

The genetic dissection of complex traits

Karl W Broman

Department of Biostatistics
Johns Hopkins University

<http://www.biostat.jhsph.edu/~kbroman>

Linkage mapping in mouse and man

Karl W Broman

Department of Biostatistics
Johns Hopkins University

<http://www.biostat.jhsph.edu/~kbroman>

The genetic approach

- Start with the phenotype; find genes that influence it.
 - Allelic differences at the genes result in phenotypic differences.
- Value: Need not know anything in advance.
- Goal
 - Understanding the disease etiology (e.g., pathways)
 - Identify possible drug targets

3

Approaches to gene mapping

- Experimental crosses in model organisms
- Linkage analysis in human pedigrees
 - A few large pedigrees
 - Many small families (e.g., sibling pairs)
- Association analysis in human populations
 - Isolated populations vs. outbred populations
 - Candidate genes vs. whole genome

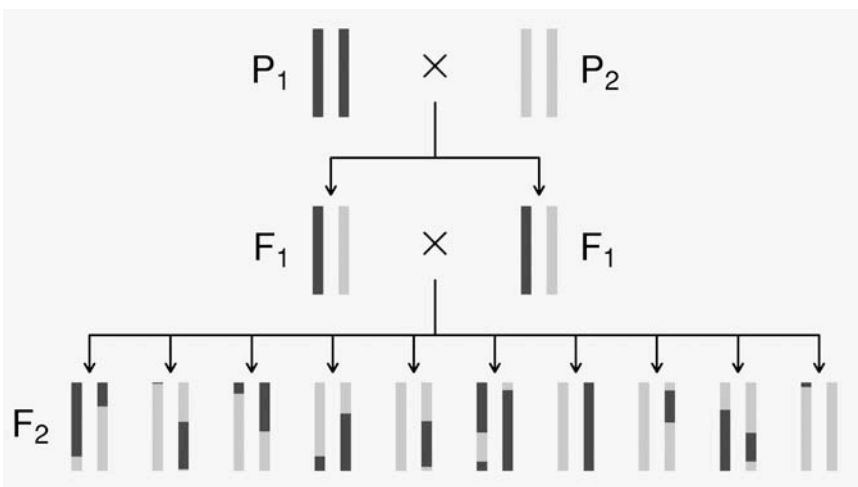
4

Outline

- A bit about experimental crosses
- Meiosis, recombination, genetic maps
- QTL mapping in experimental crosses
- Parametric linkage analysis in humans
- Nonparametric linkage analysis in humans
- ~~QTL mapping in humans~~
- ~~Association mapping~~

5

The intercross



6

The data

- Phenotypes, y_i
- Genotypes, $x_{ij} = AA/AB/BB$, at genetic markers
- A genetic map, giving the locations of the markers.

7

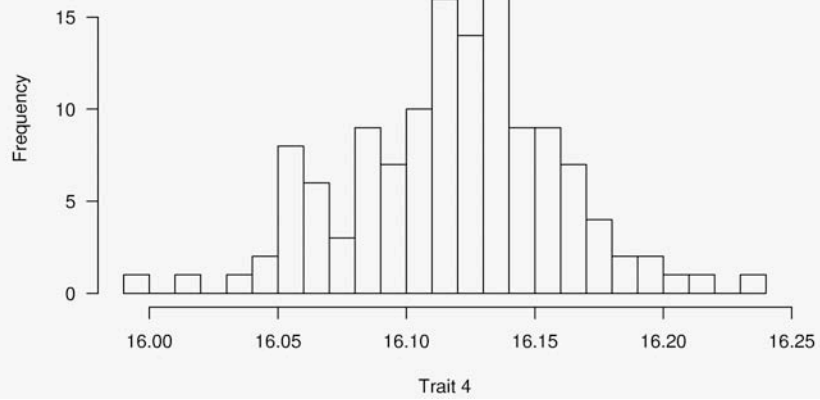
Goals

- Identify genomic regions (QTLs) that contribute to variation in the trait.
- Obtain interval estimates of the QTL locations.
- Estimate the effects of the QTLs.

8

Phenotypes

133 females
(NOD × B6) × (NOD × B6)



