*Proc. Assoc. Advmt. Anim. Breed. Genet. 20:511-514*

# A BINARY CLASSIFIER USING SNP DATA FOR PREDICTION OF PHENOTYPIC OUTCOMES IN HANWOO (KOREAN) CATTLE

**D.C. Detterer[1,2], S.H. Lee[3], P. Kwan[1] and C. Gondro[2]**

[1]School of Science and Technology, University of New England, Armidale NSW 2351, Australia
[2]The Centre for Genetic Analysis and Applications, University of New England, Armidale NSW 2351, Australia
[3]Hanwoo Experiment Station, National Institute of Animal Science, RDA, Pyeongchang, Korea

## SUMMARY

Korean *Hanwoo* cattle are prized for their high marbling ability and meat quality. Classically, these cattle possess a homogeneous yellow coat colouring, with farmers believing that *Hanwoo* with white spotted coats are crossbred and therefore unacceptable for breeding purposes. In this study we first attempted to determine if the coat spots were due to a non-*Hanwoo* genetic background or, alternatively, if the trait is intrinsic to the breed. By genotyping 232 (136 spotted) animals from half-sib families on the Illumina Bovine 50K SNP array, we compared the genotyped *Hanwoo* to other unrelated Hanwoo and European taurine breeds using principal component analysis. Results showed no evidence of crossbreeding in the spotted animals. A differential evolution algorithm was then used to evolve a classifier for the trait which selected 12 SNP with an accuracy of ~82% in separating individuals; further investigation using only haplotypes inherited from the sires resulted in a marked improvement to ~92% accuracy for these 12 SNP. This research highlights the potential for using these SNP as genetic markers to either entirely remove the trait from the population in the long term or manage matings so that the trait is not expressed in the offspring.

## INTRODUCTION

The shift towards breed analysis via large scale genomic data has provided greater accuracy in prediction and opened avenues for a more complete understanding of the biology underpinning phenotypic traits useful for selection (Hayes *et al*. 2009; Habier *et al*. 2010).

The most common markers found throughout the genome are single nucleotide polymorphisms (SNP), single points of differentiation between individuals within a population. Through linkage disequilibrium, marker SNP associated with a disease or quantitative trait can suggest areas in the genome that demand further investigation (Carlson *et al*. 2004).

In the present research, after quality control filtering, we used 37,065 SNP genotyped on the Illumina 50K array for 232 Korean *Hanwoo* cattle derived from 28 sires. 136 of the 232 cases exhibited spots in their colouring—a trait deemed undesirable by breeders (see Brown & Lawrence 2010 for a study linking coat colouring to beef carcass grading). Note that none of the sires exhibited coat spots. An additional 229 animals from 5 European taurine breeds and other unrelated *Hanwoo* were also used in the analyses.

Currently, spotted *Hanwoo* cattle are simply culled due to a belief that these animals are not purebred. Although this is effective as a *brute force* method for removing undesired animals from the population, genomic technologies offer more efficient means of removing undesired traits by informing breeding choices. Ultimately, the objective is to breed-out the undesired trait rather than waste resources breeding cattle only to cull undesired offspring.

## METHODS

A three-step process was used in this study. Firstly, principal component analysis (PCA) was conducted on the genomic data in order to ascertain the relationship between the spotted *Hanwoo*

cattle, the homogeneous *Hanwoo* population and other taurine breeds. Secondly, a differential evolution (DE) algorithm was used to search for a set of SNP within the genomic data which would classify the animals in the *Hanwoo* data set according to their status as either spotted or unspotted. Thirdly, the *Hanwoo* data set was phased into sire and dam haplotypes, with the SNP-based classifier derived in the previous step applied to the sire haplotypes.

In the first step, a Genomic Relationship Matrix (GRM) was used to characterise the relationship between individuals and the principal components of the matrix were then calculated.

In the next step of this investigation, *k*-means clustering was used to drive a DE algorithm as a strategy for stochastically selecting SNP that could separate between the spotted and unspotted animals.

