Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Unravelling Population Structures and Selection Signatures for milk production and adaptive traits in Genetically Admixed Population of Sub-Saharan African Cattle

Kazeem Ajasa Badmus
alqasimgenelibrary@gmail.com

Federal University Gashua

Jayesh Vyas
ICAR-National Dairy Research Institute

Anil Chitra
ICAR-National Dairy Research Institute

Ishmeet Kumar
ICAR-National Dairy Research Institute

Ritik Kumar Singh
ICAR-National Dairy Research Institute

Sabyasachi Mukherjee
ICAR-National Dairy Research Institute

Anupama Mukherjee
ICAR-National Dairy Research Institute

Article

Keywords: Sub-Saharan African cattle, Population structure, admixture, Random Forest, selection signatures, QTL

Posted Date: August 11th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10489550/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.

Abstract

Sub-Saharan African cattle breeds such as Kuri, N'Dama, and Bororo represent unique genetic resources with distinct adaptive traits and production potentials. Understanding their genomic diversity and population structure is critical for conservation and genetic improvement. In this study, we analysed high-density SNP genotyping data to investigate population structure, breed classification and signatures of selection for milk production and adaptation of these cattle. Random forest classification identified twenty highly informative SNPs for breed differentiation and revealed traceability and assignment of the breeds. Principal component analysis (PCA), admixture, and neighbour-joining (NJ) trees revealed distinct clustering patterns, consistent with the breeds' evolutionary histories. Manhattan plot illustration, F_{ST} and gene enrichment analyses confirmed overlaps with known milk production and adaptation QTLs. Manhattan plots and subsequent annotation of candidate regions identified genes associated with both milk production (*APP, CHODL, ROBO1, EPHA3, GAP43, KALRN, EPHB1, BRWD1, DSCAM* and *COL18A1*) and adaptive and other traits (*PDCD10, TBC1D5, CHN1, RAPGEF4, ERBB4, DOCK10, DDR2, ROR1, NGEF* and *AGAP1*). In conclusion, random forest, a machine learning approach and breed informative markers could be used for breed traceability, assignment and assessment of genetic diversity meant for cattle breed conservation and improvement. Therefore, this finding provide new insights into the genomic architecture of African cattle and support strategies for their sustainable management and utilization.

Introduction

African cattle breeds are critical for food security and rural livelihoods due to their adaptation to diverse ecological zones and tolerance to harsh conditions. N'Dama cattle are renowned for trypanotolerance, an adaptive trait enabling survival in tsetse fly-infested areas [1]. Kuri cattle in the other hand, localised around Lake Chad, are valued for unique morphological features and adaptation to aquatic environments [2]. Bororo (Zebu) cattle in addition, are widely distributed across West Africa with genetic contributions to productivity and resilience [3]. Modern genomic tools, including high-density SNP arrays, enable detailed investigations into population structure, inbreeding, and signatures of selection. High-density SNP arrays and other contemporary genomic technologies allow for in-depth studies of population structure, inbreeding, and selection signatures. Random forest machine learning approaches enhance breed classification and identify highly informative markers [4]. Such analyses are also critical for breed traceability and for identifying genetic variants contributing to adaptation and productivity in livestock [5]. A central component of population structure analysis is the identification and use of breed-informative markers-SNPs with high discriminatory power across populations. Marker informativeness can be quantified using allele frequency differences and information-theoretic measures [6]. One widely used statistic, δ (delta), measures the absolute allele frequency difference between breeds and is effective for identifying markers that maximize population differentiation [7]. Such markers have applications in breed assignment, genomic traceability, and conservation programmes [8].

Beyond classical methods for population structure analysis, machine learning approaches such as Random Forest classification [8] have gained much attention in evaluating the power of informative SNPs for breed assignment. Predictive accuracy of such classification models is often assessed using a confusion matrix, which quantifies misclassification rates and the reliability of breed assignment [9, 10].

Principal component analysis (PCA) and admixture in the other hand are widely applied basically to infer breed relationships and ancestry history [11, 12]. Understanding population structure and genetic relationships among cattle breeds using high-throughput genotyping technologies and high-density SNP arrays is fundamental for effective utilization, conservation, and genetic improvement, in evaluate admixture, and design effective management strategies [13, 14, 15]. Phylogenetic methods, such as the Neighbor-Joining (NJ) tree, summarize genetic distances to depict evolutionary relationships and divergence among populations [16]. Multivariate analyses, including Principal Component Analysis (PCA), provide a dimension-reduction framework to visualise clustering patterns of individuals and populations [11]. genome-wide visualization approaches further enhance interpretation of marker informativeness. Manhattan plots are frequently used to display genome-wide distributions of allele frequency differences or informativeness statistics, highlighting highly differentiated loci that may represent breed-defining markers or candidate regions under selection [17].

Insights into the genetic makeup and selection signature profile of admixed population of Bororo, Kuri and N'Dama cattle exposed to tropical environments is the aim of this study. Based on neutral evolutionary theory, a number of statistical techniques have been developed to identify selection footprints [18]. Various approaches have been developed to detect selection signatures within and between cattle populations, including Tajima's D, runs of homozygosity (ROH), integrated haplotype score (iHS), XP-EHH, F_{ST}, and principal component analysis (PCA) [19, 20]. Application of these methods has revealed numerous genomic regions associated with productivity and environmental adaptation in cattle [21]. Selection scans using F_{ST} and Manhattan plots have identified candidate genes linked to milk production (*PRL*, *CSN2*, *DGAT1*) and adaptation-related traits, including disease resistance and environmental resilience (*BoLA*, *TICAM1*, and *HSP* genes) [22–26]. Despite these advances, limited information exists on recent positive selection and adaptive evolution in Kuri, N'Dama, and Bororo cattle [27]. Therefore, this study integrates phylogenetic analysis, admixture inference, and genome-wide selection scans to identify genomic regions and candidate genes underlying milk production and adaptive traits in these African indigenous breeds.

Materials and Methods

A total number of 78 cattle of African origin were used for this study. The cattle are made up of 30 Kuri, 25 N'Dama and 23 Zebu Bororo breeds of Sub-Saharan African origin. The genotype data for the African cattle breeds analysed for this study were obtained from the publicly accessible WIDDE (Web-Interfaced Next Generation Database for Genetic Diversity Exploration) repository.

Data availability: The genotyping data of African cattle breeds used for conducting this research were already deposited into the WIDDE public data repository. http://widde.toulouse.inra.fr/widde/widde/main.do?module=cattle

Genotype Data and Quality Control

High-density SNP genotype data for Kuri, N'Dama, and Bororo cattle were obtained from publicly available datasets and processed using PLINK v1.9 [28]. Quality control (QC) steps included: Filtering SNPs with call rate < 95%, removing SNPs with minor allele frequency (MAF) < 0.05, excluding SNPs deviating from Hardy-Weinberg equilibrium ($p < 1 \times 10^{-6}$), removing individuals with >10% missing genotypes. After QC, the pruned dataset was used for downstream analyses. Allele frequency distributions were calculated across populations to assess variation within and between breeds.

Genetic differentiation

To identify informative SNPs for breed differentiation, Weir & Cockerham's F_{ST} [29] was calculated in PLINK to quantify genetic differentiation among breeds. Delta (Δ) allele frequency was computed as the absolute difference in allele frequency between breeds. Allele frequency differences (Δ) were calculated for each pairwise breed comparison. SNPs exhibiting high Δ values in one or more comparisons were considered informative markers for breed discrimination and were retained. Then, informativeness for assignment (I) was estimated following [6], where SNPs with higher I values are more effective for population classification. Random forest classification was conducted in R (randomForest package) [30]. Individuals were classified into their respective breeds (Kuri, N'Dama, Bororo) using genotype data. The algorithm was trained on 70% of the data and tested on the remaining 30% [8]. SNP importance scores (Mean Decrease in Gini and Mean Decrease in Accuracy) were extracted to rank highly informative SNPs [8, 30]. Classification performance was assessed using a confusion matrix, overall accuracy, precision, recall, and F1-score. These metrics quantified the ability of the random forest model to correctly assign individuals to their breeds [31].

Population Structure and admixture Analysis

Population stratification was examined using smartPCA from the EIGENSOFT v7.2.1 package [4]. PCA was performed with default settings, and the top principal components (PCs) were extracted. To determine the optimal number of informative PCs, a scree plot was generated in R (ggplot2), where eigenvalues were plotted against PC rank. Genetic distances were estimated from allele frequencies using Nei's genetic distance in PHYLIP v3.697 [32, 33] A neighbor-joining (NJ) tree was then constructed

to visualize evolutionary relationships among the three breeds. The resulting dendrogram was visualized in FigTree v1.4.4 [16]. Population admixture was evaluated using admixture v1.3.0. Cross-validation was applied to select the optimal number of ancestral clusters (K = 2–6) [1]. All statistical analyses and visualizations were conducted in R v4.3.1.

Detection of Selection Signatures and Quantitative Trait Loci

Population differentiation and selection signatures were assessed using pairwise F_{ST} estimates among Kuri, N'Dama, and Bororo cattle. Pairwise SNP-based F_{ST} values were calculated in VCFtools v0.1.17 [34] for Kuri–N'Dama, Kuri–Bororo, and N'Dama–Bororo comparisons. Genomic regions under putative selection were identified by summarizing F_{ST} values in 100-kb sliding windows with 50-kb steps across autosomes, excluding windows containing fewer than five SNPs. Mean and maximum F_{ST} values per window were visualized using Manhattan plots in R (ggplot2) [13]. Candidate regions were annotated using GALLO, with gene information obtained from Ensembl BioMart and QTL overlaps identified through the Animal QTLdb. Functional enrichment analyses (GO) were performed using clusterProfiler [13].

