Proc. Assoc. Advmt. Anim. Breed. Genet. 26: 94-97

IBS VERSUS IBD – NEW INSIGHTS FROM WHOLE GENOME SEQUENCE DATA

C.L. Warburton[1], R. Costilla[2], M.E. Goddard[3], B.J. Hayes[1] and T.H.E. Meuwissen[4]

[1] Queensland Alliance for Agriculture and Food Innovation, University of Queensland, St. Lucia Qld, 4072 Australia
[2] Cawthron Institute, Nelson 7010, New Zealand
[3] Centre for AgriBioscience, Agriculture Victoria, Bundoora, Vic, 3083, Australia
[4] Faculty of Biosciences, Norwegian University of Life Sciences, Ås Norway

SUMMARY

An implicit assumption in most methods for genomic prediction is that identical by state (IBS) genome regions identified by high density (HD), genome wide markers will also be identical by descent (IBD). With the availability of whole genome sequence on many individuals, we can now directly test this assumption. Here we describe a methodology in cattle that tests IBS haplotypes for IBD within and across breeds, and across sub-species *Bos taurus indicus* (*Bos indicus*) and *Bos taurus taurus* (*Bos taurus*). Within breeds, the number of variant alleles in the sequence data that were different for 250kb IBS haplotypes (IBS for SNP on HD array) were very small. Across *Bos taurus* breeds, this number increased slightly, and was consistent with previous estimates of the number of SNP required for multi-taurus breed predictions (360,000 SNP). Notably, in populations consisting of *Bos indicus* and *Bos taurus* breeds of cattle, our results indicate that approximately 1.5 million SNP are required to ensure 100kb IBS haplotypes are also IBD. These findings suggest that for multi-sub species predictions in cattle, a higher density of markers is required than the current HD arrays.

INTRODUCTION

Many authors have assumed that high density SNP arrays would be sufficient for multi-breed and across breed genomic predictions in cattle (de Roos *et al.* 2009; Hayes *et al.* 2019). This follows earlier work that investigated the conservation of linkage disequilibrium (LD) phase of SNP across breeds, which concluded that, for *Bos taurus* breeds, 300,000 SNP would be required for across breed predictions (de Roos *et al.* 2009; Goddard and Hayes 2009, Porto-Neto *et al.* 2014). With the availability of whole genome sequence (WGS), this hypothesis can be tested in more detail. Animals may share common alleles on a chromosome segment for two possible reasons; the chromosome segments are identical by descent (IBD) where the alleles have been inherited from a common ancestor and thus the SNP-QTL allele phase is likely to be conserved; or haplotypes may be identical by state (IBS) where they share alleles by chance, which may result in spurious associations between QTL and SNP allele phase (Powell *et al.* 2010). Haplotypes from different animals that are IBS on the HD SNP array can be compared for how many SNP they differ by in the whole genome sequence, and this gives some information about the likelihood of the haplotype being identical by descent (IBD). Few studies have investigated the density of SNP that is required to estimate IBD relationships between animals (de Roos *et al.* 2009), especially across the two cattle subspecies. The aim of this study is to determine if high density SNP panels are sufficient to identify IBD relationships, and thus maintain LD phase between SNP and QTL alleles, in genetically distinct breeds of cattle originating from two subspecies, *Bos indicus* and *Bos taurus.*

MATERIALS AND METHODS

Whole genome sequence data consisting of approximately 52 million SNP were obtained from the 1000 Bulls Genomes Run 9 dataset (Hayes and Daetwyler 2019). Quality control in sequence data involved removing non-bi-allelic SNP and SNP that had a minor allele count less than 10. High

density marker panel data (HD) were obtained by subsetting out 640K SNP in the whole genome sequence that corresponded to the BovineHD SNP chip positions (Matukumalli *et al.* 2011). Phasing was performed upon the HD SNP and the WGS SNP genotypes using Eagle version 2.4.1 (Loh *et al.* 2016). For simplicity, the results presented here are arbitrarily for chromosome 25.

