Linkage analysis: basic principles

Manuel Ferreira & Pak Sham

Boulder Advanced Course 2005

Outline

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

1. Aim

For a heritable trait...

Linkage: localizes region of the genome where a locus (loci) 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 trait locus

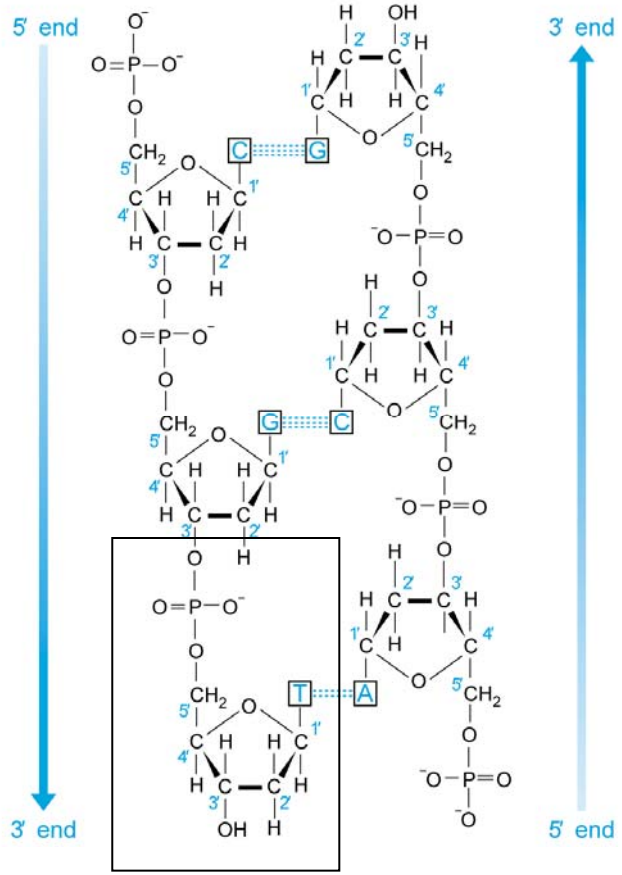
Association: identifies a locus that regulates the trait

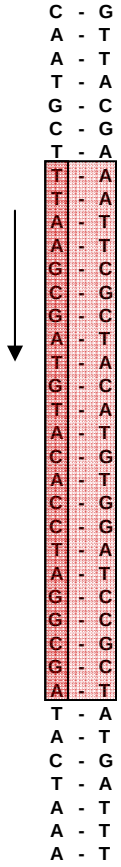
Population-specific phenomenon:
Affected individuals in a population share the same ancestral predisposing DNA segment at a given trait locus

2. Human Genome

DNA structure

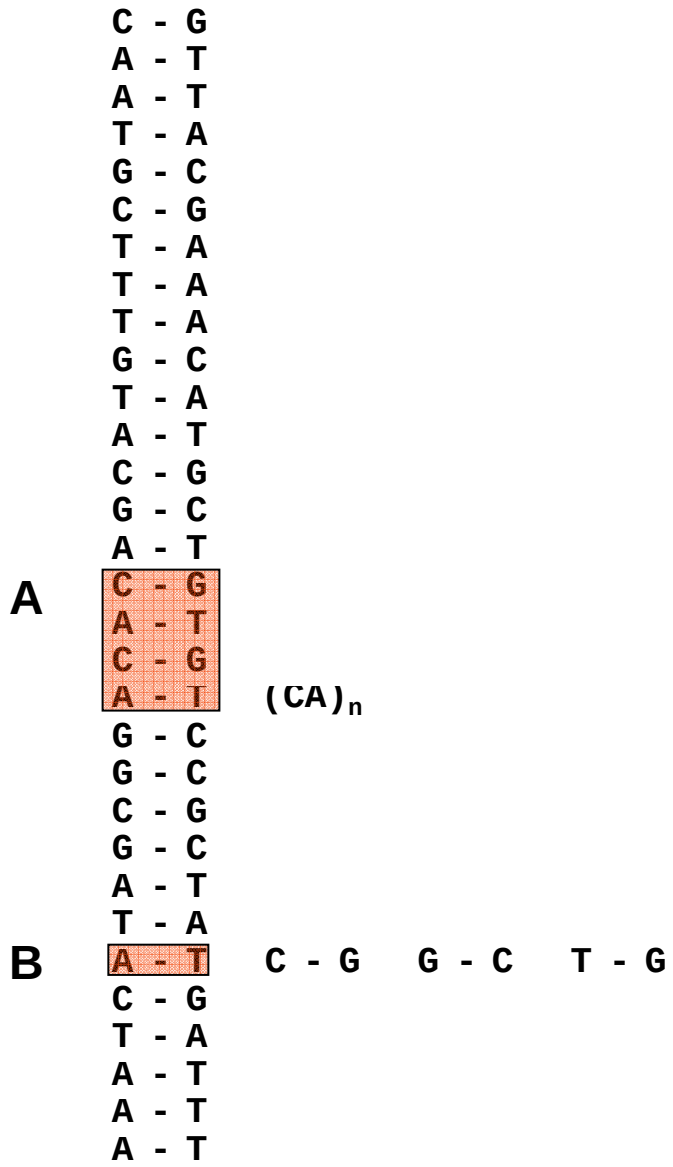
- A DNA molecule is a linear backbone of alternating sugar residues and phosphate groups
- Attached to carbon atom 1' of each sugar is a nitrogenous base: A, C, G or T
- Two DNA molecules are held together in anti-parallel fashion by hydrogen bonds between bases [Watson-Crick rules] *Antiparallel double 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



