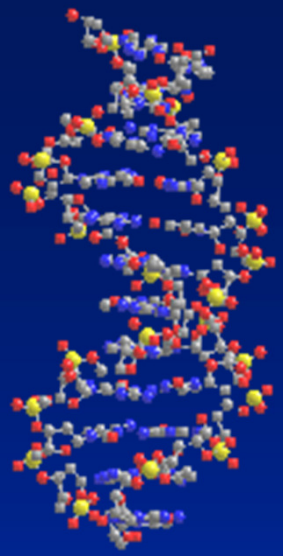
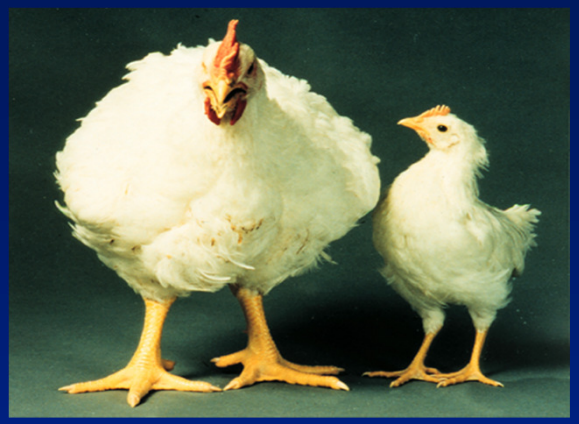
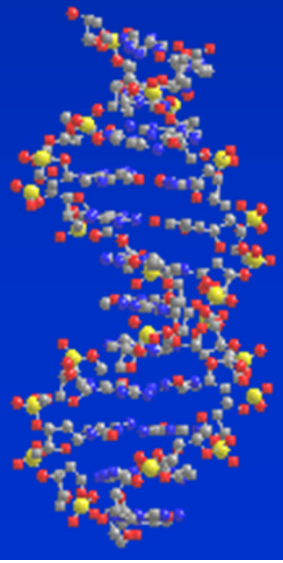




Genomic Selection in the era of Genome sequencing



Course overview

- Day 1
  - Linkage disequilibrium in animal and plant genomes
- Day 2
  - Genome wide association studies
- Day 3
  - Genomic selection
- Day 4
  - Genomic selection
- Day 5
  - Imputation and whole genome sequencing for genomic selection

Genome wide association

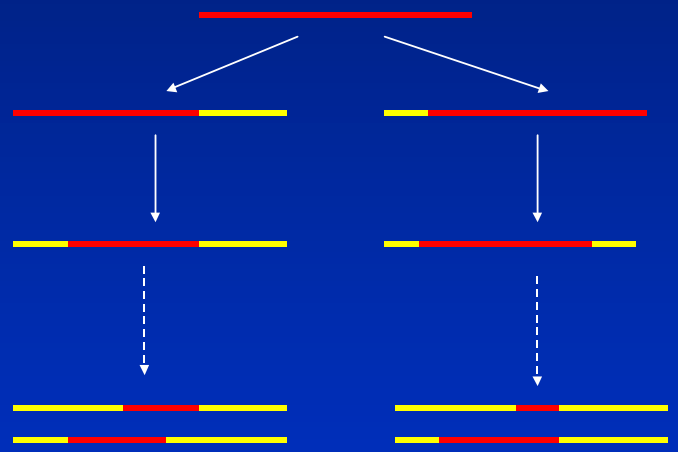
- Association testing with single marker regression
- Power of genome wide association studies
- Accounting for population structure
- LD mapping with haplotypes
- Validation

Genome wide association

- LD mapping of QTL exploits population level associations between markers and QTL.

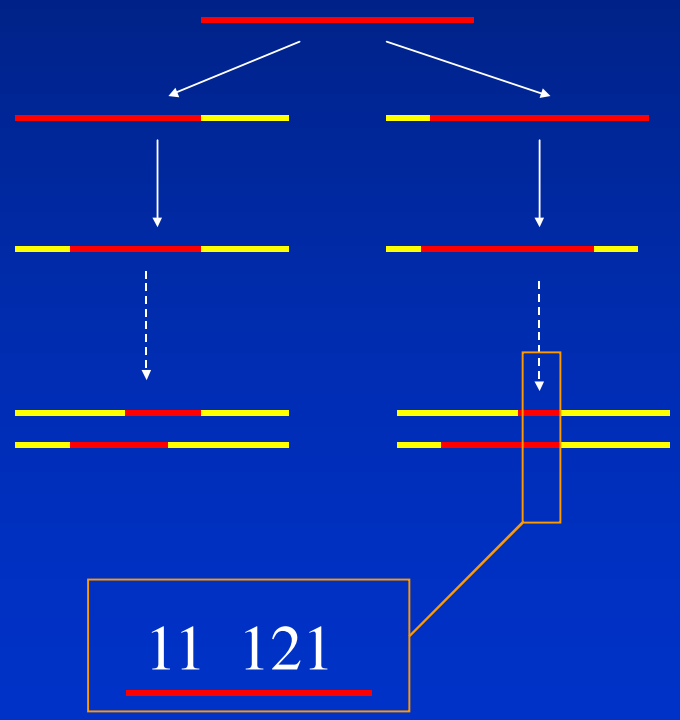
Genome wide association

- 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



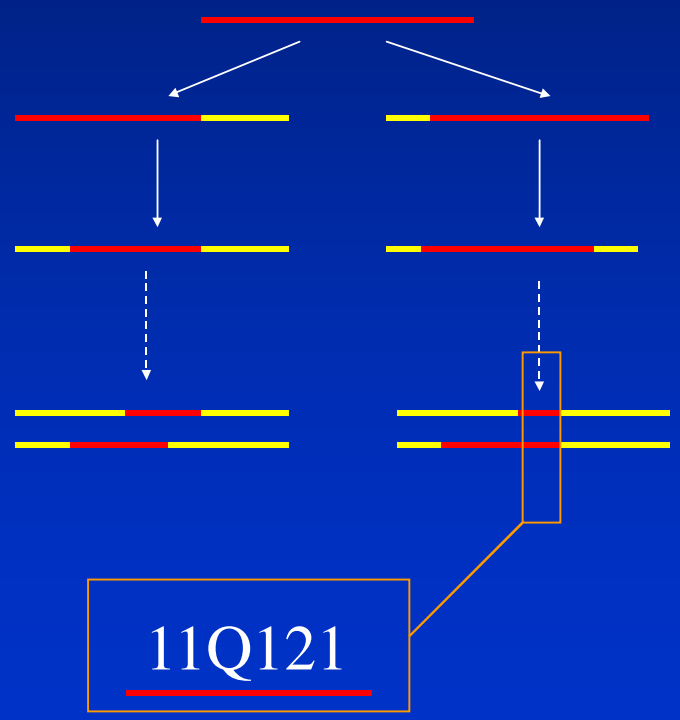
Genome wide association

- 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



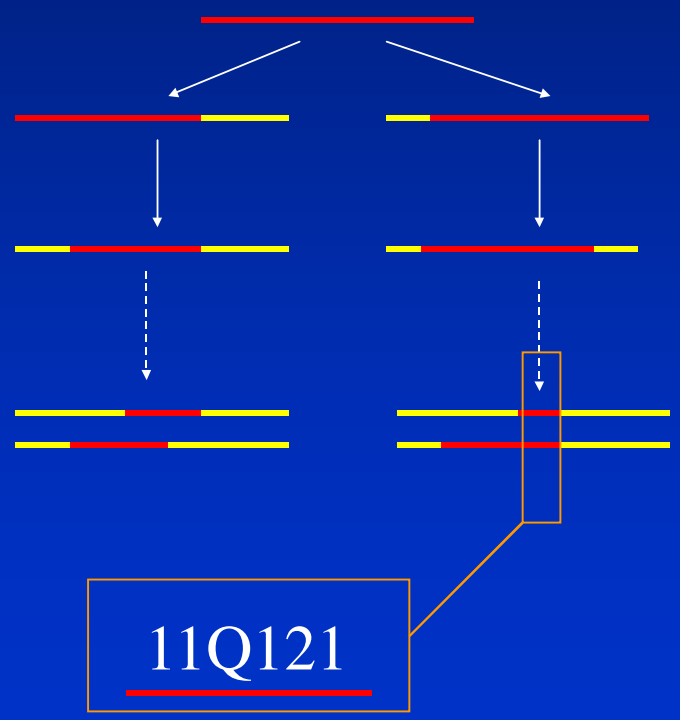
Genome wide association

- 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



Genome wide association

- 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 way to exploit these associations is by single SNP 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 2
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

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

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

Single marker regression

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

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

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:

$$\begin{bmatrix} \hat{\mu} \\ \hat{g} \end{bmatrix} = \begin{bmatrix} 2.35 \\ 1.28 \end{bmatrix}$$

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

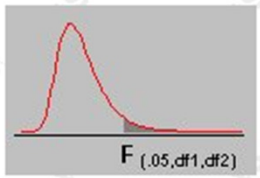
Single marker regression

- 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 .

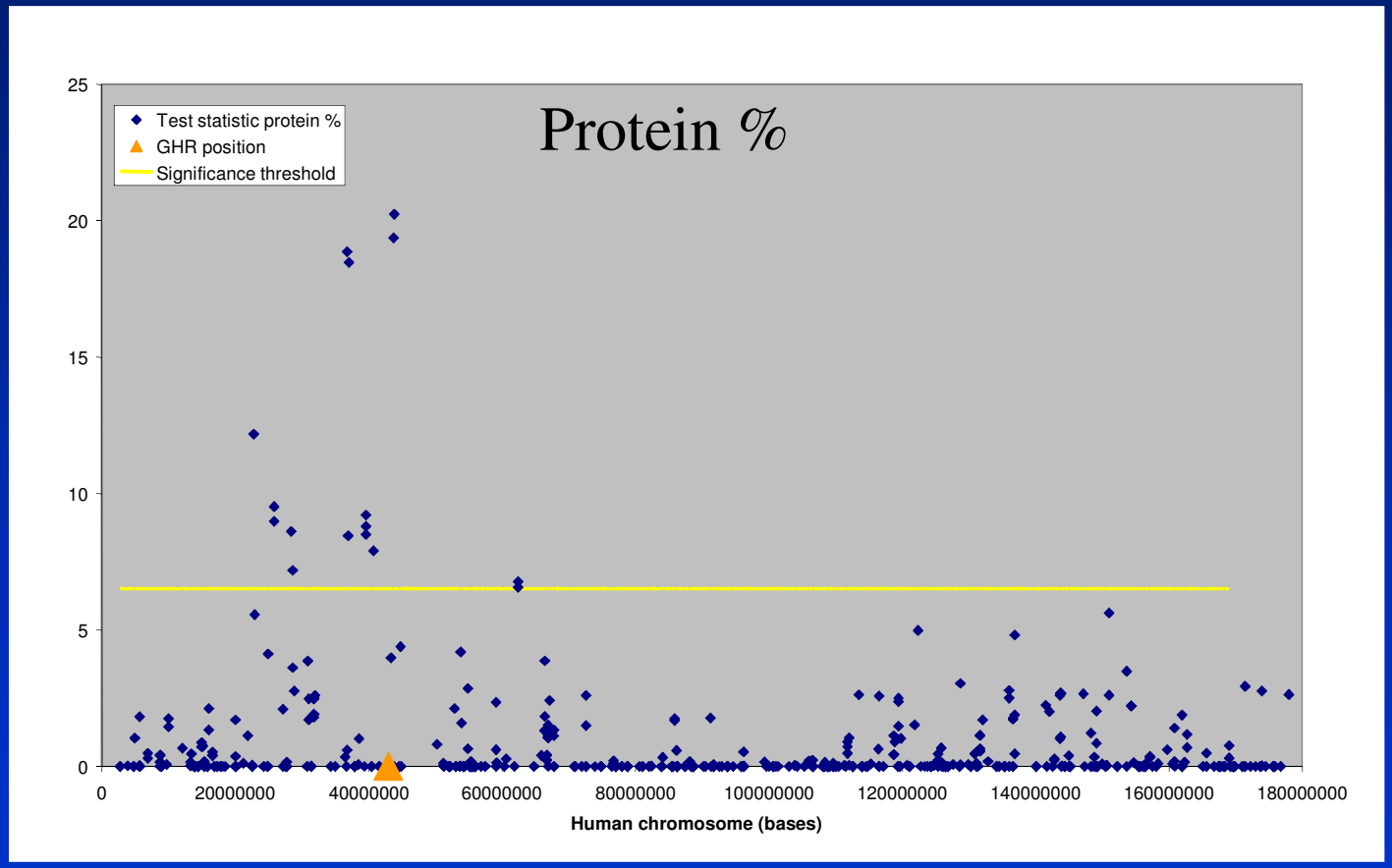


df2/df1	1	2	3	4	5	6	7	8	9	10
1	161.4476	199.5000	215.7073	224.5832	230.1619	233.9860	236.7684	238.8827	240.5433	241.8817
2	18.5128	19.0000	19.1643	19.2468	19.2964	19.3295	19.3532	19.3710	19.3848	19.3959
3	10.1280	9.5521	9.2766	9.1172	9.0135	8.9406	8.8867	8.8452	8.8123	8.7855
4	7.7086	6.9443	6.5914	6.3882	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

Experiment

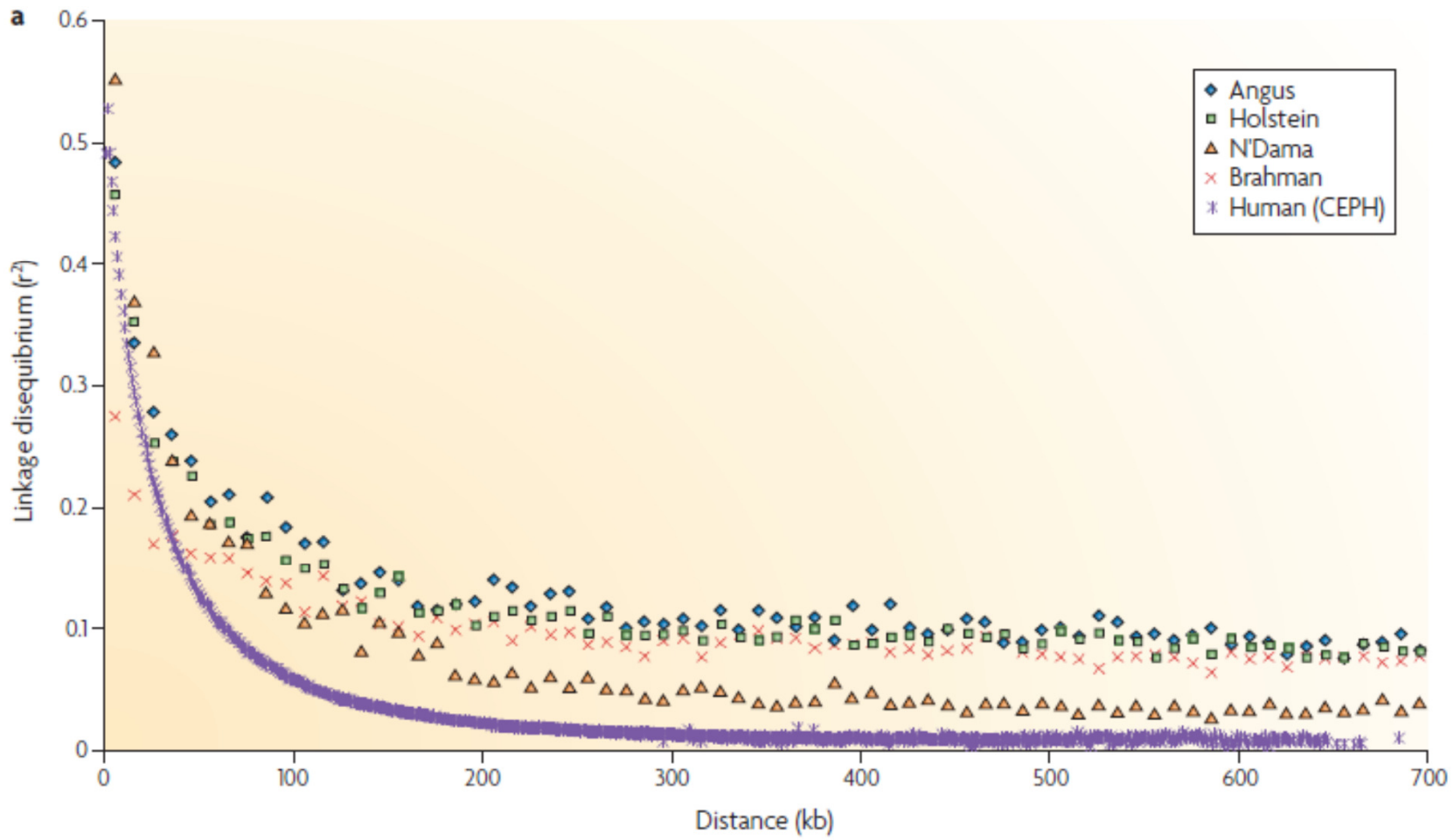
- 384 Holstein-Friesian dairy bulls selected from Australian dairy bull population
- genotyped for 10 000 SNPs
- Single marker regression with protein%

