## LANDCORP FARMING LECTURE

# Genomic selection for accelerated genetic gain in livestock

B.J. HAYES and M.E. GODDARD[1]

Animal Genetics and Genomics, Primary Industries Research Victoria, Australia
[1]Faculty of Land and Food Resources, University of Melbourne, Australia

## ABSTRACT

The sequencing of a number of the livestock genomes and developments in genotyping technology has made it possible to genotype animals for tens of thousands of markers relatively cheaply. Genomic selection exploits this information to calculate genomic breeding values (GEBVs), as the sum of the effects of small chromosome segments across the genome. Effects of chromosome segments are first estimated in a large population with phenotypic information. In subsequent generations, only marker information is required to calculate GEBVs. Accuracies of GEBVs calculated for animals with marker information only, can be as high as 0.85. Results from analysis of real data from dairy cattle suggest approximately 30,000 single nucleotide polymorphism markers will be required to calculate GEBVs with maximum accuracy. Chromosome segment effects should be estimated from at least 2000 phenotypic records. A trial of genomic selection in the Australian Holstein population, with 384 bulls typed for 9919 markers, achieved accuracies of GEBVs of up to 0.82, despite sub optimal marker density and limited number of phenotypic records. In dairy cattle, maximum genetic gain using genomic selection will be achieved by selection of young bulls for widespread use without progeny test, allowing a halving of the generation interval.

**Keywords:** Genomic selection; single nucleotide polymorphisms; accuracy.

## INTRODUCTION

The idea of using DNA markers to improve the accuracy of predicting breeding values, and therefore the rate of genetic gain, has been around for decades (*e.g.* Smith, 1967; Soller & Beckman, 1983). However, adoption of marker assisted selection (MAS) in livestock improvement has been relatively slow for a number of reasons. For most economic traits there are a large number of genes affecting the trait with any one marker only capturing a limited proportion of the total genetic variance (Shrimpton & Robertson, 1988; Hayes & Goddard, 2001). As a consequence relatively small gains are possible with the limited number of markers that have been available in most livestock species, and the cost of genotyping these markers has been high. Additionally, the complexity of calculating breeding values including marker information has been a barrier to the application of MAS.

Three recent developments are already resulting in a rapidly accelerating adoption of MAS. The sequencing of a number of livestock genomes, including cow, pig and chicken has lead to the discovery of many thousands of DNA markers for these species, in the form of single nucleotide polymorphisms, or SNPs. An SNP is a difference in DNA sequence at the same point in the genome between two animals. Concurrent with the discovery of numerous SNP markers throughout the livestock genomes has been a dramatic reduction in the cost of genotyping, which can be as less than US 1c per SNP per animal.

The third development resulting in increased adoption of MAS was the realisation and demonstration that it is possible to calculate very accurate breeding values from marker data alone using a method called *genomic selection* (Meuwissen *et al*., 2001). Typically in MAS, a limited number of regions of the genome containing genes with large effects on the trait of interest are traced with markers. Breeding values are then calculated using both pedigree and the limited marker information. In this type of MAS only a proportion of the genetic variance is captured by the markers. An alternative, if a dense marker map is available, is to divide the genome into small segments and then simultaneously estimate the effects of all these segments on the trait of interest. In subsequent generations, animals can be genotyped for the markers to determine which chromosome segments they carry, then the estimated effects of the segments the animal carries can be summed up across the whole genome to predict a breeding value. This breeding value is termed a genomic estimated breeding value (GEBV). In this way all the additive genetic variance is captured by the markers, leading to high accuracies of GEBVs. Meuwissen *et al*.

(2001) termed this "Genomic selection". Meuwissen *et al*. (2001) in simulations achieved accuracies of predicting breeding values from markers alone (the correlation between true breeding value and estimated breeding values) of 0.85.

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

1) How many markers are required?
2) How many phenotypic records are required in the initial experiment estimating the effect of chromosome segments?
3) How does genomic selection perform in real data?
4) How should a breeding program be altered to maximise the gains from genomic selection?

## HOW MANY MARKERS ARE REQUIRED FOR GENOMIC SELECTION?

Genomic selection exploits a property of livestock populations called linkage disequilibrium (LD). In order to understand how linkage disequilibrium is generated, consider an ancestral animal many generations ago, with descendants in the current population. Each generation, recombination events break down the ancestor's chromosome, until in the current generation only small regions of chromosome which trace back to the common ancestor remain. These chromosome regions in different animals will be identical if they trace back to the same common ancestor chromosome (Figure 1). They will therefore carry the same marker alleles, as well as the same alleles of any gene on the chromosome segment which affects economic traits. In genomic selection, DNA markers are used to infer which animals carry chromosome segments which trace back to a common ancestor. The number of markers required for genomic selection is determined by the size of chromosome segments which trace back to a common ancestor – the shorter the segments, the more segments must be inferred, and therefore greater the number of markers that are required.

