Current Biology, Volume *26*

Supplemental Information

Multi-locus Analyses Reveal Four Giraffe Species Instead of One

Julian Fennessy, Tobias Bidon, Friederike Reuss, Vikas Kumar, Paul Elkan, Maria A. Nilsson, Melita Vamberger, Uwe Fritz, and Axel Janke

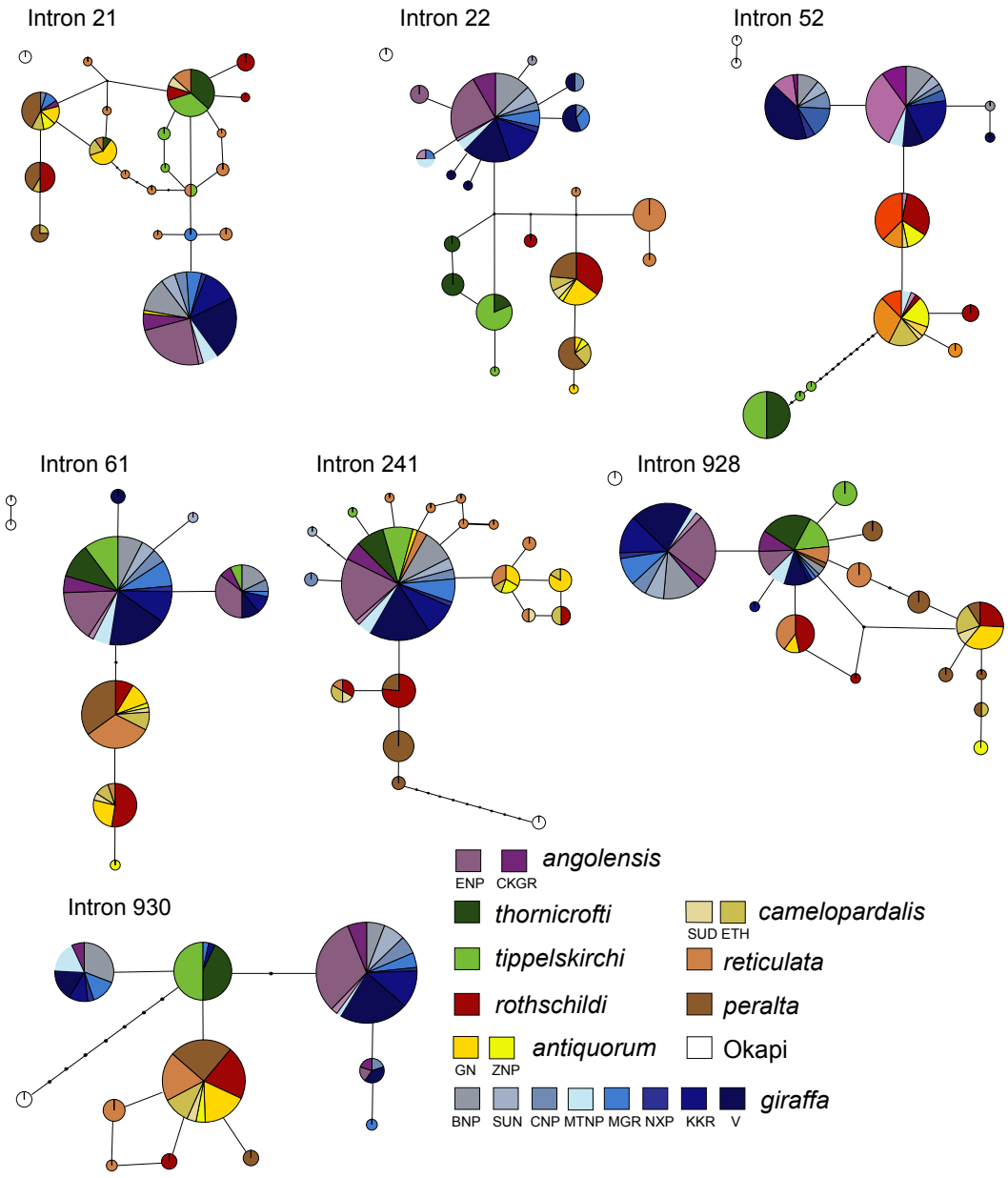

Figure S1. Haplotype networks of seven intron sequences. Related to Figure 2 and 3. The networks for 105 giraffe show that numerous single alleles are shared and that for most loci subspecies cannot be easily distinguished. Notable exceptions are intron 52 and 930 that are exclusive for Masai (*tippelskirchi*, including the formerly recognized Thornicroft's giraffe) and 241 that is nearly exclusive for the West African giraffe (*peralta*). Furthermore, a southern clade (*angolensis* plus *giraffa)* and northern clade (*antiquorum*, *peralta*, *rothschildi*) are prominent for most loci. The lack of further resolution is possibly a consequence of insufficient data and gene flow remains uncertain.

