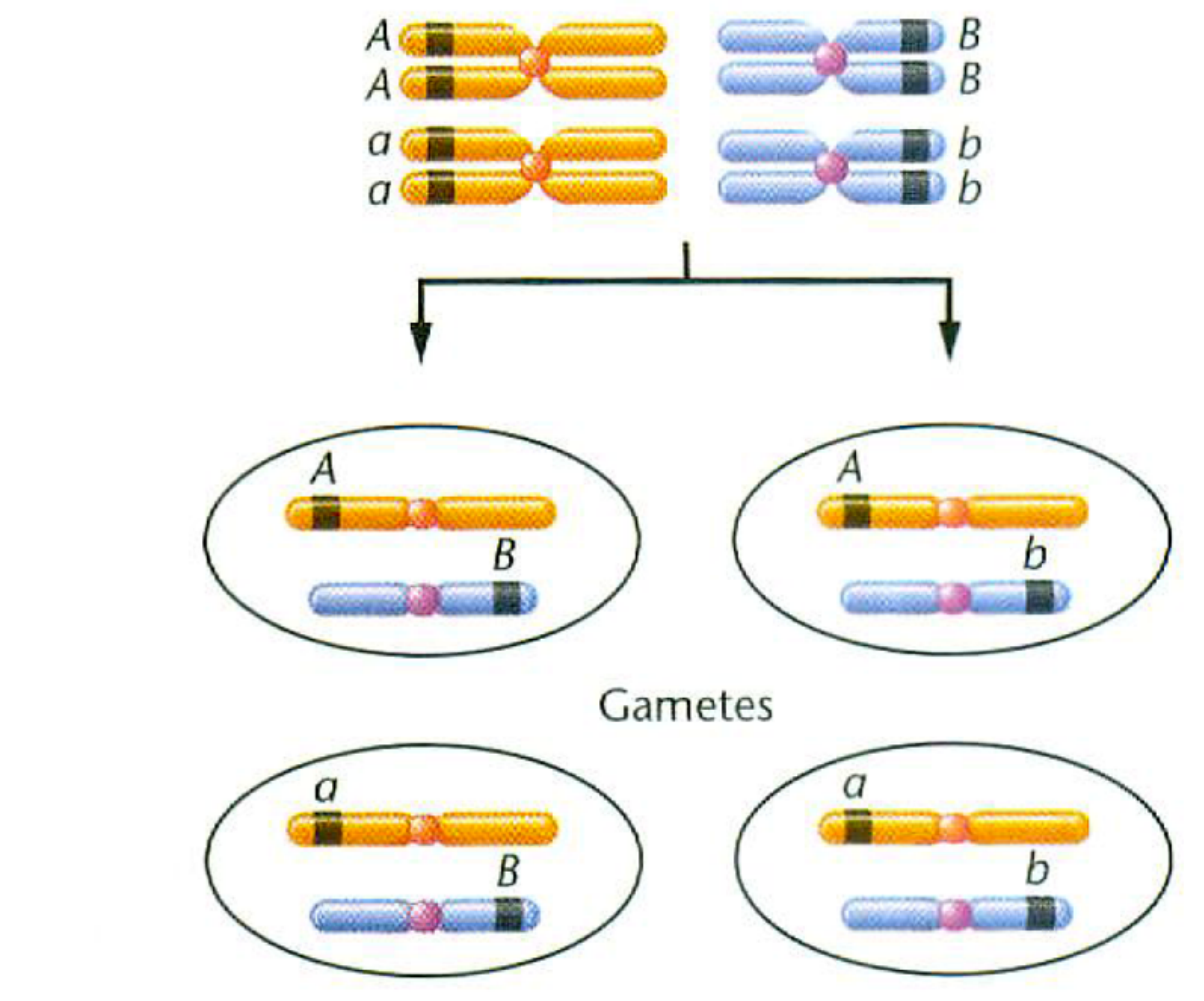
# Human linkage analysis

fundamental concepts

# Genes and chromosomes

Alelles of genes located on different chromosomes show independent assortment (Mendel's 2nd law)

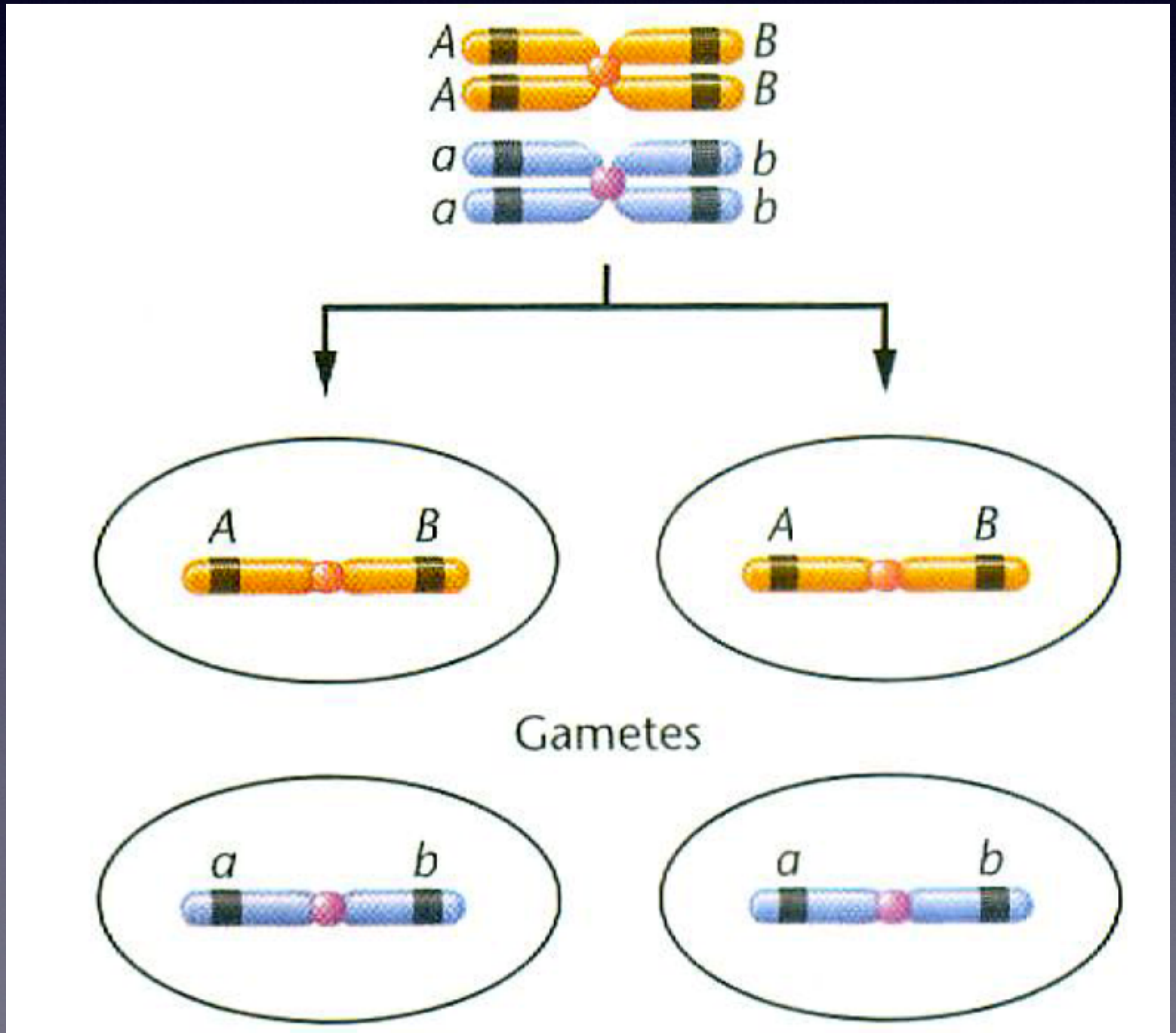


For 2 genes:
4 gamete classes with equal number

W. S Klug, M.R Cummings "Concepts of Genetics" 8th edition, Prentice Hall, 2005

# Linkage

Alleles of genes located on the same chromosome tend to segregate together - linkage

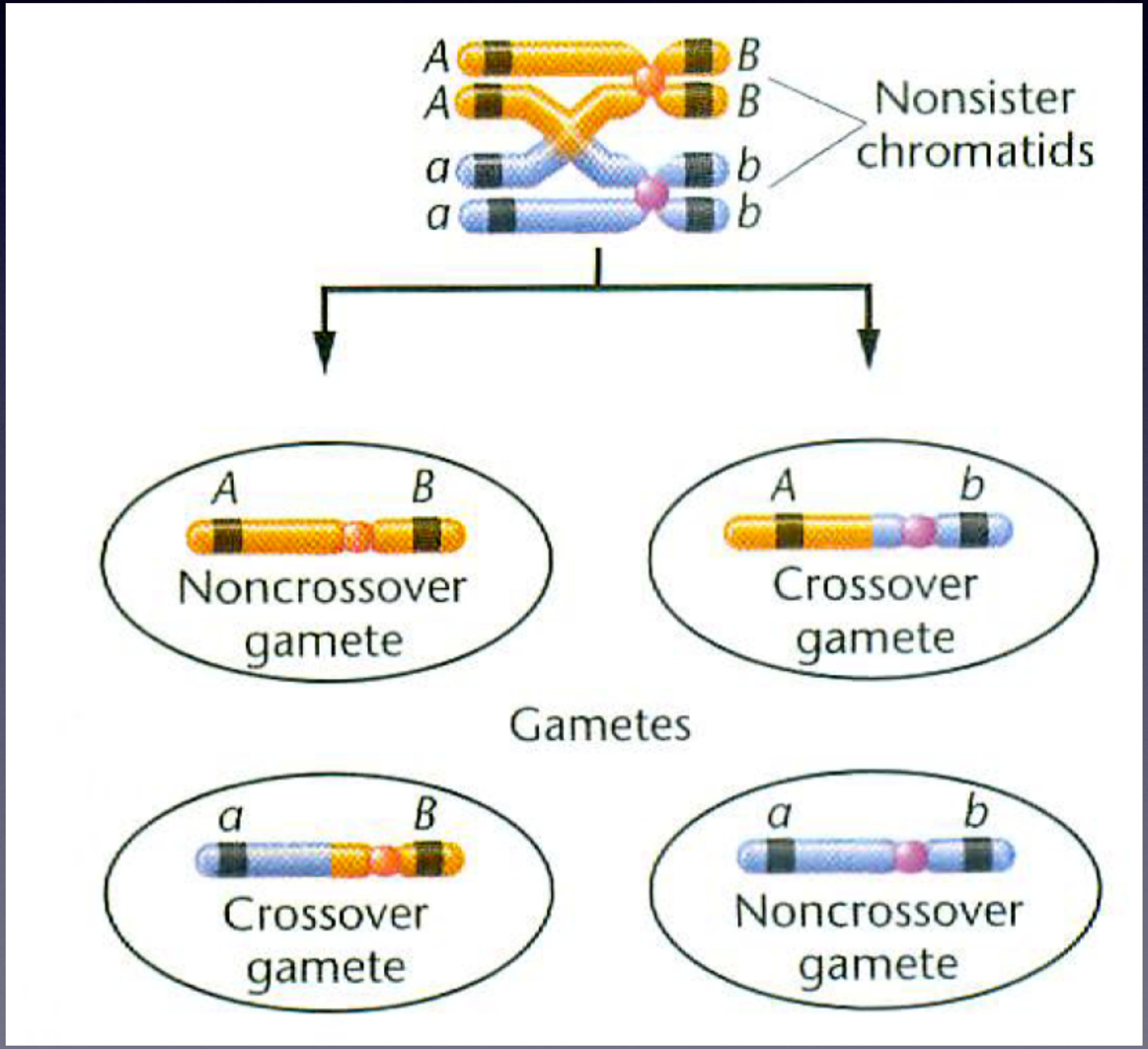


For 2 genes and complete linkage:
2 parental genotype gamete classes

W. S Klug, M.R Cummings “Concepts of Genetics” 8th edition, Prentice Hall, 2005

# Linkage

Crossing-over (non-sister chromatid exchange by meiotic recombination)

