scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Whole-genome sequences of 240 indigenous African cattle from Egypt, Uganda, and South Africa

Njabulo Dlamini[1,17], Junxin Gao[2,17] ✉, Catarina Ginja[3,4], Juha Kantanen[5], Nasser Ghanem[6], Donald R. Kugonza[7], Mahlako Makgahlela[1,8], Ahmed Elnahas[9], Avhashoni Zwane[10], Barbara Mugwanya Zawedde[11], Christine Nakkazi[11], Generous Behabura[7], Khaniysani Nxumalo[12], Maano Malima[1], Mohamed Ali Radwan[6], Mohamed Hamada Elsawy[13], Morris Agaba[7], Nadia Hamdi Fahim[6], Rana Atef Khfagy[14], Rania Agamy[6], Rodney Okwasiimire[5,15,16], Sarah Waibi[15], Simon Lashmar[1] & Richard P. M. A. Crooijmans[2] ✉

Indigenous cattle are central to livestock production in Africa, valued for their adaptability to harsh tropical environments despite lower productivity than commercial breeds. Genome analyses offer critical insights into the genetic potential for enhancing both resilience and productive traits, supporting the advancement of worldwide cattle farming systems. Here, we generated whole-genome sequence data for 240 indigenous cattle representing breeds from distinct agro-climatic regions in Egypt, Uganda, and South Africa. The dataset comprises over ten terabytes of paired-end reads generated using the Illumina NovaSeq. 6000 platform, with an average genome coverage of approximately 10×. Post-filtering reads were mapped to the ARS-UCD1.2 reference genome with a mean mapping rate of 99.2% (range: 64.5–99.9%). Variant calling identified ~43 million SNPs and 6 million indels (≤50 bp) unevenly distributed across the genome. Functional annotation indicated that many variants were located within or near known genes. This comprehensive genomic resource provides a foundation for future studies of genetic diversity, breed identity, population structure, local adaptation, breed-specific traits, or strategies for global cattle conservation.

Background & Summary

Livestock farming is crucial in developing countries, contributing significantly to livelihoods and ensuring food and nutritional security[1]. Cattle are a vital part of economies across various regions in Africa as they act as savings and insurance, provide food security through meat and milk, generate income, and supply multiple applications for horns, hides, and skins. Additionally, these animals enable households to utilize common-property resources (like communal pastures), and meet social commitments (such as in ceremonies or as part of dowry)[2,3].

[1]Agricultural Research Council, Animal Production, Irene, South Africa. [2]Animal Breeding and Genomics, Wageningen University & Research, Wageningen, The Netherlands. [3]CIBIO, Centro de Investigação em Biodiversidade e Recursos Genéticos, InBIO Laboratório Associado, Campus de Vairão, Universidade do Porto, Vairão, Portugal. [4]BIOPOLIS, Program in Genomics, Biodiversity and Land Planning, CIBIO, Campus de Vairão, Vairão, Portugal. [5]Natural Resources Institute Finland, Jokioinen, Finland. [6]Animal Production Department, Faculty of Agriculture, Cairo University, Giza, Egypt. [7]College of Agricultural and Environmental Sciences, Makerere University, Kampala, Uganda. [8]Department of Animal, Science, University of the Free State, Bloemfontein, South Africa. [9]Animal Production Department, Faculty of Agriculture, Sohag University, Sohag, Egypt. [10]Department of Genetics, University of Pretoria, Hatfield, Pretoria, South Africa. [11]National Agricultural Research Organization, Mukono Zonal Agricultural Research and Development Institute, Mukono, Uganda. [12]Mpumalanga Department of Agricultural Rural Development, Land Reform and Environmental Affairs, Animal Research Directorate, Nooitgedacht Research Development Centre, Mpumalanga, South Africa. [13]Department of Cattle, Animal Production Research Institute, Agriculture Research Center, Dokki, Giza, Egypt. [14]Department of Genetics, Faculty of Agriculture, Cairo University, Giza, Egypt. [15]College of Veterinary Medicine, Animal Resources and Biosecurity, Makerere University, Kampala, Uganda. [16]Department of Agricultural Sciences, University of Helsinki, Helsinki, Finland. [17]These authors contributed equally: Njabulo Dlamini, Junxin Gao. ✉e-mail: junxin.gao@wur.nl; richard.crooijmans@wur.nl

Archaeological and genetic evidence traces the introduction of cattle into North Africa from the Levant at around 8,000 years before present (yBP), with subsequent expansions southward into West and East Africa (~4,500 yBP) and ultimately into South Africa (~2,000 yBP)[4–6]. The earliest African cattle were taurine (*Bos taurus*), domesticated from the Near Eastern aurochs approximately 10,000 yBP[7]. Subsequent introduction and admixture with indicine cattle (*Bos indicus*, zebu) from South Asia increased genetic diversity[8]. Indicine cattle first emerged in South Asia around 8,000 yBP, approximately two millennia after the taurine cattle in the Near East, and are thought to have descended at least in part from *Bos primigenius namadicus*, a South Asian aurochs subspecies[9].

African cattle breeds exhibit notable adaptability to tropical diseases, heat stress, and limited forage quality, despite their typically lower productivity compared to high-input commercial breeds such as Angus and Holstein-Friesian[1,10,11]. Centuries of local adaptation and breeding have resulted in considerable genetic and phenotypic diversity, including traits conferring resistance to environmental stressors[12–14]. Through selective breeding and crossbreeding, cattle breeds have achieved high productivity (e.g., Black Angus, Holstein-Friesian, Simmental), but many African cattle remain indigenous, exhibiting lower productivity yet excelling in resilience and disease resistance[10].

According to the FAO Domestic Animal Diversity Information System (DAD-IS), around 3,000 cattle breeds are registered globally, with approximately 600 breeds found across Africa, predominantly in sub-Saharan regions (FAO DAD-IS, 2024). Of these, around 150 are classified as indigenous, encompassing taurine, zebu, and hybrid types such as Sanga. Sanga cattle represent stabilized ancient hybrids between *Bos taurus* and *Bos indicus* that arose following secondary introgression of zebu into African taurine populations after their introduction to the continent[4,15]. These breeds exhibit intermediate phenotypes between *Bos taurus* and *Bos indicus*, characterized by smaller humps and dewlaps[13,16]. Zenga cattle, primarily found in East and Central Africa, are generally considered crossbred populations derived from crosses between zebu and Sanga, and they often carry a higher proportion of zebu (indicine) ancestry than Sanga breeds[13]. Zenga populations were developed to enhance environmental adaptability by combining a stronger zebu genetic background with the local adaptation of Sanga cattle, enabling them to withstand harsh conditions such as heat, drought, poor pasture, and foot-and-mouth disease[13,17]. Egypt represents a confluence zone for both taurine and indicine cattle, as exemplified by the Baladi breed, which shares traits with West Asian cattle[18]. Central Uganda is notable for composite breeds (Nganda, Nyoro, and Ntuuku), developed through selection and crossbreeding with Ankole. Other breeds in the region include the East African Shorthorn Zebu (Nkedi, Kyoga, Serere, and Karamojong) and Friesians[19]. In southern Africa, indigenous Sanga breeds such as Nguni, Tuli, Drakensberger, and Afrikaner are widely used in various beef production systems[20], while humpless taurine breeds are predominant in Northwest Africa. Despite large-scale initiatives such as the 1000 Bull Genomes Project[21], publicly available cattle genomic resources remain heavily biased toward taurine breeds, whereas African indigenous cattle are still under-represented, particularly clearly characterized composite local populations from northeast, equatorial, and southern Africa. This dataset substantially enhances the representation of African-adapted cattle genomes. Although indigenous cattle remain numerically dominant in Africa, the widespread introduction of commercial taurine breeds and their crossbreds has raised concerns about the erosion of indigenous genetic diversity and the loss of locally adapted traits[4,15]. However, the spread of crossbreeding for productivity gains has raised concerns regarding the erosion of indigenous genetic diversity and loss of locally adapted traits[22,23]. This has renewed interest in the conservation and characterization of indigenous breeds, recognizing their role as reservoirs of genetic diversity and their adaptability to challenging and changing environments.

The emergence of next-generation sequencing (NGS) has revolutionized genetic research, providing unmatched opportunities for high-throughput and cost-efficient analysis of DNA and RNA molecules[24]. Whole-genome sequencing (WGS) offers unparalleled opportunities to assess genetic diversity, adaptation, and selection signatures in African cattle populations, providing essential resources for breeding, conservation, and the development of new genomic tools[25].

