How can we best use DNA data in the selection of cattle?

M.E. Goddard

Victorian Department of Primary Industries and
University of Melbourne
Australia

Abstract

Some traits are controlled by single genes, such as many genetic abnormalities, and some cattlemen already use DNA tests for these genes. However, most economically important traits are controlled by many genes and influenced by the environment. Recently panels of 50,000 genetic markers called SNPs have become available and they provide an opportunity to select for all the traits we seek to improve. The first step in using these SNPs is to estimate the effect of each SNP on all the important traits. To do this requires a ‘discovery population’ of cattle that have been genotyped for the SNPs and measured for the traits. From this population a prediction equation is derived that predicts breeding value or progeny difference for a trait, say marbling, from the 50,000 SNP genotypes (breeding values are just twice the progeny difference). This prediction equation can then be applied to animals that have SNP genotypes but no phenotypic information to calculate a predicted breeding value that is sometimes called a molecular breeding value (MBV). Usually animals will have some other source of information about their breeding value (e.g. pedigree and ultrasound scans) that are used at present to calculate a traditional estimated breeding value (EBV) or EPD. Therefore the MBV should be combined with the traditional EBV to give a prediction of breeding value that is more accurate than either source of information alone. This new estimate has been called a genomic EBV (GEBV). Obtaining a prediction equation that accurately predicts breeding value from SNP genotypes has proven difficult in beef cattle. We have often found that a prediction that works in one breed or herd does not work in other breeds and herds. Therefore we need to estimate the prediction equation from very large discovery populations which include several breeds and the test or validate the prediction equations across large populations also comprising several breeds.

Introduction

Cattle breeders seek to use the genetic variation between animals to improve their herds and breeds. Genetic variation exists for almost all traits that we can measure or record. Sometimes this variation is controlled by a single gene (e.g. red vs. black coat colour) but often there are many genes where variation affects the trait and where environmental factors also affect the phenotype or performance of the animal as we observe it (e.g. growth rate). Regardless of the number of genes involved, genetic variation is caused by differences between animals in DNA sequence. These differences in sequence arise due to mutation and sometimes the mutated version and the original version both occur at intermediate frequency in the population and we speak of a polymorphism.

Scientists have long sought to find the variations in DNA sequence that cause variations in phenotype, especially in traits that are of economic importance, in the hope that they could be used in selection of better cattle. For some traits controlled by a single gene this search has been successful. For instance, we now know the variation in DNA sequence that determines red vs. black coat colour and the variation that causes the disease mannosidosis. However, for most traits we know few, if any, of the variations in DNA sequence that cause genetic variation.

One type of DNA variation is called a single nucleotide polymorphism (SNP). It is a position in the DNA where a single letter of the DNA sequence varies. For instance, it might be a 'T' in some animals but a 'G' in others. In fact animals receive one copy of the DNA sequence from their sire and one from their dam, so an animal can have one of three genotypes at such a SNP: GG, TT or GT.

Recently commercial 'SNP chips' that can genotype 50,000 SNPs have become available and given us new hope that genetic markers for economic traits could be found. These 50,000 SNPs are unlikely to cause differences in performance between animals but they cover the whole genome of the animal so that genes that do cause variation must be close to at least one SNP on the chip. Therefore we expect the SNPs to be in linkage disequilibrium with the polymorphic genes that actually cause variation in economic traits. Consequently, the SNPs are correlated with the causal variants and so can be used to predict the genetic or breeding value of animals for those traits.

In this paper I will consider how the data from SNP chips can best be used to predict breeding values. The answer depends on several factors. Firstly, it depends on the number of genes that have an important effect on the trait. This is considered in the next section entitled 'Genetic architecture of quantitative traits'. Regardless of the genetic architecture we need a discovery population of animals that have been measured for the trait and genotyped for the SNPs, from which to estimate a prediction equation that predicts breeding value from SNP genotypes. This predictions equation must be validated in an independent sample of animals. Finally a procedure to incorporate the SNP data on selection candidates into the calculation of estimated breeding values (EBVs) or EPDs is needed (EBVs are just twice EPDs). This paper will deal with each of these topics in turn.

