Introduction to Linkage Analysis

Jurg Ott, Ph.D.
Rockefeller University, New York
ott@rockefeller.edu

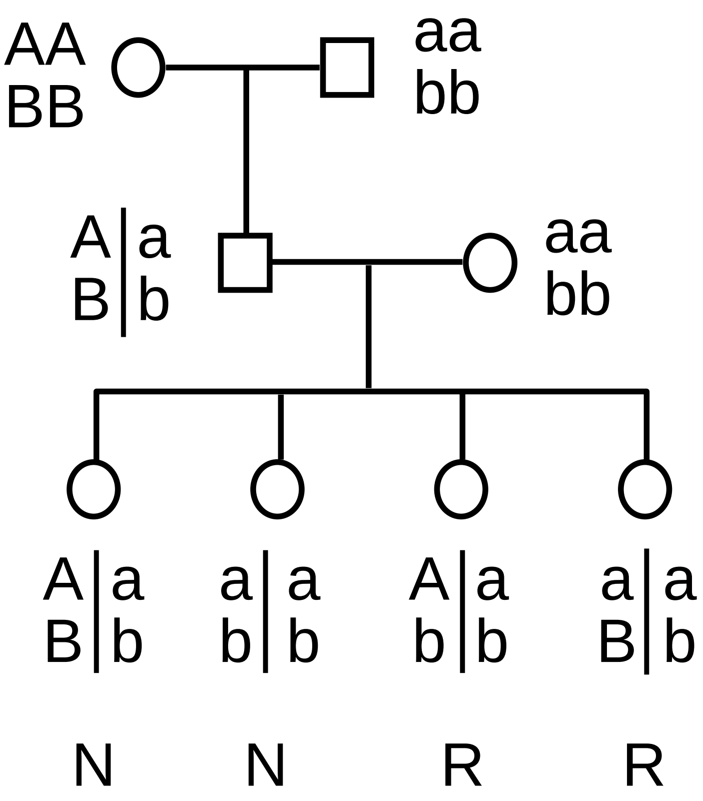
Genetic Markers

- Loci that are polymorphic (two or more alleles), inherited in a mendelian manner. "Sign posts" in gene mapping for localizing new genes. *Definition*: The most common allele has frequency <0.95.
- DNA polymorphisms: Stable differences in DNA sequence (chromosomes). 400 up to 100,000's of markers (SNPs) created.

Co-inheritance of Disease and Marker Genes in Families

- Generally, disease and genetic markers are inherited independently (Mendel’s second law).
- For a marker in close proximity to a disease locus, their genes travel together in family pedigrees (*genetic linkage*), only occasionally interrupted by co-called crossing-over.

