https://doi.org/10.1038/s42003-024-06481-2

Continent-wide genomic analysis of the African buffalo (*Syncerus caffer*)

Check for updates

Andrea Talenti[1,2,24], Toby Wilkinson[1,2,24], Elizabeth A. Cook[3,4], Johanneke D. Hemmink[1,2,3,4], Edith Paxton[1], Matthew Mutinda[5], Stephen D. Ngulu[6], Siddharth Jayaraman[1], Richard P. Bishop[3], Isaiah Obara[7], Thibaut Hourlier[8], Carlos Garcia Giron[8], Fergal J. Martin[8], Michel Labuschagne[9], Patrick Atimnedi[10], Anne Nanteza[11], Julius D. Keyyu[12], Furaha Mramba[13], Alexandre Caron[14,15,16], Daniel Cornelis[17,18], Philippe Chardonnet[19], Robert Fyumagwa[12], Tiziana Lembo[20], Harriet K. Auty[20], Johan Michaux[21], Nathalie Smitz[22], Philip Toye[3,4], Christelle Robert[1,2,23,25], James G. D. Prendergast[1,2,25] & Liam J. Morrison[1,2,25] ✉

The African buffalo (*Syncerus caffer*) is a wild bovid with a historical distribution across much of sub-Saharan Africa. Genomic analysis can provide insights into the evolutionary history of the species, and the key selective pressures shaping populations, including assessment of population level differentiation, population fragmentation, and population genetic structure. In this study we generated the highest quality de novo genome assembly (2.65 Gb, scaffold N50 69.17 Mb) of African buffalo to date, and sequenced a further 195 genomes from across the species distribution. Principal component and admixture analyses provided little support for the currently described four subspecies. Estimating Effective Migration Surfaces analysis suggested that geographical barriers have played a significant role in shaping gene flow and the population structure. Estimated effective population sizes indicated a substantial drop occurring in all populations 5-10,000 years ago, coinciding with the increase in human populations. Finally, signatures of selection were enriched for key genes associated with the immune response, suggesting infectious disease exert a substantial selective pressure upon the African buffalo. These findings have important implications for understanding bovid evolution, buffalo conservation and population management.

The African buffalo, *Syncerus caffer*, is a key member of the charismatic African megafauna, and was historically distributed across sub-Saharan Africa, inhabiting a diverse range of habitats from dry savannah to montane rainforest. Over the past century the population density and distribution has been much reduced. The population range has also become increasingly fragmented due to anthropogenic pressures, resulting in approximately 70% of the global population being restricted to protected areas[1–3].

The species has been historically divided into varying numbers of subspecies based upon distribution, habitat and morphology, the most recent update of the IUCN Red List recognising *S. caffer caffer* (Eastern and Southern African savannah), *S. c. brachyceros* (Western African savannah), *S. c. aequinoctialis* (Central African savannah), and *S. c. nanus* (Western and Central African forest)[4]. The genetic understanding of population diversity and structure across the species range mostly derives from the application of low resolution tools, such as mitochondrial D-loop sequences, microsatellites and mitogenomes[5–7], with a more recent study using genome-wide single-nucleotide polymorphisms (SNPs)[8]. The two studies to analyse diversity at the genome level, focused on South African *S. c. caffer* animals ($n = 40$) in protected areas[9] and *S. c. caffer* populations ($n = 59$) from East and Southern Africa[10]. These studies have collectively highlighted that the current subspecies classification may not be supported by genetic data, and that there is population substructuring within and between the putative subspecies. They have also indicated concerns with respect to low effective population sizes in increasingly isolated populations in some African regions. Improved genetic tools can potentially contribute to conservation management strategies, both in terms of restoring connectivity between relevant populations in order to improve or restore genetic diversity, and avoiding loss of genetic integrity (i.e. maintenance of genetic diversity relevant to local environmental adaptation) through uninformed population mixing (e.g. translocations)[6,11,12].

As well as being an iconic species of African wildlife, the African buffalo is the closest bovid relative of domesticated cattle (*Bos taurus taurus & Bos*

A full list of affiliations appears at the end of the paper. ✉e-mail: liam.morrison@roslin.ed.ac.uk

taurus indicus) in Africa. The African buffalo has co-evolved in Africa with pathogens responsible for important and impactful diseases of cattle such as African animal trypanosomosis[13] and foot and mouth disease virus (FMD)[14,15]. For trypanosomosis, in contrast to the often devastating impact that infection has on cattle, African buffalo are largely tolerant, displaying much less severe clinical signs (e.g. refs. 16,17). Additionally, African buffalo are the primary host for the tick-borne protozoan *Theileria parva*, the causative agent of East Coast fever, an often deadly disease in cattle that is asymptomatic in buffalo[18]. These diseases have impeded productivity and the expansion of African pastoralists and their cattle for centuries[19,20]. During the colonial era, European cattle also brought with them diseases then exotic to Africa, such as rinderpest, brucellosis and bovine tuberculosis[21], to which African buffalo are susceptible. African buffalo and cattle co-exist today across many wildlife/livestock interfaces that enhance mutual pathogen transmission[22], and this can result in imposition of strict veterinary controls at these interfaces that often impact local livelihoods and conservation efforts (e.g. refs. 23,24). This makes the buffalo particularly interesting in terms of host-pathogen coevolution and potentially providing a route to identifying host genes and pathways relevant to controlling these diseases in livestock.

This study aimed to develop a reference genome for the African buffalo, as a foundation to analyse the population genomic structure across the current distribution of the species in sub-Saharan Africa. Two reference genomes have previously been published, but were generated via short read sequencing, resulting in relatively fragmented final genome assemblies (scaffold N50s of 2.40 Mb and 2.32 Mb, respectively)[25,26]. Using a combination of long read (PacBio) and Hi-C sequencing, we generated and de novo assembled a substantially higher quality and more contiguous reference genome of 2.65 Gb, with a scaffold N50 of 69.17 Mb. We then sequenced the genomes of 196 African buffalo samples from across the current species distribution, which enabled the analysis of genetic substructure, admixture between populations, and effective population sizes. We also assessed *S. caffer* genomes for signatures of selection, highlighting genes that may be responsible for environmental adaptation, in particular against diseases important for both buffalo and cattle.

Results

Assembly statistics

We first generated a de novo *S. c. caffer* reference genome from a male buffalo (OPB4) sampled in Ol Pejeta Conservancy, Kenya, providing the foundation to enable the characterisation of the genetic diversity of African buffalo populations, both in terms of their geographic regions and habitats and their current subspecies classification. We applied a deep sequencing strategy, based on a combination of 60× long read (PacBio) and 75× short read (Illumina) reads, to generate a de novo reference genome ensuring high per base sequence quality and consensus to achieve good genome contiguity, with an N50 of 69.16 Mb. The long reads were assembled using FALCON (Dovetail Genomics) and polished using Arrow. Contigs were then scaffolded using ~393 million 2× 150 bp Illumina reads of HiC data, using the HiRise software. Gaps in the draft genome were addressed using PBJelly[27]. Finally, Pilon[28] was used for sequential rounds of polishing, each of which was assessed for its resulting assembly quality over previous rounds. The genome following four rounds of polishing displayed the highest assembly statistics, with a total of 3351 scaffolds, a total length of 2.65 Gb (comparable to 2.72 Gb for the *Bos taurus* genome), a scaffold N50 of 69.16 Mb and a quality value (QV) of 35.9, indicating ~1 error every 5000 bp. The assembly statistics are summarised in Fig. 1 and Supplementary Data 1.

Previous African buffalo reference genomes, generated by Glanzmann et al.[25]. and Chen et al.[26], were based solely on Illumina short read sequencing, which led to highly fragmented assemblies of 442,401 scaffolds with a scaffold N50 of 2.40 Mb, and 150,000 scaffolds with an N50 of 2.30 Mb, respectively. These very fragmented assemblies provided limited scope for downstream analysis of variants and their predicted effects on functional regions, i.e. annotated genes and regulatory regions (a comparison of the three genome assemblies is illustrated in Fig. 1).

Transcriptome analyses and genome annotation

To enable in depth characterisation of the African buffalo transcriptome and to facilitate the annotation of gene isoforms, we performed full length isoform sequencing (Iso-Seq) across samples from six different tissues (prescapular lymph node, testis, liver, kidney, lung and spleen) collected from the same animal for which the genome was assembled (OPB4). In total 51,521 distinct, high quality isoforms (defined as being supported by at least two full length reads and with >99% base composition accuracy) were detected across these samples (median of 11,520 per tissue, maximum of 27,271 in the testis). Complementing these data, we also generated Illumina RNA-seq data, from the same animal, from eight tissues (heart, prescapular and inguinal lymph nodes, testis, liver, kidney, lung and spleen). All transcriptomic data were deposited to ENA with accession numbers PRJEB36587 and PRJEB36588 for RNA-seq and Iso-Seq, respectively. Together these data have been used to provide a high quality annotation of the buffalo assembly which can be accessed through the Ensembl Rapid Release genome browser: https://rapid.ensembl.org/Syncerus_caffer_GCA_902825105.1.

African buffalo-specific sequence

After aligning the African buffalo genome to eight high quality assemblies of four different Bovidae species (cattle, water buffalo, yak and goat[29–34]), portions of the *S. caffer* genome that did not match any regions in the other assemblies were ascertained. This process identified a total of 24,336,918 intervals, for a total of 145,050,830 bp of sequence not identified in the other eight assemblies. This includes both small variations (e.g. SNPs, small indels), unplaced contigs without alignments to any other genome, and large portions of the genome lacking any alignment.

We then refined the region selection by filtering out shorter intervals (<60 bp) and regions defined as too close to a telomere (<10 Kb) or to a gap (<1 Kb), leaving a total of 113,654,400 bp in 81,357 fragments longer than 60 bp, which were neither telomeric nor neighbouring an assembly gap. These regions have an average length of 1397 bp (3772.4 bp SD) and a median size of 286 bp (min. 61 bp, max 308,890 bp). The majority of the regions (74,659 fragments accounting for 112,762,919 bp) represent sequence not found in any of the other species genomes considered in the study, whereas the remaining are classified as divergent haplotypes. Of the 113 Mb, a total of 64.9 Mb (57.1%) are putatively identified as repeats using RED[35]. To rule out the possibility of these novel regions being due to contamination, we confirmed the coverage of these regions was consistent with the rest of the genome, using short-read whole-genome sequencing data from 46 samples from the population analysis (see section below; Supplementary Fig. 1).

HOMER analysis was conducted to characterise the content of novel sequences, and considered 4286/7096 sequences with less than 60% of masked nucleotides. These sequences were enriched for 38 motif types (*P*-value < 1e−5), such as the FOSL2/MA0478.1/Jaspar (0.661) motif, originally described as a negative regulatory sequence in the differentiation-sensitive adipocyte gene (aP2); this motif has also been identified as potentially being important in viral gene regulation, as it has been found in a transcriptional enhancer for the Gibbon ape leukaemia virus[36]. We performed the feature analysis on the annotation generated by Ensembl from the Iso-Seq sequencing data previously described. We identified 7096 annotated genes and 131 pseudogenes overlapping the novel regions, of which 583 genes, 194 ncRNA genes and 71 pseudogenes were entirely included in the identified regions (Supplementary Data 2). A total of 317 of 583 genes had at least one biological term annotated. GO terms definitions were fetched using the goatools python package[37]. Out of 4088 terms in the background dataset, 17 (15 GO terms and 2 KEGG pathways) were found significantly enriched. Among the significant terms was the defence response GO term (GO:0006952, FDR-corrected *P*-value: 0.0189, Supplementary Data 2), described as the response triggered by the presence of a foreign body.

