(3) QTL and GWAS methods

By the end of this lecture you should be able to explain:

- Some of the principles underlying the statistical analysis of QTLs
- Under what conditions particular methods are suitable
- The core differences between QTL analysis and GWAS

Link: Refer to Hugo's lecture on QTL notes

Linkage LOD addendum to lecture 2

log odds or LOD score

•Compares the likelihoods of a data set exhibiting *r* crossovers out of a potential *N* between a pair of markers under the hypothesis of linkage (i.e. θ <0.5, where θ represents the recombination fraction)
versus the same observation under the hypothesis of independent segregation (i.e. θ =0.05).

- So **LOD scores** are calculated for recombination fraction θ values to determine if there is significant evidence of linkage

- For a given value of θ, the LOD score is

$$log_{10} \frac{P(\text{observed data assuming recombination fraction is } \theta)}{P(\text{observed data assuming recombination fraction is } .5)}$$

- We find the value of θ that gives the maximum LOD score

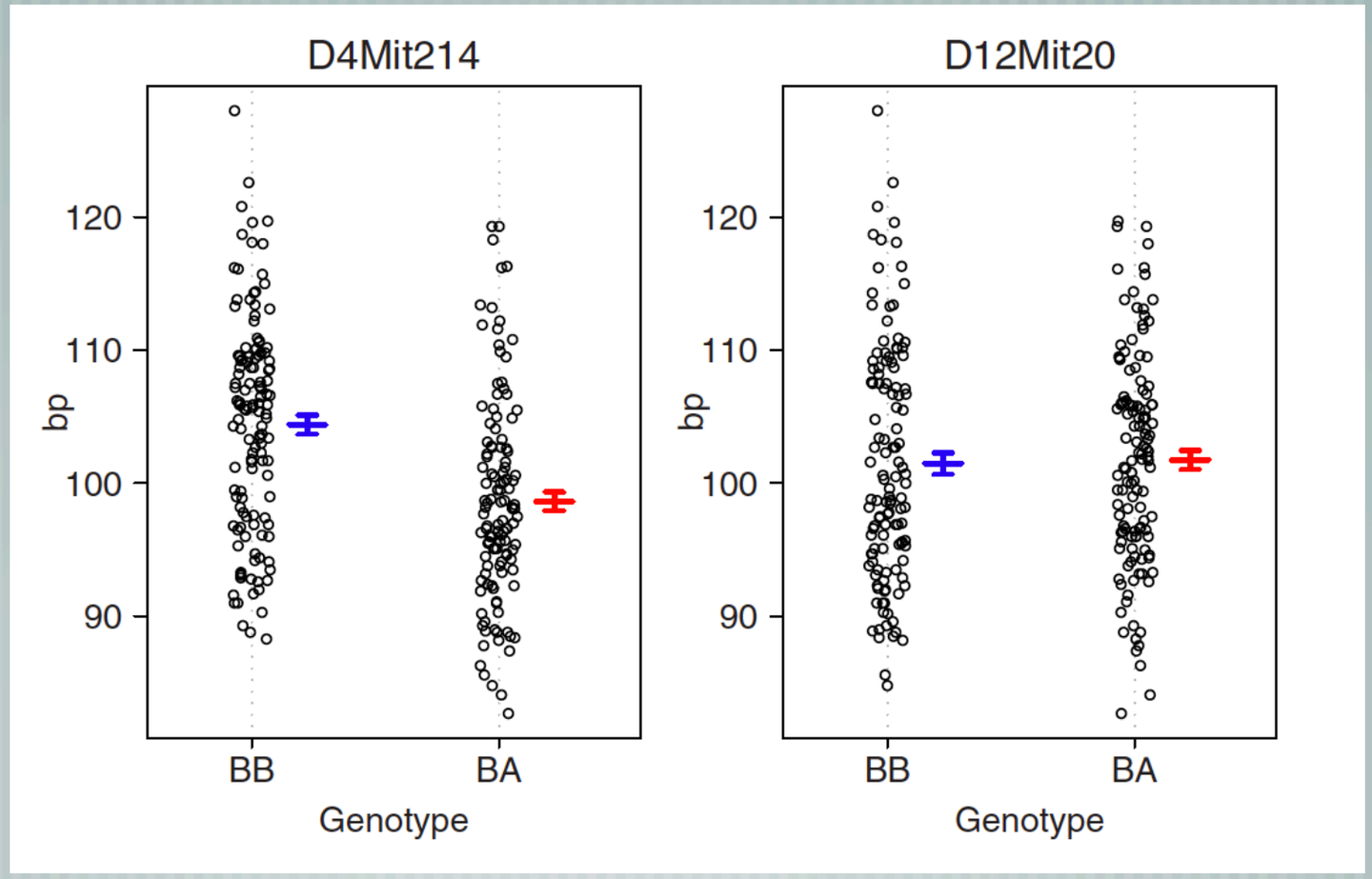
LOD scores >3 give evidence of linkage and the null hypothesis is rejected

- This value coresponds to a likelihood of observing the dataset, given the two markers are unlinked, of <1/1000

QTL analysis

Detecting an association between phenotype and the genotype of markers

Markers
partition the population into different genotypic groups based on the presence or absence of a particular marker locus



Typically: null hypothesis tested is that the mean of the trait value is independent of the genotype at the marker.

Null rejected when the test statistic is larger than a critical value, and QTL is declared linked to the marker

t-test, ANOVA and simple linear regression are equivalent when their hypotheses are testing for differences in the phenotypic means

	SNP1	TRAIT 1	TRAIT 2
IND1	A	60	19
IND10	T	7	22
IND11	A	92	18
IND12	T	9	21
IND13	T	4	24
IND14	A	77	20
IND15	A	95	20
IND16	T	7	19
IND17	T	8	22
IND18	A	84	23
IND19	A	79	21
IND2	A	90	19
IND3	T	12	20
IND4	A	100	22
IND5	T	5	19
IND6	T	6	24
IND7	T	10	25
IND8	A	80	18
IND9	A	110	21

	SNP1	TRAIT 1	TRAIT 2
IND1	A	60	19
IND2	A	90	19
IND4	A	100	22
IND8	A	80	18
IND9	A	110	21
IND11	A	92	18
IND14	A	77	20
IND15	A	95	20
IND18	A	84	23
IND19	A	79	21
IND3	T	12	20
IND5	T	5	19
IND6	T	6	24
IND7	T	10	25
IND10	T	7	22
IND12	T	9	21
IND13	T	4	24
IND16	T	7	19
IND17	T	8	22

	MEAN	MEAN
A	86.70	20.10
T	7.56	21.78
	5.275E-12	0.0780

Marker linked to QTL = less chance of recombination between them therefore marker and QTL will be inherited together

When the QTL is not linked = independent segregation of M + Q = nsd between group means

Results from basic t-test

Trait 1 SNP1 linked to trait with A allele increasing the trait

Trait 2 SNP1 not linked, no significant differences between groups

Could also use Kruskal–Wallis♪

The statistical machinery for QTL mapping

Some analysis techniques (there are many, many variants):

Simple t-test: use to evaluate presence of a QTL through statistical differences between two marker genotypes (in a BC)

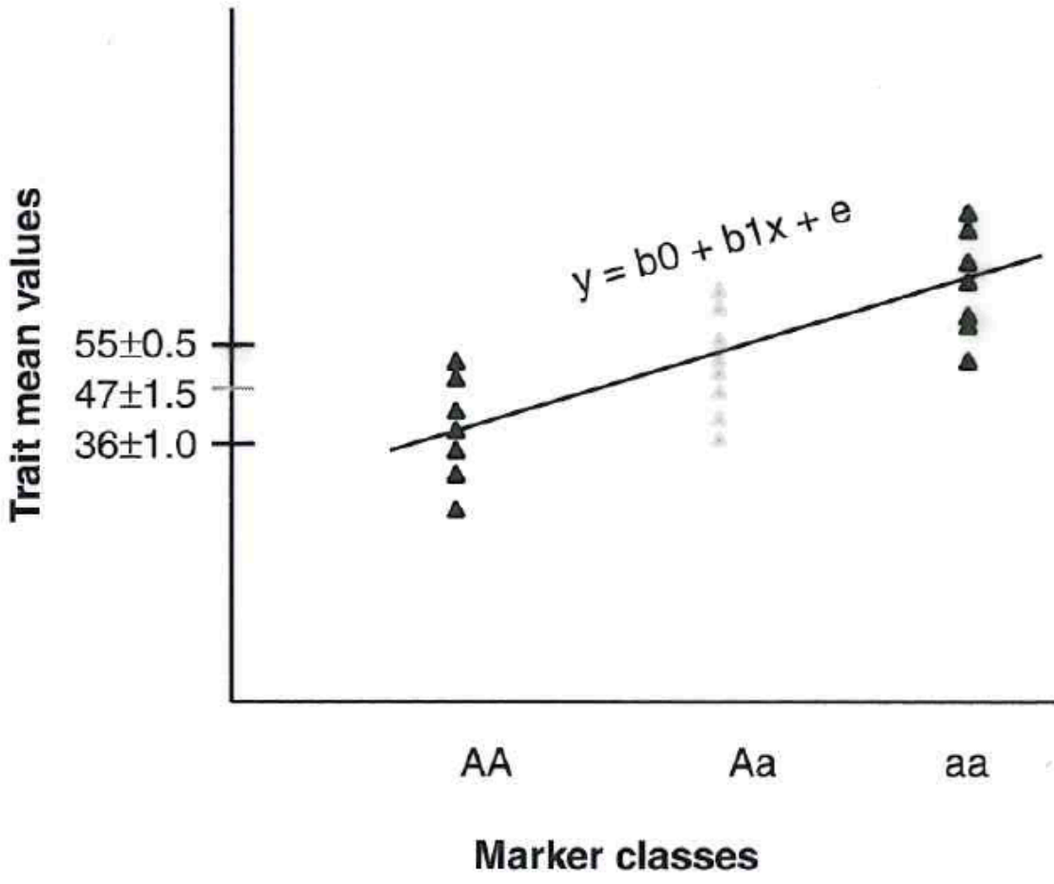
ANOVA (marker regression): detects marker differences when there are more than two marker genotypes.
Produces a ranking of genotypes, in order of phenotypic effect for the trait of interest, and tests for significant differences between each genotype (inter-cross = F)