9

NOD



10

C57BL/6



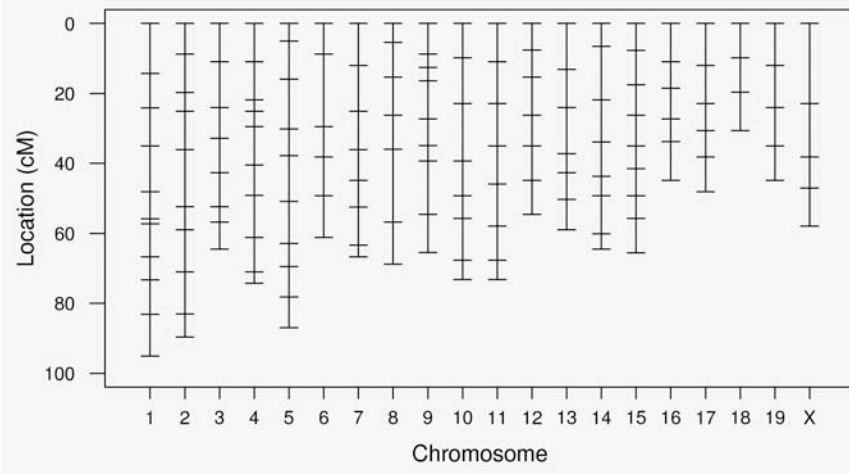
11

Agouti coat



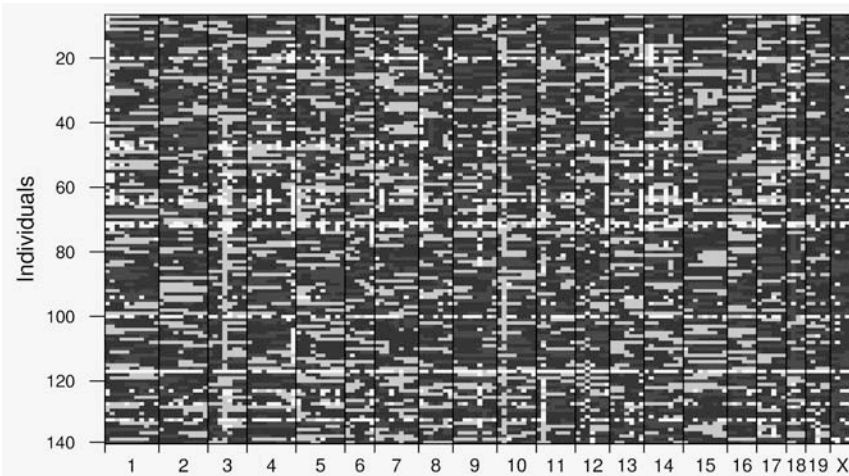
12

Genetic map



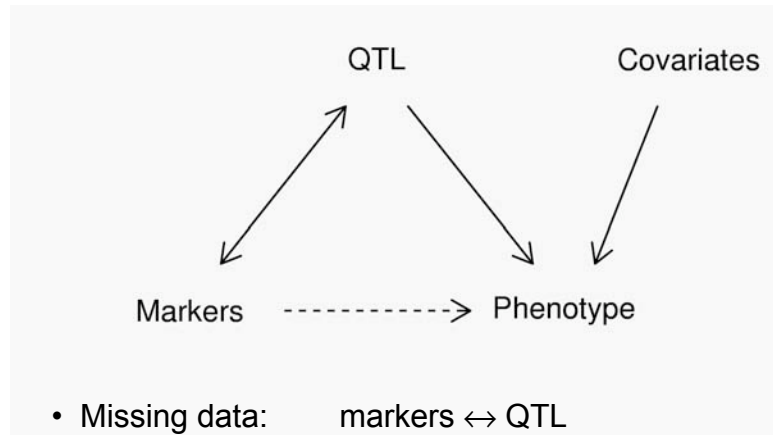
13

Genotype data



14

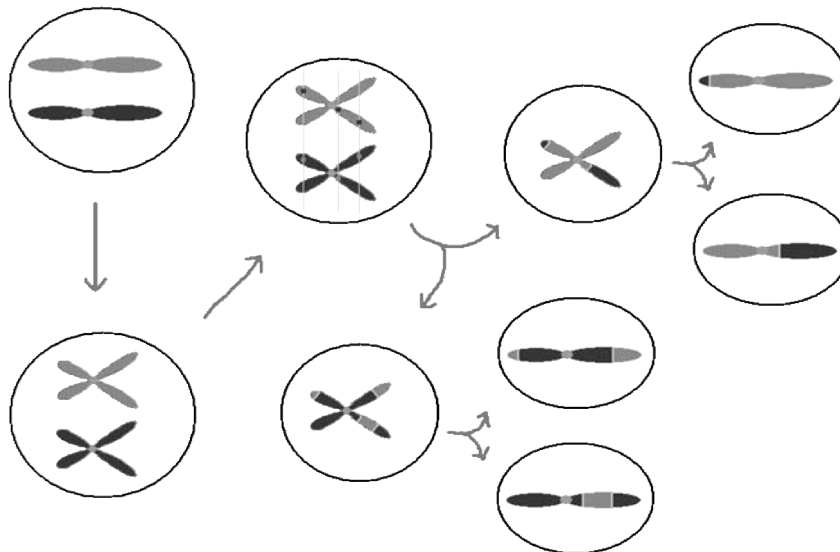
Statistical structure



- Missing data: markers \leftrightarrow QTL
- Model selection: genotypes \leftrightarrow phenotype

15

Meiosis



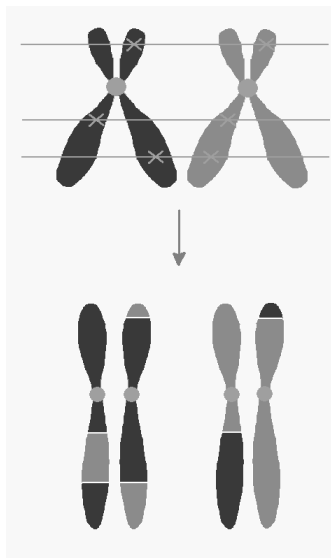
16

Genetic distance

- Genetic distance between two markers (in cM) =
Average number of crossovers in the interval
in 100 meiotic products
- “Intensity” of the crossover point process
- Recombination rate varies by
 - Organism
 - Sex
 - Chromosome
 - Position on chromosome

17

Crossover interference



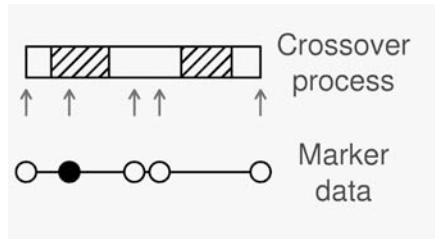
- Strand choice
→ Chromatid interference
- Spacing
→ Crossover interference

Positive crossover interference:

Crossovers tend not to occur too close together.

18

Recombination fraction



We generally do not observe the locations of crossovers; rather, we observe the grandparental origin of DNA at a set of genetic markers.

Recombination across an interval indicates an odd number of crossovers.

Recombination fraction =

$$\Pr(\text{recombination in interval}) = \Pr(\text{odd no. XOs in interval})$$

19

Map functions

- A map function relates the genetic length of an interval and the recombination fraction.

$$r = M(d)$$

- Map functions are related to crossover interference, but a map function is not sufficient to define the crossover process.
- Haldane map function: no crossover interference
- Kosambi: similar to the level of interference in humans
- Carter-Falconer: similar to the level of interference in mice

20

Models: recombination

- We assume no crossover interference
 - Locations of breakpoints according to a Poisson process.
 - Genotypes along chromosome follow a Markov chain.
- Clearly wrong, but super convenient.

21

Models: gen \leftrightarrow phe

Phenotype = y , whole-genome genotype = g

Imagine that p sites are all that matter.

$$E(y | g) = \mu(g_1, \dots, g_p)$$

$$SD(y | g) = \sigma(g_1, \dots, g_p)$$

Simplifying assumptions:

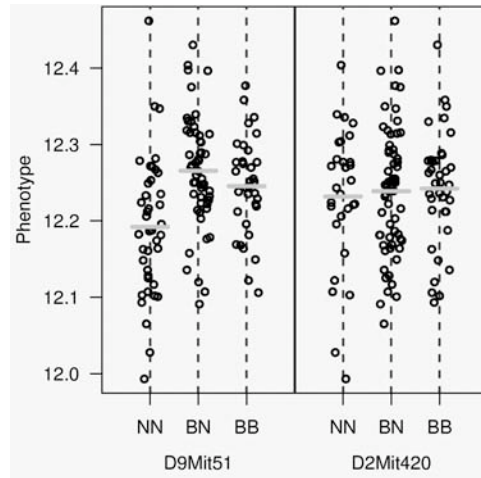
- $SD(y | g) = \sigma$, independent of g
- $y | g \sim \text{normal}(\mu(g_1, \dots, g_p), \sigma)$
- $\mu(g_1, \dots, g_p) = \mu + \sum \alpha_j 1\{g_j = AB\} + \beta_j 1\{g_j = BB\}$

22

The simplest method

“Marker regression”

- Consider a single marker
- Split mice into groups according to their genotype at a marker
- Do an ANOVA (or t-test)
- Repeat for each marker



23

Marker regression

Advantages

- + Simple
- + Easily incorporates covariates
- + Easily extended to more complex models
- + Doesn't require a genetic map

Disadvantages

- Must exclude individuals with missing genotypes data
- Imperfect information about QTL location
- Suffers in low density scans
- Only considers one QTL at a time

24

Interval mapping

Lander and Botstein 1989

- Imagine that there is a single QTL, at position z .
- Let q_i = genotype of mouse i at the QTL, and assume

$$y_i | q_i \sim \text{normal}(\mu(q_i), \sigma)$$

- We won't know q_i , but we can calculate (by an HMM)

$$p_{ig} = \Pr(q_i = g | \text{marker data})$$

- y_i , given the marker data, follows a mixture of normal distributions with known mixing proportions (the p_{ig}).
- Use an EM algorithm to get MLEs of $\theta = (\mu_{AA}, \mu_{AB}, \mu_{BB}, \sigma)$.
- Measure the evidence for a QTL via the LOD score, which is the \log_{10} likelihood ratio comparing the hypothesis of a single QTL at position z to the hypothesis of no QTL anywhere.

