

Computational Human Genetics

Itsik Pe'er

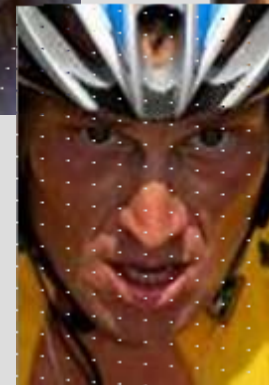
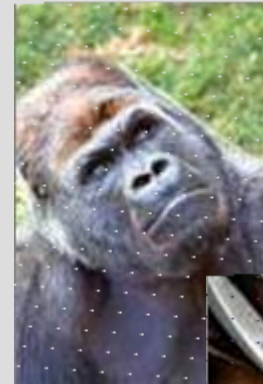
Department of Computer Science
Columbia University

Fall 2006

Reminder

- Population genetics inferences:

Modeling human history

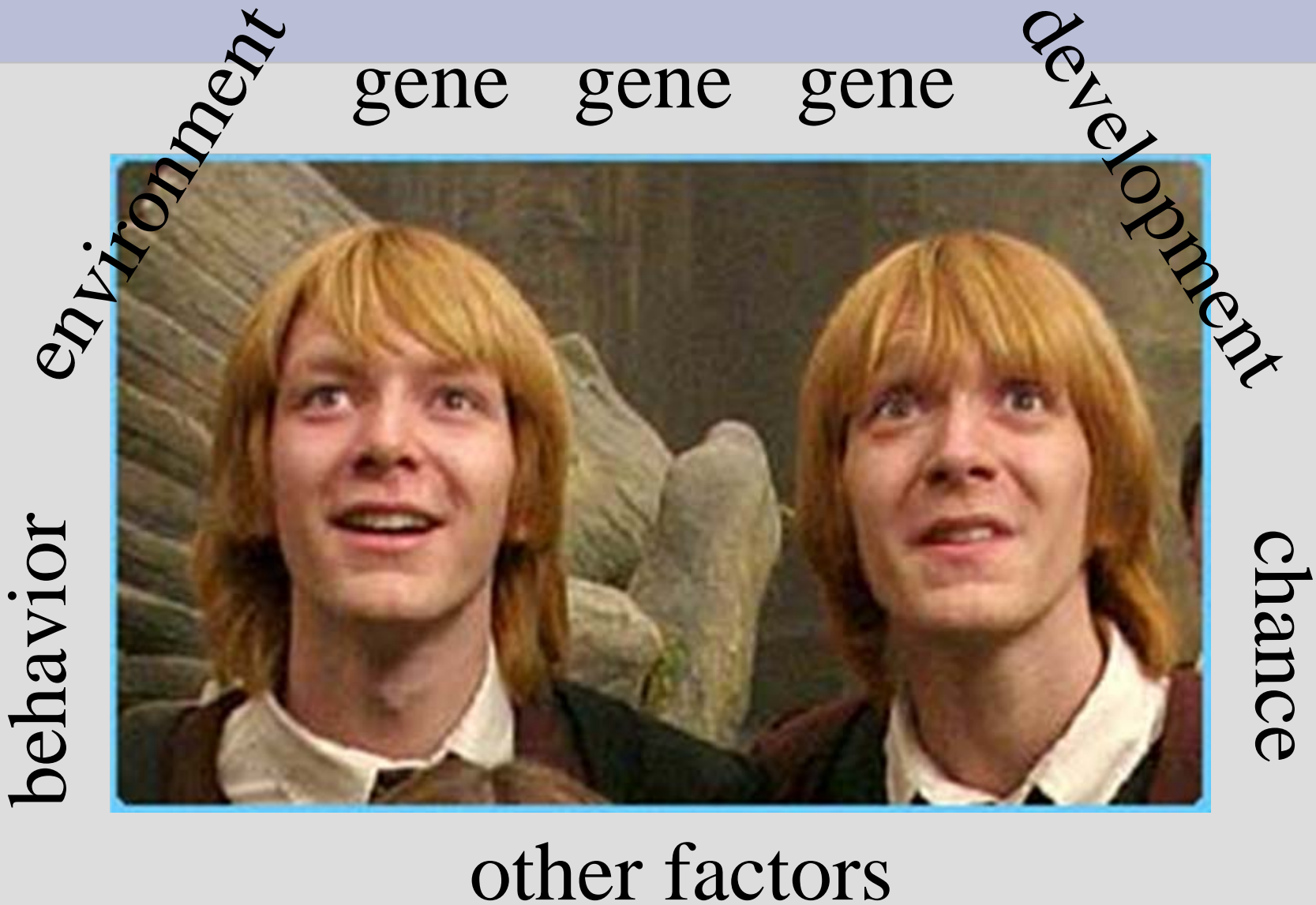


How about phenotypes?

Meeting #4

Linkage Analysis

Heritability of human phenotypes



Large genetic contribution by unknown genes:

Means to understand disease



The promise of personalized medicine

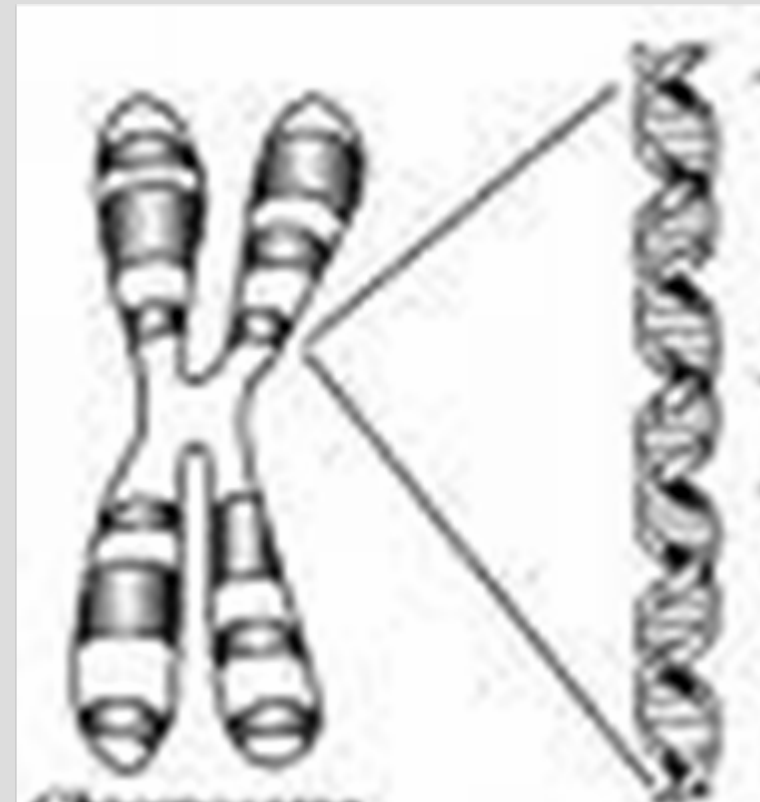
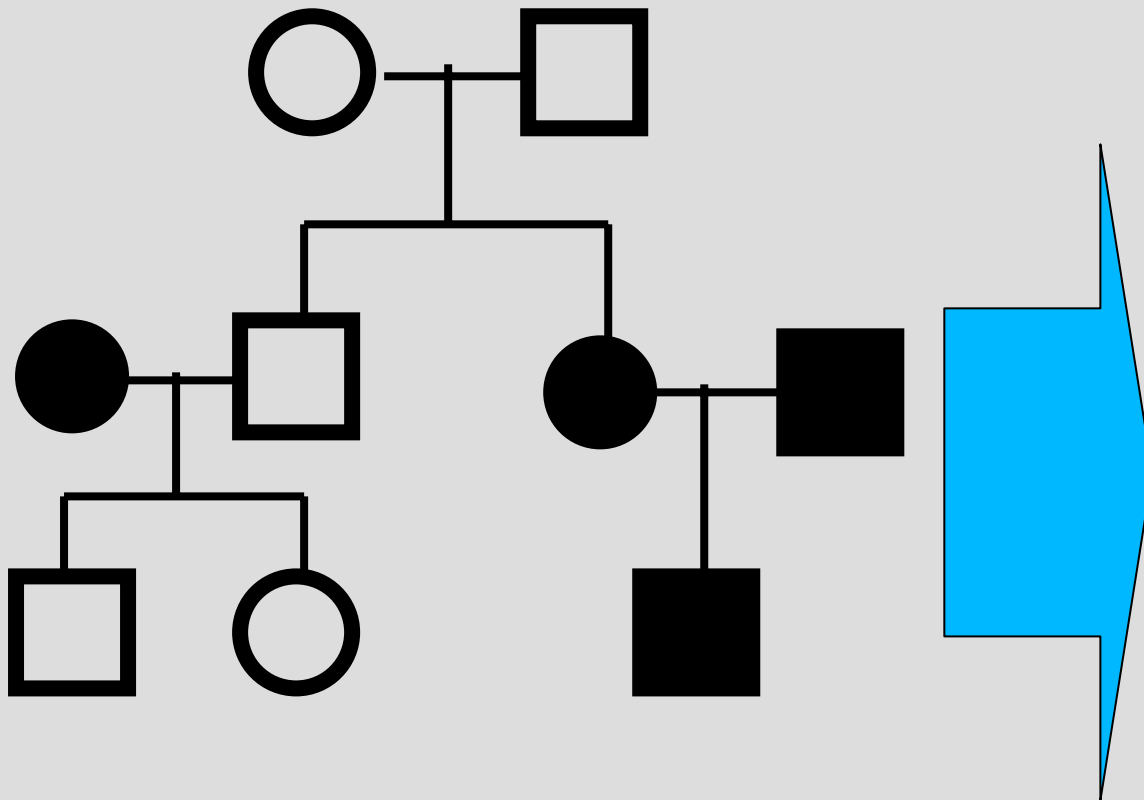


Genetics can help predict:

- Disease:
Will I become diabetic?
- Treatment:
Will this tumor respond to chemo?

