BEN HAYES
COURSE NOTES

TOULOUSE

12-16 SEPTEMBRE 2011

1. Linkage disequilibrium in livestock populations

1.1 A brief history of QTL mapping

The vast majority of economically important traits in livestock and aquaculture production systems 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 (eg 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 and 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 a many of small effect (Shrimpton and Robertson 1998, Hayes and Goddard 2001). In Figure 1.1, the size of quantitative trait loci (QTL) reported in QTL mapping experiments in both pigs and dairy cattle is shown. These histograms are not the true distribution of QTL effects however, they are only able to observe effects above a certain size determined by the amount of environmental noise, and the effects are estimated with error. In Figure 1.1. B, the distribution of effects adjusted for both these factors is displayed. The distributions in Figure 1.1 B indicate there are many genes of small effect, and few of large effect. The search for these loci, particularly those of moderate to large effect, 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.

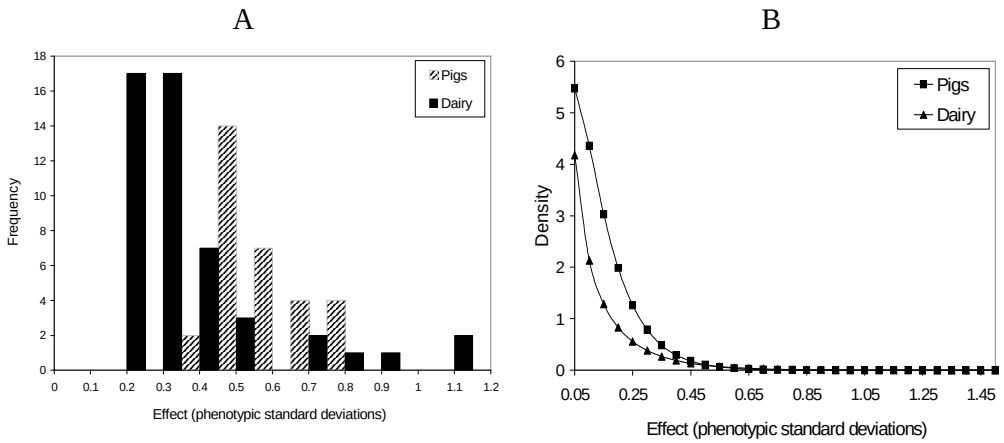


Figure 1.1 A. Distribution of additive (QTL) effects from pig experiments, scaled by the standard deviation of the relevant trait, and distribution of gene substitution (QTL) effects from dairy experiments scaled by the standard deviation of the relevant trait. B. Gamma Distribution of QTL effect from pig and dairy experiments, fitted with maximum likelihood.

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.

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 recently, the number of DNA markers identified in livestock genome was comparatively limited, and the cost of genotyping the markers was high. This 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 1.2. 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.

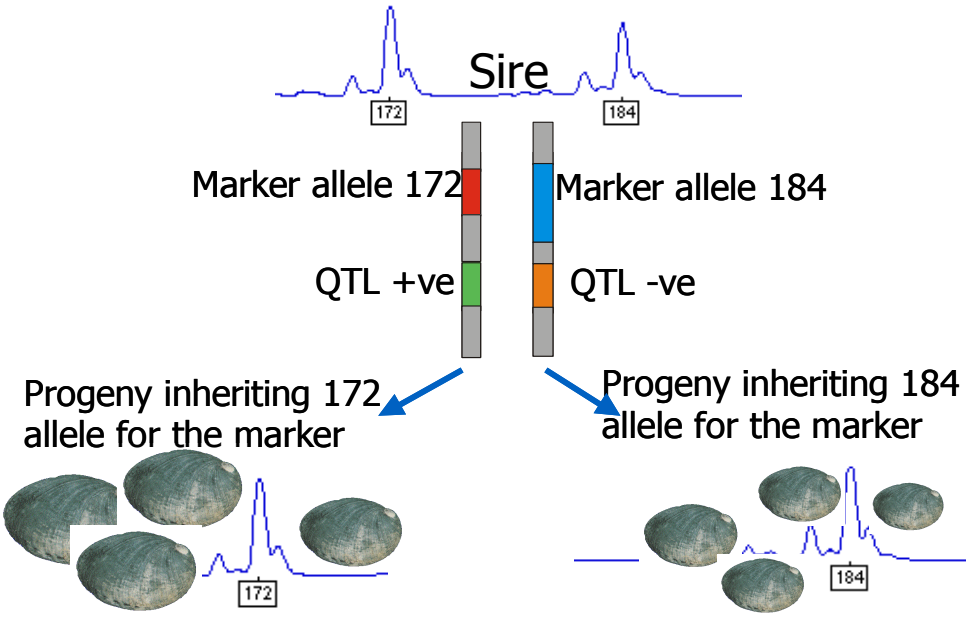


Figure 1.2. Principle of quantitative trait loci (QTL) detection, illustrated using an abalone example. A sire is heterozygous for a marker locus, and carries the alleles 172 and 184 at this locus. The sire has a large number of progeny. The progeny are separated into two groups, those that receive allele 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 has been performed in all nearly 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. To illustrate this, consider the formula that Darvasi and Soller (1997) gave for estimating the 95% CI for QTL location for simple QTL mapping designs under the assumption of a high density genetic map. The formula was $CI=3000/(kN\delta^2)$, where N is the number of individuals genotyped, δ allele substitution effect (the effect of getting an extra copy of the increasing QTL allele) in units of the residual standard deviation, k the number of informative parents per individual, which is equal 1 for half-sibs and backcross designs and 2 for F_2

progeny, and 3000 is about the size of the cattle genome in centi-Morgans. For example, given a QTL segregates on a particular chromosome within a half sib family of 1000 individuals, for a QTL with an allele substitution effect of 0.5 residual standard deviations the 95% CI would be 12 cM. Such large confidence intervals have two problems. Firstly if the aim of the QTL mapping experiment is to identify the mutation underlying the QTL effect, in a such a large interval there are a large number of genes to be investigated (80 on average with 20 000 genes and a genome of 3000cM). Secondly, use of the QTL in marker assisted selection is complicated by the fact that the linkage between the markers and QTL is not sufficiently close to ensure that marker-QTL allele relationships persist across the population, rather marker-QTL phase within each family must be established to implement marker assisted selection.

An alternative, if dense markers were available, would be to exploit linkage disequilibrium (LD) to map QTL. Performing experiments to map QTL in genome wide scans using LD has recently become possible due to the availability of 10s of thousands of single nucleotide polymorphism (SNP markers) in cattle, pigs, chickens and sheep in the near future (eg. (ftp://ftp.hgsc.bcm.tmc.edu/pub/data/Btaurus/snp/Btau20040927/bovine-snp.txt). A SNP marker is a difference in nucleotide between animals (or an animals pair of chromosomes), at a defined position in the genome, eg.

Animal 1. ACTCGGGC

Animal 2. ACTTGGGC

Rapid developments in SNP genotyping technology now allow genotyping of a SNP marker in an individual for as little as 1c US.

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 and 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, $\chi^{2'} = r^2$.

Their investigations using simulation showed out of a number of multi-allelic pair-wise measures of LD $\chi^{2'}$ was the best predictor of useable marker-QTL LD (eg. the proportion of QTL variance explained by the marker).

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 1.3 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.

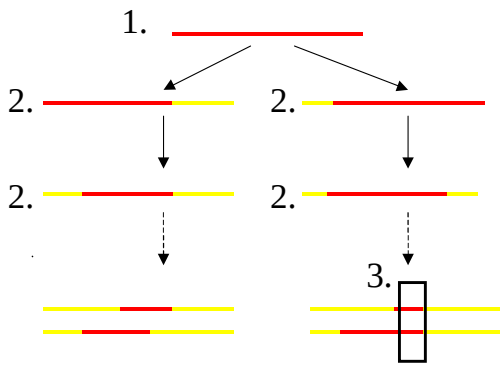


Figure 1.3 An ancestor many generations ago (1) leaves descendants (2). Each generation, the ancestors chromosome is broken down by recombination, until all that remains in the current generation are small conserved segments of the ancestor's chromosome (3). The chromosome segment homozygosity (CSH) 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 and 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 1.4. This is because by chance there may have been an ancestral recombination between one pair of markers, but not the other.

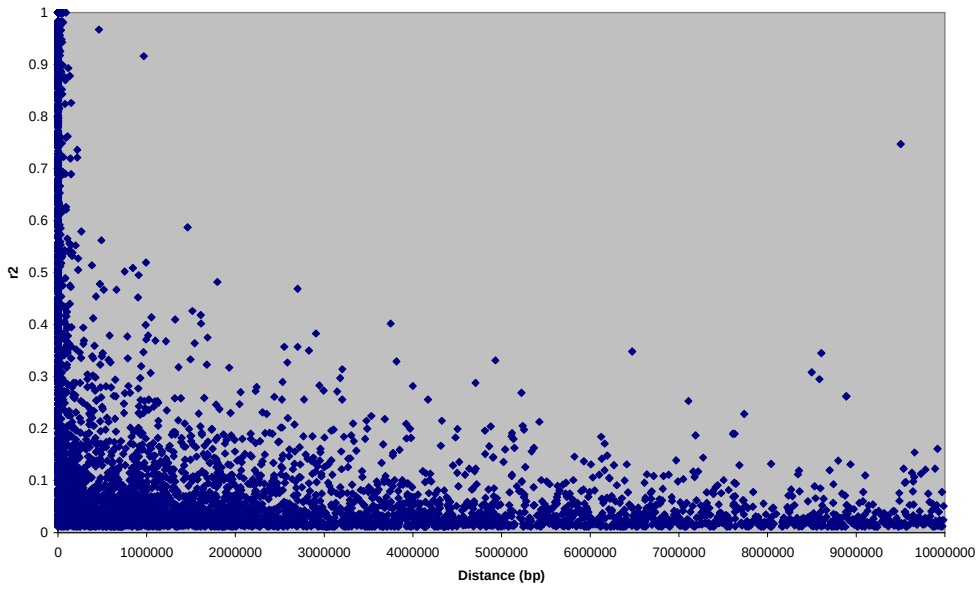


Figure 1.4. r^2 values against distance in bases between pairs of markers from 0 000 genome wide SNPs genotyped in a population of Holstein Friesian cattle. 1000000 bases is approximately 1cM.

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 recombinations. Hayes et al. (2003) investigated the variability of r^2 and CSH using simulation. They simulated a chromosome segment of 10 cM containing 11 markers was simulated with a mutation-drift model, with a constant *N* of 1000. They

found CSH was less variable than r^2 provided at least four loci were included in the calculation of CSH, Figure 1.5.

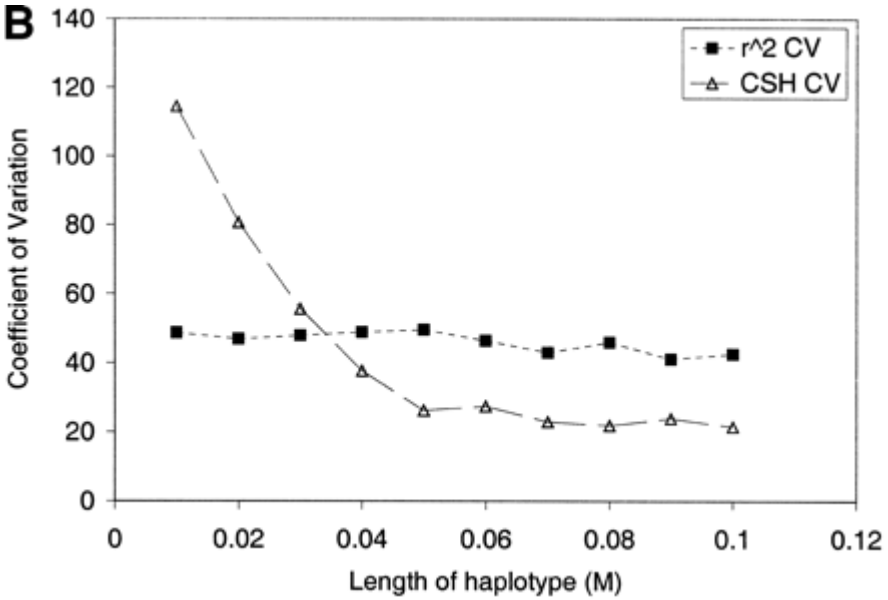


Figure 1.5 Coefficient of variation of r^2 and CSH in a simulated populations, over haplotype regions of the same length, across 200 replicates. There one marker per 0.01M (Hayes et al. 2003).

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 (eg. Lander and 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 (eg. 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 1.6.

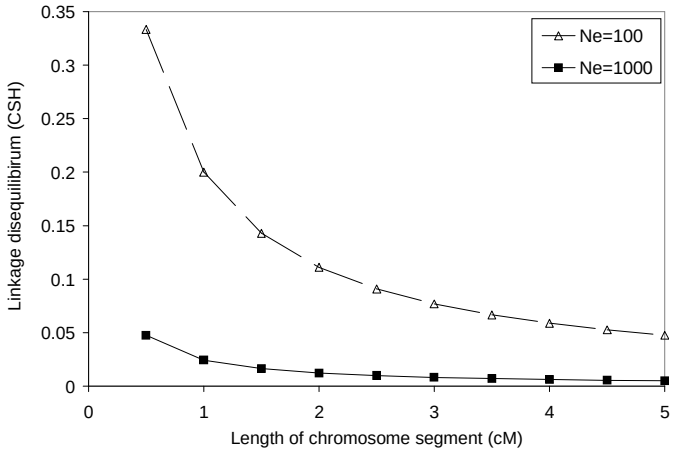
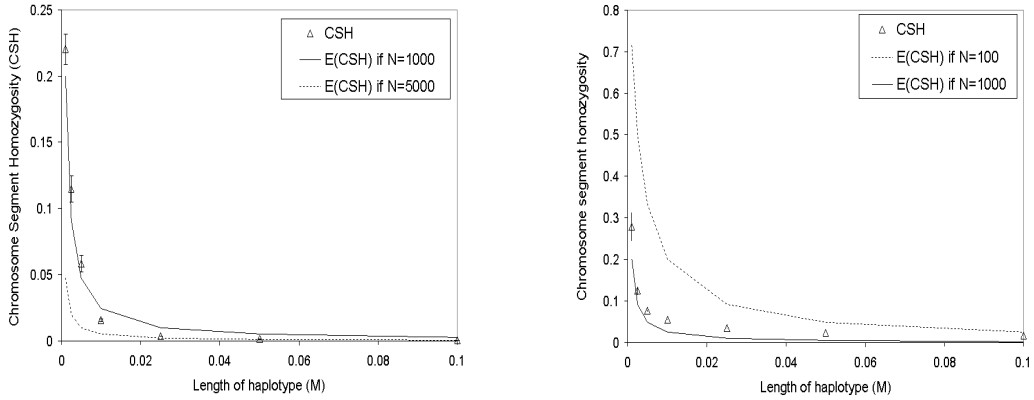


Figure 1.6. The extent of LD (as measured by chromosome segment homozygosity, CSH) for increasing chromosome segment length, for N_e=100 and N_e=1000. Note that r^2 has the same expectation as CSH.

