma_e^2\right]}{N-1}\\ &= \sigma_e^2 + \frac{\alpha'\mathbf{Q}'\left(\mathbf{I}_n - \frac{\mathbf{x}_i\mathbf{x}_i'}{\mathbf{x}_i'\mathbf{x}_i}\right)\mathbf{Q}\alpha}{N-1}\end{aligned}$$

which is biased and inconsistent, because the bias (second term in the expression above) cannot be shown to vanish with increased N since $\mathbf{I}_n - \frac{\mathbf{x}_i\mathbf{x}_i'}{\mathbf{x}_i'\mathbf{x}_i}$ grows with N as well. Now, the t^2 or $F-$ statistic used for computing $p-$ values for testing the hypothesis H_0:$\beta_i = 0$ versus the alternative is based on

$$F = t^2 = \frac{\widetilde{\beta}_i^2}{\widehat{Var}\left(\widetilde{\beta}_i\right)},$$

where $\widehat{Var}\left(\widetilde{\beta}_i\right) = \frac{\widetilde{\sigma}_e^2}{\mathbf{x}_i'\mathbf{x}_i}$is the estimate of the variance of the regression coefficient. Then, approximately

$$E(F) \approx \frac{\sigma_e^2 + \frac{\alpha'\mathbf{Q}'\mathbf{x}_i\mathbf{x}_i'\mathbf{Q}\alpha}{\left(\mathbf{x}_i'\mathbf{x}_i\right)}}{\sigma_e^2 + \frac{\alpha'\mathbf{Q}'\left(\mathbf{I}_n - \frac{\mathbf{x}_i\mathbf{x}_i'}{\mathbf{x}_i'\mathbf{x}_i}\right)\mathbf{Q}\alpha}{N-1}},$$

which is not equal to 1 under the null hypothesis $\beta_i = 0$ unless the marker is the QTL, and provided that there are no other QTLs or epistatic effects involving the trait in question. It follows that F (or t) cannot have a central distribution and that p-values in GWAS are questionable. This problem cannot be solved by any of the multiple-test corrections (such as Bonferroni) done in standard GWAS. Finally, the variance attributed to a locus in a standard GWAS is assessed (assuming that the data have been centered) as

$$R^2 = 1 - \frac{\mathbf{y}'\mathbf{y} - \mathbf{x}_i'\mathbf{x}_i\widetilde{\beta}_i^2}{\mathbf{y}'\mathbf{y}},$$

so using the preceding developments one arrives at the approximate result

$$\begin{aligned} E\left(R^2\right) &\approx 1 - \frac{(N-1)\sigma_e^2 + \alpha'\mathbf{Q}'\mathbf{Q}\alpha - \frac{\alpha'\mathbf{Q}'\mathbf{x}_i\mathbf{x}_i'\mathbf{Q}\alpha}{\mathbf{x}_i'\mathbf{x}_i}}{N\sigma_e^2 + \alpha'\mathbf{Q}'\mathbf{Q}\alpha} \\ &\approx \frac{\alpha'\mathbf{Q}'\mathbf{x}_i\mathbf{x}_i'\mathbf{Q}\alpha}{\left(N\sigma_e^2 + \alpha'\mathbf{Q}'\mathbf{Q}\alpha\right)\mathbf{x}_i'\mathbf{x}_i}, \quad \text{for large } N. \end{aligned} \tag{22}$$

Clearly, this is difficult to interpret.

In summary, the partition of variance into locus-specific contributions is not straightforward when linkage disequilibrium exists. Knowledge of the distribution of allelic effects and of frequencies is required, in addition to the entire linkage disequilibrium structure, to answer the question properly. Unfortunately, a great difficulty is that of obtaining a sensible estimate of a multi-dimensional LD structure. On the other hand, if the distribution of additive effects is symmetric and independent of that of allelic frequencies, assuming LE mat provide a reasonable approximation to the variance partition. It may be possible to refine the variance partition further by introducing models for the LD structure. For instance, the "Bulmer" effect (Bulmer 1971) produces within-chromosome gradients of negative LD, and there is empirical evidence from cattle (Henner Simianer and Saber Qanbari, personal communication) that LD tends to be negative within chromosomes, but positive when the pairs of loci involve different chromosomes. However, this problem is brought up here only for the purpose of suggesting that research may be warranted in this area. We conclude that attributions to variance contributed by a single QTL from a standard GWAS analysis may be misleading, conceptually and statistically, when the trait is complex and affected by many genes. Yet another factor to consider in the "missing heritability" saga?.

Acknowledgments Daniel Gianola acknowledges support from the Wisconsin Agriculture Experiment and from a joint grant from the Scientific Office of AgroParisTech. France, and the Animal Genetics Division of INRA, France. The authors thank two anonymous reviewers, especially "2", for a most through appraisal of the manuscript.

References

Avery PJ, Hill WG (1979) Variance in quantitative traits due to linked dominant genes and variance in heterozygosity in small populations. Genetics 91:817–844

Barton NH (2000) Estimating multilocus linkage disequilibria. Heredity 84:373–389

Beavis WD (1998) QTL analysis: Power, precision, and accuracy. pp. 145–161. In: Paterson AH (ed.) Molecular dissection of complex traits. CRC Press, Boca Ration

Bulmer MG (1971) The effect of selection on genetic variability. Am Nat 105:201–211

Bulmer MG (1976) Regressions between relatives. Genet Res 28:199–203

Bulmer MG (1980) The Mathematical Theory of Quantitative Genetics. Oxford University Press, New York

Comstock RE, Robinson HF (1952) Estimation of average dominance of genes. In JW Gowen (ed.) Heterosis, pp 494–516. Lowa State College Press, Ames

