



# Genotyping and Whole-Genome Resequencing of Welsh Sheep Breeds Reveal Candidate Genes and Variants for Adaptation to Local Environment and Socioeconomic Traits

James Sweet-Jones<sup>1</sup>, Vasileios Panagiotis Lenis<sup>2,3</sup>, Andrey A. Yurchenko<sup>4</sup>, Nikolay S. Yudin<sup>4</sup>, Martin Swain<sup>2</sup> and Denis M. Larkin<sup>1,4\*</sup>

## OPEN ACCESS

### Edited by:

Anastasia Anashkina,  
Engelhardt Institute of Molecular  
Biology (RAS), Russia

### Reviewed by:

Qianjun Zhao,  
Chinese Academy of Agricultural  
Sciences (CAAS), China  
Siroj Bakoev,  
Centre for Strategic Planning  
and Management of Biomedical  
Health, Russia  
Jianlin Han,  
International Livestock Research  
Institute (ILRI), Kenya

### \*Correspondence:

Denis M. Larkin  
[dmlarkin@gmail.com](mailto:dmlarkin@gmail.com)

### Specialty section:

This article was submitted to  
Computational Genomics,  
a section of the journal  
Frontiers in Genetics

**Received:** 30 September 2020

**Accepted:** 10 May 2021

**Published:** 18 June 2021

### Citation:

Sweet-Jones J, Lenis VP,  
Yurchenko AA, Yudin NS, Swain M  
and Larkin DM (2021) Genotyping  
and Whole-Genome Resequencing  
of Welsh Sheep Breeds Reveal  
Candidate Genes and Variants  
for Adaptation to Local Environment  
and Socioeconomic Traits.  
*Front. Genet.* 12:612492.  
doi: 10.3389/fgene.2021.612492

<sup>1</sup> Royal Veterinary College, University of London, London, United Kingdom, <sup>2</sup> Institute of Biological, Environmental and Rural Sciences, University of Aberystwyth, Aberystwyth, United Kingdom, <sup>3</sup> School of Health and Life Sciences, Teesside University, Middlesbrough, United Kingdom, <sup>4</sup> The Federal Research Center Institute of Cytology and Genetics, The Siberian Branch of the Russian Academy of Sciences (ICG SB RAS), Novosibirsk, Russia

**Background:** Advances in genetic tools applied to livestock breeding has prompted research into the previously neglected breeds adapted to harsh local environments. One such group is the Welsh mountain sheep breeds, which can be farmed at altitudes of 300 m above sea level but are considered to have a low productive value because of their poor wool quality and small carcass size. This is contrary to the lowland breeds which are more suited to wool and meat production qualities, but do not fare well on upland pasture. Herein, medium-density genotyping data from 317 individuals representing 15 Welsh sheep breeds were used alongside the whole-genome resequencing data of 14 breeds from the same set to scan for the signatures of selection and candidate genetic variants using haplotype- and SNP-based approaches.

**Results:** Haplotype-based selection scan performed on the genotyping data pointed to a strong selection in the regions of *GBA3*, *PPARGC1A*, *APOB*, and *PPP1R16B* genes in the upland breeds, and *RNF24*, *PANK2*, and *MUC15* in the lowland breeds. SNP-based selection scan performed on the resequencing data pointed to the missense mutations under putative selection relating to a local adaptation in the upland breeds with functions such as angiogenesis (*VASH1*), anti-oxidation (*RWDD1*), cell stress (*HSPA5*), membrane transport (*ABCA13* and *SLC22A7*), and insulin signaling (*PTPN1* and *GIGFY1*). By contrast, genes containing candidate missense mutations in the lowland breeds are related to cell cycle (*CDK5RAP2*), cell adhesion (*CDHR3*), and coat color (*MC1R*).

**Conclusion:** We found new variants in genes with potentially functional consequences to the adaptation of local sheep to their environments in Wales. Knowledge of these variations is important for improving the adaptative qualities of UK and world sheep breeds through a marker-assisted selection.

**Keywords:** sheep, signatures of selection, Wales, whole-genome resequencing, adaptation

## INTRODUCTION

Since the domestication in the semi-arid Fertile Crescent of Iran and Turkey, sheep (*Ovis aries*) have undergone migration and selection to form established breeds that are well-suited to various local environments (Zeder, 2008). The process of natural or artificial positive/negative selection results in genomic regions of a decreased diversity, which are known as the signatures of selection (Kijas et al., 2012). Detecting signatures of selection is important for understanding the genetic mechanisms of the adaptation of breeds to their local environments. Previous investigations have unearthed genes relating to hypoxia in sheep adapted to high altitudes on the Qinghai-Tibetan Plateau of China (*THRB*) (Yang et al., 2016), fat deposition in sheep from arid deserts of northern Africa (*PCDH9*) (Kim et al., 2016), and metabolism in Russian sheep adapted to low temperatures (*POMC*) (Yurchenko et al., 2019). Whilst these examples represent the extreme circumstances, there are examples of selection in sheep adapted to more temperate climates. For instance, three upland sheep breeds from northern England were shown to demonstrate higher than expected frequencies of the known missense mutations in genes associated with reproductive success (*PRLP*) and presence of horns (*RXFP2*) (Bowles et al., 2014). Knowledge of such selection is essential for the continued improvement of breeds better suited to their environments and with a better socioeconomic trait potential in current selection programs, which is particularly pressing in terms of climate change (Bowles, 2015). Here, we investigated another example of previously neglected breeds adapted to local environments in Wales, United Kingdom, where sheep have historically been farmed at altitudes of >1,000 feet (~300 m) above sea level on a rough pasture with shallow-rooted plants (Davies, 1935).

Sheep were introduced to Wales by Neolithic settlers, bringing a primitive breed similar to the contemporary Soay (6,000 years ago) with two likely further introductions by Roman white-faced, fine-wool breeds (2,000 years ago) and Norse black-faced breeds (1,000 years ago) (Ryder, 1964). The South Wales Welsh Mountain (SWWM) breed of today, a descendent of the Roman and Soay breeds, has been documented since the 16th century and is renowned for its hardiness, lambing ability, and sweet meat taste. Despite this potential economic gain, these mountain sheep suffer a trade-off to a kemp wool and small carcass size, leading them to be poorly exploited outside of Wales (Williams-Davies, 1981). To overcome this, the SWWM breed was cross-bred in the eighteenth century to form a number of more productive local breeds, which are now farmed on the lowlands (Ryder, 1964; Williams-Davies, 1981).

Previous work to reveal the population history of Welsh sheep showed that Welsh breeds clustered closely with each other based on haplotype sharing, forming two groups within this cluster, which aligned with the upland or lowland farming style. Exceptions to this were the Black Welsh Mountain (BWM) and Kerry Hill with Beulah, which remained distinct from all other breeds (Beynon et al., 2015). This distinction of BWM is supportive of the alternate ancestry from Norse breeds and deliberate selection based on coat color (Williams-Davies, 1981). Moreover, the divergence of Kerry Hill and Beulah may be the result of a genetic bottleneck or a founder effect that these

breeds underwent. Additional support of this can be seen in the low effective population sizes and high inbreeding coefficients of these breeds (Beynon et al., 2015). Low effective population sizes endanger these breeds to risks of increased homozygosity and lack of genetic diversity due to inbreeding. This is a potential risk for UK upland sheep breeds who have remained geographically and genetically isolated due to their adaptability to thrive on pastures that other breeds cannot (Bowles et al., 2014; Heaton et al., 2014).

With this in mind, it is important, from a perspective of cultural significance, breed conservation, and breed improvement, to study the adaptation of Welsh breeds, which potentially offers insight into genomic adaptation to upland farming, as well as lambing and meat quality. Likewise, lowland productive breeds offer a good comparison due to their stronger capabilities for traits related to socioeconomic gain. Through the use of a commercially available medium-density SNP genotyping array with a HapFLK and Decorrelated Composite of Multiple Statistics (DCMS) software and whole-genome resequencing data with a DCMS pipeline, this paper aims to identify the signatures of selection in the genomes of Welsh sheep breeds and candidate genes containing functional missense mutations in these regions.

## MATERIALS AND METHODS

### Data Source and Variant Calling

Genotyping data from 353 individuals across 18 Welsh breeds on the Illumina OvineSNP50 SNP array from Beynon et al. (2015) were used in this study. Illumina pair-ended read (150 bp) resequencing of 11 Welsh sheep samples from the same set, representing one sample per breed (**Supplementary Table 1**), was performed at the University of Aberystwyth to ~13× raw coverage using Illumina HiSeq according to the manufacturer's protocol. Three remaining Welsh sheep resequenced genomes for Welsh Hardy-Speckled Faced, Dolgellau Welsh Mountain, and Talybont Welsh Mountain were downloaded from the National Centre for Biotechnology Information Sequence Read Archive, PRJNA160933 (Heaton et al., 2014). A dataset of resequenced Russian sheep samples ( $n = 40$ ) adapted to a contrasting environment was used as an outgroup in this study (Sweet-Jones et al., 2021).

