Introduction to QTL mapping

Manuel Ferreira

Boulder Introductory Course 2006

Outline

1. Aim
2. The Human Genome
3. Principles of Linkage Analysis
4. Parametric Linkage Analysis
5. Nonparametric Linkage Analysis

1. Aim

QTL mapping

- LOCALIZE and then IDENTIFY a locus that regulates a trait (QTL)

↓

Nucleotide or sequence of nucleotides with variation in the population, with different variants associated with different trait levels.

For a heritable trait...

Linkage: localize region of the genome where a QTL that regulates the trait is likely to be harboured

Family-specific phenomenon:
Affected individuals in a family share the same ancestral predisposing DNA segment at a given QTL

Association: identify a QTL that regulates the trait

Population-specific phenomenon:
Affected individuals in a population share the same ancestral predisposing DNA segment at a given QTL

2. Human Genome

DNA structure

- ▷ A DNA molecule is a linear backbone of alternating sugar residues and phosphate groups

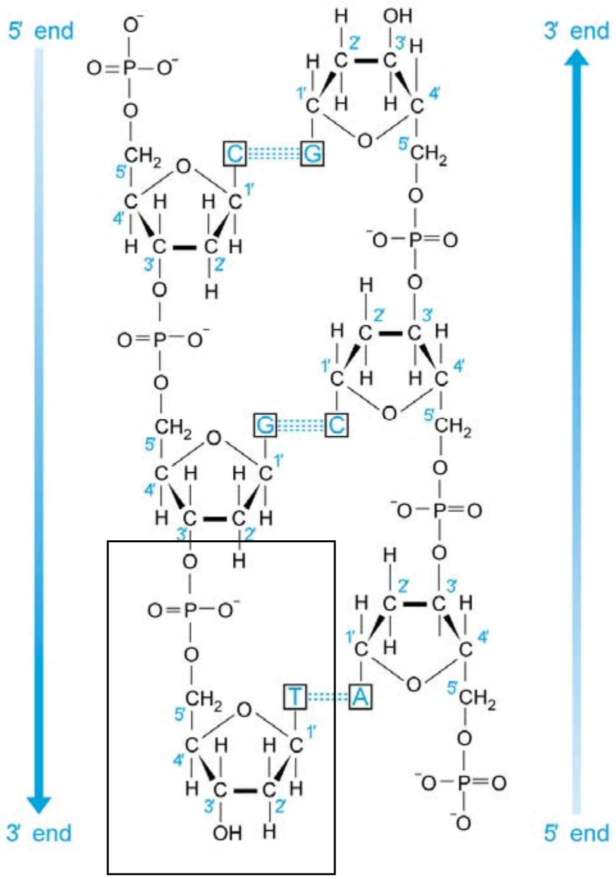
- ▷ Attached to carbon atom 1’ of each sugar is a nitrogenous base: A, C, G or T

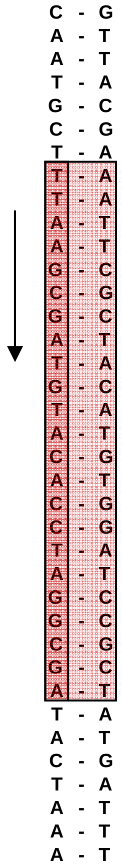
- ▷ Two DNA molecules are held together in anti-parallel fashion by hydrogen bonds between bases [Watson-Crick rules] *Antiparallel double helix*

- ▷ A gene is a segment of DNA which is transcribed to give a protein or RNA product

- ▷ Only one strand is read during gene transcription

- ▷ Nucleotide: 1 phosphate group + 1 sugar + 1 base



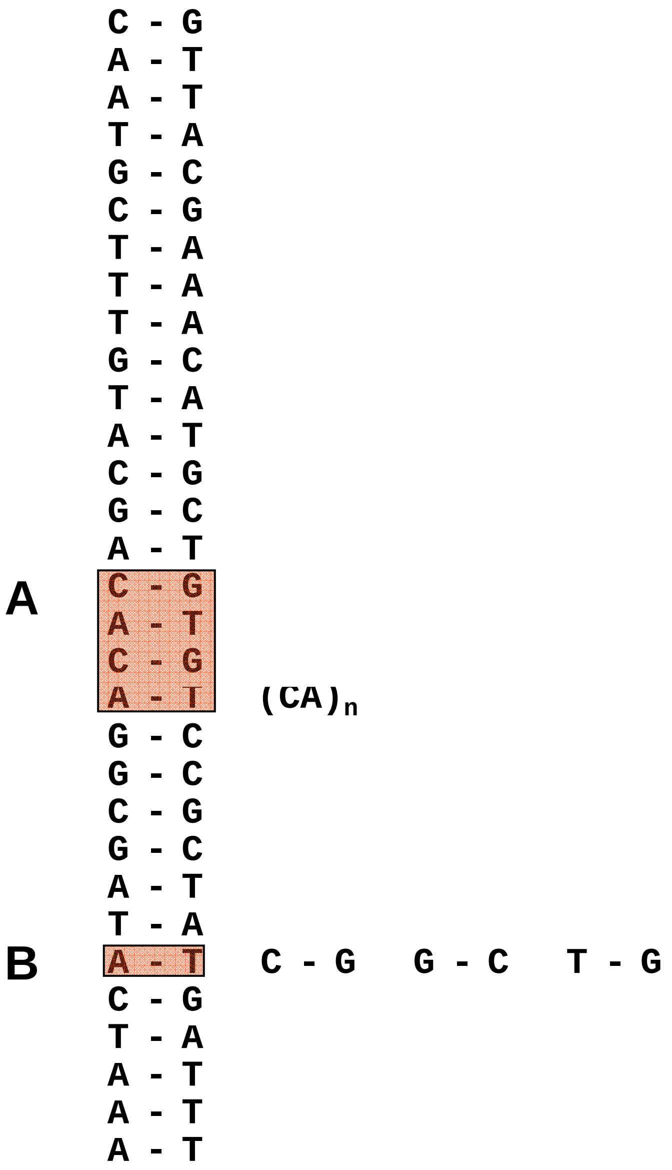


C - G
A - T
A - T
T - A
G - C
C - G
T - A
T - A
T - A
A - T
A - T
G - C
C - G
G - C
A - T
T - A
G - C
T - A
A - T
C - G
A - T
C - G
C - G
T - A
A - T
G - C
G - C
C - G
G - C
A - T
T - A
A - T
C - G
T - A
A - T
A - T
A - T

