Behavior Genetics, Vol. 30, No. 3, 2000

Some Properties of a Variance Components Model for Fine-Mapping Quantitative Trait Loci

Lon R. Cardon[1,2] and Gonçalo R. Abecasis[1]

Received 1 Nov. 1999—Final 10 Mar. 2000

Identifying etiological variants for multifactorial traits by allelic association holds promise when many markers are available in close proximity. However, evidence for or against association at any particular marker does not provide any direct information about the influence of causal variants or the frequency of the etiologic allele(s). Recently, a variance components model of linkage and association was developed for quantitative traits which is sufficiently flexible to provide some insights into these issues. We show that this combined linkage/association model provides an estimate of the additive genetic variance of a trait that is attributable to disequilibrium between the marker and QTL. We use this estimate to construct approximate boundaries of the minimum level of disequilibrium between an observed marker and unobserved QTL and to delimit the permissible range of allele frequencies at the QTL based on available data at nearby markers. This information may facilitate fine-mapping studies of complex traits that aim to localize QTLs by assessment of association with many markers in a candidate region of interest.

KEY WORDS: Quantitative trait loci; single nucleotide polymorphisms; linkage; associations; variance components.

INTRODUCTION

Recent efforts toward construction of a high-density map of single nucleotide polymorphisms (SNPs) across the human genome have generated high expectations for identification of multifactorial trait loci (Risch and Merikangas, 1996; Chakravarti, 1998; Kruglyak, 1999; Lander, 1999). For complex diseases, a number of family-based allelic association methods have been developed to accommodate SNP data while addressing potentially confounding issues such as population admixture, stratification and the combined effects of linkage and association (Spielman and Ewens, 1996). However, the development of analogous methods for quantitative traits has lagged behind that for discrete traits. This is unfortunate since many common disorders such as dyslexia, anxiety/depression, obesity, hypertension, osteoporosis, and asthma are often measured on continuous scales in both research and clinical settings (e.g., reading ability, personality scales, body size/weight, blood pressure, bone density, bronchial responsiveness). Although virtually any continuous measure can be easily dichotomized and thus analyzed using the available suite of discrete measure tests, such (arbitrary) data transformations generally result in wasted information and a subsequent loss of statistical power to detect genetic effects (Neale and Cardon, 1992).

Recently, Fulker *et al.* (1999) developed an extension of a commonly used linkage analysis approach for assessment of allelic association in quantitative traits. This method, embedded in the context of variance components modeling, makes use of means and variances between and within siblings to account for linkage and association simultaneously. For association assessments, this between/within pair model provides

[1] The Wellcome Trust Centre for Human Genetics, University of Oxford, Roosevelt Drive, Oxford OX3 7BN, U.K.
[2] To whom correspondence should be addressed. Fax: (+44) 01865 287 650. e-mail: lon@well.ox.ac.uk

0001-8244/00/0500-0235$18.00/0

a direct test of gametic phase disequilibrium, i.e., allelic association unconfounded by population substructure (Abecasis *et al.,* 2000). Although several other family-based methods have recently been proposed for quantitative trait studies (Allison, 1997; Rabinowitz, 1997; Xiong *et al.,* 1998; Allison *et al.,* 1999a; George *et al.,* 1999; Cardon, 2000), none shares all of the features of explicit modeling of linkage, direct assessment of population substructure, and flexible handling of general family structures as does the variance components approach.

As no test of association can distinguish causal variants from alleles in linkage disequilibrium, perhaps the most useful outcome of an association study is to help guide the design of further genetic and biological experiments aimed at identifying and characterizing etiological variants. Given significant evidence for association at some locus, it would be useful to know what type of markers (in terms of allele frequencies and spacing/density) should be genotyped further, how large the estimated effect size at the associated locus is, or whether the associated allele may be the etiological variant itself. The between/within variance components model has a distinctive feature that can provide some insight into these questions: namely, the model includes separate parameters for linkage and association effects. This parameterization differs from most discrete-trait [e.g., TDT (Spielman *et al.,* 1993)] and continuous-trait approaches, which tend to model linkage and association jointly and therefore evaluate the recombination fraction, disequilibrium coefficient, and effect size in a single combined test.

Because of the specific parameterization of the between/within model, positive evidence for association results in decreased linkage parameter estimates (Fulker *et al.,* 1999; Abecasis *et al.,* 2000). Consequently, in general, markers close to the QTL should be accompanied by larger association parameter estimates and therefore smaller linkage estimates than those positioned more distant to the QTL. This relationship has not been systematically explored, and it seems that there is further information in the model to assist in localization of etiological variants.

Here we evaluate the behavior of the association and linkage parameters in the model of Fulker *et al.* (1999). We show that the between/within pair modeling framework permits direct estimation of the proportion of major locus additive genetic variance explained by disequilibrium between marker and QTL, thereby providing indirect information about marker–QTL distance that can assist in fine-mapping studies. We show that one of the likelihood-ratio tests used in the original formulation of the model provides a means to assess whether evidence for association suggests complete vs. incomplete linkage disequilibrium between the marker and QTL. We also show that parameter estimates from models including vs. omitting association effects can be used to construct estimates of the minimum disequilibrium coefficient between alleles at an observed marker and unobserved QTL. Finally, we use the parameter estimates to delimit the permissible frequency range of the unobserved QTL alleles.

