The first sheep graph pan-genome reveals the spectrum of structural variations and their effects on different tail phenotypes

Ran Li[1#], Mian Gong[1#], Xinmiao Zhang[1#], Fei Wang[1#], Zhenyu Liu[1#], Lei Zhang[1#], Mengsi Xu[2], Yunfeng Zhang[2], Xuelei Dai[1], Zhuangbiao Zhang[1], Wenwen Fang[1], Yuta Yang[1], Huanhuan Zhang[1], Weiwei Fu[1], Chunna Cao[1], Peng Yang[2], Zeinab Amiri Ghanatsaman[3], Niloufar Jafarpour Negari[4], Hojjat Asadollahpour Nanaei[1,4], Xiangpeng Yue[5], Yuxuan Song[1], Xianyong Lan[1], Weidong Deng[6], Xihong Wang[1], Ruidong Xiang[7], Eveline M. Ibeagha-Awemu[8], Pat (J.S.) Heslop-Harrison[9], Johannes A. Lenstra[10], Shangquan Gan[2]*, Yu Jiang[1]*

[1]Key Laboratory of Animal Genetics, Breeding and Reproduction of Shaanxi Province, College of Animal Science and Technology, Northwest A&F University, Yangling, Shaanxi, 712100, China

[2]State Key Laboratory of Sheep Genetic Improvement and Healthy Production, Xinjiang Academy of Agricultural and Reclamation Sciences, Shihezi, Xinjiang 832000, China

[3]Department of Animal Science, Fars Agricultural and Natural Resources Research and Education Center, Agricultural Research, Education & Extension Organization (AREEO), Shiraz, Iran

[4]Department of Animal Science, Shahid Bahonar University of Kerman, Kerman, Iran

[5]State Key Laboratory of Grassland Agro-ecosystems, Key Laboratory of Grassland Livestock Industry Innovation, Ministry of Agriculture and Rural Affairs, Engineering Research Center of Grassland Industry, Ministry of Education, College of Pastoral Agriculture Science and Technology, Lanzhou University, Lanzhou, 730020, China

[6] Faculty of Animal Science and Technology, Yunnan Agricultural University, Kunming, 650201, China

[7]Faculty of Veterinary & Agricultural Science, The University of Melbourne, Parkville, 3052, VIC, Australia

[8]Agriculture and Agri-Food Canada, Sherbrooke Research and Development Centre, Sherbrooke, QC J1M 0C8, Canada

[9]Department of Genetics and Genome Biology, University of Leicester, Leicester, UK

[10]Faculty of Veterinary Medicine, Utrecht University, Utrecht, The Netherlands

[#]These authors contributed equally.

*To whom correspondence should be addressed.

Email: yu.jiang@nwafu.edu.cn; shangquangan@163.com

Abstract

Structural variations (SVs) are a major contributor of genetic diversity and phenotypic variations, however their prevalence and functions in domestic animals are largely unexplored. Here, we assembled 26 haplotype-resolved genome assemblies from 13 genetically diverse sheep breeds using PacBio HiFi sequencing. We then constructed an ovine graph pan-genome and demonstrated its advantage in discovering 142,593 biallelic SVs (Insertions and deletions), 7,028 divergent alleles and 13,419 multiallelic variations with high accuracy and sensitivity. To link the SVs to genotypes, we genotyped the SVs in 687 resequenced individuals of domestic and wild sheep using a graph-based approach and identified numerous population-stratified variants, of which expression-associated SVs were detected by integrating RNA-seq data. Taking the varying sheep tail morphology as example, we located a putative causative insertion in *HOXB13* gene responsible for the long tail and reported multiple large SVs associated with the fat tail. Beyond generating a benchmark resource for ovine structural variants, our study also highlighted that the population genetics analysis based on graph pan-genome rather than reference genome will greatly benefit the animal genetic research.

Keywords: sheep pan-genome, structural variations, *de novo* assembly, tail morphology

Introduction

Structural variations (SVs) are a major source of genetic variation that range from 50 base pairs (bp) to over megabases (Mb) in size. Compared to single-nucleotide polymorphisms (SNPs), SVs could cause larger scale genomic perturbations in genes and regulatory regions and are therefore more likely to affect gene expression and phenotypes [1, 2].

However, SV detection has been challenging primarily due to technical limitations [3]. The approach based on short reads (100-150 bp) causes poor sensitivity and high false discovery rate [4]. To overcome the limitation of short reads, the long-read sequencing such as PacBio SMRT sequencing and Oxford Nanopore sequencing produce long reads (>10 kb) and greatly facilitated the efficacy of SV discovery by directly spanning the breakpoints [5-9]. Despite that, the conventional long reads suffer from high sequencing error rate (5-15%), which not only implicated downstream analysis but also hampered discernment in highly repetitive regions. In addition, the linear reference assembly has been shown to be not fully compatible with the methodology of SV detection [10, 11].

The emergence of PacBio high fidelity (HiFi) reads is revolutionizing our methodologies in *de novo* assembly and SV detection [12]. As a major leap forward in long-read sequencing, the HiFi reads are both long（15 kb – 25 kb）and highly accurate（>99.9% base accuracy）[13], offering base-level resolution at SV breakpoints. Furthermore, the HiFi sequencing could deliver haplotype-resolved assemblies to depict the sequence diversity of the diploid genome [14], which are essential for pan-genomic studies. Compared with a linear reference genome, the pan-genomic reference models can now maximumly utilize the information from haplotype-resolved genome assemblies to faithfully capture complex regions and SVs in the sequenced individual [11, 15]. Yet the superiority of HiFi sequencing in pan-genome construction and SV detection has just revealed [6, 16, 17]. The analysis of *de novo* assemblies at population-scale using HiFi sequencing is going to transform our

understanding of SV complexity in each species.

Sheep are among the first domesticated animals with economic and cultural importance. Previous studies have shown that SVs determine the white coat color [18], fleece variations [19] and horn phenotypes (polledness) [20] – characters that have been under selection for millennia. Nevertheless, a comprehensive characterization of genetic variation characterized by SVs remains to be explored in sheep as in most animal species. In this study, we generated reference-grade *de novo* assemblies for 13 representative Eurasian sheep breeds, together with two sets of partially phased assemblies for each individual taking advantage of Pacbio HiFi sequencing. These datasets allow us to construct a graph-based ovine pangenome and explore the SV landscape with unprecedent accuracy and sensitivity. We further analyzed the SV spectrum in 686 domestic and wild sheep from short-reads sequencing and highlighted SVs as largely unexplored contributors to variations in gene expression and phenotypes.

Results

High-quality *de novo* assemblies of 13 representative sheep breeds

Domestic sheep could be generally divided into five genetic groups including European, Asian, Middle Eastern and two African lineages (one of the Dorper sheep and the other of the African landraces) [21]. We selected 13 genetically diverse accessions to cover the major genetic groups, including seven European, four East Asian, one Middle Eastern (Kermani sheep from Iran) and one African breeds (White dorper). The seven distinct European breeds covered those for meat (Suffolk, Dorset, Texel, Charollais), wool (Merino, Romney) and dairy (East Friesian) purposes and the East Asian accessions covered all the three lineages of Northern China (Ujumqin, Kazak), Qinghai-Tibetan Plateau (Tibetan) and Yunnan-Kweichow Plateau (Yunnan). The 13 individuals displayed distinct phenotypes and a good representation especially for the Eurasian breeds. A SNP analysis showed that they encompass 95.5% of the SNPs with minor allele frequency (MAF) >0.05 among the studied 645 domestic

sheep worldwide.

To obtain high-quality *de novo* assemblies, the 13 accessions were sequenced using PacBio HiFi sequencing at an average coverage of 20.9×, yielding highly accurate long reads with per-base accuracy >99.9%. We used the Hifiasm assembler **[14]** for *de novo* assembly, which could generate the primary contigs as well as two sets of partially phased contig assemblies from HiFi reads. We ultimately assembled 13 primary collapsed assemblies with average length of 2.86 Gb (2.81 Gb- 2.92 Mb) (Table 1). The average contig number were 1582 with contig N50 length of 68.2 Mb. The most complete assembly (FRI1) achieved a final contig N50 of 85.1 Mb which is much higher than the sheep reference genome (ARS-UI_Ramb_v2.0). We found that the HiFi sequencing depth is not correlated with assembly length (r^2=-0.04, Spearman rank correlation), but positively correlated with contig N50 length (r^2=0.93, Spearman rank correlation) and negatively correlated with contig N50 number (r^2=-0.67, Spearman rank correlation). We further ordered and oriented contigs of the 13 collapsed assemblies into chromosomes based on their synteny to the reference genome. As a result, on average 93.58% of the contigs per assembly were placed into chromosomes. The resulting chromosome-level collapsed assemblies for the 13 representative breeds will serve as a reference panel for future research in sheep genetics and genomics. Meanwhile, the partially phased assemblies displayed an average contig N50 length of 14.5 Mb, which could capture the variations of the diploid and will be used for the subsequent construction of the multiassembly graph.

We assessed the completeness using Benchmarking Universal Single-Copy Orthologs (BUSCO) analysis and the base accuracy using quality value (QV) for the collapsed and partially phased assemblies (Table 1). The collapsed assemblies displayed an average completeness of 94.1% (93.8%-94.4%) whereas the BUSCO scores for the partially phased assemblies were lower but still consistently >90% complete. To measure the QV of the assemblies, the 13 accessions were also sequenced to an average of 30× depth Illumina short-reads. All the collapsed and partially phased assemblies displayed higher quality value than that of the sheep reference genome assembly (QV=42.5) except for WDU1.

Construction of sheep graph pan-genome

To build a comprehensive sheep graph pan-genome, we used the reference genome (ARS-UI_Ramb_v2.0) as the backbone and the 26 haplotype-resolved assemblies from the 13 individuals were augmented into the graph individually. The partially phased assemblies, instead of the primary contigs, were used as they would encompass more comprehensive sequence variations of a haploid genome. A multigraph genome was generated using minigraph [22] which is one of the state-of-art graph constructors especially for discovering complex variations. Only autosomal sequences were considered in the multigraph genome, as the X and Y chromosomes of males were underrepresented in the sequencing.

The resulting multiassembly graph spanned 2.75 Gb including 167,565 nonreference nodes of 137.7 Mb. Based on the presence and absence of each node among reference genome and the 26 assemblies, we classified the graph pangenome into three categories: core (nodes present in all assemblies), flexible (nodes present in at least two assemblies) and private (nodes present in one assembly). The core and the flexible genome accounted for 69.6% and 29.0% of the total size, respectively, whereas the private part was minimal (1.4%) (Fig. 1B). The variations were most associated with LINEs following by those with mixed interspersed repeats (Fig. 1C).

The graph genome encompassed 91,848 structural variations with length ≥100 bp, representing bubbles that are located between a single source and a single sink node coming from the reference assembly. The paths in a bubble below 100 bp were not considered, since we found that the small variants contain abundant false positives by manual inspection. On the other hand, the extended graph coordinate system proposed by minigraph enabled us to annotate complex variations including those in highly diverse regions. We classified the structural variations into three categories: (1) biallelic variations (Insertions and deletions); (2) divergent alleles where a non-reference node of ≥100 bp from a single path were found; (3) multiallelic variations where more than one non-reference paths were present (Fig. 1D and Supplementary data 1 and 2). Notably, from the multiassembly graph, we identified 7028 divergent

loci adding up to 11.7 Mb and another 13，419 multiallelic loci of 33.3 Mb. The median length of divergent alleles (726 bp) and multiallelic loci (575 bp) were longer than that of biallelic variations (452 bp).

