Behavior Genetics, Vol. 34, No. 2, March 2004 (© 2004)

Combined Linkage and Association Tests in Mx

D. Posthuma,[1,3] E. J. C. de Geus,[1] D. I. Boomsma,[1] and M. C. Neale[2]

Statistical methods aimed at the detection of genes for quantitative traits suffer from two problems: (i) when a linkage approach is employed, relatively large sample sizes are usually required; and (ii) when an association approach is employed, effects of population stratification may blur genuine locus–trait associations. The variance components method proposed by Fulker *et al.* (1999) addressed both these problems; it is statistically powerful because it involves a combined analysis of linkage and association and can include information from multiplex families, which reduces the overall amount of necessary individual genotypes. In addition, it includes an explicit test for the presence of spurious association. After a brief illustration of the various ways in which population stratification may affect locus–trait associations, the implementation in Mx (Neale, 1997) of the method as proposed by Fulker *et al.* (1999) is discussed and illustrated. In addition, an extension to this method is proposed that allows the use of (variable) sibship sizes greater than two, the estimation of additive and dominance association effects, and the use of multiple alleles. These extensions can be implemented when parental genotypes are available or unavailable.

KEY WORDS: QTL; population stratification; structural equation modeling; variance components modeling; quantitative trait.

INTRODUCTION

Statistical methods aimed at the detection of quantitative trait loci (QTLs) have primarily focused on detecting linkage between a QTL (or a marker in linkage disequilibrium with the QTL) and a trait (e.g., Almasy and Blangero, 1998; Amos, 1994; Boomsma and Dolan, 1998; Eaves *et al.,* 1996; Fulker and Cardon, 1994; Fulker and Cherny, 1996; Goldgar, 1990; Haseman and Elston, 1972; Schork, 1993). Recently, however, attention has shifted toward methods designed to detect *associations* between QTLs and traits (e.g., Abecasis *et al.,* 2000; Fulker *et al.,* 1999; Lesch *et al.,* 1996; Plomin *et al.,* 2001). Under certain conditions, testing for association can be more powerful than testing for linkage (Risch, 2000; Risch and Merikangas, 1996; Sham *et al.,* 2000), even without assuming that one of the typed markers is the actual trait locus (Long and Langley, 1999).

A widely used design to test for an association between a locus and a trait is the case-control design. This design, however, is sensitive to the effects of population stratification that may confound genuine locus—trait associations (Hamer and Sirota, 2000). Spurious associations may arise in a population that is a mix of two or more genetically distinct subpopulations. Any trait that is more frequent in one of the subpopulations compared to the other subpopulation(s) (e.g., because of cultural differences or assortative mating) will show a statistical association with any allele that has a different frequency across those two populations (e.g., as a result of different ancestors or genetic drift). This association is called spurious because within each population the allele is unrelated to variation in the trait. In practice, more than two populations may have combined and it will not be obvious from the combined populations whether or not the sample is stratified and in what way.

[1] Department of Biological Psychology, Vrije Universiteit Amsterdam, The Netherlands.

[2] Virginia Institute of Psychiatric and Behavioral Genetics, Virginia Commonwealth University, Richmond, Virginia.

[3] To whom correspondence should be addressed at Vrije Universiteit, Department of Biological Psychology, van der Boechorststraat 1, 1081 BT, Amsterdam, The Netherlands. Tel: +31-20-444-8814; Fax: +31-20-444-8832. E-mail: danielle@psy.vu.nl

Population stratification is often considered the culprit for nonreplication of previously found associations (Cardon and Bell, 2001; Ioannidis *et al.,* 2001; Plomin and Caspi, 1999; Risch, 2000; Sullivan *et al.,* 2001). However, what is frequently overlooked is that population stratification is as likely to obscure genuine associations as it is to falsely introduce them. The first aim of this paper is to illustrate these opposing impacts of population stratification on association under various admixtures of subpopulations with different trait means and different allele frequencies.

To control for the confounding effects of population stratification, family-based tests have been developed in which locus–trait associations are compared across genetically related individuals. Because these individuals stem from the same stratum, locus–trait associations observed within genetically related individuals are genuine. Most available family-based tests for association have been developed for binary traits, such as the Haplotype Relative Risk test (HRR, Falk and Rubinstein, 1987; Terwilliger and Ott, 1992) and the Transmission Disequilibrium Test (TDT, Spielman *et al.,* 1993). Under the assumption of random ascertainment, a clinical binary diagnosis such as "depressed" or "not depressed" or "hypertensive" vs. "normotensive," however, is less powerful for gene finding than a continuous trait such as the score on a depression scale or blood pressure (Boomsma *et al.,* 2000; Van den Oord, 1999). For this reason the TDT has recently been extended to the analysis of quantitative traits (q-TDT; Allison, 1997; Rabinowitz, 1997). The TDT is based on the comparison of transmitted alleles from the parents to affected offspring with nontransmitted alleles. In its original form the TDT has some drawbacks: (i) it requires parental genotypes that complicates its application to late-onset diseases; (ii) two homozygous parents are noninformative, resulting in a decrease of the available sample size; and (iii) no more than one affected child per family can be included because siblings are not genetically independent. Recently, extensions of the TDT have been developed that deal with some of its original drawbacks (reviewed in Zhao, 2000).

Fulker *et al.* (1999) proposed a variance components sib-pair analysis for mapping QTL. This method is based on the modeling of allelic effects on the trait values as a test for association and simultaneous modeling of the sibship covariance structure as a test for linkage (Fulker *et al.,* 1999). By partitioning the association effects into a *between family* component and a *within family* component, spurious associations can be separated from genuine associations. The *between family* effects reflect both the genuine and the possible spurious association between locus alleles and a trait (or allelic association between locus alleles and trait locus alleles). The *within family* effects reflect only the genuine association.

When simultaneously modeling linkage (using identity by descent (IBD) information at positions across the genome) and association (using the alleles from candidate genes/markers) lying within the region that shows linkage), evidence for linkage in a genomic region is expected to decrease; by modeling the allelic effects on the trait values, the residual variance will show less evidence for linkage. If the evidence for linkage does not completely decrease in the presence of a significant genuine association effect of a marker within the linkage region, this could imply that the linkage derives from some other gene within that genomic region, that not all relevant alleles of that locus have been genotyped, or that (part of) the observed linkage may have been artefactual (i.e., because of marker genotype errors) (Abecasis *et al.,* 2000, 2001; Cardon and Abecasis, 2000; McKenzie *et al.,* 2001).

The second aim of this paper is to present an implementation of the combined linkage and association test, including the test for the presence of spurious associations. Although we will present this implementation in the context of using Mx software (Neale, 1997), the general algebraic formulas can also be implemented in other genetic software, such as MERLIN (Abecasis *et al.,* 2002) or SOLAR (Almasy and Blangero, 1998). Mx (Neale, 1997) is a matrix algebra interpreter that uses numerical optimization to obtain parameter estimates by maximum likelihood. Its flexibility allows the relative simple implementation of extensions to multiple (marker) alleles, dominance as well as additive association effects, and variable sibship sizes. In addition, either parental genotypes or sibling genotypes can be used to derive the coefficients used for the decomposition of the association into spurious and genuine effects. These extensions will also be discussed in algebraic terms and implemented in an example Mx script.

Effects of Population Stratification on Statistical Association

We start with a brief definition of some terms used in this paper and will mostly adhere to the definitions given by Terwilliger and Göring (2000). *Linkage between a marker and a trait locus* refers to the non-independent segregation of the marker and the trait locus, implying that the recombination fraction between them is less than 0.5. *Linkage between a locus and a* *trait* is related to this and denotes that pairs of genetically related individuals that share two locus alleles IBD are phenotypically more alike than pairs of genetically related individuals that share none of their alleles on the locus IBD. The locus may either be the trait locus itself or be a marker linked to the trait locus (i.e., a recombination fraction between the marker and the trait locus of less than 0.5); it is in *linkage disequilibrium* (LD), but not necessarily in *disequilibrium* with the trait locus, *LD* or *allelic association* refers to the situation in which certain alleles of a marker are preferentially cosegregated with certain alleles of a trait locus. LD may occur because two loci are in tight linkage but can also occur as a result of population stratification or when certain allele combinations at different loci confer enhanced reproductive fitness. In the latter two cases we speak of disequilibrium. *Association between a locus and a trait* refers to the apparent allelic effects of a locus on trait values. This locus may either be the trait locus itself (i.e., the actual gene) or be a marker in LD with the trait locus.

When several populations have combined, spurious association between a locus and a trait may arise. The size and direction of this association depend on the combination of allele frequencies and trait means in the subpopulations. Different trait means for the same genotypic category across subpopulations will generally result in a difference of the overall means across subpopulations, which is why a difference in overall trait means across subpopulations is generally given as a prerequisite for spurious association to occur. Yet, it should be kept in mind that the crucial events leading to spurious associations between alleles at a locus and a trait are a difference in allele frequencies at that locus and a difference in the trait means for a given *genotype* across subpopulations.