Simple Assumed Example



AA
BB

aa
bb

A|a
B|b

aa
bb

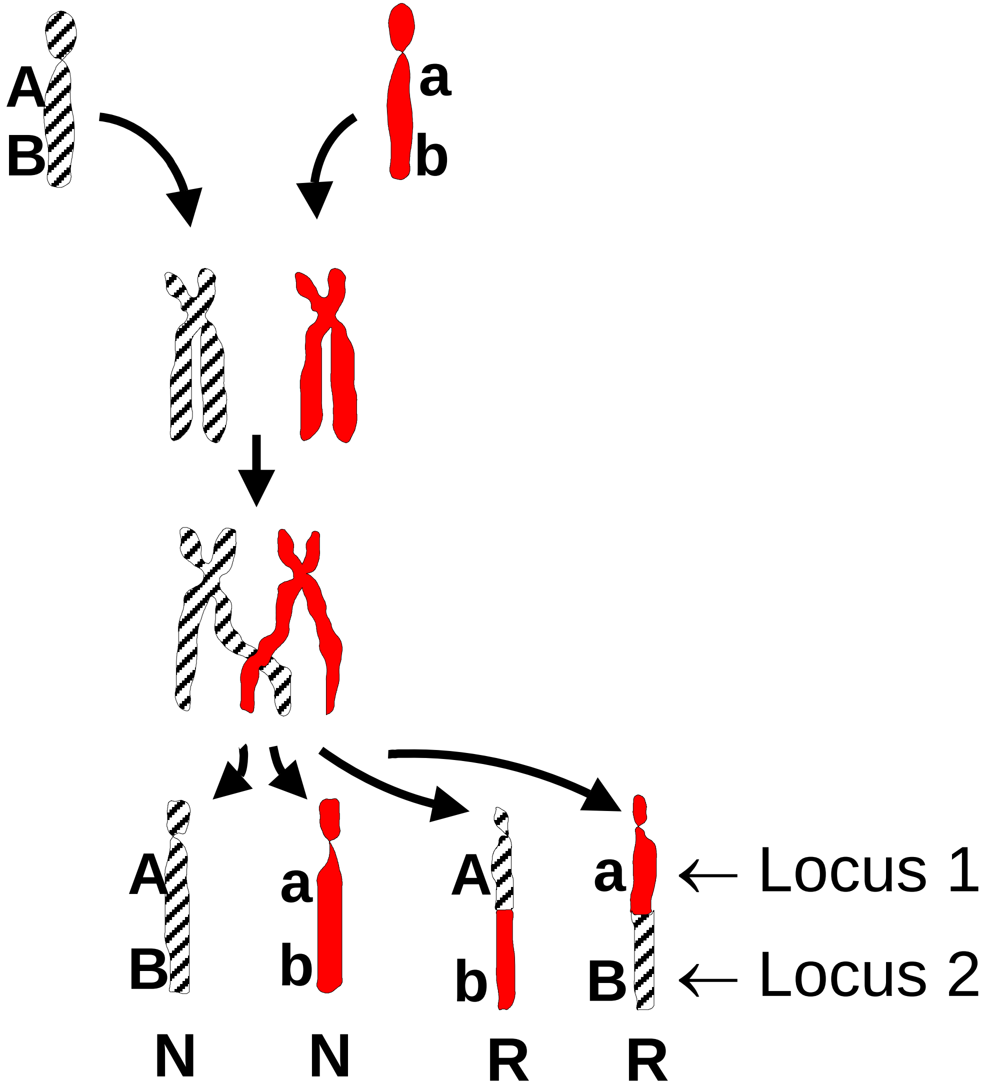
A|a
B|b

a|a
b|b

A|a
b|b

a|a
B|b

N N R R



A
B

a
b

A
B

a
b

A
b

a
B

← Locus 1

← Locus 2

N N R R

Definitions

- *Recombination* — alleles at different loci have different grandparental origin. Recombination fraction θ = proportion of recombinants = probability for recombination to occur.
- *Crossovers* cannot be observed directly, only their phenotypic expression as recombinations
- Multiple crossover points on a gamete:
  - Odd number → recombination
  - Even number → no recombination

