Title

Whole-genome sequence and assembly of the Javan gibbon (Hylobates moloch)

Journal

Journal of Heredity, 114(1)

ISSN

0022-1503

Authors

Escalona, Merly
VanCampen, Jake
Maurer, Nicholas W
et al.

Publication Date

2023-03-16

DOI

10.1093/jhered/esac043

Peer reviewed

Journal of Heredity, 2023, **114**, 35–43
https://doi.org/10.1093/jhered/esac043
Advance access publication 23 September 2022
Genome Resources

American Genetic Association

OXFORD

Genome Resources

Whole-genome sequence and assembly of the Javan gibbon (*Hylobates moloch*)

Merly Escalona[1], Jake VanCampen[2], Nicholas W. Maurer[1], Marina Haukness[1,3], Mariam Okhovat[2], Robert S. Harris[4], Allison Watwood[4], Gabrielle A. Hartley[5,6], Rachel J. O'Neill[5,6], Paul Medvedev[7,8,9,10], Kateryna D. Makova[4,7,8], Christopher Vollmers[1], Lucia Carbone[2,11,12,13,*], Richard E. Green[1,*]

[1]Department of Biomolecular Engineering, University of California–Santa Cruz, Santa Cruz, CA 95064, USA,
[2]Department of Medicine, Knight Cardiovascular Institute, Oregon Health and Science University, Portland, OR 97239, USA,
[3]University of California Santa Cruz Genomics Institute, Santa Cruz, CA 95064, USA,
[4]Department of Biology, Pennsylvania State University, University Park, PA, USA,
[5]Department of Molecular and Cell Biology, University of Connecticut, Storrs, CT 06296, USA,
[6]Institute for Systems Genomics, University of Connecticut, Storrs, CT 06296, USA,
[7]Center for Medical Genomics, Pennsylvania State University, University Park, PA, USA,
[8]Center for Computational Biology and Bioinformatics, Pennsylvania State University, University Park, PA, USA,
[9]Department of Computer Science and Engineering, Pennsylvania State University, University Park, PA, USA,
[10]Department of Biochemistry and Molecular Biology, Pennsylvania State University, University Park, PA, USA,
[11]Department of Molecular and Medical Genetics, Oregon Health and Science University, Portland, OR 97239, USA,
[12]Division of Genetics, Oregon National Primate Research Center, Beaverton, OR 97006, USA,
[13]Department of Informatics and Clinical Epidemiology, Oregon Health and Science University, Portland, OR 97239, USA.

*Corresponding authors: ed@soe.ucsc.edu; carbone@ohsu.edu

Corresponding Editor: Liliana Cortés-Ortiz

Abstract

The Javan gibbon, *Hylobates moloch*, is an endangered gibbon species restricted to the forest remnants of western and central Java, Indonesia, and one of the rarest of the Hylobatidae family. Hylobatids consist of 4 genera (*Holoock*, *Hylobates*, *Symphalangus*, and *Nomascus*) that are characterized by different numbers of chromosomes, ranging from 38 to 52. The underlying cause of this karyotype plasticity is not entirely understood, at least in part, due to the limited availability of genomic data. Here we present the first scaffold-level assembly for *H. moloch* using a combination of whole-genome Illumina short reads, 10X Chromium linked reads, PacBio, and Oxford Nanopore long reads and proximity-ligation data. This *Hylobates* genome represents a valuable new resource for comparative genomics studies in primates.

Key words: genome assembly, gibbon, Hi-C, long reads, proximity ligation

Introduction

The silvery or Javan gibbon, *Hylobates moloch* (Audebert 1798), is a small ape, specialized forest dweller that relies on closed canopy lowland evergreen forest (Andayani et al. 2001) and is restricted to small and isolated forest fragments in central and western Java, Indonesia. Like many of the gibbon species, *H. moloch* is endangered due to habitat loss and fragmentation, and the illegal pet trade (Andayani et al. 2001; Nijman 2015). Based on mitochondrial DNA (mtDNA) control-region sequence data analysis, the Javan gibbon is thought to have 2 genetically differentiated lineages: Western and Central (Andayani et al. 2001).

