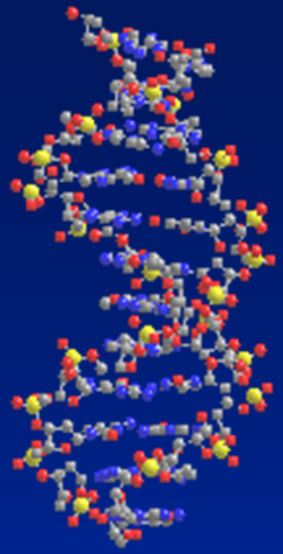
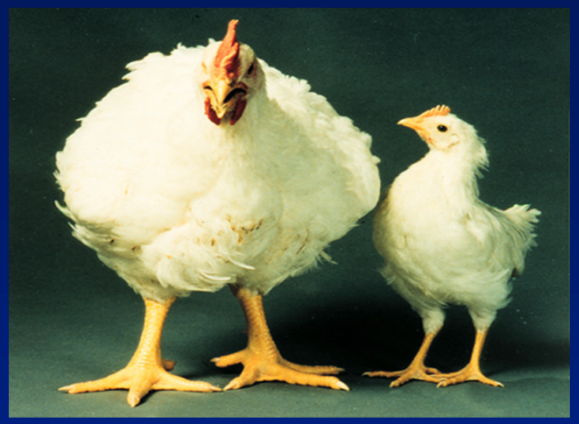
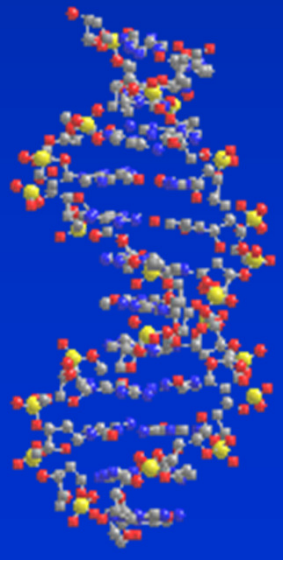




# Genomic Selection in the era of Genome sequencing

# Genomic selection in action:

Genex Cooperative, Inc.
A subsidiary of Cooperative Resources International

Search the Genex site

BEEF | DAIRY | PRODUCTS & SERVICES | AI TRAINING | LEARNING CENTER | ABOUT | NEWS | CONTACT

HOME > SO MANY NEW HIGH GENETIC MERIT BULLS?

Print Page | E-mail Page

**So Many New High Genetic Merit Bulls?**

## Why Are There So Many New High Genetic Merit Bulls?

Genex provides the answers for this question and others relating to genomics.

### Why are there so many new high genetic merit bulls?

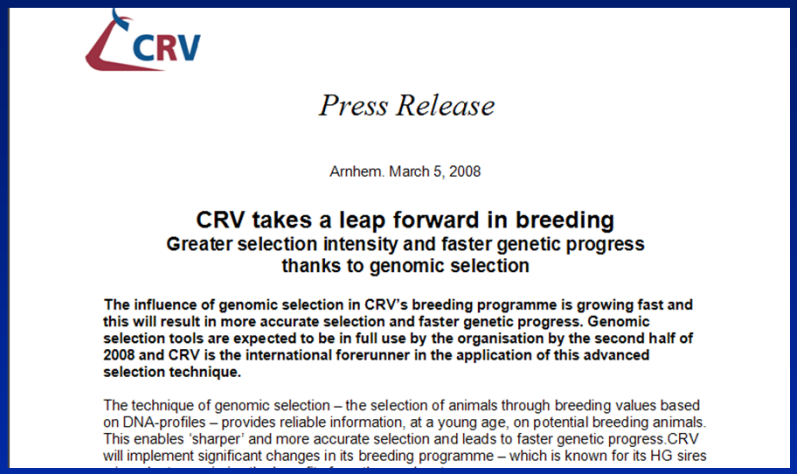
This January, it is as if we fast forwarded through nine USDA sire summaries all in one release without a base change. With genomic selection now available, 34 new Holstein sires graduated that would not have been available for another two or three years if we had continued to use the old sire evaluation system. The advantage we gain from genomics is that we can now predict the genetic merit of our young animals at a much higher reliability earlier than we could with simply the use of traditional parent averages. This gives dairy producers the potential to make significant genetic progress in their herds.

### What is genomic selection?

Genomic selection is predicting the performance of an animal's offspring based on the DNA of that animal. Recent breakthroughs from industry-supported research have resulted in new tests that look at thousands of DNA markers of an individual bull or cow. The DNA markers of the individual animal are compared to the marker profiles of thousands of bulls and cows with millions of offspring with known performance for the various traits of interest. This allows an estimation of breeding values of an animal based on associations between DNA marker profiles and milk yield, Somatic Cell Score, Productive Life and other important traits.

### What is a genomic genetic evaluation?

A genomic genetic evaluation includes the information from the DNA marker testing of that individual. The evaluation also considers parent or pedigree information and may also include progeny information or, for females, performance information. The



CRV

*Press Release*

Arnhem. March 5, 2008

**CRV takes a leap forward in breeding**
**Greater selection intensity and faster genetic progress thanks to genomic selection**

**The influence of genomic selection in CRV's breeding programme is growing fast and this will result in more accurate selection and faster genetic progress. Genomic selection tools are expected to be in full use by the organisation by the second half of 2008 and CRV is the international forerunner in the application of this advanced selection technique.**

The technique of genomic selection – the selection of animals through breeding values based on DNA-profiles – provides reliable information, at a young age, on potential breeding animals. This enables 'sharper' and more accurate selection and leads to faster genetic progress.CRV will implement significant changes in its breeding programme – which is known for its HG sires


GENE TEAM: GENETICS AUSTRALIA'S FIRST GENETIC MARKER BULLS INFUSE, DEFIER AND WATCHDOG.

# Genomic selection in action:

| | Australia | Ireland | NZ (LIC) | France | Germany | Netherlands | DK/SWE/ FIN | USA/Can |
|---|---|---|---|---|---|---|---|---|
| Year in which genomic evaluation commenced nationally | 2011 | 2009 | 2008 | 2009 | 2010 | 2010 | 2008 | 2008 |
| Size of reference population (males; production traits) | 2,247 | 4,500 | 3,600 | 19,377 | 19,377 | 19,377 | 19,377 | 12,152 |
| Reliability (total merit index)* | 43% | 54% | 55-60% | 65% | 65% | 60% | 55-60% | 62% |
| Reliability (protein yield)* | 50% | 61% | 55-60% | 65% | 72% | 66% | 63% | 71% |
| Females included in reference population | Soon (10k) | Not yet | 16,000 | Not yet | 0 | 0 | 0 | 11,473 |
| Number of young bulls genotyped per year | 300 | 1,000 | 1,500 | 12-15,000 | 6,000 | 2,100 | 1,800 | 13,070 |
| Number of bulls progeny-tested | 100 | 70 | 160 | 0 | <500 | 140 | 175 | 2,000 |
| Age at which young bulls are widely used (months) | 16 | 24 | 14 | 16 | 15 | 20 | 20 | 12 |
| Price relative to proven bulls | same | less | more | less | same | same | same | Same |
| Number of young genomically tested bulls in the top 20 bulls ranked on country's index | 11 | 10 | 20 | 20 | 17 | 11 | 12 | 20 |
| Market-share of genomically tested bulls (bulls without milking daughters) | N/A | 50% | 30-35% | 30% | <30% | 25% | 45% | 43% |

# Course overview

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

# Linkage disequilibrium

- A brief history of QTL mapping
- Measuring linkage disequilibrium
- Causes of LD
- Extent of LD in animals and plants
- The extent of LD between breeds and lines
- Strategies for haplotyping

# A brief history of QTL mapping

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

# Two models…….

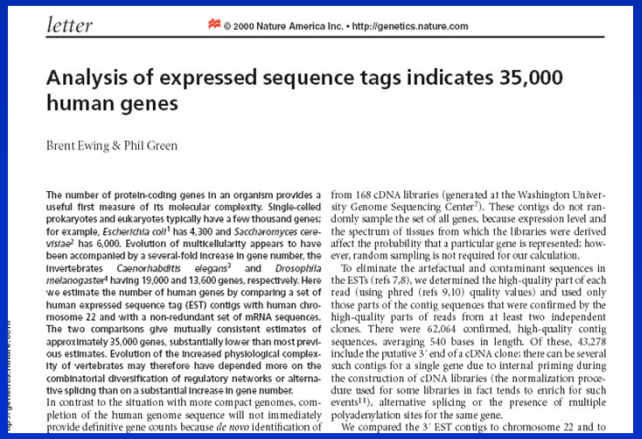
- Infinitesimal model:
  - assumes that traits are determined by an infinite number of unlinked and additive loci, each with an infinitesimally small effect
  - This model the foundation of animal breeding theory including breeding value prediction
  - Spectacularly successful in many cases!

*Time to market weight for meat chickens has decreased from 16 to 5 weeks in 30 years*

# 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

*Nature*. 2010 October 14; 467(7317): 832–838. doi:10.1038/nature09410.

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

*<10% of phenotypic variance!*

NATURE|Vol 456|6 November 2008

# 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

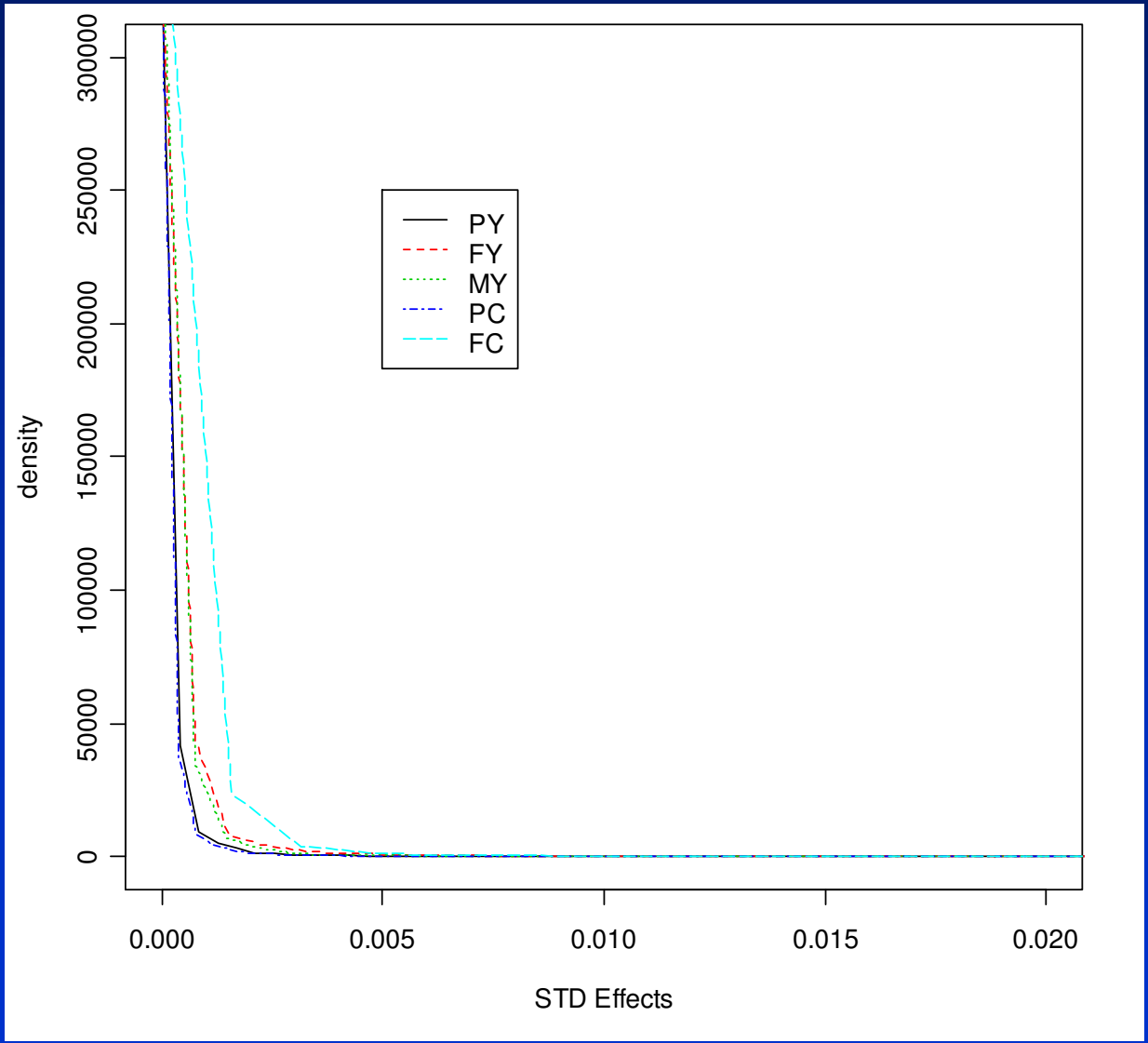

Even though these genome-wide association studies (GWAS) turned up dozens of variants, they did "very little of the prediction that you would do just by asking people how tall their parents are", says Joel Hirschhorn at the Broad Institute in Cambridge, Massachusetts, who led one of the studies[7].

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

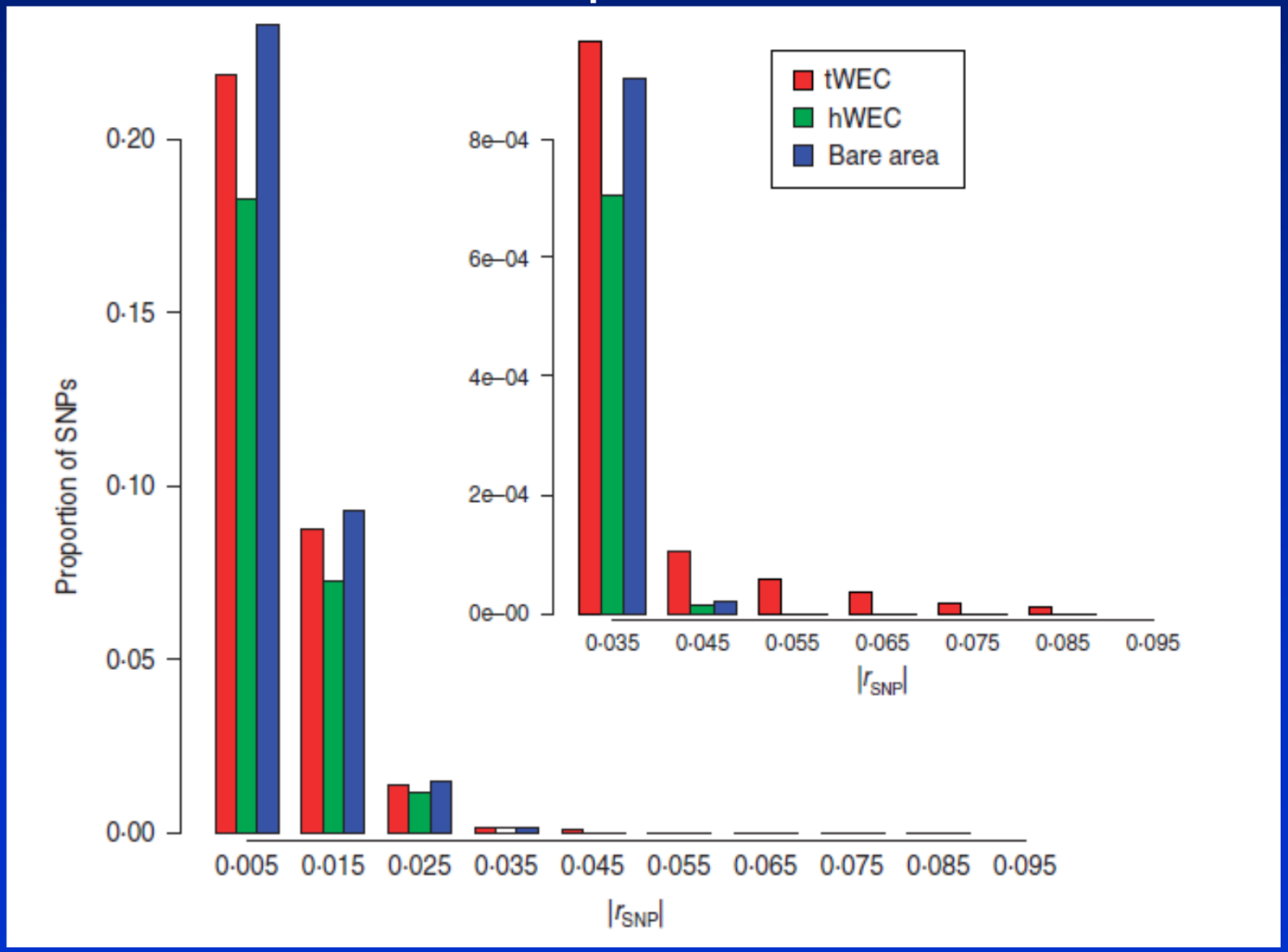
# Distribution of QTL effects

- DGAT1 40% of variation in fat% (FC)



# Distribution of QTL effects

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



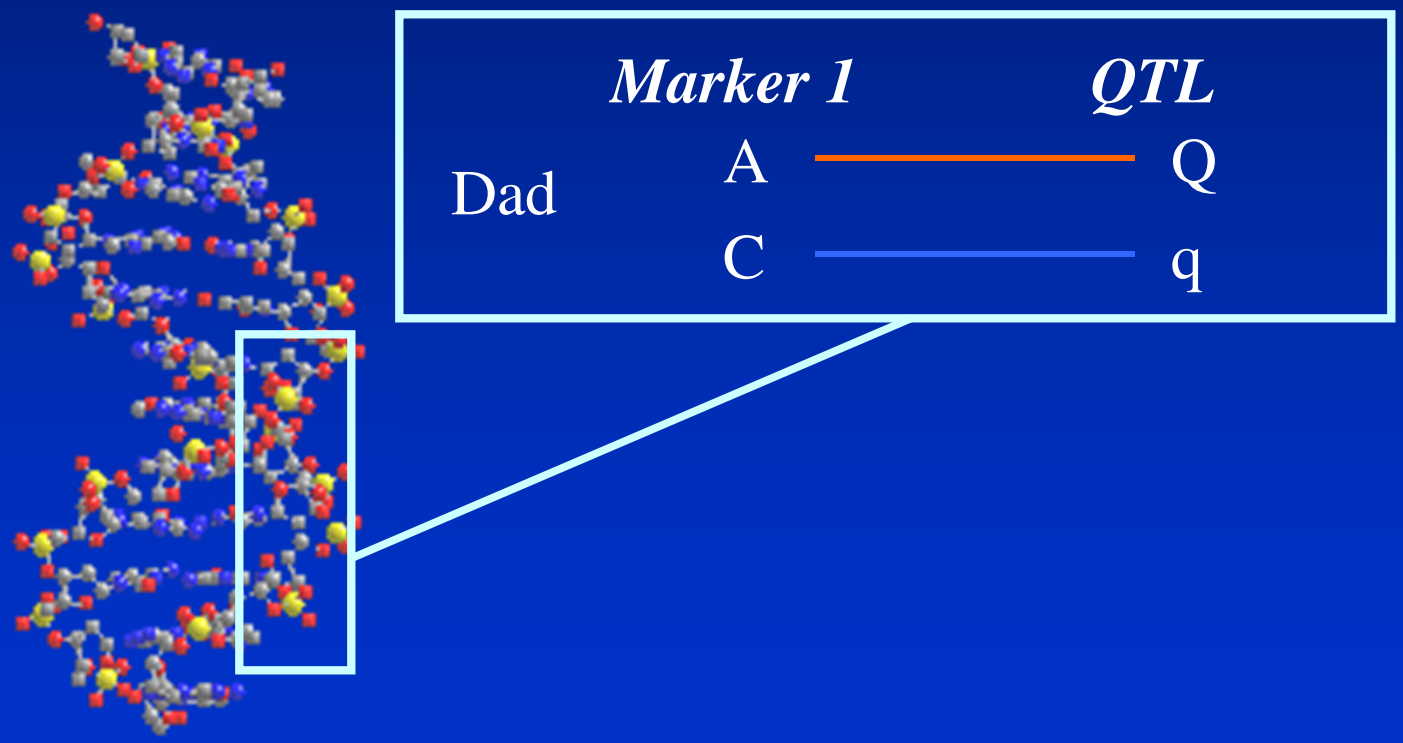
# Quantitative trait loci (QTL) detection

- If we had information on the location in the genome of the QTL we could
  - increase the accuracy of breeding values
  - improve selection response
- How to find them?