LD between a gene affecting a quantitative trait and one or several markers can be measured by $r^2$, the proportion of variation caused by the alleles at a gene which is explained by the markers. In Figure 2, the average decline of $r^2$ with distance between two SNP markers is given for three different cattle breeds (Australian Holstein, Norwegian Red and Australian Angus).

For genomic selection to be successful, the level of LD between adjacent markers should be $r^2$>=0.2 (based on Meuwissen *et al*., 2001). Figure 2 implies that in dairy cattle, there must be a marker every 100kb (kilo bases) or less to achieve this. 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! For cattle at least, a SNP chip with this many markers is already commercially available (Lien, S. pers. comm.).

**Figure 1:** An ancestor many generations ago (1) leaves descendants (2). Each generation, the ancestor's chromosome is broken down by recombination, until all that remains in the current generation are small conserved segments of the ancestor's chromosome (3). Chromosome segments descended from the same common ancestor in the current generation are identical and therefore carry the same marker alleles and alleles of genes affecting quantitative traits.

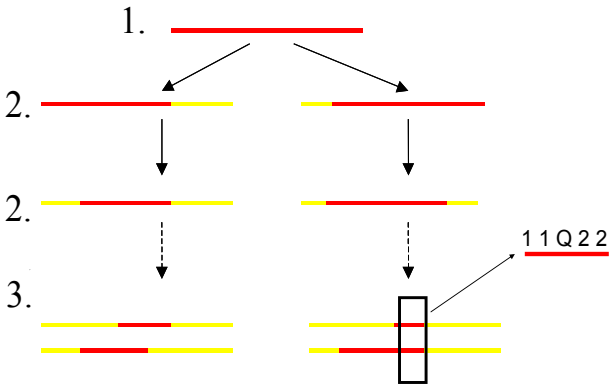
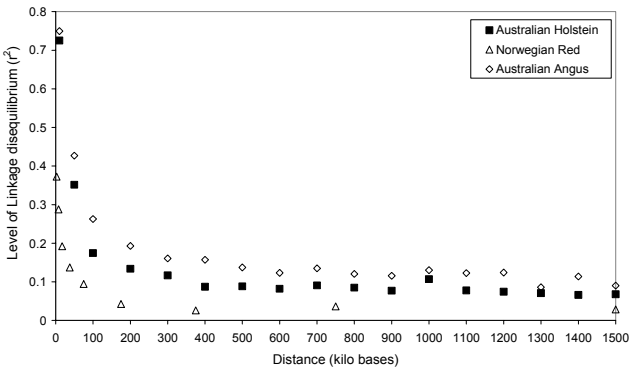



**Figure 2:** Average $r^2$ value according to the distance between the SNP markers. Results are from 10 000 SNPs distributed across the genome genotyped in 384 Holstein cattle or 384 Angus cattle, or 403 SNPs genotyped in 783 Norwegian Red cattle. Norwegian red data kindly supplied by Prof. Sigbjorn Lien, Norwegian University of Life Sciences.



## HOW MANY PHENOTYPIC RECORDS ARE REQUIRED IN THE INITIAL EXPERIMENT ESTIMATING THE EFFECT OF CHROMOSOME SEGMENTS?

So for a particular region of the genome, there may be chromosome segments descended from a number of different ancestors in the population. The accuracy of genomic selection will depend on the number of unique chromosome segments, and the number of phenotypic records per unique chromosome segment, which determines the accuracy of estimating the effect of that segment. One difficulty with genomic selection is that a very large number of chromosome segment effects must be estimated, most likely from a data set where the number of phenotypic observations is less than the number of chromosome segment effects to be estimated. The more phenotypic records available, the more observations there will be per unique chromosome segment, 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 three statistical methods for calculating GEBVs: least squares, a best linear unbiased prediction (BLUP) method assuming equal variances associated to each chromosomal segment, and a Bayesian method assuming a prior distribution of the variance associated with each chromosome segment. The methods were compared with different numbers of phenotypic records. In simulations, the effects of the chromosome segments were estimated in one generation of animals, and the breeding values for the progeny of these animals were predicted based only on the markers which they carried. The results suggested the Bayesian method was superior to the others (Table 1). The results also suggest that in the order of 2000 phenotypic records are required to accurately estimate the chromosome segment effects.