Consider two subpopulations A and B that combine to form the mixed population M. Let subpopulation A have a trait mean μ_A of 105 and subpopulation B a trait mean μ_B of 100. Consider a diallelic locus with alleles E and e and frequencies p and q, respectively, where $q = 1 - p$. Let p in subpopulation A (p_A) be 0.9 and p in subpopulation B (p_B) be 0.5. This locus contributes neither to μ_A nor to μ_B; in other words, within each subpopulation there is no association between the locus and the trait. Let μ_m and p_m denote the trait mean and the frequency of allele E, respectively, in the mixed population (M). Let P, H, and Q denote the genotypic frequencies of the three possible genotypes EE, Ee, and ee, respectively. As subpopulations A and B are in Hardy-Weinberg equilibrium (HWE), P, H, and Q may be calculated from the allele frequencies of each subpopulation

$$P_A = p_A^2,\ H_A = 2p_A q_A,\ Q_A = q_A^2$$

and

$$P_B = p_B^2,\ H_B = 2p_B q_B,\ \text{and}\ Q_B = q_B^2$$

(see also Table I).

As the locus is not related to the phenotypic trait values, the three genotypic categories have equal means within subpopulations. Across subpopulations, however, the trait means are different for individuals that have similar genotypes. Assuming equal population sizes for subpopulations A and B, mixing the subpopulations creates population M, where the genotypic frequencies P_M, H_M, and Q_M are derived from the genotypic frequencies of the two subpopulations A and B

Table I. Formulas and Hypothetical Situation Illustrating the Effects of Population Stratification in the Absence of a Genuine Association

	Population mean	Allele frequencies		Genotypic frequencies			Trait means (μ_g) for given genotype		
	μ	p(E)	q(e)	P(EE)	H(Ee)	Q(ee)	EE	Ee	ee
A	105.00	0.9	0.1	0.81	0.18	0.01	105.00	105.00	105.00
B	100.00	0.5	0.5	0.25	0.50	0.25	100.00	100.00	100.00
M	102.50	0.7	0.3	0.53	0.34	0.13	103.82	101.32	100.19

Note: Following Falconer and Mackay (1996) p denotes the frequency of allele E, $q = 1 - p$ and denotes the frequency of allele e. P, H, and Q denote the genotypic frequencies of genotypes EE, Ee, and ee, respectively. P, H, Q, p, and q in the mixed population are derived from the genotypic frequencies in the subpopulations. P_M is derived as $\sum_{t=1}^{T} P_t \times n_t / \sum_{t=1}^{T} n_t$, where n is the total sample size of subpopulation t, and $t = 1, \ldots, T$. Analogously, H_M is derived as $\sum_{t=1}^{T} H_t \times n_t / \sum_{t=1}^{T} n_t$, and Q_M is derived as $\sum_{t=1}^{T} Q_t \times n_t / \sum_{t=1}^{T} n_t$. The allele frequencies p and q in the combined population M are derived as $p_M = P_M + \frac{1}{2}H_M$ and $q_M = Q_M + \frac{1}{2}H_M$ respectively.

Two subpopulations A and B of equal size, differ both in trait means (per genotype) and in allele frequencies of a diallelic locus. Within each population no locus-trait association exists, whereas in the mixed population M a spurious locus-trait association is clearly evident.

(Table I). As is shown in Table I, P_M, H_M, and Q_M are 0.53, 0.34, and 0.13, respectively. The allele frequencies are calculated following the rules of the biometrical model (Falconer and Mackay, 1996): $p_M = P_M + \frac{1}{2} H_M$ and $q_M = Q_M + \frac{1}{2} H_M$. Note that population M is no longer in HWE.

The trait means for each genotype in population M are a function of the trait means and frequencies of each genotype in subpopulations A and B. Assuming equal population sizes, the trait mean of individuals with genotype g in population M is calculated as follows ($\mu_{g,M}$):

$$\mu_{g,M} = \frac{G_{g,A} \times \mu_{g,A} + G_{g,B} \times \mu_{g,B}}{G_{g,A} + G_{g,B}} \tag{1}$$

where $G_{g,A}$ refers to the frequency of genotype g in population A, $G_{g,B}$ refers to the frequency of genotype g in population B, $\mu_{g,A}$ refers to the trait mean for genotype g in population A, and $\mu_{g,B}$ refers to the trait mean for genotype g in population B.

For the example given in Table I, this results in different trait means for each of the three genotypic categories in population M, reflecting a spurious statistical association between the locus and the trait. Figure 1 presents this effect graphically.

In the biometrical model, which is drawn in Figure 2, a denotes the (additive) effect of genotype EE on the trait, $-a$ denotes the (additive) effect of genotype ee on the trait, and d denotes the dominance deviation for the heterozygous genotype Ee. In association analysis we aim to quantify a and d. In the situation described in Table I and Figure 1, both a and d are 0 for subpopulations A and B. From the values given in the last three cells from Table I, however, the estimated a and d in the mixed population M would be obtained as $(103.82 - 100.19)/2 = 1.82$ and $101.32 - (103.82 + 100.19)/2 = -0.69$, respectively.

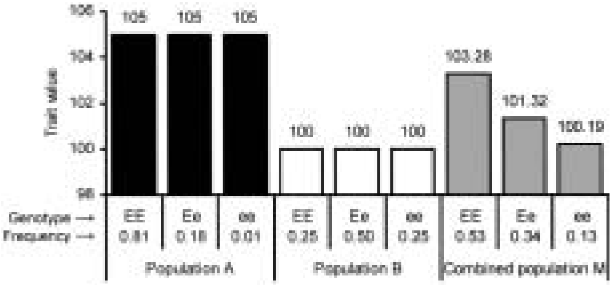

Fig. 1. Graphical representation of the effects of population stratification. Two populations A and B differ both in overall trait means (and trait means per genotype) and in allele frequencies of a diallelic locus. Within each population no locus–trait association exists, whereas in the mixed population a spurious locus–trait association is clearly evident. Specifics concerning this situation are given in Table I. Genotypes and their frequencies are given on the x-axis, whereas the trait means per genotype are scaled on the y-axis.

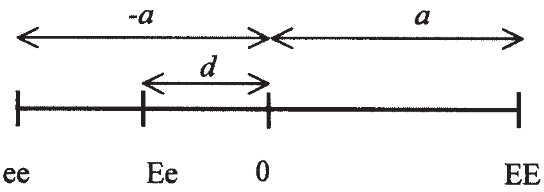

Fig. 2. Biometric model for a diallelic trait with alleles E and e. Let a be the effect of genotype EE on the trait mean, $-a$ the effect of ee, and d the dominance deviation of the heterozygous genotype Ee.

For the example given in Table I and represented in Figure 1 we used extreme allele frequency differences ($\Delta p = p_A - p_B = 0.4$) between the two subpopulations and a mean difference of 5 scale points. Figure 3a plots the effects of varying allele frequency differences between populations A and B for four $\Delta\mu$s ($\mu_A - \mu_B = 10, 5, -5,$ or -10) on the estimated value of a in the mixed population, in the absence of a genuine association (i.e., $a = 0$ in subpopulations A and B). In Figure 3b the effect on the calculated value of d in the mixed population is plotted for the same situations and a d of 0 in subpopulations A and B. The allele frequency p_B is constant at 0.5, whereas the allele frequency p_A is varied in steps of 0.01 from 0.99 to 0.01. The mean μ_B is constant at 100, whereas μ_A is 110, 105, 95, or 90.

As becomes evident from Figures 3a and b, population stratification will result in spurious associations between a locus and a trait. As the genuine a and d values were 0, the estimated a and d values in the mixed population are always biased (except when $\Delta p = 0$), and may result both in positive effects of a and d, as well as in negative values of a and d. The bias in estimation of d becomes relatively small when the difference in allele frequency between subpopulations A and B is small to moderate (between -0.3 and 0.3).

Using the same situations as described above, yet assuming a value of $+2$ for a in subpopulations A and B, shows that in the presence of a genuine association the estimated value of a in the mixed population may be overestimated, underestimated, or of reversed sign.

As the genuine dominance deviation was fixed at 0, the calculated dominance deviation from the mixed population is always biased (except when $\Delta p = 0$ or when the genotypic means are equal across populations) and is similar to the effects seen in Figure 3b. Our purpose is to clarify the different ways in which

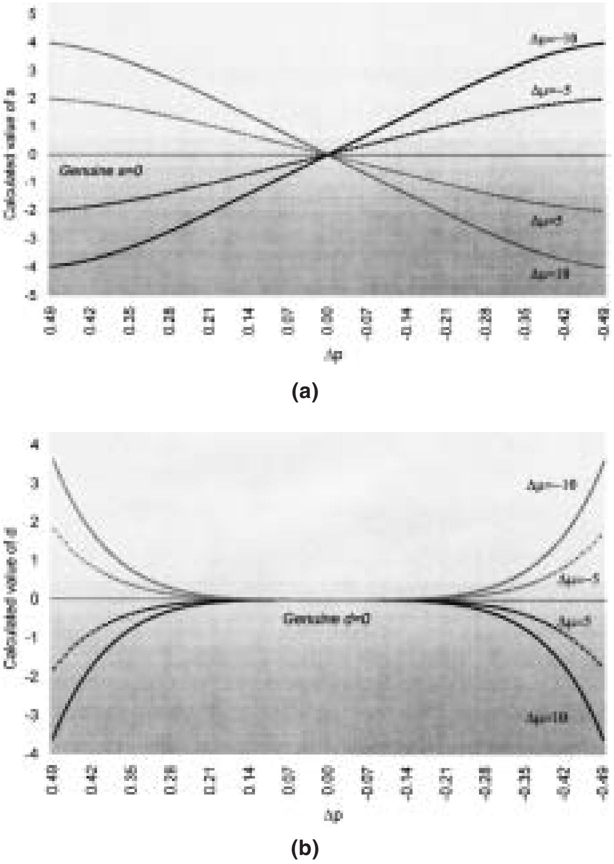

Fig. 3. a, Effect of population stratification on the calculated value of a in the absence of a genuine locus–trait association ($a = 0; d = 0$) for varying levels of allele frequency differences. The mixed population exists of populations A and B with constant μ_B (100) whereas μ_A is varied from 110, 105, 95, and 90. Allele frequency p_B is constant at 0.5. Allele frequency p_A is varied with steps of 0.01 from 0.99 to 0.01. b, Effect of population stratification on the calculated value of d in the absence of a genuine locus–trait association ($a = 0$; $d = 0$) for varying levels of allele frequency differences. The mixed population exists of populations A and B with constant μ_B(100), whereas μ_A is varied from 110, 105, 95, and 90. Allele frequency p_B is constant at 0.5. Allele frequency p_A is varied with steps of 0.01 from 0.99 to 0.01.

population stratification may affect genetic effects in general; thus we chose not to discuss situations in which a genuine dominance deviation is present.