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 1.7. Results for r^2 would look very similar.



A B

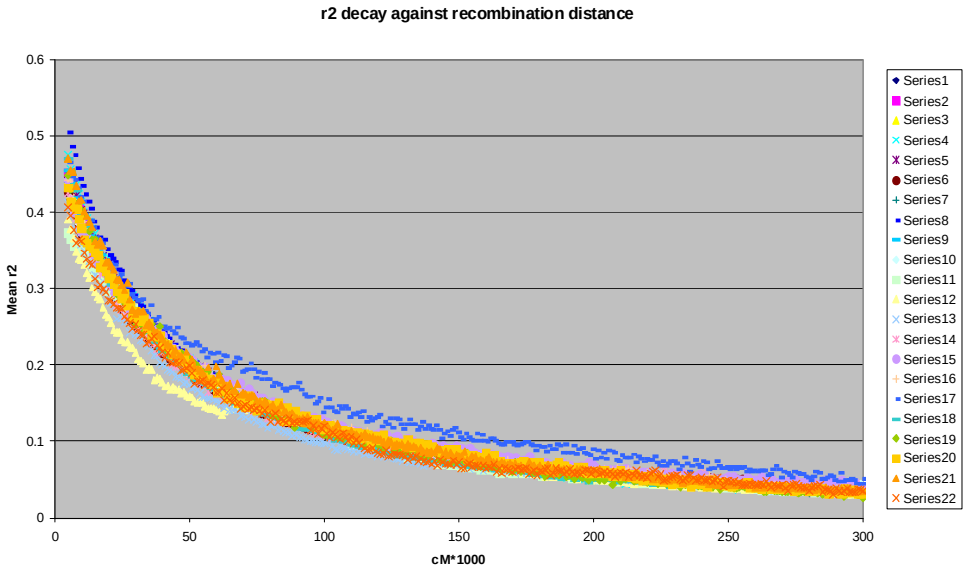
Figure 1.7. Chromosomal homozygosity for different lengths of chromosome (given the recombination rate) for populations: A. Linearly increasing population size, from N=1000 to N=5000 over 100 generations, following 6000 generations at N=1000. B. Linearly decreasing population size, from N=1000 to N=100 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 r2 with changing effective population size can be written as $E(r^2) = 1/(4N_t c + 1)$ where $t = 1/2c$.

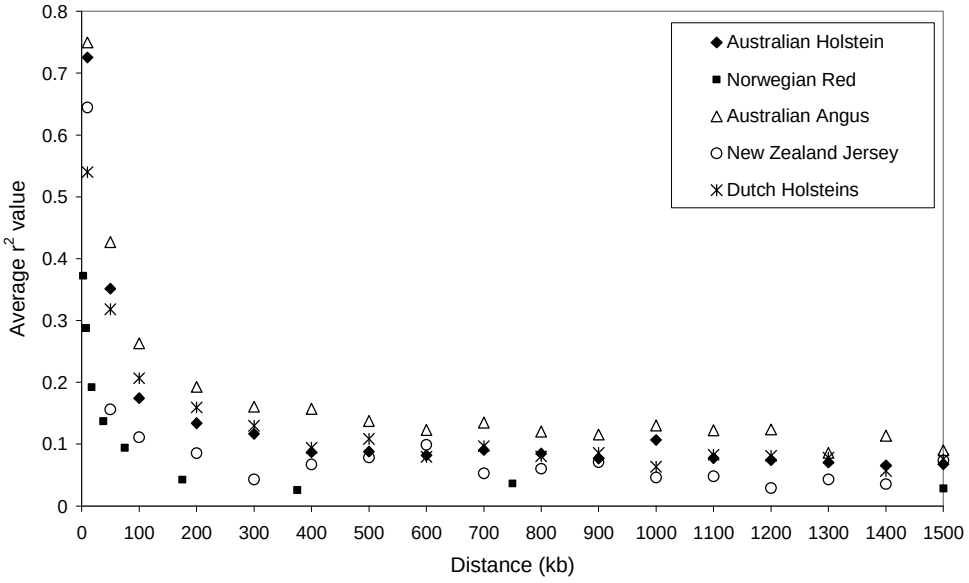
1.4 The extent of LD in livestock and human populations

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 1.8. In cattle moderate LD extends up to 100kb, Figure 1.8. However, very high levels of LD (eg. $r^2 \geq 0.8$ only extend very short distances in both humans and cattle.

It is interesting to compare the extent of LD in the different cattle populations. The Dutch and Australian Holstein populations had a very similar decline of LD, probably because these populations are highly related (eg. Zenger et al. 2007) and are similar in effective population size and history. The decline of LD in the Norwegian Reds was more rapid than in the Holstein populations. One explanation for this could be that the effective population size in Norwegian Red is higher than in Holstein, even though the global population is much smaller. Effective population size in Norwegian Reds is approximately 400 (Meuwissen et al. 2002), while for the global Holstein population effective population size is close to 150 (Zenger et al 2007), and a more limited extent of LD is expected with larger effective population size.



A



B

Figure 1.8. A. Average r^2 with distance in Caucasian humans (from Tenesa et al. 2007). 1cM is approximately 1000kb. B. Average r^2 value according to the distance between SNP markers in different cattle populations. Results are from 9918 SNPs distributed across the genome genotyped in 384 Holstein cattle or 384 Angus cattle, 403 SNPs genotyped in 783 Norwegian Red cattle, 3072 SNPs genotyped in 2430 Dutch Holstein cattle, or 351 SNPs genotyped in Jersey cattle. Norwegian red data kindly supplied by Prof. Sigbjorn Lien, Norwegian University of Life Sciences, New Zealand Jersey data kindly supplied by Dr. Richard Spelman, Livestock Improvement Co-operative.

Figure 1.8 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 Jerseys and Norwegian Reds, a larger number of markers would be required.

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.

McRae et al. (2002) evaluated the extent of LD in domestic sheep. They used the D' parameter rather than r^2, so comparison with results for other species given here is difficult. They found that high levels of LD extended for tens of centimorgans and declined with increasing marker distance. They also thoroughly investigated bias in D' under different conditions, and found that *D*' may be skewed when rare alleles are

present. They therefore recommended that the statistical significance of LD is used in conjunction with coefficients such as *D*' to determine the true extent of 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 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. 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. 2007). 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 LD-MAS or Genomic selection using the same marker set and SNP effects across the breeds or populations.

In Figure 1.9, 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 (eg. Zenger et al. 2007).

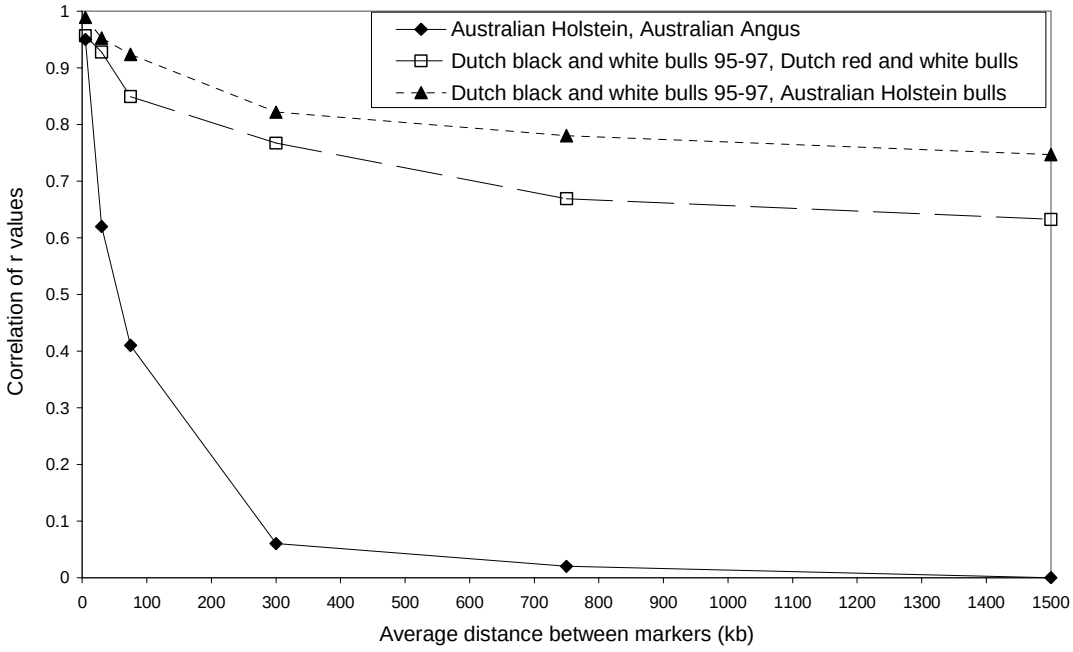


Figure 1.9. Correlation between r values for various cattle populations or sub-populations, as a function of marker distance (from De Roos et al. 2007).

1.6 Optional topic 1. Brief note on haplotyping strategies

Calculations of LD parameters like r^2 and CSH assume that the genotypes of individuals can be phased into haplotypes (ie. which marker alleles belong on the paternally inherited chromosome and which marker alleles belong on the maternally inherited chromosome). If large half sib families are available, the sires haplotypes can fairly readily be reconstructed by determining which alleles are most often co-inherited from the sire. The haplotypes which the dam passed on the to the progeny can then be inferred by 'subtracting' the alleles transmitted from the sire from the progeny genotypes. Inferring haplotypes becomes more difficult in complex

pedigrees, with missing marker information, or when there is very little pedigree information at all.

One method of inferring haplotypes in complex pedigrees is to run a *Markov Chain* on a set of *genetic descent graphs*. A genetic descent graph specifies the paths of gene flow (parents to offspring), but not the particular founder alleles travelling down the paths. See Sobel and Lange (1996) for more details on this procedure. This method is implemented in a freeware program called SimWalk (http://www.genetics.ucla.edu/software/simwalk_doc/).

In some cases, the individuals that are genotyped may be randomly sampled from the population, with no pedigree information available. Provided the markers which have been genotyped are closely spaced, it can be possible to estimate haplotypes based on linkage disequilibrium and allele frequency information alone. One such method was proposed by Stephens et al. (2001). Suppose we have a sample of *n* diploid individuals from a population (these individuals are assumed to be unrelated). Let G = (G_1,….G_n) denote the (known) genotyped for the individuals, let H = (H_1,…,H_n) denote the (unknown) corresponding haplotype pairs, let F = (F_1,….,F_M) denote the set of unknown population haplotype frequencies, and let f = (f_1,….,f_M) denote the set of unknown sample haplotype frequencies (the M possible haplotypes are labelled 1,…,M). The haplotype reconstruction method of Stephens et al. (2001) regards the unknown haplotypes as unobserved random quantities and aims to evaluate their conditional distribution in light of the genotype data. To do this, they used MCMC, to obtain an approximate sample from the posterior distribution of H given G, eg. Pr(H|G). The steps in the algorithm are:

1. Start with an initial guess for H (the haplotype pairs of all individuals), H^0. This begins by listing all haplotypes that must be present unambiguously in the sample, that is those individuals who are homozygous at every locus or are heterozygous at only one locus. For the other individuals, who have ambiguous haplotypes, the haplotypes can be allocated at random from the genotypes.
2. Choose an individual, i, at random from all the ambiguous genotypes. Sample the haplotypes for this individual for the next iteration (H_i^{t+1}). These haplotypes are

sampled from a distribution which assumes that the haplotypes in the haplotype pair H_i are likely to look either *exactly the same* or *similar to* a haplotype that has already been observed. This assumption is based on the existence of both LD and mutation – if the chromosome segment carrying the haplotypes is short enough, there will be considerable LD, greatly restricting the number of haplotypes. New haplotypes can be generated either by recombination or mutation at one of the markers. Formally, the distribution from which the new haplotypes are sampled is:

$$\pi(h \mid H) = \sum_{\alpha=1}^{M} \sum_{s=0}^{\infty} \frac{r\infty}{r} \left(\frac{\theta}{r+\theta} \right)^{s} \frac{r}{r+\theta} P_{\alpha h}^{s}$$

where r_α is the number of haplotypes of type α in the set H, r is the total number of haplotypes in H, θ is a scaled mutation rate (based on assumptions about population size, mutation rates at individual loci and length of the haplotype, relating to the expectation of LD described above), and P is mutation matrix (mapping the mutations onto markers in the haplotype). This corresponds to the next sampled haplotype, *h*, being obtained by applying a random number of mutations, s, to a randomly chosen existing haplotype, α, where *s* is sampled from a geometric distribution.

The above algorithm is implemented in a program (again free) called PHASE. At least for short haplotypes (< 1cM) it appears to construct haplotypes very accurately. A nice feature of the algorithm is that an approximate probability of each haplotype for each animal being correct can be obtained from the posterior distribution. These probabilities could potentially be used in the QTL mapping procedure. The PHASE program is now widely used in human genetics, and is likely to be used to construct the bovine haplotype map as part of the bovine genome sequencing activity.

2. Genome wide association studies

2.1 Introduction

Linkage disequilibrium (LD) mapping of QTL exploits population level associations between markers and 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, and if there is a QTL somewhere within the chromosome segment, they will also carry identical QTL alleles. There are a number of QTL mapping strategies which exploit LD, the simplest of these is the genome wide association test using single marker regression.

2.2 Genome wide association tests 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

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

Where **y** is a vector of phenotypes, $\mathbf{1_n}$ is a vector of 1s, **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. Note that the g can actually be a vector of 2 times the number of marker alleles, if both additive and dominance effects are to be estimated. The underlying assumption here is that the marker will only affect the trait if it is in linkage disequilibrium with an unobserved QTL. This model ignores fixed effects other than the mean, however they can be easily included.

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 for example by an analysis of variance (ANOVA), $F_{\alpha,v1,v2}$ 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	Phenotpe	SNP allele 1	SNP allele 2
1	2.030502	1	1
2	3.542274	1	2
3	3.834241	1	2
4	4.871137	2	2
5	3.407128	1	2
6	2.335734	1	1
7	2.646192	1	1
8	3.762855	1	2
9	3.689349	1	2
10	3.685757	1	2

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 is 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**

The mean and SNP effect can then be estimated as:

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

Where **y** is the (number of animals x 1) vector of phenotypes.
In the above example the estimated 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:

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

Using the above values, the value of F is 4.56. This can be compared to the tabulated F-value at a 5% significance value and 1 and 9 (number of records -1) degrees of freedom is 5.12. So the SNP effect in this case is not significant (not surprisingly with only 10 records!).

