scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Whole genome sequences of 289 native cattle from Finland, the Netherlands, and Portugal

Catarina Ginja[1,2,16], Junxin Gao[3,16] ✉, Juha Kantanen[4], Nasser Ghanem[5], Donald Kugonza[6], Mahlako Makgahlela[7,8], Ana Elisabete Pires[1,2,9], Anabel Usié[10,11], Bert Dibbits[3], Carolina Bruno de Sousa[12], Daniel Gaspar[1,2], Daniil Ruvinskiy[4], Etske Bijl[13], Hauke Smidt[14], Heli Lindeberg[4], Henk Bovenhuis[3], Kimberley Laport[3], Kisun Pokharel[4], Ludmilla Blaschikoff[1,2], Melak Weldenegodguad[4], Rayner Gonzalez Prendes[3], Rodney Okwasiimire[4,15], Silvia Guimarães[1,2], Ying Lui[3] & Richard P. M. A. Crooijmans[3] ✉

Native cattle breeds in Europe are vital to agricultural heritage and livestock production, combining adaptation to diverse environments with desirable traits such as high-quality beef and milk. To investigate genetic diversity, local adaptation, and productivity-related characteristics, we generated whole-genome sequences from 289 cattle representing 11 native breeds and the commercial Holstein-Friesian breed across Finland, the Netherlands, and Portugal. These breeds span diverse climates and management systems, from cold northern regions to Mediterranean environments in southern Europe. The dataset comprises over 11 terabytes of paired-end Illumina NovaSeq6000 sequencing data, with an average depth of ~10 × and an alignment rate of ~99.7% against the ARS-UCD1.2 and 2.0 cattle reference genomes. Variant calling identified about 30 million SNPs and 2.7 million small indels distributed unevenly across the genome. Annotation linked many variants to known genes. This genomic resource provides an important foundation for studying genomic diversity, environmental adaptation, small structural variants discovery, and genomic mapping of economically important traits, offering insights for future breeding and conservation programs in European cattle.

Background & Summary

Cattle descended from at least two separate aurochs domestication events: one in the Fertile Crescent that gave rise to taurine cattle (*Bos taurus*), and a later event in the Indus Valley that produced indicine cattle (*Bos indicus*)[1,2]. Following domestication, cattle dispersed quickly across distinct migration routes throughout Europe, Africa and Asia[3]. Numerous breeds became ubiquitous in all regions of the globe, arising under local selection and complex demographic processes of admixture and differentiation[4], including hybridization with aurochs[5–7].

Across Europe, uniparental (mtDNA, Y-chromosome) and autosomal genetic patterns jointly resolve lineage structure, demographic history, and signals of local adaptation. Maternal lineages are dominated by the

[1]CIBIO, Centro de Investigação em Biodiversidade e Recursos Genéticos, InBIO Laboratório Associado, Campus de Vairão, Universidade do Porto, Vairão, Portugal. [2]BIOPOLIS, Program in Genomics, Biodiversity and Land Planning, CIBIO, Campus de Vairão, Vairão, Portugal. [3]Animal Breeding and Genomics, Wageningen University & Research, Wageningen, The Netherlands. [4]Natural Resources Institute Finland, Jokioinen, Finland. [5]Animal Production Department, Faculty of Agriculture, Cairo University, Giza, Egypt. [6]Department of Agricultural Production, College of Agricultural and Environmental Sciences, Makerere University, Kampala, Uganda. [7]Agricultural Research Council-Animal Production Institute, Irene, South Africa. [8]Department of Animal, Wildlife and Grassland Sciences, University of the Free State, Bloemfontein, South Africa. [9]Faculdade de Medicina Veterinária, Universidade Lusófona, Lisboa, Portugal. [10]CEBAL, Centro de Biotecnologia Agrícola e Agro-Alimentar do Alentejo, Beja, Portugal. [11]MED, Instituto Mediterrâneo para a Agricultura, Ambiente e Desenvolvimento, Universidade de Évora, Évora, Portugal. [12]CIISA, Faculty of Veterinary Medicine, University of Lisbon, Lisbon, Portugal. [13]Food quality and Design, Wageningen University and Research, Wageningen, The Netherlands. [14]Microbiology, Wageningen University and Research, Wageningen, The Netherlands. [15]Department of Agricultural Sciences, University of Helsinki, Helsinki, Finland. [16]These authors contributed equally: Catarina Ginja, Junxin Gao. ✉e-mail: junxin.gao@wur.nl; richard.crooijmans@wur.nl

T3 mitochondrial haplogroup, reflecting an ancestral bottleneck during post-domestication dispersal from the Near East[8]. Although T3 is especially common in northern Europe, T and T2 haplotypes also occur[9]. The Iberian Peninsula (Spain and Portugal) exhibits the greatest mitochondrial diversity among European cattle, harboring multiple maternal lineages, including the rare Q and African T1 haplogroups[10]. Paternally, two major Y-chromosome clades (Y1 and Y2) prevail across northern and central Europe, encompassing breeds such as Eastern Finncattle and Icelandic cattle[9]. These are associated with two waves of dairy cattle expansion: the Y1 with pied or red breeds from the North Sea and Baltic coasts, and the Y2 with spotted, yellow, or brown breeds from Alpine/Swiss region[11]. Iberian breeds carry multiple Y1 and Y2 haplotypes[10]. Autosomal data corroborate that locally adapted Iberian cattle are genetically distinct and highly diverse, with strong influences from Northwest African taurine cattle[12]. This diversity likely enhances their adaptive potential in the face of ongoing and future environmental changes. Consistent with these patterns, signatures of positive selection related to extremely cold adaptation have been detected in Swedish mountain cattle and Finncattle, such as the Fjäll breed[13].

Genomic data remain scarce for several specific local Finn, Dutch, and Iberian breeds[14]. In Finland, Eastern, Northern, and Western Finncattle evolved under boreal to sub-arctic conditions with short growing seasons and long winter feeding[15,16]. These predominantly polled cattle are more often used for landscape management and grazing of less productive marginal lands[17]. In the Netherlands, native breeds developed in a temperate maritime climate with extensive spring-autumn grazing and winter housing, and span dairy and dual-purpose types: Deep Red (vegetation management), Dutch Belted (distinct mid-body "belt"), Dutch Friesian (traditional black-and-white dairy), Meuse-Rhine-Yssel (dual-purpose, high milk-protein, robustness), and Groningen White Headed (soundness and longevity)[18]. In Portugal (Iberian Peninsula), Barrosã, Mirandesa, and Mertolenga occupy contrasting eco-regions from humid Atlantic mountains to continental highlands and Mediterranean systems[19]. While previous studies have released genomic datasets for commercial cattle (e.g., Holstein-Friesian[20]) or limited subsets of native breeds (e.g., Western/Eastern Finncattle or Dutch native cattle[21]), our dataset represents the most comprehensive resource to date for European regional breeds, including population-level variant calls[22,23]. Such data has the potential to elucidate on the impact of distinct selective pressures in the phenotypic and genetic differences observed among native cattle breeds adapted to various environments as has been highlighted in a recent review[24].