This study presents high-coverage WGS data for hundreds of indigenous African cattle from Egypt ($N = 25$), Uganda ($N = 95$), and South Africa ($N = 120$), sampled from diverse agro-ecological zones (Table 1). These data represent the one of the largest collection of sequenced indigenous African cattle to date and provide an important addition to existing global cattle genome databases[21,26]. The data are expected to be extremely versatile, and useful for tasks such as exploring genomic diversity, identifying signatures of positive selection, analyzing genome-environment associations, and pinpointing genomic variants near candidate genes or QTLs related to disease and production traits. Furthermore, the data can assist in investigating different types of genomic variants, including small insertions/deletions and structural variants, and in developing genomic analysis tools, such as high or low-density SNP genotyping arrays, for breeding programs.

Methods

Ethics statement. All methods were conducted in accordance with relevant guidelines and regulations. Blood samples were collected during the animals' annual health inspections, conducted by licensed veterinarians. Prior to sample collection, written informed consent was obtained from each animal's owner. In Egypt, sampling was conducted in accordance with the animal welfare guidelines of the Institutional Animal Care and Use Committee, Cairo University (CU-IACUC), which approved the protocol under number CUIIF720. In Uganda, the study was approved by the Makerere University School of Biosecurity, Biotechnical and Laboratory Sciences (SBLS) Higher Degrees Research and Ethics Committee (Ref. No. SBLS/HDRC/20/001). In South Africa, sampling of blood was performed with the approval of the Animal Ethics Committee of the Agricultural Research Council (APAEC [2020/17]), according to guidelines for the handling of animals during sample collection.

Country	Breed	Geographic region	Species	Purpose	Number of herds	*N* (Female)	*N* (Male)	*N*
Egypt	Egyptian cattle	North	Hybrid	Dual-purpose	2	0	3	3
	Egyptian cattle	Central	Hybrid	Dual-purpose	5	4	6	10
	Egyptian cattle	South	Hybrid	Dual-purpose	10	4	8	12
Uganda	Ankole	Western	Sanga*	Dual-purpose	3	4	15	19
	Karamojong	Northern	African *Bos indicus*	Dual-purpose	3	5	6	11
	Nganda	Central	Zenga*	Dual-purpose	7	2	25	27
	Ntuuku	Western	African *Bos indicus*	Dual-purpose	3	2	17	19
	Nkedi	Central	African *Bos indicus*	Dual-purpose	3	5	14	19
South Africa	Afrikaner	Limpopo, Northern Cape	Sanga	Dual-purpose	3	9	31	40
	Bonsmara	Limpopo, Northern Cape, Eastern Cape,	Sanga	Dual-purpose	5	8	12	20
	Nguni	KwaZulu-Natal, Limpopo	Sanga	Dual-purpose	10	5	35	40
	Tuli	Eastern Cape, Northern Cape	Sanga	Dual-purpose	6	8	12	20
Total	—	—	—	—	60	56	184	240

Table 1. Indigenous African cattle populations collected in Egypt, Uganda, and South Africa. Detailed metadata, including sex, GPS coordinates, and ENA accession numbers, are provided in Table S1. *Sanga: Hybrid of '*Bos taurus* × *Bos indicus*' (collective name: Sanga). *Zenga: Hybrid of '*Bos taurus* × Sanga' (collective name: Zenga).

Sampling. In total, 240 healthy animals of African indigenous breeds were sampled along a north-south transect across Egypt, Uganda, and South Africa, with each breed adapted to its unique agro-climatic and ecological environment (Table 1; European Nucleotide Archive (ENA) details are provided in Table S1). Breed-specific sample sizes ranged from 3 to 40 individuals, primarily reflecting practical considerations, including animal availability, accessibility of sampling locations, and field logistics. To maximize representation of within-breed genetic diversity despite these differences, sampled animals were selected to be unrelated within two generations and were collected across multiple herds (typically 2–10 herds per breed, depending on availability). Detailed GPS coordinates of sampling locations are provided in Fig. 6 and Table S1. As this study is presented as a genomic resource descriptor rather than a hypothesis-driven association study, formal power calculations were not performed. The sampling strategy aimed to capture representative within-breed genetic diversity across agro-ecological zones. Breed classification was based on national breed registries, phenotypic characteristics, and breeder records, rather than genomic validation.

Whole-blood samples were collected from each animal in 9 ml Vacutainer® tubes containing K3EDTA as anticoagulant (Greiner Bio-One, ref. 455036). Sampling was conducted by qualified veterinarians following standardized protocols established by the OPTIBOV project (https://www.optibov.org/), in accordance with official animal healthcare guidelines and with informed consent from breeders.

In the northernmost region, Egypt, local cattle breeds were sampled from the northern ($N = 3$), central ($N = 10$), and southern ($N = 12$) parts of the country. Lower sampling in northern Egypt reflects limited herd availability and consolidation of smallholder systems. Because the Egyptian regional groups were unevenly sampled, the apparent geographic gradient observed in the PCA should be interpreted cautiously, as the limited sample size may reduce the stability of regional clustering. Egypt's climate is characterized by hot, dry summers and mild winters, with coastal regions experiencing moderate temperatures due to the influence of the Mediterranean Sea. In contrast, desert areas are subject to extreme temperature fluctuations: daytime temperatures may exceed 40 °C in summer and can fall below 0 °C during winter nights. Humidity remains low throughout much of the country, especially in the interior. Egyptian cattle are managed under a wide range of ecosystems, ranging from the humid temperate Mediterranean coast in the north (Lower Egypt) to hot and arid conditions in the southern governorates (Upper Egypt). This local breed is mainly managed under an extensive system, with few animals (1–5) kept per household, and fed fodder crops such as berseem (alfalfa) in winter, maize in summer, and concentrates year-round, with free access to water. The most common diseases are tick-borne diseases, brucellosis, and foot-and-mouth disease.

Moving south to Uganda, the sampling included Ankole ($N = 19$), Karamojong ($N = 11$), Nganda ($N = 27$), Nkedi ($N = 19$), and Ntuuku ($N = 19$) cattle. Karamojong cattle were collected in the north, Ntuuku and Ankole in the west, and Nganda in the central region. Uganda's equatorial climate is marked by two rainy seasons and mild temperatures, typically ranging from 16 °C to 28 °C. However, the northern savannah and higher-altitude regions may experience cooler nights and warmer days. Livestock production is based on grazing natural pastures, with crop residues and commercial feeds used in more intensive systems. Major diseases affecting Ugandan cattle include East Coast fever, trypanosomiasis, and helminth infections.

In South Africa, samples were collected from the Afrikaner ($N = 40$), Bonsmara ($N = 20$), Nguni ($N = 40$), and Tuli ($N = 20$) breeds. The country exhibits substantial climatic variation: temperate highlands in the interior experience warm summers and cold winters with occasional frost and subzero night temperatures, while lowland and desert regions such as the Karoo can reach extreme temperatures, with summer highs over 40 °C and winter lows below freezing. Coastal areas have milder, more humid climates with moderate temperatures. The Temperature-Humidity Index (THI) varies from moderate in coastal and highland areas to low in arid regions. Livestock in South Africa typically graze on natural pastures, with additional fodder crops and commercial feeds provided as needed. The main disease challenges include foot-and-mouth disease, tick-borne diseases, and anthrax.