Results of genome scans with dense SNP panels

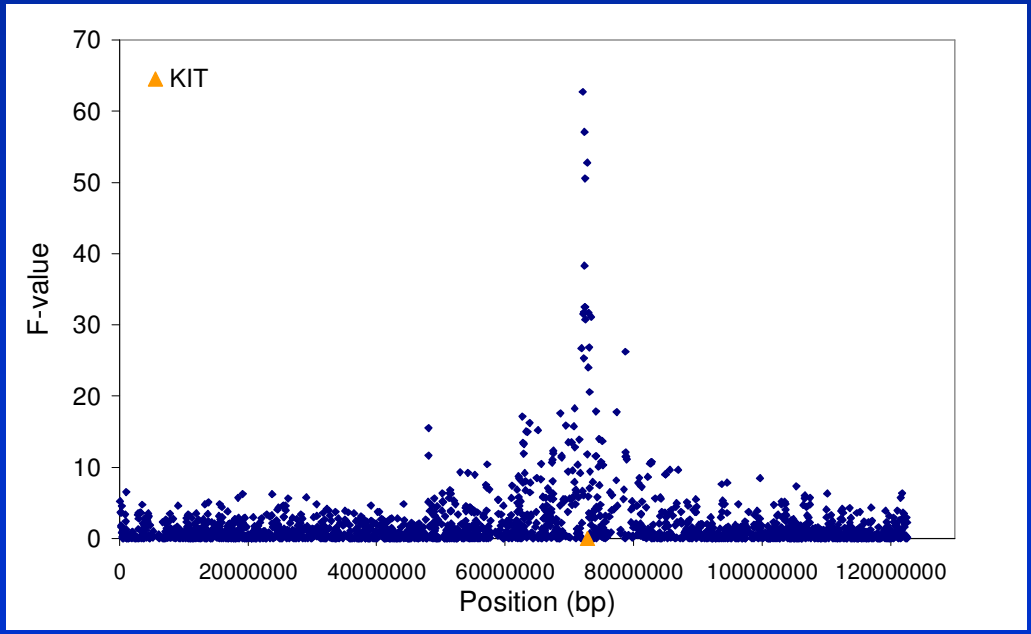


Extent of LD in humans and livestock

And cattle……



Proportion of black….



Genome wide association

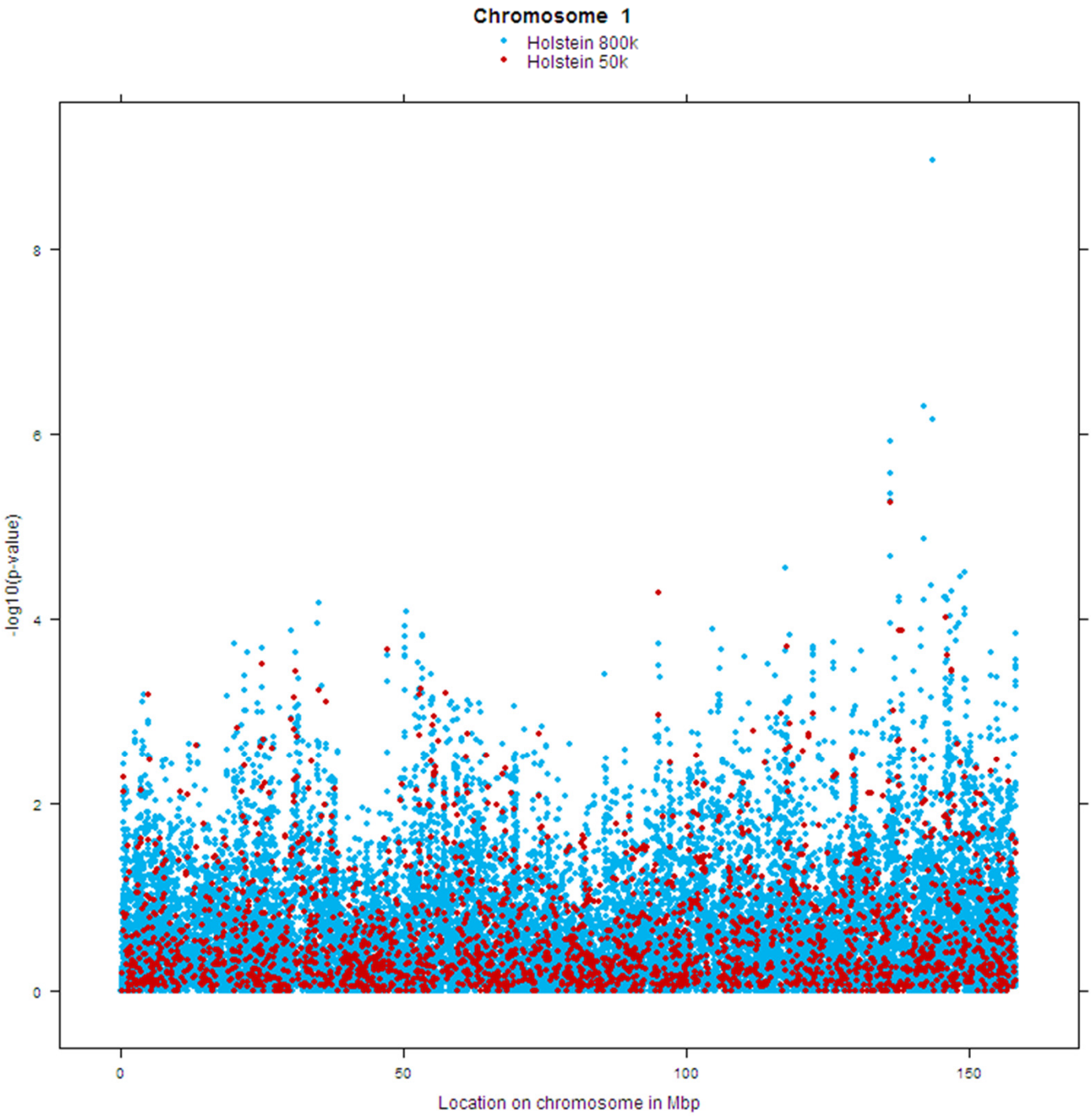
- Association testing with single marker regression
- Power of genome wide association studies
- Accounting for population structure
- LD mapping with haplotypes
- Validation

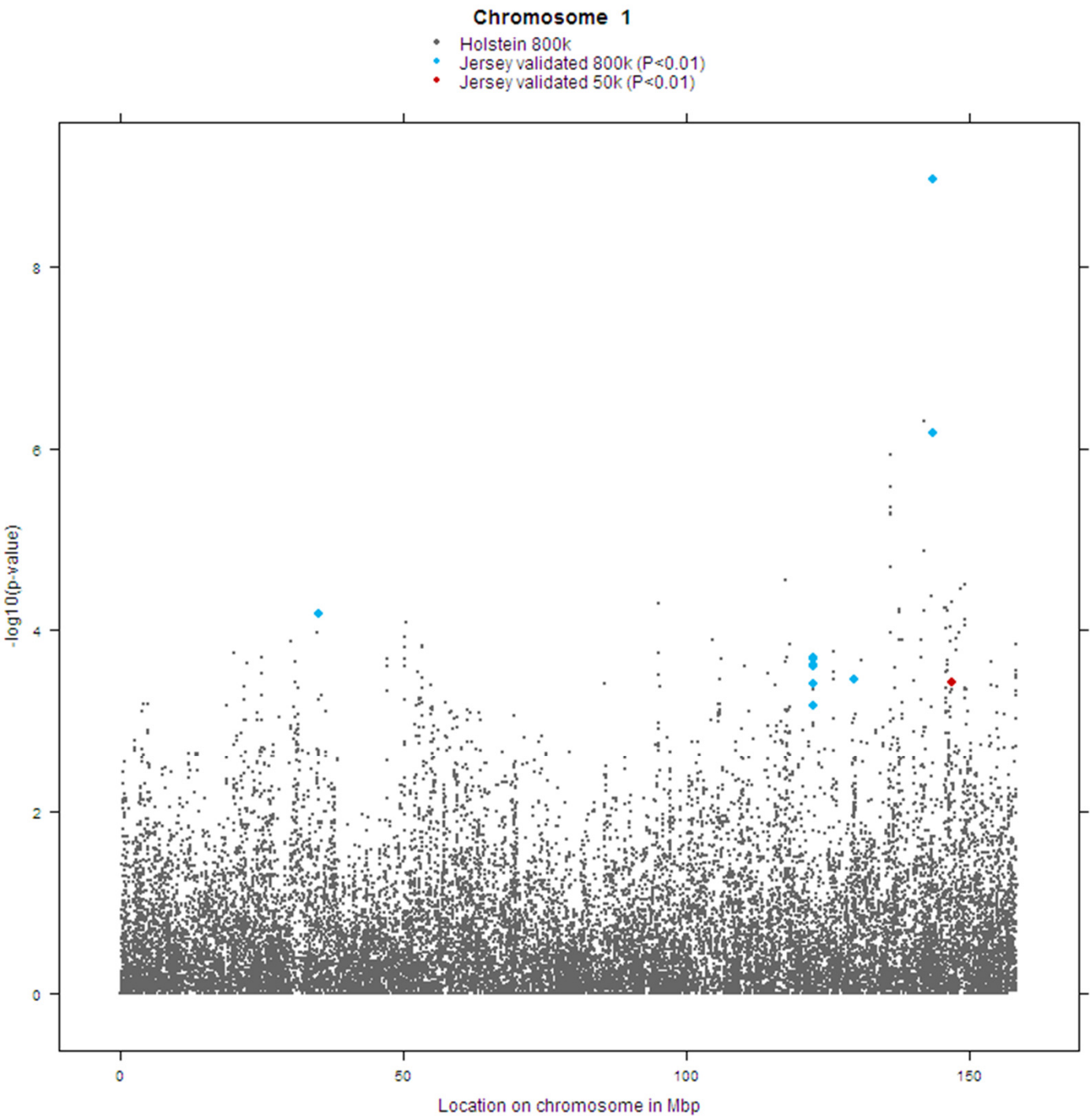
Power of GWAS

- 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
  - H_0 = no QTL
  - H_1 = there is a QTL
- How many animals do we need to genotype and phenotype?

Power of GWAS

- 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


Chromosome 1
Holstein 800k
Holstein 50k
-log10(p-value)
0
2
4
6
8
0
50
100
150
Location on chromosome in Mbp


Chromosome 1
Holstein 800k
Jersey validated 800k (P<0.01)
Jersey validated 50k (P<0.01)
-log10(p-value)
0
2
4
6
8
0
50
100
150
Location on chromosome in Mbp

Power of GWAS

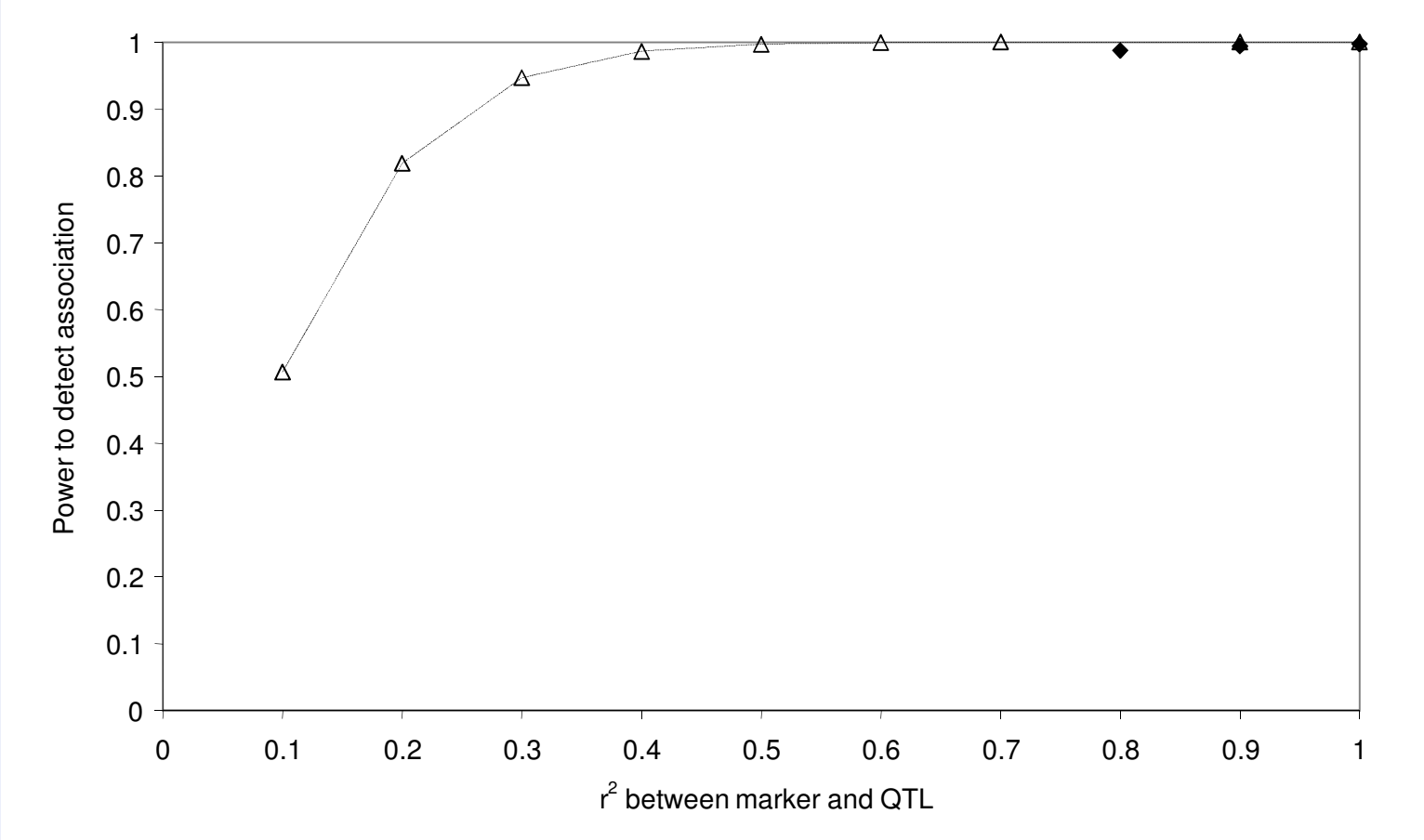
- Power is a function of:
  - r^2 between the marker and QTL
    - sample size must be increased by $1/r^2$ to detect an un-genotyped QTL, compared with sample size for testing QTL itself
  - Proportion of total phenotypic variance explained by the QTL
  - Number of phenotypic records