The use of up-to-date molecular biology and bioinformatics tools to characterize and manage farm animal biodiversity has been recommended by The Food and Agriculture Organization of The United Nations[25]. However, there is still a large gap between the current state-of-the art in the use of tools to characterize genomic resources and its application to many non-commercial and local breeds[26], hampering the consistent use of genetic and genomic data as indicators of genetic erosion. The analysis of livestock populations from distinct biogeographical regions for which comprehensive phenotypes and environmental information are available can provide key information on adaptive and selection traits. Here we generate and release whole genome sequencing and population variants data for 289 healthy cattle representing 11 native breeds (per-breed sample sizes 21–29) and the transboundary commercial Holstein-Friesian breed (5 per country), distributed across diverse agro-ecological and climatic conditions, i.e, Finland, The Netherlands, and Portugal, across the European Continent (Table 1, Table S1). Of the 289 animals, 61 are males and 228 are females. Over eleven terabytes of paired-end sequencing data, generated using the Illumina Novaseq6000 platform with an average coverage depth of approximately 10 × (range ~9 × −16×), yielded approximately 33 million variants. Data were collected as part of the OPTIBOV project (2018–2022; project pages: https://leap-agri.com/?page_id=291; https://subsites.wur.nl/en/optibov-project.htm), an ERA-Net Cofund programme under LEAP-Agri—a joint Europe-Africa research and innovation initiative on food and nutrition security and sustainable agriculture (https://cordis.europa.eu/project/id/727715).

This resource enables: (i) detection of strong candidate selective sweeps with sample sizes $\geq$ 20; robust detection of weaker sweeps will generally require integrating larger external WGS datasets (e.g., 1000 Bull Genomes Project[27]) and/or incorporating phenotypic data[28]; (ii) functional characterization of candidate sweep genes and favorable alleles introgressed into the improvement of other breeds (e.g., less diverse transboundary commercial populations). (iii) characterization of genome-environment interactions within and among breeds; and (iv) comparative analyses of evolution, adaptation, and selection in livestock to supports sustainable management of biodiversity and climate resilience. (v) identification of deleterious variants (particularly lethal alleles) by model-based methods for targeted pruning, where incorporating health-related traits can help reduce major disease outbreaks[29,30]. Few studies have assembled such a comprehensive dataset of native European cattle genomes. Therefore, this investigation represents an important data source in addition to public genome databases and provides a reference for comparative studies of cattle from other regions of the globe.

Methods

Ethics statement. All procedures complied with European and national regulations on the protection of animals used for scientific purposes. In Finland, animal handling procedures and sample collections were performed in accordance with the legislation approved by Regional State Administrative Agency for Southern Finland (ESAVI/31854/2019). In the Netherlands, the study was carried out under the animal experimentation policy of Wageningen University & Research. Cattle blood samples were collected by licensed veterinarians during routine annual herd-health inspections with written owner consent; no procedures beyond routine veterinary practice were performed, and therefore no ethics committee approval was required. In Portugal, sampling complied with Decree-Law 113/2013, as amended by Decree-Law 1/2019. Blood was collected by licensed veterinarians during routine herd-health management with written owner consent, and no procedures beyond routine veterinary practice were performed, therefore no ethical committee approval was required.

Country	Geographic region	Breed	Species	Purpose	Number of herds	*N* (Female)	*N* (Male)	*N* (Total)
Finland	Northern Europe, Nordic Country	Northern Finncattle	*Bos taurus*	Dual purpose breed	1	25	0	25
		Western Finncattle	*Bos taurus*	Dual purpose breed	1	25	0	25
		Eastern Finncattle	*Bos taurus*	Dual purpose breed	1	25	0	25
		Holstein Friesian	*Bos taurus*	Dairy	2	5	0	5
The Netherlands	Northwestern Europe	Deep Red	*Bos taurus*	Dual purpose breed	4	21	3	24
		Dutch belted	*Bos taurus*	Dual purpose breed	7	17	6	23
		Dutch Friesian	*Bos taurus*	Dual purpose breed	4	24	0	24
		Holstein Friesian	*Bos taurus*	Dairy	1	5	0	5
		Meuse-Rhine-Yssel	*Bos taurus*	Dual purpose breed	4	15	8	23
		Groningen White Headed	*Bos taurus*	Dual purpose breed	6	16	5	21
Portugal	Southwestern Europe, i.e. Iberian Peninsula	Barrosã	*Bos taurus*	Meat (denomination of origin) and draft	14	14	14	28
		Holstein Friesian	*Bos taurus*	Dairy	1	5	0	5
		Mertolenga	*Bos taurus*	Meat (denomination of origin) and draft	11	16	13	29
		Mirandesa	*Bos taurus*	Meat (denomination of origin) and draft	13	15	12	27
Total					70	228	61	289

Table 1. Overview of samples from native European cattle breeds collected in Finland, the Netherlands, and Portugal. Detailed information on sex, GPS coordinates, accession numbers, measurement phenotypes, and ENA/EVA IDs is provided in Table S1.

Sampling. A total of 289 animals from European breeds were sampled, including 274 from native European breeds adapted to its specific agro-climatic and ecological environment and 15 Holstein-Friesian animals (Table 1, Table S1). Samples were collected from nonrelated animals back to the second generation and originated from 70 herds in Finland, The Netherlands, and Portugal. Representative photographs of each breed are shown in Fig. 1. All animals were healthy at the time of collection. The dataset includes breed, sex, country/herd of origin, GPS coordinates, and sequencing coverage; relevant health-trait phenotypes (front and back height, body size, body weight, body temperature, coat color, horn status, and milk production) were also recorded (Table S1). Whole blood was collected into 9 ml Vaccutainer® tubes containing K3EDTA (Greiner Bio-one ref. 455036) by licensed veterinarians following protocols within the OPTIBOV project (https://www.optibov.org/).

From Finland (Northern Europe), three native breeds were included: Western Finncattle ($N = 25$), Northern Finncattle ($N = 25$), and Eastern Finncattle ($N = 25$). From the Netherlands (Northwestern Europe), five native breeds were included: Deep Red ($N = 24$), Dutch Belted ($N = 23$), Dutch Friesian ($N = 24$), Meuse-Rhine-Yssel ($N = 23$), and Groningen White Headed ($N = 21$); genomes for these Dutch breeds have been previously analyzed and released[21,31]. From Portugal (Southwestern Europe), three Iberian native breeds were sampled: Barrosã ($N = 28$), Mirandesa ($N = 27$), and Mertolenga ($N = 29$). Per-breed sample sizes among native breeds ranged from 21 (Groningen White Headed) to 29 (Mertolenga); The transboundary commercial Holstein-Friesian breed dataset has been previously analyzed and released across the three countries ($N = 5$ per country)[20,22].

Genomic DNA isolation and whole-genome sequencing. Blood samples were processed for DNA extraction at the Natural Resources Institute in Finland, the Animal Breeding and Genomics Department of Wageningen University (the Netherlands), and the CIBIO-InBIO laboratories of BIOPOLIS (Portugal). Genomic DNA was extracted using a salting-out precipitation method (Gentra Puregene Blood Kit, Qiagen) following the manufacturer's instructions. Genomic DNA was quantified using the double-stranded dsDNA BR assay with a Qubit® 2.0 fluorometer (Life Technologies, CA, USA) and normalized to 20–50 ng/µl in 100 µl volumes. Illumina Novaseq6000 (Illumina Inc., USA) resequencing data were obtained from dual-indexed genomic libraries with paired-end and 150 bp reads.

