Linkage analysis: an overview

26.10.2009

Yurii Aulchenko,
Erasmus MC Rotterdam

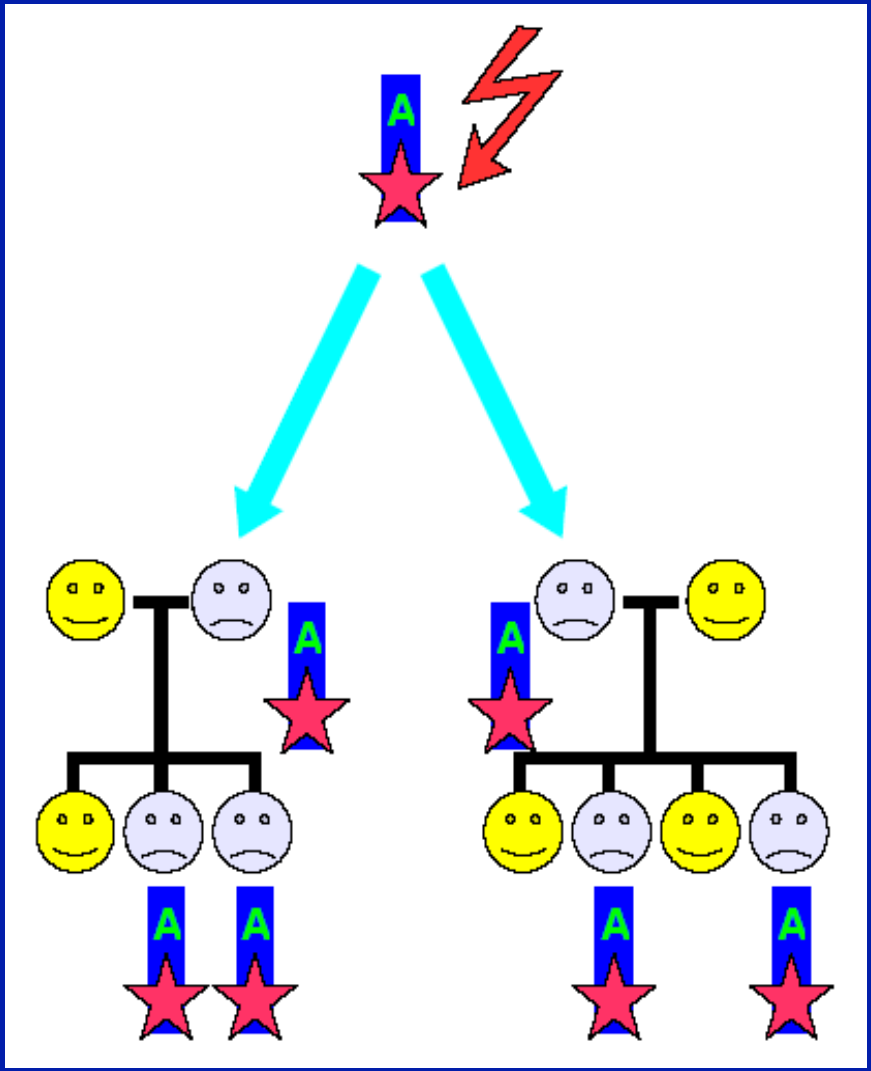
Outline

- Difference between linkage and association
- LOD score analysis
- Algorithms to compute LOD score

