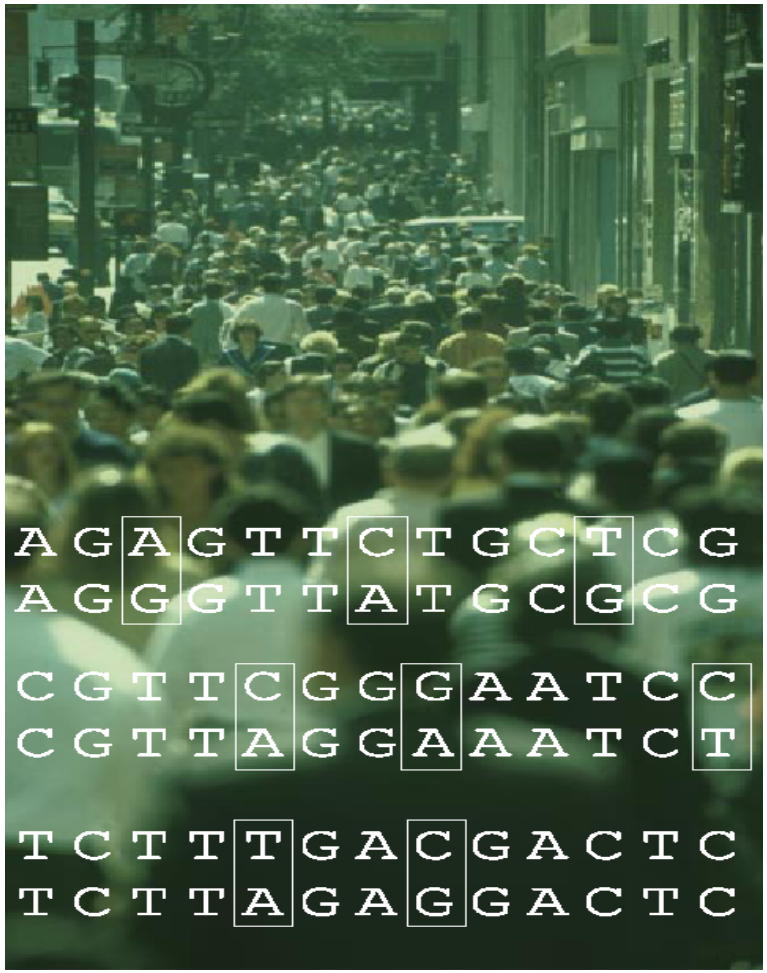
Linkage Analysis

02-710 Computational Genomics

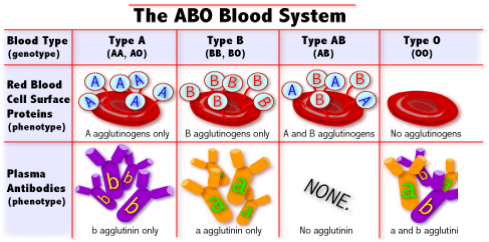
Seyoung Kim

Genome Polymorphisms

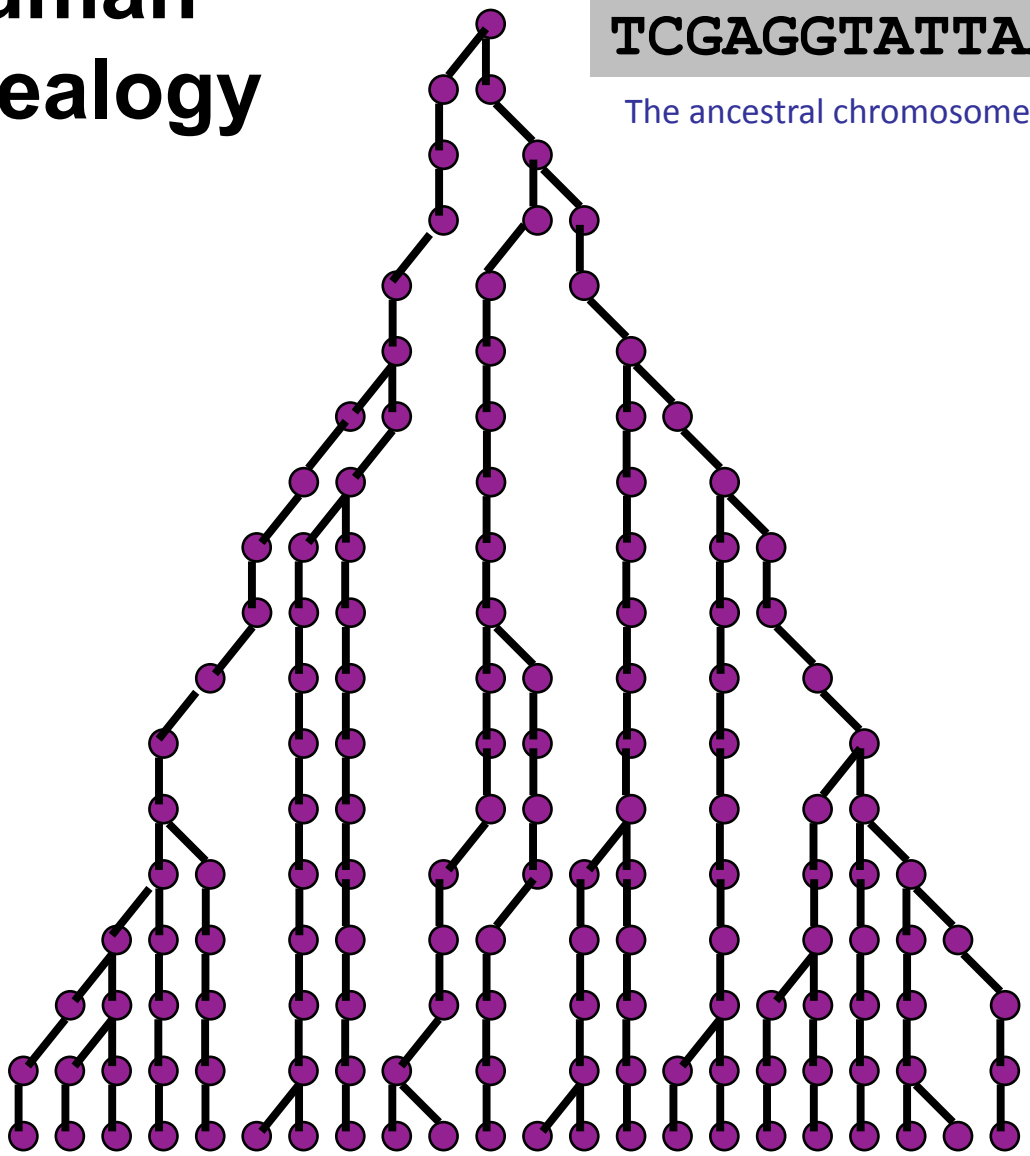
Genetic Variation



Phenotypic Variation



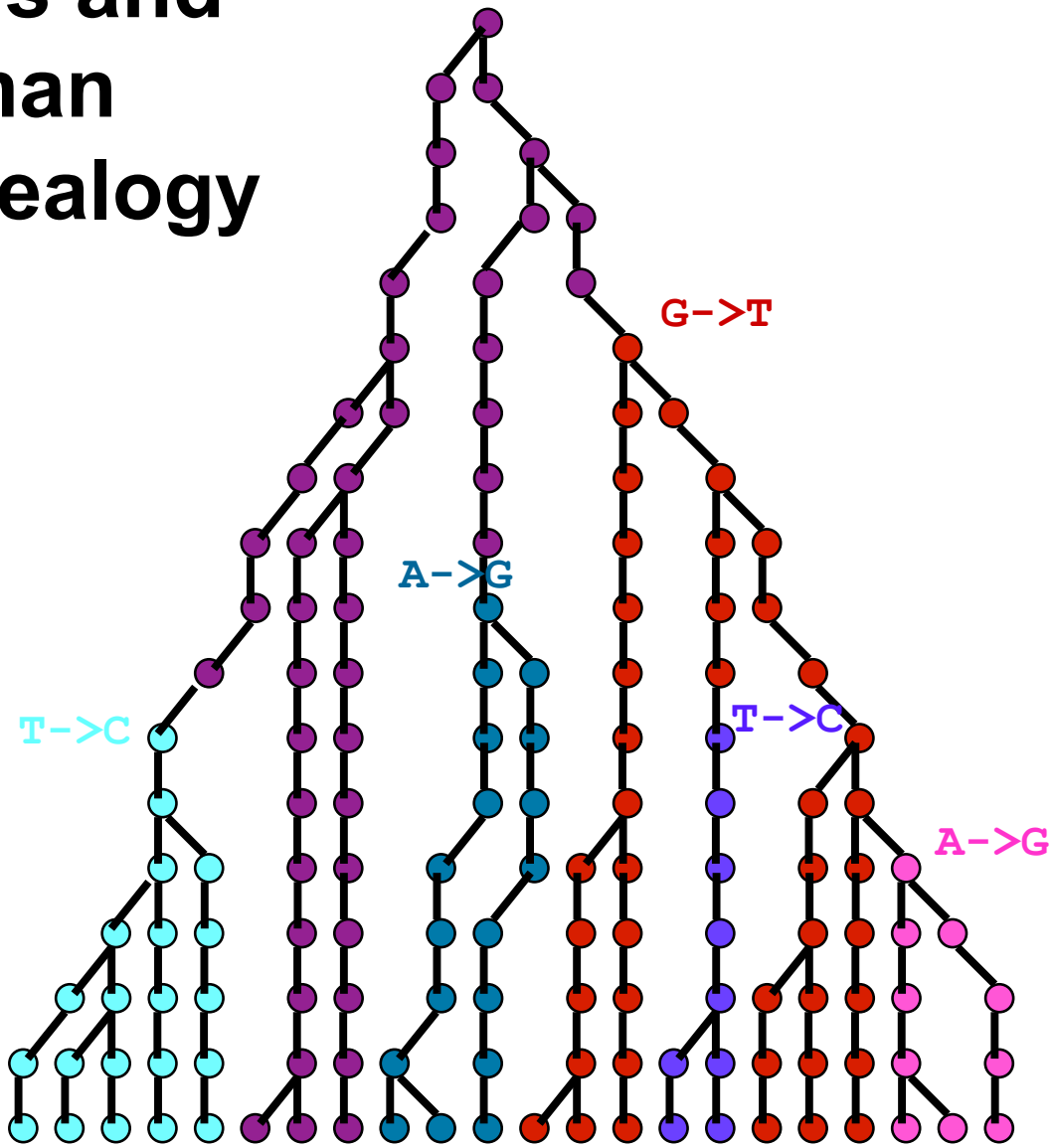
A Human Genealogy



TCGAGGTATTAAC

The ancestral chromosome

SNPs and Human Genealogy

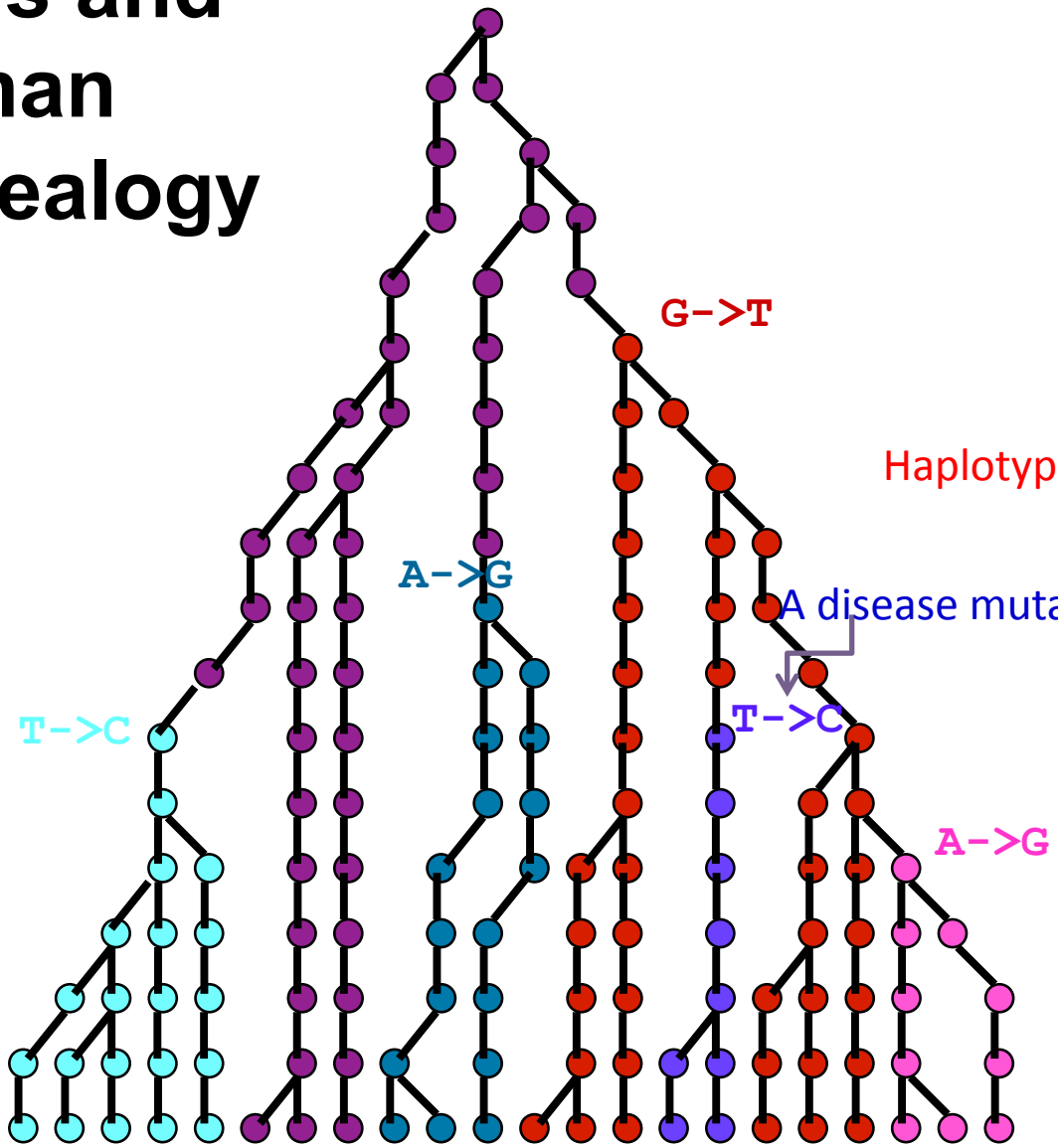


TCGAGGTATTAAC
TCTAGGTATTAAC
TCGAGGCATTAAC
TCTAGGTGTTAAC
TCGAGGTATTAGC
TCTAGGTATCAAC

* ** * *

The SNPs

SNPs and Human Genealogy



TCGAGGTATTAAC
TCTAGGTATTAAC
TCGAGGCATTAAC
TCTAGGTGTTAAC
TCGAGGTATTAGC
