Genomic Selection

Ben Hayes and Hans Daetwyler

February 2015
Armidale, Australia

1. Linkage disequilibrium in livestock populations

1.1 Models of quantitative trait variation

Many economically important traits in livestock, aquaculture and plant production are quantitative, that is they show continuous distributions. In attempting to explain the genetic variation observed in such traits, two models have been proposed, the infinitesimal model and the finite loci model. The *infinitesimal model* assumes that traits are determined by an infinite number of unlinked and additive loci, each with an infinitesimally small effect (Fischer 1918). This model has been exceptionally valuable for animal breeding, and forms the basis for breeding value estimation theory (e.g. Henderson 1984).

However, the existence of a finite amount of genetically inherited material (the genome) and the revelation that there are perhaps a total of only around 20 000 genes or loci in the genome (Ewing & Green 2000), means that there is must be some *finite number of loci* underlying the variation in quantitative traits. In fact there is increasing evidence that the distribution of the effect of these loci on quantitative traits is such that there are a few genes with large effect, and many of small effect. This genetic architecture is observed for traits as diverse as bristle number in *Drosophila*, height in humans, yield in rice, percentage of oil in maize kernels, and milk production in dairy cattle (Shrimpton & Robertson 1988, Lango-Allen et 2010, Huang et al. 2010, Laurie et al. 2004, Hayes et al. 2011). For human height for example, Lango-Allen et al. (2010) conducted a powerful experiment to find the loci affecting this trait. They found and validated polymorphisms affecting the height at 180 loci, however these loci together explained only 10% of the variation in human height! And human height is a highly heritable trait.

The search for loci affecting quantitative traits, and the use of this information to increase the accuracy of selecting genetically superior animals, has been the motivation for intensive research efforts in the last two decades. Note that in this course *any* locus with an effect on the quantitative trait is a called a QTL, not just the loci of large effect.

Two approaches have been used to uncover QTL. The *candidate gene approach* assumes that a gene involved in the physiology of the trait could harbour a mutation causing variation in that trait. The gene, or parts of the gene, are sequenced in a number of different animals, and any variations in the DNA sequences, that are found, are tested for association with variation in the phenotypic trait. This approach has had some successes – for example a mutation was discovered in the oestrogen receptor locus (*ESR*) which results in increased litter size in pigs (Rothschild *et al.* 1991). For a review of mutations which have been discovered in candidate genes see Andersson and Georges (2004). There are two problems with the candidate gene approach, however. Firstly, there are usually a large number of candidate genes affecting a trait, so many genes must be sequenced in several animals and many association studies carried out in a large sample of animals (the likelihood that the mutation may occur in non-coding DNA further increases the amount of sequencing required and the cost). Secondly, the causative mutation may lie in a gene that would not have been regarded *a priori* as an obvious candidate for this particular trait. Interestingly, a variant of the candidate gene approach called pathway analysis has recently been used with some success to detect loci underlying variation in quantitative traits. In this approach, *pathways* of genes rather than individual genes that could plausibly affect the trait are identified (eg. Li et al 2013). Then polymorphisms within the genes in the pathway are tested for association with the trait.

An alternative is the QTL mapping approach, in which chromosome regions associated with variation in phenotypic traits are identified. QTL mapping assumes the actual genes which affect a quantitative trait are not known. Instead, this approach uses neutral DNA markers and looks for associations between allele variation at the marker and variation in quantitative traits. A DNA marker is an identifiable physical location on a chromosome whose inheritance can be monitored. Markers can be expressed regions of DNA (genes) or more often some segment of DNA with no known coding function but whose pattern of inheritance can be determined (Hyperdictionary, 2003).

When DNA markers are available, they can be used to determine if variation at the molecular level (allelic variation at marker loci along the linkage map) is linked to variation in the quantitative trait. If this is the case, then the marker is linked to, or on

the same chromosome as, a quantitative trait locus or QTL which has allelic variants causing variation in the quantitative trait.

Until about 2007, the number of DNA markers identified in livestock and plant genomes was comparatively limited, and the cost of genotyping these markers was high. The scarcity of markers constrained experiments designed to detect QTL to using a linkage mapping approach. If a limited number of markers per chromosome are available, then the association between the markers and the QTL will persist only within families and only for a limited number of generations, due to recombination. For example in one sire, the *A* allele at a particular marker may be associated with the increasing allele of the QTL, while in another sire, the *a* allele at the same marker may be associated with the increasing allele at the QTL, due to historical recombination between the marker and the QTL in the ancestors of the two sires.

To illustrate the principle of QTL mapping exploiting linkage, consider an example where a particular sire has a large number of progeny. The parent and the progeny are genotyped for a particular marker. At this marker, the sire carries the marker alleles 172 and 184, Figure 1.1. The progeny can then be sorted into two groups, those that receive allele 172 and those that receive allele 184 from the parent. If there is a significant difference between the two groups of progeny, then this is evidence that there is a QTL linked to that marker.

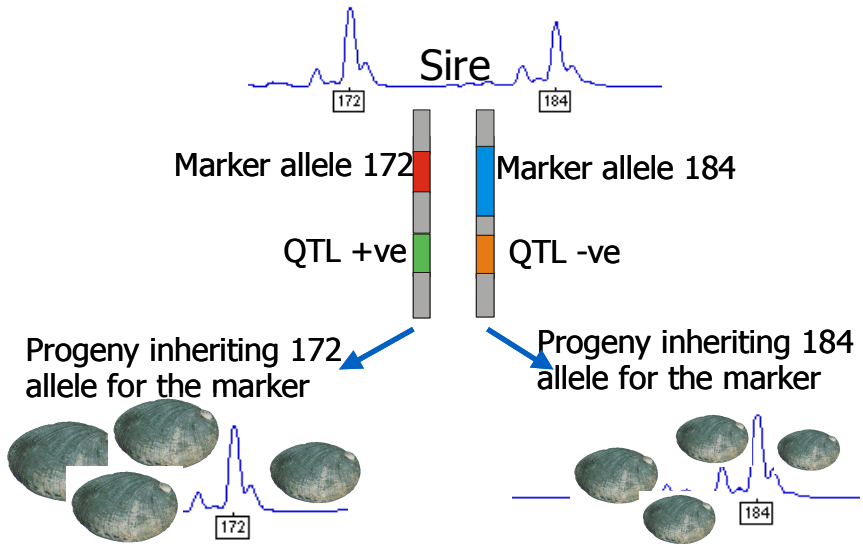


Figure 1.1. Principle of quantitative trait loci (QTL) detection, illustrated using an abalone example. A sire is heterozygous for a marker locus, and carries the alleles 172 and 184 at this locus. The sire has a large number of progeny. The progeny are separated into two groups, those that receive allele 172 and those that receive allele 184. The significant difference in the trait of average size between the two groups of progeny indicates a QTL linked to the marker. In this case, the QTL allele increasing size is linked to the 172 allele and the QTL allele decreasing size is linked to the 184 allele (Figure courtesy of Nick Robinson).

QTL mapping exploiting linkage was performed in all livestock species for a huge range of traits (for a review see Andersson and Georges (2004)). The problem with mapping QTL exploiting linkage is that, unless a huge number of progeny per family or half sib family are used, the QTL are mapped to very large confidence intervals on the chromosome.

An alternative, if dense markers are available, is exploit linkage disequilibrium between markers and QTL. Performing experiments to map QTL in genome wide scans using LD is now possible due to the availability of many thousands of single nucleotide polymorphism (SNP markers) in cattle, pigs, chickens, sheep, salmon and goats. A SNP marker is a difference in nucleotide between individuals (or an individual's pair of chromosomes), at a defined position in the genome, eg.

Animal 1. ACT**C**GGGC

Animal 2. ACT**T**GGGC

Rapid developments in SNP genotyping technology now allow genotyping of a SNP marker in an individual for less than 1c US. This makes possible large experiments to uncover the loci affecting quantitative traits which exploit LD between markers and QTL.

1.2 Definitions and measures of linkage disequilibrium.

The classical definition of linkage disequilibrium (LD) refers to the non-random association of alleles between two loci. Consider two markers, A and B, that are on the same chromosome. A has alleles A1 and A2, and B has alleles B1 and B2. Four haplotypes of markers are possible A1_B1, A1_B2, A2_B1 and A2_B2. If the frequencies of alleles A1, A2, B1 and B2 in the population are all 0.5, then we would expect the frequencies of each of the four haplotypes in the population to be 0.25. Any deviation of the haplotype frequencies from 0.25 is linkage disequilibrium (LD), ie the genes are not in random association. As an aside, this definition serves to illustrate that the distinction between linkage and linkage disequilibrium mapping is somewhat artificial – in fact linkage disequilibrium between a marker and a QTL is required if the QTL is to be detected in either sort of analysis. The difference is:

> *linkage analysis* only considers the linkage disequilibrium that exists within families, which can extend for 10s of cM, and is broken down by recombination after only a few generations.
>
> *linkage disequilibrium* mapping requires a marker to be in LD with a QTL across the entire population. To be a property of the whole population, the association must have persisted for a considerable number of generations, so the marker(s) and QTL must therefore be closely linked.

One measure of LD is *D*, calculated as (Hill 1981)

$$D = freq(A1\_B1)*freq(A2\_B2)-freq(A1\_B2)*freq(A2\_B1)$$

where freq (A1_B1) is the frequency of the A1_B1 haplotype in the population, and likewise for the other haplotypes. The *D* statistic is very dependent on the frequencies of the individual alleles, and so is not particularly useful for comparing the extent of LD among multiple pairs of loci (eg. at different points along the genome). Hill and Robertson (1968) proposed a statistic, r^2, which was less dependent on allele frequencies,

$$r^2 = \frac{D^2}{freq(A1) * freq(A2) * freq(B1) * freq(B2)}$$

Where freq(A1) is the frequency of the A1 allele in the population, and likewise for the other alleles in the population. Values of r^2 range from 0, for a pair of loci with no linkage disequilibrium between them, to 1 for a pair of loci in complete LD.

As an example, consider a situation where the allele frequencies are
freq(A1) = freq(A2) = freq (B1) = freq (B2) = 0.5
The haplotype frequencies are:
freq(A1_B1) = 0.1
freq(A1_B2) = 0.4
freq(A2_B1) = 0.4
freq(A2_B2) = 0.1
The *D* = 0.1*0.1-0.4*0.4 = -0.15
And D^2 = 0.0225.
The value of r^2 is then 0.0225/(0.5*0.5*0.5*0.5) = 0.36. This is a moderate level of r^2.

Another commonly used pair-wise measure of LD is D' (Lewontin 1964). To calculate D', the value of D is standardized by the maximum value it can obtain:

$$D' = |D|/D_{max}$$

Where D_{max}= min[freq(A1)*freq(B2), -1*freq(A2)*freq(B1)] if D>0, else
= min[freq(A1)*freq(B1),--1*freq(A2)*freq*B2)] if D<0.

The statistic r^2 is preferred over D' as a measure of the extent of LD for two reasons. If we consider the r^2 between a marker and an (unobserved) QTL, *r^2 is the proportion of variation caused by the alleles at a QTL which is explained by the markers*. The decline in r^2 with distance actually indicates how many markers or phenotypes are required in initial genome scan exploiting LD are required to detect QTL. Specifically, sample size must be increased by a factor of $1/r^2$ to detect an ungenotyped QTL, compared with the sample size for testing the QTL itself (Pritchard & Przeworski 2001). D' on the other hand does a rather poor job of predicting required marker density for a genome scan exploiting LD, as we shall see in Section 2. The second reason for using r^2 rather than D' to measure the extent of LD is that D' tends to be inflated with small sample sizes or at low allele frequencies (McRae *et al.* 2002).

The above measures of LD are for bi-allelic markers. While they can be extended to multi-allelic markers such as microsatellites, Zhao et al. (2005) recommended the $\chi^{2'}$ measure of LD for multi-allelic markers, where

$$\chi^{2'} = \frac{1}{(l-1)} \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{D_{ij}^2}{freq(A_i)\, freq(B_j)},$$

and $D_{ij} = freq(A_i \_ B_j) - freq(A_i)\, freq(B_j)$, *freq($A_i$)* is the frequency of the i^{th} allele at marker A, *freq(B_j)* is the frequency of the j^{th} allele at marker B, and *l* is the minimum of the number of alleles at marker A and marker B. Note that for bi-allelic markers, $\chi^{2'} = r^2$.

While pair-wise measures of LD are important and widely used, are not particularly illuminating with respect to the causes of LD. For example, statistics such as r^2 consider only two loci at a time, whereas we may wish to calculate the extent of LD across a chromosome segment that contains multiple markers. An alternate multi-locus definition of LD is the **chromosome segment homozygosity** (**CSH**) (Hayes *et al.* 2003). Consider an ancestral animal many generations ago, with descendants in the current population. Each generation, the ancestor's chromosome is broken down, until only small regions of chromosome which trace back to the common ancestor

remain. These chromosome regions are identical by descent (IBD). Figure 1.2 demonstrates this concept.

The CSH then is the probability that two chromosome segments of the same size and location drawn at random from the population are from a common ancestor (ie IBD), without intervening recombination. CSH is defined for a specific chromosome segment, up to the full length of the chromosome. The CSH cannot be directly observed from marker data but has to be inferred from marker haplotypes for segments of the chromosome. Consider a segment of chromosome with marker locus A at the left hand end of the segment and marker locus B at the other end of the segment (as in the classical definition above). The alleles at A and B define a haplotype. Two such segments are chosen at random from the population. The probability that the two haplotypes are identical by state (IBS) is the haplotype homozygosity (HH). The two haplotypes can be IBS in two ways,

i. The two segments are descended from a common ancestor without intervening recombination, so are identical by descent (IBD), or
ii. the two haplotypes are identical by state but not IBD

The probability of i. is CSH. The probability of ii. is a function of the marker homozygosities, given the segment is not IBD. The probabilities of i. and ii. are added together to give the haplotype homozygosity (HH):

$$HH = CSH + \frac{(Hom_A - CSH)(Hom_B - CSH)}{1 - CSH}$$

Where Hom_A and Hom_B are the individual marker homozygosities of marker A and marker B. This equation can be solved for CSH when the haplotype homozygosities and individual marker homozygosities are observed from the data. For more than two markers, the predicted haplotype homozygosity can be calculated in an analogous but more complex manner.

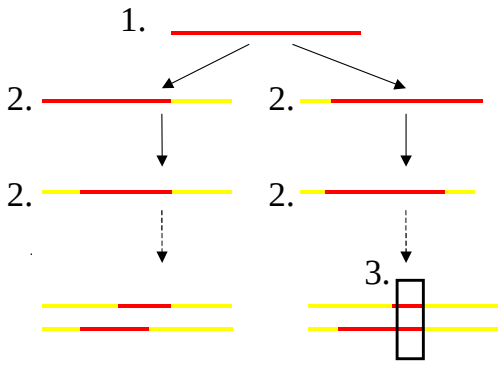


Figure 1.2 An ancestor many generations ago (1) leaves descendants (2). Each generation, the ancestor's chromosome is broken down by recombination, until all that remains in the current generation are small conserved segments of the ancestor's chromosome (3). The chromosome segment homozygosity (CSH) is the probability that two chromosome segments of the same size and location drawn at random from the population are from a common ancestor.

Another justification for using multi-locus measures of LD is that they can be less variable than pair-wise measures. The variation in LD arises from two sampling processes (Weir & Hill 1980). The first sampling process reflects the sampling of gametes to form successive generations, and is dependent on finite population size. The second sampling process is the sampling of individuals to be genotyped from the population, and is dependent on the sample size, *n*. The first sampling process contributes to the high variability of LD measures. Marker pairs at different points in the genome, but a similar distance apart, can have very different r^2 values, particularly if the marker distance is small, Figure 1.3. This is because by chance there may have been an ancestral recombination between one pair of markers, but not the other.

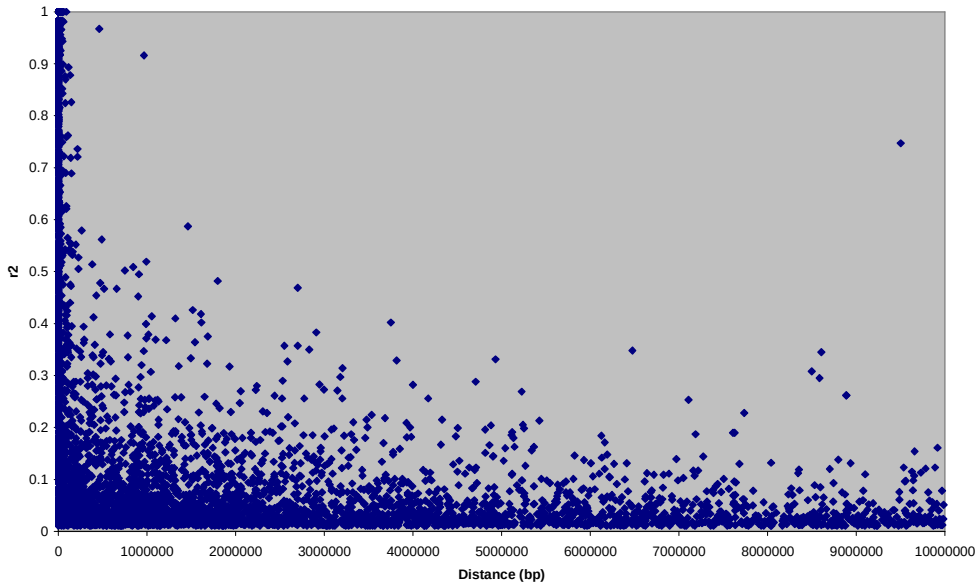


Figure 1.3. The value of r^2 against distance in bases between pairs of markers from 10 000 genome wide SNPs genotyped in a population of Holstein Friesian cattle. 1000000 bases is approximately 1cM.

Multi-locus measures of LD can have reduced variability because they accumulate information across multiple loci in an interval, thus averaging some of the effects of chance recombination.

1.3 Causes of linkage disequilibrium in livestock populations

LD can arise due to migration, mutation, selection, small finite population size or other genetic events which the population experiences (e.g. Lander & Schork 1994). LD can also be deliberately created in livestock populations. In an F2 QTL mapping experiment LD is created between marker and QTL alleles by crossing two inbred lines.

In livestock populations, finite population size is generally implicated as the key cause of LD. This is because

- effective population sizes for most livestock populations are relatively small, generating relatively large amounts of LD
- LD due to crossbreeding (migration) is large when crossing inbred lines but small when crossing breeds that do not differ as markedly in gene frequencies,

and it disappears after only a limited number of generations (e.g. Goddard 1991)

- mutations are likely to have occurred many generations ago.
- while selection is probably a very important cause of LD, it's effect is likely to be localised around specific genes, and so has relatively little effect on the amount of LD 'averaged' over the genome. The use of LD measures to detect selected areas of the genome will be discussed briefly in section 1.8.

1.3.1 Predicting the extent of LD with finite population size

If we accept finite population size as the key driver of LD in livestock populations, it is possible to derive a simple expectation for the amount of LD for a given size of chromosome segment. This expectation is (Sved 1971)

$$E(r^2) = 1/(4Nc+1)$$

where N is the finite population size, and c is the length of the chromosome segment in Morgans. The CSH has the same expectation (Hayes *et al.* 2003). This equation predicts rapid decline in LD as genetic distance increases, and this decrease will be larger with large effective population sizes, Figure 1.4.

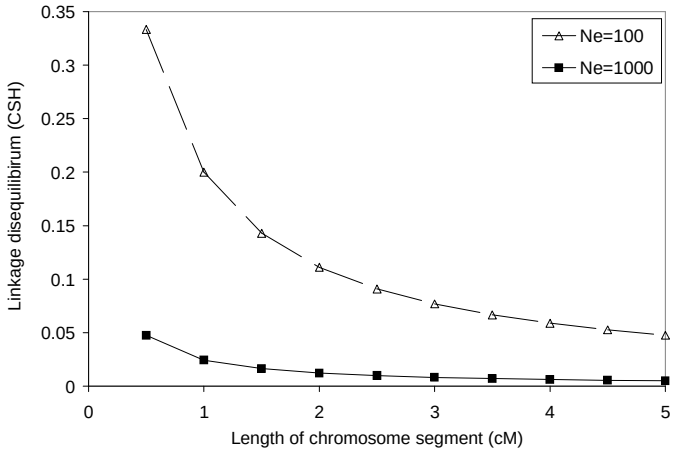
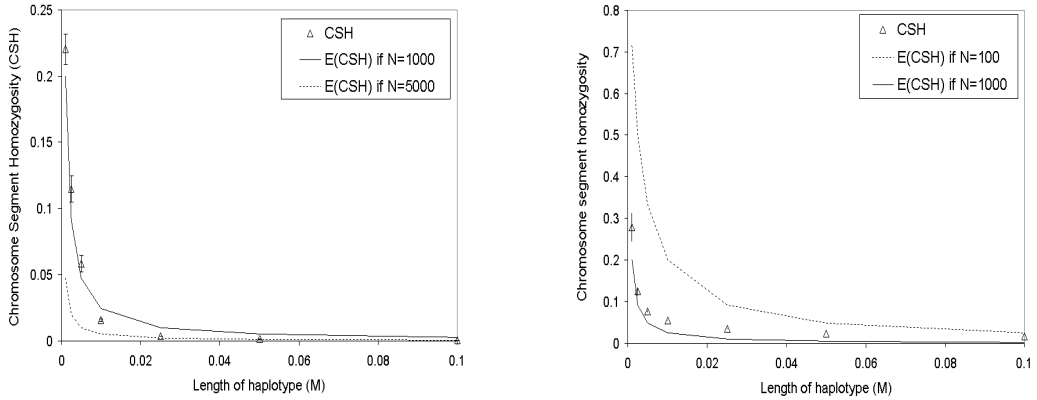


Figure 1.4. The extent of LD (as measured by chromosome segment homozygosity, CSH) for increasing chromosome segment length, for N_e=100 and N_e=1000. Note that r^2 has the same expectation as CSH.

As the extent of LD that is observed depends both on recent and historical recombinations, not only the current effective population size, but also the past effective population size are important. Effective population size for livestock species may have been much larger in the past than they are today. For example in dairy

cattle the widespread use of artificial insemination and a few elite sires has greatly reduced effective population size in the recent past. In humans, the story is the opposite; improved agricultural productivity and industrialisation have led to dramatic increases in population size. How does changing population size affect the extent of LD? To investigate this, we simulated a population which either expanded or contracted after a 6000 generation period of stability. The LD, as measured by CSH, was measured for different lengths of chromosome segment, Figure 1.5. Results for r^2 would look very similar.



A B

Figure 1.5. Chromosome segment homozygosity for different lengths of chromosome (given the recombination rate) for populations: A. Linearly increasing population size, from N=1000 to N=5000 over 100 generations, following 6000 generations at N=1000. B. Linearly decreasing population size, from N=1000 to N=100 over 100 generations, following 6000 generations at N=1000.

The conclusion is that LD at short distances is a function of effective population size many generations ago, while LD at long distances reflects more recent population history. In fact, provided simplifying assumptions such as linear change in population size are made, it can be shown that the r^2 or CSH reflects the effective population size 1/(2c) generations ago, where c is the length of the chromosome segment in Morgans. So the expectation for r^2 with changing effective population size can be written as $E(r^2) = 1/(4N_t c + 1)$ where $t = 1/2c$.

1.4 The extent of LD in livestock and human populations

If LD is a predominantly result of finite population size, then the extent of LD should be less in humans than in cattle, as in humans the effective population size is ~ 10000 (Kruglyak 1999) whereas in livestock where effective population sizes can be as low as 100 (Riquet *et al.* 1999). The picture is somewhat complicated by the fact that livestock populations have been very much larger, while the Caucasian effective population size has been very much smaller (following the out of Africa hypothesis). So what we could expect to see is that at long distances between markers, the r_2 values in livestock are much larger than in humans, while at short distances, the level of LD is more similar. This is in fact what is observed. Moderate LD (eg. $r^2 \geq 0.2$ in humans typically extends less than 5kb (~0.005cM), depending on the population studied (Dunning *et al.* 2000; Reich *et al.* 2001; Tenesa *et al.* 2007), Figure 1.6. In cattle moderate LD extends up to 100kb, Figure 1.6. However, very high levels of LD (eg. $r^2 \geq 0.8$) only extend very short distances in both humans and cattle.

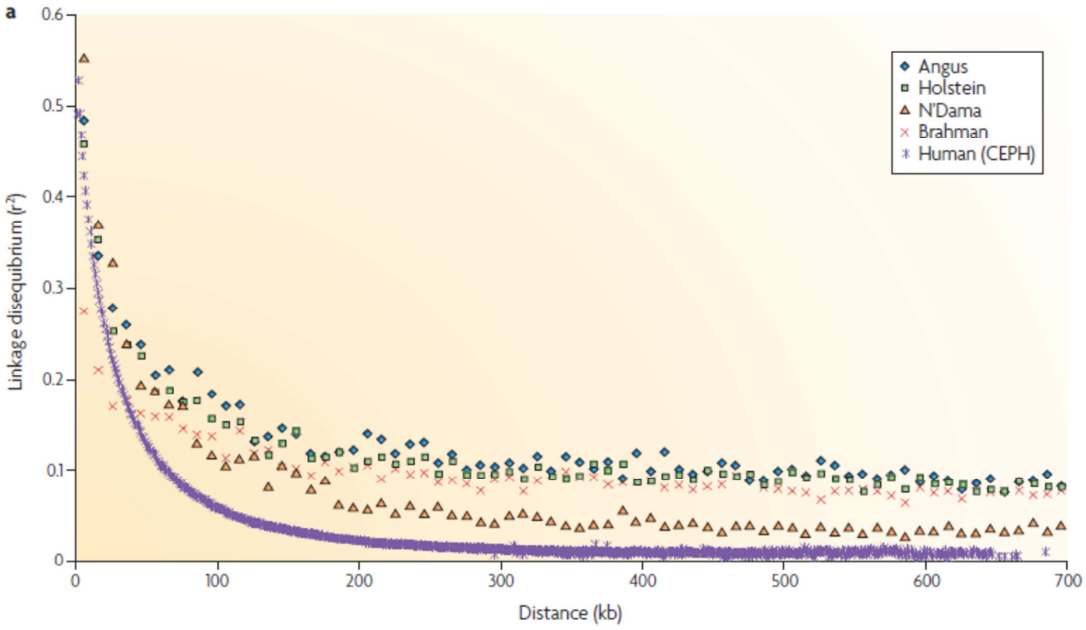


Figure 1.6. A. Average r^2 with distance in Caucasian humans (from Tenesa et al. 2007), and average r^2 value according to the distance between SNP markers in different cattle populations (from Goddard and Hayes 2009, Bovine HapMap Consortium 2009).

Figure 1.6 implies that for the Holstein populations at least, there must be a marker approximately every 100kb (kilo bases) or less to achieve an average r^2 of 0.2. This level of LD between markers and QTL would allow a genome wide association study of reasonable size to detect QTL of moderate effect. As the bovine genome is approximately 3,000,000kb, this implies that in order of 30,000 evenly spaced markers are necessary in order that every QTL in the genome can be captured in a genome scan using LD to detect QTL. In a breed like N'Dama, a larger number of markers would be required, give the lower levels of linkage disequilibrium.

Du et al. (2007) assessed the extent of LD in pigs using 4500 SNP markers genotyped in six lines of commercial pigs. Only maternal haplotypes of the commercial pigs were used to evaluate r^2 between the SNPs, as the paternal haplotypes were over-represented in the population. The results from their study indicate there may be considerably more LD in pigs than in cattle. For SNPs separate by 1cM, the average value of r^2 was approximately of 0.2. LD of this magnitude only extends 100kb in cattle. In pigs at a 100kb the average r^2 was 0.371.

Heifetz et al. (2005) evaluated the extent of LD in a number of populations of breeding chickens. They used microsatellite markers and evaluated the extent of LD with the $\chi^{2'}$ statistic. In their populations, they found significant LD extended long distances. For example 57% of marker pairs separated by 5-10cM had an $\chi^{2'} \geq 0.2$ 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, so the extent of LD in other chicken populations with larger effective population sizes may be substantially different.

In sheep, the extent of LD varies greatly between breeds, reflecting their population histories (Kijas et al. 2012). In breeds such as Border Leicester, the extent of LD is similar to that in Holstein cattle, reflecting a small recent effective population size. However in Merino sheep, the extent of LD is more similar to that observed in human populations, reflecting the fact that even recent effective population size is quite large in this breed (Kijas et al. 2012). At the extreme are Soay sheep, which have been

isolated on an island of the coast of Scotland for many generations, with a small effective population size, reflected in extensive long range LD.

1.5 Extent of LD between populations and breeds.

Marker assisted selection exploiting LD relies on the phase of LD between markers and QTL being the same in the selection candidates as in the reference population where the QTL marker associations were detected. However as the reference population and the population in which MAS is to be applied become more and more diverged, for example different breeds, the phase is less and less likely to be conserved. The statistic r is a measure for LD between two markers in a population, but can also be used to measure the persistence of the LD phases between populations, provided the same allele is designated as the first allele in both populations. While the r^2 statistic between two SNP markers at the same distance in different breeds or populations can be the same value even if the phases of the haplotypes are reversed, they will only have the same value and sign for the r statistic if the phase is the same in both breeds or populations. For marker pairs of a given distance, the correlation between r in two populations, corr(r1,r2), is equal to the correlation of the effects of the marker between both populations, for markers that have that same distance to a QTL (De Roos *et al.* 2008). If this correlation is 1, the 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. Calculating the correlation of r values across different breeds and populations as an indicator of how far the same marker phase is likely to persist between these breeds and populations (Goddard *et al.* 2006). This information can in turn be used to give an indication of marker density required to ensure marker-QTL phase persists across populations and or breeds, which would be necessary for the application MAS or Genomic selection using the same marker set and SNP effects across the breeds or populations.

In Figure 1.7, the correlation of r values is given for a number of different cattle populations. The correlation of r values for Dutch Red-and-white bulls and Dutch Black-and-white bulls was 0.9 at 30kb. This indicates at this distance r^2 is high in both populations and the sign of r is the same in both populations, so the LD phase is

the same in both populations. If one of these SNPs was actually an unknown mutation affecting a quantitative trait, the other SNP could be used in MAS and the favourable SNP allele would be the same in both breeds. For Holstein and Angus breeds, the correlation of r is above 0.9 only at 10kb or less. For Australian Holsteins and Dutch Holsteins, the correlation of r values was above 0.9 up to 100kb, reflecting the fact that there are common bulls used in the two populations (e.g. Zenger *et al.* 2007).