25

Interval mapping

Advantages

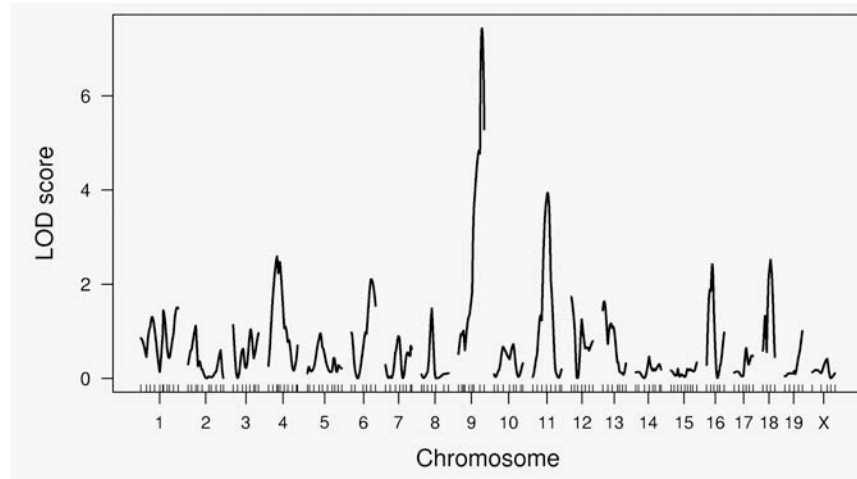
- + Takes proper account of missing data
- + Allows examination of positions between markers
- + Gives improved estimates of QTL effects
- + Provides pretty graphs

Disadvantages

- Increased computation time
- Requires specialized software
- Difficult to generalize
- Only considers one QTL at a time

26

LOD curves



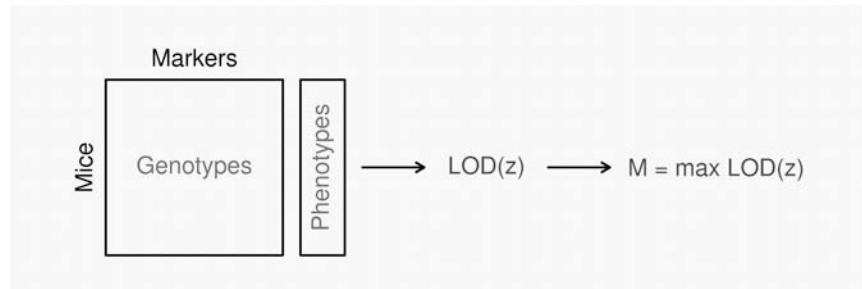
27

LOD thresholds

- To account for the genome-wide search, compare the observed LOD scores to the distribution of the maximum LOD score, genome-wide, that would be obtained if there were no QTL anywhere.
- The 95th percentile of this distribution is used as a significance threshold.
- Such a threshold may be estimated via permutations (Churchill and Doerge 1994).

28

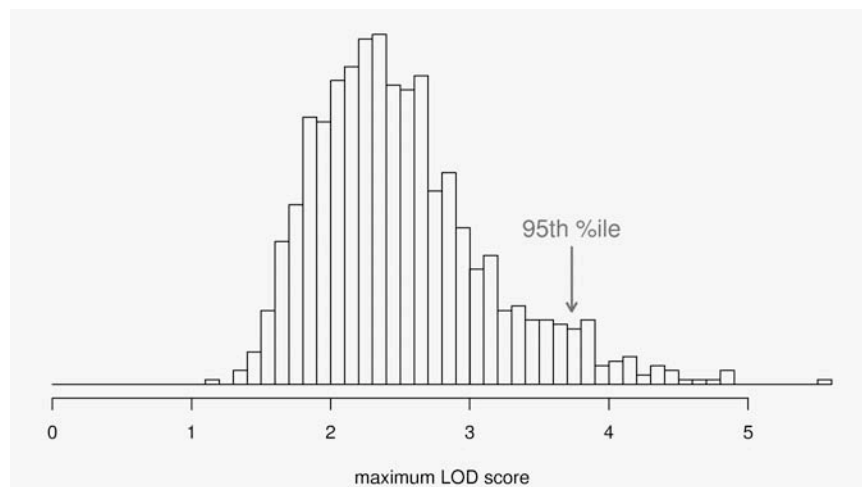
Permutation test



- Shuffle the phenotypes relative to the genotypes.
- Calculate $M^* = \max \text{LOD}^*$, with the shuffled data.
- Repeat many times.
- LOD threshold = 95th percentile of M^* .
- P-value = $\Pr(M^* \geq M)$