Linear regression: most complex point analysis method, allowing different characteristics of the QTL to be investigated. Inc.
dominance effects, additive effects
genotype-environment interactions, epistasis

Simple in terms of data analysis, performed using common statistical software, gene order and complete linkage map not needed

Reduced power to detect QTL ~ QTL position can not be determined precisely
Many False positives

In ANOVA (sometimes called marker regression) the progenies are divided into two groups based on the genotype e.g. Fig 6.1 we see individuals with aa have sig higher phenotype score than those with Aa and AA - indicating that the marker is linked to a QTL.



$y = b0 + b1x + e$,
where y is the phenotypic value of a line, $b0$ is the population mean, $b1$ is the additive effect of the locus on the trait, and e is a residual error term. x is directly related to the genotypic code at the locus being tested for the line considered, it is -1 (for female parent) or 1 (for donor or male parent).

Fig. 6.1 Principle of single-marker analysis

If the phenotype distributions of the phenotype classes are approximately the same - marker not linked to QTL
But: the greater the distance between marker and QTL - less likely to be detected =recombination may occur between them | what about missing data?

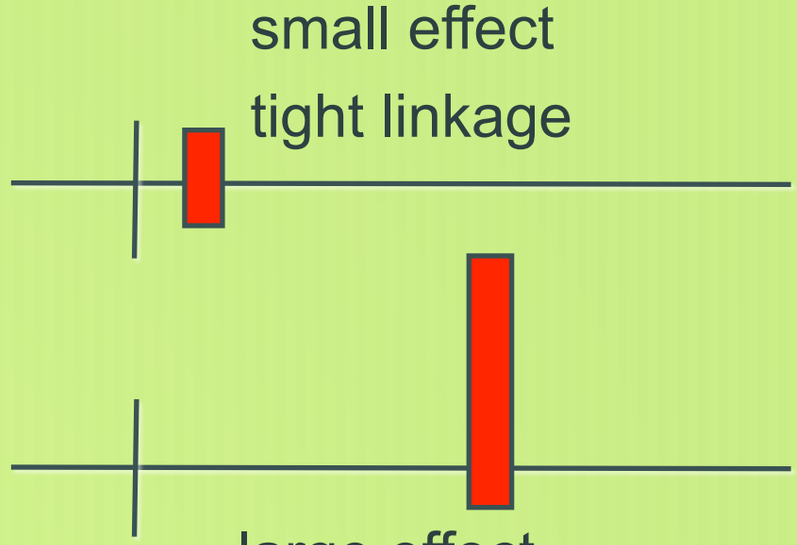
Which? guide to QTL mapping: Single marker rating

- Simplicity
- Allow detection of a QTL, but....
- poor estimates of QTL location and QTL effect due to incomplete linkage to markerMust discard individuals whose genotypes are missing at the markerWhen markers are sparse, the QTL may be quite far from all markers
- can not inspect positions between markers ...power for QTL detection will decrease

- A small MM-mm difference:

small effect

tight linkage



large effect

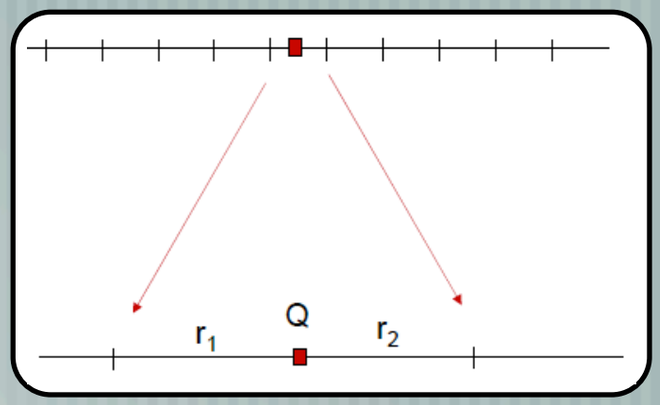
loose linkage

Interval mapping: Can use probability estimates for the genotypes in intervals between markers; most popular method

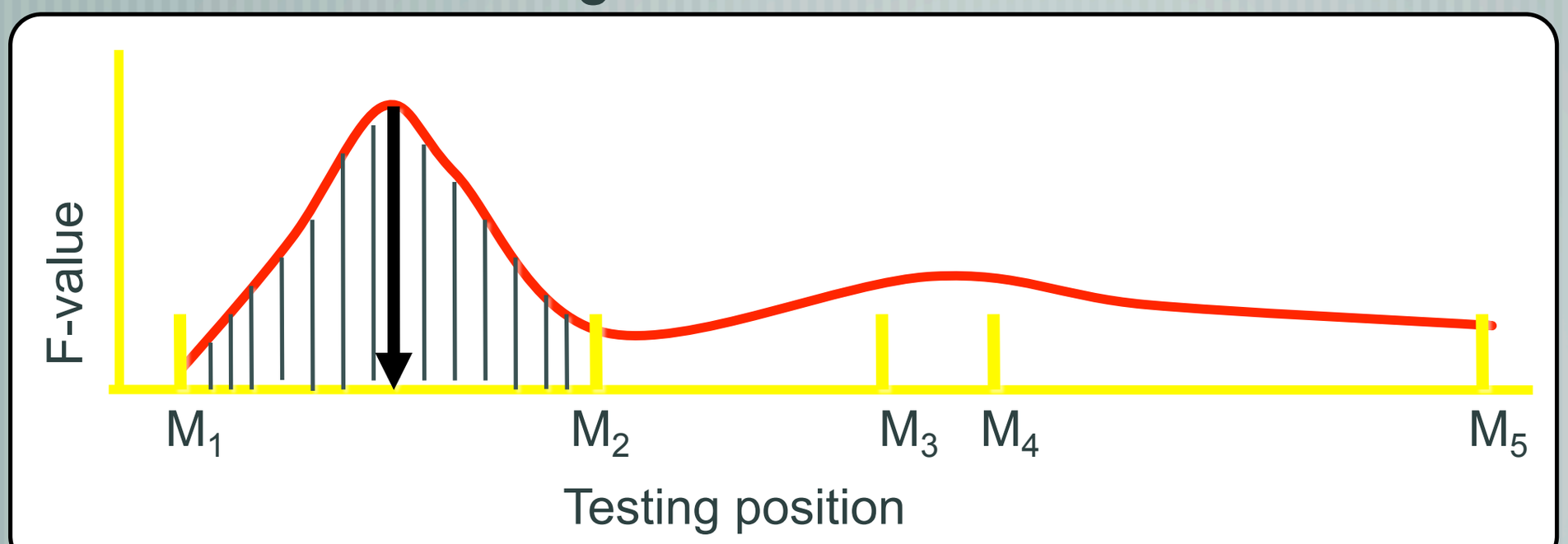
SIM: simple interval mapping
Uses map: analyses intervals between adjacent linked markers

Linked markers compensates for recombination between markers and QTL - more powerful than single point analysis

Model tests for the presence of 1 QTL at steps within the interval in a one dimensional linear scan across the genome, testing the same null hypothesis each time