Implementing the Test for Combined Linkage and Association in Mx

Modeling Spurious and Genuine Association

When allelic effects are estimated from genetically related subjects, effects of population stratification can be controlled for. The method proposed by Fulker *et al.,* 1999 uses the *within family genetic effects* on the trait value as an estimate of the genuine association. The *between family genetic effects* on the trait value include both the genuine and the possible spurious association. When the *between family genetic effects* and the *within family genetic effects* are unequal, a spurious association is said to exist, which may either be in the same direction (*between genetic effects* > *within genetic effects*) or in the opposite direction (*between genetic effects* < *within genetic effects*) compared to the genuine association. Thus, equating the *between effects* and the *within effects* serves as a test of the presence (and direction) of spurious associations between a locus and a trait in the data set. This test can be conducted on DNA markers as well as candidate genes.

Estimation of the *between genetic effects* is based on defining the contribution of each family or sibship to the population mean in terms of genetic effects. Thus, for each sibship the genetic mean needs to be calculated. Estimation of the *within genetic effects* is based on defining each individual's genetic deviation from the genetic mean of his sibship. The genetic family/sibship mean can be calculated using the sibling genotypes (if parental genotypes are unavailable) or using the parental genotypes (if available). In this section the implementation in Mx (Neale, 1997) of the combined linkage and association method for these two situations (parental genotypes unavailable and parental genotypes available) as can be applied to real data, is discussed.

Parental Genotypes Unavailable

In Table II the coefficients for the *within* (genuine) and *between* (possibly spurious and genuine) additive and dominance effects are derived for a diallelic locus using sibpairs. The general expression for the means, following Fulker *et al.* (1999) yet including both additive effects and dominance, for the observed score in sib j from the i^{th} family (y_{ij}) is:

$$y_{ij} = \mu + a_b A_{bi} + a_w A_{wij} + d_b D_{bi} + d_w D_{wij} + e_{ij} \qquad (2)$$

where μ denotes the overall trait mean (equal for all individuals), A_{bi} is the derived coefficient (e.g., $\frac{1}{2}$, or $-\frac{1}{2}$, 1, etc.) for the *between families* additive genetic effect for the i^{th} family, as calculated in the fifth column of Table II. A_{wij} denotes the coefficient by which the *within families* additive genetic effects need to be multiplied for sib j from the i^{th} family as derived in the last two columns of Table II. D_{bi} is the coefficient by which the *between families* dominant genetic effect needs to be multiplied for the i^{th} family, as calculated in the fifth column of Table II. D_{wij} denotes the coefficient as derived for the *within families* dominant genetic effects for sib j from the i^{th} family (see last two columns of Table II). Parameters a_b and a_w are the estimated additive *between* and *within* effects; parameters d_b and d_w are the estimated dominance *between* and *within* effects, and e_{ij} denotes that part of the grand mean that is not explained by the genotypic effects.

For a diallelic locus, derivation of the additive *between* and *within* coefficients and the dominance *between* and *within* coefficients is straightforward and can be taken from Table II (e.g., $\frac{1}{2}$, or $-\frac{1}{2}$, 1, etc.). For a locus with more than two alleles, however, this becomes a daunting task. We therefore chose to have the necessary coefficients calculated by the program instead of specifying them in a matrix (e.g., Neale, 2000; Neale *et al.,* 1999).

Let matrices **A** and **C** be vectors of dimensions $1 \times n$, where $n = 2, \ldots, n$ for the number of alleles at the locus. Let matrices **D** and **F** be subdiagonal matrices of dimensions $n \times n$. Matrix **A** contains the estimated combined spurious and genuine (i.e., *between*) additive allelic effects. Matrix **C** contains the estimated genuine additive (i.e., *within*) allelic effects. Matrix **D** contains the estimated spurious and genuine (i.e., *between*) dominance deviations for the heterozygous genotypes, and matrix **F** contains the estimated genuine (i.e., *within*) dominance deviations. Let matrix **I** be a vector containing one's of dimension $1 \times n$. In the Mx script language this is written (see also Appendices I and II for full Mx script example; anything after ! on the same line is not read by the Mx program and can be used for additional remarks):

```
#define n 5                !number of alleles = 5; the letter n will be substituted
                           !by the number 5, except when n occurs as part of a word
Begin matrices;            !start declaration of matrices
  A Full 1 n  free         !will contain additive allelic effects WITHIN
  C Full 1 n  free         !will contain additive allelic effects BETWEEN
  D Sdiag n n free         !will contain dominance deviations within
  F Sdiag n n free         !will contain dominance deviations between
  I Unit 1 n               !unit vector to multiply allelic effects [1 1 1 1 1]
End matrices;              !end declaration of matrices
```

With these matrices, two symmetric matrices of dimensions $n \times n$, one for the *between* (i.e., the sum of the spurious and genuine effects) and one for the *within* (i.e., the genuine effects) estimates, are calculated that contain the genotypic effects of the homozygous genotypes on the diagonal and the genotypic effects of the heterozygous genotypes on the off-diagonals.

```
Begin algebra;
  K = (A'@I) + (A@I') ;    !calculates linear combinations of the allelic effects
  L = D + D' ;             !dominance deviations below and above diagonal
  W = K + L ;              !creates one full n x n matrix containing the WITHIN
                           !genotypic effects
  M = (C'@I) + (C@I') ;    !calculates linear combinations of the allelic effects
  N = F + F' ;             !dominance deviations below and above diagonal
  B = M + N ;              !creates one full n x n matrix containing the BETWEEN
                           !genotypic effects
End algebra;
```

The symbol @ denotes the Kronecker product (⊗) in Mx and results in the multiplication of each element of the first matrix by the second matrix. For a locus with five alleles, matrix **W** is a symmetric matrix of dimension $n \times n$ containing the following estimated effects for a locus with five alleles ($n = 5$):

$$\mathbf{W} \left\{ \begin{array}{lllll} a_{w,1} + a_{w,1} & & & & \\ a_{w,1} + a_{w,2} + d_{w,12} & a_{w,2} + a_{w,2} & & & \\ a_{w,1} + a_{w,3} + d_{w,13} & a_{w,2} + a_{w,3} + d_{w,23} & a_{w,3} + a_{w,3} & & \\ a_{w,1} + a_{w,4} + d_{w,14} & a_{w,2} + a_{w,4} + d_{w,24} & a_{w,3} + a_{w,4} + d_{w,34} & a_{w,4} + a_{w,4} & \\ a_{w,1} + a_{w,5} + d_{w,15} & a_{w,2} + a_{w,5} + d_{w,25} & a_{w,3} + a_{w,5} + d_{w,35} & a_{w,4} + a_{w,5} + d_{w,45} & a_{w,5} + a_{w,5} \end{array} \right.$$

where $a_{w,1\ldots n}$ refers to the genuine additive allelic effects of the alleles labeled $1 \ldots n$, and $d_{w,12\ldots nn}$ refers to the genuine dominance deviation of the heterozygous genotypes labeled $12 \ldots nn$. Note that with this notation $a_{w,1\ldots}a_{w,n}$ refers to *allelic* effects, whereas a_w refers to *genotypic* effects. Similarly, matrix **B** will be symmetric, of dimension $n \times n$ and will contain the analogous estimated genuine and spurious additive and dominance genotypic effects (subscripted b).