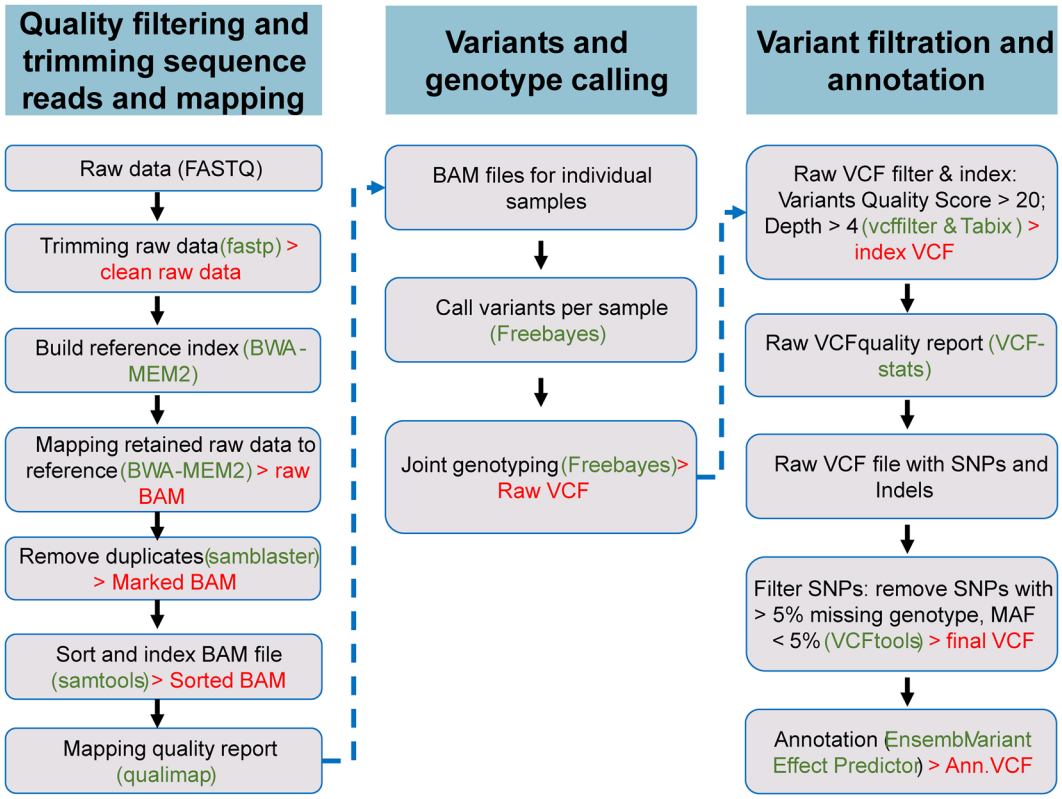

Fig. 1 Overview of the sequence quality control, processing, mapping, variant calling, filtering, and annotation pipeline used in this study. The workflow was adapted from the population variant calling pipeline (Code Availability). Software tools are shown in green, data files generated during the workflow are shown in red. Data saved to external repositories (e.g., ENA, EVA) are indicated in blue in parentheses.

Genomic DNA isolation and whole-genome sequencing. The blood samples were processed at the Wageningen University & Research Sequencing Facility. Genomic DNA was extracted using a modified salting-out precipitation method (Gentra Puregene Blood Kit, Qiagen) following the manufacturer's instructions. The concentration of DNA was measured using the Qubit® dsDNA BR Assay on a Qubit® 2.0 fluorometer (Life Technologies, CA, USA). To visualize the isolated DNA samples, agarose gel electrophoresis with ethidium bromide staining was employed. The genomic DNA from each sample was normalized to a final volume of 100 µl and a concentration of 20–50 ng/µl. NovaSeq. 6000 (Illumina Inc., USA) sequencing data were obtained from dual-indexed genomic libraries with paired-end and 150 bp reads.

Sequencing data processing, mapping and variant calling. The raw sequencing reads underwent pre-processing, which involved trimming adapter sequences, correcting mismatched bases in overlapping paired-end reads, and removing duplicate and low-quality reads, using the fastp v0.23.4 package[27]. No samples were excluded during this preprocessing step; all samples were retained for downstream alignment and variant calling. BWA-MEM2 (v2-2.2.1) was used to align sequence reads to the bovine reference genome (assembly versions ARS-UCD1.2/GCA_002263795.2 and ARS-UCD2.0/GCA_002263795.4 for the Y chromosome)[28]. The resulting SAM/BAM files were further processed, which included marking duplicate reads with samblaster v0.1.26[29], sorting by coordinates using Samtools v1.14[30], and generating a mapping quality report with QualiMap v2.0[31]. The sorted BAM files obtained were subsequently used for variant calling. Figure 1 provides an overview of the steps involved in mapping, variant calling, and variant filtering.

Both individual sample variant calling and joint genotyping across the samples were conducted using the FreeBayes (v1.3.1) pipeline for "population variant calling" (https://wiki.anunna.wur.nl/index.php/Population_variant_calling_pipeline)[32]. Previous studies showed Biallelic SNPs called from Freebayes and GATK-HC had similar concordance to those called using the bead arrays and had greater than 99% concordance to each other[33]. To improve the reliability of detected variants, several calling and filtering parameters were applied during this process. Specifically, genotype qualities were recorded (--genotype-qualities), and only the two most likely alleles per site were considered (--use-best-n-alleles 2) to reduce spurious complexity at multi-allelic positions. Reads supporting variant calls were further constrained by requiring a minimum base quality of 10 (--min-base-quality 10), a minimum alternate allele fraction of 0.2 (--min-alternate-fraction 0.2), and at least two reads supporting the alternate allele (--min-alternate-count 2). Variant calling was conducted under a diploid model (--ploidy 2), and haplotype-based complex calling was disabled (--haplotype-length 0) to focus on single-site variation. The requirement that at least 20% of reads support the alternative allele was based on the commonly used FreeBayes setting

