Linkage Disequilibrium to Genomic Selection
Course overview

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

Means for growth in Atlantic salmon families in Norwegian breeding program
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 estimation
  – 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 ?
The distribution of QTL effects

• From results of QTL mapping experiments

• Two problems
  – no small effects, effects estimated with error
  – Fit a truncated gamma distribution
The distribution of QTL effects

- Many small QTL, few QTL of large effect.
- 100 – 150 QTL sufficient to explain observed variation in quantitative traits in livestock.
The distribution of QTL effects

Proportion of variance accounted for

Pig
Dairy

QTL ranked in order of size
Common variants in the GDF5-UQCC region are associated with variation in human height

Serena Sanna1,2,19, Anne U Jackson1,19, Ramaiah Nagaraja3, Cristen J Willer1, Wei-Min Chen1,4, Lori L Bonnycastle6, Haiqing Shen5, Nicholas Timpson7,8, Guillaume Lettre9, Gianluca Usla2, Peter S Chinas5, Heather M Stringham1, Laura J Scott1, Mariano Dei2, Sandra Laia2, Giuseppe Albai2, Laura Crisponi2, Silvia Naitza2, Kimberly F Doheny10, Elizabeth W Pugh10, Yoav Ben-Shlomo7, Shah Ebrahimi11, Debbie A Lawlor7,8, Richard N Bergman12, Richard M Watanabe12,13, Manuela Ud2, Jaakko Tuomilehto14, Josef Coresh15, Joel N Hirschhorn3, Alan R Shuldiner5,16, David Schlessinger3, Francis S Collins5, George Davey Smith2,8, Eric Boerwinkle17, Antonio Cao2, Michael Boehnke1, Gonçalo R Abecasis1 & Karen L Mohlke19

Identifying genetic variants that influence human height will advance our understanding of skeletal growth and development. Several rare genetic variants have been convincingly and reproducibly associated with height in mendelian syndromes, and common variants in the transcription factor gene HMGA2 are associated with variation in height in the general population1. Here we report genome-wide association analyses, using genotyped and imputed markers, of 6,669 individuals from Finland and Sardinia, and follow-up analyses in an additional 28,801 individuals. We show that common variants in the osteoarthritis-associated locus2 GDF5-UQCC contribute to variation in height with an estimated additive effect of 0.44 cm (overall $P < 10^{-15}$). Our results indicate that there may be a link between the genetic basis of height and osteoarthritis, potentially mediated through alterations in bone growth and development.
Letter abstract

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

Common variants in the GDF5-UQCC region are associated with variation in human height
Serena Sanna,1,2,19 Anne U Jackson,1,19 Ramaiah Nagaraja,3 Cristen J Willer,4 Wei-Min Chen,1,4, Lori L Bonnycastle,6 Haiqing Shen,5 Nicholas Timpson,2,8, Guillaume Lettre,9 Gianluca Ucata,2 Peter S Chinas,3 Heather M Stringham,4 Laura J Scott,4 Mariano Dej,2, Sandra Li,2 Giuseppe Albai,2 Laura Crisponi,2 Silvia Naitza,2 Kimberly F Doheny,10 Elizabeth W Pugh,10 Yoav Ben-Shlomo,7, Shah Ebrahim,11, Debbie A Lawlor,7,8, Richard N Bergman,12, Richard M Watanabe,12,13, Manuela Uda,2, Jaakko Tuomilehto,14, Josef Coresh,15, Joel N Hirschhorn,5, Alan R Shuldiner,5,16 David Schlessinger,3, Francis S Collins,5, George Davey Smith,2,8, Eric Boerwinkle,17, Antonio Cao,2, Michael Boehnke,1, Gonçalo R Abecasis1 & Karen L Mohlke10

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

< 1% of phenotypic variance!
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

Diagram:

- **Marker 1**
  - Dad: A, C
  - Kid 1: A, Q

- **QTL**
  - Kid 2: C, q
  - Dad: Q, q
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

Sire

Marker allele 172

QTL +ve

Marker allele 184

QTL -ve

Progeny inheriting 172 allele for the marker

Progeny inheriting 184 allele for the marker
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)
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)
- ~10 mill SNPs
- ~7 mill SNPs with minor allele >5%
- ~100,000-300,000 cSNPs
- ~50,000 nonsynonymous cSNPs -> change protein structure
The Revolution

