A score test for linkage using identity by descent data from sibships

Sandrine Dudoit and Terence P. Speed

Department of Statistics
University of California, Berkeley
367 Evans Hall, #3860
Berkeley, CA 94720-3860

Technical Report No. 528
July 1998

We consider score tests of the null hypothesis $\mathrm{H}_0 : \theta = \frac{1}{2}$ against the alternative hypothesis $\mathrm{H}_1 : 0 \leq \theta < \frac{1}{2}$, 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 $T(\frac{1}{2}) = \mathbf{1}^T \alpha$, $\mathbf{1} = (1, \ldots, 1)$. 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 is 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 rather unusual class of score tests which arises naturally in human genetics. However, their essence can be described quite efficiently without any of the genetic background, and we now do so. Let $\alpha = (\alpha_1, \ldots, \alpha_m)$ and $\pi = (\pi_1, \ldots, \pi_m)$ 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 $\mathbf{1} = (1, \ldots, 1)$. 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 $\mathrm{H}_0 : \theta = \frac{1}{2}$ against the alternative $\mathrm{H}_1 : 0 \leq \theta < \frac{1}{2}$, that is, for testing $\mathrm{H}_0 : \rho = \alpha$ against alternatives along the curve $\mathcal{C}_\pi(\theta)$, based upon counts $N = (N_1, \ldots, N_m)$ 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 problem 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 to deal with the nuisance parameter π in this score test. This study was motivated by the observation that in some important cases in our genetic context, 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 and based on a widely used statistic in linkage analysis. This fact is not only convenient in applications, but also suggests that the test might enjoy some measure of model robustness, that is, perform well against a broad class of alternatives. The test has also been described as "model-free". We thought it would be of interest to learn just how far this property extended, and if possible, to understand its origins. 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 sibships. This involves describing how IBD patterns in pedigrees 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 the 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 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, the score test for affected only sibships doesn't depend on the nuisance parameter π and is based on a well-known statistic in linkage analysis, S_{pairs} (cf. Kruglyak *et al.* [10], Whittemore and Halpern [16]).

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 mosaics of the two parental chromosomes (see Figure 1). 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 *co-segregation*, 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 transmission to a child. Exceptions to co-segregation occur for loci on the same chromosome 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. This fraction is a monotonic function of the physical distance between the loci; it is 0 when they are essentially one locus, and reaches a maximum of $\frac{1}{2}$ when the loci are widely separated on the same chromosome or on different chromosomes. 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, e.g. $\theta < 0.05$ (see Ott [11] for an introduction to linkage analysis).

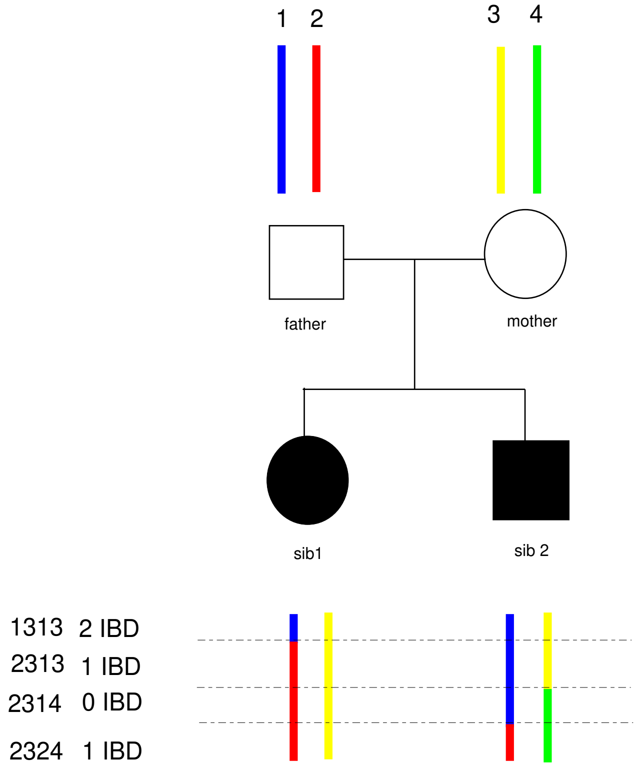

Figure 1: Segregation products for a sibship of size 2 and a single chromosome pair. Male and female individuals are represented by squares and circles, respectively, and colored symbols indicate affectedness by the disease under study. The paternal and maternal chromosome pairs are labeled by $(1, 2)$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

When mapping disease susceptibility (DS) genes, we are interested in testing whether genetic markers with known location are linked or not to DS genes, i.e. in testing a null hypothesis of the form $\mathrm{H}_0 : \theta = \frac{1}{2}$, where θ is the recombination fraction between a genetic marker and a putative DS gene. This could be done by studying the co-segregation of variants of the DS genes with those of other mapped genes or markers. (By now, there are scores of well-mapped markers along each human, mouse and many other chromosomes.) Frequent co-segregation 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 the same ancestral chromosome. This is by contrast to *identity by state*, 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 associations between the *sharing of DNA identical by descent at loci linked to DS loci* and *disease status* in families with affected individuals. At loci unlinked to DS loci, IBD sharing is independent of disease status. For