The major advantage of the multiassembly graph does not lie in the efficacy in discovering regular biallelic variations (insertions/deletions); instead, it is most efficient in recovering complex SVs, i.e., the divergent and multiallelic variations. These complex SV could be an import type of structural variations that can be trait-associated but are generally neglected by most SV callers. The divergent variations intersected with 1797 genes with the non-reference alleles predicted to harbor 27 genes. A divergent locus containing MYADM-like repeat was previously reported to be associated with mean corpuscular hemoglobin concentration traits and weight of lamb weaned [23], but the full sequence of the divergent allele was not resolved. We found this divergent allele to be 50.7 kb long and was presented in 14/26 of the partially phased haplotypes (Fig. 1E). We also noticed several large divergent alleles that affected genic region of insulin degrading enzyme (IDE), Membrane Associated Guanylate Kinase, WW And PDZ Domain Containing 2 (MAGI2), placental prolactin-related protein 3-like (Supplementary fig. 1). The multiallelic loci intersected 5092 genes and the non-reference alleles themselves were predicted to harbor 161 genes. One of the largest multiallelic region represented a deletion of 105.6 kb which is nested with an insertion of 5.5 kb (Fig. 1F). The deletion likely result in the loss of intelectin gene, that displays specific expression in lung of sheep and potentially plays a role in mucosal response in allergy [24]. Another loci (45.6 kb) harboring beta hemoglobin genes was previously reported as divergent allele and associated with tolerance of anemia [25], but it actually also belonged to multiallelic variation (Supplementary fig. 2). We provided additional examples of multiallelic loci that would result in gene gain/loss, such as ASIP, UDP-glucuronosyltransferase 2B18-like, olfactory receptor (LOC101115700), etc. (Supplementary fig. 3).

A comprehensive catalog and frequency spectrum of SVs in sheep population

Despite advantages in discovering divergent alleles and multiallelic variations, the

graph genome generated by minigraph does not involve base alignment, rendering accurate breakpoint assignments for many structural variations. In addition, minigraph is not well adapted for discovering small SVs, i.e. SVs below 100 bp [14]. We therefore used pbsv (https://github.com/PacificBiosciences/pbsv), a read based SV caller, to jointly call SVs among the 13 individuals. As a complementing strategy to the pan-genome, the reads-based approach could determine the precise breakpoints and included SVs below 100 bp. Aligning PacBio HiFi reads to the reference genome generated 147,832 SVs including 142,450 insertions/deletions, 516 inversions and 4,866 duplications (Supplementary data 3). We observed much more insertions than deletions, which contrasts with results based on short read sequencing where deletions are typically more frequent and easier to identify [26]. The pbsv call set covers 94.8% of the biallelic variations from minigraph; we therefore merged the pbsv call set and the large variants (>5 kb) with from minigraph that did not overlap with pbsv call set, resulting a final comprehensive catalogue of 147,975 SVs (insertions and deletions) in sheep genome. It should be noted that the SV catalogue include the 516 inversions, 4,866 duplications 7,028 divergent alleles and 13,419 multiallelic variations, which however were not subjected to the subsequent population-level genotyping using short-reads data.

The number of variants rapidly decreases with length, consistent with the reports in humans using long-read sequencing [5, 6, 27]. We observed noticeable peaks at 150 bp corresponding to SINES, 1,350 bp and 1,600 bp corresponding to LTR, as well as 7,750 bp corresponding to LINES (Fig 2A). The SVs were most underrepresented in untranslated region (UTR), coding sequences (CDS), and regulatory elements (H3K4me3, H3K27ac), suggesting strong negative selection that acted against them (Supplementary fig. 4). We also reported 201 SV hotspots sites covering 18.2 Mb of the autosomes (Supplementary fig. 5). The SVs were significantly clustered at the terminal 5 Mb region of each chromosome with 4.1-fold enrichment (p value<0.01, permutation test) (Supplementary fig. 6).

In order to better understand the frequency spectrum of the SVs and identify those under selection, we used the graph genotyping software Paragraph [28] to

genotype the variants in a panel of 687 resequenced samples. Only the insertions and deletions from autosomes were genotyped. Paragraph represent SVs as diverging paths in a graph and could achieve high rate of genotyping accuracy [29]. The samples included 644 domestic sheep from Europe, East Asia Middle East and Africa , 33 Asiatic mouflons and 8 argali with an average coverage of 18.1× (Fig. 2B).

To ensure a high-quality set of genotyped SVs for population genetics analysis, we filtered genotyping results based on genotyping rates and adherence to Hardy Weinberg equilibrium [29]. We excluded 12,705 SVs that failed to be genotyped in >90% of the samples and another 2011 displaying deviation from Hardy Weinberg equilibrium by displaying excess of heterozygotes, leaving 121,258 (89.2%) reliably genotyped SVs for downstream analysis (Supplementary data 4). We excluded the SVs not because that they are unreliable but because they cannot be correctly genotyped probably due to the limitation of short reads. We also estimated the genotyping accuracy of the remaining SVs by measuring the Mendelian errors, which represent genotypes that are found in the offspring but could not be inherited form either parent. The Mendelian error rate was 3.8% on average in the 11 trios of 29 individuals (each ~30x), suggesting high accuracy of the genotyping results.

PCA analysis using either insertions or deletions displays similar population structure and shows clear separation of the four groups, and in many cases, even individual breeds could be distinguished (Supplementary fig. 7), agreeing with the clusters defined by SNPs. Therefore, the SV catalogue represented a broad diversity of worldwide sheep population.

The level of linkage disequilibrium (LD) between SVs and SNPs could reflect the extent of SVs representing independent and undiscovered genetic variations relative to previous evolutionary studies. We estimated how frequently the SVs were linked to the flanking SNPs. About 41.8% SVs link strongly with surrounding SNPs (rank>200), the characteristic of which could be detected by the surrounding linked SNPs. It is worth noting that there are still 37.8% SVs display low levels of LD with nearby SNPs (rank < 100), implying variants that are poorly tagged by easily

genotyped SNP markers (Supplementary fig. 8).

In order to infer the allele state (ancestral or derived) for each SV, we additionally genotyped one snow sheep and one bighorn sheep and, together with one argali as outgroup, and could identify the state for 115,089 SVs. Based on the allele frequency of derived alleles (DAF) in domestic sheep, the variants were classified into four categories: shared (DAF ≥ 0.95), major (0.95 > DAF ≥ 0.5), polymorphic (0.5 > DAF ≥ 0.05) and rare (DAF < 0.05) SVs (Fig. 2C).

We next dissected *de novo* SVs in particular those arising to high frequency in the domestication and improvement process. The *de novo* SVs accounted for 28% of all the SVs (Fig. 2D). It should be noted that the proportion of *de novo* SVs would be overestimated due to the relative limited number of wild sheep. Among them, we found 592 high-frequency *de novo* SVs (DAF ≥ 0.5 in at least one geographic group) (Fig. 2E). The *de novo* SVs reaching high frequency in all domestic sheep might be associated with domestication. Those SVs were found to be enriched in KEGG pathways of Ras signaling and long-term depression (adjusted p value < 0.05). For example, a *de novo* mutation of 3.4 kb insertion dominating in domestic sheep were identified in the first intron of *GRID2* (Fig. 2F). The gene plays a key role in synaptogenesis and neurodevelopment [30]. Those that rapidly arose in certain geographic regions might be involved in the emergence of some domesticated traits. A 1.4 kb *de novo* insertion prevalent in European sheep was found at the 3' UTR of *IRF2BP2* gene (Fig. 2G), which was reported as the causative mutation for the woolly hair [19].

Population-stratified SVs and their potential effect on gene expression

After initial domestication at the Fertile Crescent, domestic sheep dispersed worldwide and adapted to a wide range of natural and artificial selection. In this context, we detected putative selection signals by dissecting population-stratified SVs among the 33 breeds/populations. The population differentiation was measured among each breed versus all others breeds using a modified calculation of global Fst (see methods). It should be noted that this strategy tends to discover common

selection signals present in multiple breeds. Therefore, the population-stratified SVs between Europe and East Asian breeds are preferentially detected since they are the predominant breeds in the comparison.

We identified 865 population-stratified SVs with the top 1% global Fst values (Supplementary data 4), of which 351 intersected with genic regions and 324 were within 100 kb of gene flanking region (Fig. 3A). The majority of them (698, 80.7%) were confirmed by the Ohana approach with which we detected selection in multiple populations by varying the number of ancestral admixture components. The related genes were associated with diverse phenotypic and production traits, including wool type (IRF2BP2, FGF7) [19, 31], tail morphology (HOXB13) [32], horn type (RXFP2) [33], ear type (MSRB3) [34], fat deposition (BMP2, PDGFD, PDGFA) [35], coat color (KIT) [36], etc. Notably, the highest stratified SV (Fst=0.78) was the insertion as shown in Fig. 2G, representing the most differentiated SV among European and Asian breeds. The next two highest signals affecting HOXB13 and BMP2 were both associated with tail morphology which will be described in following sections. Functional annotation of the putatively affected genes revealed enrichment in KEGG pathways involved in important signaling pathways (Rap1 signaling pathway, MAPK signaling pathway), parasitic disease (African trypanosomiasis, amoebiasis), etc. (Supplementary fig 9).

SVs could impact the gene expression of neighboring genes by altering the composition or location of cis-regulatory elements [9, 37]. To explore SV-associated expression changes, we leveraged transcriptomic data from 13 hybrids of Texel× Kazak sheep [38, 39], for each of which RNA-seq data for 8 tissues were available. Additionally, we obtained the WGS data for each of the individual as well as their parents. With these datasets, we were able to identify the candidate expression-associated SVs using two complementary strategies by either SV-eQTL mapping or allelic-specific expression mapping (see methods). For SV-eQTL mapping, we defined a list of SV-gene pairs based on their location and correlate the genotype of each SV with different expression levels of associated gene. 705 SVs were found to

be associated with gene expression change in at least two tissues (adjusted p value<0.05, two-sided Mann-Whitney U test) (Fig. 3B). For allelic-specific expression mapping, we expected that heterozygous SVs could act as cis-regulatory elements and cause allelic-specific expression in associated genes. In this context, 1108 SVs were found to be associated with allelic-specific expression of nearby genes (Fig. 3C, 3D). Therefore, based on the two approaches, we identified 1794 SVs which potentially affect gene expression (Supplementary data 5). Notably, 43 of the expression-associated SVs were among the population-stratified SVs, influencing expression of genes involved in immunity (APOL3-like, LOC114118416) [40-42], lipid and fructose metabolism (SEC31A, KHK) [43, 44], osteoblast differentiation (BMP2, METTL7A) [45, 46], and development (DOCK6, DIP2B) [47, 48], etc.

SVs as strong candidates affecting sheep tail morphology

The sheep tail is a key genetic trait that is related to domestication, adaptation, and animal welfare. The wild ancestors (Asiatic mouflon) have a thin and short tail whereas the domestic sheep developed four tail types including the primitive short thin-tailed type and the derived short fat-tailed (including fat-rumped), the long fat-tailed and long thin-tailed types. The underlying causative mutations are still unknown, but the long tail and the fat tail traits are apparently derived traits that arose in subsequent improvement process indicating that the causative mutations are likely to be *de novo* variants in domestic sheep.

Since the long tail is the derived trait, the causative SV is expected to be under selection in long tailed sheep breeds. To identify SVs associated with tail length, we performed comparisons within the fat-tailed type (the long fat-tailed breeds versus the short fat-tailed breeds) from Asian breeds and the thin-tailed type (the European long thin-tailed breeds versus the North European short thin-tailed breeds), respectively. Notably the most differentiated SV from the two independent comparisons both point to an insertion of 169 bp (chr11:37525005) close to the 5’ UTR of HOXB13 gene (Fig. 4A). Residing in a specific haplotype exclusively found in long tailed sheep, the insertion is prevalent in long-tailed breeds whereas rare in short-tailed breeds (Fig. 4B

and 4C). This gene is mainly expressed in the embryo and colon of sheep and cattle by searching the Ruminant genome database [49]. We luckily found the expression of HOXB13 in colons of the Texel x Scottish Blackface individuals presumably carrying at least one copy of the insertion [50], and individuals with short tail presumably absence of the insertion (Texel sheep and small-tail sheep) [51, 52]. As expected, the insertion was only expressed in the colons of the Texel x Scottish Blackface with no expression in Texel sheep and small-tail sheep, indicating that it belongs to the 5' UTR region of HOXB13 gene (Fig. 4D and Supplementary fig 10).

