Practical implementation of APY in swine

J. Vila T., A. E. Huisman, K. L. M. Peeters & M. C. A. M. Bink

Hendrix Genetics BU swine, Spoornstraat 69, 5830 AA, Boxmeer, The Netherlands
jordi.vila@hendrix-genetics.com (Corresponding Author)

Summary

The use of large amounts of genotypes in ssGBLUP leads to high computational demands. Here we combined the APY algorithm together with single vs. multi breed setting, attempting to find the most efficient method for animal genetic evaluation. The predictive ability analysis resulted in similar accuracies and bias across all scenarios but with significant reduction in time when the analysis included the APY algorithm in combination with a single-breed approach.

Keywords: APY, multi-breed, predictive ability, swine, pigs, genomics, genetic evaluation

Introduction

Genomic selection in animal genetic evaluations enables early accurate predictions thus boosting genetic gain (Meuwissen et al., 2001). Single-step GBLUP is currently widespread used and has become the method of choice for swine breeders. This method requires the inversion of the genomic relationship () and pedigree ( matrices, implying cubic computations and quadratic memory costs when extra individuals are added (Aguilar et al., 2010). In cattle, memory limitations have been estimated to be around 150k genotyped individuals (Aguilar et al., 2011). Differently, pig breeding companies tend to daily update breeding values and are therefore expected to run into problems to have their evaluations finished on time.

Mistral et al. (2014) developed a computationally efficient method to overcome memory and time limitations, called . It consists of dividing the genotyped individuals in two groups, core and non-core, in which EBVs for the non-core animals are derived from recursions on the core set. APY is based on the fact that the is not full rank, indicating high redundancy level of the genomic information (Mistral, 2016). All additive information in a population is encased in n particular number of independent chromosome segments (). In this context the breeding values of n animals (core set) capture all variance, the remaining breeding values being linear functions of the previous ones.

The composition of the core group is not critical and a random selection of animals delivers high accuracies (Fragomeni et al., 2015; Bradford et al., 2017). Determining the size of the core set is the same as finding the total number of and corresponds to the dimensionality of the . The latter can be addressed experimentally using eigenvalue decomposition, by finding the number of eigenvalues that explain nearly all variance and assuming the remaining variance caused by sampling noise (Pocrnic et al., 2016a). However, this is only true when the number of SNPs and individuals are ~12 times higher than the amount of (MacLeod et al., 2005; Pocrnic et al., 2016b).

In commercial breeding programs a valid question is whether several within breed evaluations or a single multi breed one should be performed. The latter can boost accuracies due to increased reference population size (Lund et al., 2014). Also, might be beneficial to
properly determine the dimensionality of the – since lots of genotyped animals are needed to achieve a full coverage of all – and consequently the size of the core group.

Here we aimed to compare the performance of the and the full approaches in combination with single-breed vs. multi-breed evaluations. Our ultimate goal was to optimize running time without giving up on accuracy so we consider that the best method is the most efficient time wise only if the approaches under study deliver similar accuracies.

Material

Data set

We used data on 465,961 records of ultrasound back fat measurements for Landrace and Large White animals that were born between 2010 and 2016. Their pedigree was traced back until 2000. A total of 10,977 Landrace and 13,948 Large White were genotyped for 51k SNPs using the GGP Porcine BeadChip (Neogen Genomics, Lincoln, NE). Data was supplied by Hendrix Genetics BV (Boxmeer, Netherlands). Table 1 shows an overview.

Scenarios

Four different scenarios were compared. Two of them used the full whereas two others used the . In addition, within each approach we compared single (SB) vs. multi (MB) breed evaluations. The genomic relationship matrices were constructed using Calc_grm software (Wageningen UR Livestock Research) and breeding values were estimated with MiXBLUP (LUKE National Resources Institute Finland and Wageningen UR Livestock Research) using ssGBLUP. For the APY scenarios, a Principal Component Analysis on the full was performed in order to determine the size of the core set of animals. The number of eigenvalues explaining 98% of the variance was chosen, given that in previous studies delivered the highest predictive ability (Pocrnic et al., 2016b).

Performance evaluation

Running time was assessed by looking at total number of PCG iterations, CPU time consumed, total wall clock time and time per iteration. Memory demand was measured by VmHWM, which reports on total RAM spent by Calc_grm. Analysis were run on a High Performance Computer cluster (HPC, Wageningen University & Research Centre) using 8 CPUs.

Correlations between breeding values were calculated comparing full and , assuming the former to be the gold standard and the latter an approximation. Also, breeding values coming out of the single vs. multi breed scenarios were compared.

Predictive ability of breeding values was assessed by regression analysis of offspring performance on sire EBV. A set of 121 Large White and 137 Landrace boars, born between 2012 and 2015, was selected. Breeding values for those boars were calculated without considering neither own nor offspring phenotypes. Then, offspring performance was regressed on sire and dam EBV and both regression coefficients () and correlation () were analysed (Formula 1 in the Appendix). A regression coefficient () of 0.5 is theoretically expected for unbiased estimates of sire breeding values as parents pass on half of their genetic merit to the offspring.
Results

Eigenvalue decomposition showed that 8,612 components explain 98% of the variance and was assigned as number of core animals for the scenarios. Its composition was determined at random resulting in 3,980 Landrace and 4,631 Large White individuals. All generations were represented.

