# INCREASE OF POWER AND EFFICIENCY TO FINE-MAP GENETIC DEFECTS USING GENOTYPE PROBABILITIES THROUGH SEGREGATION ANALYSES

**N. Duijvesteijn[1,2], S.A. Clark[2], B.P. Kinghorn[2] and J.H.J. van der Werf[2]**

[1]Hendrix Genetics Research, Technology & Services B.V., Boxmeer, the Netherlands
[2]School of Environmental and Rural Science, University of New England, Armidale, NSW, 2351 Australia

## SUMMARY

This simulation study shows a method which makes more efficient use of pedigree and genomic information to increase the chance to detect genetic disorders. We make use of Geneprob, a program which uses segregation analysis to calculate the genotype probabilities of pedigreed animals. The results show that our method, for a trait with a recessive inheritance pattern, is better in the detection of the region of the causative mutation compared to a method which used allele frequencies of cases and controls only. This method can be used across all pedigreed species.

## INTRODUCTION

In recent years, the detecting of genetic disorders and lethal recessive conditions in livestock populations through the use of genomic tools, has increased (f.e. VanRaden *et al.* 2011 and Derks *et al.* 2017). The defects are mostly spread by intensive use of elite sires which are unknowingly carrier of an autosomal recessive defect. In most populations of cattle where artificial insemination has resulted in a very efficient distribution of the genetic material of superior sires, genetic disorders and lethal recessive conditions have been detected. The success of fine mapping an observed Mendelian genetic disorder requires another approach than that classically used to detect lethal recessive conditions.

A genetic disorder often gives the animal an abnormal phenotype and deprived performance. Accurate recoding of the phenotype by the farmer is essential and often targeted genotyping or sequencing of affected animals and related family members has resulted in successful fine mapping of genetic disorders (e.g. Daetwyler *et al.* 2014). In populations with less routinely genotyping and / or large populations which are extensively managed, success of detection has been compromised. For example, in sheep very few genetic disorders or lethal recessives based on genomic information have been identified. More efficient use of pedigree information and genomic information could increase the chance of detection of genetic disorders.

In this study we show a simple but effective application with the use of Geneprob, a program which uses segregation analysis, to calculate the genotype probabilities of animals within the pedigree, to facilitate the detection of genetic disorders. All animals genotyped within the pedigree are for a GWAS where the phenotype is a linear score derived from genotype probabilities (viz. the probable number of alleles carried). A simulation is done using sheep data to illustrate the application, but the method can be used on any pedigreed population.

## MATERIALS AND METHODS

**Genotypic data.** Genotypes originated from various research flocks (Sheep Genomics, the CRC Information Nucleus Flock, and the MLA Resource Flocks) as well as from industry data collected by sheep breeders. For this study only genotypes of animals from the Merino breed were selected. In total 21,000 Merino sheep were genotyped and imputed up to sequence (Bolormaa *et al.* 2019). For the purpose of this study one chromosome was selected (OAR5) to demonstrate the detection of a recessive causative mutation.

**Detection of the recessive causative mutation.** A single simulation was run on sheep data to illustrate the concept. One SNP on OAR5 was selected to be the causative mutation for an unknown fictive genetic disorder. The minor allele frequency (MAF) of the SNP needed to be between 0.04 and 0.05 to reflect a mutation that is present within a population at low frequency. The SNP was located between 20 and 30 Mb. The randomly selected SNP was Chr5:29170109. The highest linkage disequilibrium (LD) between the causative SNP and a SNP located at the 50K SNP array was 0.333 and the SNP from the 50K SNP array was Chr5:29178193. For an unidentified recessive disease, the genotyping strategy will depend on the available budget and availability of identified cases, but in this study we assume that both cases and controls will be genotyped with the commercially available Illumina ovineSNP50 BeadChip.

From the Merino dataset, 54 cases (homozygous for the recessive allele of SNP Chr5:29170109) were identified. From those 54 cases, we selected 20 cases for our study based on criteria: 1) sires (father of the cases) needed to have more than 1 offspring genotyped, 2) the dam needed to be known and, 3) no full sibs were selected. Besides the 20 cases we selected 10 offspring from sires of cases and 10 random controls which came from the same flock and year as the cases and weren't sires or dams from cases. For this group of 40 sheep, the pedigree was pruned and phenotypes for the disease status was appointed to them. The phenotype code 0 was given to all controls and parents of cases (as they don't have the recessive disease), and all cases were appointed phenotype code 1. In total 31 animals had phenotype 0 and 20 had phenotype 1, all remaining animals from the pedigree got phenotype 8, which means they can be carrier but they are not homozygous for the recessive allele.