Results

Analysis of allele frequency differences (ΔMAF) across Kuri, N’Dama, and Bororo cattle revealed substantial divergence at multiple SNP loci (Table 1). The ΔMAF values ranged from 0.8261 to 0.873, indicating pronounced allele frequency differentiation between breeds. For example, ARS-BFGL-NGS-4016 (BTA 13) showed the highest differentiation (ΔMAF = 0.873), with a major allele frequency of 0.913 in one population compared to 0.04 in another. Similarly, SNPs such as ARS-BFGL-NGS-117151 (BTA 24) and BTA-58423-no-rs (BTA 24) exhibited ΔMAF values > 0.87, suggesting strong divergence at chromosome 24. Notably, SNPs on chromosomes (BTA) 7, 10, 11, 14, and 26 also exhibited high ΔMAF (> 0.82), implying breed-specific allele fixation or near-fixation events. These results point to selective pressures or historical genetic drift that may have contributed to differentiation among Kuri, N’Dama, and Bororo cattle.

Table 1
Delta between Kuri, N'Dama and Bororo

SNP	CHR	A1	A2	MAF_b1	MAF_b2	MAF_b3	Delta_MAF
ARS-BFGL-NGS-4016	13	C	A	0.4333	0.04	0.913	0.873
ARS-BFGL-NGS-117151	24	A	G	0.5	0.02	0.8913	0.8713
BTA-58423-no-rs	24	A	C	0.4667	0.02	0.8913	0.8713
ARS-BFGL-NGS-111395	14	A	G	0.4167	0	0.8696	0.8696
BTB-00737840	7	A	G	0.5167	0.02	0.8696	0.8496
ARS-BFGL-NGS-30236	10	A	G	0.5167	0	0.8478	0.8478
ARS-BFGL-NGS-4101	11	A	C	0.4833	0	0.8478	0.8478
BTB-00528306	13	G	A	0.3333	0	0.8478	0.8478
Hapmap31774-BTA-145190	7	G	A	0.5333	0	0.8478	0.8478
Hapmap33430-BTC-037618	6	A	G	0.5	0	0.8478	0.8478
ARS-BFGL-NGS-21471	24	A	G	0.4833	0.06	0.8913	0.8313
ARS-BFGL-NGS-104836	14	A	C	0.65	0	0.8261	0.8261
ARS-BFGL-NGS-110287	18	A	G	0.55	0	0.8261	0.8261
ARS-BFGL-NGS-36064	26	G	A	0.4333	0	0.8261	0.8261
Hapmap31876-BTA-60880	26	A	G	0.4333	0	0.8261	0.8261
MAF = Major Allelic Frequency							

Table 2 presents the breed informative markers (BIMs) detected among Kuri, N'Dama, and Bororo cattle based on allelic frequency differences and informativeness for assignment index (I). The most informative SNPs included ARS-BFGL-NGS-4016 (CHR13), ARS-BFGL-NGS-117151 (CHR24), ARS-BFGL-NGS-104836 (CHR14), and ARS-BFGL-NGS-11407 (CHR10), each exhibiting high heterozygosity (H) and substantial between-breed differentiation (H). SNPs such as ARS-BFGL-NGS-11407 and Hapmap23685-BTA-132541 also showed strong allele frequency contrasts among the three breeds, making them highly discriminative for population assignment. The informativeness index (I) ranged from 0.2358 to 0.2549, indicating moderate to high breed informativeness across loci. Notably, SNPs on chromosomes 13, 24, and 14 displayed the highest I values, suggesting their potential utility in distinguishing Bos taurus-adapted breeds (N'Dama) from Bos indicus-influenced breeds (Bororo and Kuri).

Table 2
Breed Informative Markers among Kuri, N'Dama and Bororo

SNP	CHR	A1	A2	MAF_KURI	MAF_N'DAMA	MAF_BORORO	H_T	H_S_	I_n
ARS-BFGL-NGS-4016	13	C	A	0.4333	0.04	0.913	0.497127	0.242255	0.254872
ARS-BFGL-NGS-117151	24	A	G	0.5	0.02	0.8913	0.498252	0.244323	0.253929
BTA-58423-no-rs	24	A	C	0.4667	0.02	0.8913	0.496692	0.243584	0.253109
ARS-BFGL-NGS-104836	14	A	C	0.65	0	0.8261	0.499873	0.247439	0.252434
ARS-BFGL-NGS-111395	14	A	G	0.4167	0	0.8696	0.489852	0.237638	0.252214
ARS-BFGL-NGS-11407	10	A	G	0.2833	1	0.2174	0.5	0.248786	0.251214
Hapmap41813-BTA-27442	5	A	G	0.3167	0.14	0.9565	0.498326	0.252273	0.246053
Hapmap31774-BTA-145190	7	G	A	0.5333	0	0.8478	0.496858	0.251951	0.244908
ARS-BFGL-NGS-30236	10	A	G	0.5167	0	0.8478	0.49592	0.252504	0.243416
BTB-00528306	13	G	A	0.3333	0	0.8478	0.477401	0.234164	0.243236
Hapmap33430-BTC-037618	6	A	G	0.5	0	0.8478	0.494852	0.25269	0.242162
ARS-BFGL-NGS-4101	11	A	C	0.4833	0	0.8478	0.493661	0.252504	0.241156

MAF = Major Allelic Frequency

The principal component analysis (PCA) revealed clear genetic separation among the three cattle breeds (Fig. 1). The first principal component (PC1) explained 47.53% of the total genetic variation and effectively distinguished Bororo (left cluster) from N'Dama (right cluster). The second component (PC2), accounting for 9.66% of the variation, separated Kuri from the other breeds, clustering them distinctly in the upper part of the plot. The tight clustering of individuals within each breed reflects a high degree of within-breed genetic homogeneity, while the wide separation across breeds demonstrates substantial between-breed genetic divergence. The scree plot (Fig. 2) illustrates the proportion of genetic variation explained by the first ten principal components (PCs). PC1 accounted for the largest share of the total variation (47.53%), followed by PC2 (9.66%). Together, PC1 and PC2 explained 57.19% of the total variance, indicating that the first two axes captured the majority of the genetic structure present in the dataset.

The admixture analysis at K = 3 (Fig. 3) revealed clear genetic differentiation among the three cattle breeds. Bororo individuals were predominantly assigned to the K1 (red) ancestral cluster, with only minor contributions from K2 and negligible K3 ancestry. N'Dama cattle showed almost complete assignment to the K3 (green) cluster, indicating a highly homogeneous genetic background. In contrast, Kuri cattle were primarily composed of the K2 (blue) ancestral component but exhibited noticeable contributions from both K1 and K3, suggesting a mixed ancestry pattern. The degree of admixture was highest in Kuri, whereas Bororo and N'Dama displayed relatively uniform ancestry profiles. The strong assignment of Bororo and N'Dama individuals to distinct ancestral clusters indicates substantial genetic differentiation between these breeds.

The Neighbor-Joining (NJ) tree based on pairwise genetic distances (Fig. 4) revealed distinct clustering patterns among the three cattle populations. Kuri individuals (red) formed a separate cluster, highlighting their unique genetic divergence from the other populations. N'Dama (blue) and Bororo (green) showed a closer genetic relationship, with individuals from both

populations clustering more tightly together. The tree topology indicates that Kuri is genetically distant, while N'Dama and Bororo share a more recent common ancestry. The NJ tree provides further evidence of the genetic distinctiveness of Kuri cattle, consistent with results from PCA and admixture analyses.

The Random Forest analysis identified the top 20 most informative SNPs contributing to population differentiation (Fig. 5). The SNP ARS-BFGL-NGS-... showed the highest importance with the largest mean decrease in Gini index, indicating its strong discriminatory power. Several other SNPs within the ARS-BFGL-NGS panel ranked highly, followed by Hapmap595 loci and others. The cumulative importance line (orange) shows that these top SNPs together explain a substantial proportion of the model's classification accuracy.

The confusion matrix illustrates the classification accuracy of the Random Forest model in assigning individuals to their correct breeds (Figs. 6 and 7). Out of the 9 true Kuri individuals, all were correctly classified (100% accuracy). For Bororo individuals, 4 were misclassified as Kuri and 4 as Bororo, showing mixed assignment. N'Dama individuals were misclassified entirely, with 7 being assigned to N'Dama incorrectly. Generally, the model achieved high accuracy for Kuri, but misclassification rates were higher in Bororo and N'Dama.

Manhattan plot revealed the genome-wide distribution of population differentiation (F_{ST}) across autosomes (Fig. 8). This outlined multiple regions with elevated differentiation between populations. The majority of SNPs showed moderate differentiation ($F_{ST} < 0.25$), suggesting shared ancestry and gene flow across populations. Several SNPs exceeded the threshold line ($F_{ST} > 0.5$), indicating candidate regions under strong divergent selection. Peaks of high F_{ST} values were distributed across several chromosomes, particularly on CHR6, CHR14, CHR16, and CHR20, which are well-known to harbour QTLs linked to milk production, reproduction, immunity, and adaptation in cattle. The clustering of highly differentiated SNPs suggests the presence of selective sweeps, reflecting adaptation to environmental and management pressures in the studied populations.

