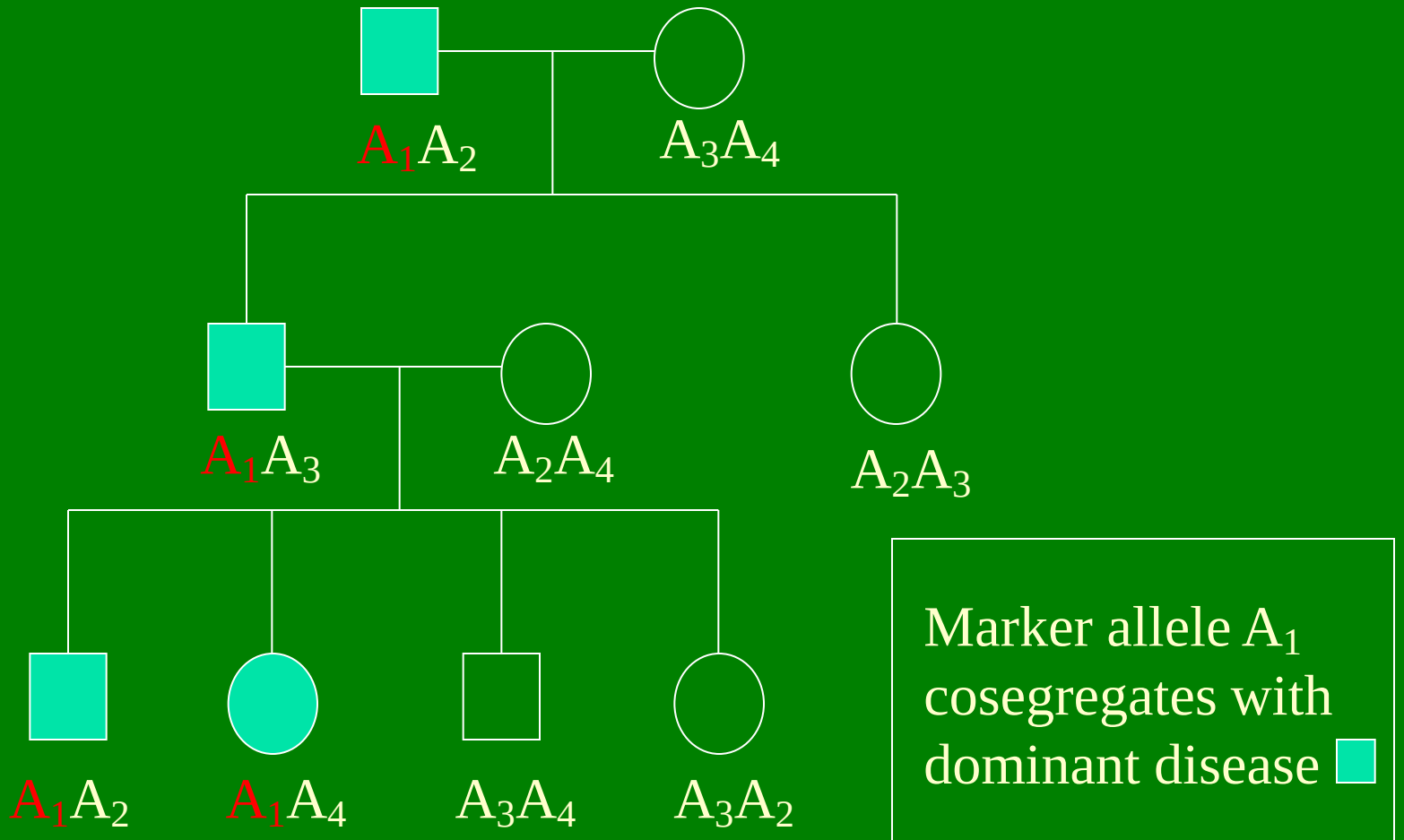
QTL linkage analysis in nuclear families

Peter M. Visscher
Queensland Institute of Medical Research
Brisbane, Australia
peter.visscher@qimr.edu.au

Overview

- QTL mapping by linear regression
  - Haseman-Elston regression
  - Using similarity scores
- Variance components & maximum likelihood
- IBD estimation from marker data
- Statistical power

Linkage = Co-segregation



All QTL mapping methods are essentially 2-stage procedures

1. Genetic markers give information on IBD sharing between relatives [genotypes]
2. Association between phenotypes and genotypes gives information on QTL location and effect [linkage]

- Need informative mapping population

Population	*Features*	*Example Species*
Inbred lines		
Backcross (BC)	Simplest design; powerful if dominance in 'right' direction	mice, plants
F_2	Estimation of additive and dominance effects; more powerful than BC for additive effects	mice, rats
Advanced intercross line (AIL)	As for F_2 but with increased resolution of map location	mice
Recombinant inbred lines (RIL)	F_1 followed by inbreeding; homozygous comparisons only; powerful for additive effects; less environmental noise	mice, plants
Congenic lines (= Nearly isogenic lines)	Backcrossing followed by inbreeding; homozygous comparisons only after inbreeding. Lines contain ~1% of donor genome	mice, rats, plants
Double haploid lines (DHL)	Instant homozygosity through doubling of F_1 gametes; homozygous comparisons only; powerful for additive effects and QTLxE interactions	plants
$F_{2:3}$	Inbred progeny of F_2; increased precision through progeny means	plants
Structured outbred populations		
BC / F_2 / AIL	As for inbred lines; mapping variation between lines	livestock, outbreeding trees/plants
Large fullsib families	Estimating contrasts between parental alleles. Allows for dominance estimation.	trees, fish, poultry
Halfsib families	Estimating contrasts between common parent alleles	cattle, pigs, poultry, trees
Nuclear families, including sibpairs	Detection of variance explained by markers	humans, livestock
Unstructured outbred populations		
Complex pedigrees	Detection of variance explained by markers	humans, livestock

Mapping populations

Human sibpairs are not the only design!

Model: QTL as a random effect

$y_i = \mu + Q_i + A_i + E_i$

Q_i = QTL genotype contribution for chrom. segment

A_i = Contribution from rest of genome

$var(y) = \sigma_q^2 + \sigma_a^2 + \sigma_e^2$

Genetic covariance between relatives

$$\mathrm{cov}(y_i, y_j) = \pi_{ij}\,\sigma_q^2 + a_{ij}\,\sigma_a^2$$

a_{ij} = average prop. of alleles shared in the genome (twice kinship coefficient)

π_{ij} = proportion of alleles IBD at QTL (0, ½ or 1)

