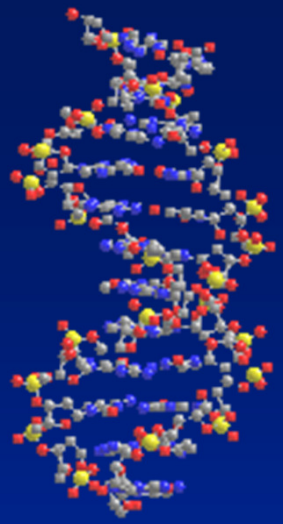
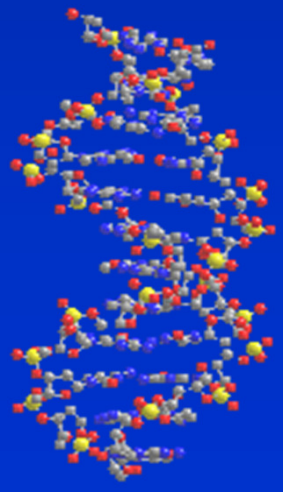


Genome wide assocation and Genomic Selection in the era of Genome sequencing




DEFIER
INFUSE

In crops

nature genetics

Genomic and metabolic prediction of complex heterotic traits in hybrid maize

Christian Riedelsheimer[1], Angelika Czedik-Eysenberg[2], Christoph Grieder[1], Jan Lisec[2], Frank Technow[1], Ronan Sulpice[2], Thomas Altmann[3], Mark Stitt[2], Lothar Willmitzer[2,4] & Albrecht E Melchinger[1]

Maize is both an exciting model organism in plant genetics and also the most important crop worldwide for food, animal feed and bioenergy production. Recent genome-wide association and metabolic profiling studies aimed to resolve quantitative traits to their causal genetic loci and key metabolic regulators.

known as heterosis[13]. Breeders, therefore, judge an inbred line not by its performance *per se* but by its potential to create superior hybrids. Unfortunately, line performance is only a weak predictor of hybrid performance[14], rendering phenotypic evaluation of crosses still inevitable.

A powerful approach for solving the first problem emerged in animal

Course overview

- Day 1
  - Quantitative traits
  - Linkage disequilibrium
  - Genome wide association studies
- Day 2 and 3
  - Genomic prediction - BLUP and GBLUP
  - Genomic prediction – Bayesian methods
- Day 4
  - Validation of genomic predictions
  - Optimal breeding program design with genomic selection
- Day 5
  - Imputation and whole genome sequencing for genomic selection

Day 1

- Introduction to Quantitative traits
- Linkage disequilibrium
- Genome wide association studies

Quantitative traits

- How to explain the genetic variation observed for many of the traits of economic importance in livestock and plant species?

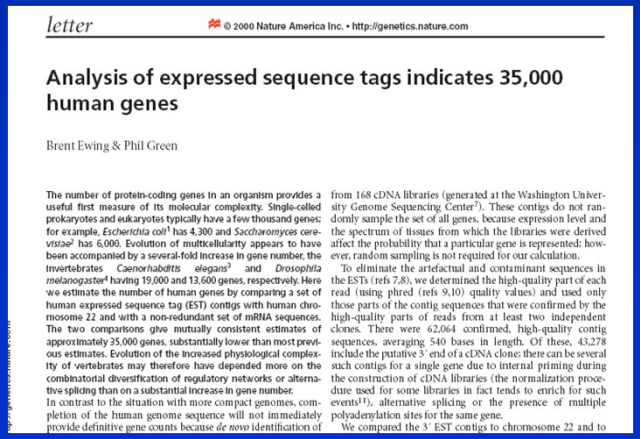
Two models……

- Infinitesimal model:
  - assumes that traits are determined by an infinite number of unlinked and additive loci, each with an infinitesimally small effect
  - This model the foundation of animal breeding theory including breeding value prediction
  - Spectacularly successful in many cases!

Farmed Atlantic salmon: Feed conversion efficiency has improved by 140% in seven generations

Two models……

- *vs* the Finite loci model…..
  - But while the infinitesimal model is very useful assumption,
  - there is a finite amount of genetic material
  - With a finite number of genes……
  - Define any gene that contributes to variation in a quantitative/economic trait as quantitative trait loci (QTL)
- A key question is *what is the distribution of the effects of QTL for a typical quantitative trait* ?



letter © 2000 Nature America Inc. • http://genetics.nature.com

Analysis of expressed sequence tags indicates 35,000 human genes

Brent Ewing & Phil Green

The number of protein-coding genes in an organism provides a useful first measure of its molecular complexity. Single-celled prokaryotes and eukaryotes typically have a few thousand genes; for example, *Escherichia coli*[1] has 4,300 and *Saccharomyces cerevisiae*[2] has 6,000. Evolution of multicellularity appears to have been accompanied by a several-fold increase in gene number; the invertebrates *Caenorhabditis elegans*[3] and *Drosophila melanogaster*[4] having 19,000 and 13,600 genes, respectively. Here we estimate the number of human genes by comparing a set of human expressed sequence tag (EST) contigs with human chromosome 22 and with a non-redundant set of mRNA sequences. The two comparisons give mutually consistent estimates of approximately 35,000 genes, substantially lower than most previous estimates. Evolution of the increased physiological complexity of vertebrates may therefore have depended more on the combinatorial diversification of regulatory networks or alternative splicing than on a substantial increase in gene number.

In contrast to the situation with more compact genomes, completion of the human genome sequence will not immediately provide definitive gene counts because *de novo* identification of

from 168 cDNA libraries (generated at the Washington University Genome Sequencing Center[7]). These contigs do not randomly sample the set of all genes, because expression level and the spectrum of tissues from which the libraries were derived affect the probability that a particular gene is represented; however, random sampling is not required for our calculation.

To eliminate the artefactual and contaminant sequences in the ESTs (refs 7,8), we determined the high-quality part of each read (using phred (refs 9,10) quality values) and used only those parts of the contig sequences that were confirmed by the high-quality parts of reads from at least two independent clones. There were 62,064 confirmed, high-quality contig sequences, averaging 540 bases in length. Of these, 43,278 include the putative 3′ end of a cDNA clone; there can be several such contigs for a single gene due to internal priming during the construction of cDNA libraries (the normalization procedure used for some libraries in fact tends to enrich for such events[11]), alternative splicing or the presence of multiple polyadenylation sites for the same gene.

We compared the 3′ EST contigs to chromosome 22 and to

Human height

NATURE | LETTER

◂ previous article next article ▸

Hundreds of variants clustered in genomic loci and biological pathways affect human height

Hana Lango Allen, Karol Estrada, Guillaume Lettre, Sonja I. Berndt, Michael N. Weedon, Fernando Rivadeneira, Cristen J. Willer, Anne U. Jackson, Sailaja Vedantam, Soumya Raychaudhuri, Teresa Ferreira, Andrew R. Wood, Robert J. Weyant, Ayellet V. Segrè, Elizabeth K. Speliotes, Eleanor Wheeler, Nicole Soranzo, Ju-Hyun Park, Jian Yang, Daniel Gudbjartsson, Nancy L. Heard-Costa, Joshua C. Randall, Lu Qi, Albert Vernon Smith, Reedik Mägi + *et al.*

Affiliations | Contributions | Corresponding authors

Nature **467**, 832–838 (14 October 2010) | doi:10.1038/nature09410
Received 23 April 2010 | Accepted 28 July 2010 | Published online 29 September 2010

Most common human traits and diseases have a polygenic pattern of inheritance: DNA sequence variants at many genetic loci influence the phenotype. Genome-wide association (GWA) studies have identified more than 600 variants associated with human traits[1], but these typically explain small fractions of phenotypic variation, raising questions about the use of further studies. Here, using 183,727 individuals, we show that hundreds of genetic variants, in at least 180 loci, influence adult height, a highly heritable and classic polygenic trait[2, 3]. The large number of loci reveals patterns with important implications for genetic studies of common human diseases and traits. First, the 180 loci are not random, but instead are enriched for genes

日本語要約
print
email
download citation
order reprints
rights and permissions
share/bookmark

Human height

NATURE | LETTER

previous article next article

Hundreds of variants clustered in genomic loci and biological pathways affect human height

Hana Lango Allen, Karol Estrada, Guillaume Lettre, Sonja I. Berndt, Michael N. Weedon, Fernando Rivadeneira, Cristen J. Willer, Anne U. Jackson, Sailaja Vedantam, Soumya Raychaudhuri, Teresa Ferreira, Andrew R. Wood, Robert J. Weyant, Ayellet V. Segrè, Elizabeth K. Speliotes, Eleanor Wheeler, Nicole Soranzo, Ju-Hyun Park, Jian Yang, Daniel Gudbjartsson, Nancy L. Heard-Costa, Joshua C. Randall, Lu Qi, Albert Vernon Smith, Reedik Mägi + *et al.*

Affiliations | Contributions | Corresponding authors

Nature **467**, 832–838 (14 October 2010) | doi:10.1038/nature09410

Received 23 Ap

180 loci explain 10% of the variance

Most common h

inheritance: DNA sequence variants at many genetic loci influence the phenotype. Genome-wide association (GWA) studies have identified more than 600 variants associated with human traits[1], but these typically explain small fractions of phenotypic variation, raising questions about the use of further studies. Here, using 183,727 individuals, we show that hundreds of genetic variants, in at least 180 loci, influence adult height, a highly heritable and classic polygenic trait[2, 3]. The large number of loci reveals patterns with important implications for genetic studies of common human diseases and traits. First, the 180 loci are not random, but instead are enriched for genes

Yield in Rice

nature genetics

Genome-wide association studies of 14 agronomic traits in rice landraces

Xuehui Huang[1,2,10], Xinghua Wei[3,10], Tao Sang[4,10], Qiang Zhao[1,2,10], Qi Feng[1,10], Yan Zhao[1], Canyang Li[1], Chuanrang Zhu[1], Tingting Lu[1], Zhiwu Zhang[5], Meng Li[5,6], Danlin Fan[1], Yunli Guo[1], Ahong Wang[1], Lu Wang[1], Liuwei Deng[1], Wenjun Li[1], Yiqi Lu[1], Qijun Weng[1], Kunyan Liu[1], Tao Huang[1], Taoying Zhou[1], Yufeng Jing[1], Wei Li[1], Zhang Lin[1], Edward S Buckler[5,7], Qian Qian[3], Qi-Fa Zhang[8], Jiayang Li[9] & Bin Han[1,2]

Uncovering the genetic basis of agronomic traits in crop landraces that have adapted to various agro-climatic conditions is important to world food security. Here we have identified ~3.6 million SNPs by sequencing 517 rice landraces and constructed a high-density haplotype map of the rice genome using a novel data-imputation method. We performed genome-wide association studies (GWAS) for 14 agronomic traits in the population of *Oryza sativa indica* subspecies. The loci identified through GWAS explained ~36% of the phenotypic variance, on average. The peak signals at six loci were tied closely to previously identified genes. This study provides a fundamental resource for rice genetics research and breeding, and demonstrates that an approach integrating second-generation genome sequencing and GWAS can be used as a powerful complementary strategy to classical biparental cross-mapping for dissecting complex traits in rice.

Yield in Rice

nature genetics

"our results suggest that multiple loci with relatively small effects contribute to the phenotypic variance"

Genome-wide association studies of 14 agronomic traits in rice landraces

Xuehui Huang[1,2,10], Xinghua Wei[3,10], Tao Sang[4,10], Qiang Zhao[1,2,10], Qi Feng[1,10], Yan Zhao[1], Canyang Li[1], Chuanrang Zhu[1], Tingting Lu[1], Zhiwu Zhang[5], Meng Li[5,6], Danlin Fan[1], Yunli Guo[1], Ahong Wang[1], Lu Wang[1], Liuwei Deng[1], Wenjun Li[1], Yiqi Lu[1], Qijun Weng[1], Kunyan Liu[1], Tao Huang[1], Taoying Zhou[1], Yufeng Jing[1], Wei Li[1], Zhang Lin[1], Edward S Buckler[5,7], Qian Qian[3], Qi-Fa Zhang[8], Jiayang Li[9] & Bin Han[1,2]

Uncovering the genetic basis of agronomic traits in crop landraces that have adapted to various agro-climatic conditions is important to world food security. Here we have identified ~3.6 million SNPs by sequencing 517 rice landraces and constructed a high-density haplotype map of the rice genome using a novel data-imputation method. We performed genome-wide association studies (GWAS) for 14 agronomic traits in the population of *Oryza sativa indica* subspecies. The loci identified through GWAS explained ~36% of the phenotypic variance, on average. The peak signals at six loci were tied closely to previously identified genes. This study provides a fundamental resource for rice genetics research and breeding, and demonstrates that an approach integrating second-generation genome sequencing and GWAS can be used as a powerful complementary strategy to classical biparental cross-mapping for dissecting complex traits in rice.

The case of the missing heritability

When scientists opened up the human genome, they expected to find the genetic components of common traits and diseases. But they were nowhere to be seen. **Brendan Maher** shines a light on six places where the missing loot could be stashed away.

If you want to predict how tall your children might one day be, a good bet would be to look in the mirror, and at your mate. Studies going back almost a century have

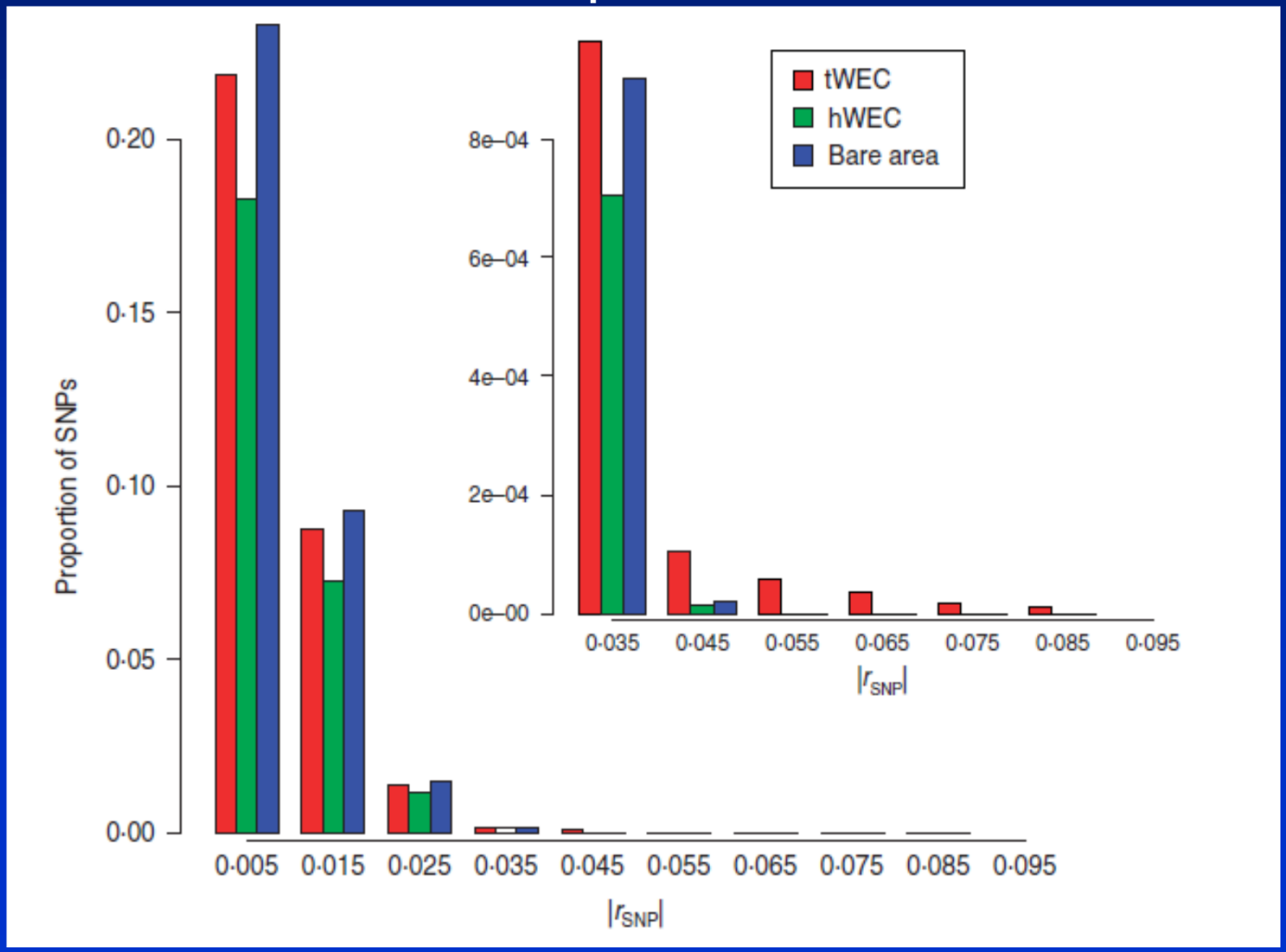
Even though these genome-wide association studies (GWAS) turned up dozens of variants, they did "very little of the prediction that you would do just by asking people how tall their parents are", says Joel Hirschhorn at the Broad Institute in Cambridge, Massachusetts, who led one of the studies[2].

contribute to a variety of traits and common diseases. But even when dozens of genes have been linked to a trait, both the individual and cumulative effects are disappointingly small and nowhere near enough to explain earlier estimates of heritability. "It is the big topic in the genetics of common disease right

ILLUSTRATIONS BY D. PARKINS

Distribution of QTL effects

- Distribution of effects for parasite resistance and bare breech area in sheep

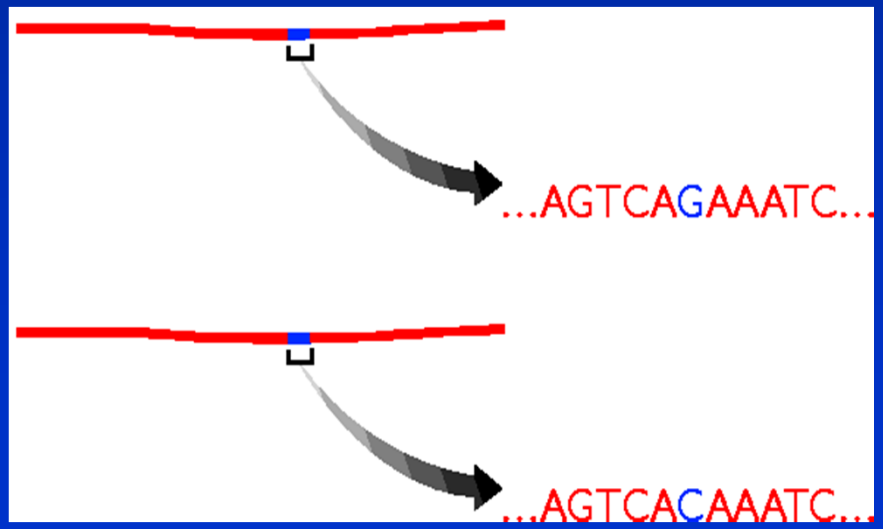


Quantitative traits

- Large number of causative mutations (quantitative trait loci, QTL) for most complex traits
- Variance explained by individual markers will be small
- Genome wide association studies -> powerful experiments!
- Genomic prediction -> Use large numbers of DNA markers to simultaneously track all QTL

The Revolution

- As a result of sequencing animal and plant genomes, have a huge amount of information on variation in the genome
  - at the DNA level
- Most abundant form of variation are Single Nucleotide Polymorphisms (SNPs)





- 1000 Genomes project (Pilot)
- ~15 mill SNPs
- ~7 mill SNPs with minor allele >5%
- ~100,000-300,000 cSNPs
- ~50,000 nonsynonymous cSNPs -> change protein structure
- Every individual carries 250-300 loss of function mutations!

The Revolution

- SNP chips available for
  - Sheep, Cattle (50K, 800K), Pigs,
  - Chickens
  - Salmon
  - Horse, Dog
- Plants
  - Maize, Wheat
  - Cotton, Soybean under development
- Cost?
  - ~ $100-200 USD for 60K SNPs
- Genotyping by re-sequencing?
  - 40 million SNPs in cattle
  - Insertion deletions
  - Copy number variants?

