

Linkage Analysis for Gene Identification

Chiara Sabatti

csabatti@mednet.ucla.edu

www.stat.ucla/~sabatti/home/teaching/teach.html

Why do we want to find a gene?

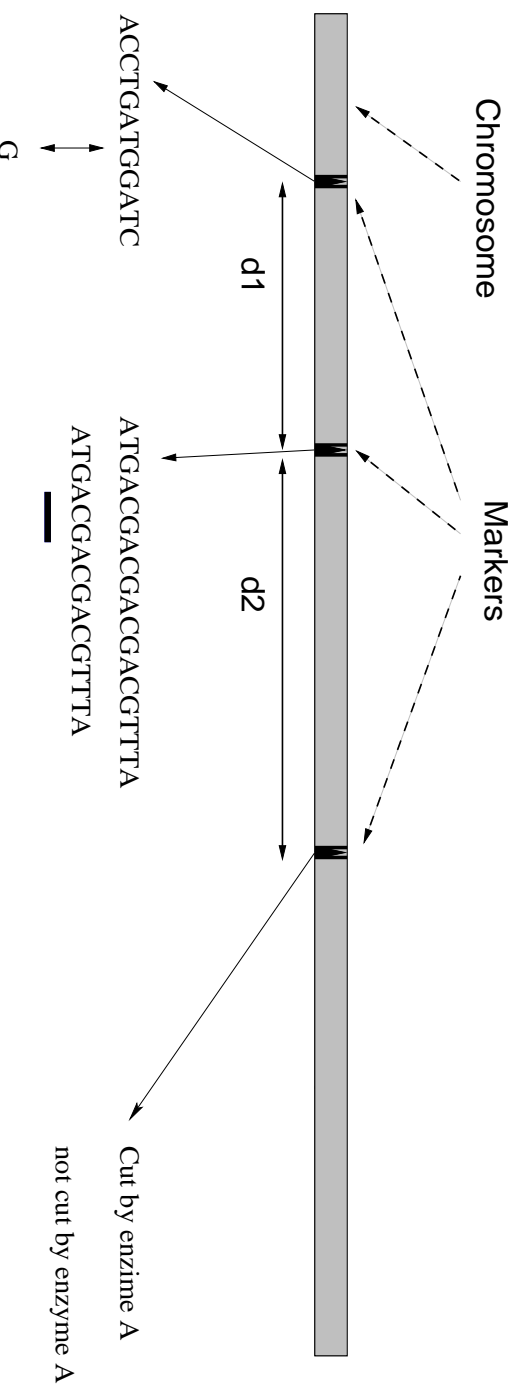
- curiosity
- relevant for medicine
 - screening and counseling
 - learn about the biological pathway affected
 - develop drugs that restore functionality of pathway
 - this same pathway could be affected in other form of the disease and the same drugs may be useful for other stuff
 - sophisticated diagnostic tool that assign the correct treatment to different form of disease
 - genetic therapy

What does it mean to find a gene?

1. Locate a chromosomal region where the gene seems to be
2. refine the possible region to a length where it can be studied in detail:
 - sequence
 - study of functions of genes involved
 - study of polymorphism
3. identify one gene that is mutated in affected individuals and not in controls
4. understand the function of that one gene and the effect of the mutation.