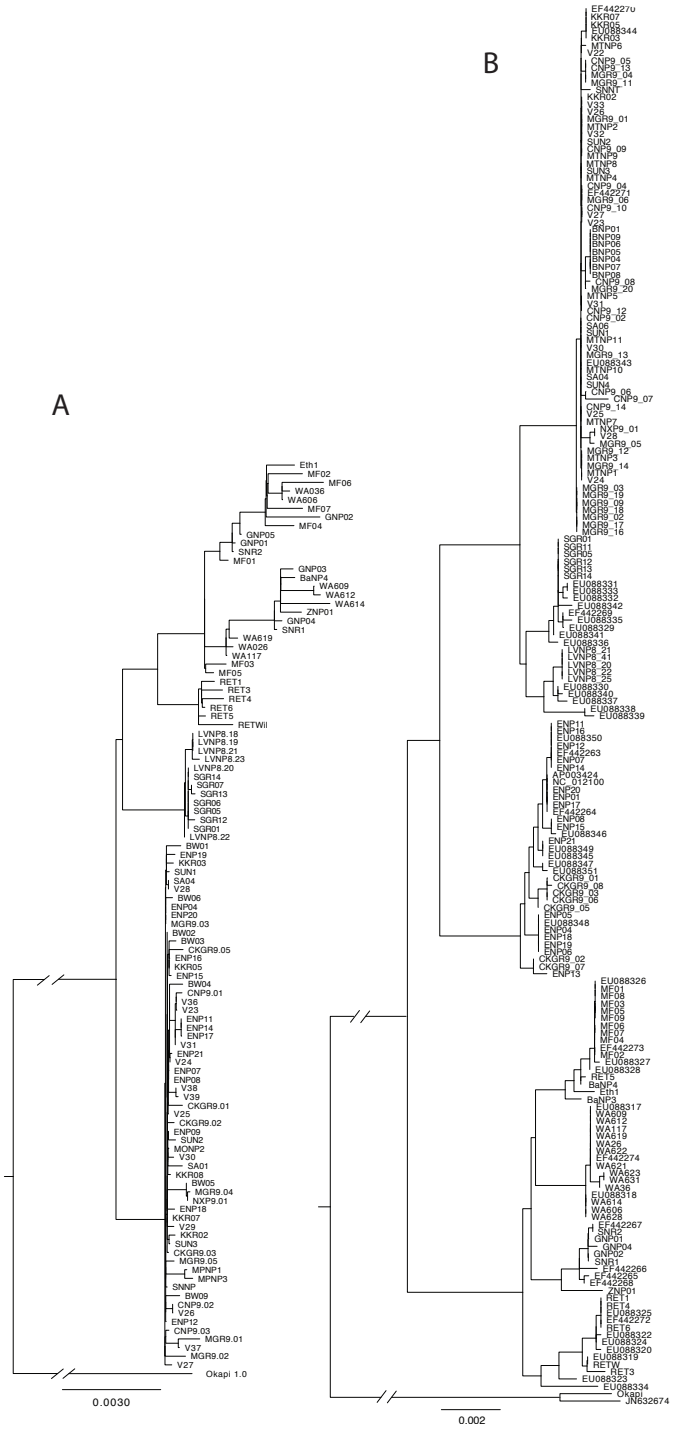

Figure S2. Evolutionary trees with details on the individual IDs. Related to Figure 2. A) ASTRAL tree from individual nuclear loci with ML branch lengths. While analysis of concatenated sequences is problematic [S1] a ML tree based on concatenated sequences, which shows the West African Giraffe separate can be found in doi:10.5061/dryad.h3tc2. B) BEAST mtDNA tree with details for accession numbers and individual IDs and their location. Note – the okapi branch (root) is not to scale in both figures to allow for better resolution among giraffe branches. Genbank accession numbers for published data are shown and are detailed in [S2] and individual ID and sample location can be found in doi:10.5061/dryad.h3tc2.

