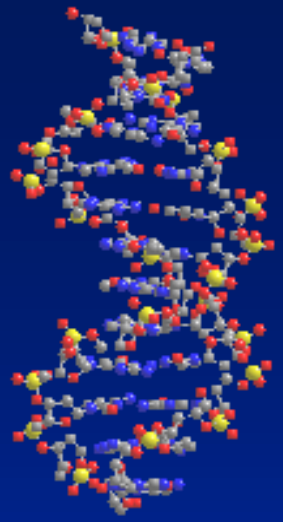
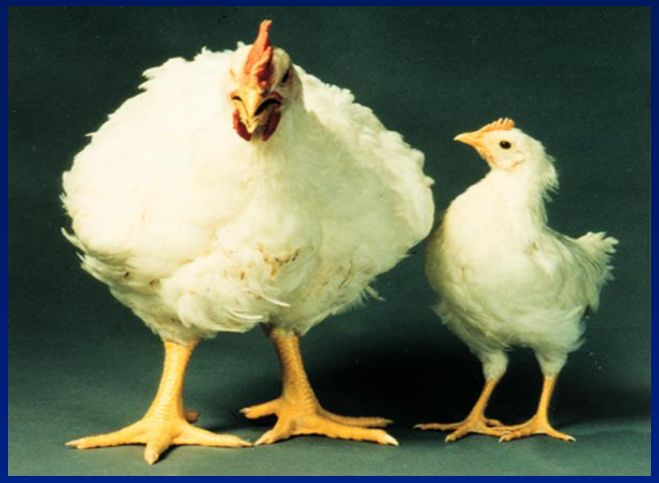
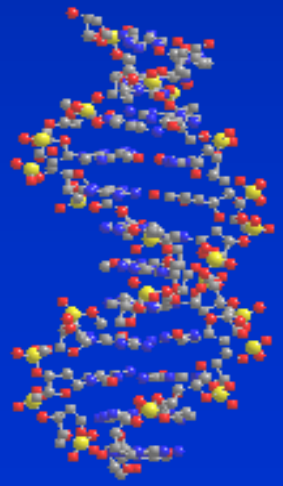




# From sequence data to genomic prediction


# Course overview

- Day 1
  - Introduction
  - Generation, quality control, alignment of sequence data
  - Detection of variants, quality control and filtering
- Day 2
  - Imputation from SNP array genotypes to sequence data
- Day 3
  - Genome wide association studies with SNP array and sequence variant genotypes
- Day 4 & 5
  - Genomic prediction with SNP array and sequence variant genotypes (BLUP and Bayesian methods)
  - Use of genomic selection in breeding programs

# Genome wide association

- Aim
  - With SNP arrays: find markers in high linkage disequilibrium with causative mutations -> candidate genes
  - With sequence data: find causative mutations (?)
  - Put these on SNP chip, GBS designs

# Genome wide association

- Linkage disequilibrium
- Models for GWAS
- Factors affecting accuracy of GWAS
- Accounting for population structure
- Examples with sequence – can we find causative mutations?
- Using biological information

# Definitions of LD

- Genome wide association studies with SNP arrays exploit linkage disequilibrium with common SNP and QTL

# 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…….

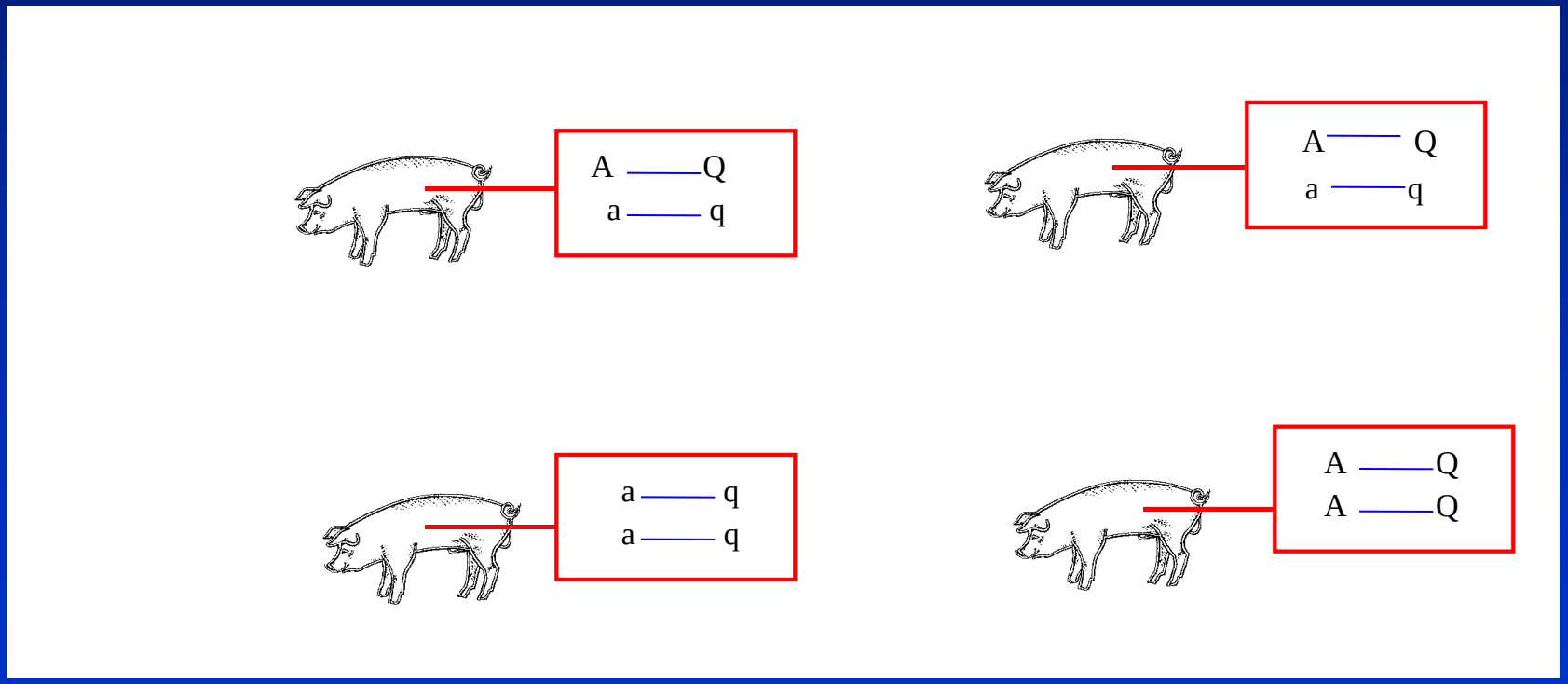
| | | *Marker A* | | |
|---|---|---|---|---|
| | | A1 | A2 | Frequency |
| *Marker B* | B1 | 0.25 | 0.25 | 0.5 |
| | B2 | 0.25 | 0.25 | 0.5 |
| | Frequency | 0.5 | 0.5 | |

# 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)]

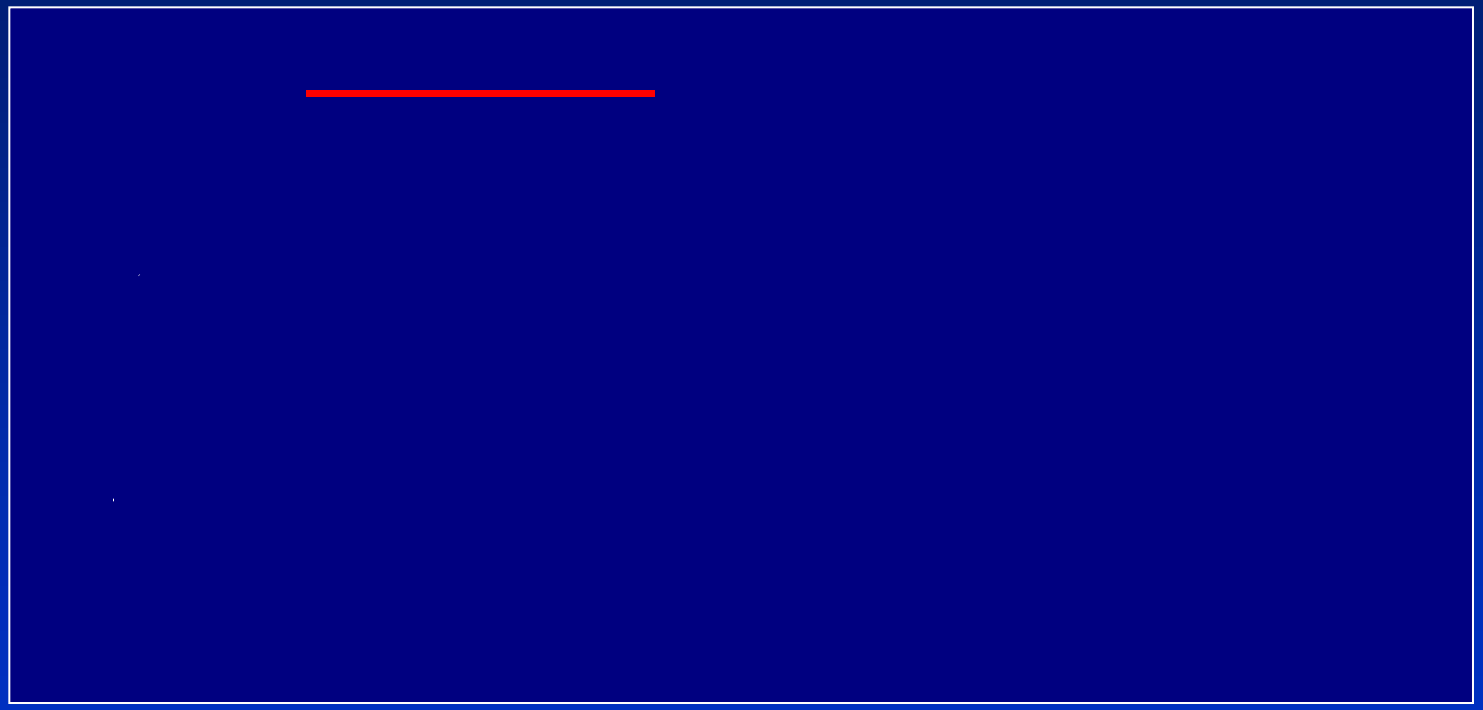
Values between 0 and 1.

