

INTRODUCTION TO GENE MAPPING METHODS

March 20, 2008

Adele A. Mitchell, PhD

Department of Genetics & Genomic Sciences

Mount Sinai School of Medicine

adele.mitchell@mssm.edu

OUTLINE

- BACKGROUND AND HISTORY
- LINKAGE ANALYSIS
- ASSOCIATION ANALYSIS
- TRANSMISSION DISTORTION TESTING

OUTLINE

- **BACKGROUND AND HISTORY**
- LINKAGE ANALYSIS
- ASSOCIATION ANALYSIS
- TRANSMISSION DISTORTION TESTING

HISTORY OF GENE MAPPING

YEAR

1900 — Mendel's work rediscovered (originally published 1865)

Mendel's Laws

-Law of Segregation

-Law of Independent Assortment

1950 —

2000 —

HISTORY OF GENE MAPPING

YEAR

1900 — Mendel's work rediscovered (originally published 1865)

1904-1928: Morgan's fly room with students Sturtevant, Muller & Bridges

Major Discoveries

-Complementation

-X-linked inheritance

-Linkage mapping

-Recessive lethality

-Nondisjunction

-“String of Beads” model
of genes on chromosomes

1950

2000

HISTORY OF GENE MAPPING

YEAR

1900 — Mendel's work rediscovered (originally published 1865)

1904-1928: Morgan's fly room with students Sturtevant, Muller & Bridges

1922: Sir R.A. Fisher

1935: Penrose

Key Statistical Ideas

Fisher: Deriving continuous traits from discrete genes

Penrose: Expected allele-sharing among affected sib pairs (ASP)

1947: Haldane&Smith

Haldane&Smith: Converting recombination fraction to physical distance

1950 — 1955: Morton

Morton: Testing for linkage in humans

2000 —

HISTORY OF GENE MAPPING

YEAR

1900 — Mendel's work rediscovered (originally published 1865)

1904-1928: Morgan's fly room with students Sturtevant, Muller & Bridges

1922: Sir R.A. Fisher

1935: Penrose

1947: Haldane&Smith

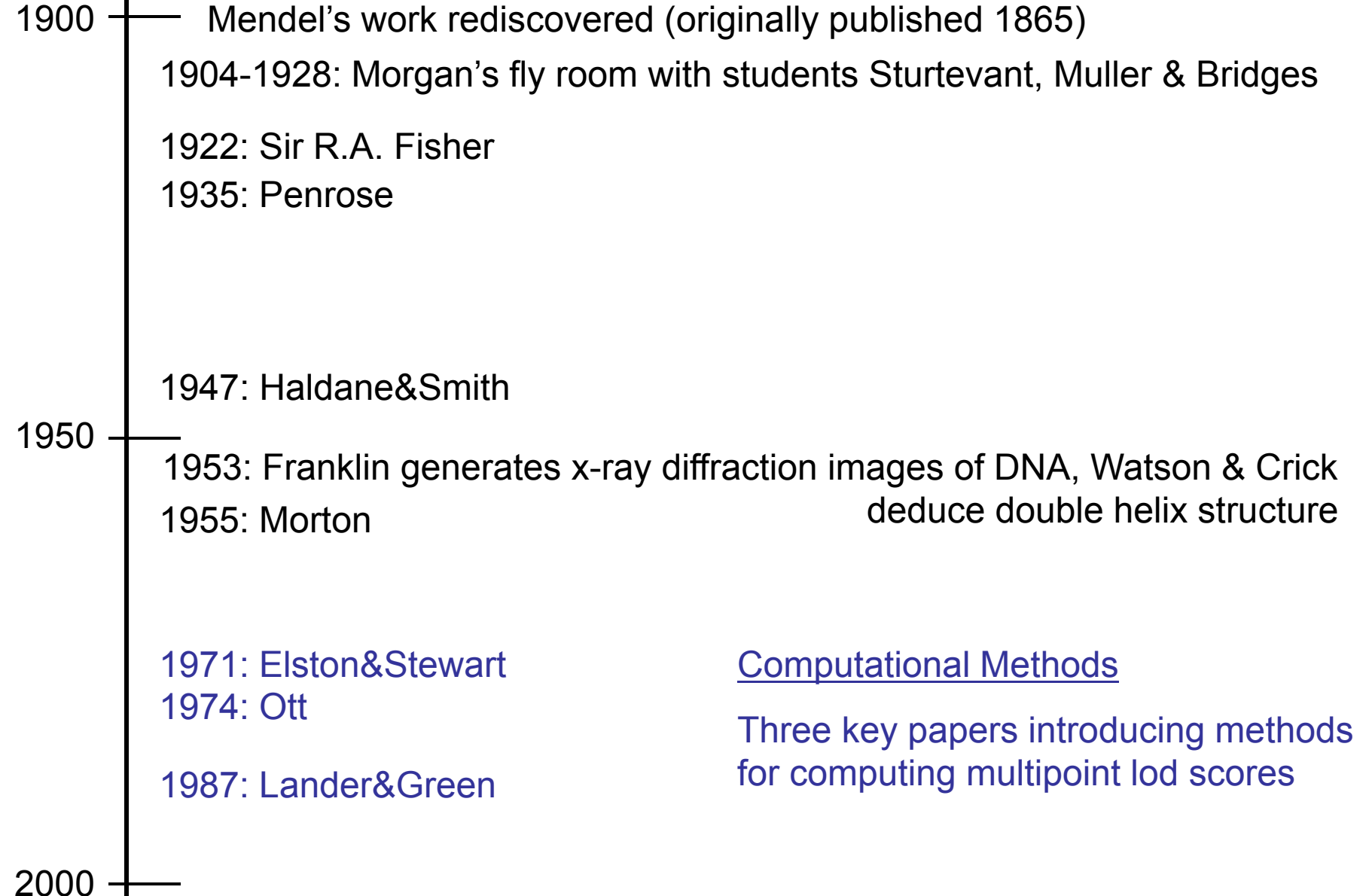
1950 — 1953: Franklin generates x-ray diffraction images of DNA, Watson & Crick deduce double helix structure

