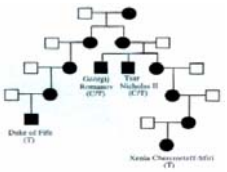
Computational Genomics

10-810/02-710, Spring 2009

Parametric and nonparametric linkage analysis



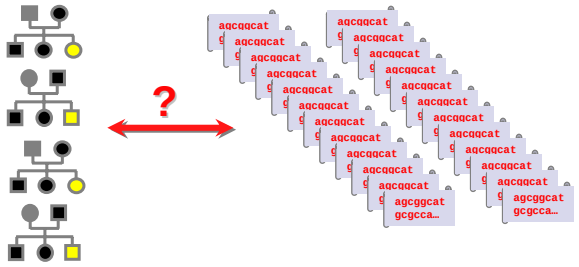
Eric Xing

Lecture 22, April 8, 2009

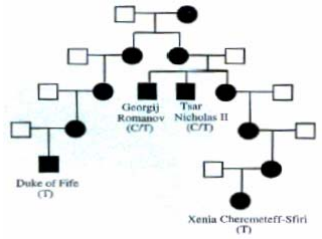
Reading: handouts

A crime or mass-disaster scene

- Given genetic fingerprints of F family pedigrees for alleged victims and genetic fingerprints of S samples found at a disaster site:
  - Who can you confirm died at the site? (legal)
  - Who died at the site that is outside the alleged set? (law enforcement)
  - Cluster the remains for burial. (closure)



Royal pedigree example



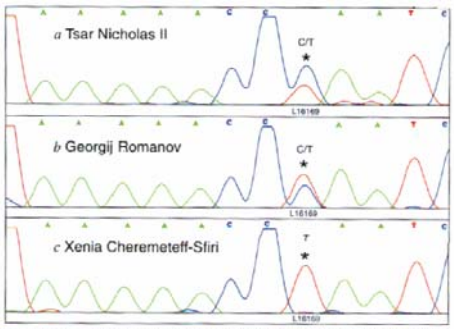
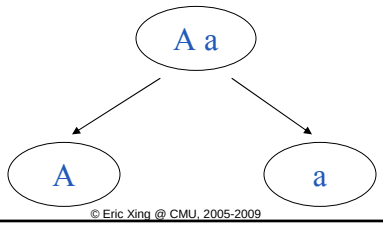


Fig. 2 Automated sequence chromatographs comparing mtDNA sequences at position 16169. a, Sequence from bones of putative Tsar Nicholas II, showing heteroplasmy with cytosine predominating thymine; b, sequence from bones of Grand Duke Georgij Romanov, showing heteroplasmy with thymine predominating cytosine; c, sequence from Countess Xenia Cheremeteff-Sfiri, homoplasmic for thymine.

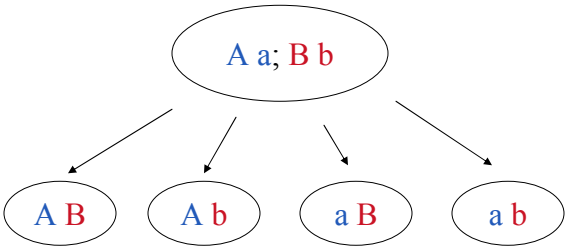
Mendel's two laws

- Modern genetics began with Mendel's experiments on garden peas. He studied seven contrasting pairs of characters, including:
  - The form of ripe seeds: round, wrinkled
  - The color of the seed albumen: yellow, green
  - The length of the stem: long, short
- **Mendel's first law**: Characters are controlled by pairs of genes which separate during the formation of the reproductive cells (meiosis)



Mendel's two laws

- **Mendel's second law**: When two or more pairs of gene segregate simultaneously, they do so independently.