Aim

- Provide you with genome wide association and genomic prediction methodologies to exploit high density SNP genotypes in livestock and plant improvement

Day 1

- Introduction to Quantitative traits
- Linkage disequilibrium
- Introduction to Genomic Prediction
- Genomic prediction with BLUP

Linkage disequilibrium

- Measuring linkage disequilibrium
- Causes of LD
- Extent of LD in animals and plants

Definitions of LD

- Why do we need to define and measure LD?
- Genomic prediction assumes markers are in LD with QTL
- Determine the number of markers required for genomic prediction

Definitions of LD

- Classical definition:
  - Two markers A and B on the same chromosome
  - Alleles are
    - marker A A1, A2
    - marker B B1, B2
  - Possible haploptypes are A1_B1, A1_B2, A2_B1, A2_B2

Definitions of LD

Linkage equilibrium…….

		Marker A		
		A1	A2	Frequency
Marker B	B1			0.5
	B2			0.5
	Frequency	0.5	0.5	

Definitions of LD

Linkage equilibrium…….

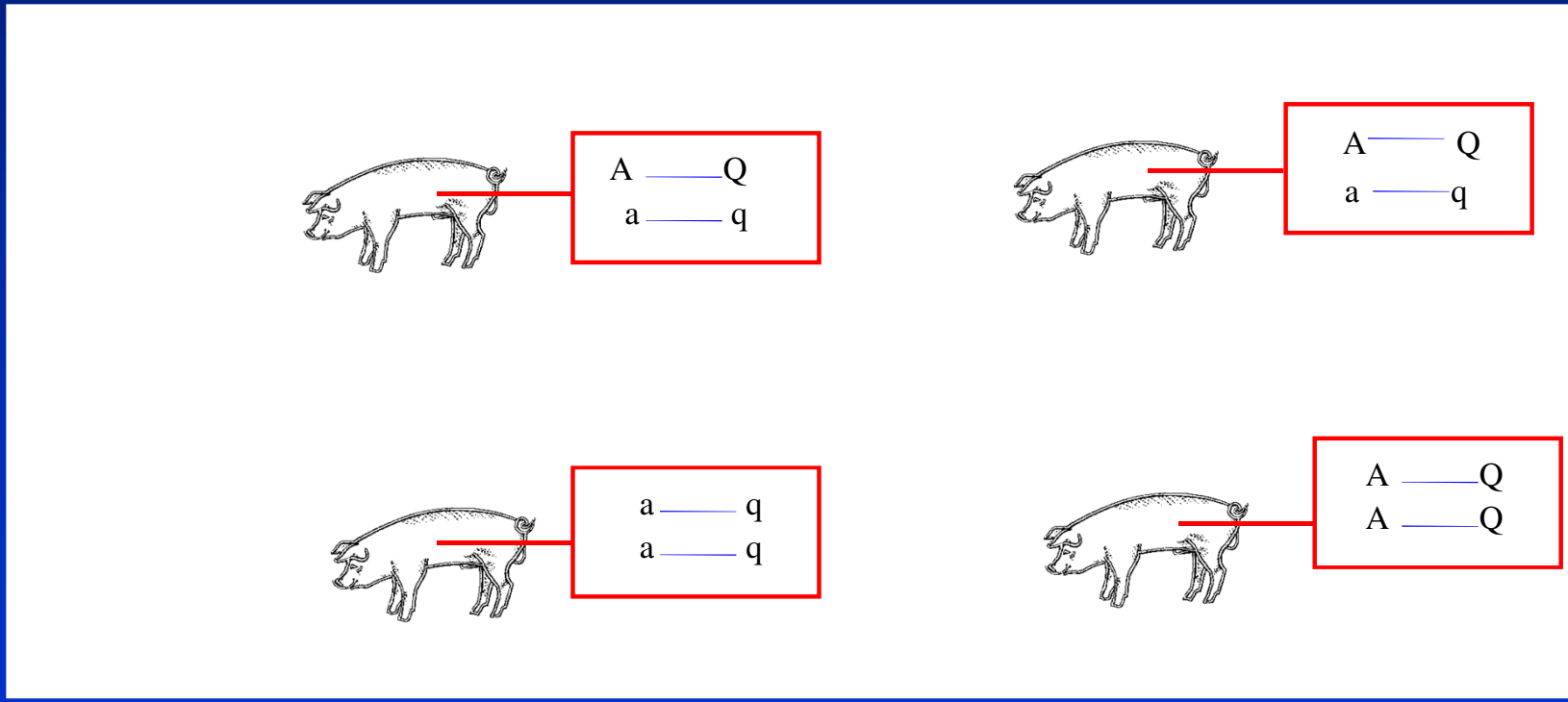
		Marker A		
		A1	A2	Frequency
Marker B	B1	0.25	0.25	0.5
	B2	0.25	0.25	0.5
	Frequency	0.5	0.5	

Definitions of LD

Linkage disequilibrium……..

		Marker A		
		A1	A2	Frequency
Marker B	B1	0.4	0.1	0.5
	B2	0.1	0.4	0.5
	Frequency	0.5	0.5	

- Linkage disequilibrium between marker and QTL



Definitions of LD

Linkage disequilibrium……..

		Marker A		
		A1	A2	Frequency
Marker B	B1	0.4	0.1	0.5
	B2	0.1	0.4	0.5
	Frequency	0.5	0.5	

D = freq(A1_B1)*freq(A2_B2)-freq(A1_B2)*freq(A2_B1)

= 0.4 * 0.4 - 0.1 * 0.1

= 0.15

Definitions of LD

- Measuring the extent of LD (determines how dense markers need to be for LD mapping)

  D = freq(A1_B1)*freq(A2_B2)-freq(A1_B2)*freq(A2_B1)

  – highly dependent on allele frequencies
    - not suitable for comparing LD at different sites

  $\mathbf{r^2}=D^2/[freq(A1)*freq(A2)*freq(B1)*freq(B2)]$

Definitions of LD

Linkage disequilibrium……..

		Marker A		
		A1	A2	Frequency
Marker B	B1	0.4	0.1	0.5
	B2	0.1	0.4	0.5
	Frequency	0.5	0.5	

D = 0.15

$r^2 = D^2/[freq(A1)*freq(A2)*freq(B1)*freq(B2)]$

$r^2 = 0.15^2/[0.5*0.5*0.5*0.5]$

= 0.36

Definitions of LD

- Measuring extent of LD
  - determines how dense markers need to be for LD mapping

  D = freq(A1_B1)*freq(A2_B2)-freq(A1_B2)*freq(A2_B1)
  - highly dependent on allele frequencies
    - not suitable for comparing LD at different sites

  $\mathbf{r^2}=D^2/[freq(A1)*freq(A2)*freq(B1)*freq(B2)]$

  Values between 0 and 1.

Definitions of LD

- If one loci is a marker and the other is QTL
- The r^2 between a marker and a QTL is the *proportion of QTL variance which can be observed at the marker*
  - eg if variance due to a QTL is 200kg^2, and r^2 between marker and QTL is 0.2, variation observed at the marker is 40kg^2.

Definitions of LD

- If one loci is a marker and the other is QTL
- The r^2 between a marker and a QTL is the *proportion of QTL variance which can be observed at the marker*
  - eg if variance due to a QTL is 200kg^2, and r^2 between marker and QTL is 0.2, variation observed at the marker is 40kg^2.

Linkage disequilibrium

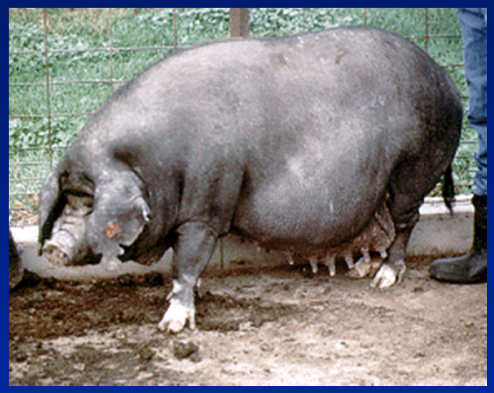
- Measuring linkage disequilibrium
- Causes of LD
- Extent of LD in animals and plants

Causes of LD

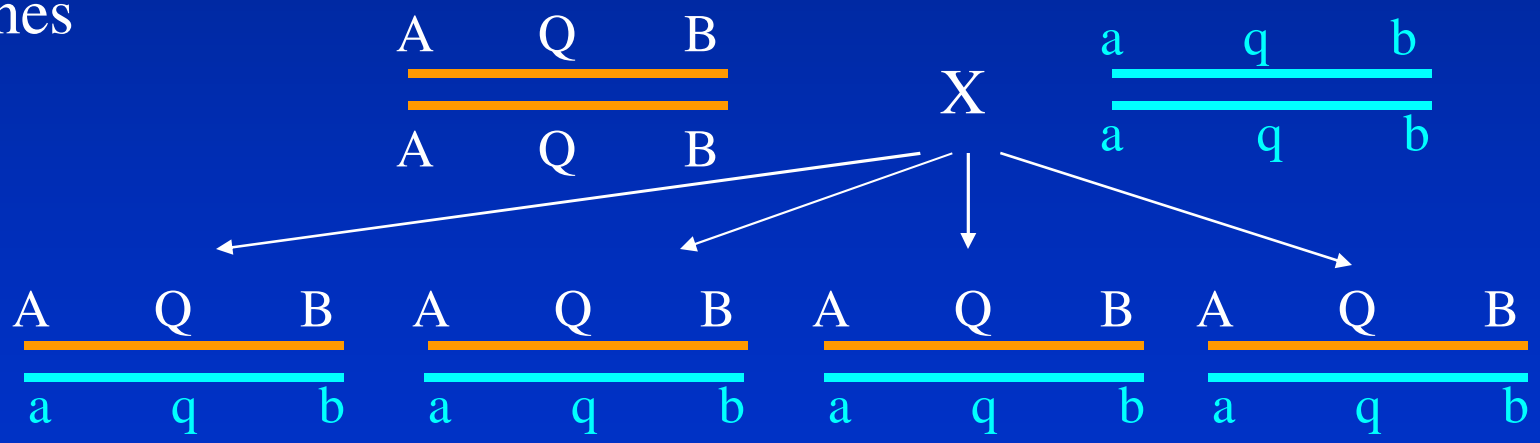
- Migration
  - LD artificially created in crosses
    - large when crossing inbred lines
    - but small when crossing breeds that do not differ markedly in gene frequencies
    - disappears after only a limited number of generations

- F2 design

X

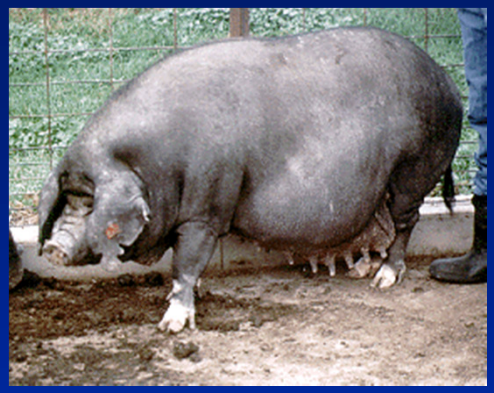


Parental Lines

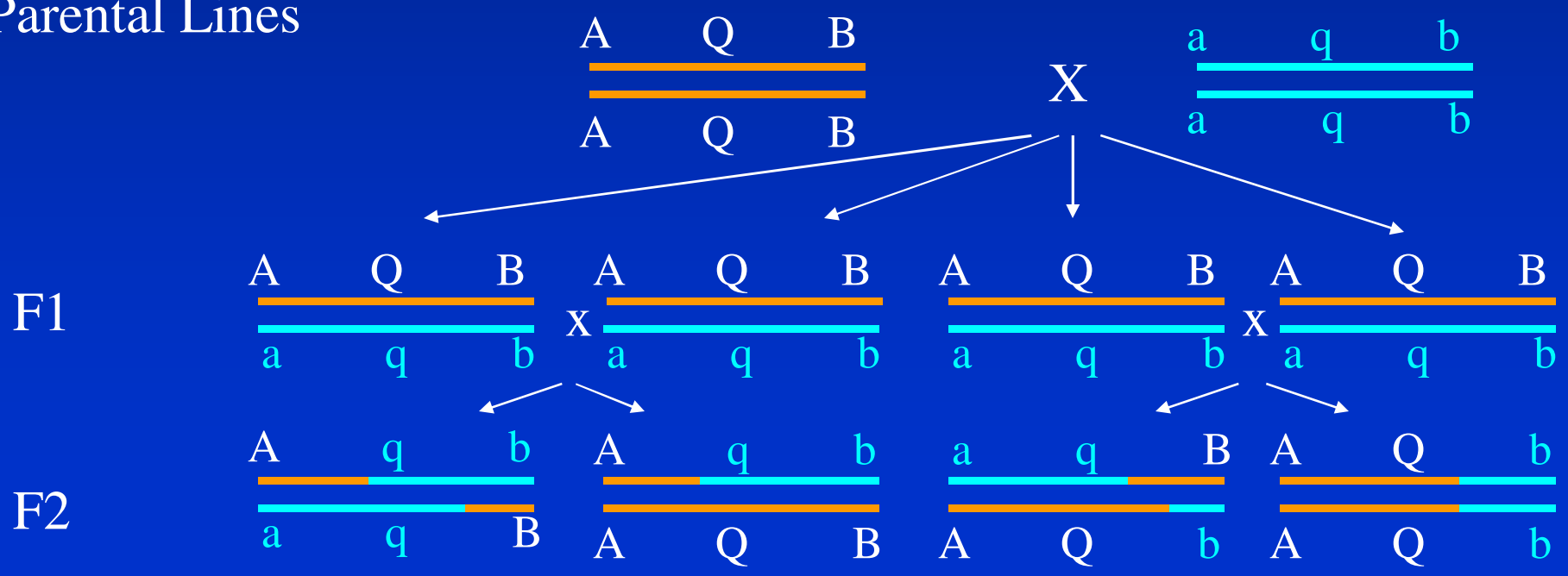


- F2 design

X



Parental Lines



Causes of LD

- Migration
  - LD artificially created in crosses designs
    - large when crossing inbred lines
    - but small when crossing breeds that do not differ markedly in gene frequencies
    - disappears after only a limited number of generations
- Selection
  - Selective sweeps

Generation 1

A_____q
A_____q
a_____q

A_____q
a_____q
a_____q

Generation 2

Generation 3

Mutation

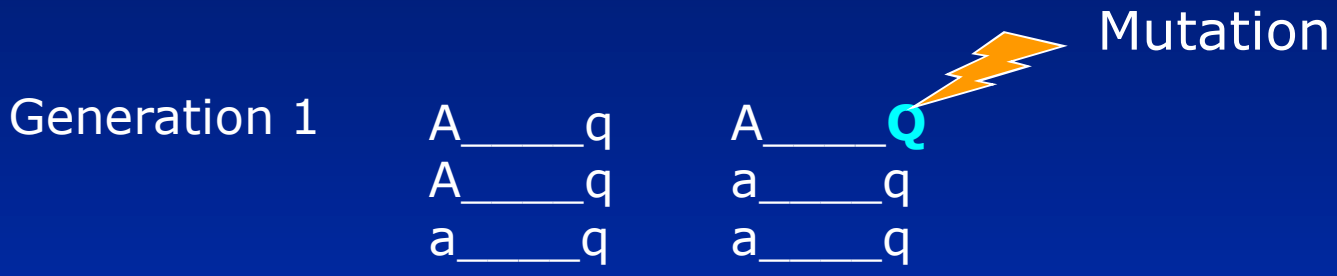
Generation 1

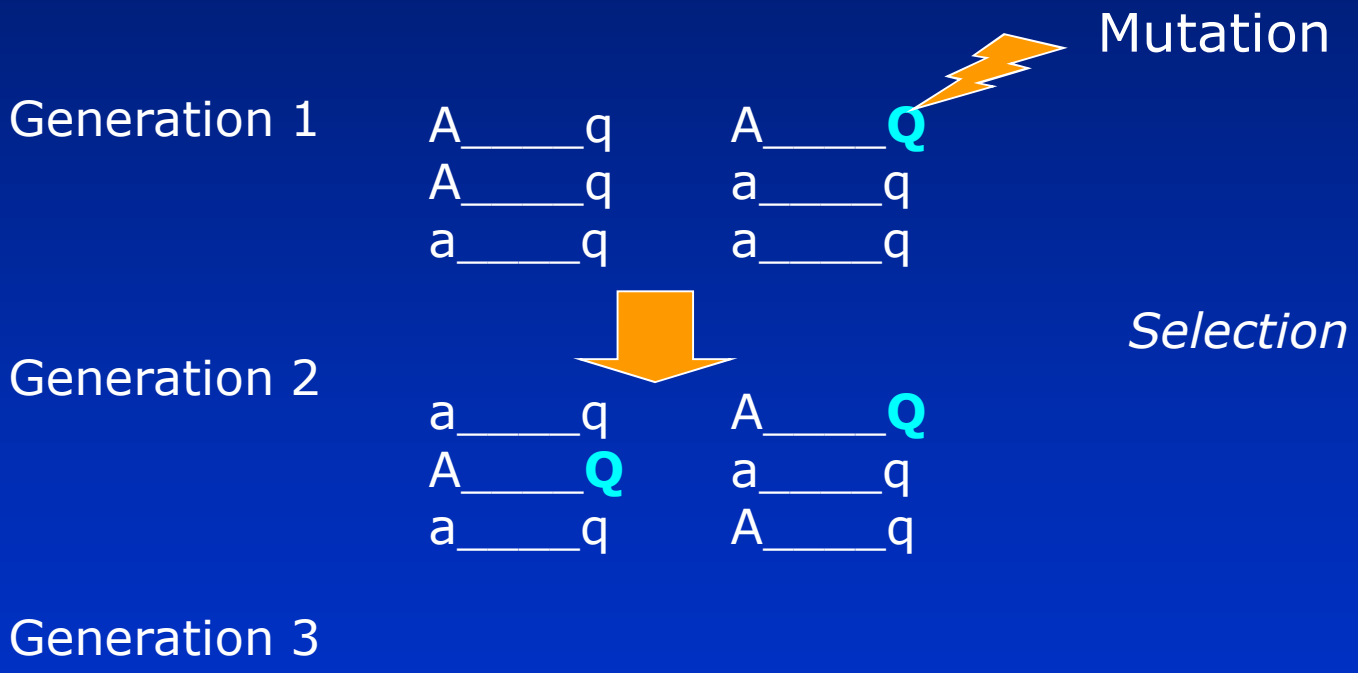
A_____q
A_____q
a_____q

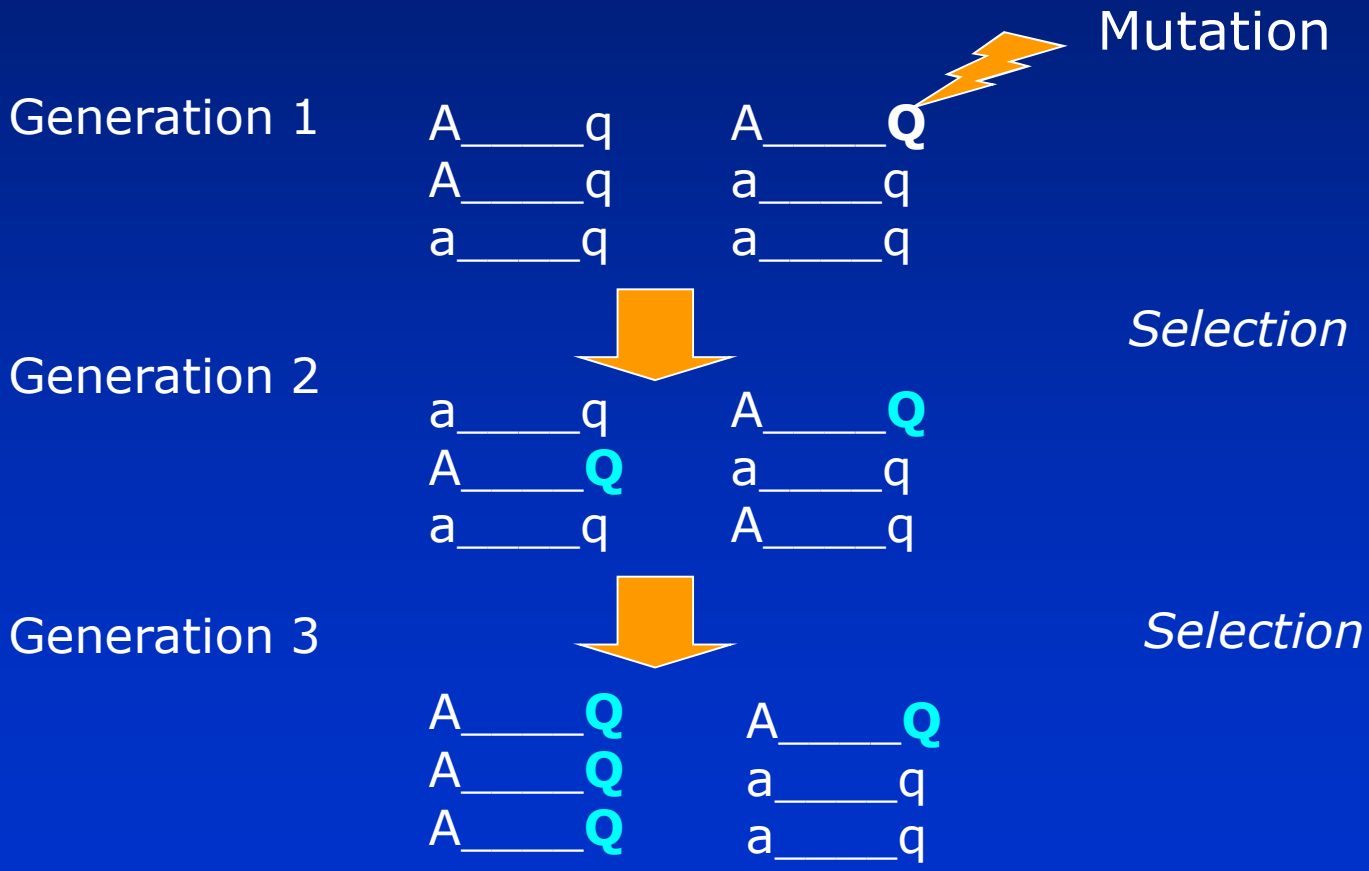
A_____q
a_____q
a_____q

Generation 2

Generation 3


Mutation
Generation 1
A_____q
A_____q
a_____q
A_____Q
a_____q
a_____q
Generation 2
Generation 3