# 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 $40kg^2$.

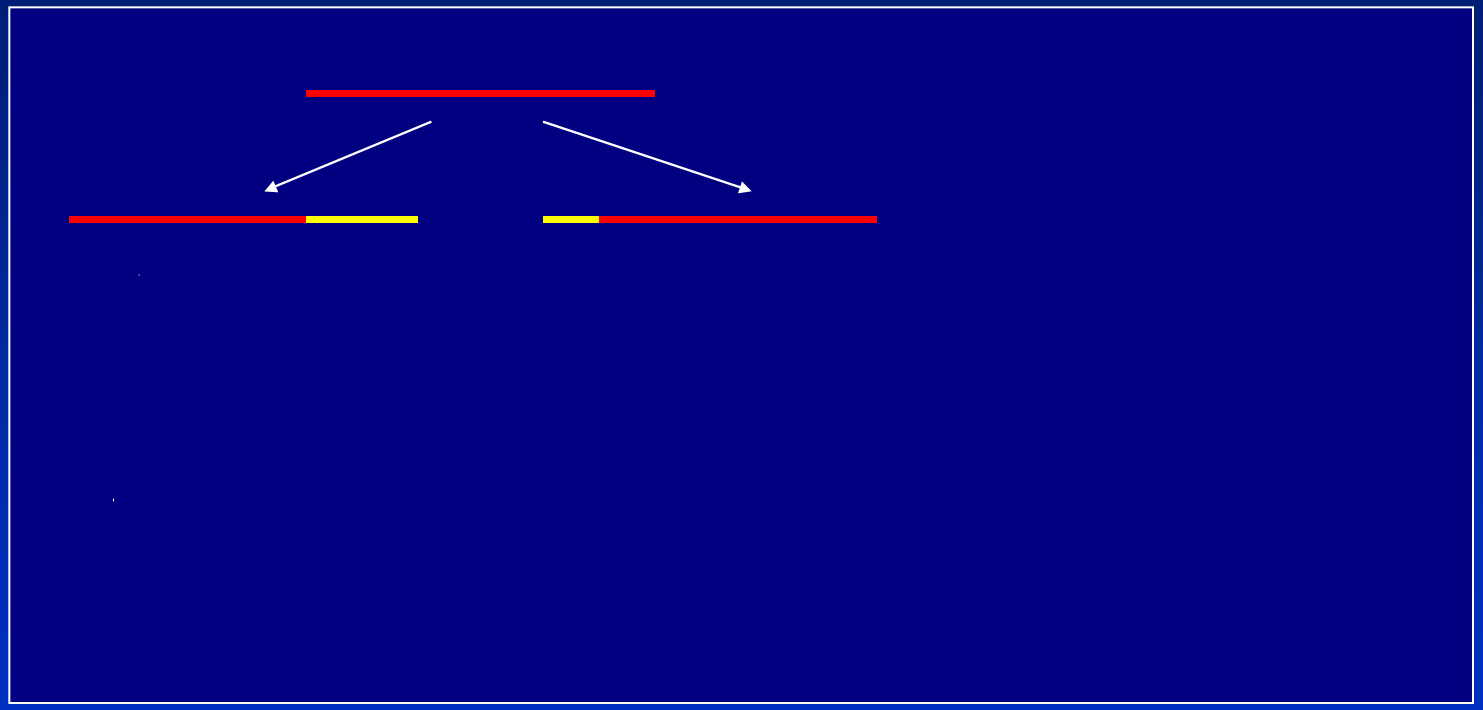
# Causes of LD

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



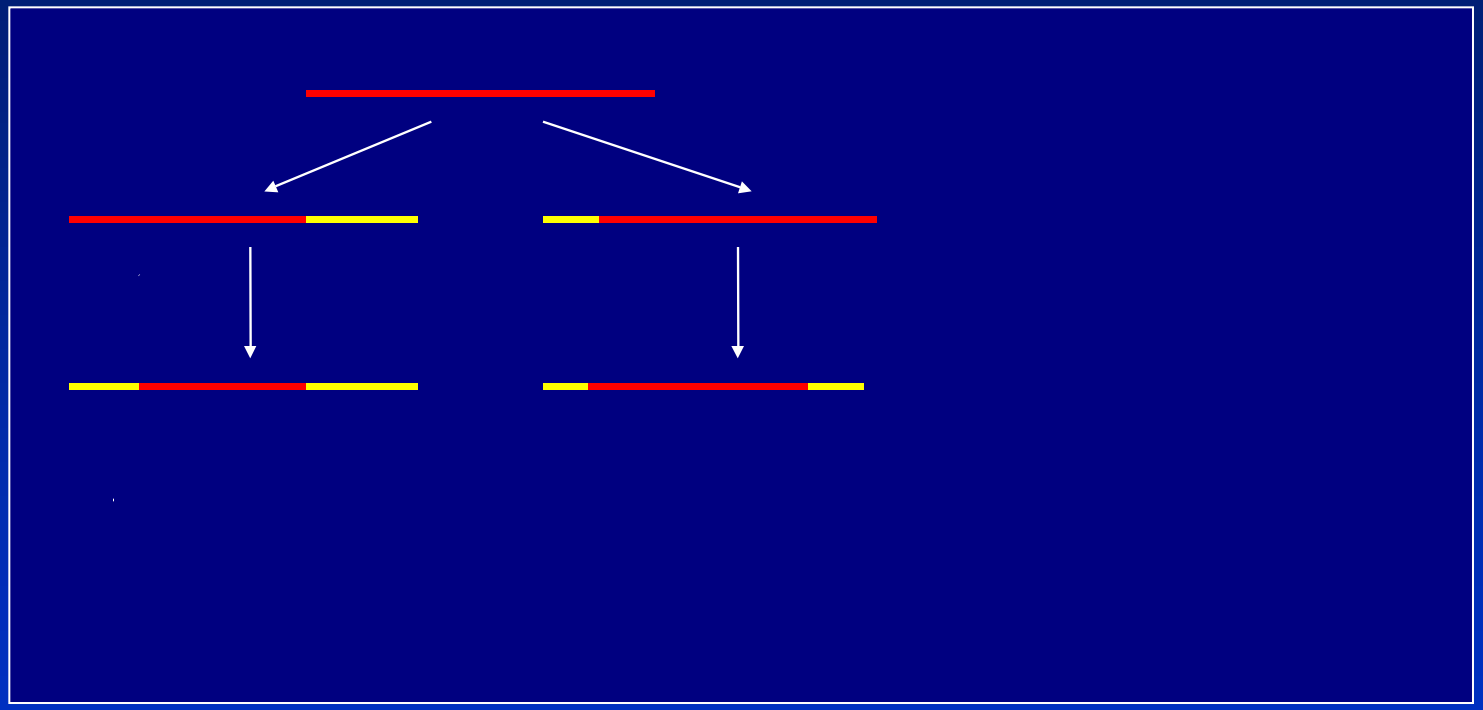
# Causes of LD

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



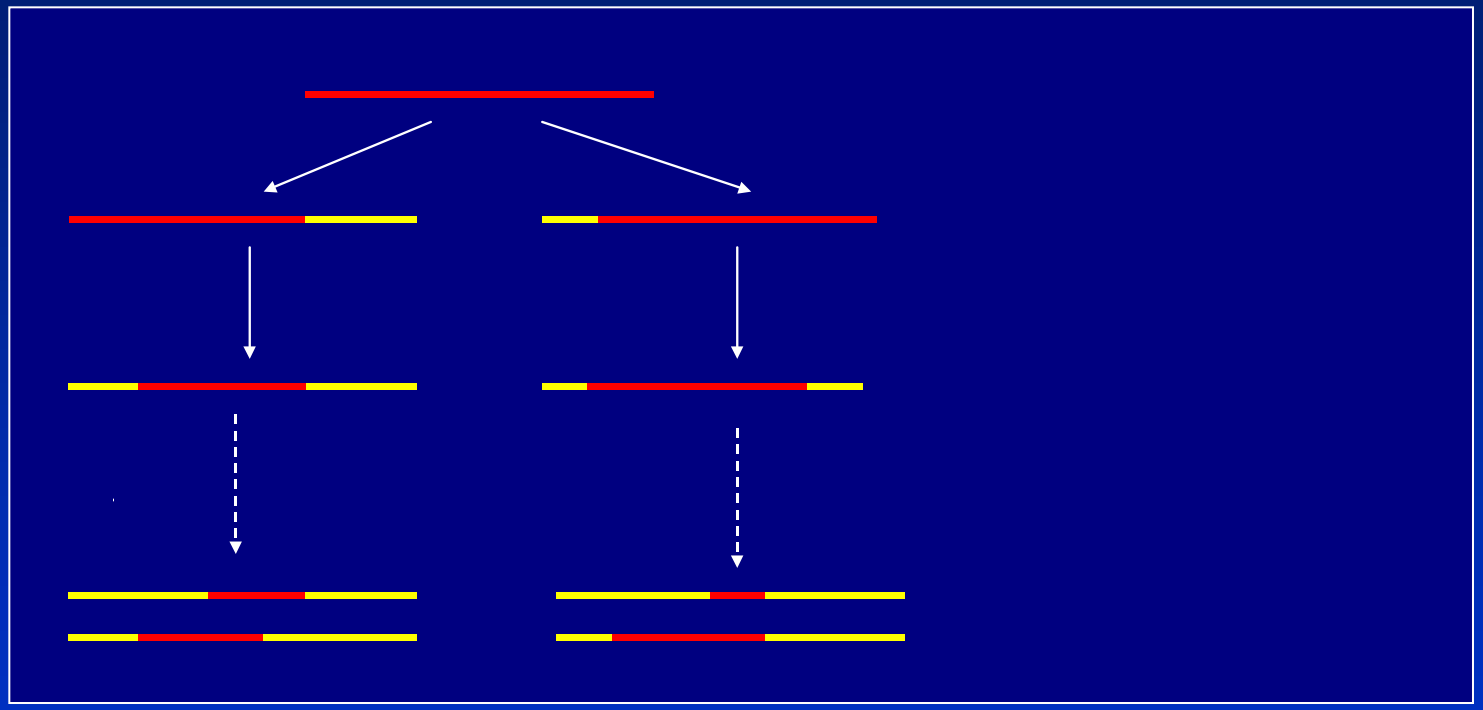
# Causes of LD

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



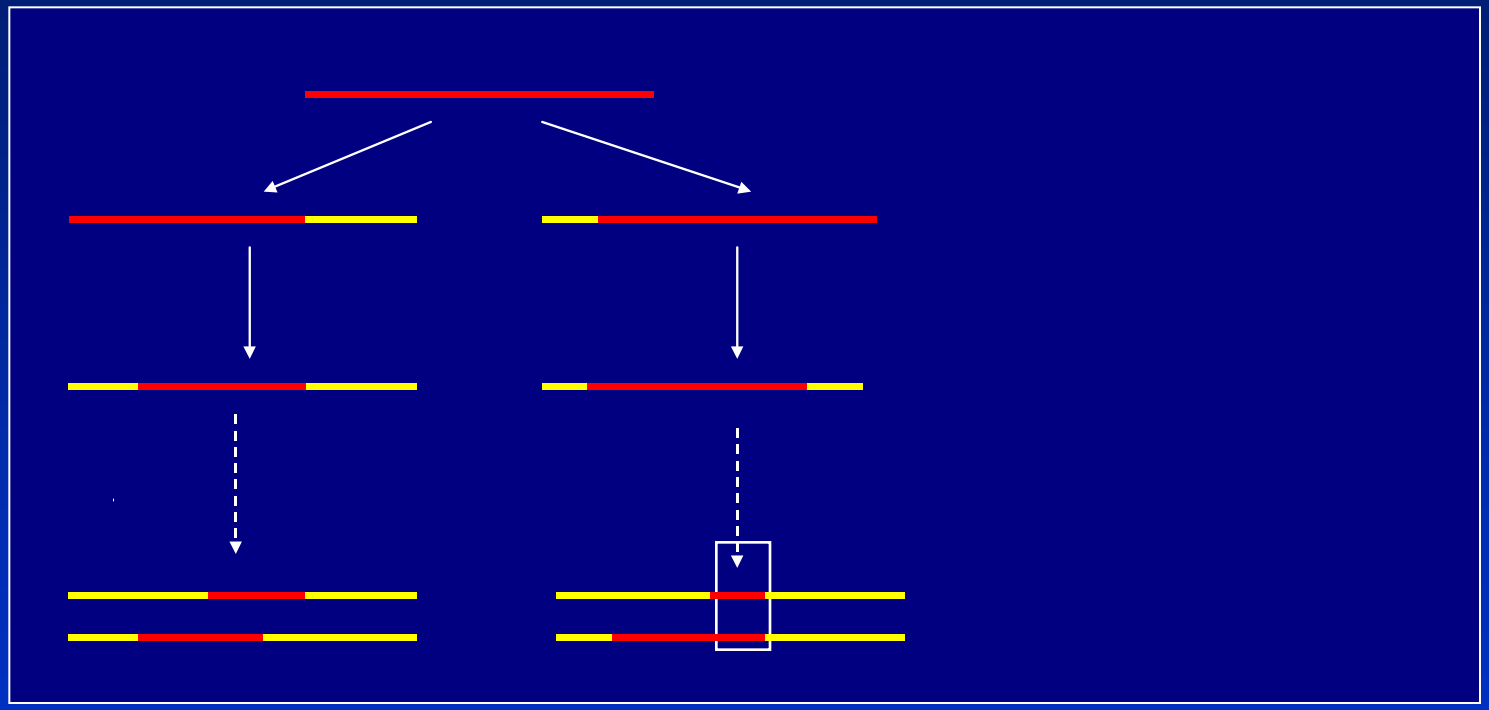
# Causes of LD

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



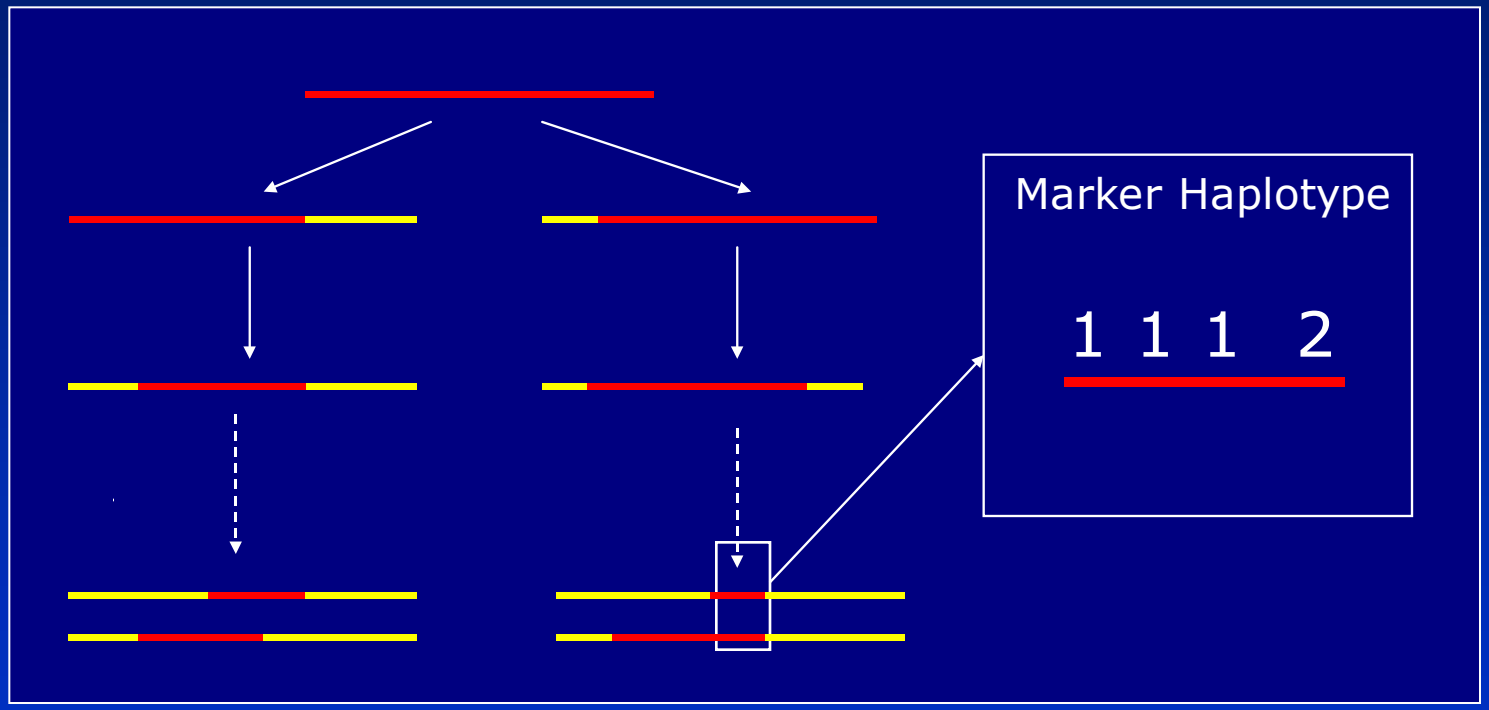
# Causes of LD

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



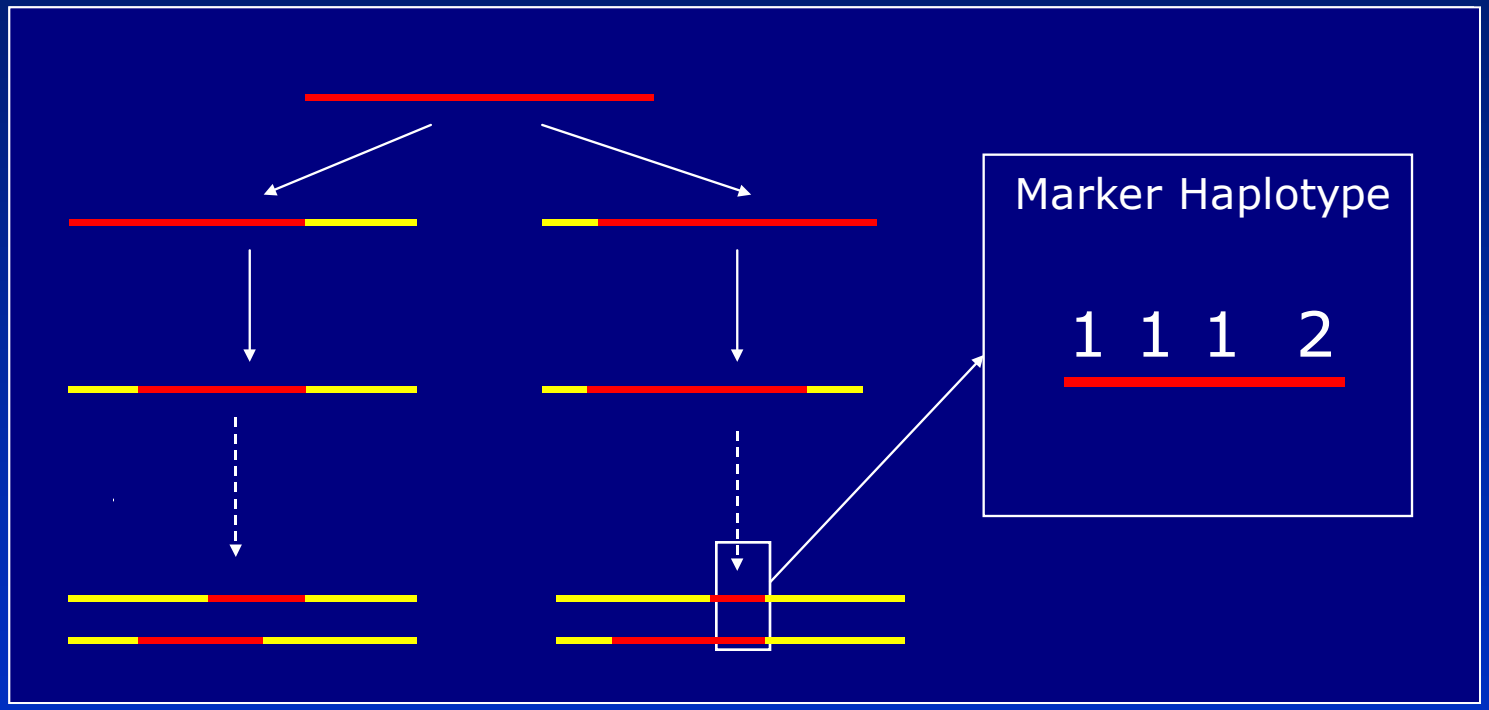
# Causes of LD

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



# Causes of LD

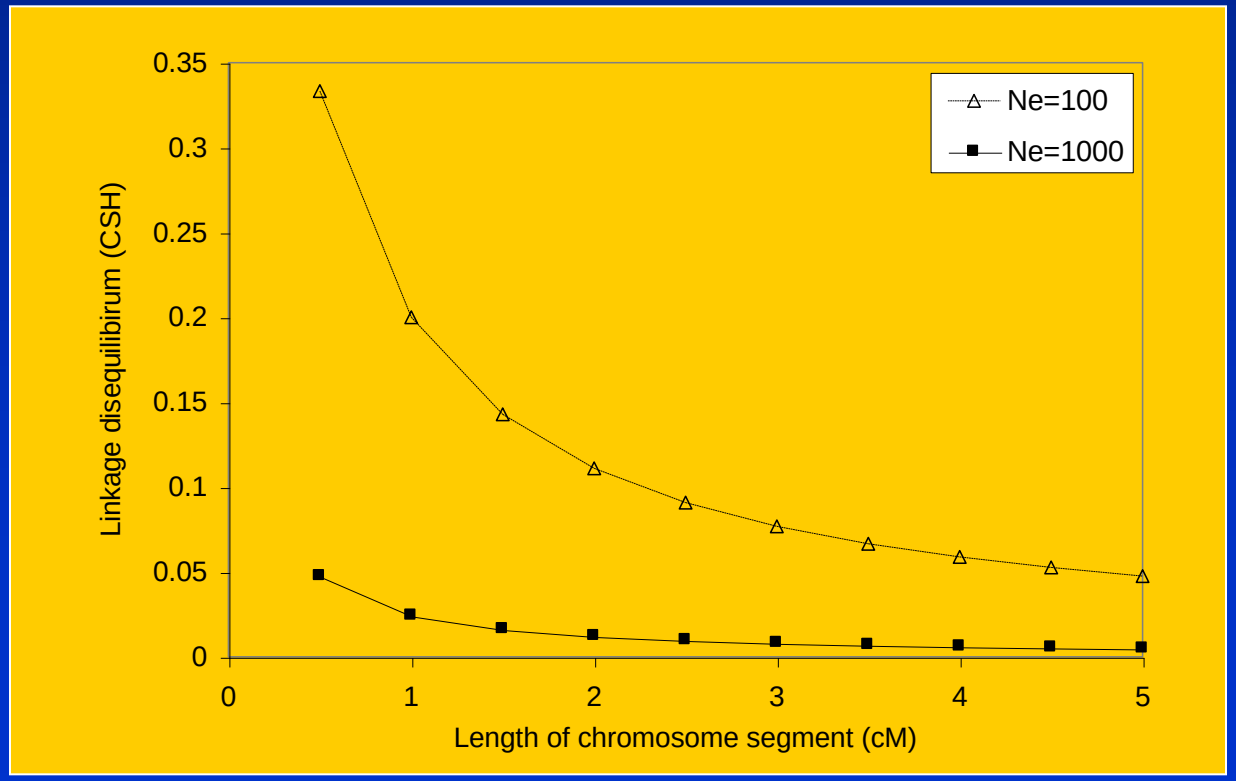
- A chunk of ancestral chromosome is conserved in the current population




