ARTICLE

Diversity of copy number variation in the worldwide goat population

Mei Liu[1,2] · Yang Zhou[3] · Benjamin D. Rosen [1] · Curtis P. Van Tassell[1] · Alessandra Stella[4] ·
Gwenola Tosser-Klopp [5] · Rachel Rupp[5] · Isabelle Palhière[5] · Licia Colli [6] · Brian Sayre[7] · Paola Crepaldi[8] ·
Lingzhao Fang[1] · Gábor Mészáros[9] · Hong Chen[2] · George E. Liu[1] · the ADAPTmap Consortium

Abstract

Goats (*Capra hircus*) are an important farm animal species. Copy number variation (CNV) represents a major source of genomic structural variation. We investigated the diversity of CNV distribution in goats using CaprineSNP50 genotyping data generated by the ADAPTmap Project. We identified 6286 putative CNVs in 1023 samples from 50 goat breeds using PennCNV. These CNVs were merged into 978 CNV regions, spanning ~262 Mb of total length and corresponding to ~8.96% of the goat genome. We then divided the samples into six subgroups per geographic distribution and constructed a comparative CNV map. Our results revealed a population differentiation in CNV across different geographical areas, including Western Asia, Eastern Mediterranean, Alpine & Northern Europe, Madagascar, Northwestern Africa, and Southeastern Africa groups. The results of a cluster heatmap analysis based on the CNV count per individual across different groups was generally consistent with the one generated from the SNP data, likely reflecting the population history of different goat breeds. We sought to determine the gene content of these CNV events and found several important CNV-overlapping genes (e.g. *EDNRA*, *ADAMTS20*, *ASIP*, *KDM5B*, *ADAM8*, *DGAT1*, *CHRNB1*, *CLCN7*, and *EXOSC4*), which are involved in local adaptations such as coat color, muscle development, metabolic processes, osteopetrosis, and embryonic development. Therefore, this research generated an extensive CNV map in the worldwide population of goat, which offers novel insight into the goat genome and its functional annotation.

Introduction

The goat (*Capra hircus*) is one of the major livestock species to be domesticated. Its domestication is estimated to have first occurred more than 10,000 years ago in southeast Asia and to have involved at least four distinct domestication events (Naderi et al. 2008). The phenotypic and genetic variability that exists between goat breeds is mainly due to

These authors contributed equally: Mei Liu and Yang Zhou

Electronic supplementary material The online version of this article (https://doi.org/10.1038/s41437-018-0150-6) contains supplementary material, which is available to authorized users.

✉ George E. Liu
George.Liu@ars.usda.gov

1   Animal Genomics and Improvement Laboratory, BARC, Agricultural Research Service, USDA, Beltsville, MD 20705, USA

2   College of Animal Science and Technology, Shaanxi Key Laboratory of Agricultural Molecular Biology, Northwest A&F University, Yangling, Shaanxi 712100, China

3   Key Laboratory of Agricultural Animal Genetics, Breeding and Reproduction, Education Ministry of China, Huazhong Agricultural University, Wuhan, Hubei 430070, China

4   Parco Tecnologico Padano, Lodi, Italy

5   INRA, UMR444, Laboratoire de Génétique Cellulaire, Castanet-Tolosan, France

6   Istituto di Zootecnica e Centro di Ricerca sulla Biodiversità e sul DNA Antico—BioDNA, Università Cattolica del Sacro Cuore di Piacenza, Piacenza, Italy

7   Virginia State University, Petersburg, Virginia, USA

8   Dipartimento di Medicina Veterinaria, Università di Milano, Via Celoria 10, Milano, Italy

9   Division of Livestock Sciences, University of Natural Resources and Life Sciences, Vienna, Austria

adaptation and artificial selection for animal production including meat, wool, and milk (Diamond 2002). During the last decade, high-throughput single nucleotide polymorphism (SNP) arrays have facilitated population genetics studies to improve our understanding of the genetic mechanisms underlying complex economic and adaptive traits in domesticated animals (Goddard and Hayes 2009; Wang et al. 2014). Compared with the other major livestock species, the goat was one of the last for which medium-density SNP chips became available. In 2012, the first medium-density Illumina CaprineSNP50 BeadChip was designed and released by the international goat genome consortium (Tosser-Klopp et al. 2014). Lashmar et al. (2015) has validated the Illumina CaprineSNP50 BeadChip in the South African Angora population. With this SNP chip, Talenti et al. (2017) has performed a genome-wide scan for the genomic signatures left by selection in the Valdostana goat breed and found evidence of selective sweep regions on three different chromosomes which contained genes involved in the immune system development and regulation. Recently, the ADAPTmap consortium, an international effort for coordination among projects for genotyping and re-sequencing of goat breeds, has been actively working on using this SNP array to understand genetic diversity, population history admixture, selection signatures, and other features in the worldwide goat populations (Colli et al. under review). Based on the ADAPTmap goat dataset, several signatures of selection in different chromosomic regions have been detected across the different breeds, sub-geographical clusters, phenotypic and climatic groups using SNP data (Bertolini et al. under review).

However, genome function and evolution are not limited to SNP. For example, copy number variation (CNV), gains and losses of large regions of genomic sequence ranging from 50 bp to 5 Mb between individuals of a species, is one of the widely dispersed forms of structural variations in mammalian genomes and independently contributes to phenotypic diversity and disease susceptibility (Conrad et al. 2010; Feuk et al. 2006; Zhang et al. 2009). Copy number variable genes are a well-established cause of gene family member differentiation and a common mechanism underpinning evolutionary change. Recently, studies of differentiated CNV in humans (Narang et al. 2014; Zarrei et al. 2015) and domesticated animals like cattle (Xu et al. 2016), sheep (Yang et al. 2017) and dog (Chen et al. 2008; Alvarez and Akey 2012) have suggested that CNV is a good genetic marker for population genetics analysis and group-differential selection signatures. Therefore, CNV-based population genomic analysis in the worldwide goat population, which is lacking till now, could offer new insights into the genomic architecture of goats and facilitate our understanding of the evolution and subsequent selection within the goat genome.

Up to now, only a few studies on goat CNV have been reported. Before the goat genome was sequenced, Fontanesi et al. (2010) made a preliminary attempt to detect goat CNV in four European breeds using a low-density Comparative Genomic Hybridization array (aCGH) which was designed from the cattle Btau_4.0 genome assembly. However, there is no published study reporting a comprehensive CNV map in the worldwide goat population, and the features of goat CNV at the population level are not well understood.

In this study, we aimed to investigate CNV in the worldwide goat populations using the dataset generated by the ADAPTmap Project. We utilized PennCNV to detect CNV from the CaprineSNP50 array genotyping data, and then performed CNV-based variance analyses. With a large-scale goat CNV map, our results provide extensive CNV information and potential candidates for further exploration on the roles of CNV underlying important traits and evolutionary adaptation in goat.

Methods

Selecting populations and animals

Data from 50 breeds and 1023 animals were retrieved from the Illumina CaprineSNP50 BeadChip array genotyping data, which has been collected by the ADAPTmap Project. This goat CNV study data collection was generated after removing mixed and hybrid breeds and related animals, which were identified by the Treemix analysis (Pickrell and Pritchard 2012). The division of subgroups were performed based on geographical areas and the genome structure analysis, which was assessed by the maximum likelihood based on the approach implemented in Admixture software v1.3.0 (Alexander et al. 2009) using SNP data (Colli et al. under review). The number of animals per population and geographic origin of breed development are described in Supplementary Table S1.