Sequencing, mapping and variant calling. Raw sequencing data were pre-processed using fastp v0.23.4[32] for adapter trimming, quality filtering, and duplicate removal. Reads with an average Phred score below 30 and those shorter than 36 bases were discarded. Clean reads were mapped to the bovine reference genome (assembly version ARS-UCD1.2/ARS-UCD2.0 Y-chromosome) using BWA-MEM2 v2.2.1[33]. SAM/BAM files were processed by marking PCR duplicates (samblaster v0.1.26)[34], coordinate sorting (Samtools v1.14)[35], and quality assessment/coverage estimation on post-QC BAMs (QualiMap v2.0)[36].

Variant calling was performed with the Freebayes (v1.3.1) pipeline "population variants calling" on population samples followed by joint genotype[22,37]. Variants with a Phred-scaled quality < 20 and sequencing depth < 4 were removed with vcffilter/vcflib v0.00.2019.07.10[38]. Further filtering was applied with VCFtools v0.1.16[39] to exclude SNPs with a missing genotype rate $>5\%$ or a minor allele frequency (MAF) $<5\%$ across samples.

Variants annotation. Variant statistics were summarized with SnpEff v5.2[40]. Variant annotation by genomic context and predicted coding consequences was performed with Ensembl Variant Effect Predictor (VEP) v111[41]

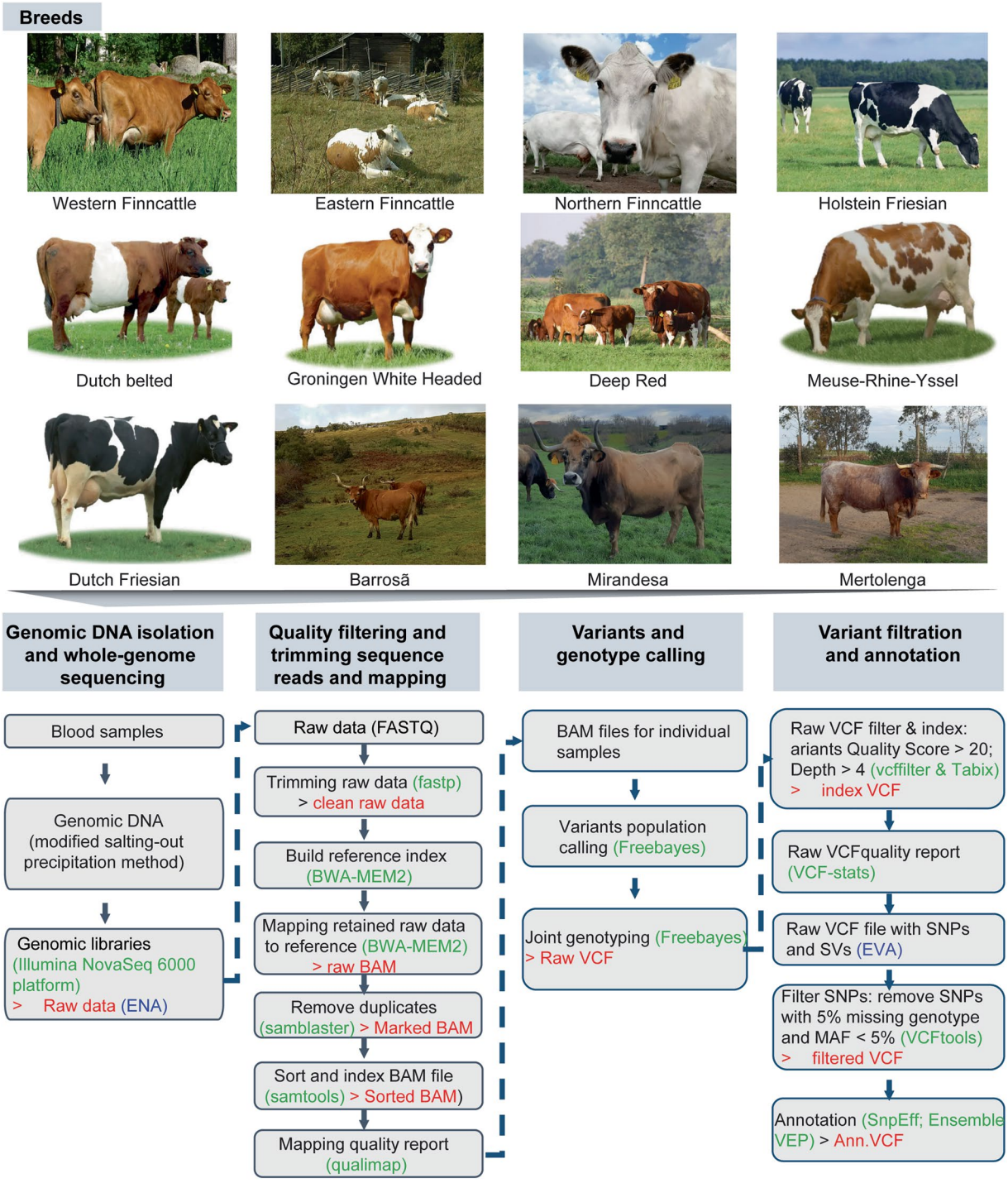

Fig. 1 Overview of sampling, quality control, mapping, variant calling, filtering, and annotation. The pipeline follows the 'population Freebayes variants_calling' (https://wiki.anunna.wur.nl/index.php/Population_variant_calling_pipeline). Text in green indicates the software used, text in blue indicates the data records, and text in red shows the output files generated in each step.

using the ARS-UCD1.2 reference genome (and ARS-UCD2.0 for chromosome Y). Population structure was assessed by principal component analysis (PCA) in PLINK v1.9[42] on genotype data from filtered variants on 289 individuals. Variants were pruned for linkage disequilibrium with–indep-pairwise 50 10 0.2, and eigenvalues/eigenvectors were obtained with –pca 4. PCA plots were generated in R/ggplot2 v3.5.1[43]. SNP density (reported as SNPs per kb) was computed in fixed windows and visualized as a Circos-style track with TBtools[44]. Figure 1 overviews sampling, mapping, variant calling, filtering, and annotation; exact commands are listed in Code Availability.

