Evaluating haplotype diversity within and between Australian sheep breeds

S.J. Goodswen[*, †], C.Gondro[†], H.N. Kadarmideen[*], and J.H.J van der Werf [†‡]

Introduction

Genomic data can be used to predict differences in phenotype or breeding values between individuals, using linkage disequilibrium (LD) between marker and putative quantitative trait loci (QTL). So far, most genome association studies are based on genotypes at individual loci, and although this has given reasonable accuracies of predicting breeding value (Roos et al. (2008)), there is limited evidence of underlying QTL effects being consistant at the basis of such predictions. One problem with single markers in dense genotypic data is that different loci can easily be in LD by random chance, and SNPs apparently linked to QTL effects may have limited predictive ability in data from individuals that are genetically less related. When using ordered genotypes and information on haplotype similarity, the power of predicting QTL effects can be increased. If two individuals share the same extended haplotype over the same genomic region, the chance that they carry the same marker-QTL allele relationship *by descent* is much higher. In this paper, we explore the extent by which various haplotype lengths are shared within and between 4 Australian sheep breeds. We count how many haplotypes are present in the breed for a given number of loci, and how many of these haplotypes are shared between breeds. Such statistics will become important when assessing identity by descent (IBD) probabilities and how well these can be separated from identity by state (IBS) probabilities.

Material and methods

Data for the haplotype analysis was provided by the CRC for Sheep Industry Innovation. The data consisted of genotypes for 3,001 animals obtained via the Illumina 50k Ovine Bead Chip. The genotypes of SNPs were phased using the program fastPhase (Scheet and Stephens (2006)). There were 48,640 SNPs distributed across 26 chromosomes. The animals were progeny of 159 industry sires: 34 Poll Dorset sires, 21 White Suffolk, 35 Border Leicester, and 69 Merino sires. For the dams, 2,500 (83.3%) were pure Merino and 501 (16.7%) were a Merino-Border Leicester cross. The sire haplotypes and maternal haplotypes were separated into groups and Table 1 shows the number of animals grouped according to their sire breed and their dam breed.

We divided the genome into haplotype block sizes of 3, 5, and 10 SNPs. For an *n* SNP block there are 2^n possible haplotypes. Figure 1 shows the haplotype count for the 4 sire breed

[*] CSIRO Livestock Industries, Davies Laboratory, University Drive, Townsville, QLD 4810, Australia

[†] School of Environmantal and Rural Science, University of New England, Armidale, NSW 2351, Australia

[‡] Cooperative Research Centre for Sheep Industry Innovation (Sheep CRC), Armidale, NSW 2351, Australia

groups for the first (out of 1,831) 3-SNP block on paternal chromosome #1. It can be noted that there is a correlation between the frequencies and SNP patterns. SNP patterns “211”, “212”, and “222” do not exist in this region of the chromosome in any breed.

Table 1: The number and breed of animals allotted to groups.

Group #	Breed	Grouping Criteria[++]	Sire Type	No. of animals in group
1	Poll Dorset	Sire	Terminal	735
2	White Suffolk	Sire	Terminal	507
3	Border Leicester	Sire	Maternal	756
4	Merino	Sire	Merino	1003
			Total	***3001***
5	Pure Merino	Dam		2500
6	Border Leicester *Merino	Dam		501
			Total	***3001***

[++] Sire = grouped according to animal’s sire breed, Dam=grouped according to animal’s dam breed.

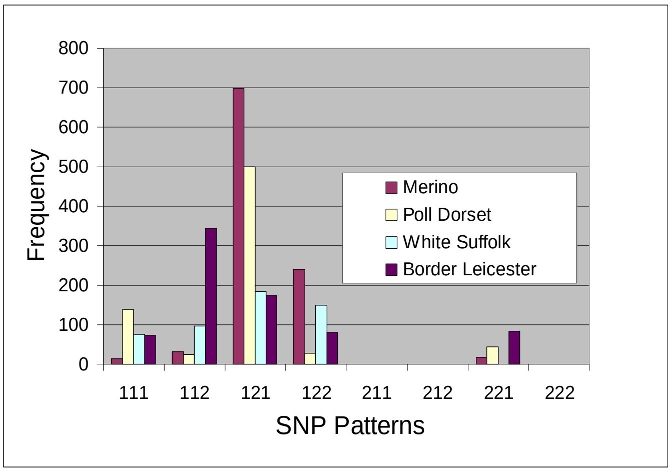

Figure 1: The number of animals per first 3-SNP block on chromosome #1 counted for 4 sire breed groups

To determine haplotype diversity: (1) we calculated the average number of different haplotypes per block. (2) Converted counts into proportions and calculated the standard deviation to assess the spread of the haplotype frequencies. (3) Calculated Euclidian distance measures between the breeds: $\sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$

Where p_i = haplotype frequency at SNP block for breed 1; q_i= haplotype frequency at SNP block for breed 2; and n = number of SNP blocks and (4) Estimated haplotype similarity across breeds. For one breed at a time, we take all haplotypes occurring in a SNP block. We then, within another breed, count how many times the same haplotype occurs in a SNP block at the same chromosomal location. From the counts, the probability that an animal within the comparison breed has the same haplotype is calculated. Calculations are repeated for each SNP block along the chromosome and an average probability per SNP block is determined.