Single QTL model

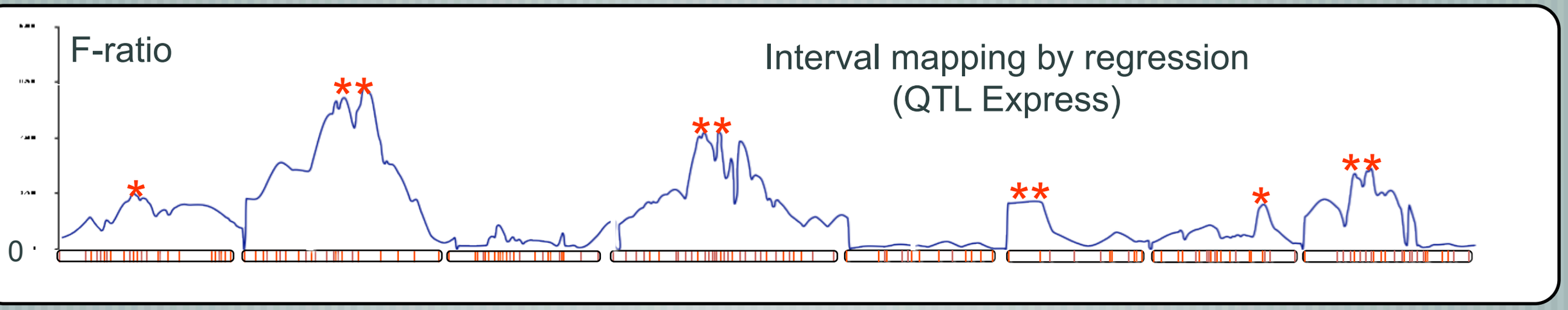


- Move the putative QTL position (Q) every 2 cM from M_1 to M_2 and draw the profile of the F value.
- The peak of the profile corresponds to the best estimate of the QTL position

Interval mapping implementation

- Carry out a QTL scan step-wise: once a significant QTL has been identified, other markers tested for their ability to explain the residual variation

- Known QTL are said to be “fixed” or “co-factors” in the regression

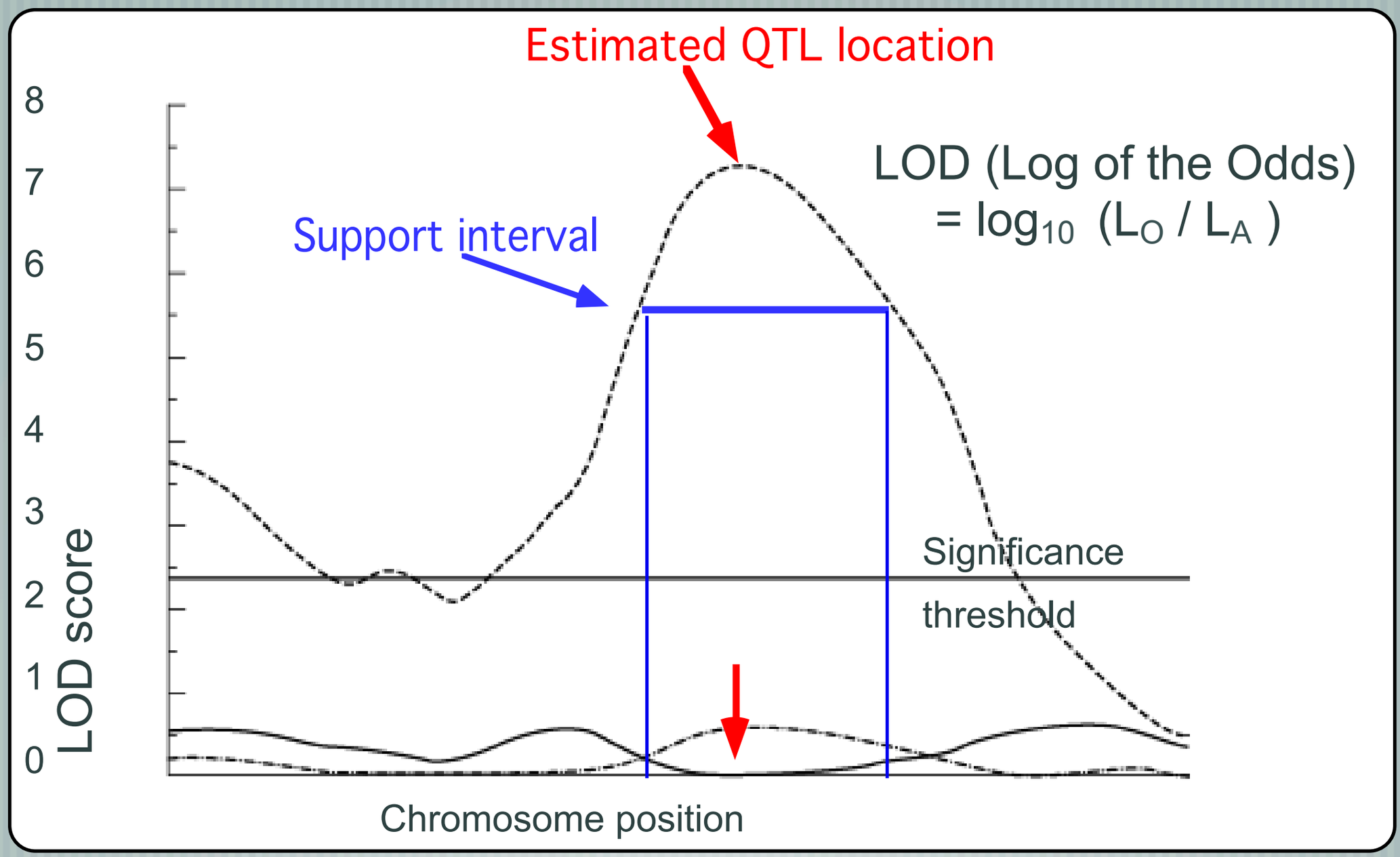


L

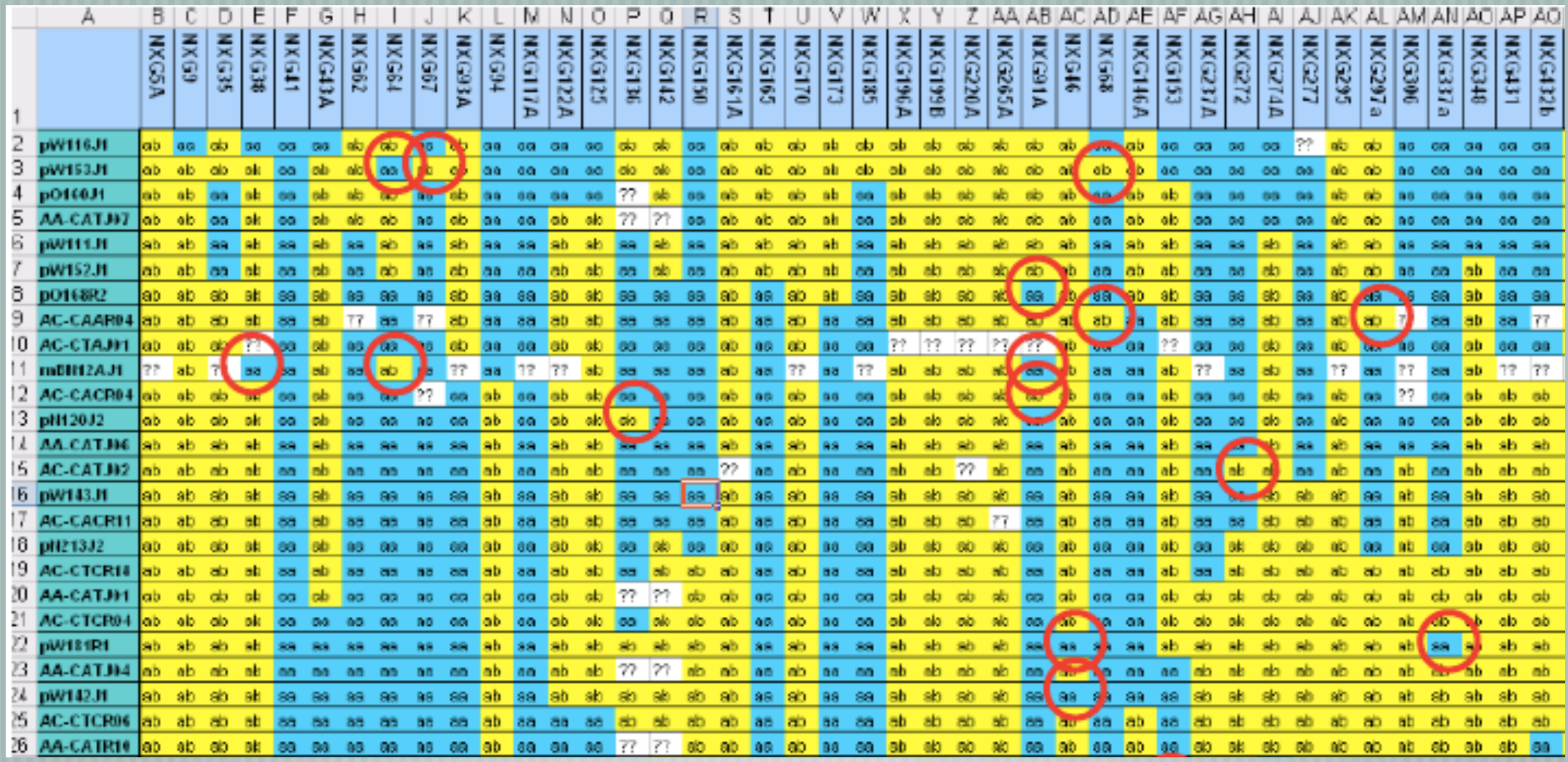
LI

LOD or LRS (=4.6xLOD)

