

UC Davis

UC Davis Previously Published Works

Title

De Novo Genome Assembly for the Coppery Titi Monkey (*Plecturocebus cupreus*): An Emerging Nonhuman Primate Model for Behavioral Research

Permalink

<https://escholarship.org/uc/item/89p8h9xj>

Journal

Genome Biology and Evolution, 16(5)

ISSN

1759-6653

Authors

Pfeifer, Susanne P

Baxter, Alexander

Savidge, Logan E

et al.

Publication Date



2024-05-02

DOI

10.1093/gbe/evae108

Peer reviewed

De Novo Genome Assembly for the Coppery Titi Monkey (*Plecturocebus cupreus*): An Emerging Nonhuman Primate Model for Behavioral Research

Susanne P. Pfeifer ^{1,2,*}, Alexander Baxter^{3,4}, Logan E. Savidge^{3,4}, Fritz J. Sedlazeck ⁵, and Karen L. Bales^{3,4,6}

¹School of Life Sciences, Arizona State University, Tempe, AZ, USA

²Center for Evolution and Medicine, Arizona State University, Tempe, AZ, USA

³Department of Psychology, University of California, Davis, CA, USA

⁴California National Primate Research Center, Neuroscience and Behavior Division, Davis, CA, USA

⁵Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX, USA

⁶Department of Neurobiology, Physiology, and Behavior, University of California, Davis, CA, USA

*Corresponding author: E-mail: susanne@spfeiferlab.org.

Accepted: May 09, 2024

Abstract

The coppery titi monkey (*Plecturocebus cupreus*) is an emerging nonhuman primate model system for behavioral and neurobiological research. At the same time, the almost entire absence of genomic resources for the species has hampered insights into the genetic underpinnings of the phenotypic traits of interest. To facilitate future genotype-to-phenotype studies, we here present a high-quality, fully annotated de novo genome assembly for the species with chromosome-length scaffolds spanning the autosomes and chromosome X (scaffold N50 = 130.8 Mb), constructed using data obtained from several orthologous short- and long-read sequencing and scaffolding techniques. With a base-level accuracy of ~99.99% in chromosome-length scaffolds as well as benchmarking universal single-copy ortholog and *k*-mer completeness scores of >99.0% and 95.1% at the genome level, this assembly represents one of the most complete Pitheciidae genomes to date, making it an invaluable resource for comparative evolutionary genomics research to improve our understanding of lineage-specific changes underlying adaptive traits as well as deleterious mutations associated with disease.

Key words: coppery titi monkey, *Plecturocebus cupreus*, Pitheciidae, platyrrhine, primate, hybrid assembly.

Significance

This study presents the first fully annotated, chromosome-level de novo genome assembly for the coppery titi monkey (*Plecturocebus cupreus*)—a critical step toward improving the genomic resources for this emerging nonhuman primate model system for behavioral and neurobiological research.

Introduction

Understanding the factors at play during behavioral development is one of the key challenges to improve our understanding of human health and disease. Of particular

interest to the behavioral research community are coppery titi monkeys (*Plecturocebus cupreus*; previously *Callicebus cupreus*; Groves 2005)—one of currently 34 recognized species of *Plecturocebus* (Byrne et al. 2016). Coppery titi

© The Author(s) 2024. Published by Oxford University Press on behalf of Society for Molecular Biology and Evolution.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact reprints@oup.com for reprints and translation rights for reprints. All other permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site—for further information please contact journals.permissions@oup.com.

monkeys live in long-term, socially monogamous (pair-bonded) family units that include the parents and up to three generations of their offspring (Kinzey 1997) and that are characterized by a high level of paternal care (Mendoza and Mason 1986; Valeggia et al. 1999). Hence, this species provides researchers with an opportunity to study complex social behaviors in the context of monogamy—an extremely rare trait among mammalian taxa (Lukas and Clutton-Brock 2013)—in particular with regards to the neurobiology underlying pair-bonding, mate-guarding, and male parenting (Bales et al. 2007, 2017, 2021). In addition, coppery titi monkeys have been utilized to study cognition, including decision-making, problem-solving, and memory, as well as the neurological mechanisms underlying vocal communication as part of their social interactions (Bales et al. 2017; Lau et al. 2020).

Yet, despite its usage as a primate model for neuroendocrine and behavioral research, genomic resources remain limited for the species, prohibiting researchers from investigating the genetic underpinnings of the phenotypic traits studied in experimentally controlled settings. To enable future genotype-to-phenotype studies, we here provide a high-quality, fully annotated, hybrid de novo genome assembly for the coppery titi monkey, *PleCup_hybrid*. By providing insights into the genomic makeup of coppery titi monkeys—a representative of the Pitheciidae family which remains relatively poorly characterized genetically—this resource serves as a foundation for future studies to advance our understanding of primate evolution on this branch of the phylogenetic tree specifically as well as behavioral research, population genetic studies, and primate comparative evolutionary analyses in general.

Materials and Methods

Sample Collection

Peripheral blood samples were collected from an adult male coppery titi monkey born and housed at the California National Primate Research Center (CNPRC). Postmortem tissue samples were collected from two additional individuals, including samples from the adrenal gland, brain, heart, lung, ovaries, pancreas, and pituitary gland from an adult female and a testes sample from an adult male. These samples were collected within 24 h of each subject's death (which was unrelated to the present study) and were immediately stored at -80°C . At a later date, the tissues were thawed to -20°C and approximately 10 mg of tissue was sectioned using a cryostat and collected for this study.

Sample Preparation and Sequencing

The de novo genome assembly was generated using several orthologous sequencing technologies: Oxford Nanopore Technologies (ONT) long-read sequencing, 10x Genomics linked-read sequencing, Illumina short-read sequencing, and

Dovetail Genomics Hi-C sequencing. In addition, RNA from eight different tissues was sequenced to facilitate genome annotation.

