SELECTION OF REFERENCE CANDIDATES FOR WHOLE GENOME SEQUENCING IN AN AUSTRALIAN WAGYU POPULATION

R.A. McEwin[1], M.L. Hebart[1], H. Oakey[2], R. Tearle[1], J. Grose[3], G.I. Popplewell[4] and W.S. Pitchford[1]

[1]Davies Research Centre, School of Animal and Veterinary Sciences, University of Adelaide, SA, 5371 Australia
[2]Biometry Hub, School of Agriculture, Food and Wine, University of Adelaide, SA, 5064 Australia
[3]3D Genetics Pty LTD, 939 Pukawidgi Rd, Bukkulla, NSW, 2360 Australia
[4]Popplewell Genetics, 33 Tom Schmidt Court, Mount Samson, Qld, 4520 Australia

SUMMARY

Denser genotypes on individuals has the potential to enhance genetic gain through accuracy of genomic selection. However denser genotypes, such as whole genome sequencing, on large numbers of animals is costly. This can be overcome by selecting suitable reference candidates for denser genotyping to describe the population, allowing for accurate imputation of the unselected candidates (target population). Two methods to select reference candidates were compared: the MCA method which utilises a pedigree based relationship matrix, and the MCG method which utilises a genomic relationship matrix. In a Wagyu population, the MCG method gave slightly superior imputation accuracies in the target population across differing reference population sizes as well as explaining 5% more of the genetic variance in the population when 100 candidates were selected. Similarity between chosen candidates was high between the two methods having selected 71 animals in common out of 100 with a high rank correlation of 0.82.

INTRODUCTION

Whole genome sequencing presents as an opportunity to capture more information about the genetic structure of a population which can be utilised in breeding program and mating decisions through genomic selection methodologies. However, it is costly with sequencing to 30x coverage costing approximately $1000 per sample. Through the use of imputation, high density genotyping does not need to be carried out population wide as "filling in the blanks" of sparsely genotyped individuals to higher densities can be completed using inferred haplotypes. One method of genomic selection is genomic best linear unbiased prediction (G-BLUP; Clark and van der Werf 2013) which utilises a relationship matrix calculated from a genotyped set of individuals. The resulting relationship matrix can be utilised to determine which individuals are the best candidates to describe variation in the population (i.e. form the reference population) and therefore suitable for genome sequencing to achieve high imputation accuracies. This is the aim of the commercial Wagyu breeding program behind this study and while other methods exist, e.g. Bickhart *et al.* (2015), the convenience of using a relationship matrix, having been already constructed for genomic evaluations, was appealing.

MATERIALS AND METHODS

Selection of candidates was carried out using two methods described by Yu *et al.* (2014). The first, denoted the MCA method, selects candidates for whole genome sequencing by minimising the genetic variation of the target population, relative to the selected pool, in order to improve their imputation accuracy. This method utilises Wrights numerator relationship matrix (A) such that;

$$\mathbf{A}_{11}^{*} = \mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}$$

where the 1 subscript denotes the set of target animals and 2 subscript denotes the set of animals selected to be sequenced. Diag($\mathbf{A}_{11}^{*}$) are the residual variances that are expected to remain if sequence data were to be obtained from the selected individuals and used to predict/impute genotypes of the target set. Animals were selected using an iterative process. **A** was constructed using an Australian Full-Blood Wagyu pedigree comprised of 10,549 individuals with a depth of up to 9 generations from the current generation using the R package *pedigreemm* (Bates and Vazquez 2014).

The second method (MCG) is akin to MCA but utilises a genomic relationship matrix (G) in place of A. G was constructed as per VanRaden (2007) method 2, utilising genotype information on 5,334 individuals genotyped with 30K GGP-LD (Neogen: GeneSeek Operations) or Bovine VersaSNP 50K (Weatherbys Scientific) chips. Animals genotyped on the Versa SNP were imputed to 30K from the approximate ~10K overlap between the chips, due to the significantly larger reference population available (4940 vs. 394), using Fimpute 2.2 (Sargolzaei *et al.* 2014). After imputation, SNPs were retained that had a minor allele frequency greater than or equal to 0.05 before building the GRM. All genotyped animals were present in the pedigree resulting in 5,334 animal overlap between the numerator (A) and genomic (G) relationship matrices.