- (L_O / L_A) = ratio of the likelihood of the null hypothesis (no QTL in the marker interval) to the likelihood of the alternative hypothesis (QTL present)

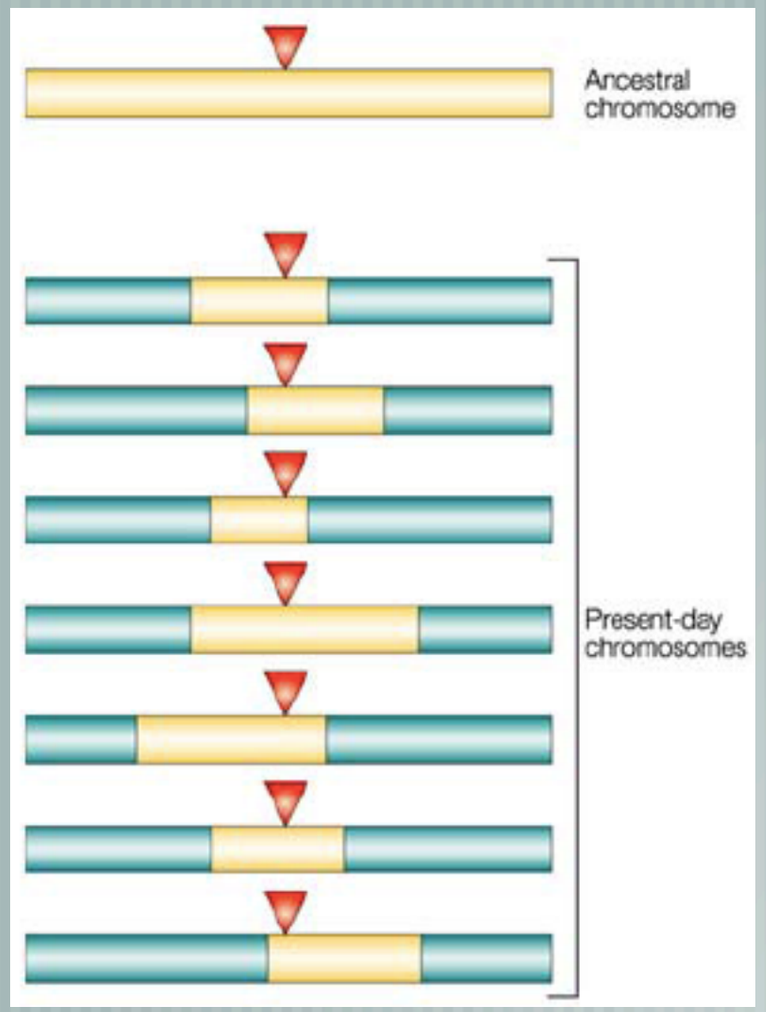


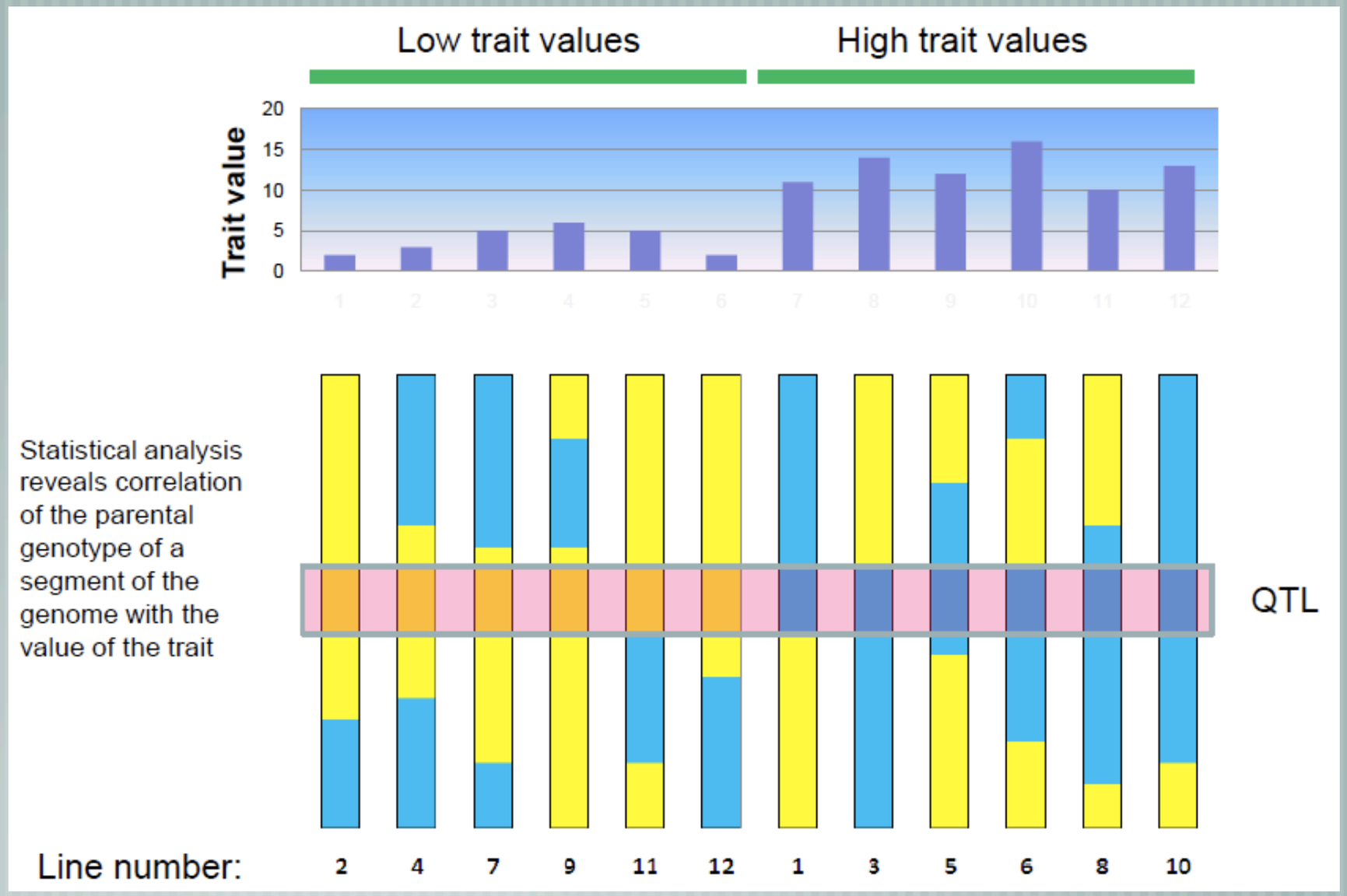
Map sorted genotype data = graphical genotypes

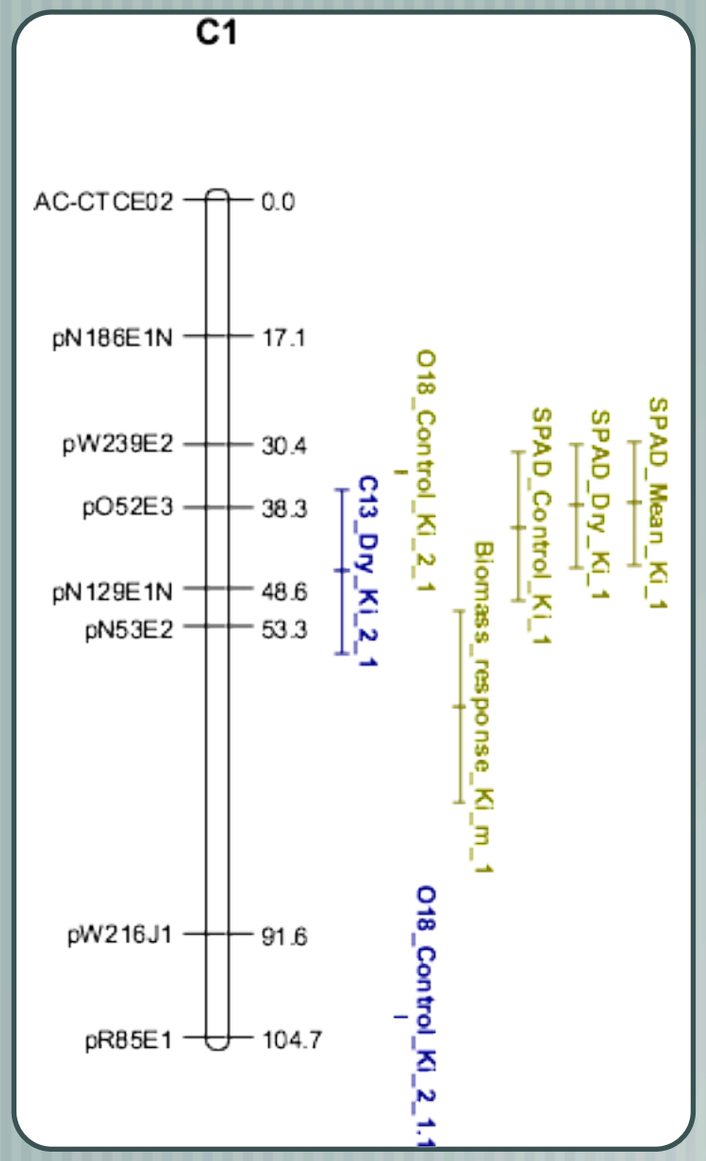


- Format and check the data
- Calculate all pairwise recombination frequencies
- Assign markers to linkage groups then map markers within each linkage group