Mutation
Generation 1
A_____q
A_____q
a_____q
A_____Q
a_____q
a_____q
Selection
Generation 2
a_____q
A_____Q
a_____q
A_____Q
a_____q
A_____q
Generation 3


Mutation
Generation 1
A_____q
A_____q
a_____q
A_____Q
a_____q
a_____q
Selection
Generation 2
a_____q
A_____Q
a_____q
A_____Q
a_____q
A_____q
Selection
Generation 3
A_____Q
A_____Q
A_____Q
A_____Q
a_____q
a_____q

Mutation

Generation 1

A_____q
A_____q
a_____q

A_____**Q**
a_____q
a_____q

Selection

Generation 2

a_____q
A_____**Q**
a_____q

A_____**Q**
a_____q
A_____q

Generation 3

Selection

A_____**Q**
A_____**Q**
A_____**Q**

A_____**Q**
a_____q
a_____q

Causes of LD

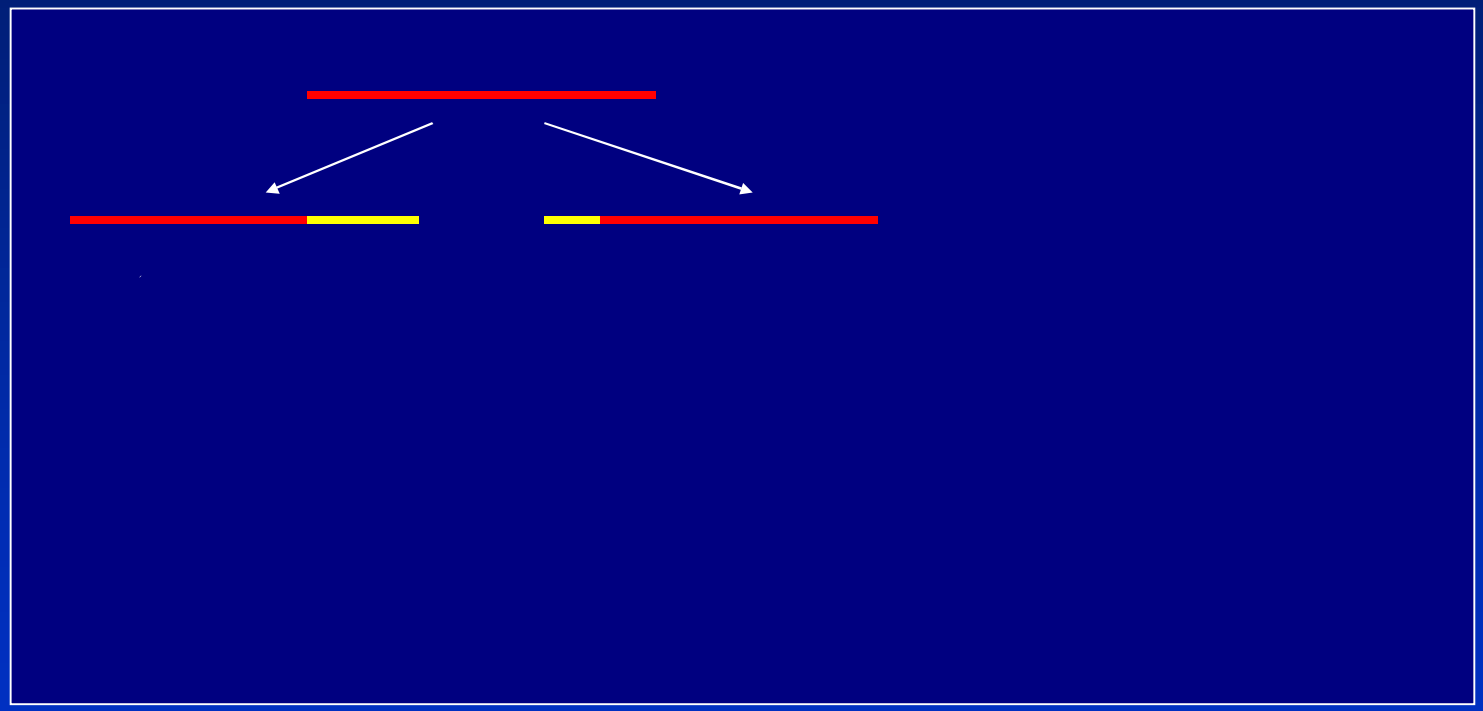
- Migration
  - LD artificially created in crosses designs
    - large when crossing inbred lines
    - but small when crossing breeds that do not differ markedly in gene frequencies
    - disappears after only a limited number of generations
- Selection
  - Selective sweeps
- Finite population size
  - generally implicated as the key cause of LD in livestock populations, where effective population size is small

