






Resequencing and signatures of selection scan in two Siberian native sheep breeds point to candidate genetic variants for adaptation and economically important traits

J. Sweet-Jones*, A. A. Yurchenko[†], A. V. Igoshin[†], N. S. Yudin[†] , M. T. Swain[‡]  and D. M. Larkin^{*†} 

*Royal Veterinary College, University of London, London NW1 0TU, UK. [†]The Federal Research Center Institute of Cytology and Genetics, The Siberian Branch of the Russian Academy of Sciences (ICG SB RAS), Novosibirsk 630090, Russia. [‡]Institute of Biological, Environmental and Rural Sciences, University of Aberystwyth, Aberystwyth SY23 3DA, UK.

Summary

Russian sheep breeds represent an important economic asset by providing meat and wool, whilst being adapted to extreme climates. By resequencing two Russian breeds from Siberia: Tuva ($n = 20$) and Baikal ($n = 20$); and comparing them with a European (UK) sheep outgroup ($n = 14$), 41 million variants were called, and signatures of selection were identified. High-frequency missense mutations on top of selection peaks were found in genes related to immunity (*LOC101109746*) in the Baikal breed and wool traits (*IDUA*), cell differentiation (*GLIS1*) and fat deposition (*AADA3L3*) in the Tuva breed. In addition, genes found under selection owing to haplotype frequency changes were related to wool traits (*DSC2*), parasite resistance (*CLCA1*), insulin receptor pathway (*SOCS6*) and DNA repair (*DDB2*) in the Baikal breed, and vision (*GPR179*) in the Tuva breed. Our results present candidate genes and SNPs for future selection programmes, which are necessary to maintain and increase socioeconomic gain from Siberian breeds.

Keywords local breeds, selection, sheep, whole-genome resequencing

Following the domestication of sheep (*Ovis aries*) and their migration with human populations, natural selection allowed improved adaptation to local environments, and artificial selection and breed formation affected economically important traits (Zeder, 2008). In Russia, where populations of livestock face environmental stresses in temperature, sheep have been selectively bred to meet production demands while being adapted to their local conditions. Using whole-genome genotyping, Deniskova *et al.* (2018) demonstrated genetic clustering of Russian breeds based on their wool type and further showed coarse wool breeds clustering with Asian breeds, and fine-wool breeds with European breeds. Therefore, it may be expected that each of these two clusters of breeds would demonstrate distinct molecular traces of adaptation. This was shown through a signatures of selection scan using high-density genotyping of 15 Russian sheep breeds (Yurchenko *et al.*,

2019). Promising candidate gene regions were identified including those linked to wool traits, environmental adaptations, and domestication. The next step is to identify candidate genetic variants contributing to adaptations and economic traits. The aim of this study was to identify candidate genes containing missense SNPs under selection from two resequenced Russian sheep breeds from Siberia – the Tuva and Baikal.

Samples of two Russian breeds, a fine-wool Baikal long-thin-tailed sheep ($n = 20$) and a coarse-wool Tuva short-fat-tailed sheep ($n = 20$) were resequenced to approximately 15× raw coverage using paired-ended Illumina reads (150 bp) by Novogene (Hong Kong). In addition, 11 samples from 11 UK sheep breeds were resequenced to approximately 13× raw coverage each, and three samples from three additional UK breeds were downloaded from the Sequence Repository Archive (Table S1). Russian sheep samples were described in Yurchenko *et al.* (2019) whereas the UK samples were described in Heaton *et al.* (2014) and Beynon *et al.* (2015). Reads were mapped to Texel reference genome Oar version 3.1 with the BWA-MEM algorithm (BWA version 0.7.10; Li, 2013), then sorted with SAMTOOLS version 0.1.18 (Li *et al.*, 2009). Duplicates were marked and libraries merged with PICARD version 2.18 (<http://broadinstitute.github.io/pica>

Address for correspondence

D. M. Larkin, Royal Veterinary College, University of London, London NW10TU, UK.
E-mail: dmlarkin@gmail.com

Accepted for publication 28 September 2020

rd/). Base-quality recalibration, SNP calling and hard filtering were performed with GENOME ANALYSIS TOOLKIT version 3.8 (McKenna *et al.*, 2010) using filter expression: 'QD < 2.0||FS > 60||MQ < 40||MQRankSum < -12.5||ReadPosRankSum < -8'. The output file containing all high-quality SNPs was further filtered in PLINK (Purcell *et al.*, 2007) using the parameters '--geno 0.1 --mind 0.05 --maf 0.0000001 --chr 1-26'. The population statistics inbreeding coefficient (F), expected heterozygosity (H_e) and proportion of polymorphic loci (P) were calculated with PLINK --hardy --freq and --het commands. The transition–transversion index was calculated using VCFTOOLS version 0.1.13 --Ts/Tv-summary command (Danecek *et al.*, 2011).