Functional enrichment analysis of candidate regions under selection (Table 3) revealed several significant Gene Ontology (GO) terms (adjusted $p < 0.05$) associated with milk production, adaptation, and heat stress responses. The top enriched categories included cell morphogenesis (GO:0000902, p.adjust = 2.64E-09) and neuron development (GO:0048666, p.adjust = 1.57E-06), both linked to milk yield, udder tissue development, and heat-stress resilience. Genes such as *APP, ROBO1, EPHA3, GAP43*, and *DSCAM* were consistently present across morphogenesis and neuron-related GO terms, suggesting their broad involvement in mammary gland development and adaptive physiology. Additionally, enrichment in positive regulation of catalytic activity (GO:0043085) and regulation of GTPase activity (GO:0043087) highlighted genes like *RAPGEF4, DOCK10, ARHGAP15*, and *ERBB4*, which are implicated in cell signalling, stress response, and metabolic regulation. Such processes are critical for maintaining cellular homeostasis under environmental stressors, including heat stress prevalent in African cattle production systems.

Table 3
F_{ST} GO Enrichment candidate Genes for milk production and adaptive traits

ID	Description	p.adjust	Classification	Genes
GO:0000902	cell morphogenesis	2.64E-09	Milk Production, Adaptation, Heat Stress	*APP, CHODL, ROBO1, EPHA3, GAP43, EPHB1, BRWD1, COL18A1*
GO:0043085	positive regulation of catalytic activity	4.42E-08	Other, Adaptation	*PDCD10, CHN1, RAPGEF4, ERBB4, DDR2, ROR1, NGEF, AGAP1*
GO:0043087	regulation of GTPase activity	3.07E-09	Other, Adaptation	*RASA2, TBC1D5, CHN1, ALS2, DOCK10, IQGAP3, VAV3*
GO:0044093	positive regulation of molecular function	4.79E-07	Other, Adaptation	*IL1RAP, MAP3K13, RIPK4, TBC1D5, CHN1, RNF25*
GO:0048666	neuron development	1.57E-06	Milk Production, Adaptation, Heat Stress	*DSCAM, CYFIP1, NEUROD1*
GO:0050790	regulation of catalytic activity	8.61E-07	Other, Adaptation	*CBLB, RASA2, SH3BP5, TFPI, ARHGAP15*
GO:0120039	plasma membrane bounded cell projection morphogenesis	1.45E-06	Milk Production, Adaptation, Heat Stress	*GAP43, KALRN, DSCAM, CYFIP1, NRP2*

Discussion

High Δ values indicate substantial allele frequency differences among breeds and identify highly informative loci for breed discrimination. The extreme allele frequency shifts observed at some loci (Δ > 0.8) suggest strong genetic differentiation, which may have arisen through directional selection acting on breed-specific adaptive traits, or through demographic processes such as genetic drift, founder effects, population bottlenecks, and restricted gene flow. Consequently, these loci may harbour genes associated with breed-specific adaptations and warrant further investigation. For instance, N'Dama cattle are renowned for trypanotolerance and disease resistance [34, 35], while Kuri are adapted to aquatic environments of Lake Chad [34, 36], and Bororo (Zebu) excel in heat tolerance and long-distance trekking [37]. Divergence at these SNPs may thus reflect adaptive differences related to environmental stress, disease, or production traits. Particularly, the clustering of highly differentiated SNPs on chromosome 24 may indicate a selective sweep or QTL region linked to important adaptive traits such as fertility or immunity [38]. Similarly, SNPs on chromosome 14, a chromosome frequently associated with milk yield and body conformation traits in cattle [39], may suggest differentiation linked to production trade-offs between taurine (N'Dama, Kuri) and indicine (Bororo) breeds. Generally, the large allele frequency differences suggest that genetic differentiation among these breeds is not random, but shaped by both natural selection (environment, disease pressure, climate adaptation) and artificial selection (management and breeding objectives). These SNPs may represent valuable markers for understanding local adaptation and for designing genomic-assisted breeding programmes tailored to African cattle production systems.

The PCA results highlight the strong genetic structuring of the cattle populations studied. The clear separation along PC1 and PC2 underscores the unique evolutionary histories of Bororo, Kuri, and N'Dama, shaped by geographical isolation, environmental adaptation, and selective breeding. The clustering pattern confirms previous reports of independent domestication and subsequent admixture events in African cattle populations [13, 40]. The distinct positioning of N'Dama supports its classification as an African taurine breed, historically known for trypanotolerance and adaptation to humid environments [1, 41]. Bororo, belonging to the indicine lineage, shows strong separation from N'Dama, consistent with its zebu ancestry and adaptation to arid Sahelian environments [40]. kuri cattle, clustering independently along PC2, reflects their unique genetic background and adaptation to the Lake Chad Basin, corroborating earlier findings of their genetic distinctiveness due to geographic isolation and local adaptation [42, 43]. Therefore, the PCA demonstrates that historical admixture, selective pressures, and environmental adaptation have contributed to shaping the genetic diversity of these African cattle breeds [44]. These results reinforce that African cattle breeds harbour both shared and unique genetic signatures, shaped by domestication, gene flow, and local adaptation [1, 24].

The admixture analysis showed near-exclusive K3 ancestry in N'Dama, confirming its distinct West African taurine origin and previously reported trypanotolerant adaptation [40, 45]. Bororo formed a separate K1 cluster, consistent with its predominantly zebu ancestry and genetic differentiation from African taurine cattle [13, 46]. In contrast, Kuri exhibited the highest level of admixture, with dominant K2 ancestry and contributions from both Bororo-like (K1) and N'Dama-like (K3) components, indicating historical taurine–zebu gene flow. The geographic position of Kuri within the Lake Chad Basin likely facilitated such admixture through livestock movement and exchange [46, 47]. While N'Dama and Bororo maintained relatively pure ancestral backgrounds, Kuri represented an admixed population derived from both taurine and Zebu lineages. This genetic admixture, coupled with habitat loss and population decline, may contribute to the threatened status of Kuri cattle. These findings underscore the importance of conserving the unique genetics.

The separation of Kuri in the neighbour joining (NJ) tree supports earlier findings of their isolation and adaptation to the Lake Chad environment, which has likely contributed to their unique genetic background [1, 43]. The close clustering of N'Dama and Bororo suggests shared evolutionary history, reflecting their taurine ancestry and historical gene flow within West African cattle populations [13]. Together, these results indicate a triangular population structure. Kuri showes a genetically distinct feature and possibly under unique selective pressures. N'Dama and Bororo: closely related, forming a taurine genetic background cluster. The concordance of NJ tree topology with admixture and PCA highlights the robustness of the findings and emphasizes the role of both demography and selection in shaping the genetic structure of these populations.

Breed informative markers (BIMs) are crucial for differentiating cattle populations, assessing admixture, and guiding conservation and breeding programmes [48, 49]. The identified SNPs exhibited strong differentiation among Kuri, N'Dama, and Bororo, reflecting their unique evolutionary histories and adaptations. N'Dama cattle, known for their trypanotolerance and West African taurine ancestry, consistently displayed low MAFs for several loci (e.g., ARS-BFGL-NGS-111395 and Hapmap31774-BTA-145190), supporting their genetic distinctiveness [1, 50]. Bororo (Zebu) cattle, by contrast, carried high MAFs at these loci, consistent with indicine ancestry and adaptive introgression [49]. Kuri cattle, with their mixed taurine-indicine background, generally showed intermediate allele frequencies, suggesting historical admixture and local adaptation [51]. The presence of highly informative SNPs on chromosomes 13 and 24 may be linked to regions under selection for adaptive traits, as previous studies have associated these chromosomes with immune response, heat tolerance, and production traits [52, 37]. The ability of these SNPs to clearly separate the three cattle breeds highlights their potential for use in breed traceability, conservation genetics, and genomic selection programmes in West Africa.

The high-ranking SNPs revealed by the Random Forest analysis are strong candidate markers for population differentiation. SNPs with high mean decrease in Gini suggest they may be under selection pressure or linked to genomic regions influencing adaptation and breed-specific traits [53]. The prevalence of ARS-BFGL-NGS markers among the top SNPs highlights their broad informativeness in cattle population genetics, consistent with earlier genomic studies [49, 54]. Such SNPs can serve as potential biomarkers for further studies, including selection signature scans, GWAS, and breed conservation programmes [55, 56]. The Random Forest results complement PCA, admixture, and NJ tree findings, reinforcing that specific loci play a key role in distinguishing Kuri, N'Dama, and Bororo cattle populations [57,58].

The confusion matrix highlights differences in breed assignment power: The perfect classification of Kuri suggests the presence of strong, distinct genetic markers unique to this population, consistent with their known genetic isolation and limited admixture [36, 59]. The misclassification of Bororo and N'Dama reflects possible genetic overlap between these breeds, potentially due to historical admixture, shared ancestry, or insufficient breed-specific markers captured in the SNP set [60, 61]. These findings align with results from PCA, admixture, and NJ tree analyses, which also showed partial clustering and overlaps between Bororo and N'Dama, but clear separation of Kuri [57,58]. From a practical perspective, while Kuri cattle can be reliably identified, improving SNP panels with more informative loci may enhance discrimination between Bororo and N'Dama [55, 56].