In order to verify the effect of this insertion on tail length, we evaluated the association between the insertion and tail length in a population of 201 sheep of the same age (nine months old). The three genotypes including homologous insertion (II), heterozygote (ID) and homologous deletion (DD) could be directly distinguished using agarose gel electrophoresis of PCR products (Fig. 4E). The tail length of the homologues (II) found to be significantly longer than that of the heterozygotes (ID) by 2.77 cm on average (Mann-Whiney U test, $p<0.0001$) (Fig. 4F). To validate the function of the insertion, we performed luciferase reporter assays. Relative to the wildtype (WT), the mutant with the insertion dramatically decreased luciferase activity by 10-fold (Fig. 4G), indicating that the insertion could reduce transcription level of HOXB13. HOXB13 has been previously identified as a repressor of caudal vertebra development [32]. Therefore, the insertion is likely to increase the tail length of sheep by depressing the expression of HOXB13.

To identify candidate SVs related with fat tail, we compared the genomes of eight fat tail breeds (SFK, KAZ, HUS, UJM, LTH, TAN, BSB, SSS) and ten thin tail breeds (FIN, DEM, EFR, GTL, SEL, SFK, SOG, POD, TIB, YBB). We applied the population branch statistic (PBS) to identify positive selection in fat tail sheep using Asiatic mouflon as outgroup. Two most differentiated region were identified by either SV or SNPs (Fig. 5A, 5B), one located at the intergenic region between BMP2 and HAO1 gene (hereafter referred to as IBH region) and the other in PDGFD gene. Notably, the most stratified SVs in these two regions were among the top selective variants of both SVs and SNPs. The most differentiated SV cluster were located at

IBH region which has also reported as a candidate selective region affecting fat tail trait [53, 54]. Notably, six highly differentiated SVs including one *de novo* SV were identified in the selective region (Fig. 5C), of which the largest one is a 7728 bp insertion. PDGFD is recognized as the most plausible candidate gene for the fat-tail phenotype in recent studies [21, 55]. We found three highly differentiated SVs residing in the selective region of PDGFD, of which the largest event is a 867 bp insertion. The three selective variants were all *de novo* SVs that were not present in Asiatic mouflon but prevalent in fat-tailed sheep. Therefore, our result suggested that both of these regions were likely to be involved in the fat tail phenotype and the SVs that we identified should be prioritized when searching for the causative mutations.

Discussion

The genome-sed analysis of genomic structural variations in diverse breeds is critical to understanding the full repertoire of genetic diversity and pinpointing their associations with phenotypic traits. However, previous resequencing studies mainly rely on mapping short reads to a linear reference genome to detect genetic variations, precluding the discovery of regions with pronounced structural variations [56]. The graph-based pan-genome model presents a new approach to assess the whole landscape of sequence diversity, and is likely to become the new guide for future genomics research. To build a high-quality ovine pan-genome, we utilized the PacBio HiFi sequencing to generate a panel of reference-level assemblies and partially phased haplotypes for 13 representative sheep breeds, providing a solid foundation for in-depth functional genomic studies. The long and accurate HiFi reads guarantee the high-standard of our pan-genome through generation of high-quality *de novo* assemblies, preservation of the two haplotypes and precise assignment of SVs' breakpoints. The development of this ovine pan-genome will be a remarkable step toward shifting in animal genomics research from a linear reference genome to a pan-genomic model comprising of multiple reference-quality representative genomes.

Application of graph-based pangenome has its clear opportunities and obstacles.

The pan-genomic approaches are not so straightforward in practice as the linear reference genome and are still not fully developed in the methodologies and implementations [11, 57]. For example, despite becoming one of the most promising graph constructors, the minigraph that we applied has limitations in performing base alignment, presenting relative high positives for small structural variants [22]. Nevertheless, a shift to the graph-based pan-genomes model will provide researchers with enhanced opportunities to systematically investigate previously inaccessible genomic variations. In our study, we found that the minigraph shows great performance in discovering divergent alleles and multiallelic variations，which account for 27.4% of total variations in length but are mostly neglected in previous studies. We further applied the read-based approach to resolve small SVs. The two complementary approaches of graph-based pan-genome and read-based calling together ensures a reference panel of high-quality SVs for downstream functional genomic studies.

The pangenome representation with an influx of high-quality assemblies will revolutionize our understanding of SV prevalence and their biological significance, giving a new dimension to the nature and extent of genomic variation, beyond SNPs, impacting traits. We demonstrated how this pangenome datasets can be utilized to link SVs to phenotypes by genotyping and assessing the SV landscape in a large population from short-read sequencing. We applied Paragraph to genotype SVs, which is a graph-based genotyper and enables efficient encoding of structural variation into the pangenome and genotyping of those variants [28]. The population-scale study allows us to trace their responses to natural and artificial selection. Taking the tail morphology evolution as an example, we identified the putative causative mutation in HOXB13 gene that causes long tail, highlighting the keys roles of SVs underlying phenotypic changes. We further reported segregating SVs in fat tail breeds and in particular *de novo* SVs that showed higher or comparative selection signals than surrounding SNPs, which warrant further investigation.

Methods

Sample collection and PacBio sequencing

Samples were selected to maximumly represent world sheep diversities, especially Eurasian breeds. Fresh blood samples were collected for each individual. High-quality genomic DNA (gDNA) was extracted as previously described [58] and assessed for purity and quantity using Nanodrop 1000 (Thermo Fisher) and Qubit (Thermo Fisher) assays. Library with an average insert size of ~15 kb was generated using the SMRTbell Express Template Prep Kit 2.0 (PacBio) and fractionated on the SageELF (Sage Science, Beverly, MA) into narrow library fractions. Libraries were then sequenced on 2-3 SMRT Cells 8M on a Sequel II instrument using Sequel II instrument (Pacbio) using 30-hour movie times. Raw data was processed using the CCS algorithm (version 6.0.0, parameters: --minPasses 3 -- minPredictedAccuracy 0.99 –maxLength 21000) to generate highly accurate HiFi reads.

De novo assembly using HiFi reads and quality assessment

Hifiasm v0.15.3-r339 was used to generate the assembly from HiFi CCS reads using default parameters [14]. Hifiasm yields one primary contig assembly and two sets of partially phased contig assemblies. The primary contig assembly for each breed was scaffolded to chromosome level using RagTag v2.0.1 with the reference sheep genome assembly [59].

The genome completeness was assessed using the BUSCO program (v.3.0.2) containing the Mammalia odb9 gene set (4104 BUSCO genes) [60]. The single plus duplicated complete BUSCO gene counts are reported. The base accuracy was measured by assembly quality value (QV) using yak (https://github.com/lh3/yak), which compares the 31-kmers found in short-reads and the assembly sequences.

Pan-genome construction

The pan-genome construction was performed using a previously reported workflow [61] with minor modifications. Minigraph v0.15-r426 with option -xggs was used to generate the graph pan-genome [22]. The sheep reference genome (ARS-

UI_Ramb_v2.0, GCF_016772045.1) was used as the backbone of the multiassembly graph and the 26 partially phased contig assemblies were added successively.

Identification of non-reference divergent alleles and multiallelic alleles from the graph pan-genome

The resulting graph by minigraph is composed of chains of bubbles with the reference as the backbone. We first used gfatools v0.5-r234 which is a bubble popping algorithm to extract bubbles from the graph (https://github.com/lh3/gfatools). Each bubble represents a structural variation, encompassing the start and end nodes of reference sequences as well as path traversing the start and end nodes.

Depending on the number of paths in a bubble, the structural variations were classified into three categories: (1) biallelic variations (Insertions/Deletions): either the reference path is longer (deletions) or non-reference path is longer (insertions). The node in the shorter path is required to be below 100 bp. (2) Divergent alleles: biallelic variations where the reference nodes and the non-reference nodes were both longer than 100 bp. (3) Multiallelic alleles: multiallelic variations where one reference path and more than one nonreference path were present. Those nonreference paths below 100 bp were excluded.

Genome annotation of non-reference sequences

The nodes in a bubble that are not derived from the reference path were referred to as non-reference alleles. The genes within non-reference sequences were annotated using a homology-based approach. Protein sequences from different species including cattle (Bos taurus), goat (Capra hircus), sheep (Ovis aries) and pig (Sus scrofa) were downloaded from NCBI and processed with Exonerate v2.2.0 [62] to find the best hit per protein sequence. EvidenceModeler v1.1.1 [63] was used to integrate all evidence into a non-redundant gene annotation.

Annotation of repeats

Interspersed repeats and low complexity DNA sequences were identified using RepeatMasker v4.0.7 (http://www.repeatmasker.org) with a combined repeat database including Dfam v.20170127 and RepBase v20170127 with parameters: -species Ruminantia -xsmall -s -no_is -cutoff 255 -frag 20000 -gff. Tandem repeats were

identified using Tandem repeat finder [64]. The identified tandem repeats as well as those low complexity sequences identified by RepeatMakser were merged as low-complexity regions (LCR). Of the LCR, regions of tandem repeat with the unit motif length ≥7 bp was classified as VNTR and those with the unit motif length ≤6 bp were classified as STR. The remaining LCR were classified as Other-LCR.

SV calling

We integrated the results from pan-genome analysis and alignment-based SV calling to generated a high-confidence SV calls. For pan-genome analysis, the SVs were generated from our multiassebly graph genome using gfatools v0.5-r234 as we described in previous section. For alignment-based SV calling, the SVs were detected using pbsv tools v2.6.2 (https://github.com/PacificBiosciences/pbsv). The CCS reads were aligned to the reference genome using pbmm2 to generate the alignment BAM file. Then the pbsv discover module was used to identify signatures of structural variation within each aligned BAM. The tandem repeat annotation file was provided to increase sensitivity and recall of this step. Then the signatures of all the samples were fed to pbsv call module to jointly call SV from structural variant signatures using the –ccs mode.

Mendelian error rate estimation

A Mendelian inheritance error represents a genotype in the offsprint that cannot be inherited from either of their paraments. The whole genome sequencing data of 11 trios were from our recently study for allelic specific expression analysis [38]. For each trio, the Mendelian error rate is the proportion of SVs from the children that deviate from Mendelian inherence.

SV hotspot analysis

SV hotspot regions were determined using primatR package “hotspotter” section (bw=200000, num.trial=1000, pval<1e-8) [65]. SVs were significantly clustered at the terminal 5 Mb region of each chromosome (pvalue < 0.01, permutation test, random shuffling of 1000 times). The random shuffling of SV intervals were performed using Bedtools v2.29.0 with the shuffle command [66].

Enrichment of SVs in the genome

The H3K4me3 and H3K27ac peak regions were retrieved from our Ruminant genome database. We only used insertions and deletions for enrichment. For insertions, the end position is the start position plus the insertion length. We first recorded the observed number of overlaps between SVs and UTR, exon, intron, H3K4me3 and H3K27ac peak regions, respectively. Then we counted the average number of overlaps between randomly shuffled SVs (1000 times) and each genomic region.

SV genotyping in resequenced samples

We collected whole genome sequencing data for 645 domestic sheep, from four geographic regions (Europe, East Asia, Middle east and Africa) as well as wild sheep including 33 Asiatic mouflons (*Ovis aries musimon*) and 8 argali sheep (*Ovis ammon*). For the domestic sheep, those breeds with at least five individuals remain unchanged whereas those samples from one country were grouped into one population if each of the breed is composed of less than 5 individuals. In this way, all the domestic sheep were assigned to 33 breeds/populations (Supplementary data 6). We genotyped SVs in these samples using Paragraph v2.4a [28]. The maximum permitted read count for each variant were set to 20 times the average sample depth in order to reduce runtime for repetitive regions. The resulting genotype files in vcf format from all samples were then combined using bcftools v1.9 [67].