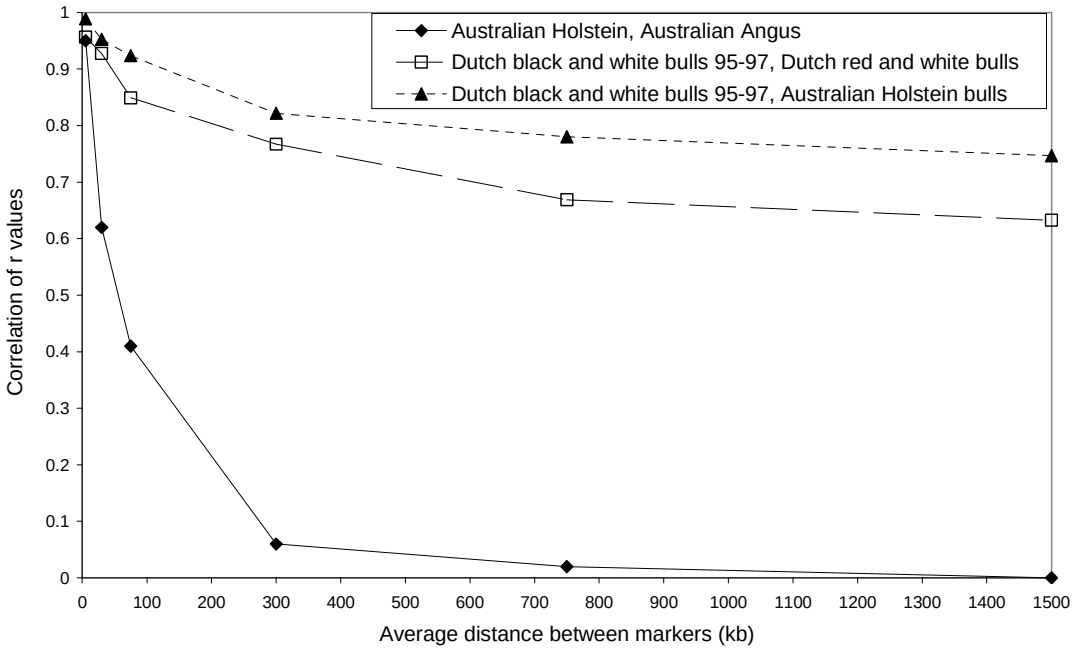


Figure 1.7. Correlation between r values for various cattle populations or sub-populations, as a function of marker distance (from (De Roos *et al.* 2008)).

2. Genome wide association studies

2.1 Introduction

This chapter provides an overview of statistical methods for genome wide association studies (GWAS) in animals, plants and humans.

The simplest form of GWAS, a marker by marker analysis, is illustrated with a small example. The problem of selecting a significance threshold that accounts for the large amount of multiple testing that occurs in GWAS is discussed. Population structure causes false positive associations in GWAS if not accounted for, and methods to deal with this are presented. Methodology for more complex models for GWAS, including haplotype based approaches, accounting for identical by descent versus identical by state, and fitting all markers simultaneously are described and illustrated with examples.

2.2 Genome wide association tests using single marker regression

Genome wide association studies exploit linkage disequilibrium, that is population level associations between markers and causative mutations (also called quantitative trait loci or QTL). These 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. There are a number of statistical methodologies which exploit these associations. The simplest of these is the genome wide association test using single marker regression.

In a random mating population with no population structure the association between a marker and a trait can be tested with single marker regression as

$$\mathbf{y} = \mathbf{W}\mathbf{b} + \mathbf{X}g + \mathbf{e}$$

Where **y** is a vector of phenotypes, **W** is a design matrix assigning phenotype records to fixed effects, **b** is a vector of fixed effects (e.g. the mean, population structure effects, age and so on), **X** is a design matrix allocating records to the marker effect, g

is the effect of the marker and **e** is a vector of random deviates $e_{ij} \sim N(0, \sigma_e^2)$, where σ_e^2 is the error variance. In this model the effect of the marker is treated as a fixed effect, and the model is additive, such that two copies of the second allele has twice as much effect as one copy, and no copies has zero effect. The underlying assumption here is that the marker will only affect the trait if it is in linkage disequilibrium with an unobserved QTL.

The null hypothesis is that the marker has no effect on the trait, while the alternative hypothesis is that the marker does affect the trait (because it is in LD with a QTL). The null hypothesis is rejected if $F > F_{\alpha,v1,v2}$, where *F* is the *F* statistic calculated from the data and $F_{\alpha,v1,v2}$ is the value from an *F* distribution at α level of significance and *v1, v2* degrees of freedom.

Consider a small example of 10 animals genotyped for a single SNP. The phenotypic and genotypic data is:

Animal	Phenotype	SNP allele 1	SNP allele 2
1	2.03	1	1
2	3.54	1	2
3	3.83	1	2
4	4.87	2	2
5	3.41	1	2
6	2.34	1	1
7	2.65	1	1
8	3.76	1	2
9	3.69	1	2
10	3.69	1	2

We need a design matrix **X** to allocate both the mean and SNP alleles to phenotypes. In this case we will use an **X** matrix with number of rows equal to the number of records, and one column for the SNP effect. We will set the effect of the "1" allele to zero, so the SNP effect column in the **X** matrix is the number of copies of the "2" allele an animal carries (**X** matrix in bold):

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**

In this case the **W** matrix is simply a vector, with each element 1, as each individual gets a dose of the mean. The mean and SNP effect can then be estimated as:

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

Where **y** is the (number of animals) vector of phenotypes.
In the above example the estimate of the mean and SNP effect are

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

This is not far from the real value of these parameters. The data above was "simulated" with a mean of 2, a QTL effect of 1, an r^2 (a standard measure of LD) between the QTL and the SNP of 1, plus a normally distributed error term.

The F-value can be calculated as:

$$F = \frac{(n-1)\left(\hat{g}\,\mathbf{X}'\mathbf{y} - 1/n\mathbf{y'y}\right)}{\mathbf{y'y} - \hat{g}\,\mathbf{X}'\mathbf{y} - \hat{u}\,\mathbf{1_n}'\mathbf{y}}$$

Using the above values, the value of F is 4.56. This can be compared to the tabulated F-value of 5.12 at a 5% significance value and 1 and 9 (number of records -1) degrees of freedom. So the SNP effect in this case is not significant (not surprisingly with

only 10 records!). F-values can of course be easily transformed into P values for comparison with significance thresholds, a topic which is addressed later.

2.3 Power of genome wide association tests using single marker regression

An important question for GWAS is how big does the study have to be to have any power to detect associations of a given size? The power of the association test to detect a QTL by testing the marker effect depends on:

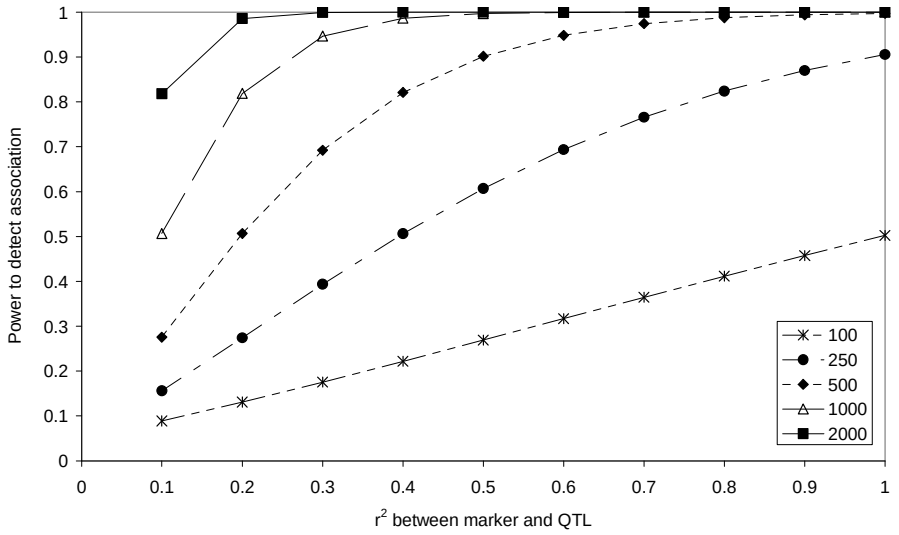
1. The r^2 between the marker and QTL. Specifically, sample size must be increased by a factor of $1/r^2$ to detect an ungenotyped QTL, compared with the sample size for testing the QTL itself (Pritchard & Przeworski 2001).
2. The proportion of total phenotypic variance explained by the QTL, termed h_Q^2.
3. The number of phenotypic records *n*
4. The allele frequency of the rare allele of the SNP or marker, *p*, which determines the minimum number of records used to estimate an allele effect. The power becomes particularly sensitive to *p* when *p* is small (e.g. <0.1).
5. The significance level α set by the experimenter.

The power is the probability that the experiment will correctly reject the null hypothesis when a QTL of a given size of effect really does exist in the population. Figure 2.1 illustrates the power of an association test to detect a QTL with different levels of r^2 between the QTL and the marker and with different numbers of phenotypic records. The power was derived using the formula of (Luo 1998).

Using both this figure, and the extent of LD in our population, we can make predictions of the number of markers required to detect QTL in a genome wide association study. For example, an r^2 of at least 0.2 is required to achieve power ≥ 0.8 to detect a QTL of $h_Q^2 = 0.05$ with 1000 phenotypic records. To illustrate, in dairy cattle, $r^2 \approx 0.2$ at 100kb. So assuming a genome length of 3000Mb in cattle, we would need at least 15 000 markers in such an experiment to ensure there is a marker 100kb from every QTL. However this assumes that the markers are evenly spaced, and all have a rare allele frequency above 0.2. In practise, the markers may not be evenly spaced and the rare allele frequency of a reasonable proportion of the markers

will be below 0.2. Taking these two factors into account, approximately 30 000 markers would be required. In practise, higher levels of r^2 than 0.2 are desirable; otherwise it is difficult to distinguish true associations from noise when 10s of thousands of markers are tested.

A



B

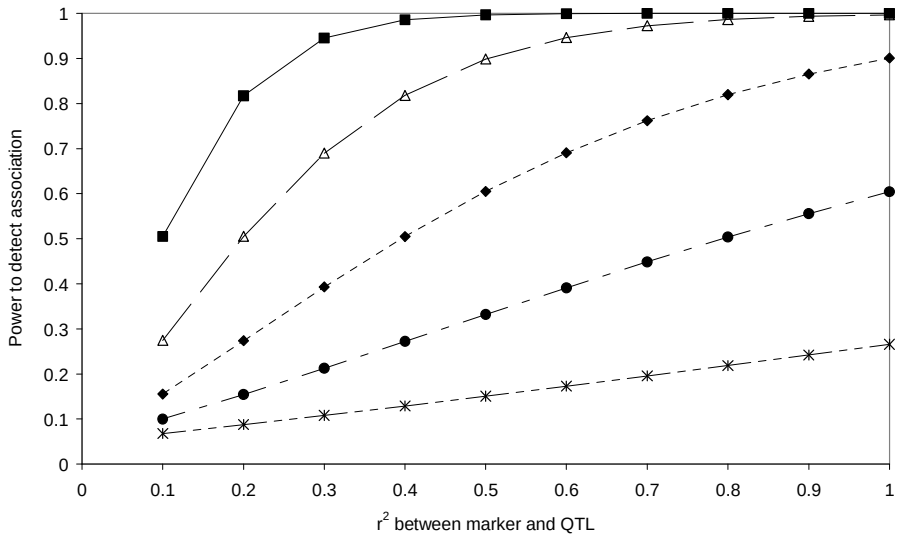


Figure 2.1 A. Power to detect a QTL explaining 5% of the phenotypic variance with a marker. B. Power to detect a QTL explaining 2.5% of the phenotypic variance with a marker, for different numbers of phenotypic records given in the legend and for different levels of r^2 between the marker and the QTL, with a *P* value of 0.05. Rare allele frequencies at the QTL and marker were both 0.2.

2.4 Choice of significance level

With such a large number of markers tested in genome wide association studies, an important question is what value of α to choose. In a genome wide association study, we will be testing tens of thousands, hundreds of thousands or with sequence data potentially millions of variants. So a major issue in setting significance thresholds is the multiple testing problem. In most QTL mapping experiments, many positions along the genome or a chromosome are analysed for the presence of a QTL. As a result, when these multiple tests are performed the "nominal" significance levels of single tests don't correspond to the actual significance levels in the whole experiment, e.g. when considered across a chromosome or across the whole genome. For example, if we set a point-wise significance threshold of 5%, we expect 5% of results to be false positives. If we analyse 100 000 markers (assuming for the moment these points are independent), we would expect 100000*0.05 = 5000 false positive results! Obviously more stringent thresholds need to be set. One option would be to adjust the significance level for the number of markers tested using a Bonferoni correction to obtain an experiment wise P-value of 0.05. However such a correction does not take account of the fact that 'tests' on the same chromosome may not be independent, as the markers can be in linkage disequilibrium with each other as well as the QTL. As a result, the Bonferoni correction tends to be very conservative, or requires some decision to be made about how many independent regions of the genome were tested.

Churchill and Doerge (1994) proposed the technique of permutation testing to overcome the problem of multiple testing in QTL mapping experiments. Permutation testing is a method to set appropriate significance thresholds with multiple testing (e.g. testing many locations along the genome for the presence of the QTL). Permutation testing is performed by analysing a large number of simulated data sets that have been generated from the real one, by randomly shuffling the phenotypes across individuals in the mapping population. This removes any existing relationship between genotype and phenotype, and generates a series of data sets corresponding to the null hypothesis. Genome scans can then be performed on these simulated data-sets. For each simulated data the highest value for the test statistic is identified and stored. The values obtained over a large number of such simulated data sets are

ranked yielding an empirical distribution of the test statistic under the null hypothesis of no QTL. The position of the test statistic obtained with the real data in this empirical distribution immediately measure the significance of the real dataset. For example if we carry out 100 000 analyses of permuted data, the F value for the 5000th highest value will represent the cut off point for the 5% level of significance. Significance thresholds can then be set corresponding to 5% false positives for the entire experiment, 5% false positives for a single chromosome, and so on. Permutation testing is an excellent method of setting significance thresholds in a random mating population. In populations with some pedigree or other structure, however, randomly shuffling phenotypes across marker genotypes will not preserve any pedigree structure that exists in the data.

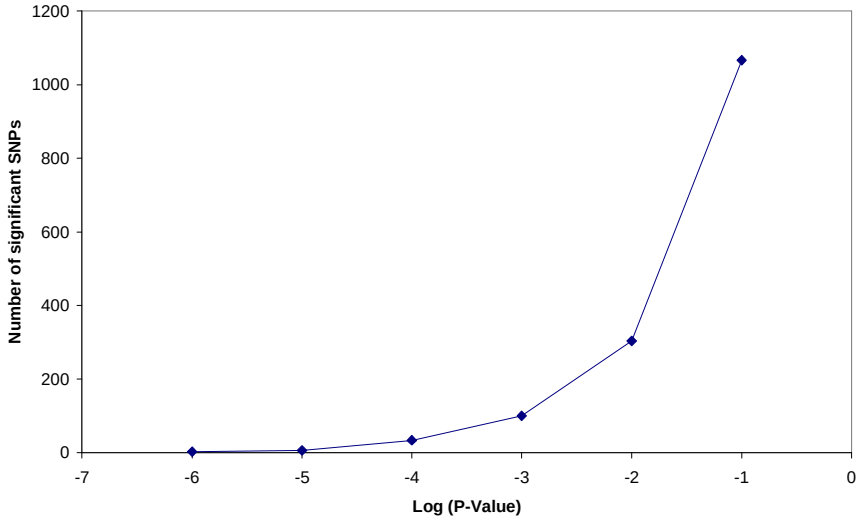
In human genetics, permutation testing has been used to determine the number of independent tests, given the SNP on standard SNP panels (typically close to a million, with >10% MAF), and for widely studies populations. Such studies derive a nominal P value in the order of $<5x10^{-8}$, in order to arrive at an experiment wise P value of 0.05 (Churchill & Doerge 1994).

An alternative to attempting to avoid false positives is to monitor the number of false positives relative to the number of positive results (Fernando *et al.* 2004). The researcher can then set a significance level with an acceptable proportion of false positives. The false discovery rate (FDR) is the expected proportion of detected QTL that are in fact false positives (Benjamini & Hochberg 1995; Weller *et al.* 1998**).** FDR can be calculated for a QTL mapping experiment as

$$mP_{max}/n,$$

where P_{max} is a chosen P value significance threshold, n is the number of QTL which exceed the significance threshold and m is the number of markers tested. Figure 2.2 shows an example of the false discovery rate in an experiment where 9918 SNPs were tested for the effect on feed conversion efficiency in 384 Angus cattle. As the significance threshold is relaxed, the number of significant SNPs increases. However, the FDR also increases.

A



B

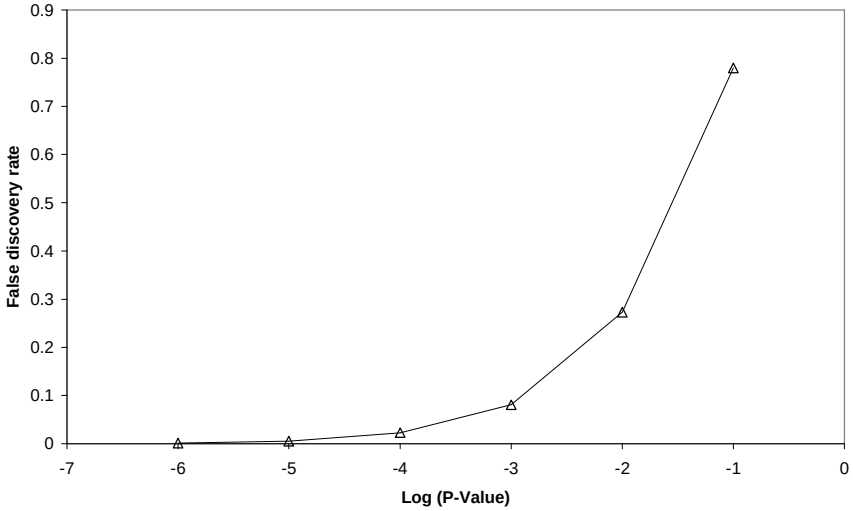


Figure 2.2. A. Number of significant markers at different P values in a genome wide association study with 9918 SNPs, using 384 Angus cattle with phenotypes for feed conversion efficiency. B. False discovery rate at the different P-values.

In this experiment, a P-value of 0.001 was chosen as a criteria to select SNPs for further investigation. At this P-value, there were 56 significant SNPs. So the false discovery rate was 9918*0.001/56 = 0.18. This level of false discovery was deemed acceptable by the researchers.

A number of other statistics have been proposed to control the proportion of false positives, including the proportion of false positives – PRP (Fernando *et al.* 2004), and the positive false discovery rate - pFDR (Storey 2002).

Quantile-Quantile plots (QQ plots) are widely used to display the proportion of significant results compared to the expected number of significant results at a given P value. An example QQ plot (Pryce *et al.* 2011b) is shown in Figure 2.3. The figure clearly demonstrated that in their study, at values greater than P<0.001, more significant SNP were observed than expected by chance.

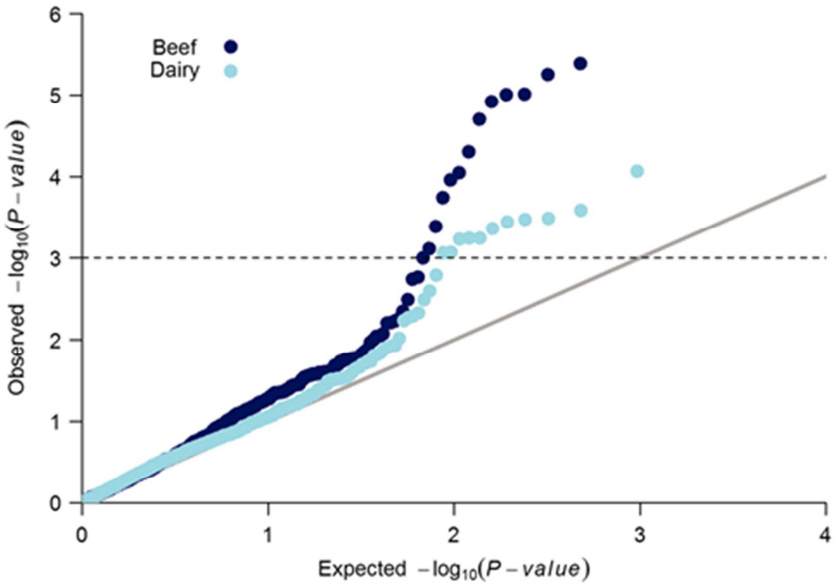


Figure 2.3. An example of a quantile-quantile plot of observed against expected by chance P values. From Pryce et al. (2011b), an association test of SNP for effect on stature in cattle, in regions of genes associated with variation in height in other species. The Quantile-quantile plot is of P-values of 879 SNPs that were 500 kbp either side of 55 orthologous genes found to be associated with height in human populations (Gudbjartsson *et al.* 2008; Lettre *et al.* 2008; Weedon *et al.* 2008; Kim *et al.* 2009). Using dairy and beef data sets, the phenotype (stature) was regressed on each SNP by using a mixed model that included pedigree (ASReml (Gilmour *et al.* 2009)). Observed and expected P-values would fall on the gray solid line if there were no association. The dashed horizontal line is the threshold selected for significance (P < 0.001). Note that a 1-Mbp window was used from which to select SNPs because, in contrast to humans, where LD is expected to persist over only 10s of kilobase pairs (Tenesa *et al.* 2007), non-zero levels of LD have been observed up to 1 Mbp in cattle (Bovine HapMapConsortium 2009).

2.5 Confidence intervals

Interestingly, there are very few reports in the literature on methods to estimate confidence intervals in genome wide association studies. A method based on cross-validation is briefly described here. To calculate approximate 95% confidence intervals for the location of QTL underlying the significant SNPs, a genome wide association study is first conducted as above. The data set is then split into two halves at random (e.g. half the animals in the first data set, the other half in the second data set). The genome wide association study is then re-run for each half of the data. When each half of the data confirmed a significant SNP in the analysis of the full data (i.e. a significant SNP in almost the same location), then a confidence interval can be calculated in the following way. The position of the most significant SNP from each split data set was designated x_{1i} and x_{2i} respectively, for the i^{th} QTL position (taken as the most significant SNP in a region from the full data set). So for n pairs of such SNPs, the standard error of the underlying QTL is calculated as

$se(\bar{x}) = \sqrt{\frac{1}{4n}\sum_{i=1}^{n} x_{1i} - x_{2i}}$. The 95% confidence interval is then the position of the most significant SNP from the full data analysis $\pm 1.96\, se(\bar{x})$.

Using this approach in a data set with 9918 SNPs genotyped on 384 Holstein-Friesian cattle, and for the trait protein kg, there were 24 significant SNP clusters (clusters of SNP putatively marking the same QTL, a cluster consists of 1 or more SNPs) in the full data, and the confidence interval for the QTL was calculated as 2Mb.

2.6 Avoiding spurious false positives due to population structure

Any unaccounted for population structure will result in false positive associations in GWAS (Pritchard *et al.* 2000). In livestock and plant populations with multiple offspring per parent, selection for specific breeding goals and breeds, strains or lines within the population all create population structure. A simple example is where the population includes a parent with a large number of progeny in the population. In our example the parent has a significantly higher estimated breeding value than other parents in the population. If a rare allele at a marker anywhere on the genome is

homozygous in the parent, the sub-population made up of it's progeny will have a higher frequency of the allele than the rest of the population. As the parent estimated breeding value is high, his progeny will also have higher than average estimated breeding values. Then in the genome wide association study, if the number of progeny of the parent is not accounted for, the rare allele will appear to have a (perhaps significant) positive effect.

Spielman et al. (1993) proposed the transmission disequilibrium test (TDT) which requires that parents of individuals in the genome wide association study are genotyped to ensure the association between a marker allele and phenotype is linked to the disease locus, as well as in linkage disequilibrium across the population with it. In this way the TDT test avoids spurious associations due to population structure. However the TDT test has a cost in that genotypes of both parents must be collected, and this is often not possible in livestock and plant populations.

An alternative is to remove the effect of population structure using a mixed model:

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

Where u is a vector of polygenic effect in the model with a covariance structure $u_i \sim N(0, \mathbf{A}\sigma_a^2)$, where **A** is the average relationship matrix built from the pedigree of the population, and σ_a^2 is the polygenic variance. **Z** is a design matrix allocating animals to records. In other words, the pedigree structure of the population is accounted for in the model. Note that this is BLUP, with the marker effect and the mean as fixed effects and the polygenic effects as random effects.

In the study of Macleod et al. (2010), they assessed the effect of including or omitting the pedigree on the number of QTL detected in the experiment, in a simulation where no QTL effects were simulated so that all QTL detected were false positives (Table 1). They found a significant increase in the number of false positives, when the polygenic effects were not fully accounted for.

Table 2.1. Detection of type I errors in data with no simulated QTL (MacLeod *et al.* 2010).

Analysis model	Significance level		
	p<0.005	p<0.001	p<0.0005
Expected type I errors	40	8	4
1. Full pedigree model	39 (SD=14)	9 (SD=5)	4 (SD=3)
2. Sire pedigree model	46* (SD=21)	11* (SD=7)	6* (SD=5.5)
3. No pedigree model	68** (SD=31)	18** (SD=11)	10** (SD=7)
4. Selected 27% - full pedigree	54** (SD=18)	12** (SD=6)	7** (SD=4)

The results indicate that the number of type 1 errors (significant SNPs detected when no QTL exist) is significantly higher when no pedigree is fitted, and even fitting sire does not remove all spurious associations due to population structure.

A problem arises if the pedigree of the population is not recorded, or is recorded with many errors. One solution in this case is to use the markers themselves to infer the genomic relationship matrix **G** (Hayes *et al.* 2007) or population structure (e.g. Pritchard *et al.* 2000). The G matrix can then be fitted in the place of A in the model above.

Principal components (of the genomic are widely used in human GWAS to take account of population structure (e.g. Patterson *et al.* 2006). In livestock and plant populations, extreme caution is recommended with principal components approaches, as unless they are specifically tested it is unclear what component of variation they are removing (McVean 2009; Daetwyler *et al.* 2012a).

One way of determining if population structure has been successfully removed is to inspect the QQ plot. If population structure has not been correctly accounted for, this

will result in an excess of associations at all levels of P-value. That is, the observed values will be greater than the expected values at all values of $-\log_{10}(P)$.

2.7 Genome wide association experiments using haplotypes

Rather than using single markers, haplotypes of markers could be used in the genome wide association. The effect of haplotypes in windows across the genome would then be tested for their association with phenotype. The justification for using haplotypes is that marker haplotypes may be in greater linkage disequilibrium with the QTL alleles than single markers. If this is true, then the r^2 between the QTL and the haplotypes is increased, thereby increasing the power of the experiment.

To understand why marker haplotypes can have a higher r^2 with a QTL than an individual marker, consider two chromosome segments containing a QTL drawn at random from the population, which happen to carry identical marker haplotypes for the markers on the chromosome segment. There are two ways in which marker haplotypes can be identical, either they are derived from the same common ancestor so they are identical by descent (IBD), or the same marker haplotypes have been regenerated by chance recombination (identical by state IBS). If the "haplotype" consists only of a single SNP the chance of being identical by state is a function of the marker homozygosity. Now as more and more markers are added into the chromosome segment, the chance of regenerating identical marker haplotypes by chance recombination is reduced. So the probability that identical haplotypes carried by different animals are IBD is increased. If the haplotypes are IBD, then the chromosome segments will also carry the same QTL alleles. As the probability of two identical haplotypes being IBD increases, the proportion of QTL variance explained by the haplotypes will increase, as marker haplotypes are more and more likely to be associated with unique QTL alleles. This is particularly true for QTL with rare (low frequency) minor alleles.

Just as for single markers, the proportion of QTL variance explained by the markers can be calculated. Let q_1 be the frequency of the first QTL allele and q_2 be the frequency of the second QTL allele. The surrounding markers are classified into *n*

haplotypes, with p_i the frequency of the i^{th} haplotype. The results can be classified into a contingency table:

	Haplotype			
	1	i	N	
QTL allele 1	p_1q_1-D_1	p_iq_1-D_i	p_nq_1-D_n	Q_1
QTL allele 2	p_1q_2+D_1	p_1q_2+D_i	p_nq_2+Dn	Q_2
	p_1	p_i	p_n	1

For a particular haplotype i represented in the data, we calculated the disequilibrium as $D_i = p_i(q_1)$-p_iq_1, where $p_i(q_1)$ is the proportion of haplotypes i in the data that carry QTL allele 1 (observed from the data), p_i is the proportion of haplotypes i, and q_1 is the frequency of QTL allele 1. The proportion of the QTL variance explained by the haplotypes, and corrected for sampling effects was then calculated as

$$r^2(h,q) = \frac{\sum_{i=1}^{n} \frac{D_i^2}{p_i}}{q_1 q_2}$$

A model for testing haplotypes in an association study could be similar to the model described above:

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

However **g** is now a vector of haplotype effects rather than the effect of a single marker. The haplotypes could be treated as random, as there are likely to be many of them and some haplotypes will occur only a small number of times. The effect of treating the haplotypes as random is to "shrink" the estimates of the haplotypes with only a small number of observations. This is desirable because it reflects the uncertainty of predicting these effects. So $g_i \sim N(0, I\sigma_h^2)$ where I is an identity matrix and σ_h^2 the variance of the haplotype effects. The g can be estimated from the equations:

$$\begin{bmatrix} \hat{\mu} \\ \hat{\mathbf{u}} \\ \hat{\mathbf{g}} \end{bmatrix} = \begin{bmatrix} \mathbf{1_n}'\mathbf{1_n} & \mathbf{1_n}'\mathbf{Z} & \mathbf{1_n}'\mathbf{X} \\ \mathbf{Z}'\mathbf{1_n} & \mathbf{Z}'\mathbf{Z} + \mathbf{A}^{-1}\lambda_1 & \mathbf{Z}'\mathbf{X} \\ \mathbf{X}'\mathbf{1_n} & \mathbf{X}'\mathbf{Z} & \mathbf{X}'\mathbf{X} + \mathbf{I}\lambda_2 \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{1_n}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \\ \mathbf{X}'\mathbf{y} \end{bmatrix}$$

Where $\lambda_1 = \frac{\sigma_e^2}{\sigma_a^2}$, and $\lambda_2 = \frac{\sigma_e^2}{\sigma_h^2}$. Note that this model assumes no-covariance between haplotype effects. In practise, the haplotype variance is unlikely to be known, so will need to be estimated . A REML program, such as ASREML (Gilmour *et al*. 2009), can be used to do this. As the haplotypes are fitted as random effects, an F value is no longer appropriate. Rather, the statistic -2*(Loglikelihood no haplotype fitted – Loglikelihood haplotype fitted) can be calculated, and compared to an inverse chi square distribution with 1 degree of freedom.

In GWAS in real data, haplotypes may have some advantage. Pryce et al. (2010a) conducted a GWAS using either 50,000 genome wide SNP or haplotypes constructed from the alleles of these SNP, in a dairy cattle population. For the trait fertility, significant effects were only detected, and subsequently validated in a different population, when haplotypes were used. There was little difference, in terms of number of effects validated for other traits like milk production.