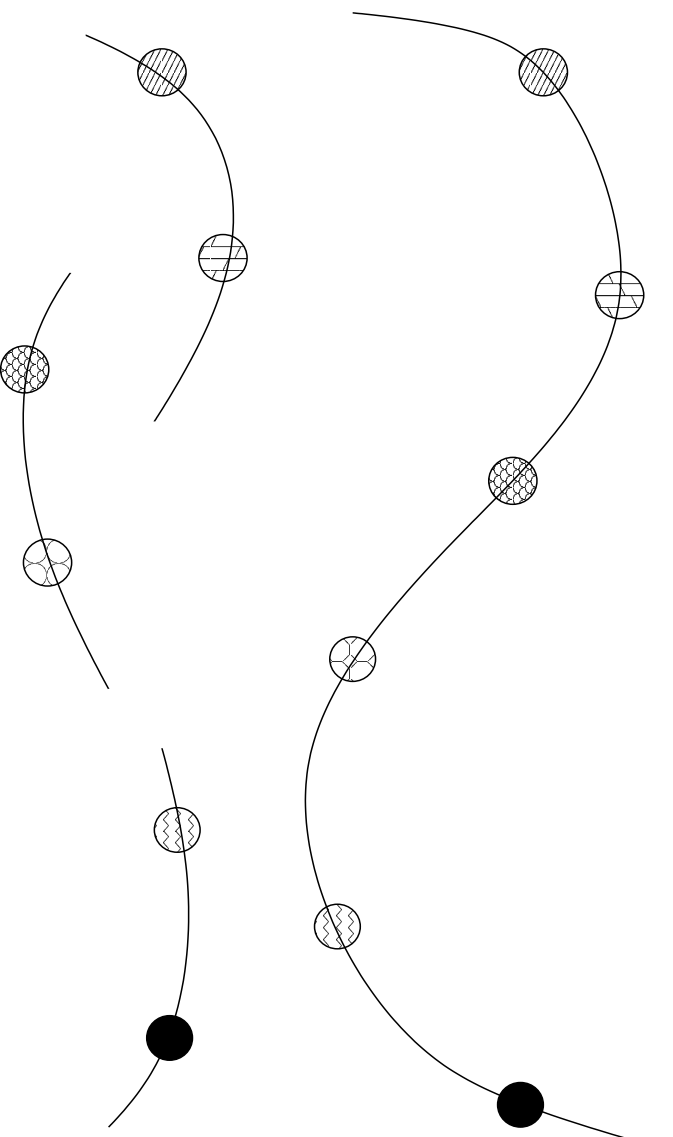
The first step: locate gene on a map

Markers are landmarks in the genome: places that we recognize, of which we know the relative positions and where DNA is polymorphic. (ex. street signs, Las Vegas and the Grand Canyon while driving in the desert)



How do we construct maps?

distance between beads proportional to number of times they are together



“Cutting processes” and map types

GENETIC MAPS

PHYSICAL MAPS

recombination

radiation

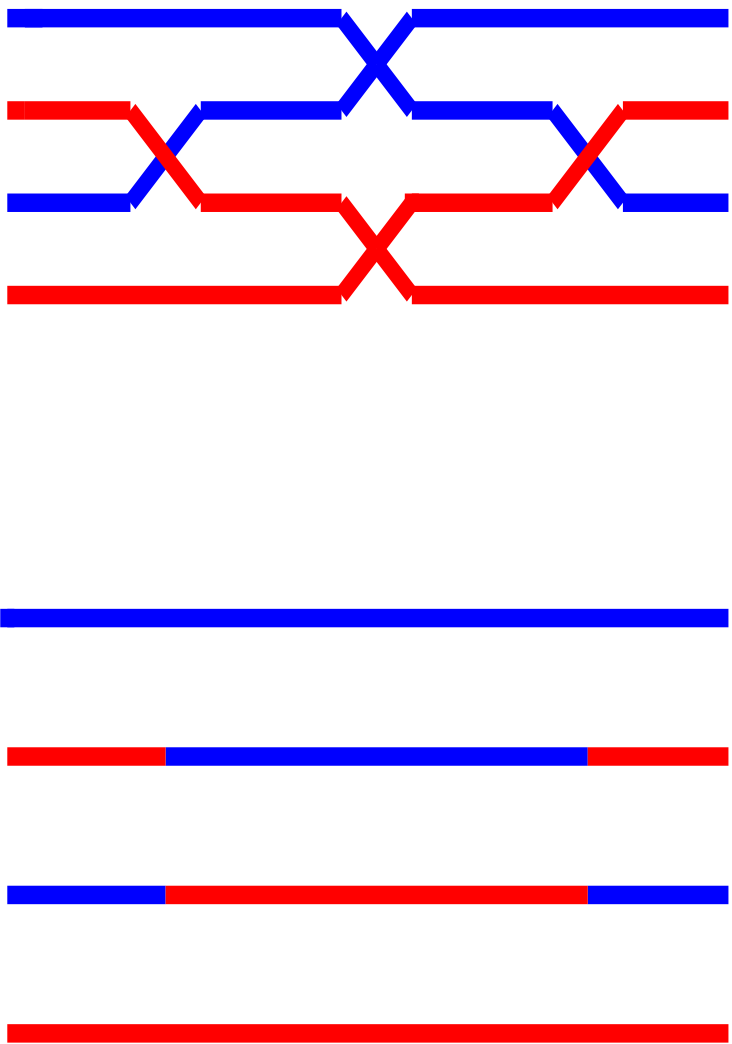
Morgans

number of bases

(more later)