"Exceptions" to Mendel's Second Law

Morgan's fruitfly data (1909): 2,839 flies

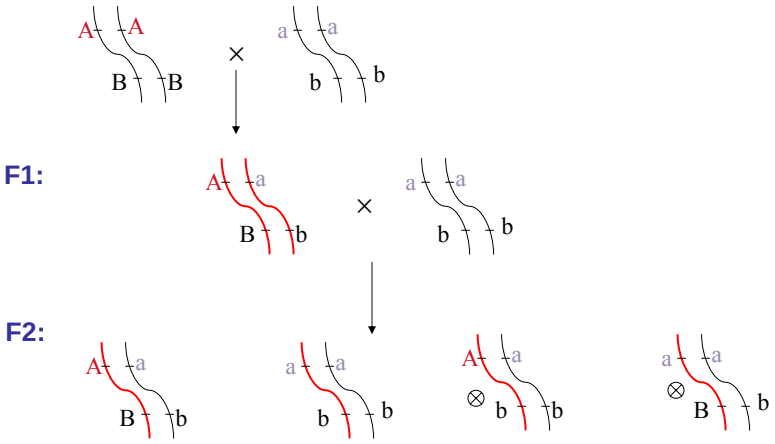
Eye color A: red a: purple
Wing length B: normal b: vestigial

AABB x aabb

AaBb x aabb

	AaBb	Aabb	aaBb	aabb
Exp	710	710	710	710
Obs	1,339	151	154	1,195

Morgan's explanation



⊗ ***Crossover has taken place***

Recombination

- *Parental types*: AaBb, aabb
- *Recombinants*: Aabb, aaBb
  - The proportion of recombinants between the two genes (or characters) is called the ***recombination fraction*** between these two genes.

- ***Recombination fraction*** It is usually denoted by r or θ. For Morgan's traits:

$$r = (151 + 154)/2839 = 0.107$$

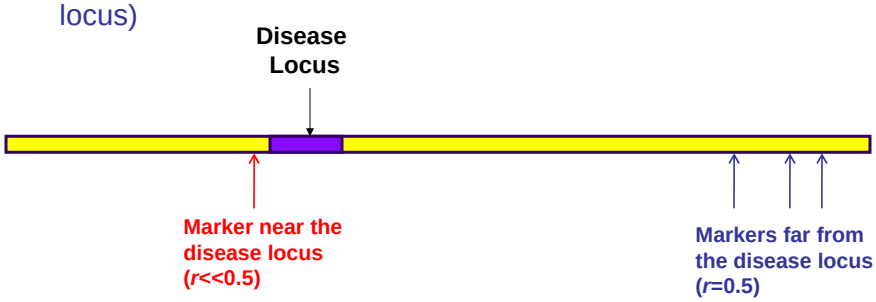
If $r < 1/2$: two genes are said to be ***linked***.

If $r = 1/2$: independent segregation (Mendel's second law).

Now we move on to (small) pedigrees.

Linkage Analysis

- Goal: Identify the unknown disease locus
- Idea: Given pedigree data and a map of genetic markers, let's look for the markers that are linked to the unknown disease locus (i.e. linkage between the disease locus and the marker locus)



DNA Polymorphisms as Genetic Markers

- Microsatellites
  - Many alleles, very informative because of the high heterozygosity (the chance that a randomly selected person will be heterozygous)
- SNPs (single nucleotide polymorphisms)
  - Variation in a single nucleotide
  - Only two alleles at each locus, less informative than microsatellites
  - Advantage: high-throughput genotyping technique is available
  - Haplotypes that combine multiple SNPs can be used as markers

Parametric vs. Nonparametric Linkage Analysis

- Parametric Linkage Analysis
  - Need to specify the disease model
    - Compute LOD-score based on the model for each marker
    - Markers with the high LOD-scores are considered as linked to disease locus
  - Highly effective for Mendelian disease caused by a single locus
  - Usually based on a large pedigree
- Nonparametric Linkage Analysis
  - No need to specify the disease model
  - Multifactorial disease caused by multiple genes
  - Usually based on a large number of small pedigrees with affected siblings and their parents

Parametric Method Based on LOD Scores

One locus: founder probabilities

- ***Founders*** are individuals whose parents are not in the pedigree.
  - They may or may not be typed. Either way, we need to ***assign probabilities*** to their actual or possible genotypes.
  - This is usually done by assuming ***Hardy-Weinberg equilibrium.*** If the **frequency** of *D* is .01, *H-W says*