- Size of conserved chunks depends on effective population size

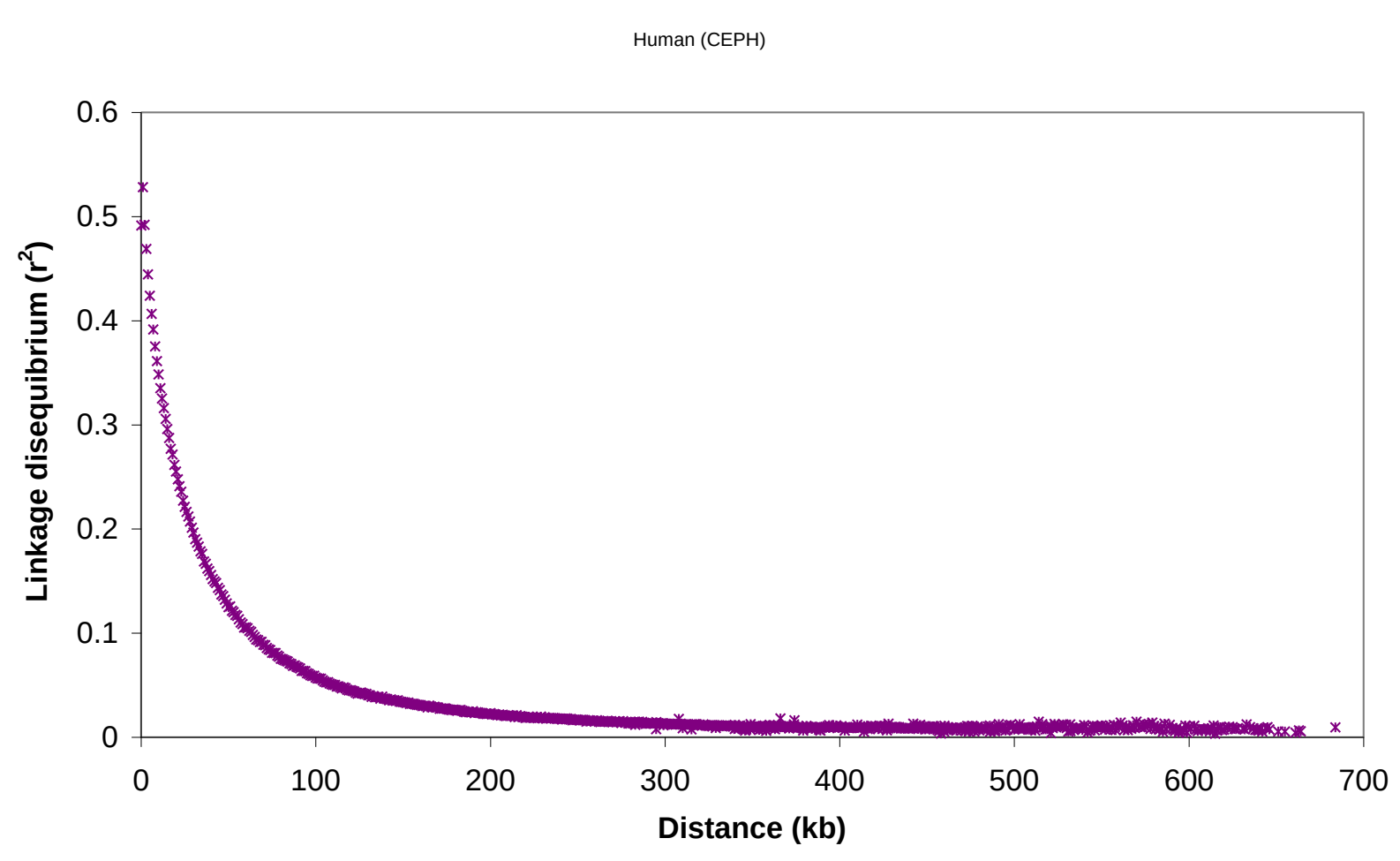
# Causes of LD

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



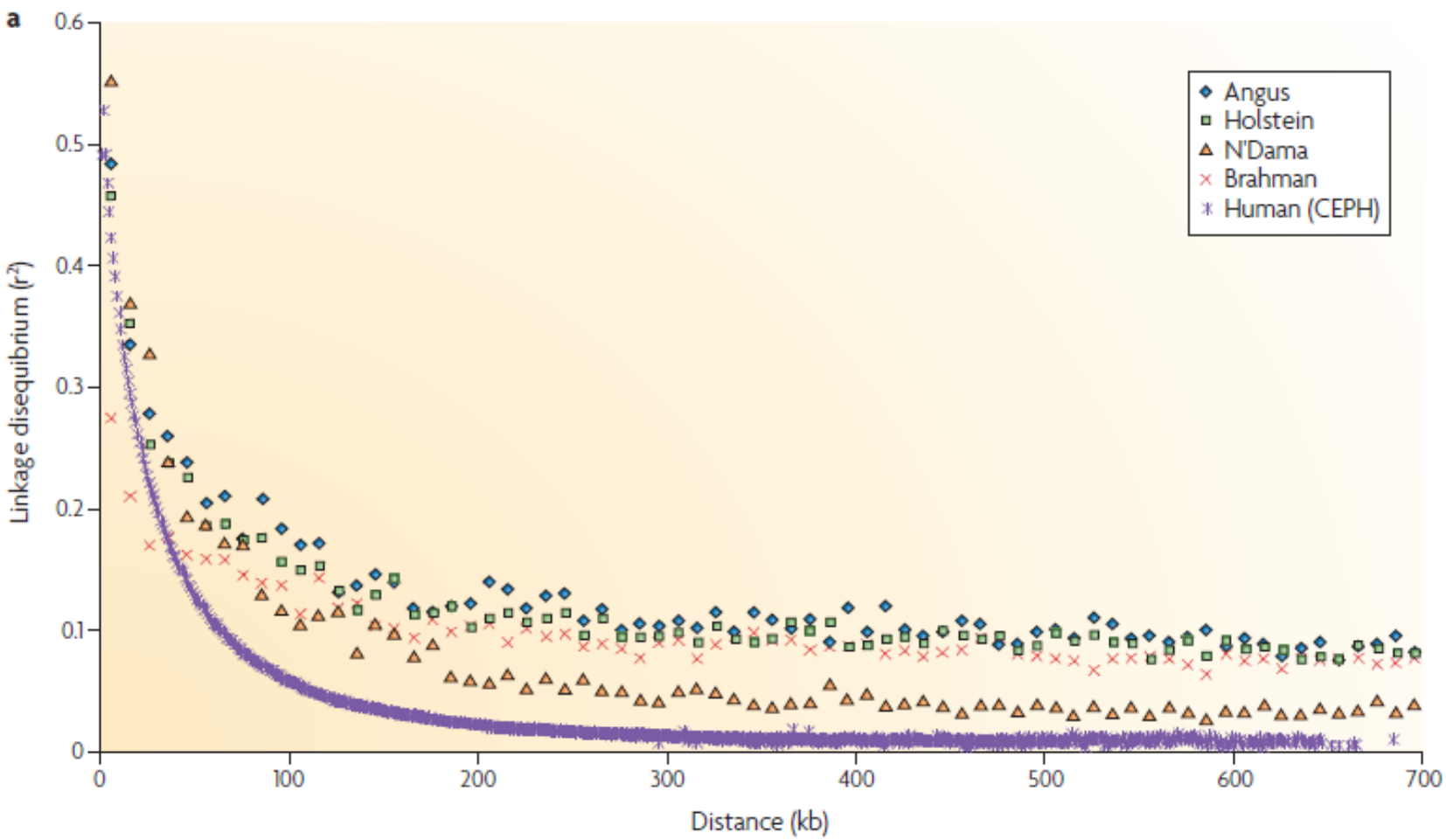
# Extent of LD in humans and livestock

Humans…….(Tenesa et al. 2007)



# Extent of LD in humans and livestock

And cattle……

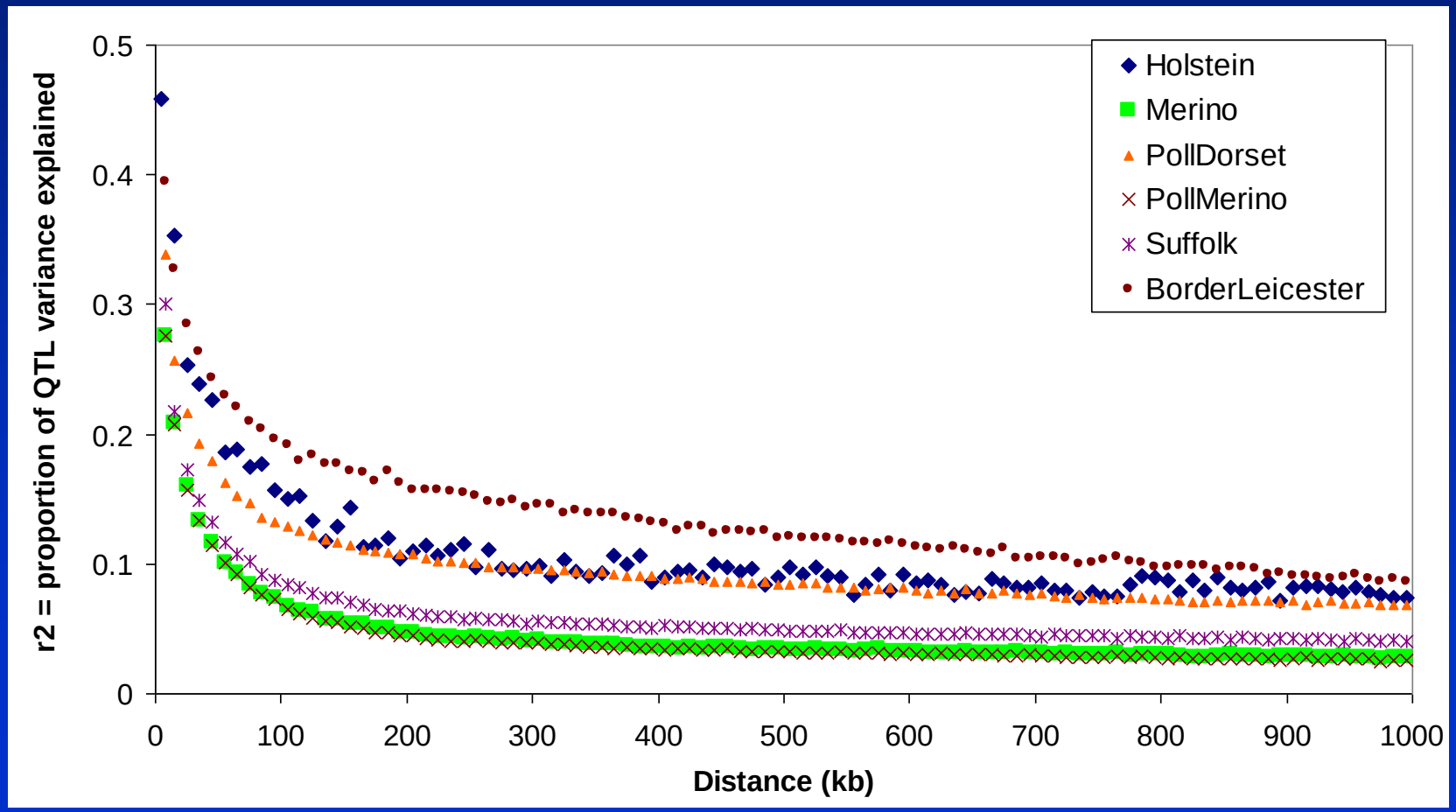


# Implications?

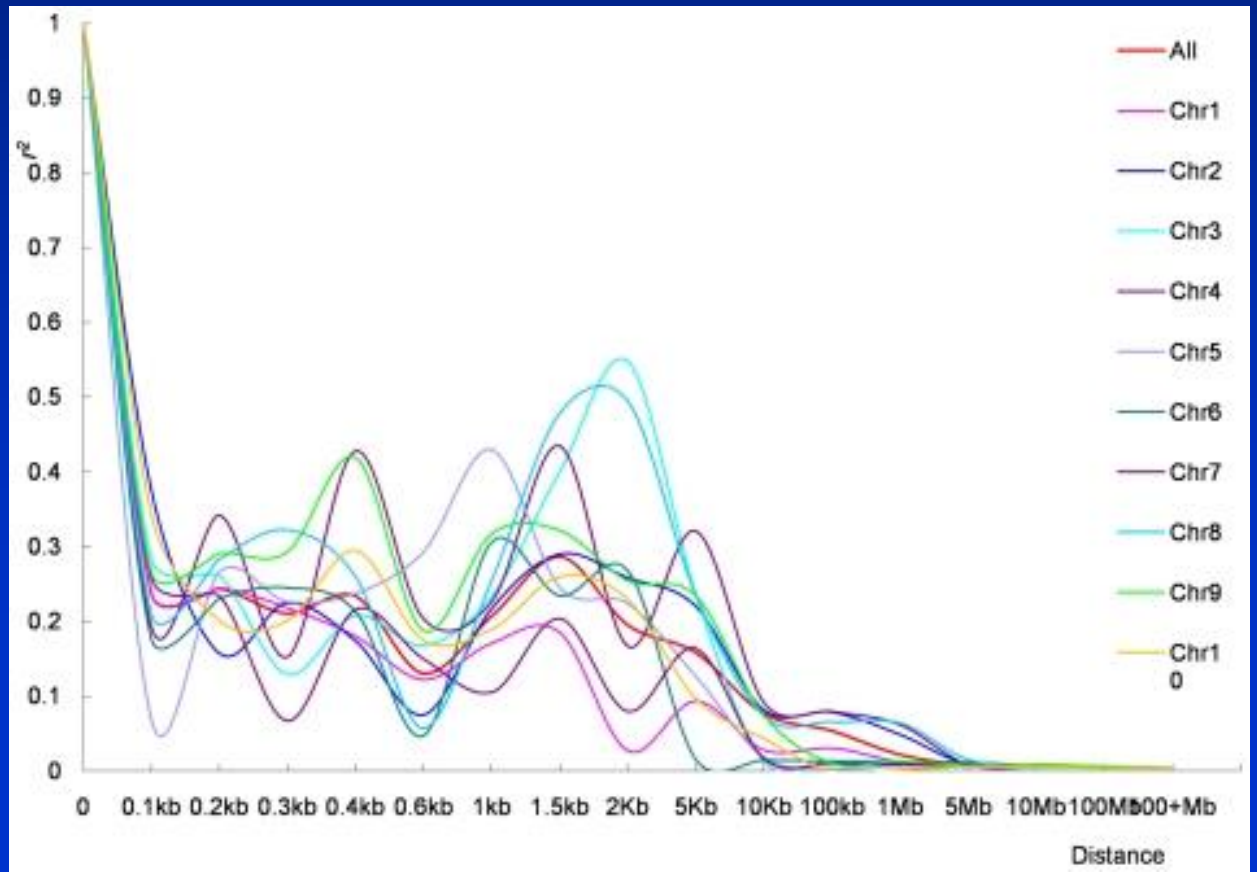
- In Holsteins, need a marker approximately every 10kb to get average $r^2$ of 0.5 between marker and QTL
- ~ 300K SNP

# Extent of LD in other species

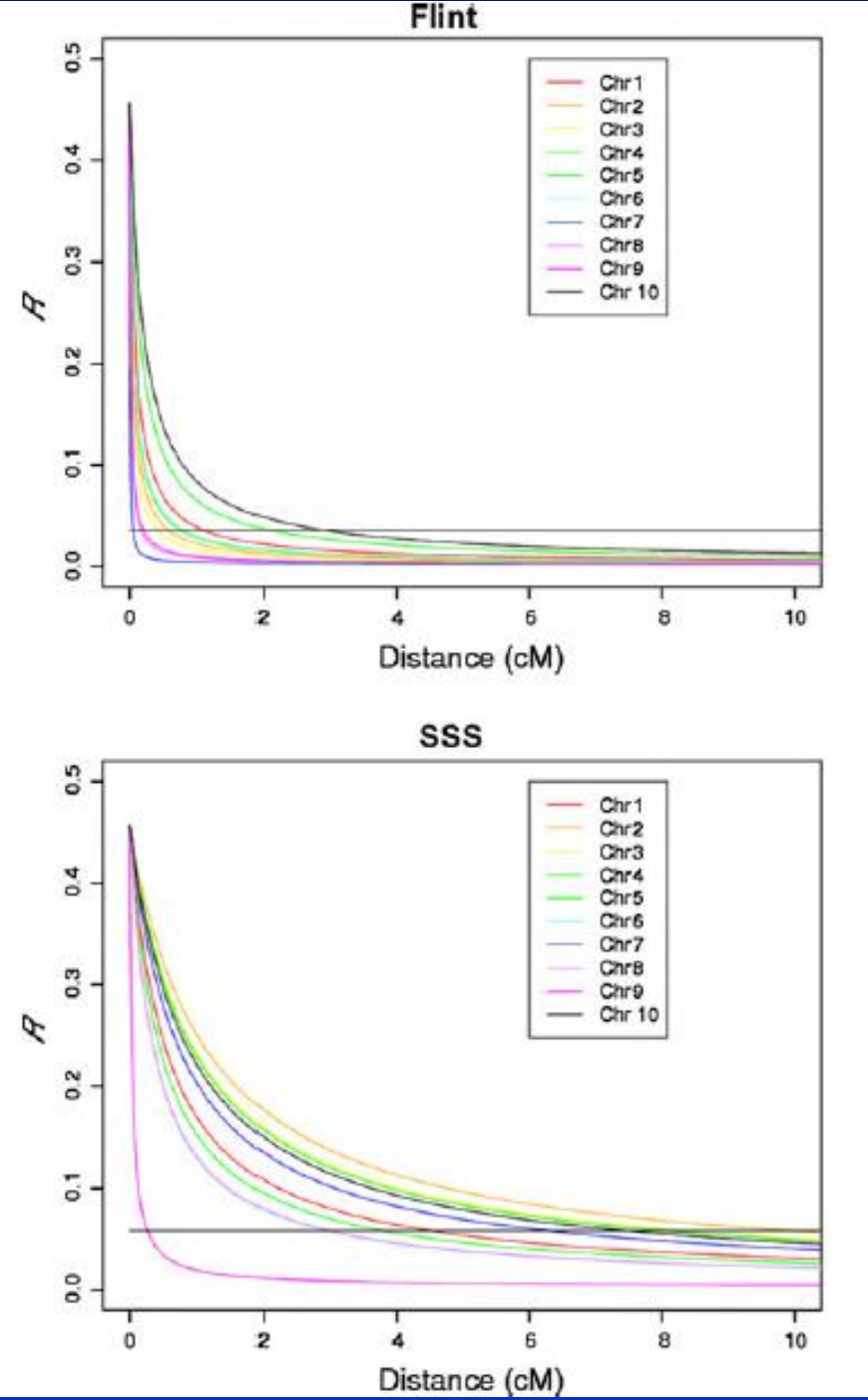
- Sheep HapMap project (Kijas et al. 2011)



- 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

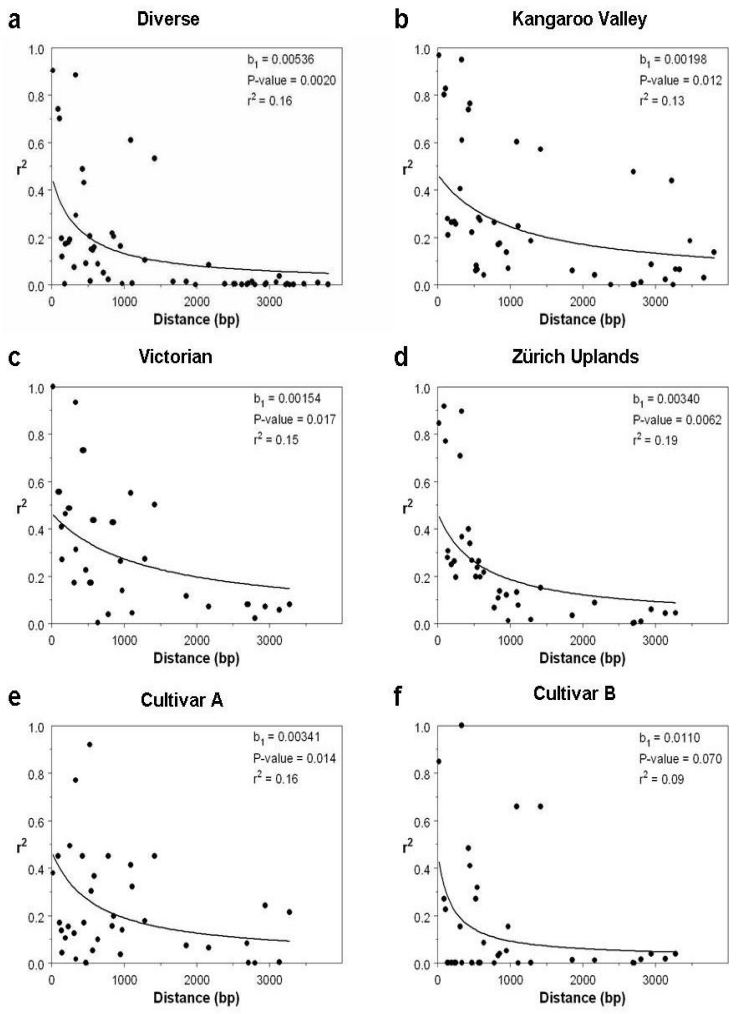


- Maize (ii)
  - Van Ingehlandt et al. 2011 TAG 123:11
  - Considerable LD among heterotic groups
  - Concluded 4000-65,000 SNPs needed for genome wide association