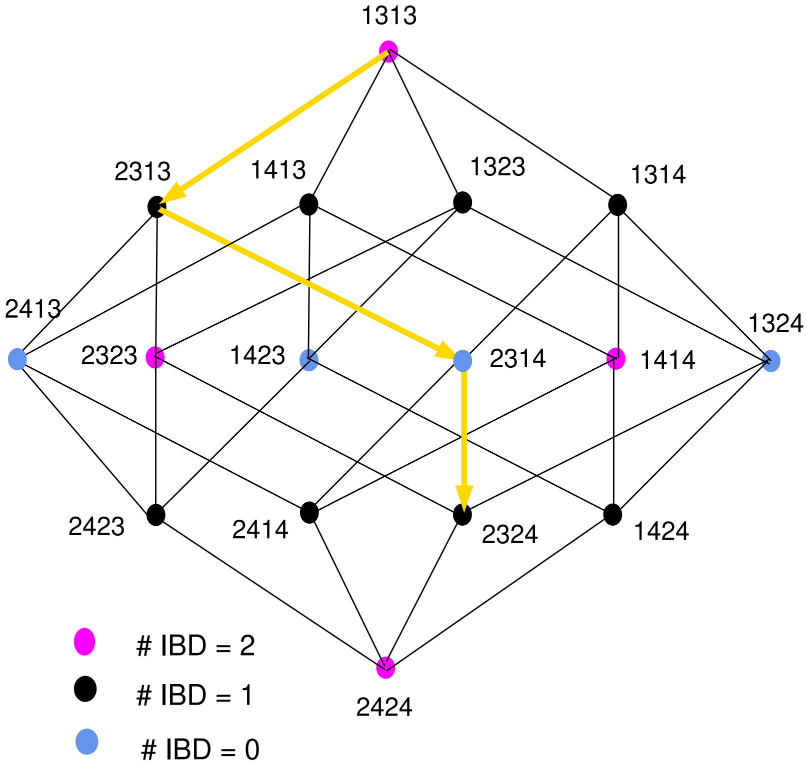

Figure 2: 4-dimensional hypercube whose vertices correspond to the 16 possible inheritance vectors for a sib-pair and whose edges represent permissible transitions. The arrows indicate the transitions for the segregation products represented in Figure 1.

example, for sibships, this approach builds on the simple notion that if susceptibility to the disease under study has a genetic component, then 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 [4] for a simple example of this association in sib-pairs). We expect it to work because full sibs get all their genes from the same source, their parents, but only 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 we can readily determine IBD status at so-called marker loci, one of which may be tightly linked to the DS locus (see Kruglyak *et al.* [10] for a treatment of incomplete IBD data). This approach will thus be successful if (and only if) (a) there is a noticeable association between disease status of the sibs and their IBD status at a DS locus, and (b) 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 disease and IBD status [4]. If we have a dense enough set of marker loci, problem (b) would appear to be solved, but in truth, there is always a trade-off between the magnitude of the association in (a) and the density of the marker set necessary for its detection. These issues were addressed in a recent paper by E.A. Thompson [14], who refers to the two components (a) and (b) as the *specificity* of the DS locus and the *scale* of the genetic distance, respectively.

The IBD pattern within a pedigree may be summarized at any point of the genome 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 $x = (x_1, x_2, \ldots, x_{2k-1}, x_{2k})$, indicating the outcome of each of the $2k$ segregations giving rise to the sibship. More precisely, for $i = 1, \ldots, k$, 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 from which sib i inherited DNA at $\mathcal{L}$, 3 or 4 (see Figure 1). 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. Since the chance of a recombination between the two loci is the recombination fraction and recombination events are independent across segregations, then the transition matrix $R(\theta)$ between inheritance vectors at loci separated by a recombination fraction θ has entries

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

where $\Delta(x, y)$ is the number of coordinates at which the inheritance vectors x and y differ, i.e. 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

$$R(\theta) = \begin{bmatrix} 1-\theta & \theta \\ \theta & 1-\theta \end{bmatrix}^{\otimes 2k}. \tag{2}$$

Note that we are assuming equal male and female recombination fractions, otherwise, we would have a transition matrix for paternal segregations and a transition matrix for maternal segregations.

The notion of inheritance vector may be extended to any type of pedigree and the transition matrix $R(\theta)$ has the same form. For one segregation, the recombination 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. Jointly, the recombination 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 [3], Proposition 1 and Figures 1 and 2). The random walk model for the recombination process is widely used and is referred to in the genetics literature as the *no interference model.*

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 and genealogies. Although not needed for an understanding of this paper, a discussion of these probabilities and the genetic model under which they are calculated may be found in [4]. For affected only sibships, and without distinguishing between sharing of maternal and paternal DNA, Ethier and Hodge [5] show how the 2^{2k} inheritance vectors may be grouped into a much smaller number of equivalence classes which we call *identity by descent (IBD) configurations.* 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: (i) interchange the paternal labels 1 and 2, (ii) interchange the maternal labels 3 and 4, (iii) interchange the parental origin of the DNA by interchanging 1 and 3 and 2 and 4, and (iv) permute the sibs. For example, for sib-pairs, one usually considers three IBD configurations, corresponding to the number of chromosomes sharing DNA IBD at the locus, instead of the 16 inheritance vectors.

Table 1: Sib-pair IBD configurations.