CNV detection with PennCNV

The Illumina CaprineSNP50 BeadChip (Tosser-Klopp et al. 2014) including 53,347 SNP markers was used to genotype the animals. Original marker positions were remapped on the new goat reference sequence ARS1 or ASM170441v1 (Bickhart et al. 2017). After removing SNPs, which were unmapped, mapped to sexual chromosomes or with low call rate ($<95\%$), a total of 51,057 autosomal markers were selected for CNV analysis. Signal intensity ratios (log R Ratio: LRR) and allelic frequencies (B allele frequency: BAF) were retrieved using Illumina GenomeStudio1.0 software for each SNP. The population frequency of B allele (PFB) file was calculated based on the

BAF of each marker in each population. The goat GC model file was generated using cal_gc_snp.pl with default settings (http://penncnv.openbioinformatics.org/en/latest/misc/faq/). The goat HapMap population was used as a reference, which included 53 Alpine, 26 Angora, 30 Boer, 38 Creole, 16 Jinlan, 15 Katjang, 59 Saanen, 20 Savanna, 27 Skopelos, and 1 Yunling goat, and four parent–parent–child trios (Tosser-Klopp et al. 2014). CNVs were inferred within each animal using PennCNV software which was based on the hidden Markov model (Wang et al. 2007). CNVs were required to span at least three probes. PennCNV quality filters were subsequently applied as follows: we kept high-quality samples with a standard deviation (SD) of LRR < 0.30, BAF drift < 0.01, and waviness factor value between −0.05 and 0.05. Appropriate LRR adjustments based on the GC model were incorporated in PennCNV. Similar to previous studies (Zhou et al. 2016; Ma et al. 2017), CNV regions (CNVRs) were obtained through merging overlapping CNVs, which were generated from all the samples, by at least 1 bp overlap. Then CNVRs were defined as three types (loss, gain, and both) according to the composed CNV types (loss and/or gain).

Population differential analysis

To explore group-differential CNVRs, we first divided the goat samples into six groups according to their geographic distribution and then constructed a comparative CNVR map across these groups. The geographical distribution of the data-set was the following: 53 animals (3 breeds) from Western Asia (WAS), 195 animals (11 breeds) from Eastern Mediterranean (EME), 103 animals (4 breeds) from Alpine & Northern Europe (ANE), 78 animals (5 breeds) from Madagascar (MAD), 182 animals (7 breeds) from Northwestern Africa (NWA), and 412 animals (20 breeds) from Southeastern Africa (SEA). In total, the CNV event frequency (defined as CNV count per individual, i.e. CNV count in each CNVR normalized by sample size) was determined and utilized as the CNV characteristics for comparison among six groups. To explore the potential differences involved with selection pressure for CNVRs, we estimated the CNV event frequency per group and the variance across all six groups. Based on the frequency across six groups, heatmap cluster analysis was performed by calculating Euclidean distances of 26 CNVRs, whose variances were ranked the top 5% among 526 CNVRs (CNV events > = 2).

Gene annotation and PANTHER analysis

Genic content of goat CNV regions was screened using the annotation release 102 for the goat assembly ARS1 (RefSeq assembly accession: GCF_001704415.1, ftp://ftp.ncbi.nlm.

nih.gov/genomes/all/GCF/001/704/415/ GCF_001704415.1_ARS1). Variants were annotated using ANNOVAR software, which identified the variants in intronic, exonic, intergenic, 5′/3′-UTR, splicing site, and upstream/downstream (less than a threshold away from a transcript, by default 1 kb) (Wang et al. 2010). Since the goat's gene list was not available, we performed gene ontology (GO) enrichment analysis with the bovine gene list using PANTHER. We only considered terms with gene count more than five and P-value < 0.05, after the Bonferroni correction for multiple testing.

To explore gene-containing CNV's potential functional impact, we queried Online Mendelian Inheritance in Animals (OMIA, http://omia.angis.org.au/home/) database to find genes which are associated with the inherited disorders and/or other traits.

CNV validation by quantitative PCR (qPCR)

We performed SYBR Green real-time qPCR on bioRad MyIQ thermocycler to test the accuracy for CNV call in this study. Based on CNV call results by PennCNV, the CNVs located in six functional genes (*AHCY, ASIP, ITCH, EDNRA, NR3C2*, and *DGAT1*) were selected for qPCR validation using several African goat samples with different copy number status (normal, gain or loss type). Primers for qPCR were designed by the NCBI Primer-BLAST webtool (http://www.ncbi.nlm.nih.gov/tools/primer-blast/index.cgi? LINK_LOC = BlastHome). Primers information is shown in Supplementary Table S2. The relative copy numbers of these genes were determined by the $2*2^{-\Delta\Delta CT}$ method using qPCR. The *MC1R* gene was used as the reference gene for all qPCR experiments (Fontanesi et al. 2009). ΔCT from all samples were normalized to control group, six randomly selected European samples (three Saanen and three Alpine goats). All reactions of 25 µL were amplified in triplicate. The PCR procedure was as follows: initial denaturation for 5 min at 95 °C; followed by 40 cycles of 94 °C for 30 s, annealing at 60 °C for 25 s, and primer extension at 72 °C for 30 s. Melting curve was completed at the end of the amplification with the conditions: a cycle of 95 °C for 1 min and then 1 cycle of 55 °C for 1 min followed by an increase rate of 0.5 °C/cycle from 55 to 95 °C.

Results

Identification of CNVs in the worldwide goat populations

Using PennCNV, we performed large-scale CNV screens in 2328 goats derived from the ADAPTmap Project dataset. After quality filtering, a total of 6286 CNV events for 2632

Table 1 Summary of CNVs and CNVRs identified from the goat ADAPTmap populations

Group	CNV							CNVR		
	Unique	Count	Avg.[a]	Gain/ Avg.[a]	Loss/ Avg.[a]	Length/ Avg.[a] (kb)	SD[b] (kb)	Count[c]	Length/ Avg.[a] (kb)	SD[b] (kb)
WAS	152	225	4	71/1	154/3	38,968/735	157	112	20,566/184	152
EME	794	1369	7	737/4	632/3	273,106/1401	197	401	91,449/228	251
ANE	245	341	3	98/1	243/2	56,351/548	140	192	39,668/207	183
MAD	421	647	8	106/1	541/7	141,478/1814	196	244	64,345/264	257
NWA	880	1750	10	251/1	1499/8	371,023/2039	183	355	91,611/2581	267
SEA	868	1954	5	683/2	1271/3	379,503/921	195	505	115,129/228	247
Total	2632	6286	6	1946/2	4340/4	1,260,428/ 1232	189	978	262,000/268	302

Samples were divided into Western Asia (WAS), Eastern Mediterranean (EME), Alpine & Northern Europe (ANE), Madagascar (MAD), Northwestern Africa (NWA,) and Southeastern Africa (SEA) groups. CNVRs in this table represent non-redundant CNVR counts after merging both gain and loss CNVs identified within each group

Note: [a]*Avg.* average. The average values are divided or normalized by sample counts except that the lengths are average lengths normalized by CNV and CNVR counts. $N_{WAS} = 53$, $N_{EME} = 195$, $N_{ANE} = 103$, $N_{MAD} = 78$, $N_{NWA} = 182$, $N_{SEA} = 412$

[b]*SD* standard deviation

[c]These numbers are non-redundant CNVR counts

unique CNVs were identified among 1023 individuals from 50 breeds. The cumulative length of all 978 merged CNVRs was ~262 Mb (262,000,202 bp), with an average length per CNVR 267,893 bp. For unique CNVs, some were detected in multiple populations while some others were specifically detected in one population. The CNV events were used to calculate the average CNV/CNVR frequencies and length in each group (Table 1). The average CNV count in all test samples was six. The highest CNV frequency was observed in NWA while the lowest frequency was in ANE. EME contained more gain CNVs than other groups. Notably, the average CNVR length in NWA was almost ten folds longer than those in other four groups. Detailed information on CNV event counts for all breeds is shown in Supplementary Table S1. We also estimated the CNV densities (CNVR length in each breed normalized by the goat genome size, which is 2,922,813 kb) for each goat group, finding varying CNV densities across six groups. For instance, the SEA group had the highest CNV density of 3.94%. Both the NWA and EME groups had ~3.13%. The WAS group had the lowest CNV density of 0.70%.