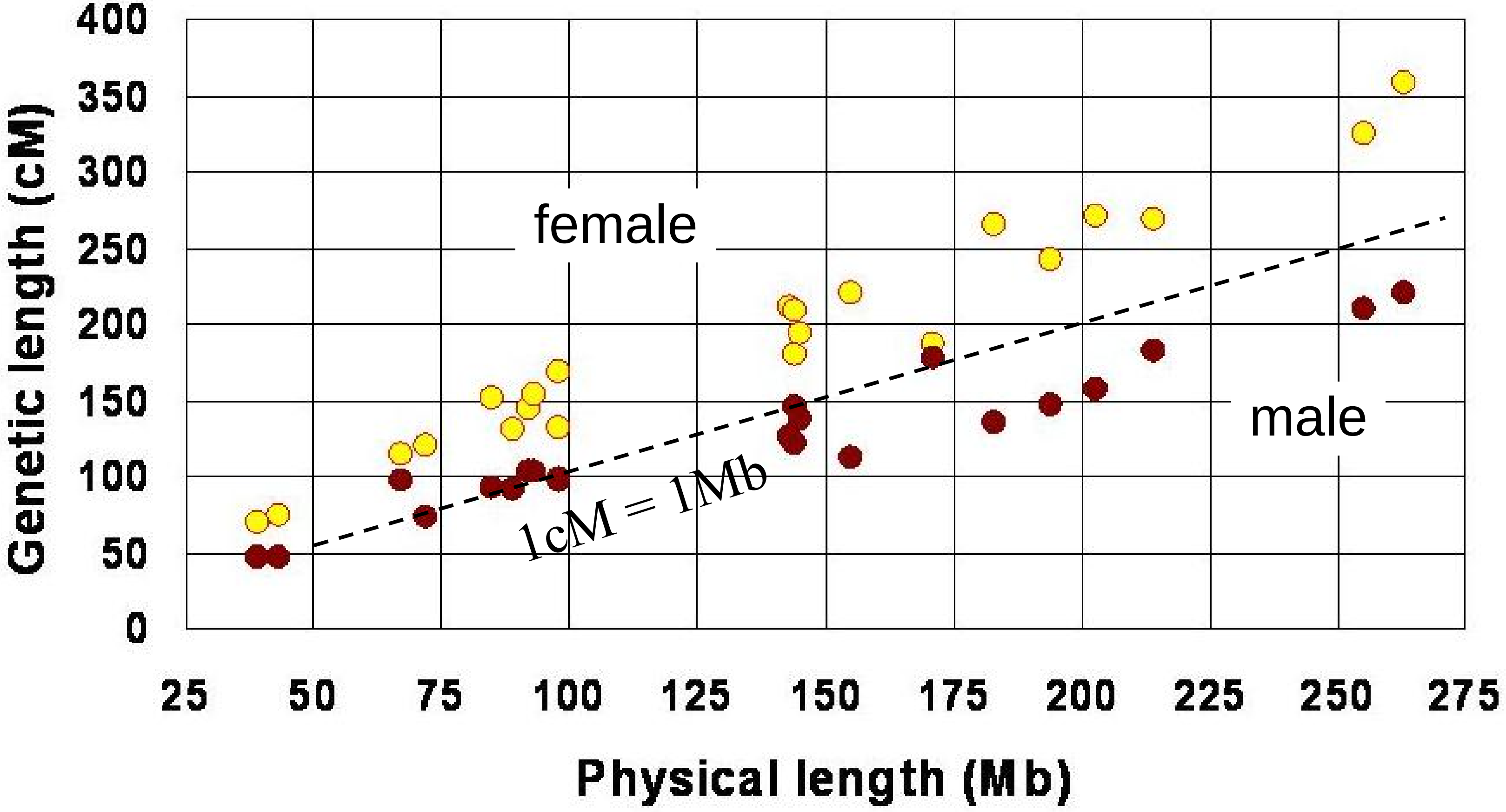
Genetic Distance

- Crossovers occur randomly on a chromosome (not necessarily uniformly distributed).
- Genetic distance (map distance) between two points = expected number of crossovers between them on a gamete. Unit of measurement = Morgan (M) = 100 cM.
- Chiasma interference, no chromatid interference
- Different crossover frequencies in females and males: Genetic distances are sex (age?) specific.
- Example: $\theta = 0.03 \rightarrow x = 0.03 \text{ M} = 3 \text{ cM}$

Physical/Genetic Lengths of Human Autosomes

Morton (1991) *PNAS* **88**, 7474 (physical lengths)

Dib et al (1996) *Nature* **380**, 152 (genetic lengths)



Recombination Fraction and Age

Haldane and Crew (1925): *Offspring of phase-known matings in poultry*. 5 cocks, doubly heterozygous *BS/bs* for two sex-linked mendelian loci, mated with *bs* hens. The four possible offspring types all distinguishable phenotypically. Total of 648 chicks.

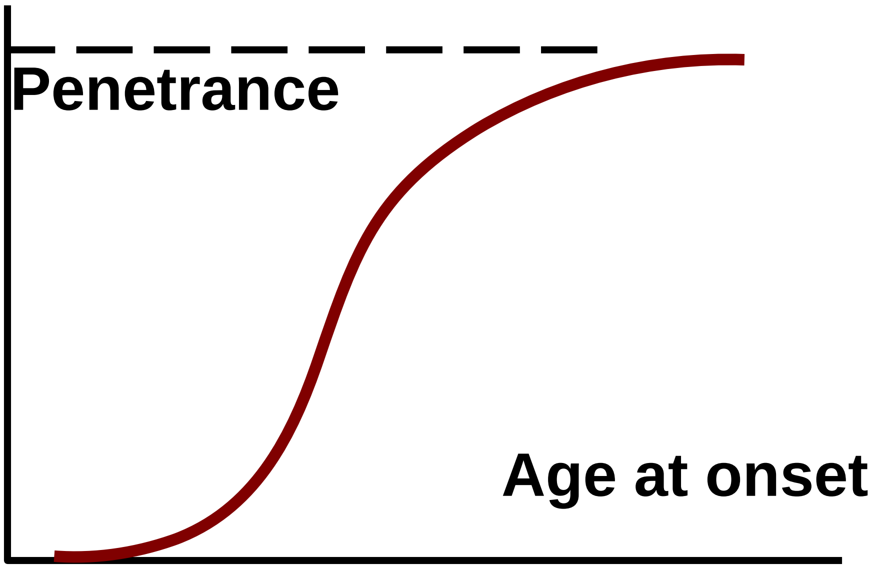
Breeding year	1	2	3
Recomb. fraction	0.229	0.369	0.476

Recombination fraction progressively larger with advancing age of the cocks.

How Do We Localize Genes for Heritable Diseases?

- Collect families with affected individuals
- Draw blood and extract DNA
- Determine genotypes for ~400 – 100,000s of markers along genome.
- Track inheritance of marker alleles and trait in pedigrees: Linkage analysis.