Gene Mapping/Positional Cloning

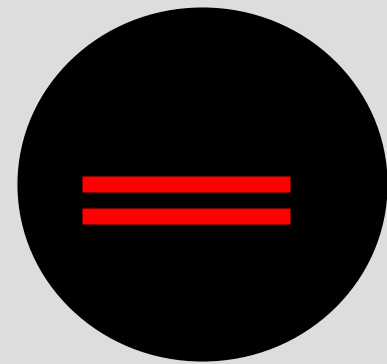
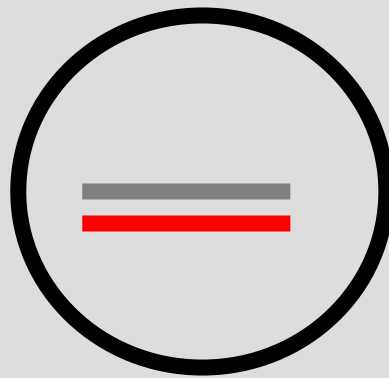
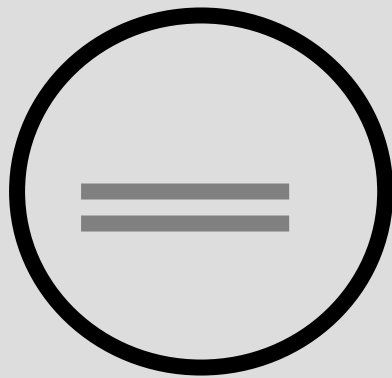
- From trait calls to functional region



Linkage Analysis

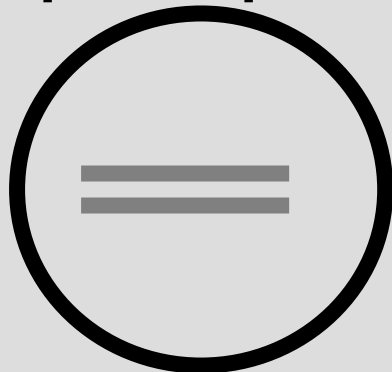
- Homozygosity mapping for rare recessives
 - Identity by state/descent
 - Probabilistic model
- The general case of linkage analysis
 - Lander-Green
 - Elston-Stewart

Recessive Disease

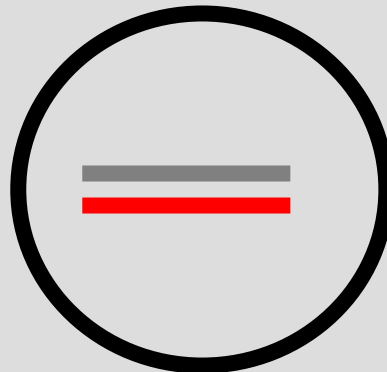


Rare Recessive Alleles

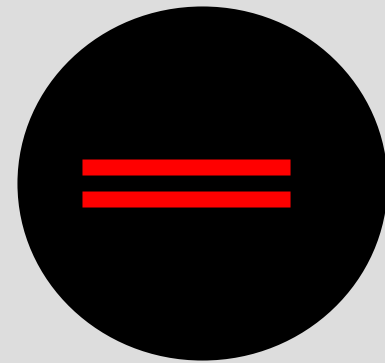
- Allele frequency: $p \ll 0.5$
- $p^2 \ll p$



q^2



$2pq$

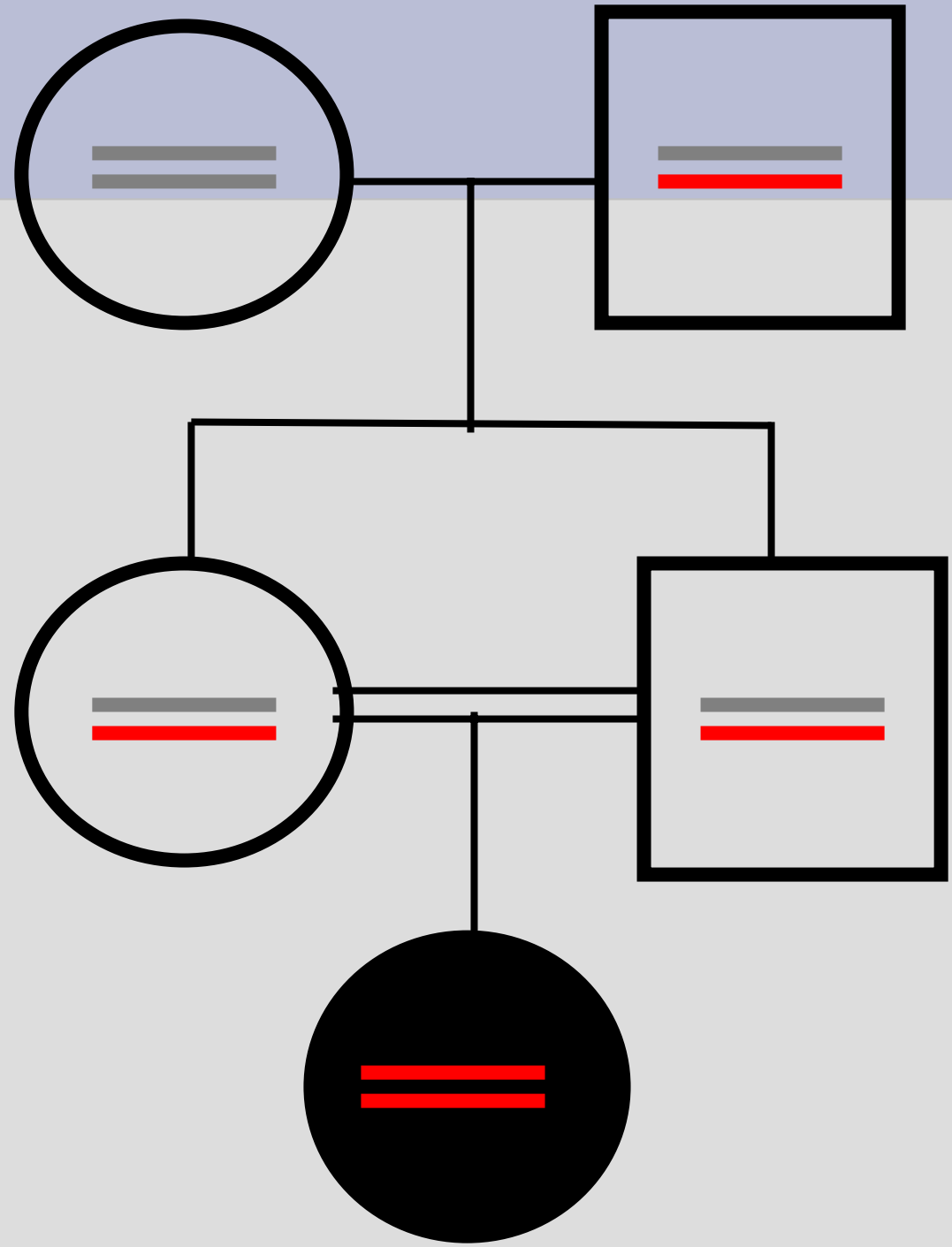


p^2

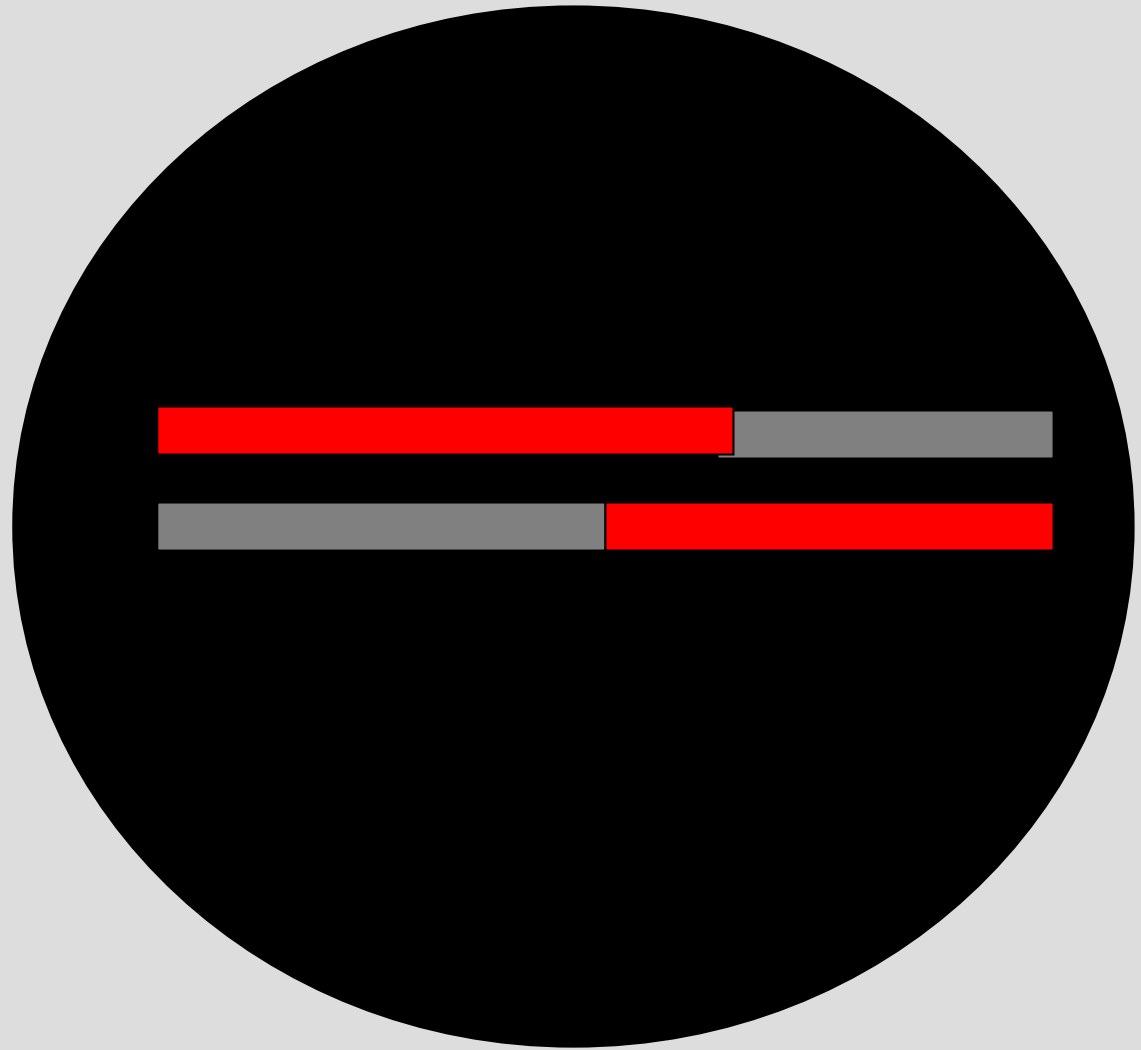
- Hardy-Weinberg Equilibrium:
 - indicator of random mating