The genome-wide F_{ST} distribution revealed both shared ancestry and population-specific selection among the three cattle breeds. Elevated F_{ST} peaks on CHR6 and CHR14 correspond to regions previously associated with milk production, carcass quality, and reproductive traits [62,63], while signals on CHR16 and CHR20 overlapped loci linked to disease resistance and environmental adaptation in African cattle [1, 41]. The moderate overall F_{ST} values reflect the historical admixture between

African taurine and indicine cattle, alongside distinct signatures of local adaptation [14, 40]. Collectively, these patterns suggest that both natural and artificial selection have shaped genomic variation related to productivity, reproduction, and adaptation. Functional enrichment analyses identified genes involved in morphogenesis, mammary gland development, neuronal regulation, and signal transduction. Candidate genes including *EPHA3*, *ROBO1*, *ERBB4*, and *DDR2* are associated with tissue development and lactogenic processes, whereas *APP*, *GAP43*, *RAPGEF4*, and *DOCK10* may contribute to stress resilience, immune function, trypanotolerant and environmental adaptation [64–68]. These results indicate that selection in indigenous cattle has targeted pathways underlying both milk production and adaptation to harsh African production environments.

Conclusion

High-density SNP analyses revealed clear genetic differentiation and population structure among African cattle breeds, consistent with their evolutionary backgrounds. PCA, admixture, and NJ tree analyses confirmed breed-specific clustering. The perfect classification of Kuri, N'Dama, and Bororo cattle highlights the presence of breed-specific genetic signatures demonstrating that machine learning approaches, combined with informative SNP markers, provide powerful tools for assessing genetic diversity, population structure, breed traceability, and the conservation and improvement of African cattle genetic resources. F_{ST}, and gene enrichment analyses highlighted genomic regions overlapping with production and adaptation QTLs, underscoring selection for both milk yield and environmental resilience. Candidate genes such as *APP, ROBO1, EPHA3*, and *GAP43* were linked to heat tolerance and adaptation, while *DOCK10, DDR2*, and *ERBB4* were associated with adaptive and functional traits. Therefore, these findings enhance understanding of the genomic architecture and adaptive evolution of African cattle, providing a strong foundation for genetic improvement, conservation, and sustainable breed management.

Declarations

Ethics

This work did not require ethical approval from animal welfare committee as all the analyses were performed using pre-existing databases and no new animal procedures were performed for this genomic analysis.

Data availability

The genotyping data of African cattle breeds used for conducting this research were already deposited into the WIDDE public data repository. http://widde.toulouse.inra.fr/widde/widde/main.do?module=cattle

Supplementary data

The supplementary data for this work were attached along with the submission of this manuscript.

Authors' contributions

B.K.A., A.M and S.M.: conceptualization, data curation, funding acquisition, investigation, project administration, resources, supervision; J.V., A.C. and I.K..: data curation, formal analysis, investigation, writing-review and editing; validation, visualization, writing-original draft; R.K.S.: methodology, resources, software, writing-review and editing; J.V.: conceptualization, writing-review and editing

All authors gave final approval for publication and agreed to be held accountable for the work performed therein.

Conflict of interest declaration

We declare no competing interests for this study.

Funding

Enabling research environment and study grant was provided by UNESCO-DBT-TWAS-2023 with UTR number-ICIC52025090900536591; date-09092025 and hosted by ICAR-National Dairy Research Institute, Karnal, India and is

acknowledged.

Acknowledgements

We gratefully acknowledged research support grant provided by UNESCO-DBT-TWAS to carry out this study and the enabling environment and teaching aid provided by the Animal Genetics and Breeding Division, ICAR-National Dairy Research Institute, India and Federal University Gashua, Yobe State, Nigeria.

References

1. Hanotte, O. et al. Mapping of quantitative trait loci controlling trypanotolerance in a cross of tolerant West African N'Dama and susceptible East African Boran cattle. *Proc. Natl. Acad. Sci. USA*. **100** (13), 7443–7448. https://doi.org/10.1073/pnas.1232392100 (2003).
2. Rege, J. E. O. & Tawah, C. L. The state of African cattle genetic resources II. Geographical distribution, characteristics and uses of present-day breeds and strains. *Anim Genet Resour Inf*. ;26:1–25. (1999). https://doi.org/10.1017/S1014233900001148 (1999).
3. Tawah, C. L., Rege, J. E. O. & Aboagye, G. S. A close look at a rare breed - The Kuri cattle of Lake Chad Basin: Origin, distribution, production and adaptive characteristics. *S Afr. J. Anim. Sci.* **27** (2), 31–40 (1997).
4. Radovan, K., Nina, M., Gábor, M., Mojca, S. & Daniel, Z. Classification of cattle breeds based on the random forest approach, *Livest. Sci*, 267, 105143. (2023). https://doi.org/10.1016/j.livsci.2022.105143 (2023).
5. Gandini, G. C. & Oldenbroek, J. K. Strategies for moving from conservation to utilization. In: (ed Oldenbroek, J. K.) Utilisation and Conservation of Farm Animal Genetic Resources. Wageningen Academic, 29–54 (2007).
6. Rosenberg, N. A., Li, L. M., Ward, R. & Pritchard, J. K. Informativeness of genetic markers for inference of ancestry. *Am. J. Hum. Genet.* **73** (6), 1402–1422. https://doi.org/10.1086/379178 (2003).
7. Sham, P. C., Cherny, S. S. & Hewitt, J. K. Power of linkage versus association analysis of quantitative traits, by use of variance-components models, for sibship data. *Am. J. Hum. Genet.* **71** (6), 1342–1354 (2000).
8. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
9. Botstein, D., White, R. L., Skolnick, M. & Davis, R. W. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am. J. Hum. Genet.* **May;32** (3), 314–331 (1980). PMID: 6247908; PMCID: PMC1686077.
10. Gondro, C., van der Werf, J. & Hayes, B. (eds). Genome-Wide Association Studies and Genomic Prediction. Humana Press. (2013). 10.1007/978-1-62703-447-0
11. Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2** (12), e190. https://doi.org/10.1371/journal.pgen.0020190 (2006).
12. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Gen. Res.* **19** (9), 1655–1664. https://doi.org/10.1101/gr.094052.109 (2009).
13. Decker, J. E. et al. Worldwide Patterns of Ancestry, Divergence, and Admixture in Domesticated cattle. *PLoS Genet.* **10** (3), e1004254. 10.1371/journal.pgen.1004254 (2014).
14. Medugorac, I. et al. Genetic diversity of European cattle breeds highlights the conservation value of traditional unselected breeds with high effective population size. *Mol. Ecol.* **18** (16), 3394–3410 (2009).
15. Gandini, G. C. & Oldenbroek, J. K. Strategies for moving from conservation to utilization. In: (ed Oldenbroek, J. K.) Utilisation and Conservation of Farm Animal Genetic Resources. Wageningen Academic, 29–54. (2007).
16. Saitou, N. & Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4** (4), 406–425 (1987).
17. Turner, S. D. et al. Quality control procedures for genome-wide association studies. *Curr. Prot. Hum Gen.* 68(1), 1.19.1–1.19.18. (2011).
18. Biswas, S. & Akey, J. M. Genomic insights into positive selection. *Trends Genet.* **22** (8), 437–446. https://doi.org/10.1016/j.tig.2006.06.005 (2006).

19. Vitti, J. J., Grossman, S. R. & Sabeti, P. C. Detecting natural selection in genomic data. *Ann. Rev. Genet.* **47**, 97–120. https://doi.org/10.1146/annurev-genet-111212-133526 (2013).
20. Randhawa, I. A. S., Khatkar, M. S., Thomson, P. C. & Raadsma, H. W. A meta-assembly of selection signatures in cattle. *PONE* **11** (4), e0153013. https://doi.org/10.1371/journal.pone.0153013 (2016).
21. Kambal, S. Candidate signatures of positive selection for environmental adaptation in indigenous African cattle: A review. *Anim. Genet.* **54** (6), 689–708. https://doi.org/10.1111/age.13353 (2023).
22. Voight, B. F., Kudaravalli, S., Wen, X. & Pritchard, J. K. A map of recent positive selection in the human genome. *PLoS Biol*. ;4(3):e72. (2006). https://doi.org/10.1371/journal.pbio.0040072 (2006).
23. Sabeti, P. C. et al. Genome-wide detection and characterization of positive selection in human populations. *Nature*. ;449(7164):913–918. (2007). https://doi.org/10.1038/nature06250 (2007).
24. Grisart, B. et al. Positional candidate cloning of a QTL in dairy cattle: identification of a missense mutation in the bovine DGAT1 gene with major effect on milk yield and composition. *Genome Res*. ;12(2):222–231. (2002). https://doi.org/10.1101/gr.224202 (2002).
25. Flori, L. et al. Adaptive admixture in the West African bovine hybrid zone: insight from the Borgou population. *Mol. Ecol.* **23** (13), 3241–3257. 10.1111/mec.12816 (2014). Epub 2014 Jun 19. PMID.
26. Kim, E. S. et al. Effect of artificial selection on runs of homozygosity in U.S. Holstein cattle. *PONE*. ;8(11):e80813. (2013). https://doi.org/10.1371/journal.pone.0080813 (2013).
27. Kim, E. S. & Rothschild, M. F. Genomic adaptation of admixed dairy cattle in East Africa. *Front. Genet.* **5**, 443. https://doi.org/10.3389/fgene.2014.00443 (2014).
28. Purcell, S. et al. A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81** (3), 559–575. https://doi.org/10.1086/519795 (2007).
29. Weir, B. S. & Cockerham, C. C. Estimating F-statistics for the analysis of population structure. *Evol* **38** (6), 1358–1370. https://doi.org/10.2307/2408641 (1984).
30. Liaw, A. & Wiener, M. Classification and Regression by RandomForest. *Forest*. 23 (2001).
31. Tawah, C. L., Rege, J. E. O. & Aboagye, G. S. A close look at a rare breed - The Kuri cattle of Lake Chad Basin: Origin, distribution, production and adaptive characteristics. *S Afr. J. Anim. Sci.* **27** (2), 31–40 (1997).
32. Zou, Y. et al. Common Methods for Phylogenetic Tree Construction and Their Implementation in R. *Bioengineering* 11, 480. (2024). https://doi.org/10.3390/bioengineering11050480 (2024).
33. Fitch, W. M. & Margoliash, E. Construction of phylogenetic trees. *Science* **155**, 279–284 (1967).
34. Weir, B. S. & Cockerham, C. C. Estimating F-statistics for the analysis of population structure. *Evol* **38** (6), 1358–1370. https://doi.org/10.2307/2408641 (1984).
35. Marshall, K. et al. Livestock genomics for developing countries – African examples in practice. *Front. Genet.* **10**, 297. https://doi.org/10.3389/fgene.2019.00297 (2019).
36. Hoffmann, I. Climate change and the characterization, breeding and conservation of animal genetic resources. *Anim. Genet.* **41** (Suppl. 1), 32–46. https://doi.org/10.1111/j.1365-2052.2010.02043.x (2010).
37. Bahbahani, H. et al. Signatures of selection for environmental adaptation and zebu × taurine hybrid fitness in East African Shorthorn Zebu. *Front. Genet.* **8**, 68. https://doi.org/10.3389/fgene.2017.00068 (2017).
38. Mbole-Kariuki, M. N. Genome-wide analysis reveals the ancient and recent admixture history of East African Shorthorn Zebu from Western Kenya. *Hered* **113** (4), 297–305. https://doi.org/10.1038/hdy.2014.31 (2014).
39. Cole, J. B. et al. Genome-wide association analysis of thirty one production, health, reproduction and body conformation traits in contemporary U.S. Holstein cows. *BMC Genom*. **12**, 408. https://doi.org/10.1186/1471-2164-12-408 (2011).
40. Kim, J. et al. The genome landscape of indigenous African cattle. *Genom Biol.* **18** (1), 34 (2017).
41. Noyes, H. et al. Genetic and expression analysis of cattle identifies candidate genes in pathways responding to trypanosome infection. *PNAS USA*. **108** (22), 9304–9309. https://doi.org/10.1073/pnas.1013486108 (2011).

