EDITORIAL

Is genomic selection now a mature technology?

A couple of years ago, I wrote 'FAQ for genomic selection' in JABG (Volume 8, 245–246), and statements there are still intact IMHO. With many new studies, many murky points became clear and new puzzles appeared.

The genetic evaluation by BLUP became mature technology after the discovery of inexpensive inverse of the numerator relationship matrix (Henderson) and computing methodologies by iteration on data (Schaeffer). Then, the largest evaluations could be conducted by BLUP. Refinements continued, but the main steps were done. One can wonder whether now the genomic selection is also a mature technology.

Many studies in the past were focused on finding the optimum mix of SNP markers (or their weights), with sometimes impressive increases in accuracy. Our group at UGA has access to perhaps the largest data sets in genomics across species anywhere. Our studies showed up to 100% increases in accuracy with a small number of genotypes but little increase or none with a larger number (>10–20k). This means that in GWA (genomewide association) studies, peaks in Manhattan plots (or SNP with large effects) obtained with a small number of genotypes are mostly due to population structure. Particularly BayesB is well known in the 'academic industry' for detecting many large 'genes' that greatly change or disappear with slightly different data sets. Why? With, say 1000 genotyped animals the rank of the SNP BLUP design matrix is 1000 even with millions of SNP, resulting in a large number of solutions with a similar fit. It would be interesting to have a comprehensive study on this topic.

With genomic analyses, often we are confronted with strange validation results, usually due to using a particular validation strategy. A 'dairy' strategy based on comparing GEBV obtained with cut data and some function of EBV with complete data is appropriate only for animals (sires) with large progeny groups. A validation based on predictability (correlation between GEBV and phenotypes) seems to be the most appropriate for animals with phenotypes but is not working well with complex models (e.g. maternal). Validation based on splitting data into many groups greatly depends on the type of splitting and often yields unrealistically high accuracies, especially with few genotyped animals. So the quest for the perfect validation continues. Properties of a particular validation are clearer by looking at the decomposition of GEBV into five components: parent average, yield deviation, progeny contribution, direct genomic value and pedigree index [e.g. Lourenco *et al.*, (2015) *Genet. Sel. Evol.*, **47**:56]. One big plus of a genomic validation is that it usually includes BLUP validation, often exposing problems in BLUP models such as excessive complexity. Good BLUP models are important as bad EBVs usually mean bad GEBVs. Realized accuracies may be very low due to strong selection [Bijma (2012) *J Anim Breed Genet.*, **129**:345–358].

Single-step GBLUP (ssGBLUP) became a universally accepted methodology, partly due to simplicity (and accuracy) and partly due to problems with de-regression when genotyped animals have different amounts of information (e.g. males and females). Particularly the adoption level of ssGBLUP is high in the pig industry, which is more dependent on complex models than any other industry. Many early problems with ssGBLUP have been addressed or solved. For instance, poor convergence rates and some biases are caused by incomplete pedigrees especially if reference genotyped animals have little information. A simple solution is removing old pedigrees and phenotypes with a surprising result of sometimes higher realized accuracy. A more complex solution could be accounting for missing pedigrees via a concept called 'metafounders'. In practice, many problems in ssGBLUP are associated with quality of genomic data. Newly developed tools detect many such problems including those associated with inadequate imputation, but there is never enough tools.

As the number of genotyped animals keeps increasing (to over 1 million in Holsteins), a serious problem in ssGBLUP was the cost of inversion of the genomic relationship matrix (G). One solution was to try a recursion on a limited number of individuals. If we have a million genotyped animals, would a recursion on only 1000, 10 000 or 100 000 individuals yield sufficiently accurate inverse of G, at a greatly reduced computing cost? In the APY algorithm developed at UGA [algorithm for proven and young – see summary in Misztal (2016), *Genetics.*, **202**, 401–409], the suggested recursion was on proven animals (=sires) only. Tests with APY in Holstein involving about 20k

© 2016 Blackwell Verlag GmbH • J. Anim. Breed. Genet. **133** (2016) 81–82

doi:10.1111/jbg.12209

sires yielded perfect GEBVs. But it seemed that a recursion on 20k cows and in fact on any random sample of >15k animals was just as good.

What does it all mean? Follow the paper trail back to Fisher. In populations with small effective size, inheritance involves large blocks of DNA (also called independent chromosome segments or linkage disequilibrium blocks), with individual SNP inseparable. How many blocks are there? The top estimate is 4 times effective population size (N_e) per Morgan [Stam (1980) *Genet. Res.*, **35**: 131–155), although blocks are not of equal size and, experimentally, 25% of the largest blocks seem to explain 90% of the SNP variation. How can we find the number of blocks? Measure the dimensionality of G (~ZZ′ where Z is gene content) or the design matrix of SNP BLUP (Z′Z) as the number of eigenvalues explaining 98% of variation. With a large number of SNP and genotyped animals, studies at UGA indicated the dimensionality of approximately 15 000 for Holsteins, 10 000 for Angus, 5000 for pigs and 3000 for broiler chicken. The different numbers are due to different N_e. If the number of SNP markers is increased from 50k to 20 M, or the number of genotyped animals from 100k to 2 M, the dimensionality will only increase marginally. The small dimensionality of genomic information and APY algorithm solve the large-data problem for ssGBLUP as computations can be done in a linear time.

Small dimensionality of the genomic information explains some strange behaviours in GWA. In simulation studies where QTL are on SNP markers, the best correlation of a simulated QTL effect is not with the actual SNP effect but with an average of adjacent SNPs. Also Manhattan plots often use a window size from 1 to 10 Mbase because of excessive noise otherwise. It seems that the number of adjacent SNP (or size of the window to avoid noise) may be dictated by an average DNA block size. With 10 000 blocks, the minimum block size is about 300 kb suggesting the maximum resolution of GWA.

Limited resolution of GWA means that finding causative SNP by GWA may be hard if not impossible, as seems to be the case with recent studies. If we find causative SNP by other means (e.g. bioinformatics), will those SNP be useful in genetic evaluation? APY inverse is derived from G, and G is derived from (possibly weighted) SNP BLUP. So if we find all, say 200 causative SNP and their variances, we can create APY inverse of G with a recursion on 200 individuals. In fact, we can use APY with any mix of causative and regular SNP.

How many causative SNP are we likely to find? For complex traits under intensive selection, most likely SNP with the big effect have already been selected for, unless there is pleiotropy. In one of our studies, a large peak in the Manhattan plot for production traits in Holsteins associated with the DGAT1 gene also showed up as the only large peak for mortality. As pleiotropy is hard to detect (e.g. editing data for mortality took a year), selection on large causative SNP needs to be treated with caution. But what about using causative SNP for selection across breeds? Look into old papers about dominance, epistasis and long-term selection experiments, and doubts proliferate.

So what about ssGBLUP with APY? Think of any model, any number of phenotypes, any pedigrees, any number of genotypes, and incorporation of causative SNP if found. Also, short of genotyping mistakes, implementation of genomic selection in a day rather than a year. For polygenic traits, I declare the commercial genomic selection a mature methodology. What about genomic engineering, gene networks, etc.? Great boon for Mendelian traits and for research, and time will tell.

Comments by or discussions with Ignacio Aguilar, Andres Legarra, Daniela Lourenco, Yutaka Masuda, Ivan Pocrnic, Shogo Tsuruta and Paul VanRaden are gratefully acknowledged.

I. Misztal
Subject Editor for JABG
University of Georgia
E-mail: ignacy@uga.edu