Species from the Hylobatidae family are endemic to the rainforests of Southeast Asia (Carbone et al. 2014; Veeramah et al. 2015). Among their features are suspensory bimanual brachiation (Reichard et al. 2016), social pair-bonding, and highly rearranged chromosomes relative to other members of the primate order (Dutrillaux et al. 1975). Most of these genomic rearrangements are specific to the Hylobatidae, differentiating them from the other members of the Hominoidea superfamily (Carbone et al. 2006, 2014).

The hylobatids are organized in 4 gibbon genera, which carry highly divergent karyotypes: *Nomascus* (crested gibbon) $2n = 52$, *Symphalangus* (siamang) $2n = 50$, *Hylobates* (Hylobates group) $2n = 44$, and *Hoolock* (hoolock gibbon) $2n = 38$ (Dutrillaux et al. 1975; Koehler et al. 1995; Mrasek et al. 2003; Carbone et al. 2006). Gibbon chromosomal structures also differ from that of their most recent common ancestor with humans, from which they diverged ~17 million years

Received May 10, 2022; Accepted September 8, 2022

ago. Thus, gibbons, which split from the human lineage after the Old World Monkeys did (Carbone et al. 2014), represent an important branch in the primate phylogeny. Although several studies shed light on the origin and mechanisms of the gibbon genome plasticity, there is only one assembled gibbon genome available, the *Nomascus leucogenys* (Asia_NLE_v1). Here, we present a scaffold-level genome assembly of *H. moloch* (HMol_V3) created using multiple sequencing technologies (i.e., short reads, linked reads, long reads, and high-throughput chromosome conformation capture data), resulting in high quality, completeness, and contiguity (Table 1). This assembly is a new genomic resource that will help us better understand the mechanisms underlying genome plasticity and help with Gibbon preservation efforts.

Methods

Biological materials

As described previously (Carbone et al. 2014), EBV-transformed cell lines were established from whole blood samples of an adult male Javan gibbon called Lionel at the Gibbon Conservation Center in Santa Clarita, CA. The blood was collected opportunistically during checkups and in agreement with protocols reviewed and approved by the Gibbon Conservation Center. Genome stability was established by karyotyping, which was carried out on metaphase chromosome spreads prepared per standard protocols. Briefly, slides were dehydrated in a 70%, 90%, and 100% ethanol row for 2 min each followed by air drying. Slides were stained with a 1:5 dilution of DAPI in Vectashield (Vector Laboratories, Inc., Newark, CA). Images were captured on an Olympus AX70 microscope and karyotyped using CytoVision software (Leica Biosystems Richmond, Inc., Richmond, IL) followed by manual review (Supplementary Fig. 1).

Nucleic acid library preparation

Illumina shotgun sequencing

We used a public shotgun whole-genome sequencing (WGS) dataset (Okhovat et al. 2020) that was generated from blood-derived genomic DNA from Lionel, as described in Carbone et al. (2014).

CHiCago library

One million Lionel lymphoblastoid cells were resuspended in cold PBS and crosslinked with 1% paraformaldehyde (Electron Microscopy Sciences [EMS], Hatfield, PA). Crosslinked chromatin was extracted via hypertonic buffer with 1% SDS and then bound to SPRI beads in 18% PEG-8000. Bead-bound chromatin was thoroughly washed and then digested with DpnII restriction enzyme (New England Biolabs; NEB, Ipswich, MA). Biotin-11-dCTP (ChemCyte, San Diego, CA) was incorporated by DNA Polymerase I, Klenow Fragment (NEB), and intra-aggregate ligation was achieved overnight by T4 DNA Ligase (NEB). Proximity-ligated DNA was isolated by SPRI bead purification after crosslink reversal with Proteinase K (Qiagen, Germantown, MD) in 8% SDS solution. Proximity-ligation products were randomly sheared to optimal Illumina library insert size via Diagenode Bioruptor NGS sonication platform before library preparation with the NEBNext Ultra II DNA Library Prep Kit for Illumina (NEB). Adaptor ligation products were SPRI bead purified before capturing biotinylated molecules using Dynabeads MyOne Streptavidin C1 beads (Thermo Fisher Scientific, Waltham, MA). After enrichment of biotinylated products by buffer exchange, indexing PCR using KAPA HiFi HotStart ReadyMix (Roche, Basel, Switzerland) and custom TruSeq indexing adaptors was performed using the streptavidin-bound DNA as input. The resulting library molecules were purified and simultaneously size-selected by SPRI bead cleanup. The resulting library was sent to Fulgent Genomics (Temple City, CA) for sequencing on an Illumina HiSeq 4000 platform (Illumina, San Diego, CA; 2x75bp).

