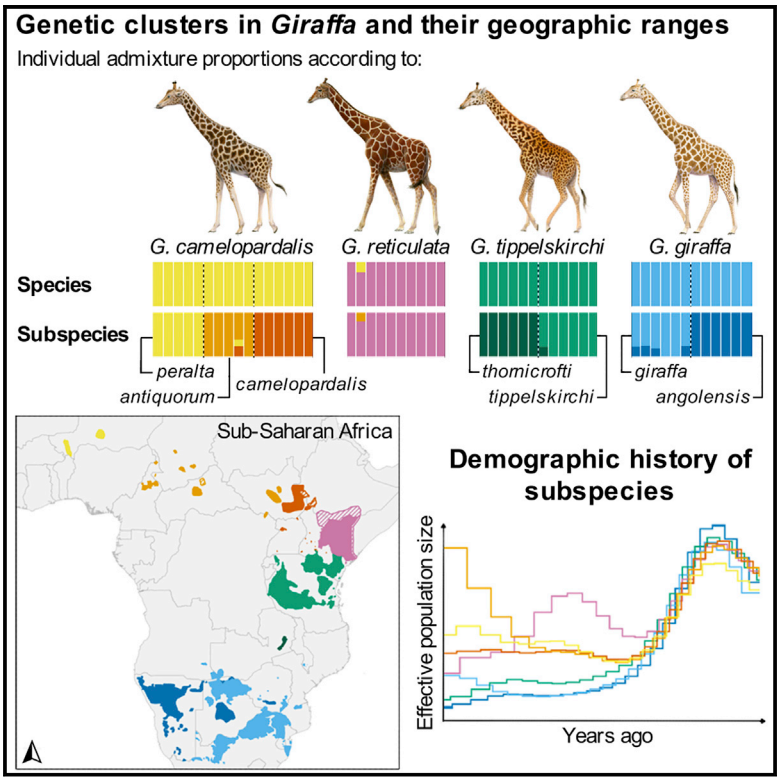
Report

Current Biology

Whole-genome analysis of giraffe supports four distinct species

Graphical abstract



Highlights

- Comprehensive whole-genome analysis of 50 individuals from all giraffe subspecies
- Results show four separate giraffe lineages with distinct demographic histories
- Luangwa giraffe should be treated as its own separate subspecies of Masai giraffe
- Chromium *de novo* genome assembly of the critically endangered Kordofan giraffe

Authors

Raphael T.F. Coimbra, Sven Winter, Vikas Kumar, ..., Pavel Dobrynin, Julian Fennessy, Axel Janke

Correspondence

raphael.coimbra@senckenberg.de (R.T.F.C.), axel.janke@senckenberg.de (A.J.)

In brief

Coimbra et al.'s assessment on giraffe taxonomy uses genome data from all traditionally recognized subspecies. Population genomic and phylogenomic analyses support four separately evolving giraffe lineages that underwent distinct demographic histories. Their current genomic diversity will inform targeted conservation efforts for these iconic taxa.

Coimbra et al., 2021, Current Biology *31*, 2929–2938
July 12, 2021 © 2021 The Authors. Published by Elsevier Inc.
https://doi.org/10.1016/j.cub.2021.04.033

CellPress

Current Biology

CellPress
OPEN ACCESS

Report

Whole-genome analysis of giraffe supports four distinct species

Raphael T.F. Coimbra,[1,2,10,*] Sven Winter,[1,2,10] Vikas Kumar,[3,4] Klaus-Peter Koepfli,[5,6] Rebecca M. Gooley,[5,6] Pavel Dobrynin,[7] Julian Fennessy,[8] and Axel Janke[1,2,9,11,*]
[1]Senckenberg Biodiversity and Climate Research Centre, Senckenberganlage 25, 60325 Frankfurt am Main, Germany
[2]Institute for Ecology, Evolution and Diversity, Goethe University, Max-von-Laue-Straße 13, 60438 Frankfurt am Main, Germany
[3]Key Laboratory of Vertebrate Evolution and Human Origins, Institute of Vertebrate Paleontology and Paleoanthropology, Chinese Academy of Sciences, Beijing 100044, China
[4]Center for Excellence in Life and Paleoenvironment, Chinese Academy of Sciences, Beijing 100044, China
[5]Smithsonian-Mason School of Conservation, Front Royal, VA, 22630, USA
[6]Smithsonian Conservation Biology Institute, Center for Species Survival, National Zoological Park, 3001 Connecticut Avenue NW, Washington, DC 20008, USA
[7]Computer Technologies Laboratory, ITMO University, 49 Kronverkskiy Pr., Saint Petersburg 197101, Russia
[8]Giraffe Conservation Foundation, PO Box 86099, Eros, Windhoek, Namibia
[9]LOEWE Centre for Translational Biodiversity Genomics, Senckenberganlage 25, 60325 Frankfurt am Main, Germany
[10]These authors contributed equally
[11]Lead contact
*Correspondence: raphael.coimbra@senckenberg.de (R.T.F.C.), axel.janke@senckenberg.de (A.J.)
https://doi.org/10.1016/j.cub.2021.04.033

SUMMARY

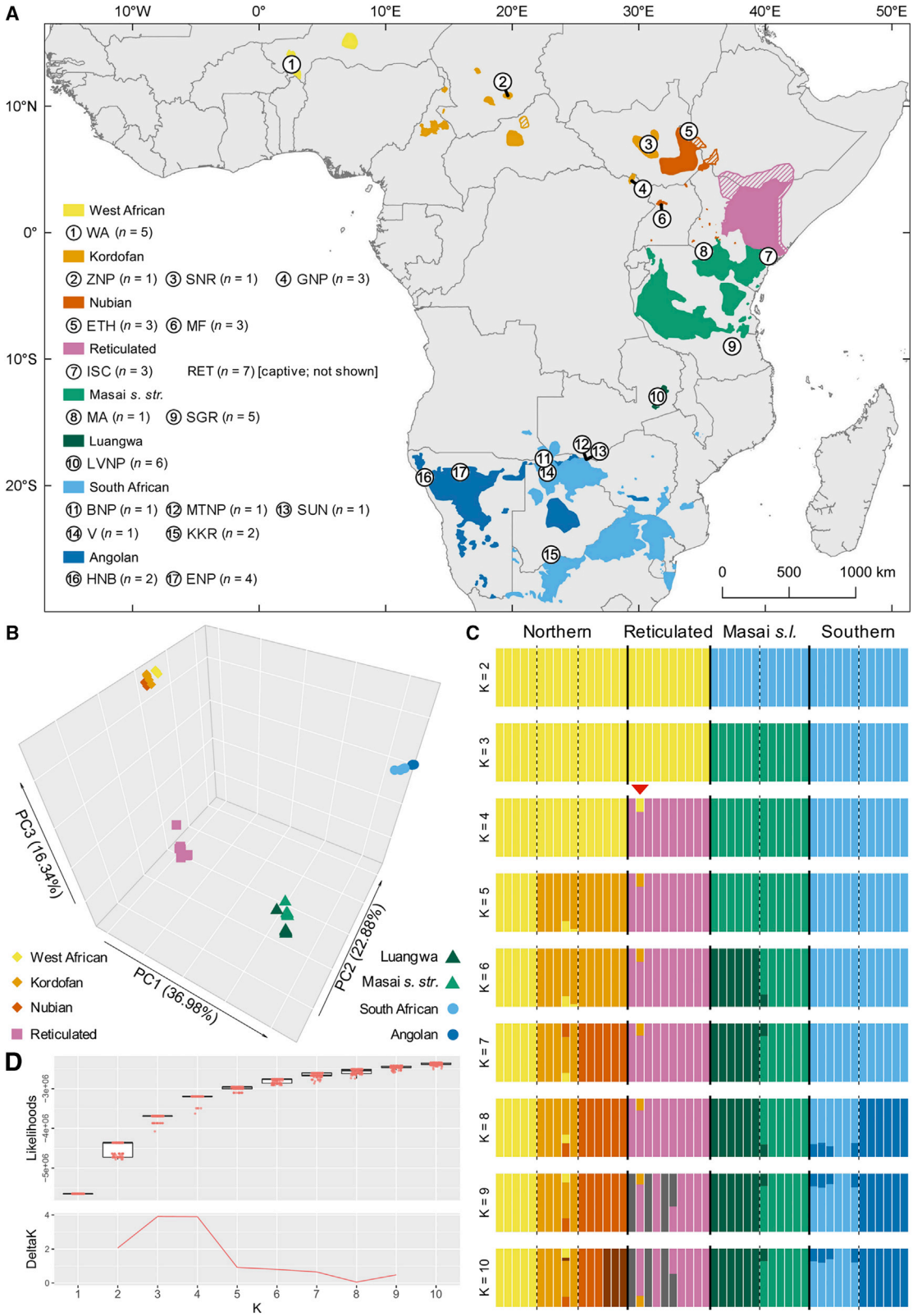
Species is the fundamental taxonomic unit in biology and its delimitation has implications for conservation. In giraffe (*Giraffa* spp.), multiple taxonomic classifications have been proposed since the early 1900s.[1] However, one species with nine subspecies has been generally accepted,[2] likely due to limited in-depth assessments, subspecies hybridizing in captivity,[3,4] and anecdotal reports of hybrids in the wild.[5] Giraffe taxonomy received new attention after population genetic studies using traditional genetic markers suggested at least four species.[6,7] This view has been met with controversy,[8] setting the stage for debate.[9,10] Genomics is significantly enhancing our understanding of biodiversity and speciation relative to traditional genetic approaches and thus has important implications for species delineation and conservation.[11] We present a high-quality *de novo* genome assembly of the critically endangered Kordofan giraffe (*G. camelopardalis antiquorum*)[12] and a comprehensive whole-genome analysis of 50 giraffe representing all traditionally recognized subspecies. Population structure and phylogenomic analyses support four separately evolving giraffe lineages, which diverged 230–370 ka ago. These lineages underwent distinct demographic histories and show different levels of heterozygosity and inbreeding. Our results strengthen previous findings of limited gene flow and admixture among putative giraffe species[6,7,9] and establish a genomic foundation for recognizing four species and seven subspecies, the latter of which should be considered as evolutionary significant units. Achieving a consensus over the number of species and subspecies in giraffe is essential for adequately assessing their threat level and will improve conservation efforts for these iconic taxa.

