BMC Genetics

Methodology article

A quantitative linkage score for an association study following a linkage analysis

Tao Wang and Robert C Elston*

Address: Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland, USA

Email: Tao Wang - txw54@case.edu; Robert C Elston* - rce@darwin.case.edu

* Corresponding author

Published: 20 January 2006

BMC Genetics 2006, **7**:5 doi:10.1186/1471-2156-7-5

Received: 04 August 2005

Accepted: 20 January 2006

This article is available from: http://www.biomedcentral.com/1471-2156/7/5

Abstract

Background: Currently, a commonly used strategy for mapping complex quantitative traits is to use a genome-wide linkage analysis to narrow suspected genes to regions on a scale of centiMorgans (cM), followed by an association analysis to fine map the genetic variation in regions showing linkage. Two important questions arise in the design and the resulting inference at the association stage of this sequential procedure: (1) how should we design an efficient association study given the information provided by the previous linkage study? and (2) can an association in a linkage region explain, in part, the detected linkage signal?

Results: We derive a quantitative linkage score (QLS) based on Haseman-Elston regression (Haseman and Elston 1972) and make use of this score to address both questions. In designing an association study, the selection of a subsample from the linkage study sample can be guided by the linkage information summarized in the QLS. When heterogeneity exists, we show that selection based on the QLS can increase the proportion of sample individuals from the subpopulation affected by a disease allele and therefore greatly improves the power of the association study. For the resulting inference, we frame as a hypothesis test the question of whether a linkage signal in a region can be in part explained by a marker allele. A simple one sided paired t-statistic is defined by comparing the two sets of QLSs obtained with/without modeling a marker association: a significant difference indicates that the marker can at least partly account for the detected linkage. We also show that this statistic can be used to detect a spurious association.

Conclusion: All our results suggest that a careful examination of QLSs should be helpful for understanding the results of both association and linkage studies.

Background

Identifying genes underlying complex quantitative traits, which are often heterogeneous and multifactorial, is still a great challenge in genetic epidemiology studies. Currently, a commonly used strategy for mapping complex traits is to use a genome-wide linkage analysis to narrow suspected genes to regions on a scale of centiMorgans (cM), followed by an association analysis to fine map the genetic variation in regions showing linkage. At the association stage of this sequential process, we are often interested in two questions: (1) how should we design a powerful and efficient association study given the information provided by the previous linkage study? and (2) can an association in a linkage region explain, in part, the detected linkage signal? Although these questions that arise respectively at the design and inference stages are two

quite different aspects of an association study, they are related because both questions essentially rely on the interdependence of linkage and association. Here, we derive a quantitative linkage score (QLS) from Haseman-Elston linkage regression [1] and make use of this score to address both questions in the scenario of analyzing a complex quantitative trait.

The loci predisposing to a complex quantitative trait are usually expected to have small effects. One important reason for this, among others, is heterogeneity of the phenotype, where an allele of interest may have no effect on some individuals because they have different genetic and environmental backgrounds. If these individuals are included in the sample used in the association study, the effect of the examined allele is "diluted" and this leads to great difficulty in detecting association. Careful selection of individuals from the sample to exclude such possible "dilution" should presumably provide greater power. Ideally, we should like to find a variable, such as age, sex or ethnicity, that indicates heterogeneous persons. Unfortunately, such an indicator variable is often unclear or unavailable for a complex trait. Nevertheless, if an association study follows a linkage study, selection of the sample for the association study may be guided by the linkage information already obtained, using the linkage signal as a natural heterogeneity indicator. This idea has long been recognized and implemented in practice [2-4]. Fingerlin et al. (2004) systematically examined the selection of cases for a case-control association study based on allele-sharing information provided by affected members of a family [5]. We focus here on sample selection for an association study of a quantitative trait and show the usefulness of the QLS when heterogeneity exists.

After an association has been detected between the trait and a marker allele in the region of linkage, the question of whether this association accounts, in part, for the previously found linkage signal is not trivial. If the allele statistically associated with the trait is partly responsible for the linkage, we may be more confident that this allele is itself functional or in linkage disequilibrium with the true functional variant, rather than a false discovery resulting from other causes. On the other hand, if the associated allele cannot explain any linkage signal, we may consider adding more association markers to the region in order to avoid missing a possible genetic variant affecting the trait of interest. In the case of affected sibs (or other affected relatives) used for linkage analysis, one approach is to examine the difference in the allele sharing identical by descent (IBD) between members of families selected on the basis of the associated marker [2,6]. We address this question for a quantitative trait by testing whether there is a significant difference between the QLS with and without including this marker in the model. We show that this test is essentially the same as examining the interaction between the linkage and association signals and therefore is related to the genotype-IBD sharing test (GIST) proposed by Li et al. (2004) for affected sibship data[6]. Fulker (1999) proposed a similar idea, in the context of a variance component model, simultaneously modeling the association and linkage in the mean and variance-covariance structure of a family [7]. They focused on testing a similar, but different, hypothesis to determine whether the allele is the true candidate or is merely in disequilibrium with the trait locus, by comparing a model with all the parameters freely estimated to a model in which the linked genetic variance of the quantitative trait locus (QTL) is set to zero, on the assumption that there is a single variant responsible for the linkage signal[8].

In this paper, we propose a linkage score derived from quantitative trait linkage analysis that has important applications when an association study follows a linkage analysis. Although the linkage score derived here can be easily extended to general families, to implement our approach we focus here on nuclear families. We first derive the linkage score in the method section. Then we perform computer simulations to examine the usefulness of this score to select a sample for an association study when heterogeneity exists, and to clarify whether the association can, at least in part, explain the linkage signal.

Methods

Our goal is to derive a score that captures the linkage information for quantitative traits in a way that will be useful for a follow-up association study. For simplicity of presentation, we assume the quantitative trait value may be affected by the presence of an allele without any other covariates present, which is not a necessary limitation for our derivation. We suppose linkage markers have been genotyped for family members and therefore the proportion of alleles shared IBD at a particular location can be estimated for all pairs of relatives in a pedigree [9,10].