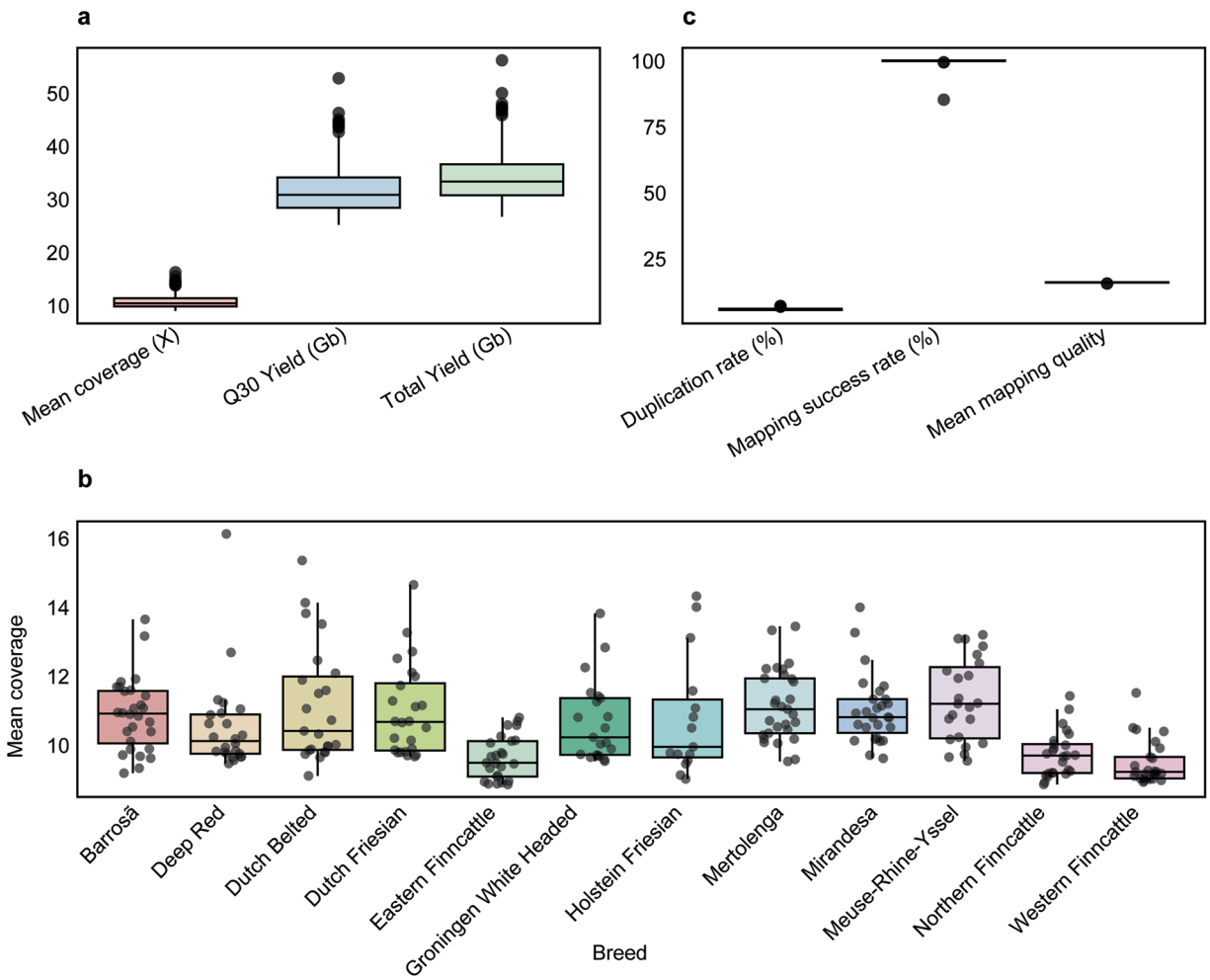

Fig. 2 Sequencing data and mapping summary statistics depicting the high quality of the data. (**a**) Boxplots showing the distributions of total sequencing yield (Gb), Q30 yield (Gb; bases with quality score >30), and mean coverage for European taurine cattle ($N = 289$). (**b**) Mean coverage per individual and breed. (**c**) Duplication rates, mapping success rates, and mean mapping quality resulting from mapping reads to the cattle reference genome. In all panels, boxes represent the interquartile range (IQR; 25th–75th percentile), the horizontal line within the box indicates the median, and whiskers extend to the range within 1.5 × IQR. Black dots indicate individual sample values.

Data Records

This study contributes 154 new FASTQ datasets and new 289-sample population variant datasets. Per-sample metadata (breed, sex, country/herd, GPS coordinates, and measurement phenotypes) are provided in the corresponding ENA/EVA records and in Table S1.

Raw sequencing reads (FASTQ) for indigenous cattle ($N = 274$) from Finland, the Netherlands, and Portugal are available in the ENA under PRJEB90816[45]. Of these, 120 Dutch native samples overlap previously released data under PRJEB56301[21]. Raw reads Holstein-Friesian animals ($N = 15$) are available from the previously published project[20].

Joint-called population VCFs comprising ~30.0 million SNPs and ~2.7 million small indels (shorter than 50 bps) across 289 animals (274 indigenous and 15 Holstein-Friesian) are deposited at the EVA under PRJEB98152 (ARS-UCD1.2)[46]. Y-chromosome variants ~0.6 million SNPs and ~0.1 million small indels for 61 male samples are available under PRJEB94144 (ARS-UCD2.0)[47].

Technical Validation

Quality control of sequencing and mapping data. For each sample, we obtained 26 to 56 Gb of sequencing data, of which about 89–95% of the bases (average 92%) had a minimum Phred quality score of 30, indicating an expected base-calling accuracy of 99.9% (Fig. 2a, Table S2). Post-QC alignment-based sequencing coverage (from BAM files) ranged from 9× to 16× per animal, with an average of 10 × across all samples (Fig. 2a, Table S3). Per-breed mean coverage is shown in Fig. 2b. The mapping success rate against the taurine reference genome (ARS-UCD1.2; ARS-UCD2.0 for the Y chromosome) ranging from 85.0% to 99.9%, with an average of 99.7%. The mean mapping quality (MAPQ) score was 15.7, indicating high-quality read alignment and ~97.3% confidence in the mapping process (Fig. 2c, Table S4).

Variant types	Count (Raw)	Count (Filtered)
SNP	30,039,938	13,381,915
MNP	534,872	315,554
Small insertion	1,306,071	803,395
Small deletion	1,456,327	750,332
Other	174,241	92,235
Total	33,511,449	15,343,431
Overlapped genes	27,215	27,148
Overlapped transcripts	43,494	43,419

Table 2. Summary statistics of population variants in the VCF file.

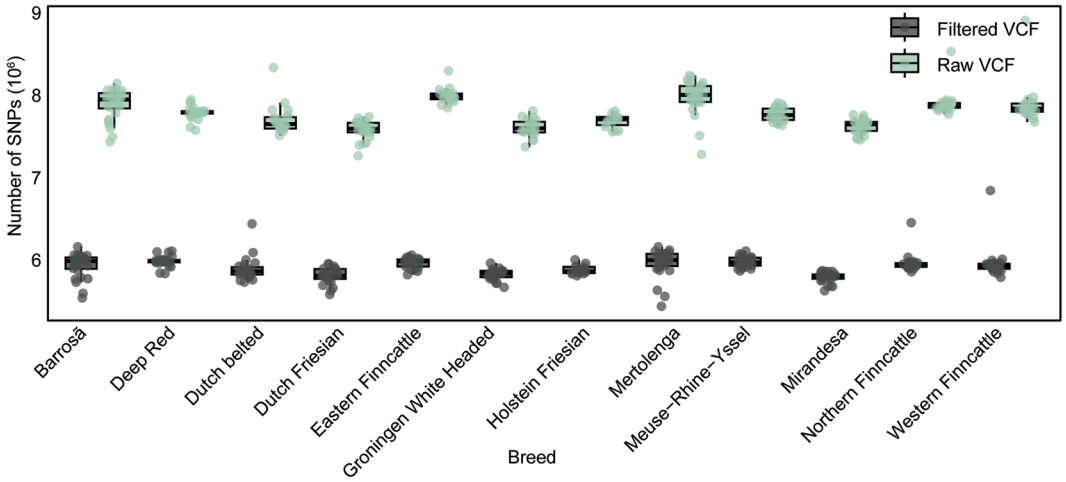

Fig. 3 Boxplot graphic depicting the number of SNPs per individual in each European cattle breed. Boxes represent the interquartile range (IQR; 25th–75th percentile), the horizontal line within the box indicates the median, and whiskers extend to the range within 1.5 × IQR. Green and black dots indicate individual sample values.