Problems

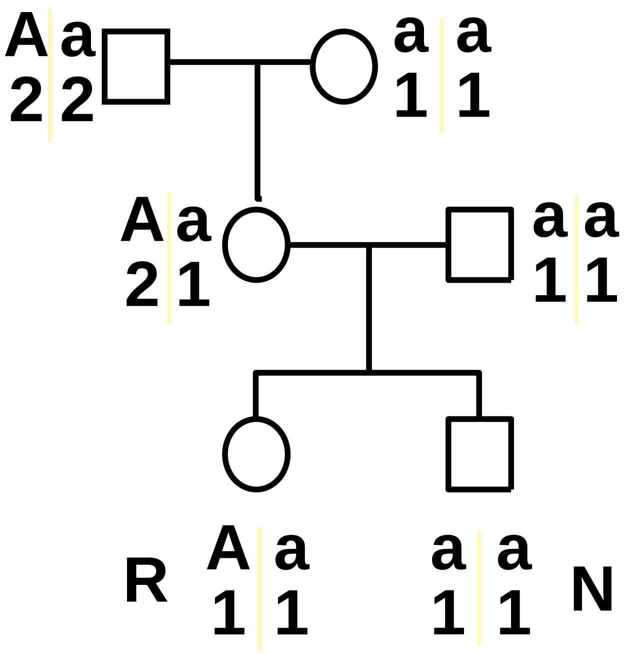


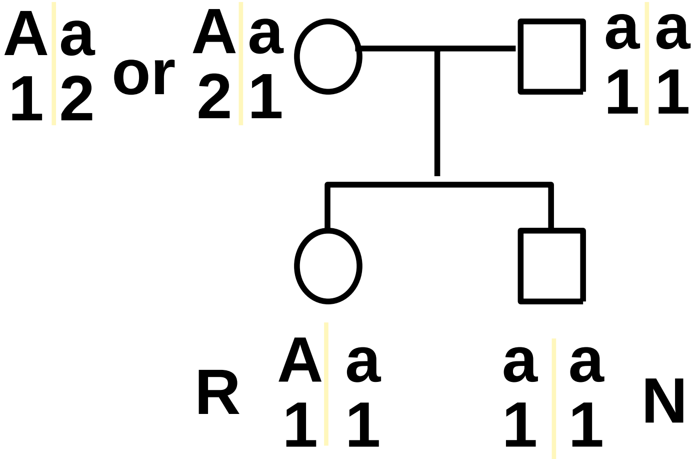
- Penetrance incomplete
- Phenocopies
- Parents unavailable
- Individuals not consenting to study

- Cannot generally count recombinants and nonrecomb.
- **Solution**: Estimate recombination fraction by maximum likelihood method

Likelihood

- Likelihood = probability of data. Depends on unknown parameters: $L(\theta) = P(\text{data}; \theta)$
- Phase known double back-cross: $L(\theta) = \theta^k(1 - \theta)^{n-k}$ k = number of recombinants, n = total number of meioses.
- Phase unknown double back-cross: $L(\theta) = \theta^k(1 - \theta)^{n-k} + \theta^{n-k}(1 - \theta)^k$





Lod Score

- *Lod score* = scaled log likelihood ratio, $Z(\theta) = \log_{10}[L(\theta)/L(\theta = ½)]$, θ = trial value for recombination fraction
- With linkage, lod score tends to increase. Maximum lod score $\geq 3 \rightarrow$ significant linkage.

Lod score calculated by hand

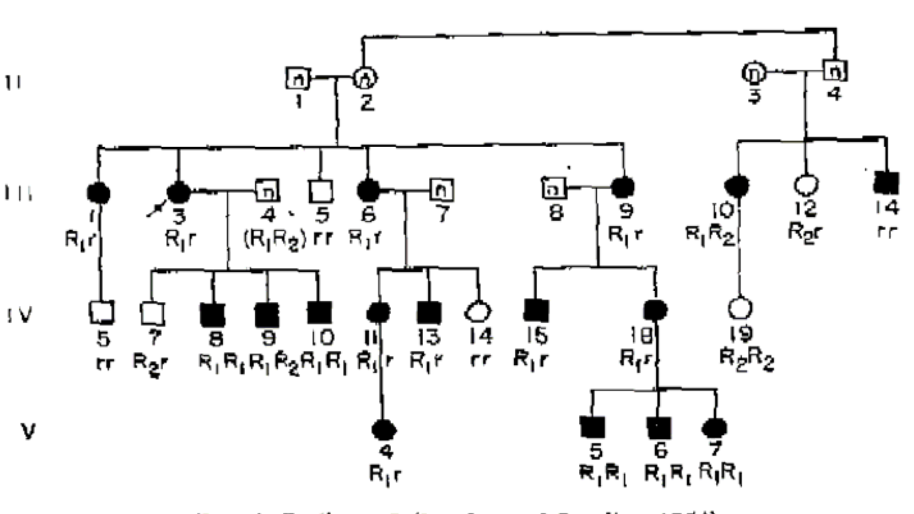


FIG. 5. Pedigree 5 (Lawler and Sandler, 1954)