**Table 1:** 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 |
| Bayes B | 0.708 | 0.787 | 0.848 |

## RESULTS FOR GENOMIC SELECTION IN REAL DATA

An experiment was conducted to test genomic selection in the Australian Holstein population. Three hundred and eighty four bulls were selected from the Australian dairy bull population for genotyping. The bulls selected were those with extreme high and extreme low estimated breeding values (EBVs) for the Australian selection index, ASI = (3.8*protein) + (0.9*fat) – (0.048*milk) and the records for the sub-components are based on performance of the bull's daughters (a progeny test). The bulls were genotyped for 9919 SNP genome wide markers.

From the 384 bulls, 300 were chosen at random for the prediction of SNP effects (using a program kindly provided by Prof. Larry Schaeffer). For the remaining 84 bulls, GEBVs were predicted based on their SNP genotypes alone. This was done for protein kg, fat kg, protein %, fat % and milk kg. The GEBVs were correlated with the average performance of their daughters for these bulls, Table 2.

There are a number of reasons why the accuracies could be lower than those achieved in simulations, including inadequate marker density (only 9919 markers were used here), and too few phenotypic records were used to estimate chromosome segment effects accurately. However, the results are promising, particularly for protein kg, and suggest that genomic selection with a larger number of markers and a larger number of phenotypic records used to estimate chromosome segment effects could give very accurate GEBVs.

**Table 2**: Correlations of genomic estimated breeding values predicted using marker genotypes only and daughter yield deviations for 84 Australian Holstein Dairy bulls. A different set of three hundred bulls with progeny test information were used to predict the effects of markers.

| **Trait** | **Accuracy** |
|---|---|
| Protein (kg) | 0.82 |
| Fat (kg) | 0.70 |
| Protein (%) | 0.73 |
| Fat (%) | 0.62 |
| Milk (kg) | 0.70 |

## OPTIMISING BREEDING PROGRAMS WITH GENOMIC SELECTION AND IMPLEMENTATION

The accuracy of traditional estimated breeding values increases as an animal ages and it and its relatives acquire phenotypic data. However,

animals can be typed for DNA markers at any age and so the gain in accuracy of EBV due to adding the marker data should be greatest at young ages. Consequently, if selection is optimized, marker data should lead to a decrease in generation length. In dairy cattle, MAS leads to greater gains if MAS selection of yearling bulls and cows is practiced than if a traditional progeny testing system is adhered to (Spelman *et al*., 1999). This is particularly true for genomic selection, where the accuracy of the GEBV is as high as the accuracy of an estimated breeding value following a progeny test. Potentially, genomic selection could lead to a doubling of the rate of genetic gain through selection and breeding from bulls at 2 years of age rather than 4 years of age.

However, the first use of genomic selection in the dairy industry is likely to be the selection of young bulls to progeny test. Once the industry is confident that the technology works, other uses will quickly follow, including:

- Selection of heifers as bull dams based on markers
- Selection of young bulls for widespread use without progeny test
- Selection of young bulls as bull sires
- Selection of replacement heifers

In conclusion, the sequencing of a number of the livestock genomes and developments in genotyping technology means animals can be genotyped for tens of thousands of markers relatively cheaply. Genomic selection exploits this information to calculate genomic breeding values, as the sum of the effects of small chromosome segments. Accuracies of GEBVs calculated in this way, for animals with marker information only, can be as high as 0.85. Results from real data in dairy and beef cattle suggest that approximately 30,000 SNPs will be required to calculate GEBVs with maximum accuracy, and chromosome segment effects should be estimated from at least 2000 phenotypic records. When we tested genomic selection in the Australian Holstein population, with 384 bulls typed for 9919 markers, we achieved accuracies of GEBVs of up to 0.82, despite the sub-optimal marker density and limited number of phenotypic records. In dairy cattle, maximum genetic gain using genomic selection will be achieved by selection of young bulls for widespread use without progeny test, allowing a halving of the generation interval. Selection of heifers as bull dams and replacement heifers could also be considered if the cost of genotyping becomes sufficiently low.

## ACKNOWLEDGEMENTS

The program for calculating GEBVs was kindly provided by Prof Larry Schaeffer, University of Guelph. Norwegian Red SNP data kindly supplied by Prof. Sigbjorn Lien, Norwegian University of Life Sciences.

## REFERENCES

Hayes, B.J.; Goddard, M.E. 2001: The distribution of the effects of genes affecting quantitative traits in livestock. *Genetics Selection Evolution* **33**: 209-229.

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

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.

Smith, C. 1967: Improvement of metric traits through specific genetic loci. *Animal Production* **9**: 349-358.

Soller, M.; Beckmann, J.S. 1983: Genetic polymorphism in variety identification and genetic improvement. *Theoretical Applied Genetics* **67**: 25-33.

Spelman, R.J.; Garrick, D.J.; van Arendonk, J.A.M. 1999: Utilization of genetic variation by marker assisted selection in commercial dairy populations. *Livestock Production Science* **59**: 51-60.