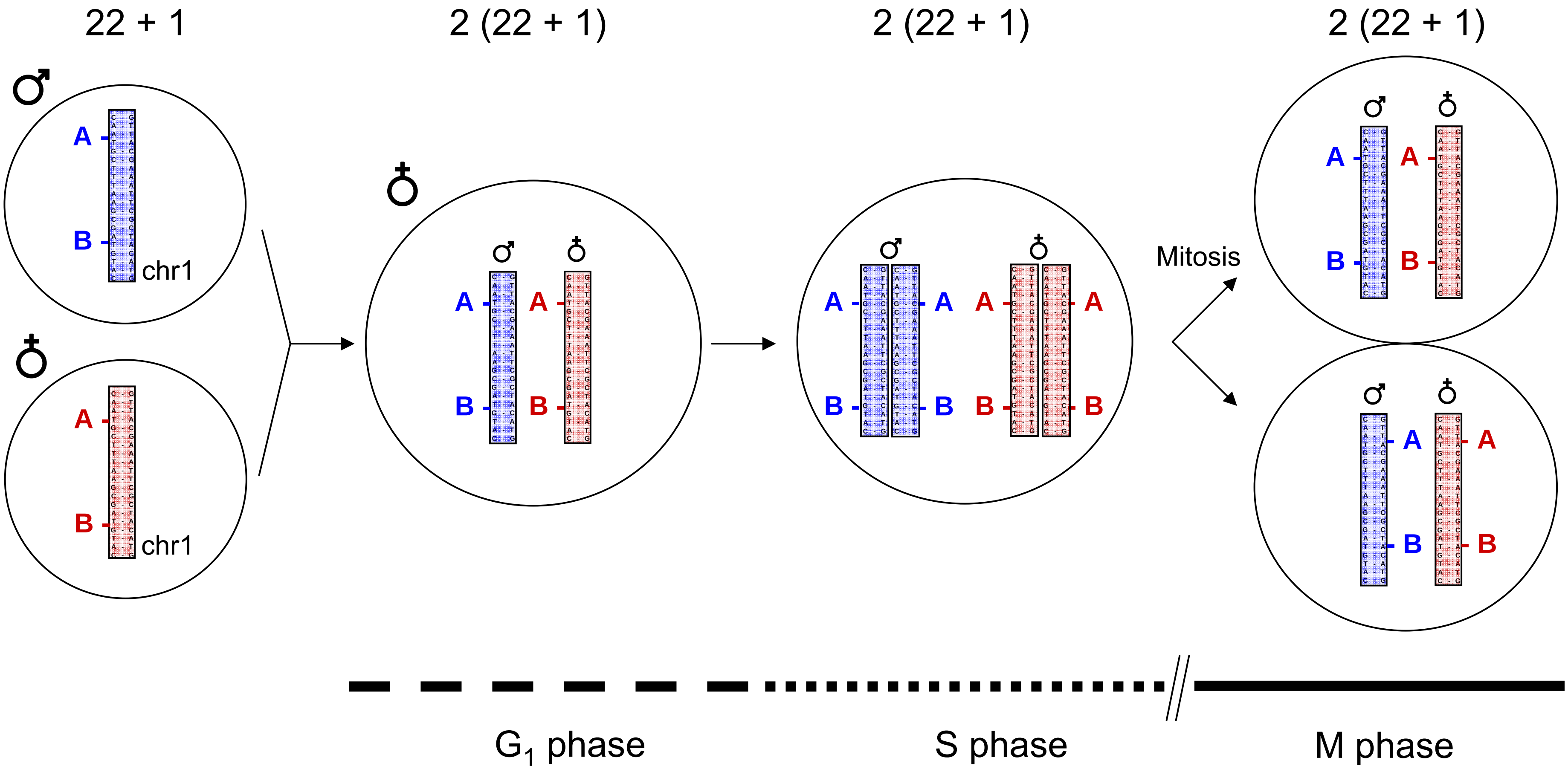
DNA polymorphisms

- ▷ Microsatellites
  >100,000
  Many alleles, $(CA)_n$, very informative, even, easily automated

- ▷ SNPs
  10,054,521 (25 Jan ‘05)
  10,430,753 (11 Mar ‘06)
  Most with 2 alleles (up to 4), not very informative, even, easily automated



DNA organization



22 + 1

2 (22 + 1)

2 (22 + 1)

2 (22 + 1)