(see book)

Crossover: the hidden process



two chiasmata;
average number of crossover per gamete is 1;

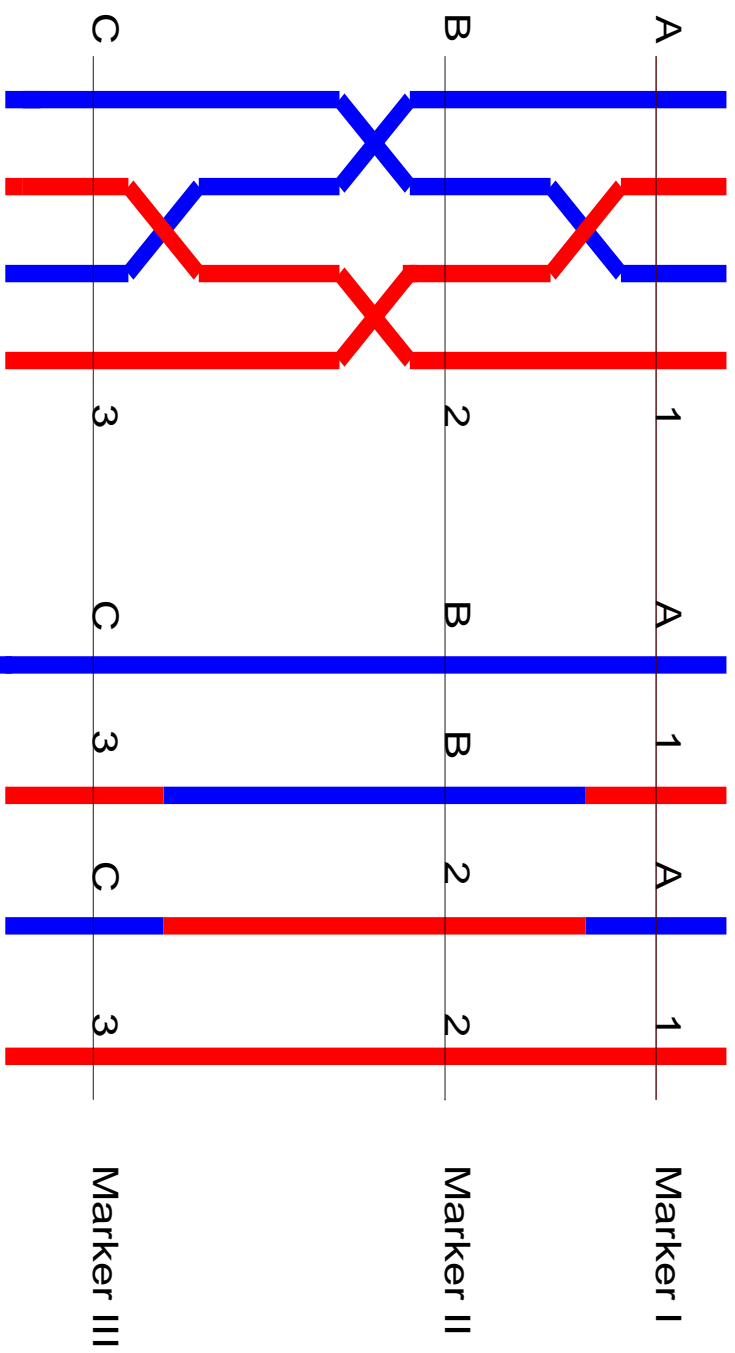
Distance and Crossovers

The distance between two position x and y on a chromosome is defined as the expected number of crossovers per gamete between x and y

$$d(x, y) = E(\text{crossover between } x \text{ and } y)$$

is measured in Morgans.

