Chapter 12

Statistical Genetics

Darlene R. Goldstein

Terry Speed has produced many interesting and important contributions to the field of statistical genetics, with work encompassing both experimental crosses and human pedigrees. He has been instrumental in uncovering and elucidating algebraic structure underlying a diverse range of statistical problems, providing new and unifying insights.

Here, I provide a brief commentary and introduction to some of the key building blocks for an understanding of the papers. Some readers may also find useful a refresher on group action (see e.g. Fraleigh [6]) and hidden Markov models [13].

Linkage mapping

Linkage analysis studies inheritance of traits in families, with the aim of determining the chromosomal location of genes influencing the trait. The 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.

A measure of the degree of linkage is the recombination fraction θ , the chance of recombination occurring between two loci. For unlinked genes, $\theta = 1/2$; for linked genes, $0 \leq \theta < 1/2$.

D.R. Goldstein

Institut de mathématiques d'analyse et applications, École Polytechnique Fédérale de Lausanne, and Swiss Institute of Bioinformatics, Switzerland
e-mail: darlene.goldstein@epfl.ch

Human pedigrees

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

This paper offers a novel and comprehensive algebraic view of sib-pair methods, fundamentally unifying a large collection of apparently ad hoc procedures and providing powerful insights into the methods.

Identical by descent allele sharing

Data for linkage analysis consist of sets of related individuals (pedigrees) and information on the genetic marker and/or trait genotypes or phenotypes. The recombination fraction is most commonly estimated by maximum likelihood for an appropriate genetic model for the coinheritance of the loci.

However, likelihood-based linkage analysis can be difficult to carry out due to the problem that the mode of inheritance may be complex and in any case is usually unknown. Nonparametric approaches are thus appealing, since they do not require a genetic inheritance model to be specified. Such methods typically 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 the trait and allele sharing at loci linked to trait susceptibility loci, which can be used to localize trait susceptibility genes.

Testing for linkage with IBD data has developed differently, depending on the type of trait. For qualitative traits, tests are based on *IBD sharing conditional on phenotypes*. Affected sib-pair methods are a popular choice; these are often described as nonparametric since the mode of inheritance does not need to be specified (see Hauser and Boehnke [10] for a review). On the other hand, for quantitative trait loci (QTL), linkage analysis is based on examination of *phenotypes conditional on sharing* (for example, the method of Haseman and Elston [9] or one of its many extensions).

Inheritance vector

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. For sibships of size k , label locus (1, 2) and maternally derived alleles (3, 4). The inheritance vector at a given locus is the vector $x = (x_1, x_2, \dots, x_{2k-1}, x_{2k})$, 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.

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 12.1). The IBD configurations can be thought of as orbits of groups acting on the set of possible inheritance vectors $\mathcal{X}$ [2].

Table 12.1 Sib-pair IBD configurations

Alleles IBD	Inheritance vectors	$ C_j $
0 IBD	(1, 3, 2, 4), (1, 4, 2, 3), (2, 3, 1, 4), (2, 4, 1, 3)	4
1 IBD	(paternal) (1, 3, 1, 4), (1, 4, 1, 3), (2, 3, 2, 4), (2, 4, 2, 3)	8
	(maternal) (1, 3, 2, 3), (1, 4, 2, 4), (2, 3, 1, 3), (2, 4, 1, 4)	
2 IBD	(1, 3, 1, 3), (1, 4, 1, 4), (2, 3, 2, 3), (2, 4, 2, 4)	4

Score test for linkage

The literature contains several proposed tests of the null hypothesis of no linkage ($H : \theta = 1/2$) based on score functions of IBD configurations for sibships and other pedigrees, with scores chosen to yield good power against a particular alternative. The score test of Dudoit and Speed to detect linkage represents a major breakthrough in that it creates a coherent, unified based approach to the linkage analysis of qualitative and quantitative traits using IBD data. 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 .

In general, the probability vector π depends on unknown genetic parameters. However, using their formulation of the problem, Dudoit and Speed [4] show rigorously that for affected sibships of a given size, the second-order Taylor series expansion of the log likelihood around the null of no linkage is independent of the genetic inheritance model. They thus provide a mathematically justified basis for affected sib-pair methods, which do not require an assumed mode of inheritance.

Practical advantages of the score test include: it is locally most powerful for alternatives close to the null; any genotype distribution can be used (i.e., Hardy-Weinberg equilibrium is not required); conditioning on phenotypes eliminates selection bias introduced by nonrandom ascertainment; and combining differently ascertained pairs is straightforward, providing the important benefit of allowing us to

avoid discarding any data. For many realistic simulation scenarios [7, 8], the score test proves to be robust and shows large power gains over commonly used nonparametric tests.

Although the paper focuses on pairs of sibs, the same score test approach is also applicable to any set of relatives [3].

Experimental crosses

N. J. Armstrong, M. S. McPeck and T. P. Speed (2006). Incorporating interference into linkage analysis for experimental crosses. *Biostatistics* 7:374–386.

This paper improves multilocus linkage analysis of experimental crosses by incorporating a realistic model of crossover interference, and implementing it by extending the Lander-Green algorithm for genetic reconstruction. It represents the culmination of a series of studies of the modeling of crossover interference.

χ^2 model of crossover interference

During the (four-strand) process of crossing over in meiosis, two types of interference (nonindependence) are distinguished: chromatid interference, a situation in which the occurrence of a crossover between any pair of nonsister chromatids affects the probability of those chromatids being involved in other crossovers in the same meiosis; and crossover interference, which refers to nonrandom location of chiasmata along a chromosome.

Most genetic mapping is carried out assuming independence; that is, no chromatid interference and no crossover interference. This assumption simplifies likelihood calculations. Although there is little empirical evidence for chromatid interference, there is substantial evidence of crossover interference. Thus, more a more realistic model incorporating crossover interference should be able to provide more accurately estimated genetic maps.

The χ^2 model of crossover interference [5] provides a dramatically improved fit over a wide range of models [12, 14]. This model assumes that recombination intermediates (structures formed after initiation of recombination) are resolved in one of two ways: either with or without crossing over. Recombination initiation events are assumed to occur according to a Poisson distribution, but constraints on the resolution of intermediates creates interference. The χ^2 model assumes m unobserved intermediates between each crossover. For $m = 1$, the model reduces to the no (crossover) interference model. This model is a special case of the more general gamma model, but has the advantage of being computationally more feasible.

Genetic reconstruction and the Lander-Green algorithm

Genetic mapping in humans can be viewed as a missing data problem, since we are typically unable to observe the complete data (the number of recombinant and non-recombinant meioses for each interval). If complete data were available, maximum likelihood estimates of a set of recombination fractions θ_i , $i = 1, \dots, T - 1$, for adjacent markers $\mathcal{M}_1, \dots, \mathcal{M}_T$ would just be the observed proportion of recombinants in an interval.

The genetic reconstruction problem is to determine the expected number of recombinations that occurred in intervals of adjacent markers, given genotypes at multiple marker loci in a pedigree and the recombination fraction for each interval. Construction is straightforward when there is complete genotype information, including the ancestral origin (paternal or maternal).

More commonly this information is not known, so a different strategy for likelihood calculation is needed to obtain recombination fraction estimates. Lander and Green [11] proposed an approach based on the use of inheritance vectors. They showed that the probability of the observed data can be calculated for any particular inheritance vector and that under no crossover interference, the inheritance vectors form a Markov chain along the chromosome. They model the pedigree and data as a hidden Markov model, where the hidden states are the (unobserved) inheritance vectors. The complexity of their algorithm for calculating likelihoods increases linearly with the number of markers but exponentially in pedigree size, making it appropriate for analysis of many markers on small to moderately sized pedigrees.

In experimental crosses, mapping is generally more straightforward since investigators can arrange crosses to produce complete data. However, the presence of unobserved recombination initiation points creates a new kind of missing data when the no interference model is not assumed. The creative insight of Armstrong et al. [1] is to model the crossover interference process as a hidden Markov model. This step works because even though crossovers resulting from initiation events do not occur independently (in the presence of crossover interference), the initiation events themselves *are* assumed to be independent. Armstrong et al. [1] are thus able to extend the Lander-Green algorithm to incorporate interference according to the χ^2 model, thereby providing more accurately estimated genetic maps.

Conclusion

Terry's work in statistical genetics has identified underlying commonalities across seemingly disparate procedures, contributing meaningful theoretical and practical improvements. An impressive aspect of these works is the fresh perspective offered by viewing the problems at a stripped-down, fundamental level. Applying an exceptional combination of extensive mathematical expertise and pragmatic sensibility, Terry provides inventive solutions and a richer structural understanding of significant questions in statistical genetics.

References

- [1] N. J. Armstrong, M. S. McPeck, and T. P. Speed. Incorporating interference into linkage analysis for experimental crosses. *Biostatistics*, 7(3):374–386, 2006.
- [2] K. P. Donnelly. The probability that related individuals share some section of genome identical by descent. *Theor. Popul. Biol.*, 23:34–63, 1983.
- [3] S. Dudoit. *Linkage Analysis of Complex Human Traits Using Identity by Descent Data*. PhD thesis, Department of Statistics, University of California, Berkeley, 1999.
- [4] S. Dudoit and T. P. Speed. A score test for linkage using identity by descent data from sibships. *Ann. Stat.*, 27(3):943–986, 1999.
- [5] E. Foss, R. Lande, F. Stahl, and C. Steinberg. Chiasma interference as a function of genetic distance. *Genetics*, 133:681–691, 1993.
- [6] J. B. Fraleigh. *A First Course in Abstract Algebra*. Addison-Wesley Pub. Co., Reading, Mass., 7th edition, 2002.
- [7] D. R. Goldstein, S. Dudoit, and T. P. Speed. Power of a score test for quantitative trait linkage analysis of relative pairs. *Genet. Epidemiol.*, 19(Suppl. 1): S85–S91, 2000.
- [8] D. R. Goldstein, S. Dudoit, and T. P. Speed. Power and robustness of a score test for linkage analysis of quantitative trait based on identity by descent data on sib pairs. *Genet. Epidemiol.*, 20(4):415–431, 2001.
- [9] J. K. Haseman and R. C. Elston. The investigation of linkage between a quantitative trait and a marker locus. *Behav. Genet.*, 2:3–19, 1972.
- [10] E. R. Hauser and M. Boehnke. Genetic linkage analysis of complex genetic traits by using affected sibling pairs. *Biometrics*, 54:1238–1246, 1998.
- [11] E. S. Lander and P. Green. Construction of multilocus genetic maps in humans. *Proc. Natl. Acad. Sci. USA*, 84:2363–2367, 1987.
- [12] M. S. McPeck and T. P. Speed. Modeling interference in genetic recombination. *Genetics*, 139:1031–1044, 1995.
- [13] L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.
- [14] H. Zhao, T. P. Speed, and M. S. McPeck. Statistical analysis of crossover interference using the chi-square model. *Genetics*, 139:1045–1056, 1995.

The Annals of Statistics
1999, Vol. 27, No. 3, 943–986

A SCORE TEST FOR LINKAGE USING IDENTITY BY DESCENT DATA FROM SIBSHIPS

BY SANDRINE DUDOIT¹ AND TERENCE P. SPEED²

University of California, Berkeley

We consider score tests of the null hypothesis $H_0: \theta = \frac{1}{2}$ against the alternative hypothesis $H_1: 0 \leq \theta < \frac{1}{2}$, based upon counts multinomially distributed with parameters n and $\rho(\theta, \pi)_{1 \times m} = \pi_{1 \times m} T(\theta)_{m \times m}$, where $T(\theta)$ is a transition matrix with $T(0) = I$, the identity matrix, and $T(\frac{1}{2}) = (1, \dots, 1)^T(\alpha_1, \dots, \alpha_m)$. This type of testing problem arises in human genetics when testing the null hypothesis of no linkage between a marker and a disease susceptibility gene, using identity by descent data from families with affected members. In important cases in this genetic context, the score test is independent of the nuisance parameter π and based on a widely used test statistic in linkage analysis. The proof of this result involves embedding the states of the multinomial distribution into a continuous-time Markov chain with infinitesimal generator Q . The second largest eigenvalue of Q and its multiplicity are key in determining the form of the score statistic. We relate Q to the adjacency matrix of a quotient graph in order to derive its eigenvalues and eigenvectors.

1. Introduction. This paper concerns a class of score tests which arise naturally in human genetics. However, their essence can be described quite efficiently without any of the genetic background, and we now do so. Let $\alpha = (\alpha_1, \dots, \alpha_m)$ and $\pi = (\pi_1, \dots, \pi_m)$ be two multinomial distributions, viewed as points in a simplex, and let $\{T(\theta): 0 \leq \theta \leq \frac{1}{2}\}$ be a one-parameter family of transition matrices such that $T(0) = I$, the identity matrix, and $T(\frac{1}{2}) = \mathbf{1}^T \alpha$, where $\mathbf{1} = (1, \dots, 1)$. These objects allow us to define the curve $\mathcal{C}_\pi(\theta)$ of distributions $\rho(\theta, \pi) = \pi T(\theta)$, $0 \leq \theta \leq \frac{1}{2}$, connecting $\pi = \rho(0, \pi)$ to $\alpha = \rho(\frac{1}{2}, \pi)$. Our interest is a score test for the null hypothesis $H_0: \theta = \frac{1}{2}$ against the alternative $H_1: 0 \leq \theta < \frac{1}{2}$, that is, for testing $H_0: \rho = \alpha$ against alternatives along the curve $\mathcal{C}_\pi(\theta)$, based upon counts $N = (N_1, \dots, N_m)$ multinomially distributed with parameters $n = \sum_i N_i$ and $\rho(\theta, \pi)$. The associated log-likelihood is $l(\theta, \pi) = \sum_i N_i \ln(\rho_i(\theta, \pi))$, and the score test in question should be based on $l'(\frac{1}{2}, \pi) = \sum_i N_i \rho'_i(\frac{1}{2}, \pi)/\alpha_i$, where $'$ denotes differentiation in θ . It turns out in our genetic context that $l'(\frac{1}{2}, \pi) \equiv 0$, and so we consider the second derivative, obtaining $l''(\frac{1}{2}, \pi) = \sum_i N_i \rho''_i(\frac{1}{2}, \pi)/\alpha_i = \sum_i N_i (\sum_j \pi_j u_{ji})/\alpha_i$, where $U = (u_{ij}) = T''(\frac{1}{2})$. Now, we would normally need

Received August 1998; revised March 1999.

¹Supported in part by a PMMB Burroughs-Wellcome Fellowship.

²Supported in part by NIH Grant R01 HG01093.

AMS 1991 subject classifications. Primary 62F03; secondary 92D30, 60J20, 15A18, 05C20, 05C30.

Key words and phrases. Score test, Markov chain, infinitesimal generator, quotient graph, adjacency matrix, eigenvalues, orbits, Pólya's theory, linkage analysis.

S. DUDOIT AND T. P. SPEED

to deal with the nuisance parameter π in this score test. This study was motivated by the observation that in some important cases in linkage analysis, U has rank 1, that is, $u_{ij} = a_i b_j$, for suitable vectors (a_i) and (b_i) . In such cases, $l''(\frac{1}{2}, \pi) = (\sum_j a_j \pi_j)(\sum_i b_i N_i / \alpha_i)$, and the score test is independent of the nuisance parameter π . Moreover, for our genetic problem, the score test is based on a widely used nonparametric statistic in linkage analysis, S_{pairs} [Whittemore and Halpern (1994a), Kruglyak, Daly, Reeve-Daly and Lander (1996)]. We thought it would be of interest to understand the origins of this property and to learn just how far it extended.

In Section 2, we present the genetic problem which motivated our study, the linkage analysis of disease susceptibility genes using identity by descent (IBD) data from sets of siblings (sibships). This involves describing how IBD patterns in pedigrees (i.e., collections of related individuals) may be summarized by inheritance vectors which correspond to the vertices of a hypercube. The inheritance vectors along a chromosome are embeddable in a continuous-time random walk on the vertices of the hypercube, with time parameter the genetic distance along the chromosome. For our purpose, the inheritance vectors may be partitioned into so-called IBD configurations, which are orbits of groups acting on the set of inheritance vectors. In Section 3, we derive a semi-group property for the IBD configuration transition matrix $T(\theta)$ and present a spectral decomposition of $T(\theta)$ in terms of the eigenvalues and eigenvectors of its infinitesimal generator Q . The second largest eigenvalue of Q and its multiplicity are key in determining the form of the linkage score statistic. In order to derive the eigenvalues and eigenvectors of the infinitesimal generator, we relate it to the adjacency matrix of a quotient graph. Finally, in Section 4, we derive score statistics for testing linkage in sibships and illustrate the results with sib-pairs and sib-trios in Section 5. Remarkably, in an affected-only analysis, where siblings not affected by the disease are ignored, the score test for sibships of a given size does not depend on the nuisance parameter π and is based on a well-known statistic in linkage analysis, S_{pairs} [Whittemore and Halpern (1994a), Kruglyak, Daly, Reeve-Daly and Lander (1996)].

2. Testing linkage using identity by descent data. Genetic mapping is based upon the phenomenon of *crossing-over*, which is the exchange of corresponding DNA between chromosomes from the same pair during gamete (egg/sperm) formation. The human genome is distributed along 23 pairs of chromosomes, 22 autosomal pairs and the sex chromosome pair (XX for females and XY for males). Each pair consists of a paternally inherited chromosome and a maternally inherited chromosome. As a result of crossovers, chromosomes passed from parent to offspring are combinations of the two grandparental chromosomes (see Figures 1 and 2). In general, the DNA variants (*alleles*) passed from parent to offspring at two nearby chromosomal locations (*loci*) have the same grand-parental origin (e.g., at both loci, the maternally inherited alleles are from the maternal grandfather). This is sometimes called *cosegregation*, as *segregation* is the process leading to the choice of one of a parent's two variants (maternal or paternal) at any given locus for transmis-

A SCORE TEST FOR LINKAGE

FIG. 1. Segregation products for a sibship of size 2 and a single chromosome pair. Male and female individuals are represented by squares and circles, respectively, and colored symbols indicate affectedness by the disease under study. The paternal and maternal chromosome pairs are labeled by (1, 2) and (3, 4), respectively. The inheritance vectors and IBD configurations of the sib-pair are indicated on the left.

sion to a child. Exceptions to cosegregation occur due to crossovers; then, the variants passed on to the child have different grand-parental origins at the two loci and the chromosome is said to be *recombinant* (e.g., for the maternally inherited chromosome, the variant from the maternal grandfather was inherited at one locus and that from the maternal grandmother was inherited at the other locus). The frequency with which this occurs is the *recombination fraction* between the two loci, conventionally denoted by θ . [Recombination fractions are assumed to be constant across conditions (e.g., age and temperature) and individuals. Under general models for crossovers, the recombination fraction between two loci belongs to the interval $[0, 1/2]$.] In general, two loci are said to be *linked* if their recombination fraction is less than $\frac{1}{2}$, and *unlinked* if it is $\frac{1}{2}$. Thus, unlinked loci may be widely separated on the same chromosome, or on different chromosomes. Loci are said to be *tightly linked* if the recombination fraction θ is close to 0, for example, $\theta < 0.05$ [see Ott (1991), McPeck (1996) and Speed (1996) for an introduction to linkage analysis].

S. DUDOIT AND T. P. SPEED

FIG. 2. Four-dimensional hypercube whose vertices correspond to the 16 possible inheritance vectors for a sib-pair and whose edges represent permissible transitions. The arrows indicate the transitions for the segregation products represented in Figure 1.

When mapping disease susceptibility (DS) genes, we are interested in testing whether genetic markers with known location are linked or not to DS genes, that is, in testing a null hypothesis of the form $H_0: \theta = \frac{1}{2}$, where θ is the recombination fraction between a genetic marker and a putative DS gene. This could be done by studying the cosegregation of variants of the DS genes with those of other mapped genes or markers. (By now, there are hundreds of well-mapped markers along each human, mouse and many other chromosomes.) Frequent cosegregation of a DS locus with a mapped marker would imply a small recombination fraction between the two loci, and hence an accurate placement of the DS locus. However, for most diseases of interest, we do not in general know, and are unable to determine, the alleles present at the DS loci prior to their being mapped. Indeed, much of the interest in mapping DS loci is to determine the variants segregating in populations. Thus, a direct approach to mapping DS loci is generally not available. Many ingenious methods have been developed by geneticists to circumvent this problem, and this paper concerns one such which studies marker identity by descent in sibships with affected members. DNA at the same locus on two chromosomes from the same pair is said to be *identical by descent* (*IBD*) if it originated from

A SCORE TEST FOR LINKAGE

the same ancestral chromosome. This is by contrast to *identity by state* (IBS), where the same DNA variant in two individuals may have entered the family under study through different ancestors and hence may not be IBD. Linkage analysis methods based on IBD data seek to exploit the association between the *sharing of DNA identical by descent at loci linked to DS loci and disease status (phenotype)* in families with affected individuals. At loci unlinked to DS loci, IBD sharing is independent of phenotype. For example, for sibships, this association arises from the fact that full sibs get all their genes from the same source, their parents. Consequently, if susceptibility to the disease under study has a genetic component, the disease status (affected or not) of the sibs should be associated with their IBD status (identical or not) at the DS loci [see Tables 3 and 4 in Dudoit and Speed (1999) for a simple example of this association in sib-pairs]. This approach is effective to the extent that disease susceptibility is affected by genes rather than, say, a shared environment, or other “random” factors. Determining IBD status at or near a DS locus is usually feasible, where determining the gene variant is not, because one can readily determine IBD status at so-called marker loci, one of which may be tightly linked to the DS locus [cf. Kruglyak and Lander (1995) and Kruglyak, Daly, Reeve-Daly and Lander (1996) for a treatment of incomplete IBD data]. IBD-based methods for detecting linkage to DS loci will thus be successful if (and only if) (1) there is a noticeable association between phenotype and IBD status of relatives at the DS loci, and (2) this association is strong enough to remain detectable when IBD status at an (unknown) DS locus is replaced by observed IBD status at a marker locus. Recombination between a DS locus and a marker locus will attenuate the association between phenotype and IBD status. If we have a dense enough set of marker loci, problem (2) would appear to be solved, but in truth there is always a close connection between the magnitude of the association in (1) and the density of the marker set necessary for its detection. These issues were addressed by Thompson (1997), who refers to the two components (1) and (2) as the *specificity* of the DS loci and the *scale* of the genetic distance, respectively. We refer the reader to Whittemore and Halpern (1994a, b), Whittemore (1996), Kruglyak, Daly, Reeve-Daly and Lander (1996), Kong and Cox (1997), Feingold and Siegmund (1997), Teng and Siegmund (1997), Dudoit (1999), and Dudoit and Speed (1999), McPeck (1999) for recent discussions of linkage analysis using IBD data.

