Extensions to a Score Test for Genetic Linkage with Identity by Descent Data

Sandrine Dudoit and Darlene R. Goldstein

Abstract

Genetic analysis aims to determine which underlying genes affect traits, their chromosomal locations and variants, and, ultimately, their modes of action at the biochemical level. Linkage analysis is an initial step in elucidating the genetic mechanisms affecting a trait of interest. This paper reviews genetic linkage analysis, with an emphasis on the score test approach developed by Dudoit and Speed [8, 10]. Two extensions of the test under current investigation are also presented: use of the test with larger sets of relatives than pairs, and generalization to allow for missing DNA identity by descent (IBD) information.

Keywords: allele sharing; complex traits; identity by descent; linkage analysis; pedigree; score test

1 Introduction

A central problem in genetic analysis is to determine which gene(s), if any, affect particular phenotypes, the chromosomal locations of these genes, their different alleles and, ultimately, their biochemical modes of action. Linkage analysis is an initial step in elucidating the genetic mechanisms affecting a trait of interest. Its goal is to determine the chromosomal location of the gene(s) influencing the trait. Linkage analysis proceeds by tracking patterns of coinheritance of the trait of interest and other traits or genetic markers, relying on the varying degree of recombination between trait and marker loci to map the loci relative to one another.

Mendel's second law of inheritance hypothesizes that different "factors" (traits or genes) segregate to gametes (sperm or egg) independently. Actually, independent assortment of gene pairs only occurs when the genes are on different chromosomes or are so far apart on the same chromosome that there is the same chance of recombination as nonrecombination. Such pairs of genes are said to be *unlinked*. Two genes are *linked* when they do not segregate independently. A measure of the degree of linkage is the *recombination fraction*, the chance of recombination occurring between two loci, denoted almost universally in the genetics literature as θ. For unlinked genes, $\theta = 1/2$; for linked genes, $0 \leqslant \theta < 1/2$. The following gives a brief introduction to linkage analysis; more substantial detail is provided by Ott [30], McPeek [27], and Speed [32].

Data for linkage analysis consist of sets of related individuals (*pedigrees*) and information on the genetic marker and/or trait *genotypes* (the two alleles at a locus) or *phenotypes* (the outward manifestation of a trait), usually selected on the basis of phenotype (*e.g.* a disease, such as diabetes, or a quantitative trait, such as glucose tolerance). For this setup, the recombination fraction is most commonly estimated by the method of maximum likelihood, the likelihood being determined by an appropriate genetic model for the coinheritance of the loci. The conventional measure of support for the hypothesis of linkage between two loci at recombination fraction θ versus that of no linkage is given by the *lod score*

$$Z(\theta) = log_{10}\left[\frac{L(\theta)}{L(1/2)}\right],$$

where $L(\theta) \propto f(\underline{X} \mid \theta)$ denotes the likelihood for θ given the observed data $\underline{X}$. Positive values of Z are evidence of linkage, while negative values indicate no linkage. With lod score linkage analysis, the null hypothesis of no linkage ($H_0 : \theta = 1/2$) is rejected for sufficiently large values of $Z(\hat{\theta}_{MLE})$, often taken to be 3. Linkage analysis based on the lod score is referred to in the genetics literature as "parametric" or "model-based" linkage analysis, as the mode of inheritance must be specified using some parametric model.

Genetic linkage mapping has been successful at mapping genes for traits following Mendelian inheritance patterns, typically recessive or dominant diseases. Identifying genes affecting complex traits, or traits not following these simple modes of inheritance, has proven to be more challenging. Lod-score linkage analysis for complex traits is difficult to carry out due to many complicating factors. Chief among these is that the mode of inheritance is rarely known. "Nonparametric," or "model-free," approaches thus have appeal, since they do not require a genetic inheritance model to be specified. Such methods usually focus on identical by descent (IBD) allele sharing at a locus between a pair of relatives. DNA at a locus is shared by two relatives *identical by descent* if it originated from the same ancestral chromosome. In families of individuals possessing the trait of interest, there is association between allele sharing at loci linked to trait susceptibility loci and the trait (see *e.g.* Dudoit and Speed [9] for examples). This association may be used to localize trait susceptibility genes. For loci unlinked to trait susceptibility loci, IBD sharing of DNA is not associated with the occurrence of the trait. Early work on linkage analysis using IBD data from sib-pairs can be found in Day and Simons [6] for qualitative traits, and in Haseman and Elston [20] for quantitative traits.