To explore the geographic pattern of CNV across the six groups, we constructed violin plots using the CNV length and the number of SNPs within CNVs. The patterns showed slight difference across groups but similar distributions for CNV length and SNP count covered by each CNV per animal (Supplementary Figure S1). Both the longest CNV and the most SNP count within CNVs were observed in the SEA group while the shortest CNV and the least SNP were in the WAS group. Across 29 pairs of goat autosomes, we found a varying distribution of CNV counts (Fig. 1). On
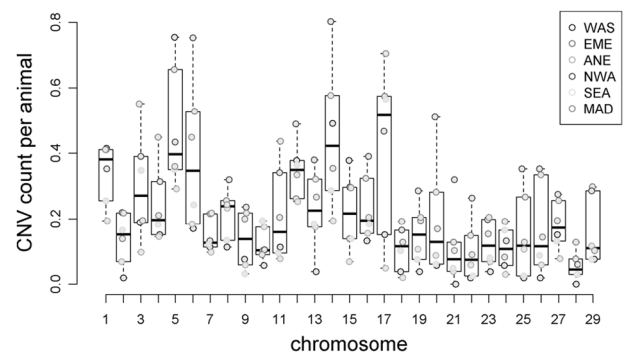

Fig. 1 Boxplot for distributions of CNVs count per autosome across six goat groups. WAS: Western Asia, $n = 53$; EME: Eastern Mediterranean, $n = 195$; ANE: Alpine & Northern Europe, $n = 103$; NWA: Northwestern Africa, $n = 182$; SEA: Southeastern Africa, $n = 412$; MAD: Madagascar, $n = 78$. Each "box" contains a thin horizontal line (25–75% range) and a horizontal black line (median)

chr1, 5, 6, 12, 14, and 17, the CNV distribution showed a high level with at least 0.35 CNV count per animal. On chr21, 22, and 28, the CNV count per animal was lower than 0.10 (median value). In addition, clear variance for CNV count across different groups were observed on chr5 (0.04), 6 (0.05), 14 (0.05), 17 (0.07), 20 (0.03).

Group-differential CNVRs

To compare the frequency of CNV across different groups, we first merged CNV for each animal into non-redundant CNVR within groups. This identified 112, 401, 192, 244, 355, and 505 CNVRs in WAS, EME, ANE, MAD, NWA, and SEA groups, respectively, with the responding CNVR
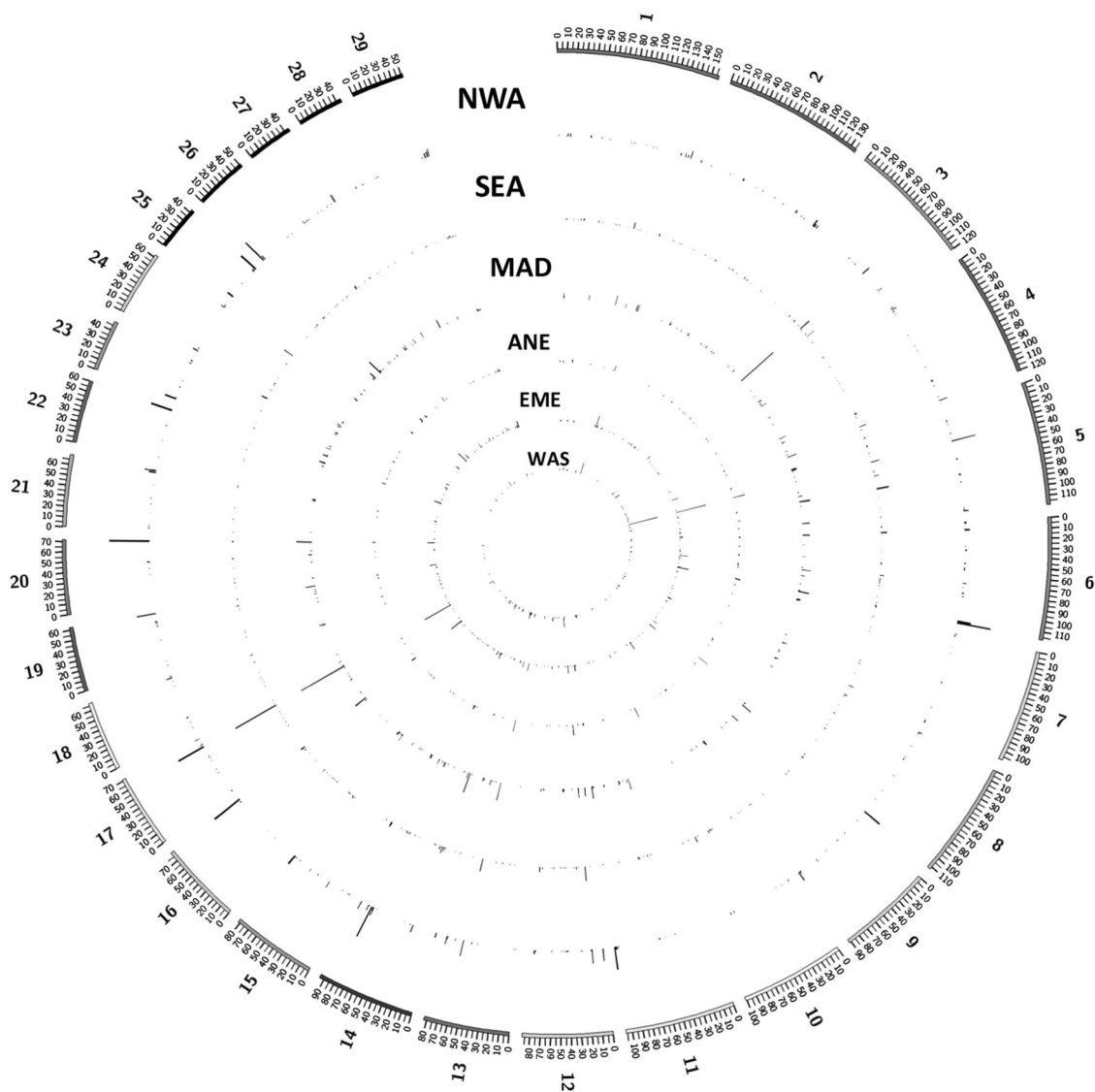
Fig. 2 Genomic distribution and frequencies of CNV regions on autosomes in six groups from the goat ADAPTmap data-set. The tracks from outside to inside are: Northwestern Africa (NWA), Southeastern Africa (SEA), Madagascar (MAD), Alpine & Northern Europe (ANE), Eastern Mediterranean (EME), Western Asia (WAS), respectively. The frequency of CNVR was plotted for the subgroups and only CNVRs contained > = 2 CNV events were used

length of 21, 91, 40, 64, 92, and 115 Mb (Table 1). CNV sharing analysis was conducted to further investigate the distribution pattern of CNVRs among different populations. We observed a large number of CNVRs shared by multiple groups, which were considered as common variants. As shown in Supplementary Figure S2, 41 CNVRs were shared by African, European and Asian groups, while 89 and 86 CNVRs were shared within African and European groups, respectively. The goat CNV sharing results among African, Asian, and European groups showed that African goats have more lineage-specific CNVs than Asian and European goats, and the number of CNVRs shared by Asian and European goats was smaller than that were shared either by

Asian and African goats or by European and African goats (Supplementary Figure S2A).