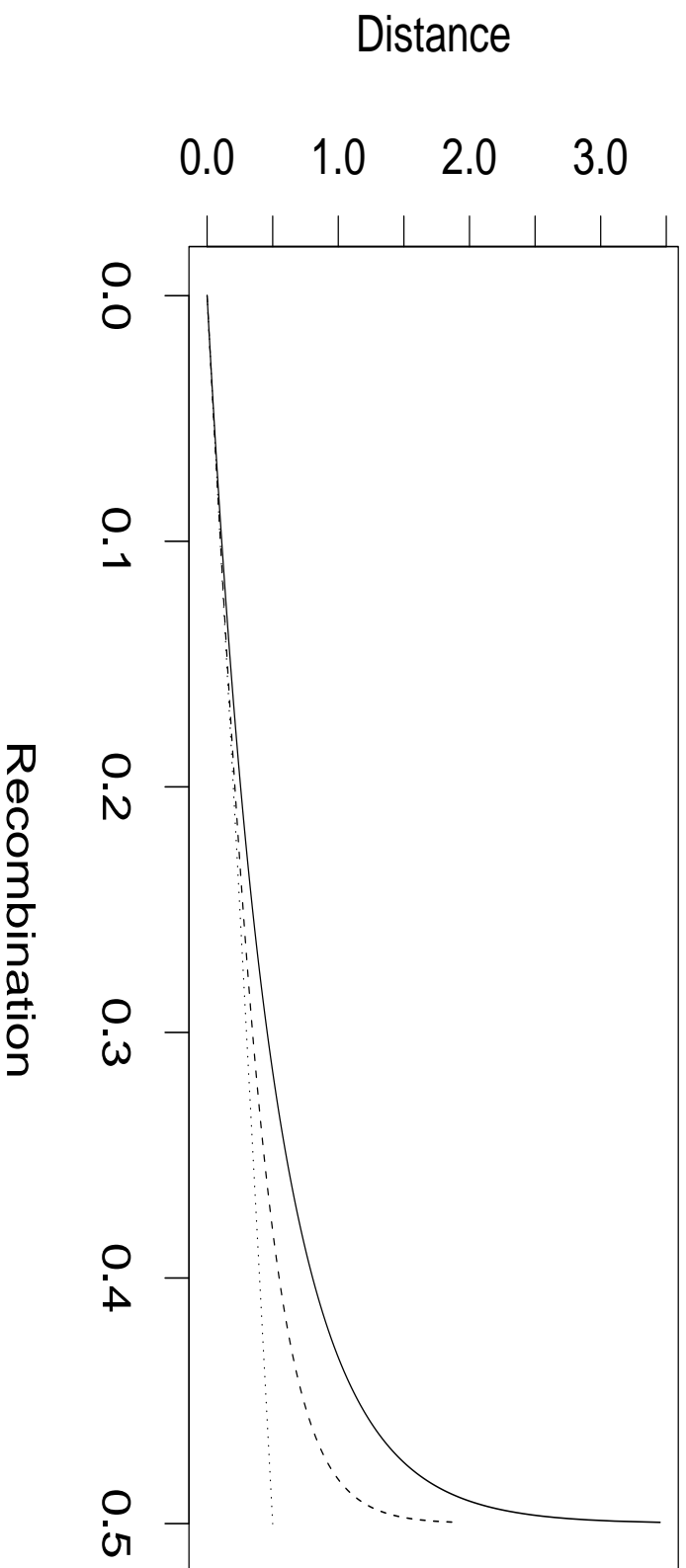
Recombination: the observable process



- 2 recombinant gametes between marker I and II;
- 2 recombinant gametes between marker II and III;
- 0 recombinant gametes between marker I and III;

