# EVOLVING TO THE BEST SNP PANEL FOR HANWOO BREED PROPORTION ESTIMATES

**C. Esquivelzeta-Rabell[1], H.A. Al-Mamun[1], S.H. Lee[2], H.K. Lee[3], K.D. Song[3] and C. Gondro[1]**

[1]School of Environmental and Rural Science, University of New England, NSW, Australia
[2]Division of Animal and Dairy Science, Chung Nam National University, Daejeon, Korea
[3]The Animal Genomics and Breeding Center, Hankyong National University, Anseong, Korea

## SUMMARY

Hanwoo is highly prized for its marbling ability and is the most important cattle breed in Korea. In order to maintain the integrity of the breed and for product certification purposes it is important to develop tools to confirm the origin of the products. Breed composition estimates based on a large number of molecular markers (e.g. HD SNP arrays) are highly accurate but expensive for routine usage. The identification of a reliable panel with a small number of markers will reduce costs and can enable broader adoption of the technology by industry. In this work a heuristic optimization method was used to find the most reliable subset of markers, from the Illumina BovineHD array, to estimate breed proportion in Hanwoo. Accuracies of breed proportion estimates above 90% can be achieved using as little as 200 markers. The best balance between accuracy and number of SNP was obtained with 500 markers achieving 94% accuracy. Rapid and cost effective breed composition prediction in Hanwoo cattle based on a SNP panel with at least 200 markers will help to certify the products with an acceptable accuracy and ensure breed purity within the breeding program. The method described herein is directly applicable to other breeds.

## INTRODUCTION

Hanwoo is the most important native Korean cattle and its history traces back 5,000 years (Jo *et al.* 2012). Over this long timespan the purpose of these cattle has evolved from farming, transportation and religious sacrifice to beef production (Lee *et al.*, 2014). Hanwoo beef has unique marbling characteristics which confer a special tenderness, juiciness and unique flavour to the meat, making it highly sought after by consumers at premium prices (Kim *et al.* 2000; Han and Lee 2010; Jo *et al.* 2012). It has also been shown that Hanwoo has a healthier fatty acid composition in comparison to other breeds (Jo *et al.*, 2012) which makes them even more attractive to consumers. In order to certify the products it is important to develop cost effective tools that allow verifying that the product truly comes from pure bred Hanwoo cattle. Breed prediction is also a useful tool for breed associations where the animals need to be purebred to be registered and, within genomic selection (GS) programmes, it can be used for quality control of research and industry samples (Dodds *et al.*, 2014).

Before the availability of marker data, breed proportion estimates could only be obtained from pedigree information. Single nucleotide polymorphism (SNP) genotypes potentially allow for more accurate estimates of breed proportion, even in the absence of pedigree records. A number of tools exist for predicting breed composition using genetic markers. Most of these implement statistical methods developed for prediction of admixture levels and use the complete set of markers. Common approaches are based on hidden Markov Model (HMM) clustering algorithms or maximum likelihood procedures (Frkonja *et al.*, 2011). To obtain estimates of breed composition in crossbred populations, a *reference population* consisting of genotypes from purebred animals that may have contributed to the composite population are used. Dodds *et al.* (2014) explored genomic selection methodology by comparing GBLUP with regression methods to develop predictions for breed proportions. This study showed that either method can be applied

but which one is better depends on the structure of the ancestral breeds that contributed to the population of interest. Blackburn *et al.* (2014) showed that, in composite populations, using a small set of 60K markers (extracted from the Bovine HD SNP chip) at high frequency in each of the founder breeds; the proportion of the founder breeds in the composite animals can be estimated. In combination these studies showed how promising SNP panels are to characterize genetic composition within a population; nevertheless if the main objective is product certification and breed verification, the use of a full high density SNP panel has economic constraints. Consequently, it is of practical importance to find a small and accurate subset of SNP to estimate breed composition. In the present study we explored the use of Differential Evolution to identify a small SNP panel that can accurately be used for Hanwoo breed composition evaluation.

## MATERIAL AND METHODS

**Data.** Genotype information from the BovineHD (700K Illumina BeadChip) array was available for a total of 2,453 animals from different cattle breeds (Hanwoo, Angus, Brahman, Charolais, Holstein and Jersey). The data set was divided into a discovery (2,253) and a validation (200) population. The discovery and validation samples were mutually exclusive. First, 200 samples were randomly selected among the 6 different breeds previously mentioned as validation samples and then the remaining samples were used as discovery population. After quality control 497,737 SNP across all populations were kept for further analysis. A second dataset consisting of genotype information from 24 Yeonbyun samples was also used to validate the proposed method. Yeonbyun are genetically highly related to Hanwoo (populations separated during the Korean War) with some level of crossing with European breeds (Gondro *et al.*, 2012a); which makes them suitable as a proxy for crossbred Hanwoo.