Quantitative linkage score (QLS)

We first derive the QLS. Suppose we have recruited N sibships. The trait value y_{ik} of sib $i(1, ..., n_k)$ in sibship $k(1, ..., N)$ is modeled by

$$y_{ik} = \mu_k + x_{ik}b + e_{ik}, \quad (1)$$

where μ_k is the sibship specific mean, which absorbs family-level effects such as polygenic and common environmental effects [11]; b is the effect of the quantitative trait locus (QTL), which may include both additive and dominant effects; x_{ik} is the corresponding vector of design variables indicating the genotype of the QTL; and e_{ik} is an individual-level random effect. For simplicity of exposition only, we assume the QTL effect is additive and there-

fore x_{ik} can be coded as one variable to indicate the number of copies of the allele of interest. Otherwise, it can be coded as a vector with two elements, for additive and dominant effects, respectively. Because in a linkage analysis the genotype of a QTL (x_{ik}) is not observed (or the marker cannot be assumed to be in linkage disequilibrium with the QTL), we are not able to estimate directly. However, we can model the QTL effect in the variance-covariance matrix at the family-level. Under the trait model (1), the variance-covariance matrix of sibship k is given by

$$E\begin{pmatrix} (y_{1k}-\mu_k)(y_{1k}-\mu_k) & \cdots & (y_{1k}-\mu_k)(y_{n_kk}-\mu_k) \\ \cdots & \cdots & \cdots \\ (y_{1k}-\mu_k)(y_{n_kk}-\mu_k) & \cdots & (y_{n_kk}-\mu_k)(y_{n_kk}-\mu_k) \end{pmatrix} = \begin{pmatrix} \sigma_b^2+\sigma_e^2 & \cdots & IBD_{1n_kk}\sigma_b^2 \\ \cdots & \cdots & \cdots \\ IBD_{1n_kk}\sigma_b^2 & \cdots & \sigma_b^2+\sigma_e^2 \end{pmatrix},$$

where σ_b^2 is the variance of the QTL, σ_e^2 is the individual random effect variance and IBD_{ijk} is the proportion of marker alleles shared IBD by sibs i and j in family k. Because both matrices are symmetric and the diagonal elements do not include linkage information, we only consider the lower triangular elements. We rearrange these elements of the above matrices as vectors of length $n_k(n_k - 1)/2$ by stacking one column on top of the other and then have

$$E\begin{pmatrix} (y_{1k}-\mu_k)(y_{2k}-\mu_k) \\ \cdots \\ (y_{ik}-\mu_k)(y_{jk}-\mu_k) \\ .. \\ (y_{(n-1)_kk}-\mu_k)(y_{n_kk}-\mu_k \end{pmatrix} = \begin{pmatrix} \sigma_b^2 IBD_{12k} \\ \cdots \\ \sigma_b^2 IBD_{ijk} \\ \cdots \\ \sigma_b^2 IBD_{(n-1)nk} \end{pmatrix}. \quad (2)$$

We can treat the above equation as a version of Haseman-Elston (HE) regression. The sibship specific mean μ_k is usually unknown and needs to be estimated; various estimates have been discussed and a shrinkage estimate $\hat{\mu}_k$ has been recommended [11,12]. For the simulations performed in this paper, the $\hat{\mu}_k$ was estimated by the function *lme* in the *R* package http://cran.us.r-project.org. In a HE regression, linkage is detected by testing whether the QLT variance $\sigma_b^2 > 0$, which is equivalent to testing the correlation between IBD_{ijk} and the trait similarity between the two sibs, as measured by $(y_{ik} - \hat{\mu}_k)(y_{jk} - \hat{\mu}_k)$ in our case. From this perspective, the linkage information provided by a sibpair can be captured by the score

$$U_{ijk} = (y_{ik}-\mu_k)(y_{jk}-\mu_k)(IBD_{ijk}-0.5) \quad (3)$$

From equation (3), we can see that for an additive trait model a positive score supports linkage and a negative score is evidence against linkage. When the inheritance model is unclear, we may take the "minmax" method to estimate the proportion of marker alleles shared IBD for a full sibpair, i.e. $IBD_{ijk} = 0.275f_{ijk1} + f_{ijk2}$ instead of $IBD_{ijk} = 0.5f_{ijk1} + f_{ijk2}$, where f_{ijk1} and f_{ijk2} are probabilities of 1 and 2 alleles shared IBD, respectively [13]. We can simply sum the scores for all the pairs in a sibship to obtain a measure of linkage evidence for this sibship, because the sibship mean absorbs any residual correlation among the sibs. We may define the QLS more generally as $U_{ijk} = S_{ijk}(IBD_{ijk} - 0.5)$, where S_{ijk} can be any measure of trait similarity, for example the squared sibpair difference, or a weighted average of the squared (mean-corrected) sum and the squared difference in trait values of two sibs, all of which are provided by different versions of HE regression implemented in the software SIBPAL of S.A.G.E. (2004). Different measures of trait similarity have been discussed in detail in the literature [e.g. [11,14-16]]. In those cases we may need to consider, in order to sum the QLSs within a sibship, a weight function appropriate for the correlation between scores among sibpairs. Note there is no difficulty in extending the QLS to qualitative traits. For example, for affected sibpairs S_{ijk} can be defined as 1 for all pairs and the linkage score is simply given by $U_{ijk} = (IBD_{ijk} - 0.5)$, which is related to the NPL score [17] and the statistic of the mean test [18].

Application of the QLS in selecting a sample for an association study

We consider selecting a set of unrelated individuals from sibships previously used for a QTL linkage analysis. In the case of a complex quantitative trait where heterogeneity exists, the goal of an association study is to detect a variant with maximum power. We emphasize that such a study would not be a classic epidemiologlcal study done to determine the attributable risk, for which subjects should be drawn randomly from a population. Rather, the study we discuss here is done for gene finding and therefore the selection of the sample should be done to provide maximum power rather than to represent the whole population.

Suppose that a population consists of two subpopulations (P1 and P2) with proportions q_1 and q_2 respectively ($q_1 + q_2 = 1$), where the gene variant has an effect in only one subpopulation (P1). To examine the usefulness of the QLS in selecting a sample for an association study, we theoretically compare the proportions of individuals affected by a disease allele selected from a homogenous subpopulation (P1) in two selected samples: one sample is obtained by randomly selecting sibships (proportion q_r) and the other is obtained by selecting sibships with QLS>0 (proportion q_{qls}). To simplify the theoretical deri-

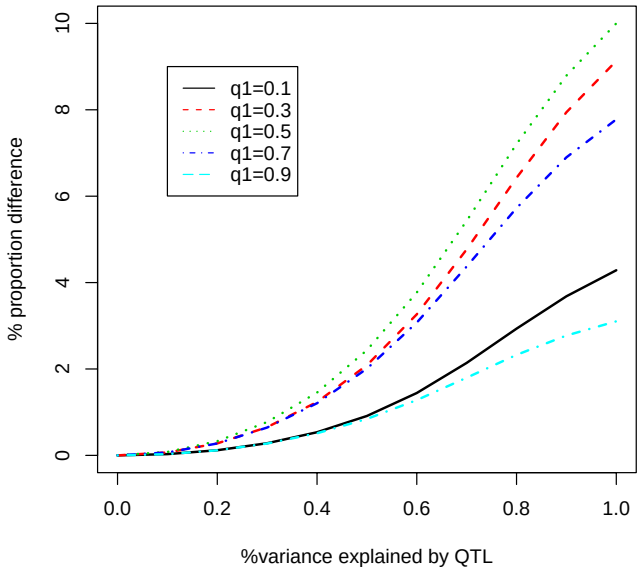

Figure 1
Difference in the proportion of individuals from subpopulation 1 between random sampling and QLS sampling. Subpopulation 1 comprises people who are affected by the QTL and q_1 is the proportion of subpopulation 1 in the whole population.

vation, we assume known IBD sharing and sibships of size 2 (independent sibpairs).