Population genetics

To better understand African buffalo genetic diversity, we generated short read sequencing data for a further 195 animals deriving from across the

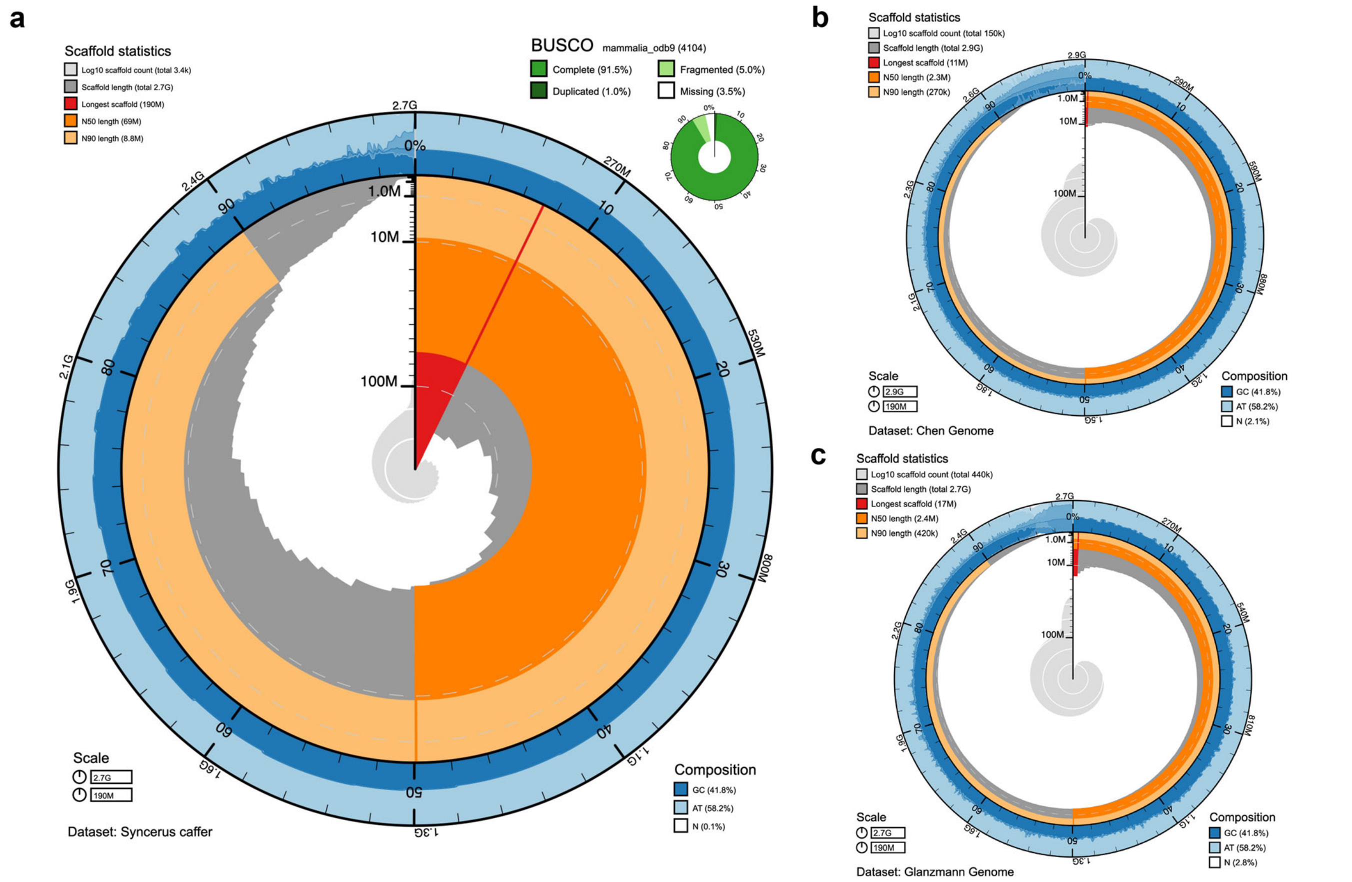

Fig. 1 | Genome assembly metrics. The BlobToolKit Snailplot shows N50 metrics and BUSCO gene completeness. **a** *Syncerus caffer* de novo assembly. The main plot represents the full genome length of 2.65 Gb. The distribution of scaffold length is shown in dark grey with the plot radius scaled to the longest chromosome present in the assembly (190 Mb, shown in red). Dark and light orange sections represent N50 and N90 (69 Mb and 8.8 Mb), respectively. The pale grey spiral shows the cumulative scaffold count on a log scale with white scale lines showing successive orders of magnitude. The blue/pale-blue/white ring graph shows the distribution of GC, AT and N percentages, respectively, for the given range in the main plot. A summary of complete, fragmented, duplicated and missing BUSCO genes in the mammalia_odb9 set is shown in the top right. **b** Chen et al. published genome. BlobToolKit Snailplot representing the *S. caffer* assembly as presented by Chen et al.[26]. **c** Glanzmann et al. published genome. BlobToolKit Snailplot representing the *S. caffer* assembly as presented by Glanzmann et al.[25]. The plot radius for both the Chen and Glanzmann genomes has been scaled to the maximum contig length (190 Mb) in the *S. caffer* genome assembled here to enable comparison of metrics.

continental range of the species (at a coverage of 15× for 146 samples, and 30× for 50 samples; Table 1 & Fig. 2a; for full sample list and metadata see Supplementary Data 3). This included samples from the currently described four subspecies; *S. c. caffer*, *S. c. nanus*, *S. c. brachyceros* and *S. c. aequinoctialis* (Table 1 & Fig. 2a), and putative *S. c. nanus* and *S. c. aequinoctialis* hybrids (based upon morphology and geography at time of sampling - labelled as 'intermediates'). Together, these samples derived from 21 sites/localities or protected areas across 12 different countries. We performed population analyses including only samples with a high call rate (>85%), and analysing only the biallelic polymorphic SNPs (minor allele frequency >5%), as well as only considering unrelated individuals (samples fourth degree or greater).

As can be seen in Fig. 2b the genetic relationships between the samples largely mirrors their geographic origin, with the first principal component (PC1) reflecting differentiation between samples from Eastern/Southern and Western Africa, which corresponds to a split between the Western/Central African subspecies (*S. c. aequinoctialis*, *S. c. brachyceros* and *S. c. nanus*) and the Eastern/Southern African subspecies *S. c. caffer*. The second component (PC2) correlates with differentiation between *S. c. caffer* samples from the Northern part of the subspecies' range (Kenya, Tanzania, Uganda) compared to *S. c. caffer* samples from Southern Africa. Notably, there was a clear signature of geography within the *S. c. caffer* data, with each geographic sub-population forming a distinct cluster in the PCA and a cline observed from Uganda to Kenya and Tanzania in the North, through Mozambique to samples from Botswana and Zimbabwe, and finally South Africa in the South. In Western/Central Africa, *S. c. aequinoctialis*, *S. c. brachyceros* and *S. c. nanus* sub-populations also formed separate clusters, although the *S. c. nanus* and *S. c. brachyceros* populations clustered closely together. Samples were initially grouped by sub-species and country of sampling. However, based on PCA results, the Tanzania and Kenya, and Botswana and Zimbabwe samples were grouped together, reflecting their geographic proximity and genetic similarity. This resulted in nine subgroups for downstream analyses; referred to hereafter as *S. c brachyceros*, *S. c. nanus*, *S. c. aequinoctialis*, intermediate (putative hybrids between *S. c. nanus*, *S. c. aequinoctialis*), *S. c. caffer* Uganda, *S. c. caffer* Kenya/Tanzania, *S. c. caffer* Mozambique, *S. c. caffer* Zimbabwe/Botswana and *S. c. caffer* South Africa. Population sample sizes post-filtering ranged from 2 for the *S. c. nanus* spp to 48 for the *S. c. caffer* from Tanzania (see Table 1), leaving a total of 163 samples for downstream analyses (see Supplementary Data 3 for samples included in these analyses).

In order to explore the relationship between these populations further, and to mitigate the different sample size between subpopulations resulting in over-representation of population-specific variation in the dataset as far as possible[38], we downsampled the larger groups to 15 representative samples (for those with less, all samples were included). Since the *S. c. brachyceros* population had a total of 16 samples, we did not perform any downsampling on this population. This resulted in a subset of 95 individuals to be considered for the population genetic analyses. As shown in the

Table 1 | Sample number by country, subspecies and pre- and post-data filtering

Sample origin	Unfiltered	Filtered missingness 0.20; Relatedness 0.0625
Botswana	17	15
Burkina Faso	9	7
Central African Republic	6	6
Chad	12	9
Gabon	7	2
Kenya	12	11
Mozambique	20	11
Niger	10	9
South Africa	8	6
Tanzania	50	48
Uganda	30	27
Zimbabwe	15	12
Total	196	163
By subspecies		
S. c. caffer	152	130
S. c. brachyceros	19	16
S. c. aequinoctialis (S.c.a)	12	9
Putative intermediate (*S.c.n/S.c.a*)	6	6
S. c. nanus (S. c. n)	7	2
Total	196	163

principal components analysis (PCA) pre- and post-downsampling (Supplementary Fig. 2), the general structure of the data was not affected by the subsampling.

Bootstrapped admixture and Evaladmix analyses (Fig. 3a, K = 2–15 tested; see Supplementary Fig. 3 for evaluation metrics and Supplementary Fig. 4 for results at multiple K) did not provide clear evidence for a particular number of defined subpopulations. However, there is an effect dependent upon geographical location of samples, and this is consistent with isolation by distance (IBD) being a main driver of genetic differentiation. For comparative purposes, the nine subgroups defined above (based on PCA and geographic proximity) were used for assessing within and between group metrics. Comparison of the genetic diversity between all pairs of populations (as represented by the F_{ST} statistic) highlights that this is largely a function of physical distance, i.e. the diversity observed between two populations increases broadly linearly with increasing distance between them (Fig. 3b, Mantel test r: 0.65, $p = 0.0018$; underlying F_{ST} data detailed in Supplementary Data 4). However, sub-structure in this isolation-by-distance analysis is observed. After excluding the *S. c. caffer* Hluhluwe-Umfolozi and *S. c. nanus* populations due to their high levels of homozygosity (see below), the relationship is even stronger, and variation in the F_{ST} values between the remaining groups can potentially largely all be explained by the distances between them (red line in Fig. 3b, Mantel test r: 0.96, $p = 0.0013$). This is consistent with the idea that these African buffalo have historically formed large continuous groups of populations with differentiation between populations simply reflecting the reduced mating probability with increasing distance. *S. c. nanus*, the forest buffalo, shows an unusually steep increase in differentiation relative to other populations (blue line in Fig. 3b). This could be for a variety of reasons, including geographical barriers reducing the gene flow between this group and the others analysed, as well as the small sample size available ($n = 2$). Animals found at the same location should exhibit little differentiation, and consistent with this, the intercept of the slopes is not significantly different from 0 in these comparisons (P >0.4 for both linear regression intercepts), i.e. when comparing *S. c. nanus* to other populations or the non- *S. c. nanus* and non-*S. c. caffer* Hluhluwe-Umfolozi populations to each other. However, this is not the case for comparisons involving the South African *S. c. caffer* Hluhluwe-Umfolozi population. Under the assumption of a simple linear relationship between genetic differentiation and geographic distance, the predicted level of diversity at a distance of 0 km is significantly higher than 0 (green line in Fig. 3b, linear regression intercept $P = 2.7 \times 10^{-4}$). This suggests that, unlike in the other population comparisons, there is elevated differentiation between this population and others, above and beyond that expected from their geographic distance apart. This is very likely to reflect the previously described bottleneck and isolation event with respect to the Hluhluwe-Umfolozi population (see Fig. 3c and below)[9,10,39].

EEMS analysis (Fig. 4a) adds to this picture of continental gene flow, with the Congo river basin likely representing a significant barrier of migration, particularly between Western/Central African *S. c. nanus* and *S. c. caffer* populations in Eastern Africa. The data also suggest that the Rift Valley potentially presents a geographical barrier to gene flow within the African buffalo.

The Relate software and genome-wide genealogies were used to estimate population-specific population sizes over time for the largest buffalo groupings (Uganda, Tanzania/Kenya and Zimbabwe/Botswana as well as all West African samples together– grouped as defined by PCA and admixture analyses; Figs. 2 and 3). As shown in Fig. 4 there has been a sharp reduction in the estimated effective population sizes across these groups in the last approximately 10,000 years, broadly mirroring the expansion of human effective population sizes over a similar time-period (Fig. 4b). There were not sufficient numbers in all individual populations for robust N_e analyses, but for the populations that did have sufficient numbers, contemporary N_e estimates were ~1300, 2000 and 3000 for Uganda, Tanzania/Kenya and Zimbabwe/Botswana, respectively. These data suggest that the effective population sizes of these Eastern and Southern African *S. c. caffer* are above the levels of conservation concern. Coalescence estimates are shown in Supplementary Fig. 5.