Testing for linkage with IBD data has developed along different lines, depending on the type of trait. For qualitative traits, the test is based on *IBD sharing conditional on phenotypes*, *e.g.* affected sib-pair methods (see [21] for a review). On the other hand, for quantitative trait loci (QTL), linkage analysis is based on examination of *phenotypes conditional on sharing*. A very widely used procedure in QTL mapping in humans is the Haseman-Elston method [20], implemented for sib-pairs and other relative pairs in the SIBPAL and RELPAL programs of the computer package S.A.G.E. [11]; many extensions of it are also available [1, 2, 3, 12, 16, 17, 24, 28, 29]. In this method, the squared difference in phenotype values for the two relatives is regressed on the (estimated) proportion of alleles they share IBD. The method can also be used with qualitative traits (binary coding), but is clearly not appropriate for analysis of relatives where the phenotypic difference is fixed by design (*e.g.* affected sib-pairs). A disadvantage of the standard Haseman-Elston method is that it uses only differences in the phenotypes rather the full joint phenotypic data, incurring possible information loss [36].

The pattern of IBD sharing at a locus within a pedigree is summarized by an *inheritance vector,* which completely specifies the ancestral source of DNA [25]. For sibships of size k, it is convenient to label paternally derived alleles at the locus (1, 2) and maternally derived alleles (3, 4). The inheritance vector at a given locus is the vector $x = (x_1, x_2, ..., x_{2k-1}, x_{2k})$, where for sib i, x_{2i-1} is the label of the paternally inherited allele (1 or 2) and x_{2i} is that of the maternally inherited allele (3 or 4) at the locus. Note that the labels 1, 2, 3, and 4 for the parental DNA only have meaning within a sibship, and may therefore correspond to different sequences of DNA in different sibships.

Inheritance vectors for sibships may be grouped into IBD configurations which can be thought of as orbits of groups acting on the set of possible inheritance vectors (Dudoit and Speed [8], Ethier and Hodge [13]). For a pair of sibs, when paternal and maternal allele sharing are not distinguished, the 16 possible inheritance vectors give rise to three IBD configurations C_j: the sibs may share 0, 1, or 2 alleles IBD at the locus (Table 1). In the case of *affected* sib-trios, that is, all three sibs are affected with the trait under study, there are four IBD configurations (Table 2); in the case of a quantitative trait on sib-trios, the number of IBD configurations is 10 (Table 3).

Table 1: Sib-pair IBD configurations

Alleles IBD	Inheritance vectors	$\mid C_j \mid$
0 IBD	(1, 3, 2, 4), (1, 4, 2, 3), (2, 3, 1, 4), (2, 4, 1, 3)	4
1 IBD	(1, 3, 1, 4), (1, 4, 1, 3), (2, 3, 2, 4), (2, 4, 2, 3) (1, 3, 2, 3), (1, 4, 2, 4), (2, 3, 1, 3), (2, 4, 1, 4)	8
2 IBD	(1, 3, 1, 3), (1, 4, 1, 4), (2, 3, 2, 3), (2, 4, 2, 4)	4

2 Score Test for Linkage

2.1 General Form of the Score Test

Dudoit and Speed [8, 10] proposed a score test to detect linkage with IBD data on sets of relatives. This approach represents a unified likelihood-based approach to the linkage

Table 2: Affected sib-trio IBD configurations

IBD configuration C_j	Pair-wise IBD sharing[a]	Representative inheritance vector	$\mid C_j \mid$
1	2, 2, 2	(1, 3, 1, 3, 1, 3)	4
2	2, 1, 1	(1, 3, 1, 3, 1, 4)	24
3	1, 1, 0	(1, 3, 1, 4, 2, 3)	24
4	2, 0, 0	(1, 3, 1, 3, 2, 4)	12

