A score test for the genetic mapping of complex human traits

Sandrine Dudoit

Department of Biochemistry, Stanford University

Outline

1. Introduction to genetic mapping

2. Sib-pair linkage score test

3. General properties of the linkage score test

Background

The human genome is distributed along 23 pairs of **chromosomes**.

In each pair, one chromosome is paternally inherited, the other maternally inherited.

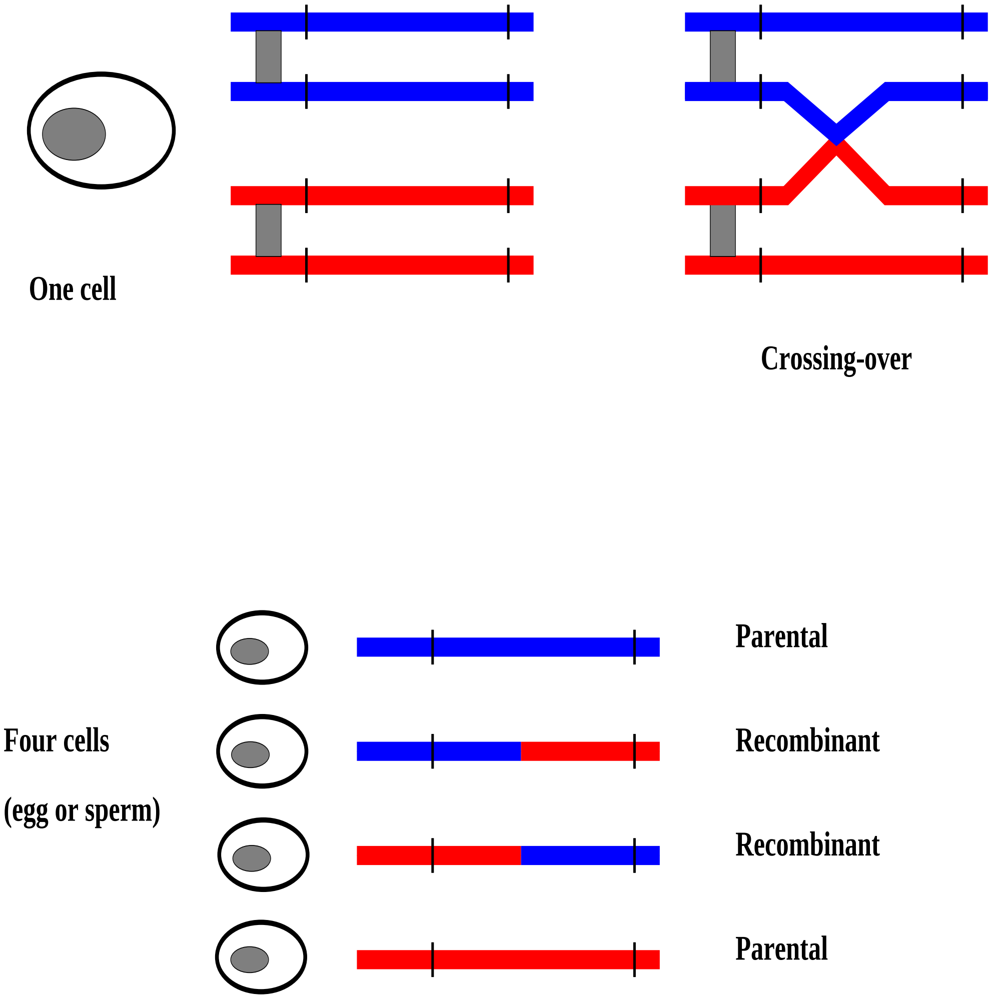
A gene is a segment of chromosomal DNA that directs the synthesis of a **protein**.

Genes have different forms which are called **alleles**.

Genotype: Specific allelic composition of a genome or of certain genes.

Phenotype: Discernible characteristics of an individual. *E.g.* blood pressure.

Meiosis and recombination



One cell

Crossing-over

Four cells

(egg or sperm)

