QTL Mapping, MAS, and Genomic Selection

Dr. Ben Hayes

Department of Primary Industries

Victoria, Australia

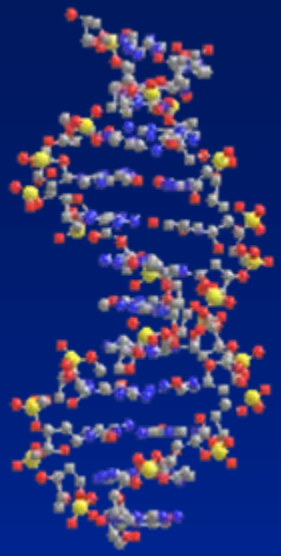
A short-course organized by

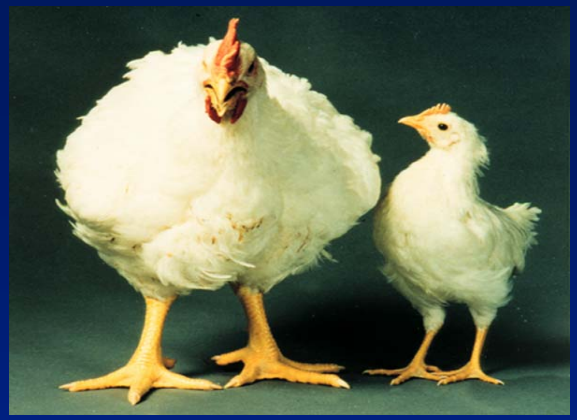
Animal Breeding & Genetics
Department of Animal Science
Iowa State University
June 4-8, 2007

With financial support from
Pioneer Hi-bred Int.

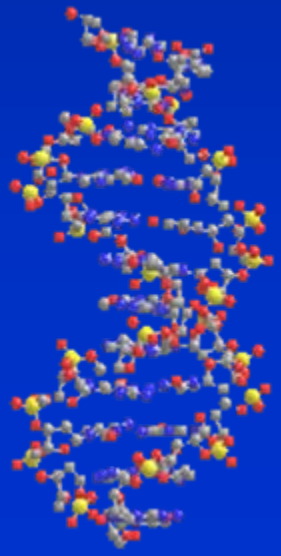
USE AND ACKNOWLEDGEMENT OF SHORT COURSE MATERIALS

Materials provided in these notes are copyright of Dr. Ben Hayes and the Animal Breeding and Genetics group at Iowa State University but are available for use with proper acknowledgement of the author and the short course. Materials that include references to third parties should properly acknowledge the original source.





Linkage Disequilbrium to Genomic Selection



Course overview

- Day 1
  - Linkage disequilibrium in animal and plant genomes
- Day 2
  - QTL mapping with LD
- Day 3
  - Marker assisted selection using LD
- Day 4
  - Genomic selection
- Day 5
  - Genomic selection continued

Mapping QTL using LD

- Association testing with single marker regression
- Accounting for population structure
- LD mapping with haplotypes
- The Identical by descent (IBD) approach
- Combined linkage-linkage disequilibrium mapping

Mapping QTL with LD

- LD mapping of QTL exploits population level associations between markers and QTL.
  - Associations arise because there are small segments of chromosome in the current population which are descended from the same common ancestor
  - These chromosome segments, which trace back to the same common ancestor without intervening recombination, will carry identical marker alleles or marker haplotypes
  - If there is a QTL somewhere within the chromosome segment, they will also carry identical QTL alleles
- The simplest mapping strategy to exploit LD is a genome wide association test using single marker regression.

Single marker regression

- Association between a marker and a trait can be tested with the model

$$\mathbf{y} = \mathbf{1_n}\boldsymbol{\mu} + \mathbf{X}g + \mathbf{e}$$

- Where
  - **y** is a vector of phenotypes
  - **1n** is a vector of 1s allocating the mean to phenotype,
  - **X** is a design matrix allocating records to the marker effect,
  - *g* is the effect of the marker
  - **e** is a vector of random deviates ~ $N(0,\sigma_e^2$
- Underlying assumption here is that the marker will only affect the trait if it is in linkage disequilibrium with an unobserved QTL.

Single marker regression

- A small example

Animal	Phenotpe	SNP allele 1	SNP allele
1	2.030502	1	1
2	3.542274	1	2
3	3.834241	1	2
4	4.871137	2	2
5	3.407128	1	2
6	2.335734	1	1
7	2.646192	1	1
8	3.762855	1	2
9	3.689349	1	2
10	3.685757	1	2

Single marker regression

- The design vector $\mathbf{1_n}$ allocates phenotypes to the mean

Animal	Phenotpe	SNP allele 1	SNP allele
1	2.030502	1	1
2	3.542274	1	2
3	3.834241	1	2
4	4.871137	2	2
5	3.407128	1	2
6	2.335734	1	1
7	2.646192	1	1
8	3.762855	1	2
9	3.689349	1	2
10	3.685757	1	2

Animal	$\mathbf{1_n}$
1	**1**
2	**1**
3	**1**
4	**1**
5	**1**
6	**1**
7	**1**
8	**1**
9	**1**
10	**1**

Single marker regression

- The design vector $\mathbf{1_n}$ allocates phenotypes to the mean
- The design vector **X** allocates phenotypes to genotypes

Animal	Phenotpe	SNP allele 1	SNP allele
1	2.030502	1	1
2	3.542274	1	2
3	3.834241	1	2
4	4.871137	2	2
5	3.407128	1	2
6	2.335734	1	1
7	2.646192	1	1
8	3.762855	1	2
9	3.689349	1	2
10	3.685757	1	2

Animal	$\mathbf{1_n}$	**X**, Number of "2" alleles
1	**1**	**0**
2	**1**	**1**
3	**1**	**1**
4	**1**	**2**
5	**1**	**1**
6	**1**	**0**
7	**1**	**0**
8	**1**	**1**
9	**1**	**1**
10	**1**	**1**

Single marker regression

- The design vector $\mathbf{1_n}$ allocates phenotypes to the mean
- The design vector **X** allocates phenotypes to genotypes

Animal	Phenotpe	SNP allele 1	SNP allele
1	2.030502	1	1
2	3.542274	1	2
3	3.834241	1	2
4	4.871137	2	2
5	3.407128	1	2
6	2.335734	1	1
7	2.646192	1	1
8	3.762855	1	2
9	3.689349	1	2
10	3.685757	1	2

y vector

Animal	$\mathbf{1_n}$	**X**, Number of "2" alleles
1	**1**	**0**
2	**1**	**1**
3	**1**	**1**
4	**1**	**2**
5	**1**	**1**
6	**1**	**0**
7	**1**	**0**
8	**1**	**1**
9	**1**	**1**
10	**1**	**1**

Single marker regression

- Estimate the marker effect and the mean as:

$$\begin{bmatrix} \hat{\mu} \\ \hat{g} \end{bmatrix} = \begin{bmatrix} \mathbf{1_n'1_n} & \mathbf{1_n'X} \\ \mathbf{X'1_n} & \mathbf{X'X} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{1_n'y} \\ \mathbf{X'y} \end{bmatrix}$$

Single marker regression

$$\begin{bmatrix} \hat{\mu} \\ \hat{g} \end{bmatrix} = \begin{bmatrix} \mathbf{1_n'1_n} & \mathbf{1_n'X} \\ \mathbf{X'1_n} & \mathbf{X'X} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{1_n'y} \\ \mathbf{X'y} \end{bmatrix}$$

$$\begin{bmatrix} 1111111111 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} = 10$$

Single marker regression

$$\begin{bmatrix} \hat{\mu} \\ \hat{g} \end{bmatrix} = \begin{bmatrix} \mathbf{1_n'1_n} & \mathbf{1_n'X} \\ \mathbf{X'1_n} & \mathbf{X'X} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{1_n'y} \\ \mathbf{X'y} \end{bmatrix}$$

$$\begin{bmatrix} 1111111111 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 1 \\ 2 \\ 1 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \end{bmatrix} = 8$$

Single marker regression

$$\begin{bmatrix} \hat{\mu} \\ \hat{g} \end{bmatrix} = \begin{bmatrix} 10 & 8 \\ 8 & 10 \end{bmatrix}^{-1} \begin{bmatrix} 33.8 \\ 31.7 \end{bmatrix}$$

Single marker regression

$$\begin{bmatrix} \hat{\mu} \\ \hat{g} \end{bmatrix} = \begin{bmatrix} 0.28 & -0.22 \\ -0.22 & 0.28 \end{bmatrix} \begin{bmatrix} 33.8 \\ 31.7 \end{bmatrix}$$

Single marker regression

- Estimates of the mean and marker effect are:

$$\begin{bmatrix} \hat{\mu} \\ \hat{g} \end{bmatrix} = \begin{bmatrix} 2.36 \\ 1.38 \end{bmatrix}$$

- In the "simulation", mean was 2, r^2 between QTL and marker was 1, and effect of 2 allele at QTL was 1.

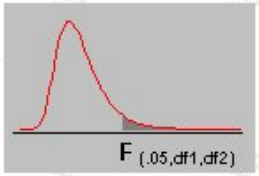
Single marker regression

- Is the marker effect significant?
- ***F*** statistic comparing between marker variance to within marker variance
- Test against tabulated value for $F_{\alpha,v1,v2}$
  - α= significance value
  - v1=1 (1 marker effect for regression)
  - v2=9 (number of records -1)

Single marker regression

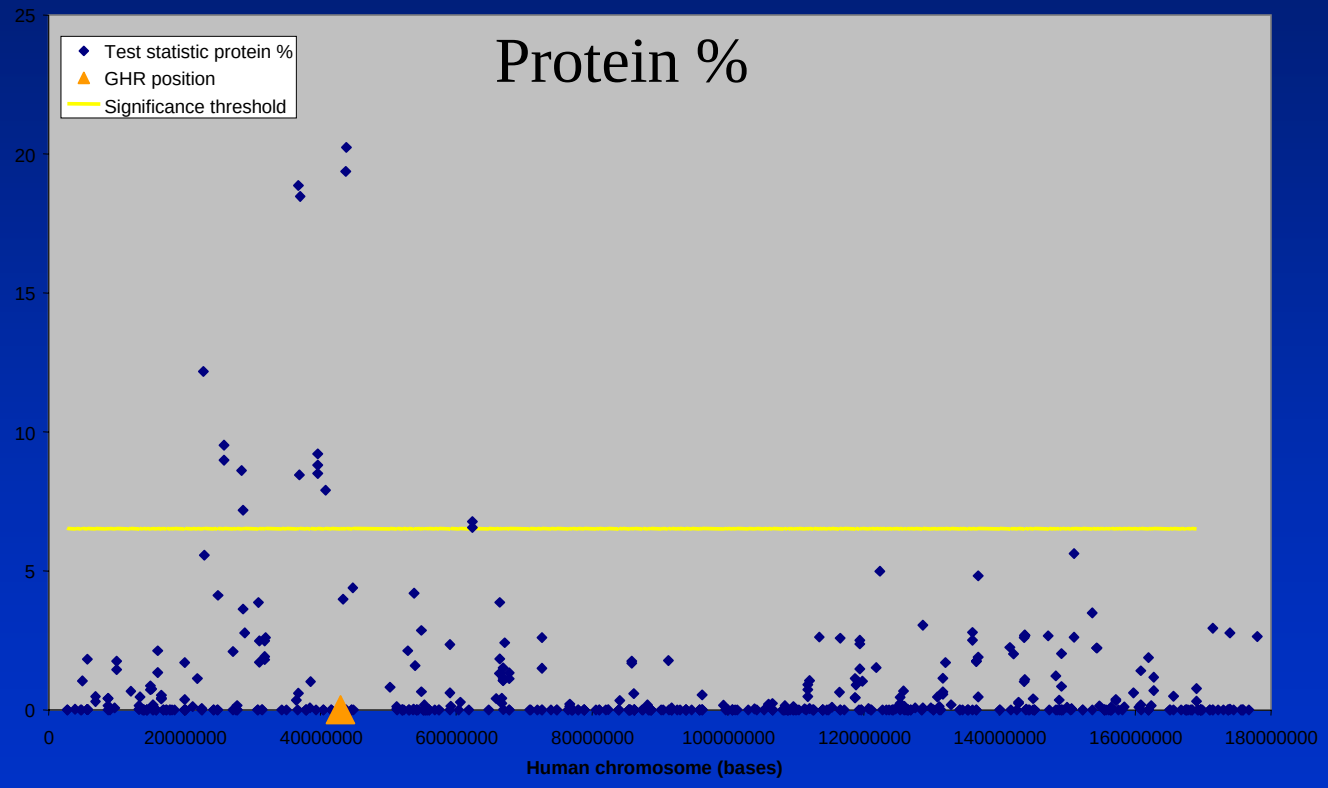
- In our simple example
  - $F_{data}=4.56$
  - $F_{0.05,1,9}=5.12$
- Not significant

F Table for alpha=.05 .



df2/df1	1	2	3	4	5	6	7	8	9	10
1	161.4476	199.5000	215.7073	224.5832	230.1619	233.9860	236.7684	238.8827	240.5433	241.8817
2	18.5128	19.0000	19.1643	19.2468	19.2964	19.3295	19.3532	19.3710	19.3848	19.3959
3	10.1280	9.5521	9.2766	9.1172	9.0135	8.9406	8.8867	8.8452	8.8123	8.7855
4	7.7086	6.9443	6.5914	6.3882	6.2561	6.1631	6.0942	6.0410	5.9988	5.9644
5	6.6079	5.7861	5.4095	5.1922	5.0503	4.9503	4.8759	4.8183	4.7725	4.7351

Results of genome scans with dense SNP panels



Experiment

- 384 Holstein-Friesian dairy bulls selected from Australian dairy bull population
- genotyped for 10 000 SNPs
- Single marker regression with protein%

Single marker regression

- What is the power of an association test with a certain number of records to detect a QTL?
- Power is probability of correctly rejecting null hypothesis when a QTL of really does exist in the population
- How many animals do we need to genotype and phenotype?

Single marker regression

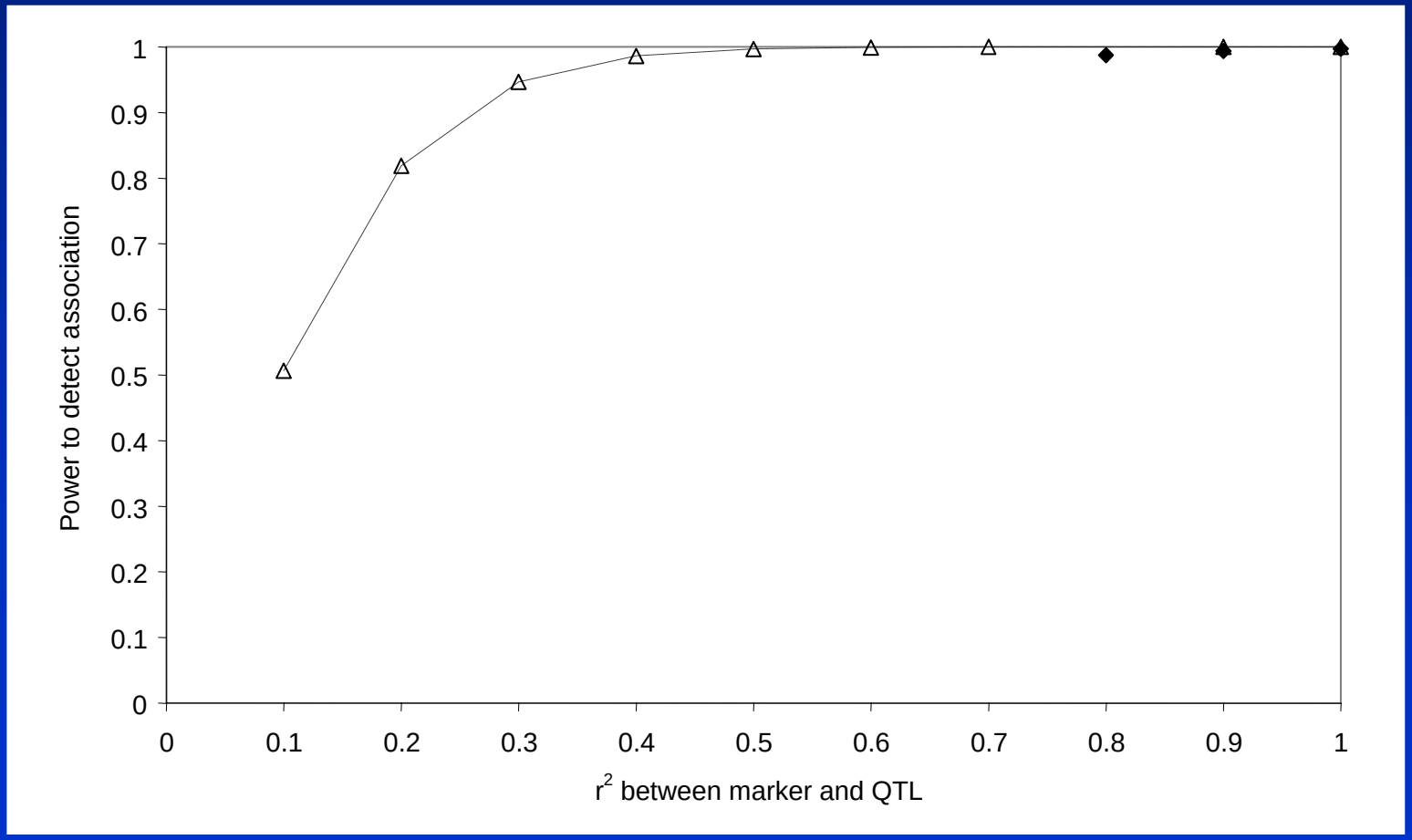
- Power is a function of:
  - r^2 between the marker and QTL
    - sample size must be increased by $1/r^2$ to detect an un-genotyped QTL, compared with sample size for testing QTL itself
  - Proportion of total phenotypic variance explained by the QTL
  - Number of phenotypic records
  - Allele frequency of the rare allele of SNP
    - determines the minimum number of records used to estimate an allele effect.
    - The power becomes particular sensitive with very low frequencies (eg. <0.1).
  - The significance level α set by the experimenter