Among 978 CNVRs in six geographic groups, we observed 526 CNVRs containing at least two CNV events. These have a cumulative length of ~191 Mb (~6.53% of the goat genome, Supplementary Table S3). The genomic distribution and frequencies of CNVRs (CNV events > = 2) on autosomes in six subgroups were shown in Fig. 2. Herein, we can clearly observe some group-differential CNVRs showing different patterns of distribution among six goat groups, e.g. CNVR1 in chr17, CNVR4 in chr13, CNVR8 in chr3, and CNVR9 in chr14. Overall, we detected eight CNV regions containing high frequency of CNV events in all
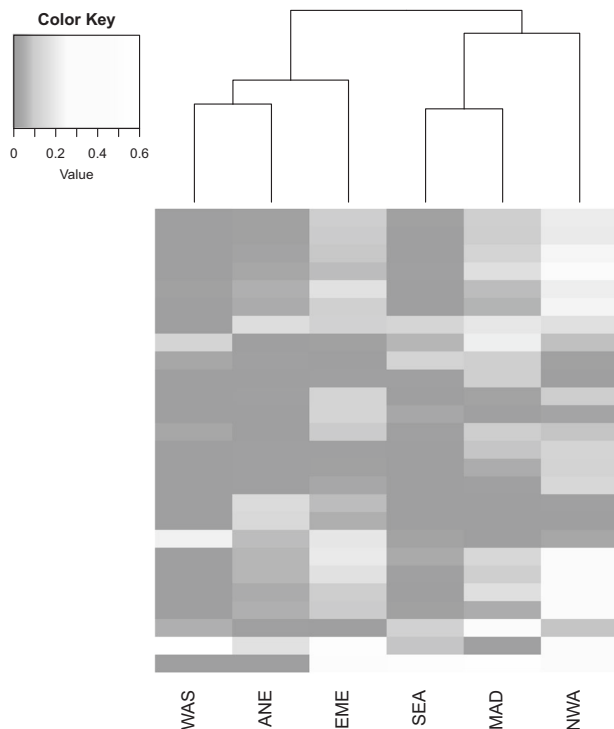
Fig. 3 Six diverse goat groups were grouped into clusters in heatmap plots based on the top 5% of CNVRs (26 CNVRs). WAS: Western Asia; EME: Eastern Mediterranean; ANE: Alpine & Northern Europe; NWA: Northwestern Africa; SEA: Southeastern Africa; MAD: Madagascar. The value in the legend of the heatmap is CNV event frequency (CNV count in each CNVR /sample size)

samples on 29 autosomes (CNV count/sample count > 10%, Supplementary Table S3). Among them, the top four CNVRs with high frequency were identified at chr17: 59,870,207–61,156,582, chr5: 36,233,025–36,842,273, chr12: 13,172,367–13,621,196, and chr13: 63,117,523–63,416,183, with corresponding total CNV event counts (frequencies) of 402 (39.30%), 232 (22.68%), 149 (14.57%), and 143 (13.98%), respectively (Supplementary Table S3). In each group, we first filtered CNVRs that were enriched by CNV by only keeping the CNVRs harboring 10 or more events. We obtained 35, 42, 8, 3, 29, and 5 common CNVs in SEA, NWA, MAD, WAS, EME, and ANE groups, respectively.

Moreover, we estimated the average CNV event count per individual (CNV event frequency) for each CNVR in every group. The variance within each CNVR was considered in each group, to explore the potential differences involved in selection pressure for CNVs (Supplementary Table S3). Our analysis revealed the distinct pattern of CNV frequency across six groups. The average frequencies for CNVR were 0.48, 0.98, 0.85, 0.43, 0.72, and 0.34% while the largest frequencies for CNVR were 50.49, 43.41, 53.85, 60.38, 49.74, and 17.48% in SEA, NWA, MAD, WAS, EME, and ANE group, respectively.

CNV differentiation analysis

To explore the evolution and selection characteristics of CNVs, we performed a cluster heatmap analysis using the CNV event frequency of CNVRs in each group. Only the 26 CNVRs whose variances were ranked the top 5% among 526 CNVRs were used. The clustering matrices, which were estimated from CNV event frequency, clearly arranged groups according to their geographic origin (Fig. 3). As expected, African groups (SEA, MAD, and NWA) were arranged separately from European and Asian groups. Within African groups, the SEA and MAD clustered better together than with NWA. Interestingly, the ANE group was grouped together with the Asian group.

Copy number variable genes

We next cataloged the gene content spanning CNVRs to predict the potential roles of these identified CNVs may play on phenotypic variations. Using the gene models present in ARS1, a total of 1437 genes were identified in 526 CNVRs (Supplementary Table S4). Among them, 154 genes were detected to overlap with group-differential CNVRs with a big variance across six groups larger than 0.01. By querying CNV-overlapped genes against OMIA, we found that 43 functional CNV-overlapped genes were related to health and diseases in animals such as cattle, goat, sheep, pig, dog, and so on. Of note, nine of them located in group-differential CNVRs with big variances larger than 0.01 and some important traits and/or disorders like coat color and abortion were associated with these copy number variable genes (Supplementary Table S4). Based on our PANTHER analysis, the enriched GO terms included molecular function terms (binding); biological process terms (such as embryonic development and metabolic processes) and cellular component terms (cytoplasmic parts and intracellular organelle parts) (Supplementary Table S5).