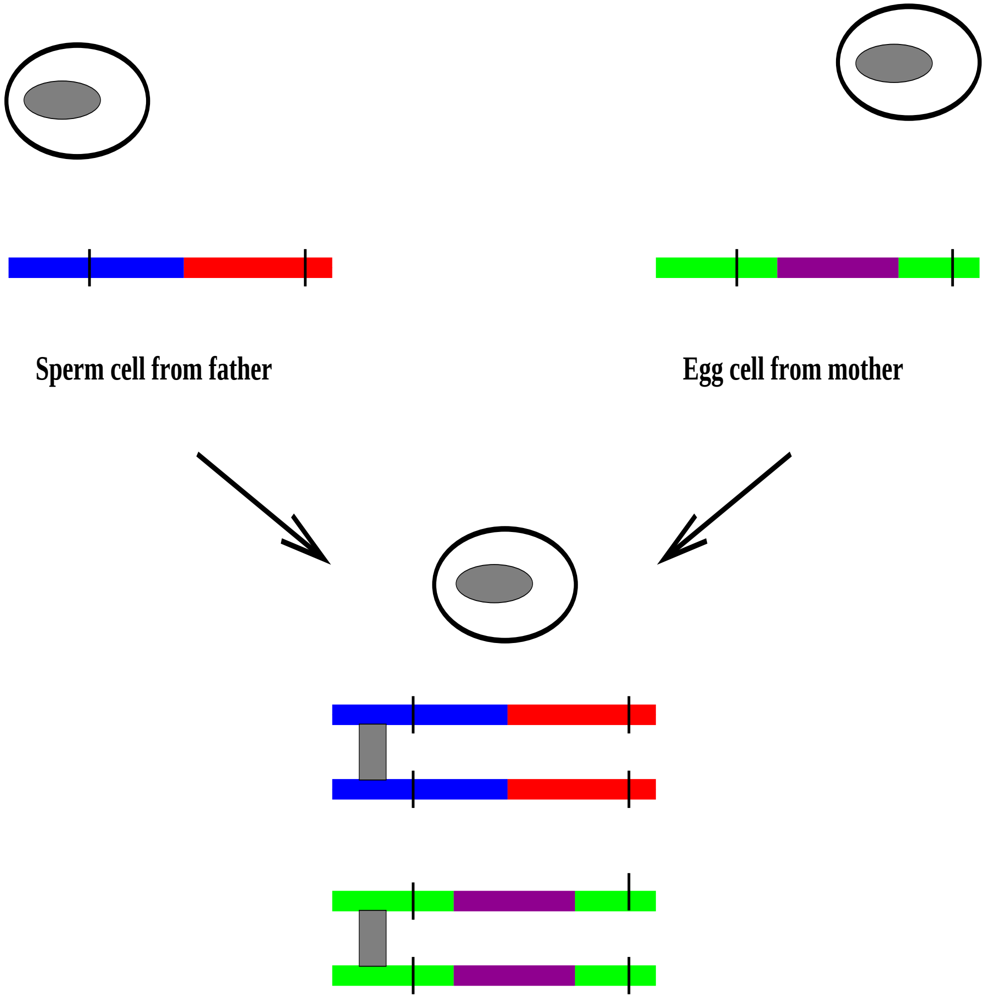
Parental

Recombinant

Recombinant

Parental

Meiosis and recombination



Sperm cell from father

Egg cell from mother

Genetic mapping

Genetic mapping consists of placing genes and other genetic markers on chromosomes $\Rightarrow$ **genetic maps**.

Genetic mapping relies on the varying degree of **recombination** between chromosomal loci to map markers relative to one another.

The distance between two loci is measured by the **recombination fraction** θ, which is the proportion of meiotic products that are recombinant at the two loci.

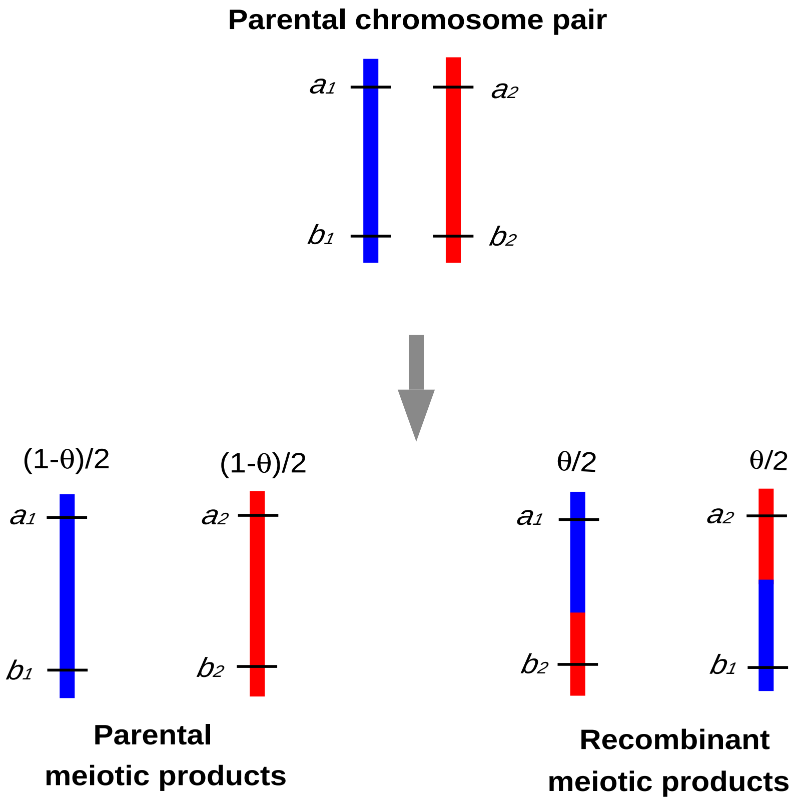
In our model, $0 \leq \theta \leq \frac{1}{2}$.