RESULTS AND DISCUSSION

We refer to giraffe (sub)species following the taxonomy used in,[7,9] which comprises: the northern giraffe (*G. camelopardalis*), including the Kordofan (*G. c. antiquorum*), the Nubian (*G. c. camelopardalis*), and the West African giraffe (*G. c. peralta*); the reticulated giraffe (*G. reticulata*); the Masai giraffe *sensu lato* (*G. tippelskirchi*), including the Masai giraffe *sensu stricto* (*G. t. tippelskirchi*) and the Luangwa (or Thornicroft's) giraffe (*G. t. thornicrofti*); and the southern giraffe (*G. giraffa*), including the Angolan (*G. g. angolensis*) and the South African giraffe (*G. g. giraffa*).

Genome sequencing of a male Kordofan giraffe using 10X Genomics' linked-reads[13,14] yielded a total of 818,750,030 filtered reads for a median effective sequencing depth of 30-fold. The *de novo* assembled pseudohaploid genome had a length of 2.46 Gbp with a 27.5 Mbp scaffold N50 (Figure S1A–S1D) and is consistent with a nearly chromosome-length genome assembly of the Masai giraffe *s. str.*[15–17] (Figure S1E). Repetitive regions comprised 45.11% of the Kordofan giraffe genome assembly (Table S1). Annotation of protein-coding genes revealed 59,083 gene models, with 6,267 alternative transcripts. Mapping of paired-end reads from 48 newly sequenced giraffe and two publicly available ones (GenBank: ERR1248124; GenBank: SRR3218456) against the Kordofan giraffe reference resulted in filtered median depths of 6–25-fold (Table S2). This dataset equally represents all currently recognized subspecies (Figure 1A and Table S2).

This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).



(legend on next page)

CellPress
OPEN ACCESS

A principal component analysis (PCA) and admixture analyses were performed to assess the genotypic clustering of individuals. A PCA of 193,073 putatively unlinked single nucleotide polymorphisms (SNPs) revealed four non-overlapping clusters (Figure 1B) congruent with the four giraffe species proposed in.[7,9] Admixture analyses based on the same SNPs and assuming varying numbers of ancestry components (K) discriminated between all giraffe subspecies (Figure 1C). Consistent with the PCA, individuals were consecutively clustered into north and south groups ($K = 2$), Masai giraffe *s. l.* and southern giraffe ($K = 3$), and northern and reticulated giraffe ($K = 4$). At $K \geq 4$, a captive reticulated giraffe individual (RET3) showed admixture with the northern giraffe. This is not unexpected as this individual is likely an offspring of a northern × reticulated giraffe cross in captivity.[9] The large number of SNPs allowed admixture analyses to detect further clustering at higher K. However, $K = 4$ showed the highest median among run likelihoods per K before the increase in median values dropped and the increase in interquartile ranges became noticeable (Figure 1D). This interpretation is analogous to the "plateau method"[20] and suggests that $K = 4$ is the optimal choice for the uppermost level of population structure as it represents the smallest value that distinguishes the primary clusters in the data. The ΔK method[21] (Figure 1D), however, indicates both $K = 3$ and $K = 4$ as equally likely. A recent re-analysis of published multi-locus sequence data for giraffe found support for different K values depending on parameter settings of the admixture analysis.[10] The authors favored the $K = 3$ solution by weighting evidence from the plateau and ΔK methods, and then selecting the smaller value of K. We argue that $K = 4$ is a more appropriate interpretation because it is consistent with both the PCA clustering and the plateauing of run likelihoods.

Genome consensus sequences were generated for all individuals. The Luangwa giraffe from the Luangwa Valley National Park (LVNP), Zambia, and one reticulated giraffe (ISC01) were excluded due to excessive amounts of undetermined bases (~25%–38%) resulting from poor cumulative genome coverage. Remaining consensus sequences were aligned by scaffold and sites with a proportion of masked bases > 0.2 were discarded. Increasingly larger genome fragments (GFs) were analyzed with an approximately unbiased (AU) test[22] to identify the minimum sequence length required to reject alternative topologies of the giraffe tree using the okapi (*Okapia johnstoni*) as an outgroup. Fourteen of the 15 alternative topologies among the four putative species of giraffe were rejected with GF lengths[3] $\geq$ 450 kbp (Figure S2), yielding 1,068 non-overlapping GFs (480.6 Mbp, ~42% of the non-repetitive genome length), which were used for phylogenomic and network analyses.

A multispecies coalescent (MSC) tree inferred from GF trees revealed fully supported monophyletic clades corresponding to the currently recognized subspecies (Figure 2A). The only exception was the South African giraffe, which was paraphyletic with respect to Angolan giraffe. Early tree branching patterns are identical to the clustering hierarchy in both PCA and admixture analyses, with the separation between north and south groups of giraffe, followed by the split between northern and reticulated, and southern and Masai giraffe *s. l.* These four clades constitute the most divergent lineages within giraffe, consistent with previous studies.[7,9] By contrast, a recent phylogenetic re-analysis of 21 intron sequences using the SuperTRI method[10] did not find consistent support for the dichotomy between northern and reticulated giraffe. However, the genome-scale MSC analysis resolved this disputed divergence with full local posterior probability and high quartet frequency support (Figure 2B, branch no. 8). Discordance among GF trees was found only within northern giraffe for the branch leading to the ancestor of Kordofan and Nubian giraffe, where quartet frequencies were similar for two of the three possible topologies around that branch (Figure 2B, branch no. 9). Thus, only the relationship between northern giraffe subspecies could not be resolved.

A phylogeny based on mitochondrial protein-coding genes complied with previously reported topologies.[7,9] The most notable difference to the MSC tree is the grouping of South African and Masai giraffe *s. l.* (Figure S3A). Such conflict between mitochondrial and nuclear genomes is not uncommon and is likely explained by mitochondrial introgression from the Masai giraffe *s. l.* into the southern giraffe via past hybridization, with later fixation of the introgressed mitochondria in the South African giraffe but maintenance of the ancestral type in the Angolan giraffe.[10] Other instances of potential mitochondrial introgression via hybridization regard two captive reticulated giraffe individuals (RET4 and RETRot1), which grouped with the Nubian giraffe. Further, the reciprocal monophyly between South African and Angolan giraffe in the mitochondrial tree (Figure S3A) and the presence of admixture between these two subspecies in the nuclear genome (Figure 1C) likely reflects female philopatry and male biased dispersal.[10]

A consensus network built from GF trees discriminates the four most divergent giraffe lineages at all assessed thresholds for conflicting edges (Figure 2C and Figure S3C–F). The distinction between the three subspecies within the northern giraffe is evident, although conflicting phylogenetic signals mask their correct relationship. At thresholds $\leq$ 15%, an increase in reticulations

Figure 1. Population structure of giraffe

(A) Map of sub-Saharan Africa indicating the ranges of each giraffe (sub)species (colored shadings),[18] with corrected distributions for Kordofan[7] and Angolan giraffe,[19] and sampling locations (numbered circles). Cross-hatched areas indicate where giraffe populations are not confirmed, but possibly occur. Acronyms for sampling locations correspond to the first letters of the sample identifiers. See also Table S2.

(B) PCA of 193,073 unlinked SNPs from 50 individuals showing four distinct clusters. PC1 separates north and south (geographical) giraffe clusters, PC2 separates southern and Masai giraffe *s. l.*, and PC3 distinguishes northern and reticulated giraffe. Shapes and colors represent (sub)species.

(C) Admixture analysis based on the same SNP dataset with K ranging from 2 to 10. All commonly accepted subspecies are partitioned. The grouping into four clusters represents the smallest value of K that explains the primary structure in the data, consistent with the PCA and the plot of likelihoods per K (C). Colors indicate an individual's cluster membership as in (A). A captive reticulated giraffe individual (RET3; red arrowhead) shows possible hybridization with a northern giraffe.

(D) Plots of run likelihoods and ΔK values per K. In the boxplot of run likelihoods per K, the increase in median values drops for $K > 4$, whereas statistical dispersion increases. In the ΔK plot, $K = 3$ and $K = 4$ are virtually indistinguishable.