The IBD pattern within a pedigree may be summarized at any chromosomal locus by the inheritance vector. Consider a sibship of $k \geq 2$ sibs and suppose we wish to identify the parental origin of the DNA inherited by each sib at a particular autosomal locus, $\mathcal{L}$ say (for loci on sex chromosomes, males and females need to be treated differently). Arbitrarily label the paternal chromosomes containing the locus of interest by (1, 2), and similarly label the maternal chromosomes by (3, 4). The *inheritance vector* of the sibship at the locus $\mathcal{L}$ is the $2k$ -vector $x = (x_1, x_2, \dots, x_{2k-1}, x_{2k})$, indicating the outcome of each of the $2k$ segregations giving rise to the sibship. More precisely, for $i = 1, \dots, k$, x_{2i-1} is the label of the paternal chromosome from which sib i inherited DNA at $\mathcal{L}$, 1 or 2, and x_{2i} is the label of the maternal chromosome

S. DUDOIT AND T. P. SPEED

from which sib i inherited DNA at $\mathcal{L}$, 3 or 4 (see Figures 1 and 2). Denote by $\mathcal{X}$ the set of all 2^{2k} inheritance vectors.

Consider now two loci, $\mathcal{L}_1$ and $\mathcal{L}_2$, separated by a recombination fraction θ , and denote the inheritance vectors at the two loci by x and y , respectively. If these two inheritance vectors differ at a particular entry, this indicates the occurrence of a recombination between $\mathcal{L}_1$ and $\mathcal{L}_2$ in the corresponding segregation. The chance of a recombination between the two loci is the recombination fraction θ , taken to be constant across conditions and individuals. the transition matrix $R(\theta)$ between inheritance vectors at loci separated by a recombination fraction θ has entries

$$(2.1) \quad r_{xy}(\theta) = \theta^{\Delta(x,y)}(1 - \theta)^{2k - \Delta(x,y)},$$

where $\Delta(x, y)$ is the number of coordinates at which the inheritance vectors x and y differ, that is, the number of recombination events between the two loci. The matrix $R(\theta)$ may be expressed as the Kronecker power of 2×2 transition matrices corresponding to transitions in each of the $2k$ coordinates,

$$(2.2) \quad R(\theta) = \left[\begin{array}{cc} 1 - \theta & \theta \\ \theta & 1 - \theta \end{array} \right]^{\otimes 2k}.$$

The notion of inheritance vector generalizes to arbitrary pedigrees, where pedigree members are separated into founders (individuals whose parents are not in the pedigree) and nonfounders (individuals whose parents are in the pedigree). In the case of sibships, the parents are founders and the sibs are nonfounders. For a pedigree with k nonfounders, the inheritance vector at a particular locus is defined to be a $2k$ -vector with coordinates describing the outcome of the paternal and maternal segregations giving rise to the k nonfounders [Lander and Green (1987); Kruglyak and Lander (1995) and Kruglyak, Daly, Reeve-Daly and Lander (1996)]. The $(2i - 1)$ st coordinate is 0 or 1 according to whether the grand-paternal or grand-maternal DNA was transmitted in the paternal segregation giving rise to the i th nonfounder. The $(2i)$ th coordinate contains the same information for the maternal segregation. For the purpose of this paper, we prefer the definition introduced earlier for sibship inheritance vectors (with labels 1, 2, 3, 4 for parental chromosomes) to the more common definition with binary labels, since this facilitates the presentation in Section 3 of group action on inheritance vectors. It is easy to show that for general pedigrees the transition matrix $R(\theta)$ between inheritance vectors also has the form given in (2.2). Although this paper is primarily concerned with IBD data from sibships, the general setup and some of the results presented here (Propositions 1, 2 and 3) apply to arbitrary pedigree types, and we discuss generalizations where appropriate.

For the purpose of linkage analysis of disease genes, certain inheritance vectors are equivalent to each other, in that they have the same probability of arising at DS genes in pedigrees with given phenotypes. Although not needed for an understanding of this paper, a discussion of these probabilities for sibships and the genetic model under which they are calculated may be found in

A SCORE TEST FOR LINKAGE

Dudoit and Speed (1999) and Dudoit (1999). For an arbitrary pedigree type, we define *IBD configurations* as any partitioning of the set of inheritance vectors. For example, for k affected sibs, Ethier and Hodge (1985) partition the 2^{2k} inheritance vectors into a much smaller number of equivalence classes as follows. Two inheritance vectors belong to the same IBD configuration if one may be obtained from the other by applying any combination of the following four operations: (1) interchange the paternal labels 1 and 2, (2) interchange the maternal labels 3 and 4, (3) interchange the parental origin of the DNA by interchanging 1 and 3 and 2 and 4, and (4) permute the sibs. With this definition, the 16 possible inheritance vectors for a sib-pair are partitioned into three IBD configurations, corresponding to the number of chromosomes sharing DNA IBD at the locus of interest (cf. Table 1). In general, inheritance vectors may be partitioned in various ways for different purposes, and we address this question in greater detail for sibships in Section 3.

For a pedigree with given phenotypes, the conditional probability vector of IBD configurations at a genetic marker $\mathcal{M}$ linked to a DS locus $\mathcal{D}$ at recombination fraction θ is given by

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

where π is the conditional probability vector of IBD configurations at the DS locus (possibly one of several unlinked DS loci), m is the number of IBD configurations and $T(\theta)$ is the transition matrix between IBD configurations θ apart. In general, π depends on unknown and numerous genetic parameters, such as penetrances and genotype frequencies. In this paper, we consider a general genetic model with multiple genes unlinked to each other, arbitrary penetrances, and no population genetic assumptions such as random mating or Hardy–Weinberg equilibrium. Under the null hypothesis of no linkage between the marker and the DS locus, the IBD distribution at the marker is

$$\alpha = \rho\left(\frac{1}{2}, \pi\right) = \frac{1}{2^{2k}}(|\mathcal{C}_1|, \dots, |\mathcal{C}_m|),$$

where $|\mathcal{C}_i|$ is the number of inheritance vectors in $\mathcal{C}_i$, the i th IBD configuration.

Thus, the IBD probabilities at the marker have two separate components: one component involving the recombination fraction θ between the marker and the DS locus (*scale*), the other depending on the mode of inheritance of

TABLE 1
Sib-pair IBD configurations

Orbits of $S_2 \times D_4$	Orbits of $S_2 \times (C_2 \times C_2)$	Inheritance vectors
0 IBD	0 IBD	(1, 3, 2, 4), (1, 4, 2, 3), (2, 3, 1, 4), (2, 4, 1, 3)
1 IBD	1 paternal IBD	(1, 3, 1, 4), (1, 4, 1, 3), (2, 3, 2, 4), (2, 4, 2, 3)
	1 maternal IBD	(1, 3, 2, 3), (1, 4, 2, 4), (2, 3, 1, 3), (2, 4, 1, 4)
2 IBD	2 IBD	(1, 3, 1, 3), (1, 4, 1, 4), (2, 3, 2, 3), (2, 4, 2, 4)

S. DUDOIT AND T. P. SPEED

the disease (*specificity*). Our score test in the recombination fraction θ focuses on the scale component and seems to achieve some robustness with respect to the specificity (π). Examples of the transition matrix $T(\theta)$ are given in Section 5 for sib-pairs and sib-trios.

Suppose we collect data on n sibships of given size and phenotypes and wish to test the null hypothesis of no linkage between a genetic marker and a DS locus. Denote by N_i the number of sibships with IBD configuration i , $i = 1, \dots, m$, at the genetic marker. Under certain sampling assumptions [Dudoit (1990)], $(N_1, \dots, N_m)$ have a Multinomial($n, \rho(\theta, \pi)$) distribution. There is no uniformly most powerful test of $H_0: \theta = \frac{1}{2}$, however, the *score test* is *locally most powerful*. Moreover, for affected-only sibships of a given size, the score statistic does not involve the nuisance parameter π and reduces to a widely used statistic in linkage analysis, S_{pairs} , which is obtained by forming all possible pairs of sibs and averaging the proportions of chromosomes on which they share DNA IBD at the marker [Whittemore and Halpern (1994a) and Kruglyak, Daly, Reeve-Daly and Lander (1996)]. This result is a corollary to Theorem 2 in Section 4.

COROLLARY 1. *For affected sib- k -tuples, using the IBD configurations of Ethier and Hodge, the score test of $H_0: \theta = \frac{1}{2}$ is based on S_{pairs} , regardless of the model for disease susceptibility, that is, regardless of π . For one affected sib- k -tuple,*

$$S_{\text{pairs}} = \frac{\sum_{i < j} S_{ij}}{k(k-1)},$$

where S_{ij} is the number of chromosomes on which the ij th sib-pair shares DNA IBD. Under the null hypothesis of no linkage, the S_{ij} 's are pairwise independent Binomial($2, \frac{1}{2}$) random variables, and thus

$$E_0[S_{\text{pairs}}] = \frac{1}{2}, \quad \text{Var}_0[S_{\text{pairs}}] = \frac{1}{4k(k-1)}.$$

For a collection of affected sib- k -tuples, S_{pairs} is summed over all sibships.

Thus, for affected-only sibships of a *given* size, S_{pairs} is locally most powerful (in θ), and may be calculated easily by considering each sib-pair one at a time and without the need for assigning sibships to IBD configurations. This finding extends the work of Knapp, Seuchter and Baur (1994a) to sibships of any size and to general genetic models with multiple unlinked DS loci and no population genetic assumptions such as random mating or Hardy-Weinberg equilibrium. Unfortunately, this simple property does not hold with all types of sibships, and we consider examples where it fails in Sections 4 and 5. Furthermore, we show that the linkage score statistic combining IBD data from sibships of *different* sizes assigns different weights to the various sibship types and these weights depend on the genetic model.

The remainder of this paper is concerned with the proof of Corollary 1 and with deriving score statistics for general sibships, with any number of affected

A SCORE TEST FOR LINKAGE

and unaffected sibs, and distinguishing the parental origin of the DNA. In general, the form of the score statistic is based on properties of the transition matrix $T(\theta)$, which in turn are determined by the pedigree type and the choice of IBD configurations. Thus, we first describe how sibship inheritance patterns may be summarized by IBD configurations which are orbits of groups acting on the set of inheritance vectors.

3. Transition matrix for IBD configurations.

3.1. Sibship IBD configurations. Let $a = (1, 3)$, $b = (1, 4)$, $c = (2, 3)$ and $d = (2, 4)$ denote all four possible segregation outcomes at a particular locus for a given sib. Then we may think of the set of inheritance vectors $\mathcal{X}$ as the set of mappings $x: \{1, \dots, k\} \rightarrow \{a, b, c, d\}$. In this setting, the IBD configurations are *orbits* of groups acting on $\mathcal{X}$, where the groups are determined by the type of operations allowed within IBD configurations [cf. Fraleigh (1989), Section 3.2 for an introduction to group action]. Let

- $\alpha = (ac)(bd)$ interchange labels 1 and 2 of paternal chromosomes,
- $\beta = (ab)(cd)$ interchange labels 3 and 4 of maternal chromosomes,
- $\gamma = (bc)$ interchange parental origin of DNA.

The group of permutations generated by α , β and γ is actually the *dihedral group*, D_4 (α and γ are sufficient to generate D_4), and the group generated by α and β is the *Klein four-group*, $C_2 \times C_2$. Figure 3 displays a square with vertices a, b, c , and d . Permutations α and β correspond to mirror images in the perpendicular bisectors of the sides, and permutations γ corresponds to a diagonal flip. The dihedral group D_4 is the group of symmetries of the square. The IBD configurations of Ethier and Hodge (1985) for affected-only sibships are the orbits of the direct product $S_k \times D_4$, of the *symmetric group* S_k on k letters and the dihedral group D_4 , acting on $\mathcal{X}$. In some situations (e.g., parental

FIG. 3. Permutations α , β and γ of the vertices of the square.

S. DUDOIT AND T. P. SPEED

imprinting, when disease susceptibility is different for maternally and paternally inherited disease alleles), it may be appropriate to distinguish between sharing of maternal and paternal DNA and exclude the group operation γ . For example, for a sib-pair, it may be appropriate to distinguish between the two inheritance vectors (1, 3, 1, 4) and (1, 3, 2, 3); for (1, 3, 1, 4) the sibs share DNA IBD on the paternal chromosome, whereas for (1, 3, 2, 3) the sibs share DNA IBD on the maternal chromosome (cf. Table 1). For sibships with both affected and unaffected individuals, similar configurations may be defined, but the sibs are permuted only among affecteds or unaffecteds. The different types of group action are listed in Table 2.

For a group $G \times H$ acting on the set $\mathcal{X}$, let m denote the number of orbits and $\mathcal{C}_i$ denote the i th orbit. In general, we may use the *Pólya theory of counting* [van Lint and Wilson (1992) and deBruijn (1964)] to find the number of orbits of groups acting on mappings, and hence determine the number of IBD configurations of each type (see Appendix, Section A). Ethier and Hodge derived the number of IBD configurations of affected sib- k -tuples, as well as the number of inheritance vectors in each IBD configuration, without reference to the group $S_k \times D_4$. Instead, they based their calculations on labels for the equivalence classes which are triples of integers [cf. pages 264, 265 in Ethier and Hodge (1985) and Appendix, Section A].

3.2. Properties of transition matrix. In this section, we derive properties of transition matrices between IBD configurations. While Theorems 1 and 2 are specific to the sibship IBD configurations defined in the previous section, Propositions 1 and 2 hold more generally for any type of pedigree, with IBD configurations defined as orbits of groups acting on the set of inheritance vectors. The transition matrix $T(\theta)$ between IBD configurations at loci separated by a recombination fraction θ is the $m \times m$ matrix with entries

$$t_{ij}(\theta) = \frac{1}{|\mathcal{C}_i|} \sum_{x \in \mathcal{C}_i} \sum_{y \in \mathcal{C}_j} r_{xy}(\theta) = \frac{1}{|\mathcal{C}_i|} \sum_{x \in \mathcal{C}_i} \sum_{y \in \mathcal{C}_j} \theta^{\Delta(x,y)} (1 - \theta)^{2k - \Delta(x,y)}.$$

However, given any two inheritance vectors in $\mathcal{C}_i$, the probability of a transition to $\mathcal{C}_j$ is the same, that is,

$$(3.1) \quad \sum_{y \in \mathcal{C}_j} r_{xy}(\theta) = \sum_{y \in \mathcal{C}_j} r_{\tilde{x}y}(\theta) \quad \text{for any } x, \tilde{x} \in \mathcal{C}_i.$$

TABLE 2
Sibship IBD configurations

# affected, # unaffected	Distinguish maternal from paternal sharing	Group $G \times H$
$k, 0$	NO	$S_k \times D_4$
	YES	$S_k \times (C_2 \times C_2)$
$h, k - h$	NO	$(S_h \times S_{k-h}) \times D_4$
	YES	$(S_h \times S_{k-h}) \times (C_2 \times C_2)$

A SCORE TEST FOR LINKAGE

This result follows by observing that if $\sim$ denotes the operation applied to x to obtain $\tilde{x}$, then $\Delta(\tilde{x}, y) = \Delta(\tilde{x}, \tilde{y}) = \Delta(x, \tilde{y})$ and $\tilde{y} \in \mathcal{C}_j$. Consequently,

$$(3.2) \quad \begin{aligned} t_{ij}(\theta) &= \sum_{y \in \mathcal{C}_j} \theta^{\Delta(x, y)} (1 - \theta)^{2k - \Delta(x, y)} \quad \text{where } x \text{ is any } x \in \mathcal{C}_i \\ &= \frac{|\mathcal{C}_j|}{|\mathcal{C}_i|} \sum_{x \in \mathcal{C}_i} \theta^{\Delta(x, y)} (1 - \theta)^{2k - \Delta(x, y)} \quad \text{where } y \text{ is any } y \in \mathcal{C}_j. \end{aligned}$$

The next two propositions relate the transition matrix $T(\theta)$ to the adjacency matrix of a quotient graph, whose eigenvalues are key in determining the form of the score statistic (see Appendix, Section B for proofs).

PROPOSITION 1. *Let $T(\theta)$ denote the transition matrix between IBD configurations which are orbits of groups acting on the set of inheritance vectors $\mathcal{X}$, and let $\theta_1 * \theta_2 = \theta_1(1 - \theta_2) + \theta_2(1 - \theta_1)$. Then, $T(\theta)$ satisfies the semigroup property*

$$T(\theta_1 * \theta_2) = T(\theta_1)T(\theta_2).$$

Thus, $T(\theta)$ may be written as

$$T(\theta) = e^{d(\theta)Q},$$

where $d(\theta) = -\frac{1}{2} \ln(1 - 2\theta)$ is the inverse of the Haldane map function and Q is the infinitesimal generator. The infinitesimal generator is given by

$$Q = B - 2kI,$$

where B is the $m \times m$ matrix with entries

$$b_{ij} = \sum_{y \in \mathcal{C}_j} I(\Delta(x, y) = 1) \quad \text{for any } x \in \mathcal{C}_i$$

$\Delta(x, y)$ is the number of coordinates at which x and y differ and $I(\cdot)$ denotes the indicator function. The stationary distribution of $T(\theta)$ is

$$\alpha = (\alpha_1, \dots, \alpha_m) = \frac{1}{2^{2k}} (|\mathcal{C}_1|, \dots, |\mathcal{C}_m|)$$

and $T(\theta)$ is reversible, that is,

$$\alpha_i t_{ij}(\theta) = \alpha_j t_{ji}(\theta).$$

Hence, for one segregation, the crossover process is embeddable in a continuous-time random walk on $\{0, 1\}$, where 0 and 1 denote, respectively, the transmission of paternal and maternal DNA to one's child, and the time parameter is $d(\theta) = -\frac{1}{2} \ln(1 - 2\theta)$. Jointly, the crossover processes are i.i.d. and hence embeddable in a continuous-time random walk on the vertices of the hypercube $\{0, 1\}^{2k}$ (cf. Donnelly (1983) and Figures 1 and 2). The

S. DUDOIT AND T. P. SPEED

random walk model for the crossover process is widely used and is referred to in the genetics literature as the *no interference model*. The *Haldane map function* relates the recombination fraction to the map distance under the no interference model. The *map distance* between two loci is the expected number of crossover events, that is, of changes of the grand-parental origin of the DNA, occurring on a segregation product in the chromosomal interval between the two loci. Under the no interference model, the crossover process on individual segregation products is a Poisson process with intensity 1 [cf. Ott (1991), McPeck (1996) and Speed (1996) for an introduction to map functions and a more detailed discussion of crossover processes]. Note that, if we have three ordered loci, and θ_1 and θ_2 are the recombination fractions between the first and second and second and third locus, respectively, then $\theta_1 * \theta_2$ is the recombination fraction between the first and third locus, under the assumption that recombination events in disjoint intervals are independent, that is, there is no crossover interference. Also, note that we did not need to assume no crossover interference to derive the semigroup property. If, however, we do assume no crossover interference, then the inheritance vectors along a chromosome form a continuous-time Markov chain with time parameter the genetic distance along a chromosome. From (3.1) and condition (15), page 63 in Rosenblatt (1974), it follows that the IBD configurations also form a continuous-time Markov chain.

In order to compute score statistics, we need derivatives of the transition matrix at $\theta = \frac{1}{2}$. These may be computed by differentiating (3.2); however, we gain more knowledge on the transition matrix $T(\theta)$ and on the form of the score statistic by using the following spectral decomposition of $T(\theta)$.

PROPOSITION 2. *Let $T(\theta)$ denote the transition matrix between IBD configurations which are orbits of groups acting on the set of inheritance vectors $\mathcal{X}$. Then $T(\theta)$ may be written as*

$$(3.3) \quad T(\theta) = \sum_h \exp(\lambda_h d(\theta)) P_h = \sum_h (1 - 2\theta)^{-\lambda_h/2} P_h,$$

where λ_h are the m real eigenvalues of the infinitesimal generator Q , and P_h are projection matrices satisfying $P_h^2 = P_h = P_h^*$, $P_h P_l = 0$, $h \neq l$, and $\sum_h P_h = I$. Here P_h^* is the adjoint of P_h with respect to the inner product $\langle x, y \rangle_\alpha = \sum_i \alpha_i x_i y_i$. The ij th entry of P_h is $\alpha_j v_{ih} v_{jh}$, where v_{ih} is the i th entry of the right eigenvector $\mathbf{v}_h$ of Q corresponding to λ_h , and the eigenvectors $\mathbf{v}_h$ are orthonormal with respect to the inner product $\langle \cdot, \cdot \rangle_\alpha$. In particular, the first two derivatives of the transition matrix with respect to θ are

$$(3.4) \quad T'(\theta) = \sum_h \lambda_h (1 - 2\theta)^{-(\lambda_h+2)/2} P_h$$

and

$$(3.5) \quad T''(\theta) = \sum_h \lambda_h (\lambda_h + 2) (1 - 2\theta)^{-(\lambda_h+4)/2} P_h.$$

A SCORE TEST FOR LINKAGE

Thus, eigenvalues of Q and their multiplicity give us information regarding the derivatives of the transition matrix $T(\theta)$ and hence, the form of the score statistic. In particular, powers of θ in $T(\theta)$ are determined by the eigenvalues of Q , and the first nonzero derivative of $T(\theta)$ at $\theta = \frac{1}{2}$ and its rank are determined by the second largest eigenvalue of Q and its multiplicity. We will relate Q to the adjacency matrix of a quotient graph in order to derive its eigenvalues. Consider the graph $\mathcal{X}$ with vertex set the set of all inheritance vectors of length $2k$ and adjacency matrix $A(\mathcal{X}) = A$ with (x, y) -entry,

$$a_{xy} = \begin{cases} 1, & \text{if } \Delta(x, y) = 1, \\ 0, & \text{otherwise.} \end{cases}$$

Here $\mathcal{X}$ is the graph defined by the first associates in the Hamming scheme $H(2k, 2)$ [Chapter 30 in van Lint and Wilson (1992)]. Consider any of the four groups $G \times H$ described in Table 2. The matrix B , defined in Proposition 1, is the adjacency matrix of the quotient graph $\mathcal{X}/(G \times H)$, which is the multigraph with the orbits of $G \times H$ as its vertices and with b_{ij} arcs going from $\mathcal{C}_i$ to $\mathcal{C}_j$. Recall that $Q = B - 2kI$; consequently, we may work with B to derive the eigenvalues of Q . The following theorem relies on general facts concerning eigenvectors and eigenvalues of adjacency matrices of quotient graphs, as well as specific facts regarding the behavior of eigenvectors of A on the orbits of the four groups $G \times H$ described in Section 3.1 (see Appendix, Section C for proof).