Reads were mapped by using the Burrows-Wheeler Aligner BWA-MEM (BWA V.0.7.10 (Li, 2013)) to the reference sheep genome Oar\_V.3.1. Reads were sorted using the Samtools V.0.1.18 (Li et al., 2009) and duplicates were marked with the Picard V.2.18<sup>1</sup>. Libraries were also merged using Picard. Base Quality Score Recalibration (BQSR) was performed, which account for the systematic errors in sequencing, using the Genome Analysis Toolkit (GATK V.3.8; McKenna et al., 2010). Samples then underwent variant calling with the GATK HaplotypeCaller function. Finally, all samples ( $n = 54$ ) underwent joint calling to merge all reported variants into a single vcf file. Following this, hard filtering for quality scores assigned by BQSR was performed by GATK, using the filter

<sup>1</sup><http://broadinstitute.github.io/picard/>

expression “[QD < 2.0] | FS > 60.00] | MQ < 40.00] | MQRankSum < -12.5] | ReadPosRankSum < -8.0].” All variants from resequenced data were converted into a Plink V.1.90 (Purcell et al., 2007) format to be run through the DCMS pipeline (Yurchenko et al., 2019).

## HapFLK Statistics

Welsh breed genotyping data were separated into three groups of related breeds based on the clustering analysis performed by Beynon et al. (2015), resulting in one group of upland breeds and two groups of lowland breeds. Of the two lowland groups, one consisted of five lowland breeds and the other consisted of only the Kerry Hill and Beulah breeds (KHB) (Table 1). Plink-formatted (Purcell et al., 2007) files for the upland and lowland breed groups had genotypes from the sex chromosomes removed and were filtered to remove the rare alleles [-maf 0.05], low called SNPs [-geno 0.01], or poorly genotyped individuals [-mind 0.05]. FastPhase V.1.4 (Scheet and Stephens, 2006) was used to estimate the number of haplotype clusters ( $k$ ) for each group (Lowland  $k = 48$ , KHB  $k = 25$ , and Upland  $k = 53$ ). HapFLK software (Bonhomme et al., 2010) was used to obtain selection statistics for each group. This test uses the hierarchical population structure to identify the haplotype-based selective sweeps, which focuses on the inherited combinations of alleles. It has the advantage of an increased statistical power, reliably detecting the hard and soft selective sweeps, and is a realistic simulation of selection through the haplotypes. HapFLK  $p$ -values

were calculated using the Python script `scaling_chi2_hapflk.py` (Bonhomme et al., 2010; Fariello et al., 2013). Adjusted  $p$ -values, or  $q$ -values, were calculated through the R `qqman`  $q$ -value function (Turner, 2014). Selected intervals were determined by boundaries of  $q < 0.05$  with SNPs within an interval of  $q < 0.01$  considered to be under a strong selection.

## De-correlated Composite of Multiple Signals (DCMS)

Five established measures of selection and genetic diversity were both used on the genotyping (15 breeds) and whole genome resequencing (upland breeds  $n = 7$ ; lowland breeds = 7) datasets as a DCMS (Table 1; Ma et al., 2015). Statistics used were: (i/ii)  $H2/H1$  and  $H12$ , which can distinguish the hard and soft selective sweeps by measuring the intensity of selection (Garud et al., 2015); (iii) Tajima's  $D$  comparing pairwise sequence differences and the number of segregating sites, detecting positive, negative, or balancing selection (Tajima, 1989); (iv) nucleotide diversity ( $Pi$ ), average number of nucleotide differences between two sequences (Nei and Li, 1979); (v)  $F_{ST}$  fixation index, comparing single SNP frequencies across a population (Weir and Cockerham, 1984). By weighting the result of each statistic and generating a combined score, regions that overlap in the analysis outcomes gain a stronger evidence to be a region under selection.

**TABLE 1** | Breed representation in the genotyping and resequencing datasets.

Breed	Abbreviation	Genotyped samples	Horns	Base color	Fleece	HapFLK group	Resequencing group
Badger Faced Welsh Mountain	BFWM	21	Yes	Black and white	Firm	Upland	NA*
Balwen	–	14	No	Black and white	Firm	Upland	Upland
Beulah	–	22	No	Black and white	Fine	KHB	Lowland
Black Welsh Mountain	BWM	24	Yes	Black	Firm	Upland	Upland
Brecknock Hill Cheviot	BHC	24	No	White	Fine	Upland	Upland
Clun Forest	–	17	No	Black	Fine	Lowland	Lowland
Dolgellau Welsh Mountain	DWM	–	Yes	White	Firm	NA*	Upland
Hardy Speckled Faced	HSF	24	Yes	White	Fine	Lowland	Lowland
Hill Radnor	–	21	No	Gray-brown	Fine	Lowland	Lowland
Kerry Hill	–	18	No	White	Fine	KHB	Lowland
Llandovery White Faced	LWF	24	No	White	Fine	Upland	Upland
Llanwenog	–	21	No	Black	Fine	Lowland	Lowland
Lleyn	–	22	No	White	Fine	Lowland	Lowland
South Wales Welsh Mountain	SWWM	17	Yes	White	Firm	Upland	Upland
Talybont Welsh Mountain	TWM	24	No	White	Firm	Upland	Upland
Welsh Mountain Hill Flock	WMHF	24	Yes	White	Firm	Upland	NA
Average/Total	–	21/317	No	–	–	102/40/173	7/7

\*NA denotes the exclusion of breed from either the HapFLK or resequencing study due to the unavailability of samples.

## H2/H1 and H12

Autosomal SNPs were filtered for  $-maf\ 0.0000001 -geno\ 0.1 -mind\ 0.1$  by Plink. SNPs were then phased by chromosome using ShapeIt2 (Delaneau et al., 2011) with 400 states and an effective population size of 100. Phased chromosomes were split into appropriate groupings per chromosome using Plink and H2/H1, H12 were calculated using the H1\_H12.py Python script (Garud et al., 2015). H2/H1 and H12 values were calculated in windows of 25 SNPs using a step size of one SNP following our previous study (Yurchenko et al., 2019).

## Tajima's D

Tajima's *D* for mutation index was calculated over the same intervals of 25 SNPs, whose lengths were calculated from the output of the H2/H1 and H12 statistics. Using the vcfTools V.0.1.13 (Danecek et al., 2011), Tajima's *D* was calculated [ $-TajimaD\ 900000000$ ] for each chromosome per group/breed file per window. Output files were concatenated per group.

## Fixation Index

$F_{ST}$  was calculated with Plink comparing each group to all others. All negative values were converted to zero, and data were smoothed with the R *runmed* function in windows of 31 SNPs.

## Nucleotide Diversity

Plink-format file was split by chromosome per breed, and nucleotide diversity was calculated with the vcfTools [ $-site-pi$ ] option. Data were then smoothed using the R *runmed* function in windows of 31 SNPs.

## Combining Statistics With DCMS

Output files from individual statistics were sorted and joined by SNP id. Genome-wide *p*-values through ranking results of each statistic were calculated in the R MINOTAUR *stat-to-p*-value function specifying the one-tailed tests (H2/H1, H12,  $F_{ST}$ -right tailed,  $\pi$ , and Tajima's *D*-left tailed). Covariance matrix was constructed based on sampling 300,000 randomly sampled SNPs using the R *CovNAMcd* function where the  $\alpha = 0.75$ . DCMS statistics were calculated using the *DCMS* function and fitted to a normal distribution to examine normality implemented by the R *MASS rlm* function. These fitted DCMS values were converted to *p*-values using *pnorm*, and adjusted for a false discovery rate with the *qvalue* function. *Q*-values were parsed to determine selection with region boundaries set at a  $q < 0.2$  and a threshold of  $q < 0.01$ .

## Candidate Gene Search

For the regions defined by our pipelines as being under selection, genes were identified using a list of 26,958 genes of the Oar\_V.3.1 genome downloaded from Ensembl BioMart v.98. Within each selected region, genes were then ranked based on their distances to the most significant SNP of that region with the closest gene being the top-ranking. The top 10 highest ranking genes from each region underwent literature review for their previous associations to adaptation to local environments or socioeconomic traits in animals. Genes with established links to these traits were identified as candidate genes in this study.

In the resequencing study, all SNPs were annotated with NGS-SNP (Grant et al., 2011) to identify missense mutations. Only genes with missense mutations in the regions under selection were considered for literature review. Additionally, missense SNPs with a strong support from the  $F_{ST}$  statistic ( $F_{ST} \geq 0.3$ ) were analyzed with PolyPhen2 to predict their effects on protein structure and function (0 = benign and 1 = deleterious; Adzhubei et al., 2013).

Functional enrichment analysis was performed using the Database for Annotation, Visualization, and Integrated Discovery (DAVID) v. 6.8 (Huang et al., 2009) using the same gene list downloaded from Ensembl Biomart v. 98. Enrichments were detected using the DAVID functional clustering tool, which verifies enrichments of similar GO terms across many different databases, to confer a stronger evidence of enrichment. Scores  $> 1.3$ -fold enrichment were considered.

## Copy Number Variant (CNV) Analysis of Resequenced Genomes

CNV detection was performed to identify the regions in the genomes of Welsh sheep that had been duplicated or deleted with respect to the Texel reference genome. We identified CNVs in the resequenced samples with the cn.Mops R package (Klambauer et al., 2012) in a window length of 700 SNPs, using sequences that had undergone BQSR. This resulted in each individual being given a raw copy number (CN) per window. CN1–CN3 were considered to be normal and discounted from the results. Raw CNVs were merged into the CNV Regions (CNVRs) with the BedOps *bedmap* function using at least 50% reciprocal overlap in at least three individuals within the same group as criteria for inclusion. Duplicate CNVRs were removed, and the neighboring CNVRs were merged. This allowed us to be confident in our results by excluding the regions where CNVRs appear to overlap the signatures of selection, as these cast doubt over their reliability.