**Method using genotype probabilities.** For the scenarios of which we wanted to improve the power by using pedigree information and genotype data, we used the software program Geneprob (Kerr and Kinghorn 1996). It uses segregation analysis to calculate the genotype probabilities of animals within the pedigree. Every animal will get assigned a probability for each genotype class (*aa*, A*a* or AA). Following convergence of Geneprob, the estimated genotype probabilities were expressed as the Most Probable Allele Count (MPAC) using the following equation:

$$MPAC = 0 * p(aa) + 1 * p(Aa) + 1 * p(aA) + 2 * p(AA),$$

where p(aa) is the genotype probability for the genotype class aa, p(Aa) is the genotype probability of the genotype class Aa, p(aA) is the genotype probability of the genotype class aA and p(AA) is the genotype probability of the genotype class AA. The value of MPAC lies between 0 and 2, similar to a SNP genotype. The MPAC was regressed on the SNP genotypes. Similar to a traditional GWAS, $-\log_{10}$(Pvalues) can be plotted to indicate a possible QTL region.

**Scenarios.** Four different scenarios compared in their success to detect the region of the causative recessive mutation. Additionally, 2 scenarios were evaluated to compare the results when very few cases were genotyped (N= 2).

The *first* scenario reflects the traditional approach. In the field 20 cases and 20 controls have been collected and the difference in MAF between cases and controls is compared. Software program PLINK (Chang *et al*. 2015) was used with the Fisher's exact test to generate p-values and $-\log_{10}$(Pvalues) are plotted to indicate a possible QTL region.

The *second* scenario uses the 40 animals with an appointed phenotype status, but pedigree information is used to increase the power of the analysis. Software program Geneprob is used to calculate the MPAC and were regressed on the SNP genotypes to indicate a possible QTL region.

For the *third* scenario, no money was available to genotype cases, but phenotype status was available on the 43 animals from the pedigree as well as 20 controls which were routinely genotyped for the breeding program. Similar to scenario 2, Geneprob was used to calculate the MPAC and regressed on the SNP genotypes.

For the *fourth* scenario, the 20 cases, 20 controls and 43 animals from the pedigree were gen-

otyped and similar to scenario 2, Geneprob was used to calculate the MPAC and regressed on the SNP genotypes.

Two scenarios were added where the traditional method (compare MAF between cases and controls) had 2 cases genotyped and 20 controls which was compared to the method where we used Geneprob to include all available data (2 cases, 20 controls and 43 animals from the pedigree were genotyped) and the MPAC was calculated and regressed on the SNP genotypes.

The 50K data of OAR5 was used for the animals in the different scenarios (1,900 SNPs).

## RESULTS AND DISCUSSION

The SNP in highest LD with the causative mutation shows an incomplete inheritance pattern of the disease (Table 1). If selection was to exclude all animals with genotype 2 (homozygous for recessive allele), four animals would be excluded, while they don't have the recessive disorder.

**Table 1. Count of animals per phenotype class and genotype class**

| Phenotype status | Genotype SNP 50K | | |
|---|---|---|---|
| | 0 | 1 | 2 |
| 0 | 8 | 18 | 4 |
| 1 | 0 | 0 | 20 |
| 8 | 27 | 6 | 0 |

The results of the chromosome-wide association study for each of the four scenario's is shown in Figure 1. The scenario which the largest -$\log_{10}$(Pvalue) was scenario 4 (-$\log_{10}$(Pvalue) =22.6), followed by scenario 2 (-$\log_{10}$(Pvalue)=8.9), then scenario 1 (-$\log_{10}$(Pvalue)=6.7), and scenario 3 has the lowest-$\log_{10}$(Pvalue) with 5.5. In all scenario's the SNP with the highest LD to the causative mutation was indicated. Although in scenario 3, another SNP along the chromosome showed a very similar -$\log_{10}$(Pvalue) and a misidentification of the region could easily have occurred.

