Linkage Analysis: Principles and Methods for the Analysis of Human Quantitative Traits

Manuel A. R. Ferreira

Queensland Institute of Medical Research, Brisbane, Australia

Currently, mapping genes for complex human traits relies on two complementary approaches, linkage and association analyses. Both suffer from several methodological and theoretical limitations, which can considerably increase the type-1 error rate and reduce the power to map human quantitative trait loci (QTL). This review focuses on linkage methods for QTL mapping. It summarizes the most common linkage statistics used, namely Haseman–Elston-based methods, variance components, and statistics that condition on trait values. Methods developed more recently that accommodate the X-chromosome, parental imprinting and allelic association in linkage analysis are also summarized. The type-I error rate and power of these methods are discussed. Finally, rough guidelines are provided to help guide the choice of linkage statistics.

Linkage analysis is one of two complementary strategies currently used for gene-mapping, the other being association analysis. Broadly speaking, linkage is designed to localize a region of the genome where a locus or loci that regulate the expression of a trait may be harbored. Typically, this region of linkage is broad and includes many different genes. By contrast, association has a higher resolution and it is designed to identify the causal gene(s) within the linkage region. Following an overview of the principles behind linkage analysis, this review summarizes the theory of common non-Bayesian statistics that test linkage between genetic loci and any human trait measured on a continuous scale. For convenience, the linkage statistics reviewed here are discussed under four groups. The first two groups of methods, Haseman–Elston and variance components, are the most popular approaches to linkage analysis; these statistics model the phenotypes of relatives conditional on the genotypic information available. In the following section, the third group of statistics reverses this approach, treating the genotypes as the dependent variable and the phenotypes as the independent variable. Finally, the fourth group summarizes additional common statistics which are implemented in popular linkage software packages or that have been developed more recently to incorporate specific effects, such as parental imprinting and allelic association. In the final section, the type-1 error rate and power of these methods are discussed.

1. Principles of Linkage Analysis

Mapping Trait Loci Through Linkage Requires Genetic Markers

Consider that L_t is a trait locus — for example, a sequence of DNA which codes for a protein that influences an observable trait. Assume that this locus exists but there is no information regarding its DNA sequence or location. The aim of linkage analysis is to localize the region where this unknown DNA sequence lies in the human genome. Now let L_{m_i} represent i marker loci — that is, known sequences of DNA which may or may not code for functional proteins — evenly distributed across the genome, covering all 22 autosomes and the X-chromosome. Linkage analysis consists of estimating the genetic distance (or the recombination fraction) between our trait locus and each of these genetic markers. As we scan the entire genome, we will eventually find a group of markers which give low recombination fractions with our trait locus, that is, which are in close proximity to it. The feasibility of such approach in humans was only made possible after the recognition of naturally occurring DNA sequence variation (Botstein et al., 1980).

Parametric Linkage Analysis

In the example described above, L_t and L_m were both genetic loci. The aim of linkage is to estimate the recombination fraction (θ) between L_t and L_m: if the loci are not linked, $\theta = 0.5$ (i.e., meiosis results on average in 50% recombinant gametes and 50% nonrecombinant gametes for L_t and L_m), if they are linked $\theta < 0.5$ (i.e., meiosis results on average in less than 50% recombinant gametes). In practical terms, however, we have direct measured data for each individual in our

Received 14 May, 2004; accepted 31 May, 2004.

Address for correspondence: Manuel A. R. Ferreira, Queensland Institute of Medical Research, P.O. Royal Brisbane Hospital, Brisbane 4029, Australia. E-mail: manuelF@qimr.edu.au

https://doi.org/10.1375/twin.7.5.513 Published online by Cambridge University Press

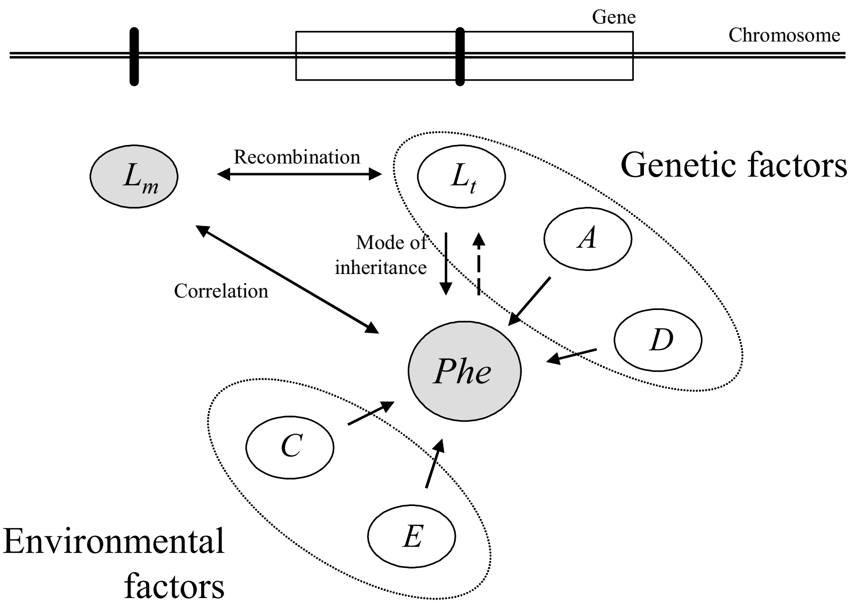


Figure 1

Parametric and nonparametric approaches to linkage analysis.
The parametric approach infers the genotypes of individuals at a trait locus (L_t) based on the observed phenotypes (*Phe*) and on the specification of a specific model of inheritance. Then, the test for linkage consists of estimating the recombination fraction between the marker locus (L_m) and the L_t. In contrast, the nonparametric approach assesses the correlation between the observed genotypic data at L_m and the observed phenotypic data (*Phe*). If L_t truly regulates the expression of *Phe*, then two individuals with the same phenotype are expected to have similar genotypic data at a close marker L_m, or vice versa. The test for linkage thus consists in comparing genotypic and phenotypic similarity between related individuals. *A* represents additive genetic factors, *D* dominance genetic factors, *C* common environmental factors and *E* specific environmental factors. Adapted with permission from Weiss and Terwilliger (2000).

sample for L_m but not for L_t. As a proxy for L_t, we measure an affection status or a quantitative value which we hypothesize L_t is controlling (*Phe*, Figure 1).

If *Phe* reflects closely the underlying genotype L_t, as it is the case with Mendelian traits, then we can determine a person's genotype at L_t by inspection of a family pedigree with phenotypic data. The test for linkage then becomes a question of estimating the recombination fraction between an observed L_m and an inferred L_t (Box 1, A).

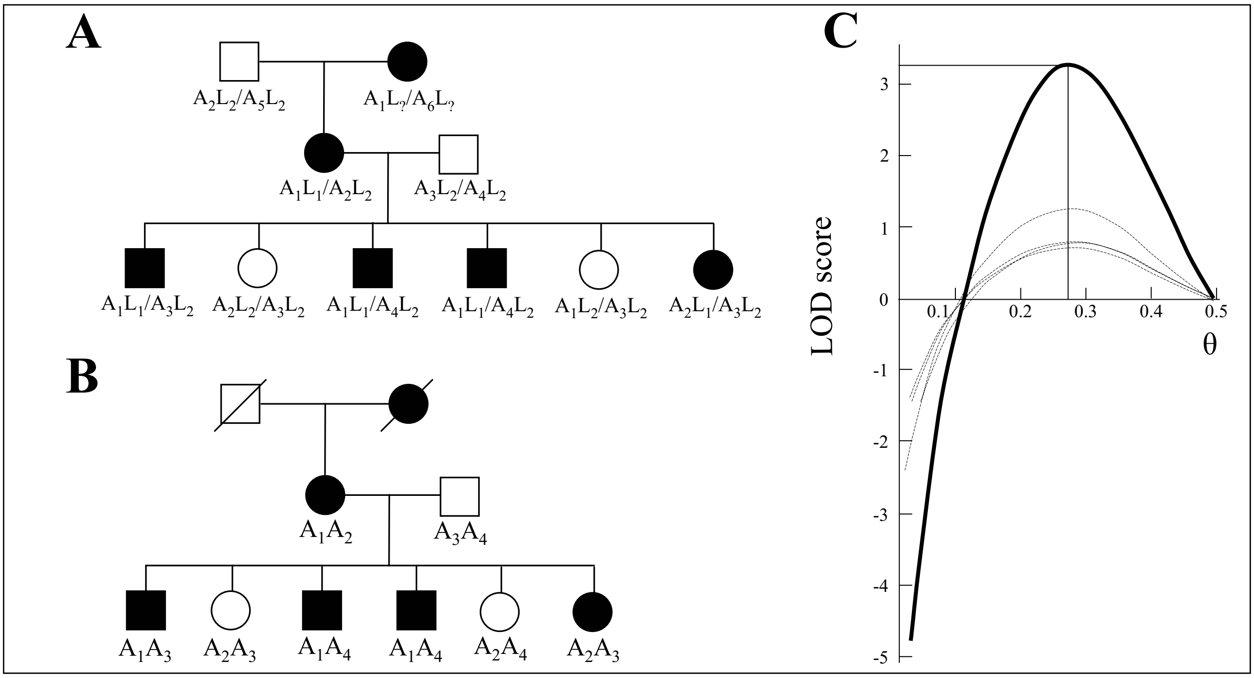
In practical terms, this consists of counting for each individual of a pedigree the number of recombinant and nonrecombinant gametes produced for L_m and L_t. This is intuitive if an individual is both informative for linkage and phase known. An individual is said to be informative for linkage if the individual's genotype is known and it is doubly heterozygous. Additionally, an individual is said to be phase known if it is possible to determine the ancestral origin of each allele, that is, if it is possible to reconstruct the haplotype of that individual.

No linkage information can be extracted from a family which does not include any individual informative for linkage. However, a family can still be used for analysis if informative individuals are present but their phase is unknown (Box 1, B). In this case, evidence for linkage between *A* and *L* is assessed by calculating the overall likelihood of the pedigree under two alternate hypotheses, that the loci are linked (with recombination fraction = θ) or that they are not linked (recombination fraction = 0.5). The ratio of these two likelihoods gives the odds of linkage, that is, how more likely the pedigree is under a model of linkage when compared to a model assuming no linkage. The logarithm of the odds is called the LOD score (Morton, 1955),

$$LOD = \log_{10} \frac{L(X \mid \theta = \hat{\theta})}{L(X \mid \theta = 0.5)} \qquad [1]$$

where *X* represents the pedigree structure and θ the recombination fraction between the marker locus and the trait locus. Being a function of the recombination fraction, LOD scores are calculated for a range of θ values (Box 1, C). The value of θ that gives the highest LOD score is the most likely recombination fraction between both loci. Traditionally, the level of significance required is set at a LOD score of 3. This is the logarithm of the likelihood ratio (1000) that is necessary to convert the odds in favor of linkage from 1:50 (prior probability) to 20:1, the latter corresponding to the conventional 0.05 threshold for statistical significance (Lander & Kruglyak, 1995; Ott, 1991). This is the typical parametric approach used to map Mendelian disease genes, where the relationship between genotype and phenotype is usually simple. The limitation is that it requires the knowledge of the underlying genetic model, namely the

https://doi.org/10.1375/twin.7.5.513 Published online by Cambridge University Press