Let $T_k = [(y_{1k} - \mu_k), (y_{2k} - \mu_k)]^T$, where the superscript T denotes transpose and the subscripts 1 and 2 indicate two sibs in a sibship. With the assumption of normal individual effects e_{ik}, $T \sim N(0, \Sigma_k)$, where

$$\sum_k = \begin{bmatrix} (\sigma_b^2 + \sigma_e^2) & IBD_{12k}\sigma_b^2 \\ IBD_{12k}\sigma_b^2 & (\sigma_b^2 + \sigma_e^2) \end{bmatrix}.$$

To further simplify the presentation, we standardize T_k as Z_k, so that the correlation matrix of Z_k is

$$\begin{pmatrix} 1 & \rho_k \\ \rho_k & 1 \end{pmatrix}$$

where $\rho_k = 0,\ 0.5\,\sigma_b^2/(\sigma_b^2 + \sigma_e^2)$ and $\sigma_b^2/(\sigma_b^2 + \sigma_e^2)$, respectively, for proportions 0, 0.5, and 1 allele sharing IBD. With the assumption that a random sample of sibpairs is used for the linkage analysis, we have $q_r = q_1$ and

$$q_{qls} = \frac{q_1}{q_1 + \frac{q_2}{\left[\frac{1}{\pi}\arctan\left(\frac{\rho^2_{IBD=1}}{1-\rho^2_{IBD=1}}\right)+1\right]}}, \qquad (4)$$

where $\rho_{IBD=1}$ is the correlation between two sibs of a pair with proportion 1 IBD sharing. (see Appendix 1). It is obvious that $\left[\frac{1}{\pi}\arctan\left(\frac{\rho^2_{IBD=1}}{1-\rho^2_{IBD=1}}+1\right)\right] \geq 1$, and so q_{qls} is always $\geq q_r$. From this inequality, we can also see that the difference between q_{qls} and q_r depends on (1) the proportion of P1: when $q_1 = 0.5$, the difference is maximum; and (2) the variance explained by the QTL: the difference is an increasing function of $\rho_{IBD=1}$, i.e. $\sigma_b^2/(\sigma_b^2 + \sigma_e^2)$. The difference between q_{qls} and q_r is presented in Figure 1, which shows that selection based on the QLS can increase the proportion of individuals from subpopulation 1 at most 10%. Nevertheless, a slight difference in this proportion is not trivial, because it may greatly improve the power of an association study (see results).

Application of the QLS to assess the correlation of association with previous linkage

To answer the question of whether a linkage signal in a region can be in part explained by a marker allele used in an association study, we compare the QLS on incorporating and not incorporating this marker into the trait model (equation 1), which we call the first (or individual) level regression, to distinguish it from the second (or family) level regression (equation 2). We frame this problem as a hypothesis test. When a marker is included in the model at the individual level, the variance-covariance matrix of sibship k is given by

$$E\begin{pmatrix} (y_{1k}-\mu_k-x_{1k}b)(y_{1k}-\mu_k-x_{1k}b) & \cdots & (y_{1k}-\mu_k-x_{1k}b)(y_{n_kk}-\mu_k-x_{n_kk}b) \\ \cdots & \cdots & \cdots \\ (y_{1k}-\mu_k-x_{1k}b)(y_{n_kk}-\mu_k-x_{n_kk}b) & \cdots & (y_{n_kk}-\mu_k-x_{n_kk}b)(y_{n_kk}-\mu_k-x_{n_kk}b) \end{pmatrix}$$
$$= \begin{pmatrix} \sigma_b^2+\sigma_e^2 & \cdots & IBD_{1n_kk}\sigma_b^2 \\ \cdots & \cdots & \cdots \\ IBD_{1n_kk}\sigma_b^2 & \cdots & \sigma_b^2+\sigma_e^2 \end{pmatrix}$$

where x_{ik} is a genotype code for the marker and b is its effect on the trait, which may arise from a "true" association (the marker is the QTL itself or is in linkage disequilibrium with the QTL), or from a "spurious" association (e.g. due to population stratification). Based on the above equation, we can obtain the corresponding QLS with the marker included in the above regression model, which is given by

$$U_{ijk} = (y_{ik} - \mu_k - x_{ik}b)(y_{jk} - \mu_k - x_{jk}b)(IBD_{ijk} - 0.5),$$

where $\hat{b}$ and $\hat{\mu}_k$ are the estimates of b and μ_k, respectively. In the following presentation, we denote the QLS obtained with and without modeling an association marker $U_{ijk}^{(a)}$ and $U_{ijk}^{(b)}$, respectively. Given these two sets of QLSs, $U_{ijk}^{(a)}$ and $U_{ijk}^{(b)}$, we expect the mean score $\bar{U}^{(b)}$ to be larger than $\bar{U}^{(a)}$ when the associated marker is the QTL, or is linked in disequilibrium with it. To compare the two means, we may apply a one-sided paired t-test. Let $\hat{U}_{ijk}^{(a)} = U_{ijk}^{(a)} - \bar{U}^{(a)}, \hat{U}_{ijk}^{(b)} = U_{ijk}^{(b)} - \bar{U}^{(b)}$ and let n be the total number of sibpairs. The statistic is then defined by

$$T = (\bar{U}^{(b)} - \bar{U}^{(a)}) \sqrt{n(n-1) / \sum (U_{ijk}^{(b)} - U_{ijk}^{(a)})^2} \qquad (5)$$

and under the null hypothesis follows a t distribution with degrees of freedom n - 1. The one sided p-value is given by $P(t_{n-1} > T)$.

It is useful to examine this statistic under various situations. When the marker modeled is not associated with the phenotype, the allelic effect b is expected to be small and therefore the statistic is likely to be close to zero. However, when there is an association between the marker and the quantitative trait in a statistical sense, but it is not related to the detected linkage (for example it is due to the well-known bias from population stratification), we may not expect the allelic effect b to be small. In this scenario, we may look upon the marker as a covariate representing to some extent population stratification, and therefore modeling this marker would reduce the residual variance of the trait similarity measure coming from population stratification, and hence strengthen the linkage signal. So we can expect the statistic T to be more likely to be negative, and our test statistic would maintain the type I error rate in a conservative fashion in the case of population stratification. Our simulation results agree with this line of reasoning (see results). In this sense, a small lower sided p-value, i.e. $P(t_{n-1} < T)$, indicates a spurious association, which is also seen in the simulations.