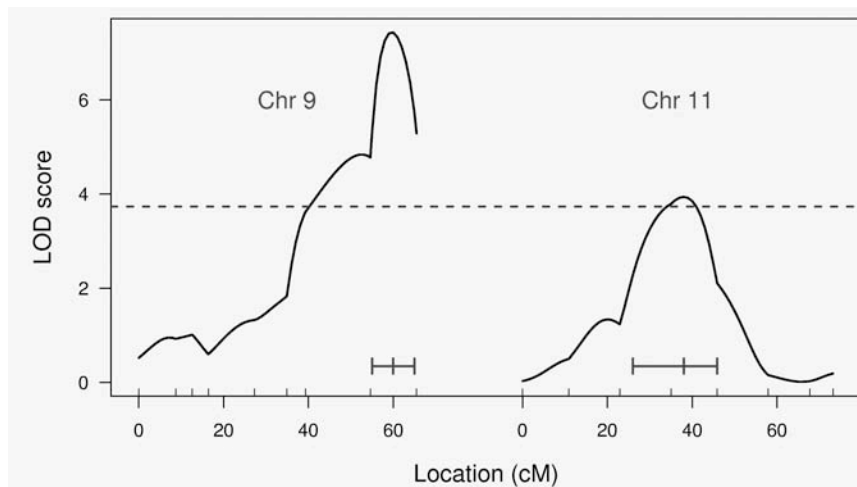
29

Permutation distribution



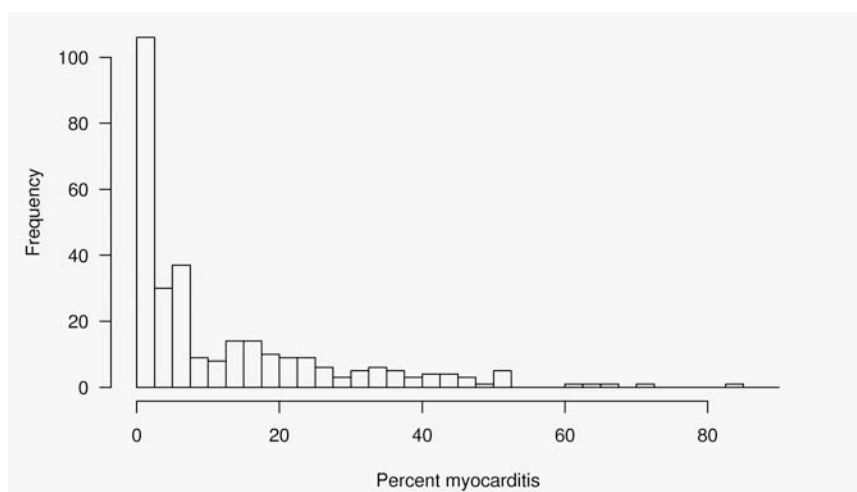
30

Chr 9 and 11



31

Non-normal traits



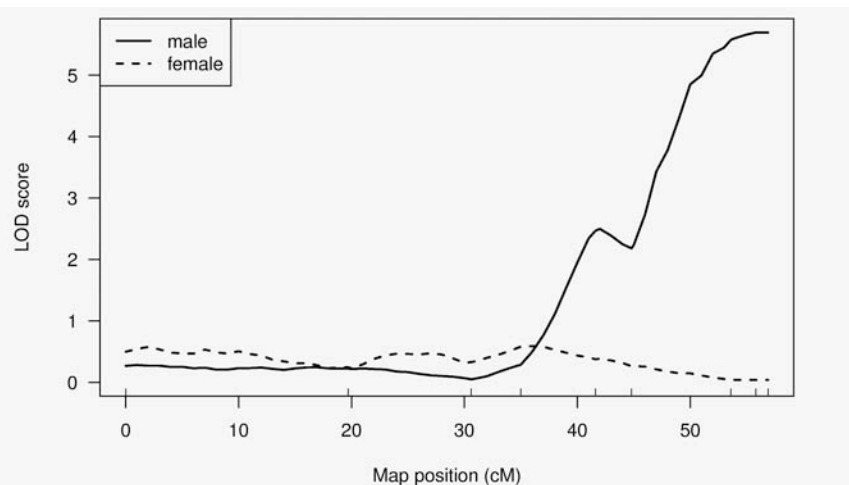
32

Non-normal traits

- Standard interval mapping assumes that the residual variation is normally distributed (and so the phenotype distribution follows a mixture of normal distributions).
- In reality: we see binary traits, counts, skewed distributions, outliers, and all sorts of odd things.
- Interval mapping, with LOD thresholds derived via permutation tests, often performs fine anyway.
- Alternatives to consider:
 - Nonparametric linkage analysis (Kruglyak and Lander 1995).
 - Transformations (e.g., log or square root).
 - Specially-tailored models (e.g., a generalized linear model, the Cox proportional hazards model, the model of Broman 2003).

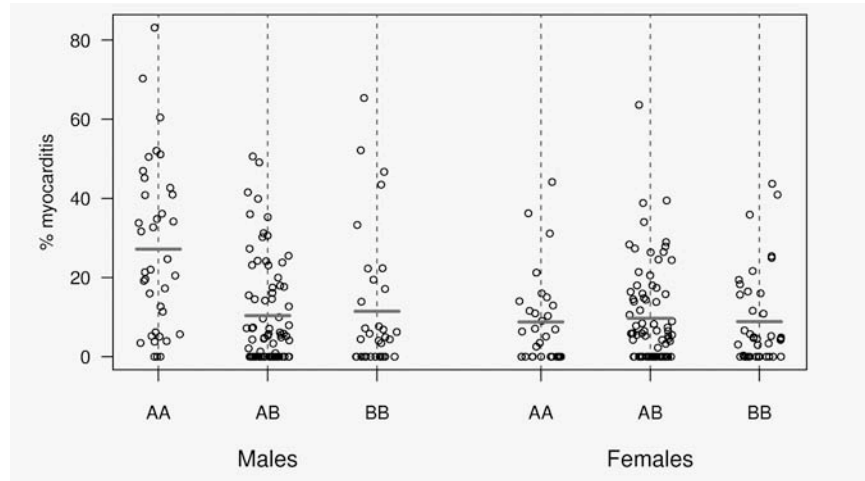
33

Split by sex



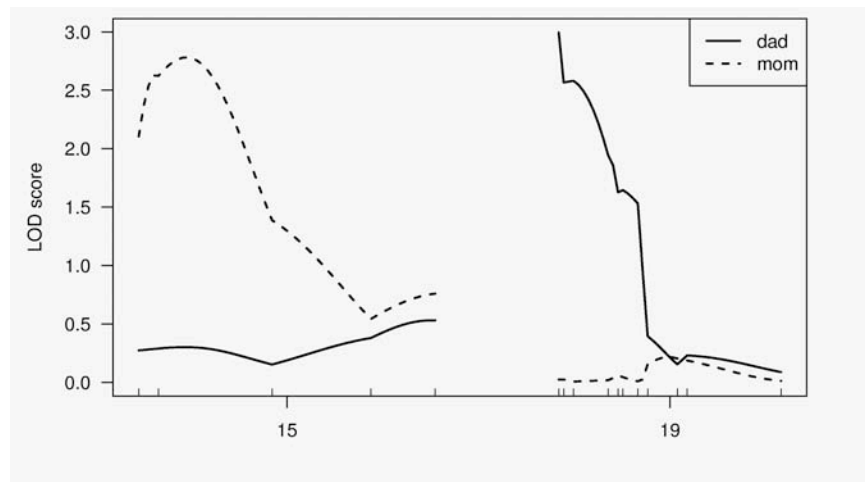
34

Split by sex



35

Split by parent-of-origin



36

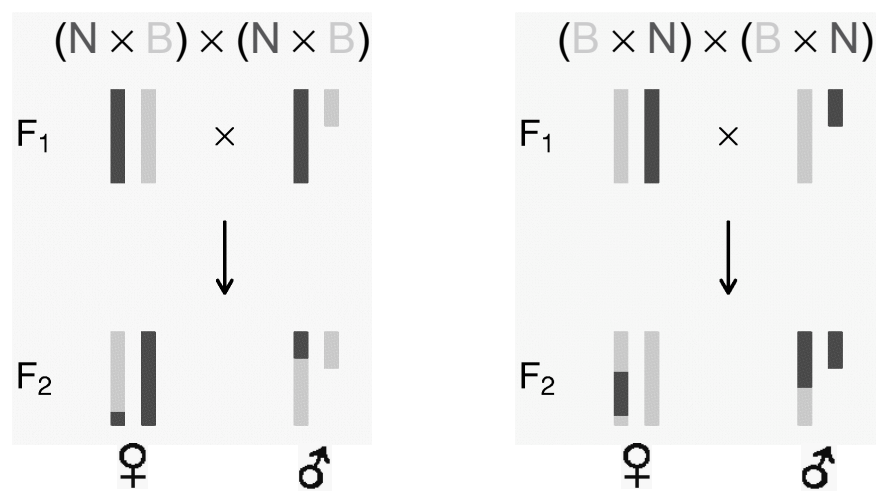
Split by parent-of-origin

Percent of individuals with phenotype

P-O-O	Genotype at D15Mit252		Genotype at D19Mit59	
	AA	AB	AA	AB
Dad	63%	54%	75%	43%
Mom	57%	23%	38%	40%