Quality control of SNP data. Variant calling across all samples yielded ~30.0 million SNPs, ~2.7 million short indels (<50 bp), 534,872 MNPs, and 174,241 other variant types (Table 2). To ensure high-confidence genotypes and minimize false positives for downstream population-structure analyses, variants were called with Freebayes v1.3.1[22,37] using both default and custom parameters: limiting detection to the two most likely alleles per site, requiring ≥ 20% of reads (and ≥2 supporting reads) for the alternate allele, and enforcing a minimum site Phred quality of QUAL ≥ 20. Subsequent filtering with VCFtools v0.1.16[39] removed sites with > 5% missing genotypes and MAF < 0.05, retaining ~13.4 million high-quality SNPs (Table 2). The number of SNPs per individual for each breed is shown in Fig. 3.

The transition and transversion ratio (Ti/Tv) was used as a quality control metric for SNP calling, with a typical value of about 2 for whole genome sequencing data[48]. Our primary and filtered dataset achieved a Ti/Tv ratio of 2.24–2.25, indicating high-quality SNP calling.

Annotation and genetic structure of SNP data. Variant calling were annotated based on their types and genomic locations (Fig. 4a, Table S5). Most variants were found in intronic (48.6%) and intergenic (38.4%) regions, while 1.8% were exonic and 10% were located upstream or downstream of genes (Fig. 4b, Table S5). The majority of the coding variants were synonymous (59.4%), which do not alter the amino acid sequence of proteins. Missense variants accounted for 38.7%, changing the amino acid sequence but not changing the full-length of the protein. We also identified 3,106 stop-gained variants, which introduce a premature stop codon into the coding sequence, resulting in truncated proteins (Fig. 4c, Table S5).

The genetic structure of these European cattle breeds was investigated through a Principal Components Analysis (PCA) on 13.4 million filtered high-quality SNPs (Fig. 5). The PCA graphic revealed a clear differentiation among cattle breeds from the different countries, with PC1 and PC2 explaining 61% of the total genomic diversity. The PC1 separated Portuguese native breeds from all others, highlighting the genomic differentiation between the Northern and Southern European breed groups. The circos plot of SNP/gene density across the bovine autosomes (BTA) shown in Fig. 6 (Table 3), depicts regions with significantly high SNP density, such as on BTA4, BTA12, BTA23, and the X-chromosome. For instance, the outlier region on BTA23 between 25 Mb and 30 Mb harbors the Major Histocompatibility Complex (MHC), also known as the Bovine Leukocyte Antigen (*BoLA*) Complex[49]. Another outlier region was located on BTA12 between 70 Mb and 73 Mb, which contains genes related to the ATP-binding cassette (ABC) family. This region encodes for ABCG4 proteins that transport various xenobiotics across the plasma membrane and cholesterol into milk[50]. In contrast, the Y-chromosome

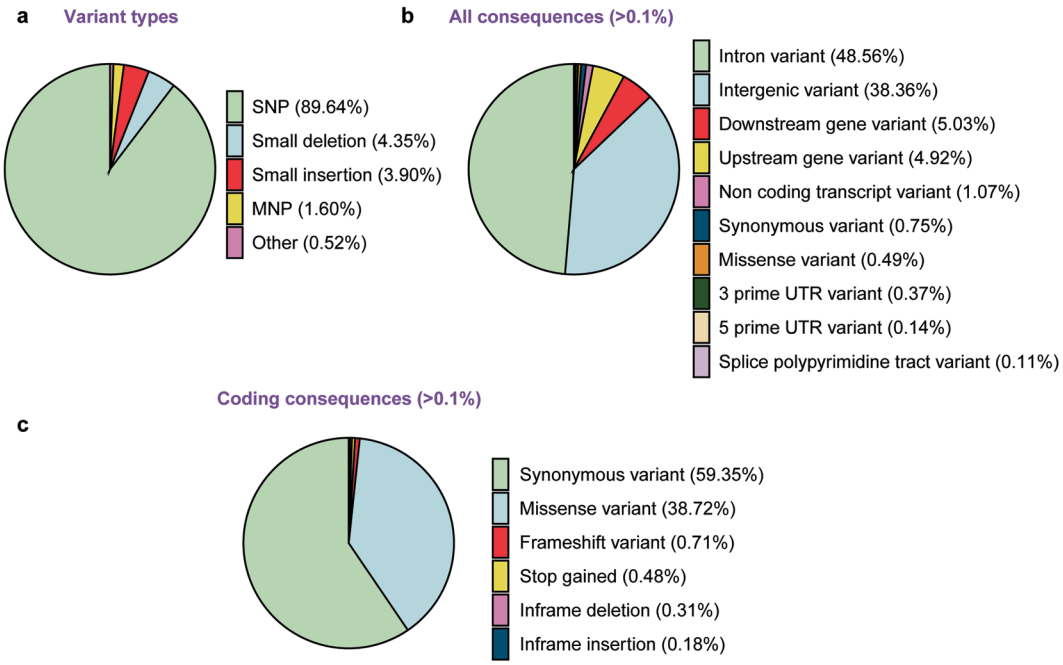

Fig. 4 Variants classified in different annotation categories. (**a**) variant types. (**b**) variants consequences. (**c**) protein-coding variants.

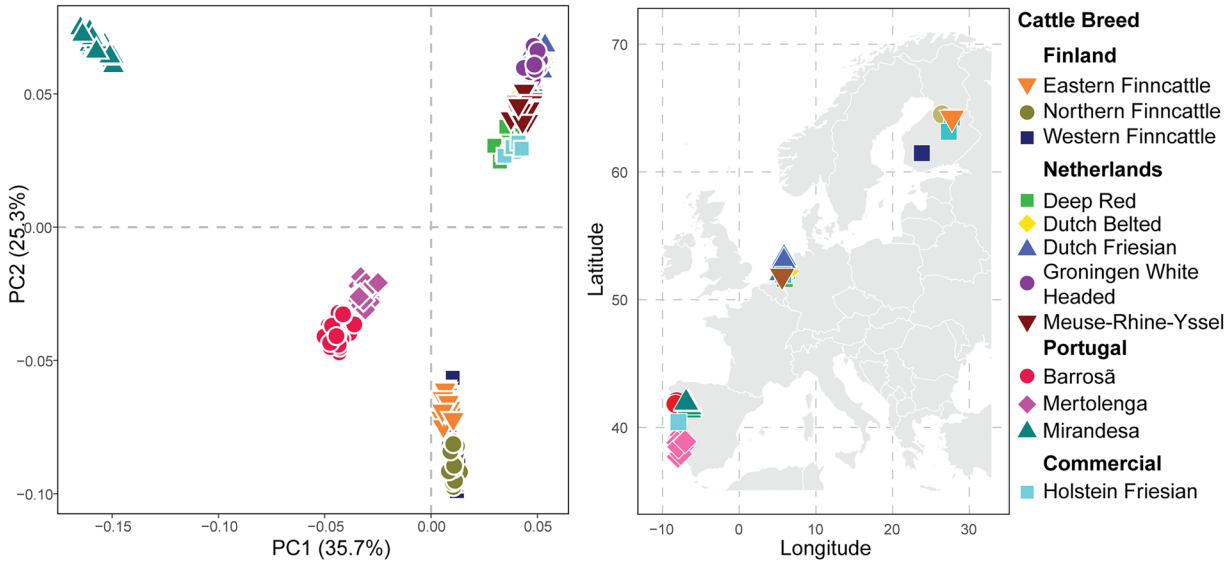

Fig. 5 Population structure (PCA) of European cattle breeds with geographic map.

exhibited a low SNP density and diversity, likely reflecting both the limited number of male samples ($N = 61$) and the small effective population size of the Y-chromosome[51].