Single marker regression

- Power is a function of:
  - r^2 between the marker and QTL
    - sample size must be increased by $1/r^2$ to detect an un-genotyped QTL, compared with sample size for testing QTL itself
  - Proportion of total phenotypic variance explained by the QTL
  - Number of phenotypic records

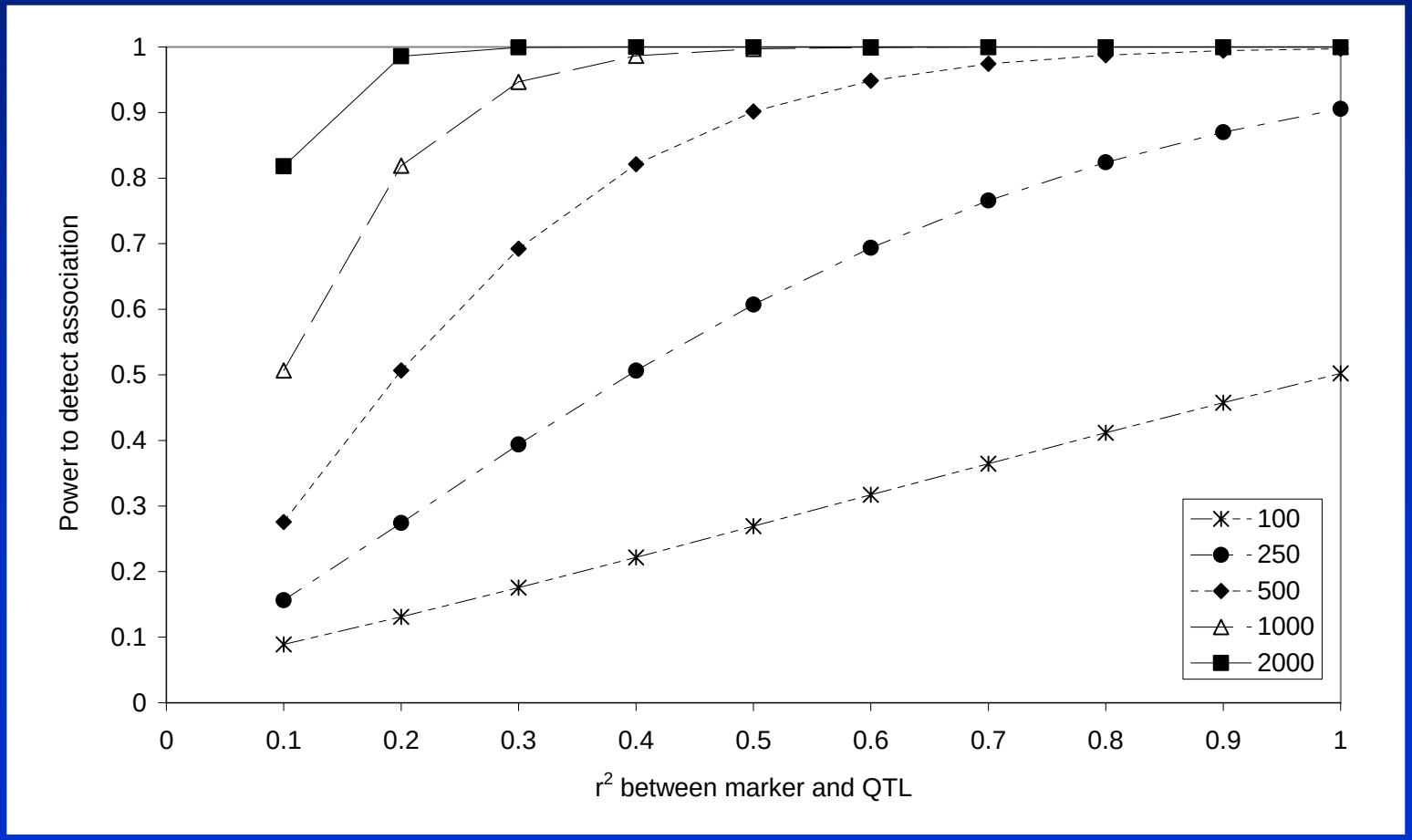
Single marker regression

- Power to detect a QTL explaining 5% of the phenotypic variance, 1000 phenotypic records



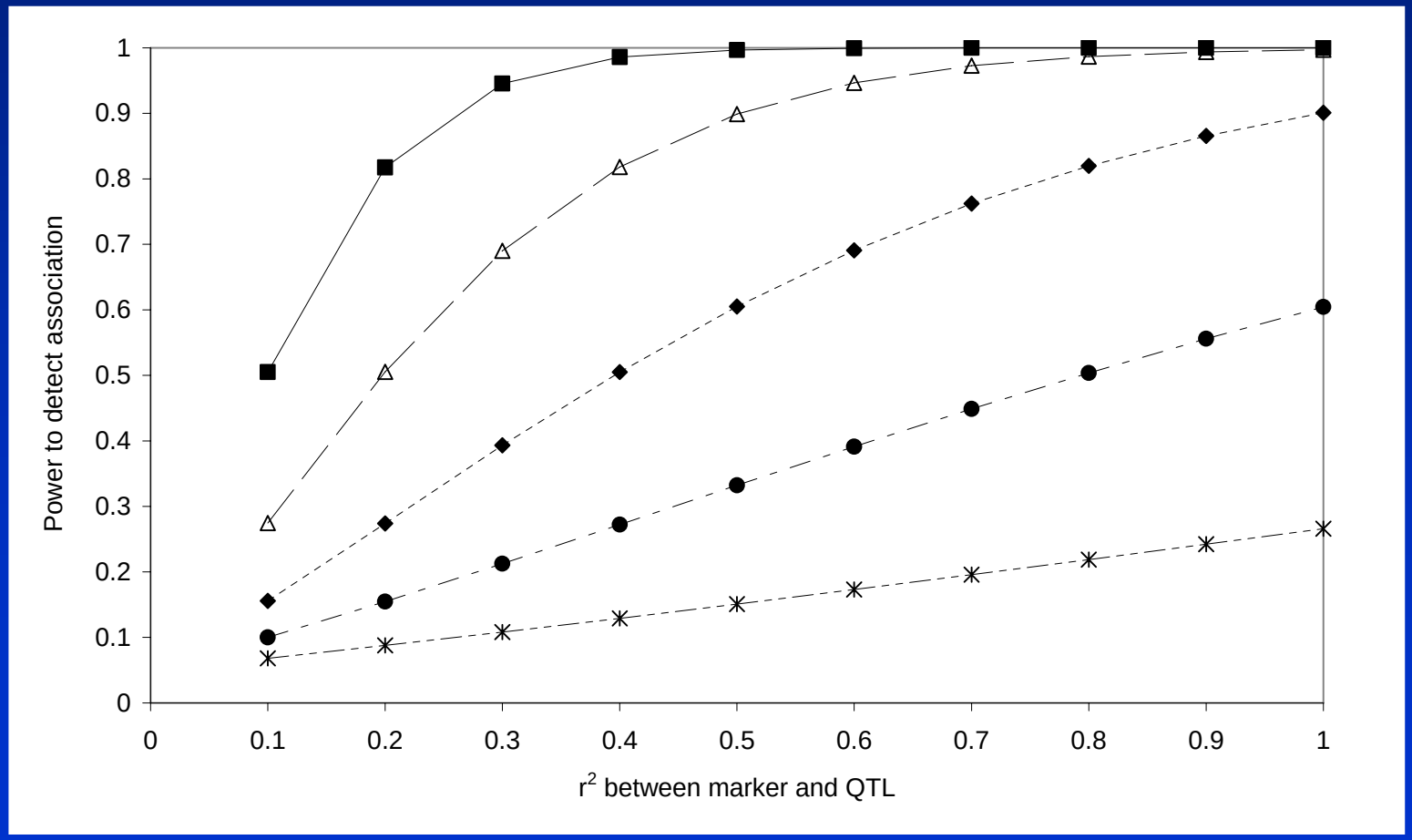
Single marker regression

- Power to detect a QTL explaining 5% of the phenotypic variance



Single marker regression

- Power to detect a QTL explaining 2.5% of the phenotypic variance



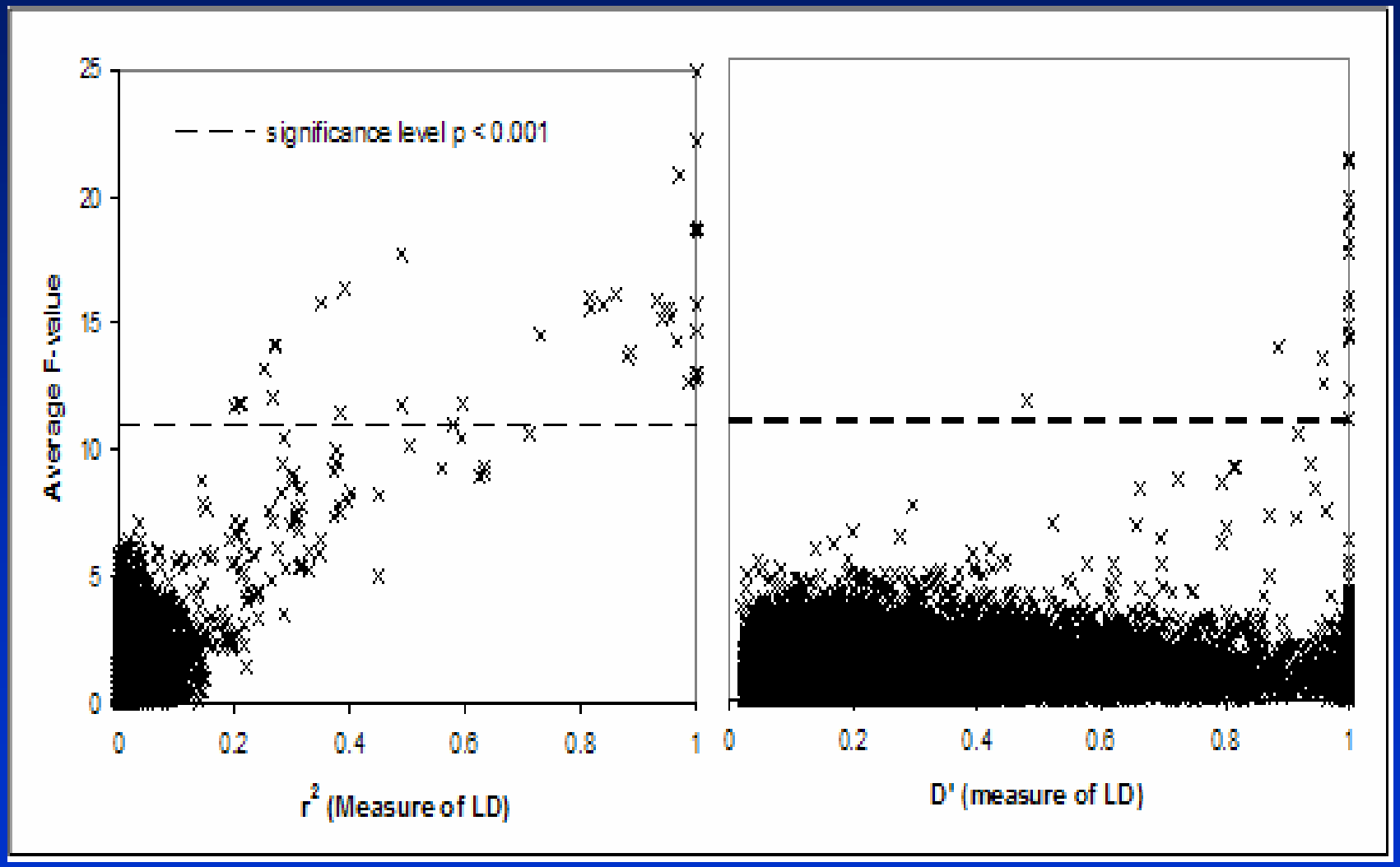
Single marker regression

- r^2 of at least 0.2 is required to achieve power ≥ 0.8 to detect a QTL of h_{QTL}=0.05 with 1000 phenotypic records.
- In dairy cattle, $r^2 \approx 0.2$ at 100kb.
- Assuming a genome length of 3000Mb in cattle, we would need at least 15 000 markers to ensure there is a marker 100kb from every QTL.
- Assumes markers are evenly spaced, all have rare allele frequency > 0.2.
- In practise, markers not be evenly spaced, rare allele frequency of some markers below 0.2.
- At least 30 000 markers required.

Single marker regression

- Another illustration of effect of r^2 on power
- An experiment to assess power of whole genome association scans in outbred livestock with commercially available SNP panels
  - 384 Angus cattle genotyped for 10,000 SNPs
  - QTL, polygenic and environmental effects were simulated for each animal
  - QTL simulated on genotyped SNPs chosen at random.
  - There was a strong correlation between F-value of significant SNPs and their r^2 with the "QTL"

Single marker regression



Single marker regression

- What significance level to use?
  - $P<0.01$, $P<0.001$?
- We have a horrible multiple testing problem
  - Eg. If test 10 000 SNP at $P<0.01$ expect 100 significant results just by chance?
- Could just correct for the number of tests
  - But is too stringent, ignores the fact that tests are on the same chromosome (eg not independent)

Single marker regression

- Could use a technique called permutation testing
  - Randomly shuffle phenotypes across genotypes
  - Test all SNPs (null hypothesis), get largest F value
  - Repeat 1000 times
  - 950th value is $P<0.05$ level corrected for multiple testing
- Difficult with pedigree structure

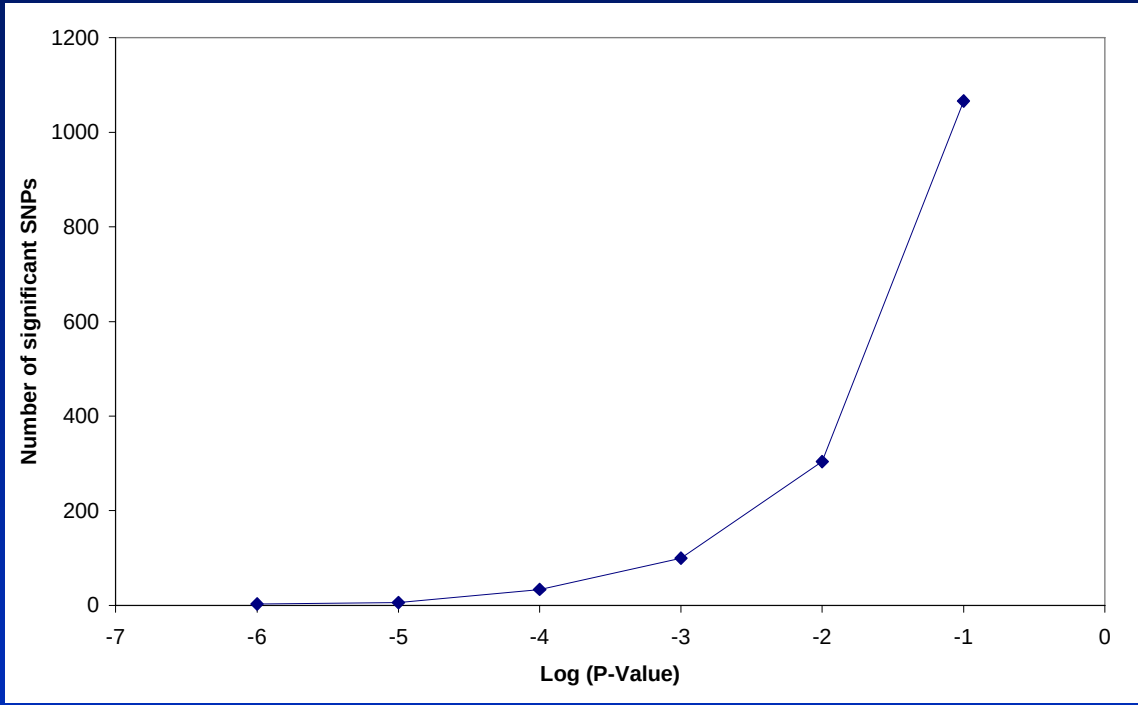
Single marker regression

- An alternative is to choose a significance level with an acceptable false discovery rate (FDR)
- Proportion of significant results which are really false positives
- FDR = mP/n
  - m = number of markers tested
  - P = significance level (eg. P=0.01)
  - n = number of markers tested