Long-Read Sequencing

For long-read sequencing, high-molecular weight (HMW) genomic DNA (gDNA) was extracted from peripheral blood using the Qiagen MagAttract HMW DNA Kit (#67563) following the manufacturer's instructions (Qiagen, Hilden, Germany) and utilizing a Thermomixer C (Eppendorf, Hamburg, Germany) with a 1.5-ml tube adapter for the mixing steps. DNA was eluted off the beads using Qiagen Elution Buffer (#1014609), and a 1× AMPure bead cleanup was performed to purify MagAttract-extracted gDNA, following a 3-pass 26-gauge needle shear. Sheared gDNA was further purified with the Pacific Biosciences SRE XL Kit (Pacific Biosciences, Menlo Park, CA, USA) to remove small, and retain larger, fragments. Quantity and quality of the gDNA were assessed using a Qubit (Invitrogen, Carlsbad, CA, USA) and Femto Pulse Fragment Analyzer (Agilent #FP-1002-0275; Agilent Technologies, Palo Alto, CA, USA), before preparing a sequencing library using the Oxford Nanopore Ligation Sequencing Kit (SQK-LSK109). This library was then sequenced on two R9 flow cells on a PromethION platform (Oxford Nanopore Technologies, Oxford, UK) to $\sim 25\times$ coverage.

Linked-Read Sequencing

For linked-read sequencing, HMW gDNA was extracted from peripheral blood following the protocol described above (see Long-Read Sequencing). Next, a library of Genome Gel Beads was combined with HMW gDNA in Master Mix and partitioning oil to create Gel Bead-In-Emulsions (GEMs) in the microfluid Genome Chip which contained millions of copies of 10× barcode primers. After completion of the preloaded program in the 10× Chromium Controller, the GEMs were transferred from the chip to polymerase chain reaction (PCR) tubes where they underwent 3 h of GEM isothermal incubation, followed by a post-GEM incubation cleanup. Thereby, silane magnetic beads were used to remove any leftover biochemical reagents from the post-GEM reaction mixture and Solid Phase Reversible Immobilization beads were used to optimize the appropriate DNA size range for Illumina library preparation. A sequencing library was prepared by end-repairing DNA, followed by A-tailing, adaptor ligation, and amplification, and sequenced (2×150 bp) on an Illumina NovaSeq 6000 (Illumina, San Diego, CA, USA) to $\sim 50\times$ coverage.

Short-Read Sequencing

For short-read sequencing, HMW gDNA was extracted using the Genra Puregene Kit (Qiagen, Hilden, Germany) and

sheared on an E220 Focused Ultrasonicator (Covaris, Woburn, MA, USA). DNA quantity was assessed using a Qubit (Invitrogen, Carlsbad, CA, USA) before preparing a sequencing library using the Kapa High Throughput Library Preparation Kit (Kapa Biosystems-Roche, Basel, Switzerland). The library was then paired-end sequenced (2 × 150 bp) on an Illumina NovaSeq 6000 (Illumina, San Diego, CA, USA) to ~65× coverage.

Hi-C Sequencing

For Hi-C sequencing, an Omni-C library (Dovetail Genomics, Scotts Valley, CA, USA) was prepared by fixing chromatin with formaldehyde in the nucleus prior to extraction. Fixed chromatin was digested with DNase I; chromatin ends were repaired and ligated to a biotinylated bridge adapter followed by proximity ligation of adapter containing ends. After proximity ligation, crosslinks were reversed and the DNA purified. Purified DNA was treated to remove biotin that was not internal to ligated fragments. Next, a sequencing library was generated using NEBNext Ultra enzymes (New England Biolabs, Ipswich, MA, USA) and Illumina-compatible adapters. Biotin-containing fragments were isolated using streptavidin beads before PCR enrichment of the library and sequencing on an Illumina HiSeqX (Illumina, San Diego, CA, USA) to ~30× coverage.

RNA Sequencing

Total RNA was extracted from eight tissue samples—adrenal gland, brain, heart, lung, ovaries, pancreas, pituitary gland, and testes—using the RNeasy Plus Universal Mini Kit and following the manufacturer's instructions (Qiagen, Hilden, Germany). RNA samples were quantified using a Qubit 2.0 Fluorometer (Life Technologies, Carlsbad, CA, USA), and RNA integrity was assessed using an Agilent TapeStation 4200 (Agilent Technologies, Palo Alto, CA, USA). Next, RNA sequencing libraries were prepared using the NEBNext Ultra RNA Library Prep Kit for Illumina following the manufacturer's instructions (NEB, Ipswich, MA, USA). Briefly, mRNAs were initially enriched with Oligo(dT) beads. Enriched mRNAs were fragmented for 15 min at 94 °C. First-strand and second-strand cDNAs were subsequently synthesized. cDNA fragments were end-repaired and adenylated at 3'-ends, and universal adapters were ligated to cDNA fragments, followed by index addition and library enrichment by PCR with limited cycles. The sequencing library was validated on an Agilent TapeStation (Agilent Technologies, Palo Alto, CA, USA) and quantified by using a Qubit 2.0 Fluorometer (Invitrogen, Carlsbad, CA, USA) as well as by quantitative PCR (KAPA Biosystems, Wilmington, MA, USA). The sequencing libraries were clustered on a single lane of a flow cell and paired-end sequenced (2 × 150 bp) on an Illumina HiSeq 4000 (Illumina, San Diego, CA, USA). Raw sequence data were demultiplexed using Illumina's

bcl2fastq software version 2.17, allowing one mismatch for index sequence identification.

Assembly

Raw ONT long-reads were validated using fastQValidator version 0.1.1a (https://genome.sph.umich.edu/wiki/FastQ_Validator), problematic reads were removed using BBMap version 38.87 (Bushnell 2014), and the quality of the remaining reads was assessed using NanoStat as embedded in pycQC version 2.5.0.23 (Leger and Leonardi 2019). An initial de novo assembly was generated from the high-quality long-reads using Flye version 2.8.1 (Kolmogorov et al. 2019) and polished using the Nanopore-based tools Racon version 1.4.13 (Vaser et al. 2017) and Medaka version 1.2.1 (<https://github.com/nanoporetech/medaka>).