COMBINED LINKAGE/ASSOCIATION MODEL

The combined linkage/association model of Fulker *et al.* (1999) is based on the standard biometrical model of an observed trait value, y_{ij}, for the jth member in the ith family as a function of an overall phenotypic mean, μ, a major additive genetic effect, a_{ij}, background genetic effects, g_{ij}, and random environmental effects, e_{ij}:

$$y_{ij} = \mu + a_{ij} + g_{ij} + e_{ij} \quad (1)$$

where the background genetic and environmental effects have mean zero. Let q_1 and q_2 represent alleles of the QTL responsible for the major genetic effect with frequencies p and q, respectively. In the absence of dominance variation, $a_{ij} = a$ for QTL genotype q_1q_1, $a_{ij} = 0$ for genotypes q_1q_2 and q_2q_1, and $a_{ij} = -a$ for genotype q_2q_2, where a is the additive genetic value of the QTL. Following standard quantitative genetics theory (Falconer, 1981), the additive QTL contribution to the phenotypic mean is $a(p - q)$ and the additive genetic variance of the QTL, V_a, is $2pqa^2$. Our assumption of no dominance variance is for simplicity of exposition only. Approaches for modeling dominance effects in the Fulker *et al.* (1999) model have recently been described (Sham *et al.,* 2000).

Given this standard biometrical model, the expected phenotypic variance for any family member is

$$V_p = V(y_{ij}) = V_a + V_g + V_e \quad (2)$$

where subscripts g and e represent residual familial and nonshared family environment effects, respectively. Let π_{ijk} represent the proportion of alleles shared identical-by-descent (ibd) at the marker locus for individuals j and k in family i, and ϕ_{ijk} represent the coefficient of relationship between the same individuals. The ibd sharing information may be derived from marker-specific or multipoint applications (Kruglyak and Lander, 1995). Also let σ^2_a, σ^2_g, and σ^2_e represent

variance parameters to be estimated, corresponding to V_a, V_g, and V_e, respectively. Then the expected variance/covariance of any two family members is

$$\sum_{ijk} = \begin{cases} \sigma_a^2 + \sigma_g^2 + \sigma_e^2 & \text{if} \quad j = k \\ f(\theta)\sigma_a^2 + \phi_{ijk}\sigma_g^2 & \text{if} \quad j \neq k. \end{cases} \qquad (3)$$

Note that the estimate of the additive genetic variance is confounded with the recombination fraction between the marker and QTL, shown as $f(\theta)\sigma_a^2$. In sibling pairs, for example, the estimated component of variance attributable to a QTL is $f(\theta)\sigma_a^2 = [\frac{1}{2} + (1 - 2\theta)^2 (\pi_{ijk} - \frac{1}{2})]\sigma_a^2$ (Amos, 1994). For simplicity of expression, henceforth we use the term σ_a^2 to refer to $f(\theta)\sigma_a^2$. This notation does not imply an assumption of $\theta = 0$.

Under an assumption of multivariate normality, σ_a^2, σ_g^2, and σ_e^2 are estimated by maximizing the log of the likelihood of the data,

$$L = c\prod_i |\Sigma_i|^{-1/2} \exp\left[-\frac{1}{2}(\mathbf{y}_i - \hat{\mathbf{y}}_i)' \Sigma_i^{-1} (\mathbf{y}_i - \hat{\mathbf{y}}_i)\right] \quad (4)$$

where c is a constant (Lange *et al.*, 1976), Σ_i is the expected covariance matrix, $\mathbf{y}_i$ the observed phenotype vector, and $\hat{\mathbf{y}}_i$ the expected phenotype vector for family i. Significance tests of the estimates are conducted by maximizing $\log(L)$ without constraints on the parameters, $\log(L_1)$, and comparing this likelihood with models in which selected parameters are fixed at zero, $\log(L_0)$. Asymptotically, the quantity $2[\log(L_1) - \log(L_0)]$ is distributed as χ^2 with degrees of freedom equal to the difference in number of parameters estimated, although violations of the multivariate normality assumption and specific parameter boundary restrictions can perturb this distribution (Hopper and Mathews, 1982; Allison *et al.*, 1999b).

The means model of allelic association for sibling pairs can be conveniently expressed in terms of gene dosage (Abecasis *et al.*, 2000). Let m_1 and m_2 represent alleles of a diallelic marker with frequencies r and s, respectively. Let c_{ij} represent the number of m_1 alleles (minus 1) at the marker for individual j in family i. For n_i siblings in family i, let $b_i = \sum_j c_{ij}/n_i$ and $w_{ij} = c_{ij} - b_i$, so that b_i is the expectation of each c_{ij} conditional on family data and w_{ij} is deviation from this expectation for offspring j. Positive values of w_{ij} indicate that an offspring inherits more copies of allele m_1 than expected, while negative values refer to excess inheritance of allele m_2. The linear model proposed by Fulker *et al.* (1999) is

$$\hat{y}_{ij} = \mu + \beta_b b_i + \beta_w w_{ij} \qquad (5)$$

so that the overall QTL mean is partitioned into between- and within-family effects defined on the basis of marker genotypes. Between-family effects may be influenced by such factors as population substructure, while within-family effects should reflect only gametic phase disequilibrium. A test of $\beta_b = \beta_w$ can be used to evaluate evidence for population substructure. If $\beta_b = \beta_w$, a more parsimonious expression of the model is

$$\hat{y}_{ij} = \mu + \beta_a(b_i + w_i) = \mu + \beta_a c_{ij} \qquad (6)$$

In all forms of this combined linkage/association model, the association parameters may be tested for significance while the standard variance parameters are left free to vary.