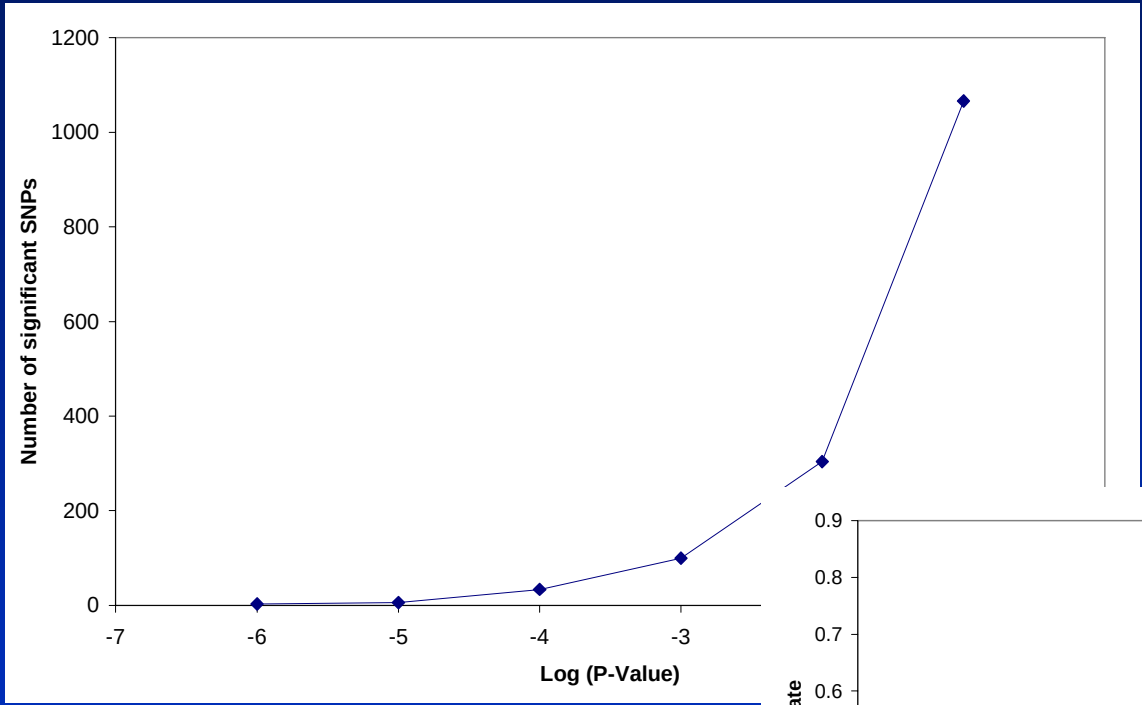
Single marker regression

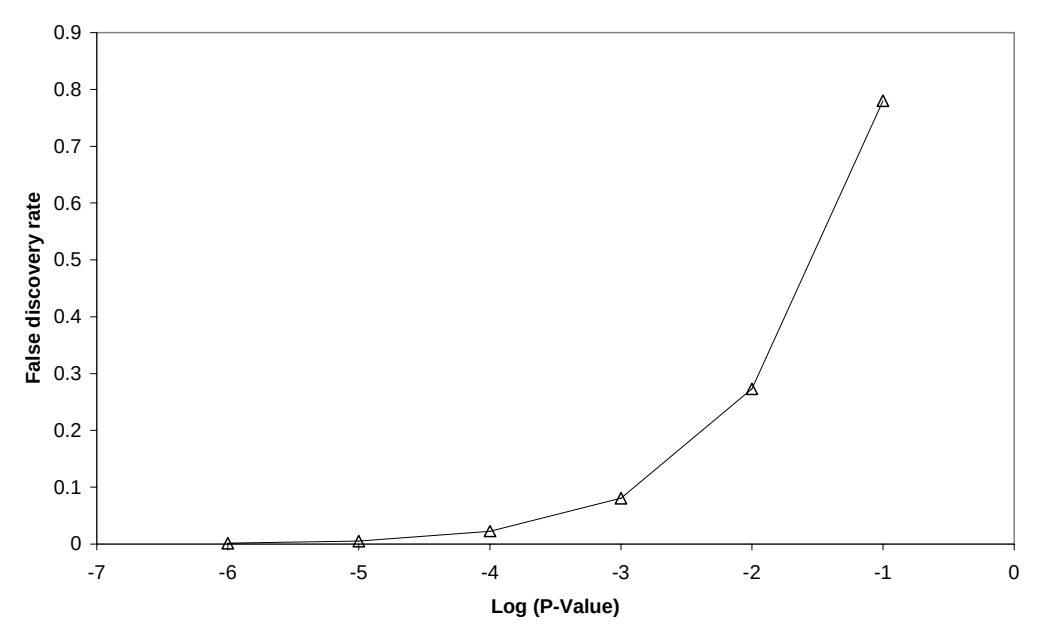
- An alternative is to choose a significance level with an acceptable false discovery rate (FDR)
- Proportion of significant results which are really false positives
- FDR = mP/n
  - m = number of markers tested
  - P = significance level (eg. P=0.01)
  - n = number of markers tested
- Example
  - 10 000 markers tested at P<0.001, and 20 significant. What is FDR?
  - FDR=10000*0.001/20 = 50%
  - Eg. 50% of our significant results are actually false positives

Single marker regression



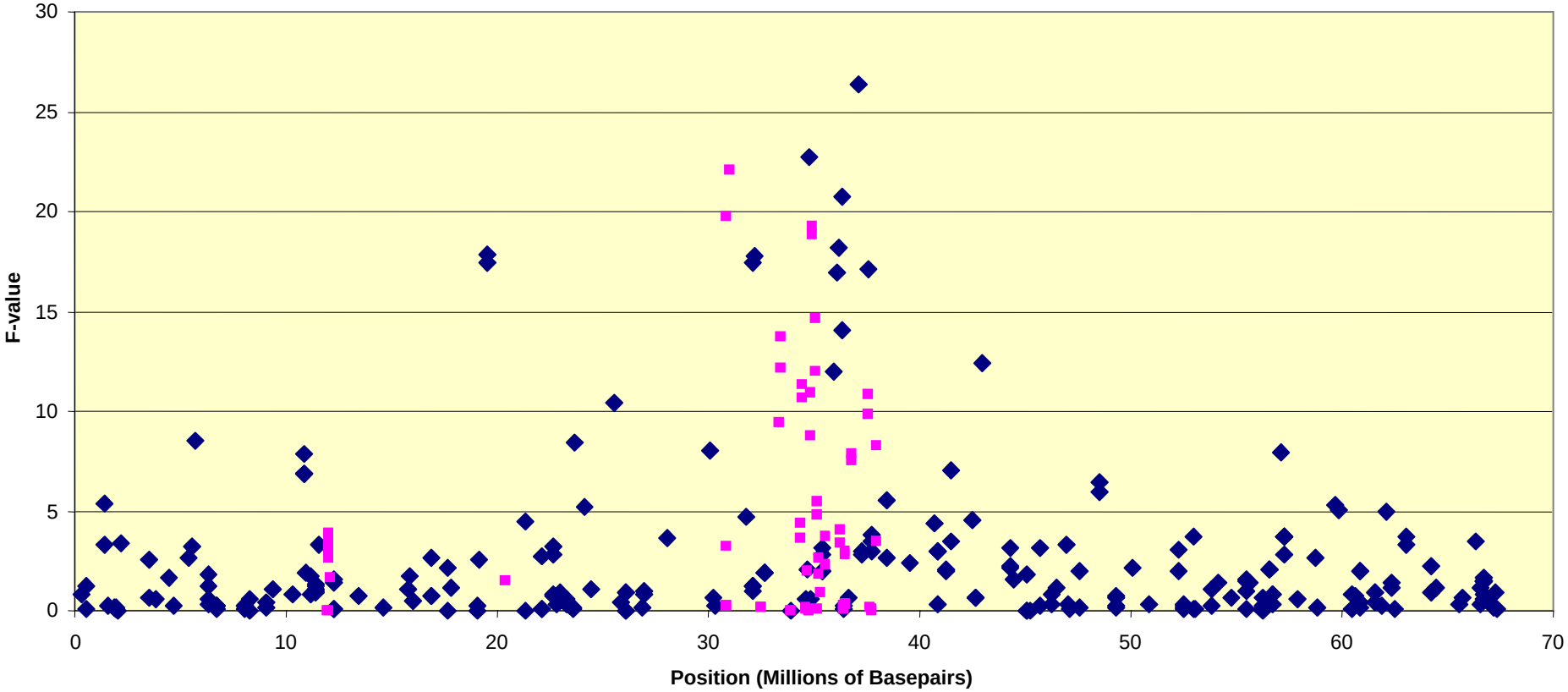
Single marker regression



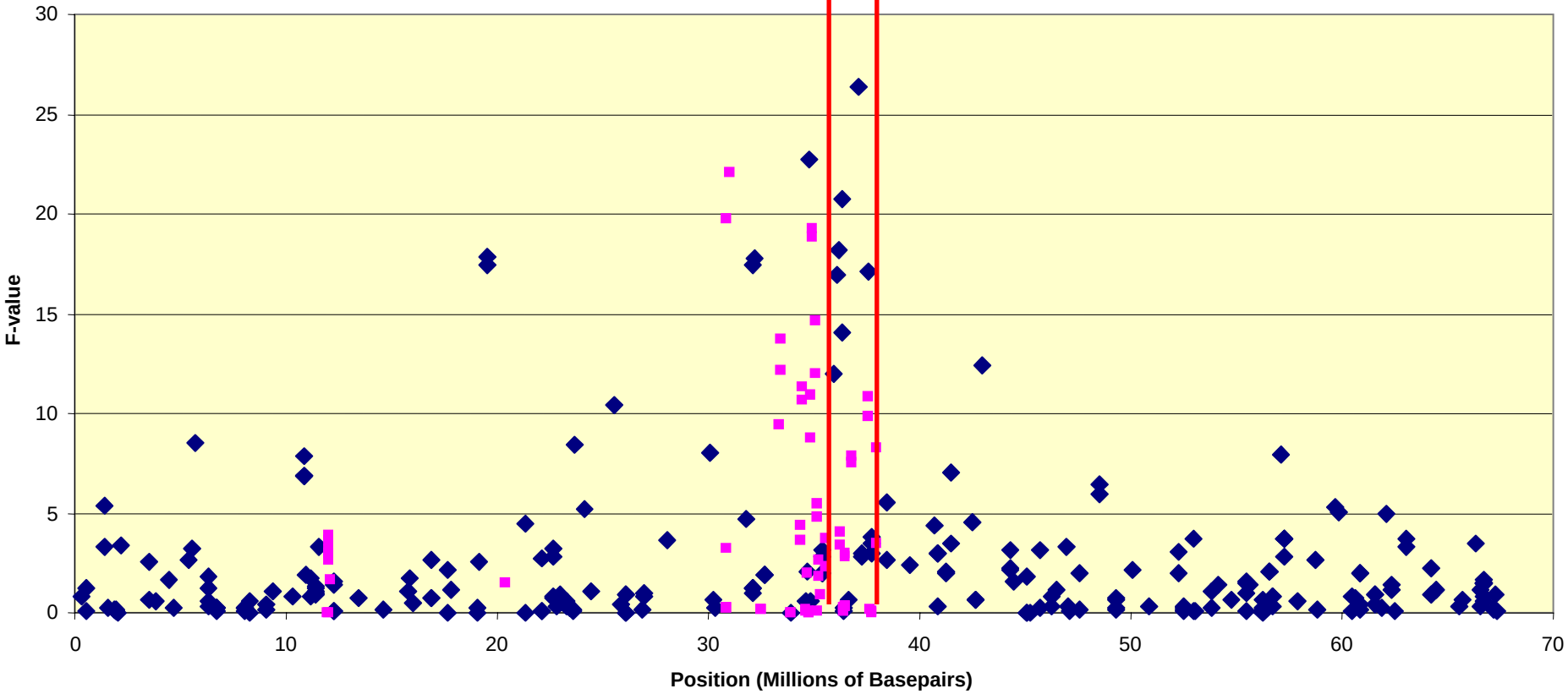


Single marker regression

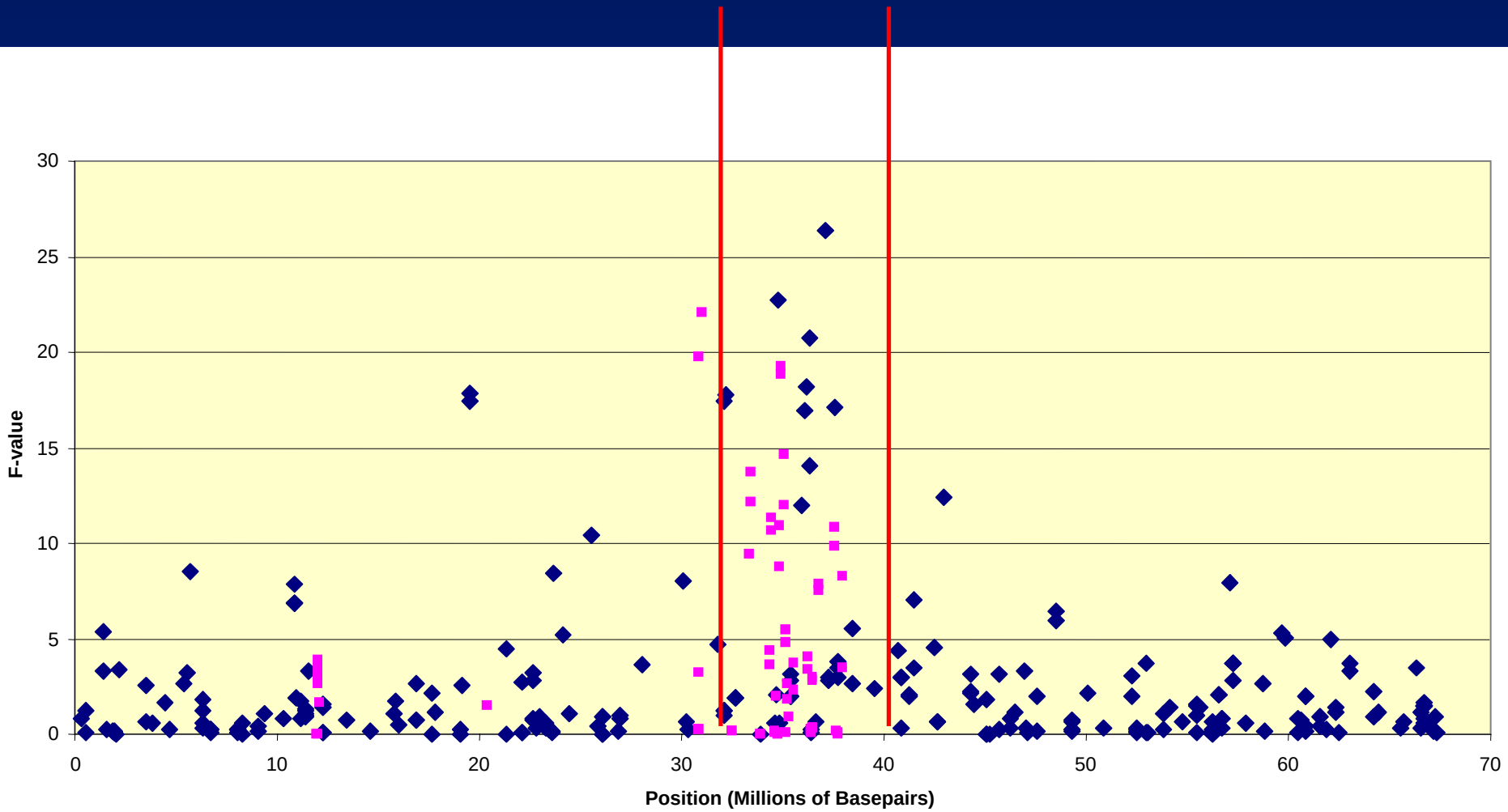
- Confidence regions
  - Following a genome wide association study, how do we decide the 95% confidence interval for the true QTL location?
- How many candidate genes to investigate?


30
25
20
15
10
5
0
F-value
0
10
20
30
40
50
60
70
Position (Millions of Basepairs)

??


30
25
20
15
10
5
0
F-value
0
10
20
30
40
50
60
70
Position (Millions of Basepairs)

??