[a] Number of alleles shared IBD between sibs 1 and 2, 1 and 3, 2 and 3, respectively for the representative vector; this order may not be the same for each vector in the configuration

Table 3: Sib-trio IBD configurations for quantitative traits

IBD configuration C_j	Pair-wise IBD sharing[a]	Representative inheritance vector	$\mid C_j \mid$
1	2, 2, 2	(1, 3, 1, 3, 1, 3)	4
2	2, 1, 1	(1, 3, 1, 3, 1, 4)	8
3	2, 0, 0	(1, 3, 1, 3, 2, 4)	4
4	1, 1, 0	(1, 3, 1, 4, 2, 3)	8
5	1, 0, 1	(1, 3, 1, 4, 2, 4)	8
6	1, 1, 2	(1, 3, 1, 4, 1, 4)	8
7	0, 0, 2	(1, 3, 2, 4, 2, 4)	4
8	0, 2, 0	(1, 3, 2, 4, 1, 3)	4
9	1, 2, 1	(1, 3, 1, 4, 1, 3)	8
10	0, 1, 1	(1, 3, 2, 4, 1, 4)	8

[a] Number of alleles shared IBD between sibs 1 and 2, 1 and 3, 2 and 3, respectively

analysis of qualitative and quantitative traits using IBD data on pedigrees. The likelihood for the recombination fraction θ, conditional on the phenotypes of the relatives, is used to form a score test of the null hypothesis of no linkage ($\theta = 1/2$).

The probability vector of IBD configurations, conditional on pedigree phenotypes, at a *marker* locus linked to a trait susceptibility locus at recombination fraction θ can be written as

$$\rho(\theta,\pi)_{1\times m} = \pi_{1\times m} T(\theta)_{m\times m},$$

where π represents the conditional probability vector for IBD configurations at the *trait* locus and the number of IBD configurations is m. $T(\theta)$ denotes the transition matrix between IBD configurations at loci separated by recombination fraction θ, and has infinitesimal generator Q. The probability vector π will in general depend on (possibly very many) unknown genetic parameters. Under the null hypothesis that the marker and trait susceptibility loci are unlinked, the IBD sharing distribution at the marker is given by the stationary distribution of $T(\theta)$, which is

$$\alpha = \rho(1/2,\pi) = \frac{1}{K}(|\, C_1 \,|, \ldots, |\, C_m \,|),$$

where $|\, C_j \,|$ is the number of inheritance vectors in IBD configuration C_j and K is the number of inheritance vectors. For general pedigrees, K is 2 raised to the number of relevant meioses, *e. g.*, for sib-pairs, $K = 2^4$.

For a given pedigree type, the form of the score test statistic is determined by the second largest eigenvalue λ_2 and corresponding eigenvector(s) of Q. The eigenvalues and their multiplicities give information regarding the form the score statistic takes. The eigenvalues are negative even integers. If $\lambda_2 = -2\kappa$, the score test is based on the κ^{th} derivative of the log-likelihood. If λ_2 has multiplicity 1, then the score statistic is independent of the genetic model for the trait. In sibships, $\lambda_2 = -4$, with multiplicity depending on the group that defines the IBD configurations (Dudoit and Speed [8]).

For sibships, because the first derivative in the Taylor series expansion of the log-likelihood about the null value $\theta = 1/2$ is 0, the score statistic is based on the second derivative $T''(1/2) = 8P_{-4}$, where P_{-4} is the projection matrix for the eigenvalue -4 and having rank the multiplicity of -4.

The score test approach is motivated by a large number of advantages, including: it is locally most powerful for alternatives close to the null; unlike a number of tests for linkage, the score test does not depend on assumptions such as population genotypes being in Hardy-Weinberg equilibrium – any genotype distribution can be used; conditioning on phenotypes eliminates selection bias introduced by nonrandom ascertainment, which is how samples are commonly obtained in practice; combining differently ascertained pairs is straightforward, which is important because otherwise some portion of the data may not be used. And as is seen below, the power and apparent robustness properties make the test an attractive alternative to nonparametric tests.

