The accuracy of genomic selection

THEO MEUWISSEN

[1]IHA-UMB, Ås, Norge

Introduction

Genomic selection is currently used in dairy cattle in wide spread manner. Other species are following this lead and consider its implementation. The aim of genomic selection is to increase the accuracy of selection especially at a young age or for traits that cannot be recorded on the candidates. It is important to recognise the factors affecting the accuracy of genomic selection and to know how important they are. For instance, how large should the training population size be, where the training population is the population in which the marker effects are estimated. Should the training population be related to the selection candidates. Should it consist of proven bulls or cows or both. Can we use other (foreign) population to increase the size of the training population. Moreover, practical limitations may reduce the accuracy of genomic selection to an extend that it is not costeffective anymore. It is good to know this before embarking on a costly marker genotyping program. Contrary to traditional selection, there is little knowledge on the factors that affect the accuracy of genomic selection. The aim of this presentation is thus to identify the factors affecting the accuracy of genomic selection and to quantify the effects of these factors. In addition it will be indicated how to improve the accuracy of genomic selection.

Use of DNA information in breeding by genomic selection

Three recent breakthroughs have resulted in the current widespread use of DNA information: the genomic selection methodology, which is a form of marker assisted selection on a genome wide scale, the discovery of large numbers of single nucleotide markers and cost effective methods to genotype them. Genomic selection estimates the effect of (hundreds of) thousands of DNA markers simultaneously, which requires assumptions about the distribution of the true marker effects. Best Linear Unbiased Prediction (BLUP) assumes the markers are normally distributed, whereas BayesB, BayesC and BayesR assume that many markers have no effect at all and some have moderate effects. The latter methods result in higher accuracy especially for traits that are known to have genes with large effects. The marker effects are estimated in a training population which is genotyped and phenotyped, and are used for the estimation of breeding values of selection candidates by combining their genotypes with the estimated marker effects. The benefits of genomic selection are greatest when selection is for traits that are not themselves recorded on the selection candidates before they can be selected, such as traits recorded in only one sex, late in life, after death, in an environment different to that it which selection candidates are kept, in crossbred offspring or which are expensive to measure.

The factors affecting the accuracy of selection

The factors that affect the accuracy of selection are (1) the marker density and through this density the linkage disequilibrium (LD) between the SNPs (Single Nucleotide Polymorphisms) and the QTL (Quantitative Trait Loci); (2) the number of phenotyped and genotyped animals that are used to estimate the SNP effects, i.e. the size of the training population; (3) the heritability of the trait or in case of proven bulls the reliability of the estimates of their daughter means, which is a measure for the size of the errors on the phenotypic data; (4) the size of the genome, which is a measure for the number of effects that need to be estimated; (5) the historical effective population size, which, if it is small, indicates that the chromosomal segments that segregate in the population are large and thus there are

few such segments that need estimation; (6) relationships between the training population and the selection candidates increases the accuracy, because even traditional selection methods achieve some accuracy if such relationships exists; (7) the number of genes and the distribution of their effects, which again indicates the number of effects that need to be estimated (if the distribution of effects is very skewed, the number of genes may be large but there are only few genes with real noticeable effects); and (8) the method used for EBV estimation.

First main determinant of accuracy : the relative information content

A main determinant of accuracy is the relative information content of the training data teta = $T*h^2/M_e$, where T is the number of training animals, h^2 is the heritability of the trait and M_e is the effective number of segments in the genome (which is approximately twice the genome size times the effective population size). The product $T*h^2$ determines the information content of the training data, i.e. their number and their reliability. Me determines the number of segments whose effects need to be estimated.

The fact that accuracy is mainly determined by teta has consequences for the comparison of accuracies of genomic selection schemes. For instance when comparing a real life scheme with h^2 is 0.25 with a computer simulation where h^2 is 0.5, we realize that the real life scheme needs twice as many training records to achieve the same accuracy (because the product $T*h^2$ remains then unchanged). Similarly, if the real life population has a twice as big effective population size, which doubles M_e, the real life scheme needs twice as many training records to achieve the same accuracy. Also, if the real life population has a twice as big a genome, again M_e is doubled and the training population size needs to double to achieve the same accuracy of genomic selection.

Second main determinant of accuracy : Marker density

If Marker density is too low, not all the genes are perfectly predicted by the markers, i.e. the linkage disequilibrium between the markers and the genes is incomplete. This has a direct effect on the accuracy of genomic selection in that the genetic variance that is explained by the SNP chip, ie. the panel of markers on a genotyping chip, is reduced. Thus, accuracy can no longer reach 100%. In addition, the unexplained genetic variance will increase the error variance of the model, and will thus reduce accuracy in the same way as that a reduced heritability does.

Conclusions

The factors affecting the accuracy of genomic selection have been identified, and are listed in the above. Two main determinants are the relative information content of the training data (teta) and the marker density. The information content of the training data is the product of the number of training animals times the heritability. Teta expresses this information content relative to the number of segments in the genome which is approximately proportional to the genome size times the effective population size. A too low marker density may make that not all genetic variance is addressed by the SNP chip, which reduces its utility. The latter seems however not a problem, with the currently commonly used dense SNP chips with 50,000 SNPs.