SHEEP PHYLOGEOGRAPHY AND DOMESTICATION AS INFERRED FROM COMPLETE GENOME SEQUENCES

M. Naval-Sanchez, S. McWilliam, A. Reverter, M. Perez-Enciso, N. J. Hudson, J. Kijas

CSIRO Agriculture Flagship, 306 Carmody Road, St Lucia, QLD 4067, Australia

SUMMARY

The phenotypic diversity present within domestic sheep breeds is the outcome of direct human selection in behavioural as well as productive traits such as meat, milk or wool. To explore the genomic diversity of domestic sheep breeds we made use of whole-genome sequences of 68 domestic sheep sampled from five major geographic regions: Africa, the Americas, Asia, Europe and the Middle East. SNP calling identified a total of 26 million variants, ranging from 22 to 25 million SNPs per individual. The Asian and African animals examined contain a higher rate of heterozygosity (32.3% and 28.4%) compared to individuals from Europe (19.9%), the Americas (19.4%) or UK (20.4%). This is most likely a consequence of the sheep reference genome being from a European breed. In the future, we aim to compare these genomes against wild ovids to give further insight into the genomic mechanisms underlying domestication across breeds as well as their functional implications.

INTRODUCTION

Specific to sheep, the process of domestication was probably initiated around 11,000 years ago to facilitate stable access to meat and subsequently 5,000 years ago human mediated selection for wool and milk production (Chessa *et al.* 2009). Early consequences of animal domestication are likely to have included changes in stature, coat pigmentation, horn morphology in ruminants and docility (Zeder 2008). Genome-wide patterns of variation have proven highly informative for detecting genes under selection, with recent examples including loci controlling digestion (Axelsson *et al.* 2012), fertility (Larkin *et al.* 2012), stature and pigmentation (Rubin *et al.* 2010, Rubin *et al*. 2012) and horn development (Kijas *et al.* 2012).

In this preliminary analysis of 68 domestic sheep genomes, we compare patterns of genetic diversity and genetic divergence between individuals sampled from major geographic regions. This represents a first step in the reconstruction of the early evolutionary history of domestic sheep and the identification of loci involved in shaping the phenotypic diversity of today's modern breeds.

MATERIALS AND METHODS

Samples. Sixty-eight domestic sheep were sequenced using Illumina paired-end technology. Of these, 46 animals were selected from the ISGC Breed Diversity Hapmap experiment genotyped using the SNP50 Beadchip (Kijas et al. 2012), 6 animals were previously used for SNP discovery in the construction of the SNP50 BeadChip and CNV detection and the remaining animals werw investigated for the first time in this work. The selected animals belong to 42 different breeds drawn for Asia (n=12), Africa (n=6), the Middle East (n=13), the Americas (n=8), the United Kingdom (n=7) and continental Europe (n=22).

Alignment and variant calling. Reads from each sample were mapped against the sheep reference assembly v3.0 (available at http://www.livestockgenomics.csiro.au/sheep/) with BWA (Li and Durbin 2009) using default parameters. Duplicate removal and sorting were performed using samtools v.0.1.18 (Li et al. 2009). Genotypes were called for each animal separately using samtools mpileup. A series of filters were applied to prune low quality variants, including minimum depth of coverage (6 fold), map quality score (> 20) and base pair quality (>20).

Variants from each animal were then combined to produce a merged VCF file. This included examination to distinguish between positions with insufficient data to assign a genotype from those that were homozygous for the allele present in the reference genome.

Sequence based diversity estimates. To examine genomic differences among breeds and to infer population diversitys we made use of two metrics, namely principal components analysis (PCA) based on genetic diversity (heterozygosity level) and the compression efficiency (CE) algorithm (Hudson et al. 2014).

CE algorithm: In brief, CE is a new measure that exploits the order and proportion of heterozygosity in SNP genotypes. First, genotypes are encoded in numerical values 0's 1's or 2's for detected in bi-allelic SNPs across samples. Second, CE is calculated as CE= $(S_b\text{-}S_a) / S_b$, where S_b and S_a correspond to the size in bytes of the SNP genotype data before and after compression by the command gzip in UNIX, respectively. This measure is a proxy for the minimum amount of information required to reproduce a dataset. CE has shown to unravel genomic patterns such as phylogeography in diverse populations including human (Hudson et al. 2014).