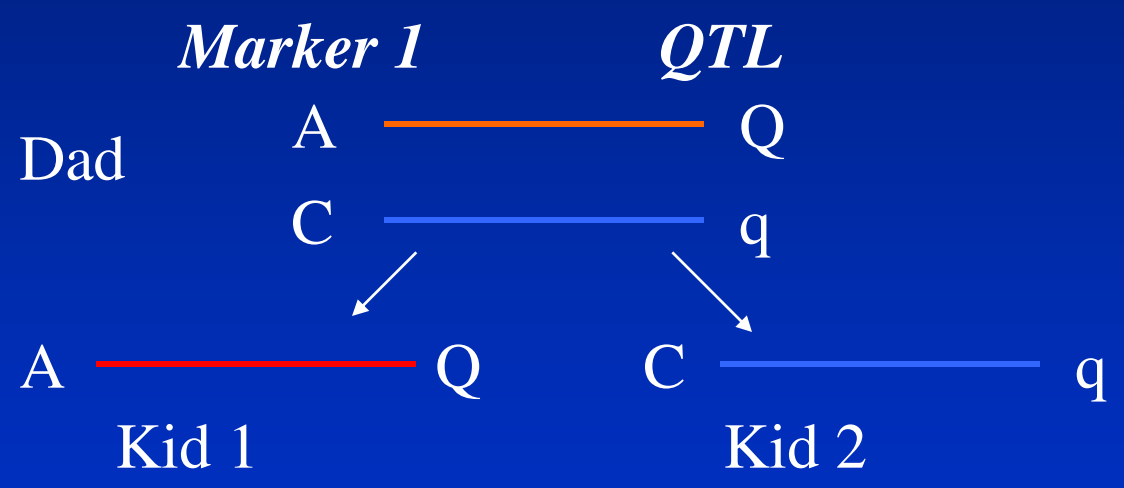
# Approaches to QTL detection

- Candidate gene approach
  - assumes a gene involved in trait physiology could harbour a mutation causing variation in that trait
  - Look for mutations in this gene
  - Some success
  - Number of candidate genes is too large
  - Very difficult to pick candidates!
- Linkage mapping
  - So use *neutral markers* and exploit linkage
    - organisation of the genome into chromosomes inherited from parents

- DNA markers: track chromosome segments from one generation to the next



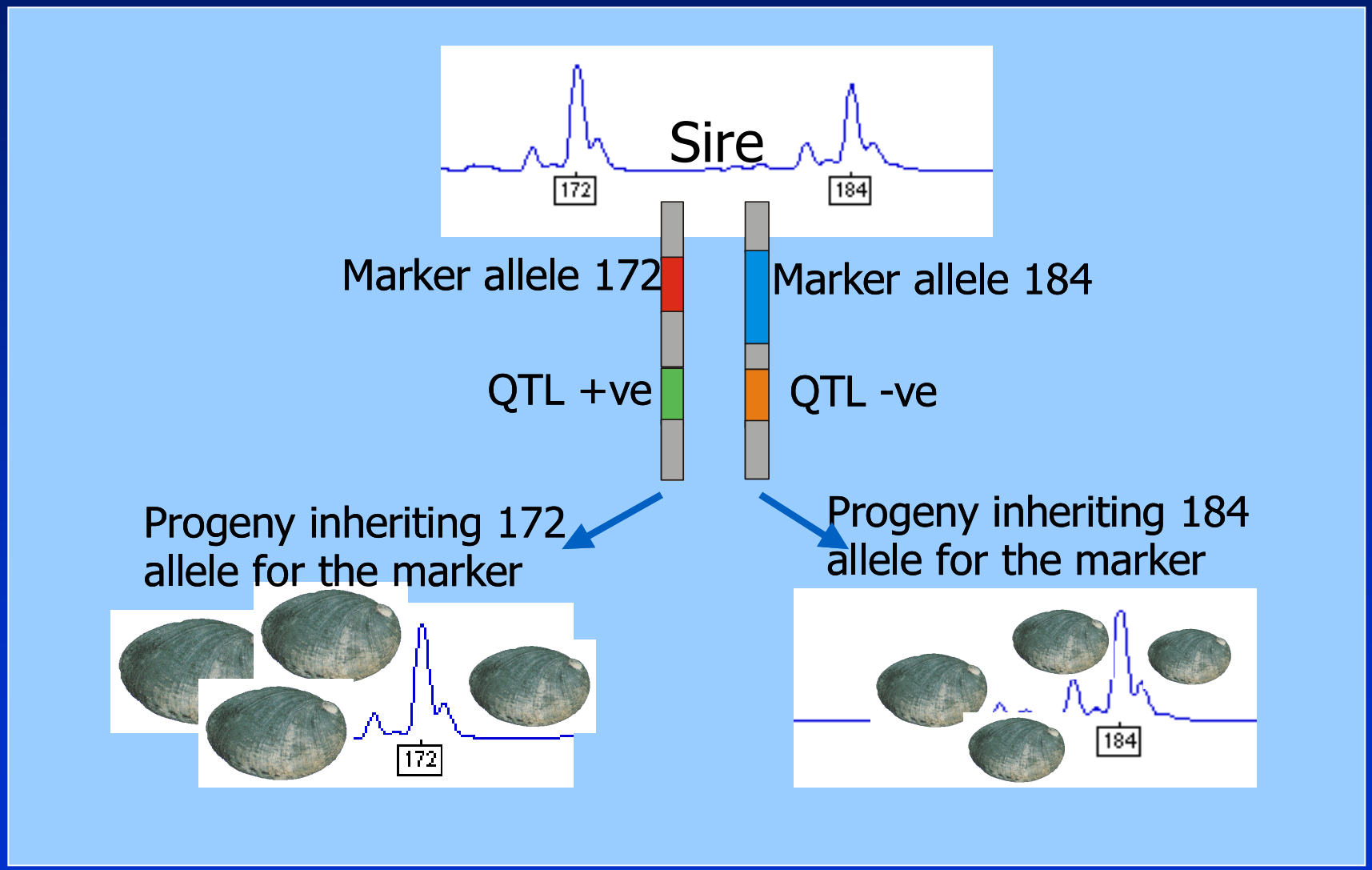
- DNA markers: track chromosome segments from one generation to the next



# Detection of QTL with linkage

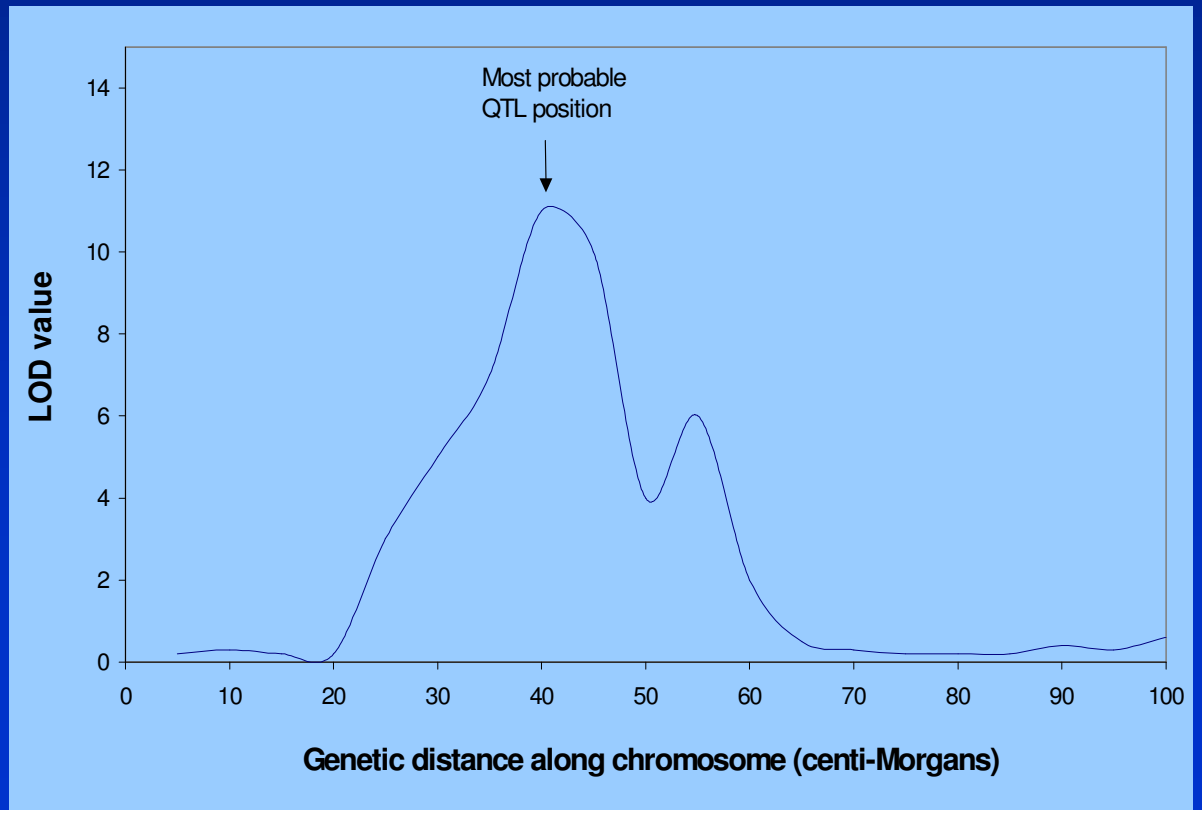
- Principle of QTL mapping
  - Is variation at the molecular level (different marker alleles) linked to variation in the quantitative trait?.
  - If so then the marker is linked to, or on the same chromosome as, a QTL

# Detection of QTL



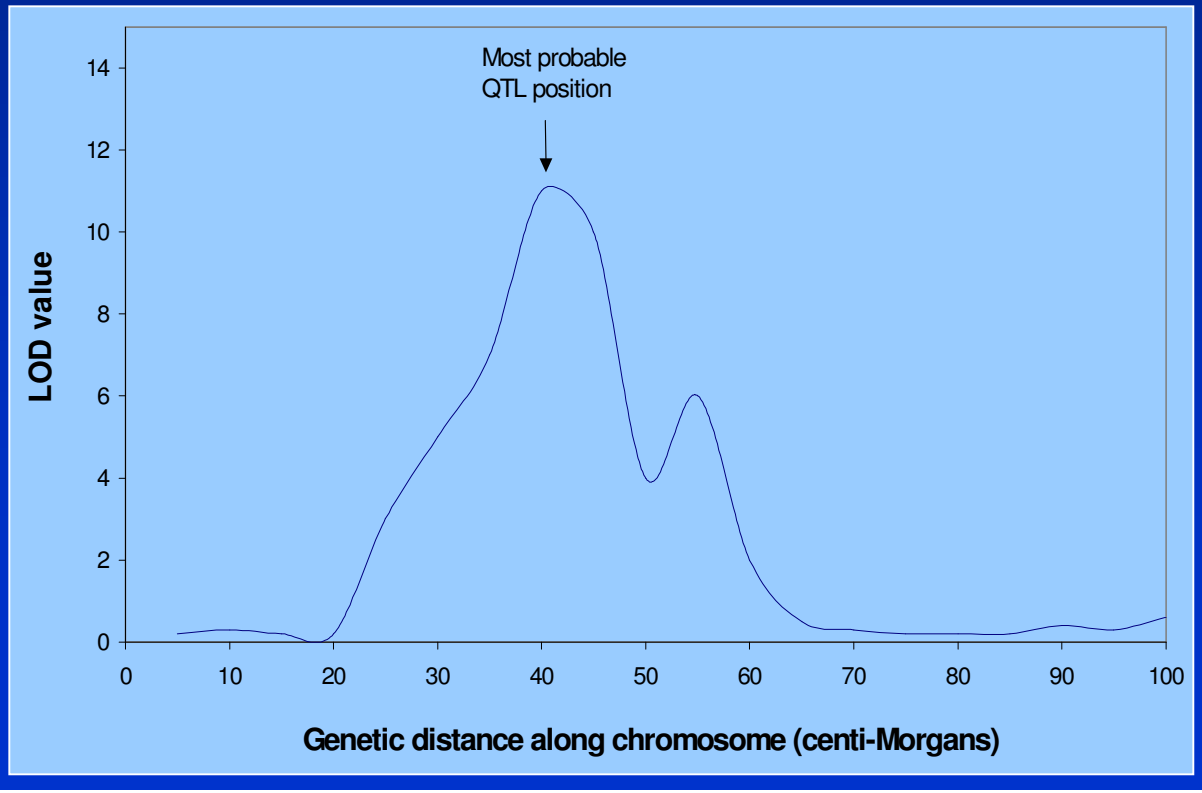
# Detection of QTL with linkage

- Can use single marker associations
- More information with multiple markers ordered on linkage maps



# Problems with linkage mapping

- QTL are not mapped very precisely
- Confidence intervals of QTL location are very wide

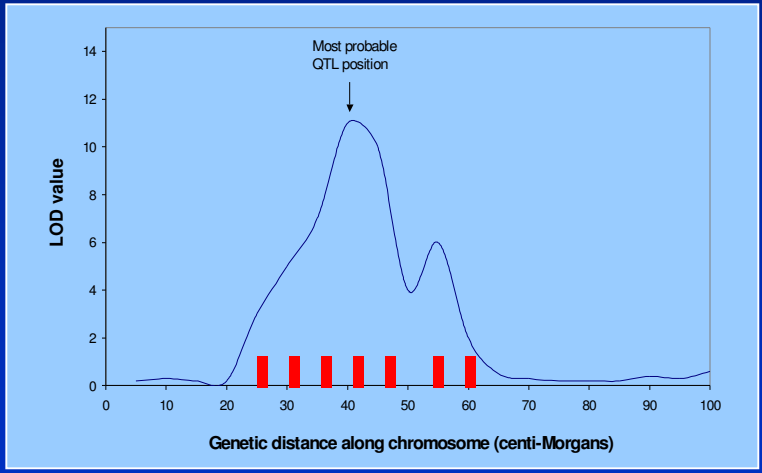


# Problems with linkage mapping

- Difficult to use information in marker assisted selection (MAS)
- Most significant marker can be 10cM or more from QTL
- The association between the marker and QTL unlikely to persist across the population
  - Eg A___Q in one sire family
  - a___Q in another sire family
- The phase between the marker and QTL has to be re-estimated for each family
- Complicates use of the information in MAS
  - Reduces gains from MAS

# Problems with linkage mapping

- Shift to fine mapping
  - Saturate confidence interval with many markers



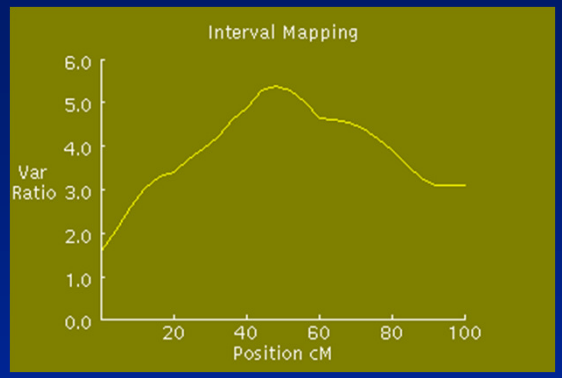

  - Use Linkage disequilibrium mapping approaches within this small chromosome segment

# Problems with linkage mapping

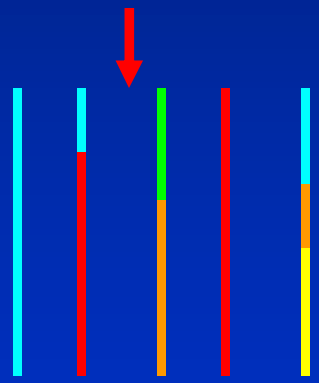
- Shift to fine mapping
  - Saturate confidence interval with many markers
  - Use Linkage disequilibrium mapping approaches within this small chromosome segment
  - Eventually find causative mutation

# DGAT1 - A success story (Grisart et al. 2002)

1. Linkage mapping detects a QTL on bovine chromosome 14 with large effect on fat % (Georges et al 1995)




2. Linkage disequilibrium mapping refines position of QTL (Riquet et al. 1999)



3. Selection of candidate genes. Sequencing reveals point mutation in candidate (DGAT1). This mutation found to be functional - substitution of lysine for analine. Gene patented. (Grisart et al. 2002)

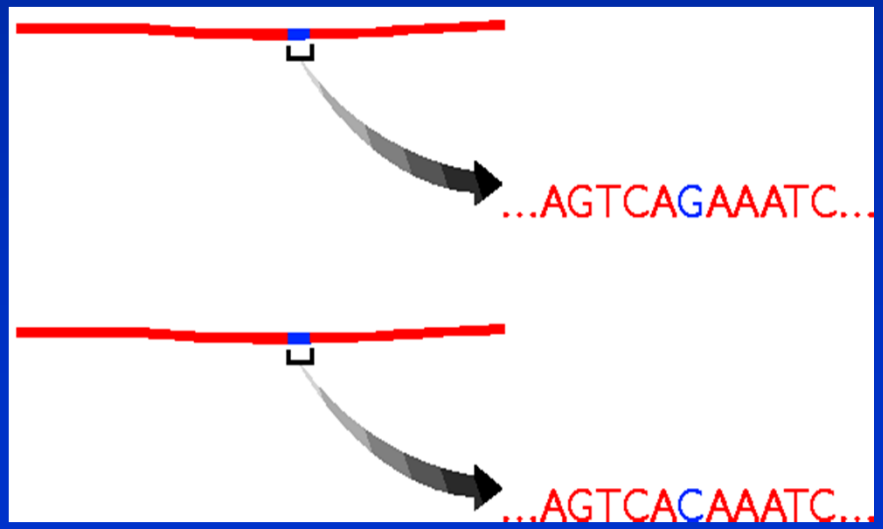
ACCTGGGAGAC
CAGGGAG

# Problems with linkage mapping

- But process is very slow
  - 10 years or more to find causative mutation
  - One limitation has been the density of markers

# The Revolution

- As a result of sequencing animal 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, Soybean under development
- Cost?
  - ~ $100-200 USD for 60K SNPs
- Genotyping by re-sequencing?
  - 40 million SNPs in cattle
  - Insertion deletions
  - Copy number variants?





# The Revolution

- Can we use SNP and sequence information to accelerate rates of genetic gain in the livestock industries?
  - Omit linkage mapping
  - Straight to genome wide association
  - Genomic selection = breeding values directly from markers or sequence ?

# Aim

- Provide you with the tools to use high density SNP and other variant genotypes in livestock and plant improvement

# Linkage disequilibrium

- A brief history of QTL mapping
- Measuring linkage disequilibrium
- Causes of LD
- Extent of LD in animals and plants
- The extent of LD between breeds and lines
- Strategies for haplotyping

# Definitions of LD

- Why do we need to define and measure LD?
- Both genomic selection and genome wide association studies assume markers to be in LD with QTL
- Determine the number of markers required for LD mapping and/or genomic selection

# 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…….

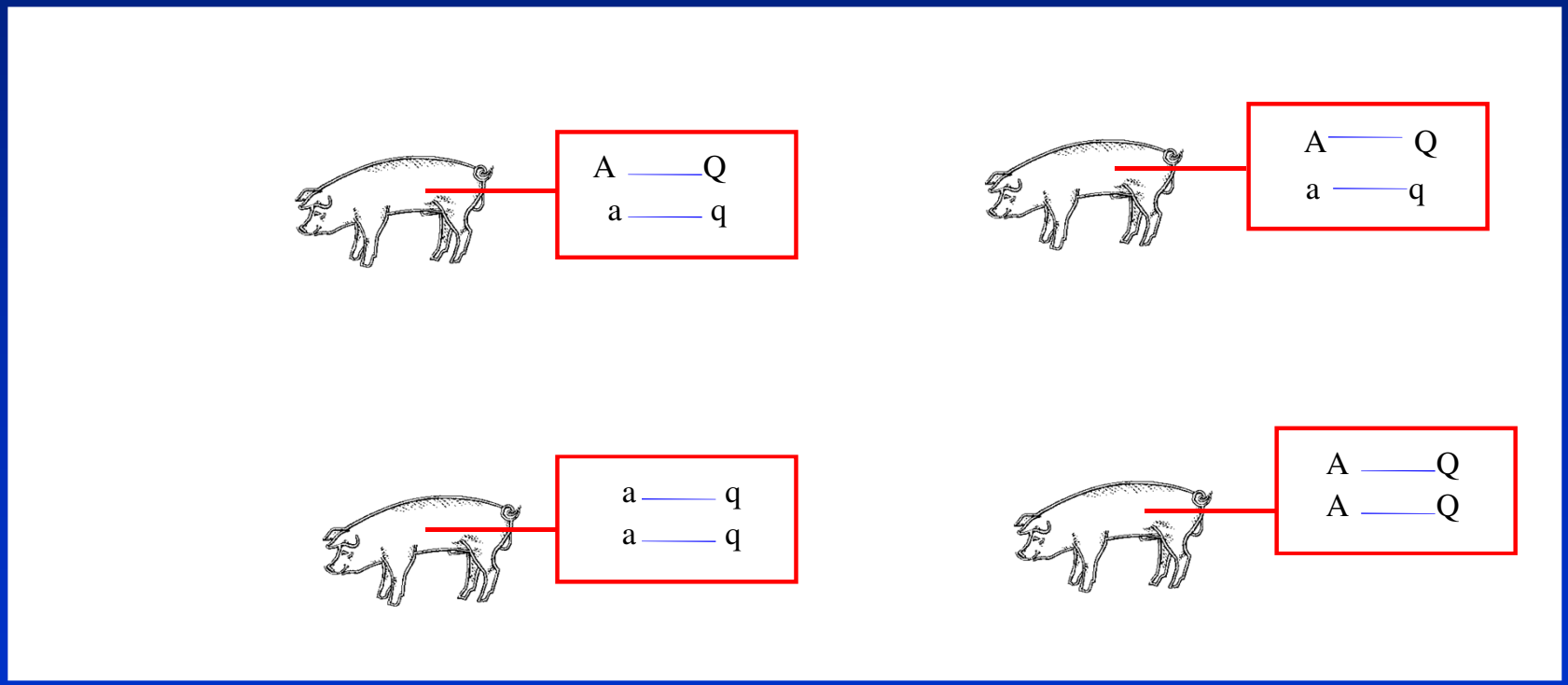
| | | *Marker A* | | |
|---|---|---|---|---|
| | | A1 | A2 | Frequency |
| *Marker B* | B1 | 0.25 | 0.25 | 0.5 |
| | B2 | 0.25 | 0.25 | 0.5 |
| | Frequency | 0.5 | 0.5 | |