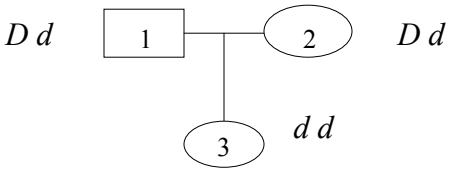
pr(*Dd*) = 2x.01x.99

- Genotypes of *founder couples* are (usually) treated as ***independent.***

pr(pop *Dd* , mom *dd*) = (2x.01x.99)x(.99)^2

One locus: transmission probabilities

- Children get their genes from their parents' genes, independently, according to **Mendel's laws**;

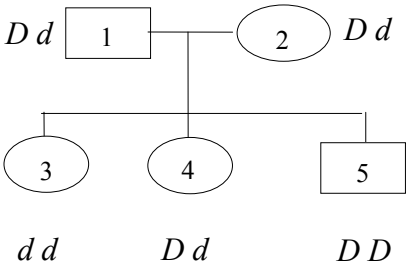


pr(kid 3 *dd* | pop 1 *Dd* & mom 2 *Dd*)

= 1/2 x 1/2

- The inheritances are independent for different children.

One locus: transmission probabilities - II



pr(3 dd & 4 *Dd* & 5 *DD* | 1 *Dd* & 2 *Dd*)
= (1/2 x 1/2)x(2 x 1/2 x 1/2) x (1/2 x 1/2).

- The factor 2 comes from summing over the two mutually exclusive and equiprobable ways 4 can get a *D* and a *d*.

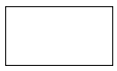
One locus: penetrance probabilities

- Independent Pedetrance Model:
  - Pedigree analyses usually suppose that, given the genotype at all loci, and in some cases age and sex, the chance of having a particular phenotype depends only on genotype at one locus, and is independent of all other factors: genotypes at other loci, environment, genotypes and phenotypes of relatives, etc.
- Complete penetrance:

DD 

pr(affected | *DD*) = 1

- Incomplete penetrance:

DD 

pr(affected | *DD*) = .8

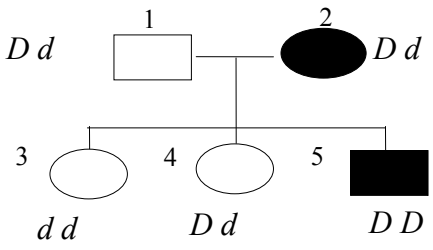
One locus: penetrance - II

- Age and sex-dependent penetrance:

D D (45)

pr(affected | *DD* , male, *45 y.o.*) = .6

One locus: putting it all together



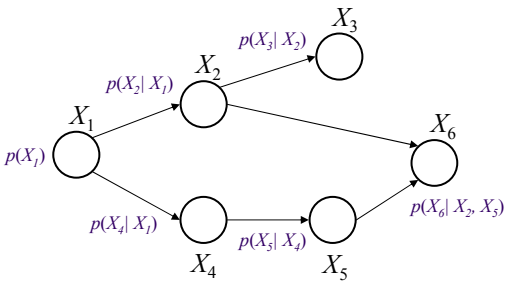
- Assume
  - Penetrances: pr(affected | *dd*) = .1, pr(affected | *Dd*) = .3 pr(affected | *DD*) = .8,
  - and that allele *D* has frequency .01.
  - In general, shaded means affected, blank means unaffected.
- The probability of this pedigree is the product:

  (2 x .01 x .99 x .7) x (2 x .01 x .99 x .3) x (1/2 x 1/2 x .9) x (2 x 1/2 x 1/2 x .7) x (1/2 x 1/2 x .8)

