GENOME WIDE SELECTION: ISSUES AND IMPLICATIONS

B. Tier[1], R. Crump[1], G. Moser, J. Sölkner[2], P.C.Thomson[2], A. Woolaston[1], H.W. Raadsma[2]

Co-operative Research Centre for Innovative Dairy Products – CRC IDP.
[1]Animal Genetics and Breeding Unit[*], University of New England, Armidale, NSW, Australia,
[2]ReproGen – Centre for Advanced Technologies in Animal Genetics and Reproduction, Faculty of Veterinary Science, University of Sydney, Camden, NSW, Australia,

SUMMARY

Single nucleotide polymorphic (SNP) chips enable us to account for variation within genomes very well. It is possible to associate this variation with variation in phenotypes and so predict genetic merit of young individuals better than ancestral indexes. This has significant implications for livestock industries as to accuracy and timing of selection decisions, and how resources are allocated to maximize the returns from investments in genotypic and phenotypic data collection. High density genotyping platforms will exacerbate the problem of the joint analysis of individuals with heterogeneous amounts of genotypic information.

INTRODUCTION

In the 1970s, best linear unbiased prediction (BLUP, Henderson 1973) methods were first applied to predict the genetic merit – estimated breeding values (EBVs) – of livestock. Now BLUP is the standard method for genetic evaluation of livestock. BLUP uses performance records to predict EBVs for all traits and individuals in a population. Genetic parameters are used to weight the records of all traits appropriately. The pedigree – through the numerator relationship matrix (**A**) – defines the expected correlations between individuals resulting from genes that are identical by descent. **A** provides the appropriate weightings in the analysis of relatives' performance records.

The first application of molecular biology to animal breeding occurred in the late 1980s, with the idea that some loci in the genome could be used as markers for other nearby loci, affecting quantitative traits (quantitative trait loci or QTL, Fernando and Grossman 1989). BLUP models were extended to incorporate marked QTL, in addition to the polygenic component for all other loci. Numerous methods were developed to detect associations between markers and phenotypes. These methods used marker information to specify part of the genetic relationships among animals with more precision than the pedigree – initially on a within-family basis, but relationships are now inferred more widely, including between base animals (Meuwissen and Goddard 2000). Considerable resources were invested in the search for QTL. At first, investigations were confined to relatively few markers per chromosome in highly structured experiments such as the grand-daughter design but now methods are available to analyse large numbers of markers from field data (e.g. QXPAK, Perez-Encisco and Misztal 2002). Regions containing QTL for many traits have been found for most livestock species, but relatively few genes and even fewer causative mutations and their effects on the phenotypes have been identified. Nevertheless, there is a small but growing number of gene or QTL tests in the marketplace for most livestock species. An animal's EBV which combines marker and polygenic effects should be more accurate than an EBV predicted without markers.

[*] AGBU is a joint venture of NSW Department of Primary Industry and the University of New England

We have recently entered a completely new era in genotyping where it is now cheaper to collect genotypic information, per unit, than phenotypic information. For most livestock species it is, or soon will be, possible to genotype individuals for thousands of single nucleotide polymorphisms (SNP) simultaneously. Furthermore we will see rapid growth in the numbers of SNP available per test. Currently this is still too expensive for routine screening of individuals, but it is not too expensive for research programs and the cost per SNP continues to decrease rapidly. High density SNP genotyping is leading to a new paradigm in animal breeding known as genome-wide selection (GWS) and is also stimulating a search for quantitative trait nucleotides (QTN) – SNP that cause phenotypic variation –, as described for dairy cattle (Raadsma *et al*. 2007). This paper explores some of the issues and implications of this new paradigm for livestock breeding.

MOLECULAR EVALUATION

The task of genetic evaluation in the new paradigm is still to associate individuals' phenotypic variation with their genotypic variation. However the moderate cost of genotyping means that, at least in the early experimental stages, there will be considerably more SNP genotypes per individual than individuals genotyped and consequently considerably more genotypic than phenotypic data on those individuals. Hence the SNP data could explain much more, even all, of the phenotypic variation than the proportion expected by the heritability. To avoid such over-parameterization we need to define one or more appropriate models for the analysis of these data and then choose among them.