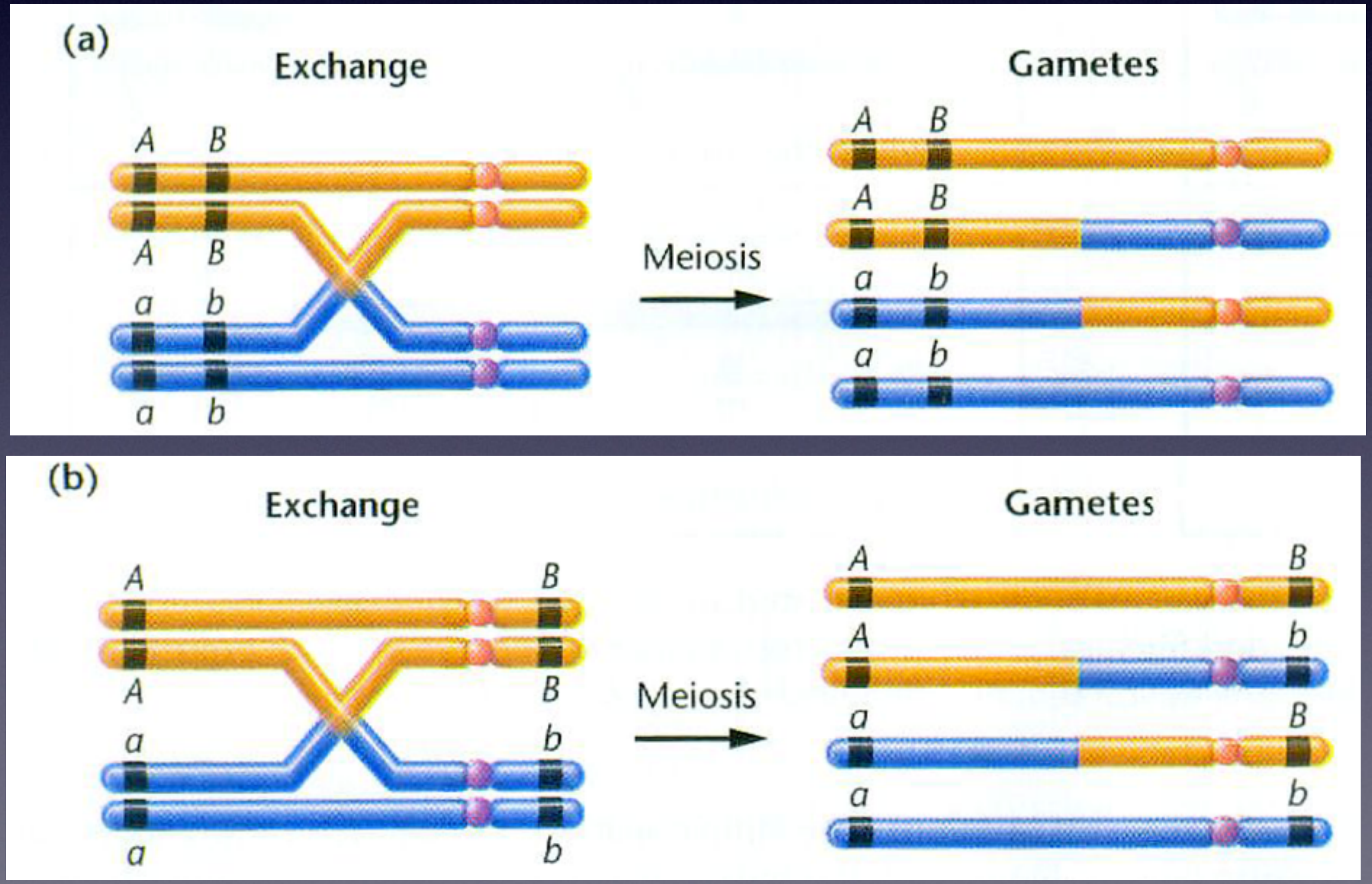


For 2 genes:
2 parental (noncrossover) classes
2 recombinant (crossover) classes
Fewer recombinant than parental gametes

W. S Klug, M.R Cummings “Concepts of Genetics” 8th edition, Prentice Hall, 2005

# Linkage mapping

To form recombinant gametes, a crossover has to occur **between** the gene loci



non-recombinant gametes form

recombinant gametes form

W. S Klug, M.R Cummings “Concepts of Genetics” 8th edition, Prentice Hall, 2005

# Principles of linkage mapping

- The crossing-over probability between gene loci is proportional to the distance separating them on the chromosome
- The number of recombinant genotypes in the offspring measures the genetic distance
- In *Drosophila* the easiest way is to cross a double heterozygous female with a double recessive male
- How about human?

# Association *vs.* linkage

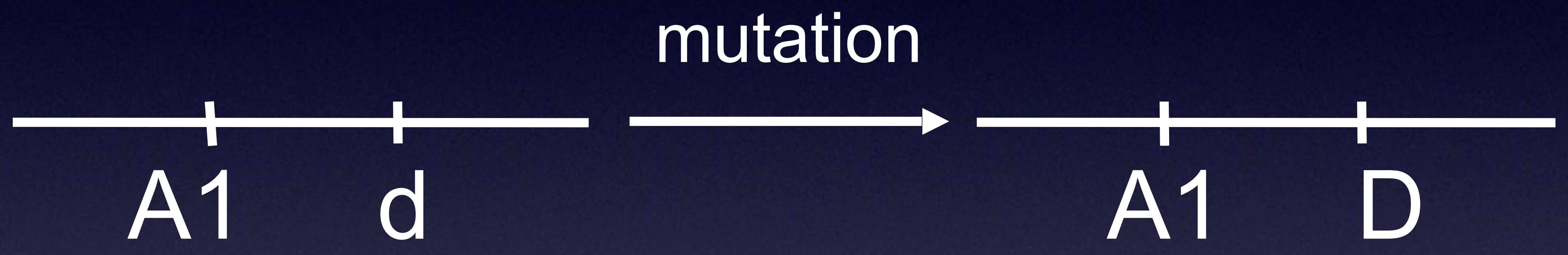
- Linkage - co-segregation of alleles of genes located on the same chromosome
    - involves gene loci, regardless of the allele
    - a simple biological mechanism (chromosomes, recombination)
    - studied in pedigrees or pairs of related individuals
    - used to study Mendelian traits - high heritability, alleles of single (or few) genes cause the phenotype

# Association *vs.* linkage

- Association - a correlation between gene alleles and traits in a population
  - always involves particular alleles
  - biological mechanism often complex or unknown - a statistical phenomenon, can be indirect
  - studied in a population of individuals, not from the same family
  - used to study multifactorial inheritance
  - can be related to linkage in a special case (linkage disequilibrium)

# LInkage disequlilbrium

Allele of the gene *d* linked with the marker locus *A* mutated to the disease allele *D* - founder event



If the A to d distance is small, then most chromosomes that carry D also carry  A1

Not vice versa (most chromosomes with A1 need not carry D)!

Linkage disequilibrium) – nonrandom association of alleles in linked loci – founder effect. Decreases over time.

# Methods

- Linkage analysis - genetic mapping
    - parametric methods
    - nonparametric methods
- Association - correlation studies (statistical)

# Linkage in the human genome

- Human genes are usually located far from each other, with large intergenic regions
- Linkage between two genes with observable phenotypes is extremely rare
- Molecular markers (RFLP, VNTR, etc.) are used
  - human genome linkage maps, e.g. CEPH
  - finding a marker linked to a disease locus

# Linkage between a marker locus and a disease gene

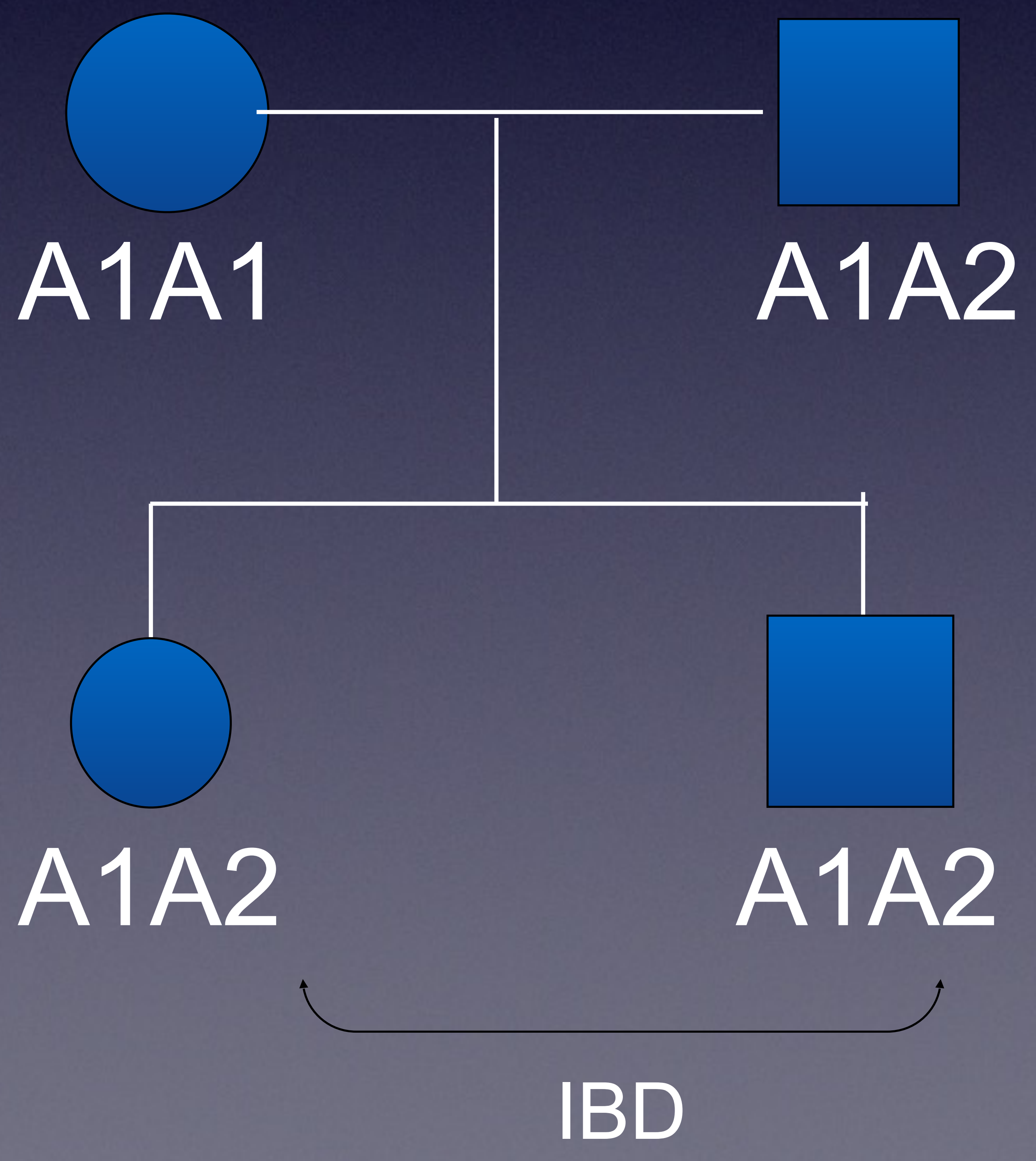
- Association in a family (among related individuals)
- Usually no population-level association
- Independent of the population structure
- Linkage disequilibrium on the population level for very rare alleles

# Mapping methods

- Parametric  (based on a model of inhertitance): lod-score analysis
  - two-point
  - multipoint
- Nonparametric linkage analysis
  - correlation between alleles in related individuals
  - IBD (identity by descent) vs. IBS (identity by state)

# Nonparametric analysis

Two alleles are identical by descent (IBD) if they are copies of the same ancestral allele



# Nonparametric methods

- Correlation of the phenotype and the coincidence of a particular marker allele
    - Twin studies
    - Affected siblings method
    - Family studies (2-3 generations)
    - Affected siblings method: in pairs of affected siblings are the marker alleles (any) identical more often, then in the control population?