For simplicity, assume the allelic effect b and the sibship mean are μ_k known and so can be specified correctly; it can then be easily shown that for sibpair (i,j) in family k, $E(U_{ijk}^{(b)} - U_{ijk}^{(a)}) = (x_{ik}x_{jk}b^2)IBD_{ijk}$ (see Appendix 2). This equation indicates that the proposed statistic essentially tests the correlation (or interaction) between the similarity of an associated marker effect, which is measured by a cross-product, and the IBD sharing between two sibs in a pair. Compared to a usual quantitative linkage analysis that detects linkage by testing the correlation between the IBD sharing and trait similarity, which may also be described as a cross-product (e.g. as in HE regressions and the variance component model), we can expect the proposed statistic to be much more powerful for detecting linkage because the noise (residual variances) from polygenic and common environmental effects is eliminated as well as the individual random effects. So, even if a usual linkage analysis fails to show signals in a region, the proposed statistic can still be useful to detect linkage when we have a candidate locus in a region.

Results

Sample selection

Because in practice the number of alleles shared IBD is generally not known with certainty, owing to partially informative markers and missing parental genotypes, we also performed computer simulations to examine the usefulness of the QLS in sample selection for an association study by comparing, in various situations, the statistics from random samples of unrelated individuals and from samples based on the rank order of the QLS. The statistic used to make the comparison is the score statistic proposed by Schaid et al. [19], which follows a χ^2 distribution with one degree of freedom for an additive model.

In our simulations, we generate 1000 sibships of size 2 from different subpopulations. A total of 6 markers, evenly space at a 2 cM density in a 10 cM range and each with 4 equally frequent alleles, are used for the linkage analysis. A QTL with 2 equally frequent alleles is located midway between marker 3 and marker 4. We assume Hardy-Weinberg equilibrium at each marker, linkage equilibrium among the markers and a Haldane no-interference map function. Trait values are constructed as the sum of a major-gene effect generated by the QTL, normal random individual effects, polygenic effects and common environmental effects. We calculate the probabilities of the number of alleles shared IBD using the program GENIBD in the S.A.G.E. package [20], removing the QTL genotype for this calculation.

We first compare random selection and the QLS selection with different sample sizes for the association study. We assume the population consists of two subpopulations, in equal proportions, from which 1,000 sibpairs have been used for the linkage analysis. In subpopulation 1, 20% of the total variance is explained by the QTL, 30% by the polygenic and common environmental effects and the rest by a random individual effect. In subpopulation 2, there is no QTL effect but the same other effects are simulated.

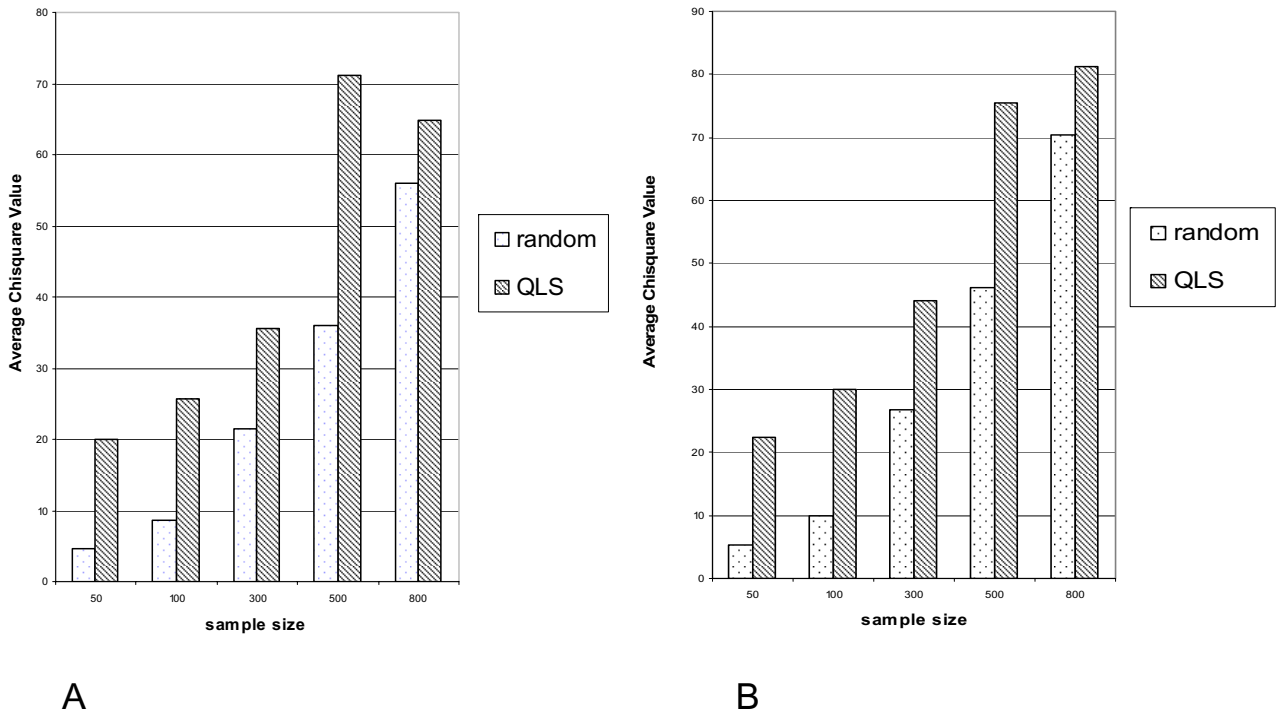

Figure 2
Comparison of average χ^2 values between random sampling and QLS sampling for various sample sizes. A. Two subpopulations: 20% of total variance is from an additive QTL in subpopulation 1, and no QTL effect exists in subpopulation 2. B. Four subpopulations with the QTL effect explaining 0%, 5%, 10%, and 20% of the total variance.

We separately sample 50, 100, 300, 500 and 800 unrelated individuals from the 1000 sibpairs by the two selection approaches and compare their score statistics. In QLS selection, we first select sibships with largest QLS and then randomly select one sib from each of these pairs, while in random selection the sibpair is selected randomly. The average χ^2 is shown in Figure 2(A). In real data, the situation may be more complex in that a population may consist of more than two subpopulations and the QTL effect could vary among subpopulations. We therefore also simulated four subpopulations with equal proportions having different QTL effects (0%, 5%, 10%, 20%) and compared the association statistics for different sample sizes. The results are shown in Figure 2(B). In both Scenarios (A and B), QLS selection can greatly increase the average value of the statistic to detect association, and this increase is larger when fewer unrelated individuals are selected.

To examine different ways of summarizing the several QLSs for a sibship, we also simulated sibships of different sizes, ranging from 2 to 4. The traits for the population with two subpopulations were simulated as before. We sampled 100 unrelated individuals from the 1000 sibships at random, or according to the rank order of the mean QLS, the minimum QLS and the maximum QLS of each sibship, respectively. Our results showed that the average χ^2 values obtained based on any of the QLSs are greater than those from random selection and that they have small differences between them ($\chi^2_{mean} > \chi^2_{max} > \chi^2_{min}$) (data not shown).