To obtain a high-confidence call set of SVs, the resulting genotypes were filtered based on overall genotyping rates and Hardy-Weinberg equilibrium [29]. Those variants that were successfully genotyped in ≥10% of all samples were removed using vcftools [68]. We also estimated Hardy-Weinberg equilibrium for each variant using vcftools and those shows excess of heterozygotes (pvalue< 1×10-5) were removed.

Quantifying linkage disequilibrium (LD) of SVs with surrounding SNPs

In order to measure the extent to which SVs are correlated with SNPs, we calculated LD (r^2 values) between the polymorphic SVs (MAF > 0.05) and the neighboring polymorphic SNPs (MAF > 0.05; 150 upstream and 150 downstream SNPs) in Mouflon and domestic sheep. SNPs located on SVs were not considered. Plink v1.9 was used to calculate LD for all pairwise cases for SV-SNP and the surrounding SNP-

SNP variant pairs. We then calculated degree of LD between each SV and flanking SNPs following a previous study [37]. We first obtained pairwise genotype correlations (r^2) for all SNP-SNP and SNP-SV pairs. For each of the 300 ranked surrounding positions, the number of times the SV rank was greater than the SNP-SNP median rank was recorded as a relative LD metric of SV to SNP. Low, middle or high LD of each SV were then classified if the number of exceed was less than 100, between 100 and 200, and greater than 200, respectively.

Inferring the derived SVs and *de novo* SVs in domestic sheep

To infer the state (ancestral or derived) for each SV, we additionally genotyped one snow sheep (12.11×, ERR4161992) and one bighorn sheep (10.3×, SRR501898). In this way, we used three wild sheep samples from one argali, one snow sheep, one bighorn sheep as outgroups. For each biallelic SV, if the same homozygous genotypes were detected in at least two outgroups, this state is defined as the ancestral state while the other is the derived state. If the derived allele is only present in domestic sheep but absent in all the wild sheep (8 Argali and 33 Asiatic mouflons), the SV was defined as a *de novo* SV in domestic sheep.

Detecting selection signals

In order to reveal selection signals in fat tail sheep, we used the method of population branch statistic (PBS) to calculate and compare the Fixation index (Fst) between the fat tail sheep, thin tail sheep and Asiatic mouflon, using the Asiatic mouflon as a distantly related population. The Fst value between populations with different traits was calculated using the python scikit-allel package v1.3.2 (https://github.com/cggh/scikit-allel). We calculated the Fst values of both SNP and SV sets. Fst values for SNPs were calculated with a 10 kb sliding window with a 5 kb step.

To scan positive selected loci between different breeds, we employed two strategies, including the global Fst and the Ohana method which identify extreme allele frequency differences by modelling ancestral admixture components. The global Fst was calculated following previously described approach with modifications [69]. For each SV site, we calculated the Fst values between all populations. Instead

of using the average of all Fst values, we took the average of the top 50% Fst values as the global FST value for each SV. The top 1% of the highest global Fst value by this method were considered as putative selective signals. For Ohana method, we first modeled the dataset with the provided workflow using the SNPs [29]. The resulting covariance was used as the neutral input for scans of selection. For each specific ancestry component, Ohana reported a likelihood ratio statistic (LRS) to quantify the likelihood of selection for each variant. We varied the number of ancestral admixture components from K=2 to K=5 to find selection signals.

Detecting expression-associated SVs

The RNA-seq data were downloaded from NCBI SRA database (accession number PRJNA485657) from 14 Texel (♂)×Kazakh (♀) sheep hybrids from 8 tissues including liver, rumen, tail fat, skin, subcutaneous fat, semitendinosus, semimembranosus, longissimus dorsi [38, 39]. The raw reads were first trimmed for adapter and low-quality sequences using Trimmomatic [70] and the high-quality reads were aligned to sheep reference genome using Hisat2 [71]. Reads counts per gene were computed using featureCounts [72] and normalized to counts per million (CPM) using edgeR package [73] and those with CPM > 1 in at least two samples were retained.

To identify expression-associated SVs, we first performed SV-eQTL mapping by measuring the association between SV genotypes and expression level of nearby genes. We defined a list of SV-gene pairs based on the distance between SV and nearby genes within 200 kb. Only the SVs with presence and absence in at least 3 samples, respectively, were considered. For each of the SV-gene pairs, the expression difference of each gene was compared between those with and without SV using a two-sided Mann-Whitney U test. A P-value threshold of < 0.05 was applied to identify candidate SV-eQTL. The candidate expression-associated SV was required to be repeatedly found in at least two tissues.

It is expected that the regulatory SVs could cause allelic specific expression. We also performed allelic specific expression (ASE) mapping by integrating the RNA-seq data as well as the SV genotyping information for each of the samples. The ASE

genes were identified following the pipeline described in our previous study [38]. First, for each sample, the ASE SNPs displaying allelic imbalance at the gene body were identified from RNA-seq data using the following parameters: (1) at least 20 total reads; (2) allele ratio (allele reads count / total reads count)> 0.65 or <0.35; (3) significant allelic imbalances with FDR<0.05. Those with at least two ASE SNPs at the gene body and were found in at least three samples were defined as candidate ASE gene. For each SV-gene pair, if the ASE gene matches with the heterozygous state for the associated SV in at least five RNA-seq samples, we considered the SV as putatively expression-associated.

Analyzing the association between the 169-bp insertion with tail length

The association analysis was conducted in a F2 population which was generated by backcrossing East Friesian sheep (♂) with the female Hu sheep (♀) x East Friesian sheep (♂). We selected 211 individuals at the same age (six months old) to measure the tail length and collect DNA samples for each individual. Since the insertion was dominant in East Friesian sheep with allele frequency of 0.92 but is completely absent in Hu sheep based on our resequencing data, the F2 population that we studied only possessed two genotypes (II and ID) due to the grading-up using East Friesian sheep. Agarose electrophoresis was used to determine genotypes of the PCR products. Primers were designed in the flanking region of the insertion (forward: TTTATGAGCTTCTCTCCGCCA; reverse: AAGTGGTATAATTGCCGGGCT). PCR amplification was performed with 94˚C for 2 min; 94˚C for 30 s, 62˚C for 30 s, and 72˚C for 30 s for 28 cycles; 72˚C for 2 min; and a hold at 4˚C.

Luciferase reporter assay

To analyze promoter activity of the SV variation by luciferase assay, fragments with (mutation) and without (wild type) the insertion was synthetized by Tsingke Biotechnology Co., Ltd. (Beijing, China). Each fragment was cloned into the pGL3-Basic vector using Sac I and Hind□ sites to assay its promoter activity on the downstream gene.

293T cells in 96-well plates were transiently transfected with 200 ng Firefly luciferase vector (pGL3) and 4 ng pRL-TK Renilla luciferase vector (Promega,

Madison, USA) using 0.5 uL Lipofectamine 3000 (Thermo Fisher Scientific, Waltham, USA). Twenty-four hours after transfection, luciferase activities were measured with a dual-luciferase reporter assay kit (Promega, Madison, USA). Renilla luciferase activity was used as the transfection control for normalizing Firefly luciferase activity. Empty pGL3-Basic vector was used as the reference for promoter activity assays. Empty pGL3-Control vector (Control) was used as the positive control for promoter activity. The data are presented as means with standard deviation.

Authors' contributions

Y.J. and S.Q.G. conceived and designed the experiments. R.L., X.M.Z. and Z.Y.L. performed *de novo* assembly and SV calling; Z.Y.L., M.S.X. and Y.F.Z. analyzed the RNA-seq and Chip-seq data; F.W., M.G. and L.Z. performed the ASE analysis; R.L., Z.Y.L., W.W.Fang and X.L.D. carried out SV genotyping and population genetics analysis; Y.T.Y. and X.Y.L. conducted the luciferase reporter assay; Z.B.Z., W.W.Fu, H.H.Z., C.N.C., P.Y., Z.A.G., N.J.N., H.A.N., X.P.Y., Y.X.S. and W.D.D. collected the samples and prepared the WGS data; R.L. wrote the manuscript and Y.J., S.Q.G., J.A.L., P.H., E.I.A., R.D.X. and X.H.W. revised it. All authors read and approved the final manuscript.

Acknowledgements

We thank the High-Performance Computing platform of Northwest A&F University for providing computing resources. We would also like to thank Prof. Xiaolong Wang for providing the nice chart of different sheep tail patterns.

Competing interests

The authors declare no competing interests.

Ethics statement

Blood samples were taken by conforming with the Helsinki Declaration of 1975 (as revised in 2008) concerning Animal Rights.

Data availability

Sequences and metadata generated in this work are publicly available. The *de novo* assemblies, HiFi data and WGS data for the 13 individuals were available at NCBI: WDU1 (PRJNA788705, PRJNA788704, PRJNA784371),QK1 (PRJNA788680, PRJNA788679, PRJNA784330), ROM1 (PRJNA788688, PRJNA788687, PRJNA784357), SFK1 (PRJNA788697, PRJNA788696, PRJNA784366), UJM1 (PRJNA788702, PRJNA788701, PRJNA784376), CHA1 (PRJNA788666, PRJNA788665, PRJNA784396), YUN1 (PRJNA788712, PRJNA788711, PRJNA784413), TEX1 (PRJNA788214, PRJNA788213, PRJNA780104), DOR1 (PRJNA788571, PRJNA788570, PRJNA782680), KER1 (PRJNA788676, PRJNA788675, PRJNA783516), MER1 (PRJNA788672, PRJNA788671, PRJNA783537), KAZ1 (PRJNA788564, PRJNA788563, PRJNA783179), FRI1 (PRJNA791176, PRJNA791175, PRJNA721520). The details are provided in Supplementary table 1.

Funding

This study was supported by research grants from the National Natural Science Foundation of China to Y.J. (U21A20247, 31822052), S.Q.G. (31760662, 31860639) and R.L. (31802027), China Postdoctoral Science Foundation to W.W.Fu (No. 2021M702690).

Tables

Table 1 Quality assessments of the primary and partially phased assemblies for the 13 individuals.

Breeds	Abbr.	Assembly	Assembly length (Gb)	Contig N50 (Mb)	Contig number	Depth	BUSCO (%)	Base accuracy
Kazak	KAZ1	Primary	2.87	73.4	1849	22.4	93.4	46.3
		hap1/hap2	2.85/2.84	10.7/13.9		-	92.3/91.9	46.6/46.4

Kermani	KER1	Primary	2.82	80.3	1675	21.7	93.8	45.2
		hap1/hap2	2.77/2.75	15.4/15.2		-	90.7/92.2	45.4/45.6
Dorset	DOR1	Primary	2.91	92.4	1293	25.0	94.3	45.5
		hap1/hap2	2.86/2.83	32.8/39.5		-	93.4/92.7	45.6/45.6
East	FRI1	Primary	2.90	85.1	982	28.0	94.4	48.2
Friesian		hap1/hap2	2.77/2.75	62.9/53.2		-	91.3/92.8	48.4/48.5
Chiniese	MER1	Primary	2.84	60.0	1773	18.7	94.2	43.3
Merino		hap1/hap2	2.82/2.83	3.7/4.8		-	92.1/90.0	42.4/43.3
Qiaoke	QK1	Primary	2.87	75.0	1654	21.5	94.0	47.1
		hap1/hap2	2.83/2.84	13.6/9.9		-	92.6/93.4	47.6/47.2
Romney	ROM1	Primary	2.86	68.3	1552	19.3	94.4	43.2
		hap1/hap2	2.81/2.82	6.3/5.7		-	92.1/92.4	43.3/43.1
Suffolk	SFK1	Primary	2.88	65.3	1513	18.8	94.3	43.4
		hap1/hap2	2.86/2.80	5.7/4.9		-	92.4/92.5	43.2/43.3
Texel	TEX1	Primary	2.82	47.6	1835	17.9	94.4	44.6
		hap1/hap2	2.76/2.75	4.52/4.49		-	91.4/90.2	44.8/44.8
White	WDU1	Primary	2.81	17.8	2125	15.4	93.8	41.7
Dorper		hap1/hap2	2.82/2.77	1.58/1.26		-	90.2/89.8	41.9/41.8
Ujumqin	UJM1	Primary	2.87	75.6	1534	21.3	94.4	44.4
		hap1/hap2	2.86/2.84	11.5/11.7		-	93.9/93.5	44.5/44.6
Charollais	CHA1	Primary	2.86	71.5	1427	19.2	94.1	44.1
		hap1/hap2	2.82/2.80	6.4/5.1		-	91.7/92.1	44.3/44.3
Yunnan	YUN1	Primary	2.91	73.9	1293	22.0	94.1	46.9
		hap1/hap2	2.89/2.85	18.7/13.1		-	93.6/93.4	46.7/46.6

Figure legends

Fig. 1 Ovine graph pan-genome construction. (A) PCA plot shows the representation of the genetic diversity of domestic sheep by the 13 individuals used for HiFi sequencing and pan-genome construction. (B) The pan-genome size changes with the increase in number of genome assemblies. (C) Repeat annotation of the structural variations stratified by SV types. (D) The ovine graph pan-genome reveals diverse SV types. (E) An example of divergent allele of 50757 bp containing a MYADM-like repeat. (F) An example of multiallelic variations containing a large deletion of

105,650 bp which is nested with another insertion of 5,518 bp.