**Breed proportion.** Breed proportion estimates were calculated using the supervised option with K=7 implemented in the ADMIXTURE software (Alexander *et al.*, 2010). From the breed proportion output we estimated the Hanwoo proportion of the validation set animals. Breed proportion was considered as the *trait*. Phenotypes of pure bred Hanwoo animals were coded as 1; animals of the other reference breeds were coded as 0; therefore prediction of the validation animals using the SNP subset was expected to be in the range of 0–1. A principal component (PC) analysis was also performed to better understand breed composition, to explore potential sub-structure within the sample and for graphic display of the data.

**Evolutionary Algorithm.** An algorithm based on Differential Evolution (DE) (Storn and Price, 1997) was used to select the best set of SNP for breed proportion estimation. Random keys were used to select the SNP panel. A random key is an evolvable vector of real numbers (one for each SNP) that are sorted in the objective function and then the ranking of the key is used to rank the SNP. The idea is that, SNP for better breed proportion estimation will evolve to higher values in the key and the rest to lower values; once the keys are sorted they reflect the relative value of a given SNP. Predefined *cutoff values* (100, 200, 300, 400, 500, 1000 and 5000) were used to select the number of SNP in the panel. Basically the DE evolves and sorts the SNP based on their key values and uses the top ranked ones up to the number defined by the *cutoff* parameter. More in-depth details on the algorithm are given in Gondro and Kwan (2012b). An objective function was used to find the *fitness* of the selected SNP panel. In the objective function, the discovery population was further divided into two subsets: i) a subset population (1,253 random samples) with known Hanwoo proportion and ii) another subset population (1,000) with unknown Hanwoo proportion (proportions were set to missing for these samples). A genomic relationship matrix (GRM) was calculated using only the selected SNP panel with the all 2,253 discovery samples. The resulting GRM was used to predict the Hanwoo proportion (using GBLUP if number of SNP > number of animals, SNP-BLUP otherwise) for the 1,000 samples with unknown breed estimates. The fitness of a selected SNP panel (accuracy) was defined as the correlation between the

observed and the predicted Hanwoo proportion for these 1,000 samples. The DE evolved for 100 to 500 generations depending on the number of SNP used in the panel; SNP panel size being inversely proportional to the number of iterations. We used 500 generations to evolve the DE for SNP panels with 100 – 400 SNP; 200 and 100 generations for SNP panels with 500 – 1,000 and 5,000 SNP respectively. Once the DE finished, the SNP panel with the highest fitness value was selected and the SNP effects were saved to perform prediction on the validation data. Prediction for Hanwoo proportion was calculated using the following equation: $\hat{y} = 1_n\mu + \sum_i \boldsymbol{X}q_i + e$, where $\mu$ is the mean, $\boldsymbol{X}$ is an incidence matrix linking observations to SNP genotypes, $q_i$ is the estimated effect of each SNP and $i$ = 1 to the number of SNP on the SNP panel.

**Random subsets**. To compare the performance of the DE Algorithm SNP were randomly selected for different panel sizes (100, 200, 300, 400, 500, 1000 and 5000) and then SNP-BLUP was performed on both validation sets. The accuracy of Hanwoo proportion estimates shown for each SNP panel is the average of 10 independent random samples.

## RESULTS AND DISCUSSION

Figure 1 shows the first two PC of the genomic relationship matrix applied to 164 animals from different cattle breeds. Hanwoo cattle is clearly separated from the European breeds and Yeonbyun animals are between Hanwoo and European breeds, showing that most of the animals are genetically highly related with Hanwoo cattle, agreeing with Gondro *et al.* (2012a) and that these cattle have potentially been crossed with European breeds. Consequently Hanwoo proportions in Yeonbyun animals are expected to be between 0 and 1. These results were confirmed when calculating the Hanwoo proportion of the validation set using the SNP panel selected with the DE Algorithm (data not shown).

