Nonparametric linkage analysis in humans

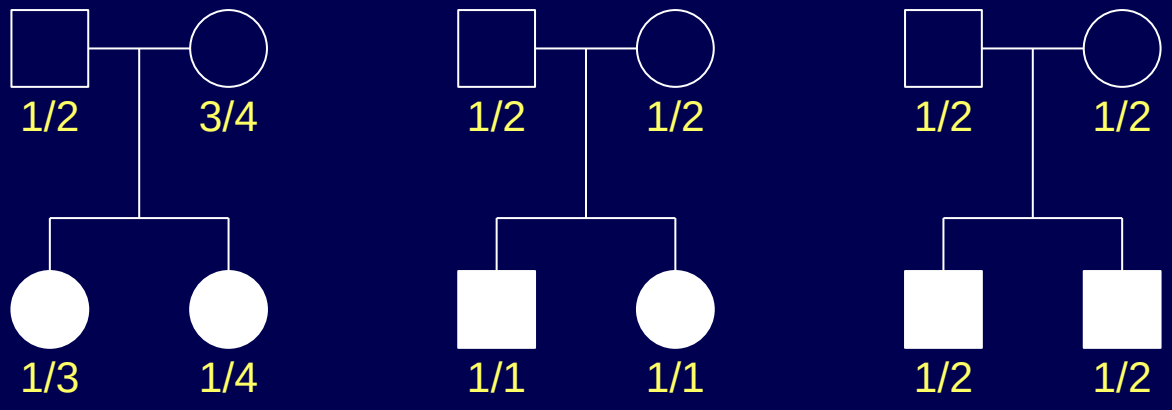
Karl W Broman
Biostatistics & Medical Informatics
University of Wisconsin – Madison

www.biostat.wisc.edu/~kbroman

Allele-sharing methods

The idea: Affected relatives will be more likely to share DNA in the region of a gene that contributes to the disease.

"Identical by descent": Two alleles are IBD if they are copies of a single ancestral allele.



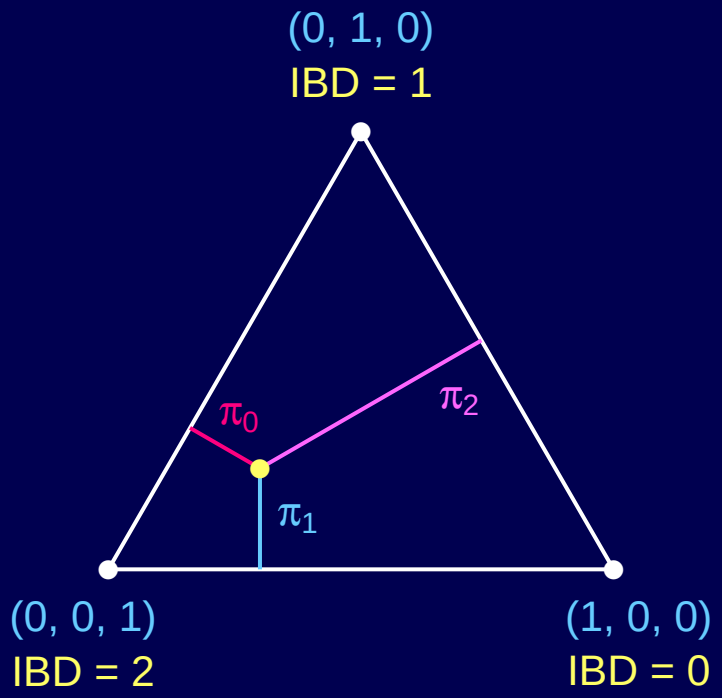
A priori, siblings share 0, 1, 2 alleles IBD with probability 1/4, 1/2, 1/4, respectively.