30
25
20
15
10
5
0
F-value
0
10
20
30
40
50
60
70
Position (Millions of Basepairs)

Confidence intervals

- One method to calculate confidence intervals
  - Count up number of "clusters"=*n*
  - Split data set into two at random (eg. half animals in one set, other half in other set)
  - Designate best SNP at a cluster location in data set 1 and data set 2 as x_{1i},x_{2i}.
  - Estimate standard error of position over best SNP as:

$$se(\bar{x}) = \sqrt{\frac{1}{4n}\sum_{i=1}^{n}\left(x_{1i} - x_{2i}\right)^2}$$

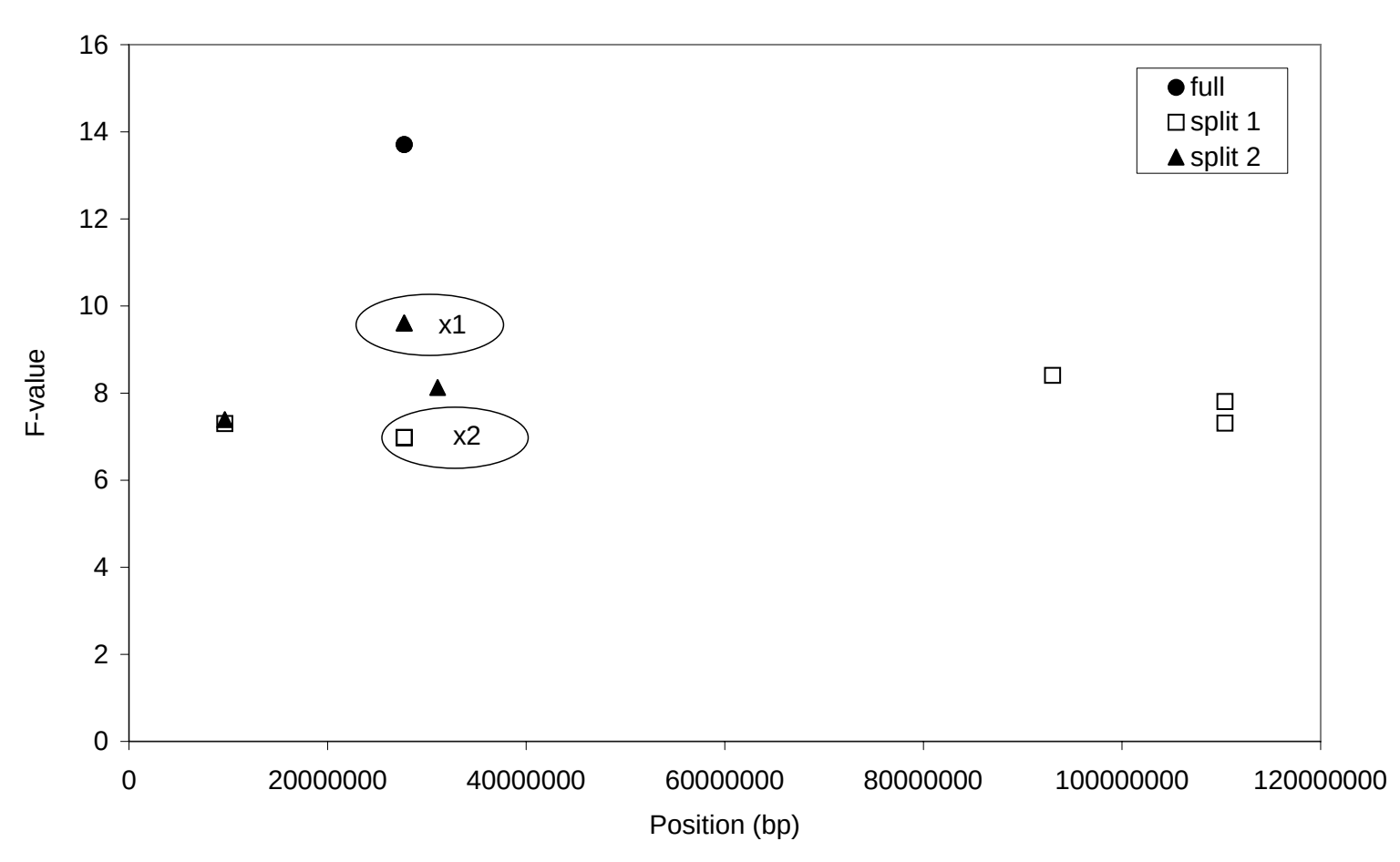
Confidence intervals

- One method to calculate confidence intervals
  - Count up number of "clusters"=*n*
  - Split data set into two at random (eg. half animals in one set, other half in other set)
  - Designate best SNP at a cluster location in data set 1 and data set 2 as x_{1i},x_{2i}.
  - Estimate standard error of position over best SNP as:

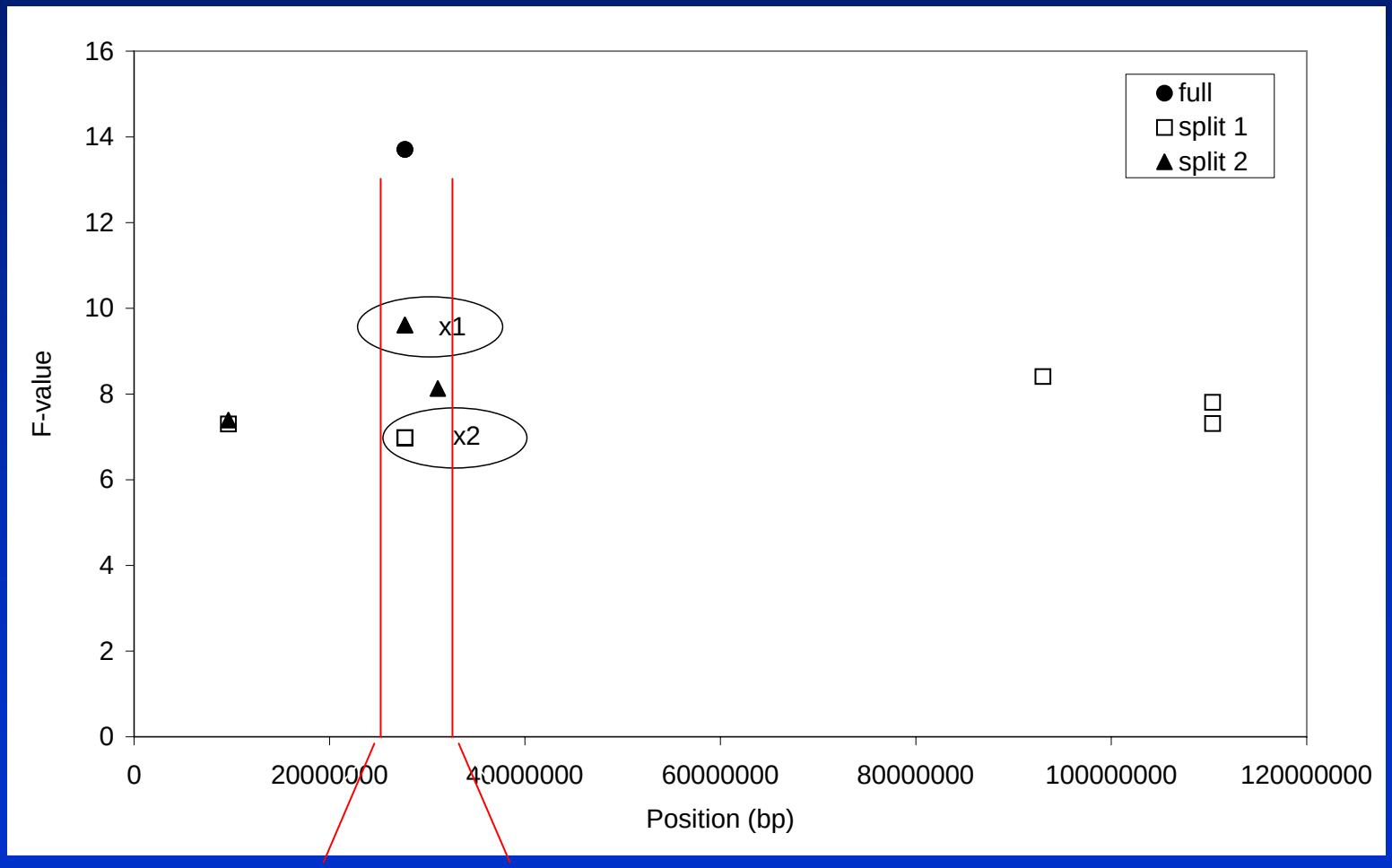
$$se(\bar{x}) = \sqrt{\frac{1}{4n}\sum_{i=1}^{n}\left(x_{1i} - x_{2i}\right)^2}$$

- 95% C.I = position of best SNP ±1.96*se(x_bar)

Confidence intervals



Confidence intervals



95% C.I

Mapping QTL using LD

- Association testing with single marker regression
- Accounting for population structure
- LD mapping with haplotypes
- The identical by descent (IBD) approach
- Combined linkage-linkage disequilibrium mapping

Population structure

- Simple model we have used assumes all animals are equally (un) related.
- Unlikely to be the case.
- Multiple offspring per sire, breeds or strains all create population structure.
- If we don’t account for this, false positives!

Population structure

- Simple example
  - a sire has many progeny in the population.
  - the sire has a high estimated breeding value
  - a rare allele at a random marker is homozygous in the sire (*aa*)
  - Then sub-population of his progeny have higher frequency of *a* than the rest of the population.
  - As the sires' estimated breeding value is high, his progeny will also have higher than average estimated breeding values.
  - If we don't account for relationship between progeny and sire the rare allele will appear to have a (perhaps significant) positive effect.

Population structure

- Can account for these relationships by extending our model…..

$$\mathbf{y} = \mathbf{1_n}'\mu + \mathbf{X}g + \mathbf{Zu} + \mathbf{e}$$

- Where
  - **u** is a vector of polygenic effect in the model with a covariance structure $u \sim N(0, A\sigma_a^2)$
  - **A** is the average relationship matrix built from the pedigree of the population
  - **Z** is a design matrix allocating animals to records.

Population structure

- Can account for these relationships by extending our model…..

$$\mathbf{y} = \mathbf{1_n}'\mu + \mathbf{X}g + \mathbf{Z}\mathbf{u} + \mathbf{e}$$

- Solutions ($\lambda = \sigma_e^2/\sigma_a^2$):

$$\begin{bmatrix} \hat{\mu} \\ \hat{\mathbf{g}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{1_n}'\mathbf{1_n} & \mathbf{1_n}'\mathbf{X} & \mathbf{1_n}'\mathbf{Z} \\ \mathbf{X}'\mathbf{1_n} & \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{1_n} & \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \mathbf{A}^{-1}\lambda \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{1_n}'\mathbf{y} \\ \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{bmatrix}$$

- An example A matrix……..

Pedigree

Animal	Sire	Dam
1	0	0
2	0	0
3	0	0
4	1	2
5	1	2
6	1	3

- An example A matrix……..

Pedigree

Animal	Sire	Dam
1	0	0
2	0	0
3	0	0
4	1	2
5	1	2
6	1	3

	Animal 1	Animal 2	Animal 3	Animal 4	Animal 5	Animal 6
Animal 1	1					
Animal 2						
Animal 3						
Animal 4						
Animal 5						
Animal 6						

- An example A matrix……..

Pedigree

Animal	Sire	Dam
1	0	0
2	0	0
3	0	0
4	1	2
5	1	2
6	1	3

	Animal 1	Animal 2	Animal 3	Animal 4	Animal 5	Animal 6
Animal 1	1					
Animal 2	0	1				
Animal 3						
Animal 4						
Animal 5						
Animal 6						

- An example A matrix……..

Pedigree

Animal	Sire	Dam
1	0	0
2	0	0
3	0	0
4	1	2
5	1	2
6	1	3

	Animal 1	Animal 2	Animal 3	Animal 4	Animal 5	Animal 6
Animal 1	1					
Animal 2	0	1				
Animal 3	0	0	1			
Animal 4						
Animal 5						
Animal 6						

- An example A matrix……..

Pedigree

Animal	Sire	Dam
1	0	0
2	0	0
3	0	0
4	1	2
5	1	2
6	1	3

Half genes from mum, half from dad

	Animal 1	Animal 2	Animal 3	Animal 4	Animal 5	Animal 6
Animal 1	1					
Animal 2	0	1				
Animal 3	0	0	1			
Animal 4	0.5	0.5	0	1		
Animal 5						
Animal 6						

- An example A matrix……..

Pedigree

Animal	Sire	Dam
1	0	0
2	0	0
3	0	0
4	1	2
5	1	2
6	1	3

	Animal 1	Animal 2	Animal 3	Animal 4	Animal 5	Animal 6
Animal 1	1					
Animal 2	0	1				
Animal 3	0	0	1			
Animal 4	0.5	0.5	0	1		
Animal 5	0.5	0.5	0	0.5	1	
Animal 6						

- An example A matrix……..

Pedigree

Animal	Sire	Dam
1	0	0
2	0	0
3	0	0
4	1	2
5	1	2
6	1	3

Animals 4 and 5 are full sibs

	Animal 1	Animal 2	Animal 3	Animal 4	Animal 5	Animal 6
Animal 1	1					
Animal 2	0	1				
Animal 3	0	0	1			
Animal 4	0.5	0.5	0	1		
Animal 5	0.5	0.5	0	0.5	1	
Animal 6						

- An example A matrix……..

Pedigree

Animal	Sire	Dam
1	0	0
2	0	0
3	0	0
4	1	2
5	1	2
6	1	3

Animals 6 is a half sib of 4 and 5

	Animal 1	Animal 2	Animal 3	Animal 4	Animal 5	Animal 6
Animal 1	1					
Animal 2	0	1				
Animal 3	0	0	1			
Animal 4	0.5	0.5	0	1		
Animal 5	0.5	0.5	0	0.5	1	
Animal 6	0.5	0	0.5	0.25	0.25	1

Population structure

- Example

Animal	Sire	Dam	Phenotype	SNP allele	SNP allele
1	0	0	10.1	1	2
2	0	0	2.2	2	2
3	0	0	2.31	2	2
4	1	2	6.57	1	2
5	1	2	6.06	1	2
6	1	3	6.21	1	2

$$\mathbf{y} = \mathbf{1_n}\boldsymbol{\mu} + \mathbf{X}g + \mathbf{e}$$

Population structure

- Example

Animal	Sire	Dam	Phenotype	SNP allele	SNP allele
1	0	0	10.1	1	2
2	0	0	2.2	2	2
3	0	0	2.31	2	2
4	1	2	6.57	1	2
5	1	2	6.06	1	2
6	1	3	6.21	1	2

$$\mathbf{y} = \mathbf{1_n}\boldsymbol{\mu} + \mathbf{X}g + \mathbf{e}$$

X	1
	2
	2
	1
	1
	1

Population structure

- Example

Animal	Sire	Dam	Phenotype	SNP allele	SNP allele
1	0	0	10.1	1	2
2	0	0	2.2	2	2
3	0	0	2.31	2	2
4	1	2	6.57	1	2
5	1	2	6.06	1	2
6	1	3	6.21	1	2

$$\mathbf{y} = \mathbf{1_n}\boldsymbol{\mu} + \mathbf{X}g + \mathbf{e}$$

$$\begin{bmatrix} \hat{\mu} \\ \hat{g} \end{bmatrix} = \begin{bmatrix} \mathbf{1_n'1_n} & \mathbf{1_n'X} \\ \mathbf{X'1_n} & \mathbf{X'X} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{1_n'y} \\ \mathbf{X'y} \end{bmatrix}$$

Population structure

- Example

Animal	Sire	Dam	Phenotype	SNP allele	SNP allele
1	0	0	10.1	1	2
2	0	0	2.2	2	2
3	0	0	2.31	2	2
4	1	2	6.57	1	2
5	1	2	6.06	1	2
6	1	3	6.21	1	2

$$\mathbf{y} = \mathbf{1_n}\boldsymbol{\mu} + \mathbf{X}g + \mathbf{e}$$

$$\begin{bmatrix} \hat{\mu} \\ \hat{g} \end{bmatrix} = \begin{bmatrix} 6 & 8 \\ 8 & 12 \end{bmatrix}^{-1} \begin{bmatrix} 33.5 \\ 38 \end{bmatrix}$$

Population structure

- Example

Animal	Sire	Dam	Phenotype	SNP allele	SNP allele
1	0	0	10.1	1	2
2	0	0	2.2	2	2
3	0	0	2.31	2	2
4	1	2	6.57	1	2
5	1	2	6.06	1	2
6	1	3	6.21	1	2

$$\begin{bmatrix} \hat{\mu} \\ \hat{g} \end{bmatrix} = \begin{bmatrix} 12.2 \\ -5 \end{bmatrix}$$

Population structure

- Example

Animal	Sire	Dam	Phenotype	SNP allele	SNP allele
1	0	0	10.1	1	2
2	0	0	2.2	2	2
3	0	0	2.31	2	2
4	1	2	6.57	1	2
5	1	2	6.06	1	2
6	1	3	6.21	1	2

$$\mathbf{y} = \mathbf{1_n}'\mu + \mathbf{X}g + \mathbf{Z}\mathbf{u} + \mathbf{e}$$