# Parametric methods

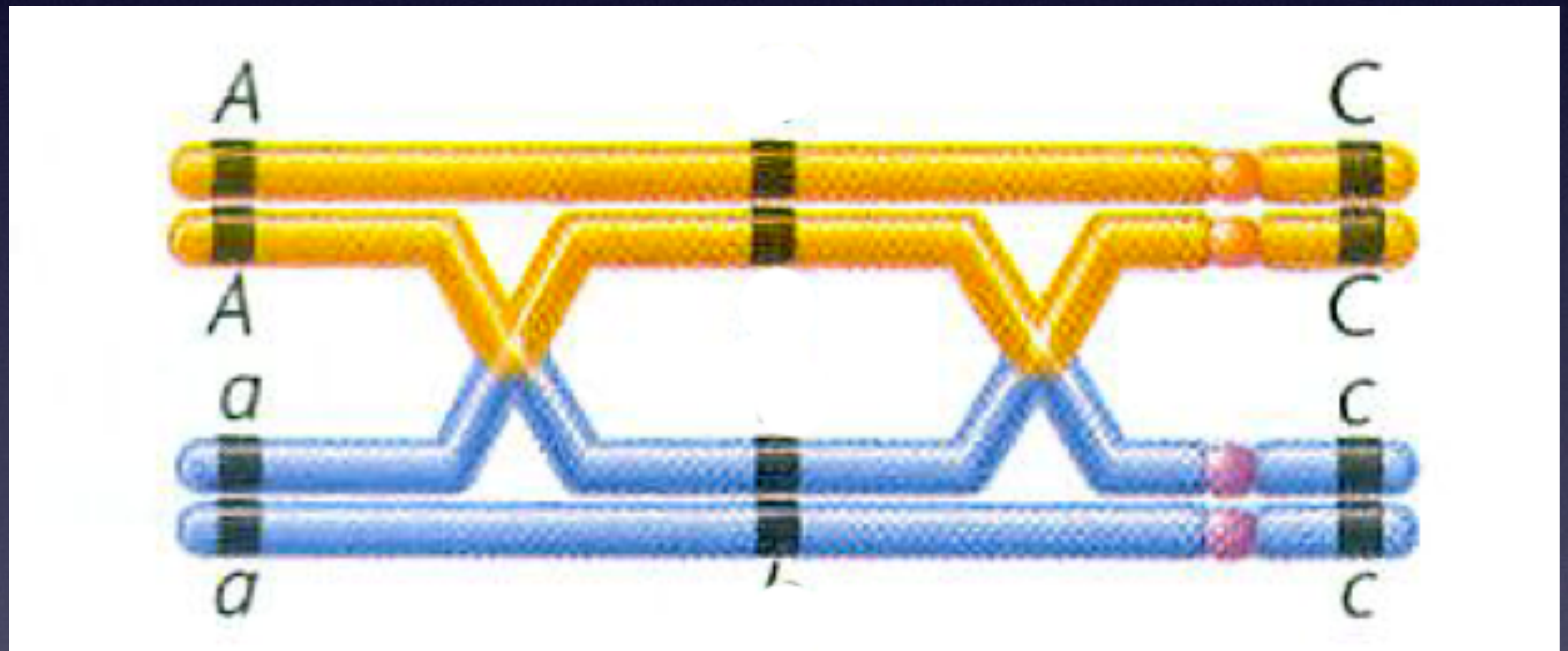
- In *Drosophila* the easiest way is to cross a double heterozygous female with a double recessive male
- How about human?

# Recombination frequency is a measure of genetic distance

- Recombination frequency $\theta$= probability of transmission of a recombinant gamete
- Loci on separate chromosomes segregate independently
  => $\theta = 0.5$
- Tightly linked loci segregate together
  => $\theta = 0$
- Therefore
  - $\theta<0.5$ linkage
  - $\theta=0.5$ no linkage

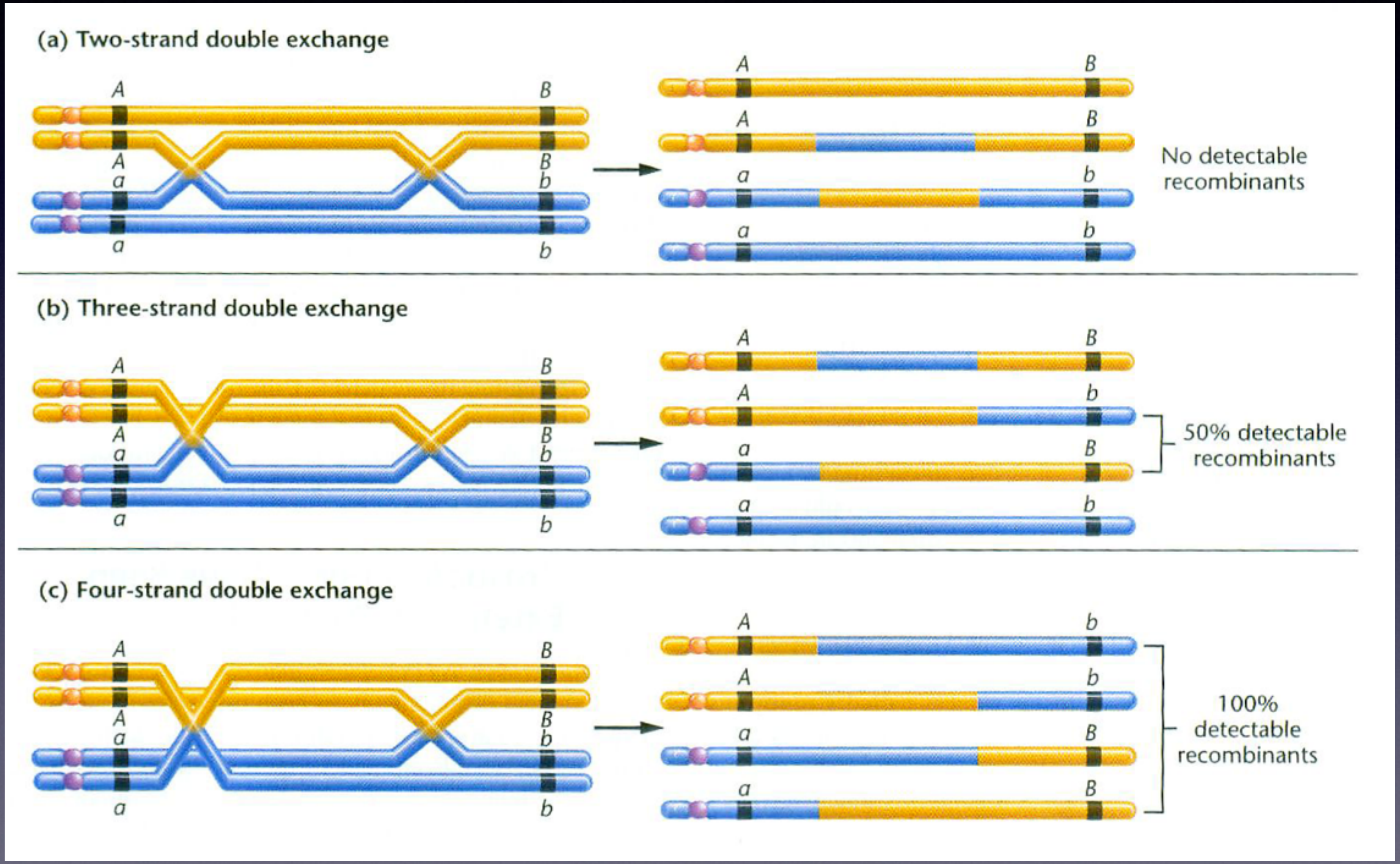
# Linkage mapping

- Unit: cM (centimorgan) = 1% recombination frequency
- The correlation is not linear



- Double crossing-over – parental type gametes
- Interference – crossing-over in one region influences the probability of c-o in nearby regions

# Double c-o – a complex picture


(a) Two-strand double exchange
A
B
A
B
a
b
a
b
A
B
A
B
a
b
a
b
No detectable recombinants
(b) Three-strand double exchange
A
B
A
B
a
b
a
b
A
B
A
b
a
B
a
b
50% detectable recombinants
(c) Four-strand double exchange
A
B
A
B
a
b
a
b
A
b
A
b
a
B
a
B
100% detectable recombinants

On average 50% recombinants. Similarly for triple, etc.

W. S Klug, M.R Cummings “Concepts of Genetics” 8th edition, Prentice Hall, 2005

# Mapping function

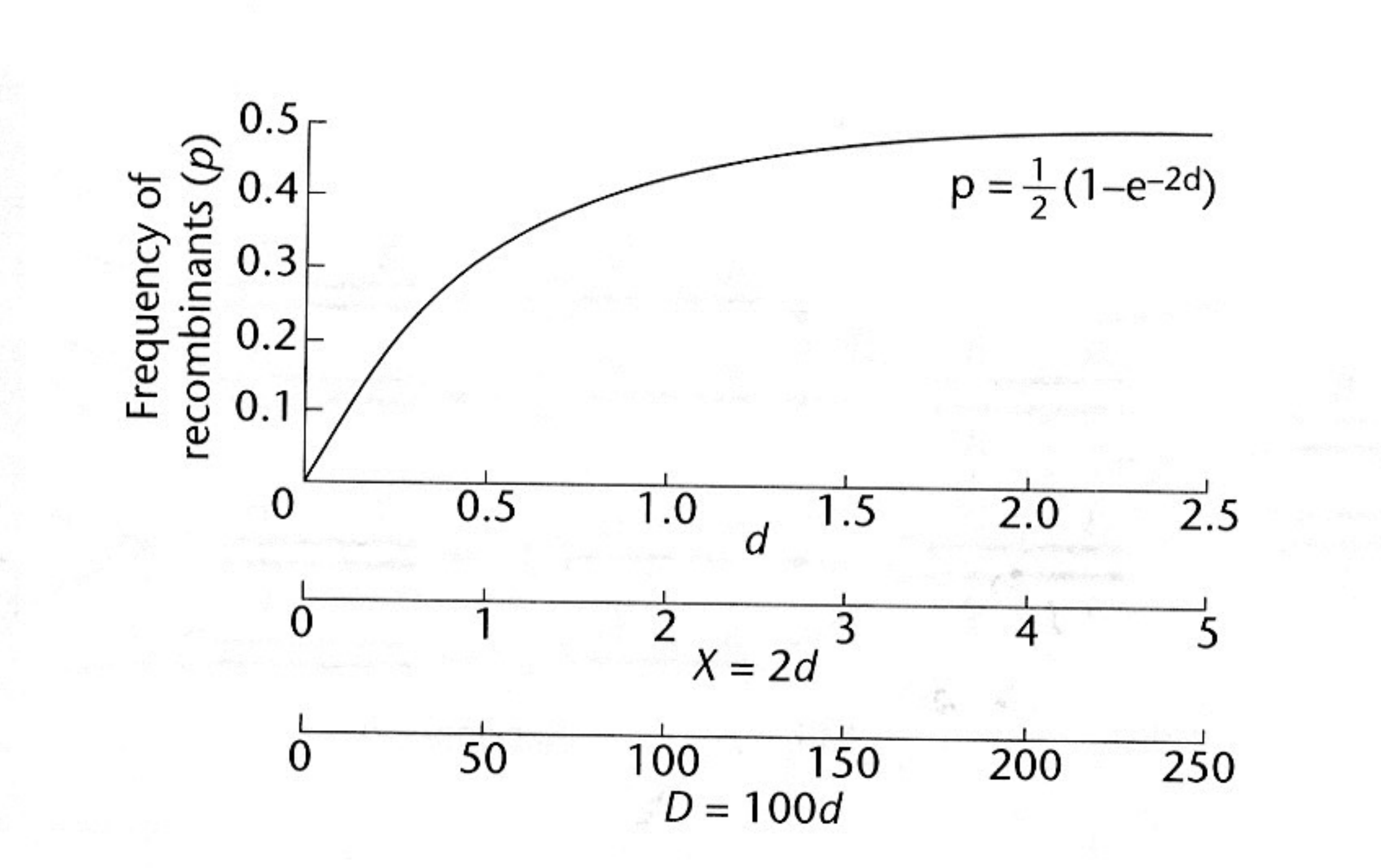
- Genetic distance as a function of observed recombinant frequency
- Haldane's function
  - multiple c-o, no interference
- Kosambi's function
  - multiple c-o, interference, commonly used
- For small θ: d≈θ

