Journal of Heredity, 2022, **113**, 568–576
https://doi.org/10.1093/jhered/esac033
Advance access publication 5 July 2022
Genome Resources

American Genetic Association

OXFORD

Genome Resources

A Chromosome-Scale Genome Assembly of the Okapi (*Okapia Johnstoni*)

Sven Winter, Raphael T. F. Coimbra, Philippe Helsen, and Axel Janke

From the Senckenberg Biodiversity and Climate Research Centre, Senckenberganlage 25, 60325 Frankfurt am Main, Germany (Winter, Coimbra, and Janke); Research Institute of Wildlife Ecology, Vetmeduni Vienna, Savoyenstraße 1, 1160 Vienna, Austria (Winter); LOEWE Centre for Translational Biodiversity Genomics (LOEWE-TBG), Senckenberganlage 25, 60325 Frankfurt am Main, Germany (Janke); Institute for Ecology, Evolution and Diversity, Goethe University, Max-von-Laue-Straße 13, 60438 Frankfurt am Main, Germany (Coimbra and Janke); and Centre for Research and Conservation, Royal Zoological Society of Antwerp, Koningin Astridplein 20, 2018 Antwerp, Belgium (Helsen).

Address correspondence to S. Winter at the address above, or e-mail: sven.winter@senckenberg.de.

Corresponding Editor: Klaus-Peter Koepfli

Abstract

The okapi (*Okapia johnstoni*), or forest giraffe, is the only species in its genus and the only extant sister group of the giraffe within the family Giraffidae. The species is one of the remaining large vertebrates surrounded by mystery because of its elusive behavior as well as the armed conflicts in the region where it occurs, making it difficult to study. Deforestation puts the okapi under constant anthropogenic pressure, and it is currently listed as "Endangered" on the IUCN Red List. Here, we present the first annotated de novo okapi genome assembly based on PacBio continuous long reads, polished with short reads, and anchored into chromosome-scale scaffolds using Hi-C proximity ligation sequencing. The final assembly (*TBG_Okapi_asm_v1*) has a length of 2.39 Gbp, of which 98% are represented by 28 scaffolds > 3.9 Mbp. The contig N50 of 61 Mbp and scaffold N50 of 102 Mbp, together with a BUSCO score of 94.7%, and 23 412 annotated genes, underline the high quality of the assembly. This chromosome-scale genome assembly is a valuable resource for future conservation of the species and comparative genomic studies among the giraffids and other ruminants.

Key words: 10X Chromium, Giraffidae, Hi-C, long-reads, PacBio

The okapi (*Okapia johnstoni*) (Figure 1A, also known as forest giraffe, is a monotypic species in the genus *Okapia* and the only extant sister group to the giraffe (*Giraffa* spp.) within the family Giraffidae. Today, okapi only occurs in central, northern, and eastern regions of the Democratic Republic of the Congo and occasionally crosses the border to Uganda. The species is classified as "Endangered" on the IUCN Red List based on an estimated decline to one-half of its original population size during the last 24 years, which equals only about three generations (Mallon et al. 2015). This elusive species is difficult to observe in its natural habitat, dense rainforest, and most population estimates rely on dung-based surveys, as ongoing political conflicts, illegal mining, and poaching hinder more accurate population counts.

Genomic data for the okapi is still limited to being sequenced from non-invasive sampling strategies (e.g., dung samples) and single molecular markers, such as a few mitochondrial and nuclear genes. Analyses of these verify the presence of okapis south-west of the Congo River (Stanton et al. 2016) and characterize the genetic diversity across their geographic distribution (Stanton et al. 2014). There are two draft genome assemblies available for the okapi (Agaba et al. 2016; Dudchenko et al. 2017; Chen et al. 2019). However, both assemblies were constructed using short-read sequencing libraries only and are thus highly fragmented.

The development and increasing accessibility of long-read sequencing technologies and long-range scaffolding methods has made chromosome-scale genome assemblies for many animal and plant species possible (Bickhart et al. 2017; Marrano et al. 2020; Winter et al. 2020; Rhie et al. 2021), including the okapi's closest relative, the giraffe (Farré et al. 2019; Liu et al. 2021). The availability of these high-contiguity genomes enables complex comparative genomic studies and the understanding of the genetic factors underpinning unique evolutionary traits, such as hypertension, which is the norm in the exceptionally tall animal (Liu et al. 2021). Furthermore, high-quality genome assemblies often serve as reference sequences to study species for which mapping of short-read sequences (e.g., from museum collections or feces) is the only source of genetic information.

Here, we report the first high-quality de novo genome assembly of the okapi based on PacBio continuous long reads, polished with short reads. Furthermore, we use existing chromatin interaction map (Hi-C) data (Dudchenko et al. 2017) to construct a chromosome-scale genome assembly as reference genome to study okapi genetic diversity, genome evolution,