Number IBD	Representative inheritance vector
0	(1,3,2,4)
1 paternal	(1,3,1,4)
1 maternal	(1,3,2,3)
2	(1,3,1,3)

For a pedigree with given genealogy and phenotype, the conditional probability vector of IBD configurations at a 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. Under the null hypothesis of no linkage, the IBD distribution at the marker is

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

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 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. Figure 4 p. 15 is a barycentric representation of curves $\mathcal{C}_\pi(\theta)$ for the sib-pair transition matrix.

Suppose we collect n pedigrees with a given genealogy and phenotype, 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 pedigrees with IBD configuration $i = 1, \ldots, m$ at the genetic marker. Under certain sampling assumptions (manuscript in preparation), $(N_1, \ldots, N_m)$ have a Multinomial$(n, \rho(\theta, \pi))$ distribution. There is no uniformly most powerful test of $\mathrm{H}_0 : \theta = \frac{1}{2}$, however, the *score test* is locally most powerful. Although this set-up applies to any type of pedigree, the IBD configurations and hence $T(\theta)$ are different depending on the genealogy and phenotype of the pedigree. Thus, different pedigrees will yield different score statistics for testing linkage. Remarkably, for affected only sibships, the score statistic doesn't 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 proportion of chromosomes on which they share DNA IBD at the marker (cf. Kruglyak *et al.* [10], Whittemore and Halpern [16]). 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* $\mathrm{H}_0 : \theta = \frac{1}{2}$ *is based on* S_{pairs}*, regardless of the model for disease susceptibility, i.e. regardless of* π*. For one affected sib-k-tuple*

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

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* $\mathrm{Binomial}(2, \frac{1}{2})$ *random variables, and thus*

$$\mathrm{E}_0\left[S_{pairs}\right] = \frac{1}{2}, \qquad \mathrm{Var}_0\left[S_{pairs}\right] = \frac{1}{4k(k-1)}.$$

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

Thus, for affected only sibships, 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 *et al.* [9] to sibships of any size and to general genetic models with multiple unlinked DS loci and no population genetic assumptions such as random mating and Hardy-Weinberg equilibrium. Unfortunately, this simple property does not hold with all types of sibships. Below, we will consider examples where it fails.

The remainder of this paper will be concerned with the proof of this result, and with deriving score statistics for general sibships, with any number of affected 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 genealogy and the choice of IBD configurations. Thus, we will first describe how inheritance patterns may be summarized by IBD configurations which are orbits of groups acting on the set of inheritance vectors.

3 Transition matrix for sibship 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 $x : \{1, \ldots, k\} \to \{a, b, c, d\}$. 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 [6] Section 3.2 for an introduction to group action). Let

α	$=$	$(ac)(bd)$	interchange labels 1 and 2 of paternal chromosomes,
β	$=$	$(ab)(cd)$	interchange labels 3 and 4 of maternal chromosomes,
γ	$=$	(bc)	interchange parental origin of DNA.

The group of permutations of the square 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, $C_2 \times C_2$. The IBD configurations of Ethier and Hodge [5] 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 of the square D_4, acting on $\mathcal{X}$. In some situations (e.g. parental 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.

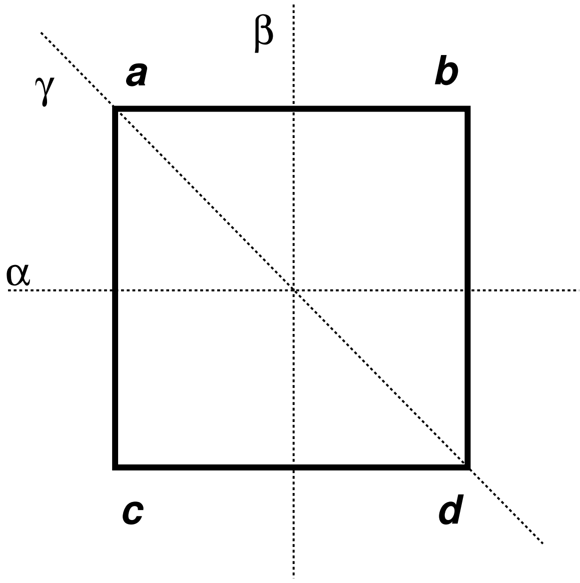

Figure 3: Permutations α, β and γ of the vertices of the square.

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. 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 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* (cf. van Lint and Wilson [15], deBruijn [2]) to find the number of orbits of groups acting on mappings, and hence determine the number of IBD configurations of each type (see Appendix 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. p.264,265 in Ethier and Hodge [5] and Appendix A).

Table 2: Sibship IBD configurations

# affected, # unaffected	Distinguish maternal from paternal sharing	Group 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	$(S_h \times S_{k-h}) \times (C_2 \times C_2)$

In the next section, we will derive properties of transition matrices between IBD configurations for the four groups just described.

3.2 Properties of transition matrix

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,

$$\sum_{y \in \mathcal{C}_j} r_{xy}(\theta) = \sum_{y \in \mathcal{C}_j} r_{\tilde{x}y}(\theta) \qquad \text{for any } x,\ \tilde{x} \in \mathcal{C}_i. \tag{3}$$

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

$$\begin{aligned} t_{ij}(\theta) &= \sum_{y \in \mathcal{C}_j} \theta^{\Delta(x,y)}(1-\theta)^{2k-\Delta(x,y)}, \qquad \text{where } x \text{ is any } x \in \mathcal{C}_i & (4) \\ &= \frac{|\mathcal{C}_j|}{|\mathcal{C}_i|} \sum_{x \in \mathcal{C}_i} \theta^{\Delta(x,y)}(1-\theta)^{2k-\Delta(x,y)}, \qquad \text{where } y \text{ is any } y \in \mathcal{C}_j. \end{aligned}$$

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 B for proofs).