Fig. 2 The repertoire of structural variations in sheep genome and frequency spectrum in domestic and wild sheep. (A) Length distribution of insertions and deletions. (B) Geographical distribution of the 645 domestic sheep and 33 Asiatic mouflons. Due to space limitation, dorper sheep from South Africa is not shown. (C) Derived allele frequency of the insertions and deletions. (D) Proportion of standing and de novo SVs. (E) High-frequency de novo SVs in domestic sheep. (F) A de novo insertion within GRID2 gene rises to high frequency in all geographical populations. (G) A de novo insertion of a retrogene rises to high frequency in European sheep.

Fig. 3 Population-stratified SVs and their effects on gene expression. (A) Genome-wide distribution of global Fst which is measured by the average value of top 50% Fst values for each SV across 33 sheep breeds. (B) Examples of expression-associated SVs identified by SV-eQTL mapping. (C) Examples of expression-associated SVs identified by allelic specific expression (ASE) mapping. Each column represents one ASE SNP in exons and each row represent one RNA-seq sample. (D) Frequency distribution of the six expression-associated SVs in 33 breeds/populations. CNM, Chinese Merino; DEM, German merino; EFR, East Friesian Dairy; ESO, Spain; FIN, Finnsheep; GTL, Gotland; NOW, Norway; OUE, Ouessant; POD, Poll Dorset; RUS, Russia; SEL, Shetland; SFK, Suffolk; SOG, Solognote; WSF, Suffolk White; ALT, Altay; BSB, Bashibai; BYK, Bayinbuluke; CLB, Cele Black; DUL, Duolang; HUS, Hu; KAZ, Kazakh; LTH, Large-tailed Han; SSS, Sishui Fur; STH, Small-tailed Han; TAN, Tan; UJM, Ujumqin; WAG, Waggir; WDS, Wadi; OLA, Oula; TIB, Tibetan; YBB, Yunnan; AFS, Afshari; MDE, Middle East; CND, Dorper; MAOA, Morocco; MOF, Mouflon.

Fig. 4 Selective SVs associated with long tail trait. (A) The genome-wide Fixation index (Fst) of long fat-tailed sheep versus short fat-tailed sheep from Asia breeds and long thin-tailed sheep versus short thin-tailed sheep from Europe breeds. (B) The selective haplotype in long-tailed sheep. The 169bp-insertion is close to the 5' UTR region of HOXB13 gene. The haplotypes include the genotypes of the insertion and surrounding SNPs. (C) The frequency of the insertion in wild sheep and domestic

sheep. LTH, Large-tailed Han; TAN, Tan; SSS, Sishui Fur; EFR, East Friesian Dairy; SOG, Solognote; DEM, German merino; SFK, Suffolk; HUS, Hu; STH, Small-tailed Han; UJM, Ujumqin; SEL, Shetland; FIN, Finnsheep; OUE, Ouessant; GTL, Gotland. (D) RNA-seq data shows that the insertion belongs to the 5' UTR region. The top panel corresponds to the expression in colon of Texel x Scottish Blackface sheep (ERS2107126) presumably carrying at least one copy of the insertion; the bottom panel corresponds to the expression in colon of one small-tail sheep (SRR5122597) presumably absence of the insertion. (E) The electrophoresis diagrams of PCR products for the three genotypes (homologous insertion: II, heterozygote: ID, homologous deletion: DD). (F)The carriers and non-carriers of the insertion differs in tail length. (G) HOXB13 promoter activity modulated by the 169 bp insertion. Luciferase activity in HEK293T cells transfected with recombinant plasmids containing the insertion or not. Each experiment was repeated at least five times. The data were the mean ± standard error of the normalized luciferase activity. **P-value<0.01.

Fig. 5 Selective SVs associated with fat tail trait. (A) Population branch statistic (PBS) values across the whole genome by comparing fat tailed sheep to thin tailed sheep using the mouflon sheep as outgroup. (B) The two most differentiated regions between fat tailed sheep and thin tailed sheep. The left panel corresponds to the intergenic region between BMP2 and HAO1 gene while the right panel corresponds to the region surrounding PDGFD gene. Those SVs with PBS value >0.8 were highlighted. (C) The haplotypes of mouflons and domestic sheep for the two most selective regions. Each column represents one SV and each row represent one individual. The black reverted triangles represent *de novo* SVs.

Supplementary files

Supplementary table 1 Accession numbers of the *de novo* assemblies.

Supplementary data

Supplementary data 1 Information of divergent variations

Supplementary data 2 Information of multiallelic variations

Supplementary data 3. List of the structural variations including insertion, deletion, inversion and duplications

Supplementary data 4 SV frequency in mouflon sheep and domestic sheep

Supplementary data 5 List of putatively expression associated SVs

Supplementary data 6 Sample list of the resequencing data

Supplementary figures

Supplementary fig. 1 Example of three representative divergent variations intersecting with IDE (a), PRP3-like (b) and MAGI2 (c), respectively.

Supplementary fig. 2 The multiallelic region harboring beta hemoglobin genes.

Supplementary fig. 3 Example of three representative multiallelic variations affecting ASIP (a), UDP-glucuronosyltransferase 2B18-like (b) and olfactory receptor (LOC101115700) (c), respectively.

Supplementary fig. 4 The number of SVs intersecting functional elements compared to randomly permuting SV locations.

Supplementary fig. 5 An ideogram showing SV hotspots using all SVs. The total number of SVs in each detected hotspot is shown by a scale going from blue to red.

Supplementary fig. 6 Enrichment analysis of SVs with respect to the last 5 Mb of each chromosome end.

Supplementary fig. 7 PCA plot showing the clusters of domestic sheep by insertions (a) and deletions (b).

Supplementary fig. 8 Histogram of the number of SV ranks (0-300) of r2 values that surpass the median SNP-based r2 value for common SVs (MAF > 0.05).

Supplementary fig. 9 Enriched KEGG pathways for the population-stratified SVs.

Supplementary fig. 10 RNA-seq mapping for the 169 bp insertion shows that the insertion belongs to 5’ UTR region.

References:

[1] Chiang C, Scott A J, Davis J R, Tsang E K, Li X, Kim Y, et al. The impact of structural variation on human gene expression. Nat Genet. 2017;49:692-9.

[2] Sudmant P H, Rausch T, Gardner E J, Handsaker R E, Abyzov A, Huddleston J, et al. An integrated map of structural variation in 2,504 human genomes. Nature. 2015;526:75-81.

[3] Mahmoud M, Gobet N, Cruz-Dávalos D I, Mounier N, Dessimoz C, Sedlazeck F J. Structural variant calling: the long and the short of it. Genome Biol. 2019;20:246.

[4] Sedlazeck F J, Lee H, Darby C A, Schatz M C. Piercing the dark matter: bioinformatics of long-range sequencing and mapping. Nat Rev Genet. 2018;19:329-46.

[5] Audano P A, Sulovari A, Graves-Lindsay T A, Cantsilieris S, Sorensen M, Welch A E, et al. Characterizing the Major Structural Variant Alleles of the Human Genome. Cell. 2019;176:663-75.e19.

[6] Ebert P, Audano P A, Zhu Q, Rodriguez-Martin B, Porubsky D, Bonder M J, et al. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. Science. 2021:eabf7117.

[7] Qin P, Lu H, Du H, Wang H, Chen W, Chen Z, et al. Pan-genome analysis of 33 genetically diverse rice accessions reveals hidden genomic variations. Cell. 2021.

[8] Liu Y, Du H, Li P, Shen Y, Peng H, Liu S, et al. Pan-Genome of Wild and Cultivated Soybeans. Cell. 2020;182:162-76.e13.

[9] Alonge M, Wang X, Benoit M, Soyk S, Pereira L, Zhang L, et al. Major Impacts of Widespread Structural Variation on Gene Expression and Crop Improvement in Tomato. Cell. 2020;182:145-61.e23.

[10] Tuzun E, Sharp A J, Bailey J A, Kaul R, Morrison V A, Pertz L M, et al. Fine-scale structural variation of the human genome. Nat Genet. 2005;37:727-32.

[11] Golicz A A, Bayer P E, Bhalla P L, Batley J, Edwards D. Pangenomics Comes of Age: From Bacteria to Plant and Animal Applications. Trends Genet. 2020;36:132-45.

[12] De Coster W, Van Broeckhoven C. Newest methods for detecting structural variations. Trends Biotechnol. 2019;37:973-82.

[13]Wenger A M, Peluso P, Rowell W J, Chang P-C, Hall R J, Concepcion G T, et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nat Biotechnol. 2019;37:1155-62.

[14]Cheng H, Concepcion G T, Feng X, Zhang H, Li H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nat Methods. 2021;18:170-5.

[15]Nurk S, Walenz B P, Rhie A, Vollger M R, Logsdon G A, Grothe R, et al. HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. Genome Res. 2020;30:1291-305.

[16]Talenti A, Powell J, Hemmink J D, Cook E A J, Wragg D, Jayaraman S, et al. A cattle graph genome incorporating global breed diversity. bioRxiv. 2021:2021.06.23.449389.

[17]Leonard A S, Crysnanto D, Fang Z-h, Heaton M P, Vander Ley B L, Herrera C, et al. Bovine pangenome reveals trait-associated structural variation from diverse assembly inputs. bioRxiv. 2021:2021.11.02.466900.

[18]Norris B J, Whan V A. A gene duplication affecting expression of the ovine ASIP gene is responsible for white and black sheep. Genome Res. 2008;18:1282-93.

[19]Demars J, Cano M, Drouilhet L, Plisson-Petit F, Bardou P, Fabre S, et al. Genome-Wide Identification of the Mutation Underlying Fleece Variation and Discriminating Ancestral Hairy Species from Modern Woolly Sheep. Mol Biol Evol. 2017;34:1722-9.

[20]Wiedemar N, Drögemüller C. A 1.8-kb insertion in the 3′-UTR of RXFP2 is associated with polledness in sheep. Anim Genet. 2015;46:457-61.

[21]Li X, Yang J, Shen M, Xie X-L, Liu G-J, Xu Y-X, et al. Whole-genome resequencing of wild and domestic sheep identifies genes associated with morphological and agronomic traits. Nat Commun. 2020;11:2815.

[22]Li H, Feng X, Chu C. The design and construction of reference pangenome graphs with minigraph. Genome Biol. 2020;21:265.

