Recombination and Linkage

Karl W Broman

Biostatistics & Medical Informatics
University of Wisconsin – Madison

http://www.biostat.wisc.edu/~kbroman

The genetic approach

- Start with the phenotype; find genes the 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

2

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

3

Linkage vs. association

Advantages

- If you find something, it is real
- Power with limited genotyping
- Numerous rare variants okay

Disadvantages

- Need families
- Lower power if common variant and lots of genotyping
- Low precision of localization

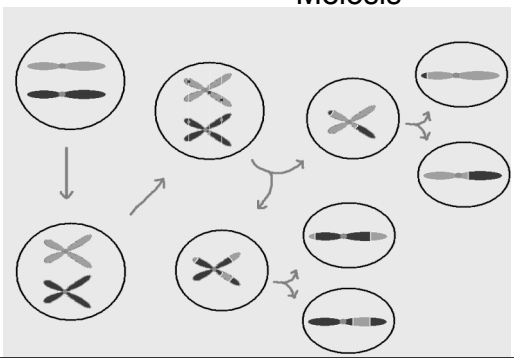
4

Outline

- Meiosis, recombination, genetic maps
- Parametric linkage analysis
- Nonparametric linkage analysis
- Mapping quantitative trait loci

5

Meiosis



6

Genetic distance

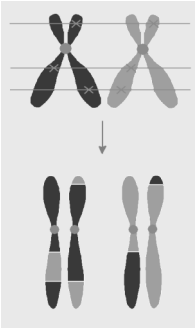
- Genetic distance between two markers (in cM) =

  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

7

Crossover interference



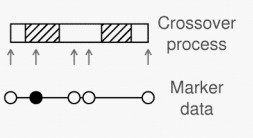
- Strand choice
  → Chromatid interference
- Spacing
  → Crossover interference

Positive crossover interference:

Crossovers tend not to occur too close together.

8

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(recombination in interval) = Pr(odd no. XOs in interval)

9

Map functions

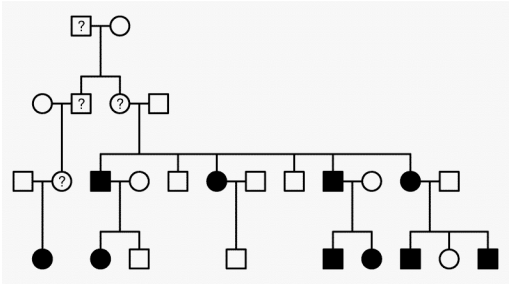
- A map function relates the genetic length of an interval and the recombination fraction.

$$r = M(d)$$

- 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

10

Linkage in large human pedigrees



11

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

12

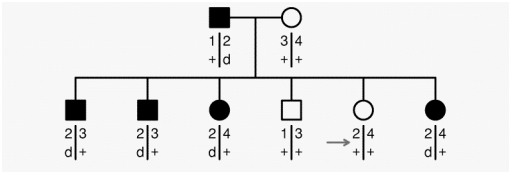
Parametric linkage analysis

- 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: θ = 1/2 vs. H_a: θ < 1/2
- Multipoint analysis
  - Consider multiple markers on a chromosome
  - θ = location of disease gene on chromosome
  - Test gene unlinked (θ = ∞) vs. θ = particular position

13

Phase known

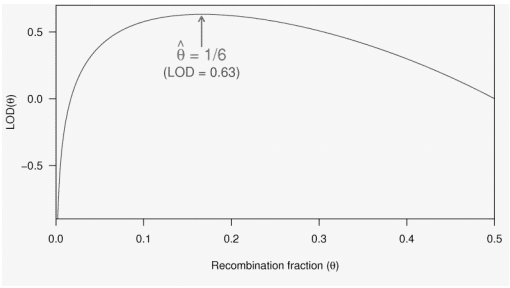


$$L(\theta) = \Pr(\text{data} \mid \theta) = \theta^1(1-\theta)^5$$

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

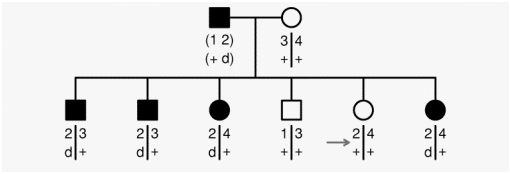
14

Likelihood function



15

Phase unknown

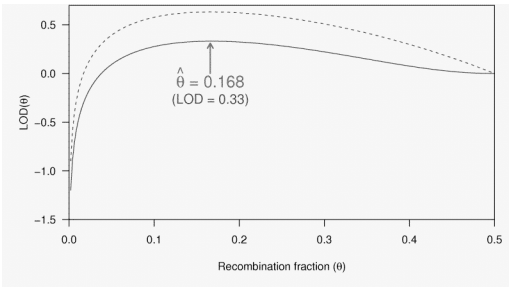


$$L(\theta) = \theta^1(1-\theta)^5 + \theta^5(1-\theta)^1$$

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

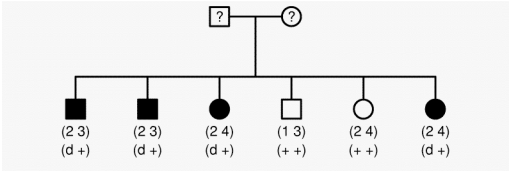
16

Likelihood function



17

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

18

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(affected | i 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