While the use of haplotypes seems initially attractive, there are a number of factors which potentially limit their value over single markers. These are:

- The requirement that the genotypes must be sorted into haplotypes and this may not be a trivial task, and it may not be 100% accurate (see chapter 5).
- The number of effects which must be estimated increases. For a single marker there is one effect to estimate if an additive model is assumed, while for marker haplotypes there are potentially a large number of effects to estimate depending on the number of markers in the haplotype.
- Some simulation results which show benefits of marker haplotypes rely on increasing the density of markers in a given chromosome segment to achieve this. This may not be possible in practise.

2.8 Mapping QTL with an Identical by descent approach

The identical by descent (IBD) is quite different from that used in single marker or haplotype regression, as now the effect of a putative QTL is explicitly modelled, rather than assuming the marker is associated with the QTL:

$$y_i = \mu + u_i + vp_i + vm_i + e_i$$

Where vp_i and vm_i are the effects of the QTL alleles carried on the ith animal's paternal and maternal chromosome respectively. In this model, the assumption is that each animal carries two unique QTL alleles, and so there are two QTL effects fitted for each animal.

Then marker haplotype information is used to infer the probability that two individuals carry the same QTL allele at a putative QTL position. As described above, the existence of LD implies there are small segments of chromosome in the current population which are descended from the same common ancestor. These IBD chromosome segments will not only carry identical marker haplotypes; if there is a QTL somewhere within the chromosome segment, the IBD chromosome segments will also carry identical QTL alleles. Therefore if two animals carry chromosomes which are likely to be IBD at a point of the chromosome carrying a QTL, then their phenotypes will be correlated. We can calculate the probability the 2 chromosomes are IBD at a particular point based on the marker haplotypes and store these probabilities in an IBD matrix (**G**). Then the v are distributed $v \sim N(0, G\sigma^2_{QTL})$, where σ^2_{QTL} is the QTL variance. If the correlation between the animals is proportional to G there is evidence for a QTL at this position.

Consider a chromosome segment which carries 10 marker loci and a single central QTL locus. Three chromosome segments were selected from the population at random, and were genotyped at the marker loci to give the marker haplotypes 11212Q11211, 22212Q11111 and 11212Q11211, where Q designates the position of the QTL. The probability of being IBD at the QTL position is higher for the first and third chromosome segments than for the first and second or second and third chromosome segments, as the first and third chromosome segments have identical marker alleles for every marker locus.

This type of information can be used, together with information on recombination rate of the chromosome segment and effective population size, for calculating an IBD matrix, **G**, for a putative QTL position from a sample of marker haplotypes. Element G_{ij} of this matrix is the probability that haplotype *i* and haplotype *j* carry the same

QTL allele. The dimensions of this matrix is (2 x the number of animals) x (2 x the number of animals), as each animal has two haplotypes.

Meuwissen and Goddard (2001) described a method to calculate the IBD matrix based on deterministic predictions which took into account the number of markers flanking the putative QTL position which are identical by state, the extent of LD in the population based on the expectation under finite population size, and the number of generations ago that the mutation occurred.

Now consider a population of effective population size 100, and a chromosome segment of 10cM with eight markers. Two animals are drawn from this population. Their marker haplotypes are 12222111, 11122111 for the first animal, and 12222111 and 11122211 for the second animal. The putative QTL position is between markers 4 and 5 (i.e. in the middle of the haplotype). The **G** matrix could look something like:

			Animal 1		Animal 2	
			Hap 1	Hap 2	Hap 1	Hap 2
			12222111	11122111	12222111	11122211
Animal 1	Hap 1	12222111	1.00			
	Hap 2	11122111	0.30	1.00		
Animal 2	Hap 1	12222111	**0.90**	0.30	1.00	
	Hap 2	11122211	0.20	0.40	0.20	1.00

To estimate the additive genetic variance, we could calculate the extent of the correlation between animals with high additive genetic relationships A_{ij}. In practise, we fit a linear model which includes additive genetic value (**u**) with $\mathbf{V(u)} = \mathbf{A}\sigma_a^2$, and then estimate σ_a^2. In a similar way, to estimate the QTL variance at a putative QTL position we fit the following linear model:

$$\mathbf{y} = \mathbf{1_n}\mu + \mathbf{Zu} + \mathbf{Wv} + \mathbf{e},$$

where ***W*** is a design matrix relating phenotypic records to QTL alleles, ***v*** is a vector of additive QTL effects, *e* the residual vector, where the random effects *v* are assumed to be distributed as $v \sim (0, \boldsymbol{G}\sigma_{QTL}^2)$. A REML program, such as ASREML (Gilmour *et al.* 2009), can be used to estimate the QTL variance and the likelihood of the data given the QTL and polygenic parameters.

QTL mapping then proceeds by proposing a putative QTL position at intervals along the chromosome. At each point, the QTL variance is estimated and the likelihood of the data given the QTL and polygenic parameters is calculated. The most likely position of the QTL is the position where this likelihood is a maximum.

The significance of the QTL at its most likely position can then be tested using a likelihood ratio test by comparing the maximum likelihood of the model with the QTL fitted and without the QTL fitted:

$$LRT = -2(LogLikelihood_{no\_QTL\_fitted} - LogLikelihood_{QTL\_fitted})$$

This test statistic has a χ_1^2 distribution. The QTL is significant at the 5% level if LR > 3.84.

Grapes et al. (2004), Grapes et al. (2006) and Zhao et al.(2007) compared single marker regression, regression on marker haplotypes and the IBD mapping approach for the power and precision of QTL mapping. Grapes et al (2004) and Grapes et al. (2006) did this assuming a QTL had already been mapped to a chromosome region, Zhao et al (2007) did this in the context of a genome wide scan for QTL. All three papers compared the approaches using simulated populations. The conclusion from these papers was that single marker regression gives greater power and precision than regression on marker haplotypes, and was comparable to the IBD method. However these results contradict those of Hayes and Goddard (2008) , who found that in real data (9323 SNPs genotyped in Angus cattle) using marker haplotypes would give greater accuracy of predicting QTL alleles than single markers. They also contradict the results of Calus et al.(2008), who found that in genomic selection, use of the IBD approach gave greater accuracies of breeding values than using either single marker regression or regression on haplotypes, particularly at low marker densities (discussed further in section 8). The explanation for the contradictory results may be that these authors (Grapes *et al.* 2004; Grapes *et al.* 2006; Zhao *et al.* 2007) were simulating a situation where single markers had very high r^2 values with the QTL, in which case using marker haplotypes would only add noise to the estimation of the QTL effect.

With current densities of markers in livestock (up to 777,000 for cattle), the high levels of r^2 obtainable would appear to make the IBD approaches redundant. However, this statement does have an implicit assumption that the distribution of allele frequencies of the QTL match that of the markers, otherwise the LD between QTL and markers will still be limited. For traits where many of the QTL have low minor allele frequencies, using haplotypes or the IBD approach may still have considerable benefits. For example, Browning and Thompson (2012) reported rare sequence variants associated with type 1 diabetes that were only detected with an IBD approach.

2.9 Fitting all markers simultaneously in GWAS

There are two disadvantages of the approaches described above that fit either single SNPs, haplotypes or single genome regions in the analysis. One of these is the multiple testing problem, that is many thousands of tests are run, so the significance level must be very stringent to take this into account. Further, the setting of a significance threshold combined with the testing of so many marker effects means that the markers most likely to exceed the threshold are those with favourable error terms, so that the significant markers have over-estimated effects. The second disadvantage, particularly of the single SNP approach, is that a region containing the true mutation can be hard to define, as a large number of SNP can be in LD with the QTL, such that significant SNP span a wide region (e.g. Pryce *et al.* 2010a). This is particularly problematic in livestock (and likely some plant species), as low, but non zero, LD extends for Mb. While a partial solution to this second problem is to jointly fit SNP in multiple or conditional regression (e.g. Yang *et al.* 2012), an even better solution to both these issues is to fit all SNP simultaneously. This involves fitting the same models that have been proposed for genomic prediction (e.g. Meuwissen *et al.* 2001), which is the subject of the next chapter.

2.10 The need for validation

The only evidence that a significant association detected in a GWAS is "real" (that is truly associated with a QTL affecting the trait) is validation in an independent population. Despite efforts to control for population structure, and use of fairly

stringent thresholds, false positives will still occur in GWAS given the enormous number of SNPs tested, which means that the chance that at least one of these is associated with some unaccounted for structure in the data is high. This means that the design of a GWAS experiment includes both discovery and validation. A validation experiment is also required to more accurately estimate the size of the QTL effect, as in the discovery experiment the effect of the QTL will be over-estimated, as described in section 2.9. The validation set must be large enough to have sufficient power (e.g. Figure 2.1), otherwise a SNP may fail to validate just because the experiment is underpowered. The relationship between the discovery and validation set should also be carefully considered. For example, if a significant SNP is discovered in a population of dairy bulls, and the SNP is "validated" in their daughters, there is high chance that the same population structure exists in both data sets, leading to apparent validation of what is really a false positive result. In livestock, the most convincing validation is across breeds (as the pedigree structure in the breeds should be independent). However, if SNP fail to validate across breeds it may be because the underlying QTL is not segregating in both breeds.

3. Genomic selection

3.1 Introduction to genomic selection

One way to use DNA marker information in livestock and plant breeding would be to first perform a GWAS, then take the most significant markers and use them in marker assisted selection. However, for traits that are controlled by a large number of loci all with small or small to moderate effects, marker assisted selection will result in only small gains in the accuracy of breeding values, as only a limited proportion of the total genetic variance will be captured by the markers. An alternative to tracing a limited number of QTL with markers is to trace all the QTL. This can be done by dividing the entire genome up into chromosome segments, for example defined by adjacent markers, and then tracing all the chromosome segments. This method was termed genomic selection by Meuwissen et al. (2001). Genomic selection exploits linkage disequilibrium – the assumption is that the effects of the chromosome segments will be the same across the population because the markers are in LD with the QTL that they bracket. Hence the marker density must be sufficiently high to ensure that all QTL are in LD with a marker or haplotype of markers. Genomic selection is now possible with the availability of many thousands of markers and high throughput genotyping technology.

Implementation of Genomic selection conceptually proceeds in two steps, 1. Estimation of the effects of chromosome segments in a reference population and 2. Prediction of genomic EBVs (GEBVs) for animals not in the reference population, for example selection candidates. This second step is straightforward: To predict GEBVs for animals with genotypes but no phenotypes the effect of the chromosome segments they carry can be summed across the genome:

$$\mathbf{GEBV} = \sum_{\mathbf{i}}^{\mathbf{n}} \mathbf{X_i}\, \hat{\mathbf{g}}_\mathbf{i}$$

Where n is the number of markers across the genome, X_i is a design matrix allocating animals to genotypes at marker i, and $\hat{\mathbf{g}}_\mathbf{i}$ is the effect of the genotype at marker i.

The difficulty in step 1. is that a very large number of marker effects must be estimated (the $\hat{\mathbf{g}}_\mathbf{i}$), most likely from a data set where the number of phenotypic observations is much less than the number of marker effects to be estimated. Most of this chapter is devoted to this problem.

Before we discuss methods to simultaneously estimate a large number of marker effects from a limited number of phenotypes, a few key points. It is important to note that genomic selection has the desirable property that because all chromosome segment effects are estimated simultaneously, the problem of over-estimation of QTL effects due to multiple testing described in section 2.9 does not occur.

Genomic selection can proceed using single markers, haplotypes of markers or using an IBD approach. The methodologies that will be described in this chapter can be applied to either single markers or haplotypes. The difference will be in the number of effects to estimate per chromosome segment (ignoring the problems of inferring haplotypes). In the case of single markers, there will be one effect per segment (eg. $\hat{g}_i$ are scalars). In the case of marker haplotypes, there will be multiple effects per segment (eg. $\hat{\mathbf{g}}_\mathbf{i}$ are a vector). Also, the following genomic selection procedures can be used to map QTL as well as predict GEBV.

3.2 Least squares for genomic selection

A number of approaches have been proposed for estimating the single marker or haplotype effects across chromosome segment effects for genomic selection. The simplest of all, and usually worst performing, is the least squares approach.

The first approach treats marker effects as fixed effects in a least squares approach. As described by Meuwissen et al. (2001) least squares genomic selection proceeds in two steps.

1. Perform single segment regression analyses for every segment, *i*, using the model

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

where ***y*** is the data vector; μ is the overall mean; $\mathbf{1}_n$ is a vector of *n* (*n*=number of records) ones; g_i represents the genetic effects of the marker or haplotypes at the i^{th} 1-cM segment (the vector of values of $\hat{g}_{ij}$ for the different *j* but at the same *i*) ; X_i is the design matrix for the i^{th} segment; and ***e*** is the error deviation. If haplotypes are fitted, the dimensions of $\mathbf{g_i}$ will be (number of haplotypes within chromosome segment *i* x 1), while the dimensions of $\mathbf{X_i}$ will be (number of records x number of haplotypes within chromosome segment *i*).

- *2*. Select the *m* most significant segments. Estimate the effects of the markers or haplotypes at these positions simultaneously using multiple regression $\mathbf{y} = \mu\mathbf{1_n} + \sum_m \mathbf{X_i g_i} + \mathbf{e}$ where summation Σ_m is over all significant QTL positions. All other haplotype effects are assumed to be zero.

The least squares approach has two major problems. One is the choice of significance level (arguments such as FDR could be used). This must not be too lenient, or else the number of chromosome segment effects to estimate will be larger than the number of phenotypic records, in which case least squares cannot be used. The other is that in the least squares approach, there is a selection of which markers or chromosome segments to include in the estimation of breeding values based on the effect of the single marker or haplotype regressions. As a result, the problem of over-estimation of effects due to multiple testing will be incurred.

3.3 SNP-BLUP and Ridge Regression

To overcome the limitations of least squares, approaches which treat marker effects or haplotype effects as random effects (effects come from a distribution of effects) have been proposed. A number of different assumptions about the distribution of effects of marker effects or haplotype effects are possible.

If we fit a model to the data where the observed phenotypes are the result of the overall mean and m marker effects (the effects are actually at the QTL of course, we are just using the associations with the markers to track them):

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

Where

y is a vector of phenotypes (number of records x 1)

$\mathbf{1_n}$ is a vector of 1s, allocating the effect of the mean to each record

μ is the overall mean

X is a design matrix, allocating records to genotypes for *m* markers (number or records x *m*)

g is a vector of the effects of the *m* markers

e is a vector of random residuals, assumed normally distributed, variance σ_e^2

We want to use this model to estimate the effects of the markers. Perhaps the simplest assumption we can make is that the marker effects are all very small, and are normally distributed, eg $V(\mathbf{g}) \sim N(0, I\sigma_g^2)$ where σ_g^2 is the variance of the marker effects for all markers. If we make this assumption, marker effects can be predicted as

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

Where $\lambda = \frac{\sigma_e^2}{\sigma_g^2}$ and ***I*** is an Identity matrix (with dimensions number of markers x number of markers). This method for predicting marker effects has been called Best linear unbiased prediction (BLUP) (Meuwissen et al. 2001) or SNP-BLUP (Moser et al. 2010) for genomic selection.

Let's now consider a small example. In the following data set there are 5 animals with phenotypes, and each animal has been genotyped for 10 markers. The genotypes have been coded as the number of copies of the second allele at the marker. For example, if the alleles at the marker were A and T, and an animal had the genotype

AA, the animal would have a 0 coded genotype. If the genotype was AT, the coded genotype would 1, and TT would be 2. These coded genotypes become the X matrix.

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

In this small example, the phenotypes were generated with a mean of 1, an effect of the second allele for the first marker of 1 (eg an animal with the coded genotype 1 would get an effect +1), and a random error term. The effect of markers 2-9 was zero.

Now let's fit the model

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

to the data, and estimate the mean and marker effects

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

To do this, we can build up the blocks of the coefficient matrix $\begin{bmatrix} \mathbf{1_n'1_n} & \mathbf{1_n'X} \\ \mathbf{X'1_n} & \mathbf{X'X + I\lambda} \end{bmatrix}^{-1}$

And the right hand side $\begin{bmatrix} \mathbf{1_n'y} \\ \mathbf{X'y} \end{bmatrix}$. The $\mathbf{1_n'}$ is the transpose of a 5 x 1 vector of 1s, eg

[1 1 1 1 1]. Using a value of 1 for lamda, we get

$$\begin{bmatrix} \hat{\mu} \\ \hat{\mathbf{g}} \end{bmatrix} = \begin{bmatrix} 0.47 \\ 0.29 \\ -0.05 \\ -0.05 \\ 0.08 \\ -0.02 \\ 0.13 \\ 0.13 \\ -0.08 \\ 0.11 \\ -0.08 \end{bmatrix}$$

That is our estimate of the mean is 0.47, the estimate of the effect of a 2 allele at the first SNP is 0.29 and so on. We can then use the vector of SNP effects (the prediction equation) to predict estimated breeding values for a group of selection candidates with genotypes only. Let's say we have five progeny for which we want GEBV, with genotypes coded into a new X matrix as below:

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

Then we can calculate their GEBV as $\mathbf{GEBV} = \mathbf{X}\hat{\mathbf{g}}$

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

Selection candidates could then be ranked on GEBV and the best selected for breeding.

When implementing SNP-BLUP in practise, the value of σ_g^2 is unlikely to be known. In this case the procedure could more correctly be referred to as Ridge Regression. There are a number of options to obtain values for σ_g^2 . One would be to first estimate the total additive genetic variance (using REML in a pedigree analysis for example) then divide by the number of markers or chromosome segments, eg $\sigma_g^2 = \sigma_a^2 / m$. Additive genetic variances have been estimated for many traits in livestock breeding. However this simple equation does not take into account the differences in marker allele frequencies. A better estimate is therefore $\sigma_g^2 = \sigma_g^2 / 2\sum_{j=1}^{m} p_j(1-p_j)$.

This is still one potential problem with this estimate, which is that it assumes the linkage disequilibrium between SNP and QTL is perfect, that is the SNP capture all the genetic variation. In practise this may not be the case. An alternative way to estimate λ which takes this into account is cross validation. In this approach, part of the data is set aside when fitting the SNP-BLUP model. The model is solved (SNP effects predicted) with different values of λ. Then GEBV are predicted for the animals that were set aside, and the value of λ is taken which minimises the mean square error between the GEBV and the y. This process can be repeated, dropping out different subsets of the data, to obtain good estimates of λ by averaging across data sets (Moser et al. 2010).

3.4 An equivalent model using the genomic relationship matrix (GBLUP)

An useful alternative method for implementing genomic selection is to predict breeding values using a genomic relationship matrix, in place of the pedigree derived relationship matrix (eg Habier et al. 2007, VanRaden et al. 2009, Hayes et al 2009). This model is actually an equivalent model to predicting individual SNP effects and calculating GEBV as the sum of these effects, provided the SNP effects are assumed to be normally distributed. If we assume a model

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

where **y** is a vector of phenotypes, μ is the mean, $\mathbf{1_n}$ is a vector of 1s**,** **Z** is a design matrix allocating records to breeding values, **u** is a vector of breeding values and **e** is a vector of random normal deviates ~ $N(0, \sigma_e^2)$. Then **u**=**Wg** where g_j is the effect of the j^{th} SNP, and $V(\mathbf{u}) = \mathbf{WW'}\sigma_g^2$. **W** is a design matrix allocating records to genotypes, as for the **X** matrix in the section above, but correacted for allele frequences. Elements of matrix **W** are w_{ij} for the i^{th} animal and j^{th} SNP, where w_{ij} = 0 – $2p_j$ if the animal is homozygous 11 at the j^{th} SNP, 1-$2p_j$ if the animal is heterozygous and 2 – $2p_j$ if the animal is homozygous 22 at the j^{th} SNP (eg $w_{ij}=x_{ij}-2p_j$, where **X** is the matrix used in SNP-BLUP above). The diagonal elements of **WW'** will be $\sum_{j=1}^{m} 2p_j(1-p_j)$ where *m* is the number of SNPs. If **WW'** is scaled such that $\mathbf{G} = \frac{n\mathbf{WW'}}{\sum_{i=1}^{n} w_{ii}}$ then $V(\mathbf{u}) = \mathbf{G}\sigma_u^2$. GEBV for both phenotyped and non-phenotyped individuals can be then predicted by solving the equations:

$$\begin{bmatrix} \hat{\mu} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} 1_n'1_n & 1_n'Z \\ Z'1_n & \mathbf{Z'Z} + \mathbf{G}^{-1}\frac{\sigma_e^2}{\sigma_u^2} \end{bmatrix}^{-1} \begin{bmatrix} 1_n'y \\ Z'y \end{bmatrix}$$

Implementing genomic selection in this way is attractive, as all that may be required is to replace the average relationship matrix with the genomic relationship matrix in the existing genetic evaluation. The method is also very attractive for populations without good pedigree records – the genomic relationship matrix will capture this information among the genotyped individuals at least. In real data, this method has been shown to be at least as good for many traits as other methods (VanRaden et al. 2009). Note that σ_u^2 may be less than the additive genetic variance for the trait, if the linkage disequilibrium between SNP and QTL is not perfect.

3.5 Bayesian methods

Both GBLUP and SNP-BLUP make the prior assumption that the effects of the SNP are all non zero, small and normally distributed. However we may wish to make different prior assumptions about the distribution of SNP effects. For example, there may be some SNP in high linkage disequilibrium with QTL of moderate to large effect. Further, for some regions of the genome there may be no QTL affecting the trait at all, and in those regions SNP effects should be zero.

If we adopt a Bayesian approach, we can capture our prior knowledge that there are some chromosome segments containing QTL of large effects, some segments with moderate to small effects, and some segments with no QTL at all when we estimate the effects of haplotypes (or single markers) within the chromosome segments.

Using Bayesian models allows us to incorporate such prior assumptions into our analysis.

3.5.1 Bayesian statistics refresher

Bayes theorem uses a simple rule about conditional probabilities

$$P(x \mid y) = P(xandy) / P(y) = P(y \mid x)P(x) / P(y)$$

This can be understood with an example. Suppose I have a jar of coins in which 99% are fair coins and 1% are double headed coins. I take a coin at random and toss it three times and observe three heads. What is the probability the coin is a double headed coin? Let y = the data, eg. 3 heads from 3 tosses, x is this is a double headed coin, x' this is a fair coin. Then P(x)=0.01,P(x')=0.99, P(y|x)=1.0 and P(y|x') =0.125 (eg. 0.5^3). Then the outcomes of the experiment can be represented in a table:

	P(x or x')	P(y|x or x')	P(y|x)*P(x)
Fair coin	0.99	0.125	0.124
Double headed coin	0.01	1.0	0.01
P(y)			0.134

Therefore the probability that this is a double headed coin given I observed three heads from three tosses is $P(x \mid y) = P(y \mid x)P(x) / P(y)$ =1.0*0.01/0.134 = 0.075. That is despite the outcome of three heads there is only a small probability of the coin being double headed because doubled headed coins are so rare.

Bayes theorem is useful because often it is easy to calculate P(y|x), while it is more difficult to calculate P(x|y), as in the above example.

After the experiment has been done, the P(y) will be a constant in all calculations we do. So we can also write Bayes theorem as

$$P(x \mid y) \propto P(y \mid x)P(x)$$

Where the symbol $\propto$ indicates is proportional to. This is useful because the calculation of P(y) may be difficult.

The probability P(x|y) is called the posterior probability because it is the probability after the experiment has been done. It is calculated from two terms. P(y|x) is the likelihood used by frequentists. P(x) is called the prior probability because it is the probability of x before the experiment was conducted. This allows us to incorporate prior knowledge into the estimate of x.

In practise, calculating the posterior distribution (and integrating out nuisance parameters) may be difficult to do. Often it is impossible to find a formula that gives the solution. Bayesians have developed a number of approaches to overcome this problem.

- Choose priors that make the algebra easy. So called conjugate prior distributions have the property that, when combined with a particular distribution for the data, they yield a recognised distribution for the posterior. For instance if the data are normally distributed, and a normal prior is used for a parameter affecting the data, then the posterior distribution of that parameter will be normally distributed.
- Numerical integration. If you can calculate the height of the posterior distribution at every point, you can integrate it over nascence parameters using numerical integration such as Simpsons rule.
- Simulation. If you can draw samples from the posterior distribution, you can use the samples to approximate the distribution. For example the mean of many samples is a good approximation to the mean of the distribution. This is what Markov Chain Monte Carlo (MCMC) methods such as Gibbs sampling do.

3.5.2 Bayesian method with a prior that assumes many QTL have a small effect and few have a large effect (BayesA)

One possible assumption about the distribution of SNP effects is that they follow a Student's *t* distribution, rather than a normal distribution. A *t* distribution has a larger probability of moderate to large effects ("a thicker tail") than a normal distribution, Figure 3.1.

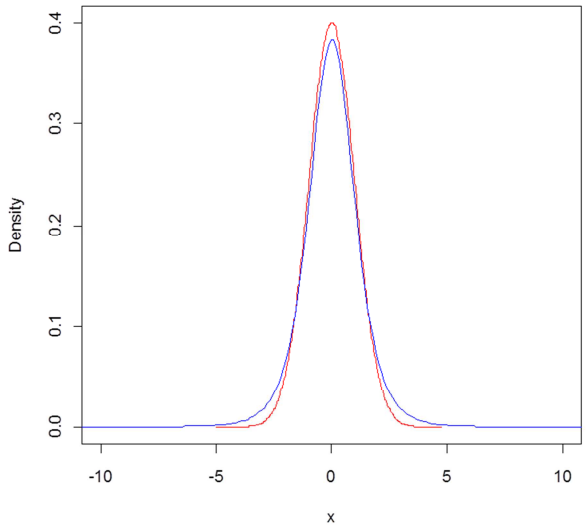


Figure 3.1 . A *t* distribution (blue line) has a higher probability of moderate to large effects than a normal distribution (red line), that is it has "thicker tails".

Unfortunately, t distributions are not as straightforward to incorporate into our predictions of marker effects as the normal distribution was. One mathematically tractable way of incorporating a *t* distribution is to assume each SNP effect comes from a normal distribution, but the σ_g^2 can be vary among the SNP. If σ_g^2 is large, then $\hat{g}$ can be large, if σ_g^2 is small, then $\hat{g}$ is likely to be small as it will be regressed back towards zero.

This leads to a hierarchical model, with one model at the level of the SNP effects and one model at the level of the variances across the SNP. Meuwissen et al. (2001) termed this approach Bayes A.

The first model is at the level of the data, and is similar to before:

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

Using the Bayesian approach, we want the posterior distribution of μ and **g**, given the data **y**, and we will get this from the likelihood of the data **y** given the parameters μ and **g**, multiplied by the priors of μ and g, $P(\mathbf{g}, \mu \mid y) \propto P(y \mid \mathbf{g}, \mu) P(\mathbf{g}, \mu)$.

In Meuwissen et al (2001), the prior distribution of the mean μ was uniform and uninformative, while the prior distribution of SNP effects (actually haplotype effects in their case) *i* was $\mathbf{g_i} \sim N(0, \sigma^2_{gi})$. Note that this is equal to BLUP estimation of the chromosome segment effects with different variances for each segment:

$$\begin{bmatrix} \hat{\mu} \\ \hat{\mathbf{g}}_1 \\ . \\ \hat{\mathbf{g}}_p \end{bmatrix} = \begin{bmatrix} \mathbf{1_n'1_n} & \mathbf{1_n'X_1} & . & \mathbf{1_n'X_p} \\ \mathbf{X_1'1_n} & \mathbf{X_1'X_1} + \mathbf{I}\frac{\sigma^2_e}{\sigma^2_{g1}} & . & \mathbf{X_1'X_p} \\ . & . & . & . \\ \mathbf{X_p'1_n} & \mathbf{X_p'X_1} & . & \mathbf{X_p'X_p} + \mathbf{I}\frac{\sigma^2_e}{\sigma^2_{gp}} \end{bmatrix}^{-1} \begin{bmatrix} 1_n'y \\ X_1'y \\ . \\ X_p'y \end{bmatrix}$$

The prior distribution of the error variance σ^2_e was $\chi^{-2}(-2, 0)$, which yields an uninformative prior (eg the prior receives little or no weight in the calculation).

The second level of model is at the variances of chromosome segment effects. In Meuwissen et al (2001), the prior distribution of the variances of effects across chromosome segments was chosen to be result in a *t* distribution at the level of the SNP effects consistent with many QTL of small effect and few of large effect. The prior distribution was used the scaled inverted chi-square distribution, $\Pr ior(\sigma^2_{gi}) \sim \chi^{-2}(v, S)$, where *S* is a scale parameter and v is the number of degrees of freedom. The values of *v* and *S* were chosen as v=4.012 and S =0.002 [these values were chosen to give a distribution similar to what would be expected from the distribution of QTL effects derived by Hayes and Goddard (2001) and the expected heterozygosity of QTL under the neutral model].

The posterior distribution of σ^2_{gi} combines information from the prior and the data. Information from the data is included by conditioning on the chromosome segment effects, eg. $P(\sigma^2_{gi} \mid \mathbf{g_i})$. An advantage of using an inverted chi-square distribution as a prior for the variances is that with normally distributed data, the posterior is also inverted chi-squared (a *conjugate* prior). In fact if the prior for our chromosome segment variances has the scale parameter *S*, and degrees of freedom *v*, then the posterior for σ^2_{gi} given the chromosome segment effects, $P(\sigma^2_{gi} \mid \mathbf{g_i})$ is an inverted chi-squared scaled by $S+\boldsymbol{g_i'g_i}$ and $v+n_i$ degrees of freedom:

$$P(\sigma^2_{gi} \mid \mathbf{g_i}) = \chi^{-2}(v + n_i, S + \mathbf{g_i}'\mathbf{g_i})$$

where n_i is the number of haplotype effects at segment *I*, or 1 if when a single effect is estimated for each SNP.

We cannot use this posterior distribution directly for estimating the σ^2_{gi} because it is conditional on the unknown $\boldsymbol{g_i}$ effects. Likewise, the values of $\boldsymbol{g_i}$ depend on σ^2_{gi} . Meuwissen et al. (2001) therefore used Gibbs sampling to estimate effects and variances. In Gibbs sampling, samples for each parameter are taken from the posterior distribution of that parameter, conditional on all the other parameters.

The Gibbs chain could proceed as follows:

Step 1. Initialise the vectors of haplotype effects for each vector of chromosome segment effects $\mathbf{g_i}$ for j=1,n_i where n_i is the number of haplotypes at the chromosome segment, with a small positive number. The overall mean μ must also be initialised.

Step 2. Update the σ^2_{gi} for the i^{th} chromosome segment by sampling it from the fully conditional distribution $\chi^{-2}(v + n_i, S + \mathbf{g_i}'\mathbf{g_i})$, where v is 4.012 and S is 0.002, and n_i is the number of haplotype effects at the i^{th} chromosome segment.