Causes of LD

- A chunk of ancestral chromosome is conserved in the current population

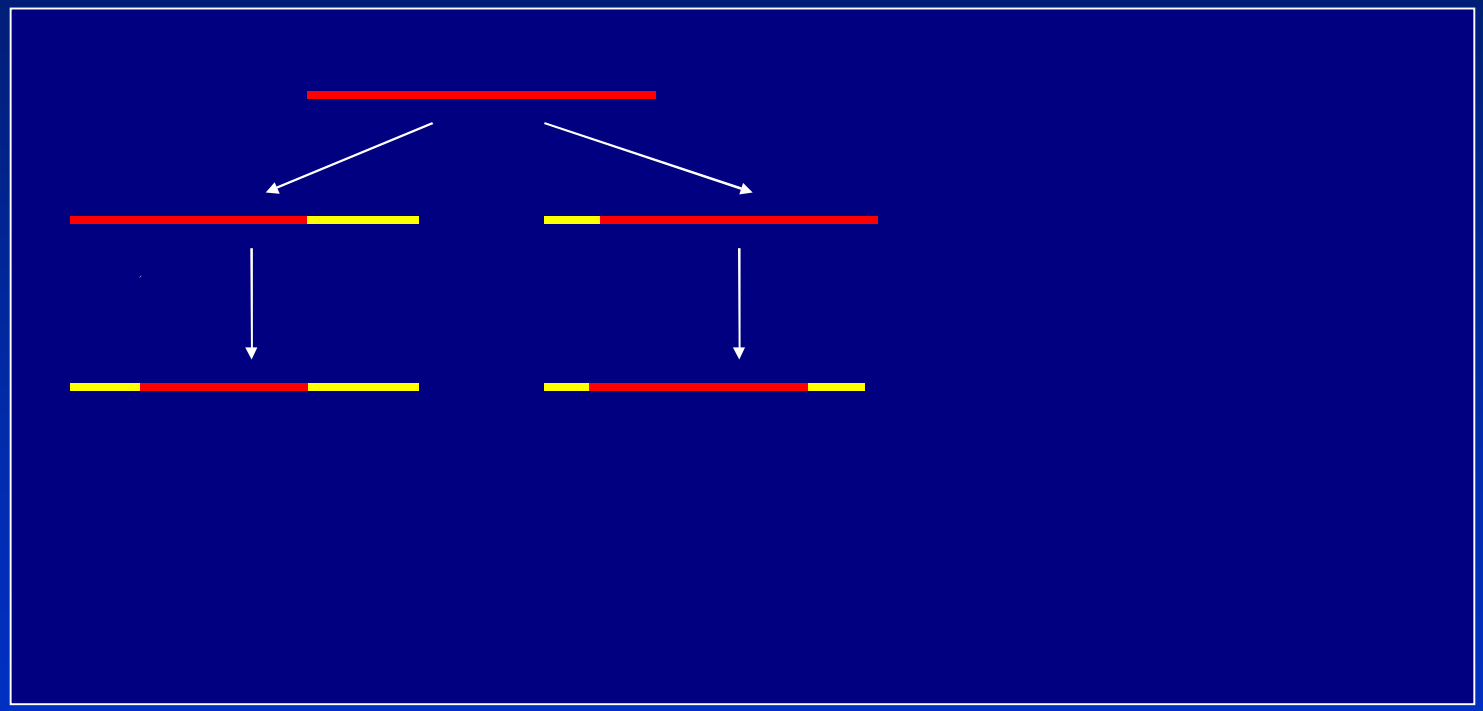
Causes of LD

- A chunk of ancestral chromosome is conserved in the current population



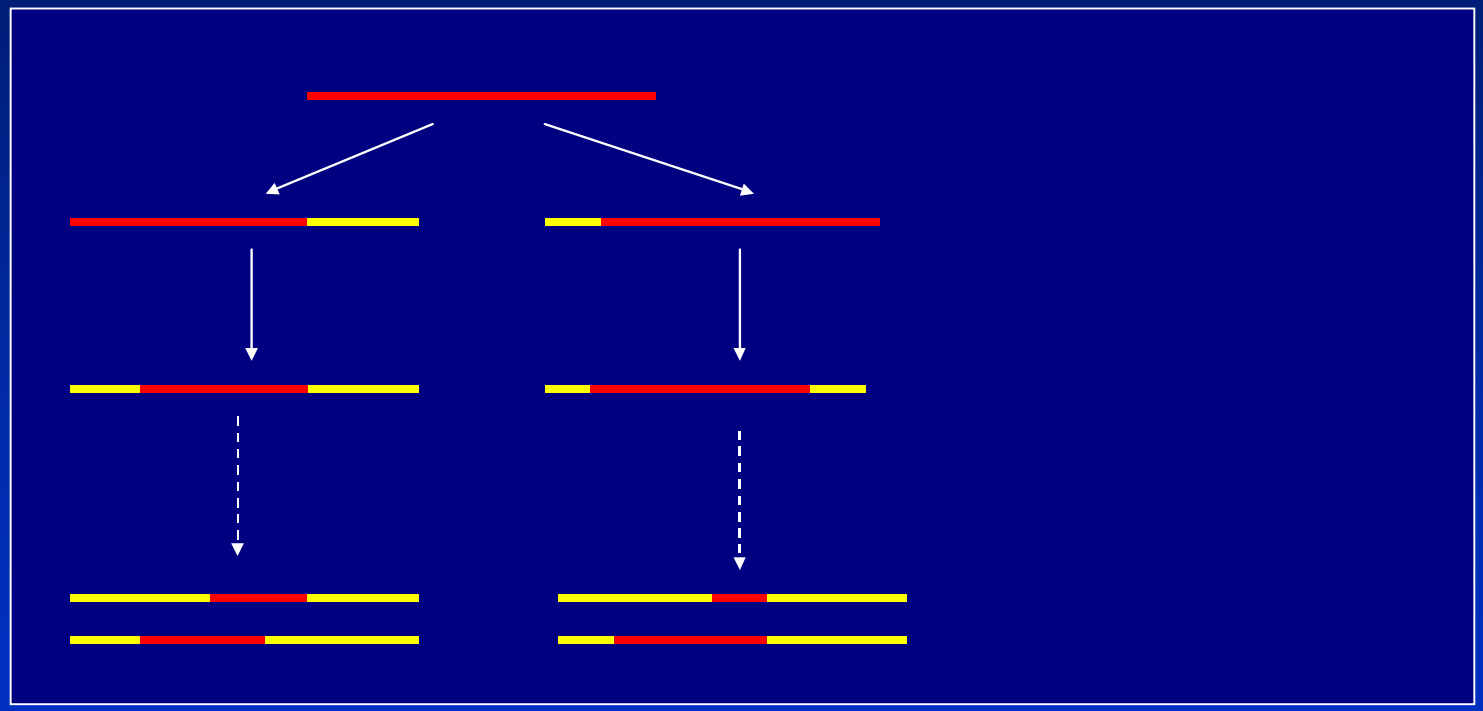
Causes of LD

- A chunk of ancestral chromosome is conserved in the current population



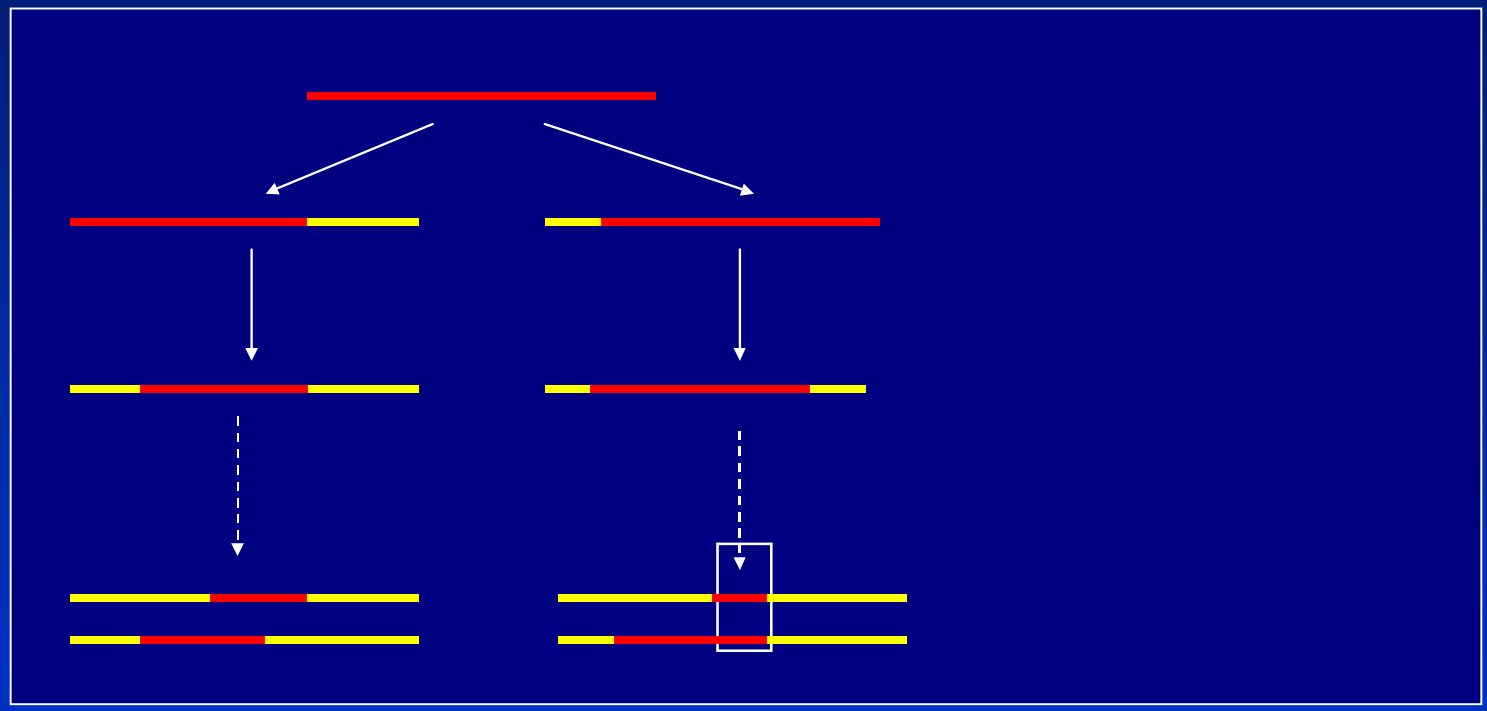
Causes of LD

- A chunk of ancestral chromosome is conserved in the current population



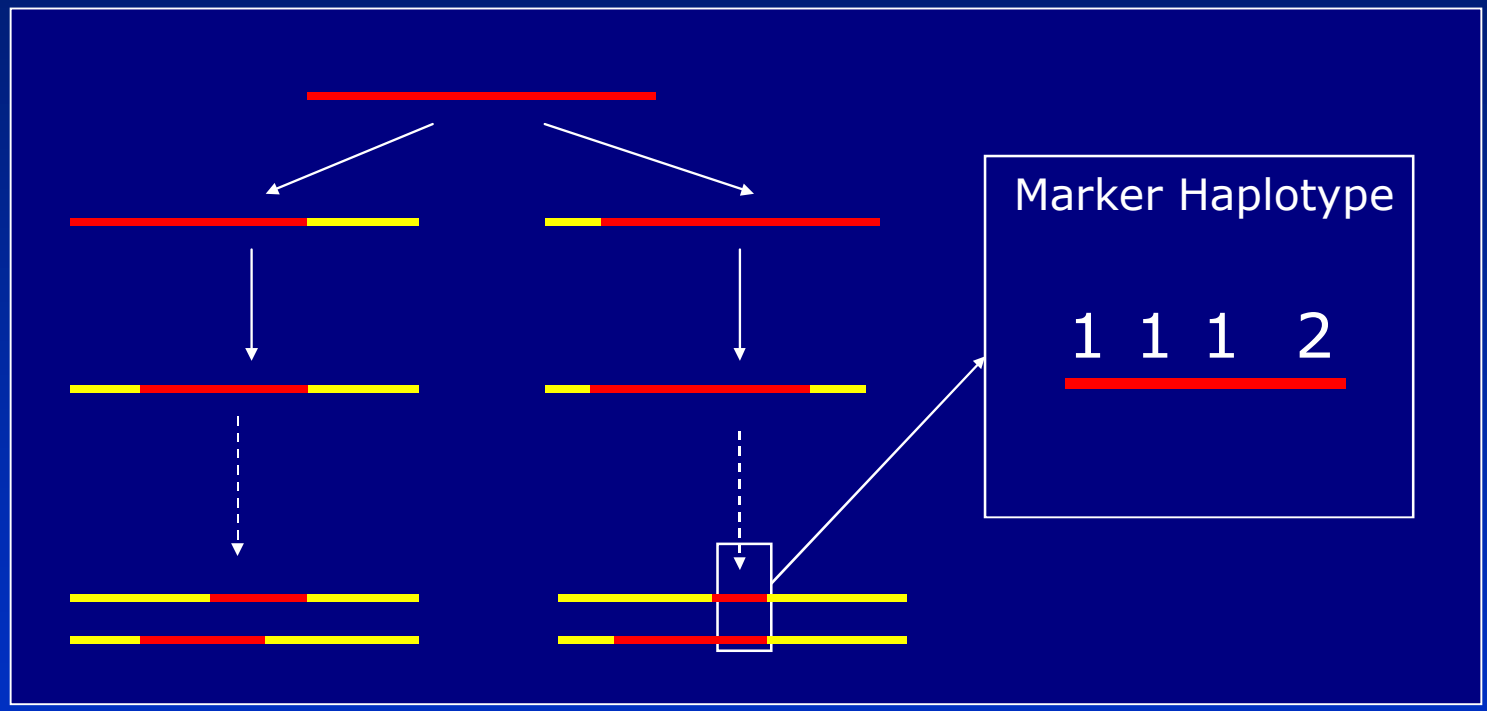
Causes of LD

- A chunk of ancestral chromosome is conserved in the current population



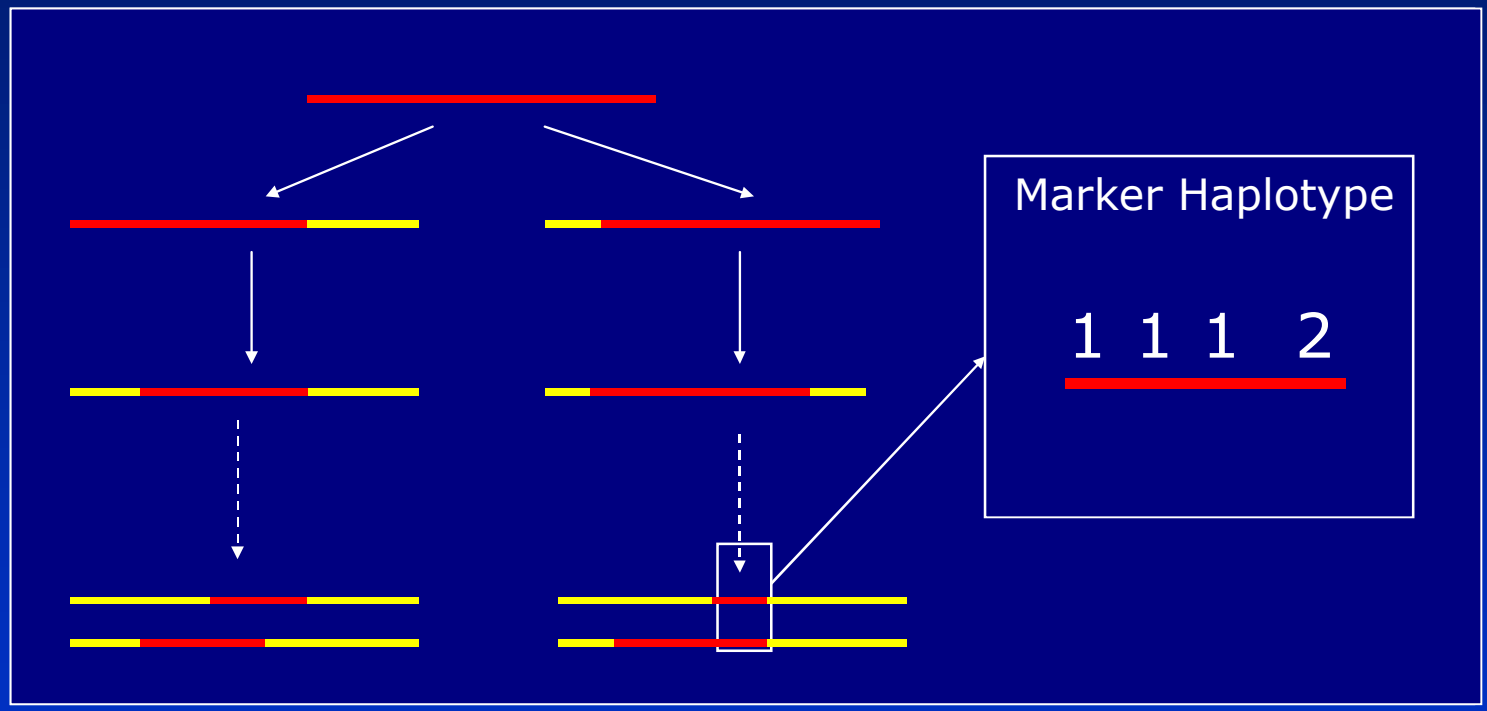
Causes of LD

- A chunk of ancestral chromosome is conserved in the current population



Causes of LD

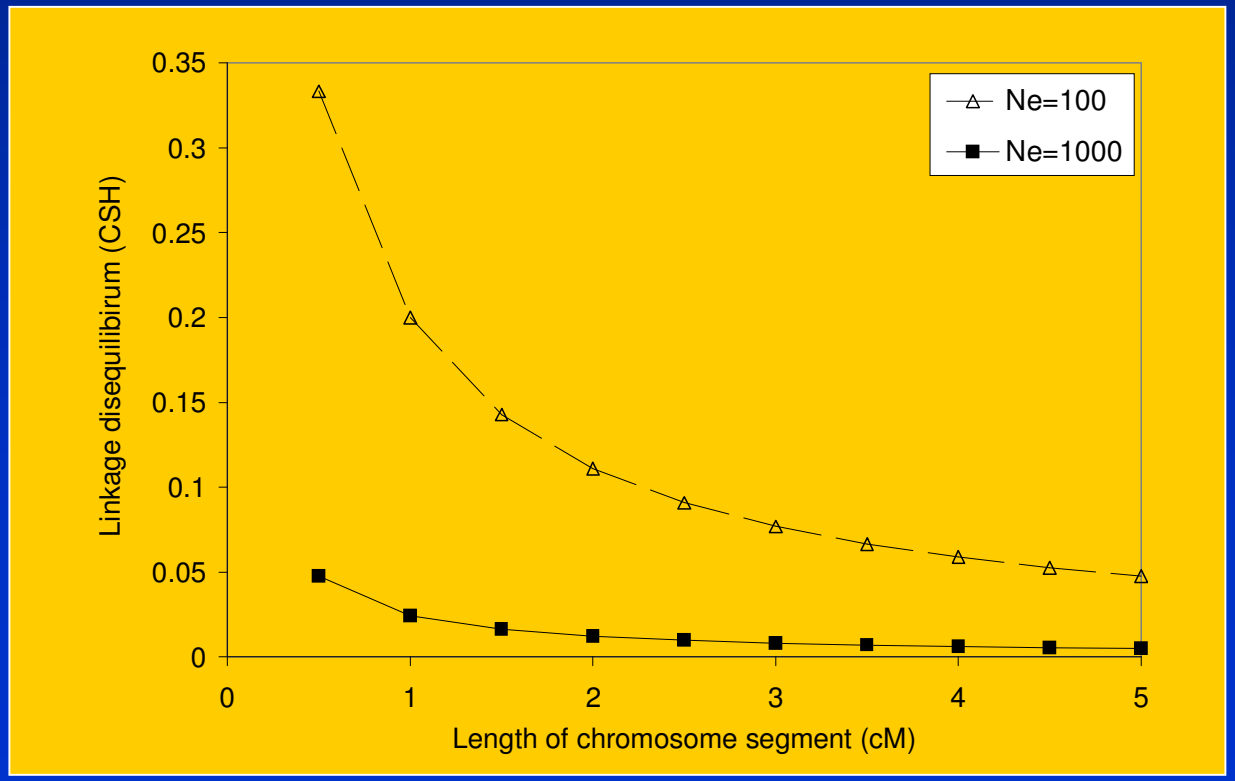
- A chunk of ancestral chromosome is conserved in the current population



- Size of conserved chunks depends on effective population size

Causes of LD

- Predicting LD with finite population size
- $E(r^2) = 1/(4Nc+1)$
  - N = effective population size
  - c = length of chromosome segment

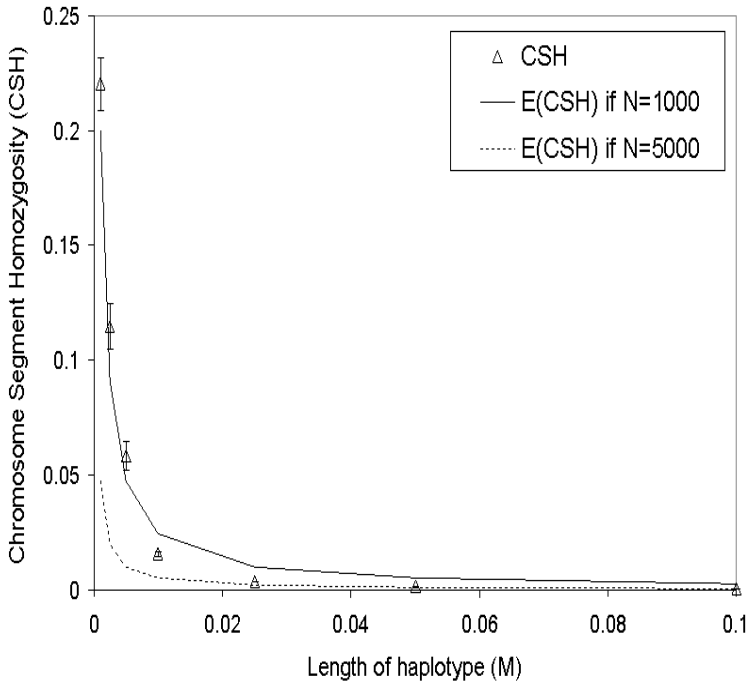


Causes of LD

- But this assumes constant effective population size over generations
- In livestock, effective population size has changed as a result of domestication
- 100 000 -> 1500 -> 100 ?
- In humans, has greatly increased
- 2000 -> 100 000 ?

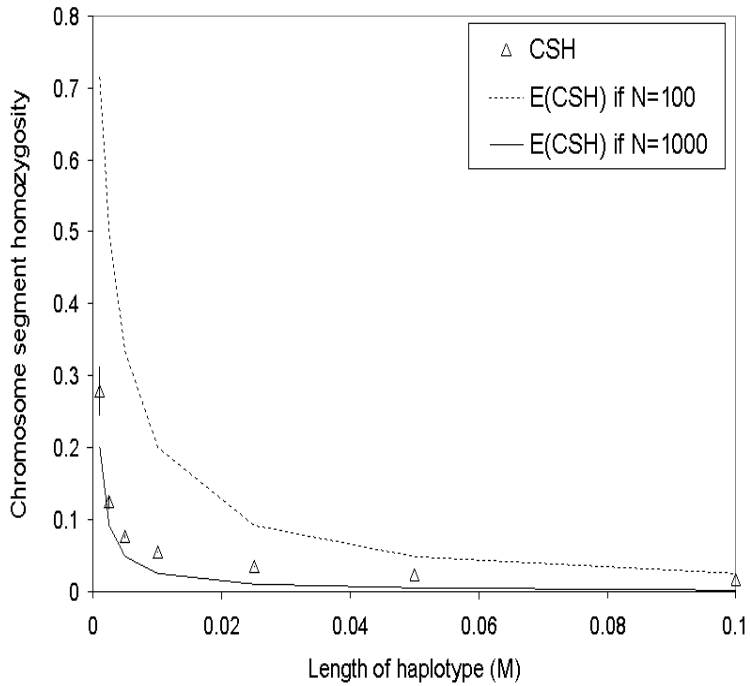
Causes of LD

1000 to 5000



A

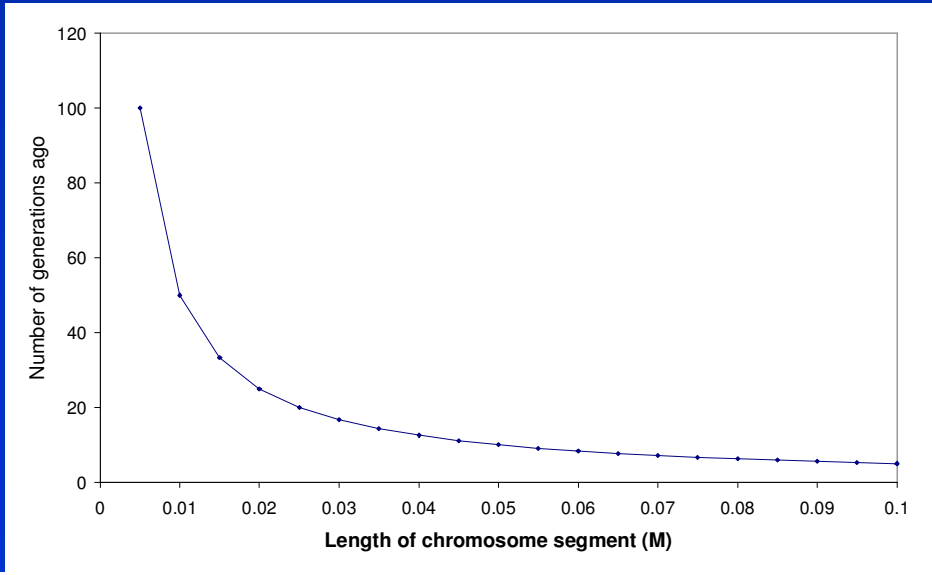
1000 to 100



B

Causes of LD

- $E(r^2) = 1/(4N_tc+1)$
- Where t = 1/(2c) generations ago
  - eg markers 0.1M (10cM) apart reflect population size 5 generations ago
  - Markers 0.001 (0.1cM) apart reflect effective pop size 500 generations ago



Causes of LD

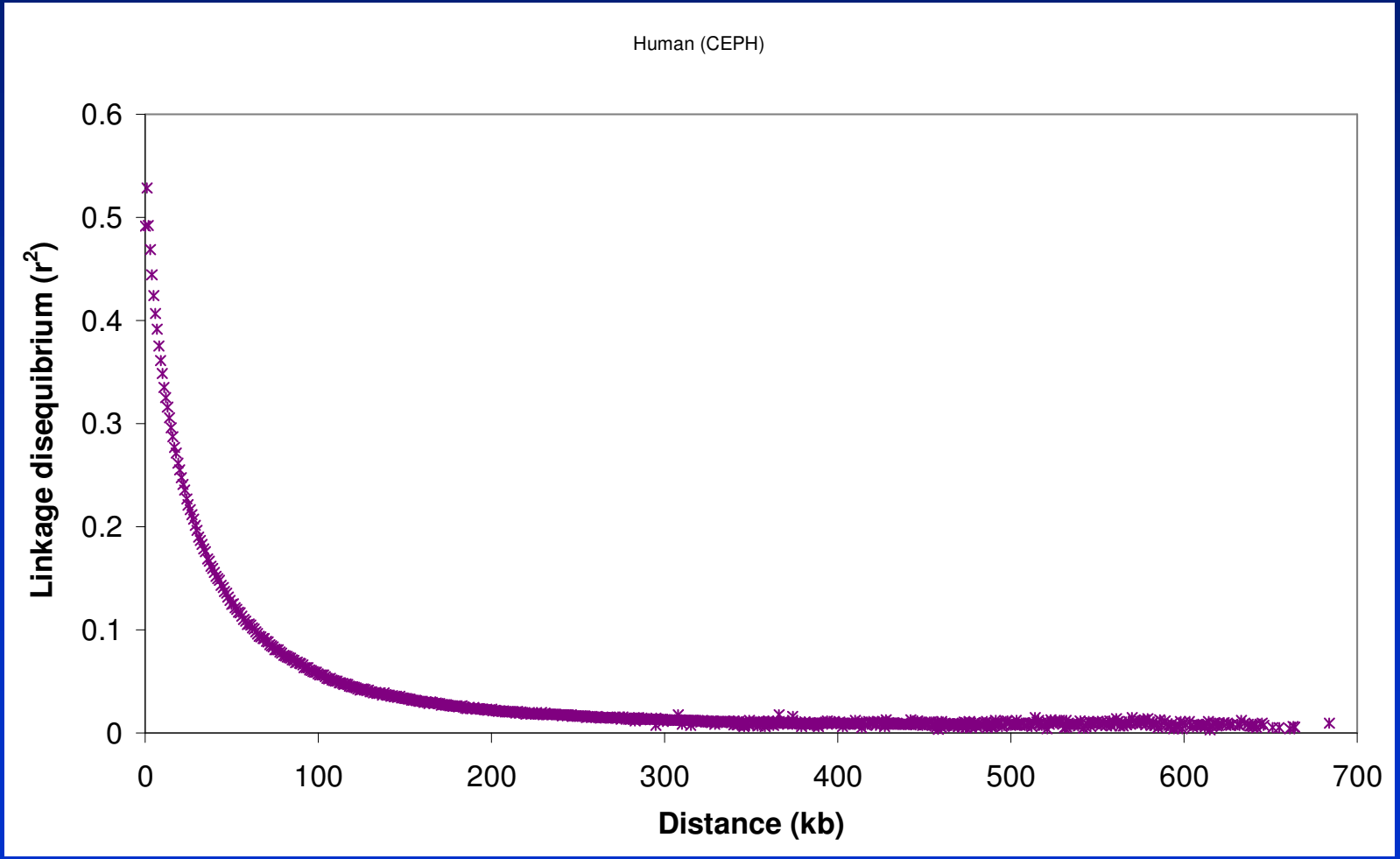
- $E(r^2) = 1/(4N_tc+1)$
- Where t = 1/(2c) generations ago
  - eg markers 0.1M (10cM) apart reflect population size 5 generations ago
  - Markers 0.001 (0.1cM) apart reflect effective pop size 500 generations ago
- LD at short distances reflects historical effective population size
- LD at longer distances reflects more recent population history

Linkage disequilibrium

- Measuring linkage disequilibrium
- Causes of LD
- Extent of LD in animals and plants

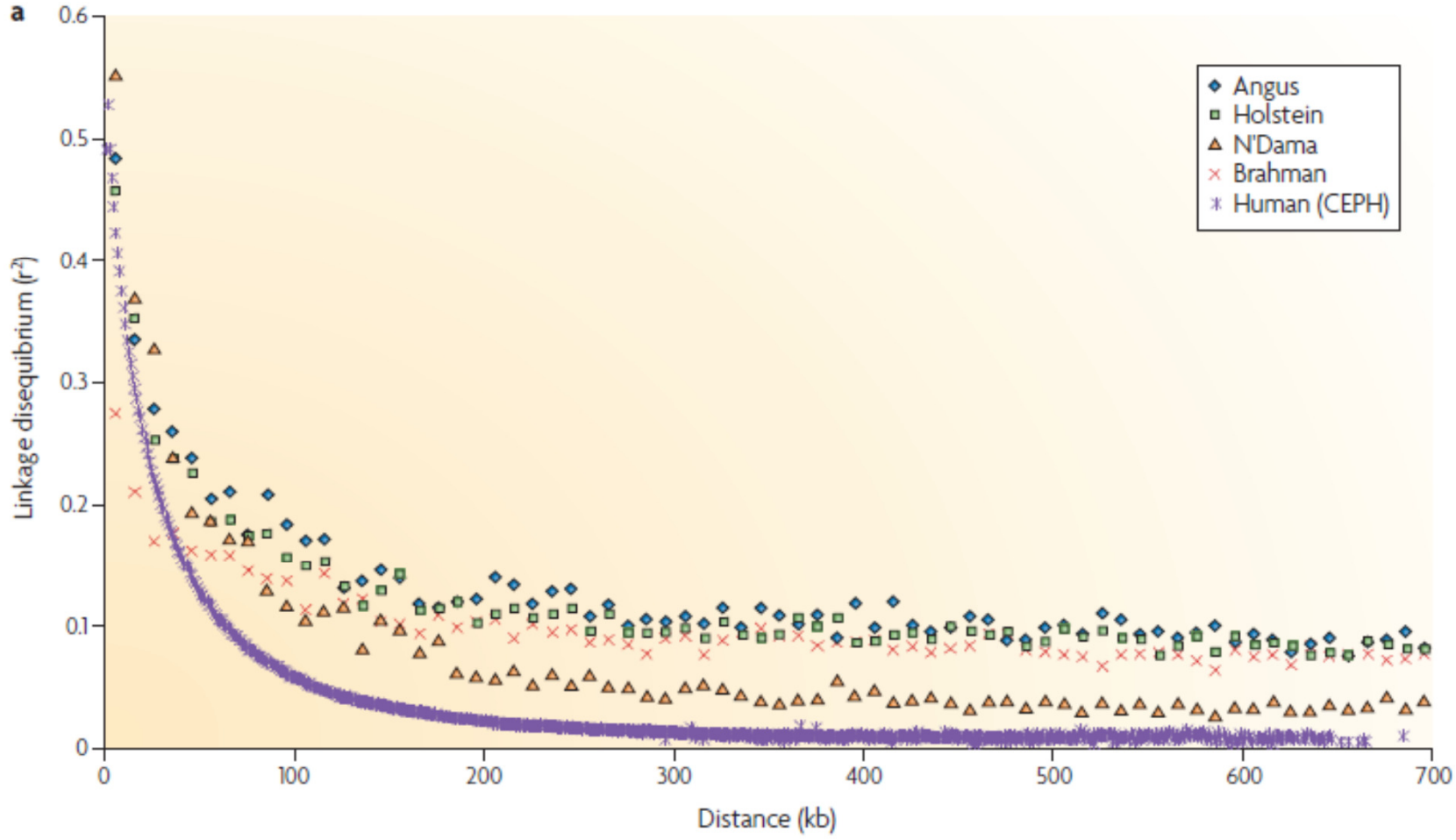
Extent of LD in humans and livestock

Humans…….(Tenesa et al. 2007)