Power of GWAS

- 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

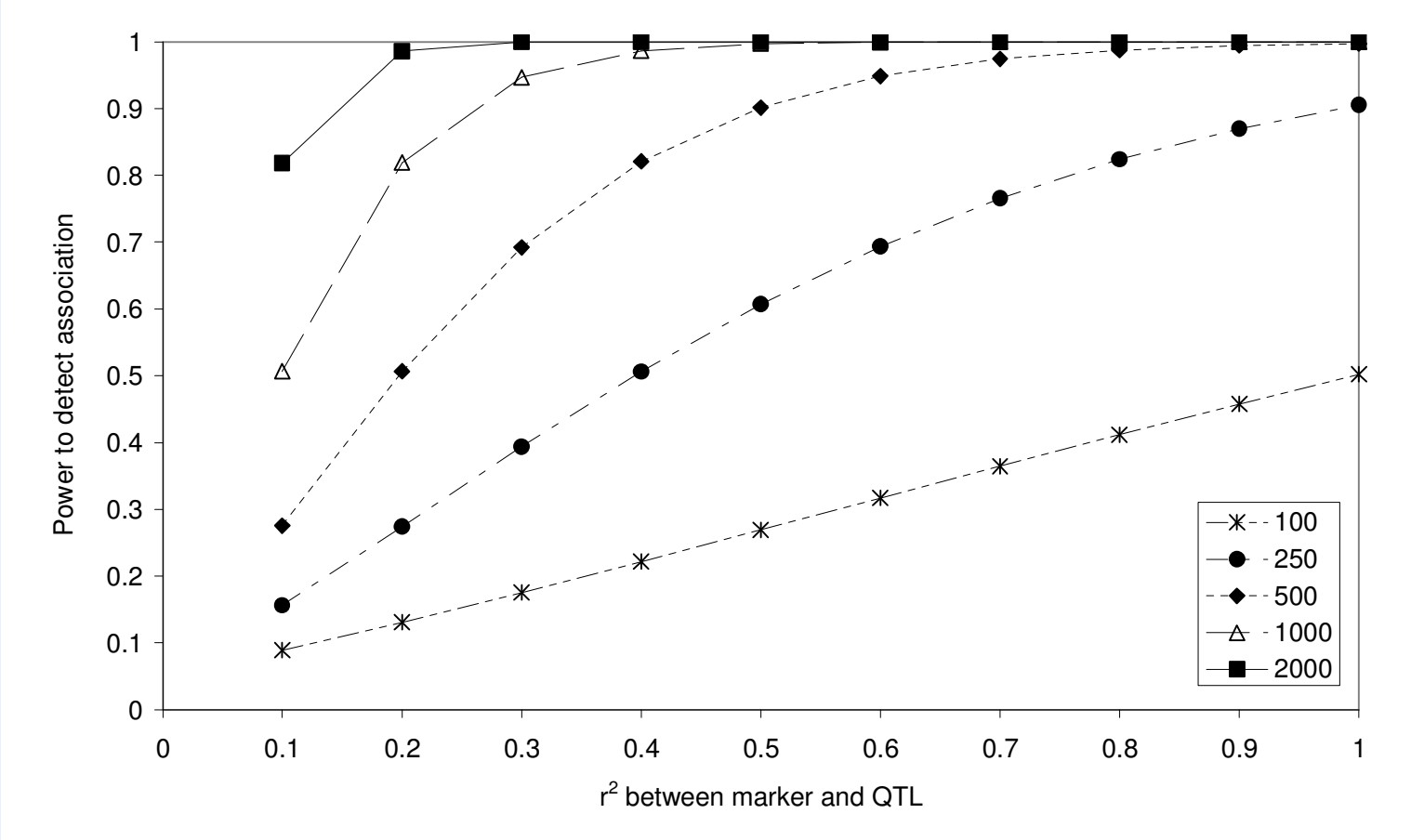
Power of GWAS

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



Power of GWAS

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



Search This journal

ne publication > Letter > Abstract

Letter abstract

Nature Genetics
Published online: 13 January 2008 | doi:10.1038/ng.74

Common variants in the *GDF5-UQCC* region are associated with variation in human height

Serena Sanna[1,2,19], Anne U Jackson[1,19], Ramaiah Nagaraja[3], Cristen J Willer[1], Wei-Min Chen[1,4], Lori L Bonnycastle[5], Haiqing Shen[6], Nicholas Timpson[7,8], Guillaume Lettre[9], Gianluca Usala[2], Peter S Chines[5], Heather M Stringham[1], Laura J Scott[1], Mariano Dei[2], Sandra Lai[2], Giuseppe Albai[2], Laura Crisponi[2], Silvia Naitza[2], Kimberly F Doheny[10], Elizabeth W Pugh[10], Yoav Ben-Shlomo[7], Shah Ebrahim[11], Debbie A Lawlor[7,8], Richard N Bergman[12], Richard M Watanabe[12,13], Manuela Uda[2], Jaakko Tuomilehto[14], Josef Coresh[15], Joel N Hirschhorn[9], Alan R Shuldiner[6,16], David Schlessinger[3], Francis S Collins[5], George Davey Smith[7,8], Eric Boerwinkle[17], Antonio Cao[2], Michael Boehnke[1], Gonçalo R Abecasis[1] & Karen L Mohlke[18]

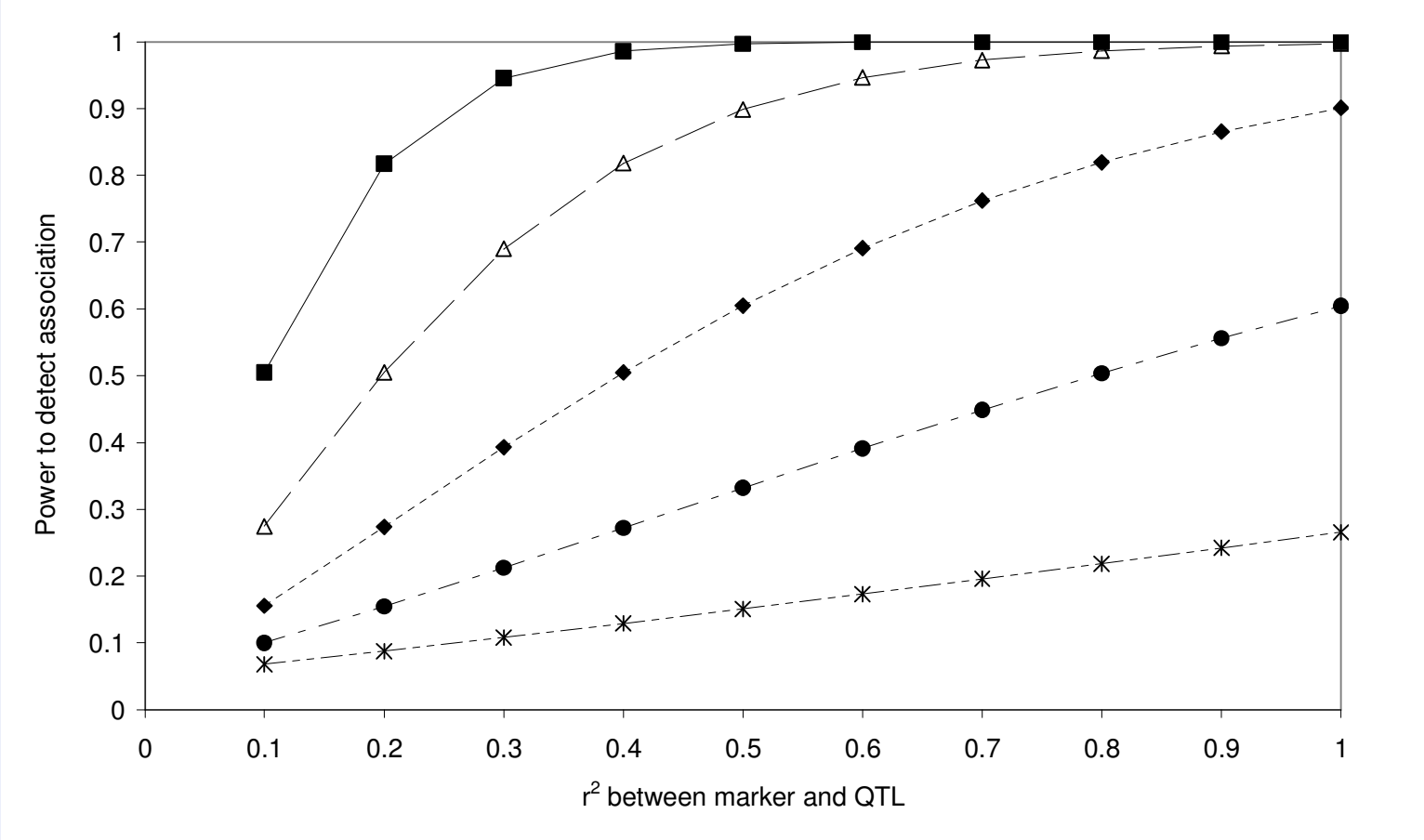
top

Identifying genetic variants that influence human height will advance our understanding of skeletal growth and development. Several rare genetic variants have been convincingly and reproducibly associated with height in mendelian syndromes, and common variants in the transcription factor gene *HMGA2* are associated with variation in height in the general population[1]. Here we report genome-wide association analyses, using genotyped and imputed markers, of 6,669 individuals from Finland and Sardinia, and follow-up analyses in an additional 28,801 individuals. We show that common variants in the osteoarthritis-associated locus[2] *GDF5-UQCC* contribute to variation in height with an estimated additive effect of 0.44 cm (overall $P < 10^{-15}$). Our results indicate that there may be a link between the genetic basis of height and osteoarthritis, potentially mediated through alterations in bone growth and development.

< 1% of phenotypic variance!

Power of GWAS

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



Power of GWAS

- 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)

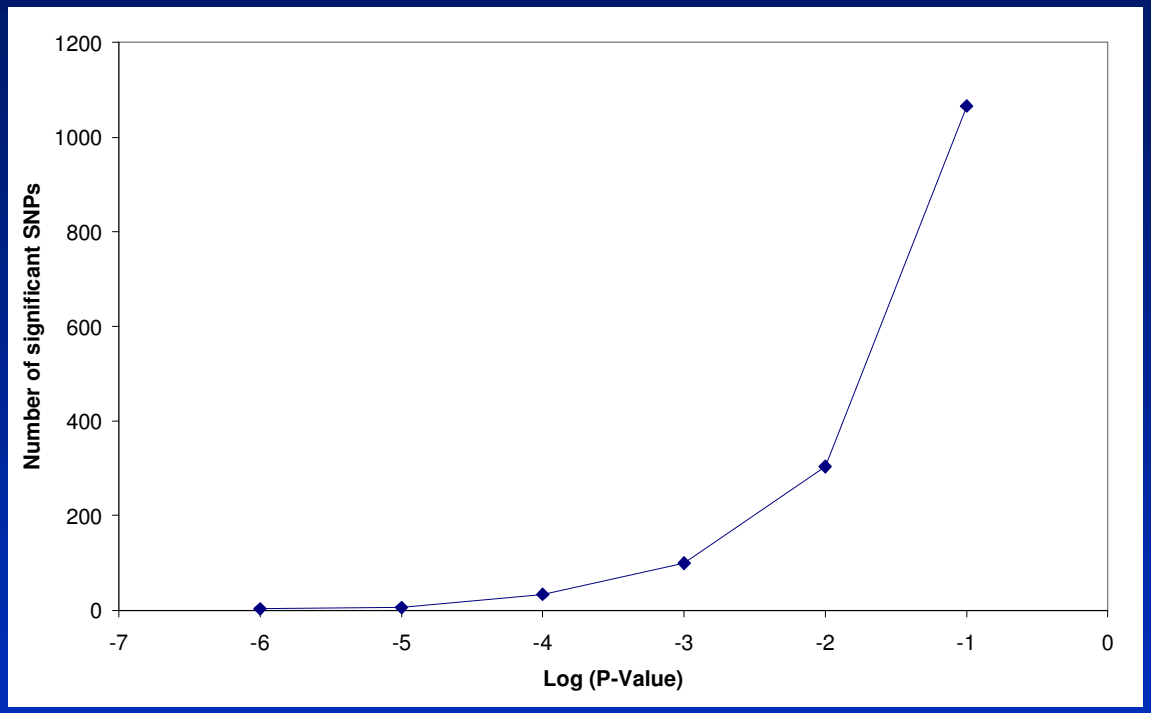
Power of GWAS

- 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 actually significant

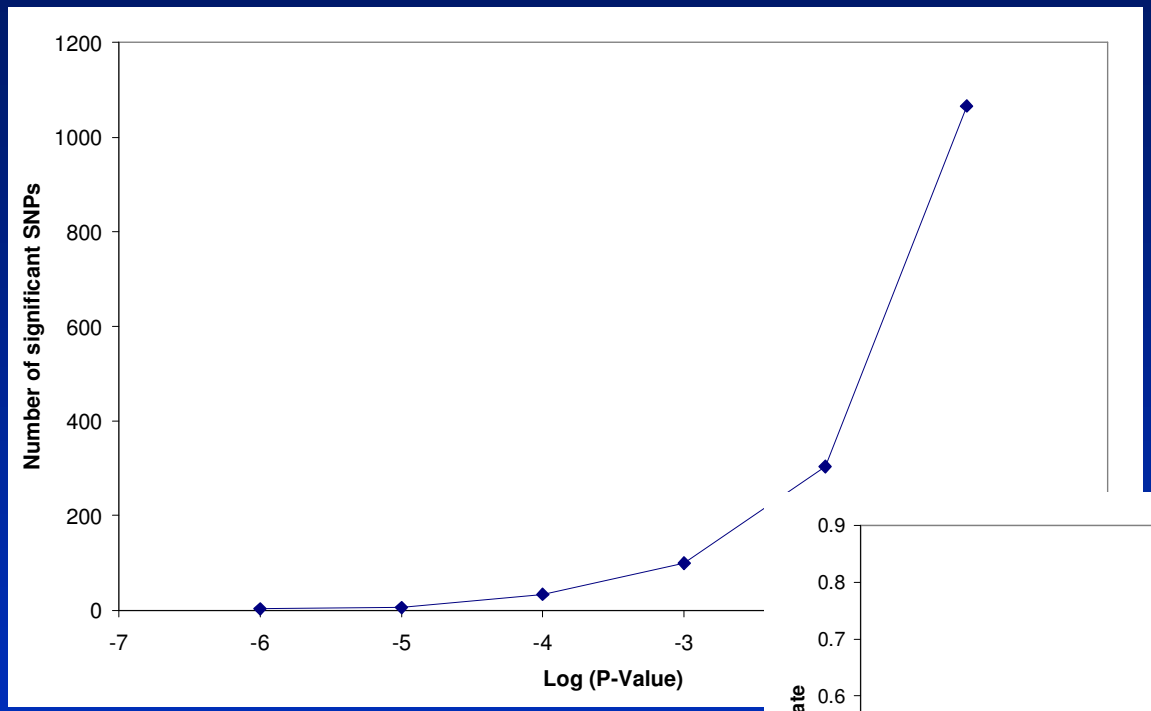
Power of GWAS

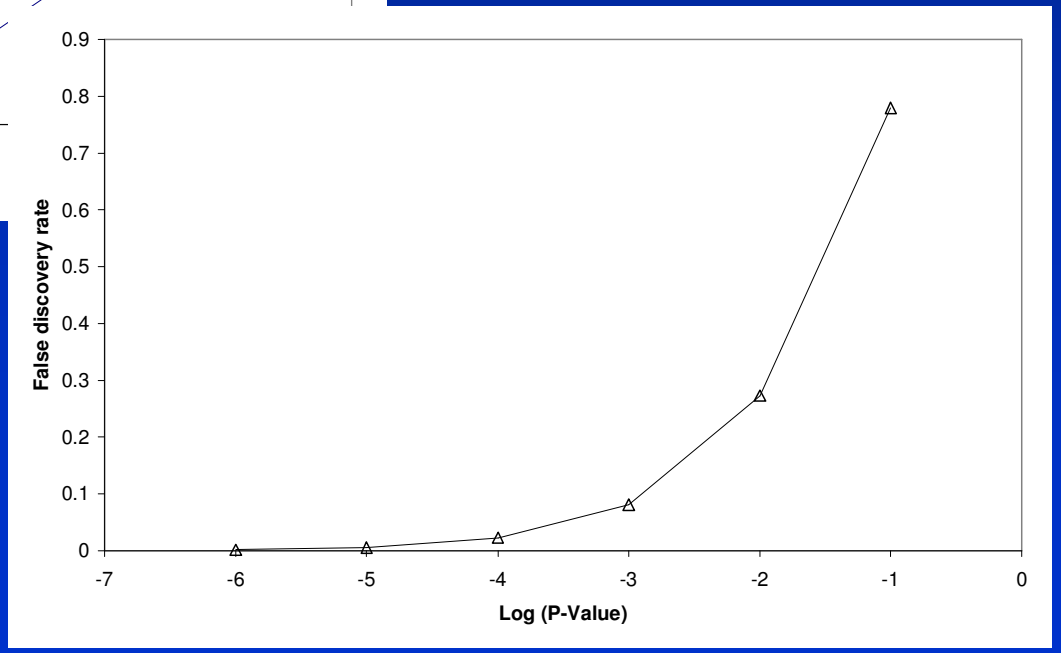
- An alternative is to choose a significance level with an acceptable false discovery rate (FDR)
- Proportion of significant results which are really false positives
- FDR = mP/n
  - m = number of markers tested
  - P = significance level (eg. P=0.01)
  - n = number of markers actually significant
-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

Power of GWAS



Power of GWAS





Genome wide association

- Association testing with single marker regression
- Power of genome wide association studies
- Accounting for population structure
- LD mapping with haplotypes
- Validation

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*)

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…..