However, analysis of all populations highlights that the *S. c. nanus* and South African *S. c. caffer* Hluhluwe-Umfolozi samples have high levels of homozygosity (F_{ROH} of 0.29 and 0.36 compared to a range of 0.12–0.21 for the other populations; Fig. 3c). This is consistent with the known extreme bottlenecks experience by the Hluhluwe-Umfolozi buffalo population[9]; the *S. c. nanus* samples derive from Lekedi NP in Gabon, and we are unaware of historical population-level data that would inform of bottlenecks – while the homozygosity analysis is obviously on individual genomes, with this population we would caution overinterpretation as we only have data from two individuals.

Selective sweeps

African buffalo are exposed to a range of different environmental pressures across their distributional range, including a range of pathogens that also impact domesticated bovids such as cattle. To investigate selective sweeps between and within the nine population groupings we calculated the XP-EHH and P_R Relate Selection Test statistics[40,41]. Due to being more susceptible to artefactual results deriving from smaller sample sizes than the XP-EHH statistic, the calculation of the P_R statistic was restricted to just the populations with more than 20 samples after filtering for relatedness (i.e. the Uganda, Zimbabwe/Botswana and Tanzanian/Kenyan populations). These two tests are complementary in that whereas the XP-EHH statistic tests for differences in haplotype homozygosity between populations, P_R characterises the speed of spread of particular genomic lineages within a population, relative to others. Supplementary Data 5 summarises the results of these two tests. In total, 73 loci of elevated XP-EHH levels overlapping a gene were identified in at least one population comparison, and 34 P_R significant loci were detected in one of the three studied populations. Of the XP-EHH loci, 9 also overlapped a significant P_R peak (Supplementary Data 5). These 9 loci spanned 11 genes, with several having strong links to immune response, including putative killer cell immunoglobulin-like receptor like protein KIR3DP1 (LOC102402296), T cell receptor beta variable 5-1-like (LOC112577699), the major histocompatibility complex

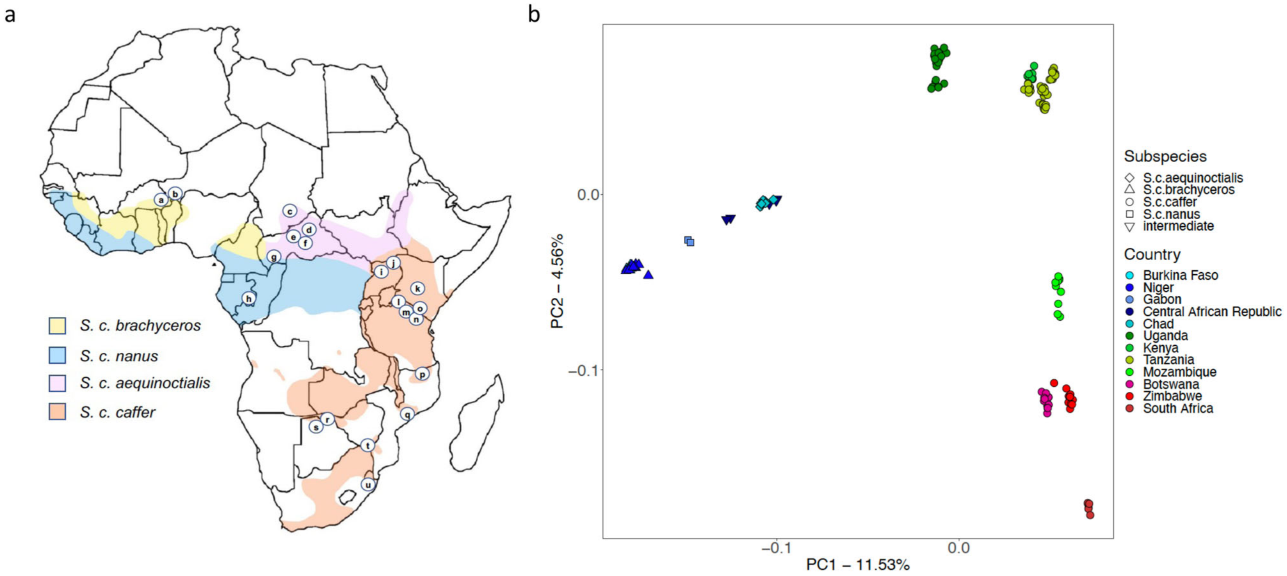

Fig. 2 | Sample source locations and Principal Component Analysis of population samples. a The sampling locations of African buffalo samples sequenced in the current study (circled letters), mapped on to the approximate current distribution of the four subspecies. a: Singou and Pama Game Reserves (GR)/Arli National Park (NP) complex, Burkina Faso ($n = 10$ samples [before data filtering]); b: W NP, Niger ($n = 10$); c: Zakouma NP, Chad ($n = 13$), d: Manovo-Gounda-St. Floris NP, Central African Republic (CAR; $n = 2$); e: Bamingui-Bangoran NP, CAR ($n = 2$); f: Sangba, CAR ($n = 1$); g: N'Gotto Forest Reserve, CAR ($n = 2$); h: Lekedi NP, Gabon ($n = 8$); i: Murchison Falls NP, Uganda ($n = 13$), j: Kidepo NP, Uganda ($n = 20$); k: Ol Pejeta Game Reserve, Kenya ($n = 12$); l: Serengeti NP, Tanzania ($n = 15$); m: Ngorongoro Conservation Area, Tanzania ($n = 15$); n: Tarangire NP, Tanzania ($n = 10$); o: Arusha NP, Tanzania ($n = 10$); p: Niassa National Reserve (NR), Mozambique ($n = 9$); q: Marromeu NR, Mozambique ($n = 9$); r: Chobe NP, Botswana ($n = 9$); s: Okavango Delta, Botswana ($n = 9$); t: Gonarezhou NP/Crook's Corner, Zimbabwe ($n = 18$), u: Hluhluwe-Umfolozi NP, South Africa ($n = 8$; for full sample data see Supplementary Data 3). **b** Principal Component Analysis of population samples, with data for components 1 and 2 illustrated. Samples are coloured by country of origin, with different symbols indicating the previously recognised subspecies.

gene TRIM26 and N-acetylneuraminic acid phosphatase (NANP). The latter is involved in sialic acid synthesis, which in turn is linked to immune response modulation, and NANP has also been observed to be under recent positive selection in both humans and cattle[42,43]. Two of these nine genes linked to both XP-EHH and P_R peaks in African buffalo were also previously linked to recent positive selection in water buffalo[43], namely myeloid-associated differentiation marker-like (LOC102403696) and tyrosine-protein phosphatase non-receptor type substrate 1-like (SIRPA-like) gene (LOC102396916). LOC102396916 was associated with significant P_R peaks in both the Uganda and Tanzania/Kenyan populations and also elevated XP-EHH scores in the South African *S. c. caffer* vs intermediate and *S. c aequinoctialis* populations (Fig. 5). SIRPA is an immunoglobulin-like cell surface receptor for CD47 (a cell surface protein that is involved in the promotion/regulation of cellular proliferation) and has been associated with a range of infectious diseases, including *Theileria annulata* infection in cattle[44] (*T. annulata* is the causative agent of tropical theileriosis across North Africa and Asia, and is closely related to *Theileria parva* found in Eastern Africa). This gene has previously been identified to be associated with selective sweeps between water buffalo breeds (elevated XP-CLR statistics between Mediterranean and Jaffrabadi, and Pandharpuri and Banni water buffalo breeds[43]). Characterisation of this gene's expression profile in the water buffalo expression atlas highlighted that it falls within a macrophage-specific cluster of genes[45]. Together these results therefore point towards this gene being a potentially important target of selection across bovids due to its role in immune response. Consequently, five of these nine genes under putative selection in African buffalo show strong links to immune response, with two of the remaining genes being uncharacterised and their function being unknown.

Discussion

African buffalo genome

The genome generated in this study represents a substantial improvement on current genomic resources available for *S. caffer*, with greater contiguity and much improved assembly and annotation – this, and the allied gene expression datasets, will hopefully serve as useful resources for the bovid and African buffalo research communities. The genome assembly is currently at the scaffold rather than chromosomal level, and so karyotype and features such as centromeres remain undefined, and the genome also contains Y chromosome and mitochondrial sequences that have not been completely resolved. There is therefore clearly scope for further improvement of the reference genome. An interesting finding was the African buffalo-specific sequence, which was identified after aligning the African buffalo genome to eight existing high quality bovid genome assemblies (cattle, water buffalo, yak and goat[29–34]). *S. caffer* sequences that that did not match any regions in the other assemblies were defined as African buffalo-specific sequence. These sequences were validated by assessing coverage of these African buffalo-specific sequences in randomly selected short read data from the population data, based on the expectation that if these were genuine African buffalo-specific sequence there would be coverage detected in multiple samples, and this was indeed the case. While 57.1% of these African buffalo-specific sequences are repeats, there are 583 genes, 71 pseudogenes, and 194 ncRNAs that are entirely within the identified regions. These were enriched for genes associated with the host defence, and the genes within these regions would clearly be of interest in further studies to identify traits that may be relevant to these African buffalo-specific sequences.

Population genomic structure: taxonomic insights

It should be noted that when dealing with real populations there is often not a simple clear answer as to the number of discrete groupings, as the samples may not represent a recent mixture of discrete ancestral populations[46]. This is broadly supported by the presented isolation-by-distance analysis. Therefore, in this study grouping of samples was largely guided by geographic sampling locations. Our analyses suggest that isolation by distance is a primary driver of the observed genetic differentiation between population samples. There is little support in our data for the current classification of the four IUCN recognised subspecies; *S. c. caffer* (Eastern and Southern African savannah), *S. c. brachyceros* (Western African savannah), *S. c. aequinoctialis* (Central African savannah) and *S. c. nanus* (Western and Central African