The decorrelated composite of multiple signals pipeline was used to calculate $H1$, $H12$, Tajima's D , Pi and F_{ST} as described in Yurchenko *et al.* (2019). The method allows integration of the major measures of the signatures of selection into a single statistic (Ma *et al.*, 2015). P -values were converted to q -values to correct for false discovery rate using BioCONDUCTOR q value package. q -Values were used to render Manhattan plots using *qqman manhattan* function in R. To identify regions under selection genome-wide, all intervals with SNPs expressing decorrelated composite of multiple signal q -values less than 0.01 were identified. Selected interval boundaries were defined by the first SNP with q -value greater than 0.2 up- and downstream. SNPs found within the regions were annotated using the NGS-SNP pipeline (Grant *et al.*, 2011). Putative effects of missense SNPs were predicted using PolyPhen score range 0–1 (0 = benign, 1 = deleterious; Adzhubei *et al.*, 2013). Representation of haplotypes was rendered by HAPLOSTRIPS (Marnetto & Huetra-Sánchez, 2017) from phased SNP data. Functional enrichment analysis was performed using the Database for Annotation, Visualization and Integrated Discovery (DAVID version 6.8; Huang *et al.*, 2009) using all genes found within regions under selection per breed against a background of all genes. Any group with enrichment score greater than 1.3 was described.

Copy number variant (CNV) analyses were conducted with cn.Mops R package (Klambauer *et al.*, 2012) in windows of 700 SNPs. These were merged into CNV regions (CNVRs) with BEDOPS *bedmap* function using at least 50% reciprocal overlap in at least three individuals per breed. Duplicate CNVRs were removed and count number was used to infer duplication or deletion. Effective

population sizes (N_e) were calculated with smc++ version 1.15.2 (Terhorst *et al.*, 2017) retrospectively excluding individuals with excessive homozygosity, all SNPs within regions under selection and CNVRs. Generation times of 4 years and mutation rate 1.0×10^{-8} were assumed from the literature (Kijas *et al.*, 2012).

Fifty-four resequenced samples with mean coverage of $11.9\times$ were aligned to the reference genome (Table S1). The GENOME ANALYSIS TOOLKIT pipeline called 41.6 million SNPs, which were pruned to 38.3 million after filtering. Population statistics (Table S2), showed high levels of polymorphic loci in both breeds as well as low inbreeding in the Baikal breed and moderate inbreeding in Tuva. Equal ranges of H_e were seen in both breeds and the N_e calculated was larger for the Tuva than the Baikal breed (Fig. S1), in line with estimates by Deniskova *et al.* (2018). Transition–transversion ratios align with those previously seen in commercial cattle breeds (Jiang *et al.*, 2008).

CNV regions covered 1 and 3% of Baikal and Tuva breed genomes respectively (Tables S3 & S4), which overlapped a list of known ovine CNVRs by 78% in Baikal and 81% in Tuva sheep. Non-overlapping CNVRs can be seen in Tables S5 & S6. Four regions under selection in Baikal and 34 in Tuva breeds overlapped CNVRs present in their respective breeds (Tables S7 & S8). These regions, clustering on OAR3, OAR6 and OAR17 in Tuva and OAR17 in Baikal breeds, were treated as artefacts of alignment and SNPs from these regions were discounted from the selection scan.

The remaining 739 selected intervals (Baikal = 296; Tuva = 443) spanned 1.0 and 1.3 Mbp in Baikal and Tuva breeds respectively containing 15 954 and 24 978 SNPs where 3084 (19%) and 9430 (38%) SNPs were not present in the NCBI SNPdb. Annotation of SNPs in the regions under selection found 31 missense mutations in Baikal (Table S9) and 12 in Tuva sheep (Table S10). DAVID functional clustering demonstrated enrichment for *ubiquitinylation*, *transmembrane helix proteins* and *keratin filament formation* terms in Baikal (Table S11) and none in the Tuva breed. A list of all genes found within the regions under selection entered for DAVID analysis is available (Tables S12 & S13).