1955: Morton

2000 —

HISTORY OF GENE MAPPING

YEAR

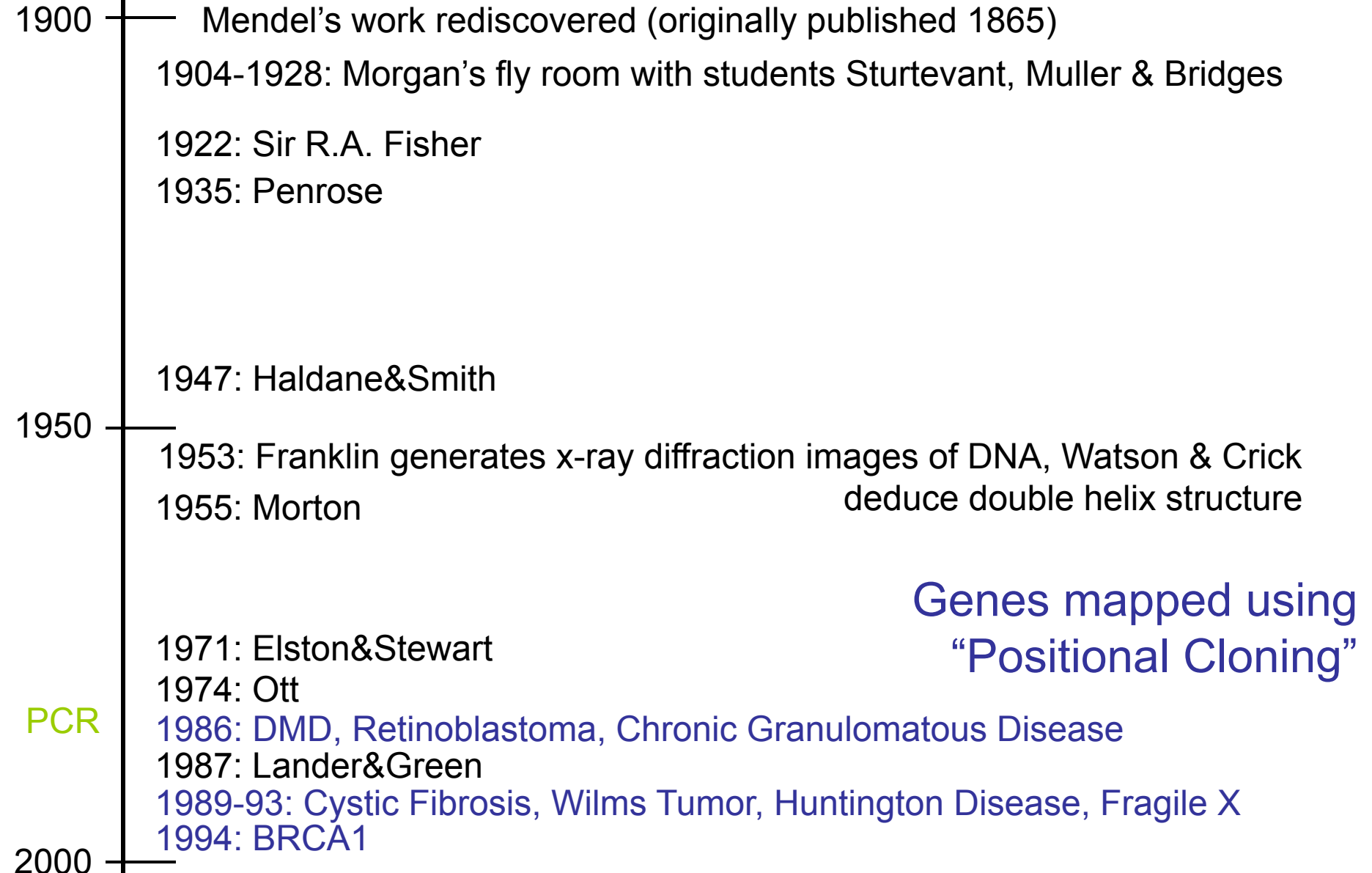


Computational Methods

Three key papers introducing methods for computing multipoint lod scores

HISTORY OF GENE MAPPING

YEAR



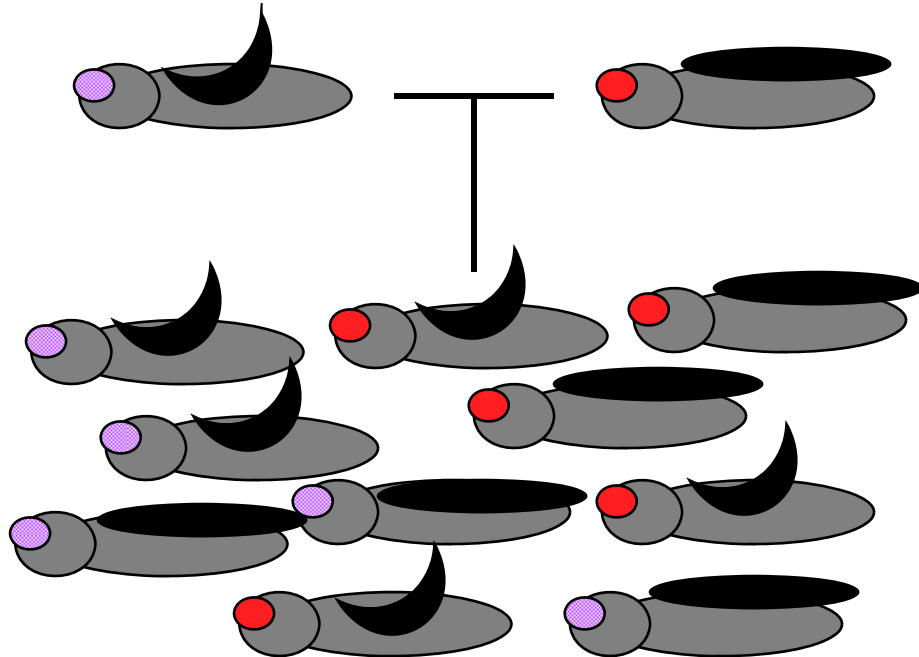
Genes mapped using
"Positional Cloning"

PCR

LINKAGE IS NON-INDEPENDENT MEIOTIC SEGREGATION

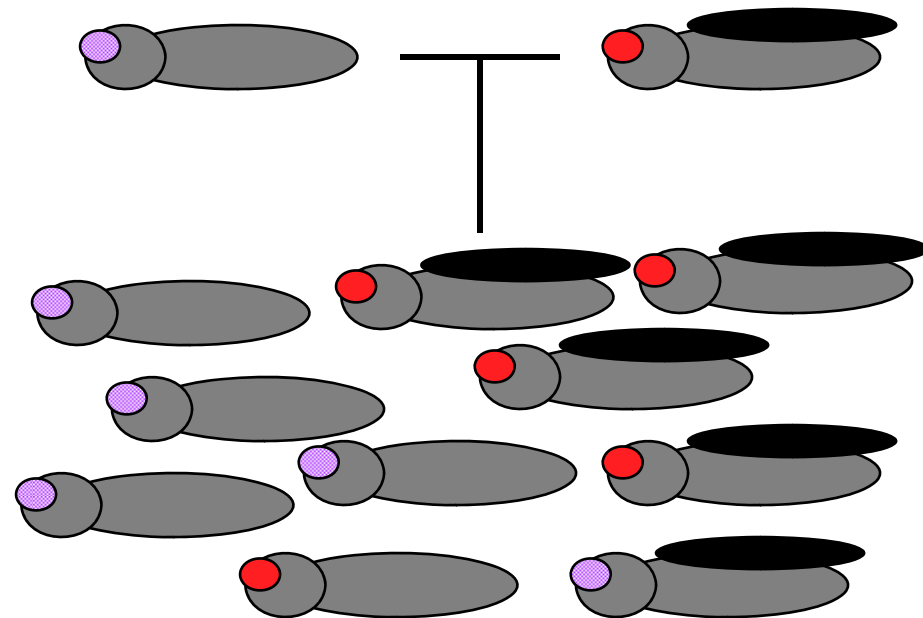
Genes on different chromosomes
segregate independently

They are “unlinked”



Genes that are physically close
tend to be inherited together

They are “linked”



ASSOCIATION IS CORRELATION BETWEEN ALLELES AT DISTINCT LOCI

- SIGNATURES OF ASSOCIATION

- If knowing an individual's genotype at locus 1 helps predict their genotype at locus 2, they are associated
- If haplotype frequencies differ from what would be predicted by allele frequencies, they are associated

- CAUSES OF ASSOCIATION

- Population structure
- Interaction between alleles
- Linkage disequilibrium
- Random chance

ASSOCIATION IS CORRELATION BETWEEN ALLELES AT DISTINCT LOCI

- **SIGNATURES OF ASSOCIATION**

- If knowing an individual's genotype at locus 1 helps predict their genotype at locus 2, they are associated
- If haplotype frequencies differ from what would be predicted from allele frequencies, they are associated

- CAUSES OF ASSOCIATION

- Population structure
- Interaction between alleles
- Linkage disequilibrium
- Random chance

ASSOCIATION EXAMPLE

- I genotyped 100 people for three SNPs in my favorite gene
- Genotypes for SNP 1 are shown below
- What is the most likely SNP 1 genotype for a randomly-chosen person from the sample?

<u>AA</u>	<u>Aa</u>	<u>aa</u>	<u>Total</u>
64	33	3	100

ASSOCIATION EXAMPLE

- I genotyped 100 people for three SNPs in my favorite gene
- Genotypes for SNP 1 are shown below
- What is the most likely SNP 1 genotype for a randomly-chosen person from the sample?

<u>AA</u>	<u>Aa</u>	<u>aa</u>	<u>Total</u>
64	33	3	100

Answer: the most likely genotype is AA (with probability 0.64)

ASSOCIATION EXAMPLE

- Genotypes at SNP1 and SNP2 are shown below
- SNP1 and SNP2 genotypes are correlated
- Knowing a person's genotype at SNP2 can give information about their genotype at SNP1

		SNP 1			Total
		AA	Aa	aa	
SNP 2	BB	60	3	0	63
	Bb	3	29	2	34
	bb	1	1	1	3
Total:		64	33	3	100

ASSOCIATION EXAMPLE

- Recall that the most likely SNP1 genotype of a randomly chosen person was AA (probability 0.64)
- What is the probability of SNP1 genotype AA for a person with SNP2 genotype BB? Bb? bb?

		SNP 1			Total
		AA	Aa	aa	
SNP 2	BB	60	3	0	63
	Bb	3	29	2	34
	bb	1	1	1	3
Total:		64	33	3	100

ASSOCIATION EXAMPLE

- Probability of SNP1 genotype AA for a person with:
 - Unknown SNP2 genotype is 0.64
 - BB at SNP 2 is $60/63 = 0.95$
 - Bb at SNP 2 is $3/34 = 0.09$
 - bb at SNP2 is $1/3 = 0.33$

		SNP 1			Total
		AA	Aa	aa	
SNP 2	BB	60	3	0	63
	Bb	3	29	2	34
	bb	1	1	1	3
Total:		64	33	3	100

ASSOCIATION EXAMPLE

- Probability of SNP1 genotype AA for a person with:
 - Unknown SNP2 genotype is 0.64
 - BB at SNP 2 is $60/63 = 0.95$
 - Bb at SNP 2 is $3/34 = 0.09$
 - bb at SNP2 is $1/3 = 0.33$

Because SNP1 genotype frequencies differ for each SNP2 genotype, we say that there is association between the alleles at these loci

		SNP 1			Total
		AA	Aa	aa	
SNP 2	BB	60	3	0	63
	Bb	3	29	2	34
	bb	1	1	1	3
Total:		64	33	3	100

ASSOCIATION EXAMPLE

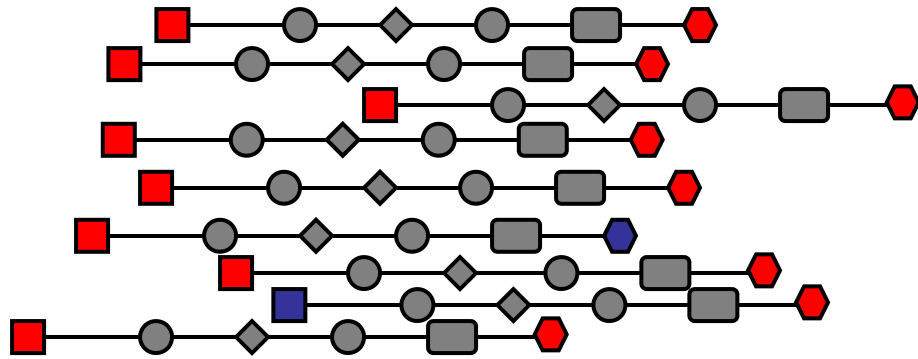
- Genotypes at SNP1 and SNP3 are shown below
- Is there correlation between SNP1 and SNP3 genotypes in this sample?
- Does knowing a person's SNP3 genotype give information about their SNP1 genotype?