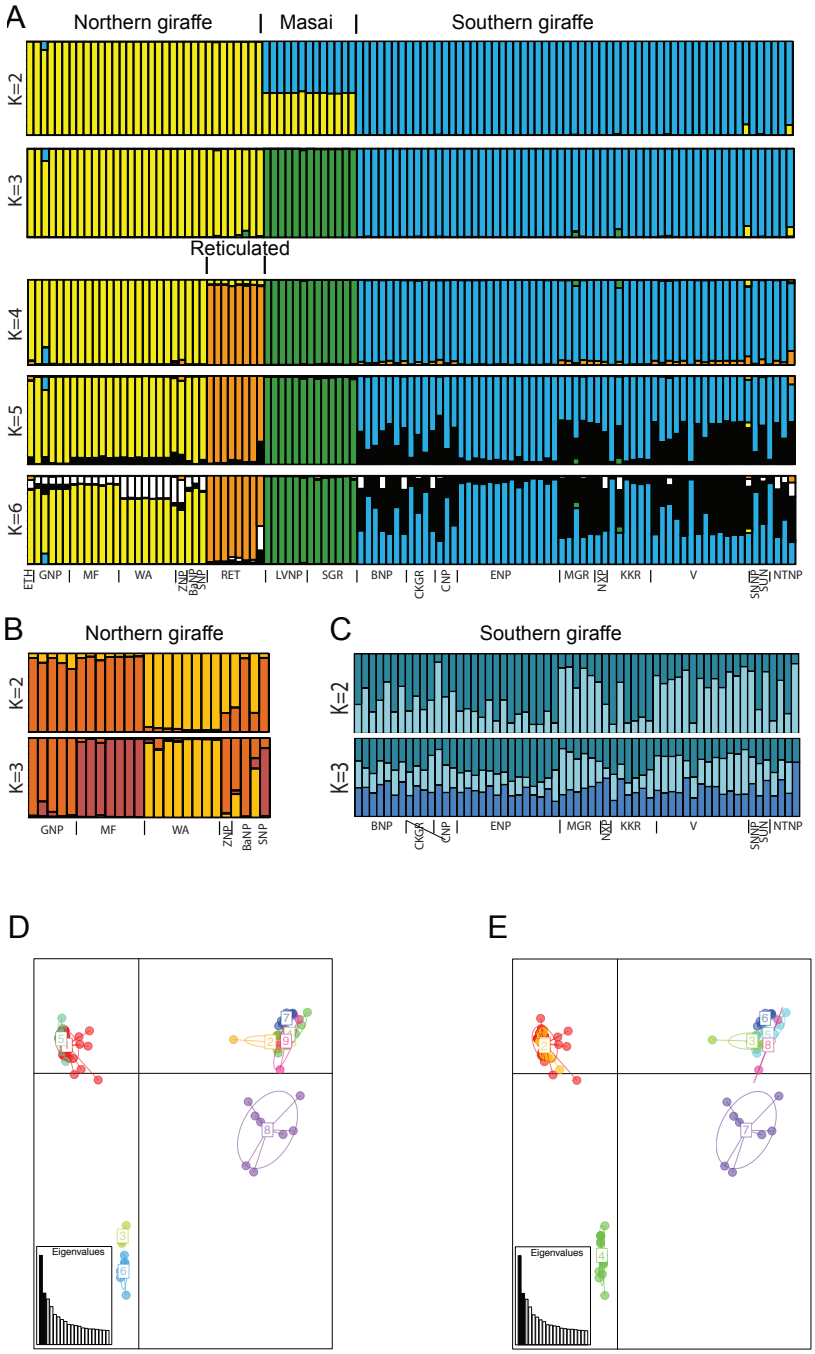

Figure S3. Additional Structure and PCA analyses. Related to Figure 3.
A) Structure analysis for all subspecies for K=2 to K=6. K=4 has the highest delta K and from K=5 increasing admixtures is evident.
B) A separate Structure analysis for the subspecies of the northern giraffe reveals evidence for additional cluster of West African (WA) and Nubian (former Rothschild's, MF) giraffe. However, in this data set haplotypes sharing with other subspecies is evident and these are not distinct in other analyses.
C) Southern giraffe do not show additional clustering when analyzed separately. This is in contrast the the clear separation of the subspecies by mtDNA sequences.
Abbreviations for the geographic origin are explained in Table S1.
D) PCAs of giraffe haplotypes with grouping according to traditional nine subspecies classification 1 - *G. c. angolensis,* 2 - *G. c. antiquorum,* 3 – *G. c. thornicrofti,* 4 - *G. c.*

rothschildi, 5 - *G. c. giraffa,* 6 - *G. c. tippelskirchi*, 7 - *G. c. peralta*, 8 - *G. c. reticulata*, 9 - *G. c. camelopardalis)*.
PCAs along Axis 1-3 (not shown) produces nearly identical results. In this analyses four distinct clades are evident and these correspond to the four giraffe species suggested by other analyses, other subspecies are overlapping.
E) PCAs of giraffe haplotypes with grouping according to mtDNA differentiation 1- *G. c. giraffa,* 2 - *G. c. angolensis*, 3 - *G. c. antiquorum*, 4 - *G. c. tippelskirchi*, 5 - *G. c. rothschildi*, 6 - *G. c. peralta*, 7 - *G. c. reticulata*, 8 - *G. c. camelopardalis*). In this analyses four distinct clades are evident and these correspond to the four giraffe species suggested by other analyses, other subspecies are overlapping.

Table S1. Origin, abbreviation, number of individuals (n) and traditional subspecies designation of analyzed giraffe sequences. Related to Table 1, Figure 1 and Figure 2.