Results and discussion

A pair-wise LD analysis was completed by the International Sheep Genomics Consortium (Raadsma et al, in press, www.sheephapmap.org). So, whilst we acknowledge the importance of pair-wise LD, we used muliple SNP blocks with frequency counts. Also, r^2 is uninformative in 2 situations: (1) In some instances r^2 can be the same between the marker and QTL in different breeds, even though the phase may have reveresed (Rocha et al. (2002)) and (2) it provides no clues to help localize the QTL. Descriptive statistics for the haplotype frequency counts are given in Table 3.

Table 3: Haplotype diversity within breeds for 3, 5 and 10 SNP blocks

Group #	Breed	Mean haplotype count per SNP block for chr 1**			Mean haplotype count per SNP block per chr ++			Standard Deviation for means for all chromosomes		
		SNP block size								
		3	5	10	3	5	10	3	5	10
		Paternal Chromosomes								
1	Poll Dorset	5.6	11.3	27.5	5.6	11.4	28.8	0.08	0.36	1.52
2	White Suffolk	5.7	11.1	24.3	5.6	11.0	24.2	0.12	0.42	1.52
3	Border Leicester	5.6	11.9	33.1	5.7	12.0	34.7	0.12	0.44	2.53
4	Merino (sire)	6.9	18.3	73.9	6.9	18.2	73.8	0.08	0.55	3.43
5	Merino (dam)	7.1	20.8	108.6	7.1	21.2	111.2	0.08	0.62	5.37
6	Border Leicester * Merino	6.0	12.9	32.1	6.0	12.9	32.6	0.09	0.45	1.79
Maximum count per block		8	32	1024	8	32	1024			
		Maternal Chromosomes								
1	Poll Dorset	7.3	21.6	101.1	7.4	21.9	106.4	0.05	0.57	6.47
2	White Suffolk	7.1	19.8	80.1	7.2	20.1	84.6	0.08	0.59	5.30
3	Border Leicester	7.6	23.3	115.9	7.6	23.5	122.6	0.06	0.59	7.60
4	Merino (sire)	7.7	24.8	144.7	7.7	24.9	147.6	0.04	0.50	7.57
5	Merino (dam)	7.7	56.7	199.0	7.7	27.3	206.7	0.03	0.59	10.70
6	Border Leicester * Merino	7.1	19.3	72.7	7.2	19.7	76.1	0.07	0.56	4.72

** Computed along chromosome #1 ++ Computed along all 26 chromosomes

The results in Table 3 highlight that the larger the SNP block size the more informative it becomes as it distinguishes the true haplotypes from frequent SNP patterns that occur by chance. The 10 SNP block indicates that there is less haplotype diversity within the Poll Dorset and White Suffolk breeds than within the Merino breed. This is consistent with the LD measures and estimates of effective population sizes found by Raadsma et al. The results of this study, however, have a more statistical implication for genomic selection. To estimate haplotype effects, it is relevant to know how many phenotypes are available per haplotype (per breed) and whether these haplotypes exist in different populations. In designing an association study, one could optimize the design by sampling individuals as much as possible across all existing haplotypes.

The haplotype diversity between breeds using Euclidian distance measure for chromosome #1: Merino-White Suffolk = 15403.69, Merino-Border Leicester = 17193.37, Merino-Poll Dorset = 15572.26, White Suffolk-Border Leicester = 15526.71, White Suffolk-Poll Dorset = 11257.13, Border Leicester-Poll Dorset = 17061.38. The breeds White Suffolk and Poll Dorset are the least diverse from each other; and Border Leicester is the most diverse.

Table 4 shows the probability that a marker-QTL located on chromosome #1 of one breed will persist in another breed using 3 and 10-SNP block haplotypes for comparison. Larger SNP blocks have dramatic reduction in probablity of carrying the same QTL than smaller SNP blocks across breeds due to high chances of recombinations in larger distances.

Table 4: Haplotype similarity across breeds using 3 and 10-SNP block on chromosome #1 (values are in percentages)

Sire breed with Marker-QTL[++]	Comparison sire breed [**]							
	Poll Dorset		White Suffolk		Border Leicester		Merino	
	3	10	3	10	3	10	3	10
Poll Dorset	69.9	2.7	62.0	0.9	59.4	0.7	67.3	1.2
White Suffolk	62.0	1.1	70.7	2.4	60.0	0.7	68.3	1.1
Border Leicester	59.4	0.7	60.0	0.7	70.8	3.2	68.3	1.5
Merino	67.3	1.2	68.3	1.1	68.3	1.5	86.6	7.2

[++] Breed carrying the haplotype containing a presumed marker-QTL

[**] Breed carrying same haplotype containing marker-QTL alleles

Conclusion

We have shown frequency counts of haplotypes of various lengths as a simple method to evaluate overall haplotype diversity. The results reveal that the breed Poll Dorset has the least haplotype diversity within the breed followed by White Suffolk, Border Leicester, and Merino. The breeds White Suffolk and Poll Dorset are the least diverse from each other; and Border Leicester is the most diverse. Finally, estimation of haplotype similarity across breeds can provide us with an expectation as to whether a SNP marker allele can predict QTL alleles across breeds or only within breeds.

Acknowledgments

Office of the Chief Executive of CSIRO for postgraduate scholarship for Stephen Goodswen.

References

Rocha, J. L., Pomp, D., and Van Vleck, L. D. (2002). *Quantitative Trait Loci: Methods and Protocols* (Humana Press Inc, Totowa) 311-346.

Roos, A. P. W. d., Hayes, B. J., Spelman, R. J. et al. (2008). *Genetics.* 179:1503-1512.

Scheet, P., and Stephens, M. (2006). *Am. J. Hum. Genet.,* 78:629-644.