$$d = \frac{\ln(1-2\theta)}{2}$$

$$d = \frac{\ln(\frac{1+2\theta}{1-2\theta})}{4}$$

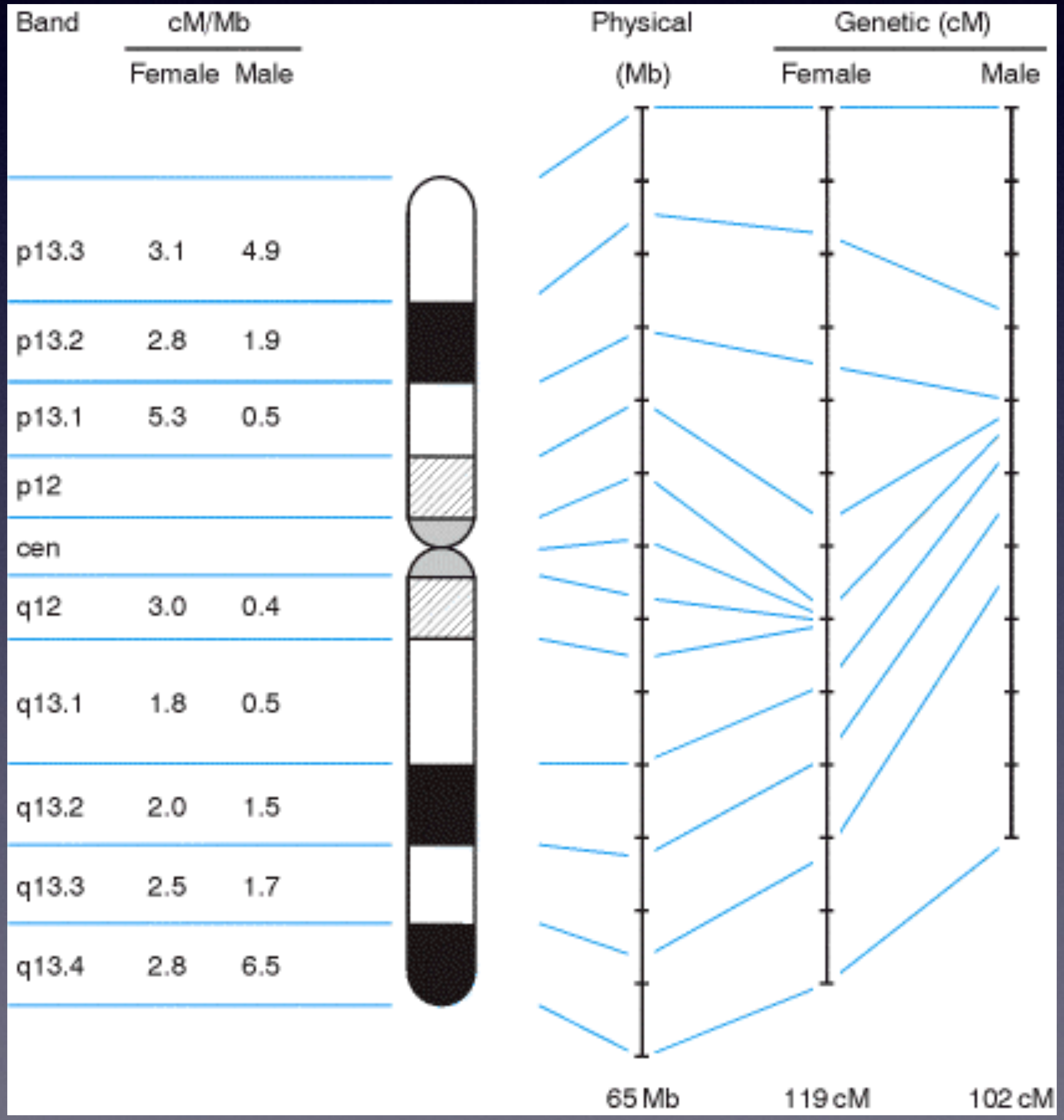
# Mapping function

- Observed frequency of recombinants approaches 0.5 with increasing distance
- For unlinked genes 50% “recombinants”, like for genes far apart on the chromosome



# Sex and recombination frequency

- Total male genetic map = 2851cM (autosomal)
- Total female genetic map = 4296 cM (autosomal)
- For ~3000Mb of autosomal genome
  - 1 cM in males ≈ 1.05 Mb
  - 1 cM in females ≈ 0.7 Mb
  - average 1 cM ≈ 0.88 Mb
- the male/female ratio varies across genome



| Band | cM/Mb | |
|---|---|---|
| | Female | Male |
| p13.3 | 3.1 | 4.9 |
| p13.2 | 2.8 | 1.9 |
| p13.1 | 5.3 | 0.5 |
| p12 | | |
| cen | | |
| q12 | 3.0 | 0.4 |
| q13.1 | 1.8 | 0.5 |
| q13.2 | 2.0 | 1.5 |
| q13.3 | 2.5 | 1.7 |
| q13.4 | 2.8 | 6.5 |

# Likelihood

- Likelihood: the probability of obtaining the observed data under assumptions of a tested model

# Likelihood in pedigree analysis

- In a fully informative pedigree
  - data: R=number of recombinants; NR=number of parental genotypes
  - the parameter: recombination frequency (probability) θ

- Null hypothesis – no linkage (θ=0.5)
- Likelihood ratio L(θ)/L(θ=0.5)
- lod score (Z) = logarithm of odds – decimal logarithm of the likelihood ratio

# Simple lod score calculations

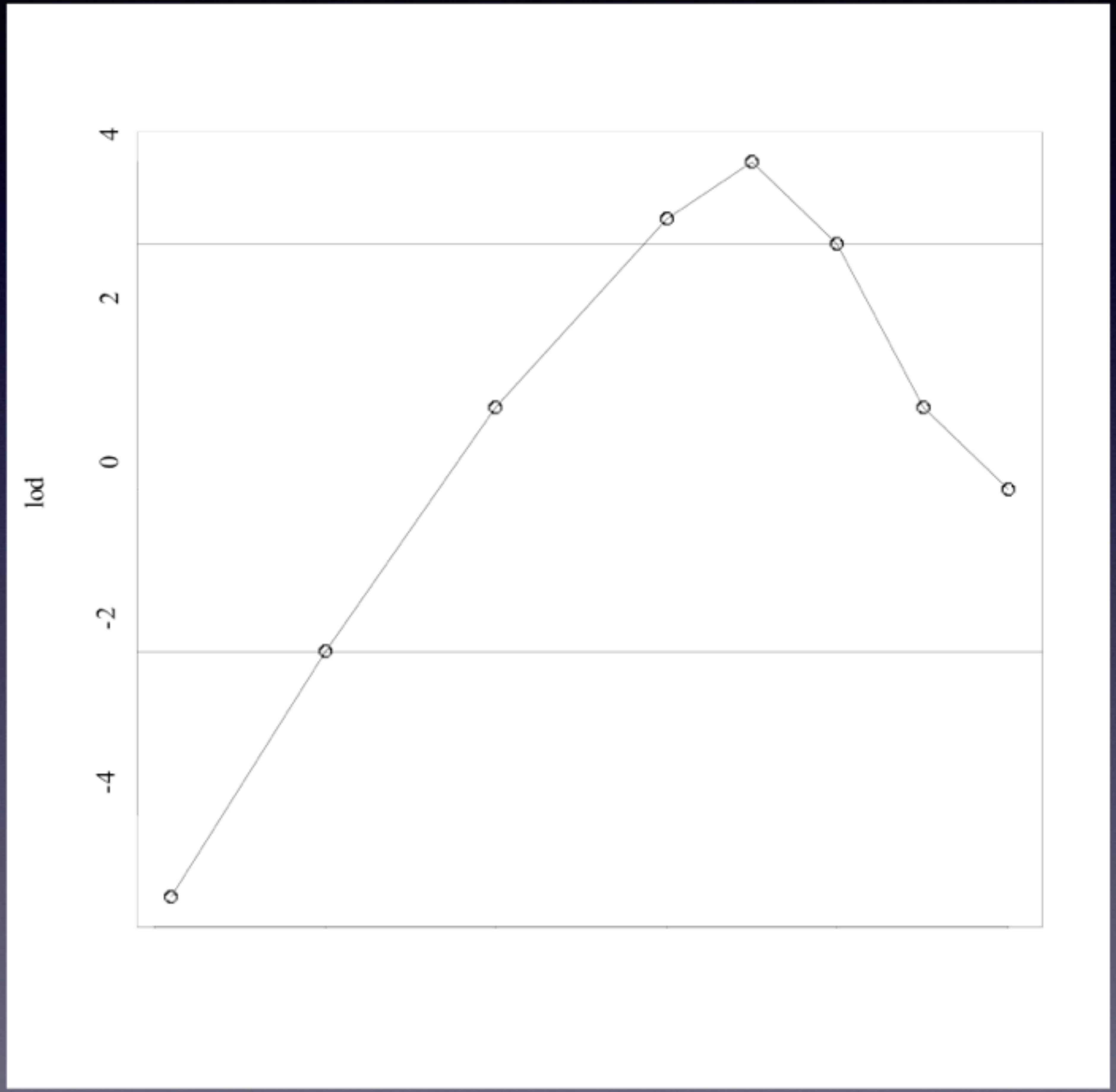
For each pedigree (i), the lod score is:

$$Z_i(\theta) = \log_{10} \frac{L(pedigree / \theta)}{L(pedigree / \theta = 0.5)}$$

For each θ, lod-score is summed across pedigrees (F):

$$Z(\theta) = \sum_{i=1}^{F} Z_i(\theta)$$

# Two-point linkage analysis



significance
(Z>3, Z>2 for X-linked)

excluded

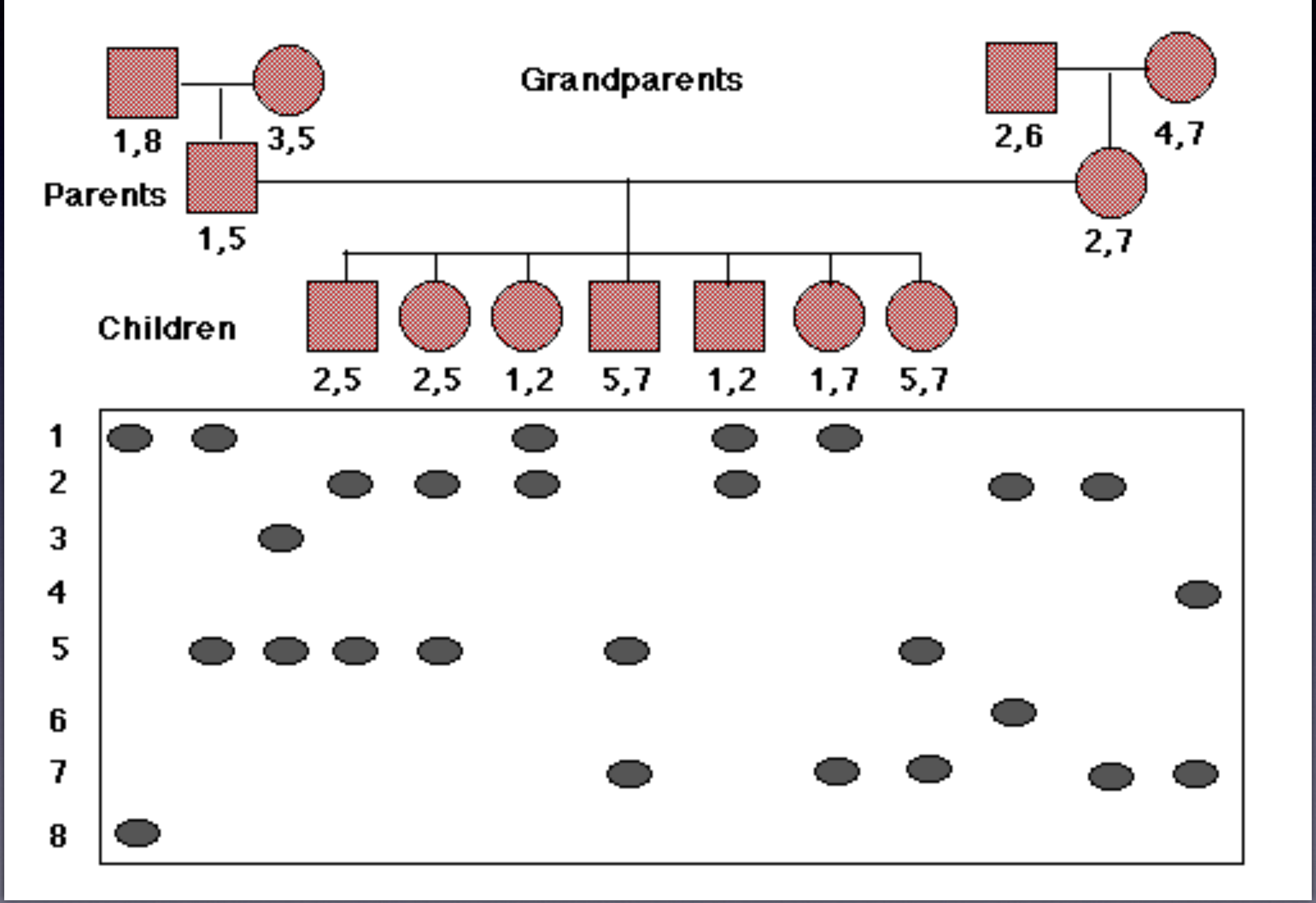
**Table**

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| θ = | 0.01, | 0.10, | 0.20, | 0.30, | 0.35, | 0.40, | 0.45, | 0.50 |
| lod= | -5.0, | -2.0, | 1.0, | 3.3, | 4.0, | 3.0, | 1.0, | 0.0 |