Population structure

- Example

Animal	Sire	Dam	Phenotype	SNP allele	SNP allele
1	0	0	10.1	1	2
2	0	0	2.2	2	2
3	0	0	2.31	2	2
4	1	2	6.57	1	2
5	1	2	6.06	1	2
6	1	3	6.21	1	2

$$\mathbf{y} = \mathbf{1_n}'\mu + \mathbf{X}g + \mathbf{Z}\mathbf{u} + \mathbf{e}$$

$$\begin{bmatrix} \hat{\mu} \\ \hat{\mathbf{g}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{1_n}'\mathbf{1_n} & \mathbf{1_n}'\mathbf{X} & \mathbf{1_n}'\mathbf{Z} \\ \mathbf{X}'\mathbf{1_n} & \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{1_n} & \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \mathbf{A}^{-1}\lambda \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{1_n}'\mathbf{y} \\ \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{bmatrix}$$

Population structure

- Example

Animal	Sire	Dam	Phenotype	SNP allele	SNP allele
1	0	0	6.51	1	1
2	0	0	2.2	2	2
3	0	0	2.31	2	2
4	1	2	4.72	1	2
5	1	2	5.02	1	2
6	3	2	2.93	2	2

$$\mathbf{y} = \mathbf{1_n}'\mu + \mathbf{X}g + \mathbf{Z}\mathbf{u} + \mathbf{e}$$

λ=0.33

Population structure

- Example

Animal	Sire	Dam	Phenotype	SNP allele	SNP allele
1	0	0	6.51	1	1
2	0	0	2.2	2	2
3	0	0	2.31	2	2
4	1	2	4.72	1	2
5	1	2	5.02	1	2
6	3	2	2.93	2	2

$$\begin{bmatrix} \hat{\mu} \\ \hat{g} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} 6 & 8 & 1 & 1 & 1 & 1 & 1 & 1 \\ 8 & 12 & 1 & 2 & 2 & 1 & 1 & 1 \\ 1 & 1 & 1.916575 & 0.416625 & 0.16665 & -0.3333 & -0.49995 & -0.3333 \\ 1 & 2 & 0.416625 & 1.749925 & 0 & -0.3333 & -0.49995 & 0 \\ 1 & 2 & 0.16665 & 0 & 1.49995 & 0 & 0 & -0.3333 \\ 1 & 1 & -0.49995 & -0.49995 & 0 & 1.6666 & 0.3333 & 0 \\ 1 & 1 & -0.3333 & -0.3333 & 0 & 0 & 1.6666 & 0 \\ 1 & 1 & -0.3333 & 0 & -0.3333 & 0 & 0 & 1.6666 \end{bmatrix}^{-1} \begin{matrix} 33.45 \\ 37.96 \\ 10.1 \\ 2.2 \\ 2.31 \\ 6.57 \\ 6.06 \\ 6.21 \end{matrix}$$

Population structure

- Example

Animal	Sire	Dam	Phenotype	SNP allele	SNP allele
1	0	0	10.1	1	2
2	0	0	2.2	2	2
3	0	0	2.31	2	2
4	1	2	6.57	1	2
5	1	2	6.06	1	2
6	1	3	6.21	1	2

$$\begin{bmatrix} \hat{\mu} \\ \hat{g} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} 10.3 \\ -3.5 \\ 1.9 \\ -1.1 \\ -0.9 \\ 0.2 \\ -0.3 \\ -0.1 \end{bmatrix}$$

Population structure

- Example of importance of accounting for population structure…….
  - 365 Angus cattle genotyped for 10,000 SNPs
  - polygenic and environmental effects were simulated for each animal
  - *No QTL fitted!*
  - Effect of each SNP tested using three models
    - SNP only
    - SNP and sire
    - SNP and full pedigree

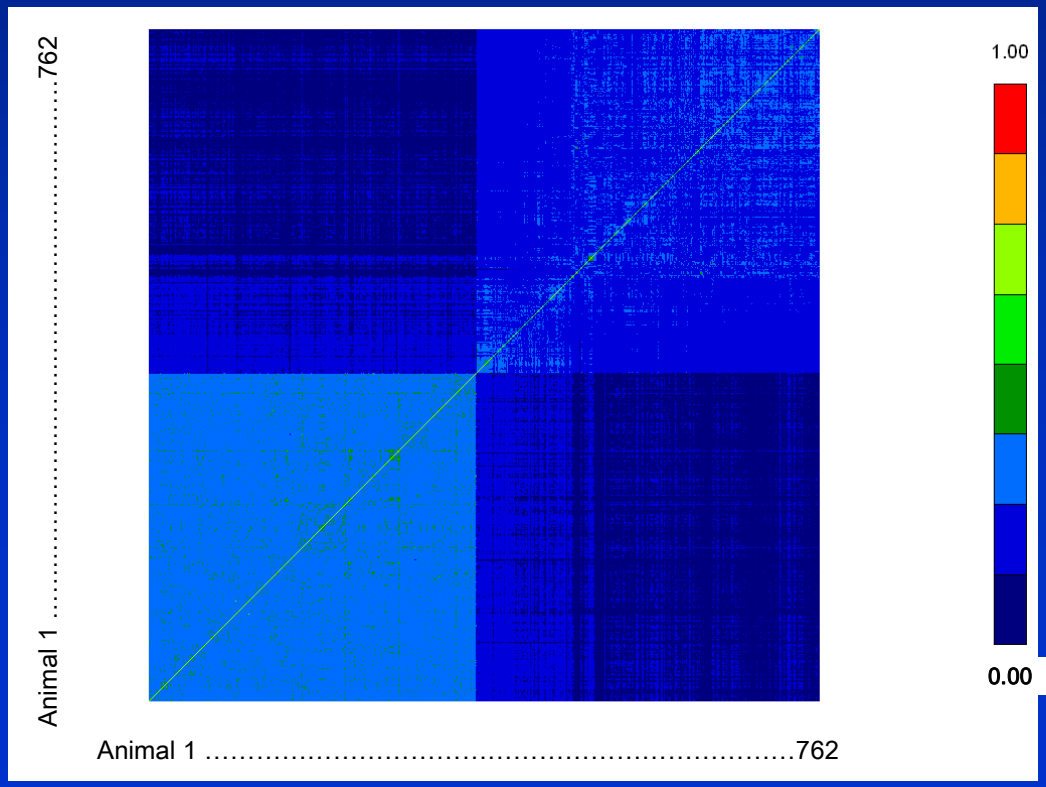
Population structure

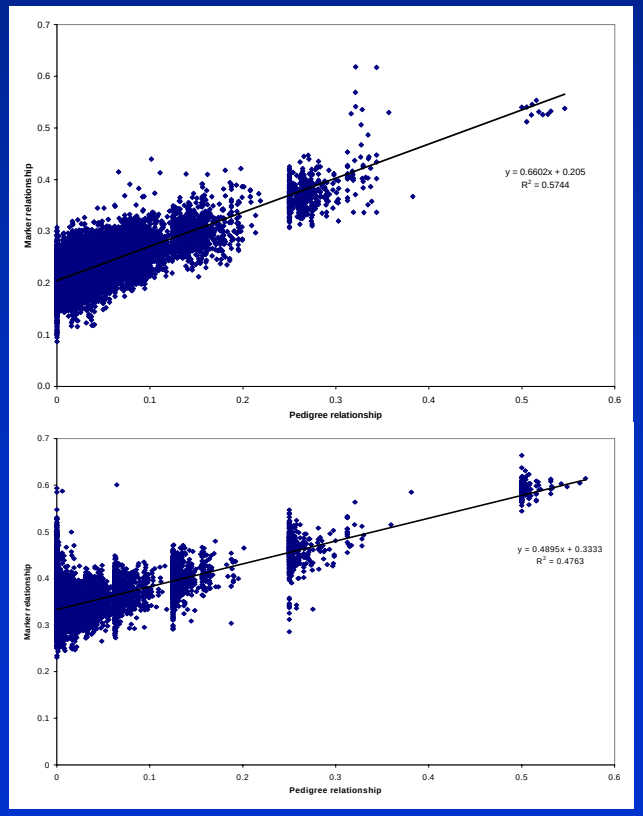
Number of false positives……

Analysis model	Significance level		
	$p<0.005$	$p<0.001$	$p<0.0005$
Expected type I errors	40	8	4
1. Full pedigree model	39 (SD=14)	9 (SD=5)	4 (SD=3)
2. Sire pedigree model	46^{*} (SD=21)	11^{*} (SD=7)	6^{*} (SD=5.5)
3. No pedigree model	68^{**} (SD=31)	18^{**} (SD=11)	10^{**} (SD=7)
4. Selected 27% - full pedigree	54^{**} (SD=18)	12^{**} (SD=6)	7^{**} (SD=4)

Population structure

- Problem when we do not have history of the population
- Solution – use the average relationship across all markers as the **A** matrix





Mapping QTL using LD

- Association testing with single marker regression
- Accounting for population structure
- LD mapping with haplotypes
- The identical by descent (IBD) approach
- Combined linkage-linkage disequilibrium mapping

LD mapping with haplotypes

- Power of association study depends on LD between markers and QTL
- One way to increase LD between QTL alleles and markers is to use *haplotypes* of markers rather than a single marker

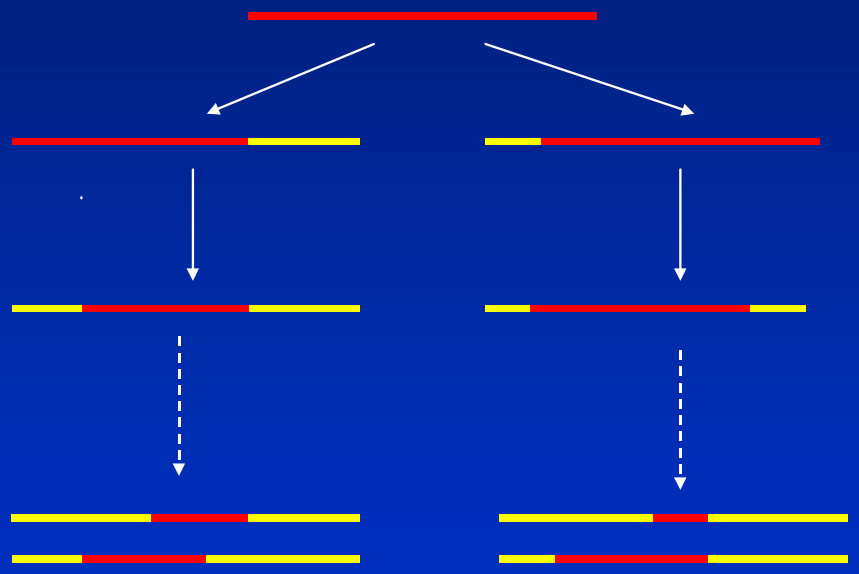
- 1_Q single marker (1 is the allele of the marker)
- 1_1_Q_2_1 Haplotype of markers

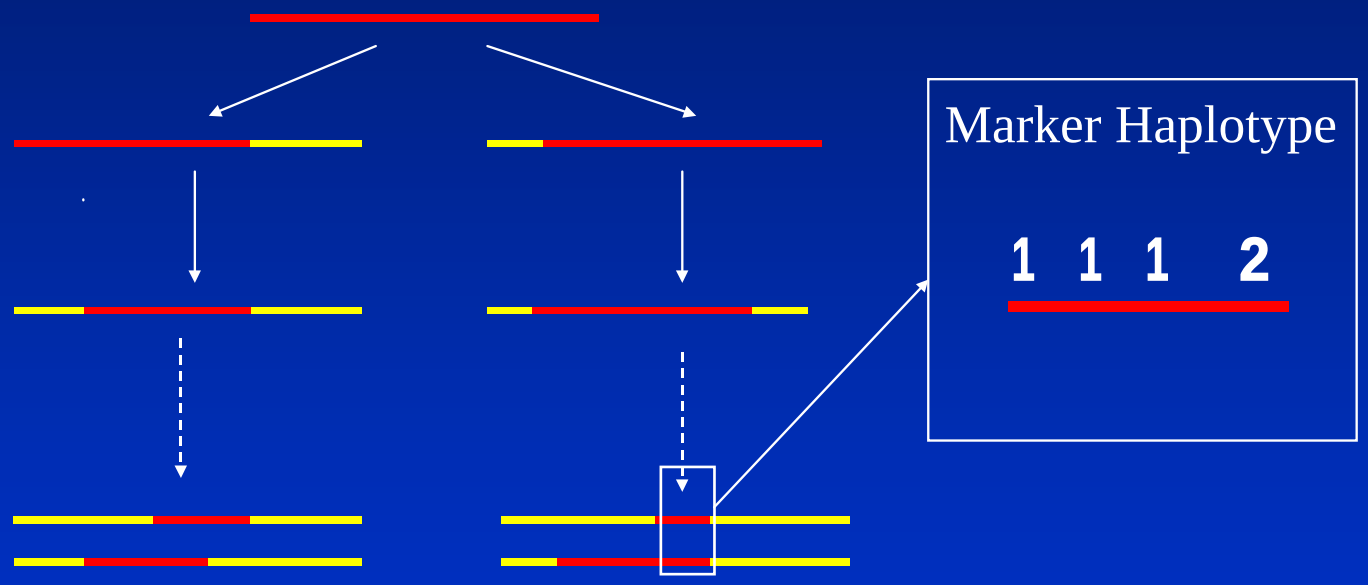
LD mapping with haplotypes

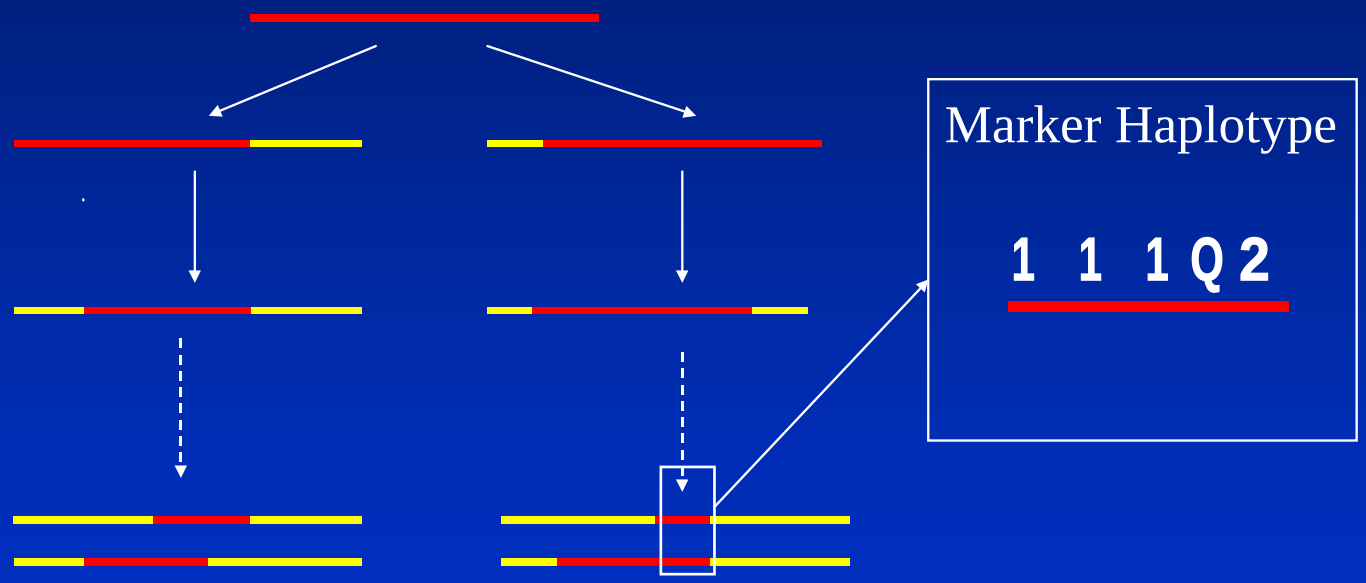
- Value of haplotypes depends on LD between haplotype and QTL
  - If we find two identical haplotypes from the population, what is the probability they carry the same QTL allele?
  - If probability is high, high level of LD between haplotype and QTL

LD mapping with haplotypes

- If we find two identical haplotypes from the population, what is the probability they carry the same QTL allele?
- Haplotypes identical either because chromosome segments from same common ancestor




Marker Haplotype
1 1 1 2


Marker Haplotype
1 1 1 Q 2

LD mapping with haplotypes

- If we find two identical haplotypes from the population, what is the probability they carry the same QTL allele?
- Haplotypes identical either because chromosome segments from same common ancestor
- Or because of chance recombination…….

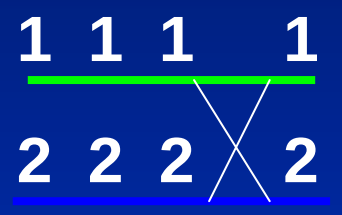
Chance recombination produces the same haplotype…..