[23]Gonzalez M V, Mousel M R, Herndon D R, Jiang Y, Dalrymple B P, Reynolds J O, et al. A Divergent Artiodactyl MYADM-like Repeat Is Associated with Erythrocyte Traits and Weight of Lamb Weaned in Domestic Sheep. PLOS ONE. 2013;8:e74700.

[24]French A T, Bethune J A, Knight P A, McNeilly T N, Wattegedera S, Rhind S, et al. The expression of intelectin in sheep goblet cells and upregulation by interleukin-4. Vet Immunol Immunopathol. 2007;120:41-6.

[25]Jiang Y, Wang X, Kijas J W, Dalrymple B P. Beta-globin gene evolution in the ruminants: evidence for an ancient origin of sheep haplotype B. Anim Genet. 2015;46:506-14.

[26]Durbin R M, Altshuler D, Durbin R M, Abecasis G R, Bentley D R, Chakravarti A, et al. A map of human genome variation from population-scale sequencing. Nature. 2010;467:1061-73.

[27]Quan C, Li Y, Liu X, Wang Y, Ping J, Lu Y, et al. Characterization of structural variation in Tibetans reveals new evidence of high-altitude adaptation and introgression. Genome Biol. 2021;22:159.

[28]Chen S, Krusche P, Dolzhenko E, Sherman R M, Petrovski R, Schlesinger F, et al. Paragraph: a graph-based structural variant genotyper for short-read sequence data. Genome Biol. 2019;20:291.

[29]Yan S M, Sherman R M, Taylor D J, Nair D R, Bortvin A N, Schatz M C, et al. Local adaptation and archaic introgression shape global diversity at human structural variant loci. eLife. 2021;10:e67615.

[30]Kalkan Z, Durasi İ M, Sezerman U, Atasever-Arslan B. Potential of GRID2 receptor gene for preventing TNF-induced neurodegeneration in autism. Neurosci Lett. 2016;620:62-9.

[31]Zhu X-J, Liu Y, Dai Z-M, Zhang X, Yang X, Li Y, et al. BMP-FGF signaling axis mediates Wnt-induced epidermal stratification in developing mammalian skin. PLoS Genet. 2014;10:e1004687.

[32]Economides K D, Zeltser L, Capecchi M R. HOXB13 mutations cause overgrowth of caudal spinal cordand tail vertebrae. Dev Biol. 2003;256:317-30.

[33]Kardos M, Luikart G, Bunch R, Dewey S, Edwards W, McWilliam S, et al. Whole□genome resequencing uncovers molecular signatures of natural and sexual selection in wild bighorn sheep. Mol Ecol. 2015;24:5616-32.

[34]Chen C, Liu C, Xiong X, Fang S, Yang H, Zhang Z, et al. Copy number variation

in the MSRB3 gene enlarges porcine ear size through a mechanism involving miR-584-5p. Genet Sel Evol. 2018;50:1-18.

[35]Pan Z, Li S, Liu Q, Wang Z, Zhou Z, Di R, et al. Rapid evolution of a retro-transposable hotspot of ovine genome underlies the alteration of BMP2 expression and development of fat tails. BMC Genomics. 2019;20:1-15.

[36]Nishikawa S, Kusakabe M, Yoshinaga K, Ogawa M, Hayashi S-l, Kunisada T, et al. In utero manipulation of coat color formation by a monoclonal anti□c□kit antibody: two distinct waves of c□kit□dependency during melanocyte development. The EMBO journal. 1991;10:2111-8.

[37]Yang N, Liu J, Gao Q, Gui S, Chen L, Yang L, et al. Genome assembly of a tropical maize inbred line provides insights into structural variation and crop improvement. Nat Genet. 2019;51:1052-9.

[38]Wang F, Shao J, He S, Guo Y, Pan X, Wang Y, et al. Allele-specific expression and splicing provides insight into the phenotypic differences between thin- and fat-tailed sheep breeds. Journal of Genetics and Genomics. 2021.

[39]Pan X, Cai Y, Li Z, Chen X, Heller R, Wang N, et al. Modes of genetic adaptations underlying functional innovations in the rumen. Sci China Life Sci. 2021;64:1-21.

[40]Nanjappa S G, Mudalagiriyappa S, Fites J S, Suresh M, Klein B S. CBLB Constrains Inactivated Vaccine–Induced CD8+ T Cell Responses and Immunity against Lethal Fungal Pneumonia. The Journal of Immunology. 2018;201:1717-26.

[41]Gaudet R G, Zhu S, Halder A, Kim B-H, Bradfield C J, Huang S, et al. A human apolipoprotein L with detergent-like activity kills intracellular pathogens. Science. 2021;373:eabf8113.

[42]Lascano J, Uchil P D, Mothes W, Luban J. TRIM5 retroviral restriction activity correlates with the ability to induce innate immune signaling. J Virol. 2015;90:308-16.

[43]Iwaki A, Moriwaki K, Sobajima T, Taniguchi M, Yoshimura S i, Kunii M, et al. Loss of Rab6a in the small intestine causes lipid accumulation and epithelial cell death from lactation. The FASEB Journal. 2020;34:9450-65.

[44]Han J, Li E, Chen L, Zhang Y, Wei F, Liu J, et al. The CREB coactivator CRTC2

controls hepatic lipid metabolism by regulating SREBP1. Nature. 2015;524:243-6.
[45]Matsubara T, Kida K, Yamaguchi A, Hata K, Ichida F, Meguro H, et al. BMP2 Regulates Osterix through Msx2 and Runx2 during Osteoblast Differentiation*. J Biol Chem. 2008;283:29119-25.
[46]Lee E, Kim J Y, Kim T K, Park S Y, Im G I. Methyltransferase-like protein 7A (METTL7A) promotes cell survival and osteogenic differentiation under metabolic stress. Cell Death Discov. 2021;7:154.
[47]Sukalo M, Tilsen F, Kayserili H, Müller D, Tüysüz B, Ruddy D M, et al. DOCK6 mutations are responsible for a distinct autosomal-recessive variant of Adams-Oliver syndrome associated with brain and eye anomalies. Hum Mutat. 2015;36:593-8.
[48]Winnepenninckx B, Debacker K, Ramsay J, Smeets D, Smits A, FitzPatrick D R, et al. CGG-repeat expansion in the DIP2B gene is associated with the fragile site FRA12A on chromosome 12q13.1. Am J Hum Genet. 2007;80:221-31.
[49]Fu W, Wang R, Nanaei H A, Wang J, Hu D, Jiang Y. RGD v2.0: a major update of the ruminant functional and evolutionary genomics database. Nucleic Acids Res. 2021.
[50]Clark E L, Bush S J, McCulloch M E B, Farquhar I L, Young R, Lefevre L, et al. A high resolution atlas of gene expression in the domestic sheep (Ovis aries). PLoS Genet. 2017;13:e1006997.
[51]Chao T, Wang G, Ji Z, Liu Z, Hou L, Wang J, et al. Transcriptome Analysis of Three Sheep Intestinal Regions reveals Key Pathways and Hub Regulatory Genes of Large Intestinal Lipid Metabolism. Scientific Reports. 2017;7:5345.
[52]Jiang Y, Xie M, Chen W, Talbot R, Maddox J F, Faraut T, et al. The sheep genome illuminates biology of the rumen and lipid metabolism. Science. 2014;344:1168-73.
[53]Pan Z, Li S, Liu Q, Wang Z, Zhou Z, Di R, et al. Rapid evolution of a retro-transposable hotspot of ovine genome underlies the alteration of BMP2 expression and development of fat tails. BMC Genomics. 2019;20:261.
[54]Baazaoui I, Bedhiaf-Romdhani S, Mastrangelo S, Ciani E. Genome-wide analyses reveal population structure and identify candidate genes associated with tail fatness in local sheep from a semi-arid area. Animal. 2021;15:100193.
[55]Dong K, Yang M, Han J, Ma Q, Han J, Song Z, et al. Genomic analysis of

worldwide sheep breeds reveals PDGFD as a major target of fat-tail selection in sheep. BMC Genomics. 2020;21:1-12.

[56]Ho S S, Urban A E, Mills R E. Structural variation in the sequencing era. Nat Rev Genet. 2020;21:171-89.

[57]Miga K H, Wang T. The Need for a Human Pangenome Reference Sequence. Annual Review of Genomics and Human Genetics. 2021;22:81-102.

[58]Li R, Yang P, Dai X, Asadollahpour Nanaei H, Fang W, Yang Z, et al. A near complete genome for goat genetic and genomic research. Genet Sel Evol. 2021;53:74.

[59]Alonge M, Soyk S, Ramakrishnan S, Wang X, Goodwin S, Sedlazeck F J, et al. RaGOO: fast and accurate reference-guided scaffolding of draft genomes. Genome Biol. 2019;20:224.

[60]Manni M, Berkeley M R, Seppey M, Simão F A, Zdobnov E M. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. Mol Biol Evol. 2021;38:4647-54.

[61]Crysnanto D, Leonard A S, Fang Z-H, Pausch H. Novel functional sequences uncovered through a bovine multiassembly graph. Proc Natl Acad Sci. 2021;118:e2101056118.

[62]Keilwagen J, Wenk M, Erickson J L, Schattat M H, Grau J, Hartung F. Using intron position conservation for homology-based gene prediction. Nucleic Acids Res. 2016;44:e89.

[63]Haas B J, Salzberg S L, Zhu W, Pertea M, Allen J E, Orvis J, et al. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biol. 2008;9:R7.

[64]Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 1999;27:573-80.

[65]Bakker B, Taudt A, Belderbos M E, Porubsky D, Spierings D C J, de Jong T V, et al. Single-cell sequencing reveals karyotype heterogeneity in murine and human malignancies. Genome Biol. 2016;17:115.

[66]Quinlan A R, Hall I M. BEDTools: a flexible suite of utilities for comparing

genomic features. Bioinformatics. 2010;26:841-2.

[67]Narasimhan V, Danecek P, Scally A, Xue Y, Tyler-Smith C, Durbin R. BCFtools/RoH: a hidden Markov model approach for detecting autozygosity from next-generation sequencing data. Bioinformatics. 2016;32:1749-51.

[68]Danecek P, Auton A, Abecasis G, Albers C A, Banks E, DePristo M A, et al. The variant call format and VCFtools. Bioinformatics (Oxford, England). 2011;27:2156-8.