# 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 $40kg^2$.
- Key parameter determining the power of LD mapping to detect QTL
  - Experiment sample size must be increased by $1/r^2$ to have the same power as an experiment observing the QTL directly

# Definitions of LD

- Another LD statistic is D’
  - |D|/Dmax
  - Where
    - Dmax
      - = min[freq(A1)*freq(B2),(1-freq(A2))(1-freq(B1))]
      - if D>0, else
      - = min[freq(A1)(1-freq(B1),(1-(freq(A2))*freq(B2)]
      - if D<0.
  - But what does it mean?
  - Biased upward with low allele frequencies
  - Overestimates $r^2$

# Definitions of LD

- Another LD statistic is D’
  - |D|/Dmax
  - Where
    - Dmax
      - = min[freq(A1)*freq(B2),(1-freq(A2))(1-freq(B1))]
      - if D>0, else
      - = min[freq(A1)(1-freq(B1),(1-(freq(A2))*freq(B2)]
      - if D<0.
  - But what does it mean?
  - Biased upward with low allele frequencies
  - Overestimates $r^2$

# Linkage disequilibrium

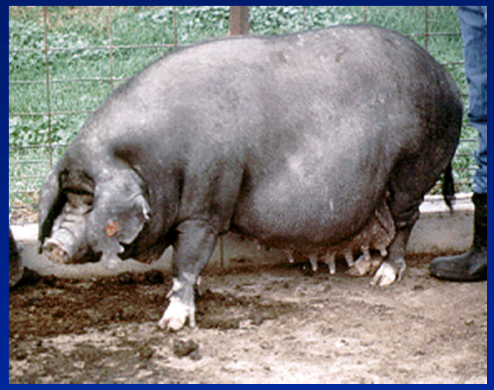
- A brief history of QTL mapping
- Measuring linkage disequilibrium
- Causes of LD
- Extent of LD in animals and plants
- The extent of LD between breeds and lines
- Strategies for haplotyping

# 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

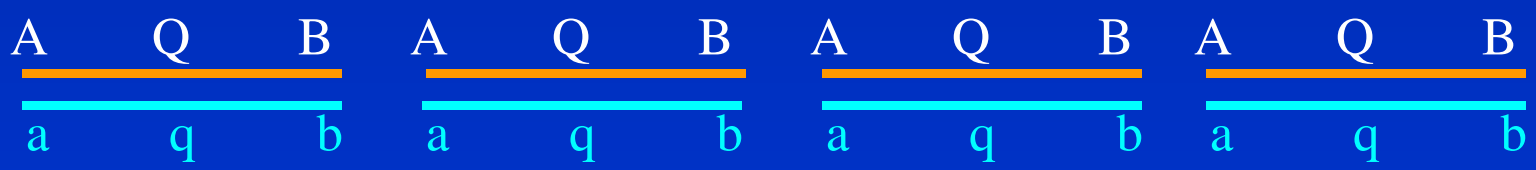
- F2 design

X



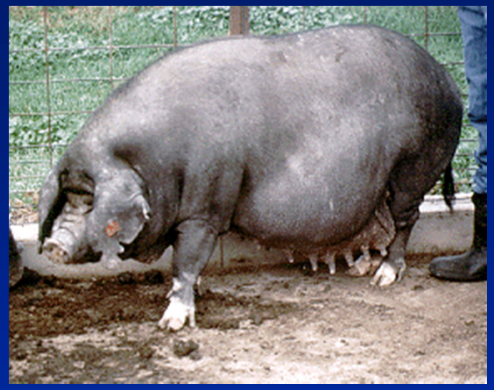
Parental Lines


F1

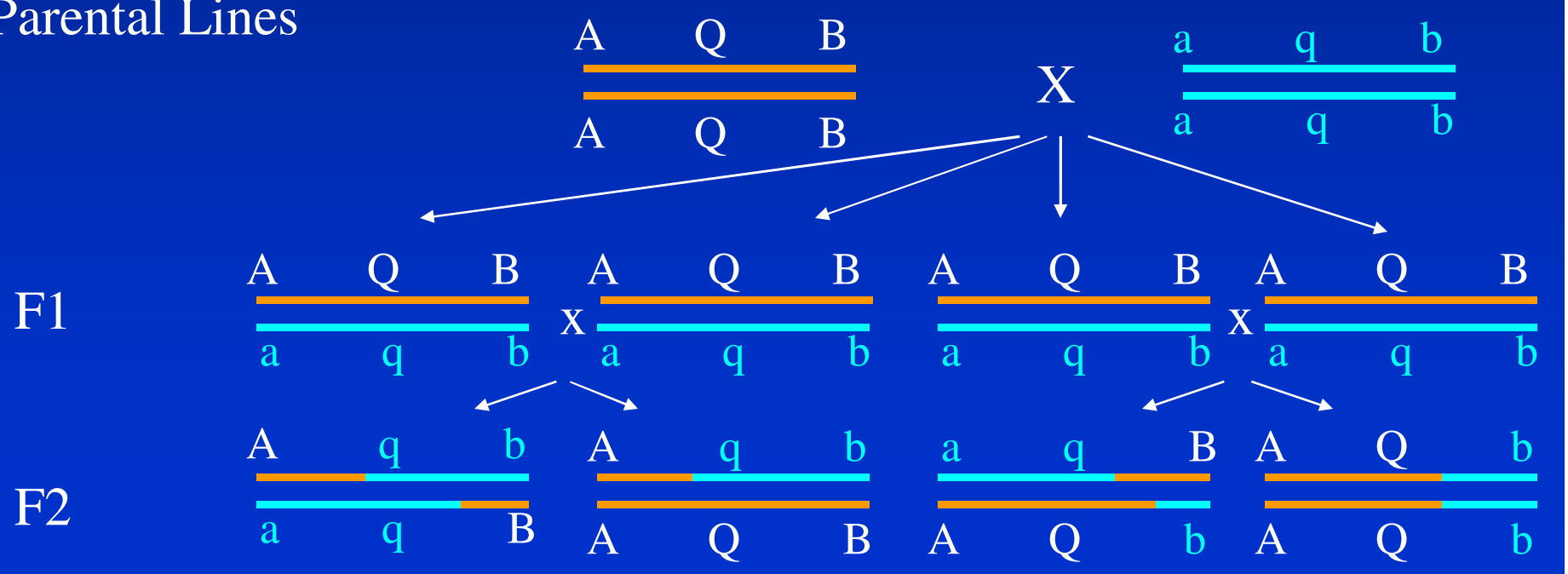


- F2 design

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

# Causes of LD

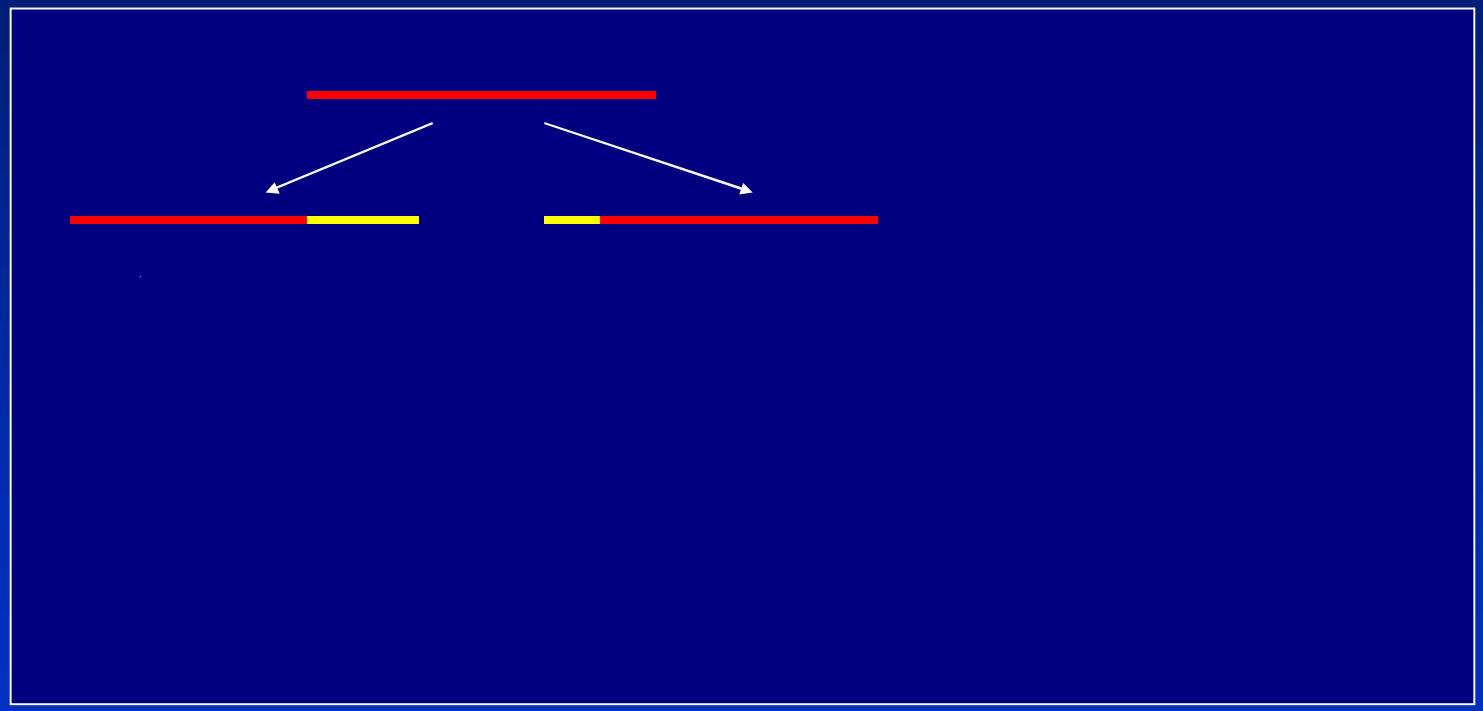
- Migration
  - LD artificially created in crosses designs
    - large when crossing inbred lines
    - but small when crossing breeds that do not differ markedly in gene frequencies
    - disappears after only a limited number of generations
- Selection
  - Selective sweeps
- Small finite population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