$E(\pi_{ij}) = a_{ij}$

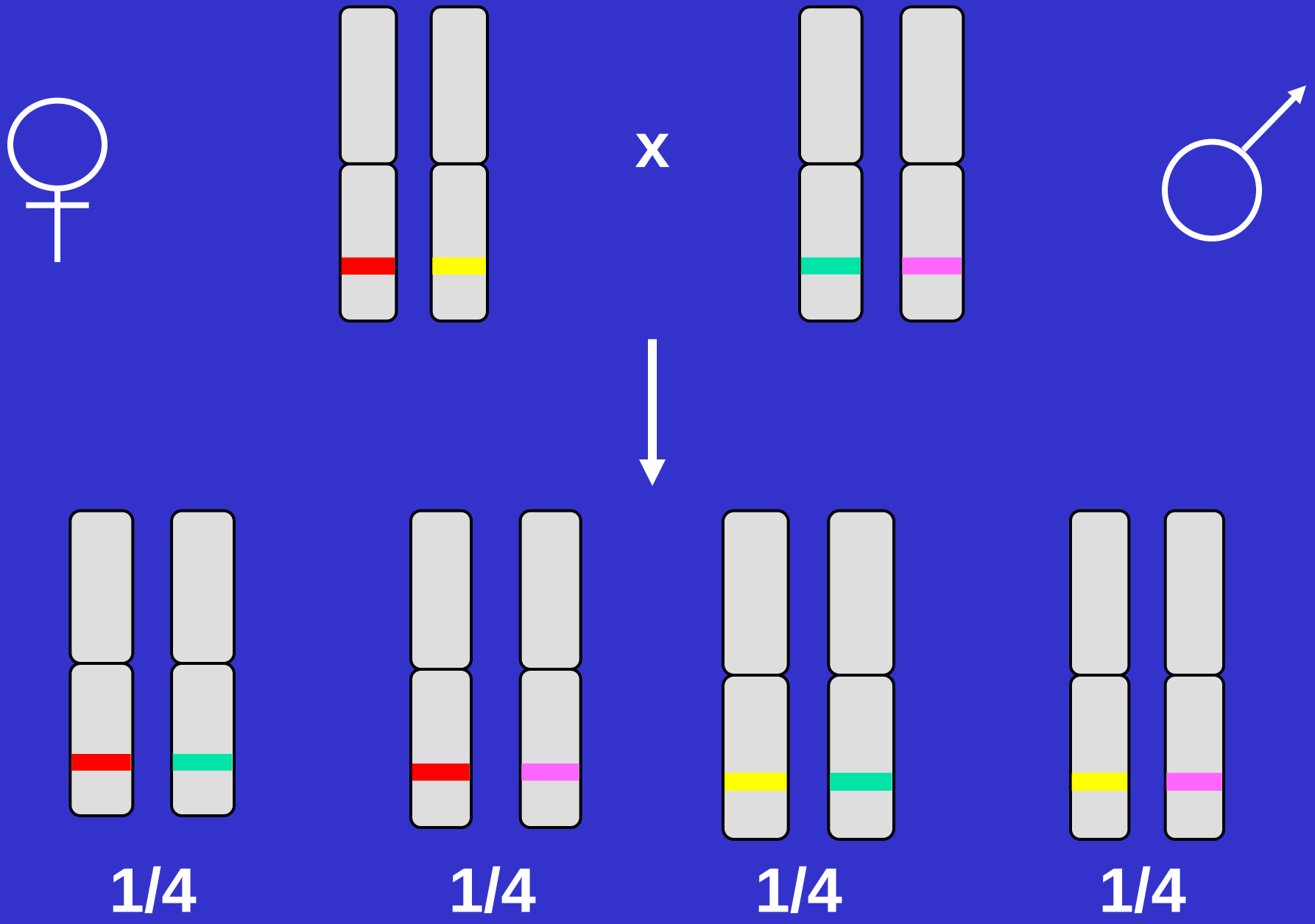
$\hat{\pi}$

$\hat{\pi}_{ij}$ = Pr(2 alleles IBD) + ½Pr(1 allele IBD)
|given data

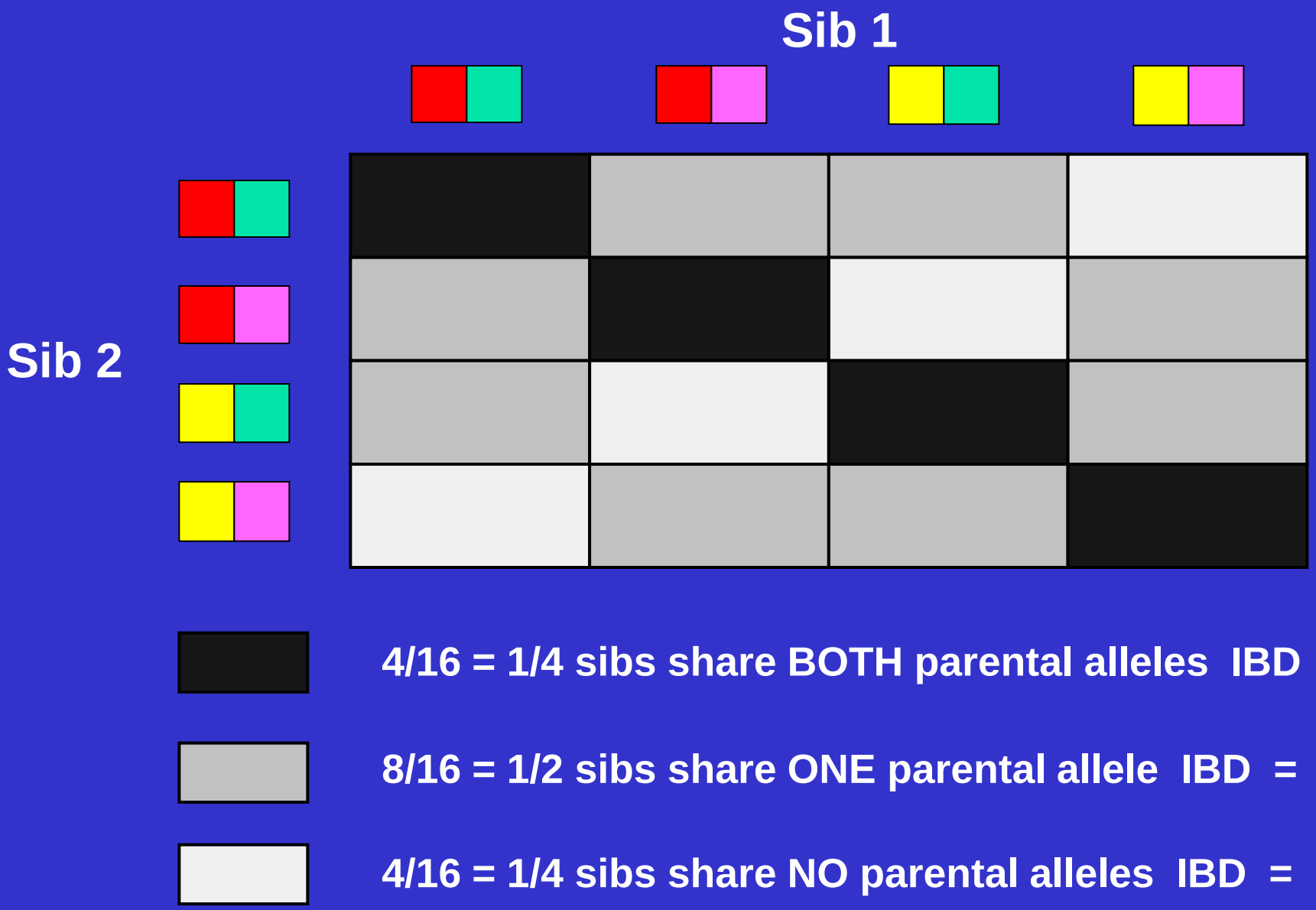
= proportion of alleles IBD in non-inbred pedigree

Estimate $\hat{\pi}_{ij}$ with genetic markers

Random segregation and identity-by-descent in sibpairs



IDENTITY BY DESCENT



Several notations

IBD	Probability	Actual
IBD0	k_0	0 or 1
IBD1	k_1	0 or 1
IBD2	k_2	0 or 1
	$\Sigma=1$	$\Sigma=1$

Realisations

k_0	k_1	k_2
1	0	0
0	1	0
0	0	1

$\pi_a = ½k_1 + k_2 = R = 2\theta$

$\pi_d = k_2 = \Delta_{xy}$

[e.g., LW Chapter 7; Weir and Hill 2011, Genetics Research]

Sibpairs & fully informative marker

# Alleles IBD	π	Pr.
0	0	¼
1	½	½
2	1	¼

$E(\pi) = \Sigma\, \pi Pr(\pi) = ½$

$E(\pi^2) = \Sigma\, \pi^2 Pr(\pi) = {}^3/_8$

$var(\pi) = E(\pi^2) - E(\pi)^2 = {}^1/_8$

CV = $0.5\sqrt{2} = 70\%$

QTL mapping by linear regression

- Simple
  - standard stats package
- Robust
  - non-normal traits
- Powerful
- Computationally fast
  - permutations
  - bootstrapping

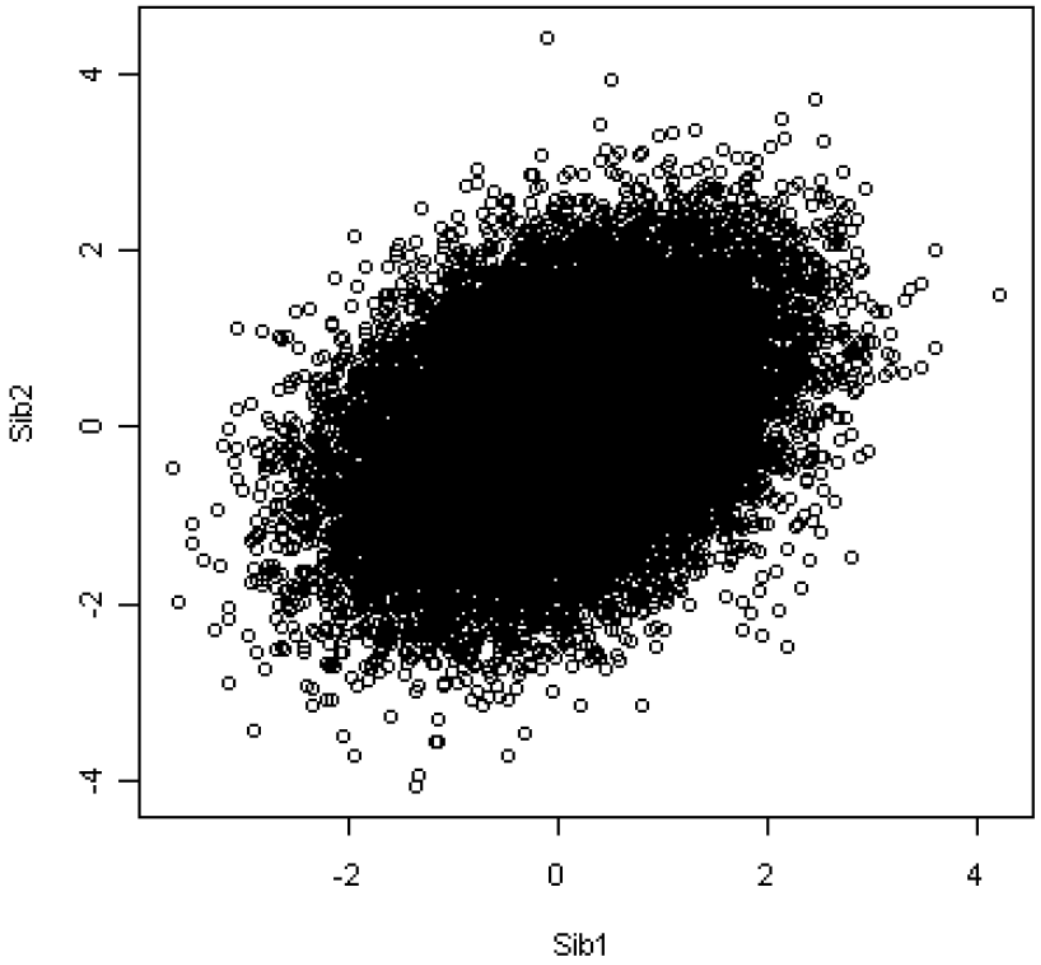
Haseman-Elston (1972)