Quantitative trait loci analysis & association mapping







- Confidence interval
- Threshold of 3 (QTL at position 1000 x more likely than no QTL), simulations confirmed this to be accurate
- Permutation test
- QTL intervals can be plotted on a map.
- For selection it is good to use two markers - one on each side
- Less chance of recombination breaking link

Which? guide to QTL mapping: Interval mapping rating

- Advantages:

the position of the QTL can be inferred by a support interval

the estimated position and effects of the QTL tend to be asymptotically unbiased if there is only one segregating QTL on a chromosome

method requires fewer individuals?

ake appropriate account of missing genotype data

- Disadvantages:

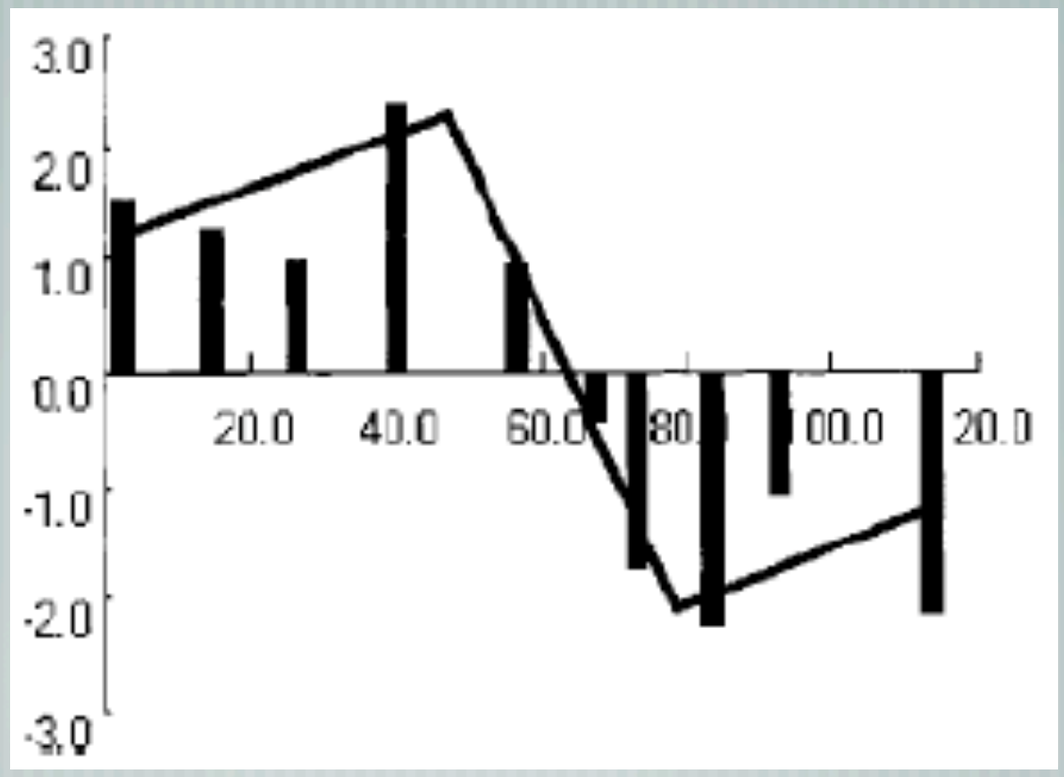
  - even when there is no QTL within an interval, the likelihood profile on the interval can still exceed the threshold if there is a QTL nearby

  - if there is more than one QTL on a chromosome, the test statistic at the position being tested will be affected by all QTL and their estimated positions

  - not efficient to use only two markers at a time for testing

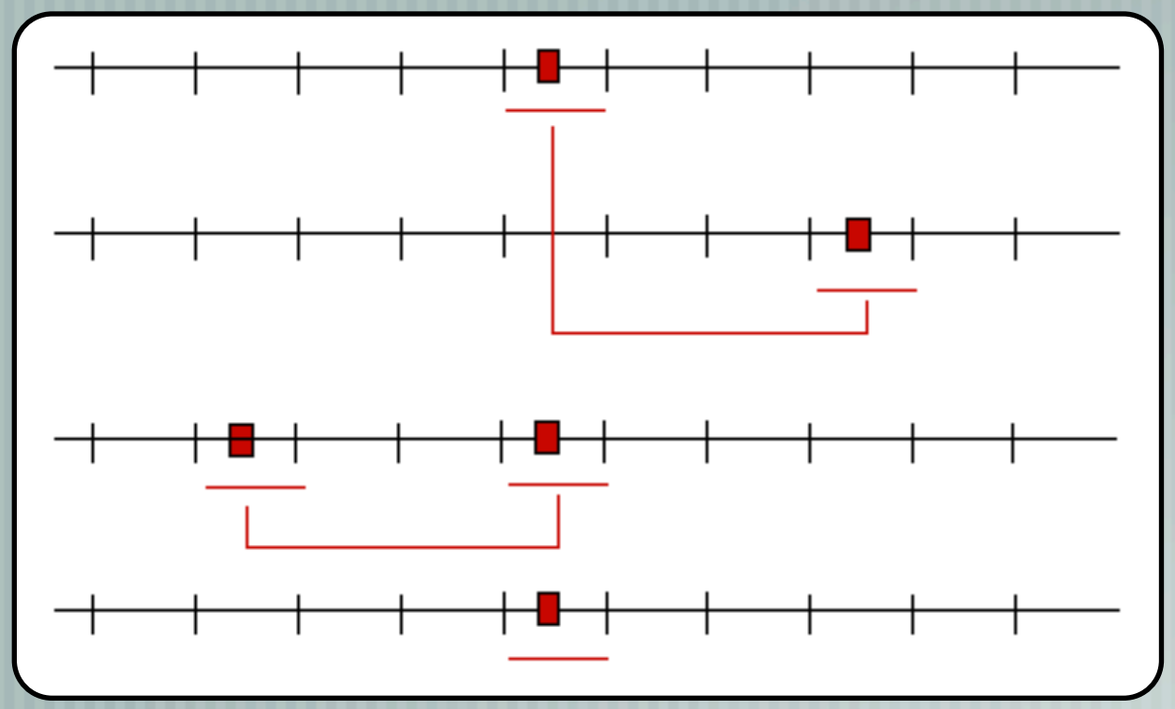
Multiple QTL mapping

- The first scan of QTL analysis usually searches for single QTLs
- Might have opposite effects - can you distinguish them?
- Might have the same effect - QTL ghosts!



Multiple interval mapping

- Uses multiple marker intervals simultaneously
- Aims to map multiple QTLs in a single step

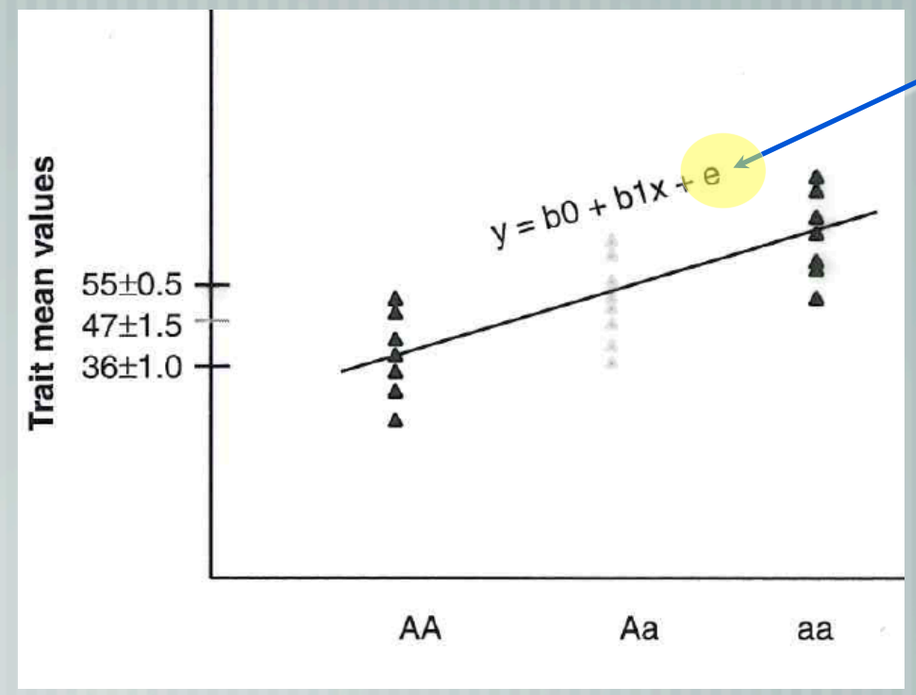


Method:

- Build regression models which include all QTLs (detected first by CIM)
- Use information content (IC) theory to evaluate alternative models
- Allows simultaneous detection and estimation of additive, dominance & epistatic effects

CIM or MQM

Some of the variability between lines that share a common QTL genotype at QTL locus 1 is due to the fact that they have different genotypes at QTL locus 2 somewhere else in the genome.



error term (e) in the CIM model
= true experimental error + variation at other loci (or genetic background segregation)

Permutation testing to determine experiment-wide signficance thresholds

- Multiple testing problem: how often are random QTL effects of a certain magnitude detected in similar datasets?
  - Method: - create a large number of ‘random empirical’ datasets
    - take your marker data and randomly reassign the
    - s back to the marker genotypes
    - QTL detection process
    - highest LR produced for a ‘random QTL’
    - n the map
    - whole process > 1000 times
    - record the magnitude of the lowest ‘random QTL’ observed in the top 5% of LR results = threshold

43	13.5
44	13.4
45	13.2
46	13.1
47	12.9
48	12.7
49	12.6
50	12.5
51	12.3
52	12.1
53	11.8
54	11.6
55	11.5
56	11.3
57	11.2
58	11.0
59	10.5
60	10.4

top 5% of random

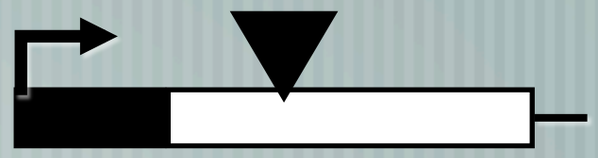
95% of random

QTL to gene

- If genome sequenced:
  - Test candidate genes in interval

- If genome not sequenced:
  - Find syntenic region in sequenced genome

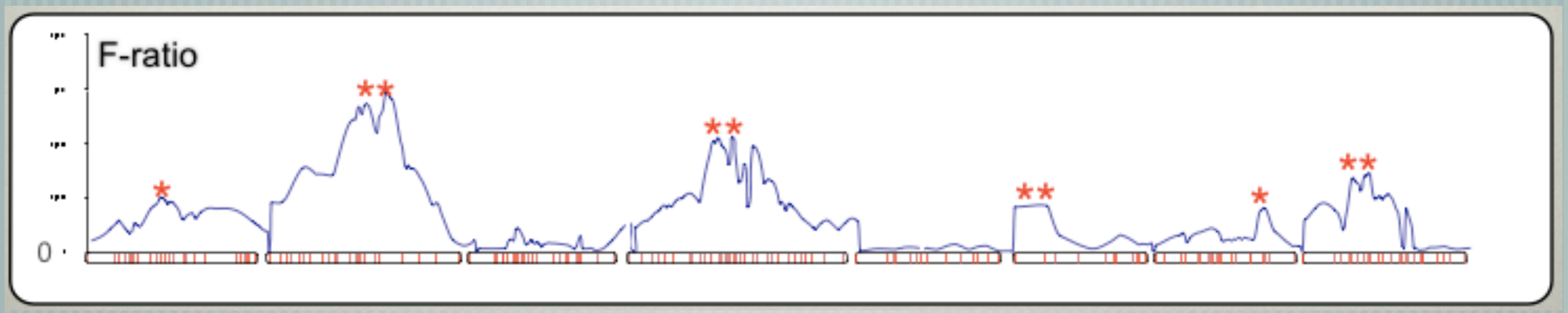
- To confirm identity:
  - Look for mutations in gene in varieties
  - Transfer gene into elite species and determine consequences



Species	Trait	Gene - Function
Tomato	Fruit Size Sugar Content Fruit shape	OFRX – regulatory (?) Lin5 – Apoplastic invertase Ovate - unknown
Rice	Flowering time	Se1- Transcription factor CK2 – protein kinase Hd3a - unknown
Maize	Apical dominance	Tb1 – Transcription factor
Arabidopsis	Flowering time	CRY2 - Cytochrome

Potential problems with QTL analysis methods

- Replication of results is often poor
- Most traits show several (4-30) QTLs of 10-30cM
- Detected QTLs = 5-50% of the observed phenotypic variation
- When isolated in inbred lines, QTLs often show strong interaction effects (G x G, G x E), that are not apparent in a normal analysis
- QTL mapping in populations can only uncover the genetic variation contributed by the parents (usually x 2)



Understanding detection vs. localisation

- One major problem is that most QTL studies are vastly underpowered

- Darvasi & Soller (1997) give an appropriate expression for the sample size required for a 95% position confidence interval (CI)

$$CI = 1500/(n\delta)^2$$

- For a QTL with d = 0.25, 0.1, and 0.05, the sample sizes needed for a 1cM CI are 1500, 3800, and 7600... but typical n = only 350!

- Effect of linkage: for d = 0.05, 0.1, 0.2, increase in sample size (over c = 0) is 1.2, 1.6, 2.8

Power and repeatibility: the Beavis effect

- QTLs with low power of detection tend to have their effects overestimated, often very dramatically

- e.g. a QTL accounting for 0.75% of total F_2 variation has only a 3% chance of being detected with 100 F_2 progeny (markers spaced at 20 cM)

- But when such a QTL is detected, the average estimated total variance it accounts for is 15%!

- The Beavis effect raises the real concern that many QTL of apparent large effect may be artifacts

- Under an infinitesimal model this is especially a concern since it suggests that a few discrete loci account for much of the variation

In practice, the power of QTL analysis depends on:

- ‘Environmental’ variance – minimise this or record differences as fixed effects
- ‘Heritability’ – proportion of total variance that is genetically determined

- Number of recombinants (# individuals, # generations, RF)
- Marker density – rarely limiting with interval mapping
- Accuracy of marker genotyping
- Accuracy of trait data – check for systematic errors
- Some methods are also sensitive to the distribution of trait data (i.e. consider transformation to get a more normal distribution)

Table 6.1 Comparison of different types of methods used in QTL analysis

Features	Single-marker analysis	Simple interval mapping	Composite interval mapping	Multiple QTL mapping
Principle	One marker is involved at a time to find the QTL-marker association	It is based on the joint frequencies of a pair of adjacent markers and a putative QTL flanked by the two markers	Multiple regression methods are integrated with interval mapping to increase the probability of including all significant QTL in the model	It uses multiple marker intervals simultaneously to fit multiple putative QTL directly in the QTL-mapping model
Methods	Simple *t*-test, ANOVA, linear regression, likelihood ratio test, maximum likelihood estimation	Likelihood approach, regression approach or combination of above two approaches	Combining simple interval mapping with multiple regression methods	Cockerham's model for interpreting genetic parameters and the method of maximum likelihood for estimating genetic parameters
Advantages	Simple in terms of data analysis Performed using common statistical software Gene order and complete linkage map are not required	QTL location can be identified	Multiple QTL in a single linkage group can be identified	More powerful and precise than all the above three methods Epistasis between QTL, genotypic values of individuals and heritabilities of quantitative traits can be readily estimated and analysed
Limitations	The putative QTL genotypic means and QTL positions are confounded, and thus it causes biased estimation of QTL effects and low power in detection of such QTL QTL positions cannot be precisely determined due to the nondependence among the hypothesis tests for linked markers that confound QTL effect and position Doing a *t*-test/ANOVA at every marker results in many false positives	Requires prior construction of good quality linkage map Considers one QTL at a time in the model for QTL mapping and hence it is biased in estimation of QTL when multiple QTL are located in the same linkage group	Inclusion of too many cofactors reduced the power to identify QTL relative to interval mapping	Sophisticated high-end systems are required with skilled manpower
Reference	Edwards et al. (1987)	Lander and Botstein (1989)	Jansen (1993), Rodolphe and Lefort (1993), and Zeng (1993)	Kao et al. (1999)

Improving on QTL mapping: Association Mapping

- Uses a random sample of individuals from the population
- applicable for human studies (see the HapMap project)

- Natural populations of a species; ecotypes

- All existing individuals descend from one original population (= Most Recent Common Ancestor, MRCA), that evolved through mutation & selection