Linkage and association

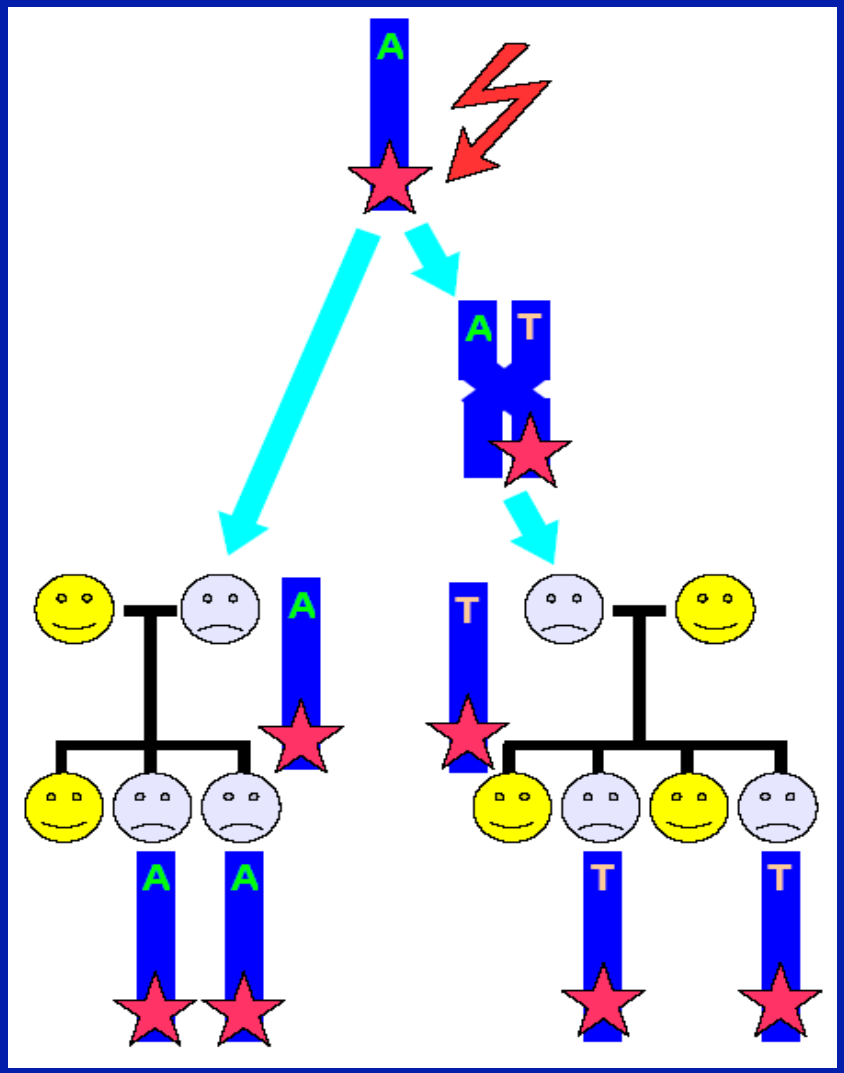
- Sampling unit
  - Sib-pair
  - Family of arbitrary structure
  - A random person from population
- Association assumes that a particular allele is associated with the trait **in all sampling units**
- Linkage assumes that a particular allele is associated with the trait **within the sampling unit, but the allele may be different across sampling units**

Dense map, tight LD



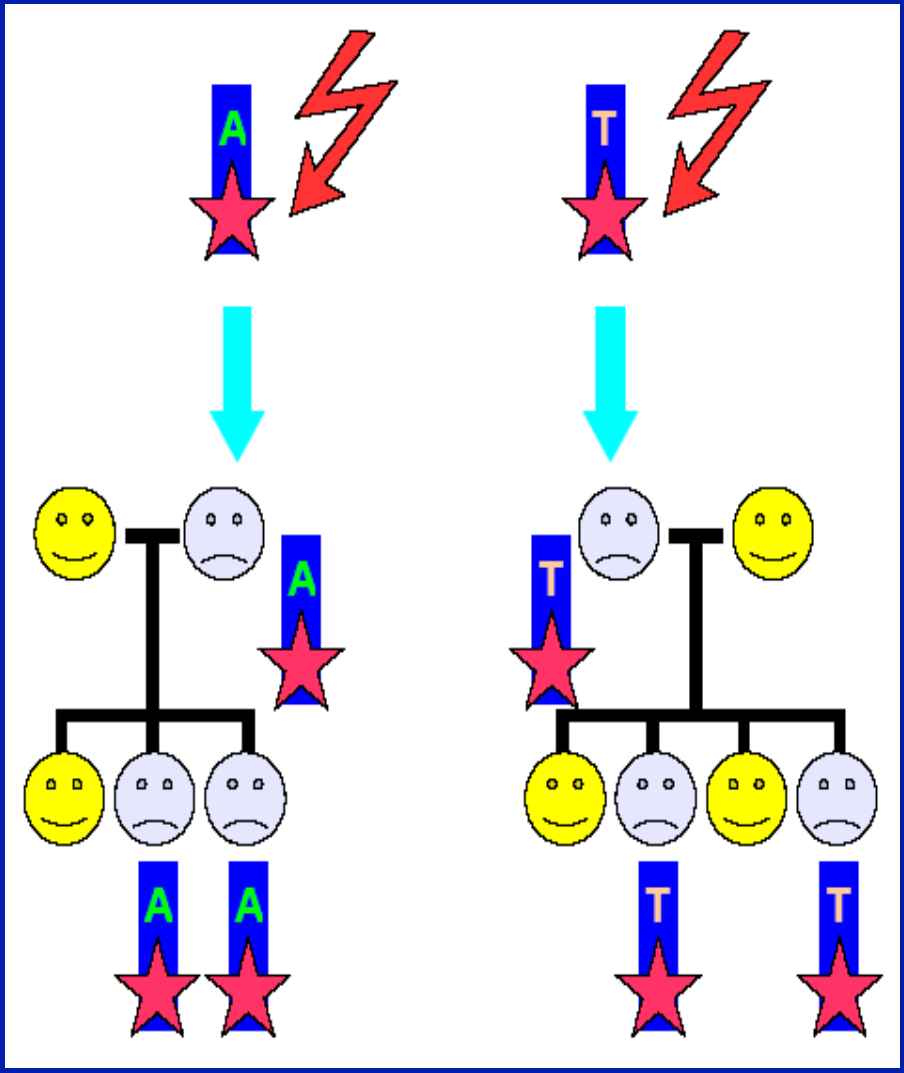
- Classical situation for association mapping
- Association mapping successful
- Linkage mapping possible (but potential problem with LD)