Example

- Single diallelic gene, disease allele frequency = 0.1
- Penetrances $f_0 = 0.01$, $f_1 = 0.1$, $f_2 = 0.5$
- At a position with r.f. 5% from gene

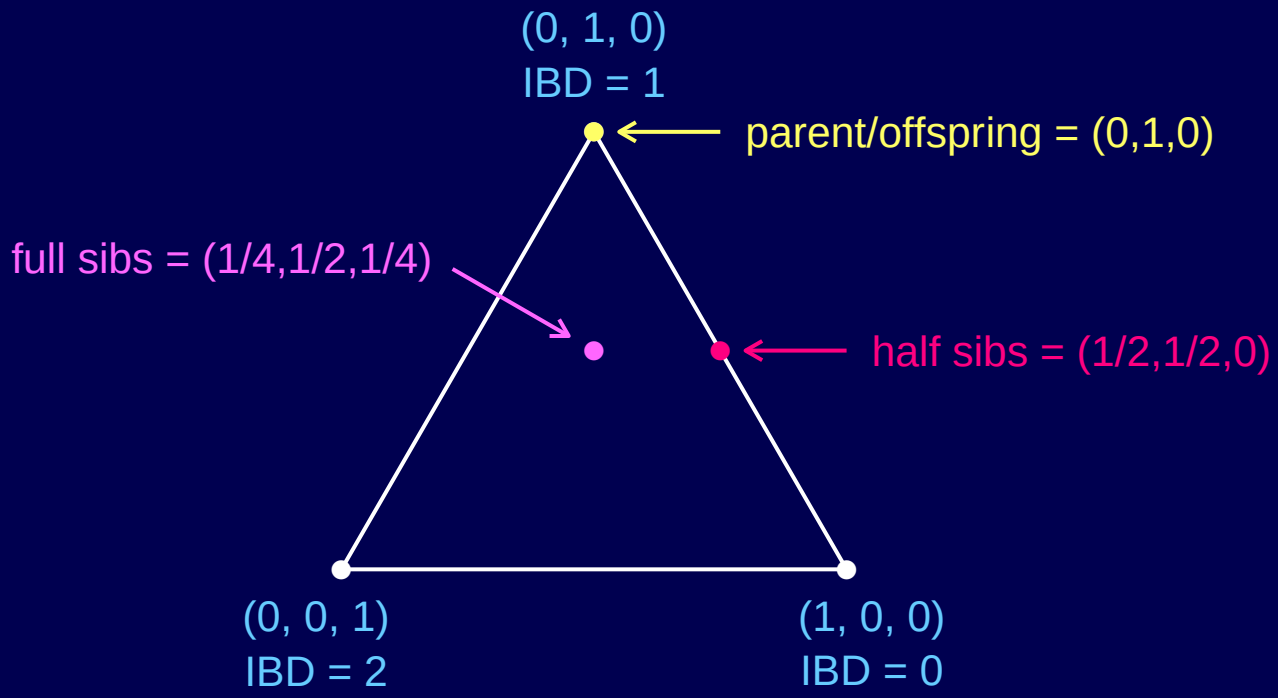
Siblings' phenotypes	Pr(IBD = v | aff. status)			ave. no. alleles IBD
	0	1	2	
Both affected	0.063	0.495	0.442	1.38
Neither affected	0.248	0.500	0.252	1.00
One aff, one not	0.368	0.503	0.128	0.76