• 100 000s of SNPs reported for cattle, chicken, pig
• Sheep, Atlantic Salmon on the way
• Plants?
The Revolution

• Can we use SNP information to greatly accelerate the application of marker assisted selection in the livestock industries?
The Revolution

• Can we use SNP information to greatly accelerate the application of marker assisted selection in the livestock industries?
  – Omit linkage mapping
  – Straight to genome wide LD mapping
  – Breeding values directly from markers?
• Genomic selection
Aim

• Provide you with the tools to use high density SNP 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
- Strategies for haplotyping
Definitions of LD

• Why do we need to define and measure LD?
• Both genomic selection and LD mapping require 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

<table>
<thead>
<tr>
<th>Marker A</th>
<th>A1</th>
<th>A2</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Marker B</td>
<td>B1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>B2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Frequency</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
</tr>
</tbody>
</table>
Definitions of LD

Linkage equilibrium........

<table>
<thead>
<tr>
<th>Marker B</th>
<th>Marker A</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>B1</td>
<td>A1</td>
<td>0.25</td>
</tr>
<tr>
<td></td>
<td>A2</td>
<td>0.25</td>
</tr>
<tr>
<td></td>
<td>Frequency</td>
<td>0.5</td>
</tr>
<tr>
<td>B2</td>
<td>A1</td>
<td>0.25</td>
</tr>
<tr>
<td></td>
<td>A2</td>
<td>0.25</td>
</tr>
<tr>
<td></td>
<td>Frequency</td>
<td>0.5</td>
</tr>
<tr>
<td>Frequency</td>
<td></td>
<td>0.5</td>
</tr>
</tbody>
</table>
**Definitions of LD**

Linkage disequilibrium

<table>
<thead>
<tr>
<th>Marker B</th>
<th>Markers</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>B1</td>
<td>A1</td>
<td>0.4</td>
</tr>
<tr>
<td></td>
<td>A2</td>
<td>0.1</td>
</tr>
<tr>
<td></td>
<td>Frequency</td>
<td>0.5</td>
</tr>
<tr>
<td>B2</td>
<td>A1</td>
<td>0.1</td>
</tr>
<tr>
<td></td>
<td>A2</td>
<td>0.4</td>
</tr>
<tr>
<td></td>
<td>Frequency</td>
<td>0.5</td>
</tr>
<tr>
<td>Frequency</td>
<td></td>
<td>0.5</td>
</tr>
</tbody>
</table>
Definitions of LD

Linkage disequilibrium........

<table>
<thead>
<tr>
<th>Marker B</th>
<th>Marker A</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>B1</td>
<td>A1</td>
<td>0.4</td>
</tr>
<tr>
<td></td>
<td>A2</td>
<td>0.1</td>
</tr>
<tr>
<td></td>
<td>Frequency</td>
<td>0.5</td>
</tr>
<tr>
<td>B2</td>
<td>A1</td>
<td>0.1</td>
</tr>
<tr>
<td></td>
<td>A2</td>
<td>0.4</td>
</tr>
<tr>
<td></td>
<td>Frequency</td>
<td>0.5</td>
</tr>
<tr>
<td>Frequency</td>
<td></td>
<td>0.5</td>
</tr>
</tbody>
</table>

within a sire family
sire haplotypes A1_B1, A2_B2
Definitions of LD

Linkage disequilibrium........

within a population
unrelated animals selected at random:
Definitions of LD

• In fact, LD required for both linkage and linkage disequilibrium mapping

• Difference is
  – linkage analysis mapping considers the LD that exists within families
    • extends for 10s of cM
    • broken down after only a few generations
  – LD mapping requires a marker allele to be in LD with a QTL allele across the whole population
    • association must have persisted across multiple generations to be a property of the population
    • so marker and QTL must be very closely linked
• Linkage between marker and QTL

Large difference indicates presence of important gene

Large difference indicates presence of important gene
• Linkage disequilibrium between marker and QTL
Definitions of LD

Linkage disequilibrium........

