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

Ran Li^{1#}, Mian Gong^{1#}, Xinmiao Zhang^{1#}, Fei Wang^{1#}, Zhenyu Liu^{1#}, Lei Zhang^{1#}, Mengsi Xu², Yunfeng Zhang², Xuelei Dai¹, Zhuangbiao Zhang¹, Wenwen Fang¹, Yuta Yang¹, Huanhuan Zhang¹, Weiwei Fu¹, Chunna Cao¹, Peng Yang², Zeinab Amiri Ghanatsaman³, Niloufar Jafarpour Negari⁴, Hojjat Asadollahpour Nanaei^{1,4}, Xiangpeng Yue⁵, Yuxuan Song¹, Xianyong Lan¹, Weidong Deng⁶, Xihong Wang¹, Ruidong Xiang⁷, Eveline M. Ibeagha-Awemu⁸, Pat (J.S.) Heslop-Harrison⁹, Johannes A. Lenstra¹⁰, Shangquan Gan^{2*}, Yu Jiang^{1*}

¹Key Laboratory of Animal Genetics, Breeding and Reproduction of Shaanxi Province, College of Animal Science and Technology, Center for Functional Genomics, Institute of Future Agriculture, Northwest A&F University, Yangling, Shaanxi, 712100, China

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

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

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

⁵State 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

- 25 Agriculture Science and Technology, Lanzhou University, Lanzhou, 730020, China
- 26 ⁶ Faculty of Animal Science and Technology, Yunnan Agricultural University,
- 27 Kunming, 650201, China
- 28 ⁷ Faculty of Veterinary & Agricultural Science, The University of Melbourne,
- 29 Parkville, 3052, VIC, Australia
- 30 ⁸ Agriculture and Agri-Food Canada, Sherbrooke Research and Development Centre,
- 31 Sherbrooke, QC J1M 0C8, Canada
- 32 ⁹ Department of Genetics and Genome Biology, University of Leicester, Leicester, UK
- 33 ¹⁰ Faculty of Veterinary Medicine, Utrecht University, Utrecht, The Netherlands
- 34 [#]These authors contributed equally.
- 35 ^{*}To whom correspondence should be addressed.
- 36 Email: yu.jiang@nwafu.edu.cn; shangquangan@163.com
- 37

38 **Abstract**

39 Structural variations (SVs) are a major contributor to genetic diversity and phenotypic
40 variations, but their prevalence and functions in domestic animals are largely
41 unexplored. Here, we assembled 26 haplotype-resolved genome assemblies from 13
42 genetically diverse sheep using PacBio HiFi sequencing. We constructed a graph-
43 based ovine pan-genome and discovered 142,422 biallelic insertions and deletions,
44 7,028 divergent alleles and 13,419 multiallelic variations. We then used a graph-based
45 approach to genotype the biallelic SVs in 684 individuals from 45 domestic breeds
46 and two wild species. Integration with RNA-seq data allows to identify candidate
47 expression-associated SVs. We demonstrate a direct link of SVs and phenotypes by
48 localizing the putative causative insertion in *HOXB13* gene responsible for the long-
49 tail trait and identifying multiple large SVs associated with the fat-tail. Beyond
50 generating a benchmark resource for ovine structural variants, our study highlights
51 that animal genetic research will greatly benefit from using a pan-genome graph
52 rather than a single reference genome.

53

54 Keywords: sheep pan-genome, structural variations, *de novo* assembly, tail
55 morphology, *HOXB13*

56

57

58

59 **Introduction**

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

65 However, SV detection has been most challenging^{3, 4}. Identification based on
 66 short reads (≤ 150 bp) has a poor sensitivity and a high false discovery rate⁵. The long
 67 reads of >10 kb by PacBio SMRT sequencing and Oxford Nanopore sequencing
 68 greatly facilitated the efficacy of SV discovery by directly spanning the breakpoints^{6, 7},
 69 ^{8, 9, 10, 11}. However, a sequencing error rate of 5-15% of the conventional long reads
 70 impede downstream analysis and is particularly problematic for analysis of highly
 71 repetitive regions. Another main restraint in efficient SV detection lies in the
 72 reference genome model. The pan-genome graph model has been shown to be much
 73 superior to the linear reference assembly in representing the genome diversity and
 74 thereby facilitating analysis of large genomic variations^{12, 13, 14, 15, 16}.

75 The novel PacBio circular consensus long-read sequencing (CCS) offers high
 76 fidelity (HiFi) reads of 15 to 25 kb with $>99.9\%$ base accuracy¹⁷, which now is
 77 revolutionizing the *de novo* assembly and SV detection. It is expected to foster the
 78 development of pan-genome applications by delivering haplotype-resolved assemblies
 79 to depict the sequence diversity of the diploid genome¹⁸. With the availability of
 80 population-scale genome assemblies, the pan-genomic reference models will allow to

determine faithfully the structure of complex regions for reliable SV detection^{7, 12, 19}.
 Sheep are among the first domesticated livestock with economic and cultural importance. Previous studies have shown that SVs determine the white coat color²⁰, fleece variations²¹ and polledness²² in sheep – characters that have been under selection for millennia. Nevertheless, a population-scale characterization of genetic variation characterized by SVs remains to be explored in sheep as in most farm animal species. In this study, we generated reference-grade *de novo* assemblies for 13 representative sheep breeds, together with two sets of phased assemblies for each individual taking advantage of the Pacbio HiFi sequencing. These datasets allow us to construct a graph-based ovine pan-genome and to explore the SV landscape with unprecedented accuracy and sensitivity. We further analyzed the SV spectrum in 684 domestic and wild sheep from short-reads sequencing and highlighted SVs as largely unexplored contributors to variations in phenotypes and gene expression.

Results

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

Domestic sheep could be generally divided into at least four genetic groups including European, Asian, Middle East and Africa with a separate position of the crossbred Dorper sheep^{23, 24, 25}. We selected 13 representative sheep (Fig. 1a) from seven European, four East Asian, one Middle Eastern (Kermani from Iran) and one African (White Dorper) breeds. The European breeds include meat (Suffolk, Dorset, Texel, Charollais), wool (Merino, Romney) as well as dairy (East Friesian) sheep, whereas

the East Asian accessions represent the three lineages of Northern China (Ujumqin, Kazak), Qinghai-Tibetan Plateau (Tibetan) and Yunnan-Kweichow Plateau (Yunnan)²⁴. A SNP analysis showed that the 13 sheep encompass 95.5% of the SNPs with minor allele frequency (MAF) >0.05 among the studied 643 domestic sheep worldwide.

The 13 sheep were sequenced using PacBio HiFi sequencing at an average coverage of 20.9×, yielding highly accurate long reads with a per-base accuracy >99.9%. We used the Hifiasm assembler¹⁸ for *de novo* assembly, which generated the primary contigs as well as two sets of phased contig assemblies from HiFi reads. We ultimately assembled 13 primary collapsed assemblies with an average length of ~2.86 Gb (Table 1). The average contig number were 1582 with contig N50 lengths of 68.2 Mb. The most complete assembly (FRI1) achieved a contig N50 of 85.1 Mb, which is much higher than for the sheep reference genome (ARS-UI_Ramb_v2.0). The HiFi sequencing depth is not correlated with assembly length ($r_s=-0.04$, Spearman rank correlation), but with contig N50 length ($r_s=0.93$) and negatively with contig N50 number ($r_s=-0.67$). We further ordered and positioned ~93.6% of the contigs from each of the 13 collapsed assemblies into chromosomes based on their synteny to the reference genome. The resulting chromosome-level collapsed assemblies for the 13 representative breeds may very well serve as a reference panel for future research in sheep genetics and genomics. In addition, the haplotype phased assemblies achieved an average contig N50 length of 14.5 Mb, which captures per diploid individual almost all sequence variants and was used for the subsequent construction of the multiassembly graph.