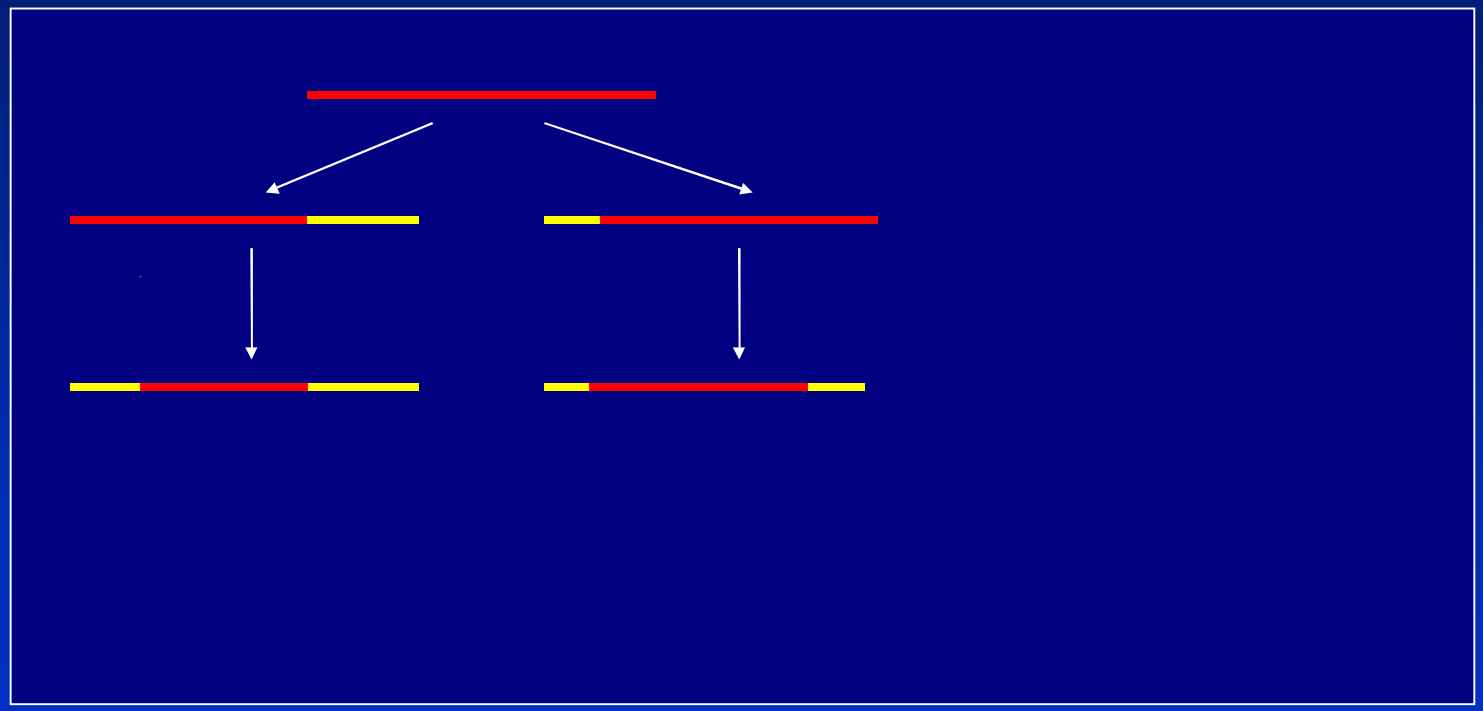
# Causes of LD

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



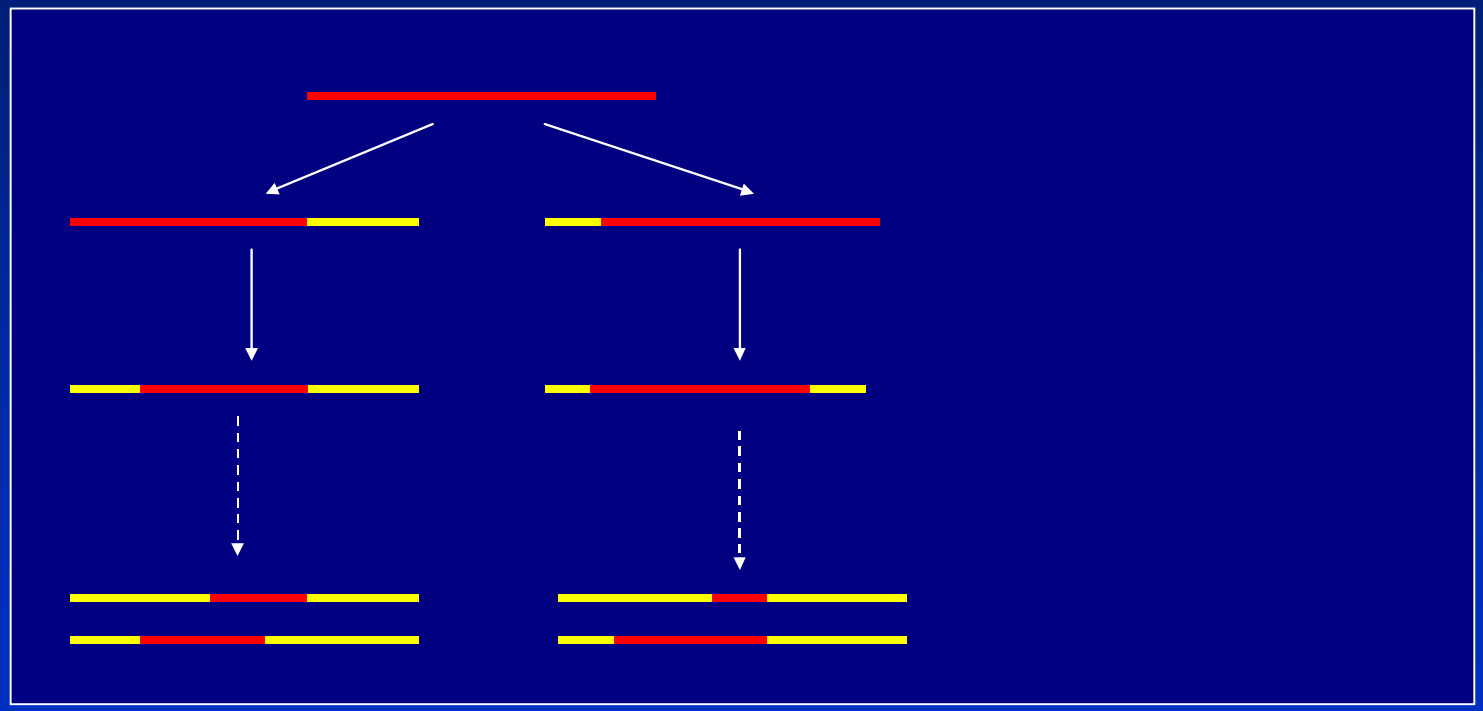
# Causes of LD

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



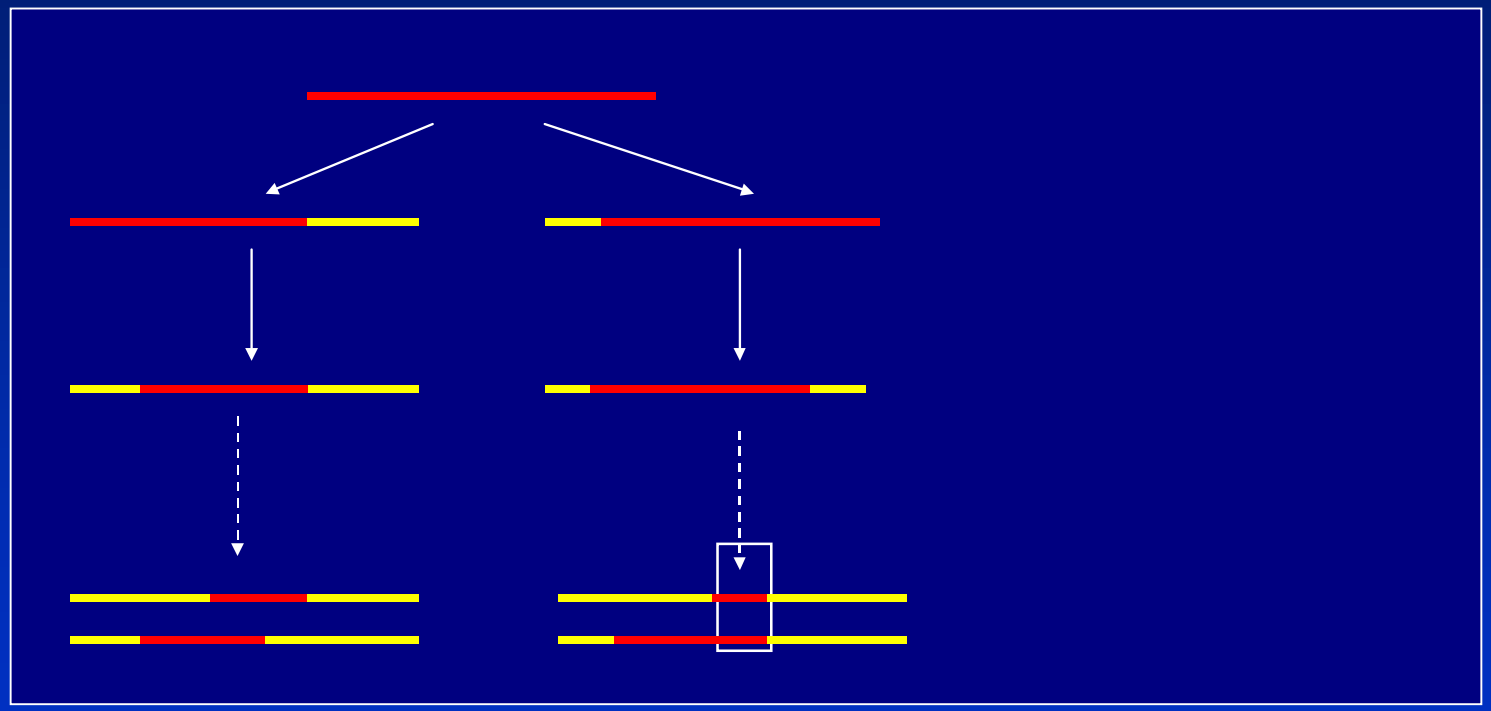
# Causes of LD

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



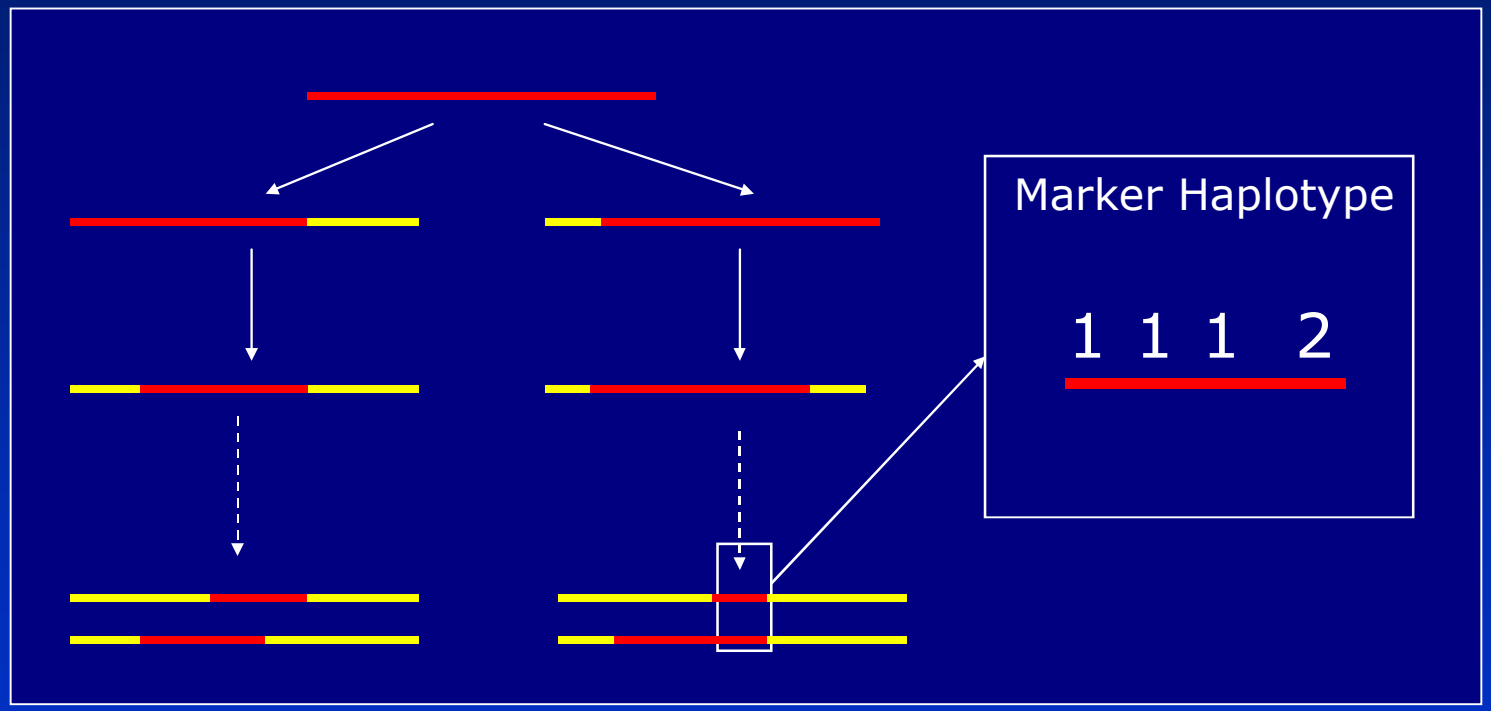
# Causes of LD

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



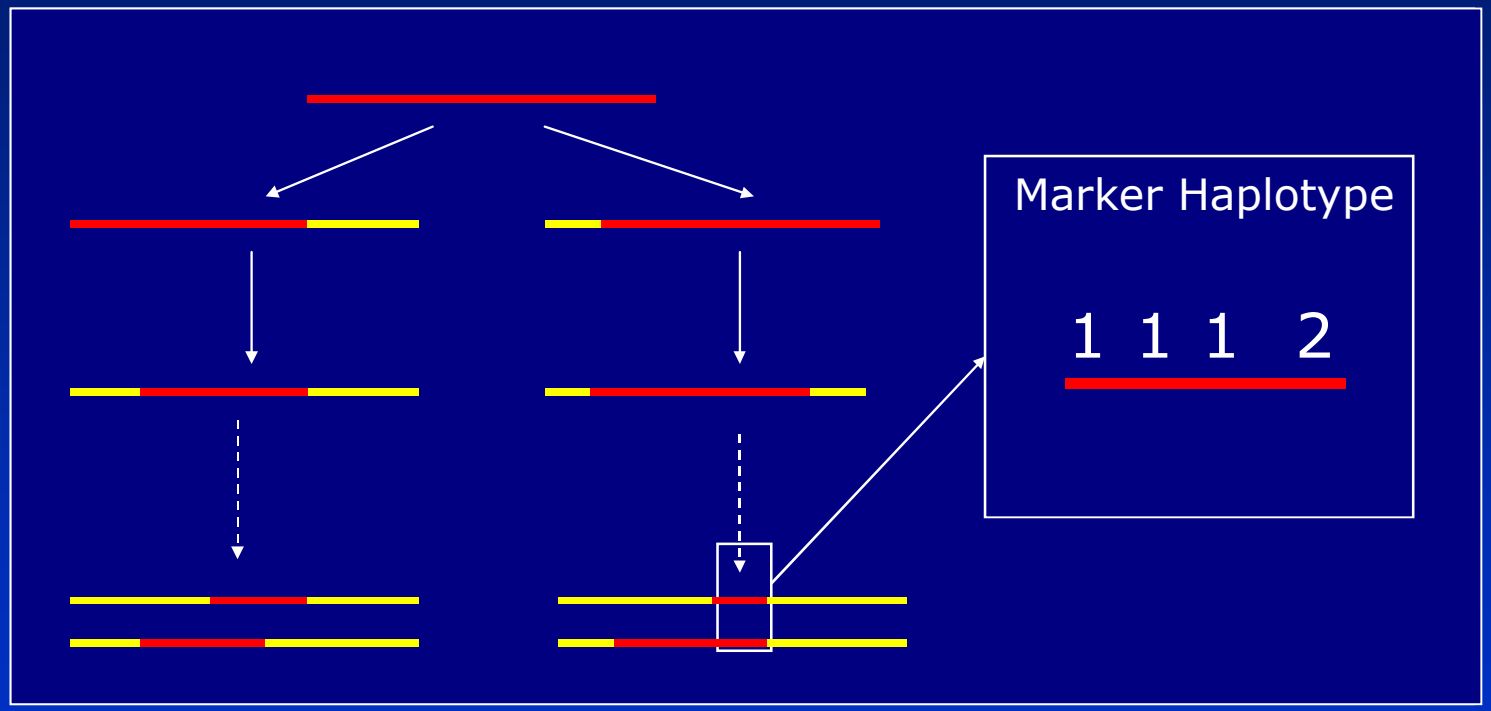
# Causes of LD

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



# Causes of LD

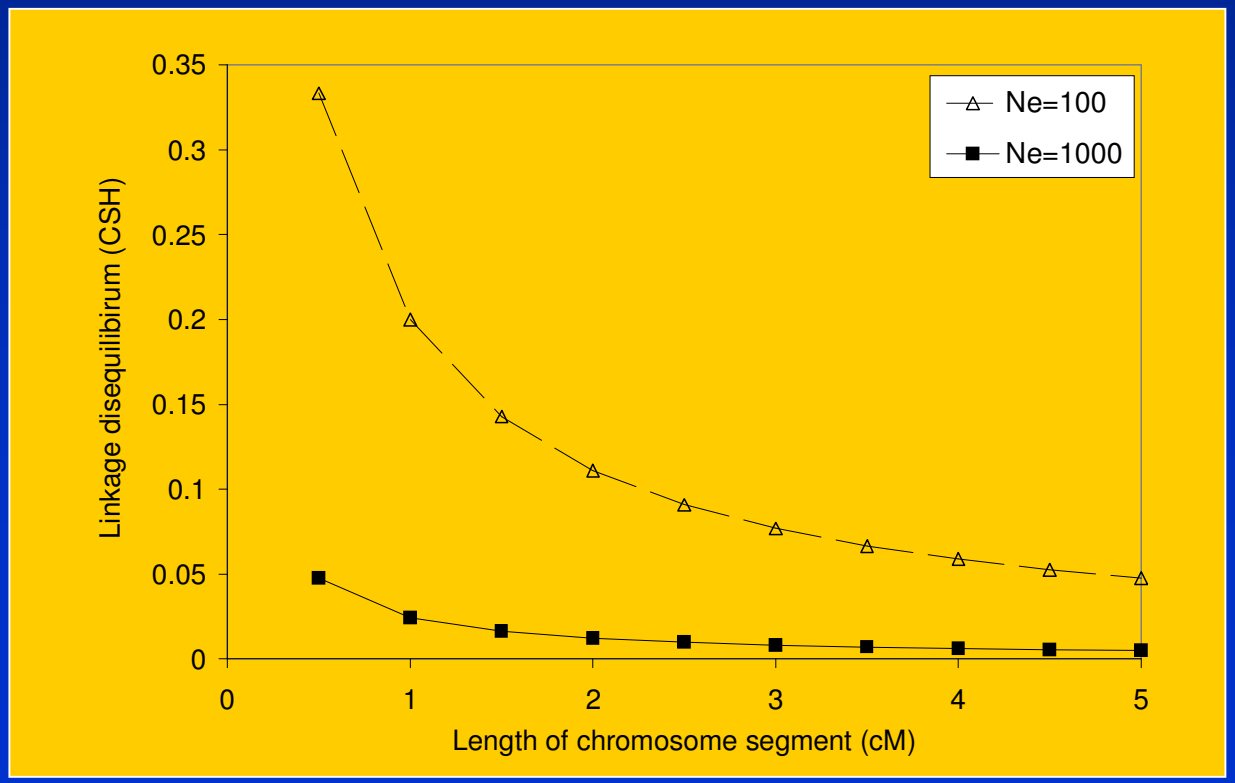
- A chunk of ancestral chromosome is conserved in the current population




- 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

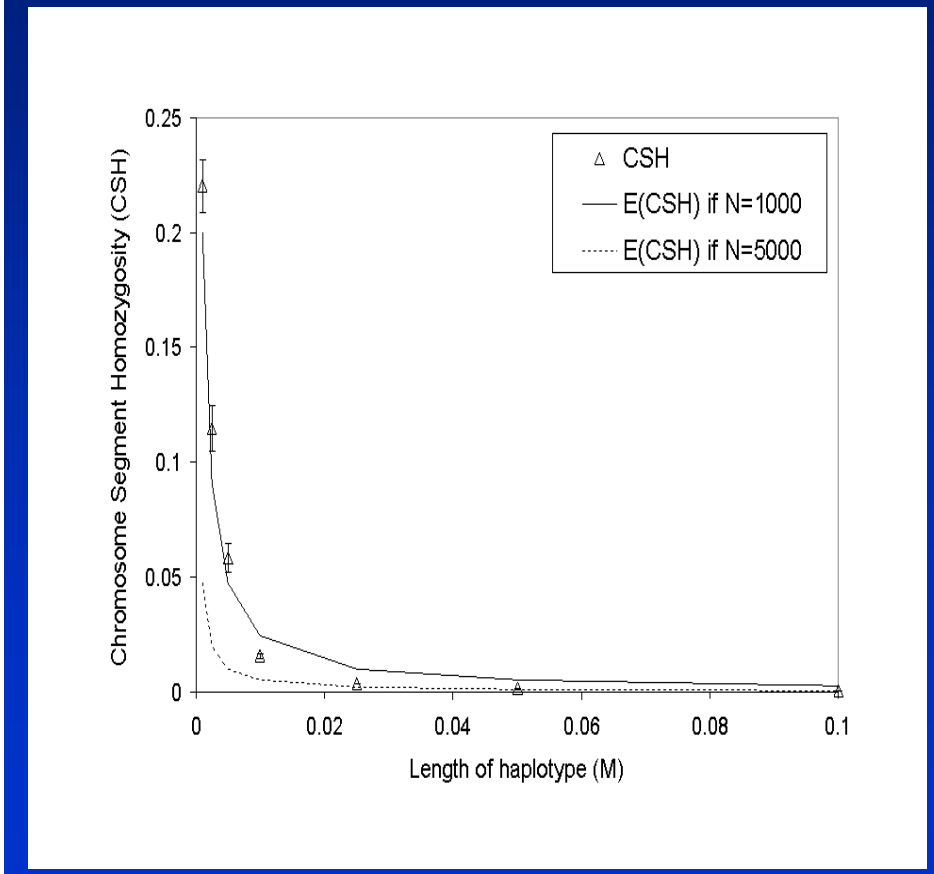


# Causes of LD

- 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 ?

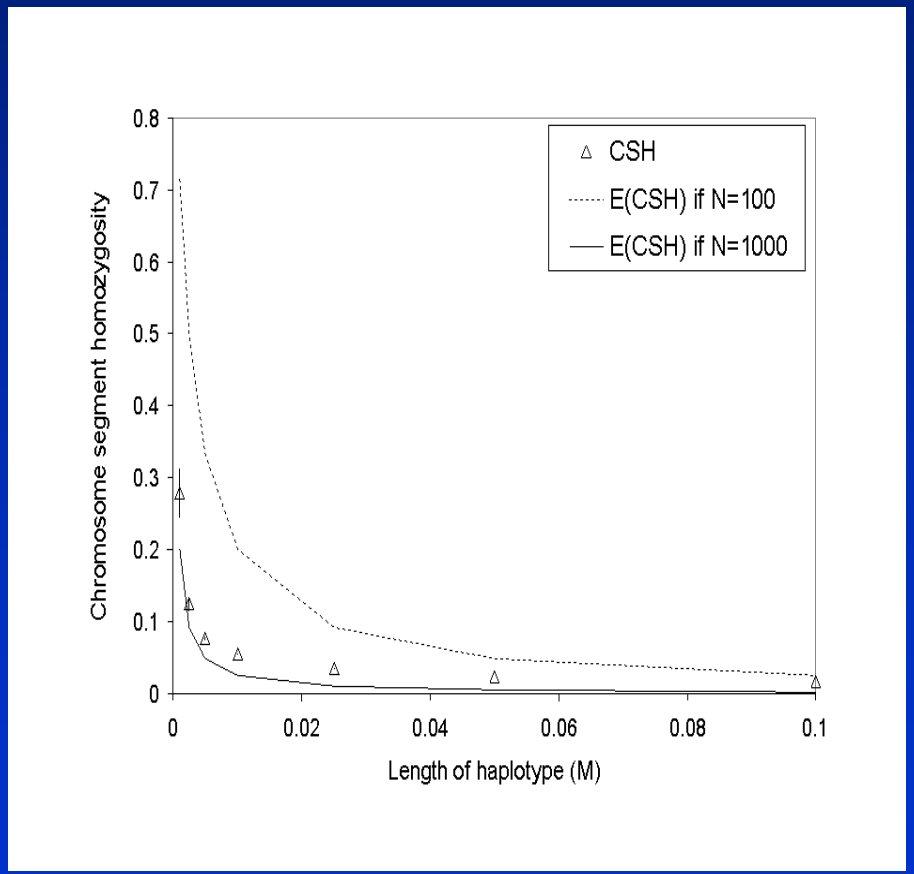
# Causes of LD

1000 to 5000




A

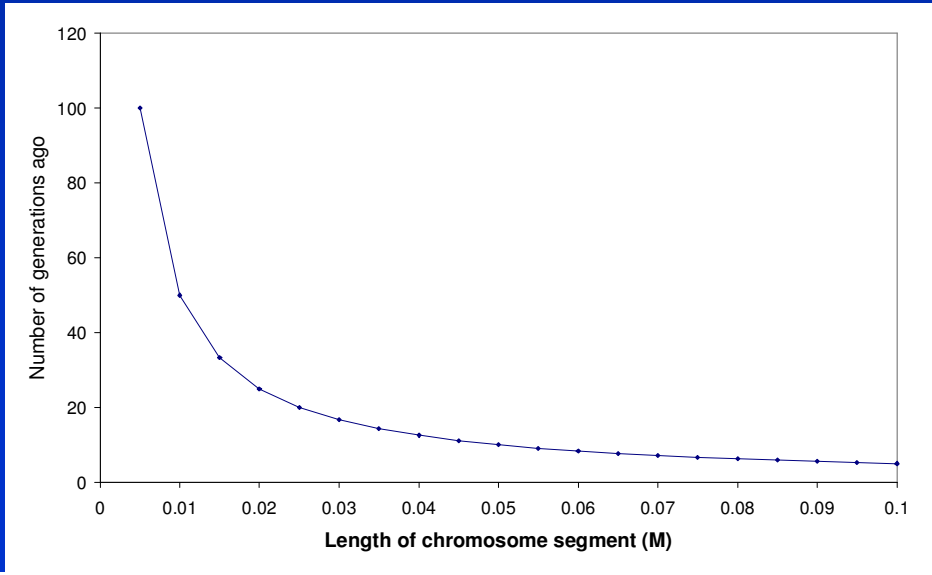
1000 to 100




B

# Causes of LD

- $E(r^2) = 1/(4N_t c+1)$
- 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

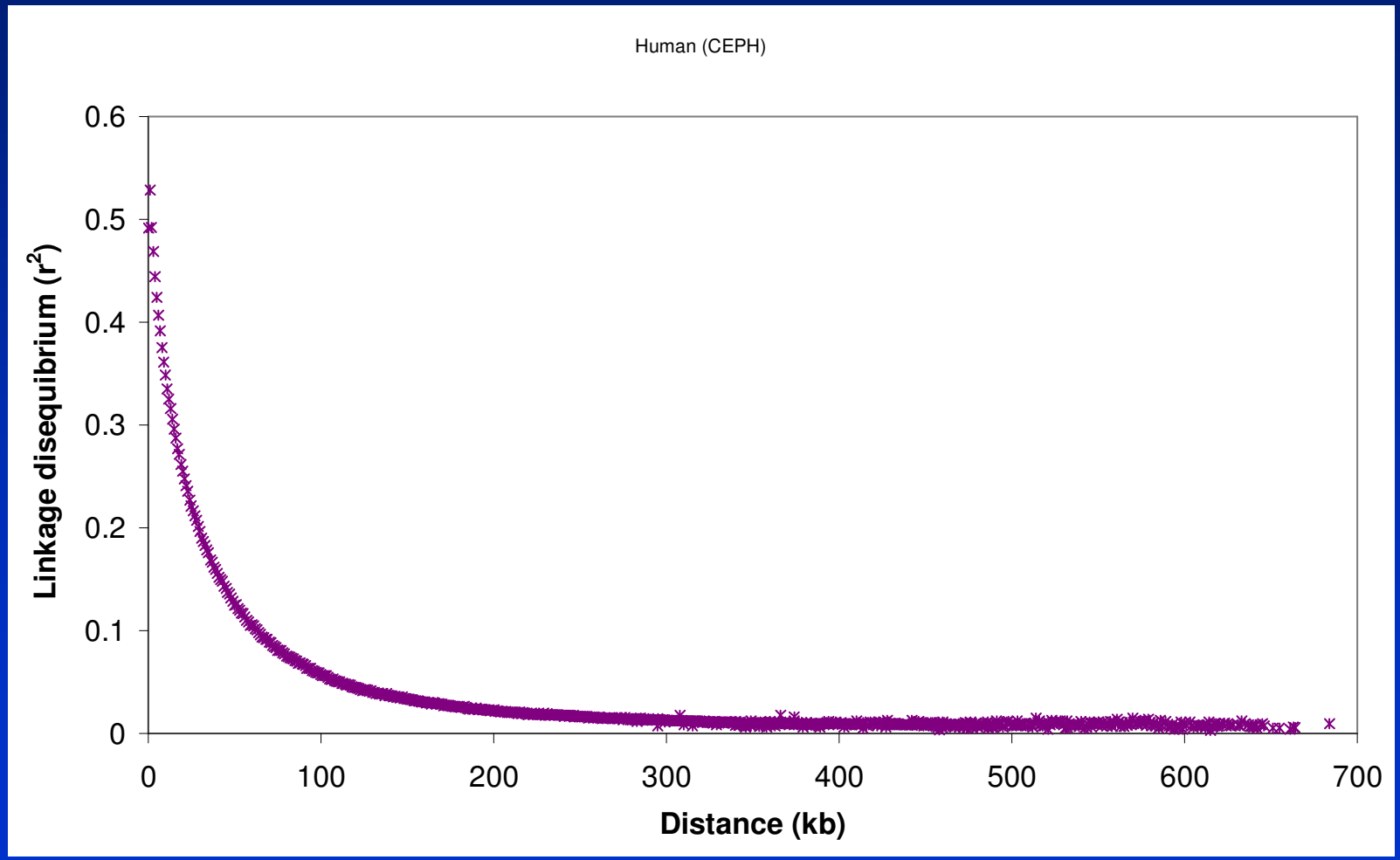
- $E(r^2) = 1/(4N_t c+1)$
- Where t = 1/(2c) generations ago
  - eg markers 0.1M (10cM) apart reflect population size 5 generations ago
  - Markers 0.001 (0.1cM) apart reflect effective pop size 500 generations ago
- LD at short distances reflects historical effective population size
- LD at longer distances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

