Detection of copy number variants in African goats using whole genome sequence data

Citation for published version:
VarGoats Consortium 2021, 'Detection of copy number variants in African goats using whole genome sequence data', BMC Genomics, vol. 22, no. 1, pp. 398. https://doi.org/10.1186/s12864-021-07703-1

Digital Object Identifier (DOI):
10.1186/s12864-021-07703-1

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Publisher's PDF, also known as Version of record

Published In:
BMC Genomics

Publisher Rights Statement:
© The Author(s). 2021 Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
Detection of copy number variants in African goats using whole genome sequence data

Wilson Nandolo1,2, Gábor Mészáros1, Maria Wurzinger1, Liveness J. Banda2, Timothy N. Gondwe6, Henry A. Mulindwa3, Helen N. Nakimbugwe6, Emily L. Clark5, M. Jennifer Woodward-Greene6,7, Mei Liu6, the VarGoats Consortium, George E. Liu6, Curtis P. Van Tassell6, Benjamin D. Rosen6* and Johann Sölkner1

Abstract

Background: Copy number variations (CNV) are a significant source of variation in the genome and are therefore essential to the understanding of genetic characterization. The aim of this study was to develop a fine-scaled copy number variation map for African goats. We used sequence data from multiple breeds and from multiple African countries.

Results: A total of 253,553 CNV (244,876 deletions and 8677 duplications) were identified, corresponding to an overall average of 1393 CNV per animal. The mean CNV length was 3.3 kb, with a median of 1.3 kb. There was substantial differentiation between the populations for some CNV, suggestive of the effect of population-specific selective pressures. A total of 6231 global CNV regions (CNVR) were found across all animals, representing 59.2 Mb (2.4%) of the goat genome. About 1.6% of the CNVR were present in all 34 breeds and 28.7% were present in all 5 geographical areas across Africa, where animals had been sampled. The CNVR had genes that were highly enriched in important biological functions, molecular functions, and cellular components including retrograde endocannabinoid signaling, glutamatergic synapse and circadian entrainment.

Conclusions: This study presents the first fine CNV map of African goat based on WGS data and adds to the growing body of knowledge on the genetic characterization of goats.

Keywords: African goats, Copy number variations, Whole genome sequence

Background

Structural variations (SV) are an important source of genetic variation [1–4]. SV are generally considered to comprise a myriad of subclasses that consist of unbalanced copy number variants (CNV), which include deletions, duplications and insertions of genetic material, as well as balanced rearrangements, such as inversions and interchromosomal and intrachromosomal translocations [5]. Deletions and insertions are referred to as unbalanced SV because they result in changes in the length of the genome. Insertions or deletions in the genome are typically considered CNV when they are at least 50–1000 base-pairs (bp) long [6–11]. CNV are not as abundant as single nucleotide polymorphisms (SNP), but because of their larger sizes, they may have a dramatic effect on gene expression in individuals [12]. Duplication or deletion in or near a gene or the regulatory region of the gene may lead to modification of the function of the gene.

* Correspondence: Ben.Rosen@usda.gov
6Animal Genomics and Improvement Laboratory, USDA-ARS, Beltsville, MD, USA
Full list of author information is available at the end of the article
CNV cover about 4.5–9.8% of the human genome [13] and are associated with many Mendelian disorders [12]. Girirajan et al. [14] found that CNV significantly determine the severity and prognosis of many genetic disorders. Approximately 14% of diseases in children with intellectual disability are caused by CNV [15]. On the other hand, some CNV have been found to be associated with adaptive fitness of individuals, such as adaptation to starch diets associated in the gene encoding α-amylase [13].

Traditionally, microarray-based comparative genomic hybridization (array CGH) or SNP genotyping arrays are used to detect CNV. Several studies have been carried out using these methods to detect and map CNV in the goat genome, including studies by Fontanesi et al. [16] in four goat breeds; Nandolo et al. [17] in 13 East African goat breeds; and Liu et al. [18] in the global goat population.

Detecting CNV using array CGH and SNP genotyping arrays suffers from shortcomings that include hybridization noise, limited coverage of the genome, low resolution, and difficulty in detecting novel and rare mutations [19–21]. The development of whole-genome sequencing (WGS) technologies has made it possible for more rigorous and accurate detection of CNV.

According to Mills et al. [22], WGS-based CNV detection methods fall into four major approaches: methods based on paired-end (PE) mapping, split reads (SR), read depth (RD) and de novo assembly of a genome (AS). The PE and SR methods are useful for detection of small-scale CNV [23], and several algorithms are loosely based on them, including BreakDancer [24], Pindel [25], and Delly [26]. RD approaches are very useful for detection of larger CNV. Algorithms using this approach include CNV-Seq [27], CNVnator [28] and the event-wise testing approach (EWT) developed by Yoon et al. [29]. The methods can also be combined. For example, LUMPY [30] is able to combine two or more of the previous approaches to refine SV detection. Assembly-based approaches are computationally intensive and are therefore not generally used with WGS data [23, 31]. Most of these SV-detection algorithms have been extensively reviewed [1, 31–34].

LUMPY implements a breakpoint prediction framework, where a breakpoint is defined as a pair of genomic regions that are adjacent in a sample, but not in the reference genome. The location of the breakpoint is determined using a probability function that considers different sources of evidence supporting the existence of a breakpoint, including information from discordant read pairs and split reads. A discordant read pair occurs when sequence from two ends of an insert are inconsistent when compared to the reference genome. These inconsistencies result from differences between mapping distance or the orientation between the pairs of sequences [35, 36]. Split reads are sequences that map to the reference genome on one end only, and, as explained by Ye and Hall [33], such reads can indicate the location of a breakpoint with a high degree of certainty. There are similar algorithms that rely heavily on the use of breakpoints to determine genome rearrangements at single-nucleotide resolution, including Delly [26] and Pindel [25].

Like LUMPY, Manta [37] incorporates use of PE and SR methods. However, Manta also uses AS analysis. Manta overcomes the computational expense of AS methods by splitting the work into many smaller workflows which can be carried out in parallel. Manta scans the genome for SV and then scores, genotypes and filters the SV based on diploid germline and somatic biological models [37]. Manta can detect all structural variant types that are identifiable in the absence of copy number analysis and large-scale de-novo assembly, which is why this approach is also a good candidate for joint analysis of small sets of diploid individuals, tumor samples, and similar analyses. Both LUMPY and Manta are good at identifying SV break points with high resolution.