Received March 16, 2022; Accepted June 30, 2022

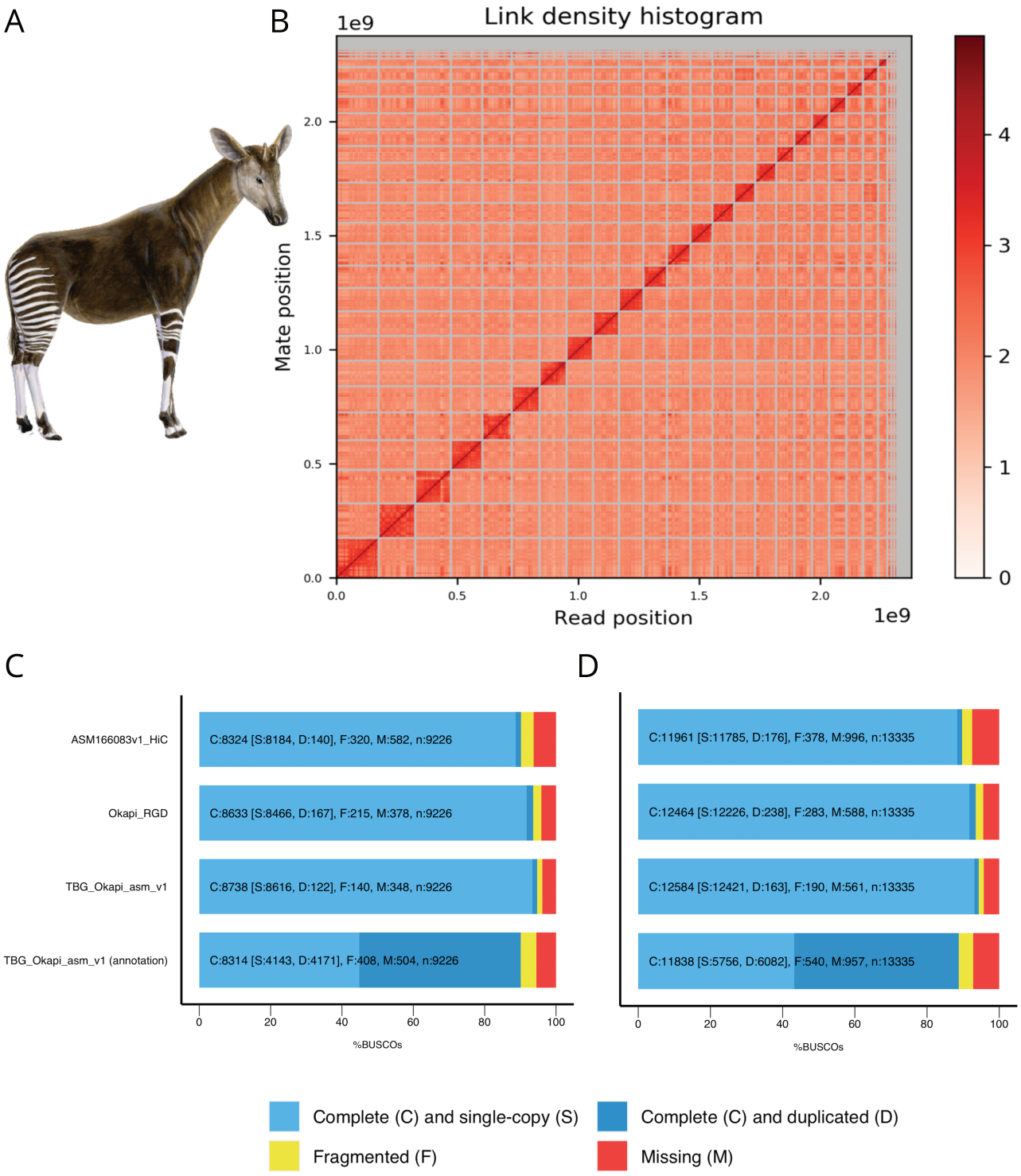

Figure 1. The okapi, Hi-C contact map, and Gene set completeness analyses. Painting of the okapi by Jon Baldur Hlidberg (www.fauna.is) (A). Proximity ligation contact map of the scaffolded assembly *TBG_Okapi_asm_v1* (B). Gene set completeness analyses (BUSCO) using both the mammalia_odb10 (C) and the cetartiodactyla_odb10 gene sets (D) for the 3 okapi assemblies and the annotation of *TBG_Okapi_asm_v1*. The assembly with the highest BUSCO scores for both gene sets is *TBG_Okapi_asm_v1*.

selection, and future comparative and population genomic analyses. This improved assembly of a genome from the sister group to the giraffe is an important resource to study the genetic basis and evolution of the unique characteristics of giraffids and will facilitate future genome-wide conservation genomic assessments of wild okapi populations.

Methods

Biological Materials

A kidney-tissue sample from a male okapi named Moyo (International Studbook Number 0385) was collected by veterinarians during post-mortem examination at the Zoo Antwerp and preserved at −80 °C until further processing.

High-molecular-weight genomic DNA (gDNA) was extracted using a standard phenol-chloroform extraction protocol (Sambrooks J, Russel DW 2001). DNA concentration and yield were measured using the Qubit dsDNA BR Assay Kit on the Qubit Fluorometer (Thermo Fisher Scientific) and DNA integrity was evaluated using the Genomic DNA ScreenTape on the Agilent 2200 TapeStation system (Agilent Technologies).

Nucleic Acid Library Preparation and Sequencing

A SMRTbell sequencing library was prepared using the SMRTbell Express Prep kit v2.0 Protocol (Pacific Biosciences—PacBio, Menlo Park, CA, USA) and sequenced on the PacBio Sequel II system on continuous long read (CLR) mode with the Sequel II Sequencing Kit 2.0 (PacBio). For preparing a Chromium sequencing library (10X Genomics, Inc., Pleasanton, CA, USA), the gDNA was size-selected using a BluePippin System (SageScience, Inc., Beverly, MA, USA) to remove DNA molecules < 30 kbp. The size-selected gDNA was then sent to SciLifeLab (Stockholm, Sweden) for library preparation and sequencing on the Illumina NovaSeq 6000 system.

Genome Assembly

The PacBio subreads sequencing output from the Sequel II run was converted from BAM to FASTQ format using BAM2fastx tools (PacBio) and de novo assembled with WTDBG2 v.2.5 (https://scicrunch.org/resources/data/record/nlx_144509-1/SCR_017225/resolver?q=SCR_017225&l=SCR_017225&i=rrid:scr_017225; RRID: SCR_017225) (Ruan and Li 2019) using the preset for PacBio Sequel reads (flag "-x sq"). The resulting assembly was subjected to a two-step polishing process to further improve the assembly accuracy. First, three iterations of long-read polishing were conducted in racon v.1.4.3 (https://scicrunch.org/resources/data/record/nlx_144509-1/SCR_017642/resolver?q=SCR_017642&l=SCR_017642&i=rrid:scr_017642; RRID: SCR_017642) (Vaser et al. 2017). For each iteration, the same PacBio reads used for the assembly were mapped to the latest assembly with minimap2 v.2.17-r941 (https://scicrunch.org/resources/data/record/nlx_144509-1/SCR_018550/resolver?q=SCR_018550&l=SCR_018550&i=rrid:scr_018550; RRID: SCR_018550) (Li 2018). Subsequently, three iterations of polishing with high-accuracy short-reads were performed with pilon v.1.23 (https://scicrunch.org/resources/data/record/nlx_144509-1/SCR_014731/resolver?q=SCR_014731&l=SCR_014731&i=rrid:scr_014731; RRID: SCR_014731) (Walker et al. 2014) to reduce single base errors. The short reads derived from the Chromium library were used for short-read polishing, after an initial assembly using Supernova v.2.1.1 (10X Genomics, Inc.; https://scicrunch.org/resources/data/record/nlx_144509-1/SCR_016756/resolver?q=SCR_016756&l=SCR_016756&i=rrid:scr_016756; RRID: SCR_016756) suggested that the library quality was not sufficient for a standalone de novo assembly. To make the reads available for analyses, they were processed with Long Ranger v.2.2.2 (10x Genomics, Inc.; https://scicrunch.org/resources/data/record/nlx_144509-1/SCR_018925/resolver?q=SCR_018925&l=SCR_018925&i=rrid:scr_018925; RRID: SCR_018925) and mapped to the latest assembly with bwa-mem v.0.7.17-r1194 (https://scicrunch.org/resources/data/record/nlx_144509-1/SCR_010910/resolver?q=SCR_010910&l=SCR_010910&i=rrid:scr_010910; RRID: SCR_010910) (Li 2013) before every iteration of pilon.