2.2 Score Test for Pairs of Relatives

The linkage information from IBD and phenotype data on n sib-pairs is combined into the score statistic

$$S_{sib}(\nu) = 16\sum_{i=1}^{n}(\pi_{2i} - \pi_{0i})(N_{2i} - N_{0i}),$$

where $\pi_{ji} = \pi_j(\phi_{1i}, \phi_{2i}; \nu)$ is the conditional probability, given phenotypes (ϕ_{1i}, ϕ_{2i}) and genetic model parameters ν, that sib-pair i shares j alleles IBD (j = 0, 1, 2) at the *trait* locus (which could be one of several unlinked loci contributing to the trait); N_{ji} is 1 if sib-pair i shares j alleles IBD at the *marker* locus and 0 otherwise; and the sum is over all sib-pairs in the sample. The null IBD distribution at the marker is (1/4, 1/2, 1/4) for sharing (0, 1, 2) alleles.

For half-sib, avuncular, and grandparental pairs, the form of the test statistics is

$$S_{rel}(\nu) \propto \sum_{i=1}^{n} (\pi_{1i} - \pi_{0i})(N_{1i} - N_{0i});$$

the constant of proportionality and its sign differs for these relative types, but cancels in standardization. For these relative pairs, the null IBD distribution at the marker is (1/2, 1/2) for sharing (0, 1) alleles. For pairs of cousins, the score test statistic is given by

$$S_{cousin}(\nu) = 12 \sum_{i=1}^{n} (\pi_{1i} - \pi_{0i}/3)(N_{1i} - N_{0i}/3).$$

In this case, the null IBD sharing probabilities are (3/4, 1/4) for (0, 1) alleles.

For these types of relative pairs, the form of the score test statistic is fairly simple and readily interpretable. The statistic can be viewed as a weighted combination of IBD scores for each pair type, where the weights are given by differences in sharing probabilities conditional on phenotypes. For qualitative traits in sib-pairs, the weights depend on the genetic model but are constant in the phenotype and hence factor out. Thus, no genetic model is required. In general, however, this is not the case and a genetic model must be assumed in order to compute the weights.

The power and robustness properties of the score test were extensively studied via simulation of sib-pair and general relative pair data on a quantitative trait (Goldstein, Dudoit and Speed [18, 19]). For these studies, data were generated under a biallelic major gene model for the quantitative trait ϕ consisting of a single gene effect g with residual variation e, so that $\phi = \mu + g + e$, with μ the overall mean. Genotypic effect values are $g = a$ (> 0) for an A_1A_1 individual, $g = d$ for an A_1A_2 individual, and $g = -a$ for an A_2A_2 individual (see *e.g.* Falconer and Mackay [14]). The error term e has mean 0 and variance σ_e^2, constant across genotypes. The joint distribution of the error terms for a pair of relatives was assumed to be bivariate normal, with correlation ρ. Thus, in the population the trait is distributed as a mixture of bivariate normals, with mixing probabilities equal to the genotype frequencies. The *heritability* of a trait due to the genetic locus is the proportion of genetic variance to total variance: $H = \sigma_g^2/(\sigma_g^2 + \sigma_e^2)$. The parameters d, p, ρ, and H were varied, along with the selection strategy used to obtain pairs. Each simulated data set was analyzed with every model under consideration (one correct, the others wrong). This set of models was chosen as it is widely used in simulation studies of methods for analyzing quantitative traits.

In many realistic simulation scenarios, the score test approach showed large power gains over commonly used nonparametric tests, even when the assumed model for analysis deviated greatly from the true generating model. Based on the simulations, a generic additive model was recommended when little is known about the true underlying model.

Although the focus here has been on pairs of relatives, Dudoit [7] showed that the same score test approach is more generally applicable to any set of relatives. In practice, families included in studies of genetic traits often consist of more than a single pair of relatives. In addition, these simulation studies, like most, have considered only the case when complete IBD information is available. Yet realistically, the genetic information necessary to determine IBD status may be incomplete for some individuals. Thus, further investigation of test properties and feasibility of implementation for these situations is warranted.

3 Some Extensions of the Score Test

The score test approach is quite general, and its implementation for pairs of relatives may be generalized in a number of ways. For example, the model for phenotypes may be expanded to include covariates. We consider here a few other extensions that we are currently researching: first, derivation and implementation of the test for larger pedigrees and an examination of test feasibility and properties in this case; second, modification of the test to accomodate data with incomplete IBD information.

