



EDITORIAL

FAQ for genomic selection

Genomic selection has been practiced in many species and in many organizations. In some cases, the results have been spectacular, and in some not. When the results fall short of expectations, questions remain as to whether they were because of inadequate statistics, too small chip size, problems with quality control or basic issues. In the end, one wonders what the limits of genomic selection are, and what will follow it. Based on published and unpublished results on genomic selection, one can prepare a FAQ sheet. Here it is. While looking at it, remember that FAQs change over time.

I have heard that with 1000 animals genotyped and phenotyped I will have accurate predictions for many generations. Is this true? *Not really. One needs more genotypes and the genomic associations decay under selection.*

How many genotypes do I need to see in order to achieve meaningful increases in accuracy? *In dairy cattle with large progeny groups, this would be approximately 2000. Larger numbers would be required if progeny groups are small and heritabilities are low. As few as 600 genotypes may be sufficient for a research paper.*

Is it better to estimate SNP effects or use a genomic relationship matrix (GBLUP)? *Both methods result in an equivalent model. It is easier to estimate weights (variances) of SNP effects with the first approach while the computations are simpler with the second one. When the weights are known, the costs for solving with optimized algorithms are similar.*

Are weights for SNP effects important? *Initial studies have indicated that they are, but as the number of genotypes increases, using equal weight for each SNP seems to provide equal or even higher accuracy for most traits. The exceptions are traits with a gene of really large effect (e.g. DGAT1 for fat content in dairy cattle).*

Are prediction equations developed for one population accurate for other populations of the breed? *It depends on links between the populations. The accuracy of prediction for an animal depends on the*

number of recent ancestors in the reference population. If that number is high and the populations are strongly linked, the accuracy may be decent. If that number is low, the accuracy will be close to 0. In the extreme, the genomic prediction for a different population, while ignoring the parent average, may be less accurate than the regular EBV.

Are prediction equations developed with one breed useful for other breeds? *They are not. They would be if SNP effects were gene effects that are similar across breeds. However, SNP effects point mostly to common haplotypes of recent ancestors, or in other words, we are getting 'better' additive relationships.*

If this is the case, what fraction of the additive variability is explained by genes or closely linked SNP? *For polygenic traits, commonly quoted fractions are from 0.05 to 0.20. Newer studies suggest that the fraction may be close to 0.50; however, these studies could be picking some genetic relationships among weakly connected individuals.*

What if prediction equations are developed with genotypes for many breeds? *The predictions can actually be pretty good for each breed present in the training data set and perhaps for simple crosses. However, scaling of SNP effects/genomic relationships is required to avoid biases and losses of accuracy across breeds and breed compositions.*

Does it help accuracy if we have multiple generations of genotyped animals? *Under selection, it may help or may not. The effect of old relationships decreases because of different values of haplotypes during selection and also because of changes in the genetic background.*

How different are genomic relationships from pedigree relationships? *If pedigrees are very deep and the population uniform, not very different, with SD of the difference <0.04 .*

There are many details in calculating the genomic relationships. Are they important? *For populations with large progeny groups (dairy cattle), nearly*

all result in similar accuracy of EBV. When progeny groups are small and when genomic and pedigree relationships are combined (single-step approach), EBV may be biased and less accurate unless the scales of genomic and pedigree relationships are the same.

The animal model uses many features such as adjustments, maternal effects, unknown parent groups and heterosis. Does the genomic model account for all these features automatically? *There is no magic. Think of the genomic selection as the animal model with more accurate relationships. Lack of adjustments would cause side effects. Some of these adjustments are not fully researched in genomic selection (e.g. unknown parent groups). Also, some adjustments are less important if animals 'compared' genomically are 'uniform'.*

Are estimates of variance components computed using the genomic relationships higher than those using the pedigree information? *They are similar if pedigrees are correct and the genomic relationships are scaled as mentioned earlier although the standard errors of the estimates should be lower. Also, with genomic information, one can estimate variance components with no pedigree.*

Will high-density chips improve the quality of prediction? *This is not clear. The genomic relationships are pretty accurate with SNP50k, so increasing the density does not help much unless causative SNPs are identified. Experience indicates a large number of causative SNPs for most traits, and their accurate estimation would require large data sets. So time will tell. Use of high-density chips with small data sets may indicate a false improvement as we are approaching the saturated model.*

Let's look into the future. If we have complete genomes, determine all causative SNPs, and estimate effects of these SNPs in a population or a mixture of populations, can we predict EBV for any animal accurately? *Good question. Genes form complex networks, and the phenotype is a highly nonlinear function of causative genes (or SNPs). In particular, the substitution value of a causative SNP depends on many other genes influencing the specific path-*

way (genetic background) and is also environment dependent. Therefore, the value of each causative SNP is not fixed but is likely to vary from line to line, over generations and even more across breeds. Using averages, over genotypes and environments will limit the accuracy. But one can get lucky, at least for some traits with strong genetic determination.

If so, then why does the animal model work so well? *The linear approximation of a highly nonlinear function is pretty good in the short run. In the animal model, animals are conditioned on parents, for one generation only. And in the short run, environments do not change much.*

The amount of genomic information is huge. Will it be overwhelming when we acquire a large quantity of genotypes? *Given genotypes of parents, the genotype of an animal can be almost completely described by the location of crossovers with some 100 numbers in total. This is exploited successfully in imputation, almost but not exactly because of mutations, genotyping errors, etc.*

What is the future for animal breeding after genomic selection? *Who knows? In the short run, genomic selection may turn out to be a more accurate animal model. There are lots of basic problems to solve, genomically or not. In the long run, we may be able to allocate matings and create genotypes optimized for specific environments.*

What is important in the implementation of genomic selection now? *Attention to detail. Use all useful phenotypes, refine models and check quality of genotypes. Recent reports indicate many problems with poor genotyping quality or genotype misidentification.*

Comments by or discussions with Ignacio Aguilar, Luc Janns, Andres Legarra, Tony Reverter and Zulma Vitezica are gratefully acknowledged.

I. Misztal,
Associate Editor
University of Georgia
E-mail: ignacio@uga.edu