Hi-C libraries

High-molecular-weight DNA was extracted from 5×10^6 lymphoblastoid cells using the Qiagen Blood and Cell Culture DNA Mini Kit, following the manufacturer's recommended protocol. DNA was processed as described in Dovetail Genomics' Chicago library preparation protocol (2017) with components from Active Motif's Chromatin Assembly Kit (Carlsbad, CA) and HeLa Core Histone product. Chromatin was assembled by first combining Active Motif's human histone chaperone NAP-1, HeLa Core Histones, and high salt buffer. After 15 min on ice, Active Motif's low salt buffer, ATP-utilizing chromatin assembly and remodeling factor (ACF), and creatine kinase-containing ATP regeneration system were added to the mixture. High-molecular-weight DNA was then added and incubated in this mixture for one hour at 27 °C. Crosslinking was achieved by adding paraformaldehyde (EMS) to a final concentration of 1% before binding the crosslinked chromatin to SPRI beads in 18% PEG-8000. Bead-bound chromatin was thoroughly washed and then digested with DpnII restriction enzyme (NEB). Biotin-11-dCTP (ChemCyte) was incorporated by DNA Polymerase I, Klenow Fragment (NEB), and intra-aggregate ligation was achieved overnight by T4 DNA Ligase (NEB). Proximity-ligated DNA was isolated by SPRI bead purification after crosslink reversal with Proteinase K (Qiagen) in an 8% SDS solution. The purified proximity-ligation products were randomly reduced to Illumina library insert size via the Diagenode Bioruptor NGS sonication platform. TruSeq libraries were made using the NEBNext Ultra II DNA Library Prep Kit for Illumina (NEB) and a custom Y-adaptor. Adaptor ligation products were SPRI bead purified before capturing biotinylated molecules using Dynabeads MyOne Streptavidin C1 beads (Thermo Fisher). After enrichment of biotinylated products by buffer exchange, streptavidin-bound DNA was used as input to index PCR with KAPA HiFi HotStart ReadyMix (Roche) and TruSeq indexing adaptors with dual unique index sequences. The resulting library molecules were purified and simultaneously size-selected using SPRI beads. The library was sent to Fulgent Genomics for sequencing on an Illumina HiSeq 4000 platform (2x100 bp).

10× Chromium linked reads

Genomic DNA was extracted from 5×10^6 lymphoblastoid cells using the Qiagen Blood and Cell Culture DNA Mini Kit and used as input to the 10× Genomics Chromium Genome Library Kit and Gel Bead Kit v2 (Pleasanton, CA), following the manufacturer's recommended protocols (Manual CG00043 Rev A) to generate the library. The library was sent to Fulgent Genomics for sequencing on an Illumina HiSeq X Ten platform (2x150 bp).

Oxford Nanopore long reads

Oxford Nanopore Technologies (ONT) sequencing libraries were prepared from genomic DNA using the LSK-109 sequencing kit with minor modifications. Namely, end-repair, A-tailing, and ligation incubation times were increased to 30 min each. Libraries were sequenced on the ONT MinION using a R9.4.1 flow cell. Fast5 raw data files were basecalled and converted into FASTQ files using the ONT research basecaller flappie [Version 1.0.0] (https://github.com/nanoporetech/flappie).