Sparse map, recombination



- Classical situation for linkage mapping
- Association mapping:
  * If some LD retained, less powerful cf. scenario 1, but still possible
  * No LD retained – no power
- Linkage mapping is powerful

Allelic heterogeneity



- Association mapping:

  * If some mutation is common, some power

  * If mutations are “private”, no power

- Linkage mapping is powerful (but potential problem with LD)

Linkage vs. association

"Fair" comparison is difficult because set-ups of linkage and association are different

Association: **a common risk variant**

- Common disease
- Collection of distantly related people (case/control sample)
- Dense marker set (array with 100s of 1000s of markers)

Linkage: **rare, possibly heterogeneous mutation(s)**

- Rare familial (form of) disease
- Small collection or even single extended family
- Sparse or thinned marker set (100s to few 1000s of markers)

Outline

- Difference between linkage and association
- **LOD score analysis**
- Algorithms to compute LOD score

LOD score

Data Y: pedigree, phenotypes, marker

LOD score at location *d* is defined as

$$LOD(d) = \log_{10} \frac{L(Y \mid d, \omega)}{L(Y \mid d = \infty, \omega)} = \log_{10} \frac{P_1}{P_0}$$

ω – the vector of genetic parameters (disease allele frequency, penetrances, …)

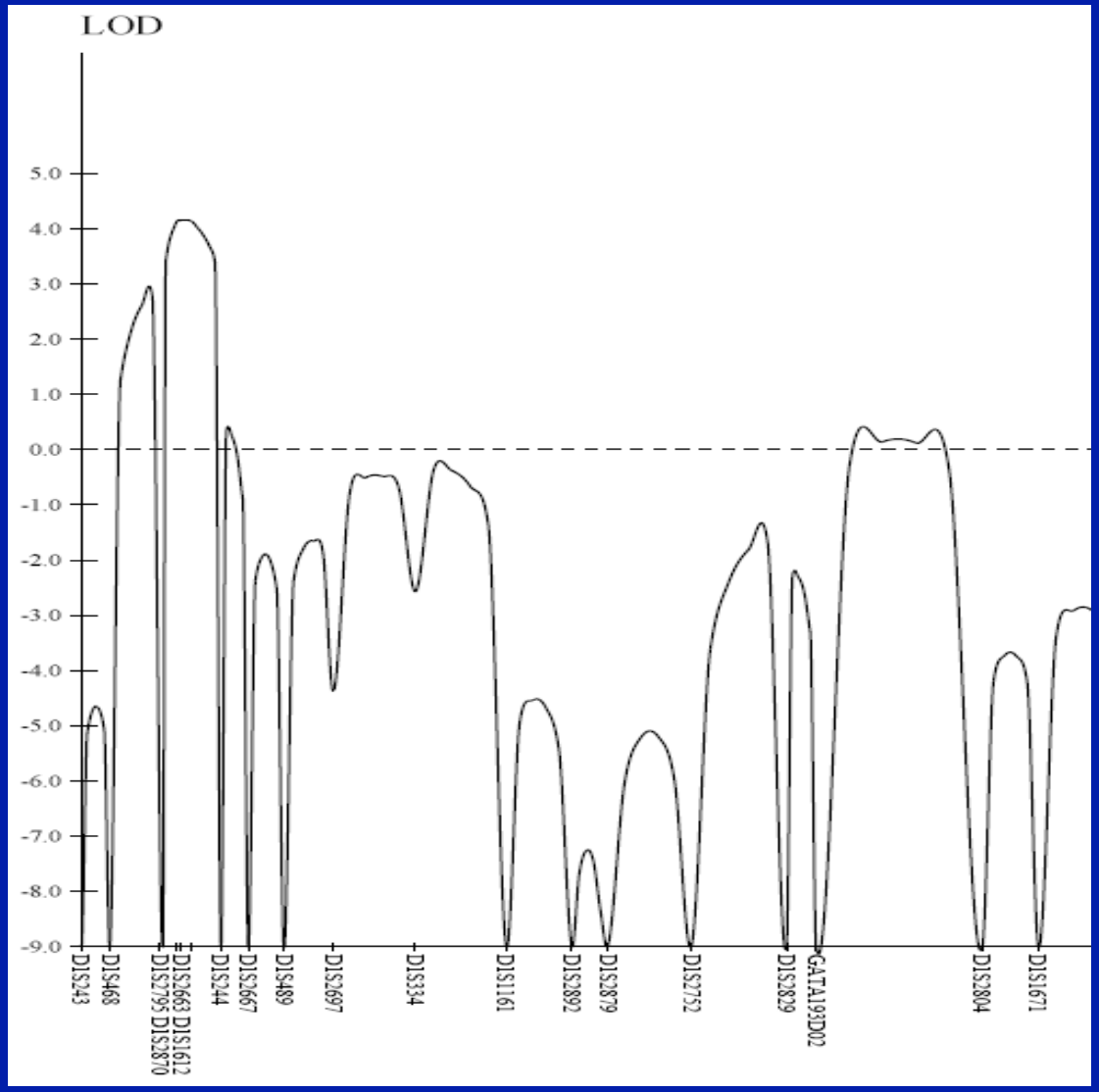
- Estimated in segregation analysis
- … or defined relatively arbitrarily

