GENOMIC SELECTION BASED ON DENSE GENOTYPES INFERRED FROM SPARSE GENOTYPES

M. E. Goddard[1,2] and B.J. Hayes[1]

[1] Department of Primary Industries and [2] University of Melbourne

SUMMARY

Genomic breeding values (GEBVs) predicted from dense marker panels are now being used in the dairy industry for bull selection. However the cost of genotyping these dense marker panels is too high to justify genomic selection for dairy cows, beef bulls and rams for meat and wool production. However if potential selection candidates were genotyped for a standard panel of less expensive sparse single nucleotide polymorphisms (SNPs) and key ancestors within their breed were genotyped for a dense panel of markers, it would be possible to trace the chromosome segments in the selection candidates back to the key ancestors and thus infer their genotypes at all markers assayed on the key ancestors. In this way genomic selection could be practiced using a large number of markers while the cost of genotyping was kept low by the widespread use of a standard SNP chip.

INTRODUCTION

Despite great advances in molecular genetics in recent years, we have identified very few of the genes that cause variation in economically important traits. These genes are known as quantitative trait loci or QTL, because most of the traits of importance are quantitative traits controlled by many genes and by environmental factors. An important recent advance has been the availability of panels of markers consisting of 10,000s to 1,000,000s of single nucleotide polymorphism (SNP) markers. This has made practicable the use of genomic selection as described by Meuwissen *et al.* (2001). Genomic selection refers to the use of a dense panel of genetic markers covering the whole genome to estimate the breeding value of selection candidates (Meuwissen et al 2001). Because the markers are closely spaced, a QTL located anywhere in the genome will be in linkage disequilibrium with at least one marker. Using the Illumina panel of approximately 50,000 SNPs, VanRaden *et al.* (2009) could estimate the breeding value for milk production traits of Holstein bull calves with an accuracy of approximately 0.8. As a result, most developed countries with large dairy industries are now implementing genomic evaluation to calculate estimated breeding values (EBVs). It is anticipated that this trend will spread to beef cattle and to other species because genomic evaluations potentially allow animals to be selected for any trait at birth with an accuracy approaching that of a progeny test. Therefore genomic selection should lead to great increases in the rate of genetic gain in livestock (Schaeffer 2006).

However, there are at least two problems with the implementation of genomic selection. Firstly, the ‘SNP chips’ used for genotyping SNPs are still too expensive (~$A400) to use on most animals. Secondly, there are likely to be many new SNP chips introduced to the market and many other markers, that are not on any SNP chip, discovered to be useful for selection for some trait. To estimate the effects of SNPs on a trait requires a large dataset of animals that have been measured for the trait and genotyped for the markers (a reference population). Therefore, each new panel of SNPs will have to be genotyped across a reference population of tens of thousands of animals. Even then, many animals will have been genotyped with the old SNP chip and will not be able to benefit from the new SNPs. It could even be necessary to genotype animals with multiple SNP chips to obtain EBVs with the maximum accuracy.

In this paper we outline an alternative approach that uses a standard SNP chip of moderate size and cost and infers the genotypes of an animal at all known markers so that they can be used for genomic selection. Firstly, we consider the number of SNPs needed for genomic selection, and then we describe the imputation of SNP genotypes at ungenotyped SNPs.

NUMBER OF MARKERS NEEDED FOR GENOMIC SELECTION

There are two approaches to genomic selection that can be distinguished. In the first, the same panel of markers is used to genotype selection candidates as was used to discover which markers are useful for predicting breeding value. For instance, the Illumina 50k panel might be used to genotype a reference population and from this a prediction equation is derived that predicts breeding value from 50,000 SNP genotypes. Selection candidates are then genotyped for the same 50,000 SNPs and the prediction equation used. In the second approach, an experiment genotypes a reference population for 50,000 SNPs but the data is used to find a subset of the SNPs that can be used to predict breeding value. The selection candidates are genotyped for only a subset of markers. This raises two questions: How many markers are needed in the discovery experiment and how many of these are needed in the final prediction equation?

In the discovery experiment we do not know where the QTL are so we need to cover the whole genome. The density of SNPs needed depends on the extent of LD. The extent of LD in turn depends on the effective population size (N_e). In a breed with a small N_e, such as Holstein (N_e<100), LD extents over considerable distances and so 50,000 markers may be enough. However, if many breeds are used, the LD is only consistent between breeds of *Bos taurus* cattle if markers are <10kb apart (de Roos et al 2008), implying the need for >300,000 markers. Even for a particular breed, the number of markers needed is not an absolute number because the accuracy of predicting breeding value increases slowly as the number of markers increases.