## Data Visualization

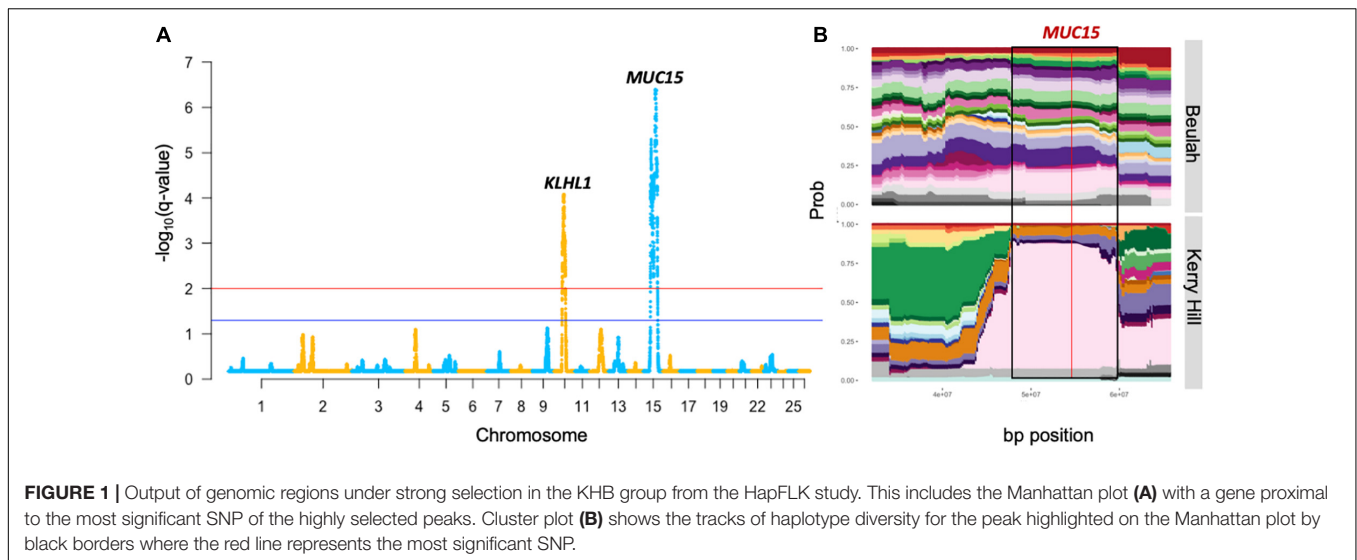
HapFLK clusterplots and haplotype trees were visualized from the prepared R scripts available (Bonhomme et al., 2010; Fariello et al., 2013). All Manhattan plots were rendered in R by the *qqman* package setting the suggestive line ( $q = 0.05$ ) and genome-wide line ( $q = 0.01$ ) to indicate selection. Haplostrips for visualization of haplotype sharing (Marnetto and Huerta-Sánchez, 2017) in the regions under selection detected by DCMS were run using phased data.

## RESULTS

### Regions Under Selection Detected From Genotyping Data

Using the HapFLK software, signatures of selection were found in each grouping using 44,711 SNPs (lowland group contained one region, KHB-two, and upland-31; **Supplementary Tables 2–4**). Cluster plots and Haplotype Trees for the most significantly selected regions in the KBH group are shown in **Figure 1** and





for other groups, in **Supplementary Figure 1**. Lengths of the regions found under selection ranged from 1.81 to 64.78 Mb with a median length of 7.33 Mb. Moreover, DCMS also found the regions under selection in all 15 breeds, with at least one region of a strong selection ( $q < 0.01$ ) in each breed using 47,366 SNPs. Lengths of the regions under selection ranged from 1 to 3.9 Mb where the median length of regions was 0.16 Mb. The number of regions detected in each breed ranged from 8 in Kerry Hill to 89 in Clun Forest (**Supplementary Tables 5–19**). In these regions under selection, 1,089 unique genes were found, with 179 occurring in multiple breeds including: *GBA3* (Hill Radnor, LWF, Lley, SWWM, TWM, and WMHF), *ENSOARG0000021104* (BFWM, Clun Forest, Hill Radnor, and Lley), *PCDH9* (BHC, Hill Radnor, and Llanwenog), and *PPARGCA1* (LWF, SWWM, and WMHF). Manhattan plots for all the breeds investigated are shown in **Figure 2** and **Supplementary Figure 2**.

There was a substantial overlap between the results for a region found on OAR6: 31.99–46.00 Mb ( $q = 2.0 \times 10^{-5}$ ) in the upland breeds in the HapFLK study to a narrower region, OAR6: 40.20–43.38 Mb, in LWF ( $q = 6.9 \times 10^{-15}$ ), WMHF ( $q = 4.3 \times 10^{-11}$ ), and SWWM ( $q = 8.73 \times 10^{-5}$ ). In all cases, the top ranked genes were *GBA3* linked to liver metabolism (Dekker et al., 2011) and *PPARGC1A* with known roles in mitochondrial biogenesis, fat deposition, and milk fatty acid composition in cattle (Fernandez-Marcos and Auwerx, 2011; Armstrong et al., 2018; Yuan et al., 2019; Li et al., 2016). On the HapFLK cluster plot for this region, low diversity was seen in the Balwen, BWM, LWF, and SWWM breeds, further supporting the idea of selection for this region. When plotted on a haplotype tree, it can be seen that Balwen had the longest branch compared to the other upland breeds, whilst the strongest signature of selection is seen in SWWM (**Supplementary Figure 1**).

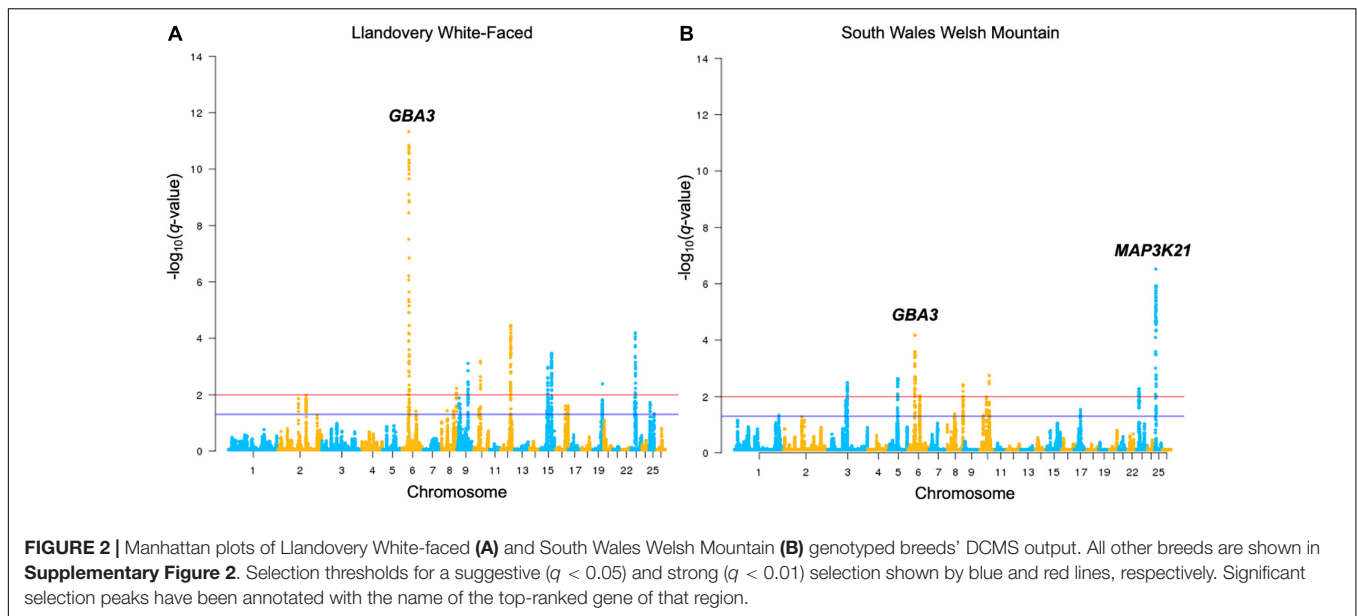
Several other candidate regions detected by HapFLK overlapped with the regions found in the Welsh breeds by DCMS. OAR3: 23.23–33.86 Mb ( $q = 5.0 \times 10^{-6}$ ) shared its top-ranked genes, *TDRD15* and *APOB*, with a region under selection in the upland Balwen breed ( $q = 0.0003$ ). These genes

are associated with cholesterol mobilization in Large White pigs (Bovo et al., 2019). OAR7: 47.55–54.20 Mb ( $q = 0.0008$ ) overlapped with another region OAR7: 51.30–52.76 Mb in the upland BWM ( $q = 5.0 \times 10^{-9}$ ), but these did not share the top-ranked genes. Strong overlapping candidate genes are presented in **Table 2**.

The most strongly selected region in the HapFLK study was seen on OAR15 32.15–72.10 Mb in the KHB lowland group, with the top-ranking gene being *MUC15* ( $q = 4.1 \times 10^{-7}$ ), associated with a low fecal egg count in Spanish Churra sheep during gastrointestinal parasite infections (Periasamy et al., 2014; Benavides et al., 2015). This was supported by a low haplotype diversity seen in the Kerry Hill breed at this locus, but this signature was not seen in Beulah or in the DCMS results in either breeds.