		SNP 1			Total
		AA	Aa	aa	
SNP 3	CC	15	7	1	23
	Cc	35	19	2	56
	cc	14	7	0	21
	Total:	64	33	3	100

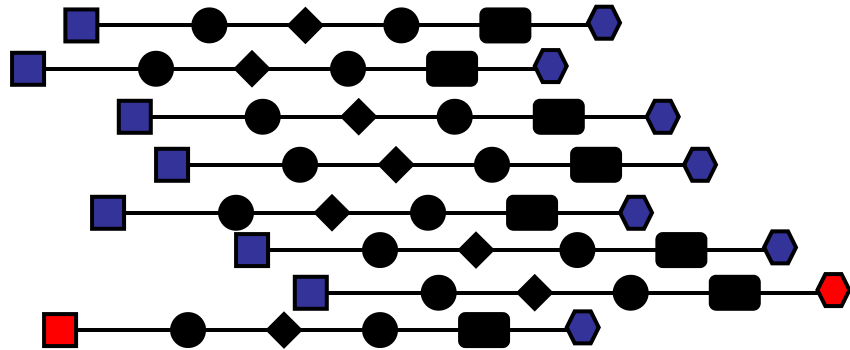
ASSOCIATION IS CORRELATION BETWEEN ALLELES AT DISTINCT LOCI

- SIGNATURES OF ASSOCIATION
 - If knowing an individual's genotype at locus 1 helps predict their genotype at locus 2, they are associated
 - If haplotype frequencies differ from what would be predicted by allele frequencies, they are associated
- **CAUSES OF ASSOCIATION**
 - Population structure
 - Interaction between alleles
 - Linkage disequilibrium
 - Random chance

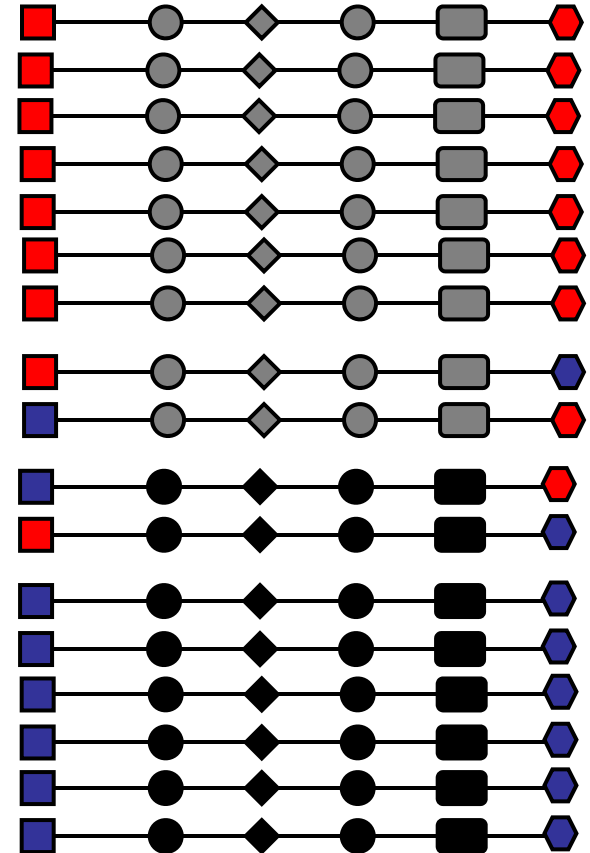
ASSOCIATION CAN RESULT WHEN TWO SEPARATE POPULATIONS MIX



In population A, the **red allele** is common at SNP1 and SNP6

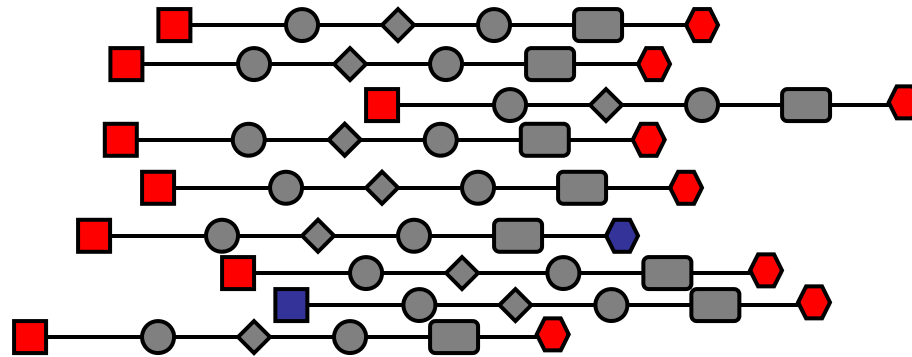


In population B, the **blue allele** is common at SNP1 and SNP6



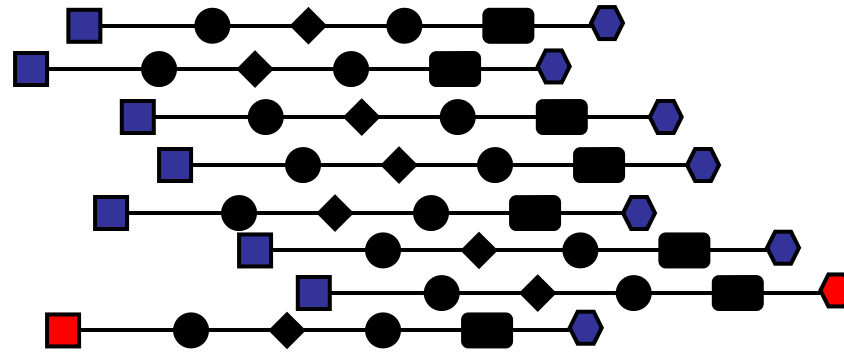
In a mixed sample, **red-red** and **blue-blue** are more common than **red-blue** or **blue-red**

ASSOCIATION CAN RESULT WHEN TWO SEPARATE POPULATIONS MIX



There is no association between the alleles at SNP1 and SNP6 in **Population A** because knowing the allele at SNP1 does not help to predict the allele at SNP 6

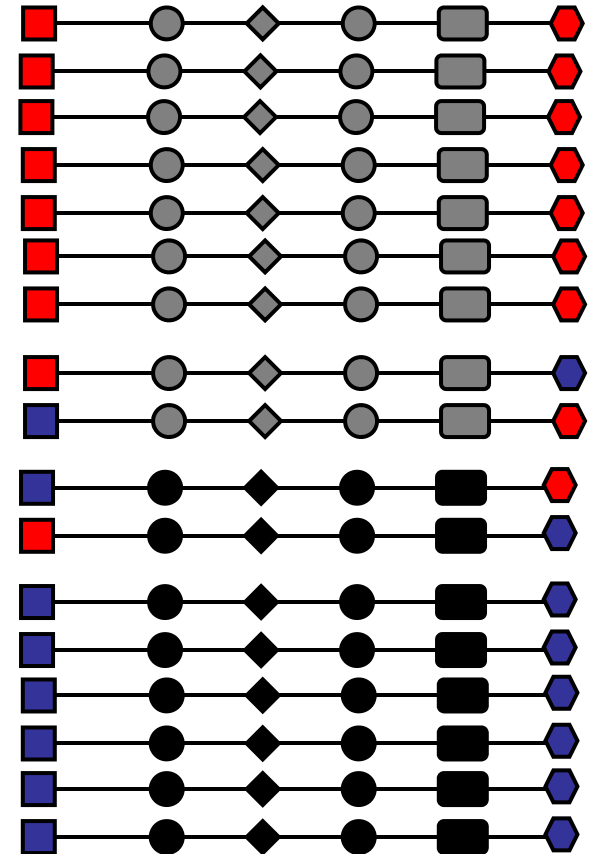
ASSOCIATION CAN RESULT WHEN TWO SEPARATE POPULATIONS MIX



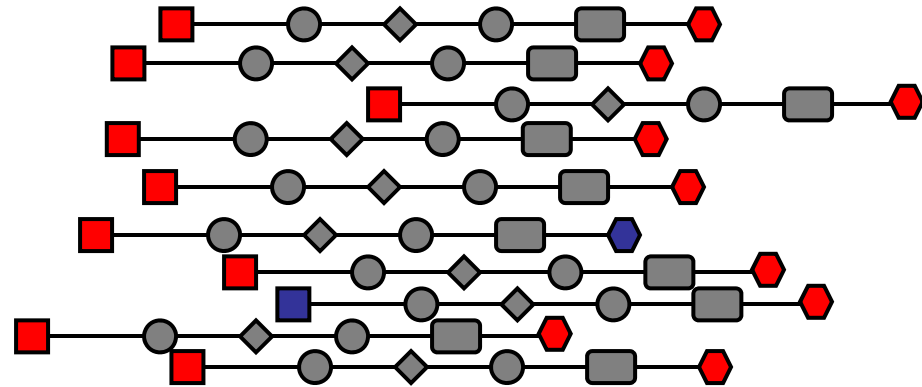
There is no association between the alleles at SNP1 and SNP6 in **Population B** because knowing the allele at SNP1 does not help to predict the allele at SNP 6

ASSOCIATION CAN RESULT WHEN TWO SEPARATE POPULATIONS MIX

- Recall that there was no association between SNP 1 and SNP 6 alleles in either individual population
- In the combined sample, the allele at SNP 1 gives information about SNP 6
- So, there is association between the alleles at these loci in the combined sample



EXAMPLE: ASSOCIATION DUE TO POPULATION MIXTURE

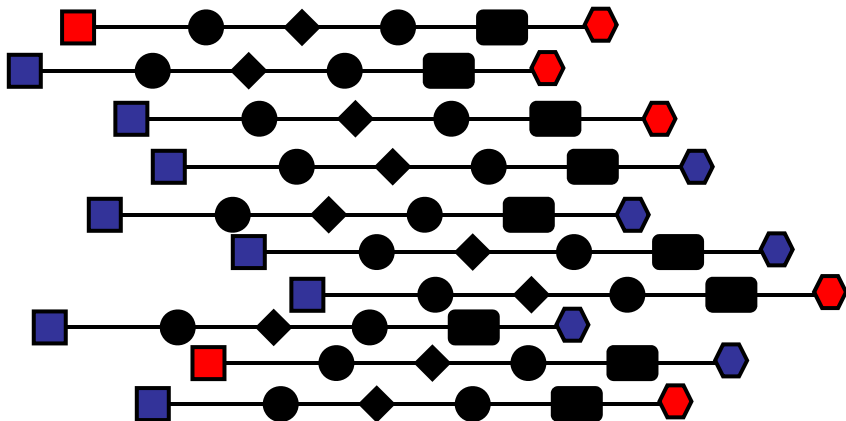


In population A:

10% of chromosomes are blue at SNP1

10% of chromosomes are blue at SNP6

1% of chromosomes are blue-blue



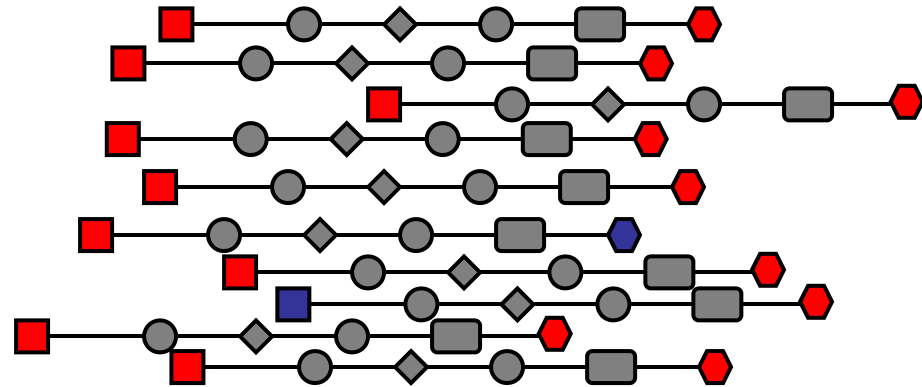
In population B:

20% of chromosomes are red at SNP1

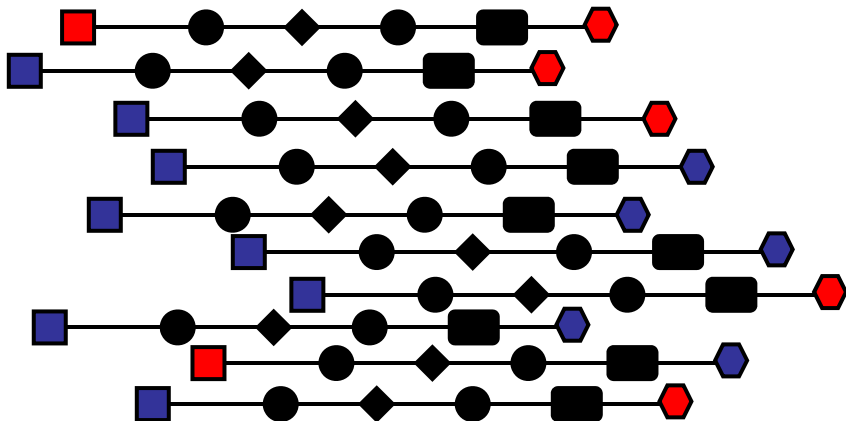
50% of chromosomes are red at SNP6

10% of chromosomes are red-red

EXAMPLE: ASSOCIATION DUE TO POPULATION MIXTURE



- Is there association between alleles at these loci in either individual population?



- In the combined sample?

SNP1 AND SNP6 ARE ASSOCIATED IN COMBINED SAMPLE

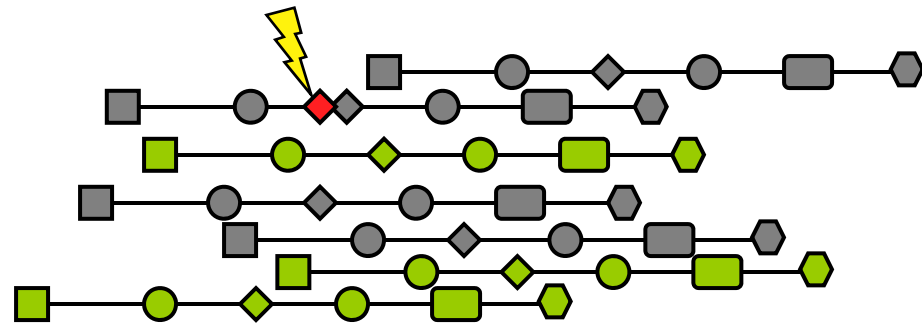
Expected allele counts for 100 chromosomes from each population

<u>SNP 1</u>	<u>SNP 6</u>	<u>POP 1</u>	<u>POP 2</u>	<u>COMBINED</u>	<u>PERCENT</u>
BLUE	BLUE	1	40	41	20.5%
BLUE	RED	9	40	49	24.5%
RED	BLUE	9	10	19	9.5%
RED	RED	81	10	91	45.5%

If a chromosome has blue at SNP1, the probability of blue at SNP6 is about the same as the probability of red at SNP6.

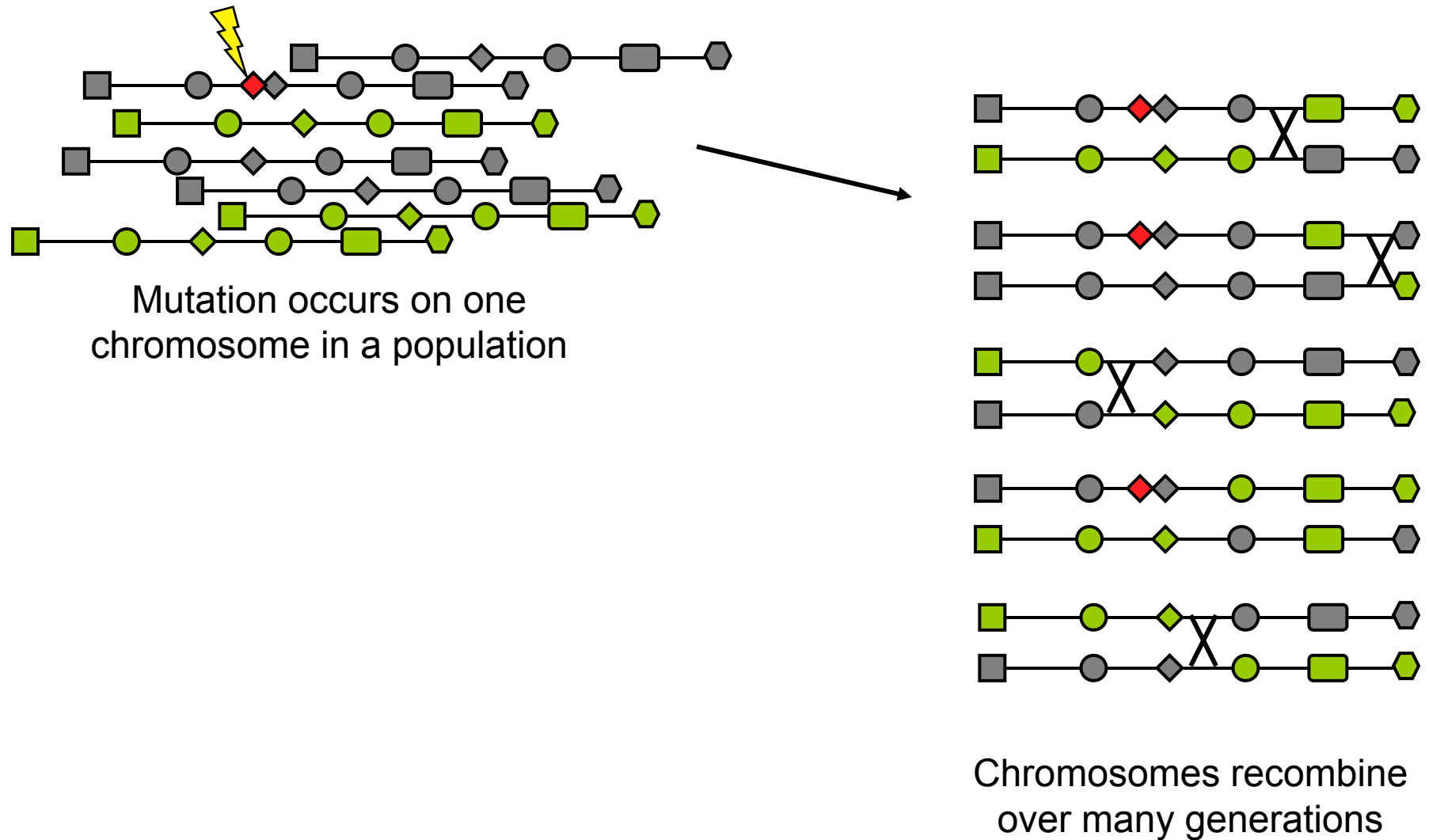
But, if a chromosome has red at SNP1, it is about 5 times more likely to have red than blue at SNP6.