CHANGE IN VARIANCE DUE TO ASSOCIATION

In the likelihood given in (4), the matrix Σ_i is used to model the expected variance of the residuals. When the predicted trait vector, $\hat{\mathbf{y}}_i$ contains only the population mean (as in a linkage-only model in which $\hat{\mathbf{y}}_i = \mu$ for all i), the quantity $\Sigma_i(\mathbf{y}_i - \hat{\mathbf{y}}_i)' (\mathbf{y}_i - \hat{\mathbf{y}}_i)$ in (4) reflects the total phenotypic variance (i.e., $V_e + V_g + V_a$). However, the between/within model in (5) or (6) expresses each $\hat{\mathbf{y}}_i$ as a function of the marker so that $\sum_i(\mathbf{y}_i - \hat{\mathbf{y}}_i)' (\mathbf{y}_i - \hat{\mathbf{y}}_i)$ no longer reflects the total phenotypic variance. To describe $\sum_i(\mathbf{y}_i - \hat{\mathbf{y}}_i)' (\mathbf{y}_i - \hat{\mathbf{y}}_i)$, we require the variances and covariance of the observed and predicted trait scores, $\mathbf{y}_i$ and $\hat{\mathbf{y}}_i$.

The variance of y is shown in (2).To determine the variance of $\hat{y}$, it is convenient to consider "additive genetic values" for the marker genotypes as α for genotype m_1m_1, 0 for genotype m_1m_2, and $-\alpha$ for genotype m_2m_2 (at present, α is undefined; in the Appendix we show that $\alpha = aD/rs$, where D is the disequilibrium coefficient between m_1 and q_1). Then, as in the usual biometrical model, the variance of $\hat{y}$ is simply $V_{\hat{y}} = V_\alpha = 2rs\alpha^2$. The covariance between y and $\hat{y}$ is $C_{y,\hat{y}} = \sum_{iikl} P(q_i q_j m_k m_l) a_{ij} \alpha_{kl}$, since a_{ij} and α_{kl} are mean deviates for y and $\hat{y}$, respectively (see the Appendix). When allelic association between q_1 and m_1 is expressed in terms of the haplotype frequency and the product of the component allele frequencies, $D = P(q_1m_1) - pr$, the probabilities of the joint marker–QTL genotypes can be simply determined (Weiss, 1993). These frequencies, together with the genetic values a and α for each possible genotype pair, are given in Table I and can be used to show that $C_{y,\hat{y}} = 2rs\alpha^2 = V_a$.

Using these quantities, the variance described by $\Sigma_i(\mathbf{y}_i - \hat{\mathbf{y}}_i)' (\mathbf{y}_i - \hat{\mathbf{y}}_i)$ in the between/within model is $V_y +$

Table I. Genetic Values (a_{ij}, α_{kl}) and Joint Genotype Probabilities for a QTL and Marker in Disequilibrium

QTL genotype	Marker genotype	a_{ij}	α_{kl}	Probability
q_1q_1	m_1m_1	a	α	$(D+pr)^2$
q_1q_1	m_1m_2	a	0	$2(D+pr)(ps-D)$
q_1q_1	m_2m_2	a	$-\alpha$	$(ps-D)^2$
q_1q_2	m_1m_1	0	α	$2(D+pr)(qr-D)$
q_1q_2	m_1m_2	0	0	$2[(D+pr)(D+qs)+(ps-D)(qr-D)]$
q_1q_2	m_2m_2	0	$-\alpha$	$2(ps-D)(D+qs)$
q_2q_2	m_1m_1	$-a$	α	$(qr-D)^2$
q_2q_2	m_1m_2	$-a$	0	$2(qr-D)(D+qs)$
q_2q_2	m_2m_2	$-a$	$-\alpha$	$(D+qs)^2$

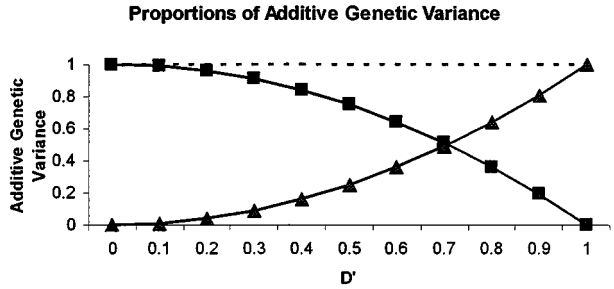

Fig. 1. Relationship between total additive genetic variance (dashed line), additive genetic variance estimated in the presence of association parameters (V_a; squares), and additive genetic variance attributable to the marker through linkage disequilibrium with the QTL (V_α; triangles).

$V_{\hat{y}} - 2C_{y,\hat{y}} = V_e + V_g + V_a - V_\alpha$. Thus, the linkage parameter in the between/within model provides an estimate of the difference between the additive genetic variance of the QTL and the variance of the QTL explained by association with the marker allele:

$$\sigma^2_a = V_a - V_\alpha \quad (7)$$

The relationship between V_α and V_a is shown in Fig. 1, where it can be seen that although the total estimated genetic variance is constant for all levels of disequilibrium, V_a decreases and V_α increases with disequilibrium levels. That is, the total variance explained by the between/within model is the same whatever the level of disequilibrium, but it is partitioned into different parameters in proportion to the degree of association between the marker and QTL.

Knowing the actual basis of the linkage parameter, σ^2_a, helps to clarify some patterns under different conditions of disequilibrium and marker/QTL allele frequencies (Table II). When $D = 0$, σ^2_a reduces to the typical additive genetic variance of a QTL, but when the absolute value of D is between 0 and D_{max} ($0 < |D| < D_{max}$), the estimate is reduced by an amount that depends on D, a, and the frequency of the marker alleles. Interestingly, when disequilibrium is complete ($|D| = D_{max}$), two different parameter estimates can arise. When the marker and trait allele frequencies are unequal, the association and linkage parameter estimates are indistinguishable from the case where linkage disequilibrium is incomplete. Thus, it is not always possible to distinguish between complete and incomplete linkage disequilibrium in this model. In contrast, when disequilibrium is complete and the marker and trait alleles are equal, $V_a = V_\alpha = 2pqa^2$, and the σ^2_a estimate in the variance components model equals zero. In the latter case, all of the linkage information is encompassed in the linear model, and the association parameter provides an unconfounded estimate of the additive genetic value of the QTL, $\beta_w = a$.