Haplotype TCTAGGTATCAAC

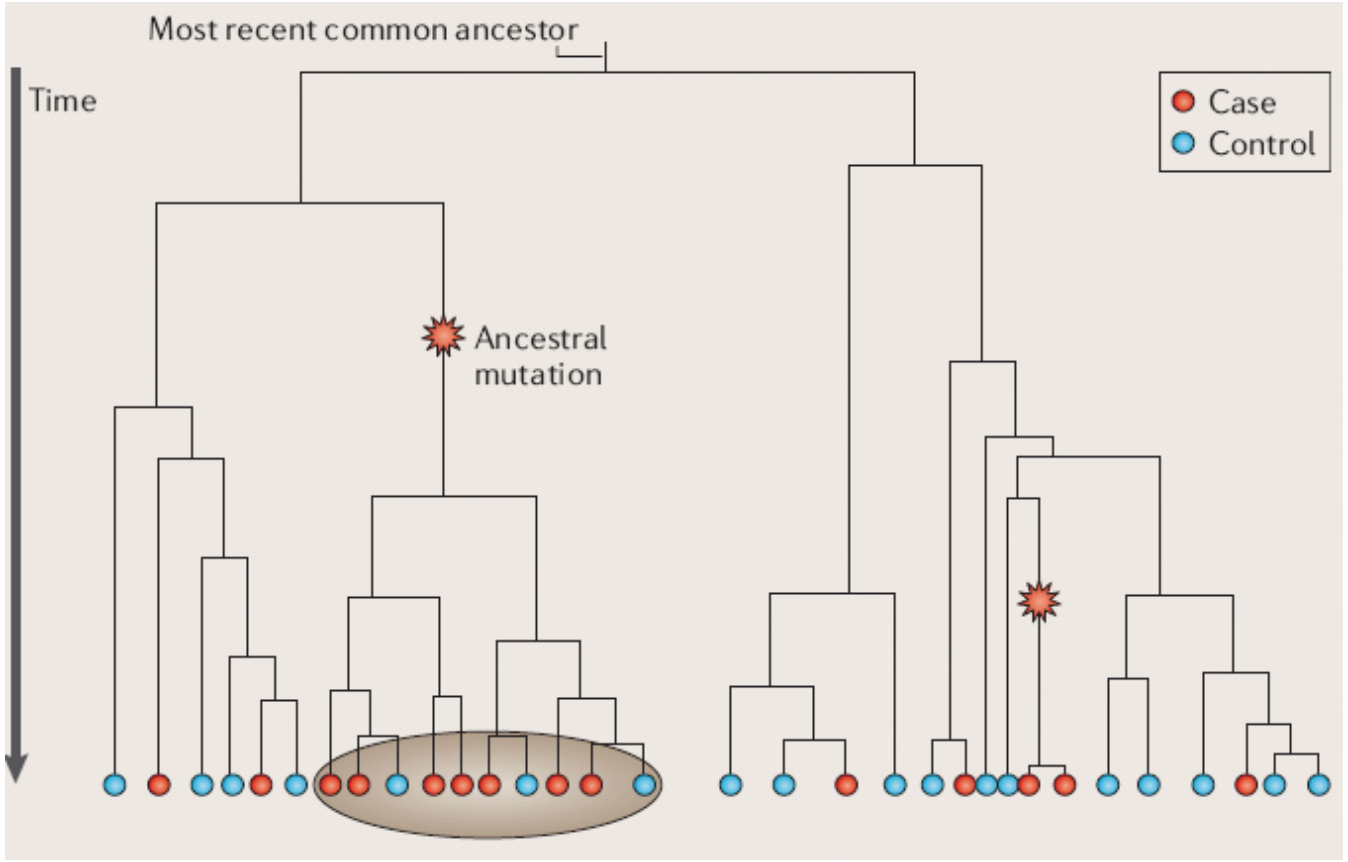
* ** * *

The SNPs

A disease mutation

Identifying Disease Loci

- All individuals are related if we go back far enough in the ancestry



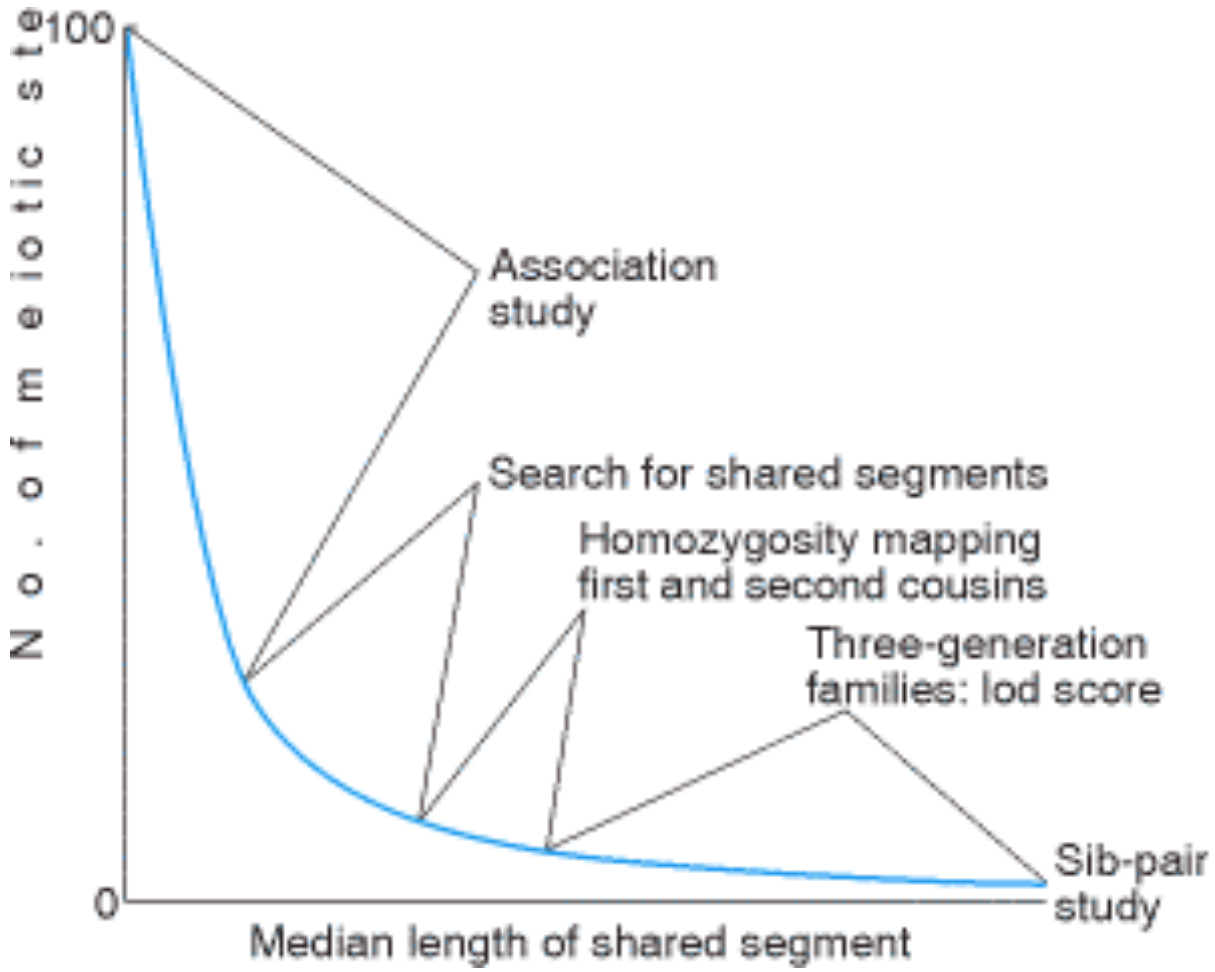
Balding, Nature Reviews Genetics, 2006

Overview

- How can we identify the genetic loci responsible for determining phenotypes?
  - Linkage analysis
    - Data are collected for family members
    - Difficult to collect data on a large number of families
    - Effective for rare diseases
    - Low resolution on the genomes due to only few recombinations
      - » a large region of linkage