124 On the basis of Benchmarking Universal Single-Copy Orthologs (BUSCO)
125 analysis, the collapsed assemblies displayed an average completeness of 94.1%
126 (93.8%-94.4%, Table 1). BUSCO scores for the partially phased assemblies were
127 lower but still consistently >90% complete. To measure the quality value (QV) of the
128 assemblies, the 13 sheep were sequenced to an average of 30× depth Illumina short-
129 reads, which showed for all the collapsed and phased assemblies except WDU1
130 higher values than obtained for the sheep reference genome assembly (QV=42.5).

131 **Construction of sheep graph-based pan-genome and SV detection**

132 To build a comprehensive sheep graph-based pan-genome, we used the reference
133 genome (ARS-UI_Ramb_v2.0) as the backbone. The 26 haplotype-resolved
134 assemblies from the 13 individuals were individually added to the graph to encompass
135 the sequence variations more comprehensively than the primary contigs. A multigraph
136 genome was generated using minigraph which is a state-of-the-art graph constructor
137 that especially is suitable for discovering complex variations^{26, 27}.

138 The resulting multiassembly graph spans 2.75 Gb and includes 167,565
139 nonreference nodes of 137.7 Mb. Based on the presence or absence of each node
140 among the reference genome and the 26 haplotype-resolved assemblies, we divided
141 the pan-genome into three categories (Fig. 1b): core (nodes present in all assemblies,
142 69.6%), flexible (nodes present in at least two assemblies, 29.0%) and private (nodes
143 present in one assembly, 1.4%). The variations were most associated with LINES
144 followed by VNTR and mixed interspersed repeats (Fig. 1c).

145 The genome graph represents SVs as bubbles that are located between a single

146 source and a single sink node coming from the reference assembly. We classified the
147 SVs into three categories: (1) biallelic insertions and deletions; (2) divergent alleles
148 with a non-reference node; (3) multiallelic variations with more than one non-
149 reference path (Fig. 1d and 1e, Supplementary Data 1 and 2). The genome graph
150 revealed 71,401 biallelic insertions and deletions, 7,028 divergent alleles and 13,419
151 multiallelic variations. However, despite advantage in recovering divergent alleles and
152 multiallelic variations, the genome graph constructed by minigraph is not well adapted
153 for the discovery of small SVs¹⁸ and thus we excluded those below 100 bp. As a
154 complementary approach to detect insertions and deletions, we used pbsv
155 (<https://github.com/PacificBiosciences/pbsv>) to jointly call among the 13 individuals
156 SVs using the PacBio HiFi reads, yielding 142,279 insertions/deletions. We observed
157 much more insertions than deletions, contrasting with results based on short-read
158 sequencing which preferentially identifies deletions²⁸. The pbsv call set covers 97.5%
159 of the insertions and deletions found by minigraph. Therefore, we merged the pbsv
160 call set and the large variants (>5 kb) specifically found by minigraph, resulting in
161 142,422 insertions and deletions, which together with the identified divergent alleles
162 and multiallelic variations constituted a final comprehensive catalogue of SVs in the
163 sheep genome (Fig. 1d and 1e, Supplementary Data 3).

164 The number of insertions/deletions rapidly decreases with length, consistent with
165 reports in humans using long-read sequencing^{6, 7, 29}. Noticeable peaks at 150 bp
166 correspond to SINES, 1,300 bp for LTRs, as well as 7,750 bp for LINEs (Fig. 1d). The
167 median length of divergent alleles (726 bp) and multiallelic loci (575 bp) are longer

168 than that of biallelic variations (230 bp).

169 The insertion sequences were predicted to harbor 588 genes whereas the deletions
170 potentially affect the exonic region of 1920 genes by containing exons. The
171 nonreference sequences of the divergent alleles contained 31 genes. We revealed the
172 full sequence of one divergent locus containing *MYADM-like* repeat (Fig. 1f), which
173 was previously reported to be associated with mean corpuscular hemoglobin
174 concentration traits and weight of lamb weaned but the sequence was unknown³⁰. We
175 also noticed several other large divergent alleles that likely to correspond to
176 phenotypic variants (Supplementary Fig. 1). The non-reference alleles of the
177 multiallelic loci were predicted to harbor 146 genes. One of the largest multiallelic
178 region represented a deletion of 105.6 kb which is nested with an insertion of 5.5 kb
179 (Fig. 1g). The deletion likely results in the loss of the intelectin genes, that displays
180 extremally high expression in lung of sheep and potentially plays a role in mucosal
181 response in allergy³¹. Additional examples of large multiallelic loci that resulted in
182 gene gain/loss, included *HBB*, *ASIP*, *UDP-glucuronosyltransferase 2B18-like*,
183 olfactory receptor (*LOC101115700*), etc. (Supplementary Fig. 2).

184 The SVs are underrepresented in untranslated region (UTR), coding sequences
185 (CDS), and regulatory elements (H3K4me3, H3K27ac), suggesting strong negative
186 selection that acted against them (Supplementary Fig. 3). We also reported 201 SV
187 hotspots sites covering 18.2 Mb of the autosomes (Supplementary Fig. 4). The SVs
188 were significantly clustered at the terminal 5 Mb region of each chromosome with
189 4.1-fold enrichment (P value<0.01, permutation test) (Supplementary Fig. 5).

190 **Frequency spectrum of SVs in the sheep populations**

191 In order to better understand the frequency spectrum of the SVs within sheep, we used
 192 the graph genotyping software Paragraph³² to genotype the biallelic insertions and
 193 deletions in autosomes in a panel of 684 resequenced samples from 45 domestic
 194 breeds or populations and two wild ovine species. The samples included 643 domestic
 195 sheep from Europe, East Asia, Middle East and Africa, 33 Asiatic mouflons and 8
 196 argali with an average coverage of 18.1× (Fig. 2a, see Methods for the assignment of
 197 the breeds and populations. Paragraph represent SVs as diverging paths in a graph and
 198 achieves a high rate of genotyping accuracy³³.

199 To ensure a high-quality set of autosomal SVs for population genetics analysis,
 200 we excluded 17,330 SVs that failed to be genotyped in >90% of the samples and
 201 another 2,011 that by an excess of heterozygotes deviated from Hardy Weinberg
 202 equilibrium, leaving 116,776 (85.6%) reliably genotyped SVs for downstream
 203 analysis. The frequency of the SVs was provided in Supplementary Data 4. We also
 204 estimated the genotyping accuracy of the remaining SVs by measuring the Mendelian
 205 errors, which represent genotypes that are found in the offspring but could not be
 206 inherited from either parent. The Mendelian error rate was 3.8% on average in the 11
 207 trios of 29 individuals (each ~30× coverage) used in our previous study³⁴.

208 PCA analysis using either insertions or deletions displays similar population
 209 structure and shows a clear separation of the four genetic groups (Supplementary Fig.
 210 6), agreeing with the clusters defined by SNPs. Therefore, the SV catalogue
 211 reproduces the genetic diversity of the worldwide sheep population.

212 The level of linkage disequilibrium (LD) between SVs and SNPs could reflect the
213 extent of SVs representing independent and undiscovered genetic variations relative
214 to previous evolutionary studies. We estimated how frequently the SVs were linked to
215 the flanking SNPs (Fig. 2b). Around 39.8% SVs link strongly with surrounding SNPs
216 ($\text{rank} > 200$), the characteristic of which could be detected by the surrounding linked
217 SNPs. It is worth noting that there are still 39.2% SVs display low levels of LD with
218 nearby SNPs ($\text{rank} < 100$), implying variants that are poorly tagged by easily
219 genotyped SNP markers.