Table II. Expected Values of Linkage and Association Parameters[a]

Disequilibrium level	QTL/marker allele frequency	$E(\beta_w)$	$E(\sigma^2_a\|\beta_w)$
Complete ($\|D\| = D_{max}$)	$p = r$	a	0
Complete ($\|D\| = D_{max}$)	$p \neq r$	α	$V_a - V_\alpha$
Incomplete ($0 < \|D\| < D_{max}$)	$p \neq r$ or $p = r$	α	$V_a - V_\alpha$
None ($D = 0$)	$p \neq r$ or $p = r$	0	V_a

[a] D is the disequilibrium coefficient between marker and QTL alleles; a represents the additive genetic value; $\alpha = aD/rs$; $V_a = 2pqa^2$; $V_\alpha = 2rs\alpha^2$; QTL and marker allele frequencies are represented by p and q and by r and s, respectively. All quantities of $E(\sigma^2_a|\beta_w)$ are expressed under the assumption that $\theta = 0$. When $\theta > 0$, all values in this column will decrease by a factor $f(\theta)$ as described by Amos (1994).

CANDIDATE GENE TEST

Given the relationship between the linkage and the association parameters under different conditions of disequilibrium, it is possible to construct a likelihood-ratio test to assess whether association to a candidate polymorphism is consistent with evidence for an etiological variant of the trait. As shown in Table II, when the marker is the QTL, the linkage parameter, σ^2_a, equals zero and the estimate of β_w reflects the additive genetic value of the QTL. Thus, given significant evidence for association, situations in which $\sigma^2_a = 0$ suggest that the variant may be an etiologic mutation (or, equivalently, a marker in complete disequilibrium with

identical allele frequencies). Comparison of a model involving all free parameters $\theta_{(A)} = (\mu, \beta_b, \beta_w, \sigma^2_a, \sigma^2_g, \sigma^2_e)$, against a model in which σ^2_a is set to zero, $\theta_{\sigma^2_a = 0} = (\mu, \beta_b, \beta_w, \sigma^2_g, \sigma^2_e)$, provides a 1-df test of residual genetic variance after accounting for the disequilibrium between the marker and QTL. Since $\sigma^2_a = 0$ only if $V_a = V_\alpha$ (see Table II), a significant difference between the models implies that $\sigma^2_a > 0$ and $V_a \neq V_\alpha$; i.e., the marker and QTL are distinct. Conversely, a nonsignificant difference indicates that the marker and QTL are indistinguishable. Fulker *et al.* (1999) used this likelihood-ratio test in their simulations but did not explore its properties with respect to parameter expectations.

DISEQUILIBRIUM AND ALLELE FREQUENCY BOUNDARIES

Rejecting the model of identity between marker and QTL may be useful for guiding fine-mapping association studies, as it provides a specific means to distinguish markers in linkage disequilibrium from actual QTLs. Still, there is further information in the parameter estimates that may assist in such endeavors. Consider the standard linkage-only variance components model involving parameters $\theta_{(L)} = (\mu, \sigma^2_a, \sigma^2_g, \sigma^2_e)$ and the full association model involving parameters $\theta_{(A)} = (\mu, \beta_b, \beta_w, \sigma^2_a, \sigma^2_g, \sigma^2_e)$. By independently fitting these two models, we obtain estimates of V_a, V_α, and β_w that can be compared to describe the genetic architecture of the QTL. Let $\sigma^2_a(L)$ and $\sigma^2_a(A)$ reflect additive genetic variance estimates from the linkage-only and full association models, respectively. As shown above, $\sigma^2_a(L)$ estimates $V_a = 2pqa^2$, $\sigma^2_a(A)$ estimates $V_a - V_\alpha$, and β_w estimates aD/rs. Thus, $\sigma^2_{a(L)} - \sigma^2_a(A)$ provides an estimate of $V_a = 2rs\alpha^2$. Algebraic rearrangement of these expectations yields the squared disequilibrium coefficient as $D^2 = V^2_a\, pq/(2\beta^2_w V_a)$, which, as a proportion of the maximum disequilibrium, $D_{\max}$, is $D'^2 = (D/D_{\max})^2 = V^2_a pq/2\beta^2_w V_a D^2_{max}$. From the model parameter estimates, all of the quantities needed to calculate this ratio are available except the frequency of QTL alleles. In their absence, we can still derive two useful pieces of information: (i) the absolute value of the minimum normalized disequilibrium coefficient, D'; and (ii) the minimum and maximum QTL allele frequencies.