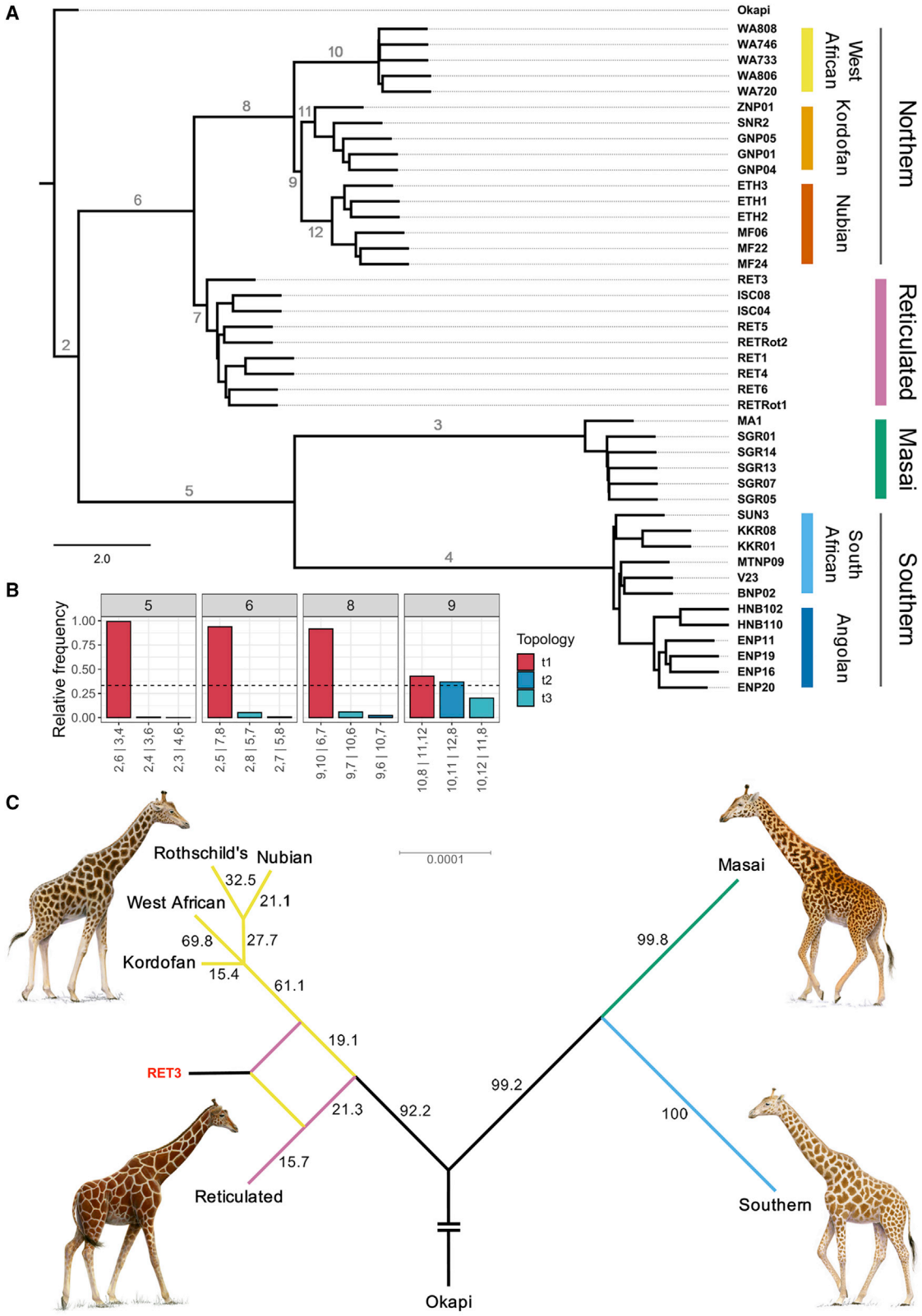


Figure 2. Nuclear phylogenomic relationships among giraffe

(A) MSC tree based on 1,068 maximum likelihood genome fragment (GF) trees showing four most divergent monophyletic clades. Northern + reticulated giraffe are sister species to Masai + southern giraffe. Kordofan and Nubian giraffe are sister to West African giraffe, whereas South African giraffe is paraphyletic. The

(legend continued on next page)

CellPress
OPEN ACCESS

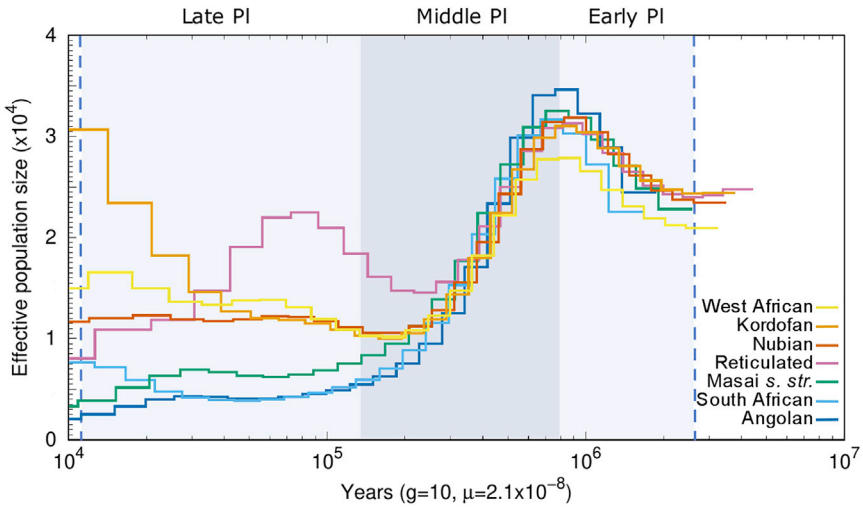


Figure 3. Demographic history of giraffe species and subspecies

Changes in N_e over time were estimated using the PSMC model. Demographic trajectories are shown for one representative of each (sub)species and are distinct for each putative species but similar within them. Plots were scaled by a mutation rate of 2.12×10^{-8} substitutions per site per generation and a generation time of 10 years. Alternating blue shadings indicate Early, Middle, and Late subepochs of the Pleistocene (Pl). See also Figure S4.

between northern and reticulated giraffe and among the subspecies of the northern giraffe implies phylogenetic discordance (Figure S3C–F). The individual RET3 is placed in a reticulation intermediate to the northern and reticulated giraffe, suggesting that this individual is a hybrid (Figure 2C and Figure S3C–F).

The genome data allowed revisiting the number of subspecies within the Masai giraffe *s. l*. The Luangwa giraffe was previously subsumed into the Masai giraffe *s. str.* based on the lack of differentiation in mitochondrial DNA.[7] However, the genome analyses found evidence of substructure among them, consistent with.[9] The mitochondrial phylogeny (Figure S3A) groups Luangwa giraffe individuals from the LVNP with one Masai giraffe *s. str.* individual (MA1) from the Masai Mara National Reserve, Kenya, and separates them from Masai giraffe *s. str.* individuals from the Selous Game Reserve, Tanzania. Moreover, the admixture analyses on genome-wide SNPs at $K \geq 6$ (Figure 1C) and a neighbor-joining tree (Figure S3B) both support two clusters within Masai giraffe *s. l.* Although the lack of samples from intermediate Masai giraffe *s. str.* populations hinders a definitive conclusion, we argue that Luangwa giraffe should be treated as a separate subspecies of Masai giraffe *s. l*. Its apparent genetic distinctiveness from Masai giraffe *s. str.* populations and its low numbers in the wild (~600 individuals[23]) reinforce the importance of enhancing conservation efforts for the taxon and justify keeping its subspecies status.

An orthology search between the Kordofan giraffe and 11 Cetartiodactyla species (Table S3) identified 2,127 protein-coding sequences (~1.49 Mbp) for reconstructing and dating the Cetartiodactyla tree (Figure S3G and Table S4). Giraffe and okapi diverged in the Middle Miocene at ~12 Ma ago. Divergences among the four suggested giraffe species occurred in the Middle Pleistocene between 370 and 230 ka, younger than previously reported.[7] Still, the divergence times are on the order of that between polar and brown bears[24] and are consistent with the period in which different demographic trends started to develop for each putative species (~300 ka; see below).

Changes in effective population size (N_e) over time were assessed with the pairwise sequentially Markovian coalescent (PSMC) model[25] for three individuals of each currently recognized subspecies (Figure 3 and Figure S4), except for the Luangwa giraffe (see STAR Methods). Demographic histories are similar among individuals of the same (sub)species but differ among the suggested species, in particular between the northern + reticulated and the Masai *s. str.* + southern giraffe (~100–200 ka). Although fluctuations in N_e can be driven by factors unrelated to speciation, e.g., climatic oscillations, diseases, and human activities,[26] the point where PMSC trajectories diverge is consistent with the cessation of gene flow among ancestral populations and subsequent splitting of lineages.[27] Moreover, distinct demographic histories are characteristic for even closely related and potentially hybridizing species such as whales[28] and big cats.[29]

A PSMC plot of all (sub)species (Figure 3) shows the highest N_e following population expansions around 1 Ma ago after a series of paleoclimatic shifts during the transition from Early to Middle Pleistocene led to increasing aridification across Africa.[30] Subsequently, the N_e of each (sub)species declines, precipitously in the case of Masai giraffe *s. str.* and southern giraffe, during the Middle and Late Pleistocene. These time periods coincided with cooler glacial and warmer interglacial periods that are marked by decreasing forest cover and the expansion and shifting of savannah grasslands associated with the increasingly arid climate

former Rothschild's giraffe grouped with the Nubian giraffe. Luangwa giraffe was not included (see STAR Methods). Internal branches are numbered. Branches defining (sub)species received full local posterior probability support. Branch lengths are in coalescent units and indicate GF tree discordance. See also Figure S2 and Figure S3.

(B) Quartet frequencies of alternative arrangements around internal branches (no. 5, 6, 8, and 9) of the giraffe tree in (A). Bars represent relative frequencies of the MSC tree topology (red) and alternative arrangements (blue). Labels in the x axis show branch configurations in each topology. Dashed lines indicate the one-third quartet frequency threshold. Similar quartet frequencies at branch no. 9 indicate phylogenetic incongruence.

(C) Consensus network at 15% threshold for conflicting edges, constructed from GF trees, discriminates four most divergent lineages within giraffe. Conflicting phylogenetic signals mask the relationship among northern giraffe subspecies. Individual RET3 is placed in a reticulation between northern and reticulated giraffe, suggesting hybridization. Southern giraffe subspecies cannot be resolved, and Luangwa giraffe is not included. Edge labels show percentage of splits among GF trees. Network tips were collapsed to the respective (sub)species. Branch coloring indicate species classification. See also Figure S3.