37

The X chromosome



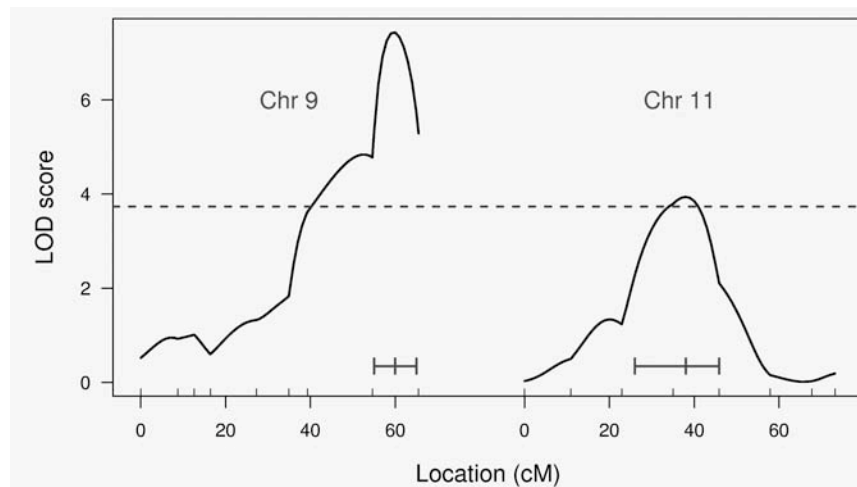
38

The X chromosome

- $BB \equiv BY?$ $NN \equiv NY?$
 - Different “degrees of freedom”
 - Autosome $NN : NB : BB$
 - Females, one direction $NN : NB$
 - Both sexes, both dir. $NY : NN : NB : BB : BY$
- ⇒ Need an X-chr-specific LOD threshold.
- “Null model” should include a sex effect.

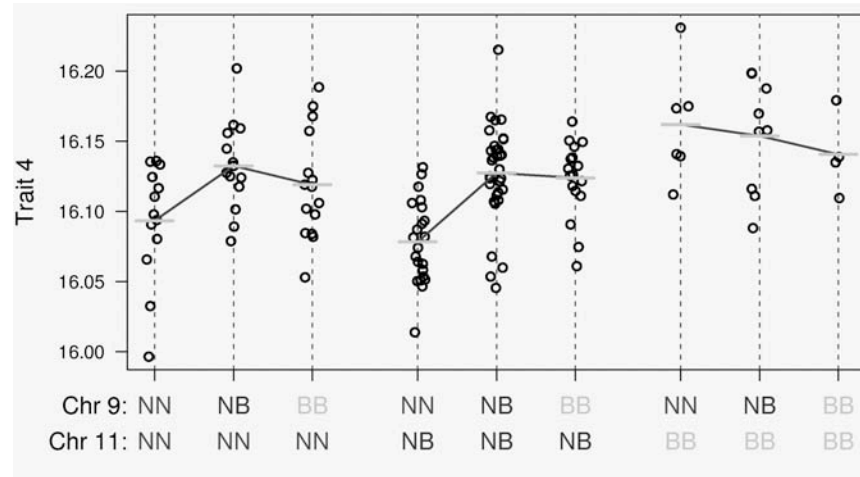
39

Chr 9 and 11



40

Epistasis



41

Going after multiple QTLs

- Greater ability to detect QTLs.
- Separate linked QTLs.
- Learn about interactions between QTLs (epistasis).

42

Model selection

- Choose a class of models.
 - Additive; pairwise interactions; regression trees
- Fit a model (allow for missing genotype data).
 - Linear regression; ML via EM; Bayes via MCMC
- Search model space.
 - Forward/backward/stepwise selection; MCMC
- Compare models.
 - $BIC_{\delta}(\gamma) = \log L(\gamma) + (\delta/2) |\gamma| \log n$

Miss important loci \leftrightarrow include extraneous loci.

43

Special features

- Relationship among the covariates
- Missing covariate information
- Identify the key players vs. minimize prediction error

44

Before you do anything...

Check data quality

- Genetic markers on the correct chromosomes
- Markers in the correct order
- Identify and resolve likely errors in the genotype data

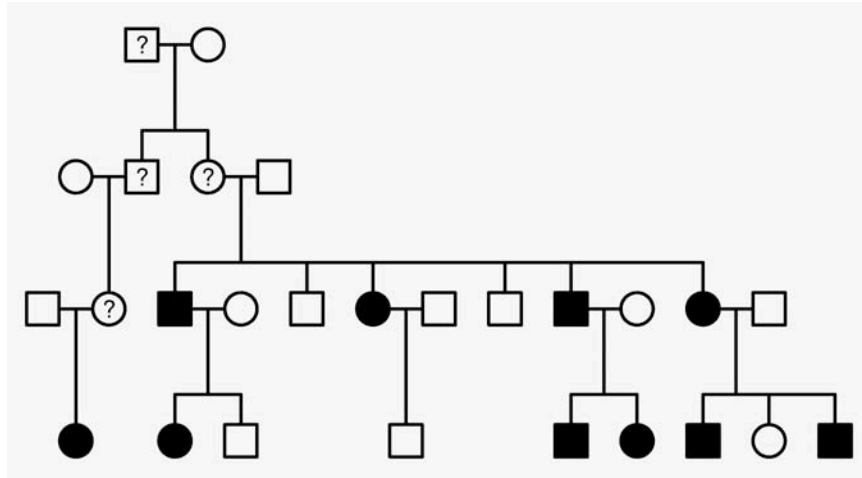
45

Software

- R/qtl
<http://www.biostat.jhsph.edu/~kbroman/qtl>
- Mapmaker/QTL
http://www.broad.mit.edu/genome_software
- Mapmanager QTX
<http://www.mapmanager.org/mmQTX.html>
- QTL Cartographer
<http://statgen.ncsu.edu/qtlcart/index.php>
- Multimapper
<http://www.rni.helsinki.fi/~mjs>

46

Linkage in large human pedigrees



47

Before you do anything...