220 In order to infer the allele state (ancestral or derived) for each SV, we additionally
221 genotyped one snow sheep (*Ovis nivicola*) and one bighorn sheep (*Ovis canadensis*)
222 and, together with one argali (*Ovis ammon*) as outgroup, and identified the state for
223 115,089 SVs and a variable allele frequency of derived alleles (DAF) in domestic
224 sheep (Fig. 2c). About 28% of the derived alleles occurs only in domestic sheep (Fig.
225 2d) and may have reached a high frequency by domestication and improvement. The
226 proportion of *de novo* SVs may be overestimated due to the relatively limited number
227 of wild sheep but this does not explain a high geographic differentiation of 592 high-
228 frequency *de novo* SVs ($\text{DAF} \geq 0.5$ in at least one geographic group (Fig. 2e). Those
229 SVs were found to be enriched in KEGG pathways of Ras signaling and long-term
230 depression (adjusted P value < 0.05). For example, we found in all regions a *de novo*
231 3.4 kb insertion in the first intron of *GRID2* (Fig. 2f), which plays a key role in
232 synaptogenesis and neurodevelopment³⁵. A 1.4 kb *de novo* insertion prevalent in
233 European sheep was found at the 3' UTR of *IRF2BP2* (Fig. 2g), which was reported

234 as the causative mutation for the wool coat²¹.

235 **Population-stratified SVs and their potential effect on gene expression**

236 After initial domestication at the Fertile Crescent, domestic sheep dispersed
237 worldwide and adapted to a wide range of natural and artificial selection. In this
238 context, we detect putative selection signals by dissecting population-stratified SVs
239 among the selected 33 out of the 45 breeds or populations to ensure a similar number
240 of breeds or populations for European and East Asian sheep. For the differentiation
241 between these groups, we used a modified calculation of global F_{ST} (see Methods).
242 Because this strategy tends to discover common selection signals present in multiple
243 breeds, the population-stratified SVs differentiating Europe and East Asian breeds
244 were preferentially detected since they are in our panel the predominant breeds.

245 We identified 865 population-stratified SVs with the top 1% global F_{ST} values
246 (Supplementary Data 5), 351 of which intersect with genic regions and 324 are within
247 100 kb of gene flanking region (Fig. 3a). The majority of them (698, 80.7%) were
248 confirmed by the Ohana approach³⁶ with which we detected selection signals in
249 multiple populations by varying the number of ancestral admixture components. The
250 related genes are associated with diverse phenotypic and production traits, including
251 wool type (*IRF2BP2*, *FGF7*)^{21, 37}, horn type (*RXFP2*)³⁸, fat deposition (*BMP2*,
252 *PDGFD*, *PDGFA*)^{23, 39}, and coat color (*KIT*)⁴⁰, etc. Notably, the highest stratified SV
253 ($F_{ST}=0.78$) is the 1.4 kb insertion downstream of *IRF2BP2* (Fig. 2g), differentiating
254 European and Asian breeds. Functional annotation of the *top* F_{ST} signals reveals
255 enrichment in several KEGG pathways, such as those involved in signaling pathways

256 (Rap1, MAPK) and parasitic disease (African trypanosomiasis, amoebiasis)

257 (Supplementary Fig 7).

258 SVs may impact the gene expression of neighboring genes via the cis-regulatory

259 elements^{10, 41, 42}. To explore SV-associated expression changes, we used 112 RNA-seq

260 data from 8 tissues of 14 hybrids of Texel×Kazak sheep as well as WGS data of the

261 same sheep^{34, 43} (see Methods). This unique combination of data enabled us to identify

262 the candidate expression-associated SVs by allelic-specific expression (ASE)

263 mapping. By ASE mapping, we found 1,141 heterozygous SVs that would putatively

264 act as cis-regulatory elements and thus cause allelic-specific expression in associated

265 genes (Fig. 3b and 3c, Supplementary Data 6). These expression-associated SVs were

266 significantly enriched in generic regions (P value $<2.2\text{e-}16$, permutation test), and 34

267 overlapped with population-stratified SVs which is greater than random chances (P

268 value $<1.1\text{e-}10$, Chi-Square test). We additionally performed SV-eQTL mapping and

269 found 260 SVs with genotypes that correlate with expression levels of a nearby gene

270 (Supplementary Data 7).

271 **SVs as strong candidates affecting sheep tail morphology**

272 The tail morphology of sheep is a key genetic trait that is related to domestication,

273 adaptation, and animal welfare⁴⁴. The wild ancestors (Asiatic mouflon) have a short

274 thin tail whereas four domestic tail types include in addition to the primitive type the

275 long thin-tailed, the short fat-tailed (including fat-rumped) and the long fat-tailed

276 types. Although several genes have been found to be associated with these traits^{23, 45,}

277 ⁴⁶, the underlying causative mutations are still unknown. However, the long and fat

278 tails are derived traits, indicating that the causative mutations are likely to be *de novo*
279 variants.

280 To identify SVs associated with tail length, we compared the long and short tailed
281 breeds of (1) the Asian long fat-tailed breeds versus Asian short fat-tailed breeds from
282 Asian breeds and (2) the European long thin-tailed breeds versus the North European
283 short thin-tailed breeds, respectively. Notably, in both independent comparisons the
284 most differentiated SVs point to an insertion of 169 bp (chr11:37525005) close to the
285 5' UTR of *HOXB13* (Fig. 4a). This resides in a specific haplotype exclusively found
286 in long tailed sheep, where the insertion is prevalent in long-tailed breeds and rare in
287 short-tailed breeds (Fig. 4b and 4c). *HOXB13* is mainly expressed in the embryo and
288 colon of sheep and cattle⁴⁷. We found the expression of *HOXB13* in colons of the
289 short-tailed Texel×long-tailed Scottish Blackface individuals presumably carrying at
290 least one copy of the insertion⁴⁸, and in colons of individuals with short tail
291 presumably without the insertion (Texel sheep and small-tail sheep)^{49, 50}. The insertion
292 was only expressed in the colons of the Texel×Scottish Blackface but not in Texel
293 sheep and small-tail sheep (Fig. 4d and Supplementary Fig. 8).

294 In order to verify the effect of this insertion on tail length, we evaluated the
295 association between the insertion and tail length in a population of 201 sheep of the
296 same age (nine months old). The three genotypes including homologous insertion (II),
297 heterozygote (ID) and homologous deletion (DD) were directly distinguished through
298 PCR amplification and agarose gel electrophoresis (Fig. 4e). Despite lacking the DD
299 individuals in the studied population (see Methods), the tail length of the homologues

(II) was found to be significantly longer than that of the heterozygotes (ID) by 2.77 cm on average (Mann-Whiney U test, P value<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 decreased luciferase activity by 10-fold (Fig. 4g), indicating that the insertion reduces the transcriptional level of *HOXB13*. The gene has been previously identified as a repressor of caudal vertebra development^{51,52}. Therefore, the insertion is likely to increase the tail length of sheep by depressing the expression of *HOXB13*.

To identify candidate SVs related to the 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 an outgroup. The two most differentiated regions are identified by either SV or SNPs (Fig. 5a and 5b), one located at the intergenic region between *BMP2* and *HAOI* (hereafter referred to as IBH region) and the other in *PDGFD*. Notably, the most stratified SVs in these two regions are among the top selective variants indicated by both SVs and SNPs. The most differentiated SV cluster are 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 IBH selective region (Fig. 5c), of which the largest one is a 7728 bp insertion. *PDGFD* has been recognized as the most plausible candidate gene for the fat-tail phenotype in recent studies^{23,55}. 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 are all domestic *de novo* SVs. Therefore, our results suggest that both regions are associated with the fat-tail phenotype and the SVs identified within these regions are the most plausible candidate causative mutations.

Discussion

The genome-wide 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, precluding the discovery of regions with pronounced structural variations⁴. The graph-based pan-genome model largely reshapes the methodology to assess the whole landscape of sequence diversity, and is likely to transform the genomic research field. To build a high-quality ovine pan-genome, we utilized the PacBio HiFi sequencing to generate a panel of reference-level assemblies and 26 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 SV' breakpoints. The development of this ovine pan-genome will contribute to shifting in animal genomics research from a linear reference genome to a pan-genomic model comprising multiple reference-quality representative genomes.