Figure 4. Genomic diversity among giraffe

(A) Global heterozygosity estimates per individual. Heterozygosity was calculated from the folded site frequency spectrum and reflects the fraction of sites where an individual was heterozygous. Error bars represent the standard deviation around the mean calculated from 200 bootstrap replicates.

(legend continued on next page)

CellPress
OPEN ACCESS

across the continent.[31,32] Approximately 30 ka ago, South African giraffe numbers started to recover, while both the Masai giraffe *s. str.* and the Angolan giraffe followed with a decline. The northern giraffe show either stabilization (Nubian giraffe) or an increase (West African and Kordofan giraffe) in N_e during the Late Pleistocene and Holocene, while reticulated giraffe show a signature of expansion followed by decline during these time periods. During the transition from Middle to Late Pleistocene and a large period of the latter, the reticulated giraffe had the highest N_e among giraffe species. Inferences of N_e for the subspecies of the northern giraffe since the Holocene are higher than their current total census estimates (~5,600 individuals; https://giraffeconservation.org/). However, these should be interpreted with caution given the lower power of the PSMC model for recent time periods.[27] Still, the small populations of northern giraffe represent only a tiny remnant of their original distribution. The declines observed for reticulated, southern, and Masai giraffe *s. str.* coincide with the early expansion of human populations and their activities, as has been suggested for other ruminant species.[33]

Genome-wide heterozygosity and the realized inbreeding coefficient (F_{ROH}) based on the proportion of the genome in runs of homozygosity (ROH) were estimated per individual (Figure 4). The differences in the inferred F_{ROH} levels among the four putative species do not correlate with their current census sizes. Despite low numbers in the northern giraffe, the amount of heterozygosity is moderate (Figure 4A) and their F_{ROH} is relatively low (Figure 4B–C), falling within the range of many mammals.[34] The northern giraffe once ranged abundantly across a large area of the sub-Saharan Africa. Regardless of their drastic decline over the last century, the danger of inbreeding in today's small and isolated populations has not yet manifested in their genomes. This is especially encouraging for the West African giraffe, where conservation efforts by the Government of Niger and partners have resulted in an increase from 49 to > 600 individuals during the last 25 years,[35] likely helping to mitigate inbreeding and genetic erosion in this population.

The reticulated giraffe has the highest heterozygosity (Figure 4A) and the lowest signs of inbreeding (Figure 4B–C). Seven individuals are from zoos and their pedigrees go back three to four generations to wild individuals (studbooks not shown). Informed breeding recommendations likely explain their low F_{ROH} levels, but these patterns are also seen in three wild individuals from Kenya. Nevertheless, reticulated giraffe numbers have declined by > 50% over the last 30 years, making the possibility of genetic erosion a serious concern for the future.

In contrast, we found unexpectedly low levels of heterozygosity (Figure 4A) and high levels of inbreeding (Figure 4B–C) in the Masai giraffe *s. l.* and southern giraffe. Both species occur in relatively large numbers (Masai *s. l.*: ~35,000; southern: 54,750; https://giraffeconservation.org/) and have not gone through any known population bottlenecks. However, the pattern of F_{ROH} is consistent with their N_e trajectories (Figure 3), both showing a precipitous decline during Middle and Late Pleistocene and toward the present. The highest levels of F_{ROH} are found in five of the six Luangwa giraffe individuals, which are geographically isolated in and around the Luangwa Valley, Zambia. The relatively small numbers and geographic isolation of this population may have contributed to the high inbreeding observed. In contrast, Masai giraffe *s. str.* and southern giraffe sampled from protected areas (Selous Game Reserve, Tanzania, and Etosha National Park, Namibia) harboring large and "connected" populations contained smaller numbers and lower amounts of ROH (Figures 4B and 4C). A denser sampling of each giraffe species will yield deeper insights into the factors contributing to the differences in inbreeding and adaptive potential among populations.

Most biologists agree that species are "separately evolving metapopulation lineages"[36], but empirically assigning species ranks is not trivial and relies on the operational criteria used.[37] Here, we showed that all population genomic and phylogenomic analyses of 50 giraffe individuals are consistent with each other and support four separately evolving lineages with distinct demographic histories. No significant signals of hybridization between those lineages were observed in wild individuals, but three captive reticulated giraffe showed signs of either admixture or mitochondrial introgression. Furthermore, morphology,[38] ecology,[39] and limited gene flow[9] suggest the existence of more than one giraffe species, although the exact number is still disputed with recent in-depth genetic assessments supporting either three[10] or four[7,9] species. We suggest that the four most divergent lineages of giraffe revealed here conform to the broader unified concept of species.[36] They are also consistent with long-term reproductive isolation (albeit incomplete, as indicated by captive individuals), the primary criterion for delineating species under the biological species concept,[40] used as the definition of species by the IUCN (https://www.iucn.org/sites/dev/files/iucn-glossary-of-definitions_en.pdf). Studies of behavior, ecology, and physiology are warranted to understand reproductive compatibility/incompatibility and divergence. Current levels of genomic diversity suggest that the threat of genetic erosion is not as concerning as expected for the northern giraffe, however, it is particularly alarming for the Luangwa giraffe. On the basis of the genomic separation among the subspecies found in the northern, Masai *s. l.*, and southern giraffe, we suggest these should be recognized as evolutionary significant units that constitute important interpopulation diversity in these lineages. Overall, our results conform with previous findings of limited admixture and gene flow among giraffe populations[6,7,9] and provide a solid genomic basis for the recognition of four distinct species of giraffe, not one.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

(B) Realized inbreeding coefficients (F_{ROH}) per individual calculated using the model-based RZooRoH method. Bars are partitioned into different colors that represent the proportion of the genome assigned to up to six length-related classes of homozygosity-by-descent (HBD). HBD classes 4 (longest), 16, 64, 320, 1280, and 5120 (shortest) correspond approximately to ancestors inbreeding 2, 8, 32, 160, 640, and 2560 generations ago. Note that the 5120 class is representative of N_e in the distant past, rather than individual inbreeding levels.

(C) Number of ROH segments versus accumulated length of ROH per individual. Note that ROH correspondent to HBD segments of class 5120 are not included. Shapes and colors represent (sub)species.

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
  - Kordofan giraffe genome assembly and annotation
  - Whole-genome re-sequencing
  - Quality control and read mapping
  - Genotype likelihoods
  - Pruning of linked sites
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Population structure and admixture analyses
  - Genetic distances and neighbor-joining tree
  - Nuclear phylogenomic inference
  - Assembly and phylogeny of mitochondrial genomes
  - Divergence time estimates
  - Demographic reconstruction
  - Heterozygosity
  - ROH and Inbreeding

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cub.2021.04.033.

ACKNOWLEDGMENTS

We thank an array of partners, government and non-government, across Africa who collaborated with the Giraffe Conservation Foundation (GCF) to permit, collect, and include samples in this analysis, including: African Parks Network; Governments of Botswana, Ethiopia, Kenya, Namibia, Niger, Tanzania, and Zambia; San Diego Zoo Global; Uganda Wildlife Authority; Wildlife Conservation Alliance; and Wildlife Conservation Society. We are grateful to D. Baranski and C. Greve (TBG Laboratory Center) for their expert support in genome sequencing, J. Jebram (Opel-Zoo Kronberg) for the studbook information, and S. Hofman at the Royal Zoological Society of Antwerp for providing the Kordofan giraffe sample. Giraffe paintings are by J.B. Hlíðberg (www.fauna.is). The present study is also a product of the Centre for Translational Biodiversity Genomics (LOEWE-TBG) as part of the "LOEWE—Landes-Offensive zur Entwicklung Wissenschaftlich-ökonomischer Exzellenz" program of Hesse's Ministry of Higher Education, Research, and the Arts as well as the Leibniz Association.

AUTHOR CONTRIBUTIONS

R.T.F.C., S.W., and A.J. conceived the study. A.J. funded the genome sequencing. J.F. collected and/or coordinated field sampling. S.W. carried out the laboratory procedures and assembled the Kordofan giraffe genome. R.T.F.C. performed most of the computational analyses in the study. V.K. performed the divergence time analysis. K.-P.K., R.M.G., and P.D. estimated the inbreeding based on runs of homozygosity. R.T.F.C. and A.J. wrote the manuscript with input from all authors.

DECLARATION OF INTERESTS

The authors declare no competing interests.

INCLUSION AND DIVERSITY

We worked to ensure diversity in experimental samples through the selection of the genomic datasets. One or more of the authors of this paper self-identifies as an underrepresented ethnic minority in science. One or more of the authors of this paper self-identifies as a member of the LGBTQ+ community.