<table>
<thead>
<tr>
<th>Marker A</th>
<th>A1</th>
<th>A2</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>B1</td>
<td>0.4</td>
<td>0.1</td>
<td>0.5</td>
</tr>
<tr>
<td>B2</td>
<td>0.1</td>
<td>0.4</td>
<td>0.5</td>
</tr>
<tr>
<td>Frequency</td>
<td>0.5</td>
<td>0.5</td>
<td></td>
</tr>
</tbody>
</table>

\[ D = \text{freq}(A_1B_1) \times \text{freq}(A_2B_2) - \text{freq}(A_1B_2) \times \text{freq}(A_2B_1) \]

\[ = 0.4 \times 0.4 - 0.1 \times 0.1 \]

\[ = 0.15 \]
Definitions of LD

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

\[ D = \text{freq}(A_1,B_1) \times \text{freq}(A_2,B_2) - \text{freq}(A_1,B_2) \times \text{freq}(A_2,B_1) \]
- highly dependent on allele frequencies
- not suitable for comparing LD at different sites

\[ r^2 = \frac{D^2}{[\text{freq}(A_1) \times \text{freq}(A_2) \times \text{freq}(B_1) \times \text{freq}(B_2)]} \]
### Definitions of LD

Linkage disequilibrium........

<table>
<thead>
<tr>
<th>Marker B</th>
<th>Marker A</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A1</td>
<td>A2</td>
</tr>
<tr>
<td>B1</td>
<td>0.4</td>
<td>0.1</td>
</tr>
<tr>
<td>B2</td>
<td>0.1</td>
<td>0.4</td>
</tr>
<tr>
<td>Frequency</td>
<td>0.5</td>
<td>0.5</td>
</tr>
</tbody>
</table>

\[
D = 0.15
\]

\[
r^2 = \frac{D^2}{[freq(A1) \times freq(A2) \times freq(B1) \times freq(B2)]}
\]

\[
r^2 = 0.15^2/[0.5 \times 0.5 \times 0.5 \times 0.5]
\]

\[
= 0.36
\]
Definitions of LD

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

\[ D = \text{freq}(A_1 \_B_1) \times \text{freq}(A_2 \_B_2) - \text{freq}(A_1 \_B_2) \times \text{freq}(A_2 \_B_1) \]
- highly dependent on allele frequencies
  - not suitable for comparing LD at different sites

\[ r^2 = \frac{D^2}{[\text{freq}(A_1) \times \text{freq}(A_2) \times \text{freq}(B_1) \times \text{freq}(B_2)]} \]

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 $200\text{kg}^2$, and $r^2$ between marker and QTL is 0.2, variation observed at the marker is $40\text{kg}^2$. 
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

- If you are using microsatellites, need a multi-allele equivalent
- Use $\chi^2$ (Zhao et al. 2005)
Definitions of LD

• Another LD statistic is $D'$
  – $|D|/D_{max}$
  – Where
    • $D_{max}$
      – $= \min[\text{freq}(A1)\times\text{freq}(B2), (1-\text{freq}(A2))(1-\text{freq}(B1))]$ if $D > 0$, else
      – $= \min[\text{freq}(A1)(1-\text{freq}(B1),(1-(\text{freq}(A2))\times\text{freq}(B2))]$ if $D < 0$.
  – But what does it mean?
  – Biased upward with low allele frequencies
  – Overestimates $r^2$
Definitions of LD

• Another LD statistic is $D'$
  – $|D|/D_{max}$
  – Where
    • $D_{max}$
      – if $D>0$, $D_{max} = \min[\text{freq}(A1)\times\text{freq}(B2), (1-\text{freq}(A2))(1-\text{freq}(B1))]$
      – if $D<0$, else $D_{max} = \min[\text{freq}(A1)(1-\text{freq}(B1)); (1-(\text{freq}(A2))\times\text{freq}(B2))]$
    – But what does it mean?
    – Biased upward with low allele frequencies
    – Overestimates $r^2$
Definitions of LD

• Multi-locus measures of LD
  – $r^2$ is useful, easy to calculate and very widely used
    • and equivalents for loci with multiple alleles exist
  – But, only considers two loci at a time
    • cannot extract LD information available from multiple loci
    • not particularly intuitive with regards to the causes of LD
Definitions of LD

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

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

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

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

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

• chromosome segment homozygosity (CSH) = Pr(Two chromosome segments randomly drawn from the population are derived from a common ancestor)
Definitions of LD

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

- chromosome segment homozygosity (CSH) = Pr(Two chromosome segments randomly drawn from the population are derived from a common ancestor)
Definitions of LD