1 1 1 1

2 2 2 2

Sire

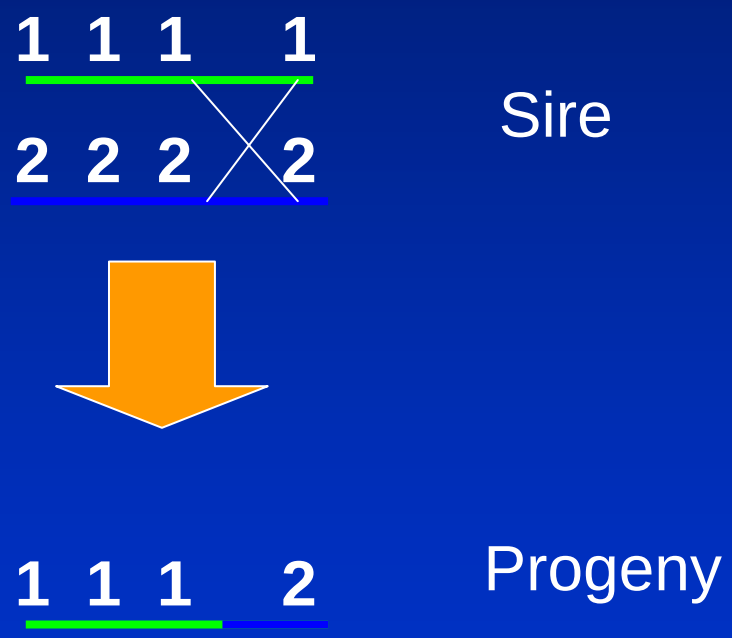
Chance recombination produces the same haplotype…..



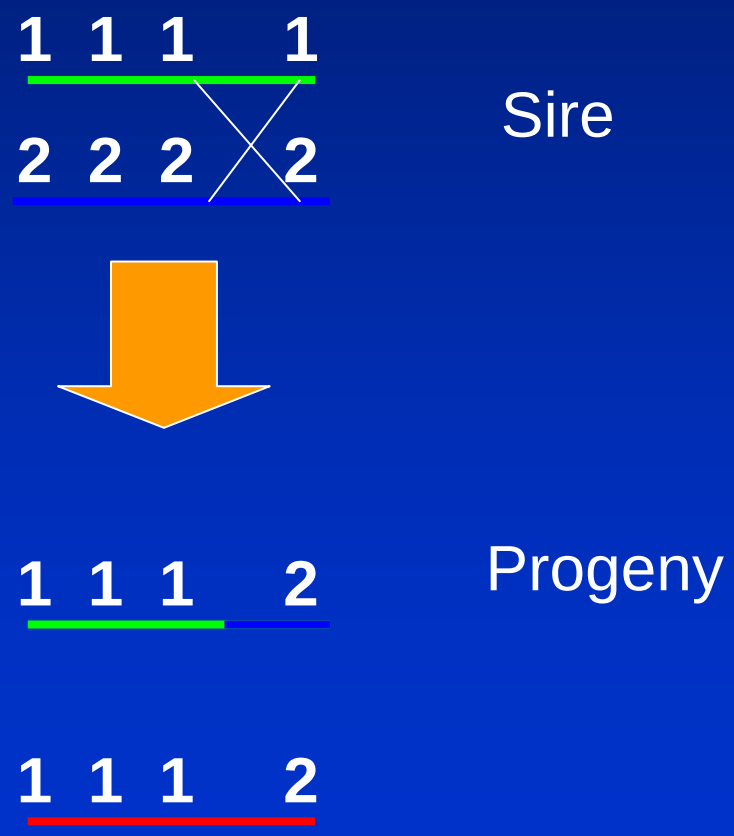
Sire

Formation of gamete

Chance recombination produces the same haplotype…..



Chance recombination produces the same haplotype…..



Chance recombination produces the same haplotype…..

1 1 1 q 1

2 2 2 q 2

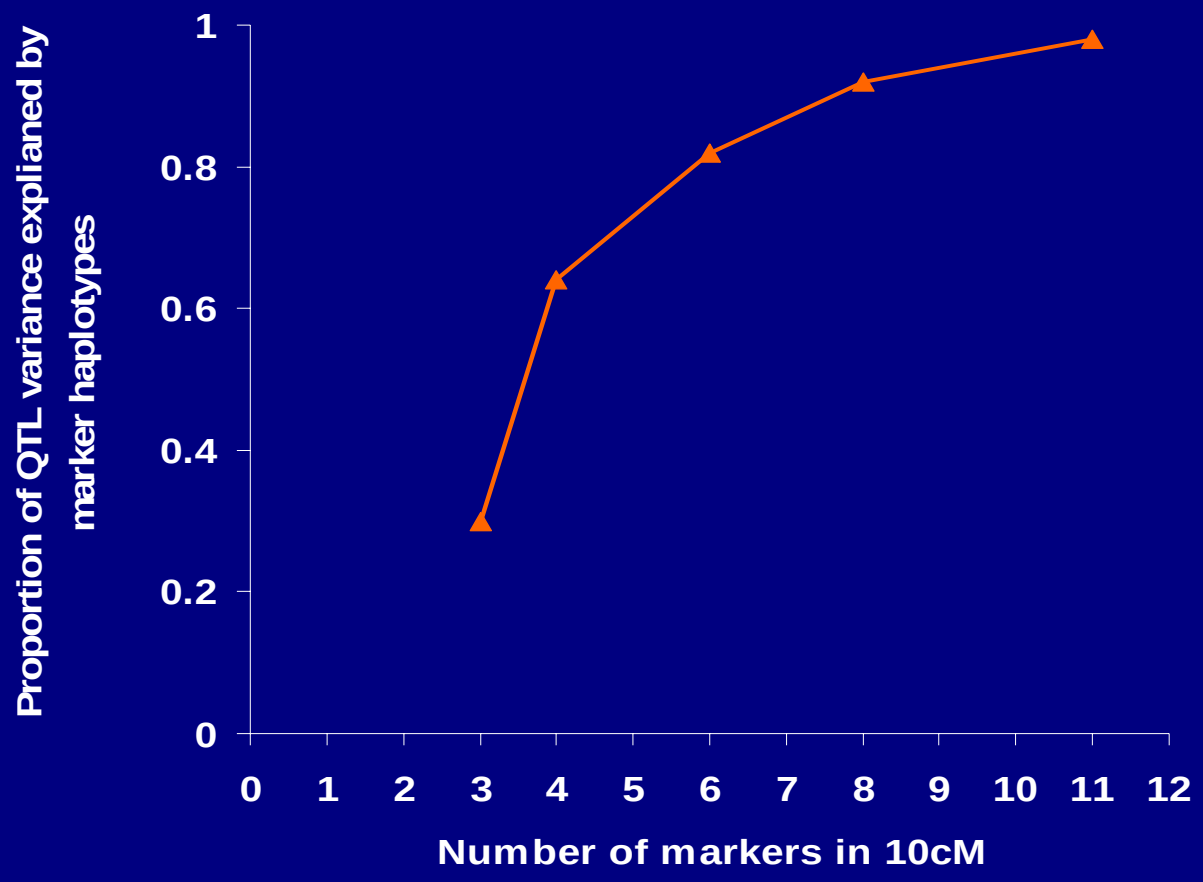
Sire

1 1 1 q 2

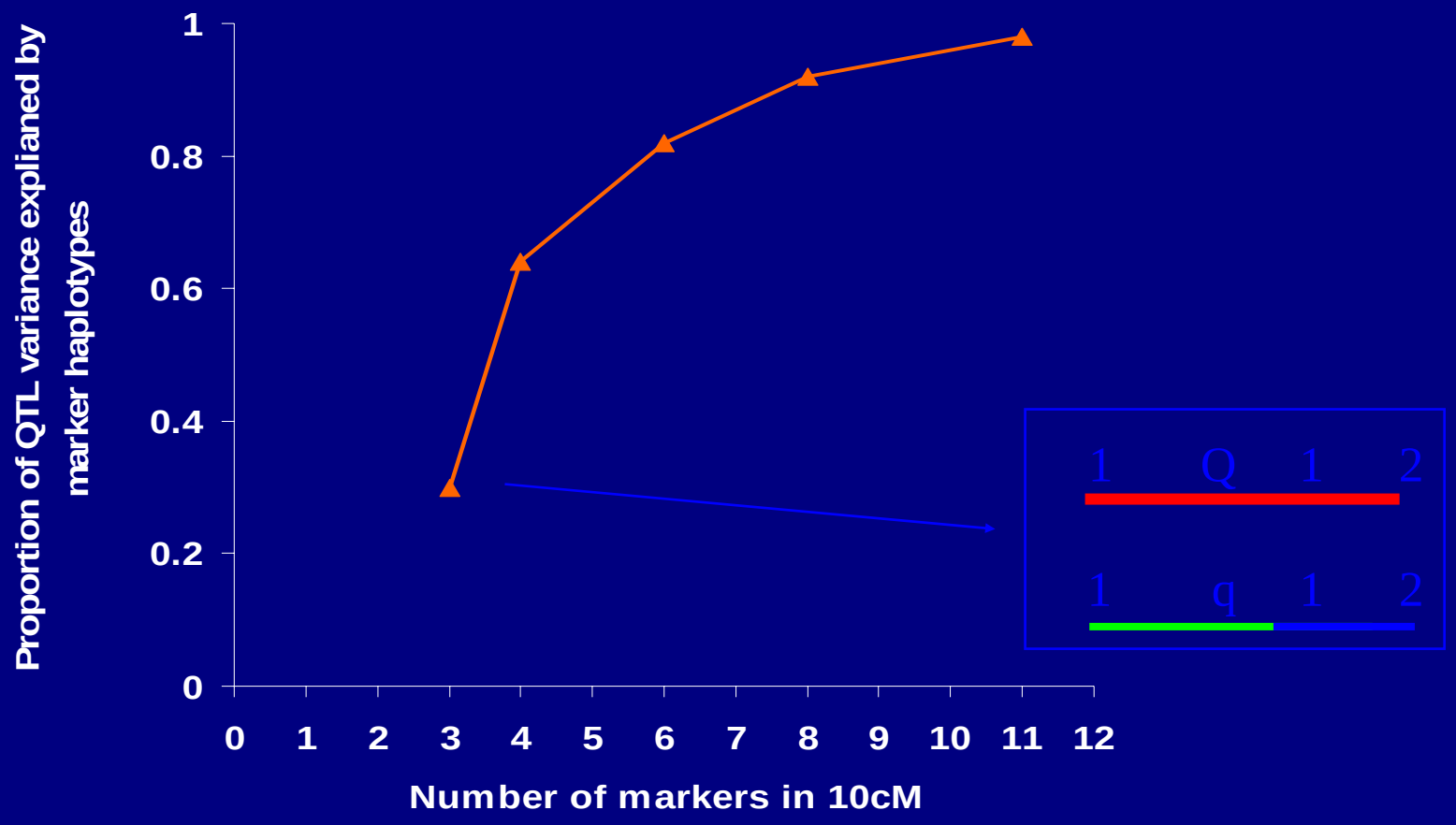
Progeny

1 1 1 Q 2

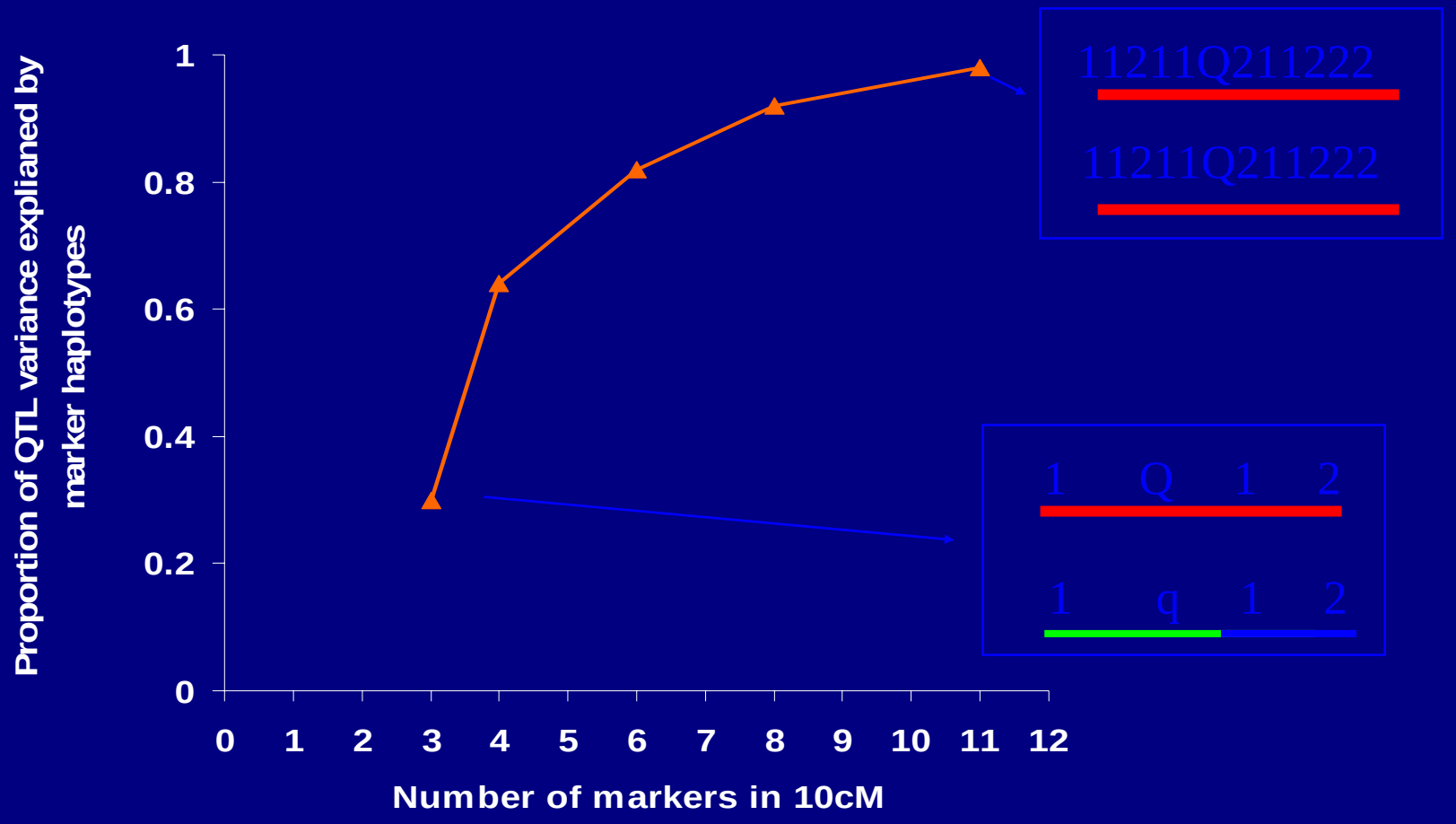
Proportion of QTL variance explained by surrounding markers



Proportion of QTL variance explained by surrounding markers



Proportion of QTL variance explained by surrounding markers



LD mapping with haplotypes

- If we find two identical haplotypes from the population, what is the probability they carry the same QTL allele?
- Haplotypes identical either because chromosome segments from same common ancestor
- Or because of chance recombination…….
- With more markers in haplotype, the chance of creating the same haplotype by recombination becomes small

LD mapping with haplotypes

- Model ?

$$\mathbf{y} = \mathbf{1_n}'\mu + \mathbf{Xg} + \mathbf{Zu} + \mathbf{e}$$

- Where **g** is now a vector of haplotype effects dimensions (number of haplotypes observed x 1)
- And **X** allocates records to haplotyes

LD mapping with haplotypes

- Example (eg after using PHASE to infer haplotype)

Animal	Paternal haplotype	Maternal haplotype
1	1	1
2	1	2
3	2	3
4	5	4
5	3	2

- **X**

LD mapping with haplotypes

- Example (eg after using PHASE to infer haplotype)

Animal	Paternal haplotype	Maternal haplotype
1	1	1
2	1	2
3	2	3
4	5	4
5	3	2

- **X**

		Haplotype				
		1	2	3	4	5
	1	**2**	**0**	**0**	**0**	**0**
	2	**1**	**1**	**0**	**0**	**0**
Animal	3	**0**	**1**	**1**	**0**	**0**
	4	**0**	**0**	**0**	**1**	**1**
	5	**0**	**1**	**1**	**0**	**0**

LD mapping with haplotypes

- Fit haplotypes as random effects
  - $\mathbf{g} \sim N(0,\sigma_h^2)$
  - Some haplotypes will be rare, very few observations
  - Fitting the haplotype effect as random regresses the effects back to account for the lack of information

$$\begin{bmatrix} \hat{\mu} \\ \hat{\mathbf{g}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{1_n'1_n} & \mathbf{1_n'X} & \mathbf{1_n'Z} \\ \mathbf{X'1_n} & \mathbf{X'X} & \mathbf{X'Z} \\ \mathbf{Z'1_n} & \mathbf{Z'X} & \mathbf{Z'Z} + \mathbf{A}^{-1}\lambda \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{1_n'y} \\ \mathbf{X'y} \\ \mathbf{Z'y} \end{bmatrix}$$

LD mapping with haplotypes

- Fit haplotypes as random effects
  - $\mathbf{g} \sim N(0,\sigma_h^2)$
  - Some haplotypes will be rare, very few observations
  - Fitting the haplotype effect as random regresses the effects back to account for the lack of information
  - $\lambda_h=\sigma_e^2/\sigma_h^2$

$$\begin{bmatrix} \hat{\mu} \\ \hat{\mathbf{g}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{1_n'1_n} & \mathbf{1_n'X} & \mathbf{1_n'Z} \\ \mathbf{X'1_n} & \mathbf{X'X}+\mathbf{I}\lambda_{QTL} & \mathbf{X'Z} \\ \mathbf{Z'1_n} & \mathbf{Z'X} & \mathbf{Z'Z}+\mathbf{A}^{-1}\lambda \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{1_n'y} \\ \mathbf{X'y} \\ \mathbf{Z'y} \end{bmatrix}$$

LD mapping with haplotypes

- There is a "cost" of using haplotypes instead of single markers
- With single markers only one effect to estimate, with haplotypes many effects
- Fewer observations per effect, lower accuracy of estimating each effect

	Proportion of QTL variance explained	Maximum number of haplotypes	Observed number of haplotypes
Nearest marker	0.10	2	2
Best marker	0.20	2	2
2 Marker haplotypes	0.15	4	3.4
4 Marker haplotypes	0.28	16	9.4
6 Marker haplotypes	0.55	64	20.8

Mapping QTL using LD

- Association testing with single marker regression
- Accounting for population structure
- LD mapping with haplotypes
- The identical by descent (IBD) approach
- Combined linkage-linkage disequilibrium mapping

The IBD approach

- Principle:
  - Existence of LD implies small segments of chromosome in population which are descended from the same common ancestor (IBD).