Step 3. Given the $\mathbf{g_i}$ and μ calculate the values for $\boldsymbol{e}$ as $\mathbf{e} = \mathbf{y} - \mathbf{Xg} - 1_n'\mu$, where $\mathbf{X}$ = [$\mathbf{X_1}$ $\mathbf{X_2}$ $\mathbf{X_3}$ ...] is the design matrix of all haplotype effects; and $\boldsymbol{g}$ is a vector of all haplotype effects across chromosome segments. Then update the error variance, σ^2_e by drawing a single sample from $\chi^{-2}(n-2, \mathbf{e_i}'\mathbf{e_i})$

Step 4. Sample the overall mean μ given the updated error variance from a normal distribution with mean $\frac{1}{n}\left(\mathbf{1_n}'\mathbf{y} - \mathbf{1_n}'\mathbf{Xg}\right)$ and variance σ^2_e / n, where $\mathbf{X}$ = [$\mathbf{X_1}$ $\mathbf{X_2}$ $\mathbf{X_3}$ ...] is the design matrix of all haplotype effects; and $\boldsymbol{g}$ is a vector of all haplotype effects.

Step 5. Sample all the haplotype effects g_{ij} given the newly sampled μ, σ^2_e and σ^2_{gi} from a normal distribution with mean $\frac{\mathbf{X}_{ij}^{'}\mathbf{y} - \mathbf{X}_{ij}^{'}\mathbf{X}\mathbf{g}_{(ij=0)} - \mathbf{X}_{ij}^{'}\mathbf{1}_n\mu}{\mathbf{X}_{ij}^{'}\mathbf{X}_{ij} + \sigma_e^2/\sigma_i^2}$, where $\boldsymbol{X_{ij}}$ is column of $\boldsymbol{X}$ of effect g_{ij}; $\mathbf{g}_{(ij=0)}$ equals $\boldsymbol{g}$ except that the effect of g_{ij} is set to zero, and variance $\sigma_e^2/(\mathbf{X}_{ij}^{'}\mathbf{X}_{ij} + \sigma_e^2/\sigma_{gi}^2)$.

Step 6. Repeat Step 2 (using the updated $\mathbf{g_i}$) to Step 5 for a large number of cycles.

Other authors have published similar methods but with different priors used for the variance of chromosome segment effects**.** In Xu (2003) this was $1/\chi_0^2$ (eg. an inverted chi-square distribution with 0 degrees of freedom). Xu (2003) also described their method for single SNP markers, rather than marker haplotypes. Therefore the matricies $\mathbf{X_i}$ are the design matricies for the effect of a single marker, so X_{ij} =1 if the i^{th} SNP genotype for individual *j* is a_1a_1, X_{ij}=0 if the i^{th} SNP genotype for individual *j* is a_1a_2, and X_{ij}=-1 if the i^{th} SNP genotype for individual *j* is a_2a_2. The implicit assumption in Xu (2003) is that the partial regression coefficient, g_i, (the effect of marker i on the trait), will absorb partly the effects of all QTL located between markers i-1 and i+1. The validity of this assumption will depend on the LD between the markers and the QTL.

Ter Braak et al. (2005) argued that prior used by Xu (2003) would result in an improper posterior distribution, in particular a posterior of g_i with infinite mass near zero. To ensure a valid posterior, they altered the prior distribution of variance of chromosome segment effects to be $1/\chi^2_{-0.002}$.

Xu (2003) actually proposed their method for QTL mapping rather than genomic selection, claiming that the method gave more precise estimates of QTL location than single QTL models. This was because the effect of a QTL was removed in adjacent marker brackets so the QTL were mapped to a smaller interval. The approach also gave more accurate estimates of QTL effect, as the problem of over-estimating the QTL effect due to multiple testing were avoided.

3.5.3 Bayesian method with a prior that assumes many SNP have a no effect, some have a moderate to large effect (BayesB)

Another possible assumption about the distribution of SNP effect is that many SNP are not in genomic regions containing QTL, and therefore have zero effect, while some SNP may be in LD with QTL having a moderate to large effect. The prior BayesA does not reflect this, the prior does not have a density peak at $\sigma^2_{gi} = 0$; in fact its probability of $\sigma^2_{gi} = 0$ is infinitesimal. Meuwissen et al. (2001) addressed this in their Method BayesB. The prior distribution of SNP effects in BayesB is a mixture distribution with many SNP with zero effect, and the rest with a *t* distribution of effects. Method BayesB used a prior that has a high density, π, at $\sigma^2_{gi} = 0$ and has an inverted chi-square distribution for $\sigma^2_{gi} > 0$; . The prior distribution was

$$\sigma^2_{gi} = 0 \text{ with probability } \pi,$$
$$\sigma^2_{gi} \sim \chi^{-2}(\nu, S) \text{ with probability } (1-\pi),$$

where ν= 4.234 and S = 0.0429 yield the mean and variance of σ^2_{gi} given that $\sigma^2_{gi} > 0$ (see Meuwissen et al. 2001 for derivation of *v* and *S* values).

Figure 3.2 Illustrates the difference between the prior distribution of variances of chromosome segment effects used in method Bayes B and that used in method BayesA.

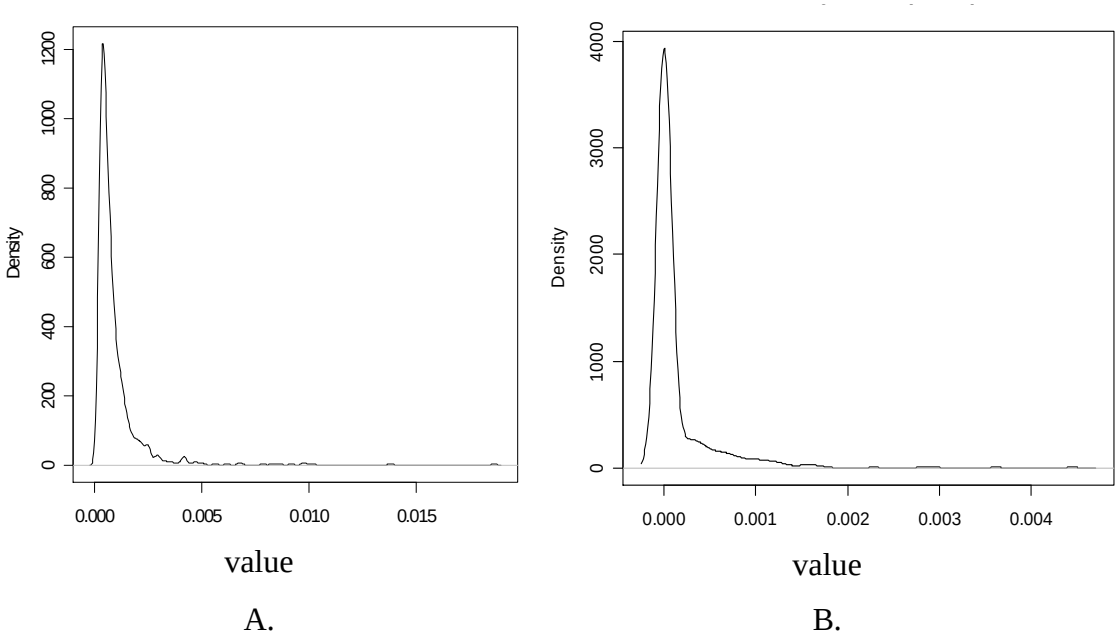


Figure 3.2 A. Prior distribution of variances of chromosome segment effects used in method BayesA, and B. Prior distribution of variances of chromosome segment effects used in method BayesB in Meuwissen et al. (2001), for 20% of chromosome segments containing QTL.

The figure illustrates the infinitesimal density of the prior used in BayesA at 0, and the much higher mass near (and actually at) zero for the prior used in BayesB. The Gibbs sampler described in Method BayesA cannot be used in method BayesB, as it will not move through entire sampling space. This is because the sampling of $\sigma^2_{gi} = 0$ from the posterior distribution of Var(of σ^2_{gi}) is not possible if $\boldsymbol{g}'_i\boldsymbol{g}_i > 0$, which it will never be as $\mathbf{g_i} = 0$ has an infinitesimal probability if $\sigma^2_{gi} > 0$. This problem was resolved by sampling σ^2_{gi} and $\boldsymbol{g_i}$ simultaneously using a Metropolis-Hastings algorithm (see Meuwissen et al. 2001 for details).

3.5.4 Other assumptions for the distribution of SNP effects

Another possible prior assumtion for the distribution of QTL effects is that they follow a double exponential distribution – very many of the SNP effects are very close to zero. This method was developed by Yi and Xu (2008) and was called the BayesianLASSO.

Another method is similar to BayesB, in that it allows some of the SNP effects to be zero, but assumes that for the SNPs with non zero effects these effects follow a normal distribution (Habier et al. 2011). This method had two potential advantages over BayesB – the proportion of SNP that had zero effect was estimated from the data, rather than assumed, and secondly there were multiple degress of freedom to estimate the variance of the normal distribution from which SNP effects were derived, rather than one per SNP as in BayesB, although the effect of this at the level of SNP effects may not be that pronounced.

3.6 Comparison of accuracy of methods of genomic prediction

Meuwissen et al. (2001) evaluated their methods (least squares, BLUP, Bayes A and Bayes B), using simulation. A genome of 1000 cM was simulated with a marker spacing of 1 cM. The markers surrounding every 1-cM region were combined into marker haplotypes. Due to finite population size (N_e = 100), the marker haplotypes were in linkage disequilibrium with the QTL located between the markers. The effects of the chromosome segments were predicted in one generation of 2000 animals, and the breeding values for the progeny of these animals were predicted based only on the markers which they carried, Table 3.1.

Table 3.1. Comparing estimated *vs.* true breeding values in progeny with no phenotypic records (from Meuwissen et al. (2001). Chromosome segments were estimated in a population of 2000 animals.

	$r_{TBV;EBV}$ + SE	$b_{TBV.EBV}$ + SE
LS	0.318 ± 0.018	0.285 ± 0.024
BLUP	0.732 ± 0.030	0.896 ± 0.045
BayesA	0.798	0.827

BayesB	0.848 + 0.012	0.946 + 0.018

Mean of five replicated simulations LS, least squares; BLUP, best linear unbiased prediction; Bayes, Bayesian method with inverse chi-square prior distribution and where the prior density of having zero QTL effects was increased; $r_{TBV;EBV}$, correlation between estimated and true breeding values (equals accuracy of selection); $b_{TBV;EBV}$, regression of true on estimated breeding value.

The least squares method does very poorly, primarily because the haplotype effects are over-estimated. The increased accuracy of the Bayesian approach occurs because this method sets many of the effects of the chromosome segments to close to zero in BayesA or zero in BayesB, and "shrinks" the estimates of effects of other chromosome segments based on a prior distribution of QTL effects.

In real data, large differences in the accuracy of BLUP, BayesA, BayesB and the other methods have not been observed. For example, Verbyla et al (2009) compared the accuracy of GEBV from BayesA, BLUP and BayesSSVS, which is very similar to BayesB, for three traits in dairy cattle, Protein kg, Fat% and Protein%. The data were 1800 bulls genotyped for 39,000 SNP markers. The phenotypes of the bulls were the average of their daughters performance for the trait. Accuracy of the methods was approximated by removing the youngest bulls from the data set when the prediction equation was derived. Then GEBV was calculated for these bulls, and correlated with their progeny test values.

Table 3.2. Correlation and Regression Coefficient between predicted GEBV and EBV in the validation data set

Method	Measure	Bayes SSVS*	Bayes A*	BLUP*
Protein kg	$\rho_{DGV,ABV}$	0.58	0.57	0.60
	$b_{ABV,DGV}$	0.99	1.00	1.06
Fat kg	$\tau_{DGV,ABV}$	0.56	0.53	0.56
	$b_{ABV,DGV}$	0.90	0.86	0.99
Protein %	$\tau_{DGV,ABV}$	0.67	0.64	0.66
	$b_{ABV,DGV}$	0.97	1.00	0.89
Fat %	$\tau_{DGV,ABV}$	0.74	0.72	0.65
	$b_{ABV,DGV}$	0.87	0.86	0.93

*Average accuracies reported over validation sets from years 2005, 2006,

2007. $\tau_{DGV,ABV}$ Correlation coefficient between the EBV and predicted DGV, $b_{ABV,DGV}$ Regression coefficient of the EBV on predicted DGV

The accuracy of the methods was surprisingly similar for most traits, except for fat%. This is probably because a mutation with large to moderate effect, in the DGAT1 gene (Grisart et al. 2002) segregates for this trait. Both BayesA and BayesB would not shrink the effect of this large mutation as severely as BLUP, and so the GEBV are more accurate.

3.7 Factors affecting the accuracy of genomic selection

While the simulations, and now real results, demonstrate genomic selection has huge potential to increase rates of genetic gain, several key questions remain regarding its implementation. These are

1) How many markers are required, determined by the extent of LD.
2) How many phenotypic records are required in the initial experiment estimating the effect of chromosome segments

To address the first question, the lower the LD the more SNPs will be required to ensure at least one SNP is in LD with each QTL. Calus et al. (2008) demonstrated that provided r^2 (a commonly used measure of LD) between adjacent SNPs was on average greater than 0.2, accurate genomic breeding values could be predicted. In Holstein Friesian (Black and white) cattle r^2 of 0.2 occurs at approximately 100kb, implying 30,000 markers should be sufficient to apply genomic selection. The extent of genome wide LD is largely determined by the past effective population size. The expectation of r^2 is $\frac{1}{4N_e c+1}$ where N_e is effective population size and c is the distance between loci in Morgans (Sved 1971). Meuwissen (2009) demonstrated by

simulation that to achieve very accurate genomic estimated breeding values, $10*N_e*L$ markers are required, where L is the length of the genome in Morgans. In Holstein Friesian cattle, N_e is approximately 100 and the length of the genome is 30 Morgans, again suggesting 30 000 markers are required. As the results in real data above suggest, 30 000 markers is indeed sufficient to predict accurate breeding values in Holstein Friesian cattle. In other species with large effective population sizes, larger numbers of markers will be required.

Provided the markers are dense enough, the accuracy of genomic breeding values will depend on the number of individuals genotyped and phenotyped in the reference population, the heritability of the trait, and the number of loci affecting the trait (Goddard 2008; Daetwyler et al. 2008). Given that there is little knowledge of the number of loci affecting the vast majority of traits important in livestock, a conservative assumption is that the number of loci is equal to the number of independent chromosome segments in the population. This can be derived from the effective population size and the length of the genome as $q=2N_eL$. (Goddard 2008, Hayes et al 2009). Then the accuracy of genomic breeding values for individuals with no phenotypes of their own is $r = \sqrt{\left[1 - \lambda/(2N\sqrt{a}) * \ln((1 + a + 2\sqrt{a})/(1 + a - 2\sqrt{a})\right]}$ where $a= 1+2\ \lambda/N$, and $\lambda= qk/h^2$, with $k = 1/log(2N_e)$, where h^2 is the heritability of the trait and N is the number of phenotypic records in the reference population (Goddard 2008). This deterministic prediction suggests large reference populations are required to predict accurate genomic estimated breeding values, particularly for low heritability traits, Figure 3.3. The deterministic predictions agree well with accuracies that have been achieved in dairy cattle experiments (Hayes et al. 2009).

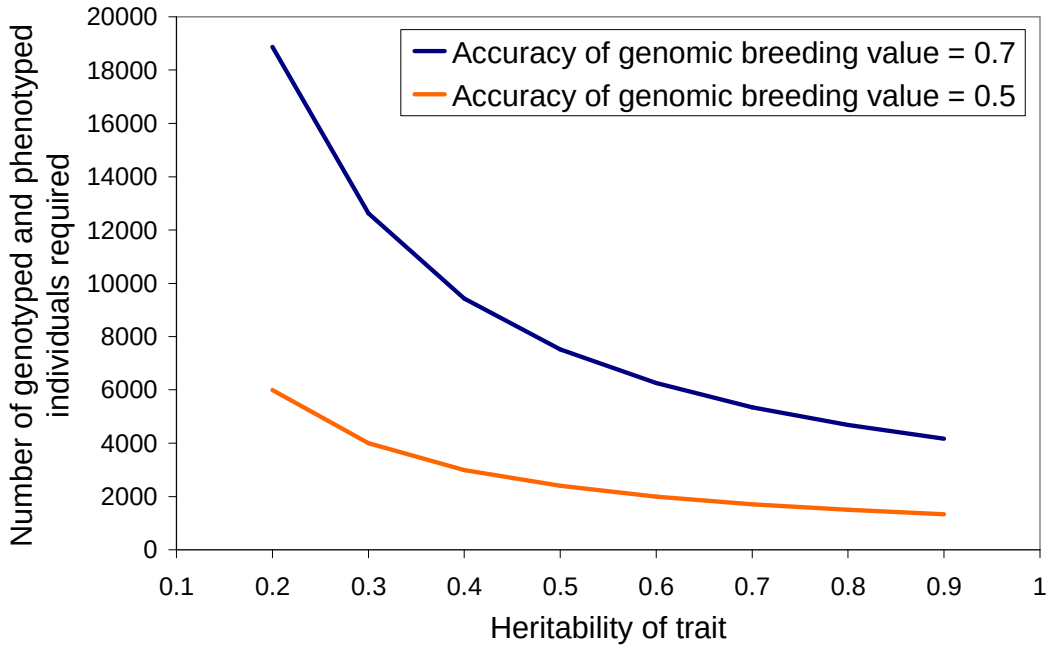


Figure 3.3. Number of genotyped and phenotyped individuals required in the reference population to reach a desired accuracy of genomic breeding value (for un-phenotyped individuals) N_e was 100.

3.8 Genomic selection across populations and breeds

In practise Genomic selection is always applied in a population that is different to the reference population where the marker effects are estimated. It might be that the selection candidates are from the same breed, but are younger than the reference population, or they could be from a different selection line or breed. Genomic selection relies on the phase of LD between markers and QTL being the same in the selection candidates as in the reference population. However as the two populations diverge, this is less and less likely to be the case, especially if the distance between markers and QTL is relatively large. In section 1.5 we used the correlation between r in two populations, corr(r_1,r_2), to assess the persistence of LD across populations. No if the chromosome segment effects are estimated in population 1, and GEBVs in that population can be predicted with an accuracy x_1, then the GEBVs of animals population 2 may be predicted from the chromosome segment effects of population 1 with an accuracy $x_2 = x_1$*corr(r_1,r_2). For each set of populations, one can work out the marker density that is required to obtain a corr(r_1,r_2) = 0.9 (De Roos *et al.* 2008).

In the above, we have assumed that effect of QTL alleles are similar in different breeds and populations. For some QTL which have been traced to known mutations, the alleles do act reasonably similarly in different breeds and populations. For example, the A allele of the DGAT1 gene results in increased fat yield and reduced protein yield and milk volume in New Zealand Holstein-Friesians, Jersey's and Ayshires (Spelman *et al.* 2002). However while the size of the effects are consistent for protein and milk volume in the Holstein-Friesian and Jersey breeds, the size of the fat response in Holstein-Friesians is nearly double that for Jerseys (Spelman et al. 2002). Another problem is that we have assumed that the same mutations affecting production traits are polymorphic in different breeds. This is true for some well characterised mutations such as the K232A mutation in DGAT1, which is polymorphic in Holsteins, Jerseys, Aryshires and some *Bos indicus* breeds (Spelman *et al.* 2002). Other mutations, such as some of the functional mutations in the myostatin gene, appear to breed specific (Dunner *et al.* 2003). One solution would be to use a multi-breed reference population, so that all the genetic variants are captured.

In practise, the observed increases from using multi-breed reference populations have been small (eg Erbe et al. 2012). One possibility is that the markers are not yet dense enough to be in the same phase with the QTL across breeds – this is a justification for using sequence data as described in chapter 5.

Finally, genotype by environment interaction may also reduce the accuracy of predicted GEBV when the chromosome segment effects are estimated from animals in another population.

3.9 How often to re-estimate the chromosome segment effects?

If the markers used in genomic selection were actually the underlying mutations causing the QTL effects, the estimation of chromosome segment effects could be performed once in the reference population. GEBVs for all subsequent generations could be predicted using these effects. A more likely situation in practise is that there will be markers with low to moderate levels of r^2 with the underlying mutations

causing the QTL effect. Over time, recombination between the markers and QTL will reduce the accuracy of the GEBV using chromosome segment effects predicted from the original reference population. Meuwissen et al. (2001) used simulations to investigate the change in accuracy of GEBV with an increasing number of generations between the reference population and the population for which GEBV were estimated, Table 3.3.

Table 3.3. The correlation between estimated and true breeding values in generations 1003–1008, where the estimated breeding values are obtained from the BayesB marker estimates in generations 1001 and 1002. From Meuwissen et al. (2001).

Generation	$r_{TBV;EBV}$
1003	0.848
1004	0.804
1005	0.768
1006	0.758
1007	0.734
1008	0.718

The generations 1004–1008 are obtained in the same way as 1003 from their parental generations.

After five generations, the decline in accuracy of GEBV was large. This suggests that with the levels of LD simulated in Meuwissen et al. (2001), re-estimation of the chromosome effects should take place every 3 generations.

De Roos et al (2008) investigated the same issue using real SNP data from both Dutch and Australian Holstein Bulls. They calculated the correlation of r values at different marker distances for sub-divisions of the same population across time, as an indicator of persistency of marker-QTL phase across generations. They found correlation of r values between Dutch Holstein bulls before 1995 and Dutch Holstein calves born in 2006 is 0.9 at 135kb. They concluded from this data that with 20,000 markers, the predictions of chromosome segment effects should be usable for two generations, as accuracy will be reduced only slightly (by a factor 0.9) by breakdown of LD phase over this time.

More results are needed in real data to determine how often SNP effects or chromosome segment effects should be re-estimated in practise.

3.10 Validation of genomic predictions

Measures of prediction accuracy *(adapted from (Daetwyler et al. 2013))* . The term accuracy refers to different statistical properties of an estimator or a predictor. The correlation between estimated and true breeding values has a linear relationship with the response to selection. Therefore correlation has emerged as the most commonly used metric to assess prediction accuracy. Most of the models used in genomic selection are designed to predict breeding values; therefore, the predictand should be the true breeding value. However, true breeding values are generally only available in simulation studies. Therefore, an important decision to be made is what should be the predictand in real-data studies. Some of the most commonly used predictands are: individual phenotypes (raw or adjusted for factors such as fixed effects), averages of offspring performance (e.g. daughter yield deviations in dairy cattle or progeny means in poultry), and estimated breeding values (EBV). Different predictands contain different signal-to-noise ratios and this requires consideration when assessing an estimate of predictive performance. A common practice to accommodate this problem is to divide the estimated correlation by the square root of the heritability of the predictand, $\sqrt{h^2}$, or more in general, by the square root of the proportion of variance of the predictand that can be attributed to additive effects (e.g. accuracy of bull daughter trait deviations or deregressed proofs).

In pedigree animal models, individual accuracies are calculated from the prediction error variance (PEV) using $r_{ind} = \sqrt{1 - PEV[VarG]^{-1}}$ (or approximations thereof (e.g. Misztal & Wiggans 1988; Hickey *et al.* 2009)), where VarG is the additive genetic variance. REML methods using a genomic relationship matirx such as GBLUP, are quite sensitive to the population structure in the sample and their allele frequencies. The main issue is that numerator relationship matrices and genomic relationship matrices assume different base populations and this may affect the estimation of the variance components. Setting the two matrices to the same numerical scale has received some attention mostly to allow fitting them together in

the same model or within creation of the H matrix in OneStep genomic selection (e.g. Christensen & Lund 2009; Aguilar *et al.* 2010). This will partly correct for the problem. However, if there are multiple breeds or diverse populations (e.g. Africans and Europeans, heterotic groups in maize) in the sample, additional adjustements of the genomic relationship matrix may be necessary to get all subgroups adjusted to the same base population (Erbe *et al.* 2012). Further research is needed to correctly scale the genomic relationship matrix in heterogeneous reference populations. Caution is advised when using accuracies from PEV.

Another important metric is the slope of the regression of true on estimated breeding values. If this slope deviates from its expection, which is usually 1, it is called bias. Biased estimated breeding values are an issue where individuals are given mating contributions that are proportional to their estimated breeding values, or where pedigree and genomic information is combined to produce one breeding value. In all cases, it is important to investigate the slope and intercept of the regression of observations on predictions as well as their expectations, because departures from expected values should point to deficiencies of the model.

Deciding on the target of prediction. *(adapted from (Daetwyler et al. 2013)).* The ultimate target individuals of genomic prediction are the selection candidates, but their accuracy of prediction cannot be computed due to the lack of predictands (e.g. phenotypes). Hence, a testing population needs to be selected, which requires giving thought to a number of factors. Likely the most important principle of selecting a testing population is that it should mimic the relationship of the selection candidates to the training population. Relatedness is an important component of prediction accuracy, as pointed out above. If the testing population is more related to the training population than the selection candidates, then the estimate of prediction accuracy will inflated. For example, in a training-testing scheme, it is not adequate to test the accuracy only in individuals one generation removed from the training population, if the selection candidates are mostly grand progeny. Similarly, in replicated cross-validation, the manner in which individuals are assigned to particular folds affects accuracy. Drawing random subsets is simple to implement, but if full and half-sib families are present in the reference population then prediction implicitly contains a within family component which increases accuracies. Achieved accuracy

may be significantly lower than within family accuracy if individuals in selection candidates do not share full or half-sib families (Legarra *et al.* 2008). A more rigorous test would be to randomly assign whole families to subsets to make prediction explicitly across families. Being cognizant of the impact of relationships on the accuracy of genomic estimated breeding values allows cross-validation procedures to be modified so that the accuracy can be calculated within and across groups of individuals such as families, generations, genetic groups, strains, lines and breeds. Saatchi et al. (2011) proposed an approach for designing cross-validation schemes that uses k-means clustering based on genomic relationships to partition the data into the various folds to minimize the relationships between training populations and testing populations.

It is also important to be cognizant of the presence and effect of population structure (e.g. breeds, lines of common origin) when designing the testing scheme. While genomic selection can make use of otherwise unknown structure to increase the response to selection, similar to applications in associations mapping (e.g. Pritchard *et al.* 2000), it is more often the case that the structure is already captured by some other means (breeders knowledge or pedigree information for example) (Malosetti *et al.* 2007). The accuracy of a structured dataset may be higher than the accuracy within its subgroups, because the 'structured data' accuracy contains a component discerning individuals based on mean genetic level of each subgroup. If the GEBV are going to be used to make selection decisions within family (i.e. chose between a number of full sibs on the basis of their Mendelian sampling terms), an effort should be made to obtain the accuracy with which this decision can be made.

Some studies have attempted to evaluate the accuracy of the estimation of the Mendelian sampling term. For example (VanRaden *et al.* 2009; Lund *et al.* 2011; Wolc *et al.* 2011) compared the accuracy of estimated breeding values predicted from parent average or genomic information. If the accuracy of the parent average is high (close to its limit of $\sqrt{0.5}$) then any increase in accuracy must relate mostly to the Mendelian sampling term (Daetwyler *et al.* 2007). If the accuracy of the parent average is low, then genomic information may be useful for predicting parent average as well as Mendelian sampling, so the distinction becomes less important. Mendelian

sampling term accuracy can also be predicted by comparison of accuracies of GEBVs predicted from average genotypes of the parents and actual individual genotypes, as shown by Wolc et al. (2011), or by correlating the residuals of GEBV and predictand when both are corrected for the parent average estimated breeding values. In the future the contribution of genomic information to evaluating the accuracy of the Mendelian sampling term needs to become more prominent in the validation of genomic prediction. For example, validation data sets could be created which contain several (e.g. 50) full sib families with each of these full sib families comprising several (e.g. 30) individuals. Plant breeding data sets may be particularly suited to this purpose because large numbers of full sibs can easily be generated.

Regardless of the applied testing strategy, comparison with accuracies obtained with pedigree based models (if available) is generally a reasonable approach to assess the additional accuracy obtained from using marker information on top of pedigree information. This difference may be evaluated at the level of reliabilities (accuracy squared), since this is a measure of the additional variance explained by the markers, on top of the variance explained by the pedigree based model. It should be noted that an accuracy obtained by testing using the Pearson correlation is never 'context-free' and this makes comparison of accuracies across studies difficult.

Common pitfalls of validation *(adapted from (Daetwyler et al. 2013; Wray et al. 2013)* The main pitfall of validating genomic prediction accuracy is the failing to ensure that the training and testing populations are independent. In this context, independence does not mean unrelated but that the information used to calculate the observations (i.e. daughter trait deviations, EBV, deregressedEBV) did <u>not</u> include phenotypic information from the testing population. As discussed earlier, the testing population should mimick the target population or selection candidates. The main aim of genomic selection is to predict (young) individuals that do not have phenotypes. Thus, to ensure proper validation, phenotypes of testing individuals should not contribute to the training observations. In this section we discuss various ways of falling into the non-independence trap, which lead to inflated genomic selection accuracy. The fundamental principle is to set up the validation as close as possible to the way genomic selection will be applied in a particular breeding program.

Case 1. Observations of training and testing population from same genetic evaluation

Prediction accuracies may be biased upwards when the phenotypes used to estimate the genomic breeding values are also included in calculation of adjusted progeny means or when estimated breeding values for training and testing that are obtained from the same evaluation (e.g. Amer & Banos 2010). One example is using progeny phenotypes we wish to predict for validation when calculating progeny means of parents in the reference, resulting in upwardly biased accuracies. A particularly bad variation of this occurs when using EBVs with low accuracy as observations from a genetic evaluation of all individuals. Here the EBV of training and selection candidates is heavily dependent on their relatives. In this case, the accuracy you calculate as r(GEBV, EBV) will be the accuracy of predicting the parent average and will contain very little accuracy due to Mendelian sampling. This is a poor measure of the efficacy of genomic selection. Another example is using phenotypes of contemporaries of testing individuals (e.g. same generation and age) to calculate the observations in the training population. This situation would not occur in a real breeding program and thus the accuracy attained is not realistic.

Case 2. Selecting subsets of loci based on GWAS in all data

A guiding principle and one of the main merits of genomic selection is its use of all loci to predict a GEBV. Nevertheless, it may be desireable to reduce the number of loci in genomic selection due to genotyping cost or to reduce the accumulation of errors associated with estimating many effects. The latter may become relevant when using sequence data for genomic selection, where the majority of variants is expected to have no effect. One simple way to choose a subset is with a GWAS. If both training and testing individuals are used to select the most significant variants for genomic selection, then the accuracy will be inflated even if the testing phenotypes are excluded from genomic selection subsequently (see Figure below). This is also called overfitting and is again due to phenotypes of testing individuals contributing to the model.