# Linkage disequilibrium

- A brief history of QTL mapping
- Measuring linkage disequilibrium
- Causes of LD
- Extent of LD in animals and plants
- The extent of LD between breeds and lines
- Strategies for haplotyping

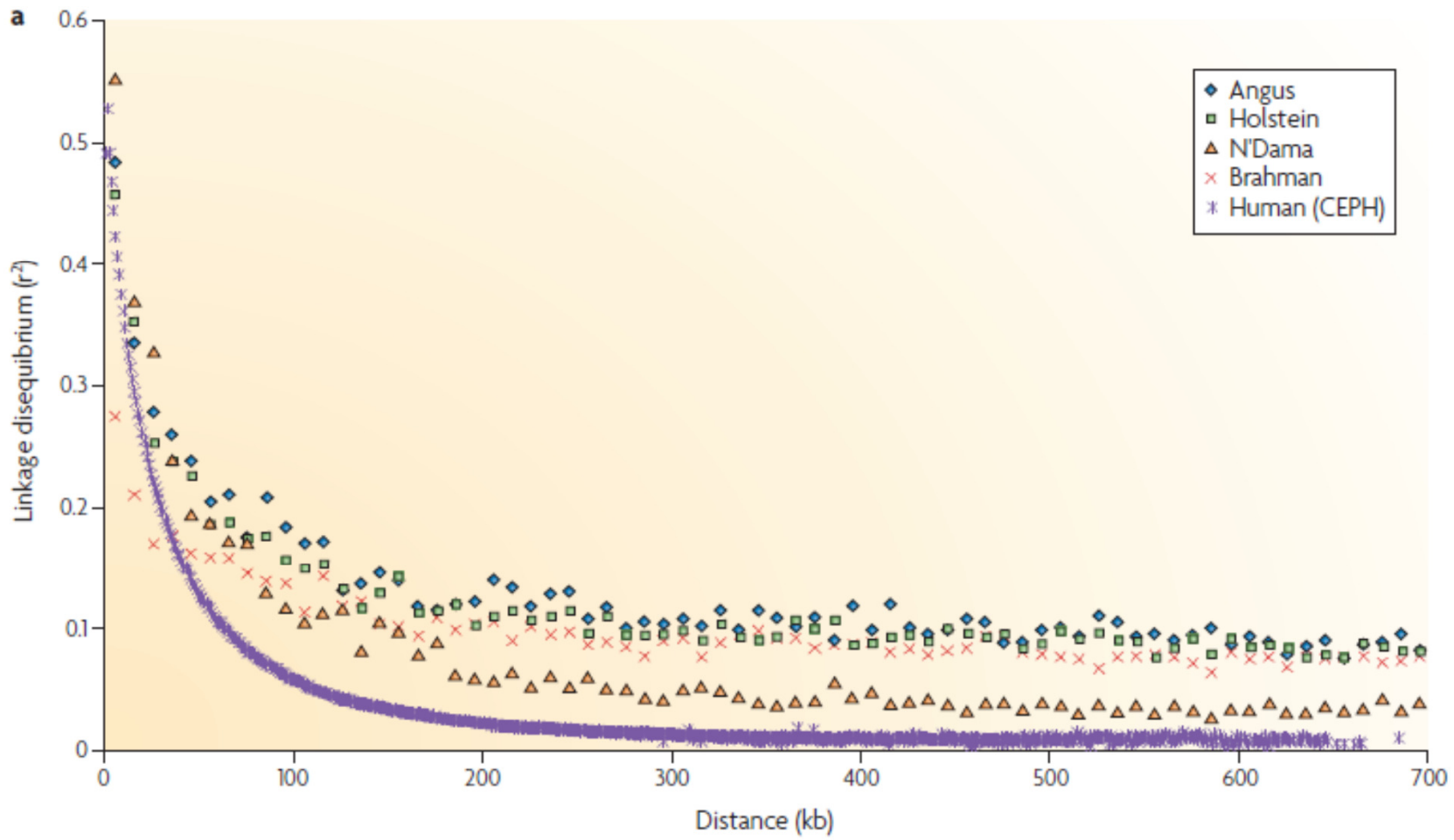
# Extent of LD in humans and livestock

Humans…….(Tenesa et al. 2007)



# Extent of LD in humans and livestock

And cattle……

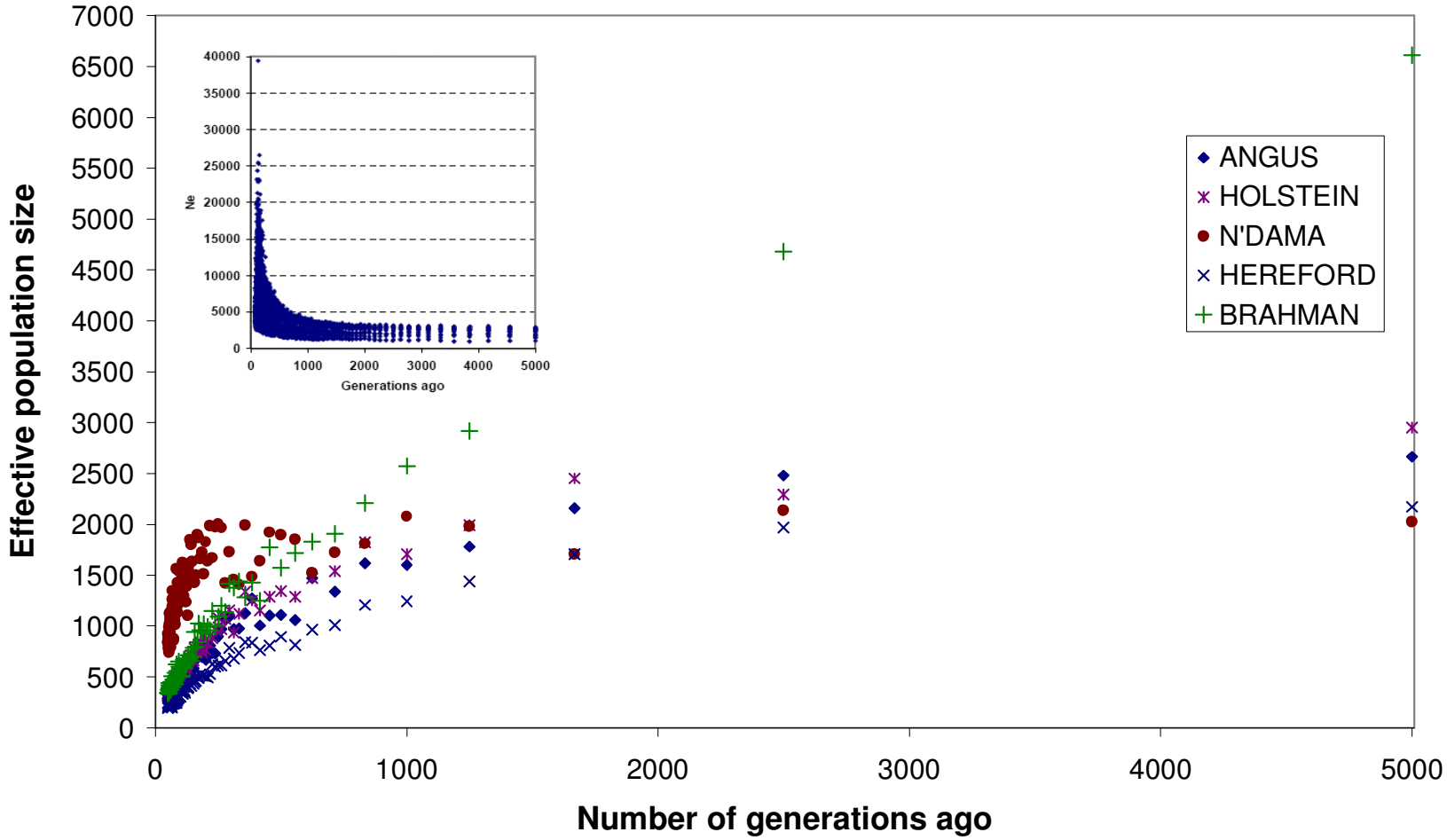


# Causes of LD

- $E(r^2) = 1/(4N_t c+1)$
- 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 a genome wide association study of reasonable size to detect QTL of moderate effect.
- 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

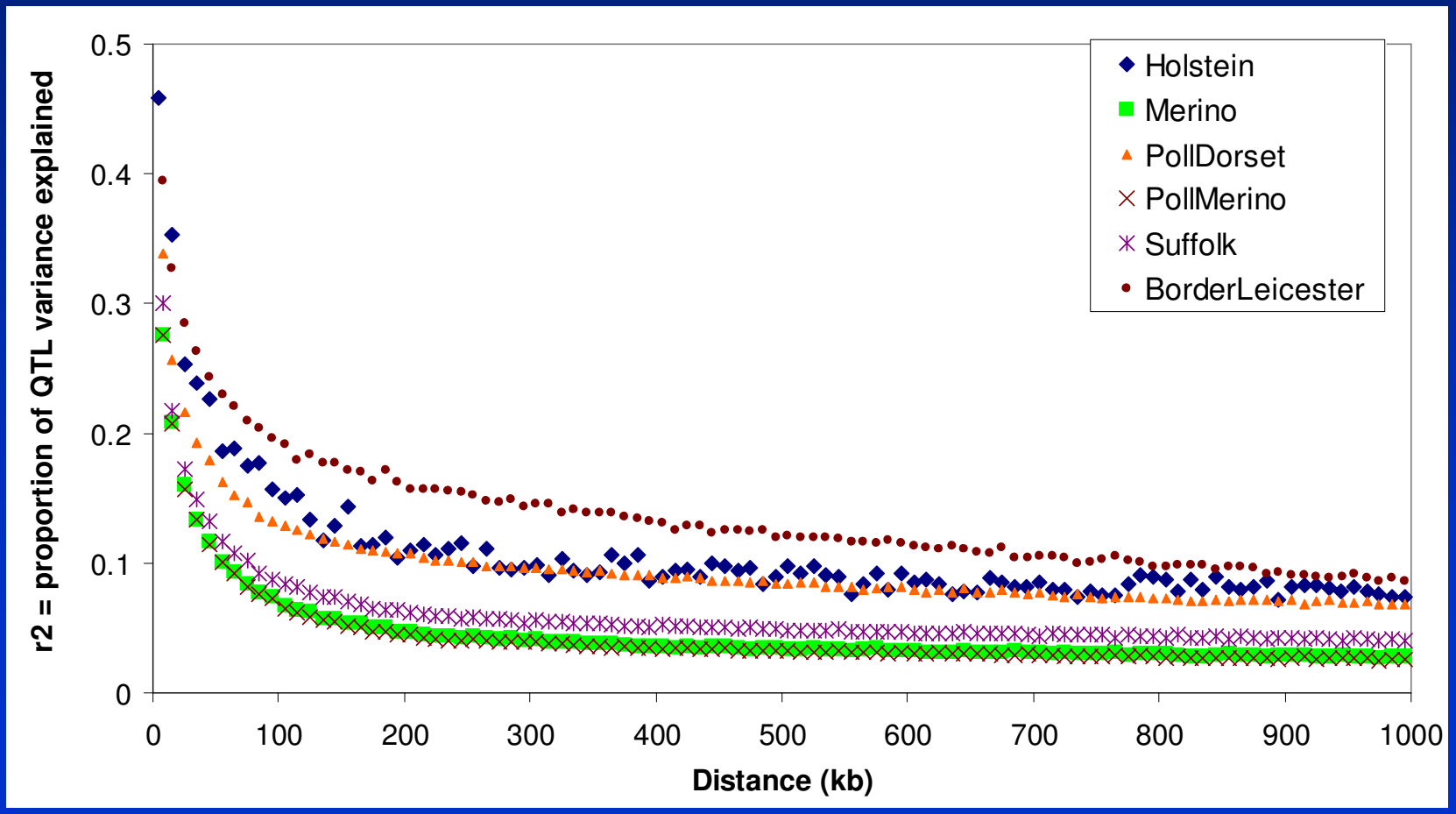
- Pigs
  - Du et al. (2007) assessed extent of LD in pigs using 4500 SNP markers in six lines of commercial pigs.
  - Their results indicate there may be considerably more LD in pigs than in cattle.
  - $r^2$ of 0.2 at 1000kb.
  - LD of this magnitude only extends 100kb in cattle.
  - In pigs at a 100kb average $r^2$ was 0.371.

# Extent of LD in other species

- Chickens
  - Heifetz et al. (2005) evaluated the extent of LD in a number of populations of breeding chickens.
  - In their populations, they found significant LD extended long distances.
  - For example 57% of marker pairs separated by 5-10cM had $\chi 2' \geq 0.2$ in one line of chickens and 28% in the other.
  - Heifetz et al. (2005) pointed out that the lines they investigated had relatively small effective population sizes and were partly inbred

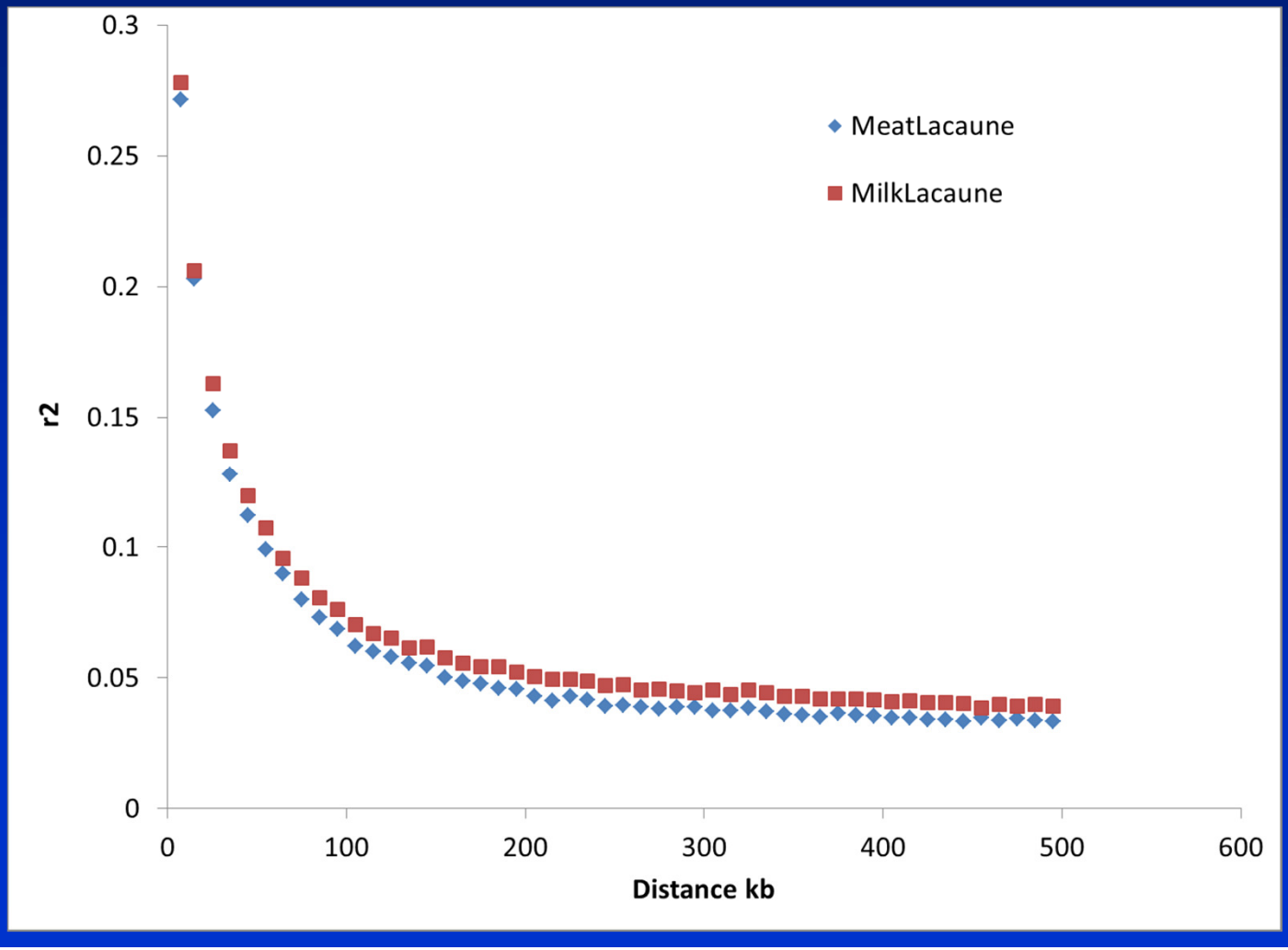
# Extent of LD in other species

- Sheep HapMap project (Kijas et al. 2011)



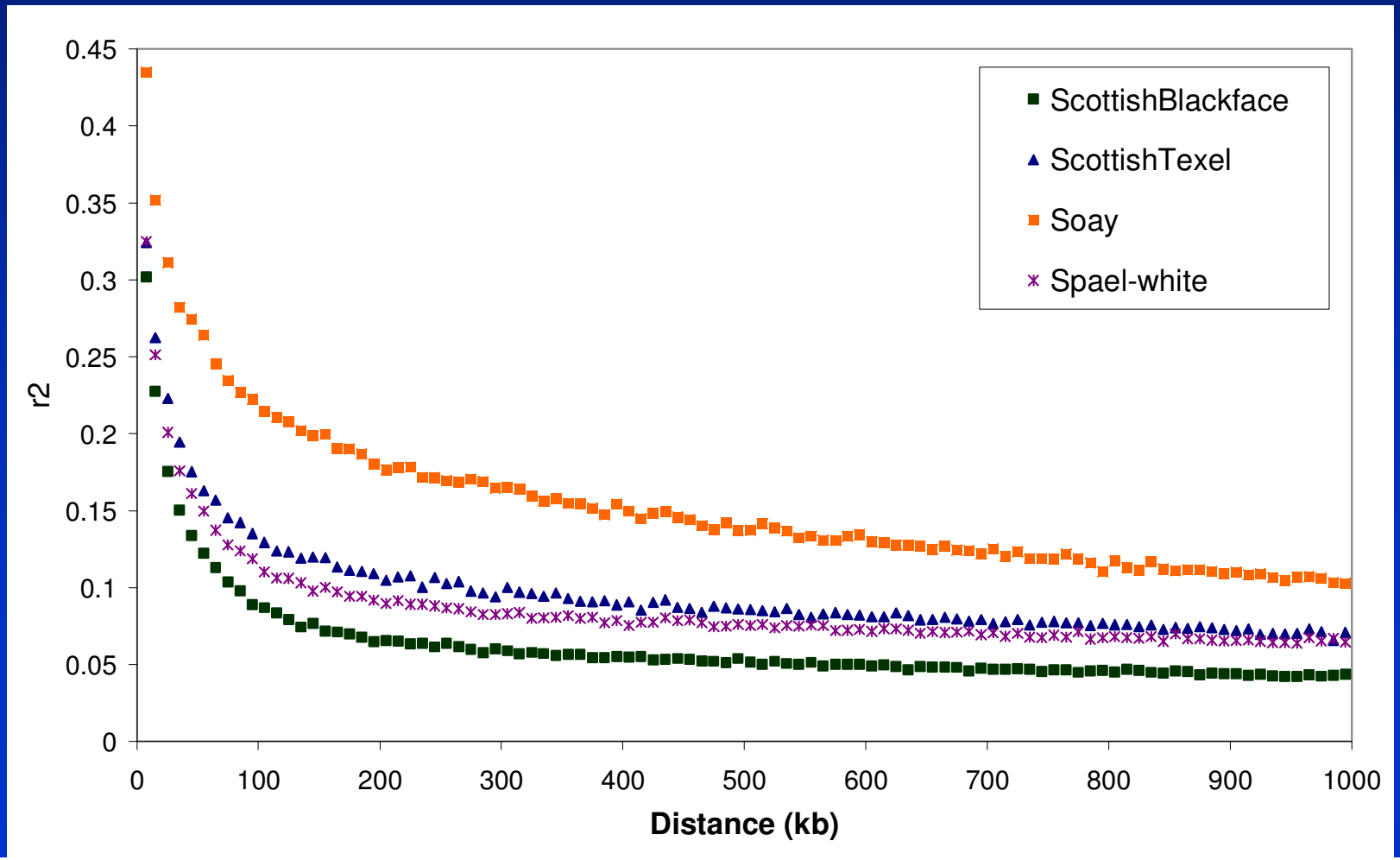
# Extent of LD in other species

- Sheep HapMap project (Kijas et al. 2011)