  - Genome-wide association studies
    - Data are collected for unrelated individuals
    - Easier to find a large number of affected individuals
    - Effective for common diseases, compared to family-based method
    - Relatively high resolution for pinpointing the locus linked to the phenotype

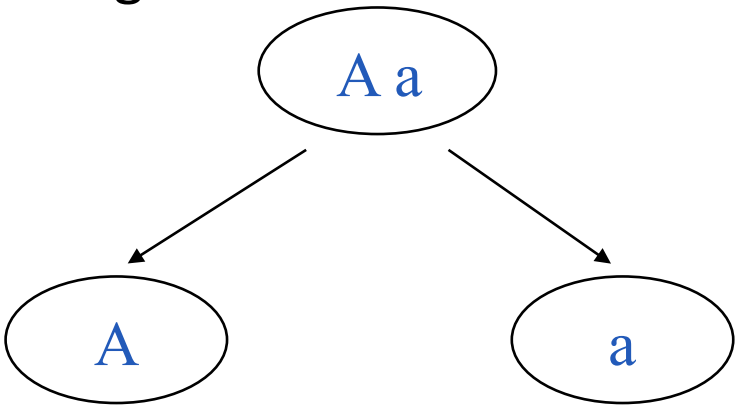
Linkage Analysis vs. Association Analysis



Strachan & Read, Human Molecular Genetics, 2001

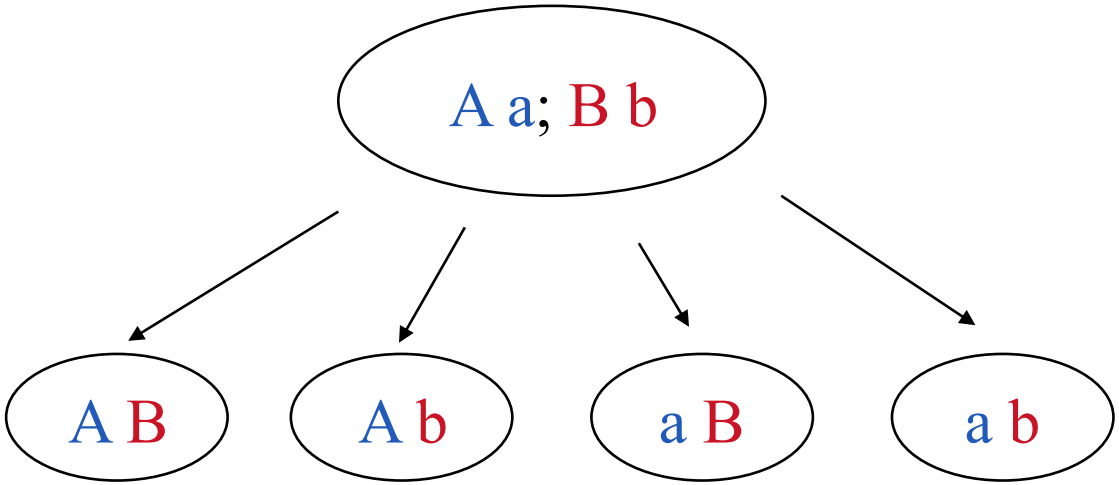
Mendel's two laws

- 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 genes segregate simultaneously, they do so independently.