3.1 Score Test for Sib Trios

The next largest pedigree to consider, after small "pedigrees" of pairs of individuals, would contain three individuals. Here we consider sib-trios, in both the case of a qualitative trait and for a quantitative trait.

For a qualitative trait, sib-trios may have the same trait values, as do affected sib-trios (ASTs), or they may instead be discordant (DSTs), where one has a different value than the other two. The IBD configurations for ASTs are given in Table 3. The infinitesimal generator for the IBD configuration transition matrix $T(\theta)$ is

$$Q_{AST} = \begin{bmatrix} -6 & 6 & 0 & 0 \\ 1 & -4 & 2 & 1 \\ 0 & 2 & -4 & 2 \\ 0 & 2 & 4 & -6 \end{bmatrix},$$

which has eigenvalues $\lambda = 0, -4, -8, -8$ (Dudoit and Speed [8]). The score statistic is

$$S_{AST} = \frac{16}{3}(3\pi_1 + \pi_2 - \pi_3 - \pi_4)(3N_1 + N_2 - N_3 - N_4),$$

where N_j denotes the number of ASTs with IBD configuration C_j at the marker. Although the form of the statistic here is a little more complicated than that for sib-pairs S_{sib}, it is not overly so. For DSTs, however, there are seven IBD configurations, and the eigenvalue $\lambda_2 = -4$ has multiplicity two, leading to a score statistic that is the sum of two statistics similar to S_{AST}, but with seven rather than four terms in each factor.

To complete the picture for sib-trios, we have derived the score test statistic in the case of a quantitative trait (QST) as well. In this case, there are 10 IBD configurations

(Table 3); the infinitesimal generator Q here is

$$Q_{QST} = \begin{bmatrix} -6 & 2 & 0 & 0 & 0 & 2 & 0 & 0 & 2 & 0 \\ 1 & -6 & 1 & 1 & 1 & 1 & 0 & 0 & 1 & 0 \\ 0 & 2 & -6 & 2 & 2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & -6 & 1 & 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & -6 & 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 1 & -6 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 2 & 2 & -6 & 0 & 0 & 2 \\ 0 & 0 & 0 & 2 & 0 & 0 & 0 & -6 & 2 & 2 \\ 1 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & -6 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & -6 \end{bmatrix},$$

with eigenvalues 0 (multiplicity 1), -4 (multiplicity 3), and -8 (multiplicity 6). The three orthonormal (unit norm with respect to the inner product $\langle\ ,\ \rangle_\alpha$) right eigenvectors corresponding to $\lambda_2 = -4$ are

$$\begin{aligned} u &= (0,-1,-2,-1,0,1,2,0,0,1) \\ v &= (2,0,-2,-1,-1,1,0,0,1,0) \\ w &= \sqrt{2}(-1,0,1,1,0,-1,-1,1,0,0). \end{aligned}$$

Thus, the score statistic for n QSTs is based on the second largest eigenvalue of Q and is given by

$$\begin{aligned} S_{QST} = 8\sum_{i=1}^{n}\left(\sum_{j=1}^{10} u_j\pi_{ji}\right)\left(\sum_{k=1}^{10} u_k N_{ki}\right) + 8\sum_{i=1}^{n}\left(\sum_{j=1}^{10} v_j\pi_{ji}\right)\left(\sum_{k=1}^{10} v_k N_{ki}\right) \\ + 8\sum_{i=1}^{n}\left(\sum_{j=1}^{10} w_j\pi_{ji}\right)\left(\sum_{k=1}^{10} w_k N_{ki}\right) \end{aligned}$$

$$\begin{aligned} &= 8\sum_{i=1}^{n}(-\pi_{2i} - 2\pi_{3i} - \pi_{4i} + \pi_{6i} + 2\pi_{7i} + \pi_{10i})(-N_{2i} - 2N_{3i} - N_{4i} + N_{6i} + 2N_{7i} + N_{10i}) \\ &\quad + 8\sum_{i=1}^{n}(2\pi_{1i} - 2\pi_{3i} - \pi_{4i} - \pi_{5i} + \pi_{6i} + \pi_{9i})(2N_{1i} - 2N_{3i} - N_{4i} - N_{5i} + N_{6i} + N_{9i}) \\ &\quad + 8\sqrt{2}\sum_{i=1}^{n}(-\pi_{1i} + \pi_{3i} + \pi_{4i} - \pi_{6i} - \pi_{7i} + \pi_{8i})(-N_{1i} + N_{3i} + N_{4i} - N_{6i} - N_{7i} + N_{8i}). \end{aligned}$$