Mitosis

G_1 phase

S phase

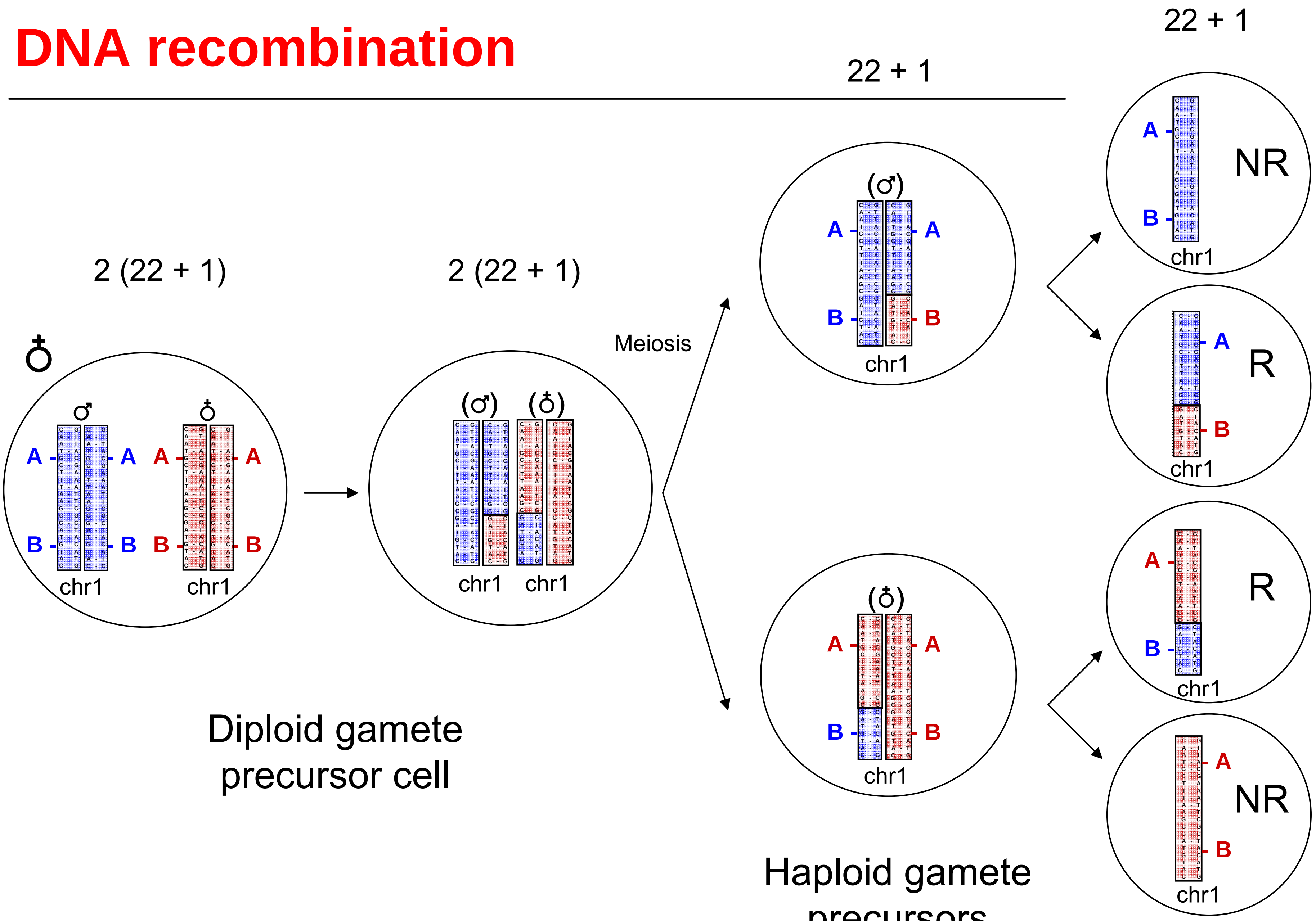
M phase

Haploid
gametes

Diploid zygote
1 cell

Diploid zygote
>1 cell

DNA recombination

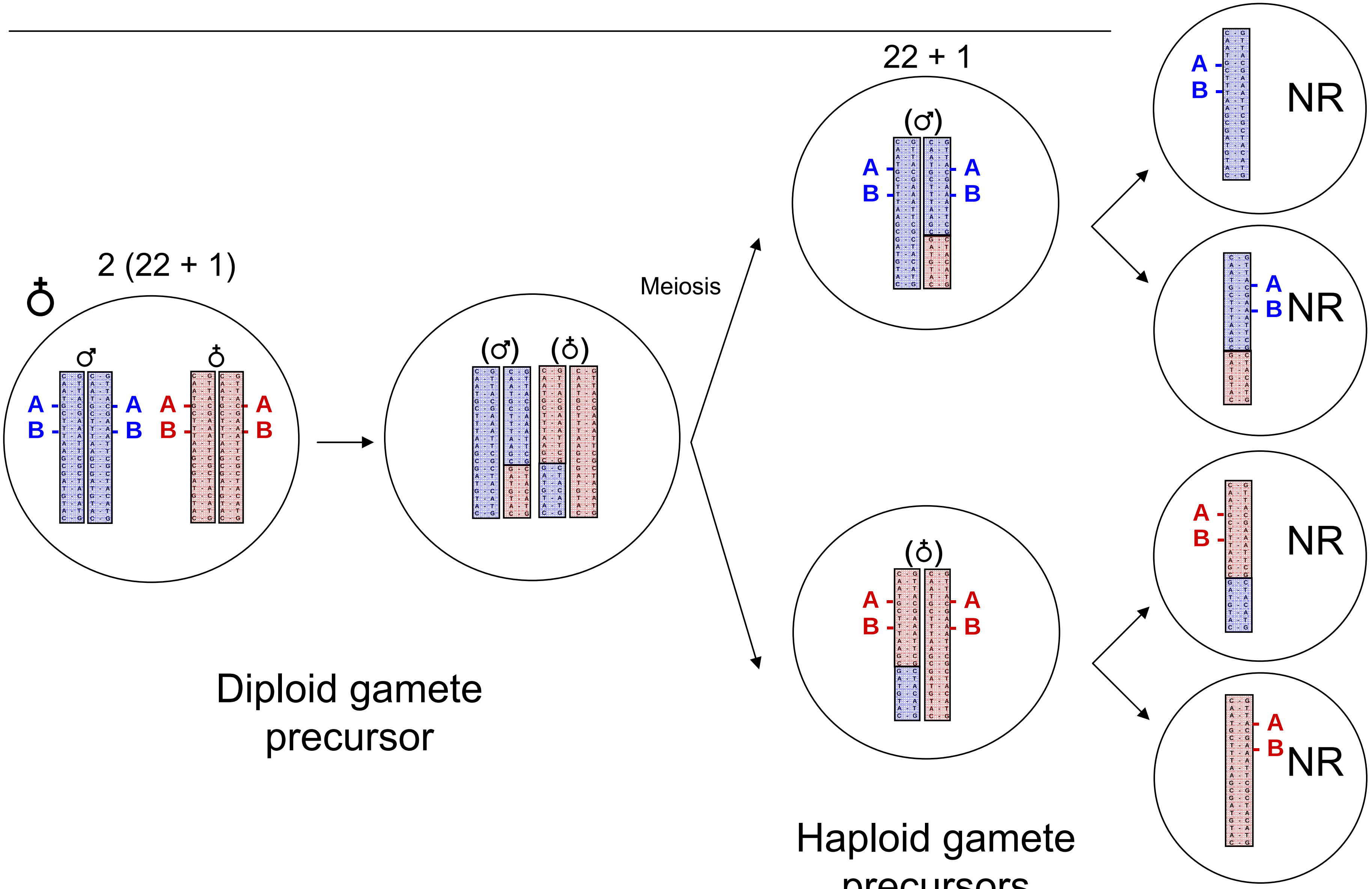


Diploid gamete precursor cell

Haploid gamete precursors

Hap. gametes

DNA recombination between linked loci



Diploid gamete precursor

Haploid gamete precursors

Hap. gametes

Human Genome - summary

- DNA is a linear sequence of nucleotides partitioned into 23 chromosomes

  Two copies of each chromosome (2x22 autosomes + XY), from paternal and maternal origins. During meiosis in gamete precursors, recombination can occur between maternal and paternal homologs

- Recombination fraction between loci A and B (θ)

  Proportion of gametes produced that are recombinant for A and B

  If A and B are very far apart: 50%R:50%NR - $\theta = 0.5$

  If A and B are very close together: <50%R - $0 \leq \theta < 0.5$

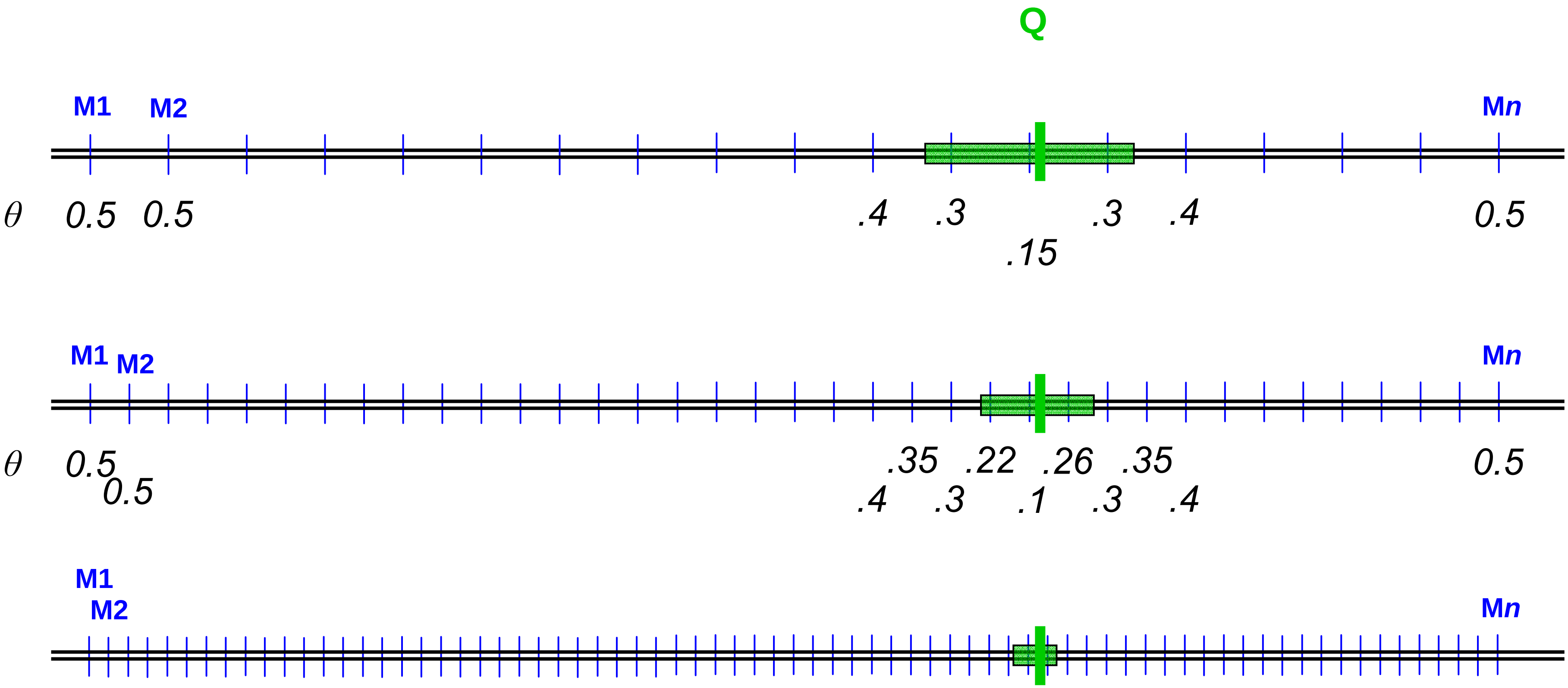
- Recombination fraction (θ) can be converted to genetic distance (cM)

  Haldane: $cM = 100 \cdot [-0.5 \cdot \ln(1 - 2 \cdot \theta)]$ eg. θ=0.17, cM=20.8