THEOREM 1 (Eigenvalues of infinitesimal generator Q for sibship IBD configurations). *The largest eigenvalue of Q is 0, with multiplicity one, and the second largest eigenvalue of Q is -4 , with multiplicity depending on the group $G \times H$ defining the IBD configurations.*

- (a) $S_k \times D_4$: -4 has multiplicity one;
- (b) $S_k \times (C_2 \times C_2)$: -4 has multiplicity two;
and for $k \geq 3$
- (c) $(S_h \times S_{k-h}) \times D_4$: -4 has multiplicity two if $h = 1$ or $h = k - 1$, and three if $2 \leq h \leq k - 2$;
- (d) $(S_h \times S_{k-h}) \times (C_2 \times C_2)$: -4 has multiplicity four if $h = 1$ or $h = k - 1$, and six if $2 \leq h \leq k - 2$.

Furthermore, all other eigenvalues of Q belong to the set $\{-2i \binom{2k}{i} : i = 3, \dots, 2k\}$, where the subscript $\binom{2k}{i}$ is the largest possible multiplicity of the eigenvalue $-2i$. Thus, from (3.4) and (3.5),

$$(3.6) \quad T'(\tfrac{1}{2}) = 0$$

and

$$(3.7) \quad U = T''(\tfrac{1}{2}) = 8P_{-4},$$

where P_{-4} is the projection matrix for the second largest eigenvalue, -4 , with rank the multiplicity of -4 . In general, the ij th entry of P_{-4} is $\alpha_j \sum v_i v_j$,

S. DUDOIT AND T. P. SPEED

where the v 's are the right orthonormal (with respect to the inner product $\langle \cdot, \cdot \rangle_\alpha$) eigenvectors of Q corresponding to -4 , and the sum is over all such eigenvectors.

Note that we may also show that $T'(\frac{1}{2}) = 0$ by simple algebra, but this approach does not yield any particular insight into other general properties of $T(\theta)$.

The projection matrix for the largest eigenvalue of Q , $\lambda_1 = 0$, is $T(\frac{1}{2})$, the matrix whose rows are equal to the stationary distribution α . From Proposition 2 and Theorem 1, transition matrices for sibship IBD configurations have the form

$$T(\theta) = T(\frac{1}{2}) + (1 - 2\theta)^2 P_{-4} + o((1 - 2\theta)^2),$$

and the rate of convergence of $T(\theta)$ to $T(\frac{1}{2})$ as $\theta \rightarrow \frac{1}{2}$ is $O((1 - 2\theta)^2)$. Under the no interference model, the rate of convergence in terms of the map distance $d = -\ln(1 - 2\theta)/2$ is $O(e^{-4d})$.

More generally, since the matrix $R(\theta)$ has the same form for any type of pedigree, Propositions 1, 2 and 3 apply to arbitrary pedigrees, as long as the IBD configurations are defined as orbits of groups. Thus, for $2k$ segregations, the transition matrix for IBD configurations has the general form

$$T(\theta) = \sum_h (1 - 2\theta)^{-\lambda_h/2} P_h,$$

where the eigenvalues λ_h of Q belong to the set $\{-2i \binom{2k}{i} : i = 0, \dots, 2k\}$. The i th derivative, $i = 0, \dots, 2k$, of $T(\theta)$ is given by

$$T^{(i)}(\theta) = \sum_h \left\{ \prod_{j=0}^{i-1} (\lambda_h + 2j) \right\} (1 - 2\theta)^{-(\lambda_h + 2i)/2} P_h.$$

The first nonzero derivative of $T(\theta)$ at $\theta = \frac{1}{2}$ and its rank are determined by the second largest eigenvalue of Q , λ_2 , and its multiplicity. If $\lambda_2 = -2i$, the first nonzero derivative is the i th derivative,

$$T^{(i)}(\frac{1}{2}) = (-2)^i i! P_{-2i},$$

and the rank of this i th derivative is the multiplicity of $\lambda_2 = -2i$. Furthermore, the rate of convergence to $T(\frac{1}{2})$ is determined by the second largest eigenvalue,

$$T(\theta) = T(\frac{1}{2}) + (1 - 2\theta)^{-\lambda_2/2} P_{\lambda_2} + o((1 - 2\theta)^{-\lambda_2/2}).$$

The smaller the second largest eigenvalue, the faster the convergence to $T(\frac{1}{2})$. We proved that for sibships, the second largest eigenvalue is -4 , but this is not the case for all types of relatives. It turns out that for grandparent/grandchild pairs, who can share DNA IBD on either 0 or 1 chromosome, the second largest eigenvalue is -2 [Dudoit (1999)]. It is noteworthy that for some types of relative pairs (e.g., cousin pairs), the usual IBD configurations (0 or 1 IBD) are not orbits of groups and the transition matrix for these usual configura-

A SCORE TEST FOR LINKAGE

tions does not satisfy the semigroup property [Donnelly (1983) and Dudoit (1999)].

In the next section, we will explore the implications of Theorem 1 on score tests of the null hypothesis of no linkage between a marker and a gene using IBD data from sibships.

4. Linkage score test for sibships. Suppose we have data on n sibships of a given type (e.g., affected sib- k -tuples with orbits of $S_k \times D_4$), in the form of multinomial counts N_i , $i = 1, \dots, m$, for the number of sibships with IBD configuration i at a marker $\mathcal{M}$. We wish to test the null hypothesis of no linkage between the marker $\mathcal{M}$ and a DS locus $\mathcal{D}$, which could be one of several DS loci unlinked to each other; that is, we wish to test

$$H_0: \theta = \frac{1}{2} \text{ (no linkage)} \quad \text{versus} \quad H_1: 0 \leq \theta < \frac{1}{2} \text{ (linkage)},$$

where θ denotes the recombination fraction between $\mathcal{M}$ and $\mathcal{D}$. Note that $\theta = \frac{1}{2}$ and $\pi = \alpha$ are not identifiable.

The log-likelihood of the IBD data, conditional on the phenotypes, is

$$l(\theta, \pi) = \sum_i N_i \ln(\rho_i(\theta, \pi)),$$

where

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

The *score test* is based on the first nonzero derivative in the Taylor series expansion of the log-likelihood about $\theta = \frac{1}{2}$. In our problem, the first derivative vanishes, so we turn to the second derivative of the log-likelihood with respect to θ , which yields a test that maximizes the second derivative of the power function at the null. We find the score statistic for the given sibship type to be

$$S = \frac{\partial^2 l(\theta, \pi)}{\partial \theta^2} \bigg|_{\theta=1/2} = \sum_{i=1}^m N_i \frac{\partial^2 \rho_i(\theta, \pi) / \partial \theta^2}{\rho_i(\theta, \pi)} \bigg|_{\theta=1/2} = \sum_{i=1}^m N_i \frac{\sum_{j=1}^m \pi_j u_{ji}}{\alpha_i},$$

where $U = T''(\frac{1}{2}) = 8P_{-4} = (8\alpha_j \sum v_i v_j)$, the v 's are right orthonormal (with respect to $\langle \cdot, \cdot \rangle_\alpha$) eigenvectors of Q corresponding to the eigenvalue -4 , and the sum in U is over all such eigenvectors. The null hypothesis of no linkage is rejected for large values of the score statistic S . We show next that for affected sib- k -tuples, with the orbits of $S_k \times D_4$, the second largest eigenvalue has multiplicity one, and as a result, the score statistic for affected-only sibships of a given size is independent of the nuisance parameter π .

4.1. Affected sib- k -tuples, orbits of $S_k \times D_4$. A very widely used statistic in linkage analysis is S_{pairs} [cf. Kruglyak, Daly, Reeve-Daly and Lander (1996), S_P of Whittemore and Halpern (1994a), and *PAIRS* and *WP* of Suarez and van Eerdewegh (1984)]. For a sibship of size k , S_{pairs} is obtained by forming all

S. DUDOIT AND T. P. SPEED

possible pairs of sibs and averaging the proportions of chromosomes on which they share DNA IBD at the locus of interest; that is,

$$S_{\text{pairs}} = \frac{\sum_{i < j} S_{ij}}{k(k-1)},$$

where S_{ij} is the number of chromosomes on which the ij th sib-pair shares DNA IBD. The corollary in Section 2 results from the following theorem.

THEOREM 2 (Affected sib- k -tuple score statistic, orbits of $S_k \times D_4$). *Suppose we have IBD data on n affected sib- k -tuples, with IBD configurations defined as the orbits of $S_k \times D_4$. Then the score test of the null hypothesis of no linkage between a marker and a DS locus, $H_0: \theta = \frac{1}{2}$, is based on S_{pairs} . The contribution of n affected sib- k -tuples to the overall score statistic is*

$$\begin{aligned} S &= 8 \left(\sum_{j=1}^m v_j \pi_j \right) \left(\sum_{i=1}^m v_i N_i \right) \\ (4.1) \quad &= 2^{2k-2} \left(\sum_{j=1}^m u_{j1} \pi_j \right) (2S_{\text{pairs}} - n) \\ &= 8\sqrt{k(k-1)} \left(\sum_{j=1}^m v_j \pi_j \right) (2S_{\text{pairs}} - n), \end{aligned}$$

where u_{i1} is the i th entry of the first column of $U = T''(\frac{1}{2})$, v_i is the i th entry of the right eigenvector of Q corresponding to the eigenvalue -4 and $u_{i1} = 2^{5-2k} \sqrt{k(k-1)} v_i$. For n sibships, S_{pairs} is summed over all sibships.

The proof of Theorem 2 may be found in the Appendix, Section D, and relies on Theorem 1 and the following identity. For a sibship with inheritance vector x ,

$$\begin{aligned} (4.2) \quad S_{\text{pairs}} &= \frac{\sum_{i=1}^4 a_i(x)(a_i(x) - 1)}{2k(k-1)} \\ &= \frac{a_1(x)^2 + a_2(x)^2 + a_3(x)^2 + a_4(x)^2 - 2k}{2k(k-1)}, \end{aligned}$$

where $a_i(x)$ is the number of i labels in the inheritance vector x of the sibship, $i = 1, 2, 3, 4$, and $a_1(x) + a_2(x) + a_3(x) + a_4(x) = 2k$. Without loss of generality, we let the first IBD configuration be the one for which all sibs inherited the same maternal and paternal DNA, that is, with representative inheritance vector $(1, 3, 1, 3, \dots, 1, 3)$ and label $(0, 0, 0)$ in the notation of Ethier and Hodge (1985). The entries of the first column of U are easily computed, as seen in the proof.

Thus, for affected sib- k -tuples and without distinguishing between sharing of maternal and paternal DNA, the score test is based on S_{pairs} , regardless

A SCORE TEST FOR LINKAGE

of the genetic model, and may be calculated easily by considering each sib-pair one at a time and without the need for assigning sibships to IBD configurations. For commonly studied genetic models, the affected sib-pair “possible triangle” constraints hold [Dudoit and Speed (1999)], and as a result, $\sum_j v_j \pi_j \geq 0$. This follows by noting that $n \sum_j v_j \pi_j$ is the expected value of $\sum_j v_j N_j = \sqrt{k(k-1)}(2S_{\text{pairs}} - n)$ when $\theta = 0$. Also, for each sibship, S_{pairs} is an average of sib-pair statistics, each with expected value $\pi_2 + \pi_1/2 \geq 1/2$, where in this expression π_i , $i = 0, 1, 2$, is the probability that an affected sib-pair shares DNA IBD on i chromosomes at the DS locus. For IBD data from sibships of a *given* size, it is thus appropriate to reject the null hypothesis of no linkage for large values of S_{pairs} . However, the score statistic for combining IBD data from sibships of *different* sizes involves weights which do depend on the genetic model ($\sum_i v_i \pi_i$).

4.2. *Affected sib- k -tuples, orbits of $S_k \times (C_2 \times C_2)$.* For affected sib- k -tuples and distinguishing between sharing of maternal and paternal DNA, the second largest eigenvalue of the infinitesimal generator Q , -4 , has multiplicity two (Theorem 1). Hence, the second derivative of the transition matrix at $\theta = \frac{1}{2}$ has rank 2 and entries

$$u_{ij} = 8\alpha_j(v_i v_j + \tilde{v}_i \tilde{v}_j),$$

where $v = (v_1, \dots, v_m)^T$ and $\tilde{v} = (\tilde{v}_1, \dots, \tilde{v}_m)^T$ are the right orthonormal (with respect to the inner product $\langle \cdot, \cdot \rangle_\alpha$) eigenvectors of Q corresponding to the second largest eigenvalue. These eigenvectors are based on V_e and V_o , respectively (see Appendix, Section C). The score statistic is given by

$$\begin{aligned} S &= \sum_{i=1}^m N_i \frac{\sum_{j=1}^m \pi_j 8\alpha_i(v_i v_j + \tilde{v}_i \tilde{v}_j)}{\alpha_i} \\ &= 8 \left(\sum_{j=1}^m v_j \pi_j \right) \left(\sum_{i=1}^m v_i N_i \right) + 8 \left(\sum_{j=1}^m \tilde{v}_j \pi_j \right) \left(\sum_{i=1}^m \tilde{v}_i N_i \right). \end{aligned}$$

Thus, in general, the score test depends on the parameters of the genetic model through π . However, in some situations (e.g., no imprinting), this score statistic reduces to S_{pairs} .

4.3. *Discordant sib- k -tuples.* For sibships of size at least 3, with both affected and unaffected individuals [orbits of $(S_h \times S_{k-h}) \times D_4$ or of $(S_h \times S_{k-h}) \times (C_2 \times C_2)$], the second largest eigenvalue of the infinitesimal generator has multiplicity at least two (Theorem 1). Hence, in general, the score statistic depends on the genetic model and is a sum of terms of the form $(\sum_{j=1}^m v_j \pi_j)(\sum_{i=1}^m v_i N_i)$, where the v 's are the right orthonormal eigenvectors of Q corresponding to the second largest eigenvalue.

In the next section, we consider the examples of sib-pairs and sib-trios and present the transition matrix $T(\theta)$, the infinitesimal generator Q and the score statistic for these sibship types.

S. DUDOIT AND T. P. SPEED

5. Examples.

5.1. *Sib-pairs, orbits of $S_2 \times D_4$.* For sib-pairs with either 0, 1 or 2 affected individuals and without distinguishing between sharing of maternal and paternal DNA, there are three distinct IBD configurations, labeled 0, 1, 2, according to the number of chromosomes sharing DNA IBD at the locus of interest. The transition matrix is

$$T(\theta) = \begin{bmatrix} \psi^2 & 2\psi\bar{\psi} & \bar{\psi}^2 \\ \psi\bar{\psi} & \psi^2 + \bar{\psi}^2 & \psi\bar{\psi} \\ \bar{\psi}^2 & 2\psi\bar{\psi} & \psi^2 \end{bmatrix},$$

where $\psi = \theta^2 + (1-\theta)^2$ and $\bar{\psi} = 1 - \psi$. Figure 4 is a barycentric representation of curves $\mathcal{C}_\pi(\theta) = \{\rho(\theta, \pi) = \pi T(\theta) : 0 \leq \theta \leq \frac{1}{2}\}$ for the sib-pair transition matrix $T(\theta)$ and for $\pi = (\pi_0, \pi_1, \pi_2)$ on the boundaries of the simplex. The infinitesimal generator is

$$Q = \begin{bmatrix} -4 & 4 & 0 \\ 2 & -4 & 2 \\ 0 & 4 & -4 \end{bmatrix}.$$

Q has eigenvalues $\lambda = 0, -4$ and -8 . The left and right eigenvectors of Q corresponding to $\lambda_2 = -4$ are $(1/2\sqrt{2}, 0, -1/2\sqrt{2})$ and $(\sqrt{2}, 0, -\sqrt{2})$, respectively

FIG. 4. $S_2 \times D_4$ -Barycentric representation of curves $\mathcal{C}_\pi(\theta)$, $0 \leq \theta \leq \frac{1}{2}$, for π on boundaries of simplex.

A SCORE TEST FOR LINKAGE

(right eigenvector has unit norm with respect to the inner product $\langle \cdot, \cdot \rangle_a$). Hence

$$U = 8P_{-4} = 8 \begin{bmatrix} \sqrt{2} \\ 0 \\ -\sqrt{2} \end{bmatrix} \left[\frac{1}{2\sqrt{2}}, 0, -\frac{1}{2\sqrt{2}} \right] = \begin{bmatrix} 4 & 0 & -4 \\ 0 & 0 & 0 \\ -4 & 0 & 4 \end{bmatrix}.$$

If we let N_i denote the number of affected sib-pairs sharing DNA IBD on i chromosomes at the marker, $i = 0, 1, 2$, then the score statistic for affected sib-pairs is

$$16(\pi_2 - \pi_0)(N_2 - N_0),$$

similarly for discordant and unaffected sib-pairs. Note that $N_2 - N_0$ may be rewritten as $N_2 - (n - N_1 - N_2) = 2(N_2 + \frac{1}{2}N_1) - n$. Thus, the score test is based on $S_{\text{pairs}} = N_2 + \frac{1}{2}N_1$, also known as the *mean IBD statistic*. These findings extend the work of Knapp, Seuchter and Baur (1994a), who proved the local optimality of the mean IBD statistic for affected sib-pairs under a single DS locus model with random mating and Hardy–Weinberg equilibrium. The score statistic for combining IBD data from affected, discordant and unaffected sib-pairs is a linear combination of the mean IBD statistics for each type of sib-pair, with weights $\pi_2 - \pi_0$ depending on the genetic model.

Note that this setup also applies to testing linkage to quantitative trait loci. In this case, each sib-pair may have different continuous phenotypes, and the score statistic is given by

$$16 \sum_i (\pi_{2i} - \pi_{0i})(N_{2i} - N_{0i}),$$

where $N_{ji} = 1$ if the i th sib-pair shares DNA IBD on $j = 0, 1, 2$ chromosomes at the marker and $N_{ji} = 0$ otherwise, and π_{ji} is the conditional probability that the i th sib-pair shares DNA IBD on j chromosomes at the trait locus given the phenotypes of the sibs [Dudoit (1999)].

5.2. Sib-pairs, orbits of $S_2 \times (C_2 \times C_2)$. For sib-pairs with any number of affecteds and distinguishing between sharing of maternal and paternal DNA, there are four distinct IBD configurations, conveniently labeled by the pair (i, j) , $i, j = 0, 1$, where i and j denote the number of paternally and maternally inherited chromosomes sharing DNA IBD, respectively. The transition matrix is

$$T(\theta) = \begin{bmatrix} \psi^2 & \psi\bar{\psi} & \psi\bar{\psi} & \bar{\psi}^2 \\ \psi\bar{\psi} & \psi^2 & \bar{\psi}^2 & \psi\bar{\psi} \\ \psi\bar{\psi} & \bar{\psi}^2 & \psi^2 & \psi\bar{\psi} \\ \bar{\psi}^2 & \psi\bar{\psi} & \psi\bar{\psi} & \psi^2 \end{bmatrix}$$

S. DUDOIT AND T. P. SPEED

and the infinitesimal generator is

$$Q = \begin{bmatrix} -4 & 2 & 2 & 0 \\ 2 & -4 & 0 & 2 \\ 2 & 0 & -4 & 2 \\ 0 & 2 & 2 & -4 \end{bmatrix}.$$

Q has eigenvalues $\lambda = 0, -4, -4$ and -8 . The two orthonormal right eigenvectors corresponding to $\lambda_2 = -4$ are $(\sqrt{2}, 0, 0, -\sqrt{2})$ and $(0, \sqrt{2}, -\sqrt{2}, 0)$, hence

$$\begin{aligned} U = 8P_{-4} &= 8 \begin{bmatrix} \sqrt{2} \\ 0 \\ 0 \\ -\sqrt{2} \end{bmatrix} \begin{bmatrix} \frac{1}{2\sqrt{2}}, 0, 0, -\frac{1}{2\sqrt{2}} \end{bmatrix} + 8 \begin{bmatrix} 0 \\ \sqrt{2} \\ -\sqrt{2} \\ 0 \end{bmatrix} \begin{bmatrix} 0, \frac{1}{2\sqrt{2}}, -\frac{1}{2\sqrt{2}}, 0 \end{bmatrix} \\ &= \begin{bmatrix} 4 & 0 & 0 & -4 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ -4 & 0 & 0 & 4 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 4 & -4 & 0 \\ 0 & -4 & 4 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}. \end{aligned}$$

Let N_{ij} denote the number of affected sib-pairs sharing DNA IBD on i paternal and j maternal chromosome at the marker, $i, j = 0, 1$. The score statistic is given by

$$16(\pi_{11} - \pi_{00})(N_{11} - N_{00}) + 16(\pi_{10} - \pi_{01})(N_{10} - N_{01}),$$

and in general depends on the genetic model, similarly for discordant and unaffected sib-pairs. When $\pi_{10} = \pi_{01}$, the score test is based on $N_{11} - N_{00}$, that is, $N_2 - N_0$ in the more usual notation.