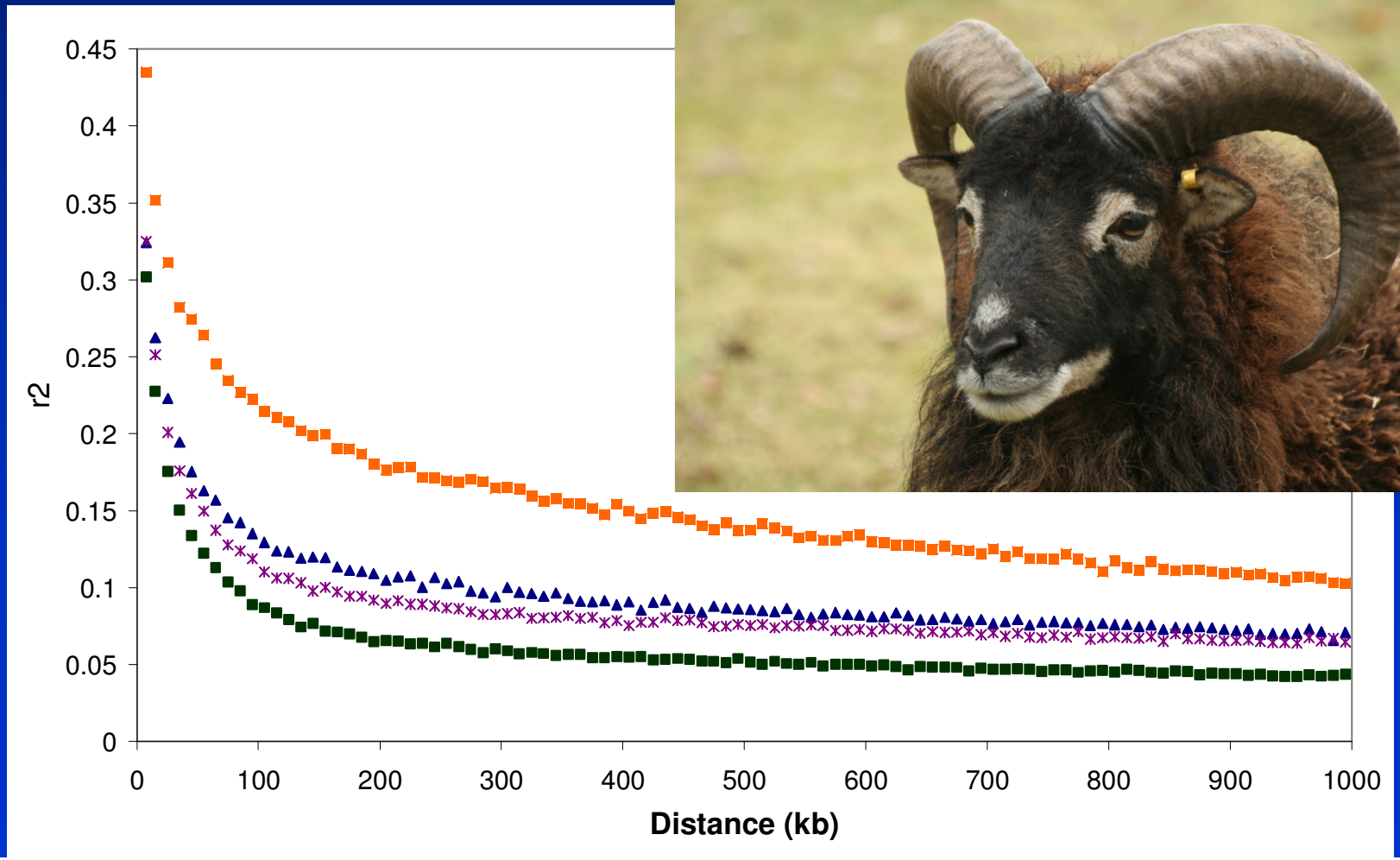
# Extent of LD in other species

- Sheep HapMap project (Kijas et al. 2011)



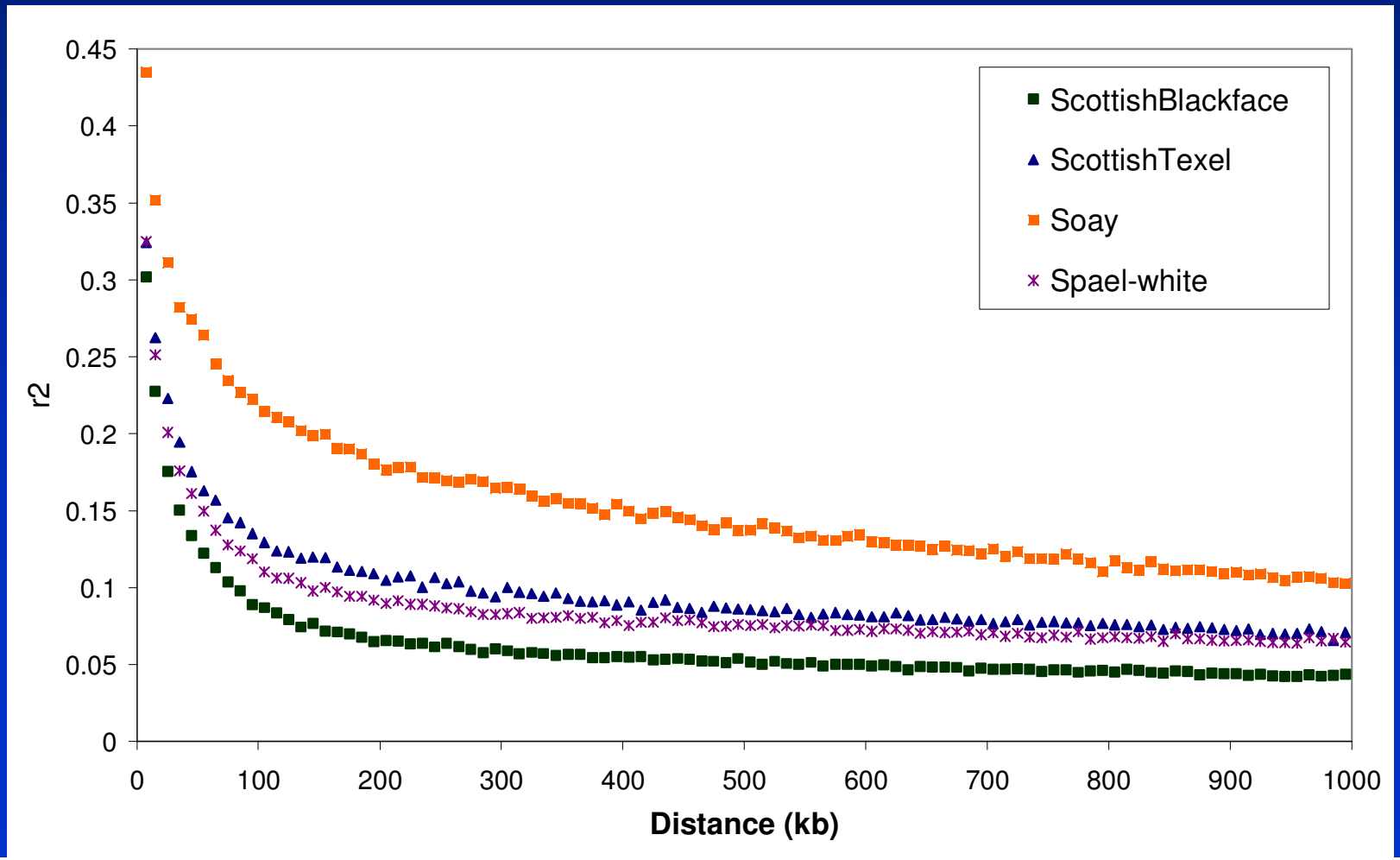
# Extent of LD in other species

- Sheep HapMap project (Kijas et al. 2011)



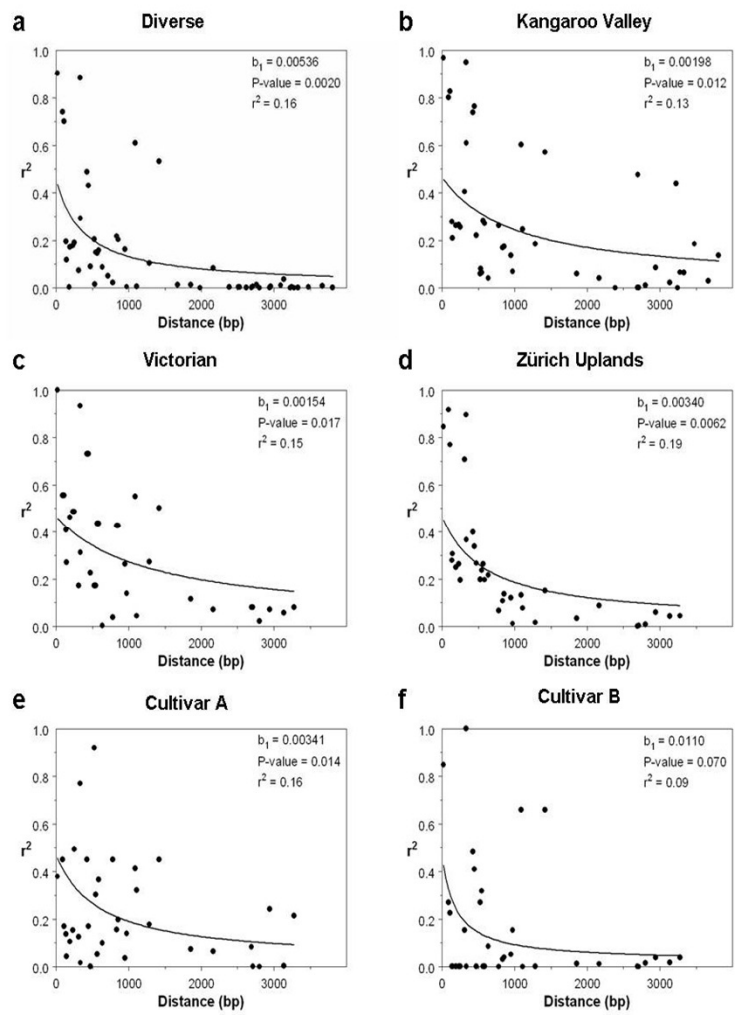

# Extent of LD in other species

- Sheep HapMap project (Kijas et al. 2011)

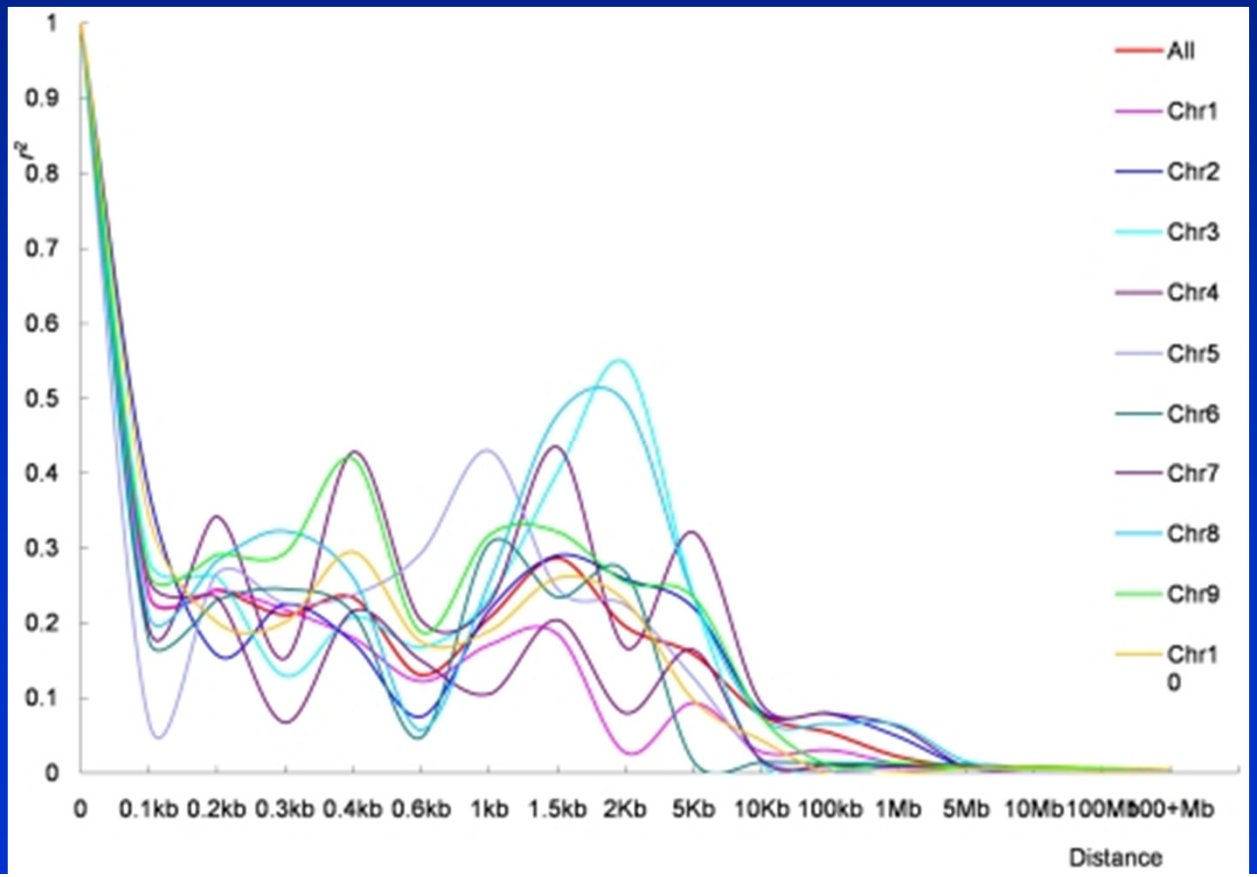


# Extent of LD in other species

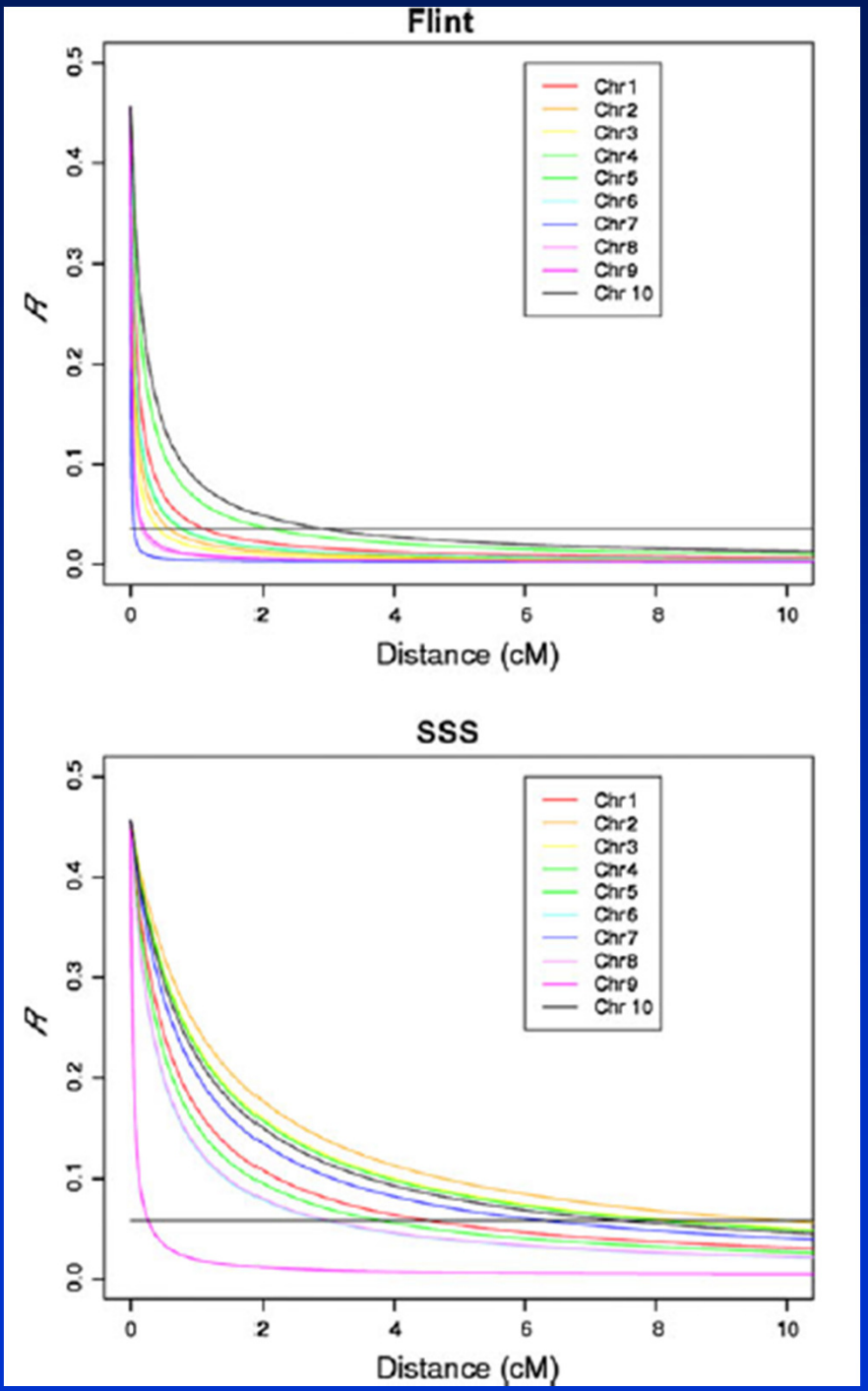
- Perennial ryegrass
  - Ponting et al. 2007
  - an outbreeder
  - very little LD
  - Extremely large effective population size?



- Maize (i)
  - Yan et al. 2009 (PLoS One. 4:e8451).
  - Inbreeder
  - 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
  - Inbreeder
  - Considerable LD among heterotic groups
  - Concluded 4000-65,000 SNPs needed for genome wide association



# Linkage disequilibrium

- A brief history of QTL mapping
- Measuring linkage disequilibrium
- Causes of LD
- Extent of LD in animals and plants
- The extent of LD between breeds and lines
- Strategies for haplotyping

# Persistence of LD across breeds

- Can the same marker be used across breeds?
  - Genome wide LD mapping expensive, can we get away with one experiment?
- The $r^2$ statistic between two SNP markers at same distance in different breeds can be same value even if phases of haplotypes are reversed
- However they will only have same value and sign for *r* statistic if the phase is same in both breeds or populations.

# Persistence of LD across breeds

| | | *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 | |

*Breed 1*

$$r = \frac{\left(freq(A1\_B1) * freq(A2\_B2) - freq(A1\_B2) * freq(A2\_B1)\right)}{\sqrt{freq(A1) * freq(B2) * freq(B1) * freq(B2)}}$$

# Persistence of LD across breeds

| | | *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 | |

*Breed 1*

$$r = \frac{(0.4*0.4 - 0.1*0.1)}{\sqrt{0.5*0.5*0.5*0.5}}$$

# Persistence of LD across breeds

| | | *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 | |

*Breed 1*

$$r = 0.6$$

# Persistence of LD across breeds

| | | *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 | |

*Breed 1*

$$r = 0.6$$

| | | *Marker A* | | |
|---|---|---|---|---|
| | | A1 | A2 | Frequency |
| *Marker B* | B1 | 0.3 | 0.2 | 0.5 |
| | B2 | 0.2 | 0.3 | 0.5 |
| | Frequency | 0.5 | 0.5 | |

*Breed 2*

$$r = 0.2$$

# Persistence of LD across breeds

| | | *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 | |

*Breed 1*

$$r = 0.6$$

| | | *Marker A* | | |
|---|---|---|---|---|
| | | A1 | A2 | Frequency |
| *Marker B* | B1 | 0.2 | 0.3 | 0.5 |
| | B2 | 0.3 | 0.2 | 0.5 |
| | Frequency | 0.5 | 0.5 | |

*Breed 2*

# Persistence of LD across breeds

| | | *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 | |

*Breed 1*

$$r = 0.6$$

| | | *Marker A* | | |
|---|---|---|---|---|
| | | A1 | A2 | Frequency |
| *Marker B* | B1 | 0.2 | 0.3 | 0.5 |
| | B2 | 0.3 | 0.2 | 0.5 |
| | Frequency | 0.5 | 0.5 | |

*Breed 2*

$$r = -0.2$$

# Persistence of LD across breeds

- For marker pairs at a given distance, the correlation between their r in two populations, corr(r1,r2), is equal to correlation of effects of the marker between both populations
  - If this correlation is 1, marker effects are equal in both populations.
  - If this correlation is zero, a marker in population 1 is useless in population 2.
  - A high correlation between r values means that the marker effect persists across the populations.