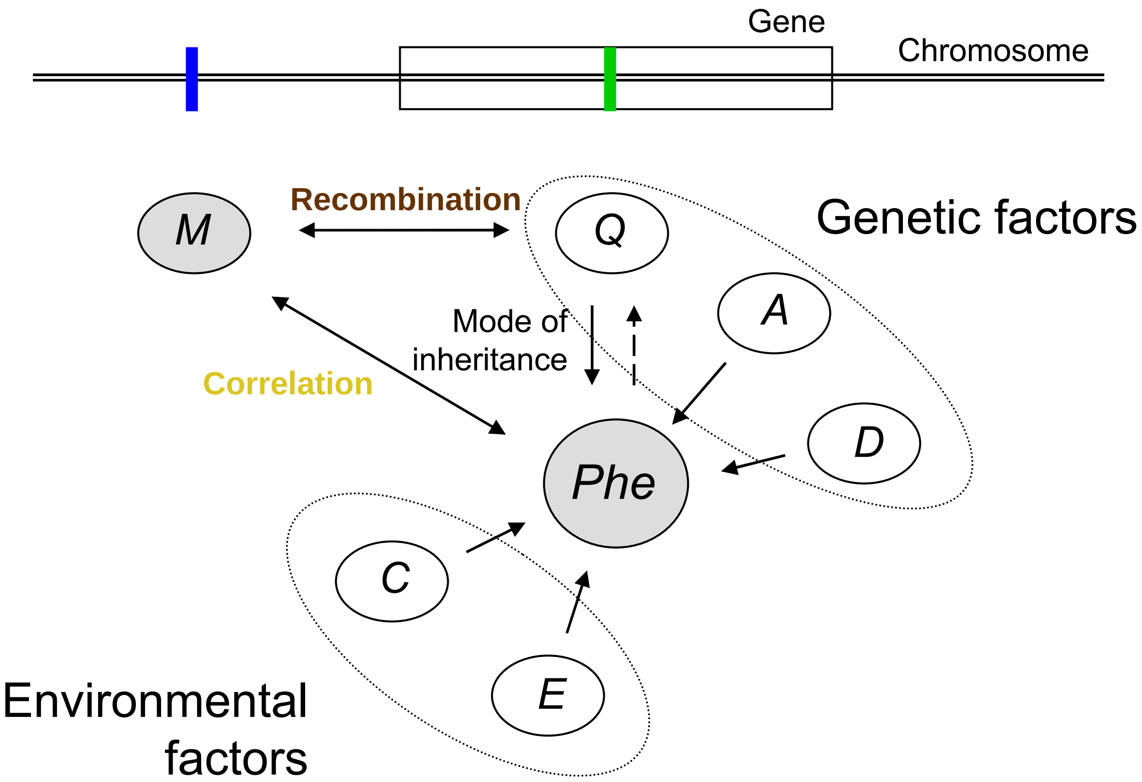
  Kosambi: $cM = 100 \cdot [0.25 \cdot \ln((1 + 2 \cdot \theta)/(1 - 2 \cdot \theta))]$ eg. θ=0.17, cM=17.7

3. Principles of Linkage Analysis

Linkage Analysis requires genetic markers



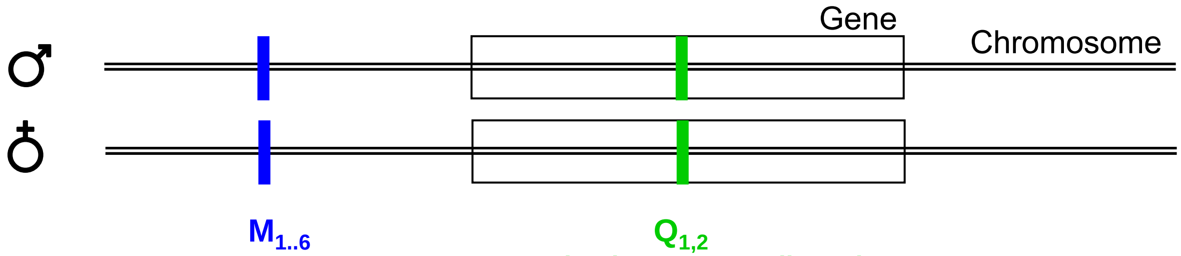
Linkage Analysis: Parametric vs. Nonparametric