Imputation accuracy, described here as the correlation between true and imputed genotypes (r), was calculated for the 4,940 individuals genotyped on the 30K chip by masking their true genotypes to a ~10K density. Seven rounds total of single replicate genotype imputation (Fimpute 2.2) was then carried out using 4 reference population sizes (100, 50, 25, 10) of animals selected for whole genome sequencing from the 2 methods (MCA or MCG respectively).

RESULTS AND DISCUSSION

The degree of similarity between the MCA and MCG methods was very high with MCA selecting 71/100 individuals that were selected by MCG. Of the animals that were selected by both methods, they were ranked very similarly with a strong positive rank correlation of 0.82 (Figure 1). This is a stronger relationship than previously reported Yu *et al.* (2014), however with approximately half of the animals in the pedigree having been genotyped and the target population also being the potential selection pool it is less surprising the lists are similar. The MCG method did account for slightly more genetic variance reaching 35% when 100 animals were selected compared to 30% accounted for using the MCA method. The first 20 selected animals accounted for 19% and 21% of the genetic variance for the MCG and MCA method respectively with each additional animal there after contributing less information (Figure 2).

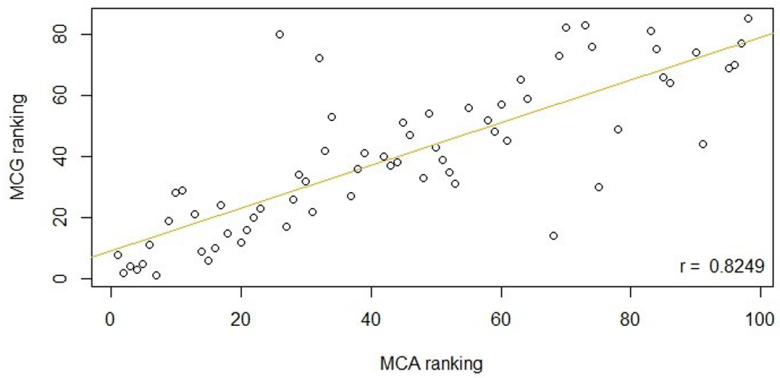

Figure 1. Correlation between ranks of candidates selected for whole genome sequencing using the MCA or MCG methods respectively

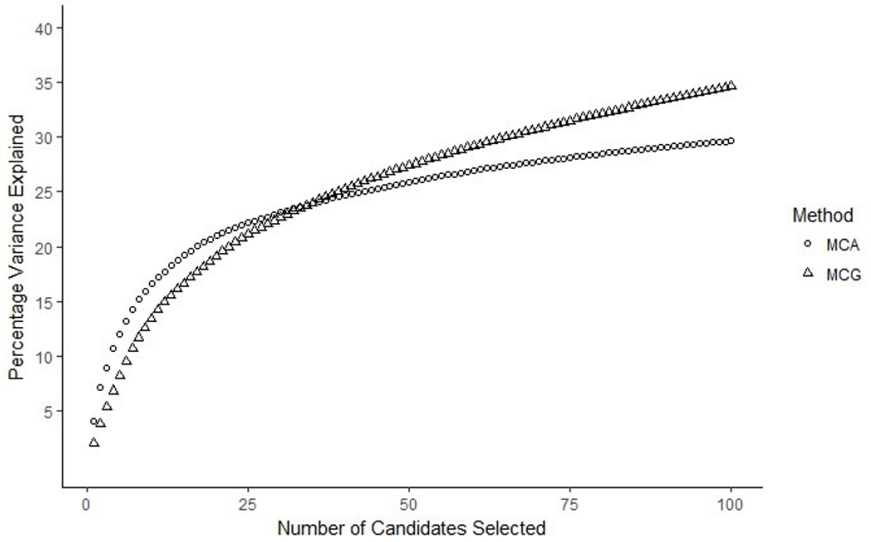

Figure 2. Diagonal values of A* representing the percentage of genetic variance explained for each additional selected candidate for whole genome sequencing using the MCG method (top) or MCA method (bottom)

A common logic is to identify animals that have a higher number of descendants, i.e. are considered influential, to be selected for sequenicng. For 100 genotyped sires (with effective progeny numbers of 1 to 437, mean = 47, in this population) the amount of genetic variation accounted for was 30%, equivalent to the MCA method but lower than MCG method.