Two loci are said to be **linked** if $\theta < \frac{1}{2}$, and **unlinked** if $\theta = \frac{1}{2}$.

Model for one meiosis

θ = probability that a meiotic product is recombinant across the interval spanned by the loci $\mathcal{A}$ and $\mathcal{B}$.

There are four different types of meiotic products at loci $\mathcal{A}$ and $\mathcal{B}$.



Model for one meiosis

Joint distribution of meiotic products at loci $\mathcal{A}$ and $\mathcal{B}$

Locus $\mathcal{B}$

		b_1	b_2	
Locus $\mathcal{A}$	a_1	$\frac{1}{2}(1-\theta)$	$\frac{1}{2}\theta$	$\frac{1}{2}$
	a_2	$\frac{1}{2}\theta$	$\frac{1}{2}(1-\theta)$	$\frac{1}{2}$
		$\frac{1}{2}$	$\frac{1}{2}$	

$\theta = \frac{1}{2}$: *Mendel's Second Law*

independent segregation at the two loci,

$\mathcal{A}$ and $\mathcal{B}$ are unlinked.

$\theta = 0$: *Mendel's First Law*

$\mathcal{A}$ and $\mathcal{B}$ behave like one locus.

Model for k meioses

At each locus, summarize the outcome of the k meioses using an **inheritance vector** $x = (x_1, \ldots, x_k)$, where for the ith meiosis

$$x_i = \begin{cases} 0, & \text{grand-paternal DNA transmitted,} \\ 1, & \text{grand-maternal DNA transmitted.} \end{cases}$$

Assume:

1. Independence of all meioses.
2. Constant recombination fractions across individuals and conditions.

Model for k meioses

Mendel's First Law. At any given locus, the 2^k inheritance vectors are equally likely.

Generalization of Mendel's Second Law. Conditional distribution of inheritance vectors at locus $\mathcal{B}$ given $\mathcal{A}$:

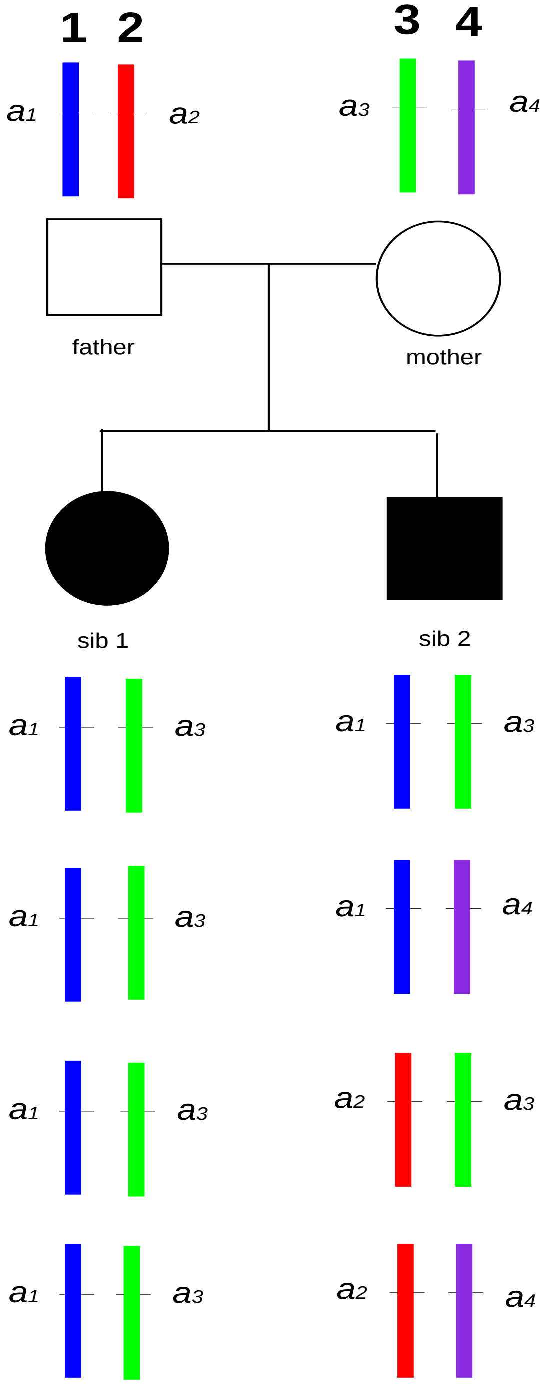
$$R(\theta) \;=\; \begin{bmatrix} 1-\theta & \theta \\ \theta & 1-\theta \end{bmatrix}^{\otimes k}.$$

Identity by descent

DNA at the same locus on two different chromosomes is said to be **identical by descent (IBD)** if it originated from the same ancestral chromosome.