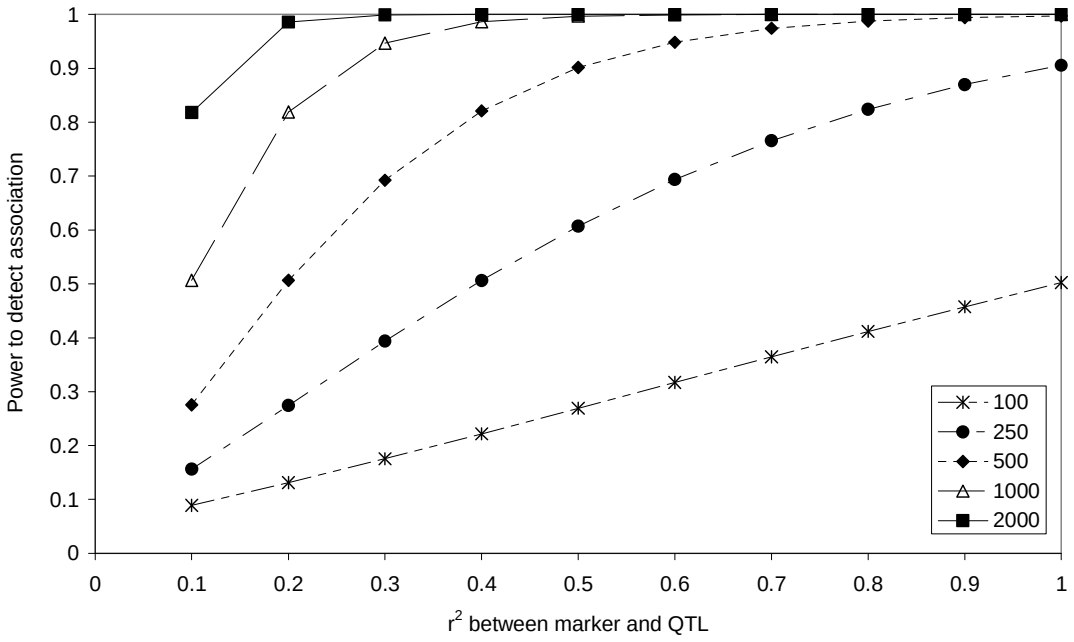
The power of the association test to detect a QTL by testing the marker effect depends on:

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 and 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 particular sensitive to *p* when *p* is small (eg. <0.1).
5. The significance level α set by the experimenter.

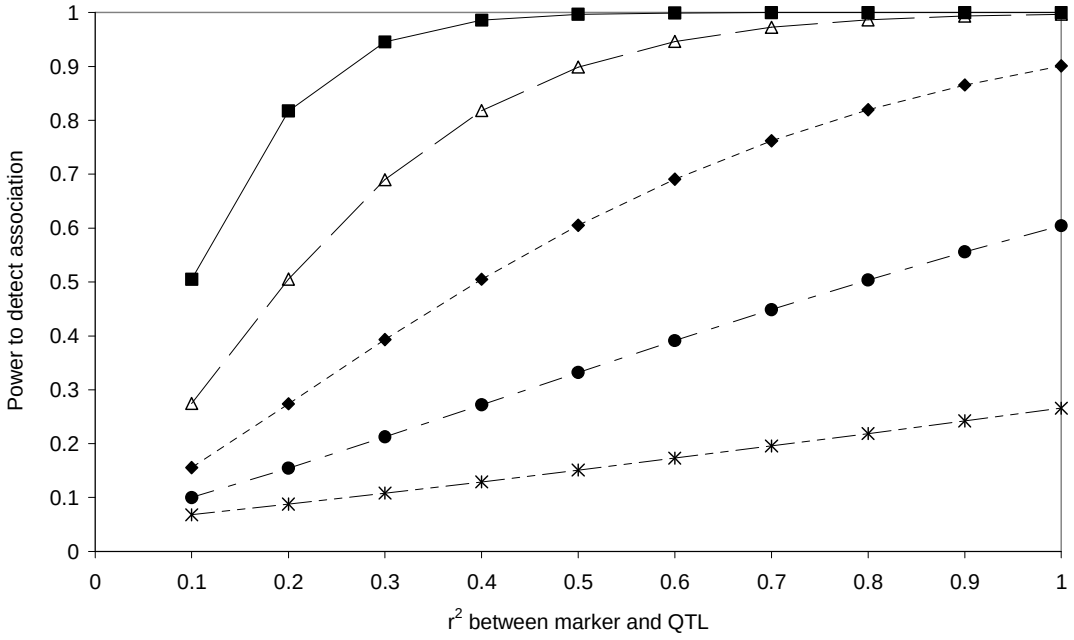
The power is the probability that the experiment will correctly reject the null hypothesis when a QTL of a given size of effect really does exist in the population. Figure 2.1 illustrates the power of an association test to detect a QTL with different levels of r^2 between the QTL and the marker and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

Using both this figure, and the extent of LD in our livestock species, 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. 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, at least 30 000 markers would be required.

To demonstrate the dependence of power on r^2 between a QTL and SNP in another way, consider the results of Macleod et al. (2007). They attempted to assess the power of whole genome association scans in outbred livestock with commercially available SNP panels. In their study, 365 cattle were genotyped using a 10,000 SNP panel while QTL, polygenic and environmental effects were simulated for each animal, with QTL simulated on genotyped SNPs chosen at random. The power to detect a QTL accounting for 5% of the phenotypic variance with 365 animals genotyped, was 37% ($p<0.001$). There was a strong correlation between the F-value of significant SNPs and their r^2 with the "QTL", Figure 2.2. The correlation of F-values with D' was almost zero.



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.

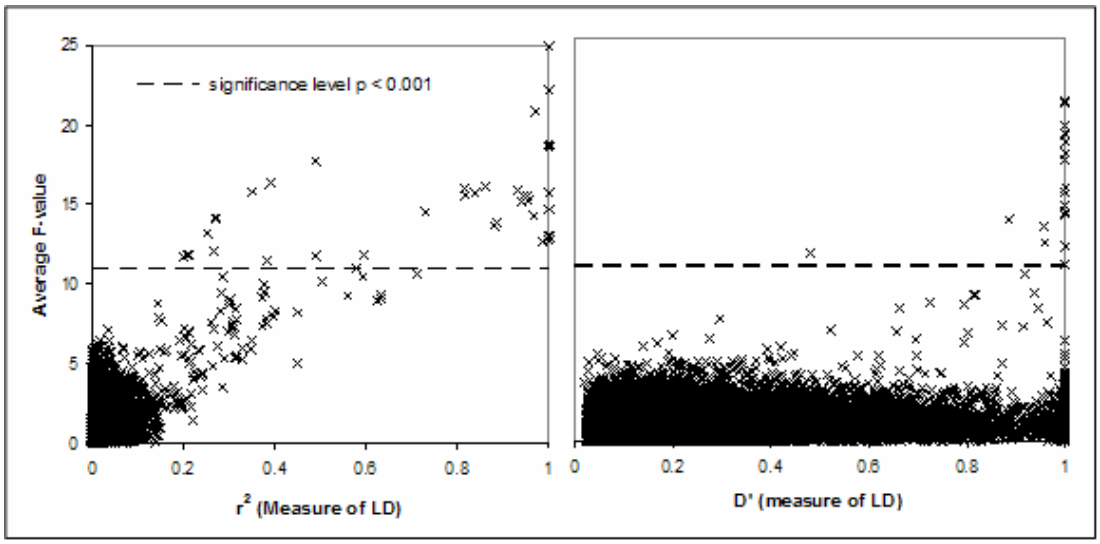


Figure 2.2 Plots of F-values of SNPs tested for association, against r^2 and D' of the tested SNP with the QTL. The QTL accounted for 5% of the phenotypic variance. From Macleod et al. 2007.

2.2.1 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 10s or possibly 100s of thousands of markers. 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 test don't correspond to the actual significance levels in the whole experiment, eg. 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 10 000 markers (assuming for the moment these points are independent), we would expect 10000*0.05 = 500 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

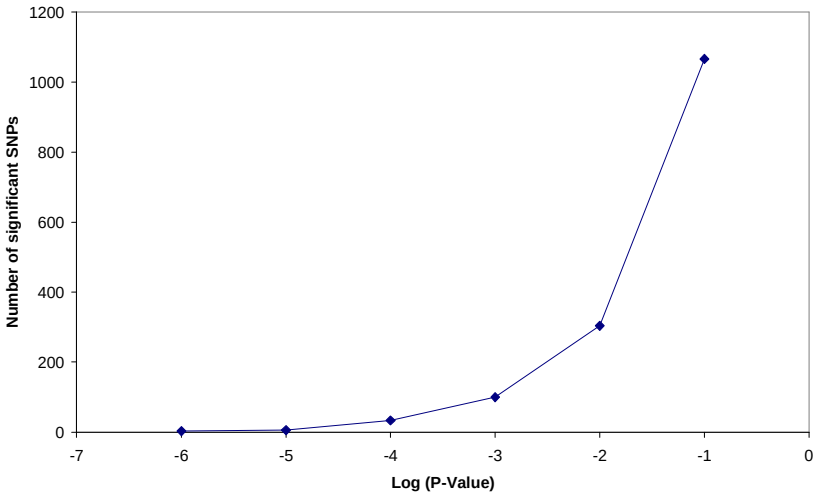
correction tends to be very conservative, and requires some decision to be made about how many independent regions of the genome were tested.

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 (eg 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 5000^{th} 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 structure however, randomly shuffling phenotypes across marker genotypes will not preserve any pedigree structure that exists in the data.

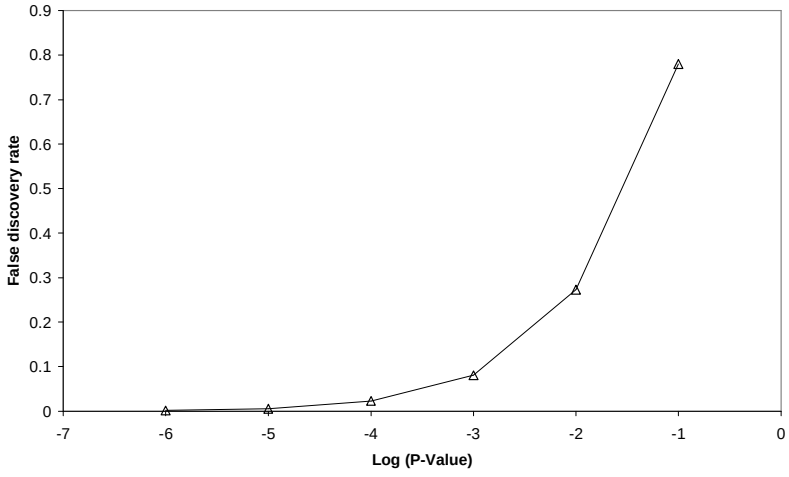
An alternative to attempting to avoid false positives is to monitor the number of false positives relative to the number of positive results (Fernando et al. 2004). The researcher can then set a significance level with an acceptable proportion of false positives. The false discovery rate (FDR) is the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

$$mP_{max}/n,$$

where P_{max} is the largest P value of QTL which exceed the significance threshold, n is the number of QTL which exceed the significance threshold and m is the number of markers tested. Figure 2.3 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 2.3 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).

2.2.2 Confidence intervals.

There are few reports in the literature on methods to estimate confidence intervals in genome wide association studies. A method based on cross-validation is 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 (eg. 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 (ie a significant SNP in almost the same location), the information is used 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.

2.2.3 Avoiding spurious false positives due to population structure

The very simple model above for testing association of a marker to phenotype assumes there is no structure in the population, that is it assumes all animals are equally related. In livestock populations, or any population for that matter, this is unlikely to be the case. Multiple offspring per sire, selection for specific breeding goals and breeds or strains within the population all create population structure. Failure to account for population structure can cause spurious associations (false positives) in the genome wide association study (Pritchard 2000). A simple example is where the population includes a sire with a large number of progeny in the population. In this case the sire has a significantly higher estimated breeding value than other sires in the population. If a rare allele at a marker any where on the genome is homozygous in the sire, the sub-population made up of his progeny will have a higher frequency of the allele than the rest of the population. As the sires' estimated breeding value is high, his progeny will also have higher than average estimated breeding values. Then in the genome wide association study, if the number of progeny of the sire 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 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 (2007) described in section 2.2.1, 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 all QTL detected are false positives), Table 2.1. 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.

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 average relationship matrix (Hayes et al. 2007) or population structure (eg. Pritchard 2000).

For a given marker single locus, a similarity index S_{xy} between two individuals x and y is calculated, where $S_{xy} = 1$ when genotype $x = ii$ (i.e. both alleles at loci l are identical) and genotype $y = ii$, or when $x = ij$ and $y = ij$. $S_{xy} = 0.5$ when $x = ii$ and $y = ij$, or vice versa, $S_{xy} = 0.25$ when $x = ij$ and $y = ik$, and $S_{xy} = 0$ when the two individuals have no alleles in common at the locus. The similarity as a result of chance alone was $s = \sum_{i=1}^{a} p_i^2$ where p_i is the frequency of allele i in the (random mating) population, and a is the number of alleles at the locus. Then the relationship between individuals x and y at locus l is calculated as $r_l = (S_{xy} - s)/(1 - s)$. The average relationship between the individuals is calculated as the r_l averaged over all loci.

With large numbers of markers, average relationship matrices derived from markers can be very accurate, and can even capture mendelian sampling effects (eg. Two full sibs may be more or less related than 0.5 because they have more or less paternal and maternal chromosome segments than expected by chance. This approach can also be used to correct for population structure across breeds or lines. In Figure 2.4, the average relationship matrix derived from markers is shown for a combined Angus Holstein Population.

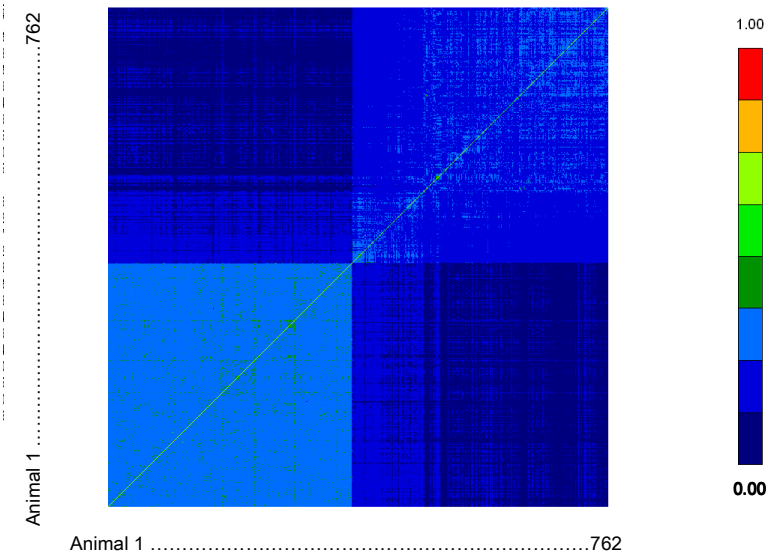


Figure 2.4. Average-relationship matrix derived from 9323 SNP loci where the population consists of two breeds. The diagonal element for the first Angus animal is in the bottom left hand corner and the element for the last Holstein animal in the top left hand corner.