Geographical origin	**Abbreviation**	**n**	**Previous subspecies designation**	**MtDNA subspecies (this study)**
Badingilo National Park, South Sudan	BaNP	2	*camelopardalis*	*camelopardalis*
Bwabwata National Park, Namibia	BNP	7	*angolensis*	*giraffa*
Chobe National Park, Botswana	CNP	11	*angolensis*	*giraffa*
Gambella National, Ethiopia	ETH	1	*camelopardalis*	*camelopardalis*
Etosha National Park, Namibia	ENP	17	*angolensis*	*angolensis*
Garamba National Park, DRC	GNP	3	*antiquorum*	*antiquorum*
Khamab Kalahari Reserve, South Africa	KKR	6	*giraffa*	*giraffa*
Luangwa Valley National Park, Zambia	LVNP	5	*thornicrofti*	*tippelskirchi*
Moremi Game Reserve, Botswana	MGR	16	*angolensis*	*giraffa*
Mosi-oa-Tunya National Park, Zambia	MTNP	11	*angolensis*	*giraffa*
Murchison Falls National Park, Uganda	MF	9	*rothschildi*	*camelopardalis*
Nxai Pans, Botswana	NXP	1	*angolensis*	*giraffa*
Nuernberg/Stuttart Zoo	RET	6	*reticulata*	*reticulata*
Selous Game Reserve, Tanzania	SGR	6	*tippelskirchi*	*tippelskirchi*
Shambe National Park, South Sudan	SNR	2	*camelopardalis*	*antiquorum*
Sioma Ngwezi NP, Zambia	SNNP	1	*angolensis*	*giraffa*
Sun hotel, Livingstone, Zambia	SUN	4	*giraffa*	*giraffa*
Koure, Niger	WA	13	*peralta*	*peralta*
Vumbura Concession, Botswana	V	11	*angolensis*	*giraffa*
Zakouma National Park, Chad	ZNP	1	*antiquorum*	*antiquorum*

Table S2. Detailed divergences time estimates. Related to Figure 3.
A: Estimated divergence times and confidence intervals.

Divergence	Estimated divergence time (Ma)
Southern giraffe – (Masai giraffe, (reticulated giraffe, northern giraffe)	2.00 (1.23-3.12)
Masai giraffe – (reticulated giraffe, northern giraffe)	1.87 (1.16-2.90)
Reticulated giraffe – northern giraffe	1.25 (0.74-1.97)

Note – divergence times were estimated based on a divergence of okapi to 11.5 million years ago (Ma) using BEAST and nuclear loci. For more information see: doi:10.5061/dryad.h3tc2.