Daetwyler, HD, Pong-Wong R, Villanueva B, Wooliams JA (2010) The impact of genetic architecture on genome-wide evaluation methods. Genetics 185:1021–1031

de los Campos G, Gianola D, Allison DAB (2010) Predicting genetic predisposition in humans: the promise of whole-genome markers. Nat Rev Genet 11:880–886

Emigh TH (1977) Partition of phenotypic variance under unknown dependent association of genotypes and environments. Biometrics 33:505–514

Falconer DS, Mackay TFC (1996) Introduction to Quantitative Genetics. 4th edn. Longman, New York

Fisher RA (1918) The correlation between relatives on the suppostion of Mendelian inheritance. Trans Royal Soc Edinburgh 52: 399–433

Gianola D, de los Campos G, Hill WG, Manfredi E, Fernando RL (2009) Additive genetic variability and the Bayesian alphabet. Genetics 183:347–363

Goldberger AS (1977) Models and methods in the IQ debate, Part I. Social Systems Research Institute Workshop Series, Number 7710. University of Wisconsin, Madison

Goddard ME, Hayes BJ (2009) Mapping genes for complex traits in domestic animals and their use in breeding programmes. Nat Rev Genet 10:381–391

Hayes JF, Hill WG (1981) Modification of estimates of parameters in the construction of genetic selection indices. Biometrics 37: 483–493

Hedrick PW (1987) Gametic disequilibrium measures: proceed with caution. Genetics 117:331–341

Henderson CR (1953) Estimation of variance and covariance components. Biometrics 9:226–252

Heslot N, Yang HP, Sorrells ME, Jannink JL (2012) Genomic selection in plant breeding: a comparison of models. Crop Sci 52:146–160

Hill WG, Robertson A (1966) The effect of linkage on limits to artificial selection. Genet Res 8:269–294

Hill WG, Robertson, A (1968) Linkage disequilibrium in finite populations. Theor Appl Genet 38:226–231

Hospital F (1992) Effets de la liaison genique et des effectifs finis sur la variabilité des caracteres quantitatifs sous selection. These de Doctorat. Universite de Motpellier II, Academie de Montpellier

Kathiresan S, Melander O, Guiducci O, Surti A, Burtt N, Rieder MJ, Cooper GM, Roos C, Voight BF, Havulinna AS, Wahlstrand B, Hedner T, Corella D, Shyong T, Ordovas JM, Berglund G, Vartiainen E, Jousilahti P, Hedblad B, Taskinen MR, Newton-Cheh C, Salomaa V, Peltonen L, Groop L, Altshuler DM, Orho-Melander M (2008) Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. Nat Genet 40:189–196

Kempthorne O (1978) Logical, epistemological and statistical aspects of nature-nurture data interpretation. Biometrics 34:1–23

Lewontin RC, Rose A, Kamin LJ (1984) Not in Our Genes: Biology, Ideology, and Human Nature. New York, Penguin

Lewontin RC (1988) On measures of gametic disequilibrium. Genetics 120:849–852

Lynch M, Walsh B (1998) Genetics and Analysis of Quantitative Traits. Sinauer, Sunderland

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TF, McCarroll SA, Visscher PM (2009) Finding the missing heritability of complex diseases. Nature 8. doi:10.1038/nature08494

Marchetti GM, Drton M (2010) ggm: Graphical Gaussian Models. R package version 1.0.4. http://CRAN.R-project.org/package=ggm

Marsaglia G, Olkin I (1984) Generating correlation matrices. SIAM J Sci Stat Comput 5:470–475

Meuwissen TH, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–1829

Ober U, Ayroles JF, Stone EA, Richards S, Zhu D, Gibbs RA, Stricker C, Gianola D, Schlather M, Mackay TFC, Simianer H (2012) Using whole-genome sequence data to predict quantitative trait phenotypes in Drosophila melanogaster. PLos Genet 8:e1002685

Powell JE, Kranis A, Floyd J, Dekkers JCM, Knott S, Haley CS (2011) Optimal use of regression models in genome-wide association studies. Anim Genet 43:133–143

Sabatti C, Risch N (2002) Homozygosity and linkage disequilibrium. Genetics 160:1707–1719

Searle SR (1971) Linear Models. Wiley, New York

Speliotes EK, Willer CJ, Berndt SI, Monda KL, Thorleifsson G et al (2010) Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. Nat Genet 42: 937–948

Stranger BE, Stahl EA, Raj T (2011) Progress and promise of genome-wide association studies for human complex trait genetics. Genetics 187:367–383

Thompson R (1979) Sire evaluation. Biometrics 35:339–353

Turelli M, Barton NH (1990) Dinamycs of polygenic characters under selection. Theor Popul Biol 38:1–57

Weir B (2008) Linkage disequilibrium and association mapping. Annu Rev Genom Human Genet 9:129–142

Wu X, Ye Y, Rosell R, Amos CI et al (2011) Genome-wide association study of survival in non–small cell lung cancer patients receiving platinum-based chemotherapy. J Natl Cancer Inst 103:817–825

Xu S (2003) Theoretical basis of the Beavis effect. Genetics 165:2259–2268

Zhang X-S, Wang J, Hill WG (2002) Pleiotropic model of maintenance of quantitative genetic variation at mutation–selection balance. Genetics 161:419–433

Zhao H, Nettleton D, Soller M, Dekkers JCM (2005) Evaluation of linkage disequilibrium measures between multi-allelic markers as predictors of linkage disequilibrium between markers and QTL. Genet Res Camb 86:77–87

Zuk O, Hechter E, Sunyaev SR, Lander ES (2012) The mystery of missing heritability: genetic interactions create phantom heritability. Proc Natl Academy Sci 109:1193–1198