Extent of LD in humans and livestock

And cattle……

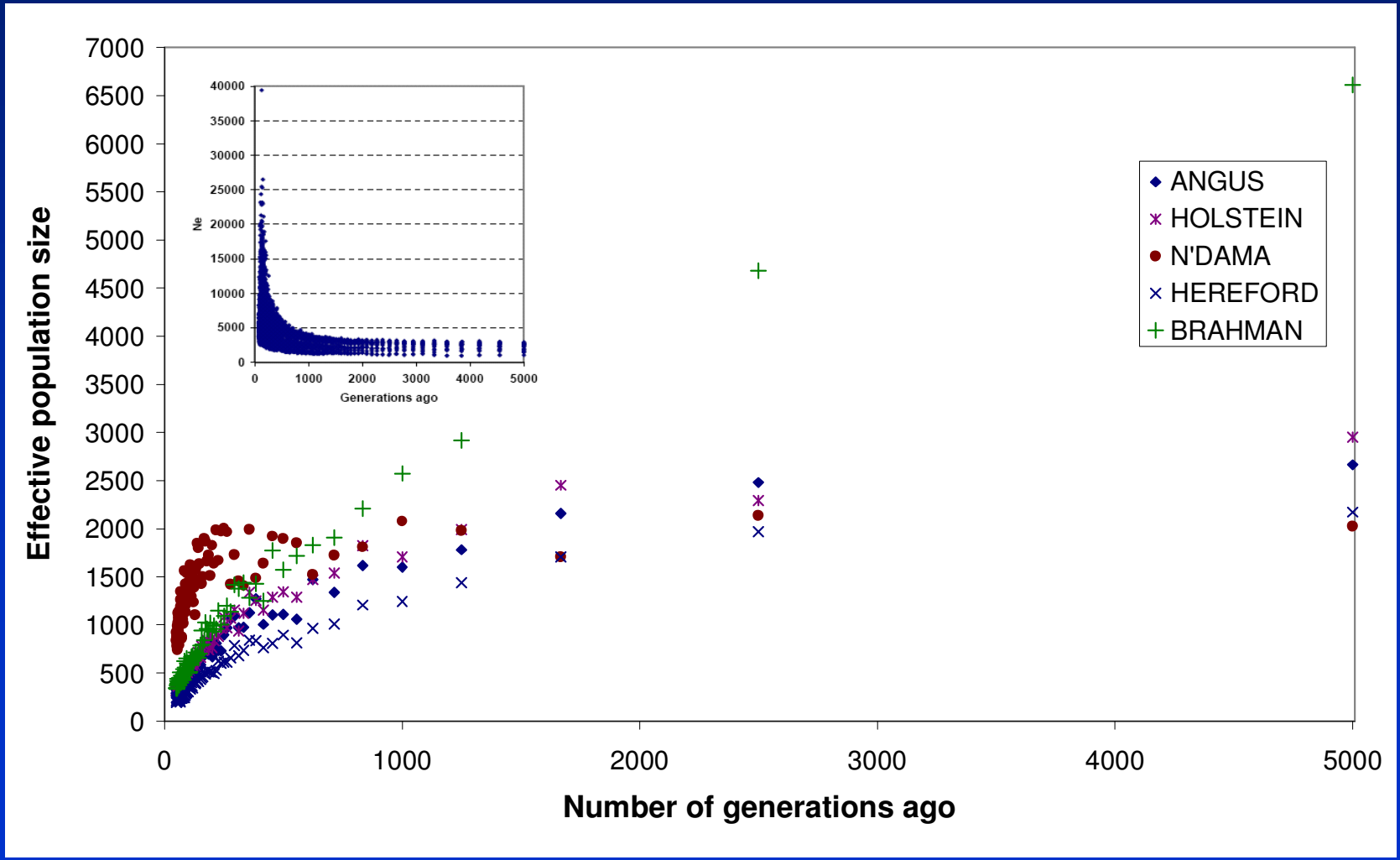


Causes of LD

- $E(r^2) = 1/(4N_tc+1)$
- Where t = 1/(2c) generations ago
  - eg markers 0.1M (10cM) apart reflect population size 5 generations ago
  - Markers 0.001 (0.1cM) apart reflect effective pop size 500 generations ago
- LD at short distances reflects historical effective population size
- LD at longer distances reflects more recent population history

Extent of LD in humans and livestock

Population size humans and cattle…..



Implications?

- In Holsteins, need a marker approximately every 50kb to get average r^2 of 0.3 between marker and QTL (eg. 25kb marker-QTL).

Implications?

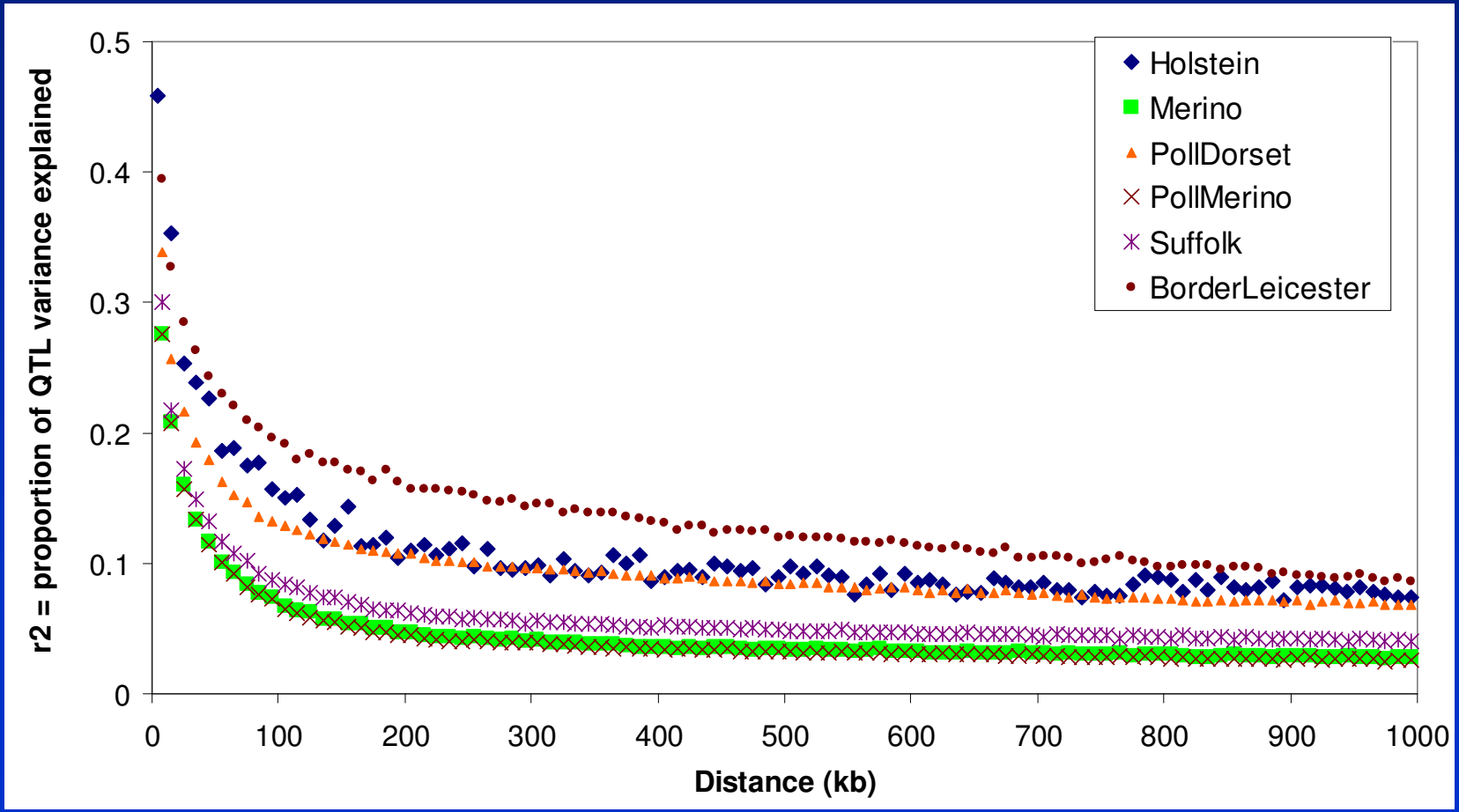
- In Holsteins, need a marker approximately every 50kb to get average r^2 of 0.3 between marker and QTL (eg. 25kb marker-QTL).
- This level of marker-QTL LD would allow a genome wide association study of reasonable size to detect QTL of moderate effect.

Implications?

- In Holsteins, need a marker approximately every 50kb to get average r^2 of 0.3 between marker and QTL (eg. 25kb marker-QTL).
- This level of marker-QTL LD would allow genomic prediction to capture reasonable proportion of genetic variance
- Bovine genome is approximately 3,000,000kb
  - 60,000 evenly spaced markers to capture every QTL in a genome scan

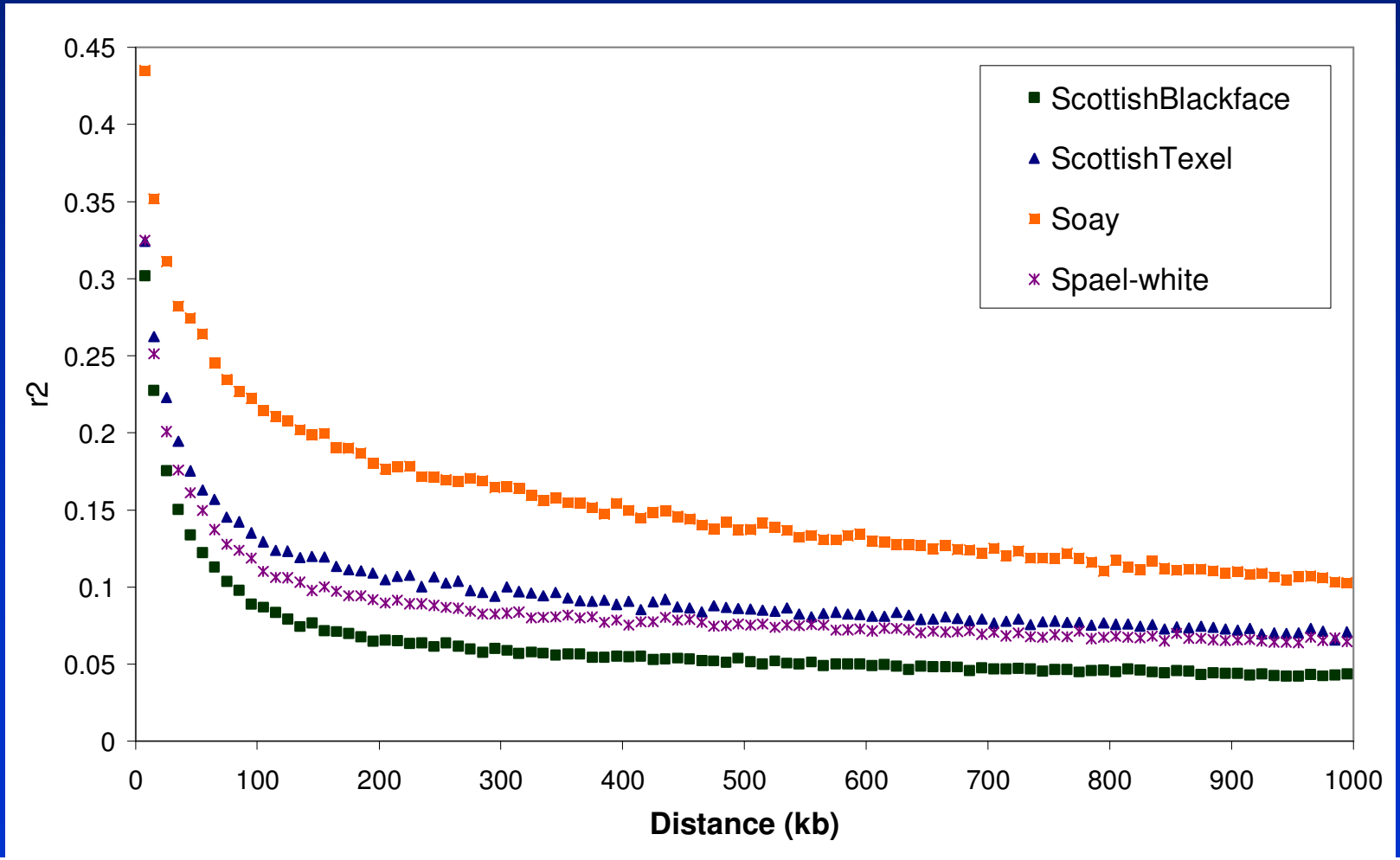
Extent of LD in other species

- Sheep HapMap project (Kijas et al. 2011)



Extent of LD in other species

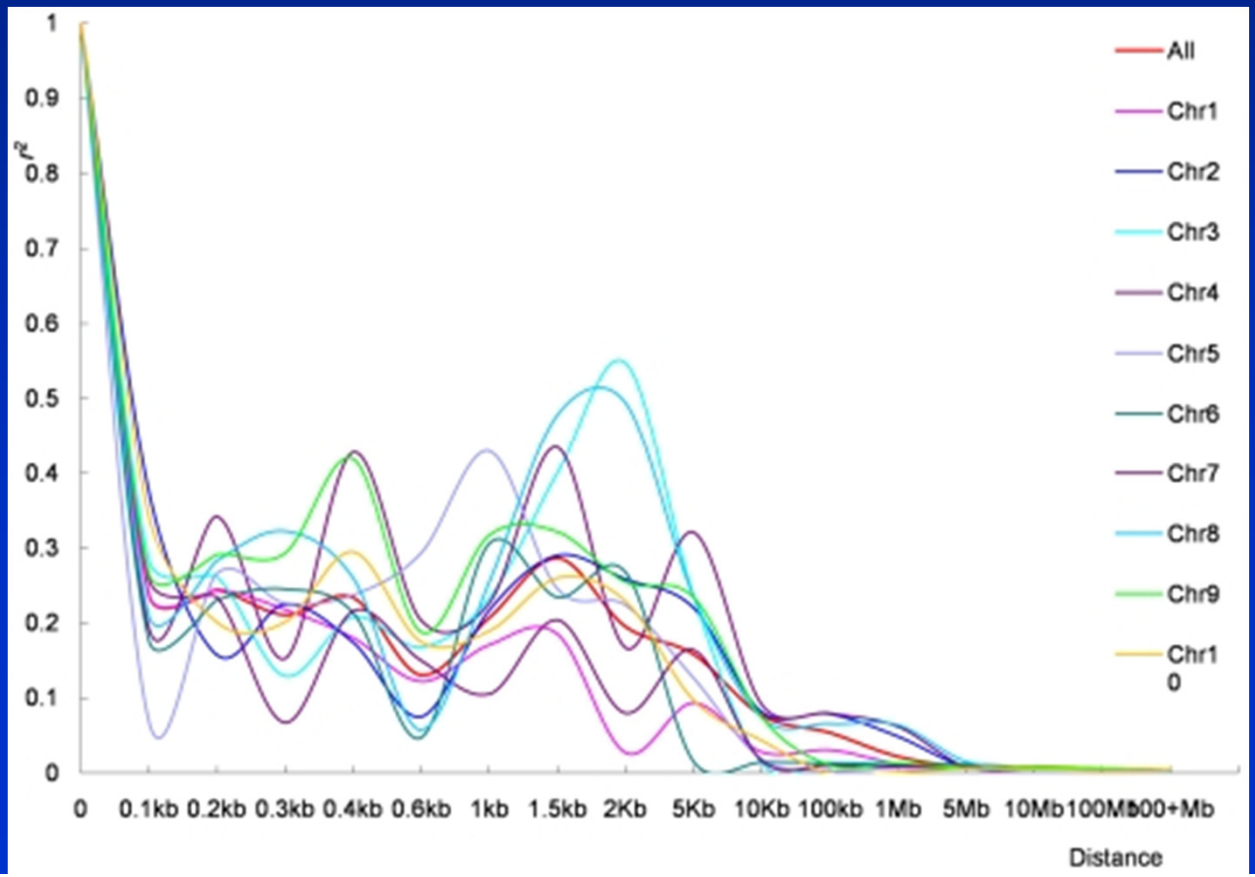
- Sheep HapMap project (Kijas et al. 2011)



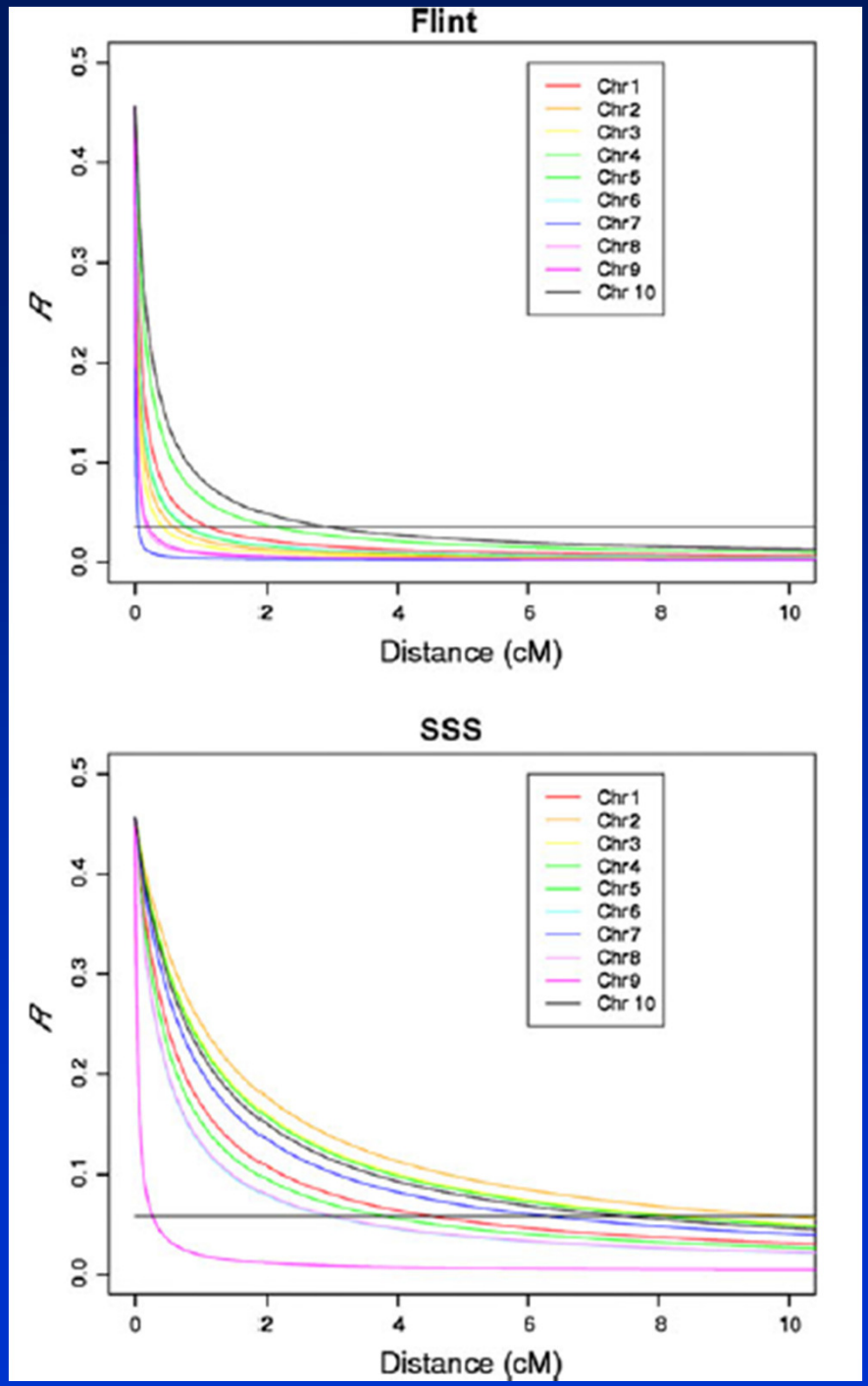
Extent of LD in other species

- Sheep HapMap project (Kijas et al. 2011)

- Maize (i)
  - Yan et al. 2009 (PLoS One. 4:e8451).
  - Relatively low LD across 632 inbred lines
  - Concluded up to 480,000 SNPs needed for genome wide association

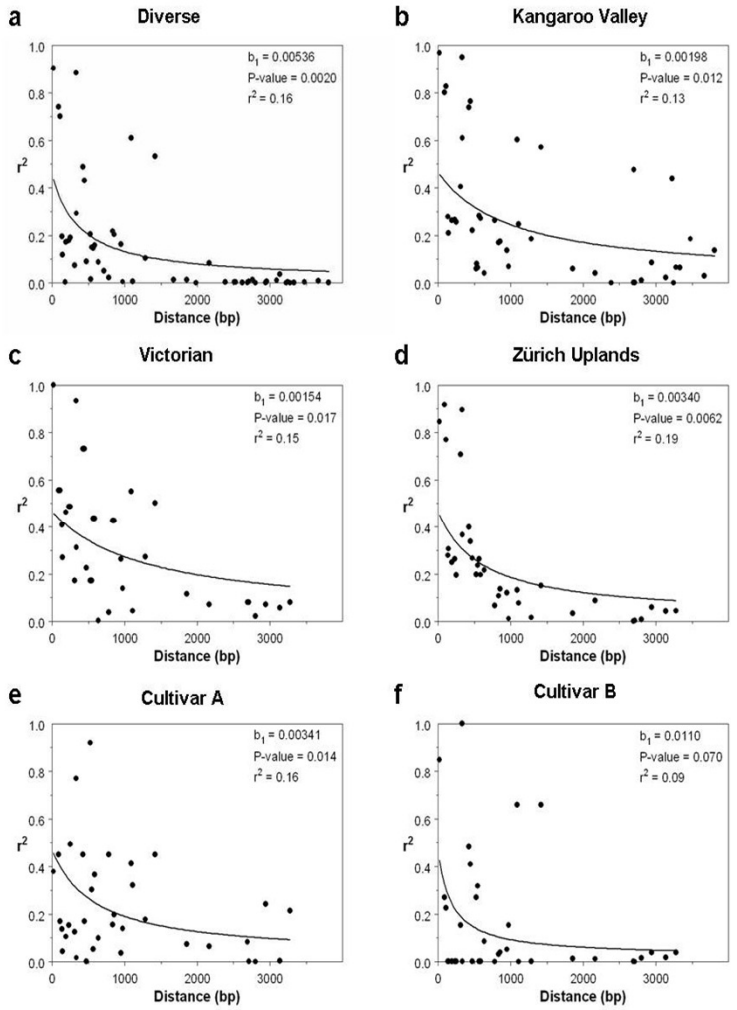


- Maize (ii)
  - Van Ingehlandt et al. 2011 TAG 123:11
  - Considerable LD among heterotic groups
  - Concluded 4000-65,000 SNPs needed for genome wide association



Extent of LD in other species

- Perennial ryegrass
  - outbreeder
  - very little LD (Ponting et al 2007)
  - Extremely large effective population size?