19

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:

$$g_i = \begin{matrix} 3 & 2 \\ 1 & 2 \\ d & + \\ 5 & 4 \end{matrix} \qquad x_i = \left\{ \begin{matrix} (2\ 3) \\ (1\ 2) \\ \text{unaffected} \\ (-\ -) \end{matrix} \right\}$$

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

20

The parts

- Prior = $Pop(g_i)$ — Founding genotype probabilities
- Penetrance = $Pen(x_i \mid g_i)$ — Phenotype given genotype
- Transmission — Transmission parent $\rightarrow$ child

  $= Tran(g_i \mid g_{m(i)}, g_{f(i)})$

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

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

21

Examples

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

$$\text{Pen}\left(x_i = \begin{Bmatrix} (1\ 2) \\ \text{affected} \end{Bmatrix} \;\middle|\; g_i = \begin{matrix} 1 & 2 \\ d & + \end{matrix}\right) = f_1$$

$$\text{Tran}\left(g_i = \begin{matrix} 1 & 2 \\ d & + \end{matrix} \;\middle|\; g_{m(i)} = \begin{matrix} 1 & 3 \\ + & d \end{matrix} \quad g_{f(i)} = \begin{matrix} 4 & 2 \\ + & + \end{matrix}\right) = \left(\tfrac{1}{2}\theta\right) \cdot \tfrac{1}{2}$$

22

The likelihood

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

$$\Pr(x \mid g) = \prod_i \text{Pen}(x_i \mid g_i)$$

Phenotypes conditionally independent given genotypes

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

F = set of "founding" individuals

23

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

24

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.

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

$$= \frac{\max_\theta \Pr(M \mid D,\theta)\Pr(D \mid \theta)}{\Pr(\text{M} \mid D,\theta=\frac{1}{2})\Pr(D \mid \theta=\frac{1}{2})} = \frac{\max_\theta \Pr(M \mid D,\theta)}{\Pr(\text{M} \mid D,\theta=\frac{1}{2})}$$

25

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)

26

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)

27

Model misspecification

- In single-point linkage, the LOD score is relatively robust to misspecification of:
  - Phenocopy rate
  - Effect size
  - Disease allele frequency

  However, 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.

28

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.

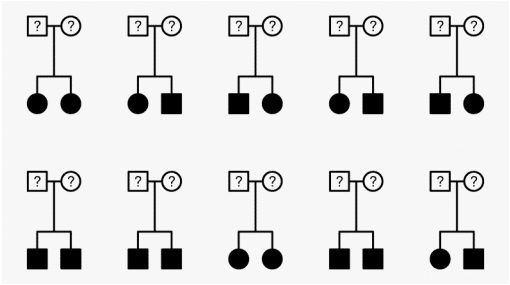
29

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`
- Merlin
  `http://www.sph.umich.edu/csg/abecasis/Merlin`

30

Linkage in affected sibling pairs



31

Nonparametric linkage

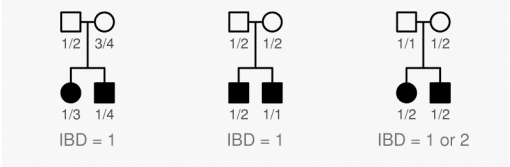
Underlying principle

- 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).

32

Identity by descent (IBD)

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



33

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.

34

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

	IBD probabilities			
Type of sibpair	0	1	2	Ave. IBD
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

35

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
  $(n_0, n_1, n_2) = (15, 38, 47)$

36

Affected sibpair tests

- Mean test

  Let $S = n_1 + 2\,n_2$.

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

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

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

  Example: S = 132
  Z = 4.53
  LOD = 4.45

37

Affected sibpair tests

- χ^2 test

  Let $\boldsymbol{\pi}_0 = (1/4, 1/2, 1/4)$

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

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

38

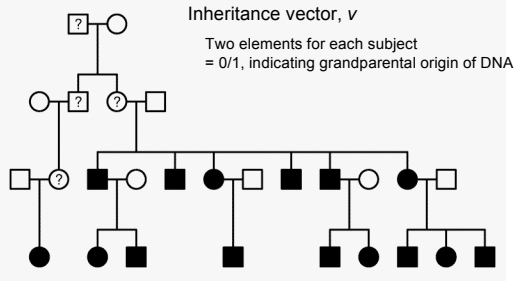
Incomplete data

- We seldom know the alleles shared IBD for a sib pair exactly.
- We can calculate, for sib pair *i*,

  p_{ij} = Pr(sib pair i has IBD = *j* | marker data)
- For the means test, we use $\sum_i p_{ij}$ in place of n_j
- Problem: the deminator 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)