“Exceptions” to Mendel’s Second Law

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

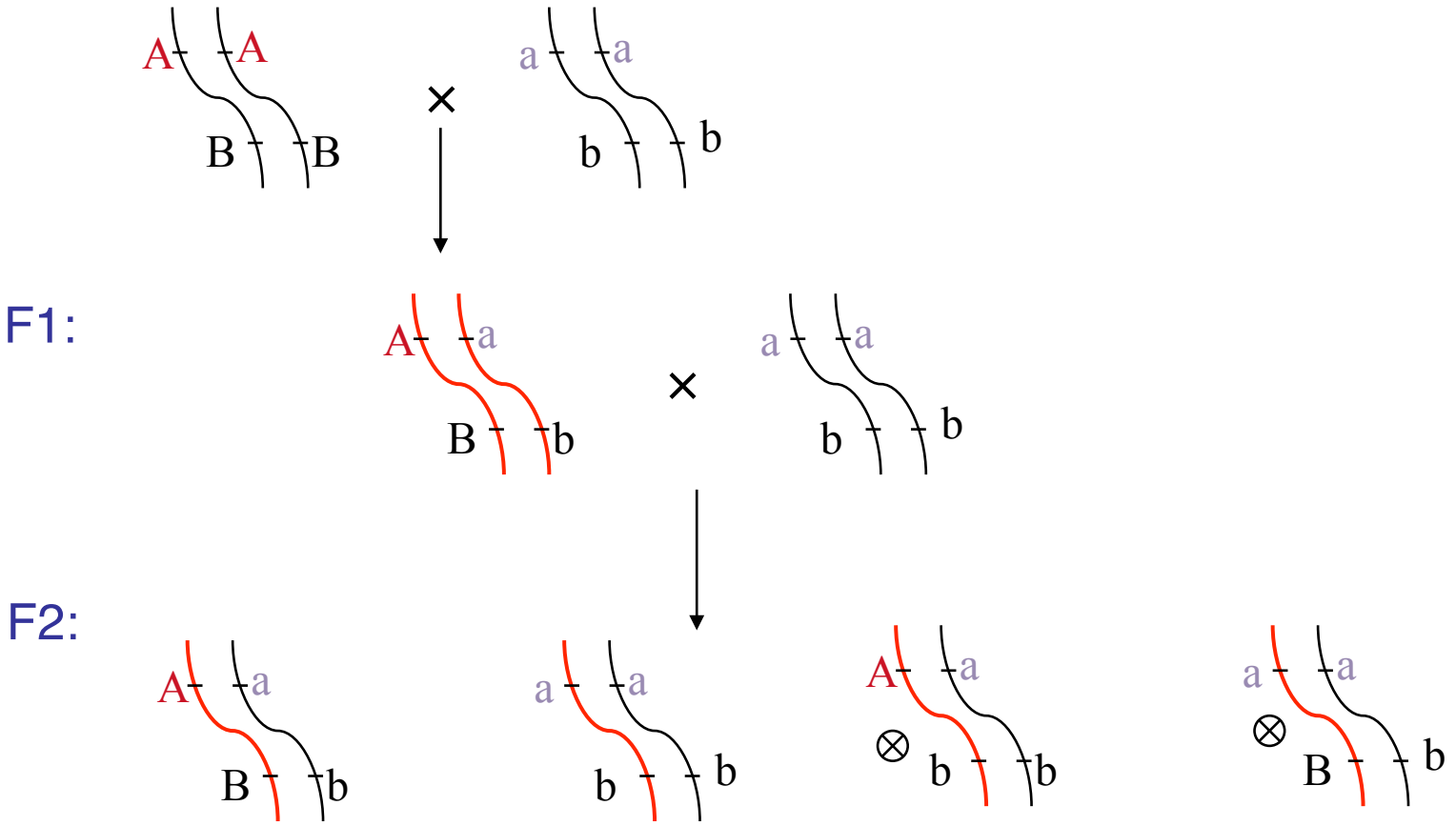
Eye color A: red a: purple
Wing length B: normal b: vestigial

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:

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

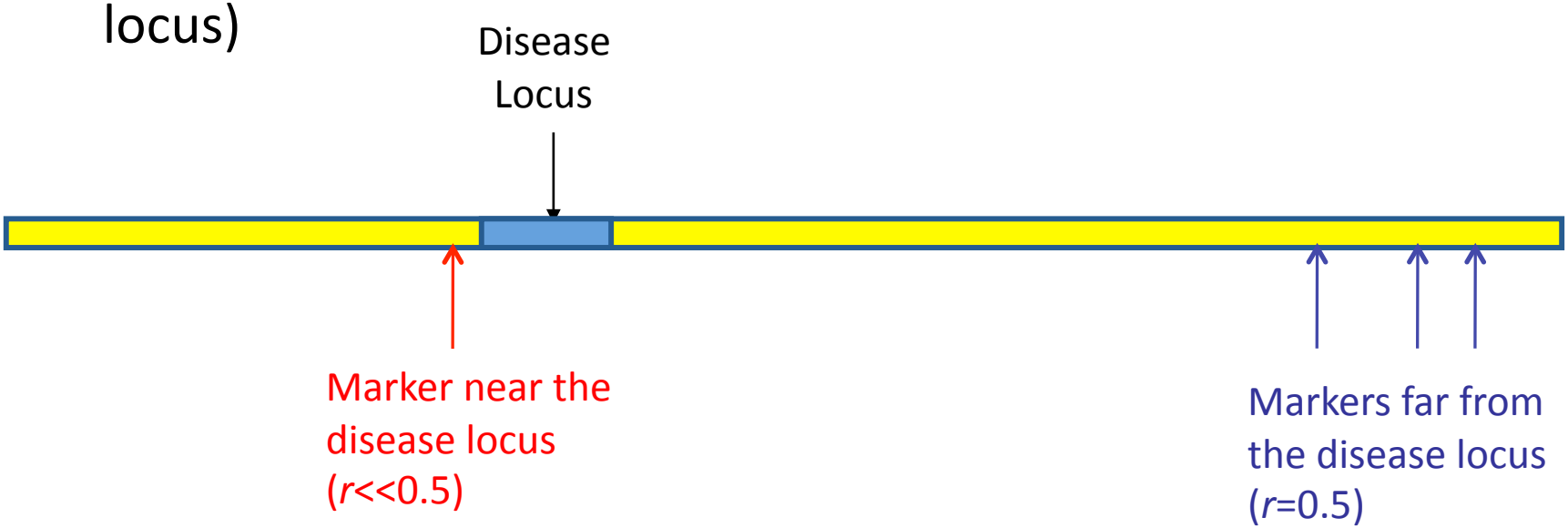
If $r < 1/2$: two genes are said to be ***linked***.

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

Now we move on to (small) pedigrees.

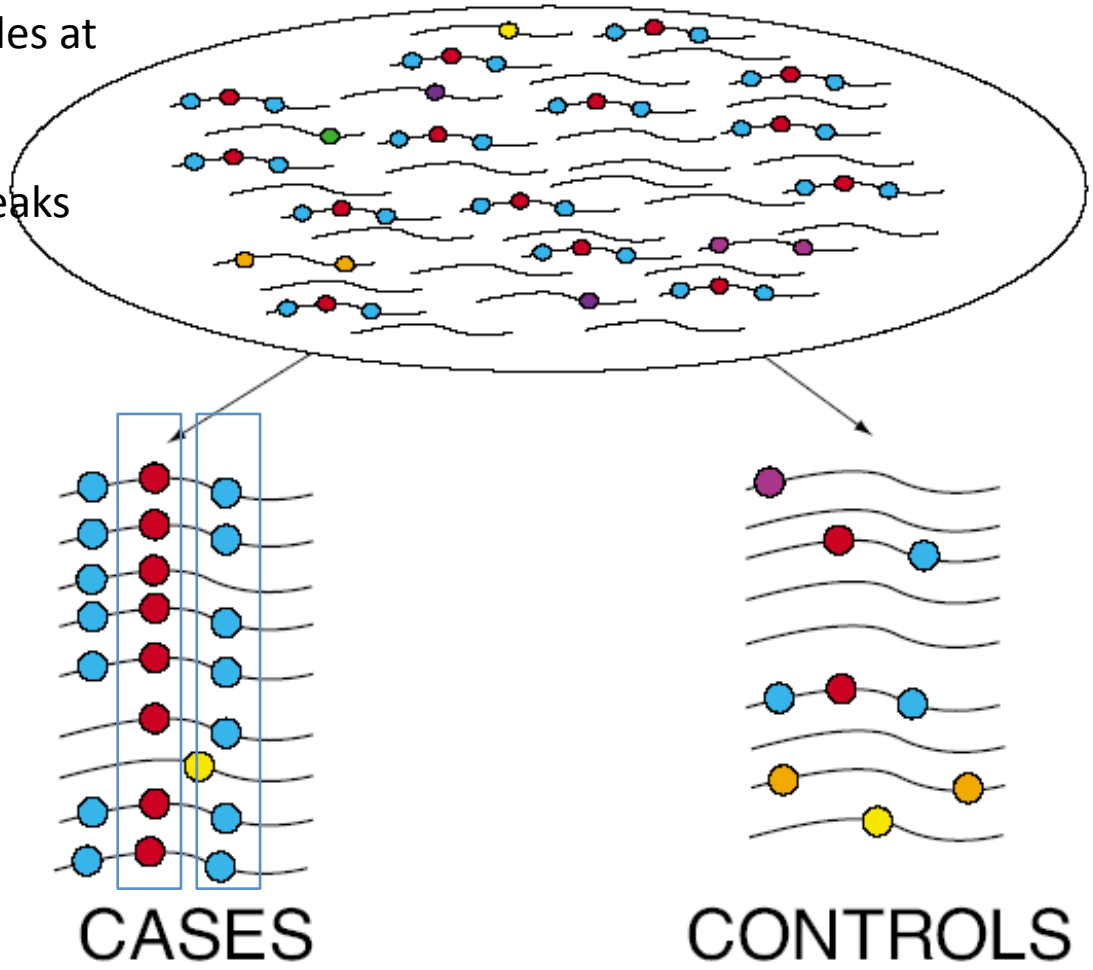
Linkage Analysis

- 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)