One locus: putting it all together - II

- To write the likelihood of a pedigree:
  - we begin by multiplying founder gene frequencies,
  - followed by founder penetrances.
  - next we multiply transmission probabilities,
  - followed by penetrance probabilities of offspring, using their independence given parental genotypes.
  - If there are missing or incomplete data, we must sum over all mutually exclusive possibilities compatible with the observed data.
- Two algorithms:
  - The general strategy of beginning with founders, then non-founders, and multiplying and summing as appropriate, has been codified in what is known as the Elston-Stewart algorithm for calculating probabilities over pedigrees. It is one of the two widely used approaches.
  - The other is termed the Lander-Green algorithm and takes a quite different approach.
  - Both are hidden Markov models, both have compute time/space limitations with multiple individuals/loci (see next) , and extending them beyond their current limits is the ongoing outstanding problem.

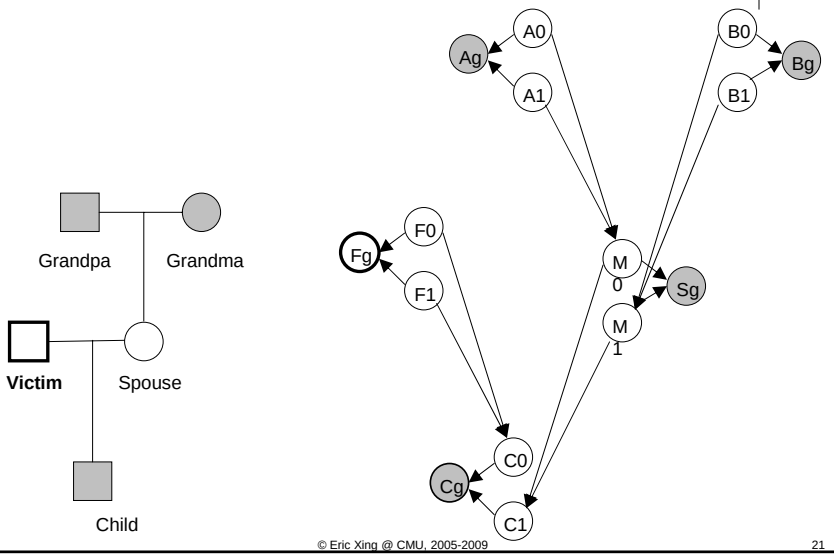
Probabilistic Graphical Models



- The joint distribution on $(X_1, X_2, \ldots, X_N)$ factors according to the "parent-of" relations defined by the edges E :

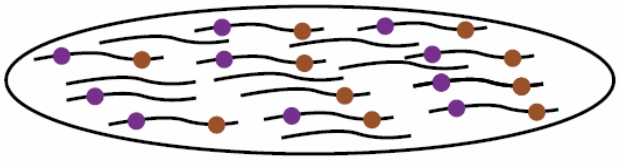
$$p(X_1, X_2, X_3, X_4, X_5, X_6) = p(X_1)\, p(X_2 \mid X_1) p(X_3 \mid X_2)\, p(X_4 \mid X_1) p(X_5 \mid X_4) p(X_6 \mid X_2, X_5)$$

Pedigree as Graphical Models: the allele network

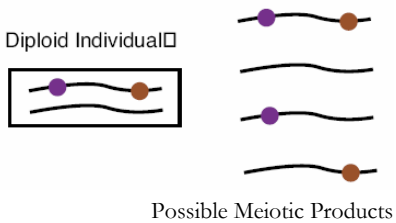


Linkage Disequilibrium

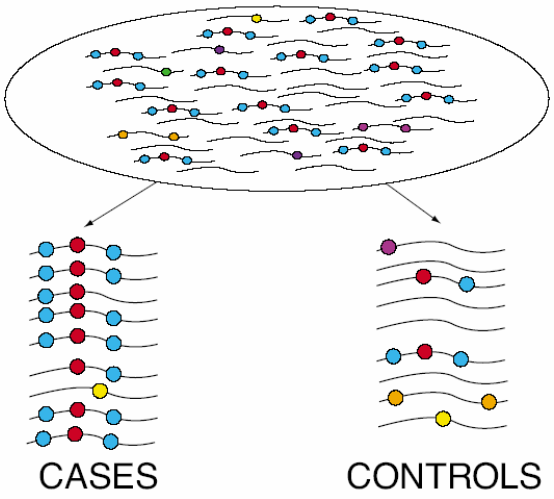
- LD is the non-random association of alleles at different sites