Using DNA samples from African and European goats, we performed qPCR to validate several important CNVs overlapping with functional genes involved in local adaptations such as coat color, behavior, embryonic development, osteopetrosis, and metabolic processes. The results showed that approximately 57% of the qPCR results were consistent with the PennCNV prediction in terms of copy number change directions (Supplementary Table S2). Meanwhile, we provided an example for the underlying data of CNV calling by PennCNV. As shown in Fig. 4, within CNVR9 (chr14: 80,740,427–81,602,889), 21 SNP probes were used to define the CNV status based on the signal intensity of LRR and BAF. Among them, probes 14 and 15 were mapped to *DGAT1* (the Diacylglycerol O-Acyltransferase 1). Consistent with the CNV call results,
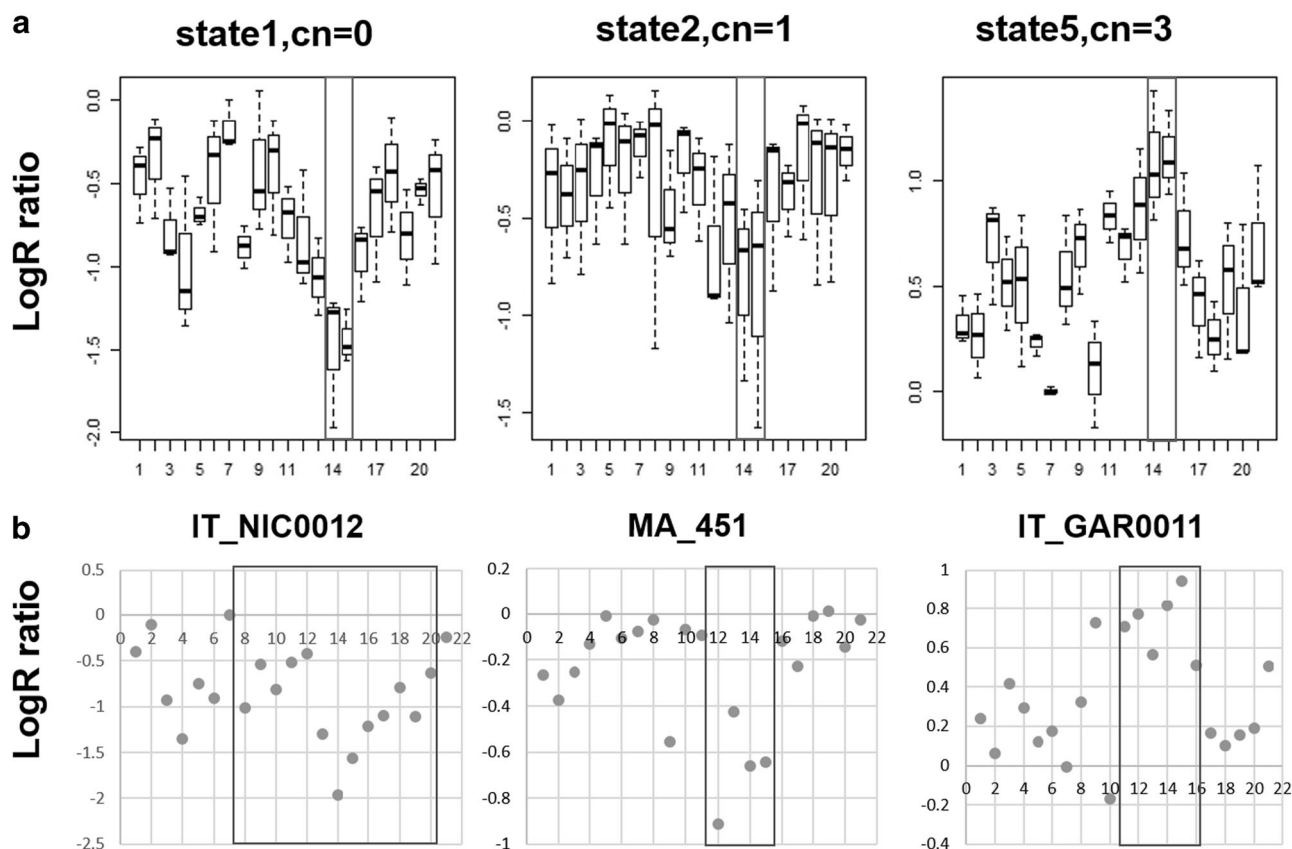
Fig. 4 PennCNV plot of Log R Ratio for three different status of *DGAT1* CNV within CNVR9. **a** Boxplot for distributions of LRRs of 21 probes within CNVR9. Probes 14 and 15 in the red box were located in *DGAT1*. The mean and standard deviation were generated by three sample groups with *DGAT1* CNV state 1 (*cn* = 0), 2 (*cn* = 1) and 5 (*cn* = 3), respectively. **b** Scatter plot of LRR for a CNV event

harboring *DGAT1* (plots in blue box) in three individuals randomly selected. IT_NIC0012, MA_451, and IT_GAR0011 were defined as state1 (*cn* = 0), state2 (*cn* = 1), and state5 (*cn* = 3) by PennCNV, respectively. Only CNVs, spanning at least 3 probes with a standard deviation (*SD*) of LRR < 0.30, were reported

three different CNV statuses (*cn* = 0, 1, 3) at *DGAT1* CNV locus showed obvious difference for the LRR value.

Discussion

Our study is the first systematic effort to perform comprehensive analyses of the CNV population genetic properties for the worldwide goat populations across different geographical areas. Using the large-scale data from the ADAPTmap Project, we identified a total of 6286 CNVs in 1023 goats (an average of 6 CNVs per goat) and 978 CNVRs spanning 262 Mb, or 8.96% of the goat genome. In nine goats of four European breeds (Saanen and Camosciata from the Alps region, Girgentana from Sicily in the Mediterranean area and Murciano Granadina from Spain), Fontanesi et al. (2010) identified 161 CNVs (an average of 18 CNVs per goat) and 127 CNVRs covering 11.47 Mb by a cross species cattle–goat aCGH experiment. Compared with the previous goat CNV studies, our study explored many more breeds and individuals. These newly discovered CNVs in specific goat breeds can be further explored to

understand their potential functional and evolutionary features in the goat genome. The large and comprehensive goat data offer a valuable platform to explore the population genetic characteristics and evolutionary selection of CNVs. Our results indicated that segregating CNV, just like SNP, express a certain degree of diversity across different breeds. Similar to the published studies on CNV population genetic properties in some other species (human, dog, cattle, and sheep), this study supports that the diversity of goat CNVs is involved in the differential selective pressure during goat domestication, which are probably related to functional and evolutionary aspects of the goat genome.

In this study, the new goat reference ARS1 assembly, which was generated by long reads data with the combination of 3rd sequencing technologies (Celera Assembler v. 8.2, BioNano Genomics Irys and Lachesis Hi-C) (Bickhart et al. 2017), was utilized to remap original marker positions and 978 CNVRs were detected in the goat genome. However, we found a minimal overlap with the previous published results based on the Btau_4.0 assembly (Fontanesi et al. 2010). We first converted their genomic coordinates based on the ARS1 assembly using liftOver (https://

genome.ucsc.edu/cgi-bin/hgLiftOver), at several relaxed thresholds (minimum ratio of bases that must be remapped, varying from 10 to 50%). At 10%, a total of 42 CNVRs (1,837,187 bp) were converted to ARS1 and only one CNVR (chr16: 34,326,169–34,331,914) overlapped with our results (chr16: 34,261,365–34,856,843) (Supplementary Table S6). There could be three reasons for this low overlap: 1) the liftOver conversion between assemblies was inefficient for DNA segments within copy number variable regions; 2) different sample sizes were screened between previous study (nine animals from four European breeds) and this study (1023 animals from 50 worldwide goat breeds); 3) the species difference between cattle and goat. For goat CNV calling, the previous study used a bovine-based CGH platform while this study used the goat medium-high SNP array.

In terms of clustering analysis, the results which were generated by CNV data in this study are generally consistent with those based on SNPs among these goat breeds (Colli et al. under review). The clustering results provide a glimpse of the migration of goats and further support the hypothesis of a putative migration of domestic livestock from the Fertile Crescent towards western and northern Europe, which was validated by the SNP study (Colli et al. under review). Only a slight difference was observed. In Colli et al.'s study, the goat breeds in the European groups (ANE and EME) were grouped first, but this CNV study showed that WAS and ANE were grouped first, and then they were grouped with EME. The difference may be due to tyhe fact that SNP's location and status are well defined whereas CNV's location and status are not well defined. The CNV-based population study is still more challenging because of the difficulty of genotyping them. Most CNV calling algorithms were developed to detect CNV sample by sample, which often lead to identification of CNVs with various coordinates (start and end positions) and make it difficult to explore CNV at population levels (Hou et al. 2012). Thus, CNV-based population genetics studies should be improved with precise CNV mapping via high-throughput sequencing and new CNV calling methods, which can balance the tradeoffs between individual CNV calling and common CNV calling at the population level (Bickhart et al. 2016; Sudmant et al. 2015).