# 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 GWAS/genomic prediction
- In cattle, $r^2$~0.4 at 5kb ~ 300 000 markers necessary for GWAS
- In humans, LD lower, need many more markers

# Genome wide association

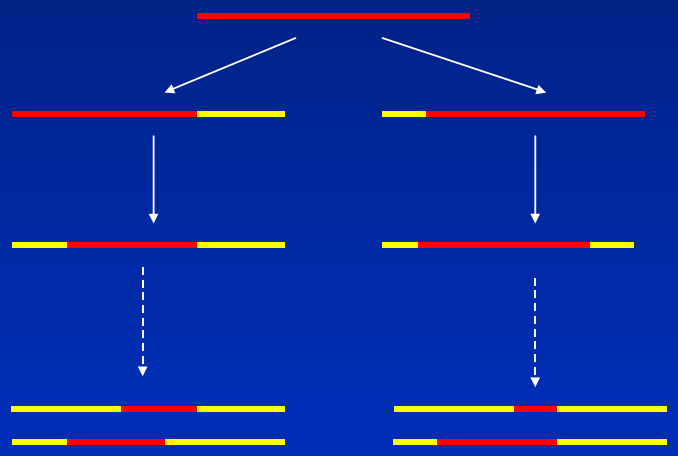
- Linkage disequilibrium
- Models for GWAS
- Factors affecting accuracy of GWAS
- Accounting for population structure
- Examples with sequence – can we find causative mutations?
- Using biological information

# Genome wide association

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

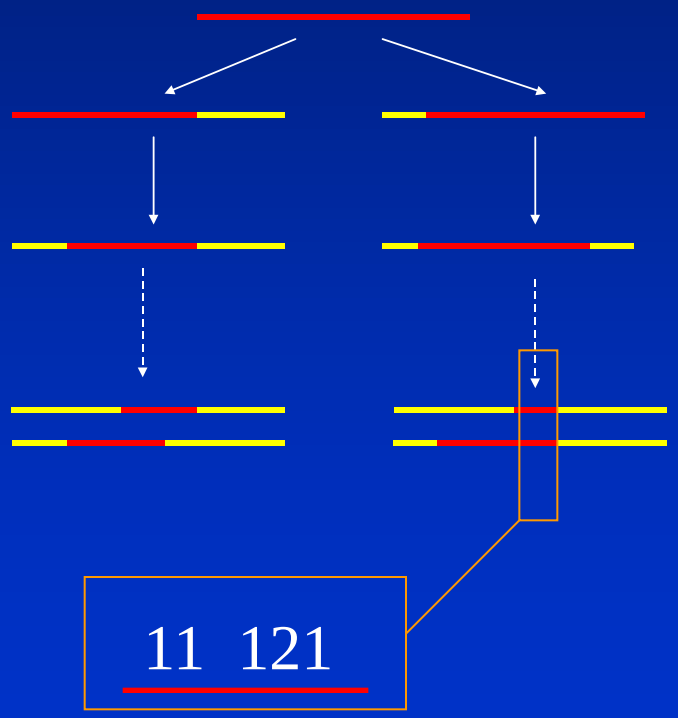
# Genome wide association

- 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



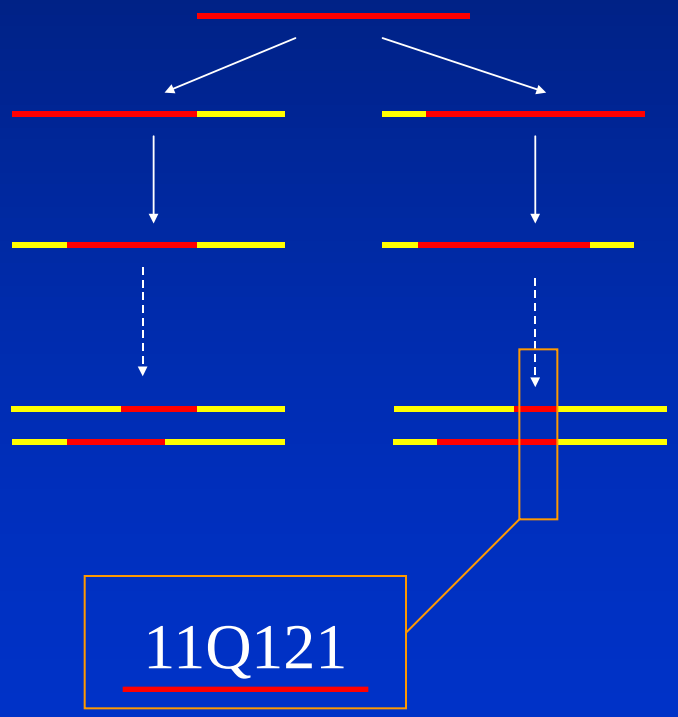
# Genome wide association

- 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



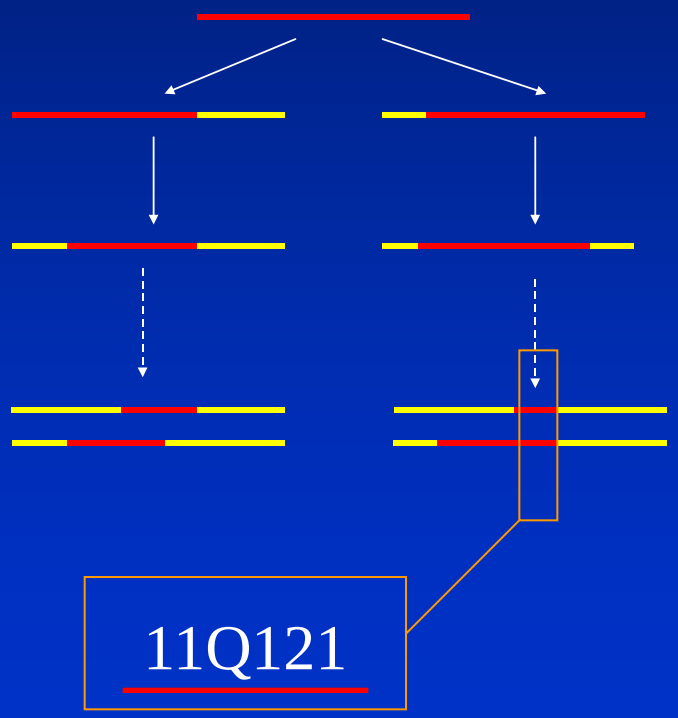
# Genome wide association

- 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.

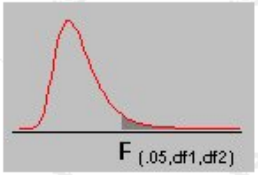
# Single marker regression

- Is the marker effect significant?
- ***F*** statistic comparing between marker variance to within marker variance
- Test against tabulated value for $F_{\alpha,v1,v2}$
  - $\alpha$= significance value
  - v1=1 (1 marker effect for regression)
  - v2=8 (number of records -2)

# Single marker regression

- In our simple example
  - $F_{data}=4.56$
  - $F_{0.05,1,8}=5.12$
- 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 |

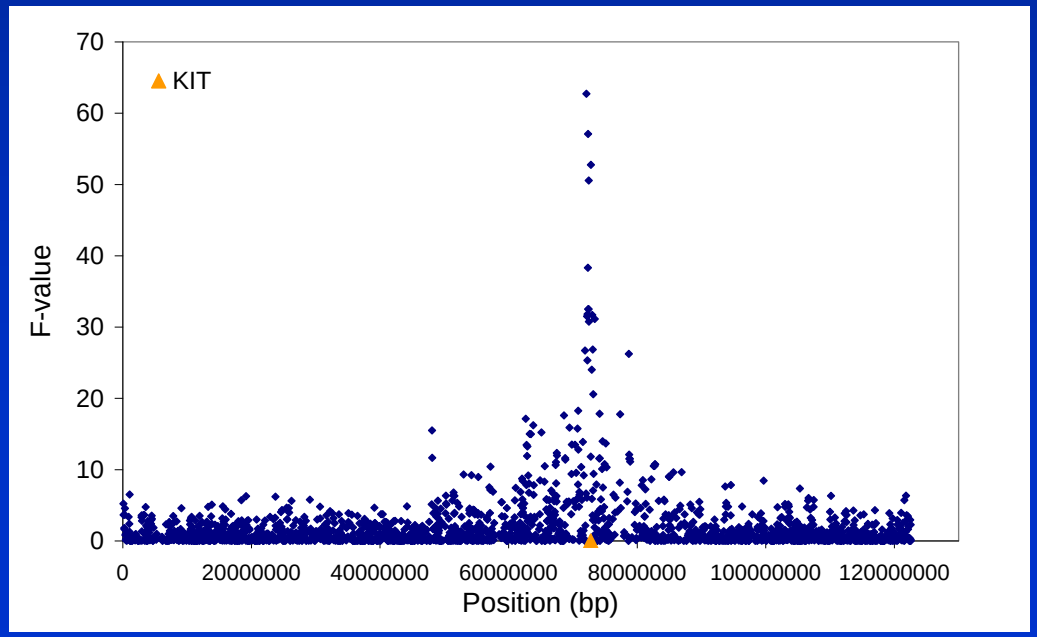
# Proportion of black….



- 600 Holstein-Friesian dairy bulls scored proportion of black
- genotyped for 50 000 SNPs
- Single marker regression

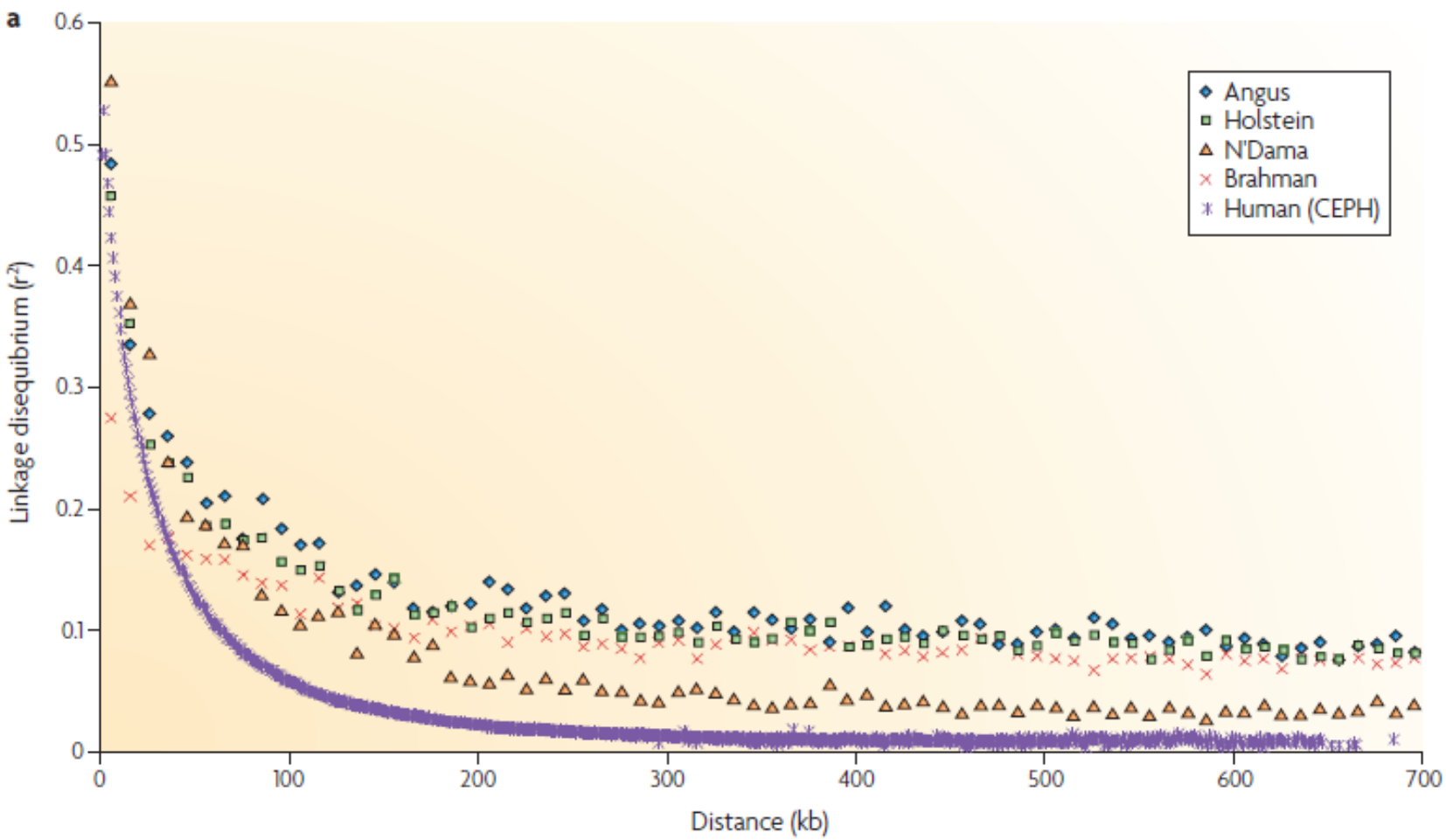
# Proportion of black….





# Extent of LD in humans and livestock

And cattle……



# Genome wide association

- Linkage disequilibrium
- Models for GWAS
- Factors affecting accuracy of GWAS
- Accounting for population structure
- Examples with sequence – can we find causative mutations?
- Using biological information

# 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 individuals 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

# Power of GWAS

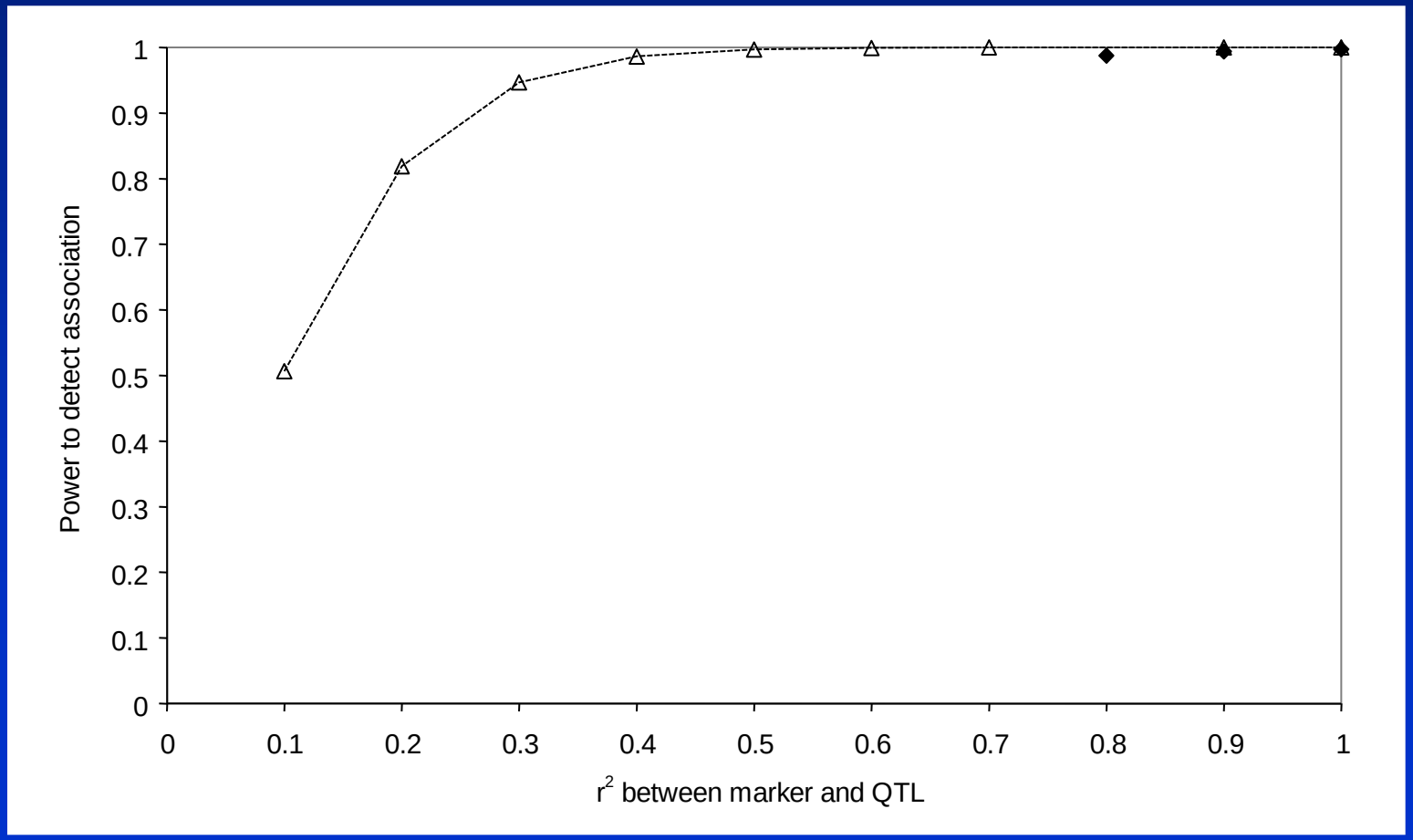
- Power is a function of:
  - $r^2$ between the marker and QTL
    - sample size must be increased by $1/r^2$ to detect an un-genotyped QTL, compared with sample size for testing QTL itself
  - Proportion of total phenotypic variance explained by the QTL
  - Number of phenotypic records