In the other lowland group, the only region found under selection, OAR13: 47.16–54.37 Mb ( $q = 0.01$ ), overlapped with another region found in the upland group, OAR13: 46.62–72.94 Mb ( $q = 2.0 \times 10^{-6}$ ), but the top candidate genes were different. In the lowland breeds, the top-ranked genes were *RNF24* and *PANK2*, which have been found under selection in world sheep breeds in association to a loss of vision following domestication (Naval-Sanchez et al., 2018; Wang et al., 2019). Cluster plots for this region showed a decreased haplotype diversity amongst the selected region on OAR13 in the lowland breeds, especially Clun Forest and HSF. Furthermore, significant selection at this locus in HSF ( $q = 0.0006$ ) was confirmed using the DCMS pipeline. In the case of the upland breeds, the top-ranked genes were *DHX35* and *PPP1R16B* which are related to innate viral immunity (Rahman et al., 2017) and endothelial cell proliferation (Pszczola et al., 2018), respectively.

The genes found within the regions under selection from HapFLK, functional enrichments seen for the KHB group included the DENN domain and connexin gap junctions, enriched 1.6- and 1.4-fold, respectively. The most highly enriched terms in the upland breeds were bactericidal permeability protein, major intrinsic protein, and semaphorins, which were all enriched over threefold. The most enriched cluster seen in



**TABLE 2 |** Candidate genomic regions and genes from the HapFLK analysis and corresponding DCMS region.

OAR	Region start	Region end	Group	q-value	Candidate genes (rank)	Function	Overlapping DCMS result
1	258,588,662	275,499,089	Upland	0.0002	<i>SIM2(1)</i> , <i>CLDN14(2)</i>		
3	23,229,012	33,863,536	Upland	$4.0 \times 10^{-6}$	<i>TDRD15(1)</i> , <i>APOB(2)</i>	Cholesterol efflux	Balwen
6	31,991,037	45,997,407	Upland	$1.0 \times 10^{-5}$	<i>GBA3(1)</i> , <i>PPARGC1A(2)</i> , <i>DHX15(3)</i> , <i>SOD3(4)</i> , <i>KCNIP4(6)</i>	Metabolism	LWF, WMHF, SWWM
10	35,062,152	51,872,705	KHB	$8.0 \times 10^{-5}$	<i>KLHL1(1)</i> , <i>PCDH9(2)</i>	Neurodevelopment, cell adhesion	
13	46,620,758	72,941,741	Upland	$2.0 \times 10^{-6}$	<i>DHX35(1)</i> , <i>PPP1R16B(3)</i> , <i>ADIG(6)</i>	Innate immunity, angiogenesis	BFWM
13	47,164,400	54,366,328	Lowland	0.01	<i>RNF24(1)</i> , <i>PANK2(2)</i> , <i>MAVS(3)</i>	Vision	HSF
15	32,152,675	65,710,885	Upland	$4.0 \times 10^{-7}$	<i>MUC15(1)</i> , <i>ANO3(3)</i>	Mucous production	

the genotyping DCMS analysis was the Type II keratin filaments from the Hill Radnor breed, which was enriched fivefold. This was followed by the Ribonuclease A in Clun forest, enriched 2.8-fold and leucine rich-repeats in Lley, enriched 2.7-fold (**Supplementary Table 20**).

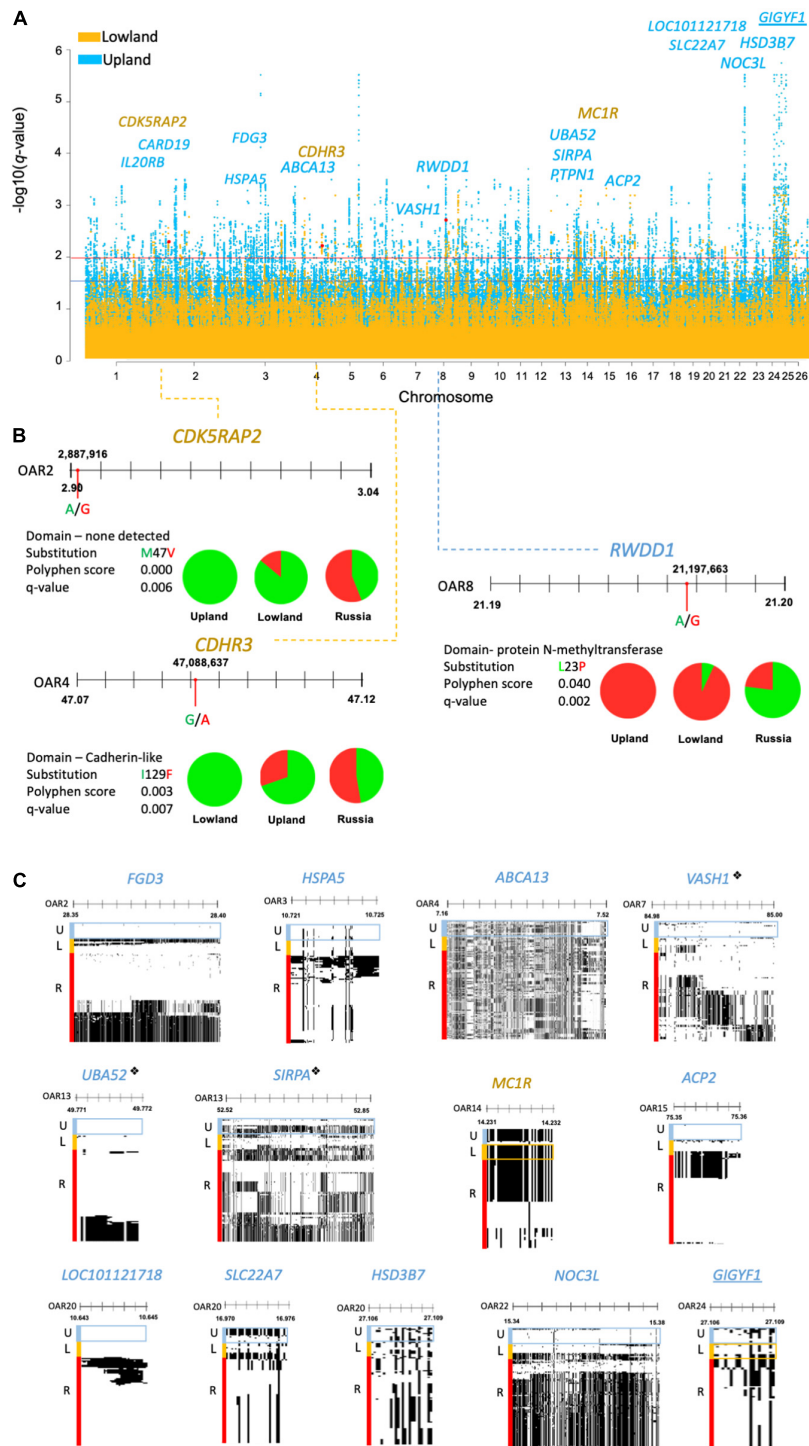
## Signatures of Selection Detected From Resequencing Data

Fifty-four (14 Welsh and 40 Russian sheep) resequenced genomes were aligned to the Oar\_V.3.1 genome with a mean filtered coverage of  $11.9\times$  (**Supplementary Table 1**). A total of 41,643,098 SNPs were called, which were pruned to 38,276,494 SNPs after filtering. CNVRs covered 0.27% of the lowland and 0.24% of the upland genomes, overlapping 852 and 669 genes, respectively (**Supplementary Tables 21,22**). Three CNVRs from the lowland breeds and 52 CNVRs from the upland breeds overlapped the regions of selection. Some of these CNVRs had a high frequency in the population, including the regions under a strong selection on OAR24 in the lowland breeds, spanning the *CLCN7* gene and on OAR17, whereas in the upland breeds, these spanned the *IGLV4-69*, *ZNF280B*, and *PRAME* genes.

After excluding the regions overlapping CNVRs, DCMS found 2,996 regions under selection in the Welsh breeds (lowland = 514, upland = 2,482; **Supplementary Tables 23,24** and **Figure 3A**). These regions overlapped 104 and 430 genes in the lowland and upland breeds, respectively. The most significantly selected region in the upland group was OAR22: 15.3676–15.3679 Mb, which overlapped the *NOC3L* gene ( $q = 1.8 \times 10^{-8}$ ), and for the lowland group, it was a single synonymous SNP located in *MYH11* ( $q = 0.0006$ ).

## Identification of Candidate Genes and Missense Mutations

In regions under selection, there were 12 missense mutations found in the lowland breeds (**Supplementary Table 25**) and 85 missense mutations found in the upland breeds (**Supplementary Table 26**). Of these, only missense mutations and their enclosing genes, which were top-ranking SNPs in their selected intervals are discussed below, leading to a total of 4 in the lowland breeds and 14 in the upland breeds. Clarification of the type of selection relied on a strong support from the *H2/H1* and *H12* statistics, which were considered to be haplotype-based selection, or *F<sub>ST</sub>*,



**FIGURE 3 | (A)** Manhattan plot of DCMS  $q$ -values of the lowland (yellow) and upland (blue) breeds showing missense mutations found under selection, highlighted in red with the corresponding gene names. Selection thresholds for a suggestive ( $q < 0.05$ ) and strong ( $q < 0.01$ ) selection shown by blue and red lines, respectively. Underlined gene names show selection in both the lowland and upland breeds. **(B)** Allele frequencies for missense mutations identified by a strong  $F_{ST}$  score represented by pie charts (green = reference allele, red = derived allele). This shows the location of missense mutations along their gene with nucleotide substitution highlighted by a red circle. Blue and yellow dotted lines point to the corresponding peak positions of the Manhattan plot. Amino acid substitution is shown, alongside the Polyphen score and  $q$ -value. **(C)** Haplostrip plots spanning genes containing the missense mutations but were selected on the basis of the H2/H1 and H12 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 colors on the Manhattan plot. “♦” is used to denote genes under selection by DCMS and HapFLK.

which was considered fixation of a variant in a population and was seen as a selection acting on individual SNPs.