Data availability

Raw sequencing reads (FASTQ) for indigenous cattle ($N = 274$) are available in the ENA under PRJEB90816[45]. A subset of 120 Dutch native samples overlaps previously released data under PRJEB56301212[1]. Raw reads Holstein-Friesian animals ($N = 15$) are available from the previously published project[20]. Joint-called population variant call sets (VCFs) across 289 animals are deposited at the at the EVA under PRJEB98152 (ARS-UCD1.2)[46]. Y-chromosome variants for 61 male samples are available under PRJEB94144 (ARS-UCD2.0)[47]. All metadata are provided in Table S1.

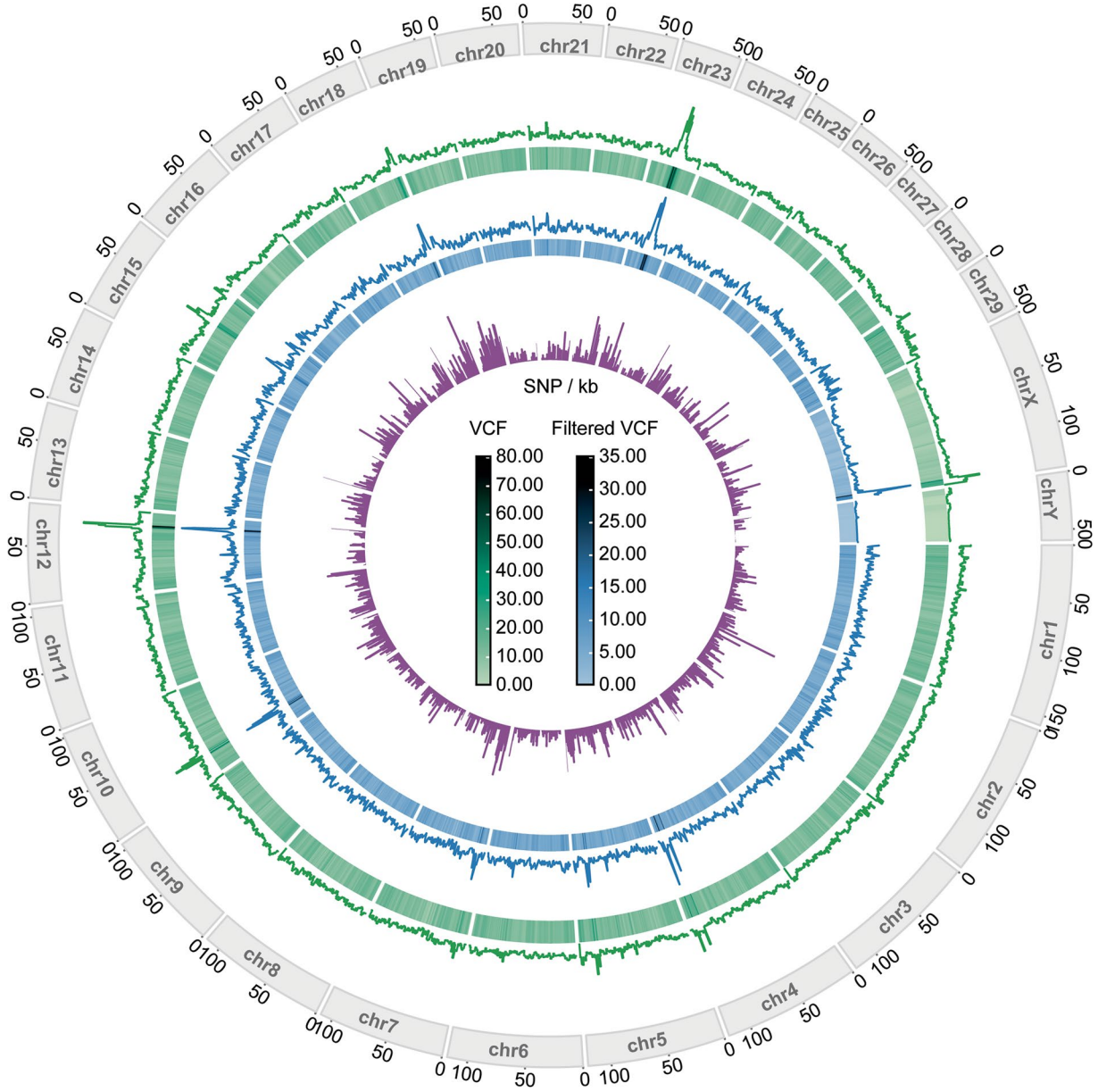

Fig. 6 Chromosome-wise distribution of SNP density in European cattle. The green heatmap/line shows raw SNPs (~30 million; per kb, 1-Mb windows). The blue heatmap/line shows filtered SNPs (~13.4 million; per kb, 1-Mb windows). The purple bars indicate gene density (genes per 1-Mb window). The x-axis is genomic position along each chromosome; the y-axes show density per window (Reference genome: ARS-UCD1.2/ARS-UCD2.0 Y-chromosome).

Code availability

Most of the data analyses were completed by standard bioinformatic tools running on the Linux system. The version and code/parameters of the main software tools are described below.

Fastp (v0.23.4): Codes for quality control of raw data.

This process involved filtering out low-quality reads (Phred quality score of 30 or higher), trimming reads shorter than 36 bases, correcting base errors, eliminating duplicates, and exporting the high-quality reads to output files.

```
fastp -i "${input_dir}/${SAMPLE}/RAW1.fq.gz \
        -o "${output_dir}/${SAMPLE}/${READS_1} " \
    -I "${input_dir}/${SAMPLE}/RAW2.fq.gz" \
    -O "${output_dir}/${SAMPLE}/${READS_2} " \
    -q 30 -l 36 -c -D
```

Chromosome	Variants (Raw)	Variants (Filtered)	Variants density (Raw, count/kb)	Variants density (Filtered, count/kb)
1	2,069,563	928,288	13.05	5.86
2	1,696,657	749,531	12.45	5.5
3	1,486,597	675,576	12.29	5.58
4	1,597,429	731,202	13.31	6.09
5	1,576,894	704,229	13.13	5.86
6	1,545,842	726,613	13.12	6.17
7	1,386,878	630,270	12.53	5.69
8	1,401,258	617,099	12.37	5.45
9	1,307,868	604,016	12.40	5.73
10	1,351,510	630,055	13.08	6.1
11	1,289,986	586,784	12.06	5.48
12	1,309,678	607,952	15.02	6.97
13	1,007,198	451,994	12.07	5.41
14	1,032,308	469,791	12.53	5.7
15	1,247,645	546,345	14.68	6.43
16	1,020,516	472,493	12.60	5.83
17	965,778	453,175	13.20	6.19
18	859,262	406,075	13.05	6.17
19	772,356	348,433	12.17	5.49
20	933,294	431,165	12.97	5.99
21	879,478	397,825	12.59	5.69
22	745,402	326,087	12.27	5.37
23	974,787	463,978	18.57	8.84
24	800,979	369,205	12.85	5.92
25	549,600	253,746	12.98	5.99
26	675,422	306,989	12.99	5.9
27	637,813	301,208	13.98	6.6
28	620,377	288,490	13.50	6.28
29	796,265	390,809	15.58	7.65
X	972,809	356,508	7.00	2.56
Y	68,912	10,427	1.16	0.21

Table 3. Summary statistics of SNPs in the VCF files for each chromosome.