So for quantitative traits, even with only one extra individual, the form of the score statistic is already much more complicated, and correspondingly much less interpretable, than it is for pairs. In addition, specification of a joint phenotypic model is more cumbersome for larger pedigrees, and may also be unstable due to the larger number of IBD configurations. Furthermore, even once a model is specified, exact calculation of the statistic also becomes more difficult.

In nonparametric linkage analysis, the problem of dealing with larger sets of relatives than pairs has been approached in a number of different ways [26, 33, 34, 35]; for reviews, see [5, 15]. A widely used method to handle the issue is to consider the set of relatives only pairwise, typically by considering all possible pairs [22]. We have begun to compare exact treatment of QSTs with approximations based on pairwise score statistics. We hope to arrive at a weighting scheme based on pairs which will provide a good approximation to the exact treatment, yet is simpler and faster to compute and interpret.

3.2 Score Test with Missing IBD Information

Computing the score statistic relies on availability of complete inheritance vectors, so that there is sufficient genotypic information to determine IBD allele sharing status. In practice, however, the available genotype data may be limited to information on the allele states (identity by state, or IBS) and thus there is some information missing. IBD status may also be missing due to failure of the genotyping method for some individuals or unavailability of connecting individuals in the pedigree. It is therefore desirable to modify the score test to allow for the case of incomplete genotypic information.

When IBD information is incomplete, partial information obtained from marker data may be summarized by the *inheritance distribution*, a conditional probability distribution over possible inheritance vectors at the marker locus [23, 24]. Now, rather than counting the number of pedigrees with IBD configuration C_j, let

$$r_{ji} = P(\text{Pedigree has IBD configuration } C_j \text{ at the marker} \mid M_i),$$

where M_i denotes available marker information. Then a natural test statistic $\widetilde{S}(\nu)$ may be obtained from the complete data score statistic $S(\nu)$, by replacing the IBD indicators by their expectation given the marker data. When the trait and marker loci are in linkage equilibrium, then

$$\widetilde{S}(\nu) = E_0[S(\nu) \mid M, \phi]$$

for marker data M and phenotypes ϕ. Kruglyak *et al.* [23] use a similar statistic with a "perfect data" approximation, which consists of substituting the null variance of the complete data statistic, $Var_0[S(\nu) \mid \phi]$, for the null variance $Var_0[\widetilde{S}(\nu) \mid \phi]$ of the incomplete data statistic. This approximation is conservative, as $Var_0[\widetilde{S}(\nu) \mid \phi] \leqslant Var_0[S(\nu) \mid \phi]$.

In fact, the true inheritance distribution $\{r_{ji}\}$ will rarely be known; rather, it must be estimated from the data, for example with the program GENEHUNTER [23]. Call these estimated probabilities $\{\hat{r}_{ji}\}$. Then the incomplete data statistic for sib-pairs (ignoring the multiplicative constant 16) is

$$\widetilde{S}(\nu) = \sum_{i=1}^{n} (\pi_{2i} - \pi_{0i})(\hat{r}_{2i} - \hat{r}_{0i}).$$

The null expectation and variance of $\widetilde{S}(\nu)$ may be estimated using sample moments of $\hat{r}_{2i} - \hat{r}_{0i}$ from the data. This approach may be problematic, though, as there must be a sufficient number of sib-pairs with the same missing genotype pattern to give reliable estimates. This aspect is even worse with larger pedigrees.