Inheritance vectors may be partitioned into a smaller number of **IBD configurations** which are defined as orbits of groups acting on the set of inheritance vectors.

Sib-pair IBD configurations



Inheritance vector	# IBD
(1,3,1,3)	2
(1,3,1,4)	1
(1,3,2,3)	1
(1,3,2,4)	0

Sib-pair transition matrix $T(\theta)$

Transition matrix for the 2^4 *inheritance vectors*:

$$R(\theta) \;=\; \begin{bmatrix} 1-\theta & \theta \\ \theta & 1-\theta \end{bmatrix}^{\otimes 4}.$$

Transition matrix for the 3 *IBD configurations*:

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

where $\psi = \theta^2 + (1-\theta)^2$ and $\bar{\psi} = 1 - \psi$.

Genetic mapping using IBD data

- The IBD configuration of related individuals at a locus **linked** to a gene is **associated** with their phenotypes.
- The IBD configuration of related individuals at a locus **unlinked** to any genes is **independent** of their phenotypes.

$$\Downarrow$$

Sample groups of related individuals with particular phenotypes and compare the frequencies of IBD configurations at marker loci to the frequencies expected under Mendel's First Law.

Association of phenotype and IBD configuration

E.g. Sib-pair, single gene $\mathcal{D}$, with two alleles D and d, "disease" allele D is fully recessive w.r.t. d - DD individuals are affected, and Dd and dd individuals are unaffected.

Heterozygous parents (Dd).

Table 1: Joint probability of # affected sibs and # chromosomes sharing DNA IBD at the gene.

		# affected sibs			
		0	1	2	
# chromosomes	0	$\frac{1}{8}$	$\frac{1}{8}$	0	$\frac{1}{4}$
sharing DNA IBD	1	$\frac{1}{4}$	$\frac{1}{4}$	0	$\frac{1}{2}$
at gene	2	$\frac{3}{16}$	0	$\frac{1}{16}$	$\frac{1}{4}$
		$\frac{9}{16}$	$\frac{3}{8}$	$\frac{1}{16}$	

Note **association**.

Affected sib-pair method

Sample affected sib-pairs and compare the proportions of sib-pairs sharing 0, 1, 2 IBD at a marker to the Mendelian proportions of $(\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$.

E.g. Cudworth & Woodrow (1975). 15 sib-pairs affected with juvenile-onset diabetes, IBD sharing in the human leucocyte antigen region.

	Observed	Expected
# IBD	# of sib-pairs	# of sib-pairs
		(Mendel's First Law)
0	1	$15 \times \frac{1}{4}$
1	4	$15 \times \frac{1}{2}$
2	10	$15 \times \frac{1}{4}$

$$\chi_2^2 = 14.$$

Unified approach for qualitative and quantitative phenotypes

- Likelihood analysis of IBD data **conditional on phenotypes**.
  - More natural and appropriate.
  - Deals with problematic random sampling assumptions.
  - Single likelihood analysis for IBD data from different pedigree types obtained by various ascertainment mechanisms.
- Test null hypothesis of no linkage between a marker locus and a gene using a **score test in the recombination fraction θ**.
  - Some optimality properties from theory.
  - Some robustness properties apparent in practice.
- Derive score test under **general genetic models for the trait**, which may include multiple unlinked genes and do not make population genetic assumptions.

Sib-pair conditional IBD probabilities at a gene

$\phi = (\phi_1, \phi_2)$ = phenotypes of sib-pair, qualitative or quantitative,
$x = (x_1, x_2, x_3, x_4)$ = inheritance vector of sib-pair at the gene $\mathcal{D}$,
pg = parental genotypes at the gene $\mathcal{D}$.

By Bayes' theorem:

$$pr(x|\phi) = \frac{\sum_{pg} pr(\phi|x, pg)\ pr(x|pg)\ pr(pg)}{\sum_x \sum_{pg} pr(\phi|x, pg)\ pr(x|pg)\ pr(pg)}.$$

For $j = 0,\ 1,\ 2$, let

$$\begin{aligned} \pi_j(\phi_1, \phi_2; \nu) &= pr\big(\text{Sib-pair shares } j \text{ IBD at } \mathcal{D} \mid \phi_1, \phi_2\big) \\ &= \sum_{\{x:\ \#IBD=j\}} pr(x|\phi). \end{aligned}$$

ν: parameters of the genetic model for the trait.

