EDITORIAL

Genomic selection : marker assisted selection on a genome wide scale

Genomic selection (GS) has become a very intense field of research during recent years. GS may be defined as the simultaneous selection for many (tens or hundreds of thousands of) markers, which cover the entire genome in a dense manner so that all genes are expected to be in linkage disequilibrium with at least some of the markers. In a sense, GS is marker assisted selection on a genome wide scale. Methodology for GS was first presented by Meuwissen *et al.* (*Genetics* **157**:1819–1829), but the term GS was actually first introduced by Haley and Visscher at Armidale *WCGALP* in 1998 (as pointed by Sijne van der Beek, personal communication). Fortunately, Haley and Visscher used the term in exactly the same way as we did in the 2001 paper. GS sounded like a crazy idea in 2001, when a major aim in marker assisted selection research was to reduce the costs of genotyping. However, the development in the technology for single nucleotide polymorphism (SNP) genotyping has been tremendous, and in humans 500 000 SNPs are genotyped routinely nowadays. It is this incredible development of the genotyping technology that makes GS feasible.

This issue of *Journal of Animal Breeding and Genetics* contains a series of papers (the references without journal name) presenting the state of the art on genomic selection. I will not discuss every paper separately here, but will consider some issues that are across the papers. One issue that is mentioned by Goddard and Hayes and was also described this year by Guillaume *et al.* (Dublin *EAAP*) is the equivalence between the BLUP-GS model, where the marker effects are estimated by BLUP (see Kolbehdari *et al.* for details), and the traditional BLUP model, where the usual pedigree based relationship matrix is replaced by a relationship matrix estimated by the markers. Basically, if $\mathbf{X}$ denotes the design matrix of the marker effects, then $\mathbf{XX}'$ is used as a marker estimated relationship matrix. Replacing the pedigree relationship matrix by a marker estimated relationship matrix has been tested before in non-GS marker assisted selection schemes and the results may be summarized as (e.g. Villanueva *et al.*, *J. Anim. Sci.* **83**:1747–1752): (1) the extra genetic gains are quite small, (2) they decrease with increasing genome size and (3) the pedigree should be used in addition to the markers to infer the relationship matrix. Result (3) suggests that we should perhaps use more sophisticated marker estimated kinship matrices in GS than simply $\mathbf{XX}'$. Results (1) and (2) suggest that it is perhaps really needed to use more sophisticated GS models than BLUP-GS, such as the Bayesian model 'BayesB' where *a priori* many marker effects are assumed to be zero. In effect, the latter reduces the genome size by concentrating on those parts of the genome where there are quantitative trait loci (QTL).

A variety of methods have been suggested for the calculation of GS-estimated breeding values (EBV), ranging from BLUP (Kolbehdari *et al.*), Bayesian methods such as BayesB (Meuwissen *et al.*, *Genetics* **157**:1819–1829), and machine learning techniques (Long *et al.*). These methods differ in their assumptions about the underlying genetic model: BLUP assumes the infinitesimal model, i.e. a large number of genes each with small effects scattered along the chromosomes; the machine learning technique assumes that there are a limited number of genes, and thus also a limited number of SNPs, worthwhile to be fitted; and BayesB is in between, i.e. it assumes few genes with large effects and many genes with small effect. The method that reflects the biological nature of the gene effects closest is expected to yield the most accurate GS-EBV. Thus, more research into the distribution of gene effects is warranted.

The paper of Muir shows that selection has a large decreasing effect on the accuracy of GS-EBV. The question arises: is the reduction of the accuracy of selection due to the Bulmer effect larger for GS-EBV than for conventional EBV? Dekkers' paper describes a general framework for answering this question. If we assume a heritability of 0.5 and only phenotypic and pedigree information are available, the pseudo-BLUP accuracy of selection is 0.77 when the Bulmer effect was ignored. This is reduced to 0.72 when accounting for the Bulmer effect (assuming selection reduced the variance of the EBV by 80% in both parents). If we assume that GS-EBV explained 59% of the total genetic variance, ignoring the Bulmer effect leads to an accuracy of GS-EBV of 0.77 [=√(0.59)]. Accounting for the Bulmer effect leads to an accuracy of GS-EBV of 0.66. Thus, the accuracy of conventional EBV is reduced by 7.5% while that of GS-EBV is reduced by 14.3%, i.e. an almost twice as large reduction. The reason for this difference is that conventional EBV obtain new information on the entire genome every generation, while GS selects every generation very accurately for the same part of the genome. This effect may be alleviated by frequently re-estimating marker effects, in the hope that new marker-QTL associations can be exploited.

When fitting of 10–100 thousands of marker (haplotype) effects, there is a severe risk of overfitting the data, i.e. errors in the data will be explained by marker effects. Cross-validation is an important tool to guard against this problem, where the data are split into two parts, one training set, where effects are estimated, and one test set, where the predictions are compared with the real data. As the test set is not involved in the estimation, its errors are independent and any correlation between predicted and realized values is due to prediction of true effects. The downside is that the training set is smaller than the total data set, so that predictions are not as accurate as they could have been. The latter can however be solved by making the test set small, e.g. 10% of the total data. As standard errors of prediction are proportional to 1/√n, where n is number of records, using 90% of the data increases the standard errors by only ~5%, i.e. a rather small increase. By performing 10 analyses with 10 different test-sets of 10% of the data, all records are left out once, and thus all records contribute to the comparison of predicted and realized values.

Traditional selection is successfully improving a large number of traits in ongoing breeding schemes, but it requires widespread (all selection candidates), reasonably accurate and preferably early in life recording of the traits. Especially, functional traits only partially fulfill these requirements, and their rates of genetic gain are therefore much lower than that of production traits. As mentioned in several of the papers, GS may overcome these problems, although Calus and Veerkamp found substantially reduced accuracies of GS-EBV for lowly heritable traits. The reason is that the accuracy of traditional EBV predictions reduces rather quickly with genetic distance (reduced relationship between the animals), while marker effects may be valid in quite unrelated animals, which makes that recordings on some animals can be used for the prediction of EBV in the entire population. As GS has the potential to achieve a more balanced selection response (balanced over production and functional traits), and substantially reduced inbreeding rates (Daetwyler *et al.*), its main effect on future breeding schemes may be an increased sustainability.

Theo. Meuwissen
Norwegian University of Life Sciences, Ås, Norway