- Since this original founding event, many crossing-overs have occurred

- This method uses historical recombinants: linkage disequilibrium analysis

Improving on QTL mapping: Association Mapping

- Key is the expected number of recombinants; c = recombination fraction

  Probability (number recombinants) in n individuals is $(1-c)^n$

- LD mapping uses the 'historical recombinants' in a sample; t = time to MRCA

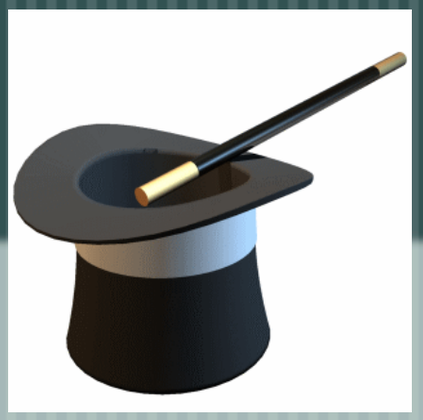
  Probability (number recombinants) = $(1-c)^{2t}$

- Hence, if t is large, many more expected recombinants in random sample and hence more power for very fine mapping (i.e. c < 0.01)

- To perform association mapping with linkage disequilibrium you need dense markers (normally limited to sequenced model organisms)

- Since the linkage disequilibrium region is very small, this approach has more power for fine mapping (get closer to the important gene), even with small sample sizes

Material for association mapping



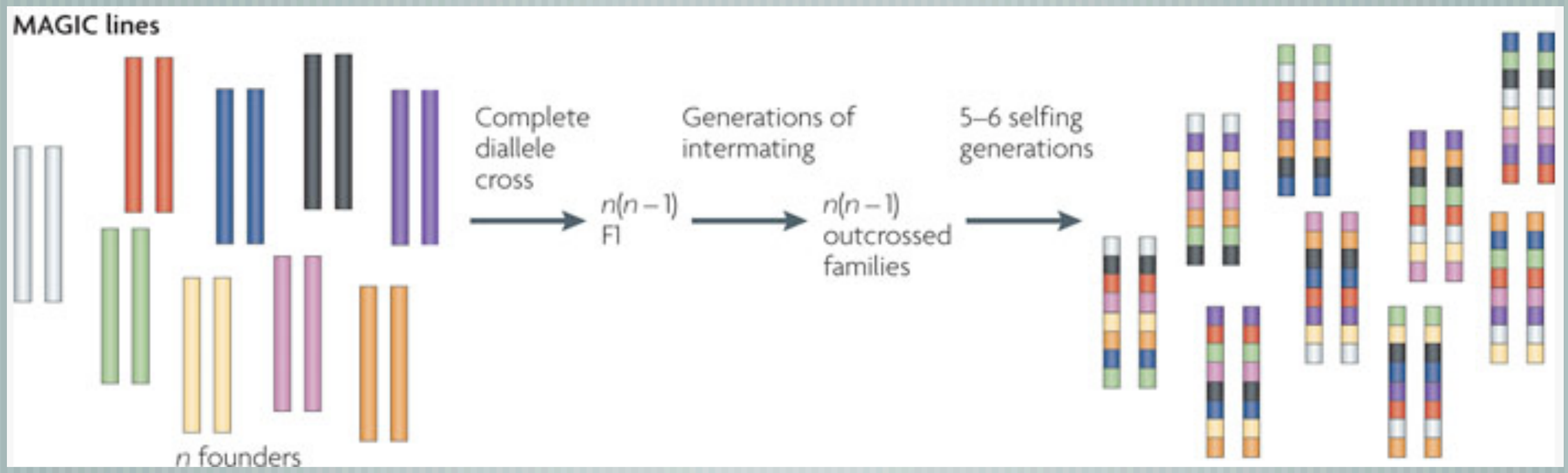
- MAGIC lines: Multi-parent Advanced Generation Intercross
- Give genetic map resolution of about 300kb or 50 genes
- Kover et al (2009)

~ Mackay and Powell (2006) Methods for linkage disequilibrium mapping in crops. Trends in Plant Science 12: 57-63

~ 19 genetically diverse founders | Intercrossed for 4 generations | inbred for 6 generations = 1026 MLs

~MLs are homozygous - replication

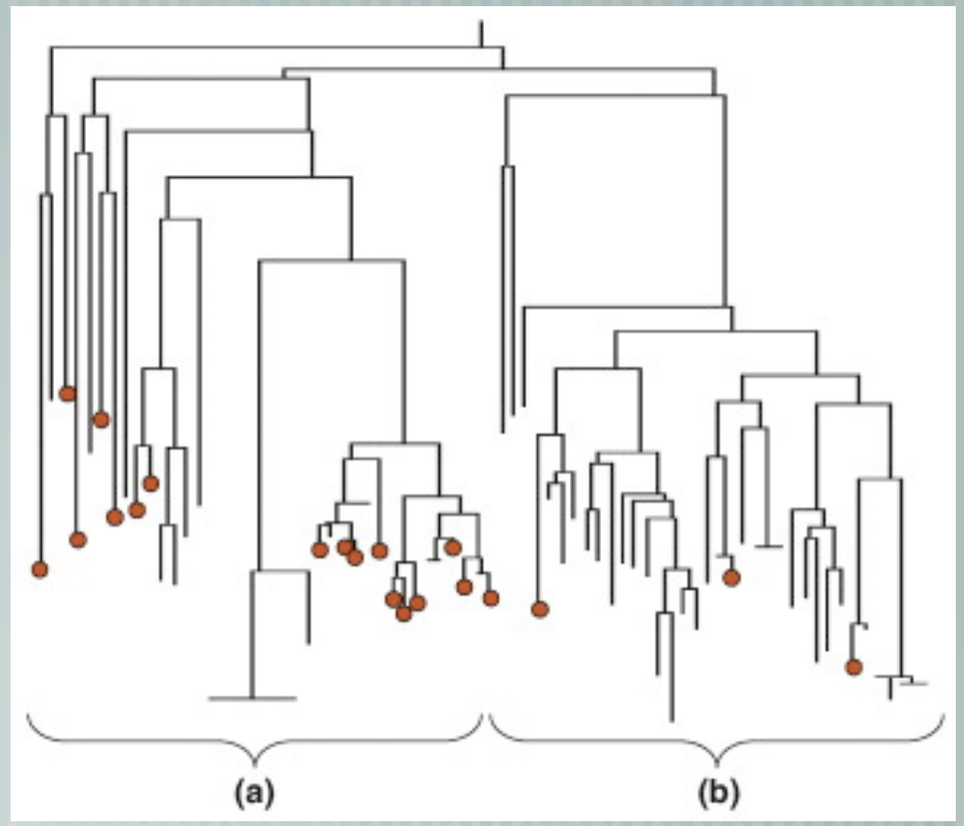
Rafalski(2010) Association genetics in crop improvement. Current Opinion in Plant Biology 13:174-180Zhu et al (2008) Status and prospects of association mapping in plants. The Plant Genome 1:2-20



Which? guide: association mapping rating

- Can have marker-trait associations in the absence of linkage
- e.g. if a marker predicts group membership, and being in that group gives you a different trait value, then a marker-covariance will occur
- Since a sample population actually consists of several distinct subpopulations we have lumped together marker alleles may provide information as to which group an individual belongs