Linkage Disequilibrium in Gene Mapping

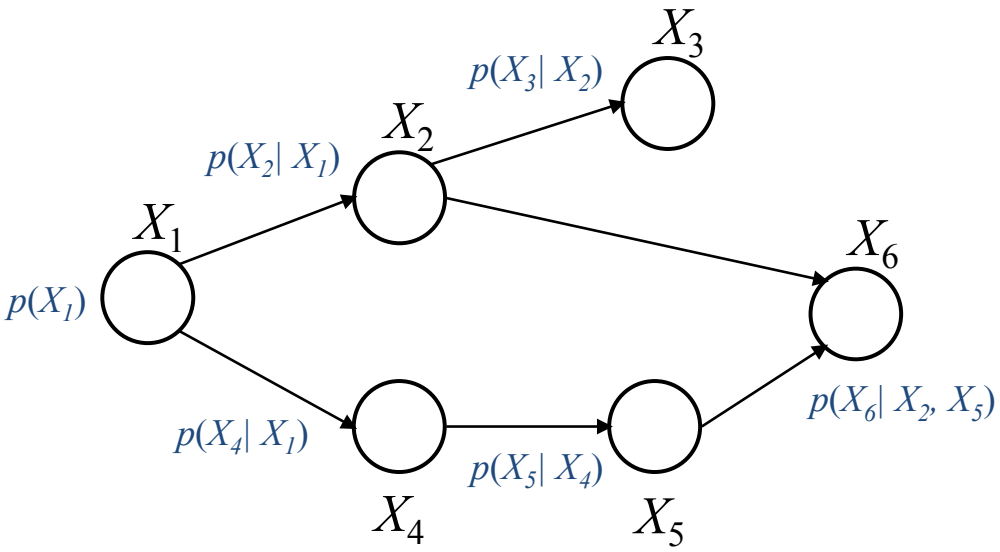
- LD is the non-random association of alleles at different loci
- Genetic recombination breaks down LD



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

Probabilistic Graphical Models

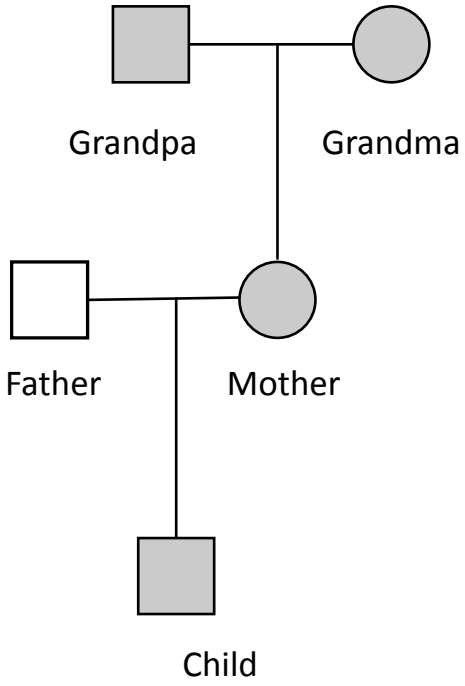


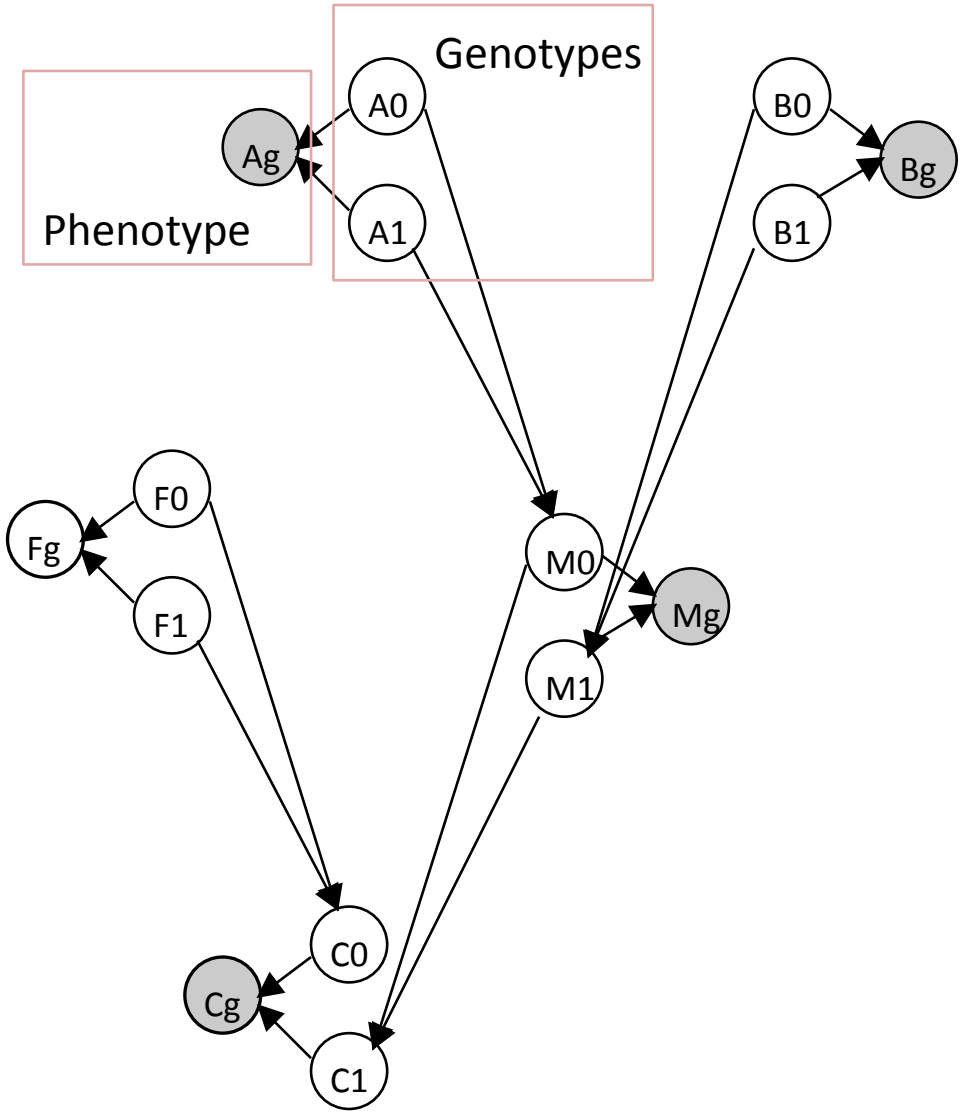
- The joint distribution on (X_1, X_2,…, X_N) factors according to the "parent-of" relations defined by the edges E :

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

Pedigree as Graphical Models: the Allele Network

Shaded means affected,
blank means unaffected.