Identity by Descent

- Same chromosome region transmitted through parallel lineages

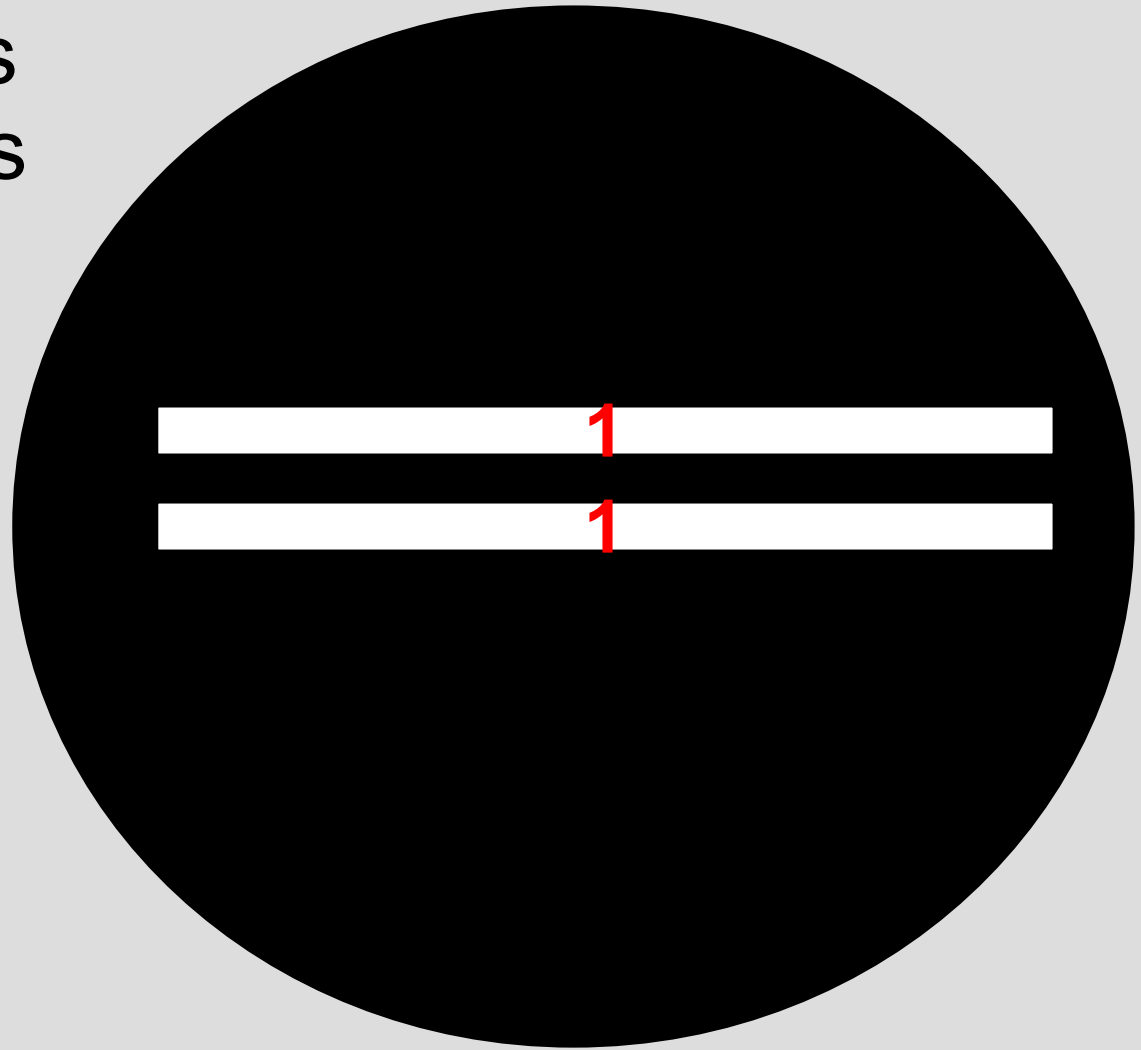


Changes in IBD



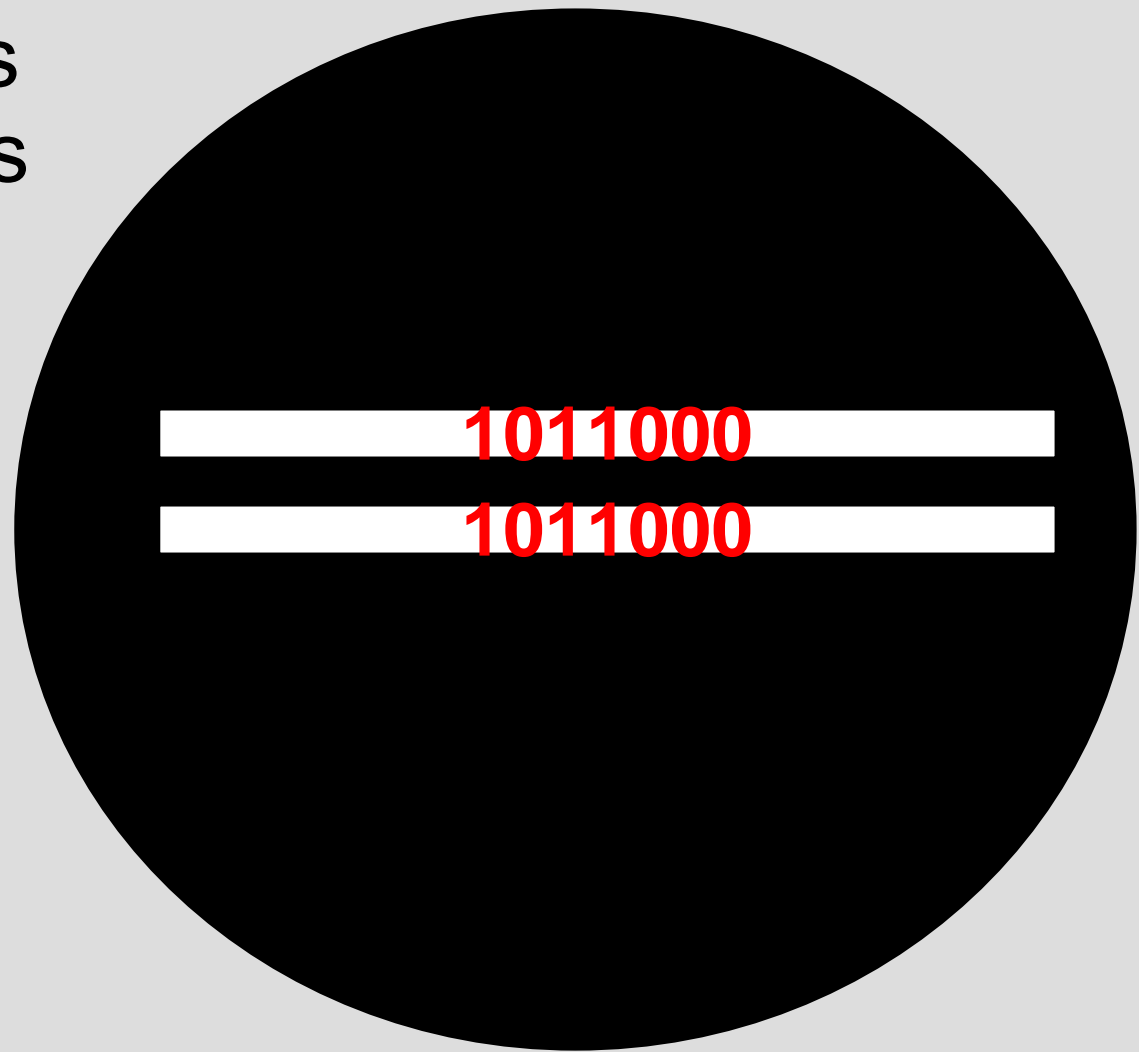
Identity by State

- Observation is a homozygous genotype



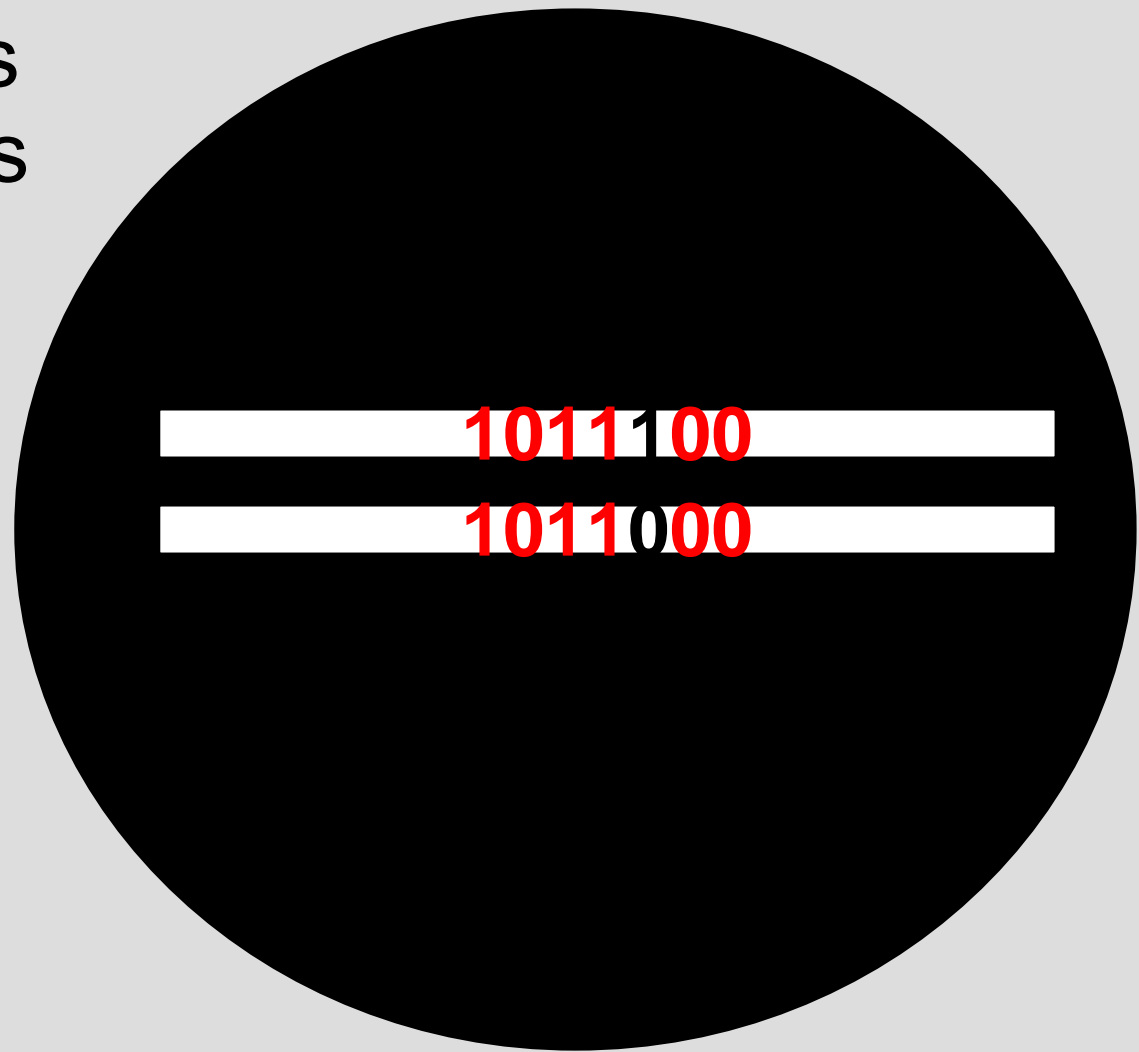
Identity by State

- Observation is a homozygous genotype
- Across a region



Identity by State

- Observation is a homozygous genotype
- Across a region
- With errors

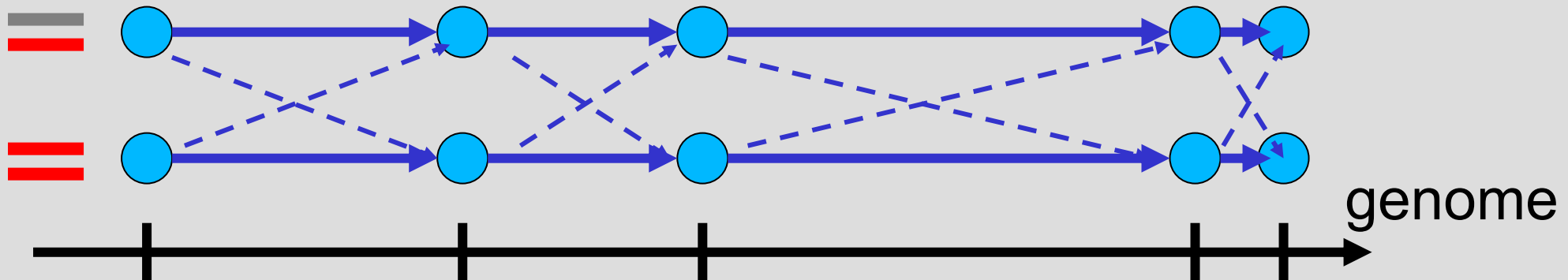


Linkage Analysis