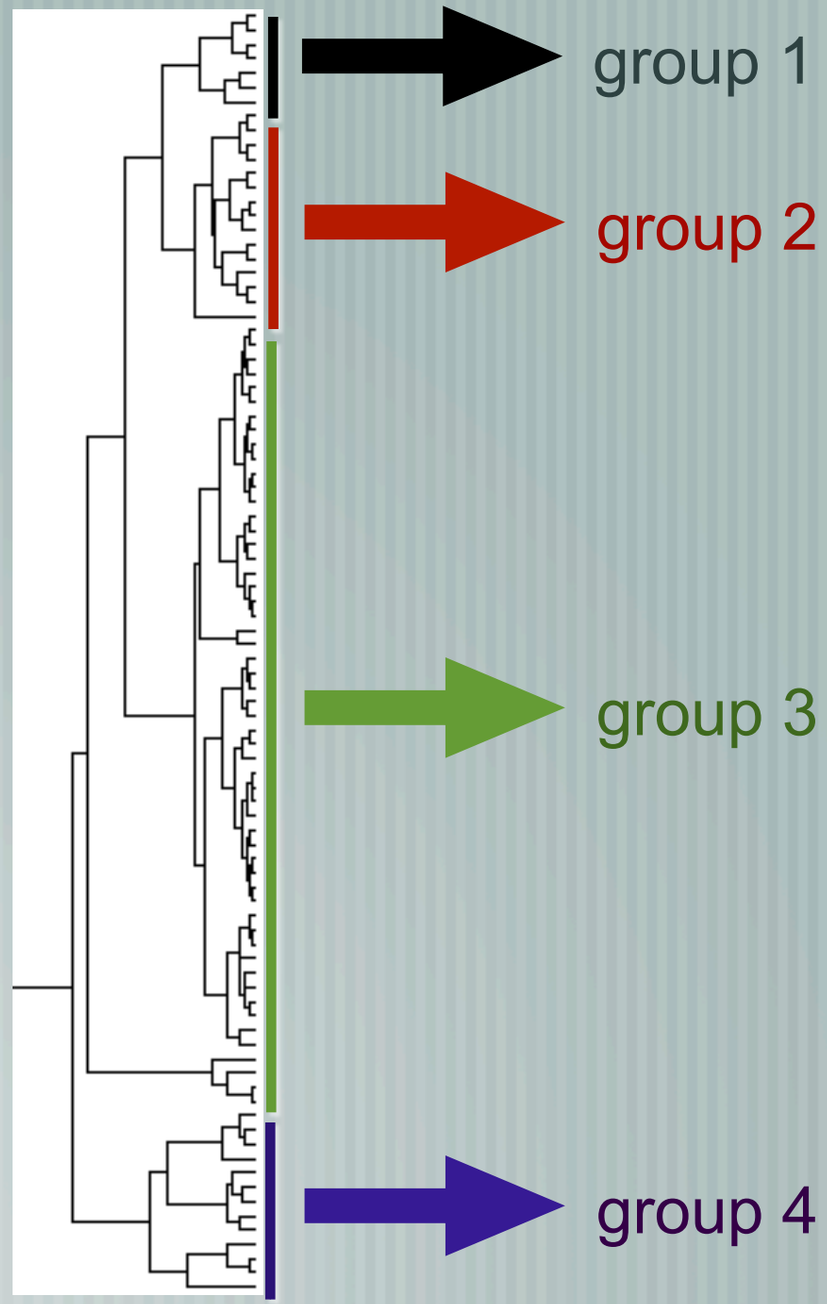
=stratification (population substructure)



Two ways to adjust for population stratification

- (1) Use molecular makers to classify individuals into groups

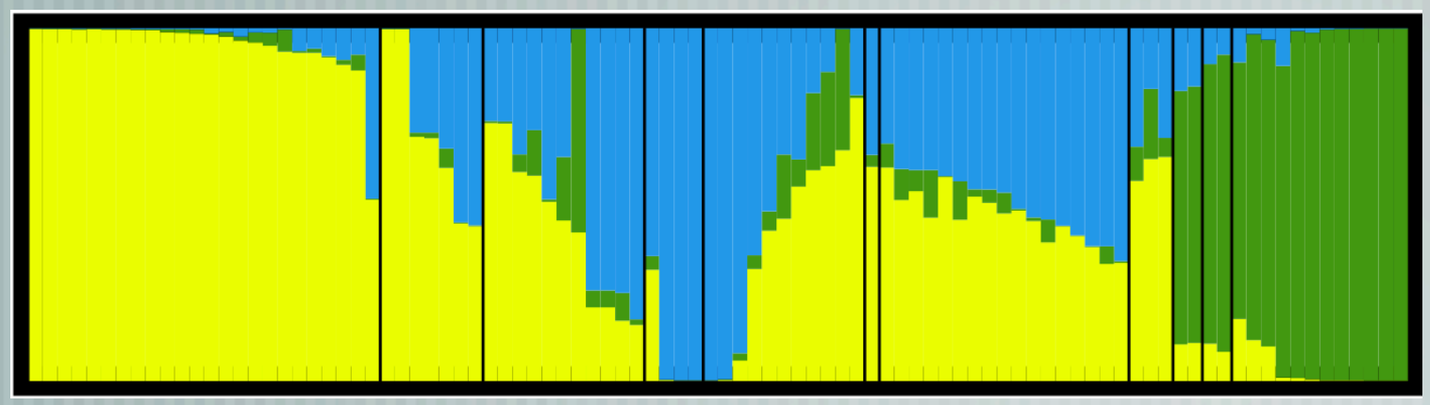
  Then run association mapping within each group (structured AM)

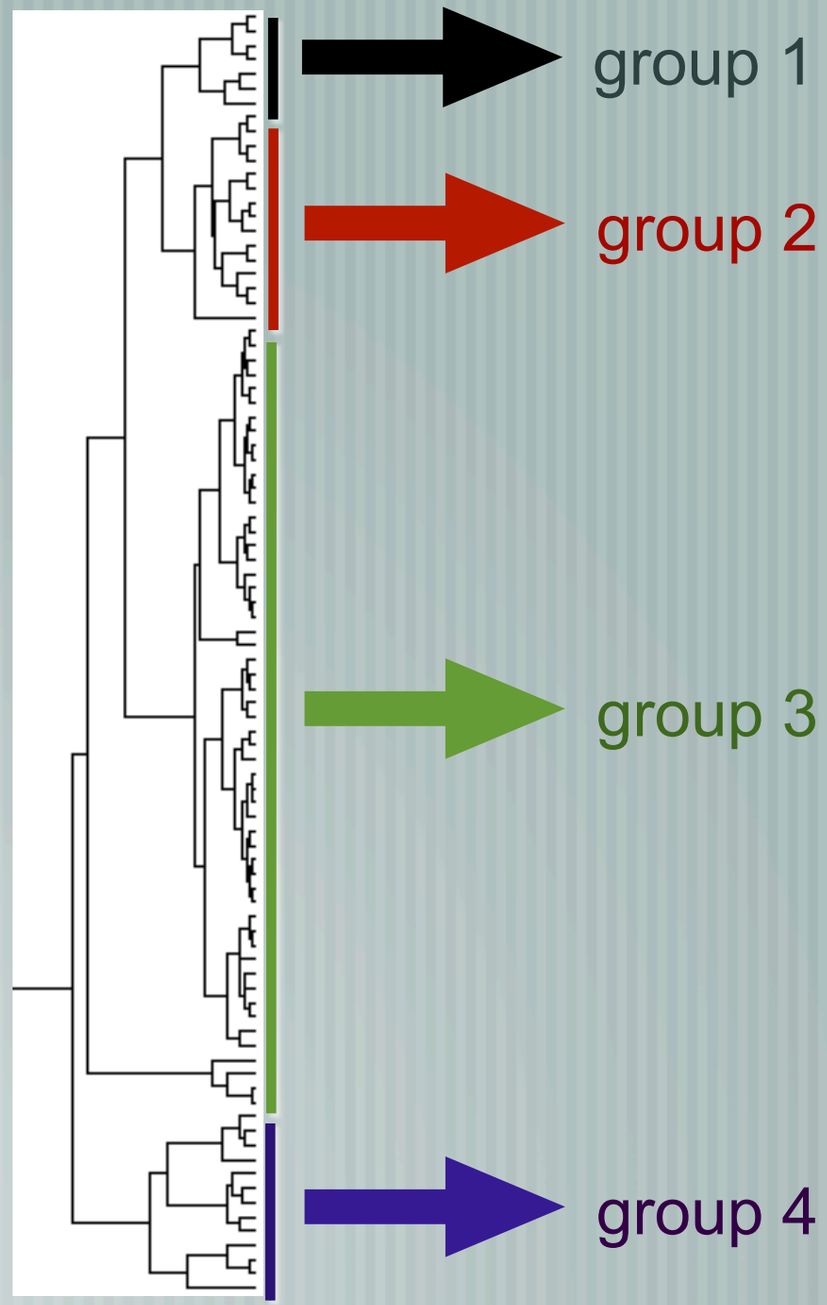


Two ways to adjust for population stratification

- (2) Use a simple regression approach, adding additional markers as cofactors for group membership, removing the group effect

- Generate a Q matrix using STRUCTURE





Method	Advantages	Disadvantages
QTL mapping	No population structure Allows identification of rare alleles Doesn't require large numbers of markers	Coarse mapping Limited genetic diversity
Association mapping (candidate gene approach)	Fine mapping	Existing knowledge required of trait Biased for previously identified genes
Association mapping (genome scan)	Fine mapping Blind approach Identification of common alleles	False positives/false negatives Not good at identifying rare alleles
Dual approach (QTL and association)	Fine mapping Blind approach Can identify false positives and negatives	Resource intensive
MAGIC lines	High resolution Increased genetic diversity Genotyped once – unlimited replication	Genetic and allelic variation. Requires ~10 generations to set up

SNPs	ecotypes	crossing over	population stratification
dense SNPs	genetic linkage	historical recombinants	Haldane function
genotyping errors	Beavis effect	independent assortment	linkage disequilibrium
Kosambi function	recombination frequency	low population sample size	random sample of individuals

ecotypes	Kosambi function	population stratification	SNPs
dense SNPs	genetic linkage	Beavis effect	crossing over
historical recombinants	Haldane function	genotyping errors	linkage disequilibrium
random sample of individuals	recombination frequency	low population sample size	independent assortment

things to do with association mapping	things to do with making a genetic map	reasons for wrong QTL identification	things to do with genetic variation