To improve accuracy, the draft genome assembly was further polished using HyPo version 1.0.3 (Kundu et al., in preprint), together with high-quality 10x Genomics linked-reads and Illumina short-reads. In brief, linked-reads were mapped to the draft assembly using LongRanger version 2.2.2 (Marks et al. 2019), whereas short-reads were first quality and adapter trimmed using Trim Galore version 0.6.1 (<https://github.com/FelixKrueger/TrimGalore>) before mapping them to the draft assembly using minimap2 version 2.17 (Li 2018). Duplicates were removed using the MarkDuplicates tool implemented in the Genome Analysis Toolkit version 4.1.8.1 (van der Auwera and O'Connor 2020), and properly paired reads with a minimum mapping quality of 40 were extracted using SAMtools version 1.9 (Li et al. 2009). Using these read data sets as input, HyPo polishing of the draft assembly was performed, assuming an approximate genome size of 2.7 Gb and an approximate mean coverage of 36× and 48× for 10x Genomics linked-reads and Illumina short-reads, respectively.

The polished draft de novo assembly was combined with Dovetail Omni-C library reads to scaffold the assembly using HiRise—a software pipeline designed for scaffolding genome assemblies using long-range proximity ligation data (Putnam et al. 2016). In brief, Dovetail Omni-C library sequences were aligned to the draft input assembly using BWA version 0.7.17 (Li and Durbin 2009). Next, separations of Dovetail Omni-C read pairs that mapped within draft scaffolds with a minimum mapping quality of 50 were analyzed by HiRise to produce a likelihood model for genomic distance between read pairs. This model was then used to identify and break putative misjoins, to score prospective joins, and make new joins.

Lastly, the assembly was screened for contaminations using the National Center for Biotechnology Information (NCBI) Foreign Contamination Screen tool suite (<https://github.com/ncbi/fcs>), resulting in the exclusion of four scaffolds of bacterial or viral origin that ranged from 318 to 6,495 bp in size.

Annotation

Repeat Annotation

Repeat families were identified de novo and classified using the software package RepeatModeler version 2.0.1 (Flynn et al. 2020), with RECON version 1.08 (Bao and Eddy 2002) and RepeatScout version 1.0.6 (Price et al. 2005) embedded within. Based on the species-specific repeat model obtained from RepeatModeler, repeats in the assembly were masked using RepeatMasker version 4.1.0 (<https://repeatmasker.org>). Variant telomeric repeats were characterized using Computel version 1.2 (Nersisyan and Arakelyan 2015), with a genome length of 2,764,079,242 bp (“-lgenome 2764079242”).

Gene Annotation

Publicly available coding sequences from human (genome assembly: GRCh38.p13; Schneider et al. 2017), bonobo (Mhudiblu_PPA_v2; Mao et al. 2021), rhesus macaque (rheMac10; Warren et al. 2020), olive baboon (Panu_3.0; Rogers et al. 2019), and common marmoset (mCalJac1; Yang et al. 2021) were used to train an ab initio species-specific hidden Markov model for *P. cupreus* using both the gene prediction software AUGUSTUS version 2.5.5 (Stanke et al. 2008) with six rounds of prediction optimization and the gene finder SNAP version 2006-07-28 (Korf 2004). In addition, RNA sequencing reads obtained from the adrenal gland, brain, heart, lung, ovaries, pancreas, pituitary gland, and testes were mapped onto the genome using STAR version 2.7 (Dobin et al. 2013) to provide information about the location of exons, introns, and exon–intron junctions using the *bam2hints* tool integrated within AUGUSTUS. To improve the accuracy of gene boundaries, ab initio models from AUGUSTUS and SNAP were then used together with the intron–exon boundary hints to predict genes in the repeat-masked genome, retaining only those genes predicted by both tools. The quality of the gene prediction was assessed using annotation edit distance (AED) scores generated for each of the predicted genes as part of the MAKER pipeline version 3.01.03 (Cantarel et al. 2008; Campbell et al. 2014). To help guide the annotation process, Swiss-Prot peptide sequences from the UniProt database were downloaded and used in conjunction with the protein sequences from *Homo sapiens*, *Pan paniscus*, *Macaca mulatta*, *Papio anubis*, and *Callithrix jacchus* to generate peptide evidence in MAKER. Genes were further characterized for their putative function by performing a BLAST (Altschul et al. 1990) search of the peptide sequences against the UniProt database.

tRNA Annotation

tRNAs were predicted using the software tRNAscan-SE version 2.05 (Chan et al. 2021).

Quality Assessment

The genome assembly was evaluated using three characteristics: contiguity, completeness, and correctness. To obtain summary statistics and assess contiguity, QAST version 5.0.2 (Mikheenko et al. 2018) was used to calculate the sequence length of the shortest contig at 50% and 90% of the total genome length (N50 and N90, respectively) as well as the smallest number of contigs whose combined lengths are at least 50% and 90% of the genome size (L50 and L90, respectively). compleasm version 0.2.6 (Huang and Li 2023) was used to assess genome and transcriptome completeness of the assembly by searching for highly conserved single-copy orthologous genes previously identified in eukaryotes, mammals, and primates (i.e. utilizing the eukaryota_odb10 data set containing 255 BUSCOs from 70 species, the mammalia_odb10 data set containing 9,226 BUSCOs from 24 species, and the primates_odb10 data set containing 13,780 BUSCOs from 25 species, respectively; Manni et al. 2021). Lastly, correctness was evaluated by comparing the *k*-mers in the genome assembly to those observed in the unassembled, highly accurate short-reads obtained from the same individual using Merqury version 1.3 (Rhie et al. 2020), together with a *k*-mer database of the recommended size ($k = 21$) built from the Illumina short-reads using Meryl version 1.4 (<https://github.com/marbl/meryl>).