For a minimum D' value, note that the proportion of additive genetic variance explained by marker–QTL association is

$$\frac{V_\alpha}{V_a} = \frac{2rs\alpha^2}{2pqa^2} = \frac{rs(aD/rs)^2}{pqa^2} = \frac{D^2}{pqrs}$$

which is simply the χ^2 statistic for testing $D = 0$ for $2n$ gametes, divided by $2n$ (Weir, 1996). Given any value of V_α/V_a, the minimum D'^2 consistent with V_α/V_a occurs when the allele frequencies of the marker and QTL are equal. In this case, the denominator of this expression simplifies to $pqrs = p^2q^2 = D^2_{\max}$, so $V_\alpha/V_a = D'^2$. Thus, an absolute value of the minimum normalized disequilibrium coefficient is readily available from the parameter estimates as

$$D'_{\min} = \sqrt{V_\alpha/V_a} \tag{8}$$

In order to construct QTL allele frequency boundaries, we make use of the conditions that $D_{\max} = ps$ if $p \leq r$ and $D_{\max} = qr$ if $p \geq r$. In these situations, we can rearrange the model parameters to obtain $p = 1/(1 + \varphi/r^2)$ if $p \leq r$ and $p = 1 - [1/(1 + \varphi/s^2)]$ if $p \geq r$, where $\varphi = V_a D'^2/(2\beta_w{}^2)$. Note that the quantity $1/(1 + \varphi/r^2)$ will be smallest when $D' = 1$, and similarly, $1 - 1/(1 + \varphi/s^2)$ will be largest when $D' = 1$. Thus, setting $D' = 1$ and solving for p, the boundaries of the frequency of the QTL allele associated with allele m_1 are

$$\begin{aligned} P_{\min} &= \frac{1}{1 + V_a/(2\beta^2_w r^2)} \\ P_{\max} &= 1 - \left(\frac{1}{1 + V_a/(2\beta^2_w s^2)}\right) \end{aligned} \tag{9}$$

Table III shows the QTL allele range according to various marker allele frequencies and the proportion of variance explained by linkage disequilibrium between marker and QTL. Clearly, little information is available concerning QTL allele frequencies when the marker explains only a small amount of the QTL variance. For example, when $V_\alpha/V_a = .25$, nearly all QTL allele frequencies are consistent with the data unless the marker frequencies are extreme. Conversely, when the variance explained by the marker is substantial, reasonably precise information is available. In the case of $V_\alpha/V_a \geq .75$, narrow QTL allele boundaries are implied by the observed marker allele frequencies. Considered from a different perspective, these ranges emphasize the restricted relationship between QTL and marker as a consequence of allele frequencies. For example, a SNP with equifrequent alleles cannot explain 50% (or more) of the variance of a QTL that has a minor allele frequency of .33 or less (Table III). This type of QTL allele frequency information could assist in fine mapping by focusing the marker identification strategy and the genotyping burden on only those markers that have likely allele frequencies in the range of the QTL.

Table III. Permissible QTL Allele Frequencies as a Function of the Amount of Variance Explained by QTL–Marker Disequilibrium and the Allele Frequency of the Marker[a]

Marker allele freq. (r)	Proportion of additive genetic variance explained by marker–QTL association							
	$V_\alpha/V_a = .25$		$V_\alpha/V_a = .5$		$V_\alpha/V_a = .75$		$V_\alpha/V_a = .95$	
	p_{min}	p_{max}	p_{min}	p_{max}	p_{min}	p_{max}	p_{min}	p_{max}
.01	.003	.039	.005	.020	.008	.013	.010	.011
.1	.027	.308	.053	.182	.077	.129	.095	.105
.2	.059	.500	.111	.333	.158	.250	.192	.208
.3	.097	.632	.176	.462	.243	.364	.289	.311
.4	.143	.728	.250	.571	.333	.471	.388	.412
.5	.200	.800	.333	.667	.429	.571	.487	.513
.6	.273	.857	.429	.750	.529	.667	.588	.612
.7	.368	.903	.538	.824	.636	.757	.689	.711
.8	.500	.941	.667	.889	.750	.842	.792	.808
.9	.692	.973	.818	.947	.871	.923	.895	.905
.99	.961	.997	.980	.995	.987	.992	.989	.990

[a] p_{min} and p_{max} are the minimum and maximum possible frequencies of a QTL allele in linkage disequilibrium with a marker allele having frequency r. V_α and V_a are the additive genetic variance attributable to linkage disequilibrium between marker and QTL and the traditional additive genetic variance, respectively.

SIMULATIONS

A number of simulations were conducted to evaluate the QTL allele frequency and disequilibrium quantities in a more practical setting. Following the biometrical model in (1), a QTL having allele frequencies p = .30 and q = .70 and explaining 25% of the phenotypic variance was simulated in 200 families, each comprising four offspring and both parents. Residual polygenic and nonshared environmental effects accounted for 50 and 25% of the phenotypic variance, respectively. The frequency of the marker allele (m_1) in disequilibrium with the increasing QTL allele (q_1) was varied between .10 and .90. For each marker allele frequency, disequilibrium was induced in the parental generation according to D' coefficients of 1.0, .5 and .0. One thousand simulations were conducted for all comparisons.

Figure 2 shows the average D'_{min} values (with standard deviations as error bars) calculated from application of expression (8) to the simulated data. Although there is substantial variability at all marker allele frequencies, average D'_{min} levels are always less than the true value simulated, and only when the marker and QTL allele frequencies coincide (r = .30) do the average D'_{min} values closely approximate the actual values. Thus, it appears that D'_{min} does indeed capture the desired information. It is noteworthy that the average D'_{min} values obtained in the absence of disequilibrium (shown as filled circles in Fig. 2) are all greater than zero. This deviation from zero reflects the fact that variance parameters (which define D'_{min} in this case) are constrained to be greater than or equal to zero (Searle *et al.*, 1992).