In our study, variances of the CNV event frequency for all CNVRs were investigated among six groups, to understand the diversity of CNV in the worldwide goat population. Among the 978 CNVRs, CNVR1 (chr17: 59,870,207–61,156,582) showed the highest frequency and variance. Its frequencies in all three Africa groups (NWA, SEA, and MAD) and south Europe (EME) were high, while those in Asia (WAS) and north Europe (ANE) were low, suggesting that it might be a crucial group-differential candidate CNVR involved in the evolution and selection

characteristics among goat breeds. We then explored for the underlying functional genes and found it was overlapped by *NR3C2, ARHGAP10, PRMT9, TMEM184C*, and *EDNRA* genes. Consistently, these five genes were also mapped to a 1 Mb CNV among three different phenotype Boer goats (solid-colored, spotted and traditional) (Menzi et al. 2016). The genomic amplification of *EDNRA* (endothelin receptor type A) has been demonstrated to be positively associated with the degree of white spotting (OMIA 000214-9925). On the other hand, *EDNRA* is also related to branchial arch development and fetal muscle development (Sato et al. 2008). *NR3C2*(Nuclear Receptor Subfamily 3 Group C Member 2) encodes a mineralocorticoid receptor involved in blood pressure regulation. Mutations in *NR3C2* are related to disorders like autosomal dominant pseudohypoaldosteronism type I (Geller et al. 1998), early onset hypertension (Martinez et al. 2009) in humans, and anxiety-like behavior in mice (Rozeboom et al. 2007). Also, it was reported to overlap with the CNVR No. 90 on BTA17 and its additional copies of the *NR3C2* gene might relate to specific adaptation traits to harsh and dry environments and to the very mild behavior of the Girgentana goats (Fontanesi et al. 2010). Besides, the CNVR2 (chr5: 36233025–36842273) also caught our attention with the second higher frequency (22.68%, Supplementary Table S3) and variance among six goat subgroups (0.05, Supplementary Table S4). In the two genes overlapped by the CNVR2, we noticed that *ADAMTS20* (ADAM Metallopeptidase with Thrombospondin Type 1 Motif 20) played a critical role in coat color variation during goat domestication based on the color gene database (Dong et al. 2015). *ADAMTS20* encodes a preprotein that is proteolytically processed to generate the mature peptide. From a genome-wide association study in dogs and humans, Wolf et al. (2015) have identified *ADAMTS20* as a risk variant for cleft lip and palate (OMIA 001140-9615). The SNP detected in *ADAMTS20* has been demonstrated to associate with melanocyte development by alignment of the sequences of eight goats (Crepaldi and Nicoloso 2007). It is the first time to observe the *ADAMTS20* CNV in goats. These results indicated that the diversity of *ADAMTS20* CNV could be related to coat color adaptations in the worldwide goat breeds. In addition, we identified the CNVR4 (chr13: 63,117,523–63,416,183) as a hotspot CNV region with the 13.98% frequency of CNV events among all test animals. The CNVR4 harbored the *ASIP* (Agouti signaling peptide), *AHCY* (adenosyl homocysteine) and a part of *ITCH* (itchy E3 ubiquitin protein ligase) genes. It was consistent with the results as previously published in sheep (Norris and Whan 2008), which identified a 190-kb tandem duplication encompassing the ovine *ASIP* and *AHCY* coding regions and the *ITCH* promoter region as the genetic cause of white coat color of dominant white/tan (A(Wt)) agouti sheep. This

indicated that the CNVR4 could represent a recurrent interspecies CNVR. In Girgentana and Saanen goat breeds, Fontanesi et al. (2009) also detected a CNV overlapped with *ASIP* and *AHCY* and found it may be responsible for variation in coat color. Recently, a signature of selection contrasting goat breeds with different coat colors also reported the *ASIP* and *ADAMTS20* genes (Bertolini et al., under review). In this study, the copy number gain type of *EDNRA*, *NR3C2*, *ASIP*, *AHCY*, and *ITCH* CNV loci were validated by PCR test in African indigenous goats (Supplementary Table S2). Overall, these results indicated that the evolution and selection of CNVR1, CNVR2 and CNVR4 might have affected the crucial roles of the overlapping functional genes in goat color and growth adaptations.

Our study identified 154 genes overlapping with eight group-differential CNVRs with a big variance higher than 0.1, suggesting these genes may undergo differential evolutionary pressures and selection in goat domestication. Some of the group-differential genes were novel and related to goat metabolic processes, osteopetrosis, and embryonic development. For example, *DGAT1* was detected in CNVR9 (chr14: 80,740,427–81,602,889), with a high frequency (9.97%, Supplementary Table S3) and a large variance for its frequency across six groups (1.73%, Supplementary Table S3). The CNVR9 is almost 1 Mb and overlaps with 49 genes. Among all the samples, the CNV at the *DGAT1* locus was observed in 88 individuals from four goat subgroups (EME, ANE, MAD, NWA, and SEA). Interestingly, 83 of them were copy number loss type (0 or 1 copy) and the other five samples with copy gain were all from EME subgroup. *DGAT1* encodes a multipass transmembrane protein that functions as a key metabolic enzyme catalyzing the last and limiting step of triglyceride synthesis (Mayorek et al. 1989). The high protein sequence conservation for *DGAT1* between ruminant species supports the hypothesis that this protein plays a major role in biological functions, and some functional variants in the *DGAT1* gene have been demonstrated to have major effects on milk yield and composition in cattle, sheep and goat (Martin et al. 2017; Grisart et al. 2002). However, in some previous sheep and goat studies, it was chosen as the control gene for CNV validation by qPCR with the assumption that there were two copies of DNA segment in this region (Fontanesi et al. 2011; Fontanesi et al. 2009; Yan et al. 2017). Herein, we carried out qPCR to validate the *DGAT1* CNV locus. Similar to the CNV call results by PennCNV, the qPCR experiments only detected the copy number loss type of *DGAT1* CNV in African goats. Therefore, our study provides a new venue for researchers to understand the detailed roles of *DGAT1* copy number change in ruminants. In addition, we also found some meat production related genes such as *KDM5B* (Lysine Demethylase 5B) in CNVR5,

ADAM8 (ADAM Metallopeptidase Domain 8) in CNVR11, and *SHH* (Sonic Hedgehog) in CNVR73. *KDM5B* is a zinc finger transcription factor and encodes a lysine-specific histone demethylase that demethylates Lys-4 of histone H3. *KDM5B* can act as a transcriptional corepressor for *PAX9*, which is related to Wnt / Hedgehog / Notch and mesodermal commitment pathway. *ADAM8* encodes a member of the ADAM (a disintegrin and metalloprotease domain) family, the members of which are implicated in biological processes including fertilization, muscle development, and neurogenesis. Nishimura et al. (2015) has demonstrated that *ADAM8* is involved in invasion of neutrophils into injured muscle fibers, and is required for their efficient elimination. *SHH* is a critical moderator inhibiting fat formation and promoting myogenic and osteogenic differentiation (Bentzinger et al. 2012; James et al. 2010). The deletion in *SHH* long-range enhancers has been demonstrated to associate with Holoprosencephaly and Currarino syndrome (Horn et al. 2004). Recently, the *SHH* CNV was also identified in Chinese cattle and was associated with body status (Liu et. al, under review). With the Livestock and Poultry Pangenome Database (http://animal.nwsuaf.edu.cn/code/index.php/panLiv), we also found the evidences for the CNVs at *DGAT1*, *KDM5B*, and *SHH* loci in ruminants. These findings indicated that the CNVs mentioned above could be considered as candidates for improving productive traits (meat and milk) in goat breeding.