Three missense mutations were found in the regions under selection ( $q < 0.01$ ), with a strong support from the  $F_{ST}$  statistic ( $F_{ST} \geq 0.3$ ). In the lowland breeds, these included the reference alleles of OAR2:2,887,916 in the *CDK5RAP2* gene ( $q = 0.006$ ;  $F_{ST} = 0.3$ ) and OAR4:47,087,846 in *CDHR3* ( $q = 0.007$ ;  $F_{ST} = 0.3$ ). For the upland breeds, only one missense mutation was found, relating to the derived allele of OAR8:21,197,663 in the *RWDD1* gene ( $q = 0.002$ ;  $F_{ST} = 0.5$ ). The Polyphen score of this selected mutation, L23P, was low and did not support the large change in the protein function (Figure 3B).

Missense SNPs in the regions under selection supported by the haplotype statistics ( $H2/H1$ ,  $H12$ ) were only found in the *MC1R* gene in the lowland breeds, and in the upland breeds: *FGD3*, *HSPA5*, *ABCA13*, *PTPN1*, *ACP2*, *LOC101121718*, *NOC3L*, *VASH1*, *SIRPA*, and *UBA52* genes. Alternatively, some missense SNPs received strong support from Tajima's  $D$  and  $Pi$ , including one in the *GIGYF1* gene found in the selected regions in both the upland and lowland breeds, as well as the *SLC22A7* and *HSD3B7* genes in the upland breeds (Figure 3C).

In the upland breeds, three of the genes found with missense SNPs were also identified in the regions under selection defined in the upland breeds of the genotyping analysis. These included: *VASH1* ( $q = 0.003$ ) on OAR7, which also overlapped with BWM from the genotyping DCMS data; *UBA52* ( $q = 0.01$ ) on OAR13, which also overlapped with BHC from the genotyping data, and *SIRPA* ( $q = 0.007$ ). No missense mutations found in the lowland breeds overlapped with the regions under selection detected from the genotyping dataset.

## Candidate Genes in Welsh Lowland Sheep Breeds

*MC1R*, found in a region under selection in the lowland breeds (Table 3), is known to cause an upregulation of tyrosinase in hair melanocytes, which led to an increase of black eumelanin pigment; however, in sheep, typically, white pheomelanin is produced due to selection for mutations in *MC1R* (Weatherhead and Logan, 1981; Yang et al., 2013). Another gene, *CDK5RAP2*, with a strong support from allele frequency statistics, is related to human neurodevelopment by recruiting tubulin subunits, which are important for cortical gyration (Issa et al., 2013). Mutations in this gene have also been linked to Hertwig's anemia mutant mouse models displaying blood cytopenia, aneuploidy due to impaired cell-cycle spindle checkpoints, and increased neuronal cell death (Lizarraga et al., 2010). Finally, *CDHR3*, encoding a cell adhesion protein is linked to childhood asthma and rhinovirus-C susceptibility, was also located in a region under selection in Asian sheep by HapFLK (Fariello et al., 2014; Bochkov et al., 2015).

## Candidate Genes in Welsh Upland Sheep Breeds

Within the upland breeds, many genes found in the selected regions had functions related to cell stress and metabolism (Table 3). The most highly supported by DCMS, *RWDD1*, encodes a transcription factor related to sulfide metabolism in metazoans (Kang et al., 2008; Li et al., 2018). *HSPA5* also responds to cell stress mechanisms by promoting protein refolding in the endoplasmic reticulum in cancers or virally-infected cells (Booth et al., 2015). This gene has also been found under selection in

**TABLE 3** | Candidate genes and missense mutations in the lowland and upland sheep breeds.

OAR	Region start	Region end	SNP	Breed	Gene (rank)	Reference allele	Alternative allele	Mutation	PolyPhen	q-value	Function
2	2,887,916	2,887,916	2,887,916	Lowland	<i>CDK5RAP2</i> (1)	A*	G	M47V	0	0.006	Cell cycle
2	28,318,917	28,318,992		Upland	<i>FGD3</i> (1)					0.003	GTP binding
3	10,725,499	10,725,501		Upland	<i>HSPA5</i> (1)					0.007	Cell stress
4	7,476,204	7,476,211	7,476,211	Upland	<i>ABCA13</i> (1)	T*	C	K2619Q	0.9	0.006	Membrane transport
4	47,088,764	47,089,276		Lowland	<i>CDHR3</i> (1)					0.004	Cell adhesion
7	84,989,884	84,989,669		Upland	<i>VASH1</i> (1)					0.004	Angiogenesis
8	21,197,663	21,198,137	21,197,663	Upland	<i>RWDD1</i> (1)	A	G*	L23P	0.04	0.002	Anti-oxidation
13	49,771,782	49,771,855		Upland	<i>UBA52</i> (1)					0.1	Ubiquitinylation
13	52,848,831	52,848,943		Upland	<i>SIRPA</i> (1)					0.006	Cell recognition
13	52,848,831	52,848,943		Upland	<i>PTPN1</i> (1)					0.1	Insulin pathway
14	14,231,667	14,232,187		Lowland	<i>MC1R</i> (1)					0.004	Coat color
15	75,351,718	75,352,196		Upland	<i>ACP2</i> (1)					0.006	Development
20	10,644,119	10,644,602		Upland	<i>LOC10110726</i> (1)					0.009	Fat deposition
20	16,972,103	16,972,103		Upland	<i>SLC22A7</i> (1)					0.009	Cyclic nucleotide transport
22	15,368,530	15,368,926		Upland	<i>NOC3L</i> (1)					$2.0 \times 10^{-8}$	Cell cycle
24	27,106,763	27,106,886		Upland	<i>HSD3B7</i> (1)					0.002	Bile synthesis
24	35,794,428	35,794,428		Lowland/Upland	<i>GIGYF1</i> (1)					0.007	Insulin pathway

\*Denotes selected allele.



Chinese Yellow-Feathered chickens with regard to meat quality (Huang et al., 2020) and muscling in world pig breeds (Li et al., 2011).

Of the three genes found in the selected regions in both the genotyping and resequencing datasets, two related to cell stress mechanisms. The most significant region contains *VASH1*, encoding the vasohibin 1 signaling molecule, with known roles of negatively regulating angiogenesis (Chen et al., 2020), as well as promoting expression of antioxidation enzymes (Miyashita et al., 2012). Age-related downregulation of *VASH1* leads to a lower endothelial cell stress tolerance, posing as a risk factor for human vascular diseases in later life (Takeda et al., 2016). Secondly, *UBA52* encodes a protein with an ubiquitinate activity, with its downregulation causing cell cycle arrest and reduced protein synthesis essential for pre-implantation embryogenesis success in mouse models (Kobayashi et al., 2016).

Several genes related to metabolism and growth were found under selection in the upland breeds. The most significant of these, *HSD3B7*, is part of the *bile biosynthesis* pathway (Cheng et al., 2003). *PTPN1* is a risk-factor gene linked to diabetes and obesity (Olivier et al., 2004), which has a direct involvement in the *insulin* and *leptin signaling pathways*, and that mice lacking this gene were resistant to weight gain and intolerant to glucose (Elchebly et al., 1999). *GIGYF1* also has roles in enhancing the insulin receptor pathway, but additionally has been linked to translational repression (Giovannone et al., 2003; Tollenaere et al., 2019). Similar effects have been seen with *APC2*, which is linked to muscle mass in mice (Kärst et al., 2011). The Rho-GEF-containing gene *FGD3*, expressed in the growth plate of long bones, was previously found under selection in French Trotter and Gidran horses, as well as in Jutland and Japanese black cattle in association to birth weight (Takasuga et al., 2015; Grilz-Seger et al., 2019; Stronen et al., 2019).

Two membrane transport proteins, *ABCA13* and *SLC22A7*, were found to be in the regions under selection in the upland breeds. *ABCA13* encodes a member of the ATP-binding cassette membrane transporter family, responsible for the active transport of biological substrates across cell membranes (Prades et al., 2002). Secondly, *SLC22A7*, a transmembrane solute carrier, with roles in cAMP and cGMP transport in mammalian tissues and, therefore, is important for intracellular signaling which may mobilize intracellular  $Ca^{2+}$ , activate protein kinases, or activate transcription factors (Kobayashi et al., 2005; Yan et al., 2016). The final gene shared between the genotyping and resequencing data was *SIRPA*, which is expressed by macrophages and polarizes M1 phagocytic macrophages to M2 antiphagocytic macrophages, which is a key survival strategy for tumors (Barclay and van den Berg, 2014).