39

Larger families



Inheritance vector, *v*

Two elements for each subject
= 0/1, indicating grandparental origin of DNA

40

Score function

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

41

Combining families

- Calculate the normalized score for each family
  $Z_i = \{S_i - \mu_i\} / \sigma_i$
- Combine families using weights $w_i \geq 0$
  $Z = \sum_i w_i Z_i \Big/ \sqrt{w_i^2}$
- Choices of weights
  - w_i = 1 for all families
  - w_i = 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(inheritance vector v | marker data)

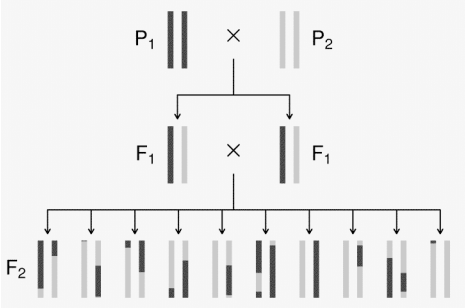
42

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`

43

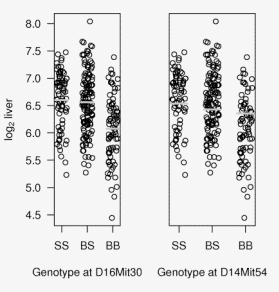
Intercross



44

ANOVA at marker loci

- Split mice into groups according to genotype at marker
- Do a t-test / ANOVA
- Repeat for each marker



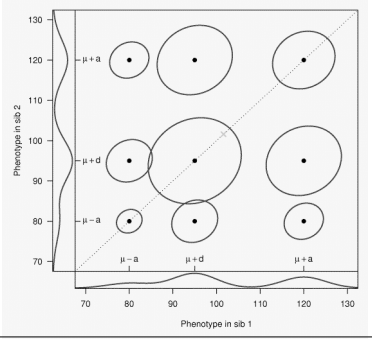
45

Humans vs Mice

- More than two alleles
- Don't know QTL genotypes
- Unknown phase
- Parents may be homozygous
- Markers not fully informative
- Varying environment

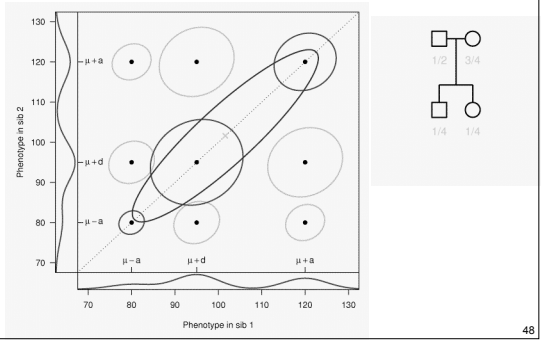
46

Diallelic QTL



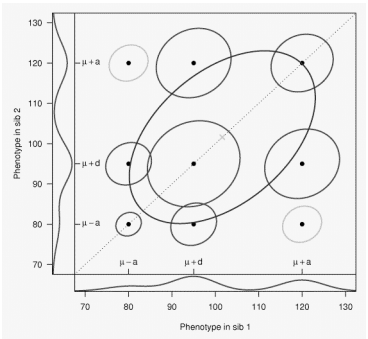
47

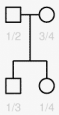
IBD = 2



48

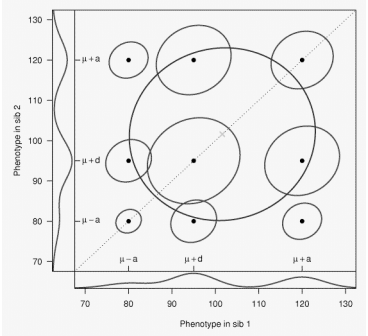
IBD = 1

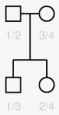




49

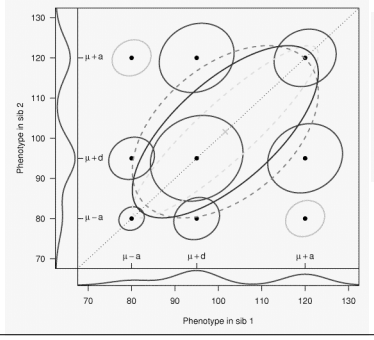
IBD = 0

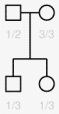




50

IBD = 1 or 2





51

Haseman-Elston regression

For sibling pairs with phenotypes (y_{i1}, y_{i2}),

- Regression the squared difference $(y_{i1} - y_{i2})^2$ on IBD status
- If IBD status is not known precisely, regress on the expected IBD status, given the available marker data

There are a growing number of alternatives to this.

52

Challenges

- Non-normality
- Genetic heterogeneity
- Environmental covariates
- Multiple QTL
- Multiple phenotypes
- Complex ascertainment
- Precision of mapping

53

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

54

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
- McPeek 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

55