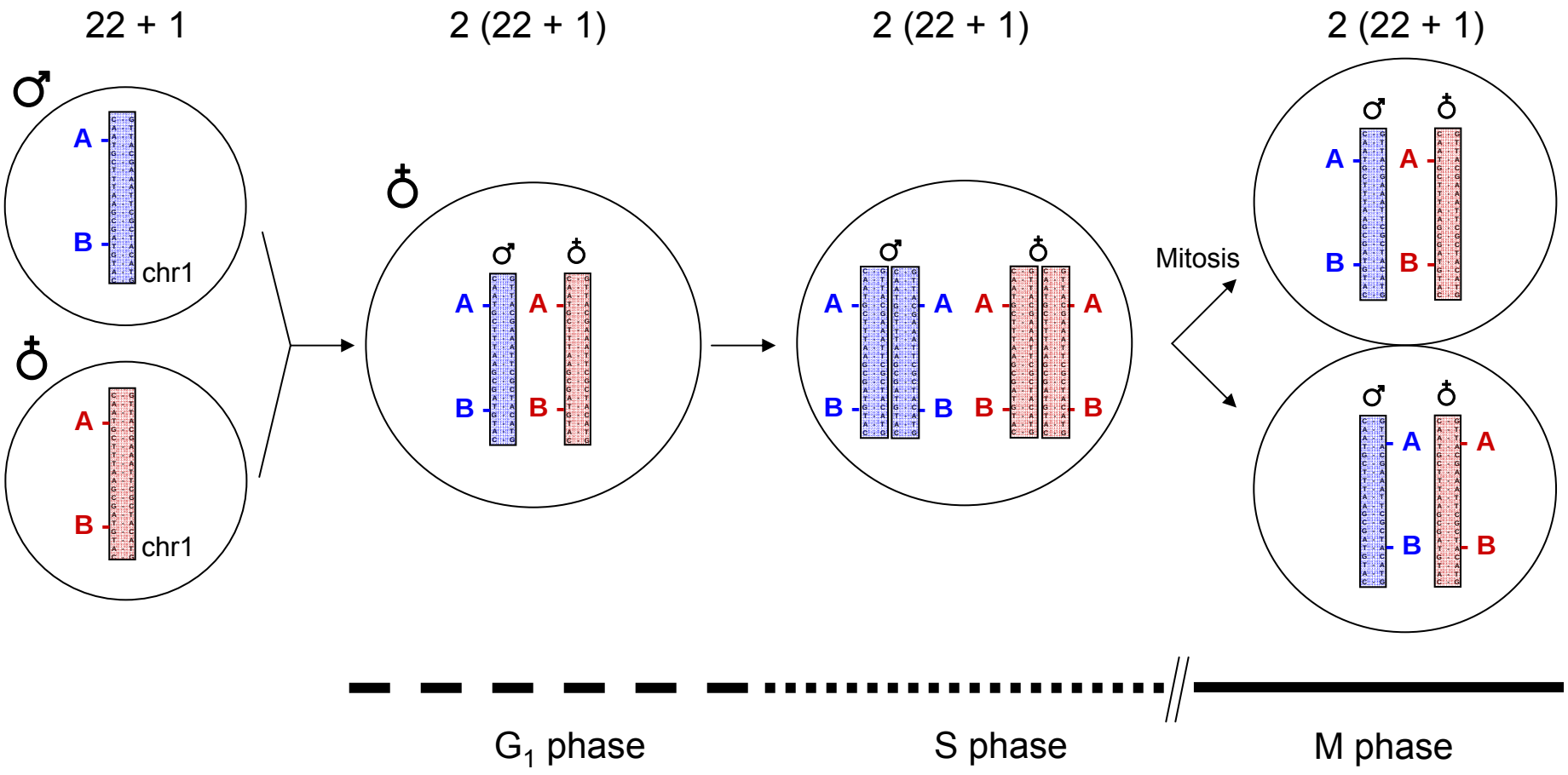


DNA polymorphisms

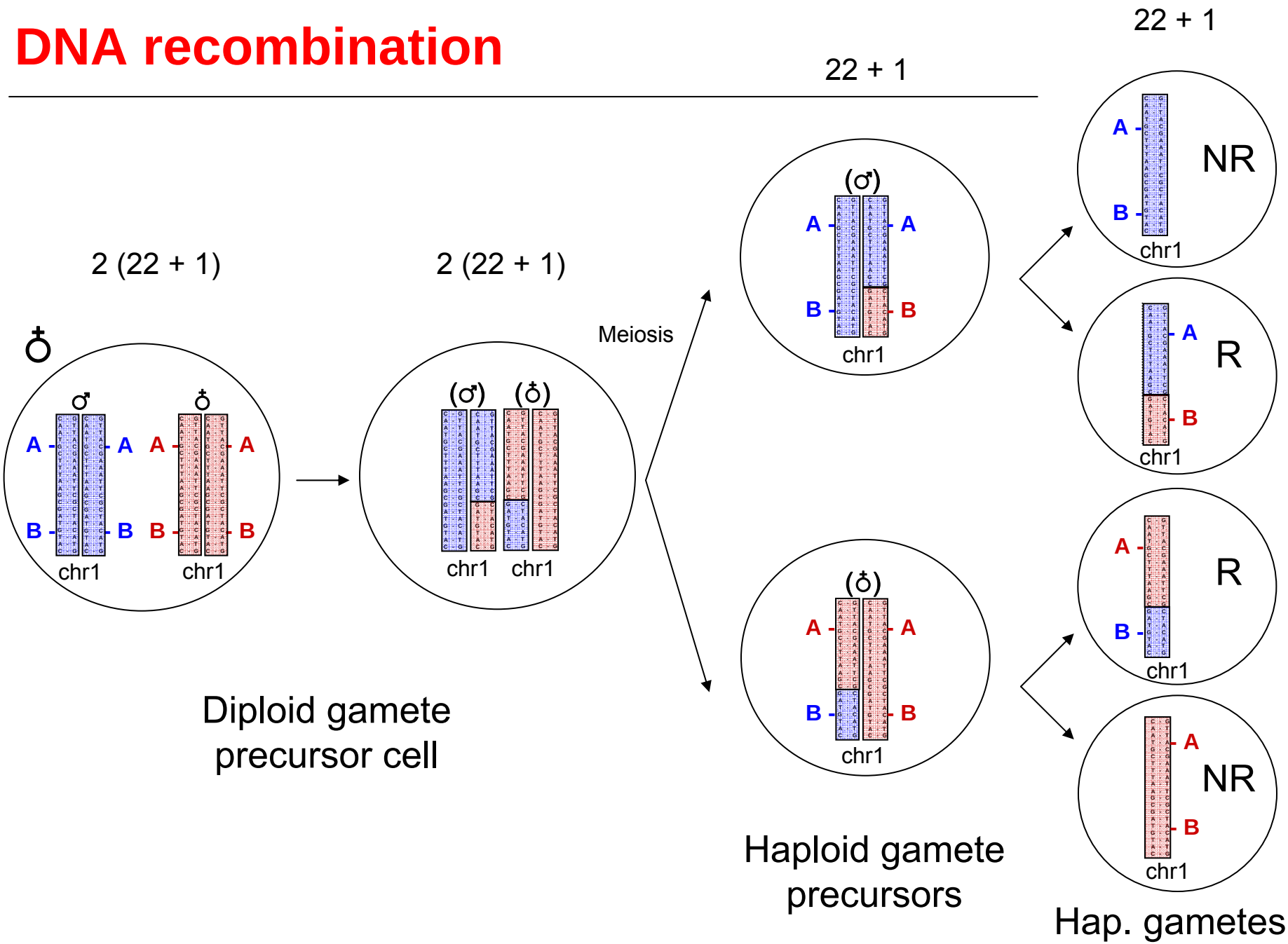
- RFLPs
- Minisatellites
- Microsatellites
  >100,000
  Many alleles, $(CA)_n$, very informative, even, easily automated
- SNPs
  10,054,521 (25 Jan '05)
  Most with 2 alleles (up to 4), not very informative, even, easily automated