Proposition 1 *Let $\theta_1 * \theta_2 = \theta_1(1-\theta_2) + \theta_2(1-\theta_1)$. Then $T(\theta)$ satisfies the semi-group 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.

$$Q = T'(0) = B - 2kI,$$

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

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

and $I(\,)$ denotes the indicator function. T has stationary distribution

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

and T is reversible, i.e.

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

See Ott [11] for an introduction to map functions. 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, i.e. no crossover interference. Also, note that we didn't need to assume no crossover interference to derive the semi-group property. If however we 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 equation (3) and condition (15) p. 63 in Rosenblatt [12], 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 equation (4), but we gain more knowledge on the form of the score statistic by using the following spectral decomposition of $T(\theta)$.

Proposition 2 *The transition matrix $T(\theta)$ may be written as*

$$T(\theta) = \sum_h e^{\lambda_h d(\theta)}\, P_h = \sum_h (1-2\theta)^{-\lambda_h/2}\, P_h, \tag{5}$$

where the λ_h are the m real eigenvalues of the infinitesimal generator Q, and the P_h are projection matrices satisfying $P_h^2 = P_h = P_h^*$, $P_h P_l = 0$, $h \neq l$, and $\sum_h P_h = I$. P_h^* is the adjoint of P_h with respect to the inner product $< x, y >_\alpha = \sum_i \alpha_i x_i y_i$. In particular, the first two derivatives of the transition matrix with respect to θ are

$$T'(\theta) = \sum_h \lambda_h (1-2\theta)^{-(\lambda_h+2)/2} P_h, \tag{6}$$

and

$$T''(\theta) = \sum_h \lambda_h(\lambda_h+2)(1-2\theta)^{-(\lambda_h+4)/2} P_h. \tag{7}$$

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 of Q corresponding to λ_h and with unit norm with respect to the inner product $<,>_\alpha$.

Thus, eigenvalues of Q and their multiplicity give us information regarding the derivatives of the transition matrix T 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 non-zero 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

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

$\mathcal{X}$ is the graph defined by the first associates in the Hamming scheme $H(2k, 2)$ (cf. Chapter 30 in van Lint and Wilson [15]). Consider any of the four groups H described in Table 2. The matrix B, defined in Proposition 1, is the *adjacency matrix of the quotient graph $\mathcal{X}/H$*, which is the multi-digraph with the orbits of 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 H (see Appendix C for proof).

Theorem 1 Eigenvalues of infinitesimal generator Q.
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 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 $\left\{-2i_{\binom{2k}{i}} : i = 3, \ldots, 2k\right\}$, where the subscript $\binom{2k}{i}$ is the largest possible multiplicity of the eigenvalue $-2i$. Thus, from equations (6) and (7)

$$T'\Big(\frac{1}{2}\Big) = 0 \tag{8}$$

and

$$U = T''\Big(\frac{1}{2}\Big) = 8P_{-4}, \tag{9}$$

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 $\alpha_j \sum v_i v_j$ where the v's are right eigenvectors of Q with unit norm with respect to the inner product $<,>_\alpha$ and the sum is over all such eigenvectors.

Note that we may also show that $T'(\frac{1}{2}) = 0$ by simple algebra.

4 Linkage score test

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, $i = 1, \ldots, m$, 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 unlinked DS loci, that is, we wish to test

$$\mathrm{H}_0:\ \theta = \frac{1}{2} \qquad \text{(no linkage)} \qquad \text{versus} \qquad \mathrm{H}_1:\ 0 \leq \theta < \frac{1}{2} \qquad \text{(linkage)},$$

where θ denotes the recombination fraction between $\mathcal{M}$ and $\mathcal{D}$.

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 non-zero 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 to be

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

where $U = T''(\frac{1}{2}) = 8P_{-4} = (8\alpha_j \sum v_i v_j)$, the v's are right eigenvectors of Q corresponding to the eigenvalue -4, with unit norm with respect to the inner product $<,>_\alpha$, and the sum in U is over all such eigenvectors. We will see 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 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 *et al.* [10], S_P of Whittemore and Halpern [16], and $PAIRS$ and WP of Suarez and Van Eerdewegh [13]). For a sibship of size k, S_{pairs} is obtained by forming all possible pairs of sibs and averaging the proportions of chromosomes on which they share DNA IBD, that is,

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

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.
For affected sib-k-tuples, without distinguishing between sharing of maternal and paternal DNA, the linkage score test is the same as the test based on S_{pairs}. The contribution of n affected sib-k-tuples to the overall score statistic is

$$S = \frac{2^{4k-7}}{k(k-1)}\Big(\sum_{j=1}^{m} u_{j1}\pi_j\Big)\Big(\sum_{i=1}^{m} u_{i1}N_i\Big) = 2^{2k-2}\Big(\sum_{j=1}^{m} u_{j1}\pi_j\Big)(2S_{pairs} - n). \tag{10}$$