We focused on the genes found in the top selected intervals with the difference in allele frequencies between populations supported by F_{ST} or haplotype analysis ($H1/H2$ statistics; Fig. 1; Table 1). In the Baikal sheep, *DSC2* (q -

Figure 1 (a) Manhattan plot of decorrelated composite of multiple signals q -values of Baikal (yellow) and Tuva (blue) breeds showing missense mutations found under selection, highlighted in red with corresponding gene names. Selection thresholds for suggestive ($q < 0.05$) and strong ($q < 0.01$) selection are shown by blue and red lines respectively. The asterisk denotes peaks excluded from analysis owing to CNVR overlap, coordinated by colour to the corresponding breed. (b) Allele frequencies for missense mutations identified by strong F_{ST} score represented by pie charts (green, reference allele; red, derived allele). This shows location of missense mutations along their gene with nucleotide substitution highlighted by red circle. Coloured dotted lines point to corresponding peak positions on the Manhattan plot. Functional domain of SNP, amino acid substitution, taken from NGS-SNP, is shown alongside PolyPhen score and q -value. (c) HAPLOSTRIP plots spanning genes containing missense mutations but selected on the basis of $H1/H2$ haplotype statistics. Similar haplotypes are clustered together per population to demonstrate selection within these regions across the whole gene. These show the presence of reference (white) or derived (black) alleles making up different haplotypes. Populations of interest are highlighted in boxes corresponding to their colours on the Manhattan plot