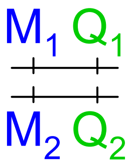
Adapted from Weiss & Terwilliger 2000

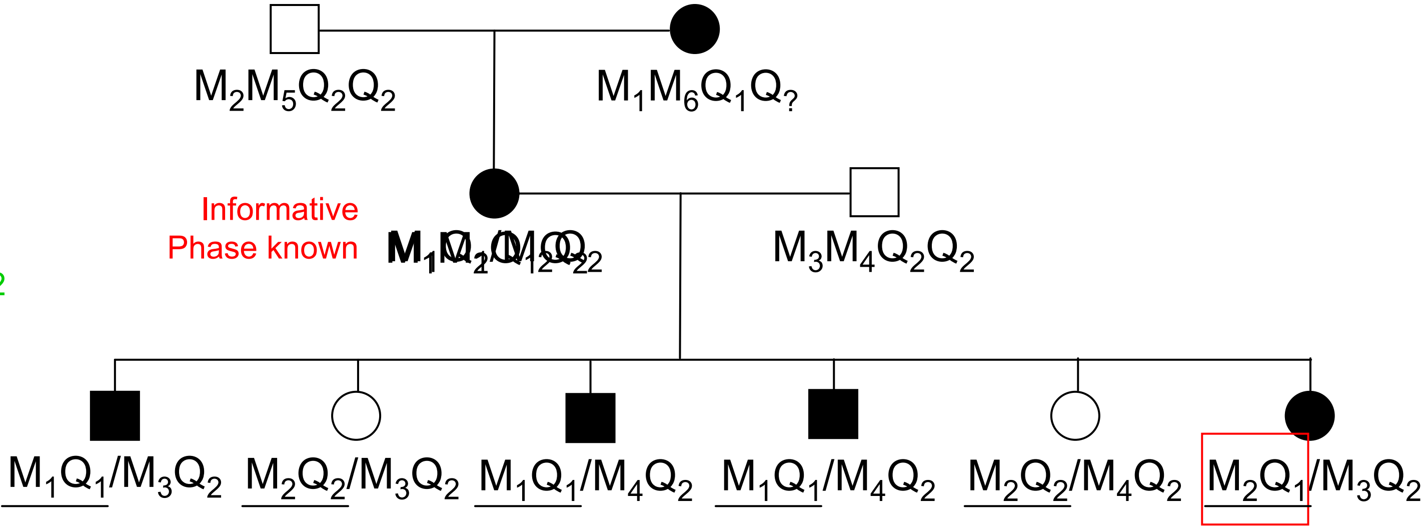
4. Parametric Linkage Analysis

Linkage with informative phase known meiosis



Autosomal dominant, Q_1 predisposing allele





NR: M_1Q_1

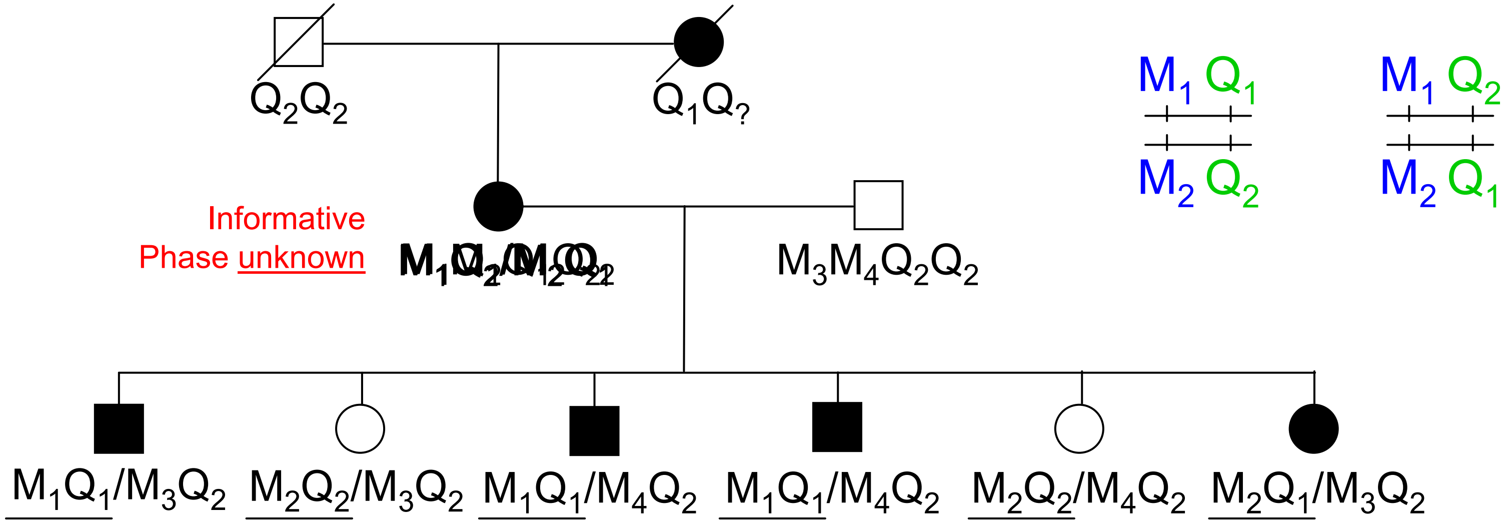
NR: M_2Q_2

R: M_1Q_2

R: M_2Q_1

$$\theta_{MQ} = 1/6 = 0.17 \quad (\sim 20.8\ cM)$$

Linkage with informative phase unknown meiosis



M_1Q_1/M_2Q_2	*P*	*N*
NR: M_1Q_1	½(1-θ)	3
NR: M_2Q_2	½(1-θ)	2
R: M_1Q_2	½θ	0
R: M_2Q_1	½θ	1

M_1Q_2/M_2Q_1	*P*	*N*
R: M_1Q_1	½θ	3
R: M_2Q_2	½θ	2
NR: M_1Q_2	½(1-θ)	0
NR: M_2Q_1	½(1-θ)	1

$$L(X\mid\theta)=\frac{1}{2}\cdot\left[\theta^1\cdot(1-\theta)^5\right]+\frac{1}{2}\cdot\left[\theta^5\cdot(1-\theta)^1\right]$$

$$L(X\mid\theta=0.5)=\frac{1}{2}\cdot\left[0.5^1\cdot(1-0.5)^5\right]+\frac{1}{2}\cdot\left[0.5^5\cdot(1-0.5)^1\right]=(0.5)^6$$

Parametric LOD score calculation

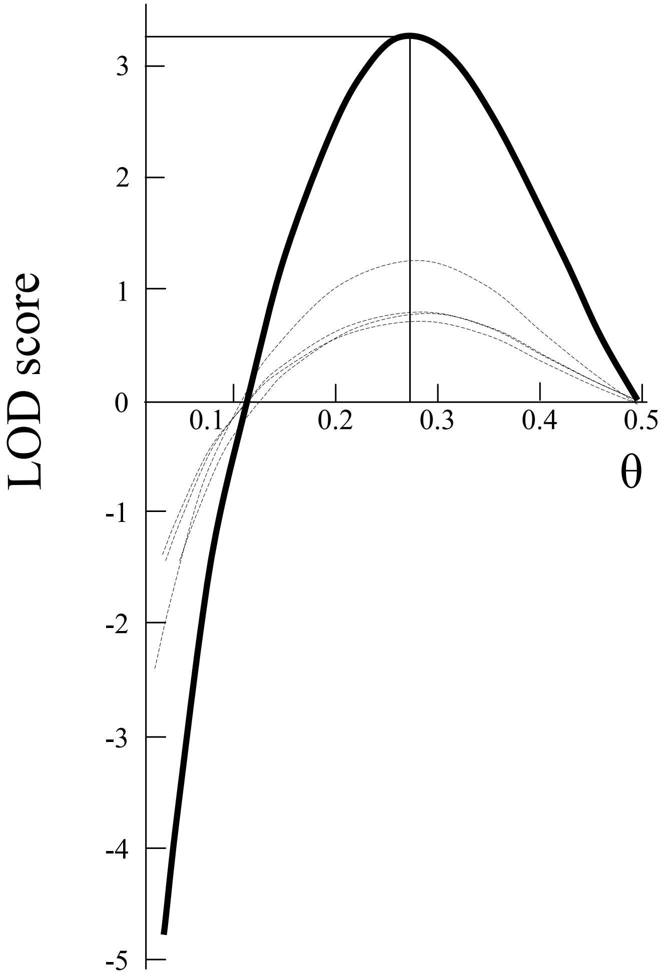
$$OD = \frac{L(X \mid \theta)}{L(X \mid \theta = 0.5)} \qquad LOD = \log_{10} \frac{L(X \mid \theta)}{L(X \mid \theta = 0.5)}$$

$$LOD = \log_{10} \frac{\frac{1}{2} \cdot \left[\theta^1 \cdot (1-\theta)^5\right] + \frac{1}{2} \cdot \left[\theta^5 \cdot (1-\theta)^1\right]}{(0.5)^6}$$

$$OD = \prod_{i=1}^{n} \frac{L(X_i \mid \theta)}{L(X_i \mid \theta = 0.5)}$$

$$LOD = \log_{10} \left(\prod_{i=1}^{n} \frac{L(X_i \mid \theta)}{L(X_i \mid \theta = 0.5)} \right)$$

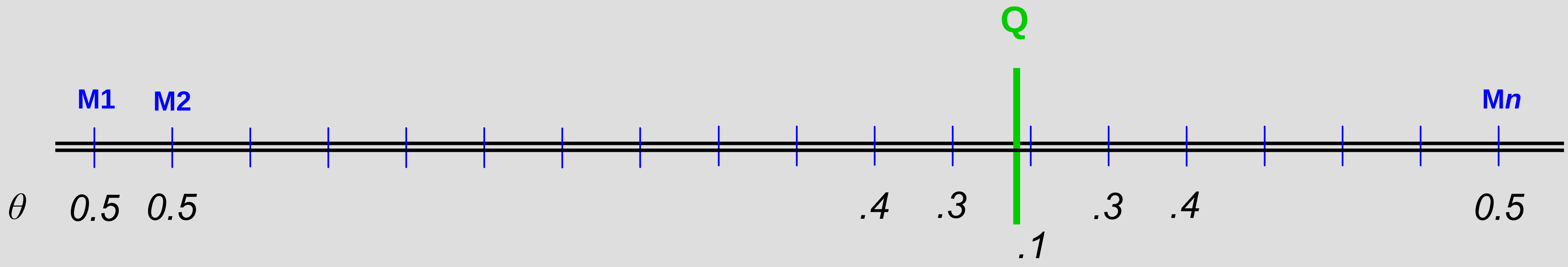
$$LOD = \sum_{i=1}^{n} \log_{10} \left(\frac{L(X_i \mid \theta)}{L(X_i \mid \theta = 0.5)} \right) = \sum_{i=1}^{n} LOD_i$$



- ▷ Overall LOD score for a given θ is the sum of all family LOD scores at θ

eg. LOD=3 for θ=0.28

Parametric Linkage Analysis - summary



- For each marker, estimate the θ that yields highest LOD score across all families

- This θ (and the LOD) will depend upon the mode of inheritance assumed

  MOI determines the genotype at the trait locus Q and thus determines the number of meiosis which are recombinant or nonrecombinant. Limited to Mendelian diseases.

- Markers with a significant parametric LOD score (>3) are said to be linked to the trait locus with recombination fraction θ

Outline

1. Aim
2. The Human Genome
3. Principles of Linkage Analysis
4. Parametric Linkage Analysis
5. Nonparametric Linkage Analysis

5. Nonparametric Linkage Analysis

Approach

- Parametric: genotype marker locus & genotype trait locus

  (latter inferred from phenotype according to a specific disease model)

  Parameter of interest: θ between marker and trait loci

- Nonparametric: genotype marker locus & phenotype

  If a trait locus truly regulates the expression of a phenotype, then two relatives with similar phenotypes should have similar genotypes at a marker in the vicinity of the trait locus, and vice-versa.