We now proceed to the calculation of the sibship genetic means and each individual's deviation from the sibship's genetic mean. For sibships of size two, each individual's deviation from the sibship genetic mean can easily be deducted by precalculating half the difference between the genetic effects of each sib (as is done in Table II). For sibship sizes larger than two, the *within* component is no longer simply "half the difference," but instead is mathematically represented by the deviation of sib *j* from the i^{th} sibship mean. The individual genotypes should be in the datafile (which is the "raw" datafile and not a variance/covariance matrix). These are selected from the list of input variables to be *used* and will be specified in a matrix. They need to be treated differently from variables that are to be *analyzed* (the phenotype). The `definition variable` function in Mx can be used to separate variables that are used as covariates (such as sex, age, and allelic effects) from the dependent variables.

```
G2: datagroup
Select pheno1 pheno2 pheno3 a1s1 a2s1 a1s2 a2s2 a1s3 a2s3  !Select all variables to
                                                           !be used or analysed
..
Definition_variables a1s1 a2s1 a1s2 a2s2 a1s3 a2s3         !define which variables
                                                           !need to be treated as a
                                                           !covariate
Begin matrices ;                  !begin declaration of matrices for group 2
..
  K Full 1 4 Fixed                !Will contain first and second allele of sib1
  L Full 1 4 Fixed                !Will contain first and second allele of sib2
  M Full 1 4 Fixed                !Will contain first and second allele of sib3
..
End matrices ;                    !end declaration of matrices for group 2
Specify K a1s1 a2s1 a1s1 a2s1     !put alleles of sib 1 into vector
Specify L a1s2 a2s2 a1s2 a2s2     !put alleles of sib 2 into vector
Specify M a1s3 a2s3 a1s3 a2s3     !put alleles of sib 3 into vector
```

For each individual, two alleles need to be present in the data file. The alleles should be coded as 1, 2, 3, ..., *n*. For each sibship, different elements need to be taken from matrices **B** and **W** to calculate the family genetic mean and each individual's deviation from that mean. The definition variables that have now been put into matrices (**K, L,** and **M**) that contain numbers that correspond to the specific alleles from the respective individual. For example, if the first sib has genotype 11, the second sib has genotype 34, and the third sib has genotype 13 at a marker locus, matrix **K** contains [1 1 1 1], matrix **L** contains [3 4 3 4], and matrix **M** contains [1 3 1 3].

Matrices **K, L,** and **M** can now be used to select the relevant cells from matrices **B** and **W:**

```
!For sibships of size 3 for a univariate trait
  Begin matrices
    B Computed n n  = B1              !spurious and genuine genotypic effects,
                                      !precalculated in previous Mx group
    W Computed n n  = W1              !genuine genotypic effects
    S Full 1 1 Fixed                  !to contain sibshipsize (3)
    G Full 1 1 Free                   !grand mean, to be estimated
                                      !dimensions 1 x number of variables
  End matrices
   Matrix S 3                         !sibship size = 3
  Begin Algebra;
      V = (\part(B,K) + \part(B,L) + \part(B,M) ) % S ;
     !sib genetic mean: between effects (spurious and genuine)
     D = (\part(W,K) + \part(W,L) + \part(W,M) ) % S ;
     !used for individual's deviation from sib mean: within effects (genuine)
  End Algebra;
  Means G + V + (\part(W,K)-D) | G + V + (\part(W,L)-D) | G + V + (\part(W,M)-D) ;
     !means model: grand mean + sib genetic mean effects + individual's deviation
     !from sib genetic mean, for three sibs
```

The `\part` statement in Mx allows one to select a rectangular submatrix from a larger matrix. For example, \part(**B,K**) tells Mx to select from matrix **B** the part specified in matrix **K**. Matrix **K** should always be of dimension 1 × 4 (start row, start column, end row, end column) and specifies the elements of matrix **B** where the

Table II. Partitioning of Additive and Dominance Genotypic Effects into Between and Within Components for a Diallelic Locus with Alleles E and e in Sib-pairs

Genotype		Genotypic effect		Additive		Dominance		Partitioned genotypic effects	
Sib 1	*Sib 2*	*Sib 1*	*Sib 2*	*Mean*	*Difference/2*	*Mean*	*Difference/2*	*Sib 1*	*Sib 2*
EE	EE	a	a	a_b	0	0	0	a_b	a_b
EE	Ee	a	d	$\frac{1}{2}a_b$	$\frac{1}{2}a_w$	$\frac{1}{2}d_b$	$-\frac{1}{2}d_w$	$(\frac{1}{2}a_b + \frac{1}{2}a_w) + (\frac{1}{2}d_b - \frac{1}{2}d_w)$	$(\frac{1}{2}a_b - \frac{1}{2}a_w) + (\frac{1}{2}d_b + \frac{1}{2}d_w)$
EE	ee	a	−a	0	a_w	0	0	a_w	$-a_w$
Ee	EE	d	a	$\frac{1}{2}a_b$	$-\frac{1}{2}a_w$	$\frac{1}{2}d_b$	$\frac{1}{2}d_w$	$(\frac{1}{2}a_b - \frac{1}{2}a_w) + (\frac{1}{2}d_b + \frac{1}{2}d_w)$	$(\frac{1}{2}a_b + \frac{1}{2}a_w) + (\frac{1}{2}d_b - \frac{1}{2}d_w)$
Ee	Ee	d	d	0	0	d_b	0	d_b	d_b
Ee	ee	d	−a	$-\frac{1}{2}a_b$	$\frac{1}{2}a_w$	$\frac{1}{2}d_b$	$\frac{1}{2}d_w$	$(-\frac{1}{2}a_b + \frac{1}{2}a_w) + (\frac{1}{2}d_b + \frac{1}{2}d_w)$	$(-\frac{1}{2}a_b - \frac{1}{2}a_w) + (\frac{1}{2}d_b - \frac{1}{2}d_w)$
ee	EE	−a	a	0	$-a_w$	0	0	$-a_w$	a_w
ee	Ee	−a	d	$-\frac{1}{2}a_b$	$-\frac{1}{2}a_w$	$\frac{1}{2}d_b$	$-\frac{1}{2}d_w$	$(-\frac{1}{2}a_b - \frac{1}{2}a_w) + (\frac{1}{2}d_b - \frac{1}{2}d_w)$	$(-\frac{1}{2}a_b + \frac{1}{2}a_w) + (\frac{1}{2}d_b + \frac{1}{2}d_w)$
ee	ee	−a	−a	$-a_b$	0	0	0	$-a_b$	$-a_b$

relevant submatrix (which can also be a single element) starts and ends. Because matrix **K** contains the alleles of an individual, the submatrix is selected conditional on that individual's genotype.

In our example, in which the first sib is of genotype 11, the second sib has genotype 34, and the third sib has genotype 13, the mean of the estimates in cells (denoted by row and column) 11, 34, and 13 from matrix **B** is calculated as the sibship genetic mean (representing the *between family* effects of that sibship, in matrix **V**). Similarly, for the first sib the *within family* effect is calculated by subtracting the estimate in cell 11 from matrix **W** from the mean of the parameters in cells 11, 34, and 13 from matrix **W** (i.e., `(\part (W,K)-D)`).

Because of linear dependency between the allelic effects, two constraint groups (one for the *within* effects and one for the *between* effects) are needed in which the sum of all the allelic effects is constrained to be 0 (see Appendices I and II).

Abecasis *et al.* (2000) showed that calculation of the sibship genetic mean based on both parental genotypes is less error prone than calculation of the sibship genetic mean based on available sibling genotypes. For sibship sizes of four and above the two methods are equally powerful and error rates are closer to nominal significance rates. The above method can be used when genotype information from *both* parents is unavailable.

Parental Genotypes Available

When both parental genotypes are available, the expected mean additive genotypic value of the offspring (a_{bi}) equals the midparental genotypic value

$$a_{bi} = \frac{G_{iF} + G_{iM}}{2}, \tag{3}$$

where $G_{i,F}$ is the additive genotypic value of the father in family i, and $G_{i,M}$ is the additive genotypic value of the mother in family i.

When dominance effects are also considered, the midparental genotypic value is no longer an estimate of the expected offspring mean, because parents and offspring are uncorrelated in terms of dominance effects. The genotypes of the parents, however, do provide information on the expected dominance effects in the offspring. For example, when one parent is of genotype EE, with a corresponding genotypic value of a, and the other parent is of genotype ee, with a corresponding value of $-a$, the midparental genetic value will be 0. However, all of their offspring will be of genotype Ee, with a corresponding genetic value of d.

For each type of parental mating we therefore need to calculate all possible genotypes in the offspring and their probability, given the parental mating type. The mean value in terms of a and d of the possible genotypes in the offspring weighted by their probability gives the expected offspring (i.e., sibling) genetic mean. In Table III the coefficients for additive and dominance between and within effects are derived, conditional on the parental genotypes.

Extending this to a multiallele locus quickly becomes a large undertaking, and it is more convenient to use a program such as Mx that can calculate the necessary coefficients (A_{bi}, A_{wijk}, D_{bi}, and D_{wijk}) by which the effects (a_b, a_w, d_b, and d_w) need to be multiplied conditional on the parental genotypes. For a given parental mating type, the possible genotypes of offspring and their probabilities may be calculated in Mx by using the parental alleles to select elements from the matrices that contain the *between* and *within* effects (matrices **B** and **W**). Whereas in the previous section both alleles that were used to select from matrices **B** and **W** were from the same person (i.e., one sib), we now pair paternal and maternal alleles to obtain all possible genotypes of the offspring. The maximum number of genotypic categories in the offspring from one mating type is four (i.e., when both parents are heterozygous and have four different alleles). We thus specify in Mx the following matrices:

```
Specify N a1p1 a1p2 a1p1 a1p2    !first allele parent one first allele parent two
Specify O a1p1 a2p2 a1p1 a2p2    !first allele parent one second allele parent two
Specify X a2p1 a1p2 a2p1 a1p2    !second allele parent one first allele parent two
Specify Y a2p1 a2p2 a2p1 a2p2    !second allele parent one second allele parent two
```

These are used to select relevant submatrices from matrix **B** and **W** to calculate the genetic offspring (i.e., sibship) mean and each offspring's individual deviation from that mean (see Appendix II for the full Mx script). The additive and dominance coefficients can be calculated in Mx in this manner for an arbitrary number of alleles and an arbitrary number of offspring.