Many studies have been carried out to detect CNV using WGS data in various domesticated species: cattle [38], cats [39], chickens [40], dogs [41], etc. So far, there is no report of goat CNV discoveries using WGS data. The goal of this study was to identify CNV in the goat genome through the intersection of LUMPY and Manta outputs as a part of the characterization of African goats in conjunction with the ADAPTmap project [42]. Goats are a very important farm animal genetic resource for the livelihoods of African smallholders, and a deeper understanding of the goat genome is necessary to facilitate the improvement of goats in the region. This study aimed to generate a fine-scale CNV map for the goat genome.

**Results**

**Number and distribution of CNV**

The number of CNV detected depended on the filter levels (low, medium, or stringent) and the cut-off point for CNV length (3 Mb or 10 Mb) as given in Supplementary Figure 11 (Additional file 2). Using precise SV only with moderate filters (PE + SR ≥ 5), LUMPY detected 8563 duplications and 230,497 deletions while Manta detected 24,088 duplications and 320,374 deletions. A combined data set with 244,876 deletions and 8677 duplications (totaling 253,553, translating into an average of 1393 CNV per animal) was derived from the intersection of the LUMPY and Manta sets after removal of variants shorter than 50 bp or longer than 3 Mb. The combined data set had more observations than the LUMPY data set (which had fewer raw CNV) because
for some individuals, many short CNV from Manta intersected with few long CNV from LUMPY.

The CNV were distributed across the 29 autosomes as shown in Fig. 1. A vast majority of the CNV (96.6%) were losses. This is not unexpected, because all CNV detection methods suffer from an inherent deficiency in detecting insertions. In the case of CNV detection using WGS data, this limitation is even more pronounced with PE methods, because they detect insertions when the mapped reads are at a distance shorter than the fragment length, so they are not able to detect insertions larger than the insert size of the reference library [43]. This has also been supported by the observation that recall percentage is lower than 2 and 5% for medium (1–100 kb) and large (100 kb-1 Mb) duplications, respectively, for most of the SV-calling algorithms currently in use, including Manta and LUMPY used in this study [44].

Overall, the mean CNV length was about 3.3 kb, with a median of 1.3 kb. The distribution of the lengths of the CNV for each population are shown in Fig. 2 by CNV length category. A summary of the descriptive statistics of the CNV for the populations are given in Table 1. Most of the CNV losses (99.92%) were less than 100 kb long while 6.3% of CNV gains were longer than 100 kb. Despite the overwhelming proportion of losses over gains, there were more CNV gains observed over 100 kb than losses. Similarly, only 1.04% of the loss CNV were longer than 10 kb, while almost one-quarter (22.99%) of all gain CNV were over 10 kb. As a result, CNV gains were longer than CNV losses and had larger range in length. Deletions and duplications averaged about 2.3 and 31.5 kb long, with median lengths of 1.3 and 1.4 kb, respectively. There were no significant differences in the distribution of CNV across the five populations as shown in the percentile and sample QQ plots in Fig. 3.

Population CNV differentiation

Analysis of population differentiation ($V_{ST}$) as described by Redon et al. [11] showed that several CNV were highly differentiated between and across the populations. Some of these CNV overlapped with genes of importance in goats. Results for the pairwise population $V_{ST}$ tests and the $V_{ST}$ test across all the populations with their respective 99th percentile CNV $V_{ST}$ thresholds are given in Supplementary Table 1 (Additional file 1). $V_{ST}$ values for the pairwise tests are given in Supplementary Figures 1–10 (Additional file 2). The $V_{ST}$ values for genes that were in CNV that were highly differentiated across all populations are shown in Fig. 4. The gene DST was in a CNV with a very high $V_{ST}$ threshold across all the populations. DST has been associated with herpes virus and respiratory disease (BRD) in cattle [45]. Some CNV were highly differentiated both between and across populations. CNV with high differentiation between only some populations include the CNV corresponding to the genes BCO2, CCSE1 (FAM190A), COL24A1, CPNE4, CWC22, IMMP2L, KBTBD12, LAMA3, NAAALDL2, RFX3, SEMA3D, SLC2A13, STPG2 (C4orf37), TAF4 (FAM19A2), TMEM117, TMEM161B and VPS13B. The rest of the genes were in CNV that were highly differentiated across all populations.

Number and distribution of CNV regions (CNVR)

The lists of CNV regions (CNVR) by population are given in Supplementary Table 2 (Additional file 1) and their locations on the goat genome are shown in Fig. 5. Plots of the CNVR for each breed (with more than 2 animals) are given in Supplementary Figures 12 to 40 (Additional file 2). Descriptive statistics of the CNVR for each population are given in Supplementary Table 3 (Additional file 1) while a distribution of CNVR by size and populations is given in Fig. 6. Over 92% of the CNVR were copy losses. There was a wide variation in the number and sizes of the CNVR between and among populations.
the populations. The fraction of copy gains or gains and losses was highest in the group of CNVR of at least 10 kbp, with 25% copy gains and 19% for losses/gains (Fig. 6).

**Number and distribution of global CNVR**

Global CNVR for different levels of SV filter parameters are given in Supplementary Figures 41 to 64 (Additional file 2). Only the PE and SR filter levels and the CNV length cut-off point affected CNVR coverage. Inclusion of imprecise SV led to an increase in the proportion of called duplications, but the additional duplications were much longer than the upper cut-off point for CNV length. A total of 6231 global CNVR were found across all animals. A list of the global CNVR is given in Supplementary Table 4 (Additional file 1) and a summary is given in Table 2. There were 5742 CNVR with copy losses, 280 with copy gains and 209 with both copy losses and gains in different individuals. The locations of the global CNVR are given in Fig. 7. CNVR with both gains and losses were much longer (mean 185.8 kb) and constituted a significant proportion of the total CNVR coverage (65.6%). Sixteen of these were longer than 1 Mb (on chromosomes 1, 2, 6, 7, 12, 14 (two regions), 17, 19, 21, 23 (two regions), 27 and 29).