"The more alleles pairs of relatives share at a QTL, the greater their phenotypic similarity"

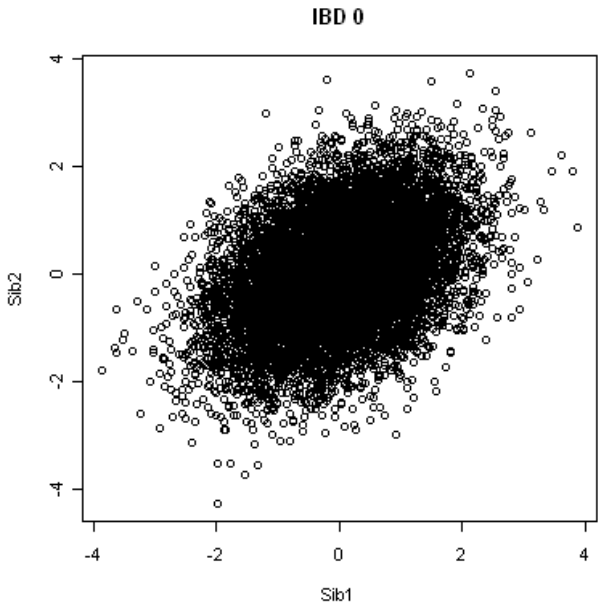
or

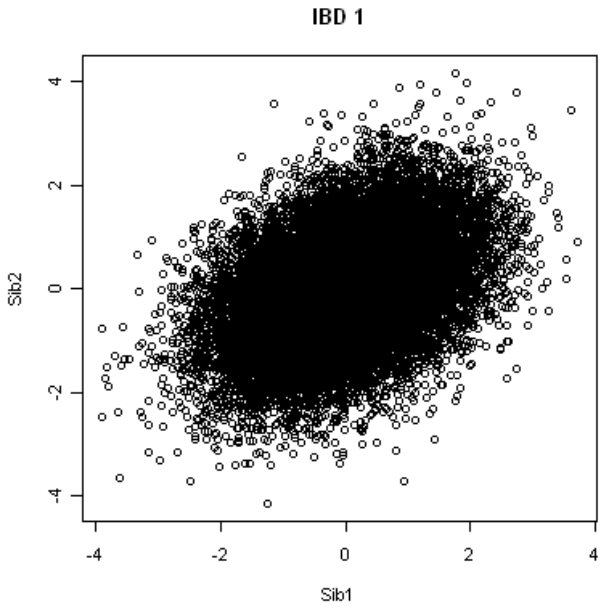
"The more alleles they share IBD, the smaller the difference in their phenotype"

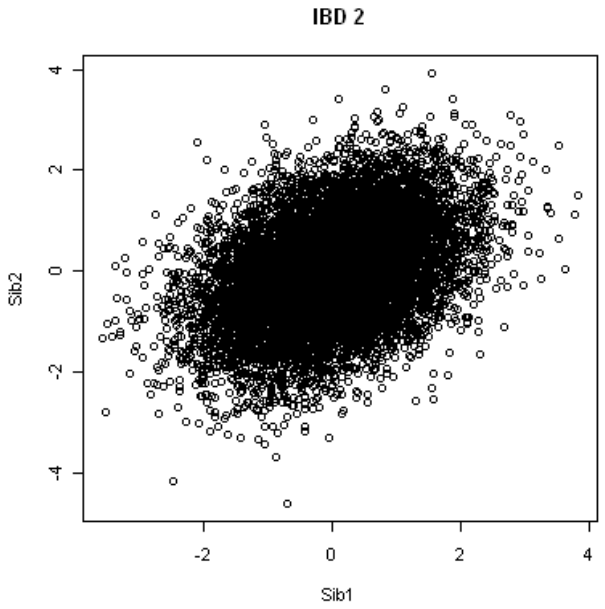
Population sib-pair trait distribution



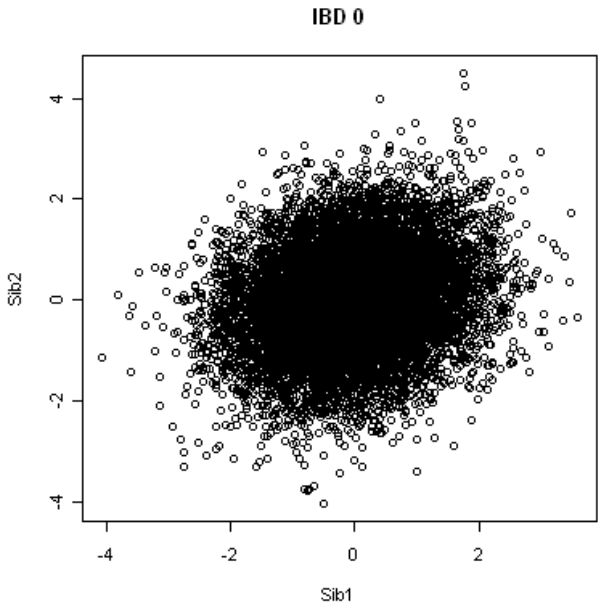
No linkage

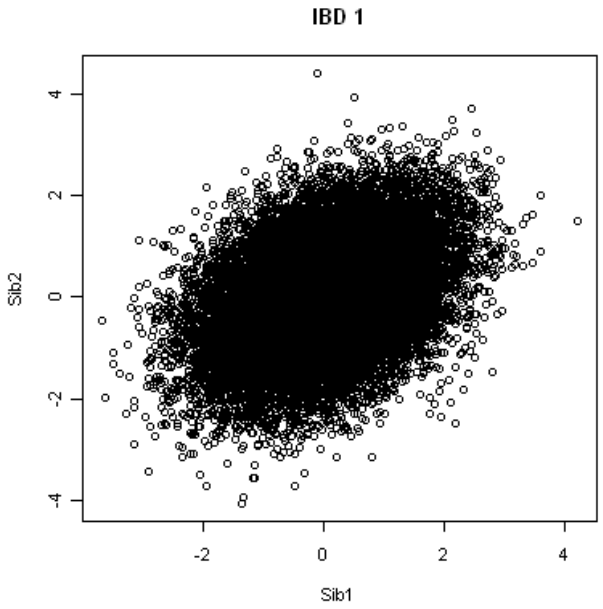

IBD 0
Sib2
Sib1
-4
-2
0
2
4

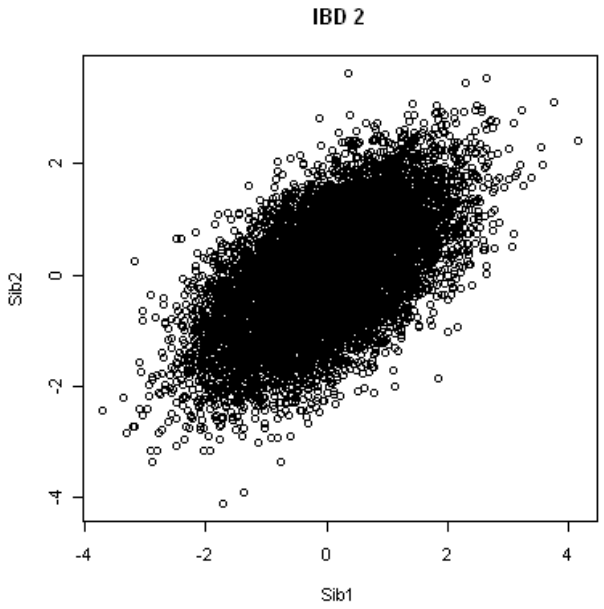

IBD 1
Sib2
Sib1
-4
-2
0
2
4


IBD 2
Sib2
Sib1
-4
-2
0
2
4

Under linkage







Sib pair design to map QTL

- Multiple ‘families’ of two (or more) sibs
- Phenotypes on sibs
- Marker genotypes on sibs (& parents)

- Correlate phenotypes and genotypes of sibs

Data structure is simple

Pair	Phenotypes		Prop. alleles IBD
1	y_{11}	y_{12}	π_1
2	y_{21}	y_{22}	π_2
.....			
n	y_{n1}	y_{n2}	π_n