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

- Where
  - **u** is a vector of polygenic effects 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

$g=-3$

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	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

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

$\lambda = 0.33$

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

$$\begin{bmatrix} \hat{\mu} \\ \hat{g} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} 6 & 8 & 1 & 1 & 1 & 1 & 1 & 1 \\ 8 & 12 & 1 & 2 & 2 & 1 & 1 & 1 \\ 1 & 1 & 1.825 & 0.33 & 0.165 & -0.33 & -0.33 & -0.33 \\ 1 & 2 & 0.33 & 1.66 & 0 & -0.33 & -0.33 & 0 \\ 1 & 2 & 0.165 & 0 & 1.495 & 0 & 0 & -0.33 \\ 1 & 1 & -0.33 & -0.33 & 0 & 1.66 & 0 & 0 \\ 1 & 1 & -0.33 & -0.33 & 0 & 0 & 1.66 & 0 \\ 1 & 1 & -0.33 & 0 & -0.33 & 0 & 0 & 1.66 \end{bmatrix}^{-1} \begin{bmatrix} 33.45 \\ 37.96 \\ 10.1 \\ 2.2 \\ 2.31 \\ 6.57 \\ 6.06 \\ 6.21 \end{bmatrix}$$

Population structure

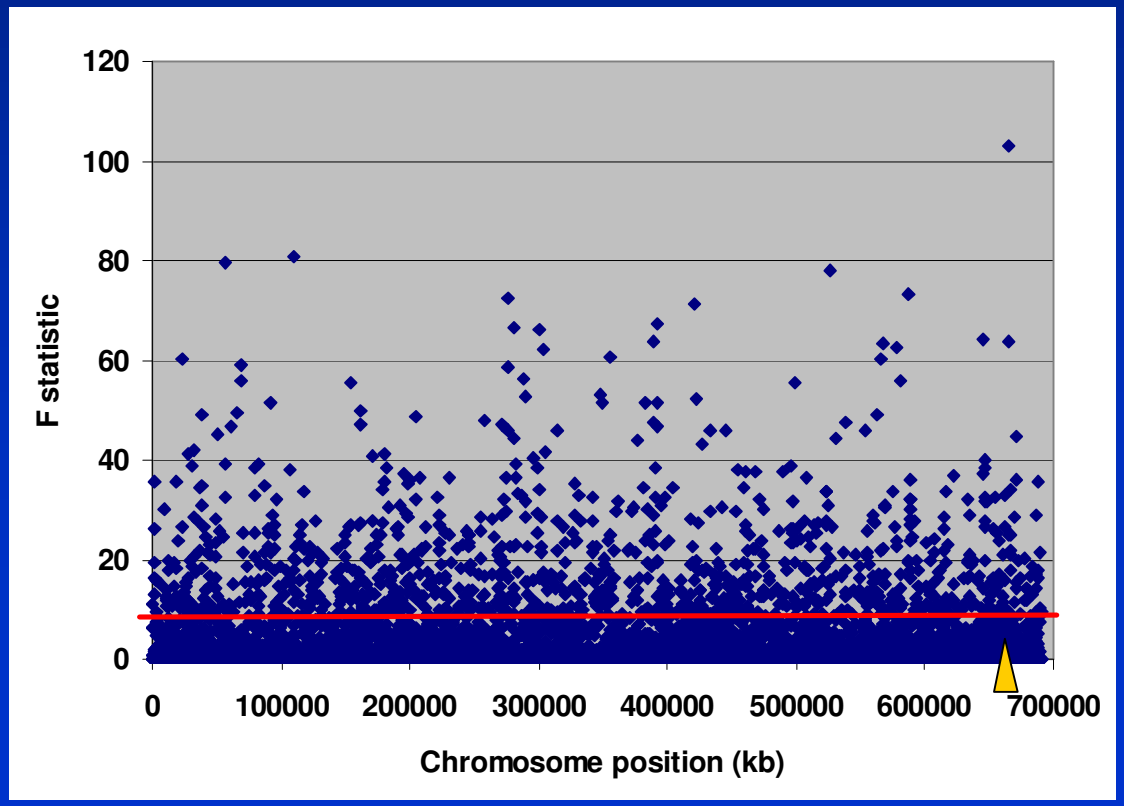
- Example

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

$$\begin{bmatrix} \hat{\mu} \\ \hat{g} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} 10.6 \\ -3.7 \\ 1.9 \\ -1.1 \\ -0.9 \\ 0.2 \\ -0.3 \\ -0.2 \end{bmatrix}$$

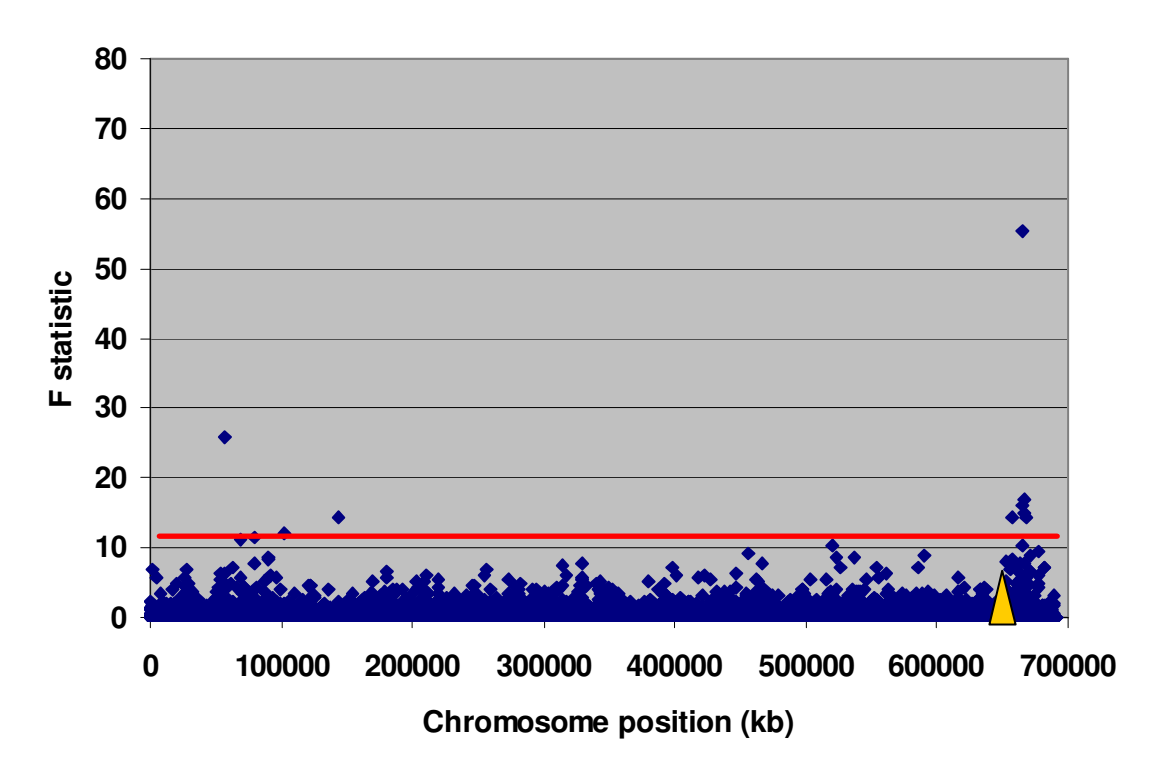
Population structure

- A simulated data set with a half sib family structure, one QTL simulated



Population structure

- A simulated data set with a half sib family structure, one QTL simulated



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

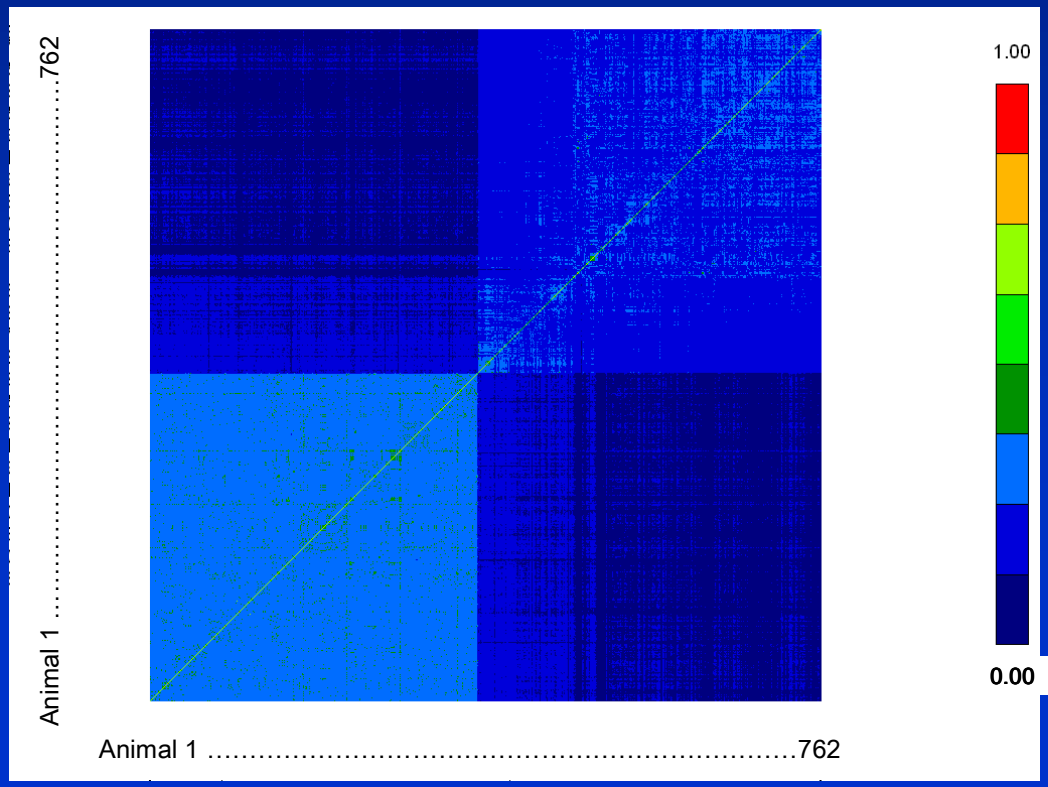
Population structure

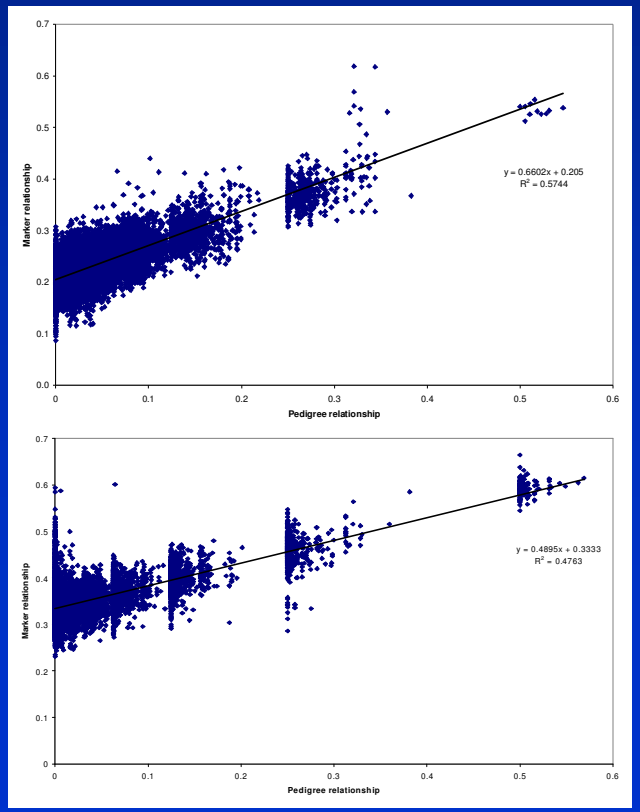
Number of false positives.....

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





Genomic relationship matrix

- Rescale X to account for allele frequencies
  - $w_{ij} = x_{ij} - 2p_j$
- Then

$$\mathbf{G} = \mathbf{WW}'/2\sum_{j=1}^{p} p_j(1-p_j)$$

Genome wide association

- Association testing with single marker regression
- Power of genome wide association studies
- Accounting for population structure
- LD mapping with haplotypes
- Validation

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

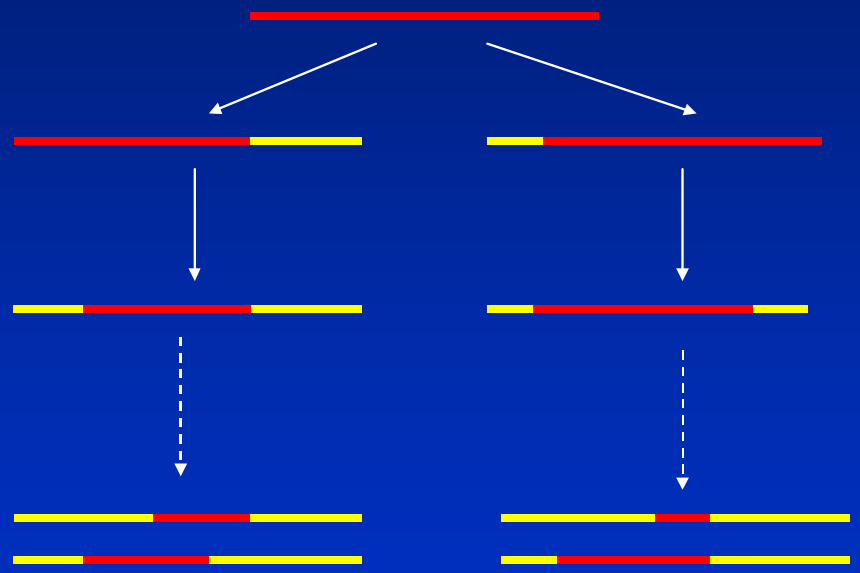
- 1_Q single marker (1 is the allele of the marker)
- 1_1_Q_2_1 Haplotype of markers

