GENETIC DIVERSITY AND POPULATION STRUCTURE OF FOUR SOUTH AFRICAN SHEEP BREEDS

L. Sandenbergh[1,2], S.W.P. Cloete[1,3], R. Roodt-Wilding[2], M.A. Snyman[4], and A.E. Van der Merwe[2]

[1]Directorate Animal Sciences: Elsenburg, Private Bag X1, Elsenburg, 7607, South Africa
[2]Department of Genetics, Stellenbosch University, Private Bag X1, Matieland, 7602, South Africa
[3]Department of Animal Sciences, Stellenbosch University, Private Bag X1, Matieland, 7602, South Africa
[4]Grootfontein Agricultural Development Institute, Private Bag X 529, Middelburg EC 5900, South Africa

SUMMARY

Prior knowledge of the genetic diversity, extent of linkage disequilibrium (LD) and population structure is necessary to determine the sample size and number of SNPs necessary to ensure sufficient power of detection in genome-wide association studies (GWAS) and genomic prediction. The OvineSNP50 chip was used to genotype Dorper, Namaqua Afrikaner (NA), South African Mutton Merino (SAMM) and 2 flocks of South African Merino to determine the genetic diversity, differences in LD across breeds and population differentiation. The NA samples exhibited the least number of polymorphic loci and was also the least genetically diverse breed tested. The South African Merino samples exhibited high levels of diversity comparable to results of international Merinos. The NA samples exhibited the longest stretches of LD in comparison to the 3 other breeds, while the Merino had the most rapid decay in LD. Dorper and SAMM samples exhibited intermediate LD length in comparison to the 2 aforementioned breeds. A principal component analysis (PCA) indicated 4 distinct clusters in the data representing the 4 breeds. The inclusion of additional SAMM and other Merino-based breed samples may aid in increasing the resolution and clearly defining breeds and subtypes.

INTRODUCTION

Genomic prediction and GWAS rely on sufficient marker coverage of the genome and a representative sample cohort (Goddard and Hayes 2009). Estimates relating to the genetic diversity, extent of LD and population differentiation is vital in selecting representative samples and determining the number of markers required for genomic prediction and GWAS (Goddard and Hayes 2009; Zhang *et al*. 2012; Kijas *et al.* 2014).

The South African Merino is the primary fine wool producing breed in South Africa and is also utilised for meat production. The SAMM was originally developed from the German Merino and has become the major dual-purpose breed in South Africa (Cloete and Olivier 2010; Schoeman *et al*. 2010). The Dorper, a 50-50 composite of the Dorset Horn and Persian breeds, is the major meat producing breed in the country (Cloete and Olivier 2010). The NA is a hardy, fat-tailed sheep indigenous to South Africa and is primarily maintained for conservation purposes (Schoeman *et al*. 2010; Qwabe *et al.* 2013). The breed is considered endangered with <1000 breeding ewes and <20 breeding rams remaining (FAO 2000; Qwabe *et al.* 2013). Although the genetic diversity and population structure of South African sheep breeds have been explored previously using microsatellite markers (Soma *et al*. 2012; Qwabe *et al.* 2013), a fine-scale investigation is necessary to confirm the genetic diversity and the breed structure, and determine the extent of LD for future genomic studies (Kijas *et al*. 2012). The current study used the OvineSNP50 chip to genotype 160 Dorper, NA, South African Merino and SAMM samples to investigate differences in genetic diversity, LD and population differentiation across the breeds and sampling groups.

MATERIALS AND METHODS

Samples and genotyping. The Dorper (n=20), NA (n=20) and SAMM (n=20) samples were obtained from a resource flock on the west coast of the Western Cape Province of South Africa at the Nortier Research Farm. The South African Merino samples were obtained from the resource flocks maintained at Cradock (n=50) and Grootfontein (n=50) in the Eastern Cape Province. Blood samples were obtained through venipuncture of the jugular vein and stored between -20°C and-80°C. Samples were thawed and applied to bloodcards for transport. Genotyping was done with the OvineSNP50 beadchip at GeneSeek Inc. (Lincoln, NE, USA).