Box 1

Parametric linkage analysis. Estimating the recombination fraction between a marker locus and a disease locus.
A Pedigree with 3 founders and 7 nonfounders, all genotyped for a genetic marker A and phenotyped for an autosomal dominant disease. The evidence for linkage provided by this pedigree consists in counting the proportion of recombinant and nonrecombinant gametes produced by informative individuals and testing if it is different from 0.5. Of the four gamete-producing individuals of this pedigree, only individual II$_1$ is informative for linkage: she is heterozygous for both the marker locus and the disease locus. Additionally, she is phase-known, since we know that she inherited alleles A$_1$ and L$_1$ from the mother and A$_2$L$_2$ from the father; thus, her haplotype can be reconstructed as A$_1$L$_1$/A$_2$L$_2$. The question is now simply to count the number of recombinant and non-recombinant gametes that individual II$_1$ produced. There are four possible gametes: A$_1$L$_1$, A$_2$L$_2$, A$_1$L$_2$ and A$_2$L$_1$. The first two are nonrecombinants, whereas the latter two are recombinants. By inspection of the generation III, we conclude that individual II$_1$ produced 5 nonrecombinant gametes (three A$_1$L$_1$ and two A$_2$L$_2$) and only 1 recombinant gamete (A$_2$L$_1$). The recombination fraction between A and L is therefore 1 in 6 gametes, i.e. $\theta = 0.17$. **B** The same pedigree as in A, but with no genotypic data for the grandparents. In this case, individual II$_1$ is still informative for linkage but she is now phase-unknown. As a result, it is not possible to identify recombinants in generation III unambiguously and count them: there are either 1 or 5 recombinants in generation III. In this situation, assessing evidence for linkage requires likelihood-based methods. If the loci are truly linked, with recombination fraction θ, the probability of a gamete being recombinant is θ and the probability of it being nonrecombinant is 1-θ. Thus, the likelihood of observing 1 recombinant and 5 nonrecombinant gametes is $\theta^1\cdot(1-\theta)^5$; in the same way, the likelihood of observing 5 recombinants and 1 nonrecombinant is $\theta^5\cdot(1-\theta)^1$. Since both these possibilities are equally likely (individual II$_1$ is either A$_1$L$_1$/A$_2$L$_2$ or A$_1$L$_2$/A$_2$L$_1$), the likelihood of the pedigree given that the loci are truly linked is $1/2\cdot[\theta^1\cdot(1-\theta)^5] + 1/2\cdot[\theta^5\cdot(1-\theta^1)]$. The alternate hypothesis is that the loci are unlinked. If this is the case, the probability that a gamete will be recombinant or non-recombinant is 1/2; therefore, the probability of observing m recombinants and n recombinants is $(1/2)^m\cdot(1/2)^n$. The likelihood of the pedigree given that the loci are unlinked is thus $(1/2)\cdot(1/2)^6 + (1/2)\cdot(1/2)^6$, that is $(1/2)^6$. Following formula [1], the LOD score for this example would be given by $\log_{10}(1/2[\hat{\theta}^1\cdot(1-\hat{\theta})^5] + 1/2[\hat{\theta}^5\cdot(1-\hat{\theta}^1)] - \log_{10}(1/2)^6$. The LOD score would then be calculated for a range of θ values, and a LOD curve for the family constructed. An identical approach would be applied to other families. **C** Since the overall likelihood of a given set of pedigrees is the product of the likelihoods of each individual family, the LOD curve of individual families (thin lines), being logarithms, can be added up across families to produce an overall LOD score curve (thick line). For example, a LOD score of 3 for a $\theta = 0.28$ indicates that overall our pedigrees are 1000 ($3 = \log_{10}1000$) times more likely to be observed if we assume that both loci are linked with a recombination fraction of 0.28 then if we assume that they are not linked ($\theta = 0.5$).

mode of genetic inheritance, gene frequencies and penetrance of each genotype.

Nonparametric Linkage Analysis

Parametric linkage analysis requires the specification of a precise genetic model. To some extent, this limits this type of analysis to discrete traits with Mendelian inheritance. However, many discrete traits (for example, diabetes, atopy) and certainly most continuous traits (e.g., height, eosinophil levels) may involve the action of multiple genes: they are said to have a complex mode of inheritance. In this case, specifying a genetic model becomes less tractable and linkage analysis must revert to model-free methods.

There are two types of model-free approaches to linkage analysis. The first type of approach, known as parametric model-free, retains the parametric framework in the sense that it specifies a genetic model, though this is only an approximation to reality. Since the true disease model is typically unknown, the alternatives are either to assume a particular genetic model even though this may be the wrong model (e.g., Clerget-Darpoux et al., 1986; Tiwari et al., 1980) or to conduct the analysis under multiple models, so that

https://doi.org/10.1375/twin.7.5.513 Published online by Cambridge University Press

one of these is likely to be close to the true model (e.g., Clerget-Darpoux et al., 1986; Elston, 1989; Greenberg, 1989; Risch, 1984). The use of multiple models, however, raises numerous problems (Hodge & Elston, 1994; Sham, 1998). The other model-free approach to linkage analysis of complex traits is known as nonparametric linkage. This approach abandons the conventional LOD score parametric method, in the sense that it does not formally test if the recombination fraction θ between a marker and a trait locus is significantly different from 0.5. Rather, the rationale of this group of methods is the following: if a sequence of DNA truly regulates the expression of a trait, two individuals with the same DNA sequence are expected to have similar trait values, or vice versa. If, by contrast, the locus is not involved in the regulation of the phenotype, the genotypic and phenotypic similarity between two related individuals will be independent. The focus of the remaining sections of this review will be on this second approach to nonparametric linkage analysis.

Nonparametric linkage methods avoid the need to specify an inheritance model for the trait but they require the estimation of both the phenotypic and genotypic similarities between two individuals. Phenotypic similarity can be expressed in different ways, namely by calculating squared differences, squared sums or normalized products, or by estimating the trait covariance between two individuals. Genotypic similarity at a given locus can be expressed in two different ways: the number of alleles that both individuals share identical by state (IBS) or identical by descent (IBD). IBS alleles look the same, and may have the same DNA sequence, but they are not necessarily derived from a known common ancestor. Alleles IBD are copies of the same ancestral allele. For rare alleles, two independent origins are unlikely, so IBS generally implies IBD. For common alleles this may not be true. Thus, though both IBS and IBD data can be used for linkage analysis, IBD is the more powerful and generally preferable. Since IBD information is an essential component of most nonparametric linkage methods, details on its calculation are presented in the next section.

2. Calculation of IBD

Several methods have been proposed for estimating the number of alleles shared IBD between two related individuals at a marker locus. The most general are the Elston–Stewart algorithm (Elston & Stewart, 1971) and the Lander–Green algorithm (Lander & Green, 1987). The Lander–Green algorithm handles smaller pedigrees but a large number of loci; in this way, it is particularly appropriate for the analysis of pedigrees collected by most linkage studies. In addition, the most popular linkage software packages (e.g., Allegro, Genehunter, Merlin) have implemented this algorithm, albeit with some modifications to improve computational issues. For these reasons, this section describes IBD estimation using the Lander–Green algorithm.

Singlepoint IBD Estimation

Consider a pedigree with f founders (individuals with no ancestors in the pedigree) and n nonfounders (individuals with at least one parent in the pedigree). For simplicity, assume that $n = f = 2$, that is, the pedigree consists of two siblings and both parents. In the absence of any genotypic information, there are 2^{2n} equally likely genotypic conformations for the sib-pair, according to Mendel's first law of segregation (Figure 2, A and B). Each conformation is specified by a unique inheritance vector $v(x) = (p_1,m_1;p_2,m_2;\ldots;p_n,m_n)$, that is, a binary vector whose coordinates describe the outcome of the two meioses which produced each nonfounder of the pedigree for a particular locus (Lander & Green, 1987). Specifically, $p_i = 0$ or 1, according to whether the grand-paternal or grand-maternal allele was transmitted in the paternal meiosis giving rise to the i^{th} nonfounder; m_i carries the same information for the corresponding maternal meiosis.

Since each inheritance vector clearly specifies which of the distinct $2f$ founder alleles was inherited by each nonfounder, it describes a unique pattern of gene flow through the pedigree. As mentioned above, in the absence of any genotypic information, there are 2^{2n} equally likely gene flow patterns (Figure 2, B). However, as genotypic information is added to the pedigree, the probability distribution is concentrated on certain inheritance vectors: genotypic data renders some vectors inconsistent, others less and others more likely to be observed (Figure 2, C–E). Indeed, one can apply Bayes' theorem to compute the probability of each inheritance vector given the genotypic data observed (see Appendix A in Kruglyak et al., 1996).

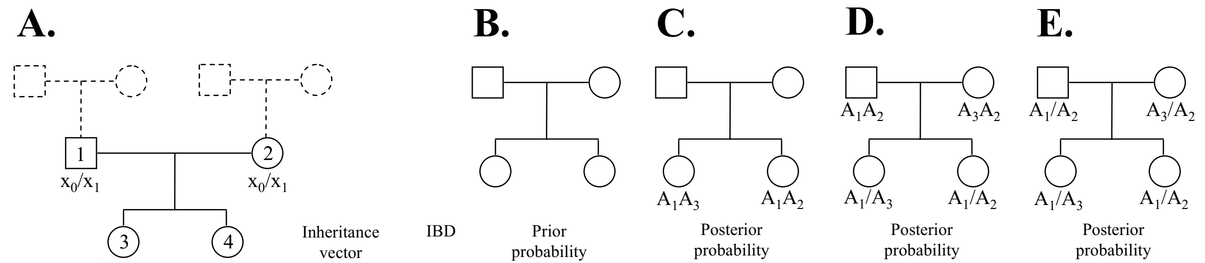
If more than one inheritance vector is found to be compatible with the genotypic data at a single locus, the overall likelihood of the pedigree has to be formulated in a way that accommodates this uncertainty. The likelihood of the pedigree is thus calculated as the sum of the probabilities of the 2^{2n} inheritance vectors, and can be written in matrix form as

$$P(x) = 1^T \cdot Q \cdot 1 \qquad [2]$$

where x represents the observed genotypic data at the locus, 1 is a column vector with 2^{2n} elements equal to 1 and Q is a 2^{2n}-by-2^{2n} diagonal matrix with the 2^{2n} probabilities, one for each inheritance vector. This is the general formula for the likelihood calculation of pedigree data at a single locus. How can this singlepoint approach to pedigree likelihood be used to calculate IBD between nonfounders?

Consider the special case in which the inheritance vector is known with certainty (Figure 2, E). The inheritance vector fully determines which of the $2f$ founder alleles was inherited by each nonfounder and, thus, completely specifies IBD sharing at a single locus between each nonfounder. In this example, the

https://doi.org/10.1375/twin.7.5.513 Published online by Cambridge University Press



3	4	Inheritance vector	IBD	Prior probability	Posterior probability	Posterior probability	Posterior probability
x_0/x_0	x_0/x_0	0000	2	1/16	0	0	0
x_0/x_0	x_0/x_1	0001	1	1/16	1/6	1/4	1
x_0/x_0	x_1/x_0	0010	1	1/16	0	0	0
x_0/x_0	x_1/x_1	0011	0	1/16	1/12	0	0
x_0/x_1	x_0/x_0	0100	1	1/16	1/6	1/4	0
x_0/x_1	x_0/x_1	0101	2	1/16	0	0	0
x_0/x_1	x_1/x_0	0110	0	1/16	1/12	0	0
x_0/x_1	x_1/x_1	0111	1	1/16	0	0	0
x_1/x_0	x_0/x_0	1000	1	1/16	0	0	0
x_1/x_0	x_0/x_1	1001	0	1/16	1/12	0	0
x_1/x_0	x_1/x_0	1010	2	1/16	0	0	0
x_1/x_0	x_1/x_1	1011	1	1/16	1/6	1/4	0
x_1/x_1	x_0/x_0	1100	0	1/16	1/12	0	0
x_1/x_1	x_0/x_1	1101	1	1/16	0	0	0
x_1/x_1	x_1/x_0	1110	1	1/16	1/6	1/4	0
x_1/x_1	x_1/x_1	1111	2	1/16	0	0	0
			P (IBD=0)	1/4	1/3	0	0
			P (IBD=1)	1/2	2/3	1	1
			P (IBD=2)	1/4	0	0	0

Figure 2

Singlepoint IBD estimation using inheritance vectors in pedigrees with variable genotypic data available.