5.3. Affected sib-trios, orbits of $S_3 \times D_4$. For affected sib-trios (ASTs), there are four IBD configurations, with representative inheritance vectors and labels [defined as in Ethier and Hodge (1985)] listed in Table 3.

TABLE 3
IBD configurations for affected sib-trios

IBD configuration i	Representative inheritance vector	Label	$ \mathcal{C}_i $
1	(1, 3, 1, 3, 1, 3)	(0, 0, 0)	4
2	(1, 3, 1, 3, 1, 4)	(0, 0, 1)	24
3	(1, 3, 1, 4, 2, 3)	(0, 1, 1)	24
4	(1, 3, 1, 3, 2, 4)	(1, 1, 1)	12

A SCORE TEST FOR LINKAGE

The transition matrix $T(\theta)$ is given by

$$\begin{bmatrix} (1-3\theta+3\theta^2)^2 & 6\theta\bar{\theta}(1-3\theta+3\theta^2) & 6\theta^2\bar{\theta}^2 & 3\theta^2\bar{\theta}^2 \\ \theta\bar{\theta}(1-3\theta+3\theta^2) & 1-4\theta+10\theta^2-12\theta^3+6\theta^4 & 2\theta\bar{\theta}(1-\theta+\theta^2) & \theta\bar{\theta}(1-\theta+\theta^2) \\ \theta^2\bar{\theta}^2 & 2\theta\bar{\theta}(1-\theta+\theta^2) & 1-4\theta+10\theta^2-12\theta^3+6\theta^4 & \theta\bar{\theta}(2-5\theta+5\theta^2) \\ \theta^2\bar{\theta}^2 & 2\theta\bar{\theta}(1-\theta+\theta^2) & 2\theta\bar{\theta}(2-5\theta+5\theta^2) & 1-6\theta+17\theta^2-22\theta^3+11\theta^4 \end{bmatrix}$$

and the infinitesimal generator is

$$Q = \begin{bmatrix} -6 & 6 & 0 & 0 \\ 1 & -4 & 2 & 1 \\ 0 & 2 & -4 & 2 \\ 0 & 2 & 4 & -6 \end{bmatrix}.$$

Q has eigenvalues $\lambda = 0, -4, -8, -8$, and the left and right eigenvectors corresponding to $\lambda_2 = -4$ are $\frac{1}{16}\sqrt{\frac{2}{3}}(3, 6, -6, -3)$ and $\sqrt{\frac{2}{3}}(3, 1, -1, -1)$, respectively. Hence

$$U = 8P_{-4} = 8\sqrt{\frac{2}{3}} \begin{bmatrix} 3 \\ 1 \\ -1 \\ -1 \end{bmatrix} \frac{1}{16}\sqrt{\frac{2}{3}}[3, 6, -6, -3] = \begin{bmatrix} 3 & 6 & -6 & -3 \\ 1 & 2 & -2 & -1 \\ -1 & -2 & 2 & 1 \\ -1 & -2 & 2 & 1 \end{bmatrix}.$$

Let N_i denote the number of ASTs with IBD configuration i at the marker, $i = 1, 2, 3, 4$. Then the score statistic for testing linkage is

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

Note that $3N_1 + N_2 - N_3 - N_4$ may be rewritten as $2(3N_1 + 2N_2 + N_3 + N_4) - 3n = 6S_{\text{pairs}} - 3n$.

5.4. Discordant sib-trios, orbits of $(S_1 \times S_2) \times D_4$. For discordant sib-trios (DSTs), where the first sib is the “odd” sib (i.e., the only affected sib or the only unaffected sib), there are seven IBD configurations, with representative inheritance vectors listed in Table 4.

TABLE 4
IBD configurations for discordant sib-trios

IBD configuration i	Representative inheritance vector	$ \mathcal{C}_i $
1	(1, 3, 1, 3, 1, 3)	4
2	(1, 3, 1, 3, 1, 4)	16
3	(1, 3, 1, 3, 2, 4)	8
4	(1, 3, 1, 4, 2, 3)	8
5	(1, 3, 1, 4, 2, 4)	16
6	(1, 3, 1, 4, 1, 4)	8
7	(1, 3, 2, 4, 2, 4)	4

S. DUDOIT AND T. P. SPEED

The infinitesimal generator is

$$Q = \begin{bmatrix} -6 & 4 & 0 & 0 & 0 & 2 & 0 \\ 1 & -5 & 1 & 1 & 1 & 1 & 0 \\ 0 & 2 & -6 & 2 & 2 & 0 & 0 \\ 0 & 2 & 2 & -6 & 2 & 0 & 0 \\ 0 & 1 & 1 & 1 & -5 & 1 & 1 \\ 1 & 2 & 0 & 0 & 2 & -6 & 1 \\ 0 & 0 & 0 & 0 & 4 & 2 & -6 \end{bmatrix}.$$

Q has eigenvalues $\lambda = 0, -4, -4, -8, -8, -8, -8$, and the two orthonormal right eigenvectors corresponding to $\lambda_2 = -4$ are $v = \sqrt{\frac{2}{3}}(-1, -1, -1, -1, 1, 1, 3)$ and $\tilde{v} = (1/\sqrt{3})(4, 1, -2, -2, -1, 2, 0)$. Hence

$$U = 8P_{-4} = \begin{bmatrix} 3 & 4 & -2 & -2 & -4 & 2 & -1 \\ 1 & 2 & 0 & 0 & -2 & 0 & -1 \\ -1 & 0 & 2 & 2 & 0 & -2 & -1 \\ -1 & 0 & 2 & 2 & 0 & -2 & -1 \\ -1 & -2 & 0 & 0 & 2 & 0 & 1 \\ 1 & 0 & -2 & -2 & 0 & 2 & 1 \\ -1 & -4 & -2 & -2 & 4 & 2 & 3 \end{bmatrix}.$$

Denote the i th column of U by u_i , then $u_1 + u_7 = -u_3 = -u_4 = u_6$ and $u_1 - u_7 = u_2 = -u_5$. Let N_i denote the number of DSTs with IBD configuration i at the marker, $i = 1, \dots, 7$. Then, the score statistic for testing linkage is

$$\begin{aligned} S &= 8 \left(\sum_{j=1}^7 v_j \pi_j \right) \left(\sum_{i=1}^7 v_i N_i \right) + 8 \left(\sum_{j=1}^7 \tilde{v}_j \pi_j \right) \left(\sum_{i=1}^7 \tilde{v}_i N_i \right) \\ &= \frac{8}{3} \left(2(-\pi_1 - \pi_2 - \pi_3 - \pi_4 + \pi_5 + \pi_6 + 3\pi_7) \right. \\ &\quad \times (-N_1 - N_2 - N_3 - N_4 + N_5 + N_6 + 3N_7) \\ &\quad + (4\pi_1 + \pi_2 - 2\pi_3 - 2\pi_4 - \pi_5 + 2\pi_6) \\ &\quad \left. \times (4N_1 + N_2 - 2N_3 - 2N_4 - N_5 + 2N_6) \right). \end{aligned}$$

6. Discussion. In this paper, we have derived score statistics for testing the null hypothesis of no linkage between a marker and a disease gene using identity by descent (IBD) data from sibships. We considered IBD configurations which are orbits of groups acting on the set of inheritance vectors, and proved that the transition matrix between IBD configurations satisfies a semi-group property (Proposition 1). For general pedigree types, we derived the

A SCORE TEST FOR LINKAGE

following spectral decomposition for the IBD configuration transition matrix $T(\theta)$:

$$T(\theta) = \sum_h (1 - 2\theta)^{-\lambda_h/2} P_h,$$

where λ_h and P_h are the eigenvalues and projection matrices of the infinitesimal generator Q , respectively, and λ_h are negative even integers (Propositions 2 and 3). By relating Q to the adjacency matrix of a quotient graph, we derived properties of its eigenvalues and eigenvectors. In general, the second largest eigenvalue of Q and its multiplicity determine the form of the score statistic for a given pedigree type. If the second largest eigenvalue of Q is $-2i$, the score test is based on the i th derivative of the log-likelihood, and if it has multiplicity one, the score statistic is independent of the genetic model (i.e., of the nuisance parameter π). For affected-only sibships of a given size, the second largest eigenvalue is -4 and has multiplicity one. As a result, the score test is based on the second derivative of the log-likelihood and is independent of the genetic model. Furthermore, the score statistic for affected-only sibships of a given size reduces to a well-known nonparametric statistic in linkage analysis, S_{pairs} .

Testing the null hypothesis of no linkage using IBD data, as we do here, is an instance of a general class of testing problems in which the null hypothesis is that a Markov chain has reached its stationary distribution [Diaconis (1988)]. The second largest eigenvalue of the infinitesimal generator not only determines the rate of convergence to the stationary distribution, but also plays an important role in hypothesis testing, as illustrated by our study.

Since the mid-1970's, linkage statistics have been the subject of numerous studies in the genetics literature. Earlier research focussed on IBD data from sib-pairs, and the sib-pair case still remains popular for theoretical work on linkage analysis and in practice for genetic studies. Several test statistics have been examined to test the null hypothesis of no linkage between a marker and a DS locus, based on the numbers (N_0, N_1, N_2) of affected sib-pairs sharing DNA IBD on 0, 1 and 2 chromosomes, respectively. These include the mean IBD statistic, that is, $S_{\text{pairs}} = N_2 + \frac{1}{2}N_1$ [Blackwelder and Elston (1985), Schaid and Nick (1990), Knapp, Seuchter and Baur (1994a, b)], N_2 [Day and Simons (1976), Blackwelder and Elston (1985) and Schaid and Nick (1990)], $N_2 + \frac{1}{4}N_1$ [Feingold and Siegmund (1997)], likelihood ratio statistics and χ^2 goodness-of-fit statistics, either unrestricted or restricted to the "possible triangle" [Risch (1990a, b), Holmans (1993), Faraway (1993), Feingold, Brown and Siegmund (1993), Holmans and Clayton (1995), Knapp, Seuchter and Baur (1994c) and Kruglyak and Lander (1995)]. Knapp, Seuchter and Baur (1994a) first pointed out the relationship of S_{pairs} to likelihood-based tests, in the special case of affected sib-pairs and genetic models with a single disease locus with random mating and Hardy-Weinberg equilibrium. In the past five years, nonparametric allele-sharing statistics for small pedigrees have been actively studied. Current methods for analyzing IBD data from small pedigrees rely on "scoring functions" which are defined as functions of the IBD configurations of affected pedigree members. [Whittemore and Halpern

S. DUDOIT AND T. P. SPEED

(1994a), Kruglyak, Daly, Reeve-Daly and Lander (1996), Kong and Cox (1997), Teng and Siegmund (1997) and McPeck (1999)]. A few scoring functions have been suggested on empirical grounds, such as S_{pairs} and S_{all} [Whittemore and Halpern (1994a)], but the optimality and relationship of these scoring functions to likelihood analysis deserve further study. In addition, while it is clear that different pedigree types differ in their linkage information, IBD data from different pedigrees are typically combined by assigning equal weights to the standardized allele-sharing statistics of the various pedigrees. Different standardization and weighing schemes have been suggested, but the issue of combining IBD data from different pedigree types has not yet been thoroughly examined in the literature and remains an open problem. Whittemore (1996) proposed a likelihood-based approach to linkage analysis using multi-point marker data from general pedigrees. In the context of a genome scan, Whittemore considered the null hypothesis of no DS gene linked to the candidate locus against the alternative hypothesis of a DS gene *at* the candidate locus ($\theta = 0$). The null and alternative hypotheses were expressed in terms of a specificity parameter β , which is a function of genetic parameters, such as penetrances and allele frequencies. Whittemore proposed testing the null hypothesis of no DS gene linked to the candidate locus using a score test in β , that is, in the *specificity* parameter. Like our score test in the recombination fraction θ , this is a likelihood-based test which simultaneously treats the IBD data from all pedigrees. However, this type of score test depends intrinsically on the parameterization of the genetic model. This is a potentially serious shortcoming for complex diseases, for which there typically is no established knowledge regarding the number of DS loci, the penetrances, or the frequency of genotypes in the study population. Similarly, McPeck (1999) derived allele-sharing statistics for different types of alternative genetic models by focusing on specificity parameters.

It is important to note that in genome scans, where marker density is high, one expects to be testing near $\theta = 0$ rather than near $\theta = \frac{1}{2}$. If one has some knowledge concerning the mode of inheritance of the disease (e.g., rare recessive disease), then making use of this information in deriving allele-sharing statistics as in McPeck (1999) is desirable. However, the main obstacle in obtaining optimal statistics for an alternative hypothesis where $\theta = 0$ is precisely their dependence on the genetic model, which is usually unknown. By contrast, our score test in the recombination fraction θ focuses on the *scale* component and seems to achieve some robustness with respect to specificity. Although our score statistic in θ is locally most powerful near $\theta = \frac{1}{2}$, S_{pairs} was also found to be optimal for some classes of genetic models when $\theta = 0$ by McPeck (1999). A recent study by Davis and Weeks (1997) on affected sib-pairs found S_{pairs} to perform well for a variety of two-locus genetic models. Our preliminary work on applying the sib-pair score test in θ to the linkage analysis of quantitative traits also demonstrates the power of this type of score test for nonlocal alternatives (Dudoit (1999)). However, when the alternative hypothesis is such that $\theta = 0$, the optimal statistic depends on the genetic model, and there are inevitably classes of models for which S_{pairs} is outperformed by

A SCORE TEST FOR LINKAGE

other statistics. For example, Feingold and Siegmund (1997) reported that for sib-pairs N_2 performed better than S_{pairs} for rare heterogeneous traits. For larger sibships, the performance of S_{pairs} needs to be further investigated.

As mentioned earlier, an important issue is the combination of IBD data from different pedigree types. The score test in θ presented in this paper produces weights for combining IBD data from different types of sibships ($\sum_j v_j \pi_j$), but these weights are linear combinations of the IBD probabilities, π , which depend on the usually unknown genetic model. Our work on quantitative traits revealed that the sib-pair score test was robust to the choice of model for computing the weights $\pi_2 - \pi_0$. For sib-pairs sampled randomly with respect to their phenotypes, the score test was found to be similar in power to the widely used Haseman–Elston test, and for sib-pairs selected with extreme concordant or discordant phenotypes, the score test was found to be far more powerful than the Haseman–Elston test [cf. Haseman and Elston (1972), Risch and Zhang (1995, 1996), Dudoit (1999)].

We believe that score statistics in θ as described in this paper have the potential to be useful in practice for linkage analysis using IBD data from sibships and relative pairs, when the mode of inheritance is unknown. We have derived score statistics in θ for IBD data from other types of relative pairs (e.g., cousin pairs), using the IBD configurations considered by Donnelly (1983) and the general framework of Propositions 1 and 2 [Dudoit (1999)]. These score statistics may readily be extended to accommodate incomplete IBD data by replacing the IBD counts by their expected value given multipoint marker data [cf. inheritance distribution of Kruglyak and Lander (1995) and Kruglyak, Daly, Reeve-Daly and Lander (1996)]. We were very encouraged by our preliminary studies with sib-pairs and will further explore the issue of combining IBD data from different sibship types and relative pairs.

APPENDIX

A. Pólya theory of counting. Let A and B be finite sets, $|A| = n$, and let G and H be finite groups, G acting on A and H on B . By Theorem 35.3 in van Lint and Wilson (1992), the number of orbits of $G \times H$ acting on B^A , the set of mappings from A to B , is given by

$$\frac{1}{|H|} \sum_{\tau \in H} Z_G(m_1(\tau), \dots, m_n(\tau)),$$

where

$$m_i(\tau) := \sum_{j|i} j z_j(\tau), \quad i = 1, \dots, n,$$

$$z_j(\tau) := \text{number of cycles of } \tau \text{ having length } j, \quad j = 1, \dots, |B|.$$

For a group G acting on a set of n elements, the *cycle index* Z_G is a polynomial in n letters, $X_1, \dots, X_n$, defined by

$$Z_G(X_1, \dots, X_n) := \frac{1}{|G|} \sum_{\sigma \in G} X_1^{z_1(\sigma)} \dots X_n^{z_n(\sigma)}.$$

S. DUDOIT AND T. P. SPEED

The cycle index for the symmetric group on n letters is

$$Z_{S_n}(X_1, \dots, X_n) = \sum_{(1^{k_1} \dots n^{k_n})} \frac{1}{1^{k_1} \dots n^{k_n} k_1! \dots k_n!} X_1^{k_1} \dots X_n^{k_n},$$

where $(1^{k_1} \dots n^{k_n})$ denotes a partition of n with k_i parts of size i , $i = 1, \dots, n$. $Z_{S_n}(X_1, \dots, X_n)$ is also the coefficient of z^n in the expansion of $\exp(\sum_{i=1}^{\infty} (z^i/i) X_i)$ [cf. deBruijn (1964), page 147]; this is the formula which is most appropriate for our problem. In our problem, we wish to determine the number of orbits of $G \times H$ acting on the set of mappings $\{a, b, c, d\}^{\{1, \dots, k\}}$ (i.e., the set of inheritance vectors), where $H = D_4$ or $C_2 \times C_2$ and $G = S_k$ or $S_h \times S_{k-h}$. Table 5 lists for each permutation in D_4 the number of cycles of length j , $j = 1, 2, 3, 4$. The m_i 's of Table 6 are calculated using Table 5, and the fact that for $\tau \in D_4$ and $i \geq 0$,

$$m_4(\tau) = m_{4i+4}(\tau),$$

$$m_1(\tau) = m_{4i+1}(\tau) = m_{4i+3}(\tau),$$

$$m_2(\tau) = m_{4i+2}(\tau).$$

Table 7 lists for each permutation in D_4 the cycle index $Z_{S_k}(m_1(\tau), \dots, m_k(\tau))$. When $H = C_2 \times C_2$, only the first and third rows of Table 7 are used. When $G = S_h \times S_{k-h}$, we note that the cycle index polynomial of the direct product of two groups is simply the product of the cycle indices of the two groups and Table 7 may be used again. The number of IBD configurations for the four groups are listed below.

$S_k \times D_4$:

$$m = \begin{cases} (k+1)(k+3)(k+5)/48, & k \text{ odd,} \\ (k+2)(k^2+7k+18)/48, & k \text{ even and } k/2 \text{ odd,} \\ (k+4)(k^2+5k+12)/48, & k \text{ even and } k/2 \text{ even,} \end{cases}$$

which agrees with equation (5) of Ethier and Hodge (1985).

TABLE 5
Cycles of D_4 . $z_j(\tau)$ denotes the number of cycles of τ having length j *

Permutation τ	$z_1(\tau)$	$z_2(\tau)$	$z_3(\tau)$	$z_4(\tau)$
ι	4	0	0	0
$\rho_1 = (cadb)$	0	0	0	1
$\rho_2 = (ad)(bc)$	0	2	0	0
$\rho_3 = (bdac)$	0	0	0	1
$\mu_1 = (ab)(cd)$	0	2	0	0
$\mu_2 = (ac)(bd)$	0	2	0	0
$\delta_1 = (bc)$	2	1	0	0
$\delta_2 = (ad)$	2	1	0	0

*The elements of D_4 are listed according to the notation of Fraleigh [(1989), page 70].