Sib-pair conditional IBD probabilities at a marker

- Marker $\mathcal{M}$ linked to a gene $\mathcal{D}$ at recombination fraction θ.
- (ϕ_1, ϕ_2): phenotypes of sib-pair, qualitative or quantitative.
- For $j = 0, 1, 2$,

$$\pi_j(\phi_1, \phi_2; \nu) = pr\big(\text{Sib-pair shares } j \text{ IBD at } \mathcal{D} \mid \phi_1, \phi_2\big),$$

$$\rho_j(\phi_1, \phi_2; \theta, \nu) = pr\big(\text{Sib-pair shares } j \text{ IBD at } \mathcal{M} \mid \phi_1, \phi_2\big).$$

- ν: parameters of the genetic model for the trait.

Then

$$\Big(\rho_0,\ \rho_1,\ \rho_2\Big) = \Big(\pi_0,\ \pi_1,\ \pi_2\Big) \begin{bmatrix} \psi^2 & 2\psi\bar{\psi} & \bar{\psi}^2 \\ \psi\bar{\psi} & \psi^2 + \bar{\psi}^2 & \psi\bar{\psi} \\ \bar{\psi}^2 & 2\psi\bar{\psi} & \psi^2 \end{bmatrix},$$

where $\psi = \theta^2 + (1-\theta)^2$ and $\bar{\psi} = 1 - \psi$.

Sib-pair conditional IBD probabilities at a marker

$\theta = 0$:

$$T(0) = I_3,$$

$$\Big(\rho_0, \rho_1, \rho_2\Big) = \Big(\pi_0, \pi_1, \pi_2\Big).$$

$\theta = \frac{1}{2}$:

$$T\Big(\frac{1}{2}\Big) = \begin{bmatrix} \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \end{bmatrix},$$

$$\Big(\rho_0, \rho_1, \rho_2\Big) = \Big(\frac{1}{4}, \frac{1}{2}, \frac{1}{4}\Big).$$

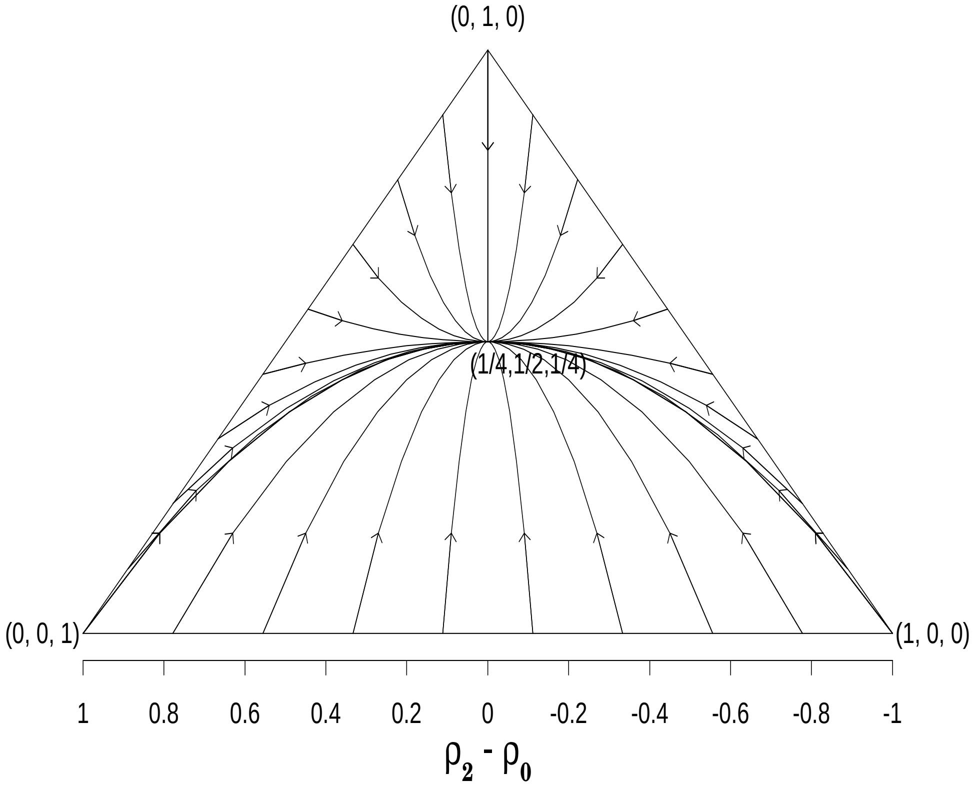


Figure 1: Barycentric representation of curves $\{\rho = \pi T(\theta) : 0 \leq \theta \leq \frac{1}{2}\}$, for $\pi = (\pi_0, \pi_1, \pi_2)$ on boundaries of simplex.