Genome Sequence Comparison with Human

To gain a better understanding of genome evolution, a genome comparison was performed between the de novo genome assembly for *P. cupreus* (PleCup_hybrid) and the telomere-to-telomere human genome assembly (T2T-CHM13v2.0 downloaded from NCBI RefSeq; Nurk et al. 2022) using MUMmer version 3.23 (Kurtz et al. 2004). Following Kurtz et al. (2004), nucleotide alignments of the chromosome-length scaffolds were performed using the *nucmer* program with a minimum match length of 25 (“-l 25”) to encompass matching exons. To account for introns, neighboring matches were joined into clusters if gaps between exact matches were no longer than 3 kb (“-g 3000”). As matching genes are expected to be longer, the maximum-length colinear chain of matches was required to be at least 100 nucleotides (“-c 100”). Next, the longest increasing subset of alignments was calculated and filtered (using *delta-filter*) to maintain alignments with a minimum length of 1 kb (“-l 1000”) and a minimum percent identity of 85% (“-l 85”) to reduce poor alignments. Syntenic regions were plotted in R version 4.3.0 using the *circlize* package (Gu et al. 2014).

Results and Discussion

The genome of the coppery titi monkey was sequenced from peripheral blood DNA isolated from a male individual

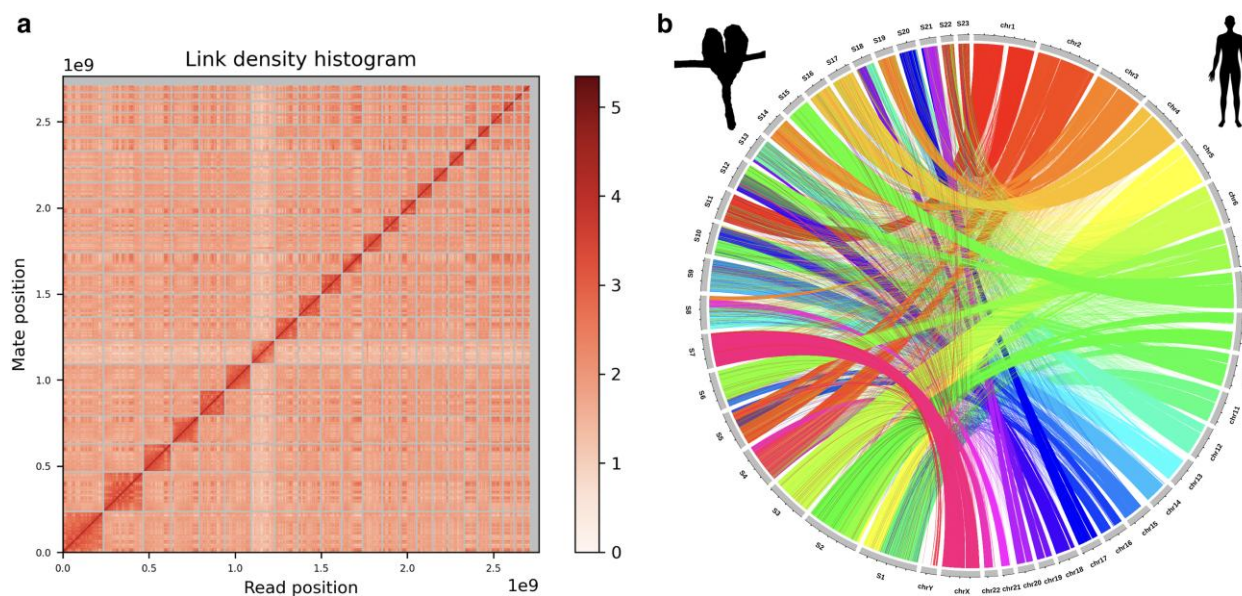


FIG. 1.—De novo genome assembly for the coppery titi monkey (*P. cupreus*). a) Chromatin interaction map of the de novo genome assembly for the coppery titi monkey (*P. cupreus*), PleCup_hybrid, generated in this study. Chromosome-length scaffolds are represented as squares along the diagonal, ordered by size from bottom left to top right. b) Genome comparison between the coppery titi monkey (with chromosome-length scaffolds [S] depicted on the left-hand side) and human (with chromosomes [chr] depicted on the right-hand side). Syntenic regions are color-coded by human chromosome. Silhouette images were obtained from PhyloPic (*H. sapiens* by Katy Lawler; CC BY 4.0 DEED) and generated by S.P.P. (*P. cupreus*).

housed at the CNPRC and de novo assembled using a combination of orthologous technologies—single-molecule long-read ONT sequencing, 10x Genomics linked-read sequencing, and Illumina short-read sequencing, as well as a chromosome-contact map from Dovetail Genomics—to generate chromosome-scale scaffolds. In brief, the initial draft assembly was constructed from high-quality long-reads, sequenced on two R9 PromethION flow cells. Specifically, after removing 456 problematic reads with a length smaller than 10 bp from the flow cell readout, the remaining 3,412,944 reads, exhibiting a mean length of 20.7 kb (read N50: 36.0 kb) and totaling 25.6-fold (X) whole-genome sequencing coverage, were assembled into a draft genome using Flye (Kolmogorov et al. 2019). To further improve accuracy, the draft assembly, containing 3,030 contigs (contig N50: 18.1 Mb), was first polished using the Nanopore-based tools Racon (Vaser et al. 2017) and Medaka (<https://github.com/nanoporetech/medaka>), before performing a final polish using HyPo (Kundu et al., in preprint) together with whole-genome 150 bp paired-end 10x Genomics linked-reads (sequenced to 52.5x coverage) and 150 bp paired-end Illumina short-reads (65.4x coverage). Contigs were then scaffolded using long-range Dovetail proximity ligation data (31.1x coverage) by making 720 new joins and breaking 9 misjoins, and contaminations were removed using the NCBI Foreign Contamination Screen tool suite (<https://github.com/ncbi/fcs>).