Morton (1956) *Am J Hum Genet* **8**, 80-96

c = recombination fraction

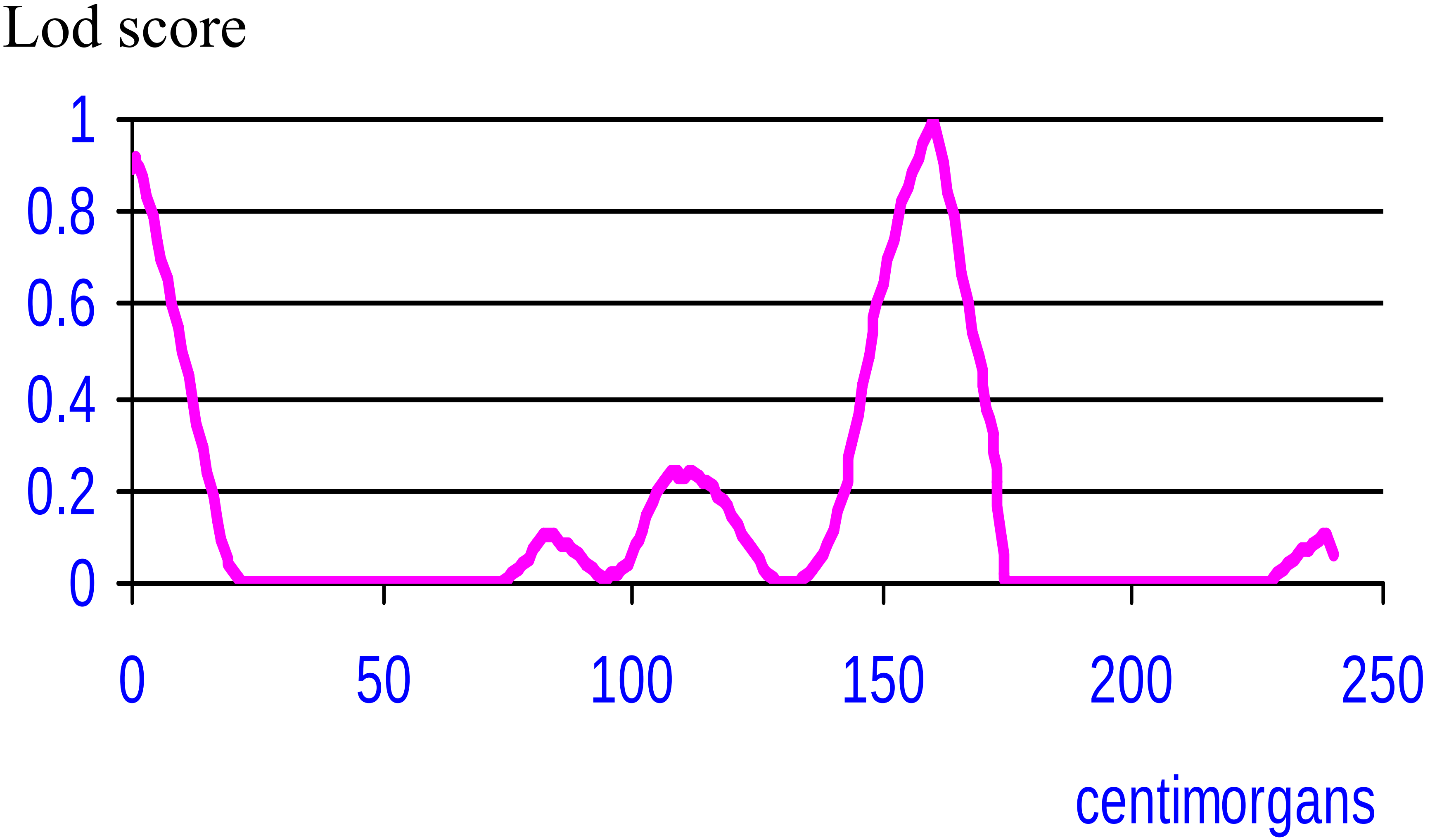
$$Z = \log_{10} 220/39168\ \{810c(1\text{-}c)^{19} + 324c(1\text{-}c)^{18} + 180c(1\text{-}c)^{17} + 72c(1\text{-}c)^{16} + 90c^3(1\text{-}c)^{17} + 72c^3(1\text{-}c)^{16} + 40c^3(1\text{-}c)^{15} + 24c^3(1\text{-}c)^{14} + 90c^4(1\text{-}c)^{15} + 20c^4(1\text{-}c)^{13} + 90c^5(1\text{-}c)^{15} + 432c^5(1\text{-}c)^{14} + 20c^5(1\text{-}c)^{13} + 104c^5(1\text{-}c)^{12} + 1800c^6(1\text{-}c)^{14} + 558c^6(1\text{-}c)^{13} + 440c^6(1\text{-}c)^{12} + 176c^6(1\text{-}c)^{11} + 90c^7(1\text{-}c)^{13} + 324c^7(1\text{-}c)^{12} + 120c^7(1\text{-}c)^{10} + 360c^8(1\text{-}c)^{12} + 378c^8(1\text{-}c)^{11} + 80c^8(1\text{-}c)^{10} + 76c^8(1\text{-}c)^{9} + 4c^8(1\text{-}c)^{4} + 180c^9(1\text{-}c)^{11} + 522c^9(1\text{-}c)^{10} + 80c^9(1\text{-}c)^{9} + 100c^9(1\text{-}c)^{8} + 10c^9(1\text{-}c)^{3} + 180c^{10}(1\text{-}c)^{10} + 846c^{10}(1\text{-}c)^{9} + 40c^{10}(1\text{-}c)^{8} + 216c^{10}(1\text{-}c)^{7} + 18c^{10}(1\text{-}c)^{4} + 4c^{10}(1\text{-}c)^{2} + 1170c^{11}(1\text{-}c)^{9} + 378c^{11}(1\text{-}c)^{8} + 260c^{11}(1\text{-}c)^{7} + 72c^{11}(1\text{-}c)^{6} + 45c^{11}(1\text{-}c)^{3} + 180c^{12}(1\text{-}c)^{8} + 396c^{12}(1\text{-}c)^{7} + 40c^{12}(1\text{-}c)^{5} + 18c^{12}(1\text{-}c)^{2} + 270c^{13}(1\text{-}c)^{7} + 234c^{13}(1\text{-}c)^{6} + 40c^{13}(1\text{-}c)^{5} + 52c^{13}(1\text{-}c)^{4} + 180c^{14}(1\text{-}c)^{6} + 108c^{14}(1\text{-}c)^{5} + 80c^{14}(1\text{-}c)^{4} + 16c^{14}(1\text{-}c)^{3} + 90c^{15}(1\text{-}c)^{5} + 162c^{15}(1\text{-}c)^{4} + 20c^{15}(1\text{-}c)^{3} + 180c^{16}(1\text{-}c)^{4} + 72c^{16}(1\text{-}c)^{3} + 90c^{17}(1\text{-}c)^{3}\}$$

