scientific data

Check for updates

OPEN

DATA DESCRIPTOR

A near telomere-to-telomere phased reference assembly for the male mountain gorilla

David R. Nelson[1]✉, Richard Muvunyi[2,3], Khaled M. Hazzouri[4], Jean-Claude Tumushime[2,5], Gaspard Nzayisenga[5], Nziza Julius[5], Wim Meert[6], Latifa Karim[7], Wouter Coppieters[7], Katherine M. Munson[8], DongAhn Yoo[8], Evan E. Eichler[8,9], Kourosh Salehi-Ashtiani[1]✉ & Jean-Claude Twizere[1,2]✉

The endangered mountain gorilla, *Gorilla beringei beringei*, faces numerous threats to its survival, highlighting the urgent need for genomic resources to aid conservation efforts. Here, we present a near telomere-to-telomere, haplotype-phased reference genome assembly for a male mountain gorilla generated using PacBio HiFi (26.77× ave. coverage) and Oxford Nanopore Technologies (52.87× ave. coverage) data. The resulting non-scaffolded assembly exhibits exceptional contiguity, with contig N50 of ~95 Mbp for the combined pseudohaplotype (3,540,458,497 bp), 56.5 Mbp (3.1 Gbp) and 51.0 Mbp (3.2 Gbp) for each haplotype, an average QV of 65.15 (error rate = 3.1×10^{-7}), and a BUSCO score of 98.4%. These represent substantial improvements over most other available primate genomes. This first high-quality reference genome of the mountain gorilla provides an invaluable resource for future studies on gorilla evolution, adaptation, and conservation, ultimately contributing to the long-term survival of this iconic species.

Background & Summary

The mountain gorilla (*Gorilla beringei beringei*, see Fig. 1) is an endangered subspecies of the eastern gorilla, with a population of approximately 1,063 individuals remaining in the wild as of 2018[1]. These great apes are found exclusively in the high-altitude forests of the Virunga Massif, Sarambwe Reserve and Bwindi Impenetrable National Parks, spanning the borders of Rwanda, Uganda, and the Democratic Republic of Congo, at elevations ranging from 1,100 to 4,500 meters above sea level[2]. As one of our closest living relatives, sharing approximately 98% of their DNA with humans[2], the mountain gorilla holds deep evolutionary, ecological, and conservation importance. However, no reference genome was available for this subspecies. Understanding their genome is crucial for deciphering the genetic basis of their unique adaptations, such as their ability to thrive at high altitudes, their population history, and susceptibility to diseases, as well as for informing conservation strategies to protect this endangered species. Obtaining high-quality genomic samples from mountain gorillas is exceptionally challenging due to their limited population size, remote habitats, and strict conservation regulations[3]. Furthermore, until recently, producing enough long-read sequencing data for high-quality mammalian genomes has been costly, limiting candidate species for genomic projects. These factors have hindered the generation of a comprehensive reference genome for this subspecies. Previous genomic studies on mountain gorillas generally used sequencing reads from *G. beringei beringei* to align with other reference genomes, including the Eastern lowland gorilla (*Gorilla beringei graueri*)[4], limiting the scope of genetic analyses and comparative studies. A near telomere-to-telomere (T2T) phased reference assembly would provide a cutting-edge resource for investigating

[1]Laboratory of Algal, Synthetic, and Systems Biology, Division of Science and Math, New York University Abu Dhabi (NYUAD), Abu Dhabi, UAE. [2]Viral Interactomes Laboratory, GIGA Institute, University of Liege, Liege, Belgium. [3]Veterinary Unit, Conservation Department, Rwanda Development Board (RDB), Kigali, Rwanda. [4]Khalifa Center for Genetic Engineering & Biotechnology (KCGEB/UAEU), United Arab Emirates University, Al Ain, Abu Dhabi, UAE. [5]Mountain Gorilla Veterinary Project, "Gorilla Doctors", Musanze, Rwanda. [6]Genomics Core Leuven, Center for Human Genetics, KU Leuven, Leuven, Belgium. [7]Genomics Platform, GIGA Institute, University of Liege, Liege, Belgium. [8]Department of Genome Sciences, University of Washington School of Medicine, Seattle, Washington, USA. [9]Howard Hughes Medical Institute, University of Washington, Seattle, Washington, USA. ✉e-mail: drn2@nyu.edu; ksa3@nyu.edu; jean-claude.twizere@uliege.be

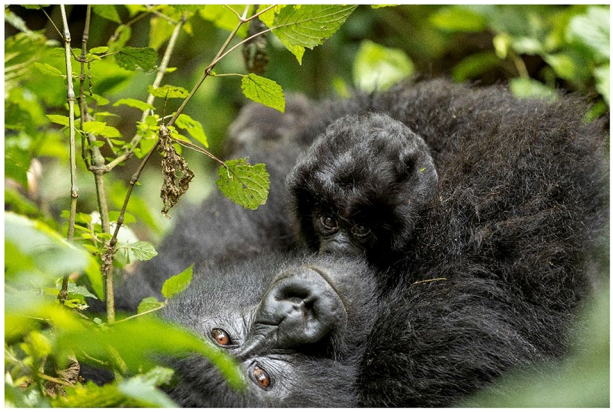

Fig. 1 Male mountain gorilla (*Gorilla beringei beringei*) donor of blood sample used for DNA extraction for sequencing photographed with mother. Opportunities for biological specimen sample collection in these endangered animals is typically restricted to periods where human intervention may be necessary to prevent life-threatening conditions. The two-year-old gorilla, Igicumbi (13.4 kg), underwent critical veterinary intervention, allowing for an opportunistic blood sampling for samples with high molecular weight (HMW) DNA extraction.

the complex evolutionary history, population dynamics, and genetic diversity of mountain gorillas, similar to recent advances in human genomics[5].

Recent advancements in long-read sequencing technologies, such as Pacific Biosciences (PacBio; Menlo Park, CA, USA) HiFi (high-fidelity) and Oxford Nanopore Technologies (ONT; Oxford, UK), have revolutionized the field of genomics by enabling the generation of highly contiguous and accurate genome assemblies[6]. These technologies, combined with state-of-the-art assembly algorithms and phasing methods, have made it possible to construct near-T2T assemblies for various species, including humans[7] and other great apes[8,9]. However, the application of these technologies to the mountain gorilla genome has been limited by the scarcity of high-quality samples and the challenges associated with obtaining them from wild populations[3].

We present the first near-T2T phased reference assembly for the male mountain gorilla, overcoming the challenges associated with obtaining high-quality genomic samples from this endangered subspecies. The presented haplotype and combined reference assemblies include all autosomes and the sex chromosomes for the sampled male mountain gorilla. We used a combination of long-read sequencing technologies and advanced assembly and k-mer-based phasing approaches to sequence the *G. beringei beringei* genome. By leveraging the power of PacBio HiFi reads, which offer both long read lengths (15–20 kbp) and high accuracy (>99.9%)[10], and ONT ultra-long reads, which can span hundreds of kilobases[11], we were able to generate accurate haplotype separation in the absence of parental information.