The final de novo assembly of the coppery titi monkey exhibits a total length of 2.76 Gb—similar to the genome sizes

reported for other Pitheciidae for which short-read (150 bp paired-end Illumina) assemblies were recently released (ranging from 2.61 Gb for the Aracá uakari [*Cacajao ayresii*] to 2.74 Gb for the red-nosed bearded saki [*Chiropotes albinasus*]; Kuderna et al. 2023)—and comprises 1,288 scaffolds with a scaffold N50 of 130.8 Mb. In concordance with previous cytogenetic studies based on morphology from classical staining, chromosome banding, and chromosome flow sorting that reported a diploid karyotype of $2n = 46$ in the species with a submetacentric X chromosome and a very small acrocentric Y chromosome (de Boer 1974; Benirschke and Bogart 1976; Dumas et al. 2005), 23 scaffolds spanning the autosomes and chromosome X were observed that contained 98.2% of the assembly (Fig. 1a). The quality of the highly contiguous hybrid assembly for this behavioral model system is comparable to several recently released primate genome assemblies, including that of the common marmoset (*C. jacchus*)—a platyrrhine species frequently used for neurological and regenerative medicinal research (Okano et al. 2012; Kishi et al. 2014)—comprising 1,233 scaffolds with a scaffold N50 of 137.0 Mb (Yang et al. 2021), and that of rhesus macaque (*M. mulatta*)—the most widely used nonhuman primate model in biomedical research (Rogers 2022)—comprising 2,979 scaffolds, with a scaffold N50 of 82.4 Mb (Warren et al. 2020).

A total of 38.71% of the genome is repetitive, with 34.38% of regions corresponding to class I transposable element (TE) repeats, 2.22% to class II TE repeats, 1.15% to simple repeats, and 0.26% to low complexity repeats. Similar to humans (Coleman et al. 1999; Varley et al.

Table 1

Contiguity and completeness of the coppery titi monkey genome assembly

Contiguity		Completeness		
Genome length	2,764,079,242 bp	No. of genes	40,825	
No. of scaffolds	1,288	Complete BUSCOs (C)	255 (100.0%) ^d	Eukaryota^a
No. of contigs ≥ 1 kb	1,276	Complete and single-copy BUSCOs (S)	237 (92.9%)	
No. of gaps	720	Complete and duplicated BUSCOs (D)	18 (7.1%)	
No. of N's per 100 kb	2.74	Fragmented BUSCOs (F)	0 (0.0%)	
		Missing BUSCOs (M)	0 (0.0%)	
Scaffold N50	130,776,705 bp	Complete BUSCOs (C)	9,204 (99.8%) ^d	Mammalia^b
Scaffold N90	73,180,778 bp	Complete and single-copy BUSCOs (S)	9,094 (98.6%)	
Scaffold L50	9	Complete and duplicated BUSCOs (D)	110 (1.2%)	
Scaffold L90	20	Fragmented BUSCOs (F)	10 (0.1%)	
Largest scaffold	235,103,535 bp	Missing BUSCOs (M)	12 (0.1%)	
GC content	40.79%	Complete BUSCOs (C)	13,684 (99.3%) ^d	Primates^c
		Complete and single-copy BUSCOs (S)	13,531 (98.2%)	
		Complete and duplicated BUSCOs (D)	153 (1.1%)	
		Fragmented BUSCOs (F)	31 (0.2%)	
		Missing BUSCOs (M)	65 (0.5%)	

^aBased on the eukaryota_odb10 data set (containing 255 BUSCOs from 70 species). ^bBased on the mammalia_odb10 data set (containing 9,226 BUSCOs from 24 species). ^cBased on the primates_odb10 data set (containing 13,780 BUSCOs from 25 species). ^dAt the transcript level: 80.8%, 70.2%, and 64.7% for eukaryota, mammalia, and primates, respectively.

2002; Lee et al. 2014), the majority of telomeric repeats in the coppery titi monkey genome consists of noncoding t-type (TTAGGG) repeats (96.4%), with additional c-, g-, and j-type telomeric variant repeats present (TCAGGG: 0.3%, TGAGGG: 0.1%, and TTGGGG: 0.1%, respectively).

After masking repeat regions, protein-coding and non-coding genes were annotated using a species-specific ab initio prediction model, based on coding sequences obtained from high-quality genome assemblies of primates spanning hominids (human [*H. sapiens*] and bonobo [*P. paniscus*]), catarhines (rhesus macaque and olive baboon [*P. anubis*]), and platyrrhines (common marmoset), and further refined using transcriptome (RNA sequencing) data from eight different tissues—adrenal gland, brain, heart, lung, ovaries, pancreas, pituitary gland, and testes (~53 to 59 million reads per tissue)—obtained from two titi monkeys from the CNPRC. Overall, the coppery titi monkey genome assembly contains 40,825 gene models (total predicted coding sequence: 54,347,078 bp)—similar to those predicted in common marmosets (43,572 gene models) based on the distribution of exons, introns, and intergenic regions as well as transcriptional, translational, and splicing signals using the GENSCAN software (Burge and Karlin 1997). Out of these models, 20,523 genes exhibit an AED (Eilbeck et al. 2009; Yandell and Ence 2012) of 0.5 or smaller, with an average length of 1,331 bp—similar to the total number of protein-coding genes observed in rhesus macaque (20,389 and 20,605 genes with an average length of 1,564 and 1,489 bp for Chinese and Indian rhesus macaques, respectively; He et al. 2019).