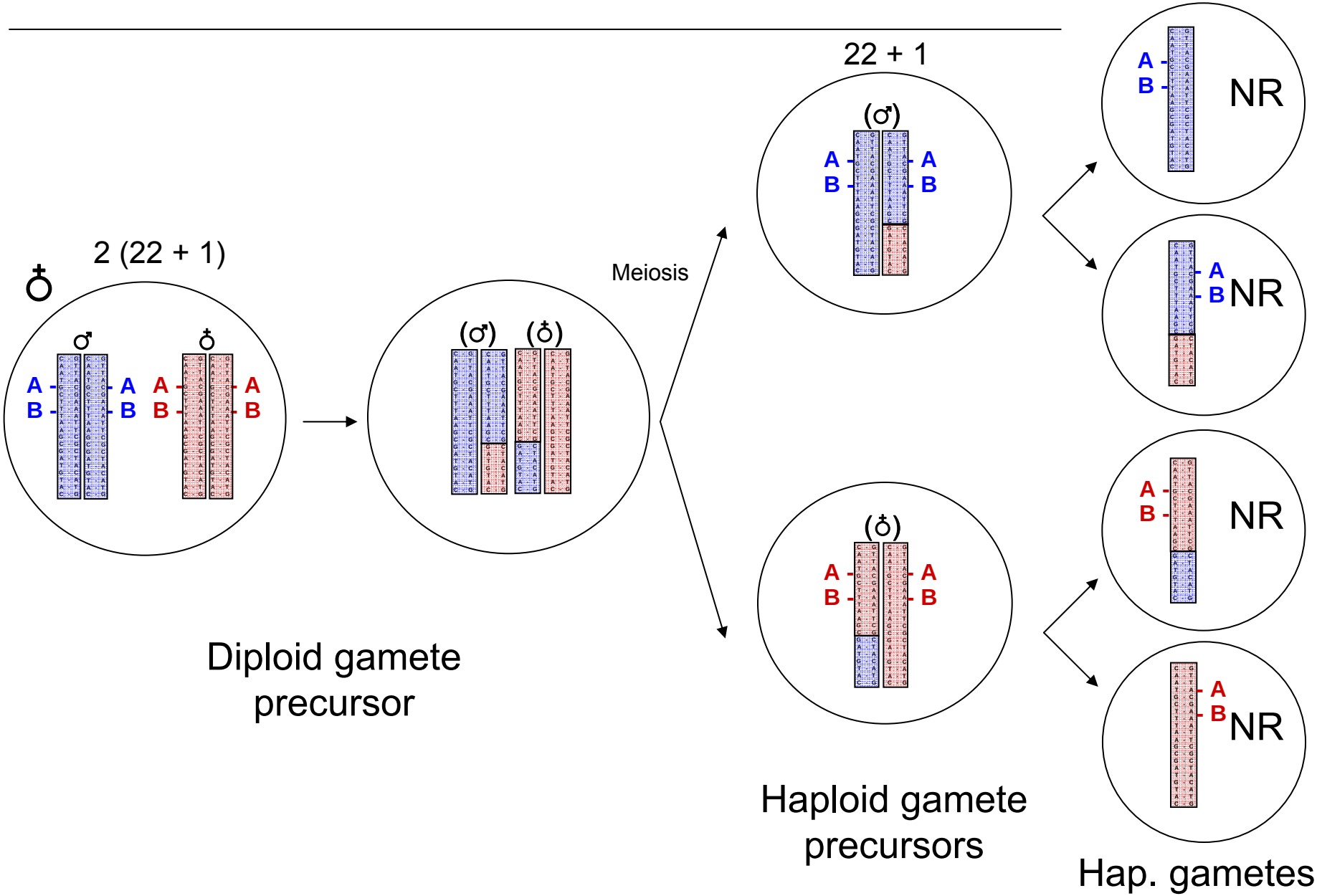
DNA organization



DNA recombination



DNA recombination between linked loci



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 - $0 \leq \theta < 0.5$

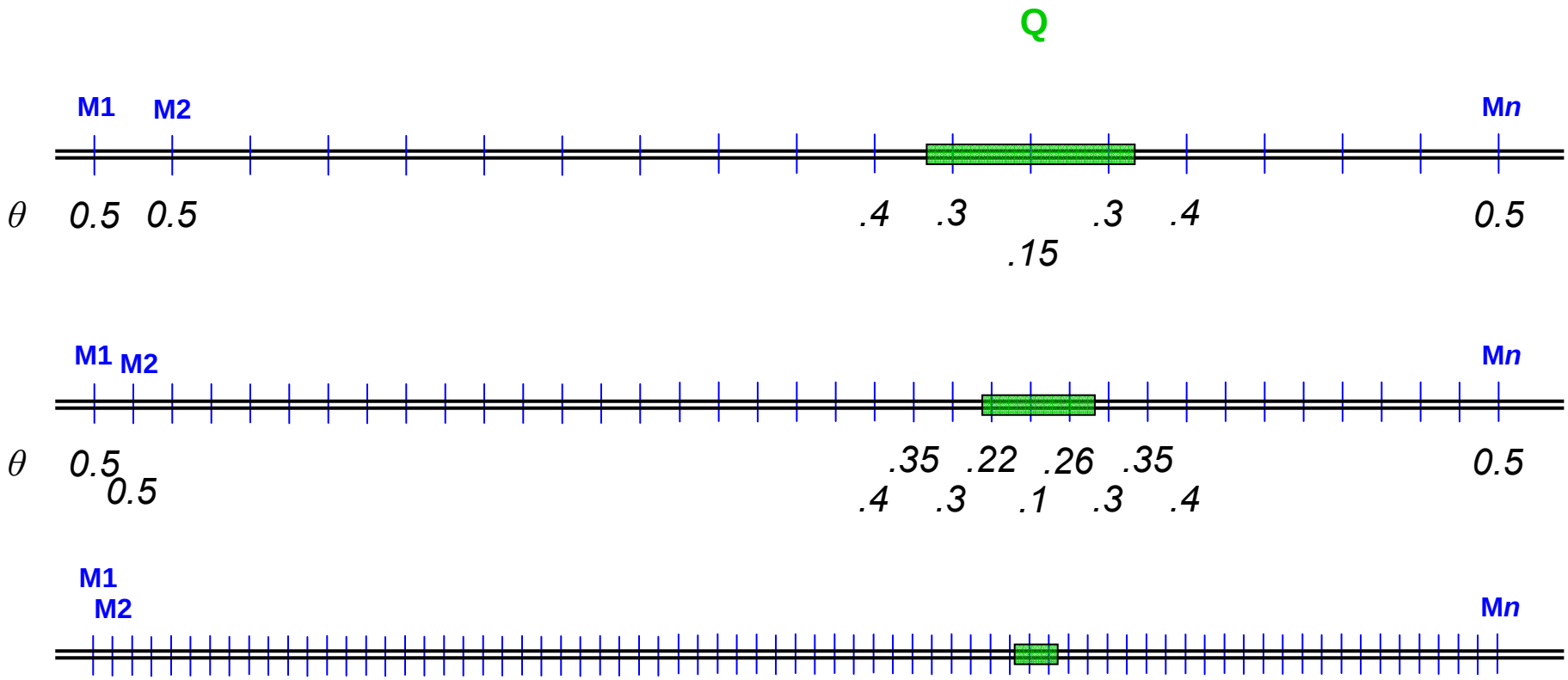
- Recombination fraction (θ) can be converted to genetic distance (cM)

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

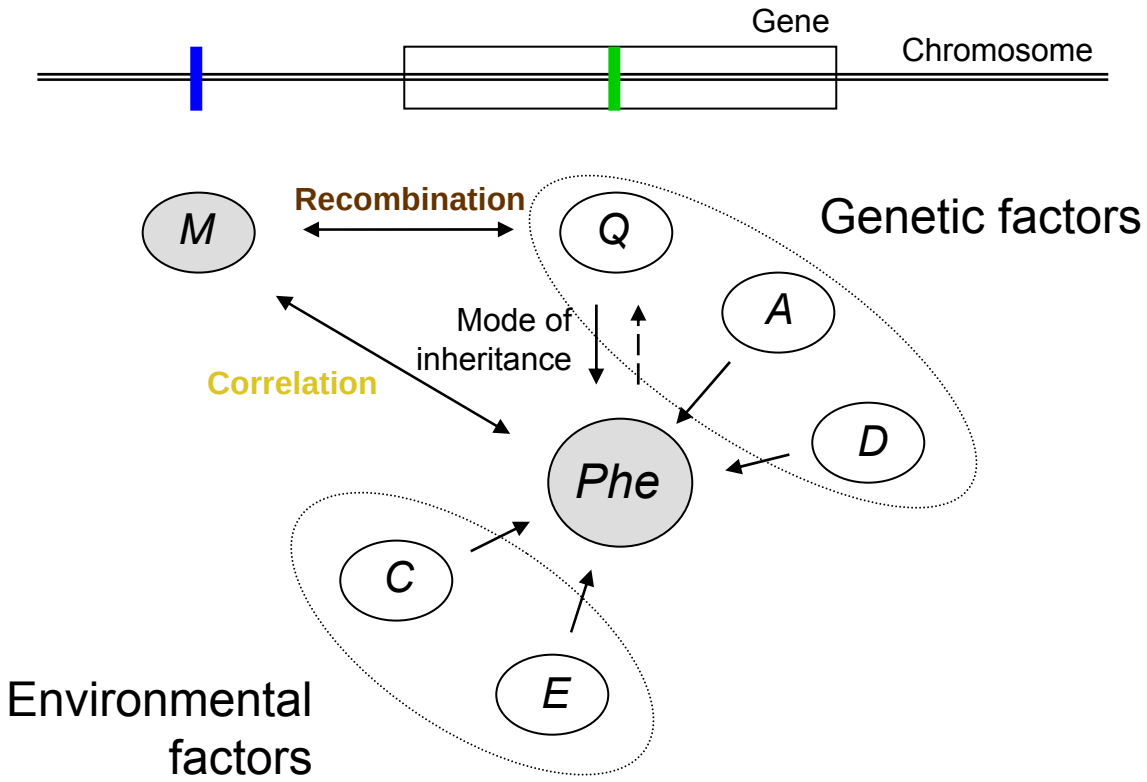
  Kosambi: $cM = 100 \cdot [0.25 \cdot \ln((1 + 2 \cdot \theta)/(1 - 2 \cdot \theta))]$ eg. θ=0.17, cM=17.7