Genetic architecture of quantitative traits

Most traits of economic importance are controlled by many genes and by environmental factors: but how many genes? The best data comes from genome wide association studies (GWASs) on human height which is a classical quantitative trait. These GWASs have measured tens of thousands of people for height and genotyped them for over 300,000 SNPs. The SNPs with the biggest effects explain 0.3% of the phenotypic variance (Visscher 2008). Much of the variance must be due to genes with even smaller effects, so it seems likely that there are >1000 genes affecting height. Our own studies on milk production traits in dairy cattle give a similar result.

Of course there are traits and populations where a single gene causes a large part of the genetic variance. The mutations that cause double muscling have a large effect and would explain a large part of the variation in muscling in a population in which the mutation occurred at intermediate frequency. However, it appears that such cases are rare. Probably there is a spectrum of effects that mutations can have on a trait. Most mutations have a very small effect but occasionally there are mutations of large effect (figure 1).

How common are the mutant alleles? If mutations have no effect on the phenotype or fitness of an animal (so called neutral mutations) their frequency drifts by chance. In this case, most mutant alleles are at very low frequency but some, by chance, become common. If the effective population size of the population is constant there is a simple formula for the distribution of allele frequencies (figure 2). However, cattle have experienced periods of low effective population size or inbreeding due to domestication and stud book formation. This disperses allele frequencies so that rare alleles are not as common as expected from figure 2 but more like figure 3.

The SNPs on the SNP chip are likely to be neutral and so there distribution of allele frequencies would be like those of fig 3 except that rare SNPs are hard to discover, so the SNPs on the chip are biased towards intermediate allele frequencies.

The mutations affecting economic traits are unlikely to be completely neutral. If they affect a trait they are likely to have an effect on natural fitness or on selection by human cattle breeders. If the mutant allele is more fit than the original allele, it will increase in frequency and may become fixed in the population (i.e. the only allele present). More commonly the mutant will be less fit and will exist at low frequency most of the time until it is eliminated. This is important because it means that variants we are trying to find, because they affect economic traits, are likely to have one rare allele and one common allele.

Further evidence for this comes from the study of the variance in quantitative traits caused by new mutations each generation. This has been measured in laboratory animals such as mice and is typically 0.001 times the environmental variance. If the mutations were neutral the genetic variance would build up each generation and reach a value of 2 x effective population size x 0.001 times environmental variance. In populations where the effective population size is >10000, this would lead to heritabilities > 0.95 which is not what we observe. Therefore selection must be acting to eliminate some of the variation introduced by mutation. Before the mutant is lost completely it will add to the genetic variation but it will usually be rare.

In summary, for most economic traits, we will need to track very many genes in order to explain a large part of the genetic variance. Although, in some cases, a single gene may have a large effect, typically we will need hundreds of genes to explain most of the genetic variance and at many of these genes one allele is likely to be rare.

The discovery population

Unfortunately we do not know the mutations that cause variation in traits of interest, so we have to rely on random SNPs spread over the whole genome, so that each causative mutation is correlated with at least one SNP. We genotype the SNPs and we measure the trait on the discovery population and from this experiment we calculate a prediction equation that predicts the trait from the SNP genotypes. This is the central concept of 'genomic selection' (Meuwissen et al 2001). They considered several statistical methods for calculating this prediction equation. The best method depends on the genetic architecture of the trait. If there are very many genes with very small effects on the trait, then the best method is the one Meuwissen *et al*. called BLUP. If only some of the SNPs are needed to track genes for the trait, then the best method is the one they called Bayes B. In practice we find that the accuracy of the prediction equation is similar whether we use BLUP or Bayes B. In either case a large proportion, if not all, of the 50,000 SNPs are needed to maximise the accuracy of prediction.

The factors that have the biggest effect on the accuracy of prediction are the size of the discovery population and the heritability of the trait. The discovery population should be at least several thousand animals. The lower the heritability the larger the discovery population needs to be. This is because the phenotype measured on an animal is a poor guide to its genetic value if the heritability is low. The highest 'heritability' occurs when we use the result of a progeny test on each animal instead of its own phenotype. (The EPD calculated from a progeny test is a better guide to the sire's genetic value than his own phenotype). The dairy industry is fortunate to have many progeny tested sires to use in its discovery population, but it is still necessary to use thousands of bulls to achieve an accuracy >0.7.

We (Goddard 2008, Hayes et al 2009) have developed a theory for predicting the accuracy of prediction equations that use the BLUP method of estimation. Fig 4 illustrates the number of records needed to achieve accuracies of 0.5 or 0.7. If we use the phenotype of animals in the discovery population for a trait with $h^2 = 0.3$, we will need 4,000 animals to achieve an accuracy of 0.5.