Received: July 4, 2020
Revised: January 6, 2021
Accepted: April 14, 2021
Published: May 5, 2021

REFERENCES

1. Shorrocks, B. (2016). Present distribution and geographical races. In The Giraffe: Biology, Ecology, Evolution and Behaviour (John Wiley & Sons), pp. 26–41.
2. Dagg, A.I. (1971). Giraffa camelopardalis. Mamm. Species, 1–8.
3. Lackey, L.B. (2011). Giraffe, *Giraffa camelopardalis*, Eighth Edition (North American Regional/Global Studbook).
4. Gray, A.P. (1972). Mammalian Hybrids: A Check-List with Bibliography, Second Edition (Commonwealth Agricultural Bureaux).
5. Stott, K. (1959). Giraffe intergradation in Kenya. J. Mammal. *40*, 251.
6. Brown, D.M., Brenneman, R.A., Koepfli, K.-P., Pollinger, J.P., Milá, B., Georgiadis, N.J., Louis, E.E., Jr., Grether, G.F., Jacobs, D.K., and Wayne, R.K. (2007). Extensive population genetic structure in the giraffe. BMC Biol. *5*, 57.
7. Fennessy, J., Bidon, T., Reuss, F., Kumar, V., Elkan, P., Nilsson, M.A., Vamberger, M., Fritz, U., and Janke, A. (2016). Multi-locus analyses reveal four giraffe species instead of one. Curr. Biol. *26*, 2543–2549.
8. Bercovitch, F.B., Berry, P.S.M., Dagg, A., Deacon, F., Doherty, J.B., Lee, D.E., Mineur, F., Muller, Z., Ogden, R., Seymour, R., et al. (2017). How many species of giraffe are there? Curr. Biol. *27*, R136–R137.
9. Winter, S., Fennessy, J., and Janke, A. (2018). Limited introgression supports division of giraffe into four species. Ecol. Evol. *8*, 10156–10165.
10. Petzold, A., and Hassanin, A. (2020). A comparative approach for species delimitation based on multiple methods of multi-locus DNA sequence analysis: A case study of the genus *Giraffa* (Mammalia, Cetartiodactyla). PLoS ONE *15*, e0217956.
11. Supple, M.A., and Shapiro, B. (2018). Conservation of biodiversity in the genomics era. Genome Biol. *19*, 131.
12. Fennessy, J., and Marais, A. (2018). Giraffa camelopardalis ssp. antiquorum. In The IUCN Red List of Threatened Species 2018: e.T88420742A88420817 (International Union for Conservation of Nature and Natural Resources).
13. Marks, P., Garcia, S., Barrio, A.M., Belhocine, K., Bernate, J., Bharadwaj, R., Bjornson, K., Catalanotti, C., Delaney, J., Fehr, A., et al. (2019). Resolving the full spectrum of human genome variation using Linked-Reads. Genome Res. *29*, 635–645.
14. Zheng, G.X.Y., Lau, B.T., Schnall-Levin, M., Jarosz, M., Bell, J.M., Hindson, C.M., Kyriazopoulou-Panagiotopoulou, S., Masquelier, D.A., Merrill, L., Terry, J.M., et al. (2016). Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. Nat. Biotechnol. *34*, 303–311.
15. Agaba, M., Ishengoma, E., Miller, W.C., McGrath, B.C., Hudson, C.N., Bedoya Reina, O.C., Ratan, A., Burhans, R., Chikhi, R., Medvedev, P., et al. (2016). Giraffe genome sequence reveals clues to its unique morphology and physiology. Nat. Commun. *7*, 11519.
16. Dudchenko, O., Batra, S.S., Omer, A.D., Nyquist, S.K., Hoeger, M., Durand, N.C., Shamim, M.S., Machol, I., Lander, E.S., Aiden, A.P., and Aiden, E.L. (2017). De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. Science *356*, 92–95.
17. Dudchenko, O., Shamim, M.S., Batra, S., Durand, N.C., Musial, N.T., Mostofa, R., Pham, M., Glenn St. Hilaire, B., Yao, W., Stamenova, E., et al. (2018). The Juicebox Assembly Tools module facilitates de novo assembly of mammalian genomes with chromosome-length scaffolds for under $1000. bioRxiv, 254797.
18. O'Connor, D., Stacy-Dawes, J., Muneza, A., Fennessy, J., Gobush, K., Chase, M.J., Brown, M.B., Bracis, C., Elkan, P., Zaberirou, A.R.M., et al.

(2019). Updated geographic range maps for giraffe, *Giraffa* spp., throughout sub-Saharan Africa, and implications of changing distributions for conservation. Mammal Rev. *49*, 285–299.

19. Winter, S., Fennessy, J., Fennessy, S., and Janke, A. (2018). Matrilineal population structure and distribution of the Angolan giraffe in the Namib desert and beyond. Ecol. Genet. Genomics *7–8*, 1–5.
20. Pritchard, J.K., Wen, X., and Falush, D. (2010). Documentation for STRUCTURE software: version 2.3.
21. Evanno, G., Regnaut, S., and Goudet, J. (2005). Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Mol. Ecol. *14*, 2611–2620.
22. Shimodaira, H. (2002). An approximately unbiased test of phylogenetic tree selection. Syst. Biol. *51*, 492–508.
23. Bercovitch, F., Carter, K., Fennessy, J., and Tutchings, A. (2018). Giraffa camelopardalis ssp. thornicrofti. In The IUCN Red List of Threatened Species 2018: e.T88421020A88421024 (International Union for Conservation of Nature and Natural Resources).
24. Hailer, F., Kutschera, V.E., Hallström, B.M., Klassert, D., Fain, S.R., Leonard, J.A., Arnason, U., and Janke, A. (2012). Nuclear genomic sequences reveal that polar bears are an old and distinct bear lineage. Science *336*, 344–347.
25. Li, H., and Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. Nature *475*, 493–496.
26. Ellegren, H., and Galtier, N. (2016). Determinants of genetic diversity. Nat. Rev. Genet. *17*, 422–433.
27. Nadachowska-Brzyska, K., Burri, R., Smeds, L., and Ellegren, H. (2016). PSMC analysis of effective population sizes in molecular ecology and its application to black-and-white *Ficedula* flycatchers. Mol. Ecol. *25*, 1058–1072.
28. Árnason, Ú., Lammers, F., Kumar, V., Nilsson, M.A., and Janke, A. (2018). Whole-genome sequencing of the blue whale and other rorquals finds signatures for introgressive gene flow. Sci. Adv. *4*, eaap9873.
29. Figueiró, H.V., Li, G., Trindade, F.J., Assis, J., Pais, F., Fernandes, G., Santos, S.H.D., Hughes, G.M., Komissarov, A., Antunes, A., et al. (2017). Genome-wide signatures of complex introgression and adaptive evolution in the big cats. Sci. Adv. *3*, e1700299.
30. DeMenocal, P.B. (2004). African climate change and faunal evolution during the Pliocene–Pleistocene. Earth Planet. Sci. Lett. *220*, 3–24.
31. Bonnefille, R. (2010). Cenozoic vegetation, climate changes and hominid evolution in tropical Africa. Global Planet. Change *72*, 390–411.
32. Castañeda, I.S., Caley, T., Dupont, L., Kim, J.-H., Malaizé, B., and Schouten, S. (2016). Middle to Late Pleistocene vegetation and climate change in subtropical southern East Africa. Earth Planet. Sci. Lett. *450*, 306–316.
33. Chen, L., Qiu, Q., Jiang, Y., Wang, K., Lin, Z., Li, Z., Bibi, F., Yang, Y., Wang, J., Nie, W., et al. (2019). Large-scale ruminant genome sequencing provides insights into their evolution and distinct traits. Science *364*, eaav6202.
34. Brüniche-Olsen, A., Kellner, K.F., Anderson, C.J., and DeWoody, J.A. (2018). Runs of homozygosity have utility in mammalian conservation and evolutionary studies. Conserv. Genet. *19*, 1295–1307.
35. Fennessy, J., Marais, A., and Tutchings, A. (2018). Giraffa camelopardalis ssp. peralta. In The IUCN Red List of Threatened Species 2018: e.T136913A51140803 (International Union for Conservation of Nature and Natural Resources).
36. De Queiroz, K. (2007). Species concepts and species delimitation. Syst. Biol. *56*, 879–886.
37. Zachos, F.E. (2018). Mammals and meaningful taxonomic units: the debate about species concepts and conservation. Mammal Rev. *48*, 153–159.
38. Groves, C., and Grubb, P. (2011). Giraffidae. In Ungulate Taxonomy (Johns Hopkins Univ. Press), pp. 64–70.
39. Thomassen, H.A., Freedman, A.H., Brown, D.M., Buermann, W., and Jacobs, D.K. (2013). Regional differences in seasonal timing of rainfall discriminate between genetically distinct East African giraffe taxa. PLoS ONE *8*, e77191.
40. Mayr, E. (1942). Systematics and the Origin of Species (Columbia University Press).
41. Hassanin, A., Delsuc, F., Ropiquet, A., Hammer, C., Jansen van Vuuren, B., Matthee, C., Ruiz-Garcia, M., Catzeflis, F., Areskoug, V., Nguyen, T.T., and Couloux, A. (2012). Pattern and timing of diversification of Cetartiodactyla (Mammalia, Laurasiatheria), as revealed by a comprehensive analysis of mitochondrial genomes. C. R. Biol. *335*, 32–50.
42. Keane, M., Semeiks, J., Webb, A.E., Li, Y.I., Quesada, V., Craig, T., Madsen, L.B., van Dam, S., Brawand, D., Marques, P.I., et al. (2015). Insights into the evolution of longevity from the bowhead whale genome. Cell Rep. *10*, 112–122.
43. Weisenfeld, N.I., Kumar, V., Shah, P., Church, D.M., and Jaffe, D.B. (2017). Direct determination of diploid genome sequences. Genome Res. *27*, 757–767.
44. Mikheenko, A., Prjibelski, A., Saveliev, V., Antipov, D., and Gurevich, A. (2018). Versatile genome assembly evaluation with QUAST-LG. Bioinformatics *34*, i142–i150.
45. Waterhouse, R.M., Seppey, M., Simão, F.A., Manni, M., Ioannidis, P., Klioutchnikov, G., Kriventseva, E.V., and Zdobnov, E.M. (2018). BUSCO applications from quality assessments to gene prediction and phylogenomics. Mol. Biol. Evol. *35*, 543–548.
46. Chu, J. (2018). Jupiter Plot: a Circos-based tool to visualize genome assembly consistency (version 1.0). Zenodo. https://doi.org/10.5281/zenodo.1241235.
47. Hoff, K.J., Lomsadze, A., Borodovsky, M., and Stanke, M. (2019). Whole-Genome Annotation with BRAKER. In Gene Prediction: Methods and Protocols, M. Kollmar, ed. (Humana), pp. 65–95.
48. Gremme, G. (2012). Computational Gene Structure Prediction. Doctoral dissertation (University of Hamburg).
49. Stanke, M., Diekhans, M., Baertsch, R., and Haussler, D. (2008). Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics *24*, 637–644.
50. Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114–2120.
51. Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. ArXiv, 1303.3997v2.
52. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The sequence alignment/map format and SAMtools. Bioinformatics *25*, 2078–2079.
53. Okonechnikov, K., Conesa, A., and García-Alcalde, F. (2016). Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. Bioinformatics *32*, 292–294.
54. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M.A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. *20*, 1297–1303.
55. Korneliussen, T.S., Albrechtsen, A., and Nielsen, R. (2014). ANGSD: analysis of next generation sequencing data. BMC Bioinformatics *15*, 356.
56. Fox, E.A., Wright, A.E., Fumagalli, M., and Vieira, F.G. (2019). ngsLD: evaluating linkage disequilibrium using genotype likelihoods. Bioinformatics *35*, 3855–3856.
57. Meisner, J., and Albrechtsen, A. (2018). Inferring population structure and admixture proportions in low-depth NGS data. Genetics *210*, 719–731.
58. R Core Team (2019). R: a language and environment for statistical computing.
59. Skotte, L., Korneliussen, T.S., and Albrechtsen, A. (2013). Estimating individual admixture proportions from next generation sequencing data. Genetics *195*, 693–702.