3. Principles of Linkage Analysis

Linkage Analysis requires genetic markers



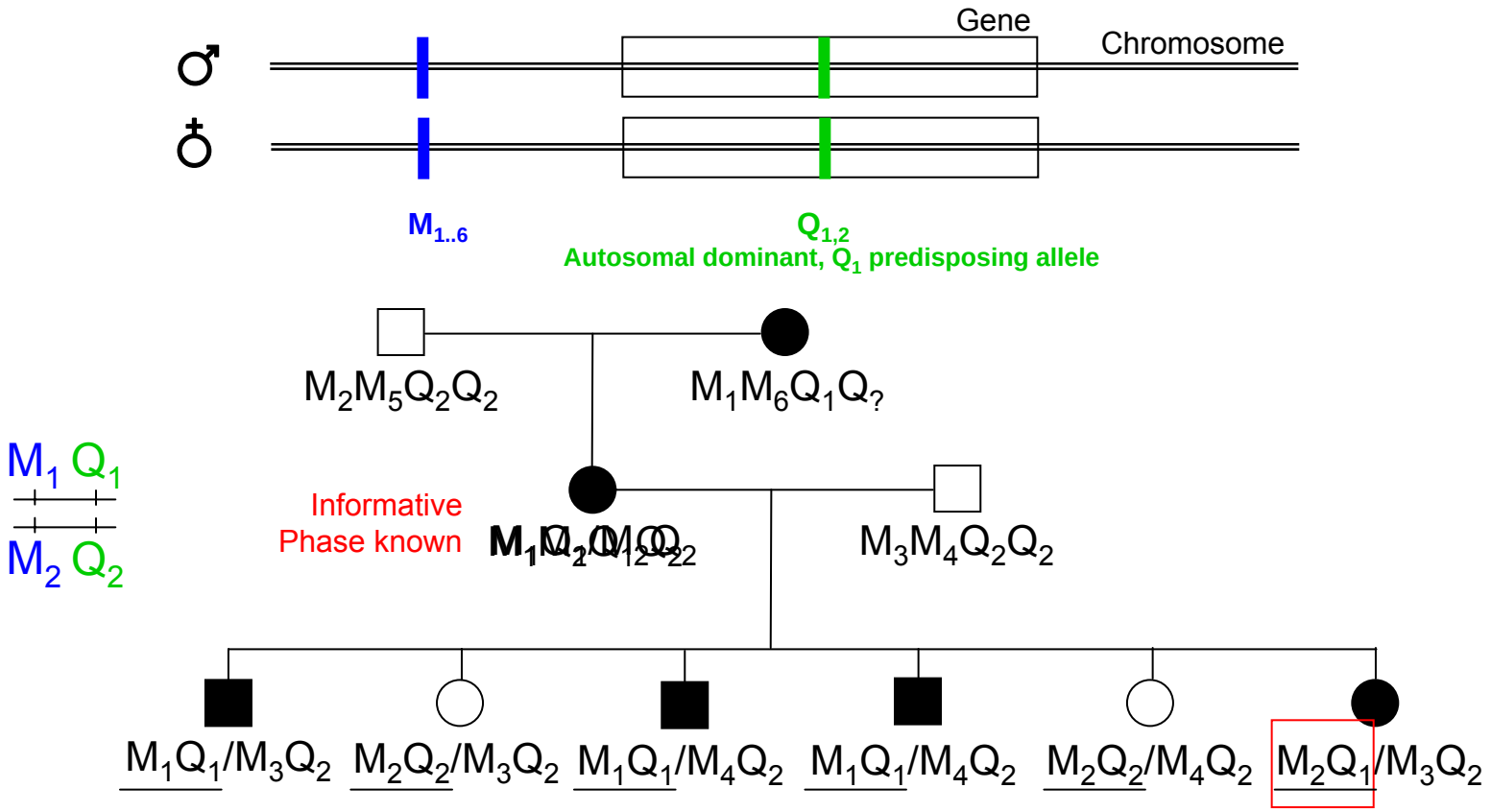
Linkage Analysis: Parametric vs. Nonparametric



Adapted from Weiss & Terwilliger 2000

4. Parametric Linkage Analysis

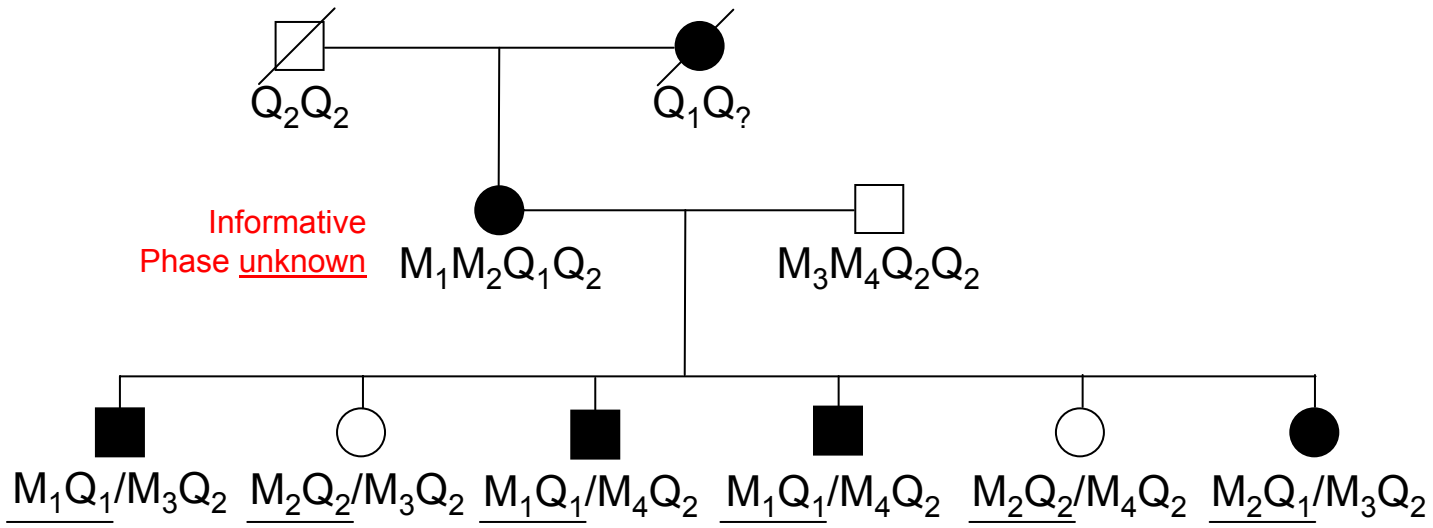
Linkage with informative phase known meiosis