Can the prediction equation calculated from one breed be used in other breeds? The prediction equation depends on linkage disequilibrium between the SNPs on the chip and the mutations that actually cause variation in the trait. We have found that this linkage disequilibrium (LD) is only consistent from one *Bos taurus* breed to another if the SNP and the gene are <10 kb apart. The SNPs on the 50k chip are about 60kb apart and do not show consistent LD between breeds (de Roos et al 2008). Confirming this, we have found that a prediction equation estimated in Holsteins does not work in Jerseys.

Fig 5 illustrates the problem. In the fig 5, one breed has the '+' allele for the trait carried on the same chromosome as the 'T' allele at the SNP. However, in another breed the '+' allele is on the same chromosome as the 'G' allele or the causal gene is not even polymorphic (i.e. all animals carry the same allele, '-' or '+').

If you include both Holsteins and Jerseys in the discovery population it is possible to find a prediction equation that works in both breeds. This is not surprising if in fact there are many SNPs in LD with a causal gene and so it is possible to find one that is in consistent LD with the causal gene in both breeds. However, if you demand that the prediction equations works across 3 or 4 or more breeds, it becomes increasingly difficult to find a SNP on the 50k SNP chip that will work. We have used many breeds in a beef cattle discovery population and found that it is hard to find SNPs consistently associated with a trait presumably because associations that occur in one breed are diluted or cancelled out by associations in other breeds.

The solution is to use a SNP chip with many more SNPs so that there is always a SNP very close (<10kb) to each causal gene and in consistent LD with it across *B. taurus* breeds. Hopefully a SNP chip with >300,000 SNPs will be available in the future. Even then I believe we will need different prediction equations for *B.indicus* breeds and crosses between *B. taurus* and *B. indicus*.

Validation of the prediction equations

The accuracy of a prediction equation cannot be judged in the discovery population from which it was derived. With 50,000 SNPs to choose from it is always possible to find some combination of them that works. You must assess the accuracy of the prediction equation by applying it to another independent group of animals that have been genotyped and measured for the trait. This can be a part of the discovery population that has been 'put aside' and not used for calculating the prediction equation.

However, it should not be a random subset of the discovery population that has been put aside. Generally the prediction equation works much better in a random subset of the discovery population than in a completely new group of animals even if they are from the same breed. Since it is not in the discovery population that we want to use the prediction equation but in new groups of 'selection candidates', it is important that the validation be done in a group that are representative of the selection candidates where the prediction equation will be applied.

We have found that prediction equations frequently do not work as well in a validation test as expected from the discovery population. I suspect that one reason for this is that one allele at the causal gene is rare. Consequently a gene that is polymorphic (i.e. variable) in one sample of a breed may not be polymorphic in another sample of the same breed, or one allele may be so rare in the validation sample that we have little power to detect it.

In summary, the frequent disappointments in the validation of prediction equations appear to be for four reasons:

- the effects of individual genes are very small and so hard to estimate accurately
- linkage disequilibrium is different in different breeds and so a prediction equation doesn't work in a new breed
- causal genes often have one rare allele whose frequency varies widely between samples of a breed and between breeds so that the variance caused by the gene varies widely. In extreme cases the gene may not be variable in some breeds.
- causal genes with one rare allele are in poor LD with the SNPs on the SNP chip and so not well predicted by a prediction equation based on these SNPs.

To improve on the past disappointing history of validation we need:

- Very large discovery populations so that the effects of SNPs are estimated accurately
- Discovery populations that sample a breed widely and include all breeds in which the prediction equation is to be used
- A large enough validation population so that the standard error of the estimated accuracy of the prediction equation is small
- A validation population that represents the cattle where it will be applied
- A chip with >300,000 SNPs
- Better methods to deal with causal genes with rare alleles
- Better understanding of the genetic architecture of the traits we work on so that the method of estimating the prediction equation can reflect this architecture.