We believe that *multiple imputation* provides a more promising approach to estimation of the linkage score statistic with missing IBD data. Rubin [31] details multiple imputation procedures in the context of survey nonresponse; for multiple imputation in genetics problems, see Clayton [4]. With single imputation, one value is chosen for the missing information. With multiple imputation, missing data are replaced with at least two values representing the distribution of possibilities. Multiple imputation methods allow standard complete data methods to be applied, have increased efficiency over single imputation methods, and also more realistically reflect the increase in uncertainty due to the missing information. Thus, we are currently working to extend the applicability of the score test using multiple imputation to estimate missing IBD sharing.

Sampling from the imputation distribution $[M_{missing} \mid M_{observed}]$ of the marker information under the null T times yields multiple copies of "complete" data. Each of these produces a statistic $S^{(t)}$, $t = 1, ..., T$ (we now suppress the dependence of S on genetic model ν to avoid cumbersome notation below). Then we can define the multiple imputation "score" statistic S^* as the average value of $S^{(t)}$ over the T copies:

$$S^* = \frac{1}{T} \sum_{t=1}^{T} S^{(t)}.$$

Under the null, $E(S^*) = 0$ and

$$Var(S^*) = V - \frac{(T-1)}{T} E_{observed} \left\{ Var_{[M_{missing}|M_{observed}]}(S) \right\},$$

where $V = Var_0[S \mid \phi]$ is the complete data score statistic variance under the null [4]. The second term may be estimated using the sample variance of the imputation statistics $S^{(t)}$.

This method of extending the score test for linkage may be viewed as a compromise between the conservative "perfect data" approximation and exact calculation by enumeration of all possible states for the missing marker data. Such evaluation quickly becomes infeasible when several markers, each with large numbers of alleles, are used, as is common in linkage studies. There will be a reduction of power attributable to missing information, but preliminary simulations using the multiple imputation approach are encouraging. We are working toward a more complete implementation with the aim of broadening the class of problems to which the score test approach may be applied.

Sandrine Dudoit, Division of Biostatistics, University of California, Berkeley, `sandrine@stat.berkeley.edu`

Darlene R. Goldstein, Bioinformatics Core Facility, Institut Suisse de Recherche Expérimentale sur le Cancer (ISREC), 1066 Epalinges, Switzerland; and Institut de mathematiques (IMA), École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland, `darlene.goldstein@isrec.unil.ch`

References

[1] C. I. Amos. Robust variance-components approach for assessing genetic linkage in pedigrees. *American Journal of Human Genetics*, 54:535–543, 1994.

[2] C. I. Amos and R. C. Elston. Robust methods for the detection of genetic linkage for quantitative data from pedigrees. *Genetic Epidemiology*, 6:349–360, 1989.

[3] C. I. Amos, R. C. Elston, A. F. Wilson, and J. E. Bailey-Wilson. A more powerful robust sib-pair test of linkage for quantitative traits. *Genetic Epidemiology*, 6:435–449, 1989.

[4] D. Clayton. Tests for genetic linkage and association with incomplete data. 2001. Invited talk; available at `http://www-gene.cimr.cam.ac.uk/clayton/talks/enar01.pdf`.

[5] S. Davis and D. E. Weeks. Comparison of nonparametric statistics for detection of linkage in nuclear families: single-marker evaluation. *American Journal of Human Genetics*, 61:1431–1444, 1997.

[6] N. E. Day and M. J. Simons. Disease-susceptibility genes — their identification by multiple case family studies. *Tissue Antigens*, 8:109–119, 1976.

[7] S. Dudoit. *Linkage analysis of complex human traits using identity by descent data.* Ph.D. thesis, Department of Statistics, University of California, Berkeley, 1999.

[8] S. Dudoit and T. P. Speed. A score test for linkage using identity by descent data from sibships. *Annals of Statistics*, 27:943–986, 1999.

[9] S. Dudoit and T. P. Speed. Triangle constraints for sib-pair identity by descent probabilities under a general multilocus model for disease susceptibility. In M. E. Halloran and S. Geisser, editors, *Statistics in Genetics*, volume 112 of *IMA Volumes in Mathematics and its Applications*, pages 181–221. Springer-Verlag, New York, 1999.

[10] S. Dudoit and T. P. Speed. A score test for linkage analysis of qualitative and quantitative traits based on identity by descent data on sib-pairs. *Biostatistics*, 1:1–26, 2000.