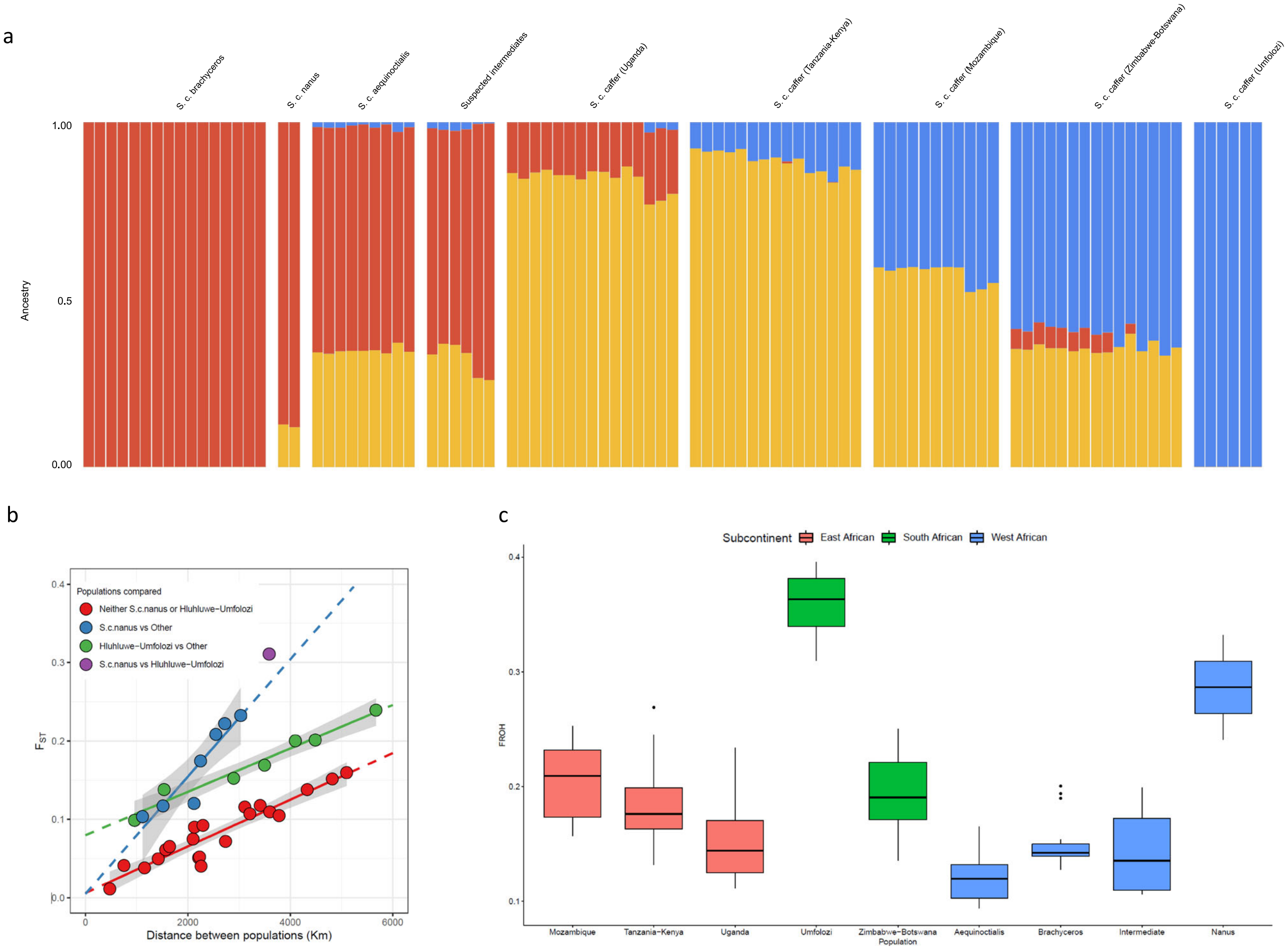

Fig. 3 | Population genetic analyses based on genome sequences. a Admixture analysis for K = 3 (Western/Central Africa in red, Eastern Africa in yellow and Southern Africa in blue). **b** Isolation by distance (IBD) analysis of African buffalo populations. The F_{ST} values were calculated between all pairs of populations and plotted against their geographic distance apart. Pairwise comparisons involving *S. c. nanus* are indicated in blue, pairwise comparisons involving the Hluhluwe-Umfolozi population are shown in green, the single pairwise comparison comparing *S. c. nanus* and Hluhluwe-Umfolozi in purple, and all remaining pairwise comparisons in red. The predicted pairwise F_{ST} values outside of the observed distances are indicated by dashed lines. **c** The proportion of homozygous segments per sample (FROH) indicating that the Hluhluwe-Umfolozi population has unusually high levels of homozygosity.

forest), with *S. c. brachyceros* and *S. c. aequinoctialis* sometimes being lumped and treated as a single subspecies, viz. *S. c. brachyceros*. Historically these classifications have been based on a combination of geographical distribution, habitat preferences and morphological features. *Syncerus c. nanus*, the forest buffalo, is the most divergent morphologically, being on average much smaller, predominantly rufous in colour as opposed to black, and with a different horn shape. From this perspective it is perhaps surprising that we could not detect substantial genetic divergence from the Western/Central African savannah buffalo. However, this finding agrees with previous genetic analyses using mitochondrial D-loop sequence markers, which similarly indicated a lack of support for differentiation between Western/Central African 'subspecies'[5]. However, the limited number of samples assigned to *S. c. nanus* did not enable balanced analyses, with in general a smaller number of populations sampled for the Western and Central African regions compared to Eastern and Southern Africa. This may have resulted in some bias in our population analyses. We did attempt to mitigate this bias to some extent by reducing populations to balanced numbers of samples per population where relevant and possible. The status of the samples termed 'intermediate', which were suggested to be putative hybrids between *S. c. nanus* and *S. c. aequinoctialis* at the time of sampling (based on morphology, area and habitat) is not completely clear from our analyses – from PCA and admixture analyses it is not clear that these samples are indeed intermediates, and the data suggest that these samples are closer to *S. c. aequinoctialis* than *S. c. nanus*. In summary, the present database provides genome-level and -wide resolution on variation (based upon 23,454,419 identified variants relative to the assembled reference genome); a much more robust basis for identifying genetic differentiation than previous methods used to identify genetic substructuring in this species. These insights have parallels with previous genome/multilocus genetic data studies on African ungulates with similar pan-Sub-Saharan distribution, the giraffe (*Giraffa camelopardis*) and zebra (*Equus quagga*), which indicated a lack of correlation of genetic data with morphology-based speciation, in those cases resulting in the identification of cryptic speciation[47,48].

Admixture and EEMS analyses indicate that the population genomic structure is shaped by geographical barriers, which limit where migration and therefore where cluster and population mixing can happen. This is evidenced by Ugandan buffalo demonstrating ancestry from both Eastern and Western African populations, and there being some signal of East African ancestry in Central African buffalo (*S. c. aequinoctialis*, *S. c. nanus* and intermediate; Fig. 3a and Supplementary Fig. 4). Both admixture and EEMS data indicate that Uganda is likely to act as an interface zone between these clusters, although further sampling in relevant populations (for example, known buffalo populations in Eastern CAR and DRC, South Sudan and Western Ethiopia) would help resolve the extent of gene flow. EEMS analyses suggests that any divergence between the East and West African populations was most likely driven by geography, with the Congo Basin and River effectively creating a barrier to North-South gene flow in the

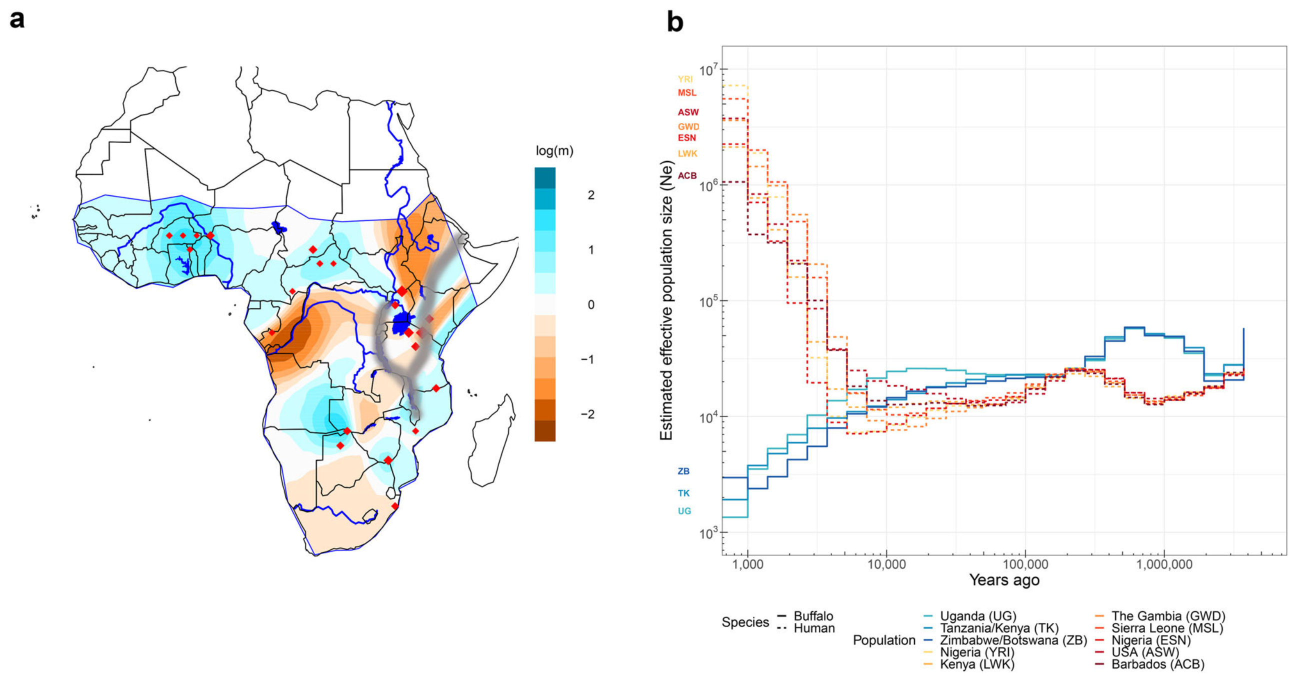

Fig. 4 | Continental gene flow and effective population sizes. a The contour map shows the mean of two independent Estimating Effective Migration Surfaces (EEMS) posterior migration rate estimates between 400 demes modelled over the land surface of Sub-Saharan Africa. A value of 1 (blue) indicates a tenfold greater migration rate over the average; −1 (orange) indicates tenfold lower migration than average. The courses of the major river systems (Niger, Congo, Nile and Orange rivers), as well as water bodies with a surface area greater than 5000 km^2 are included to highlight their potential relationships with migratory routes and barriers; grey shading indicates the Great Rift Valley. Red diamonds indicate geographical location of samples in the dataset. **b** Estimated effective population sizes of African buffalo (solid lines) and human (dashed lines) populations over time. The countries of sampling for each population are indicated in the legend along with the three letter 1000 Genomes consortium population code for the human data. Only human populations from the 1000 Genomes consortium dataset of recent African origin are shown.

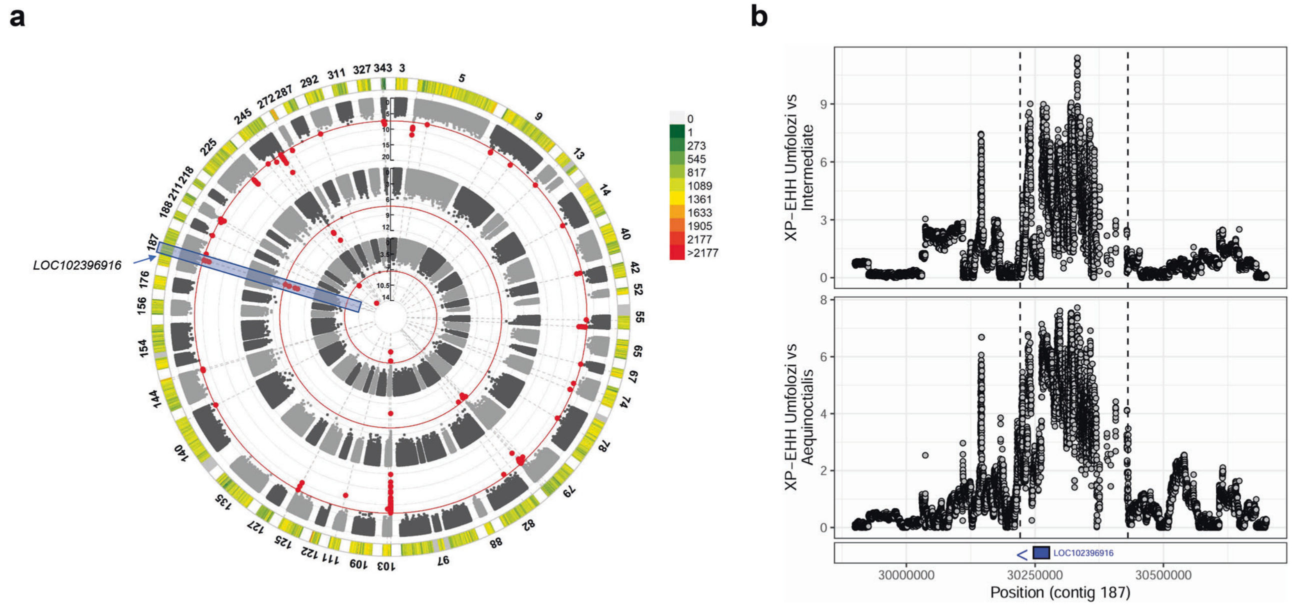

Fig. 5 | Selective sweep analysis. a The coloured outermost track and legend indicates the SNP density across 41 large contigs. The next three tracks show the P_R scores in the Uganda (centremost), Zimbabwe/Botswana (middle) and Tanzania/Kenya (outer) populations. Red points indicate SNPs with a *P*-value less than 5×10^{-8}. The peaks at LOC102396916 are highlighted. **b** Absolute XP-EHH scores across Contig 187 for the Hluhluwe-Umfolozi versus intermediate and Hluhluwe-Umfolozi versus *S. c. aequinoctialis* populations, indicating the peak detected at the LOC102396916 locus.