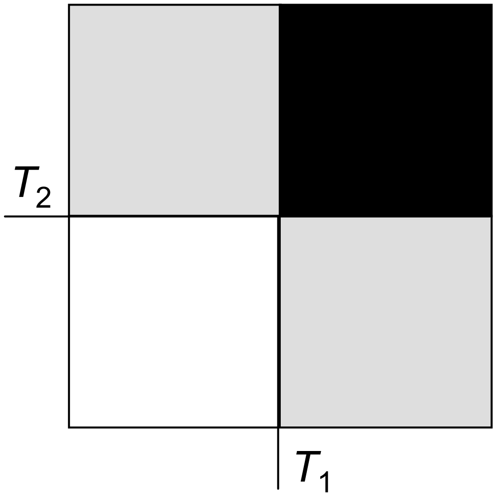
  Interest: correlation between phenotypic similarity and marker genotypic similarity

  No need to specify mode of inheritance, allele frequencies, etc...

Phenotypic similarity between relatives

- Squared trait differences $(X_1 - X_2)^2$
- Squared trait sums $(X_1 + X_2)^2$
- Trait cross-product $[(X_1 - \mu)\cdot(X_2 - \mu)]$
- Trait variance-covariance matrix $\left\{\begin{matrix} Var(X_1) & Cov(X_1X_2) \\ Cov(X_1X_2) & Var(X_2) \end{matrix}\right\}$

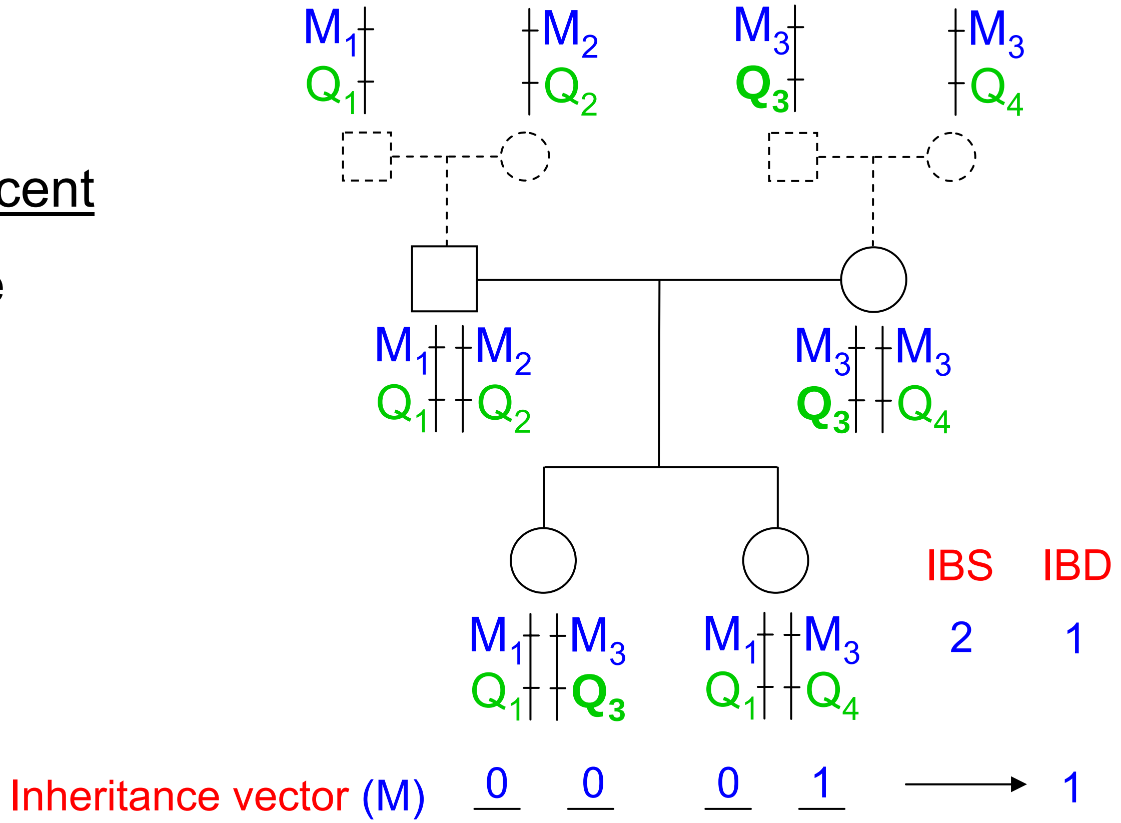
- Affection concordance



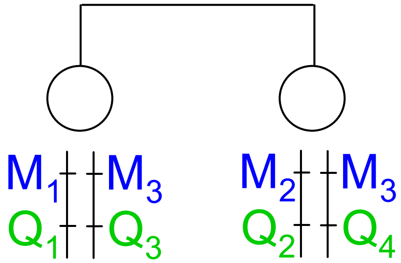
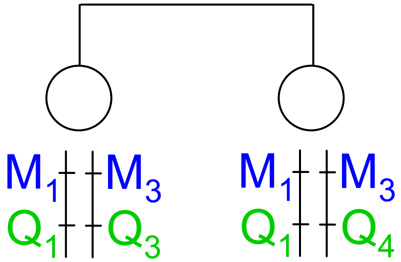
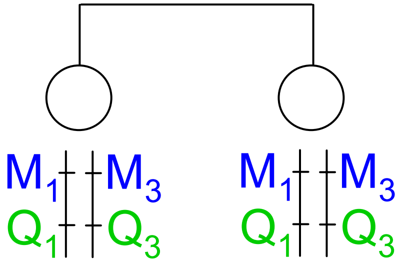
Genotypic similarity between relatives

- IBS Alleles shared Identical By State “look the same”, may have the same DNA sequence but they are not necessarily derived from a known common ancestor

- IBD Alleles shared Identical By Descent are a copy of the same ancestor allele

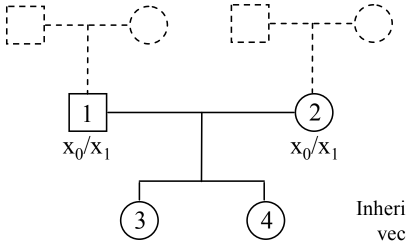


Genotypic similarity between relatives - π

	Inheritance vector (M)	Number of alleles shared IBD	Proportion of alleles shared IBD - π
	0 0 1 1	0	0
	0 0 0 1	1	0.5
	0 0 0 0	2	1

Genotypic similarity between relatives - $\hat{\pi}$

A **B** **C** **D**



2^{2n}

3	4	Inheritance vector	IBD
x_0/x_0	x_0/x_0	0000	2
x_0/x_0	x_0/x_1	0001	1
x_0/x_0	x_1/x_0	0010	1
x_0/x_0	x_1/x_1	0011	0
x_0/x_1	x_0/x_0	0100	1
x_0/x_1	x_0/x_1	0101	2
x_0/x_1	x_1/x_0	0110	0
x_0/x_1	x_1/x_1	0111	1
x_1/x_0	x_0/x_0	1000	1
x_1/x_0	x_0/x_1	1001	0
x_1/x_0	x_1/x_0	1010	2
x_1/x_0	x_1/x_1	1011	1
x_1/x_1	x_0/x_0	1100	0
x_1/x_1	x_0/x_1	1101	1
x_1/x_1	x_1/x_0	1110	1
x_1/x_1	x_1/x_1	1111	2

P (IBD=0)
P (IBD=1)
P (IBD=2)

$$\hat{\pi} =$$

Practical

▷ **Aim** (1) Estimate IBD with MERLIN; (2) IBD estimation can be influenced by genotyped individuals and allele frequencies; (3) compute $\hat{\pi}$

```
H:\manuel  -     Copy folder “Linkage” to C:\
```

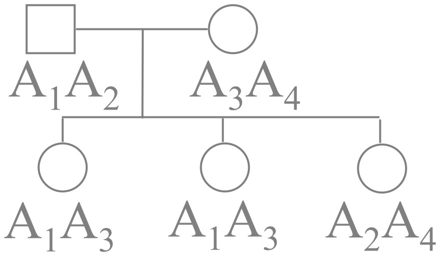
1. Open with Notepad: pr1.ped pr1.dat pr1.map pr1.freq
2. Start>Run>C:/Linkage/pfe32.exe
3. Run Command Prompt
4. Keep a File Explorer window open