Although in this paper we focus on the usefulness of the QLS in the situation where a significant linkage region has already been identified, we are also interested in the situation where the linkage signal is not so clear, because in the case of a complex quantitative trait we expect only weak linkage signals when using customary sample sizes. To show the usefulness of QLS selection in this scenario, we also simulated 500 sibpairs from two subpopulations in which different proportions of the variance (5%, 10%, 15%, and 20%) are explained by the QTL in just one subpopulation. In this simulation, linkage signals are quite small and even cannot be detected. We sampled 100 unrelated individuals for an association study. The results

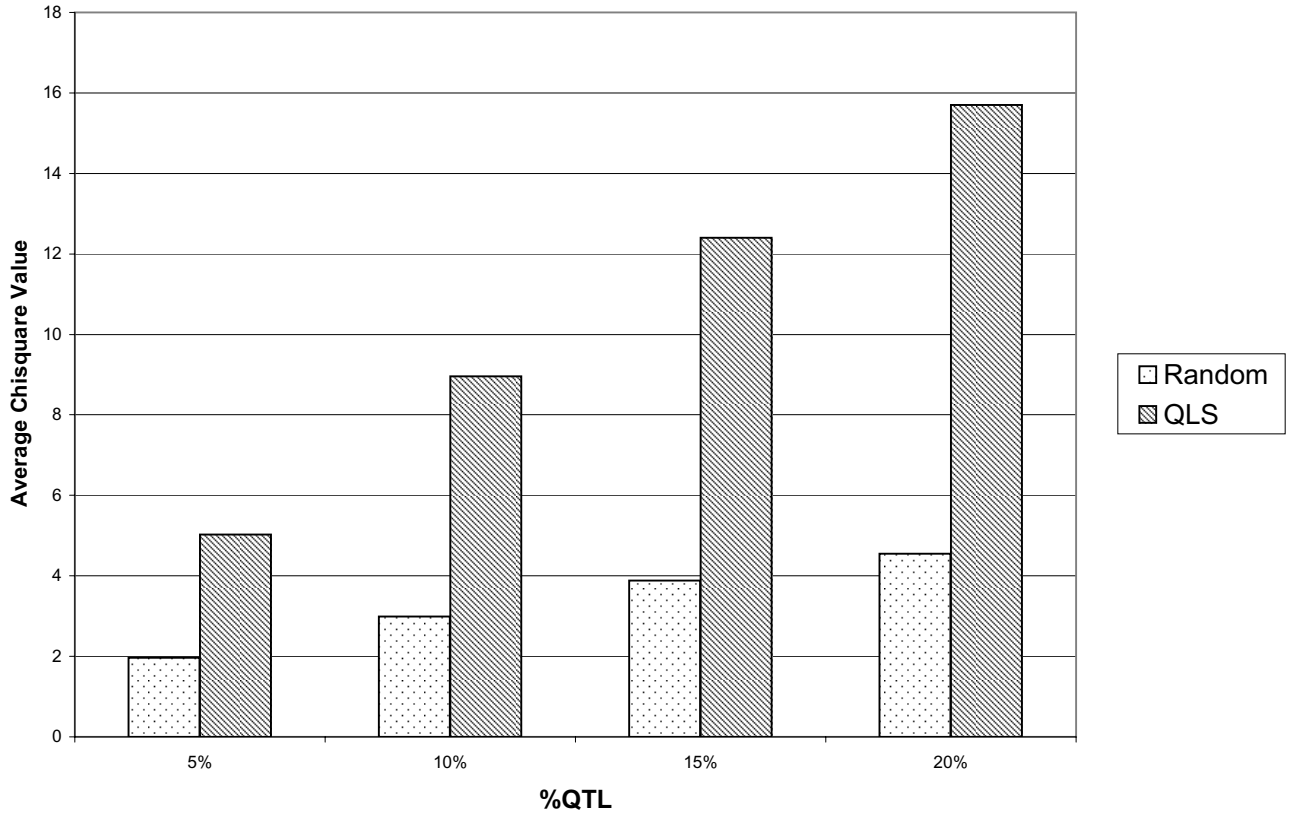

Figure 3
Comparison of average χ^2 values between random sampling and QLS sampling when the power to detect linkage is small.

show that the sampling based on the QLS still improves the power of an association study, even in the case that the power to detect linkage is negligible (see Figure 3).

At the stage of the association study, a family sample is also often used and then a joint linkage/association analysis can be applied in this case. One advantage of the joint linkage/assocation model is that, when it detects association, this method can simultaneously take account of the linkage information. We also performed a simulation study to examine the usefulness of QLS based sample selection in this case. A total of 500 nuclear families of size 4 from two subpopulations, in equal proportion, were generated for the previous linkage study. We further sampled 50, 100 and 150 families for fining mapping. Different QTL effects were simulated in subpopulation 1 (0%, 10%, 20%, 30% and 40%) and subpopulation 2 (0%). We compared the statistics of a commonly used joint linkage/association method (awbw) for a random sample and QLS based sample of families [21]. The results show the power of this joint analysis can also be greatly improved by the QLS selection approach (see Figure 4).

Testing the correlation between association and a previous linkage

To assess the properties of our tests to determine whether an association is responsible in part for the linkage of a complex quantitative trait, we carried out a limited simulation study. We examined the type I error rate of the proposed test under two scenarios: (1) no trait-marker association and (2) trait-marker association due to population stratification. Under no trait-marker association, we simulated 10,000 replicate data sets of 500 sibpairs or 500 sibships (200, 200 and 100 sibships of sizes 2, 3 and 4, respectively). Trait values were constructed as the sum of a major-gene effect generated by the QTL that explains 10% of the variance, and various proportions of random individual, polygenic and common environmental effects. An association marker with two equally frequent alleles was simulated to be in complete linkage equilib-