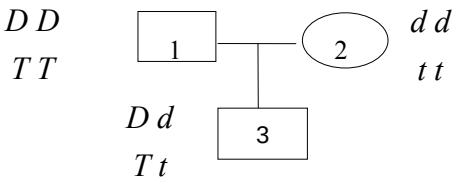
- Genetic recombination breaks down LD



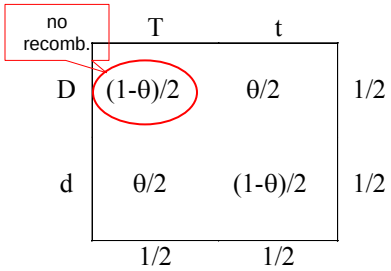
Linkage Disequilibrium in Gene Mapping



Two loci: linkage and recombination



- Son 3 produces sperm with *D-T, D-t, d-T* or *d-t* in proportions:



Two loci: linkage and recombination - II

- Son produces sperm with *DT, Dt, dT* or *dt* in proportions:

	T	t	
D	(1-θ)/2	θ/2	1/2
d	θ/2	(1-θ)/2	1/2
	1/2	1/2	

θ = 1/2 : independent assortment (*cf* Mendel) **unlinked** loci

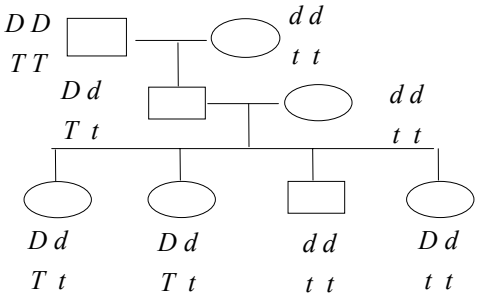
θ < 1/2 : **linked** loci

θ ≈ 0 : **tightly linked** loci

Note: θ > 1/2 is never observed

If the loci are linked, then *D-T* and *d-t* are *parental*, and *D-t* and *d-T* are *recombinant* haplotypes.

Two loci: estimation of recombination fractions

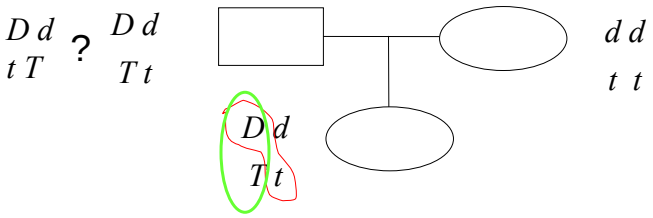


Recombination only discernible in the father. Here $\hat{\theta}$ = 1/4 (why?)

This is called the ***phase-known double backcross*** **pedigree.**

Two loci: phase

- Suppose we have data on two linked loci as follows:



- Was the daughter's *D-T* from her father a parental or recombinant combination?
  - This is the problem of **phase**: did father get *D-T* from one parent and *d-t* from the other? If so, then the daughter's paternally derived haplotype is parental.
  - If father got *D-t* from one parent and *d-T* from the other, these would be parental, and daughter's paternally derived haplotype would be recombinant.

Two loci: dealing with phase

- Phase is usually regarded as unknown genetic information, specifically, in parental origin of alleles at heterozygous loci.
- Sometimes it can be inferred with certainty from genotype data on parents.
- Often it can be inferred with high probability from genotype data on several children.
- In general genotype data on relatives helps, but does not necessarily determine phase.
- In practice, probabilities must be calculated under all phases compatible with the observed data, and added together. The need to do so is the main reason linkage analysis is computationally intensive, especially with multilocus analyses.