Modeling Linkage

Implementation of the linkage component in the variance components model is straightforward and can be done by using the "pi-hat" ($\hat{\pi}$) approach, in which the covariance resulting from the marker or trait locus for a sibpair is modeled as a function of the IBD status of that sibpair. Generally, for sibships, the phenotypic variance is decomposed in familial variance (σ_f^2), variance resulting from nonshared environmental influences (σ_e^2), additive variance from the QTL or marker in LD with the QTL (σ_a^2), and dominance variance resulting from the QTL, or a marker in LD with the QTL (σ_d^2). The variance-covariance matrix for the i^{th} family, Ω_{ijk} is then given by

$$\Omega_{ijk} = \begin{cases} \sigma_f^2 + \sigma_a^2 + \sigma_d^2 + \sigma_e^2 & \text{if } j = k \\ \sigma_f^2 + \hat{\pi}_{ijk}\sigma_a^2 + \hat{z}_{ijk}\sigma_d^2 & \text{if } j \neq k \end{cases} \tag{4}$$

where $\hat{\pi}_{ijk}$ is the estimated proportion of alleles shared IBD between sibs j and k for the i^{th} family, and $\hat{z}_{ijk}$ is the probability of complete IBD sharing between sibs j and k for the i^{th} family. The estimated proportion of alleles shared IBD between sibs j and k ($\hat{\pi}_{ijk}$) is based on the probabilities that sibs j and k share 0, one, or two alleles IBD ($p_{(\text{IBD}=0)}$, $p_{(\text{IBD}=1)}$, $p_{(\text{IBD}=2)}$, respectively) that can be obtained from genetic software such as Genehunter (Kruglyak *et al.,* 1996). The formula to obtain $\hat{\pi}_{ijk}$ for the i^{th} family is given by

$$\hat{\pi}_{ijk} = 0 \times p_{(\text{IBD}=0)_{ijk}} + 0.5 \times p_{(\text{IBD}=1)_{ijk}} + 1 \times p_{(\text{IBD}=2)_{ijk}} \tag{5}$$

The probability of complete IBD sharing between sibs j and k for the i^{th} family simply equals pIBD2$_{ikj}$:

$$\hat{z}_{ijk} = p_{(\text{IBD}=2)_{ijk}} \tag{6}$$

Tests

The test for spurious association consists of the joint test that matrix **A** equals matrix **C** (from the first group in our example script), and that matrix **D** equals matrix **F** (from the first group in our example script). If the parameters in these matrices cannot be constrained to be equal, there is evidence of spurious association. The conservative test for the presence of a genuine association is to test whether matrices **A** and **D** are significantly different from 0.

The test for the presence of dominance effects can be conducted by comparing the minus two loglikelihoods (−2LL's) from the full model and a model without the subdiagonal matrices **D** and **F** from the first group in the example Mx script that contain the deviations of the heterozygous genotypes from the mid value of the two corresponding homozygous genotypes. This can be done conservatively only for the presence of the genuine dominance effects (i.e., dropping matrix **D**) or for the presence of both the genuine and spurious dominance effects (dropping matrices **D** and **F**).

Three models may be evaluated to test whether linkage is present and whether the linkage component

Table III. Partitioning of Additive and Dominance Genotypic Effects into Between and Within Components for a Diallelic Locus with Alleles E and e Conditional on Parental Genotypes

	Offspring						
	EE		Ee		ee		
	a		*d*		$-a$		
Parental mating type ↓	Probability	Deviation from family mean (*Within*)	Probability	Deviation from family mean (*Within*)	Probability	Deviation from family mean (*Within*)	Family mean (*Between*)
EE × EE	1	0	0	n.p.	0	n.p.	a_b
EE × Ee	$\frac{1}{2}$	$\frac{1}{2}a_w - \frac{1}{2}d_w$	$\frac{1}{2}$	$-\frac{1}{2}a_w + \frac{1}{2}d_w$	0	n.p.	$\frac{1}{2}a_b + \frac{1}{2}d_b$
EE × ee	0	n.p.	1	0	0	n.p.	d_b
Ee × Ee	$\frac{1}{4}$	$a_w - \frac{1}{2}d_w$	$\frac{1}{2}$	$\frac{1}{2}d_w$	$\frac{1}{4}$	$-a_w - \frac{1}{2}d_w$	$\frac{1}{2}d_b$
Ee × ee	0	n.p.	$\frac{1}{2}$	$\frac{1}{2}d_w + \frac{1}{2}a_w$	$\frac{1}{2}$	$-\frac{1}{2}d_w - \frac{1}{2}a_w$	$\frac{1}{2}d_b - \frac{1}{2}a_b$
ee × ee	0	n.p.	0	n.p.	1	0	$-a_b$

Note: The expectation for an individual sibling is the sum of the *Between* and *Within* components.
n.p., not possible.

can be partly or completely explained by the association: (i) a model with a linkage component only; (ii) a model with both linkage and association; (iii) a model with the association component only. If the linkage component is reduced in model (ii) as compared to model (i), but still significant, this may indicate that within the linkage region another gene, besides the gene used for the association component, is also influencing the trait, that not all relevant alleles of that locus have been genotyped, or that LD between the marker and the trait locus is incomplete. If the linkage component disappears when modeled simultaneously with association, it indicates that the linkage is completely explained by the association effects of the tested locus or by the effects of another locus that is in complete LD with the tested locus.

Practical Considerations

The implementation in Mx of the analysis as proposed by Fulker *et al.* (1999) is flexible in terms of the number of alleles it can incorporate, variable sibship sizes, the inclusion of both additive and dominance effects and can be used both when parental genotypes are available or unavailable. Theoretically, it may include loci with an unlimited number of alleles. With an increasing number of alleles, however, the chance increases that not all possible genotypes are present in the sample. This should be explored beforehand, and the corresponding elements in matrices **A, C, D,** and **F** containing the allelic effects and dominance deviations should be fixed to prevent nonidentification. For example when alleles labelled 3 and 4 do not exist in a heterozygous genotype, the dominance deviation for genotype “3,4” cannot be estimated. Element 3,4 from matrices **D** and **F** needs to be constrained at 0. If, on the other hand, two alleles only occur in a heterozygote, the additive effects cannot be distinguished from the dominance deviation and either one cannot be estimated. Related to this, it is also possible to group certain alleles as if they were one allele (or different alleles with the same effect) and to contrast the effect of one allele against the effects of all other alleles. This can be implemented in Mx by using constraints on the corresponding matrix elements containing the allelic effects. If alleles that differ in size are used (e.g., variable number tandem repeats [VNTRs], a linear regression of allele size may be incorporated into the model (see for example Zhu *et al.,* 1999).

Sibship size may vary across families. In this case one may use the `variable length` datafile option in Mx and use sibship size (specified in Matrix **S** from the second Mx group in the example script) as a `definition variable`, which is read from the datafile and varies across families. The simultaneous implementation of an arbitrary number of alleles, for an arbitrary sibship size, using parental genotypes or sibling genotypes, and decomposing both the additive effect and the dominance deviations into genuine and spurious effects is unique to Mx.

CONCLUSION

We have illustrated the effects of population stratification on quantitative traits and have shown that in the absence of a genuine association, population stratification may result in a spurious association between any trait that differs in mean between subpopulations and any locus that differs in allele frequency between subpopulations. This situation is illustrated by the well-known "chopsticks gene" example as described by Hamer and Sirota (2000). As was also mentioned by Witte *et al.* (1999; for binary traits), population stratification may not only result in overestimation of allele effects on quantitative traits, but also in an underestimation. More specifically, in the presence of a genuine association population, stratification may result in: (i) an overestimation of the genuine association effects, (ii) an underestimation of the genuine association effects, or (iii) a reversal or incorrect direction of allelic effects.

Genuine association effects will be overestimated because of the effects of population stratification when within the subpopulations' higher trait values are associated with a higher frequency of the increaser allele and lower trait values are associated with a lower increaser allele frequency. Or, in other words, a positive $\Delta p(p_A - p_B)$ is related to a positive $\Delta\mu(\mu_A - \mu_B)$, and a negative Δp to a negative $\Delta\mu$ (see also Figure 4). In this case we may speak of *concordant pairing* of allele frequency and trait value. In practice, such a situation may exist, for example, as a result of assortative mating within subpopulations that differ in trait means and allele frequencies. Differences in trait means and allele frequencies may exist as a result of historical or cultural differences or as a result of natural selection. For example, when in one population high trait values increase reproductive fitness, the frequency of the increaser allele for that trait and the overall trait mean may increase in that population. In the other population, in which high trait values are irrelevant for reproductive fitness, the increaser allele frequency and the overall trait mean remain the same. Assortative mating within subpopulations ensures that eventually *concordant pairing* between increaser allele and trait value will exist, and neglecting the stratified nature of the complete sample will lead to an overestimation of genetic effects.