The polished assembly was scaffolded by Dovetail Genomics' HiRise pipeline (Putnam et al. 2016) using publicly available Hi-C data for the okapi (accession numbers: SRR8616855, SRR8616856) from the DNA Zoo Consortium (Dudchenko et al. 2017). Three iterations of gap-closing were performed with TGS-GapCloser v.1.1.1 (https://scicrunch.org/resources/data/record/nlx_144509-1/SCR_017633/resolver?q=SCR_017633&l=SCR_017633&i=rrid:scr_017633; RRID: SCR_017633) (Xu et al. 2020) to further improve the continuity of the assembly.

Assembly QC

Assembly statistics were calculated with Quast v.5.0.2 (https://scicrunch.org/resources/data/record/nlx_144509-1/SCR_001228/resolver?q=SCR_001228&l=SCR_001228&i=rrid:scr_001228; RRID: SCR_001228) (Gurevich et al. 2013) and a gene set completeness analyses was conducted with BUSCO v.5.1.3 (https://scicrunch.org/resources/data/record/nlx_144509-1/SCR_015008/resolver?q=SCR_015008&l=SCR_015008&i=rrid:scr_015008; RRID: SCR_015008) (Seppey et al. 2019) using both the Cetartiodactyla (cetartiodactyla_odb10) and the Mammalia (mammalia_odb10) datasets. The results were compared to 2 publicly available okapi genome assemblies: *ASM166083v1_HiC* from the DNA Zoo Consortium (Agaba et al. 2016; Dudchenko et al. 2017) and *Okapi_RGD* from the Ruminant Genome Project (Chen et al. 2019).

QualiMap v.2.2.1 (https://scicrunch.org/resources/data/record/nlx_144509-1/SCR_001209/resolver?q=SCR_001209&l=SCR_001209&i=rrid:scr_001209; RRID: SCR_001209) (Okonechnikov et al. 2016) was used to assess the mapping rate and coverage distribution across the genome assembly after mapping both the short and long reads back to the assembly using bwa-mem and minimap2, respectively. The resulting mapping files were used together with the output of a BLASTN v.2.11.0 + (https://scicrunch.org/resources/data/record/nlx_144509-1/SCR_001598/resolver?q=SCR_001598&l=SCR_001598&i=rrid:scr_001598; RRID: SCR_001598) (Zhang et al. 2000) search against NCBI's Nucleotide database to check the assembly for contamination with BlobTools v.1.1.1 (https://scicrunch.org/resources/data/record/nlx_144509-1/SCR_017618/resolver?q=SCR_017618&l=SCR_017618&i=rrid:scr_017618; RRID: SCR_017618) (Laetsch and Blaxter 2017). We also estimated the completeness, quality value (QV), and base-level error rate of the assembly using Merqury v.1.1 (Rhie et al. 2020) with a *k*-mer size of 21.

Genome Size Evaluation

The okapi's haploid genome size was estimated by two different approaches: one based on *k*-mer counting and another based on read depth. In the first approach, we estimated the genome size using GenomeScope v.2.0 (https://scicrunch.org/resources/data/record/nlx_144509-1/SCR_017014/resolver?q=SCR_017014&l=SCR_017014&i=rrid:scr_017014; RRID: SCR_017014) (Vurture et al. 2017) based on the 21-mer count calculated from the short reads with Jellyfish v.2.2.10 (https://scicrunch.org/resources/data/record/nlx_144509-1/SCR_005491/resolver?q=SCR_005491&l=SCR_005491&i=rrid:scr_005491; RRID: SCR_005491) (Marçais and

Kingsford 2011). In the second approach, we estimated the genome size based on the coverage of the short reads mapped onto the assembly using backmap v.0.3 (Pfenninger et al. 2021).

Annotation

To annotate the repeats in the assembly, a de novo repeat library was generated using RepeatModeler v.2.0.1 (https://scicrunch.org/resources/data/record/nlx_144509-1/SCR_015027/resolver?q=SCR_015027&l=SCR_015027&i=rrid:scr_015027; RRID: SCR_015027) (Flynn et al. 2020) and combined with the Cetartiodactyla repeat database from RepBase (Bao et al. 2015). This custom library was then used to annotate, and in a second step, mask the repeats in the genome using RepeatMasker v.4.1.0 (https://repeatmasker.org/RepeatMasker/; https://scicrunch.org/resources/data/record/nlx_144509-1/SCR_012954/resolver?q=SCR_012954&l=SCR_012954&i=rrid:scr_012954; RRID: SCR_012954). Interspersed repeats were hard-masked and simple repeats soft-masked to increase the accuracy of gene annotation.

The GeMoMa pipeline v.1.7.1 (https://scicrunch.org/resources/data/record/nlx_144509-1/SCR_017646/resolver?q=SCR_017646&l=SCR_017646&i=rrid:scr_017646; RRID: SCR_017646) (Keilwagen et al. 2016; Keilwagen et al. 2018) was used for homology-based gene prediction with the alignment tool MMseqs2 (Steinegger and Söding 2017). Ten mammalian genomes and corresponding annotations were used as references: *Bos taurus* (GCF_002263795.1), *Homo sapiens* (GCF_000001405.39), *Mus musculus* (GCF_000001635.27), *Sus scrofa* (GCF_000003025.6), *Camelus dromedarius* (GCF_000803125.2), *Equus caballus* (GCF_002863925.1), *Ovis aries* (GCF_002742125.1), *Tursiops truncatus* (GCF_011762595.1), *Cervus hanglu yarkandensis* (GCA_010411085.1), and *Capra hircus* (GCF_001704415.1). The predicted genes were annotated by a BLASTP v.2.11.0 + (https://scicrunch.org/resources/data/record/nlx_144509-1/SCR_001010/resolver?q=SCR_001010&l=SCR_001010&i=rrid:scr_001010; RRID: SCR_001010) (Zhang et al. 2000) search against the Swiss-Prot database (https://scicrunch.org/resources/data/record/nlx_144509-1/SCR_002380/resolver?q=SCR_002380&l=SCR_002380&i=rrid:scr_002380; RRID: SCR_002380; release 2021-01) with an e-value cutoff of 10^{-6}. Gene ontology (GO) terms, motifs, and domains were further annotated with InterProScan v.5.50.84 (https://scicrunch.org/resources/data/record/nlx_144509-1/SCR_005829/resolver?q=SCR_005829&l=SCR_005829&i=rrid:scr_005829; RRID: SCR_005829) (Quevillon et al. 2005; Jones et al. 2014).