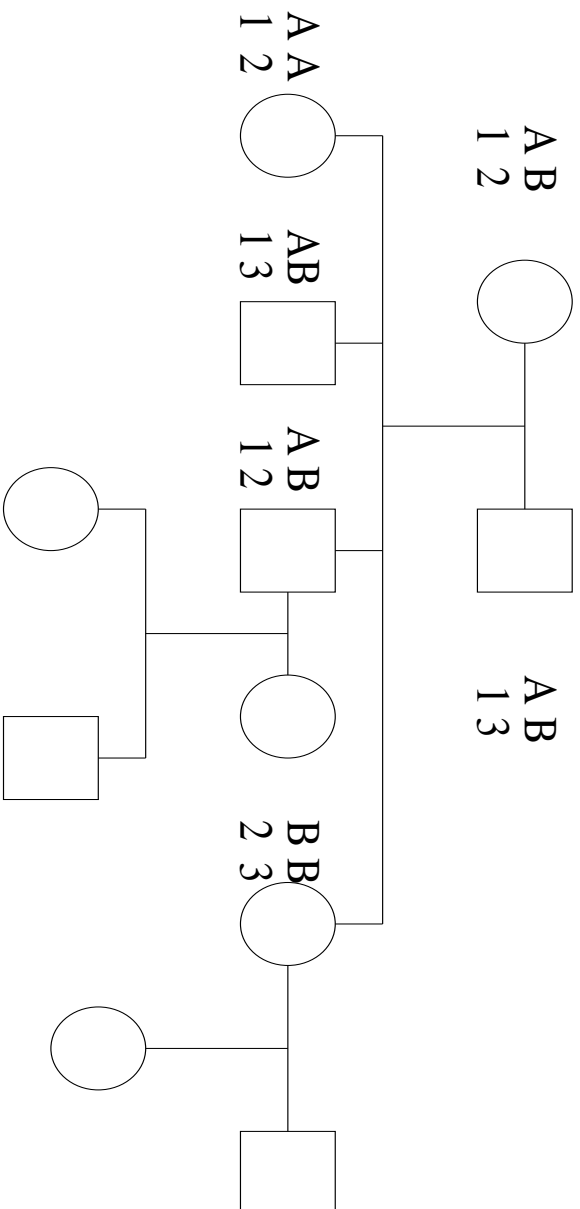
Recombination and distance

We can go from Recombination to distance with mathematical formulas



How do we measure Recombination?

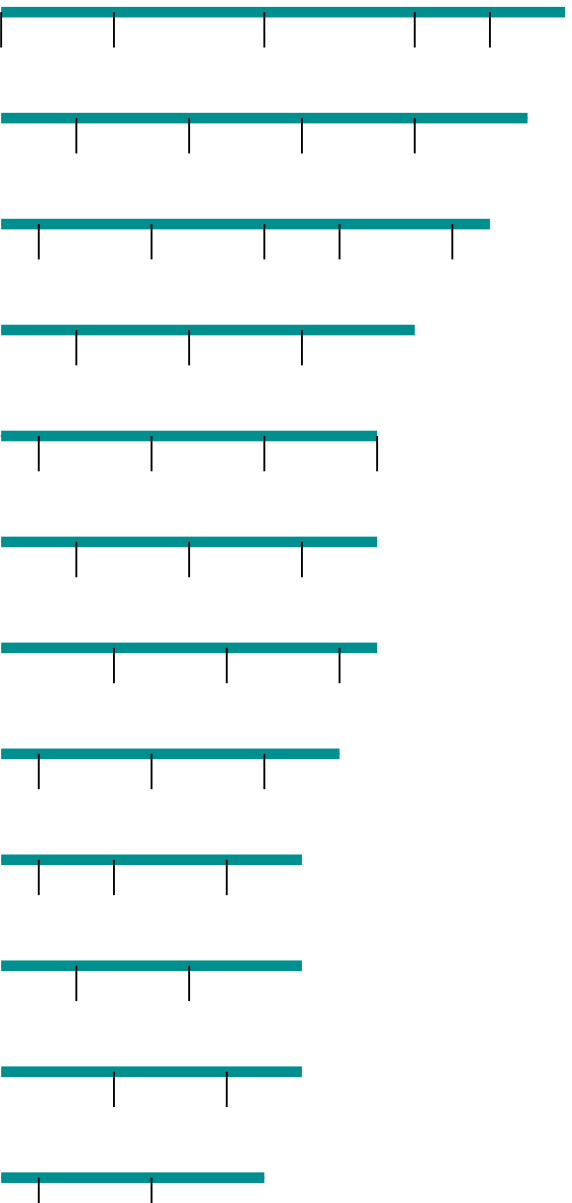
By looking at genotypes of Pedigrees



(Marshfield, CEPH families)

Mapping a Disease Gene

- Collect families with affected individuals
- genotypes markers that cover the entire genome
- Estimate recombination between disease gene and markers



When is a disease “linked” to a marker?