- Verify relationships between individuals
- Identify and resolve genotyping errors
- Verify marker order, if possible
- Look for apparent tight double crossovers, indicative of genotyping errors

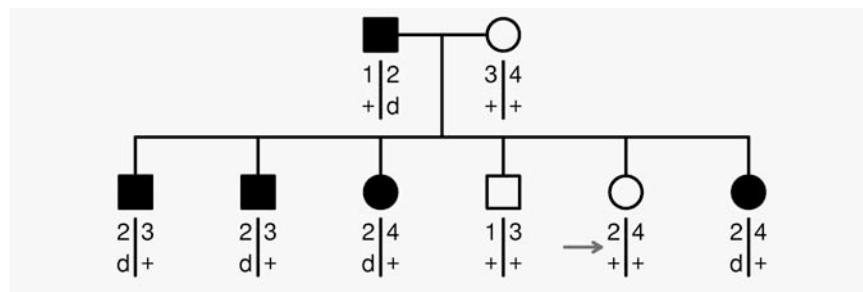
48

Parametric linkage analysis

- Assume a specific genetic model. For example:
 - One disease gene with 2 alleles
 - Dominant, fully penetrant
 - Disease allele frequency known to be 1%.
- Single-point analysis (aka two-point)
 - Consider one marker (and the putative disease gene)
 - θ = recombination fraction between marker and disease gene
 - Test $H_0: \theta = 1/2$ vs. $H_a: \theta < 1/2$
- Multipoint analysis
 - Consider multiple markers on a chromosome
 - θ = location of disease gene on chromosome
 - Test gene unlinked ($\theta = \infty$) vs. θ = particular position

49

Phase known

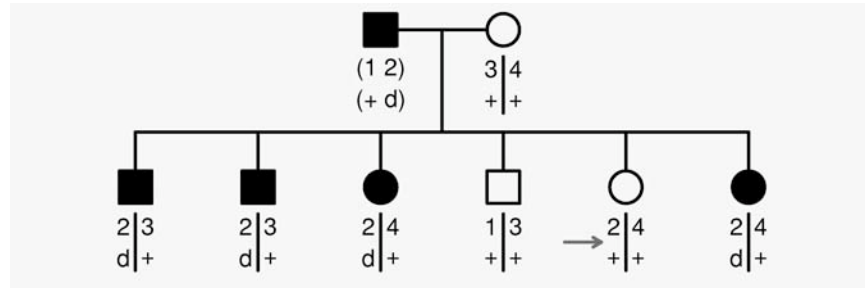


$$L(\theta) = \Pr(\text{data} | \theta) = \theta^1(1-\theta)^5$$

$$\text{LOD score} = \log_{10} \left\{ \frac{\max_{\theta} L(\theta)}{L(\theta = 1/2)} \right\}$$

50

Phase unknown

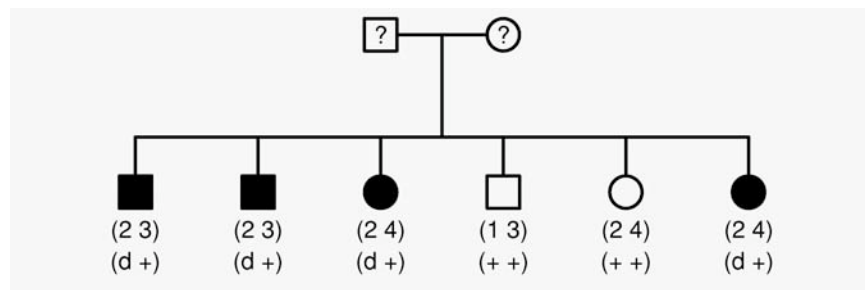


$$L(\theta) = \theta^1(1-\theta)^5 + \theta^5(1-\theta)^1$$

$$\text{LOD score} = \log_{10} \left\{ \frac{\max_{\theta} L(\theta)}{L(\theta = 1/2)} \right\}$$

51

Missing data



The likelihood now involves a sum over possible parental genotypes, and we need:

- Marker allele frequencies
- Further assumptions: Hardy-Weinberg and linkage equilibrium

52

More generally

- Simple diallelic disease gene
 - Alleles d and + with frequencies p and 1-p
 - Penetrances f_0, f_1, f_2 , with $f_i = \Pr(\text{affected} \mid i \text{ d alleles})$
- Possible extensions:
 - Penetrances vary depending on parental origin of disease allele
 $f_1 \rightarrow f_{1m}, f_{1p}$
 - Penetrances vary between people (according to sex, age, or other known covariates)
 - Multiple disease genes
- We assume that the penetrances and disease allele frequencies are known

53

Likelihood calculations

- Define
 - g = complete ordered (aka phase-known) genotypes for all individuals in a family
 - x = observed “phenotype” data (including phenotypes and phase-unknown genotypes, possibly with missing data)

- For example:

$$g_i = \begin{array}{c|c} 3 & 2 \\ 1 & 2 \\ d & + \\ 5 & 4 \end{array} \quad x_i = \left\{ \begin{array}{l} (2 \ 3) \\ (1 \ 2) \\ \text{unaffected} \\ (- \ -) \end{array} \right\}$$

- Goal: $L(\theta) = \Pr(x \mid \theta) = \sum_g \Pr(g) \Pr(x \mid g, \theta)$

54

The parts

- Prior = $\text{Pop}(g_i)$ Founding genotype probabilities
- Penetrance = $\text{Pen}(x_i | g_i)$ Phenotype given genotype
- Transmission Transmission parent \rightarrow child
 $= \text{Tran}(g_i | g_{m(i)}, g_{f(i)})$

Note: If $g_i = (u_i, v_i)$, where u_i = haplotype from mom and v_i = that from dad

Then $\text{Tran}(g_i | g_{m(i)}, g_{f(i)}) = \text{Tran}(u_i | g_{m(i)}) \text{Tran}(v_i | g_{f(i)})$

55

Examples

$$\text{Pop}\left(g_i = \begin{array}{c|c} 1 & 2 \\ \hline d & + \end{array}\right) = p_1 \cdot p_2 \cdot p \cdot (1-p)$$

$$\text{Pen}\left(x_i = \begin{array}{c} (1 \ 2) \\ \text{affected} \end{array} \mid g_i = \begin{array}{c|c} 1 & 2 \\ \hline d & + \end{array}\right) = f_1$$

$$\text{Tran}\left(g_i = \begin{array}{c|c} 1 & 2 \\ \hline d & + \end{array} \mid g_{m(i)} = \begin{array}{c|c} 1 & 3 \\ \hline + & d \end{array} \quad g_{f(i)} = \begin{array}{c|c} 4 & 2 \\ \hline + & + \end{array}\right) = \left(\frac{1}{2}\theta\right) \cdot \frac{1}{2}$$

56

The likelihood

$$\Pr(x) = \sum_g \Pr(g) \Pr(x | g)$$

$$\Pr(x | g) = \prod_i \text{Pen}(x_i | g_i) \quad \text{Phenotypes conditionally independent given genotypes}$$

$$\Pr(g) = \prod_{i \in F} \text{Pop}(g_i) \prod_{i \notin F} \text{Tran}(g_i | g_{m(i)}, g_{f(i)})$$

F = set of "founding" individuals

57

That's a mighty big sum!