60. Kopelman, N.M., Mayzel, J., Jakobsson, M., Rosenberg, N.A., and Mayrose, I. (2015). Clumpak: a program for identifying clustering modes and packaging population structure inferences across K. Mol. Ecol. Resour. *15*, 1179–1191.

61. Vieira, F.G., Lassalle, F., Korneliussen, T.S., and Fumagalli, M. (2016). Improving the estimation of genetic distances from next-generation sequencing data. Biol. J. Linn. Soc. Lond. *117*, 139–149.

62. Lefort, V., Desper, R., and Gascuel, O. (2015). FastME 2.0: a comprehensive, accurate, and fast distance-based phylogeny inference program. Mol. Biol. Evol. *32*, 2798–2800.

63. Nguyen, L.-T., Schmidt, H.A., von Haeseler, A., and Minh, B.Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol. *32*, 268–274.

64. Yu, G., Smith, D.K., Zhu, H., Guan, Y., and Lam, T.T.-Y. (2017). ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods Ecol. Evol. *8*, 28–36.

65. Zhang, C., Rabiee, M., Sayyari, E., and Mirarab, S. (2018). ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. BMC Bioinformatics *19* (*Suppl 6*), 153.

66. Sayyari, E., Whitfield, J.B., and Mirarab, S. (2018). DiscoVista: Interpretable visualizations of gene tree discordance. Mol. Phylogenet. Evol. *122*, 110–115.

67. Huson, D.H., and Bryant, D. (2006). Application of phylogenetic networks in evolutionary studies. Mol. Biol. Evol. *23*, 254–267.

68. Meng, G., Li, Y., Yang, C., and Liu, S. (2019). MitoZ: a toolkit for animal mitochondrial genome assembly, annotation and visualization. Nucleic Acids Res. *47*, e63–e63.

69. Katoh, K., and Standley, D.M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. *30*, 772–780.

70. Lechner, M., Findeiss, S., Steiner, L., Marz, M., Stadler, P.F., and Prohaska, S.J. (2011). Proteinortho: detection of (co-)orthologs in large-scale analysis. BMC Bioinformatics *12*, 124.

71. Talavera, G., and Castresana, J. (2007). Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst. Biol. *56*, 564–577.

72. Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. *24*, 1586–1591.

73. Puttick, M.N. (2019). MCMCtreeR: functions to prepare MCMCtree analyses and visualize posterior ages on trees. Bioinformatics *35*, 5321–5322.

74. Bertrand, A.R., Kadri, N.K., Flori, L., Gautier, M., and Druet, T. (2019). RZooRoH: an R package to characterize individual genomic autozygosity and identify homozygous-by-descent segments. Methods Ecol. Evol. *10*, 860–866.

75. Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O., and Walichiewicz, J. (2005). Repbase Update, a database of eukaryotic repetitive elements. Cytogenet. Genome Res. *110*, 462–467.

76. Li, H. (2011). Improving SNP discovery by base alignment quality. Bioinformatics *27*, 1157–1158.

77. Gascuel, O. (1997). BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. Mol. Biol. Evol. *14*, 685–695.

78. Hoang, D.T., Chernomor, O., von Haeseler, A., Minh, B.Q., and Vinh, L.S. (2018). UFBoot2: improving the ultrafast bootstrap approximation. Mol. Biol. Evol. *35*, 518–522.

79. Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K.F., von Haeseler, A., and Jermiin, L.S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. Nat. Methods *14*, 587–589.

80. Sayyari, E., and Mirarab, S. (2016). Fast coalescent-based computation of local branch support from quartet frequencies. Mol. Biol. Evol. *33*, 1654–1668.

81. Chernomor, O., von Haeseler, A., and Minh, B.Q. (2016). Terrace aware data structure for phylogenomic inference from supermatrices. Syst. Biol. *65*, 997–1008.

82. Muller, Z., Bercovitch, F., Brand, R., Brown, D., Brown, M., Bolger, D., Carter, K., Deacon, F., Doherty, J.B., Fennessy, J., et al. (2018). Giraffa camelopardalis (amended version of 2016 assessment). In The IUCN Red List of Threatened Species 2018: e.T9194A136266699 (International Union for Conservation of Nature and Natural Resources).

CellPress
OPEN ACCESS

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biological Samples		
1 Kordofan giraffe blood sample	This paper	Table S2
48 giraffe blood or skin biopsy samples	[7,9,19]	Table S2
Chemicals, Peptides, and Recombinant Proteins		
Proteinase K	Sigma-Aldrich	Cat#P8044-1G
Roti-Phenol	Carl Roth	Cat#0038.3
Chloroform	VWR	Cat#22711.260
Critical Commercial Assays		
NucleoSpin Tissue kit	Macherey-Nagel	Cat#740952.50
Deposited Data		
Raw sequencing reads	This paper	NCBI BioProject: PRJNA635165
Kordofan giraffe genome assembly	This paper	GenBank: JABSTO000000000
49 giraffe mitochondrial genomes	This paper	GenBank: MT605012–MT605060
Masai giraffe genome assembly	[15–17]	https://www.dnazoo.org/
Kordofan giraffe sequencing reads	[7]	GenBank: ERR1248124
Masai giraffe sequencing reads	[15]	GenBank: SRR3218456
Okapi sequencing reads	[15]	GenBank: SRR3217625, SRR3217884
Giraffe mitochondrial genome	[41]	GenBank: JN632645
Okapi mitochondrial genome	[41]	GenBank: JN632674
Arabian camel protein-coding sequences	Ensembl	GenBank: GCA_000803125.2
Pig protein-coding sequences	Ensembl	GenBank: GCA_000003025.6
Sperm whale protein-coding sequences	Ensembl	GenBank: GCA_002837175.2
Bowhead whale protein-coding sequences	[42]	http://www.bowhead-whale.org
Lesser mouse-deer protein-coding sequences	[33]	http://animal.nwsuaf.edu.cn/code/index.php/RGD
Siberian musk deer protein-coding sequences	Ensembl	GenBank: GCA_004024705.2
Cow protein-coding sequences	Ensembl	GenBank: GCA_002263795.2
Goat protein-coding sequences	Ensembl	GenBank: GCA_001704415.1
Sheep protein-coding sequences	Ensembl	GenBank: GCA_000298735.1
Pronghorn protein-coding sequences	[33]	http://animal.nwsuaf.edu.cn/code/index.php/RGD
Okapi protein-coding sequences	[33]	http://animal.nwsuaf.edu.cn/code/index.php/RGD
Cow protein sequences	NCBI	GenBank: GCA_002263795.2
Software and Algorithms		
Supernova v.2.1.1	[43]	https://support.10xgenomics.com/de-novo-assembly/software/overview/latest/welcome; RRID:SCR_016756
QUAST-LG v.5.0.2	[44]	http://quast.sourceforge.net/quast-lg.html; RRID:SCR_001228
BUSCO v.3.0.3	[45]	https://busco-archive.ezlab.org/v3/; RRID:SCR_015008
JupiterPlot v.1.0	[46]	https://github.com/JustinChu/JupiterPlot/tree/1.0
RepeatMasker v.4.0.7	Institute for Systems Biology	http://www.repeatmasker.org/; RRID:SCR_012954
RepeatModeler v.1.0.11	Institute for Systems Biology	http://www.repeatmasker.org/RepeatModeler/; RRID:SCR_015027
BRAKER v.2.1.2	[47]	https://github.com/Gaius-Augustus/BRAKER
GenomeThreader v.1.7.1	[48]	http://genomethreader.org/
AUGUSTUS v.3.3.2	[49]	http://bioinf.uni-greifswald.de/augustus/; RRID:SCR_008417
proc10xG	UC Davis Bioinformatics Core	https://github.com/ucdavis-bioinformatics/proc10xG