Fixation Index (Fst): Fst to calculate the genetic distance between populations was calculated as in Weir and Cockerman 1984 paper, using vcftools *–weir-fst-pop* option.

RESULTS AND DISCUSSION

Whole genomes of 68 domestic breeds from different geographical regions Africa (n=6), Americas (n=8), Asia (n=12), Europe (n=22), Middle East (n=13), United Kingdom (n=7) were sequenced at an average depth of 8X in all groups (7.6-8.2) (Table1). SNP calling resulted in the discovery of a total of 26 million SNPs across the collection of animals. The average number of variants observed was calculated after grouping individuals into the geographically defined groups. The highest average number was identified in European animals, however this reflects the larger number of genomes sequenced. Next, we examined the percentage of heterozygous SNPs between populations and discovered that Asian and African populations contain a higher rate of heterozygosity (32.3 and 28.4) compared to breeds in Europe (19.9), Americas (19.4) and UK (20.4). This is most likely a consequence of the sheep reference genome reference being from a European breed (Jiang *et al.* 2014). Therefore, rather than considering it a measure of heterozygosity within breeds it reflects that Asian and African sheep are more genetically divergent to the reference genome in comparison to European, UK and American. Also, we calculated the Fixation Index (Fst) of each population compared to Middle East breeds, where first sheep domestication took place In all comparisons we observe very low Fst values showing a very weak population structure across sheep breeds (Table 1).

Table 1. Summary statistics on samples depth, number of called SNPs and percentage of heterozygosity

Region	Number of Samples	Average Depth	Average Number of called SNPs, millions	% Heterozygosity	Fst
Africa	6	8.2	24.34	28.3	0.024
Americas	8	8.1	23.18	19.4	0.021
Asia	12	8.2	23.86	32.3	0.020
Europe	22	8.1	25.05	19.9	0.023
Middle East	13	7.8	23.00	24.6	-
United Kingdom	7	7.6	22.23	20.4	0.030

Previous analysis based on mitochondrial haplotypes and SNP chip datasets have suggested globally distributed populations of sheep exhibit generally weak population substructure in comparison to other domestic species (Meadows et al. 2005; Kijas et al. 2012). We sought to determine if the much higher density (and unbiased) SNP collection obtained here is able to provide additional detail about the relatedness amongst a global collection of domestic sheep. We performed PCA of pairwise allele sharing to infer global patterns of genetic structure, with the results shown in Figure 1A. PC1 separated European and UK sheep from African, Asian, and Middle East. This largest PC only explained 4.2% of the total variance, indicating geographic origin is not a major source of variation. The second PC (2.6%) separated African, Middle East and Asian sheep. Finally, sheep from the Americas do not form a discrete cluster, but were rather distributed throughout the clusters ofAfrican or European animals. Thus, likely reflecting the highly admixed population history of the animals sampled from the Americas. Also, we analysed the 68 domestic sheep genotypes on basis of their compression efficiency (CE) and heterozygosity. The CE algorithm provides a new alternative to cluster populations based on the allele order and proportion across individuals (Hudson et al. 2014). It has been previously shown to reveal population structure in human populations, as well as cattle, mouse, dog and feral versus domestic sheep (Hudson et al. 2014). Here, we explore only domestic breeds which present relatively similar heterozygosity and CE levels (Figure 1B). Therefore, surprisingly, CE does not capture the same population structure as PCA and it is not able to clearly differentiate the phylogeography of the different breeds. Finally, the CE presents two clear outliers. The first corresponds to an Asian Garut animal with very low heterozygosity and high CE, whereas the second belongs to an American sheep from Santa Ines, with high heterozygosity and low CE. Possibly reflecting the level of admixture in different breeds.