As known, animal health and production are both important for livestock, and single gene diseases are easier to explain as compared to complex traits. OMIA is a comprehensive, annotated catalog of inherited disorders and other familial traits in animals (Lenffer et al. 2006). As listed in OMIA database, we identified a series of functional CNV-overlapped genes related to animal health, growth development and fertility traits. For example, *EXOSC4* (Exosome component 4), encoding a component of the RNA exosome machinery, is a potential factor involved in the maintenance of genome stability and it is deregulated in lung cancer. *EXOSC4* has been identified as one of the functional gene associated with embryonic lethality (OMIA 002042-9913, Charlier et al. 2016). *MED22* (mediator Complex Subunit 22) is also related to abortion (embryonic lethality) in *Bos taurus* (OMIA 002043-9913). The related pathways of *MED22* are regulation of lipid metabolism by *PPARα* (peroxisome proliferator-activated receptor alpha) and metabolism. *KDSR* (3-ketodihydrosphingosine Reductase) is related to metabolism and sphingolipid metabolism. Cole et al. (2016) has identified the *KDSR* mutation as a causative mutation for haplotype BHM (spinal muscular atrophy, a reduced-fertility haplotype) in *Bos taurus* (OMIA 000939-9913). *TMEM95* (transmembrane protein 95 encoding) was associated with male subfertility in Bos taurus (OMIA 001902-9913). A GWAS by Pausch et al.

(2014) for male reproductive ability on 7961 Fleckvieh AI bulls highlighted a region on BTA19, and the whole-genome re-sequencing data revealed the candidate causal nonsense mutation in the *TMEM95* gene. *CHRNB1* (cholinergic receptor nicotinic beta 1 subunit) is associated with the bovine arthrogryposis multiplex congenita (AMC) phenotype (OMIA 002022-9913), which is a syndromic term for a congenital condition characterized by multiple joint contractures (Agerholm et al. 2016). *CLCN7* (chloride Voltage–Gated Channel 7) plays an important role in osteopetrosis with gingival hamartomas in Belgian Blue cattle (OMIA 001887-9913, Sartelet et al. 2013). Chloride–proton exchange by the lysosomal anion transporter ClC-7/Ostm1 is of pivotal importance for the physiology of lysosomes and bone resorption. The mutations in *ADAMTS2* (ADAM metallopeptidase with thrombospondin type 1 motif 2), which encodes for the enzyme procollagen I amino proteinase, has been identified to cause the Dermatosparaxis in sheep (OMIA 000328-9940) and Belgian Blue cattle (OMIA 000328-9913, Colige et al. 1999). Overall, these functional genes observed in CNVRs allow this study to play a fundamental role in the support of goat biomedical and genomic research.

In conclusion, this study constructed an extensive CNV map and performed the first CNV population genetic analysis with thousands of samples from diverse goat breeds. The promising CNVRs, which were of high-variance across different geographic goat groups and/or overlapped with critical functional genes related to animal productive traits or health, were identified and could be further used for exploring the functional and evolutionary aspects of the goat genome. The CNV-based population genetics analyses reveal goat population structures, reflecting the population history of different goat breeds. These results represent a valuable molecular resource, highlighting the importance of the goat CNV to understand the genome architecture of livestock. Deeper understanding of CNV will have the potential to help breeders design effective selection strategies and enhance genetic improvement.

Data archiving

This article does not report new empirical data or software.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

Agerholm JS, McEvoy FJ, Menzi F, Jagannathan V, Drögemüller C (2016) A CHRNB1 frameshift mutation is associated with familial arthrogryposis multiplex congenita in Red dairy cattle BMC Genom 17:479

Alexander DH, Novembre J, Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals Genome Res 19:1655–1664

Alvarez CE, Akey JM (2012) Copy number variation in the domestic dog Mamm Genome 23:144–163

Bentzinger CF, Wang YX, Rudnicki MA (2012) Building muscle: molecular regulation of myogenesis Cold Spring Harb Perspect Biol 4:a008342

Bickhart DM, Rosen BD, Koren S, Sayre BL, Hastie AR, Chan S et al (2017) Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome Nat Genet 49:643–650

Bickhart DM, Xu L, Hutchison JL, Cole JB, Null DJ, Schroeder SG et al (2016) Diversity and population-genetic properties of copy number variations and multicopy genes in cattle DNA Res 23:253–262

Charlier C, Li W, Harland C, Littlejohn M, Coppieters W, Creagh F et al (2016) NGS-based reverse genetic screen for common embryonic lethal mutations compromising fertility in livestock Genome Res 26:1333–1341

Chen W K, Swartz J D, Rush L J, Alvarez C. E (2008) Mapping DNA structural variation in dogs Genome Res 19:500–509

Cole JB, Null DJ, VanRaden PM (2016) Phenotypic and genetic effects of recessive haplotypes on yield, longevity, and fertility J Dairy Sci 99:7274–7288

Colige A, Sieron AL, Li S-W, Schwarze U, Petty E, Wertelecki W et al (1999) Human Ehlers-Danlos syndrome type VII C and bovine dermatosparaxis are caused by mutations in the procollagen I N-proteinase gene Am J Hum Genet 65:308–317

Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y et al (2010) Origins and functional impact of copy number variation in the human genome Nature 464:704–712

Crepaldi P, Nicoloso L (2007) SNPs in coat colour genes in goats. Ital J Anim Sci 6(sup1):91–93

Diamond J (2002) Evolution, consequences and future of plant and animal domestication Nature 418:700–707

Dong Y, Zhang X, Xie M, Arefnezhad B, Wang Z, Wang W et al (2015) Reference genome of wild goat (*Capra aegagrus*) and sequencing of goat breeds provide insight into genic basis of goat domestication BMC Genom 16:431

Feuk L, Carson AR, Scherer SW (2006) Structural variation in the human genome Nat Rev Genet 7:85–97

Fontanesi L, Beretti F, Martelli P, Colombo M, Dall'Olio S, Occidente M et al (2011) A first comparative map of copy number variations in the sheep genome Genomics 97:158–165

Fontanesi L, Beretti F, Riggio V, González EG, Dall'Olio S, Davoli R et al (2009) Copy number variation and missense mutations of the agouti signaling protein (ASIP) gene in goat breeds with different coat colors Cytogenet Genome Res 126:333–347

Fontanesi L, Martelli PL, Beretti F, Riggio V, Dall'Olio S, Colombo M et al (2010) An initial comparative map of copy number variations in the goat (Capra hircus) genome BMC Genom 11:639

Geller DS, Rodriguez-Soriano J, Boado AV, Schifter S, Bayer M, Chang SS et al (1998) Mutations in the mineralocorticoid receptor gene cause autosomal dominant pseudohypoaldosteronism type I Nat Genet 19:279–281