A Possible genotypic combinations for the sib-pair (3–4), assuming that the parents (1–2) are phase-known, where x_0 indicates that the parent inherited the allele from the grandfather and x_1 indicates that the allele was inherited from the grandmother. The inheritance vector for the sib-pair fully determines which of the four paternal alleles was inherited by each sib. For example, the vector 0100 specifies that the first nonfounder inherited one allele from the father's father (0) and the other allele from the mother's mother (1), and that the second nonfounder inherited one allele from the father's father (0) and the other allele from the mother's father (0). Thus, an inheritance vector has 2*n* digits (i.e., meiosis) and each digit can only assume two values: 0 if the allele was inherited from the parent's father and 1 if it was inherited from the parent's mother; therefore, there are 2^{2n} possible inheritance vectors per pedigree. Note that each inheritance vector fully specifies how many alleles IBD are shared by both sibs. **B** Prior to considering any genotypic data, all inheritance vectors are equally likely, according to Mendel's second law of segregation. C–D However, as genotypic data is added to the pedigree, some vectors become incompatible, others more likely and others less likely to be observed. Note that pedigree D contains no information about founder phase; in this case, inheritance vectors that differ only by phase changes in the founders are completely equivalent and must therefore have equal probabilities (e.g., 0001 and 1110). As a consequence, one can reduce the inheritance vector space from 2^{2n} to 2^{2n-f}. E In the extreme case where the phase of both parents is known, the inheritance vector can be determined unambiguously. For all pedigrees, the probability that two nonfounders share *i* alleles IBD at a given locus is simply obtained by adding the probabilities of the appropriate inheritance vectors.

siblings clearly have 1 allele IBD at locus *A*. At the opposite end of the scale is, of course, the case of a pedigree with no genotypic information (Figure 2, B); there are 16 equally likely inheritance vectors that result in three possible IBD states: 0, 1 and 2, with probabilities 1/4, 1/2 and 1/4, respectively. Extending this to the general case, the probability that two nonfounders share *k* alleles IBD at a given locus is simply obtained by adding the probabilities of the appropriate inheritance vectors (Kruglyak & Lander, 1995). More formally, if *V* denotes all possible 2^{2n} inheritance vectors that $v(x)$ can assume, then

$$P(IBD = k) = \sum_{w \in V} P[IBD = k \mid v(x) = w] \cdot P[v(x) = w] \quad [3]$$

where $P[IBD = k \mid v(x) = w]$ takes the value of 1 or 0 if the vector w is compatible or incompatible with IBD = k, respectively, and $P[v(x) = w]$ is the posterior probability of observing the inheritance vector w. Finally, once the three probabilities of sharing 0, 1 or 2 alleles IBD have been calculated, conventionally denoted as π_0, π_1, π_2, the proportion of alleles shared IBD at the locus is estimated by $\hat{\pi} = \pi_1/2+\pi_2$.

Multipoint IBD Estimation

Formula [2] indicates how to calculate the likelihood of a given pedigree given genotypic data at a single locus. Frequently, however, we have collected data at several ordered loci for each pedigree. Though it is possible to calculate singlepoint likelihoods (and IBDs) at all marker loci individually, this approach does not extract the full information from a data set. For example, if a family is uninformative or has no genotypic data for a marker locus, the singlepoint IBD estimation for a sib-pair at that locus will correspond to the prior probabilities π_0 = 1/4, π_1 = 1/2, π_2 = 1/4, and,

https://doi.org/10.1375/twin.7.5.513 Published online by Cambridge University Press

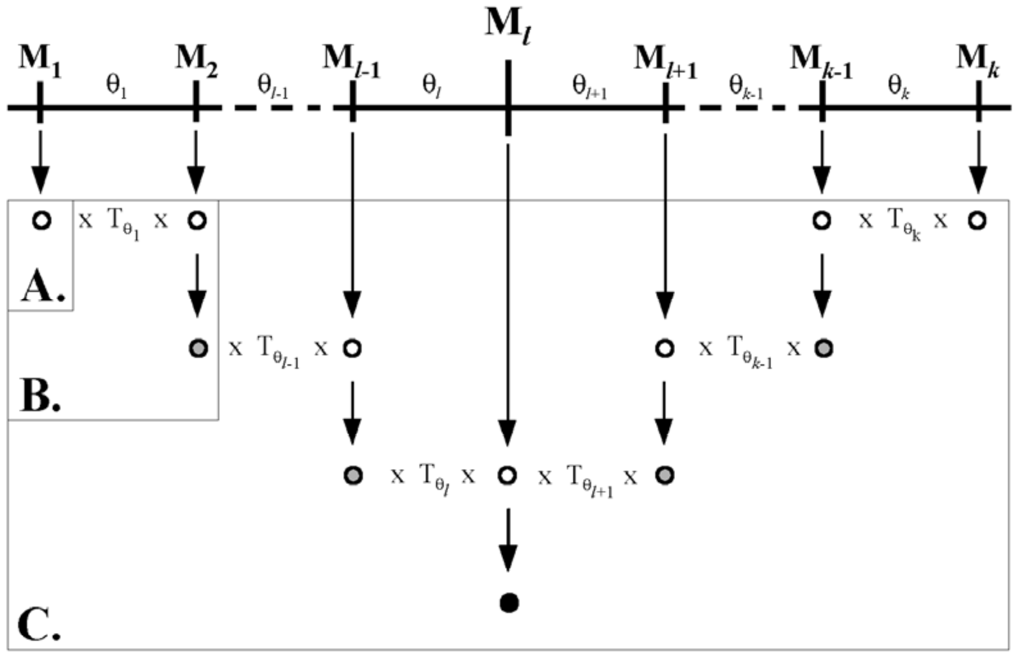


Figure 3

Pictorial representation of multipoint calculation of pedigree likelihood.
The aim is to calculate the probability of observing each of the 2^{2n} possible inheritance vectors $v(x)$ at an arbitrary marker location l. This can be done in three ways: singlepoint, unilateral (left or right) multipoint and bilateral multipoint. **A** Singlepoint likelihood calculations. Only the genotypic data at the locus is used to calculate the probability distribution of the inheritance vectors (open circles). **B** Unilateral multipoint likelihood calculations (grey circles). The probability distribution of $v(x)$ at a locus l is a function of both the genotypic data observed at all loci on its left, and the genotypic data at the locus itself. The probability distribution of $v(x)$ at marker l conditional on the genotypes of all preceding l–1 loci is obtained by multiplying the probability distribution of $v(x)$ at l–1 by a 2^{2n}-by-2^{2n} transition matrix T_{θ_l} with elements $\theta_l^{\,r} \cdot (1-\theta_l)^{2n-r}$, where r specifies the number of differences between inheritance vectors at locations l–1 and l. Note that the probability distribution at l–1 is again expressed as a unilateral multipoint likelihood, so that it includes all genotypic information from locus 1 to locus l–1. The same approach can be used to calculate right unilateral likelihoods. **C** Bilateral multipoint likelihood calculation (black circle). The probability distribution of $v(x)$ at locus l is now a function of the genotypic data observed at the l–1 preceding loci, the genotypic data at l, and the genotypic data at the following k–l loci. Thus, the genotypic information of all loci is used to calculate the probability distribution of $v(x)$ at location l.

thus, that family will have no contribution toward the overall linkage signal. To avoid this limitation, the Lander–Green algorithm (but also the Elston–Stewart algorithm) has been designed to optionally use all marker data available to estimate the IBD distribution at an arbitrary marker location. Other multipoint algorithms have been developed for large complex pedigrees, namely using Markov–Chain Monte-Carlo methods (Heath, 1997) or average sharing methods (Almasy & Blangero, 1998; Fulker et al., 1995). In contrast to the Lander–Green and the Elston–Stewart algorithms, the two latter approaches calculate approximate and not exact IBD distributions. Though multipoint exact calculations are preferable, they may be computationally prohibitive in large pedigrees.

The essential concept of multipoint IBD estimation is that the inheritance pattern at a location l can be inferred not only using the genotyping data at locus l but, complementary, by inspection of the inheritance patterns of adjacent loci. There are three sources of information regarding the likelihood of the pedigree at a given marker l: using genotypic data from marker l-1, from marker l and from marker l+1. This can be specified as a Markov chain across all available genotypic information which is then used to compute a bilateral multipoint likelihood of pedigree data at any arbitrary location l (Figure 3). The overall likelihood of a pedigree given the k loci data is

$$P(x_1, x_2, ..., x_k) = 1^T Q_1 \cdot T_{\theta 1} \cdot Q_2 \cdot T_{\theta 2} ... T_{\theta k} \cdot Q_k \cdot 1 \qquad [4]$$

where Q_1 ... Q_k are diagonal matrices with the probabilities for each inheritance vector calculated using the genotypic data at locations 1 to k, and T_{θ_1} ... T_{θ_k} are the transition matrices between consecutive markers which allow to calculate the probabilities of each possible inheritance vector at a given location k using the genotypic data from all preceding loci. Because matrix multiplication is associative, this formula can be computed from the left or from the right. Thus, if we want to calculate the bilateral multipoint likelihood at a location l given all the genotypic data at the k loci, we would have to factorize this probability has a left conditional likelihood, a single marker likelihood and a right conditional likelihood (Figure 3).

As with the singlepoint approach, the multipoint calculation of IBD between any pair of relatives is

https://doi.org/10.1375/twin.7.5.513 Published online by Cambridge University Press

straightforward once the probabilities of the 2^{2n} inheritance vectors have been determined: it simply consists in summing the probabilities of the appropriate vectors. In the same way, the proportion of alleles shared IBD at the locus is estimated by $\hat{\pi} = \pi_1/2+\pi_2$

Sections 1 and 2 have introduced some basic concepts of linkage analysis, including the rationale of parametric and nonparametric approaches, and IBD estimation. The subsequent sections will present the statistical fundamentals of different methodological frameworks which incorporate genotypic and phenotypic information from relatives to test for linkage.

3. Haseman–Elston Regression and Appropriate Extensions

Original Haseman–Elston Regression

This method was suggested for the analysis of sib-pairs by Haseman and Elston (1972). Let X_1 and X_2 represent the quantitative trait values of a sib-pair, and $\hat{\pi}$ the respective proportion of alleles identical-by-descent (IBD) at a locus L. The central idea of this method is the theoretical decomposition of the expected squared trait difference given $\hat{\pi}$

$$E\left[(X_1 - X_2)^2 \mid \hat{\pi}\right] = E\left[\left(X_1^2 + X_2^2 - 2\cdot X_1 \cdot X_2\right)\mid \hat{\pi}\right] \\ = Var(X_1) + Var(X_2) - 2Cov(X_1X_2 \mid \hat{\pi}) \qquad [5]$$

Assuming that the variance of the trait can be factorized into genetic (additive and dominance effects — V_A and V_D, both at the quantitative trait loci (QTL) and from residual contributions) and environmental components (shared and nonshared effects — V_C and V_E), the variances of X_1 and X_2 are given by

$$Var(X_1) = Var(X_2) = V_{A_{QTL}} + V_{D_{QTL}} + V_{A_{residual}} \\ + V_{D_{residual}} + V_C + V_E \qquad [6]$$

The variances are, of course, independent of the sib-pair $\hat{\pi}$. The overall covariance between X_1 and X_2, which can be derived from path diagrams (see Figure 4 for an example), is given by

$$Cov(X_1, X_2 \mid \hat{\pi}) = \hat{\pi}\cdot V_{A_{QTL}} + \pi_2 \cdot V_{D_{QTL}} + 2\cdot\Phi\cdot V_{A_{residual}} \\ + \Delta\cdot V_{D_{residual}} + V_C \qquad [7]$$

where $2\cdot\Phi$ is twice the kinship coefficient (i.e., twice the probability that two alleles drawn at random, one from each relative, will be IBD; also equivalent to the expected proportion of alleles IBD), and Δ represents the expected probability of sharing two alleles IBD; that is, both are the theoretical values without considering the genotypic data. For sib-pairs, $2\cdot\Phi = 1/2$ and $\Delta = 1/4$. Thus, from [5] it follows that

$$E\left[(X_1 - X_2)^2 \mid \hat{\pi}\right] = -2\cdot V_{A_{QTL}}\cdot\hat{\pi} - 2\cdot V_{D_{QTL}}\cdot\pi_2 + 2\cdot V_{A_{QTL}} \\ + 2\cdot V_{D_{QTL}} + V_{A_{residual}} + 3/2\cdot V_{D_{residual}} + 2\cdot V_E \qquad [8]$$

And, assuming an additive model where $V_{D_{QTL}}$ and $V_{D_{residual}}$ are both 0, the expression becomes

$$E\left[(X_1 - X_2)^2 \mid \hat{\pi}\right] = -2\cdot V_{A_{QTL}}\cdot\hat{\pi} + 2\cdot V_{A_{QTL}} \\ + V_{A_{residual}} + 2\cdot V_E \qquad [9]$$