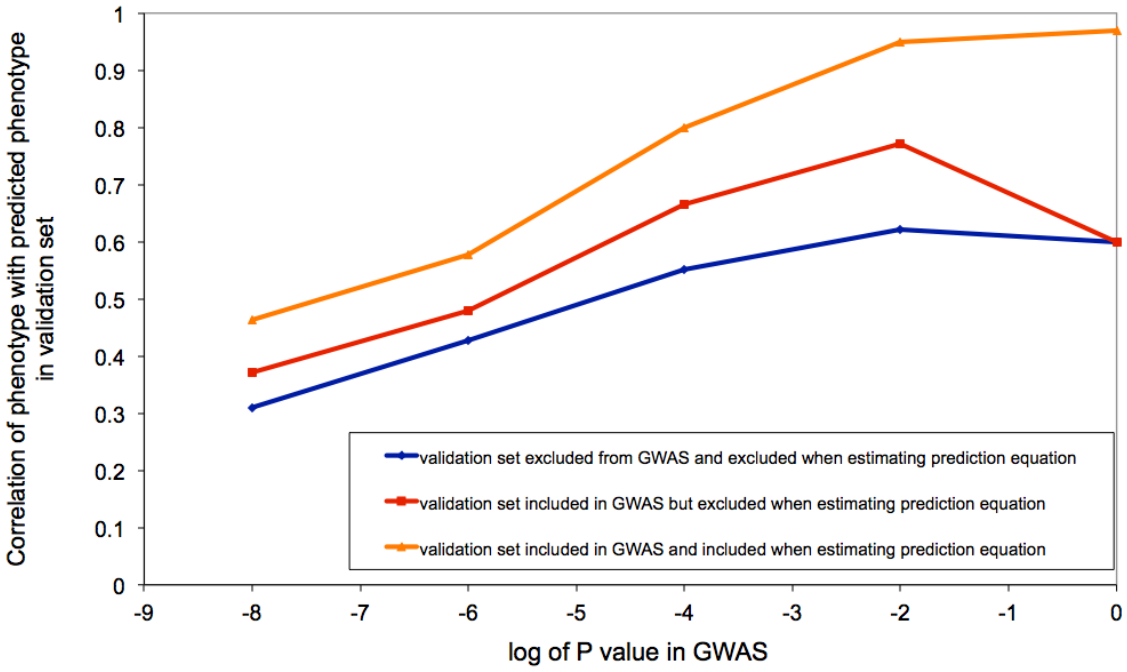


Figure from Wray et al (2013): Example of Case 2: overlap of discovery and testing samples. An example using dairy cattle data to show the impact of leaving the testing cohort in the discovery set, either at both SNP selection (GWAS) and SNP effect estimation stages or at the effect size estimation stage only, leads to considerable bias. Data were on 2,732 dairy bulls with ~500K SNPs phenotyped for average milk yield of their daughters' milk production. The bulls were split into a discovery sample (bulls born during or before 2003), N_d = 2,458, and a validation sample (bulls born after 2003) of N_v = 274. As an aside, it also demonstrates that string subset selection based on GWAS leads to lower accuracy than using the whole set of SNP.

3.10 Optimal breeding program design with genomic selection

Adapted from (Pryce & Daetwyler 2012). Genomic selection allows prediction of very accurate EBVs for young individuals. This has substantial implications for the design of breeding schemes. For example in a dairy cattle breeing scheme, rather than waiting until a bull has daughters with phenotypic records, a process that typically takes 5-6 years, young bulls with no progeny can be used as sires. The development of high-throughput genotyping methods and reduced genotyping cost has made the application of genomic selection feasible. Here we concentrate on dairy cattle breeding schemes, with brief reference to other livestock species.

The main dairy genomic breeding schemes. Breeding schemes concentrate on changing three terms in the breeder's equation, $\Delta G = ir\sigma_a [L]^{-1}$, where ΔG is the genetic gain per year, i is the selection intensity, r is the accuracy of selection, σ_a is the genetic standard deviation, and L is the generation interval. Assuming that the genetic variance is constant, one can calculate the rate of genetic gain by increasing the selection intensity and/or the accuracy of selection, or by decreasing the generation interval. Genomic selection can potentially affect all three of these components at various points in the four pathways of selection found in dairy cattle. Most of the studies on breeding scheme design under genomic selection have compared rates of genetic gain and rates of inbreeding to those achieved in conventional progeny-testing schemes to allow for fair comparisons to current rates of genetic gain.

One genomic breeding scheme design that has already gained popularity is partially replacing progeny-testing with genomic selection. Here, young bulls are genotyped and genomic breeding values are used to select and reduce the number of progeny-test candidates. The advantage with this scheme is that the number of bulls entering progeny-testing is reduced, thereby offsetting the cost of genotyping young bulls.

Another approach is to screen a large number of bulls and then select the best 10-20 for widespread use as young sires (Schaeffer 2006; Pryce *et al.* 2010b; Winkelman & Spelman 2010; Buch 2011; de Roos *et al.* 2011; Mc Hugh *et al.* 2011). Most studies assumed that bulls would be genotyped once. The exception was Winkelman and Spelman (2010) who also included schemes where bulls were pre-screened with a low density SNP chip to identify candidates for the full-screen. This second scheme is more aggressive than the pre-screening scheme and eliminates progeny-testing completely.

The most intensive selection intensity in female pathways is likely to be achieved through nucleus breeding schemes. Selection intensity can be increased further and generation intervals reduced by using reproductive technologies such as MOET or

juvenile *in-vitro* embryo transfer (JIVET) and sexed semen (Pedersen *et al.* 2009b; Pryce *et al.* 2010b).

Rates of genetic gain and inbreeding achieved by altering breeding scheme designs. Using a pre-screening breeding scheme it is possible to increase the rate of genetic gain by up to 12% (de Roos *et al.* 2011). Similar results were obtained by Pryce et al. (2010) who used a deterministic model with a reliability of GEBV of 60%. Here the rate of genetic gain achieved was 16% more than a conventional progeny test scheme. The rates of inbreeding per year in PRE-SCREEN ranged from 0.10 and 0.20% and were either very similar or reduced to inbreeding from progeny testing (Buch 2011; de Roos *et al.* 2011; Lillehammer *et al.* 2011). These results show that the rate of genetic gain can be increased or maintained through introducing genomically estimated breeding values (GEBVs), but without making substantial alterations to the design of breeding schemes. Furthermore, the effect on annual inbreeding levels is small because generation intervals remain unchanged from conventional schemes.

Large-scale screening and use of young bulls could replace progeny-testing completely. The models used to estimate rates of genetic gain achievable range between +28% and +108% improvement over progeny-testing (Table 3.4). The rate of genetic gain depends on the number of bulls genotyped versus the number selected as sires (selection intensity), the accuracy of selection and the generation interval. The highest selection intensity was 2.67 (König & Swalve 2009) and was achieved when the top 0.1% of animals were selected. Exactly the same selection intensity and response to selection can be achieved if the screened population is 10,000 and the best 100 are selected, which is probably a more realistic scenario. König and Swalve (2009) assumed that older females would be selected as parents, which is why the generation interval is longer than other schemes. Harris *et al.*(2008) suggested that bulls should not be used widely until two years of age, so that congenital birth defects can be checked. However, reducing the generation interval will result in greater rates of genetic gain, as demonstrated by McHugh *et al*. (2011) who evaluated breeding schemes where bulls were parents at either 2 or 3 years of age. These schemes resulted in the highest rates of inbreeding per year ranging from 0.18 to 0.70%, mainly due to shortened generation intervals (Table 3.4).

Table 3.4. Rates of genetic gain and inbreeding for breeding schemes where young bulls are genotyped and used

Authors	Bulls screened	SC	SI	Reliability	ΔG/year	ΔG as % conv	ΔF/year	ΔF/gen	L
König and Swalve (2009)	50,000	500	2.67	56%	0.31	+44%*			4.60
Pryce et al. (2010)	1,000	20	2.42	60%	0.40	+59%	0.07%	0.20%	2.67
Winkelman and Spelman (2010)	500	10	2.42	52%	0.36	+44%			3.75
Buch (2011)	2,000	30	2.52	50%	0.29	+65%	0.31%	0.74%	2.38
Lillehammer et al. (2011)	750	20	2.31	37%	0.28	+28%	0.18%		3.04
de Roos et al (2011)	1,000	20	2.42	58%	0.50	+108%	0.52%	1.14%	2.20
McHugh et al. (2011)	500	30	1.99	59%	0.34	+100%	0.70%	1.73%	2.48

SC is sires of cows

SI is selection intensity in SC-pathway

ΔG %: is % increase of genomic selection over conventional progeny testing

* compared to the rate of genetic gain of conventional progeny testing of Schaeffer (2006)

De Roos et al. (2011) markers explain 40% gen var

NUCLEUS breeding schemes where the male and female pathways are controlled are another option to structure breeding schemes. Pryce *et al.* (2010) considered a nucleus with 300 females selected for JIVET at 3 months and becoming parents at 1 year of age, 20 sires were selected, becoming parents at 2 years of age. The scheme referred to in Table 3.4 by de Roos *et al.* (2011) was actually a closed nucleus of 200 cows where each dam had 10 offspring, generating 1000 males and 1000 females. MOET was used in this scheme, so cows would be 3 years of age when her ET calves were born and the 20 selected sires would be 5 years old. The rate of inbreeding was 0.52% per year (Table 1) which was almost three times the annual rate of inbreeding under conventional schemes.

Pryce *et al.* (2010) showed that using reproductive technologies aggressively could result in very high rates of genetic gain (double the rate of genetic gain when compared to progeny testing). However, this was also associated with comparatively high rates of inbreeding, making implementation of this type of scheme less attractive.

Overview of impact of genomic breeding schemes on rates of inbreeding. The rate of inbreeding per year ranged between 0.07 and 0.70% per year (Table 3.4). The relatively low estimates of inbreeding per year reported by Pryce et al. (2010) were calculated using a deterministic model. While sufficient to compare schemes within their study, they are not directly comparable to estimates of inbreeding rate in other studies using stochastic methods.

The source of the increased accuracy of genomic selection over traditional methods is a better estimation of the Mendelian sampling term. This allows for a reduction in co-selection of relatives. Consider the selection of candidates for a progeny test scheme, where 2 or more young full brothers will have the same set of EBVs. Therefore, under truncation selection all the full brothers will be selected. In contrast, GEBVs will differ among full brothers and only the best will be selected. This leads to a reduction of inbreeding per generation as seen in the pre-screening scenarios. The extent to which co-selection is improved depends how well the Mendelian sampling terms can be estimated (i.e. the accuracy of genomic selection). Therefore, improvements in genomic prediction methods should decrease inbreeding per generation. It matters of course what genomic selection is compared with. In dairy cattle, progeny test schemes already predict Mendelian Sampling terms with high accuracy. Thus, when comparing use of young genomically tested bulls to progeny test schemes, the accuracy of young bull GEBVs is generally lower than the accuracy of progeny test bull EBVs. This results in increased co-selection in the young genomic scheme versus the progeny test schemes leading to higher inbreeding per generation in the young genomic scheme (de Roos *et al.* 2011; Mc Hugh *et al.* 2011). This trend is moderated if GEBVs are available on female selection candidates resulting in less co-selection because the GEBVs will be more accurate than traditional EBVs in cows (Schaeffer 2006; Daetwyler *et al.* 2007; Sorensen & Sorensen 2009).

Implications for the reference population. Continuous re-estimation of marker effects in a genotyped reference population with accurate phenotypes is necessary for a successful genomic selection program (e.g. Habier *et al.* 2007). One risk with replacing progeny-testing with breeding schemes that screen large numbers of young bulls and only select a small number of these for widespread use, is that fewer bulls will be added to the reference population on an annual basis than in the past. This

would decrease the accuracy of genomic prediction as the distance between the current dairy population and the majority of animals in the reference population increases (Lillehammer *et al.* 2010). Countries with small populations may be more affected by this issue than larger populations (McHugh *et al.* 2011). Considerable effort has gone into increasing the size of current reference populations and this effort must continue to ensure reference populations remain relevant to selection candidates. One of the strategies used to increase reference populations is to share genotypes. Currently, the Eurogenomics (France, Germany, the Netherlands and Denmark/Norway/Sweden) and North American (USA and Canada) reference populations include approximately 20,000 and 12,000 males respectively (Table 2).

Genotyping of cows is another way in which a larger reference population can be achieved. However, currently genotyping costs are too high to genotype commercial dairy cows. This means that only high merit (or elite) cows will be genotyped. High merit cows may have been preferentially treated and therefore their phenotypes could be biased. Therefore, adding cows to the reference population, in some cases could be detrimental. However, there are examples of research projects where females are being genotyped specifically to become part of the reference population. For example, in Australia the Dairy Futures Cooperative Research Centre's 10,000 Holstein Cow Genomes project, where 10,000 cows (from commercial herds) have been genotyped to become part of the reference populations. In fact collecting data on cows may actually be more important in the genomic era than ever before, as cows may become a key part of future reference populations. Decreasing genotyping costs may allow all females to be genotyped in the future. Buch (2011) compared using progeny tested bulls in a reference population to using their genotyped daughters and phenotypes in the reference. The accuracy of genomic selection was higher when using the cows due to a loss of information when using the progeny tested bull. Possibly because the 'phenotypes' used for progeny test bulls are daughter trait deviations which are the mean of a bull's daughter group adjusted for fixed effects, thus ignoring variation around the mean. Whether this increase in accuracy alone warrants genotyping of very large numbers of cows remains to be investigated.

Another attractive aspect of having females in the reference population is that novel traits that are difficult or expensive to measure could be included in breeding

programs. Examples include health disorders, such as hoof-diseases recorded by hoof trimmers (Buch *et al.* 2011), residual feed intake (Pryce *et al.* 2011a), milk fatty acid composition (Soyeurt *et al.* 2011) or detailed recordings of reproductive measurements, such as pregnancy diagnosis data. One option could be to set up managed groups of information herds, selected for impeccable record-keeping.

In the pig, beef, sheep, and poultry industries, a major impact of genomic selection is likely to be increased genetic gain for hard to select for traits. This would include traits like disease resistance in poultry and meat quality in pigs. A sheep information nucleus has been implemented in Australia, where many difficult to measure traits are recorded (van der Werf *et al.* 2010). This population serves as the reference population for genomic selection (Daetwyler *et al.* 2012b). In pigs, sheep and beef cattle, genomic selection is often applied in cross-bred systems (Saatchi *et al.* 2011; Cleveland *et al.* 2012).

4. Imputation of genotypes in animal breeding

4.1 Introduction

If we knew the haplotypes individuals carried at every point on the genome, and we knew what SNP alleles were contained within with each unique haplotype in the population, then we could infer or impute the genotypes an individual carries for any SNP locus.

This would be useful for a number of reasons.

- Although the SNP array technology is that typically greater than 99.9% of all SNP are called per individual, at high quality, this still leaves a considerable number of SNP genotypes missing per individual. For example, with 50,000 SNP, this would result in 50 missing genotypes. For larger arrays, the number missing will be even higher. Missing genotypes complicate the implementation of genomic selection and genome wide association studies – the X matrix will be incomplete. Imputation can be used to infer these missing genotypes
- Imputation could be used to recover the high density genotypes for animals genotyped with a low density array. For example, we may be able to impute 50K genotypes for an individual from actual genotypes from a 7K array.
- Combining data sets. This particularly useful if one group of individuals are genotyped for one panel of SNPs, and another group is genotyped for another panel. Provided there is sufficient overlap between the two panels, the full set of SNPs can be imputed into all individuals, and genomic prediction or genome wide association studies can proceed, potentially with greater power.
- Imputation could be used to recover genotypes calls for full genome sequence data (eg. very dense SNP /insertions and deletions, copy number variants, to enable genomic predictions or genome wide association studies from this full sequence data.
- As will be described in the next chapter, there is uncertainty in calling genotypes from full sequence data, particularly if the coverage of sequence

is low. For example, if a region of the genome is sequenced at a depth of two sequences, it is difficult to determine if the individual is heterozygous or homozygous, as both sequences may be derived from the paternal or maternal chromosome. Imputation is used to take advantage of the linkage disequilibrium in the population to improve the probability of correctly calling genotypes from sequence data.

4.2 How does imputation work – Hidden Markov Models

As described above, if we knew the haplotypes individuals carried at every point on the genome, and we knew what SNP alleles were associated with each unique haplotype in the population, then we could infer or impute the genotypes an individual carries for any SNP locus.
In practice of course, we don't know the true haplotypes that each individual carries. Hidden Markov Models (HMM), are a useful approach here. In a HMM, the hidden state, the true haplotypes in the population, generate the observations, which are the genotypes. HMM have been widely used to estimate the probability that an individual carries a particular genotype at a particular SNP, given the genotype data for that individual at the other SNP and the rest of the population.

Many of the methods for imputation that use HMM also take advantage of a reference population, genotyped for all SNPs, that has been previously phased. These h reference haplotypes are designated H. Then the haplotypes carried by the target individuals for imputation (eg. those genotyped at a low density SNP array) are considered as a mosaic of the haplotypes in the reference. "Mosaic" means that the target individual must comprise of haplotypes from the reference population, with some crossovers between the haplotypes, and some rare mutation. This is illustrated in Figure 1. Some methods assume this population has been previously phased from haplotypes to genotypes, using the PHASE program for example.

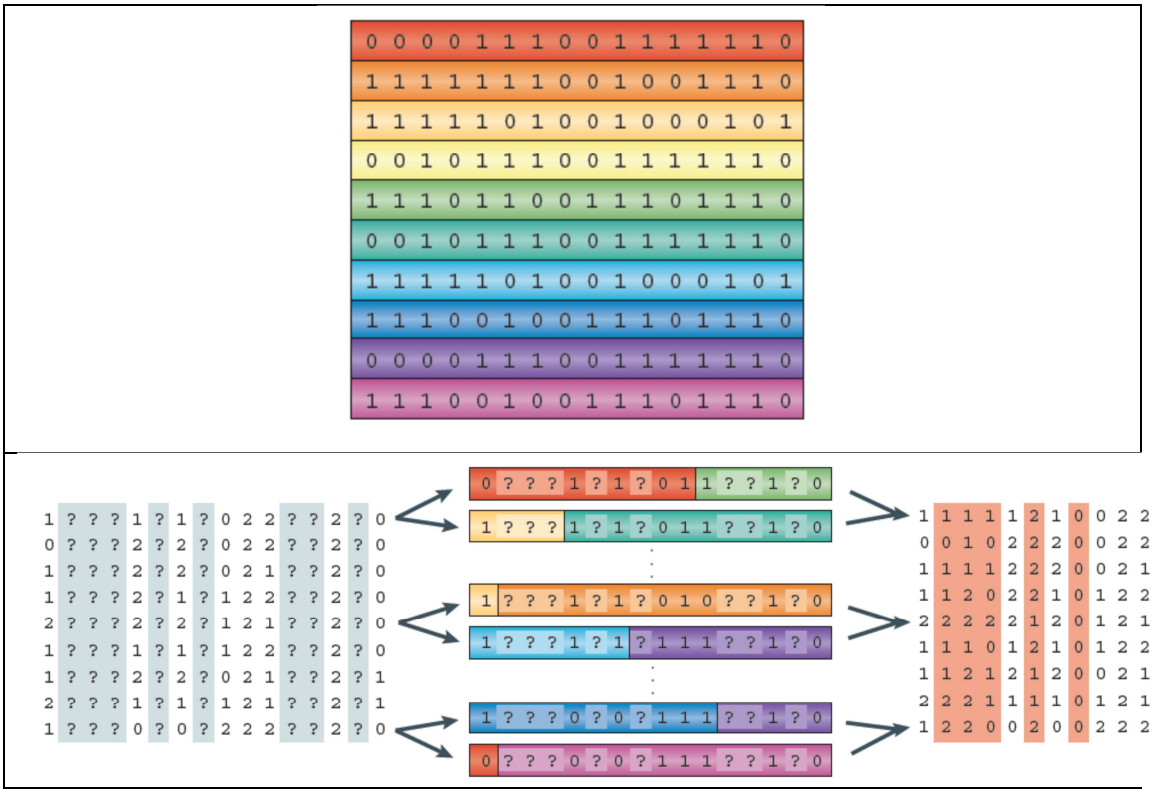


Figure 1. From (Marchini & Howie 2010). A cartoon of genotype imputation. A. A phased reference population is the a requirement in many imputation programs. B. The genotypes in the target population are phased, then assigned a mosaic of the reference haplotypes via a hidden markov model.

If we consider a chromosome with L loci, then the five components of a Hidden Markov Models are

- hidden states (S). In this case these are indicator variables assigning the alleles at the reference haplotypes to target individuals. There are one 1 to L indicator variables, and each indicator variable comprises two numbers, one for the paternal and one for the maternal chromosome. For example, in the Figure above, the value of S1 for target individual one would be 1,2.
- observed values. In this case these are the genotypes **G**, of which some may be missing.
- state transition probabilities. This is a h x h matrix, describing the probability of moving from one haplotype to another (for example through recombination or mutation).
- emission probabilities. In the HMM, the underlying state (haplotypes) are said to "emit" the observations, the genotypes. So the emission

probabilities are the probability of observing the genotype carried by an individual for a particular underlying hidden state. For example, if the genotype at a particular locus was AT, and the underlying hidden state was A**A**CG, with the bold allele the allele at the current SNP, the emission probability would be 1 (assuming no genotyping error).

- Initial state probabilities. This is the probability the HMM starts in a particular state, eg at a particular haplotype.

The methods for imputation differ in their assumptions about the hidden states, the way state transition probabilities are derived, emission probabilities, and initial state probabilities.

The major strategies for imputation described in the literature will be reviewed briefly here. Much of the material is from two reviews (Marchini & Howie 2008; Marchini & Howie 2010) . Both reviews are suggested further reading.

IMPUTE1.0 uses a reference population as described above (eg a set of phased haplotypes), and parameters describing the recombination rate to estimate the probability of genotypes.
The probability of the genotypes for an individual Gi to be imputed, given the reference haplotypes H, is then

$$P(Gi|H,\theta,\rho) = \sum_{S} P(Gi|S,\theta)P(S|H,\rho)$$

Where ρ is the recombination rate map across the genome, θ is a mutation parameter that (rarely) allows the genotype vector for individual i to differ through mutation from the reference haplotypes that they are derived from, and S is the hidden states (haplotypes). S can also be thought of as a design matrix which "copies" the selected reference haplotypes to the target genotypes. For example, if there are 5 loci, and individual i is a mosaic of haplotypes from the reference 1 and 2, with a crossover between the third and fourth loci, then S would be
11100
00011
The probability is calculated by integrating over all possible states the probability of the observed genotypes given the states and the mutation rate, and the transition

between states P(S|H,ρ). This term is the probability of the States given the reference haplotypes and the recombination rate.

The recombination rate map must be supplied to IMPUTE1.0. A forward-backward algorithm for HMM is used to estimate the probability distributions (Rabiner 1989).

IMPUTE2.0 is a modification of IMPUTE1.0. This method first estimates the phase of SNP in the target population, then compares these phased haplotypes to those in the reference population to impute the missing alleles. As this algorithm uses haploid imputation (eg haplotypes in the target are compared to the haplotypes of the reference, rather than comparing genotypes), the authors of this method (Howie *et al.* 2009) demonstrate that this leads to much faster imputation.

FastPHASE. FastPHASE (Scheet & Stephens 2006), is an modification of the PHASE program already discussed. The hidden states in the model are clusters of haplotypes rather than the haplotypes themselves. For example, a cluster may be a group of haplotypes that are almost identical, with the exception of a (rare) single mutation. Clustering very similar haplotypes greatly reduces the number of hidden states that must be considered, which decreases computation time. The default setting for the number of clusters at a given genomic location in fastPHASE is 20.
The probability haplotype I for the current individual comes from the k^{th} cluster is weighted according to how many haplotypes of type k have been observed:

$$P(Gi|\alpha,\theta,r) = \sum_{s} P(Gi|Si)P(Si|\alpha,r)$$

Where α is a vector of the proportion of times each of the haplotype clusters is occurs, eg. The weight for the kth haplotype cluster may be 0.2. In this case θ is the frequency of alleles within each cluster. The transition probabilities, the probability of switching between a cluster for an individual, is the term P(Zi,α,r). r is a combination of recombination rates and mutation rates, both of which are estimated in the fastPHASE program.
The likelihood of genotype Gi is then

$$L(Gi|H,\alpha,\theta,r) = \prod P(Gi|\alpha,\theta,r) \prod P(Hi|\alpha,\theta,r)$$

An Expectation-Maximisation algorithm is used to fit the model, and compute genotype probabilities.

MACH (Li *et al.* 2010). MACH has some similarities with FastPHASE, however it uses the full set of haplotypes as hidden states rather than haplotype clusters. During each EM iteration of the model fitting, the current estimates of haplotype phase, except for the individual being fitted, are used as the reference haplotypes. Individuals are removed from the set of reference haplotypes one at a time and are updated, with the updated pair of haplotypes for the individual is sampled from the posterior probability distribution, based on the current reference haplotypes:

$$P(Gi|D-i,\theta,\tau) = \sum_{s} P(Gi|S,\tau)P(S|D-i,\theta)$$

where D–i is the set of estimated haplotypes of all individuals except i, S denotes the hidden states of the HMM, η is an 'error' parameter that controls how similar Gi is to the copied haplotypes (to account for genotyping error) and θ is a 'crossover' parameter that controls transitions between the hidden states. The parameters η and θ are during each iteration (eg estimated from the data) based on counts of the number and location of the change points in the hidden states S and counts of the concordance between the observed genotypes to those implied by the sampled hidden states. Imputation of unobserved genotypes using a reference panel of haplotypes, H, is naturally accommodated in this method by adding H to the set of estimated haplotypes D–I (Marchini & Howie 2010).

BEAGLE (Browning & Browning 2009). BEAGLE uses a different approach to define the hidden states to the methods defined above. Local clustering of haplotypes is used- that is, for a given genomic location, the possible hidden states are reduced to those that are observed in the reference. This is in contrast to IMPUTE and MACH, where at any position the number of states is the number of reference haplotypes squared. So the number of hidden states in BEAGLE varies with location. In addition, a haplotype cluster can only emit a single allele (eg A or T) – haplotypes carrying different alleles are assigned to different clusters, and there is 0 probability of

genotyping error assumed. The idea behind these conditions is to reduce computation. A final difference is that many haplotype configurations are assigned a probability of zero by the Browning model. This allows the model to be more parsimonious (eg better fit to the data), but means that the haplotype model must be constructed from all sampled individuals, rather than from a subset acting as a reference panel. Otherwise if a new haplotype is encountered in the target individuals, there may be no haplotype configuration in the model that is consistent with the individual's genotype. Some of the differences between BEAGLE and MACH/IMPUTE and fastPHASE are summarized in Figure 2 (from (Browning & Browning 2009)).

One key difference between BEAGLE and MACH/IMPUTE/fastPHASE is that no use is made in BEAGLE of population parameters recombination rates or mutation rates. When the reference population is small, this is a disadvantage for BEAGLE, as the only information is from the data in the current genomic location, while MACH/IMPUTE/fastPHASE can gain accuracy from the additional information on the population and genome wide parameters such recombination rates and mutation rates. However when the data set is large, estimating these parameters can incur additional computational cost, and using the parameters when they are inaccurate may actually decrease the accuracy of imputation.

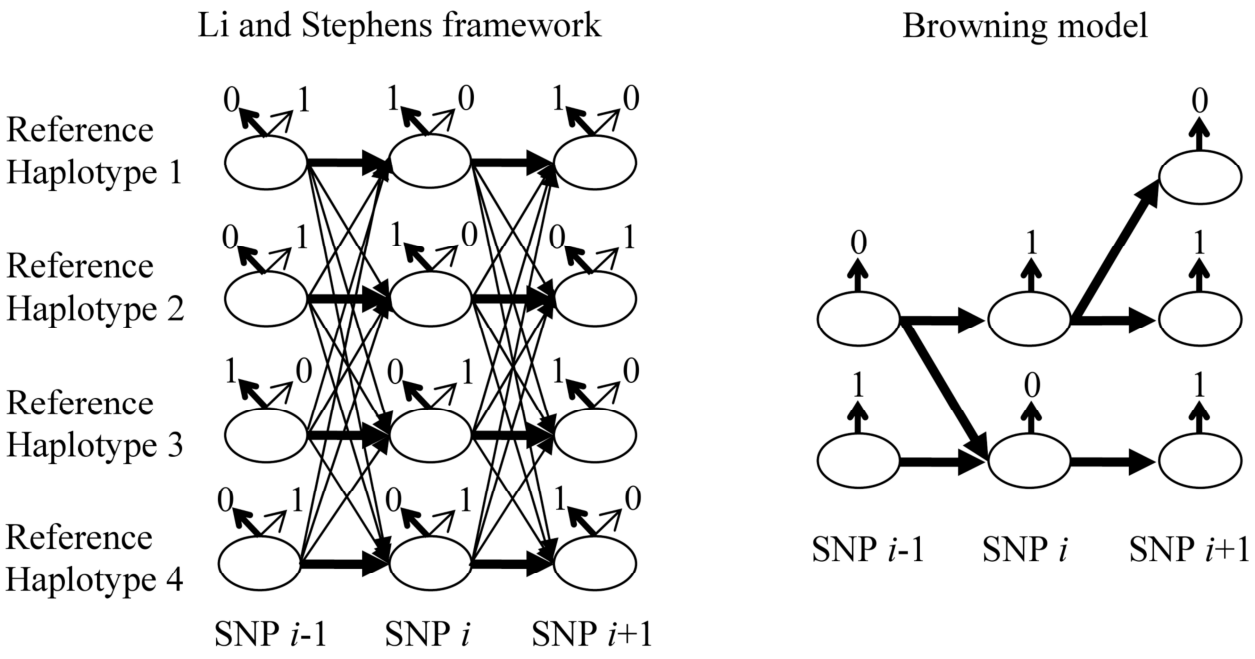


Figure 2. Illustration highlighting major differences between models based on the Li and Stephens framework (2003), the basis for MACH, IMPUTE and fastPHASE, and the Browning model (Browning 2006), the basis for BEAGLE. Excerpts of the models covering three markers (SNPs i-1, i and i+1) are shown. Ovals are hidden states of the models. For the Li and Stephens framework, these states are defined by the reference haplotypes, while for the Browning model the states are localized clusters of haplotypes. Note that the graphical representation of the Browning model is that given in Browning (2008), while earlier representations had states as edges rather than as nodes of the graph. The Browning model will tend to have fewer states at any given marker than will unconstrained models based on the Li and Stephens framework, and the number of states can vary from marker to marker for the Browning model but is fixed in the Li and Stephens framework. Arrows between states from one SNP to the next are transitions of the HMM. For the Li and Stephens framework, transitions with highest prior probability (those seen in the reference haplotypes) are shown with bold arrows, while thin arrows allow for historical recombination. For the Browning model, there are at most k transitions coming out of a state, where k is the number of alleles at the next marker (i.e. 2 for SNPs), which helps to keep the model parsimonious. Arrows coming out of the top of the states are possible emissions of the HMM, which are the observed alleles. For the Li and Stephens framework, emissions with highest prior probability (the alleles on the reference haplotypes) are shown with bold arrows, while thin arrows represent mutations to other alleles. The reference haplotypes here are 011, 010, 101 and 001. For the Browning model, there is only one possible emission from each state, which helps to keep the model parsimonious. The models shown are illustrative only. The actual form of the Browning model will vary depending on the alleles of the reference haplotypes outside this window of markers..