BWA-mem (v2-2.2.1): Code for indexing the reference genome and mapping reads.
${REF} INDEX
bwa-mem2 index ${REF}
Mapping clean ${READS}
bwa-mem2-2.2.1_x64-linux/bwa-mem2 mem ${REF} ${READS_1} ${READS_2} > ${SAMPLE}.sam

Samblaster (v0.1.26): Codes for marking duplicate reads.
samblaster -r -i ${SAMPLE}.sam

samtools (v2.9.2): Codes for sorting/index and converting to BAM.
samtools sort -m 16 G -@ {resources.cpus} -O bam ${SAMPLE}.sam > ${SAMPLE}.bam
samtools index -@ {resources.cpus} ${SAMPLE}.bam

Qualimap (v2.0): Codes for mapping quality report.
unset DISPLAY & & qualimap bamqc -bam ${SAMPLE}.bam --java-mem-size = *G -nt * -outdir ${SAMPLE}

Freebayes (v1.3.1): Codes for variant calling, joint genotypes, and VCF file filtering/indexing.
{params.scripts_dir}/freebayes-parallel.sh < ({params.scripts_dir}/fasta_generate_regions.py ${REF}.fai {params.chunksize}) * \

-f ${REF} \
--use-best-n-alleles 2 --haplotype-length 0 --ploidy 2 --min-alternate-count 2 --min-base-quality 10 --min-alternate-fraction 0.2 --min-alternate-count 2 --genotype-qualities \

-L ${SAMPLE.bam} | vcffilter -f 'QUAL > 20 & DP > 4' | bgzip -c > ${SAMPLE.vcf.gz}
Index ${SAMPLE.vcf.gz}
tabix -p vcf ${SAMPLE.vcf.gz}

vcfstats (v0.00.2019.07.10): Codes for generating variant statistics report on each chromosome.
vcfstats --vcf ${SAMPLE.vcf} --outdir ${SAMPLE} --formula 'chromosome in ${REF}'
VCFtools(v0.1.16): Codes for variant filtration.
vcftools --gzvcf ${SAMPLE} --maf 0.05 --min-meanDP 4 --max-missing 0.05 --recode
--recode-INFO-all--out ${SAMPLE}

Ensembl Variant Effect Predictor (v111): Codes for variant annotation.
vep -i ${SAMPLE} -o ${SAMPLE_ANNOTATION} --cache --species bos_taurus --assembly ${REF}

PLINK (v1.9): Codes for population structure (PCA)
Convert VCF to PLINK binary format
plink --vcf ${SAMPLE}.vcf.gz --allow-extra-chr–make-bed --out ${SAMPLE} --indep-pairwise 50 10 0.2
--snps-only
Perform PCA with top 4 components
plink -- bfile ${SAMPLE} --pca 4 --out ${SAMPLE}.pca --extract ${SAMPLE}.prune.in

VCFtools (v0.1.16): Codes for SNP density in Mb windows
vcftools --gzvcf ${SAMPLE}.vcf.gz --SNPdensity 1000000 --out ${SAMPLE}.snp_density

Received: 29 July 2025; Accepted: 23 October 2025;
Published online: 03 December 2025

References

1. Park, S. D. *et al*. Genome sequencing of the extinct Eurasian wild aurochs, Bos primigenius, illuminates the phylogeography and evolution of cattle. *Genome biology* **16**, 1–15 (2015).
2. Di Lorenzo, P. *et al*. Uniparental genetic systems: a male and a female perspective in the domestic cattle origin and evolution. *Electronic Journal of Biotechnology* **23**, 69–78 (2016).
3. Pitt, D. *et al*. Domestication of cattle: Two or three events? *Evolutionary applications* **12**, 123–136 (2019).
4. Scherf, B. D. & Pilling, D. The second report on the state of the world's animal genetic resources for food and agriculture (2015).
5. Pérez-Pardal, L. *et al*. Legacies of domestication, trade and herder mobility shape extant male zebu cattle diversity in South Asia and Africa. *Scientific reports* **8**, 18027 (2018).
6. Ginja, C. *et al*. Iron age genomic data from Althiburos–Tunisia renew the debate on the origins of African taurine cattle. *Iscience* **26** (2023).
7. Verdugo, M. P. *et al*. Ancient cattle genomics, origins, and rapid turnover in the Fertile Crescent. *Science* **365**, 173–176 (2019).
8. Lenstra, J. A. *et al*. Meta-analysis of mitochondrial DNA reveals several population bottlenecks during worldwide migrations of cattle. *Diversity* **6**, 178–187 (2014).
9. Kantanen, J. *et al*. Maternal and paternal genealogy of Eurasian taurine cattle (Bos taurus). *Heredity* **103**, 404–415 (2009).
10. Ginja, C. *et al*. The genetic ancestry of American Creole cattle inferred from uniparental and autosomal genetic markers. *Scientific reports* **9**, 11486 (2019).
11. Edwards, C. J. *et al*. Dual origins of dairy cattle farming–evidence from a comprehensive survey of European Y-chromosomal variation. *PLoS One* **6**, e15922 (2011).
12. da Fonseca, R. R. *et al*. Consequences of breed formation on patterns of genomic diversity and differentiation: the case of highly diverse peripheral Iberian cattle. *BMC genomics* **20**, 1–13 (2019).
13. Ghoreishifar, S. M. *et al*. Signatures of selection reveal candidate genes involved in economic traits and cold acclimation in five Swedish cattle breeds. *Genetics Selection Evolution* **52**, 1–15 (2020).
14. Weldenegodguad, M. *et al*. Whole-genome sequencing of three native cattle breeds originating from the northernmost cattle farming regions. *Frontiers in genetics* **9**, 728 (2019).
15. Hämet-Ahti, L. The boreal zone and its biotic subdivision. *Fennia-International Journal of Geography* **159** (1981).
16. Bläuer, A. *et al*. Inferring prehistorical and historical feeding practices from δ15N and δ13C isotope analysis on Finnish archaeological domesticated ruminant bones and teeth. *Fennoscandia Archaeologica* (2016).
17. Virkajärvi, P. *et al*. Dairy production systems in Finland. *Grassland Science in Europe* **20**, 51–66 (2015).
18. Felius, M. *Cattle breeds of the World*. (BRILL, 2024).
19. Tavares, J. C. D. & Almeida, A. M. D. The Portuguese mertolenga cattle breed: a review. *Tropical Animal Health and Production* **56**, 129 (2024).
20. *ENA European Nucleotide Archive* https://identifiers.org/ena.embl:PRJEB76602 (2024).
21. *ENA European Nucleotide Archive* https://identifiers.org/ena.embl:PRJEB56301 (2022).
22. Gao, J. *et al*. Evidence of early genomic selection in Holstein Friesian across African and European ecosystems. *BMC genomics* **26**, 615 (2025).
23. Daetwyler, H. D. *et al*. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. *Nature genetics* **46**, 858–865 (2014).
24. Xia, X. *et al*. Global dispersal and adaptive evolution of domestic cattle: a genomic perspective. *Stress Biology* **3**, 8 (2023).
25. Ajmone-Marsan, P., Boettcher, P., Ginja, C., Kantanen, J. & Lenstra, J. *Genomic characterization of animal genetic resources: Practical guide*. (Food & Agriculture Org., 2023).
26. Bruford, M. W. *et al*. Prospects and challenges for the conservation of farm animal genomic resources, 2015-2025. *Frontiers in genetics* **6**, 314 (2015).
27. 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).
28. Harris, A. M. & DeGiorgio, M. A likelihood approach for uncovering selective sweep signatures from haplotype data. *Molecular biology and evolution* **37**, 3023–3046 (2020).
29. Kircher, M. *et al*. A general framework for estimating the relative pathogenicity of human genetic variants. *Nature genetics* **46**, 310–315 (2014).
30. Groß, C. *et al*. pCADD: SNV prioritisation in Sus scrofa. *Genetics Selection Evolution* **52**, 4 (2020).
31. Gonzalez-Prendes, R. *et al*. Integrative QTL mapping and selection signatures in Groningen White Headed cattle inferred from whole-genome sequences. *PloS one* **17**, e0276309 (2022).
32. Chen, S. F. Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp. *Imeta* **2**, https://doi.org/10.1002/imt2.107 (2023).