We constructed highly contiguous and accurate near-T2T assemblies that capture the majority of the mountain gorilla genome, including complex regions such as centromeres and telomeres (Table 1, Fig. 2). Compared to the T2T *G. gorilla* assembly[9], we find that ~90% of each of the chromosomes align with as few as two contigs (Fig. 4 and Table 2). The resulting assembly exhibits exceptional contiguity, with contig N50 of ~95 Mbp for the combined pseudohaplotype (3.5 Gbp), 56.5 Mbp (3.1 Gbp) and 51.0 Mbp (3.2 Gbp) for each haplotype, and an average QV of 65.15 (error rate = 3.1×10^{-7}). These high QV scores reflect the assembly's superior base-level accuracy and are consistent with the latest standards in long-read sequencing projects.

The generation of a near-T2T phased reference assembly for the mountain gorilla will be important for conservation biology, evolutionary genomics, and comparative studies among great apes. This resource will enable investigation of the genetic basis of adaptive traits, uncover the evolutionary history of mountain gorillas, and identify genomic regions associated with disease susceptibility and resilience. Moreover, the assembly will serve as a valuable reference for future population genomic studies, facilitating the development of targeted conservation strategies to protect this endangered subspecies.

Methods

Sample collection. Sample collection was conducted under the strict guidelines of a research permit issued by the Rwanda Development Board (RDB) for the project titled "*Obtaining high-quality genomic information and understanding the genetic diversity of mountain gorillas in Rwanda*," obtained from the Wildlife Authority of Rwanda. A two-year-old infant male gorilla named Igicumbi exhibited lethargy and reduced feeding following his mother's death from a severe respiratory infection. Due to his concerning condition, veterinarians conducted an immediate veterinary health intervention. After taking the animal's body weight measurement (13.4 kg), the veterinary team performed a chemical immobilization (anesthesia). A combination of a half-dose of ketamine (23.45 mg) and dexmedetomidine (0.2 mg) was administered intramuscularly into the right thigh. The anesthesia took effect within three minutes, with full sedation achieved within five minutes. Igicumbi remained stable throughout the procedure. A physical examination was performed, and samples, including swabs and blood, were taken and stored in various sample media. The blood sample (approximately 20 mL) was drawn from the left femoral vein using BD butterfly catheter and kept in four 5 ml EDTA tubes. The samples were immediately

Assembly	Total length (Gbp)	GC (%)	N50 (Mbp)	L50	Contigs	BUSCOs % complete	Reference	Sequencing technology
Gorilla gorilla (GCF_029281585.2)[17]	3.5	40.49	150	10	26	98.4	NHGRI, NIH	PacBio Sequel; ONT PromethION
G. beringei beringei Joined	3.5	40.59	95	14	350	98.4	This study	PacBio Sequel; ONT PromethION
G. beringei beringei Hap 2	3.2	40.49	51	16	665	89.1	This study	PacBio Sequel; ONT PromethION
G. beringei beringei Hap 1	3.1	40.29	57	18	986	83.1	This study	PacBio Sequel; ONT PromethION
Gorilla gorilla (GCA_030174185.1)[44] (Kamilah)	3.6	40.53	38	27	1053	98	UWashington	PacBio Sequel
G. beringei (GCA_963575185.1)	2.7	40.87	0.055	14910	109953	68.9	IBE	Illumina

Table 1. Comparison of assembly metrics, including non-scaffolded contigs, from the new *G. beringei beringei* assemblies and three other assemblies used for comparisons. The main *G. beringei beringei assembly* has had contiguous regions integrated from each haplotype to form a singular, pseudohaplotype representative assembly[15] for the species. Non-BUSCO[20] metrics were generated with QUAST[22].