π = 0, ½ or 1 for fully informative markers

Properties of squared differences

$E(Y_1 - Y_2)^2 = var(Y_1 - Y_2) + (E(Y_1 - Y_2))^2$

$var(Y_1 - Y_2) = var(Y_1) + var(Y_2) - 2cov(Y_1, Y_2)$

If $E(Y_i) = E(Y_j)$ and $var(Y_1) = var(Y_2)$, then

$E(Y_1 - Y_2)^2 = 2(1 - \rho)var(Y)$ and

$Var(Y_1 - Y_2)^2 = 8(1 - \rho)^2 var(Y)^2$

If Y~N[E(Y),var(Y)]

Haseman-Elston method

- Phenotypes on a relative pair:

$$Y = (y_1 - y_2)^2$$

$$E(Y) = E[(Q_1 - Q_2)^2 + (A_1 - A_2)^2 + (E_1 - E_2)^2]$$

$$= E[(Q_1 - Q_2)^2] + \{2(1-a_{12})\sigma_a^2 + 2\sigma_e^2\}$$

$$= 2[\sigma_q^2 - cov(Q_1,Q_2)] + \{\sigma_\varepsilon^2\}$$

$$= (2\sigma_q^2 + \sigma_\varepsilon^2) - 2\pi_{12}\sigma_q^2$$

π_{12} = proportion of alleles IBD at QTL for the pair of relatives

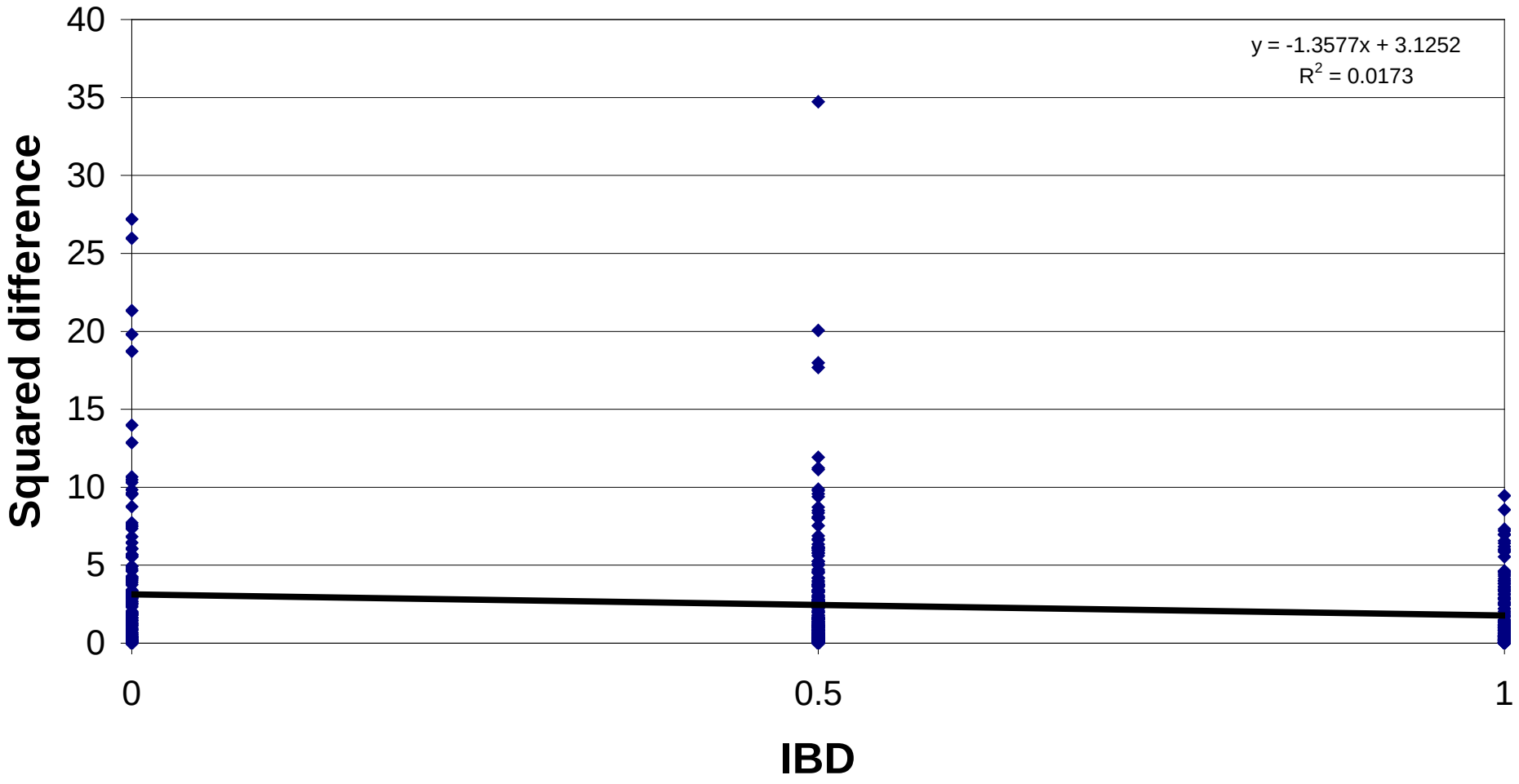
Conditional expectation

$E(Y_j | \pi_j) = (2\sigma_q^2 + \sigma_\varepsilon^2) - (2\sigma_q^2)\pi_j$

- negative slope of Y on π if $\sigma_q^2 > 0$
- estimate π_j from marker data [$\hat{\pi}_j$]
- use simple linear regression to detect QTL:

$$E(Y_j | \hat{\pi}_j) = \alpha + \beta\hat{\pi}_j$$

Haseman-Elston regression



A significant negative slope indicates linkage to a QTL

Single fully informative marker

$\beta = -2(1 - 2r)^2 \sigma_q^2$

$\alpha = 2[1 - 2(1-r)r] \sigma_q^2 + \sigma_\varepsilon^2$

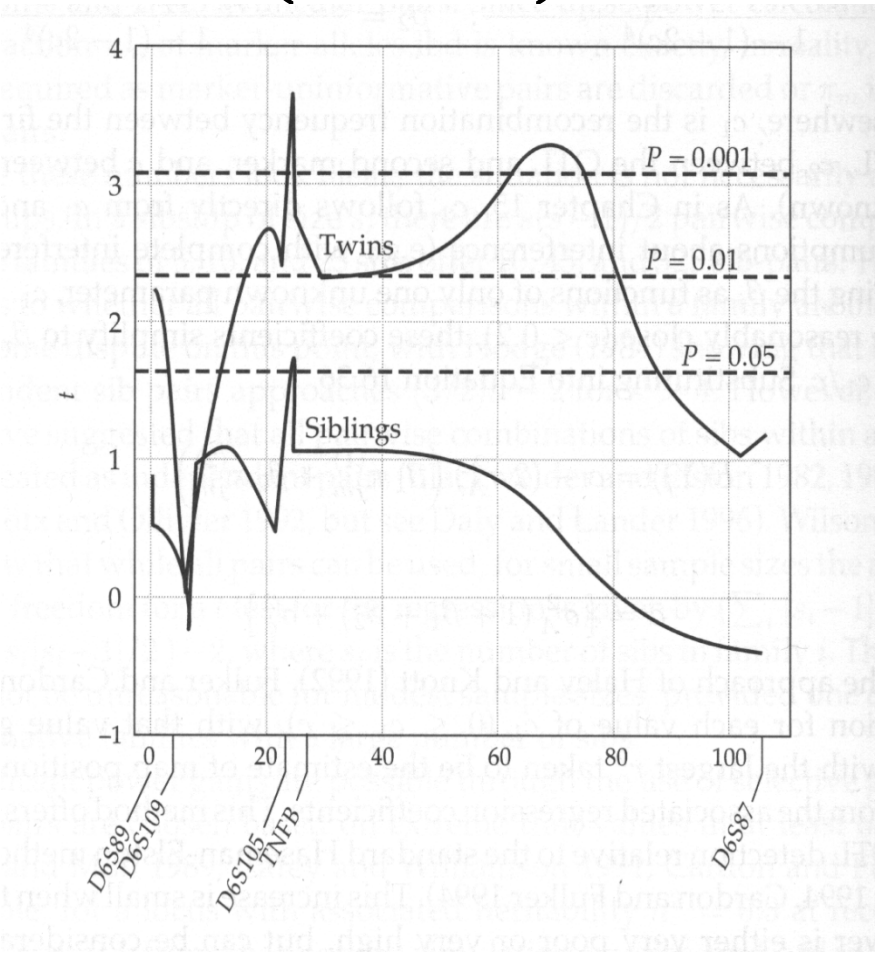
r = recombination fraction between marker & QTL

- Disadvantage of method
  - not powerful
  - confounding between QTL location and effect

- => Use multipoint estimates of π

Example from Cardon et al. (1994)

Be sceptical if you see plots like this!