To ensure unbiased comparisons between breeds, we standardised the number of animals per breed by randomly selecting 50 animals per breed. If this is not done, it would be difficult to determine if differences in results were due to different LD patterns or just sampling. The data included 3 purebred *Bos taurus* breeds: Hereford (average WGS read depth 11.52, minimum 1.35, maximum 37.97), Charolais (average WGS read depth 10.75, minimum 5.92, maximum 20.74) and Angus (average WGS read depth 11.96, minimum 2.03, maximum 29.02), 1 *Bos indicus* breed: Brahman (average WGS read depth 9.54, minimum 1.85, maximum 47.62); and 1 stabilised composite breed of *Bos indicus* x *Bos taurus* origins: Droughtmaster (average WGS read depth 9.34, minimum 5.77, maximum 19.45).

Haplotypes were generated for both the WGS data and the HD genotypes for all animals using fixed, non-overlapping haplotype windows of 100kb and 250kb, as previously described in Warburton *et al.* (2023). For each haplotype window size, haplotypes generated with the HD markers were compared across all haplotype loci for each pair of animals and identical haplotypes were considered to be IBS. If a haplotype was identified as IBS using the HD marker panel, the equivalent WGS haplotype loci was compared between the same pair of animals. Identical by descent haplotypes were identified by calculating the proportion of WGS SNP different between a haplotype pair. Proportion of WGS SNP different was calculated by counting the number of SNP different on the WGS haplotype, then dividing this number by the total number of SNP within the WGS haplotype window. A higher proportion of SNP different between haplotype pairs suggests that WGS haplotypes are less likely to be IBD at a potential QTL. For each pair of animals with IBS haplotypes, the breed of the animals was recorded as a breed-pair, for example if both animals were Angus then the breed-pair would be Angus_Angus. The proportion of WGS SNP different were averaged across each of the breed-pair comparisons to investigate probability of IBD across breeds for a given haplotype size. Each haplotype within the 100kb and 250kb window was further allocated to a haplotype bin based on the number of HD SNP within the haplotype; <10 (0 – 9 SNP), 10 (10 – 19 HD SNP), 20 (20 – 29 HD SNP), 30 (30 – 39 HD SNP), 40 (40 – 49 HD SNP) and 50 (50 – 59 HD SNP). The proportion of SNP different was averaged across each of the haplotype bins to calculate the average proportion of SNP different for haplotypes with varying density of HD SNP. The average proportion of SNP different for each haplotype window and bin was calculated for each of the breed-pairs and plotted using ggplot in R (Wickham 2016).

There are approximately 3 billion base pairs in the bovine genome (Zimin *et al*. 2009) and the number of SNP required can be calculated by dividing the size of the genome in base pairs by haplotype length in base pairs (hap_size) to obtain the total number of haplotypes required to span the genome. Once the number of HD SNP required per haplotype to identify and IBD segment (nSNP_hap) is obtained this number can be multiplied by the number of haplotypes required to span the genome, giving the estimated total number of SNP required to identify IBD haplotypes in multi-breed beef populations (see Equation 1).

$$Number\ of\ SNP\ required = \frac{3{,}000{,}000{,}000}{hap\_size} * nSNP\_hap \qquad \textit{Equation 1}$$

It is important to consider sequencing error rate, as this will affect both the number and proportion of SNP different per haplotype window. Daetwyler *et al*. (2014) showed that the concordance between HD SNP genotypes and sequence genotypes in the 1000 Bulls data was approximately 99% after BEAGLE correction, and the rate of opposing homozygotes was, on average, 0.6%. Therefore, a threshold of 1% of SNP different between WGS haplotypes would allow for this estimated rate of genotyping error in the 1000 Bulls data. As such, in this study, whole

genome sequence haplotypes that have less than or equal to 1% SNP different between animals at a loci will be considered to be IBD for the purposed of calculating the density of SNP required to identify IBD relationships in multi-breed and multi-subspecies populations.