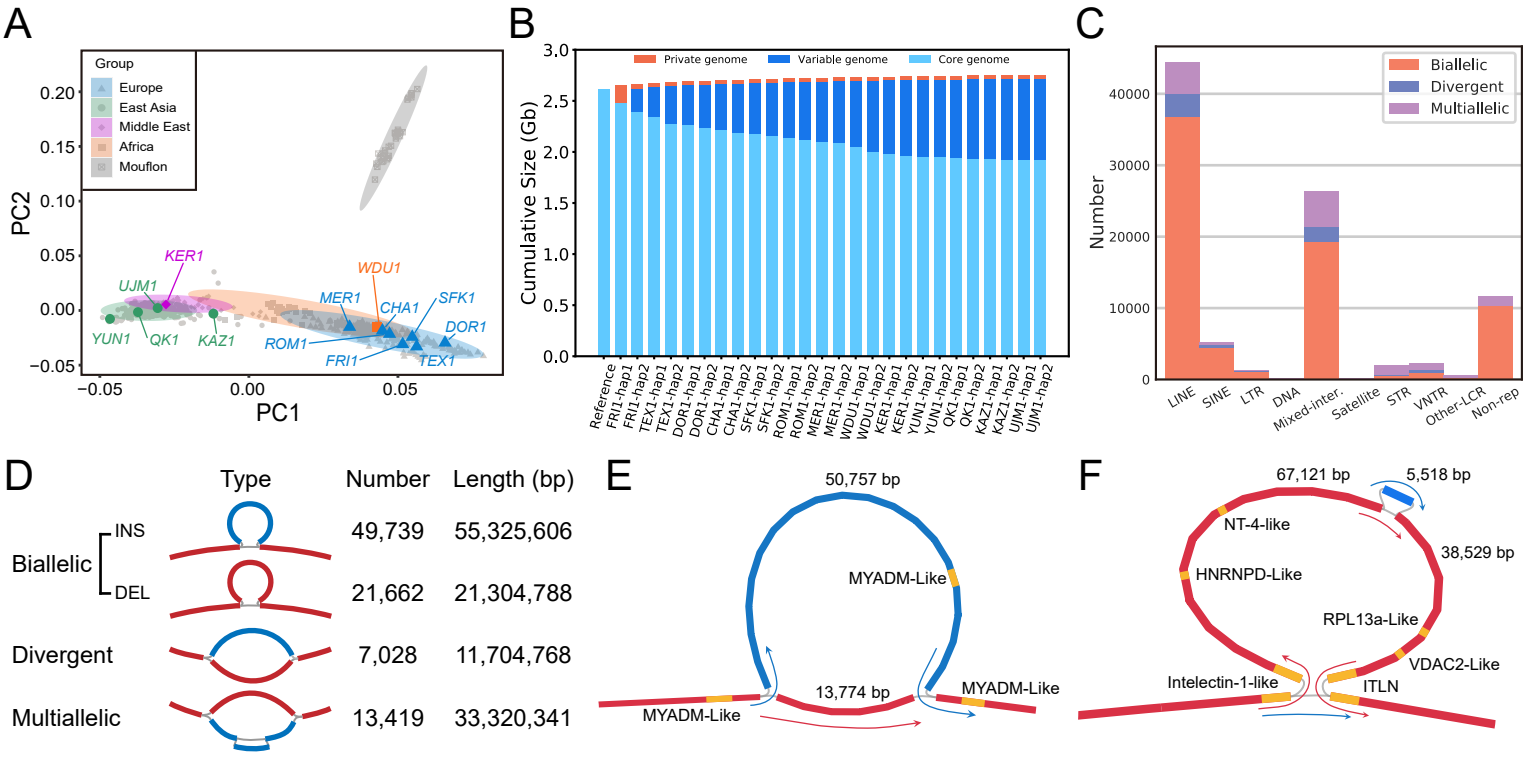
[69]Kijas J W, Lenstra J A, Hayes B, Boitard S, Neto L R P, San Cristobal M, et al. Genome-wide analysis of the world's sheep breeds reveals high levels of historic mixture and strong recent selection. PLoS Biol. 2012;10:e1001258.

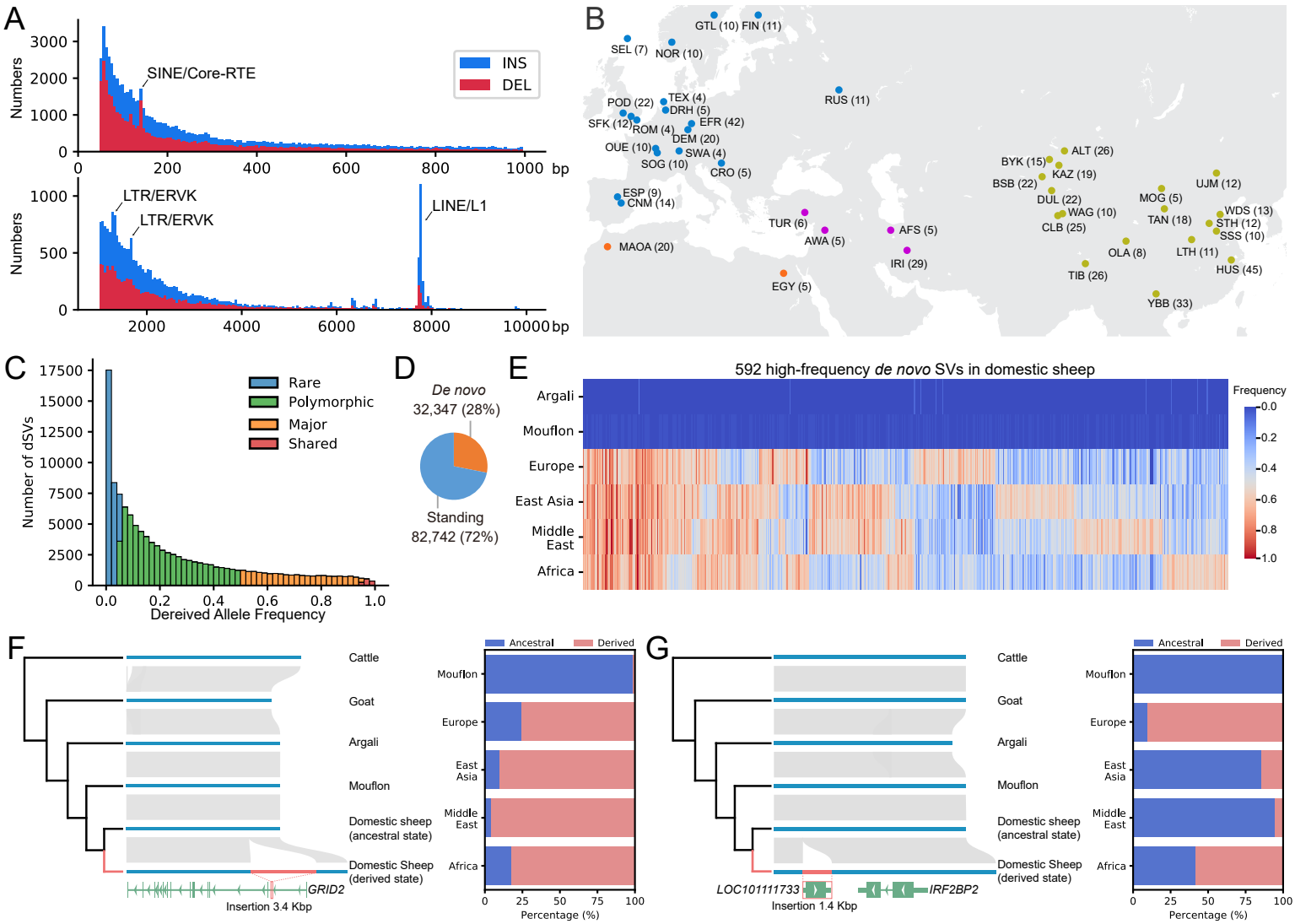
[70]Bolger A M, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30:2114-20.

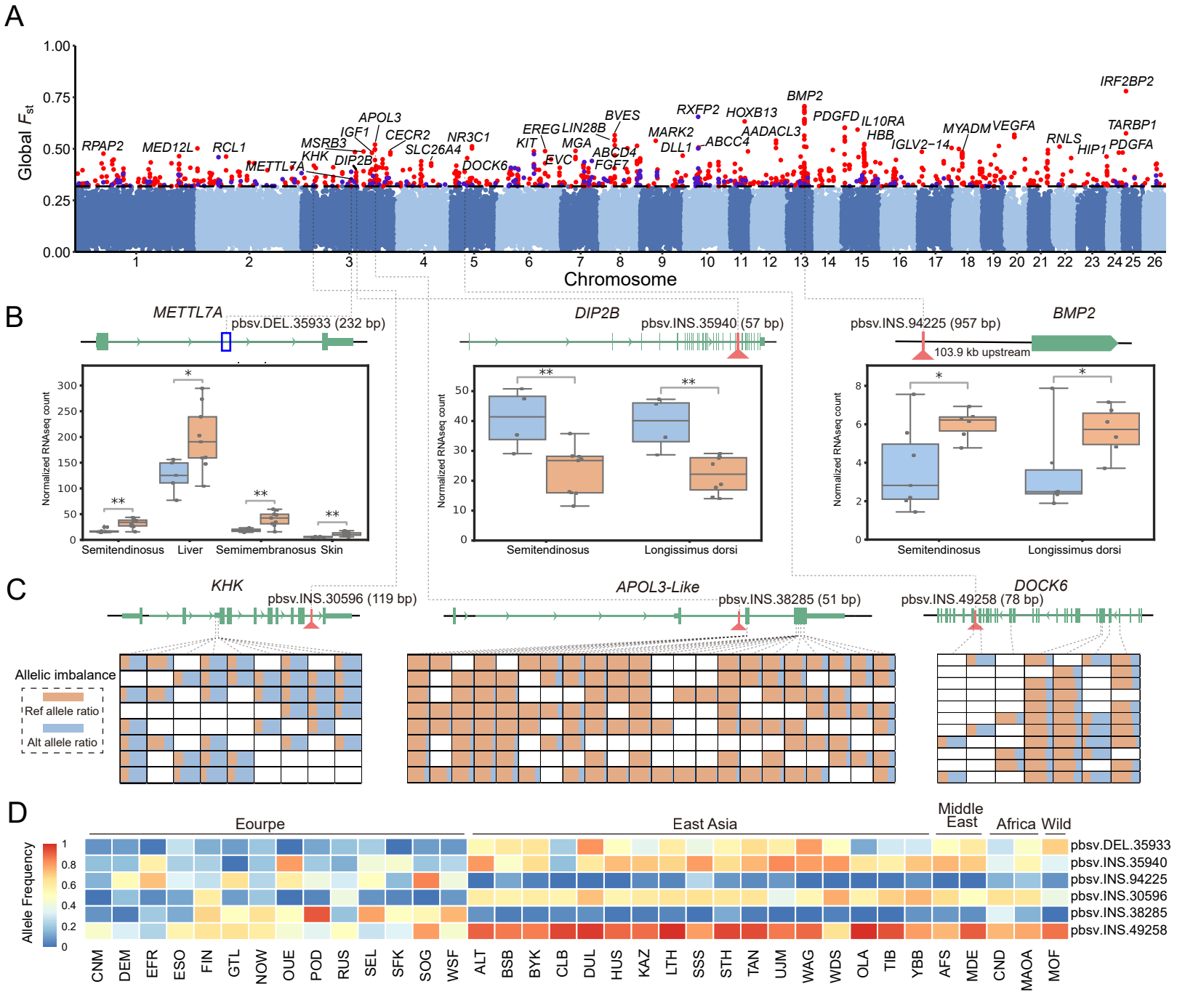
[71]Kim D, Langmead B, Salzberg S L. HISAT: a fast spliced aligner with low memory requirements. Nat Methods. 2015;12:357-60.