B: Estimated divergence times and confidence intervals.

Species pair	Estimated divergence time
Whale-Dolphin	34.85 (34.0-36.0)
Sheep- Antelope	10.85 (7.1-14.6)
(Sheep,Antelope)-Cow	20.0 (18.0-22.0)
(Bovidae)-Giraffe	28.2 (22.8-35.2)
(Ruminantia,(Cetacea))-Pig	61.91 (54.2-66.1)
Cetartiodactyla-Dog	83.04 (67.8-98.1)
(Cetartiodactyla,Dog)-Human	90.85 (74.7-103.1)

Note – on four fossil based, independent divergences times included: whale-dolphin divergence at 34-36 Ma [S3], cattle-antelope 18-22 Ma [S4], divergence time of pig – remaining Artiodactyla at 52.4-65.8 Ma [S5], and Carnivora (dog)-Artiodactyla 62.5-131.0 [S5]. The analysis was for a sample size of 200,000, burn-in of 20,000 and sampling trees from every second iteration.

Supplemental Experimental Procedures

Coding sequences were extracted from scaffolds >10,00 bp. For annotation the scaffolds were repeat-masked using Repeat masker [S6]. From these scaffolds, genes were predicted by AUGUSTUS [S7] and coding sequences were extracted using the perl script provided with AUGUSTUS. For phylogenetic analysis orthology searches were made using using the recursive Blast method [S8] from nine genomes: human (*Homo sapiens)*, dog (*Canis familiaris*), pig (*Sus scrofa*), cattle (*Bos taurus*), Tibetan antelope (*Pantholops hodgsonii*), sheep (*Ovis aries*), bottlenose dolphin (*Tursiops truncates*), and bowhead whale (*Balaena mysticetus*) available from the Ensembl database (http://www.ensembl.org/info/data/ftp/index.html) and from the bow head whale (*Balaena mysticetus*) genome (http://alfred.liv.ac.uk/downloads/bowhead_whale/bowhead_whale_coding_sequences.zip) and Tibetan antelope (*Pantholops hodgsonii*) from ftp://climb.genomics.cn/pub/10.5524/100001_101000/100027/. Only the orthologous groups with data from seven species, including giraffe, were selected for further phylogenetic analysis. The selected orthologous sequences were translated using custom perl script and aligned using MAFFT [9]. PAL2NAL [S10] generated a nucleotide alignment from the aligned amino acids sequences and alignment-gaps were removed by the program GBLOCKS [S11]. Alignments with more than 25% observed nucleotide distance among any species were removed to avoid artifacts from multiple substitutions and unrecognized alignment artifacts. Only coding sequences larger than 180 bp were

selected for further phylogenetic analyses. The best evolutionary model was predicted by JMODELTEST. RAxML version 8.2.4 [S12] was used to reconstruct a maximum likelihood (ML) tree using the GTR+G+I model.

The giraffe divergence time was estimated using MCMC tree in PAML version 4 [S13]. The molecular clock was calibrated on four fossil based, independent divergences times: whale-dolphin divergence at 34-36 Mya [S3], cattle-antelope 18-22 Mya [S4], divergence time of pig – remaining Artiodactyla at 52.4-65.8 Mya [S5], and Carnivora (dog)-Artiodactyla 62.5-131.0 [S5]. The analysis was for a sample size of 200,000, burn-in of 20,000 and sampling trees from every second iteration.

Analysis on 18 giraffe individuals covering all the major subspecies of giraffe were used to estimate the divergence time using BEAST on seven nuclear loci. Corresponding Okapi nuclear loci were Sanger sequenced and ortholog nuclear loci of cow (*Bos taurus*) genome were fetched from UCSC https://genome.ucsc.edu/ database. All the sequences were aligned using MAFFT [9]. Later BEAST was run with the settings of 100 million generations, HKY+I+G model, log normal relaxed clock and tree prior of Yule process. We used the molecular calibration point of 11.5 Ma [14] with standard deviation of 0.5 for the okapi and giraffe split with cow as the outgroup. Convergence was checked and confirmed with Tracer [S15].