## Gene Ontology Enrichments Show Adaptations to Environment in the Resequenced Lowland and Upland Breeds

Nine functional category enrichments were found from genes within CNVRs in the lowland breeds and 14 enrichments in the upland breeds (Supplementary Tables 27,28). Some of these enrichments were shared between the lowland and

upland breeds. These were semaphorins; ion transport, pleckstrin homology domain; SAND domains; and EGF-like domains. Exclusive enrichments in the lowland breeds' CNVRs included cell surface receptors, neuromuscular process, and DNA binding whereas exclusive enrichments in the upland breeds included Src homology-3 domain Rho signal transduction, Notch signaling, and chondrocyte differentiation (1.4-fold). Genes within the regions under selection in the upland breeds showed significantly enriched clusters including interleukin-1 and ATP-binding (Supplementary Table 29). No functional category enrichments were found in genes in the regions under selection in the lowland breeds.

## DISCUSSION

Our study has demonstrated that regions under putative selection in Welsh sheep genomes contain candidate genes for adaptation to their local environment and production of socioeconomic traits. We used a large set of animals genotyped on a relatively small number of SNPs, applying the haplotype and point-based selection scan algorithms. This was combined with a relatively small number of resequenced individuals subjected to point-based selection scan. As a result, we detected genomic regions under selection in individual and groups of breeds, including candidate missense variants within these regions. Regions under selection detected by the three approaches followed the expected patterns seen previously that the haplotype-based approach would detect larger but fewer regions than the point-based approach, which detected smaller, but more numerous regions under selection (Yurchenko et al., 2019).

Exposure to altitudes has a range of deleterious effects caused by hypoxia, exposure to ultraviolet radiation, and generation of oxygen radicals. These, in turn, have been linked to a negative energy balance, dysregulated proteostasis, cellular stress mechanisms, and DNA damage (Askew, 2002; Pasiakos et al., 2017). Therefore, the presence of genes related to cell-stress and anti-oxidation in the regions under selection gives reassurance that the results from this study are relevant to the adaptations of Welsh sheep to altitudes. Genes related to hypoxia, however, were not identified in the regions under selection in the upland breeds, so it can be assumed that it is not a stress factor for these breeds because the altitudes they are farmed at are only moderate. Furthermore, body conditioning genes, such as *FDG3* and *ACP2*, in the upland breeds also suggest physical mechanisms of adaptation, such as increased fat deposition and muscle mass, however, these could also be linked to socioeconomic performance (Giovannone et al., 2003; Bento et al., 2004; Gu et al., 2011; Kärst et al., 2011; Grilz-Seger et al., 2019).

We observed selection at the loci of other top-ranking candidates from the haplotype analysis with known roles in energy consumption, liver metabolism, milking, fat deposition, and angiogenesis (Yang et al., 2016; Armstrong et al., 2018; Pszczola et al., 2018). These findings, showing the top-ranking genes sharing functions of that in the resequencing study, provide many candidate genes relating to survivability and socioeconomic traits in Welsh upland sheep.

Further evidence of selection in these breeds can be seen from functional term enrichments in genes found in the selected regions and in CNVRs.

Lowland breeds showed less signatures of selection, however, they mainly had selection in regions containing genes known to be associated with domestication, which are commonly reported signatures of selection in world sheep breeds (Kijas et al., 2012; Wei et al., 2015; Wang et al., 2019; Li et al., 2020). By demonstrating the lowland breeds sharing the signatures of selection with other productive breeds, this indicates that they are better suited for productive qualities but do not show an adaptation for the Welsh uplands. Despite this, selection for *CDK5RAP2* seen in the lowland breeds of the resequencing study may be linked to upland adaptation as mice models with truncating mutations have lower red blood cell counts, however, this is true for white blood cells too, and so, may be linked to immunity traits as well as neurodevelopment (Barker and Bernstein, 1983).

Differences in the results when using the genotyping and resequencing data were expected and can be attested to an increased density of SNPs in the resequencing data and a different composition of populations used for each study. The former effect would be expected to lead to narrower regions being detected under selection in the resequencing dataset, which could shift the most significant SNP away from candidate genes detected from the genotyping data. Secondly, using groups of multiple breeds could mean that whilst haplotypes often remain similar amongst closely related breeds, certain point mutations that differentiate those breeds from each other may become diluted in frequency, and so, would not be considered under selection by this method when a small number of animals is used. This is likely the case where, lowland breeds show less signatures of selection than upland breeds, suggesting that they share less signatures when grouped. This postulates that Welsh lowland breeds are more diverse than the upland breeds, supported by data from Beynon et al. (2015), which could be in response to the demand for socioeconomic traits and lack of selection pressures in comparison to the upland breeds, where there seem to be a selective pressure on the same region, leading to shared signatures of selection. This further suggests that the lowland breeds have been selected for the production of socioeconomic traits, rather than adaptation to their local environment.

Lower costs of genome resequencing have allowed a deeper insight to the individual mutations that could have functional roles within a region under selection; however, this is not always a realistic approach when investigating many related breeds in a single study. Our method here has demonstrated reliability in using resequencing data from a small number of individuals of different breeds but applied to similar environments can be supported by genotyping many individuals of these breeds. This is truer with the upland Welsh breeds, which is most likely

due to the higher environmental selective pressures applied when compared to the lowland breeds. This has greatly eluted candidate genes relating to hardiness and survivability in both the genotyping and resequencing data.

## CONCLUSION

Here, we have seen the first investigation into signatures of selection in Welsh sheep breeds using a large number of genotyped individuals and a small number of whole-genome resequenced individuals. Statistical pipelines have shown selection in Welsh upland breeds in regions containing genes relating to adaptation to the local environment, including candidate genetic variants, as well as some genes related to the production of socioeconomic traits in the lowland breeds. In turn, this information is useful, not only for the conservation of these culturally important breeds, but also for the improved production capabilities of mountain breeds and adaptation of productive breeds through a marker-assisted selection.

## DATA AVAILABILITY STATEMENT

The original contributions presented in the study are publicly available in NCBI using accession number PRJNA646642.

## AUTHOR CONTRIBUTIONS

DL: leading the project, sample collection, and writing the manuscript. JS-J: running the analyses and drafting the manuscript. VL: genome sequencing of Welsh sheep samples and initial analysis. AY: analysis pipeline development. NY: sample collection and manuscript editing. MS: Welsh sheep sequencing and initial analysis. All authors edited the manuscript.

## FUNDING

The collection and sequencing of Russian sheep samples used in this study were funded by the Russian Scientific Foundation grant 19-76-20026 to DL. Resources from HPC Wales and Supercomputing Wales supported the contributions of VL and MS.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2021.612492/full#supplementary-material>

## REFERENCES

Adzhubei, I., Jordan, D. M., and Sunyaev, S. R. (2013). Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Genet.* Chapter 7:Unit7.20. doi: 10.1002/0471142905.hg0720s76

Armstrong, E., Ciappesoni, G., Iriarte, W., Da Silva, C., Macedo, F., Navajas, E. A., et al. (2018). Novel genetic polymorphisms associated with carcass traits in grazing Texel sheep. *Meat Sci.* 145, 202–208. doi: 10.1016/j.meatsci.2018.06.014

Askew, E. W. (2002). Work at high altitude and oxidative stress: antioxidant nutrients. *Toxicology* 180, 107–119. doi: 10.1016/S0300-483X(02)00385-2