- Haplotype homozygosity = CSH + Identical chance (and not IBD)
- For two loci
  \[ HH = CSH + (\text{Hom}_A - CSH)(\text{Hom}_B - CSH)/(1 - CSH) \]
- Derivation for multiple loci similar, but more complex
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
- 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

Parental Lines

F1

A  Q  B  A  Q  B  A  Q  B  A  Q  B
a  q  b  a  q  b  a  q  b  a  q  b
• F2 design

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
Generation 1
A____q   A____q
A____q   a_____q
a_____q   a_____q

Mutation

Generation 2

Generation 3
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
a_____q  A_____Q
A_____Q  a_____q
a_____q  A_____q

Generation 3

Mutation
Selection
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

• Size of conserved chunks depends on effective population size
Causes of LD

- Predicting LD with finite population size
- \( E(r^2) \) and \( E(CSH) = \frac{1}{4Nc+1} \)
  - \( N \) = effective population size
  - \( c \) = length of chromosome segment

![Diagram showing linkage disequilibrium (CSH) vs. length of chromosome segment (cM) for different effective population sizes (Ne=100 and Ne=1000).]
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

1000 to 100

A

B
Causes of LD

- \( E(r^2) = \frac{1}{(4N_t c + 1)} \)
- Where \( t = \frac{1}{(2c)} \) 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
- Strategies for haplotyping
Extent of LD in humans and livestock

Humans........(Tenesa et al. 2007)

r^2 decay against recombination distance

Distance (kb)

Mean r^2

Human series compare across multiple populations.
Extent of LD in humans and livestock

And cattle......
Extent of LD in humans and livestock

Population size humans and cattle.....
Implications?

• In Holsteins, need a marker approximately every 200kb to get average $r^2$ of 0.2 between marker and QTL (eg. 100kb marker-QTL).
Implications?

- In Holsteins, need a marker approximately every 200kb to get average $r^2$ of 0.2 between marker and QTL (eg. 100kb 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 200kb to get average $r^2$ of 0.2 between marker and QTL (eg. 100kb 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
  - 30,000 evenly spaced markers to capture every QTL in a genome scan
  - Markers not evenly spaced ~ 60 000 markers required
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

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

<table>
<thead>
<tr>
<th>Marker B</th>
<th>Marker A</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>B1</td>
<td>A1 0.4</td>
<td>A2 0.1</td>
</tr>
<tr>
<td></td>
<td>A1 0.1</td>
<td>A2 0.4</td>
</tr>
<tr>
<td>Frequency</td>
<td>0.5</td>
<td>0.5</td>
</tr>
</tbody>
</table>

**Breed 1**

\[
\begin{align*}
\text{Pearson correlation coefficient } r &= \frac{(\text{freq}(A1 \_ B1) \times \text{freq}(A2 \_ B2) - \text{freq}(A1 \_ B2) \times \text{freq}(A2 \_ B1))}{\sqrt{\text{freq}(A1) \times \text{freq}(B2) \times \text{freq}(B1) \times \text{freq}(B2)}}
\end{align*}
\]
**Persistence of LD across breeds**

<table>
<thead>
<tr>
<th>Marker B</th>
<th>B1</th>
<th>B2</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Marker A</td>
<td>A1</td>
<td>A2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.4</td>
<td>0.1</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td>0.1</td>
<td>0.4</td>
<td>0.5</td>
</tr>
<tr>
<td>Frequency</td>
<td>0.5</td>
<td>0.5</td>
<td></td>
</tr>
</tbody>
</table>

\[
r = \frac{(0.4 \times 0.4 - 0.1 \times 0.1)}{\sqrt{0.5 \times 0.5 \times 0.5 \times 0.5}}
\]
Persistence of LD across breeds

### Marker A

<table>
<thead>
<tr>
<th>Marker B</th>
<th>B1</th>
<th>B2</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>0.4</td>
<td>0.1</td>
<td>0.5</td>
</tr>
<tr>
<td>A2</td>
<td>0.1</td>
<td>0.4</td>
<td>0.5</td>
</tr>
<tr>
<td>Frequency</td>
<td>0.5</td>
<td>0.5</td>
<td></td>
</tr>
</tbody>
</table>

**Breed 1**

\[ r = 0.6 \]
# Persistence of LD across breeds

![Persistence of LD across breeds](image)