Two loci: founder probabilities

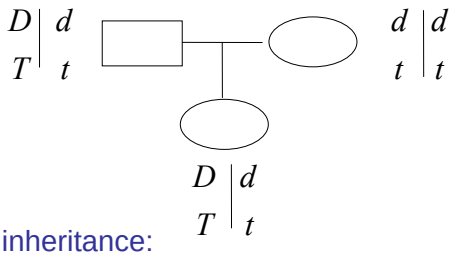
- Two-locus founder probabilities are typically calculated assuming **linkage equilibrium**, i.e. independence of genotypes across loci.
- If *D* and *d* have frequencies .01 and .99 at one locus, and *T* and *t* have frequencies .25 and .75 at a second, linked locus, this assumption means that *DT, Dt, dT* and *dt* have frequencies .01 x .25, .01 x .75, .99 x .25 and .99 x .75 respectively. Together with Hardy-Weinberg, this implies that

$$\begin{array}{c} Dd \\ Tt \end{array} \quad \square$$

pr(*DdTt*) = (2 x .01 x .99) x (2 x .25 x .75)
= 2 x (.01 x .25) x (.99 x .75) + 2 x (.01 x .75) x (.99 x .25).

- This last expression adds haplotype pair probabilities.

Two loci: transmission probabilities



- Haplotype inheritance:
  - Initially, this must be done with haplotypes, so that account can be taken of recombination.
  - Then terms like that below are summed over possible phases.
    - Here only the father can exhibit recombination: mother is **uninformative**.

pr(kid *DT/dt* | pop *DT/dt* & mom *dt/dt*)
= pr(kid *DT* | pop *DT/dt*) x pr(kid dt | mom *dt/dt*)
= (1-θ)/2 x 1.

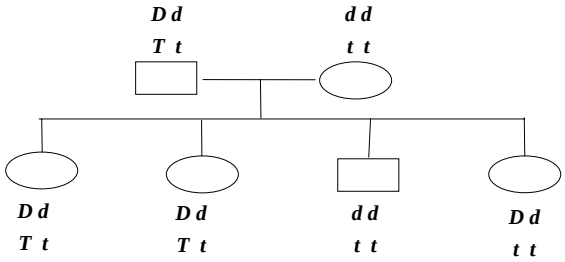
Two Loci: Penetrance

- In all standard linkage programs, different parts of phenotype are conditionally independent given all genotypes, and two-loci penetrances split into products of one-locus penetrances.
- Assuming the penetrances for DD, Dd and dd given earlier, and that T,t are two alleles at a co-dominant marker locus.

$$\begin{aligned} & \Pr(\text{affected } \& \; Tt \mid DD, Tt) \\ = \; & \Pr(\text{affected} \mid DD, Tt) \times \Pr(Tt \mid DD, Tt) \\ = \; & 0.8 \times 1 \end{aligned}$$

Two loci: phase unknown double backcross

- We assume below pop is as likely to be *DT / dt* as *Dt / dT*.



$$\begin{aligned} & \Pr(\text{all data} \mid \theta) \\ = \; & \Pr(\text{parents' data} \mid \theta) \times \Pr(\text{kids' data} \mid \text{parents' data}, \theta) \\ = \; & \Pr(\text{parents' data}) \times \{[((1-\theta)/2)^3 \times \theta/2]/2 + [(\theta/2)^3 \times (1-\theta)/2]/2\} \end{aligned}$$

This is then maximised in θ, in this case numerically. Here $\hat{\theta} = 0.25$

Log (base 10) odds or LOD scores

- Suppose pr(data | θ) is the likelihood function of a recombination fraction θ generated by some 'data', and pr(data | 1/2) is the same likelihood when θ= 1/2.
- Statistical theory tells us that the ratio

$$L = pr(data \mid \theta^*) / pr(data \mid 1/2)$$

provides a basis for deciding whether $\theta = \theta^*$ rather than $\theta = 1/2$.