There are a number of situations in which marker derived relationship matrices will be especially valuable. When there is limited or no pedigree recorded in a population, marker genotypes may be the only source of information available to build relationship matrices. For example, in livestock, there are many traits which can only be recorded in animals which are not candidates for selection, such as meat quality. If there is no recorded pedigree linking selection candidates and commercial animals on which the trait is recorded, marker derived relationship matrices could be used in estimation of QTL effects for marker assisted selection. Another example is populations where multiple sires are used in the same paddock of dams, such that recording pedigree is difficult. Finally, in multi-breed populations including crosses between breeds, the marker derived relationship matrix offers a way to account for the different breed composition of the animals.

2.3 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.

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

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

For example, in a simulated population of Ne=100, and a chromosome segment of length 10cM, the proportion of the QTL variance accounted for by marker haplotypes when there were 11 markers in the haplotype was close to one, Figure 2.5. [Note that if the effective population size was larger, the proportion of genetic variance explained by a 10cM haplotype would be reduced (Goddard 1991).]

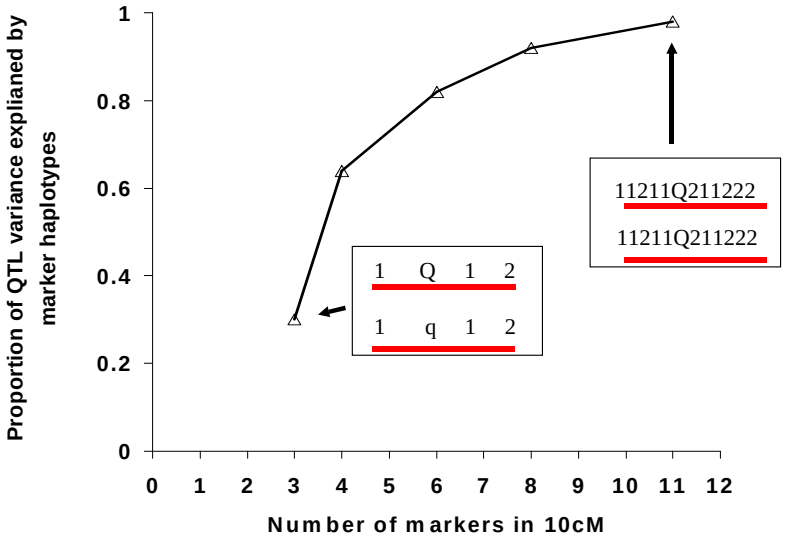


Figure 2.5. Proportion of QTL variance explained by marker haplotypes with an increasing number of markers in a 10 cM interval

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 2002), can be used to do this.

3. Genomic selection

3.1 Introduction to genomic selection

One problem with LE-MAS, LD-MAS or Gene-MAS is that only a limited proportion of the total genetic variance is 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 has become possible very recently with the availability of 10s of 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 straightfoward: 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 chromosome segments across the genome, X_i is a design matrix allocating animals to the haplotype effects at segment i, and $\hat{\mathbf{g}}_\mathbf{i}$ is the vector of effects of the haplotypes within chromosome segment i.

The difficulty in step 1. is that a very large number of haplotype effects across the chromosome segments must be estimated (the $\hat{\mathbf{g}}_\mathbf{i}$), most likely from a data set where the number of phenotypic observations is less than the number of chromosome segment effects to be estimated.

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 3.2.2 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 section 4.2 can be applied to either single markers or haplotypes. The only 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). We will describe the IBD approach separately.

It is important to note that the following genomic selection procedures can be used to map QTL as well as predict GEBV. Procedures such as the LDLA approach as described yesterday assume one QTL per chromosome. Given the distribution of QTL effects, there are likely to be 100 or more QTL throughout the genome affecting a particular quantitative trait (eg. Hayes et al. 2006). Therefore most chromosomes will carry at least two QTL affecting the trait, though one of these may have a very small effect. Both estimates of effects and position of a QTL can be biased by other QTL on the same chromosome, especially if the QTL are closely linked. The worst case scenario is that two linked QTL cancel each others effects, so none of the QTL are detected. Alternatively, a 'ghost' QTL, with a very large confidence interval, can be positioned between two real QTL (Martinez and Curnow 1992). Because genomic selection approaches can fit all QTL simultaneously, they can remove the effect of the QTL in brackets adjacent to the true QTL position, giving tighter confidence intervals.

3.2 Methodologies 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. A key difference between these approaches is the assumption they make about the variances of haplotype or single marker effects across chromosome segments.

The simpler assumption is that the variance of haplotype effects is equal across all chromosome segments. This is analogous to estimating breeding values where we assume that the breeding values are distributed $V(\mathbf{u}) \sim N(0, A\sigma_a^2)$. In the case of the chromosome segment effects, they would be distributed $V(\mathbf{g}) \sim N(0, I\sigma_g^2)$ where σ_g^2 is the variance of the effects across all segments.

However this assumption does not capture our "prior" knowledge that some chromosome segments will contain QTL with large effects, some chromosome segments will contain QTL with small effects, and some chromosome segments will contain no QTL. We can capture this prior knowledge by modelling the data at two levels. The first level is at the level of the data including the overall mean, the error variance and the chromosome segment effects. In this model, each chromosome segment has it's own variance of haplotype or marker effects $V(\mathbf{g}_i) \sim N(0, I\sigma_{gi}^2)$. The second model is at the level of the variance of chromosome segment effects, to allow these to be different for each approach.

We shall consider genomic selection approaches with the simpler assumption of equal variances of effects across chromosome segments first.

3.2.1.1 Least squares

The first approach actually makes no assumptions regarding the distribution of chromosome segment effects, because it treats these effects as fixed in a least squares approach. The approach is identical to that described for LD-MAS. 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 $\boldsymbol{y}$ is the data vector; μ is the overall mean; $\mathbf{1}_n$ is a vector of n (n=number of records) ones; $\boldsymbol{g_i}$ represents the genetic effects of the 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) ; $\boldsymbol{X_i}$ is the design matrix for the i^{th} segment; and $\boldsymbol{e}$ is the error deviation. 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 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 can 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 chromosome segment effects to include in the estimation of breeding values based on the effect of the chromosome segment estimated from single segment regression. As a result, the problem of over-estimation of segment effects due to multiple testing will be incurred.

3.2.2 Ridge regression and BLUP

To overcome the problem of over-estimation of segment effects in the context of marker assisted selection, Whittaker et al. (2000) applied ridge regression. In ridge regression, estimates of the $\mathbf{g_i}$ are shrunk towards the mean, in an attempt to avoid the over-estimation of these effects. This shrinkage can also allow all effects to be estimated simultaneously. In ridge regression, all $\mathbf{g_i}$ have a common variance. Ridge regression can be applied to genomic selection:

$$\hat{\mathbf{g}} = (\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}'\mathbf{y}$$

where $\mathbf{X}$ is a matrix allocating all marker genotypes or haplotypes to phenotypes, and y is a vector of phenotypes. The difficulty with ridge regression is that the choice of λ is arbitrary. Further, if a very small value λ is chosen, there may not be a unique

solution for the model with the large number of $\mathbf{g_i}$ fitted. Methods for selecting values of λ are given in Xu (2003) and Whittaker et al. (2000). Xu (2003) concluded that ridge regression was not a viable choice for QTL mapping if the model includes markers across the entire genome. There reason was that ridge regression treats all effects equally across all loci, whereas in fact many markers have negligible effects. However ridge regression may still perform reasonably well in the context of estimating genomic breeding values, as the effects are accumulated across many segments.

If $\lambda = \sigma_e^2 / \sigma_g^2$ in the equation for ridge regression, this is in fact BLUP as used by Meuwissen et al. (2001). The BLUP assumes the variance of haplotype effects at each chromosome segment is the same.

An important question is what value of σ^2_g should be used in the BLUP (eg. the variance of haplotype effects at a chromosome segment). Meuwissen et al. (2001) dealt with this problem by calculating the genetic variance expected from a genetic drift-mutation model, and assuming the distribution of QTL effects was as given by Hayes and Goddard (2001). See their paper in the appendix for details.

Another way of estimating σ^2_g would be to first estimate the total additive genetic variance (using REML for example) then divide by the number of chromosome segments.

An example of genomic selection using BLUP follows. Consider the following data set for animals with a single chromosome, with 4 markers defining three chromosome segments. The markers are SNPs, so there are 4 possible haplotypes per segment. Phenotypes were "simulated" with an overall mean of 2, an effect of haplotype 1 in the first segment of 1, an effect of haplotype 1 in the second segment of -0.5, and a normally distributed error term with mean 0 and variance 1. The data is as follows:

	Haplotype Segment 1		Haplotype Segment 2		Haplotype Segment 3		Phenotype
Animal	P	M	P	M	P	M	
1	1	1	2	2	1	1	3.41
2	1	2	1	2	1	1	2.47
3	2	2	1	2	1	2	2.32
4	1	3	2	3	2	1	2.32
5	1	4	1	3	2	1	1.75

Note that there are 9 haplotypes observed in total (4 for the first segment, 3 for the second segment, and 2 for the third segment), while there are only 5 phenotypic records.

The design matrix (**X**) for this data set is (in bold):

	Segment 1 haplotypes				*Segment 2 haplotypes*			*Segment 3 haplotypes*	
Animal	*1*	*2*	*3*	*4*	*1*	*2*	*3*	*1*	*2*
1	**2**	**0**	**0**	**0**	**0**	**2**	**0**	**2**	**0**
2	**1**	**1**	**0**	**0**	**1**	**1**	**0**	**2**	**0**
3	**0**	**2**	**0**	**0**	**1**	**1**	**0**	**1**	**1**
4	**1**	**0**	**1**	**0**	**0**	**1**	**1**	**1**	**1**
5	**1**	**0**	**0**	**1**	**1**	**0**	**1**	**1**	**1**

The vector $\mathbf{1_n}$' is [1 1 1 1 1]'

The mixed model equations are

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

Where $\lambda = \frac{\sigma_e^2}{\sigma_g^2}$ and ***I*** is an Identity matrix (total number of haplotypes x total number of haplotypes).

Assuming a value of 1 for λ, the mixed model equations with our data are:

5	5	3	1	1	3	5	2	7	3
5	8	1	1	1	2	6	2	8	2
3	1	6	0	0	3	3	0	4	2
1	1	0	2	0	0	1	1	1	1
1	1	0	0	2	1	0	1	1	1
3	2	3	0	1	4	2	1	4	2
5	6	3	1	0	2	8	1	8	2
2	2	0	1	1	1	1	3	2	2
7	8	4	1	1	4	8	2	12	3
3	2	2	1	1	2	2	2	3	4

$$\begin{bmatrix} \hat{\mu} \\ \hat{g} \end{bmatrix}$$

12.27
13.36
7.11
2.32
1.75
6.54
13.93
4.07
18.15
6.39

Giving the following estimates of the mean and Haplotype effects:

Effect		Estimate
Mean		2.11
Segment 1	Haplotype 1	0.15
	Haplotype 2	0
	Haplotype 3	-0.05
	Haplotype 4	-0.1
Segment 2	Haplotype 1	-0.16
	Haplotype 2	0.31
	Haplotype 3	-0.15
Segment 3	Haplotype 1	0.09
	Haplotype 2	-0.09

With so few records, the accuracy of estimating the haplotype effects is low.

Now if we genotype a group of young animals we can estimate their GEBV from the haplotypes they carry:

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

Consider the following animals:

	Haplotype segment 1		Haplotype segment 2		Haplotype segment 3	
Animal	Paternal	Maternal	Paternal	Maternal	Paternal	Maternal
6	1	2	1	2	1	1
7	1	1	2	2	1	2
8	2	3	2	2	1	2
9	1	4	3	1	1	2
10	2	4	2	2	1	2

The **X** matrix for the new animals is:

	Segment 1 haplotypes				*Segment 2 haplotypes*			*Segment 3 haplotypes*	
Animal	*1*	*2*	*3*	*4*	*1*	*2*	*3*	*1*	*2*
6	1	1	0	0	1	1	0	2	0
7	2	0	0	0	0	2	0	1	1
8	0	1	1	0	0	2	0	1	1
9	1	0	0	1	1	0	1	1	1
10	0	1	0	1	0	2	0	1	1

Using the values of $\hat{\mu}$ and $\hat{\mathbf{g}}$ from above gives the following vector of GEBV

Animal	**GEBV**	**TBV**
6	0.48	0.5
7	0.91	2
8	0.57	0
9	-0.26	0.5
10	0.52	0

As the data was simulated, we also have a true breeding value (TBV) for these animals (the sum of the true haplotype effects described above). We can correlate the GEBV and TBV to get the accuracy of genomic selection in this case, which is 0.43 in this case.

With BLUP the chromosome segment (or QTL) with the largest variance will tend to have it's variance over-estimated, and this will still decrease the accuracy of genomic selection somewhat although much less than when the **g** are treated as a fixed effect. Better estimates of breeding value can be obtained by methods that allow the variance of the chromosome segment effects to vary between chromosome segments.

3.2.4 Bayesian methods

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.

3.2.4.1 Optional topic: 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.

3.2.4.2 Bayesian method with a prior that assumes many QTL have a small effect and few have a large effect

If we allow the variance of the effects across chromosome segments to vary, then the variances $V(\mathbf{g_i}) = \sigma^2_{gi}$ must be estimated. Meuwissen et al. (2001) described a Bayesian method they termed Bayes Method A to estimate chromosome segments effects and their variances simultaneously.

The method modelled the data at two levels. The first is at the level of the data as above:

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

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 prior distribution of the mean μ was uniform and uninformative, while the prior distribution of haplotype effects within chromosome segment *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.

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 consistent with many QTL of small effect and few of large effect. The prior distribution was 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 (see Appendix for details).

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. 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 $S+\boldsymbol{g_i}'\boldsymbol{g_i}$ 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*.

We cannot use this posterior distribution directly for estimating the σ^2_{gi} because it is conditional on the unknown $\boldsymbol{g_i}$ effects. Meuwissen et al. (2001) therefore used Gibbs sampling to estimate effects and variances.

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 $\frac{\mathbf{X}_{ij}'\mathbf{y} - \mathbf{X}_{ij}'\mathbf{X}\mathbf{g}_{(ij=0)} - \mathbf{X}_{ij}'\mathbf{1}_n\mu}{\mathbf{X}_{ij}'\mathbf{X}_{ij} + \sigma^2_e/\sigma^2_i}$, 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 $\sigma^2_e/(\mathbf{X}_{ij}'\mathbf{X}_{ij} + \sigma^2_e/\sigma^2_{gi})$.

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 $1/\chi^2_0$ (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 ith SNP genotype for individual *j* is a_1a_1, X_{ij}=0 if the ith SNP genotype for individual *j* is a_1a_2, and X_{ij}=-1 if the ith 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. Xu (2003) describe applications for plant populations for QTL mapping such as backcross, double haploid, or F2.