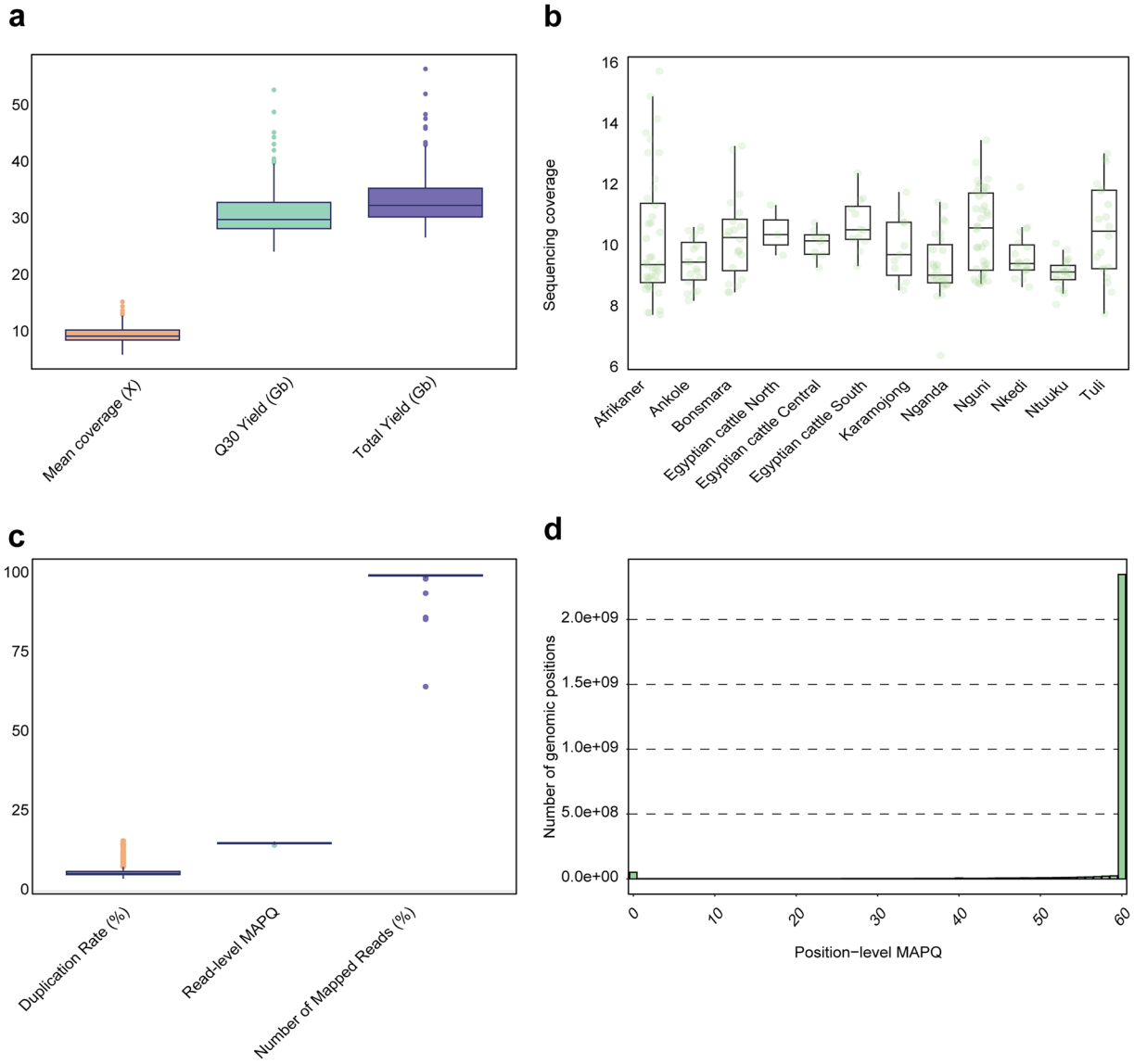

Fig. 2 High-quality sequencing data and mapping results. (**a**) Boxplots show total sequencing yield (Gb), Q30 yield (Gb), and mean coverage for African indigenous cattle samples ($N = 240$). (**b**) Mean coverage per individual is shown for each breed. (**c**) Duplication rates, mapping success rates, and average read-level mapping quality (MAPQ). In all boxplots, boxes represent the interquartile range (IQR; 25th–75th percentile), the horizontal line within each box indicates the median, whiskers extend to the most extreme data points within 1.5 × IQR, and dots represent individual animals. (**d**) Histogram of position-level MAPQ values across the genome. The x-axis shows MAPQ scores (0–60), the y-axis shows the number of genomic positions, and each bar represents the mean number of positions across samples at each MAPQ value.

for diploid samples with moderate sequencing depth, and was intended to reduce spurious alternate observations while retaining true variants supported by multiple reads[32]. Following variant calling, initial site-level filtering was performed using vcffilter/vcflib v0.0.2019.07.1032[34], retaining only variants with a quality score greater than 20 (QUAL > 20) and a total depth greater than 4 (DP > 4). These thresholds were used to remove low-confidence sites and variants supported by insufficient read depth, which are more likely to arise from sequencing noise, local misalignment, or ambiguity in repetitive and structurally complex genomic regions. We acknowledge that the DP > 4 threshold is relatively permissive and may retain some false-positive variants. We did not apply an additional per-genotype depth filter in this analysis, although such filtering could be considered in future work for more stringent analyses of individual SNPs. Because the same filtering criteria were applied uniformly across all populations, and downstream analyses were based on genome-wide variant patterns after minor allele frequency (MAF) and missingness filtering, this strategy is unlikely to introduce systematic ancestry-specific filtering bias.

For the filtered datasets, variants with a missing genotype rate exceeding 5% or a MAF below 0.05 were removed using VCFtools v0.1.16[35]. This ensured that only high-confidence variants were retained for subsequent population structure analyses. Both the primary VCF files and the filtered VCF files generated after this

General statistics of raw and filtered variants		
Variant class	**Raw count**	**Filtered count (MAF ≥ 0.05, missing rate ≤ 0.05)**
SNPs	43,529,353	26,939,288
Indels	3,408,559	469,367
Substitutions	1,928,480	1,178,349
Insertions	1,686,400	1,019,008
deletion	2,069,151	1,222,658
Overlapped genes	27,213	27,208
Overlapped transcripts	43,492	43,487

Table 2. Summary statistics for variants identified in the raw and filtered VCF files (excluded Y), including counts of SNPs, small indels (≤50 bp), and substitution categories.

step have been deposited in the European Variation Archive (EVA) to support transparency and future reuse. Y-chromosomal estimates were based on the 56 male samples included in this dataset. Due to the low diversity and limited coverage of the Y chromosome, Y-chromosomal variants were excluded from downstream analyses.

For the raw and filtered variant analysis, we used the Ensembl Variant Effect Predictor (VEP, release 110.1) for annotating variants identified in genomic regions[36]. Variant annotation was performed in offline mode using the *Bos taurus* ARS-UCD1.2 reference assembly. In the current analysis, VEP annotation was limited to standard gene-based consequence categories, including coding, UTR, intronic, intergenic, splice-region, upstream/downstream gene, and non-coding transcript consequences. Regulatory region annotations, such as promoter and enhancer annotations, were not included in this version of the analysis. Detailed information about the specific codes used for mapping, variant calling, filtering, and annotation processes can be found in the "Code Availability" section.

Population structure. A Principal Component Analysis (PCA) based on filtered autosomal SNP data (MAF ≥ 0.05, missing rate ≤ 0.05) was conducted to investigate the genetic structure of African indigenous cattle, incorporating reference groups African indigenous cattle breeds relative to reference groups from the 1000 Bull Genomes Project Run 9[21]: Reference panels include African *Bos taurus* (N'Dama and Muturu), European *Bos taurus* (Hereford, Holstein Friesian, Jersey, and Angus), African *Bos indicus* (Goffa, Boran, Kenana, and Ogaden), and Asian *Bos indicus* (Nelore, Hariana, Tharparkar, Gir, and Red Sindhi). LD-pruned SNPs was performed with PLINK (--indep-pairwise 50 10 0.2), and the top four principal components were extracted (--pca 4). Eigenvectors were visualized in ggplot2 v3.5.1[37].

Genome-wide ancestry composition was inferred using ADMIXTURE v1.3.0[38]. LD-pruned genotypes were used to reduce redundancy. Ancestry proportions were estimated for *K* values ranging from 2 to 4 with cross-validation error. The *K* values were selected based on biological interpretability rather than the lowest cross-validation error. Admixture plots were visualized in R with ggplot2 v3.5.1[37].

Data Records

The complete raw sequencing data, in FASTQ format, have been deposited in the ENA under Project accession PRJEB90914[39]. The corresponding variant call files (VCFs), including both the primary dataset and the filtered dataset used for downstream analyses, have been deposited in the EVA under project accession PRJEB110435[40] (ARS-UCD1.2; raw data), PRJEB94226[41] (ARS-UCD1.2; filtered data) and PRJEB93868[42] (ARS-UCD2.0 Y chromosome; raw data). Publicly available reference variant datasets were obtained from the 1000 Bull Genomes Project (PRJNA391427[43]) and the African Genomic Reference Resource (PRJEB74565[44]; Muturu).

Technical Validation

Quality control of sequencing data. Each sample yielded a sequencing output ranging from 27 to 57 Gb, with 89–95% of the bases (averaging 93%) achieving a minimum Phred quality score of 30. This score indicates an anticipated base calling accuracy of 99.9%. No samples were excluded during the initial preprocessing step, and all samples were retained for downstream alignment and variant calling. The average coverage for the samples ranged from 6× to 15×, with an overall average of 10× (Fig. 2a, Table S2). Additionally, the average coverage for each breed per individual is illustrated in Fig. 2b.

Following filtering, the proportion of reads mapped to the taurine reference genome ranged from 64.5% to 99.9%, with a mean mapping rate of 99.2% (Table S3). As alignment was performed against the predominantly *Bos taurus* reference genome (ARS-UCD1.2/2.0), we also examined breed-specific mapping performance in relation to ancestry and admixture background. Overall mapping rates were high, but a modest trend was observed: the lowest average mapping rates occurred in Ugandan indicine/Zenga breeds (Nganda: 98.07%; Ntuuku: 98.17%; Karamojong: 99.18%; Nkedi: 99.41%), consistent with increased divergence from the taurine reference. Mapping rates increased in admixed Sanga and Sanga-derived breeds (Ugandan Ankole: 99.48%; South African Afrikaner: 99.58%; Bonsmara: 99.57%; Nguni: 99.51%; Tuli: 99.53%). Egyptian cattle displayed the highest mapping rates among African populations (North: 99.79%; Middle: 99.62%; South: 99.60%), consistent with comparatively higher taurine ancestry and/or closer similarity to the reference (Table S3).