Synteny Analyses

Synteny between the 2 chromosome-scale okapi assemblies, as well as between okapi and cattle (GCF_002263795.1), sheep (GCF_002742125.1), and the Masai giraffe (GCA_013496395.1) (Farré et al. 2019) was analyzed using JupiterPlot v.3.81 (Chu 2018).

Demographic History

The demographic history of the okapi was inferred for the two chromosome-scale assemblies of the three available okapi genomes using PSMC (https://scicrunch.org/resources/data/record/nlx_144509-1/SCR_017229/resolver?q=SCR_017229&l=SCR_017229&i=rrid:scr_017229; RRID: SCR_017229) (Li and Durbin 2011). Before running PSMC, we annotated the repeats in each assembly as described above and hard-masked the transposable elements (TEs) and simple repeats. Next, we trimmed the short-read data with fastp v.0.20.1 (https://scicrunch.org/resources/data/record/nlx_144509-1/SCR_016962/resolver?q=SCR_016962&l=SCR_016962&i=rrid:scr_016962; RRID: SCR_016962) (Chen et al. 2018). Trimming and filtering were performed with low complexity filter and base correction enabled. Sequencing adapters and polyG stretches at the end of reads were removed and a sliding window of 4 bp was applied to detect poor quality regions (Phred score < 15). Reads were discarded if they were shorter than 36 bp, had > 40% low-quality bases, or more than five undetermined bases. Subsequently, the trimmed reads were mapped to the assembly using bwa-mem, duplicates were marked with Picard v.2.20.8 (https://scicrunch.org/resources/data/record/nlx_144509-1/SCR_006525/resolver?q=SCR_006525&l=SCR_006525&i=rrid:scr_006525; RRID: SCR_006525) (Broad Institute 2019), and all reads from the BAM files that were not properly mapped in non-repeat regions with expected insert-sizes were removed. The resulting BAM files were used to generate a diploid consensus sequence for each assembly as input for PSMC using Samtools v.1.9 (https://scicrunch.org/resources/data/record/nlx_144509-1/SCR_002105/resolver?q=SCR_002105&l=SCR_002105&i=rrid:scr_002105; RRID: SCR_002105) (Li et al. 2009) and BCFtools v.1.9 (RRID: SCR_002105) (Danecek et al. 2021). PSMC was run with the parameters -N25 -t15 -r5 -p "4 + 25*2 + 4+6" and 100 bootstrap replicates. To scale the results, we applied a generation time of 8 years and a mutation rate of 1.82×10^{-8} substitutions per site per generation (Chen et al. 2019).

Genome-Wide Heterozygosity

A folded Site Frequency Spectrum (SFS) was used to estimate the genome-wide heterozygosity of the two chromosome-scale assemblies from the BAM files generated for PSMC. First, ANGSD v.0.933 (Korneliussen et al. 2014) (flag -doSaf 1) was used to estimate site allele frequencies. We set the minimum score for both mapping and base quality to 30, used the 95th percentile of the sample's depth distribution as maximum depth cut-off, and enabled the extended Base Alignment Quality (BAQ) adjustment (flag -baq 2). The folded SFS was then generated with ANGSD's companion program realSFS (flag -fold 1) with 200 bootstrap replicates and used to calculate heterozygosity as a percentage of heterozygous sites in R v.4.1.2 (R Core Team. 2015).

Results and Discussion

Genome Sequencing and Assembly

Sequencing on the PacBio Sequel II generated 196.7 Gbp of long-read data with a mean subread length of 9.79 kbp, yielding a coverage of approximately 79-fold. For the 10X Genomics Chromium library, we received 101.6 Gbp of raw sequencing data or 677 million short reads (41-fold).

The initial supernova run to assemble the Chromium data resulted in an assembly with a total length of only 1.9 Gbp and a scaffold N50 of 47.40 kbp. Therefore, we opted for a PacBio-based de novo assembly and used the Chromium data only for polishing. The final scaffolded and gap-closed chromosome-scale assembly (*TBG_Okapi_asm_v1*) has a

Table 1. Assembly statistics for 3 okapi assemblies (A) and repeat content of TBG_Okapi_asm_v1 (B).

	TBG_Okapi_asm_v1	ASM166083v1_HiC	Okapi_RGD	TBG_Okapi_asm_v1 contig-level[a]	ASM166083v1_HiC contig-level[a]	Okapi_RGD contig-level[a]
A						
No. of scaffolds/contigs	3653	172 277	42 448	3728	277 832	91 907
No. of scaffolds/contigs (> 1 kbp)	3651	86 360	25 028	3726	164 046	73 000
L50	10	11	230	13	16 645	7876
LG50[b]	10	10	217	15	16 286	7625
N50 (bp)	102 048 473	101 359 828	3 620 116	61 700 678	43 239	95 185
NG50[b](bp)	102 048 473	104 281 413	3 764 328	53 333 850	43 889	97 169
Max. scaffold/contig length (bp)	174 459 755	181 025 969	14 878 189	148 083 009	469 718	825 140
Total length (bp)	2 387 733 189	2 890 536 570	2 621 199 438	2 387 725 973	2 557 339 957	2 574 443 623
GC (%)	41.81	41.50	41.54	41.81	41.49	41.54
No. of *N*'s	7976	152 208 624	46 728 290	0	296 936	37 810
No. of *N*'s per 100 kbp	0.33	5613.91	1782.71	0.0	11.61	1.47

B

Type of element	Number of elements	Length (bp)	Percentage of assembly
SINEs	477 903	68 662 602	3.33%
LINEs:	2 621 069	725 760 611	31.09%
L1/LINE1	1 275 373	351 185 177	13.70%
RTE/Bov-B	1 113 880	331 287 642	15.27%
LTR elements	415 791	114 098 712	4.64%
DNA transposons	332 213	62 260 499	2,55%
Unclassified	222 957	75 076 405	0.96%
Small RNA	241 675	36 024 135	1.84%
Satellites	217	184 067	0.04%
Simple repeats	397 628	17 447 967	0.74%
Low complexity	69 534	3 598 861	0.15%
Total	7 168 240	1 785 586 678	43.53%