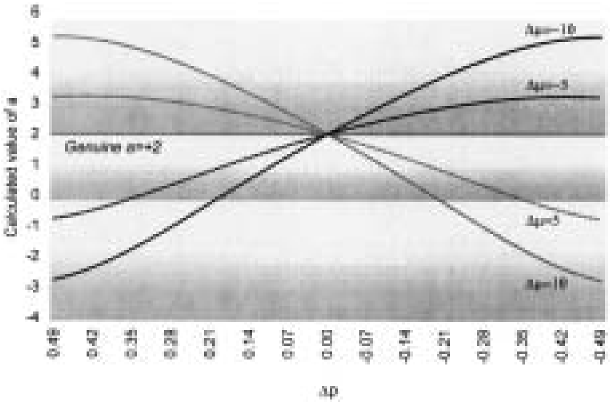

Fig. 4. Effect of population stratification on the calculated value of *a* in the presence of a genuine locus–trait association ($a = +2; d = 0$) for varying levels of allele frequency differences. The mixed population exists of populations A and B with constant μ_B (100), whereas μ_A is varied from 110, 105, 95, and 90. Allele frequency p_B is constant at 0.5. Allele frequency p_A is varied with steps of 0.01 from 0.99 to 0.01.

In the presence of a genuine association, underestimation of the additive genetic effects will occur when, within subpopulations, relatively higher trait values tend to go together with a lower frequency of the increaser allele, or vice versa (either a positive Δp and a negative $\Delta\mu$, or a negative Δp and a positive $\Delta\mu$). In this case we may speak of *discordant pairing* of allele frequency and trait value. This situation may be understood by considering that the overall mean of a subpopulation may also influenced by other (non-) genetic factors. For example, it is well known from mouse model systems, that the same allele at the same locus may cause a major disease in one mouse strain, but no phenotype in a strain with a different genetic background (e.g., Linder, 2001; Liu *et al.,* 2001; Montagutelli *et al.,* 2000). The same has been reported for effects on gene expression in different environmental backgrounds (Cabib *et al.,* 2000; Crabbe *et al.,* 1999). Put differently, in one strain the presence of the particular allele leads to crossing a certain threshold value above which a disease will evolve, whereas in the other strain, because of a different genetic or environmental background, this threshold is not reached. The frequency of the disease-predisposing allele may therefore rise in the population with the genetic or environmental background that prevents the individuals within that population from reaching a threshold. In humans, the presence of different genetic (or environmental) backgrounds that derive from mixed ethnicity may cause the allele frequency of the increaser allele of a subpopulation with a relatively low trait mean to be higher than the allele frequency of the increaser allele in a population with a higher overall trait mean.

Non-Mendelian traits are likely to be influenced by multiple (risk) factors of which the presence differs across subpopulations; thus *discordant pairing* may realistically hide genuine allele–trait associations when the effects of population stratification are neglected. When the difference in trait means between subpopulations and the difference in increaser allele frequencies becomes extreme in the presence of *discordant pairing,* the genuine allelic effects will appear reversed in sign as a result of population stratification. This suggests that in the mixed population individuals who are homozygous for the increaser alleles (EE) have a lower trait value than individuals who are homozygous for the decreaser allele (ee), whereas in the subpopulations the opposite is true. This statistical effect is known as Simpson's paradox (Simpson, 1951; Yule, 1900) and refers to the reversal of the direction of an association when data from several groups are combined to form a single group. Its importance to gene hunting studies may well have been illustrated by the numerous association studies for schizophrenia, in which the same allele of the same locus has both been associated with increased and decreased risk for schizophrenia (Baron, 2001; Bray and Owen, 2001).

Family-based tests of association explicitly model the consequences of population stratification, by looking at allelic effects within genetically related subjects. In the method proposed by Fulker *et al.* (1999) spurious association is defined as the difference between the allelic effects as estimated from the comparison of unrelated subjects (*between effects*) and the allelic effects as estimated from the comparison of genetically related subjects (*within effects*). This method, which was originally proposed to include sibpairs, diallelic markers, and additive effects, has now explicitly been extended to include variable sibship sizes, multiallele markers, and dominance deviations, using the parental genotypes (if available) or the sibling genotypes.

It is known that the use of multivariate phenotypes may provide more statistical power than univariate phenotypes (e.g., Allison *et al.,* 1998; Boomsma and Dolan, 1998). The method as implemented in Mx can easily be extended to multivariate phenotypes. One can then model the association as an effect on the factor mean of multivariate measurements. In this case it may be assumed that the allelic association effects on the multivariate measurements are all proportionally related. Covariance among the traits resulting from the association will lead to a decrease in the estimated amount of covariance because of the linkage component.

With the rapidly increasing availability of large amounts of genomic data, the detection of linkage and/or association between a marker (and all the linked loci surrounding the marker that are in LD with it) and a trait becomes a realistic tool in the hunt for genes for complex traits. Combining linkage analysis and association analysis has already proved to be a powerful tool in gene finding (e.g., Neale *et al.,* 1999; Trembath *et al.,* 1997; Zhu *et al.,* 1999; see Beekman *et al.,* 2003 for a practical implementation of the method described in the present paper). Particularly when fine mapping is a goal of interest this method is invaluable, because the effect of linkage will be reduced when estimated in the presence of association, thereby providing information on the specific region where the QTL is expected to reside (Cardon and Abecasis, 2000). An explicit test for population stratification is crucial to rule out spurious associations. The Fulker *et al.* (1999) method has all these advantages and, as was shown in the present paper, can easily be conducted in a statistical package such as Mx.

APPENDIX I: PARENTAL GENOTYPES UNAVAILABLE

```
Mx scripts can also be downloaded from the Mx homepage or from the Mx Scripts' Library:
http://www.vcu.edu/mx
http://www.psy.vu.nl/mxbib

!Mx script for the conduction of the combined linkage and association method
!testing for spurious association
!extended to sibships>2, additive and dominance association, multiple alleles
!using sibling genotypes to calculate the mean genotypic value within a sibship

#define n 5          !number of alleles is 5
#define nvar 1       !univariate
#define nsibs 3      !sibshipsize = 3
#ngroups 4           !one precalculation group, one data group, two constraint groups

G1: calculation group between and within effects
```

```
Data Calc
  Begin matrices;          !start declaration of matrices
  A  Full  1  n  free      !will contain additive allelic effects within
  C  Full  1  n  free      !will contain additive allelic effects between
  D  Sdiag  n  n free      !will contain dominance deviations within
  F  Sdiag  n  n free      !will contain dominance deviations between
  I  Unit 1 n              !unit vector to multiply allelic effects [1 1 1 1 1]
 End matrices;             !end declaration of matrices

 Begin algebra;
  K = (A'@I) + (A@I') ;
  L = D + D' ;
  W = K + L ;
  M = (C'@I) + (C@I') ;
  N = D + D' ;
  B = M + N ;
 End algebra ;
st .2 all
end

G2: datagroup: sibship size three
  Data NInput=12
  Missing =-99.00
  Rectangular File=myfile.dat
  Labels ph1 ph2 ph3 a1s1 a2s1 a1s2 a2s2 a1s3 a2s3 pi12 pi13 pi23 z12 z13 z23
  Select ph1 ph2 ph3 a1s1 a2s1 a1s2 a2s2 a1s3 a2s3 pi12 pi13 pi23 z12 z13 z23;
       !selects 3 phenotypes; one for each sib
       !selects 6 allele variables, a1s1 is allel #1 from sib #1
       !selects pi's and z's
  Definition_variables
       a1s1 a2s1 a1s2 a2s2 a1s3 a2s3 pi12 pi13 pi23 z12 z13 z23;
       !declare the allele variables and the pIBD=2 as definition variables
 Begin Matrices;
  F Lower nvar nvar Free                ! familial variance
  Q Lower nvar nvar Free                ! QTL additive variance
  R Lower nvar nvar Free                ! QTL dominance variance
  E Lower nvar nvar Free                ! non-shared environmental variance
  B Computed n n = B1                   ! spurious and genuine genotypic effects
  W Computed n n = W1                   ! genuine genotypic effects
  I Ident nsibs nsibs Fix               !
  P Sym nsibs nsibs Fix                 ! To contain pi-hats
  Z Sym nsibs snibs Fix                 ! To contain pIBD2’s
  T Stand nsibs nsibs Fix
  K Full 1 4 Fix                        ! First and second allele of sib1
  L Full 1 4 Fix                        ! First and second allele of sib2
  M Full 1 4 Fix                        ! First and second allele of sib3
  S Full 1 1 Fix                        ! to contain nsibs
  G Full 1 nvar Free                    ! grand mean
 End Matrices;
 Matrix S 3                             ! sibship size 3
 Matrix K 1 1 1 1
 Matrix L 1 1 1 1
 Matrix M 1 1 1 1
 Matrix P
       0
       1 0
       1 1 0
Matrix  Z
       0
       1 0
       1 1 0
 Specify K a1s1 a2s1 a1s1 a2s1   !genotype sib1 to be used for \part
 Specify L a1s2 a2s2 a1s2 a2s2   !genotype sib2 to be used for \part
 Specify M a1s3 a2s3 a1s3 a2s3   !genotype sib3 to be used for \part
Specify P   1
            pi12 1
            pi13 pi23 1
Specify Z   1
            z12 1
            z13 z23 1
Specify T   .5                      ! when familial variance is modeled as
            .5 .5                   ! add gen variance

Begin Algebra;
 V = (\part(B,K) + \part(B,L) + \part(B,M) ) % S ; !"B"
 D = (\part(W,K) + \part(W,L) + \part(W,M) ) % S ; !used for deviation: W
End Algebra;

Means G + V + (\part(W,K)-D) | G + V + (\part(W,L)-D) | G + V + (\part(W,M)-D);
Covariance T@(F*F') + P@(Q*Q') + Z@(R*R') + I@(E*E') ;

End

Constrain sum allelic effects = 0
Constraint ni=1
Begin Matrices;
  A full 1 n = A1
  O zero 1 1
End Matrices;
Begin algebra;
  B = \sum(A) ;
End Algebra;
Constraint O = B ;
end

Constrain sum allelic effects = 0
Constraint ni=1
Begin Matrices;
 C full 1 n = C1
 O zero 1 1
End Matrices;
Begin algebra;
 B = \sum(C) ;
End Algebra;
Constraint O = B ;
option multiple issat
end

save full.mxs

!test for spurious association W=B
Specify 1 A 101 102 103 104 105
Specify 1 C 101 102 103 104 205 !first 4 equal to within; last unequal but because
                                !of second constrain 205 will be equal to 105
Specify 1 D 801 802 803 804 805 806 807 808 809 810
Specify 1 F 801 802 803 804 805 806 807 808 809 810
end

!Drop dominance: non-conservative test (i.e. genuine and spurious)
Specify 1 D 801 802 803 804 805 806 807 808 809 810
Specify 1 F 801 802 803 804 805 806 807 808 809 810
Drop @0 801 802 803 804 805 806 807 808 809 810
end

!Drop all allelic effects: non-conservative test (i.e. genuine and spurious)
```

```
Specify 1 A 101 102 103 104 105
Specify 1 C 101 102 103 104 205
Specify 1 D 801 802 803 804 805 806 807 8O8 809 810
Specify 1 F 801 802 803 804 805 806 807 808 809 810
Drop @0 101 102 103 104 105 801 802 803 804 805 806 807 808 809 810
end

get full mxs

!drop QTL linkage effect while keeping association effects in the model
Drop Q 2 1 1 !QTL additive variance
Drop R 2 1 1 !QTL dominance variance
end
```