Conditional likelihood of IBD data for sib-pair$^{\mathrm{S}}$

Phenotype and IBD data on n sib-pairs. For ith sib-pair:

- phenotypes (ϕ_{1i}, ϕ_{2i}), qualitative or quantitative,
- IBD indicators $(N_{0i},\ N_{1i},\ N_{2i})$

$$N_{ji} = \begin{cases} 1, & \text{sib-pair shares } j \text{ IBD at marker } \mathcal{M}, \\ 0, & \text{otherwise.} \end{cases}$$

When (i) sib-pairs are unrelated, and (ii) the phenotype of a sib-pair is independent of phenotype and genotype data on other sib-pairs given the genotype of this sib-pair, then

$$L(\theta, \nu) = pr(\text{IBD data}|\text{phenotype data}) = \prod_{i=1}^{n} \rho_{0i}^{N_{0i}} \rho_{1i}^{N_{1i}} \rho_{2i}^{N_{2i}},$$

where $\rho_{ji} = pr(\text{Sib-pair shares } j \text{ IBD at } \mathcal{M}|\phi_{1i}, \phi_{2i})$, $j = 0, 1, 2.$

Sib-pair linkage score test

Test $\quad \mathrm{H}_0 : \theta = \dfrac{1}{2}$ **no linkage** *vs.* $\mathrm{H}_1 : 0 \leq \theta < \dfrac{1}{2}$ **linkage**.

The sib-pair linkage score test for qualitative and quantitative traits is based on the **second derivative** of the log-likelihood w.r.t. θ evaluated at $\theta = \frac{1}{2}$.

$$ST = \frac{\partial^2}{\partial\theta^2} \log L(\theta, \nu)\Big|_{\theta=\frac{1}{2}} = 16 \sum_{i=1}^{n} (\pi_{2i} - \pi_{0i})(N_{2i} - N_{0i}),$$

where for the ith sib-pair and $j = 0,\ 1,\ 2$

$$\pi_{ji} = pr\big(\text{Sib-pair shares } j \text{ IBD at } \mathcal{D} \mid \phi_{1i}, \phi_{2i}\big).$$

In practice

- **Incomplete IBD data**

Use a Hidden Markov Model to infer IBD status from marker genotype data ⇒ Inheritance distribution.

- **Genome scans**

Test for linkage at hundreds of markers simultaneously

⇒ adjust for multiple testing

- Ornstein-Uhlenbeck approximation;
- resampling methods.

General setting

For general pedigree types

- Partition the set of inheritance vectors into a smaller number of IBD configurations.
- Collapse the transition matrix $R(\theta)$ for inheritance vectors into the smaller transition matrix $T(\theta)$ for IBD configurations.
- Compute IBD probabilities given phenotypes:

$$\rho \;=\; \pi\, T(\theta).$$

- Derive the score test in θ by computing derivatives of $T(\theta)$.

General setting

- Define IBD configurations as **orbits of groups** acting on the set of inheritance vectors.

  E.g. k affected sibs: orbits of $S_k \times D_4$;

  unilineal relative pairs (*e.g.* cousins): as in Donnelly (1983).

- Count IBD configurations using **Pólya theory**.

- $R(\theta)$ for inheritance vectors: large and simple,

  $T(\theta)$ for IBD configurations: smaller and more complicated.

  $\Longrightarrow$ work with $R(\theta)$, then use properties of **quotient graphs** to deal with $T(\theta)$.

- Properties of the score test in θ are based on the **second largest eigenvalue** and corresponding eigenvector(s) of $T(\theta)$.

IBD configurations for k affected sibs

Label the paternal and maternal chromosomes containing the locus of interest by $(1,2)$ and $(3,4)$, respectively.
Let $a=(1,3)$, $b=(1,4)$, $c=(2,3)$, and $d=(2,4)$.

Inheritance vectors. Set $\mathcal{X}$ of mappings
$x:\{1,2,\ldots,k\}\rightarrow\{a,b,c,d\}$.

IBD configurations. Orbits of $S_k\times D_4$ acting on $\mathcal{X}$.

- S_k: permutations of the "genotypes" of the k sibs.
- D_4: permutations of $\{a,b,c,d\}$

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

Pólya theory: counting IBD configurations

The number of orbits of $S_k \times D_4$ acting on $\mathcal{X}$ is

$$\frac{1}{|D_4|}\sum_{\tau\in D_4} Z_{S_k}(m_1(\tau),\ldots,m_k(\tau)),$$

$$\begin{aligned}
Z_{S_k}(X_1,\ldots,X_k) &= \frac{1}{|S_k|}\sum_{\sigma\in S_k} X_1^{z_1(\sigma)}\ldots X_k^{z_k(\sigma)} \qquad \text{cycle index,}\\
m_i(\tau) &= \sum_{j|i} j z_j(\tau), \qquad i=1,\ldots,k,\\
z_j(\tau) &= \text{number of cycles of } \tau \text{ having length } j.
\end{aligned}$$

Pólya theory: counting IBD configurations

For k affected sibs, the number of IBD configurations is

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

E.g. Affected sib-trios, $m = 4$

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

Properties of transition matrix $T(\theta)$

- $T(\theta)$ satisfies the semi-group property

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

where $\theta_1 * \theta_2 = \theta_1(1-\theta_2) + \theta_2(1-\theta_1)$.