West of the continent, and Uganda being the pinch point at which Central African savannah and forest populations can intersect with Eastern African savannah buffalo.

The driving forces shaping the differentiation between Northern and Southern populations of *S. c. caffer* (i.e. between the Kenyan, Ugandan and Tanzanian cluster and the Mozambique, Botswana, Zimbabwe and South Africa cluster) is less clear from our analyses. A potential role of the Great Rift Valley acting as historical barrier to gene flow has been suggested within other large savannah mammals[49–51]. However, all Tanzanian samples included in the present study originated from the North of the country (the closest population in the Southern cluster being Niassa Special Reserve in Mozambique – approximately 1000 km from the Northern Tanzanian parks); additional samples from Central and Southern Tanzania where substantial buffalo populations exist (e.g. in Ruaha and Nyerere NPs) could potentially identify animals that are genetically intermediate, and reveal that there is a steady cline of differentiation within *S. c. caffer* from North to South, as supported by the isolation-by-distance analysis. The data are broadly consistent with the findings of a previous genomic study of *S. c. caffer* across its range, which also concluded that there was a primary split between northern and southern *S. c. caffer* populations approximately 50,000 years ago, followed by gene flow[10].

Effective population sizes

Although effective population size estimates are difficult to estimate accurately and can be confounded by population structure, the effective population size data interestingly suggests a coincident drop in N_e with the rise in human N_e (obtained through the 1000 Genomes data[52]). This is observed in similar analyses applied to both other individual African ungulates (giraffe)[53] and collated global ruminant data[26]. In the case of African buffalo, previous studies based on both microsatellite and mitochondrial DNA data have suggested an expansion approximately 80,000 years ago coincident with the spread of grassland habitat, which was followed by a significant decline ~3–7000 years ago, probably resulting from an overall increase in arid areas across Africa that are inhospitable to African buffalo[7,54,55] – our findings are consistent with the conclusion of a significant decrease in N_e ~10,000 years ago, although a historical population expansion ~80,000 years ago was not apparent from our data. For the African buffalo, it was anticipated that the greater resolution provided by genomic data may detect a drop in N_e observed as a result of the rinderpest virus epidemic of the 1890s[56], which anecdotally caused very high mortality of the buffalo populations through Eastern and Southern Africa in particular[57,58]. However, given the relatively recent timing of the rinderpest epidemic and the fact that the N_e was reducing across the relevant timeframe in our analysis, from the genome data we are not able to infer the impact of rinderpest upon population sizes. Other analyses using lower resolution genetic markers[55,59,60] were also not able to detect a drop in N_e that correlated with the timing of the rinderpest epidemic, although a recent genomic study using samples from *S. c caffer* did identify a very significant drop in N_e over the past 500 years, which could plausibly be explained by a combination of colonial activities and rinderpest[10] – notably the decline was particularly steep in samples from Hluhluwe-Umfolozi. While we did not have sufficient numbers in each population to robustly test N_e, for the closely related groupings of Uganda, Tanzania/Kenya and Zimbabwe/Botswana N_e estimates were approximately of 1300, 2000 and 3000 individuals, respectively. In these clusters at least, there is limited evidence for inbreeding depression, in agreement with previous studies[6]. However, the *S. c. nanus* and South African *S. c. caffer* Hluhluwe-Umfolozi samples showed high levels of homozygosity, meaning that further population-specific work is required in order to assess inbreeding risk. The *S. c. caffer* Hluhluwe-Umfolozi population is known to derive from very small number of founder animals, and our finding is in agreement with previous data that has indicated high inbreeding coefficients and low genome-wide heterozygosity levels in this population[9,10]. Although different generation intervals were used, impacting the precise estimates of timings, the coalescence estimates in this study are in broad agreement with a previous study indicating that the West and East African buffalo populations split tens of thousands of years ago[10].

While we have very limited numbers of *S. c. nanus* samples, the finding of high levels of homozygosity may perhaps be explained by the very different features of forest buffalo behaviour, in that relative to savannah buffalo forest buffalo have smaller home ranges, shorter daily movements, negligible seasonal movements and live in significantly smaller group sizes[2]. This is linked to the forest habitat likely generally acting as a greater barrier to gene flow than savannah environments, limiting migration/dispersal and resulting in comparatively small and isolated populations[5]. Genetic diversity metrics such as heterozygosity/homozygosity and effective population size will clearly be an important feature for future studies, particularly where there are increasingly fragmented and isolated populations, as is the case for the West African Savannah buffalo.

Selective sweeps

The selective sweep analyses identified tyrosine-protein phosphatase non-receptor type substrate 1-like (SIRPA-like) as being under selection, independently detected using two distinct and complementary methodologies (P_R and XP-EHH), and across several population groupings (Ugandan, Tanzanian/Kenyan, South African *S. c. caffer*, intermediate and *S. c aequinoctialis* populations). The same locus was identified in selective sweep analyses of the Asian buffalo *Bubalus bubalis*[43], and expression analysis in this species identified upregulated gene expression in a macrophage-specific cluster. Interestingly SIRPA has been associated with *Theileria annulata* infection in cattle[44], and its gene expression has been shown in independent studies to be significantly upregulated in host cells following infection and the cellular transformation associated with *T. annulata* infection[61,62]. While SIRPA will clearly be involved in the immune response to other pathogens, it is notable that *B. bubalis* is the primary host of *T. annulata* (the tick-borne causative agent of tropical theileriosis across North Africa and Asia). *Syncerus caffer* is similarly the primary host for the related parasite *Theileria parva* (and the related *Theileria* sp. buffalo[63]), and it is therefore plausible to link the described function of this gene with the long co-existence and co-evolution of *S. caffer* with *T. parva*. Although only the Ugandan, Tanzanian/Kenyan and South African *S. c. caffer* populations are within the current distribution of the tick vector (*Rhipicephalus appendiculatus*) of *T. parva*, the historical range and selection of *T. parva* cannot likely be inferred by the current vector distribution. Several other genes detected in the selective sweep analysis have been implicated in the host response to apicomplexan protozoa (which includes *Theileria* species), which lends credence to the hypothesis that the ancient co-evolution and selection pressure exerted by *T. parva* in *S. caffer* may have played a role in shaping the patterns of diversity in relevant regions of the current *S. caffer* genome. The long relationship between *T. parva* and *S. caffer* is reflected in the limited pathology caused by infection of *T. parva* in *S. caffer*, which is in stark contrast to the severe and often fatal disease caused by *T. parva* infection in other hosts such as domestic cattle[18,64]. The latter have only co-existed with *T. parva* for 5000–10000 years[65]. This finding may provide a route to identifying genes and pathways important in controlling disease during infections by *Theileria* species, that can, for example, be translated to mitigating the effect of these pathogens upon cattle or Asian buffalo owned by resource-poor farmers.

Conclusion

For the first time we have analysed genome-level data from all extant recognised African buffalo subspecies, covering the majority of the remaining geographical distribution of the species. Our findings demonstrate that the African buffalo population differentiation is largely driven by the isolation by distance effect of geographic location. While current subspecies nomenclature is likely to still have utility in terms of Management or Conservation Units, more samples and data, particularly from *S. c. nanus*, *S. c. brachyceros* and *S. c. aequinoctialis*, would help resolve the status of taxonomic units across the population range of African buffalo. The data also demonstrated that genetic connectivity between populations has historically been constrained by geographical barriers that have shaped the modern population structure (particularly the Congo basin), and that human influence has been for ~10,000 years and remains a main pressure on effective population size and population fragmentation. While most populations do not show signs of inbreeding, particular populations do, and this has implications for conservation and management of the species. Finally, through analyses of selective sweeps, we identified infectious diseases as a likely substantial contributor to historical selection, and hypothesise that protozoan pathogens for which the buffalo has been primary host for millennia may be responsible for driving some of this selection.

Materials and methods

Sample collection

DNA samples were obtained through (1) active sampling of animals for this project; this was done in collaboration with the Kenya Wildlife Service at the Ol Pejeta Conservancy, Kenya, or (2) secondary use of DNA samples previously collected; this included samples previously collected and published from Tanzania[14], Uganda[66], and Mozambique, Botswana, Zimbabwe, South Africa, Niger, Burkina Faso, Gabon Central African Republic and Chad[5,6,8]. For sample collection in Kenya, buffalo were darted and sedated by qualified veterinary personnel from KWS, and 10 ml blood collected into Paxgene Blood DNA tubes from peripheral venous sampling. DNA was extracted from the Paxgene Blood DNA tubes using the Paxgene Blood DNA kit (Qiagen) according to the manufacturer's instructions. Tissue pieces (OPB4) were snap frozen in liquid nitrogen in the field. Tissue pieces were homogenised using mortar and pestle over liquid nitrogen. The powder was resuspended in Trireagent (Sigma-Aldrich) and RNA was isolated using the RNeasy kit (Qiagen) according to the manufacturer's instructions. We have complied with all relevant ethical regulations for animal use.

Relevant research approvals were obtained in all instances; for the active sampling within this study, approval was obtained from the Kenya Wildlife Service (permit number KWS/BRM/5001). For secondary use of DNA samples previously collected, relevant permits are Tanzania Wildlife Research Institute and Tanzania Commission for Science and Technology (permit number 2021-262-NA-2021-066)[14] and Uganda Wildlife Authority (permit number COD/96/05)[66], or details are provided in refs. 5,6,8.

Genome sequencing

For the reference genome, a buffalo sample from Ol Pejeta in Kenya (OPB4) was sequenced using a combination of Illumina HiSeq (Dovetail Genomics & Edinburgh Genomics) and Pacific BioSciences approaches (Dovetail Genomics & Edinburgh Genomics) to a final sequencing coverage of 75× (Illumina) and 60× (PacBio). The same sample was also sequenced using Illumina Hi-C (Dovetail Genomics) in order to facilitate scaffolding. For the population samples, approximately 2.5 µg of total DNA from 196 animals sampled across Africa (Kenya, Uganda, Tanzania, Mozambique, Botswana, Zimbabwe, South Africa, Niger, Burkina Faso, Gabon, Central African Republic and Chad; Table 1, Supplementary Data 3) was subjected to whole-genome sequencing by Illumina HiSeq; this was performed at a coverage of 30× for 50 samples from Tanzania, with the remaining samples being sequenced at 15×.

Genome assembly

A primary assembly of the single molecule PacBio sequencing from OPB4 (mean read lengths >10 Kb) was generated using FALCON and consisted of 7269 contigs and an N50 of 1.9 Mb. This primary assembly was scaffolded using the Hi-C libraries and the HiRise software by Dovetail. The resulting scaffold-level assembly was further improved via gap filling and polishing steps performed with PBJelly[27] and Pilon[28], respectively, as described below. Gap filling: 7085 gaps (both inter- and intra-scaffolds) were identified in the scaffold-level assembly. A total of 78 inter-scaffold gaps were partially filled (i.e. extended on one side) using PBJelly, with 476,665 bases added in total, while none of the identified gaps were fully closed. This observation confirmed the high quality of the primary assembly achieved from PacBio reads including a post-processing step using Arrow (part of the GenomicConsensus package from PacBio). Polishing: An additional 75× Illumina short read sequencing (101 bp paired-end reads) of DNA from the same individual used to build the reference genome assembly (OPB4), was used to polish the de novo scaffold-level reference genome assembly. Polishing allows the correction of artefacts due to sequencing errors in assemblies, using the pile up of short reads that are associated with low sequencing error (~1%). This process was performed multiple times and improvement upon quality metrics (i.e. reduced numbers of ambiguous bases, corrected SNPs, resolved small indels, closed gaps) were assessed after each round of Pilon (see Supplementary Data 1a). The rate of improvements reached a plateau between the third (P3) and the fourth (P4) rounds of Pilon, and therefore the resulting P4 polished assembly was considered optimal and used for downstream analysis. Given the reference genome should not contain any homozygote alternate variant calls relative to the short read data from the same sample, we compared how the number of these changed following polishing. The Illumina short reads, sequenced from the same animal as that used to generate the reference genome assembly (OPB4), were mapped with bwa-mem (BWA v0.7.17) against the polished genome assemblies (P2–P4). The percentages of mapped reads were extremely high (>99%) and comparable across the P2, P3 and P4 assemblies.