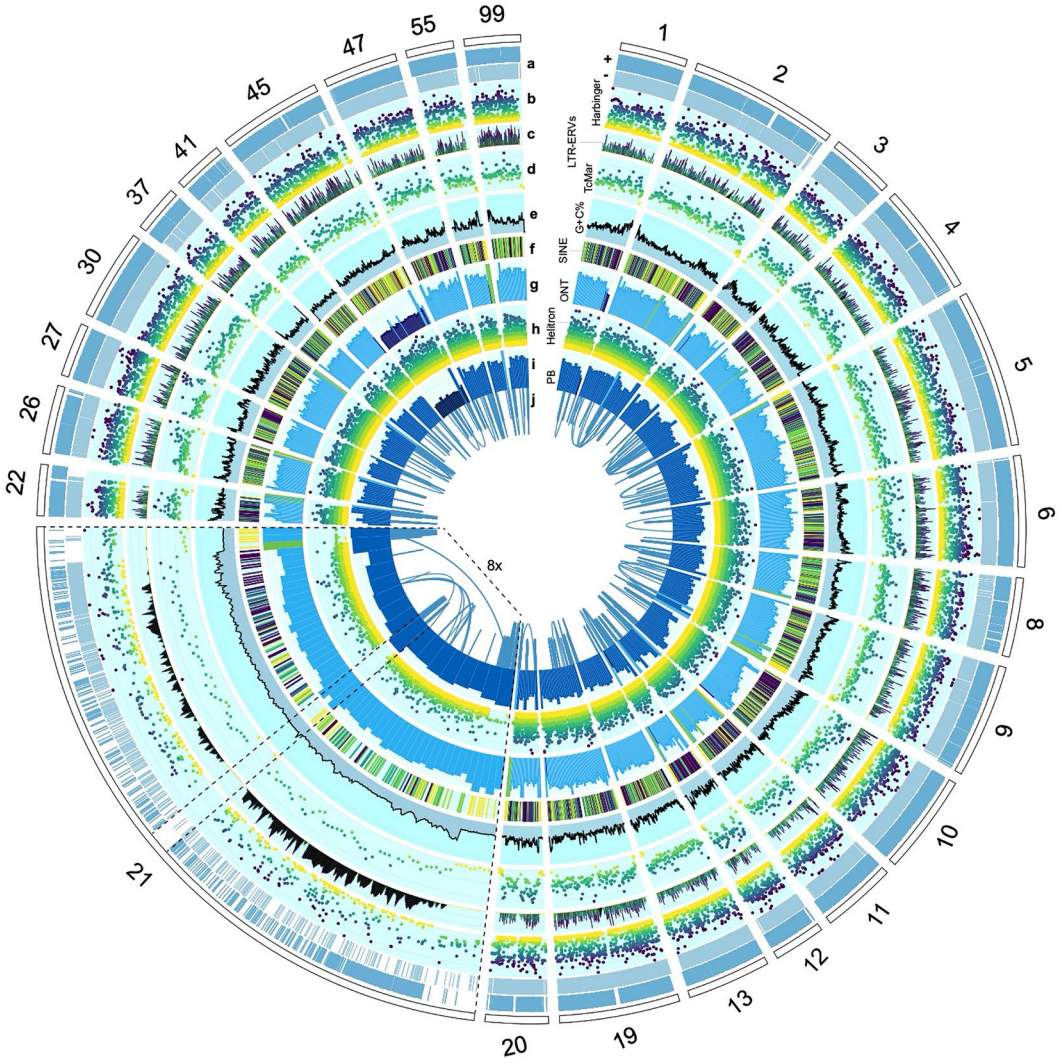

Fig. 2 Genome map showing sequencing read coverage and annotations for the new *de novo*, near-T2T male *G. beringei beringei* genome (gberber-nt2t). The 25 largest contigs from the pseudohaplotype assembly are shown. An example T2T-sequenced chromosome is displayed (chromosome 17, contig ptg000021l) at 8× zoom with the telomeres and centromere regions, identifiable by having high coverage and low gene/LTR/SINE content, bounded in dashed lines. Circos tracks represent (**a**) plus (top) and minus (bottom) strand *ab initio* gene predictions. (**b**) Harbinger elements. (**c**) LTR/ERV repeats, where the y-axis = % divergence from the reference. (**d**) TcMar elements. Repeat element tracks have a y-axis ranging from zero to 50% divergence. (**e**) G + C % with a y-axis ranging from 20–80 G + C %. (**f**) SINE elements. (**g**) ONT read coverage (ave. = 52.87× ; see also Fig. 3). (**h**) Helitron elements. (**i**) PacBio HiFi long-read coverage (ave. = 26.77×). (**j**) Extended tracks of high homology/similarity (i.e., identical long repeat sequences). Circos data tracks and configurations are in Dataset S4.

T2T ape chrom	Primary			haplotype 1			haplotype 2		
	# of contigs covered	Coverage (bp)	% of reference	# of contigs	Coverage (bp)	% of reference covered	# of contigs	Coverage (bp)	% of reference covered
chr1	4	237,707,687	95.24%	14	180,940,045	72.49%	4	231,546,192	92.77%
chr2A	1	137,378,233	90.72%	10	112,909,482	74.56%	1	132,178,282	87.29%
chr2B	1	136,062,712	90.56%	4	125,838,800	83.75%	6	128,478,292	85.51%
chr3	2	200,939,185	93.43%	2	177,378,364	82.47%	5	195,416,211	90.86%
chr4	3	204,062,820	96.47%	3	193,087,438	91.28%	5	196,609,431	92.94%
chr5	3	185,082,376	95.53%	6	161,707,587	83.46%	7	152,792,470	78.86%
chr6	1	190,508,696	95.52%	12	158,865,908	79.65%	1	187,323,059	93.92%
chr7	4	170,487,976	99.82%	10	137,133,811	80.29%	5	165,073,671	96.65%
chr8	1	155,108,187	93.96%	1	149,650,825	90.66%	4	154,695,553	93.71%
chr9	2	124,320,708	90.56%	2	122,796,978	89.45%	6	86,568,840	63.06%
chr10	1	141,277,892	93.21%	6	132,036,850	87.12%	1	140,296,339	92.57%
chr11	2	138,838,189	93.25%	4	134,675,392	90.46%	5	115,150,124	77.34%
chr12	1	136,227,336	88.33%	1	136,261,006	88.35%	4	136,381,984	88.43%
chr13	2	116,586,957	82.81%	5	101,931,986	72.40%	7	108,227,660	76.87%
chr14	4	103,190,271	70.76%	4	98,010,849	67.21%	8	100,928,580	69.21%
chr15	1	89,872,724	84.96%	2	84,764,302	80.13%	10	72,861,251	68.87%
chr16	2	124,758,120	95.44%	11	63,326,728	48.44%	3	112,952,899	86.41%
chr17	1	109,562,186	97.19%	10	91,383,663	81.06%	1	106,719,769	94.67%
chr18	1	102,543,862	88.36%	1	100,177,088	86.32%	2	75,947,298	65.44%
chr19	3	61,716,027	76.84%	5	34,808,876	43.34%	6	46,182,474	57.50%
chr20	1	64,674,541	76.55%	5	46,429,018	54.95%	2	65,036,363	76.98%
chr21	1	46,498,953	78.14%	2	30,311,096	50.94%	1	41,473,323	69.69%
chr22	1	45,547,538	83.16%	2	17,052,716	31.14%	1	34,484,037	62.96%
chrX	4	160,481,058	90.38%	5	112,678,151	63.46%	3	54,936,868	30.94%
chrY	3	36,331,738	53.90%	2	11,160,400	16.56%	3	34,553,210	51.26%
Average	**2**	**128,790,639**	**87.80%**	**5.16**	**108,612,694**	**71.60%**	**4.04**	**115,072,567**	**77.79%**

Table 2. Mapping statistics of *G. beringei beringei* chromosomes to a reference T2T Gorilla gorilla assembly[9].

placed in a cooler with ice packs and transported to the RAB - One Health Laboratory, Rubirizi Station, Rwanda for further analysis.

DNA isolation and sequencing. For genomic DNA isolation, white blood cells were isolated from 5 ml of blood samples 30 hours post-collection, using a red blood cell (RBC) lysis solution (Qiagen, Hilden, Germany) and several cycles of centrifugation (2000 g at 4 °C)/washes in phosphate-buffered saline (PBS, pH 7.4). Peripheral blood mononuclear cells (PMBCs) were also isolated from 5 ml of blood using the standard Percoll ® (GE Healthcare, Chicago, USA) centrifugation method. Briefly, plasma, PBMCs, and RBCs are separated following a gentle centrifugation at 350 g at room temperature, and PBMCs are washed several times with PBS to remove RBC contaminants. To isolate high molecular weight (HMW) DNA, two different kits were employed according to the manufacturers' instructions: NucleoBond HMW DNA (Macherey-Nagel, Dueren, DE) or Monarch HMW DNA Extraction kit for tissue (New England Biolabs, Ipswich, MA, USA). DNA quality was evaluated using Qubit fluorometric quantitation (Thermo Fisher Scientific, Waltham, MA, USA) and a DNA Fragment Analyzer (5200 fragment Analyzer System, Agilent, Santa Clara, CA, USA) with an Agilent HS Genomics Kit (Agilent, DNF-468 HS Genomic DNA 50 kb Kit).