The mean read-level mapping quality (MAPQ) score was 15.10 across samples (range: 14.74–15.29; Fig. 2c, Table S3), whereas the mean position-level MAPQ score was 57.22 across samples (range: 56.53–57.52;), with a consistently high histogram peak at MAPQ 60 (Fig. 2d). Read-level MAPQ includes all aligned reads, including

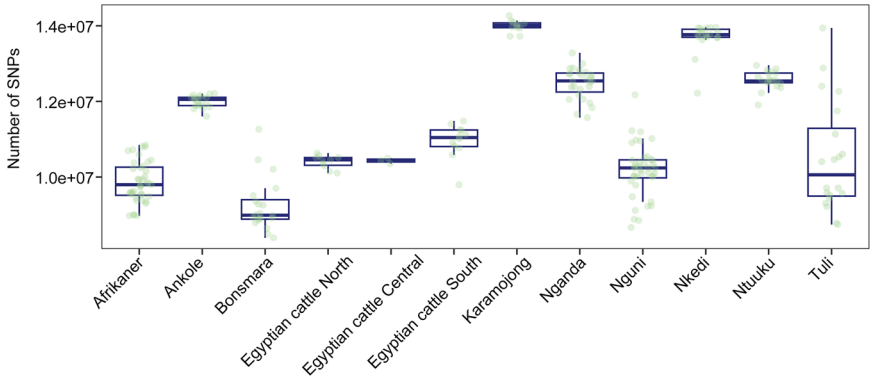

Fig. 3 Distribution of SNP counts per individual across African indigenous cattle breeds ($N = 240$). Boxplots summarize the number of SNPs per breed, with green dots indicating individual animals. Boxes represent the interquartile range (IQR; 25th–75th percentile), the horizontal line within each box indicates the median, and whiskers extend to the most extreme data points within 1.5 × IQR.

Chromosomes	Raw count	Filtered count (MAF ≥ 0.05, missing rate ≤ 0.05)	Raw SNP density (count/kb)	Filtered SNP density (count/kb)
1	3,054,950	1,905,767	15.92	12.02
2	2,509,697	1,560,383	15.17	11.45
3	2,182,855	1,372,405	14.81	11.34
4	2,346,938	1,514,437	16.27	12.62
5	2,234,247	1,424,304	15.28	11.86
6	2,294,109	1,420,024	16.10	12.05
7	2,026,336	1,291,838	15.17	11.67
8	2,084,021	1,286,116	15.20	11.35
9	1,986,777	1,201,054	15.52	11.39
10	1,953,440	1,250,852	15.65	12.11
11	1,956,381	1,197,983	15.21	11.2
12	1,896,883	1,212,848	17.99	13.91
13	1,508,773	954,999	14.95	11.44
14	1,540,191	958,304	15.38	11.63
15	1,791,782	1,143,478	17.60	13.45
16	1,523,736	955,272	15.34	11.79
17	1,441,388	880,146	16.17	12.03
18	1,249,042	795,619	15.78	12.09
19	1,158,209	717,422	15.00	11.31
20	1,422,343	881,320	16.38	12.24
21	1,347,469	840,117	16.05	12.03
22	1,139,953	689,747	15.46	11.35
23	1,277,536	774,090	20.15	14.74
24	1,241,658	775,387	16.42	12.44
25	812,331	485,597	15.74	11.47
26	1,021,410	629,008	16.36	12.1
27	989,448	612,686	17.90	13.43
28	986,407	619,813	17.97	13.49
29	1,138,383	704,382	18.30	13.78
X	1,437,550	850,456	8.30	6.12
Y	131,300	82,311	2.19*	—

Table 3. Chromosome-wise summary statistics of SNPs in the raw and filtered VCF file, including the number of variants and their distribution across the 29 autosomes and the sex chromosomes. * Y-chromosomal SNP density was calculated using the 56 male samples included in this dataset. The Y chromosome showed lower SNP density and diversity (mean: 2.19, range: 0.05–12.09) than the autosomes and X chromosome in males (mean: 15.56, range: 0.01–69.72).

multimapped or ambiguously aligned reads from repetitive or complex genomic regions, whereas position-level MAPQ summarises the average mapping quality of reads supporting each genomic position. Thus, although

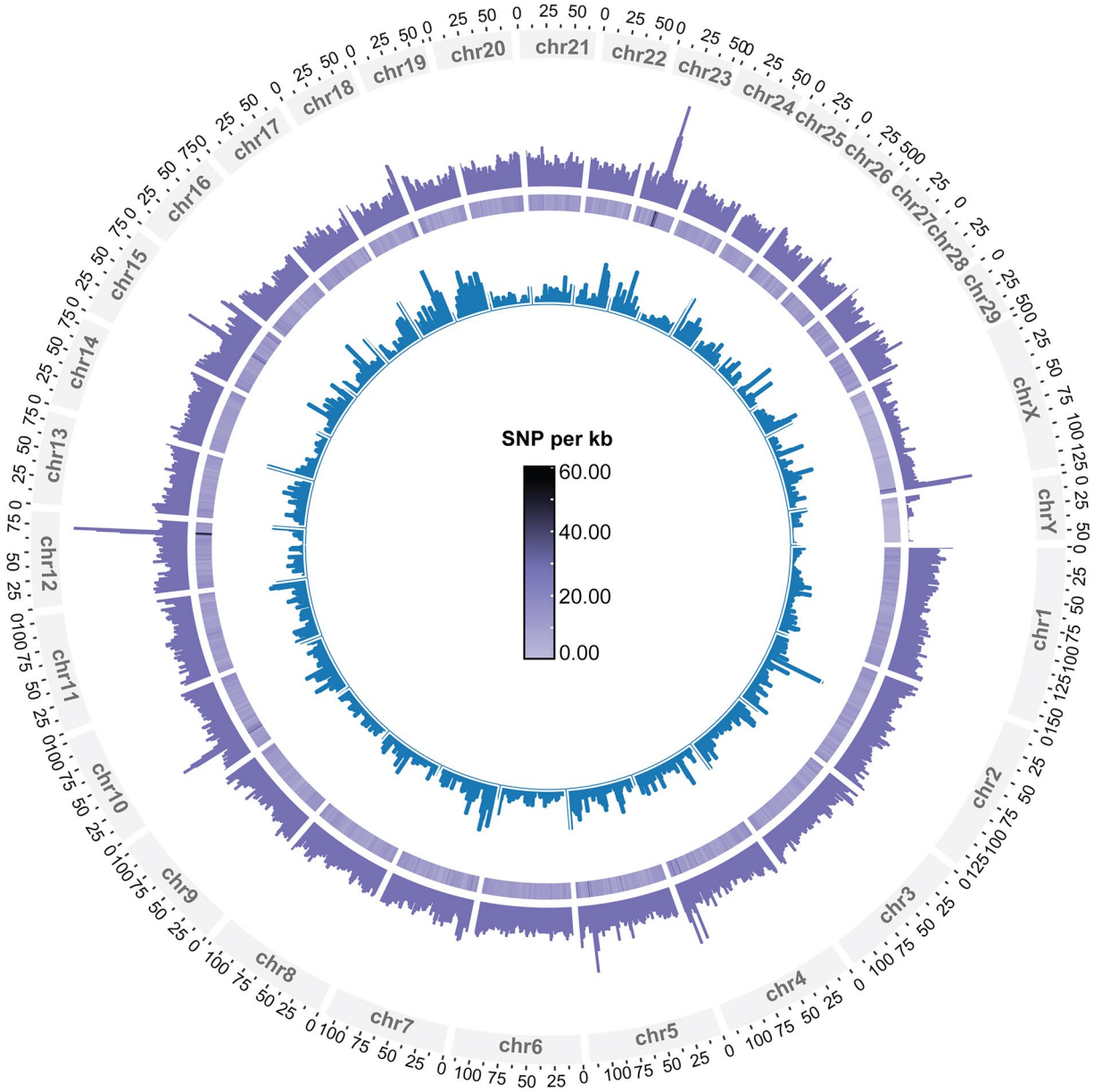

Fig. 4 Chromosomal distribution of SNPs across African cattle genomes based on ~27 million variants. Chromosomes are ordered by size on the X-axis. The purple heatmap and accompanying bar plot show SNP density (SNP per kb), while the blue bar plot indicates gene density across 1 Mb windows.

read-level MAPQ value is lower than expected for uniquely mapped reads generated by BWA-MEM2[28], most callable genomic positions were predominantly supported by confidently mapped reads, it is comparable to values reported in previous cattle whole-genome sequencing studies involving diverse European populations[45]. The relatively low read-level MAPQ may reflect mapping ambiguity in repetitive or structurally complex genomic regions, as well as genetic divergence from the taurine-based reference genome, particularly in indicine-related populations. Consistent with this interpretation, read-level MAPQ showed a modest ancestry-associated pattern, with the lowest values observed in indicine-related populations (e.g., Karamojong, 14.74) and the highest values in taurine-indicine admixed/crossbred populations (e.g., Bonsmara, 15.29). Duplication rates (mean: 6%, range: 3.89–15.78%, Table S3) and were also comparable to those reported in previous cattle WGS studies[45]. Therefore, variants located in repetitive or complex genomic regions should be interpreted with caution. Nevertheless, downstream analyses were based on variant-level quality filtering, including variant quality, read depth, missingness, and MAF thresholds, to reduce the impact of low-confidence variant calls.