PacBio CCS long reads

Genomic DNA was extracted from 5×10^6 Lionel lymphoblastoid cells using the Qiagen Blood and Cell Culture DNA Mini Kit and sent to the Vincent J. Coates Genomics Sequencing Lab (Berkeley, CA) for sequencing on 4 SMRT cells.

RNA sequencing

We used public bulk RNA-seq data (Hartley et al. 2021) that was obtained from fresh frozen Lionel lymphoblastoid cell pellets using mirVana Total RNA Isolation kit (Thermo Fisher) and prepared using Illumina TruSeq stranded total RNA with Ribo-depletion.

DNA sequencing and genome assembly

Mitochondrial genome assembly

The mitochondrial genome was assembled from the shotgun Illumina data using a reference-guided iterative approach (Green et al. 2008). The *Hylobates agilis* mitochondrial genome (NC_014042) was used as the starting reference sequence.

Nuclear genome assembly

We generated the initial assembly with 10X Chromium linked reads using the Supernova assembler [version 2.0.1, --style pseudohap] (Weisenfeld et al. 2017). Then, we ran a first gap closing round with both PacBio and ONT long reads using minimap2 [Version 2.12] (Li 2018) and YAGCloser (https://www.github.com/merlyescalona/yagcloser). Next, we preprocessed both Chicago and Hi-C data by trimming to the DpnII junction sequence (GATCGATC). We then ran HiRise [Version 2.1.6] (Putnam et al. 2016) to scaffold the assembly using the Chicago data and the short reads. We closed gaps and re-scaffolded the assembly with HiRise, this time using the Hi-C data and the short reads. We polished the scaffolds in 2 rounds using short and linked reads with Pilon [Version 1.22] (Walker et al. 2014). To align the shotgun short reads to the assembly, we used BWA-MEM [version 0.7.17-r1188] (Li and Durbin 2009), and to align the linked reads, we followed the pipeline from https://github.com/ucdavis-bioinformatics/proc10xG, which is a set of scripts that extract the GEM barcodes and trim primer sequences of the linked reads.

We remapped the Hi-C data to our scaffolded assembly with BWA-MEM [with options -5SP]. Then, we identified ligation junctions and generated Hi-C pairs using the pairtools [Version 0.20] (Goloborodko et al. 2018). From the Hi-C pairs, we generated a multi-resolution Hi-C matrix in a binary form (contact map) with cooler [Version 0.8.10] (Abdennur and Mirny 2020). Finally, to identify and manually correct structural errors generated in the scaffolding process, we used HiGlass [Version 2.1.11] (Kerpedjiev et al. 2018) to visualize the Hi-C contact map, and D-GENIES [Version 1.2.0.1] (Cabanettes and Klopp 2018) to compare our genome to the closest genome reference available (Asia_NLE_v1) (Fig. 1A and B).

Next, we removed scaffolds that had at least 80% identity to the mitochondrial genome and that were equal to or shorter than the length of the mitochondrial sequence. Finally, we removed duplicated scaffolds, trimmed scaffold endings that contained N sequences and masked sequencing adaptors.

Analysis of sex chromosomes

The individual used for this assembly is a male. Thus, we can identify scaffolds from both sex chromosomes. Ideally, chromosomes X and Y should be assembled as separate scaffolds. We would expect those scaffolds to be primarily covered by reads at a depth approximately half the median coverage across the genome (CN = 1) and that X and Y homolog genes will fall into separate scaffolds, except in psuedoautosomal regions (PAR) at the ends of the sex chromosomes.

To determine which scaffolds correspond to sex chromosomes, we aligned the human PAR genes (Mangs and Morris 2007), human Y genes and their X homologs (Godfrey et al. 2020) (Supplementary Table 1) to the present assembly with BLAT (Kent 2002), keeping the alignments with the highest sequence identity per gene. We then aligned the Illumina shotgun data to the assembly using BWA-MEM and calculated depth of coverage across the entire genome using bedtools [Version 2.27.1, options -bga -split] (Quinlan and Hall 2010). We also generated copy number (CN) calls with Control-FREEC [Version 11.6] (Boeva et al. 2012) using a pileup file created with samtools mpileup (Li et al. 2009) from the aligned Illumina reads and a window size 1 Kb. Finally, we used the joint information from the gene annotation, gene alignments, depth of coverage, and the Hi-C contact map to manually correct sex-chromosome scaffolds.