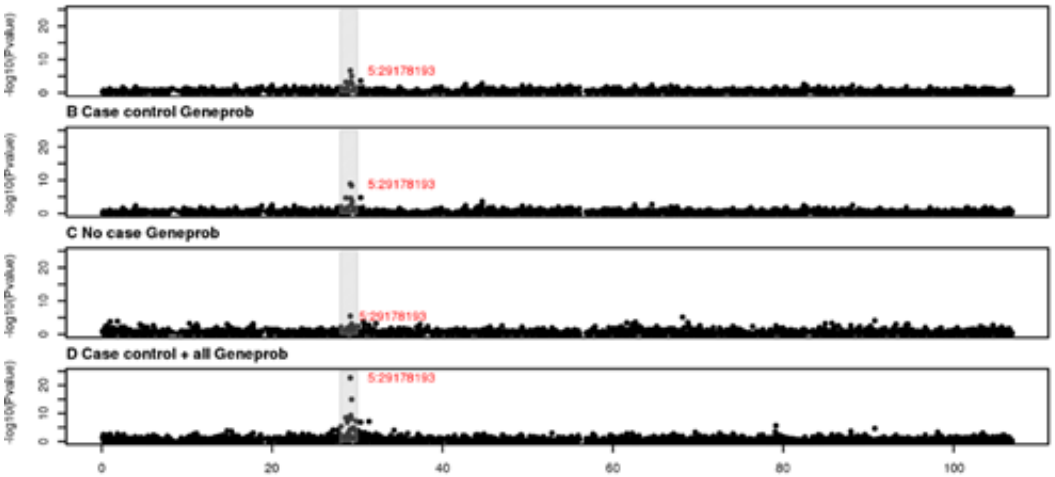


**Figure 1. Chromosome-wide association of OAR 5. A) Association analyses of 20 cases and controls using PLINK. B) Association analyses using Geneprob on all 20 cases and 20 controls. C) Association analyses using Geneprob on only the controls and genotyped animals from the pedigree. D). Association analyses using Geneprob on all cases, controls and genotyped animals from the pedigree**

Also, the 'value' of only genotyping 2 cases has been investigated (Figure 2) and the method using genotype probabilities had increased power compared to the traditional method using Fischer's exact test. The traditional method did not detect the region with the causative mutation (Figure 2A), while the method using Geneprob did detect the region with a clear signal (Figure 2B).

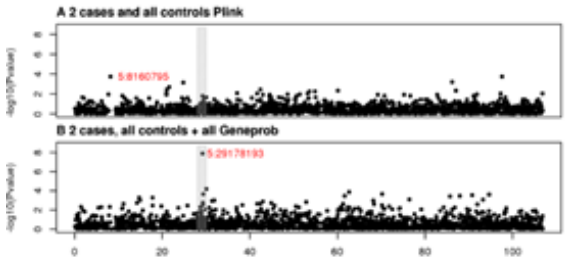


**Figure 2. Chromosome-wide association of OAR 5. A) Association analyses of 2 cases and 20 controls using PLINK. B) Association analyses using Geneprob on 2 cases, 20 controls and genotyped animals from the pedigree**

For the scenarios tested, we have shown the added power to detect a recessive mutation through the use of a segregation analyses and use all available data (pedigree, genotype data and phenotypic information; scenario 4). The results are especially valuable to use for pedigreed species where genotyping is still costly and additional genotyping of affected animals is not covered by available budgets. This study is relatively small and further testing is needed to determine to which extend this method is more beneficial compared to more traditional methods.

## CONCLUSIONS

To conclude, we have demonstrated in this small simulation study that segregation analysis of a trait with a recessive inheritance pattern can lead to considerably power in a GWAS and therefore is better in the detection of the region of the causative mutation compared to a method which used allele frequencies of cases and controls only. We advise at least some cases need to be genotyped to be able to accurately determine the region of the recessive genetic disorder. This method can be used across all pedigreed species and is especially valuable for species where genotyping is still relatively expensive.

## ACKNOWLEGEMENTS

The authors acknowledge the contributions of people from breeders and many CRC participants that contributed to the Sheep CRC Information Nucleus flocks. Also Nasir Moghaddar is acknowledged for his help retrieving data and helpful discussions.

## REFERENCES

Bolormaa S., Chamberlain A. J., Khansefid M., Stothard P., Swan A. A., Mason B., ... and Daetwyler H. D. (2019) *Genet. Sel. Evol.* **51:** 1.

Chang C.C., Chow C.C., Tellier L.C., Vattikuti S., Purcell, S.M. and Lee, J.J. (2015) Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience*. **4**: 7.

Daetwyler H. D., Capitan A., Pausch H., Stothard P., Van Binsbergen R., Brøndum R. F., ... and Esquerré D. (2014) *Nat. Genet.* **46:** 858.

Derks M. F., Megens H. J., Bosse M., Lopes M. S., Harlizius B. and Groenen M. A. (2017) *BMC Genomics* **18:** 858.

Kerr R. J. and Kinghorn B. P. (1996) *J. Anim. Breed. Genet.* **113**: 457.

VanRaden P. M. , Olson K.M., Null D.J. and Hutchison J.L. (2011) *J. Dairy Sci.* **94**: 6153.