Proc. Assoc. Advmt. Anim. Breed. Genet. 22:317-320

DEVELOPMENT OF THE BEEF GENOMIC PIPELINE FOR BREEDPLAN SINGLE STEP EVALUATION

N.K. Connors, J. Cook, C.J. Girard, B. Tier, K.P. Gore, D.J. Johnston and M.H. Ferdosi

Animal Genetics and Breeding Unit*, University of New England, Armidale, NSW 2351

SUMMARY

Single step genomic BLUP (SS-GBLUP) for BREEDPLAN beef cattle evaluations is currently being tested for implementation across a number of breeds. A genomic data pipeline has been developed to enable efficient analysis of the industry-recorded SNP genotypes for incorporation in SS-GBLUP analyses. Complex data collection, along with format and/or naming convention inconsistencies challenges efficient data processing. This pipeline includes quality control of variable formatted data, and imputation of genotypes, for building the genomic relationship matrix required for implementation into single step evaluation.

INTRODUCTION

Genomic information from high density SNP panels has been incorporated into the Australian beef cattle genetic evaluation system, BREEDPLAN, since 2011, by "blending" EBVs from the standard analysis with direct genomic values (DGVs) from independent genomic prediction analyses using selection index theory. The ultimate goal has been to include all available information including pedigrees, phenotypes, and genotypes in a single analysis, known as single step genomic BLUP (Legarra *et al.* 2014).

One of the major practical challenges for including genomic information in genetic evaluations has been the development of scalable data-management systems (Swan *et al.* 2012) which can handle the increasing number of genotypes with increasing density of SNPs (Johnston *et al.* 2012). Quality control of the data becomes increasingly important, as inclusion of genotypes raises questions with regards to existing pedigree and potential breed. This paper describes the data pipeline developed for incorporating genomic information into SS-GBLUP analyses for BREEDPLAN, from on-farm DNA collection, through data quality control and building the genomic relationship matrix (GRM), to implementation within single step evaluation.

INDUSTRY DATA STRUCTURE

The genomic pipeline from sampling DNA on-farm to genomic evaluation is the most complex data recording process involved in genetic evaluation, and is regularly subject to errors. Samples are often handled by several people at different points in the pipeline, genotyping can be carried out by a number of different research and commercial entities using a variety of platforms, and ensuring data consistency has proved difficult.

Currently, Australian beef cattle genetic evaluations are organised individually by breed societies using databases that are maintained by the Agricultural Business Research Institute (ABRI) in most cases and using BREEDPLAN evaluation software licenced to ABRI (Graser *et al.* 2005), apart from Angus Australia who maintain their own database. At the time of writing, the role of the Animal Genetics and Breeding Unit (AGBU) within the single step genomic pipeline is to collate genotypes from various breed societies and construct a GRM which is used in

*AGBU is a joint venture of NSW Department of Primary Industries and the University of New England

single step evaluations conducted routinely by ABRI. In future it is intended for the pipeline to be incorporated into the recording and processing at breed societies and ABRI for routine SS-GBLUP evaluation.

GENOMIC DATA PROCESSING

The genomic pipeline begins upon receiving raw genotypes from a genotyping lab (Figure 1). The DNA sample must be assigned to an animal ID, usually provided by the breed society, either by name, society ID, or BREEDPLAN database number. This process has significant issues with regards to mismatching of samples to animals, particularly with historic data. Often issues with animal identities (e.g. additions/changes to suffix/prefix, duplicate names, etc.) has meant DNA samples have been attributed to the wrong animal. Thus far this has been a major hurdle in the roll-out of SS-GBLUP, as animals with simple identity changes/errors, which in turn lead to pedigree errors, will be rejected from the GRM downstream. Ensuring consistent sample identification is critical but not always successful.

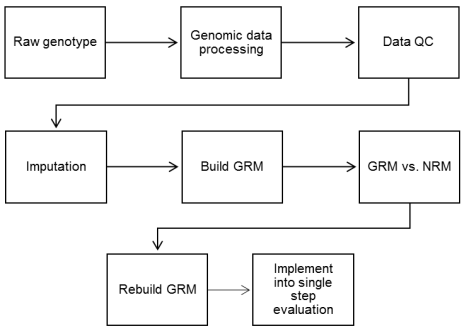

Figure 1. Genomic pipeline flow chart

DATA QUALITY CONTROL

For quality assurance purposes, raw genotypes should be provided with GenCall (GC) scores for each SNP and a SNP map file to ensure consistency across SNP panels. The SNP maps may be used for imputation and checking recombination events, and allow the genotypes to be readily converted to a consensus 150K wide format genotype. The 150K formats allow consistency across all genotypes regardless of panels/chips, and enable high-throughput data management and quality control.

With a consistent format across all genotypes, the data undergoes a quality control (QC) analysis, with filters including average GC score, missing SNPs, SNPs with low GC scores, and allele frequencies. Animals are removed from the dataset used to construct the GRM based on the following criteria:

- Less than 79% calls with a GC score > 0.6
- More than 20% missing SNPs on the observed panel
- Average GC score less than 0.6
- Sire or dam younger than 550 days (based on recorded pedigree and date of birth)
- More than 50% SNPs heterozygous
- Minor and major allele frequencies are higher than 80% or lower than 20%
- Inconsistency between assigned sex and genotype determined sex

In each case, where a genotype fails due to poor data quality, the sample/animal is flagged with the breed society and/or lab for either re-genotyping the sample or re-sampling if possible.