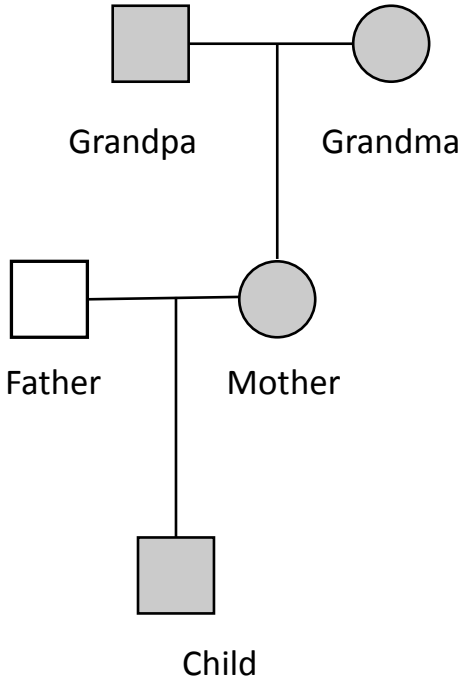




Founders and Non-founders

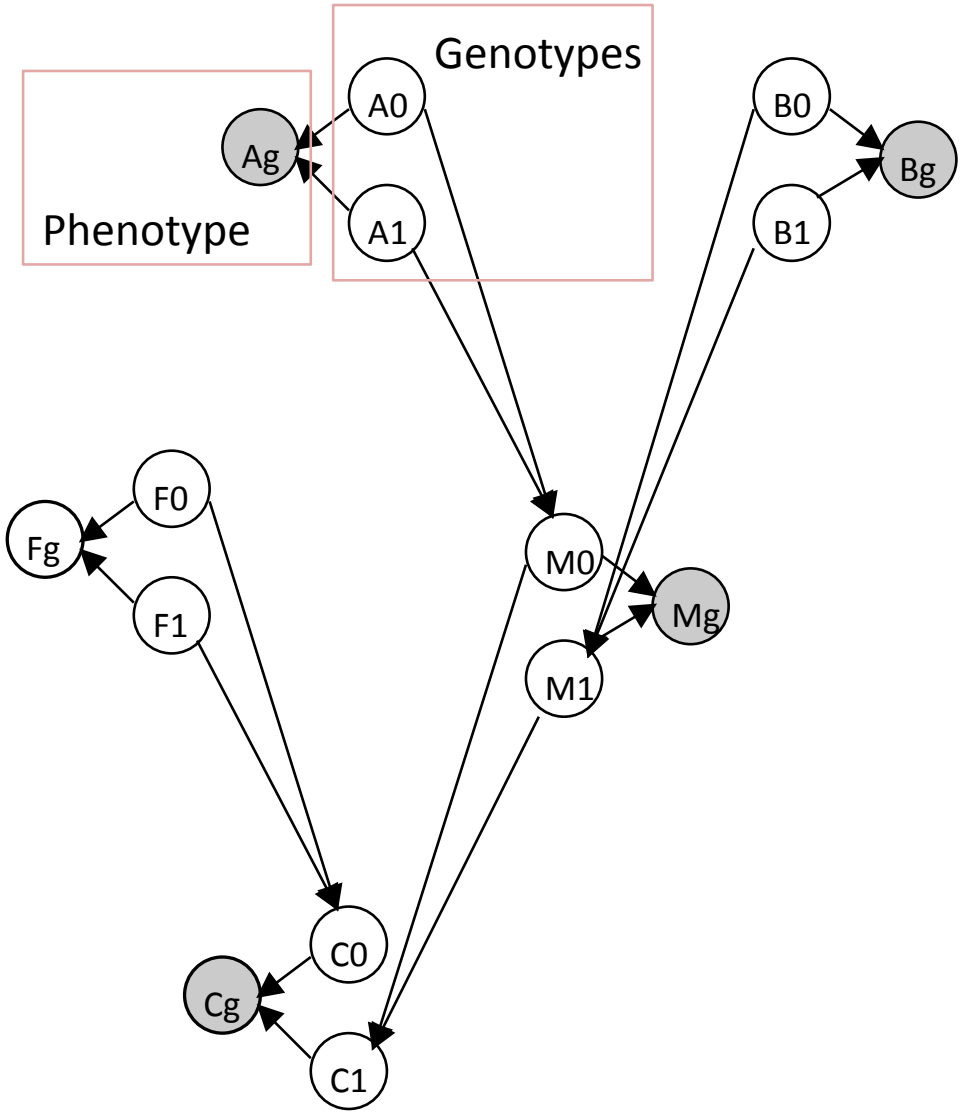
- **Founders:** individuals whose parents are not in the pedigree.

- **Non-founders**: individuals whose parents are not in the pedigree.



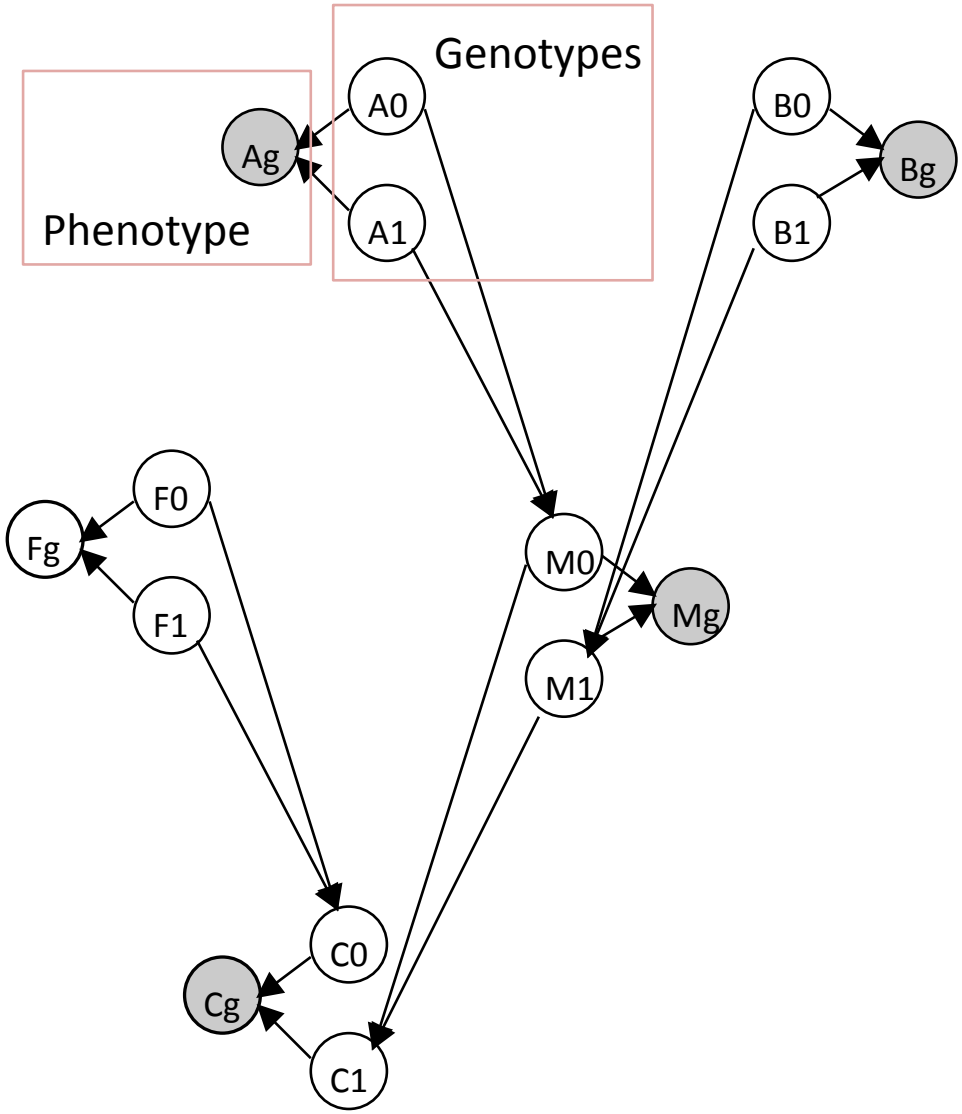
Probability Models over Pedigree

- **Founder genotype probabilities**
- **Transmission probabilities**: P(child's genotype | father's genotype, mother's genotype)
- **Penetrance model**: P(phenotype|genotype) for each individual



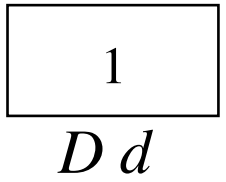
Probability Models over Pedigree

- Genotype probabilities are independent
  - across different founders
  - Across siblings of the same parents
- Phenotype probability of each individual is independent of all other individuals genotypes, conditional on their own genotype

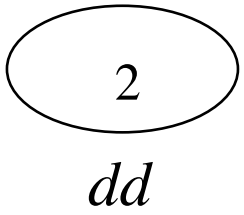


One Locus: Founder Genotype Probabilities

- ***Assign founder probabilities*** to their genotypes, assuming ***Hardy-Weinberg equilibrium***
  - Example: If the frequency of *D* is .01*, HWE says*



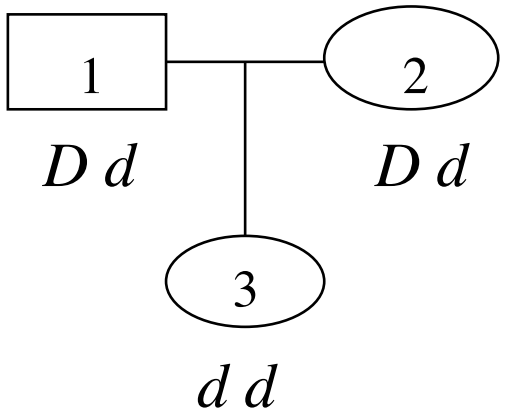
P(*Dd*) = 2x.01x.99



P(*dd*) = (.99)^2

One Locus: Transmission Probabilities

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



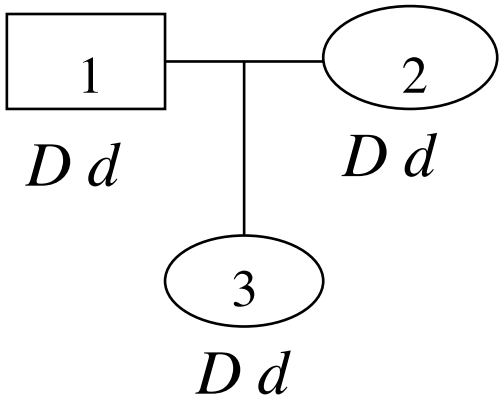
$$P(G_{ch3} = dd \mid G_{pop1} = Dd, G_{mom2} = Dd)$$

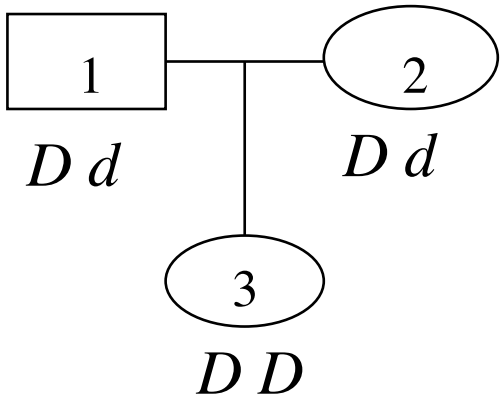
$$= 1/2 \times 1/2$$

- The inheritances are independent for different children.