A SCORE TEST FOR LINKAGE

TABLE 6
 $m_i(\tau)$ for $\tau \in D_4$

Permutation τ	$m_{4i+1}(\tau) = m_{4i+3}(\tau)$	$m_{4i+2}(\tau)$	$m_{4i+4}(\tau)$
ι	4	4	4
ρ_1, ρ_3	0	0	4
ρ_2, μ_1, μ_2	0	4	4
δ_1, δ_2	2	4	4

$$S_k \times (C_2 \times C_2):$$

$$m = \begin{cases} (k+3)(k+2)(k+1)/24, & k \text{ odd,} \\ (k+2)(k^2+4k+12)/24, & k \text{ even.} \end{cases}$$

$$(S_h \times S_{k-h}) \times D_4:$$

$$m = \frac{1}{8} \left[\binom{h+3}{3} \binom{k-h+3}{3} + 2I(4|h)I(4|k-h) \right. \\ \left. + \frac{3}{4}I(2|h)I(2|k-h)(h+2)(k-h+2) \right. \\ \left. + \frac{1}{8}(I(2|h) + (h+1)(h+3))(I(2|k-h) + (k-h+1)(k-h+3)) \right],$$

where $I(j|i) = 1$ if j divides i and 0 otherwise.

$$(S_h \times S_{k-h}) \times (C_2 \times C_2):$$

$$m = \frac{1}{4} \left[\binom{h+3}{3} \binom{k-h+3}{3} + \frac{3}{4}I(2|h)I(2|k-h)(h+2)(k-h+2) \right].$$

For any inheritance vector x , let pat denote the less frequent of the paternal labels 1 and 2, and similarly let mat denote the less frequent of the maternal labels 3 and 4. The number of paternal labels pat in the inheritance vector x is denoted by $|pat|$, similarly for the maternal labels. The number of sibs with a pair of labels (pat, mat) is denoted by $|(pat, mat)|$. Ethier and Hodge (1985) define the label of the particular inheritance vector x to be the triple (l_1, l_2, l_3) where

$$l_1 = |(pat, mat)|, \quad l_2 = \min(|pat|, |mat|), \quad l_3 = \max(|pat|, |mat|).$$

TABLE 7
 $Z_{S_k}(m_1(\tau), \dots, m_k(\tau))$ for $\tau \in D_4$. $I(\cdot)$ is the indicator function

Permutation τ	$\exp\left(\sum_{i=1}^{\infty} m_i(\tau) \frac{z^i}{i}\right)$	$Z_{S_k}(m_1(\tau), \dots, m_k(\tau))$
ι	$(1-z)^{-4}$	$\binom{k+3}{3}$
ρ_1, ρ_3	$(1-z^4)^{-1}$	$I(4 k)$
ρ_2, μ_1, μ_2	$(1-z^2)^{-2}$	$I(2 k)(k+2)/2$
δ_1, δ_2	$(1-z)^{-2}(1-z^2)^{-1}$	$\frac{1}{4}(I(2 k) + (k+1)(k+3))$

S. DUDOIT AND T. P. SPEED

For example, if $x = (1, 3, 1, 4, 2, 3)$, the less frequent of the paternal labels is 2, thus $pat = 2$ and $|pat| = 1$. Similarly, $mat = 4$ and $|mat| = 1$. $(mat, pat) = (2, 4)$ and the number of sibs with pair of labels $(2, 4)$ is 0. Thus $l_1 = 0$, $l_2 = 1$ and $l_3 = 1$. In ambiguous cases such as $|1| = |2| = k/2$ or $|3| = |4| = k/2$, Ethier and Hodge suggest making a choice that results in $l_1 \geq l_2/2$. Then the triple satisfies

$$0 \leq l_1 \leq l_2 \leq l_3 \leq k/2, \quad \text{and} \quad l_1 \geq l_2/2 \text{ if } l_3 = k/2.$$

We can modify the labeling of Ethier and Hodge for the orbits of $S_k \times (C_2 \times C_2)$ and let

$$l_1 = |(pat, mat)|, \quad l_2 = |pat|, \quad l_3 = |mat|.$$

B. Transition matrix for IBD configurations.

B.1. Proof of Proposition 1. Let $\bar{\theta} = 1 - \theta$. We first prove the semigroup property for the transition matrix $R(\theta)$ of inheritance vectors. Let $\Delta = \Delta(x, y)$, then

$$\begin{aligned} r_{xy}(\theta_1 * \theta_2) &= (\theta_1 * \theta_2)^\Delta (1 - (\theta_1 * \theta_2))^{2k-\Delta} \\ &= (\theta_1 \bar{\theta}_2 + \bar{\theta}_1 \theta_2)^\Delta (\theta_1 \theta_2 + \bar{\theta}_1 \bar{\theta}_2)^{2k-\Delta} \\ &= \sum_{i=0}^{\Delta} \sum_{j=0}^{2k-\Delta} \binom{\Delta}{i} \binom{2k-\Delta}{j} \theta_1^{i+j} \bar{\theta}_1^{2k-(i+j)} \theta_2^{\Delta-i+j} \bar{\theta}_2^{2k-(\Delta-i+j)}. \end{aligned}$$

Also,

$$\sum_z r_{xz}(\theta_1) r_{zy}(\theta_2) = \sum_z \theta_1^{\Delta(x,z)} \bar{\theta}_1^{2k-\Delta(x,z)} \theta_2^{\Delta(y,z)} \bar{\theta}_2^{2k-\Delta(y,z)}.$$

Now, for $i = 0, \dots, \Delta$, $j = 0, \dots, 2k - \Delta$, divide the set of all 2^{2k} inheritance vectors into groups of $\binom{\Delta}{i} \binom{2k-\Delta}{j}$ inheritance vectors z , such that z differs from x at i of the Δ positions at which x and y differ and z differs from x at j of the $2k - \Delta$ positions at which x and y agree. Then, $\Delta(x, z) = i + j$, $\Delta(y, z) = (\Delta - i) + j$, and

$$\begin{aligned} \sum_z \theta_1^{\Delta(x,z)} \bar{\theta}_1^{2k-\Delta(x,z)} \theta_2^{\Delta(y,z)} \bar{\theta}_2^{2k-\Delta(y,z)} \\ = \sum_{i=0}^{\Delta} \sum_{j=0}^{2k-\Delta} \binom{\Delta}{i} \binom{2k-\Delta}{j} \theta_1^{i+j} \bar{\theta}_1^{2k-(i+j)} \theta_2^{\Delta-i+j} \bar{\theta}_2^{2k-(\Delta-i+j)}. \end{aligned}$$

Therefore,

$$r_{xy}(\theta_1 * \theta_2) = \sum_z r_{xz}(\theta_1) r_{zy}(\theta_2).$$

A SCORE TEST FOR LINKAGE

Consider now the transition matrix for IBD configurations. From (3.2),

$$\begin{aligned}
 t_{ij}(\theta_1 * \theta_2) &= \sum_{y \in \mathcal{C}_j} r_{xy}(\theta_1 * \theta_2) \quad (\text{where } x \text{ is any } x \in \mathcal{C}_i) \\
 &= \sum_{y \in \mathcal{C}_j} \sum_z r_{xz}(\theta_1) r_{zy}(\theta_2) = \sum_{y \in \mathcal{C}_j} \sum_l \sum_{z \in \mathcal{C}_l} r_{xz}(\theta_1) r_{zy}(\theta_2) \\
 &= \sum_l \sum_{z \in \mathcal{C}_l} r_{xz}(\theta_1) \sum_{y \in \mathcal{C}_j} r_{zy}(\theta_2) \\
 &= \sum_l t_{lj}(\theta_2) \sum_{z \in \mathcal{C}_l} r_{xz}(\theta_1) \\
 &= \sum_l t_{il}(\theta_1) t_{lj}(\theta_2).
 \end{aligned}$$

Hence, $T(\theta)$ satisfies the semigroup property $T(\theta_1 * \theta_2) = T(\theta_1)T(\theta_2)$. Now $T(\theta)$ is differentiable and for $\theta \neq \frac{1}{2}$,

$$\begin{aligned}
 \frac{T(\theta + h(1 - 2\theta)) - T(\theta)}{h(1 - 2\theta)} &= \frac{T(\theta * h) - T(\theta)}{h(1 - 2\theta)} \\
 &= \left(\frac{T(\theta)}{1 - 2\theta} \right) \left(\frac{T(h) - I}{h} \right) = \left(\frac{T(h) - I}{h} \right) \left(\frac{T(\theta)}{1 - 2\theta} \right).
 \end{aligned}$$

Thus $T'(\theta)$, the matrix of first derivatives of the transition probabilities, is given by

$$T'(\theta) = \lim_{h \rightarrow 0} \frac{T(\theta + h) - T(\theta)}{h} = \lim_{h \rightarrow 0} \frac{T(\theta + h(1 - 2\theta)) - T(\theta)}{h(1 - 2\theta)},$$

that is,

$$T'(\theta) = \frac{T(\theta)}{1 - 2\theta} T'(0) = T'(0) \frac{T(\theta)}{1 - 2\theta},$$

and hence

$$T(\theta) = e^{d(\theta)Q},$$

where $d(\theta) = -\frac{1}{2} \ln(1 - 2\theta)$ is the inverse of the Haldane map function and $Q = T'(0)$ is the infinitesimal generator. Q has entries,

$$q_{ij} = \sum_{y \in \mathcal{C}_j} (-2k I(\Delta(x, y) = 0) + I(\Delta(x, y) = 1)) = \sum_{y \in \mathcal{C}_j} I(\Delta(x, y) = 1) - 2k \delta_{ij},$$

where x is any inheritance vector in $\mathcal{C}_i$ and $\delta_{ij} = 1$ if $i = j$ and 0 otherwise. Then Q may be written as $Q = B - 2kI$, where B is the $m \times m$ matrix with entries

$$b_{ij} = \sum_{y \in \mathcal{C}_j} I(\Delta(x, y) = 1) \quad \text{for any } x \in \mathcal{C}_i.$$

S. DUDOIT AND T. P. SPEED

$T(\theta)$ satisfies

$$|\mathcal{C}_i|t_{ij}(\theta) = |\mathcal{C}_j|t_{ji}(\theta);$$

hence, the stationary distribution of $T(\theta)$ is

$$\alpha = (\alpha_1, \dots, \alpha_m) = \frac{1}{2^{2k}}(|\mathcal{C}_1|, \dots, |\mathcal{C}_m|),$$

since

$$\sum_i \alpha_i t_{ij}(\theta) = \sum_i \alpha_j t_{ji}(\theta) = \alpha_j. \quad \square$$

B.2. Proof of Proposition 2. Q satisfies the reversibility condition $\alpha_i q_{ij} = \alpha_j q_{ji}$, hence Q is self-adjoint with respect to the real inner product $\langle x, y \rangle_\alpha = \sum_i \alpha_i x_i y_i$ on $\mathbb{R}^m$. Hence, from the principal axis theorem [cf. Jacob (1990), page 288], Q has an orthonormal basis of eigenvectors with only real eigenvalues, λ_h , $h = 1, \dots, m$ (not necessarily distinct). Denote the h th right eigenvector by $\mathbf{v}_h$ and its i th entry by v_{ih} . Then $\langle \mathbf{v}_h, \mathbf{v}_l \rangle_\alpha = \sum_i \alpha_i v_{ih} v_{il} = \delta_{hl}$. Since Q is reversible, the row vector $\mathbf{w}_h$ with i th entry $w_{hi} = \alpha_i v_{ih}$ is the left eigenvector of Q corresponding to the h th eigenvalue. Hence Q may be written as

$$Q = \sum_h \lambda_h P_h,$$

where

$$(P_h)_{ij} = v_{ih} w_{hj} = \alpha_j v_{ih} v_{jh},$$

that is,

$$P_h = \mathbf{v}_h \mathbf{w}_h.$$

The projection matrices satisfy $P_h^2 = P_h = P_h^*$, $P_h P_l = 0$, $h \neq l$ and $\sum_h P_h = I$, where P_h^* is the adjoint of P_h with respect to $\langle \cdot, \cdot \rangle_\alpha$. It follows that

$$T(\theta) = \sum_h \exp(\lambda_h d(\theta)) P_h = \sum_h (1 - 2\theta)^{-\lambda_h/2} P_h. \quad \square$$

C. Adjacency matrix of quotient graph $\mathcal{X}/(G \times H)$. Consider the graph $\mathcal{X}$ with vertex set the set of all inheritance vectors of length $2k$ and adjacency matrix $A(\mathcal{X}) = A$ with (x, y) -entry,

$$a_{xy} = \begin{cases} 1, & \text{if } \Delta(x, y) = 1, \\ 0, & \text{otherwise.} \end{cases}$$

To describe the eigenvectors of A it is convenient to code the inheritance vectors $x = (x_1, x_2, \dots, x_{2k})$ as in a 2^{2k} factorial experiment, where $x_{2i-1} = 1$ when factor $2i - 1$ is absent and 2 when it is present, and $x_{2i} = 3$ when factor $2i$ is absent and 4 when it is present. The eigenvectors of A have the following patterns.

A SCORE TEST FOR LINKAGE

PROPOSITION 3 (Eigenvalues and eigenvectors of adjacency matrix A). *The eigenvector corresponding to the eigenvalue $\lambda = 2k$ is the grand mean term $V_0 = (1, 1, \dots, 1)^T$. The eigenvectors corresponding to the eigenvalue $\lambda = 2k - 2$ are the $2k$ main effect terms, $V_1, V_2, \dots, V_{2k}$, where*

$$V_{2i-1}(x) = I(x_{2i-1} = 2) - I(x_{2i-1} = 1),$$

$$V_{2i}(x) = I(x_{2i} = 4) - I(x_{2i} = 3).$$

The eigenvectors corresponding to the eigenvalue $\lambda = 2k - 4$ are the $\binom{2k}{2}$ 2-factor interactions, V_{ij} , $1 \leq i < j \leq 2k$, where

$$V_{ij}(x) = V_i(x)V_j(x).$$

In general, the eigenvectors corresponding to the eigenvalue $\lambda = 2(k - i)$, $i = 0, \dots, 2k$, are the $\binom{2k}{i}$ i -factor interactions, $V_{j_1, j_2, \dots, j_i}$, $1 \leq j_1 < j_2 < \dots < j_i \leq 2k$, where

$$V_{j_1, j_2, \dots, j_i}(x) = V_{j_1}(x)V_{j_2}(x) \cdots V_{j_i}(x).$$

Let H denote the matrix with rows the 2^{2k} eigenvectors of A described above. Then, H is an Hadamard matrix; that is, its entries are 1 and -1 and $HH^T = 2^{2k}I$.

PROOF (Partial). We need not distinguish the parental origin of the DNA, hence, for simplicity, denote 1's and 3's by 0's and 2's and 4's by 1's. Then

$$V_i(x) = I(x_i = 1) - I(x_i = 0) = 2I(x_i = 1) - 1.$$

$\lambda = 2k$: the rows of A sum to $2k$ hence $\lambda = 2k$ is an eigenvalue of A with eigenvector V_0 .

$\lambda = 2k - 2$:

$$\begin{aligned} \sum_y a_{xy} V_i(y) &= \sum_y I(\Delta(x, y) = 1)(2I(y_i = 1) - 1) \\ &= 2 \sum_y I(\Delta(x, y) = 1, y_i = 1) - 2k \\ &= 2(I(x_i = 1)(2k - 1) + I(x_i = 0)) - 2k \\ &= 2((2k - 2)I(x_i = 1) + 1) - 2k \\ &= (2k - 2)(2I(x_i = 1) - 1) = (2k - 2)V_i(x). \end{aligned}$$

Hence $\lambda = 2k - 2$ is an eigenvalue of A with eigenvectors V_i , $i = 1, \dots, 2k$. It is easy to show that $\langle V_i, V_j \rangle = 2^{2k} \delta_{ij}$.

S. DUDOIT AND T. P. SPEED

$$\lambda = 2k - 4:$$

$$\begin{aligned}
& \sum_y a_{xy} V_i(y) V_j(y) \\
&= \sum_y I(\Delta(x, y) = 1) (2I(y_i = 1) - 1) (2I(y_j = 1) - 1) \\
&= \sum_y I(\Delta(x, y) = 1) (4I(y_i = 1, y_j = 1) - 2I(y_i = 1) - 2I(y_j = 1) + 1) \\
&= 4(I(x_i = 1, x_j = 1)(2k - 2) + I(x_i = 1, x_j = 0) + I(x_i = 0, x_j = 1)) \\
&\quad - 2(I(x_i = 1)(2k - 1) + I(x_i = 0)) \\
&\quad - 2(I(x_j = 1)(2k - 1) + I(x_j = 0)) + 2k \\
&= 4I(x_i = 1, x_j = 1)(2k - 2) \\
&\quad + 4(I(x_i = 1) - I(x_i = 1, x_j = 1)) \\
&\quad + 4(I(x_j = 1) - I(x_i = 1, x_j = 1)) \\
&\quad - 2((2k - 2)I(x_i = 1) + 1) - 2((2k - 2)I(x_j = 1) + 1) + 2k \\
&= (2k - 4)(4I(x_i = 1, x_j = 1) - 2I(x_i = 1) - 2I(x_j = 1) + 1) \\
&= (2k - 4)V_i(x)V_j(x).
\end{aligned}$$

Hence $\lambda = 2k - 4$ is an eigenvalue of A with eigenvectors V_{ij} . $\square$

In order to prove Theorem 1, we rely on the following general facts concerning quotient graphs [Chapter 5 in Godsil (1993)]. Consider a group $G \times H$ acting on the vertices of $\mathcal{X}$, as described in Table 2. Then, by the same argument as that leading to (3.1), the orbits of $G \times H$, $\mathcal{C}_i$, $i = 1, \dots, m$, form an equitable partition of the vertex set of $\mathcal{X}$. The matrix B defined in Proposition 1 is the *adjacency matrix of the quotient graph* $\mathcal{X}/(G \times H)$, which is the multidigraph with the orbits of $G \times H$ as its vertices and with b_{ij} arcs going from $\mathcal{C}_i$ to $\mathcal{C}_j$. Let C denote the *characteristic matrix* of the partition $(\mathcal{C}_i)$; C is a $2^{2k} \times m$ matrix, with ij th entry 1 or 0 according as the i th vertex of $\mathcal{X}$ is contained in the orbit $\mathcal{C}_j$ or not.

FACT 1 [Based on Lemma 2.2 in Godsil (1993)]. The eigenvalues of B are a subset of the eigenvalues of A .

FACT 2 [Based on Lemma 2.2 in Godsil (1993)]. If v is an eigenvector of B , then Cv is an eigenvector of A which is constant over the orbits of $G \times H$, with entry v_i on $\mathcal{C}_i$.

FACT 3. If V is an eigenvector of A which is constant over the orbits of $G \times H$, with $V(x) = v_i \forall x \in \mathcal{C}_i$, then the vector v , with i th entry v_i , is an eigenvector of B .

A SCORE TEST FOR LINKAGE

PROOF. For any $x \in \mathcal{C}_i$

$$\lambda v_i = \lambda V(x) = \sum_y a_{xy} V(y) = \sum_j v_j \sum_{y \in \mathcal{C}_j} a_{xy} = \sum_j v_j b_{ij}. \quad \square$$

The proof of Theorem 1 also relies on the following specific properties of the eigenvectors of A on the orbits of $G \times H$.

C.1. Quotient graph $\mathcal{X}/(S_k \times D_4)$.

FACT 4. The $2k$ eigenvectors of A corresponding to the eigenvalue $2k - 2$ sum to 0 over the orbits of $S_k \times D_4$, that is, $\forall i = 1, \dots, 2k$, and any orbit $\mathcal{C}$,

$$\sum_{x \in \mathcal{C}} V_i(x) = 0.$$

PROOF. Let $\iota \in S_k$ denote the identity permutation and as before, let $\alpha = (ac)(bd)$ denote the permutation of D_4 which corresponds to interchanging the paternal labels 1 and 2. Let $\tilde{x} = (\iota, \alpha)(x)$ denote the inheritance vector obtained from x by interchanging the paternal labels. Then, for $1 \leq i \leq k$,

$$\begin{aligned} V_{2i-1}(x) &= (I(x_{2i-1} = 2) - I(x_{2i-1} = 1)) \\ &= (I(\tilde{x}_{2i-1} = 1) - I(\tilde{x}_{2i-1} = 2)) = -V_{2i-1}(\tilde{x}), \end{aligned}$$

and since applying (ι, α) to the elements of $\mathcal{C}$ results in a permutation of the inheritance vectors in $\mathcal{C}$, then

$$\sum_{x \in \mathcal{C}} V_{2i-1}(x) = - \sum_{x \in \mathcal{C}} V_{2i-1}(\tilde{x}) = - \sum_{x \in \mathcal{C}} V_{2i-1}(x).$$

Consequently,

$$\sum_{x \in \mathcal{C}} V_{2i-1}(x) = 0.$$

The proof for V_{2i} is similar, but uses the permutation β instead of α . $\square$

FACT 5. The k^2 eigenvectors of A corresponding to the eigenvalue $2k - 4$ and involving “odd” and “even” factors sum to 0 over the orbits of $S_k \times D_4$, that is, $\forall i, j = 1, \dots, k$, and any orbit $\mathcal{C}$,

$$\sum_{x \in \mathcal{C}} V_{2i-1}(x) V_{2j}(x) = 0.$$

PROOF. Here again, let $\tilde{x} = (\iota, \alpha)(x)$. Then

$$V_{2i-1}(x) V_{2j}(x) = (-V_{2i-1}(\tilde{x})) V_{2j}(\tilde{x})$$

and

$$\sum_{x \in \mathcal{C}} V_{2i-1}(x) V_{2j}(x) = - \sum_{x \in \mathcal{C}} V_{2i-1}(\tilde{x}) V_{2j}(\tilde{x}) = - \sum_{x \in \mathcal{C}} V_{2i-1}(x) V_{2j}(x).$$

S. DUDOIT AND T. P. SPEED

Hence

$$\sum_{x \in \mathcal{C}} V_{2i-1}(x) V_{2j}(x) = 0. \quad \square$$

FACT 6. Let

$$V(x) = \sum_{(i,j)} \{V_{2i-1, 2j-1}(x) + V_{2i, 2j}(x)\},$$

where the sum is over all $\binom{k}{2}$ unordered pairs (i, j) of distinct integers ranging from 1 to k . Then V is an eigenvector of A corresponding to the eigenvalue $2k - 4$. Furthermore, V is constant over the orbits of $S_k \times D_4$, that is, for any orbit $\mathcal{C}$,

$$V(x) = V(\tilde{x}) \quad \text{whenever } x, \tilde{x} \in \mathcal{C}.$$

PROOF. Members of the same orbit are obtained by a combination of any of the following three operations: a permutation $\sigma \in S_k$ of the sibs, and permutations α and γ of the pairs of labels of all sibs simultaneously. We consider a particular configuration $\mathcal{C}$ and the effect of each operation separately on $x \in \mathcal{C}$.

$\tilde{x} = (\iota, \alpha)(x)$, where ι is the identity in S_k and $\alpha = (ac)(bd)$: for each pair (i, j) ,

$$V_{2i-1}(\tilde{x}) V_{2j-1}(\tilde{x}) + V_{2i}(\tilde{x}) V_{2j}(\tilde{x}) = (-V_{2i-1}(x))(-V_{2j-1}(x)) + V_{2i}(x) V_{2j}(x),$$

hence $V(\tilde{x}) = V(x)$.

$\tilde{x} = (\iota, \gamma)(x)$, where ι is the identity in S_k and $\gamma = (bc)$: for each $1 \leq i \leq k$,

$$\begin{aligned} I(\tilde{x}_{2i-1} = 1) &= I(\tilde{x}_{2i-1} = 1, \tilde{x}_{2i} = 3) + I(\tilde{x}_{2i-1} = 1, \tilde{x}_{2i} = 4) \\ &= I(x_{2i-1} = 1, x_{2i} = 3) + I(x_{2i-1} = 2, x_{2i} = 3) \\ &= I(x_{2i} = 3), \end{aligned}$$

and similarly

$$I(\tilde{x}_{2i-1} = 2) = I(x_{2i} = 4).$$

Hence, for $1 \leq i \leq k$,

$$(C.1) \quad V_{2i-1}(\tilde{x}) = V_{2i}(x),$$

and consequently $V(x) = V(\tilde{x})$.