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

- $T(\theta) = \sum_h (1-2\theta)^{-\lambda_h/2} P_h$, where λ_h are real eigenvalues of Q and P_h are projection matrices.

Properties of transition matrix $T(\theta)$

Idea. Use graph theoretic arguments to derive eigenvalues of Q.

- $\mathcal{X}$ graph with vertex set the set of inheritance vectors and adjacency matrix $A = (a_x{}^{y})$

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

where $\Delta(x,y)$ is the Hamming distance.

- $Q = B - kI$, where B is the adjacency matrix of the quotient graph $\mathcal{X}/G \times H$ and $G \times H$ is the group defining the IBD configurations.

- The eigenvalues of Q belong to $\{-2i_{\binom{k}{i}} : i = 0, \ldots, k\}$.

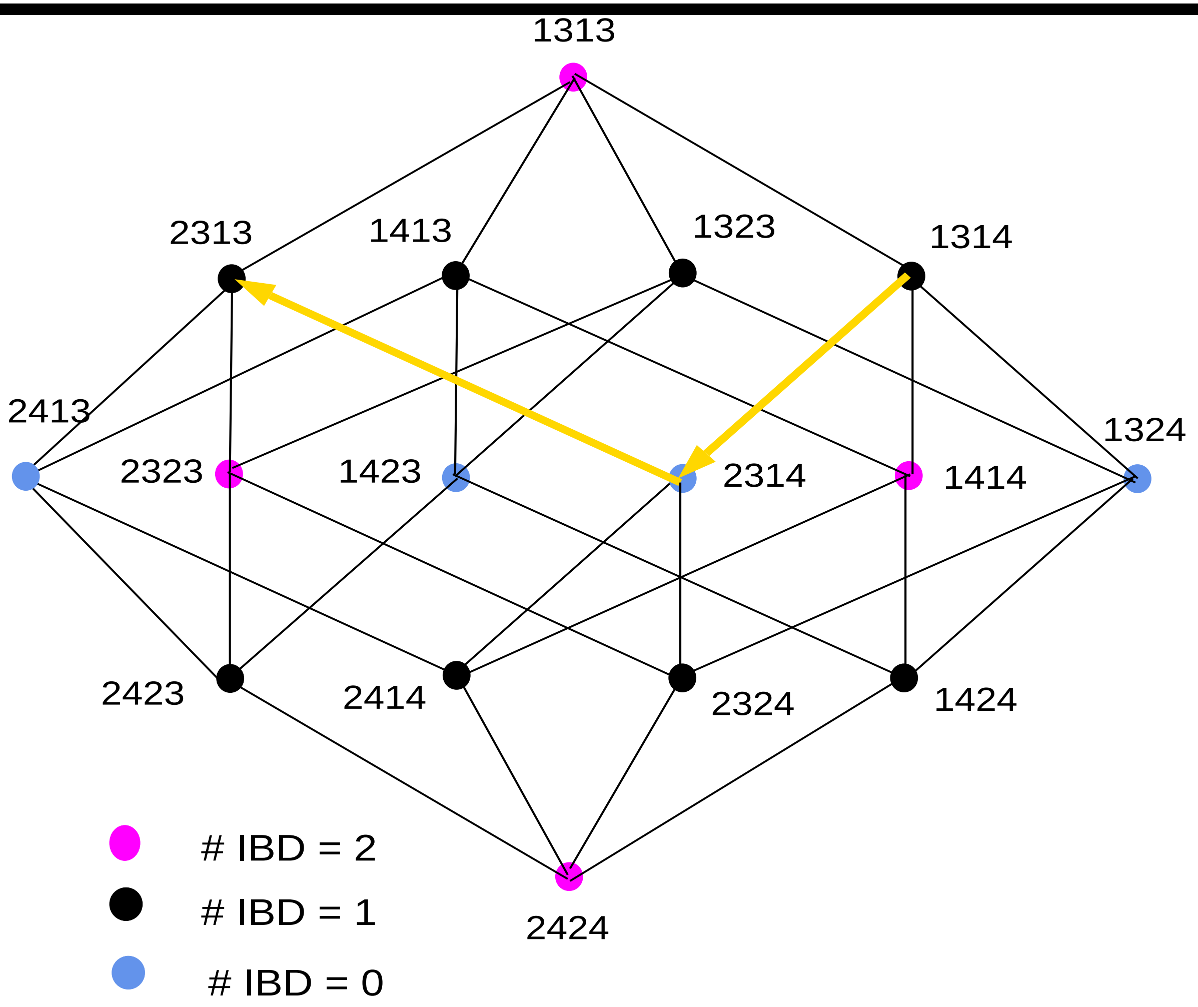


Figure 2: Sib-pair graph $\mathcal{X}$: 4-dimensional hypercube whose vertices correspond to the 16 possible inheritance vectors for a sib-pair and whose edges represent permissible transitions.