# 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 $\alpha$ set by the experimenter

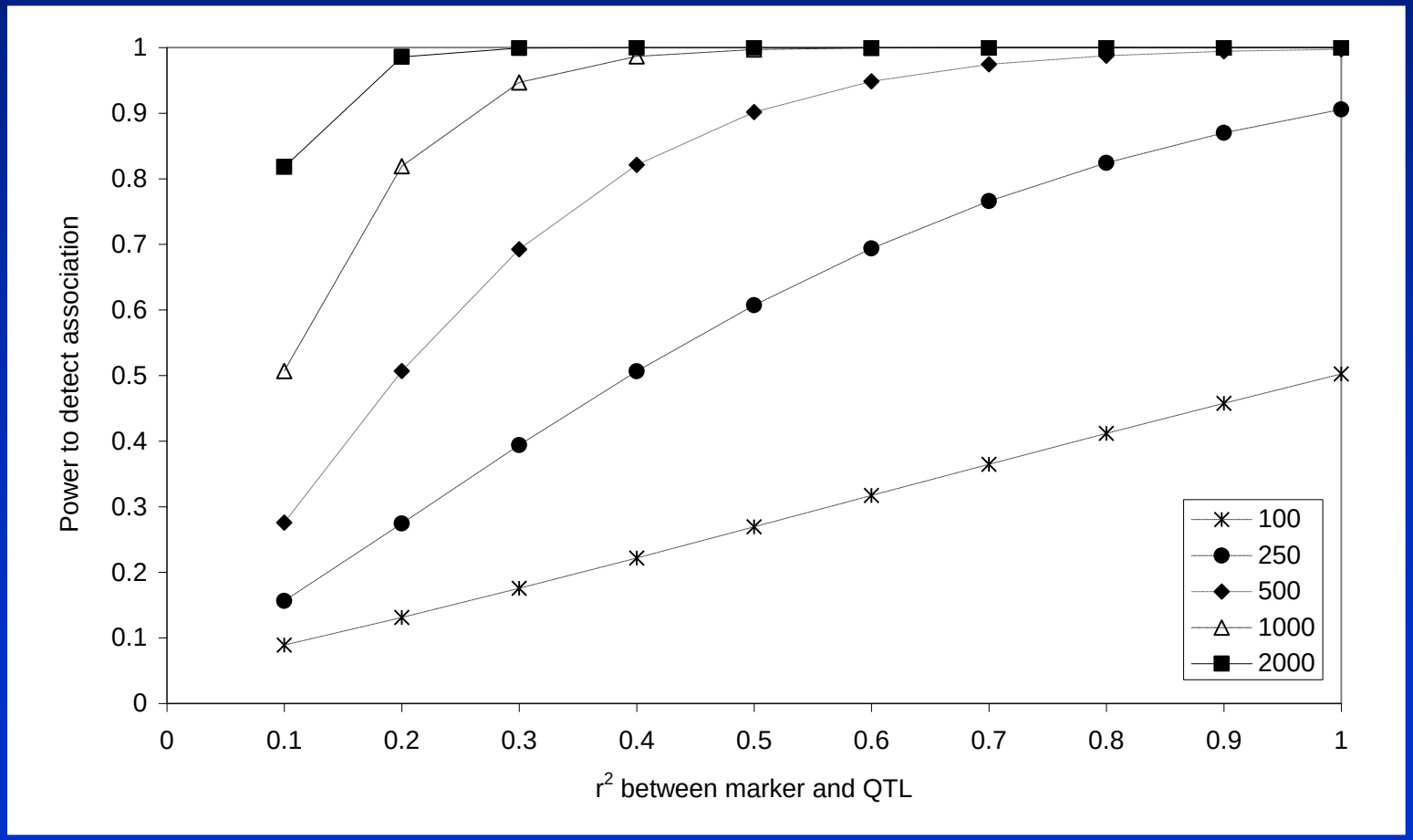
# Power of GWAS

- 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

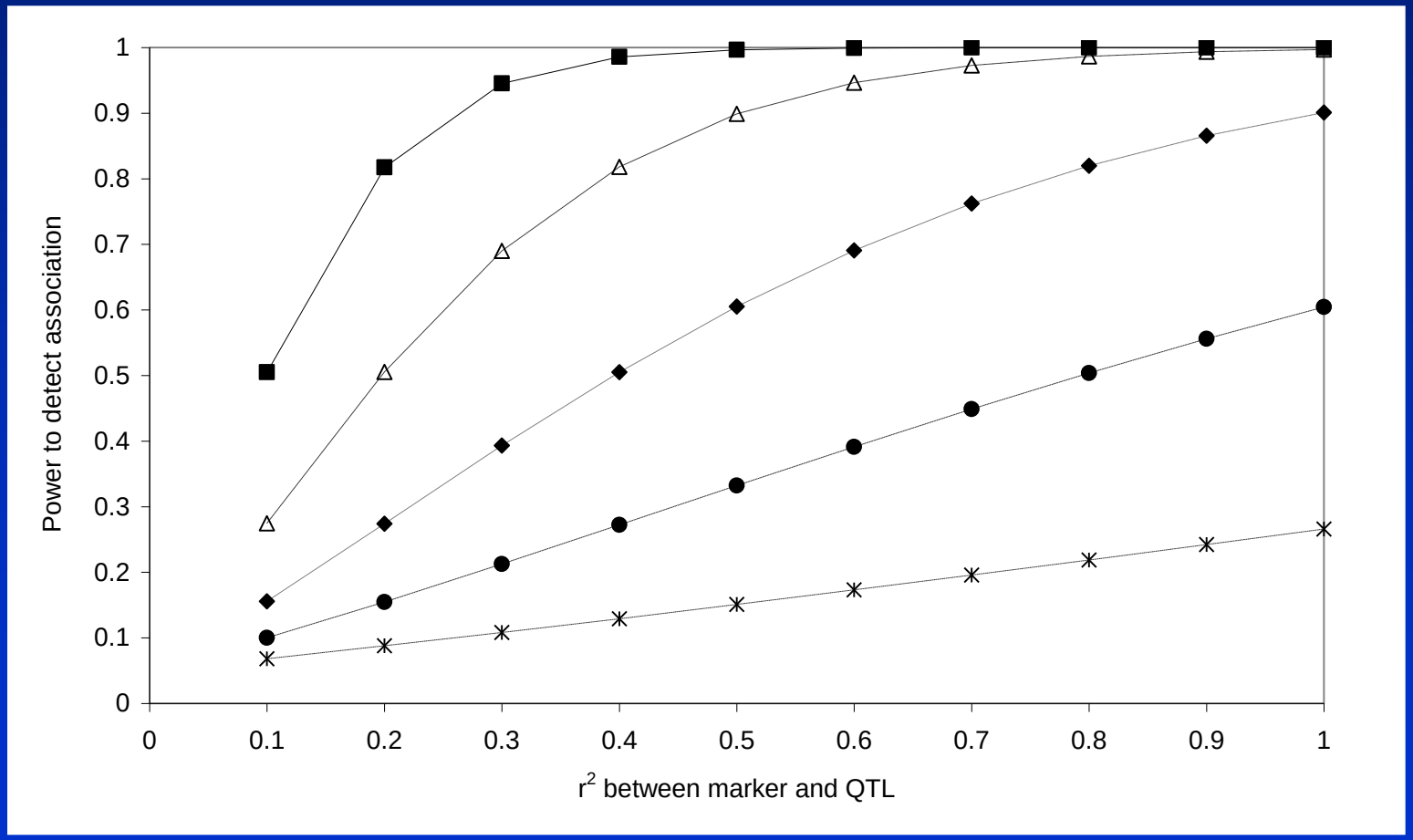
***180 loci explain 10% of the variance***

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

- 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
- **In practise, $P<5x10^{-8}$**

# Genome wide association

- Linkage disequilibrium
- Models for GWAS
- Factors affecting accuracy of GWAS
- Accounting for population structure
- Examples with sequence – can we find causative mutations?
- Using biological information

# 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 |

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

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

# 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 |

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

$\lambda$=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

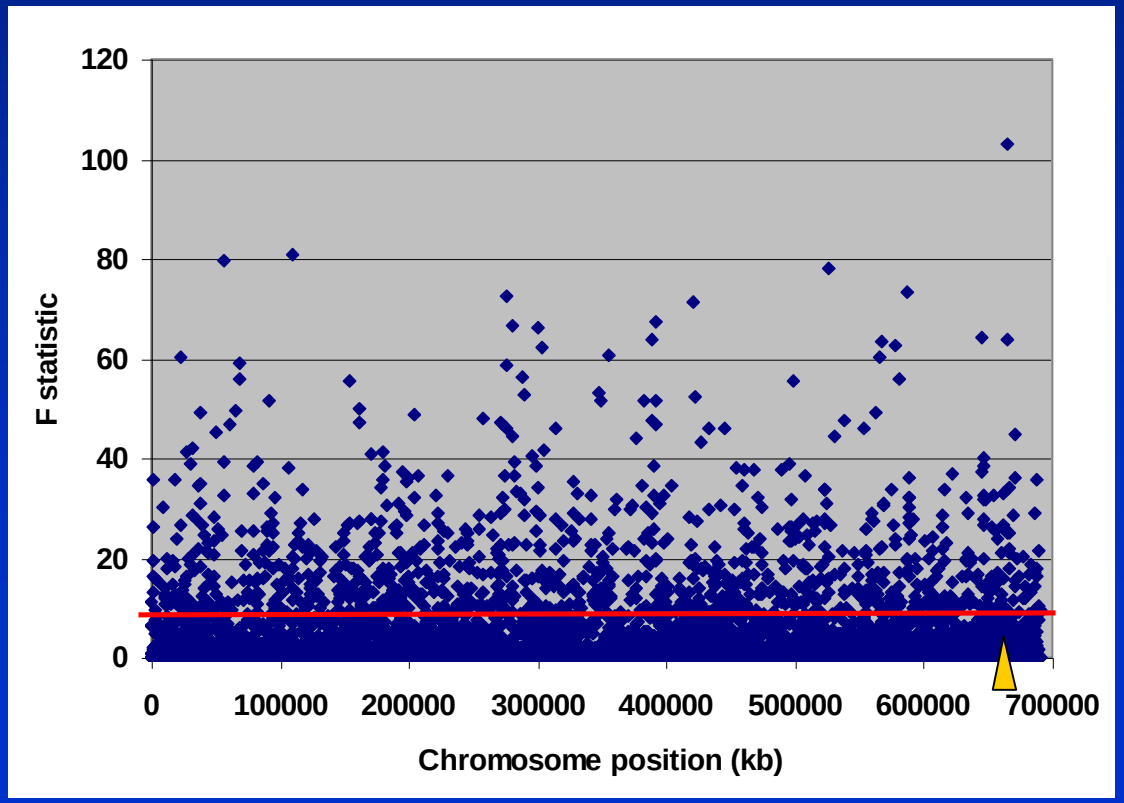
- Example

| 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 |

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

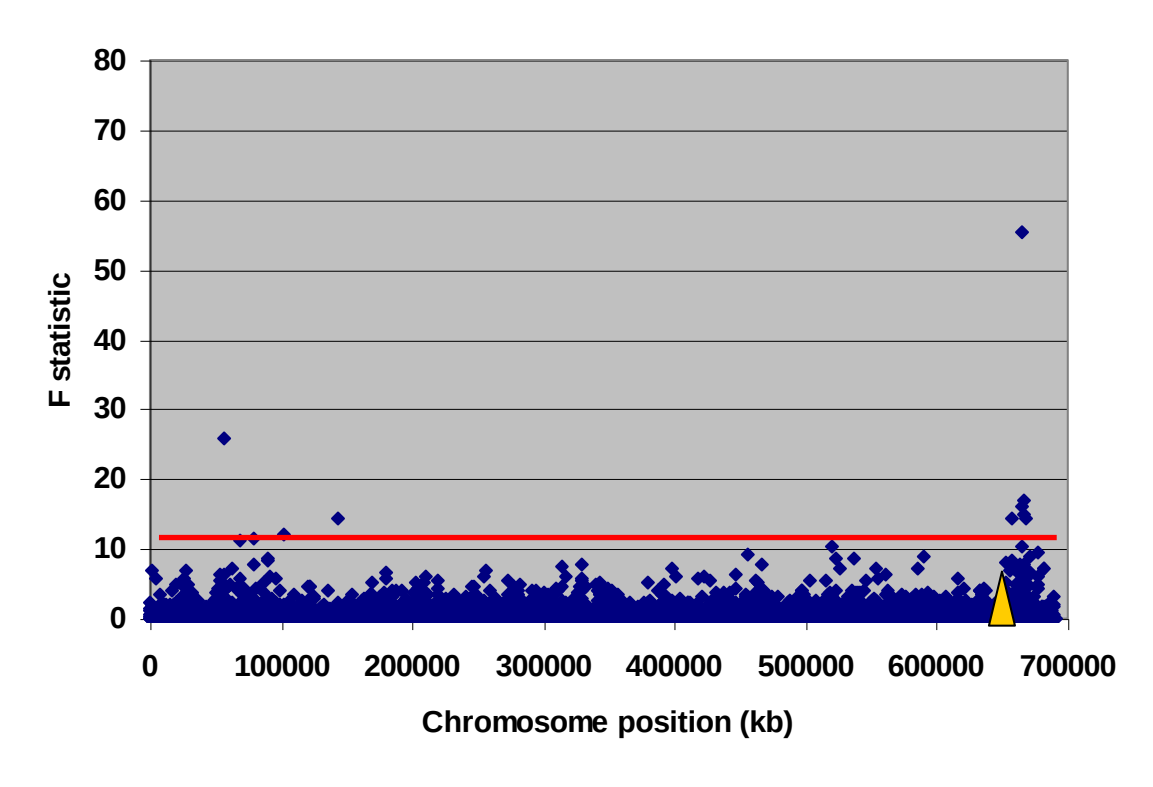
# Population structure

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



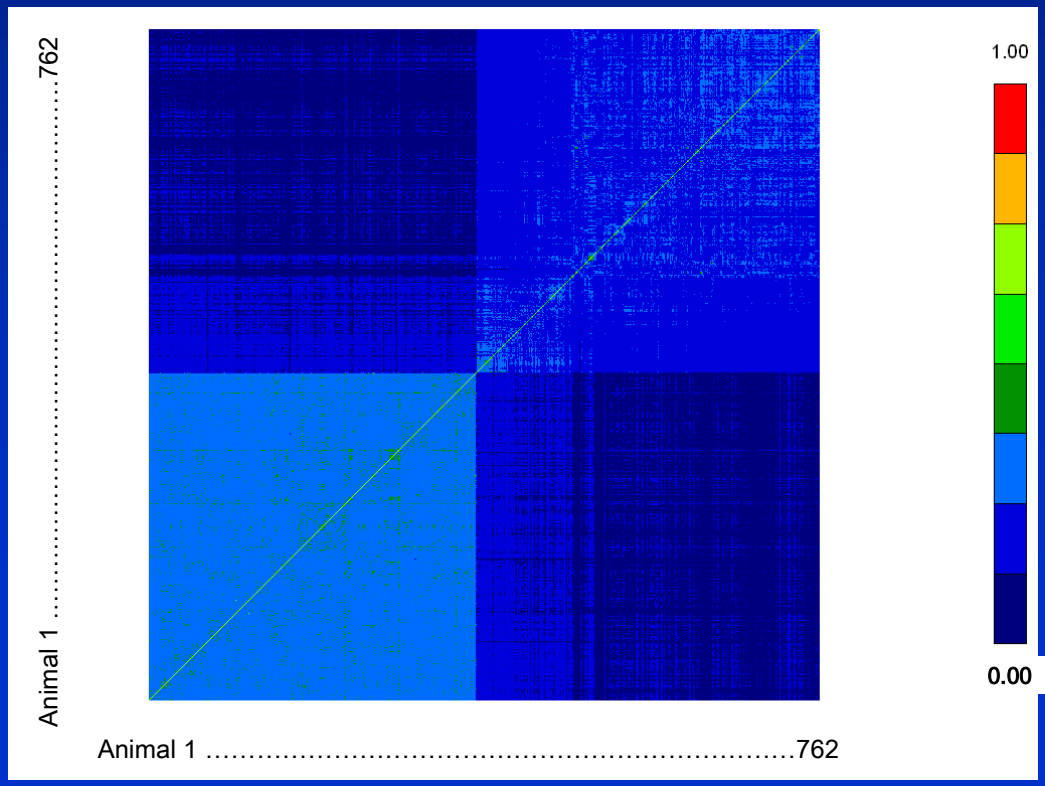
# Population structure

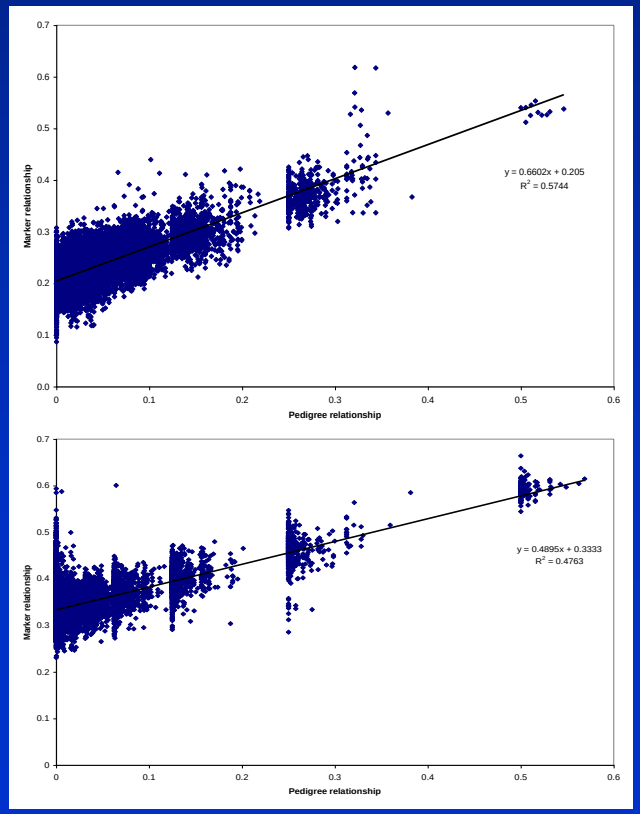
- A simulated data set with a half sib family structure, one QTL simulated