- Homozygosity mapping for rare recessives
 - Identity by state/descent
 - Probabilistic model
- The general case of linkage analysis
 - Lander-Green
 - Elston-Stewart

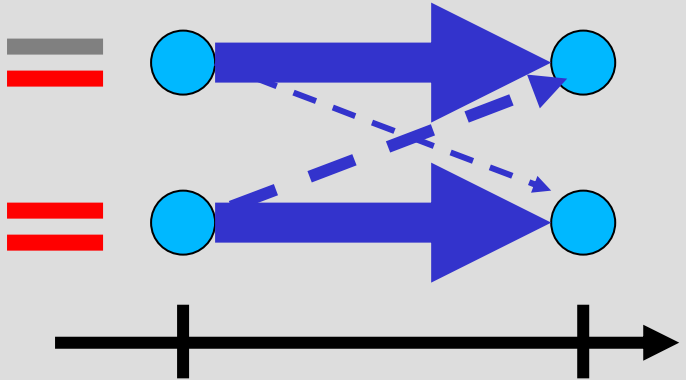
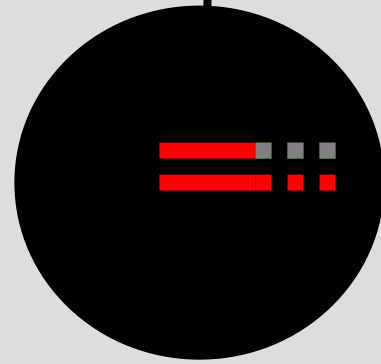
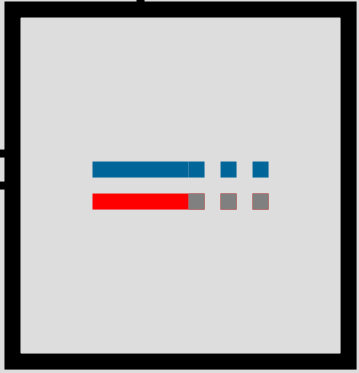
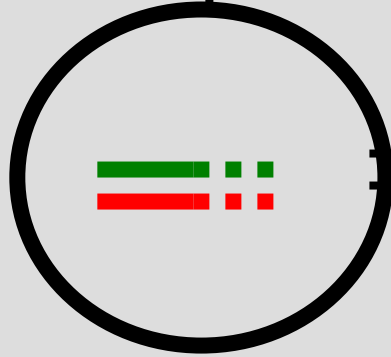
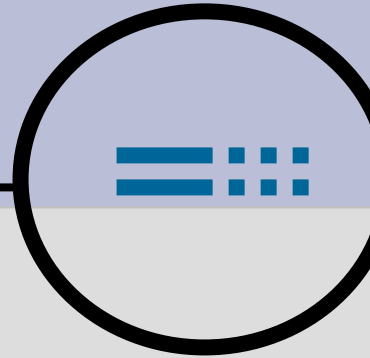
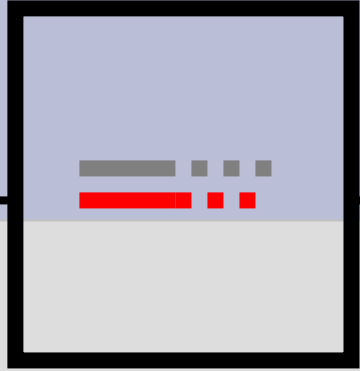
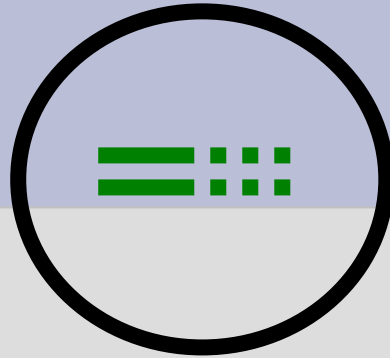
General Framework

- States:
 - IBD Sharing \times Markers
- Transition:



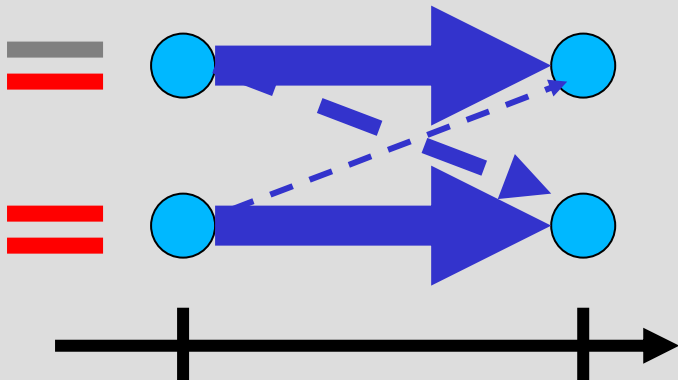
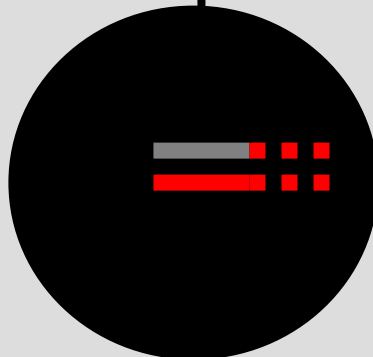
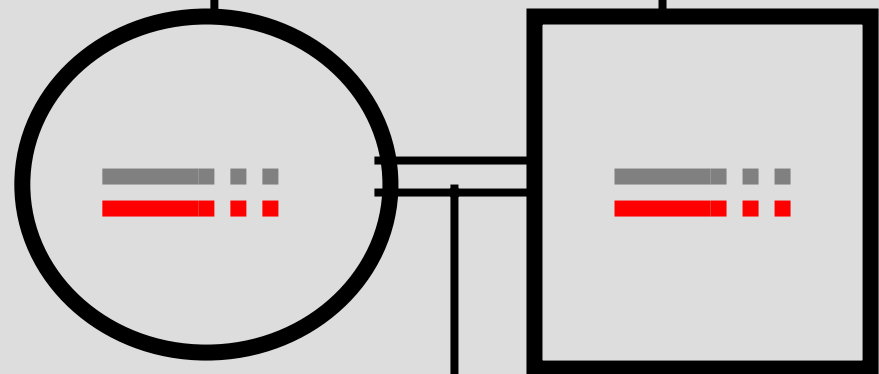
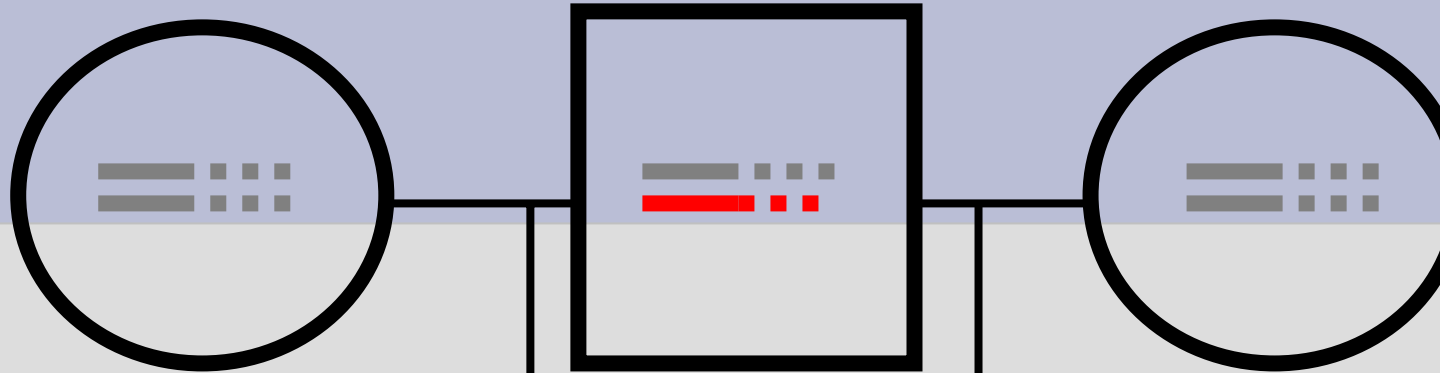
HMM

- States:
 - IBD Sharing \times Markers
- Transition:
 - From sharing:
 - rL for each meiosis:
 - $4rL$



HMM

- States:
 - IBD Sharing \times Markers
- Transition:
 - From sharing: $4rL_j$
 - To sharing: $\frac{3}{4} rL_j$



Emission

- Symbols:

{00,Het,11}

- If not IBD:

$$\Pr(00) = p^2$$

$$\Pr(\text{HET}) = 2pq$$

$$\Pr(11) = q^2$$

- If IBD:

$$\Pr(00) = p$$

$$\Pr(\text{HET}) = 0$$

$$\Pr(11) = q$$

Emission

- Symbols:

{00,Het,11}

- If not IBD:

$$\Pr(00) = p^2(1-3\varepsilon)+\varepsilon$$

$$\Pr(\text{HET}) = 2pq(1-3\varepsilon)+\varepsilon$$

$$\Pr(11) = q^2(1-3\varepsilon)+\varepsilon$$

- If IBD:

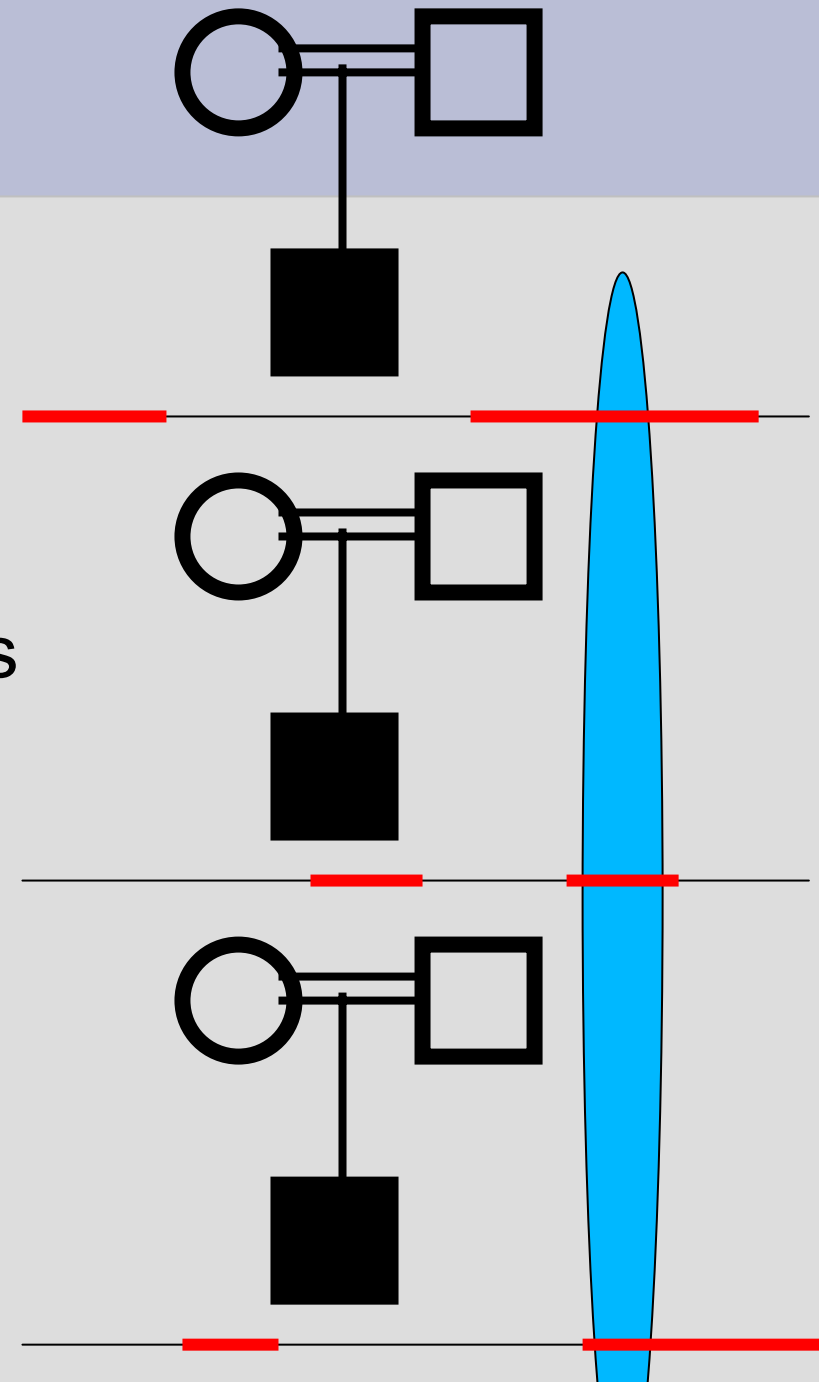
$$\Pr(00) = p(1-3\varepsilon)+\varepsilon$$

$$\Pr(\text{HET}) = 0(1-3\varepsilon)+\varepsilon$$

$$\Pr(11) = q(1-3\varepsilon)+\varepsilon$$

Multiple Families

- Null hypothesis:
Independent IBD
- Alternative:
Same region IBD in all families