Exercice1

(1) Estimate IBD for pedigrees **A**, **B** and **C** in the previous slide

(2) Change allele frequencies (pr1.freq) from 0.25 0.25 0.25 0.25 to
(i) 0.45 0.25 0.25 0.05 and
(ii) 0.05 0.25 0.25 0.45

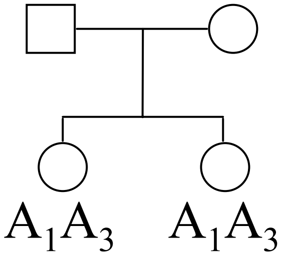
Practical



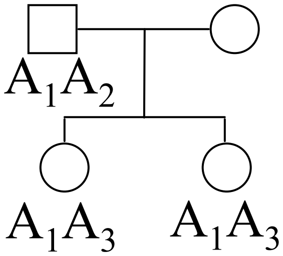
Exercice 2

(1) Modify pr1.ped and estimate IBD probabilities and $\hat{\pi}$ between twin 1 and twin 2 for pedigrees **E**, **F** and **G**:

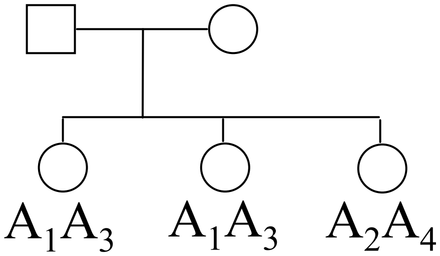
E



F



G



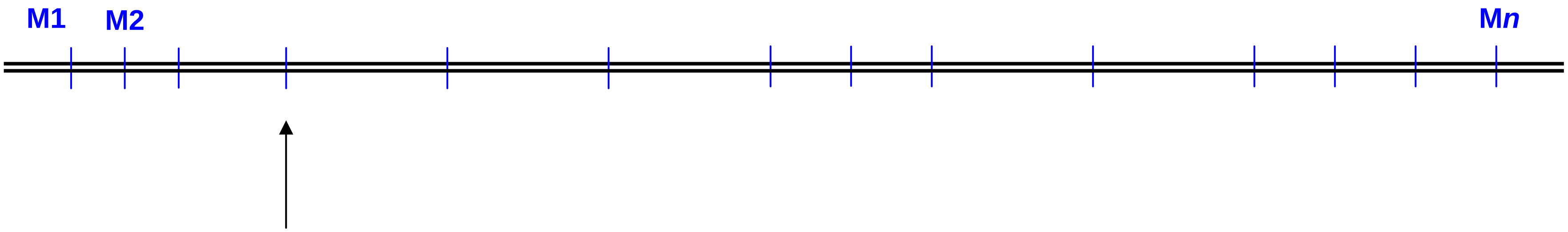
P(IBD=0)

P(IBD=1)

P(IBD=2)

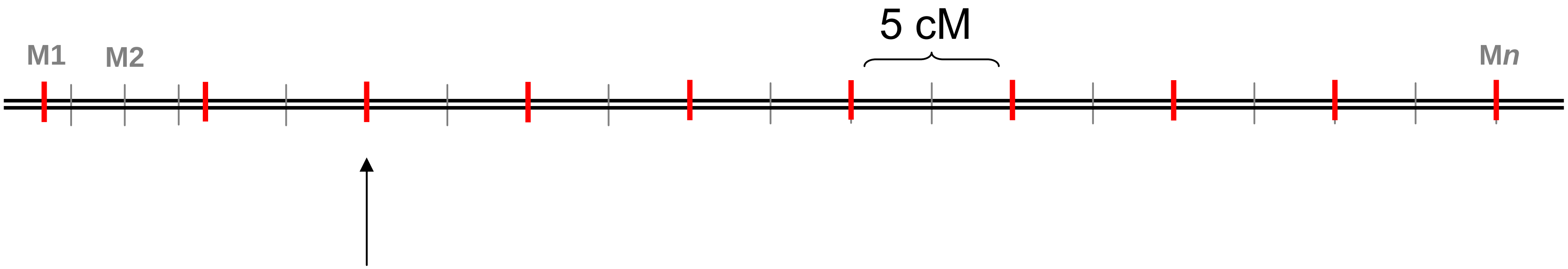
$\hat{\pi}$

Allele frequencies on pr1.freq: 0.25 0.25 0.25 0.25



IBD at a marker

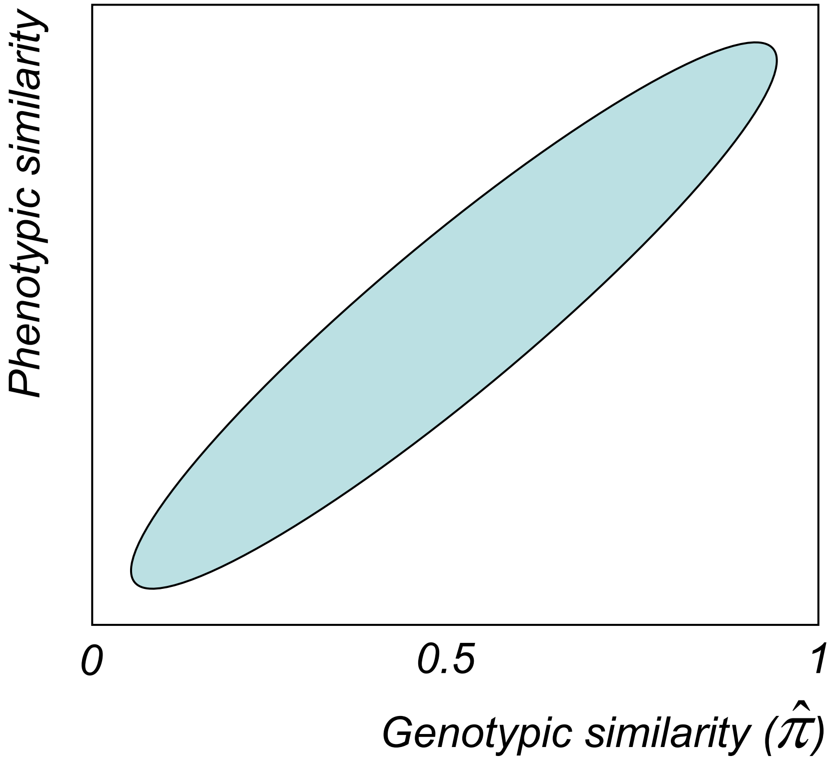
Singlepoint IBD



IBD at a 'grid'

Multipoint IBD

Statistics that incorporate both phenotypic and genotypic similarities

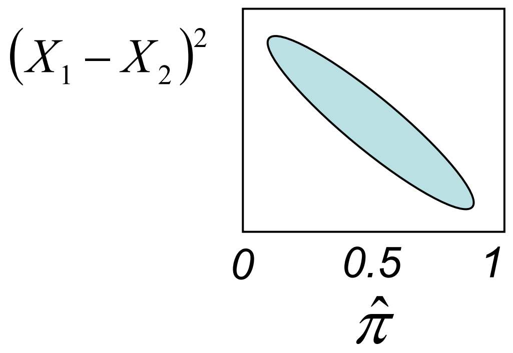


Haseman-Elston regression – Quantitative traits

$$E\left[\left(X_1 - X_2\right)^2 \mid \hat{\pi}\right]$$

$$= E\left[\left(X_1^2 + X_2^2 - 2 \cdot X_1 \cdot X_2\right) \mid \hat{\pi}\right]$$

$$= Var\left(X_1\right) + Var\left(X_2\right) - 2Cov\left(X_1 X_2 \mid \hat{\pi}\right)$$



$$Var\left(X_1\right) = Var\left(X_2\right) = V_Q + V_A + V_C + V_E$$

$$Cov\left(X_1, X_2 \mid \hat{\pi}\right) = \hat{\pi} \cdot V_Q + \frac{1}{2} \cdot V_A + V_C$$