# Persistence of LD across breeds

- Example

| Marker 1 | Marker 2 | Distance kb | r Breed 1 | r Breed 2 |
|---|---|---|---|---|
| A | B | 20 | 0.8 | 0.7 |
| C | D | 50 | -0.4 | -0.6 |
| E | F | 30 | 0.5 | 0.6 |
| | | | | |
| | Average kb | 33 | corr(r1,r2) | 0.98 |

# Persistence of LD across breeds

- Example

| Marker 1 | Marker 2 | Distance kb | r Breed 1 | r Breed 2 |
|---|---|---|---|---|
| A | B | 20 | 0.8 | 0.7 |
| C | D | 50 | -0.4 | -0.6 |
| E | F | 30 | 0.5 | 0.6 |
| | | | | |
| | Average kb | 33 | corr(r1,r2) | 0.98 |

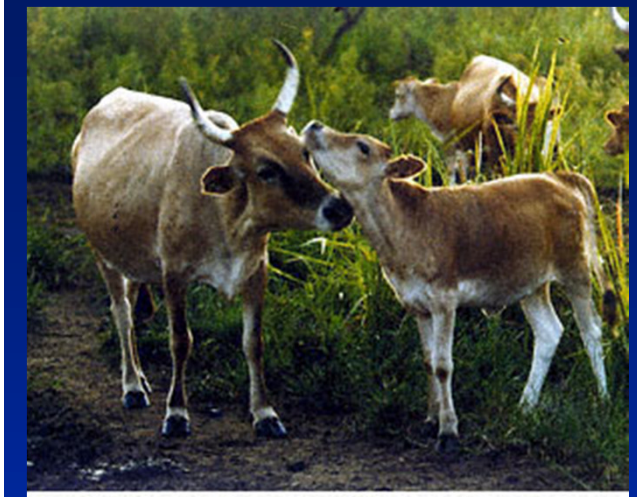
| Marker 1 | Marker 2 | Distance kb | r Breed 1 | r Breed 2 |
|---|---|---|---|---|
| A | B | 500 | 0.4 | 0.2 |
| C | D | 550 | -0.4 | -0.2 |
| E | F | 450 | 0.2 | -0.3 |
| | | | | |
| | Average kb | 500 | corr(r1,r2) | 0.54 |

# The International Bovine Haplotype Map project

- A follow on from the bovine genome sequencing project
- Bovine hap map project aims to characterise LD within and between cattle breeds
- 19 breeds from around the world genotyped for 32 000 Single Nucleotide markers (25 animals from each breeds)

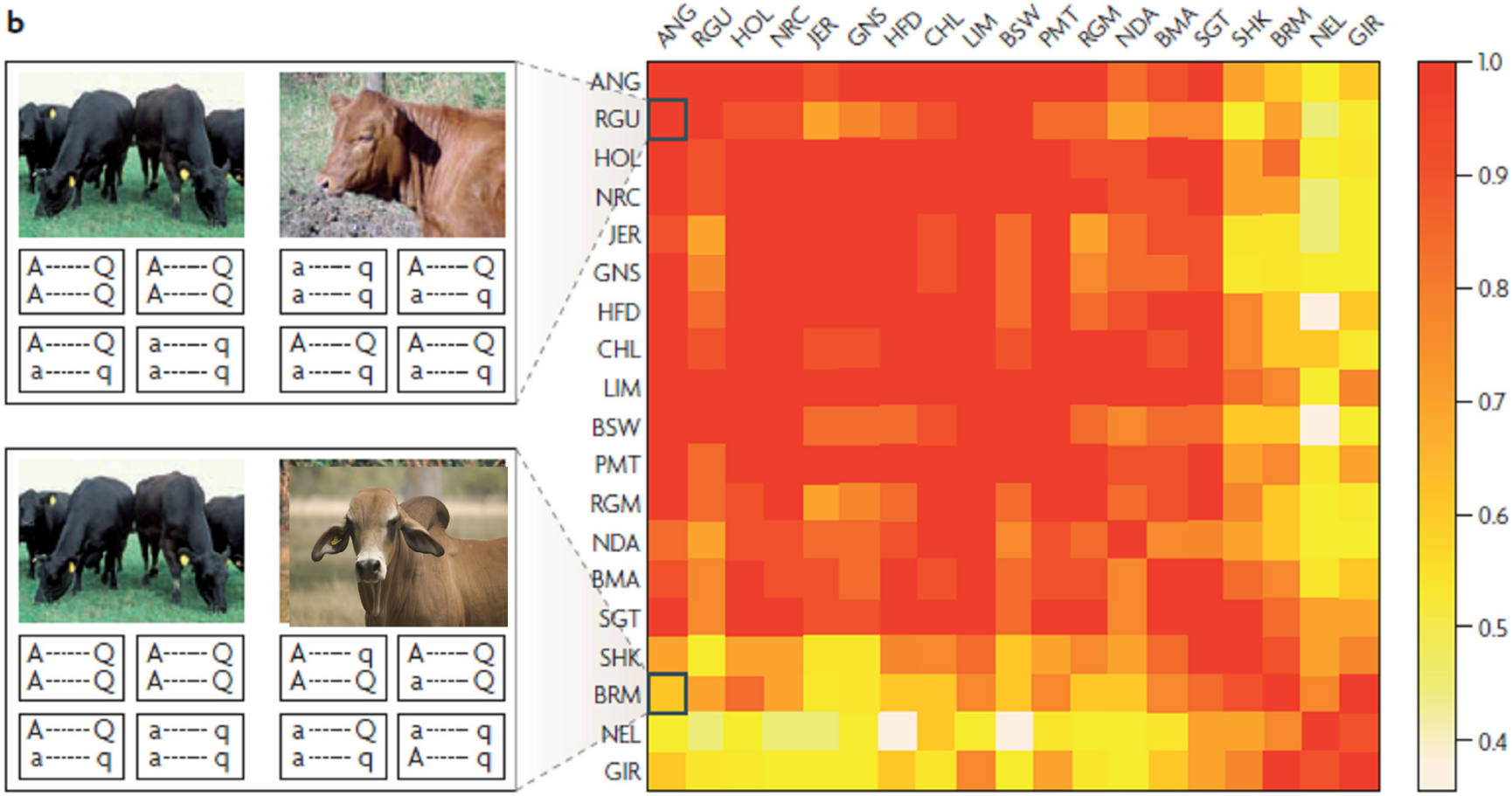
# Breeds sampled….

| Species and Breed | Land of origin | Primary purpose |
|---|---|---|
| *Bos taurus* | | |
| Angus | Scotland | Beef |
| Brown Swiss | Switzerland | Dairy |
| Charolais | France | Beef |
| Guernsey | Channel Islands | Dairy |
| Hereford | UK | Beef |
| Holstein | Netherlands | Dairy |
| Jersey | Channel Islands | Dairy |
| Limousin | France | Beef |
| N'dama | West Africa | Multi-purpose |
| Norwegian Red | Norway | Dairy/Dual purpose |
| Piedmontese | Italy | Beef/ Dual purpose |
| Red Angus | Scotland | Beef |
| Romagnola | Italy | Beef |
| Sheko | Ethiopia | Multi-purpose |
| Bos indicus | | |
| Brahman | USA | Beef |
| Gir | India | Beef |
| Nellore | Brazil | Beef |
| Hybrid | | |
| Beefmaster | USA | Beef |
| Santa Gertrudis | USA | Beef |

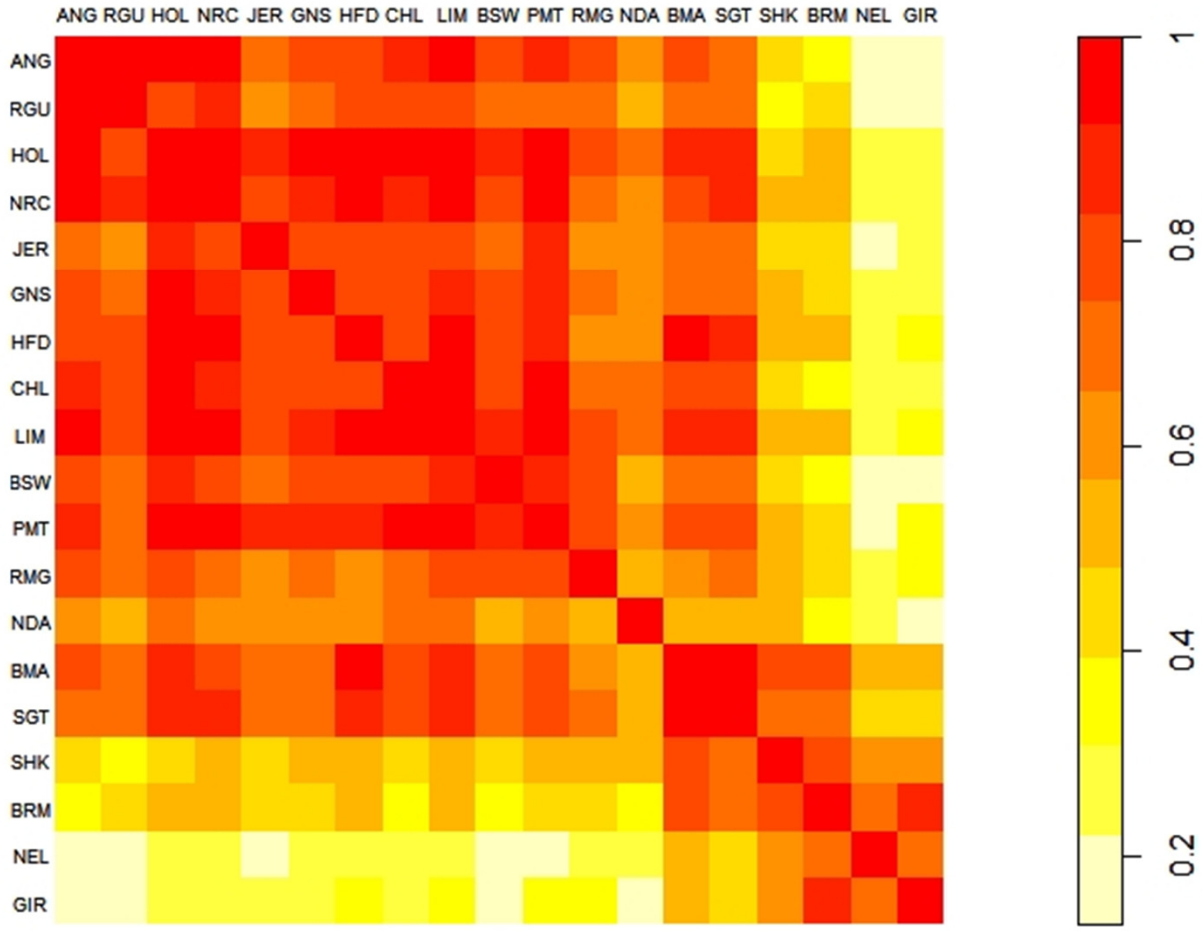


provided by Campagnie J. Van Lancker

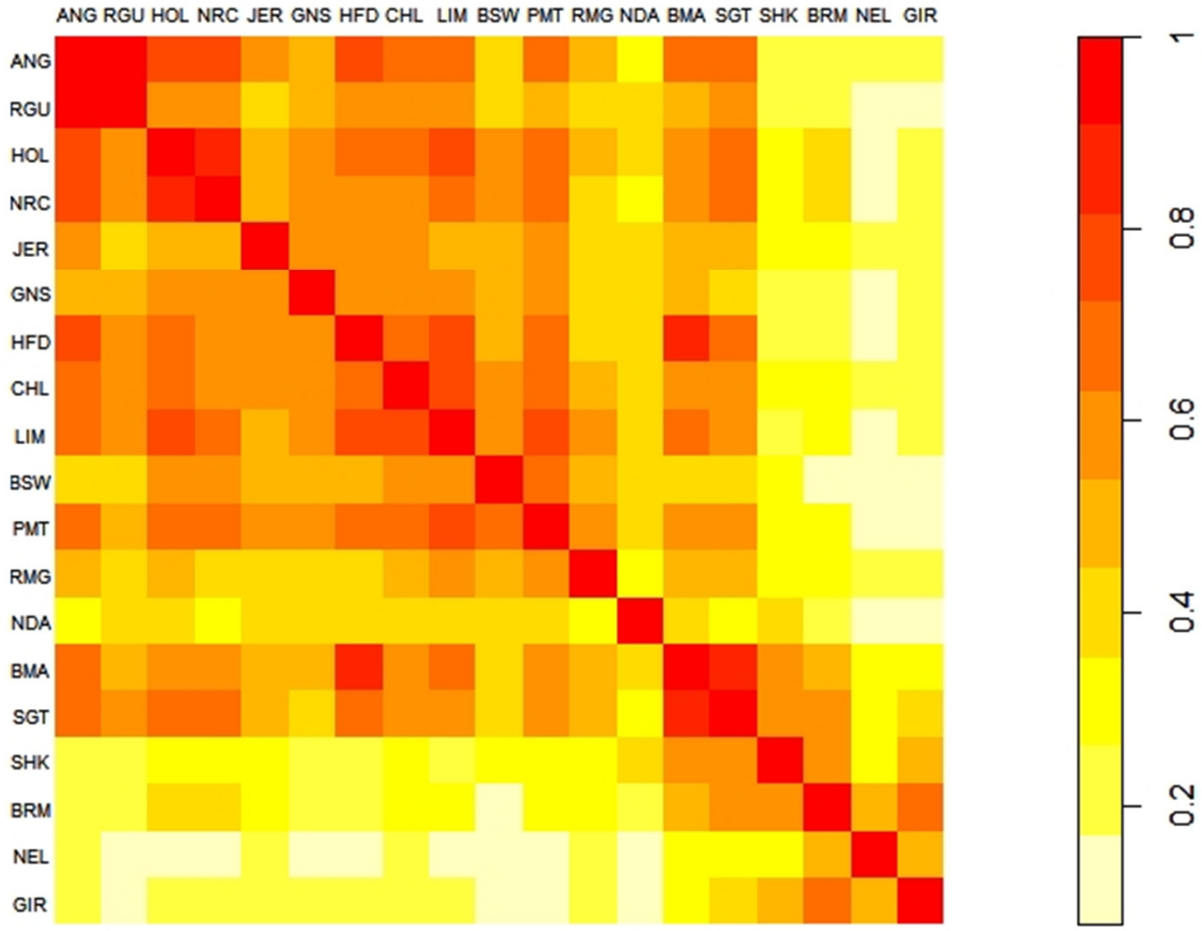
# LD across breeds (10kb)



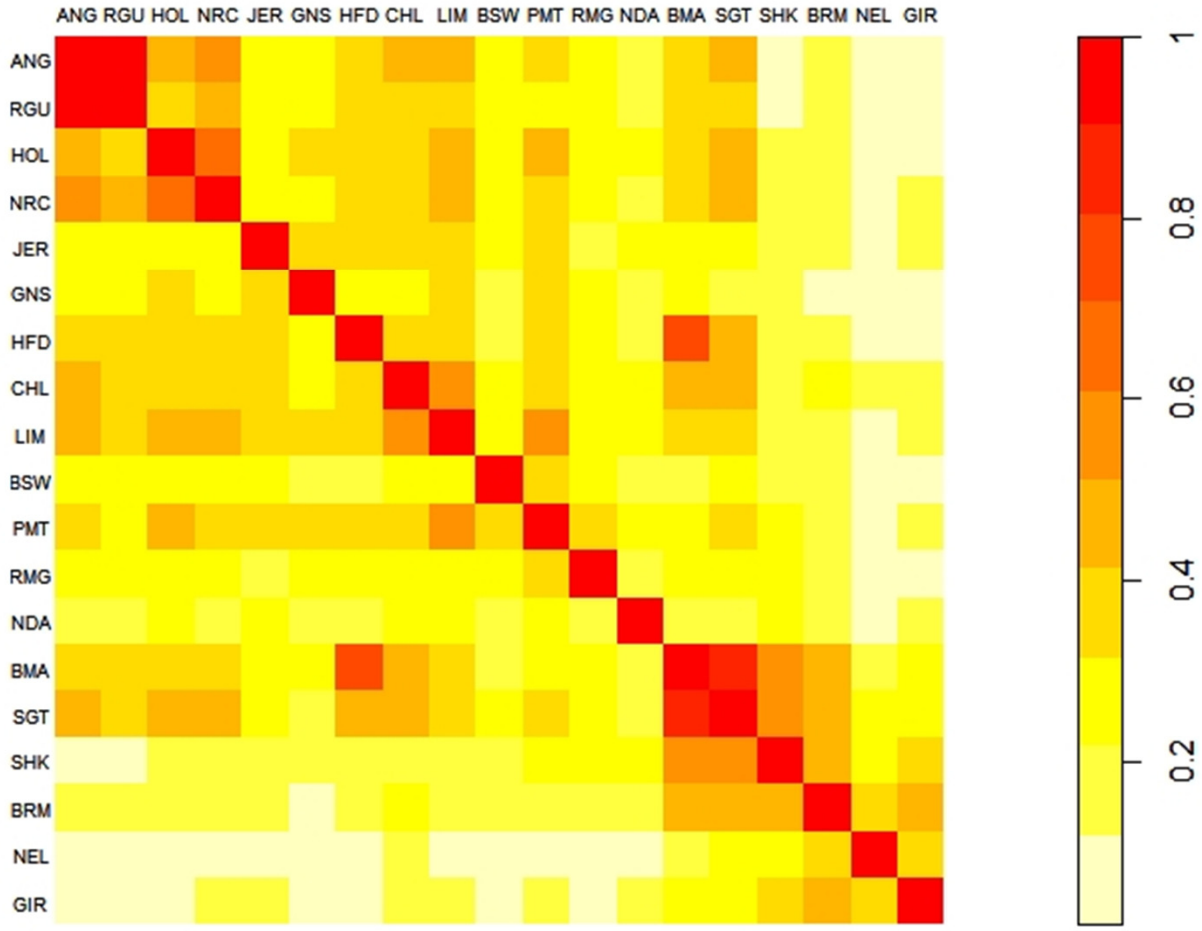
# 10 - 50kb


ANG RGU HOL NRC JER GNS HFD CHL LIM BSW PMT RMG NDA BMA SGT SHK BRM NEL GIR
ANG
RGU
HOL
NRC
JER
GNS
HFD
CHL
LIM
BSW
PMT
RMG
NDA
BMA
SGT
SHK
BRM
NEL
GIR
1
0.8
0.6
0.4
0.2

## 50 - 100kb


ANG RGU HOL NRC JER GNS HFD CHL LIM BSW PMT RMG NDA BMA SGT SHK BRM NEL GIR
ANG
RGU
HOL
NRC
JER
GNS
HFD
CHL
LIM
BSW
PMT
RMG
NDA
BMA
SGT
SHK
BRM
NEL
GIR
1
0.8
0.6
0.4
0.2

# 100 - 250kb


ANG RGU HOL NRC JER GNS HFD CHL LIM BSW PMT RMG NDA BMA SGT SHK BRM NEL GIR
ANG
RGU
HOL
NRC
JER
GNS
HFD
CHL
LIM
BSW
PMT
RMG
NDA
BMA
SGT
SHK
BRM
NEL
GIR
1
0.8
0.6
0.4
0.2

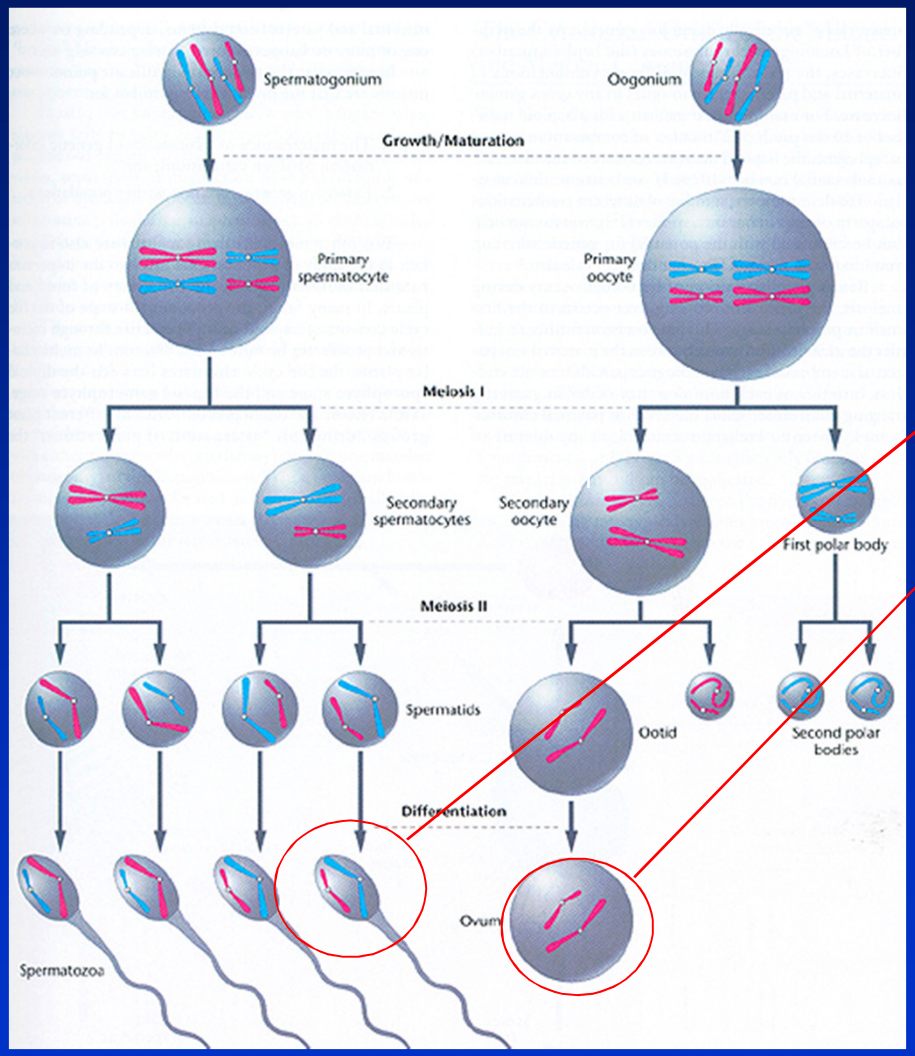
# Persistence of LD across breeds

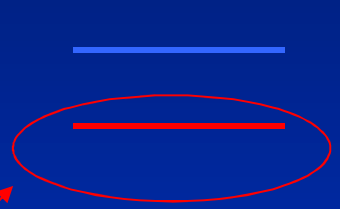
- Recently diverged breeds/lines, good prospects of using a marker found in one line in the other line
- More distantly related breeds, will need very dense marker maps to find markers which can be used across breeds
  - In *Bos taurus* cattle, marker every 10kb = 300,000 markers
- Important in multi breed/multi line populations
  - eg. beef, sheep, pigs
  - Across inbred lines in plant species