Linkage Analysis

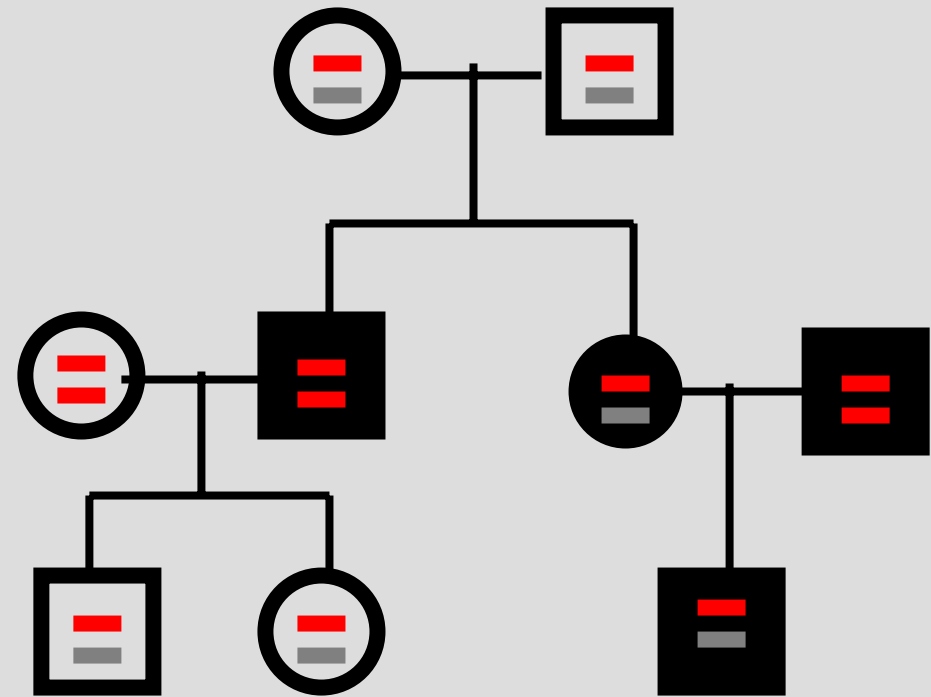
- Homozygosity mapping for rare recessives
 - Identity by state/descent
 - Probabilistic model
- The general case of linkage analysis
 - Lander-Green
 - Elston-Stewart

Generalizations

- Non-deterministic, arbitrary effect
 - *Pentrance* of genotype G :
 $f_G = \mathbf{Pr}(\text{Affected} \mid G)$
Recessive: $f_{\text{het}} = f_{00}$
Dominant: $f_{\text{het}} = f_{11}$
- General pedigrees

If We Typed the Mutation...

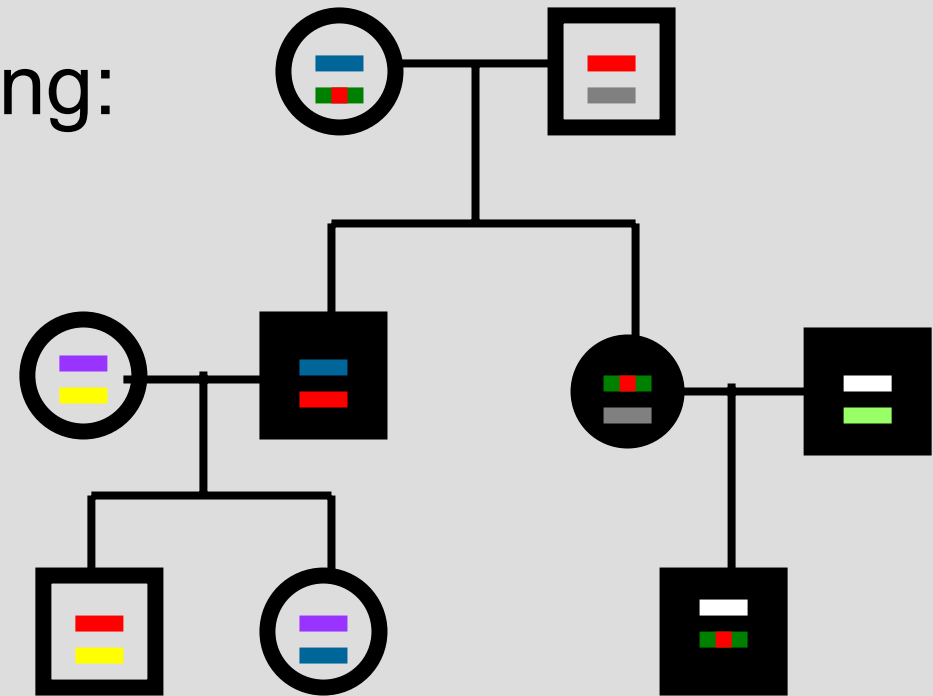
- Single point analysis
- Likelihood:



$$L(G) = \prod_{\text{affected } i} f_{G(i)} \prod_{\text{unaffected } i} (1 - f_{G(i)})$$

If We Knew the Meiosis Outcomes

- Relies on segment sharing:
Multi point analysis
- Likelihood depends on alleles a at founder chromosomes



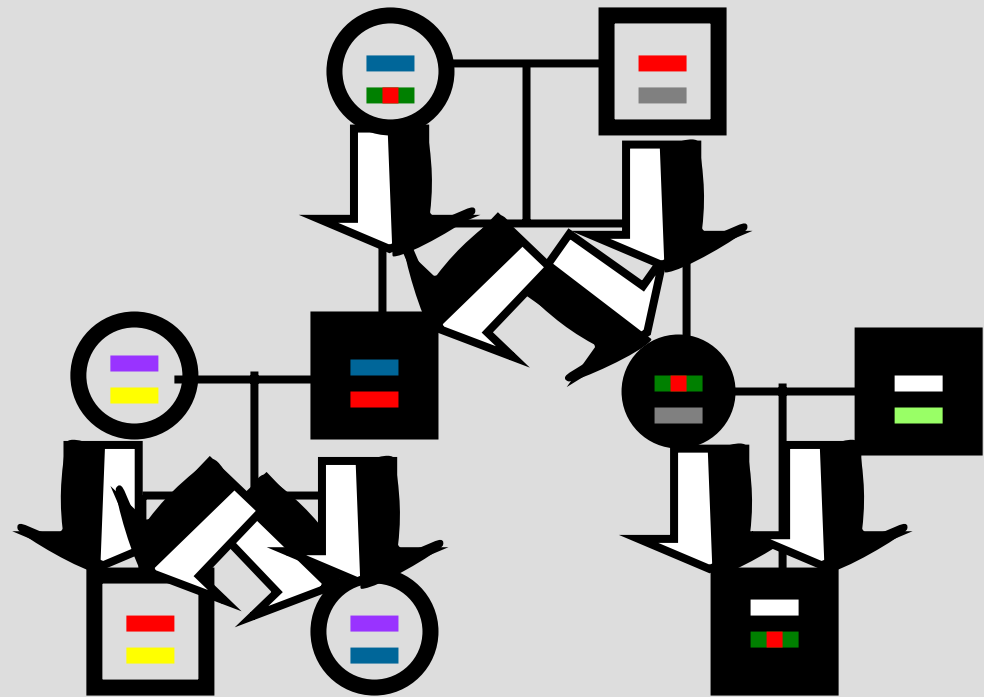
Allele frequencies

Penetrances

$$L(I) = \sum_a \left(\prod_{\text{founder } i} \Pr(a_i) \prod_{\text{any } i} \Pr(X_i | I, a) \right)$$

IBD BitVector, Descent Graph

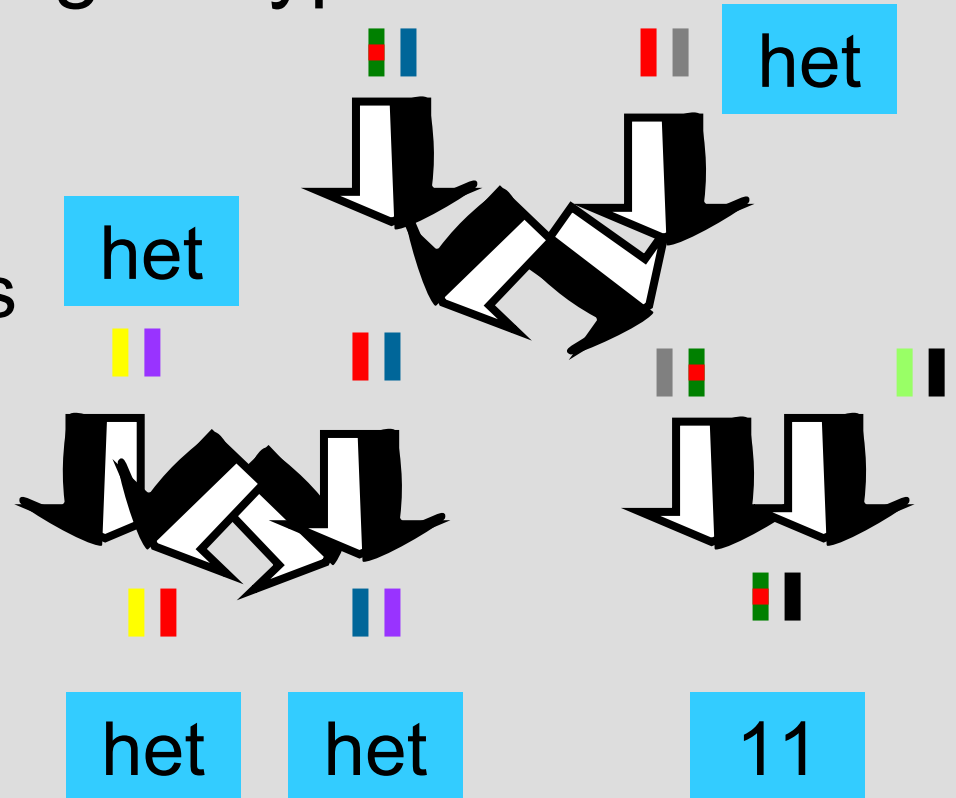
- Bit-entry per meiosis:
Which chromosome is transmitted
- Determines classes of same allele



Inheritance Vector

- Given IBD vector + some genotype data:

- Fixed founder alleles
- Variable alleles
- Don't-care founder alleles



- Viable configurations:

11 , 10101/01010

$$p^2 (p^3 q^2 + p^2 q^3)$$

- Inheritance vector lists all 2^{2n} probabilities

Inheritance Vectors as Emission Probabilities

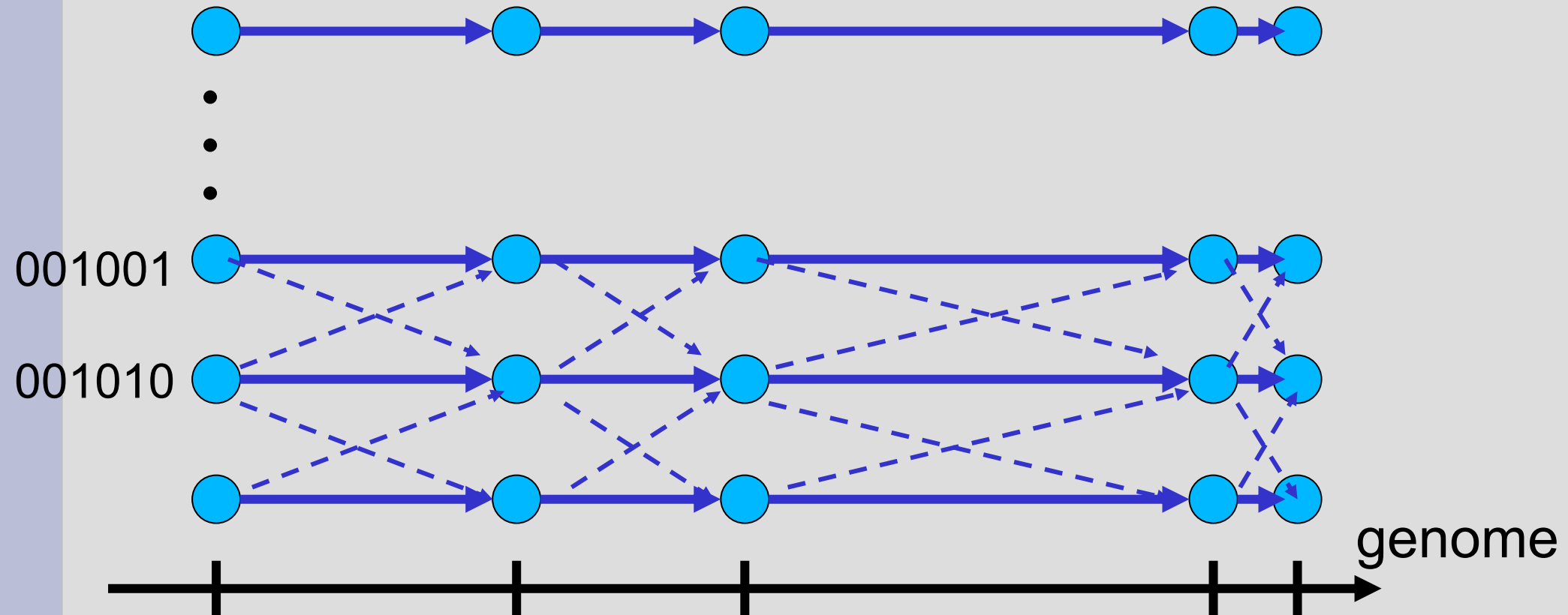
- Hidden state:
 - IBD BitVector

- Emitted observation:
 - Genotypes



HMM of Changing Inheritance Vectors

- Transition \cong a set of recombinations
 $\Pr(\text{specific } k \text{ recombinations}) = \theta^k(1-\theta)^{2n-k}$
where $\theta = rL_j$



Putting it Together

- Construct the Lander-Green HMM
- Compute $\Pr(G|I)$ for all I at all sites j
- Compute induced distribution of $\Pr(I|G)$
- Compute likelihood of phenotype under the alternative hypothesis for site j : $\sum \Pr(X|I)$

Limitations

- Parametric: assumes penetrances
- Complexity: $O(m2^{4n})$
Reductions:
 - $O(m(2n)2^{2n})$:
break transition into single meiosis events
 - Reduce n by inevitable symmetries, don't-cares

Non Parametric Linkage

- Summary statistic instead of penetrance model
- Example:

$$\sum_{\text{affected } i, j} IBD(i, j)$$

- Score by distribution under the null

Linkage Analysis

- Homozygosity mapping for rare recessives
 - Identity by state/descent
 - Probabilistic model
- The general case of linkage analysis
 - Lander-Green
 - Elston-Stewart

Pedigree Likelihood

- G_i : genotype vector for individual i
- Founders: $1..k$
- Non founders: $i \rightarrow m(i), f(i)$

Segregation+
recombination
probabilities

Founder priors
by Hardy-Weinberg

$$\prod_{\text{founder } i} \Pr(G_i)$$

$$L(X) = \sum_{G_1} \sum_{G_2} \cdots \sum_{G_k} \prod_{\text{nonfounder } i} \Pr(G_i | G_{m(i)}, G_{f(i)})$$

Penetrances

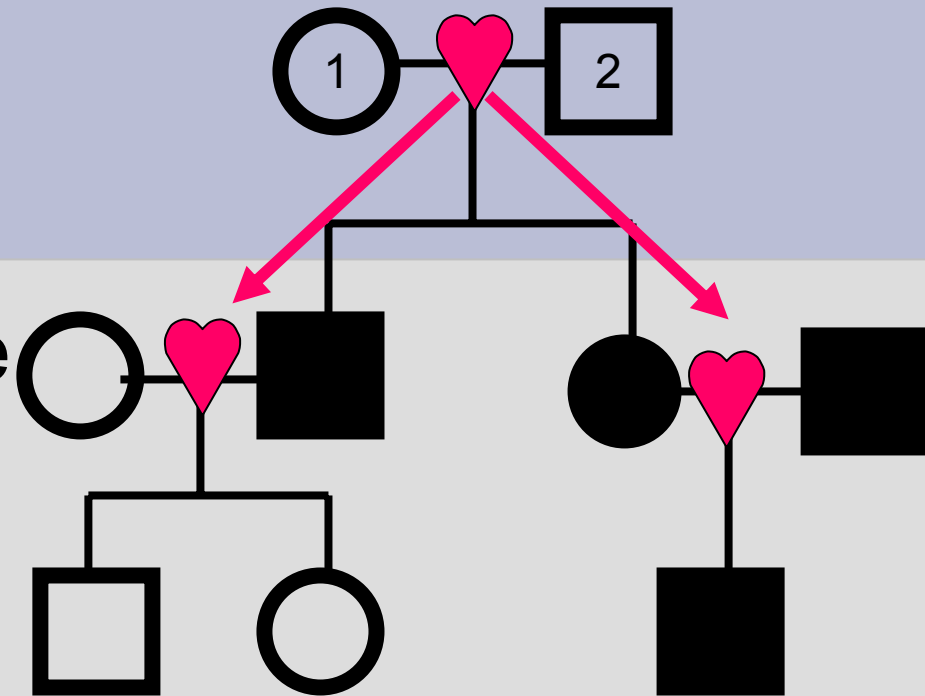
$$\prod_{\text{any } i} \Pr(X_i | G_i)$$

Double Exponential

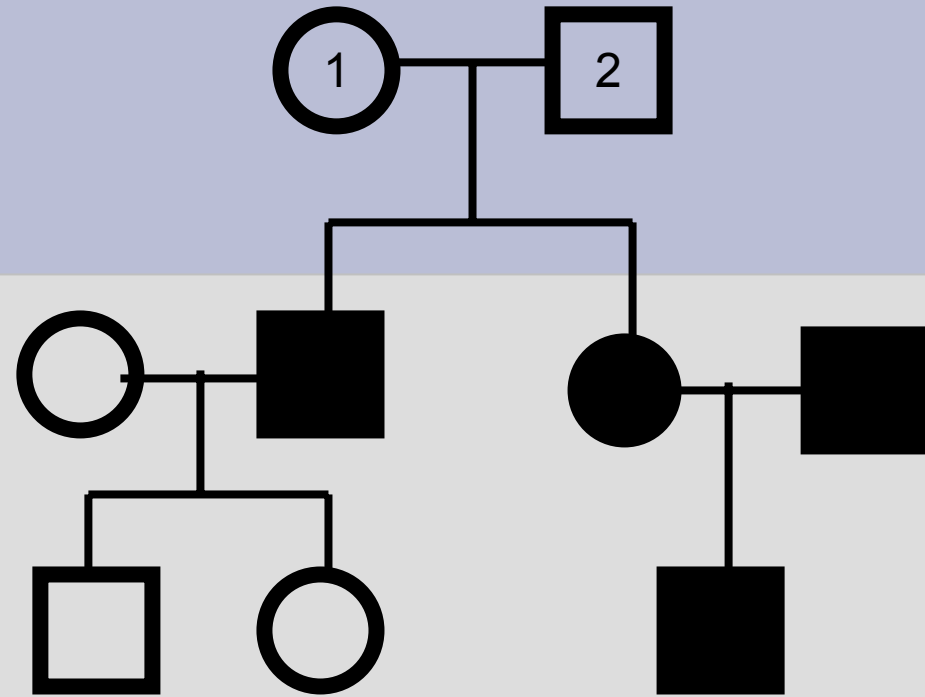
- Complexity disaster:
 - Exponential in #markers
 - Exponential in #individuals

Simple Pedigrees

- A founder in each couple
 - No inbreeding
 - Rooted tree of couples
- \forall founder f, s define subtree T_s