Thus, the pair squared trait difference can be regressed on $\hat{\pi}$ with the slope being an estimate of $-2\cdot V_{A_{QTL}}$ This assumes that $\hat{\pi}$ is estimated at the true trait locus or at a locus so close to it that has the same IBD distribution (that is, with $\theta = 0$). If, however, IBD is being estimated at a marker not tightly linked to the trait locus, the linear relationship between the squared trait difference and $\hat{\pi}$ is now only an imperfect estimate of $V_{A_{QTL}}$ Indeed, the closer the marker is to the true trait locus the better the regression slope should approximate $V_{A_{QTL}}$ This can be corrected in the linear model by multiplying the regression coefficient by $(1-2\cdot\theta)^2$, this term being the correlation between $\hat{\pi}$ at the marker locus and $\hat{\pi}$ at the trait locus. Therefore, if we consider the pair squared trait difference as the dependent variable and $\hat{\pi}$ at any arbitrary location L as the independent variable for a given number of pedigrees, the regression coefficient (β) is an estimate of $V_{A_{QTL}}(1-2\cdot\theta)^2$. A significant negative regression coefficient implies that there is either a relatively large genetic effect at a moderate distance from the marker or that there is a smaller genetic effect close to the marker. Thus, the test for linkage is a one-sided t test of the null hypothesis H_0: $\beta = 0$

Extensions to the Original Haseman–Elston Regression

Wright (1997) reexamined the original Haseman–Elston approach (HE–SD) and showed that the pair squared trait difference and the mean-corrected squared trait sum are statistically independent and, hence, can provide complementary information for linkage analysis. Following this observation, Drigalenko (1998) suggested that a more accurate estimate of β could be obtained by simply averaging the estimates from two regressions, one using the squared differences and the other the squared sums. This approach is analogous to the cross-product model (HE–CP) suggested by Elston et al. (2000) that uses $Y = [(X_1-\mu)\cdot(X_2-\mu)]$ as the dependent variable. However, Forrest (2001) pointed out that weighting the two slope estimates equally is not optimal, since the slope estimates from both regressions have different variances when the sibling correlation is positive. To correct this problem, new weighted methods (HE–W) have been proposed that estimate β as the weighted sum of both regression slopes (Forrest, 2001; Visscher & Hopper, 2001; Xu et al., 2000). The weights used by Xu et al. (2000) and Forrest (2001) are the slope variance estimates obtained directly from the regression models, whereas Visscher and Hopper (2001) used the inverse of the respective empirical variance estimates. Finally, Sham and Purcell (2001) simplified the method by Xu et al. (2000) by expressing the two slope variances as a function of the sibling correlation (HE–COM). Irrespective of the nature of the extension, the test for linkage with all the Haseman–Elston-based approaches is a one-sided t test of the regression slope.

https://doi.org/10.1375/twin.7.5.513 Published online by Cambridge University Press

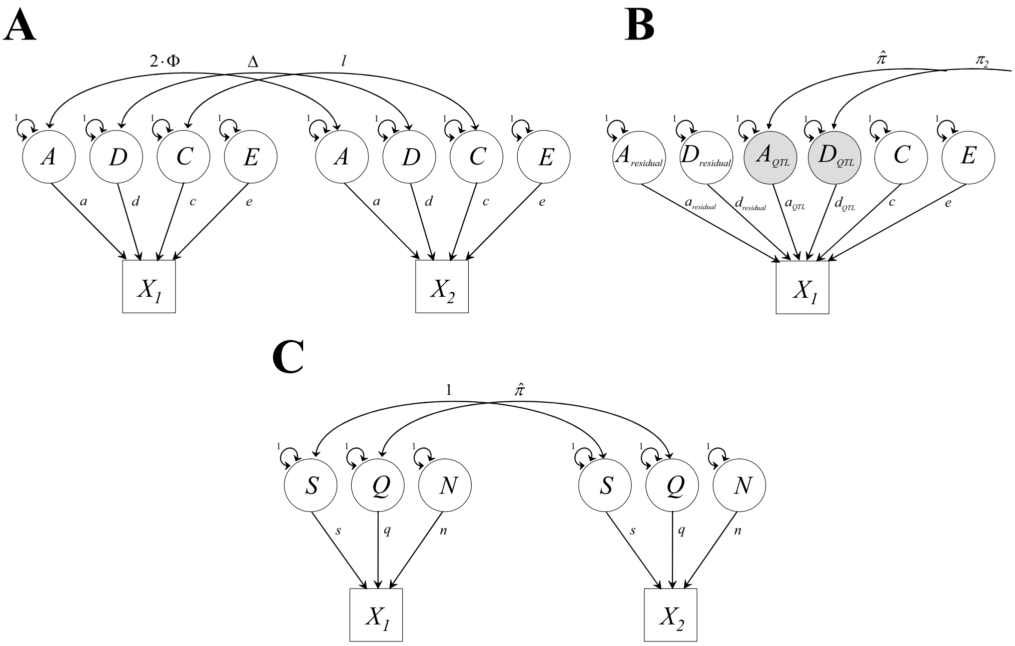


Figure 4

Path diagrams for variance components modelling with the 'pi-hat approach'

A Decomposition of the variance and covariance of a trait X in a sib-pair into four latent factors, additive genetic effects *(A)*, dominance genetic effects (*D*), common environmental effects and specific environmental effects (*E*). Each of these is constrained to have a variance equal to 1 and to load on the phenotype X with the coefficients *a, d, c* and *e*, respectively. The additive genetic effects are correlated between individuals by twice the kinship coefficient (2·Φ), the dominance genetic effects correlated by Δ (which represents the expected probability of sharing two alleles IBD), and the common environment components by *l*, the household indicator. For sib-pairs, 2·Φ = 1/2, Δ = 1/4 and *l* = 1. **B** The model in **A** modified to include the effects of QTL, with the corresponding additive (A_{QTL}) and dominance (D_{QTL}) contributions. Only one sibling shown; $A_{residual}$, $D_{residual}$ and *C* are correlated between siblings as specified for *A*, *D* and *C* in **A**; $\hat{\pi}$ represents the proportion of alleles IBD at the locus and π_2 the probability of sharing two alleles IBD. **C** The model represented in **B** but simplified so that it is identified when applied to sib-pair data. *S* represents shared latent factors, *Q* the QTL latent factor and *N* the nonshared latent factor. See text for details.

4. Variance Components (VC) Maximum Likelihood

The 'Pi-Hat' Approach to Linkage Analysis

This is an intuitive approach to QTL linkage based on an extension to the traditional *ADCE* model (Neale & Maes, 1999). For a more detailed review of the 'pi-hat' approach and some of its extensions see Posthuma et al. (2003). Take a sample of sib-pairs which have been phenotyped for trait *X* and genotyped at markers evenly spaced across the genome; for each of these markers, the three IBD probabilities have been estimated and $\hat{\pi}$ calculated. In addition to the traditional *ADCE* components of variance (Figure 4A), we now want to model the effect on the phenotype *X* of an additional latent factor, an individual genomic locus *Q*. The additive effect of this locus is correlated between siblings by $\hat{\pi}$, the proportion of alleles shared IBD, whereas the dominance effect is correlated by π_2 the probability of sharing two alleles IBD (Figure 4B). Following standard tracing rules of path analysis, the variance for X_1 or X_2 can be expressed as

$$Var(X_1) = Var(X_2) = a^2_{residual} + d^2_{residual} + c^2 + e^2 + a^2_{QTL} + d^2_{QTL} = V_{A_{residual}} + V_{D_{residual}} + V_C + V_E + V_{A_{QTL}} + V_{D_{QTL}} \quad [14]$$

and the covariance between X_1 and X_2 is given by

$$Cov(X_1, X_2) = a_{residual} \cdot 2 \cdot \Phi \cdot a_{residual} + d_{residual} \cdot \Delta \cdot d_{residual} + c \cdot l \cdot c + a_{QTL} \cdot \hat{\pi} \cdot a_{QTL} + d_{QTL} \cdot \pi_2 \cdot d_{QTL} = 2 \cdot \Phi \cdot V_{A_{residual}} + \Delta \cdot V_{D_{residual}} + l \cdot V_C + \hat{\pi} \cdot V_{A_{QTL}} + \pi_2 \cdot V_{D_{QTL}} \quad [15]$$

where *l* represents the household indicator (1 if individuals share the same household, 0 if they do not; if we are modeling the sib-pair common environment, *l* = 1), with the remaining notations being equivalent to those used in formula [7]. The likelihood of observing the phenotypic data of the sib-pair in the i^{th} pedigree conditional on the genotypic data is given by

$$-2 \cdot \ln[L(x_i \mid \hat{\pi})] = -n \cdot \ln(2 \cdot \pi) + \ln|\Sigma_i| + (x_i - \mu)' \cdot \Sigma_i^{-1} \cdot (x_i - \mu) \quad [16]$$

where x_i is the vector of observed phenotypes for the i^{th} pedigree, n = 2 (number of observed phenotypes), μ is the vector that models the means, and Σ_i is the expected variance–covariance matrix specified by [14] and [15]. Note that in this formula π represents the conventional ratio of a circle's perimeter to its diameter (~ 3.14). The expected covariance matrix includes six free parameters. Thus, if fitted to sib-pair data (which supply only two independent statistics, one variance and one covariance), it is under identified. In this case,

https://doi.org/10.1375/twin.7.5.513 Published online by Cambridge University Press

the problem can be solved by grouping $1/2 \cdot V_{A_{residual}} + V_C$ under the shared latent factor S, and $1/2 \cdot V_{A_{residual}} + V_E$ under the nonshared latent factor N (Sham, 1998) (Figure 4C). Assuming an additive model (i.e., $V_{D_{residual}}$ $V_{D_{QTL}} = 0$), the variance–covariance matrix for siblings j and k of the ith pedigree is now reduced to

$$\Sigma_{jk} = \begin{cases} V_N + V_S + V_{QTL} & for \quad j = k \\ V_S + \hat{\pi} \cdot V_{QTL} & for \quad j \neq k \end{cases} \qquad [17]$$

Though this model involves three parameters and predicts only two statistics when applied to the sib-pair design, it is still identified because the covariance equation is a simple linear regression on $\hat{\pi}$. Similarly, if the QTL dominance component ($V_{D_{QTL}}$) was also included, the model would still be identified, since this parameter is a linear regression on π_2. However, the improvement obtained by including dominance or other gene-by-gene interaction parameters is still a controversial issue in linkage analysis. Due to power issues, modeling the QTL dominance component seems to be appropriate only when the marker locus is very close to or is suspected to be the true QTL itself (Almasy & Blangero 1998; Sham et al., 2000a). If data are collected under the classical twin design, model [17] predicts an additional independent statistic (the MZ covariance) and, hence, $V_{A_{residual}}$ and V_C could be estimated independently.

To test for linkage between a given marker and the phenotype X, a saturated model H_0 and a nested submodel H_1 are fitted separately to the same dataset. The submodel H_1 differs from the saturated model H_0 in that the QTL factor has been dropped, that is, the effect of the locus on the phenotype has been fixed to zero. Therefore, the statistic $2 \cdot [\ln(L_{H_0}) – \ln(L_{H_0})]$ provides a relative measure of fit of H_1: a significant chi-square indicates that the submodel H_1 fits the data significantly worse than H_0. Asymptotically, under the null hypothesis of no linked QTL, this test statistic is 0 with a probability of 0.5 and it follows a χ^2 distribution with 1-*df* with a probability of 0.5 (Hopper & Matthews, 1982). A significant drop in fit when dropping $V_{A_{QTL}}$ suggests that the similarity of two individuals at the marker locus significantly influences their phenotypic similarity; in other words, the marker locus is the trait locus or it is linked to it.