42. Lauvie, A., Audiot, A., Couix, N., Casabianca, F. & Brives, H. Diversity of rare breed management programs: Between conservation and development. *Livest. Sci.* **140** (1–3), 161–170 (2011).
43. Kugonza et al. Genetic diversity and differentiation of Ankole cattle populations in Uganda inferred from microsatellite data. *Livest. Sci.* **135**, 140–147. 10.1016/j.livsci.2010.06.158 (2011).
44. Bahbahani, H. et al. Signatures of positive selection in African Butana and Kenana dairy zebu cattle. *PONE*, 13(1), e0190446. (2018). https://doi.org/10.1371/journal.pone.0190446 (2018).
45. Hanotte, O. et al. *African Pastoralism: Genetic Imprints of Origins and Migrations* Vol. 296, 336–339 (Science, 2002). 10.1126/science.1069878
46. Edea, Z. et al. Genome-wide genetic diversity, population structure and admixture analysis in African and Asian cattle breeds. *Animal* **9** (2), 218–226. https://doi.org/10.1017/S1751731114002560 (2015).
47. Mwai, O., Hanotte, O., Kwon, Y. J. & Cho, S. African indigenous cattle: unique genetic resources in a rapidly changing world. *Asian-Australas J. Anim. Sci.* **28** (7), 911–921 (2015).
48. Qanbari, S. & PMCID. On the Extent of Linkage Disequilibrium in the Genome of Farm Animals. *Front Genet.* 17;10:1304. (2020). 10.3389/fgene.2019.01304. PMID: 32010183; : PMC6978288.
49. Decker, J. E. et al. Worldwide Patterns of Ancestry, Divergence, and Admixture in Domesticated cattle. *PLoS Genet.* **10** (3), e1004254. 10.1371/journal.pgen.1004254 (2014).
50. Flori, L. et al. Adaptive admixture in the West African bovine hybrid zone: insight from the Borgou population. *Mol Ecol*. ;23(13):3241-57. (2014). 10.1111/mec.12816. Epub PMID: 24888437 (2014).
51. Mwai, O., Hanotte, O., Kwon, Y. J. & Cho, S. African indigenous cattle: unique genetic resources in a rapidly changing world. *Asian-Australas J. Anim. Sci.* **28** (7), 911–921 (2015).
52. Gautier, M. & Naves, M. Footprints of selection in the ancestral admixture of a New World Creole cattle breed. *Mol. Ecol.* **20** (15), 3128–3143 (2011).
53. Hayes, B. J., Pryce, J., Chamberlain, A. J., Bowman, P. J. & Goddard, M. E. Genetic architecture of complex traits and accuracy of genomic prediction: coat colour, milk-fat percentage, and type in Holstein cattle as contrasting model traits. *PLoS Genet.* **5**(9), e1000396 (2009).
54. Bovine HapMap Consortium. Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds. *Science* **324** (5926), 528–532. https://doi.org/10.1126/science.1167936 (2009).
55. Garrick, D., Fernando, R., Garrick, D. J. & Fernando, R. L. Implementing a QTL detection study (GWAS) using genomic prediction methodology. Methods Mol Biol 1019: 275–298. *Meth. Mol. Biol.* (Clifton, N.J.). 1019. 275 – 98. (2013). 10.1007/978-1-62703-447-0_11
56. Boichard, D. et al. Design of a bovine low-density SNP array optimized for imputation. *PLoS ONE*. **7** (3), e34130. https://doi.org/10.1371/journal.pone.0034130 (2016).
57. Makina, S. Genome wide scan for signatures of selection in six cattle breeds in South Africa. *Genet. Selection Evol.* **47** 10.1186/s12711-015-0173-x (2015).
58. Kim, J. et al. The genome landscape of indigenous African cattle. *Genom Biol.* **18** (1), 34 (2017).
59. Hanotte, O. et al. Geographic distribution and frequency of a taurine Bos taurus and an indicine Bos indicus Y specific allele amongst sub-saharan African cattle breeds. *Mol Ecol.* 9(4):387 – 96. (2000). 10.1046/j.1365-294x.2000.00858.x. PMID: 10736042.
60. Ajmone-Marsan, P., Garcia, J. F. & Lenstra, J. A. On the origin of cattle: How aurochs became cattle and colonized the world. *Evol. Anthr : Issues News Reviews*. **19** (4), 148–157. https://doi.org/10.1002/evan.20267 (2010).
61. Hayes, B. J., Pryce, J., Chamberlain, A. J., Bowman, P. J. & Goddard, M. E. Genetic architecture of complex traits and accuracy of genomic prediction: coat colour, milk-fat percentage, and type in Holstein cattle as contrasting model traits. *PLoS Genet.* **5**(9), e1000396 (2009).
62. Kemper, K. E. et al. The distribution of SNP marker effects for faecal worm egg count in sheep, and the feasibility of using these markers to predict genetic merit for resistance to worm infections. *Genet. Res.* **96**, e002 (2014).

63. Li, R. et al. Genome-wide Scan of Selection Signatures in Dehong Humped Cattle for Heat Tolerance and Disease Resistance. *Anim. Genet*. 51, 292–299 (2020). (2020).
64. Utsunomiya, Y. T. et al. Genomic data as the hitchhiker's guide to cattle adaptation: Tracking the milestones of past selection in the bovine genome. *Front. Genet.* **11**, 554. https://doi.org/10.3389/fgene.2020.00554 (2015).
65. Jayawardana, J. M. D. R., Lopez-Villalobos, N., McNaughton, L. R. & Hickson, R. E. Genomic Regions Associated with Milk Composition and Fertility Traits in Spring-Calved Dairy Cows in New Zealand. *Genes* **14** (4), 860. https://doi.org/10.3390/genes14040860 (2023).
66. Boland, A., Côté, J-F. & Barford, D. Structural biology of DOCK-family guanine nucleotide exchange factors. *FEBS Lett.* **597** (6), 794–810. https://doi.org/10.1002/1873-3468.14523 (2018).
67. Ruiz-Lafuente, N., Minguela, A., Muro, M. & Parrado, A. The role of DOCK10 in the regulation of the transcriptome and aging. *Heliyon* **5** (2), e01224. https://doi.org/10.1016/j.heliyon.2019.e01224 (2018).