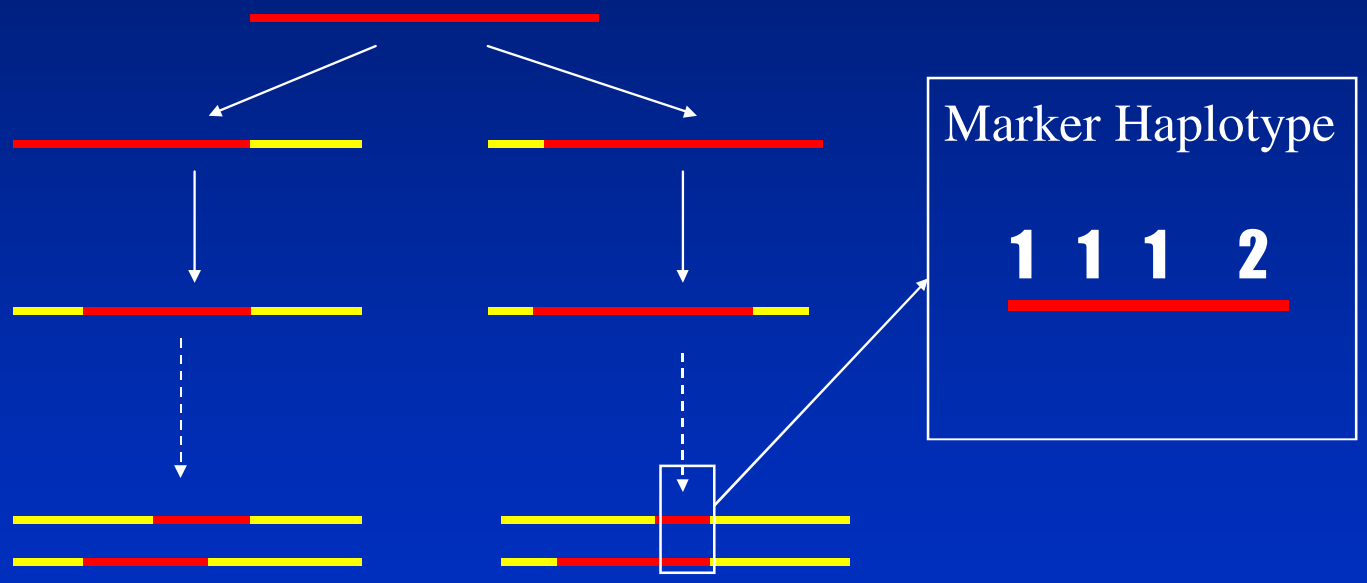
LD mapping with haplotypes

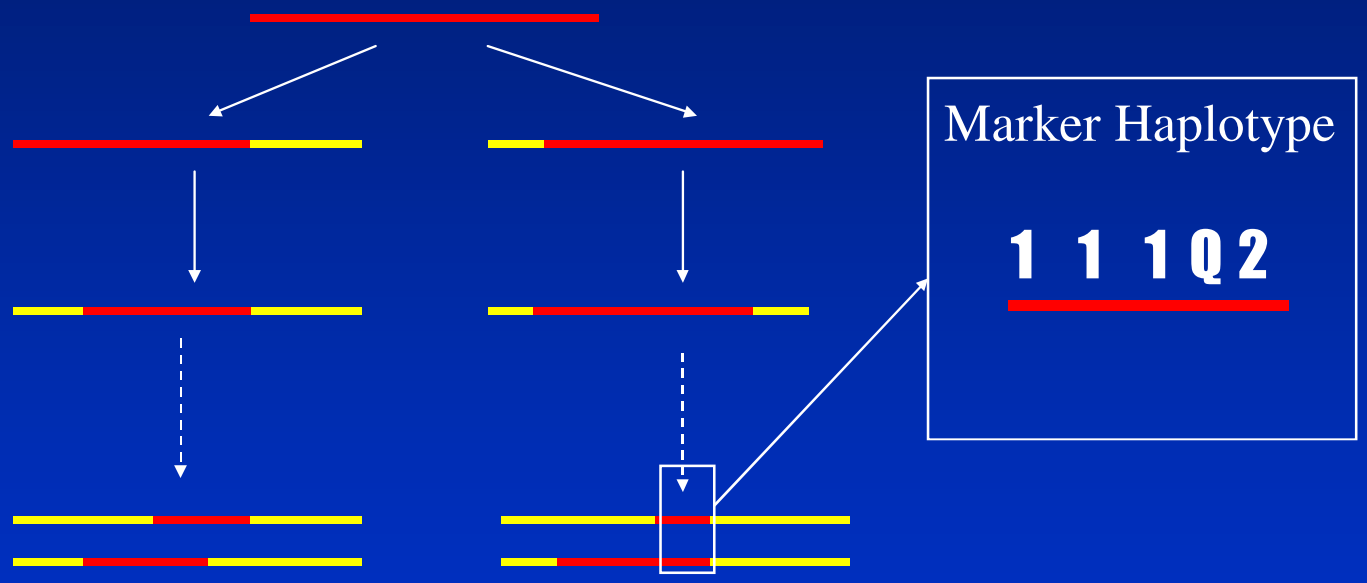
- Value of haplotypes depends on LD between haplotype and QTL
  - If we find two identical haplotypes from the population, what is the probability they carry the same QTL allele?
  - If probability is high, high level of LD between haplotype and QTL

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…….

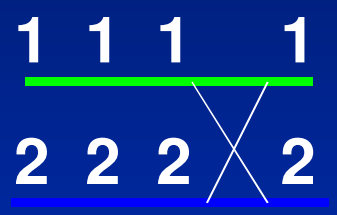
Chance recombination produces the same haplotype…..

1 1 1 1

2 2 2 2

Sire

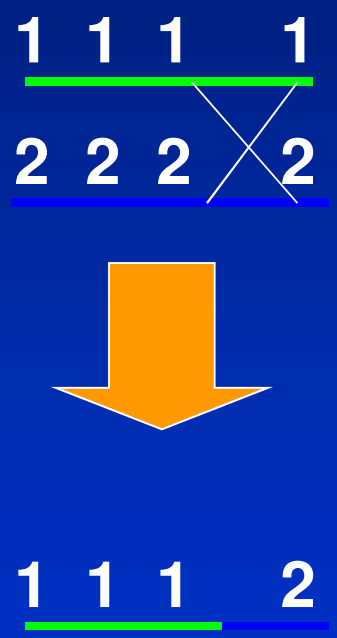
Chance recombination produces the same haplotype…..



Sire

Formation of gamete

Chance recombination produces the same haplotype…..



Sire

Progeny

Chance recombination produces the same haplotype…..

Sire

Progeny

Chance recombination produces the same haplotype…..

1 1 1 q 1

2 2 2 q 2

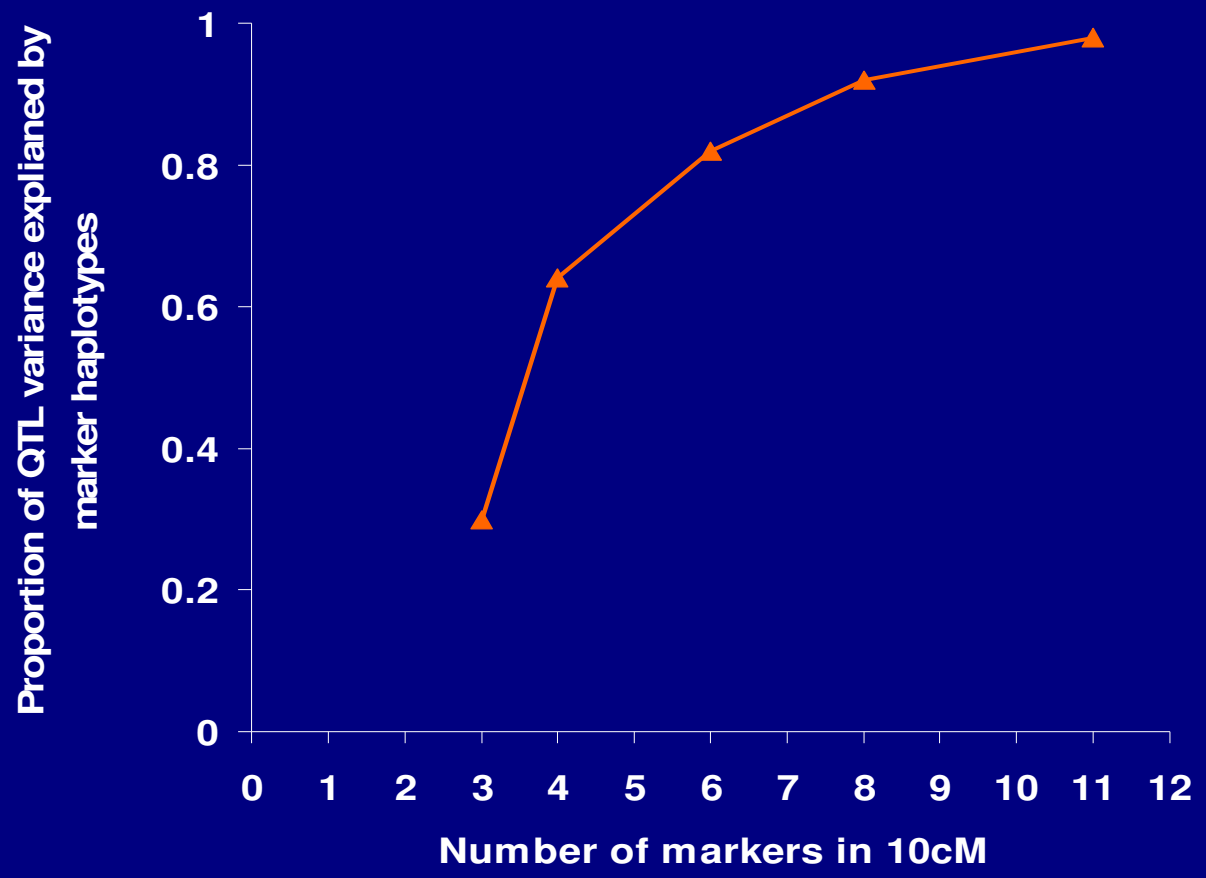
Sire

1 1 1 q 2

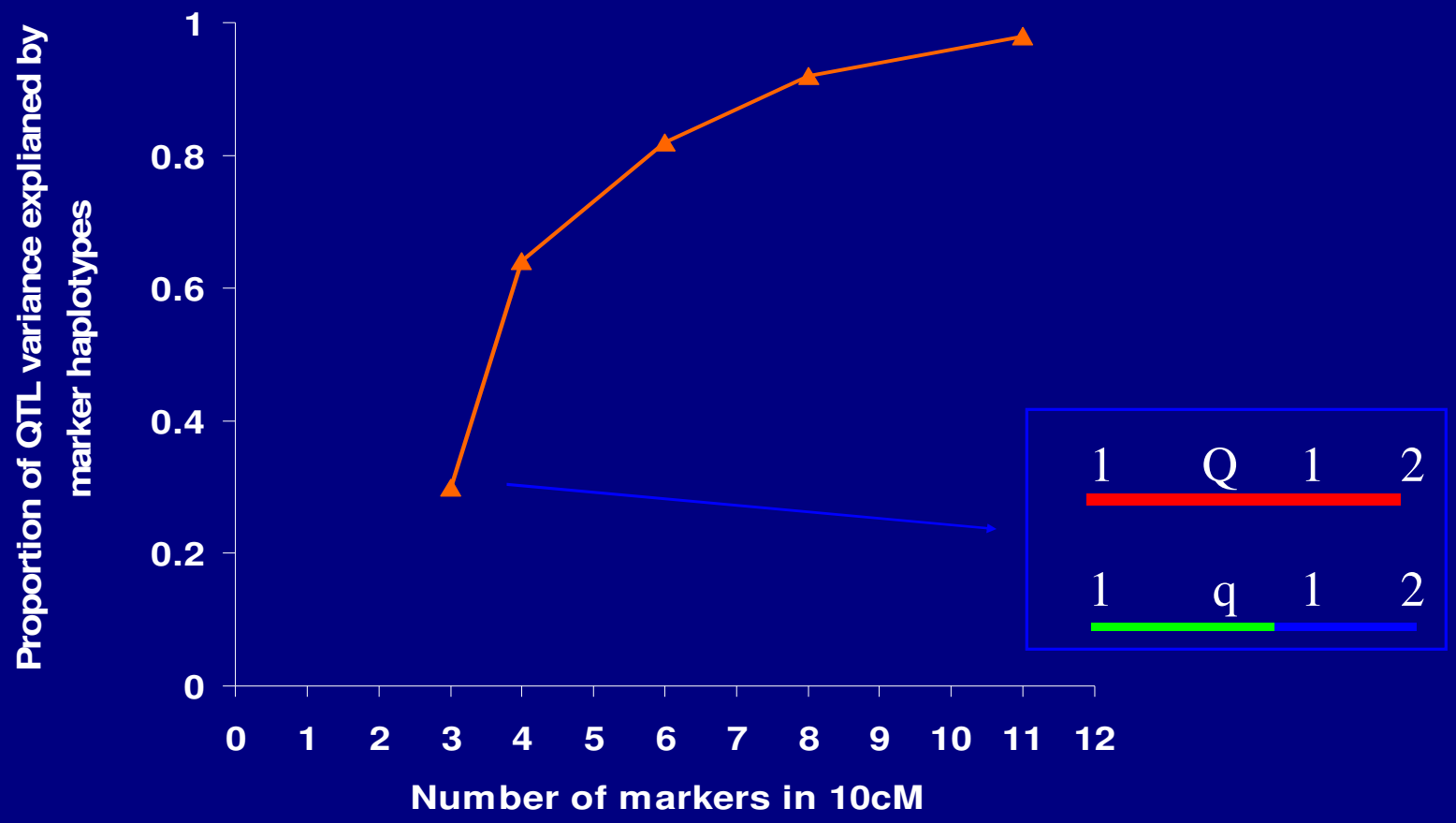
Progeny

1 1 1 Q 2

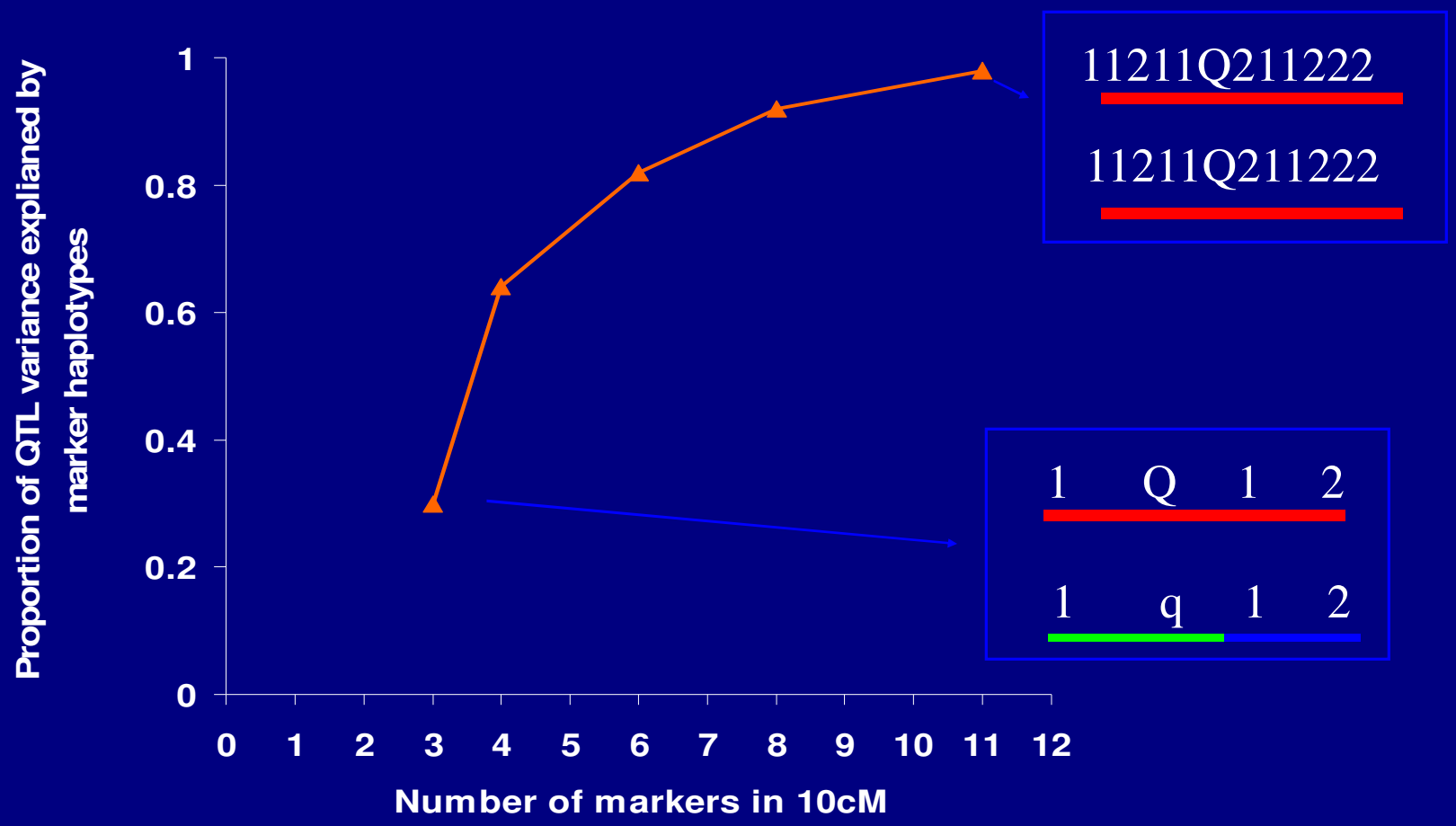
Proportion of QTL variance explained by surrounding markers