A good example is given in Browning and Browning (2009). They compared the performance of IMPUTE1.0 and BEAGLE, in the Wellcome Trust Case Control Consortium (WTCCC) data, which includes approximately 2000 cases for each of seven diseases (bipolar disorder, coronary artery disease, Crohn's disease, hypertension, rheumatoid arthritis, type 1 diabetes, and type 2 diabetes) and approximately 3000 shared controls. The comparison used data from chromosome 1 with 53,683 markers genotyped A subset of 24,705 markers was masked and imputed with either BEAGLE or IMPUT1.0 in 188 individuals, using a reference panel of 600, 300 or 60 individuals with full genotypes. The authors found that while IMPUTE1.0 was more accurate with smaller reference set sizes, BEAGLE was more accurate when the reference size was bigger. The allele-frequency correlations were 0.990 (BEAGLE) and 0.992 (IMPUTE) with a reference panel of 60 individuals, 0.997 (BEAGLE) and 0.998 (IMPUTE) with a reference panel of 300 individuals, and 0.998

(BEAGLE) and 0.998 (IMPUTE) with a reference panel of 600 individuals. The authors concluded that the difference in accuracy between IMPUTE and BEAGLE is substantially smaller than the gain in accuracy obtained from using larger reference panels.

4.3 Including information from pedigree to improve the accuracy of imputation

There is additional information for phasing, and therefore imputation, if the pedigree amongst the individuals in the target and reference populations are known. For example, if a sire has large number of offspring, his genotypes can be phase into haplotypes by simply counting the alleles across the markers that occur together (allelic co-segregation). Trios, which consist of father, mother and offspring, and sometimes used in human genetics for the same purpose. When this information is known, the number of hidden states that must be considered can be reduced to four, corresponding to the paternal and maternal alleles of both the mother and father. Druet and Georges (2010) extended both BEAGLE and fastPHASE to take advantage of pedigree structures more typical of livestock and crop populations, for example large half sib or full sib families. In their approach, sires with six or more offspring or individuals with five or more sibs were phased using alleleic co-segregation and linkage approach. Then these "known" haplotypes were used in 1) fastPHASE, to estimate the parameters of the EM algorithm or 2) BEAGLE, to generate the directed acyclic graph (DAG) describing the hidden states, transition and emission probabilities. Either BEAGLE or fastPHASE are then run. In dairy cattle, recent results suggest that using the pedigree information in this way, prior to running BEAGLE, can improve the accuracy of imputation (Druet pers com).

4.4 An alternative approach to phasing and imputation: Long range phasing

An alternative approach to phasing and imputation is to exploit the fact that some individuals share a recent common ancestor, and therefore share long chromosome segments which are identical by descent. This is particularly true of livestock

populations, where some sires have very large number of descendants. As described by Kong et al (2008), this leads to a phasing approach based on the key observation that if animals have non-conflicting homozygote genotypes over a long string of consecutive loci, they have at least one long haplotype in common. This requirement, of a long string of loci, leads to a high probability that the common long haplotype has originated in a common ancestor (eg is identical by descent as well as identical by state). The method proceeds by considering one individual at a time, and identify either real or "surrogate" parents (if the real parents are unknown). As describe by Kong et al. (2008) and Hickey et al. (2011), surrogate parents are individuals who share a haplotype with the individual being considered, identified as those individuals that do not have any opposing homozygote genotypes with the current individual. Inference of the phase at each locus for the current individual within the paternal/maternal haplotype is attempted by stepping through the paternal/maternal surrogates until a surrogate is found that is homozygous at that locus and thus can be used to declare the phase. If the surrogates that are one degree removed from the current individual cannot be used to declare phase, eg they are heterozygous, surrogates of the surrogates are collected, and so on, until a homozygote is found, Figure 3. Hickey et al. (2011) demonstrated that using a modified long range phasing algorithm in livestock populations led to extremely accurate phasing, in reasonable computing time. This is likely because livestock populations have relatively small Ne, so large segments of chromosome are shared between individuals.

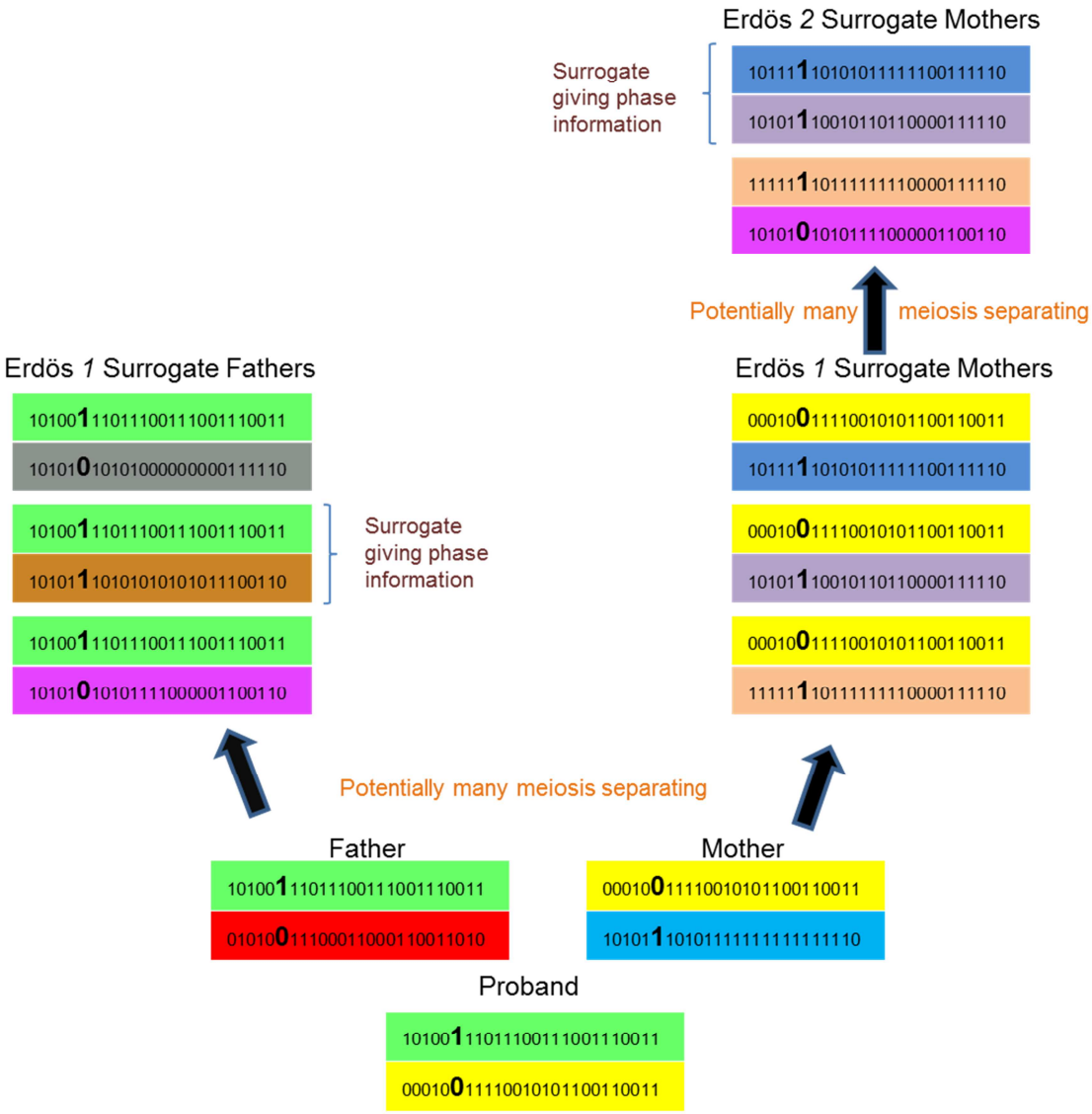


Figure 3. From Hickey et al. (2011). Illustration of the long range phasing process.

As demonstrated by Daetwyler et al.(2011), and Hickey (pers com), the principle of comparing long stretches of chromosomes between individuals to identify common segments can also be used to impute and phase missing genotypes. They demonstrated this approach gave more accurate imputation results than fastPHASE in a dairy cattle population, in a fraction of the computing time.3.

4.5 Results of imputation in livestock populations.

In dairy cattle, accurate imputation from low density markers to 50K SNPs has been described by a number of authors. Weigel et al. (2010) evaluated the accuracy of imputing up to 43,385 SNP in Jersey cattle, when in 1, 2, 5, 10, 20, 40, or 80% of these loci were genotyped in a target population. Both IMPUTE2.0 and fastPHASE were used for imputation. They found the accuracy of imputation was low (<0.80) when fewer than 1,000 SNP are used, but when 4,000 SNP were used the accuracy of imputation was 0.95. Weigel et al (2010) also assessed the effect of imputation on the accuracy of genomic estimated breeding values (GEBV). They concluded that provided the target population was genotyped for at least 3000SNP, with imputation to 43,000 SNP, GEBV were predicted with an accuracy of 95% of what was possible with the real 43,000 SNP. They also demonstrated that using the imputed genotypes resulted in GEBV that were approximately 5% more accurate than using the 3000 SNP alone, without imputation.

Similar results for the accuracy of imputing 3,000 SNP to approximately 50,000 SNP have been found in Holstein-Friesian dairy cattle. Zhang and Druet (2010) reported error rates of 3-4% in this situation using DAGPHASE (Druet & Georges 2010), though their main conclusion was that the accuracy of imputation was dependent on the genetic relationship between the target individual and the reference population (discussed below). Dassonneville et al.(2011) using the same method observed similar error rates when imputing 3K to 50K in European Holstein cattle, and went on to demonstrate that the loss in accuracy of GEBV using the imputed genotypes rather than 50K genotypes was only 0.02. Daetwyler et al. (2011) reported slightly higher error rates with their implementation of the long range phasing algorithm, although the used as smaller reference population, and the algorithm outperformed fastPHASE. Using BEAGLE in the same population gave error rates of 5%.

Another interesting potential application of imputation was demonstrated by Druet et al. (2010), where two populations, each genotyped for separate panels of

approximately 28,000 SNP, and overlapping by approximately 9,000 SNP were imputed up to 60,000 SNP with very low error rates (note that in this study all animals were actually genotyped with 60,000SNP, but the results do demonstrate the possibility of meta-analysis of populations genotyped with different SNP panels.

In pig and chicken breeding, moderate sized full sib families are the norm. In such populations, another imputation strategy is possible, whereby parents are genotyped for a dense (say 50K) marker panel, and the offspring are genotyped with a very low density marker panel (say 384 SNP), as outlined by Habier et al.(2009). Given the limited number of recombinations that occur between parents and offspring, this very limited number of markers is sufficient to determine whether progeny have inherited maternal or paternal chromosomes from each parent. The rest of the markers can then be "imputed" if the haplotypes of the parents are known. Habier et al. (2009) demonstrated this very low cost strategy could result in prediction of genomic breeding values with accuracies nearly as high as if the progeny had been genotyped for the full 50K SNP. This strategy is now being used in pig and chicken breeding programs (Dekkers, pers com).

In sheep, few results have been published. Hayes et al. (2011) reported fairly low accuracies of imputation in three sheep breeds, albeit with very small reference populations (80 to 200). Accuracies of imputing 48,000 SNP from 5,000 SNP was 80% for Poll Dorsets, White Suffolks and Border Leicesters. For Merino sheep, even though a much larger reference set was used, the accuracy of imputation was only 71%, likely due to the very large effective population size for this breed (see below). While imputation is likely to be an important strategy in crop species, no results have been published to date.

4.6 Factors affecting accuracy of imputation

4.6.1 Size of the reference population.

It is critical that the reference population is large enough to capture all the haplotypes in the population. If a target haplotype is encountered which has not been previously observed in the reference population, the imputation of missing genotypes is unlikely

to be accurate. The size of the reference is also important for other consideration – in fastPHASE for example, haplotype (actually cluster) frequencies are used in the model, and these will be inaccurately estimated with a low number of markers. In BEAGLE, the accuracy of imputation is very dependent on the size of the reference population as this determines how well the directed acyclic graph (DAG) describes the population. If the reference is too small, there may be haplotypes in the target which are not represented in the reference, so the alleles on these haplotypes will be poorly imputed. Browning and Browning (2009) demonstrated that increasing the size of the reference had a large impact on the accuracy of imputation, as was larger than the differences between methods.

4.6.2 Density of markers and effective population size.

If the markers are not sufficiently dense that there is substantial linkage disequilibrium between them, the methods using population level algorithms (eg MACH, BEAGLE, IMPUTE2.0, fastPHASE), will perform very poorly. This is because haplotypes encountered in the reference and haplotypes encountered in the target population, although they have a limited number of alleles in common, could be identical by chance rather than identical by chance, so the identity of the missing marker alleles in the target does not match those in the full genotyped animals. In dairy cattle population, linkage disequilibrium is sufficiently high (due to the low effective population size) that 3K SNP can be used to impute 50K with low error rates, provided the reference population is sufficiently large. However in a number of sheep breeds, the same number of markers cannot be successfully used for imputation using population based methods, as the level of linkage disequilibrium is too low, a result of higher effective population size than in dairy cattle (e.g. Hayes *et al.* 2011). Even if the marker density is too low for successful imputation using the population algorithms, within family linkage can still be exploited in some situations to obtain accurate imputations (e.g. Habier *et al.* 2009).

4.5.3 Genetic distance from the reference population.

Particularly when imputing from low marker densities (eg 3K to 50K), the accuracy of imputation is likely to be highly dependent on the genetic distance of the target individual from the reference population (e.g. Zhang & Druet 2010). If for example the individual has a sire in the reference, his or her 3K marker haplotypes will be readily identifiable among the 50K haplotypes. However if the individual does not have a sire, or a more distant relative in the reference, the chance his or her 3K haplotype has previously been observed (without intervening recombination) diminishes rapidly. In a sheep population, Hayes et al. (2011) demonstrated that 64% of the variation in accuracy of imputation among target individuals was accounted for by average genetic relationship to the reference.

Allele frequency. Another reason for using a large reference population is to ensure rare alleles are captured, and can be accurately imputed into the target individuals. For rare alleles, the probability of imputing the correct genotype by chance is high, as the majority of the individuals will be homzogygous for the common allele. However if the accuracy of imputation is corrected for the homzygosity of the markers, it is clear that the accuracy of imputation is actually lower for rare alleles, Figure 4. Another way of interpreting this is to think of the consequences for GWAS association study. If an allele is rare, the number of phenotype observations on that allele is low. If a significant proportion of these are actually incorrect due to the imputation, the already limited power will be greatly reduced.

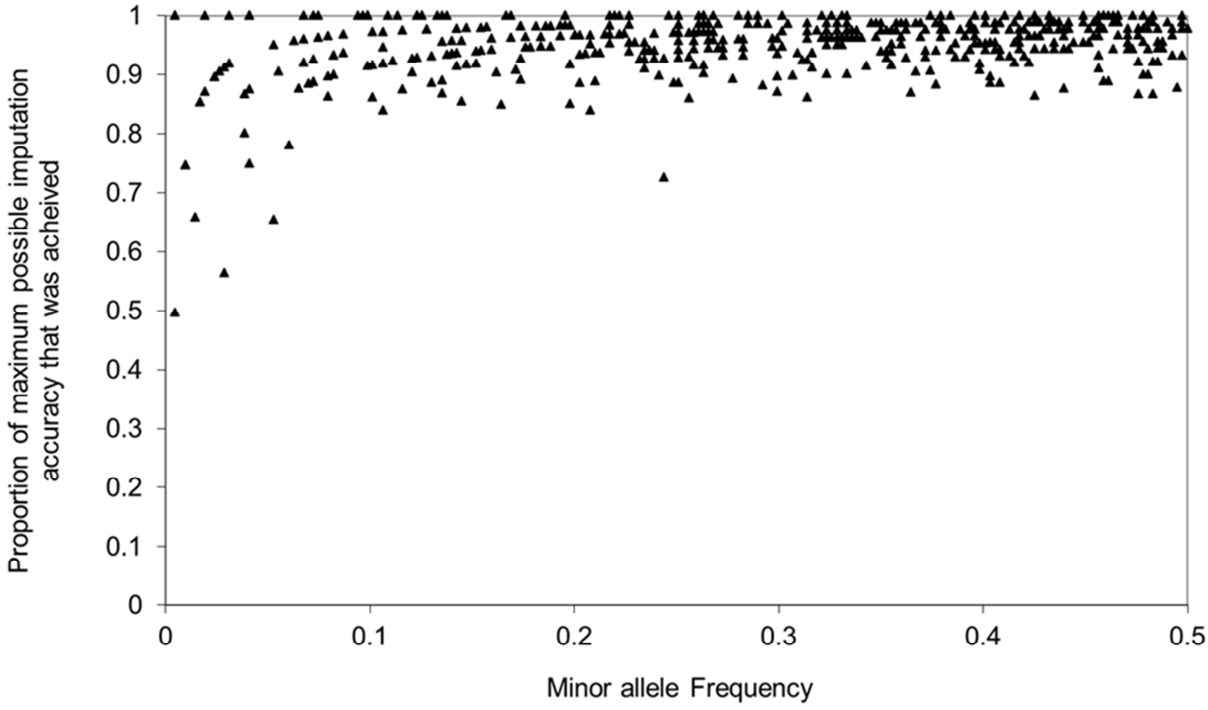


Figure 4. From Hayes et al. (2011). Proportion of maximum possible imputation accuracy that was achieved (50K to high density genotypes) by minor allele frequency, in a terminal sire sheep breed. The proportion of maximum possible imputation accuracy was calculated as the accuracy of imputation that was achieved minus the accuracy of imputation that would be achieved by chance, that is random sampling of genotypes conditional on genotype frequencies for each marker divided by one minus the accuracy of imputation that would be achieved by chance.

4.6.4 Why does imputation lead to more statistical power?

An obvious question is, if there is already enough information in haplotypes of low density markers to accurately impute up to higher density markers, why would the imputed genotypes add any power to genome wide association studies or increase the accuracy of genomic estimated breeding values? One explanation is that while testing the haplotypes themselves would require a factor with multiple levels, with degrees of freedom lost corresponding to the number of haplotypes-1, testing a SNP with two alleles leads to the loss of only one degree of freedom.

Further, if the GWAS is done across breeds, the marker density may be such that imputation is from the sparse markers is only possible within breeds (eg the haplotypes only persist within breeds), this can lead to the same SNP allele being imputed across breeds, such that an across breed test can be carried out.

5. Genome sequencing for genomic selection and Genome wide association studies

This short chapter suggests some potential advantages of using whole genome sequence in genome wide association studies and genomic selection. As there are very papers or results with full genome sequence data, the suggestions here should be considered hypothesis for testing, rather than results based on evidence. This area is unfolding very rapidly, so some of the ideas proposed below may well be out of date shortly after the time of writing (2012)!

5.1 Motivation

If all the individuals in a population could be sequenced, all the genomic variants in the population would be captured. This includes SNPs, small insertions and deletions, and copy number variants (CNVs). Why would this benefit genome wide association studies and genomic selection?

For genome wide association studies, the advantage is obvious. If full sequence data is used rather than a panel of SNP markers, then the actual mutation affecting the trait will be present in the data. So potentially, the GWAS could lead to direct identification of the causal variant. In practise, there may be other variants in complete LD with the causal variant, so that functional information has to be used to refine the candidates.

For genomic selection, the advantage of using full genome sequence data is less obvious. If genomic predictions are already based on a large number of SNP in high LD with QTL, using full genome sequence may not add much to accuracy and may with some methods in fact decrease accuracy, given the very large increase in the number of effects that need to be estimated from perhaps the same number of phenotypic records. However, the sequence data could increase the accuracy of genomic predictions in a number of situations

1) If LD between the QTL and SNP is incomplete. In this situation, the full QTL effect is captured only by the sequence data and not by the SNP data (as the

actual causative mutation is now in the data set). This is especially likely if some of the QTL alleles are very rare, while the majority of the SNP alleles on the widely used arrays have quite high minor allele frequencies. Meuwissen and Goddard (2010), using simulation, demonstrated a 5% increase in accuracy from using full sequence data over the densest SNP panel they simulated

2) If genomic predictions are made across breeds. In multi-breed populations, using full sequence data is likely to be particularly advantageous, as there is no longer the need to rely on SNP-QTL associations which may not persist across breeds.
3) Persistence of accuracy of genomic predictions. With current marker densities, for example the 50K SNP array in cattle, the accuracy of genomic predictions decays surprisingly rapidly with either generations removed from the reference set, or genetic distance from the reference set (Habier *et al.* 2007). This is because, with SNPs spaced every ~ 60kb, the SNP-QTL associations break down quite quickly. With full sequence, the QTL themselves should underlie the prediction equation, so that the decay in accuracy is greatly reduced. In their simulations, Meuwissen and Goddard (2010) demonstrated there was very little decay in accuracy over generations when full genome sequence was used. This is particularly important for expensive to measure traits, like feed conversion efficiency and methane emissions, where the cost of updating the prediction equation could be prohibitive.

5.2 Which individuals to sequence?

As sequencing is still expensive compared to the cost of genotyping (though this cost has declined more than one million fold in ten years, as is likely to keep declining), it is unlikely, at the time of writing at least, that the entire reference population will be sequenced. Rather, a likely strategy is that a few hundred or few thousand individuals will be sequenced, and imputation used to impute the variants in the sequence (including SNP, indels and CNV) into the full reference population (e.g. Meuwissen & Goddard 2010; Le & Durbin 2011). One obvious way to choose the individuals

then is to choose those that will maximise the accuracy of imputation, or equivalently, capture the highest proportion of genetic variation in the target population. This leads to sequencing of key ancestors. To choose amongst the possible ancestors, the following algorithm could be used (Hayes and Goddard 2007).

Let the number of potential key ancestors be *n* and let **A** be an *n*x*n* matrix which is the additive relationship matrix among the *n* animals in the whole population. Let **c** be an *n*x1 vector with the *n* animals ordered in the same way as in **A**, and c_i = the average relationship between animal *i* and the whole population. Consider a sub matrix of **A** ($\mathbf{A_m}$) containing the relationship between a subset *m* of the animals, to be sequenced, and let $\mathbf{c_m}$ be the equivalent sub vector of **c**. Then $\mathbf{p}=\mathbf{A_m}^{-1}\mathbf{c_m}$ is a vector whose i^{th} element is the proportion of the genes in the whole population that derive only from animal *i* amongst the *m* key ancestors and **p'1** is the total of the elements of **p** and is the proportion of genes in the whole population that derive from the *m* key ancestors (where **1** is a vector of 1s). Therefore to select the *m* ancestors that capture the most genetic variation in the population find the subset that maximise **p'1**. This can be done either by stepwise regression, which can be done by finding the single individual with the largest value of p, choosing the next individual by setting the individual with the previous highest contribution to 0 in $\mathbf{c_m}$, recalculating **p**, and so on. A genetic algorithm can also be used.

An example of the use of this algorithm applied to real data is given in Figure 1 for the Poll Dorset Sheep breed. Sequencing 50 key ancestors would capture 35% of the genetic variation in this breed.

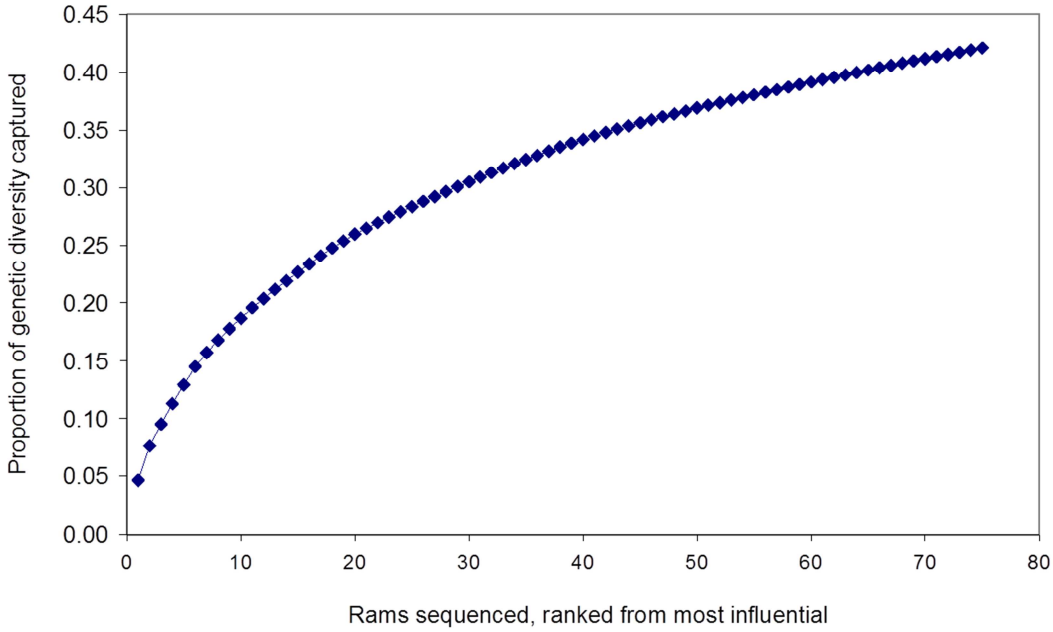


Figure 1. Proportion of genetic diversity (as measured by pedigree), captured by subsets of groups of rams, ranked from the most to least influential.

5.3 Imputation of full sequence data

Once a subset of individuals, perhaps the key ancestors, are sequenced, the next task is to impute the variants that occur in the sequence into the reference population for GWAS or genomic selection.

The first step is to sequence a reasonable number of individual, then variants are identified between the individuals and between the two chromosomes (paternal and maternal) of the individuals, followed by calling of genotypes in the each sequenced individual. To identify variants and call genotypes, the properties of the sequence data must be taken into account. While it is beyond the scope of this chapter to fully describe these and algorithms that have been used for this purpose, the properties of the sequence data that must be dealt with are variable coverage of each base in the genome, and variable quality of the sequence data. The variable coverage arises because of the process used to sequence genomes, which is to shatter each genome into small pieces (perhaps 150bp long), sequence these, and then align the reads to a reference genome (a genome that has been assembled previously). The probability that each small piece of genome is sequenced is random, and many genome locations

are sequenced multiple times. When the reads are aligned to a reference genome, this results in a depth of coverage (the number of times each base is sequenced) which is approximately poisson distributed, with mean the mean coverage set by the laboratory ("the depth of sequencing"). For example, if the average fold coverage is 4, then 1.8% of the genome will not be covered at all (eg. $P(0) = 4^0e^{-4}/0!$). One of the major challenges for calling variants, and genotypes, is that for truly heterozygous sites, the probability that both alleles are observed in the sequence data is low at low fold coverage. A further challenge is the high rate of sequence errors, these occur approximately one every hundred base pairs with the Illumina technology at least. Algorithms have been devised to take both sources of error into account when calling genotypes from the sequence data. The best algorithms give probabilities of each genotype (for example AA,AT and TT) at a putative variant for each individual, rather than an absoloute genotype call. These probabilities take into account the depth of sequence reads, the quality of the reads at that location. A recent paper (Danecek *et al.* 2011) describes software implementing such an algorithm. The 1000 Genomes paper (1000 Genome consortium (2010), supplementary reading is also recommended reading here.

Population level information can also be used to increase the accuracy of calling genotypes from the sequence data. Both MACH and BEAGLE, described in the Chapter 4.0, have been modified to take in genotype probabilities calculated from sequence data, run imputation and therefore exploit population level information to improve the accuracy of genotype calls. Again, both these approaches are well described in the 1000 Genomes consortium paper (2010), Supplementary methods.

Once the genotypes have been called in the sequence individuals, they can be used as a reference population for imputing the variants in the sequence into the group of animals with 50K or 800K genotypes. This can be done using any of the imputation programs, provided they are computationally efficient, as the number of variants is likely to be very large! Note that it may be worthwhile to use genotype probabilities here rather than absolute genotypes, to account for any uncertainty in imputation.

5.4 Methods for genomic prediction with full sequence data

Once the variants in the sequence data have been imputed into the animals with SNP array genotypes and phenotypes, a prediction equation can be derived. The question is which genomic prediction method is appropriate for this data? At the time of writing, this question had not been answered in real data, so what follows is speculation. If we assume that quantitative traits are controlled by perhaps a few thousand loci, then we would like our genomic prediction method to attribute effects to these 1000s of loci, and set the rest of the effects of the variants (which may be in LD with the causative mutations, but are not the causative mutations themselves, to zero. In this case, a BLUP method, which assumes the effect of all variants is small, non zero, and normally distributed, is inappropriate. A method such as BayesB, or BayesCpi, which allow for a large number of variant effects to be set to zero, would seem to be a much more appropriate method.

In their simulation of a population with sequence data, with a tens of QTL, and very large number SNP, Meuwissen and Goddard (2010) demonstrated very considerable advantage in the accuracy of GEBV for BayesB over BLUP (up to 40%). However it must again be pointed out that this is simulated data with a simple genetic architecture, and the methods need to be tested in real data set.

5.5 An example of using full sequence data. A genome wide association study in Rice.

An elegant example of the power of a genome wide association study with full sequence data was provided by Huang et al. (2010) "Genome wide association studies of 14 agronomic traits in rice landraces". A key advantage they had was they were using inbred lines, so there were no heterozygous genotypes for any variant in the data, so very low coverage sequencing could be used. They sequenced 517 rice landraces (inbred lines!) at only 1x coverage. These lines represented ~ 82% of diversity in the world's rice cultivars. Each line was well characterised for 14 agronomic traits including grain yield and growth rates. The sequence from each line was stacked, or piled up, for the calling of sequencing variants. 3.6 million SNP were

detected in these pileups. However, with 1x coverage, they could only call genotypes at ~ 20% of the SNP for each landrace. So imputation was used to fill in the missing genotype. Then GWAS were performed for each of the traits using the 3.6 million imputed genotypes in the 517 lines. The authors demonstrated that they found already known mutations with effects on some of these traits, place a host of new mutations with very significant effects for future investigation.

6. Practical Exercises

6.1 Assessing the extent of linkage disequilibrium in HaploView

We will use the HaploView program to calculate r^2 values. The data set we will use is 10 SNP markers on a section of chromosome 20 genotyped in 325 bulls.

The genotype (in linkage format) file for HaploView has the following format

Pedigree_ID Individual_ID Sire_ID Dam_ID Sex Affected Marker1_Allele1 Marker2_Allele2

You can find out more about the genotype input file in the Help tab of haploview

The map file consists of two columns, the marker name and the position, eg

Marker1 19992222
Marker2 23100202

Import the genotype file "325_bulls_genos.txt". Import the file "map.txt". Set the minumum distance to calculate r^2 to markers less than 5000kb apart.

Are all the markers in Hardy-Weinberg equilibrium? Which marker has the lowest minor allele frequency?

Set the HW cuttoff to 0.0000, and click on the box to make sure they are all included.

Then click on the LD plot tab. To make sure the values are r^2, click Display -> Show LD values -> R–squared. The boxes show the r^2 values between the markers from 0 to 100. If the markers are in 100% LD, there will be a red box with no number.

Which markers are in the highest LD? Are there any markers in perfect LD?

Does the LD decay uniformly across the chromosome segment (for example look at marker 1 versus the rest)? How would you describe the pattern of LD with distance in this small chromosome segment?

6.2 Genome wide association study

Now we will conduct a genome wide association study using the same data and phenotypes.

Before we use go further, let's take a moment to get acquainted with R. We will use a simple example of multiplication of two matricies to obtain another matrix. Open the R graphical user interface by clicking on it. You should see the command prompt.

Let's multiply two matricies a and b to get a third matrix c.