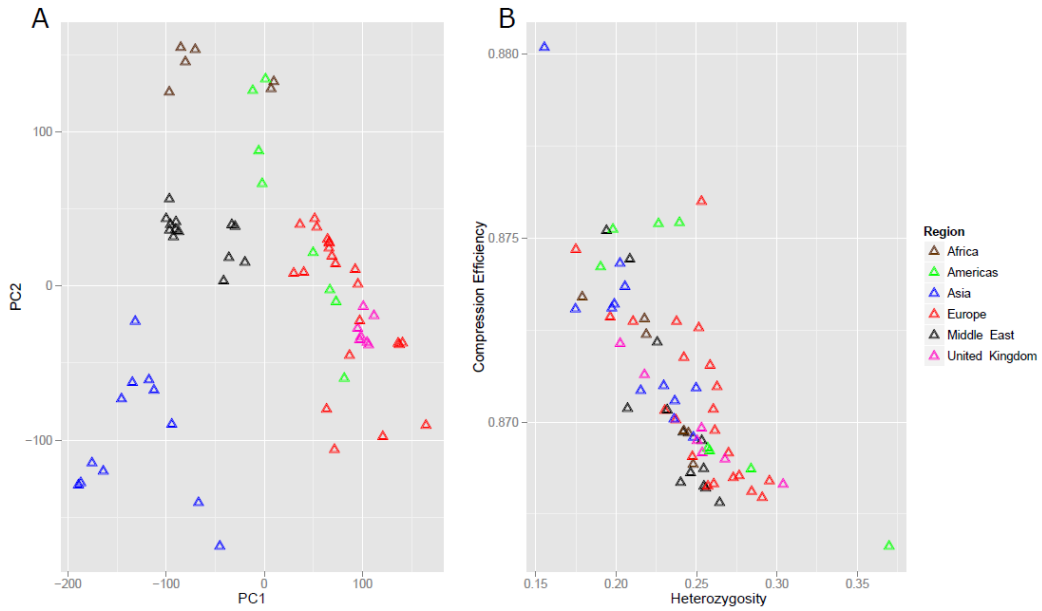

Figure1. Population structure. Breeds were coloured by origin Africa, Americas, Asia, Europe, Middle East and United Kingdom. A) Principal Component Analysis of genetic distance and B) plot CE versus Heterozigosity based on 550,048 SNPs without missing genotypes across the 68 animals.

Future prospects in our analysis is to study the genomic features selected in particular domestic breeds together with the addition of 18 wild ovid genotypes which would allow us to study the impact of domestication by defining genomic regions and the associated functional traits selected across domestic breeds.

REFERENCES

Axelsson E., Ratnakumar A., Arendt M-L., Maqbool K., Webster M.T., *et al*. (2013) *Nature* **495**: 360.

Chessa B., Pereira F., Arnaud F., Amorim A., Goyache F., *et al.* (2009) *Science* **324**: 532.

Danecek P., Auton A., Abecasis G., Albers C.A., Banks E., *et al.* (2011). *Bioinforma Oxf Engl* **27**: 2156.

Gautier M.m Laloe D. and Moazami-Goudarzi K. (2010) *PlOS ONE* **5**:e13038

Hudson N.J., Porto-Neto L.R., Kijas J., McWilliam S., Taft R.J., *et al*. (2014) *BMC Bioinformatics* **15**: 66.

Kijas J.W., Lenstra J.A., Hayes B., Boitard S., Porto Neto L.R., *et al.* (2012) *PLoS Biol* **10**: e1001258.

Jiang Y., Xie M., Chen W., Tallbot R. Maddox J.F., *et al.* (2014) *Science* **6**:344(6188):1168.

Larkin D.M., Daetwyler H.D., Hernandez A.G., Wright C.L., Hetrick L.A., *et al.* (2012) *Proc Natl Acad Sci U S A* **109**: 7693.

Li H. and Durbin R. (2009). *Bioinforma Oxf Engl* **25**: 1754.

Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R., 1000 Genome Project Data Processing Subgroup. (2009) *Bioinforma Oxf Engl* **25**: 2078.

Rubin C-J., Zody M.C., Eriksson J., Meadows J.R.S., Sherwood E., *et al.* (2010) *Nature* **464**: 587.

Rubin C-J., Megens H-J., Martinez Barrio A., Maqbool K., Sayyab S., *et al.* (2012) *Proc Natl Acad Sci U S A* **109**: 19529.

Tishkoff S.A., Reed F.A., Friedlaender F.R., Ehret C., Ranciaro A. , *et al.* (2009) *Science* **324**:528.

Weir B.S. and Cockerham C.C. (1984) *Evolution* **38**:1358.

Zeder M.A. (2008) *Proc Natl Acad Sci U S A* **105**: 11597.