Gene annotation and identification of repetitive elements

The genome was annotated by NCBI according to the NCBI Eukaryotic Genome Annotation Pipeline [Version 8.3] (Thibaud-Nissen et al. 2013) using publicly available RNA-seq data generated for the *Hylobates lar* (Xu et al. 2018) and *H. moloch* (Okhovat et al. 2020). Repetitive elements were identified and soft-masked using RepeatMasker (Version 4.1.0, RRID:SCR_012954) (Smit et al. 2013) with the Dfam database [Version 3.1, -e ncbi -species Primates -xsmall] (Hubley et al. 2016). We compared our final annotation with other primates (Supplementary Table 2) and plotted the results (Fig. 1D) in R [Version 3.6.3] (R Core Team 2020) using the ggplot2 package [Version 3.3.2] (Wickham 2016).

Quality assessment

We ran BUSCO [Version 5] (Simão et al. 2015) to evaluate genome quality and completeness by quantifying the number of universal single-copy orthologs present in the assembly. Specifically, we used the mammalia ortholog database (mammalia_odb10), which contains 9,226 genes. This assessment was done on both gibbon assemblies available (HMol_V3 and Asia_NLE_v1). We measured the base level

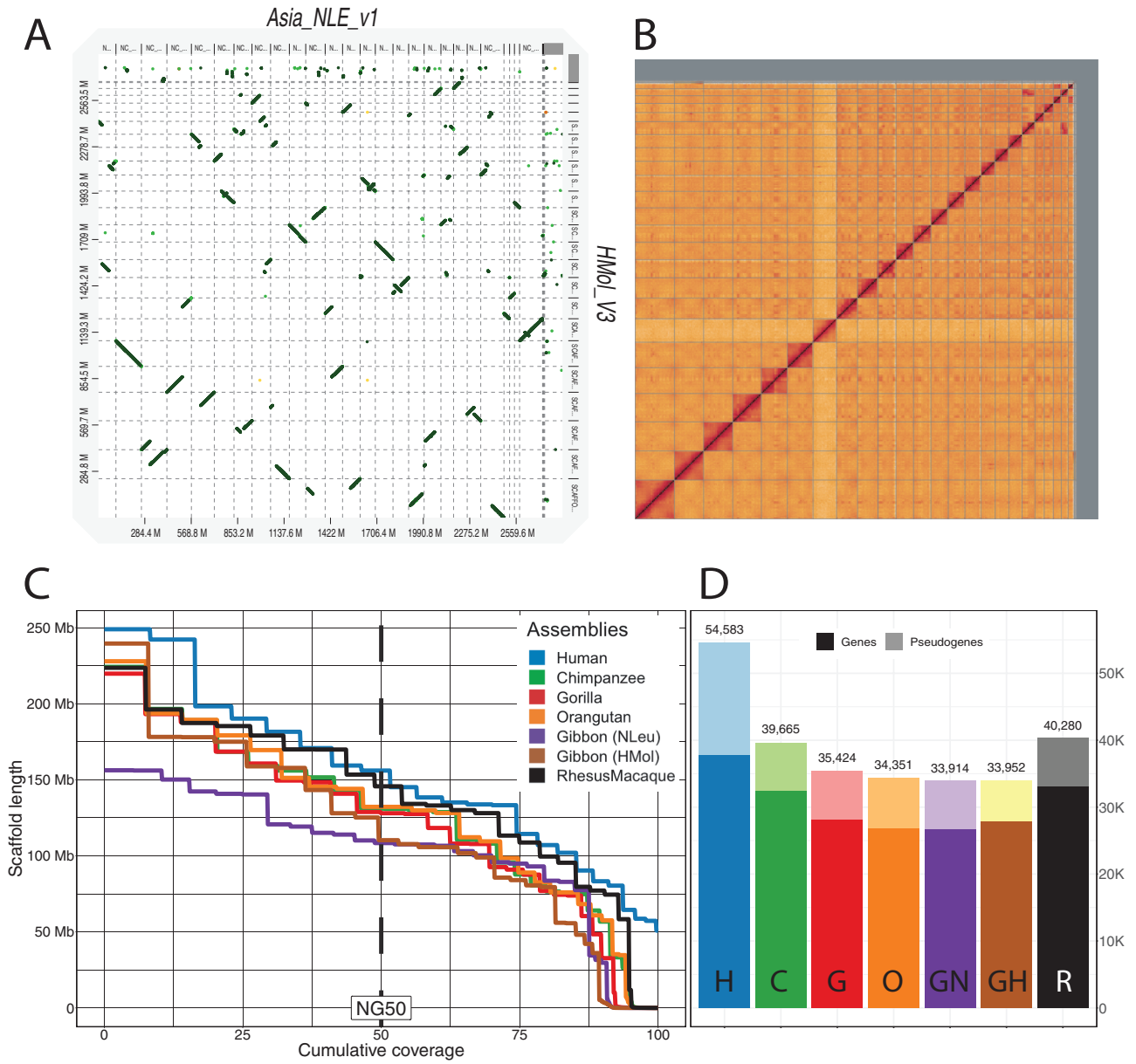

Fig. 1. (A, B) Visual support for the identification of mis-assemblies and curation of HMol_V3. (A) Dotplot showing a comparison of HMol_V3 assembly vs. Asia_NLE_v1, chromosome scale assembly of the *Nomascus leucogenys*, the closest gibbon genome available on NCBI. (B) Hi-C contact map from HiGlass showing validation of the assembly. (C, D) Comparison of HMol_V3 versus other primate genomes. (C) Comparison of NGx statistics among genomes from great apes, gibbon, and rhesus. (D) Comparison of the number of annotated genes (dark) and pseudogenes (light) for the same primate species as in (C).