RESULTS AND DISCUSSION

Figure 1 shows the proportion of WGS SNP different averaged across haplotype bin for each haplotype window and breed-pair comparison. There were no 250kb haplotypes with <10 SNP or 20 SNP in a bin, and as such, comparisons between proportion of SNP different are unable to be drawn for this bin. As the number of HD SNP in a haplotype bin increases there is a trend for the proportion of WGS SNP different to decrease across all haplotype windows and breeds, suggesting that the density of SNP within a haplotype will affect the accuracy of IBD detection from IBS haplotypes. Haplotypes with lower SNP density are more likely to be identified as IBS because there are few SNP per haplotype that are being used to identify haplotypes that are IBS. However, by increasing the number of SNP within a haplotype window, it is more likely that haplotypes that are identified as IBS using marker panel haplotypes will be IBD in WGS.

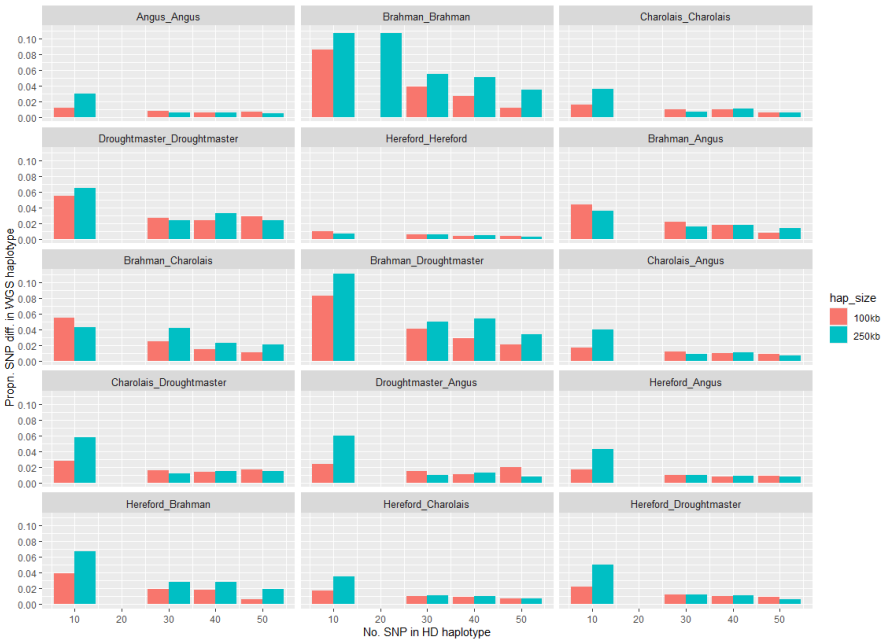

Figure 1. Proportion of SNP different (y-axis) in whole genome sequence haplotype as a function of the number of SNP in the high density marker panel genotype (x-axis) for each breed-pair comparison, across the 100kb and 250kb haplotype windows

The optimal number of SNP per haplotype and haplotype window size can be identified by the haplotypes with a proportion of WGS SNP different that is less than 1% across all breed pairs. In the results presented, the 250kb haplotype window with 30 HD SNP resulted in a 1% or less proportion of WGS SNP different across the multi-breed *Bos taurus* comparisons. Therefore using Equation 1, the estimate of the number of SNP required to identify IBD relationships that are likley to maintain LD phase between SNP and QTL in multi-breed *Bos taurus* populations is 360,000. This finding is corroborated by earlier studies that have demonstrated that 300,000 SNP are required to maintain LD phase between SNP in QTL in *Bos taurus* populations (de Roos *et al*. 2009; Goddard

and Hayes 2009). In this study, the 100kb haplotype window with 50 HD SNP resulted in a 1% proportion of WGS SNP different in the multi-subspecies (*Bos indicus* and *Bos taurus*) breed-pair comparisons. Using Equation 1, it is estimated that 1.5 million evenly spaced SNP would be required to achieve haplotypes are IBD in multi-breed and multi-subspecies beef populations. Currently, the Bovine HD marker panel consists of approximately 728,000 SNP, so this estimate is double the number of SNP that are currently available on the HD marker panel.