- With a marker having k alleles and a diallelic disease gene, we have a sum with $(2k)^{2n}$ terms.
- Solution:
 - Take advantage of conditional independence to factor the sum
 - Elston-Stewart algorithm: Use conditional independence in pedigree
 - Good for large pedigrees, but blows up with many loci
 - Lander-Green algorithm: Use conditional independence along chromosome (assuming no crossover interference)
 - Good for many loci, but blows up in large pedigrees

58

Ascertainment

- We generally select families according to their phenotypes. (For example, we may require at least two affected individuals.)
- How does this affect linkage?

If the genetic model is known, it doesn't: we can condition on the observed phenotypes.

$$\begin{aligned} \text{LOD} &= \frac{\max_{\theta} \Pr(\text{data} \mid \theta)}{\Pr(\text{data} \mid \theta = \frac{1}{2})} = \frac{\max_{\theta} \Pr(M, D \mid \theta)}{\Pr(M, D \mid \theta = \frac{1}{2})} \\ &= \frac{\max_{\theta} \Pr(M \mid D, \theta) \Pr(D \mid \theta)}{\Pr(M \mid D, \theta = \frac{1}{2}) \Pr(D \mid \theta = \frac{1}{2})} = \frac{\max_{\theta} \Pr(M \mid D, \theta)}{\Pr(M \mid D, \theta = \frac{1}{2})} \end{aligned}$$

59

Model misspecification

- To do parametric linkage analysis, we need to specify:
 - Penetrances
 - Disease allele frequency
 - Marker allele frequencies
 - Marker order and genetic map (in multipoint analysis)
- Question: Effect of misspecification of these things on:
 - False positive rate
 - Power to detect a gene
 - Estimate of θ (in single-point analysis)

60

Model misspecification

- Misspecification of disease gene parameters (f 's, p) has little effect on the false positive rate.
- Misspecification of marker allele frequencies can lead to a greatly increased false positive rate.
 - Complete genotype data: marker allele freq don't matter
 - Incomplete data on the founders: misspecified marker allele frequencies can really screw things up
 - BAD: using equally likely allele frequencies
 - BETTER: estimate the allele frequencies with the available data (perhaps even ignoring the relationships between individuals)

61

Model misspecification

- In single-point linkage, the LOD score is relatively robust to misspecification of:
 - Phenocopy rate
 - Effect size
 - Disease allele frequencyHowever, the estimate of θ is generally too large.
- This is less true for multipoint linkage (i.e., multipoint linkage is not robust).
- Misspecification of the degree of dominance leads to greatly reduced power.

62

Other things

- Phenotype misclassification (equivalent to misspecifying penetrances)
- Pedigree and genotyping errors
- Locus heterogeneity
- Multiple genes
- Map distances (in multipoint analysis), especially if the distances are too small.

All lead to:

- Estimate of θ too large
- Decreased power
- Not much change in the false positive rate

Multiple genes generally not too bad as long as you correctly specify the marginal penetrances.

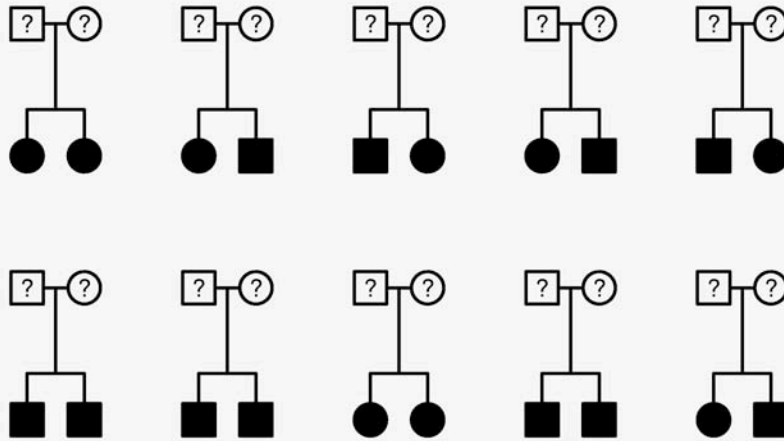
63

Software

- Liped
`ftp://linkage.rockefeller.edu/software/liped`
- Fastlink
`http://www.ncbi.nlm.nih.gov/CBBresearch/Schaffer/fastlink.html`
- Genehunter
`http://www.fhcrc.org/labs/kruglyak/Downloads/index.html`
- Allegro
Email `allegro@decode.is`

64

Linkage in affected sibling pairs



65

Nonparametric linkage

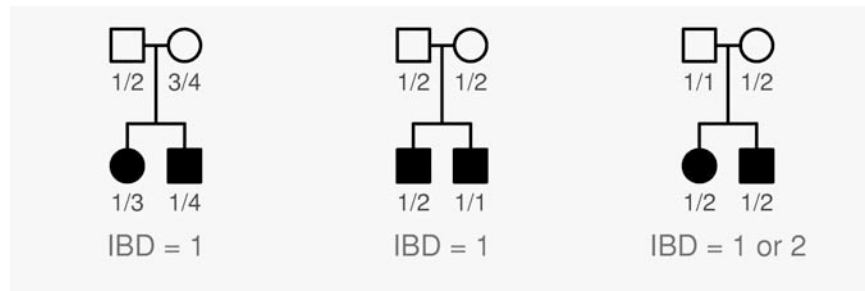
Underlying principle

- Relatives with similar traits should have higher than expected levels of sharing of genetic material near genes that influence the trait.
- “Sharing of genetic material” is measured by identity by descent (IBD).

66

Identity by descent (IBD)

Two alleles are identical by descent if
they are copies of a single ancestral allele



67

IBD in sibpairs

- Two non-inbred individuals share 0, 1, or 2 alleles IBD at any given locus.
- *A priori*, sib pairs are IBD=0,1,2 with probability $1/4$, $1/2$, $1/4$, respectively.
- Affected sibling pairs, in the region of a disease susceptibility gene, will tend to share more alleles IBD.

68

Example

- Single diallelic gene with disease allele frequency = 10%
- Penetrances $f_0 = 1\%$, $f_1 = 10\%$, $f_2 = 50\%$
- Consider position rec. frac. = 5% away from gene

Type of sibpair	IBD probabilities			Ave. IBD
	0	1	2	
Both affected	0.063	0.495	0.442	1.38
Neither affected	0.248	0.500	0.252	1.00
1 affected, 1 not	0.368	0.503	0.128	0.76

69

Complete data case

Set-up

- n affected sibling pairs
- IBD at particular position known exactly
- n_i = no. sibpairs sharing i alleles IBD
- Compare (n_0, n_1, n_2) to $(n/4, n/2, n/4)$
- Example: 100 sibpairs
 $(n_0, n_1, n_2) = (15, 38, 47)$

70

Affected sibpair tests

- Mean test

Let $S = n_1 + 2 n_2$.