Assembly statistics

To directly compare the quality of the genome assembly at each step during the assembly process, and to highlight improvements, QUAST (v 5.0.2)[67] was used to produce genome assembly metrics for each iteration of the genome assembly, pre and post gap filling with PBJelly, and for each successive round of polishing with Pilon (Supplementary Data 1b). QUAST further compares a given genome assembly to a reference genome, and for this the genome assembly for the water buffalo *Bubalus bubalis* (GCF_003121395.1)[29] was used, to produce genome alignment metrics and details of suspected misassemblies (Supplementary Data 1c). A custom Python script (https://raw.githubusercontent.com/evotools/CattleGraphGenomePaper/master/Assembly/ABS.py) was used to calculate scaffold metrics, N, L, NG, LG and GC content for a given proportion of the scaffold-level P4 genome assembly, in 5% increments (5–100, Supplementary Data 1d). The scaffold-level P4 genome assembly contains a total of 3351 scaffolds, of which 1381 scaffolds are greater than 10 kb. Quality values (QV) representative of the single-base accuracy were computed using Merqury (v1.1)[68]with the K-mer counts generated by Meryl (v1.2; https://github.com/marbl/meryl). For downstream analysis we selected 1381 contigs with a length of 10 kb or greater, representing 99.68% (2.653 Gb) of the total length of the assembled genome. This subset of contigs were used for downstream analyses.

Detection of novel genomic sequences

Following completion of the assembly, we identified the novel sequences in the genome in comparison with other ruminant species. We selected a set of nine genome assemblies for five species, and calculated the distances among them using mash v2.2[69], using a K-mer size of 32. We used the following genome assemblies to generate the alignment graph: *Syncerus caffer* (accession number GCA_902825105.1), *Bubalus bubalis* Mediterranean (GCF_003121395.1)[29], *Capra hircus* San Clemente (GCF_001704415.1)[34], *Bos grunniens* (GCA_005887515.2)[30], *Bos taurus indicus* Brahman (GCF_003369695.1), *Bos taurus taurus* Angus (GCA_003369685.2)[31], *Bos taurus taurus* Hereford (GCF_002263795.1)[33], *Bos taurus taurus* N'Dama (GCA_905123515) and *Bos taurus indicus* Ankole (GCA_905123885)[32]. We then generated a phylogenetic tree using the neighbour-joining algorithm included in the neighbour software from Phylip (v3.698)[70] which was used to create the following guide tree for CACTUS[71]:

((angus:0.00187,hereford:0.00115)Anc1:0.0004,(ankole:0.00317,((yak:0.00671,((abuffalo:0.01228,wbuffalo:0.0095)Anc6:0.00438,goat:0.04443) Anc5:0.01195)Anc4:0.00254,brahman:0.00256)Anc3:0.00023) Anc2:0.0004,ndama:0.00195)Anc0;

The HAL archive of multiple whole-genome alignments (mWGA) was generated using the software CACTUS[71], and then converted to PackedGraph format using the hal2vg software (v.2.1)[72] with the African buffalo genome as reference. We then used the nf-GraphSeq workflow (https://github.com/evotools/CattleGraphGenomePaper/tree/master/detectSequences/nf-GraphSeq) described in Talenti et al.[32]. based on libbdsg[73] to identify the nodes (i.e. the fragment of genome) that are found exclusively in the backbone of the graph (i.e. African buffalo genome), excluding all intervals overlapping a gap. We combined all interval regions less than 5 bp apart using BEDTools (v.2.30.0)[74]. We then annotated the regions by length (short if <10 bp, intermediate if <60 bp and large if >60 bp), position (labelled telomeric if <10Kb from the end of a scaffold larger than 5 Mb, flanking a gap if <1Kb from an N-mer), type of sequence (novel if >95% of the bases in the region are not found in any other genome, or haplotype if <95% of the bases were found only in the African Buffalo) and proportion of masked bases. We filtered out regions if they 1) were not classified as long, 2) contained less than 50% novel bases, and 3) were not telomeric or were not flanking a gap.

To validate that these regions corresponded to buffalo sequence, and did not derive, for example, from contamination, 46 of the population WGS samples were randomly selected and their coverage examined at these regions, with the assumption that if these regions corresponded to contamination in our reference sample, they would not have aligned reads from multiple buffalo samples. Mean read depth was calculated for each of the 74,659 novel regions within the reference genome, for the 46 population samples, using Mosdepth (v0.3.4)[75]. The distribution of average coverage values across the population samples, for each novel region, is shown in Supplementary Fig. 1. There are only 1494 novel regions with a mean read depth <1 and 419 regions with no reads mapped across these 46 samples, suggesting that these putative African buffalo-specific regions do not derive from an artefact such as contamination.

We characterised the content of the novel regions by 1) performing a motif analysis using HOMER (v4.11.1)[76], and 2) by detecting the novel features. To identify these features, we used the annotation generated by Ensembl and available in the rapid release database (http://www.ensembl.info/2020/06/25/ensembl-rapid-release/; accession GCA_902825105.1). We identified all gene features overlapping a novel sequence using bedtools intersect (v2.30.0)[74], and identified only these fully overlapping a novel region still with bedtools intersect with the -f 1.0 option (100% of overlap between the feature and the novel region).

Once we identified these fully new gene features, we extracted the GO term and KEGG pathways present in the annotation itself in embl format. To do so, we first converted the file in GenBank format, and then extracted for each gene the transcript IDs, protein IDs and biological terms. For these terms, we performed an enrichment analysis in R using a binomial test with the genes not in novel regions as background.

Reference genome annotation

Genome annotation was undertaken at EMBL-EBI by Ensembl, primarily using RNA-seq and full-length isoform sequencing (Iso-Seq) data generated from the animal for which the genome was assembled. A TruSeq stranded total RNA-seq library with one round of Ribo-Zero Gold kit (Illumina) was prepared from one pooled library consisting of RNA samples from eight tissues (heart, prescapular and inguinal lymph nodes, testis, liver, kidney, lung and spleen) collected from the animal for which the genome was assembled. RNA-seq was performed at Edinburgh Genomics on an S2 lane of an Illumina NovaSeq 6000 platform generating 100 bp paired-end reads. Iso-Seq was performed at the Centre for Genomic Research at the University of Liverpool, using RNA samples from six different tissues (prescapular lymph node, testis, liver, kidney, lung and spleen) collected from the same animal. Full-length cDNA from total RNA was generated using TeloPrime full-length cDNA amplification kit (v2) from Lexogen. A total of six barcoded TeloPrime libraries from six RNA samples were multiplexed. Iso-seq was performed on the resulting multiplexed library using six PacBio Sequel SMRT cells. The RNA-seq data were aligned to the reference genome using STAR[77]. For loci where the structures derived from the transcriptomic data appeared to be fragmented or absent, gap-filling using cross-species protein data was carried out. For more information on the annotation process see Supplementary Note 1.

Detection of variants in WGS samples across Africa

For all 196 WGS samples from *S. caffer* across Africa (raw data is available at ENA via accession numbers PRJEB59220 and ERP144275), reads were mapped with bwa-mem (BWA v0.7.17) against the reference genome generated as above. The GATK (v4.0.11.0) pipeline, following the best practices as outlined at https://gatk.broadinstitute.org/hc/en-us/articles/360036194592-Getting-started-with-GATK4, was used with HaplotypeCaller to identify variants (SNPs and Indels). The GATK best practice includes a Variant Quality Score Recalibration (VQSR) step that compares all variant calls to those in a high quality set to identify and flag potential false positives. Unlike in well-characterised species no gold-standard set of variants is available for the African buffalo. We therefore used a consensus set of 6,806,905 variants called from the Illumina data generated for the same sample as the reference genome using three software tools (GATK, Arrow and Longshot[78]). Although we do not expect this set to be free of false variant calls, we expect it to be enriched for true positives and this was therefore used in VQSR. Three VQSR tranches, 99, 99.9 and 100 (each representing the proportion of gold-standard variants that are retained at each quality threshold), were assessed. The variant set resulting from the 99.9 tranche was selected for downstream analyses with a Ti/Tv ratio of 2.07 and >120 M variants. The variant set was further filtered for GQ (Phred-scaled Probability that the call is incorrect) values less than 30 and site missingness of 0.9 (at least 90% of the samples contain data at this site). PLINK (v1.90) was used to calculate sample missingness, the proportion of variant sites missing from each sample, and vcftools (v0.1.13) to calculate the relatedness of all individuals. For downstream analyses, individual samples with a missingness greater than 0.15 were removed, and additionally individuals that were closer than fourth degree relatedness (relatedness value 0.0625), were also removed, resulting in a variant dataset covering 163 individual animals. We checked for any mapping biases due to use of an East African reference genome, by randomly sampling three animals per country and comparing how read mapping rates differed by longitude (Supplementary Fig. 6). No obvious mapping bias was observed among the West African samples when mapping to the reference genome obtained from an East African sample.

Genomic diversity analyses

The VCF file for the set of unrelated samples was first filtered through bcftools (v1.9; https://samtools.github.io/bcftools/) to keep only unrelated individuals according to the KING method implemented in vcftools[79,80]. A cutoff of 0.0625 was applied to exclude 3rd degree relatives or closer. Furthermore only biallelic SNPs in large contigs (>10 Kb) were retained. Variants were further filtered using plink (v1.90b4)[81] to restrict to those with a minor allele frequency >0.05. This dataset was then used to carry out analyses of migration events and effective population size. ADMIXTURE can benefit from having an even sample size for the different populations/samples deriving from the same location that were tested[38]. Therefore, for these analyses we identified a representative subsample for the populations with more than 15 animals. Downsampling (sample size reduction) was carried out using the BITE R package[82] to select a representative set of individuals for each population. BITE uses multi-dimensional scaling from identity-by-state distances to select a subset of individuals whose genetic structure reflect that of the total set. The downsampling process was performed on each population separately. For each group we selected the variants with very high call rate (99%) and highly polymorphic (--maf 0.3). The downsampling step in BITE was performed considering only individuals with 95% call rate and up to 10K markers to compute the kinship matrix (options n.trials = 100,000, ibs.marker = 10,000, n.k = 2, ibs.thr = 0.95, id.cr = 0.95). Principal component analysis (PCA) was performed post downsampling using plink v1.90b4. Admixture analysis was performed using ADMIXBoots (https://github.com/RenzoTale88/ADMIXBoots), a Nextflow workflow that performs bootstrapped admixture (v1.3.0)[83], defining a consensus of the different K at different iterations using CLUMPP[84] and generating plots in R. The workflow was run pruning for variants in linkage (plink --indep-pairwise 5000 100 0.3), testing every K between 2 and 15, and with 100 bootstraps of 100,000 markers each. A consensus of the different bootstraps was called using CLUMPP in LargeKGreedy mode. Bar charts for each consensus K, boxplots for the distribution of the CV errors and line plots of the H' scores of each K were generated from the pipeline automatically. EvalAdmix[46] was run within

ADMIXBoots on the admixture results computed on the pruned genotypes. The EvalAdmix plots were generated using the plotting scripts provided in the software website [https://www.popgen.dk/software/index.php/EvalAdmix]. For the isolation by distance analysis, pairwise F_{ST} values between populations were calculated using vcftools, and the Haversine formula was used to calculate the distances between the centre points of population sampling sites.