$\tilde{x} = (\sigma, \iota)(x)$, where $\sigma \in S_k$ and ι is the identity in D_4 : For $1 \leq i \leq k$, $\tilde{x}_{2i-1} = x_{2\sigma^{-1}(i)-1}$ and $\tilde{x}_{2i} = x_{2\sigma^{-1}(i)}$, thus

$$\begin{aligned} V(\tilde{x}) &= \sum_{(i,j)} \{V_{2\sigma^{-1}(i)-1}(x) V_{2\sigma^{-1}(j)-1}(x) + V_{2\sigma^{-1}(i)}(x) V_{2\sigma^{-1}(j)}(x)\} \\ &= \sum_{(i,j)} \{V_{2i-1}(x) V_{2j-1}(x) + V_{2i}(x) V_{2j}(x)\} = V(x). \end{aligned}$$

In particular, for $k > 1$,

$$V(1, 3, 1, 3, \dots, 1, 3) = \sum_{(i,j)} (-1)(-1) + (-1)(-1) = 2 \binom{k}{2} = k(k-1) \neq 0.$$

Hence, since V is a linear combination of eigenvectors of A which is nonzero, then V is an eigenvector of A corresponding to the eigenvalue $2k - 4$. Furthermore, V is constant on the orbits of $S_k \times D_4$. $\square$

FACT 7. The $k(k-1)$ 2-factor eigenvectors $\{V_{2i-1, 2j-1}, V_{2i, 2j} : 1 \leq i < j \leq k\}$ have the same sums over the orbits of $S_k \times D_4$, that is, for any orbit $\mathcal{C}$ and $1 \leq i_1 < j_1 \leq k$, $1 \leq i_2 < j_2 \leq k$,

$$\sum_{x \in \mathcal{C}} V_{2i_2-1, 2j_2-1}(x) = \sum_{x \in \mathcal{C}} V_{2i_1-1, 2j_1-1}(x) = \sum_{x \in \mathcal{C}} V_{2i_1, 2j_1}(x) = \sum_{x \in \mathcal{C}} V_{2i_2, 2j_2}(x).$$

PROOF. Let $x \in \mathcal{C}$, then $\tilde{x} = (\iota, \gamma)(x) \in \mathcal{C}$, and by (C.1), for each $1 \leq i < j \leq k$,

$$\sum_{x \in \mathcal{C}} V_{2i-1}(x) V_{2j-1}(x) = \sum_{x \in \mathcal{C}} V_{2i}(\tilde{x}) V_{2j}(\tilde{x}) = \sum_{x \in \mathcal{C}} V_{2i}(x) V_{2j}(x).$$

Also, consider any permutation $\sigma \in S_k$, then $\tilde{x} = (\sigma, \iota)(x) \in \mathcal{C}$ and

$$\sum_{x \in \mathcal{C}} V_{2i}(x) V_{2j}(x) = \sum_{x \in \mathcal{C}} V_{2i}(\tilde{x}) V_{2j}(\tilde{x}) = \sum_{x \in \mathcal{C}} V_{2\sigma^{-1}(i)}(x) V_{2\sigma^{-1}(j)}(x).$$

Similarly for $V_{2i-1, 2j-1}$.

PROPOSITION 4 [Eigenvalues of adjacency matrix B of quotient graph $\mathcal{X}/(S_k \times D_4)$]. *The two largest eigenvalues of B are $2k$ and $2k - 4$, and have multiplicity one. All other eigenvalues of B are strictly less than $2k - 4$ and belong to the set $\{2(k-i) \binom{2k}{i} : i = 3, \dots, 2k\}$, where $\binom{2k}{i}$ is the largest possible multiplicity of the eigenvalue $2(k-i)$. The eigenvector v corresponding to $2k - 4$ may be obtained from*

$$V(x) = \sum_{(i,j)} \{V_{2i-1, 2j-1}(x) + V_{2i, 2j}(x)\},$$

by letting

$$v_i = V(x) \quad \text{where } x \text{ is any } x \in \mathcal{C}_i.$$

PROOF. From Proposition 3 and Fact 1, the eigenvalues of B belong to the set $\{2(k-i) \binom{2k}{i} : i = 0, \dots, 2k\}$.

$\lambda = 2k$: the rows of B sum to $2k$, hence $2k$ is an eigenvalue of B with corresponding eigenvector $\mathbf{1} = (1, 1, \dots, 1)^T$.

$\lambda = 2k - 2$: from Fact 4, eigenvectors of A corresponding to the eigenvalue $2k - 2$ sum to 0 over the orbits of $S_k \times D_4$, hence no eigenvector of A can

S. DUDOIT AND T. P. SPEED

be constant and nonzero over the orbits. Hence, from Fact 2, $2k - 2$ is not an eigenvalue of B .

$\lambda = 2k - 4$: we have shown with Fact 6 that V is an eigenvector of A , corresponding to the eigenvalue $2k - 4$, which is constant over the orbits. Hence, by Fact 3, V yields an eigenvector of B . It remains to show that B has no other eigenvector, that is, V is the only eigenvector of A which is constant over the orbits. The orthogonal complement of V in the eigenspace of A for $\lambda = 2k - 4$ is spanned by the following $2k^2 - k$ vectors:

$$\begin{aligned} W_{2i-1, 2j} &= V_{2i-1, 2j} - \frac{\langle V_{2i-1, 2j}, V \rangle}{|V|^2} V \\ &= V_{2i-1, 2j}, \quad 1 \leq i, j \leq k, \\ W_{2i-1, 2j-1} &= V_{2i-1, 2j-1} - \frac{\langle V_{2i-1, 2j-1}, V \rangle}{|V|^2} V \\ &= V_{2i-1, 2j-1} - \frac{1}{k(k-1)} V, \quad 1 \leq i < j \leq k, \\ W_{2i, 2j} &= V_{2i, 2j} - \frac{\langle V_{2i, 2j}, V \rangle}{|V|^2} V \\ &= V_{2i, 2j} - \frac{1}{k(k-1)} V, \quad 1 \leq i < j \leq k. \end{aligned}$$

By Fact 5, for any orbit $\mathcal{C}$,

$$\sum_{x \in \mathcal{C}} W_{2i-1, 2j}(x) = 0.$$

Also, by Fact 7,

$$\begin{aligned} &\sum_{x \in \mathcal{C}} W_{2i-1, 2j-1}(x) \\ &= \sum_{x \in \mathcal{C}} V_{2i-1, 2j-1}(x) - \frac{1}{k(k-1)} \sum_{(i, j)} \sum_{x \in \mathcal{C}} \{V_{2i-1, 2j-1}(x) + V_{2i, 2j}(x)\} = 0, \end{aligned}$$

and similarly

$$\sum_{x \in \mathcal{C}} W_{2i, 2j}(x) = 0.$$

Hence, no eigenvector in the orthogonal complement of V in the eigenspace of A for $\lambda = 2k - 4$ is constant over the orbits of $S_k \times D_4$. Consequently, by Fact 2, $2k - 4$ is an eigenvalue of B with multiplicity 1. $\square$

C.2. Quotient graph $\mathcal{X}/(S_k \times (C_2 \times C_2))$. Facts 4 and 5 also apply to the orbits of $S_k \times (C_2 \times C_2)$. Facts 6 and 7 may be modified as follows.

A SCORE TEST FOR LINKAGE

FACT 8. Let

$$V_o(x) = \sum_{(i,j)} V_{2i-1, 2j-1}(x)$$

and

$$V_e(x) = \sum_{(i,j)} V_{2i, 2j}(x),$$

where the sums are over all $\binom{k}{2}$ unordered pairs (i, j) of distinct integers ranging from 1 to k . Then, V_e and V_o are two eigenvectors of A corresponding to the eigenvalue $2k - 4$. Furthermore, V_e and V_o are constant over the orbits of $S_k \times (C_2 \times C_2)$.

FACT 9. The $k(k-1)/2$ 2-factor eigenvectors $\{V_{2i-1, 2j-1}: 1 \leq i < j \leq k\}$ have the same sums over the orbits of $S_k \times (C_2 \times C_2)$, that is, for any orbit $\mathcal{C}$ and $1 \leq i_1 < j_1 \leq k$, $1 \leq i_2 < j_2 \leq k$,

$$\sum_{x \in \mathcal{C}} V_{2i_1-1, 2j_1-1}(x) = \sum_{x \in \mathcal{C}} V_{2i_2-1, 2j_2-1}(x).$$

Similarly for the $k(k-1)/2$ 2-factor eigenvectors $\{V_{2i, 2j}: 1 \leq i < j \leq k\}$.

PROPOSITION 5 [Eigenvalues of adjacency matrix B of quotient graph $\mathcal{X}/(S_k \times (C_2 \times C_2))$]. *The largest eigenvalue of B is $2k$, with multiplicity one, and the second largest eigenvalue is $2k - 4$, with multiplicity two. All other eigenvalues of B are strictly less than $2k - 4$ and belong to the set $\{2(k-i)\binom{2k}{i}: i = 3, \dots, 2k\}$. The eigenvectors corresponding to $2k - 4$ may be obtained from V_e and V_o .*

PROOF. From Proposition 3 and Fact 1, the eigenvalues of B belong to the set $\{2(k-i)\binom{2k}{i}: i = 0, \dots, 2k\}$.

$\lambda = 2k$: the rows of B sum to $2k$, hence $2k$ is an eigenvalue of B with corresponding eigenvector $\mathbf{1} = (1, 1, \dots, 1)^T$.

$\lambda = 2k - 2$: from Fact 4, eigenvectors of A corresponding to the eigenvalue $2k - 2$ sum to 0 over the orbits of $S_k \times (C_2 \times C_2)$, hence no eigenvector of A can be constant and nonzero over the orbits. Hence, from Fact 2, $2k - 2$ is not an eigenvalue of B .

$\lambda = 2k - 4$: from Fact 8, V_o and V_e are eigenvectors of A , corresponding to the eigenvalue $2k - 4$, which are constant over the orbits. Hence, by Fact 3, V_e and V_o yield two eigenvectors of B . It remains to show that B has only two eigenvectors, that is, V_e and V_o are the only eigenvectors of A which are constant over the orbits. The orthogonal complement of $\text{Span}\{V_o, V_e\}$ in the eigenspace of A for $\lambda = 2k - 4$ is spanned by the following $2k^2 - k$ vectors:

$$\begin{aligned} W_{2i-1, 2j} &= V_{2i-1, 2j} - \frac{\langle V_{2i-1, 2j}, V_e \rangle}{|V_e|^2} V_e - \frac{\langle V_{2i-1, 2j}, V_o \rangle}{|V_o|^2} V_o \\ &= V_{2i-1, 2j}, \quad 1 \leq i, j \leq k, \end{aligned}$$

S. DUDOIT AND T. P. SPEED

$$\begin{aligned}
W_{2i-1, 2j-1} &= V_{2i-1, 2j-1} - \frac{\langle V_{2i-1, 2j-1}, V_e \rangle}{|V_e|^2} V_e - \frac{\langle V_{2i-1, 2j-1}, V_o \rangle}{|V_o|^2} V_o \\
&= V_{2i-1, 2j-1} - \frac{2}{k(k-1)} V_o, \quad 1 \leq i < j \leq k, \\
W_{2i, 2j} &= V_{2i, 2j} - \frac{\langle V_{2i, 2j}, V_e \rangle}{|V_e|^2} V_e - \frac{\langle V_{2i, 2j}, V_o \rangle}{|V_o|^2} V_o \\
&= V_{2i, 2j} - \frac{2}{k(k-1)} V_e, \quad 1 \leq i < j \leq k.
\end{aligned}$$

By Fact 5, for any orbit $\mathcal{C}$,

$$\sum_{x \in \mathcal{C}} W_{2i-1, 2j}(x) = 0.$$

Also, by Fact 9,

$$\sum_{x \in \mathcal{C}} W_{2i-1, 2j-1}(x) = \sum_{x \in \mathcal{C}} V_{2i-1, 2j-1}(x) - \frac{2}{k(k-1)} \sum_{(i, j)} \sum_{x \in \mathcal{C}} V_{2i-1, 2j-1}(x) = 0,$$

and similarly,

$$\sum_{x \in \mathcal{C}} W_{2i, 2j}(x) = 0.$$

Hence, no eigenvector in the orthogonal complement of $\text{Span}\{V_o, V_e\}$ in the eigenspace of A for $\lambda = 2k - 4$ is constant over the orbits of $S_k \times (C_2 \times C_2)$. Consequently, by Fact 2, $2k - 4$ is an eigenvalue of B with multiplicity 2. $\square$

C.3. Quotient graph $\mathcal{X}/((S_h \times S_{k-h}) \times D_4)$. Facts 4 and 5 also apply to the orbits of $\mathcal{X}/((S_h \times S_{k-h}) \times D_4)$. The proof for sibships with both affected and unaffected sibs is similar to that for affected-only sibships, but involves new combinations of eigenvectors. Without loss of generality, order the sibs such that the first h are affected and the last $k - h$ unaffected. For $k \geq 3$, define

$$\begin{aligned}
V^a(x) &= \sum_{1 \leq i < j \leq h} \{V_{2i-1, 2j-1}(x) + V_{2i, 2j}(x)\}, \quad h \geq 2, \\
V^u(x) &= \sum_{h+1 \leq i < j \leq k} \{V_{2i-1, 2j-1}(x) + V_{2i, 2j}(x)\}, \quad h \leq k-2, \\
V^{au}(x) &= \sum_{1 \leq i \leq h, h+1 \leq j \leq k} \{V_{2i-1, 2j-1}(x) + V_{2i, 2j}(x)\}.
\end{aligned}$$

Facts 6 and 7 are then modified as follows.

FACT 10. For $k \geq 3$, V^a ($h \geq 2$), V^u ($h \leq k - 2$) and V^{au} are eigenvectors of A corresponding to the eigenvalue $2k - 4$. Furthermore, these are constant over the orbits of $\mathcal{X}/((S_h \times S_{k-h}) \times D_4)$.

FACT 11. For any orbit $\mathcal{C}$ of $\mathcal{X}/((S_h \times S_{k-h}) \times D_4)$ and $1 \leq i_1 < j_1 \leq h$, $1 \leq i_2 < j_2 \leq h$,

$$\sum_{x \in \mathcal{C}} V_{2i_2-1, 2j_2-1}(x) = \sum_{x \in \mathcal{C}} V_{2i_1-1, 2j_1-1}(x) = \sum_{x \in \mathcal{C}} V_{2i_1, 2j_1}(x) = \sum_{x \in \mathcal{C}} V_{2i_2, 2j_2}(x).$$

Similarly for $h+1 \leq i_1 < j_1 \leq k$, $h+1 \leq i_2 < j_2 \leq k$ and $1 \leq i_1, i_2 \leq h$, $h+1 \leq j_1, j_2 \leq k$.

PROPOSITION 6 [Eigenvalues of adjacency matrix B of quotient graph $\mathcal{X}/((S_h \times S_{k-h}) \times D_4)$]. *The largest eigenvalue of B is $2k$, with multiplicity one, and the second largest eigenvalue is $2k-4$, with multiplicity three if $2 \leq h \leq k-2$ and two otherwise. All other eigenvalues of B are strictly less than $2k-4$ and belong to the set $\{2(k-i)\binom{2k}{i} : i = 3, \dots, 2k\}$. The eigenvectors corresponding to $2k-4$ may be obtained from V^a , V^u and V^{au} .*

PROOF. From Proposition 3 and Fact 1, the eigenvalues of B belong to the set $\{2(k-i)\binom{2k}{i} : i = 0, \dots, 2k\}$. We give the proof for $\lambda = 2k-4$; for the other eigenvalues, the proof is as for Propositions 4 and 5. From Fact 10, V^a ($h \geq 2$), V^u ($h \leq k-2$) and V^{au} are eigenvectors of A , corresponding to the eigenvalue $2k-4$, which are constant over the orbits. Hence, by Fact 3, they yield eigenvectors of B . It remains to show that these are the only eigenvectors of B , that is, V^a , V^u and V^{au} are the only eigenvectors of A which are constant over the orbits. The orthogonal complement of $\text{Span}\{V^a, V^u, V^{au}\}$ in the eigenspace of A for $\lambda = 2k-4$ is spanned by the $2k^2-k$ vectors $W_{i,j}$, $1 \leq i, j \leq 2k$, $i \neq j$, defined as follows

$$W_{i,j} = V_{i,j} - \frac{\langle V_{i,j}, V^a \rangle}{|V^a|^2} V^a - \frac{\langle V_{i,j}, V^u \rangle}{|V^u|^2} V^u - \frac{\langle V_{i,j}, V^{au} \rangle}{|V^{au}|^2} V^{au}.$$

The $W_{i,j}$'s simplify to

$$\begin{aligned} W_{2i-1, 2j} &= V_{2i-1, 2j}, \quad i, j = 1, \dots, k, \\ W_{2i-1, 2j-1} &= V_{2i-1, 2j-1} - \frac{1}{h(h-1)} V^a, \quad 1 \leq i < j \leq h, \\ &= V_{2i-1, 2j-1} - \frac{1}{(k-h)(k-h-1)} V^u, \quad h+1 \leq i < j \leq k, \\ &= V_{2i-1, 2j-1} - \frac{1}{2h(k-h)} V^{au}, \quad 1 \leq i \leq h, \quad h+1 \leq j \leq k, \\ W_{2i, 2j} &= V_{2i, 2j} - \frac{1}{h(h-1)} V^a, \quad 1 \leq i < j \leq h, \\ &= V_{2i, 2j} - \frac{1}{(k-h)(k-h-1)} V^u, \quad h+1 \leq i < j \leq k, \\ &= V_{2i, 2j} - \frac{1}{2h(k-h)} V^{au}, \quad 1 \leq i \leq h, \quad h+1 \leq j \leq k. \end{aligned}$$

S. DUDOIT AND T. P. SPEED

By Fact 5, for any orbit $\mathcal{C}$,

$$\sum_{x \in \mathcal{C}} W_{2i-1, 2j}(x) = 0.$$

Also, by Fact 11,

$$\sum_{x \in \mathcal{C}} W_{2i-1, 2j-1}(x) = 0$$

and

$$\sum_{x \in \mathcal{C}} W_{2i, 2j}(x) = 0.$$

Hence, no eigenvector in the orthogonal complement of $\text{Span}\{V^a, V^u, V^{au}\}$ in the eigenspace of A for $\lambda = 2k - 4$ is constant over the orbits of $\mathcal{X}/((S_h \times S_{k-h}) \times D_4)$. Consequently, by Fact 2, $2k - 4$ is an eigenvalue of B with multiplicity three if $2 \leq h \leq k - 2$ and two otherwise. $\square$

C.4. Quotient graph $\mathcal{X}/((S_h \times S_{k-h}) \times (C_2 \times C_2))$. For $(S_h \times S_{k-h}) \times (C_2 \times C_2)$ we again separate “even” and “odd” eigenvectors and consider six new combinations of eigenvectors,

$$V_e^a(x) = \sum_{1 \leq i < j \leq h} V_{2i, 2j}(x), \quad h \geq 2,$$

$$V_o^a(x) = \sum_{1 \leq i < j \leq h} V_{2i-1, 2j-1}(x), \quad h \geq 2,$$

$$V_e^u(x) = \sum_{h+1 \leq i < j \leq k} V_{2i, 2j}(x), \quad h \leq k - 2,$$

$$V_o^u(x) = \sum_{h+1 \leq i < j \leq k} V_{2i-1, 2j-1}(x), \quad h \leq k - 2,$$

$$V_e^{au}(x) = \sum_{1 \leq i \leq h, h+1 \leq j \leq k} V_{2i, 2j}(x),$$

$$V_o^{au}(x) = \sum_{1 \leq i \leq h, h+1 \leq j \leq k} V_{2i-1, 2j-1}(x).$$

Facts 6 and 7 may then be suitably modified.

D. Score statistic: Proof of Theorem 2. From Theorem 1, -4 is an eigenvalue of the infinitesimal generator Q with multiplicity 1. Hence, the second derivative of the transition matrix at $\theta = \frac{1}{2}$ has rank 1 and entries

$$u_{ij} = 8\alpha_j v_i v_j,$$

A SCORE TEST FOR LINKAGE

where $v = (v_1, \dots, v_m)^T$ is the right eigenvector of Q with unit norm with respect to the inner product $\langle \cdot, \cdot \rangle_\alpha$. The score statistic is given by

$$S = \sum_{i=1}^m N_i \frac{\sum_{j=1}^m \pi_j 8\alpha_i v_i v_j}{\alpha_i} \\ = 8 \left(\sum_{j=1}^m v_j \pi_j \right) \left(\sum_{i=1}^m v_i N_i \right).$$

It is convenient to express the score statistic in terms of the first column of U , $8\alpha_1 v_1 v$. Without loss of generality, we let the first IBD configuration be the one for which all sibs inherited the same maternal and paternal DNA, that is, with representative inheritance vector $(1, 3, 1, 3, \dots, 1, 3)$ and label $(0, 0, 0)$ in the notation of Ethier and Hodge (1985):

$$S = \frac{8}{8(8\alpha_1 v_1^2)\alpha_1} \left(\sum_{j=1}^m u_{j1} \pi_j \right) \left(\sum_{i=1}^m u_{i1} N_i \right) \\ = \frac{2^{2k}}{u_{11} |\mathcal{C}_1|} \left(\sum_{j=1}^m u_{j1} \pi_j \right) \left(\sum_{i=1}^m u_{i1} N_i \right).$$

By differentiating (3.2) we find that

$$u_{ij} = 2^{4-2k} \sum_{y \in \mathcal{C}_j} ((\Delta(x, y) - k)^2 - k/2), \text{ where } x \text{ is any inheritance vector in } \mathcal{C}_i$$

and in particular, $u_{11} = 2^{5-2k} k(k-1)$. The contribution of an affected sib- k -tuple with inheritance vector $x \in \mathcal{C}_i$ to the score statistic is based on

$$u_{i1} = 2^{4-2k} \sum_{y \in \mathcal{C}_1} ((\Delta(x, y) - k)^2 - k/2) \\ = 2^{4-2k} ((a_2(x) + a_4(x) - k)^2 + (a_2(x) + a_3(x) - k)^2 \\ + (a_1(x) + a_4(x) - k)^2 + (a_1(x) + a_3(x) - k)^2 - 2k) \\ = 2^{4-2k} (2(a_1(x)^2 + a_2(x)^2 + a_3(x)^2 + a_4(x)^2) \\ + 2(a_2(x)a_4(x) + a_2(x)a_3(x) + a_1(x)a_4(x) + a_1(x)a_3(x)) \\ - 2k(2a_1(x) + 2a_2(x) + 2a_3(x) + 2a_4(x)) + 4k^2 - 2k) \\ = 2^{5-2k} (a_1(x)^2 + a_2(x)^2 + a_3(x)^2 + a_4(x)^2 \\ + (a_1(x) + a_2(x))(a_3(x) + a_4(x)) - 4k^2 + 2k^2 - k) \\ = 2^{5-2k} (a_1(x)^2 + a_2(x)^2 + a_3(x)^2 + a_4(x)^2 + k^2 - 2k^2 - k) \\ = 2^{5-2k} (a_1(x)^2 + a_2(x)^2 + a_3(x)^2 + a_4(x)^2 - k(k+1)).$$

S. DUDOIT AND T. P. SPEED

Hence, from (4.2),

$$u_{i1} = 2^{5-2k} k(k-1)(2S_{\text{pairs}} - 1).$$

Now, $u_{i1} = 8\alpha_1 v_1 v_i$, $u_{11} = 2^{5-2k} k(k-1)$ and $\alpha_1 = 2^{2-2k}$. Thus

$$v_1 = \sqrt{k(k-1)} \quad \text{and} \quad v_i = \sqrt{k(k-1)(2S_{\text{pairs}} - 1)}.$$

Hence

$$S = 2^{2k-2} \left(\sum_j u_{j1} \pi_j \right) (2S_{\text{pairs}} - n) = 8\sqrt{k(k-1)} \left(\sum_j v_j \pi_j \right) (2S_{\text{pairs}} - n),$$

where S_{pairs} is summed over all sibships with k affected sibs. $\square$

Acknowledgments. We are grateful to Steve Evans for many helpful discussions on Markov chains and to Cheryl Praeger, Alice Niemeyer and Nick Wormald for graph-related ideas. We also thank the referees for their valuable comments and suggestions on a previous version of this paper.