ASSOCIATION CAN RESULT FROM LINKAGE AFTER MANY GENERATIONS

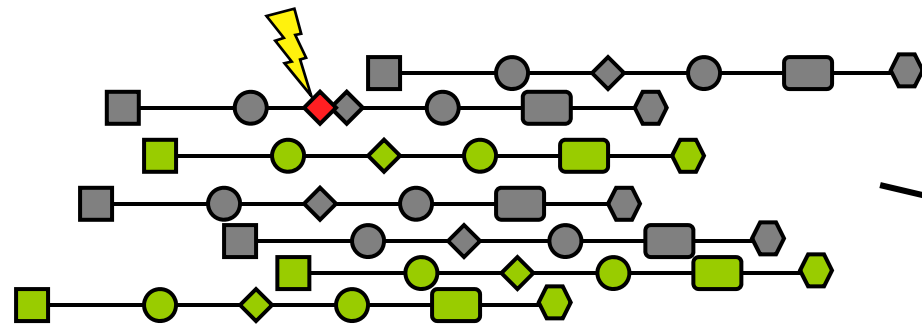


Mutation occurs on one chromosome in a population

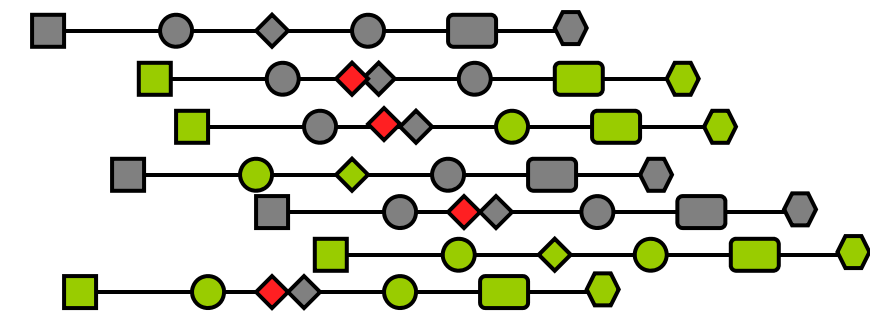
ASSOCIATION CAN RESULT FROM LINKAGE AFTER MANY GENERATIONS



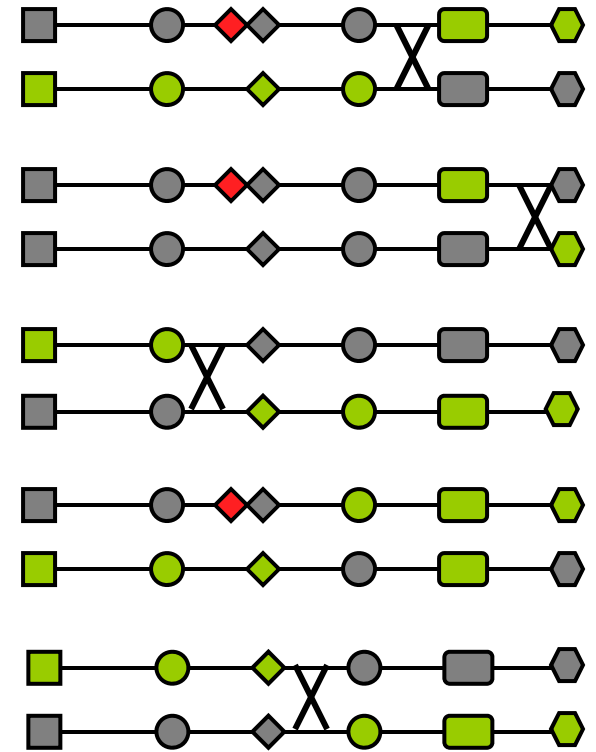
ASSOCIATION CAN RESULT FROM LINKAGE AFTER MANY GENERATIONS



Mutation occurs on one chromosome in a population



Markers near the mutation resemble the chromosome on which mutation occurred



Chromosomes recombine over many generations

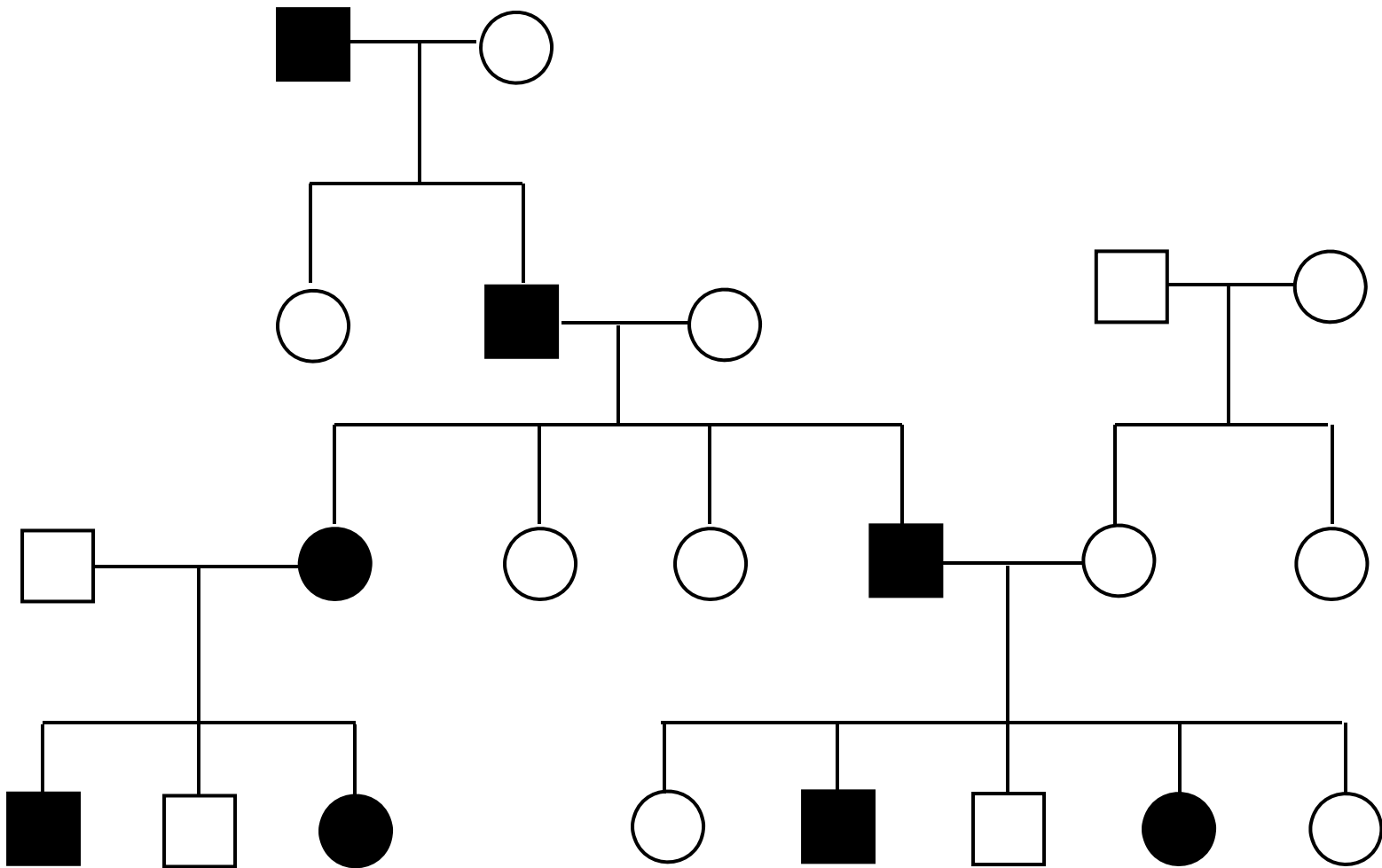
LINKAGE DISEQUILIBRIUM (LD)

- Association that is due to physical proximity of SNPs is called LINKAGE DISEQUILIBRIUM
- **WARNING! LD is a misleading name!**
 - “Linkage Disequilibrium” is not the same as linkage
 - LD requires both linkage and association
- Population-based association methods for gene mapping require LD between a marker allele and the risk allele at an unknown locus
- Linkage methods require physical proximity only