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

$$\begin{aligned} S_{pairs} &= \frac{\sum_{i=1}^{4} a_i(x)(a_i(x)-1)}{2k(k-1)} \\ &= \frac{a_1(x)^2 + a_2(x)^2 + a_3(x)^2 + a_4(x)^2 - 2k}{2k(k-1)}, \end{aligned} \tag{11}$$

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, i.e. with representative inheritance vector $(1,3,1,3,\ldots,1,3)$ and label $(0,0,0)$ in the notation of Ethier and Hodge [5]. 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 independent of the nuisance parameter π (i.e. doesn't depend on 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. On the other hand, the score statistic for combining IBD data from sibships of different sizes involves weights which do depend on the genetic model ($\sum_{i=1}^m u_{i1}\pi_i$ = expected value of $\sum_{i=1}^m u_{i1}N_i/n$ when $\theta = 0$, i.e. "right on top" of the gene), but require the computation of only the first column of the matrix of second derivatives. As we will see next, it is not always the case that the score test for a particular type of sibship is independent of the genetic model.

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 (cf. 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 eigenvectors of Q corresponding to the second largest eigenvalue and with unit norm with respect to the inner product $< .,. >_\alpha$. These eigenvectors are based on V_e and V_o, respectively (see Appendix C). The score statistic is given by

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

Thus, in general, the score test depends on the parameters of the genetic model π. In some situations however (e.g. no imprinting, that is same maternal and paternal contribution to disease susceptibility), 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 (cf. Theorem 1). Hence, in general, the score statistic depends on the genetic model and is a sum of terms of the form $\left(\sum_{j=1}^m \pi_j v_j\right)(\sum_{i=1}^m N_i v_i)$, where v_i is one of the eigenvectors of Q corresponding to $\lambda = -4$.

In the next section, we will 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.

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 $\bar{\psi} = 1 - \psi$. 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 $\lambda = -4$ are $(\frac{1}{2\sqrt{2}}, 0, -\frac{1}{2\sqrt{2}})$ and $(\sqrt{2}, 0, -\sqrt{2})$, respectively (right eigenvector has unit norm with respect to the inner product $< .,. >_\alpha$). 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 in the more common form $N_2 + \frac{1}{2}N_1$, known as the "mean IBD" statistic (cf. Knapp *et al.* [9]).

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

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

and the infinitesimal generator is

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

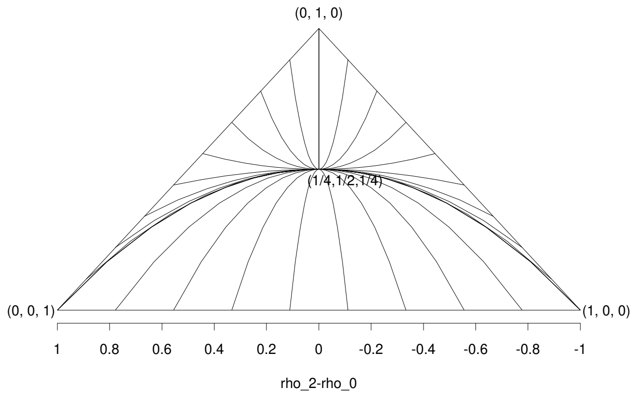

Figure 4: $S_2 \times D_4$ - Barycentric representation of curves $\mathcal{C}_\pi(\theta)$, $0 \leq \theta \leq \frac{1}{2}$, for π on boundaries of simplex.

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

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

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 parameters of the model for disease susceptibility. 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 [5]) listed in Table 3.

Table 3: IBD configurations for affected sib-trios

IBD Configuration i	Representative inheritance vector	Label	$\lvert\mathcal{C}_i\rvert$
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

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 $\lambda = -4$ 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

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

5.4 Discordant sib-trios, orbits of $(S_1 \times S_2) \times D_4$

For discordant sib-trios, 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.

The infinitesimal generator is

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

Table 4: IBD configurations for discordant sib-trios

IBD Configuration i	Representative inheritance vector	$\lvert\mathcal{C}_i\rvert$
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

Q has eigenvalues $\lambda = 0, -4, -4, -8, -8, -8, -8$, and the two right eigenvectors corresponding to $\lambda = -4$ are $v = \sqrt{\frac{2}{3}}(-1,-1,-1,-1,1,1,3)$ and $\tilde{v} = \frac{1}{\sqrt{3}}(4,1,-2,-2,-1,2,0)$, respectively. 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, $i = 1, \ldots, 7$. Then, the score statistic for testing linkage is

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

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 defined IBD configurations to be the orbits of groups acting on the set of inheritance vectors and proved that the change in IBD configurations along a chromosome was embeddable in a continuous time Markov chain, without the assumption of no crossover interference. In our genetic context, the i-th derivative, $i = 0, \ldots, 2k$, of the IBD configuration transition matrix $T(\theta)$ is given by

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

where the λ_h and P_h are the eigenvalues and projection matrices of the infinitesimal generator Q , respectively (cf. Proposition 2). 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. If the second largest eigenvalue of Q is $-2i$, the score statistic is based on the i-th derivative of the log-likelihood. For affected only sibships, the second largest eigenvalue -4 has multiplicity one, and as a result, the score test is based on the second derivative of the log-likelihood and is independent of the genetic model (i.e. of the nuisance parameter π). Furthermore, the score statistic reduces to a well-known statistic in linkage analysis, S_{pairs}.