Figures

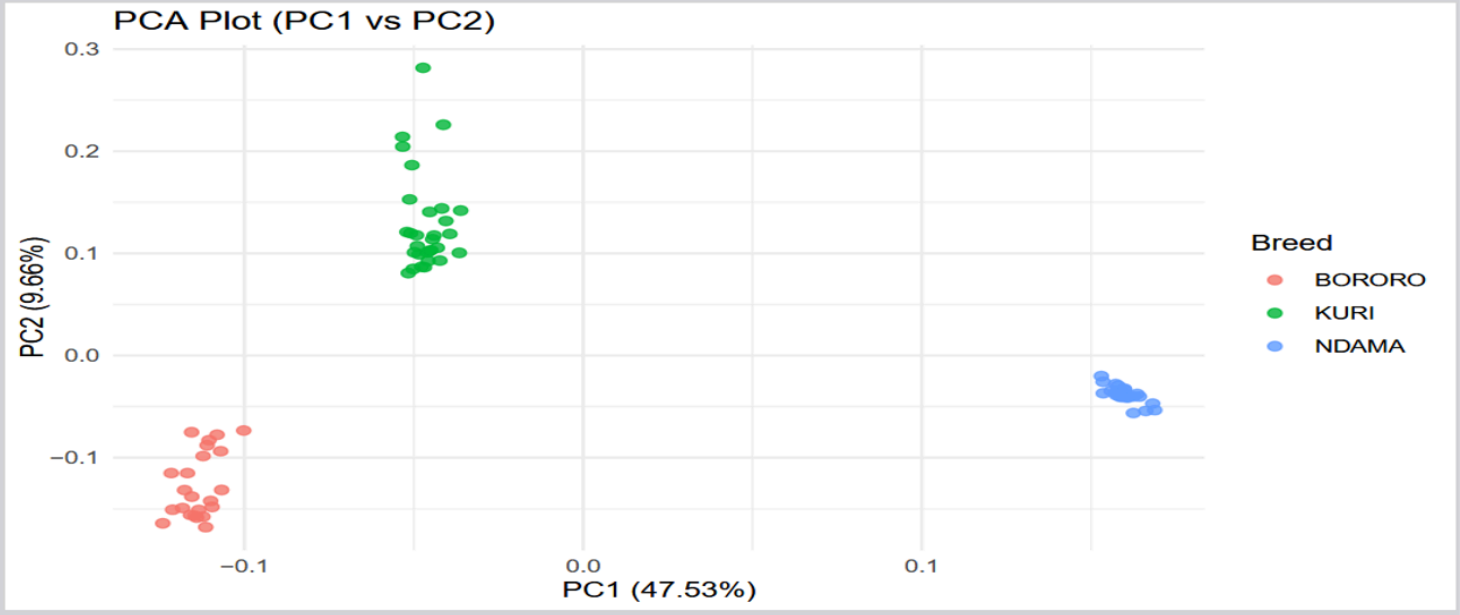


Figure 1

PCA plot showing genetic distances among the three breeds

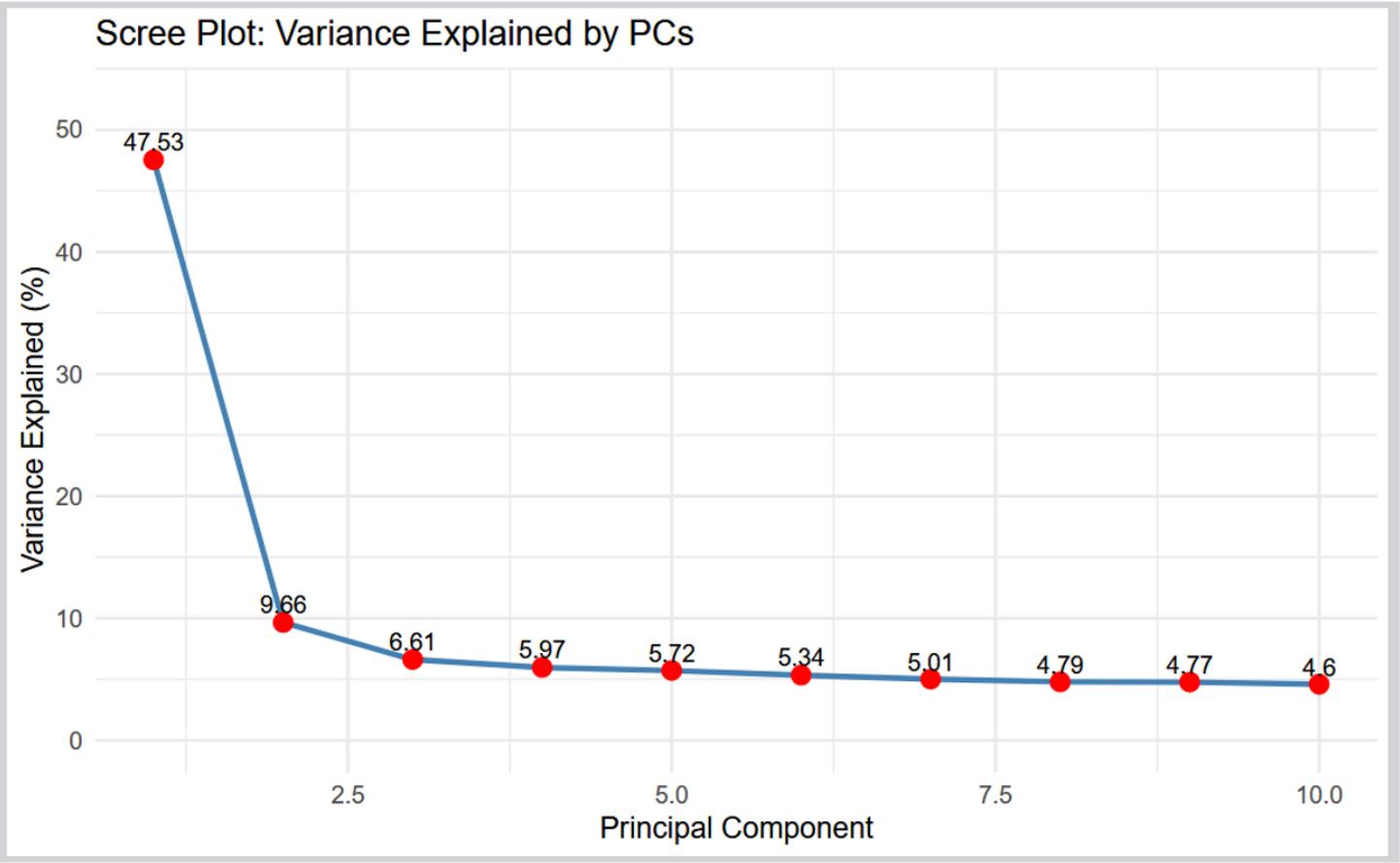


Figure 2

PCA scree plot revealing proportion of genetic variance explained among the breeds

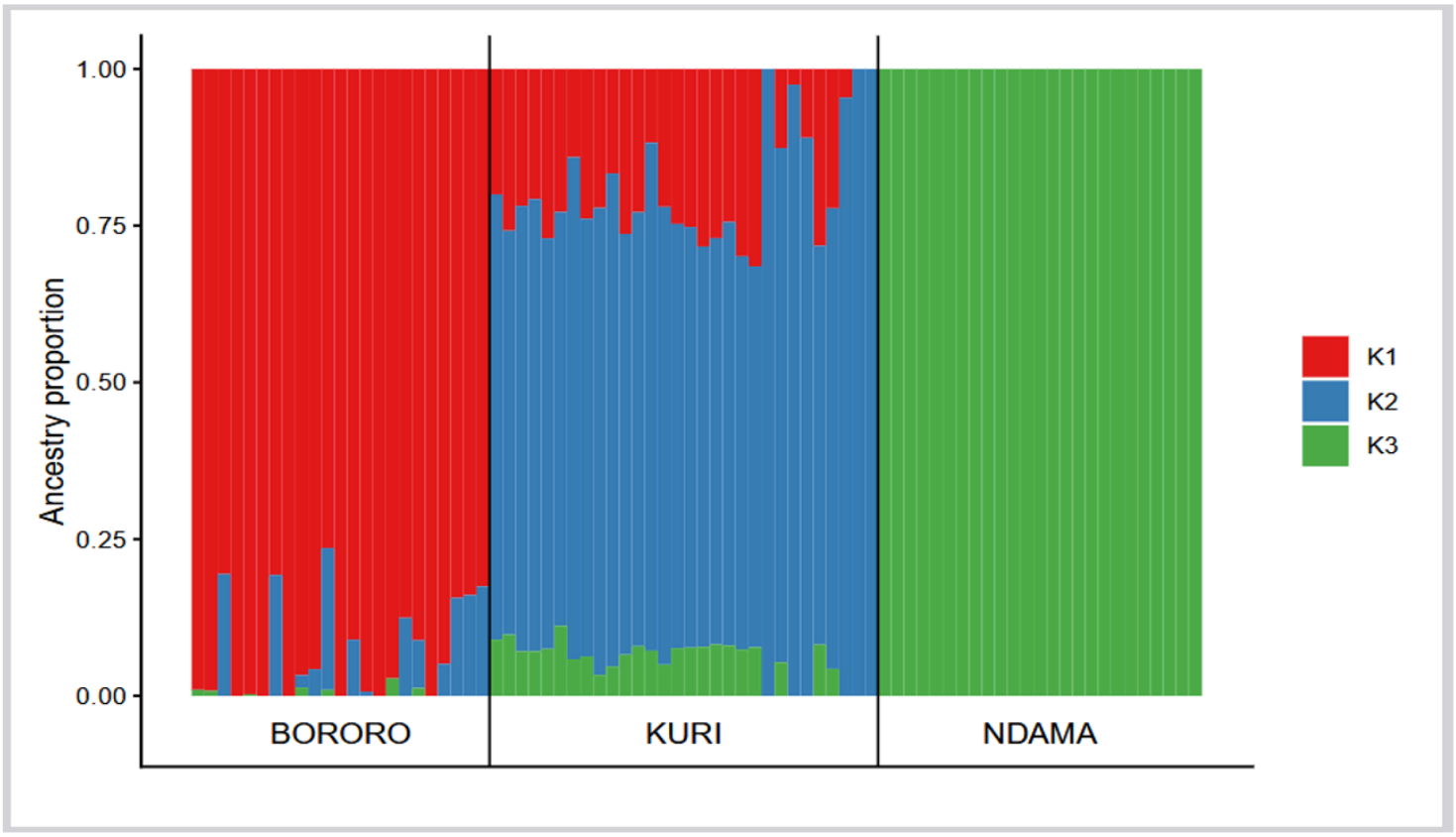


Figure 3

Admixture plots showing genetic clustering and differentiation among African Kuri, N'Dama and Bororo