Additional alignment-level quality metrics further supported the reliability of the sequencing data. Across samples, the median insert size 339 bp (range: 253–397 bp), consistent with typical short-insert Illumina whole-genome libraries. GC content was highly stable, with a mean of 44.28% (range: 41.98–46.03%), indicating uniform sequence composition across samples and no obvious GC-related bias. The general error rate was low overall, averaging 0.0075 (range: 0.0053–0.0098), further supporting good read and alignment quality.

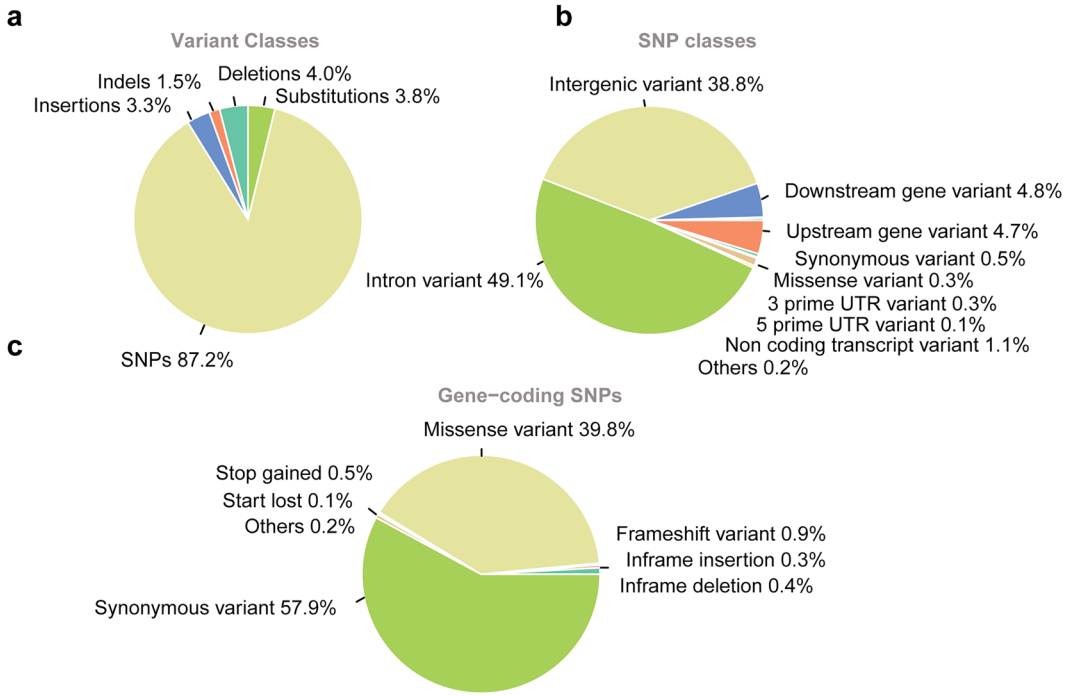

Fig. 5 Functional annotation of ~27 million variants identified in African indigenous cattle genomes. (**a**) Variant classes. (**b**) Distribution of SNPs by genomic context. (**c**) Gene-coding SNPs.

Taken together, the high mapping rates, consistent coverage, acceptable mapping performance, and favorable alignment-level QC metrics demonstrate that the sequencing dataset was of sufficient quality for downstream genetic analyses.

Quality control of SNP data. Variant calling across all samples identified approximately 43 million SNPs and 6 million small indels (≤50 bp; Table 2). To ensure variant accuracy and minimize false positives, a series of stringent filtering steps were applied. Only variants located on autosomes and sex chromosomes were retained, and only biallelic SNPs were considered. In addition, the variant-calling and filtering criteria required a minimum base quality of 10, at least 20% of reads and a minimum of two reads supporting the alternative allele, and retention of variants with a quality score greater than 20 and a total read depth greater than 4. No explicit per-sample missingness threshold was applied. However, per-individual missing genotype rates were low and homogeneous across the 240 individuals before filtering, with a mean of 0.6% and a range of 0.2–3.2%. As no individual showed a missing genotype rate greater than 5%, no samples were excluded from downstream analyses based on per individual missingness. The transition/transversion (Ti/Tv) ratio, a standard measure of variant calling quality, was 2.29 in the post-filtering variant dataset, consistent with expectations for whole-genome sequencing and indicative of high data quality[46]. The number of SNPs per individual for each breed is shown in Fig. 3 and Table S4.

Subsequent filtering for downstream analyses targeted variants with high missingness or low MAF. The SNPs with a missing genotype rate > 0.05 and those with MAF < 0.05 were excluded. After all quality control steps, approximately 27 million high-quality SNPs and 2 million small indels (≤50 bp) remained (Table 2). Among these, 27,208 genes and 43,487 transcripts were overlapped by at least one SNP. Analysis of SNP distribution in both the raw and filtered datasets revealed a genome-wide but uneven pattern across all *Bos taurus* chromosomes (Table 3, Fig. 4). The Y chromosome showed markedly lower SNP density and diversity (mean: 2.19, range: 0.05–12.09) than the autosomes and X chromosome in males (mean: 15.56, range: 0.01–69.72), with Y-chromosomal estimates based on the 56 male samples included in this dataset[47]. Due to the low diversity of the Y chromosome, Y-chromosomal variants were excluded from downstream analyses.

Annotation and genetic structure of SNP Data. Raw and filtered variants from the autosomes and X chromosome were classified into variant types, including SNPs, indels, substitutions, insertions, and deletions (Table 3; Fig. 5a), and were further functionally annotated according to their genomic locations and predicted effects (Table S5). The overall annotation profile was consistent with large-scale cattle whole-genome sequencing studies[21,45,48], including the 1000 Bull Genomes Project and other bovine WGS datasets, in which most SNPs are located in non-coding regions, particularly intergenic and intronic regions. In the 1000 Bull Genomes reference dataset (Run 6)[21], approximately 66.1% of SNPs were intergenic, 26.2% were intronic, and less than 1% were classified as missense or synonymous variants. Similarly, in our dataset, most SNPs were in intronic (49.1%) and intergenic (38.8%) regions, whereas only a small proportion were located in exonic regions (0.8%; Fig. 5b). The differences in the relative proportions of intergenic and intronic variants between datasets may reflect differences in the reference genome assembly and genome annotation used in the 1000 Bull Genomes