# Population structure

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



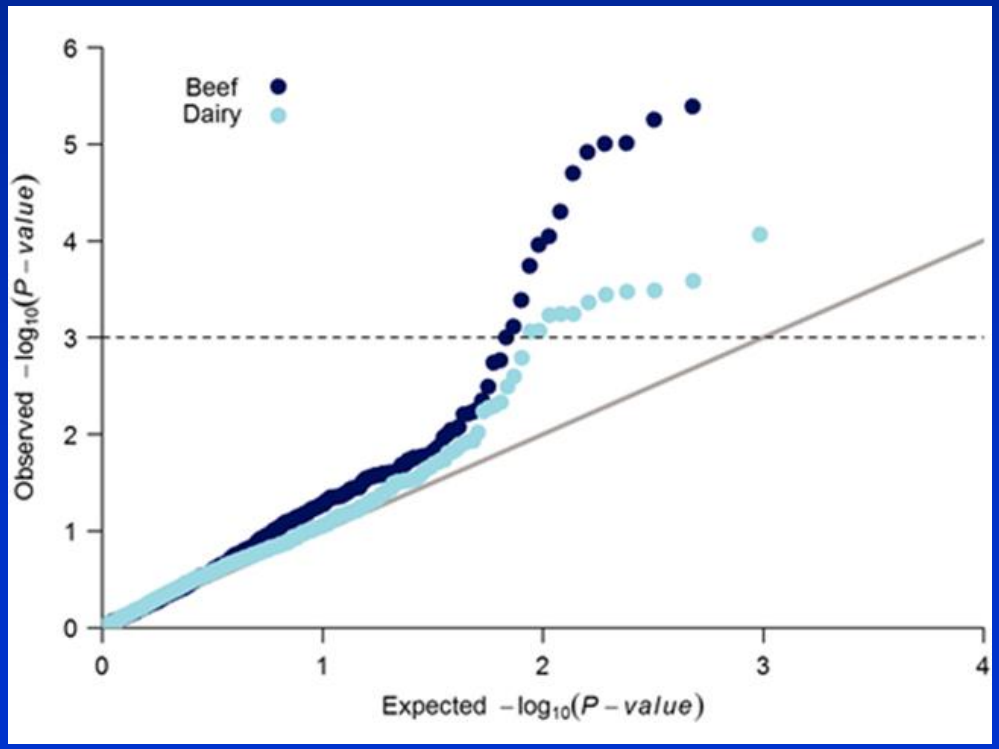
# Genomic relationship matrix

- 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

- Linkage disequilibrium
- Models for GWAS
- Factors affecting accuracy of GWAS
- Accounting for population structure
- Examples with sequence – can we find causative mutations?
- Using biological information

# GWAS with sequence

- Step 1. Impute sequence data into all individuals with phenotypes
  - Target region
  - Whole genome
- Step 2. Run GWAS
  - Single SNP regression?
  - Use genotype probabilities to account for inaccuracy in imputation

# 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.

# GWAS with sequence

ARTICLES

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]

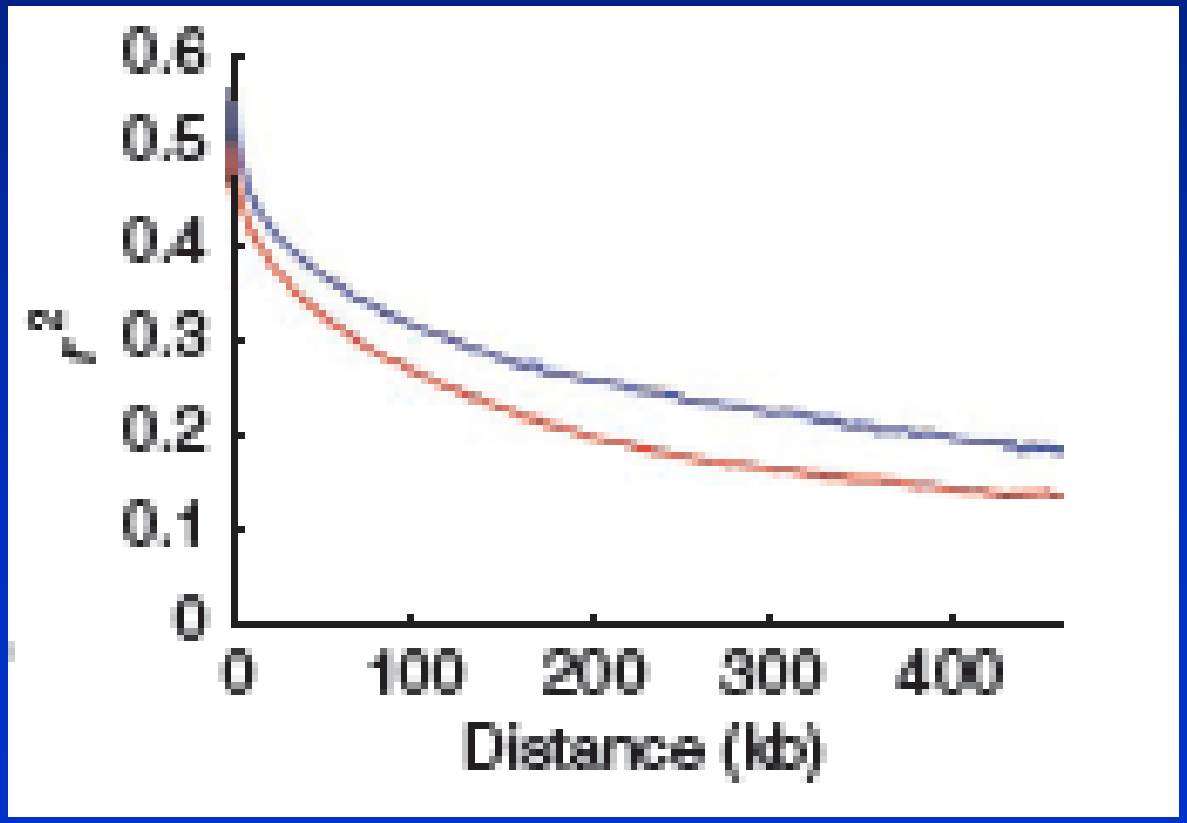
Uncovering the genetic basis of agronomic traits in crop landraces that have adapted to various agro-climatic conditions is important to world food security. Here we have identified ~3.6 million SNPs by sequencing 517 rice landraces and constructed a high-density haplotype map of the rice genome using 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.

# GWAS with sequence

- Huang et al. (2010)
  - Sequenced 517 rice landraces (inbred lines!) at 1x coverage
  - Represent ~ 82% of diversity in worlds rice cultivars
  - Called SNP in sequence pileups
    - 3.6 million SNP
  - With 1x coverage, could only call genotypes at ~ 20% of SNP
  - Therefore use imputation to fill in missing genotype
  - Example

# GWAS with sequence

- Huang et al. (2010)
  - Extent of LD
    - Red indica, blue japonica



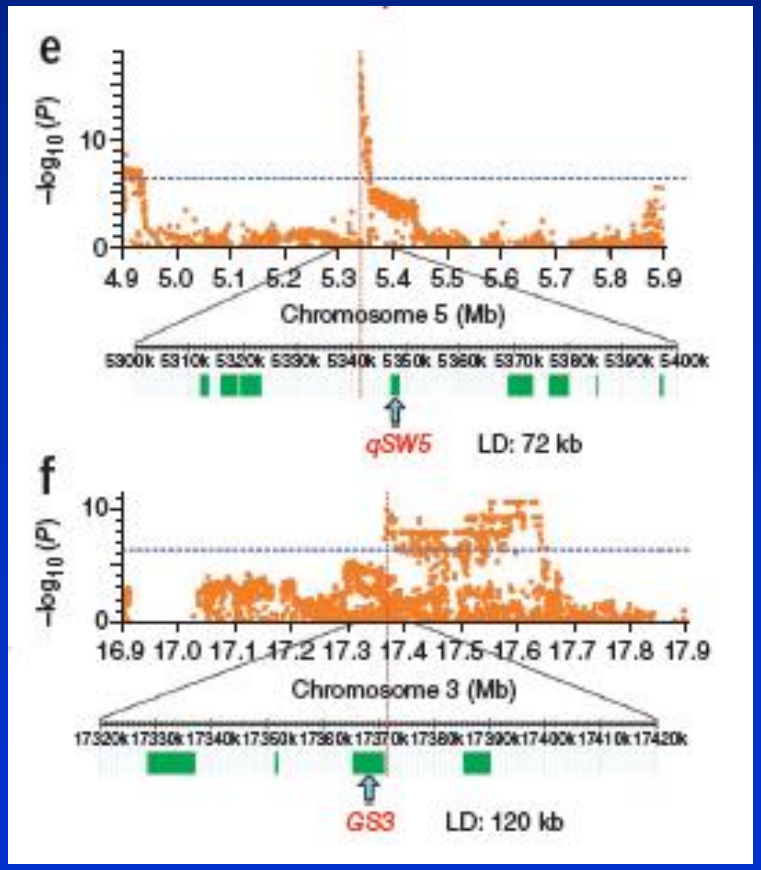
# GWAS with sequence

- Huang et al. (2010)
  - Now have 517 lines with 3.6 million SNP genotyped
  - Well characterised phenotypes for 14 agronomic traits
    - Grain size, flowering date, etc

  – Perform GWAS!

  – Confirmed known mutations

  – Many new mutations



# GWAS with sequence

- Can we detect known mutations with imputed sequence data?
- *DGAT1 -> Chr14, large effect on fat% in milk*
- *GHR -> Chr20, large effect on protein%*

# GWAS with sequence

- Hubert Pausch (Technical University of Munich)
- Impute sequence variants into 2 populations with 650K SNP data
- 2327 Holstein bulls
- 3513 Fleckvieh bulls
- Accuracy of imputation DGAT1 mutation 99.8%

# 1000 bull genomes Run 3.0

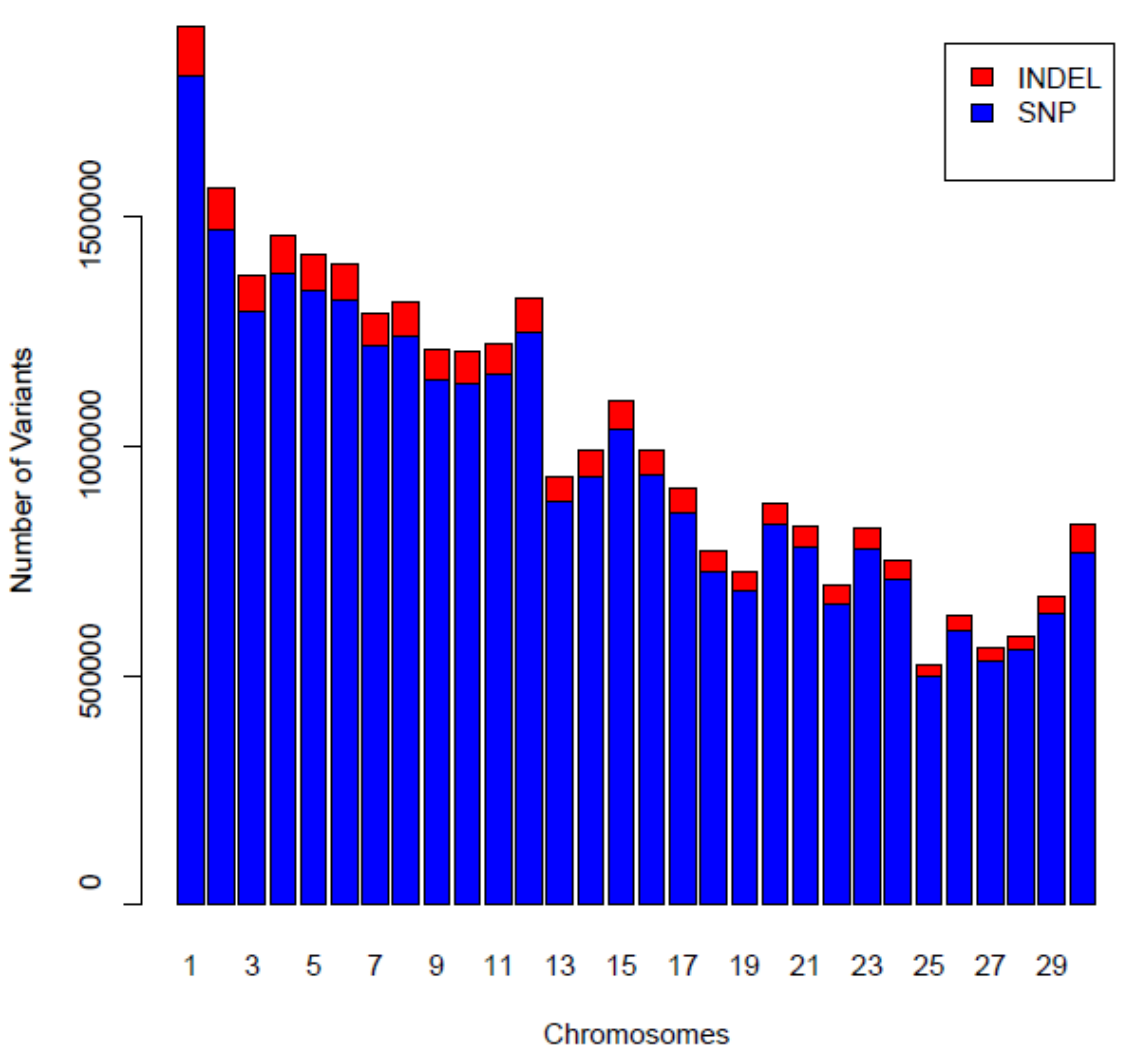
- 14 Partners
- Average 10.1X

| Breed | Number |
|---|---|
| Holstein | 122 |
| Jersey | 26 |
| Simmental | 87 |
| Angus | 54 |
| Swedish Reds | 16 |
| Piedmontese | 2 |
| Limousin | 25 |
| Hereford | 1 |
| Guelph Composite | 9 |
| Finnish Ayrshire | 17 |
| Charolais | 8 |
| Brown Swiss | 43 |
| Belgian Blue | 10 |
| Beef Booster | 8 |
| **All** | **429** |

# 1000 bull genomes project

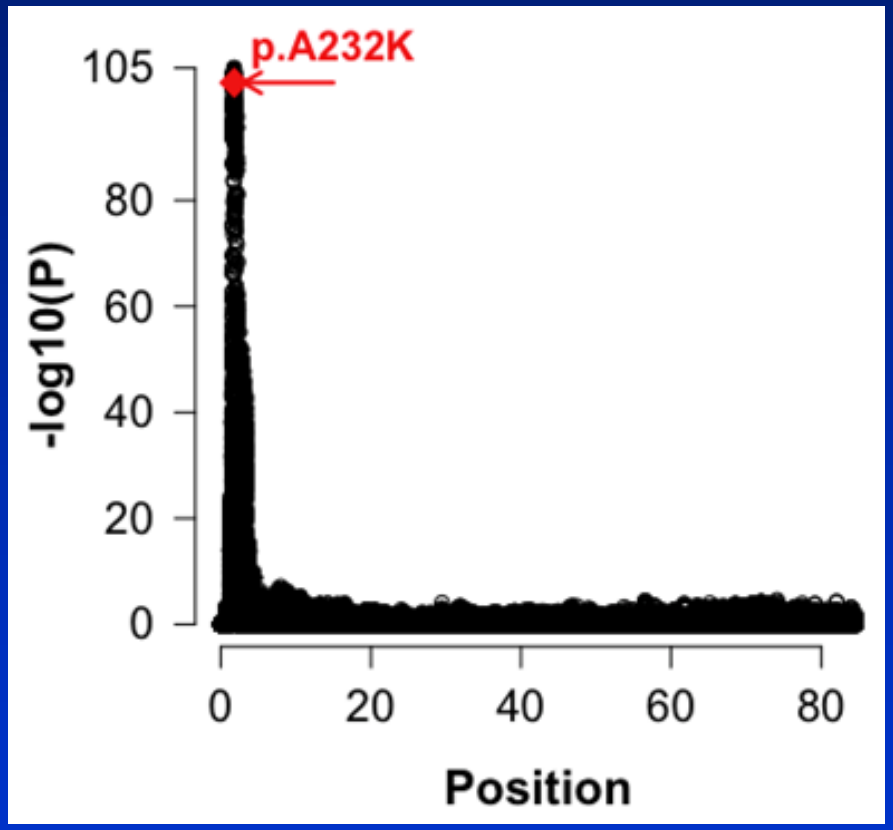
- 30.8 million filtered variants
- 29.1 million SNP
- 1.7 million INDEL