[a]broken into contigs at gaps with a length of ≥ 10 *N*'s. Statistics for these columns are based on contigs, the remaining columns are based on scaffolds.
[b]Based on an estimated reference length of 2 526 089 448 bp.

total length of 2 387 734 269 bp in 3653 scaffolds and 3728 contigs, resulting in a contig N50 of over 61 Mbp and a scaffold N50 of more than 102 Mbp (Table 1A). The 28 largest scaffolds (>3.9 Mbp), which is five or six more than the expected haploid number of chromosomes for the okapi (2n = 44–46) (Ulbrich and Schmitt 1969; Petit et al. 1994) make up 98.0% of the total assembly length (Figure 1B). The remaining contigs are smaller than 500 kbp. The five or six scaffolds that could not be anchored into the expected 44-46 chromosome-scale scaffolds lack enough evidence in the Hi-C contact map to be placed properly either by the algorithm of the HiRise pipeline or by manual curation. One reason for this could be the use of previously published Hi-C data from a different individual, which could exhibit a different karyotype than the individual used for the new assembly.

With a contig NG50 of 53.3 Mbp, the new okapi genome assembly represents an approximate 1200-fold improvement in contiguity over *ASM166083v1_HiC* (NG50 = 43.9 kbp) and about 550-fold compared to *Okapi_RGD* (NG50 = 97.2 kbp) (Table 1A).

Assembly Completeness and Quality Assessment

GenomeScope estimated the genome size of the okapi at 2 526 089 448 bp, which is 140 Mbp larger than the total assembly length of *TBG_Okapi_asm_v1* but 360 Mbp smaller than *ASM166083v1_HiC*. The coverage-based approach used in backmap estimated the haploid genome size at 3.07 Gbp, nearly 500 Mbp larger than the *k*-mer based estimate. Despite the discrepancy in genome size estimates, Merqury estimated an assembly completeness for *TBG_Okapi_asm_v1* of 91.89% and a QV of 32.79 based on k-mers, corresponding to a base error rate of 0.05%.

The gene set completeness analyses in BUSCO also suggest high completeness of the assembly identifying 94.7% of complete BUSCO genes in the Mammalia and 94.3% in the Cetartiodactyla dataset, the highest BUSCO scores of the

three available assemblies (Figure 1C,D). Furthermore, long and short reads map to the assembly with a mapping rate of 95.7% and 95.5%, respectively, and no contamination is evident in the BlobPlot (Supplementary Figure S1).

Genome-Wide Heterozygosity

The genome-wide heterozygosity values for the chromosome-scale assemblies from *ASM166083v1_HiC* and the new assembly *TBG_Okapi_asm_v1* were estimated at 0.120% (sd 9.24×10^{-7}) and 0.173% (sd 1.22×10^{-6}), respectively. These new estimates are in the same range as a previous estimate (0.132%) (Brüniche-Olsen et al. 2018) derived from the same short-read data as *ASM166083v1_HiC*. Even though these genomes are derived from captive individuals, the current levels of heterozygosity are likely to be representative for wild okapi in that the fully resolved captive okapi pedigree suggests no recent inbreeding. For example, Moyo, the individual sampled for *TBG_Okapi_asm_v1*, is the offspring of a female founder individual and an unrelated 1.5-generation male.

In comparison to other species, the heterozygosity of the two okapi individuals are around three to four times higher than the heterozygosity of wild individuals of the northern giraffe *Giraffa camelopardalis* (0.03-0.045%), which have a census population of approximately 5000 individuals (Coimbra et al. 2021, 2022). However, these estimates were derived using a different BAQ adjustment (-baq 1) in ANGSD, which is known to underestimate the heterozygosity by approximately three to four times (Prasad et al. 2022). Estimates from Hu et al. (Hu et al. 2020) place the heterozygosity of the okapi in the same range as other endangered species such as the Sumatran Rhinoceros *Dicerorhinus sumatrensis* (0.130%), the Giant Panda *Ailuropoda melanoleuca* (0.132%), and the Western lowland gorilla *Gorilla gorilla gorilla* (0. 144%). Yet, estimates of heterozygosity vary greatly depending on the methods used; hence, comparisons between studies must be evaluated with caution. Furthermore, heterozygosity alone may be an inaccurate representation of a species' level of threat of extinction from genetic depletion (Teixeira and Huber 2021), thus requiring complementary assessments of mutational load, inbreeding, and population sizes.

Annotation

Repeat Annotation

Repeat annotation of *TBG_Okapi_asm_v1* identified a repeat content of 43.53% or 1.04 Gbp of the assembly (Table 1B).

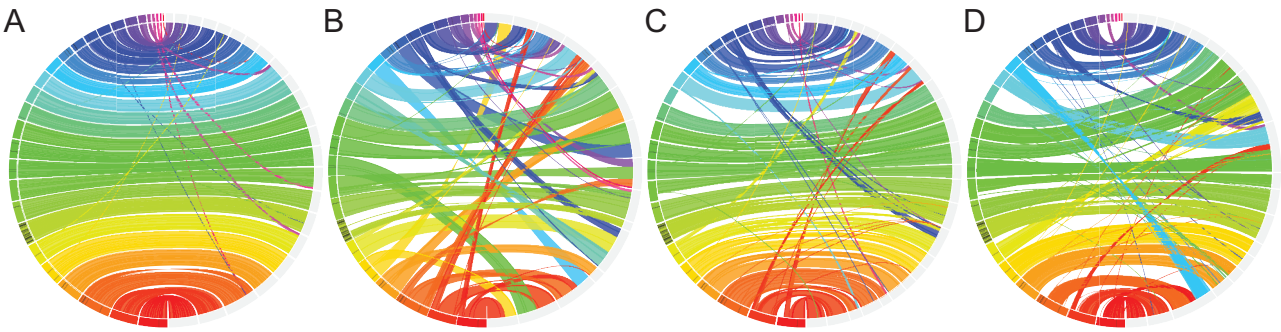

Figure 2. Synteny comparison between okapi and other ruminants. Synteny Plot generated with JupiterPlot between the okapi assembly *TBG_Okapi_asm_v1* (left) and (A) the okapi assembly *ASM166083v1_HiC*, (B) Masai giraffe, (C) cattle, and (E) sheep. Boxes in the outer circle of the plot indicate scaffolds of the reference (color) and query (grey). Black lines within the colored boxes represent gaps in the reference assembly.