Proportion of QTL variance explained by surrounding markers



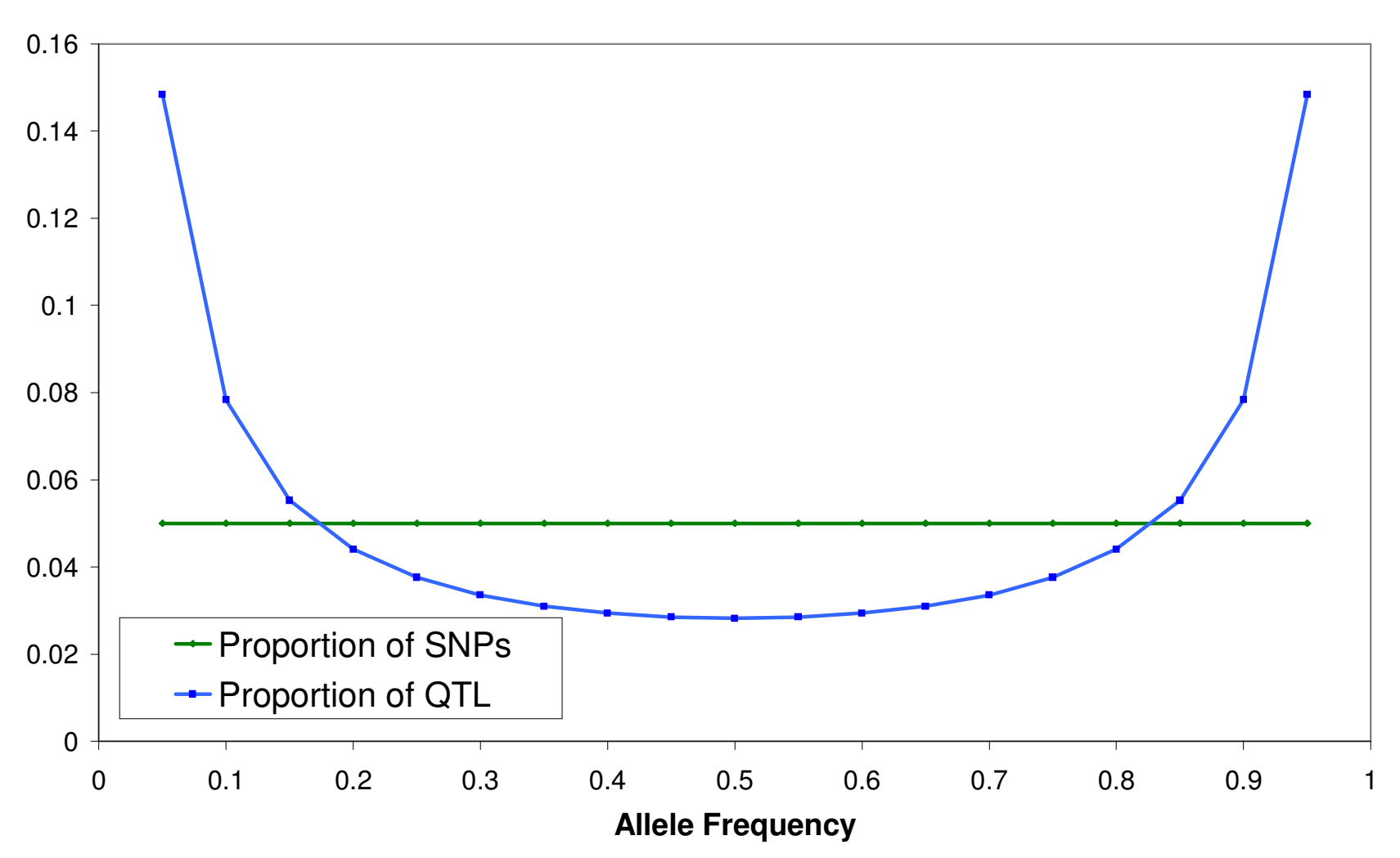
Proportion of QTL variance explained by surrounding markers



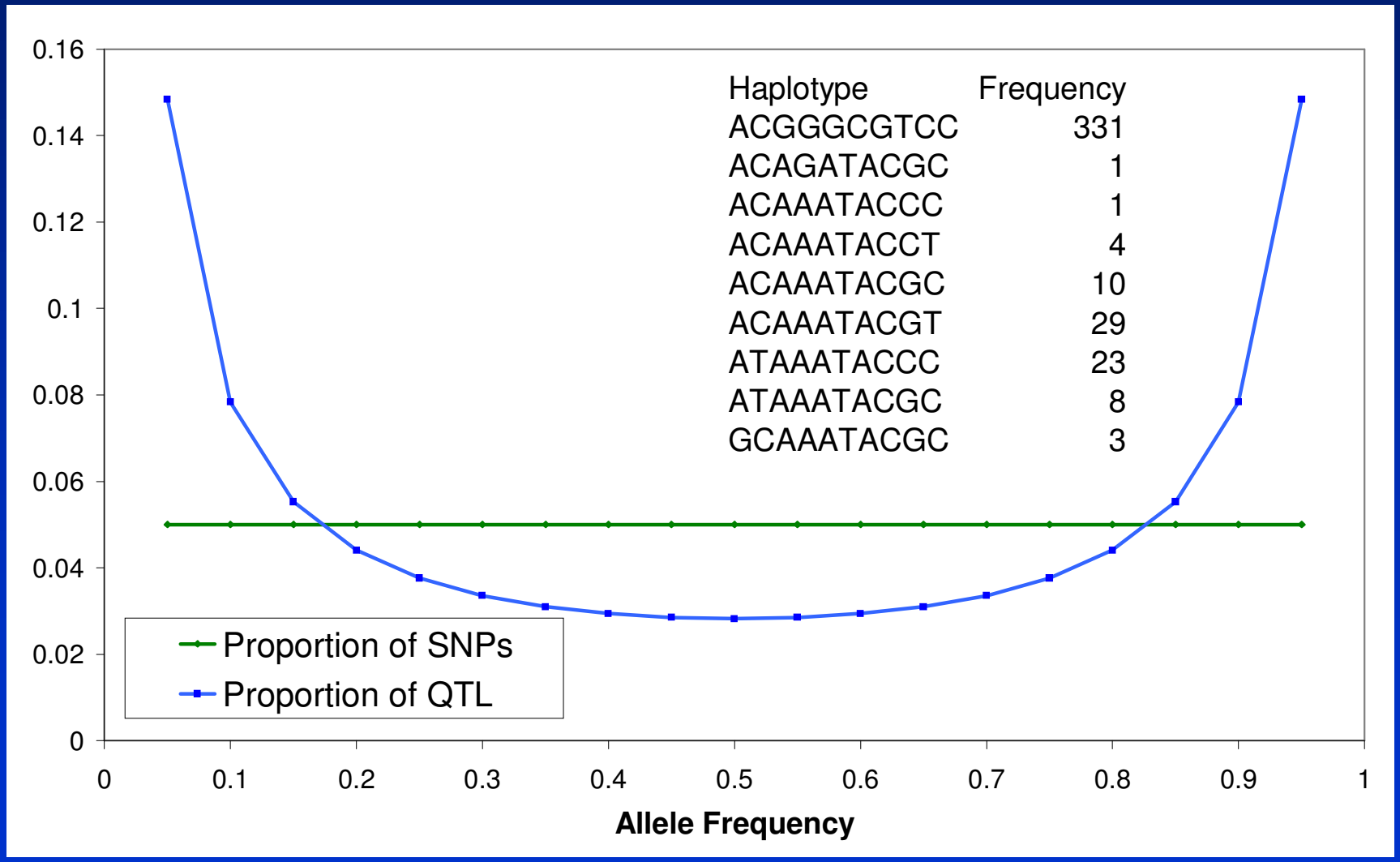
LD mapping with haplotypes

- 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

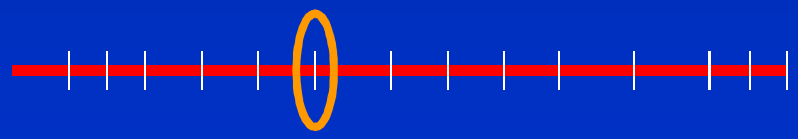
SNP/QTL allele frequency mismatch?



SNP/QTL allele frequency mismatch?



LD mapping with haplotypes



LD mapping with haplotypes

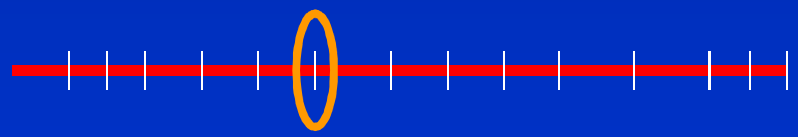
```
Animal_1 1 1 1 1 1 2 1
         1 1 1 1 1 2 1

Animal_2 1 1 1 1 1 2 1
         1 2 1 1 1 2 1

Animal_3 1 2 1 1 1 2 1
         1 2 2 1 1 2 1

Animal_4 2 2 1 1 1 2 1
         2 2 2 1 1 2 1

Animal 5 1 2 1 1 1 2 1
         1 2 2 1 1 2 1
```



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
  - $\lambda_h=\sigma_e^2/\sigma_h^2$