Implementation of genomic selection

The purpose of the prediction equation is to apply it to selection candidates in order to more accurately estimate their breeding value or calculate their EPD. When the prediction equation is applied to the SNP genotypes of selection candidates it generates an estimate of their breeding values which is sometimes called a molecular breeding value or MBV. However, typically these selection candidates have other sources of information on their breeding value which are used to calculate a traditional EPD. It is desirable to combine the MBV and the traditional EPD to give a better estimate of breeding value than either of them alone. Some people are calling this combined value a genomic EBV or genomic EPD (GEPD). However, in reality it is just an EPD calculated by including a new source of information (i.e. SNP genotypes). This view is important because it emphasises that the SNP genotypes are only valuable for selecting breeding stock, to the extent that they increase the accuracy of the EPD.

Thus the implementation of genomic selection might involve the following steps: A cattle owner takes a tissue sample from each selection candidate and submits it to a company that provides a DNA service. The company genotypes each animal and applies its prediction equation to generate a MBV for each animal. The MBV is transmitted to the genetic evaluation centre that calculates EPDs and they combine it with the traditional data to generate GEPDs. In order to use the MBVs the genetic evaluation centre needs to know the genetic correlation between the MBV and the trait that is being predicted.

This can be estimated from the validation population but not so easily from the discovery population because it is biased upwards in the discovery sample.

This is the process planned for beef cattle in Australia and, I believe, in USA and Canada. It is not, in my opinion, the ideal process nor the process planned for dairy cattle in the three countries. One disadvantage is that there are likely to be many MBVs. Each company will have different prediction equations based on different SNPs, each company will update their SNP panel regularly and change their prediction equation regularly and there will be different MBV formulae for different breeds. As a result, for some trait, there may be 60 MBVs that will be used by different cattle owners over the next 6 years. A genetic evaluation centre using this data will need a genetic correlation matrix among all these 60 MBVs and the trait itself. This correlation matrix will be very difficult to estimate and to maintain up to date and the risk of errors is high. (e.g. Will DNA companies remember to tell all genetic evaluation centres every time they change a prediction equation and supply the new MBV for lots of old cattle so that genetic covariances can be re-estimated?). Secondly, the commercial phenotypes collected on the cattle that are genotyped as part of the DNA service will not be automatically available to update the prediction equations.

The dairy industry has gone down a different path. Raw genotypes will be submitted to the genetic evaluation centre. These will be used to calculate a prediction equation across all data. As more data accumulates through the commercial use of DNA testing it will be added to the database and used to update the prediction equation. Raw genotypes are stable data; they do not change when the prediction equation is changed unlike the MBVs calculated from them. Therefore storing the raw genotypes is much more logical.

However, there are other disadvantages with the dairy industry model. In USA only the AI companies that participated in the USDA experiment have the right to use the prediction equation on bulls. In other countries there are similar restrictions. In all cases these occur because companies believe they have an advantage over their competitors that they are unwilling to sacrifice. Since nearly everyone is investing in the same technology, in the near future nobody will have much advantage. The world's dairy farmers benefitted greatly from the open system of exchange of information that has prevailed for traditional genetic evaluation based on pedigrees and phenotypes. They would also be best served by a similar approach to genetic evaluations that include genomic data. So, I believe, would beef producers.

We are currently in the first generation of genomic EPDs and EBVs. In years to come I predict that many animals will have DNA genotypes and the methods used to calculate EPDs will change drastically from today's methods to ones driven by the effects of genes. This will lead to more accurate EPDs but is only possible if the genetic evaluation centre has access to the raw genotype data.

References

Goddard, M.E. (2009) *Genomic selection: prediction of accuracy and maximisation of long-term response.* Genetica, **Epub Aug 14 2008**.

Hayes BJ, Visscher PM and Goddard ME (2009) Increased accuracy of artificial selection by using the realised relationship matrix. Genetics Research 91: 47-60.

Meuwissen, T. H. E., Hayes, B. J. & Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157**, 1819–1829.

de Roos, A.P.W., et al. (2008), *Linkage disequilibrium and persistence of phase in Holstein-Friesian, Jersey and Angus cattle.* Genetics, **179**(3): p. 1503-1512.

Visscher, P.M. (2008) *Sizing up human height variation.* Nature Genetics, **40**(5): p. 489-490.