From blood samples, we generated 4,892,167 HiFi reads; of which, 3,860,717 were 'covered' and 2,111,243 were 'chained'. For HiFi library preparation and sequencing, we followed the protocol described by PacBio (https://www.pacb.com/wp-content/uploads/Procedure-checklist-Preparing-whole-genome-and-metagenome-libraries-using-SMRTbell-prep-kit-3.0.pdf) with a few modifications: (i) we used Megaruptor 3 at a slower speed that gives a longer fragment length distribution: specifically, we shear once at setting 28 to set the mode length, then immediately repeat the shear program at setting 31 to reduce the long tail of molecules; (ii) we used the DNeasy PowerClean Pro cleanup kit (Qiagen, Hilden, Germany) to achieve high-purity DNA; and (iii) after library preparation, we performed a strict size selection with the PippinHT instrument (SageScience, Beverly, MA, USA) using a high-pass cutoff of 13 kbp with a 30-minute elution time (and the longer program "0.75% 9–30 kb R + T 75E") for removal of short fragments. We used a total of 4 SMRT Cells 8 M on a Sequel IIe PacBio instrument.

For ONT libraries preparation and sequencing, we followed the protocol described on ONT community (https://community.nanoporetech.com/docs/prepare/library_prep_protocols/ligation-sequencing-v14-human-cfdna-multiplex-sqk-nbd114-24/v/cfm_9208_v114_revb_15may2024). Quality control of gorilla genomic DNA was performed using a Fragment Analyzer with the Agilent HS Genomics Kit (Agilent, DNF-468 HS Genomic DNA 50 kb Kit). Library preparation was conducted with approximately 200 fmol of DNA using the Ligation Sequencing Kit V14 (LSK114, ONT) following the high duplex ligation protocol. The final libraries

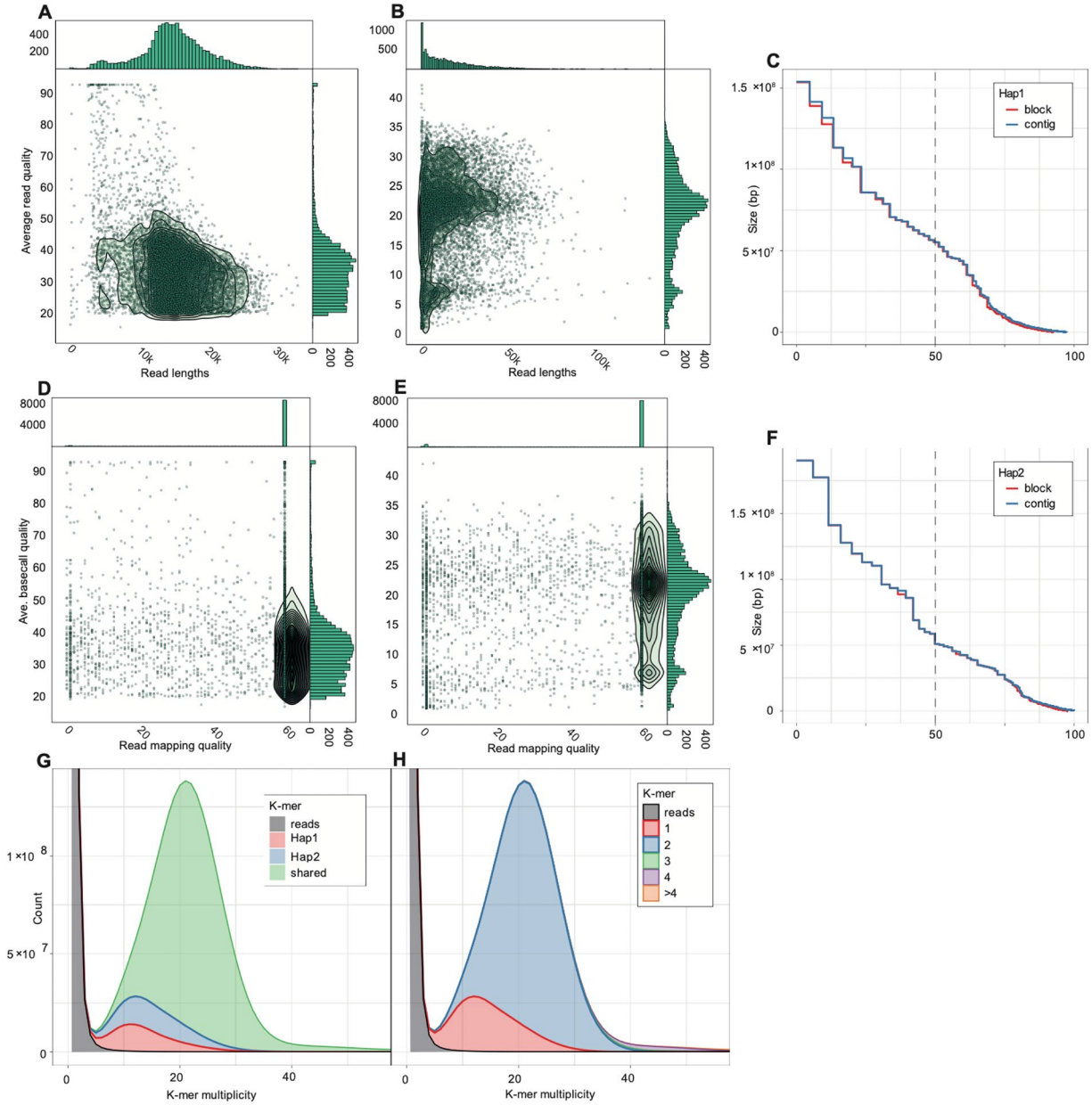

Fig. 3 Quality assessment metrics for the phased mountain gorilla genome assembly. These results demonstrate the assembly quality and haplotype resolution across both haplotypes (Hap1 and Hap2). See Fig. 4 and Table 2 for mapping of this assembly to a recent T2T *G. gorilla* genome[9]. (**A,B**): Sequencing read lengths (x-axis) compared to the average read quality (y-axis) for (**A**) PacBio HiFi and (**B**) ONT reads. (**C,F**): NG graphs: Bottom panels present NG (length-weighted median) curves for both haplotypes. The x-axis represents NG percentages, while the y-axis shows contig/block sizes. Separate curves for contigs (blue) and phased blocks (orange) are plotted. The high NG50 values and gradual curve slopes suggest excellent contiguity and completeness for both haplotypes. (**D,E**): Read mapping quality (QV, x-axis) compared to the average base-call quality for (**D**) PacBio HiFi and (**E**) ONT reads. QV, calculated as −10 * log10 (error rate), quantifies base-level accuracy. (**G,H**): K-mer spectra: These plots display k-mer frequency distributions for the Hap1 and Hap2 haplotype assemblies. The x-axis represents k-mer multiplicity (occurrence frequency), while the y-axis shows the count of distinct k-mers at each multiplicity. The bimodal distribution with peaks at 1× and 2× coverage indicates effective separation of haplotypes and high completeness.

were sequenced on a PromethION HD flow cell. We used the native barcoding Kit 24 V14, and to increase the coverage, we used four flow cells on a PromethION 2 solo instrument. Libraries were recovered after 24 hours and reloaded onto a washed flow cell for a total of three loadings over 72 hours. The total number of duplex reads was 2,479,196 (~89 gigabytes) and the simplex reads were 6,532,927 (~196 gigabytes). The ONT reads passing filtering were: 9,000,542 reads, of which 4,861,406 were fully corrected and 154,854 were nearly fully corrected. Overall, 1,033,384 ultra-long reads had full chains. Read and assembly QC metrics are included in the NanoPlot[12] and Merqury[13,14] analyses in the supplementary datasets.