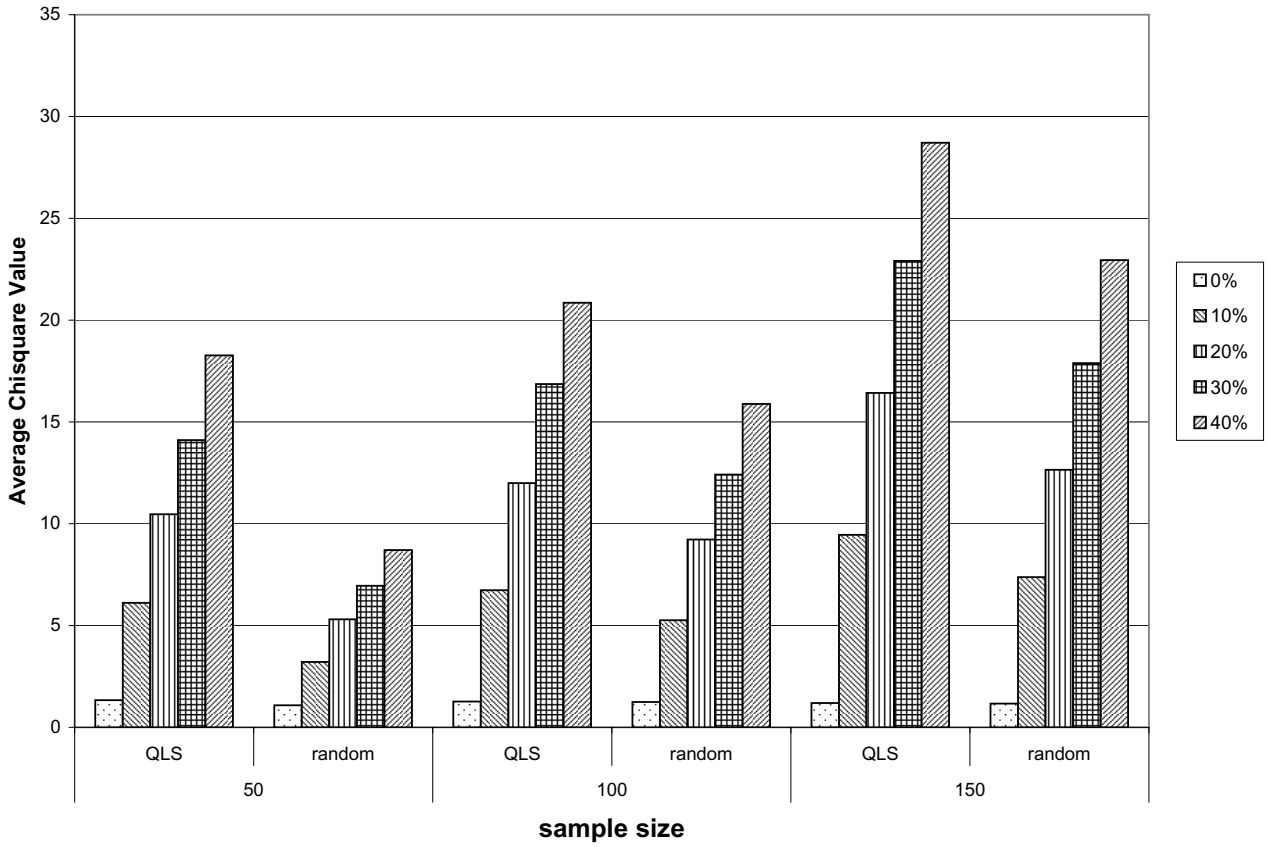

Figure 4
Comparison of average χ^2 values between random sampling and QLS sampling for a joint linkage/association method. Two subpopulations: 0%, 10%, 20%, 30% and 40% of total variance is from an additive QTL in subpopulation 1, and no QTL effect exists in subpopulation 2.

rium with the QTL and a fully informative linkage marker (with 100 equal frequent alleles) was also simulated at the same location. For the case of linkage but no trait-marker association, the results show that the type I error rate of the proposed statistic is generally good for a complex quantitative trait for both sibpair data and sibship data (see Table 1). Under a spurious association, we generated 10,000 replicate datasets of 500 sibpairs from two subpopulations. The trait mean and the frequencies of the marker alleles were different in the two subpopulations. The results (see Table 2) show that the power to detect linkage is consistent in various situations suggesting that the linkage test is quite robust to population stratification (the"linkage" column). For the proposed test examining the linkage-association correlation, the type I error rate is controlled conservatively (the "association-linakge" colmun). When the effect of population stratification is small, the empirical type I error rate is close to the correct level (0.05). We also examined the usefulness of the proposed statistic for detecting spurious association due to population stratification by using the lower sided t-test (shown in the "population stratification" column of Table 2). The results suggest that in practice when the association cannot explain any of the linkage, this statistic may nevertheless be useful to determine whether the association is "false".

We also performed simulations to assess the power of the proposed statistic to detect the correlation between the gene effect and IBD sharing. We compared this statistic with a revised HE regression that we have shown is one of the most powerful versions of HE [11]. An associated marker with two equal allele frequencies was simulated as the QTL itself. We generated trait values with various different QTL effects, keeping fixed polygenic, common environmental effects and individual random effects. We

Table 1: Empirical type I error rate of the proposed test at the nominal 5% level. A diallelic marker is completely linked to the QTL under HW equilibrium.

Effects [1]	500 sibpairs		500 sibships [2]	
	Linkage	Association-Linkage	Linkage	Association-Linkage
10%/80%	0.195	0.056	0.391	0.059
20%/70%	0.211	0.059	0.310	0.057
30%/60%	0.236	0.055	0.328	0.055
40%/50%	0.248	0.054	0.346	0.058
50%/40%	0.270	0.056	0.368	0.056

[1] Common environmental & polygenic effects/individual random effects
[2]There are 200 sibships of size 2, 200 sibships of size 3 and 100 sibships of size 4.

considered two sets of linkage markers: fully informative and partially informative (six markers were used for the linkage analysis, evenly spaced at a 2 cM density in a 10 cM range around the QTL). For each situation, we generated 1,000 replicate samples of data on 500 sibpairs. Table 3 shows that by incorporating information on the candidate marker the proposed test is much more powerful than quantitative linkage analysis. In general, for a complex quantitative trait, usual linkage analysis may lack power and therefore miss an important region, because the noise from other genetic and environmental effects masks the linked gene effect. When no linkage is detected in a region where an important candidate gene is located, it is not wise to discard this region from further study. We may use the proposed statistic to assess whether the "negative" linkage result is true.

Discussion

There is great interest in QTL mapping because many important diseases themselves, or intermediate phenotypes, are measured on a continuous scale. Although trait-marker association studies are expected to be soon conducted genome-wide, because of cost considerations currently an association study often focuses on candidate regions determined by a previous linkage study. For such an association study, we should utilize the information available in the previous linkage study to optimize its design and to facilitate its interpretation. We have proposed a quantitative linkage score, based on the widely used HE regression, to provide quantitative linkage information useful for a follow-up association study. This score is not limited to continuous traits, but can also be used for binary (affected/unaffected) traits. We illustrated the usefulness of this score to answer two different questions posed by an association study: (1) how to select samples at the design stage when heterogeneity exists; and (2) how to test at the inference stage whether an observed association can explain in part a previous linkage signal. In this paper, we are not necessarily advocating a two-stage approach to analyze family data on which we have information on both linkage markers and association markers. For such data a joint linkage and association framework could be of more interest than a two-stage analysis approach. Recent work on this kind of joint analysis has included work on both regression-based methods [22] and variance-component methods [23,24]. However, in the presence of heterogeneity any advantage such a joint analysis may have when performed using all the data available may be lost, because those families that are not affected because of segregation at a linked locus will "dilute" the effect and result in loss of power. Therefore, even for analyzing data with information from both link-

Table 2: Empirical type I error rate of the proposed test at the nominal 5% level when population stratification exists. A diallelic marker is completely linked to the QTL with HW equilibrium in an admixed population. Total 500 sibpairs (250/250) are selected from two subpopulations. p_1 and p_1 are frequencies of the rarer marker allele in the two subpopulations. *d* is the difference in trait means between two subpopulations.

p_1 - p_2	*d* = 10			*d* = 20		
	Linkage	Association-Linkage	Population stratification	Linkage	Association-Linkage	Population stratification
0.4	0.200	0.00055	0.403	0.165	0	0.758
0.3	0.191	0.0016	0.311	0.166	0	0.536
0.2	0.200	0.0054	0.177	0.159	0.001	0.283
0.1	0.192	0.025	0.084	0.154	0.012	0.106
0	0.194	0.050	0.019	0.150	0.047	0.013

Table 3: The power of the proposed test for 500 sibpairs when linkage signal is weak. A diallelic marker is completely linked to the QTL in perfect disequilibrium. The trait value is generated by a QTL with variance of varying size, together with polygenic and common environmental effects (with variance 0.3) and a random individual effect (with variance 0.5).