The 'Mixture Distribution' Approach

The 'pi-hat' approach calculates $\hat{\pi}$ for the sib-pair at each marker and uses this value in the variance–covariance matrix when computing the likelihood function [16] for each observation. This matrix, specifically the covariance element $V_S + \hat{\pi} \cdot V_{QTL}$, determines the shape of the bivariate probability density function (PDF) which returns the likelihood of each observation. With fully informative marker data, the PDF can assume three different shapes according to the three possible $\hat{\pi}$ values: 0, 0.5 and 1. For example, if $\hat{\pi} = 0$, then the appropriate PDF specifies that any combination of trait values for a sib-pair is equally likely to occur; in the other extreme, however, if $\hat{\pi} = 1$, then the PDF determines that sib-pairs which are concordant for a trait are more likely to be observed than pairs which are discordant. The limitation of the 'pi-hat' approach arises in the presence of incomplete marker information, that is when IBD cannot be determined with certainty and $\hat{\pi}$ can assume values other than 0, 0.5 and 1. In this case we no longer have three possible PDF distributions but many which are not biologically meaningful. One alternative to this approach is the finite 'mixture distribution' method (Eaves et al., 1996). Take the same example as above, where we have phenotypic data for sib-pairs and genotypic data collected at several markers evenly spaced across the genome. In this case, however, only the three IBD probabilities π_0, π_1 and π_2 are estimated; $\hat{\pi}$ is not calculated. For a given observed vector, three individual likelihoods are calculated, respectively assuming that $\hat{\pi}$ is 0 (i.e., the covariance element is simply V_S), $\hat{\pi}$ is 0.5 (covariance $V_S + 0.5 \cdot V_{QTL}$) and $\hat{\pi}$ is 1 (covariance $V_S + V_{QTL}$). Thus, this method forces the likelihood to be read in the three meaningful PDFs; the overall likelihood of each vector then simply consists of the weighted sum of the three likelihoods, where the weights are respectively π_0, π_1 and π_2. More formally, the overall likelihood of a vector of observed trait values $x_i = [x_{i1}, x_{i2}, \ldots, x_{in}]$ for the ith pedigree containing n members, conditional on the IBD information is

$$-2 \cdot \ln[L(x_i \mid IBD)] = \sum_{l=1}^{m} [-2 \cdot \ln(w_l \cdot L_{il})] \qquad [18]$$

where w_l is the weight for the mth model, and L_{il} the likelihood of the trait vector x_i under the mth model. As with the previous approach, the test for linkage involves fitting a saturated model which includes the effect of the marker locus (H_0) and a submodel where this component has been dropped (H_1). Then, the statistic $2 \cdot [\ln(L_{H_0}) – \ln(L_{H_0})]$ provides a test for the significance of the QTL contribution to the phenotypic correlation.

Note that with complete IBD information, the 'pi-hat' and the 'mixture distribution' approaches are equivalent, giving exactly the same results. For example, the likelihood of a given vector of traits $[x_1\ x_2]$ for a sib-pair which is IBD 2 will be obtained with the 'pi-hat' approach from a PDF with a distribution specified by a $\hat{\pi}$ of 1 (L_1). Similarly, if the 'mixture distribution' is used, the overall likelihood of the same trait vector is $\pi_0 \cdot L_0 + \pi_1 \cdot L_{0.5} + \pi_2 \cdot L_1 = L_1$ since $\pi_0 = \pi_1 = 0$ and $\pi_2 = 1$.

5. Statistics that Model IBD Conditional on Trait Values

All methods discussed so far model the siblings' trait values conditional on the siblings' IBD status. In other words, the phenotypic similarity is treated as the dependent variable and the genotypic similarity as the independent variable. However, it has been pointed out that this form of relating these two sets of variables may result in biased results (Sham et al., 2000a; Sham et al., 2002). Sample selection is usually done through trait

https://doi.org/10.1375/twin.7.5.513 Published online by Cambridge University Press

values but not through genotypes: as a result, a significant departure from bivariate normality in the sib-pair trait distribution can be observed. If, in spite of this, the trait value is considered the dependent variable, both regression analysis and variance components can result in inflated type-1 error. If, on the other hand, the trait value is considered the independent variable and the IBD status the dependent variable, the assumption of trait normality can be avoided and the type-1 error is predicted to be correct. This is the general approach of the second group of methods described here.

Reverse Haseman–Elston Regression

The basic idea of the approach proposed by Sham et al. (2002) is to reverse the original Haseman–Elston paradigm and regress the IBD sharing on the trait squared sums and squared differences simultaneously. This idea had already been proposed by Henshall and Goddard (1999). Sham et al.'s (2002) approach is applicable to pedigrees of arbitrary size, but requires the correct specification of the population mean, variance and heritability of the trait. Consider a pedigree i with n members. Let S_{jk} and D_{jk} represent two vectors of dimension n·(n–1)/2 which include the trait squared sums (s_{jk}) and squared differences (d_{jk}), respectively, for all j and k pairs of the pedigree, for $j \neq k$. S_{jk} and D_{jk} represent the independent variables. For simplicity, assume that they are placed in the same vector $Y = [S, D]'$, and that Y is mean-centered. The dependent variable is $(\hat{\pi})_{jk}$, that is, the proportion of alleles IBD between member j and k of the i^{th} pedigree. The array $[(\hat{\pi})_{jk}]$ is inserted into the vector Π, which, again, has dimension $n \cdot (n–1)/2$ and has been mean-centered. Then, the multivariate regression equation of Π on Y is

$$\Pi = \Sigma_{Y\Pi}{}' \cdot \Sigma_Y^{-1} \cdot Y + e \qquad [19]$$

where $\Sigma_{Y\Pi}$ is the covariance matrix between Y and Π, Σ_Y the covariance matrix of Y, and e a vector of residuals. The covariance matrix between Y and Π is composed of two blocks stacked horizontally, where the first block is the covariance matrix between S and Π and the second block is the covariance matrix between D and Π. The diagonal elements of these matrices can be thought to represent a pair's phenotypic similarity (s_{jk} or d_{jk}) in terms of the pair's genotypic similarity ($\hat{\pi}_{jk}$): Wright (1997) and Drigalenko (1998) showed that this equals $2 \cdot Q$ or $-2 \cdot Q$ respectively, where Q is the phenotypic variance explained by the additive effects of the QTL. In addition, the off-diagonal elements of both matrices can be seen as a pair's phenotypic similarity (s_{jk} or d_{jk}) in terms of the genotypic similarity of every other possible pair in the pedigree ($\hat{\pi}_{lm}$). This demonstrates one important property of this statistic: the IBD sharing of a pair of relatives is modeled by the squared sums and squared differences of all relative pairs in the pedigree. These off-diagonal elements can be shown to be defined as $2 \cdot Q \cdot Cov(\hat{\pi}_{jk}, \hat{\pi}_{lm})$ or $-2 \cdot Q \cdot Cov(\hat{\pi}_{jk}, \hat{\pi}_{lm})$, for the squared sums and squared difference matrices, respectively, where $Cov(\hat{\pi}_{jk}, \hat{\pi}_{lm})$, represents the genotypic similarity between pair jk and pair lm of the i^{th} pedigree.

Thus, the matrix $\Sigma_{Y\Pi}$ can be factorized into $Q \cdot \Sigma_\Pi \cdot H$ where Q is a diagonal matrix for the phenotypic variance due to the QTL, Σ_Π the covariance matrix for $\hat{\pi}$, and H a matrix being composed of two matrices stacked horizontally, the first being a diagonal matrix with elements of 2 and off-diagonal elements of 0 and the second a similar matrix with diagonal elements –2. Thus

$$\Pi = Q \cdot \Sigma_\Pi \cdot \left(H \cdot \Sigma_Y^{-1} \cdot Y\right) + e = Q \cdot \Sigma_\Pi \cdot (B) + e$$

$$\Leftrightarrow B' \cdot \Pi = B' \cdot Q \cdot \Sigma_\Pi \cdot B + B' \cdot e$$

$$\Leftrightarrow Q = \frac{B' \cdot \Pi}{B' \cdot \Sigma_\Pi \cdot B} \qquad [20]$$

ignoring the residual contribution. Therefore, for a given family the scalars $B' \cdot \Pi$ and $B' \cdot \Sigma_\Pi \cdot B$ are calculated and their ratio gives an estimate of Q. Across all pedigrees the estimate of Q is given by

$$Q = \frac{\Sigma(B' \cdot \Pi)}{\Sigma(B' \cdot \Sigma_\Pi \cdot B)} \qquad [21]$$

The test statistic that in large samples has asymptotically a chi-square distribution with 1-df under the null hypothesis is

$$T = Q \cdot \Sigma(B' \cdot \Pi) = Q^2 \cdot \Sigma(B' \cdot \Sigma_\Pi \cdot B) \qquad [22]$$

Reverse VC Maximum Likelihood

Sham et al. (2000b) proposed to reverse the 'pi-hat' VC approach by defining the likelihood of the genotype data of a sib-pair conditional on the trait values, $L(G \mid x_i)$. Applying Bayes' theorem and assuming that the likelihood is dependent on G only through π (equivalent to $2 \cdot \Phi$, as defined in [7]), then

$$L(G \mid x_i) \approx \frac{L(x_i \mid \hat{\pi})}{\sum_{\pi} \left[L(x_i \mid \pi) P(\pi)\right]} \qquad [23]$$

where $L(x_i \mid \hat{\pi})$ is calculated as in [16] and the denominator is the weighted sum of the three likelihoods under the theoretical π values of 0, 0.5 and 1. As with other VC approaches, the test for linkage consists in fitting two different models which differ in the covariance structure: H_0, which includes the effect of the QTL and H_1, which does not. The statistic $2 \cdot [\ln(L_{H0}) – \ln(L_{H1})]$ then provides a chi-square test for linkage with 1 degree of freedom. This method requires the correct specification of the phenotypic mean, variance and correlation, which can be obtained from previous studies of the same trait or from preliminary analysis of the sib-pair data.

6. Additional Statistics

There are six additional groups of linkage statistics that I will briefly discuss here. These are frequently mentioned in the literature and have been implemented in popular linkage software packages. In addition, the last three groups address important emergent issues in linkage analysis.

Mean IBD Sharing Statistic for Discordant or Concordant Sib-Pairs

Risch and Zhang (1995, 1996) introduced the mean IBD sharing statistic for the analysis of discordant and concordant sib-pairs. As with the reverse Haseman–Elston and the reverse variance components approaches, the variable being modeled is the IBD information. However, in this case, the trait values are not taken as an independent variable, but rather as a constant. The sib-pairs included for analysis are particular pairs that have been selected on the basis of their joint trait scores. The sample can consist of extreme discordant sibling pairs (EDSPs, defined as one sibling with a trait value above a threshold Z_b and the other with a trait value below Z_l), or high and low concordant pairs (both siblings above Z_b or both below Z_l). If a marker is linked to the trait, the pair's IBD sharing will deviate from the expected value of 1/2 under the null hypothesis (H_0) of no linkage. For discordant pairs, the alternative hypothesis (H_1) is that the mean sharing is less than 1/2. For concordant pairs, H_1 is that the mean sharing is greater than 1/2. Thus, the statistical significance of the IBD sharing deviation can be tested with a one-sample Z test.

Statistics Based on IBD Scoring Functions

Another alternative method for the analysis of quantitative traits based on allele sharing statistics is implemented in the framework of Whittemore and Halpern (1994) and Kong and Cox (1997). This framework is most appropriate for the analysis of binary traits, but it has been adapted for quantitative traits (Abecasis et al., 2002). The basic idea is to define some function S to score each possible inheritance vector for a given pedigree according to the evidence for linkage they provide: the larger the value of S for a given vector w, the greater the evidence for linkage. The weighted scores of all vectors in a pedigree are summed to produce an overall score which reflects that pedigree's contribution to the linkage signal. The standardized overall scores of all pedigrees in the sample are then used to calculate a LOD score based on the Kong and Cox linear or exponential models. Different scoring functions S have been proposed. For quantitative traits, one possible scoring function can be defined as $S(w) = \sum_a (S_a)^2$, where $S_a = \sum_c (y_c - \mu)^2$. That is, the score for each vector w is calculated by summing the squared scores of all the founder alleles (a) present in the vector. The score for each founder allele in the vector w is calculated as the mean deviate for all individuals c who carry that allele in that pedigree (note that y_c is the continuous trait for individual c in the pedigree and μ is the population mean).