NR: M_1Q_1
NR: M_2Q_2
R: M_1Q_2
R: M_2Q_1

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

Linkage with informative phase unknown meiosis



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

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

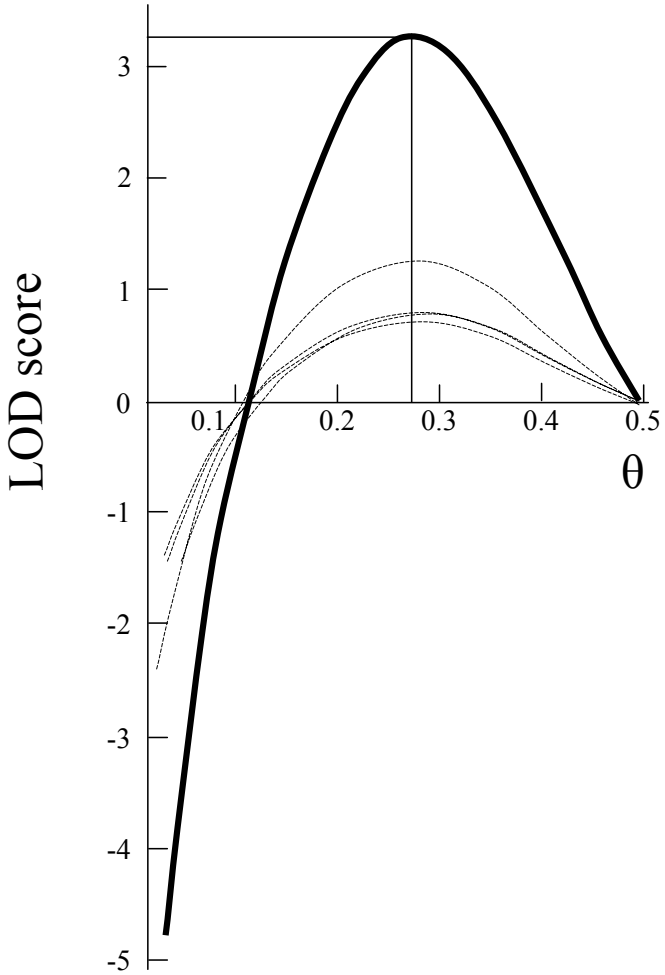
$$OD = \frac{L(X \mid \theta)}{L(X \mid \theta = 0.5)} \qquad LOD = \log_{10} \frac{L(X \mid \theta)}{L(X \mid \theta = 0.5)}$$

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

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

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

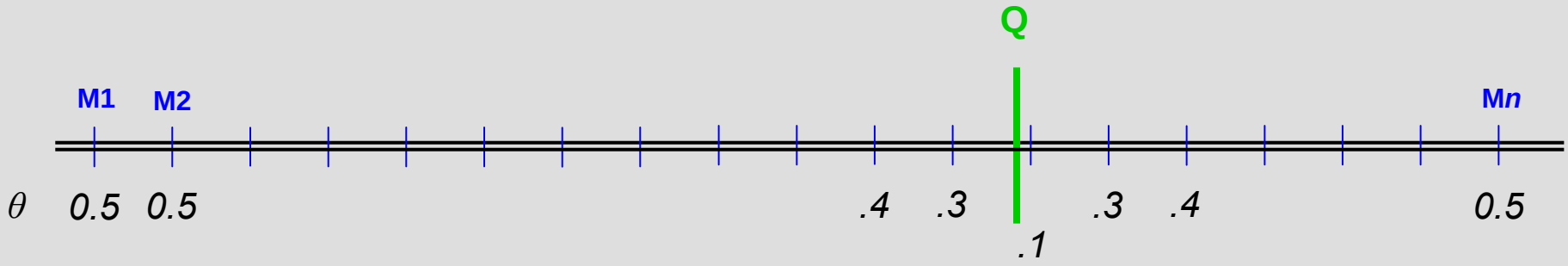
$$LOD = \sum_{i=1}^{n} \log_{10}\left(\frac{L(X_i \mid \theta)}{L(X_i \mid \theta = 0.5)} \right) = \sum_{i=1}^{n} LOD_i$$



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

  eg. LOD=3 for θ=0.28

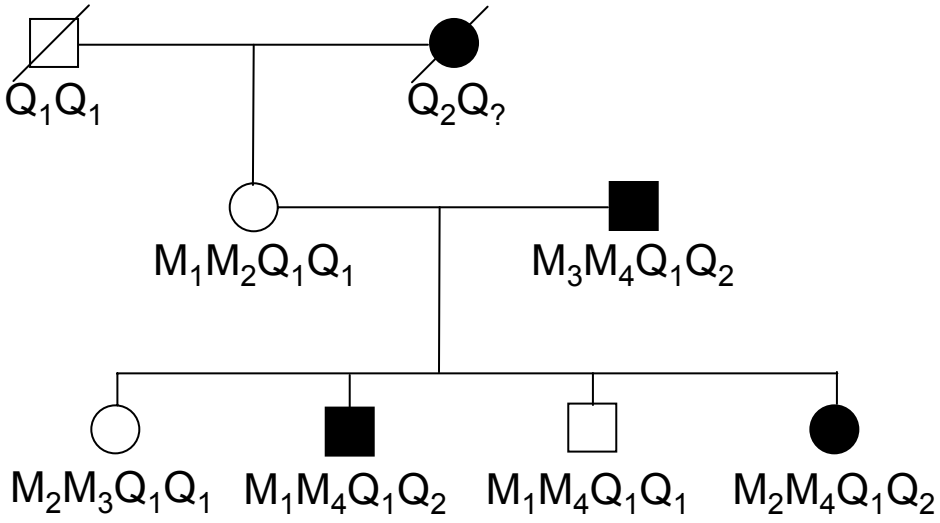
Parametric Linkage Analysis - summary



- 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

Practical



1. Identify informative individual(s)
2. Reconstruct possible phase(s)
3. Classify gametes as R or NR
4. Count R and NR gametes
5. Express $L(X \mid \theta)$ $L(X \mid \theta = 0.5)$
6. Express LOD score $f(\theta)$

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

Practical II

- Talk example

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

- Practical example

$$LOD = \log_{10} \frac{\frac{1}{2} \cdot \left[\theta^{1} \cdot (1-\theta)^{3}\right] + \frac{1}{2} \cdot \left[\theta^{3} \cdot (1-\theta)^{1}\right]}{(0.5)^{4}}$$

Graph each…

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.

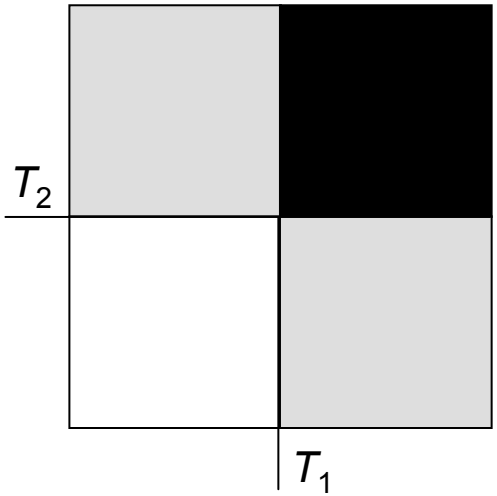
  Interest: correlation between phenotypic similarity and marker genotypic similarity

  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 $\begin{Bmatrix} Var(X_1) & Cov(X_1X_2) \\ Cov(X_1X_2) & Var(X_2) \end{Bmatrix}$