Rapid Summation



- Define conditional subtree likelihood:
 $C(X, s, G_s) = \Pr(X[T_s] | G_s)$

- Rearrange summation

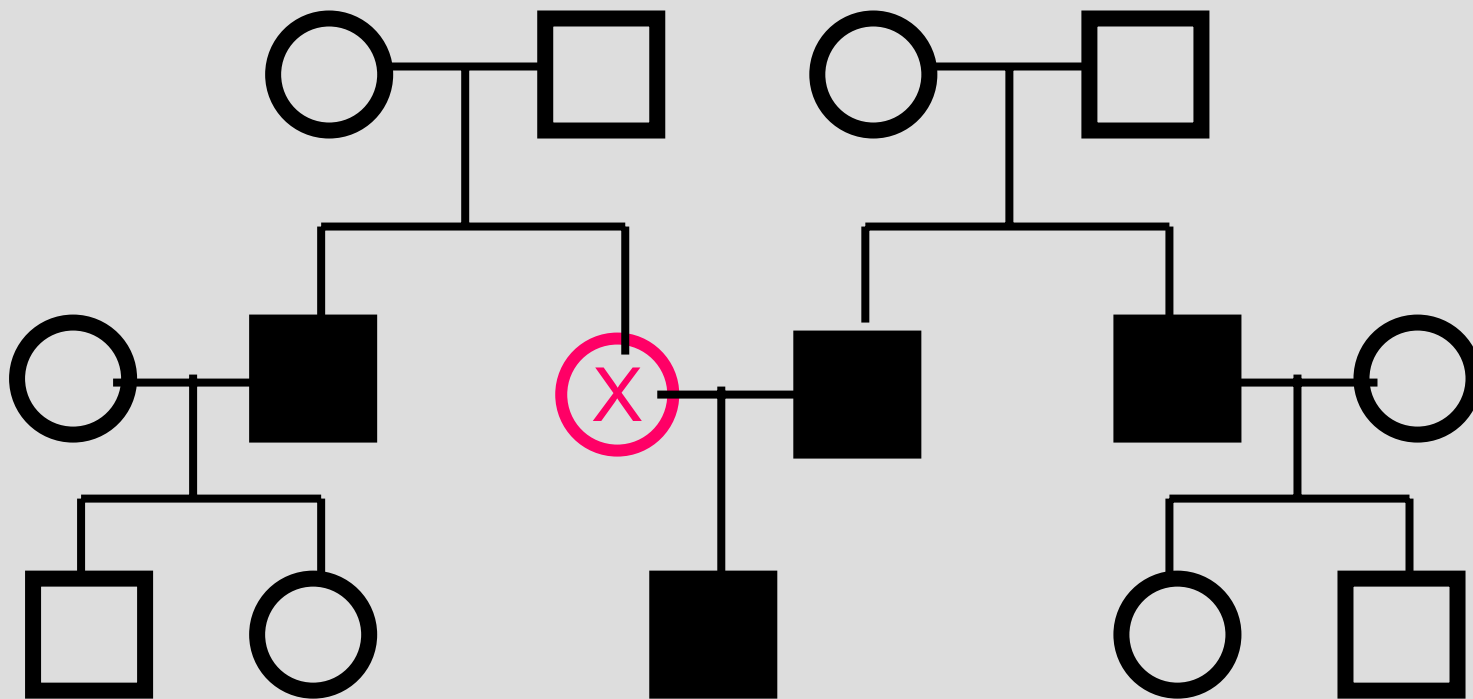
$$L(X) = \sum_{G_1} [\Pr(G_1) \Pr(X_1 | G_1) C(X, 1, G_1)]$$

- Recursively compute:

$$C(X, s, G_s) = \sum_{G_f} \left(\Pr(G_f) \Pr(X_f | G_f) \prod_{m(i)=1, f(i)=2} \sum_{G_i} \Pr(G_i | G_{m(i)}, G_{f(i)}) C(X, i, G_i) \right)$$

General Loopless Pedigrees

- Can work upwards as well, e.g.:
 $\Pr(\text{subtree upper-left of } X \mid G_x)$

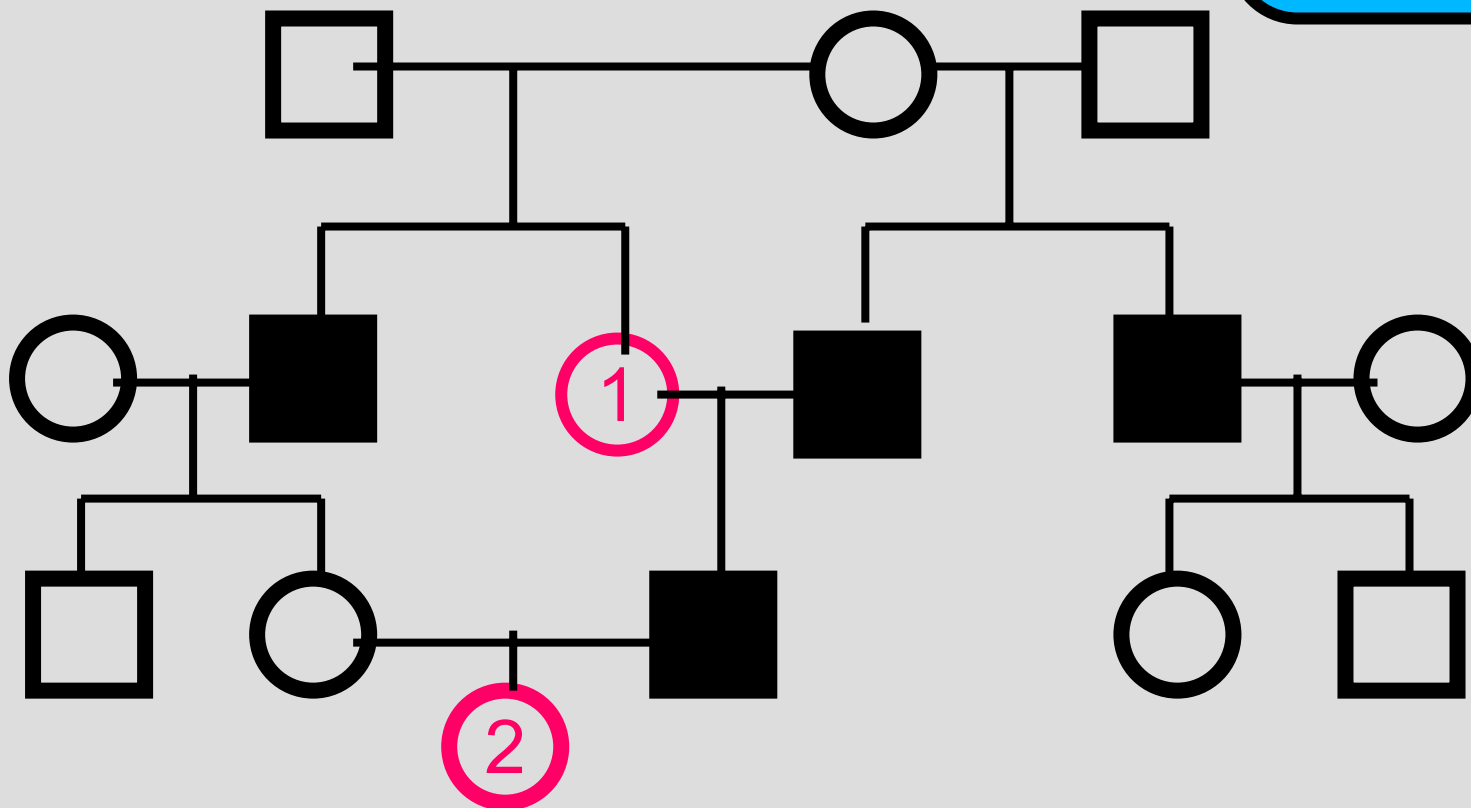


Handling Loops

- Exhaust loop-breakers:

$$L(X) = \sum_{G_1} \sum_{G_2} \cdots \sum_{G_k} L(\text{rest})$$

Exponential in:
#markers,
#loop breakers



Summary

- Homozygosity mapping for rare recessives
 - Probabilities for IBD/IBS
- Linkage analysis
 - Lander-Green across the chromosome
 - Elston-Stewart along the pedigree

Further Reading

- **Lander & Green**, Construction of multilocus genetic linkage maps in humans. Proc Natl Acad Sci U S A. 1987 Apr;84(8):2363-7
- **Kruglyak L, Daly MJ, Reeve-Daly MP, Lander ES** Parametric and nonparametric linkage analysis: a unified multipoint approach. Am J Hum Genet. 1996 Jun;58(6):1347-63.
- **Elston RC, Stewart J.** A general model for the genetic analysis of pedigree data. Hum Hered. 1971;21(6):523-42.
- <http://www.sph.umich.edu/csg/abecasis/class/>
Lessons 22-24

Extra Credit

1. Given a population out of Hardy Weinberg Equilibrium, how many generations of random mating are needed to bring it back to equilibrium?
2. Would you prefer to homozygosity-map using 5 sib-couples? 5 3rd cousins? 5 10th cousins?
3. Given trait with 1% prevalence, and a single, 4% causal allele, with penetrances, f_{het} and f_{hom} , what is the relative increase in risk to children of an affected individual? Siblings? Half siblings? Niblings?

Project Suggestion

- Implement homozygosity mapping
 - Assume you have a quantitative recessive trait ($\mu_{11} \gg \mu_{01} = \mu_{00}$) known for many contemporary individuals
 - Assume you have a large pedigree, with occasional inbreeding loops of arbitrary size
 - Assume data on 10^5 - 10^6 SNPs for many contemporary individuals