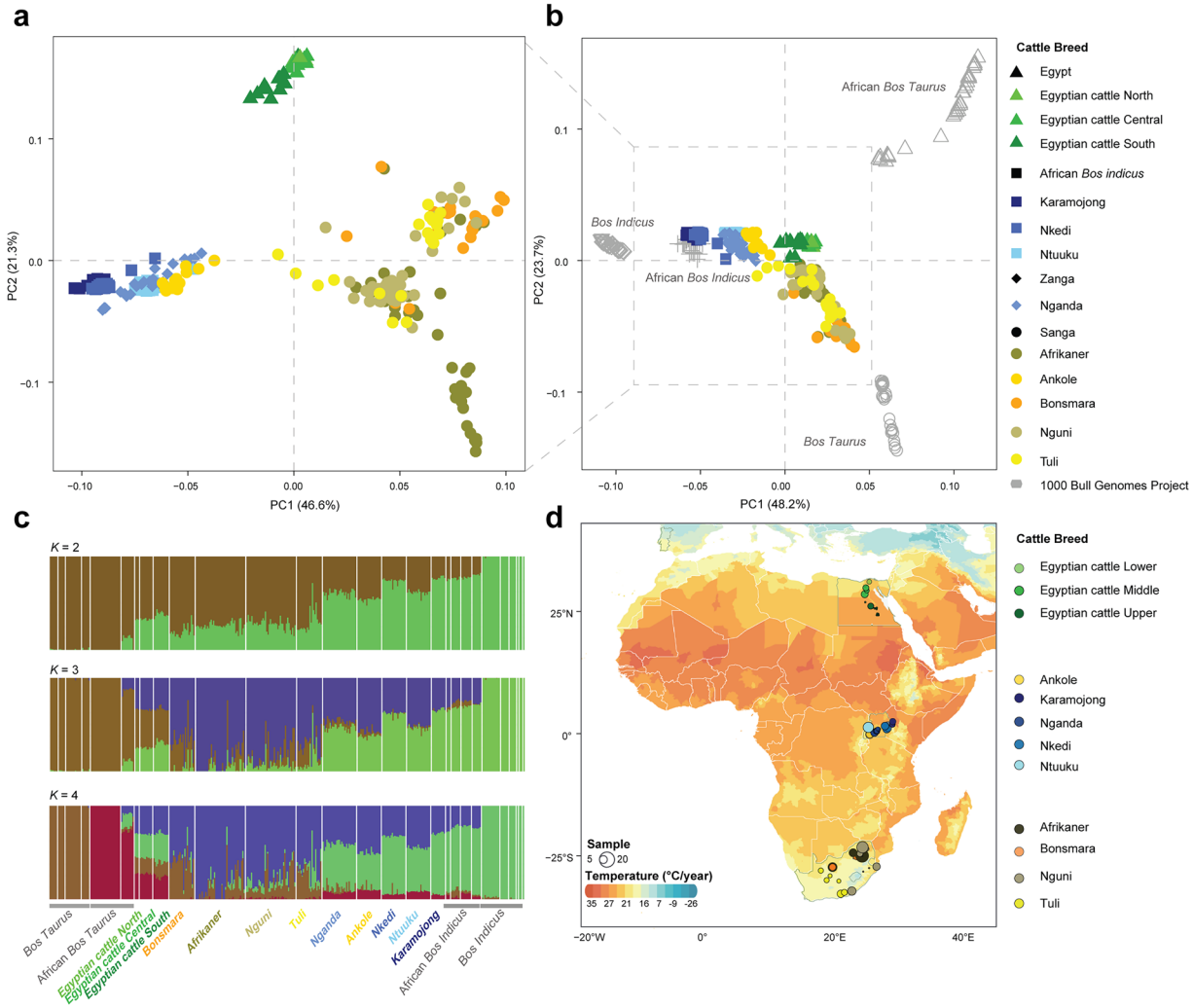

Fig. 6 Genetic and geographic population structure of African indigenous cattle breeds. (**a**) Principal component analysis (PCA) of African indigenous cattle breeds. (**b**) PCA of African indigenous cattle breeds in relation to reference populations from the 1000 Bull Genomes Project Run 9[21]. Axis labels indicate the proportion of variance explained by each principal component. (**c**) ADMIXTURE analysis of African indigenous cattle breeds together with reference populations from the 1000 Bull Genomes Project Run 9[21] Reference panels include African *Bos taurus* (N'Dama and Muturu), *Bos taurus* (Hereford, Holstein Friesian, Jersey, and Angus), African *Bos indicus* (Goffa, Boran, Kenana, and Ogaden), and *Bos indicus* (Nelore, Hariana, Tharparkar, Gir, and Red Sindhi). (**d**) Geographic sampling locations of the African indigenous cattle breeds included in this study (Table S1).

Project Run 6 compared with the current reference and annotation used in this study[21]. To benchmark the functional annotation results, we compared the coding consequence distribution in the present dataset with the annotation summary previously reported for 289 European native cattle using a comparable reference and annotation framework[45]. In the present dataset, among gene-coding SNPs, 57.9% were synonymous variants and 39.8% were missense variants, corresponding to a synonymous-to-missense ratio of approximately 1.46. This distribution was broadly comparable to that reported for European native cattle, where coding consequences were similarly dominated by synonymous variants (59.35%) and missense variants (38.72%), with a synonymous-to-missense ratio of approximately 1.53. In addition, we identified 1,957 stop-gained variants introducing premature stop codons in the present dataset (Fig. 5c). Overall, the similar distribution of annotated SNP consequence classes supports the consistency of our functional annotation results with previous cattle genomic resources.

PCA was performed to provide an overview of the genetic composition of the African cattle samples in relation to reference populations from the 1000 Bull Genomes Project Run 9[21] (Fig. 6a,b). In both PCA panels, Egyptian, Ugandan, and South African cattle formed three broad groups, reflecting the major genetic differences represented in the dataset. Egyptian cattle showed a tentative geographic pattern, with individuals from Upper Egypt, in the southern region, tending to position closer to *Bos indicus*, whereas individuals from Lower Egypt, in the northern region, were positioned closer to *Bos taurus*. However, this pattern should be interpreted

cautiously because of the limited and uneven sampling across Egyptian regions. Southern African Sanga breeds, including Nguni and Bonsmara, were located between the taurine and indicine reference groups, consistent with their known admixed ancestry. ADMIXTURE was used as a complementary exploratory approach to visualize broad ancestry patterns. The *K* values presented in Fig. 6c ($K = 2$, 3, and 4) were selected based on biological interpretability rather than the lowest cross-validation error. Consistent with the PCA results, ADMIXTURE revealed heterogeneous ancestry patterns across African cattle populations, with variable contributions from taurine-related ancestry (African and European taurine) and indicine-related ancestry components (Fig. 6c). The geographic locations of these samples are shown in Fig. 6d (Table S1). These analyses are intended to provide a descriptive overview of the genetic structure represented in this resource and to facilitate interpretation and reuse of the dataset in future studies.

Data availability

Raw sequencing data are available in the ENA under Project accession PRJEB90914[39]. VCFs, including both the primary dataset and the filtered dataset used for downstream analyses, are available in EVA under the following project accessions: PRJEB110435[40] (ARS-UCD1.2; raw data), PRJEB94226[41] (ARS-UCD1.2; filtered data), and PRJEB93868[42] (ARS-UCD2.0 Y chromosome; raw data). Publicly available reference variant datasets were obtained from the 1000 Bull Genomes Project (PRJNA391427[43]) and the African Genomic Reference Resource (PRJEB74565[44]).

Code availability

The data analyses were conducted using standard bioinformatics tools on a Linux system. Representative command-line scripts and parameter settings for the main sequencing quality control, read preprocessing, mapping, variant calling, variant filtering, annotation, and population structure analyses are publicly available on GitHub (https://github.com/junxingao888/BovWGS-Pipeline) and Zenodo (https://doi.org/10.5281/zenodo.19189075)[49].