Searching for highly conserved single-copy orthologous genes (BUSCOs) previously identified in eukaryotes, mammals,

and primates (Manni et al. 2021) highlighted that the coppery titi monkey assembly is nearly complete at the genome and transcriptome level (Table 1). To further evaluate the accuracy and completeness of the coppery titi monkey assembly, a comparison of *k*-mers (i.e. genomic regions of length *k*) in the de novo genome assembly with those observed in the high-accuracy Illumina short-reads obtained from the same individual was performed. This comparison showed that the coppery titi monkey genome exhibits a high base-level accuracy with a mean consensus quality of 39.7 in chromosome-length scaffolds, corresponding to a ~99.99% accuracy. Moreover, a *k*-mer completeness score of 95.1% suggests that only a small number of heterozygous variants are missing from the de novo genome assembly (supplementary fig. S1, Supplementary Material online). Finally, a whole-genome alignment between the 23 chromosome-length *P. cupreus* scaffolds and the chromosomes in the human telomere-to-telomere assembly (Nurk et al. 2022) revealed fundamental genetic similarities and large-scale shared sequence homology between the two species (Fig. 1b). This result is consistent with a previously published comparative cytogenetic study (Dumas et al. 2005) and further highlights the importance of this emerging non-primate model system.

Conclusion

The high-quality, highly contiguous, fully annotated genome assembly presented here is a critical step toward improving the coppery titi monkey as a behavioral model system. By facilitating genomic analyses, this novel resource will open new avenues of biomedical and behavioral research, enabling the

design of gene expression studies as well as the direct investigation of the genetic underpinnings of phenotypic traits of interest. In addition, insights into the content and structure of the coppery titi monkey genome will facilitate the discovery of naturally occurring genetic variation and population genetic analyses in the species and, as one of the most complete Pitheciidae genomes to date, will provide a novel comparative evolutionary perspective on the genetic divergence among primate lineages, important to advance our understanding of primate evolution.

Supplementary Material

Supplementary material is available at *Genome Biology and Evolution* online.

Acknowledgments

Oxford Nanopore and 10x Genomics sequencing were carried out at the Cold Spring Harbor Laboratory Genome Center. Illumina sequencing was carried out at the DNA Technologies and Expression Analysis Core at the UC Davis Genome Center (supported by NIH Shared Instrumentation Grant 1S10OD010786-01). RNA sequencing was carried out at Genewiz. Hi-C sequencing, scaffolding of the polished draft de novo assembly, and annotation were carried out at Dovetail Genomics. Computations were performed on Arizona State University's High Performance Cluster. The authors would like to thank Philipp Rescheneder at Oxford Nanopore Technologies (ONT) for the helpful feedback and discussions as well as Cyril Versoza for the help with the quality control of the raw ONT long-read data.

Funding

This work was supported by the California National Primate Research Center Pilot Program (NIH P51OD011107), the Dovetail Genomics Tree of Life Program, National Science Foundation CAREER award (DEB-2045343), and National Institute of General Medical Sciences of the National Institutes of Health ESI-MIRA award (R35GM151008) to S.P.P., as well as NIH R01HD092055 and MH125411 to K.L.B., and by the Good Nature Institute. The content is solely the responsibility of the authors and does not necessarily represent the official views of the funders.

Conflict of Interest

F.J.S. receives research support from Pacific Biosciences, Illumina, Genentech, and Oxford Nanopore Technologies.

Data Availability

This Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under the accession

JBDJOS000000000. The version described in this paper is version JBDJOS010000000. All sequence data has been deposited under NCBI BioProject PRJNA1068557.

Literature Cited

- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol.* 1990;215(3):403–410. [https://doi.org/10.1016/S0022-2836\(05\)80360-2](https://doi.org/10.1016/S0022-2836(05)80360-2).
- Bales KL, Ardekani CS, Baxter A, Karaskiewicz CL, Kuske JX, Lau AR, Savidge LE, Saylor KR, Witczak LR. What is a pair bond? *Horm Behav.* 2021;136:105062. <https://doi.org/10.1016/j.yhbeh.2021.105062>.
- Bales KL, Del Razo RA, Conklin QA, Hartman S, Mayer HS, Rogers FD, Simmons TC, Smith LK, Williams A, Williams DR, et al. Titi monkeys as a novel non-human primate model for the neurobiology of pair bonding. *Yale J Biol Med.* 2017;90(3):373–387.
- Bales KL, Mason WA, Catana C, Cherry SR, Mendoza SP. Neural correlates of pair-bonding in a monogamous primate. *Brain Res.* 2007;1184:245–253. <https://doi.org/10.1016/j.brainres.2007.09.087>.
- Bao Z, Eddy SR. Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Res.* 2002;12(8):1269–1276. <https://doi.org/10.1101/gr.88502>.
- Benirschke K, Bogart MH. Chromosomes of the tan-handed titi (*Callicebus torquatus*, Hoffmannsegg, 1807). *Folia Primatol.* 1976;25(1):25–34. <https://doi.org/10.1159/000155705>.
- Burge C, Karlin S. Prediction of complete gene structure in human genomic DNA. *J Mol Biol.* 1997;268(1):78–94. <https://doi.org/10.1006/jmbi.1997.0951>.
- Bushnell B. BMap: a fast, accurate, splice-aware aligner. Lawrence Berkeley National Laboratory; 2014. LBNL Report #: LBNL-7065E.
- Byrne H, Rylands AB, Carneiro JC, Alfaro JWL, Bertuol F, da Silva MNF, Messias M, Groves CP, Mittermeier RA, Farias I, et al. Phylogenetic relationships of the New World titi monkeys (*Callicebus*): first appraisal of taxonomy based on molecular evidence. *Front Zool.* 2016;13(1):10. <https://doi.org/10.1186/s12983-016-0142-4>.
- Campbell MS, Holt C, Moore B, Yandell M. Genome annotation and curation using MAKER and MAKER-P. *Curr Protoc Bioinformatics.* 2014;48(1):4.11.1–4.11.39. <https://doi.org/10.1002/0471250953.bi0411s48>.
- Cantarel BL, Korf I, Robb SMC, Parra G, Ross E, Moore B, Holt C, Sánchez Alvarado A, Yandell M. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res.* 2008;18(1):188–196. <https://doi.org/10.1101/gr.6743907>.
- Chan PP, Lin BY, Mak AJ, Lowe TM. tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. *Nucleic Acids Res.* 2021;49(16):9077–9096. <https://doi.org/10.1093/nar/gkab688>.
- Coleman J, Baird DM, Royle NJ. The plasticity of human telomeres demonstrated by a hypervariable telomere repeat array that is located on some copies of 16p and 16q. *Hum Mol Genet.* 1999;8(9):1637–1646. <https://doi.org/10.1093/hmg/8.9.1637>.
- de Boer LEM. Cytotaxonomy of the platyrrhini (primates). *Genen Phaenen.* 1974;17:1–155.
- Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics.* 2013;29(1):15–21. <https://doi.org/10.1093/bioinformatics/bts635>.
- Dumas F, Bigoni F, Stone G, Sineo L, Stanyon R. Mapping genomic rearrangements in titi monkeys by chromosome flow sorting and multidirectional in-situ hybridization. *Chromosome Res.* 2005;13(2):85–96. <https://doi.org/10.1007/s10577-005-7063-y>.