A number of approaches have been developed to model the phenotypic data (**y**) as a function (*f*) of the genotypic data (**g**): $\mathbf{y}=\mathbf{b}+f(\mathbf{g})+\mathbf{e}$, where **b** and **e** are vectors of systematic effects and residuals and $f(\mathbf{g})$ can be considered a molecular breeding value (MBV). All methods share the idea of limiting consideration of the variation in **g** in some way. This is done by either transforming the inherent variation (Moser *et al.* 2007) or by choosing among subsets of possible variables in direct (Gianola *et al.* 2006, Woolaston 2007) or indirect (Crump *et al.* 2007) ways. However, the key question with genetic evaluation is how well we predict the genetic merit of future animals.

Without any marker data BLUP EBVs calculated as mid-parent values can have accuracies as high as 0.71. To achieve an accuracy of 0.5 for a mid-parent EBV requires a sire with an accuracy of 0.8 and a dam with 0.6, or similar. Inclusion of single markers can increase the accuracy of EBVs when the effect of the genotypes is known with high precision. However, when a number of markers with effects estimated with moderate precision are used, the resulting EBVs can be less accurate than those from a standard BLUP analysis.

Ideally knowing **g**, and $f(\mathbf{g})$, will provide more accurate knowledge of an embryo's EBV. This is one of the early promises that accompanied the revolution in molecular biology. The question is "will it come true?" and the corollary is that, if we can know $f(\mathbf{g})$ completely, then performance need no longer be recorded. Consequently, it would be unnecessary to progeny-test sires leading to shorter generation intervals, higher accuracy of selection and faster genetic progress (Schaeffer 2006).

There are generally only 2 alleles at any SNP. Tests count the number of copies of 1 of the 2 alleles expected at that locus. The thousands of SNP in a single test may span the genome, but some parts of the genome may only be sparsely represented. Accompanying each test are overall quality results for each SNP and individual in the tested group. Bi-allelic tests for SNP with 3 or 4 alleles are currently unreliable, but as there are relatively few such SNP they can be ignored. Using the appropriate SNP map, haplotypes can be determined over small sections of chromosome and linkage disequilibrium (LD) between proximal SNP used to assign 'failures' to genotypes. Pedigree

information can enhance this process (Zenger *et al.,* 2007). Thus SNP technology will enable us to know **g** with high accuracy.

Most of the approaches for estimating MBVs discussed above (Crump *et al*. 2007, Moser *et al*. 2007) were applied to data consisting of 1546 dairy bulls genotyped for 15036 SNP using EBVs for some traits as performance records. These data were partitioned into two sets. A training set used to develop prediction function and a test set used to evaluate the prediction function. High correlations (>0.9) between the EBVs and MBVs were found for animals in the training set across methods, traits and replicates. However, smaller correlations ranging from 0.5-0.85 were found between the EBVs and MBVs for individuals in the test set. Variation in these correlations could be ascribed mostly to the traits, with the correlations being worst for a lowly heritable trait. In another study, phenotypes for a small set of individuals (n=1600) for a modestly heritable trait (0.3) were simulated. Principal component analysis (Woolaston *pers. comm.* 2007) was used to predict MBVs which were compared with the simulated BVs. High correlations were found for individuals in both analysed set (with records, r=0.98) and predicted sets (r=0.96). These are similar to the correlations found by Gianola *et al.* (2006) in a simulated set of data. However, one expects high correlations when the same models are used for both simulation and analysis.

None of these models predicted the performance of the dairy sires in the test data set perfectly. However, the data were EBVs and not true breeding values. We may not know **g** very well because there were insufficient variable SNP in parts of the genome. Specifying **g** as haplotypes rather than SNP may provide a better representation of the genotypic variation. Alternatively it may be difficult to specify a single $f(.)$ for the population as there could be variation in gene action and/or incomplete LD across the population. One source of variation in gene action could be interactions between active genes and other effects, either genetic or environmental. More complex models that account for such potential problems may improve accuracy of prediction, but how long will $f(\mathbf{g})$ hold up without subsequent recalibration and will it be a good predictor of performance for a correlated trait – such as the same trait in a different environment, or a different breed? Currently we have insufficient data to answer these questions; however, while there is uncertainty about $f(\mathbf{g})$, regular recording of performance is required until we do.