The number of markers needed to predict breeding value in selection candidates should depend on the number of QTL affecting the trait. If there were only a small number of QTL one would expect that a few SNPs near each QTL would be enough. Unfortunately we have not identified the QTL so we cannot answer this question directly. The methods that have been tried for genomic selection vary in the assumptions they make about the number of QTL. The method called BLUP by Meuwissen et al (2001) assumes that all markers have an effect and that these effects are drawn from a single normal distribution. By contrast a method called BayesB assumes that only a fraction of the markers are needed and the others have zero effect. When compared on real data these methods do almost equally well suggesting that there are a large number of QTL affecting most traits. However, it may be possible to choose a small panel of markers that does almost as well as the full 50,000 SNPs. Even a slight loss of accuracy is undesirable but might be acceptable if the cost of genotyping the small panel was much less than the cost of the 50,000 SNP panel. For instance, a panel of 384 SNPs might give acceptable accuracy for one trait but, since many traits are important to most breeders, this still implies a 'small' panel of >1500 SNPs.

The accuracy of genomic selection improves as the number of animals in the reference population (animals with both genotypes and phenotypes) increases, and reference populations in the order of 10s of thousands of animals are required to subsequently predict GEBV with accuracy say greater than 0.7. The only practical way to obtain a very large reference population is to include the animals on which the SNP panel is used commercially. For instance, animals that are selection candidates may, in the future, have phenotypes recorded and could be added to the reference population so that the effects of SNPs on traits recorded could be continually re-estimated and updated. This is analogous to updating the EBV of a sire as his number of recorded offspring increases. However, if the selection candidates are genotyped for only a small panel, it is only the effects of these SNPs that can be re-estimated. Conversely, if the selection candidates had been genotyped for all SNPs, the choice of the best SNPs could be continually improved.

Therefore, we would like to use a small panel of SNPs to minimise the cost of genotyping, but we would like to use all available SNPs to maximise the accuracy of estimated breeding values.

The next section describes a proposal to achieve this. The method put forward has been described by Goddard and Hayes (2008), Goddard (2008) and Habier *et al.* (2009).

INFERRING SNP GENOTYPES FROM A STANDARD PANEL OF SNPS

In most breeds of livestock Ne is quite small (100-200). This means that all the animals of the breed trace back to the equivalent of 100-200 ancestors in each generation. In practice, we find that there are a small number of ancestors that contribute disproportionally to the modern breed. For instance, in Holsteins there are 6 sires who each contribute over 12% of the genes in modern cattle. These 'key ancestors' are only a few generations removed from the current population. Consequently, modern animals inherit large chromosome segments from key ancestors without any recombination. If a key ancestor is t generations removed from a modern animal, the average length of a chromosome segment inherited from the key ancestor is 1/(t+1) M. For instance, if the key ancestor is 3 generation ago, the average chromosome segment inherited is 25 cM. If this segment contained ~7 SNPs, it should be possible to trace the segment from the modern animal back to the key ancestor. If we know the alleles that key ancestor carried at other markers on this 25 cM segment, then we know that the modern animal must have inherited these alleles along with the rest of the segment. Thus we can infer the genotype of a modern animal at all markers that have been genotyped on the key ancestors provided we have genotyped enough markers on the modern animal to trace his or her chromosome segments back to a key ancestor. There is an additional requirement – the genotypes of the key ancestors must be phased so that we know which alleles are carried on the same chromosome. This can be achieved by a linkage analysis using known relatives or by a non-pedigree method relying on LD between markers such as fastPhase (Scheet and Stephens 2006).