Meuwissen et al. (2001) pointed out that in reality, the distribution of genetic variances across chromosome segments is that there are many chromosome segments which contain no QTL, and relatively few chromosome segments which do contain QTL. However, the prior density of method BayesA does not actually 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. 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 Appendix for derivation of *v* and *S* values).

Figure 4.1 Illustrates the difference between the prior distribution of variances of chromosome segment effects used in method Bayes B and that used in method BayesA.

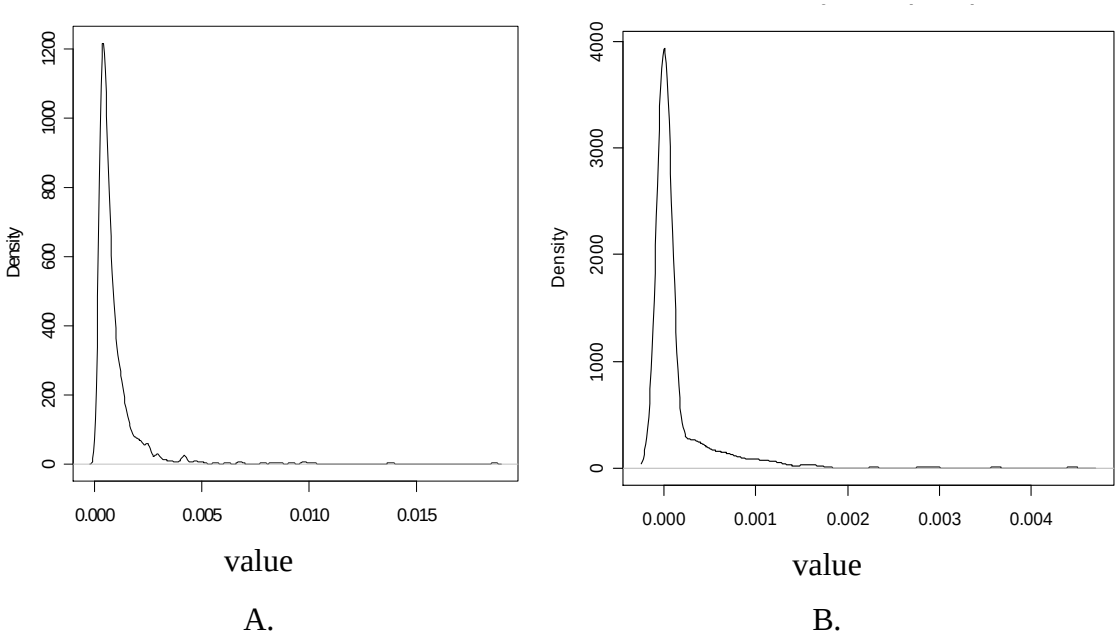


Figure 4.1 A. Prior distribution of variances of chromosome segment effects used in method BayesA, and B. Prior distribution of variances of chromosome

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 Appendix for details).

3.2.5 Evaluation of accuracy of genomic selection methods

To evaluate their methods (least squares, BLUP, Bayes A and Bayes B), 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 4.1.

Table 4.1. 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.

3.3 Factors affecting the accuracy of genomic selection

While the simulations demonstrate genomic selection has huge potential to increase rates of genetic gain, several key questions remain regarding it's 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
3) How do non-additive effects affect the accuracy of genomic selection.

3.3.1 Extent of linkage disequilibrium and number of markers required

The arguments here are similar to those given in chapter 3 for the number of markers required for LD-MAS. For genomic selection to work, the haplotypes or single markers must be in sufficient LD with the QTL such that the haplotype or single markers will predict the effects of the QTL across the population. For genomic selection to be as successful as in the simulations of Meuwissen et al. (2001), the level of LD between adjacent markers should be r^2>=0.2, as this was the level of LD there simulations generated. Solberg et al. (2006) used simulation of a population with N_e 100 to assess the effect of marker spacing on the accuracy of genomic selection (with BayesMethodB). They found a drop in accuracy of 20% as marker spacing was increased from one marker every 0.5cM to one marker every 4cM. Calus et al. (2007)

used simulation to assess the effect of the average r^2 between adjacent marker pairs on the accuracy of genomic selection (where the accuracy was the correlation of true breeding values and GEBV for a group of un-phenotyped animals). They found that accuracy increased dramatically as the average r^2 between adjacent markers increased, from 0.68 when the average r^2 between adjacent markers was 0.1, to 0.82 when the average r^2 between adjacent markers was 0.2, Figure 4.2.

In dairy cattle populations, an average r^2 of 0.2 between adjacent markers is only achieved when markers are spaced every 100kb. As the bovine genome is approximately 3 000 000kb, this implies that in order of 30 000 markers are required for genomic selection to be successful!

3.3.2 Haplotypes or single markers

Closely related to the effect of the extent of linkage disequilibrium on the accuracy of genomic selection is the effect of using single markers rather than haplotypes. The advantage of haplotypes over single markers in genomic selection is dependent on how accurately the haplotypes identify identical by descent chromosome segments compared to the accuracy with which single markers identify identical by descent chromosome segments. This can be quantified as the proportion of QTL variance which is explained by the haplotype effects compared to the proportion of QTL variance which is explained by single marker effects, as discussed in section 2.3. Calus et al. (2007) compared the accuracy of GEBV for progeny without phenotypic records from genomic selection using single markers or marker haplotypes, in simulated data.

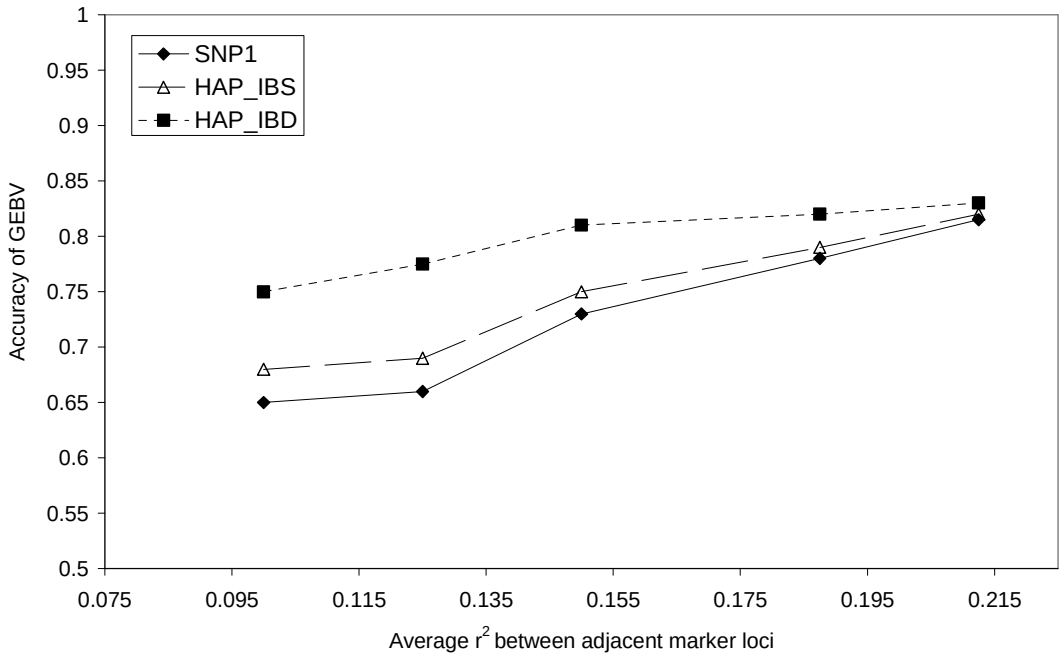


Figure 4.2 Accuracies of genomic breeding values estimated for animals with no phenotypic information with three different models of genomic selection: SNP1, using the single marker approach of Xu (2003), with the addition of a polygenic effect in the model; HAP_IBS using haplotypes of adjacent markers and method BayesB of Meuwissen et al (2001) to estimate haplotype effects, with the addition of a polygenic effect; HAP_IBD using windows of haplotypes of 10 markers in the approach of Meuwissen and Goddard (2004). With permission from the authors., Calus et al. (2007).

They constructed haplotypes from the two adjacent markers defining each chromosome segment. They found that the advantage of using haplotypes increased at lower marker densities (or lower r^2 values between adjacent makers). When the r^2 between adjacent markers was 0.2 or greater, there was little advantage in using marker haplotypes, Figure 4.2 Presenting the accuracy as a function of the average r^2 between adjacent markers, as Calus et al. (2007) do, is appealing as the results can be used to infer the number of markers required to achieve a desired accuracy of genomic selection given the extent of LD observed in the livestock species in question. However, in all cases the accuracy achieved with the IBD approach was higher than regression on single markers or markers haplotypes. This was particularly true at low densities of markers, probably due to the contribution from linkage.

3.3.4 Number of phenotypic records in the reference population

The accuracy of genomic selection will depend on the number of haplotype effects at the chromosome segments, and the number of phenotypic records per unique haplotype, or per marker allele if single markers are used. The more phenotypic records available, the more observations there will be per haplotype and the higher the accuracy of genomic selection. There are also large differences between statistical methodologies in the accuracy achieved with a low number of records. Meuwissen *et al*. (2001) compared the accuracy of least squares, BLUP and BayesB with different numbers of phenotypic records, Table 4.2. Their results also suggest that in the order of 2000 phenotypic records are required to accurately estimate the haplotype effects. In their simulation, a heritability of 0.3 was used. If the heritability were higher, so that phenotype was a more accurate predictor of genotype, fewer records may be required. For example, in dairy cattle, daughter yield deviations (DYDs) are often used as the phenotype. DYDs can have an accuracy of 0.99.

Table 4.2: Correlations between true and estimated breeding values when the number of phenotypic records is varied (from Meuwissen et al. 2001, with permission from the authors)

	No. of phenotypic records		
	500	1000	2200
Least squares	0.124	0.204	0.318
Best linear unbiased prediction (BLUP)	0.579	0.659	0.732
BayesB	0.708	0.787	0.848

3.4 Non additive effects in genomic selection

While breeding values by definition should include only additive effects (genetic merit which is passed from one generation to the next), in some cases it may be desirable to predict genetic merit which better predict an animals actual phenotype, for example through the inclusion of dominance and epistatic effects. If phenotypes are used in the estimation of chromosome segment effects (rather than DYDs for example), inclusion of epistatic and dominance effects in the model could improve the

accuracy of estimating the additive effect of the chromosome segment effects. Further, dominance and epistatic effects can be exploited to produce sets of progeny with maximum genetic merit, through mate selection for example (Kinghorn 1998).

Estimates of dominance effects with single markers is straight forward, requiring extension of the genetic model to estimate two effects per SNP, rather than one:

$$y_j = \mu + \sum_i^p x_{ij} g_i + \sum_i^p w_{ij} d_i + e_i$$

Where x_{ij} and w_{ij} are defined as $x_{ij} = \sqrt{2}$ and $w_{ij} = -1$ for genotype A_1A_1, $x_{ij} = 0$ and $w_{ij} = 1$ for A_1A_2 and , $x_{ij} = -\sqrt{2}$ and $w_{ij} = -1$ for A_2A_2. If G_{11}, G_{12} and G_{22} are the genotypic coefficients for the three genotypes, then $g_i = G_{11} - G_{22}$ for the additive effect and $d_i = 2G_{12} - G_{22} - G_{11}$. The *x* and *w* coded in this way are independent and each has a zero expectation and unity variance. The Bayesian estimation method of Xu (2003) can then be extended to estimate d_i as well as g_i.

Estimation of epistatic effects is more difficult, due to extremely large number of marker by marker interactions in the single marker approach, or haplotype by haplotype interactions in the haplotype approach. Xu (2007) extended the single marker Bayesian approach in Xu (2003) to account for epistatic effects.
A model including epistatic effects can be written as:

$$y = \sum_{l=1}^{k} g_l \alpha_l + \sum_{l'>l}^{k} (g_l \times g_{l'}) \alpha_{ll'} + \varepsilon$$

Where $g_l \times g_{l'}$ is the element wise multiplication of vectors g_l and $g_{l'}$, α_l is the main effect of locus *l*, and $\alpha_{ll'}$ is the epistatic effect between locus *l* and *l'*. The model can be simplified to fit into the methodology of Xu (2003) by using *j* to index the j^{th} genetic effect for $j=1,q$, where $q=k(k+1)/2$. The model can then be re-written

$$\mathbf{y} = \sum_{\mathbf{j}=1}^{\mathbf{q}} \mathbf{Z_j} \gamma_\mathbf{j} + \boldsymbol{\varepsilon}$$

For example, $Z_j = g_l$ and $\gamma_j = \alpha_l$ if the j^{th} effect is a main effect, and $Z_j = g_l \times g_{l'}$ and $\gamma_j = \alpha_{ll'}$ if the j^{th} effect is an epistatic effect.

Xu (2007) used a similar approach to that in Xu (2003) to estimate the γ_j. A normal prior was assigned to the γ_j, where $\gamma_j \sim N(0, \sigma_j^2)$. The prior assigned to the σ_j^2 was $\sigma_j^2 \sim 1/\chi^2_{(\tau,\omega)}$. For details on this prior distribution see Xu (2007).

Xu (2007) showed that epistatic effects could be estimated both in simulated data with this approach using 600 records in a back-cross design. They also applied the method to real data from a barley backcross experiment.

Gianola et al. (2006) presented semi-parametric procedures for genomic selection which allowed them to estimate interactions between potentially hundreds of thousands of markers. Their methods included kernel regression, which regress marker effects according to a smoothing parameter *h*, embedded into the standard mixed model equations. Their model treated the variance of effects across chromosome segments as equal. In a small example, they achieved accuracies of up to 0.85 for predicted genotypic values in selection candidates with no phenotypes, when both dominance and epistasis were simulated. For more details see their paper.

3.5 Genomic selection with low marker density

The IBD methodology for genomic selection is particularly suited to cases where marker density is low, as in this case there will be some advantage in including the linkage information in the estimation of chromosome segment effects carried by each animal. Calus et al. (2007) demonstrated that use of the IBD approach can achieve high accuracies of genomic selection even with levels of r^2 between adjacent markers as low as 0.1, Figure 4.2. This result is however dependent on population structure. For example large sire half sib groups in the population will allow accurate estimation of sire haplotypes, such that linkage information contributes considerably to the accuracy of genomic selection.

In LD-MAS, a polygenic breeding value is included in the GEBV to pick up genetic variance not captured by the markers. In genomic selection as specified by Meuwissen et al. (2001), a polygenic component is not included in the prediction of GEBVs. However if the available marker density is not sufficient to ensure all QTL are in high LD with a marker of haplotype, inclusion of a polygenic component in the

GEBV from genomic selection would recapture some of the effects of the QTL which are not in sufficient LD with markers.

Even with a sparse marker map, genomic selection can also be used to increase the efficiency of development of composite lines (Piyasatian et al. 2006). Crosses between breeds will exhibit much greater levels of LD than within breed populations. Piyasatian et al. (2006) demonstrated that the genetic merit of composite lines can be improved by using genomic selection to capture chromosome segments with the largest effects from the contributing breeds, even with a sparse marker map.