DE is a heuristic in the family of evolutionary algorithms which also includes genetic algorithms, evolution strategies and evolutionary programming (Price *et al*. 2005; Fogel 2000). DE is a relatively straightforward algorithm to implement; it evolves real-valued vectors of parameters against an objective function. The purpose of the objective function is, in turn, to evaluate the "fitness" of a given vector in relation to how successfully its parameters solve a given problem.

In the present research, the objective function was based on *k*-means clustering, testing a set of SNP as to their ability to effectively separate the cases into spotted and unspotted cattle. Each real-valued vector in the DE algorithm represents a set of SNP using relative position indexing (Onwubolu & Davendra 2009); over the course of a run, SNP with greater predictive value in the clustering are given greater weight so that by the end of a run, the SNP selected collectively perform better as predictors. 100 runs of the DE/*k*-means algorithm were carried out in order to discover the most commonly selected SNP across the genome. From SNP selected 3 or more times, further exploratory runs of the algorithm were conducted to experimentally find a set of SNP that offered the greatest separation between spotted and unspotted cases.

As a final step of this analysis, an attempt was made to apply the classifier based on the selected SNP to the haplotypes each individual inherited from its sire.

The SNP data from the 232 *Hanwoo* cattle were phased into haplotypes inherited from the sire and dam respectively for each offspring. After removing any animals having unphased alleles on any of the selected SNP on the sire-inherited haplotypes, 60 animals remained, 33 of which were spotted.

**RESULTS**

As the initial hypothesis was that the spotted cattle were not purebred *Hanwoo* cattle but instead crossbreds, we ran a PCA of the genomic data. The purpose of this step in the overall analysis was to ascertain the relationship between the spotted *Hanwoo* cattle, the homogeneous *Hanwoo* population and other taurine breeds.

Applying PCA to the data resulted in a clear clustering separating the *Hanwoo* cattle from other breeds. Furthermore, spotted and unspotted Hanwoo form the same tight cluster—this suggests that the spotted *Hanwoo* are purebreds and not the result of crossbreeding. It is thus reasonable to assume that the genetics determining spotted and unspotted cattle are intrinsic to the breed itself. In Figure 1, the spotted and unspotted *Hanwoo* cattle clustered indistinguishably together on the left, with European beef and milk breeds forming clusters on the right.