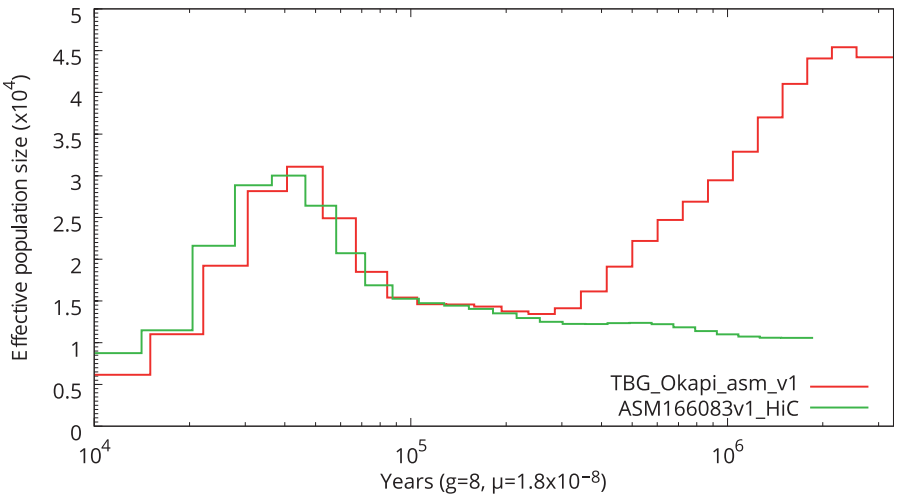

Figure 3. Demographic history of 2 okapi individuals estimated by PSMC.

The most abundant retroelements are Long Interspersed Nuclear Elements (LINEs), which account for 31.09% of the assembly, whereas Long Terminal Repeats (LTR) elements, Short Interspersed Nuclear Elements (SINEs), and DNA transposons each only account for < 5%. The most common LINEs in *TBG_Okapi_asm_v1* are RTE/Bov-B (15.27%) and L1/LINE1 (13.70%). Repeat statistics calculated for the two previously published assemblies show no major differences in repeat content compared to *TBG_Okapi_asm_v1*, except of an approximately 2% lower repeat content in *ASM166083v1_HiC* likely attributed to the more fragmented assembly (Supplementary Table S1).

Gene Annotation

The homology-based gene prediction identified 23 412 genes in the *TBG_Okapi_asm_v1* assembly with a median gene length of 7408 bp resulting in BUSCO scores of 88.8% and 90.1% for the Cetartiodactyla and Mammalia datasets, respectively (Figure 1C,D). Of all 55 344 predicted proteins, 54 708 (98.85%) were functionally annotated by InterProScan, and in 41 684 proteins (75.31%) at least one GO term was identified. Furthermore, 53 839 proteins (97.27%) were assigned to the Swiss-Prot database.

Genome Synteny With Other Ruminants

Synteny analyses revealed high synteny between the two okapi assemblies (Figure 2A) with a few inversions and five additional scaffolds in *TBG_Okapi_asm_v1*. Manual curation of the Hi-C-based scaffolding yielded no evidence for the placements of the five additional scaffolds or the correct orientation of contigs responsible for the inversions in the synteny plot. As expected from cytological studies, the comparison between okapi and the Masai giraffe genome (Figure 2B) revealed major rearrangements. In fact, not a single okapi scaffold could be found in the giraffe without being split into multiple parts or combined into the 15 much larger giraffe chromosomes. These chromosomal rearrangements, especially in the giraffe, are caused by Robertsonian-centric fusions of acrocentric chromosomes (Huang et al. 2008; Cernohorska et al. 2013; Agaba et al. 2016; Liu et al. 2021). In contrast, the synteny between okapi and cattle or sheep was much higher (Figure 2C,D), despite these species being more distantly related to the okapi than the giraffe. Many okapi scaffolds showed perfect synteny with scaffolds of the non-giraffid ruminants, indicating slow karyological evolution/changes.

PSMC

The estimation of effective population size (*Ne*) over time with PSMC revealed different trajectories for the two okapi genome assemblies *TBG_okapi_asm_v1 and ASM166083v1_HiC*. For recent times (10–250 kya), both assemblies show congruent demographic patterns, with a maximum *Ne* around 40 kya. However, they diverge drastically further in the past (250 kya–2 Mya) reaching a 3-fold difference in *Ne* around 1.5 Mya (Figure 3). Differences in Ne estimates for such ancient time spans should be interpreted with caution, as they may reflect an artifact of the method. Further, both datasets are based on different library preparation methods (standard Illumina vs. 10× Genomics' Chromium), and to our knowledge, the effects of these methods on an analysis such as PSMC have not been tested. Yet, bootstrap analyses showed clear support for each of the two PSMC trajectories (Supplementary Figure S2).

The third assembly (*Okapi_RGD*) could not be analyzed in this study, due to major differences in insert sizes in the available datasets, resulting in a much lower coverage after applying the same filtering of the BAM file as for the other datasets. However, *Okapi_RGD* has been previously analyzed with PSMC and showed a trajectory more similar to TBG*_okapi_asm_v1* than *ASM166083v1_HiC* but with much more variability in the bootstrap replicates (Chen et al. 2019).

Conclusion

High-quality genome assemblies are an important resource for a variety of studies, especially for comparative genomics and conservation/ population genomics analyses. Currently, references are still missing for most non-model organisms, making the high-quality chromosome-scale genome assembly of the okapi presented here an important contribution to the field of evolutionary genomics. This is the first long-read-based assembly for this species, with the highest contiguity and completeness of BUSCO genes among the available okapi genome assemblies. Long-read-based assemblies are more likely to span complex repeat regions, such as telomeric or centromeric regions, which are usually collapsed in short-read-based assemblies, allowing for more in-depth analyses of structural variation (Gordon et al. 2016; Korlach et al. 2017; Paris et al. 2020).

Supplementary Material

Supplementary material can be found at *Journal of Heredity* online.

Funding

The present study was funded by the Centre for Translational Biodiversity Genomics (LOEWE-TBG) through the program "LOEWE – Landes-Offensive zur Entwicklung Wissenschaftlich-ökonomischer Exzellenz" of Hesse's Ministry of Higher Education, Research, and the Arts.

Conflict of Interest

The authors declare that they have no conflicting interests.

Acknowledgments

We thank the Genome Technology Center (RGTC) at Radboudumc for the use of the Sequencing Core Facility (Nijmegen, The Netherlands), which provided the PacBio SMRT sequencing service on the Sequel II platform. We are grateful to the TBG laboratory team (Carola Greve, Damian Baranski, and Alexander Ben Hamadou) for the help with the DNA extraction and the PacBio library preparation.