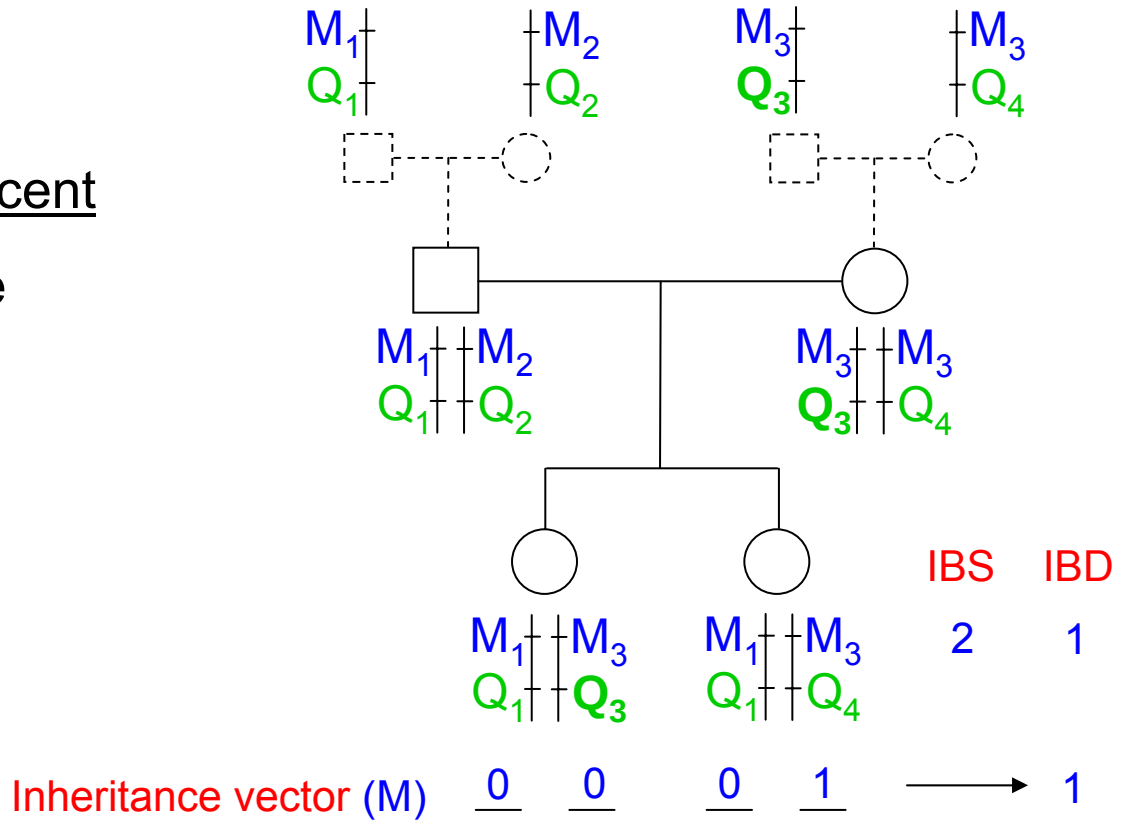
- Affection concordance



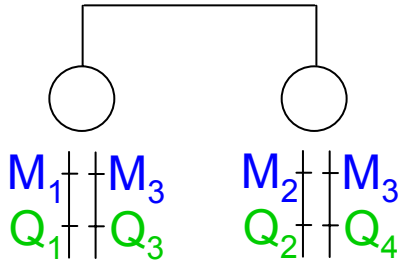
Genotypic similarity between relatives

- 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

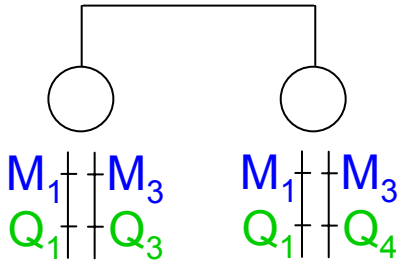
- IBD — Alleles shared Identical By Descent are a copy of the same ancestor allele

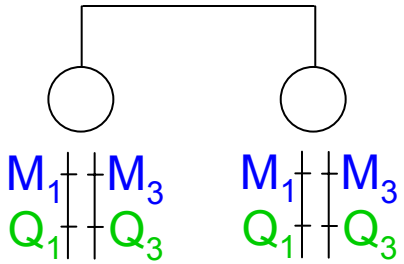


Genotypic similarity between relatives - π

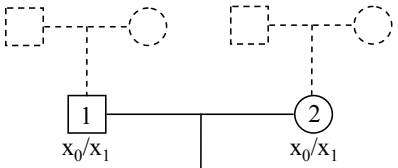


Inheritance vector (M)	Number of alleles IBD	Proportion of alleles IBD - π
0 0 1 1	0	0
0 0 0 1	1	0.5
0 0 0 0	2	1





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



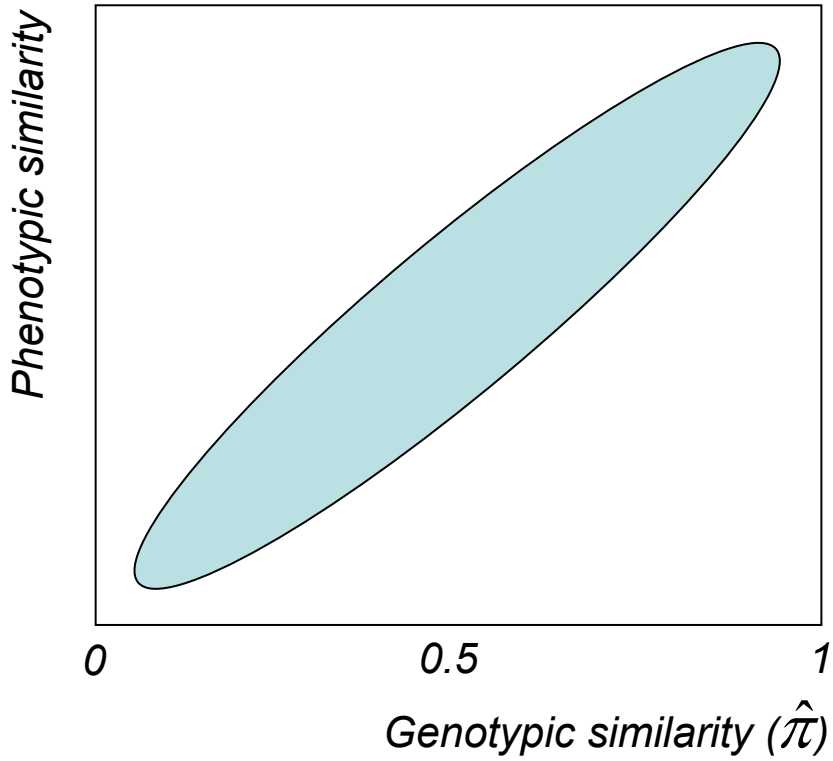
2^{2n}

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)

$$\hat{\pi} = \pi_0 \cdot \frac{0}{2} + \pi_1 \cdot \frac{1}{2} + \pi_2 \cdot \frac{2}{2} = \frac{\pi_1}{2} + \pi_2$$

Statistics that incorporate both phenotypic and genotypic similarities



Haseman-Elston regression – Quantitative traits

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

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

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

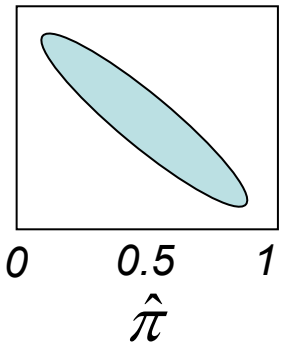
$$Var(X_1) = Var(X_2) = V_Q + V_A + V_C + V_E$$