Linkage disequilibrium

- Extent of LD in a species determines marker density necessary for genomic prediction
- In cattle, r^2~0.3 at 50kb ~ 60 000 markers necessary for genome scan
- In humans, LD lower, need many more markers for genomic prediction

Day 1

- Introduction to Quantitative traits
- Linkage disequilibrium
- Genome wide association studies

Genome wide association

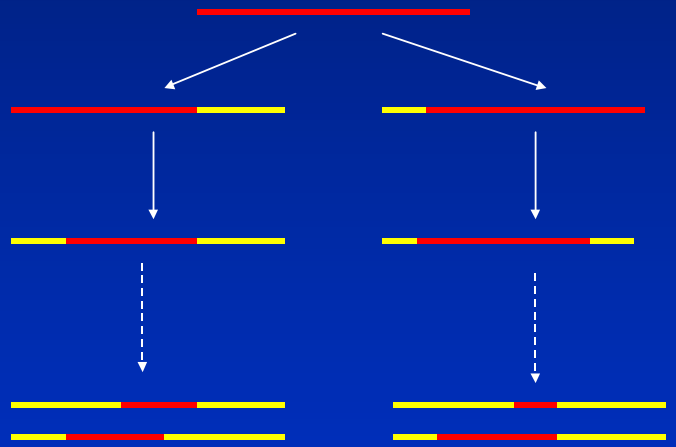
- Association testing with single marker regression
- Power of genome wide association studies
- Accounting for population structure
- Validation

Genome wide association

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

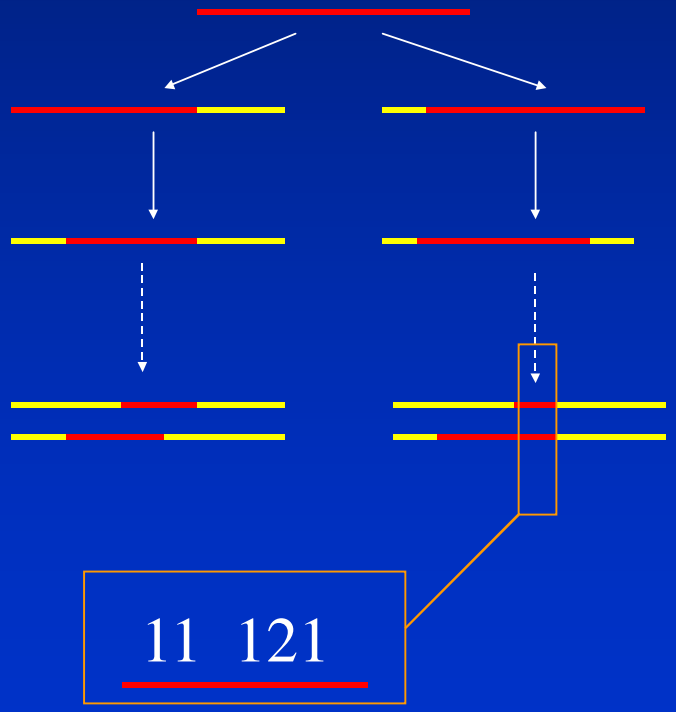
Genome wide association

- 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



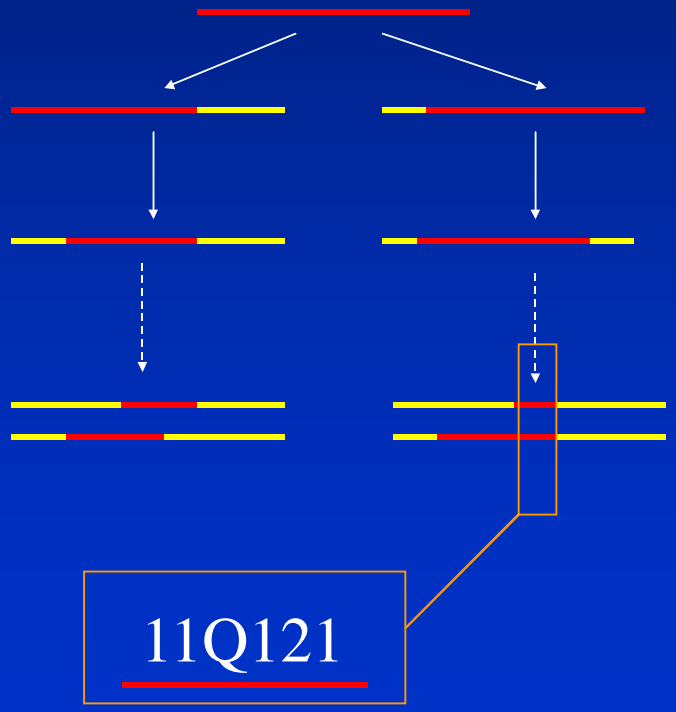
Genome wide association

- 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



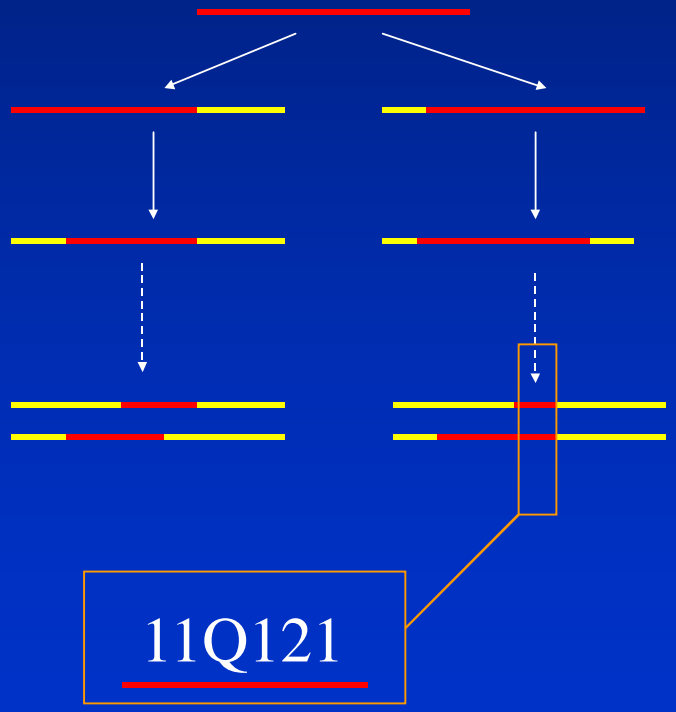
Genome wide association

- 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.

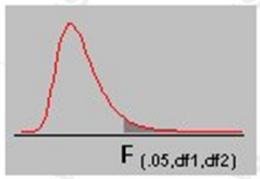
Single marker regression

- 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

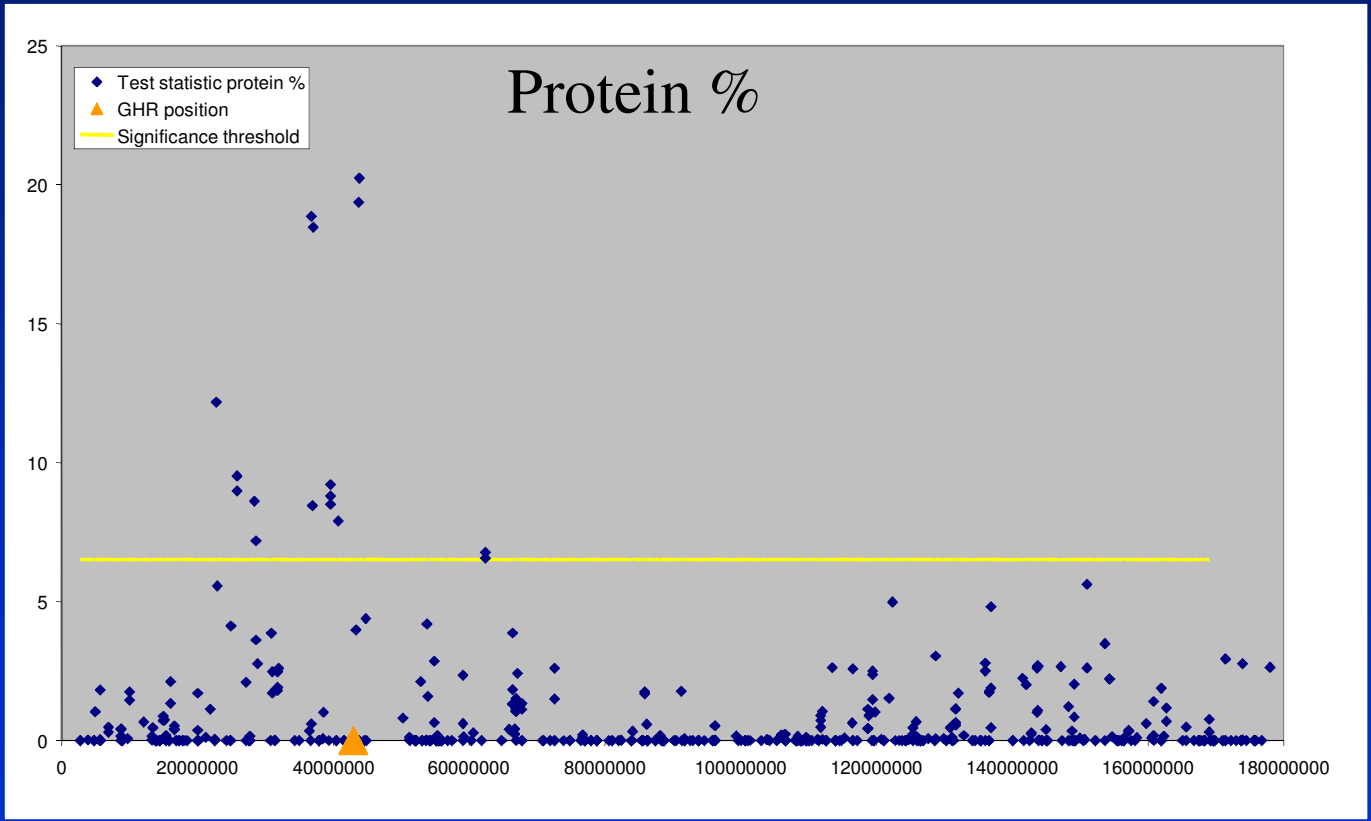


df2/df1	1	2	3	4	5	6	7	8	9	10
1	161.4476	199.5000	215.7073	224.5832	230.1619	233.9860	236.7684	238.8827	240.5433	241.8817
2	18.5128	19.0000	19.1643	19.2468	19.2964	19.3295	19.3532	19.3710	19.3848	19.3959
3	10.1280	9.5521	9.2766	9.1172	9.0135	8.9406	8.8867	8.8452	8.8123	8.7855
4	7.7086	6.9443	6.5914	6.3882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

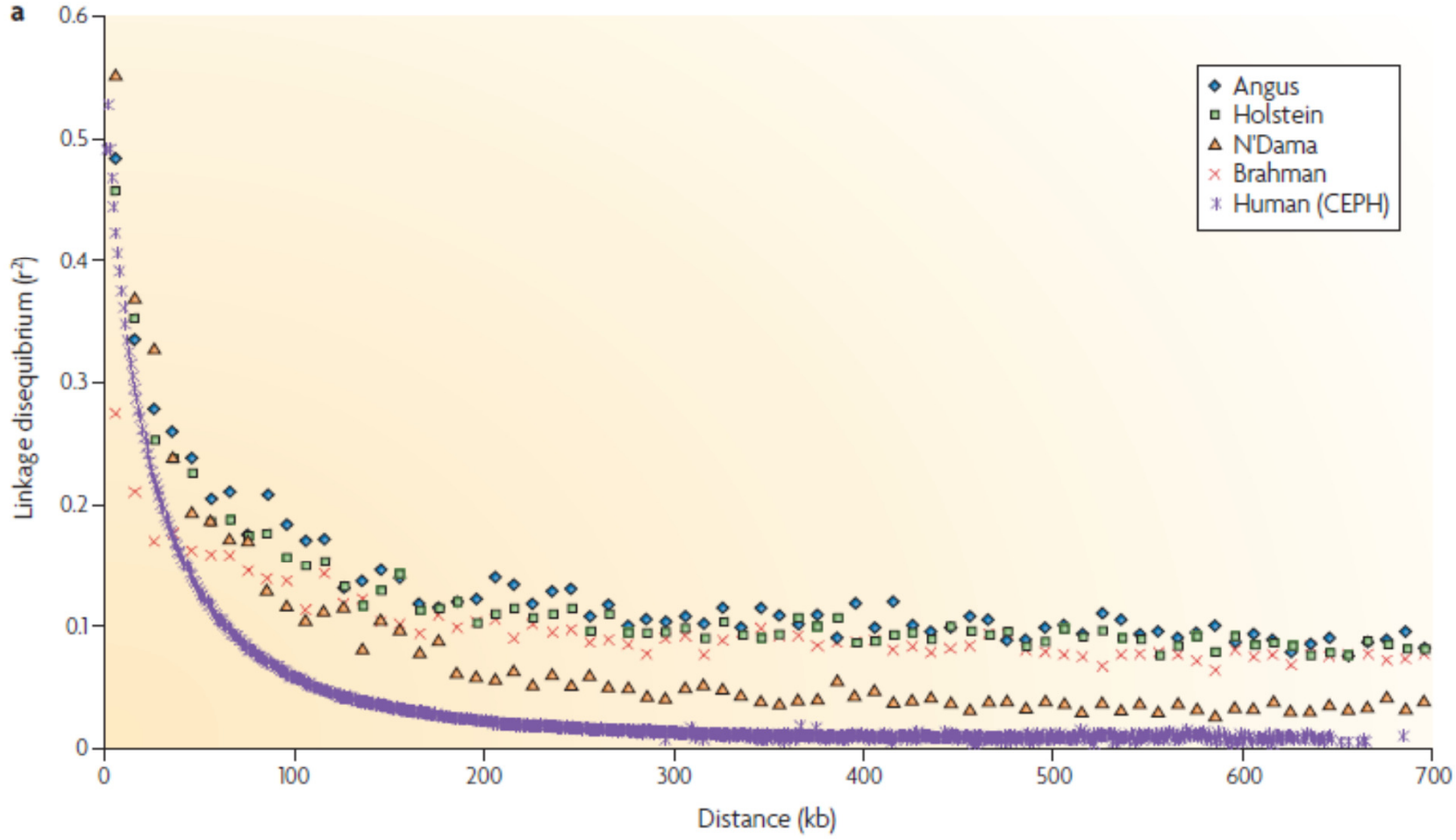
- 384 Holstein-Friesian dairy bulls selected from Australian dairy bull population
- genotyped for 10 000 SNPs
- Single marker regression with protein%

Results of genome scans with dense SNP panels

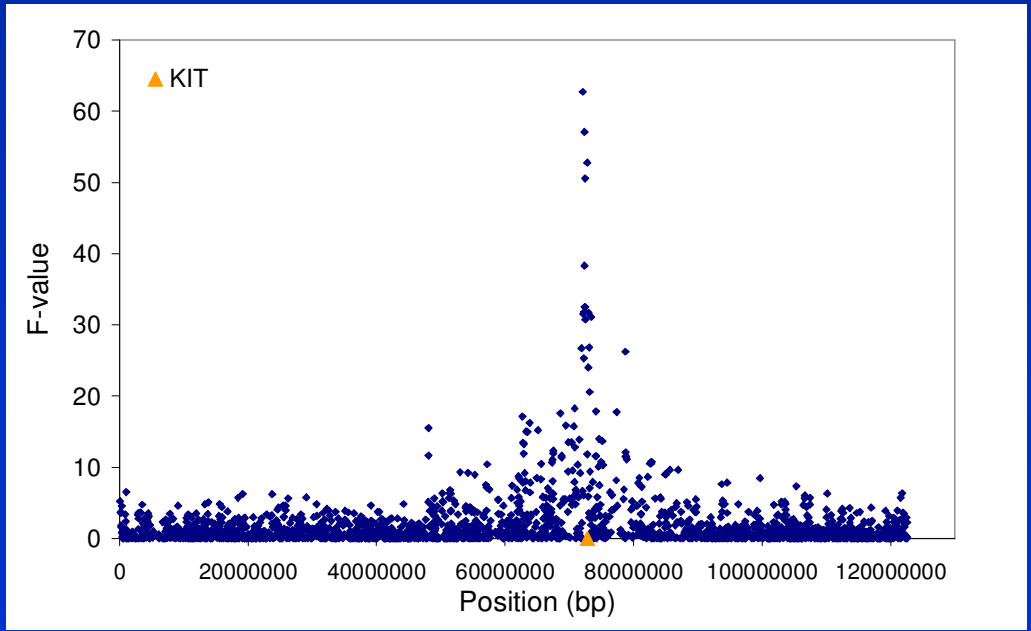


Extent of LD in humans and livestock

And cattle……



Proportion of black….