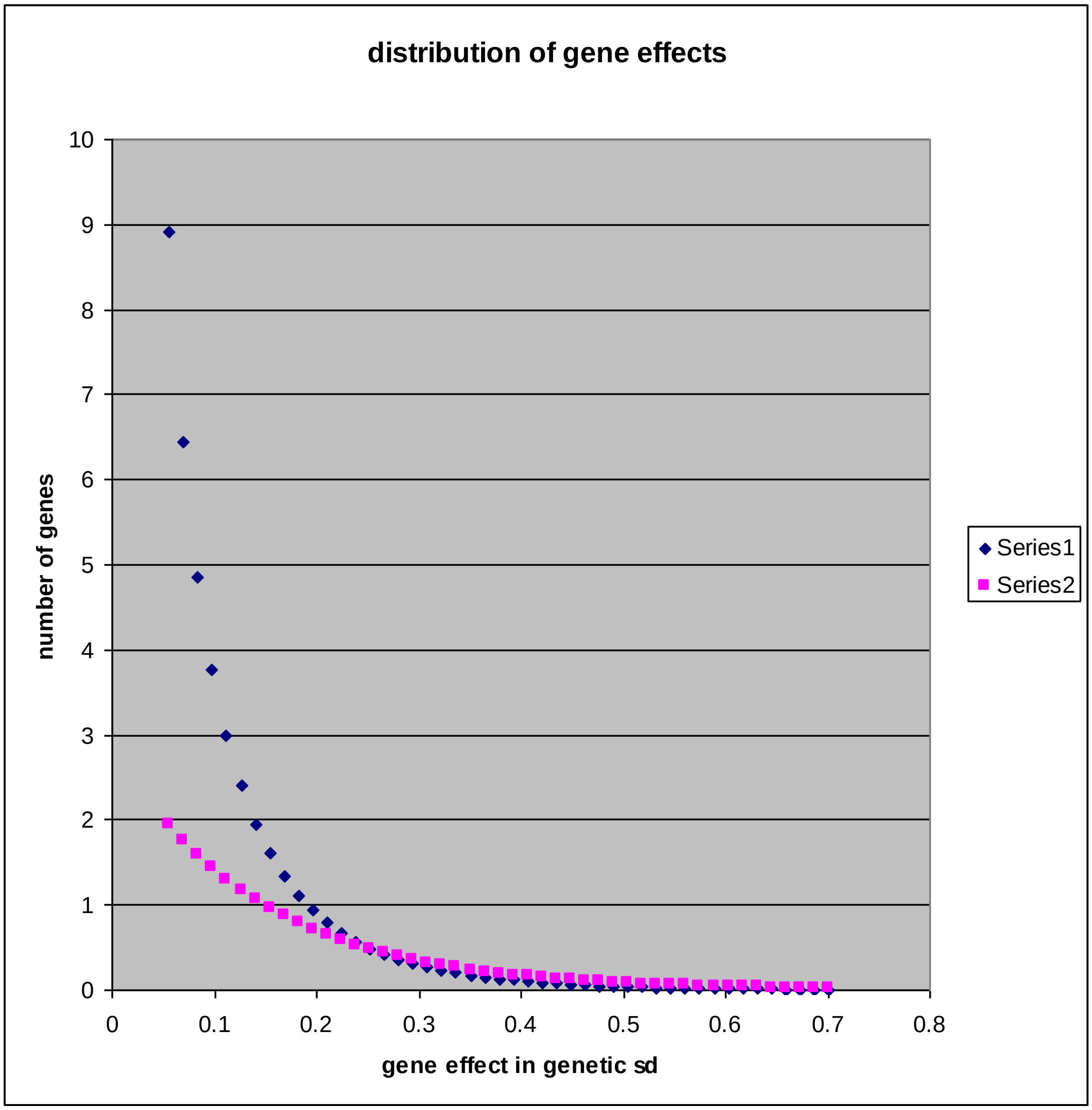

Figure 1. A model of the distribution of gene effects.
Series 1 = the number of genes of a given effect size that are heterozygous in the average animal after selection where the selection coefficient is proportional to the effect of the gene.
Series 2 = the number of new mutations of each size added per 1000 generations to the average animal. This distribution is assumed to be exponential.