Simulation results for p_{min} and p_{max} are shown in Fig. 3. Results obtained for the case of complete disequilibrium are shown in the top panel, where it may be seen that the true QTL allele frequency of .30 is bounded quite tightly when the associated marker allele is of a similar frequency (e.g., r = .25 – .35). While this precision diminishes rapidly as the difference between QTL and marker allele frequencies increases, even when r = .70, the data correctly indicate that the associated QTL allele is not exceedingly rare. This outcome is encouraging for studies that face the need for detection of many new polymorphic sites. When disequilibrium is more modest, as in the middle panel in Fig. 3, markers with frequencies in the range of .20–.50 can still exclude extreme QTL allele frequencies. Of course, when disequilibrium is absent, as in the bottom panel in Fig. 3, QTL allele frequencies are of no use. These simulation outcomes closely mirror the analytical results shown in Table III, emphasizing the restrictive relationships implied by disequilibrium levels and marker/QTL allele frequency differences.

DISCUSSION

We have shown that when linkage disequilibrium is apparent in the model of Fulker *et al.* (1999), the linkage

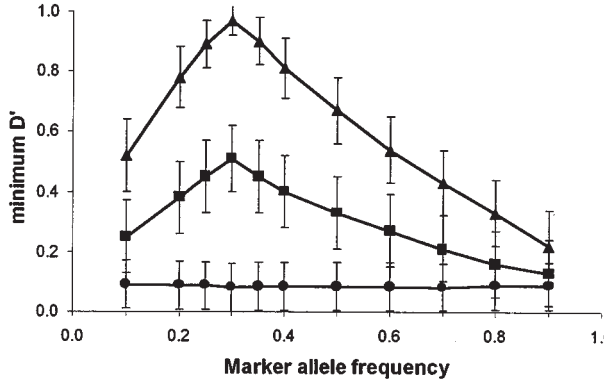

Fig. 2. Average D'_{min} values obtained from simulations of disequilibrium in small nuclear families. The simulated QTL explains 25% of the phenotypic variance and has increasing allele frequency p = .30. Residual genetic and environmental effects were simulated to account for 50 and 25% of the phenotypic variance, respectively. The three lines show the results for simulated D' levels of 1.0 (triangles), .5 (squares), and .0 (circles). Standard deviations of 1000 D'_{min} levels calculated according to expression (8) are shown as error bars.

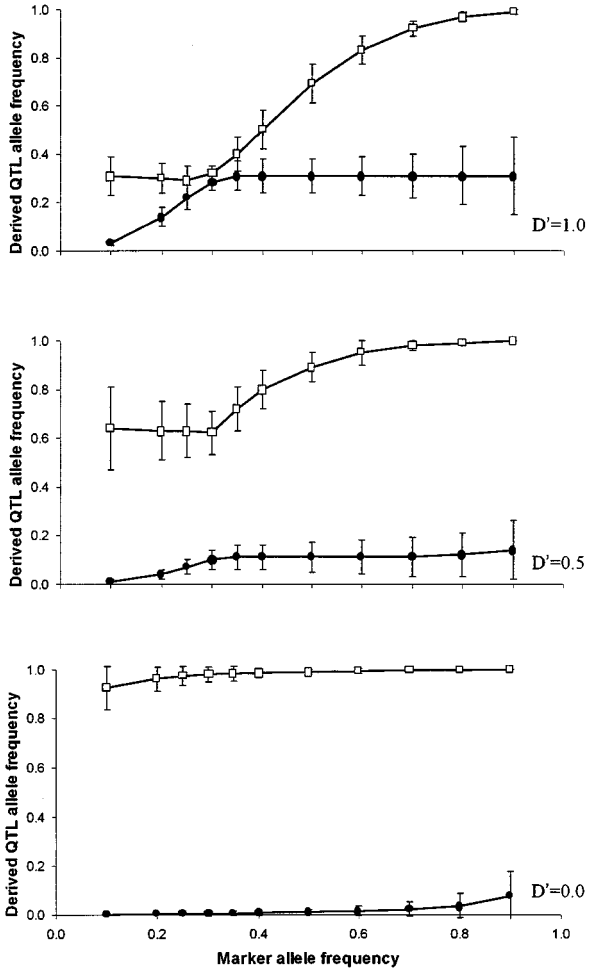

Fig. 3. Average minimum and maximum QTL allele frequencies obtained from 1000 simulations of small nuclear families. The three panels show complete disequilibrium (top), D' = .50 (middle), and no simulated disequilibrium (bottom). The properties of the QTL are as described in the text and in the legend to Fig. 2. The average minimum and maximum QTL allele frequencies derived according to expression (9) are shown as filled circles and open squares, respectively. Average p_{min} and p_{max} values are shown ±1 SD for all simulations conducted.

and association parameters have a simple relationship that provides extra information about the underlying QTL. In contrast to most other family-based association approaches, which evaluate only whether or not a marker allele is associated with variability in a trait (and thus do not directly assist in the design of further experiments), this information could help determine whether additional markers should be examined, what allele frequencies they should have, and how large a region they should cover.

In fine-mapping studies of multifactorial traits, one of the many problems that must be overcome is that of defining a mutation. Naturally, evidence for association, however strong, does not imply etiology effect so other factors must be used to draw conclusions. The likelihood-ratio test proposed here can be used to distinguish QTLs from other associated markers. Rejecting a model of QTL/marker identity indicates that the associated allele is not the etiological variant or that there are multiple mutations in the QTL, and therefore, cloning research should proceed further. However, rejecting QTL/marker identity may require large samples, owing to the dependence on linkage effects in this test. Consequently, this approach is well suited to the evaluation of candidate polymorphisms in a data set for which linkage evidence has been obtained. Careful consideration of the linkage/association patterns, especially in the context of successive markers, may facilitate selection of appropriate markers or functional experiments for further study.