Genome wide association

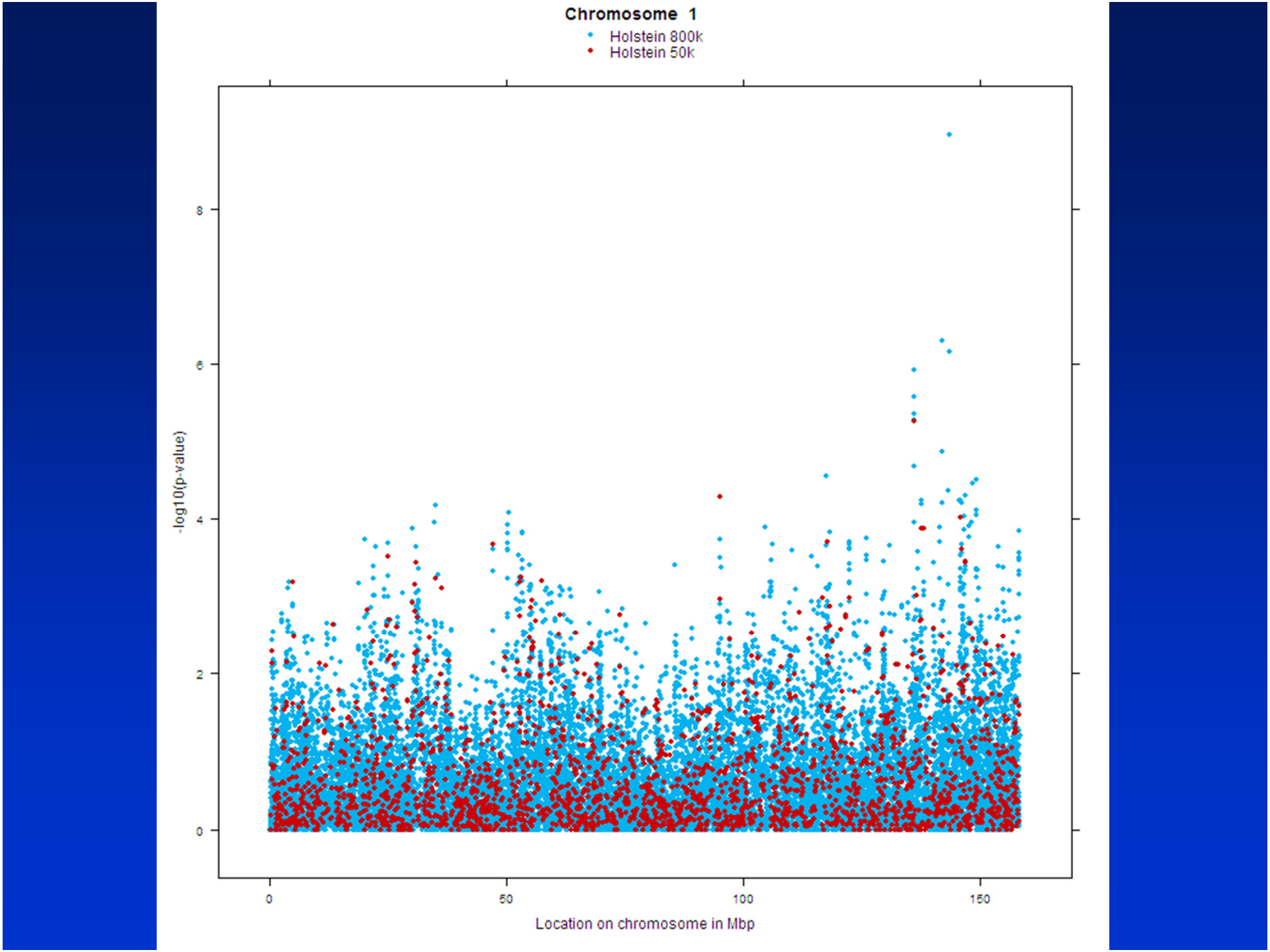
- Association testing with single marker regression
- Power of genome wide association studies
- Accounting for population structure
- Validation

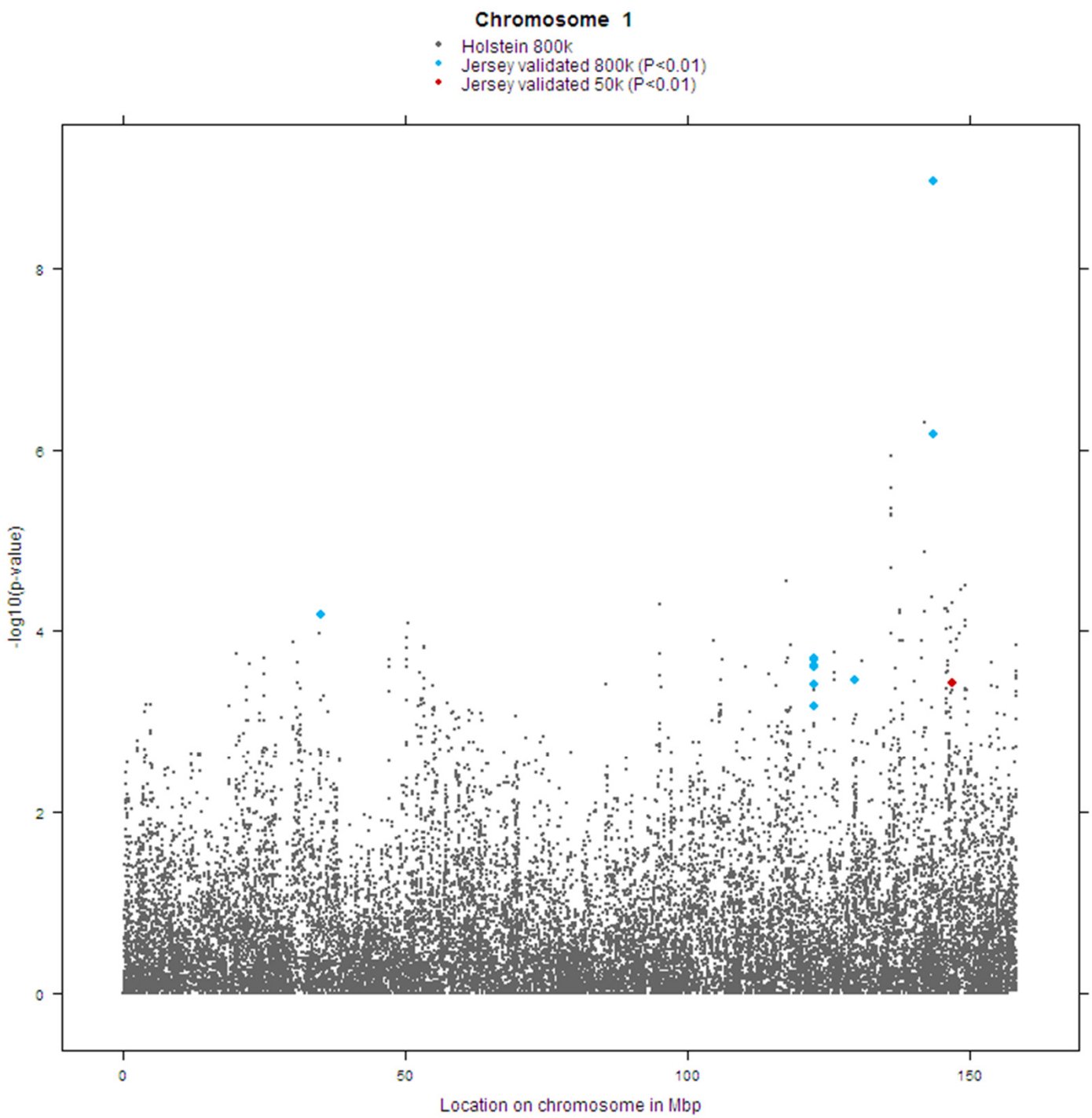
Power of GWAS

- 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

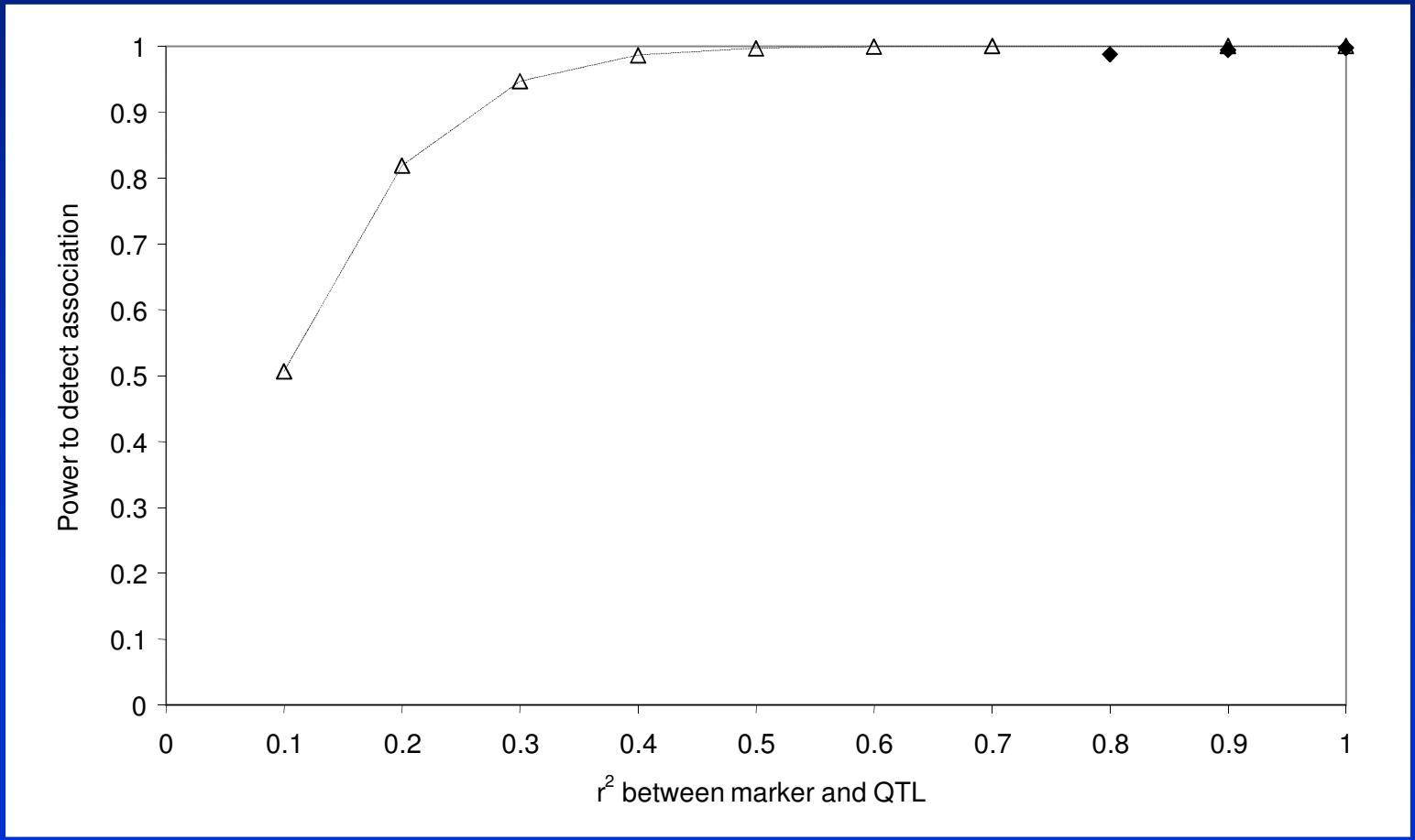
- Power is a function of:
  - r^2 between the marker and QTL
    - sample size must be increased by $1/r^2$ to detect an un-genotyped QTL, compared with sample size for testing QTL itself
  - Proportion of total phenotypic variance explained by the QTL
  - Number of phenotypic records

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

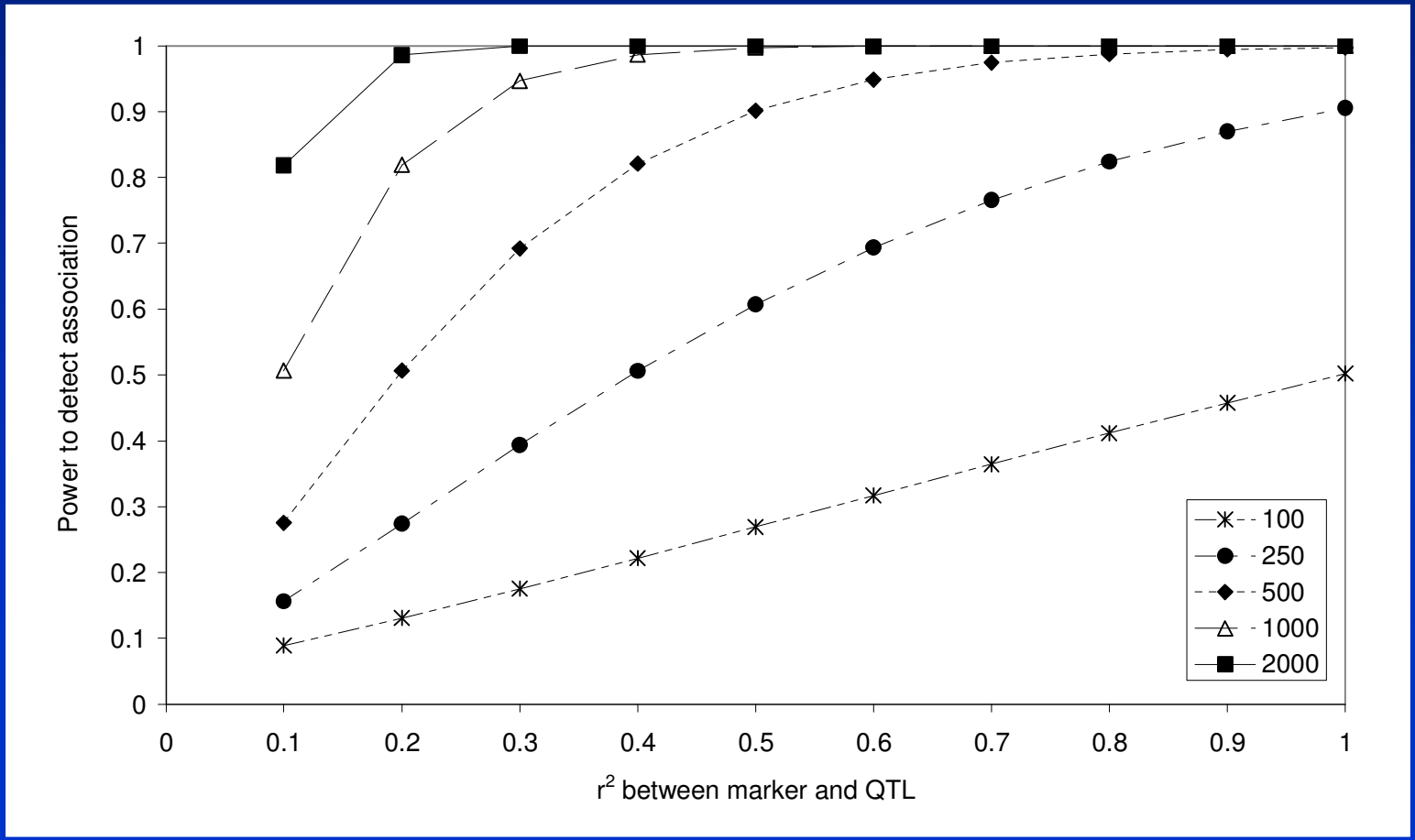
Power of GWAS

- 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



Human height

NATURE | LETTER

previous article next article

Hundreds of variants clustered in genomic loci and biological pathways affect human height

Hana Lango Allen, Karol Estrada, Guillaume Lettre, Sonja I. Berndt, Michael N. Weedon, Fernando Rivadeneira, Cristen J. Willer, Anne U. Jackson, Sailaja Vedantam, Soumya Raychaudhuri, Teresa Ferreira, Andrew R. Wood, Robert J. Weyant, Ayellet V. Segrè, Elizabeth K. Speliotes, Eleanor Wheeler, Nicole Soranzo, Ju-Hyun Park, Jian Yang, Daniel Gudbjartsson, Nancy L. Heard-Costa, Joshua C. Randall, Lu Qi, Albert Vernon Smith, Reedik Mägi + *et al.*

Affiliations | Contributions | Corresponding authors

Nature **467**, 832–838 (14 October 2010) | doi:10.1038/nature09410

Received 23 Ap

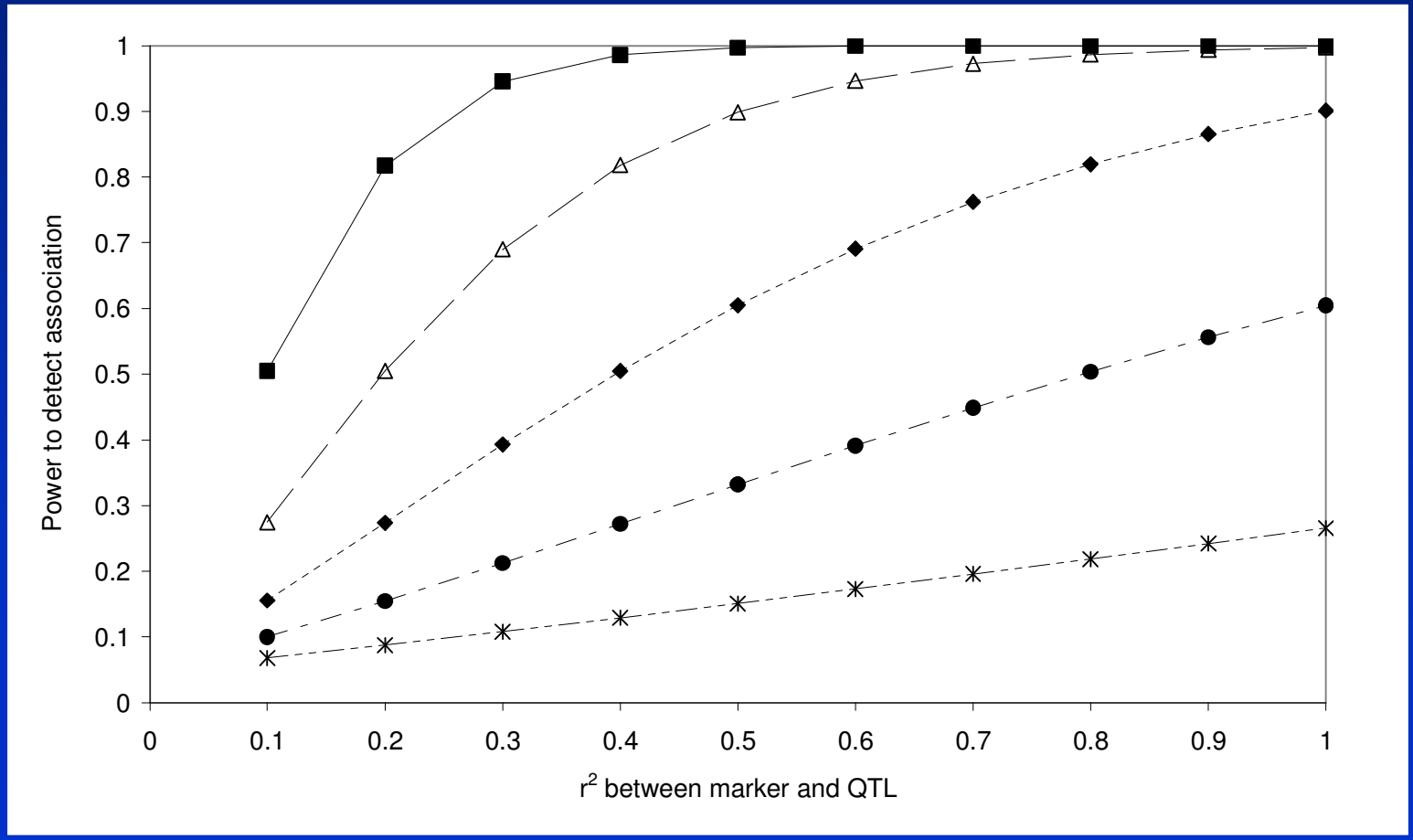
180 loci explain 10% of the variance

Most common h

inheritance: DNA sequence variants at many genetic loci influence the phenotype. Genome-wide association (GWA) studies have identified more than 600 variants associated with human traits[1], but these typically explain small fractions of phenotypic variation, raising questions about the use of further studies. Here, using 183,727 individuals, we show that hundreds of genetic variants, in at least 180 loci, influence adult height, a highly heritable and classic polygenic trait[2, 3]. The large number of loci reveals patterns with important implications for genetic studies of common human diseases and traits. First, the 180 loci are not random, but instead are enriched for genes

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)