- Barclay, A. N., and van den Berg, T. K. (2014). The interaction between signal regulatory protein alpha (SIRP $\alpha$ ) and CD47: structure, function, and therapeutic target. *Annu. Rev. Immunol.* 32, 25–50. doi: 10.1146/annurev-immunol-032713-120142
- Barker, J., and Bernstein, S. (1983). Hertwig's anemia: characterization of the stem cell defect. *Blood* 61, 765–769. doi: 10.1182/blood.v61.4.765.bloodjournal614765
- Benavides, M. V., Sonstegard, T. S., Kemp, S., Mugambi, J. M., Gibson, J. P., Baker, R. L., et al. (2015). Identification of novel loci associated with gastrointestinal parasite resistance in a Red Maasai x Dorper backcross population. *PLoS One* 10:e0122797. doi: 10.1371/journal.pone.0122797
- Bento, J. L., Palmer, N. D., Mychaleckyj, J. C., Lange, L. A., Langefeld, C. D., Rich, S. S., et al. (2004). Association of protein tyrosine phosphatase 1B gene polymorphisms with type 2 diabetes. *Diabetes* 53:3007. doi: 10.2337/diabetes.53.11.3007
- Beynon, S. E., Slavov, G. T., Farré, M., Sunduimijid, B., Waddams, K., Davies, B., et al. (2015). Population structure and history of the Welsh sheep breeds determined by whole genome genotyping. *BMC Genet.* 16:65. doi: 10.1186/s12863-015-0216-x
- Bochkov, Y. A., Watters, K., Ashraf, S., Griggs, T. F., Devries, M. K., Jackson, D. J., et al. (2015). Cadherin-related family member 3, a childhood asthma susceptibility gene product, mediates rhinovirus C binding and replication. *Proc. Natl. Acad. Sci. U S A* 112, 5485–5490. doi: 10.1073/pnas.1421178112
- Bonhomme, M., Chevalet, C., Servin, B., Boitard, S., Abdallah, J., Blott, S., et al. (2010). Detecting selection in population trees: the Lewontin and Krakauer test extended. *Genetics* 186, 241–262. doi: 10.1534/genetics.104.117275
- Booth, L., Roberts, J. L., and Dent, P. (2015). HSPA5/Dna K may be a useful target for human disease therapies. *DNA Cell Biol.* 34, 153–158. doi: 10.1089/dna.2015.2808
- Bovo, S., Mazzoni, G., Bertolini, F., Schiavo, G., Galimberti, G., Gallo, M., et al. (2019). Genome-wide association studies for 30 haematological and blood clinical-biochemical traits in Large White pigs reveal genomic regions affecting intermediate phenotypes. *Sci. Rep.* 9:7003. doi: 10.1038/s41598-019-43297-1
- Bowles, D. (2015). Recent advances in understanding the genetic resources of sheep breeds locally adapted to the UK uplands: opportunities they offer for sustainable productivity. *Front. Genet.* 6. doi: 10.3389/fgene.2015.00024
- Bowles, D., Carson, A., and Isaac, P. (2014). Genetic distinctiveness of the Herdwick sheep breed and two other locally adapted hill breeds of the UK. *PLoS One* 9:e87823. doi: 10.1371/journal.pone.0087823
- Chen, C. Y., Salomon, A. K., Caporizzo, M. A., Curry, S., Kelly, N. A., Bedi, K. C., et al. (2020). Depletion of vasohibin 1 speeds contraction and relaxation in failing human cardiomyocytes. *Circ. Res.* 127, e14–e27. doi: 10.1161/CIRCRESAHA.119.315947
- Cheng, J. B., Jacquemin, E., Gerhardt, M., Nazer, H., Cresteil, D., Heubi, J. E., et al. (2003). Molecular genetics of 3 $\beta$ -hydroxy- $\delta$ -5-c27-steroid oxidoreductase deficiency in 16 patients with loss of bile acid synthesis and liver disease. *J. Clin. Endocrinol. Metab.* 88, 1833–1841. doi: 10.1210/jc.2002-021580
- Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., et al. (2011). The variant call format and VCFtools. *Bioinformatics* 27, 2156–2158. doi: 10.1093/bioinformatics/btr330
- Davies, E. (1935). Sheep farming in Upland Wales. *Geography* 20, 97–111.
- Dekker, N., Voorn-Brouwer, T., Verhoek, M., Wennekes, T., Narayan, R. S., Speijer, D., et al. (2011). The cytosolic  $\beta$ -glucosidase GBA3 does not influence type 1 Gaucher disease manifestation. *Blood Cells Mol. Dis.* 46, 19–26. doi: 10.1016/j.bcmd.2010.07.009
- Delaneau, O., Marchini, J., and Zagury, J.-F. (2011). A linear complexity phasing method for thousands of genomes. *Nat. Methods* 9, 179–181. doi: 10.1038/nmeth.1785
- Elchebly, M., Payette, P., Michaliszyn, E., Cromlish, W., Collins, S., Loy, A. L., et al. (1999). Increased insulin sensitivity and obesity resistance in mice lacking the protein tyrosine phosphatase-1B gene. *Science* 283:1544. doi: 10.1126/science.283.5407.1544
- Fariello, M. I., Boitard, S., Naya, H., SanCristobal, M., and Servin, B. (2013). Detecting signatures of selection through haplotype differentiation among hierarchically structured populations. *Genetics* 193, 929–941. doi: 10.1534/genetics.112.147231
- Fariello, M.-I., Servin, B., Tosser-Klopp, G., Rupp, R., Moreno, C., Consortium, I. S. G., et al. (2014). Selection signatures in worldwide sheep populations. *PLoS One* 9:e103813. doi: 10.1371/journal.pone.0103813
- Fernandez-Marcos, P. J., and Auwerx, J. (2011). Regulation of PGC-1 $\alpha$ , a nodal regulator of mitochondrial biogenesis. *Am. J. Clin. Nutr.* 93, 884S–890S. doi: 10.3945/ajcn.110.001917
- Garud, N. R., Messer, P. W., Buzbas, E. O., and Petrov, D. A. (2015). Recent selective sweeps in north american *Drosophila melanogaster* show signatures of soft sweeps. *PLoS Genet.* 11:e1005004. doi: 10.1371/journal.pgen.1005004
- Giovannone, B., Lee, E., Laviola, L., Giorgino, F., Cleveland, K. A., and Smith, R. J. (2003). Two novel proteins that are linked to insulin-like growth factor (IGF-I) receptors by the Grb10 adapter and modulate IGF-I signalling. *J. Biol. Chem.* 278, 31564–31573. doi: 10.1074/JBC.M211572200
- Grant, J. R., Arantes, A. S., Liao, X., and Stothard, P. (2011). In-depth annotation of SNPs arising from resequencing projects using NGS-SNP. *Bioinformatics* 27, 2300–2301. doi: 10.1093/bioinformatics/btr372
- Grilz-Seger, G., Neuditschko, M., Ricard, A., Velie, B., Lindgren, G., Mesarič, M., et al. (2019). Genome-wide homozygosity patterns and evidence for selection in a set of European and Near Eastern horse breeds. *Genes (Basel)* 10:491. doi: 10.3390/genes10070491
- Gu, X., Feng, C., Ma, L., Song, C., Wang, Y., Da, Y., et al. (2011). Genome-wide association study of body weight in chicken F2 resource population. *PLoS One* 6:e21872. doi: 10.1371/journal.pone.0021872
- Heaton, M. P., Leymaster, K. A., Kalbfleisch, T. S., Kijas, J. W., Clarke, S. M., McEwan, J., et al. (2014). SNPs for parentage testing and traceability in globally diverse breeds of sheep. *PLoS One* 9:e94851. doi: 10.1371/journal.pone.0094851
- Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* 4, 44–57. doi: 10.1038/nprot.2008.211
- Huang, X., Otecko, N. O., Peng, M., Weng, Z., Li, W., Chen, J., et al. (2020). Genome-wide genetic structure and selection signatures for color in 10 traditional Chinese yellow-feathered chicken breeds. *BMC Genomics* 21:316. doi: 10.1186/s12864-020-6736-4
- Issa, L., Mueller, K., Seufert, K., Kraemer, N., Rosenkotter, H., Ninnemann, O., et al. (2013). Clinical and cellular features in patients with primary autosomal recessive microcephaly and a novel CDK5RAP2 mutation. *Orphanet J. Rare Dis.* 8:59. doi: 10.1186/1750-1172-8-59
- Kang, N., Chen, D., Wang, L., Duan, L., Liu, S., Tang, L., et al. (2008). Rwd11, a thymus aging related molecule, is a new member of the intrinsically unstructured protein family. *Cell. Mol. Immunol.* 5, 333–339. doi: 10.1038/cmi.2008.41
- Kärst, S., Cheng, R., Schmitt, A. O., Yang, H., de Villena, F. P. M., Palmer, A. A., et al. (2011). Genetic determinants for intramuscular fat content and water-holding capacity in mice selected for high muscle mass. *Mamm. Genome* 22, 530–543. doi: 10.1007/s00335-011-9342-6
- Kijas, J. W., Lenstra, J. A., Hayes, B., Boitard, S., Porto Neto, L. R., San Cristobal, M., et al. (2012). Genome-wide analysis of the world's sheep breeds reveals high levels of historic mixture and strong recent selection. *PLoS Biol.* 10:e1001258. doi: 10.1371/journal.pbio.1001258
- Kim, E.-S., Elbeltagy, A. R., Aboul-Naga, A. M., Rischkowsky, B., Sayre, B., Mwacharo, J. M., et al. (2016). Multiple genomic signatures of selection in goats and sheep indigenous to a hot arid environment. *Heredity (Edinb)* 116, 255–264. doi: 10.1038/hdy.2015.94
- Klambauer, G., Schwarzbauer, K., Mayr, A., Clevert, D.-A., Mitterecker, A., Bodenhofer, U., et al. (2012). cn.MOPS: mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. *Nucleic Acids Res.* 40:e69. doi: 10.1093/nar/gks003
- Kobayashi, M., Oshima, S., Maeyashiki, C., Nibe, Y., Otsubo, K., Matsuzawa, Y., et al. (2016). The ubiquitin hybrid gene UBA52 regulates ubiquitination of ribosome and sustains embryonic development. *Sci. Rep.* 6:36780. doi: 10.1038/srep36780
- Kobayashi, Y., Ohshiro, N., Sakai, R., Ohbayashi, M., Kohyama, N., and Yamamoto, T. (2005). Transport mechanism and substrate specificity of human organic anion transporter 2 (hOat2 [SLC22A7]). *J. Pharm. Pharmacol.* 57, 573–578. doi: 10.1211/0022357055966