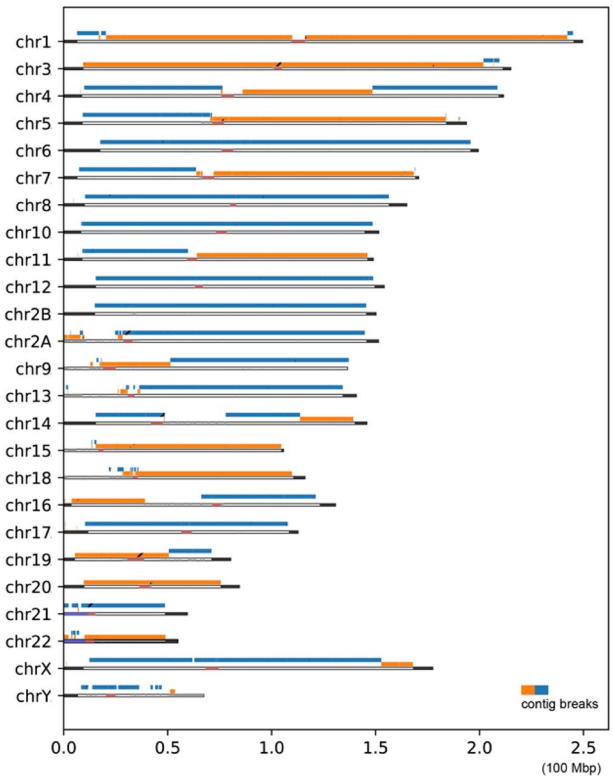

Fig. 4 Mapping of the *G. beringei beringei* contigs to a recently published T2T *G. gorilla* genome[9]. The *G. beringei beringei* assembly aligns with an average of 2 contigs mapping per chromosome, with incomplete coverage of some satellite repeat regions (Table 2). The change of contigs is indicated by a transition of orange and blue. The T2T genomes at the bottom of each row are annotated by the following color scheme: red: centromere, purple: acrocentric p-arm, dark gray: pCht satellites, light gray: other satellites. Figure S1 shows a zoomed-in view of Chr 1 and Table S1 shows detailed mapping statistics.

Genome assembly. We used hifiasm[15] (version 0.19.8-r603, https://github.com/chhylp123/hifiasm), a tool designed specifically for assembling long reads, to assemble the mountain gorilla genome[16]. This approach allowed us to leverage the high accuracy and long read lengths of HiFi and ONT sequencing data to generate highly contiguous assemblies, achieving an average of 2 non-scaffolded contigs covering each chromosome of a reference *Gorilla* assembly[9]. Recent gorilla assembly projects by Yoo *et.al.*[9] and Makova *et.al.*[17] used VERKKO for de novo assembly. Hifiasm uses a hybrid assembly approach that integrates HiFi and ultra-long ONT reads in a graph-based framework to achieve highly contiguous and complete near T2T assemblies. VERKKO combines ultra-long ONT reads and HiFi data in a hierarchical, haplotype-aware strategy to resolve complex regions and produce phased assemblies. VERKKO is known to erroneously assemble extra non-biological sequences (https://github.com/marbl/verkko/issues/59) and result in lower assembly base accuracy in some instances[18]. Thus, our decision to use Hifiasm was based on its proven strong performance, especially in the absence of Hi-C data. The command was 'hifiasm -o assembly.asm -t 32–ul ONT_reads.fq HiFi_reads.fq'. Hifiasm generates three main outputs: two haplotype assemblies (hap1 and hap2) and a pseudohaplotype assembly in graph format (.gfa) which is converted to fasta with 'awk '/^S/{print "＞"$2;print $3}' $LINE ＞ "${LINE%.gfa}".fa.' We tested all possible kmer configurations, where i = $(sed -n "$SLURM_ARRAY_TASK_ID"p ilist.txt) \\ hifiasm -o./k-mer-cycle/gorilla-kmer-"$i".asm -t 7 gorilla.fq -k "$i", and found the default k-mer assignment to be optimal for contiguity. After initial contig assembly where hifiasm assembles all reads into contigs without separating haplotypes, the heterozygous SNPs are used to phase these contigs into two sets, creating the haplotype-resolved assemblies Hap1 and Hap2, which represent the two separate chromosome sets of the diploid genome[16]. Then, the best contig for each region is selected, based on length, quality, and consistency with surrounding areas to produce the combined, main assembly. Hifiasm uses sequences from either haplotype to fill gaps or resolve conflicts, resulting in a single, haploid-like (pseudohaplotype) representation of the genome optimized for contiguity.

Genome quality assessment. To evaluate the quality and completeness of our assembly, we employed several tools: QUAST for evaluating assembly contiguity, Merqury[13,14] for k-mer-based quality assessment, SAMtools (https://github.com/samtools) for alignment statistics, NanoPlot[12,19] for long-read sequencing and assembly metrics, BUSCO[20,21] for assessing the completeness of conserved genes, and QUAST[22] to evaluate assembly contiguity and generate standard assembly statistics. These analyses were carried out in various conda environments (see 'analysis_dependencies.tar.gz')[16] and primarily performed on the pseudohaplotype reference assembly, providing a comprehensive assessment of its quality and completeness.

Comprehensive sequencing read, assembly, and mapping statistics were generated using NanoPlot[16]. The ONT dataset comprised 9,304,794 reads, totaling 155,097,187,007 bases, of which 150,069,675,303 bases (96.8%) were successfully aligned to the assembly. These reads demonstrated good length and moderate accuracy, with a read length N50 of 29,503 bp and a mean read length of 16,668.5 bp (standard deviation: 16,164.8 bp). The reads exhibited an average identity of 95.4% when aligned to the assembly, with a median identity of 98.5%. Read quality was variable, with a mean read quality score of 12.5 and a median of 21.4. In terms of quality distribution, 85.4% of reads (7,948,633) surpassed Q10, 76.8% (7,144,515) exceeded Q15, 59.6% (5,548,344) achieved Q20 or higher, 22.2% (2,066,884) reached Q25, and 6.3% (587,910) attained Q30 or above. This substantial set of long ONT reads complemented the HiFi dataset, contributing to the generation of our near-T2T *G. beringei beringei* assembly. The consistently high QV scores across multiplicities, with the majority of k-mers showing QV > 60, indicate superior sequence quality and low error rates in both haplotypes[16].