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

$$E(S | H_0) = n \quad \text{var}(S | H_0) = n/2$$

$$\text{Let } Z = (S - n) / \sqrt{n/2} \quad \text{LOD} = Z^2 / (2 \ln 10)$$

Example: $S = 132$
 $Z = 4.53$
 $\text{LOD} = 4.45$

71

Affected sibpair tests

- χ^2 test

Let $\pi_0 = (1/4, 1/2, 1/4)$

$$X^2 = \sum_i (n_i - \pi_{0i} n)^2 / \pi_{0i} n$$

Example: $X^2 = 26.2$
 $\text{LOD} = X^2 / (2 \ln 10) = 5.70$

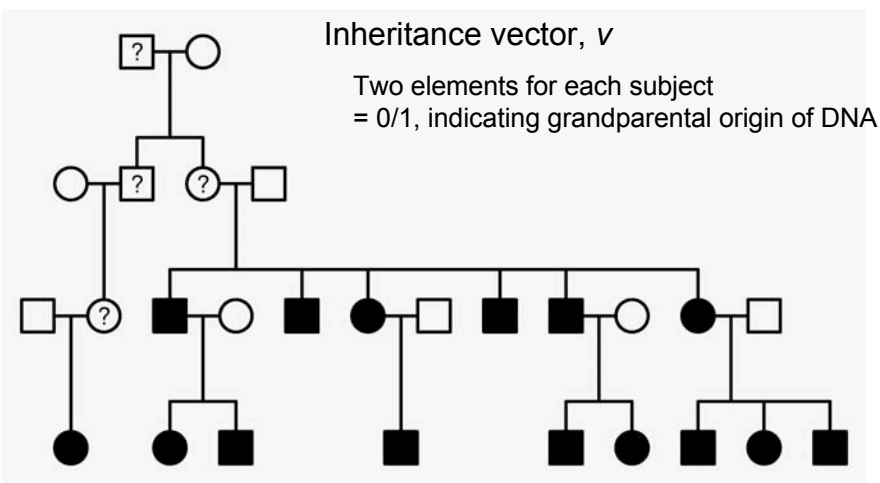
72

Incomplete data

- We seldom know the alleles shared IBD for a sib pair exactly.
- We can calculate, for sib pair i ,
$$p_{ij} = \Pr(\text{sib pair } i \text{ has IBD} = j \mid \text{marker data})$$
- For the means test, we use $\sum_i p_{ij}$ in place of n_j
- Problem: the denominator in the means test, $\sqrt{n/2}$, is correct for perfect IBD information, but is too small in the case of incomplete data
- Most software uses this perfect data approximation, which can make the test conservative (too low power).
- Alternatives: Computer simulation; likelihood methods (e.g., Kong & Cox AJHG 61:1179-88, 1997)

73

Larger families



74

Score function

- $S(v)$ = number measuring the allele sharing among affected relatives
- Examples:
 - $S_{\text{pairs}}(v)$ = sum (over pairs of affected relatives) of no. alleles IBD
 - $S_{\text{all}}(v)$ = a bit complicated; gives greater weight to the case that many affected individuals share the same allele
 - S_{all} is better for dominance or additivity; S_{pairs} is better for recessiveness
- Normalized score, $Z(v) = \{S(v) - \mu\} / \sigma$
 - $\mu = E\{S(v) \mid \text{no linkage}\}$
 - $\sigma = SD\{S(v) \mid \text{no linkage}\}$

75

Combining families

- Calculate the normalized score for each family

$$Z_i = \{S_i - \mu_i\} / \sigma_i$$
- Combine families using weights $w_i \geq 0$

$$Z = \sum_i w_i Z_i / \sqrt{w_i^2}$$
- Choices of weights
 - $w_i = 1$ for all families
 - $w_i = \text{no. sibpairs}$
 - $w_i = \sigma_i$ (i.e., combine the Z_i 's and then standardize)
- Incomplete data
 - In place of S_i , use $\bar{S}_i = \sum_v S_i(v) p(v)$
where $p(v) = \Pr(\text{inheritance vector } v \mid \text{marker data})$

76

Software

- Genehunter
<http://www.fhcrc.org/labs/kruglyak/Downloads/index.html>
- Allegro
Email allegro@decode.is
- Merlin
<http://www.sph.umich.edu/csg/abecasis/Merlin>

77

Summary

- Experimental crosses in model organisms
 - + Cheap, fast, powerful, can do direct experiments
 - The “model” may have little to do with the human disease
- Linkage in a few large human pedigrees
 - + Powerful, studying humans directly
 - Families not easy to identify, phenotype may be unusual, and mapping resolution is low
- Linkage in many small human families
 - + Families easier to identify, see the more common genes
 - Lower power than large pedigrees, still low resolution mapping
- Association analysis
 - + Easy to gather cases and controls, great power (with sufficient markers), very high resolution mapping
 - Need to type an extremely large number of markers (or very good candidates), hard to establish causation

78

References

- Broman KW (2001) Review of statistical methods for QTL mapping in experimental crosses. *Lab Animal* 30:44–52
- Jansen RC (2001) Quantitative trait loci in inbred lines. In Balding DJ et al., *Handbook of statistical genetics*, Wiley, New York, pp 567–597
- Lander ES, Botstein D (1989) Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* 121:185 – 199
- Churchill GA, Doerge RW (1994) Empirical threshold values for quantitative trait mapping. *Genetics* 138:963–971
- Broman KW (2003) Mapping quantitative trait loci in the case of a spike in the phenotype distribution. *Genetics* 163:1169–1175
- Miller AJ (2002) *Subset selection in regression*, 2nd edition. Chapman & Hall, New York

79

References

- Lander ES, Schork NJ (1994) Genetic dissection of complex traits. *Science* 265:2037–2048
- Sham P (1998) *Statistics in human genetics*. Arnold, London
- Lange K (2002) *Mathematical and statistical methods for genetic analysis*, 2nd edition. Springer, New York
- Kong A, Cox NJ (1997) Allele-sharing models: LOD scores and accurate linkage tests. *Am J Hum Gene* 61:1179–1188
- McPeck MS (1999) Optimal allele-sharing statistics for genetic mapping using affected relatives. *Genetic Epidemiology* 16:225–249
- Feingold E (2001) Methods for linkage analysis of quantitative trait loci in humans. *Theor Popul Biol* 60:167–180
- Feingold E (2002) Regression-based quantitative-trait-locus mapping in the 21st century. *Am J Hum Genet* 71:217–222

80