3.6 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. 2007).

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, Kaupe et al. 2004). 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. 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.

3.7 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 r2 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 4.3.

Table 4.3. 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 (2007) 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.

3.8 Cost effective genomic selection

Depending on the genotyping technology used, the cost of genotyping animals for ~ 30 000 SNPs may be $500, while the cost of genotyping animals for ~ 50 SNPs may be as low as $20. If the number of markers required to apply genomic selection can be reduced, this could represent a large saving to the breeding program (and may make the difference between applying or not applying genomic selection).

There are two possibilities to reduce the number of markers in genomic selection. When the method BayesB of Meuwissen et al. (2001) is applied to estimate chromosome segment effects in the reference population, many of the chromosome segment effects will be set to close to zero. So genotyping the markers in these

segments in animals where GEBVs are to be predicted using generations has no value. In other words only the subset of markers in chromosome segments with a non-zero effect need be genotyped. One problem with this approach occurs when genomic selection is extended to multiple traits. If the selection criteria includes say 30 traits, and there are 50 markers per trait with non-zero effects, then the total number of markers to be genotyped may be ~ 1500. For most genotyping platforms, the cost of genotyping 1500 markers is close to the cost of genotyping 30000 markers!

3.9 Optimal breeding program design with genomic selection

Genomic selection allows prediction of very accurate EBVs for young animals. The effect of such information on the optimal breeding program design for the different livestock industries could be profound.

In dairy cattle breeding, progeny testing is currently used to identify bulls of high genetic merit. A good description of the progeny test scheme was given by Schaeffer et al. (2006) "In the progeny test scheme, a number of elite cows are identified each year as the dams of young bulls, and these cows are mated to specific sires". At one year of age, the young bulls are test mated to a large number of cows in the population, in order that they will have about 100 daughters with their first EBVs for production and other traits. Approximately 43 months later the daughters from these matings complete their first lactations and the young bull EBVs for production are produced with an accuracy of approximately 75%. At this point the young bull is proven or returned to service." As suggested by Schaeffer et al. (2006), genomic selection allows GEBVs with an accuracy of 0.75 or greater to be calculated for bull calves. Bull calves can therefore be selected at this stage, rather than following progeny testing. This reduces the generation interval by at least half. Further genetic gains can be made by genotyping the elite bull dams and selecting a smaller number for mating to specific sires. Schaffer (2006) suggested the effect of genomic selection may be to shift the structure of the dairy cattle breeding industry to a model similar to that used by the poultry and swine industries, where companies maintain a nucleus of elite animals "within house". Another effect of genomic selection may be

more appropriate balance in the direction of genetic gain. Currently in the dairy industry, large gains are made for production traits, while the gains in fertility are relatively smaller, in part due to the lower accuracy of fertility EBVs (and also because production and fertility are unfavourable correlated). Genomic selection could increase the accuracy of fertility EBVs, if sufficient records were taken in the initial experiment to estimate chromosome segment effects, allowing greater contribution of this trait to the total breeding objective.

In the pig. 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.

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.

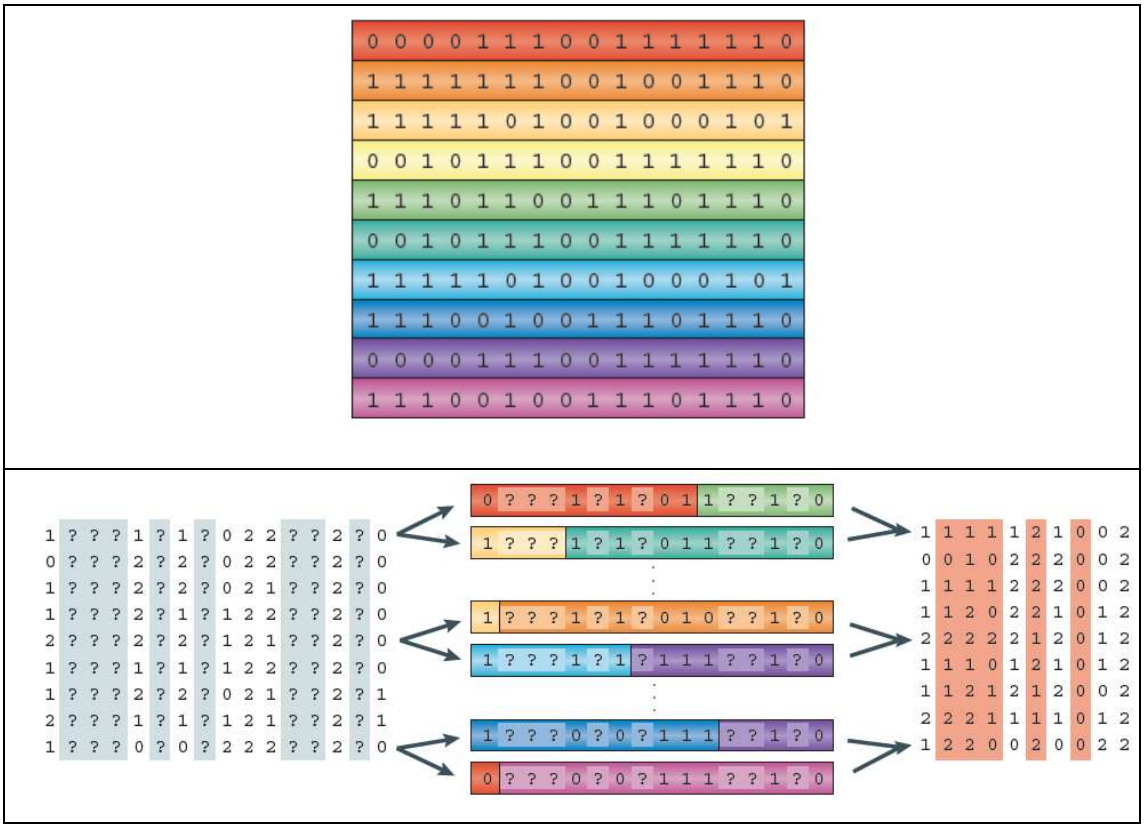


Figure 1. From Marchini and Howie (2010). A cartoon of genotype imputation. A. A phased reference population is the a requirement in many imputation programs. B. The genotypes in the target population are phased, then assigned a mosaic of the reference haplotypes via a hidden markov model.

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; one in Nature Reviews Genetics (Marchini and Howie 2010), and another in Human Genetics (2008). 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 et al. 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 and Stephens2009), 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

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

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 and Howie 2010).

BEAGLE (Browning and Browning 2008). 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 2008).

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.

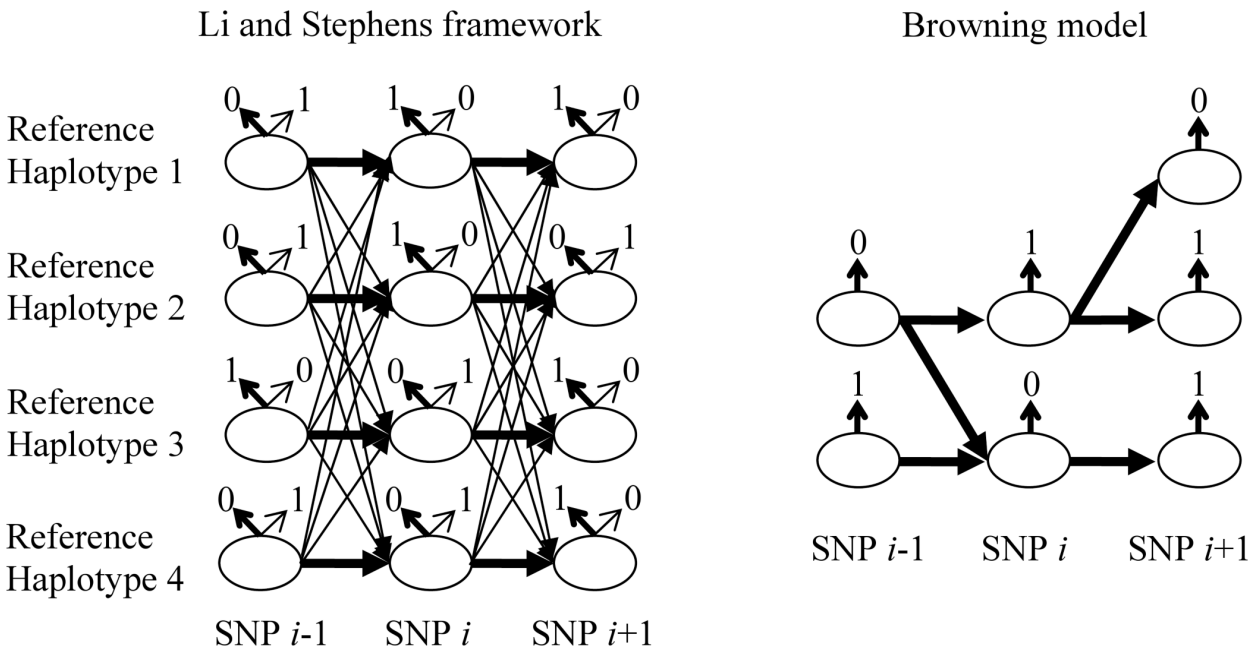


Figure 2. Illustration highlighting major differences between models based on the Li and Stephens framework (2003), the basis for MACH, IMPUTE and fastPHASE, and the Browning model (Browning 2006), the basis for BEAGLE. Excerpts of the models covering three markers (SNPs i-1, i and i+1) are shown. Ovals 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 Kon 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.

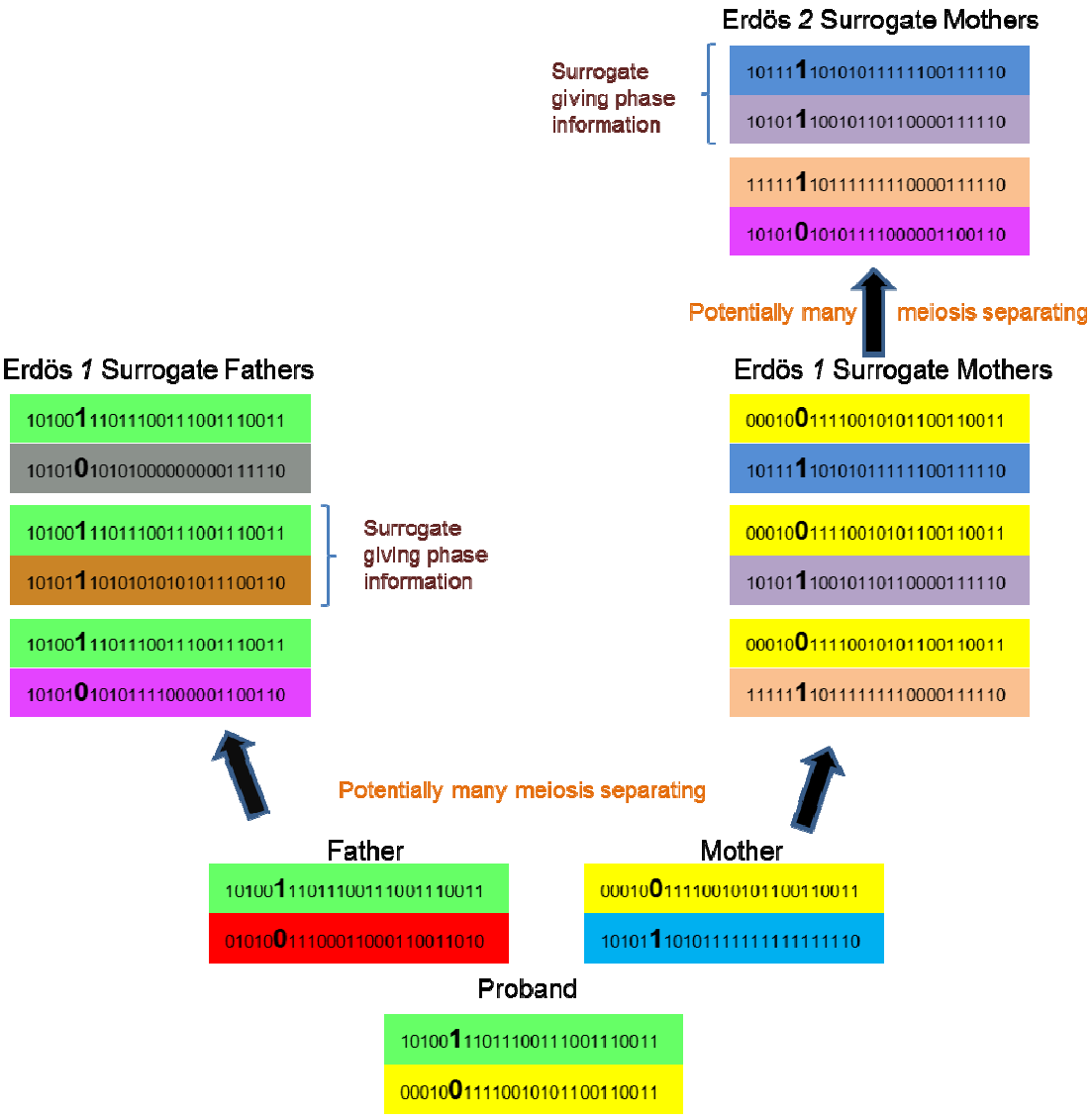


Figure 3. From Hickey et al. (2011). Illustration of the long range phasing process.

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. In a follow up paper, Weigel et al (2011) 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 and Georges 2009), 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 outline 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. (2010) 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 (eg. 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 (eg. 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 (eg Zhang and 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.