Computer Programs for Linkage Analysis

- *LIPED* (Ott 1974, 1976), 2-point analysis
- *PAP* (Hasstedt 1982)
- *LINKAGE* (Lathrop et al. 1986). *FastLINK*
- *Mapmaker* (Lander et al. 1987)
- *CRI-MAP* (Phil Green)
- *Mendel* (Lange et al. 1988)
- *Vitesse* (O'Connell and Weeks 1995)
- *Genehunter* (Kruglyak et al. 1996, Kong and Cox 1997); *Aspex* (Risch); *Loki* (Heath 1997); *SAGE* (Elston)
- *Allegro* (Gudbjartsson et al. 2000)
- *Merlin* (Abecassis)
- *Simwalk2* (D. Weeks)

Chrom. 5 scan with bipolar ASPs

(Ginns et al. [1996] *Nat Genet* **12**, 431)



Sequential likelihood ratio test of $\theta = 0.5$ vs. $\theta = 0.2$

Morton (1955) *Am J Hum Genet* **7**, 277-318

- Accept linkage when combined lod score, $Z(0.2) \geq 3$
- Reject linkage when $Z(0.2) < -2$ ("excluded" θ values)
- Otherwise, continue sampling

Locus Heterogeneity

Morton (1956) *Am J Hum Genet* **8**, 80

- Each family has its own different recombination fraction.

  θ_i = recombination in *i*-th family
- Easy likelihood ratio test:

$$\chi^2 = 4.6 \times \left(\sum_i Z_{\max,i} - Z_{\max,total} \right)$$

- *n* – 1 df, *n* = number of families

Locus Heterogeneity

Smith (1963) *Ann Hum Genet* **27**, 175

- Realistically, only 2 recombination fractions, $\theta < 0.50$ and $\theta_0 = 0.50$ → some families with linkage and others without linkage. Mixture of these two family types.
- Solution: Estimate 2 parameters:

  α = proportion of families with linkage

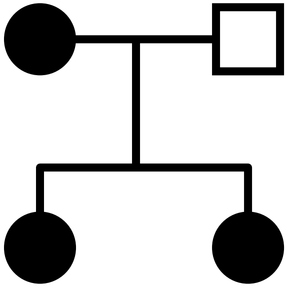
  θ = recombination fraction in "linked" families
- Computer program: HOMOG

Example: Osteogenesis Imperfecta and GC Blood Types

(Vogel and Motulsky 1986)

	Female rec. fraction, θ_f				
θ_m	0.05	0.10	0.20	0.30	0.50
0.50	0.28	0.70	1.19	1.01	0
0.30	2.74	3.98	4.73	4.68	3.42
0.20	4.30	5.62	6.42	6.37	5.08
0.10	5.50	6.84	7.64	7.59	6.26
0.05	5.74	7.08	7.88	7.83	6.48