# Linkage disequilibrium

- A brief history of QTL mapping
- Measuring linkage disequilibrium
- Causes of LD
- Extent of LD in animals and plants
- The extent of LD between breeds and lines
- Strategies for haplotyping

# Definition of Haplotype



Paternal gamete

Maternal gamete

SNP1 SNP2 SNP3 SNP4

----A—----T—----C-——-G—

# Haplotyping

- LD statistics such as $r^2$ use haplotype frequencies

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

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

- Need to infer haplotypes

# Haplotyping

- In large half sib families
  - which of the sire alleles co-occur in progeny most often
    - Dam haplotypes by subtracting sire haplotype from progeny genotype
- Complex pedigrees
  - Much more difficult, less information per parent, account for missing markers, inbreeding
  - *SimWalk*
- Randomly sampled individuals from population
  - Infer haplotypes from LD information!
  - *PHASE*

# Haplotyping

- PHASE program:
  - Start with group of unphased individuals

| | *Genotypes* |
|---|---|
| *Anim1* | 121122<br>121122 |
| *Anim2* | 122122<br>121122 |
| *Anim3* | 122222<br>121122 |
| *Anim4* | 121222<br>122122 |

# Haplotyping

- PHASE program:
  - Sort haplotypes for unambiguous animals

| | | | |
|---|---|---|---|
| *Anim1* | 121122<br>121122 | → | 121122<br>121122 |
| *Anim2* | 122122<br>121122 | → | 122122<br>121122 |
| *Anim3* | 122222<br>121122 | | |
| *Anim4* | 121222<br>122122 | | |

# Haplotyping

- PHASE program:
  - Add to list of haplotypes in population

| | | | |
|---|---|---|---|
| *Anim1* | 121122<br>121122 | → 121122<br>121122 | |
| *Anim2* | 122122<br>121122 | → 122122<br>121122 | |
| *Anim3* | 122222<br>121122 | | |
| *Anim4* | 121222<br>122122 | | |

*Haplotype list*

121122

122122

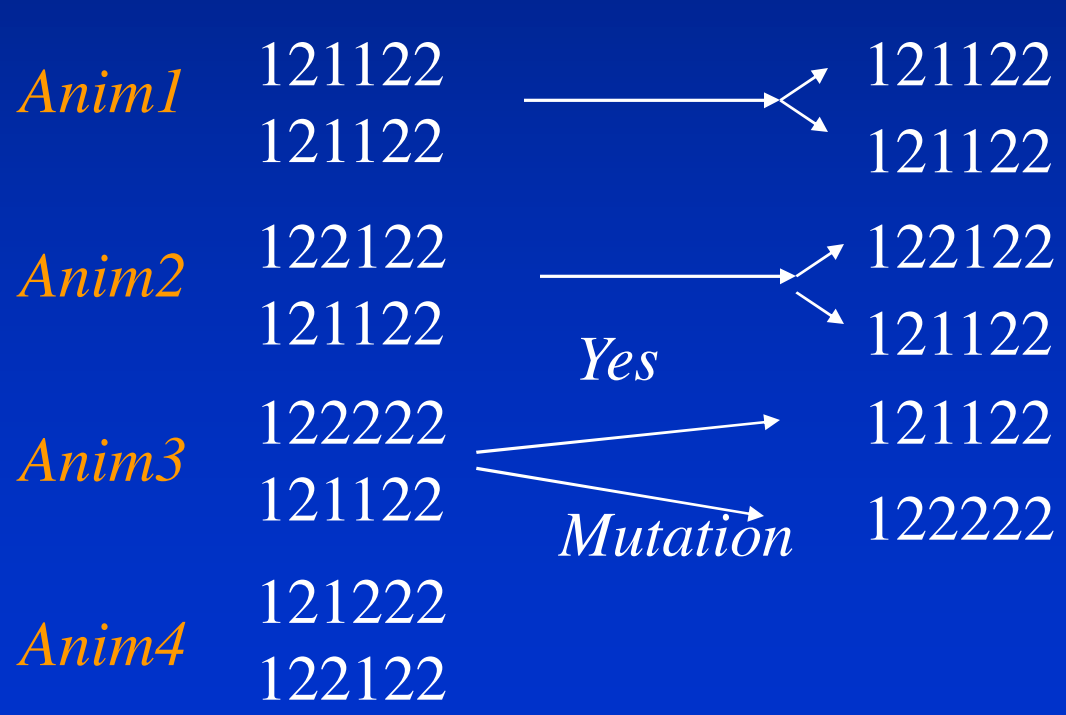
# Haplotyping

- PHASE program:
  - For an ambiguous individual, can haplotypes be same as those in list (most likely=most freq)?

| | | | | *Haplotype list* |
|---|---|---|---|---|
| *Anim1* | 121122<br>121122 | → | 121122<br>121122 | 121122 |
| *Anim2* | 122122<br>121122 | → *Yes* | 122122<br>121122 | 122122 |
| *Anim3* | 122222<br>121122 | → → *No* | 121122 | |
| *Anim4* | 121222<br>122122 | | | |

# Haplotyping

- PHASE program:
  - If no, can we produce haplotype by recombination or mutation (likelihood on basis of length of segment and num markers)



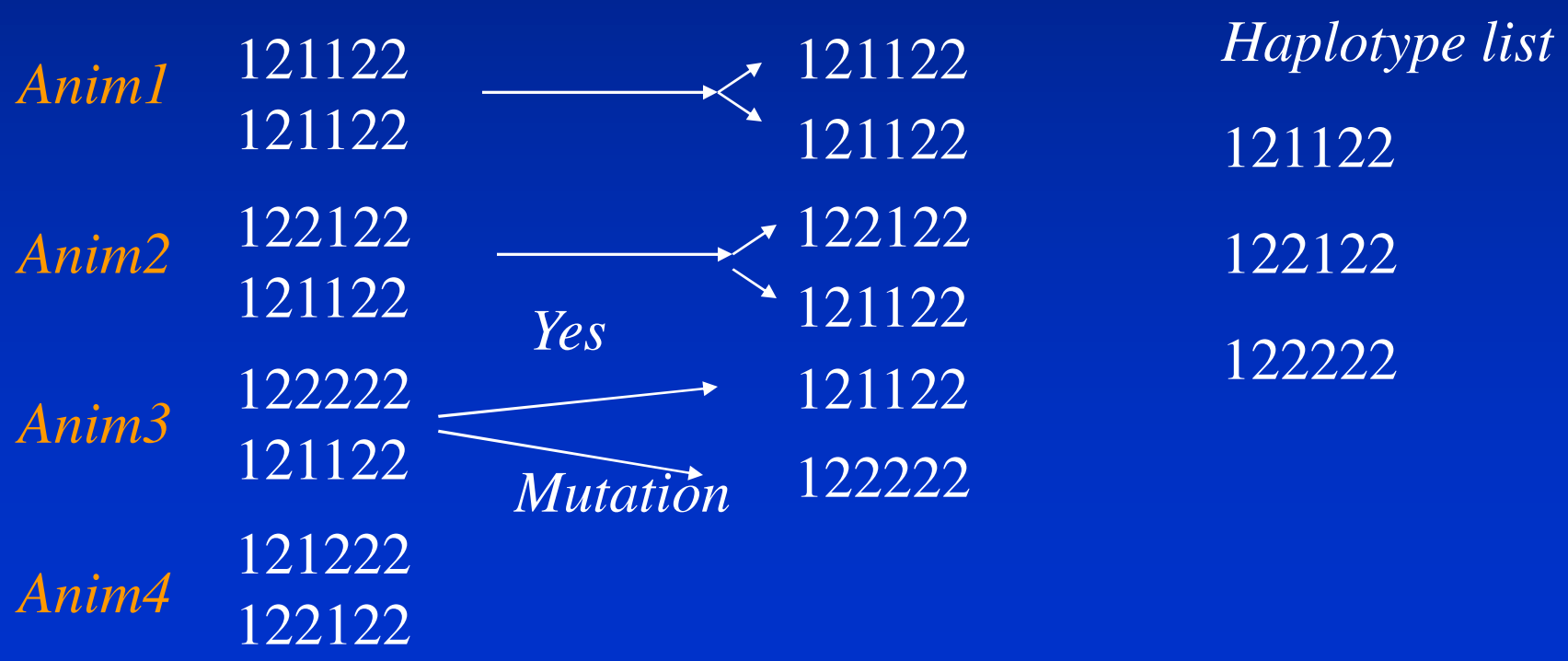

*Haplotype list*

121122

122122

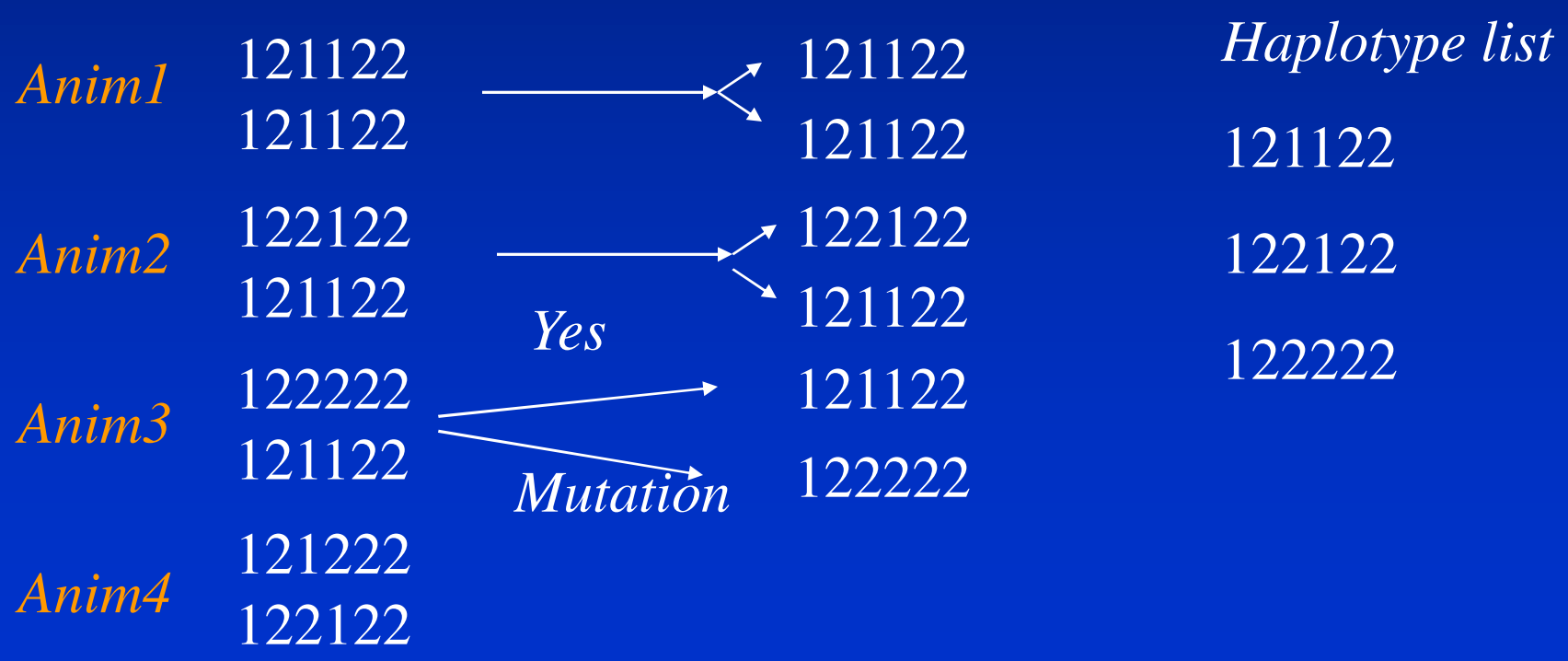
# Haplotyping

- PHASE program:
  - Update list



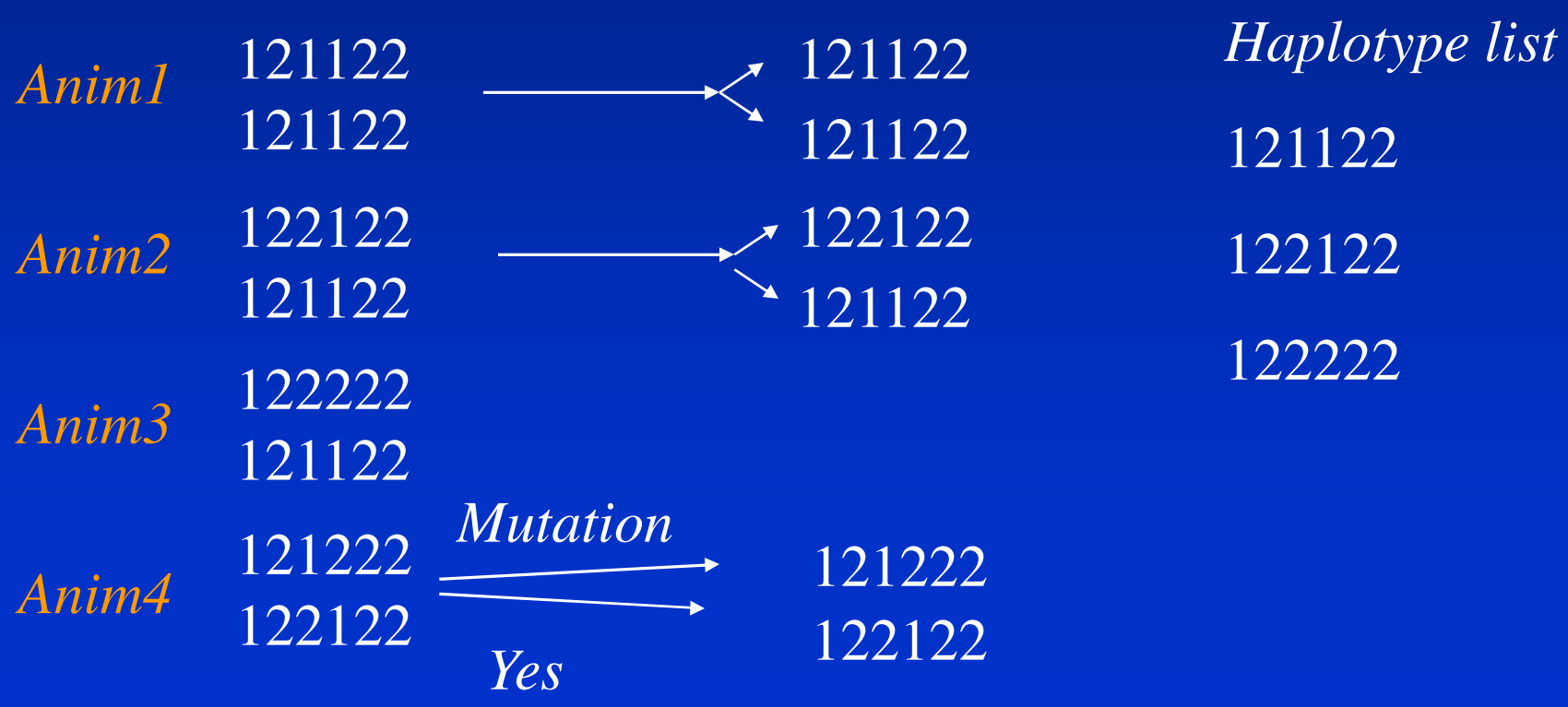
# Haplotyping

- PHASE program:
  - If we randomly choose individual each time, produces Markov Chain



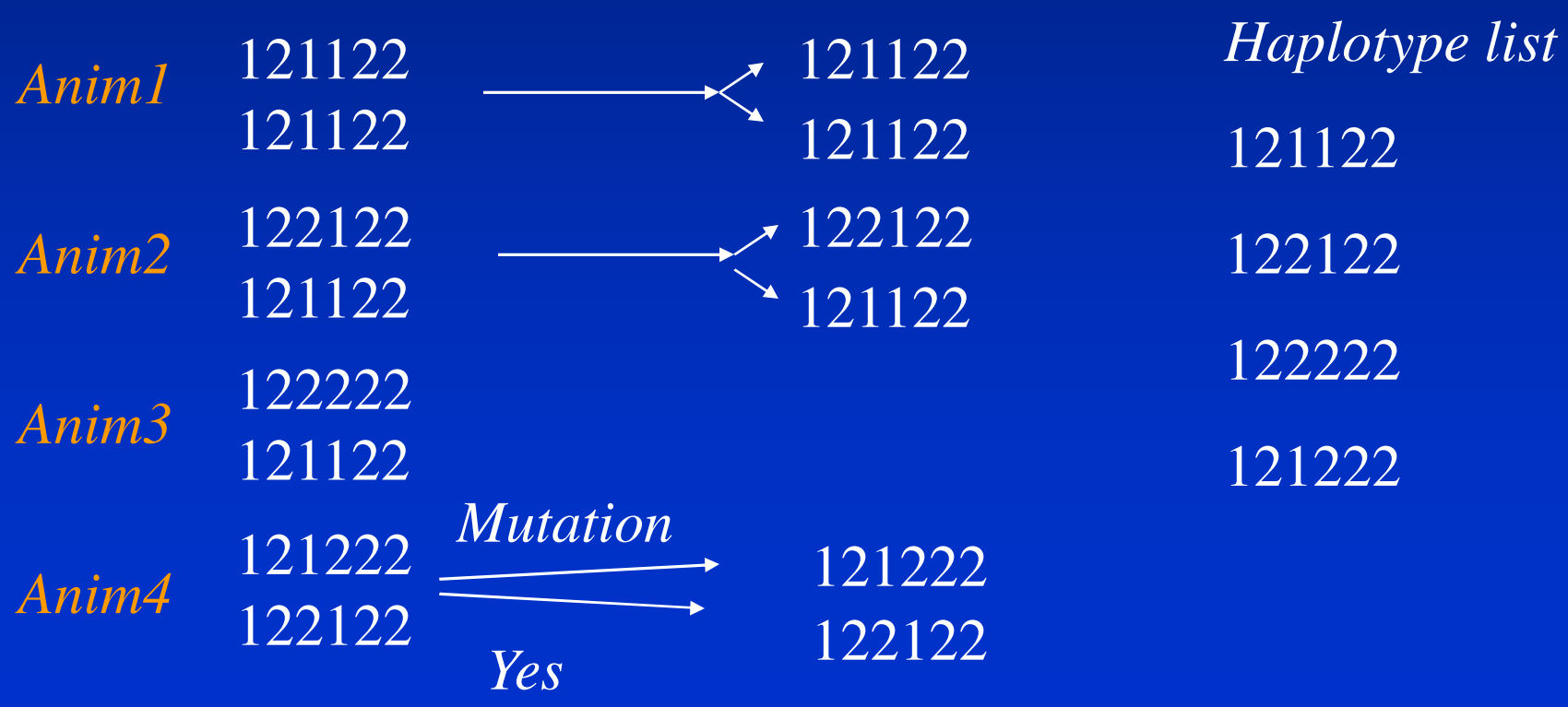
# Haplotyping

- PHASE program:
  - If we randomly choose individual each time, produces Markov Chain



# Haplotyping

- PHASE program:
  - If we randomly choose individual each time, produces Markov Chain



# Haplotyping

- PHASE program
  - After running chain for large number of iterations,
    - End up with most likely haplotypes in the population, haplotype pairs for each animal (with probability attached)
  - Only useful for *very short intervals, dense markers*!
  - But very accurate in this situation
  - Used to construct human hap map, bovine hap map
  - Very good for imputing missing genotypes
- fastPHASE, BEAGLE for large data sets

# Linkage disequilibrium

- Extent of LD in a species determines marker density necessary for LD mapping
- Extent of LD determined by population history
- In cattle, $r^2$~0.3 at 50kb ~ 60 000 markers necessary for genome scan
- Extent of across breed/line LD indicates how close a marker must be to QTL to work across breeds/lines
  - LD persists for ~ 10kb across *Bos Taurus*, 300 000 markers needed?