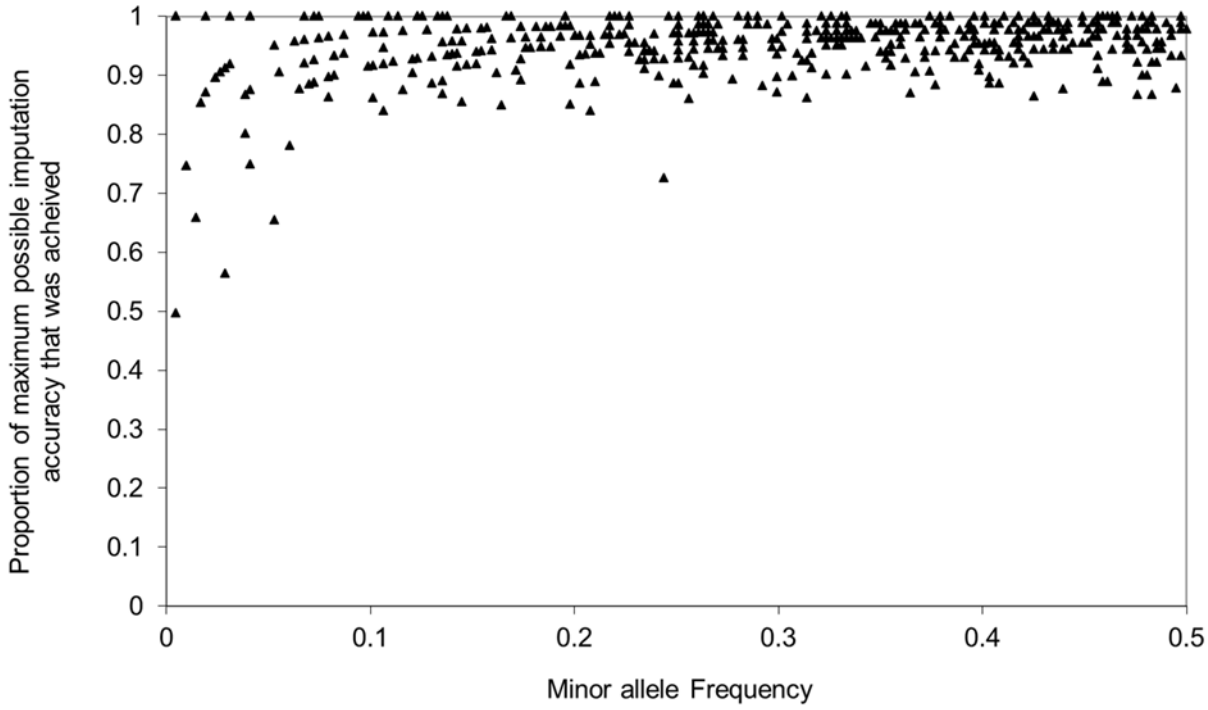


Figure 4. From Hayes et al. (2011). Proportion of maximum possible imputation accuracy that was achieved (50K to high density genotypes) by minor allele frequency, in a terminal sire sheep breed. The proportion of maximum possible imputation accuracy was calculated as the accuracy of imputation that 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 (2011)!

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. 2009). 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 (eg Meuwissen and Goddard 2010, 1000 Genomes consortium 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 amonst 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.

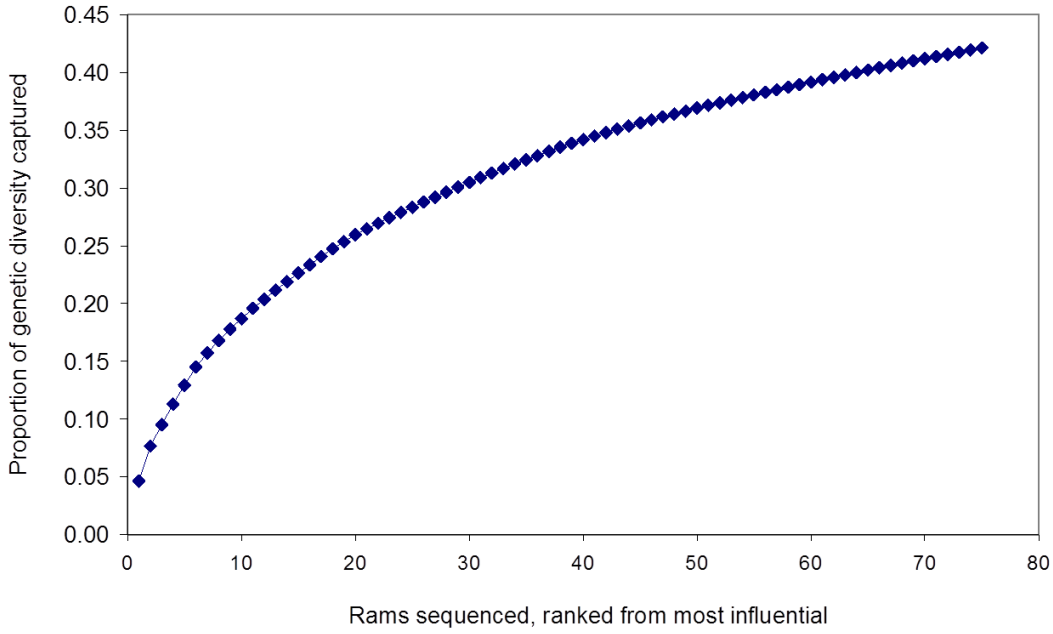


Figure 1. Proportion of genetic diversity (as measured by pedigree), captured by subsets of groups of rams, ranked from the most to least influential.

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 2011, 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 (2011), 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, 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 Haplotyping with the PHASE program

The above exercise assumes that the genotypes of each animal have already been sorted into haplotypes. In a real data set, this will not be the case. If the population has large half sib family structure, resolving the genotypes is relatively straightforward. In some situations pedigree information may no be available, or you may deliberately choose to randomly sample animals from the population for LD mapping. With this type of data it is possible to use the PHASE program (Stephens et al 2001). The program is available for free download at http://www.stat.washington.edu/stephens/software.html. Note that the program is only designed to resolve short range haplotypes, eg many markers in a single cM.

Exercise 3.2.1. Haplotyping with the PHASE program.
The casein genes in goats are good candidates for harbouring a mutation affecting milk production, as casein constitutes around 80% of caprine milk protein. You have found 10 SNPs in the goat casein genes, and want to sort the genotypes into haplotypes for LD analysis. 205 goats have been genotyped for the 10 SNPs.

The PHASE input file (goat_geno.txt) has the following format:

```
205
10
P       264     866     888     1105    1169    1379    1470    6075    6091    9889
SSSSSSSSSS
38362
        A       C       G       G       G       C       G       T       C       C
        A       C       G       G       G       C       G       T       C       C
38393
        A       C       G       G       G       T       G       ?       ?       T
        A       C       A       A       A       C       A       ?       ?       C
38421
        A       C       G       ?       G       T       G       ?       G       T
        A       C       A       ?       A       C       A       ?       C       C
38452
        ?       C       G       G       G       T       G       T       ?       C
        ?       C       A       A       A       C       A       C       ?       C
```

Where the first line is the number of animals in the analysis, the second line is the number of SNPs, the third line is the position of the SNPs (begin this line with P) in bases, the next line is the type of marker for each marker (S=SNP,M=microsatellite). Missing alleles are coded as ?.

Then for each animal, there is an ID, followed by the genotypes at each SNP
Marker1 allele1, marker2 allele 1 …..
Marker1 allele2, marker2 allele 1……

To run the phase program, you will need to type the following:

PHASE <filename.inp> <filename.out> <number of iterations> <thinning interval> <burn in>.

For the data set goat_geno.txt, 100 iterations with thinning interval 2 and burn in 10 iterations should be sufficient.

Run the PHASE program. Go to output file. How many unique haplotypes are there? Do they have the same frequency in the population?.

The PHASE program usually predicts a number of haplotypes with very low frequency. What we want to now if the probability that these haplotypes really exist. So, take one of the rare haplotypes from the *.out file. Then, in the *.pairs file, you can see the probability for each animal of carrying a certain haplotype configuration. Are you satisfied that your rare haplotype really exists?

6.2 Estimating the extent of linkage disequilibrium

The GOLD program (for Graphical Overview of Linkage Disequilibrium) calculates linkage disequilibrium statistics from haplotype data ((Abecasis and Cookson, 2000). The statistics calculated are D, r^2 (which the authors call delta) and D'. The program also gives a nice graphical picture of the extent of linkage disequilibrium. The program is freely available from http://www.sph.umich.edu/csg/abecasis/GOLD/.

We will use the **haploxt** program from GOLD to calculate the extent of linkage disequilibrium between pairs of SNPs in practical 5.1. The inputs to the program are marker haplotypes (eg output from the phase program) and a map of the markers.

The file map.gm should look like (the header must be included):

```
MARKERID    NAME       LOCATION
       1    SNP1   266
       2    SNP2   864
       3    SNP3   888
……….
```

You can create this file from the positions line in the file goat_geno.txt

Next you will need to create a file with a list of haplotypes, called HAPLO.LST.

The format of this file is:

```
HAPLO_1    1 1 1 2 1 2 1
HAPLO_2    1 2 2 1 2 1 2
HAPLO_3    2 1 1 2 1 0 0
```

Eg. one line for each haplotype. You can create this file from the list of best pairs from PHASE, using excel for example. Note that in PHASE, we have used A,C,T and G to code the SNP alleles. These must be replaced with 1,2,3 and 4 in the *.lst file. Save your file to a location on your c: drive. If you used excel to create the file, save it as a text file, and then remove the .txt extension.

Open a DOS window, and go to the directory containing the *.lst file. Run the program **haploxt** in this directory by typing **haploxt**.

The program will produce a file called LD.XT. This a table of LD values for each marker pair. Open this file. Plot the values D' and r^2 against each other in excel. Is the value of D' usually larger or smaller than r^2?

Now open the gold program (click on the gold icon). Load the disequilibrium data (the file you have just produced, LD.XT). The load the map file (map.gm). View the LD across the segment with the delta squared statistic. Are there any regions of very high LD? Why do you think this is? In general, what is the relationship between distance between the SNPs and LD? Are there any exceptions to this across the chromosome segment?

Another useful program in the GOLD package is **ldmax**. This program estimates r^2 values from genotypes. So there is no need to haplotype the data first. The "cost" of using this program could be less accurate estimates of r^2 values. To investigate the effect of using genotype data to estimate r^2, we will use the genotype data from practical 5.1 The data format for ldmax is

```
<famid> <pid> <fatid> <motid> <sex> <genotype_1> ... <genotype_n>
```

<famid> is a unique identifier for each family, and within each family **<pid>** is a unique identifier for an individual. **<fatid>** and **<motid>** identify the individuals father and mother (if this line refers to a founder, these should be set to zero). **<sex>** denotes the individuals sex, using the convention 1=male, 2=female. Each **<genotype>** is encoded as two integer allele numbers. The pedigree columns should be separated by spaces.

An example pedigree file fragment, describing a single nuclear family genotyped at 3 markers would be:

```
100 1 0 0 1  1 2  1 2  1 2
100 2 0 0 2  1 2  1 2  1 2
100 3 1 2 1  1 1  2 2  1 1
```

This describes a family (labelled 100), contains two founders (1 and 2), and their single offspring (3). The founders are heterozygous at all marker loci, while the offspring is homozygous at all loci.

In the case of the goat genotype data, we will assume all animals are founders. The input file for **ldmax**, *qtdt.ped* has already been made for you. First rename your LD.XT file so you don't lose it. Then run the **ldmax** program, which also produces the file LD.XT. Now plot the delta^2 values from **haploxt** and **ldmax** (using excel for example). How similar are they? Do you think ldmax is giving reliable estimates of r^2 in this case?

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.

Before we use this design pack, lets 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.

Lets 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)

Now we will return to the ldDesign package. Hit the "packages" button on the top of the screen. Then click load pacjages 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 using a prior for chromosome segment variances with 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?

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. And thank you to Mario Calus for providing his unpublished manuscript.

8. References

Andersson L, Georges M. 2004. Domestic-animal genomics: deciphering the genetics of complex traits. *Nat Rev Genet*. **5**(3):202-212.

Benjamini, Y., and Y. Hochberg. 1995 Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* **57**: (1): 289–300.

Bennewitz, J., N. Reinsch, J. Szyda, F. Reinhardt, C. Kuhn, M. Schwerin, G. Erhardt, C. Weimann, and E. Kalm. 2003. Marker assisted selection in German Holstein dairy cattle breeding: Outline of the program and marker-assisted breeding value estimation. Page 5 in *Book of Abstr. 54th Annu. Mtg. Eur. Assoc. Anim. Prod*. Y. van der Honing, ed. Wageningen Academic Publishers, Wageningen, The Netherlands.

Boichard, D., S. Fritz, M. N. Rossignol, M. Y. Boscher, A. Malafosse, and J. J. Colleau. 2002. Implementation of marker-assisted selection in French dairy cattle. Electronic communication 22–03 in *Proc. 7th World Cong. Genet. Appl. Livest. Prod*., Montpellier, France.

Bovenhuis, H. and Meuwissen, T. 1996. Course *Detection and Mapping of quantitative trait loci*, 16-19 April, University of New England, Armidale, NSW, Australia.

Calus, M. P. L., Meuwissen, T. H. E., de Roos, A. P. W. and Veerkamp, R. F. 2007. Accuracy of genomic selection using different methods to define haplotypes. *Genetics.* Submitted.

Churchill, G. A. and Doerge, R. W. 1994. Empirical threshold values for quantitative trait mapping. *Genetics* **138**:963-971.

Cohen-Zinder M, Seroussi E, Larkin DM, Loor JJ, Everts-van der Wind A, Lee JH, Drackley JK, Band MR, Hernandez AG, Shani M, Lewin HA, Weller JI, Ron M. 2005. Identification of a missense mutation in the bovine ABCG2 gene with a major effect on the QTL on chromosome 6 affecting milk yield and composition in Holstein cattle. *Genome Res* 15:936-944.

Darvasi, A. and Soller, M. 1997. A simple method to calculate resolving power and confidence interval of QTL map location. *Behavior Genetics* **27:** 125-132.

Dekkers JC. 2004. Commercial application of marker- and gene-assisted selection in livestock: strategies and lessons. *J Anim Sci.* **82** E-Suppl:E313-328.

Dekkers, J. C. M., and J. A. M. van Arendonk. 1998. Optimum selection for quantitative traits with information on an identified locus in outbred populations. *Genet. Res*. **71**:257–275.

DeRoos, A. P. W, Goddard, M. E. And Hayes, B. J. 2007. Extent of linkage disequilibrium within and across dairy breeds. *J. Dairy Sci*. Submitted.

DeRoos, A. P. W., Schrooten, C., Mullart, E., Calus, M., Veerkamp, R. 2007. Genomic selection for fat percentage using markers on BTA14. *J. Dairy Sci*. Submitted.

Du, F-X, Clutter, A. C and Lohuis, M. M. 2007. Characterizing linkage disequilibrium in pig populations. *Int. J. Biol. Sci***. 3**:166-178.

Dunner. S, Miranda, M.E., Amigues, Y. *et al.* (2003) *Genet Sel Evol*.**35**:103

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. and Ponder, B.A.J. 2000. The extent of linkage disequilibrium in four populations with distinct demographic histories. *American Journal of Human Genetics* **67:** 1544-1554.

Ewing B, Green P. 2000. Analysis of expressed sequence tags indicates 35,000 human genes. *Nat Genet.* **25:**232-4.

Farnir, F., Coppieters, W., Arranz, J.J., Berzi, P., Cambisano, N., Grisart, B., Karim, L., Marcq, F., Moreau, L., Mni, M., Nezer, C., Simon, P., Vanmanshoven, P., Wagenaar, D. and Georges, M. 2000. Extensive genome-wide linkage disequilibrium in cattle. *Genome Research* **10:** 220-227.