Additional analysis using the Bayesian program, BPP version 3.2 was done to test the delimitation of different species in giraffe [S16 –18]. First method (algorithm) A00 was run to estimate the alpha and beta parameters. Slight deviations from alpha=2 and beta=2000 for θ and τ turned out to be non-crucial to the analyses. Other parameters were left to default. We used the method (algorithm) A11 to search various species delimitation models and different species phylogeny. The sequences from seven nucleotide loci were clustered into four groups of southern, Masai giraffe, Reticulate and northern giraffe according to the multi-locus coalescent tree and clades suggested by Structure and PCA analyses. In addition, the probability of five species (*G. c. peralta* being separate) the classic and mtDNA grouping was calculated. Each analysis was run with 1,000,000 generations and a burn-in of 10,000 with gamma (G) priors of $\theta \approx (2, 2000)$ and $\tau \approx (2, 2000)$. Convergence was checked by repeated analysis and with different guide trees.

The search for the microsatellites using SCIROKO version 3.4 [S19] identified useful microsatellites for future studies for Giraffe is shown in doi:10.5061/dryad.h3tc2. In addition, the repeat masking of the selected Scaffolds >5kb after removal of duplicates, identified different types of repeat elements similar to other ruminants such as mouse deer, sheep, Tibetan antelope and cow shown in doi:10.5061/dryad.h3tc2.

***De novo* assembly statistics for the 10X coverage giraffe genome assembly from 125bp paired-end Illumina reads.**

Total number of raw reads	549,679,536
Quality trimmed reads	464,882,128
Total size include N after assembly	2,432,441,945
Total size without N after assembly	2,397,009,050
Total no. of scaffold	30,24,215
Mean Size	804
Median Size	244
Longest Seq	35,995
Shortest Seq	100
Scaffolds>=10kb	5,042
N50	2,201

Note – additional information is provided in doi:10.5061/dryad.h3tc2.

List of Primer sequences and PCR conditions for amplification of nuclear introns and mtDNA in giraffe.

Name & locus	Sequence 5'-3'	PCR conditions
Intron 21 RASSF4, 9th intron	for: CAGTGTCCATCACACAAC rev: GCACCGGCATTTCAAACTTA	TD-PCR (T_a=65-55°C; 10 cycles), standard PCR (T_a=55°C; 30 cycles)
Intron 22 ACP5, 6th intron	for: CAGCAGCCAAGGAGGACTAC rev: ATCTCCTTGGGGCTGATCTC	TD-PCR (T_a=67-57°C); 10 cycles), standard PCR (T_a=57°C; 30 cycles)
Intron 52 UBN2, 4th intron	for: ACTGGCACTCTCCAGTTTCG rev: CTTCCTCTTTCCGCTTCCTC	INT_{UBN2} Bock et al. 2014b
Intron 52_140825 UBN2, 4th intron	for: GACAACCAAAAGCACAAACC rev: CACTTACCCCAGTTGTTTGG	TD-PCR (T_a=69-62°C; 14 cycles), standard PCR (T_a=62°C; 26cycles)
Intron 61 CWF19L1, 9th intron	for: GCTGGGAGGAAGGTAGCAATG rev: AATGTTGACCACCAAATGC	TD-PCR (T_a=66-59°C); 14 cycles), standard PCR (T_a= 59°C; 26 cycles)
Intron 241 NUP155, 23rd intron	for: GCTGCTGTTGATGGCATTAG rev: GGTCCACCTGATTGCTGATT	See Intron 61
Intron 928 OTOF, 12th intron	for: GCAGAGCACCAGTTCCA rev: GCTCGGTAGATCTTCACGTAG	INT_{OTOF} Bock et al. 2014b
Intron 930 SOS1, 11th intron	for: CAAAGTCCAAAGCACCCTG rev: CATGTTACTTCCTCCTTGCTTG	TD-PCR (T_a=67-60°C; 14 cycles), standard PCR (T_a=60°C; 26 cycles)
Control Region, mtDNA	F: TACACTGGTCTTGTAAGC R: TCGCTTTGGTGTTTAAGC	Bock et al. 2014b
Cytochrome b, mtDNA	F: GAAAAACCATCGTTGTCG R: TGGGAGTATATTAATAGC	Bock et al. 2014b

Note – for: forward primer. rev: reverse primer. TD-PCR: touchdown PCR. T_a: primer annealing temperature. The locus is the gene name of the human ortholog and the respective intron.