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

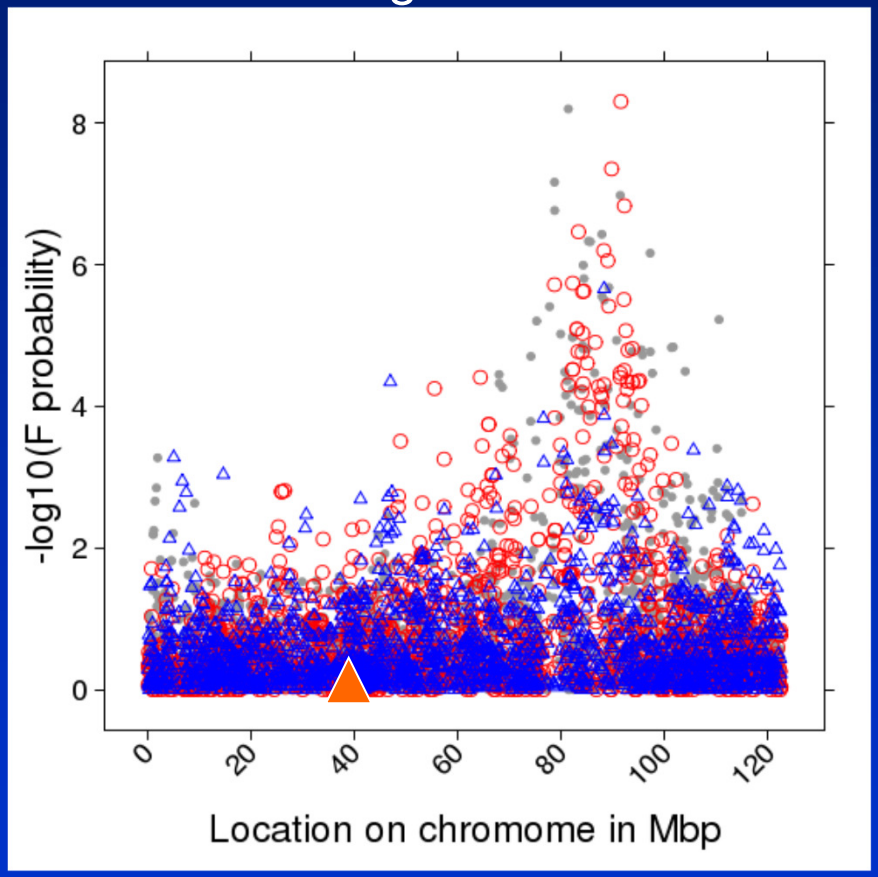
LD mapping with haplotypes

- 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

Single SNPs vs Haplotypes

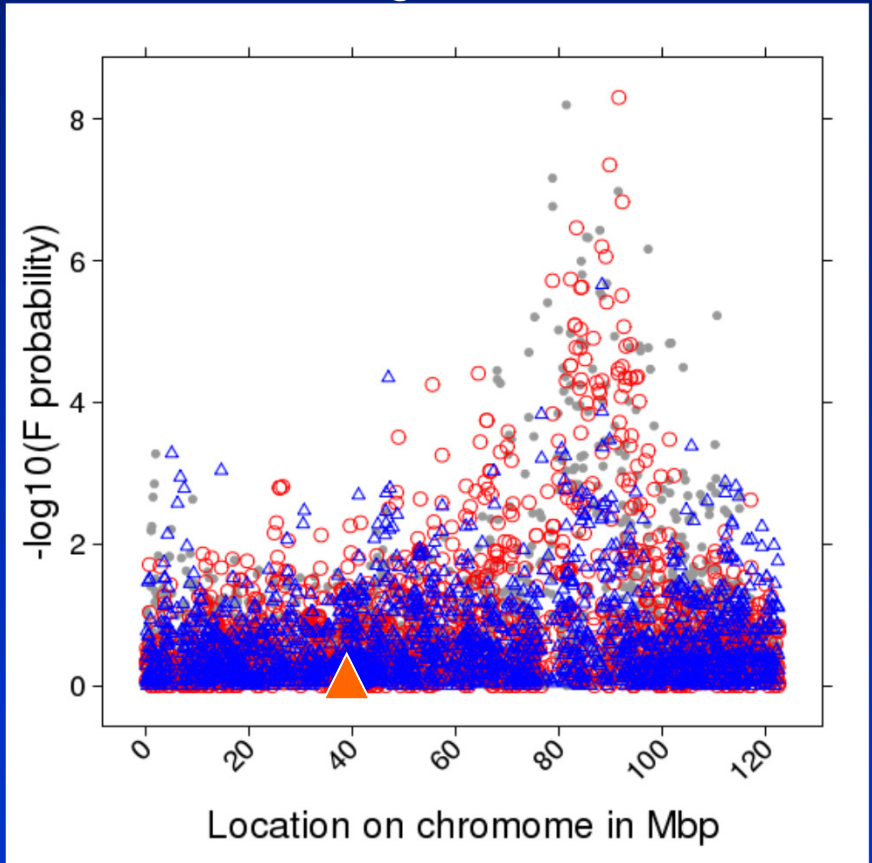
Single SNPs



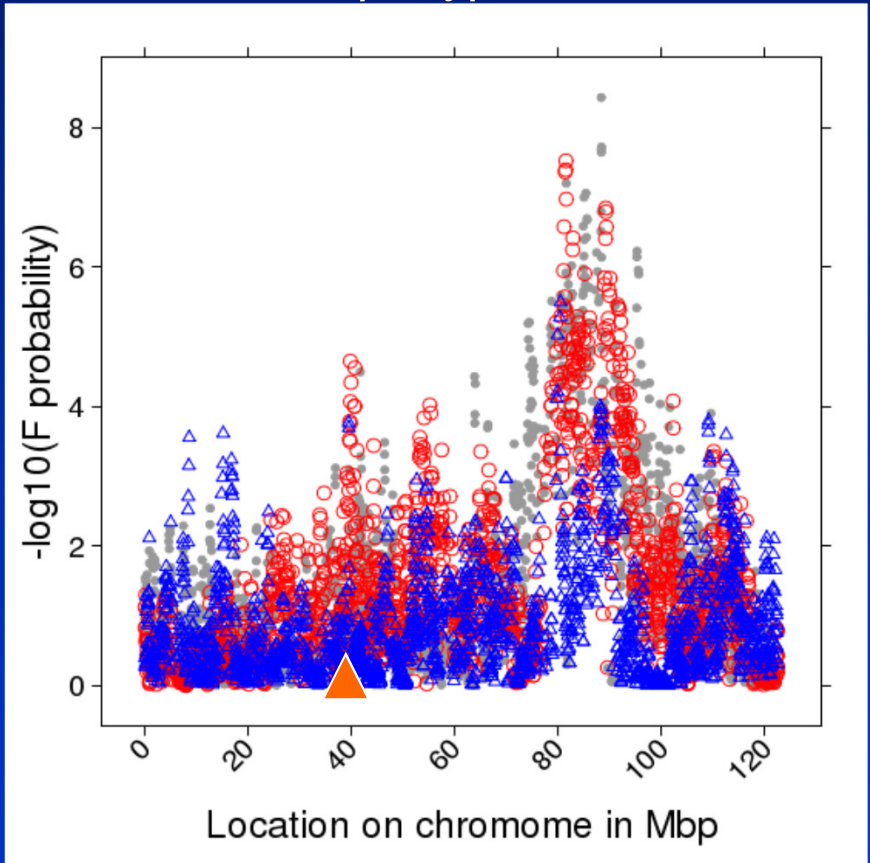
ABCG2

Single SNPs vs Haplotypes

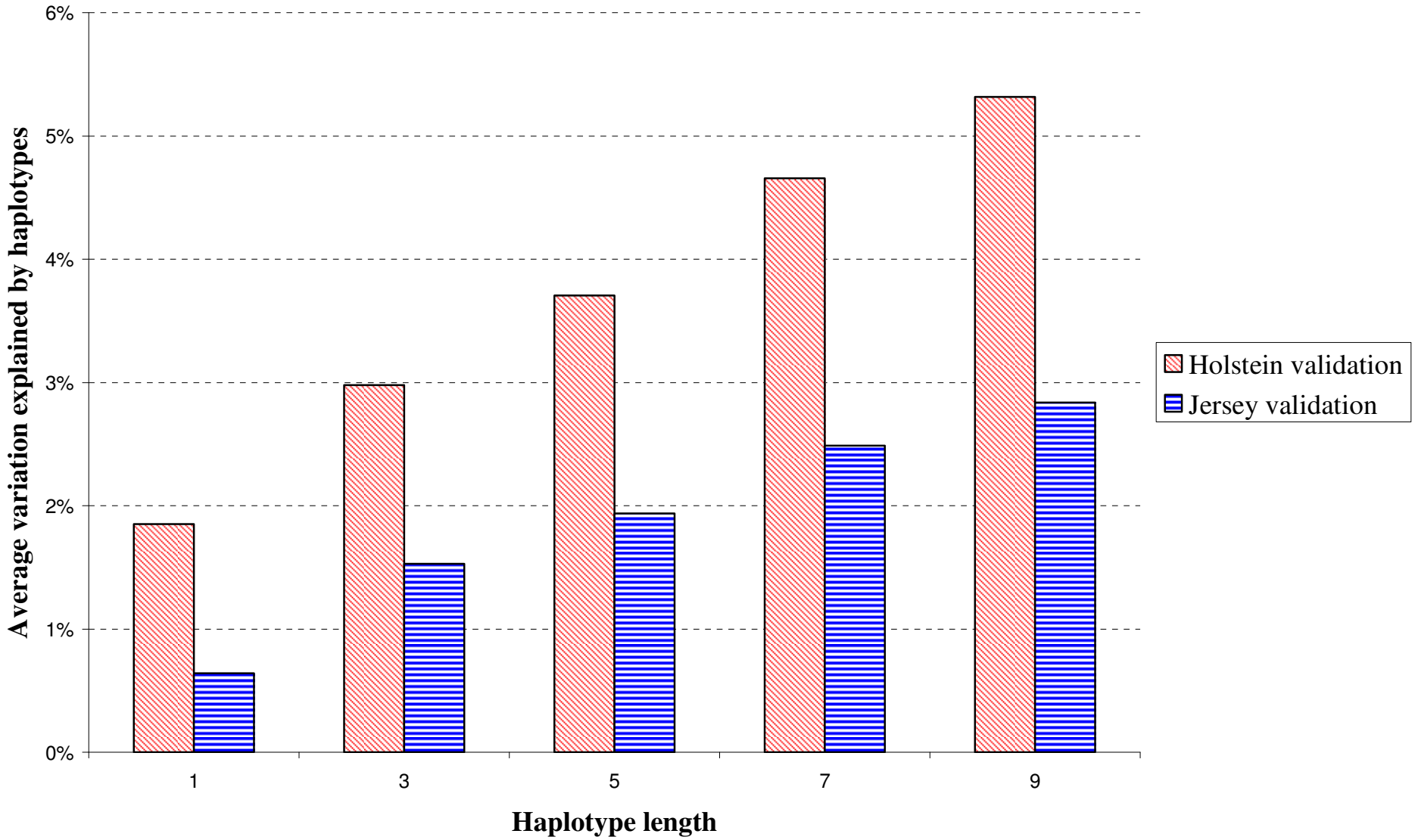
Single SNPs



Haplotypes



▲ ABCG2


6%
5%
4%
3%
2%
1%
0%
Average variation explained by haplotypes
1
3
5
7
9
Haplotype length
Holstein validation
Jersey validation

Genome wide association

- Association testing with single marker regression
- Power of genome wide association studies
- Accounting for population structure
- LD mapping with haplotypes
- Validation

Validation, validation, validation

- Must validate significant associations in ***independent*** population
  - Another breed?
  - Remove false positives
- Design of genome wide association study is ***discovery + validation***
- Make validation set large, limit number of markers to test
  - QTL effects likely to be small
  - Avoid over-estimation of QTL effect due to multiple testing

Genome wide assocation

- Take home points
- Power depends on extent of LD/marker density and number of phenotypic records
  - Knowledge of extent of LD critical
  - Use haplotypes?
- Validation, validation, validation

Course overview

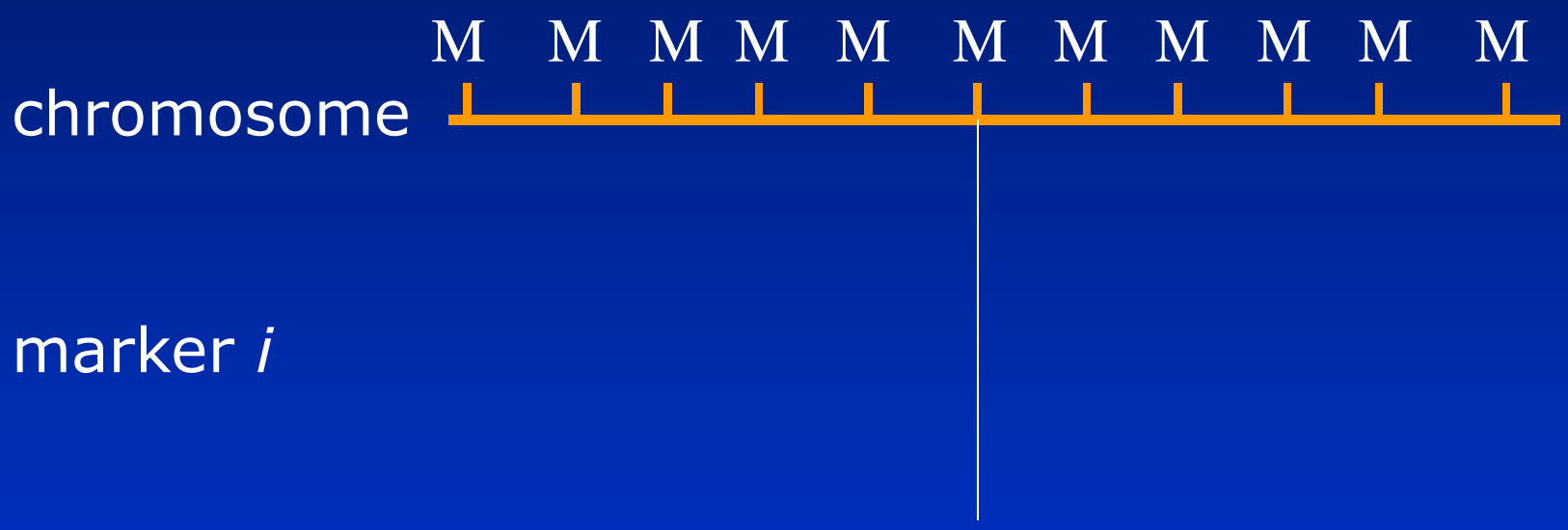
- Day 1
  - Linkage disequilibrium in animal and plant genomes
- Day 2
  - Genome wide association studies
- Day 3
  - Genomic selection
- Day 4
  - Genomic selection
- Day 5
  - Imputation and whole genome sequencing for genomic selection

Genomic selection

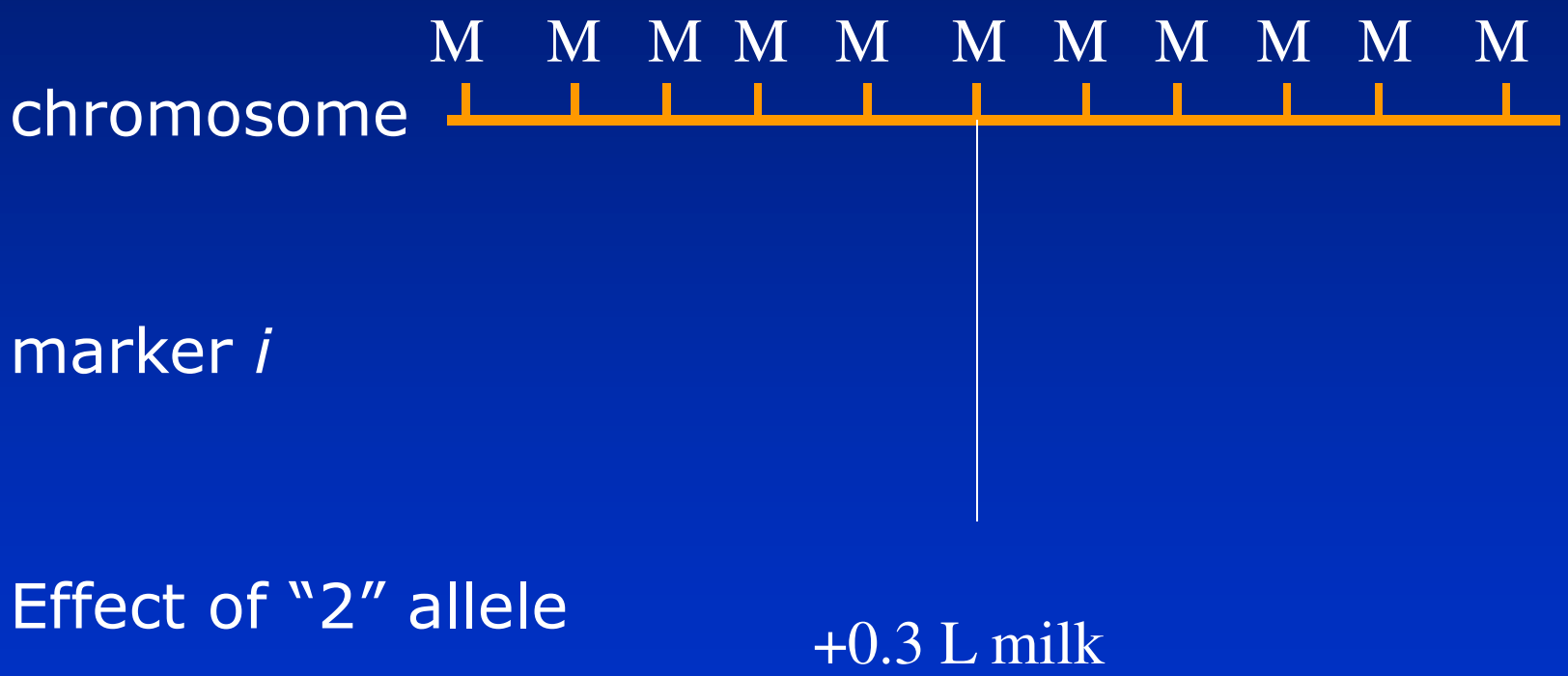
- Problem marker assisted selection is only a proportion of genetic variance is tracked with markers
  - Eg. 10 QTL << 5% of the genetic variance
- Alternative is to trace all segments of the genome with markers
  - Divide genome into chromosome segments based on marker intervals?
  - Capture all QTL = all genetic variance

Genomic selection

Genomic selection



Genomic selection



Genomic selection

Genomic selection

- Predict genomic breeding values as sum of effects over *all* SNP

$$\mathbf{GEBV} = \sum_{i}^{p} \mathbf{X}_i \hat{\mathbf{g}}_i$$

Genomic selection

- Predict genomic breeding values as sum of effects over *all* SNP

$$\mathbf{GEBV} = \sum_{i}^{p} \mathbf{X}_i \hat{\mathbf{g}}_i$$

Number of SNP

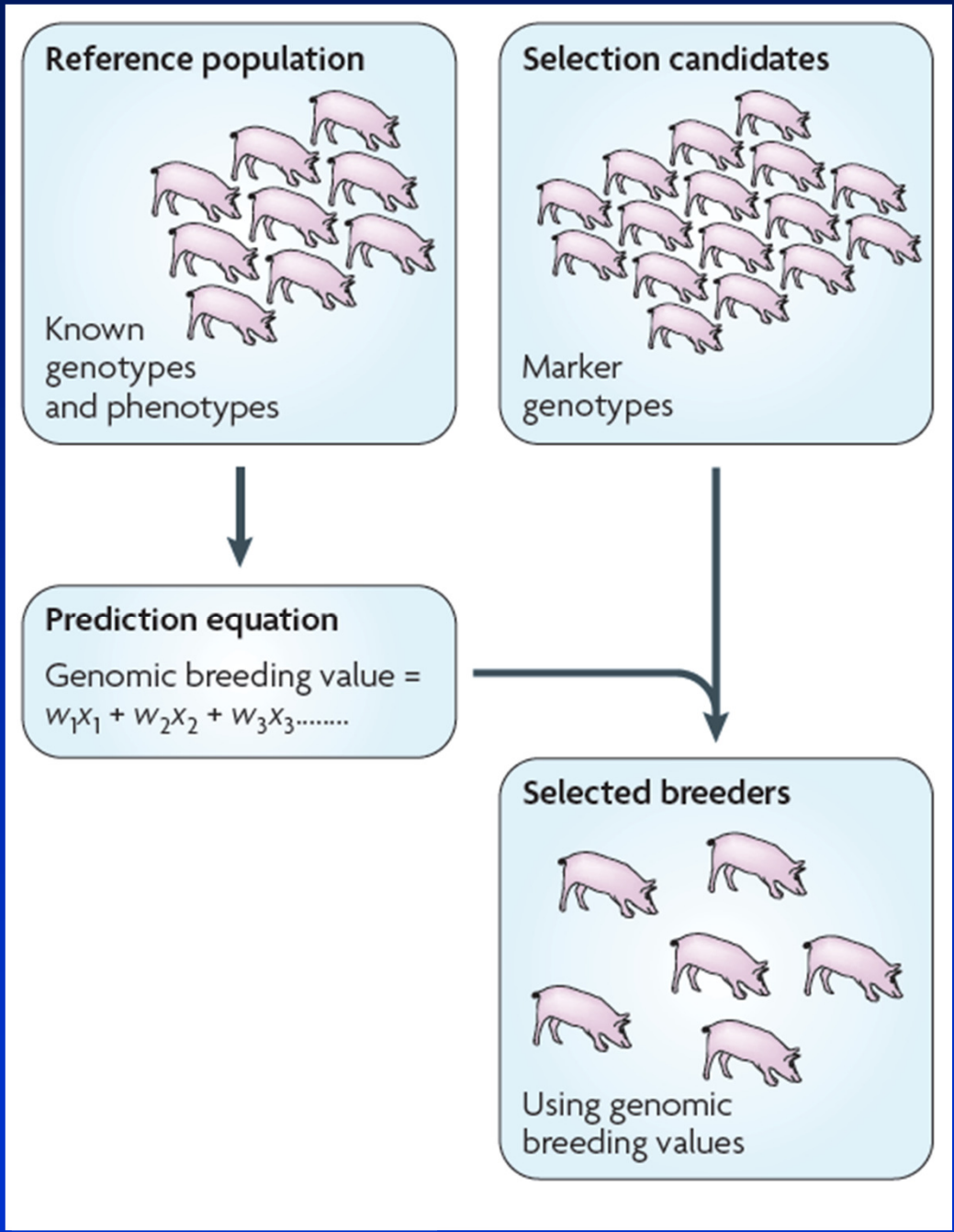
Genomic selection

- Genomic selection exploits linkage disequilibrium
  - Assumption is that markers picking up QTL and will have same effect across the whole population
- Possible within dense marker maps now available