343 The application of graph-based pan-genome has its clear opportunities and
344 obstacles. The pan-genomic approaches are in practice not so straightforward as the
345 linear reference genome and the methodologies and implementations are still in active
346 development^{12, 14, 56, 57}. For example, minigraph is one of the most promising graph
347 constructors, but has limitations in performing base alignment, generating relatively
348 high false positives for small SVs²⁶. Nevertheless, a shift to the graph-based pan-
349 genomes model will enhance the opportunities to systematically investigate
350 previously inaccessible genomic variations. In our study, we found that the minigraph
351 performed very well in discovering divergent alleles and multiallelic variations, which
352 accounted for 27.4% of the total variations in length but are previously neglected.
353 Combining the graph-based pan-genome with read-based calling yielded a reference
354 panel of high-quality SVs for downstream functional genomic studies.

355 We also demonstrated how these pan-genome datasets can be utilized to link SVs
356 to phenotypes by genotyping and assessing the SV landscape in a large population
357 from short-read sequencing. We genotype SVs using the graph-based Paragraph,
358 which enables efficient encoding of structural variation into the pan-genome³². The
359 population-scale study allows us to trace their responses to natural and artificial
360 selection. Taking the tail morphology traits as an example, we identified a plausible
361 causative mutation in *HOXB13* for the long-tail trait, which illustrates the key roles of
362 SVs underlying phenotypic changes. We further reported segregating SVs in fat-tailed
363 breeds. The observation that *de novo* SVs showed higher or comparative selection
364 signals than surrounding SNPs warrants further investigation.

365 In conclusion, the pan-genome representation with an influx of high-quality
366 assemblies will revolutionize our understanding of SV prevalence and their biological
367 significance, giving a new dimension to the nature and extent of genomic variation,
368 beyond SNPs, impacting traits.
369

370 **Methods**

371 **Sample collection and PacBio sequencing**

372 Samples were selected to for an optimal representation of the global sheep diversities.
 373 The sampling information were provided in Supplementary Table 1. High-quality
 374 genomic DNA (gDNA) was extracted from fresh blood samples as previously
 375 described⁵⁸ and assessed for purity and quantity using Nanodrop 1000 (Thermo Fisher
 376 Scientific, CA, USA) and Qubit (Thermo Fisher Scientific, CA, USA) assays. Library
 377 with an average insert size of ~15 kb was generated using the SMRTbell Express
 378 Template Prep Kit 2.0 (Pacific Biosciences, CA, USA) and fractionated on the
 379 SageELF (Sage Science, Beverly, MA, USA) into narrow library fractions. Libraries
 380 were then sequenced on 2-3 SMRT Cells 8M on a Sequel II instrument (Pacific
 381 Biosciences, CA, USA) using 30-hour movie times at Annoroad Gene Technology
 382 Co., Ltd (Beijing, China). Raw data was processed using the CCS algorithm (version
 383 6.0.0, parameters: --minPasses 3 -- minPredictedAccuracy 0.99 --maxLength 21000)
 384 to generate highly accurate HiFi reads.

385 ***De novo* assembly using HiFi reads and quality assessment**

386 Hifiasm v0.15.3-r339 was used to generate the assembly from HiFi CCS reads using
 387 default parameters¹⁸. Hifiasm yields one primary contig assembly and two sets of
 388 partially phased contig assemblies. The primary contig assembly for each breed was
 389 scaffolded to chromosome level using RagTag v2.0.1 with the reference sheep
 390 genome assembly⁵⁹.

391 The genome completeness was assessed using the BUSCO program (v.3.0.2)
 392 containing the Mammalia odb9 set of 4,104 genes⁶⁰. The single plus duplicated
 393 complete BUSCO gene counts are reported. The base accuracy was measured by
 394 assembly quality value (QV) using yak (<https://github.com/lh3/yak>), which compares
 395 the 31-kmers found in short-reads and the assembly sequences.

396 **Pan-genome construction**

397 The pan-genome construction was performed using a previously reported workflow¹⁵
 398 with minor modifications. Minigraph v0.15-r426 with option -xggs was used to
 399 generate the graph pan-genome²⁶. The sheep reference genome (ARS-UI_Ramb_v2.0,
 400 GCF_016772045.1) was used as the backbone of the multiassembly graph and the 26
 401 partially phased contig assemblies were added successively. Only autosomal
 402 sequences were considered in the multigraph genome, as the X and Y chromosomes
 403 of males were underrepresented.

404 **Identification of non-reference divergent alleles and multiallelic alleles from the** 405 **pan-genome graph**

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

411 Depending on the number of paths in a bubble, the structural variations were
 412 classified into three categories: (1) biallelic variations (insertions/deletions): either the

reference path is longer (deletions) or the non-reference path is longer (insertions).
The node in the shorter path is required to be below 100 bp. (2) Divergent alleles: the reference nodes and the non-reference nodes were both longer than 100 bp. (3) Multiallelic alleles: 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

We used RepeatMasker v4.0.5 (<http://www.repeatmasker.org>) to softmask the non-reference sequences. Then we performed an ab initio gene structure prediction using Augustus v3.3.3⁶¹ and searched the protein sequence predicted by Augustus against the local protein database using DIAMOND⁶² blastp (parameters: --more-sensitive --evaluate 1e-10 --max-target-seqs 1) and kept hits with minimum coverage of 70% and identity of 80%. The local protein database was built using DIAMOND⁶² makedb with RefSeq protein sequences from sheep, goat and cattle.

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 v4.09⁶³. The identified tandem repeats as well as those low complexity sequences identified by RepeatMasker were merged as low-complexity regions (LCR). Furthermore, LCR with unit motif lengths ≥ 7 bp were classified as VNTR and those with unit motif lengths ≤ 6 bp were classified as STR.

435 **Structural variant calling**

436 We integrated the results from graph-based pan-genome analysis and alignment-based
 437 SV calling to generated high-confidence SV calls. For pan-genome analysis, the SVs
 438 were generated from our multiassembly graph genome using gfatools v0.5-r234 as
 439 described above. For alignment-based SV calling, the SVs were detected using pbsv
 440 tools v2.6.2 (<https://github.com/PacificBiosciences/pbsv>). The CCS reads were
 441 aligned to the reference genome using pbmm2 to generate the alignment BAM file.
 442 Then the pbsv discover module was used to identify signatures of structural variation
 443 within each aligned BAM. The tandem repeat annotation file was provided to increase
 444 the sensitivity and recall of this step. Finally, the signatures of all the samples were
 445 fed to the pbsv call module to jointly call SV from structural variant signatures using
 446 the `-ccs` mode.

447 **Mendelian error rate estimation**

448 A Mendelian inheritance error represents a genotype in the offspring that cannot be
 449 inherited from either of their parents. The whole genome sequencing data of 11
 450 trios were from our recently study for allelic specific expression analysis³⁴. For each
 451 trio, the Mendelian error rate is the proportion of SVs from the children that deviate
 452 from Mendelian inheritance.

453 **SV hotspot analysis**

454 SV hotspot regions were determined using primatR package “hotspotter” section
 455 (`bw=200000`, `num.trial=1000`, `pval<1e-8`)⁶⁴. SVs were significantly clustered at the
 456 terminal 5 Mb region of each chromosome (P value<0.01, permutation test, random

457 shuffling of 1,000 times). The random shuffling of SV intervals was performed using
458 BEDTools v2.29.0 with the shuffle command⁶⁵.

459 **Enrichment of SVs in the genome**

460 The H3K4me3 and H3K27ac peak regions were retrieved from our Ruminant genome
461 database. We only used insertions and deletions for enrichment. For insertions, the
462 end position is the start position plus the insertion length. We first recorded the
463 observed number of overlaps between SVs and UTR, exon, intron, H3K4me3 and
464 H3K27ac peak regions, respectively. Then we counted the average number of
465 overlaps between randomly shuffled SVs (1,000 times) and each genomic region.