REFERENCES

- BLACKWELDER, W. C. and ELSTON, R. C. (1985). A comparison of sib-pair linkage tests for disease susceptibility loci. *Genet. Epidemiol.* **2** 85–97.
- DAVIS, S. and WEEKS, D. E. (1997). Comparison of nonparametric statistics for detection of linkage in nuclear families: single-marker evaluation. *Amer. J. Hum. Genet.* **61** 1431–1444.
- DAY, N. E. and SIMONS, M. J. (1976). Disease-susceptibility genes—their identification by multiple case family studies. *Tissue Antigens* **8** 109–119.
- DEBRUIJN, N. G. (1964). Pólya's theory of counting. In *Applied Combinatorial Mathematics* (E. F. Beckenbach, ed.) 144–184. Wiley, New York.
- DIACONIS, P. (1988). *Group Representations in Probability and Statistics*. IMS, Hayward, CA.
- DONNELLY, K. P. (1983). The probability that related individuals share some section of genome identical by descent. *Theoret. Population Biol.* **23** 34–63.
- DUDOIT, S. (1999). Linkage analysis of complex human traits using identity by descent data. Ph.D. dissertation, Univ. California, Berkeley.
- DUDOIT, S. and SPEED, T. P. (1999). Triangle constraints for sib-pair identity by descent probabilities under a general multilocus model for disease susceptibility. In *Statistics in Genetics* (M. E. Halloran and S. Geisser, eds.). Springer, New York.
- ETHIER, S. N. and HODGE, S. E. (1985). Identity-by-descent analysis of sibship configurations. *Amer. J. Med. Genet.* **22** 263–272.
- FARAWAY, J. J. (1993). Improved sib-pair linkage test for disease susceptibility loci. *Genet. Epidemiol.* **10** 225–233.
- FEINGOLD, E., BROWN, P. O. and SIEGMUND, D. (1993). Gaussian models for genetic linkage analysis using complete high-resolution maps of identity by descent. *Amer. J. Hum. Genet.* **53** 234–251.
- FEINGOLD, E. and SIEGMUND, D. (1997). Strategies for mapping heterogeneous recessive traits by allele sharing methods. *Amer. J. Hum. Genet.* **60** 965–978.
- FRALEIGH, J. B. (1989). *A First Course in Abstract Algebra*, 4th ed. Addison-Wesley, Reading, MA.
- GODSIL, C. D. (1993). *Algebraic Combinatorics*. Chapman and Hall, New York.

A SCORE TEST FOR LINKAGE

- HASEMAN, J. K. and ELSTON, R. C. (1972). The investigation of linkage between a quantitative trait and a marker locus. *Behavior Genetics* **2** 3–19.
- HOLMANS, P. (1993). Asymptotic properties of affected sib-pair linkage analysis. *Amer. J. Hum. Genet.* **52** 362–374.
- HOLMANS, P. and CLAYTON, D. (1995). Efficiency of typing unaffected relatives in an affected-sib-pair linkage study with single-locus and multiple tightly linked markers. *Amer. J. Hum. Genet.* **57** 1221–1232.
- JACOB, B. (1990). *Linear Algebra*. Freeman, New York.
- KNAPP, M., SEUCHTER, S. A. and BAUR, M. P. (1994a). Linkage analysis in nuclear families I: optimality criteria for affected sib-pair tests. *Hum. Hered.* **44** 37–43.
- KNAPP, M., SEUCHTER, S. A. and BAUR, M. P. (1994b). Linkage analysis in nuclear families II: relationship between affected sib-pair tests and lod score analysis. *Hum. Hered.* **44** 44–51.
- KNAPP, M., SEUCHTER, S. A. and BAUR, M. P. (1994c). Two-locus disease models with two marker loci: the power of affected-sib-pair tests. *Amer. J. Hum. Genet.* **55** 1030–1041.
- KONG, A. and COX, N. J. (1997). Allele-sharing models: lod scores and accurate linkage tests. *Amer. J. Hum. Genet.* **61** 1179–1188.
- KRUGLYAK, L., DALY, M. J., REEVE-DALY, M. P. and LANDER, E. S. (1996). Parametric and non-parametric linkage analysis: a unified multipoint approach. *Amer. J. Hum. Genet.* **58** 1347–1363.
- KRUGLYAK, L. and LANDER, E. S. (1995). Complete multipoint sib-pair analysis of qualitative and quantitative traits. *Amer. J. Hum. Genet.* **57** 439–454.
- LANDER, E. S. and GREEN, P. (1987). Construction of multilocus genetic maps in humans. *Proc. Nat. Acad. Sci. U.S.A.* **84** 2363–2367.
- MCPEEK, M. S. (1996). An introduction to recombination and linkage analysis. In *Genetic Mapping and DNA Sequencing* (T. P. Speed and M. S. Waterman, eds.) 1–14. Springer, New York.
- MCPEEK, M. S. (1999). Optimal allele-sharing statistics for genetic mapping using affected relatives. *Genet. Epidemiol.* **16** 225–249.
- OTT, J. (1991). *Analysis of Human Genetic Linkage*, rev. eds. Johns Hopkins Univ. Press, Baltimore.
- RISCH, N. (1990a). Linkage strategies for genetically complex traits II. The power of affected relative pairs. *Amer. J. Hum. Genet.* **46** 229–241.
- RISCH, N. (1990b). Linkage strategies for genetically complex traits III. The effect of marker polymorphism on analysis of affected relative pairs. *Amer. J. Hum. Genet.* **46** 242–253.
- RISCH, N. and ZHANG, H. (1995). Extreme discordant sib pairs for mapping quantitative trait loci in humans. *Science* **268** 1584–1589.
- RISCH, N. and ZHANG, H. (1996). Mapping quantitative trait loci with extreme discordant sib pairs: sampling considerations. *Amer. J. Hum. Genet.* **58** 836–843.
- ROSENBLATT, M. (1974). *Random Processes. Graduate Texts in Mathematics* **17**, 2nd ed. Springer, New York.
- SCHAID, D. J. and NICK, T. G. (1990). Sib-pair linkage tests for disease susceptibility loci: common tests vs. the asymptotically most powerful test. *Genet. Epidemiol.* **7** 359–370.
- SPEED, T. P. (1996). What is a genetic map function? In *Genetic Mapping and DNA Sequencing* (T. P. Speed and M. S. Waterman, eds.) 65–88. Springer, New York.
- SUAREZ, B. K. and VAN EERDEWEGH, P. (1984). A comparison of three affected-sib-pair scoring methods to detect HLA-linked disease susceptibility genes. *Amer. J. Med. Genet.* **18** 135–146.
- TENG, J. and SIEGMUND, D. O. (1997). Combining information within and between pedigrees for mapping complex traits. *Amer. J. Hum. Genet.* **60** 979–992.
- THOMPSON, E. A. (1997). Conditional gene identity in affected individuals. In *Genetic Mapping of Disease Genes* (I.-H. Pawlowitzki, J. H. Edwards and E. A. Thompson, eds.) 137–146. Academic Press, San Diego.

S. DUDOIT AND T. P. SPEED

- VAN LINT, J. H. and WILSON, R. M. (1992). *A Course in Combinatorics*. Cambridge Univ. Press.
- WHITTEMORE, A. S. (1996). Genome scanning for linkage: an overview. *Amer. J. Hum. Genet.* **59** 704–716.
- WHITTEMORE, A. S. and HALPERN, J. (1994a). A class of tests for linkage using affected pedigree members. *Biometrics* **50** 118–127.
- WHITTEMORE, A. S. and HALPERN, J. (1994b). Probability of gene identity by descent: computation and applications. *Biometrics* **50** 109–117.

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA, BERKELEY
BERKELEY, CALIFORNIA 94720-3860
E-MAIL: sandrine@stat.berkeley.edu
terry@stat.berkeley.edu

Biostatistics (2006), 7, 3, pp. 374–386

doi:10.1093/biostatistics/kxj012

Advance Access publication on December 14, 2005

Incorporating interference into linkage analysis for experimental crosses

NICOLA J. ARMSTRONG*

*Division of Radiotherapy, Netherlands Cancer Institute, Plesmanlaan 121,
1066 CX Amsterdam, The Netherlands
n.armstrong@nki.nl*

MARY SARA MCPEEK

Department of Statistics, University of Chicago, Chicago, IL 60637, USA

TERENCE P. SPEED

*Department of Statistics, University of California at Berkeley, Berkeley, CA 94720, USA and
Division of Genetics and Bioinformatics, Walter and Eliza Hall Institute of Medical Research,
Melbourne, Australia*

SUMMARY

The phenomenon of interference in genetic recombination is well-known and studied in a wide variety of organisms. Multilocus linkage analysis, which makes use of recombination patterns among all genetic markers simultaneously, is routinely used with data on humans and experimental organisms to build genetic maps. It is also used to try to determine the genes involved in traits of interest, such as common diseases. Most linkage analyses performed today ignore the occurrence of genetical interference. We present an extension to the Lander–Green algorithm for experimental crosses (backcross and intercross) to incorporate crossover interference according to the χ^2 model. Simulation results show the impact of using this model on the accuracy of estimated genetic maps.

Keywords: Crossover interference; Hidden Markov model; Linkage analysis.

1. INTRODUCTION

Genetical interference was first observed in Thomas Hunt Morgan's *Drosophila* laboratory early last century (Muller, 1916). During meiosis, chromosomes replicate, producing sister chromatids. The homologous chromosomes then pair and synapse, forming a four-strand bundle. Once pairing is complete, crossing over, the reciprocal exchange of chromosomal segments among nonsister chromatids, begins. As the chromosomes separate, the crossover positions become visible as chiasmata. Two types of interference are distinguished in this crossover process: chromatid interference, the nonrandom choice of chromatids involved in adjacent chiasma, and crossover interference, the nonrandom placement of chiasma along a chromosome. Experimental evidence in various organisms has shown that crossovers do not occur at

*To whom correspondence should be addressed.

Incorporating interference into linkage analysis

random, rather they are more evenly spaced along a chromosome (e.g. Hultén, 1974; Blank *et al.*, 1988). Chromatid interference can only be detected if all four products of meiosis can be recovered. However, to date there has been little evidence of this type of interference in experimental organisms.

For experimental crosses, the most common method of linkage analysis in use is the Lander–Green (LG) algorithm (Lander and Green, 1987). This algorithm makes use of a hidden Markov model (HMM) to construct a genetic map using multilocus linkage analysis. The LG approach is based on the assumption that there is no interference, either crossover or chromatid (NI model). The crossover positions are assumed to follow a Poisson distribution and the nonsister chromatids are chosen at random for each crossover.

Various models for the crossover process have been introduced over the years (e.g. Karlin and Liberman, 1979; Risch and Lange, 1979; King and Mortimer, 1990; McPeck and Speed, 1995). The χ^2 distribution was first used in a model for crossing over in Fisher *et al.* (1947). We consider here a four-strand version of this model. Foss *et al.* (1993) gave a simple biological motivation and interpretation for the model based on gene conversions. They proposed that gene conversions are intermediate events that can resolve in reciprocal crossing over but do not need to do so. The χ^2 model assumes that m noncrossover resolutions of gene conversion occur between each crossover. The gene conversion events follow a Poisson distribution, and the distance between actual crossovers follows a χ^2 distribution with $2(m+1)$ degrees of freedom. This model has been found to fit observed data from a variety of organisms as well (Foss *et al.*, 1993; Zhao *et al.*, 1995; Broman *et al.*, 2002), indeed better than other models (McPeck and Speed, 1995; Broman and Weber, 2000). Yet, despite the literature on alternative crossover models, in practice the phenomenon of interference is routinely ignored in genetic-mapping exercises. At most, the Kosambi map function is used after recombination fractions have been estimated under the NI model.

Geneticists have also known for many years that the rate of recombination differs, often substantially, between males and females for organisms such as humans and the mouse. In other species, such as the Lepidoptera *Bombyx mori* and *Drosophila melanogaster*, recombination does not occur at all in one of the sexes. This is equivalent to the genetic distances between markers being zero since no recombination is observed between them. Unfortunately, due to lack of parameter identifiability it is not possible to estimate sex-specific distances from an F_2 cross. The two heterozygous genotypes are indistinguishable, so that if both recombination fractions are greater than zero, the chromosome on which the crossover occurred cannot be determined. In the case of no recombination in one sex, genetic distances can be estimated from a group of F_2 progeny. However, if a model of equal recombination between the sexes is used, the genetic distance estimates will be too small as the model produces, in effect, sex-averaged genetic distance estimates. When recombination is absent in one sex, these distances will be much smaller than the true distances in the recombinant sex.

In this paper, we present an extension of the LG algorithm to incorporate crossover interference according to the χ^2 model. The following section presents the algorithm for both the backcross (BC) and F_2 cases. The special case of no recombination in one sex is also considered for the F_2 progeny. In the results section, the accuracy of this extended algorithm in estimating genetic maps is evaluated through simulation studies and compared to the original LG algorithm and the Kosambi transformation. The extra time required to obtain genetic maps when incorporating crossover interference is also considered.

2. MODEL

Let $\mathcal{M}_1, \dots, \mathcal{M}_T$ denote T genetic loci listed in order along a chromosome. The observed genotypic information can be denoted $O = (O_1, \dots, O_T)$, where O_t contains the genotypes of all individuals at $\mathcal{M}_t$. Define $\lambda_t = 2(m+1)d_t$, where d_t is the genetic distance, in Morgans, between $\mathcal{M}_t$ and $\mathcal{M}_{t+1}$ and m is a defined constant. An important difference between the LG and χ^2 algorithms is that LG estimates the recombination fraction while the χ^2 algorithm estimates λ_t . If $m = 0$, this gives a direct estimate of

N. J. ARMSTRONG ET AL.

the genetic distance and no transformation is necessary. For $m > 0$, it estimates a fixed multiple of the genetic distance.

Under the χ^2 model, there are assumed to be m unobserved crossover intermediates between each crossover. The hidden chain y_i for the χ^2 model keeps track of both the number of crossover intermediates that have occurred since the last crossover and the origin of the DNA, grandmaternal (gm) or grandpaternal (gp). Specifically, the states $i = 0, \dots, m$ represent the event of i crossover intermediates and gm origin of the DNA, while states $i = m+1, \dots, 2m+1$ represent the event of $i - (m+1)$ crossover intermediates and gp origin of the DNA. The process along a chromosome can be described by the embedded chain in Figure 1. Every $(m+1)$ st event results in a crossover that involves the strand of interest with probability $\frac{1}{2}$. For example, starting in state 0 we must progress through m steps to state m , the next intermediate which will result in a crossover. When in state m , there is either a crossover and switch to gp DNA (state $m+1$) or there is no crossover and we stay on gm DNA, returning to state 0. Note that if $m = 0$, the chain is that of the NI model with the hidden states simply representing regions of gm and gp DNA on the chromosome.

Conditional on parental genotypes, the genotypes of each offspring are independent. Hence, the likelihood may be calculated for each individual separately and then combined. In the discussion that follows, the HMM is defined for one individual. The superscript k is used when necessary to denote the k th offspring and omitted otherwise. It is also assumed, without loss of generality, that the F_1 parent is crossed with individuals from the grandmaternal strain to produce BC progeny.

2.1 Backcross

Given the inbred strains AA and aa, at any locus the BC offspring will have genotype Aa or the genotype of the homozygous parent (either aa or AA depending on which line the homozygous parent came from). We define the observed genotypes at a given locus of a BC animal to be A if homozygous and H if

Fig. 1. Embedded chain along a chromosome for the χ^2 model with parameter m .

Incorporating interference into linkage analysis

heterozygous. Hence, for a BC experiment, the elements of O are A, H and —, the three possible observed genotypes, where ‘—’ denotes missing. In this case, we need consider only the chromosome inherited from the F_1 parent.

The initial-state distribution is uniform over all states. The transition probability matrices, $A(\lambda_t)$, at each marker locus have $2(m+1) \times 2(m+1)$ entries of which only $3m+2$ are distinct. The entries $a_{ij}(t)$ can be related to the distinct elements $a_l(t)$ by grouping the pairs (i, j) into sets I_l :

1. If $0 \leq i, j \leq m$ or $i, j > m$, $(i, j) \in I_l$, where $l = j - i$.
2. If $0 \leq i \leq m, j > m$, then for $j > i + m$, $(i, j) \in I_l$, where $l = j - i$, else $(i, j) \in I_l$, where $l = j - i - (m + 1)$.
3. If $i > m, 0 \leq j \leq m$, then for $j + (m + 1) > i - 1$, $(i, j) \in I_l$, where $l = j - i + 2m + 2$, else $(i, j) \in I_l$, where $l = j - i + (m + 1)$.

For example, consider $a_{12}(t)$ when $m = 2$. The chain progresses from state 1 to state 2 if there has been a single intermediate event, if there have been four intermediate events but no crossover has occurred ($1 \rightarrow 2 \rightarrow 0 \rightarrow 1 \rightarrow 2$), or if any multiple of four intermediate events and an even number of crossovers have occurred. This is also the same set of circumstances under which the chain could move from state 4 to state 5. In this case, the transition matrix is given by:

$$A(\lambda_t) = \begin{pmatrix} a_0 & a_1 & a_2 & a_3 & a_4 & a_5 \\ a_{-1} & a_0 & a_1 & a_{-1} & a_3 & a_4 \\ a_{-2} & a_{-1} & a_0 & a_{-2} & a_{-1} & a_3 \\ a_3 & a_4 & a_5 & a_0 & a_1 & a_2 \\ a_{-1} & a_3 & a_4 & a_{-1} & a_0 & a_1 \\ a_{-2} & a_{-1} & a_3 & a_{-2} & a_{-1} & a_0 \end{pmatrix}.$$

Defining

$$f_{m,s} = \frac{1}{2} e^{-\lambda_t} \sum_{k=1}^{\infty} \frac{\lambda_t^{k(m+1)+s}}{(k(m+1)+s)!},$$

the distinct $a_l(t)$ are

$$a_l(t) = \begin{cases} f_{m,l}(\lambda_t), & -m \leq l \leq -1, \\ \frac{e^{-\lambda_t} \lambda_t^l}{l!} + f_{m,l}(\lambda_t), & 0 \leq l \leq m, \\ f_{m,l-(m+1)}(\lambda_t), & m+1 \leq l \leq 2m+1. \end{cases}$$

The probability that a particular genotype is observed at a given marker is conditional on the hidden state at that marker, that is $b_i(O_t) = P(O_t | y_t = i)$. Random errors in the genotyping process of rate ϵ are allowed. If there is no genotyping error, $0 \leq y_t \leq m$ implies $O_t = A$ and $m+1 \leq y_t \leq 2m+1$ implies $O_t = H$.

Likelihood calculations are done by way of the ‘forward variables’:

$$\alpha_t(i) = P(O_1, \dots, O_t, y_t = i | \lambda).$$

N. J. ARMSTRONG ET AL.

For n offspring, the likelihood is given by

$$P(O|\lambda) = \prod_{k=1}^n \sum_{i=0}^{2m+1} \alpha_T^k(i), \quad (2.1)$$

where $\alpha_T^k(i)$ denotes the value of $\alpha_T(i)$ for the k th offspring. The forward variables give the probability of observing the partial genotype sequence up to $\mathcal{M}_t$ and being in hidden state i at $\mathcal{M}_t$. Likewise, there are corresponding ‘backward variables’, $\beta_t(i)$, giving the probability of observing the partial genotype sequence from $\mathcal{M}_{t+1}$ to $\mathcal{M}_T$ given the hidden state i at $\mathcal{M}_t$. Both the forward and backward variables are computed recursively.