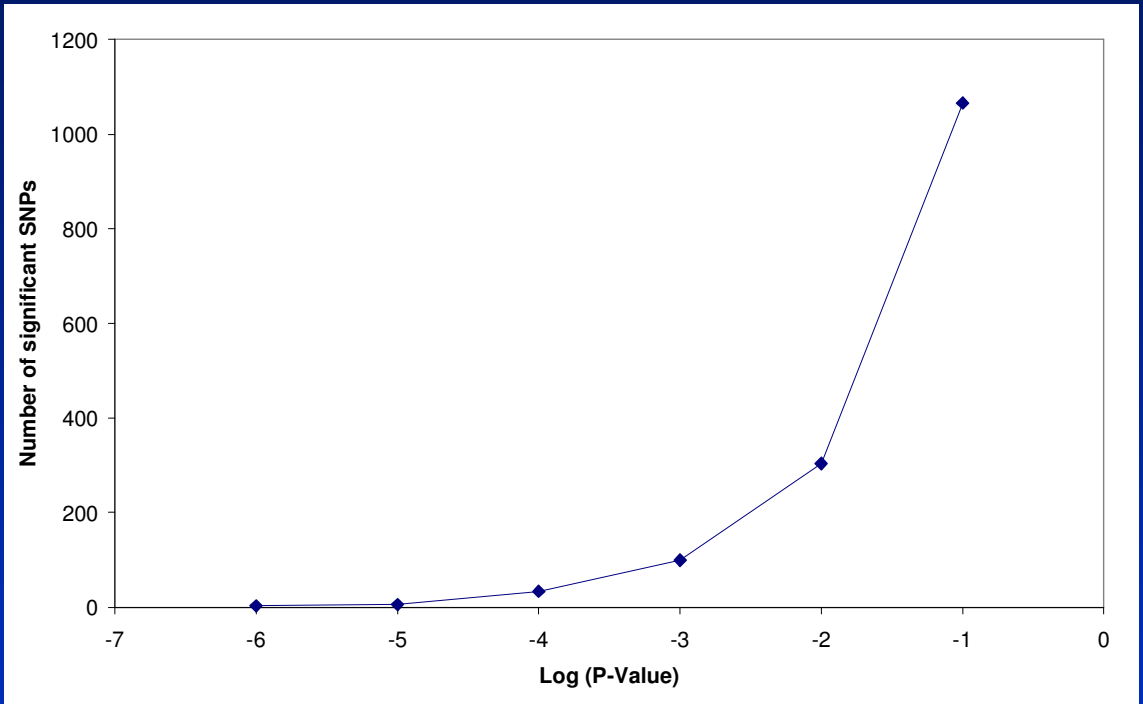
Power of GWAS

- 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

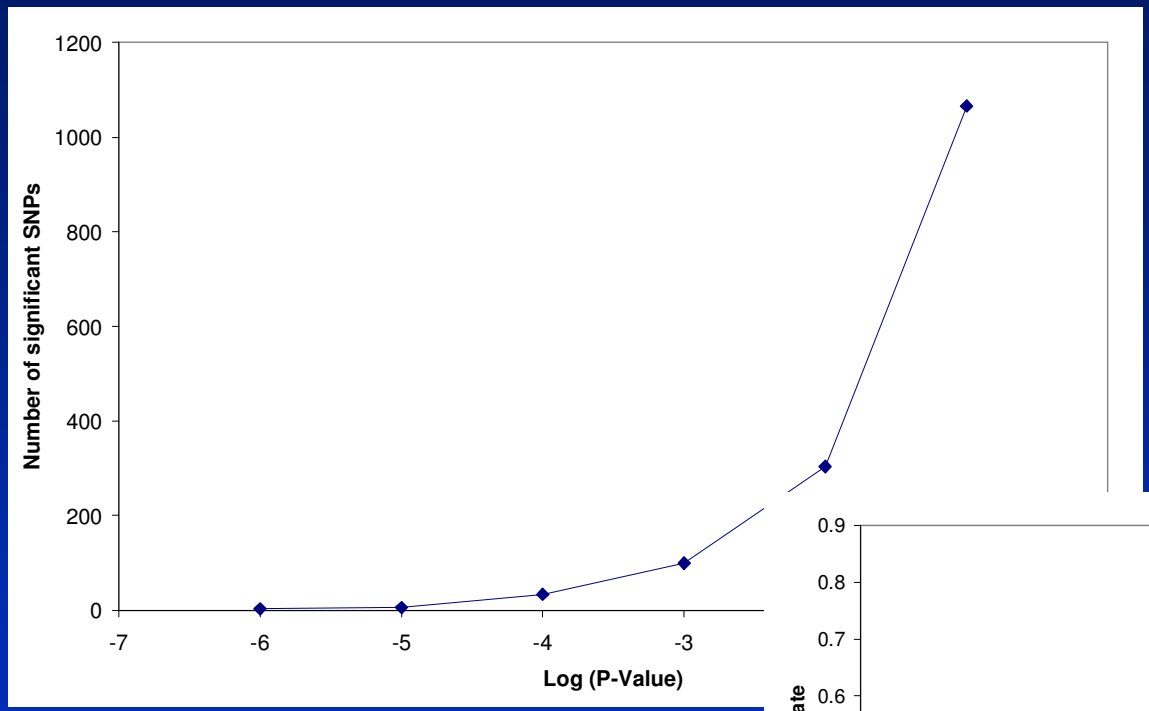
Power of GWAS

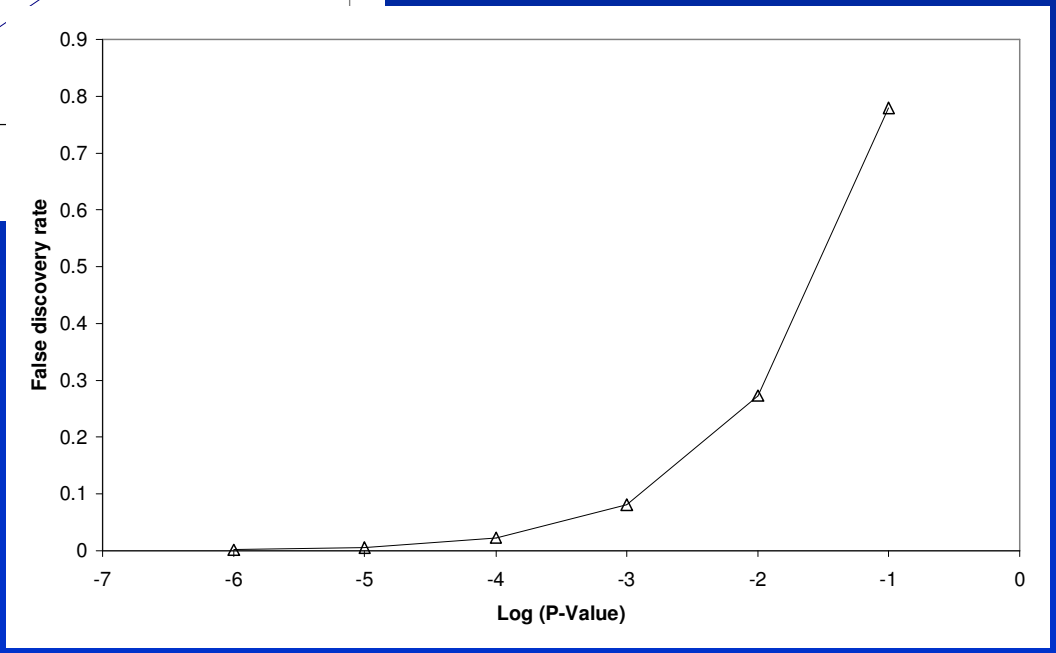
- An alternative is to choose a significance level with an acceptable false discovery rate (FDR)
- Proportion of significant results which are really false positives
- FDR = mP/n
  - m = number of markers tested
  - P = significance level (eg. P=0.01)
  - n = number of markers actually significant
-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

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

- 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…..

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

- 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

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

$$\mathbf{X} \quad \begin{matrix} 1 \\ 2 \\ 2 \\ 1 \\ 1 \\ 1 \end{matrix}$$

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

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

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

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

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

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

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

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

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

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

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

$\lambda = 0.33$

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

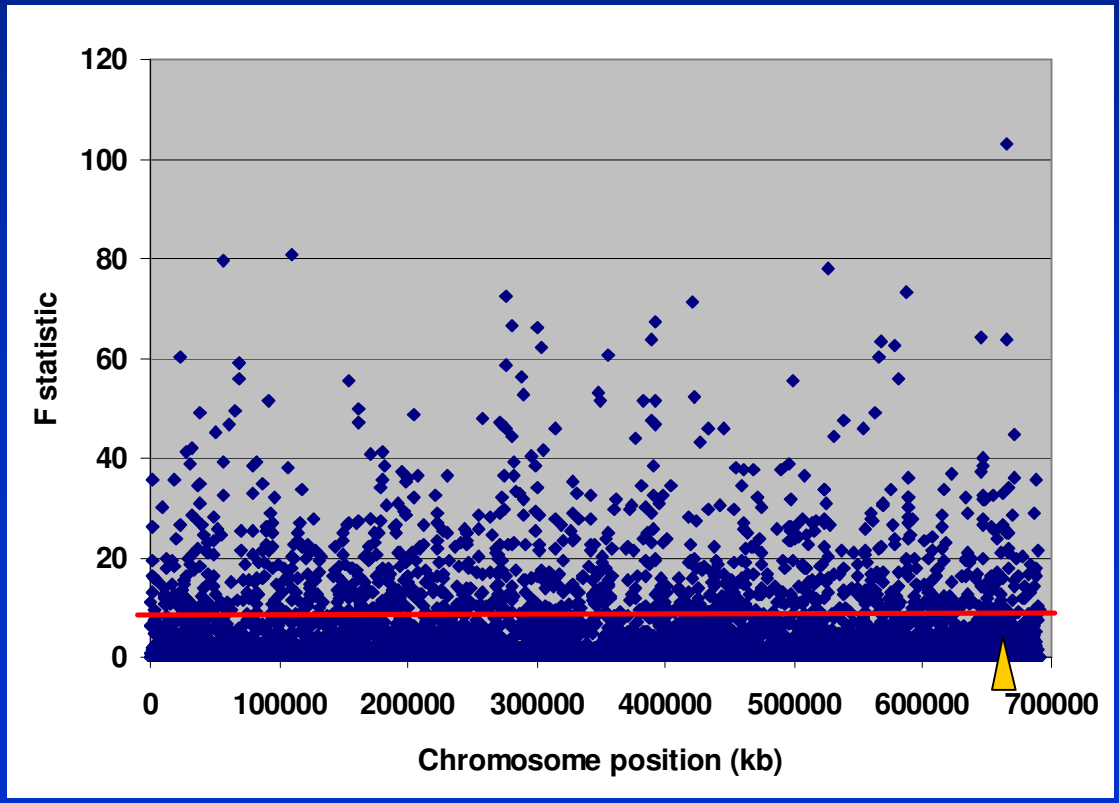
- Example

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

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

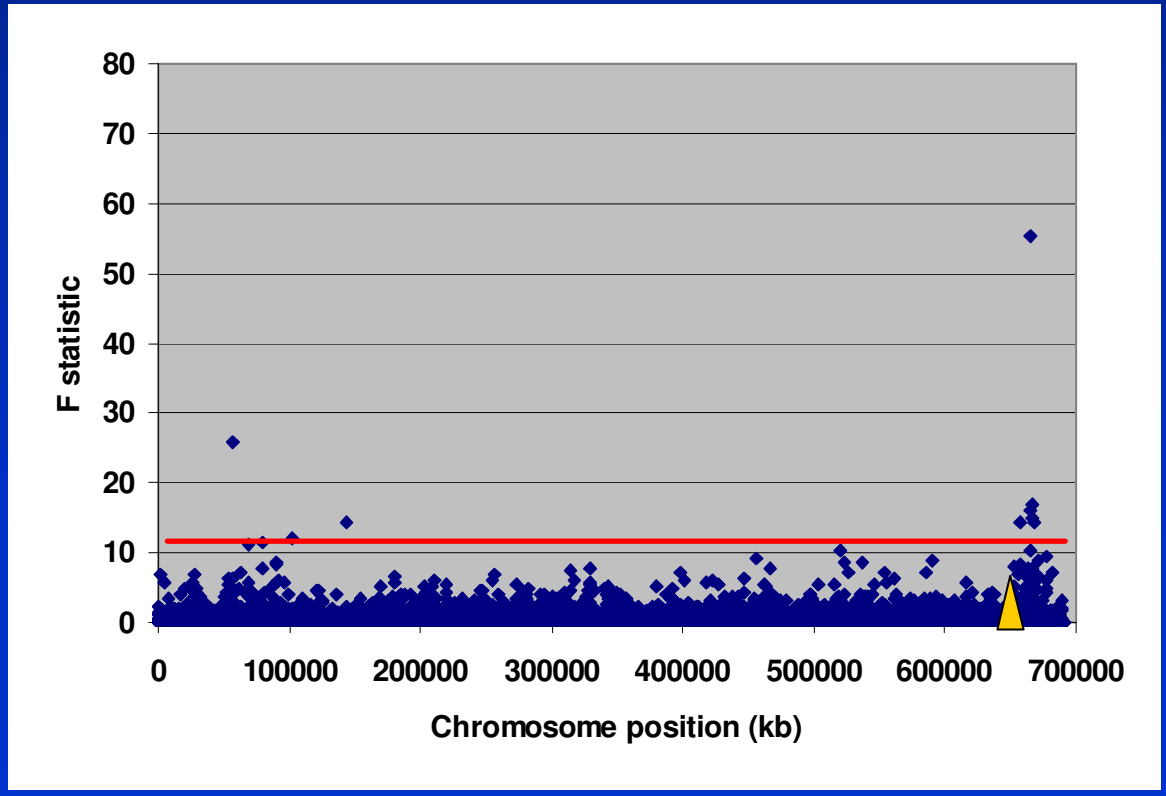
Population structure

- 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

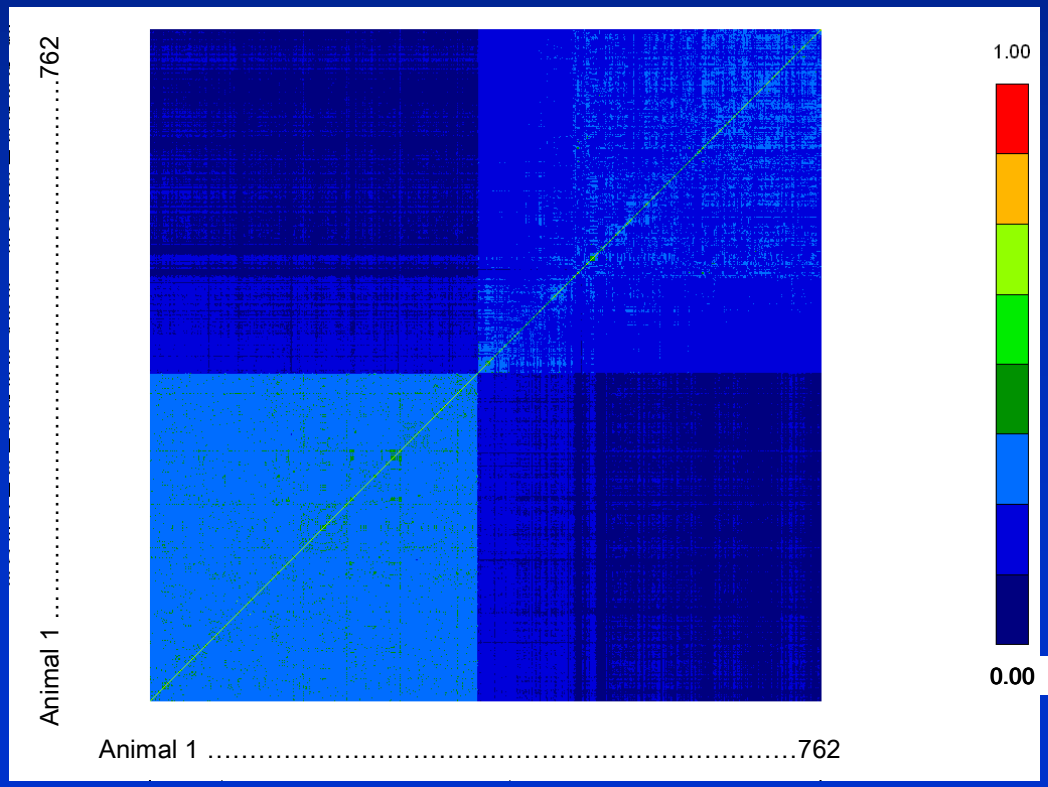
Population structure

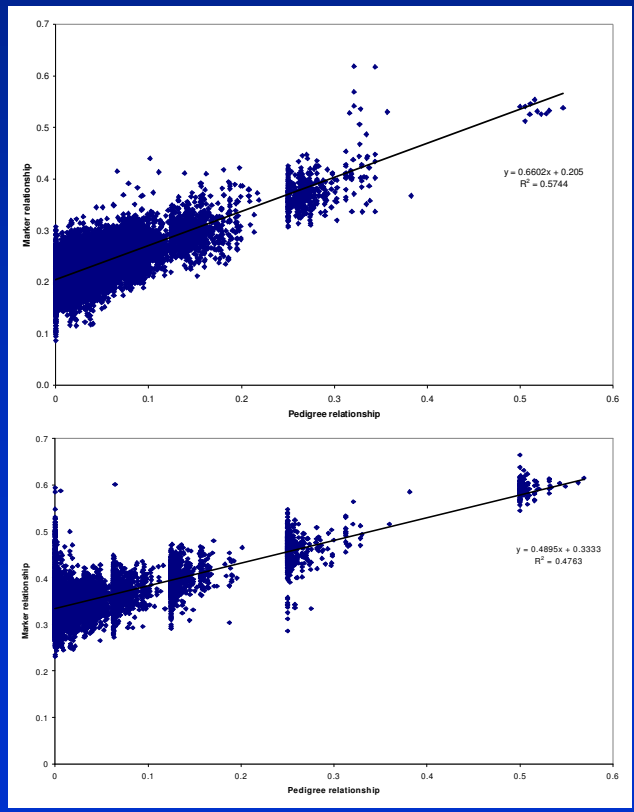
Number of false positives……

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





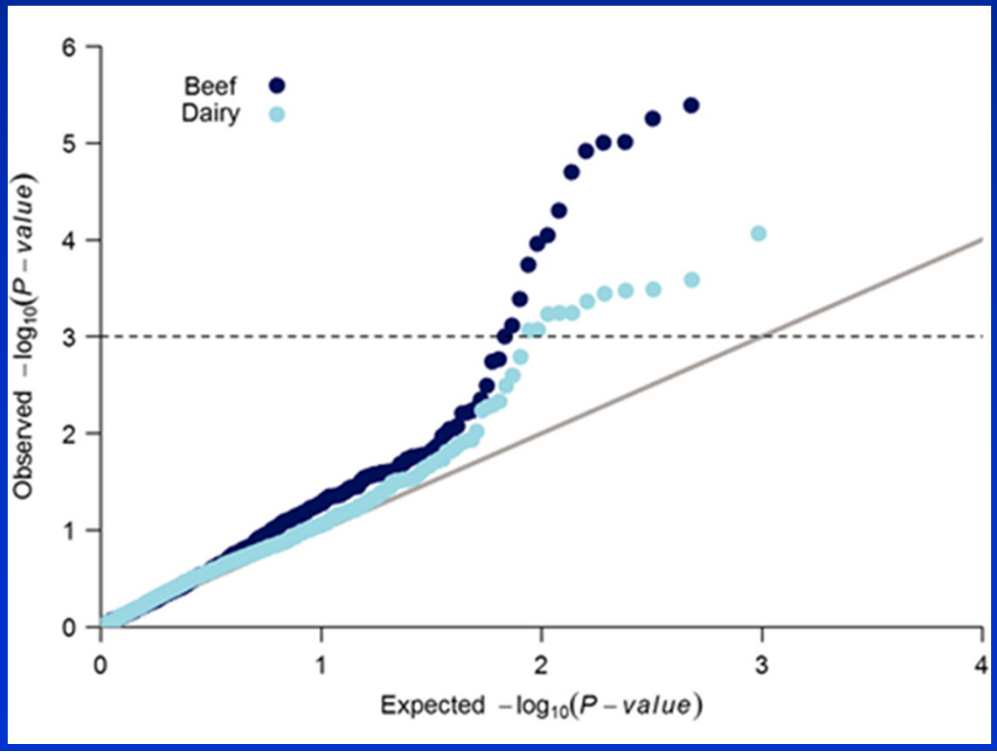
Genomic relationship matrix

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

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

Population structure

- Use a Quantile-quantile (QQ) plot to assess if we have accounted for population structure
- Rank SNPs on observed, –log10(Pvalue), then plot observed against expected
- Population structure removed if observed, expected approximately equal for large P values



Genome wide association

- Association testing with single marker regression
- Power of genome wide association studies
- Accounting for population structure
- 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
- Any population structure results in spurious associations

- Validation, validation, validation