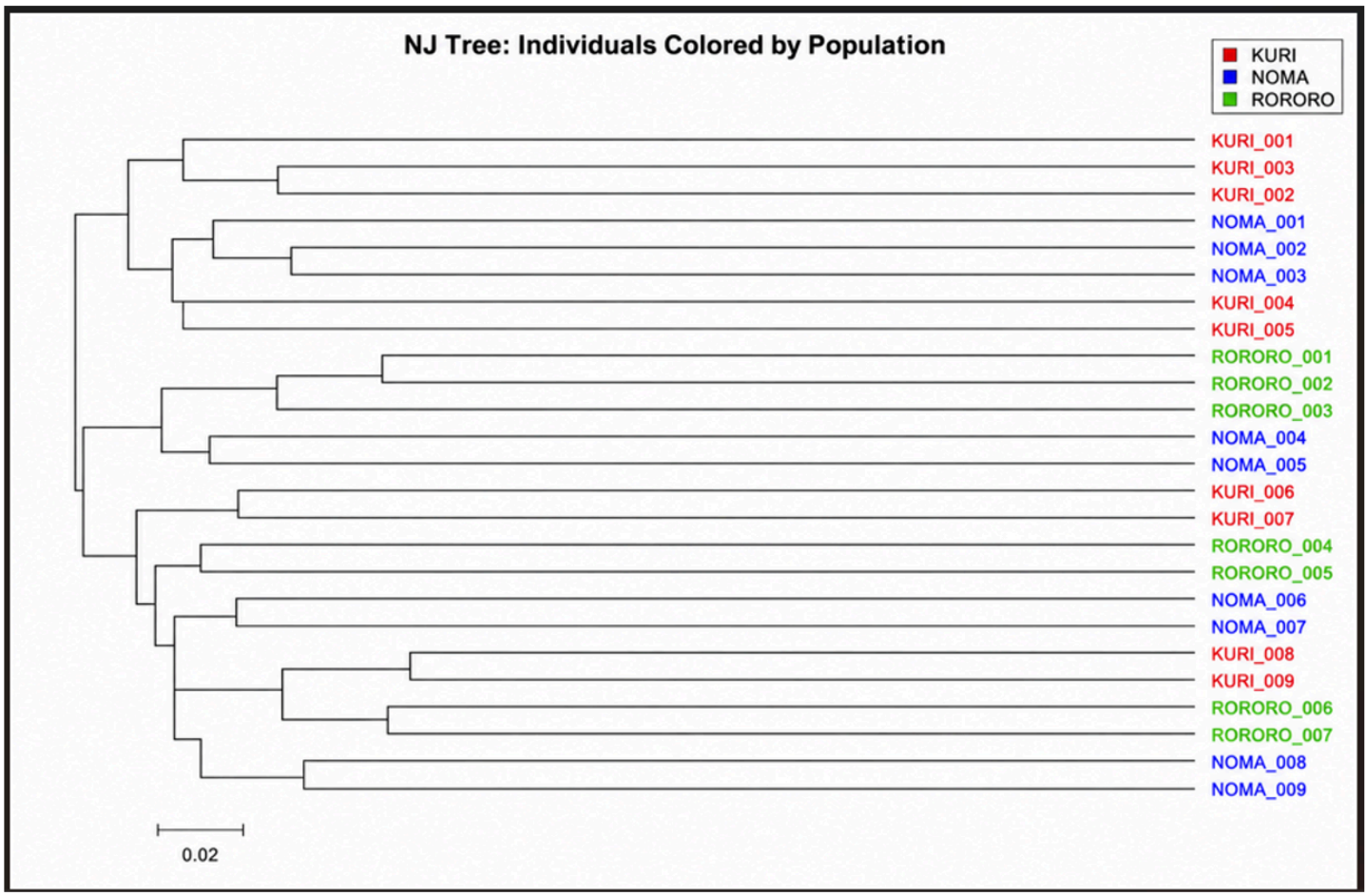


Figure 4

Neighbour Joining (NJ) Tree revealing genetic distance among Kuri, N'Dama and Zebu Bororo