Estimated Effective Migration Surfaces (EEMS)

The EEMS package developed by Petkova et al.[85] was used (https://github.com/dipetkov/eems) to estimate effective migration surfaces. The runeems_snps program was used to visualise spatial population structure in the African buffalo populations and to identify the geographic barriers to migration preventing gene flow across these populations. The runeems_snps program requires the following data as input files: (1) a matrix of average pairwise genetic dissimilarities, (2) sample coordinates, and (3) a list of habitat coordinates, here covering the natural distribution of African buffalo populations on the African continent, and listed as a sequence of vertices organised as a closed polygon. For the input files for EEMS analysis, a matrix of average pairwise genetic dissimilarities was generated from the pruned set of SNP data, using the bed2diffs_v1 program within the EEMS package. The locations of all African buffalo animals, from which DNA samples were collected for WGS and variant detection, were inputted as longitude and latitude coordinates, indicating either specific sampling locations or the centre of specified geographical regions (e.g. national parks) when no other information was available. The list of habitat coordinates was generated based on the known past and present natural distribution of the four subspecies of African buffalo populations (as described in ref. 2) and using the https://www.latlong.net/ website to identify the latitude and longitude geocoding of point locations on the African continent. EEMS analysis was run using the runeems_snps program within the EEMS package based on the African buffalo pruned SNP data. Parameters used to run EEMS analysis were set as follows: nIndiv = 163; nSites = 6000; nDemes = 400; diploid = true; numMCMCIter = 4,000,000; numBurnIter = 1,000,000; numThinIter = 9999. Description for all parameters used are defined in the EEMS instruction manual (v.0.0.0.9000). Results of EEMS analysis were plotted using the rEEMSplot package in R to generate contour plots of effective migration and effective diversity surfaces from EEMS outputs. Additionally, posterior probability trace plots (pilogl) were used to check the MCMC sampler had successfully converged using four million MCMC iterations. The effective migration and diversity surfaces plots also include the addition of lakes and rivers depicted in blue based on data extracted from the Natural Earth website (https://www.naturalearthdata.com/download/50m/physical/).

Estimating effective population sizes and selective sweeps

To calculate the XP-EHH scores the African buffalo genotype data was first phased using Beagle 5.1[86]. A recombination rate of 1 cM/Mb was assumed and XP-EHH scores calculated between each pair of populations using hapbin[87]. Peaks were called as previously described[43]. Briefly, XP-EHH scores were smoothed by averaging across 1000 SNP windows and putative selective sweep regions were those with an absolute XP-EHH > 4, with the start and end coordinates defined where the XP-EHH scores fell back below two. The locations of XP-EHH peaks in the water buffalo and cattle genomes were obtained from Dutta et al.[43] and the peaks for all three species mapped to the orthologous regions of the water buffalo genome.

Within population and between population coalescence rates for the three largest African buffalo populations were calculated using Relate v1.1.6[41] using the same phased haplotypes from Beagle. Population-wise effective populations sizes being the inverse of these coalescence rate estimates. An estimated generation time of 11 years for the African buffalo was used in this analysis[88]. Previously calculated estimated effective population sizes for human African populations were obtained from Speidel et al.[41]. The P_R statistic was also calculated using Relate[41] and the same Beagle haplotype files using an estimated mutation rate of 1.25×10^{-8}. Variants with a P less than 5×10^{-8} were retained. The circular Manhattan plot was created using the CMplot R package[89]. The water buffalo genes were lifted over to the African buffalo genome to identify which genes fell under putative selective sweep peaks.

Data availability

The buffalo reference genome, encompassing the assembly, annotation, and supplementary flat files, is retrievable through GenBank with the accession ID GCA_902825105.1. The underlying raw datasets, originating from PacBio, Illumina, and Hi-C sequencing methods utilised in the assembly process, are archived under the European Nucleotide Archive (ENA) project accession ID PRJEB59220, corresponding to sample ERS14551691. Additionally, population-level whole-genome sequencing (WGS) data for the buffalo species is accessible via the ENA, catalogued under accession number PRJEB59220. Transcriptomic data, including RNA sequencing (RNA-seq) and Isoform sequencing (Iso-seq) datasets, are available through the ENA, under the accession IDs PRJEB36588 and PRJEB36587.

Received: 28 December 2022; Accepted: 21 June 2024;
Published online: 29 June 2024

References

1. East, R. *African Antelope Database 1999*. (Gland, Switzerland and Cambridge, UK, 1999).
2. Cornelis, D. et al. in *Ecology, evolution and behaviour of wild cattle: implications for conservation*. (eds M. Melletti & J. Burton) (Cambridge University Press, 2014).
3. Cornelis, D. et al. in *Ecology and Management of the African buffalo* (eds A. Caron, D. Cornelis, P. Chardonnet, & H. H. T. Prins) (Cambridge Univeristy Press, 2023).
4. Michaux, J., Smitz, N. & Van Hooft, P. in *Ecology and Management of the African buffalo* (eds A. Caron, D. Cornelis, P. Chardonnet, & H. H. T. Prins) (Cambridge University Press, 2023).
5. Smitz, N. et al. Pan-African genetic structure in the African buffalo (*Syncerus caffer*): investigating intraspecific divergence. *PLoS One* **8**, e56235 (2013).
6. Smitz, N. et al. Genetic structure of fragmented southern populations of African Cape buffalo (*Syncerus caffer caffer*). *BMC Evol. Biol.* **14**, 203 (2014).
7. Heller, R., Bruniche-Olsen, A. & Siegismund, H. R. Cape buffalo mitogenomics reveals a Holocene shift in the African human-megafauna dynamics. *Mol. Ecol.* **21**, 3947–3959 (2012).
8. Smitz, N. et al. Genome-wide single nucleotide polymorphism (SNP) identification and characterization in a non-model organism, the African buffalo (*Syncerus caffer*), using next generation sequencing. *Mamm. Biol.* **81**, 595–603 (2016).
9. de Jager, D. et al. High diversity, inbreeding and a dynamic Pleistocene demographic history revealed by African buffalo genomes. *Sci. Rep.* **11**, 4540 (2021).
10. Quinn, L. et al. Colonialism in South Africa leaves a lasting legacy of reduced genetic diversity in Cape buffalo. *Mol. Ecol.* **32**, 1860–1874 (2023).
11. Pizzutto, C. S., Colbachini, H. & Jorge-Neto, P. N. One Conservation: the integrated view of biodiversity conservation. *Anim. Reprod.* **18**, e20210024 (2021).
12. Hohenlohe, P. A., Funk, W. C. & Rajora, O. P. Population genomics for wildlife conservation and management. *Mol. Ecol.* **30**, 62–82 (2021).
13. Auty, H., Torr, S. J., Michoel, T., Jayaraman, S. & Morrison, L. J. Cattle trypanosomosis: the diversity of trypanosomes and implications for disease epidemiology and control. *Rev. Sci. Tech. Int. Off. Epizootics* **34**, 587–598 (2015).

14. Casey-Bryars, M. et al. Waves of endemic foot-and-mouth disease in eastern Africa suggest feasibility of proactive vaccination approaches. *Nat. Ecol. Evol.* **2**, 1449–1457 (2018).
15. Bengis, R. et al. in *Ecology and Management of the African buffalo* (eds A. Caron, D. Cornelis, P. Chardonnet, & H. H. T. Prins) (Cambridge University Press, 2023).
16. Dwinger, R. H., Grootenhuis, J. G., Murray, M., Moloo, S. K. & Gettinby, G. Susceptibility of buffaloes, cattle and goats to infection with different stocks of *Trypanosoma vivax* transmitted by *Glossina morsitans centralis*. *Res Vet. Sci.* **41**, 307–315 (1986).
17. Grootenhuis, J. G., Dwinger, R. H., Dolan, R. B., Moloo, S. K. & Murray, M. Susceptibility of African buffalo and Boran cattle to *Trypanosoma congolense* transmitted by *Glossina morsitans centralis*. *Vet. Parasitol.* **35**, 219–231 (1990).
18. Morrison, W. I., Hemmink, J. D. & Toye, P. G. Theileria parva: a parasite of African buffalo, which has adapted to infect and undergo transmission in cattle. *Int. J. Parasitol.* **50**, 403–412 (2020).
19. Gifford-Gonzalez, D. Animal disease challenges to the emergence of pastoralism in Sub-Saharan Africa. *Afr. Archaeological Rev.* **17**, 95–139 (2000).
20. Lankester, F. & Davis, A. Pastoralism and wildlife: historical and current perspectives in the East African rangelands of Kenya and Tanzania. *Rev. Sci. Tech. Int. Off. Epizootics* **35**, 473–484 (2016).
21. Michel, A. L. Implications of tuberculosis in African wildlife and livestock. *Ann. N. Y. Acad. Sci.* **969**, 251–255 (2002).
22. Kock, R., Kock, M., de Garine-Wichatitsky, M., Chardonnet, P. & Caron, A. in *Ecology, Evolution and Behaviour of Wild Cattle* (eds M. Melletti & J. Burton) Chapter 26, 431–425 (Cambridge University Press, 2014).
23. Caron, A. et al. Relationship between burden of infection in ungulate populations and wildlife/livestock interfaces. *Epidemiol. Infect.* **141**, 1522–1535 (2013).
24. Kock, R. et al. in *Ecology and Management of the African buffalo* (eds A. Caron, D. Cornelis, P. Chardonnet, & H. H. T. Prins) (Cambridge University Press, 2023).
25. Glanzmann, B. et al. The complete genome sequence of the African buffalo (*Syncerus caffer*). *BMC Genomics* **17**, 1001 (2016).
26. Chen, L. et al. Large-scale ruminant genome sequencing provides insights into their evolution and distinct traits. *Science* **364**, https://doi.org/10.1126/science.aav6202 (2019)
27. English, A. C. et al. Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. *PLoS One* **7**, e47768 (2012).
28. Walker, B. J. et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9**, e112963 (2014).
29. Low, W. Y. et al. Chromosome-level assembly of the water buffalo genome surpasses human and goat genomes in sequence contiguity. *Nat. Commun.* **10**, 260 (2019).
30. Zhang, S. et al. Structural Variants Selected during Yak Domestication Inferred from Long-Read Whole-Genome Sequencing. *Mol. Biol. Evol.* **38**, 3676–3680 (2021).
31. Koren, S. et al. De novo assembly of haplotype-resolved genomes with trio binning. *Nat. Biotechnol*. https://doi.org/10.1038/nbt.4277 (2018)
32. Talenti, A. et al. A cattle graph genome incorporating global breed diversity. *Nat. Commun.* **13**, 910 (2022).
33. Rosen, B. D. et al. De novo assembly of the cattle reference genome with single-molecule sequencing. *Gigascience* **9**, https://doi.org/10.1093/gigascience/giaa021 (2020)
34. Bickhart, D. M. et al. Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. *Nat. Genet* **49**, 643–650 (2017).
35. Girgis, H. Z. Red: an intelligent, rapid, accurate tool for detecting repeats de-novo on the genomic scale. *BMC Bioinforma.* **16**, 227 (2015).
36. Franza, B. R. Jr., Rauscher, F. J. 3rd, Josephs, S. F. & Curran, T. The Fos complex and Fos-related antigens recognize sequence elements that contain AP-1 binding sites. *Science* **239**, 1150–1153 (1988).
37. Klopfenstein, D. V. et al. GOATOOLS: A Python library for Gene Ontology analyses. *Sci. Rep.* **8**, 10872 (2018).
38. Meirmans, P. G. Subsampling reveals that unbalanced sampling affects STRUCTURE results in a multi-species dataset. *Heredity* **122**, 276–287 (2019).
39. O'Ryan, C. et al. Microsatellite analysis of genetic diversity in fragmented South African buffalo populations. *Anim. Conserv.* **1**, 85–94 (1998).
40. Sabeti, P. C. et al. Genome-wide detection and characterization of positive selection in human populations. *Nature* **449**, 913–918 (2007).
41. Speidel, L., Forest, M., Shi, S. & Myers, S. R. A method for genome-wide genealogy estimation for thousands of samples. *Nat. Genet* **51**, 1321–1329 (2019).
42. Moon, J. M., Aronoff, D. M., Capra, J. A., Abbot, P. & Rokas, A. Examination of Signatures of Recent Positive Selection on Genes Involved in Human Sialic Acid Biology. *G3 Bethesda* **8**, 1315–1325 (2018).
43. Dutta, P. et al. Whole genome analysis of water buffalo and global cattle breeds highlights convergent signatures of domestication. *Nat. Commun.* **11**, 4739 (2020).
44. Prajapati, B. M., Gupta, J. P., Pandey, D. P., Parmar, G. A. & Chaudhari, J. D. Molecular markers for resistance against infectious diseases of economic importance. *Vet. World* **10**, 112–120 (2017).
45. Young, R. et al. A Gene Expression Atlas of the Domestic Water Buffalo (*Bubalus bubalis*). *Front. Genet.* **10**, 668 (2019).
46. Garcia-Erill, G. & Albrechtsen, A. Evaluation of model fit of inferred admixture proportions. *Mol. Ecol. Resour.* **20**, 936–949 (2020).
47. Fennessy, J. et al. Multi-locus Analyses Reveal Four Giraffe Species Instead of One. *Curr. Biol.* **26**, 2543–2549 (2016).
48. Pedersen, C. T. et al. A southern African origin and cryptic structure in the highly mobile plains zebra. *Nat. Ecol. Evol.* **2**, 491–498 (2018).
49. Lohay, G. G., Weathers, T. C., Estes, A. B., McGrath, B. C. & Cavener, D. R. Genetic connectivity and population structure of African savanna elephants (*Loxodonta africana*) in Tanzania. *Ecol. Evol.* **10**, 11069–11089 (2020).
50. Bertola, L. D. et al. Phylogeographic Patterns in Africa and High Resolution Delineation of Genetic Clades in the Lion (*Panthera leo*). *Sci. Rep.* **6**, 30807 (2016).
51. Smitz, N. et al. A genome-wide data assessment of the African lion (*Panthera leo*) population genetic structure and diversity in Tanzania. *PLoS One* **13**, e0205395 (2018).
52. Genomes Project, C. et al. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
53. Coimbra, R. T. F., Winter, S., Mitchell, B., Fennessy, J. & Janke, A. Conservation Genomics of Two Threatened Subspecies of Northern Giraffe: The West African and the Kordofan Giraffe. *Genes* **13**, https://doi.org/10.3390/genes13020221 (2022)
54. Van Hooft, W. F., Groen, A. F. & Prins, H. H. Phylogeography of the African buffalo based on mitochondrial and Y-chromosomal loci: Pleistocene origin and population expansion of the Cape buffalo subspecies. *Mol. Ecol.* **11**, 267–279 (2002).
55. Heller, R., Lorenzen, E. D., Okello, J. B., Masembe, C. & Siegismund, H. R. Mid-Holocene decline in African buffalos inferred from Bayesian coalescent-based analyses of microsatellites and mitochondrial DNA. *Mol. Ecol.* **17**, 4845–4858 (2008).
56. Mack, R. The great African cattle plague epidemic of the 1890's. *Trop. Anim. Hlth. Prod.* **2**, 210–219 (1970).

