



EDITORIAL

Can we make genomic selection 100% accurate?

Genomic selection or genomic prediction has been widely adopted in dairy cattle breeding and increasingly in other livestock and plants. As it uses DNA information directly, rather than relying on phenotype data, it is possible that the prediction of breeding value could be 100% accurate even without a progeny test. In reality, we would be very happy with 99% accuracy, but because 100% is a limit we cannot exceed, it is a useful goal to aim for. Genomic prediction can be described as a prediction equation that takes DNA genotypes (usually SNP genotypes) as input and outputs estimated breeding values. The prediction relies on the variants that cause differences in phenotype (causal variants) or CVs being in linkage disequilibrium [LD] with one or more of the genotyped SNPs. The accuracy of the prediction equation depends on the amount of data available to train the prediction. Below I will consider two different pathways by which we might approach 100% accuracy depending on the nature of the data available to train the prediction equation.

In Holstein dairy cattle, the commercial use of DNA genotyping is so extensive that training populations of over 100,000 animals will exist in some countries for traits for which data are routinely collected through herd recording systems, for example milk yield. In this case, a prediction equation, based on say 50,000 SNPs, might approach 100% accuracy. The accuracy of BLUP to predict breeding value depends on two factors—the proportion of genetic variance explained by the SNPs and the accuracy with which the SNP effects are estimated. If the CVs have similar properties to the genotyped SNPs, so that the LD between CVs and SNPs is the same as between SNPs themselves, the proportion of variance explained by the SNPs is $M/(M + M_e)$ where M = the number of SNPs and M_e = the effective number of chromosome segments in the genome. In Holsteins, M_e is about 3,000 and therefore 50,000 SNPs might explain 94% of the variance. However, CVs might be in weaker LD with the SNPs than the SNPs are with each other because, for instance, the CVs have lower minor allele frequencies (MAF), causing the proportion of variance explained by the SNPs to be less 94%. The maximum accuracy of EBVs possible is the square root of the proportion of genetic variance explained by the SNPs.

In practice, we estimate the proportion of variance explained by the SNPs by fitting to the data a statistical model that includes the SNP effects and an effect explained

by the pedigree. The proportion of genetic variance explained by the SNPs from such analyses varies widely from 0.5 to 1 due to real differences among traits and populations and due to biases in the data. For instance, if there are common environmental effects within families, this will bias downwards the proportion of genetic variance explained by the SNPs. Reliable estimates of the proportion of genetic variance explained by the SNPs would be very useful.

The accuracy with which SNP effects are estimated (r) depends on $\theta = T h^2/M_e$ where T = number of animals in the training population and h^2 is the “heritability” of the phenotype (e.g., the reliability when the phenotype is the result of a progeny test). Given $T = 100,000$, $h^2 = 0.3$ and $M_e = 3,000$, $\theta = 10$ and $r = \sqrt{(10/11)} = 0.95$ assuming that the SNPs explain all the genetic variance. If T reaches 1,000,000, then r could be 0.995.

However, such large training populations will not be reached for all breeds neither for all traits. An alternative way to reach high accuracy might be to directly genotype the CVs or SNPs in complete LD with them. Unfortunately, in the past we have had very limited success in finding the CVs for complex traits. Will we have greater success in the future? There now exist vastly better data and technologies than existed in the past. For instance, the 1,000 bull genome project has full genome sequence on >2,000 cattle. The use of a genome sequence instead of SNP chip genotypes has the advantage that the CVs are actually in the data. However, finding the CVs is made difficult by two factors: the small effect size of most CVs and LD between CVs and other polymorphisms. Both of these difficulties can be reduced by combining data from multiple breeds. This increases the size of the data giving greater power to detect small effects while LD is reduced.

In some cases, the small effect size of some CVs might be overcome by finding a trait on which a given CV has a larger effect. For instance, a polymorphism near the gene *SLC37A1* explains 0.001 of the variance in milk yield but 0.1 of the variance in milk phosphorus concentration and 0.2 of the variance in expression of the gene (Kemper et al 2016 BMC Genomics 17: 858). An alternative approach is to use information on the functionality of particular sites in the genome independent of genetic variation in them. This type of information exists for the human genome (ENCODE) and is being collected for livestock (FAANG).



However, information that 80% of the genome is functional would be of little use because it excludes only 20% from being CVs. We need functional variants identified in a manner that helps us to identify CVs, for instance by connecting an eQTL to an economic trait.

Even if we identify CVs or SNPs in complete LD with them, we will still need to estimate their effect on traits of interest. To do this accurately will require large data sets. It seems likely that much of the genetic variance for complex traits is explained by CVs that individually explain only about 0.0001 of the genetic variance. To estimate these effects accurately, we will need data sets of about 100,000 animals.

The analysis of data sets containing full sequence data on 100,000 animals is challenging but not impossible. Fortunately, we can separate the research analyses that use full sequence data and aim to find the CVs, from routine genetic evaluation for industry which uses the CVs but not the rest of the sequence. In analysing the sequence data set, it is best to use a statistical model that fits all the sequence variants simultaneously but allows many to have zero effect and some to have a larger effect. The various nonlinear or Bayesian methods could do this. The routine evaluation could then use BLUP, perhaps with known but different variances for different CVs.

There are limitations to both approaches described above (within breed BLUP using 50,000 SNPs and across breed Bayesian analysis to find CVs). EBVs based on BLUP analysis of 50,000 SNPs within breed are not stable in space and time. For instance, a prediction equation for Australian Holsteins does not work for New Zealand

Holsteins and may not work in Australian Holsteins sometime in the future. When using sequence data to find CVs, we hope that the CVs are included in the data but currently structural variants are not included. Imputation from SNP genotypes to include sequence variants leads to imputation errors which erode the accuracy of the EBVs. This last limitation may be overcome by directly genotyping identified CVs or SNPs in complete LD with them.

Fortunately, the two approaches are not mutually exclusive. We can continue to use 50,000 SNPs from the chip while gradually finding more and more CVs or SNPs that tag them. In either case, we need large data sets. This can be helped greatly by combining breeds and by collaborating across countries or organizations that control the data.

The most immediate benefit from identifying CVs would be an increase in the accuracy of EBVs. This is especially desirable for breeds that are not as numerous as Holsteins. If smaller breeds are forced to rely on within breed prediction equations, they will fall further behind the largest breeds and become economically uncompetitive. However, there may be other benefits. In the future, it may be possible to change CV genotypes by gene editing or to manipulate their biology by non-genetic means such as feed additives.

M.E. Goddard

*Faculty of Veterinary and Agricultural Sciences, University
of Melbourne, Parkville, VIC, Australia
Email: Mike.Goddard@ecodev.vic.gov.au*