Overall, the CNVR covered about 59.2 Mb of the goat genome. Previous work on genome-wide CNV discovery

| Table 1 Descriptive statistics of CNV and CNV length for each population |
|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| Population      | Number of samples | CNV State   | Number | CNV length (bp) | Mean  | Median | Minimum | Maximum |
| BOE             | 9                | Loss            | 9079   | 2227.1          | 1326   | 67     | 254,129 |
|                 |                  | Gain            | 331    | 20,165.9        | 1500   | 161    | 631,262 |
|                 |                  | Overall         | 9410   | 2858.1          | 1330   | 67     | 631,262 |
| EAF             | 80               | Loss            | 108,051| 2244.7          | 1293   | 52     | 2,161,018|
|                 |                  | Gain            | 3544   | 30,979.2        | 1316.5 | 118    | 2,777,398|
|                 |                  | Overall         | 111,595| 3157.2          | 1293   | 52     | 2,777,398|
| MAD             | 27               | Loss            | 31,426 | 2475.3          | 1295   | 84     | 2,069,909|
|                 |                  | Gain            | 1078   | 28,384.1        | 1446   | 84     | 1,660,243|
|                 |                  | Overall         | 32,504 | 3334.6          | 1296   | 84     | 2,069,909|
| SAF             | 44               | Loss            | 67,099 | 2368.9          | 1285   | 51     | 2,539,701|
|                 |                  | Gain            | 2514   | 31,000.7        | 1192   | 101    | 1,959,154|
|                 |                  | Overall         | 69,613 | 3402.9          | 1283   | 51     | 2,539,701|
| WAF             | 22               | Loss            | 29,221 | 2491.4          | 1280   | 52     | 2,457,795|
|                 |                  | Gain            | 1210   | 40,255.3        | 1234   | 65     | 2,788,546|
|                 |                  | Overall         | 30,431 | 3993            | 1280   | 52     | 2,788,546|
in goats using SNP data done by Liu et al. [18] showed that CNVR cover approximately 262 Mb of the goat genome. Of the 978 CNVR reported in that study, 540 CNVR intersected with 819 CNVR identified in our study. The amount of the overlap between the CNVR in the two studies was 217.1 Mb, covering 38.6 Mb (65.1%) in this study, and 194.2 Mb (74.1%) in the other study.

**Common and rare CNVR**

Most of the CNVR (>95.9%) were found in at least 2 breeds. Out of the 6231 CNVR, 98 (1.6%) were present in all the 34 breeds and 1790 (28.7%) were present in all the populations (Fig. 8a and b). The most frequent CNVR observed was on chromosome 6 from 115,822,332 bp to 115,825,687 bp with a frequency of 96.2%. There were 259 CNVR private to 30 breeds, and 1018 private to all 5 populations, distributed as shown in Fig. 8c and Fig. 8d. BOE (Tanzania and Zimbabwe), KEF (Ethiopia) and MLY (Tanzania) breeds had the highest numbers of private CNVR (20, 21 and 31, respectively).

**Functional annotation and gene enrichment analysis**

Functional annotation was carried out for genes in global and private CNVR. Up to 2980 genes overlapped with the 6321 CNVR identified in this study. Up to 755 of these genes formed 24 clusters, with enrichment scores ranging from 0.0 to 1.89. Higher enrichment scores imply higher overrepresentation of the genes in the gene set for the gene enrichment term [46]. The top 3 clusters with the highest enrichment scores are given in Table 3 while the full list is given in Supplementary Table 5 (Additional file 1). The most significant GO terms identified in the analysis included retrograde endocannabinoid signaling; glutamatergic synapse; circadian entrainment; dopaminergic synapse; gastric acid secretion; long-term potentiation; salivary secretion; and calcium signaling pathway.

CNVR private to populations and breeds overlapped with 172 and 620 genes, respectively. The GO terms associated with these genes based on functional analysis are listed in Supplementary Table 6 (Additional file 1). The genes that overlapped with the CNVR private to breeds were not significantly enriched in biological processes, molecular functions and cellular components, while the ones that overlapped with the CNVR private to populations were significantly enriched \((P \leq 0.05)\) with such terms as aldosterone synthesis and secretion; glucagon signaling pathway; insulin secretion; glutamatergic synapse; thyroid hormone synthesis; gastric acid secretion and phosphatidylinositol signaling system. The most common CNVR (chr11:115,822,332-115,825,687) includes the gene TMEM129 (transmembrane protein 129) that has been reported to be responsible for ubiquitination and proteasome-mediated degradation of misformed or unassembled proteins in the cytosol [47–49], and belongs to a network responsible for cellular assembly and organization, cellular function and maintenance, and cell cycle [50].

**Discussion**

This study identified CNV and CNVR in the goat genome using WGS data. Use of WGS for CNV detection is highly encouraged, because it overcomes many of the shortcomings of the other CNV detection methods such as the ones using array CGH and SNP data [19–21]. Genome-wide studies to discover CNV have already been done in other domesticated species, such as in *Sus scrofa* [51], *Bos taurus* [38, 52] and *Felis catus* [39]. Here we provide a first glimpse of the goat genome CNV map at a dense genome coverage, using animals from 34 diverse breeds from the African continent. This addition is an important contribution, as goats are an important
source of income and high-quality animal protein for small holder farmers in Africa.

We used two software suites (LUMPY [30] and Manta [37]) for detecting SV to increase our confidence in the SV calls. Both software packages use split read and read-pair methods. They complement each other in that LUMPY makes use of read depth methods, while Manta draws heavily on genome assembly methods. Taking the intersection of SV calls from the two methods gives us confidence that the

![Fig. 4](image.png)

**Fig. 4** Population CNV differentiation, estimated by \( V_{ST} \) computed across all populations, plotted for each chromosome. The dotted line represents the \( V_{ST} \) threshold value for this test (0.601)

![Fig. 5](image.png)

**Fig. 5** Location of the CNVR for the 29 autosomes by population. The outermost numbers are the autosomes, and the other numbers are the start and end positions of each autosome.
number of false positives in the SV calls was kept to a minimum, although this means that some true SV were possibly filtered out.

This study has shown that there are wide variations in the number and sizes of CNV in the goat genome between chromosomes, individuals and breeds. However, considering the small and variable numbers of samples within breeds, breed comparisons are not particularly meaningful. The results suggest that there are negligible differences in the sizes of CNV between populations. Some of the CNV displayed large differences between populations, suggestive of population-specific selective pressures.