A 4K subset of SNPs consistent across all panels is used as a further check for the animal's suitability for the GRM, checking for breed composition, parentage/pedigree, and duplicate genotypes (greater than 90% similarity). Currently the GRM is built for purebred animals only, and as such only animals with a minimum 80% of a single breed proportion (Boerner 2017) are included. At this point, any obvious pedigree errors will be identified and either corrected or the animals will be removed from the dataset. Animals failing to meet the required criteria of the data QC will be rejected from the GRM dataset, and provided a diagnostic code describing the cause of rejection. An example of the number of genotypes removed from a GRM dataset after quality control filters are applied is shown in Table 1.

IMPUTATION

In some instances, multiple genotypes of half-sib families with the same sire are available, enabling the sire's genotype to be imputed. Previous studies have shown that the imputation accuracy depends on the SNP density and the number of half-sibs for that sire (Ferdosi *et al.* 2014). Although un-genotyped sires with small half-sib families can be imputed, the imputed genotype will contain considerable amount of missing markers and the accuracy of imputation will be low. For inclusion in the GRM dataset, half-sib families larger than 11 individuals were considered for sire imputation, with imputation and haplotyping methods similar to those implemented in the "hsphase" algorithm (Ferdosi *et al.* 2014). The phased offspring are retained in a haplotype library for FImpute (Sargolzaei *et al.* 2014). SNP loci with more than 80% missing genotypes across animals are removed, and the missing SNPs are imputed using the haplotype library and the corrected pedigree.

Table 1. Number of genotypes removed from a GRM dataset after quality control process

Quality control filter	Number of genotypes
Total	12169
Less than 79% SNPs with GC score above 0.6	167
More than 20% SNPs missing	4
Average GC score less than 0.6	8
Extreme major/minor allele frequencies (>80% and <20%)	15
Breed proportion less than 80%	489
Duplicate genotype and sample id - multiple platforms	730
Duplicate genotype - different sample id	21
Duplicate sample id - different genotype	7
Inconsistent sex (pedigree vs genotype)	82
Incorrect sire or dam	282

GENOMIC VS PEDIGREE RELATIONSHIP QUALITY CONTROL

The GRM is built using VanRaden's method 1 (VanRaden 2008). With the inclusion of genomic information, previously unidentified relationships are discovered. These relationships may simply be previously unknown or not recorded, or may be an artefact of inbreeding within the population. Regardless of the reason, the additional information provided by the GRM to identify relationships not seen in the NRM will increase the accuracy of EBVs.

However, there will also be discrepancies between genomic and pedigree relationships, most likely due to incorrect recording; even well recorded herds have a fraction of their calves (3-5%) with incorrect pedigree (Johnston *et al.* 2012). It is possible that the recorded sire of an animal

appears 'disproven' using genomic information, in which case there are a number of possible scenarios. The recorded pedigree may be incorrect, or the genotype sample may be of the wrong animal (e.g. sampling mix up, sample identity error, etc.). The issue is knowing which scenario is correct. There are a number of actions possible with the information available:

- Ignore the genotype and continue with the pedigree relationship (i.e. genotype wrong)
- Use the genotype to fix the pedigree relationship (i.e. pedigree wrong)
- Remove animal (i.e. uncertain whether pedigree or genotype is correct).

If the genomic relationship is ignored, a new genotype and/or sample should be requested. If the pedigree is corrected based on the genotype, this correction must be performed at the breed society level. It is possible that additional genotyping may change the GRM over time, as more half-sib relationships become available and new pedigree discrepancies will appear, or animals may be re-genotyped. In some instances, duplicate genotypes will occur, whereby the sample ID are the same, and the genotypes different; or the genotypes are the same, but the sample IDs are different. In these instances, it is difficult to identify which is correct, and as such both genotypes are unrecoverable. Table 1 provides an example of the number of genotypes removed from a GRM dataset after identifying duplicate samples and pedigree errors.

There are a number of assumptions in the building of the GRM with respect to using an unselected base population with little inbreeding, which can affect the genomic relationships (VanRaden 2008). Thus the issue of genomic and pedigree relationship discrepancies remains contentious, as the 'correct' action is not always obvious.

CONCLUSIONS

Increasing use of high density genomic information has the potential to improve the accuracy of genetic evaluations, and rates of genetic gain in the beef industry. This must be supported with efficient data pipelines which automate the quality control and analysis of genotypic data for inclusion into routine genetic evaluations. The genomic pipeline described here aims to do this, though difficulties arise due to complex data recording processes, multiple sample/data handling points, multiple laboratories, commercial entities and breed societies. Carefully structured and consistent data handling among the various participants will enable a smooth transition to SS-GBLUP, providing a repeatable, traceable, and auditable process, which is documented to ensure the highest quality and to identify changes over time for the Australian beef industry.

ACKNOWLEDGEMENTS

This research includes data generated by the Beef CRC. This research is supported by Meat and Livestock Australia (MLA) project B.BFG.0050. and L.GEN.0174.

REFERENCES

Boerner V. (2017) *Proc. Assoc. Advmt. Anim. Breed. Genet.* **22:** "(in press)".
Ferdosi M.H., Kinghorn B.P., van der Werf J.H., Lee S.H., Gondro C. (2014) *BMC Bioinformatics* **15:** 172.
Graser H.-U., Tier B., Johnston D.J., Barwick S.A. (2005) *Aust. J. Exp. Agr.* **45:** 913.
Johnston D.J., Tier B., Graser H.-U. (2012) *Anim. Prod. Sci.* **52:** 106.
Legarra A., Christensen O.F., Aguilar I., Misztal I. (2014) *Livest. Sci.* **166:** 54.
Sargolzaei M., Chesnais J.P., Schenkel F.S. (2014) *BMC Genomics* **15:** 478.
Swan A.A., Johnston D.J., Brown D.J., Tier B., Graser H.-U. (2012) *Anim. Prod. Sci.* **52:** 126.
VanRaden P.M. (2008) *J. Dairy Sci.* **91:** 4414.