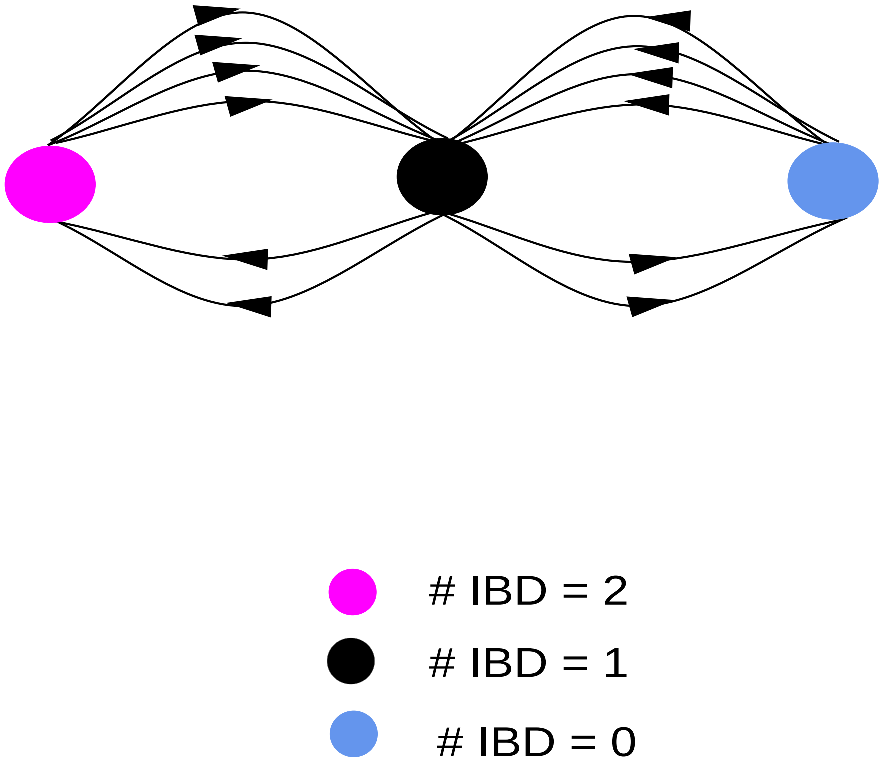


Figure 3: Sib-pair quotient graph $\mathcal{X}/(S_2 \times D_4)$.

Role of second largest eigenvalue of Q

- If the second largest eigenvalue of Q is $\lambda_2 = -2i$

$$T(\theta) = T\left(\frac{1}{2}\right) + (1-2\theta)^i P_2 + o((1-2\theta)^i).$$

- λ_2 and its multiplicity determine the first non-zero derivative of $T(\theta)$ at $\theta = \frac{1}{2}$ and its rank.

- The rate of convergence to $T(\frac{1}{2})$ as $\theta \to \frac{1}{2}$ is determined by λ_2.

Role of second largest eigenvalue of Q

- If $\lambda_2 = -2i$, the score test for a given pedigree type is based on the ith derivative of the log-likelihood at $\theta = \frac{1}{2}$.

- $$ST \propto \sum_i \sum_h \Big(\sum_j v_{jh}\pi_{ji}\Big)\Big(\sum_j v_{jh}N_{ji}\Big).$$

- If λ_2 has multiplicity one and phenotypes are constant across pedigrees, the score statistic is independent of the genetic model for the trait.

- Under the Poisson model for crossovers, the auto-correlation function for score statistics computed at loci t Morgans apart is $e^{\lambda_2 t}$.

Special case I - k affected sibs

The IBD configurations are the orbits of $S_k \times D_4$ acting on the set of 2^{2k} inheritance vectors.

$\lambda_2 = -4$, with multiplicity one.

The score test is based on the second derivative of the log-likelihood and is independent of the genetic model for the trait.

The score statistic is the widely used non-parametric statistic S_{pairs}.

Special case II - Unilineal relative pairs

The IBD configurations are defined as in Donnelly (1983).

$\lambda_2 = -4$ for half-sib, avuncular, and cousin pairs.

$\lambda_2 = -2$ for pairs of the grand-parent/grand-child type, and more distant relatives of the half-sib, avuncular, and cousin types.

Ongoing work

- Apply linkage score test to IBD data on endometriosis.
- Freely available software package for the linkage score test.
- Linkage score test for survival data.
- Combining IBD data from relative pairs with different λ_2.