QTL variance	Fully informative marker		Partially informative marker	
	Linkage	Association-Linkage	Linkage	Association-Linkage
0.03	0.102	0.307	0.083	0.285
0.05	0.148	0.402	0.145	0.475
0.10	0.284	0.666	0.260	0.575
0.20	0.563	0.874	0.520	0.838

age markers and association markers, we may consider first selecting families based on the QLS to exclude such "dilution" as much as possible.

The idea of selecting families with linkage evidence for further genotyping in a follow-up association study is not new and has been successfully implemented in practice. In the context of quantitative traits, the proposed score can conveniently be used to summarize quantitative linkage information from a sibpair (or sibship). We have shown that in a heterogeneous population, which is expected to commonly occur for a complex trait, selecting a sample of unrelated persons based on the order of the QLS magnitude results in a more homogeneous sample for an association study than does a random sample, and therefore can improve power for a given sample size. Other approaches to identifying sibpairs with linkage are available, for example using a regression diagnostic [25]. Careful comparison of these methods would merit further study.

Another use of the QLS investigated in this paper is to test whether association can account in part for a detected linkage. To address this question, we simply compare two sets of QLSs, before and after incorporating an association marker into the individual level regression model. Essentially, the proposed test evaluates the interaction of the allele effect of an associated marker and IBD sharing. In this sense it may be likened to other methods, for example the regression model proposed by Cardon [26], though our statistic emphasizes more whether an association is correlated with a previous linkage finding. This test may also be used as a substitute for the usual quantitative trait linkage analysis test when the latter fails to detect linkage. The gain in power to detect linkage by using the proposed test arises from eliminating possible environmental or other genetic noise. However, this gain is not automatic, but depends on the relationship of the associated marker to the true variant. If there is only weak linkage disequilibrium between an associated marker and the true variant, the test will be less powerful. We also showed that this statistic may be applied to detect spurious association, although that was not our primary aim. The ways commonly used in practice to detect population stratification are to use genomic control [27] or test for Hardy-Weinberg equilibrium [28]. Using IBD sharing information to test and control for population stratification provides a new approach and further study of this approach will be conducted in our future work.

Conclusion

In conclusion, as proved by our simulations, the QLS is useful for the design of, and resulting inference from, an association study following a linkage study. We suggest that careful examination of the QLS should be helpful for understanding the results of both association and linkage studies.

Authors' contributions

TW contributed to the conception of the study, performed the simulation analysis, and wrote the manuscript. RCE contributed to the conception of the study and the writing of the manuscript.

Appendix

Appendix 1 the derivation of q_{qls}

For sibpair k comprising sib 1 and sib 2, $Z_k(z_{1k}, z_{2k})$ follows the distribution $f(z_{1k}, z_{2k})$, which we assume to be a bivariate normal distribution. With the assumption that a random sample of full sibpairs is used for the linkage analysis, the proportions of pairs for which the number of alleles shared IBD is 0, 1 and 2 are $\pi_0 = 1/4$, $\pi_1 = 1/2$ and $\pi_2 = 1/4$, respectively. Let P1 and P2 refer to subpopulation 1 and 2 and let their proportions be denoted q_1 and q_2. We then have

$$\begin{aligned} Pr(QLS > 0 \mid P1) &= 2Pr(z_{1k} > 0, z_{2k} > 0, IBD = 1 \mid P1) + 2Pr(z_{1k} > 0, z_{2k} < 0, IBD = 0 \mid P1) \\ &= 2\pi_2 \int_0^\infty \int_0^\infty f(z_{1k}, z_{2k} \mid IBD = 1, P1) d_{z_{1k}} d_{z_{2k}} + 2\pi_0 \int_0^\infty \int_{-\infty}^0 f(z_{1k}, z_{2k} \mid IBD = 0, P1) d_{z_{1k}} d_{z_{2k}} \\ &= \frac{1}{4}\left[\frac{1}{\pi} \arctan\left(\frac{\rho^2_{IBD=1}}{1-\rho^2_{IBD=1}}\right) + 1\right]. \end{aligned}$$

Thus,

$$\begin{aligned} Pr(QLS > 0, P1) &= Pr(QlS > 0 \mid P1)Pr(P1) \\ &= \frac{1}{4}q_1\left[\frac{1}{\pi}\arctan\left(\frac{\rho^2_{IBD=1}}{1-\rho^2_{IBD=1}}\right)+1\right], \end{aligned}$$

which is an increasing function of q_1 and ρ. We note that ρ depends on the size of the effect and allelic frequencies of the QTL. On the other hand,

$$\begin{aligned} Pr(QLS > 0 \mid P2) &= 2Pr(z_{1k} > 0, z_{2k} > 0, IBD = 1 \mid P2) + 2Pr(z_{1k} > 0, z_{2k} < 0, IBD = 0 \mid P2) \\ &= \frac{1}{4}, \end{aligned}$$

so that $Pr(QLS > 0, P2) = Pr(QlS > 0|P2)Pr(P2) = \frac{1}{4}q_2$.