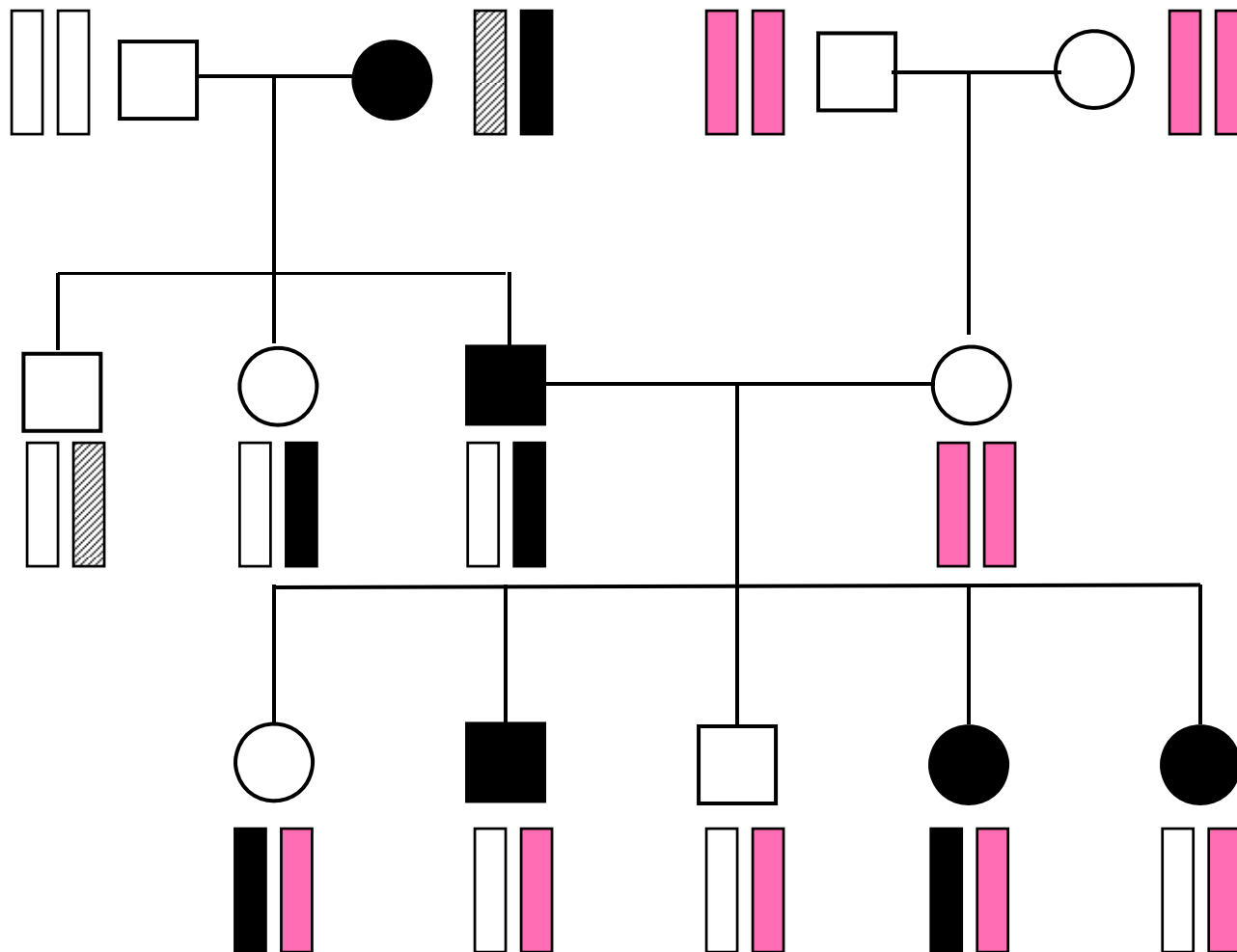
OUTLINE

- BACKGROUND AND HISTORY
- **LINKAGE ANALYSIS**
- ASSOCIATION ANALYSIS
- TRANSMISSION DISTORTION TESTING

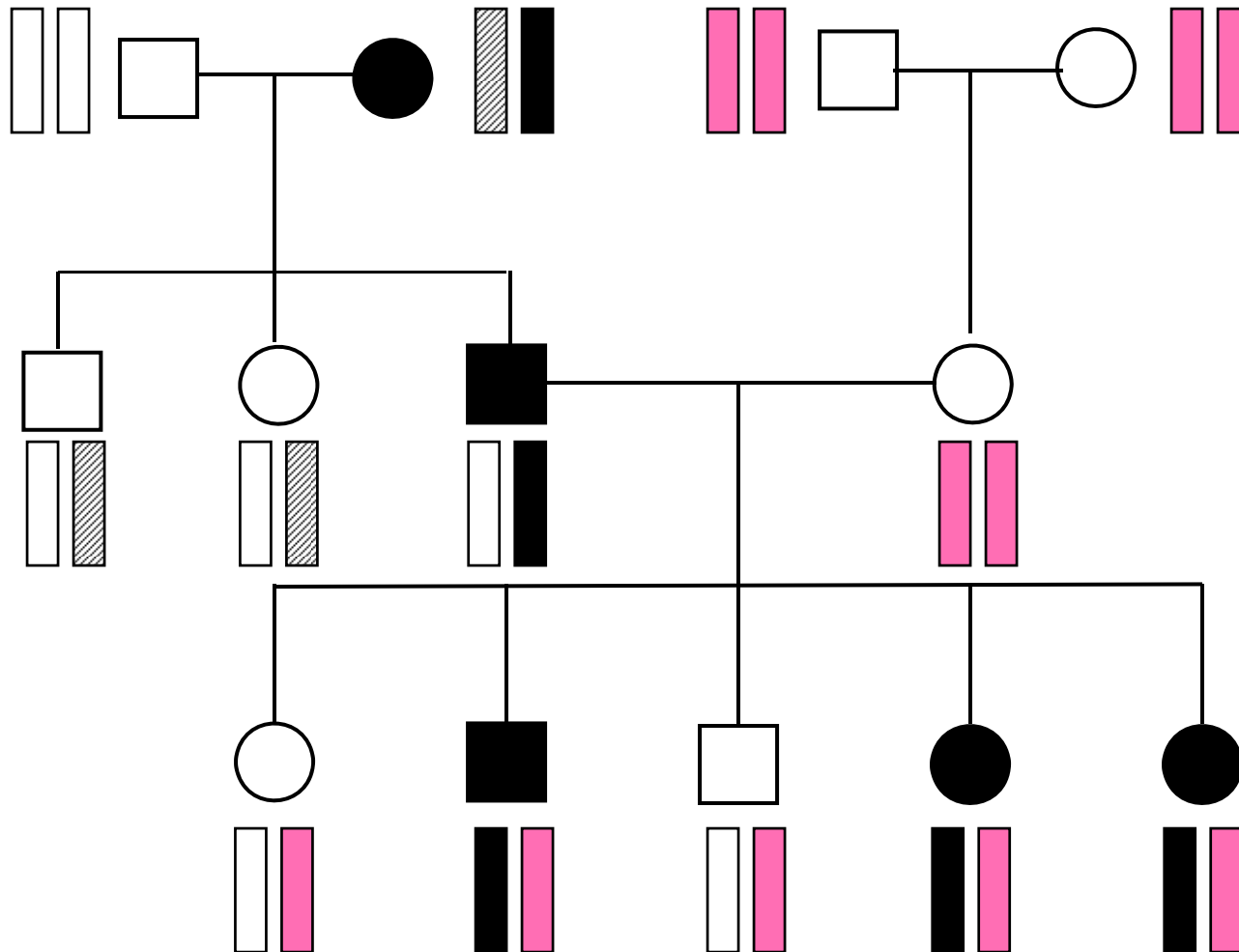
LINKAGE ANALYSIS: WHICH CHROMOSOMAL REGIONS SEGREGATE WITH DISEASE?