The software run for 932 and 934 iterations in the and scenarios respectively; whereas it took a few more rounds, 975 and 971, to reach convergence for the and . Total wall clock time clearly showed the efficiency of the . Substituting the full for the resulted in time savings of 41% and 52% for the multi and single breed approaches respectively. CPU time results pointed to the same direction.

The RAM used by Calc_grm was significantly lower in the scenarios compared to the full ones. Memory savings were 46.7% when the approach was utilized compared to the . Higher memory savings were observed under the single-breed scenarios as storage was reduced by 50.2% when was compared to .

Overall correlations were high. Among genotyped animals, Pearson correlations were i) 0.987 – 0.998 for the core group, ii) 0.979 – 0.992 for the non-core group and iii) 0.946 – 0.986 when looking at 300 young candidates without records (Table 4). Correlations were always higher when comparing breeding values from and analysis (0.983 – 0.999) rather than contrasting multi-breed with single-breed scenarios (0.963 – 0.997).

Respectively for the following scenarios, and i) regression slopes () of our predictive ability model were approximately 0.54 for all scenarios with standard errors lower than 0.006, and ii) correlations () were approximately 0.69 for all scenarios (Figure 1 in the Appendix).

Discussion

The four scenarios delivered similar results in terms of prediction accuracy and bias but showed significant differences in convergence performance.

The number of PCG iterations was similar across the four cases, indicating that the equations are well-conditioned. Nevertheless, the two scenarios showed slightly higher number of rounds compared to the full . This may be caused by differences in the composition of the core group. Wall clock time was greatly reduced using the approach, indicating that savings are the result of less time per iteration. This is most likely due to i) less reading and writing as well as ii) lower number of elements in the equations to be solved. Both reasons are explained by the high level of sparsity in the . Furthermore, it is expected that the more genotyped animals included the larger time savings will be. This is because the size of the core set does not increase linearly with the amount of extra genotypes but only if new appear in the population as a consequence of new recombination events or acquisition of new genetic material. The same phenomena was observed in the multi-breed analysis, which required more time to reach convergence than the single-breed run but with a similar amount of iterations. The reasons mentioned above apply, as the relationship matrices constructed for the single-breed scenarios are much more sparse.

The correlation between breeding values estimated with the four approaches were very similar. However, higher correlations were observed if you compare than . Therefore, it seems that implementing the approach has negligible impact on the EBVs while opting for a single or multi breed evaluation has a slightly larger effect.

The predictive ability analysis delivered regression slope values close to our expectation

"Proceedings of the World Congress on Genetics Applied to Livestock Production, 11.416"
of 0.5. This indicates that half of the sire differences in breeding values can be expected in the progeny. Additionally, barely any differences can be observed across the scenarios studied neither in regression slope () nor in correlation (), implying that predictive ability is equal regardless of the approach taken. It seems that there is almost no gain in exploiting the relationship between breeds and, in fact, only increases computation time by considering non-informative relationships. The same concept applies for drawing conclusions on the . It appears that recursion on the core set leads to efficient and accurate computation of the inverse. We observed that the genomic information is highly redundant and considering that the EBV of an animal is conditional on all other genotyped animals does not increase its predictive ability but adds unnecessary computation costs.

Conclusions

Given the results presented, the most efficient implementation of ssGBLUP in a swine breeding program includes the approach in combination with single-breed evaluations. In such way, running time is significantly improved without giving up on accuracy of breeding values.

List of References

Appendix

Table 1. Overview of data

<table>
<thead>
<tr>
<th></th>
<th>Landrace</th>
<th>Large White</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observations</td>
<td>203,877</td>
<td>261,387</td>
</tr>
<tr>
<td>Animals in pedigree</td>
<td>553,412</td>
<td>783,679</td>
</tr>
<tr>
<td>Genotyped animals</td>
<td>10,977</td>
<td>13,948</td>
</tr>
</tbody>
</table>

Table 2. Number of iterations

<table>
<thead>
<tr>
<th>Relationship matrix</th>
<th>Multi (MB)</th>
<th>Single (SB)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>932</td>
<td>934</td>
</tr>
<tr>
<td></td>
<td>975</td>
<td>971</td>
</tr>
</tbody>
</table>

Table 3. Total wall clock (CPU) time

<table>
<thead>
<tr>
<th>Relationship matrix</th>
<th>Multi (MB)</th>
<th>Single (SB)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3h 12' (615.1'')</td>
<td>2h 22' (457.0'')</td>
</tr>
<tr>
<td></td>
<td>1h 54' (446.9'')</td>
<td>1h 08' (377.5'')</td>
</tr>
</tbody>
</table>

Table 4. Correlation of breeding values for core / non-core / candidate genotyped animals

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>0.988 / 0.988 / 0.965</th>
<th>0.998 / 0.992 / 0.986</th>
<th>0.987 / 0.979 / 0.987 / 0.991 / 0.988 / 0.988 / 1</th>
</tr>
</thead>
</table>

The following linear model was fitted:

Formula 1. Observations are back fat records. corresponds to fixed effects including Line, Way of test, Sex, Batch, Scan device and a composite of Line–Year–Sex. The set of dependent variables include as well the breeding values of the sire () and dam (), and being the respective regression slopes.
Figure 1. Regression of offspring performance on sire breeding value. Offspring phenotypes were first corrected for fixed effects and the resulting residuals were plotted against sire breeding value.