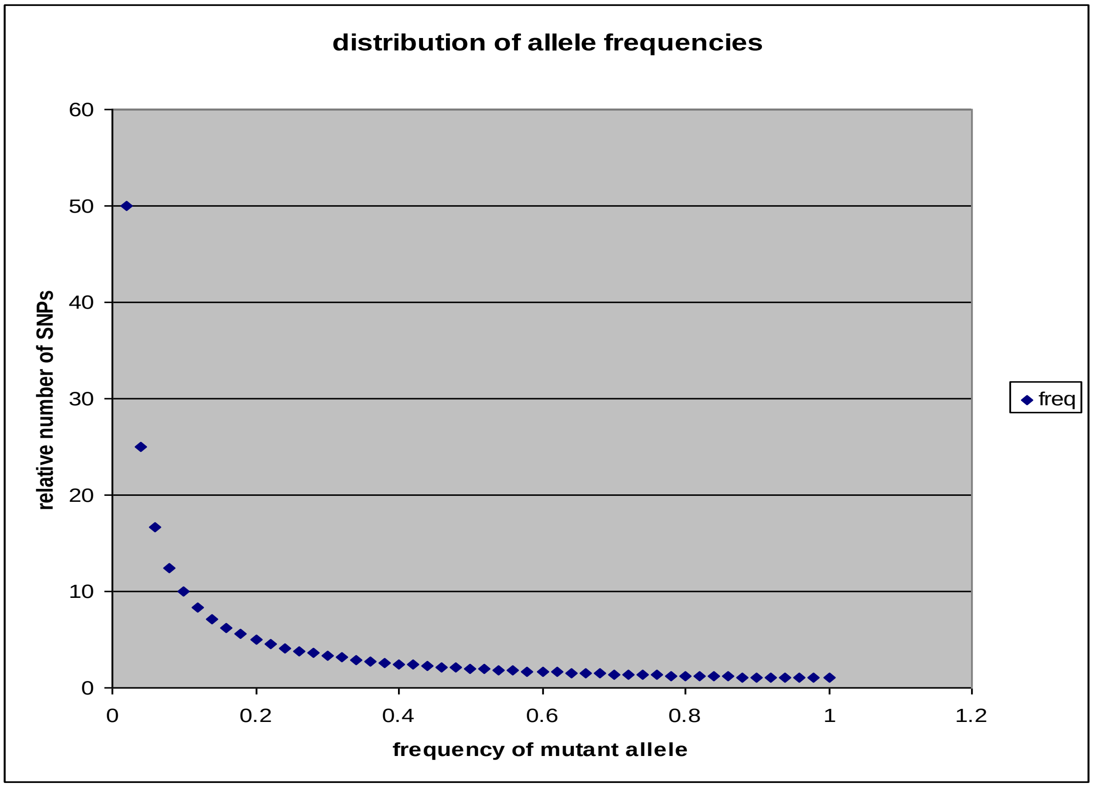

Figure 2. Distribution of frequency of mutant allele across many SNPs. This assumes that effective population size remains constant.