466 **SV genotyping in whole genome sequencing data**

467 We collected whole-genome sequencing data for 643 domestic sheep, from four
468 geographic regions (Europe, East Asia, Middle East and Africa) as well as wild sheep
469 including 33 Asiatic mouflons (*Ovis aries musimon*) and 8 argali sheep (*Ovis ammon*).
470 For the domestic sheep, we selected breeds with at least five individuals and defined
471 populations by combining breeds from the same country with less than 5 individuals.
472 In this way, all the domestic sheep were assigned to 45 breeds or populations
473 (Supplementary Data 7). We genotyped SVs in these samples using Paragraph v2.4a³².
474 The maximum permitted read count for each variant was set to 20 times the average
475 sample depth in order to reduce runtime for repetitive regions. The resulting genotype
476 files in vcf format from all samples were then combined using bcftools v1.9⁶⁶.

477 To obtain a high-confidence call set of SVs, the resulting genotypes were filtered
478 based on overall genotyping rates and Hardy-Weinberg equilibrium³³. Those variants

479 that were successfully genotyped in $\geq 90\%$ of all samples were retained using
480 vcfTools⁶⁷ for further analysis. We also estimated Hardy-Weinberg equilibrium for
481 each variant using vcfTools and those with excess of heterozygotes (P value $< 1 \times 10^{-5}$)
482 were removed.

483 **Quantifying linkage disequilibrium (LD) of SVs with surrounding SNPs**

484 In order to measure the extent to which SVs are correlated with SNPs, we calculated
485 LD (r^2 values) between the polymorphic SVs (MAF > 0.05) and the neighboring
486 polymorphic SNPs (MAF > 0.05 ; 150 upstream and 150 downstream SNPs) in
487 Mouflon and domestic sheep. SNPs located on SVs were not considered. Plink v1.9
488 was used to calculate LD for all pairwise cases for SV-SNP and the surrounding SNP-
489 SNP variant pairs. According to a previous study⁴¹, we first obtained pairwise
490 genotype correlations (r^2) for all SNP-SNP and SNP-SV pairs. For each of the 300
491 ranked surrounding positions, the number of times the SV rank was greater than the
492 SNP-SNP median rank was recorded. Low, middle or high LD of each SV were then
493 classified if the relative LD metric of SV to SNP was ≤ 100 , > 100 and ≤ 200 , and > 200 ,
494 respectively, respectively.

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

496 To infer the state (ancestral or derived) for each SV, we genotyped one snow sheep
497 ($12.11 \times$ depth, ERR4161992) and a bighorn sheep ($10.3 \times$ depth, SRR501898), which
498 together with one argali sample served as outgroups. For each biallelic SV, if the same
499 homozygous genotypes were detected in at least two outgroups, this state was defined
500 as the ancestral state while the other was the derived state. If the derived allele is only

501 present in domestic sheep but absent in all the wild sheep (8 Argali and 33 Asiatic
502 mouflons), the SV was defined as a *de novo* SV in domestic sheep.

503 **Detecting selection signals**

504 To detect population-stratified SVs, we employed two strategies, the global F_{ST} and
505 the Ohana method, which identifies extreme allele frequency differences by modelling
506 ancestral admixture components³⁶. The global F_{ST} was calculated following a
507 previously described approach with modifications⁶⁸. For each SV site, we calculated
508 the F_{ST} values between all populations. Instead of using the average of all F_{ST} values,
509 we took the average of the top 50% F_{ST} values as the global F_{ST} value for each SV.
510 The top 1% of the highest global F_{ST} value by this method were considered as
511 putative selective signals. For the Ohana method³⁶, we first modeled the dataset with
512 the provided workflow using the SNPs³³. The resulting covariance was used as the
513 neutral input for scans of selection. For each specific ancestry component, Ohana
514 reported a likelihood ratio statistic (LRS) to quantify the likelihood of selection for
515 each variant. We varied the number of ancestral admixture components from K=2 to
516 K=5 to find selection signals.

517 In order to reveal selection signals in fat-tailed sheep, we used the method of
518 population branch statistic (PBS) to calculate and compare the Fixation index (F_{ST})
519 between the fat-tailed sheep, thin-tailed sheep and Asiatic mouflon, using the Asiatic
520 mouflon as a distantly related population. The F_{ST} value between populations with
521 different traits was calculated using the python scikit-allel package v1.3.2
522 (<https://github.com/cggh/scikit-allel>). We calculated the F_{ST} values of both SNP and

SV sets. F_{ST} values for SNPs were calculated with a 10 kb sliding window with a 5 kb step.

525 **Detecting expression-associated SVs**

526 The RNA-seq data of 8 tissues (liver, rumen, tail fat, skin, subcutaneous fat,
527 semitendinosus, semimembranosus, longissimus dorsi) from 14 Texel (♂)×Kazak (♀)
528 sheep hybrids with accession number PRJNA485657 as well as their whole genome
529 sequencing data were downloaded from NCBI Sequence Read Archive^{34, 43}. The raw
530 reads were first trimmed for adapter and low-quality sequences using Trimmomatic⁶⁹
531 and the high-quality reads were aligned to the sheep reference genome using Hisat2⁷⁰.
532 Reads counts per gene were computed using featureCounts⁷¹ and normalized to
533 counts per million (CPM) using edgeR package⁷² and those with CPM>1 in at least
534 two samples were retained.

535 It is expected that the regulatory SVs could cause allelic-specific expression. We
536 thus performed allelic specific expression (ASE) mapping by integrating the RNA-seq
537 data as well as the SV genotyping information for each of the samples. The ASE
538 genes were identified following the pipeline described in our previous study³⁴. First,
539 for each sample, the ASE SNPs displaying allelic imbalance at the gene body were
540 identified from RNA-seq data using the following parameters: (1) at least 20 total
541 reads; (2) allele ratio (allele reads count/total reads count)> 0.65 or <0.35; (3)
542 significant allelic imbalances with FDR<0.05 (Chi-squared test). Those with at least
543 two ASE SNPs at the gene body and found in at least three samples were defined as
544 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 plausibly expression-associated.

The SV-based eQTL analysis was performed using the R package Matrix eQTL v.2.3⁷³. The midpoint where SV starts and ends were defined as the position of each SV. Cis-eQTL analysis was tested within each tissue against all SVs located within the 200 kb upstream and downstream of neighboring genes. The gender and population structure (the first three principal components from PCA analysis of the 14 individuals) were included as covariates to account for confounding effects using an additive linear model. Statistical significance was considered with FDR<0.05.

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

The association analysis was conducted in a F2 population which was generated by backcrossing East Friesian sheep (♂) with the female Hu sheep (♀) × East Friesian sheep (♂). A total of 211 individuals of the similar age (six months old) were used. The tail length (cm) of each individual was measured and DNA samples were collected. 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 by denaturation of 94°C for 2 min followed by 28 cycles of 94°C for 30 s, 62°C for 30 s,

567 and 72°C for 30 s and an extension of 72°C for 2 min; and a hold at 4°C.

568 **Luciferase reporter assay**

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

574 The 293T cells in 96-well plates were transiently transfected with 200 ng Firefly
575 luciferase vector (pGL3) and 4 ng pRL-TK Renilla luciferase vector (Promega,
576 Madison, USA) using 0.5 uL Lipofectamine 3000 (Thermo Fisher Scientific, Waltham,
577 USA). Twenty-four hours after transfection, luciferase activities were measured with a
578 dual-luciferase reporter assay kit (Promega, Madison, USA). Renilla luciferase
579 activity was used as the transfection control for normalizing Firefly luciferase activity.
580 The pGL3-Basic vector without insertion was used as the reference for promoter
581 activity assays and the pGL3-Control vector was used as the positive control for
582 promoter activity.