One Locus: Transmission Probabilities

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





$P(G_{ch3} = Dd \mid G_{pop1} = Dd, G_{mom2} = Dd) = (1/2 \times 1/2)\times 2$

$P(G_{ch3} = DD \mid G_{pop1} = Dd, G_{mom2} = Dd) = 1/2 \times 1/2$

The factor 2 comes from summing over the two mutually exclusive and equiprobable ways Child3 can get a *D* and a *d*.

One Locus: Penetrance Probabilities

- Complete penetrance:

  P(Ph = affected | G=*DD*) = 1

  DD 

- Incomplete penetrance:

  P(Ph = affected | G=*DD*) = .8

  DD 

- Independent Penetrance 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.

One Locus: Penetrance Probabilities

- Age and sex-dependent penetrance:

D D (45 years old)

P(Ph = affected | G = *DD* , sex = male, age = *45 y.o.*) = .6

One Locus: Putting it All Together

- The overall pedigree likelihood is given as

$$L = \prod_{f \in Founders} P(G_f) \prod_{ch \in Nonfounders} P(G_{ch} \mid G_{pop}, G_{mom}) \prod_{i \in \{Founders, Nonfounders\}} P(Ph_i \mid G_i)$$

- If founder or non-founder genotypes are **unavailable/missing**, we sum over all possible genotypes for those individuals with missing genotypes to obtain the likelihood

$$L = \sum_{G_f, G_{ch}} \prod_{f \in Founders} P(G_f) \prod_{ch \in Nonfounders} P(G_{ch} \mid G_{pop}, G_{mom}) \prod_{i \in \{Founders, Nonfounders\}} P(Ph_i \mid G_i)$$

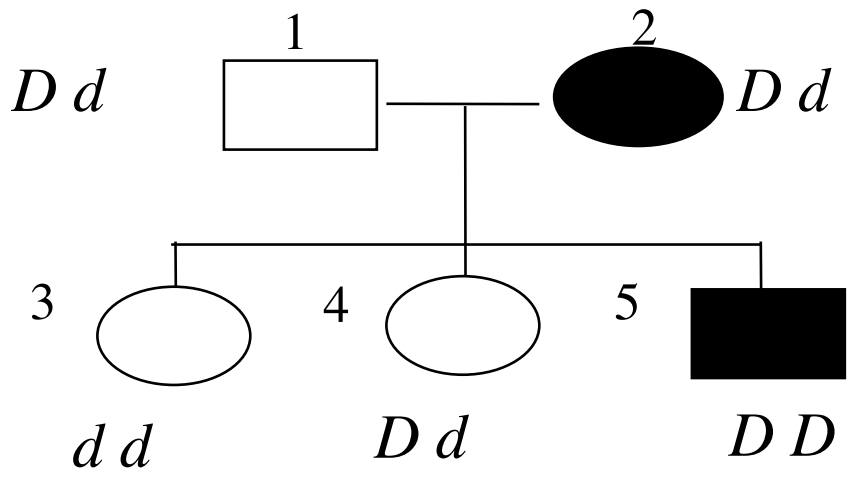
One Locus: LOD Score

- Null hypothesis: the disease locus is **unlinked** to the given marker locus being tested

- Alternative hypothesis: the disease locus is **linked** to the given marker locus being tested

- LOD Score = Log_{10} (Likelihood under the alternative hypothesis) – Log_{10} (Likelihood under the null hypothesis)
  - Likelihood under the null hypothesis can be obtained by summing the pedigree likelihood over all possible genotypes of the all pedigree individuals: Computationally expensive but efficient algorithms exist

One Locus Example



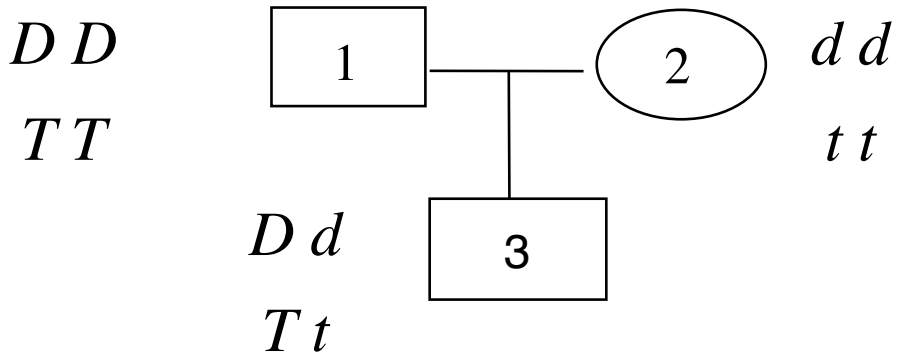
- Assume
  - Penetrances: P(affected | *dd)* = .1, P(affected | *Dd*) = .3, P(affected | *DD*) = .8.
  - Allele *D* has frequency .01.
- The probability of this pedigree is given as

  (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)

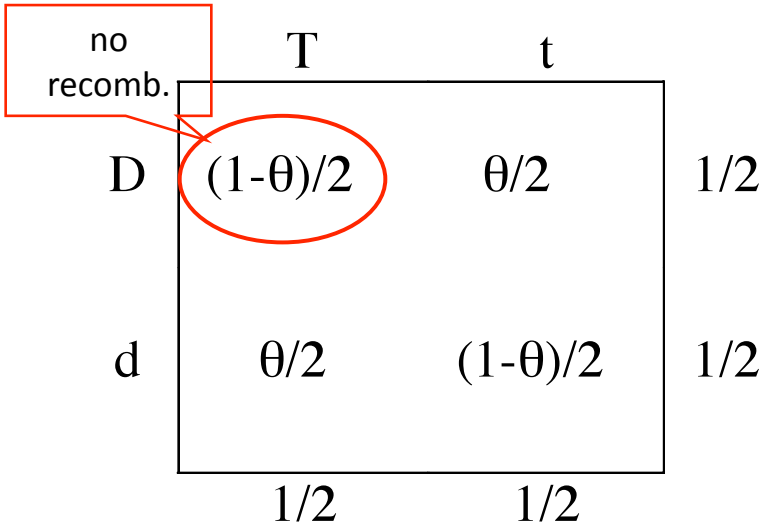
One Locus Analysis

- 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 called the Lander-Green algorithm and takes a quite different approach. Lander-Green algorithm uses hidden Markov models to model multiple loci

Two Loci: Linkage and Recombination



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



Two Loci: Linkage and Recombination

- 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

Two Loci: Linkage and Recombination

- 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	

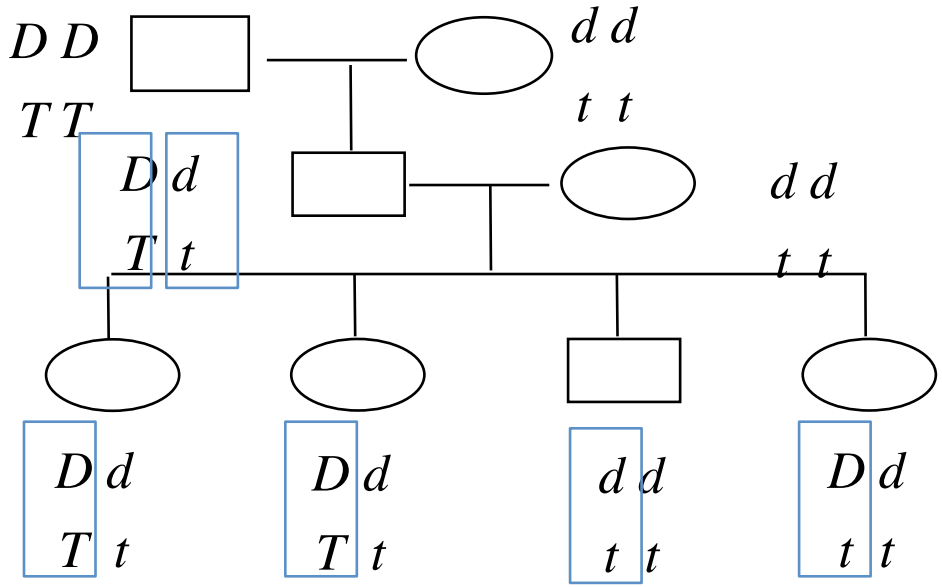
- If the loci are linked,
  - *D-T* and *d-t* are ***parental*** *haplotypes*
  - *D-t* and *d-T* are ***recombinant*** haplotypes