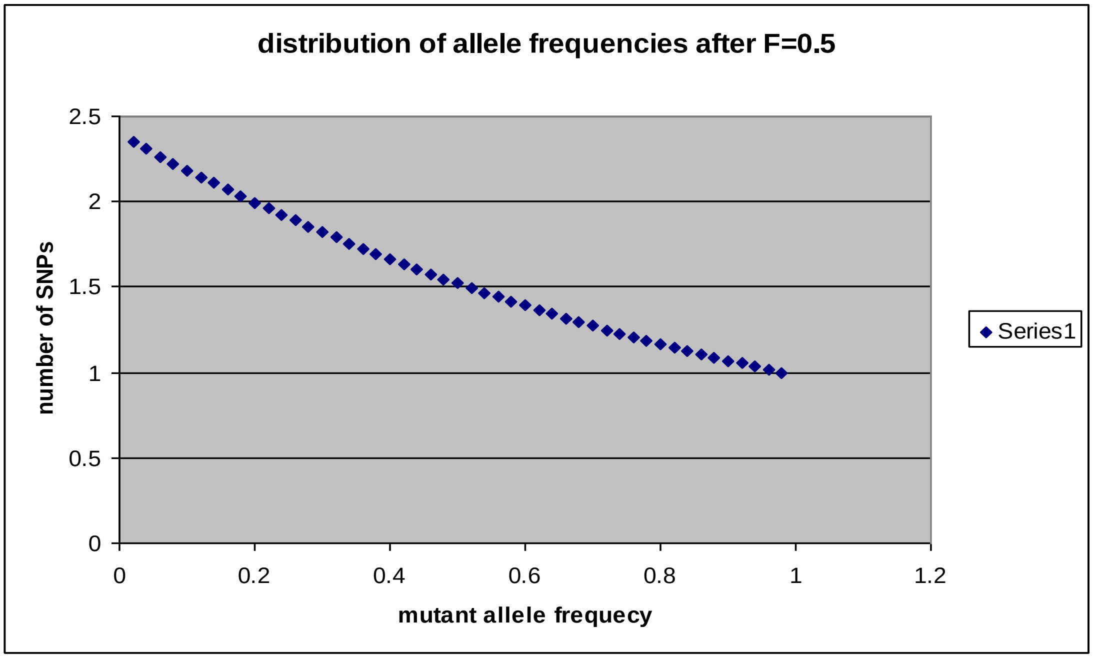

Figure 3. Distribution of mutant allele frequencies across SNPs when inbreeding has reached F=0.5 after starting with the distribution in figure 2. Not shown in this figure are the 131 SNPs where the mutant allele has been lost and the 16 SNPs where it has been fixed.

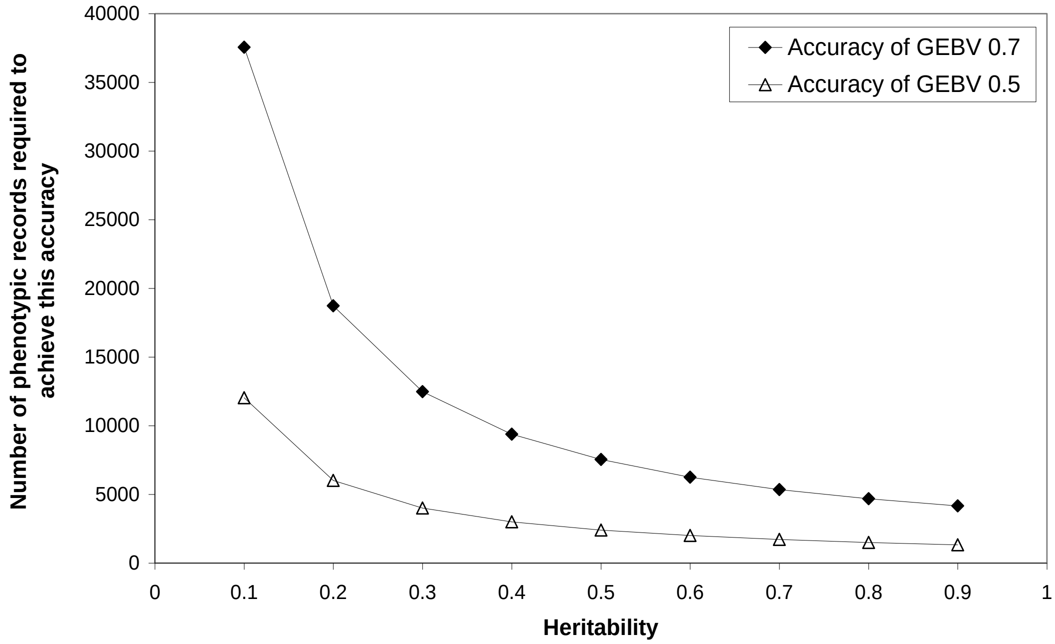

Figure 4. The number of records needed for GEBVs with accuracy of 0.7 or 0.5 in a population with Ne=100.

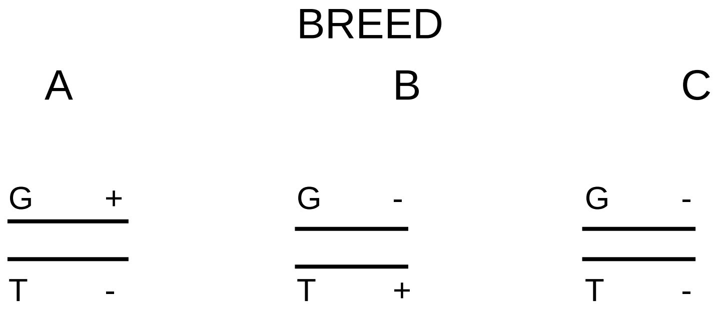

Figure 5. Linkage phase can vary between breeds. In breed A the SNP allele G will be inherited along with the ‘+’ allele for a trait and the ’T’ allele with the ‘-‘ allele for the trait. In breed B this pattern is reversed and in breed C the trait gene is not polymorphic.