- Eilbeck K, Moore B, Holt C, Yandell M. Quantitative measures for the management and comparison of annotated genomes. *BMC Bioinformatics*. 2009;10(1):67. <https://doi.org/10.1186/1471-2105-10-67>.
- Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, Smit AF. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc Natl Acad Sci U S A*. 2020;117(17):9451–9457. <https://doi.org/10.1073/pnas.1921046117>.
- Groves CP. Species of *Callicebus* (*Callicebus cupreus*). In: Wilson DE, Reeder DM, editors. *Mammal species of the world: a taxonomic and geographic reference*. 3rd ed. Baltimore: Johns Hopkins University Press; 2005. p. 142–143.
- Gu Z, Gu L, Eils R, Schlesner M, Brors B. Circlize implements and enhances circular visualization in R. *Bioinformatics*. 2014;30(19):2811–2812. <https://doi.org/10.1093/bioinformatics/btu393>.
- He Y, Luo X, Zhou B, Hu T, Meng X, Audano PA, Kronenberg ZN, Eichler EE, Jin J, Guo Y, et al. Long-read assembly of the Chinese rhesus macaque genome and identification of ape-specific structural variants. *Nat Commun*. 2019;10(1):4233. <https://doi.org/10.1038/s41467-019-12174-w>.
- Huang N, Li H. Compleasm: a faster and more accurate reimplementa-tion of BUSCO. *Bioinformatics*. 2023;39(10):btad595. <https://doi.org/10.1093/bioinformatics/btad595>.
- Kinzey W. *New World primates: ecology, evolution, and behavior*. New York: Aldine de Gruyter; 1997.
- Kishi N, Sato K, Sasaki E, Okano H. Common marmoset as a new model animal for neuroscience research and genome editing technology. *Dev Growth Differ*. 2014;56(1):53–62. <https://doi.org/10.1111/dgd.12109>.
- Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using repeat graphs. *Nat Biotechnol*. 2019;37(5):540–546. <https://doi.org/10.1038/s41587-019-0072-8>.
- Korf I. Gene finding in novel genomes. *BMC Bioinformatics*. 2004;5(1):59. <https://doi.org/10.1186/1471-2105-5-59>.
- Kuderna LFK, Ulirsch JC, Rashid S, Ameen M, Sundaram L, Hickey G, Cox AJ, Gao H, Kumar A, Aguet F, et al. Identification of constrained sequence elements across 239 primate genomes. *Nature*. 2023;625(7996):735–742. <https://doi.org/10.1038/s41586-023-06798-8>.
- Kundu R, Casey J, Sung W-K. HyPo: super fast and accurate polisher for long-read genome assemblies. *BioRxiv*. <https://doi.org/10.1101/2019.12.19.882506>, 2019 preprint: not peer reviewed.
- Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL. Versatile and open software for comparing large genomes. *Genome Biol*. 2004;5(2):R12. <https://doi.org/10.1186/gb-2004-5-2-r12>.
- Lau AR, Clink DJ, Bales KL. Individuality in the vocalizations of infant and adult coppery titi monkeys (*Plecturocebus cupreus*). *Am J Primatol*. 2020;82(6):e23134. <https://doi.org/10.1002/ajp.23134>.
- Lee M, Hills M, Conomos D, Stutz MD, Dagg RA, Lau LMS, Reddel RR, Pickett HA. Telomere extension by telomerase and ALT generates variant repeats by mechanistically distinct processes. *Nucleic Acid Res*. 2014;42(3):1733–1746. <https://doi.org/10.1093/nar/gkt1117>.
- Leger A, Leonardi T. pycoQC, interactive quality control for Oxford Nanopore Sequencing. *J Open Source Softw*. 2019;4(34):1236. <https://doi.org/10.21105/joss.01236>.
- Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*. 2018;34(18):3094–3100. <https://doi.org/10.1093/bioinformatics/bty191>.
- Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*. 2009;25(14):1754–1760. <https://doi.org/10.1093/bioinformatics/btp324>.
- Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. The sequence alignment/map format and SAMtools. *Bioinformatics*. 2009;25(16):2078–2079. <https://doi.org/10.1093/bioinformatics/btp352>.
- Lukas D, Clutton-Brock TH. The evolution of social monogamy in mam-mals. *Science*. 2013;341(6145):526–530. <https://doi.org/10.1126/science.1238677>.
- Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. BUSCO up-date: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol Biol Evol*. 2021;38(10):4647–4654. <https://doi.org/10.1093/molbev/msab199>.
- Mao Y, Catacchio CR, Hillier LW, Porubsky D, Li R, Sulovari A, Fernandes JD, Montinaro F, Gordon DS, Storer JM, et al. A high-quality bonobo genome refines the analysis of hominid evolution. *Nature*. 2021;594(7861):77–81. <https://doi.org/10.1038/s41586-021-03519-x>.
- Marks P, Garcia S, Barrio AM, Belhocine K, Bernate J, Bharadwaj R, Bjornson K, Catalanotti C, Delaney J, Fehr A, et al. Resolving the full spectrum of human genome variation using linked-reads. *Genome Res*. 2019;29(4):635–645. <https://doi.org/10.1101/gr.234443.118>.
- Mendoza SP, Mason WA. Parental division of labour and differentia-tion of attachments in a monogamous primate (*Callicebus cu-preus*). *Anim Behav*. 1986;34(5):1336–1347. [https://doi.org/10.1016/S0003-3472\(86\)80205-6](https://doi.org/10.1016/S0003-3472(86)80205-6).
- Mikheenko A, Pribelski A, Saveliev V, Antipov D, Gurevich A. Versatile genome assembly evaluation with QAST-LG. *Bioinformatics*. 2018;34(13):i142–i150. <https://doi.org/10.1093/bioinformatics/bty266>.
- Nersisyan L, Arakelyan A. Computel: computation of mean telomere length from whole-genome next-generation sequencing data. *PLoS One*. 2015;10(4):e0125201. <https://doi.org/10.1371/journal.pone.0125201>.
- Nurk S, Koren S, Rhie A, Rautiainen M, Bizkadze AV, Mikheenko A, Vollger MR, Altemose N, Uralsky L, Gershman A, et al. The com-plete sequence of a human genome. *Science*. 2022;376(6588):44–53. <https://doi.org/10.1126/science.abj6987>.
- Okano H, Hikishima K, Iriki A, Sasaki E. The common marmoset as a novel animal model system for biomedical and neuroscience re-search applications. *Semin Fetal Neonatal Med*. 2012;17(6):336–340. <https://doi.org/10.1016/j.siny.2012.07.002>.
- Price AL, Jones NC, Pevzner PA. De novo identification of repeat fam-ilies in large genomes. *Bioinformatics*. 2005;21(Suppl 1):i351–i358. <https://doi.org/10.1093/bioinformatics/bti1018>.
- Putnam NH, O’Connell BL, Stites JC, Rice BJ, Blanchette M, Calef R, Troll CJ, Fields A, Hartley PD, Sugnet CW, et al. Chromosome-scale shot-gun assembly using an in vitro method for long-range linkage. *Genome Res*. 2016;26(3):342–350. <https://doi.org/10.1101/gr.193474.115>.
- Rhie A, Walenz BP, Koren S, Phillippy AM. Merqury: reference-free quality, completeness, and phasing assessment for genome as-semblies. *Genome Biol*. 2020;21(1):245. <https://doi.org/10.1186/s13059-020-02134-9>.
- Rogers J. Genomic resources for rhesus macaques (*Macaca mulatta*). *Mamm Genome*. 2022;33(1):91–99. <https://doi.org/10.1007/s00335-021-09922-z>.
- Rogers J, Raveendran M, Harris RA, Mailund T, Leppälä K, Athanasiadis G, Schierup MH, Cheng J, Munch K, Walker JA, et al. The compara-tive genomics and complex population history of *Papio* baboons. *Sci Adv*. 2019;5(1):eaau6947. <https://doi.org/10.1126/sciadv.aau6947>.
- Schneider VA, Graves-Lindsay T, Howe K, Bouk N, Chen H-C, Kitts PA, Murphy TD, Pruitt KD, Thibaud-Nissen F, Albracht D, et al. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly.