Results of two-point analysis

Single marker (two-point analysis): table of LODs across recombination fraction (θ)

	Theta							
Marker	0	0,01	0,05	0,1	0,2	0,3	0,4	0,5
D1S243	-inf	-4,96	-2,59	-1,56	-0,65	-0,25	-0,07	0
D1S468	-inf	-4,37	-1,94	-0,99	-0,28	-0,07	-0,03	0
D1S2795	-inf	1,84	2,21	2,09	1,54	0,89	0,32	0

Results of multipoint analysis



Significance of LOD score

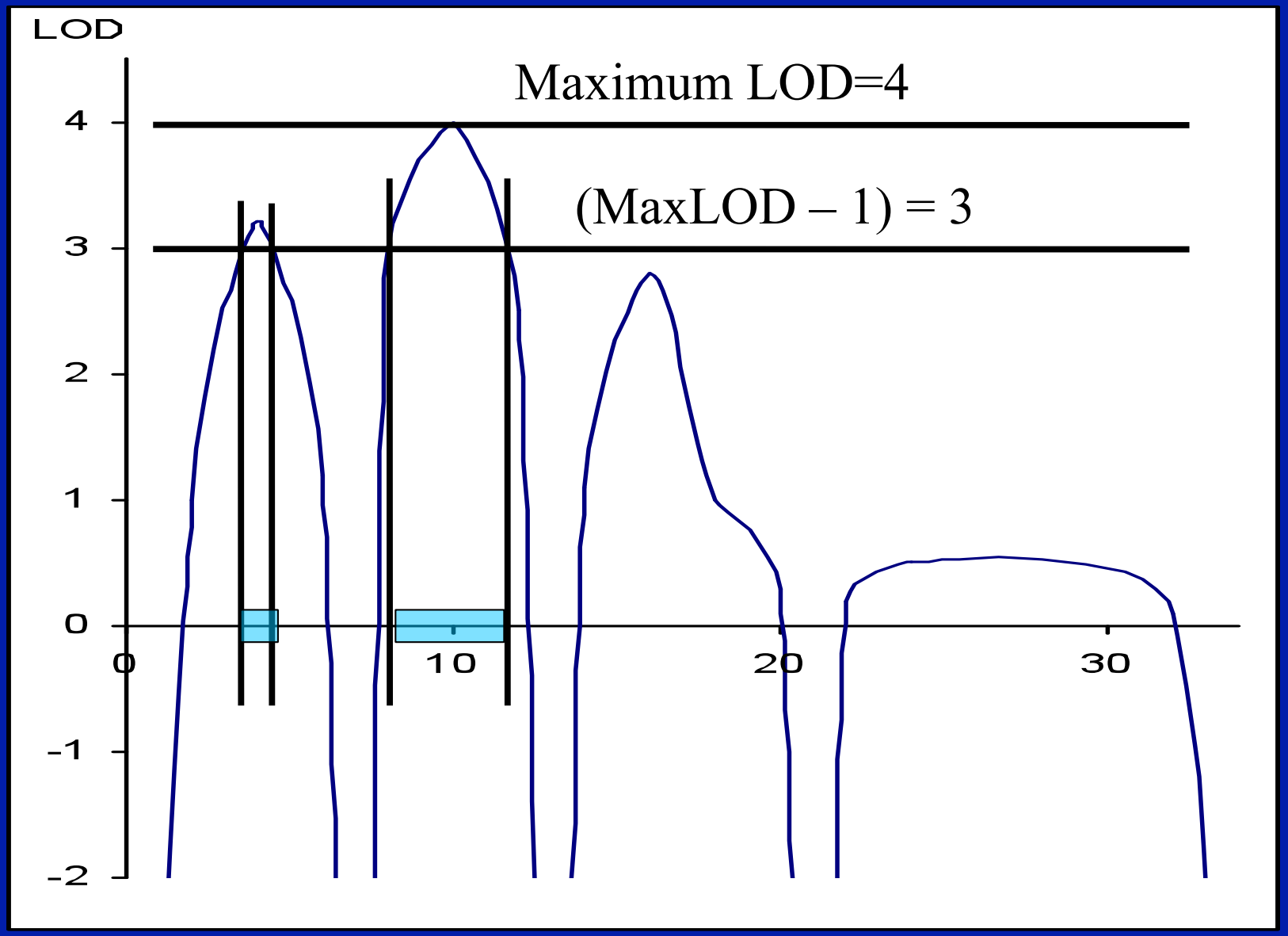
LOD ≥ 3 is considered significant in single-marker analysis