The matrix a is a two by two matrix with elements:
1 1
2 2
The matrix b is a two by three matrix with elements:
1 2 2
2 3 4

We can input these matricies into the computer memory as:
> a <- matrix(c(1,1,2,2),ncol=2,byrow=TRUE)
> b <- matrix(c(1,2,2,2,3,4),ncol=3,byrow=TRUE)

To check the dimensions of a and be are correct type:
> dim(a)
> dim(b)

You can print a matrix at any time, eg
> print(a)

Now lets multiple matricies a and b to get a new matrix c:
> c <- a%*%b (%*% is the symbol for matrix multiplication)

Check the dimensions of c are correct,
> dim(c)
And that the c matrix has the correct elements:
> print(c) (you can compare this to the result in excel for example)

A matrix can be transposed using t(a), eg
> d <-t(a)

For convenience, a genotype file with genotypes re-coded as 0, 1 or 2 (the number of copies of the second allele) is given in xvec_day4.inp. For the 325 bulls, phenotypes for protein % in their daughters milk are given in the file yvec_day4.inp.

Now we write a small R script to read in the data, and fit a regression on the number of 2 alleles for each SNP.

To start a new script, click file and then New Script. Remember to save your script.

Then read in the data. The easiseast way to do this is to set your work directory to whever the files are stored first, then read in the data as a table:

```
setwd("C:/course_piacenza")
phenotypes <- read.table("yvec_day4.inp",header=F)  #No header on file
genotypes <- read.table("xvec_day4.inp",header=F)
```

Now for each SNP we are going to fit the model

$$\mathbf{y} = mu + \mathbf{X}b + \mathbf{e}$$

Where **y** are the phenotypes, mu is the mean, **X** is the design matrix allocating phenotypes to genotypes for each SNP, b is the effect of the SNP and **e** is a vecot of random residuals. This can be done in R with the lm command (for linear model)

Lets fit the first SNP. We can do this as

```
lm(phenotypes[,1] ~ genotypes[,1])
```

The [,1] for genotypes tells R to use the first column of genotypes, eg the first SNP

The result gives the intercept (mean), and the regression coefficent, which in our case is the effect of the 2 allele. If you want just the regression coefficent returned,
`lm(phenotypes[,1] ~ genotypes[,1])$coeff[2]`

Now we would like to know how significant the SNP is. We can get this with the anova command,

```
anova(lm(phenotypes[,1] ~ genotypes[,1]))
```

If you want just the P value returned,

```
anova(lm(phenotypes[,1] ~ genotypes[,1]))$P[1]
```

Now to run the genome wide association study, get the effect of each SNP and it's P value and store them. This can be done by writing a loop for the number of SNP (10) and fitting the models above each time.

Now read in the map file (map_10_markers.txt).

Plot –log10(P value) against map position for the SNP. Which is the most siginficant SNP(s). Can you explain this result in terms of the linkage disequilibrium among the SNP in the previous practical?

Now plot the SNP effects against –log10 of their P values. Are the SNP with the largest effects the most significant? Why/why not. Will this always be the case in a GWAS study? And why do some of the SNP have the same effect?

6.3 Power of association studies

As we discussed in section 2, the power of association studies depends on the r^2 between the QTL and the marker we are trying to detect the QTL with, the frequency of the rare allele of the marker and the QTL, the number of phenotypic records, and the significance level we are testing the association at.

There is a program which calculates the power of an association study given all these parameters called ldDesign. The package is written in the R language.
By way of background, R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. We will use R in a windows environment. R provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, ...) and graphical techniques. There are a very large number of "packages" available for R, and one of these is the ldDesign pack.

Hit the "packages" button on the top of the screen. Then click load packages and click on ldDesign. If the package does not appear, you can install it by typing

```
> install.packages("ldDesign")
```

Then the package can be loaded.

The documentation for the ldDesign package can be found here:
(http://bg9.imslab.co.jp/Rhelp/R-2.4.0/src/library/ldDesign.html)

We will use the **luo.ld.power** function in the ldDesign package. This function performs a classical deterministic power calculation for power to detect linkage disequilibrium between a bi-allelic QTL and a bi-allelic marker, at a given significance level in a population level association study. This is based on the 'fixed model' power calculation from Luo (1998, Heredity 80, 198–208), with corrections described in Ball (2003).

To run the function:
> luo.ld.power(n, p, q, D, h2, phi, Vp = 100, alpha)

Where:

- *n* The sample size, i.e. number of individuals genotyped and tested for the trait of interest
- *p* Bi-allelic marker allele frequency
- *q* Bi-allelic QTL allele frequency
- *D* Linkage disequilibrium coefficient
- *h2* QTL `heritability', i.e. proportion of total or phenotypic variance explained by the QTL
- *phi* Dominance ratio: phi = 0 denotes purely additive, phi = 1 denotes purely dominant allele effects
- *Vp* Total or phenotypic variance: and arbitrary value may be used
- *alpha* Significance level for hypothesis tests

The function returns the power, or probability of detecting an effect, with the given parameters, at the given significance level.

One problem we will have is that the program takes as an input D instead of r^2, which is more useful to us. We can run the program at a desired level of r^2 between the marker and QTL by inputting for the value of $D = \sqrt{p(1-p)(q(1-q)r^2}$ where p and q are defined above.

For example, if we want to evaluate power at a level of r^2 of 0.2, with p=q=0.2, we would use a value of $\sqrt{0.2*(1-0.2)*0.2*(1-0.2)*0.2} = 0.072$. Now say we have n= 500 phenotypic records, the QTL explains 2.5% of the phenotypic variance, the QTL is purely additive (phi=0), and alpha is 0.05. Assume of a value of Vp of 100, though the value assumed will not affect the calculations. Then the power of the experiment is:

> luo.ld.power(500, 0.2, 0.2, 0.072, 0.025, 0, 100, 0.05)
Which should return a value of 0.277.

Now run the program with 1000 phenotypic records, p=q=0.2,h2=0.025,phi=0,Vp=100 an alpha =0.05 for r^2=0.1,0.2,0.3-1.0.

You can either do this by calculating the value of D at each level of r2 and rerunning the program, or you can write a small "script" which loops through the values of r2.

You can write such a script in notepad. The script might look like:

Script to calculate power at different levels of r2.

```
# Script to calculate power at different levels of r2.
n <- 1000
p_val <- 0.2
q_val <- 0.2
h2 <- 0.025
phi <- 0
Vp <- 100
alpha <- 0.05
for (i in 1:10) {
 r2 <- i/10
 D <- sqrt(p_val*(1-p_val)*q_val*(1-q_val)*r2)
 luo.ld.power(n, p_val, q_val, D, h2, phi, Vp, alpha)
}
```

Save your script with a *.R extension, eg power.R. To open the script, click the file tab and select "open script". You can run the script by clicking the edit tab and selection "Run all".

At what level of r^2 does the power reach 0.9 with these parameters? To determine this, you can plot the power against the level of r^2 in excel for example.

Now plot the power with 500 and 2000 records as well. What does the level of r2 need to be to get a power of 0.9 if 500 records are used. If 2000 records are used?

The next exercise is to determine the number of phenotypic records necessary to detect a QTL with power 0.9 with different levels of r^2. You can do this by looping through different numbers of phenotypic records (increments of 100 for example) in your script and keeping the r^2 constant. Plot the minimum number of records required to reach a power of 0.9 with r^2=0.1,0.2,0.3,0.4….1.0. (eg r^2 on the x axis, and number of phenotypic records required to reach a power of 0.9 with this level of r^2 on the y axis).

Do the results agree with the statement that the number of records must be increased by a factor of $1/r^2$ in order to achieve the same power as observing the QTL itself?

6.4 Genomic selection using BLUP

In this practical you will perform genomic selection in a small data set using BLUP. The data set consists of a reference population of 325 bulls with daughter yield deviations (DYDs) for protein %. This phenotype is an accurate predictor of genotype, eg the heritability is close to one. The bulls have been genotyped for 10 SNPs.

Then there are a set of 31 calves who are selection candidates for this years progeny test team. They are genotyped for the same 10 markers. Your task is to predict GEBV for these 31 selection candidates. To do this we will need to predict the effects of the 10 SNPs in the reference population, using the equations:

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

Where g are the SNP effects, 1n is a vector of ones (325 x 1, **X** is a design matrix allocating SNP genotype to records, μ is the overall mean. We will use R to solve these equations. The **X** matrix has already been built for you, and is contained in the file xvec_day4.inp. The y matrix is contained in the file yvec_day4.inp.

What you need to do is write a small R script to solve the equations. This can be done by starting the script in notepad, then opening it in the R console.

The first lines should declare the parameters number of markers and number of records. A this point we will also specify the value of lamda as 10.

```
nmarkers <- 10      #number of markers
nrecords <- 325     #number of records
lamda    <- 10     #value for lamda
```

Next we will read in the files. Change the path to the location where you have stored the files. Note that these statements should all be on one line. Have a look at these files before opening them.

```
x <-
matrix(scan("d:/iowacourse/practicals/day4/realDataExample/xvec_day4.
inp"),ncol=nmarkers,byrow=TRUE)
y <-
matrix(scan("d:/iowacourse/practicals/day4/realDataExample/yvec_day4.
inp"),byrow=TRUE)
```

So now we have the matrix x, the vector y. We still need a vector of ones and a identity matrix dimension number of markers x number of markers…..

```
ones <- array(1,c(nrecords))
ident_mat <-diag(nmarkers)
```

The next step is to build the coefficient matrix. This can be done in blocks, eg….

```
coeff <- array(0,c(nmarkers+1,nmarkers+1))
coeff[1:1, 1:1] <- t(ones)%*%ones
coeff[1:1,2:(nmarkers+1)] <- t(ones)%*%x
```

You will need to build the other blocks. You will also need to build the right hand side of the equation.

The solutions can be obtained easily by using the inbuilt function solve,

```
solution_vec <- solve(coeff,rhs)
```

Print out this vector of solutions (eg print(solution_vec)). What is the solution for the mean? Which SNP has the largest effect?

Next we want to print GEBV for the selection candidates. This is done with the equation:

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

The g_hat are the solutions for the SNP effects you have just solved. The xvector for the selection candidates is in the file xvec_prog.inp. Can you write a small R script to calculate the GEBV?

Fours years later, all the selection candidates receive a phenotypic record from a progeny test. The results are in the file yvec_prog.inp. What is the correlation between your GEBV and the TBV? (Don't expect this to be to high with only 10 SNPs).

6.5 Genomic selection using a Bayesian approach

For the first exercise, we will analyse a small data set using the method BayesA of Meuwissen et al. (2003). We will analyse the data with a script written in the R language, meuwissenBayesA.R. The script considers single markers rather than marker haplotypes, but would be easy to extend to haplotypes. The script estimates single marker effects (**g**), a variance for each of these effects (**gvar**), and overall mean **mu** and the error variance (**vare**). A description of the program is given here (descriptions in bold).

R coding of genomic selection from Meuwissen et al. (2001)

Set the number of markers, the number of markers and the number of iterations #

```
nmarkers <- 3      #number of markers
nrecords <- 25     #number of records
numit    <- 1000   #number of iterations
```

The next section reads in the data from two files. The first is the x vector, with 0 for the 1 1 SNP genotype, 1 for 1 2 and 2 for 2 2. The second file is a vector of phenotypic records. Set the path to the location of your files. -

```
x <-
matrix(scan("d:/iowacourse/practicals/day5/smallExample/xvec.inp"),nc
ol=nmarkers,byrow=TRUE)
y <-
matrix(scan("d:/iowacourse/practicals/day5/smallExample/yvec.inp"),by
row=TRUE)
```

Set up some storage vectors and matricies to store parameter values across iterations

```
gStore <- array(0,c(numit,nmarkers))
gvarStore <- array(0,c(numit,nmarkers))
vareStore <- array(0,c(numit))
muStore <- array(0,c(numit))
ittstore <- array(0,c(numit))
```

The Gibbs cycles begin.

Step 1. Initialization of g and mu, declaration of other arrays.

```
g <- array(0.01,c(nmarkers))
mu <- 0.1
gvar <- array(0.1,c(nmarkers))
ones <- array(1,c(nrecords))
e <- array(0,c(nrecords))
```

Begin the iterations

```
for (l in 1:numit) {
```

Step 2. Sample vare from an inverse chi-square posterior

```
      e <- y - x%*%g - mu   # First calculate the vector of residuals
      vare <- (t(e)%*%e)/rchisq(1,nrecords-2)
```

Step 3 Sample the mean from a normal posterior

```
      mu <- rnorm(1,(t(ones)%*%y -
t(ones)%*%x%*%g)/nrecords,sqrt(vare/nrecords))
```

Step 4. Sample the gvar from the inverse chi square posterior

```
      for (j in 1:nmarkers) {

#       gvar[j] <- (0.002+g[j]*g[j])/rchisq(1,4.012+1)  # Meuwissen
#et al. (2001) prior
#       gvar[j] <- (g[j]*g[j])/rchisq(1,1)        # Xu (2003)  #prior
        gvar[j] <- (g[j]*g[j])/rchisq(1,0.998)    # Te Braak et # al.
(2006) prior
      }
```

Step 5 Sample the g from a normal distribution

```
      z <- array(0,c(nrecords))
      for (j in 1:nmarkers) {
       gtemp <- g
       gtemp[j] <- 0
       for (i in 1:nrecords) {
        z[i] <- x[i,j]
       }
       mean <- ( t(z)%*%y-t(z)%*%x%*%gtemp-t(z)%*%ones*mu ) /
(t(z)%*%z+vare/gvar[j])   # Calculating the mean of the distribution
       g[j] <- rnorm(1,mean,sqrt(vare/(t(z)%*%z+vare/gvar[j])))
      }
```

The final step in each iteration is to store the parameter values

```
      for (j in 1:nmarkers) {
        gStore[l,j] <- g[j]
        gvarStore[l,j] <- gvar[j]
      }
      vareStore[l] <- vare
      muStore[l] <- mu
      ittstore[l] <- l
}
```

This is the end of the program.

Consider a data set with three markers. The data set was simulated as: the effect of a 2 allele at the first marker is 3, the effect of a 2 allele at the second marker is 0, and the effect of a 2 allele at the third marker was -2. The mu was 3 and the vare was 1. The data set is:

Animal	Phenotype	Marker1 allele 1	Marker1 allele 2	Marker2 allele 1	Marker 2 allele 2	Marker3 allele 1	Marker 3 allele 2
1	9.68	2	2	2	1	1	1
2	5.69	2	2	2	2	2	2
3	2.29	1	2	2	2	2	2
4	3.42	1	1	2	1	1	1
5	5.92	2	1	1	1	1	1
6	2.82	2	1	2	1	2	2
7	5.07	2	2	2	1	2	2
8	8.92	2	2	2	2	1	1
9	2.4	1	1	2	2	1	2
10	9.01	2	2	2	2	1	1
11	4.24	1	2	1	2	2	1
12	6.35	2	2	1	1	1	2
13	8.92	2	2	1	2	1	1
14	-0.64	1	1	2	2	2	2
15	5.95	2	1	1	1	1	1
16	6.13	1	2	2	1	1	1
17	6.72	2	1	2	1	1	1
18	4.86	1	2	2	1	1	2
19	6.36	2	2	2	2	2	2
20	0.81	1	1	2	1	1	2
21	9.67	2	2	1	2	1	1
22	7.74	2	2	2	1	1	2
23	1.45	1	1	2	2	2	1
24	1.22	1	1	2	1	2	1
25	-0.52	1	1	2	2	2	2

The first step is to make the files yvec.inp and xvec.inp. In the case of yvec.inp, this is simply the list of phenotypes (no headers or identifiers). For xvec.inp, the number of 2 alleles at each marker for each animal, as a 25 x 3 matrix. The first line of this file would be (for animal 1) "2 1 0".

Save these files in a convenient location. Next open the R graphical interface, and open the script "meuwissenBayesA.R". Check the number of markers is set to 3, and the number of records 25. You will have to change the path of the files as well.

Choose a number of iterations, say 1000.

Run the script using the run all command. As the script runs, it stores values for g, gvar, mu and vare for each iteration. After the script has run, you can use the plotting facilities in R to investigate changes in the parameters over iterations.

For example, to look at the effect of the third marker across iterations, you would enter the command

```
> plot(ittstore[1:1000],gStore[1:1000,1])
```

Use this command to investigate each of the parameters in turn, and determine if they appear to be fluctuating about the correct values.

We can also plot the posterior distribution, for example for the effect of the third marker. We would discard the first 100 iterations of the program as "burn in":

```
> plot(density(gStore[100:1000,1]))
```

Does the distribution appear to be normal? What about the distributions of the other parameters?

To get the mean of the distribution, you would type:
mean(gStore[100:1000,1])
Do the means of the parameters agree with the true value of these parameters?

Now a new set of animals (selection candidates without phenotypes) are genotyped for the three markers. Their genotypes are:

Animal	Marker1 allele 1	Marker1 allele 2	Marker2 allele 1	Marker2 allele 2	Marker3 allele 1	Marker3 allele 2	TBV
26	2	2	2	1	2	1	4
27	2	1	1	2	2	1	1
28	1	1	1	2	2	2	-4
29	1	2	2	2	2	1	1
30	1	1	2	2	1	2	-2
31	2	1	1	2	2	1	1
32	2	2	2	2	2	2	2
33	2	2	2	2	1	2	4
34	2	2	2	1	1	2	4
35	1	1	1	2	2	2	-4

Calculate the GEBV for these animals as:

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

What is the correlation with the True breeding values ? (given in the table above, TBV).

Next we will use the script to estimate SNP effects in the reference population in practical 5.6. So you will need to read in the x matrix in xvec_day4.inp, the y vector in yvec_day4.inp. The number of markers in the program will need to be changed to 10 and the number of records to 325.

Run the script.

The next thing you want to do is extract SNP solutions. After the script has run, you can do this by typing:
> mean(gStore[100:1000,1])
This will give you the mean value of the SNP effect for SNP 1 from iterations 100 to 1000 (eg, excluding burn in). So for SNP 6 you would type
>mean(gStore[100:1000,6]).

Compare your SNP solutions from the Bayes program to those from BLUP (practical 5.6). One of the reasons for using the Bayesian approach is to allow different variances of SNP effect across chromosome segments. In particular, the Bayes approach should set some variances (and so SNP effects) to very close to zero. Does this seem to have happened? How many QTL would you say are on the chromosome segment?

Can you predict GEBV for the selection candidates in practical 5.6 using the SNP solutions from the Bayesian approach? Are they more highly correlated with the TBV than the GEBV from the BLUP approach?

Now change the R script to use the prior distribution of chromosome segment variances of effects to that of Meuwissen et al. (2001), eg. $\chi^{-2}(4.012,0.002)$. Now re-run the script. How do the SNP solutions compare with those when the Xu (2003) prior is used? Are the accuracy of the GEBV improved?

6.6 Bayesian approach a large weight at zero (BayesB)

In this exercise, we will modify the BayesA script from the previous exercise to sample from a prior distribution for the chromosome segment variances with a large weight at zero. This incorporates our prior knowledge that many of the chromosome segments will not contain any QTL with an effect on the quantitative trait.

The prior of the variance of chromosome segment effects is now

$$\sigma^2_{gi} = 0 \text{ with probability } \pi,$$
$$\sigma^2_{gi} \sim \chi^{-2}(\nu, S) \text{ with probability } (1-\pi),$$

Unlike BayesA, the posterior of the variance of chromosome segment effects does not have a known distribution and cannot be sampled directly in the Gibbs chain. We will therefore implement a Metropolis Hastings (MH) step with the Gibbs chain to sample **gvar** and **g** simultaneously.

To modify the code, you will need first specify the number of MH cycles you wish to do:

```
# Parameters
nmarkers <- 10      #number of markers
nrecords <- 325     #number of records
numit    <- 1000   #number of iterations
propSegs <- 0.66    #Prior proportion of segments having a non zero
effect
numMHCycles = 20 # Number of metropolis hastings cycles when sampling
variance of segments
```

The next step is to correct the phenotypic records for all number of MH cycles when sampling the gvar and g (Steps 4 and 5). We will store the corrected records in a vector called ycorr:

```
# Step 4.  Sample the gvar and g using Metropolis Hastings algorithm
(Independance sampling)
      for (j in 1:nmarkers) {

# First correct records for all other effects including mean and
other markers
       gtemp <- g
       gtemp[j] <- 0
       ycorr <- array(0,c(nrecords,1))
       Ival <- array(0,c(nrecords,nrecords))
       for (i in 1:nrecords) {
        ycorr[i] <- y[i] - mu
        Ival[i,i] <- vare
        for (k in 1:nmarkers) {
         ycorr[i] = ycorr[i] - x[i,k]*gtemp[k]
```

```
        }
      }
```

In this step we have also built a matrix which is nrecords x nrecords and has **vare** on the diagonal, as we will need this in the next step.

The next step is to calculate the likelihood of the data given the current gvar, before we sample a new one. The formula for the likelihood is:

$L(\mathbf{y}^* \mid \sigma^2_{gi} = \frac{1}{2\pi^{1/2n} \mid \mathbf{V} \mid^{1/2}} e - 1/2(\mathbf{y}corr' \mathbf{V}^{-1} \mathbf{y}corr)$ where $\mathbf{V} = \mathbf{X}i(\mathbf{I}\sigma^2_{gi})\mathbf{X}i' + \mathbf{I}\sigma^2_{\mathbf{e}})$ and

|V| is the determinant of **V**. In R we can do this calculation as:

```
# Now calculate likelihood with current gvar[j] p(gvar[j]|ycorr)
going into the chain
       V = (x[,j]*gvar[j])%*%t(x[,j])+Ival
       LH1 <- 1/(2*pi^(1/2*nrecords)*sqrt(det(V)))*exp(-
0.5*t(ycorr)%*%ginv(V)%*%ycorr)
```

The ginv function calculates the generalised inverse of V. You will have to load the R package MASS to get this function. (Load packages in the

It is also useful to calculate the likelihood of the data when the gvar is zero, as we will sample gvar=0 many times in the MH cycles.

```
# And likelihood if variance is zero
       V = Ival
       LH0 <-  1/(2*pi^(1/2*nrecords)*sqrt(det(V)))*exp(-
0.5*t(ycorr)%*%ginv(V)%*%ycorr)
```

Now we can run the MH cycles, sampling a new gvar, comparing the likelihood of the data with the new gvar to the old gvar. If the likelihood improves, we will replace the old gvar with the new gvar. If it does not improve, we will replace it with a probability LH(new **gvar**)/LH(old **gvar**). If we do replace gvar, we will also sample the effect of the SNP with the new gvar.

```
       for (kk in 1:numMHCycles) {
        if (runif(1,0,1)<propSegs) {      # sample segment variance
from (1-progSegs)*0 + propSegs*chi-square
# Sample new gvar[j] from driver distribution
         gvar_new <- 1/rchisq(1,4.012)
         V = (x[,j]*gvar_new)%*%t(x[,j])+Ival
         LH2 <- 1/(2*pi^(1/2*nrecords)*sqrt(det(V)))*exp(-
0.5*t(ycorr)%*%ginv(V)%*%ycorr)
```

```
         alpha <- min(LH2/LH1,1) # replace gvar with prob LH(new
#gvar)/LH(old gvar).
         if (runif(1)<alpha) {
# Acceptance
          gvar[j] = gvar_new
          LH1 <- LH2
         }
        }
        else {         # if zero variance sampled
         alpha <- min(LH0/LH1,1)
         if (runif(1)<alpha) {
# Acceptance
          gvar[j] = 0
          LH1 <- LH0
         }
        }
       }
       if (gvar[j]>0) {
        meanval <- ( t(x[,j])%*%y-t(x[,j])%*%x%*%gtemp-
t(x[,j])%*%ones*mu ) / (t(x[,j])%*%x[,j]+(vare)/gvar[j])
        g[j] <-
rnorm(1,meanval,sqrt((vare)/(t(x[,j])%*%x[,j]+(vare)/gvar[j])))
       }
       else {
        g[j] <-0
       }
      }
```

Once you have finished coding the method, save your R script as a new file (BayesB.R for example).

Now run the script with the small data set from practical 5.7 (3 markers and 25 records) Use 20 MH cycles. Look at the values sampled for each of 3 segments across the Gibbs chain. Do any of the **g** get set consistently to zero? Now choose different values for the proportion of segments set to zero and the parameters of the inverse chi square parameters where gvar new is sampled from (both these for the prior of the gvar). How sensitive are the results to the parameters of the prior distribution of the variances of chromosome segment effects?

6.7 Using Beagle to impute missing genotypes

In this practical you will use the BEAGLE program (Browning and Browning 2007) to impute from sparse genotypes to denser genotypes in a data set from a 50K dairy cattle data set.

Inspect the data. The first file, reference_50.txt, contains genotypes for 22 animals that have been genotyped for all markers. These genotypes are from chromosome 1, the first 50 markers. The first line of the file is the animal ids, from one to 22. There are two columns for each animal, one column for each allele at each marker. The second row is the genotypes for marker one, two alleles per individual. The third row is the genotypes for the second marker and so on. The genotypes are unphased at this point. The alleles are coded 1,2, and 0, with 0 for missing.

The second file to check is target_50.txt. These are genotypes for 3 animals for 5 markers, which are an evenly spaced subset of the 50 markers above (eg this would be an approximately 5K array).

The other file you will need is the map file, telling the BEAGLE program the alleles at each marker. The map file is reference_map_50.txt, the first three lines of which are

```
Hapmap43437-BTA-101873 113641                1 2
ARS-BFGL-NGS-16466 244698                    1 2
Hapmap34944-BES1_Contig627_1906 369418       1 2
```

Now to run the BEAGLE program you will need to open a command prompt, and make sure that the BEAGLE executable beagle.jar and the data files are in the same location.

Change directory
C:> D:

Change folder
D: cd <foldername>

See all files in a directory
dir

The command for running beagle with the data above, with a reference and target population, is

java -Xmx1000m -jar beagle.jar unphased=reference_50.txt unphased=target_50.txt markers=reference_map_50.txt missing=0 out=5K

Note that command is all on one line. The out command will in this case give all the out files the prefix 5K.

You will need to use the 7z program to look at the output files, as they have been zipped using a program called gzip.

The file 5K.target_50.txt.phased.gz contains the imputed, and phased genotypes. Again, there are two alleles for each marker, but these are now phased, eg the first allele of the first marker is on the same chromosome segment as the first allele of the second marker.

Now we will check how accurately BEAGLE has imputed the missing genotypes. The file target_true.txt contains the real genotypes at all markers for our three "target" animals.

Compare the true and imputed genotypes to calculate an accuracy of imputation. There are a few steps to doing this, which can be done either in excel or R. For excel, paste the true genotypes and the imputed genotypes beside each other in a spreadsheet. As the true genotypes are unphased (eg the alleles could be in any order), in order to compare the genotypes you will need to calculate basically an X matrix for both three true and imputed genotypes. This matrix has dimensions number of markers (50) by number of individuals (3). The elements are the number of two alleles, which can be calculated from the genotypes as allele1 + allele 2 – 2.

Eg. If for the first animal at the first marker, the genotype was 1,1, the element of X would be zero.

Calculate a separate X for both the true genotypes and the imputed genotypes.

Then count up the number of genotypes that are the same in the imputed and true genotypes.

For example, for two markers the true and imputed X matrix could be (each column is an animal)

True Imputed
1 0 1 1 1 0
1 2 1 1 2 1

Then the accuracies of imputation for each animal are 2/2=100% for animal one, 1/2 = 50% for animal 2 and 1/2 =50% for animal 3. The counting up can be done with an IF statement in excel.

What are the accuracies of imputation for our three target animals? What are some possible reasons for the differences in accuracy of imputation?
Finally, have at look at the file 50_SNP.reference_50.txt.gprobs.gz. This file gives for each animal, the probability of each genotype for each animal. This is a measure of the uncertainty of the imputation. Each line of the file contains the marker name, the two alleles at the marker (1 and 2 for all markers in our case), then for each animal the probabilities of the three genotypes 11, 12 and 22 (or 0,1,2 in our X matrix).

Can you find maker where there is a lot of uncertainty in the imputation?

How would you build an X matrix for the genomic selection methods that takes account of uncertainty of imputation?

6.8 Validation of Genomic Prediction

In this practical you will combine what you learnt in Practicals 6.2 Genome-wide Association and 6.4 genomic BLUP.
Validating the performance of genomic predictions is important to demonstrate that it works. There are some principles to follow that will ensure proper validation. These are outlined in Section 3.10 of the notes. The main principle is to consider who the selection candidates are and what information they will have available when we want to predict their GEBV. In the majority of cases, the selection candidates will have no phenotypic information.
We then choose a validation population with phenotypes and genotypes. However, we need to make sure that the validation phenotypes are not influencing our results. There are several validation mistakes that can upwardly bias the genomic selection accuracy. Here we will do two examples, both of which are cases where phenotypic information of the validation individuals inappropriately influenced the accuracy of genomic selection. You will use the same input files as Practical 6.4.

Example 1: Validation individuals' phenotypes are used in the reference population

You can either use the BLUP with SNP effects or genomic BLUP with a relationship matrix to predict the GEBVs. In either case, you will need to combine the two genotype files (xvec_day4.inp and xvec_prog.inp) and the two phenotype files (yvec_day4.inp and yvec_prog.inp) to create one larger reference population.

To combine them use the commands below after you have read them in:

```
ycombined <- rbind(yref,yprog)
xcombined <-rbind(xref,xprog)
```

You can then use the same process as Practical 6.4 to predict GEBVs and calculate the accuracies. Compare these accuracies with the accuracies in Practical 6.4. Why are they higher?

Example 2: Choosing a subset of SNP for genomic selection in a non-independent GWAS

It is sometimes desirable to reduce the number of SNP in an analysis. For example, when the genotyping budget is limited. One way to choose a subset of SNP is with an association study. In Practical 6.2 you ran an association study on the same data. Here you will run the association study in two ways, once on the exact same data as Practical 6.2 and once on the combined files from above. In both cases, you choose the best 8 SNP based on p values to take forward into your genomic prediction analysis.

1. Run association study only with reference genotypes and phenotypes

First, you run the association study with only the reference data (yvec_day4.inp, xvec_day4.inp) as you did in P6.2. Then you only use the top 8 SNP based on pvalue.

You can take a subset of a matrix by sorting the vector of pvalues and creating an index that selects only the best 8 SNP by pvalue.

```
p2=sort(pvalues)  #sorts pvalues
index=pvalues <= p2[8]  #creates TRUE or FALSE index
nmarkers=length(index[index==TRUE]) #sets nmarker to count of TRUE in
index
xnew=array(0,c(nrecords,nmarkers))
xnew[,1:nmarkers]<-x[,index]        #puts subset of SNP into xnew
```

You then estimate the marker effects using the new x and in the end you validate with xvec_prog.inp and yvec_prog.inp

2. Run association study in combined data

Here you run the association study in the combined data. You can combine the reference and validation files as you did in Example 1. You then again select the best 8 SNP based on pvalue and put them in a new xmatrix. You run the genomic prediction step with ONLY the reference genotypes (the subset of xvec_day4.inp) and phenotypes. In the end you validate with xvec_prog.inp and yvec_prog.inp.