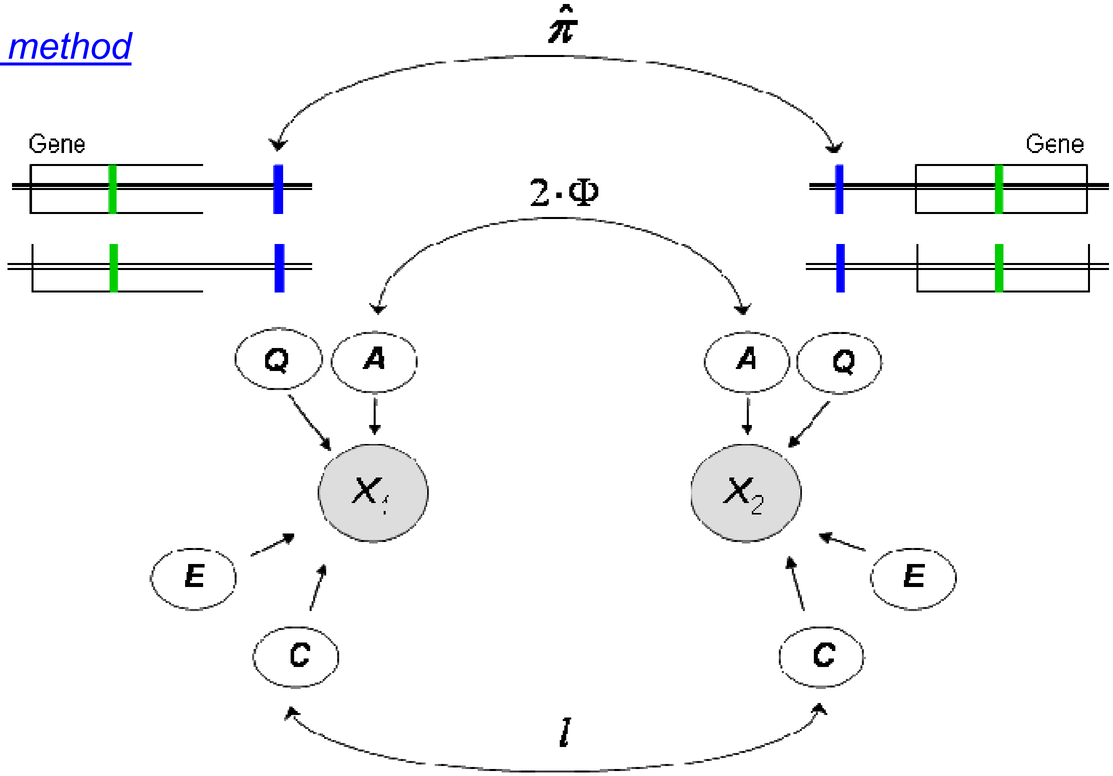
	X_1	X_2	$(X_1-X_2)^2$	$\hat{\pi}$
1	2.2	2.1	0.01	0.9
2	1.9	2.3	0.16	0.6
3	2.3	2.6	0.09	0.7
4	3.4	1.6	3.24	0.1
5	2.5	2.3	0.04	0.8
…				
1000	2.4	2.4	0	0.9

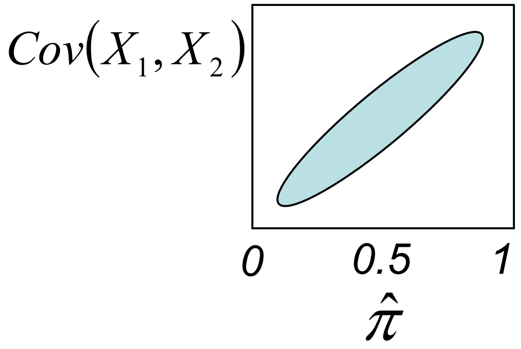
$$E\left[\left(X_1 - X_2\right)^2 \mid \hat{\pi}\right] = -2 \cdot V_Q \cdot \underset{\uparrow}{\hat{\pi}} + \underset{\uparrow}{\underline{2 \cdot V_Q + V_A + 2 \cdot V_E}}$$

Phenotypic dissimilarity = ***b* ×** Genotypic similarity **+ *c***

VC ML – Quantitative & Categorical traits

$\hat{\pi}$ method





H_1:

$$Var(X_1) = Var(X_2) = V_Q + V_A + V_C + V_E$$

$$Cov(X_1, X_2 \mid \hat{\pi}) = \hat{\pi} \cdot V_Q + 2 \cdot \Phi \cdot V_A + l \cdot V_C$$

H_0:

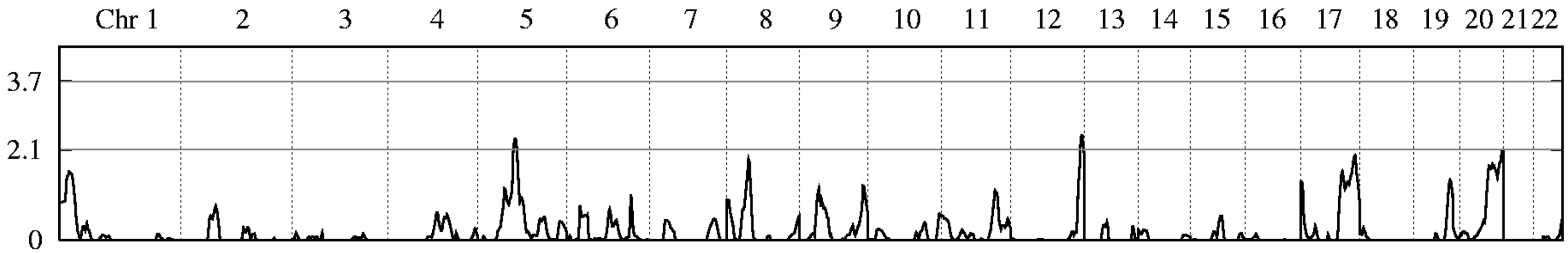
$$Var(X_1) = Var(X_2) = \qquad V_A + V_C + V_E$$

$$Cov(X_1, X_2 \mid \hat{\pi}) = \qquad 2 \cdot \Phi \cdot V_A + l \cdot V_C$$

$$LOD = \log_{10} \frac{L(H_1)}{L(H_0)}$$

e.g. LOD=3

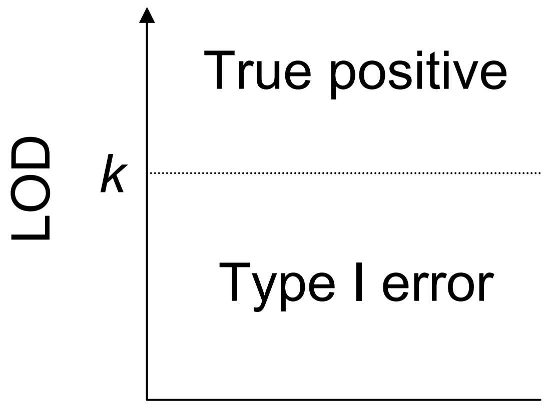
Genome-wide linkage analysis (e.g. VC)



Individual LOD scores can be expressed as *P* values (Pointwise)

LOD $\xrightarrow{(x4.6)}$ Chi-sq (n-df) → *P* value

2.1 9.67 0.0009



True positive

Type I error

→ Theoretical (Lander & Kruglyak 1995)

LOD = 3.6, Chi-sq = 16.7, *P* = 0.000022

Nonparametric Linkage Analysis - summary

- No need to specify mode of inheritance
- Models phenotypic and genotypic similarity of relatives
- Expression of phenotypic similarity, calculation of IBD
- HE and VC are the most popular statistics used for linkage of quantitative traits
- Other statistics available, specially for affection traits