33. Vasimuddin, M., Misra, S., Li, H. & Aluru, S. in *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*. 314–324.
34. Faust, G. G. & Hall, I. M. SAMBLASTER: fast duplicate marking and structural variant read extraction. *Bioinformatics* **30**, 2503–2505 (2014).
35. Danecek, P. *et al*. Twelve years of SAMtools and BCFtools. *Gigascience* **10**, giab008 (2021).
36. Okonechnikov, K., Conesa, A. & García-Alcalde, F. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics* **32**, 292–294 (2016).
37. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. *arXiv preprint arXiv:1207.3907* (2012).
38. 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).
39. Danecek, P. *et al*. The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158, https://doi.org/10.1093/bioinformatics/btr330 (2011).
40. Cingolani, P. *et al*. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *fly* **6**, 80–92 (2012).
41. McLaren, W. *et al*. The Ensembl Variant Effect Predictor. *Genome Biol* **17**, 122, https://doi.org/10.1186/s13059-016-0974-4 (2016).
42. Purcell, S. *et al*. PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American journal of human genetics* **81**, 559–575 (2007).
43. Gómez-Rubio, V. ggplot2-elegant graphics for data analysis. *Journal of statistical software* **77**, 1–3 (2017).
44. Chen, C. *et al*. TBtools: an integrative toolkit developed for interactive analyses of big biological data. *Molecular plant* **13**, 1194–1202 (2020).
45. *ENA European Nucleotide Archive* https://identifiers.org/ena.embl:PRJEB90816 (2025).
46. *EVA European Variation Archive* https://identifiers.org/ena.embl:PRJEB98152 (2025).
47. *EVA European Variation Archive* https://identifiers.org/ena.embl:PRJEB94144 (2025).
48. 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).
49. Pandya, M. *et al*. A modern approach for epitope prediction: identification of foot-and-mouth disease virus peptides binding bovine leukocyte antigen (BoLA) class I molecules. *Immunogenetics* **67**, 691–703, https://doi.org/10.1007/s00251-015-0877-7 (2015).
50. Lali, F., Anilkumar, K. & Aravindakshan, T. Novel SNP and Unique Sequences in ATP-binding Cassette Super Family-G Member-2 Transporter (ABCG2) Gene of Vechur cattle (Bos indicus). *Indian Journal of Animal Research* **52**, 1413–1415 (2018).
51. Hellborg, L. & Ellegren, H. Low levels of nucleotide diversity in mammalian Y chromosomes. *Molecular Biology and Evolution* **21**, 158–163 (2004).

Acknowledgements

We thank all members of the OPTIBOV consortium for their invaluable contributions to sample collection and sequencing. This work was supported by the EU-Africa LEAP-Agri (OPTIBOV project, LEAP-Agri-326), the European Union's Horizon 2020 research and innovation program (grant No. 727715). For Finland, we express our gratitude to the Research Council of Finland (No. 319987) and the national organizations managing the LEAP-Agri funding for their financial support. We also thank the breeders and breed associations for preserving these important local genetic resources and providing access to animals. We are particularly grateful to the Natural Resources Institute Finland, especially Tiina Reilas and Tuula-Marjatta Hamama, for their essential role in coordinating sampling and supporting the work involving Eastern Finncattle, Northern Finncattle, Western Finncattle, and Holstein breeds. For the Netherlands, we acknowledge the Dutch Belted Breeders' Association (Vereniging lakenvelder Runderen), Groninger White headed (Blaarkop Stichting), Meuse-Rhine-Yssel Study (MRY Studieverenigingen Zuid en Oost), Deep Red Cattle (Vereniging Het Brandrode Rund), and Dutch Friesian (Fries Hollands Rundvee Stamboek). For Portugal, we acknowledge Fundação Nacional para a Ciência e a Tecnologia (FCT) (2020.02754.CEECIND/CP1601/CP1649/CT0008 and Leap Agri-326/LEAPAgri/0003/2017). We acknowledge the collaboration of breeders and breed associations in Portugal for preserving local genetic resources and providing access to animals, including Filipe Ribeiro and Ricardo Loureiro (Escola Profissional de Agricultura e Desenvolvimento Rural de Vagos, Holstein-Friesian), José Leite and Rui Dantas (AMIBA – Associação dos Criadores de Bovinos de Raça Barrosã), José Pais and Nuno Henriques (ACBM - Associação de Criadores de Bovinos Mertolengos), and Valter Raposo (Associação de Criadores de Bovinos de Raça Mirandesa). We also thank Susana Lopes, Sofia Mourão, and Patrícia Ribeiro from CTM – Centro de Testagem Molecular, CIBIO, Vairão, for their support in sample collection across Barrosã, Mirandesa, Mertolenga, and Holstein breeds. Additional funding from the China Scholarship Council (CSC, grant 202208610017) is acknowledged.

Author contributions

C.G. and R.C. conceived the study. C.G. and J.G. drafted the manuscript and interpreted the data. C.G. defined reference data sets. J.G. participated in data analysis. C.G., J.K., N.G., D.K., M.M., and R.C. were responsible for sample collection. A.P., A.U., B.D., C.S., D.G., D.R., E.B., H.S., H.L., H.B., K.L., K.P., L.B., M.W., R.P., R.O., S.G., and Y.L. of the European OPTIBOV consortium contributed to the sample collection and data generation. R.C. supervised the study. All the authors read and approved the 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/s41597-025-06188-x.

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