# Markers in human linkage analysis

- Linkage of two genes with an observable phenotype - extremely rare
    - exception – NPS – Nail Patella Syndrome and AB0 blood groups
    - MHC loci
- Molecular markers
    - PCR, RFLP

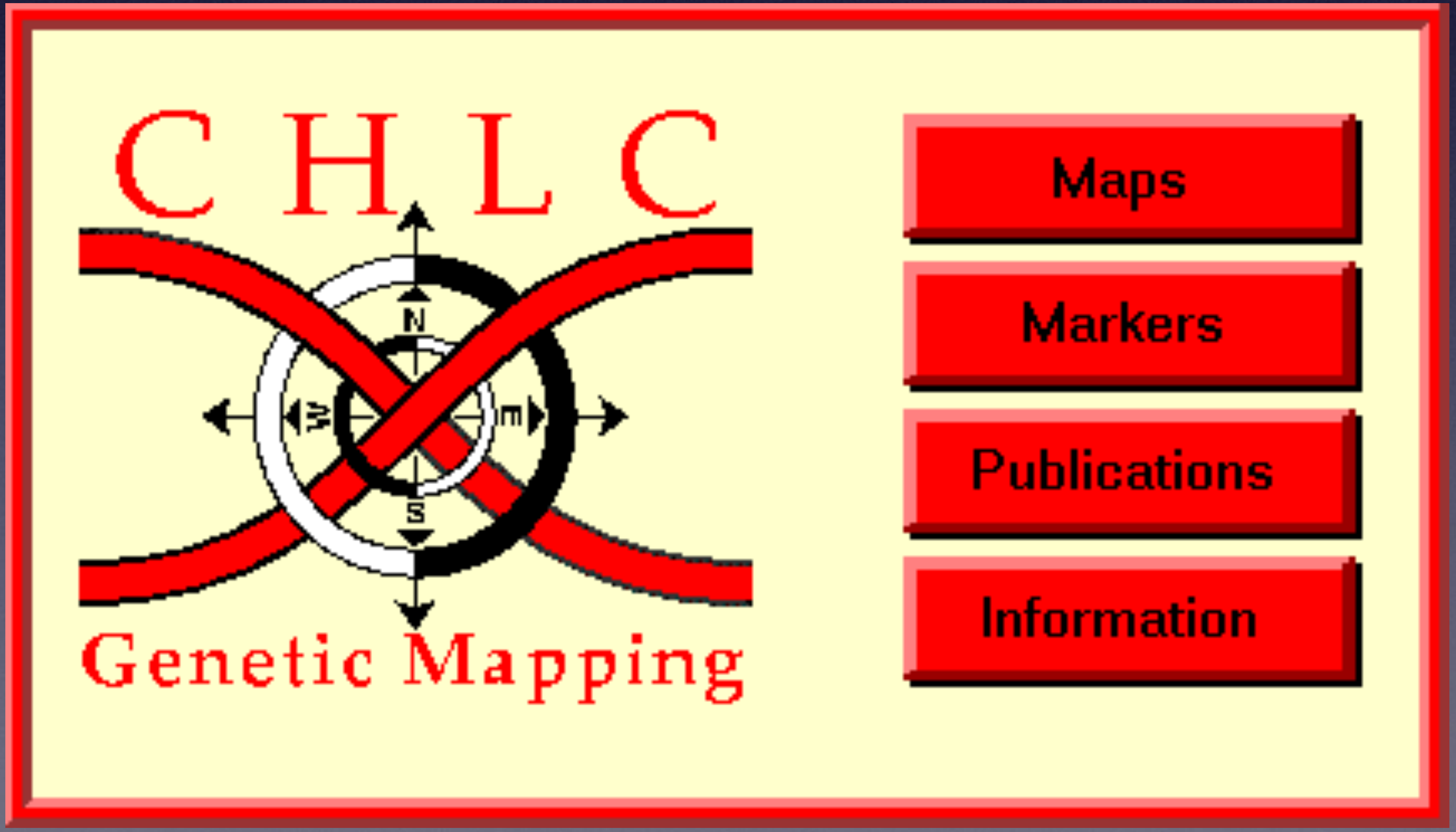
# Markers



Grandparents
1,8
3,5
2,6
4,7
Parents
1,5
2,7
Children
2,5
2,5
1,2
5,7
1,2
1,7
5,7
1
2
3
4
5
6
7
8

# Finding a gene

- I stage – general (markers spaced 8-20 cM) – define the chromosome, is it a single locus, etc.
- II stage – fine-mapping (markers spaced 1-4 cM)



The Cooperative Human Linkage Center, www.chlc.org

# Linkage in the age of genomics

- Whole genome sequencing is becoming more and more powerful and available
- Is linkage analysis still necessary?

**Table 1 | Summary of 1000 Genomes Project phase I data**

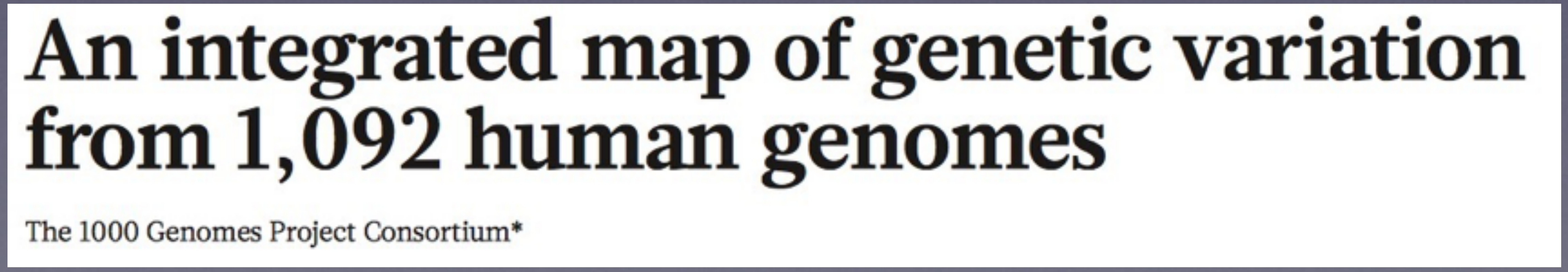
| | Autosomes | Chromosome X | GENCODE regions* |
|---|---|---|---|
| Samples | 1,092 | 1,092 | 1,092 |
| Total raw bases (Gb) | 19,049 | 804 | 327 |
| Mean mapped depth (×) | 5.1 | 3.9 | 80.3 |
| SNPs | | | |
| No. sites overall | 36.7 M | 1.3 M | 498 K |
| Novelty rate† | 58% | 77% | 50% |
| No. synonymous/non-synonymous/nonsense | NA | 4.7/6.5/0.097 K | 199/293/6.3 K |
| Average no. SNPs per sample | 3.60 M | 105 K | 24.0 K |
| Indels | | | |
| No. sites overall | 1.38 M | 59 K | 1,867 |
| Novelty rate† | 62% | 73% | 54% |
| No. inframe/frameshift | NA | 19/14 | 719/1,066 |
| Average no. indels per sample | 344 K | 13 K | 440 |
| Genotyped large deletions | | | |
| No. sites overall | 13.8 K | 432 | 847 |
| Novelty rate† | 54% | 54% | 50% |
| Average no. variants per sample | 717 | 26 | 39 |

NA, not applicable.

* Autosomal genes only.

† Compared with dbSNP release 135 (Oct 2011), excluding contribution from phase I 1000 Genomes Project (or equivalent data for large deletions).

*Lists of participants and their affiliations appear at the end of the paper.