Data Availability

The genome assembly and the underlying read data can be accessed at NCBI under BioProject PRJNA708170. All data, including the annotation, can be accessed as DRYAD

dataset (Winter et al. 2022, https://doi.org/10.5061/dryad.37pvmcvp3).

References

Agaba M, Ishengoma E, Miller WC, McGrath BC, Hudson CN, Bedoya Reina OC, 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. doi:10.1038/ncomms11519.

Bao W, Kojima KK, Kohany O. 2015. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob DNA* 6 doi:10.1186/s13100-015-0041-9.

Bickhart DM, Rosen BD, Koren S, Sayre BL, Hastie AR, Chan S, Lee J, Lam ET, Liachko I, Sullivan ST, et al. 2017. Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. *Nat Genet* 49:643–650. doi:10.1038/ng.3802.

Broad Institute. 2019. Picard toolkit. *Broad Inst*.

Brüniche-Olsen A, Kellner KF, Anderson CJ, DeWoody JA. 2018. Runs of homozygosity have utility in mammalian conservation and evolutionary studies. *Conserv Genet* 19:1295–1307. doi:10.1007/s10592-018-1099-y.

Cernohorska H, Kubickova S, Kopecna O, Kulemzina AI, Perelman PL, Elder FFB, Robinson TJ, Graphodatsky AS, Rubes J. 2013. Molecular cytogenetic insights to the phylogenetic affinities of the giraffe (*Giraffa camelopardalis*) and pronghorn (*Antilocapra americana*). *Chromosome Res* 21:447–460. doi:10.1007/s10577-013-9361-0.

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* 364:1–12. doi:10.1126/science.aav6202.

Chen S, Zhou Y, Chen Y, Gu J. 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 34: i884–i890. doi:10.1093/bioinformatics/bty560.

Chu J. 2018. Jupiter Plot: A Circos-based tool to visualize genome assembly consistency (1.0). *Zenodo*. doi:10.5281/zenodo.1241235.

Coimbra RTF, Winter S, Kumar V, Koepfli K-P, Gooley RM, Dobrynin P, Fennessy J, Janke A. 2021. Whole-genome analysis of giraffe supports four distinct species. *Curr Biol* 31: P2929–2938. doi:10.1016/j.cub.2021.04.033.

Coimbra RTF, Winter S, Mitchell B, Fennessy J, Janke A. 2022. Conservation Genomics of Two Threatened Subspecies of Northern Giraffe: The West African and the Kordofan Giraffe. *Genes* 13:1–14. doi: 10.3390/genes13020221.

Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, et al. 2021. Twelve years of SAMtools and BCFtools. *GigaScience* 10:1–4. doi:10.1093/gigascience/giab008.

Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, Durand NC, Shamim MS, Machol I, Lander ES, Aiden AP, et al. 2017. De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science* 356:92–95. doi:10.1126/science.aal3327.

Farré M, Li Q, Darolti I, Zhou Y, Damas J, Proskuryakova AA, Kulemzina AI, Chemnick LG, Kim J, Ryder OA, et al. 2019. An integrated chromosome-scale genome assembly of the Masai giraffe (*Giraffa camelopardalis tippelskirchi*). *GigaScience* 8:1–9. doi:10.1093/gigascience/giz090.

Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, Smit AF, et al. 2020. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc Natl Acad Sci. National Academy of Sciences* 117:9451–9457 doi:10.1073/pnas.1921046117.

Gordon D, Huddleston J, Chaisson MJP, Hill CM, Kronenberg ZN, Munson KM, Malig M, Raja A, Fiddes I, Hillier LW, et al. 2016. Long-read sequence assembly of the gorilla genome. *Science* 352. doi: 10.1126/science.aae0344.

Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* 29:1072–1075. doi:10.1093/bioinformatics/btt086.

Hu J-Y, Hao Z-Q, Frantz L, Wu S-F, Chen W, Jiang Y-F, Wu H, Kuang W-M, Li H, Zhang Y-P, et al. 2020. Genomic consequences of population decline in critically endangered pangolins and their demographic histories. *Natl Sci Rev* 7:798–814. doi:10.1093/nsr/nwaa031.

Huang L, Nesterenko A, Nie W, Wang J, Su W, Graphodatsky AS, Yang F. 2008. Karyotype evolution of giraffes (*Giraffa camelopardalis*) revealed by cross-species chromosome painting with Chinese muntjac (*Muntiacus reevesi*) and human (*Homo sapiens*) paints. *Cytogenet Genome Res* 122:132–138. doi: 10.1159/000163090.

Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, et al. 2014. InterProScan 5: genome-scale protein function classification. *Bioinformatics* 30:1236–1240. doi:10.1093/bioinformatics/btu031.

Keilwagen J, Hartung F, Paulini M, Twardziok SO, Grau J. 2018. Combining RNA-seq data and homology-based gene prediction for plants, animals and fungi. *BMC Bioinf* 19. doi:10.1186/s12859-018-2203-5.

Keilwagen J, Wenk M, Erickson JL, Schattat MH, Grau J, Hartung F. 2016. Using intron position conservation for homology-based gene prediction. *Nucleic Acids Res* 44:e89. doi:10.1093/nar/gkw092.

Korlach J, Gedman G, Kingan SB, Chin C-S, Howard JT, Audet J-N, Cantin L, Jarvis ED. 2017. De novo PacBio long-read and phased avian genome assemblies correct and add to reference genes generated with intermediate and short reads. *GigaScience* 6:1–6. doi:10.1093/gigascience/gix085.

Korneliussen TS, Albrechtsen A, Nielsen R. 2014. ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinf* 15:1–13. doi:10.1186/s12859-014-0356-4.

Laetsch DR, Blaxter ML. 2017. BlobTools: Interrogation of genome assemblies. *F1000Research* 6:1–18. doi:10.12688/f1000research.12232.1.

Li H. 2013 Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv* ArXiv: 13033997. doi:10.48550/arXiv.1303.3997, preprint: not peer reviewed.

Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34:3094–3100. doi:10.1093/bioinformatics/bty191.

Li H, Durbin R. 2011. Inference of human population history from individual whole-genome sequences. *Nature* 475:493–496. doi:10.1038/nature10231.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25:2078–2079. doi:10.1093/bioinformatics/btp352.

Liu C, Gao J, Cui X, Li Z, Chen L, Yuan Y, Zhang Y, Mei L, Zhao L, Qiu Q, et al. 2021. A towering genome: Experimentally validated adaptations to high blood pressure and extreme stature in the giraffe. *Sci Adv. American Association for the Advancement of Science* 7:eabe9459. doi:10.1126/sciadv.abe9459.