Affected sibpairs



T|t
1|2

or

T|t
2|1

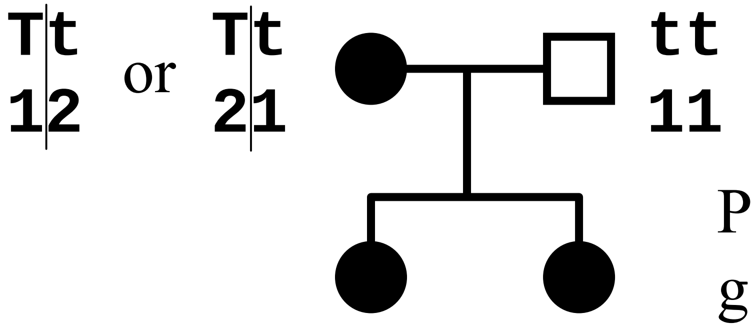
tt
11

c = recombination fraction

Probability of offspring genotypes:

Tt 11	Tt 11	$[c^2 + (1 - c)^2]/2$
Tt 21	Tt 21	$[c^2 + (1 - c)^2]/2$
Tt 11	Tt 21	$c(1 - c)$
Tt 21	Tt 11	$c(1 - c)$

Affected sibpairs



c = recombination fraction

Probability of offspring genotypes:

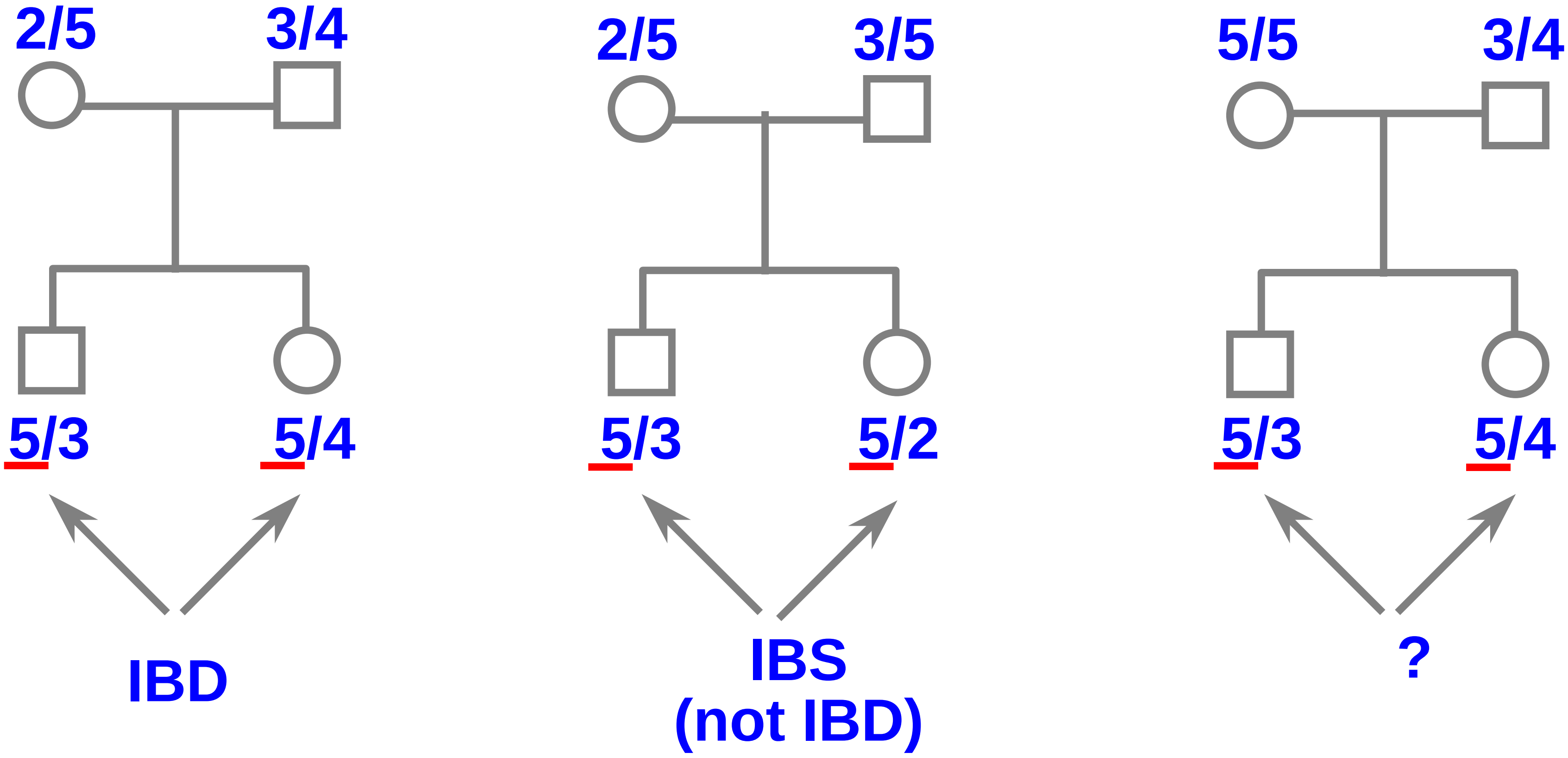
Tt 11	Tt 11	$[c^2 + (1 – c)^2]/2$	Share allele from given parent
Tt 21	Tt 21	$[c^2 + (1 – c)^2]/2$	
Tt 11	Tt 21	$c(1 – c)$	Do not share allele from given parent
Tt 21	Tt 11	$c(1 – c)$	