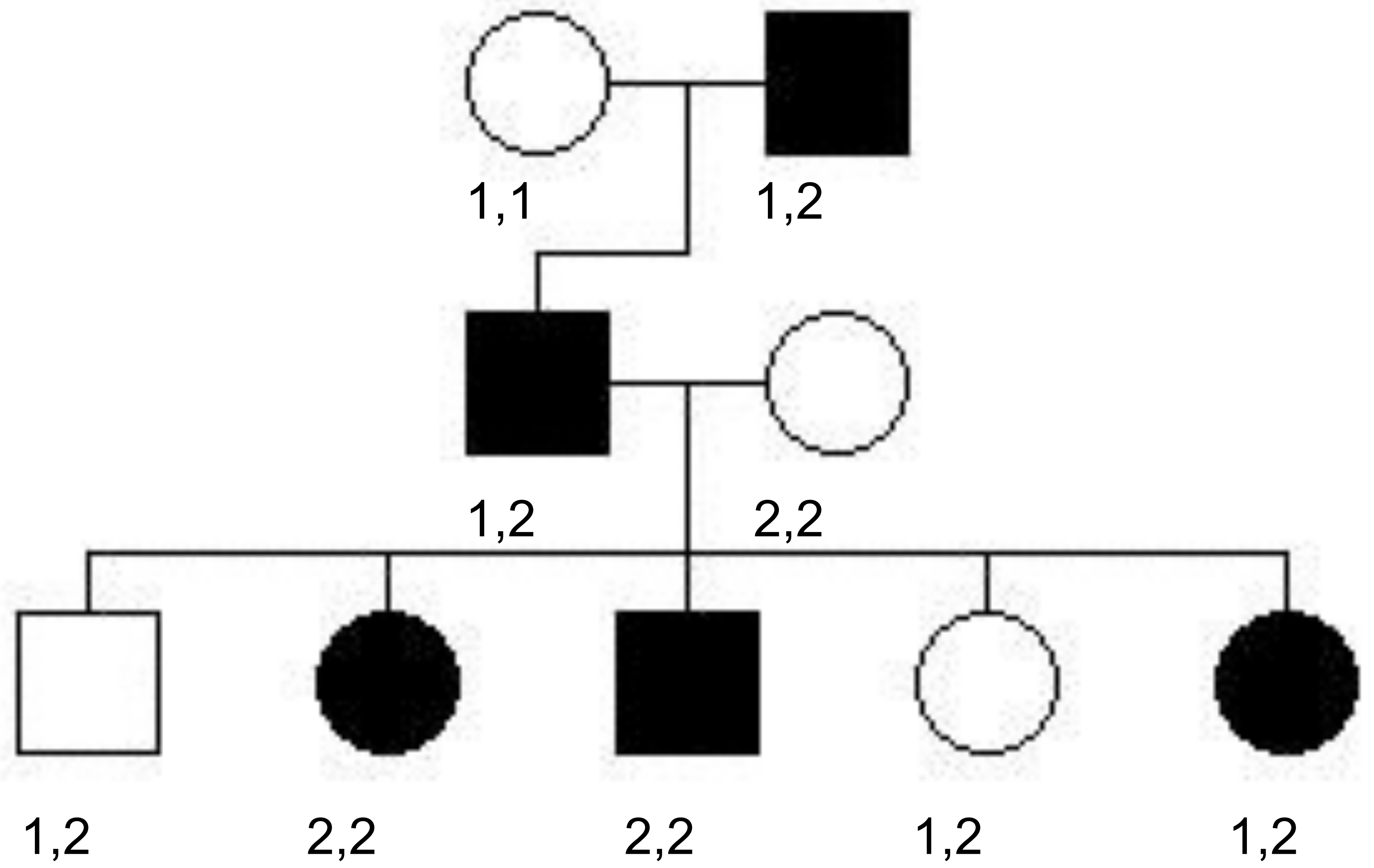


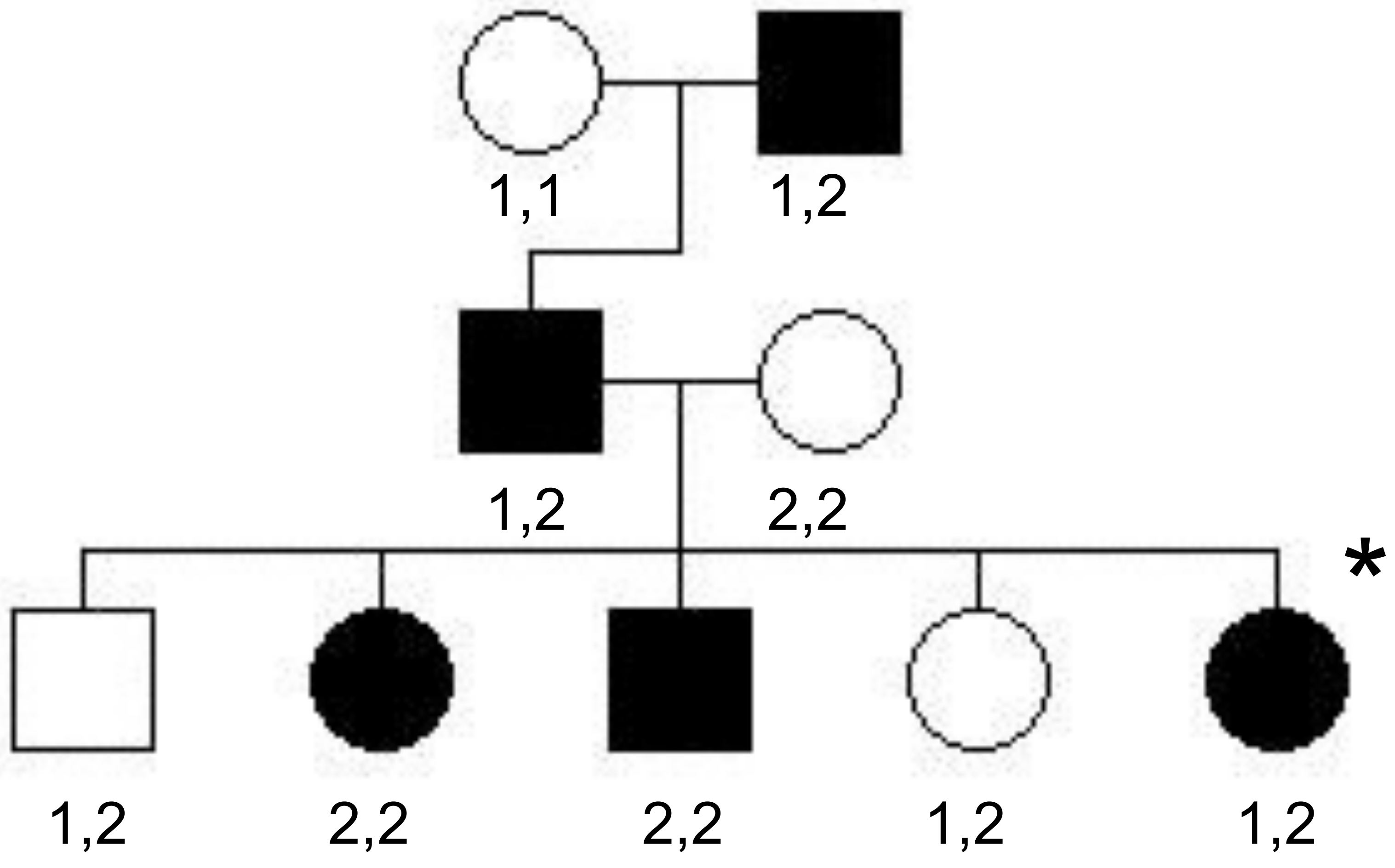
# An integrated map of genetic variation from 1,092 human genomes

The 1000 Genomes Project Consortium*

# Linkage in the age of genomics

- We can expect millions of sequence differences between two individuals
  - Less in close relatives, but still a lot
- Which of these differences is responsible for a phenotype is not evident
  - Easier in coding regions
- Whole genome (or exome) sequencing is used for very rare disorders (not enough cases for linkage)

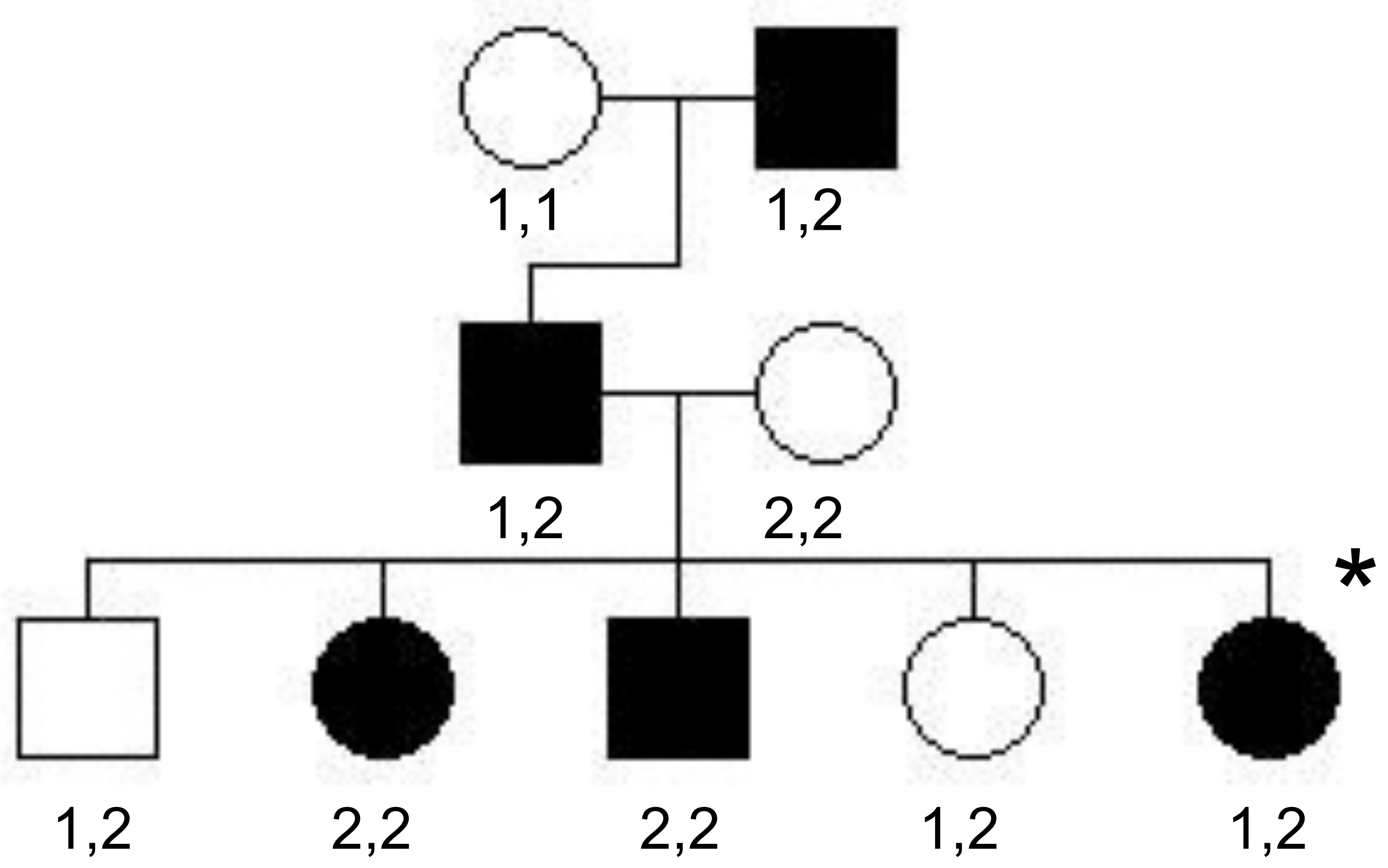

1,1
1,2
1,2
2,2
1,2
2,2
2,2
1,2
1,2



1 recombinant (R); 4 non-recombinant (NR)

Assuming no linkage (θ=0.5) probability of getting either R i NR is the same and equals ½

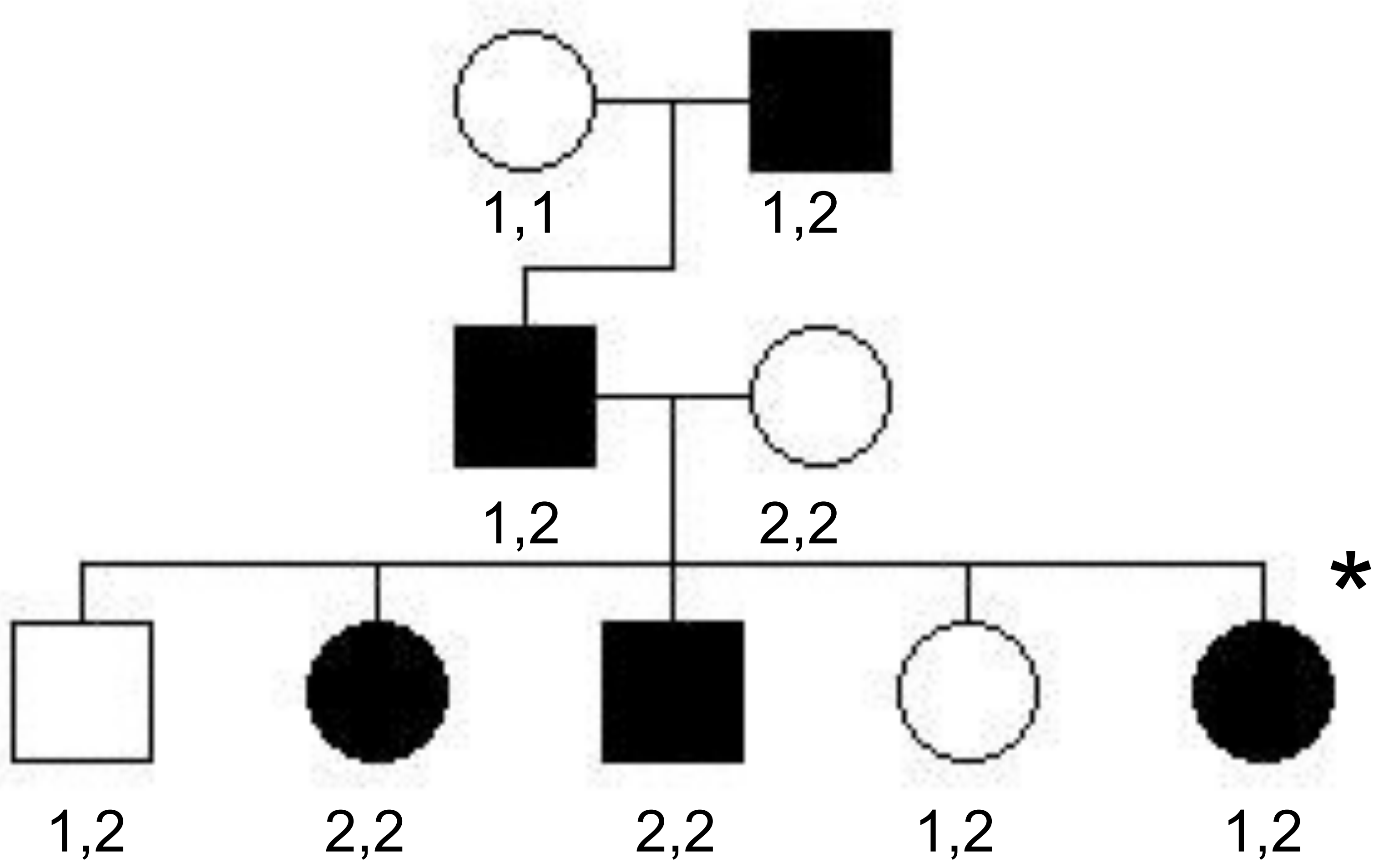
$L(\theta=0,5)= (½)^5$



1 recombinant (R); 4 non-recombinant (NR)