Goddard ME, Hayes BJ (2009) Mapping genes for complex traits in domestic animals and their use in breeding programmes Nat Rev Genet 10:381–391

Grisart B, Coppieters W, Farnir F, Karim L, Ford C, Berzi P et al (2002) Positional candidate cloning of a QTL in dairy cattle: identification of a missense mutation in the bovine DGAT1 gene with major effect on milk yield and composition Genome Res 12:222–231

Horn D, Tönnies H, Neitzel H, Wahl D, Hinkel GK, Moers AV et al. (2004) Minimal clinical expression of the holoprosencephaly spectrum and of Currarino syndrome due to different cytogenetic rearrangements deleting the Sonic Hedgehog gene and the HLXB9 gene at 7q36. 3. Am J Med Genet Part A 128(1):85–92

Hou Y, Bickhart DM, Hvinden ML, Li C, Song J, Boichard DA et al (2012) Fine mapping of copy number variations on two cattle genome assemblies using high density SNP array BMC Genom 13:376

James AW, Leucht P, Levi B, Carre AL, Xu Y, Helms JA et al (2010) Sonic Hedgehog influences the balance of osteogenesis and adipogenesis in mouse adipose-derived stromal cells Tissue Eng Part A 16:2605–2616

Lashmar S, Visser C, van Marle-Köster E (2015) Validation of the 50k Illumina goat SNP chip in the South African Angora goat South Afr J Anim Sci 45:56–59

Lenffer J, Nicholas FW, Castle K, Rao A, Gregory S, Poidinger M et al. (2006) OMIA (Online Mendelian Inheritance in Animals): an enhanced platform and integration into the Entrez search interface at NCBI. Nucleic Acids Res 34(suppl_1):D599–D601

Martin P, Palhière I, Maroteau C, Bardou P, Canale-Tabet K, Sarry J et al (2017) A genome scan for milk production traits in dairy goats reveals two new mutations in Dgat1 reducing milk fat content Sci Rep 7:1872

Martinez F, Mansego ML, Escudero JC, Redon J, Chaves FJ (2009) Association of a mineralocorticoid receptor gene polymorphism with hypertension in a Spanish population Am J Hypertens 22:649–655

Mayorek N, Grinstein I, Bar-Tana J (1989) Triacylglycerol synthesis in cultured rat hepatocytes FEBS J 182:395–400

Ma Q, Liu X, Pan J, Ma L, Ma Y, He X et al. (2017) Genome-wide detection of copy number variation in Chinese indigenous sheep using an ovine high-density 600 K SNP array. Sci Rep 7(1):912

Menzi F, Keller I, Reber I, Beck J, Brenig B, Schütz E et al. (2016) Genomic amplification of the caprine EDNRA locus might lead to a dose dependent loss of pigmentation. Sci Rep 6:28438

Naderi S, Rezaei H-R, Pompanon F, Blum MG, Negrini R, Naghash H-R et al (2008) The goat domestication process inferred from large-scale mitochondrial DNA analysis of wild and domestic individuals Proc Natl Acad Sci 105:17659–17664

Narang A, Jha P, Kumar D, Kutum R, Mondal AK, Consortium IGV et al (2014) Extensive copy number variations in admixed Indian population of African ancestry: potential involvement in adaptation Genome Biol Evol 6:3171–3181

Nishimura D, Sakai H, Sato T, Sato F, Nishimura S, Toyama-Sorimachi N et al. (2015) Roles of ADAM8 in elimination of injured muscle fibers prior to skeletal muscle regeneration. Mech Dev 135:58–67

Norris BJ, Whan VA (2008) A gene duplication affecting expression of the ovine ASIP gene is responsible for white and black sheep Genome Res 18:1282–1293

Pausch H, Kölle S, Wurmser C, Schwarzenbacher H, Emmerling R, Jansen S et al. (2014) A Nonsense Mutation in TMEM95 encoding a nondescript transmembrane protein causes idiopathic male subfertility in cattle. PLoS Genet 10:e1004044

Pickrell JK, Pritchard JKJPg (2012) Inference of population splits and mixtures from genome-wide allele frequency data PLoS Genet 8:e1002967

Rozeboom AM, Akil H, Seasholtz AF (2007) Mineralocorticoid receptor overexpression in forebrain decreases anxiety-like behavior and alters the stress response in mice Proc Natl Acad Sci 104:4688–4693

Sartelet A, Stauber T, Coppieters W, Ludwig CF, Fasquelle C, Druet T et al (2013) A missense mutation accelerating the gating of the lysosomal Cl-/H+-exchanger ClC-7/Ostm1 causes osteopetrosis with gingival hamartomas in cattle Dis Models Mech 7:119–128

Sato T, Kawamura Y, Asai R, Amano T, Uchijima Y, Dettlaff-Swiercz DA et al (2008) Recombinase-mediated cassette exchange reveals the selective use of Gq/G11-dependent and-independent endothelin 1/endothelin type A receptor signaling in pharyngeal arch development Development 135:755–765

Sudmant PH, Mallick S, Nelson BJ, Hormozdiari F, Krumm N, Huddleston J et al (2015) Global diversity, population stratification, and selection of human copy-number variation Science 349:aab3761

Talenti A, Bertolini F, Pagnacco G, Pilla F, Ajmone-Marsan P, Rothschild MF et al (2017) The Valdostana goat: a genome-wide investigation of the distinctiveness of its selective sweep region Mamm Genome 28:114–128

Tosser-Klopp G, Bardou P, Bouchez O, Cabau C, Crooijmans R, Dong Y et al (2014) Design and characterization of a 52K SNP chip for goats PloS ONE 9:e86227

Wang K, Li M, Hakonarson H (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data Nucleic Acids Res 38:e164

Wang K, Li M, Hadley D, Liu R, Glessner J, Grant SF et al (2007) PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data Genome Res 17:1665–1674

Wang G-D, Xie H-B, Peng M-S, Irwin D, Zhang Y-P (2014) Domestication genomics: evidence from animals Annu Rev Anim Biosci 2:65–84

Wolf Z T, Brand H A, Shaffer J R, Leslie E J, Arzi B, Willet C E et al (2015) Genome-wide association studies in dogs and humans identify ADAMTS20 as a risk variant for cleft lip and palate PLoS Genet 11:e1005059

Xu L, Hou Y, Bickhart DM, Zhou Y, Song J, Sonstegard TS et al. (2016) Population-genetic properties of differentiated copy number variations in cattle. Sci Rep 6:23161

Yan J, Blair HT, Liu M, Li W, He S, Chen L et al. (2017) Genome-wide detection of autosomal copy number variants in several sheep breeds using Illumina OvineSNP50 BeadChips. Small Rumin Res 155:24–32

Yang L, Xu L, Zhou Y, Liu M, Wang L, Kijas JW et al (2017) Diversity of copy number variation in a worldwide population of sheep Genomics 110:143–148

Zarrei M, MacDonald JR, Merico D, Scherer SW (2015) A copy number variation map of the human genome Nat Rev Genet 16:172–183

Zhang F, Gu W, Hurles ME, Lupski JR (2009) Copy number variation in human health, disease, and evolution. Annu Rev Genom Human Genet 10:451–481

Zhou Y, Utsunomiya YT, Xu L, Bickhart DM, Sonstegard TS, Van Tassell CP et al. (2016) Comparative analyses across cattle genders and breeds reveal the pitfalls caused by false positive and lineage-differential copy number variations. Sci Rep 6:29219