Data analysis. GenomeStudio Software v. 1.0 (Genotyping Module, Illumina) was used to call genotypes from SNP intensity data and to ensure the stringency of quality control parameters. The following quality control measures were implemented: >0.25 GenCall score; >0.5 GenTrain score; >0.01 minor allele frequency (MAF); >0.95 call rate and a sample call rate >0.95 across all samples. Samples with more than 10% missing data were excluded. Genotype data that met the quality control criteria were used to determine the number of polymorphic loci and the MAF distribution for the 5 respective sampling groups and an additional group comprising 20 Cradock and 20 Grootfontein Merino samples. The observed heterozygosity and inbreeding coefficient (F_{IS}) was calculated for each group in PLINK v.1.07 (Purcell *et al.* 2007). Allelic richness (Ar) and private allelic richness (Par) was determined using ADZE v. 1.0 (Szpiech *et al.* 2008). As SNP ascertainment bias may inflate LD values, LD was calculated for subsets of SNP data pruned within each breed and across breeds. The --indep-pairwise 50 5 0.5 command in PLINK was used to calculate pairwise LD within a 50 SNP window and remove one SNP from a pair where the LD exceeds 0.5 before moving on 5 SNPs and repeating the procedure. Linkage disequilibrium (r^2) was calculated for all SNP pairs remaining after LD pruning using the --r2 command. A principal component analysis (PCA) was conducted in the R package (R Core Team 2015), adegenet v. 1.4-2 (Jombart and Ahmed 2011) to identify population structure within and between the sampling groups and to identify potential outliers. Equal sample numbers (n=20) from each group were included in the PCA. Loci were pruned across all samples and the MAF cut-off was increased to 0.1 to mitigate the possible effect of SNP ascertainment bias. File formatting was conducted in R, PLINK or PGDspider v. 2.0.8.0 (Lischer and Excoffier 2012).

RESULTS AND DISCUSSION

From the total of 160 samples, 16 samples (2 from the Cradock Merino, 13 from the Grootfontein Merino and 1 from the SAMM sampling groups) were excluded. The remaining samples had an average call rate of 99.72% and 91% (of the total of 54 241) of the SNPs met quality control measures (Table 1). The Merino samples (Cradock and Grootfontein) were polymorphic for approximately 89% of SNPs, while NA samples were polymorphic for only 69% of SNP loci. The Dorper and SAMM samples were intermediate to these values, at 83% and 81%, respectively. The MAF distribution of the Merino, Dorper and SAMM were relatively similar and most loci exhibited MAFs of more than 30%. In contrast, the NA samples exhibited a large number of non-polymorphic loci and an equal distribution in the number of polymorphic loci across the MAF range. The NA samples also had the lowest allelic richness, private allelic richness and observed heterozygosity in comparison to the other 3 breeds (Table 1). These low levels of genetic diversity in the NA have also been observed with the microsatellite-based studies (Qwabe *et al.* 2013) and OvineSNP50 genotype information (Kijas *et al.* 2012).

Table 1. Genetic diversity estimates of the 5 sampling groups and a combination sample consisting of an equal number of Cradock (n=20) and Grootfontein (n=20) Merino samples. NA: Namaqua Afrikaner; SAMM: South African Mutton Merino, n: number of samples, MAF: Minor allele frequency; Pn: Percentage of polymorphic loci; SE: Standard error; Ar: Allelic richness; Par: Private allelic richness; He: Observed heterozygosity; F_{IS}: Inbreeding coefficient.