KEY ANCESTORS

What is the smallest set of animals that can be genotyped with the dense SNP panel? We have called these animals key ancestors to indicate a set of animals from which most of the genes in modern animals are descended. A method to find the optimum set of key ancestors is now described. Suppose that you wanted to predict the mean breeding value (m) of the modern population from knowledge of the breeding values of a set of key ancestors (**g**). A regression equation could be derived, m = **b'g,** where $\mathbf{b} = \mathbf{A}^{-1}\mathbf{c}$, **A** = numerator relationship matrix among the key ancestors, **c**= vector of relationships between key ancestors and the modern population. However, m could also be predicted from a weighted average of the breeding values of the key ancestors where the weights are the proportion of genes contributed directly from a key ancestor to the modern population ie genes not contributed through another key ancestor. In fact, **b'g** is exactly this weighted average and so the sum of the elements of **b** (**1'b**) is the proportion of genes in the modern population that can be traced to one or other of the key ancestors. Therefore the optimum set of key ancestors is the one that maximizes **1'b.** A good approximation to this optimum can be found by the equivalent of forward regression. That is, at each iteration the animal that would increase **1'b** the most is added to the list of key ancestors.

This choice of key ancestors maximises the proportion of genes in the modern population derived from the key ancestors but it ignores the problem of phasing their genotypes. If phasing is to be carried out using linkage analysis, then the list of key ancestors needs to contain related animals. The ability to trace chromosome segments from modern animals to key ancestors will decrease as the number of generations between them increases (Habier *et al.* 2009). This problem can be overcome in two ways. Firstly, one could use denser markers for the sparse SNP panel. Habier *et al.* (2009) used only one SNP per 10 cM so it is not surprising that recombinations eroded the ability to trace chromosome segments. Secondly, one could continually add animals, that occur in the pedigree between the key ancestors and the modern animals, to the list of densely genotyped animals. Habier *et al.* (2009) found that densely genotyping all parents each generation overcame the problem. However, if denser markers were used to genotype key ancestors it would not be necessary to densely genotype so many animals. For instance in the dairy industry, one might only densely genotype AI sires. Nor is it necessary to genotype these intermediate ancestors for all SNPs in the dense panel used on key ancestors. The intermediate ancestors merely provide a link between the modern animals and the key ancestors. For instance, the key ancestors might be genotyped for all known SNPs whereas the intermediate ancestors are genotyped for 50,000 SNPs.

IN SILICO GENOME SEQUENCING

The ultimate density of markers might be provided by full genome sequencing. As the cost of sequencing drops this will become possible on a list of key ancestors. Then by genotyping modern animals for a panel of sparse markers it would be possible to infer the full genome sequence on each modern animal.

A VISION FOR THE FUTURE

Before this approach to *in silico* sequencing becomes practical, it is likely that the number of SNPs and other polymorphisms available will increase steadily and this will include an increasing number of functional or causal mutations. This could present a considerable problem because SNP panels would rapidly become out of date because they did include recently discovered polymorphisms and the genotypes on older animals would be missing all recent discoveries. The proposal to use a standard sparse SNP panel overcomes this problem. Most animals are genotyped with the sparse panel which stays constant over time. Only the key ancestors need to be genotyped for all newly discovered and useful markers. In this way the cost of genotyping is low but the modern animals receive inferred genotypes for all markers thought to be important enough to genotype on the key ancestors. The same sparse panel could potentially be used for all cattle, all over the world, for many years. This would mean the sparse SNP chip would be manufactured in huge numbers and so should be relatively inexpensive and less costly than SNP chips designed for a limited market even if they genotyped far fewer SNPs.

REFERENCES

De Roos, A. P. W., Hayes, B. J., Spelman, R. and Goddard, M. E. (2008). *Genetics.* **179**:1503.

Goddard, M.E. (2008) In Pinard M-H, Gay C., Pastoret, P-P and Dodet, B. (eds): Animal Genomics for Animal Health. Dev Biol. (Basel), Basel, Karger **132**:383.

Goddard, M.E. and Hayes, B.J. (2008) Artificial selection methods and reagents. Patent application WO/2008/074101.

Habier, D., Fernando, R.L. and Dekkers, J.C.M. (2009) Genomic selection using low-density marker panels. Genetics published ahead of print as 10.1534/genetics.108.100289.

Meuwissen, T. H .E., Hayes, B. J. And Goddard, M. E. (2001) *Genetics* **157**:1819.

Schaeffer, L. R. (2006) *J. Anim. Breed. Genet.* **123**:218.

Scheet, P., and M. A. Stephens (2006) *Am. J. Hum. Genet.* **78**:629.

VanRaden, P. M., Van Tassell, C. P., Wiggans, G. R., Sonstegard, T. S., Schnabel, R. D., Taylor, J. F. and Schenkel, F. S. (2009) *J Dairy Sci* **92**: 16.