Additional problems in fine-mapping studies of multifactorial traits concern the optimal marker density and allele frequencies. Regarding the former, some have argued (Kruglyak, 1999) that markers must be spaced extremely close together in order to have any chance of highlighting complex trait loci, although real data suggest that the distances are likely to be region-specific (Keavney *et al.*, 1998; Nickerson *et al.*, 1998; Rieder *et al.*, 1999). Regarding the latter, it has been proposed that common diseases imply that the underlying mutations are also common (Cargill *et al.*, 1999; Halushka *et al.*, 1999), although others contend that this is not likely to be the case (Weiss, 1996; Terwilliger

and Weiss, 1998). Regardless of one's viewpoint, the best evidence for both of these issues will come from empirical data for each trait studied. From this perspective, empirical estimates of linkage disequilibrium levels and trait allele frequencies for any trait are desirable. We have shown here that the variance components framework lends itself to such estimates. In particular, estimates from models including vs. excluding association effects indicate the minimum level of disequilibrium between a marker and QTL, as well as the boundaries of the possible allele frequencies of the QTL.

Both of these quantities are useful in fine-mapping applications; the former for helping to determine what genetic distance must be saturated with markers to ensure coverage of the QTL, the latter to guide selection of markers that most closely match those of the QTL (and therefore maximize both the statistical power and the likelihood that a marker will be an etiologic variant). Moreover, the between/within model can be readily extended to allow for multiple alleles, such as microsatellites, or multiple nearby polymorphisms, such as specific haplotypes (Fulker *et al.*, 1999; Abecasis *et al.*, 2000). For these assessments, one can either test each allele/haplotype separately in a series of association models or evaluate all alleles simultaneously by including a separate pair of between/within parameters for each variant. The serial approach is fully consistent with the present derivations of $p_{\min}$, $p_{\max}$, and D'_{min}, in that the QTL frequencies and disequilibrium estimated refer to any trait variant associated with the specific marker allele examined. In the simultaneous-alleles approach, interpreting $p_{\min}$ and $p_{\max}$ is less obvious, but since the variance model is unchanged, the estimate of $D'_{\min}$ should still provide information on QTL location.

We note that our approach for calculating $D'_{\min}$, $p_{\min}$, and $p_{\max}$ [see Eqs. (8) and (9)] could be improved. The method requires parameter estimates from two nested models applied to the same data. It would be preferable to reparameterize a single model to specify these effects directly, thereby providing a mechanism to estimate the effects and calculate their standard errors. We do not see an obvious reparameterization of the model, however, and are left with the approximate solutions presented here. Bootstrapping or other repeated sampling techniques applied to the two models involved could provide information about the distribution of the boundary estimates. Although the performance of such resampling procedures remains unexplored, we might expect that while estimates from individual markers will vary (as shown in our simulations), obtaining convergent estimates from multiple markers would represent rather compelling evidence for both disequilibrium level and trait allele frequencies. This information could be quite useful for assessment of the pattern of linkage disequilibrium across multiple loci in the context of evolutionary modeling or multipoint/haplotype estimation.

APPENDIX

The value of α can be derived in terms of genetic parameters using the conditional disequilibrium probabilities

$$P(q_1|m_1) = D/r + p$$
$$P(q_2|m_1) = q - D/r$$
$$P(q_1|m_2) = p - D/s$$
$$P(q_2|m_2) = D/s + q$$

which are simply rearrangements of the disequilibrium coefficient $D = P(m_1q_1) - pr$. Consider the average effect of the QTL conditional on any marker genotype m_im_j as $\gamma_{QTL'\,m_km_l} = \sum_{ij} P(q_i, q_j \mid m_k, m_l)a_{ij,}$ where a_{ij} is used as in (1). In the absence of dominance variance, $a_{kl} = 0$ when $k \neq 1$. Thus,

$$\gamma_{QTL|m_1m_1} = P(q_1|m_1)^2 a - P(q_2|m_1)^2 a = a\left(\frac{D}{r} + p\right)^2 - a\left(q - \frac{D}{r}\right)^2 = a(p - q) + \frac{2Da}{r}$$

$$\gamma_{QTL|m_2m_2} = P(q_1|m_2)^2 a - P(q_2|m_2)^2 a = a\left(p - \frac{D}{s}\right)^2 - a\left(\frac{D}{s} + q\right)^2 = a(p - q) - \frac{2Da}{s}$$

and

$$\gamma_{QTL|m_1m_2} = P(q_1|m_1)P(q_1|m_2)a - P(q_2|m_1)P(q_2|m_2)a = a\left(\frac{D}{r} + p\right)\left(p - \frac{D}{s}\right) - a\left(q - \frac{D}{r}\right)\left(\frac{D}{s} + q\right) = a(p - q) + \frac{Da}{rs}(s - r)$$

.

When the marker is the QTL (e.g., in an assessment of a specific mutation), $r = p$ and $D = pq$, and these expressions are equivalent to the "breeding value" in

classical quantitative genetics terminology, though not centered around the population mean. Using any of these breeding values and their genotype relationships, $\gamma_{q1q1} = 2s\alpha$ or $\gamma_{q1q2} = (s - r)\alpha$ or $\gamma_{q2q2} = -2r\alpha$ (Falconer, 1981), solving for α gives

$$\alpha = \frac{aD}{rs}$$

which is the quantity noted by Fulker *et al.* (1999) to be the expected value of the association parameter, β_w.

ACKNOWLEDGMENTS

This work was supported by the Wellcome Trust. G.R.A. is a Wellcome Trust prize student. We wish to thank Drs. Robert Elston, Chris Amos, and John Hopper for helpful discussions. Upon completion of the work presented here, Dr. Pak Sham showed us an alternative derivation of the genetic variance attributed to linkage disequilibrium between a marker and QTL. Here we have referred to this quantity as V_α; Dr. Sham's derivation, which includes extensions to include dominance variance, yields an identical outcome that he describes as the "apparent genetic variance."