For a genome scan (Lander & Kruglyak 1995)

- LOD > 3.3 is considered genome-wide significant (P<0.05)
- LOD > 1.9 is considered suggestive (expected to appear once per genome-scan)

- Threshold depends on design and marker density!
- Best derived using empirical techniques

1-LOD support interval (~ 90% CI)



Outline

- Difference between linkage and association
- LOD score analysis
- **Algorithms to compute LOD score**

Likelihood of pedigree data

$$L(Y) = \Sigma_{\text{all } G} P(Y|G)P(G)$$

where G is a matrix of underlying (unobserved) genotypes of pedigree members

Number of possible genotypic combinations:

Genotypes possible for founders

$=((\text{No. possible haplotypes})^2)^{(\text{No. of founders})}$

by

Number of inheritance patterns

$= 2^{(\text{No. meioses})}$

Computation time

Sibship of 5, trait locus + …

- 3 SNPs => 2 sec.
- 7 SNPs => 4.5 hours
- 10 SNPs => 5 years

Trait locus + 7 SNPs in a pedigree of …

- 6 sibs => 18 hours
- 8 sibs => 12 days
- 10 sibs => ½ year

Elston-Stewart algorithm

For parts of pedigree, compute probability conditional on all possible genotypes of members who connect this part to the rest

Computation time grows

- Linear with no. people
- Exponential with no. markers
- Exponential with no. of loops

Limit: 2-4 MS (or 5-20 SNPs) and 2-3 loops

Lander-Green algorithm

For particular marker (phenotype), compute probability for all pedigree members conditional on flanking genotypes

Computation time grows

- Exponential with pedigree

  bit-size = 2 x (no. non-founders) – (no. founders)
- Linear with no. markers

Limit: bit-size 18 to 36

Markov Chain Monte-Carlo

A technique to compute approximate probabilities based on sampling from the model space

Computation time grows with

- Larger proportion of missing data
- Loops
- Denser marker maps

Limit: few hundreds of people and few dozens of loops (takes days to finish) using linkage panels

Results may depend on where you start to explore the space

Software

Elston-Stewart: Linkage, FastLINK, Vitesse, Superlink

- Binary trait linkage analysis

Lander-Green: GH, Merlin, Allegro (LG)

- Binary trait linkage analysis
- IBD estimation

MCMC: SimWalk2

- Binary trait linkage analysis
- IBD estimation

MCMC: Loki

- Bayesian quantitative trait linkage analysis
- IBD estimation

Problems with dense marker sets

Standard linkage programs assume markers are in linkage equilibrium. Type 1 error increased if

- Markers are in LD
- There are founding pedigree members with missing genotypic data

Possible practical solutions

- Model LD – implemented in Merlin; computationally complex, thus possible only for small pedigrees
- "Thinning" marker map (select markers not in LD) –implemented in MASEL