A recent study of Davis and Weeks [1] on sib-pairs found S_{pairs} to perform well for a variety of genetic models. However, the performance of S_{pairs} for larger sibships still needs to be studied. When combining IBD data from different types of sibships, the weights depend on the genetic model through the IBD probabilities π and the robustness of the test to misspecifications of the genetic model remains to be addressed.

It would be useful to derive score tests of linkage for IBD data from other types of small pedigrees (e.g. 1 cousin and 2 sibs). Donnelly [3] considered relative pairs, such as cousin-pairs, grand-parent/grandchild and uncle/nephew pairs, and partitioned the inheritance vectors into a much smaller number of orbits. Note that these orbits are not necessarily the usual IBD configurations corresponding to sharing DNA IBD on 0 or 1 chromosome, and the transition matrix for these usual configurations doesn't always satisfy the semi-group property.

The problem of testing linkage using IBD data is part of a general type of testing problems in which we wish to test whether a Markov chain has reached its stationary distribution. 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 this study.

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 [15], the number of orbits of $G \times H$ acting on B^A, the set of mappings from A to B, is given by

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

where

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

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

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

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

The cycle index for the symmetric group on n letters is

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

where $(1^{k_1} \ldots n^{k_n})$ denotes a partition of n with k_i parts of size i, $i = 1, \ldots, n$. $Z_{S_n}(X_1, \ldots, X_n)$ is also the coefficient of z^n in the expansion of $\exp\left(\sum_{i=1}^{\infty} \frac{z^i}{i} X_i\right)$ (deBruijn [2] p.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 $\{a, b, c, d\}^{\{1,\ldots,k\}}$ (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 are calculated using Table 5, and the fact that for $\tau \in D_4$ and $i \geq 0$

$$\begin{aligned} m_4(\tau) &= m_{4i+4}(\tau), \\ m_1(\tau) &= m_{4i+1}(\tau) = m_{4i+3}(\tau), \\ m_2(\tau) &= m_{4i+2}(\tau). \end{aligned}$$

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 [5].

Table 5: Cycles of D_4. $z_j(\tau)$ denotes the number of cycles of τ having length j. The elements of D_4 are listed according to the notation of Fraleigh p. 70.

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

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

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

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

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

$$\begin{aligned} m = \frac{1}{8}\Bigg[&\binom{h+3}{3}\binom{k-h+3}{3} + 2I(4|h)I(4|k-h) \\ &+ \frac{3}{4}I(2|h)I(2|k-h)(h+2)(k-h+2) \\ &+ \frac{1}{8}\Big(I(2|h) + (h+1)(h+3)\Big)\Big(I(2|k-h) + (k-h+1)(k-h+3)\Big)\Bigg]. \end{aligned}$$

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

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

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. Ethier and Hodge [5]

Table 7: $Z_{S_k}(m_1(\tau),\ldots,m_k(\tau))$ for $\tau\in D_4$. $I(\,)$ is the indicator function.

Permutation τ	$\exp(\sum_{i=1}^{\infty} m_i(\tau)\frac{z^i}{i})$	$Z_{S_k}(m_1(\tau),\ldots,m_k(\tau))$
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	$(1-z)^{-2}(1-z^2)^{-1}$	$\frac{1}{4}(I(2\|k)+(k+1)(k+3))$

define the label of the particular inheritance vector x to be the triple (l_1,l_2,l_3) where

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

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

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

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

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

B Transition matrix for sibship IBD configurations

B.1 Proof of Proposition 1

Let $\bar\theta=1-\theta$. We first prove the semi-group 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 $i=0,\ldots,\Delta$, $j=0,\ldots,2k-\Delta$, 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

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

Therefore,

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

Consider now the transition matrix for IBD configurations. From equation (4)

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

Hence, $T(\theta)$ satisfies the semi-group 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

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

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

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

where x is any inheritance vector in $\mathcal{C}_i$. Q may be written as $Q = B - 2kI$ where B is the matrix with entries

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

$T(\theta)$ satisfies

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

hence, the stationary distribution of T is

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

since

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

□

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 $< x, y >_\alpha = \sum_i \alpha_i x_i y_i$ on $\Re^m$. Hence, from the Principal Axis Theorem (Jacob [8], p.288), Q has an orthonormal basis of eigenvectors with only real eigenvalues, λ_h, $h = 1, \ldots, m$ (not necessarily distinct). Denote the h-th (column) eigenvector by $\mathbf{v}_h$ and its i-th entry by v_{ih}. Then, $< \mathbf{v}_h, \mathbf{v}_l >_\alpha = \sum_i \alpha_i v_{ih} v_{il} = \delta_{hl}$. 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

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

i.e.

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

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 $<,>_\alpha$. It follows that

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

□

C Adjacency matrix of quotient graph $\mathcal{X}/H$

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 $x = (x_1, x_2, \ldots, x_{2k})$ 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.