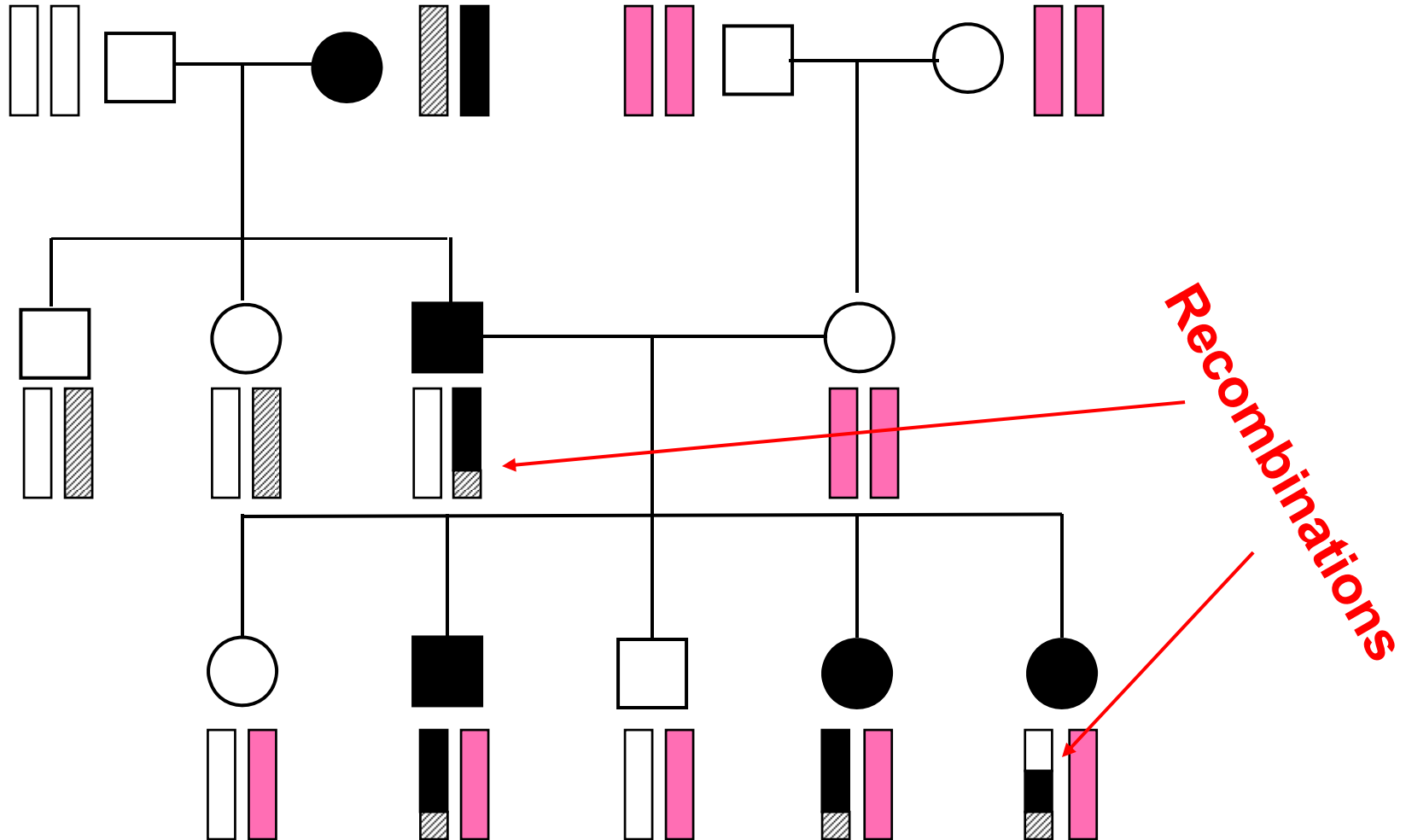
LINKAGE ANALYSIS: AN UNLINKED REGION



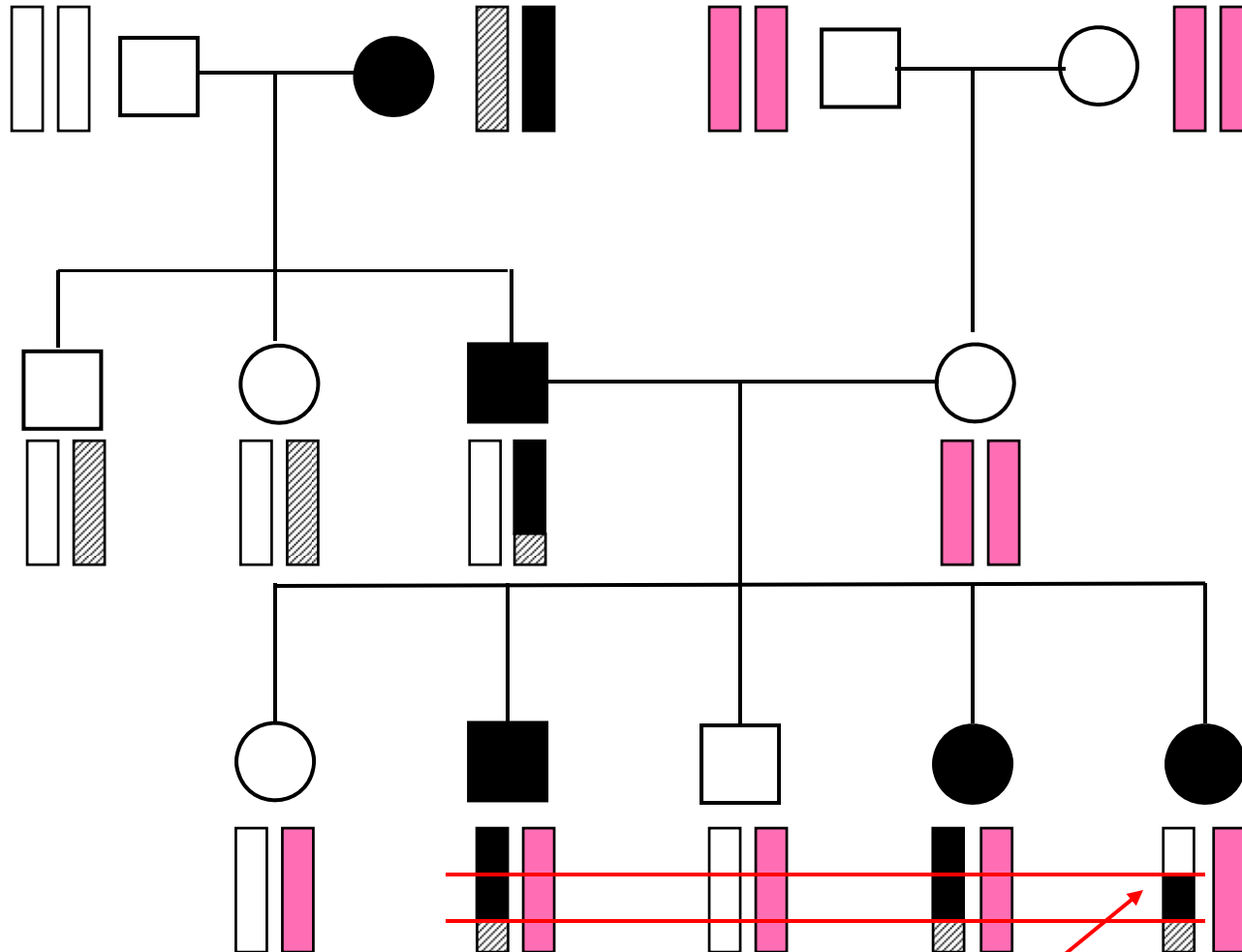
LINKAGE ANALYSIS: A LINKED REGION



RECOMBINATION EVENTS NARROW THE REGION OF LINKAGE



RECOMBINATION EVENTS NARROW THE REGION OF LINKAGE



Region of Linkage

RECOMBINATION IS TRACKED USING MICROSATELLITE MARKERS

- Variable number of short (1 - 4 bp) repeats
- Highly polymorphic
- Not associated with any known phenotypes
- Many thousands in human genome
- Three major centers perform high-throughput microsatellite genotyping
 - Center for Inherited Disease Research (CIDR), Baltimore MD
 - Mammalian Genotyping Service, Marshfield, WI
 - deCODE Genetics, Iceland

MICROSATELLITE MARKERS

One person's chromosomes:

tccagc**TGCT**TGCT**TGCT**TGCT**TGCT**TGCT**TGCT**TGCggctac

tccagc**TGCT**TGCT**TGCT**TGCT**TGCT**TGCT**TGCT**TGCT**TGCT**TGCT**TGCT**TGCggctac

Genotype: 6, 9

Another person's chromosomes:

tccagc**TGCT**TGCT**TGCT**TGCT**TGC**ggctac

tccagc**TGCT**TGCT**TGCT**TGCT**TGCT**TGCT**TGCT**TGCT**TGCT**TGCT**TGCT**TGCT**TGC**ggctac

Genotype: 4, 11

MICROSATELLITE MARKERS

- 400 - 1000 microsatellite markers adequate for genome-wide linkage study
- One marker at least every 10 cM (for 400 markers) or every 2.5 cM (for 1000 markers)
- 1 centiMorgan (cM) is the genetic distance over which there is a 1% probability of recombination in a single meiosis
- 1 cM \approx 1Mb
- Limiting factor in linkage analysis is usually the number of meioses, not the number of markers

LINKAGE TEST STATISTIC IS THE “LOD SCORE”

- Lod score weighs evidence for and against the presence of a disease-causing mutation in the region (i.e., evidence for and against linkage)
- Calculation is: $\log_{10} \left(\frac{\textit{Odds in favor of linkage}}{\textit{Odds against linkage}} \right)$
- Lod > 3 usually considered significant
- Lod < -2 usually taken as significant evidence against linkage to that region
- Lod = 0 when meioses are uninformative

LINKAGE METHODS

- “Parametric” linkage methods require specification of an inheritance model
 - Frequency of disease-causing mutation
 - Prevalence of the disease in the population
 - Penetrance (probability of disease) for carriers of zero, one or two copies of the mutation
- “Non-parametric” linkage methods do not require model specification
 - Parametric analysis using the correct inheritance model is more powerful than non-parametric analysis
 - Non-parametric analysis is usually more powerful than parametric analysis using an incorrect model

LINKAGE METHODS

- Linkage analysis works well when:
 - Disease is rare
 - Inheritance is clearly Mendelian
 - Mutations in one gene account for most of disease
 - Childhood onset of disease, straightforward diagnosis
- Linkage analysis has been largely unsuccessful for complex diseases such as schizophrenia, autism, major depression, bipolar disorder, autoimmune disorders, etc., in spite of clear evidence of high heritability.

OUTLINE

- BACKGROUND AND HISTORY
- LINKAGE ANALYSIS
- **ASSOCIATION ANALYSIS**
- TRANSMISSION DISTORTION TESTING

ASSOCIATION ANALYSIS

- Gene-mapping by association methods can be family-based or population-based
- Linkage methods are always family-based
- The association method covered in this lecture is population-based, as it is the most commonly used method and the calculations are intuitive.

POPULATION-BASED ASSOCIATION ANALYSIS

- **SAMPLE COMPOSITION:** Unrelated patients (cases) and unaffected individuals (controls)
- **GOAL:** Identify markers with alleles that are more common among cases than controls (or vice versa)
- **IDEA:** Susceptibility alleles may be in LD with the marker alleles that are over-represented in cases; protective alleles may be in LD with those that are over-represented in controls.

EXAMPLE: POPULATION-BASED ASSOCIATION ANALYSIS

- 100 cases and 100 controls were genotyped for a SNP in a candidate gene
- Minor allele frequencies were
 - 0.34 among cases
 - 0.25 among controls
- There appears to be a difference in allele frequency between cases and controls. Is this difference statistically significant?

EXAMPLE: POPULATION-BASED ASSOCIATION ANALYSIS

Note: chromosomes, not individuals, are tabulated in a test of allele frequencies

Observed Data:

	CASES	CONTROLS	
ALLELE 1	68	50	118
ALLELE 2	132	150	282
	200	200	400

EXAMPLE: POPULATION-BASED ASSOCIATION ANALYSIS

Observed Data:

	CASES	CONTROLS	
ALLELE 1	68	50	118
ALLELE 2	132	150	282
	200	200	400

Expected Data under H_0 (no association):

	CASES	CONTROLS	
ALLELE 1	59	59	118
ALLELE 2	141	141	282
	200	200	400

POPULATION-BASED ASSOCIATION USES CHI SQUARE TEST STATISTIC

$$T = \sum \frac{(\text{Observed} - \text{Expected})^2}{(\text{Expected})} \sim \chi_1^2$$

Under H_0 (no association), the expected value of T is 1

COMPUTING THE TEST STATISTIC

Observed:

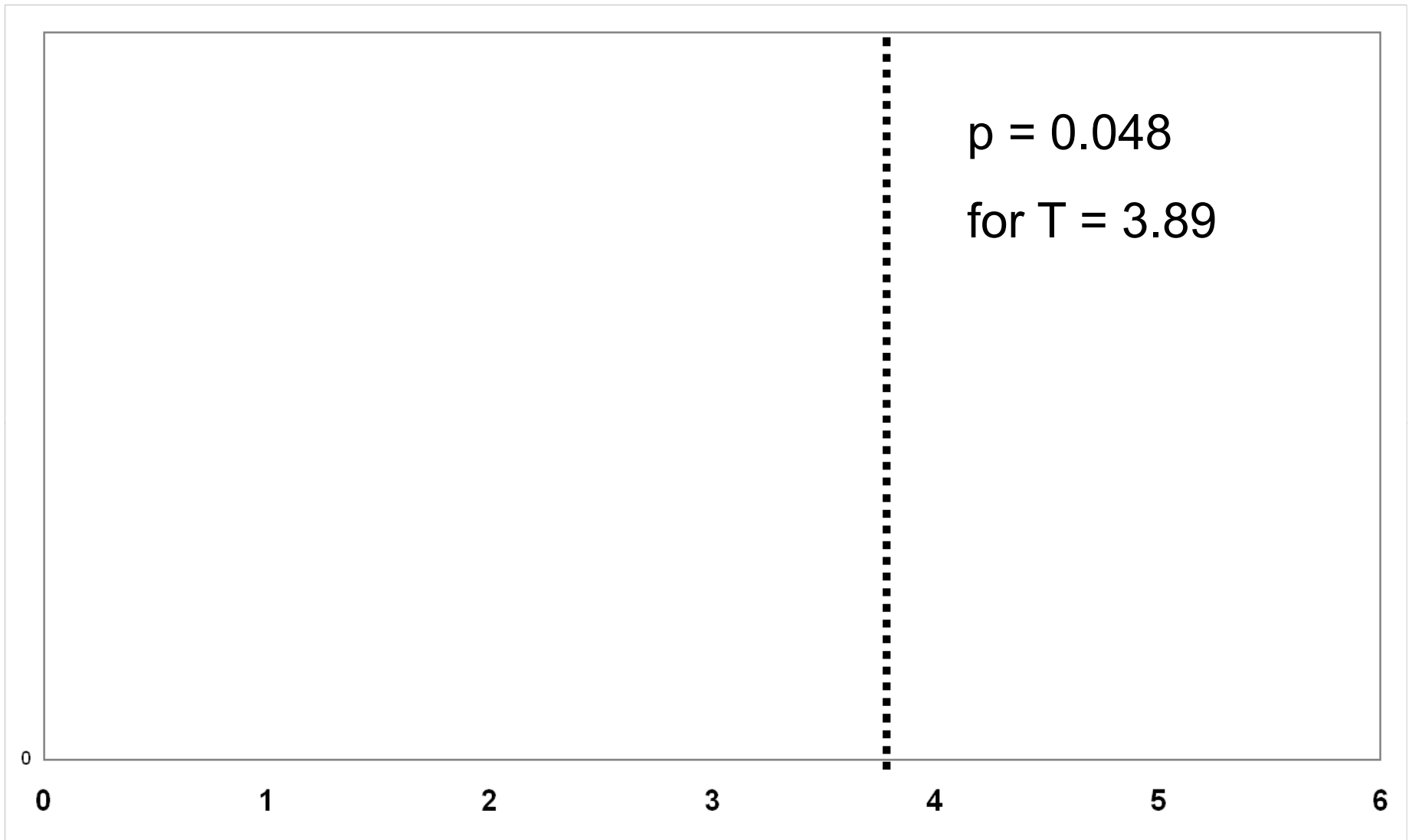
	CASES	CONTROLS	
ALLELE 1	68	50	118
ALLELE 2	132	150	282
	200	200	400

Expected:

	CASES	CONTROLS	
ALLELE 1	59	59	118
ALLELE 2	141	141	282
	200	200	400

$$T = \sum \frac{(\text{Observed} - \text{Expected})^2}{(\text{Expected})}$$
$$= \frac{(68 - 59)^2}{59} + \frac{(50 - 59)^2}{59} + \frac{(132 - 141)^2}{141} + \frac{(150 - 141)^2}{141} = 3.89$$

χ^2 DISTRIBUTION



ASSOCIATION ANALYSIS

- Population-based case-control analysis has long been employed in candidate gene studies
 - Identify genes with biologically plausible role in disease pathology
 - Genotype SNPs in these genes and test for association with disease status
- Genome-wide association studies (GWAS) were not possible until very recently
 - Genome coverage for association analysis requires 300,000+ markers
 - Genome coverage for linkage analysis requires 400 markers

GENOME-WIDE ASSOCIATION ANALYSIS

- Association methods known to be more powerful than linkage methods for common diseases with low-penetrance risk alleles
- Technological hurdle:
 - Whole genome linkage requires only 400 markers
 - Whole genome association requires 300,000+ markers
- Rapid development in SNP genotyping technology
 - In 2002, 10k SNPs genotyping chip introduced
 - In 2008, 1M SNP chips available

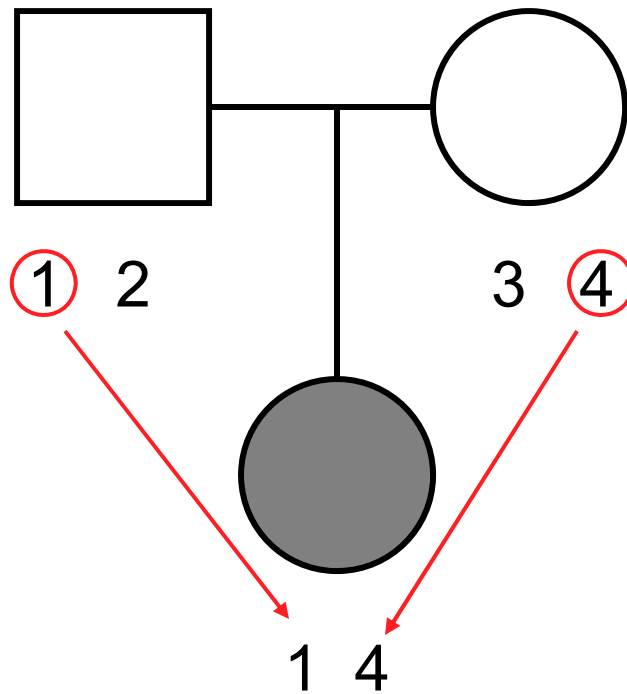
OUTLINE

- BACKGROUND AND HISTORY
- LINKAGE ANALYSIS
- ASSOCIATION ANALYSIS
- **TRANSMISSION DISTORTION TESTING**

TRANSMISSION DISEQUILIBRIUM TEST (TDT)

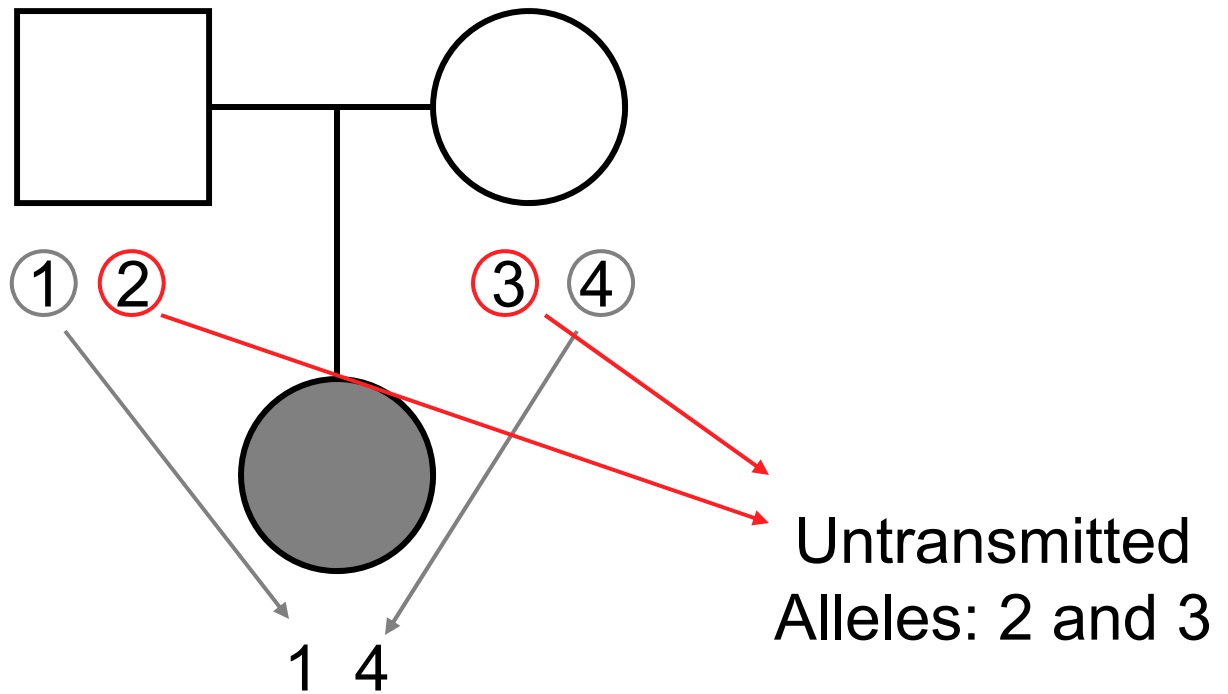
- SAMPLE COMPOSITION: Case-parent trios
- The test uses only heterozygous parents
- For each allele at a marker locus, count number of times:
 - the allele was transmitted (T) from heterozygous parents to affected children
 - the allele was not transmitted (U) from heterozygous parents to affected children
- Under H_0 , T and U have equal probability

TRANSMISSION DISEQUILIBRIUM TEST (TDT)



Transmitted Alleles: 1 and 4

TRANSMISSION DISEQUILIBRIUM TEST (TDT)



Transmitted Alleles: 1 and 4

TDT TEST STATISTIC: McNEMAR CHI SQUARE

$$\frac{(T - U)^2}{(T + U)} \sim \chi_1^2$$

Under H_0 , the expected value of the test statistic is 1

TDT IS A TEST OF LINKAGE DISEQUILIBRIUM

- Tracking meioses within families (linkage)
- Looking for transmission of the same allele to unrelated patients (association)
- H_0 : No linkage disequilibrium
- H_A : Linkage and association between marker and disease loci

SUMMARY

- Non-random segregation of genes during meiosis is called linkage
- Linkage analysis tracks recombinations in family data and is effective for rare Mendelian diseases
- Genome-wide linkage analysis requires 400 - 1000 markers
- Correlation between alleles at distinct loci is called association
- Association can be due to many phenomena including LD between marker allele and mutation or population structure