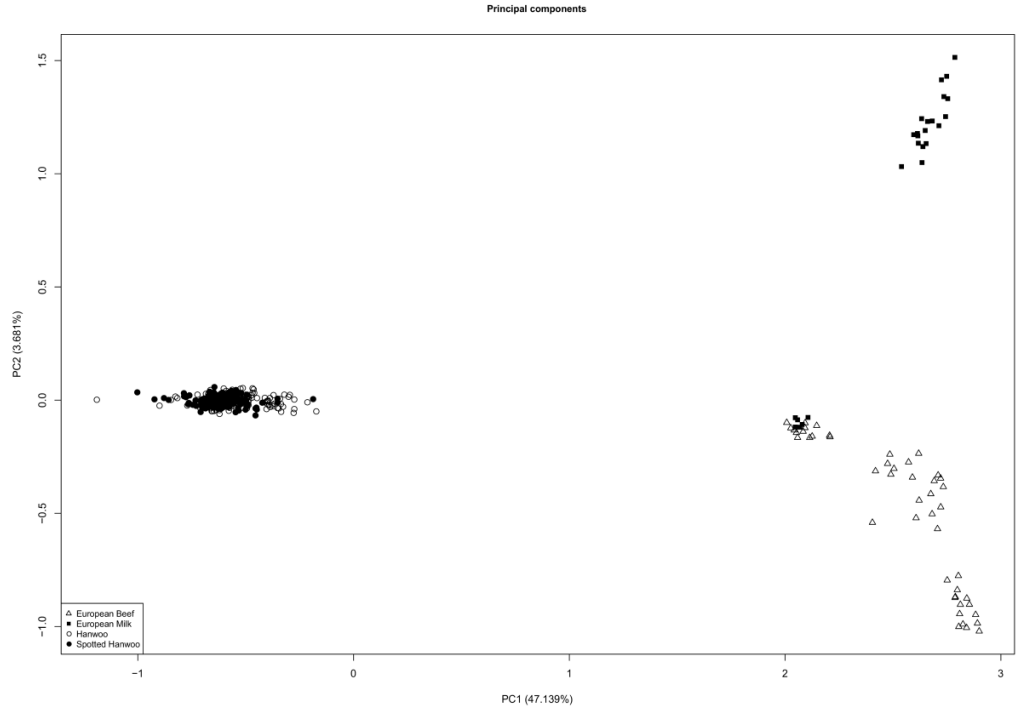


**Figure 1. Distribution of animals according to the first two Principal Components of the SNP data set. Values in parenthesis are the variance explained by each component.**

After applying the DE/k-means algorithm to the data set, 12 SNP were selected, giving a classification accuracy of ~82%; selecting more SNP did not increase the classification accuracy. In contrast, 10,000 random sets of 12 SNP selected from the 37,065 initial SNP resulted in classification accuracy with a mean of ~53% and a standard deviation of ~1.5%.

Table 1 lists the 12 selected SNP, the chromosome where each SNP is located and the position of the SNP on its respective chromosome.

**Table 1. Selected SNP used in *Hanwoo* cattle classifier**

| SNSeP | Chromosome | Position on Chromosome |
|---|---|---|
| 6149 | 3 | 120898833 |
| 7282 | 4 | 69056689 |
| 7827 | 4 | 106325551 |
| 7874 | 4 | 110856142 |
| 8800 | 5 | 58579368 |
| 13603 | 8 | 37461264 |
| 19453 | 12 | 901956 |
| 21220 | 13 | 42667485 |
| 27491 | 18 | 55895369 |
| 29257 | 20 | 44031834 |
| 34170 | 26 | 1632525 |
| 36876 | 29 | 41247528 |

Finally, performing *k*-means clustering on the phased sire haplotype data of the remaining 60 animals using the 12 SNP yielded a classification accuracy of ~92%. Given that the sires' coats were unspotted, this suggests the classifier may be effective at predicting the potential for spotted offspring given unspotted sires.

## CONCLUSION

In this paper, an evolutionary algorithm was used to build a binary classifier for a phenotypic trait relevant to the Korean *Hanwoo* industry. Principal component analysis demonstrated that coat spotting in *Hanwoo* cattle is not due to crossing with other breeds, indicating that the trait is intrinsic to the breed. Using a differential evolution algorithm, 12 SNP were selected which were able to classify the cattle with an accuracy of ~82% via *k*-means clustering. Furthermore, classification on sire-inherited haplotypes gave an increase in accuracy to ~92%, suggesting that selective breeding based on SNP data is a viable path for removing the spotted trait from the population.

This is a promising start to a larger investigation into the utilisation of genomic markers to remove spotting from the population. However, the small sample size combined with a possible over-parameterization of the data means that independent validation is needed before the markers are adopted by industry.

Future work will involve validation of these results, including the use of cross-validation methods and further data gathering, which is currently underway. In addition, an investigation into the biology underpinning the relationship between the selected SNP and the spotted phenotypic trait will be undertaken.

Ultimately, the aim is to provide industry with a marker set to enable breeding decisions in the *Hanwoo* cattle industry.

## ACKNOWLEDGEMENTS

This study was funded by grants from the Next-Generation BioGreen 21 Program (No. PJ008196), Rural Development Administration, Republic of Korea and the Australian Research Council project DP130100542.

## REFERENCES

Brown, T.R. and Lawrence, T.E., (2010) *Professional Animal Scientist* **26**: 611.
Carlson C.S. et al., 2004. *The American Journal of Human Genetics* **74:** 106.
Fogel D.B. (2000) *IEEE Spectrum*, **37:**26.
Habier D. et al. (2010) *Genetics Selection Evolution* **42:**5.
Hayes B.J. et al. (2009) **41:** 51.
Onwubolu G. and Davendra D. (2009) *Differential Evolution: A Handbook for Global Permutation-Based Combinatorial Optimization*, pp.13–34.
Price K.V., Storn R.M. and Lampinen J.A. (2005) *Differential evolution*, Springer Berlin, Germany.