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

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

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

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

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

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

Let H *denote the matrix with rows the* 2^{2k} *eigenvectors of* A *described above. Then,* H *is an Hadamard matrix, i.e. 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

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

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

$\lambda = 2k - 2$:

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

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

$\lambda = 2k - 4$:

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

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

In order to prove Theorem 1, we rely on the following general facts concerning quotient graphs (cf. Godsil [7], Chapter 5). Consider a group H acting on the vertices of $\mathcal{X}$, as described in Table 2. Then, by the same argument as that leading to equation (3), the orbits of H, $\mathcal{C}_i$, $i = 1, \ldots, m$, 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* $\mathcal{X}/H$, which is the multi-digraph with the orbits of 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 [7])*
The eigenvalues of B are a subset of the eigenvalues of A.

Fact 2 *(based on Lemma 2.2 in Godsil [7])*
If v is an eigenvector of B, then Cv is an eigenvector of A which is constant over the orbits of 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 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.*

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

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

□

The proof of Theorem 1 also relies on the following specific properties of the eigenvectors of A on the orbits of 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 $S_k \times D_4$, i.e. $\forall\ i = 1, \ldots, 2k$, 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 $S_k \times D_4$, i.e. $\forall\ i, j = 1, \ldots, k$, 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

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

Hence

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

$\square$

Fact 6 *Let*

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

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 $S_k \times D_4$, i.e. for any orbit $\mathcal{C}$

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

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 will 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 $1 \leq i \leq k$

$$V_{2i-1}(\tilde{x}) = V_{2i}(x), \tag{12}$$

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

$\tilde{x} = (\sigma, \iota)(x) \in \mathcal{C}$, 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

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

In particular, for $k > 1$

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

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

Fact 7 *The $k(k-1)$ 2-factor eigenvectors $\{V_{2i-1,2j-1}, V_{2i,2j} : i < j\}$ have the same sums over the orbits of $S_k \times D_4$, i.e. 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 equation (12) for each $1 \le i < j \le k$

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

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 $\mathcal{X}/(S_k \times D_4)$.
$2k$ and $2k-4$ are eigenvalues of B with multiplicity one. All other eigenvalues of B are strictly less than $2k-4$ and belong to the set $\left\{2(k-i)_{\binom{2k}{i}} : i = 3, \dots, 2k\right\}$, 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

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

by letting

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

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

$\lambda = 2k$: 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 be constant and non-zero 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

$$\begin{aligned}
W_{2i-1,2j} &= V_{2i-1,2j} - \frac{< V_{2i-1,2j}, V >}{|V|^2} V \\
&= V_{2i-1,2j}, \qquad 1 \le i, j \le k, \\
W_{2i-1,2j-1} &= V_{2i-1,2j-1} - \frac{< V_{2i-1,2j-1}, V >}{|V|^2} V \\
&= V_{2i-1,2j-1} - \frac{1}{k(k-1)} V, \qquad 1 \le i < j \le k, \\
W_{2i,2j} &= V_{2i,2j} - \frac{< V_{2i,2j}, V >}{|V|^2} V = V_{2i,2j} - \frac{1}{k(k-1)} V, \qquad 1 \le i < j \le k.
\end{aligned}$$

By Fact 5, for any orbit $\mathcal{C}$
$$\sum_{x\in\mathcal{C}} W_{2i-1,2j}(x) = 0.$$
Also, by Fact 7
$$\sum_{x\in\mathcal{C}} W_{2i-1,2j-1}(x) = \sum_{x\in\mathcal{C}} V_{2i-1,2j-1}(x) - \frac{1}{k(k-1)} \sum_{(i,j)} \sum_{x\in\mathcal{C}} \{V_{2i-1,2j-1}(x) + V_{2i,2j}(x)\} = 0,$$
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.

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 $\{V_{2i-1,2j-1} : i < j\}$ have the same sums over the orbits of $S_k \times (C_2 \times C_2)$, i.e. for any orbit $\mathcal{C}$ and $1 \le i_1 < j_1 \le k$, $1 \le i_2 < j_2 \le k$*
$$\sum_{x\in\mathcal{C}} V_{2i_1-1,2j_1-1}(x) = \sum_{x\in\mathcal{C}} V_{2i_2-1,2j_2-1}(x).$$
Similarly for the $k(k-1)/2$ 2-factor eigenvectors $\{V_{2i,2j} : i < j\}$.

Proposition 5 Eigenvalues of adjacency matrix B of quotient graph $\mathcal{X}/(S_k \times (C_2 \times C_2))$.
$2k$ and $2k-4$ are eigenvalues of B with multiplicities one and two, respectively. All other eigenvalues of B are strictly less than $2k-4$ and belong to the set $\left\{2(k-i)_{\binom{2k}{i}} : i = 3, \ldots, 2k\right\}$. 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 $\left\{2(k-i)_{\binom{2k}{i}} : i = 0, \ldots, 2k\right\}$.