- All variants annotated

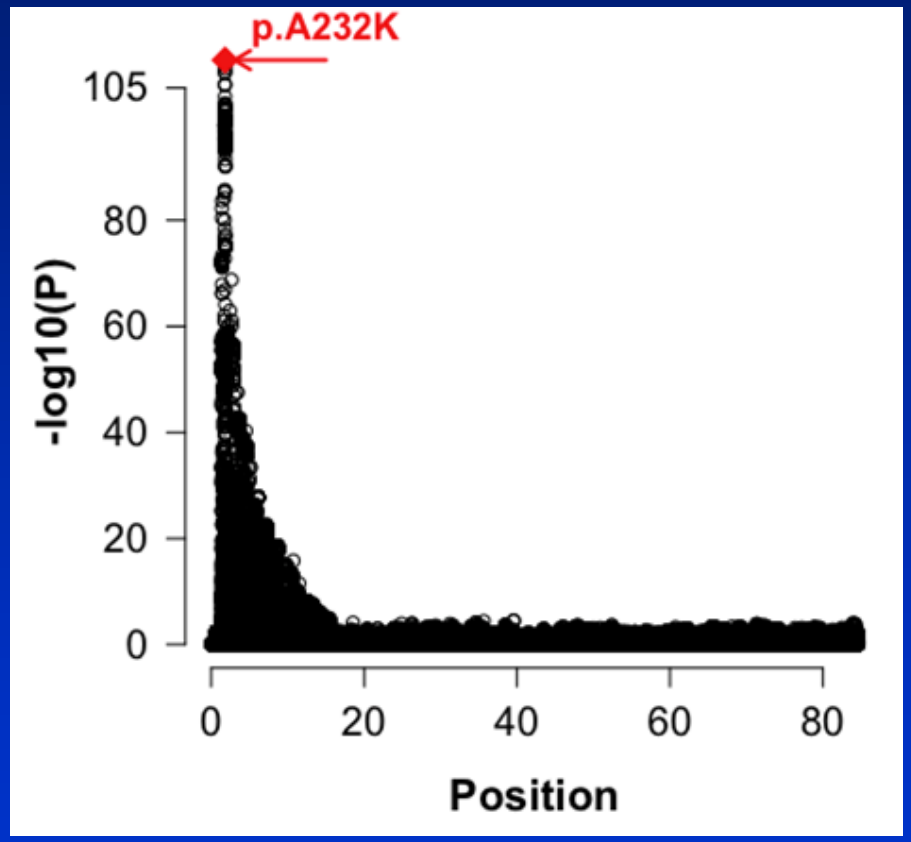


# GWAS with sequence

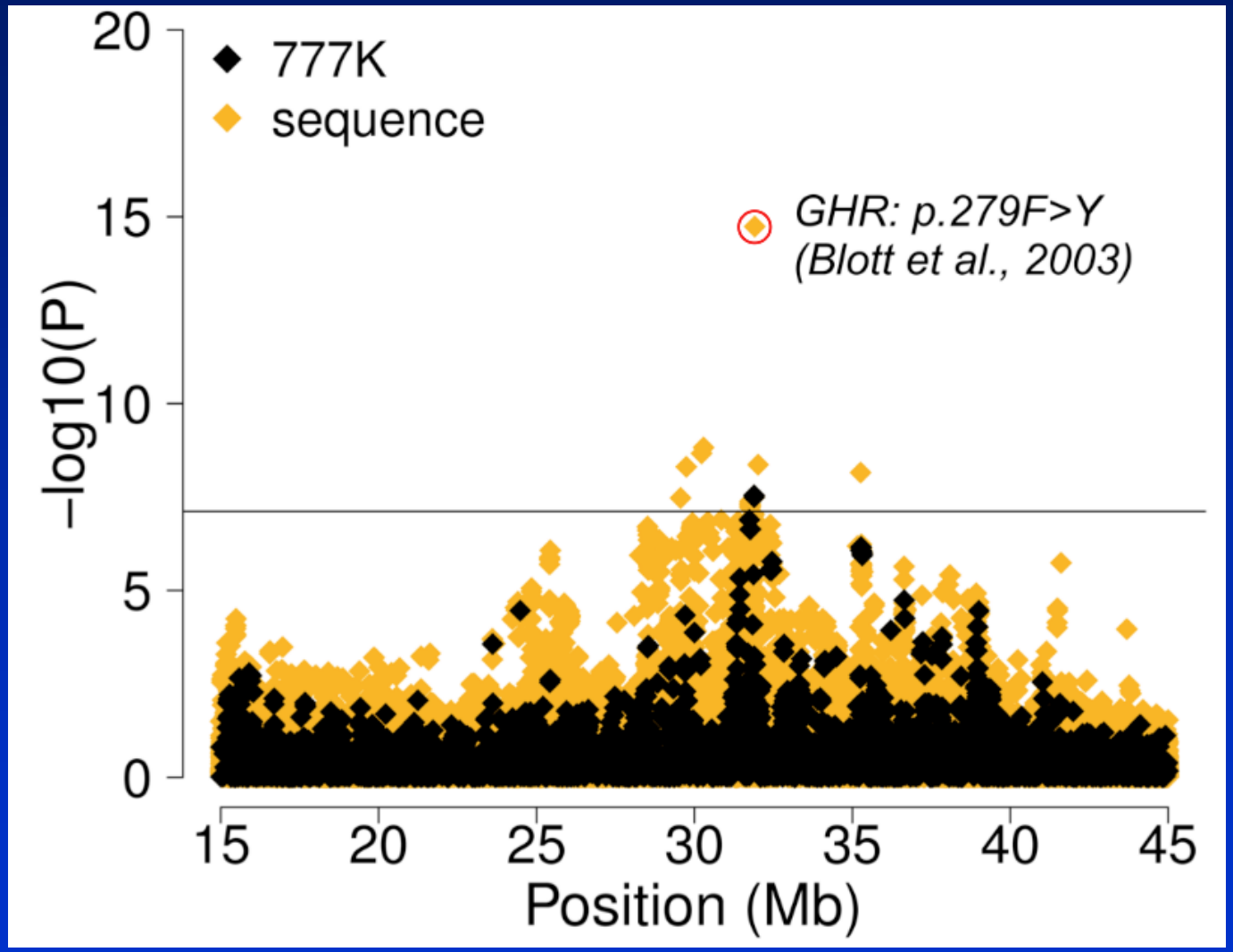
Holstein




Fleckvieh



# GWAS with sequence

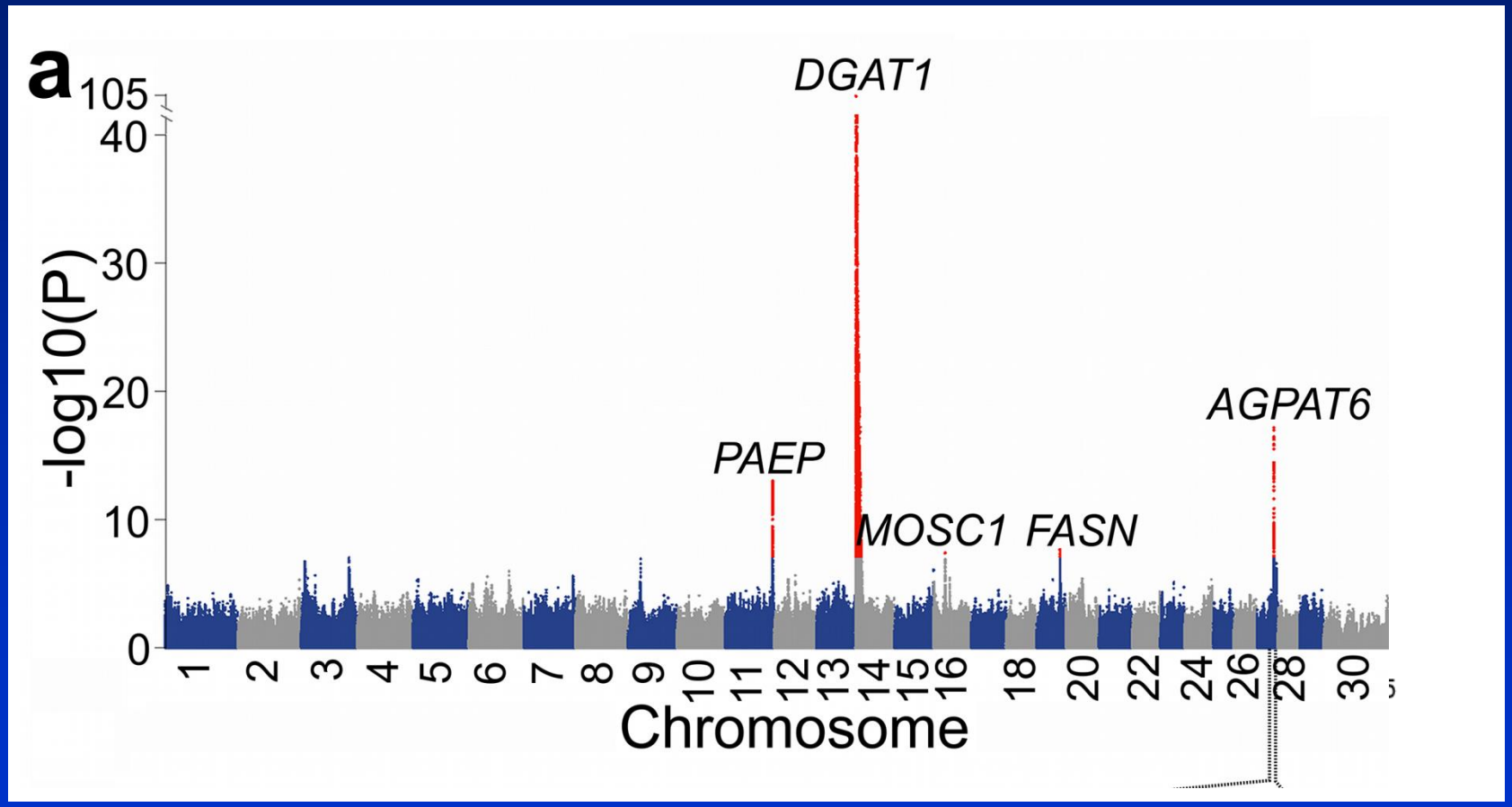


# GWAS with sequence

- Causative mutations detected
- Imputed sequence variants often more significant than original 650K
- However even with accurate imputation, causative mutation not always most significant -> sampling error
- Use additional information, multi-traits, multi-breeds, gene expression?

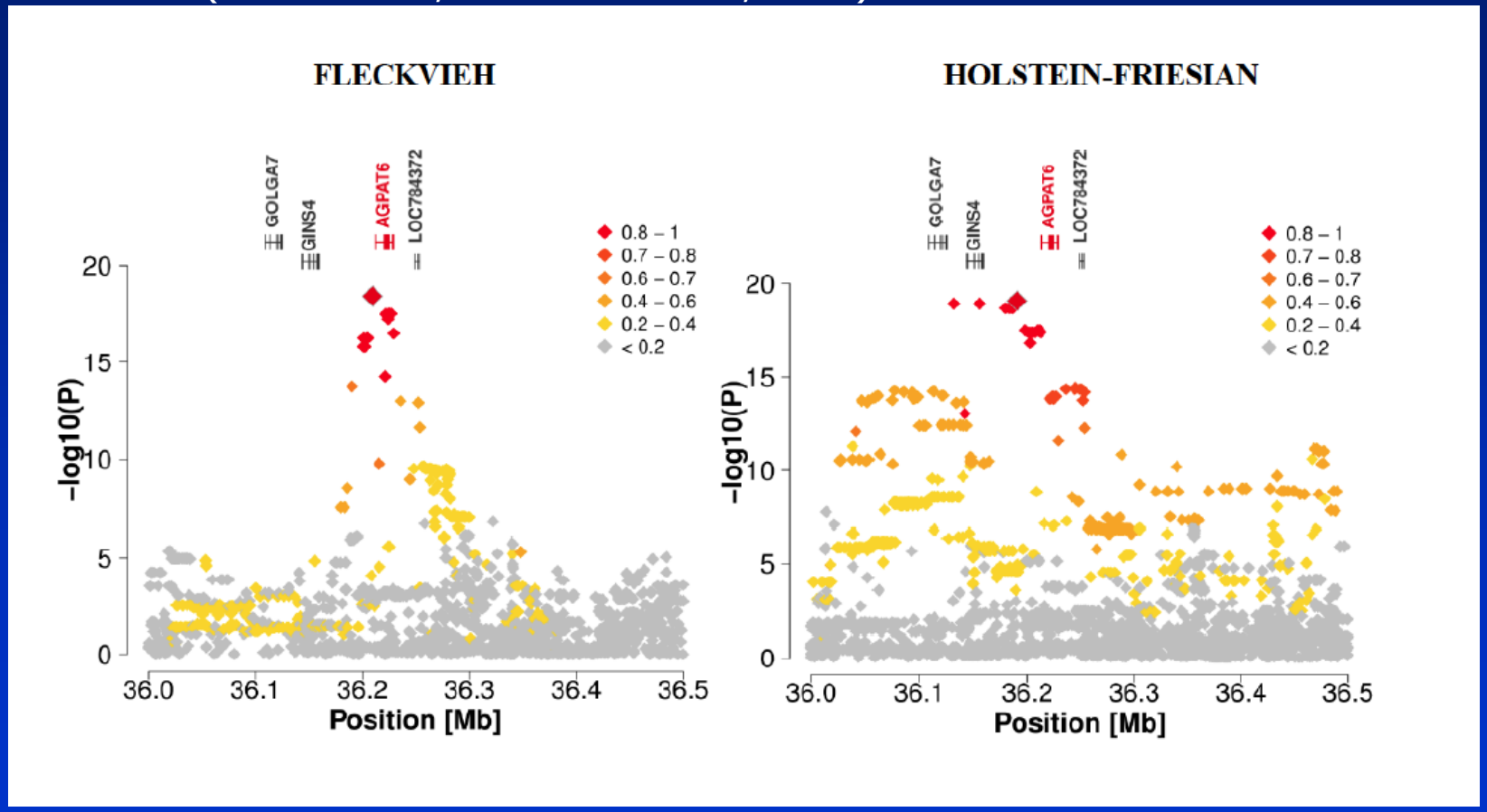
# GWAS with sequence

- Early lactation fat content (Ruedi Fries, Hubert Pausch, TUM)



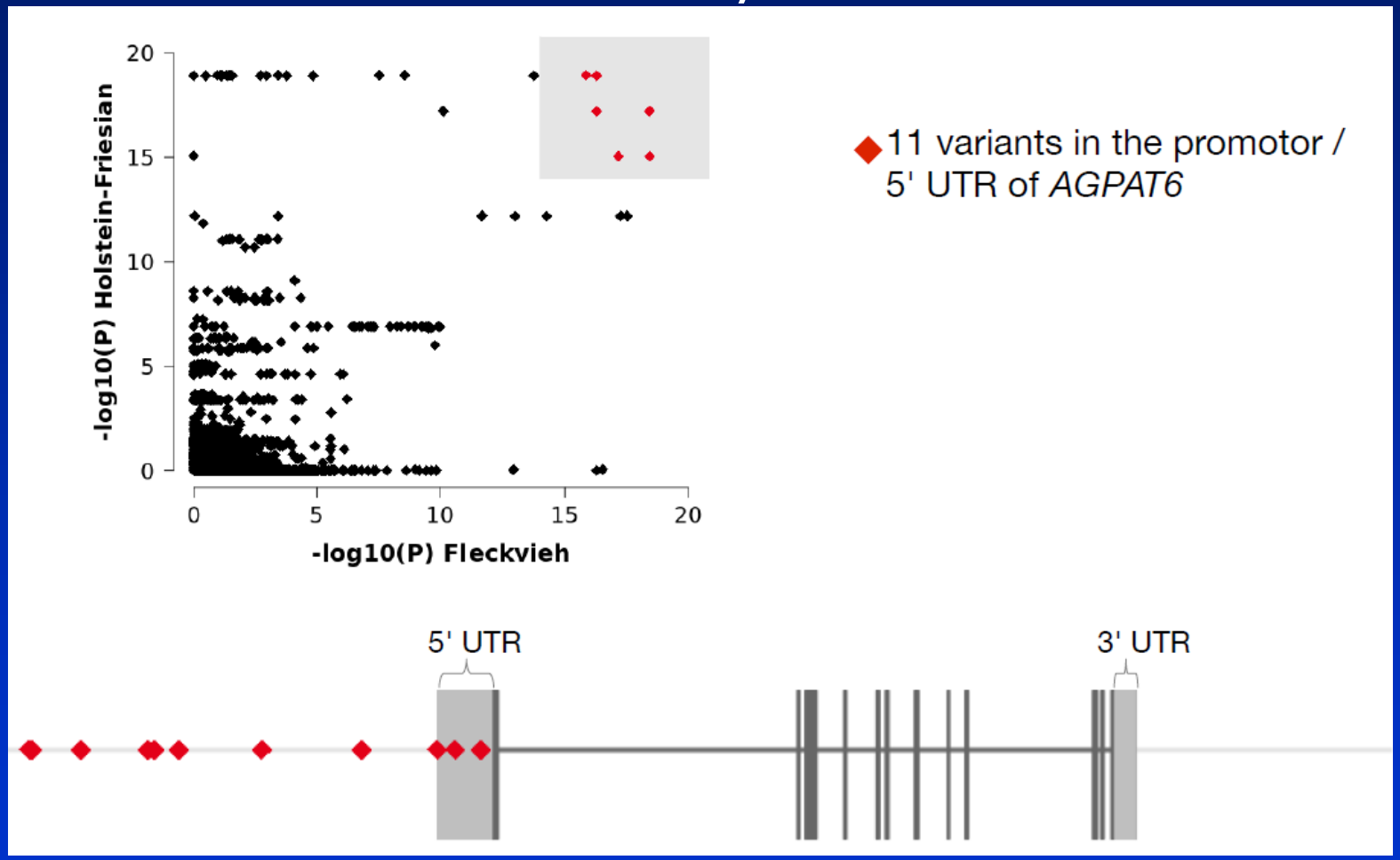
# GWAS with sequence

- Chromosome 27 -> Early lactation fat content
  - (Ruedi Fries, Hubert Pausch, TUM)