In total, PacBio HiFi reads ($n = 4{,}916{,}103$) had a read length N50 of 16,319 bp, with 74,713,988,666 bases of which 74,625,957,927 were aligned to the genome (99.9%). The reads showed an average of 99.4% identity to the assembly. The mean read length was 15,197.8, the mean read quality was 27.5, and the standard deviation of the read length was 4,775. Nearly all reads (99.7%, 4,914,842 reads) were higher than Q20, 82.4% (4,052,820 reads) were higher than Q25, and 63.8% of reads had read quality scores higher than Q30 (3,136,324 reads). The large number of high-quality HiFi and ONT reads formed the basis of the near-T2T *G. beringei beringei* assembly.

We assessed the quality of our PacBio HiFi and ONT assembled mountain gorilla genome[16] using Merqury[13,14], a reference-free method for evaluating genome assemblies. Merqury utilizes k-mer-based methods to estimate the completeness, consensus quality, and base-level accuracy of the assembly. We analyzed two haplotype assemblies (Hap1 and Hap2) separately. Meryl (https://github.com/marbl/meryl/releases/tag/v1.4.1) was used to generate k-mer databases from Illumina sequencing reads, which were then compared to the genome assemblies. This facilitated the assessment of assembly completeness, consensus quality, and error rates through downstream analyses with Merqury. The analysis provided key metrics for assessing assembly quality from long reads, including quality value (QV), error rate, number of bases, and estimated number of errors. For Hap1, we obtained a QV of 65.1013 and an error rate of 3.0894×10^{-7}. Hap2 showed a QV of 65.1958, an error rate of 3.02288×10^{-7}. These results indicate high-quality assemblies for both haplotypes, with QV scores above 60 suggesting a high degree of base-level accuracy. The slightly lower QV and higher error rate in Hap2 may be attributed to its larger size, potentially incorporating more complex or repetitive regions of the genome. The complete Merqury output, including detailed statistics for individual contigs, was examined to identify any specific regions of lower quality that may require further attention in the assembly process.

The completeness of the gorilla genome assemblies was with regard to complete universally conserved ortholog marks was assessed[16] using BUSCO v5.7.0[20] with the primates_odb10 lineage dataset (created on 2024-01-08, comprising 13,780 BUSCOs from 25 genomes). The analysis was performed in eukaryotic genome mode using miniprot[23] as the gene predictor. Results indicated a high level of completeness in the main, merged, genome, with 98.4% of BUSCOs identified as complete (97.2% single-copy and 1.2% duplicated). Only 1.1% of BUSCOs were fragmented, and 0.5% were missing. These findings suggest that the assembled genome captures the majority of expected single copy orthologs for primates, indicating a high-quality and nearly complete assembly.

Comparison with the T2T gorilla genome was performed by aligning each of the assemblies to the T2T assembly (mGorGor1) using minimap2 (v2.26)[24]; parameters used are as follows: "*minimap2 -c -L --eqx --cs -cx asm20 --secondary = no --eqx -Y*". We used all maternal chromosomes and Y chromosome as the reference[16]. The total number of contigs that mapped by alignment length >1 Mbp were counted, and the number of reference bases covered by the whole-genome alignment were quantified for each chromosome to assess relative completeness and contiguity. The alignment visualization was performed using SafFire (https://github.com/mrvollger/SafFire) and SVbyEye (https://github.com/daewoooo/SVbyEye). The shown breakpoints are regions of discontinuity and not misassembled regions. It is not known whether they correspond to structural variants, but given the sparsity and regions of the breaks, most gaps represent discontinuity at the centromere (e.g., Chr 4,5,7,11, and 13). In addition to the high coverage previously noted, we note, however that the acrocentric p-arms (11.9 Mbp or 53.2% covered) as well as satellite sequence corresponding to the centromeres and subterminal heterochromatic caps are not yet completely assembled (29 identified out of 42 expected) in this version of the mountain gorilla genome.

Genomic landscape of repetitive elements. For the annotation of transposable elements, we implemented a multifaceted approach. HiTE (https://github.com/CSU-KangHu/HiTE/tree/master), a fast and accurate dynamic boundary adjustment tool, was used for full-length transposable element detection and annotation in our genome assembly[16]. We ran RepeatMasker[25] (https://www.repeatmasker.org/) with RMblast (https://www.repeatmasker.org/rmblast/) for comprehensive repeat element identification and masking. Our findings revealed that 34.8% (1,232,080,350 bp) of the genome is composed of repetitive elements, with total interspersed repeats accounting for 29.17% (1,032,857,806 bp) of the genome sequence. Retroelements were the most abundant class of repetitive sequences, occupying 26.58% (941,219,695 bp) of the genome. This class primarily consisted of long interspersed nuclear elements (LINEs, 14.06%), short interspersed nuclear elements (SINEs, 7.70%), and long terminal repeat (LTR) elements (4.82%). LINEs, particularly L1/CIN4 elements, dominated the retroelement landscape, potentially influencing genomic plasticity and adaptive processes[26–28]. SINEs, known to play roles in gene regulation and genome evolution[29–31], were also prevalent. LTR elements, including a substantial proportion of retroviral elements (4.53% of the genome), may contribute to genomic diversity and potentially impact immune responses. In summary, the repeat elements outlined here may influence gene regulation, genomic plasticity, and the evolution of physiological traits necessary for survival in low-oxygen environments. For instance, the abundance of LINEs and SINEs might facilitate the evolution of genes involved in oxygen transport or utilization, similar to adaptations observed in other high-altitude species[32,33].

Gene prediction and transcript and protein annotation. We performed *ab initio* gene prediction on the new mountain gorilla assemblies[16], as well as four other gorilla genome assemblies downloaded from NCBI (including one assembly consisting of solely the Y chromosome from *G. beringei beringei*). This was done using SNAP and TransDecoder. SNAP[34] (Semi-HMM-based Nucleic Acid Parser) is a general-purpose gene finding program based on a generalized hidden Markov model, offering flexibility, speed, and accuracy in predicting gene structures across various organisms. TransDecoder, designed to identify coding regions within transcript sequences, excels at detecting open reading frames (ORFs) with high coding potential. TransDecoder's ability to integrate protein homology evidence and identify multiple ORFs from alternatively spliced transcripts complements SNAP's *ab initio* approach. We used SNAP with default parameters and the 'mam39.hmm' hidden Markov model and the latest TransDecoder Docker instance in Singularity (docker://trinityrnaseq/transdecoder) with default parameters to generate *ab initio* coding sequence and protein predictions.

Data visualization. A suite of visualization tools was used to represent different aspects of the *G. beringei beringei* genome assembly. Mercury (https://github.com/marbl/merqury) was used to generate k-mer spectra plots and QV distribution graphs, providing visual insights into assembly completeness and accuracy. QUAST[22] produced contiguity plots and size distribution charts, illustrating the assembly's structural characteristics. NanoPlot created read length histograms, quality score distributions, and alignment identity plots for both PacBio HiFi and ONT reads. HiTE (https://github.com/CSU-KangHu/HiTE/tree/master) and RepeatMasker (https://www.repeatmasker.org/) with RMblast (https://www.repeatmasker.org/rmblast/) generated visualizations of transposable element distributions and divergence landscapes across the genome.