[11] S. A. G. E. Statistical analysis for genetic epidemiology. *Genetic Epidemiology*, 1998.

[12] R. C. Elston, S. Buxbaum, K. B. Jacobs, and J. M. Olson. Haseman and elston revisited. *Genetic Epidemiology*, 19:1–17, 2000.

[13] S. N. Ethier and S. E. Hodge. Identity-by-descent analysis of sibship configurations. *American Journal of Medical Genetics*, 22:263–272, 1985.

[14] D. S. Falconer and T. F. C. Mackay. *Introduction to Quantitative Genetics*. Longman, Essex, England, 4th edition, 1996.

[15] E. Feingold, K. K. Song, and D. E. Weeks. Comparison of allele-sharing statistics for general pedigrees. *Genetic Epidemiology*, 19 Suppl 1:S92–S98, 2000.

[16] D. W. Fulker and L. R. Cardon. A sib-pair approach to interval mapping of quantitative trait loci. *American Journal of Human Genetics*, 54:1092–1103, 1994.

[17] D. W. Fulker, S. S. Cherny, and L. R. Cardon. Multipoint interval mapping of quantitative trait loci, using sib pairs. *American Journal of Human Genetics*, 56:1224–1233, 1995.

[18] D. R. Goldstein, S. Dudoit, and T. P. Speed. Power of a score test for quantitative trait linkage analysis of relative pairs. *Genetic Epidemiology*, 19 Suppl 1:S85–S91, 2000.

[19] D. R. Goldstein, S. Dudoit, and T. P. Speed. Power and robustness of a score test for linkage analysis of quantitative traits using identity by descent data on sib pairs. *Genetic Epidemiology*, 20:415–431, 2001.

[20] J. K. Haseman and R. C. Elston. The investigation of linkage between a quantitative trait and a marker locus. *Behavior Genetics*, 2:3–19, 1972.

[21] E. R. Hauser and M. Boehnke. Genetic linkage analysis of complex genetic traits by using affected sibling pairs. *Biometrics*, 54:1238–1246, 1998.

[22] S. Hodge. The information contained in multiple sibling pairs. *Genetic Epidemiology*, 1:109–122, 1984.

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

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

[25] E. S. Lander and P. Green. Construction of multilocus genetic maps in humans. *Proceedings of the National Academy of Sciences, USA*, 84:2363–2367, 1987.

[26] K. Lange. A test statistic for the affected sibset method. *Annals of Human Genetics*, 50:283–290, 1986.

[27] M. S. McPeek. An introduction to recombination and linkage analysis. In T. P. Speed and M. S. Waterman, editors, *Genetic Mapping and DNA Sequencing*, volume 81 of *IMA Volumes in Mathematics and its Applications*. Springer-Verlag, New York, 1996.

[28] J. M. Olson, S. Rao, K. B. Jacobs, and R. C. Elston. Linkage of chromosome 1 markers to alcoholism related phenotypes by sib pair linkage analysis of principal components. *Genetic Epidemiology*, 17 Suppl 1:S271–S276, 1999.

[29] J. M. Olson and E. Wijsman. Linkage between quantitative trait and marker locus: methods using all relative pairs. *Genetic Epidemiology*, 10:87–102, 1993.

[30] J. Ott. *Analysis of Human Genetic Linkage*. Johns Hopkins University Press, Baltimore, 3rd edition, 1999.

[31] D. B. Rubin. *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York, 1987.

[32] T. P. Speed. What is a genetic map function? In T. P. Speed and M. S. Waterman, editors, *Genetic Mapping and DNA Sequencing*, volume 81 of *IMA Volumes in Mathematics and its Applications*. Springer-Verlag, New York, 1996.

[33] P. J. Ward. Some developments on the affected-pedigree-member method of linkage analysis. *American Journal of Human Genetics*, 52:1200–1215, 1993.

[34] D. E. Weeks and K. Lange. The affected-pedigree-member method of linkage analysis. *American Journal of Human Genetics*, 42:315–326, 1988.

[35] A. S. Whittemore and J. Halpern. A class of tests for linkage using affected pedigree members. *Biometrics*, 50:118–127, 1994.

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