Sample group	n	Loci with MAF<0.01	Pn	Ar (SE)	Par (SE)	He	F_{IS}
NA	20	11921	69.20	1.75 (0.001)	0.007 (0.0003)	0.28	0.25
Dorper	20	4026	83.55	1.89 (0.001)	0.012 (0.0004)	0.34	0.11
SAMM	19	5174	81.16	1.88 (0.001)	0.012 (0.0003)	0.33	0.12
Cradock Merino	48	1120	87.12	1.99 (0.001)	0.014 (0.0004)	0.36	0.05
Grootfontein Merino	37	1120	84.43	1.99 (0.001)	0.011 (0.0003)	0.35	0.08
Merino (combined)	40	1120	89.01	1.94 (0.001)	0.012 (0.0003)	0.35	0.06

The extent of LD varied according to the manner in which LD pruning was applied to the dataset (Figure 1). The unpruned dataset exhibited LD over longer stretches, while pruning within each breed markedly reduced the LD values between SNPs. A less extreme reduction in the extent of LD was observed when SNPs were pruned across breeds. In all datasets, the Merino, followed by the Dorper and SAMM displayed the most rapid decay in LD. The NA samples had the longest stretches of LD overall. High levels of genetic diversity and LD decay over short distances has been reported for international Merino samples and may be a consequence of the large effective population size and variation maintained within the breed (Kijas *et al*. 2012; 2014)

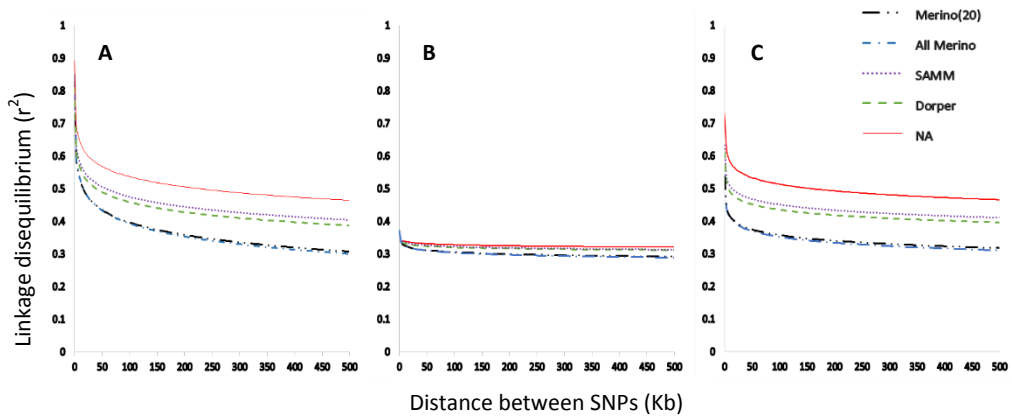

Figure 1. Linkage disequilibrium (r^2) determined for 4 South African sheep breeds prior to pruning SNPs in strong linkage disequilibrium (LD) (A); LD pruning within each breed (B); and LD pruning across all samples (C). Merino (20): Cradock (n=20) and Grootfontein Merino (n=20) samples; All Merino: Cradock (n=48) and Grootfontein Merino (n=37) samples; SAMM: South African Mutton Merino; NA: Namaqua Afrikaner.

The first principal component accounted for 12.29% of the variation in the sample, while the second and third principal components accounted for 7.93% and 6.84%, respectively. Across the first principal component, the NA and Dorper samples clustered separately while substantial overlap was seen between the other sampling groups. The third principal component separated the 4 breeds tested into separate clusters. The Grootfontein and Cradock Merino samples remained clustered together across all principal components. Inclusion of additional Merino samples (48

Cradock Merino, 37 Grootfontein Merino) and the full set of (unpruned) SNPs, resulted in the SAMM samples clustering separately from the Grootfontein and Cradock Merino for all principal components (data not shown).