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

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

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

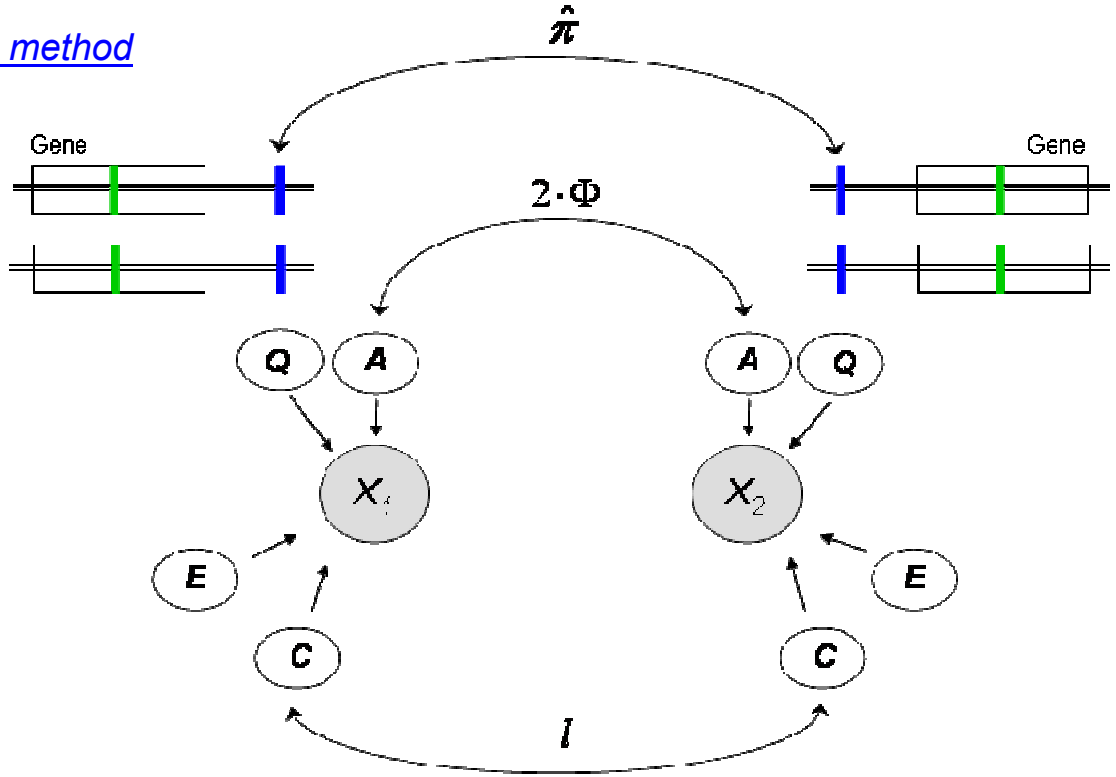
$(X_1 - X_2)^2$

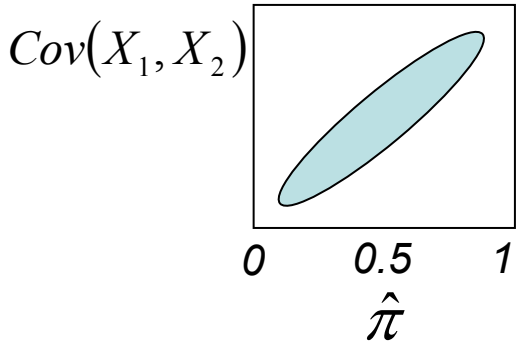


	X_1	X_2	$(X_1-X_2)^2$	$\hat{\pi}$
1	2.2	2.1	0.01	0.9
2	1.9	2.3	0.16	0.6
3	2.3	2.6	0.09	0.7
4	3.4	1.6	3.24	0.1
5	2.5	2.3	0.04	0.8
…				
1000	2.4	2.4	0	0.9

VC ML – Quantitative & Categorical traits

$\hat{\pi}$ *method*





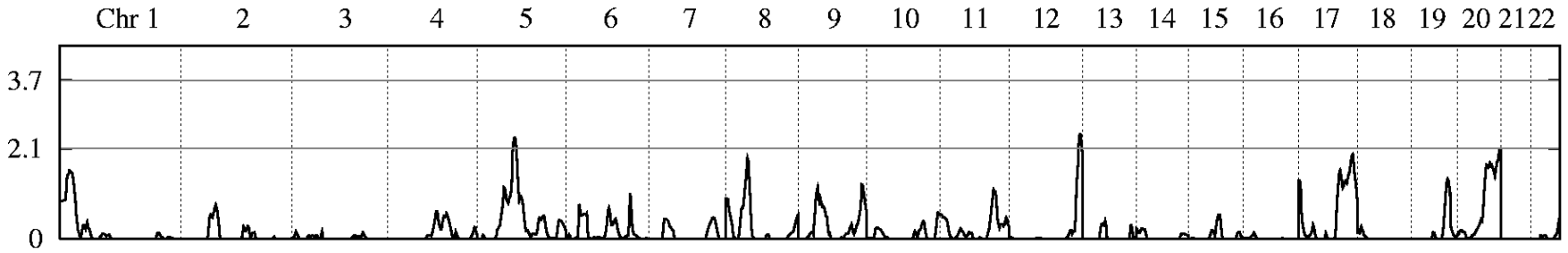
H_1: $Cov(X_1, X_2 \mid \hat{\pi}) = \hat{\pi}\cdot V_Q + 2\cdot\Phi\cdot V_A + I\cdot V_C$

H_0: $Cov(X_1, X_2 \mid \hat{\pi}) = 2\cdot\Phi\cdot V_A + I\cdot V_C$

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

e.g. LOD=3

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

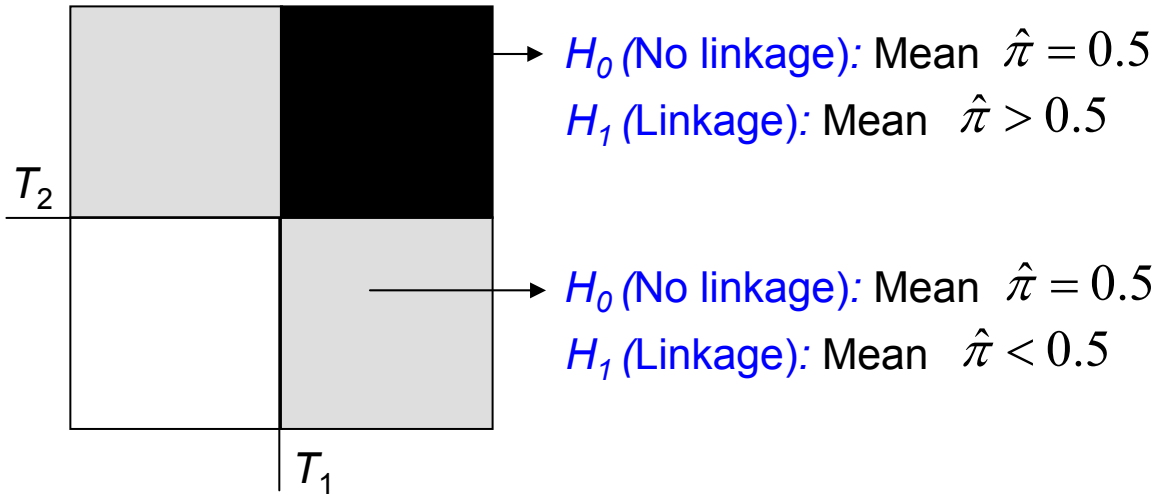


Individual LOD scores can be expressed as *P* values (Pointwise)

LOD	(x4.6) →	Chi-sq (n-df)	→	*P* value
2.1		9.67		0.0009

Statistics for selected samples

- *Mean IBD sharing statistics* *(Risch & Zhang 1995, 1996)*



Other Linkage statistics

Dependent variable: Phenotypes Independent variable: $\hat{\pi}$

- *Extensions to Haseman Elston* *(Wright 1997, Drigalenko 1998, Elston et al. 2000, Forrest 2001, Visscher & Hopper 2001, Xu et al. 2000, Sham & Purcell 2001)*
- *VC ML with mixture distribution* *(Eaves et al. 1996)*

Dependent variable: $\hat{\pi}$ Independent variable: Phenotypes

- *Pedwide-regression Analysis ("reverse HE")* *(Sham et al. 2002)*
- *Reverse VC ML* *(Sham et al. 2000)*

Statistics for affection traits

- *Based on IBD scoring functions eg.* S_{all} (Whittemore & Halpern 1994, Kong & Cox 1997)
- *Forrest & Feingold 2000 Mixed statistic*