57. Plowright, W. The effects of rinderpest and rinderpest control on wildlife in Africa. *Symposia Zool. Soc. Lond.* **50**, 1–28 (1982).
58. Estes, R. D. *The Behaviour Guide to African Mammals* (University of California Press, 1991).
59. Van Hooft, W. F., Groen, A. F. & Prins, H. H. Microsatellite analysis of genetic diversity in African buffalo (*Syncerus caffer*) populations throughout Africa. *Mol. Ecol.* **9**, 2017–2025 (2000).
60. Simonsen, B. T., Siegismund, H. R. & Arctander, P. Population structure of African buffalo inferred from mtDNA sequences and microsatellite loci: high variation but low differentiation. *Mol. Ecol.* **7**, 225–237 (1998).
61. Stephens, S. A. & Howard, C. J. Infection and transformation of dendritic cells from bovine afferent lymph by *Theileria annulata*. *Parasitology* **124**, 485–493 (2002).
62. Glass, E. J., Crutchley, S. & Jensen, K. Living with the enemy or uninvited guests: functional genomics approaches to investigating host resistance or tolerance traits to a protozoan parasite, *Theileria annulata*, in cattle. *Vet. Immunol. Immunopathol.* **148**, 178–189 (2012).
63. Bishop, R. P. et al. The African buffalo parasite *Theileria* sp. (buffalo) can infect and immortalize cattle leukocytes and encodes divergent orthologues of *Theileria parva* antigen genes. *Int. J. Parasitol. Parasites Wildl.* **4**, 333–342 (2015).
64. Wragg, D. et al. A locus conferring tolerance to *Theileria* infection in African cattle. *PLoS Genet.* **18**, e1010099 (2022).
65. Decker, J. E. et al. Worldwide patterns of ancestry, divergence, and admixture in domesticated cattle. *PLoS Genet* **10**, e1004254 (2014).
66. Obara, I. et al. The *Rhipicephalus appendiculatus* tick vector of *Theileria parva* is absent from cape buffalo (*Syncerus caffer*) populations and associated ecosystems in northern Uganda. *Parasitol. Res.* **119**, 2363–2367 (2020).
67. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072–1075 (2013).
68. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol.* **21**, 245 (2020).
69. Ondov, B. D. et al. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.* **17**, 132 (2016).
70. *PHYLIP (Phylogeny Inference Package) v. 3.7a* (Department of Genome Sciences, University of Washington, Seattle., 2009).
71. Armstrong, J. et al. Progressive Cactus is a multiple-genome aligner for the thousand-genome era. *Nature* **587**, 246–251 (2020).
72. Hickey, G., Paten, B., Earl, D., Zerbino, D. & Haussler, D. HAL: a hierarchical format for storing and analyzing multiple genome alignments. *Bioinformatics* **29**, 1341–1342 (2013).
73. Eizenga, J. M. et al. Efficient dynamic variation graphs. *Bioinformatics* **36**, 5139–5144 (2020).
74. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
75. Pedersen, B. S. & Quinlan, A. R. Mosdepth: quick coverage calculation for genomes and exomes. *Bioinformatics* **34**, 867–868 (2018).
76. Heinz, S. et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).
77. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
78. Edge, P. & Bansal, V. Longshot enables accurate variant calling in diploid genomes from single-molecule long read sequencing. *Nat. Commun.* **10**, 4660 (2019).
79. Danecek, P. et al. The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
80. Manichaikul, A. et al. Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010).
81. Chang, C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
82. Milanesi, M. et al. BITE: an R package for biodiversity analyses. *BioRxiv*. https://doi.org/10.1101/181610 (2017)
83. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
84. Jakobsson, M. & Rosenberg, N. A. CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics* **23**, 1801–1806 (2007).
85. Petkova, D., Novembre, J. & Stephens, M. Visualizing spatial population structure with estimated effective migration surfaces. *Nat. Genet.* **48**, 94–100 (2016).
86. Browning, S. R. & Browning, B. L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.* **81**, 1084–1097 (2007).
87. Maclean, C. A., Chue Hong, N. P. & Prendergast, J. G. hapbin: An Efficient Program for Performing Haplotype-Based Scans for Positive Selection in Large Genomic Datasets. *Mol. Biol. Evol.* **32**, 3027–3029 (2015).
88. Pacifici, M. et al. Database on generation length of mammals. *Nat. Conserv.* **5**, 89–94 (2013).
89. Yin, L. et al. rMVP: A Memory-efficient, Visualization-enhanced, and Parallel-accelerated tool for Genome-Wide Association Study. *Genomics, Proteomics & Bioinformatics* **4**, 619–628 (2021).

Acknowledgements

This research was funded in part by the Bill & Melinda Gates Foundation (USA) and with UK aid from the UK Foreign, Commonwealth and Development Office (Grant Agreement OPP1127286) under the auspices of the Centre for Tropical Livestock Genetics and Health (CTLGH), established jointly by the University of Edinburgh (UK), Scotland's Rural College (SRUC, UK), and the International Livestock Research Institute (UK). The findings and conclusions contained within are those of the authors and do not necessarily reflect positions or policies of the Bill & Melinda Gates Foundation nor the UK Government. The work was also supported by grants BBS/OS/GC/000012C, BBS/E/D/10002070, BBS/E/D/20002172 and BBS/E/D/20002174 from the Biotechnology and Biological Sciences Research Council (BBSRC, UK), and grant OB 490/2–1 from Deutsche Forschungsgemeinschaft (DFG, Germany). This research was conducted as part of the Consultative Group on International Agricultural Research (CGIAR) Program on Livestock (Kenya). International Livestock Research Institute (ILRI), Kenya, is supported by contributors to the CGIAR Trust Fund. CGIAR is a global research partnership for a food-secure future. Its science is carried out by 15 Research Centres in close collaboration with hundreds of partners across the globe (www.cgiar.org). Work carried out at EMBL, EBI was supported by the Wellcome Trust (WT222155/Z/20/Z) and BBSRC (BB/S020152/1).

Author contributions

L.J.M., J.G.D.P., and C.R. conceived the idea. E.A.C., J.D.H., M.M., S.D.N., E.P. and P.T. generated novel samples and data for the study. R.P.B., I.O., M.L., P.A., A.N., J.D.K., F.M., R.F., A.C., D.C., P.C., T.L., H.K.A., J.M. and N.S. contributed samples for the generation of genomic data. A.T., T.W., S.J., J.G.D.P., T.H., C.G.C., F.J.M., C.R. and L.J.M. led the analysis. A.T., T.W., J.G.D.P., C.R. and L.J.M. led the writing of the manuscript, and all authors were involved in the drafting of 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/s42003-024-06481-2.

Correspondence and requests for materials should be addressed to Liam J. Morrison.

Peer review information *Communications Biology* thanks Hubert Pausch, Deon de Jager and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editors: George Inglis and Luke R. Grinham.

Reprints and permissions information is available at http://www.nature.com/reprints

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

[1]The Roslin Institute, Royal (Dick) School of Veterinary Studies, University of Edinburgh, Midlothian EH25 9RG, United Kingdom. [2]Centre for Tropical Livestock Genetics and Health (CTLGH), Roslin Institute, University of Edinburgh, Easter Bush Campus, Roslin EH25 9RG, United Kingdom. [3]International Livestock Research Institute, P.O. Box 30709 Nairobi 00100, Kenya. [4]Centre for Tropical Livestock Genetics and Health (CTLGH), ILRI Kenya, P.O. Box 30709 Nairobi 00100, Kenya. [5]Kenya Wildlife Service, P.O. Box 40241 Nairobi 00100, Kenya. [6]Ol Pejeta Conservancy, Private Bag, Nanyuki 10400, Kenya. [7]Institute for Parasitology and Tropical Veterinary Medicine, Freie Universität Berlin, Robert-von-Ostertag-Str. 7-13, 14163 Berlin, Germany. [8]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridgeshire CB10 1SD, United Kingdom. [9]Clinomics, Uitzich Road, Bainsvlei, Bloemfontein 9338, South Africa. [10]Uganda Wildlife Authority, Kampala, Uganda. [11]College of Veterinary Medicine, Animal Resources and Biosecurity, Makerere University, Kampala, Uganda. [12]Tanzania Wildlife Research Institute, Box 661 Arusha, Tanzania. [13]Vector and Vector-Borne Diseases Institute, Tanga, Tanzania. [14]ASTRE, University of Montpellier (UMR), CIRAD, 34090 Montpellier, France. [15]CIRAD, UMR ASTRE, RP-PCP, Maputo 01009, Mozambique. [16]Faculdade Veterinaria, Universidade Eduardo Mondlan, Maputo, Mozambique. [17]CIRAD, Forêts et Sociétés, 34398 Montpellier, France. [18]Forêts et Sociétés, University of Montpellier, CIRAD, 34090 Montpellier, France. [19]IUCN SSC Antelope Specialist Group co-chair, 92100 Boulogne, France. [20]School of Biodiversity, One Health and Veterinary Medicine, College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow, United Kingdom. [21]Laboratoire de Génétique de la Conservation, Institut de Botanique (Bat. 22), Université de Liège (Sart Tilman), Chemin de la Vallée 4, B4000 Liège, Belgium. [22]Royal Museum for Central Africa (BopCo), Leuvensesteenweg 13, 3080 Tervuren, Belgium. [23]Centre for Genomic and Experimental Medicine, Institute of Genetics and Cancer, University of Edinburgh, Crewe Road South, Edinburgh EH4 2XU, United Kingdom. [24]These authors contributed equally: Andrea Talenti, Toby Wilkinson. [25]These authors jointly supervised this work: Christelle Robert, James G.D. Prendergast, Liam J. Morrison. ✉e-mail: liam.morrison@roslin.ed.ac.uk