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
FastQC v.0.11.7	Babraham Institute	https://www.bioinformatics.babraham.ac.uk/projects/fastqc/; RRID:SCR_014583
Trimmomatic v.0.38	[50]	http://www.usadellab.org/cms/index.php?page=trimmomatic; RRID:SCR_011848
BWA-MEM v.0.7.17	[51]	https://github.com/lh3/bwa; RRID:SCR_010910
SAMtools v.1.9	[52]	https://www.htslib.org/; RRID:SCR_002105
Picard v.2.18.21	Broad Institute	https://broadinstitute.github.io/picard/; RRID:SCR_006525
QualiMap v.2.2.2	[53]	http://qualimap.bioinfo.cipf.es/; RRID:SCR_001209
GATK v.3.8.1	[54]	https://github.com/broadgsa/gatk/releases/tag/3.8-1; RRID:SCR_001876
ANGSD v.0.929	[55]	http://www.popgen.dk/angsd/index.php/ANGSD
ngsLD v.1.0.0	[56]	https://github.com/fgvieira/ngsLD
PCAngsd v.0.97	[57]	http://www.popgen.dk/software/index.php/PCAngsd
R v.3.6.0	[58]	https://www.r-project.org/; RRID:SCR_001905
NGSadmix v.32	[59]	http://www.popgen.dk/software/index.php/NgsAdmix; RRID:SCR_003208
CLUMPAK	[60]	http://clumpak.tau.ac.il/
ngsDist v.1.0.5	[61]	https://github.com/fgvieira/ngsDist
FastME v.2.1.5	[62]	http://www.atgc-montpellier.fr/fastme/
IQ-TREE v.1.6.11	[63]	http://www.iqtree.org/
ggtree v.1.16.1	[64]	https://bioconductor.org/packages/release/bioc/html/ggtree.html RRID:SCR_018560
ASTRAL-III v.5.6.1	[65]	https://github.com/smirarab/ASTRAL
FigTree v.1.4.4	N/A	https://github.com/rambaut/figtree; RRID:SCR_008515
DiscoVista v.1.0	[66]	https://github.com/esayyari/DiscoVista
SplitsTree v.4.15.1	[67]	http://www.splitstree.org/; RRID:SCR_014734
Seqtk v.1.3-r106	N/A	https://github.com/lh3/seqtk
MitoZ v.2.3	[68]	https://github.com/linzhi2013/MitoZ/tree/master/version_2.3
MAFFT v.7.407	[69]	https://mafft.cbrc.jp/alignment/software/; RRID:SCR_011811
Proteinortho v.5.15	[70]	https://www.bioinf.uni-leipzig.de/Software/proteinortho/
Gblocks v.0.91b	[71]	http://molevol.cmima.csic.es/castresana/Gblocks.html; RRID:SCR_015945
PAML v.4.9	[72]	http://abacus.gene.ucl.ac.uk/software/paml.html; RRID:SCR_014932
MCMCtreeR v.1.1	[73]	https://github.com/PuttickMacroevolution/MCMCtreeR
PSMC v.0.6.5-r67	[25]	https://github.com/lh3/psmc; RRID:SCR_017229
BCFtools v.1.9	N/A	https://samtools.github.io/bcftools/; RRID:SCR_005227
RZooROH v.0.3.0	[74]	https://cran.r-project.org/package=RZooRoH

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Axel Janke (axel.janke@senckenberg.de).

Materials availability

This study did not generate new unique reagents.

Data and code availability

The raw sequencing reads generated during this study are available at NCBI Short Read Archive under the accession BioProject: PRJNA635165. The nucleotide sequence of the Kordofan giraffe genome assembled during this study is available at DDBJ/ENA/GenBank under the accession GenBank: JABSTO000000000. The nucleotide sequences of mitochondrial genomes assembled during this study are available at DDBJ/ENA/GenBank under the accessions GenBank: MT605012–MT605060.

CellPress
OPEN ACCESS

The annotation files pertaining to the Kordofan giraffe genome assembly and the dataset of 1,068 genome fragment alignments used in the phylogenomic inference are available at Zenodo (http://doi.org/10.5281/zenodo.4781356). The code used to process and analyze the data generated in this study is available at GitHub (https://github.com/rtfcoimbra/Coimbra-et-al-2021_CurrBiol).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

To perform phylogenomic and population analyses, we sequenced and *de novo* assembled a high quality reference genome of the critically endangered Kordofan giraffe[12] and re-sequenced the whole genomes of 48 giraffe individuals from previous studies.[7,9,19] These individuals equally represent all currently recognized (sub)species: the northern giraffe (n = 15; Kordofan n = 4, Nubian n = 6, West African n = 5), the reticulated giraffe (n = 10), the Masai giraffe *s. l.* (n = 11; Masai *s. str.* n = 5, Luangwa n = 6), and the southern giraffe (n = 12; Angolan n = 6, South African n = 6). The sampling method, research permits, ethical guidelines, and DNA extraction are described elsewhere.[9] Additionally, we retrieved publicly available paired-end read data of a Kordofan giraffe (GenBank: ERR1248124), a Masai giraffe *s. str.* (GenBank: SRR3218456), and an okapi (GenBank: SRR3217625, SRR3217884). For a map of sampling locations and details about samples see Figure 1A and Table S2.

METHOD DETAILS

Kordofan giraffe genome assembly and annotation

Genome sequencing

The genome of a male Kordofan giraffe from Zoo Planckendael (Belgium) was sequenced from a fresh blood sample using 10X Genomics' linked-reads technology.[13,14] High molecular weight DNA was purified by standard phenol-chloroform extraction, and size selected to remove DNA molecules < 40 kbp on a BluePippin (Sage Science). A Chromium Genome library (10X Genomics) was prepared and sequenced at BGI (Hong Kong). Sequencing was conducted in a single lane on an Illumina HiSeq X Ten (2 × 151 bp).

De novo assembly

A pseudohaplotype *de novo* genome assembly was produced with Supernova v.2.1.1[43] based on pre-filtered reads received from BGI. Assembly statistics were generated using QUAST-LG v.5.0.2[44] (Figures S1A–S1C). BUSCO v.3.0.3[45] was used to assess the completeness of the genome assembly using the mammal specific gene set (mammalia_odb9) and the '--long' option (Figure S1D). Genome assembly consistency plots compared to an available nearly chromosome-length genome assembly of the Masai giraffe (https://www.dnazoo.org/; Figure S1E) were generated with JupiterPlot v.1.0.[46]

Annotation of repetitive regions

The annotation of repeats in the Kordofan giraffe genome was done in three steps. First, we used RepeatMasker v.4.0.7 (http://www.repeatmasker.org/) to annotate and hard mask known Cetartiodactyla repeats from the RepBase library.[75] Then, we created a *de novo* repeat library for the Kordofan giraffe from the hard-masked genome assembly using RepeatModeler v.1.0.11 (http://www.repeatmasker.org/RepeatModeler/). Repetitive elements with unknown classification were removed from that library, and the remaining predicted repeats were annotated with a second run of RepeatMasker on the hard-masked assembly obtained in the first run. Genome feature files obtained in both RepeatMasker runs were combined. A summary of the repeat elements was generated with the script 'buildSummary.pl' included in RepeatMasker (Table S1).

Annotation of coding regions

Prior to gene prediction, we soft-masked the Kordofan giraffe genome assembly for the previously identified repetitive regions. We then used BRAKER v.2.1.2[47] in combination with GenomeThreader v.1.7.1[48] and AUGUSTUS v.3.3.2[49] for the automated prediction and annotation of protein-coding genes. Protein sequences derived from *Bos taurus* (GenBank: GCA_002263795.2) were used as extrinsic evidence for homolog-based predictions.

Whole-genome re-sequencing

Library preparation and whole-genome re-sequencing were carried out either at BGI (2 × 150 bp, 300 bp insert size, HiSeq 4000) or at Novogene (2 × 150 bp, 350 bp insert size, NovaSeq 6000). The exception was sample ENP11, which was sequenced from a Chromium Genome library prepared at BGI. In that case, the DNA size selection failed, resulting in molecules shorter than the required size for the linked-reads technology. However, we were able to use the paired-end reads for mapping after processing them with 'process_10xReads.py' from proc10xG (https://github.com/ucdavis-bioinformatics/proc10xG) using default options to remove barcodes, primers and filter out reads that do not match the program's barcode whitelist.

Quality control and read mapping

Read quality before and after trimming was assessed with FastQC v.0.11.7 (www.bioinformatics.babraham.ac.uk/projects/fastqc/). Reads were filtered to remove adapters and trimmed for quality using Trimmomatic v.0.38[50] with options 'ILLUMINACLIP:TruSeq3-PE-2.fa:2:30:10', 'SLIDINGWINDOW:4:20', and 'MINLEN:40'. Remaining read pairs were mapped against our Kordofan giraffe assembly with BWA-MEM v.0.7.17[51] and sorted with SAMtools v.1.9.[52] Hereafter, lane level BAMs were merged to sample level with SAMtools. PCR and optical duplicates were marked in read alignments with the 'MarkDuplicates' tool from Picard v.2.18.21 (http://broadinstitute.github.io/picard/) using the recommended optical duplicate pixel distance for patterned and unpatterned flow-cells

accordingly. The publicly available reads retrieved for the Masai giraffe and the okapi did not have proper headers, thus preventing optical duplicates detection. Mapping quality was assessed with Qualimap v.2.2.2.[53] We then performed realignment around indels with GATK v.3.8.1.[54] BAM files were cleaned with SAMtools to remove reads flagged as any of the following: unmapped, not primary alignment, supplementary alignment, failed quality checks, or PCR/optical duplicate; keeping only unique reads mapped in a proper pair. Reads mapped to scaffolds shorter than 1 Mbp (~5% of the total assembly length) were also excluded. Genomic regions identified as repeat elements by RepeatMasker were removed from all downstream analyses after merging repeats detected within 1 kbp from each other.

Genotype likelihoods

Genotype likelihoods at variant sites were calculated using the SAMtools model in ANGSD v.0.929.[55] Base alignment qualities (BAQs)[76] were computed, mapping qualities were adjusted for excessive mismatches (option '-C 50'), and reads with mapping quality < 30 and nucleotides with base quality < 30 were discarded. Minimum and maximum thresholds for the global site depth were set to $d \pm (5 \times MAD)$, where d is the median of the global site depth distribution, and MAD is the median absolute deviation. Sites with a p-value $< 1 \times 10^{-6}$ for strand bias, heterozygous bias, or Hardy-Weinberg equilibrium tests were also excluded. Only biallelic SNPs called with a p-value $< 1 \times 10^{-6}$, a minimum minor allele frequency of 0.05, and with data for at least 90% of the individuals were retained.