APPENDIX II: PARENTAL GENOTYPES AVAILABLE

```
!Mx script for the conduction of the combined linkage and association method
!testing for spurious association
!extended to sibships>2, additive and dominance association, multiple alleles
!using parental genotypes to calculate the mean genotypic value within a sibship

#define n 5        !number of alleles is 5
#define nvar 1     !univariate
#define nsibs 3    !sibshipsize = 3
#ngroups 4         !one precalculation group, one data group, two constraint groups

G1: calculation group between and within effects
Data Calc
   Begin matrices;       !start declaration of matrices
   A Full 1 n free       !will contain additive allelic effects within
   C Full 1 n free       !will contain additive allelic effects between
   D Sdiag n n free      !will contain dominance deviations within
   F Sdiag n n free      !will contain dominance deviations between
   I Unit 1 n            !unit vector to multiply allelic effects [1 1 1 1 1]
  End matrices;          !end declaration of matrices
  Begin algebra;
   K = (A'@I)+(A@I') ;
   L = D + D' ;
   W = K + L ;
   M = (C'@I)+(C@I') ;
   N = F + F' ;
   B = M + N ;
  End algebra ;
st .2 all
end

G2: datagroup: sibship size three
  Data NInput=12
  Missing =-99.00
  Rectangular File=myfile.dat
  Labels ph1 ph2 ph3 a1p1 a2p1 a1p2 a2p2 a1s1 a2s1 a1s2 a2s2 a1s3 a2s3 pi12 pi13
pi23 z12 z13 z23
  Select ph1 ph2 ph3 a1p1 a2p1 a1p2 a2p2 a1s1 a2s1 a1s2 a2s2 a1s3 a2s3 pi12 pi13
pi23 z12 z13 z23;
       !selects 3 phenotypes; one for each sib
       !selects 6 allele variables for sib, a1s1 is allel #1 from sib #1
       !selects 4 allele variables for parents a1p1 is allel #1 parent #1
       !selects pi's and z's
  Definition_variables
       a1p1 a2p1 a1p2 a2p2 a1s1 a2s1 a1s2 a2s2 a1s3 a2s3 pi12 pi13 pi23 z12 z13 z23;
       !declare the allele variables and the pIBD=2 as definition variables
```

```
Begin Matrices;
   F Lower nvar nvar Free                !familial variance
   Q Lower nvar nvar Free                !QTL additive variance
   R Lower nvar nvar Free                !QTL dominance variance
   E Lower nvar nvar Free                !non-shared environmental variance
   B Computed n n = B1                   !spurious and genuine genotypic effects
   W Computed n n = W1                   !genuine genotypic effects
   I Ident nsibs nsibs Fix               !To multiply E
   P Sym nsibs nsibs Fix                 !To contain pi-hats and to multiply Q
   Z Sym nsibs snibs Fix                 !To contain pIBD2's and to multiply R
   T Stand nsibs nsibs Fix               !To multiply F
   K Full 1 4 Fix                        !First and second allele of sib1
   L Full 1 4 Fix                        !First and second allele of sib2
   M Full 1 4 Fix                        !First and second allele of sib3
   N Full 1 4 Fix                        !a1p1 a1p2
   O Full 1 4 Fix                        !a1p1 a2p2
   X Full 1 4 Fix                        !a2p1 a1p2
   Y Full 1 4 Fix                        !a2p1 a2p2
   S Full 1 1 Fix                        !to contain 4: maximum of 4 possible
                                         !genetically different offspring

   G Full 1 nvar Free                    !grand mean
End Matrices;
Matrix S 4
Matrix K 1 1 1 1
Matrix L 1 1 1 1
Matrix M 1 1 1 1
Matrix N 1 1 1 1
Matrix O 1 1 1 1
Matrix X 1 1 1 1
Matrix Y 1 1 1 1
Matrix P
       0
       1 0
       1 1 0
Matrix  Z
       0
       1 0
       1 1 0
Specify K a1s1 a2s1 a1s1 a2s1    !genotype sib1
Specify L a1s2 a2s2 a1s2 a2s2    !genotype sib2
Specify M a1s3 a2s3 a1s3 a2s3    !genotype sib3
Specify N a1p1 a1p2 a1p1 a1p2    !parental alleles
Specify O a1p1 a2p2 a1p1 a2p2    !parental alleles
Specify X a2p1 a1p2 a2p1 a1p2    !parental alleles
Specify Y a2p1 a2p2 a2p1 a2p2    !parental alleles
Specify P    1
             pi12 1
             pi13 pi23 1
Specify Z    1
             z12 1
             z13 z23 1
Specify T    .5                        ! when familial variance is modeled as
             .5 .5                     ! add gen variance

Begin Algebra;
 V = (\part(B,N) + \part(B,O) + \part(B,X) + \part(B,Y)) % S ; !Between effects
 D = (\part(W,N) + \part(W,O) + \part(W,X) + \part(W,Y)) % S ; !for Within effects
End Algebra;

Means G + V + (\part(W,K)-D) | G + V + (\part(W,L)-D) | G + V + (\part(W,M)-D);
Covariance T@(F*F') + P@(Q*Q') + Z@(R*R') + I@(E*E') ;
End
```

```
Constrain sum allelic effects = 0
Constraint ni=1
Begin Matrices;
  A full 1 n = A1
  O zero 1 1
End Matrices;
Begin algebra;
  B = \sum(A) ;
End Algebra;
Constraint O = B ;
end

Constrain sum allelic effects = 0
Constraint ni=1
Begin Matrices;
 C full 1 n = C1
 O zero 1 1
End Matrices;
Begin algebra;
 B = \Sum(C) ;
End Algebra;
Constraint O = B ;
end
```

ACKNOWLEDGMENTS

The financial support of the Universitair Stimulerings Fonds (Grant 96/22), the Human Frontiers Science Program (Grant rg0154/1998-B), and the Netherlands Organization for Scientific Research (NWO, Grant 904-61-090) is greatly appreciated. The Netherlands Organization for Scientific Research (NWO, travel fund R 56-454), and the Simons Stichting (traveling fund) provided travel grants to facilitate collaboration with Dr. Neale, who was supported by MH-01458. The authors wish to thank Drs. Pak Sham, Conor Dolan, and Meike Bartels for their valuable comments on topics related to this paper.

REFERENCES

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

Abecasis, G. R., Cherny, S. S., and Cardon, L. R. (2001). The impact of genotyping error on family-based analysis of quantitative traits. *Eur. J. Hum. Genet.* **9**:130–134.

Abecasis, G. R., Cherny, S. S., Cookson, W. O., and Cardon, L. R. (2002). Merlin: Rapid analysis of dense genetic maps using sparse gene flow trees. *Nat. Genet.* **30**:97–101.

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

Allison, D. B., Thiel, B., St. Jean, P., Elston, R. C., Infante, M. C., and Schork, N. J. (1998). Multiple phenotype modeling in gene-mapping studies of quantitative traits: Power advantages. *Am. J. Hum. Genet.* **63**:1190–1201.

Almasy, L., and Blangero, J. (1998). Multipoint quantitative-trait linkage analysis in general pedigrees. *Am. J. Hum. Genet.* **62**: 1198–1211.