accuracy (QV, a log scaled probability of error of the consensus base calls) with a k-mer based approach, using meryl [Version snapshot v1.1 + 34] and merqury [Version 2020-01-29] (Rhie et al. 2020).

For general contiguity statistics, we ran QUAST [Version 5.0.2] (Gurevich et al. 2013). To compare the contiguity of our genome to other primates, we calculated the NGx statistics (cumulative coverage of the genome versus scaffold length) among great apes, gibbon, and rhesus macaque genomes (Supplementary Table 2). We extracted scaffold sizes of the genomes with samtools faidx, calculated the cumulative coverage using 3 Gb as the estimated genome size for all species and plotted it (Fig. 1C) in R and the ggplot2 package.

SNP Calling

We called single nucleotide polymorphism (SNPs) with the genome analysis toolkit (GATK) [Version 4.1.4.1] (Mckenna et al. 2010) according to GATK Best Practices (DePristo et al. 2011; Van der Auwera et al. 2013). First, we aligned all the shotgun Illumina data to the final reference with BWA-MEM. We then used samtools to sort it by coordinates and samtools fixmates, sort and markdup to mark duplicates. Next, we called variants with GATK HaplotypeCaller [-ERC GVCF -G StandardAnnotation -G AS_StandardAnnotation -G StandardHCAnnotation] and consolidated the output GVCF file with GenotypeGVCFs. SNPs were filtered based on depth of coverage [mean ± (5* standard deviation)] using vcftools [Version 0.1.15] (Danecek et al. 2011).

Table 1. Assembly pipeline and software usage.