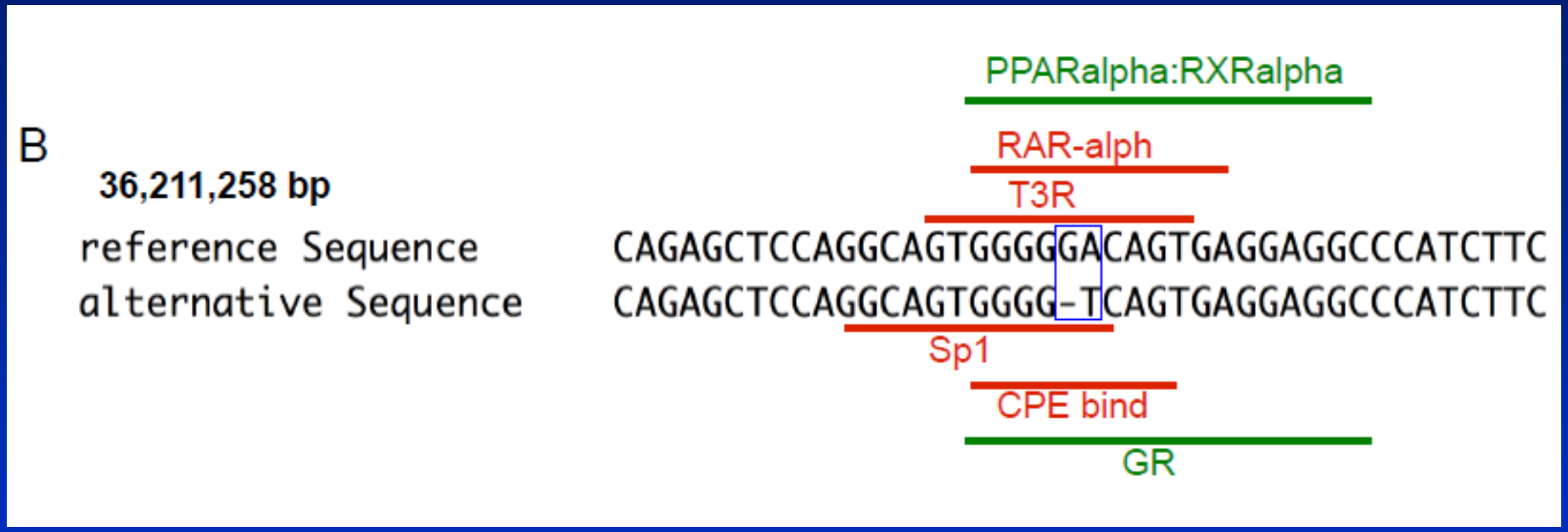
# GWAS with sequence

- Chromosome 27 -> Early lactation fat content



# GWAS with sequence

- Chromosome 27 -> Early lactation fat content

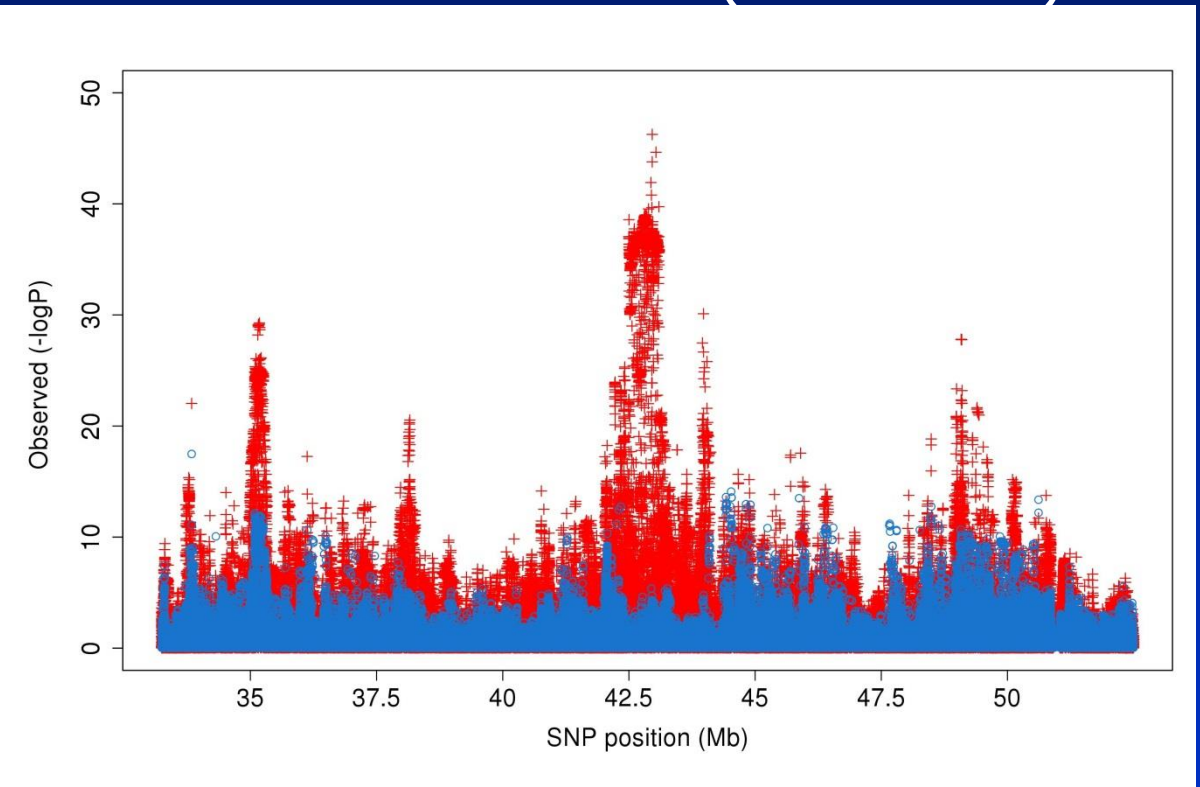


# GWAS + Biological Info

- Gene expression
  - is gene expressed in a tissue associated with phenotype
  - is the mutation associated with a change in level of expression of a gene associated with the phenotype (eQTL, Allele specific expression)
- Proteomics/Metabolomics
  - Is the mutation associated with change in a protein/metabolitite linked to the trait
- Mouse/Arabidopsis knockouts
  - Does knockout of the gene cause a phenotype similar to the one under study

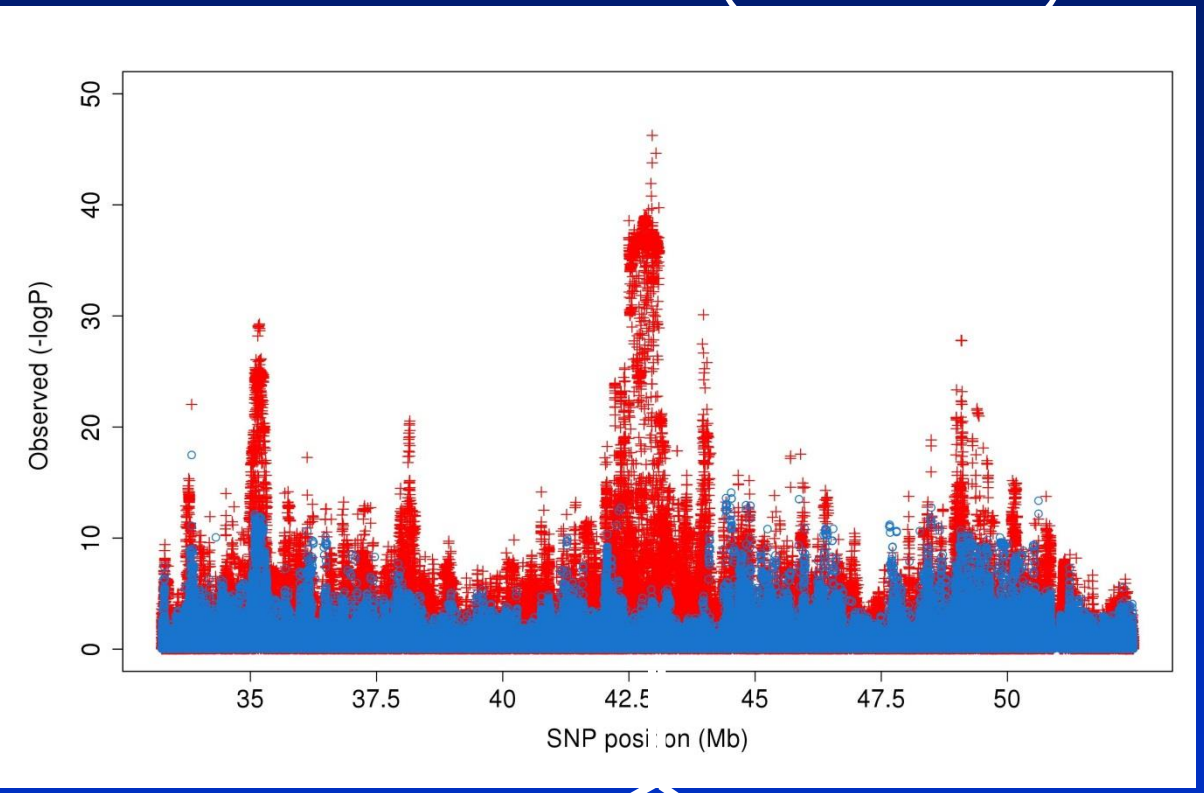
# GWAS + Biological Info

- Chromosome 19 (Protein%)



# GWAS + Biological Info

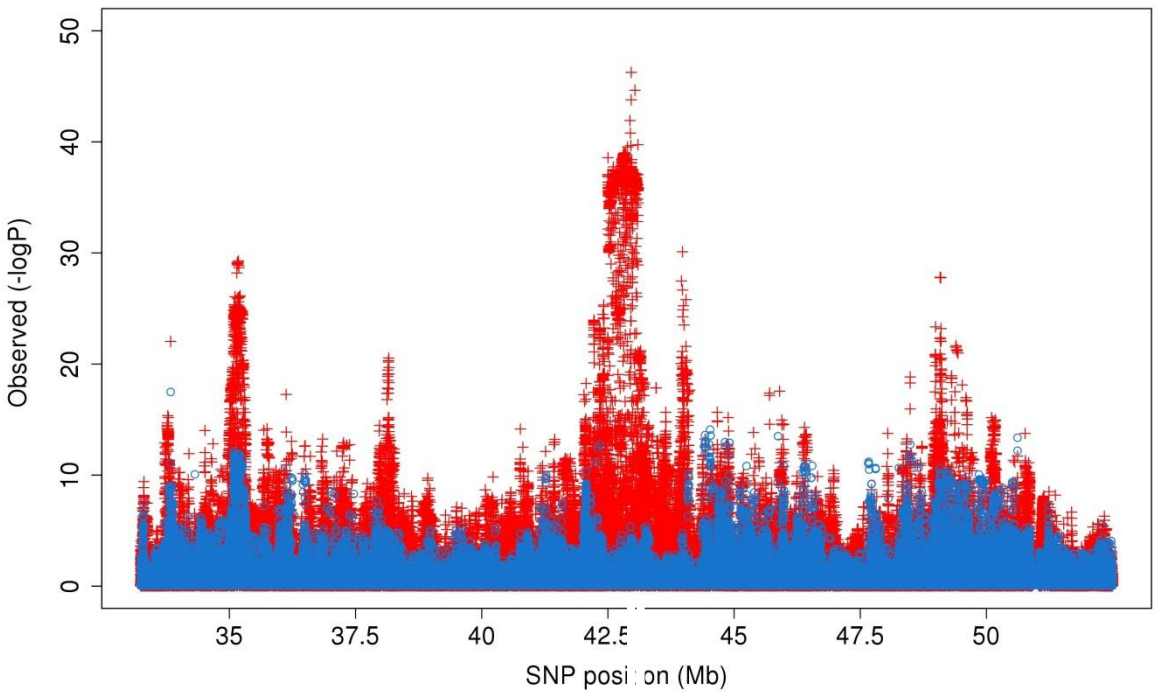
- Chromosome 19 (Protein%)



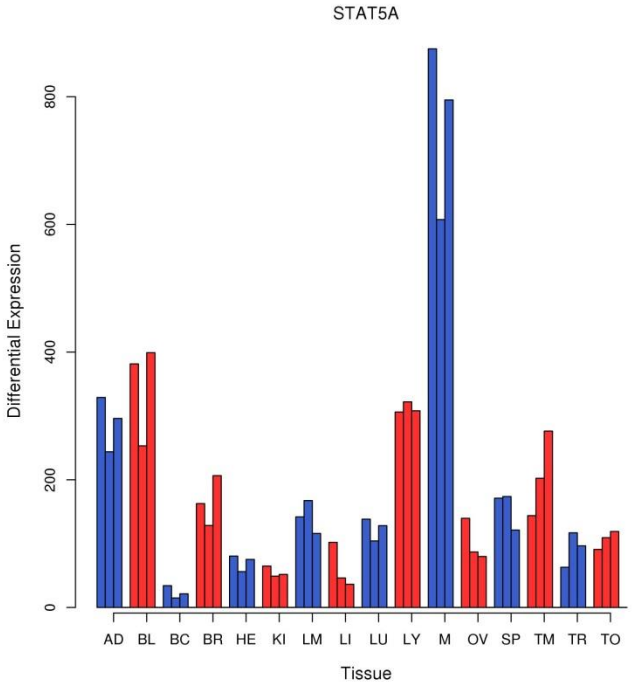

*STAT5A STAT5B*

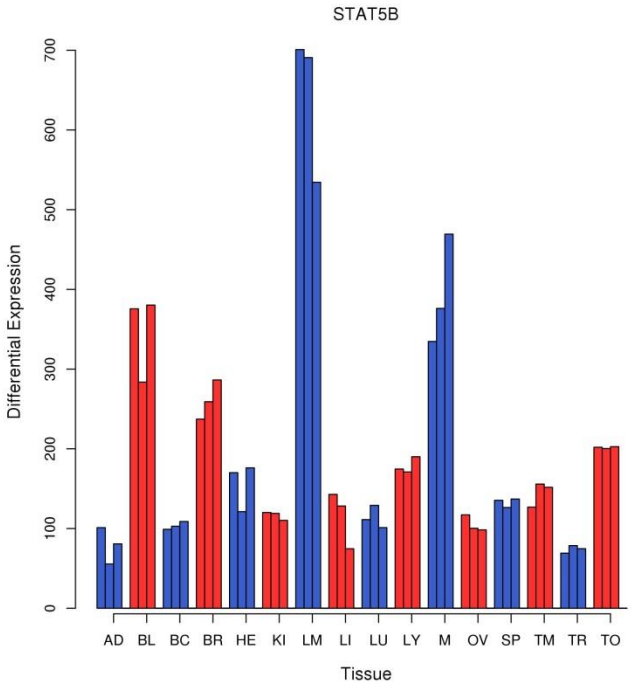
# GWAS + Biologica

- Chromosome 19 (Protein%)



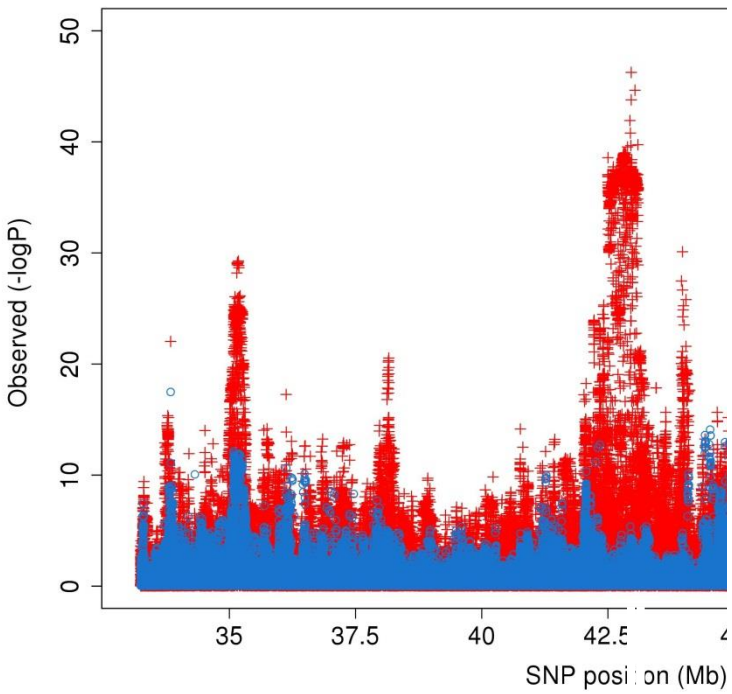

*STAT5A STAT5B*



# GWAS + Biological Info

- Chromosome 19 (Protein%)




## Stat5a is mandatory for adult mammary gland development and lactogenesis


Xiuwen Liu,[1] Gertraud W. Robinson,[1] Kay-Uwe Wagner, Lisa Garrett,[2] Anthony Wynshaw-Boris,[2] and Lothar Hennighausen[1,3]

[1]Laboratory of Biochemistry and Metabolism, National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), National Institutes of Health (NIH), Bethesda, Maryland 20892-1812 USA; [2]Laboratory of Genetic Disease Research, National Center for Human Genome Research, Bethesda, Maryland 20892 USA



Prolactin (PRL) induces mammary gland development (defined as mammopoiesis) and lactogenesis. Binding of PRL to its receptor leads to the phosphorylation and activation of STAT (signal transducers and activators of transcription) proteins, which in turn promote the expression of specific genes. The activity pattern of two STAT proteins, Stat5a and Stat5b, in mammary tissue during pregnancy suggests an active role for these transcription factors in epithelial cell differentiation and milk protein gene expression. To investigate the function of Stat5a in mammopoiesis and lactogenesis we disrupted this gene in mice by gene targeting. Stat5a-deficient mice developed normally and were indistinguishable from hemizygous and wild-type littermates in size, weight, and fertility. However, mammary lobuloalveolar outgrowth during pregnancy was curtailed, and females failed to lactate after parturition because of a failure of terminal differentiation. Although Stat5b has a 96% similarity with Stat5a and a superimposable expression pattern during mammary gland development it failed to counterbalance for the absence of Stat5a. These results document that Stat5a is the principal and an obligate mediator of mammopoietic and lactogenic signaling.


*STAT5A STAT5B*

# GWAS with sequence

- Causative mutations detected
- Imputed sequence variants often more significant than original 650K
- However even with accurate imputation, causative mutation not always most significant -> sampling error
- Use additional information, multi-traits, multi-breeds, biological information?

# GWAS Software

| Software | A matrix | G matrix | Weights | Genotype probabilities | Reference |
|---|---|---|---|---|---|
| **SNPSnappy** | Yes | No | Yes | No | Meyer K, Tier B. Genetics 2012;190:275-277. |
| **GCTA** | No | Yes | No | No | Yang J Am J Hum Genet. 2011 7;88:76-82. |
| **Emmax** | No | Yes | No | Yes | Kang HM Nat Genet. 2010;42:348–354 |

# 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

# GWAS take home points

- Large data sets needed, QTL explain 1% of variance for many traits
- Multi-breed to break down LD
- Any population structure results in spurious associations
- With SNP arrays
  - Power depends on extent of LD/marker density and number of phenotypic records
  - Knowledge of extent of LD critical
- With sequence
  - Some cases direct to causal mutation
  - Sampling error, inaccurate imputation
- Validation, validation, validation

# Results of genome scans with dense SNP panels