Holmans triangle

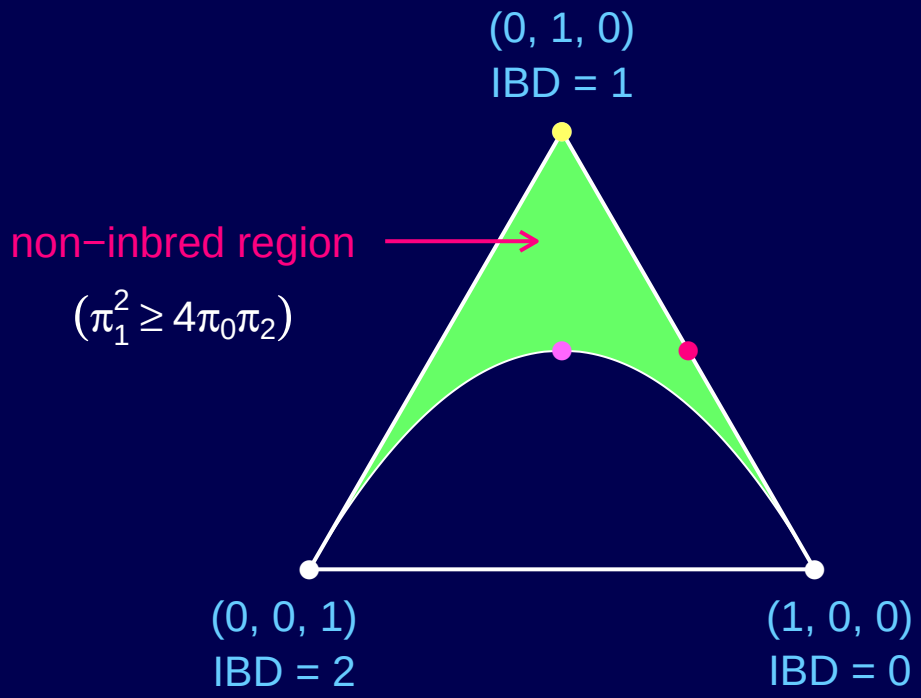


Holmans, Am J Hum Genet, 52:362–374, 1993

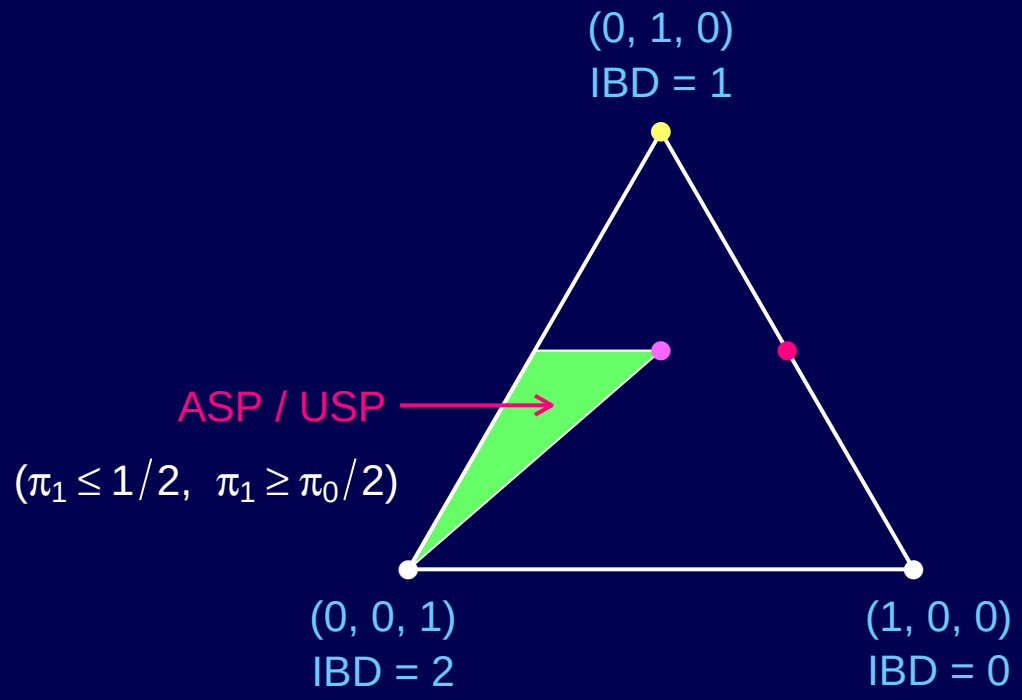
Holmans triangle



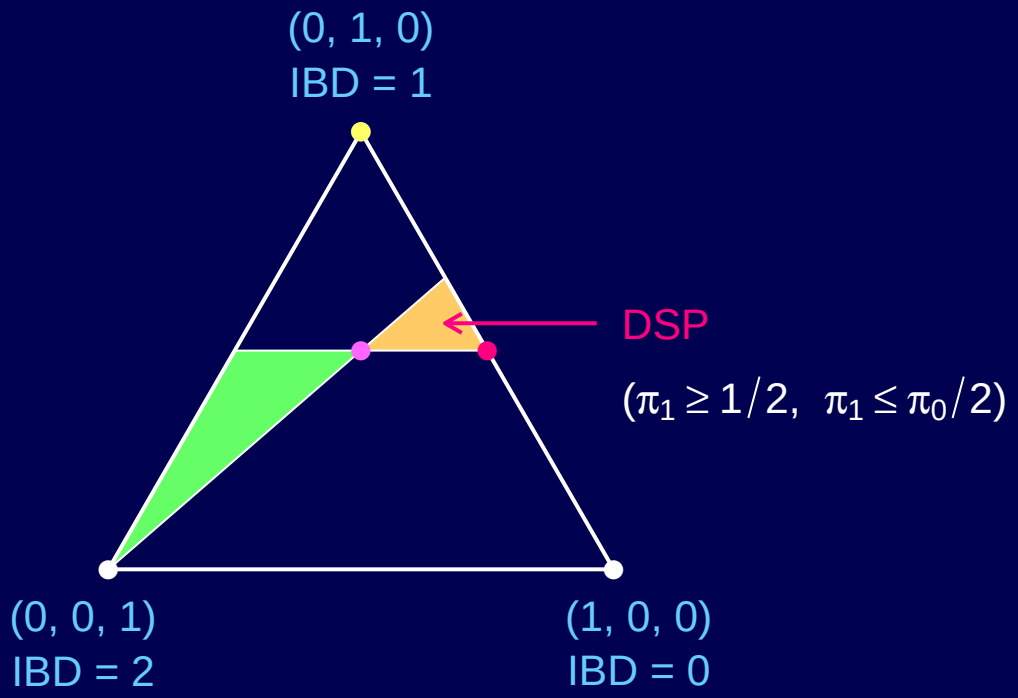
Holmans triangle



Holmans triangle



Holmans triangle



Affected sib pairs

The complete data case:

- n affected sib pairs
- IBD sharing, at a particular position, is known exactly
- n_j = no. sib pairs sharing j alleles IBD

Compare $(n_0/n, n_1/n, n_2/n)$ to $(1/4, 1/2, 1/4)$

Example: 100 affected sib pairs with $(n_0, n_1, n_2) = (15, 38, 47)$

Tests

1. Mean test

Let $S = n_1 / n + 2\, n_2 / n$

Under H_0: $\pi = (1/4, 1/2, 1/4)$,

$E(S \mid H_0) = 1$, $\mathrm{var}(S \mid H_0) = \ldots = 1/(2n)$

Let $Z = \frac{S - E(S \mid H_0)}{SD(S \mid H_0)} = \sqrt{2n}(S - 1)$.

Take $LOD = Z^2/(2 \ln 10)$

Example: $S = 1.32$, $Z = 4.53$, $LOD = 4.45$, P-value $= 6 \times 10^{-6}$

2. Proportion test

$S' = n_2 / n$ $\qquad$ $E(S' \mid H_0) = 1/4$, $\mathrm{var}(S \mid H_0) = 3/(16\, n)$