Compare the accuracies in 1 and 2. Why are they different? Which way is truly independent?

7. Acknowledgments

The assistance of a number of people in preparing these notes is gratefully acknowledged. Many thanks to Mike Goddard, for inspiration and a continuous flow of excellent ideas. Thank you to Sander De Roos, Iona MacLeod and Kathryn Kemper for reading earlier versions of the notes.

8. References

Aguilar I., Misztal I., Johnson D.L., Legarra A., Tsuruta S. & Lawlor T.J. (2010) Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. *Journal of Dairy Science* **93**, 743-52.

Amer P.R. & Banos G. (2010) Implications of avoiding overlap between training and testing data sets when evaluating genomic predictions of genetic merit. *Journal of Dairy Science* **93**, 3320-30.

Andersson L. & Georges M. (2004) Domestic-animal genomics: deciphering the genetics of complex traits. *Nat Rev Genet* **5**, 202-12.

Benjamini Y. & Hochberg T. (1995) Controlling the false discovery rate: a practical and powerful approach to muliple testing. *J.Royal Stat.Soc.* **85**, 289-300.

Boichard D., Ducrocq V., Fritz S. & Colleau J.J. (2010) Where is dairy breeding going? A vision of the future. *Interbull Bulletin* **41**, 63-8.
Browning B.L. & Browning S.R. (2009) A Unified Approach to Genotype Imputation and Haplotype-Phase Inference for Large Data Sets of Trios and Unrelated Individuals. *American Journal of Human Genetics* **84**, 210-23.
Browning S.R. & Thompson E.A. (2012) Detecting Rare Variant Associations by Identity by Descent Mapping in Case-control Studies. *Genetics*.
Buch L.H. (2011) Genetic improvement of functional traits in dairy cattle breeding schemes with genomic selection. In: *Aarhus Faculty of Science and Technology*. Aarhus University, Aarhus.
Buch L.H., Sørensen A.C., Lassen J., Berg P., Eriksson J.Å., Jakobsen J.H. & Sørensen M.K. (2011) Hygiene-related and feed-related hoof diseases show different patterns of genetic correlations to clinical mastitis and female fertility. *Journal of Dairy Science* **94**, 1540-51.
Calus M.P.L., Meuwissen T.H.E., de Roos A.P.W. & Veerkamp R.F. (2008) Accuracy of genomic selection using different methods to define haplotypes. *Genetics* **178**, 553-61.
Christensen O. & Lund M. (2009) Genomic prediction when some animals are not genotyped. *Genetics Selection Evolution* **42**, 2.
Churchill G.A. & Doerge R.W. (1994) Empirical threshold values for quantitative trait mapping. *Genetics* **138**, 963-71.
Cleveland M.A., Hickey J.M. & Forni S. (2012) A Common Dataset for Genomic Analysis of Livestock Populations. *G3: Genes|Genomes|Genetics* **2**, 429-35.
Consortium G.P. (2010) A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-73.
Consortium T.B.H. (2009) Genome-Wide Survey of SNP Variation Uncovers the Genetic Structure of Cattle Breeds. *Science* **324**, 528-32.
Daetwyler H.D., Calus M.P.L., Pong-Wong R., de los Campos G. & Hickey J.M. (2013) Genomic Prediction in Animals and Plants: Simulation of Data, Validation, Reporting, and Benchmarking. *Genetics* **193**, 347-65.
Daetwyler H.D., Kemper K.E., van der Werf J.H.J. & Hayes B.J. (2012a) Components of the accuracy of genomic prediction in a multi-breed sheep population. *Journal of Animal Science* **90**, 3375-84.
Daetwyler H.D., Swan A.A., van der Werf J.H.J. & Hayes B.J. (2012b) Accuracy of pedigree and genomic predictions of carcass and novel meat quality traits in multi-breed sheep data assessed by cross-validation. *Genetics Selection Evolution* **44**, 33.
Daetwyler H.D., Villanueva B., Bijma P. & Woolliams J.A. (2007) Inbreeding in genome-wide selection. *J.Anim.Breed.Genet.* **124**, 369-76.
Daetwyler H.D., Wiggans G.R., Hayes B.J., Woolliams J.A. & Goddard M.E. (2011) Imputation of Missing Genotypes From Sparse to High Density Using Long-Range Phasing. *Genetics* **189**, 317-27.
Danecek P., Auton A., Abecasis G., Albers C.A., Banks E., DePristo M.A., Handsaker R., Lunter G., Marth G., Sherry S.T., McVean G., Durbin R. & Group G.P.A. (2011) The Variant Call Format and VCFtools. *Bioinformatics*.
Darvasi A. & Soller M. (1997) A Simple Method to Calculate Resolving Power and Confidence Interval of QTL Map Location. *Behavior Genetics* **27**, 125-32.
Dassonneville R., Brøndum R.F., Druet T., Fritz S., Guillaume F., Guldbrandtsen B., Lund M.S., Ducrocq V. & Su G. (2011) Effect of imputing markers from a

low-density chip on the reliability of genomic breeding values in Holstein populations. *Journal of Dairy Science* **94**, 3679-86.
De Roos A.P.W., Hayes B.J., Spelman R.J. & Goddard M.E. (2008) Linkage disequilibrium and persistence of phase in Holstein-Friesian, Jersey and Angus cattle. *Genetics* **179**, 1503-12.
de Roos A.P.W., Schrooten C., Veerkamp R.F. & van Arendonk J.A.M. (2011) Effects of genomic selection on genetic improvement, inbreeding, and merit of young versus proven bulls. *Journal of Dairy Science* **94**, 1559-67.
Dekkers J.C.M. (2007) Prediction of response from marker-assisted and genomic selection using selection index theory. *J.Anim.Breed.Genet.* **124**, 331-41.
Druet T. & Georges M. (2010) A Hidden Markov Model Combining Linkage and Linkage Disequilibrium Information for Haplotype Reconstruction and Quantitative Trait Locus Fine Mapping. *Genetics* **184**, 789-98.
Druet T., Schrooten C. & de Roos A.P.W. (2010) Imputation of genotypes from different single nucleotide polymorphism panels in dairy cattle. *Journal of Dairy Science* **93**, 5443-54.
Du F.X., Clutter A.C. & Lohuis M.M. (2007) Characterizing linkage disequilibrium in pig populations. *Int J Biol Sci* **3**, 166-78.
Dunner S., Miranda M.E., Amigues Y., Canon J., Georges M., Hanset R., Williams J. & Menissier F. (2003) Haplotype diversity of the myostatin gene among beef cattle breeds. *Genetics Selection Evolution* **35**, 103 - 18.
Dunning A.M., Durocher F., Healey C.S., Teare M.D., McBride S.E., Carlomagno F., Xu C.-F., Dawson E., Rhodes S., Ueda S., Lai E., Luben R.N., Van Rensburg E.J., Mannermaa A., Kataja V., Rennart G., Dunham I., Purvis I., Easton D. & Ponder B.A.J. (2000) The Extent of Linkage Disequilibrium in Four Populations with Distinct Demographic Histories. *The American Journal of Human Genetics* **67**, 1544-54.
Erbe M., Hayes B.J., Matukumalli L.K., Goswami S., Bowman P.J., Reich C.M., Mason B.A. & Goddard M.E. (2012) Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. *Journal of Dairy Science* **95**, 4114-29.
Ewing B. & Green P. (2000) Analysis of expressed sequence tags indicates 35,000 human genes. *Nat Genet* **25**, 232-4.
Fernando R.L., Nettleton D., Southey B.R., Dekkers J.C., Rothschild M.F. & Soller M. (2004) Controlling the proportion of false positives in multiple dependent tests. *Genetics* **166**, 611-9.
Fischer R.A. (1918) The correlation between realtives: the supposition of Mendelian inheritance. *Trans Royal Soc Edin* **52**, 399.
Gianola D., Fernando R.L. & Stella A. (2006) Genomic-Assisted Prediction of Genetic Value With Semiparametric Procedures. *Genetics* **173**, 1761-76.
Gilmour A.R., Gogel B., Cullis B.R. & Thompson R. (2009) *2009 ASReml user guide release 3.0*. VSN International Ltd., Hemel Hempstead.
Goddard M. (1991) Mapping genes for quantitative traits using linkage disequilibrium. *Genetics Selection Evolution* **23**, S131 - S4.
Goddard M.E., Chamberlain A.J. & Hayes B.J. (2006) Can the same markers be used across multiple breeds? In: *8th WCGALP*, Belo Horizonte, Brazil.
Grapes L., Dekkers J.C., Rothschild M.F. & Fernando R.L. (2004) Comparing linkage disequilibrium-based methods for fine mapping quantitative trait loci. *Genetics* **166**, 1561-70.

Grapes L., Firat M.Z., Dekkers J.C., Rothschild M.F. & Fernando R.L. (2006) Optimal haplotype structure for disequilibirum-based fine mapping of quantitative trait loci using identity by descent. *Genetics* **172**, 1955-65.

Grundy B., Villanueva B. & Woolliams J.A. (1998) Dynamic selection procedures for constrained inbreeding and their consequences for pedigree development. *Genetical Research* **72**, 159-68.

Gudbjartsson D.F., Walters G.B., Thorleifsson G., Stefansson H., Halldorsson B.V., Zusmanovich P., Sulem P., Thorlacius S., Gylfason A., Steinberg S., Helgadottir A., Ingason A., Steinthorsdottir V., Olafsdottir E.J., Olafsdottir G.H., Jonsson T., Borch-Johnsen K., Hansen T., Andersen G., Jorgensen T., Pedersen O., Aben K.K., Witjes J.A., Swinkels D.W., Heijer M.d., Franke B., Verbeek A.L.M., Becker D.M., Yanek L.R., Becker L.C., Tryggvadottir L., Rafnar T., Gulcher J., Kiemeney L.A., Kong A., Thorsteinsdottir U. & Stefansson K. (2008) Many sequence variants affecting diversity of adult human height. *Nat Genet* **40**, 609-15.

Habier D., Fernando R.L. & Dekkers J.C. (2009) Genomic Selection Using Low-density Marker Panels. *Genetics*.

Habier D., Fernando R.L. & Dekkers J.C.M. (2007) The impact of genetic relationship information on genome-assisted breeding values. *Genetics* **177**, 2389-97.

Habier D, Fernando RL, Kizilkaya K, Garrick DJ. 2011. Extension of the bayesian alphabet for genomic selection. BMC Bioinformatics. 12:186.

Harris B.L., Johnson D.L. & Spelman R.J. (2008) Genomic selection in New Zealand and the implications for national genetic evaluation. In: *36th ICAR Conference*, Niagara Falls, USA.

Hayes B.J., Bowman P.J., Daetwyler H.D., Kijas J.W. & van der Werf J.H.J. (2011) Accuracy of genotype imputation in sheep breeds. *Animal Genetics*, no-no.

Hayes B.J., Chamberlain A.J. & Goddard M.E. (2006) Use of linkage markers in linkage disequilibrium with QTL in breeding programs. In: *8th WCGALP*, Belo Horizonte, Brazil.

Hayes B.J., Chamberlain A.J., McPartlan H., MacLeod I., Sethuraman L. & Goddard M.E. (2007) Accuracy of marker-assisted selection with single markers and marker haplotypes in cattle. *Genet.Res.* **89**, 215-20.

Hayes B.J. & Goddard M.E. (2001) The distribution of the effects of genes affecting quantitative traits in livestock. *Genetics Selection Evolution* **33**, 209-29.

Hayes B.J. & Goddard M.E. (2008) Technical note: prediction of breeding values using marker-derived relationship matrices. *J.Anim Sci.* **86**, 2089-92.

Hayes B.J., Visscher P.M., McPartlan H.C. & Goddard M.E. (2003) Novel Multilocus Measure of Linkage Disequilibrium to Estimate Past Effective Population Size. *Genome Res.* **13**, 635-43.

Heifetz E.M., Fulton J.E., O'Sullivan N., Zhao H., Dekkers J.C.M. & Soller M. (2005) Extent and Consistency Across Generations of Linkage Disequilibrium in Commercial Layer Chicken Breeding Populations. *Genetics* **171**, 1173-81.

Henderson C.R. (1984) *Applications of linear model in animal breeding*. University of Guelph, Guelph.

Hickey J., Kinghorn B., Tier B., Wilson J., Dunstan N. & van der Werf J. (2011) A combined long-range phasing and long haplotype imputation method to impute phase for SNP genotypes. *Genetics Selection Evolution* **43**, 12.

Hickey J.M., Veerkamp R.F., Calus M.P.L., Mulder H.A. & Thompson R. (2009) Estimation of prediction error variances via Monte Carlo sampling methods using different formulations of the prediction error variance. *Genetics Selection Evolution* **41**.

Hill W.G. (1981) Estimation of effective population size from data on linkage disequilibrium. *Genetics Research* **38**, 209-16.

Hill W.G. & Robertson A. (1968) Linkage disequilibrium in finite populations. *Theoretical Applied Genetics* **38**, 226-31.

Howie B.N., Donnelly P. & Marchini J. (2009) A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. *PLoS Genet* **5**, e1000529.

Huang X., Wei X., Sang T., Zhao Q., Feng Q., Zhao Y., Li C., Zhu C., Lu T., Zhang Z., Li M., Fan D., Guo Y., Wang A., Wang L., Deng L., Li W., Lu Y., Weng Q., Liu K., Huang T., Zhou T., Jing Y., Li W., Lin Z., Buckler E.S., Qian Q., Zhang Q.-F., Li J. & Han B. (2010) Genome-wide association studies of 14 agronomic traits in rice landraces. *Nat Genet* **42**, 961-7.

Kim J.-J., Lee H.-I., Park T., Kim K., Lee J.-E., Cho N.H., Shin C., Cho Y.S., Lee J.-Y., Han B.-G., Yoo H.-W. & Lee J.-K. (2009) Identification of 15 loci influencing height in a Korean population. *J Hum Genet* **55**, 27-31.

Kinghorn B.P. (1998) Mate selection by groups. *J Dairy Sci* **81**, 55-63.

Kong A., Masson G., Frigge M.L., Gylfason A., Zusmanovich P., Thorleifsson G., Olason P.I., Ingason A., Steinberg S., Rafnar T., Sulem P., Mouy M., Jonsson F., Thorsteinsdottir U., Gudbjartsson D.F., Stefansson H. & Stefansson K. (2008) Detection of sharing by descent, long-range phasing and haplotype imputation. *Nature Genetics* **40**, 1068-75.

König S. & Swalve H.H. (2009) Application of selection index calculations to determine selection strategies in genomic breeding programs. *Journal of Dairy Science* **92**, 5292-303.

Kruglyak L. (1999) Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nat Genet* **22**, 139-44.

Lander E.S. & Schork N.J. (1994) Genetic dissection of complex traits. *Science* **265**, 2037-48.

Lango-Allen H (2010) Hundreds of genomic loci clustered in genomic loci and biological pathways affect human height. Nature 467: 832.

Le S.Q. & Durbin R. (2011) SNP detection and genotyping from low-coverage sequencing data on multiple diploid samples. *Genome Research*.

Legarra A., Robert-Granie C., Manfredi E. & Elsen J.-M. (2008) Performance of Genomic Selection in Mice. *Genetics* **180**, 611-8.

Lettre G., Jackson A.U., Gieger C., Schumacher F.R., Berndt S.I., Sanna S., Eyheramendy S., Voight B.F., Butler J.L., Guiducci C., Illig T., Hackett R., Heid I.M., Jacobs K.B., Lyssenko V., Uda M., Boehnke M., Chanock S.J., Groop L.C., Hu F.B., Isomaa B., Kraft P., Peltonen L., Salomaa V., Schlessinger D., Hunter D.J., Hayes R.B., Abecasis G.R., Wichmann H.E., Mohlke K.L. & Hirschhorn J.N. (2008) Identification of ten loci associated with height highlights new biological pathways in human growth. *Nat Genet* **40**, 584-91.

Lewontin R.C. (1964) The interaction of selection and linkage. I. General considerations; heterotic models. *Genetics* **49**, 49-67.

Li H, Peng Z, Yang X, Wang W, Fu J, Wang J, Han Y, Chai Y, Guo T, Yang N, Liu J, Warburton ML, Cheng Y, Hao X, Zhang P, Zhao J, Liu Y, Wang G, Li J,

Yan J. 2013. Genome-wide association study dissects the genetic architecture of oil biosynthesis in maize kernels. Nat Genet. 45:43-50

Li Y., Willer C.J., Ding J., Scheet P. & Abecasis G.R. (2010) MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genetic Epidemiology* **34**, 816-34.

Lillehammer M., Meuwissen T.H.E. & Sonesson A.K. (2011) A comparison of dairy cattle breeding designs that use genomic selection. *Journal of Dairy Science* **94**, 493-500.

Lund M., de Ross S., de Vries A., Druet T., Ducrocq V., Fritz S., Guillaume F., Guldbrandtsen B., Liu Z., Reents R., Schrooten C., Seefried F. & Su G. (2011) A common reference population from four European Holstein populations increases reliability of genomic predictions. *Genetics Selection Evolution* **43**, 43.

Luo Z.W. (1998) Detecting linkage disequilibrium between a polymorphic marker locus and a trait locus in natural populations. *Heredity* **80**, 198-208.

MacLeod I.M., Hayes B.J., Savin K.W., Chamberlain A.J., McPartlan H.C. & Goddard M.E. (2010) Power of a genome scan to detect and locate quantitative trait loci in cattle using dense single nucleotide polymorphisms. *Journal of Animal Breeding and Genetics* **127**, 133-42.

Malosetti M., van der Linden C.G., Vosman B. & van Eeuwijk F.A. (2007) A Mixed-Model Approach to Association Mapping Using Pedigree Information With an Illustration of Resistance to Phytophthora infestans in Potato. *Genetics* **175**, 879-89.

Marchini J. & Howie B. (2008) Comparing Algorithms for Genotype Imputation. *American Journal of Human Genetics* **83**, 535-9.

Marchini J. & Howie B. (2010) Genotype imputation for genome-wide association studies. *Nat Rev Genet* **11**, 499-511.

Martinez O. & Curnow R.N. (1992) Estimating the locations and sizes of the effects of quantitative trait loci using flanking markers. *Theor Appl Genet* **85**, 480-8.

Mc Hugh N., Meuwissen T.H.E., Cromie A.R. & Sonesson A.K. (2011) Use of female information in dairy cattle breeding programs. *Journal of Dairy Science* **in press**.

McRae A.F., McEwan J.C., Dodds K.G., Wilson T., Crawford A.M. & Slate J. (2002) Linkage disequilibrium in domestic sheep. *Genetics* **160**, 1113-22.

McVean G. (2009) A Genealogical Interpretation of Principal Components Analysis. *PLoS Genet* **5**, e1000686.

Meuwissen T. & Goddard M. (2004) Mapping multiple QTL using linkage disequilibrium and linkage analysis information and multitrait data. *Genetics Selection Evolution* **36**, 261 - 79.

Meuwissen T. & Goddard M. (2010) Accurate Prediction of Genetic Values for Complex Traits by Whole-Genome Resequencing. *Genetics* **185**, 623-31.

Meuwissen T.H. & Goddard M.E. (2001) Prediction of identity by descent probabilities from marker-haplotypes. *Genet.Sel Evol.* **33**, 605-34.

Meuwissen T.H.E. (1997) Maximizing the response of selection with a predefined rate of inbreeding. *Journal of Animal Science* **75**, 934-40.

Meuwissen T.H.E., Hayes B.J. & Goddard M.E. (2001) Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157**, 1819-29.

Meuwissen T.H.E., Karlsen A., Lien S., Olsaker I. & Goddard M.E. (2002) Fine mapping of a quantitative trait locus for twinning rate using combined linkage and linkage disequilibrium mapping. *Genetics* **161**, 373-9.

Misztal I. & Wiggans G.R. (1988) Approximation of prediction error variance in large-scale animal models. *J.Dairy Sci.* **71**, 27-32.
Nicholas F.W. & Smith C. (1983) Increased Rates of Genetic Change in Dairy-Cattle by Embryo Transfer and Splitting. *Animal Production* **36**, 341-53.
Patterson N., Price A.L. & Reich D. (2006) Population Structure and Eigenanalysis. *PLoS Genet* **2**, e190.
Pedersen L.D., Sørensen A.C. & Berg P. (2009a) Marker-assisted selection can reduce true as well as pedigree-estimated inbreeding. *Journal of Dairy Science* **92**, 2214-23.
Pedersen L.D., Sørensen A.C. & Berg P. (2010) Marker-assisted selection reduces expected inbreeding but can result in large effects of hitchhiking. *Journal of Animal Breeding and Genetics* **127**, 189-98.
Pedersen L.D., Sørensen A.C., Henryon M., Ansari-Mahyari S. & Berg P. (2009b) ADAM: A computer program to simulate selective breeding schemes for animals. *Livestock Science* **121**, 343-4.
Piyasatian N., Fernando R.L. & Dekkers J.C. (2007) Genomic selection for marker-assisted improvement in line crosses. *Theor.Appl.Genet.*
Pritchard J.K. & Przeworski M. (2001) Linkage Disequilibrium in Humans: Models and Data. *The American Journal of Human Genetics* **69**, 1-14.
Pritchard J.K., Stephens M., Rosenberg N.A. & Donnelly P. (2000) Association Mapping in Structured Populations. *The American Journal of Human Genetics* **67**, 170-81.
Pryce J.E., Arias J., Bowman P.J., Macdonald K.A., Waghorn G.C., Wales W.J., Williams Y.L., Spelman R.J. & Hayes B.J. (2011a) Accuracy of genomic selection for residual feed intake and 250-day liveweight in dairy heifers using high-density (630k) SNP. In: *Association for Advancement of Animal Breeding and Genetics*, Perth, Australia.
Pryce J.E., Bolormaa S., Chamberlain A.J., Bowman P.J., Savin K., Goddard M.E. & Hayes B.J. (2010a) A validated genome-wide association study in 2 dairy cattle breeds for milk production and fertility traits using variable length haplotypes. *Journal of Dairy Science* **93**, 3331-45.
Pryce J.E. & Daetwyler H.D. (2012) Designing dairy cattle breeding schemes under genomic selection: a review of international research. *Animal Production Science*, -.
Pryce J.E., Goddard M.E., Raadsma H.W. & Hayes B.J. (2010b) Deterministic models of breeding scheme designs that incorporate genomic selection. *Journal of Dairy Science* **93**, 5455-66.
Pryce J.E., Hayes B.J., Bolormaa S. & Goddard M.E. (2011b) Polymorphic Regions Affecting Human Height Also Control Stature in Cattle. *Genetics* **187**, 981-4.
Rabiner L.R. (1989) A tutorial on hidden markov models and selected applications in speech recognition. *Proc IEEE* **77**, 257-86.
Reich D.E., Cargill M., Bolk S., Ireland J., Sabeti P.C., Richter D.J., Lavery T., Kouyoumjian R., Farhadian S.F., Ward R. & Lander E.S. (2001) Linkage disequilibrium in the human genome. *Nature* **411**, 199-204.
Riquet J., Coppieters W., Cambisano N., Arranz J.-J., Berzi P., Davis S.K., Grisart B., Farnir F., Karim L., Mni M., Simon P., Taylor J.F., Vanmanshoven P., Wagenaar D., Womack J.E. & Georges M. (1999) Fine-mapping of quantitative trait loci by identity by descent in outbred populations: Application to milk production in dairy cattle. *Proceedings of the National Academy of Sciences* **96**, 9252-7.

Robertson A. & Rendel J.M. (1950) The use of progeny testing with artificial insemination in dairy cattle. *Journal of Genetics* **50**, 21-31.
Rothschild M.F., Larson R.G., Jacobson C. & Pearson P. (1991) Pvull polymorphism at the porcine oestrogen receptor locus (ESR). *Animal Genetics* **22**, 448.
Saatchi M., McClure M., McKay S., Rolf M., Kim J., Decker J., Taxis T., Chapple R., Ramey H., Northcutt S., Bauck S., Woodward B., Dekkers J., Fernando R., Schnabel R., Garrick D. & Taylor J. (2011) Accuracies of genomic breeding values in American Angus beef cattle using K-means clustering for cross-validation. *Genetics Selection Evolution* **43**, 40.
Schaeffer L.R. (2006) Strategy for applying genome-wide selection in dairy cattle. *J.Anim Breed.Genet.* **123**, 218-23.
Scheet P. & Stephens M. (2006) A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase. *American Journal of Human Genetics* **78**, 629-44.
Shrimpton A.E. & Robertson A. (1988) The isolation of polygenic factors controlling bristle score in Drosophila melanogaster. II. Distribution of third chromosome bristle effects within chromosome sections. *Genetics* **118**, 445-59.
Sonesson A.K., Woolliams J.A. & Meuwissen T.H.E. (2010) Maximising Genetic Gain Whilst Controlling Rates of Genomic Inbreeding Using Genomic Optimum Contribution Selection. In: *World Congress of Genetics Applied to Livestock Production*, Leipzig, Germany.
Sorensen A.C. & Sorensen M.K. (2009) Inbreeding rates in breeding programs with different strategies for using genomic selection. *Interbull Bulletin* **40**, 94-7.
Soyeurt H., Dehareng F., Gengler N., McParland S., Wall E., Berry D.P., Coffey M. & Dardenne P. (2011) Mid-infrared prediction of bovine milk fatty acids across multiple breeds, production systems, and countries. *Journal of Dairy Science* **94**, 1657-67.
Spelman R.J., Ford C.A., McElhinney P., Gregory G.C. & Snell R.G. (2002) Characterization of the DGAT1 Gene in the New Zealand Dairy Population. *Journal of Dairy Science* **85**, 3514-7.
Spielman R.S., McGinnis R.E. & Ewens W.J. (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet* **52**, 506-16.
Storey J.D. (2002) A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **64**, 479-98.
Sved J.A. (1971) Linkage disequilibrium and homozygosity of chromosome segments in finite populations. *Theoretical Population Biology* **2**, 125-41.
Tenesa A., Navarro P., Hayes B.J., Duffy D.L., Clarke G.M., Goddard M.E. & Visscher P.M. (2007) Recent human effective population size estimated from linkage disequilibrium. *Genome Research* **17**, 520-6.
ter Braak C.J.F., Boer M.P. & Bink M.C.A.M. (2005) Extending Xu's Bayesian model for estimating polygenic effects using markers of the entire genome. *Genetics* **170**, 1435-8.
van der Werf J.H.J., Kinghorn B.P. & Banks R.G. (2010) Design and role of an information nucleus in sheep breeding programs. *Animal Production Science* **50**, 998-1003.
VanRaden P.M., Van Tassell C.P., Wiggans G.R., Sonstegard T.S., Schnabel R.D., Taylor J.F. & Schenkel F.S. (2009) Invited review: Reliability of genomic predictions for North American Holstein bulls. *Journal of Dairy Science* **92**, 16-24.

Villanueva B., Bijma P. & Woolliams J.A. (2000) Optimal mass selection policies for schemes with overlapping generations and restricted inbreeding. *Genet.Sel Evol.* **32**, 339-55.

Weedon M.N., Lango H., Lindgren C.M., Wallace C., Evans D.M., Mangino M., Freathy R.M., Perry J.R.B., Stevens S., Hall A.S., Samani N.J., Shields B., Prokopenko I., Farrall M., Dominiczak A., Johnson T., Bergmann S., Beckmann J.S., Vollenweider P., Waterworth D.M., Mooser V., Palmer C.N.A., Morris A.D., Ouwehand W.H., Caulfield M., Munroe P.B., Hattersley A.T., McCarthy M.I. & Frayling T.M. (2008) Genome-wide association analysis identifies 20 loci that influence adult height. *Nature Genetics* **40**, 575-83.

Weigel K.A., de los Campos G., Vazquez A.I., Rosa G.J.M., Gianola D. & Van Tassell C.P. (2010) Accuracy of direct genomic values derived from imputed single nucleotide polymorphism genotypes in Jersey cattle. *Journal of Dairy Science* **93**, 5423-35.

Weir B.S. & Hill W.G. (1980) Effect of mating structure on variation in linkage disequilibrium. *Genetics* **95**, 477-88.

Weller J.I., Song J.Z., Heyen D.W., Lewin H.A. & Ron M. (1998) A new approach to the problem of multiple comparisons in the genetic dissection of complex traits. *Genetics* **150**, 1699-706.

Whittaker J.C., Thompson R. & Denham M.C. (2000) Marker-assisted selection using ridge regression. *Genet.Res.* **75**, 249-52.

Winkelman A.M. & Spelman R.J. (2010) Response using genome-wide selection in dairy cattle breeding schemes. In: *World Congress for Genetics Applied to Livestock Production*, p. 0290, Leipzig, Germany.

Wolc A., Stricker C., Arango J., Settar P., Fulton J., O'Sullivan N., Preisinger R., Habier D., Fernando R., Garrick D., Lamont S. & Dekkers J. (2011) Breeding value prediction for production traits in layer chickens using pedigree or genomic relationships in a reduced animal model. *Genetics Selection Evolution* **43**, 5.

Woolliams J.A., Pong-Wong R. & Villanueva B. (2002) Strategic optimisation of short- and long-term gain and inbreeding in MAS and non-MAS schemes. In: *7th world congress of genetics applied to livestock production*, pp. 23-02.

Wray N.R. & Goddard M.E. (1994) Increasing Long-Term Response to Selection. *Genetics Selection Evolution* **26**, 431-51.

Wray N.R., Yang J., Hayes B.J., Price A., Goddard M.E. & Visscher P.M. (2013) Working title: Pitfalls of predicting complex traits from SNPs. *submitted*.

Xu S. & Jia Z. (2007) Genomewide Analysis of Epistatic Effects for Quantitative Traits in Barley. *Genetics* **175**, 1955-63.

Xu S.Z. (2003) Estimating polygenic effects using markers of the entire genome. *Genetics* **163**, 789-801.

Yang J., Ferreira T., Morris A.P., Medland S.E., Madden P.A.F., Heath A.C., Martin N.G., Montgomery G.W., Weedon M.N., Loos R.J., Frayling T.M., McCarthy M.I., Hirschhorn J.N., Goddard M.E. & Visscher P.M. (2012) Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat Genet* **44**, 369-75.

Zenger K.R., Khatkar M.S., Cavanagh J.A.L., Hawken R.J. & Raadsma H.W. (2007) Genome-wide genetic diversity of Holstein Friesian cattle reveals new insights into Australian and global population variability, including impact of selection. *Animal Genetics* **38**, 7-14.

Zhang Z. & Druet T. (2010) Marker imputation with low-density marker panels in Dutch Holstein cattle. *Journal of Dairy Science* **93**, 5487-94.
Zhao H., Nettleton D., Soller M. & Dekkers J.C.M. (2005) Evaluation of linkage disequilibrium measures between multi-allelic markers as predictors of linkage disequilibrium between markers and QTL. *Genetics Research* **86**, 77-87.
Zhao H.H., Fernando R.L. & Dekkers J.C.M. (2007) Power and Precision of Alternate Methods for Linkage Disequilibrium Mapping of Quantitative Trait Loci. *Genetics* **175**, 1975-86.