583

584 **Author contributions**

585 Y.J. and S.Q.G. conceived and designed the experiments. R.L., X.M.Z. and Z.Y.L.
586 performed *de novo* assembly and SV calling; Z.Y.L., M.S.X. and Y.F.Z. analyzed the
587 RNA-seq and Chip-seq data; F.W., M.G. and L.Z. contributed to the ASE analysis;

588 R.L., Z.Y.L., W.W.Fang and X.L.D. carried out SV genotyping and population
589 genetics analysis; Y.T.Y. and X.Y.L. conducted the luciferase reporter assay; Z.B.Z.,
590 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.
591 collected the samples and prepared the WGS data; R.L. wrote the manuscript and Y.J.,
592 S.Q.G., J.A.L., P.H., E.I.A., R.D.X. and X.H.W. revised it. All authors read and
593 approved the final manuscript.

594 **Acknowledgements**

595 We thank the High-Performance Computing platform of Northwest A&F University
596 for providing computing resources. We would also like to thank Prof. Xiaolong Wang
597 for providing the nice chart of different sheep tail patterns. This study was supported
598 by research grants from the National Natural Science Foundation of China to Y.J.
599 (U21A20247, 31822052), S.Q.G. (31760662, 31860639) and R.L. (31802027),
600 Natural Science Basic Research Program of Shaanxi (K3030121801), and China
601 Postdoctoral Science Foundation to W.W.Fu (No. 2021M702690).

602 **Competing interests**

603 The authors declare no competing interests.

604 **Ethics statement**

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

607 **Data availability**

608 Sequences and metadata generated in this work are publicly available. The *de novo*
609 assemblies, HiFi data and WGS data for the 13 individuals are available at NCBI:
610 WDU1 (PRJNA788705, PRJNA788704, PRJNA784371), QK1 (PRJNA788680,
611 PRJNA788679, PRJNA784330), ROM1 (PRJNA788688, PRJNA788687,
612 PRJNA784357), SFK1 (PRJNA788697, PRJNA788696, PRJNA784366), UJM1
613 (PRJNA788702, PRJNA788701, PRJNA784376), CHA1 (PRJNA788666,
614 PRJNA788665, PRJNA784396), YUN1 (PRJNA788712, PRJNA788711,
615 PRJNA784413), TEX1 (PRJNA788214, PRJNA788213, PRJNA780104), DOR1
616 (PRJNA788571, PRJNA788570, PRJNA782680), KER1 (PRJNA788676,
617 PRJNA788675, PRJNA783516), MER1 (PRJNA788672, PRJNA788671,
618 PRJNA783537), KAZ1 (PRJNA788564, PRJNA788563, PRJNA783179), FRI1
619 (PRJNA791176, PRJNA791175, PRJNA721520). The details are provided in
620 Supplementary Table 1.
621
622

623

624 **References:**

- 625 1. Chiang C, *et al.* The impact of structural variation on human gene expression. *Nat*
626 *Genet* **49**, 692-699 (2017).
- 627 2. Sudmant PH, *et al.* An integrated map of structural variation in 2,504 human
628 genomes. *Nature* **526**, 75-81 (2015).
- 629 3. Mahmoud M, Gobet N, Cruz-Dávalos DI, Mounier N, Dessimoz C, Sedlazeck FJ.
630 Structural variant calling: the long and the short of it. *Genome Biol* **20**, 246 (2019).
- 631 4. Ho SS, Urban AE, Mills RE. Structural variation in the sequencing era. *Nat Rev Genet*
632 **21**, 171-189 (2020).
- 633 5. Sedlazeck FJ, Lee H, Darby CA, Schatz MC. Piercing the dark matter: bioinformatics
634 of long-range sequencing and mapping. *Nat Rev Genet* **19**, 329-346 (2018).
- 635 6. Audano PA, *et al.* Characterizing the Major Structural Variant Alleles of the Human
636 Genome. *Cell* **176**, 663-675.e619 (2019).
- 637 7. Ebert P, *et al.* Haplotype-resolved diverse human genomes and integrated analysis of
638 structural variation. *Science*, eabf7117 (2021).
- 639 8. Qin P, *et al.* Pan-genome analysis of 33 genetically diverse rice accessions reveals
640 hidden genomic variations. *Cell* **184**, 3542-3558.e3516 (2021).
- 641 9. Liu Y, *et al.* Pan-Genome of Wild and Cultivated Soybeans. *Cell* **182**, 162-176.e113
642 (2020).
- 643 10. Alonge M, *et al.* Major Impacts of Widespread Structural Variation on Gene
644 Expression and Crop Improvement in Tomato. *Cell* **182**, 145-161.e123 (2020).
- 645 11. Beyter D, *et al.* Long-read sequencing of 3,622 Icelanders provides insight into the

646 role of structural variants in human diseases and other traits. *Nat Genet* **53**, 779-786
647 (2021).

648 12. Golicz AA, Bayer PE, Bhalla PL, Batley J, Edwards D. Pangenomics Comes of Age:
649 From Bacteria to Plant and Animal Applications. *Trends Genet* **36**, 132-145 (2020).

650 13. Talenti A, *et al.* A cattle graph genome incorporating global breed diversity. *bioRxiv*,
651 2021.2006.2023.449389 (2021).

652 14. Miga KH, Wang T. The need for a human pangenome reference sequence. *Annual*
653 *Review of Genomics and Human Genetics* **22**, 81-102 (2021).

654 15. Crysanto D, Leonard AS, Fang Z-H, Pausch H. Novel functional sequences
655 uncovered through a bovine multiassembly graph. *Proc Natl Acad Sci* **118**,
656 e2101056118 (2021).

657 16. Leonard AS, *et al.* Bovine pangenome reveals trait-associated structural variation
658 from diverse assembly inputs. *bioRxiv*, 2021.2011.2002.466900 (2021).

659 17. De Coster W, Van Broeckhoven C. Newest methods for detecting structural variations.
660 *Trends Biotechnol* **37**, 973-982 (2019).

661 18. Cheng H, Concepcion GT, Feng X, Zhang H, Li H. Haplotype-resolved de novo
662 assembly using phased assembly graphs with hifiasm. *Nat Methods* **18**, 170-175
663 (2021).

664 19. Nurk S, *et al.* HiCanu: accurate assembly of segmental duplications, satellites, and
665 allelic variants from high-fidelity long reads. *Genome Res* **30**, 1291-1305 (2020).

666 20. Norris BJ, Whan VA. A gene duplication affecting expression of the ovine ASIP gene
667 is responsible for white and black sheep. *Genome Res* **18**, 1282-1293 (2008).

- 668 21. Demars J, *et al.* Genome-Wide Identification of the Mutation Underlying Fleece
669 Variation and Discriminating Ancestral Hairy Species from Modern Woolly Sheep. *Mol*
670 *Biol Evol***34**, 1722-1729 (2017).
- 671 22. Wiedemar N, Drögemüller C. A 1.8-kb insertion in the 3'-UTR of RXFP2 is associated
672 with polledness in sheep. *Anim Genet***46**, 457-461 (2015).
- 673 23. Li X, *et al.* Whole-genome resequencing of wild and domestic sheep identifies genes
674 associated with morphological and agronomic traits. *Nat Commun* **11**, 2815 (2020).
- 675 24. Yang J, *et al.* Whole-genome sequencing of native sheep provides insights into rapid
676 adaptations to extreme environments. *Mol Biol Evol***33**, 2576-2592 (2016).
- 677 25. Lv F-H, *et al.* Whole-genome resequencing of worldwide wild and domestic sheep
678 elucidates genetic diversity, introgression and agronomically important loci. *Mol Biol*
679 *Evol*, (2021).
- 680 26. Li H, Feng X, Chu C. The design and construction of reference pangenome graphs
681 with minigraph. *Genome Biol***21**, 265 (2020).
- 682 27. Crysnanto D, Pausch H. Bovine breed-specific augmented reference graphs facilitate
683 accurate sequence read mapping and unbiased variant discovery. *Genome Biol* **21**,
684 184 (2020).
- 685 28. Durbin RM, *et al.* A map of human genome variation from population-scale
686 sequencing. *Nature***467**, 1061-1073 (2010).
- 687 29. Quan C, *et al.* Characterization of structural variation in Tibetans reveals new
688 evidence of high-altitude adaptation and introgression. *Genome Biol***22**, 159 (2021).
- 689 30. Gonzalez MV, *et al.* A Divergent Artiodactyl MYADM-like Repeat Is Associated with

690 Erythrocyte Traits and Weight of Lamb Weaned in Domestic Sheep. *PLoS One* **8**,
691 e74700 (2013).

692 31. French AT, *et al.* The expression of intelectin in sheep goblet cells and upregulation
693 by interleukin-4. *Vet Immunol Immunopathol* **120**, 41-46 (2007).