$Z' = \cdots = \sqrt{\frac{n}{3}}\left(4S' - 1\right)$

Example: $S' = 0.47$, $Z' = 5.08$, $LOD' = 5.61$, P-value $= 4 \times 10^{-7}$

Tests

3. χ^2 test

$$X^2 = \sum_j \frac{(n_j - n\ \pi_{0j})^2}{n\ \pi_{0j}}$$

where π_{0j} = null IBD probabilities

Example: X^2 = 26.2, LOD = 5.70, P-value = 2×10^{-6}

4. Likelihood ratio test

$$l(\pi) = n_0 \ln(\pi_0) + n_1 \ln(\pi_1) + n_2 \ln(\pi_2)$$

$$\text{LRT} = 2 \left\{ \sum_j n_j \ln(n_j/n) - \sum_j n_j \ln(\pi_{0j}) \right\}$$

Example: LRT = 23.2, LOD = 5.03, P-value = 9×10^{-6}

Note: We might restrict the alternative to the “possible triangle”

Incomplete data

- We seldom know the alleles shared IBD exactly

- We can calculate, for sib pair i,

  $$p_{ij} = \Pr(\text{sib pair } i \text{ has IBD}=j \mid \text{marker data})$$

- In the Mean test, use $\sum_i p_{ij}$ in place of n_j

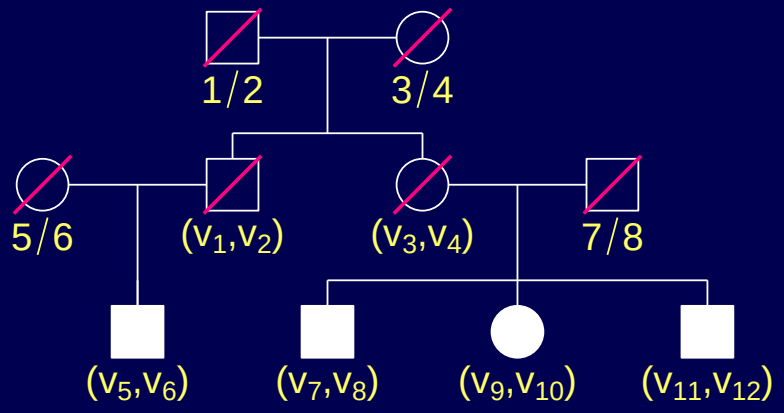
  Problem: The denominator, $\sqrt{n/2}$, is correct for perfect IBD information but is too large in the case of incomplete data.