The IBD approach

- Principle:
  - Existence of LD implies small segments of chromosome in population which are descended from the same common ancestor (IBD).
  - IBD chromosome segments will not only carry identical marker haplotypes; if there is a QTL within chromosome segment, IBD chromosome segments will also carry identical QTL alleles.

The IBD approach

- Principle:
  - Existence of LD implies small segments of chromosome in population which are descended from the same common ancestor (IBD).
  - IBD chromosome segments will not only carry identical marker haplotypes; if there is a QTL within chromosome segment, IBD chromosome segments will also carry identical QTL alleles.
  - If two animals carry chromosomes which are IBD at a QTL position, their phenotypes will be correlated.

The IBD approach

- The model

$$y_i = \mu + u_i + vp_i + vm_i + e_i$$

- Where vp_i and vm_i are the effects of the paternal and maternal QTL alleles respectively
- modelling the effect of the QTL directly rather than assuming a haplotype or marker is in LD with the QTL

The IBD approach

- Each animal has it's own QTL alleles
- There is a probability that different QTL alleles are actually IBD
- This is captured in the IBD (**G**) matrix
- Elements g_{ij} is the probability that QTL allele i and j are IBD.
- This probability is inferred from marker haplotypes
- Dimensions (2*number of animals * 2*number of animals)
- $u \sim (0, \mathbf{A}\sigma_a^2)$, $v \sim (0, \mathbf{G}\sigma_v^2)$, $e \sim \sim (0, \mathbf{I}\sigma_e^2)$

The IBD approach

- Building **IBD matrix** from marker haplotypes
  - Consider three haplotypes drawn from population at random (P is putative QTL position)
    - A 112P112
    - B 212P112
    - C 222P222
  - P(IBD at QTL A,B) >P(IBD at QTL B,C), as longer identical haplotype

The IBD approach

- Building IBD matrix from marker haplotypes
  - Parameters which determine IBD coefficients are
    - extent of LD
    - length of haplotype and
    - number of markers in the haplotype

The IBD approach

- Building IBD matrix from marker haplotypes
- Algorithm of Meuwissen and Goddard (2001)
  - deterministically predicts IBD coefficients at putative QTL positions from marker haplotypes

The IBD approach

- Building IBD matrix from marker haplotypes
- Algorithm of Meuwissen and Goddard (2001)
  - deterministically predicts IBD coefficients between two marker haplotypes using
    - number of markers flanking QTL position which are identical by state
    - probability identical by chance ~ marker homozygosity
    - extent of LD based on length of haplotype, effective population size

The IBD approach

- Building IBD matrix from marker haplotypes
  - An example with Ne = 100
  - 6 markers in 10cM, putative QTL position in centre M_M_M_Q_M_M_M
  - Sample four haplotypes from the population
  - 112112, 112112, 122112, 222122
  - IBD matrix is:

	112112	112112	122112	222122
112112	1			
112112	0.82	1		
122112	0.63	0.63	1	
222122	0.49	0.49	0.56	1

The IBD approach

- A two stage approach for linkage disequilibrium mapping
  1. For each putative QTL position, **IBD** or **G** matrix. IBD matrix has elements g_{ij}=Prob(QTL alleles i and j are identical by descent or IBD)
  2. For each position considered in step 1, construct the linear model to estimate QTL variances and other parameters, test for presence of QTL

The IBD approach

- The model

$$y_i = \mu + u_i + vp_i + vm_i + e_i$$

- $u \sim (0, \mathbf{A}\sigma_a^2)$, $v \sim (0, \mathbf{G}\sigma_v^2)$, $e \sim \sim (0, \mathbf{I}\sigma_e^2)$

The IBD approach

- Use variance component estimation procedures to find the
  - Estimate of σ_u^2
  - Estimate of σ_v^2
  - Estimate of σ_e^2
  - Which maximise the Log likelihood (LogL) of the data given these parameters
  - Eg. ASREML

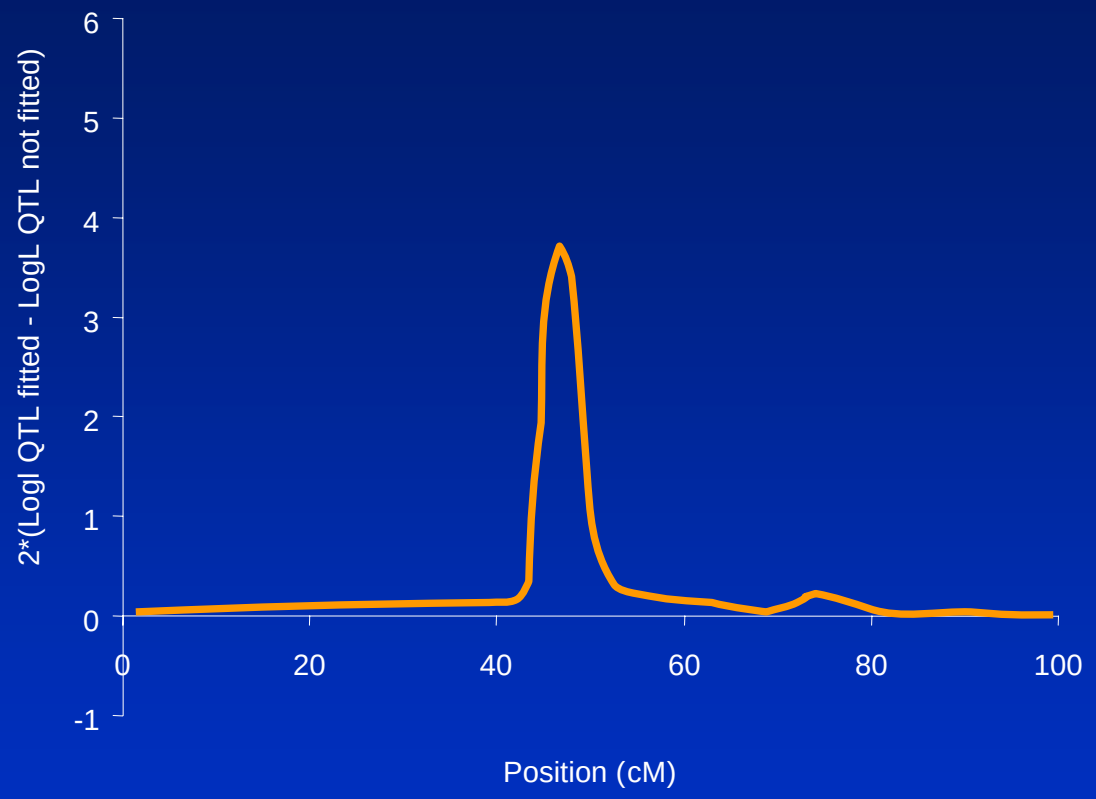
The IBD approach

- How do we test if the QTL is significant or not?
- Fit the model with no QTL:

$$y_i = \mu + u_i + e_i$$

- Plot -2*(LogL QTL fitted - LogL QTL not fitted) against position

The IBD approach



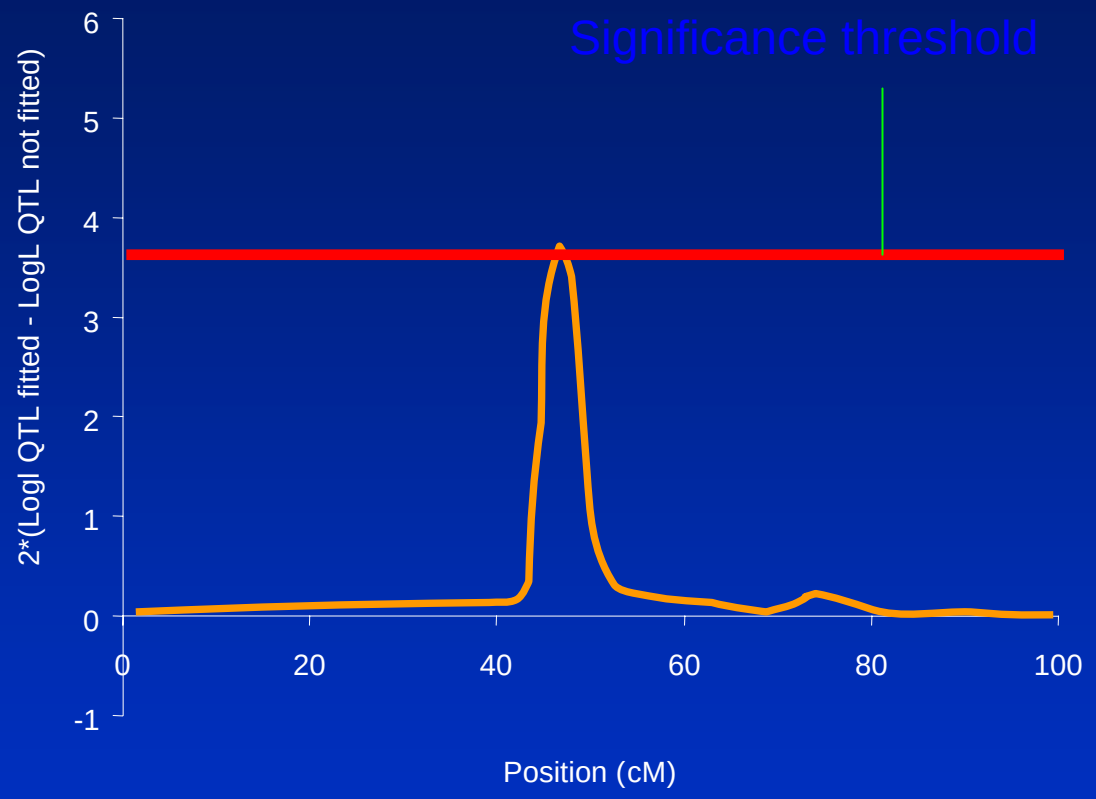
The IBD approach

- How do we test if the QTL is significant or not?
- Fit the model with no QTL:

$$y_i = \mu + u_i + e_i$$

- Plot -2*(LogL QTL fitted - LogL QTL not fitted) against position
- Is distributed as a $\chi^2_{1,\alpha}$ where α is the desired significance level
- at α=0.05 is 3.84)

Linkage mapping in complex pedigrees



The IBD approach

- Confidence interval
  - Drop of 2 of test statistic from maximum point

Comparison of approaches

- Zhao et al. (2007) compared power and precision of QTL mapping with single marker regression, haplotypes and IBD approach
- They found in simulated data that single marker regression and the IBD approach had similar power and precision
- Calus et al. (2007) found that haplotypes gave slightly greater accuracy than single markers, and that the IBD approach gave much higher accuracies at low marker densities
- Hayes et al. (2007) tried to use real data, and results indicated 6 marker haplotypes were better than single marker regression
- Level of LD, simulation assumptions??

Mapping QTL using LD

- Association testing with single marker regression
- Accounting for population structure
- LD mapping with haplotypes
- The Identical by descent (IBD) approach
- Combined linkage-linkage disequilibrium mapping

Combined LD-LA mapping

- Extent of LD very variable
- LD can exist between loci on different chromosomes!!
- Combine LD and linkage information to filter spurious peaks

Combined LD-LA mapping

- Consider a half sib design
  - LD information from sire haplotypes, maternal hapotypes of progeny
  - Linkage information from paternal haplotypes of progeny
- IBD matrix:

	SH	MHP	PHP
SH	[a]	[a]	[b]
MHP	[a]	[a]	[b]
PHP	[b]	[b]	[b]

  - a = LD (Meuwissen and Goddard 2001)
  - b = linkage

Combined LD-LA mapping

- In linkage analysis (LA) consider founder alleles (sires, dams) to be unrelated, eg…..

Sire 1211, 1212
Dam 1222, 1211
Progeny 1211, 1222

LA

		Sire		Dam		Progeny	
		Pat	Mat	Pat	Mat	Pat	Mat
Sire	Pat	1					
	Mat	0	1				
Dam	Pat	0	0	1			
	Mat	0	0	0	1		
Progeny	Pat	1	0	0	0	1	
	Mat	0	0	1	0	0	1

Sire 1211, 1212
Dam 1222, 1211
Progeny 1211, 1222

LA

		Sire		Dam		Progeny	
		Pat	Mat	Pat	Mat	Pat	Mat
Sire	Pat	1					
	Mat	0	1				
Dam	Pat	0	0	1			
	Mat	0	0	0	1		
Progeny	Pat	1	0	0	0	1	
	Mat	0	0	1	0	0	1

LD

		Sire		Dam	
		Pat	Mat	Pat	Mat
Sire	Pat	1			
	Mat	0.8	1		
Dam	Pat	0.5	0.5	1	
	Mat	0.9	0.5	0.5	1

Sire 1211, 1212
Dam 1222, 1211
Progeny 1211, 1222

LA

		Sire		Dam		Progeny	
		Pat	Mat	Pat	Mat	Pat	Mat
Sire	Pat	1					
	Mat	0	1				
Dam	Pat	0	0	1			
	Mat	0	0	0	1		
Progeny	Pat	1	0	0	0	1	
	Mat	0	0	1	0	0	1

LD

		Sire		Dam	
		Pat	Mat	Pat	Mat
Sire	Pat	1			
	Mat	0.8	1		
Dam	Pat	0.5	0.5	1	
	Mat	0.9	0.5	0.5	1

Sire 1211, 1212
Dam 1222, 1211
Progeny 1211, 1222

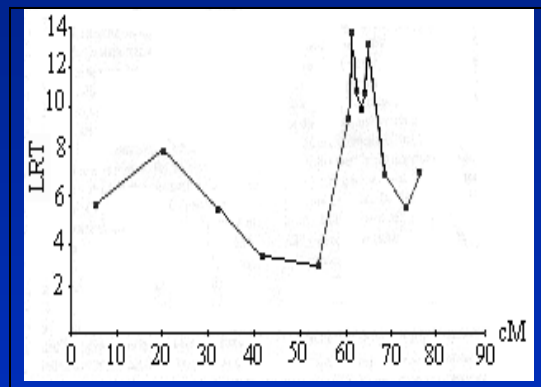
LDLA

		Sire		Dam		Progeny	
		Pat	Mat	Pat	Mat	Pat	Mat
Sire	Pat	1					
	Mat	0.8	1				
Dam	Pat	0.5	0.5	1			
	Mat	0.9	0.5	0.5	1		
Progeny	Pat	1	0.8	0.5	0.9	1	
	Mat	0.5	0.5	1	0.5	0.8	1

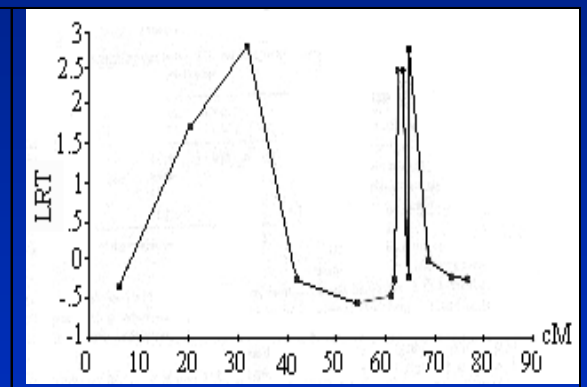
Combined LD-LA mapping

- Example of twinning QTL in Norwegian dairy cattle (Meuwissen et al. 2002)