Genomic selection

- Genomic selection avoids bias in estimation of effects due to multiple testing, as all effects fitted simultaneously

Genomic selection



Genomic selection

- First step is to predict the chromosome segment effects in a reference population
- Number of effects >>> than number of records
- Eg. 50,000 SNPs
- From ~ 2000 records?
- Need methods that can deal with this

Genomic selection with Best Linear Unbiased Prediction

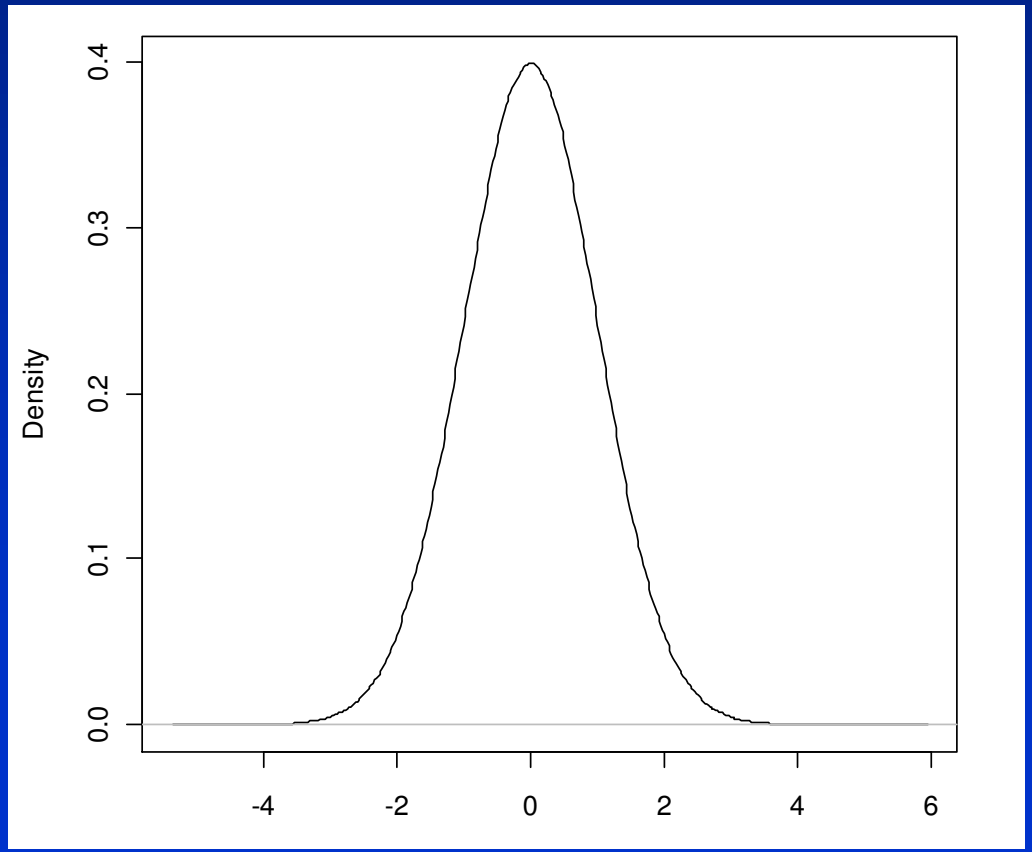
- BLUP = best linear unbiased prediction
- Model:

$$\mathbf{y} = \mu \mathbf{1_n} + \sum_{i=1}^{p} \mathbf{X_i g_i} + \mathbf{e}$$

- In BLUP we assume SNP effects come from normal distribution with same variance $E(\mathbf{g}) \sim N(0,\sigma_g^2)$

Genomic selection with BLUP

- BLUP assumes normal distribution of SNP effects



Genomic selection with BLUP

- **BLUP** = best linear unbiased prediction
- Then we can estimate segment effects as:

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

- $\lambda = \sigma_e^2 / \sigma_g^2$

Genomic selection with BLUP

- Example
- A “simulated” data set
- Single chromosome, with 10 markers
- Phenotypes “simulated”
  - overall mean of 1
  - an effect for SNP 1 of 2 allele of 1
  - normally distributed error term with mean 0 and variance 1.

Genomic selection with BLUP

- Example

		X									
Animal	Y	1	2	3	4	5	6	7	8	9	10
1	0.19	0	0	0	0	0	0	1	2	0	2
2	1.23	1	0	0	1	1	1	2	1	0	1
3	0.86	1	0	0	1	0	0	1	1	1	1
4	1.23	1	1	1	1	0	1	2	1	1	1
5	0.45	0	1	1	1	1	1	2	1	0	1

- 10 SNPs
- Only 5 phenotypic records.

Genomic selection with BLUP

- Example

		X									
Animal	Y		1	2	3	4	5	6	7	8	9	10
1	0.19		0	0	0	0	0	0	1	2	0	2
2	1.23		1	0	0	1	1	1	2	1	0	1
3	0.86		1	0	0	1	0	0	1	1	1	1
4	1.23		1	1	1	1	0	1	2	1	1	1
5	0.45		0	1	1	1	1	1	2	1	0	1

- Assume value of 1 for λ
- $1_n = [1\ 1\ 1\ 1\ 1]$

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

Genomic selection with BLUP

- Example

Mean	0.47
SNP1	0.29
SNP2	-0.05
SNP3	-0.05
SNP4	0.08
SNP5	-0.02
SNP6	0.13
SNP7	0.13
SNP8	-0.08
SNP9	0.11
SNP10	-0.08

Genomic selection with BLUP

- Now we want to predict GEBV for a group of young animals without phenotypes.

$$\mathbf{GEBV} = \mathbf{X}\hat{\mathbf{g}}$$

- We have the g_hat, and we can get **X** from their haplotypes (after genotyping)……

Progeny	X									
1	1	1	1	1	1	1	2	1	0	1
2	1	0	0	1	1	1	2	1	0	1
3	1	0	0	1	1	1	2	1	0	1
4	1	0	0	1	1	1	2	1	0	1
5	0	0	0	0	0	0	1	2	0	2

Genomic selection with BLUP

- GEBV

$$\mathbf{GEBV} = \mathbf{X}\hat{\mathbf{g}}$$

X										$\hat{\mathbf{g}}$		GEBV
1	1	1	1	1	1	2	1	0	1	0.29		0.47
1	0	0	1	1	1	2	1	0	1	-0.05		0.58
1	0	0	1	1	1	2	1	0	1	-0.05		0.58
1	0	0	1	1	1	2	1	0	1	0.08		0.58
0	0	0	0	0	0	1	2	0	2	-0.02		-0.20
										0.13		
										0.13		
										-0.08		
										0.11		
										-0.08		

Genomic selection with BLUP

- Where do we get σ_g^2 from?
- Can estimate total additive genetic variance and divide by number of segments, eg $\sigma_g^2 = \sigma_a^2 / p$
- If using single markers take account of heterozygosity

$$\sigma_g^2 = \sigma_a^2 / 2\sum_{i=1}^{p} q_i(1-q_i)$$

- Ridge regression (Bayesian approach)
- Cross validation

Genomic selection with BLUP

- An equivalent model
- If there are many QTLs whose effects are normally distributed with constant variance,
- Then genomic selection equivalent to replacing the expected relationship matrix with the realised or genomic relationship matrix (**G**) estimated from DNA markers in normal BLUP equations.
  - G_{ij} = proportion of genome that is IBD between animals i and j

Genomic selection with BLUP

- An equivalent model
- Rescale X to account for allele frequencies
  - $w_{ij} = x_{ij} - 2p_j$
- Then breeding values are
  - $\mathbf{v} = \mathbf{W}\mathbf{g}$ ($\mathbf{GEBV} = \mathbf{X}\hat{\mathbf{g}}$)
- And

$$\mathbf{G} = \mathbf{W}\mathbf{W}'/2\sum_{j=1}^{p} p_j(1-p_j)$$

- Then

$$V(\mathbf{v}) = \mathbf{G}\sigma_a^2$$

Genomic selection with BLUP

- An equivalent model

$$\mathbf{y} = \mu \mathbf{1_n} + \mathbf{Zv} + \mathbf{e}$$

$$\begin{bmatrix} \hat{\mu} \\ \hat{\mathbf{v}} \end{bmatrix} = \begin{bmatrix} \mathbf{1_n'1_n} & \mathbf{1_v'Z} \\ \mathbf{Z'1_n} & \mathbf{Z'Z} + \mathbf{G}^{-1}\dfrac{\sigma_e^2}{\sigma_a^2} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{1_n'y} \\ \mathbf{Z'y} \end{bmatrix}$$

Genomic selection with BLUP

- An equivalent model
  - Model 1.

$$\mathbf{y} = \mu \mathbf{1_n} + \sum_{i=1}^{p} \mathbf{X_i g_i} + \mathbf{e}$$

$$\begin{bmatrix} \hat{\mu} \\ \hat{\mathbf{g}} \end{bmatrix} = \begin{bmatrix} \mathbf{1_n'1_n} & \mathbf{1_n'X} \\ \mathbf{X'1_n} & \mathbf{X'X} + \mathbf{I}\dfrac{\sigma_e^2}{\sigma_g^2} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{1_n'y} \\ \mathbf{X'y} \end{bmatrix}$$

$$\mathbf{GEBV} = \mathbf{X}\hat{\mathbf{g}}$$

  - Model 2.

Genomic selection with BLUP

- An equivalent model
  - Model 1.

$$\mathbf{y} = \mu \mathbf{1_n} + \sum_{i=1}^{p} \mathbf{X_i g_i} + \mathbf{e}$$

$$\begin{bmatrix} \hat{\mu} \\ \hat{\mathbf{g}} \end{bmatrix} = \begin{bmatrix} \mathbf{1_n'1_n} & \mathbf{1_n'X} \\ \mathbf{X'1_n} & \mathbf{X'X} + \mathbf{I}\dfrac{\sigma_e^2}{\sigma_g^2} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{1_n'y} \\ \mathbf{X'y} \end{bmatrix}$$

$$\mathbf{GEBV} = \mathbf{X}\hat{\mathbf{g}}$$

  - Model 2.

$$\mathbf{y} = \mu \mathbf{1_n} + \mathbf{Zv} + \mathbf{e}$$

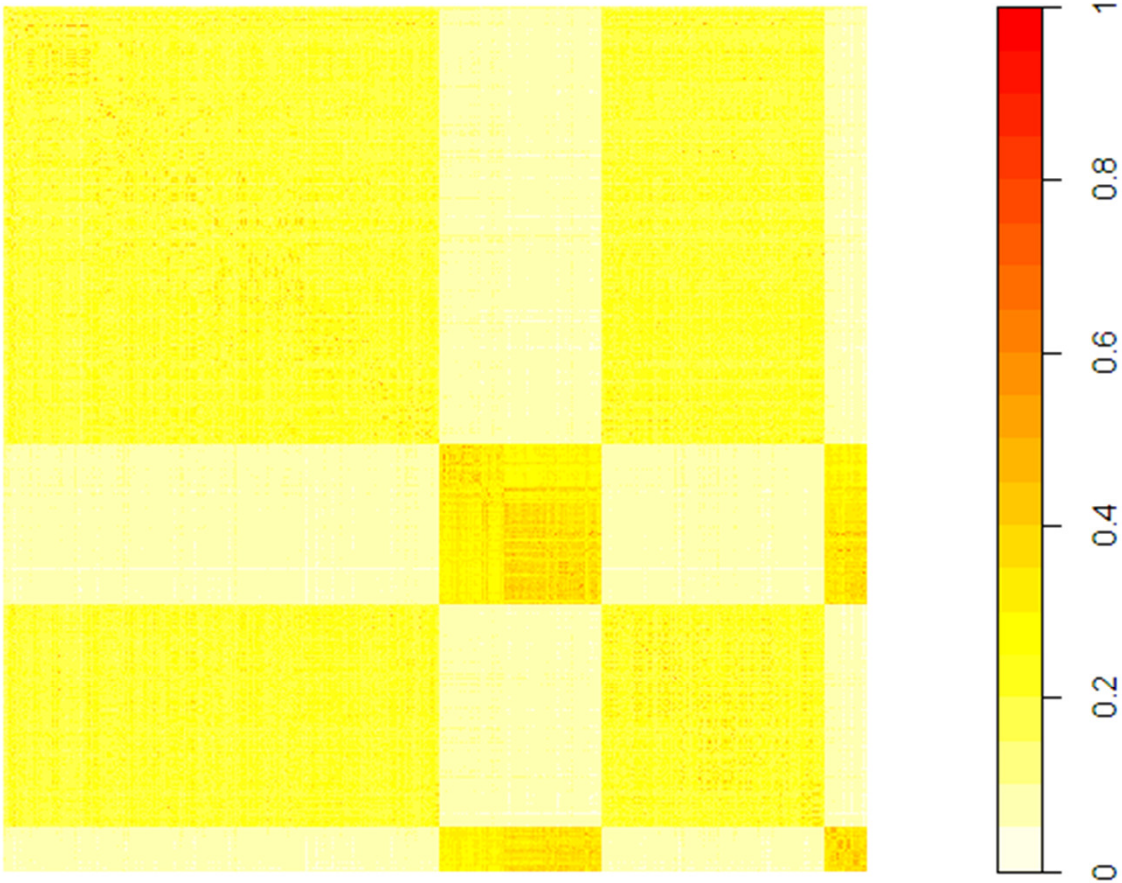
$$\begin{bmatrix} \hat{\mu} \\ \hat{\mathbf{v}} \end{bmatrix} = \begin{bmatrix} \mathbf{1_n'1_n} & \mathbf{1_v'Z} \\ \mathbf{Z'1_n} & \mathbf{Z'Z} + \mathbf{G}^{-1}\dfrac{\sigma_e^2}{\sigma_v^2} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{1_n'y} \\ \mathbf{Z'y} \end{bmatrix}$$

Holstein reference *n = 781*

Jersey reference *n = 287*

Holstein validation *n = 400*

Jersey validation *n = 77*



Genomic selection with BLUP

- An equivalent model
- Why use model 2.
  - If number of markers >>> large than number of animals, more computationally efficient
  - Can be integrated into national evaluations more readily?
  - Calculate accuracy of GEBV from inverse coefficient matrix