- This can equally well be done with $Log_{10}L$, i.e.

$$LOD(\theta^*) = Log_{10}\{pr(data \mid \theta^*) / pr(data \mid 1/2)\}$$

measures the relative strength of the data for $\theta = \theta^*$ rather than $\theta = 1/2$. Usually we write θ, not θ^* and calculate the function **LOD(θ).**

Facts about/interpretation of LOD scores

1. Positive LOD scores suggests stronger support for θ^* than for 1/2, negative LOD scores the reverse.
2. Higher LOD scores means stronger support, lower means the reverse.
3. LODs are additive across independent pedigrees, and under certain circumstances can be calculated sequentially.
4. For a single two-point linkage analysis, the threshold LOD $\approx$ 3 has become the de facto standard for "establishing linkage", i.e. rejecting the null hypothesis of no linkage.
5. When more than one locus or model is examined, the remark in 4 must be modified, sometimes dramatically.

Assumptions underpinning most 2-point human linkage analyses

- ***Founder Frequencies**:* Hardy-Weinberg, random mating at each locus. Linkage equilibrium across loci, **known** allele frequencies; founders independent.
- ***Transmission**:* Mendelian segregation, no mutation.
- ***Penetrance**:* single locus, no room for dependence on relatives' phenotypes or environment. **Known** (including phenocopy rate).
- ***Implicit**:* phenotype and genotype data **correct**, marker order and location correct
- ***Comment**:* Some analyses are *robust*, others can be *very sensitive* to violations of some of these assumptions. Non-standard linkage analyses can be developed.

Beyond two-point human linkage analysis

- The real challenge is multipoint linkage analysis, but going there would take more time than we have today.
- Next in importance is dealing with two-locus penetrances.

Nonparametric Methods for Linkage Analysis

Why Nonparametric Linkage Analysis?

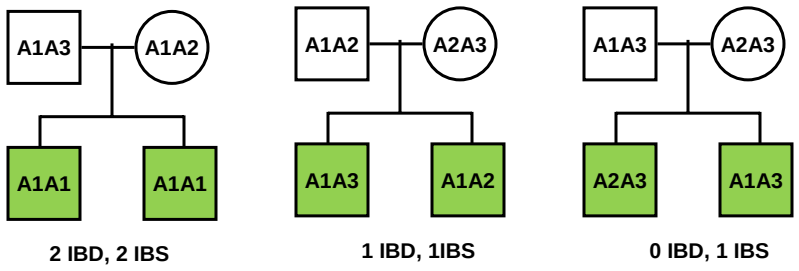
- Disadvantages of the LOD-score method
  - What if the model (allele frequency, penetrance etc.) is incorrect?
  - Works well for single-locus and high-penetrance diseases, but many diseases are multifactorial
  - Data on large pedigrees are rare
- Affected sib-pair analysis
  - Nonparametric method – no genetic model
  - Data: Genotypes of affected pair of siblings and their parents

Affected Sib-Pair Analysis

If the given genetic marker is linked to the disease locus, affected siblings share more identity-by-descent (IBD) alleles at the marker locus than expected. (i.e., affected siblings are likely to share the segment of the chromosome containing the disease locus.)

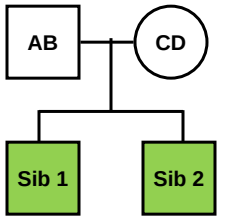
- IBD (identity by descent) : Alleles are demonstrably copies of the same ancestral allele.
- IBS (identity by state) : Alleles look the same, but they are not derived from a known common ancestor

IBD and IBS



When There is No Linkage

- Under the null hypothesis of no linkage between the marker locus and the disease locus (random segregation), the probabilities of a sib-pair sharing alleles IBD are given as:
  - P(0 IBD) = (1-0.5)*(1-0.5) = 0.25
  - P(1 IBD) = 0.5*(1-0.5) + (1-0.5)*0.5 = 0.5
  - P(2 IBD) = 0.5*0.5 = 0.25
  - Expected number of IBD alleles
    = 0*0.25+1*0.5+2*0.25 = 1

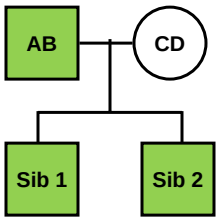


AC

AC (2 IBD)
AD (1 IBD)
BC (1 IBD)
BD (0 IBD)