[Lynch & Walsh, page 520]

Squared difference (again)

$E(Y_1 - Y_2)^2 = 2(1-\rho)var(Y)$

$Var(Y_1 - Y_2)^2 = 8(1 - \rho)^2\ var(Y)^2$

$SD(Y_1 - Y_2)^2 = (2\sqrt{2})(1 - \rho)\ var(Y)$

NB: Linkage Practical

Distribution of test statistic under the null hypothesis of no linkage

Method	Null	Alternative	Action if wrong alternative
HE	$\beta=0$	$\beta<0$	Set b=0
VC	$\sigma_{QTL}^2=0$	$\sigma_{QTL}^2>0$	Set $\sigma_{QTL}^2=0$

- The statistical tests are one-sided.
  - If Null is true then expect b>0 or $\sigma_{QTL}^2<0$ with a probability of ½
  - The test statistic follows a mixture of distributions with mixing proportion of ½
  - LRT ~ 0 with prob= ½ and ~$\chi^2_{(1)}$ with prob= ½
  - For data analysis: take p-values from $\chi^2_{(1)}$ and divide by two

QTL mapping using similarity scores

- Sometimes we only know (dis)similarity between pairs of relatives
  - mosquito bites
  - eye colour
- H-E principle still applies: more similar implies more alleles shared IBD

Models

(d= dissimilarity score; s= similarity score)

$d = \mu_d + \beta_d \pi$

$H_1: \beta_d < 0$

$s = \mu_s + \beta_s \pi$

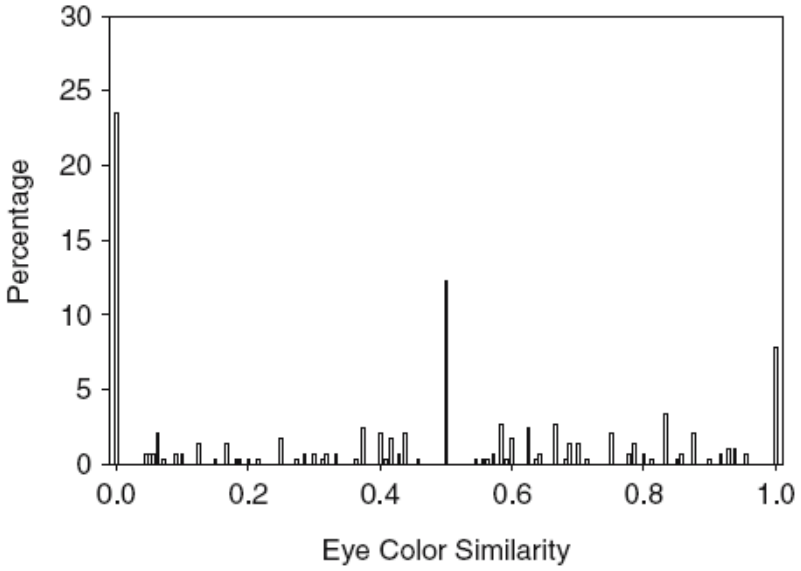
$H_1: \beta_s > 0$

Behavior Genetics (© 2005)
DOI: 10.1007/s10519-005-9007-x

Replicated Linkage for Eye Color on 15q Using Comparative Ratings of Sibling Pairs

Danielle Posthuma,[1,4] Peter M. Visscher,[2] Gonneke Willemsen,[1] Gu Zhu,[2] Nicholas G. Martin,[2] P. Eline Slagboom,[3] Eco J. C. de Geus,[1] and Dorret I. Boomsma[1]

sions. Eye color similarity was rated on a three point scale ("*not at all alike*"—"*somewhat alike*"—"*completely alike*"). The probability that twins were alike for eye color (eye color similarity, *s*) was calculated from the response pattern on all questionnaires and all informants, by summing over the product of the three possible answer categories (where "*not at all alike*" was coded 0, "*somewhat alike*" was coded 0.5, and "*completely alike*" was coded 1) and their



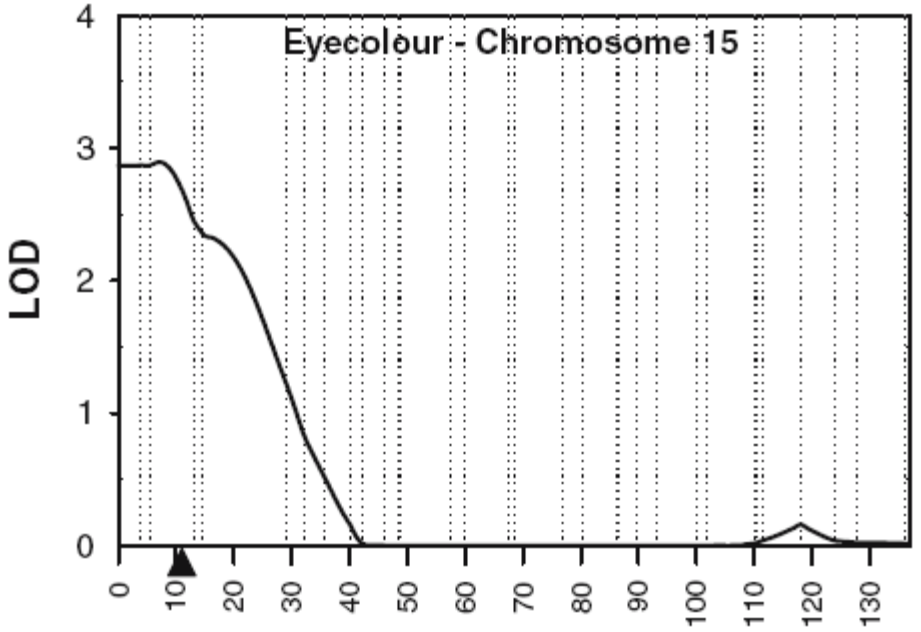


Fig. 3. Chromosome 15 region of significant linkage for eye color. The triangle marks the location of the linkage peak from Zhu *et al.* (2004) on 15q. Dotted lines represent the positions of the markers. The *x*-axis is in centiMorgan.

Linkage by Variance Component Analysis

- Why?
  - More powerful (in theory)
  - Logical extension of analysis of resemblance between relatives
  - Applicable to general (& large) pedigrees

Variance-Covariance Matrix for a pair of relatives

$$\Sigma = \begin{bmatrix} V(y_1) & Cov(y_1, y_2) \\ Cov(y_1, y_2) & V(y_2) \end{bmatrix}$$

Model must describe not only variance of each observation but also covariance for pairs of observations

Variance-Covariance Matrix

$$\Sigma_{jk} = \begin{cases} \sigma_q^2 + \sigma_a^2 + \sigma_c^2 + \sigma_e^2 & if \quad j = k \\ \hat{\pi}_{jk}\sigma_q^2 + a_{jk}\sigma_a^2 + \sigma_c^2 & if \quad j \neq k \end{cases}$$

a_{jk} is twice the kinship coefficient,
= i.e. twice the probability that two genes sampled at random from a pair of individuals are identical. (1 for MZ twins and 0.5 for fullsibs)

Variance-Covariance Matrix

$$\Sigma_{jk} = \begin{cases} \sigma_q^2 + \sigma_a^2 + \sigma_c^2 + \sigma_e^2 & if \quad j = k \\ \hat{\pi}\sigma_q^2 + \rho\sigma_a^2 + \sigma_c^2 & if \quad j \neq k \end{cases}$$

σ^2 = variation due to:

q = QTL;

a = polygenic;

e = individual-specific

c = shared environment

Bivariate density function

- Normal density function

$$L(y) = \frac{1}{\sqrt{2\pi}} \sigma^{-1} e^{-\frac{1}{2}(y-\mu)^2/\sigma^2}$$

- Bivariate normal density function

$$L(\mathbf{y}) = \frac{1}{2\pi} |\Sigma|^{-1/2} \; e^{-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})'\Sigma^{-1}(\mathbf{y}-\boldsymbol{\mu})}$$

Alternate hypothesis of linkage for sibpairs (Likelihood function):

$$L_1 = \prod_i (2\pi)^{-1} |\Sigma_i|^{-1/2} e^{-1/2(\mathbf{y_i}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{y_i}-\boldsymbol{\mu})}$$

$$\Sigma = \begin{bmatrix} \sigma_q^2 + \sigma_a^2 + \sigma_c^2 + \sigma_e^2 & \hat{\pi}_i \sigma_q^2 + a_\iota \sigma_a^2 + \sigma_c^2 \\ \hat{\pi}_i \sigma_q^2 + a_\iota \sigma_a^2 + \sigma_c^2 & \sigma_q^2 + \sigma_a^2 + \sigma_c^2 + \sigma_e^2 \end{bmatrix}$$

Note three uses of 'pi'!