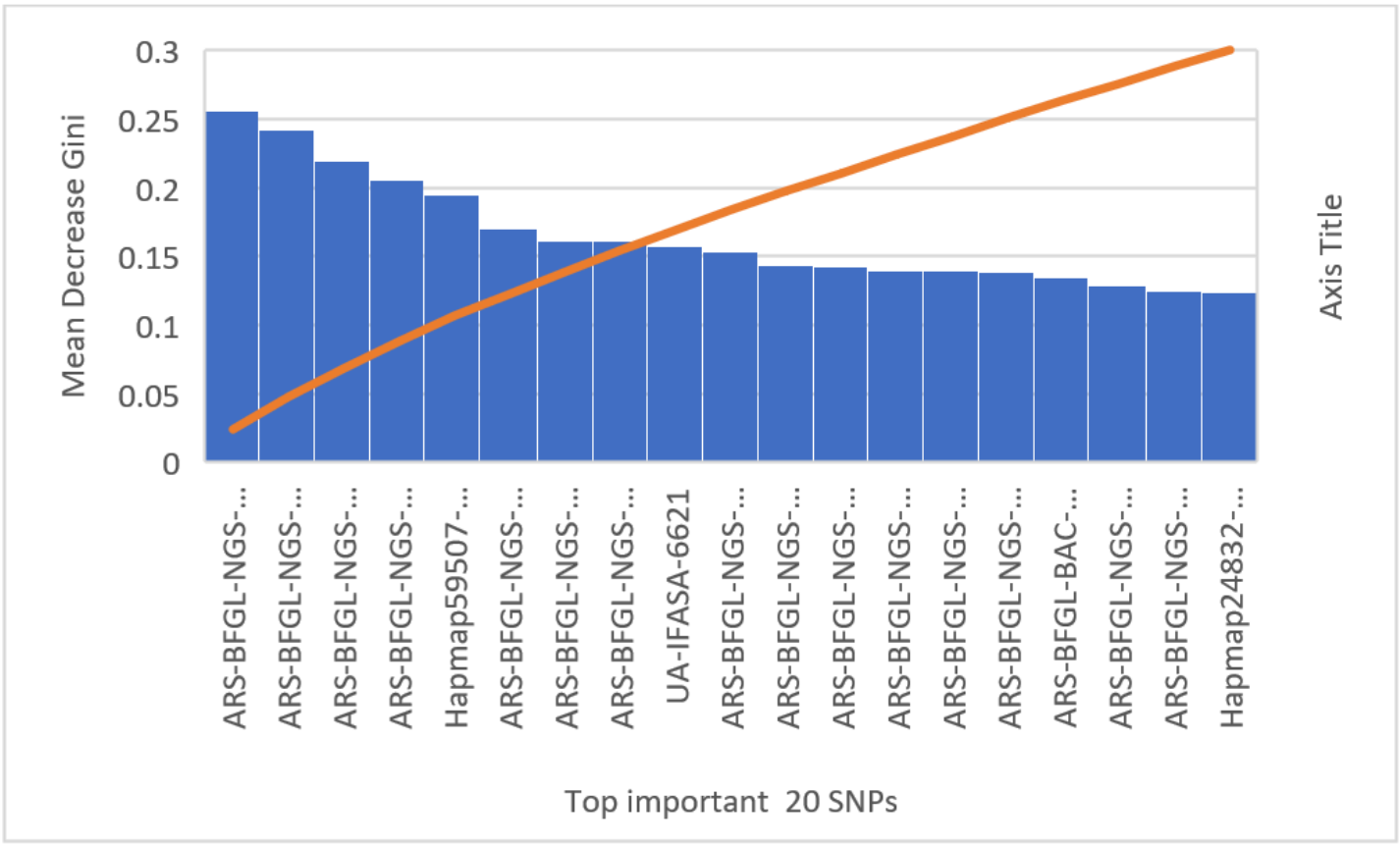


Figure 5

Random Forest of top important 20 SNPs

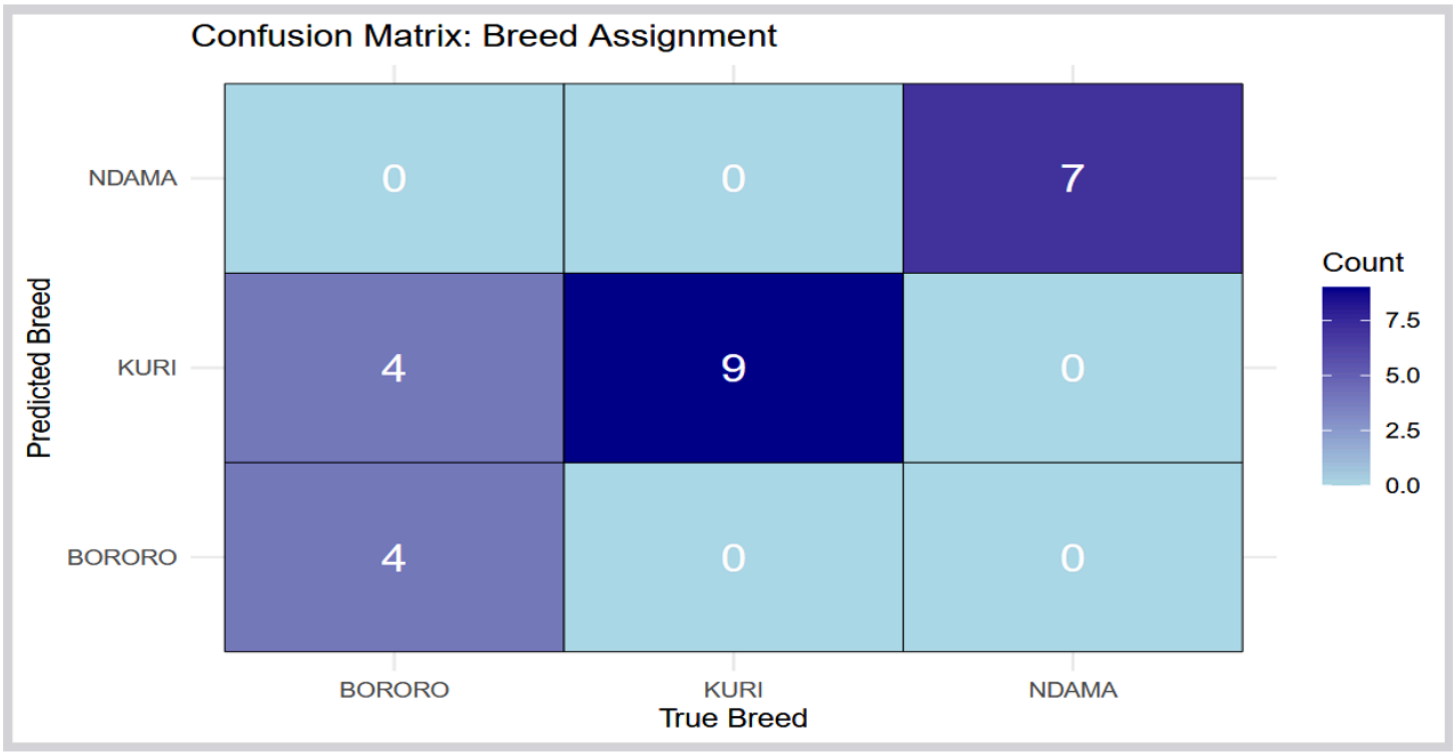


Figure 6

Confusion matrix

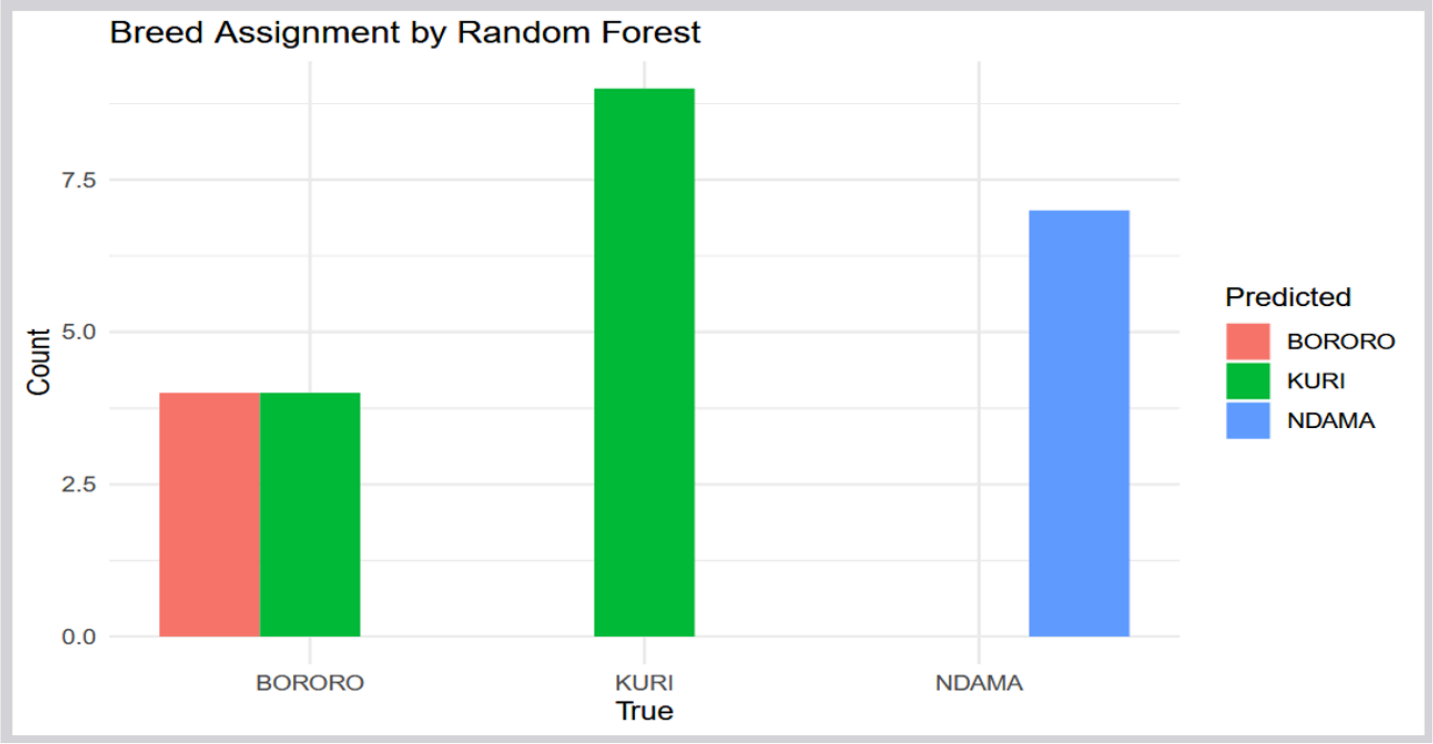


Figure 7

Breed assignment from random forest

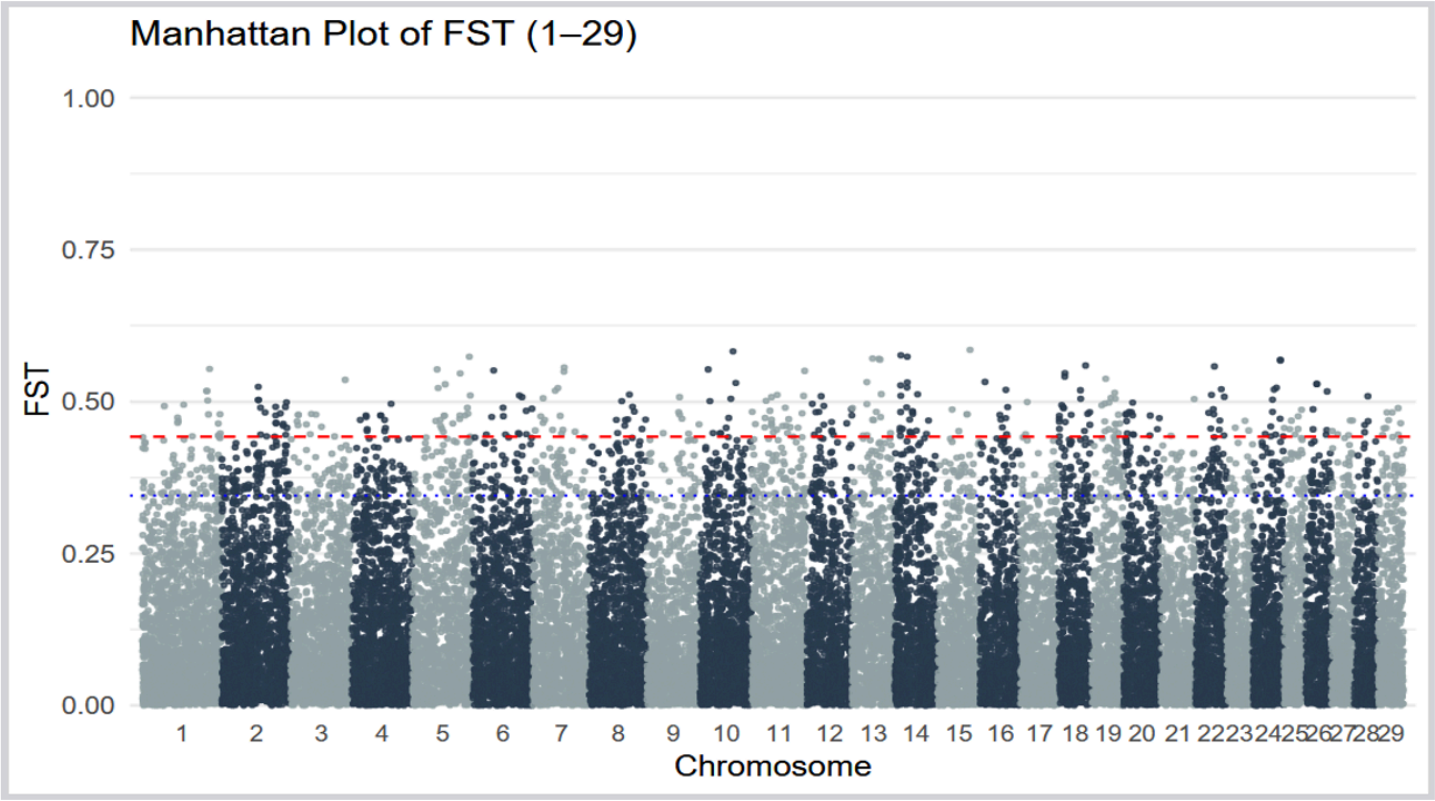


Figure 8

F_{ST} Manhattan plots for Kuri, N'Dama and Bororo cattle

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementarydata.xlsx