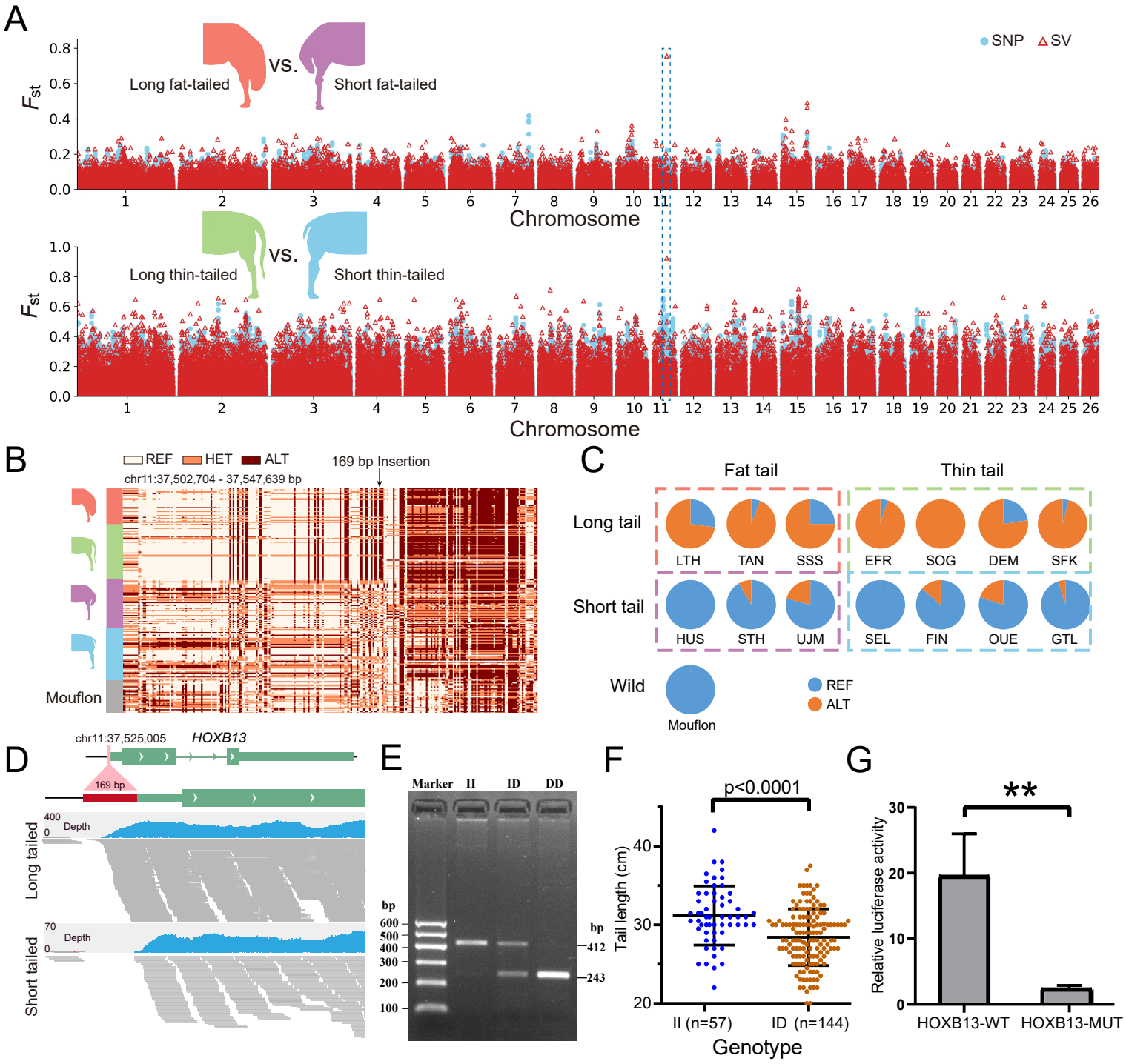
[72]Liao Y, Smyth G K, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2013;30:923-30.

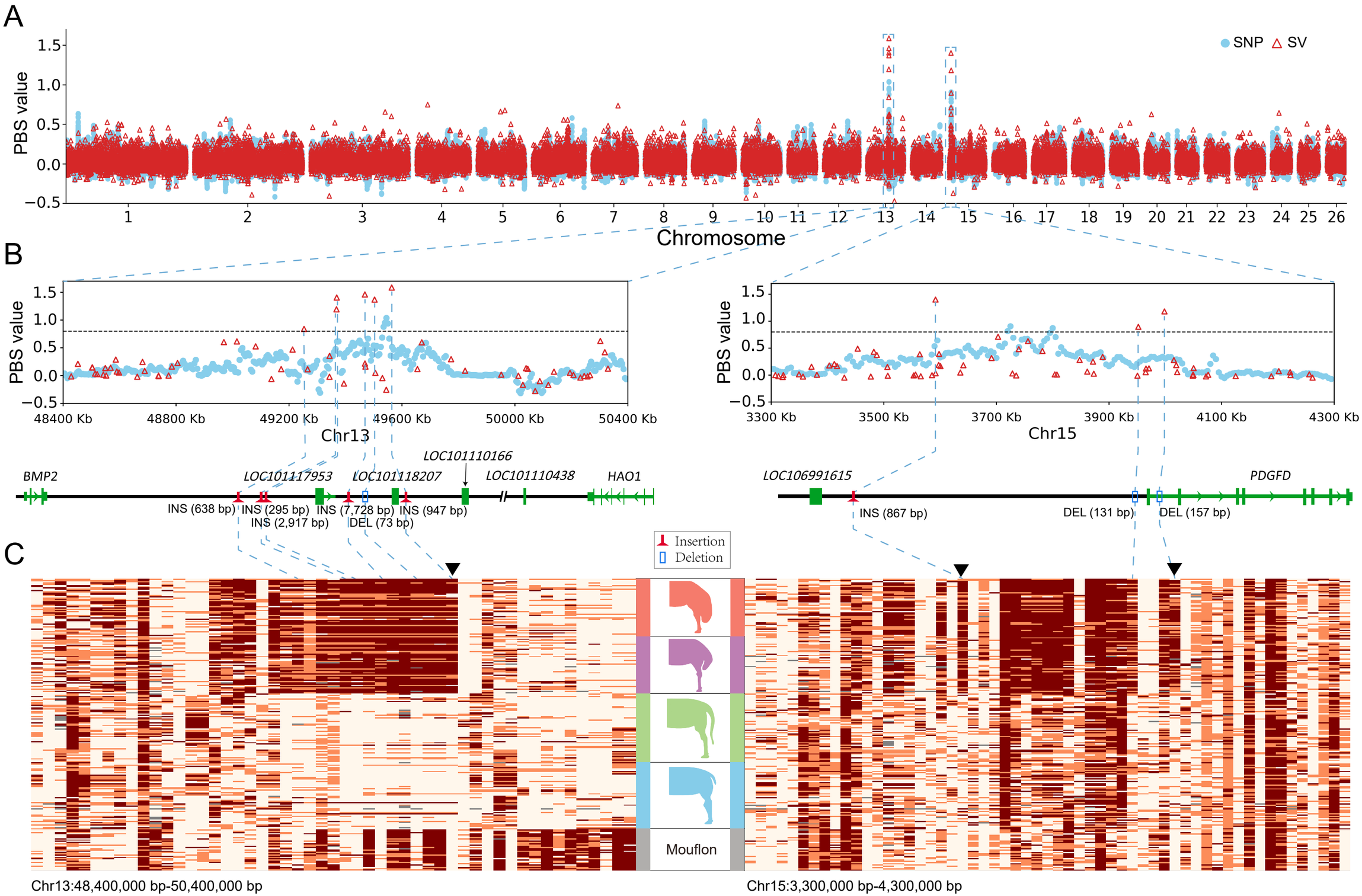
[73]Robinson M D, McCarthy D J, Smyth G K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26:139-40.


A
Group
Europe
East Asia
Middle East
Africa
Mouflon
PC2
PC1
KER1
UJM1
YUN1
QK1
KAZ1
WDU1
MER1
CHA1
SFK1
DOR1
ROM1
FRI1
TEX1
B
Private genome
Variable genome
Core genome
Cumulative Size (Gb)
C
Biallelic
Divergent
Multiallelic
Number
LINE
SINE
LTR
DNA
Mixed-inter.
Satellite
STR
VNTR
Other-LCR
Non-rep
D
Type
Number
Length (bp)
Biallelic
INS
DEL
49,739
55,325,606
21,662
21,304,788
Divergent
7,028
11,704,768
Multiallelic
13,419
33,320,341
E
50,757 bp
MYADM-Like
13,774 bp
MYADM-Like
MYADM-Like
F
67,121 bp
5,518 bp
NT-4-like
38,529 bp
HNRNPD-Like
RPL13a-Like
VDAC2-Like
Intelectin-1-like
ITLN


A
Numbers
SINE/Core-RTE
INS
DEL
bp
LTR/ERVK
LTR/ERVK
LINE/L1
B
GTL (10) FIN (11)
SEL (7)
NOR (10)
RUS (11)
POD (22)
TEX (4)
DRH (5)
SFK (12)
ROM (4)
EFR (42)
DEM (20)
OUE (10)
SWA (4)
SOG (10)
CRO (5)
ESP (9)
CNM (14)
MAOA (20)
TUR (6)
AWA (5)
AFS (5)
IRI (29)
EGY (5)
ALT (26)
BYK (15)
KAZ (19)
BSB (22)
DUL (22)
WAG (10)
CLB (25)
UJM (12)
MOG (5)
TAN (18)
WDS (13)
STH (12)
SSS (10)
OLA (8)
LTH (11)
HUS (45)
TIB (26)
YBB (33)
C
Number of dSVs
Rare
Polymorphic
Major
Shared
Dereived Allele Frequency
D
De novo
32,347 (28%)
Standing
82,742 (72%)
E
592 high-frequency de novo SVs in domestic sheep
Argali
Mouflon
Europe
East Asia
Middle East
Africa
Frequency
F
Cattle
Goat
Argali
Mouflon
Domestic sheep (ancestral state)
Domestic Sheep (derived state)
GRID2
Insertion 3.4 Kbp
Ancestral
Derived
Percentage (%)
G
LOC101111733
IRF2BP2
Insertion 1.4 Kbp


A
Global F_{st}
Chromosome
RPAP2
MED12L
RCL1
METTL7A
MSRB3
KHK
DIP2B
IGF1
APOL3
CECR2
SLC26A4
NR3C1
DOCK6
KIT
EREG
EVC
LIN28B
MGA
ABCD4
FGF7
BVES
MARK2
DLL1
RXFP2
ABCC4
HOXB13
AADACL3
BMP2
PDGFD
IL10RA
HBB
IGLV2-14
MYADM
VEGFA
RNLS
HIP1
IRF2BP2
TARBP1
PDGFA
B
METTL7A
pbsv.DEL.35933 (232 bp)
Normalized RNAseq count
Semitendinosus
Liver
Semimembranosus
Skin
DIP2B
pbsv.INS.35940 (57 bp)
Semitendinosus
Longissimus dorsi
BMP2
pbsv.INS.94225 (957 bp)
103.9 kb upstream
Semitendinosus
Longissimus dorsi
C
KHK
pbsv.INS.30596 (119 bp)
APOL3-Like
pbsv.INS.38285 (51 bp)
DOCK6
pbsv.INS.49258 (78 bp)
Allelic imbalance
Ref allele ratio
Alt allele ratio
D
Allele Frequency
Eourpe
East Asia
Middle East
Africa
Wild
pbsv.DEL.35933
pbsv.INS.35940
pbsv.INS.94225
pbsv.INS.30596
pbsv.INS.38285
pbsv.INS.49258
CNM
DEM
EFR
ESO
FIN
GTL
NOW
OUE
POD
RUS
SEL
SFK
SOG
WSF
ALT
BSB
BYK
CLB
DUL
HUS
KAZ
LTH
SSS
STH
TAN
UJM
WAG
WDS
OLA
TIB
YBB
AFS
MDE
CND
MAOA
MOF


A
Long fat-tailed
vs.
Short fat-tailed
SNP
SV
Fst
Chromosome
Long thin-tailed
vs.
Short thin-tailed
B
REF
HET
ALT
169 bp Insertion
chr11:37,502,704 - 37,547,639 bp
Mouflon
C
Fat tail
Thin tail
Long tail
Short tail
Wild
LTH
TAN
SSS
EFR
SOG
DEM
SFK
HUS
STH
UJM
SEL
FIN
OUE
GTL
Mouflon
REF
ALT
D
chr11:37,525,005
HOXB13
169 bp
Long tailed
Short tailed
Depth
E
Marker
II
ID
DD
bp
600
500
400
300
200
100
412
243
F
p<0.0001
Tail length (cm)
II (n=57)
ID (n=144)
Genotype
G
**
Relative luciferase activity
HOXB13-WT
HOXB13-MUT


A
SNP
SV
PBS value
Chromosome
B
Chr13
Chr15
BMP2
LOC101117953
LOC101118207
LOC101110166
LOC101110438
HAO1
INS (638 bp)
INS (295 bp)
INS (2,917 bp)
INS (7,728 bp)
DEL (73 bp)
INS (947 bp)
LOC106991615
PDGFD
INS (867 bp)
DEL (131 bp)
DEL (157 bp)
C
Insertion
Deletion
Mouflon
Chr13:48,400,000 bp-50,400,000 bp
Chr15:3,300,000 bp-4,300,000 bp