To provide a comprehensive overview of the genome assembly and its features, we created a Circos (circos.ca) diagram using the base Circos package[35] and Circosviz[36]. This circular plot displays various genomic elements of the near-T2T male *G. beringei beringei* genome, focusing on the 25 largest contigs, which comprise approximately 80% of the full genome. The Circos plot integrates multiple tracks of genomic information: (A) *ab initio* gene predictions on both plus (outer) and minus (inner) strands, (B) G + C content percentage (range: 20–80%), (C) LTR/ERV repeats and (D) SINE elements, both showing percentage divergence from the reference (range: 0–50%), (E) ONT read coverage, and (F) PacBio HiFi long-read coverage, with both coverage tracks ranging from 0–50×. This comprehensive visualization allows for an intuitive understanding of the genome's structure, gene distribution, repetitive element landscape, and sequencing coverage (see Fig. S2 for a workflow diagram of the methods used to generate and validate the genome assembly).

Data Records

The data associated with this project are available at the NCBI Sequencing Read Archive (SRA, Bioprojects PRJNA1214581 [ONT reads][37] and PRJNA1214493: [HiFi reads][38] and Zenodo ([ONT reads][39] and [HiFi reads][40]). The genome assemblies are available at Genbank (NCBI, Bioprojects PRJNA1242239 [haplotype 1][41], PRJNA1242299 [haplotype 2][42], and PRJNA1242300 [pseudohaplotype][43]. The assemblies, annotations, and genome analysis results (BUSCO, NanoPlot, Merqury, Circos, SafFire) are available at Zenodo[16].

Technical Validation

The quality of the *G. beringei beringei* assembly was rigorously evaluated using multiple complementary approaches to ensure high accuracy, completeness, and contiguity. We assessed the base-level accuracy and completeness using Merqury, which revealed high QV for both haplotypes: the Hap1 assembly had a QV of 65.1013 and an error rate of 3.0894×10^{-7}. Hap2 had a QV of 65.1958 and an error rate of 3.02288×10^{-7}. The assemblies' contiguity were evaluated using QUAST[22], showing a contig N50 of approximately 95 Mbp for the combined pseudohaplotype (3,540,458,497 bp), and 56.5 Mbp (3.1 Gbp) and 51.0 Mbp (3.2 Gbp) for the Hap1 and Hap2 haplotype assemblies. These values represent a substantial improvement over most available primate genomes.

Further validation was performed by aligning the sequencing reads back to the assembled genome. The PacBio HiFi reads showed an average of 99.4% identity to the assembly, with 99.88% of bases successfully mapped, while 96.8% of ONT read bases were successfully aligned with an average identity of 95.4%. The high quality of the input sequencing data, with 100% of PacBio HiFi reads surpassing Q20 and 63.8% achieving Q30 or higher, further supports the reliability of our assembly. Collectively, these comprehensive validation metrics indicate that our *G. beringei beringei* genome assembly is of high accuracy, completeness, and contiguity, providing a reliable resource for future genetic and evolutionary studies of this endangered species.

Finally, the completeness of the gorilla genome assembly regarding universally conserved, single-copy markers was assessed using BUSCO v5.7.0[20] with the primates_odb10 lineage dataset (created on 2024-01-08, comprising 13,780 BUSCOs from 25 genomes). The analysis was performed in eukaryotic genome mode using miniprot[23] as the gene predictor. Results indicated a high level of completeness, with 98.4% of BUSCOs identified as complete (97.2% single-copy and 1.2% duplicated). Only 1.1% of BUSCOs were fragmented, and 0.5% were missing. These findings suggest that the assembled genome captures the majority of expected single-copy orthologs for primates, indicating a high-quality and nearly complete (T2T) assembly.

Code availability

All scripts used in this work are hosted as supplementary datasets at Zenodo[16].

Received: 28 October 2024; Accepted: 28 April 2025;
Published online: 22 May 2025