Null hypothesis:

$$L_0 = \prod_i (2\pi)^{-1} |\Sigma_i|^{-1/2} e^{-1/2(\mathbf{y_i}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{y_i}-\boldsymbol{\mu})}$$

$$\Sigma = \begin{bmatrix} \sigma_a^2 + \sigma_c^2 + \sigma_e^2 & a_\iota \sigma_a^2 + \sigma_c^2 \\ a_\iota \sigma_a^2 + \sigma_c^2 & \sigma_a^2 + \sigma_c^2 + \sigma_e^2 \end{bmatrix}$$

Test statistic ML

$$LRT = 2\ln(ML_{full}) - 2\ln(ML_{reduced})$$

$$H_0: LRT \sim \tfrac{1}{2}\chi^2(1) + \tfrac{1}{2}(0)$$

IBD calculating algorithms

- Elston-Stewart algorithm

  Handles large pedigrees, but small nr of loci, exact IBD distributions (Elston and Stewart, 1971)

- Lander-Green algorithm

  Handles small pedigrees, but large nr of loci, exact IBD distributions (Lander and Green, 1987)

- MCMC methods

  Calculates approximate IBD distributions (Heath, 1997)

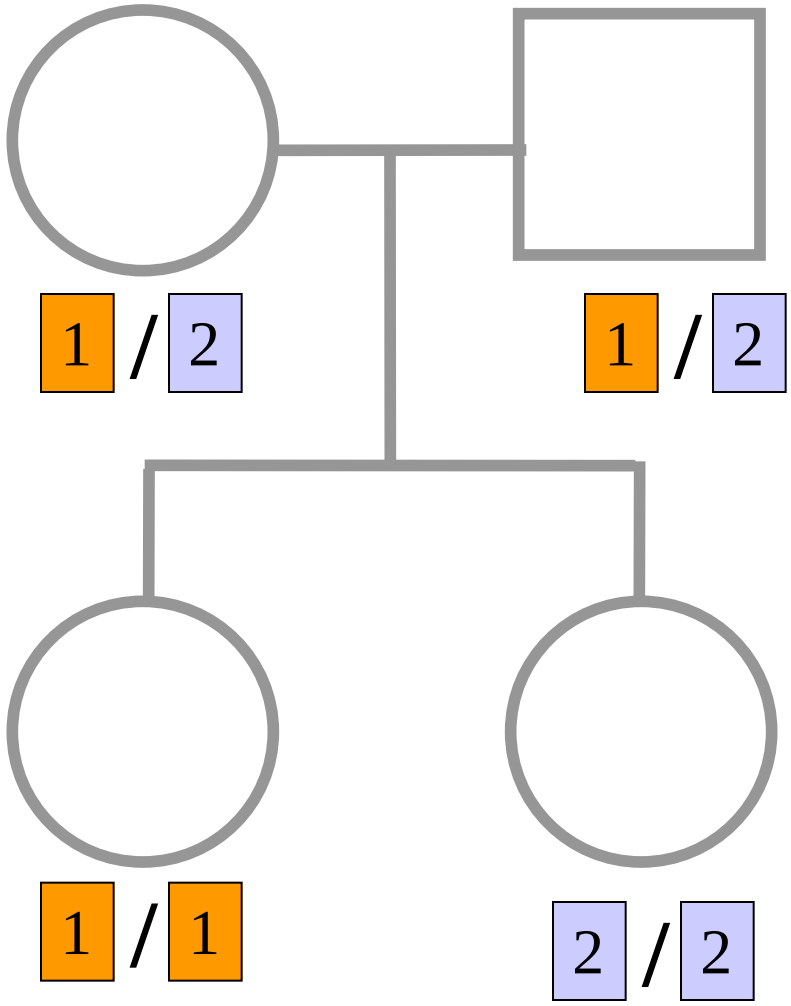
- Average sharing methods

  Calculates approximate IBD distributions (Fulker et al., 1995; Almasy and Blangero, 1998)

Estimating π when marker is not fully informative

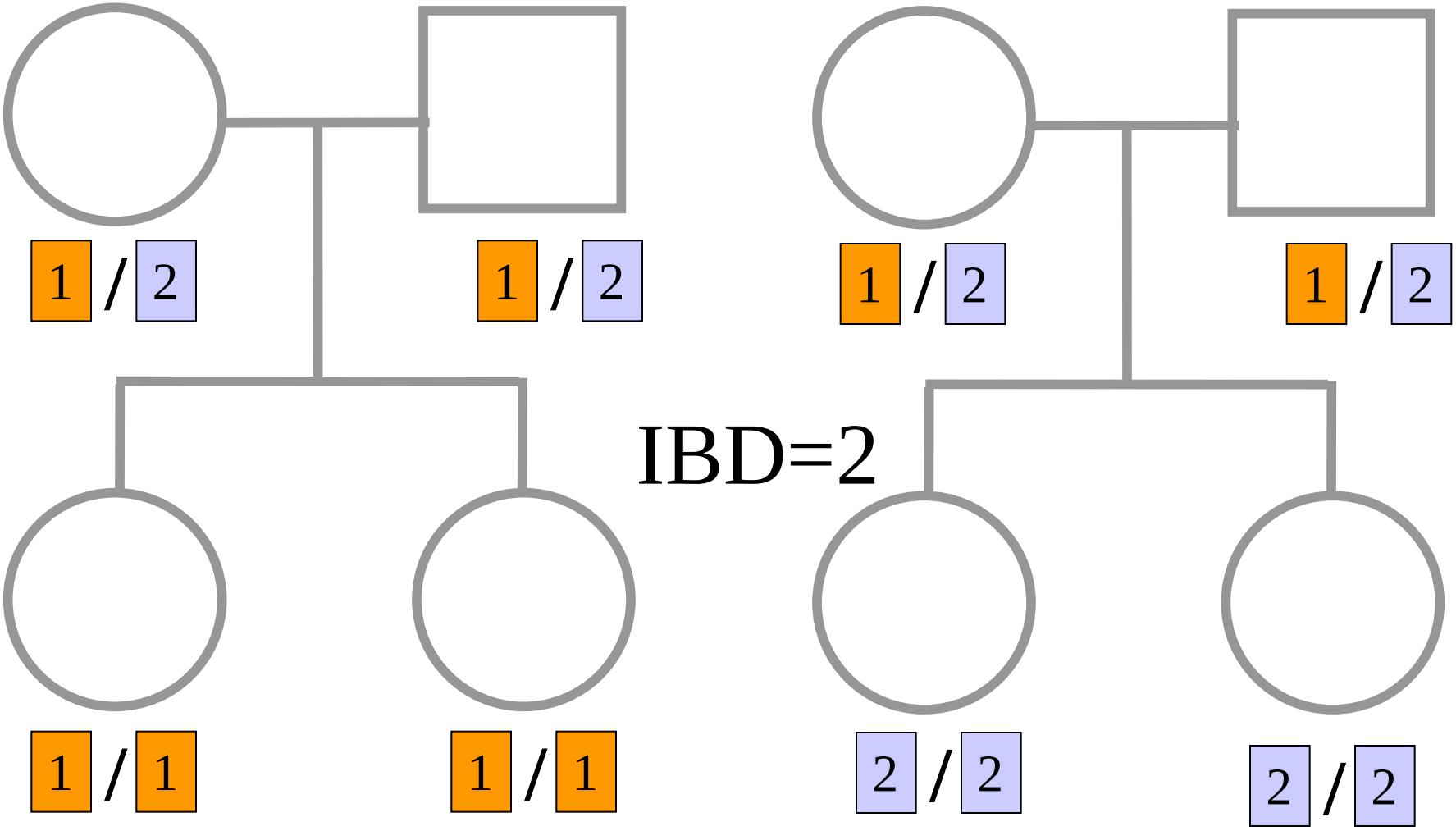
- Using:
  - Mendelian segregation rules
  - Marker allele frequencies in the population