Allele sharing

- **Per parent**. Proportion of parents transmitting same allele, $S = c^2 + (1 - c)^2$, $\frac{1}{2} \leq S \leq 1$. H_0: $S = \frac{1}{2}$.
- **Per sibship**. H_0: proportion of sibships sharing 0, 1, and 2 alleles = ½, ¼, ½, respectively.
- **Test** for $S > \frac{1}{2}$ carried out for any disease.
- Extension to **other relatives**: Whittemore statistic, implemented in *Genehunter*

Identity by descent (IBD)

Alleles shared IBD: Copies of ancestral allele



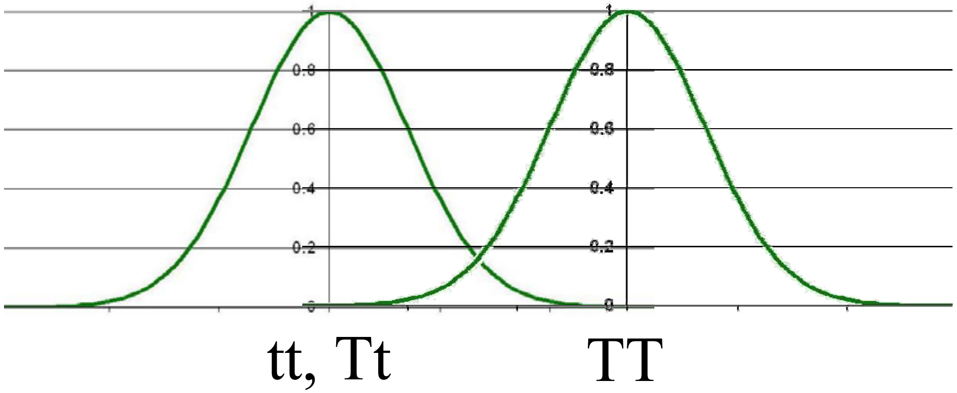
Equivalence with recessive inheritance

Knapp *et al* (1994) *Hum Hered* **44**, 44-51

- ASP analysis completely equivalent with lod score analysis under recessive inheritance, full penetrance, parents of unknown phenotype
- Elegantly allows for multiple affected offspring. No need for analysis of all pairs and complicated weighting schemes.

Quantitative phenotypes

- Mean depends on genotype

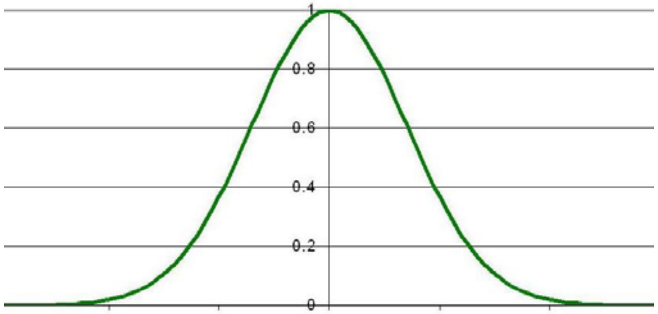


Dominant example: Only TT genotypes have elevated mean levels

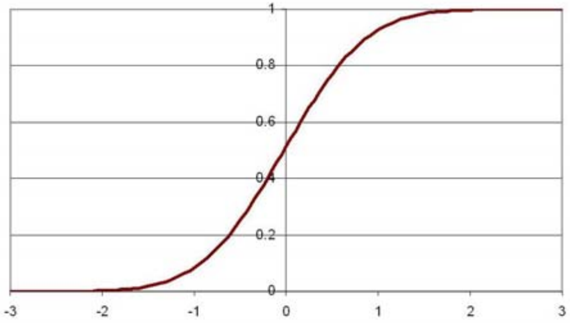
- Test whether means are different for different genotypes → ANOVA (association)
- Linkage analysis in families

Age of disease onset

- Assume normal distribution for a = onset age. Often, f(a) unknown.



- Use A = current age. P(affected by age A|risk), F(A) = P($a \leq A$ | at risk) → cumulative, sigmoid curve



- Implementation is complicated, particularly when penetrance is incomplete at high age.

Linkage between QTL and marker

Haseman & Elston (1972) *Behav Genet* **2**, 3-19

- Regress the square of the difference between sib-pair trait values on the estimated proportion of marker alleles that the sib pair shares IBD.
- Various extensions published