Forrest and Feingold Mixed Statistic

Forrest and Feingold (2000) showed that IBD sharing statistics, which model the IBD distribution conditional on trait values, are statistically independent of statistics that model trait values conditional on the IBD information, such as the original Haseman-Elston regression and variance components. Indeed, both approaches contribute complementary rather than redundant information and, thus, they can be combined to form more powerful tests of linkage. They proposed a simple composite statistic for discordant pairs that essentially just adds the standardized traditional Haseman–Elston regression coefficient (β_{HE}) with the standardized mean IBD sharing statistic, both multiplied by appropriate weights (w_{HE} and w_{IBD}, respectively). Formally, the composite statistic is defined by

$$w_{HE}\cdot\frac{\beta_{HE}}{\sqrt{Var(\beta_{HE})}}+w_{IBD}\cdot\frac{(\pi_1+2\cdot\pi_2)-1}{\sqrt{Var(\pi_1+2\cdot\pi_2)}} \qquad [24]$$

where $\pi_1+2\cdot\pi_2$ is the average number of alleles IBD for the sib-pair sample. The sum of the squared weights is constrained to be equal to 1 and both components are normalized, so that both have an expected value of zero and unit variance. In this way, the composite statistic follows a standard normal distribution under the null and, therefore, the test for linkage is a simple t test. Appropriate weights for the composite test can be chosen with knowledge of the ascertainment scheme.

X-Chromosome Linkage Statistics

There are very few descriptions of adaptations of common methods for the analysis of autosomal loci to the analysis of sex-chromosome loci, and those that have been proposed may not as yet have fully grasped the complexity of the analysis. Wiener et al. (2003) described an extension of the revised Haseman–Elston method for the analysis of X-linked loci in sib-pairs. As with adaptations of other methods described below, Wiener et al. (2003) first described the appropriate trait variance parameterization for a two-allele locus, and then derived the appropriate linkage statistic. In this case, it involved the derivation of the expressions for the expected squared trait differences and expected squared trait sum conditional on the IBD information, for sister–sister, brother–brother, and sister–brother pairs. They also showed that singlepoint IBD estimation for the X chromosome is straightforward, even when parental genotypes are unavailable. Ekstrøm (2004) similarly modified the variance components (VC) model to detect QTLs located on X, this time accommodating for multipoint IBD estimation, either using the regression approach of Fulker et al. (1995) and Almasy and Blangero (1998), or the hidden Markov model (HMM) of Kruglyak and Lander (1995). Finally, it is worth mentioning that almost 10 years ago Cordell et al. (1995) provided a simple adaptation of the Risch (1990) allele sharing method to X-linked loci. However, this approach was limited to binary traits.

https://doi.org/10.1375/twin.7.5.513 Published online by Cambridge University Press

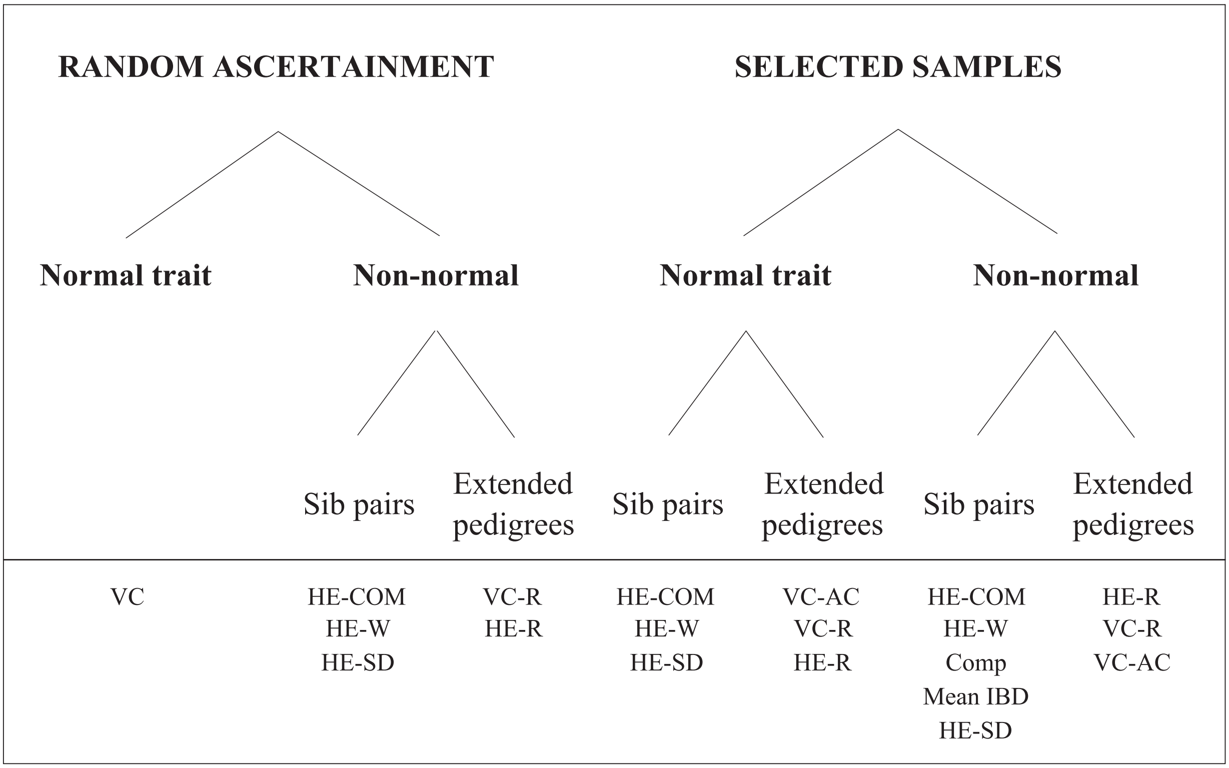


Figure 5

Robust linkage statistic, according to ascertainment scheme, trait distribution and pedigree structure.
Note that these rough guidelines should be used only as a suggestion of methods most likely to provide a test for linkage with correct type-1 error. However, both type-1 error rate and power should always be investigated empirically or theoretically. VC: traditional variance components. HE–COM: Sham and Purcell (2001) Haseman–Elston weighted extension. HE–W: Xu et al. (2000) Haseman–Elston weighted extension. HE–SD: traditional Haseman and Elston (1972) squared difference regression. VC–R: Sham et al. (2000) reverse variance components. HE–R: Sham et al. (2002) reverse Haseman–Elston regression. VC–AC: variance components with 'point-probability' or 'cumulative-probability' ascertainment corrections. Comp: Forrest and Feingold (2000) composite statistic for moderate discordant sib pairs. Mean IBD: Risch and Zhang (1995) mean IBD sharing statistic.

Linkage Statistics that Incorporate Parental Imprinting Effects

Incorporating parent-of-origin effects in linkage statistics involves reparameterizing the components of the phenotypic variance, adjusting the IBD estimation to account for imprinting, and specifying the appropriate null and alternative hypotheses to be tested. Hanson et al. (2001) have done so both for the original Haseman–Elston method and for VC. The traditional Haseman–Elston method was modified by estimating two separate β coefficients according to the source of allele sharing, whereas in the VC approach the QTL component was partitioned into maternal and paternal contributions. The maternal and paternal b coefficients and the maternal and paternal QTL variance components are appropriately multiplied by $\hat{\pi}_{mo}$ and $\hat{\pi}_{fa}$, which represent the proportion of alleles shared IBD derived from the mother and from the father, respectively, with $\hat{\pi}_{mo} + \hat{\pi}_{fa} = \hat{\pi}$. Recently, Shete et al. (2003) extended this model to large pedigrees. Finally, Strauch et al. (2000) and Knapp and Strauch (2004) developed imprinting models for binary traits.

Statistics that Test for Linkage in the Presence of Association

This last group of statistics provides a powerful linkage approach for fine-mapping of a candidate chromosomal region. A locus providing large evidence for linkage may be the true trait locus or it may be in linkage disequilibrium with it. If the locus is indeed the true trait locus, then most or the entire linkage signal at that locus should disappear when the allelic effects of that locus on the trait mean have been removed (Fulker et al., 1999). Fulker et al. (1999) extended the 'pi-hat' VC approach to include this joint test of linkage and association for sib-pairs without parental genotypes. Their approach is the following: consider a single additive two-allele locus, with the effects of the three genotypes A_2A_2, A_1A_2 and A_1A_1 being $-a$, 0 and a. The covariance structure of the VC likelihood model is retained unchanged (see formula [17]); however, the method additionally models the sib-pair expected mean vector in the likelihood function [16] as a function of an overall mean m, the pair mean s_m, and the individual deviation from the pair's mean s_d, as $\mu_1 = m + s_m + (s_d/2)$ and $\mu_2 = m+s_m - (s_d/2)$. Since s_m and s_d are expressed only as a function of the additive allelic effects at a given locus (see Table 1 in Fulker et al., 1999), a test for allelic association simply consists of

https://doi.org/10.1375/twin.7.5.513 Published online by Cambridge University Press

dropping s_m and s_d from the model (for 1-*df*, because s_m and s_d are a function of the same parameter *a*). However, since population stratification can influence pair means (but not each sibling's deviation from the sibship mean), Fulker et al. (1999) pointed out that such test may result in spurious associations. To overcome this limitation, they modified this simple approach to allow the gene effect *a* to be different for the pair means (a_b, between siblings effect) and the pair differences (a_w, within siblings effect). A more robust test of association (albeit less powerful) can be obtained by dropping only a_w (that is, s_d). In this way, the Fulker et al. model can be used to implement different tests, according to the null and alternate hypotheses specified. The following two 1-*df* tests are of particular importance here, assuming no dominance effects: (1) a test for linkage without modeling association, with the null hypothesis having the parameters V_{AQTL}, a_b and a_w fixed to 0, and the alternative hypothesis obtained by setting free V_{AQTL}; and (2) a test for linkage in the presence of association, the null having V_{AQTL} fixed to 0, and the alternate with V_{AQTL} set free (a_b and a_w free in both models). The Fulker et al. (1999) method has been extended to nuclear families of any size (Abecasis et al., 2000), and its theoretical power derived, allowing power calculations to be performed without the need for simulations (Sham et al., 2000a). Finally, Fan and Xiong (2003) have recently suggested an alternative approach to combine VC linkage analysis and association analysis. Their method incorporates LD coefficients and gene effects on the means model, as well as recombination fractions between flanking markers and a putative QTL in the covariance model.

7. Choice of Linkage Statistics

This review had two main goals: first, to introduce basic concepts of linkage analysis, and second, to summarize the statistical fundamentals of the currently most common linkage statistics. Nonetheless, it would be rather incomplete if it did not discuss the relative strengths and weaknesses of each method. This daunting task, which demands a dedicated review by itself, will be briefly addressed here. A number of references are provided that point to some of the original articles addressing this complex topic.

There are three main issues to consider when choosing which method to use. First, and most obviously, the type of linkage analysis to be performed. For example, different statistics will be chosen if the analysis includes a test of imprinting or, alternatively, a test of association. In the same way, statistics should be chosen in accordance with the ascertainment scheme used (for example, the mean IBD sharing statistic or the Forrest & Feingold statistic for discordant sib-pairs). Once this issue has been addressed, the other two factors to consider when selecting the method of analysis are the type-1 error rate and the power provided by the test. Put simply, type-1 error measures how often a significant result would occur when the null hypothesis of no linkage is true (i.e., by chance alone); by contrast, power measures how often a significant result would occur if the alternative hypothesis of linkage was true. The power estimates presented below are based on $\alpha = 0.0001$, corresponding to a central χ^2 statistic of 13.8 (see Sham et al., 2000a; and Williams & Blangero, 1999 for a discussion of this).

Type-1 Error Rate

The linkage statistic for all regression-based methods discussed in section 3 is a *t* test. For this reason, for large sample sizes, these methods have robust type-1 error rates (i.e., close to the nominal levels), even when analyzing selected or nonnormal samples (Feingold, 2002). On the other hand, Sham et al. (2000b) showed that standard variance components analysis of selected samples has inflated type-1 error rate, whether the trait follows a normal distribution or not. Appropriate ascertainment corrections can nonetheless be used to control the type-1 error rates of VC (Andrade & Amos, 2000; Sham et al., 2000a). Similarly, Allison et al. (1999) and Blangero et al. (2001) showed in a range of simulations that standard VC has inflated type-1 error rate when analyzing nonnormal data from a random sample. This effect was aggravated in the presence of strong residual sibling correlation ($r = 0.5$). In practice, Blangero et al. (2001) suggested that an appropriate transformation should be applied for traits where kurtosis ≥2, but this is not guaranteed to always work. If, even after the best transformation, the trait displays a large deviation from normality, other more robust methods should be used for analysis (e.g., regression methods).