For a given θ the probability of obtaining R is θ (by definition), therefore the probability of obtaining NR is 1- θ

$$L(\theta)= \underbrace{\theta}_{1R}\cdot\underbrace{(1- \theta)^4}_{4NR}$$



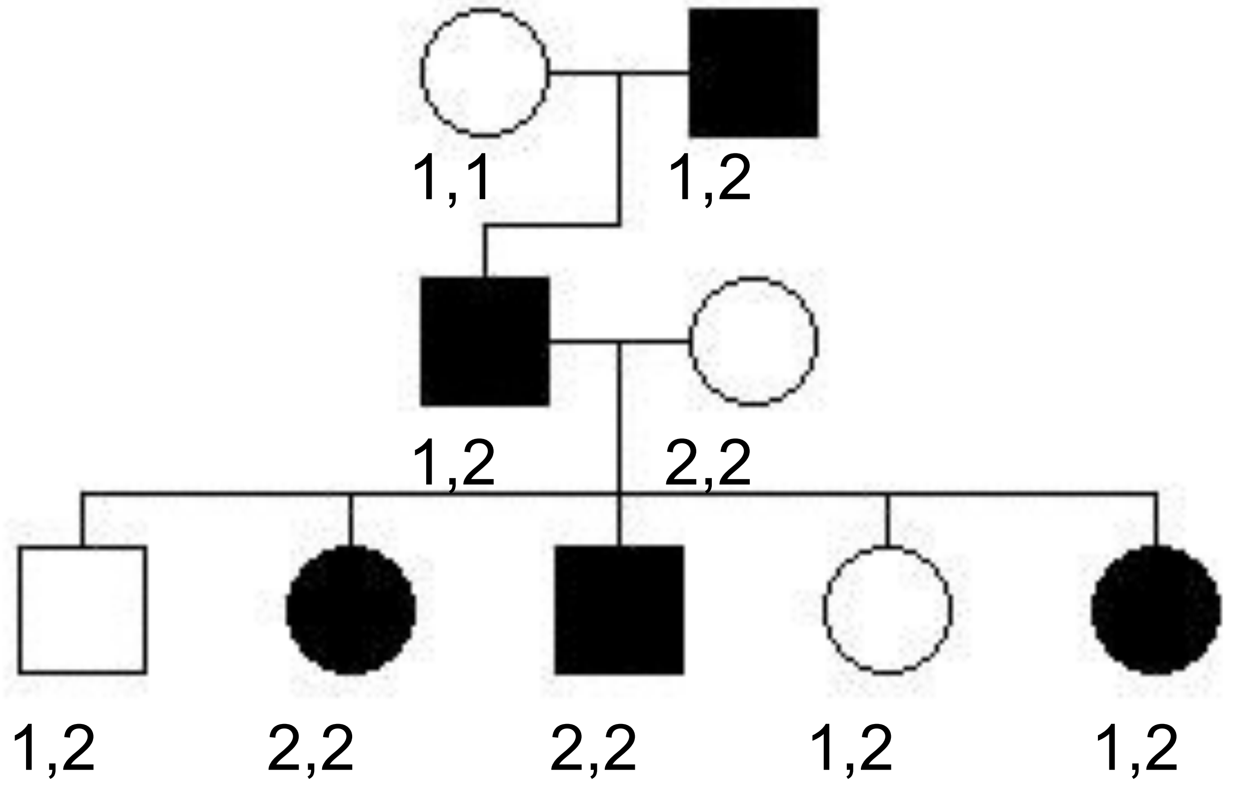
1 recombinant (R); 4 non-recombinant (NR)

$L(\theta=0.5) = (½)^5$

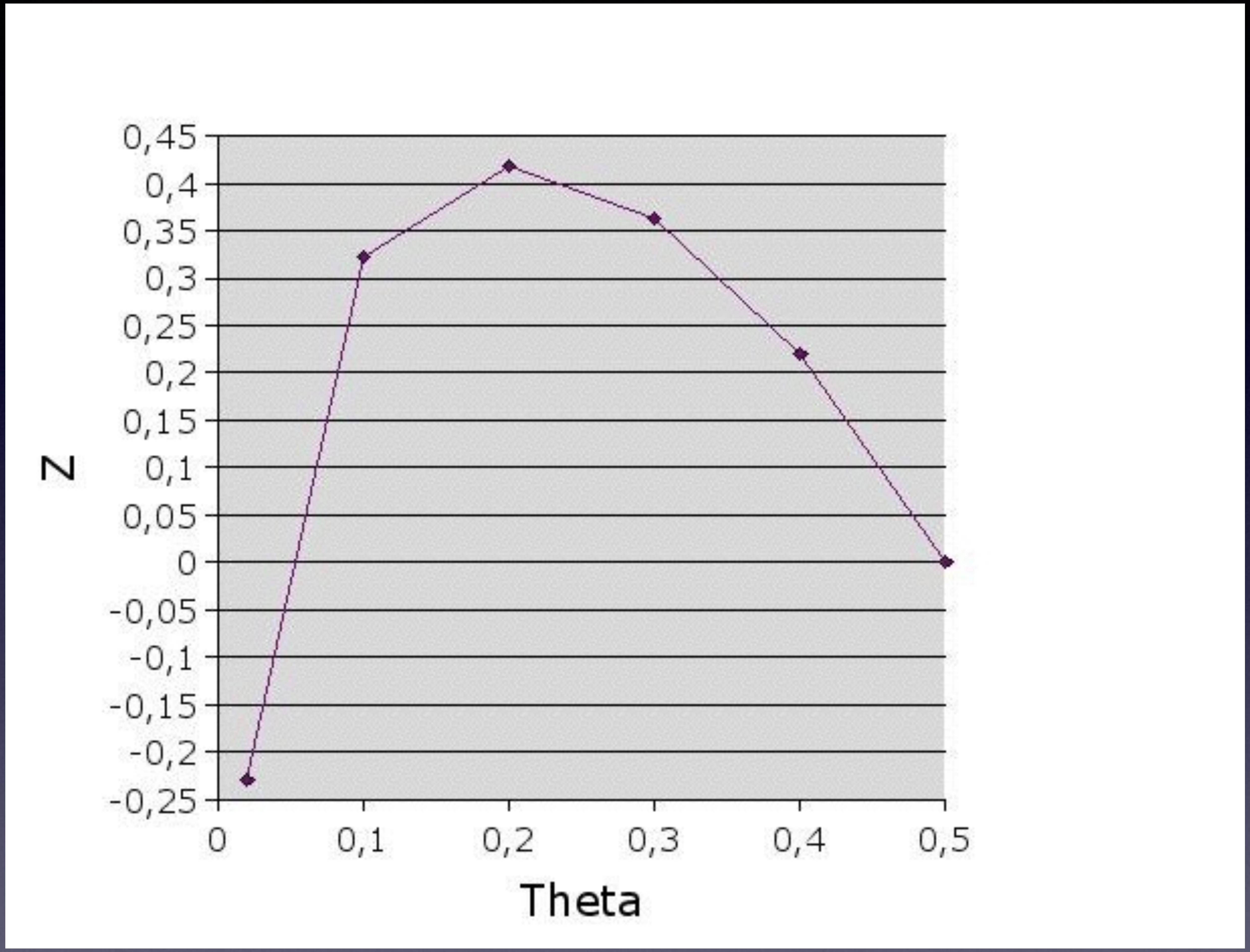
$L(\theta) = \theta \cdot (1 - \theta)^4$

For $\theta=0.1$ $L(\theta=0.1) = 0.1 \cdot (0.9)^4$

$$Z(\theta = 0{,}1) = \log_{10}\left(\frac{0{,}1 \cdot 0{,}9^4}{0{,}5^5}\right) \approx 0{,}32$$