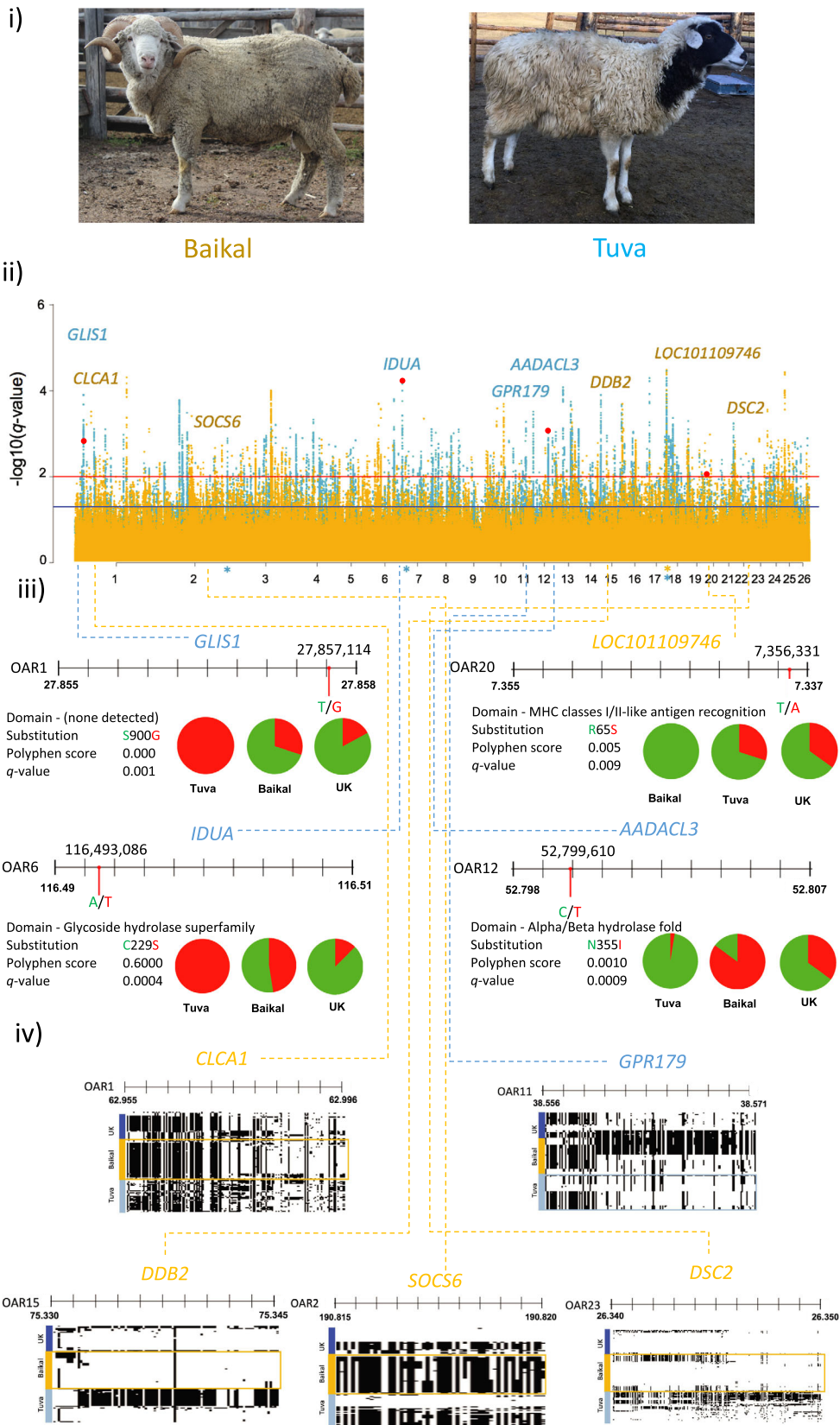


Table 1 Positions of loci under selection in the Baikal and Tuva breed genomes

OAR	Interval	Missense position	SNP breed	Gene	Reference allele	Alternate allele	Mutation	PolyPhen score	Strongest (value)	statistic	q -Value	Function	Accession number
1	27 855 091– 27 857 655	27 857 114	Tuva	<i>GLIS1</i>	T	G ¹	T215P	0.000	F_{ST} (0.3)		0.001	Cell fate	rs426118206
1	62 996 080– 62 996 303		Baikal	<i>CLCA1</i>					$H1/H12(\max)^2$		0.009	Calcium ion transport	
2	190 814 967– 190 816 753		Baikal	<i>SOC56</i>					$H1/H12/Pi/Tajima's D(\max)$		0.003	Immunity	
6	116 488 669– 116 512 658	116 493 086	Tuva	<i>IDUA</i>	A	T ¹	C229S	0.600	F_{ST} (0.5)		0.0004	Metabolism	rs159996479
11	38 565 876– 38 568 609		Tuva	<i>GPR179</i>					$H1/H12/Tajima's D(\max)$		0.001	Vision	
12	52 798 667– 52 807 569	52 799 610	Tuva	<i>AADACL3</i>	C ¹	T	V355I	0.001	F_{ST} (0.6)		0.0009	Metabolism	rs405926468
15	75 333 055– 75 345 371		Baikal	<i>DDB2</i>					$H1/H12/Pi/Tajima's D(\max)$		0.001	DNA repair	
20	7 355 294– 7 356 865	7 356 331	Baikal	<i>LOC101109746</i>	T	A ¹	R65S	0.005	F_{ST} (0.3)		0.009	Immunity	rs416908264
23	26 347 439– 26 352 033		Baikal	<i>DSC2</i>					$H1/H12/Pi/Tajima's D(\max)$		0.001	Wool traits	

¹Denotes selected allele.²Denotes listed statistics are at maximum: $H1/H12 = 1$, $Pi = 0$, Tajima's $D = -2$.

value = 0.001), which encodes a desmosomal protein found in hair follicles, linked to cashmere traits in goats (Simpson *et al.*, 2009; Wang *et al.*, 2016), overlapped a selected interval with the strongest signal originating from the *H1/H2* statistics. We failed to identify missense mutations with large F_{ST} values in the gene, suggesting that selection probably acts on haplotypes. This locus was previously found under selection in a scan for selected regions in Russian long-haired sheep (Yurchenko *et al.*, 2019). Three additional genes in Baikal breed overlapped intervals recognised by *H1/H2* statistics: *CLCA1*, a chloride channel regulatory protein, upregulated in sheep resistant to *Teladorsagia* infection (Chitneedi *et al.*, 2018); *SOCS6*, which is known to regulate the insulin receptor pathway (Krebs *et al.*, 2002); and *DDB2*, which recruits DNA repair factors after ultraviolet radiation damage (Nag *et al.*, 2001). Only one of the top signatures of selection contained missense mutation with a high F_{ST} ($F_{ST} = 0.3$) in the *LOC101109746* gene which encodes the HLA class II histocompatibility antigen DM beta chain, needed for the major histocompatibility complex for antigen presentation to the adaptive immune system (Fling *et al.*, 1994; Fig 1).