When There is Linkage

- Dominant disease
  - Pairs of siblings share one or two disease-related alleles
  - Expected number of IBDs > 1
    - This can be detected in the linkage analysis



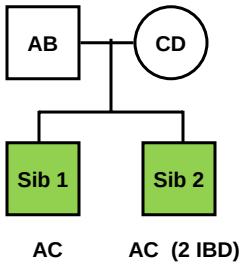
AC

AC (2 IBD)
AD (1 IBD)

Under the dominant disease model where A is linked to the disease locus, given Sib1 = (A,C), the only possible allele combinations for Sib2 are (A,C) or (A,D)

When There is Linkage

- Recessive disease
  - Pairs of siblings share both disease-related alleles
  - Expected number of IBDs > 1
    - This can be detected in the linkage analysis
  - Parents are carriers



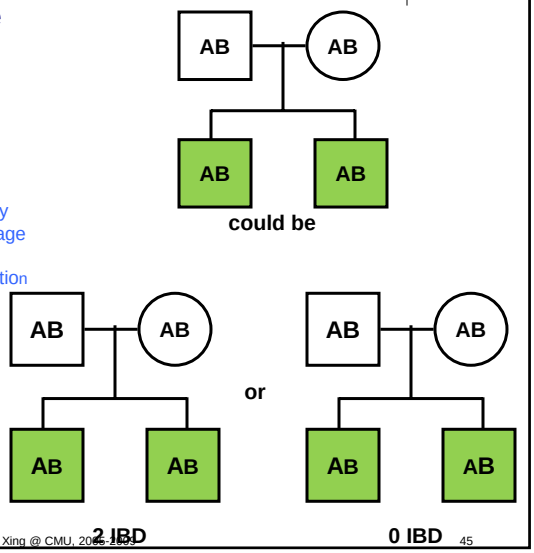
Under the recessive disease model, given Sib1 = (A,C), the only possible allele combination for Sib2 is (A,C)

Affected Sib-Pair Analysis

- Data : genotypes of the pair of affected siblings and their parents
- Two different approaches for hypothesis testing
  - Compare the expected and observed frequency of siblings with 0, 1, and 2 IBDs under the null hypothesis H0 = (0.25, 0.5, 0.25)
    - χ^2 test with 2 degrees of freedom
  - Compare the expected and observed average number of IBDs under the null hypothesis H0 = 1 ("mean test")
    - χ^2 test with 1 degree of freedom
- Note: we do not make assumptions on the genetics of disease (dominant or recessive)

Ambiguity in IBDs

- Highly polymorphic markers are preferable in order to determine IBDs
- Sometimes it is not possible to determine IBDs unequivocally
  - The mean test has been extended by estimating the ibd score as the average of the ibd scores under the various possible parental genotype combination



Other Nonparametric Methods

- Affected pedigree member method
  - Extends the affected sib-pair analysis to other relationships, such as pairs of affected people in a complex pedigree
  - Uses IBS instead of IBDs - It does not use all of the available information on linkage
- Extensions to multiple-marker loci
  - Multiple markers are more informative in determining IBDs accurately
  - Assume that the marker loci in the multiple marker are in linkage disequilibrium

Acknowledgements

Melanie Bahlo, WEHI
Hongyu Zhao, Yale
Karl Broman, Johns Hopkins
Nusrat Rabbee, UCB

References

www.netspace.org/MendelWeb

HLK Whitehouse: **Towards an Understanding of the Mechanism of Heredity,** 3rd ed. Arnold 1973

Kenneth Lange: **Mathematical and statistical methods for genetic analysis,** Springer 1997

Elizabeth A Thompson: **Statistical inference from genetic data on pedigrees,** CBMS, IMS, 2000**.**

Jurg Ott : **Analysis of human genetic linkage**, 3rd edn Johns Hopkins University Press 1999

JD Terwilliger & J Ott : **Handbook of human genetic linkage**, Johns Hopkins University Press 1994