Imputation accuracy was calculated for both the MCG and MCA method. Larger reference populations gave the highest imputation accuracies which is to be expected. For animals selected using the MCG method, selecting 100 animals was fairly comparable to selecting 50 animals with a noticeable drop in the mean accuracy from ~0.96 to 0.94 and 0.83 when 25 and 10 animals are selected respectively (Table 1). For MCA, out of the 100 selected animals, only 75 were genotyped and so could be used to calculate imputation accuracy. For comparisons sake, only reference populations of 50, 25 and 10 were constructed. MCA selected animals who weren't genotyped were generally lowly ranked, however a few non-genotyped candidates were present in the higher ranks. Therefore the MCA reference populations are not "perfectly" ranked as per the MCA method. Higher ranked animals that were not genotyped were replaced by the next available ranked animal until the desired number was sampled. The mean accuracies for MCA at 50, 25 and 10 reference animals were comparable to MCG although MCG was slightly superior. Noticeably though MCA did have a much higher minimum accuracy of 0.67 compared to 0.55 for MCG which indicates a narrower spread of imputation accuracies giving more successful imputation overall (Table 1).

Table 1. Imputation accuracy calculated for sparse 11K genotypes imputed to 30K using differing reference populations of different sizes selected from two methods

MCG					MCA			
Ref size	100	50	25	10	-*	50	25	10
Min	0.5495	0.5482	0.5474	0.5101	-	0.6724	0.5317	0.5139
Mean	0.9756	0.9659	0.9395	0.8331	-	0.9627	0.9271	0.834
Max	0.9996	0.9992	0.9968	0.9725	-	0.9995	0.9978	0.9863

***** For MCA, out of the 100 selected animals, only 75 were genotyped and so could be used to calculate imputation accuracy. For comparisons sake, only reference populations of 50, 25 and 10 were constructed. For MCA the next available candidate was selected if no genotype was available and so reference populations do not display perfect ranking but can be used as an example.

Both the MCA and MCG method assumed that all potential selection candidates had DNA available for sequencing and in a commercial pedigree this is not always the case. This fact became partially evident in the imputation study where not all MCA selected candidates had genotypes to form the reference. This is an important consideration and both methods could be easily modified to account for this. Within an iteration, the animal that is selected is logically the one that reduces the residual genetic variance of the target population i.e. Diag($\mathbf{A}_{11}^{*}$), the most. Multiplying each candidates impact on the residual by a simple vector of 0 (no DNA available) or 1 (DNA available) would ensure that only candidate animals with DNA are selected. This would also prevent bias when selecting sequence candidates to form the reference if you were just to remove animals with no DNA from the analysis all together.

CONCLUSIONS

For a full-blood Australian Wagyu herd their appeared to be little difference between the MCG and MCA methods for selection of candidates for whole genome sequencing. Both methods accounted for greater than 30% of the genetic of the target population when selecting 100 candidates and had comparable imputation accuracies up to 30K. Given the large volume of genotypes and deep, complete pedigree, either method would be suitable to select whole genome sequencing candidates to form the reference population for imputation. At the commercial level, the MCG method was selected to sample 73 candidates for sequencing due to the higher likely hood of selecting candidates with DNA sources (hair or semen) available in the first instance and the need to QC each individual hair sample in store prior to DNA extraction if semen was not available.

REFERENCES

Bates D. and Vazquez A. (2014). pedigreemm: Pedigree-based mixed-effects models. R package version 0.3-3.https://CRAN.Rproject.org/package=pedigreemm

Bickhart DM., Hutchison DJ., Null DJ., VanRaden PM. and Cole JB. (2015) *J. Dairy Sci.* **99**:5526.

Clark S., van der Werf J. (2013) In: Genome-Wide Association Studies and Genomic Prediction. Methods in Molecular Biology (Methods and Protocols), pp. 321- 330, editor Gondro C., van der Werf J. and Hayes B. Humana Press, Totowa, NJ

Neogen: GeneSeek Operations. https://genomics.neogen.com/pdf/ag151_ggp_ts.pdf

Sargolzaei M., Chesnais J. and Schenkel F. (2014) *BMC Genomics* **15**:478

VanRaden PM (2008) *J. Dairy Sci.* **91**: 4414.

Weatherbys Scientific. 2017, https://weatherbysscientific.com/versasnp/

Yu X., Woollimas J. and Meuwissen T. (2014) *Gen. Sel. Evol.* **46**: 46.