Tuva sheep showed multiple missense mutations on the top peaks of selected regions supported by the F_{ST} statistics. Of these, the strongest signature of selection was found in the derived allele of *IDUA* (q -value = 0.0004; $F_{ST} = 0.5$), encoding α -L-iduronidase. Human pathologies in this gene lead to mucopolysaccharidosis type I, which presents a global phenotype that includes coarse hair (Scott *et al.*, 1995; Kloska *et al.*, 2005). Furthermore, an alternative allele of *GLIS1*, an enhancer of pluripotency markers (Maekawa *et al.*, 2011), and the reference allele of *AADA3L3*, which is associated with fat deposition in Chinese sheep breeds (Lu *et al.*, 2020), were found in the regions under selection with support from F_{ST} . The reference allele of *GPR179*, highlighted by *H1/H2* statistics, which is linked to genetic blindness, was also found under selection in Tuva sheep (Audo *et al.*, 2012).

The results of this study, based on whole-genome resequencing, point to stronger and often more narrow signatures of selection than previously reported in the literature using the same two Russian sheep breeds but with high-density genotyping data (Yurchenko *et al.*, 2019). Owing to whole-genome resequencing data, we were able to point to novel candidate SNPs and haplotypes under selective pressure in both breeds. These include haplotypes in candidate genes and missense SNP candidates for wool traits, fat deposition and immunity in the Baikal and Tuva sheep breeds. Knowledge of these genetic variants confers insight into the adaptations of each breed to its local environment, highlights economically important traits and provides variations for marker-assisted breeding.

Acknowledgements

The work was supported by the Russian Science Foundation grant RSF 19-76-20026.

Data availability statement

Sequence data for Russian and UK breeds are available from the NCBI SRA with the Bioproject ID: PRJNA646642.

References

- Adzhubei I., Jordan D.M. & Sunyaev S.R. (2013) Predicting functional effect of human missense mutations using PolyPhen-2. *Current Protocols in Human Genetics* **76**, 7–20.
- Audo I., Bujakowska K., Orhan E. *et al.* (2012) Whole-exome sequencing identifies mutations in GPR179 leading to autosomal-recessive complete congenital stationary night blindness. *American Journal of Human Genetics* **90**, 321–30.
- Beynon S.E., Slavov G.T., Farré M. *et al.* (2015) Population structure and history of the Welsh sheep breeds determined by whole genome genotyping. *BMC Genetics* **16**, 65.
- Chitneedi P. K., Suárez-Vega A., Martínez-Valladares M., Arranz J. J. & Gutiérrez-Gil B. (2018) Exploring the mechanisms of resistance to *Teladorsagia circumcincta* infection in sheep through transcriptome analysis of abomasal mucosa and abomasal lymph nodes. *Veterinary Research*, **49**(1), e39.
- Danecek P., Auton A., Abecasis G. *et al.* (2011) The variant call format and VCFtools. *Bioinformatics* **27**, 2156–8.
- Deniskova T.E., Dotsev A.V., Selionova M.I. *et al.* (2018) Population structure and genetic diversity of 25 Russian sheep breeds based on whole-genome genotyping. *Genetics Selection Evolution* **50**, 29.
- Fling S.P., Arp B. & Pious D. (1994) HLA-DMA and -DMB genes are both required for MHC class II/peptide complex formation in antigen-presenting cells. *Nature* **368**, 554–8.
- Grant J.R., Arantes A.S., Liao X. & Stothard P. (2011) In-depth annotation of SNPs arising from resequencing projects using NGS-SNP. *Bioinformatics* **27**, 2300–1.
- Heaton M.P., Leymaster K.A. & Kalbfleisch T.S. (2014) SNPs for parentage testing and traceability in globally diverse breeds of sheep. *PLoS One* **9**, e94851.
- Huang D.W., Sherman B.T. & Lempicki R.A. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols* **4**, 44–57.
- Jiang Z., Wu X.-L., Zhang M., Michal J.J. & Wright R.W. Jr (2008) The complementary neighborhood patterns and methylation-to-mutation likelihood structures of 15,110 single-nucleotide polymorphisms in the bovine genome. *Genetics* **180**, 639–47.
- Kijas J.W., Lenstra J.A., Hayes B. *et al.* (2012) Genome-wide analysis of the world's sheep breeds reveals high levels of historic mixture and strong recent selection. *PLoS Biology* **10**, e1001258.
- Klambauer G., Schwarzbauer K., Mayr A., Clevert D.-A., Mitterecker A., Bodenhofer U. & Hochreiter S. (2012) cn.MOPS: mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. *Nucleic Acids Research* **40**, e69.
- Kloska A., Bohdanowicz J., Konopa G., Tyłki-Szymńska A., Jakóbkiewicz-Banecka J., Czartoryska B., Liberek A., Węgrzyn