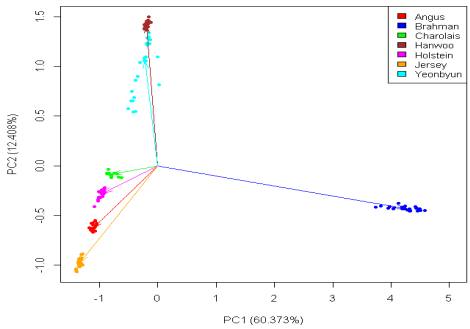


**Figure 1. Top 2 axes of variation from principal component analysis of the breeds used to select the marker panel for breed proportion estimates.**

Accuracy of breed proportion estimates using different number of markers selected with the DE Algorithm ranged between 0.83 and 0.99 for sets of 100-5000 SNP (Table 1). When using Hanwoo and other European breeds as a validation set, the accuracies didn't change much among SNP subsets using the DE Algorithm (100 to 300 SNP 98% and >300 99%) or random selection (93% with 100 SNP and 96 to 99% with >100 SNP). Results show that the number of SNP included in the different panels is sufficient to extract information about breed proportion in the population, being better than what previous studies suggest (5K SNP, Frkonja *et al.*, 2011; and 60K, Blackburn *et al.*, 2014) and demonstrating that using only a fraction of SNP from the HD SNP panel we could predict the phenotype or Hanwoo proportion which is comparable with the prediction accuracies achieved when all SNP are used (0.99). However if the accuracy is important then we need to use larger SNP panels (i.e. panels with about 1,000 SNP). On the contrary if the cost is the main concern then we could use panels with a lower number of SNP by accepting a small decrease in accuracy. It should be noted though that Hanwoo is genetically quite distinct from European breeds and panels to resolve breed composition within European breeds will probably need to be larger.

**Table 1. Accuracy of breed proportion estimates using Differential Evolution (DE) Algorithm and random SNP with different number of markers in the Yeonbyun validation set.**

| SNP | DE | Random |
|---|---|---|
| 100 | 0.83 | 0.51 |
| 200 | 0.91 | 0.72 |
| 300 | 0.91 | 0.76 |
| 400 | 0.91 | 0.83 |
| 500 | 0.94 | 0.81 |
| 1000 | 0.94 | 0.91 |
| 5000 | 0.99 | 0.98 |

Knowledge of animal breed composition in livestock populations is also important to identify the best candidates for selection. In crossbred populations it allows effective exploitation of heterosis effects by enabling accurate decisions about the best mattings to be performed within the population. Further, breed composition of crossbred animals in livestock populations provides information on the type and level of crossbreeding as well as on the level of recombination loss (e.g. VanRaden and Sanders 2003). Use of SNP panels increases the level of resolution at which the genetic diversity of composite breeds can be managed. Breed prediction also becomes possible in the case of incomplete or missing pedigrees and in the search for the best type of cross or composite of breeds.

## CONCLUSION

The method presented in this study suggests that small, accurate and cost effective SNP panels can be identified for breed proportion evaluation. The results represent a promising approach for product certification and to ensure breed purity in Hanwoo at a low cost. This method can be ported seamlessly to other breeds as well.

## AKNOWLEDMENTS

This work was supported by a grant from the Next-Generation BioGreen 21 Program (Project No. PJ01134906), Rural Development Administration, Republic of Korea and an Australian Research Council Discovery Project DP130100542.

## REFERENCES

Alexander D.H., Novembre J. and Lange K. (2010) ADMIXTURE 1.04 Software manual, Version 1.04.

Blackburn H.D., Paiva S.R., Sollero B.P., Biegelmeyer P., Caetano A.R. and Cardoso F. (2014) Proc 10th WCGALP. Manuscript 470.

Bovine HapMap Consortium. (2009) *Science.* **324**:528.

Dodds K.G., Auvray B., Newman S.N. and McEwan J.C. (2014) *BMC Gen.* **15**:92.

Frkonja A., Gredler B., Schnyder U., Curik I. and Solkner J. (2012) *Anim. Gen*. **43**:696.

Gondro C., Jang G.W., Lee S.H., Yeon S.H. and Seong H.H. (2012a) Proc. 15th AAAP. Paper Code: C15-OP-161.

Gondro G. and Kwan P. (2012b) in Multidisciplinary Computational Intelligence Techniques: Applications in Business, Engineering and Medicine, pp. 351-377 Editor IGI Global.

Han S.W. and Lee B.O. (2010) *J. Agric. Life Sci.* **22**:73.

Jo C., Cho S.H., Chang J., Nam K.C. (2012) *Anim. Front.* **2**(4):32.

Kim K. H., Kim Y. S., Lee Y .K. and Baik M. G. (2000) *Meat Sci.* **55**:47.

Lee S.H., Park B.H., Sharma A., Dang C.G., Lee S.S. *et al.* (2014) *J. Anim. Sci. and Tech.* **56**:2.

R Development Core Team. (2012) R Foundation for Statistical Computing.

Storn R. and Price K. (1997) *J. Global Optim.* **11**(4):341.

VanRaden P.M. and Sanders A.H. (2003) *J. Dairy Sci.* **86**:1036.