By considering the trait values as the dependent variable, both ‘reverse’ methods discussed in section 5 are no longer bound to tight trait distributional assumptions, and seem to have correct type-1 error under common experimental conditions. The simulations performed by Sham et al. (2002) suggest that the type-1 error rate of their regression method is not biased when analyzing either random samples, selected samples with a normally distributed trait, or a nonnormal trait if in the presence of complete IBD information. However, it produced inflated type-1 error when analyzing a nonnormally distributed trait with incomplete IBD information. Similarly, Sham et al. (2000b) showed that their ‘reverse’ VC method leads to a likelihood ratio test with the appropriate type-1 error when analyzing normal or nonnormal data, from either random or selected samples. This is a great improvement when compared to the traditional VC approach.

The type-1 error of the various statistics summarized in section 6 has been less extensively investigated. Forrest and Feingold (2000) showed that the type-1 error of their composite statistic was adequate under any ascertainment scheme simulated. For the X-chromosome, Wiener et al. (2003) showed that under

https://doi.org/10.1375/twin.7.5.513 Published online by Cambridge University Press

an ideal scenario where IBD sharing could be determined unambiguously, their regression-based approach had slightly inflated type-1 error rate if it used the cross-trait product or the linear combination of squared trait difference and squared trait sum were used as the dependent variable. The type-1 error of their method was correct with the traditional use of squared trait differences. Ekstrøm (2004) did not investigate the type-1 error of their VC approach to linkage analysis of the X-chromosome. Finally, Hanson et al. (2001) showed that their imprinting extensions to both regression and VC approaches had type-1 error rates close to the nominal values when testing linkage to a marker locus which was either unlinked to the true QTL or which was linked but had no imprinted effect.

In face of the above, the guidelines presented in Figure 5 may be used as a suggestion of methods most likely to provide a robust test for linkage. Nonetheless, it is important to stress that increased type-1 error rate is perhaps the major obstacle for gene mapping, either through linkage or association analysis. For this reason, it should always be investigated empirically.

Power

There is extensive literature on the power of Haseman–Elston regression-based approaches and variance components. For the regression-based methods HE–SD, HE–CP, HE–W and HE–COM (see section 3 for abbreviations), examples include Elston et al. (2000), Forrest and Feingold (2000), Palmer et al. (2000), Sham and Purcell (2001), Visscher and Hopper (2001) and Yu et al. (2004). See Feingold (2002) for a good discussion of the power of regression-based methods. Together, these simulations suggest that when analyzing normal data from a random sample, the different HE–W extensions and HE–COM provide virtually the same power as variance components and, in some situations, increased power when compared to HE–SD and HE–CP (Figure 6). Thus, when analyzing a normal trait from a population sample, there is no reason to use regression-based methods, but rather, variance components. By contrast, when analyzing selected samples and/or nonnormal traits, the analysis may have to revert to robust regression-based methods. In this case, the methods that seem to provide increased power are HE–COM and Xu et al.'s (2000) HE–W method.

An underlying limit to the power of the different Haseman–Elston methods described in section 3 lies in the fact that they do not accommodate larger sibships and complex pedigrees. This is one of the main strengths of variance components. However, as mentioned above, standard VC is limited to the analysis of normal data from a population sample. Different studies have investigated the power of VC under this condition, including Dolan et al. (1999), Williams and Blangero (1999), Blangero et al. (2001), Sham et al. (2000b), and Sham and Purcell (2001). The simulations provided by Blangero et al. (2001) seem to suggest that VC analysis with less than 1000 sib-pairs will only have enough power (~0.8) to detect a 40% or 20% QTL (residual shared variance fixed at 0.3, α = 0.0001), depending on whether the ascertainment is at random or based on affected sib-pairs (in this case with appropriate ascertainment correction). However, larger sibships provided increased power (see also Dolan et al., 1999; Williams & Blangero, 1999). Finally, Sham et al. (2000a) and Sham and Purcell (2001) derived the theoretical power of VC and showed that it can be approximated by

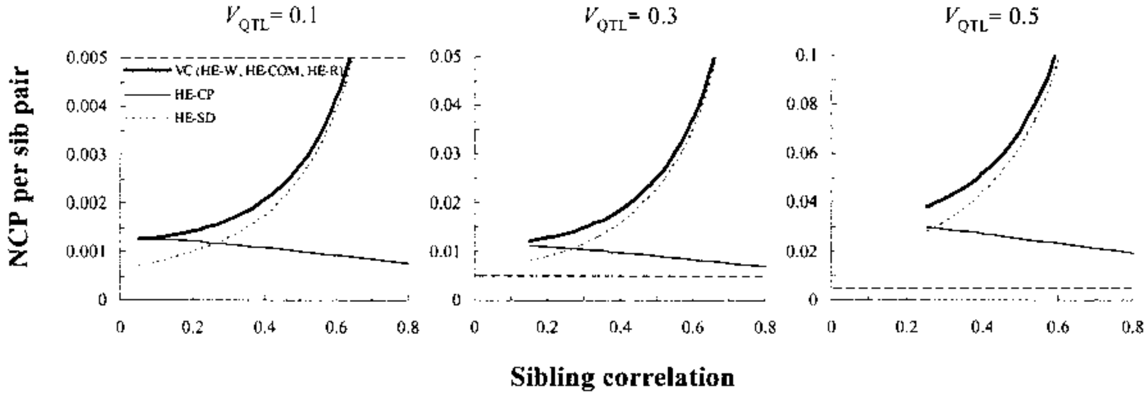


Figure 6

Theoretical power of different methods for a normal trait from a random sample of sib pairs.
Points obtained using the NCP equations derived by Sham and Purcell (2001), assuming perfect IBD information ($\mathrm{var}(\hat{\pi}) = {}^1/_8$). Note the different *y*-axis scales; the horizontal dashed line represents an NCP of 0.005 for comparison across graphs. The overall power of a test based on sib-pairs can be obtained by multiplying the respective NCP per sib-pair by the number of sib-pairs in the sample. An overall NCP of 15.75, 20.76 or 24.96 should be achieved to provide 60%, 80% or 90% power to detect linkage, respectively. These power estimates are based on α = 0.0001, corresponding to a central χ^2 statistic of 13.8. HE-SD: traditional Haseman and Elston (1972) squared trait-difference regression. HE–CP: Elston et al. (2000) revised Haseman–Elston using the cross-product. VC: traditional variance components. HE–W: Xu et al. (2000) Haseman–Elston weighted extension. HE–COM: Sham and Purcell (2001) Haseman–Elston weighted extension. HE–R: Sham et al. (2002) reverse Haseman–Elston regression.

https://doi.org/10.1375/twin.7.5.513 Published online by Cambridge University Press

$$NCP \approx \frac{s\cdot(s-1)}{2}\cdot\frac{\left(1+r^2\right)}{\left(1-r^2\right)^2}\cdot\left[V_A^2\cdot Var(\hat{\pi})+V_D^2\cdot Var(z) + 2\cdot V_A\cdot V_D\cdot Cov(\hat{\pi},z)\right] \qquad [25]$$

Thus, the asymptotical power of VC is proportional to the square of the number of pairs in the sibship (s) – as observed with the simulations described before – to the sibling correlation (r), to the squared variance due to the additive QTL component (V_{AQTL}), to the marker informativeness (as reflected in the variance of $\hat{\pi}$ and π_2), and to the squared variance due to the dominance QTL component (V_{AQTL}). In addition, they showed that if a QTL is additive, the attenuation of the NCP with increasing incomplete linkage is by a factor of $(1 - 2\cdot\theta)^4$, where θ is the recombination fraction between the marker and the trait loci. This raises the important problem of marker density and the power to detect linkage (see Atwood & Heard-Costa, 2003; Kruglyak 1997; Terwilliger et al. 1992). Formula [25] calculates the contribution of a particular sibship to the VC likelihood ratio statistic under a specific range of model parameters, and it is implemented in the software GPC (Purcell et al., 2003 ; Sham et al., 2000a).

The power of the two 'reverse' methods described in section 5 were discussed by Sham et al. (2002), Sham et al. (2000b), and Sham and Purcell (2001). The main strength of the 'reverse' regression method (HE–R) of Sham et al. (2002) compared to the conventional regression methods of section 3 is that this method is applicable to pedigrees of arbitrary size. When compared to standard VC, the strength lies in that selected samples can be analyzed without incurring an inflated type-1 error (as long as the trait is normally distributed). When analyzing a normally distributed trait from a random sample, the power of HE–R (as expressed by simulated mean test statistics) was the same as VC for sibships of size two and three, but greater for larger sibships (Sham et al., 2002). This property, however, was challenged recently by Yu et al. (2004). They showed that for sibships of size four or larger, the asymptotic distribution of the HE–R under the hypothesis of linkage is not a noncentral χ^2, and that in fact, this method seems to provide almost the same empirical power as VC. When analyzing a nonnormal trait from a random sample, the HE–R provides very low power. Nonetheless, this may still be comparable to the power provided by the original Haseman–Elston method or any of its extensions. Lastly, misspecification of the trait mean can reduce the power of HE–R considerably. The major strength of the 'reverse' VC method is that it is robust when analyzing nonnormal data, either from a random or a selected sample. Nonetheless, the power to detect linkage under nonnormality seems to be extremely low. Thus, although the risk of false-positives is minimized, false-negatives are very likely to be observed.

Finally, a brief note on the power of some of the methods discussed in section 6. Different studies have documented the increase in power provided by extreme selection methods (Cardon & Fulker, 1994; Carey & Williamson, 1991; Gu et al., 1996; Risch & Zhang ,1995). However, Allison et al. (1998) showed that under particular conditions, such extreme designs do not always result in increased power to detect a QTL. Forrest and Feingold (2000) showed that the power of their composite method exceeds that of the mean IBD sharing statistic or the original Haseman–Elston regression when sib-pairs are chosen to be moderately discordant (trait values below the 35% or above the 65% quantiles). For the X-chromosome statistics, Wiener et al. (2003) reported a power of ~0.6 (male QTL heritability 0.4) with 500+ sib-pairs for their regression method. On the other hand, Ekstrøm (2004) reported a power of ~0.2 (male QTL heritability 0.5, for a 10-cM map) for their variance components extension, using 100 nuclear families with two male and two female siblings each. For the imprinting methods, both Hanson et al. (2001) and Shete and Amos (2002) concluded that modeling imprinting will only provide a significant improvement in the power to detect linkage when the imprinting effect was moderate to large. They suggested the use of imprinting models only in regions where evidence for linkage has been previously observed. Lastly, Fulker et al. (1999) showed that if a marker locus was the trait locus itself or was in complete linkage disequilibrium with it, their method of testing for linkage while modeling association resulted in a significant drop in the linkage signal, when compared to a method which did not model association. This highlighted the importance of their method to determine whether a marker locus is the true trait locus or simply in very close proximity to it.

In summary, many linkage methods have been developed, with varying strengths and weaknesses. Which method to use depends on factors such as the ascertainment scheme, data properties, and the aim of the analysis. Ultimately, however, it depends on the type-1 error and the power provided by the different alternatives. Both the type-1 error and the power of a test should always be investigated to assess the likelihood of observing false-positive and false-negative results.

Acknowledgments

Invaluable suggestions to improve the original manuscript were provided by David Duffy, Nick Martin, Peter Visscher, Benjamin Neale and Daniëlle Posthuma. The author is supported by the doctorate grant SFRH/BD/4824/2001 from the Fundação para a Ciência e Tecnologia, Portugal.

References

Abecasis, G. R., Cardon, L. R., & Cookson, W. O. (2000). A general test of association for quantitative

https://doi.org/10.1375/twin.7.5.513 Published online by Cambridge University Press

traits in nuclear families. *American Journal of Human Genetics, 66*, 279–292.