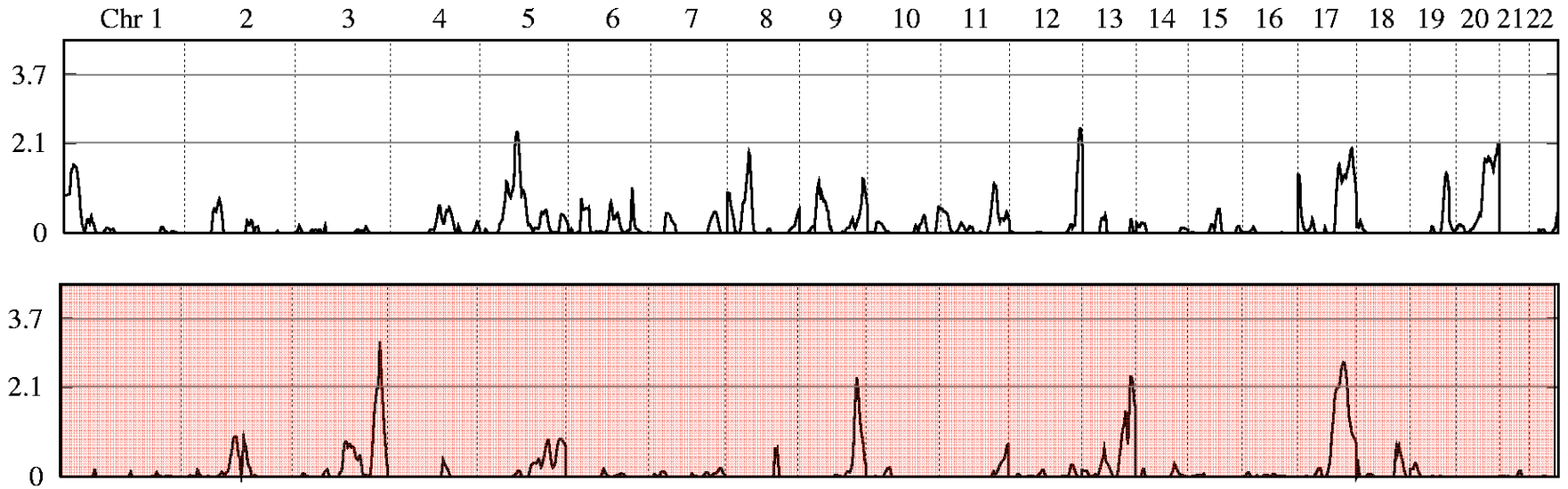
Nonparametric Linkage Analysis - summary

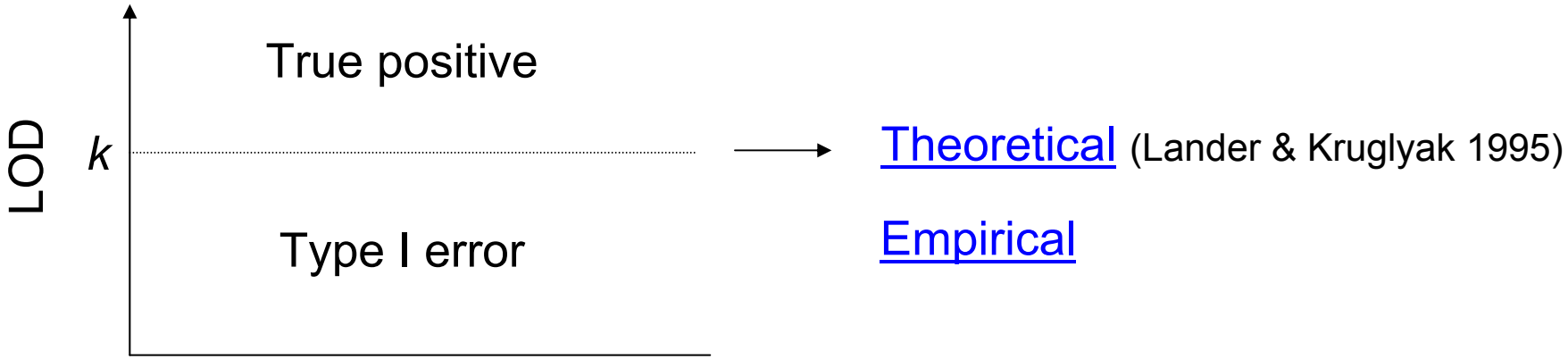
- No need to specify mode of inheritance
- Models phenotypic and genotypic similarity of relatives
- Expression of phenotypic similarity, calculation of IBD
- HE and VC are the most popular statistics used for linkage of quantitative traits
- Other statistics available, specially for affection traits

Type I error?

Power?

Type I error





Theoretical (Lander & Kruglyak 1995)

Empirical

Theoretical genome-wide thresholds

- Genome-wide threshold for significant linkage

  LOD score that occurs by chance alone on average once per 20 scans

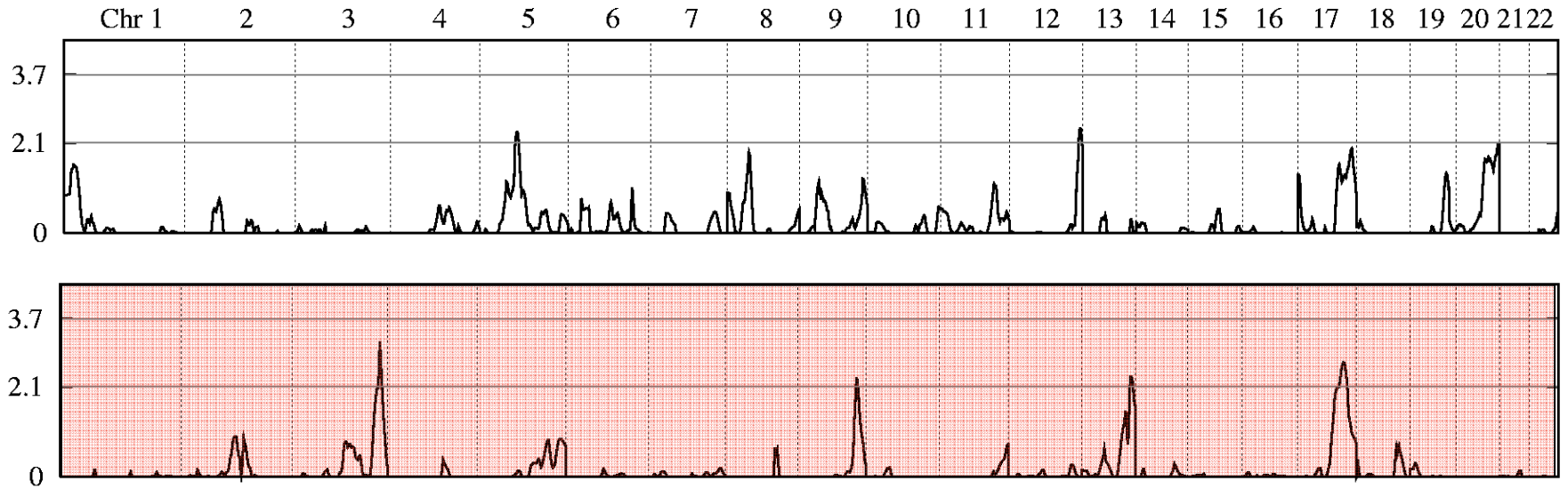
  LOD = 3.6, Chi-sq = 16.7, Pointwise *P* = 0.000022

- Genome-wide threshold for suggestive linkage

  LOD score that occurs by chance alone on average once per scan

  LOD = 2.2, Chi-sq = 10.1, Pointwise *P* = 0.00074

Empirical genome-wide thresholds



- Genome-wide threshold for significant linkage

  LOD score that occurs by chance alone on average once per 20 scans

- Genome-wide threshold for suggestive linkage

  LOD score that occurs by chance alone on average once per scan