- Li, C., Sun, D., Zhang, S., Yang, S., Alim, M. A., Zhang, Q., et al. (2016). Genetic effects of FASN, PPARGC1A, ABCG2 and IGF1 revealing the association with milk fatty acids in a Chinese Holstein cattle population based on a post genome-wide association study. *BMC Genet.* 17:110. doi: 10.1186/s12863-016-0418-x
- Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv [preprint]* arXiv:1303.3997
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352
- Li, X., Kim, S.-W., Do, K.-T., Ha, Y.-K., Lee, Y.-M., Yoon, S.-H., et al. (2011). Analyses of porcine public SNPs in coding-gene regions by re-sequencing and phenotypic association studies. *Mol. Biol. Rep.* 38, 3805–3820. doi: 10.1007/s11033-010-0496-1
- Li, X., Liu, X., Qin, Z., Wei, M., Hou, X., Zhang, T., et al. (2018). A novel transcription factor Rwdd1 and its SUMOylation inhibit the expression of sqrr, a key gene of mitochondrial sulfide metabolism in *Urechis unicinctus*. *Aquat. Toxicol.* 204, 180–189. doi: 10.1016/j.aquatox.2018.09.012
- Li, X., Yang, J., Shen, M., Xie, X.-L., Liu, G.-J., Xu, Y.-X., et al. (2020). Whole-genome resequencing of wild and domestic sheep identifies genes associated with morphological and agronomic traits. *Nat. Commun.* 11:2815. doi: 10.1038/s41467-020-16485-1
- Lizarraga, S. B., Margossian, S. P., Harris, M. H., Campagna, D. R., Han, A.-P., Blevins, S., et al. (2010). Cdk5rap2 regulates centrosome function and chromosome segregation in neuronal progenitors. *Development* 137, 1907–1917. doi: 10.1242/dev.040410
- Ma, Y., Ding, X., Qanbari, S., Weigend, S., Zhang, Q., and Simianer, H. (2015). Properties of different selection signature statistics and a new strategy for combining them. *Heredity (Edinb)* 115, 426–436. doi: 10.1038/hdy.2015.42
- Marnetto, D., and Huerta-Sánchez, E. (2017). Haplostrips: revealing population structure through haplotype visualization. *Methods Ecol. Evol.* 8, 1389–1392. doi: 10.1111/2041-210X.12747
- McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20, 1297–1303. doi: 10.1101/gr.107524.110
- Miyashita, H., Watanabe, T., Hayashi, H., Suzuki, Y., Nakamura, T., Ito, S., et al. (2012). Angiogenesis inhibitor vasohibin-1 enhances stress resistance of endothelial cells via induction of SOD<sub>2</sub> and SIRT1. *PLoS ONE* 7, e46459–e46459. doi: 10.1371/journal.pone.0046459
- Naval-Sanchez, M., Nguyen, Q., McWilliam, S., Porto-Neto, L. R., Tellam, R., Vuocolo, T., et al. (2018). Sheep genome functional annotation reveals proximal regulatory elements contributed to the evolution of modern breeds. *Nat. Commun.* 9:859. doi: 10.1038/s41467-017-02809-1
- Nei, M., and Li, W. H. (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc. Natl. Acad. Sci. U.S.A.* 76, 5269–5273. doi: 10.1073/pnas.76.10.5269
- Olivier, M., Hsiung, C. A., Chuang, L.-M., Ho, L.-T., Ting, C.-T., Bustos, V. I., et al. (2004). Single nucleotide polymorphisms in protein tyrosine phosphatase 1beta (PTPN1) are associated with essential hypertension and obesity. *Hum. Mol. Genet.* 13, 1885–1892. doi: 10.1093/hmg/ddh196
- Pasiakos, S., Berryman, C., Carrigan, C., Young, A., and Carbone, J. (2017). Muscle protein turnover and the molecular regulation of muscle mass during hypoxia. *Med. Sci. Sport. Exerc.* 49, 1340–1350. doi: 10.1249/mss.0000000000001228
- Periasamy, K., Pichler, R., Poli, M., Cristel, S., Cetrá, B., Medus, D., et al. (2014). Candidate gene approach for parasite resistance in sheep—variation in immune pathway genes and association with fecal egg count. *PLoS One* 9:e88337. doi: 10.1371/journal.pone.0088337
- Prades, C., Arnould, I., Annino, T., Shulenin, S., Chen, Z. Q., Orosco, L., et al. (2002). The human ATP binding cassette gene ABCA13, located on chromosome 7p12.3, encodes a 5058 amino acid protein with an extracellular domain encoded in part by a 4.8-kb conserved exon. *Cytogenet. Genome Res.* 98, 160–168. doi: 10.1159/000069852
- Pszczola, M., Strabel, T., Mucha, S., and Sell-Kubiak, E. (2018). Genome-wide association identifies methane production level relation to genetic control of digestive tract development in dairy cows. *Sci. Rep.* 8:15164. doi: 10.1038/s41598-018-33327-9
- Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81, 559–575. doi: 10.1086/519795
- Rahman, M. M., Bagdassarian, E., Ali, M. A. M., and McFadden, G. (2017). Identification of host DEAD-box RNA helicases that regulate cellular tropism of oncolytic Myxoma virus in human cancer cells. *Sci. Rep.* 7:15710. doi: 10.1038/s41598-017-15941-1
- Ryder, M. (1964). The history of sheep breeds in Britain. *Agric. Hist. Rev.* 12, 1–12. doi: 10.2307/40273081
- Scheet, P., and Stephens, M. (2006). A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am. J. Hum. Genet.* 78, 629–644. doi: 10.1086/502802
- Stronen, A. V., Pertoldi, C., Iacolina, L., Kadarmideen, H. N., and Kristensen, T. N. (2019). Genomic analyses suggest adaptive differentiation of northern European native cattle breeds. *Evol. Appl.* 12, 1096–1113. doi: 10.1111/eva.12783
- Sweet-Jones, J., Yurchenko, A. A., Igoshin, A. V., Yudin, N. S., Swain, M. T., and Larkin, D. M. (2021). Resequencing and signatures of selection scan in two Siberian native sheep breeds point to candidate genetic variants for adaptation and economically important traits. *Anim. Genet.* 52, 126–131. doi: 10.1111/age.13015
- Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123, 585–595. doi: 10.1093/genetics/123.3.585
- Takasuga, A., Sato, K., Nakamura, R., Saito, Y., Sasaki, S., Tsuji, T., et al. (2015). Non-synonymous FGD3 variant as positional candidate for disproportional tall stature accounting for a carcass weight QTL (CW-3) and skeletal dysplasia in Japanese Black cattle. *PLoS Genet.* 11:e1005433. doi: 10.1371/journal.pgen.1005433
- Takeda, E., Suzuki, Y., and Sato, Y. (2016). Age-associated downregulation of vasohibin-1 in vascular endothelial cells. *Aging Cell* 15, 885–892. doi: 10.1111/ace.12497
- Tollenaere, M. A. X., Tiedje, C., Rasmussen, S., Nielsen, J. C., Vind, A. C., Blasius, M., et al. (2019). GIGYF1/2-driven cooperation between ZNF598 and TTP in posttranscriptional regulation of inflammatory signaling. *Cell Rep.* 26, 3511–3521.e4. doi: 10.1016/j.celrep.2019.03.006
- Turner, S. D. (2014). qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. *bioRxiv [preprint]* doi: 10.1101/005165 bioRxiv 005165.
- Wang, W., Zhang, X., Zhou, X., Zhang, Y., La, Y., Zhang, Y., et al. (2019). Deep Genome resequencing reveals artificial and natural selection for visual deterioration, plateau adaptability and high prolificacy in Chinese domestic sheep. *Front. Genet.* 10:300. doi: 10.3389/fgene.2019.00300
- Weatherhead, B., and Logan, A. N. N. (1981). Interaction of  $\alpha$ -melanocyte-stimulating hormone, melatonin, cyclin AMP and cyclic GMP in the control of melanogenesis in hair follicle melanocytes in vitro. *J. Endocrinol.* 90, 89–96. doi: 10.1677/joe.0.0900089
- Wei, C., Wang, H., Liu, G., Wu, M., Cao, J., Liu, Z., et al. (2015). Genome-wide analysis reveals population structure and selection in Chinese indigenous sheep breeds. *BMC Genomics* 16:194. doi: 10.1186/s12864-015-1384-9
- Weir, B. S., and Cockerham, C. C. (1984). Estimating F-statistics for the analysis of population structure. *Evolution (N. Y)* 38, 1358–1370. doi: 10.2307/2408641
- Williams-Davies, J. (1981). *Welsh Sheep and Their Wool*, 1st Edn. Dyfed: Gormer Press.
- Yan, K., Gao, L.-N., Cui, Y.-L., Zhang, Y., and Zhou, X. (2016). The cyclic AMP signalling pathway: Exploring targets for successful drug discovery (Review). *Mol. Med. Rep.* 13, 3715–3723. doi: 10.3892/mmr.2016.5005
- Yang, G.-L., Fu, D.-L., Lang, X., Wang, Y.-T., Cheng, S.-R., Fang, S.-L., et al. (2013). Mutations in MC1R gene determine black coat color phenotype in Chinese sheep. *Sci. World J.* 2013:675382. doi: 10.1155/2013/675382
- Yang, J., Li, W.-R., Lv, F.-H., He, S.-G., Tian, S.-L., Peng, W.-F., et al. (2016). Whole-Genome sequencing of native sheep provides insights into rapid adaptations to extreme environments. *Mol. Biol. Evol.* 33, 2576–2592. doi: 10.1093/molbev/msw129



- Yuan, Z., Li, W., Li, F., and Yue, X. (2019). Selection signature analysis reveals genes underlying sheep milking performance. *Arch. Anim. Breed.* 62, 501–508. doi: 10.5194/aab-62-501-2019
- Yurchenko, A. A., Deniskova, T. E., Yudin, N. S., Dotsev, A. V., Khamiruev, T. N., Selionova, M. I., 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. doi: 10.1186/s12864-019-5537-0
- Zeder, M. A. (2008). Domestication and early agriculture in the Mediterranean Basin: origins, diffusion, and impact. *Proc. Natl. Acad. Sci.* 105, 11597–11604. doi: 10.1073/pnas.0801317105

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2021 Sweet-Jones, Lenis, Yurchenko, Yudin, Swain and Larkin. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.