Pruning of linked sites

Pairwise linkage disequilibrium (LD) was estimated in ngsLD v.1.0.0[56] assuming a maximum distance of 1 Mbp between SNPs. To decide on thresholds for LD pruning, we randomly sampled 256,636 (0.1%) of all pairwise SNP combinations and used the script 'fit_LDdecay.R', included in ngsLD, to fit an LD decay model for r^2 values of SNP pairs up to 200 kbp apart assuming a bin size of 250. Linked sites were pruned with the script 'prune_graph.pl', also included in ngsLD, considering a maximum distance of 50 kbp between SNPs and a minimum r^2 of 0.1.

QUANTIFICATION AND STATISTICAL ANALYSIS

Population structure and admixture analyses

A covariance matrix was calculated from genotype likelihoods of LD-pruned SNPs with PCAngsd v.0.97[57] and used to perform a PCA with the 'prcomp' function in R v.3.6.0.[58] Admixture proportions of individuals were estimated for the same set of SNPs with NGSadmix v.32[59] assuming K values from 1 to 10 for which we ran 100 replicates per K. Statistical dispersion of run likelihoods per K were depicted in a boxplot in addition to the ΔK method,[21] which was performed on the CLUMPAK webserver.[60] The highest likelihood run for each $K \geq 2$ was shown as an admixture bar plot.

Genetic distances and neighbor-joining tree

We calculated pairwise genetic distance matrices for the whole genome plus 1,000 bootstrap replicates with replacement blocks of 500 SNPs in ngsDist v.1.0.5.[61] Genotype likelihoods estimated in ANGSD were used as input with filter settings as described previously. The okapi was included as an outgroup. Neighbor-joining trees were estimated in FastME v.2.1.5[62] using the BioNJ algorithm[77] with subtree pruning and regrafting (SPR) tree topology improvement. Branch support values based on the bootstrapped matrices were assigned to the main tree with IQ-TREE v.1.6.11.[63] The resulting tree was plotted with ggtree v.1.16.1.[64]

Nuclear phylogenomic inference

We generated consensus sequences for each sample using the option '-doFasta 4' in ANGSD. Sites with mapping or base qualities < 30, minimum depth < 4, maximum depth above the 95th percentile of the sample's depth distribution, or with a total depth ratio for IUPAC assignment < 0.33 were masked. After removing repetitive regions from the consensus sequences, samples that still consisted of > 20% of masked bases due to poor cumulative genome coverage distribution (all LVNP individuals and ISC01) were excluded from further analyses. The remaining sequences were aligned by scaffold and sites with a ratio of masked bases > 0.2 were discarded. Phylogenomic inference proceeded in a maximum likelihood framework using the GF approach described in[28] with the okapi as an outgroup. We determined the appropriate GF size to statistically reject alternative topologies based on an AU test[22] (Figure S2A). For computational feasibility, we randomly sampled two individuals per (sub)species and limited the analysis to the 15 possible topologies at the four species level (Figure S2B). The AU test was performed with 10,000 replicates in IQ-TREE for fragment sizes of up to 600 kbp in steps of 50 kbp, with 200 randomly sampled GFs per fragment size. Consensus genome sequences of all samples that passed the initial filtering were aligned, and the alignments were split into non-overlapping GFs of the appropriate size removing sites as described above. GFs shorter than the selected size were discarded. Phylogenetic trees were inferred for each GF in IQ-TREE using 1,000 ultrafast bootstrap replicates[78] and the automatic ultrafast model selection.[79] From the resulting GF trees, we constructed an MSC tree in ASTRAL-III v.5.6.1[65] reporting local posterior probabilities.[80] The MSC tree was rooted with the okapi and visualized with FigTree v.1.4.4 (http://tree.bio.ed.ac.uk/software/figtree/). Discordance among GF trees was assessed through quartet frequencies in DiscoVista v.1.0.[66] A consensus network for the GF trees was generated in SplitsTree v.4.15.1[67] using different median thresholds (5%–20%).

CellPress
OPEN ACCESS

Assembly and phylogeny of mitochondrial genomes

Raw sequencing reads of 49 giraffe individuals, excluding ENP11 due to its library type, were randomly subsampled to 10%–20% of all read pairs with Seqtk v.1.3-r106 (https://github.com/lh3/seqtk). Their mitogenomes were assembled *de novo* using the module 'all' from MitoZ v.2.3[68] in quick mode. We also retrieved publicly available mitogenome assemblies for a giraffe (GenBank: JN632645) and an okapi (GenBank: JN632674). Giraffe mitogenomes were aligned with MAFFT v.7.407[69] and the alignments were visually checked. For seven individuals, a cytosine repeat (position 16,291–16,301 of the reference JN632645) in the control region could not be resolved due to a ± 1 bp variation, thus we assumed the reference state. Sequences of all 13 mitochondrial protein-coding genes were extracted from the assemblies and aligned to those of the okapi, which was used as an outgroup for phylogenetic inference. A maximum likelihood tree for the partitioned[81] protein-coding gene alignment was constructed in IQ-TREE using 1,000 ultrafast bootstrap replicates and the ultrafast model selection between codon models, assuming the vertebrate mitochondrial genetic code. The tree was plotted with the ggtree.

Divergence time estimates

Proteinortho v.5.15[70] identified orthologous protein-coding sequences for the Kordofan giraffe and 11 published Cetartiodactyla species (Table S3). Orthologs found in at least 10 species, including the Kordofan giraffe, were kept for further analysis. The corresponding orthologs for additional giraffe (sub)species were extracted using the genomic coordinates of the Kordofan giraffe genome assembly. The coding sequences were aligned using MAFFT with default parameters. Gaps were removed using Gblocks v.0.91b[71] and only coding sequence alignments > 300 bp were selected for further analysis. Trimmed alignments were concatenated and analyzed with MCMCTree in PAML v.4.9[72] to estimate divergence times among Cetartiodactyla species using four calibration points (Table S4). The analysis was run for a sample size of 20,000, a burn-in of 2,000, and tree sampling every 10 iterations using default parameters. The dated tree was plotted with MCMCtreeR v.1.1.[73]

Demographic reconstruction

Changes in N_e through time were assessed with the PSMC model v.0.6.5-r67.[25] Consensus genome sequences in FASTQ format were generated for three individuals with the best sequencing coverage of each giraffe (sub)species using the 'mpileup' command in BCFtools v.1.9 (https://samtools.github.io/bcftools/) and the included script 'vcfutils.pl'. For the reticulated giraffe, we prioritized wild individuals. Luangwa giraffe individuals were excluded due to a high proportion of undetermined bases in the consensus sequence resulting from poor cumulative genome coverage. Sites with read depth below 10 or above twice the sample's median depth and sites with a root-mean-squared mapping quality or a consensus base quality < 30 were removed. PSMC was run for 25 iterations with a maximum $2N_0$-scaled coalescent time of 15, an initial θ/ρ ratio of 5, and 64 atomic time intervals (4 + 25 × 2 + 4 + 6). We performed 100 bootstrap replicates by randomly sampling with replacement 1 Mbp blocks from the consensus sequence. Results were scaled by a mutation rate of 2.12×10^{-8} substitutions per site per generation estimated for the giraffe[33] and a generation time of 10 years.[82]

Heterozygosity

We calculated heterozygosity as the proportion of sites where an individual was heterozygous based on the per-sample folded site frequency spectrum (SFS). First, we estimated the site allele frequency at every site for each giraffe individual in ANGSD using the options '-doSaf 1 -fold 1' and the reference genome as ancestral. BAQs were computed, mapping qualities were adjusted for excessive mismatches (option '-C 50'), and reads with mapping quality < 30 and sites with base quality < 30 or depth above the 95th percentile of the sample's depth distribution were discarded. Then, we used ANGSD's subprogram realSFS to generate the per-sample folded SFS, with 200 bootstrap replicates, from which we calculated the proportion of heterozygous sites.

ROH and Inbreeding

To investigate more recent demographic history and the level of inbreeding among different populations of giraffe, we analyzed genome-wide ROH among the 50 re-sequenced individuals. First, SNPs were called in ANGSD as described above. The obtained SNP sites were then passed to BCFtools 'mpileup' with the option '--targets-file'. BCFtools 'view' and 'filter' were used to retain only biallelic SNPs called in at least 90% of the individuals with QUAL > 30, MQ > 30, and the same depth thresholds used in ANGSD. No LD pruning was performed prior to ROH detection. Realized inbreeding coefficients (F_{ROH}) for each individual were calculated from a set of 737,246 SNPs using the R package RZooRoH v.0.3.0.[74] This package implements a hidden Markov model that identifies segments of homozygosity-by-descent (HBD; evident from ROH) and non-HBD segments.[74] As ROH and the lengths of such runs are informative of past inbreeding events, RZooRoH categorizes the lengths of HBD segments into approximate generation classes. Longer ROH correspond to more recent inbreeding events while shorter ROH correspond to more ancestral inbreeding events. As the density of our filtered SNP panel allowed us to detect both long and short HBD segments, we set our model to six HBD classes (k), spanning a large range of HDB segment sizes. We created a predefined 'zoomodel' with R_k equal to 4, 16, 64, 320, 1280, and 5120 for HBD segments and > 5120 for non-HBD segments. R_k values are approximately double the number of generations from the time of inbreeding, such that our R_k values correspond approximately to ancestors inbreeding 2, 8, 32, 160, 640, and 2560 generations ago. However, we note that the 5120 class is representative of past N_e, rather than individual inbreeding levels.