Våge and Boman [54] reported that BCO2 is associated with the accumulation of carotenoids in the adipose tissue of sheep, leading to the yellow fat syndrome. The quality of semen (including total sperm motility, average path velocity and beat cross frequency) in Holstein-Friesian bulls has been associated with CCSR1 (FAM190A) as well as FAM155A [55]. GNRHR has been associated with number of days to first service after calving in dairy cattle [56] while IMMP2L is associated with cow conception rate [57]. The partial deletion of LAMA3 is responsible for epidermolysis bullosa in horses [58]; NAALADL2 is believed to be responsible for immune homeostasis [59], and TAF2 (FAM19A2) is believed to be responsible for the regulation of feed intake and metabolic activities in mice [60]. Yamano et al. [61] reported that

### Table 2 CNVR summary statistics for each CNV state based on CNV occurring in at least 2 individuals

| Copy state | Number of CNVR | Length (bp) | CNVR coverage (bp) |
|------------|----------------|-------------|---------------------|
|            |                | Mean        | Median              | Minimum | Maximum   |              |
| Loss       | 5742           | 3041.3      | 1140.5              | 52      | 1,177,087 | 17,463,236  |
| Gain       | 280            | 10,377.9    | 1008.0              | 302     | 236,347   | 2,905,806   |
| Both       | 209            | 185,755.2   | 1731.0              | 616     | 2,956,746 | 38,822,839  |
| Overall    | 6231           | 9499.6      | 1157.0              | 52      | 2,956,746 | 59,191,881  |
TOMM70 is responsible for integral mitochondrion proteins and for metabolism.

Functional annotation and clustering analysis revealed that the CNVR identified in the study have genes that are significantly enriched with many biological processes, molecular functions and cellular components, some of the most significant of which are retrograde endocannabinoid signaling, circadian entrainment and long-term potentiation. The retrograde endocannabinoid signaling system is a complex and diverse regulator of synaptic function [62], and is responsible for many diseases in the nervous system and peripheral organs. In the human genome, this system is widely considered as a potential target for treating conditions such as alcoholism [63]. A CNVR in the cannabinoid receptor 2 (CNR2) region has been reported in the human genome, but its effect has not been fully characterized [64]. Zajkowska et al. [65] suggested that there is need to explore genetic variation in the system from the perspective of copy number of variations.

Circadian entrainment is an important aspect of animal behavior and adaptation, especially considering the wide range of environmental conditions the animals are exposed to. An example of goat adaptation to the environment is their ability to rapidly change the size of their foreguts in response to changes in the environment [66]. Goats tend to be active during some parts of the day only [67], and this varies with season [67], suggesting a considerable amount of circadian entrainment. The increased importance of the biological process “response to stimulus” (GO:0050896) in the highly differentiated CNV may also support the hypothesis of the importance of circadian entrainment in goats.

**Conclusions**

This study presents the first fine CNV map of the African goats based on WGS data. This information will prove invaluable for further improvement of goats, especially on African continent, as more phenotype data becomes available, through CNV or CNVR association analyses and other approaches.

**Methods**

**Sample description**

The data used in this study was generated from 182 goats representing 34 breeds from 9 Sub-Saharan African countries (Ethiopia, Kenya, Madagascar, Malawi, Mali, Mozambique, Tanzania, Uganda, and Zimbabwe), and these countries were grouped into four populations based on geographic locations and a fifth population of Boer goats obtained in Tanzania and Zimbabwe. The Boer goat is a special breed widely used in Africa and much of the world [68]. The samples were previously genotyped using the Illumina Goat SNP50 BeadChip [69] as described by Bertolini et al. [70], Cardoso et al. [71]
and Colli et al. [72], and some of them were also used for detection of CNV using 50 K SNP chip data, as reported by Liu et al. [18]. A list of the breeds, populations and samples sizes used in the analysis is given in Table 4.

Sample processing was done by Edinburgh Genomics using the Edinburgh Clinical Genomics method. This approach uses Illumina SeqLab products and services, including, Illumina TruSeq library preparation, Illumina cBot2 cluster generation, Illumina HiSeqX sequencing, Hamilton Microlab STAR integrative automation, and Genologics Clarity LIMS X Edition as outlined in Supplementary Table 7 (Additional file 1). Quality control information for the samples is given in Supplementary Table 8 (Additional file 1).

### Sequence alignment

Sequence alignment was done using the Burrow-Wheeler Alignment (BWA) tool version 0.7.13-r1126 with the maximal exact matches (MEM) “mem” option [73]. The reads were aligned to the ARS1 *Capra hircus* (goat) reference assembly (https://www.ncbi.nlm.nih.gov/assembly/GCF_001704415.1) [74]. The aligned reads were processed into binary sequence alignment map (BAM) format using SAMTools version 1.8 [75].

### Detection of SV

SV were detected using LUMPY version 0.2.13–85-gc1bca1 and Manta version 1.5.1, which are two of the most used algorithms for detecting SV. In LUMPY, the “lumpyexpress” script was used. This script runs automated breakpoint detection for standard analyses. It uses SAMBLASTER [76] to extract split and discordant reads from BWA-MEM-aligned Binary Sequence Alignment Map (BAM) files. Default options were used, including minimum non-overlap and minimum sample weight set to 20 and 4, respectively. In Manta, the “configManta.py” script was used to process each sample, with default options including minimum variant candidate size (8); minimum candidate spanning count (3); minimum scored variant size (50); minimum diploid variant score (10); minimum diploid variant score pass point (20); minimum somatic score (10); and minimum somatic score pass point (30). The “runWorkflow.py” scripts were run in parallel to extract the SV for each sample.

### Post-processing of SV

SV from LUMPY were genotyped with svtyper version 0.6.1 [77], which uses a Bayesian maximum likelihood algorithm to determine the most likely genotype of each base-pair. Variant call format (VCF) files from the two
software packages were converted to browser extensible data (BED) format for downstream analysis using svtools version 0.5.0 [78]. Various levels of SV post-processing parameters were used to come up with the CNV calls from the SV calls. The parameters included: 1) precision of SV calls (whether imprecise SV were included in computation of the CNV calls); 2) point of application of the lower SV length cut-off point (before or after merging Manta and LUMPY SV); 3) stringency of the SV call filters (low, medium, and high stringency); and 4) upper SV length cut-off (3 or 10 Mb). Stringency of SV call filters was in terms of the number of PE and SR required as evidence supporting an SV. Consensus SV were obtained by identifying the intersection of the SV from LUMPY and Manta using BEDTools version 2.26.0 [79] with default settings.