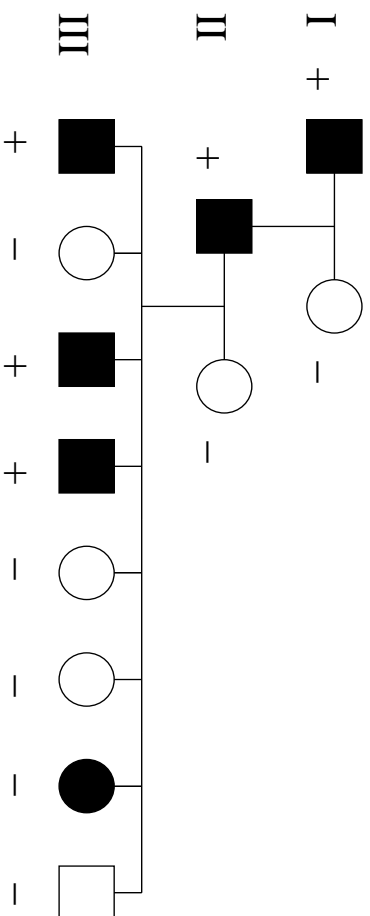
linked = recombination fraction $\theta < 1/2$

- the estimated recombination fraction $\hat{\theta}$ between disease and marker is $< 1/2$
- this result is statistically significant

$$\Rightarrow \log_{10} \frac{\text{Prob}(\text{Observation}|\hat{\theta})}{\text{Prob}(\text{Observation}|\theta = 1/2)} > 3$$

An example

Dominant disease and Rh



disease
 (D,d) or (D,D)
 normal
 (d,d)
 +
 (+,-) or (+,+)
 -
 (-,-)

Generation II haplotypes are

$$F = \{(D, +), (d, -)\} \quad M = \{(d, -), (d, -)\}$$

1 recombinant

7 non recombinant

θ = probability of recombination

Prob(observation| θ) = $\theta(1 - \theta)^7$

$\hat{\theta} = 1/8$

$$\text{LodScore} = \log_{10} \left(\frac{1/8(1 - 1/8)^7}{1/2^8} \right) = \log_{10} (12.56) = 1.1$$

\Rightarrow there is no evidence for linkage

Why?

- not enough data to be conclusive
- LodScore is log of ratio of probability of observations under best estimate or recombination over probability of observation under recombination = $1/2$;
- it is very unlikely that two things are linked, so one has to collect very strong evidence
- if I calculate the possible linkage with a big number of markers I have to worry about multiple comparisons

More sophistication

- In most cases we have to impute phases
- The above was done assuming dominant inheritance, but in many cases inheritance is not Mendelian: non parametric models
- It is important to consider the information from proximal markers as one piece of information: multipoint analysis

Some issues

The success of a linkage screen will depend on

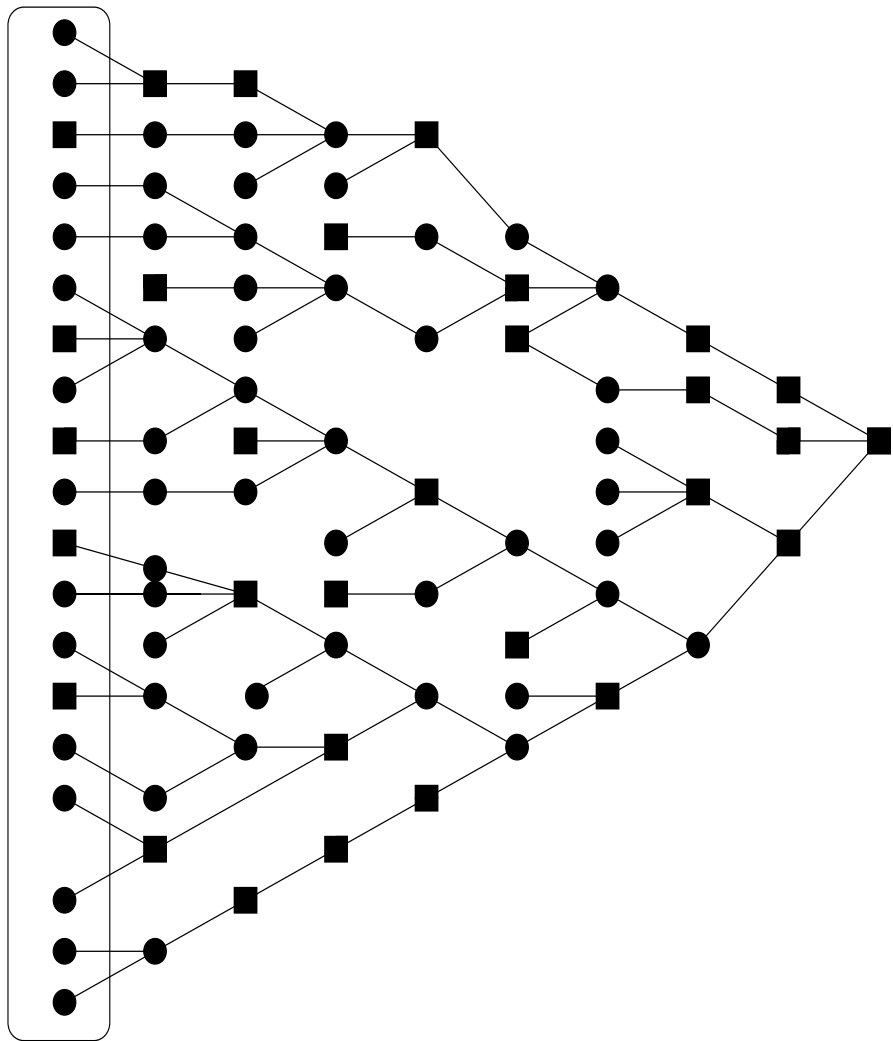
- how well the markers span the genome
- how good is the map information
- how informative are the markers
- how many recombination we can observe per family
- how well defined is a phenotype