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

Baron, M. (2001). Genetics of schizophrenia and the new millennium: Progress and pitfalls. *Am. J. Hum. Genet.* **68**:299–312.

Beekman, M., Posthuma, D., Heijmans, B. T., Lakenberg, N., Suchiman, H. E. D., Snieder, H., de Knijff, P., Frants, R. R., van Ommen, G. J. B., Kluft, C., Vogler, G. P., Slagboom, P. E., and Boomsma, D. I. (in press). Combined association and linkage analysis applied to the APOE locus. *Genet. Epidemiol.*

Boomsma, D. I., Beem, A. L., van den Berg, M., Dolan, C. V., Koopmans, J. R., Vink, J. M., de Geus, E. J. C., and Slagboom, P. E. (2000). Netherlands twin family study of anxious depression (NETSAD). *Twin Res.* **3**:323–334.

Boomsma, D. I., and Dolan, C. V. (1998). A comparison of power to detect a QTL in sib-pair data using multivariate phenotypes, mean phenotypes, and factor scores. *Behav. Genet.* **28**:329–340.

Bray, N. J., and Owen, M. J. (2001). Searching for schizophrenia genes. *Trends Mol Med.* **7**:169–174.

Cabib, S., Orsini, C., Le Moal, M., and Piazza, P. V. (2000). Abolition and reversal of strain differences in behavioral responses to drugs of abuse after a brief experience. *Science* **289**: 463–465.

Cardon, L. R., and Abecasis, G. R. (2000). Some properties of a variance components model for fine-mapping quantitative trait loci. *Behav. Genet.* **30**:235–243.

Cardon, L. R., and Bell, J. I. (2001). Association study designs for complex diseases. *Nat. Rev. Genet.* **2**:91–99.

Crabbe, J. C., Wahlsten, D., and Dudek, B. C. (1999). Genetics of mouse behavior: Interactions with laboratory environment. *Science* **284**:1670–1672.

Eaves, L. J., Neale, M. C. H., and Maes, H. (1996). Multivariate multipoint linkage analysis of quantitative trait loci. *Behav. Genet.* **26**:519–525.

Falconer, F. S., and Mackay, T. F. C. (1996). *Introduction to quantitative genetics.* Essex, UK: 4th ed. Longman Group, Ltd.

Falk, C. T., and Rubinstein, P. (1987). Haplotype relative risks: An easy reliable way to construct a proper control sample for risk calculations. *Ann. Hum. Genet.* **51**:227–233.

Fulker, D. W., and Cardon, L. R. (1994). A sib-pair approach to interval mapping of quantitative trait loci. *Am. J. Hum. Genet.* **54**:1092–1103.

Fulker, D. W., and Cherny, S. S. (1996). An improved multipoint sib-pair analysis of quantitative traits. *Behav. Genet.* **26**:527–532.

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.

Goldgar, D. E. (1990). Multipoint analysis of human quantitative genetic variation. *Am. J. Hum. Genet.* **47**:957–967.

Goldgar, D. E., and Oniki, R. S. (1992). Comparison of a multipoint identity-by-descent method with parametric multipoint linkage analysis for mapping quantitative traits. *Am. J. Hum. Genet.* **50**:598–606.

Hamer, D., and Sirota, L. (2000). Beware the chopsticks gene. *Mol. Psychiatry* **5**:11–13.

Haseman, J. K., and Elston, R. C. (1972). The investigation of linkage between a quantitative trait and a marker locus. *Behav. Genet.* **2**:3–19.

Ioannidis, J. P., Ntzani, E. E., Trikalinos, T. A., and Contopoulos-Ioannidis, D. G. (2001). Replication validity of genetic association studies. *Nat. Genet.* **29**:306–309.

Kruglyak, L., Daly, M. J., Reeve-Daly, M. P., and Lander, E. S. (1996). Parametric and nonparametric linkage analysis: A unified multipoint approach. *Am. J. Hum. Genet.* **58**:1347–1363.

Lesch, K-P., Bengel, D., Heils, A., Sabol, S. Z., Greenberg, B. D., Petri, S., Benjamin, J., Muller, C. R., Hamer, D. H., and Murphy, D. L. (1996). Association of anxiety-related traits with a polymorphism in the serotonin transporter gene regulatory region. *Science* **29**:274:1527–1531.

Linder, C. C. (2001). The influence of genetic background on spontaneous and genetically engineered mouse models of complex diseases. *Lab. Anim. NY* **30**:34–39.

Liu, J., Corton, C., Dix, D. J., Liu, Y., Waalkes, M. P., and Klaassen, C. D. (2001). Genetic background but not metallothionein phenotype dictates sensitivity to cadmium-induced testicular injury in mice. *Toxicol. Appl. Pharmacol.* **176**:1–9.

Long, A. D., and Langley, C. H. (1999). The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits. *Genome Res.* **9**:720–731.

Lynch, M., and Walsh, B. (1998). *Genetics and analysis of quantitative traits.* Sinauer Associates, Sunderland.

McKenzie, C. A., Abecasis, G. R., Keavney, B., Forrester, T., Ratcliffe, P. J., Julier, C., Connell, J. M. C., Bennett, F., McFarlane-Anderson, N., Lathrop, G. M., and Cardon, L. R. (2001). Trans-ethnic fine mapping of a quantitative trait locus for circulating angiotensin I-converting enzyme (ACE). *Hum. Mol. Genet.* **10**:1077–1084.

Montagutelli, X. (2000). Effect of the genetic background on the phenotype of mouse mutations. *J. Am. Soc. Nephrol.* **11** (Suppl. 16):S101–S105.

Neale, M. C. (1997). *Mx: Statistical modeling.* 3rd ed. Richmond, VA: Virginia Commonwealth University.

Neale, M. C. (2000). The use of Mx for association and linkage analysis. *GeneScreen* **1**:107–111.

Neale, M. C., Cherny, S. S., Sham, P. C., Whitfield, J. B., Heath, A. C., Birley, A. J., and Martin, N. G. (1999). Distinguishing population stratification from genuine allelic effects with Mx: Association of ADH2 with alcohol consumption. *Behav. Genet.* **29**:233–243.

Plomin, R., and Caspi, A. (1999). Behavioral genetics and personality. In Pervin, L. A., and John, O. P. (eds.), *Handbook of personality: Theory and research.* New York: Guilford Press.

Plomin, R., Hill, L., Craig, I., McGuffin, P., Purcell, S., Sham, P. C., Lubinski, D., Thompson, L., Fisher, P. J., Turic, D., and Owen, M. J. (2001). A Genome-wide scan of 1847 DNA markers for allelic associations with general cognitive ability: A five-stage design using DNA pooling. *Behav. Genet.* **31:**497–509.

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

Risch, N. J. (2000). Searching for genetic determinants in the new millennium.*Nature* **405**:847–856.

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

Schork, N. J. (1993). Extended multipoint identity-by-descent analysis of human quantitative traits: Efficiency, power, and modeling considerations. *Am. J. Hum. Genet.* **53**:1306–1319.

Sham, P. C. (1998). *Statistics in human genetics.* London: Arnold Publishers.

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

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *J. R. Stat. Soc. B* **13**:238–241.

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

Sullivan, P. F., Eaves, L. E., Kendler, K. S., and Neale, M. C. (2001). Genetic case-control association studies in neuropsychiatry. *Arch. Gen. Psychiatry* **58**:1015–1024.

Terwilliger, J. D., and Göring, H. H. H. (2000). Gene mapping in the 20th and 21st centuries: Statistical methods, data analysis, and experimental design. *Hum. Biol.* **72**:63–132.

Terwilliger, J., and Ott, J. (1992). A haplotype-based "haplotype relative risk" approach to detecting allelic associations. *Hum. Hered.* **42**:337–346.

Trembath, R. C., Clough, R. L., Rosbotham, J. L., Jones, A. B., Camp, R. D., Frodsham, A., Browne, J., Barber, R., Terwilliger, J., Lathrop, G. M., and Barker, J. N. (1997). Identification of a major susceptibility locus on chromosome 6p and evidence for further disease loci revealed by a two stage genome-wide search in psoriasis. *Hum. Mol. Genet.* **6**:813–820.

Van den Oord, E. J. C. G. (1999). A comparison between different designs and tests to detect QTLs in association studies. *Behav. Genet.* **29**:245–256.

Van den Oord, E. J. C. G. (2000). Framework for identifying quantitative trait loci in association studies using structural equation modeling. *Genet. Epidemiol.* **18**:341–359.

Witte, J. S., Gauderman, W. J., and Thomas, D. C. (1999). Asymptotic bias and efficiency in case-control studies of candidate genes and gene-environment interactions: Basic family designs. *Am. J. Epidemiol.* **149**:693–705.

Yule, G. U. (1900). On the association of attributes in statistics. *Phil. Trans. R. Soc. Lond. A* **194**:257–319.

Zhao, H. (2000). Family based association studies. *Stat. Methods Med. Res.* **9**:563–587.

Zhu, G., Duffy, D. L., Eldridge, A., Grace, M., Mayne, C., O'Gorman, L., Aitken, J. F., Neale, M. C., Hayward, N. K., Green, A. C., and Martin, N. G. (1999). A major quantitative-trait locus for mole density is linked to the familial melanoma gene CDKN2A: A maximum-likelihood combined linkage and association analysis in twins and their sibs. *Am. J. Hum. Genet.* **65**:483–492.