INDUSTRY IMPLICATIONS

The power and expense of genotyping technology, together with the inability of these models to predict MBVs precisely, suggests that a series of centralized recording populations be established in representative environments for most species. These populations can be intensively recorded for routine and important traits that are expensive to measure (e.g. meat tenderness). Parts of these populations can also be intensively genotyped. SNP chips for routine use containing small subsets of the large SNP chips can be developed so that all individuals in these representative populations can be genotyped for some SNP. Some of these lower density chips could be used to genotype individuals in commercial herds or flocks, where routine recording of traits that are relatively cheap to measure should continue. Most of the genotypes of progeny of parents with high density genotypic data could be inferred from the results of such low density chips. Establishing these centralized populations will demonstrate how useful $f(\mathbf{g})$ estimated in one population is in another. Centralized populations will lead to fewer people making important breeding decisions in the more extensive livestock industries. This may be a problem for more extensive industries, where there are now many decision makers breeding for a variety of goals. The number of sub-populations and the allocation of individuals to be

recorded and/or genotyped is an issue that needs research to ensure that any resources are allocated efficiently.

Genotypic data have the potential to be used in industry to select individuals for environments or markets (such as those based on meat quality), or feeding regimes for individuals, as in the dairy industry. Similarly, genotypic data from industry may contribute to estimating $f(\mathbf{g})$, if collected in large amounts together with sufficient management information.

Genotypic data on the SNP chip scale will enhance the search for QTN, but it should not be a major motivation for livestock industries. The ability to predict MBVs without understanding the precise effect of each genotype at every locus relieves us of the problem of estimating the effect of a number of markers with sufficiently high accuracy to provide more accurate EBVs than those estimated in a simple polygenic BLUP evaluation. However, the problem relating to heterogeneity of genotypic data will be exacerbated. We are now in a position where some individuals have been genotyped for a small number of markers, some have been genotyped for thousands of SNP but most have no genotypic data at all. Early approaches applying markers concentrated on an extra analysis for genotyped individuals only. High density SNP chips have greatly magnified this problem and it will be further exacerbated as new SNP chips are developed. We can now estimate MBVs for genotyped individuals and/or EBVs for all individuals. We need to be able to undertake both models in a single analysis so that the maximum amount of information is used to estimate molecular assisted breeding values (or MABVs). Such heterogeneous methods are required if we are to get the most out of our genotyping investments.

CONCLUSION

A start has been made to using high density SNP marker information to estimate genetic merit. It shows early promise but its expense requires careful consideration of which individuals should be genotyped and phenotyped. It may cause some centralization in the breeding structure of extensively raised livestock. To maximize the return on investment it is important that methods for the joint analysis of all individuals – genotyped or not – be developed.

REFERENCES

Crump, R. *et al* (2007) *Proc. Assoc. Adv. Anim. Breed. Genet.***17**:304.
Fernando R.L. andGrossman M. (1989) *Genet. Sel. Evol.* **21**:467.
Gianola, D., Fernando R.L. and Stella A. (2006) *Genetics* **173**:1761.
Henderson C.R. (1973) In *Proc. Anim. Breed. Genet. Symp. in Honor of Dr Jay L. Lush* pp. 10-41. ASAS and ADSA, Champaign Illinois.
Meuwissen, T.H.E. and Goddard M.E. (2000) *Genet. Sel. Evol.* **33**:605.
Moser, G. *et al*. (2007) *Proc. Assoc. Adv. Anim. Breed. Genet.* **17**: 227.
Perez-Encisco M. and Misztal I. (2004) *Bioinformatics* **20**:2792.
Raadsma,H.W. *et al.* (2007) *Proc. Assoc. Adv. Anim. Breed. Genet.* **17**:231.
Schaeffer L.R.S. (2006) *J. Anim. Breed. Genet.* **123**:218.
Woolaston A. (2007) PhD Thesis, University of New England.
Zenger, K., *et al*. (2007) *Proc. Assoc. Adv. Anim. Breed. Genet.* **17**:123.