694 32. Chen S, *et al.* Paragraph: a graph-based structural variant genotyper for short-read
695 sequence data. *Genome Biol* **20**, 291 (2019).

696 33. Yan SM, *et al.* Local adaptation and archaic introgression shape global diversity at
697 human structural variant loci. *eLife* **10**, e67615 (2021).

698 34. Wang F, *et al.* Allele-specific expression and splicing provides insight into the
699 phenotypic differences between thin- and fat-tailed sheep breeds. *Journal of Genetics*
700 *and Genomics*, (2021).

701 35. Kalkan Z, Durasi İM, Sezerman U, Atasever-Arslan B. Potential of GRID2 receptor
702 gene for preventing TNF-induced neurodegeneration in autism. *Neurosci Lett* **620**,
703 62-69 (2016).

704 36. Cheng JY, Stern AJ, Racimo F, Nielsen R. Detecting Selection in Multiple
705 Populations by Modeling Ancestral Admixture Components. *Mol Biol Evol*, (2021).

706 37. Zhu X-J, *et al.* BMP-FGF signaling axis mediates Wnt-induced epidermal stratification
707 in developing mammalian skin. *PLoS Genet* **10**, e1004687 (2014).

708 38. Kardos M, *et al.* Whole-genome resequencing uncovers molecular signatures of
709 natural and sexual selection in wild bighorn sheep. *Mol Ecol* **24**, 5616-5632 (2015).

710 39. Pan Z, *et al.* Rapid evolution of a retro-transposable hotspot of ovine genome
711 underlies the alteration of BMP2 expression and development of fat tails. *BMC*

- 712 *Genomics* **20**, 1-15 (2019).
- 713 40. Nishikawa S, *et al.* In utero manipulation of coat color formation by a monoclonal anti-
714 c - kit antibody: two distinct waves of c - kit - dependency during melanocyte
715 development. *The EMBO journal* **10**, 2111-2118 (1991).
- 716 41. Yang N, *et al.* Genome assembly of a tropical maize inbred line provides insights into
717 structural variation and crop improvement. *Nat Genet* **51**, 1052-1059 (2019).
- 718 42. Võsa U, *et al.* Large-scale cis- and trans-eQTL analyses identify thousands of genetic
719 loci and polygenic scores that regulate blood gene expression. *Nat Genet* **53**, 1300-
720 1310 (2021).
- 721 43. Pan X, *et al.* Modes of genetic adaptations underlying functional innovations in the
722 rumen. *Science China Life Sciences* **64**, 1-21 (2021).
- 723 44. Kalds P, Luo Q, Sun K, Zhou S, Chen Y, Wang X. Trends towards revealing the
724 genetic architecture of sheep tail patterning: Promising genes and investigatory
725 pathways. *Anim Genet* **52**, 799-812 (2021).
- 726 45. Moioli B, Pilla F, Ciani E. Signatures of selection identify loci associated with fat tail in
727 sheep1. *J Anim Sci* **93**, 4660-4669 (2015).
- 728 46. Xu S-S, *et al.* Genome-wide association analysis identifies the genetic basis of fat
729 deposition in the tails of sheep (*Ovis aries*). *Anim Genet* **48**, 560-569 (2017).
- 730 47. Fu W, Wang R, Nanaei HA, Wang J, Hu D, Jiang Y. RGD v2.0: a major update of the
731 ruminant functional and evolutionary genomics database. *Nucleic Acids Res*,
732 gkab887 (2021).
- 733 48. Clark EL, *et al.* A high resolution atlas of gene expression in the domestic sheep

- 734 (Ovis aries). *PLoS Genet* **13**, e1006997 (2017).
- 735 49. Chao T, *et al.* Transcriptome Analysis of Three Sheep Intestinal Regions reveals Key
736 Pathways and Hub Regulatory Genes of Large Intestinal Lipid Metabolism. *Scientific*
737 *Reports* **7**, 5345 (2017).
- 738 50. Jiang Y, *et al.* The sheep genome illuminates biology of the rumen and lipid
739 metabolism. *Science* **344**, 1168-1173 (2014).
- 740 51. Economides KD, Zeltser L, Capecchi MR. HOXB13 mutations cause overgrowth of
741 caudal spinal cord and tail vertebrae. *Dev Biol* **256**, 317-330 (2003).
- 742 52. Aires R, *et al.* Tail Bud Progenitor Activity Relies on a Network Comprising Gdf11,
743 Lin28, and Hox13 Genes. *Dev Cell* **48**, 383-395.e388 (2019).
- 744 53. Pan Z, *et al.* Rapid evolution of a retro-transposable hotspot of ovine genome
745 underlies the alteration of BMP2 expression and development of fat tails. *BMC*
746 *Genomics* **20**, 261 (2019).
- 747 54. Baazaoui I, Bedhiaf-Romdhani S, Mastrangelo S, Ciani E. Genome-wide analyses
748 reveal population structure and identify candidate genes associated with tail fatness
749 in local sheep from a semi-arid area. *Animal* **15**, 100193 (2021).
- 750 55. Dong K, *et al.* Genomic analysis of worldwide sheep breeds reveals PDGFD as a
751 major target of fat-tail selection in sheep. *BMC Genomics* **21**, 1-12 (2020).
- 752 56. Hickey G, *et al.* Genotyping structural variants in pangenome graphs using the vg
753 toolkit. *Genome Biol* **21**, 35 (2020).
- 754 57. Sirén J, *et al.* Pangenomics enables genotyping of known structural variants in
755 diverse genomes. *Science* **374**, abg8871 (2021).

- 756 58. Li R, *et al.* A near complete genome for goat genetic and genomic research. *Genet*
757 *Sel Evol* **53**, 74 (2021).
- 758 59. Alonge M, *et al.* RaGOO: fast and accurate reference-guided scaffolding of draft
759 genomes. *Genome Biol* **20**, 224 (2019).
- 760 60. Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. BUSCO Update: Novel
761 and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage
762 for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Mol Biol Evol* **38**, 4647-
763 4654 (2021).
- 764 61. Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B. AUGUSTUS: ab
765 initio prediction of alternative transcripts. *Nucleic Acids Res* **34**, W435-W439 (2006).
- 766 62. Buchfink B, Reuter K, Drost H-G. Sensitive protein alignments at tree-of-life scale
767 using DIAMOND. *Nat Methods* **18**, 366-368 (2021).
- 768 63. Benson G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic*
769 *Acids Res* **27**, 573-580 (1999).
- 770 64. Bakker B, *et al.* Single-cell sequencing reveals karyotype heterogeneity in murine and
771 human malignancies. *Genome Biol* **17**, 115 (2016).
- 772 65. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic
773 features. *Bioinformatics* **26**, 841-842 (2010).
- 774 66. Narasimhan V, Danecek P, Scally A, Xue Y, Tyler-Smith C, Durbin R. BCFtools/RoH:
775 a hidden Markov model approach for detecting autozygosity from next-generation
776 sequencing data. *Bioinformatics* **32**, 1749-1751 (2016).
- 777 67. Danecek P, *et al.* The variant call format and VCFtools. *Bioinformatics (Oxford,*

778 *England*) **27**, 2156-2158 (2011).