Farnir, F., Grisart, B., Coppieters, W., Riquet, J., Berzi, P., Cambisano, N., Karim, L., Mni, M., Moisio, S., Simon, P., Wagenaar, D., Vilkki, J. and Georges, M. 2002. Simultaneously mining of linkage and linkage disequilibrium to fine map quantitative trait loci in outbred half-sib pedigrees: Revisiting the location of a quantitative trait locus with major effect on milk production on bovine chromosome 14. *Genetics* **161:** 275-287.
Fernando, R. and Grossman, M. 1989. Marker assisted selection using best linear unbiased prediction. *Genetics Selection Evolution* **21:** 467-477.
Fernando, R. L., and M. Grossman. 1989. Marker-assisted selection using best linear unbiased prediction. *Genet. Select. Evol.* **21**:467–477.
Fernando, R. L., D. Nettleton, B. R. Southey, J. C. M. Dekkers, M. F. Rothschild et al. 2004. Controlling the proportion of false positives in multiple dependent tests. *Genetics* **166**: 611–619.
Fischer, R. A. 1918. The correlation between relatives: the supposition of mendelain inheritance. *Transactions of the royal society of Edinburgh*. **52**:399.
Galloway, S. M., McNatty, K. M., Ritvos, O. and Davis, G. H. 2002. Inverdale: a case study in gene discovery. *Proc. Assoc. Anim. Breed. Genet.* 14:7-10.
George, A.W., Visscher, P.M. and Haley, C.S. 2000. Mapping quantitative trait loci in complex pedigrees: A two-step variance component approach. *Genetics* **156:** 2081-2092.
Georges, M., and J. M. Massey. 1991. Velogenetics, or the synergistic use of marker assisted selection and germ-line manipulation. *Theriogenology* **25**:151–159.: evidence for the trans interaction of reciprocally imprinted genes. *Trends in Genetics* **19**: 248–252.
Georges, M., Nielsen, D., Mackinnon, M., Mishra, A., Okimoto, R., Pasquino, A.T., Sargent, L.S., Sorensen, A., Steele, M.R., Zhao, X., Womack, J.E. and Hoeschele, I. 1995. Mapping quantitative trait loci controlling milk production in dairy cattle by exploiting progeny testing. *Genetics* **139:** 907-920.
Gianola, D., Fernando, R. L., Stella, A. 2006. Genomic-Assisted Prediction of Genetic Value With Semiparametric Procedures. *Genetics* **173**: 1761-1776.
Gianola, D., Perez-Enciso, M. Toro, M. A. 2003. *Genetics* **163**:347-365.
Gibson, J. P. 1994. *Proc. 5th World Congr. Genet. Appl. Livest. Prod*. **21**:201-204.
Gilmour, A.R., Gogel, B.J., Cullis, B.R., Welham, S.J. and Thompson, R. 2002. *ASReml user guide release 1.0*. VSN International Ltd, Hemel Hempstead, HP11ES, UK.
Goddard, M. E., Chamberlain, A. C. and Hayes, B. J. 2006. Can the same markers be used in multiple breeds? *Proc 8th World Cong. Genet. Appl. Livest.* Belo Horizonte, Brasil.
Goddard, M.E. 1991. Mapping genes for quantitative traits using linkage disequilibrium. *Genetics Selection Evolution* **23:** 131s-134s.
Grapes, L., Dekkers, J.C., Rothschild, M.F., Fernando, R.L. (2004) *Genetics*. **166**:1561
Grapes, L., Firat, M.Z., Dekkers, J.C., Rothschild, M.F. and Fernando RL. (2006) *Genetics*. **172**:1955
Haley, C. S. and Visscher, P. M. 1998. *J. Dairy Sci.* **81**: 85-97.
Haley, C.S. and Knott, S.A. 1992. A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity* **69:** 315-324.
Haley, C.S., Knott, S.A. and Elsen, J.M. 1994. Mapping quantitative trait loci in crosses between outbred lines using least squares. *Genetics* **136:** 1195-1207.
Hayes, B. J, Kelly, M. J. and Miller, S. P. 2007. BMC Genetics. In Prep.
Hayes, B. J. and Goddard, M.E. 2001. The distribution of the effects of genes affecting quantitative traits in livestock. *Genetics Selection Evolution* **33:** 209-229.
Hayes, B. J. Visscher, P. M., McPartlan, H. and Goddard, M. E. 2003. A novel multi-locus measure of linkage disequilibrium and it use to estimate past effective population size. *Genome Research* **13:635.**
Hayes, B. J., Chamberlain, A. and Goddard, M. E. 2006. Use of linkage markers in linkage disequilibrium with QTL in breeding programs. *Proc. 8th World. Congr. Genet. Appl. Livest. Prod.* Belo Horizonte, Brazil, Vol. pp.
Hayes, B. J., Chamberlain, A. C., McPartlan, H., McLeod, I., Sethuraman, L., Goddard, M. E. 2007. Accuracy of marker assisted selection with single markers and marker haplotypes in cattle. *Genetical Research* Submitted.
Hayes, B., and M. E. Goddard. 2003. Evaluation of marker assisted selection in pig enterprises. *Livest. Prod. Sci*. **81**:197–211.
Heifetz EM, Fulton JE, O'Sullivan N, Zhao H, Dekkers JC, Soller M 2005. Extent and Consistency Across Generations of Linkage Disequilibrium in Commercial Layer Chicken Breeding Populations. *Genetics*. **171**: 1173-1181.

Henderson, C. R. 1984. Applications of linear models in animal breeding. *Can. Catal. Publ. Data, Univ Guelph, Canada*.
Henshall, J.M. and Goddard, M.E 1997. *Proc. 12th Assoc. Advanc. Anim. Breed. Genet*. **12**:217-221.
Hill, W. G. 1981. Estimation of effective population size from data on linkage disequilibrium. *Genetical Research* **38:** 209--216.
Hill, W. G. and Robertson, A. 1968. Linkage disequilibrium in finite populations. *Theor. Appl. Genet.* 38:226-231.
Jeon JT, Carlborg O, Tornsten A, Giuffra E, Amarger V, Chardon P, Andersson-Eklund L, Andersson K, Hansson I, Lundstrom K, Andersson L. 1999. A paternally expressed QTL affecting skeletal and cardiac muscle mass in pigs maps to the IGF2 locus. *Nat Genet*. **21**(2):157-8.
Kaupe, B., Winter, A., Fries, R. and Erhardt G. (2004) *J Dairy Res*. **71**:182
Khatkar, M S., Zenger, K. R. Hobbs, M., Hawken, R. J. Cavanagh, J. A. L. Barris, W., McClintock, A. E. McClintock, S. Thomson, P. C., Tier, B. Nicholas F. W. and Raadsma. **H. W**. 2007. A primary assembly of a bovine haplotype block map based on a 15,000 single nucleotide polymorphism panel genotyped in Holstein-Friesian cattle. *Genetics*. In Press.
Kinghorn, B.P. 1998. Mate selection by groups. *J Dairy Sci*. **81**: Suppl 2:55-63.
Kruglyak, L. 1999. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature Genetics* **22:** 139-144.
Lander, E.S. and Botstein, D. 1989. Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* **121:** 185-199.
Lander, E.S. and Schork, N.J. 1994. Genetic dissection of complex traits. *Science* **265:** 2037-2048.
Lee SH, van der Werf JH. 2004. The efficiency of designs for fine-mapping of quantitative trait loci using combined linkage disequilibrium and linkage. *Genet Sel Evol.* **36:**145.
Luo, Z. W. 1998. Linkage disequilibrium in a two-locus model. Heredity **80:** 198–208.
MacLeod, I. M., Hayes, B. J., Savin, K., Chamberlain, A. J., McPartlan, H. and Goddard, M. E. 2007. Power of dense bovine single nucleotide polymorphisms (SNPs) for genome scans to detect and position quantitative trait loci (QTL). *Genetics.* Submitted.
Mangin, B., Goffinet, B. and Rebai, A. 1994. Constructing confidence intervals for QTL location. *Genetics* **138:** 1301-1308.
Maynard Smith J, Haigh J. 1974. The hitch-hiking effect of a favourable gene. *Genet Res Camb*. **23**:23–35.
McRae, A.F., McEwan, J.C., Dodds, K.G., Wilson, T., Crawford, A.M. and Slate, J. 2002. Linkage disequilibrium in domestic sheep. *Genetics* **160:** 1113-1122.
Meuwissen TH, Goddard ME. 2004. Mapping multiple QTL using linkage disequilibrium and linkage analysis information and multitrait data. *Genet Sel Evol.* **36**(3):261-79.
Meuwissen, T. H. E., B. Hayes, and M. E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157**:1819–1829
Meuwissen, T.H.E. and Goddard, M.E. 1996. The use of marker haplotypes in animal breeding schemes. *Genetics Selection Evolution* **28:** 161-176.
Meuwissen, T.H.E. and Goddard, M.E. 2001. Prediction of identity by descent probabilities from marker-haplotypes. *Genetics Selection Evolution* **33:** 605-634.
Meuwissen, T.H.E., Hayes, B.J. and Goddard, M.E. 2001. Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157:** 1819-1829.
Meuwissen, T.H.E., Karlsen, A., Lien, S., Olsaker, I. and Goddard. M.E. 2002. Fine mapping of a quantitative trait locus for twinning rate using combined linkage and linkage disequilibrium mapping. *Genetics* **161:** 373-379.
Olsen HG, Lien S, Gautier M, Nilsen H, Roseth A, Berg PR, Sundsaasen KK, Svendsen M, Meuwissen TH. 2005. Mapping of a milk production quantitative trait locus to a 420-kb region on bovine chromosome 6. *Genetics*. **169**:275-83
Phillips MS, Lawrence R, Sachidanandam R, Morris AP, Balding DJ, Donaldson MA, Studebaker JF, et al. 2003. Chromosome-wide distribution of haplotype blocks and the role of recombination hotspots. *Nat Genet* **33**:382–387
Piyasatian, N. Fernando, R, L. Dekkers, J. C. M. 2006. Genomic selection for composite line development using low density marker maps. *Proc. 8th World. Congr. Genetics. Appl. Livest Prod. Be*lo Horizonte, Brasil.
Plastow, G., S. Sasaki, T-P. Yu, N. Deeb, G. Prall, K. Siggens, and E. Wilson. 2003. Practical application of DNA markers for genetic improvement. Pages 151–154 in *Proc. 28th Annu. Mtg. Natl. Swine Improve*. Fed., Iowa State Univ., Ames.
Pritchard JK, Przeworski M. 2001. Linkage disequilibrium in humans: models and data. *Am J Hum Genet* **69**:1–14.

Pritchard JK, Stephens M, Rosenberg NA, Donnelly P. 2000. Association Mapping in Structured Populations. *Am J Hum Genet.* 67: 170-181.
Reich, D.E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P.C., Richter, D.J., Lavery, T., Kouyoumjlan, R., Farhadian, S.F., Ward, R. and Lander, E.S. 2001. Linkage disequilibrium in the human genome. *Nature* **411:** 199-204.
Rabiner, L. R. A tutorial on hidden Markov models and selected applications in speech recognition. Proc. IEEE 77, 257–286 (1989).
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. and Georges, M. 1999. Fine-mapping of quantitative trait loci by identity by descent in outbred populations: Application to milk production in dairy cattle. *Genetics* **96:** 9252-9257.
Rothschild MF, Larson RG, Jacobson C, Pearson P. 1991. PvuII polymorphisms at the porcine oestrogen receptor locus (ESR). *Anim Genet*. **22**(5):448.
Sabeti PC, Reich DE, Higgins JM, Levine HZP, Richter DJ, et al. 2002. Detecting recent positive selection in the human genome from haplotype structure. *Nature*. **419**:832–837.
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**: 445-459.
Sobel E, Lange K. 1996. Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker-sharing statistics. *Am J Hum Genet.* **58**:1323-37.
Solberg, T. R., Sonesson, A. Wooliams, J. Meuwissen, T. H. E. 2006. Genomic selection using different marker types and density. *Proc. 8th World. Congr. Genetics. Appl. Livest Prod. Be*lo Horizonte, Brasil.
Spelman, R. J., and D. J. Garrick. 1998. Genetic and economic responses for within-family markers-assisted selection in dairy cattle breeding schemes. *J. Dairy Sci*. **81**:2942–2950
Spelman, R. J., and H. Bovenhuis. 1998. Moving from QTL experimental results to the utilisation of QTL in breeding programmes. *Anim. Genet*. **29**:77–84.
Spelman, R. J., Garrick, D. J. and van Arendonk, J. A. M. (1999) *Livest. Prod. Sci*. **59**: 51-60.
Spelman, R.J., Ford, C.A., McElhinney,*et al. et al.* (2002) *J Dairy Sci*. **85**:3514.
Spielman RS, McGinnis RE, Ewens WJ. 1993. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet* **52**:506–513.
Stephens M, Smith NJ, Donnelly P. 2001. A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet*. **68**:978-89.
Storey, J. D. 2002. A direct approach to false discovery rates. *J. R. Stat. Soc. Ser. B* **64:** 479–498.
Sved, J.A. 1971. Linkage disequilibrium and homozygosity of chromosome segments in finite populations. *Theoretical Population Biology* **2:** 125-141.
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 Res*. **17**: 520 - 526
ter Braak CJ, Boer MP, Bink MC. 2005. Extending Xu's Bayesian Model for Estimating Polygenic Effects Using Markers of the Entire Genome. *Genetics.* **170**: 1435-1438.
Voight BF, Kudaravalli S, Wen X, Pritchard JK. 2006. A map of recent positive selection in the human genome. *PLoS Biol* **4**:e72.
Wall JD, Pritchard JK. 2003. Assessing the Performance of the Haplotype Block Model of Linkage Disequilibrium. *Am J Hum Genet*. **73**: 502-515.
Weller, J. I. Shlezinger, M. and Ron, M. 2005. Correcting for bias in estimation of quantitative trait loci effects. *Genet. Sel. Evol.* 37: 501-522.
Weller, J. I., Song, J. Z., Heyen, D. W., Lewin, H. A. and Ron, M. 1998. A new approach to the problem of multiple comparisons in the genetic dissection of complex traits. *Genetics* **150**:1699-1706.
Weller, J.I., Kashi, Y. and Soller, M. 1990. Power of "daughter" and "granddaughter" designs for genetic mapping of quantitative traits in diary cattle using genetic markers. *Journal of Dairy Science* **73:** 2525-2537.
Whittaker, J. C., Haley, C. Thompson, R. 1997. *Genet. Res*. **69**:137-144.
Whittaker, J. C., Thompson, R., Denham, M. C. 2000. *Genet. Res.* **75**:249-252.
Xu, S. 2003. 2003. Estimating polygenic effects using markers of the entire genome. *Genetics*. **163**: 789-801.
Xu, S. Jia, Z. 2007. Genome-wide Analysis of Epistatic Effects for Quantitative Traits in Barley. *Genetics.* In Press.
Zeng, Z.B. 1994. Precision mapping of quantitative trait loci. *Genetics* **136:** 1457-1486.

Zenger, K.R., Khatkar, M.S., Cavanagh, J.A,, Hawken, R.J., Raadsma, H.W. (2007) *Anim Genet*. **38**:7
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**: 1975-1986
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. *Genet. Res*. 86: 77–87.
Zou, F. 2001. Efficient and robust statistical methodologies for quantitative trait loci analysis. PhD dissertation. University of Wisconsin – Madison, USA.