- Genome Res. 2017;27(5):849–864. <https://doi.org/10.1101/gr.213611.116>.
- Stanke M, Diekhans M, Baertsch R, Haussler D. Using native and syntetically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics*. 2008;24(5):637–644. <https://doi.org/10.1093/bioinformatics/btn013>.
- Valeggia CR, Mendoza SP, Fernandez-Duque E, Mason WA, Lasley B. Reproductive biology of female titi monkeys (*Callicebus moloch*) in captivity. *Am J Primatol*. 1999;47(3):183–195. [https://doi.org/10.1002/\(SICI\)1098-2345\(1999\)47:3<183::AID-AJP1>3.0.CO;2-J](https://doi.org/10.1002/(SICI)1098-2345(1999)47:3<183::AID-AJP1>3.0.CO;2-J).
- van der Auwera GA, O'Connor BD. *Genomics in the cloud: using Docker, GATK, and WDL in Terra*. 1st Edition. Sebastopol, CA: O'Reilly Media; 2020.
- Varley H, Pickett HA, Foxon JL, Reddel RR, Royle NJ. Molecular characterization of inter-telomere and intra-telomere mutations in human ALT cells. *Nat Genet*. 2002;30(3):301–305. <https://doi.org/10.1038/ng834>.
- Vaser R, Sović I, Nagarajan N, Šikić M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res*. 2017;27(5):737–746. <https://doi.org/10.1101/gr.214270.116>.
- Warren WC, Harris RA, Haukness M, Fiddes IT, Murali SC, Fernandes J, Dishuck PC, Storer JM, Raveendran M, Hillier LW, et al. Sequence diversity analyses of an improved rhesus macaque genome enhance its biomedical utility. *Science*. 2020;370(6523):eabc6617. <https://doi.org/10.1126/science.abc6617>.
- Yandell M, Ence D. A beginner's guide to eukaryotic genome annotation. *Nat Rev Genet*. 2012;13(5):329–342. <https://doi.org/10.1038/nrg3174>.
- Yang C, Zhou Y, Marcus S, Formenti G, Bergeron LA, Song Z, Bi X, Bergman J, Rousselle MMC, Zhou C, et al. Evolutionary and biomedical insights from a marmoset diploid genome assembly. *Nature*. 2021;594(7862):227–233. <https://doi.org/10.1038/s41586-021-03535-x>.

Associate editor: Adam Eyre-Walker