Mallon D, Kümpel N, Quinn A, Shurter S, Lukas J, Hart J, Mapilanga J, Beyers R, Maisels F. 2015. *Okapia johnstoni*. The IUCN Red List of Threatened Species 2015: e.T15188A51140517. doi:10.2305/IUCN.UK.2015-4.RLTS.T15188A51140517.en. Accessed on 27 July 2022.

Marçais G, Kingsford C. 2011. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* 27:764–770. doi:10.1093/bioinformatics/btr011.

Marrano A, Britton M, Zaini PA, Zimin AV, Workman RE, Puiu D, Bianco L, Di Pierro EA, Allen BJ, Chakraborty S, et al. 2020. High-quality chromosome-scale assembly of the walnut (*Juglans regia* L.) reference genome. *GigaScience* 9:1–6. doi:10.1093/gigascience/giaa050.

Okonechnikov K, Conesa A, García-Alcalde F. 2016. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics* 32:292–294. doi:10.1093/bioinformatics/btv566.

Paris M, Boyer R, Jaenichen R, Wolf J, Karageorgi M, Green J, Cagnon M, Parinello H, Estoup A, Gautier M, et al. 2020. Near-chromosome level genome assembly of the fruit pest *Drosophila suzukii* using long-read sequencing. *Sci Rep* 10:11227. doi:10.1038/s41598-020-67373-z.

Petit P, De Bois H, De Meurichy W. 1994. Chromosomal reduction in an okapi pedigree (*Okapia johnstoni*). *Z Saugetierkunde. KOCH NEFF UND OETINGER* 59:153.

Pfenninger M, Schönenbeck P, Schell T (2021). ModEst: Accurate estimation of genome size from next generation sequencing data. *Mol Ecol Resour*: 1–11. https://doi.org/10.1111/1755-0998.13570

Prasad A, Lorenzen ED, Westbury MV. 2022. Evaluating the role of reference-genome phylogenetic distance on evolutionary inference. *Mol Ecol Resour* 1:45–55. doi:10.1111/1755-0998.13457.

Putnam NH, O'Connell BL, Stites JC, Rice BJ, Blanchette M, Calef R, Troll CJ, Fields A, Hartley PD, Sugnet CW, et al. 2016. Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *Genome Res* 7:342–350. doi:10.1101/gr.193474.115.

Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, Lopez R. 2005. InterProScan: protein domains identifier. *Nucleic Acids Res* 33: W116–W120. doi:10.1093/nar/gki442.

R Core Team. *R: A language and environment for statistical computing: R Foundation for Statistical Computing*. Vienna (Austria). 2015. Available from http://www.R-project.org.

Rhie A, McCarthy SA, Fedrigo O, Damas J, Formenti G, Koren S, Uliano-Silva M, Chow W, Fungtammasan A, Kim J, et al. 2021. Towards complete and error-free genome assemblies of all vertebrate species. *Nature* 592: 737–746. doi:10.1038/s41586-021-03451-0.

Rhie A, Walenz BP, Koren S, Phillippy AM. 2020. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol* 21: 1–27. doi:10.1186/s13059-020-02134-9.

Ruan J, Li H. 2019. Fast and accurate long-read assembly with wtdbg2. *Nat Methods* 17:155–158. doi:10.1038/s41592-019-0669-3.

Sambrook J, Russel DW. 2001. *Molecular Cloning: A Laboratory Manual*, 3rd ed.; Cold Spring Harbor Laboratory Press: Cold Spring Harbor, NY, USA. ISBN 0879695773. .

Seppey M, Manni M, Zdobnov EM. 2019. BUSCO: assessing genome assembly and annotation completeness. *Methods Mol Biol Clifton NJ* doi:10.1007/978-1-4939-9173-0_14.

Stanton DWG, Hart J, Galbusera P, Helsen P, Shephard J, Kümpel NF, et al. 2014. Distinct and diverse: range-wide phylogeography reveals ancient lineages and high genetic variation in the endangered okapi (*Okapia johnstoni*). *PLOS ONE* 9: e101081. Public Library of Science. doi:10.1371/journal.pone.0101081.

Stanton DWG, Hart J, Vosper A, Kümpel NF, Wang J, Ewen JG, Bruford MW. 2016. Non-invasive genetic identification confirms the presence of the Endangered okapi *Okapia johnstoni* southwest of the Congo River. *Oryx* 50: 134–137. doi: 10.1017/S0030605314000593.

Steinegger M, Söding J. 2017. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol* 35:1026–1028. doi:10.1038/nbt.3988.

Teixeira JC, Huber CD. 2021. The inflated significance of neutral genetic diversity in conservation genetics. *Proc Natl Acad Sci. Proceedings of the National Academy of Sciences* 118: 1–10. doi:10.1073/pnas.2015096118.

Ulbrich F, Schmitt J. 1969. [The chromosomes of *Okapia johnstoni* (Sclater, 1901)]. *Acta Zool Pathol Antverp* 49:123–124.

Vaser R, Sović I, Nagarajan N, Šikić M. 2017. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res* 27:734–746. doi:10.1101/gr.214270.116.

Vurture GW, Sedlazeck FJ, Nattestad M, Underwood CJ, Fang H, Gurtowski J, Schatz MC. 2017. GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics* 33:2202–2204. doi:10.1093/bioinformatics/btx153.

Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, et al. 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* 9:e112963. doi:10.1371/journal.pone.0112963.

Winter S, Prost S, De Raad J, Coimbra RTF, Wolf M, Nebenführ M, Held A, Kurzawe M, Papapostolou R, Tessien J, et al. 2020. Chromosome-level genome assembly of a benthic associated Syngnathiformes species: the common dragonet, *Callionymus lyra*. *Gigabyte*:1–10. doi:10.46471/gigabyte.6.

Winter S, Coimbra RTF, Helsen P, Janke A. 2022. A chromosome-scale genome assembly of the okapi (Okapia johnstoni), Dryad, Dataset. https://doi.org/10.5061/dryad.37pvmcvp3.

Xu M, Guo L, Gu S, Wang O, Zhang R, Peters BA, Fan G, Liu X, Xu X, Deng L, et al. 2020. TGS-GapCloser: A fast and accurate gap closer for large genomes with low coverage of error-prone long reads. *GigaScience* 9:1–11. doi:10.1093/gigascience/giaa094.

Zhang Z, Schwartz S, Wagner L, Miller W. 2000. A Greedy Algorithm for Aligning DNA Sequences. *J Comput Bio* 7:203–214; doi: 10.1089/10665270050081478.