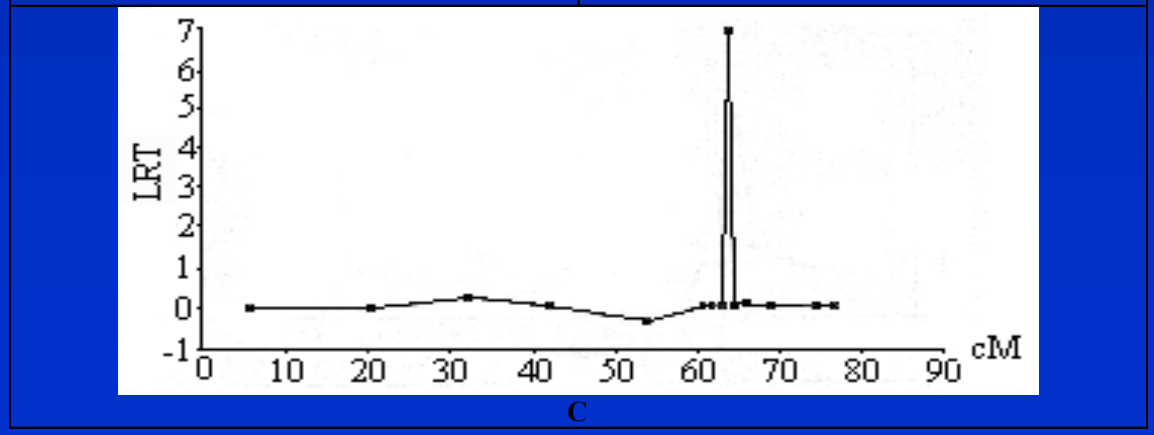
LD



LA



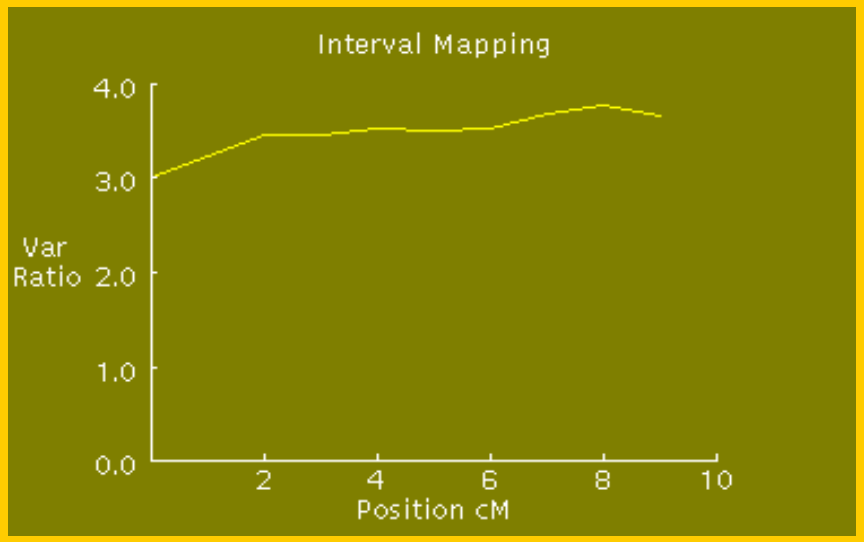
LD-LA



Combined LD-LA mapping

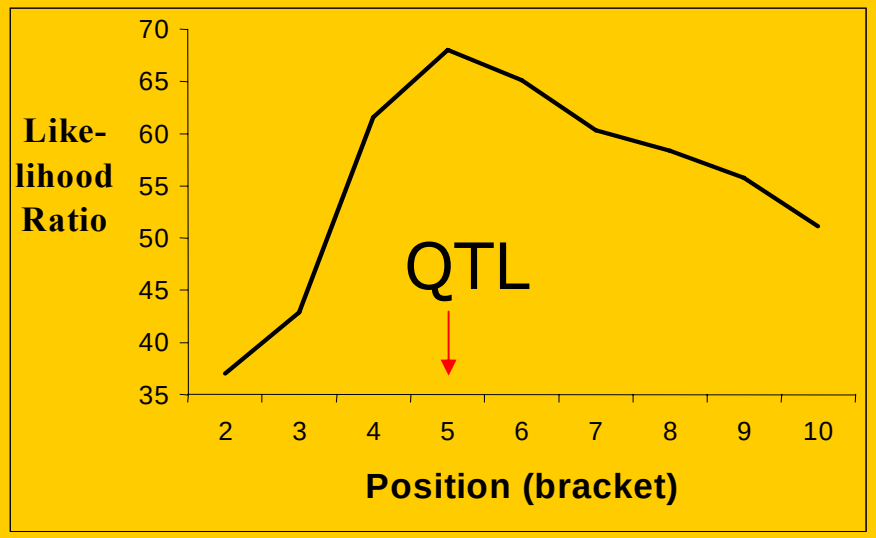
- How much information does LD add to the analysis?
  - Depends on marker spacing and extent of LD

LA



A

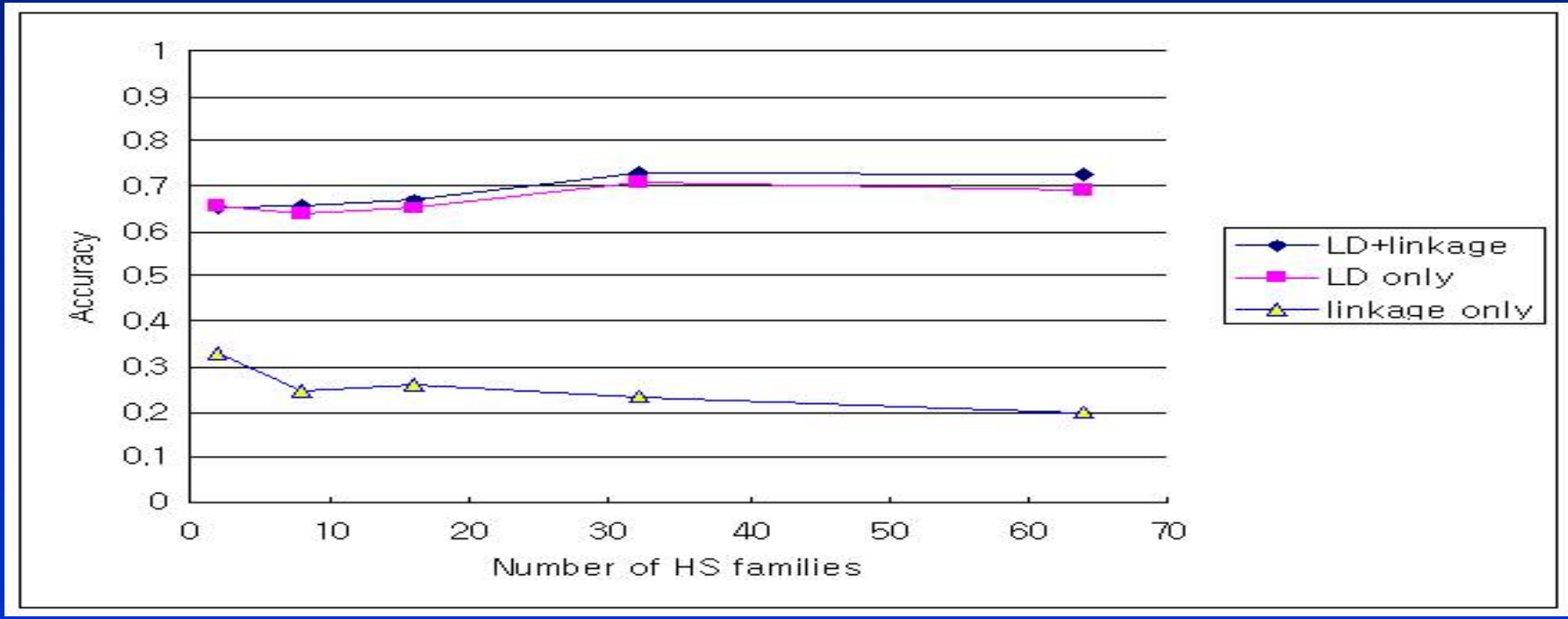
LD-LA



B

Combined LD-LA mapping

- Can we use half-sib families for LD analysis?



- Yes
  - Dam haplotypes provide LD information

Combined LD-LA mapping

- Number of progeny required to position QTL to a 95% C.I. 3cM interval with different designs:

Population	Number of genotyped progeny required to map QTL to 3cM 95% C.I.
F2	7407
Full sib	12685
Commercial (LDLA)	900

- Of course depends on assumptions about extent of LD ~ determined by N_e

Combined LD-LA mapping

- Can we use half-sib families for LD analysis?
  - + selective genotyping ?

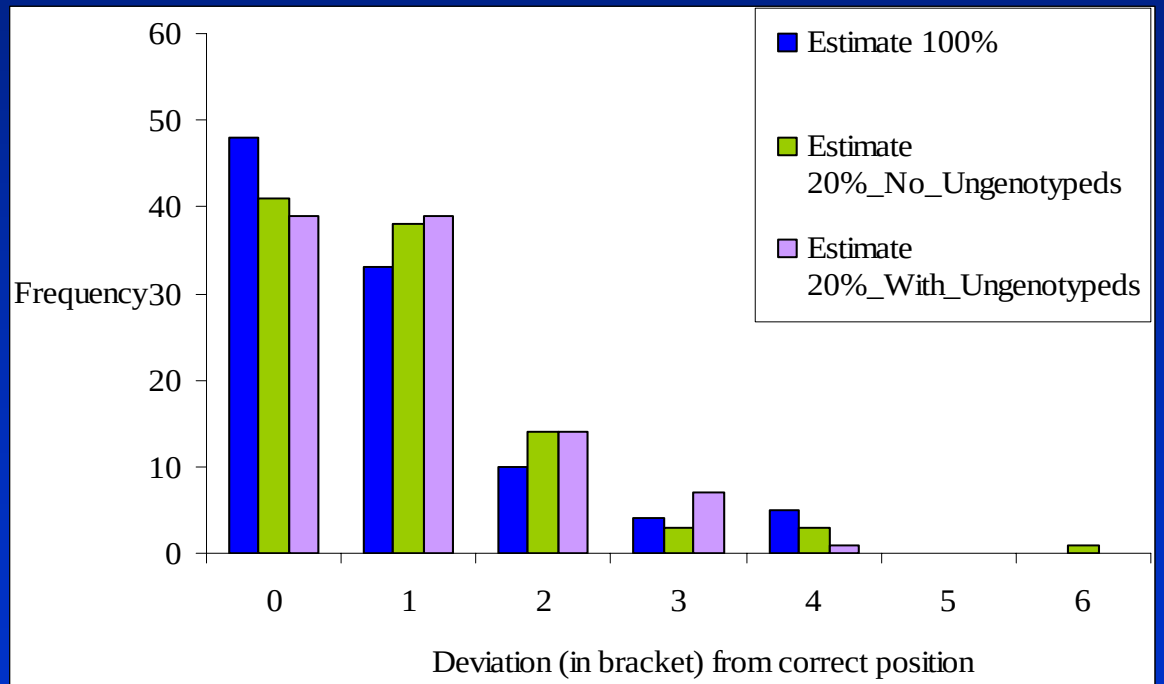


Figure 1. Precision of QTL position estimates, 15 sire design

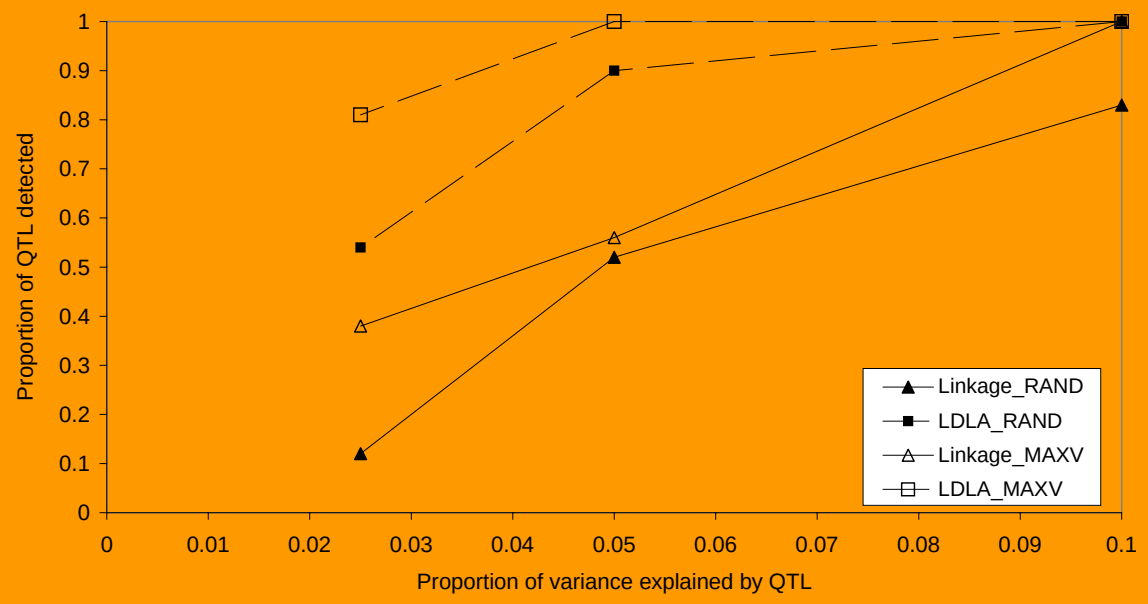
- Yes

Combined LD-LA mapping

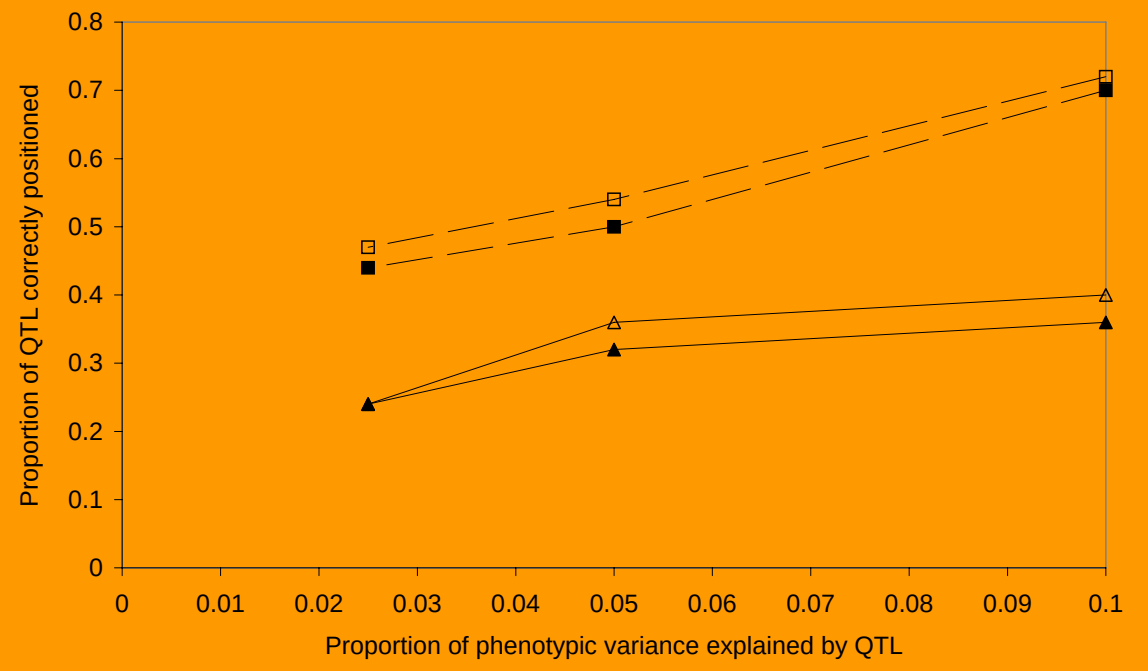
- LDLA analysis + selective genotyping = Cheap? experiment able to position QTL with high degree of precision

Combined LD-LA mapping

- Which families to pick for LD analysis?
  - Those with maximum within half sib family variance
  - Maximise chance of QTL segregating
  - Another ‘selective genotyping’ strategy



A



Power of a mapping design (30 sires mated to 60 dams in a half sib design and 10 progeny per dam, no recombination in males) to

A. detect QTL and
B. Position QTL within 3cM of the true QTL position,

for QTL explaining different proportions of the phenotypic variance. The 30 half sib families were either randomly selected (RAND) from the breeding population for genotyping, or the half sib families with the largest within half sib family variance for the trait (MAXV) were selected.

LD mapping of QTL

- Take home points
  - LD mapping uses information on historical recombinants to narrow QTL C.I.
  - Power depends on extent of LD and marker density
    - Knowledge of extent of LD critical
  - Some suggestion that single marker regression a good approach, with high marker density?
  - IBD approach allows extension to capture LA information
    - v. important with lower marker density >> power
    - filter spurious peaks
  - Half sib designs ideal for LDLA mapping
    - Use LD info from dam haplotypes