## Breast 1

<table>
<thead>
<tr>
<th>Marker B</th>
<th>A1</th>
<th>A2</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>B1</td>
<td>0.4</td>
<td>0.1</td>
<td>0.5</td>
</tr>
<tr>
<td>B2</td>
<td>0.1</td>
<td>0.4</td>
<td>0.5</td>
</tr>
<tr>
<td>Frequency</td>
<td>0.5</td>
<td>0.5</td>
<td></td>
</tr>
</tbody>
</table>

\[ r = 0.6 \]

## Breast 2

<table>
<thead>
<tr>
<th>Marker B</th>
<th>A1</th>
<th>A2</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>B1</td>
<td>0.3</td>
<td>0.2</td>
<td>0.5</td>
</tr>
<tr>
<td>B2</td>
<td>0.2</td>
<td>0.3</td>
<td>0.5</td>
</tr>
<tr>
<td>Frequency</td>
<td>0.5</td>
<td>0.5</td>
<td></td>
</tr>
</tbody>
</table>

\[ r = 0.2 \]
## Persistence of LD across breeds

### Breed 1

<table>
<thead>
<tr>
<th>Marker B</th>
<th>A1</th>
<th>A2</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>B1</td>
<td>0.4</td>
<td>0.1</td>
<td>0.5</td>
</tr>
<tr>
<td>B2</td>
<td>0.1</td>
<td>0.4</td>
<td>0.5</td>
</tr>
<tr>
<td>Frequency</td>
<td>0.5</td>
<td>0.5</td>
<td></td>
</tr>
</tbody>
</table>

### Breed 2

<table>
<thead>
<tr>
<th>Marker B</th>
<th>A1</th>
<th>A2</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>B1</td>
<td>0.2</td>
<td>0.3</td>
<td>0.5</td>
</tr>
<tr>
<td>B2</td>
<td>0.3</td>
<td>0.2</td>
<td>0.5</td>
</tr>
<tr>
<td>Frequency</td>
<td>0.5</td>
<td>0.5</td>
<td></td>
</tr>
</tbody>
</table>

$r = 0.6$
## Persistence of LD across breeds

<table>
<thead>
<tr>
<th>Marker B</th>
<th>Marker A</th>
<th>Breed 1</th>
<th>Breed 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A1</td>
<td>A2</td>
<td>A1</td>
</tr>
<tr>
<td>B1</td>
<td>0.4</td>
<td>0.1</td>
<td>0.2</td>
</tr>
<tr>
<td>B2</td>
<td>0.1</td>
<td>0.4</td>
<td>0.3</td>
</tr>
<tr>
<td>Frequency</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
</tr>
</tbody>
</table>

$r = 0.6$

$r = -0.2$
Persistence of LD across breeds

- For marker pairs at a given distance, the correlation between their $r$ in two populations, $\text{corr}(r_1, r_2)$, 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

<table>
<thead>
<tr>
<th>Marker 1</th>
<th>Marker 2</th>
<th>Distance kb</th>
<th>r Breed 1</th>
<th>r Breed 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>B</td>
<td>20</td>
<td>0.8</td>
<td>0.7</td>
</tr>
<tr>
<td>C</td>
<td>D</td>
<td>50</td>
<td>-0.4</td>
<td>-0.6</td>
</tr>
<tr>
<td>E</td>
<td>F</td>
<td>30</td>
<td>0.5</td>
<td>0.6</td>
</tr>
</tbody>
</table>

Average kb 33  corr(r1,r2) 0.98
### Persistence of LD across breeds

**Example**

<table>
<thead>
<tr>
<th>Marker 1</th>
<th>Marker 2</th>
<th>Distance kb</th>
<th>r Breed 1</th>
<th>r Breed 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>B</td>
<td>20</td>
<td>0.8</td>
<td>0.7</td>
</tr>
<tr>
<td>C</td>
<td>D</td>
<td>50</td>
<td>-0.4</td>
<td>-0.6</td>
</tr>
<tr>
<td>E</td>
<td>F</td>
<td>30</td>
<td>0.5</td>
<td>0.6</td>
</tr>
</tbody>
</table>