### Derivation of copy number variations

CNV were defined as SV duplications and deletions longer than 50 bp [80]. SV longer that 3 Mb were also

### Table 3 Functional annotation clusters of the genes found in the global CNVR based on analysis in DAVID

| Cluster (Enrichment score, database) | Enrichment term                                    | Gene count | p-value |
|-------------------------------------|---------------------------------------------------|------------|---------|
| 1 (1.89, KEGG_PATHWAY)             | chx04020: Calcium signaling pathway                | 31         | 0.001   |
|                                     | chx04970: Salivary secretion                        | 16         | 0.011   |
|                                     | chx04022: cGMP-PKG signaling pathway                | 25         | 0.016   |
|                                     | chx04270: Vascular smooth muscle contraction        | 19         | 0.037   |
|                                     | chx04261: Adrenergic signaling in cardiomyocytes    | 21         | 0.046   |
|                                     |                                                   |            |         |
| 2 (1.62, KEGG_PATHWAY)             | chx04723: Retrograde endocannabinoid signaling      | 24         | 0.000   |
|                                     | chx04724: Glutamateric synapse                      | 24         | 0.000   |
|                                     | chx04713: Circadian entrainment                    | 21         | 0.001   |
|                                     | chx04728: Dopaminergic synapse                      | 25         | 0.002   |
|                                     | chx04971: Gastric acid secretion                    | 16         | 0.004   |
|                                     | chx04720: Long-term potentiation                    | 14         | 0.009   |
|                                     | chx04970: Salivary secretion                        | 16         | 0.011   |
|                                     | chx04925: Aldosterone synthesis and secretion       | 15         | 0.014   |
|                                     | chx04730: Long-term depression                      | 13         | 0.014   |
|                                     | chx04540: Gap junction                              | 16         | 0.019   |
|                                     | chx04750: Inflammatory mediator regulation of TRP channels | 18         | 0.020   |
|                                     | chx04921: Oxytocin signaling pathway                | 23         | 0.028   |
|                                     | chx04922: Glucagon signaling pathway                | 17         | 0.028   |
|                                     | chx04972: Pancreatic secretion                      | 16         | 0.033   |
|                                     | chx04270: Vascular smooth muscle contraction        | 19         | 0.037   |
|                                     | chx04725: Cholinergic synapse                       | 18         | 0.043   |
|                                     | chx04911: Insulin secretion                         | 14         | 0.053   |
|                                     | chx04726: Serotonergic synapse                      | 17         | 0.072   |
|                                     | chx04915: Estrogen signaling pathway                | 15         | 0.091   |
|                                     | chx04961: Endocrine and other factor-regulated calcium reabsorption | 8         | 0.094   |
|                                     | chx04912: GnRH signaling pathway                    | 13         | 0.131   |
|                                     | chx04918: Thyroid hormone synthesis                 | 11         | 0.137   |
|                                     | chx04924: Renin secretion                           | 10         | 0.161   |
|                                     | chx04611: Platelet activation                       | 16         | 0.273   |
|                                     | chx04916: Melanogenesis                            | 11         | 0.458   |
|                                     | chx04310: Wnt signaling pathway                     | 14         | 0.599   |
|                                     |                                                   |            |         |
| 3 (1.14, KEGG_PATHWAY)             | chx05204: Chemical carcinogenesis                  | 12         | 0.033   |
|                                     | chx00980: Metabolism of xenobiotics by cytochrome P450 | 10         | 0.075   |
|                                     | chx00982: Drug metabolism - cytochrome P450         | 10         | 0.075   |
|                                     | chx00830: Retinol metabolism                        | 9          | 0.152   |
filtered out, because putative CNV in the goat genome are usually much shorter than this length. Visualization of the SV was done using R [81] package circlize version 0.4.7 [82].

**Population CNV differentiation**

A measure of population differentiation ($V_{ST}$) as described by Redon et al. [11] was computed based on normalized read count values for each CNV, similar to the method used in PECNV as described by Liu et al. [83], which was in turn based on clustering algorithms described by Cridland et al. [84] and transposable element detection algorithms described by Rogers et al. [85]. Read count values were corrected for size of the consensus CNV, batch effect, variable GC content and genomic mappability as described by Liu et al. [83]. Regional and batch effect correction was done by computing reads per kb per million mapped reads (RPKM) as described by