IBD can be trivial…

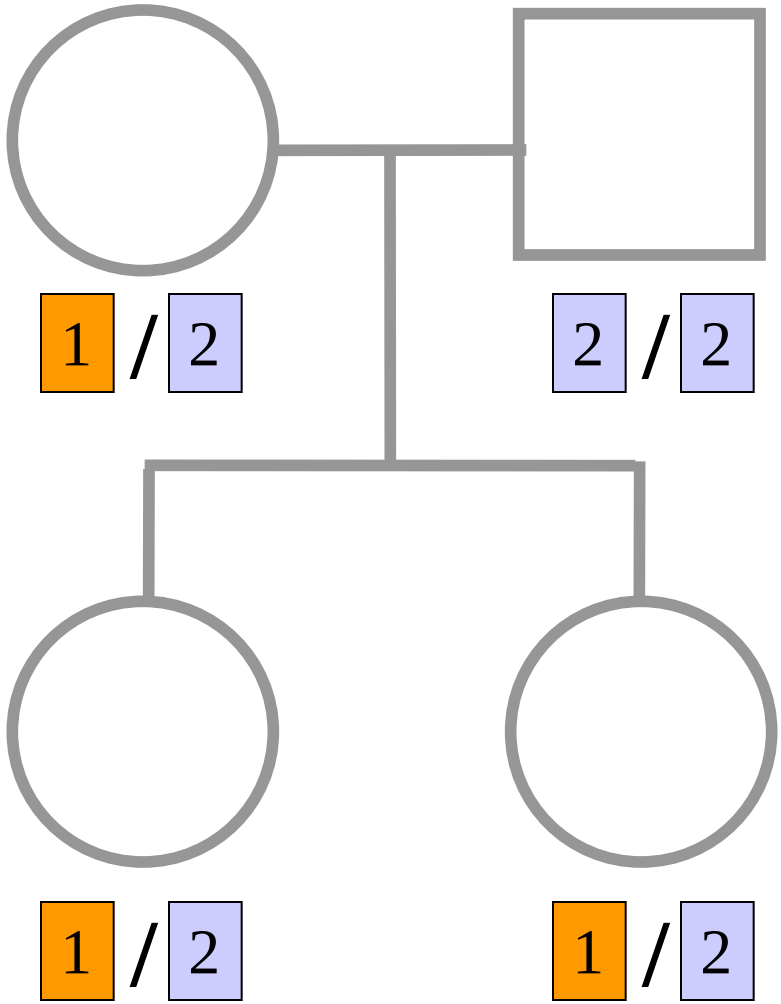


IBD=0

Two Other Simple Cases…



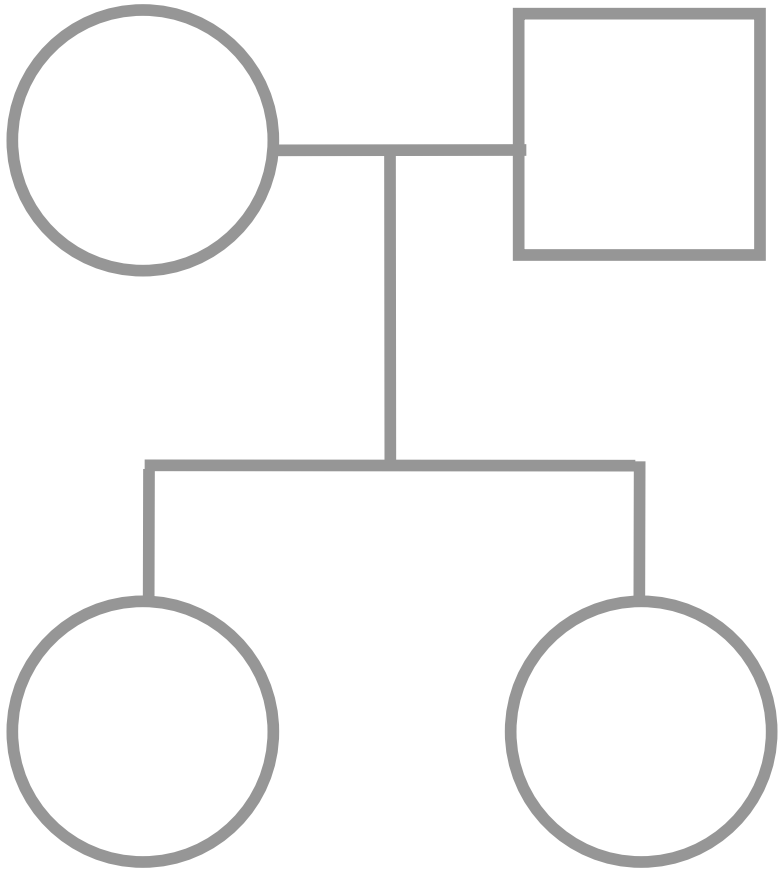
A little more complicated…



IBD=1
(50% chance)

IBD=2
(50% chance)

And even more complicated…



IBD=?

1 / 1

1 / 1

Bayes Theorem for IBD Probabilities

posterior

$$P(IBD = i \mid G) = \frac{\mathrm{P}(\mathrm{IBD} = i, G)}{P(G)}$$

prior

$$= \frac{P(IBD = i)P(G \mid IBD = i)}{P(G)}$$

Prob(data)

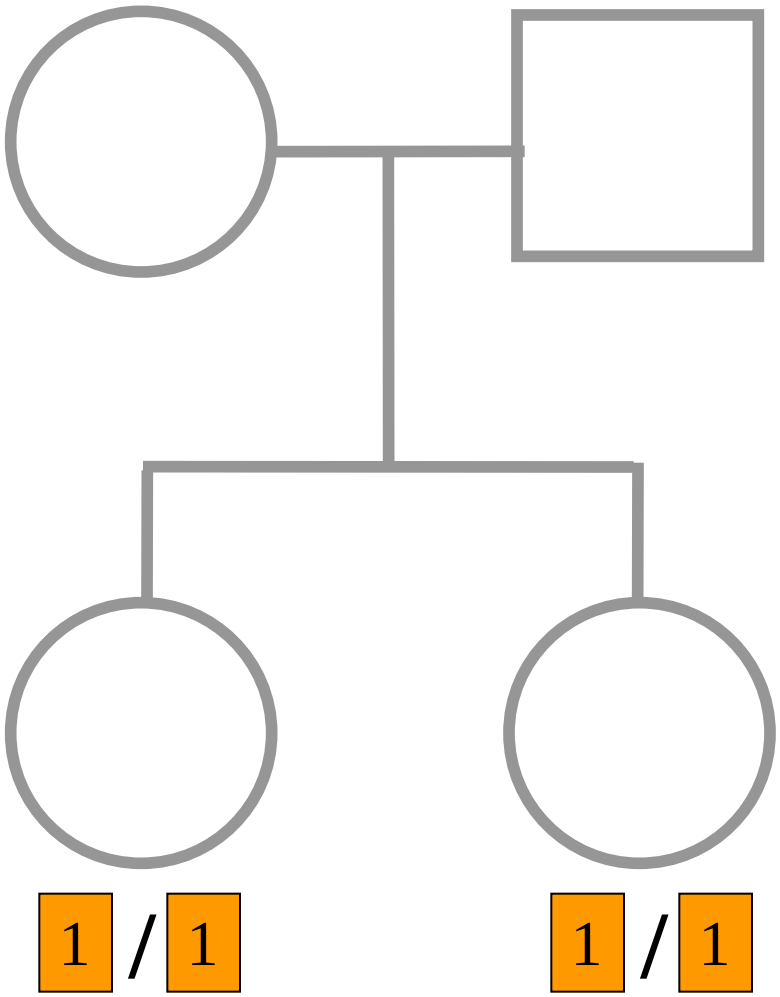
$$= \frac{P(IBD = i)P(G \mid IBD = i)}{\sum_j P(IBD = j)P(G \mid IBD = j)}$$

P(Marker Genotype|IBD State)

		IBD		
Sib	**CoSib**	0	1	2
(a,b)	(c,d)	$p_a p_b p_c p_d$	0	0
(a,a)	(b,c)	$p_a^2 p_b p_c$	0	0
(a,a)	(b,b)	$p_a^2 p_b^2$	0	0
(a,b)	(a,c)	$p_a^2 p_b p_c$	$p_a p_b p_c$	0
(a,a)	(a,b)	$p_a^3 p_b$	$p_a^2 p_b$	0
(a,b)	(a,b)	$p_a^2 p_b^2$	$p_a p_b^2 + p_a^2 p_b$	$p_a p_b$
(a,a)	(a,a)	p_a^4	p_a^3	p_a^2
Prior Probability		¼	½	¼

[Assumes Hardy-Weinberg proportions of genotypes in the population]

Worked Example



$$p_1 = 0.5$$

$$P(G \mid IBD = 0) = p_1^4 = 1/16$$

$$P(G \mid IBD = 1) = p_1^3 = 1/8$$

$$P(G \mid IBD = 2) = p_1^2 = 1/4$$

$$P(G) = 1/4\, p_1^4 + 1/2\, p_1^3 + 1/4\, p_1^2 = 9/64$$

$$P(IBD = 0 \mid G) = \frac{1/4\, p_1^4}{P(G)} = 1/9$$

$$P(IBD = 1 \mid G) = \frac{1/2\, p_1^3}{P(G)} = 4/9$$

$$\hat{\pi} = 2/3$$

$$P(IBD = 2 \mid G) = \frac{1/4\, p_1^2}{P(G)} = 4/9$$

Statistical power

- Calculate the expected value of the test statistic under the alternative hypothesis
- Calculate the variance of the test statistic under the alternative hypothesis

→ Distribution of test statistic under null and alternative hypotheses → power of detection

Linear regression (HE)

$$D = (y_1 - y_2)$$

$$D^2 = (y_1 - y_2)^2$$

Regression

$$Y = \mu + \beta\pi + e$$

$$\text{Test statistic} = \hat{\beta}^2 / \text{var}(\hat{\beta})$$

$$\sim \text{(non)central } \chi^2$$

If we know β, var(Y) and var(π), we can predict the expected test statistic & power

β for additive QTL model (assume $\sigma_y = 1$)

$\beta = -2q^2$

With q^2 the proportion of phenotypic variation due to the QTL

var($\hat{\beta}$) for additive model

$$Y = \mu + \beta\pi + e$$

$$\text{var}(\hat{\beta}) \approx \text{var}(e) / [(n-2)\text{var}(\pi)]$$

$$\text{var}(e) \approx \text{var}(Y) - \beta^2\text{var}(\pi)$$

Analytical predictions

$\beta = -2q^2$

$var(\pi) = {}^1/_8$

$var(Y) =$ Depends on population parameters;

$var(y) = f^2 + q^2 + r^2$

variance due to family effects

variance due to QTL

residual variance

QTL models

- Random QTL effects

  $Q_1 \quad \sim \quad N(0, q^2)$

  $Q_2|\pi \quad \sim \quad N(\pi Q_1, (1-\pi^2)q^2)$

- Fixed QTL effect
  - bi-allelic additive QTL with frequency p
  - $q^2 = 2p(1-p)a^2$