Average kb: 33  corr(r1,r2): 0.98

<table>
<thead>
<tr>
<th>Marker 1</th>
<th>Marker 2</th>
<th>Distance kb</th>
<th>r Breed 1</th>
<th>r Breed 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>B</td>
<td>500</td>
<td>0.4</td>
<td>0.2</td>
</tr>
<tr>
<td>C</td>
<td>D</td>
<td>550</td>
<td>-0.4</td>
<td>-0.2</td>
</tr>
<tr>
<td>E</td>
<td>F</td>
<td>450</td>
<td>0.2</td>
<td>-0.3</td>
</tr>
</tbody>
</table>

Average kb: 500  corr(r1,r2): 0.54
Experiment

• Beef cattle
  - 384 Angus animals chosen for genotyping from Trangie net feed intake selection lines
  - genotyped for 10,000 SNPs

• Dairy Cattle
  - 384 Holstein-Friesian dairy bulls selected from Australian dairy bull population
  - genotyped for 10,000 SNPs
Holstein-Angus example

Marker spacing 10kb-50kb

\[ y = 0.8652x - 0.021 \]

\[ R^2 = 0.6175 \]
Holstein-Angus example

Marker spacing 10kb-50kb

![Dairy data r vs Beef data r for marker spacing 10kb-50kb]

Marker spacing 1000kb-2000kb

![Dairy data r vs Beef data r for marker spacing 1000kb-2000kb]
LD across breeds

Correlation of r values

Australian Holstein, Australian Angus
Dutch black and white bulls 95-97, Dutch red and white bulls
Dutch black and white bulls 95-97, Australian Holstein bulls
Dutch black and white bulls <1995, Dutch black and white calves
Australian bulls < 1995, Australian bulls >=1995
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
- Important in multi breed populations – eg. beef, sheep, pigs
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
- 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

\[
D = \text{freq}(A_1B_1) \cdot \text{freq}(A_2B_2) - \text{freq}(A_1B_2) \cdot \text{freq}(A_2B_1)
\]

\[
r^2 = \frac{D^2}{\text{freq}(A_1) \cdot \text{freq}(A_2) \cdot \text{freq}(B_1) \cdot \text{freq}(B_2)}
\]

• 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
  121122
  121122
  122122
  122122
  121122
  121122
  122222
  121122
  121222
  121222
  122122
  122122
Haplotyping

• PHASE program:
  – Sort haplotypes for unambiguous animals

121122  121122  121122
121122  121122
122122  122122
121122  121122
122222
121122
121222
122122
Haplotyping

• PHASE program:
  – Add to list of haplotypes in population

121122 121122 122122 121122 121122 122222 121122

121122
121122
122122
121122
122122
121122
122222
121122
121222
122122

Haplotyping

Haplotype list

121122 121122 122122 121122 121122

122122
121122
122222
121122
121222
122122

Haplotype list
Haplotyping

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

121122 121122 121222 122122
122122 121122 122222 121122
121122 122122
122122 121122
121122 121122
122222 121122
121122 121122
121222
122122

Haplotype list
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 121122 121122
121122 121122 121122
122122 122122 122122
121122 121122 122122
122222 121122 122222
121122 122222 122222
121222 122122 122122
Haplotyping

• PHASE program:
  – Update list

121122 121122 121222 122122
122122 121122 122222 121122
121122 121122

Haplotype list

121122 121122 122122 122122
121122 121122
122222 121122
121122

Mutation

Yes

122222
Haplotyping

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

121122 121122 121122 121122
121122 121122 122122 121122
122122 122122 122122 122122
121122 121122 121122 122222
122222 122222 122222 122222
121122 121122 121122 122222
121222 121222 121222 121222
122122 122122 122122 122122

Haplotype list

Yes

Mutation
**Haplotyping**

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

```
121122 121122 121122 121122
121122 121122 121122 121122
122122 122122 122122 122122
121122 121122 121122 121122
122222 122222 122222 122222
121122 121122 121122 121122
122222 122222 122222 122222
```

*Haplotype list*
Haplotyping

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

Haplotype list

\[
\begin{align*}
121122 & \quad 121122 \\
122122 & \quad 122122 \\
121122 & \quad 121122 \\
122222 & \quad 122222 \\
121122 & \quad 121222 \\
121222 & \quad 121222 \\
122122 & \quad 122122 \\
121222 & \quad 122122
\end{align*}
\]
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
Linkage disequilibrium

- Extent of LD in a species determines marker density necessary for LD mapping
- Extent of LD determined by population history
- In cattle, $r^2 \sim 0.2$ at 100kb $\sim 60\,000$ 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