Received: 29 August 2025; Accepted: 12 May 2026;
Published online: 19 May 2026

References

1. Marshall, K. *et al*. Livestock genomics for developing countries–African examples in practice. *Frontiers in genetics* **10**, 297 (2019).
2. Herrero, M., Thornton, P. K., Gerber, P. & Reid, R. S. Livestock, livelihoods and the environment: understanding the trade-offs. *Current Opinion in Environmental Sustainability* **1**, 111–120 (2009).
3. Alonso, S., Dominguez-Salas, P. & Grace, D. The role of livestock products for nutrition in the first 1,000 days of life. *Animal Frontiers* **9**, 24–31 (2019).
4. Hanotte, O. *et al*. African pastoralism: genetic imprints of origins and migrations. *Science* **296**, 336–339 (2002).
5. Kim, K. *et al*. Inference of admixture origins in indigenous African cattle. *Molecular biology and evolution* **40**, msad257 (2023).
6. Smith, A. B. in *Oxford Research Encyclopedia of African History* (2021).
7. Rossi, C. *et al*. The genomic natural history of the aurochs. *Nature* **635**, 136–141 (2024).
8. Verdugo, M. P. *et al*. Ancient cattle genomics, origins, and rapid turnover in the Fertile Crescent. *Science* **365**, 173–176 (2019).
9. Utsunomiya, Y. *et al*. Genomic clues of the evolutionary history of Bos indicus cattle. *Animal genetics* **50**, 557–568 (2019).
10. Ndihokubwayo, F. & Koç, A. Breeding pure and crossbreeds of European cattle breeds for milk production improvement under tropical climate conditions—a review. *Tropical Animal Health and Production* **56**, 23 (2024).
11. Bahbahani, H. *et al*. Signatures of positive selection in African Butana and Kenana dairy zebu cattle. *PloS one* **13**, e0190446 (2018).
12. Mirkena, T. *et al*. Genetics of adaptation in domestic farm animals: A review. *Livestock Science* **132**, 1–12 (2010).
13. Mwai, O., Hanotte, O., Kwon, Y.-J. & Cho, S. African indigenous cattle: unique genetic resources in a rapidly changing world. *Asian-Australasian journal of animal sciences* **28**, 911 (2015).
14. Ayalew, W. *et al*. Whole-genome sequencing of copy number variation analysis in Ethiopian cattle reveals adaptations to diverse environments. *BMC genomics* **25**, 1088 (2024).
15. Rege, J. & Tawah, C. The state of African cattle genetic resources II. Geographical distribution, characteristics and uses of present-day breeds and strains. *Animal Genetic Resources/Resources génétiques animales/Recursos genéticos animales* **26**, 1–25 (1999).
16. Scholtz, M. & Theunissen, A. The use of indigenous cattle in terminal cross-breeding to improve beef cattle production in Sub-Saharan Africa. *Animal Genetic Resources/Resources génétiques animales/Recursos genéticos animales* **46**, 33–39 (2010).
17. Kim, K. *et al*. The mosaic genome of indigenous African cattle as a unique genetic resource for African pastoralism. *Nat Genet* **52**, 1099–1110, https://doi.org/10.1038/s41588-020-0694-2 (2020).
18. Blench, R. & MacDonald, K. *The origins and development of African livestock: archaeology, genetics, linguistics and ethnography*. (Routledge, 2006).
19. Kathambi, E. K., Sonstegard, T. S. & Larsen, P. A. Review: Cross-breeding, advanced reproductive technologies, and genetic selection in twelve dairy production systems in Africa. *Animal* **19**, 101424, https://doi.org/10.1016/j.animal.2025.101424 (2025).
20. Van Marle-Köster, E., Visser, C., Sealy, J. & Frantz, L. Capitalizing on the potential of South African indigenous beef cattle breeds: a review. *Sustainability* **13**, 4388 (2021).
21. Hayes, B. J. & Daetwyler, H. D. 1000 bull genomes project to map simple and complex genetic traits in cattle: applications and outcomes. *Annual review of animal biosciences* **7**, 89–102 (2019).
22. Chebo, C. & Alemayehu, K. Trends of cattle genetic improvement programs in Ethiopia: Challenges and opportunities. *Livestock Research for Rural Development* **24**, 1–17 (2012).
23. Hanotte, O., Dessie, T. & Kemp, S. Time to tap Africa's livestock genomes. *Science* **328**, 1640–1641 (2010).
24. Hudson, M. E. Sequencing breakthroughs for genomic ecology and evolutionary biology. *Molecular ecology resources* **8**, 3–17 (2008).
25. Midgley, G. F. & Bond, W. J. Future of African terrestrial biodiversity and ecosystems under anthropogenic climate change. *Nature Climate Change* **5**, 823–829 (2015).
26. Tijjani, A. *et al*. Genomic reference resource for african cattle: genome sequences and high-density array variants. *Scientific Data* **11**, 801 (2024).
27. Chen, S. F. Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp. *Imeta* **2** (2023). ARTN e107 https://doi.org/10.1002/imt2.107
28. Vasimuddin, M., Misra, S., Li, H. & Aluru, S. in *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*. 314–324.

29. Faust, G. G. & Hall, I. M. SAMBLASTER: fast duplicate marking and structural variant read extraction. *Bioinformatics* **30**, 2503–2505 (2014).
30. Danecek, P. *et al*. Twelve years of SAMtools and BCFtools. *Gigascience* **10**, giab008 (2021).
31. Okonechnikov, K., Conesa, A. & García-Alcalde, F. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics* **32**, 292–294 (2016).
32. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. *arXiv preprint arXiv:1207.3907* (2012).
33. Stegemiller, M. R. *et al*. Using whole genome sequence to compare variant callers and breed differences of US sheep. *Frontiers in genetics* **13**, 1060882 (2023).
34. Garrison, E., Kronenberg, Z. N., Dawson, E. T., Pedersen, B. S. & Prins, P. A spectrum of free software tools for processing the VCF variant call format: vcflib, bio-vcf, cyvcf2, hts-nim and slivar. *PLoS computational biology* **18**, e1009123 (2022).
35. Danecek, P. *et al*. The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158, https://doi.org/10.1093/bioinformatics/btr330 (2011).
36. McLaren, W. *et al*. The Ensembl Variant Effect Predictor. *Genome Biol* **17**, 122, https://doi.org/10.1186/s13059-016-0974-4 (2016).
37. Valero-Mora, P. M. ggplot2: elegant graphics for data analysis. *Journal of statistical software* **35**, 1–3 (2010).
38. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome research* **19**, 1655–1664 (2009).
39. *ENA European Nucleotide Archive*. https://identifiers.org/ena.embl:PRJEB90914 (2025).
40. *EVA European Variation Archive*. https://identifiers.org/ena.embl:PRJEB110435 (2026).
41. *EVA European Variation Archive*. https://identifiers.org/ena.embl:PRJEB94226 (2026).
42. *EVA European Variation Archive*. https://identifiers.org/ena.embl:PRJEB93868 (2025).
43. *ENA European Nucleotide Archive*. https://identifiers.org/ena.embl:PRJNA391427 (2017).
44. *ENA European Nucleotide Archive*. https://identifiers.org/ena.embl:PRJEB74565 (2024).
45. Ginja, C. *et al*. Whole genome sequences of 289 native cattle from Finland, the Netherlands, and Portugal. *Scientific data* **12**, 1897 (2025).
46. Wang, J., Raskin, L., Samuels, D. C., Shyr, Y. & Guo, Y. Genome measures used for quality control are dependent on gene function and ancestry. *Bioinformatics* **31**, 318–323, https://doi.org/10.1093/bioinformatics/btu668 (2015).
47. Hellborg, L. & Ellegren, H. Low levels of nucleotide diversity in mammalian Y chromosomes. *Molecular Biology and Evolution* **21**, 158–163 (2004).
48. Weldenegodguad, M. *et al*. Whole-Genome Sequencing of Three Native Cattle Breeds Originating From the Northernmost Cattle Farming Regions. *Frontiers in Genetics* **Volume 9 – 2018**, https://doi.org/10.3389/fgene.2018.00728 (2019).
49. Gao, J. BovWGS-Pipeline [Workflow]. *Zenodo* https://doi.org/10.5281/zenodo.19189075 (2026).

Acknowledgements

We thank all members of the OPTIBOV consortium for their invaluable contributions to sample collection, sequencing, and expert input. This study was funded by the Long-term EU-Africa Research and Innovation Partnership on Food and Nutrition Security and Sustainable Agriculture (LEAP-Agri) as part of the OPTIBOV project (LEAP-Agri-326), and by the European Union's Horizon 2020 Research and Innovation Program under grant agreement No. 727715. The founders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. We gratefully acknowledge the collaboration of the breeders and breed associations in Africa for maintaining these local genetic resources and providing access to the animals. For Uganda, we thank Damian Munirwa (College of Veterinary Medicine, Makerere University) for sampling and laboratory work with Nkedi and Karamojong breeds; Emmanuel Tayebwa (Ankole Cattle Conservation Trust), Kasoro Charles (Butuku Cattle Marketing Cooperative Society), and Birungi Edward (Butungama Cattle Marketing Cooperative Society) for their assistance with Ankole and Ntuuku cattle. For South Africa, we appreciate the support of the Agricultural Research Council (ARC) and breed associations for facilitating sampling of indigenous and commercial breeds. We thank Marina Botes and Mokgadi Seshoka (Afrikaner), Julius Sebei and Lucas Msiza (Bonsmara and Nguni), Jim Bredenkamp, Ben Raath, Christo Rothmann, and Kevin Cook (Tuli), Sandra Erasmus (Holstein-Friesian, ARC), and Simphiwe Nini (Nguni, ARC). Additional funding for Junxin Gao was provided by the China Scholarship Council (CSC), China (grant number 202208610017).

Author contributions

N.D., J.G., and R.C. conceived the study. N.D. and J.G. drafted the manuscript, J.G. participated in data analysis, J.G. and C.G. interpreted the results. C.G. curated the reference datasets. N.D., C.G., J.K., N.G., D.K., M.M., and R.C. were responsible for sample collection. A.E., A.Z., B.Z., C.N., G.B., K.N., M.M., M.R., M.E., M.A., N.F., R.K., R.A., R.O., S.W., and S.L. of the African OPTIBOV consortium contributed to the sample collection and data generation. R.C. supervised the study. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-026-07458-y.

Correspondence and requests for materials should be addressed to J.G. or R.P.M.A.C.