- Alternatives: computer simulations; likelihood methods

  (Kong and Cox, Am J Hum Genet, 61:1179–1188, 1997)

General pedigrees

(e.g., larger sibships)



Inheritance vector

$$v = \{(v_1, v_2), \ldots, (v_{11}, v_{12})\}$$

Indicates which parental alleles were transmitted

Score function, S(v), measures allele sharing

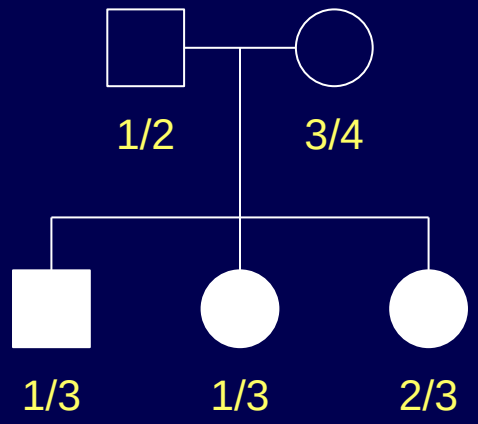
$$S_{pairs}(v) = \sum_{\text{aff. rel. pairs}} \#\{\text{alleles shared IBD}\}$$

$$S_{all}(v) = \frac{1}{2^a} \sum_{h} \left\{ \prod_{i=1}^{2f} b_i(h)! \right\}$$

a = # aff. relatives
h = vector with one random allele selected from each affected relative
$b_i(h)$ = # times ith founder allele appears in h

Example

Affected sib trio



$$S_{pairs} = 2 + 1 + 1 = 4$$

$$S_{all} = \frac{1}{2^3}\{(2 \cdot 1 \cdot 1 \cdot 1) + (2 \cdot 1 \cdot 1 \cdot 1) + (1 \cdot 1 \cdot 1 \cdot 1) + (2 \cdot 1 \cdot 1 \cdot 1) + 1 + 2 + 2 + 6\}$$

$$= 2.25$$

For S_{all}, the h vectors are:

1,1,2	1,1,3	1,3,2	1,3,3
3,1,2	3,1,3	3,3,2	3,3,3

Normalized score

$$Z(v) = \{S(v) - \mu_0\} / \sigma_0$$

where $\mu_0 = E[S(v) \mid H_0]$, $\sigma_0 = SD[S(v) \mid H_0]$

$H_0 \equiv$ no linkage (v's equally likely)

For example, for 3 affected sibs:

S_{pairs} : $\mu_0 = 3, \sigma_0 = \sqrt{3/2} \approx 1.22$

S_{all} : $\mu_0 = 15/8 = 1.875, \sigma_0 = (3/8)\ \sqrt{3/2} \approx 0.459$

For our previous example:

$S_{pairs} = 4 \longrightarrow Z \approx 0.817$

$S_{all} = 9/4 \longrightarrow Z \approx 0.817$

(For 3 aff. sibs, $S_{pairs} \equiv S_{all}$)

More generally they differ, and S_{all} is better for additive/dominant model, S_{pairs} better for recessive model.

Combining families

- Calculate normalized score for each family, $\mathsf{Z}_i = (\mathsf{S}_i - \mu_{0i})/\sigma_{0i}$

- Combine families with weights $\gamma_i \geq 0$

$$\mathsf{Z} = \sum_i \gamma_i \mathsf{Z}_i / \sqrt{\sum_i \gamma_i^2}$$

- Choices of γ:
  - $\equiv$ 1 for all families
  - $\equiv$ no. sib pairs
  - $\equiv \sigma_{0i}$ (i.e., combine S's then normalize)

Incomplete data

In place of S, use

$$\bar{S} = \sum_v S(v) \Pr(v \mid \text{marker data})$$

$\longrightarrow$ Really need the σ_0's to take account of the uncertainty in the v's.