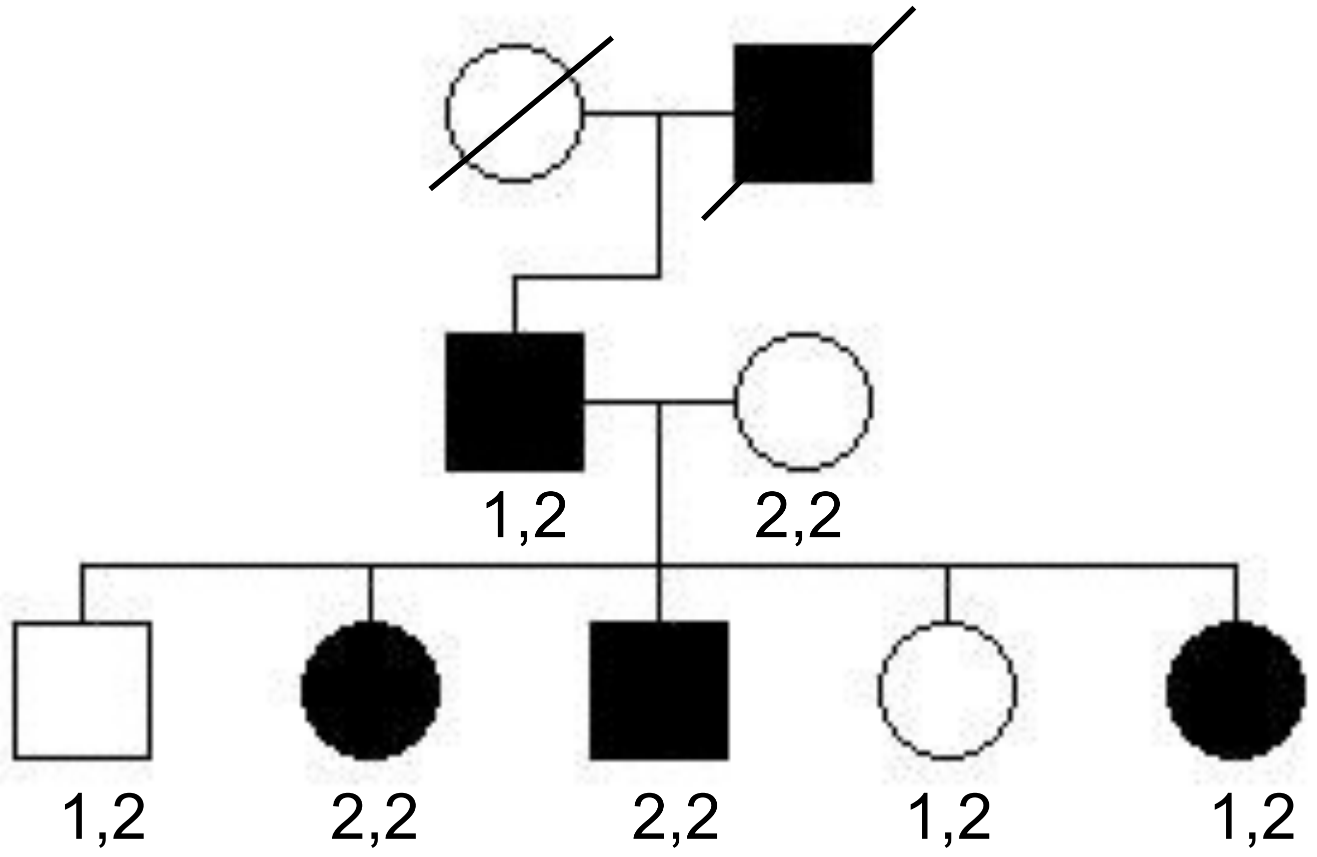


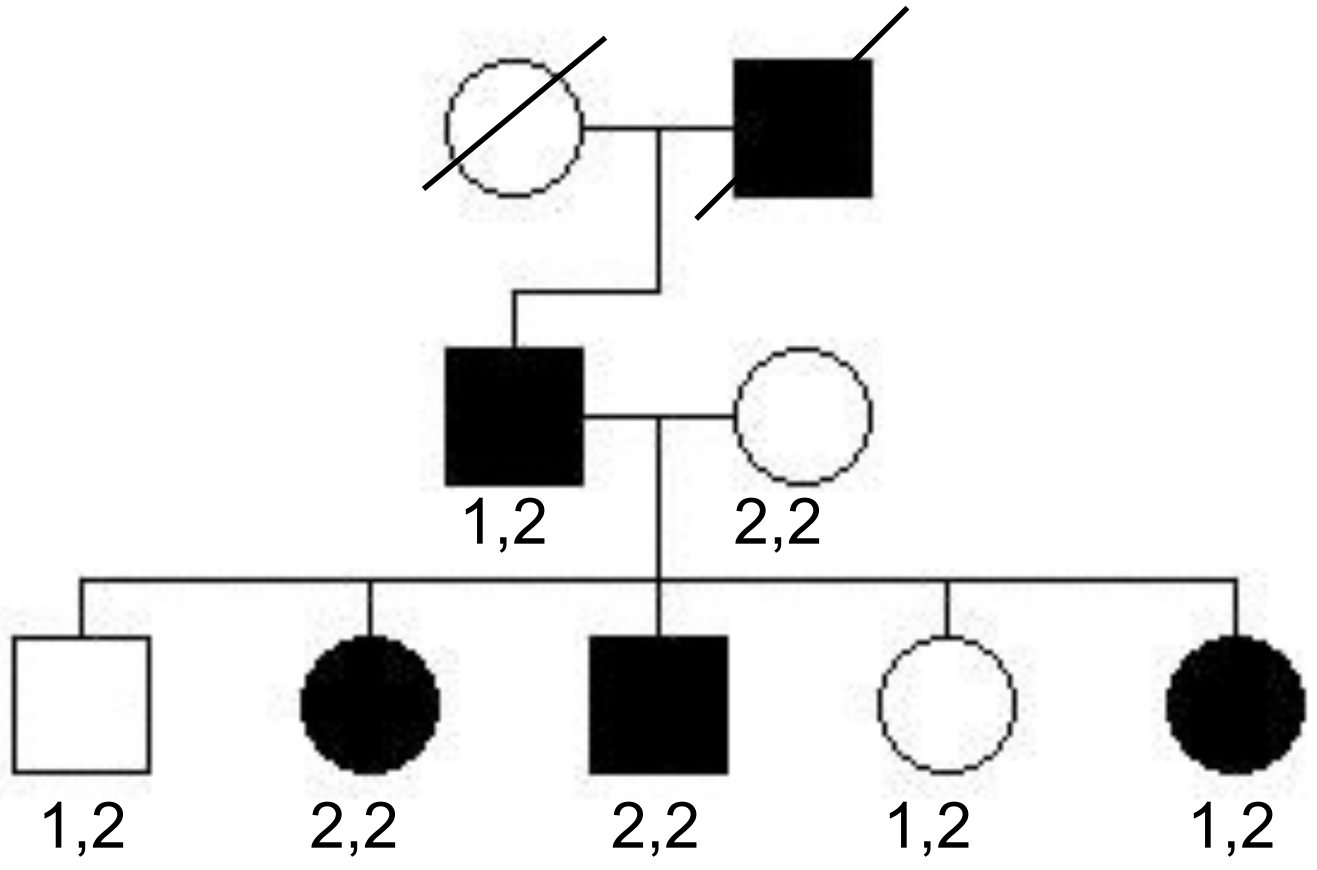
|  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|
| 0 | 0.02 | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 |
| -∞ | -0.23 | 0.32 | 0.42 | 0.36 | 0.22 | 0 |



| | | | | | | |
|---|---|---|---|---|---|---|
| 0 | 0.02 | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 |
| -∞ | -0.23 | 0.32 | 0.42 | 0.36 | 0.22 | 0 |

# Marker phase in the father unknown





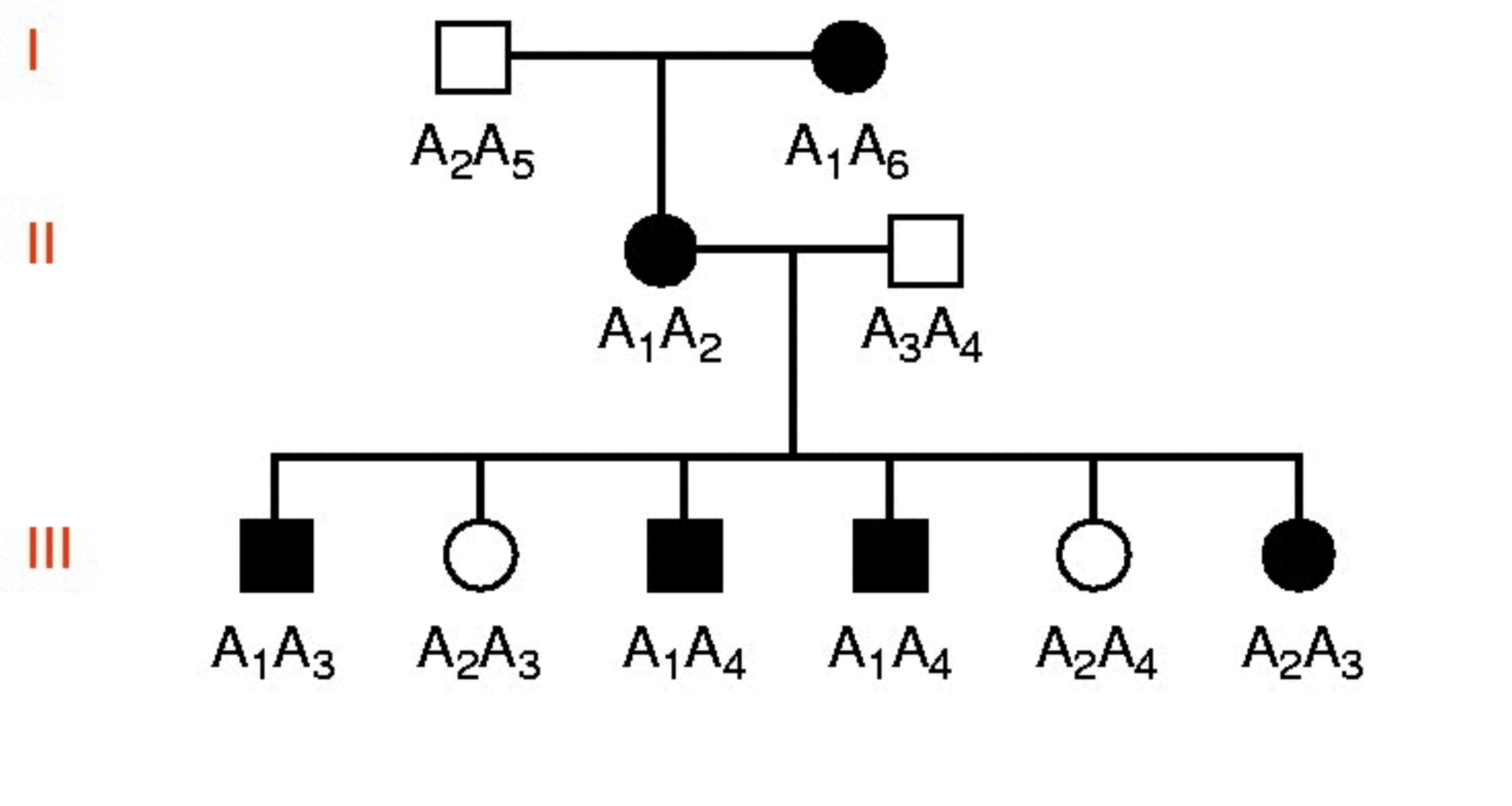
■ $\dfrac{1\ \ +}{2\ \ -}$ or ■ $\dfrac{2\ \ +}{1\ \ -}$

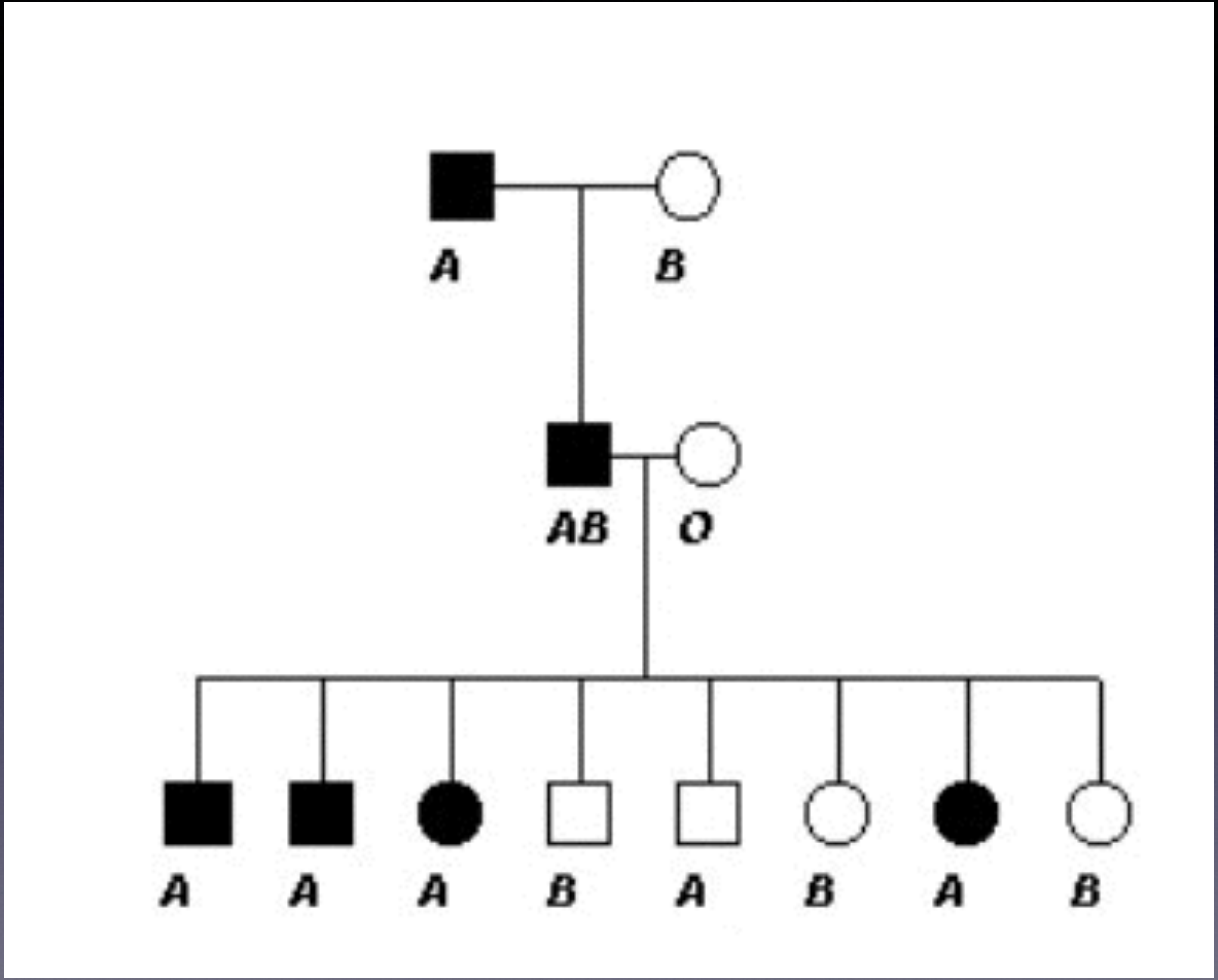
1R 4NR

1NR 4R

$$L(\theta = 0,2) = \left(\frac{0,2 \cdot 0,8^4}{2}\right) \quad + \quad L(\theta = 0,2) = \left(\frac{0,2^4 \cdot 0,8}{2}\right)$$

$$Z(\theta = 0,2) = \log_{10}\left(\frac{\dfrac{0,2 \cdot 0,8^4}{2} + \dfrac{0,2^4 \cdot 0,8}{2}}{0,5^5}\right) \approx 0,12$$


I
II
III
$A_2A_5$
$A_1A_6$
$A_1A_2$
$A_3A_4$
$A_1A_3$
$A_2A_3$
$A_1A_4$
$A_1A_4$
$A_2A_4$
$A_2A_3$


A
B
AB
O
A
A
A
B
A
B
A
B

# Nail-patella syndrome