779 68. Kijas JW, *et al.* Genome-wide analysis of the world's sheep breeds reveals high
780 levels of historic mixture and strong recent selection. *PLoS Biol* **10**, e1001258 (2012).

781 69. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina
782 sequence data. *Bioinformatics* **30**, 2114-2120 (2014).

783 70. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory
784 requirements. *Nat Methods* **12**, 357-360 (2015).

785 71. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for
786 assigning sequence reads to genomic features. *Bioinformatics* **30**, 923-930 (2013).

787 72. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for
788 differential expression analysis of digital gene expression data. *Bioinformatics* **26**,
789 139-140 (2010).

790 73. Shabalin AA. Matrix eQTL: ultra fast eQTL analysis via large matrix operations.
791 *Bioinformatics* **28**, 1353-1358 (2012).

792

793

794 Tables

795 **Table 1** Quality assessments of the primary and partially phased assemblies for the 13

796 individuals. For each sheep, we show the values for the primary assembly and for

797 both partially phased haploid assemblies.

798

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 Friesian	FRI1	Primary	2.90	85.1	982	28.0	94.4	48.2
		hap1/hap2	2.77/2.75	62.9/53.2		-	91.3/92.8	48.4/48.5
Chinese Merino	MER1	Primary	2.84	60.0	1773	18.7	94.2	43.3
		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 Dorper	WDU1	Primary	2.81	17.8	2125	15.4	93.8	41.7
		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

799

800 **Figure legends**

801 **Fig. 1** Ovine pan-genome graph construction. (a) PCA plot showing the representation
802 of the genetic diversity of domestic sheep by the 13 individuals used for HiFi
803 sequencing and pan-genome construction. (b) The pan-genome size changes with
804 increase in the number of genome assemblies. (c) Repeat annotation of the structural
805 variations stratified by SV types. (d) Length distribution of different type of SVs. (e)
806 Diagram and number of different SV types. The red bold lines indicate the reference
807 sequence and the blue bold lines represent the non-reference. (f) An example of a
808 divergent allele of 50,757 bp containing an MYADM-like repeat. (g) An example of a
809 multiallelic variation containing a large deletion of 105,650 bp which is nested with
810 another insertion of 5,518 bp. The red lines with arrows show the reference path and
811 the blue lines with arrows refer to the non-reference path.

812 **Fig. 2** The repertoire of structural variations in sheep genome and frequency spectrum
813 in domestic and wild sheep. (a) Geographical distribution of the 45 breeds or
814 populations (see Supplementary Data 7 for the assignment of the breeds or
815 populations). The Dorper sheep from South Africa and the white Suffolk from
816 Australia are not shown on the map. (b) Histogram of the number of SV ranks (0-300)
817 of r^2 values that surpass the median SNP-based r^2 value for common SVs
818 (MAF>0.05). (c) Derived allele frequencies of the insertions and deletions. (d) The
819 proportion of standing and *de novo* SVs. (e) High-frequency *de novo* SVs in domestic
820 sheep. (f) A *de novo* insertion within *GRID2* occurs at high frequencies in all studied
821 populations. (g) A *de novo* insertion of a retrogene occurs at high frequencies in

822 European sheep.

823 **Fig. 3** Population-stratified SVs and their effects on gene expression. (a) Genome-
824 wide distribution of global F_{ST} which is measured by the average value of the top 50%
825 F_{ST} values for each SV across 33 sheep breeds or populations. The red dots represent
826 population-stratified SVs that were also found by Ohana approach. (b) Examples of
827 expression-associated SVs identified by allelic specific expression (ASE) mapping.
828 Each column represents one ASE SNP in exons and each row represents one RNA-seq
829 sample. (c) Allele distributions of the three expression-associated SVs in diverse
830 breeds or populations.

831 **Fig. 4** Selective SVs associated with the long tail trait. (a) The genome-wide Fixation
832 index (F_{ST}) of long fat-tailed versus short fat-tailed Asian sheep breeds and long thin-
833 tailed versus short thin-tailed European sheep breeds. (b) The selective haplotype in
834 long-tailed sheep. The 169 bp-insertion is located upstream of the 5' UTR region of
835 *HOXB13*. The haplotypes include the genotypes of the insertion and surrounding
836 SNPs. (c) The frequency of the insertion in wild sheep and domestic sheep. (d) RNA-
837 seq data shows that the expression of the insertion. The top panel corresponds to the
838 expression in the colon of Texel × long-tailed Scottish Blackface sheep
839 (ERS2107126); the bottom panel corresponds to the expression in colon of one small-
840 tail sheep (SRR5122597) without the insertion. (e) The electrophoresis diagrams of
841 PCR products for the three genotypes (homologous insertion: II, heterozygote: ID,
842 homologous deletion: DD). (f) The carriers and non-carriers of the insertion differ in
843 tail length. (g) *HOXB13* promoter activity is modulated by the 169 bp insertion.

844 Luciferase activity in HEK293T cells transfected with recombinant plasmids
845 containing the insertion or not. Experiments were repeated at least five times. The
846 data are presented as mean $\pm$ standard error of the normalized luciferase activity.
847 **Fig. 5** Selective SVs associated with the fat-tail trait. (a) Population branch statistic
848 (PBS) values across the whole genome by comparing fat-tailed sheep to thin-tailed
849 sheep using the mouflon sheep as an outgroup. (b) The two most differentiated
850 regions between fat-tailed sheep and thin-tailed sheep. The left panel shows the
851 intergenic region between *BMP2* and *HAOI* while the right panel corresponds to the
852 region surrounding *PDGFD*. SVs with PBS value >0.8 are highlighted by blue dotted
853 lines. (c) The haplotypes of mouflons and domestic sheep for the two most selective
854 regions. Each column represents one SV and each row represent one individual. The
855 black reverted triangles represent *de novo* SVs.

856

857 **Supplementary Files**

858 **Supplementary Table 1** Accession numbers of the *de novo* assemblies.

859

860 **Supplementary Data**

861 **Supplementary Data 1** Information of divergent variations.

862 **Supplementary Data 2** Information of multiallelic variations.

863 **Supplementary Data 3** List of the biallelic insertions and deletions identified in this
864 study.

865 **Supplementary Data 4** SV frequency in 45 sheep breeds or populations.

866 **Supplementary Data 5** List of population-stratified SVs.

867 **Supplementary Data 6** List of putatively expression-associated SVs identified by

868 ASE mapping.

869 **Supplementary data 7** List of putatively expression-associated SVs identified by
870 eQTL mapping.

871 **Supplementary Data 8** Sample list of the resequencing data.

872

873 **Supplementary figures**

874 **Supplementary Fig. 1** Example of three representative divergent variations.

875 intersecting with IDE (a), PRP3-like (b) and MAGI2 (c), respectively.

876 **Supplementary Fig. 2** Example of three representative multiallelic variations

877 affecting ASIP (a), UDP-glucuronosyltransferase 2B18-like (b) and olfactory receptor
878 (LOC101115700) (c), respectively.

879 **Supplementary Fig. 3** The number of SVs intersecting functional elements compared
880 to randomly permuting SV locations.

881 **Supplementary Fig. 4** An ideogram showing SV hotspots using all SVs. The total
882 number of SVs in each detected hotspot is shown by a scale going from blue to red.

883 **Supplementary Fig. 5** Enrichment analysis of SVs with respect to the last 5 Mb of
884 each chromosome end.

885 **Supplementary Fig. 6** PCA plot showing the clusters of domestic sheep by insertions
886 (a) and deletions (b).

887 **Supplementary Fig. 7** Enriched KEGG pathways for the population stratified SVs.

888 **Supplementary Fig. 8** RNA-seq mapping for the 169 bp insertion shows that the
889 insertion belongs to 5' UTR region.

890

a

b

c