| Population   | Population Code | Breed name  | Breed Code | Country     | Number of samples |
|--------------|-----------------|-------------|------------|-------------|-------------------|
| Boer         | BOE             | Boer        | BOE        | Tanzania    | 2                 |
| Boer         | BOE             | Boer        | BOE        | Zimbabwe    | 7                 |
| East African | EAF             | Abergelle   | ABR        | Ethiopia    | 6                 |
| East African | EAF             | Galla       | GAL        | Kenya       | 7                 |
| East African | EAF             | Gogo        | GOG        | Tanzania    | 7                 |
| East African | EAF             | Gumez       | GUM        | Ethiopia    | 4                 |
| East African | EAF             | Keffa       | KEF        | Ethiopia    | 7                 |
| East African | EAF             | Landin      | LND        | Mozambique  | 5                 |
| East African | EAF             | Maasai      | MAA        | Tanzania    | 7                 |
| East African | EAF             | Manica      | MAN        | Mozambique  | 3                 |
| East African | EAF             | Malya       | MLY        | Tanzania    | 7                 |
| East African | EAF             | Norwegian   | NRW        | Tanzania    | 3                 |
| East African | EAF             | Pare White  | PRW        | Tanzania    | 6                 |
| East African | EAF             | Saanen      | SAA        | Tanzania    | 4                 |
| East African | EAF             | Small East African | SEA | Kenya    | 7                 |
| East African | EAF             | Small East African | SEA | Mozambique | 6                 |
| East African | EAF             | Sonjo       | SNJ        | Tanzania    | 2                 |
| East African | EAF             | Woyito Guji | WYG        | Ethiopia    | 7                 |
| Madagascar   | MAD             | Androy      | AND        | Madagascar  | 4                 |
| Madagascar   | MAD             | Diana       | DIA        | Madagascar  | 3                 |
| Madagascar   | MAD             | Menabe      | MEN        | Madagascar  | 7                 |
| Madagascar   | MAD             | Sofia       | SOF        | Madagascar  | 6                 |
| Madagascar   | MAD             | SudOuest    | SOU        | Madagascar  | 7                 |
| Southern African | SAF    | Balaka-Ulongwe | BAW | Malawi    | 2                 |
| Southern African | SAF    | Dedza       | DZD        | Malawi      | 4                 |
| Southern African | SAF    | Lilongwe    | LGW        | Malawi      | 3                 |
| Southern African | SAF    | Mashona     | MSH        | Zimbabwe    | 7                 |
| Southern African | SAF    | Matebele    | MTB        | Zimbabwe    | 7                 |
| Southern African | SAF    | Nsanje      | NSJ        | Malawi      | 6                 |
| Southern African | SAF    | Thyo       | THY        | Malawi      | 7                 |
| West African | WAF             | Guerra      | GUE        | Mali        | 6                 |
| West African | WAF             | Maure       | MAU        | Mali        | 1                 |
| West African | WAF             | Naine       | NAI        | Mali        | 5                 |
| West African | WAF             | Peulh       | PEU        | Mali        | 1                 |
| West African | WAF             | Soudanaise  | SDN        | Mali        | 7                 |
| West African | WAF             | Targui      | TAR        | Mali        | 2                 |
Mortazavi et al. [86], where $\text{RPKM} = \frac{10^6 \text{RC}}{\text{TRC} \times \text{S}}$, where RC is the read count of a region, S is the size of the region and TRC is the total number of mapped reads in the library. GC content and mappability correction was done on the RPKM using the formula used by Yoon et al. [29], where adjusted read count is given by $\text{RPKM}_{\text{mGC}}$, where $mGC$ is the median GC content of all regions with the same read count and $m$ is the median GC of all regions. This approach is similar to the read depth approaches used in CNVnator [28] and in CNVcaller [87]. The normalized read count values were treated as proxies of log R ratio (LRR) values normally obtained from array analysis. As defined by Redon et al. [11], $V_{ST}$ was computed as $\frac{V_{T}-V_{S}}{V_{T}}$, where $V_T$ is the variance in LRR among all unrelated individuals and $V_S$ is the average variance in LRR within each population. CNV $V_{ST}$ testing was done pairwise (for each combination of two populations) and (separately) across all the 5 populations. CNV with $V_{ST}$ values above the 99th percentile of all $V_{ST}$ values for each comparison were treated as being highly differentiated. We searched for these highly differentiated CNV in the Golden Helix Genome Browse® software (version 3.0.0) (https://www.goldenhelix.com/) using the ARS1 caprine genome reference assembly to identify the genes in the CNV.

Determination of CNV regions

CNV regions (CNVR) were obtained by merging CNV that overlapped by at least 1bp within populations (population CNVR) and across all the individuals (global CNVR) using the “merge” function in BEDTools version 2.26.0 [79].

CNVR functional annotation and gene enrichment analysis

A list of genes for the goat genome was downloaded from the NCBI website (https://www.ncbi.nlm.nih.gov/gene). The Database for Annotation, Visualization, and Integrated Discovery (DAVID) Bioinformatics Resources (version 6.8) [88–90] was used to identify if genes in the CNVR have significant biological, cellular or molecular function. Functional analysis was done using default parameters, with significance of enriched terms determined at $P \leq 0.05$. Further information about various genes was obtained from the GeneCards (www.genecards.org) database.

Acknowledgements

The VarGoats Consortium - http://www.goatgenome.org/vargoats.html Whole genome sequencing libraries were prepared and sequenced by Edinburgh Genomics and funded via Biotechnology and Biological Sciences Research Council grant research (BBS/OS/GC/0000125) ‘Reference genome and population sequencing of African goats’ awarded to The Roslin Institute.

Authors’ contributions

CPVD, BDR, GM and JS conceived the idea of this work. ELK and the VarGoats Consortium carried out the sequencing of the samples and pre-processing of the sequence data. WN, BDR, CPVT, GM and JS carried out the analysis and drafted the manuscript. WN, GM, MW, IJB, THG, HAM, HN, ELK, MMG, ML, GEL, CPVT, BDR and JS reviewed the manuscript. The authors read and approved the final manuscript.

Funding

This research was done with funding from the Feed the Future project: Improving Livestock Productivity through Enhanced Breeding Programs (Agricultural Research Service/USDA Project No: 8042-31000-001-20S). The data were collected by the African Goat Improvement Network (AGIN) (https://www.ars.usda.gov/office-of-international-research-programs/ftf-livestock-improvement/). This research used resources provided by the SCINet project of the USDA Agricultural Research Service, ARS project number 0500–00093–001-00-D.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-021-07703-1.

Additional file 1 Supplementary Table 1 Genes in CNV with $V_{ST}$ values above the 99th percentile for each comparison. Supplementary Table 2 List of CNVR by population. Supplementary Table 3 Summary of CNVR by population. Supplementary Table 4 List of global CNVR with frequency above 1%. Supplementary Table 5 Functional annotation and clustering analysis for global CNVR. Supplementary Table 6 Functional analysis for CNVR private to breeds and to populations. Supplementary Table 7 Sample processing details. Supplementary Table 8 Quality control details for the samples used in the study. Additional file 2 Supplementary Figures 1–10 CNV differentiation between the following populations, respectively: Boer and East African; Boer and Madagascar; Boer and Southern African; Boer and West African; East African and Madagascar; East African and Southern African; East African and West African; Madagascar and Southern African; Madagascar and Western Africa.

Acknowledgements

This research used resources provided by the SCINet project of the USDA Agricultural Research Service, ARS project number 0500–00093–001-00-D.
Availability of data and materials
The dataset supporting the conclusions of this article is available at the National Center for Biotechnology Information Sequence Read Archive as detailed in Additional file 1 (Supplementary Table 7).

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that there are no competing interests.

Author details
1. University of Natural Resources and Life Sciences, Vienna, Austria. 2. ULMong University of Agriculture and Natural Resources, LiLongwe, Malawi. 3. National Livestock Resources Research Institute, Tororo, Uganda. 4. Kyambogo University, Kyambogo, Uganda. 5. The Roslin Institute, University of Edinburgh, Edinburgh, Scotland, UK. 6. Animal Genomics and Improvement Laboratory, USDA-ARS, Beltsville, MD, USA. 7. National Agricultural Library, USDA-ARS, Beltsville, MD, USA.