Variance of Y for random QTL (exact)

$$\mathrm{var}(D^2) = 8r^2(1 - f^2) + (^7/_2)q^4$$

[Visscher & Hopper 2001, Annals Human Genetics]

If QTL effects is small (approximation)

- Bivariate normality, sib-correlation ρ

$$\mathrm{var}(D^2) \sim 8(1-\rho)^2$$

$$\rho = f^2 + \tfrac{1}{2}q^2$$

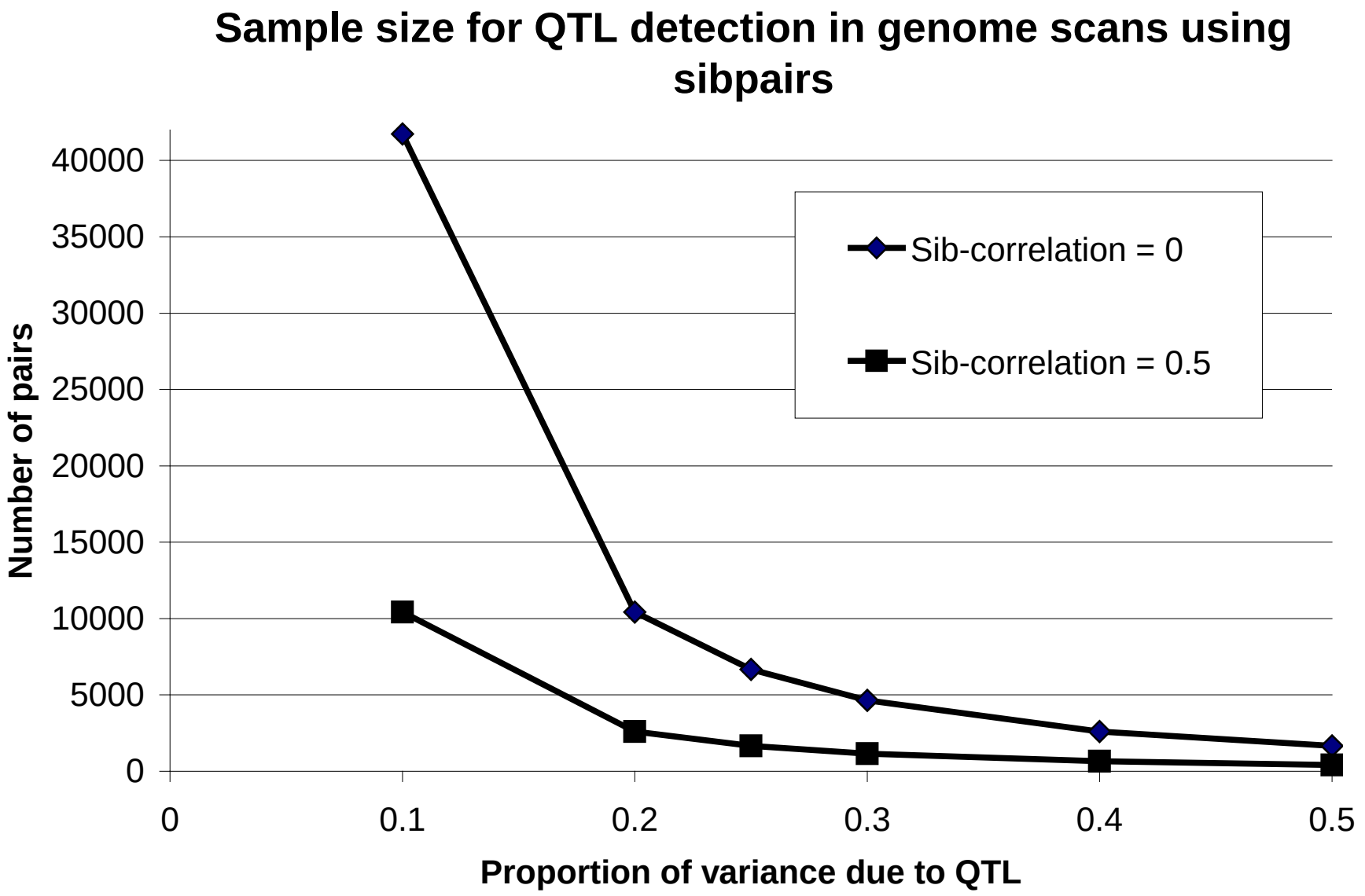
Haseman-Elston regression

$$E(\text{test statistic} \mid \text{QTL}) = E[\hat{\beta}^2/\text{var}(\hat{\beta})]$$

$$= \quad \tfrac{1}{2} \quad + \quad \tfrac{1}{2} n q^4/\text{var}(D^2)$$

$$E(T) = \quad \tfrac{1}{2} \quad + \quad n\lambda$$

$$\lambda \quad = \quad q^4 / [16(1-\rho)^2]$$



Power = 90%. Type-I error = 10^{-5}

Maximum likelihood
(assuming bivariate normality | π)

Full model:

$$-2\ln(L) = \Sigma n_{\pi} \ln|V_{\pi}| + \Sigma(y-\mu)' V_{\pi}^{-1}(y-\mu)$$

$$V_{\pi} = \begin{pmatrix} f^2 + q^2 + r^2 & f^2 + \pi q^2 \\ f^2 + \pi q^2 & f^2 + q^2 + r^2 \end{pmatrix}$$

Maximum likelihood

Reduced model:

$$-2\ln(L) = n\ln|V| + (y-\mu)'V^{-1}(y-\mu)$$

$$V = \begin{pmatrix} f^2 + r^2 & f^2 \\ f^2 & f^2 + r^2 \end{pmatrix}$$

[Fulker & Cherny 1996; Wright 1997]

Maximum Likelihood

$$LRT = 2\ln(ML_{full}) - 2\ln(ML_{reduced})$$

$$E(LRT|QTL) \approx \frac{1}{2} + n\lambda$$

$$\lambda = \ln\{(1-(f^2+\tfrac{1}{2}q^2)^2]^{1/2} / [(1-f^4)(1-(f^2+q^2)^2]^{1/4}\}$$

$$\sim q^4/8 \text{ if all correlations are small}$$

[Fulker & Cherny 1996; Wright 1997]

Genetic Power Calculator (PGC)
http://pngu.mgh.harvard.edu/~purcell/gpc/

Genetic Power Calculator

Genetic Power Calculator

S. Purcell & P. Sham, 2001-2009

This site provides automated power analysis for variance components (VC) quantitative trait locus (QTL) linkage and association tests in sibships, and other common tests. Suggestions, comments, etc to *Shaun Purcell*.

If you use this site, please reference the following Bioinformatics article:

```
Purcell S, Cherny SS, Sham PC. (2003) Genetic Power Calculator:
design of linkage and association genetic mapping studies of complex
traits. Bioinformatics, 19(1):149-150.
```

Modules

Case-control for discrete traits	Notes
Case-control for threshold-selected quantitative traits	Notes
QTL association for sibships and singletons	Notes
TDT for discrete traits	Notes
TDT and parenTDT with ascertainment	Notes
TDT for threshold-selected quantitative traits	Notes
Epistasis power calculator	Notes
QTL linkage for sibships	Notes
Probability Function Calculator	Notes

Genetic Power Calculator

QTL Linkage for Sibships

QTL additive variance :
QTL dominance variance : ☑ No dominance (* see below)
Residual shared variance :
Residual nonshared variance :
Recombination fraction :

Sample Size :
Sibship Size : 2

User-defined type I error rate : 0.05 (0.00000001 - 0.5)
User-defined power: determine N : 0.80 (0 - 1)
(1 - type II error rate)

Process Reset

LOD score and likelihood-ratio test statistic (LRT)

$$LRT = 2\ln(L_1/L_0) = 2[\ln(L_1)-\ln(L_0)] = -2[\ln(L_0)-\ln(L_1)]$$

$$\begin{aligned} LOD &= \log_{10}(L_1/L_0) \\ &= \ln(L_1/L_0) / \ln(10) \\ &= 2\ln(L_1/L_0) / 2\ln(10) \\ &= LRT / 4.605 \\ &= 0.217\ LRT \end{aligned}$$

[$\log_a(x) = \log_b(x) / \log_b(a)$]

Test statistics for linkage analysis

Method	Test	Asymptotic Distr.
Linear Regression	t-test or F-test	$t_n \approx \sim N(0,1)$ $F_{k,n} \approx \sim (1/k)\chi^2_{(k)}$
Maximum likelihood	LRT = $2\ln(L_1/L_0)$	$\sim \chi^2_{(k)}$
Maximum likelihood	LOD = $\log_{10}(L_1/L_0)$	$\sim 0.217\chi^2_{(k)}$
$-\log_{10}$(p-value)	Arbitrary	$-2\ln(p) \sim \chi^2_{(2)}$
Non-parametric	Z-score	$Z \sim N(0,1)$ $Z^2 \sim \chi^2_{(1)}$

Software for linkage analyses

- Genehunter
- Mendel
- Vitesse
- Allegro
- Simwalk
- Loki
- **Merlin (Computer Practical)**
- Solar