- A. & Węgrzyn G. (2005) Changes in hair morphology of mucopolysaccharidosis I patients treated with recombinant human α -L-iduronidase (Iaronidase, Aldurazyme). *American Journal of Medical Genetics* **139A**, 199–203.
- Krebs D.L., Uren R.T., Metcalf D. *et al.* (2002) SOCS-6 binds to insulin receptor substrate 4, and mice lacking the SOCS-6 gene exhibit mild growth retardation. *Molecular and Cellular Biology* **22**, 4567–78.
- Li, H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv*, **1303.3997v2**. <https://arxiv.org/abs/1303.3997v2>
- Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G. & Durbin R. (2009) The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–9.
- Lu Z., Yue Y., Yuan C. *et al.* (2020) Genome-wide association study of body weight traits in chinese fine-wool sheep. *Animals* **10**, 170.
- Ma Y., Ding X., Qanbari S., Weigend S., Zhang Q. & Simianer H. (2015) Properties of different selection signature statistics and a new strategy for combining them. *Heredity* **115**, 426–36.
- Maekawa M., Yamaguchi K., Nakamura T., Shibukawa R., Kodanaka I., Ichisaka T., Kawamura Y., Mochizuki H., Goshima N. & Yamanaka S. (2011) Direct reprogramming of somatic cells is promoted by maternal transcription factor Glis1. *Nature* **474**, 225–9.
- Marnetto D. & Huerta-Sánchez E. (2017) Haplostrips: revealing population structure through haplotype visualization. *Methods in Ecology and Evolution* **8**, 1389–92.
- McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytzky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M. & DePristo, M. A. (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, **20**, 1297–1303.
- Nag A., Bondar T., Shiv S. & Raychaudhuri P. (2001) The xeroderma pigmentosum group E gene product DDB2 is a specific target of cullin 4A in mammalian cells. *Molecular and Cellular Biology* **21**, 6738–47.
- Purcell S., Neale B., Todd-Brown K. *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics* **81**, 559–75.
- Scott H. S., Bunge S., Gal A., Clarke L. A., Morris C. P. & Hopwood J. J. (1995) Molecular genetics of mucopolysaccharidosis type I: Diagnostic, clinical, and biological implications. *Human Mutation*, **6**(4), 288–302.
- Simpson M.A., Mansour S., Ahnood D., Kalidas K., Patton M.A., McKenna W.J., Behr E.R. & Crosby A.H. (2009) Homozygous mutation of desmocollin-2 in arrhythmogenic right ventricular cardiomyopathy with mild palmoplantar keratoderma and woolly hair. *Cardiology* **113**, 28–34.
- Terhorst J., Kamm J.A. & Song Y.S. (2017) Robust and scalable inference of population history from hundreds of unphased whole genomes. *Nature Genetics* **49**, 303–9.
- Wang X., Liu J., Zhou G. *et al.* (2016) Whole-genome sequencing of eight goat populations for the detection of selection signatures underlying production and adaptive traits. *Scientific Reports* **6**, 38932.
- Yurchenko A.A., Deniskova T.E., Yudin N.S. *et al.* (2019) High-density genotyping reveals signatures of selection related to acclimation and economically important traits in 15 local sheep breeds from Russia. *BMC Genomics* **20**, 294.
- Zeder M.A. (2008) Domestication and early agriculture in the Mediterranean Basin: origins, diffusion, and impact. *Proceedings of the National Academy of Sciences* **105**, 11597–604.

Supporting information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Figure S1. Demographic inference of Baikal and Tuva Russian sheep breeds scaled to standard mutation rate (1.0×10^{-8}) and generation years (4)

Table S1. List of genes entered for DAVID clustering analysis of the Tuva breed