Received: 24 February 2020 Accepted: 11 May 2021
Published online: 29 May 2021

References
1. Escaramís G, Docampo E, Rabionet R. A decade of structural variants: description, history and methods to detect structural variation. Brief Funct Genomics. 2015;14(5):305–14. https://doi.org/10.1093/bfgp/elv014.
2. Matsuzaki H, Wang PH, Hu J, Rava R, Fu GK. High resolution discovery and confirmation of copy number variants in 90 Yoruba Nigerians. Genome Biol. 2009;10:1–18.
3. Fan S, Meyer A. Evolution of genomic structural variation and genomic architecture in the adaptive radiations of African cichlid fishes. Front Genet. 2014;5(JUN):163.
4. Conrad DF, Hurles ME. The population genetics of structural variation. Nat Rev Genet. 2007;8(7):530–6. https://doi.org/10.1038/nrg2042.
5. Ho SS, Urban AE, Mills RE. Structural variation in the sequencing era. Nat Rev Genet. 2019;20(13):171–89. https://doi.org/10.1038/s41587-019-0180-9.
6. Sebat J, Levy D, McCarthy SE. Rare structural variants in schizophrenia: one disorder, multiple mutations; one mutation, multiple disorders. Trends Genet. 2009;25(12):528–35. https://doi.org/10.1016/j.tig.2009.10.004.
7. Sebat J, Laskihi B, Troge J, Alexandor J, Young J, Lundin P, et al. Large-scale copy number polymorphism in the human genome. Science. 2004;305(5683):525–8. https://doi.org/10.1126/science.109198.
8. Iafrite AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, et al. Detection of large-scale variation in the human genome. Nat Genet. 2004;36(9):949–51. https://doi.org/10.1038/ng1416.
9. Feuk L, Marshall CR, Wintle RF, Scherer SW. Structural variants: changing the landscape of chromosomes and design of disease studies. Hum Mol Genet. 2006;15(suppl_1):R57–66.
10. Huwang CR, Schneider AM, Lu Y, Niranjan T, Shen P, Robinson MA, et al. Mobile interspersed repeats are major structural variants in the human genome. Cell. 2010;141(1):171–82. https://doi.org/10.1016/j.cell.2010.05.026.
11. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, et al. Global variation in copy number in the human genome. Nature. 2006;444(7118): 444–54. https://doi.org/10.1038/nature05329.
12. Beckmann JS, Estivill X, Antonaraki SE. Copy number variants and genetic traits: closer to the resolution of phenotypic to genotypic variability. Nat Rev Genet. 2007;8(8):639–46. https://doi.org/10.1038/nrg2149.
13. Zarei M, MacDonald JR, Merico D, Scherer SW. A copy number variation map of the human genome. Nat Rev Genet. 2015;16(3):172–83. https://doi.org/10.1038/nrg38071.
14. Girirajan S, Rosenfeld JA, Coe BP, Parikh S, Friedman N, Goldstein A, et al. Phenotypic heterogeneity of genomic disorders and rare copy-number variants. N Engl J Med. 2012;367(14):1321–31. https://doi.org/10.1056/NEJMoa1200395.
cancer sequencing applications. Bioinformatics. 2016;32(8):1220–2. https://doi.org/10.1093/bioinformatics/btw710.

38. Bickhart DM, Liu GE, Eichler EE, Matukumalli LK, Song J, Van Tassel CP, et al. Copy number variation of individual cattle genomes using next-generation sequencing. Genome Res. 2012;22(4):778–90. https://doi.org/10.1101/gr.133657.111.

39. Genova F, Longeri M, Lyons LA, Bagnato A, Gandolfi B, Abdellin D, et al. First genome-wide CNV mapping in Felis catus using next generation sequencing data. BMC Genomics. 2018;19:1–2.

40. Scharli SS, Mohammadbadi M, Wu D-D, Esmailzadeh A. Detection of breed-specific copy number variations in domestic chicken genome. Genome. 2018;61(3):14–19. https://doi.org/10.1016/j.genome.2017.07.016.

41. Freedman AH, Grönau I, Schweizer RM, Ortega-Del Vecchyo D, Han E, Silva PM, et al. Genome sequencing highlights the dynamic early history of dogs. PLoS Genet. 2014;10(1):e1004016. https://doi.org/10.1371/journal.pgen.1004016.

42. Stella A, Nicolazzi EL, Van Tassell CP, Rothschild MF, Colli L, Rosen BD, et al. Refinement of the DST locus associated with bovine respiratory disease complex in Holstein calves. J Anim Sci. 2016(94(Suppl. 5):3832. https://doi.org/10.1093/banana/orw005.

43. Teo SM, Pawitan Y, Ku CS, Chia KS, Salmi A. Statistical challenges associated with detecting copy number variations with next-generation sequencing. Bioinformatics. 2012;28(21):2711–8. https://doi.org/10.1093/bioinformatics/bts335.

44. Kosugi S, Momozawa Y, Liu X, Terao C, Kubo M, Kamatani Y. Comprehensive identification of genome regions determining semen quality in Holstein-Friesian bulls using information theory. Anim Reprod Sci. 2018;192:206–15. https://doi.org/10.1016/j.anireprosci.2018.03.012.

45. Liu R, Hadley D, Liu J, Bucan M, Li M, Glessner J, et al. PennCNV: an integrated high-confidence shRNA screen identifies TMEM129 as an E3 ubiquitin ligase in ER-associated protein degradation. Nat Commun. 2014;5(1):117. https://doi.org/10.1038/ncomms9509.

46. Yu F, Shen XY, Fan L, Yu ZC. Genome-wide analysis of genetic variations with detecting copy number variations with next-generation sequencing. Front Genet. 2019;10:915. https://doi.org/10.3389/fgene.2019.00915.

47. Jiang J, Ma L, Prakapenka D, VanRaden PM, Cole JB, Da Y. A large-scale genome-wide association study in U.S. Holstein cattle. Front Genet. 2019;10:41721.

48. Yuan WY, Pasmooy AMG, Stellingens C, Jonkman MF. Enamel defects in carriers of a novel LAMA3 mutation underlying epidermolysis bullosa. Acta Derm Venereol. 2019;129(6):695–6. https://doi.org/10.2340/00015555-1341.

49. Onouchi Y. Genetics of Kawasaki disease: what we know and don’t know. Circ J. 2012;76(7):1581–6. https://doi.org/10.1253/circj/CJ-12-0568.

50. Yu F, Shen XY, Fan L, Yu ZC. Genome-wide analysis of genetic variations with detecting copy number variations with next-generation sequencing. Bioinformatics. 2018;34(14):2297–305. https://doi.org/10.1093/bioinformatics/btx677.