CONCLUSIONS

In multi-breed and multi-subspecies populations of beef cattle, high density SNP are required to ensure IBS haplotypes are IBD, and therefore carry the same QTL allele. Currently the Bovine HD SNP panel consists of ~728,000 SNP, which was sufficient to identify IBD haplotypes between *Bos taurus* animals. However, higher density of SNP are required when the population consists of *Bos indicus* and *Bos taurus* animals. From our analysis we conclude that approximately 1.5 million evenly spaced SNP are required to identify IBD haplotypes and maintain SNP-QTL LD across these multi-subspecies and multi-breed populations.

ACKNOWLEDGEMENTS

The authors would like to acknowledge and sincerely thank to 1000 Bulls Genomes Project Consortium for the data provided in this manuscript.

REFERENCES

Daetwyler H.D., Capitan A., Pausch H., Stothard P., van Binsbergen R., Brøndum R.F., Liao X., Djari A., Rodriguez S.C., Grohs C., Esquerré D., Bouchez O., Rossignol M-N., Klopp C., Rocha D., Fritz S., Eggen A., Bowman P.J., Coote D., Chamberlain A.J., Anderson C., VanTassell C.P., Hulsegge I., Goddard M.E., Guldbrandtsen B., Lund M.S., Veerkamp R.F., Boichard D.A., Fries R. and Hayes B.J. (2014) *Nat. Genet.* **46**: 858.

Goddard M.E. and Hayes B.J (2009) *Nat. Rev. Genet.* **10**: 381.

Hayes B.J., Corbet N.J., Allen J.M., Laing A.R., Fordyce G., Lyons R., McGowan M.R. and Burns B.M. (2019) *J. Anim. Sci.* **97**: 55.

Hayes B.J. and Daetwyler H.D. (2019). *Annu. Rev. Anim. Biosci.* **7**: 89.

Loh P.R., Danecek P., Palamara P.F., Fuchsberger C., Reshef Y.A., Finucane H.K., Schoenherr S., Forer L., McCarthy S., Abecasis G.R., Durbin R. and Price A.L. (2016) *Nat. Genet.* **48**: 1443.

Matukumalli, L. K., Schroeder S., DeNise S.K, Sonstegard T., Lawley C.T., Georges M., Coppieters W., Gietzen K., Medrano J.F, Rincon G., Lince D., Eggen A., Glaser L., Cam G., and Van Tassel C. (2011) *Illumina Inc.*

Porto-Neto L.R., Kijas J.W. and Reverter A. (2014) *Genet. Sel. Evol*. **46**: 22.

Powell J.E., Visscher P.M. and Goddard M.E. (2010) *Nat. Rev. Genet.* **11**: 800.

de Roos A.P.W., Hayes B.J. and Goddard M.E. (2009). *Genetics.* **183**: 1545.

Warburton C.L., Costilla R., Engle B.N., Moore S.S., Corbet N.J., Fordyce G., McGowan M.R., Burns B.M. and Hayes B.J. (2023). *Heredity.* **131**: 350.

Wickham H. (2016) (Springer-Verlag New York) https://ggplot2.tidyverse.org.

Zimin A.V., Delcher A.L., Florea L., Kelley D.R., Schatz M.C., Puiu D., Hanrahan F., Pertea G., Van Tassell C.P., Sonstegard T.S., Marçais G., Roberts M., Subramanian P., Yorke J.A. and Salzberg S.L. (2009). *Genome Biol.* **10**: R42.