List of Fst values for seven nuclear loci of four giraffe species.

	Southern	Northern	Masai	Reticulated
Southern	0			
Northern	0.559**	0		
Masai	0.591**	0.522**	0	
Reticulated	0.608**	0.273**	0.595**	0

Note – Double asterisks indicate all Fst values are significant at $p<0.05$. Southern giraffe (*G. giraffa*) comprises the historic Angolan (*G. c. angolensis*) and South African giraffe (*G. c. giraffa*). Northern giraffe (*G. camelopardalis*) includes the historic Nubian giraffe (*G. c. camelopardalis*), Rothschild's giraffe (*G. c. rothschildi)*, Kordofan giraffe (*G. c. antiquorum*) and West African giraffe (*G. c. peralta*). Masai giraffe (*G. tippelskirchi*) includes historic Thornicroft's giraffe (*G. c. thornicrofti*). Reticulated giraffe (*G. reticulata*) includes only itself.

References

S1. Edwards SV, Xi Z, Janke A, Faircloth BC, McCormack JE, Glenn TC, Zhong B, Wu S, Lemmon EM, Lemmon AR, et al. (2016). Implementing and testing the multispecies coalescent model: A valuable paradigm for phylogenomics. Mol. Phylogenet. and Evol.. *231*:447-62.

S2. Bock, F., Fennessy, J., Bidon, T., Tutchings, A., Marais, A., Deacon, F., and Janke, A. (2014). Mitochondrial sequences reveal a clear separation between Angolan and South

African giraffe along a cryptic rift valley. BMC Evol. Biol. *14*. 219
S3. de Muizon, C. (2009). L'origine et l'histoire évolutive des Cétacés. Comptes Rendus Palevol 8, 295–309.
S4. Vrba, E., and Schaller, G. (2000). Antelopes, deer, and relatives. fossil record, behavioral ecology, systematics and conservation (Yale University Press, New Haven, London).
S5. Benton, M., Benton, P., and Asher, R. (2009). Calibrating and constraining molecular clocks. In The Timetree of Life, Hedges SB, Kumar S. ed. (Oxford University Press).
S6. Smit, A., Hubley, R., and Green, P. (2015). RepeatMakser Open-4.0.2013-2015. Available from http://www.repeatmasker.org.
S7. Stanke, M., Schöffmann, O., Morgenstern, B., and Waack, S. (2006). Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. BMC Bioinformatics *7*, 62.
S8. Hallström, B.M., and Janke, A. (2008). Resolution among major placental mammal interordinal relationships with genome data imply that speciation influenced their earliest radiations. BMC Evol. Biol. *8*, 162.
S9. Katoh, K., Misawa, K., Kuma, K., and Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. *30*, 3059–3066.
S10. Suyama, M., Torrents, D., and Bork, P. (2006). PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res. *34*, W609-612.
S11. 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.
S12. Stamatakis, A. (2014). RAxML Version 8: A tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies. Bioinformatics *30*,1312-1313.
S13. Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. *24*, 1586–1591.
S14. Agaba M, Ishengoma E, Miller WC, McGrath BC, et al. (2016). Giraffe genome sequence reveals clues to its unique morphology and physiology. Nat. Commun. 7, 11519.
S15. Rambaut, A., Suchard, M., Xie, D., and Drummond, A. (2014). Tracer v1.6, Available from http://beast.bio.ed.ac.uk/Tracer.
S16. Rannala B, Yang Z. (2003). Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. Genetics *164*, 1645–1656.
S17. Yang Z, Rannala B. (2010). Bayesian species delimitation using multilocus sequence data. Proc. Natl. Acad. Sci. U. S. A. *107*, 9264–9269.
S18. Yang Z. (2015). The BPP program for species tree estimation and species delimitation. Curr. Zool. *61*, 854-865.
S19. Kofler, R., Schlötterer, C., and Lelley, T. (2007). SciRoKo: a new tool for whole genome microsatellite search and investigation. Bioinformatics *23*, 1683–1685.