REFERENCES

Abecasis, G. R., Cardon, L. R., and Cookson, W. O. C. M. (2000). A general test of association for quantitative traits in nuclear families. *Am. J. Hum. Genet.* **66**:279–292.

Allison, D. B. (1997). Transmission-disequilibrium tests for quantitative traits. *Am. J. Hum. Genet.* **60**:676–690.

Allison, D. B., Heo, M., Kaplan, N., and Martin, E. R. (1999a). Sibling-based tests of linkage and association for quantitative traits. *Am. J. Hum. Genet.* **64**:1754–1763.

Allison, D. B., Neale, M. C., Zannolli, R., Schork, N. J., Amos, C. I., *et al.* (1999b). Testing the robustness of the likelihood-ratio test in a variance-component quantitative-trait loci-mapping procedure. *Am. J. Hum. Genet.* **65**:531–544.

Amos, C. I. (1994). Robust variance-components approach for assessing genetic linkage in pedigrees. *Am. J. Hum. Genet.* **54**: 535–543.

Cardon, L. R. (2000). A family-based regression model of linkage disequilibrium for quantitative traits. *Hum. Hered.* **50**: 350–358.

Cargill, M., Altshuler, D., Ireland, J., Sklar, P., Ardlie, K., *et al.* (1999). Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nature Genet.* **22**:231–238.

Chakravarti, A. (1998). It's raining SNPs, hallelujah? *Nature Genet.* **19**:216–217.

Falconer, D. S. (1981). *Introduction to Quantitative Genetics*, Longman Group, Harlow, U.K.

Fulker, D. W., Cherny, S. S., Sham, P. C., and Hewitt, J. K. (1999). Combined linkage and association sib-pair analysis for quantitative traits. *Am. J. Hum. Genet.* **64**:259–267.

George, V., Tiwari, H. K., Zhu, X., and Elston, R. C. (1999). A test of transmission/disequilibrium for quantitative traits in pedigree data, by multiple regression. *Am. J. Hum. Genet.* **65**:236–245.

Halushka, M. K., Fan, J. B., Bentley, K., Hsie, L., Shen, N., *et al.* (1999). Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. *Nature Genet.* **22**:239–247.

Hopper, J. L., and Mathews, J. D. (1982). Extensions to multivariate normal models for pedigree analysis. *Ann. Hum. Genet.* **46**:373–383.

Keavney, B., McKenzie, C. A., Connell, J. M., Julier, C., Ratcliffe, P. J., *et al.* (1998). Measured haplotype analysis of the angiotensin-I converting enzyme gene. *Hum. Mol. Genet.* **7**: 1745–1751.

Kruglyak, L. (1999). Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature Genet.* **22**:139–144.

Kruglyak, L., and Lander, E. S. (1995). Complete multipoint sib-pair analysis of qualitative and quantitative traits. *Am. J. Hum. Genet.* **57**:439–464.

Lander, E. S. (1999). Array of hope. *Nature Genet.* **21**:3–4.

Lange, K., Westlake, J., and Spence, M. A. (1976). Extensions to pedigree analysis. III. Variance components by the scoring method. *Ann. Hum. Genet.* **39**:485–491.

Neale, M. C., and Cardon, L. R. (1992). *Methodology for Genetic Studies of Twins and Families*, Kluwer Academic Press, Boston.

Nickerson, D. A., Taylor, S. L., Weiss, K. M., Clark, A. G., Hutchinson, R. G., *et al.* (1998). DNA sequence diversity in a 9.7-kb region of the human lipoprotein lipase gene. *Nature Genet.* **19**:233–240.

Rabinowitz, D. (1997). A transmission disequilibrium test for quantitative trait loci. *Hum. Hered.* **47**:342–350.

Rieder, M. J., Taylor, S. L., Clark, A. G., and Nickerson, D. A. (1999). Sequence variation in the human angiotensin converting enzyme. *Nature Genet.* **22**:59–62.

Risch, N., and Merikangas, K. (1996). The future of genetic studies of complex human diseases. *Science* **273**:1516–1517.

Searle, S. R., Casella, G., and McCulloch, C. E. (1992). *Variance Components*, John Wiley & Sons, New York.

Sham, P. C., Cherny, S. S., Purcell, S., and Hewitt, J. K. (2000). Power of linkage versus association analysis of quantitative traits using variance components models for sibship data. *Am. J. Hum. Genet.* **66**:1616–1630.

Spielman, R. S., and Ewens, W. J. (1996). The TDT and other family-based tests for linkage disequilibrium and association. *Am. J. Hum. Genet.* **59**:983–989.

Spielman, R. S., McGinnis, R. E., and Ewens, W. J. (1993). Transmission test for linkage disequilibrium: The insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am. J. Hum. Genet.* **52**:506–516.

Terwilliger, J. D., and Weiss, K. M. (1998). Linkage disequilibrium mapping of complex disease: Fantasy or reality? *Curr. Opin. Biotechnol.* **9**:578–594.

Weir, B. S. (1996). *Genetic Data Analysis II*, Sinauer Associates, Sunderland, MA.

Weiss, K. M. (1993). *Genetic Variation and Human Disease*, Cambridge University Press, Cambridge.

Weiss, K. M. (1996). Is there a paradigm shift in genetics? Lessons from the study of human diseases. *Mol. Phylogenet. Evol.* **5**:259–265.

Xiong, M. M., Krushkal, J., and Boerwinkle, E. (1998). TDT statistics for mapping quantitative trait loci. *Ann. Hum. Genet.* **62**:431–452.

Edited by Michael Neale