The expectation maximization (EM) algorithm is employed to determine the maximum likelihood estimates (MLEs) of the intermarker genetic distances, d_t , $t = 1, \dots, T-1$. In particular, Baum’s lemma is used, so that $Q(\lambda, \lambda')$ is explicitly maximized, where

$$Q(\lambda, \lambda') = \sum_y P(y, O|\lambda) \log P(y, O|\lambda')$$

and $y = (y_1, \dots, y_T)$. In terms of λ'_t this is equivalent to maximizing, for n BC offspring,

$$\sum_{k=1}^n \sum_{y_t^k, y_{t+1}^k} P(y_t^k, y_{t+1}^k | O^k, \lambda) \log P(y_{t+1}^k | y_t^k, \lambda'_t) = \sum_{k=1}^n \sum_{l=-m}^{2m+1} E(n_l^k(t) | O, \lambda) \log a_l(\lambda'_t),$$

where $n_l^k(t) \in \{0, 1\}$ is defined to be the number of transitions from state i to state j for a pair $(i, j) \in I_l$ for the k th offspring. The E step then involves calculating

$$E(n_l^k(t) | O, \lambda) = \sum_{i, j \in I_l} \xi_t^k(i, j),$$

where $\xi_t(i, j) = \frac{a_t(i)a_{t+1}(j)b_j(O_{t+1})\beta_{t+1}(j)}{\sum_{i, j} a_t(i)a_{t+1}(j)b_j(O_{t+1})\beta_{t+1}(j)}$.

For general m there is no closed-form solution to the maximization step. Instead, we must search for the maximum,

$$\operatorname{argmax}_{\lambda'_t} \sum_{k=1}^n \sum_{l=-m}^{2m+1} \log a_l(\lambda'_t) E(n_l^k(t) | O, \lambda).$$

Brent’s (1973) method is used to find a solution numerically, iterating the E and M steps until the scaled genetic distance estimates, λ_t , change by less than a prespecified tolerance level.

2.2 F_2 intercross

Given the two inbred strains AA and aa, at a given locus each F_2 offspring will be homozygous AA, homozygous aa, or heterozygous Aa. The observed genotypes are defined as A or B (homozygous AA or aa) and H if heterozygous. We can further define two incomplete genotypes C and D to be not B (i.e. AA or Aa) and not A (i.e. aa or Aa), respectively. For an F_2 intercross, both chromosomes must be considered because it is impossible to determine, at each heterozygous locus, which allele came from which parent. There are $4(m+1)^2$ hidden states, since on each chromosome the chain could be in any one of the states $0, \dots, 2m+1$. For example, if $m = 1$, there are 16 hidden states, shown here with their corresponding

Incorporating interference into linkage analysis

states on each individual chromosome:

$0 \leftrightarrow 0, 0$	$4 \leftrightarrow 1, 0$	$8 \leftrightarrow 2, 0$	$12 \leftrightarrow 3, 0$
$1 \leftrightarrow 0, 1$	$5 \leftrightarrow 1, 1$	$9 \leftrightarrow 2, 1$	$13 \leftrightarrow 3, 1$
$2 \leftrightarrow 0, 2$	$6 \leftrightarrow 1, 2$	$10 \leftrightarrow 2, 2$	$14 \leftrightarrow 3, 2$
$3 \leftrightarrow 0, 3$	$7 \leftrightarrow 1, 3$	$11 \leftrightarrow 2, 3$	$15 \leftrightarrow 3, 3$

Again, the convention that states $0, \dots, m$ represent gm DNA and $m+1, \dots, 2m+1$ gp DNA on a single chromosome is used. The initial-state distribution is assumed uniform. The conditional distribution of the observed genotypes given the hidden state i at $\mathcal{M}_t$ allows for the incomplete marker genotypes C and D, and incorporates a random error rate of ϵ .

The transitions on the maternal and paternal chromosomes are independent; hence, the joint probability is simply the product of the marginal probabilities of the transitions on each chromosome. Formally, the transition probability matrix for the states is given by

$$C(\lambda_t^f, \lambda_t^p) = A(\lambda_t^f) \otimes A(\lambda_t^p),$$

where $A(\lambda_t^f)$ and $A(\lambda_t^p)$ are the transition matrices on the maternal and paternal chromosomes, respectively, and $\otimes$ the Kronecker product. It is generally assumed that the recombination rates for the two sexes are equal, i.e. $\lambda_t^f = \lambda_t^p = \lambda_t$ and $C(\lambda_t) = A(\lambda_t) \otimes A(\lambda_t)$. In general, the (nondistinct) $c_{ij}(t)$ are obtained by translating the indices i and j into the corresponding indices for the transition matrix $A(\lambda_t)$ for each individual chromosome. The algorithm detailed in the previous section is used to determine the distinct quantities in both cases ($a_{s_1}(t)$ and $a_{s_2}(t)$), giving

$$c_{ij}(t) = a_{s_1}(t)a_{s_2}(t), \quad -m \leq s_1 \leq s_2 \leq 2m+1.$$

For example, when $m = 1$ the states are given above and the probability of moving from state 1 to state 4 involves a transition from states 0 to 1 on the first chromosome and 1 to 0 on the second chromosome. Using the notation from Section 2.1, the marginal probabilities for these transitions are $a_1(t)$ and $a_{-1}(t)$, respectively. Therefore, in this case, $c_{14}(t) = a_1(t)a_{-1}(t)$.

The special case of no recombination in one sex only affects the transition matrix in the model. Without loss of generality, if $\lambda_t^f = 0$ for $t = 1, \dots, T-1$ (i.e. no female recombination), then:

$$A(\lambda_t^f) = \mathcal{I}_{2m+1},$$

where $\mathcal{I}_{2m+1}$ is the $(2m+1) \times (2m+1)$ identity matrix. The joint transition matrix then simplifies to

$$C(\lambda_t^f, \lambda_t^p) = \begin{pmatrix} A(\lambda_t^p) & 0 & \cdots & 0 \\ 0 & A(\lambda_t^p) & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & 0 & A(\lambda_t^p) \end{pmatrix}.$$

In other words, only the block diagonal elements of $C(\lambda_t^f, \lambda_t^p)$ are nonzero.

Having defined the elements of the HMM, calculation of the forward, backward, and ξ variables can then be carried out as in Section 2.1, substituting $c_{ij}(t)$ in place of $a_{ij}(t)$. In order to find the MLEs of genetic distance, we invoke the EM algorithm once more. For n F₂ offspring and equal recombination

N. J. ARMSTRONG ET AL.

rates for the sexes, i.e. $\lambda^f = \lambda^p$, $Q(\lambda, \lambda')$ in terms of λ'_t becomes, in the general case,

$$\begin{aligned} & \sum_{k=1}^n \sum_{y_t^k, y_{t+1}^k} P(y_t^k, y_{t+1}^k | O^k, \lambda) \log P(y_{t+1}^k | y_t^k, \lambda'_t) \\ &= \sum_{k=1}^n \sum_{I_1, I_2=-m}^{2m+1} \log[a_{I_1}(\lambda'_t) a_{I_2}(\lambda'_t)] E(n_{I_1, I_2}^k(t) | O, \lambda). \end{aligned}$$

$n_{I_1, I_2}^k(t)$ is the number of transitions from i to j for a pair (i, j) related to I_{I_1} and I_{I_2} on the individual chromosomes. The E step consists then of calculating $E(n_{I_1, I_2}^k(t) | O, \lambda)$ and the M step is carried out numerically using Brent's (1973) method. To avoid underflow problems, the forward and backward variables are scaled by way of logarithms for all values of m and for all progeny numbers, when there are more than 100 markers in the data set for the BC HMM and, similarly, when there are more than 70 markers in the F_2 case.

3. RESULTS

To test the performance of the χ^2 HMM, genotype data were simulated for both F_2 and BC breeding schemes. The data were simulated under the assumption of crossover interference. To mimic biological reality, the χ^2 model with $m = 6$ was used. This model was previously estimated to provide the best fit for mouse data (Lin and Speed, 1996).

The programs were first verified by comparing all results (estimated genetic distances and likelihood values) obtained when $m = 0$ against both MapMaker and R/qtl (Broman *et al.*, 2003) which assume no crossover interference. Furthermore, for the BC case, the theoretical and empirical map functions were compared for a variety of m values. For fixed m , 1000 BC mice were simulated and genotyped every centimorgan along a single chromosome. Using this genotype data, the empirical recombination fractions could then be computed. Plotting the empirical recombination fractions against the true genetic distance and the theoretical map function, given by

$$\theta = \frac{1}{2} \left[1 - e^{-2d} \sum_{i=0}^m \frac{(2d)^i}{i!} \left(1 - \frac{i}{m+1} \right) \right],$$

on the same set of axes found close agreement. Figure 2 depicts the results for $m = 0, 1, 2$, and 3.

3.1 *Bombyx mori*

A set of F_2 genotype data for *B. mori*, the domesticated silkworm, was obtained from D. G. Heckel (Shi *et al.*, 1995). In Lepidoptera, females are the heterogametic sex and have achiasmatic meiosis (no crossing over). The purpose of the original study was to create a genetic linkage map for *B. mori*. The data consist of the genotypes of 52 F_2 progeny at a total of 58 markers. The markers are both dominant and codominant. Fifty of the autosomal loci were grouped into a total of 15 primary linkage groups (PLGs), ranging from 2 to 8 loci each, by Shi *et al.* (1995). The PLGs were taken as originally assigned and the genetic maps were re-estimated using the HMM with $m = 0$ in order to verify the published map. In addition to estimating the intermarker genetic distances, the log-likelihood values for all possible map orders were calculated conditional on their PLG assignment to verify the published orders for all PLGs. The results, both estimated intermarker genetic distances and map orders, matched the published maps for all PLGs except PLG 9 and 15 (Figure 3).

Incorporating interference into linkage analysis

Fig. 2. Empirical and theoretical map functions for various values of m .

Fig. 3. Genetic maps for PLG 9 and 15, as estimated by Shi *et al.* (1995) (listed first) and by using the HMM (second).

N. J. ARMSTRONG ET AL.

PLG 15 consists of three dominant markers. The slight difference in estimated distances can be explained by the different estimation methods used. Shi *et al.* (1995) analyzed the linkage data under the assumptions of no crossing over in the female and no crossover interference, using maximum likelihood estimation and the EM algorithm. For some triple genotype combinations of dominant markers the Markov property of the observed genotype information O is lost, e.g. $P(O_3 = C | O_1 = C, O_2 = C) \neq P(O_3 = C | O_1 = A, O_2 = C)$, making it incorrect to calculate the likelihood using a Markov chain, as was done by Shi *et al.* Hence, multilocus probabilities must be calculated some other way, such as by using an HMM. The inclusion of the hidden chain allows for the correct calculation of multilocus probabilities when there is dominance or missing data at some loci. Our method therefore makes more efficient use of the data for this PLG.

The largest PLG was PLG 9, consisting of eight loci. For this PLG, Shi *et al.* (1995) used MapMaker to estimate an approximation to the maximum likelihood map, as their program was too slow to calculate the maps for all possible orders in a reasonable time. Furthermore, the recombination fractions between loci were estimated from the MapMaker sex-averaged estimates using a linear regression (for details see Shi *et al.*, 1995). In contrast, using the HMM approach, the likelihood values for all possible orders were calculated and the genetic distances estimated under the assumption of no crossing over in the female. We concluded that the order of loci was the same as the published order but that the genetic distance estimates were different. Given their ad hoc method of estimating the distances for this PLG, it is not surprising that the HMM distance estimates differ from their published ones (Figure 3).

3.2 Timing

As m increases, the number of states in the HMM also increases. Recall there are $2(m+1)$ states for BC progeny and $4(m+1)^2$ states for F_2 progeny. The increase in the number of states leads, necessarily, to an increase in the time taken to converge to the maximum likelihood estimates due to more computations being required to calculate the forward, backward, and ξ variables. The time required to perform the E step also increases with m . However, finding the values of λ_i which maximize $Q(\lambda, \lambda')$ at each iteration (the M step) requires only a slight increase in time. An important issue with regard to the time taken to find the estimated genetic distances is that λ_i , not d_i , is estimated. The shape of $Q(\lambda_i, \lambda'_i)$ rapidly becomes very flat as the value of λ_i approaches the maximum likelihood estimate, so that in order to get an accurate estimate of d_i many iterations are often required. It is the increased number of iterations of the EM algorithm coupled with the increase in E step time which gives rise to the large overall increase in time.

The time taken to find the maximum likelihood estimates for varying m values was investigated for both BC and F_2 cases. The impact of increasing the number of mice in the data set and, separately, the impact of increasing the number of genetic markers were investigated. The value of m used for simulating the genotype data was equal to the value of m used in the HMM. Initial values for the distances were taken to be the NI ($m = 0$) estimates.

All programs were run on 400-MHz UltraSparc machines with 256 MB RAM. The HMM does not require huge amounts of memory, but the time taken depends quite heavily on the speed of the machine. Under each set of conditions investigated, the differences in the time taken for each data set were directly attributable to the number of iterations of the EM algorithm required before the convergence tolerance was satisfied. The time taken rises linearly as both the number of markers and the number of mice rise, with times also increasing with m . Kinks in Figure 4 can be attributed to the fact that the machines used were not available solely for our purposes. However, the graphs give a rough indication of the increase in time required.

The relationship between number of mice, time taken, and m is very similar for BC (Figure 4) and F_2 mice (not shown), except that the times involved are much larger in the F_2 case. For 1600 mice, it took

Incorporating interference into linkage analysis

Fig. 4. Impact of increasing the number of mice on the time taken to find the distance MLEs for 10 intervals in the BC case.

the BC HMM 0.4 s, on average, to compute distance estimates for 10 intervals when $m = 0$, whereas when $m = 6$ the time taken rose to 321 s (5.5 min), an 800-fold increase. In the F_2 case, it took 12.2 s for $m = 0$, and 235 880 s (65.5 h) for $m = 6$, an almost 20 000-fold increase. When investigating the impact of the number of markers in a data set on time taken to find the distance MLEs, the number of mice was fixed at 400 in the BC case but reduced to 100 for F_2 data. Again, the relationship between the number of markers, m , and the amount of time taken to calculate the MLEs was very similar for F_2 and BC mice (data not shown). The time taken to calculate distance estimates rises linearly with an exponential increase in markers, similar to that shown in Figure 4 for the number of mice. For example, for 32 markers, the time taken to produce distance estimates when $m = 0$ was, on average, 4.4 s and increased to approximately 13 h if $m = 6$, for F_2 data, a 10 000-fold increase. In the BC case, under the no-interference model, estimates were obtained in 0.18 s compared to 163 s (2.7 min) if $m = 6$.

3.3 Accuracy

The most important gain that we hope to make by extending the LG algorithm to incorporate crossover interference is in the accuracy of the estimates. In order to investigate this, both BC and F_2 mice were simulated under the χ^2 model with $m = 6$. In each instance, 1000 sets of 300 BC mice and 100 sets of 300 F_2 mice were simulated and the distance estimates obtained using both the NI model and the χ^2 model recorded. The accuracies of distance estimates obtained using the Kosambi map function to transform the recombination fractions obtained under the NI model were also compared.

The entries in Table 1 demonstrate that as the distance between markers increases, the gains in precision from using the χ^2 model also increase when the data are fully informative, even if the wrong level

N. J. ARMSTRONG ET AL.

Table 1. Root mean squared error in centimorgans of the distance estimates obtained under various models for fully informative BC and F₂ data simulated under the χ^2 model with $m = 6$

Model used	True distance (cM)					
	BC data			F ₂ data		
	5	10	20	5	10	20
NI	1.38	2.28	3.90	0.87	1.58	3.04
$\chi^2_{10}(m = 4)$	1.24	1.82	2.46	0.78	1.27	1.93
$\chi^2_{14}(m = 6)$	1.23	1.81	2.38	0.78	1.26	1.87
$\chi^2_{18}(m = 8)$	1.23	1.81	2.34			
NI + Kosambi	1.25	1.89	2.78	0.79	1.32	2.17

Fig. 5. Box plot of distance estimates obtained when true distance is 10 cM. In this case 300 BC meioses were simulated with 1% random genotyping error.

of interference is used in the HMM. Introducing a random genotyping error into the genotype data when either no error or the wrong amount of error is allowed for in the HMM produces similar results. For all distances the χ^2 model is more accurate, but the gains in accuracy are sometimes slight, especially if the genetic distances involved are small. Figure 5 shows the results found under one scenario, the results being typical of those found under other scenarios also. Simulations with 5% and 10% missing genotype data, and both 2% and 4% incomplete genotypes for the F₂ case were also performed. We found that this conclusion holds true in those cases also (data not shown). The results presented in Figure 6 are representative of those found under the different conditions investigated, for both F₂ and BC data.

4. CONCLUSIONS

We have developed and tested an HMM for estimating genetic distances that incorporates crossover interference, specifically under the χ^2 model. The HMM for F₂ data was extended to incorporate the situation where there is no crossing over in one of the sexes. The χ^2 HMM produces more accurate estimates of

Incorporating interference into linkage analysis

Fig. 6. Box plot of distance estimates obtained when true distance is 5 cM. In this case, 300 F_2 meioses were simulated with 5% missing and 4% incomplete genotyping data.

genetic distance, in the presence of crossover interference, than the NI model which is commonly implemented.

The gains in accuracy achieved by using the χ^2 model, with $m = 6$, instead of the Kosambi map function, are small for distances under 10 cM. With the ever-increasing number of genetic markers available and the falling costs of genotyping, in future we can expect many markers to be typed for a particular experimental cross, leading to a reduction in the typical intermarker genetic distance. The use of single-nucleotide polymorphisms and microarray technology for genotyping purposes will also mean smaller distances between markers. Using a lower level of interference than true ($m = 4$ instead of $m = 6$) resulted in very little difference in the level of accuracy obtained, likewise if a slightly higher m is used in the HMM. The problem of estimation of m has been treated in detail by Zhao *et al.* (1995) and Goldstein *et al.* (1997). The method they use is to compute the maximum likelihood for different values of m and compare them. The HMM presented in this paper may also be used to estimate the level of interference using the same approach of comparing likelihood values for different m values (NJ Armstrong and TP Speed, in preparation). The computational cost is equal to the sum of the computational costs of maximizing the likelihood for the individual values of m . The accuracy of the estimate obtained is in large part dependent on the number of individuals in the cohort and the number of markers genotyped. A more realistic model for the distribution of crossovers along a chromosome is a mixture of the χ^2 and NI models (Copenhaver *et al.*, 2002). For small-mixture probabilities, the results on accuracy of distance estimates are likely to be similar to those presented here.

The time taken to find estimates of genetic distance for F_2 data with either a large number of markers or progeny, for large levels of interference ($m = 6$), is at least 1000 times more than if no interference is assumed, and can take as much as 20 000 times longer. The increase in time required in conjunction with the small increases in accuracy suggests that this model will not become widely implemented in practice, even with the rapid increase in computing power. Development of the HMM allowing for crossing over in only one sex, for both the χ^2 model and the NI model, will hopefully be of use for researchers involved with organisms such as those in the order Lepidoptera.

The software used in this paper is available on request from the first author.

N. J. ARMSTRONG ET AL.

ACKNOWLEDGMENTS

We thank David Heckel for kindly allowing us access to the *Bombyx mori* data. This work was supported by National Institutes of Health grant GM59506-01 (to Speed).

REFERENCES

- BLANK, R., CAMPBELL, G., CALABRO, A. AND D'EUSTACHIO, P. (1988). A linkage map of mouse chromosome 12: localization of Igh and effects of sex and interference on recombination. *Genetics* **120**, 1073–1083.
- BRENT, R. (1973). *Algorithms for Minimization Without Derivatives*. Englewood Cliffs, NJ: Prentice-Hall.
- BROMAN, K., ROWE, L., CHURCHILL, G. AND PAIGEN, K. (2002). Crossover interference in the mouse. *Genetics* **160**, 1123–1131.
- BROMAN, K. AND WEBER, J. (2000). Characterization of human crossover interference. *American Journal of Human Genetics* **66**, 1911–1926.
- BROMAN, K., WU, H., SEN, S. AND CHURCHILL, G. (2003). R/qtl: Qtl mapping in experimental crosses. *Bioinformatics* **19**, 889–890.
- COPENHAVER, G., HOUSWORTH, E. AND STAHL, F. (2002). Crossover interference in Arabidopsis. *Genetics* **160**, 1631–1639.
- FISHER, R., LYON, M. AND OWEN, A. (1947). The sex chromosome in the house mouse. *Heredity* **1**, 335–365.
- FOSS, E., LANDE, R., STAHL, F. AND STEINBERG, C. (1993). Chiasma interference as a function of genetic distance. *Genetics* **133**, 681–691.
- GOLDSTEIN, D., ZHAO, H. AND SPEED, T. (1997). The effects of genotyping errors and interference on estimation of genetic distance. *Human Heredity* **47**, 86–100.
- HULTÉN, M. (1974). Chiasma distribution at diakinesis in the normal human male. *Hereditas* **76**, 55–78.
- KARLIN, S. AND LIBERMAN, U. (1979). A natural class of multilocus recombination processes and related measures of crossover interference. *Advances in Applied Probability* **11**, 479–501.
- KING, J. AND MORTIMER, R. (1990). A polymerization model of chiasma interference and corresponding computer simulation. *Genetics* **126**, 1127–1138.
- LANDER, E. AND GREEN, P. (1987). Construction of multilocus genetic linkage maps in humans. *Proceedings of the National Academy of Science of the United States of America* **84**, 2363–2367.
- LIN, S. AND SPEED, T. (1996). Incorporating crossover interference into pedigree analysis using the χ^2 model. *Human Heredity* **46**, 315–322.
- MCPEEK, M. AND SPEED, T. (1995). Modeling interference in genetic recombination. *Genetics* **139**, 1031–1044.
- MULLER, H. (1916). The mechanism of crossing-over. *The American Naturalist* **50**, 193–221, 284–305, 350–366, 421–434.
- RISCH, N. AND LANGE, K. (1979). An alternative model of recombination and interference. *Annals of Human Genetics* **43**, 61–70.
- SHI, J., HECKEL, D. AND GOLDSMITH, M. (1995). A genetic linkage map for the domesticated silkworm, *Bombyx mori*, based on restriction fragment length polymorphisms. *Genetic Research (Cambridge)* **66**, 109–126.
- ZHAO, H., SPEED, T. AND MCPEEK, M. (1995). Statistical analysis of crossover interference using the chi-square model. *Genetics* **139**, 1045–1056.

[Received April 19, 2005; first revision August 12, 2005; second revision October 7, 2005; third revision November 21, 2005; accepted for publication December 7, 2005]