Phase

- Phase is known for an individual if you can tell whether the gamete was parental or recombinant

- Phase is unknown if you cannot tell whether the gamete was parental or recombinant

Two Loci: Phase Known Pedigree



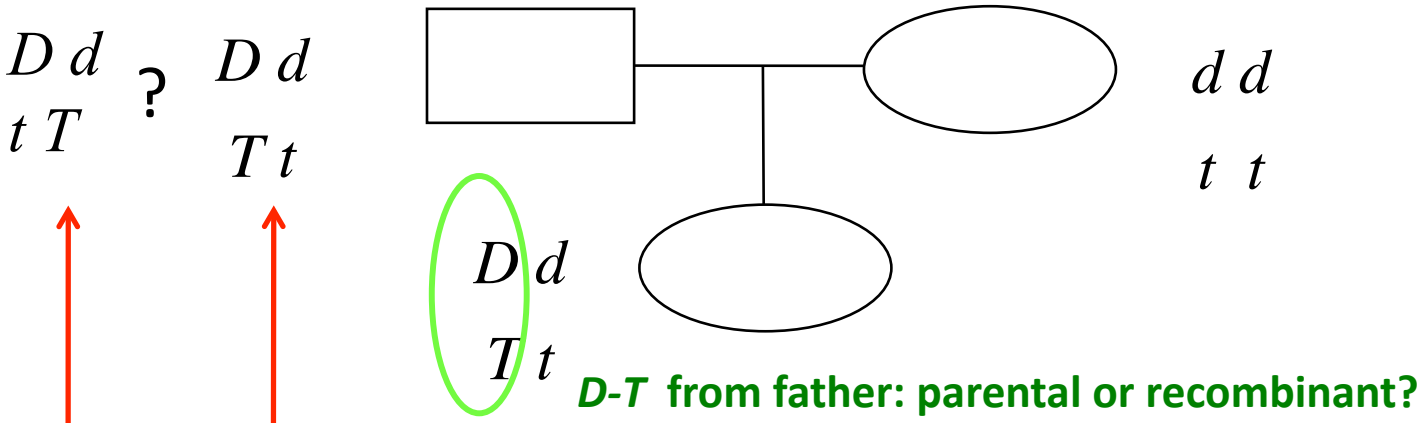
Recombination only discernible in the father. Here $\hat{\theta} = 1/4$ (why?)

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

What if the grandparents' genotypes are not known?

Two Loci: Phase Unknown Pedigree

- Suppose the grandparents' genotypes are unavailable in the double backcross pedigree



Two Loci: Dealing with Phase

- Phase is usually regarded as unknown genetic information
- Sometimes, but not always, phase can be inferred with certainty from genotype data on parents, multiple children, relatives.
- In practice, probabilities must be calculated under all phases compatible with the observed data, and added together: computationally intensive, especially with multilocus analyses.

Two Loci: Founder Probabilities

- Assume **linkage equilibrium**, i.e. independence of genotypes across the two loci.

- Allele frequencies at locus one: D = .01, and d = .99

  Allele frequencies at locus two: T = .25 and t =.75

  – Haplotype frequencies
    - DT = .01 x .25
    - Dt = .01 x .75
    - dT = .99 x .25
    - dt = .99 x .75

  – Together with Hardy-Weinberg, this implies that

  P(G = $DdTt$) = (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).

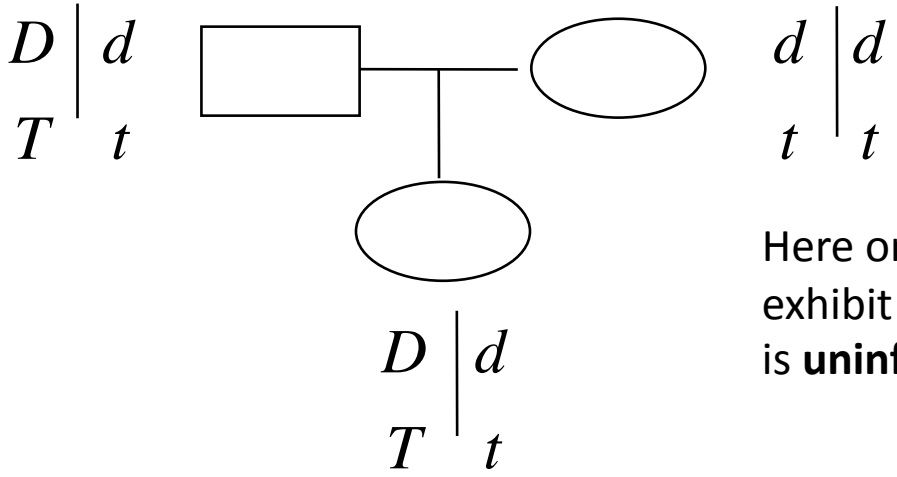
  adds haplotype pair probabilities.

Dd

Tt

Two Loci: Transmission Probabilities

- For a given haplotype inheritance:



Here only the father can exhibit recombination: mother is **uninformative**.

$P(G_{ch} = DT/dt \mid G_{pop} = DT/dt, G_{mom} = dt/dt)$

$= P(G_{ch} = DT \mid G_{pop} = DT/dt) \times P(G_{ch} = dt \mid G_{mom} = dt/dt)$

$= (1-\theta)/2 \times 1.$

- Sum the probabilities over all possible phases/haplotypes.

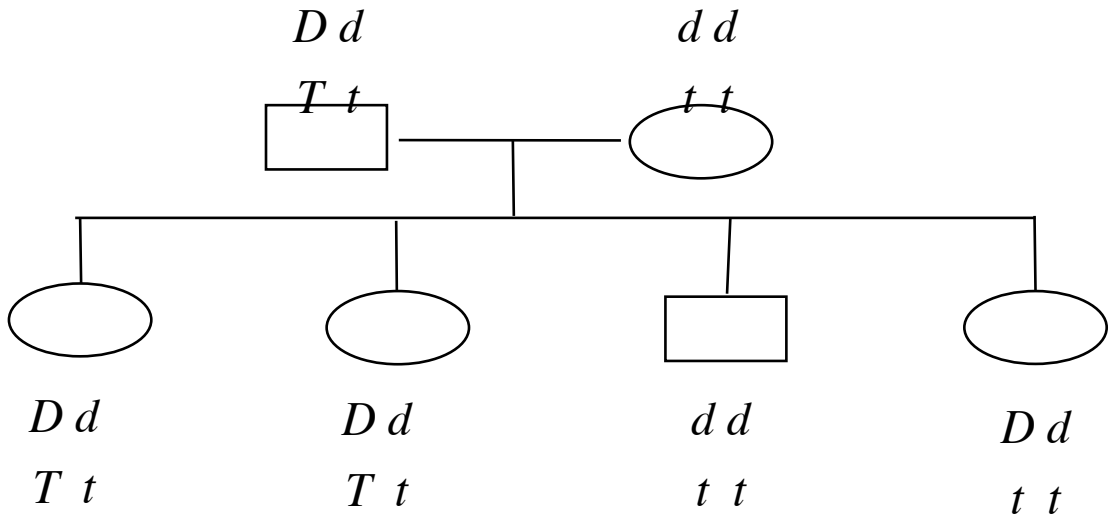
Two Loci: Penetrance

- 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.

$$P(\, Ph_1 = \text{affected}, Ph_2 = Tt \mid G_1 = DD, G_2 = Tt \,)$$
$$= \Pr(\, Ph_1 = \text{affected} \mid G_1 = DD, G_2 = Tt \,) \times \Pr(Ph_2 = Tt \mid G_1 = DD, G_2 = Tt \,)$$
$$= 0.8 \times 1$$

Two Loci: Phase Unknown Double Backcross

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



P(all data | θ)
= P(parents' data | θ) × P(kids' data | parents' data, θ)
= P(parents' data) × {[((1-θ)/2)^3 × θ/2]/2+ [(θ/2)^3 × (1-θ)/2]/2}

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

Log (base 10) Odds or LOD Scores

- Suppose P(data | θ) is the likelihood function of a recombination fraction θ generated by some 'data', and P(data | 1/2) is the same likelihood when θ= 1/2.

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

  $LOD(\theta^*) = Log_{10}P(data \mid \theta^*) - Log_{10}P(data \mid ½)$

  measures the relative strength of the data for $\theta = \theta^*$ (optimal θ) rather than θ = 1/2.

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 ≈ 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.