51. Yamano K, Matsuoka N, Tanaka K. The ubiquitin signal and autophagy: an orchestrated dance leading to mitochondrial degradation. EMBO Rep. 2016;17(3):300–16. https://doi.org/10.15252/embr.201541486.

52. Castello PE, Yount TJ, Chávez AE, Hashimotodani Y. Endocannabinoid signaling and synaptic function. Neurosci. 2012;76(1):170–81. https://doi.org/10.1016/j.neuroscience.2012.09.020.

53. Basavarajappa BS. The endocannabinoid signaling system: a potential target for next-generation therapeutics for alcoholism. Min Rev Med Chem. 2007;7(8):769–79. https://doi.org/10.2174/1389557077818378920.

54. Ishiguro H, Leonard CM, Sgro S, Onai K. Cannabinoid receptor gene variations in neuropsychiatric disorders. In: Muñoz-Rodríguez E, editor. Endocannabinoids: molecular, pharmacological, behavioral and clinical features. 1st ed. Baltimore: Bentham Science Publishers; 2013. p. 3–24. https://doi.org/10.2174/9781608050284113010006.

55. Zajkowska ZE, Englund A, Zunszain PA. Towards a personalized treatment in depression: endocannabinoids, inflammation and stress response. Pharmacogenomics. 2014;15(5):687–98. https://doi.org/10.2174/1389557148013600006.

56. Silinkove N. The physiological basis of adaptation in goats to harsh environments. Small Rumin Res. 2000;35(3):181–93. https://doi.org/10.1016/S0167-1987(00)00096-6.

57. Piccone G, Giannetto C, Casella S, Caola G. Seasonal change of daily motor activity rhythms in Capra hircus. Can J Anim Sci. 2008;88(3):351–5. https://doi.org/10.1014/1411.CAJAS7125.

58. Malan SW. The improved Boer goat. Small Rumin Res. 2000;36(2):162–70. https://doi.org/10.1016/S0921-4488(00)80164-1.

59. Tosser-Klopp G, Baudou P, Bouchez O, Calau B, Croijmans R, Dong Y, et al. Design and characterization of a 52K SNP chip for goats. PLoS One. 2014;9(1):e86227. https://doi.org/10.1371/journal.pone.0086227.

60. Bertolino F, Cardoso TF, Marras G, Nicolazzi EL, Rothschild MF, Amills M. Genome-wide patterns of homozygosity provide clues about the population history and adaptation of goats. Genet Sel Evol. 2018;50(1):59. https://doi.org/10.1186/s12711-018-0424-8.

61. Cardoso TF, Amills M, Bertolino F, Rothschild M, Marras G, Boink G, et al. Patterns of homozygosity in insular and continental goat breeds. Genet Sel Evol. 2018;50:1–11.

62. Colli L, Milanesi M, Talento A, Bertolino F, Chen M, Crisâ A, et al. Genome-wide SNP profiling of worldwide goat populations reveals strong partitioning of diversity and highlights post-domestication migration routes. Genet Sel Evol. 2018;50(1):58. https://doi.org/10.1186/s12711-017-0422-x.

63. Li H, Durbin R. Fast and accurate long-read alignment with burrows-wheeler transform. Bioinformatics. 2010;26(5):889–95. https://doi.org/10.1093/bioinformatics/btp698.

64. Bickhart DM, Rosen BD, Koren S, Sayre BL, Hastie AR, Chan S, et al. Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. Nat Genet. 2017;49(4):643–50. https://doi.org/10.1038/ng.3802.

65. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25(16):2078–9. https://doi.org/10.1093/bioinformatics/btp352.

66. Faust GG, Hall IM. SAMBLASTER: fast duplicate marking and structural variant read extraction. Bioinformatics. 2014;30(17):2503–5. https://doi.org/10.1093/bioinformatics/btu314.

67. Chiанг C, Layer RM, Faust GG, Lindberg MR, Rose DB, Garrison EP, et al. SpeedSeq: ultra-fast personal genome analysis and interpretation. Nat Methods. 2015;12(10):966–8. https://doi.org/10.1038/nmeth.3505.

68. Lanson DE, Abel HJ, Chi앙 C, Badve A, Das I, Eldred JM, et al. SVtools: Population-Scale Analysis of Structural Variation. Bioinformatics. 2019;35:4782.
79. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26(6):841–2. https://doi.org/10.1093/bioinformatics/btp033.

80. Liu GE, Bickhart DM. Copy number variation in the cattle genome. Funct Integr Genomics. 2012;12(4):609–24. https://doi.org/10.1007/s10142-012-0289-9.

81. R Development Core team. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for statistical Computing; 2016.

82. Gu Z, Gu L, Ellis R, Schlesner M, Brors B. Circize implements and enhances circular visualization in R. Bioinformatics. 2014;30(19):2811–2. https://doi.org/10.1093/bioinformatics/btu393.

83. Liu X, Li A, Xi J, Feng H, Wang M. Detection of copy number variants and loss of heterozygosity from impure tumor samples using whole exome sequencing data. Oncol Lett. 2018;16(4):4713–20. https://doi.org/10.3892/ol.2018.9150.

84. Cridland JM, Macdonald SJ, Long AD, Thornton KR. Abundance and distribution of transposable elements in two drosophila QTL mapping resources. Mol Biol Evol. 2013;30(10):2311–27. https://doi.org/10.1093/molbev/mst129.

85. Rogers RL, Cridland JM, Shao L, Hu TT, Andolfatto P, Thornton KR. Landscape of standing variation for tandem duplications in drosophila yakuba and drosophila simulans. Mol Biol Evol. 2014;31(7):1750–66. https://doi.org/10.1093/molbev/mst129.

86. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods. 2008;5(7):621–8. https://doi.org/10.1038/nmeth.1226.

87. Wang X, Zheng Z, Cai Y, Chen T, Li C, Fu W, et al. CNVcaller: highly efficient and widely applicable software for detecting copy number variations in large populations. GigaScience. 2017;6(12):1–12. https://doi.org/10.1093/gigascience/gix115.

88. Dennis G, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: pathways toward the comprehensive functional analysis of large gene lists. Genome Biol. 2003;4(5):P3. https://doi.org/10.1186/gb-2003-4-5-p3.

89. Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 2009;37(1):1–13. https://doi.org/10.1093/nar/gkn923.

90. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2009;4(1):44–57. https://doi.org/10.1038/nprot.2008.211.

Publisher’s Note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.