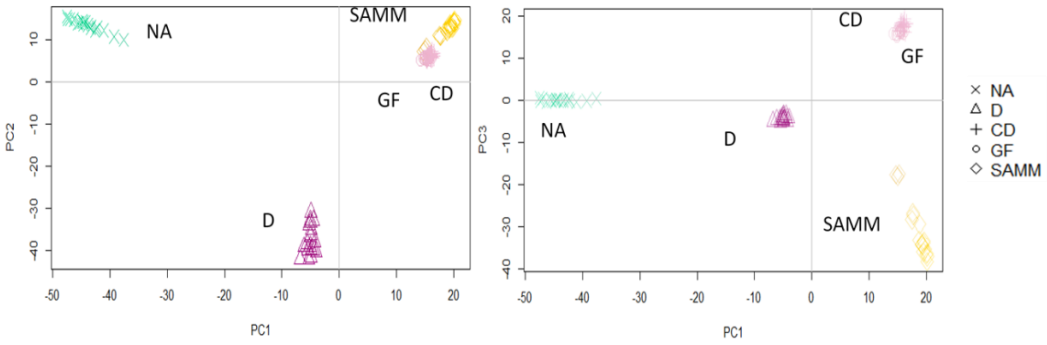

Figure 2. Principal component analysis of 4 South African sheep breeds from 5 sampling groups. (NA: Namaqua Afrikaner; D: Dorper; CD: Cradock Merino; GF: Grootfontein Merino; SAMM: South African Mutton Merino).

The NA samples exhibited large stretches of LD and the least genetic diversity of the breeds tested. Fewer SNPs would therefore be necessary to achieve the same level of coverage of the NA genome than more diverse breeds. Fewer individuals may also be needed to establish a representative sampling cohort for this breed. Despite SNP pruning, the effect of SNP ascertainment bias should still be considered when interpreting whole-genome SNP data from NA as indigenous breeds had limited representation during SNP discovery (Clark *et al*. 2005). The South African Merino samples exhibited high levels of genetic variability and a rapid decay in LD that were comparable to results of international Merino breeds (Kijas *et al*. 2012; 2014). A relatively large sample cohort and a large number of SNPs will be required of future genomic studies to adequately capture all variation contained in this breed. The 4 breeds tested appear to be genetically distinct, however, the inclusion of additional SAMM samples may elucidate the relationship between the SAMM and South African Merino further. Scope exists for further studies that include additional South African sheep breeds, such as the Dormer and Dohne Merino, to clarify the relationship between the South African sheep breeds.

REFERENCES

Clark A.G., Hubisz M.J., Bustamante C.D., Williamson S.H. and Nielsen R. (2005) *Genome Res*. **15**: 1496.
Cloete S.W.P. and Olivier J.J. (2010) The international sheep and wool handbook. Cottle D.J. (ed.), pp. 95.
FAO (2000) World watch list for domestic animal diversity, 3rd ed. Scherf B. (ed.), Rome.
Goddard M.E. and Hayes B.J. (2009) *Nature Reviews Genetics* **10**: 381.
Jombart T. and Ahmed I. (2011) *Bioinformatics* **27**: 3070.
Kijas J.W., Lenstra J.A., Hayes B., Boitard S., Porto Neto L.R., *et al.* (2012) *PLoS Biology* **10**: e1001258.
Kijas J.W., Porto-Neto L., Dominik S., Reverter A., Bunch R., *et al.* (2014) *Anim. Genet*. **45**: 754.
Lischer H.E.L. and Excoffier L. (2012) *Bioinformatics* **28**: 298.
Purcell S., Neale B., Todd-Brown K., Thomas L., Ferreira M.A.R., *et al.* (2007) *Am. J. Hum. Genet*. **81**: 559.
Qwabe S.O., Van Marle-Köster E. and Visser C. (2013) *Trop. Anim. Health Prod.* **45**: 511.
R Core Team (2015). R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.
Schoeman S.J., Cloete S.W.P. and Olivier J.J. (2010) *Livest. Sci*. **130**: 70.
Soma P., Kotze A., Grobler J.P. and Van Wyk J.B. (2012) *Small Rum. Res*. **103**: 112.
Szpiech Z.A., Jokabsson M. and Rodenberg N.A. (2008) *Bioinformatics* **24**: 2498.
Zhang H., Wang A., Wang S. and Li H. (2012) *J. Anim. Sci. Biotech* **3**: 26.