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

Allison, D. B., Heo, M., Schork, N. J., Wong, S. L., & Elston, R. C. (1998). Extreme selection strategies in gene mapping studies of oligogenic quantitative traits do not always increase power. *Human Heredity, 48*, 97–107.

Allison, D. B., Neale, M. C., Zannolli, R., Schork, N. J., Amos, C. I., & Blangero, J. (1999). Testing the robustness of the likelihood-ratio test in a variance-component quantitative-trait loci-mapping procedure. *American Journal of Human Genetics, 65*, 531–544.

Almasy, L., & Blangero, J. (1998). Multipoint quantitative-trait linkage analysis in general pedigrees. *American Journal of Human Genetics, 62*, 1198–1211.

Atwood, L. D., & Heard-Costa, N. L. (2003). Limits of fine-mapping a quantitative trait. *Genetic Epidemiology, 24*, 99–106.

Blangero, J., Williams, J. T., & Almasy, L. (2001). Variance component methods for detecting complex trait loci. *Advances in Genetics, 42*, 151–181.

Botstein, D., White, R. L., Skolnick, M., & Davis, R. W. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American Journal of Human Genetics, 32*, 314–331.

Cardon, L. R., & Fulker, D. W. (1994). The power of interval mapping of quantitative trait loci, using selected sib pairs. *American Journal of Human Genetics, 55*, 825–833.

Carey, G., & Williamson, J. (1991). Linkage analysis of quantitative traits: increased power by using selected samples. *American Journal of Human Genetics, 49*, 786–796.

Clerget-Darpoux, F., Bonaiti-Pellie, C., & Hochez, J. (1986). Effects of misspecifying genetic parameters in LOD score analysis. *Biometrics, 42*, 393–399.

Cordell, H. J., Kawaguchi, Y., Todd, J. A., & Farrall, M. (1995). An extension of the Maximum LOD Score method to X-linked loci. *Annals of Human Genetics, 59*(Pt 4), 435–449.

de Andrade, M, & Amos, C. I. (2000). Ascertainment issues in variance components models. *Genetic Epidemiology, 19*(4), 333–344.

Dolan, C. V., Boomsma, D. I., & Neale, M. C. (1999). A note on the power provided by sibships of sizes 2, 3, and 4 in genetic covariance modeling of a codominant QTL. *Behaviour Genetics, 29*, 163–170.

Drigalenko, E. (1998). How sib pairs reveal linkage. *American Journal of Human Genetics, 63*, 1242–1245.

Eaves, L. J., Neale, M. C., & Maes, H. (1996). Multivariate multipoint linkage analysis of quantitative trait loci. *Behaviour Genetics, 26*, 519–525.

Ekstrom, C. T. (2004). Multipoint linkage analysis of quantitative traits on sex-chromosomes. *Genetic Epidemiology, 26*, 218–230.

Elston, R. C. (1989). Man bites dog? The validity of maximizing LOD scores to determine mode of inheritance. *American Journal of Medical Genetics, 34*, 487–488.

Elston, R. C., Buxbaum, S., Jacobs, K. B., & Olson, J. M. (2000). Haseman and Elston revisited. *Genetic Epidemiology, 19*, 1–17.

Elston, R. C., & Stewart, J. (1971). A general model for the genetic analysis of pedigree data. *Human Heredity, 21*, 523–542.

Fan, R., & Xiong, M. (2003). Combined high resolution linkage and association mapping of quantitative trait loci. *European Journal of Human Genetics, 11*, 125–137.

Feingold, E. (2002). Regression-based quantitative-trait-locus mapping in the 21st century. *American Journal of Human Genetics, 71*, 217–222.

Forrest, W. F. (2001). Weighting improves the 'new Haseman–Elston' method. *Human Heredity, 52*, 47–54.

Forrest, W. F., & Feingold, E. (2000). Composite statistics for QTL mapping with moderately discordant sibling pairs. *American Journal of Human Genetics, 66*, 1642–1660.

Fulker, D. W., Cherny, S. S., & Cardon, L. R. (1995). Multipoint interval mapping of quantitative trait loci, using sib pairs. *American Journal of Human Genetics, 56*, 1224–1233.

Fulker, D. W., Cherny, S. S., Sham, P. C., & Hewitt, J. K. (1999). Combined linkage and association sib-pair analysis for quantitative traits. *American Journal of Human Genetics, 64*, 259–267.

Greenberg, D. A. (1989). Inferring mode of inheritance by comparison of lod scores. *American Journal of Medical Genetics, 34*, 480–486.

Gu, C., Todorov, A., & Rao, D. C. (1996). Combining extremely concordant sibpairs with extremely discordant sibpairs provides a cost effective way to linkage analysis of quantitative trait loci. *Genetic Epidemiology, 13*, 513–533.

Hanson, R. L., Kobes, S., Lindsay, R. S., & Knowler, W. C. (2001). Assessment of parent-of-origin effects in linkage analysis of quantitative traits. *American Journal of Human Genetics, 68*, 951–962.

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

Heath, S. C. (1997). Markov chain Monte Carlo segregation and linkage analysis for oligogenic models. *American Journal of Human Genetics, 61*, 748–760.

https://doi.org/10.1375/twin.7.5.513 Published online by Cambridge University Press

Henshall, J. M., & Goddard, M. E. (1999). Multiple-trait mapping of quantitative trait loci after selective genotyping using logistic regression. *Genetics, 151*, 885–894.

Hodge, S. E., & Elston, R. C. (1994). Lods, wrods, and mods: the interpretation of LOD scores calculated under different models. *Genetic Epidemiology, 11*, 329–342.

Hopper, J. L., & Mathews, J. D. (1982). Extensions to multivariate normal models for pedigree analysis. *Annals of Human Genetics, 46* (4), 373–383.

Knapp, M., & Strauch, K. (2004). Affected-sib-pair test for linkage based on constraints for identical-by-descent distributions corresponding to disease models with imprinting. *Genetic Epidemiology, 26*, 273–285.

Kong, A., & Cox, N. J. (1997). Allele-sharing models: LOD scores and accurate linkage tests. *American Journal of Human Genetics, 61*, 1179–1188.

Kruglyak, L. (1997). The use of a genetic map of biallelic markers in linkage studies. *Nature Genetics, 17*, 21–24.

Kruglyak, L., Daly, M. J., Reeve-Daly, M. P., & Lander, E. S. (1996). Parametric and nonparametric linkage analysis: a unified multipoint approach. *American Journal of Human Genetics, 58*, 1347–1363.

Kruglyak, L., & Lander, E. S. (1995). Complete multipoint sib-pair analysis of qualitative and quantitative traits. *American Journal of Human Genetics, 57*, 439–454.

Lander, E., & Kruglyak, L. (1995). Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nature Genetics, 11*, 241–247.

Lander, E. S., & Green, P. (1987). Construction of multilocus genetic linkage maps in humans. *Proceedings of the National Academy of Sciences of the United States of America, 84*, 2363–2367.

Morton, N. E. (1955). Sequential tests for the detection of linkage. *American Journal of Human Genetics, 7*, 227–318.

Neale, M. C., & Maes, H. H. M. (1999). *Methodology for genetic studies of twins and families*. Dordrecht, The Netherlands: Kluwer Academic.

Ott, J. (1991). *Analysis of human genetic linkage*. Baltimore, MD: Johns Hopkins University Press.

Palmer, L. J., Jacobs, K. B., & Elston, R. C. (2000). Haseman and Elston revisited: the effects of ascertainment and residual familial correlations on power to detect linkage. *Genetic Epidemiology, 19*, 456–460.

Posthuma, D., Beem, A. L., de Geus, E. J., van Baal, G. C., von Hjelmborg, J. B., Iachine, I., & Boomsma, D. I. (2003). Theory and practice in quantitative genetics. *Twin Research, 6*, 361–376.

Purcell, S., Cherny, S. S., & Sham, P. C. (2003). Genetic Power Calculator: Design of linkage and association genetic mapping studies of complex traits. *Bioinformatics, 19*, 149–150.

Risch, N. (1984). Segregation analysis incorporating linkage markers. I. Single-locus models with an application to type I diabetes. *American Journal of Human Genetics, 36*, 363–386.

Risch, N. (1990). Linkage strategies for genetically complex traits. III. The effect of marker polymorphism on analysis of affected relative pairs. *American Journal of Human Genetics, 46*, 242–253.

Risch, N., & Zhang, H. (1995). Extreme discordant sib pairs for mapping quantitative trait loci in humans. *Science, 268*, 1584–1589.

Risch, N. J., & Zhang, H. (1996). Mapping quantitative trait loci with extreme discordant sib pairs: sampling considerations. *American Journal of Human Genetics, 58*, 836–843.

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

Sham, P. C., Cherny, S. S., Purcell, S., & Hewitt, J. K. (2000a). Power of linkage versus association analysis of quantitative traits, by use of variance-components models, for sibship data. *American Journal of Human Genetics, 66*, 1616–1630.

Sham, P. C., & Purcell, S. (2001). Equivalence between Haseman-Elston and variance-components linkage analyses for sib pairs. *American Journal of Human Genetics, 68*, 1527–1532.

Sham, P. C., Purcell, S., Cherny, S. S., & Abecasis, G. R. (2002). Powerful regression-based quantitative-trait linkage analysis of general pedigrees. *American Journal of Human Genetics, 71*, 238–253.

Sham, P. C., Zhao, J. H., Cherny, S. S., & Hewitt, J. K. (2000b). Variance-components QTL linkage analysis of selected and non-normal samples: Conditioning on trait values. *Genetic Epidemiology, 19* (Suppl. 1), 22–28.

Shete, S., & Amos, C. I. (2002). Testing for genetic linkage in families by a variance-components approach in the presence of genomic imprinting. *American Journal of Human Genetics, 70*, 751–757.

Shete, S., Zhou, X., & Amos, C. I. (2003). Genomic imprinting and linkage test for quantitative-trait Loci in extended pedigrees. *American Journal of Human Genetics, 73*, 933–938.

Strauch, K., Fimmers, R., Kurz, T., Deichmann, K. A., Wienker, T. F., & Baur, M. P. (2000). Parametric and nonparametric multipoint linkage analysis with imprinting and two-locus-trait models: application to mite sensitization. *American Journal of Human Genetics, 66*, 1945–1957.

Terwilliger, J. D., Ding, Y., & Ott, J. (1992). On the relative importance of marker heterozygosity and intermarker distance in gene mapping. *Genomics, 13*, 951–956.

Tiwari, J. L., Hodge, S. E., Terasaki, P. I., & Spence, M. A. (1980). HLA and the inheritance of multiple scle-

https://doi.org/10.1375/twin.7.5.513 Published online by Cambridge University Press

rosis: linkage analysis of 72 pedigrees. *American Journal of Human Genetics, 32*, 103–111.

Visscher, P. M., & Hopper, J. L. (2001). Power of regression and maximum likelihood methods to map QTL from sib-pair and DZ twin data. *Annals of Human Genetics, 65*, 583–601.

Weiss, K. M., & Terwilliger, J. D. (2000). How many diseases does it take to map a gene with SNPs? *Nature Genetics, 26*, 151–157.

Whittemore, A. S., & Halpern, J. (1994). A class of tests for linkage using affected pedigree members. *Biometrics, 50*, 118–127.

Wiener, H., Elston, R. C., & Tiwari, H. K. (2003). X-linked extension of the revised Haseman–Elston algorithm for linkage analysis in sib pairs. *Human Heredity, 55*, 97–107.

Williams, J. T., & Blangero, J. (1999). Power of variance component linkage analysis to detect quantitative trait loci. *Annals of Human Genetics, 63*, 545–563.

Wright, F. A. (1997). The phenotypic difference discards sib-pair QTL linkage information. *American Journal of Human Genetics, 60*, 740–742.

Xu, X., Weiss, S., & Wei, L. J. (2000). A unified Haseman–Elston method for testing linkage with quantitative traits. *American Journal of Human Genetics, 67*, 1025–1028.

Yu, X., Knott, S. A., & Visscher, P. M. (2004). Theoretical and empirical power of regression and maximum-likelihood methods to map quantitative trait Loci in general pedigrees. *American Journal of Human Genetics, 75*, 17–26.

https://doi.org/10.1375/twin.7.5.513 Published online by Cambridge University Press