Thus

$$q_{qls} = \frac{Pr(QLS > 0, P1)}{Pr(QLS > 0, P1) + Pr(QLS > 0, P2)}$$

$$= \frac{1}{1+\frac{q_2}{\frac{1}{\pi}\arctan\left(\frac{\rho^2_{IBD=1}}{1-\rho^2_{IBD=1}}+1\right)}}.$$

Appendix 2 - $E(U^{(b)} - U^{(a)})$

Under the trait model $\gamma_{ik} = \mu_k + x_{ik}b + e_{ik}$, we assume the e_{ik} are identically and independently distributed with mean 0. Suppose μ_k and b are known. Let the subscripts 1 and 2 indicate the two sibs of a sibpair in family k. Then

$$\begin{aligned} E(U_k^{(b)} - U_k^{(a)} &= E[(\gamma_{1k} - \mu_k)(\gamma_{2k} - \mu_k) - (\gamma_{1k} - \mu_k - x_{1k}b)(\gamma_{2k} - \mu_k - x_{2k}b)]IBD_{12k} \\ &= E[(\gamma_{1k} - \mu_k)x_{2k}b + (\gamma_{2k} - \mu_k)x_{1k}b - x_{1k}x_{2k}b^2]IBD_{12k} \\ &= E[x_{1k}x_{2k}b^2 + e_{1k}x_{2k}b + e_{2k}x_{1k}b]IBD_{12k} \\ &= (x_{1k}x_{2k}b^2)IBD_{12k} \end{aligned}$$

Acknowledgements

This work was supported in part by a U.S. Public Health Service Resource Grant from the National Center for Research Resources (RR03655) and a Research Grant from the National Institute of General Medical Sciences (GM28356). Tao Wang is supported by a fellowship from the Merck Foundation.

References

1. Haseman JK, Elston RC: **The investigation of linkage between a quantitative trait and a marker locus.** *Behavior Genet* 1972, **2:**3-19.
2. Horikawa Y, Oda N, Cox NJ, Li X, Orho-Melander M, Hara M, Hinokio Y, Lindner TH, Mashima H, Schwarz PE, del Bosque-Plata L, Horikawa Y, Oda Y, Yoshiuchi I, Colilla S, Polonsky KS, Wei S, Concannon P, Iwasaki N, Schulze J, Baier LJ, Bogardus C, Groop L, Boerwinkle E, Hanis CL, Bell GI: **Genetic variation in the gene encoding calpain-10 is associated with type 2 diabetes mellitus.** *Nat Genet* 2000, **26:**163-175.
3. Van Eerdewegh P, Little RD, Dupuis J, Del Mastro RG, Falls K, Simon J, Torrey D, Pandit S, McKenny J, Braunschweiger K, Walsh A, Liu Z, Hayward B, Folz C, Manning SP, Bawa A, Saracino L, Thackston M, Benchekroun Y, Capparell N, Wang M, Adair R, Feng Y, Dubois J, FitzGerald MG, Huang H, Gibson R, Allen KM, Pedan A, Danzig MR, Umland SP, Egan RW, Cuss FM, Rorke S, Clough JB, Holloway JW, Holgate ST, Keith TP: **Association of the ADAM33 gene with asthma and bronchial hyperresponsiveness.** *Nature* 2002, **418:**426-430.
4. Kim UK, Jorgenson E, Coon H, Leppert M, Risch N, Drayna D: **Positional cloning of the human quantitative trait locus underlying taste sensitivity to phenylthiocarbamide.** *Science* 2003, **299:**1221-1225.
5. Fingerlin TE, Boehnke M, Abecasis GR: **Increasing the power and efficiency of disease-marker case-control association studies through use of allele-sharing information.** *Am J Hum Genet* 2004, **74:**432-443.
6. Li C, Scott LJ, Boehnke M: **Assessing whether an allele can account in part for a linkage signal: the genotype-IBD sharing test (GIST).** *Am J Hum Genet* 2004, **74:**418-431.
7. Fulker DW, Cherny SS, Sham PC, Hewitt JK: **Combined linkage and association sib-pair analysis for quantitative traits.** *Am J Hum Genet* 1999, **64:**259-267.
8. Cardon LR, Abecasis GR: **Some properties of a variance components model for fine-mapping quantitative trait loci.** *Behavior Genetics* 2000, **30:**235-243.
9. Amos CI, Dawson DV, Elston RC: **The probabilistic determination of identity-by-descent sharing.** *Am J Hum Genet* 1990, **47:**842-853.
10. Lander ES, Green P: **Construction of multilocus genetic linkage maps in human.** *Proceedings of the National Academy of Science of the Unite States of America* 1987, **84:**2363-2367.
11. Wang T, Elston RC: **A Modified Revisited Haseman-Elston Method to Further Improve Power.** *Hum Hered* 2004, **57:**109-116.
12. Tritchler D, Liu Y, Fallah S: **A test of linkage for complex discrete and continuous traits in nuclear families.** *Biometrics* 2003, **59:**382-392.
13. Whittemore AS, Tu I: **Simple, robust linkage tests for affected sibs.** *Am J Hum Gene* 1998, **62:**1228-1242.
14. Wright FA: **The phenotypic difference discards sib-pair QTL linkage information.** *Am J Hum Genet* 1997, **60:**740-742.
15. Elston RC, Buxbaum S, Jacobs KB, Olson JM: **Haseman and Elston revisited.** *Genet Epidemiol* 2000, **19:**1-17.
16. Shete S, Jacobs KB, Elston RC: **Adding further power to the Haseman and Elston method for detecting linkage in larger sibships: Weighting sums and differences.** *Hum Hered* 2003, **55:**79-85.
17. Kruglyak L, Daly MJ, Reeve-Daly MP, Lander ES: **Parametric and nonparametric linkage analysis: a unified multipoint approach.** *Am J Hum Genet* 1996, **58:**1347-1363.
18. Blackwelder WC, Elston RC: **A comparison of sib-pair linkage tests for disease susceptibility loci.** *Genet Epidemiol* 1985, **2:**85-97.
19. Schaid DJ, Rowland CM, Tines DE, Jacobson RM, Poland GA: **Score tests for association between traits and haplotypes when linkage phase is ambiguous.** *Am J Hum Genet* 2002, **70:**425-434.
20. **Statistical Analysis for Genetic Epidemiology** [http://darwin.cwru.edu/sage/]
21. Abecasis GR, Cardon LR, Cookson WO: **A general test of association for quantitative traits in nuclear families.** *Am J Hum Genet* 2000, **66:**279-292.
22. Wang T, Elston RC: **Two-level Haseman-Elston regression for general pedigree data analysis.** *Genet Epidemiol* 2005, **29:**12-22.
23. Fan R, Spinka C, Jin L, Jung J: **Pedigree linkage disequilibrium mapping of quantitative trait loci.** *Eur J Hum Genet* 2005, **13:**216-231.
24. Jung J, Fan R, Jin L: **Combined linkage and association mapping of quantitative trait loci by multiple markers.** *Genetics* 2005, **170:**881-898.
25. Davis CC, Brown WM, Lange EM, Rich SS, Langefeld CD: **Nonparametric linkage regression II: Identification of influential pedigrees in tests for linkage.** *Genet Epidemiol* 2001, **21(Suppl 1):**S123-S129.
26. Cardon LR: **A sib-pair regression model of linkage disequilibrium for quantitative traits.** *Hum Hered* 2000, **50:**350-358.
27. Devlin B, Roeder K: **Genomic control for association studies.** *Biometrics* 1999, **55:**997-1004.

28. Tiret L, Cambien F: **Letter: Departure from Hardy-Weinberg equilibrium should be systematically tested in studies of association between genetic markers and disease.** *Circulation* 1995, **92:**3364-3365.