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

$\lambda = 2k-2$: 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 non-zero 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 $Span\{V_o, V_e\}$ in the eigenspace of A for $\lambda = 2k-4$ is spanned by the following $2k^2 - k$ vectors

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

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 $Span\{V_o, V_e\}$ 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

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

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 $1 \leq i_1 < j_1 \leq h$, $1 \leq i_2 < j_2 \leq h$*

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

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 $\mathcal{X}/((S_h \times S_{k-h}) \times D_4)$.
$2k$ is an eigenvalue of B with multiplicity one. $2k-4$ is an eigenvalue of B 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 $\left\{2(k-i)_{\binom{2k}{i}} : i = 3, \ldots, 2k\right\}$. 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 $\left\{2(k-i)_{\binom{2k}{i}} : i = 0, \ldots, 2k\right\}$. 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 $(h \leq k-2)$, 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 $Span\{V^a, V^u, V^{au}\}$ 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

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

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

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

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

Also, by Fact 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 $Span\{V^a, V^u, V^{au}\}$ 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 $2 \le h \le k-2$ 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:

$$
\begin{aligned}
V_e^{a}(x) &= \sum_{1 \le i < j \le h} V_{2i,2j}(x), \qquad h \ge 2, \\
V_o^{a}(x) &= \sum_{1 \le i < j \le h} V_{2i-1,2j-1}(x), \qquad h \ge 2, \\
V_e^{u}(x) &= \sum_{h+1 \le i < j \le k} V_{2i,2j}(x), \qquad h \le k-2, \\
V_o^{u}(x) &= \sum_{h+1 \le i < j \le k} V_{2i-1,2j-1}(x), \qquad h \le k-2, \\
V_e^{au}(x) &= \sum_{1 \le i \le h,\ h+1 \le j \le k} V_{2i,2j}(x), \\
V_o^{au}(x) &= \sum_{1 \le i \le h,\ h+1 \le j \le k} V_{2i-1,2j-1}(x).
\end{aligned}
$$

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

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

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

$$\begin{aligned} S &= \sum_{i=1}^{m} N_i \frac{\sum_{j=1}^{m} \pi_j 8\alpha_i v_i v_j}{\alpha_i} \\ &= 8\Big(\sum_{j=1}^{m} \pi_j v_j\Big)\Big(\sum_{i=1}^{m} N_i v_i\Big). \end{aligned}$$

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, i.e. with representative inheritance vector $(1, 3, 1, 3, \ldots, 1, 3)$ and label $(0, 0, 0)$ in the notation of Ethier and Hodge.

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

By differentiating equation (4) we find that

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

Thus, the contribution of an affected sib-k-tuple with inheritance vector $x \in \mathcal{C}_i$ to the score statistic is based on

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

Hence, from equation (11) p. 12

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

Since $u_{11} = 2^{5-2k}k(k-1)$ and $|\mathcal{C}_1| = 4$, then

$$S = 2^{2k-2} \Big(\sum_{j=1}^{m} \pi_j u_{j1} \Big) (2S_{pairs} - n),$$

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

□

Ackowledgments. We would like to thank Steve Evans for many helpful discussions on Markov chains, and Cheryl Praeger, Alice Niemeyer and Nick Wormald for graph-related ideas.

References

[1] S. Davis and D. E. Weeks. Comparison of nonparametric statistics for detection of linkage in nuclear families: single-marker evaluation. *Am. J. Hum. Genet.*, 61:1431–1444, 1997.

[2] N. G. deBruijn. Pólya's theory of counting. In E. F. Beckenbach, editor, *Applied Combinatorial Mathematics*, University of California Engineering and Physical Sciences Extension Series. John Wiley & Sons, New York, 1964.

[3] K. P. Donnelly. The probability that related individuals share some section of genome identical by descent. *Theor. Pop. Biol.*, 23:34–63, 1983.

[4] S. Dudoit and T. P. Speed. Triangle constraints for sib-pair identity by descent probabilities under a general multilocus model for disease susceptibility. In M. E. Halloran and S. Geisser, editors, *Statistics in Genetics*, volume 112 of *IMA Volumes in Mathematics and its Applications*. Springer-Verlag, New York, 1998.

[5] S. N. Ethier and S. E. Hodge. Identity-by-descent analysis of sibship configurations. *Am. J. Med. Genet.*, 22:263–272, 1985.

[6] J. B. Fraleigh. *A first course in abstract algebra.* Addison-Wesley Pub. Co., Reading, Mass., 4th edition, 1989.

[7] C. D. Godsil. *Algebraic Combinatorics.* Chapman & Hall mathematics. Chapman & Hall, New York, 1993.

[8] B. Jacob. *Linear Algebra.* W.H. Freeman and Company, New York, 1990.

[9] M. Knapp, S. A. Seuchter, and M. P. Baur. Linkage analysis in nuclear families 1: Optimality criteria for affected sib-pair tests. *Hum. Hered.*, 44:37–43, 1994.

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

[11] J. Ott. *Analysis of Human Genetic Linkage.* Johns Hopkins University Press, Baltimore, revised edition, 1991.

[12] M. Rosenblatt. *Random processes.* Number 17 in Graduate texts in mathematics. Springer-Verlag, New York, 2nd edition, 1974.

[13] B. K. Suarez and P. van Eerdewegh. A comparison of three affected-sib-pair scoring methods to detect HLA-linked disease susceptibility genes. *Am. J. Med. Genet.*, 18:135–146, 1984.

[14] E. A. Thompson. Conditional gene identity in affected individuals. In I-H. Pawlowitzki, J. H. Edwards, and E. A. Thompson, editors, *Genetic mapping of disease genes.* Academic Press Inc., San Diego, London, 1997.

[15] J. H. van Lint and R. M. Wilson. *A Course in Combinatorics.* Cambridge University Press, Cambridge, New York, 1992.

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