From family to population data

- The power of resolution comes from recombination.
- In families is quite difficult to observe recombination between markers that are less than 1 cM apart (probability of crossover per gamete = 0.01).
- 1 cM \approx 1000000 bases
- sequencing machines give you a reliable read for 350–400 bases
- need a narrower map interval!
- In some cases we can treat the population of diseased individuals as a big family and observe there the effects of recombination.

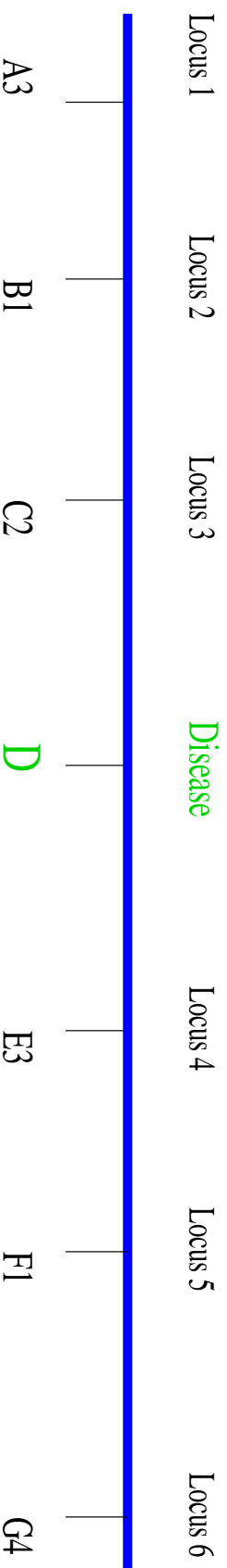
Family vs Population data

- By looking at pedigrees, we can actually observe recombination events. When we look at the current population of diseased individuals, we are looking at only 1 generation, so we do not directly observe recombinations.
- In a given family, the disease status is associated with one allele at linked markers; this is not true in general for populations.
- If the population has derived from 1 ancestor, we can observe the effects of recombination on hundreds of generations by looking at the association between one allele and the disease status.



Founder effect

Suppose that in a population of 100 chromosomes, 1 undergoes a mutation in a gene that causes a disease



- Initially all the chromosomes that inherit the disease inherit this haplotype;
- recombination and mutation will erode the haplotype;
- the markers really close to the disease locus will tend to have the same allele as in the founder \implies different distribution from the general population.

Linkage Disequilibrium

We look for association between allele distribution and disease status at various markers and locate the disease near the marker whose allele distribution varies the most with disease status.

- works well in population isolates
- it is connected to population history
- it is quite complicated to model mathematically
- it may be applied as a genome-screen technique with very dense markers.

The future of gene localization

- Linkage has been very successful for Mendelian diseases.
- For Complex diseases, no strategy has yet won the battle. Linkage disequilibrium may be more powerful, but it is still more a promise than a reality.
- Consequences of the human genome project: we will have lots of markers and very good maps.
- Gene expression data may help defining the phenotype of complex diseases.