References

1. Hickey, J. R. *et al*. Gorilla beringei ssp. beringei, Mountain Gorilla. *The IUCN Red List of Threatened Species*. T39999A17989719 (2018).
2. Scally, A. *et al*. Insights into hominid evolution from the gorilla genome sequence. *Nature* **483**, 169–175, https://doi.org/10.1038/nature10842 (2012).
3. Xue, Y. *et al*. Mountain gorilla genomes reveal the impact of long-term population decline and inbreeding. *Science* **348**, 242–245 https://doi.org/10.1126/science.aaa3952 (2015).
4. Prado-Martinez, J. *et al*. Great ape genetic diversity and population history. *Nature* **499**, 471–475, https://doi.org/10.1038/nature12228 (2013).
5. Nurk, S. *et al*. The complete sequence of a human genome. *Science* **376**, 44–53, https://doi.org/10.1126/science.abj6987 (2022).
6. Logsdon, G. A., Vollger, M. R. & Eichler, E. E. Long-read human genome sequencing and its applications. *Nat Rev Genet* **21**, 597–614, https://doi.org/10.1038/s41576-020-0236-x (2020).
7. Miga, K. H. *et al*. Telomere-to-telomere assembly of a complete human X chromosome. *Nature* **585**, 79–84 (2020).
8. Kronenberg, Z. N. *et al*. High-resolution comparative analysis of great ape genomes. *Science* **360**, https://doi.org/10.1126/science.aar6343 (2018).
9. Yoo, D. *et al*. Complete sequencing of ape genomes. *bioRxiv*, 2024.2007.2031.605654 https://doi.org/10.1101/2024.07.31.605654 (2024).
10. Wenger, A. M. *et al*. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat Biotechnol* **37**, 1155–1162, https://doi.org/10.1038/s41587-019-0217-9 (2019).
11. Jain, M. *et al*. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat Biotechnol* **36**, 338–345, https://doi.org/10.1038/nbt.4060 (2018).
12. De Coster, W. & Rademakers, R. NanoPack2: population-scale evaluation of long-read sequencing data. *Bioinformatics* **39**, https://doi.org/10.1093/bioinformatics/btad311 (2023).
13. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol* **21**, 245, https://doi.org/10.1186/s13059-020-02134-9 (2020).
14. https://github.com/marbl/merqury.
15. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature Methods* **18**, 170–175 (2021).
16. Nelson, D. R. *et al*. Supplemental datasets for 'A near telomere-to-telomere phased reference assembly for the male mountain gorilla (Gorilla beringei beringei)'. *Zenodo* https://doi.org/10.5281/zenodo.12631956 (2025).
17. Makova, K. D. *et al*. The complete sequence and comparative analysis of ape sex chromosomes. *Nature* **630**, 401–411, https://doi.org/10.1038/s41586-024-07473-2 (2024).
18. Yu, W. *et al*. Comprehensive assessment of 11 de novo HiFi assemblers on complex eukaryotic genomes and metagenomes. *Genome Res* **34**, 326–340, https://doi.org/10.1101/gr.278232.123 (2024).
19. https://github.com/wdecoster/NanoPlot.
20. Simao, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212, https://doi.org/10.1093/bioinformatics/btv351 (2015).
21. https://github.com/metashot/busco.
22. https://github.com/ablab/quast.
23. Li, H. Protein-to-genome alignment with miniprot. *Bioinformatics* **39** https://doi.org/10.1093/bioinformatics/btad014 (2023).
24. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100, https://doi.org/10.1093/bioinformatics/bty191 (2018).
25. Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr Protoc Bioinformatics* Chapter 4, Unit 4 10, https://doi.org/10.1002/0471250953.bi0410s05 (2004).
26. Cordaux, R. The human genome in the LINE of fire. *Proc Natl Acad Sci USA* **105**, 19033–19034, https://doi.org/10.1073/pnas.0810202105 (2008).
27. Han, K. *et al*. Genomic rearrangements by LINE-1 insertion-mediated deletion in the human and chimpanzee lineages. *Nucleic Acids Res* **33**, 4040–4052, https://doi.org/10.1093/nar/gki718 (2005).
28. Lee, J. *et al*. Different evolutionary fates of recently integrated human and chimpanzee LINE-1 retrotransposons. *Gene* **390**, 18–27, https://doi.org/10.1016/j.gene.2006.08.029 (2007).
29. Ray, D. A. *et al*. Differential SINE evolution in vesper and non-vesper bats. *Mob DNA* **6**, 10, https://doi.org/10.1186/s13100-015-0038-4 (2015).
30. Shimamura, M., Nikaido, M., Ohshima, K. & Okada, N. A SINE that acquired a role in signal transduction during evolution. *Mol Biol Evol* **15**, 923–925, https://doi.org/10.1093/oxfordjournals.molbev.a025997 (1998).
31. Suh, A. *et al*. De-novo emergence of SINE retroposons during the early evolution of passerine birds. *Mob DNA* **8**, 21, https://doi.org/10.1186/s13100-017-0104-1 (2017).
32. Storz, J. F. & Bautista, N. M. Altitude acclimatization, hemoglobin-oxygen affinity, and circulatory oxygen transport in hypoxia. *Mol Aspects Med* **84**, 101052, https://doi.org/10.1016/j.mam.2021.101052 (2022).
33. Storz, J. F. & Signore, A. V. Introgressive Hybridization and Hypoxia Adaptation in High-Altitude Vertebrates. *Front Genet* **12**, 696484, https://doi.org/10.3389/fgene.2021.696484 (2021).
34. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5**, 59, https://doi.org/10.1186/1471-2105-5-59 (2004).
35. http://circos.ca/.
36. https://github.com/MadsAlbertsen/multi-metagenome/tree/master/circosviz.
37. *European Nucleotide Archive*. http://identifiers.org/insdc.sra:SRP559342 (2025).
38. *European Nucleotide Archive*. http://identifiers.org/insdc.sra:SRP559291 (2025).
39. Nelson, D. R., Salehi-Ashtiani, K. S.-A. & Twizere, J. C. A near telomere-to-telomere phased reference assembly for the male mountain gorilla (Gorilla beringei beringei) -Oxford Nanopore reads. *Zenodo* https://doi.org/10.5281/zenodo.12633417 (2025).
40. Nelson, D. R., Salehi-Ashtiani, K. S.-A. & Twizere, J. C. A near telomere-to-telomere phased reference assembly for the male mountain gorilla (Gorilla beringei beringei) - Pacbio HIFI reads. *Zenodo*. https://doi.org/10.5281/zenodo.12634531 (2025).
41. Nelson, D. R., Salehi-Ashtiani, K. S.-A. & Twizere, J. C. *A near telomere-to-telomere phased reference assembly for the male mountain gorilla (Gorilla beringei beringei) Genbank* http://identifiers.org/insdc:JBMMNA000000000 (2025).
42. Nelson, D. R., Salehi-Ashtiani, K. S.-A. & Twizere, J. C. *A near telomere-to-telomere phased reference assembly for the male mountain gorilla (Gorilla beringei beringei) Genbank* http://identifiers.org/insdc:JBMMNB000000000 (2025).
43. Nelson, D. R., Salehi-Ashtiani, K. S.-A. & Twizere, J. C. *A near telomere-to-telomere phased reference assembly for the male mountain gorilla (Gorilla beringei beringei) Genbank* http://identifiers.org/insdc:JBMMNC000000000 (2025).
44. Mao, Y. *et al*. Structurally divergent and recurrently mutated regions of primate genomes. *bioRxiv* https://doi.org/10.1101/2023.03.07.531415 (2023).

Acknowledgements
We thank the Rwanda Development Board (RDB) for authorizing this study, and the Oxford Nanopore Technologies (ONT) for their support under ORG.one project. We thank Tonia Brown for her help in copy editing the manuscript. We also thank the Genomics platform of the GIGA Institute (University of Liege), and Genomics Core of the Center for Human Genetics (KU Leuven). This work was supported by the Académie de Recherche et d'Enseignement Supérieur – Commission de la Coopération au Développement (ARES-CCD, Belgium), ARES-PRD2024-Twizere grant and fellowships to R.M and J-C.Tumushime. This research was also supported by NYUAD Faculty Research Funds (AD060) from the Division of Science (NYUAD, UAE). J-C. Twizere is a Senior Investigator of the Fonds de la Recherche Scientifique (FRS-FNRS). This research was supported, in part, by funding from the National Institutes of Health (NIH) grant R01 HG002385 to E.E.E. E.E.E. is an investigator of the Howard Hughes Medical Institute.

Author contributions
D.R.N.: Experimental design, data analysis, writing - review and editing. R.M.: Sample collection, review and editing, and supervision. K.M.H.: Data analysis. J.-C.Tum.: Sample collection, review and editing. G.N.: Sample collection. J.N.: Sample collection, review and editing, and supervision. W.M.: Methodology high-molecular-weight isolation, library preparation, and PacBio HiFi sequencing. L.K.: High-molecular-weightprepartion, library preparation, and Oxford Nanopore Technologies (ONT) sequencing. W.C.: ONT Data analysis, and supervision. K.M.M.: Methodology, review and editing. D.Y.: Data analysis, review and editing. E.E.E.: Data analysis, review and editing, supervision, funding acquisition. K.S.A.: Experimental design, review and editing, supervision, funding acquisition. J.-C.Tw.: Experimental design, review and editing, supervision, funding acquisition.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-025-05114-5.

Correspondence and requests for materials should be addressed to D.R.N., K.S.-A. or J.-C.T.