Sequencing	Software	Version
ONT Basecaller	flappie https://github.com/nanoporetech/flappie	0.1.0
Assembly		
10× linked-reads assembler	Supernova	2.0.1
Long-reads aligner	minimap2	Commit 64d1c7b
Dovetail Genomics scaffolder	HiRise	2.1.6
Gap closing	YAGCloser https://github.com/merlyescalona/yagcloser	1.0
Hi-C Contact map generation		
Short-read alignment	Bwa	0.7.17-r1188
SAM/BAM processing	Samtools	1.11
SAM/BAM filtering	pairtools	0.3.0
Pairs indexing	pairix	0.3.7
Matrix generation	Cooler	0.8.10
Contact map visualization	HiGlass	2.1.11
Genome assembly quality assessment		
Processing of linked reads	proc10xG https://github.com/ucdavis-bioinformatics/proc10xG	Commit 7afbfcf
Dotplot generation	D-GENIES	1.2.0.1
Assessment tool for genome assemblies	QUAST	5.0.2
K-mer counter	meryl	v1.1 + 34
K-mer based assembly evaluation	merqury	2020-01-29
Genome assembly quality assessment		
Genome Annotation Pipeline	NCBI Eukaryotic Genome Annotation Pipeline	8.3
Screening for interspersed repeats and low-complexity DNA sequences	RepeatMasker	4.1.0
Database of transposable element and repetitive DNA families	Dfam	3.1
Heterozygosity estimation	GenomeScope	2.0
Variant caller	GATK	4.1.4.1
VCF processing	vcftools	0.1.15
VCF processing	bcftools	1.7 (htslib1.7-2)
Sex-chromosome analysis		
Fast sequence search	blat	36x9
Toolset for genomic arithmetic	bedtools	2.27.1
Copy number annotation	Control-FREEC	11.6

Software citations are listed in the text.

Heterozygosity estimation

We used 2 approaches to estimate the heterozygosity levels: one based on SNP Calling and the other based on k-mer count. For the first, we filtered the SNPs to keep only those within the autosomes. We aligned the HMol_V3 genome to the GRCh38.p13 human reference genome using minimap2 [-ax asm5]. Then, we filtered the alignments with MAPQ ≥ 60 and based on the alignment score tag (AS), we kept the best alignment per scaffold and identified scaffolds that aligned to chromosomes X and Y. We then filtered out the SNPs corresponding to those scaffolds and calculated the heterozygosity for the autosomes as the number of SNPs found in the scaffolds corresponding to the autosomes divided by the total length of autosomal scaffolds. For the second approach, we used GenomeScope [Version 2.0] (Ranallo-Benavidez et al. 2020) to fit the k-mer count histograms from the Illumina shotgun reads to the models for estimating heterozygosity of Lionel and compared this result with other available primate genome assemblies. Further statistics referring to transversion/transition were calculated using vcftools and bcftools [Version 1.7 (using htslib 1.7-2)] (Li et al. 2009; Li 2011a, 2011b).

Results

Description of sequencing datasets

We used multiple technologies to sequence the genome of a male *H. moloch*. In total, we generated over 499 million read pairs of whole-genome shotgun Illumina, 751 million read pairs of 10X Chromium linked reads, 184 million read pairs for Chicago, and 194 million read pairs for Hi-C. The long reads we generated included 751 thousand PacBio CCS reads

(N50 read length 5,927 bp; minimum read length 5 bp, mean read length 4,549 bp and maximum read length 47,777 bp) and 2.8 million Oxford Nanopore long reads (N50 read length 6,523 bp, minimum read length 2 bp, mean read length 4,846 bp and maximum read size 362,572 bp) (Table 2).

Genome assembly quality assessment

The resulting genome assembly (HMol_V3) contains 18,400 scaffolds with a total span of 2.84 Gb, a contig N50 size of ~265 Kb and a scaffold N50 size of ~125 Mb. The assembly has ~8,657 gaps/per Gb of genome. It has a consensus quality value (QV score) of 46 and a k-mer completeness score of 95.1%. BUSCO analysis of the genome assembly shows 94.5% of complete ortholog genes, compared to 95.6% for the *Nomascus leucogenys* (Asia_NLE_v1) genome.

Analysis of sex chromosomes

We identified a scaffold on our genome assembly version HMol_V2 that incorrectly joined elements of chromosome X, Y and pseudoautosomal regions. To identify these segments and break the misjoins between them, we aligned a gene set of PAR genes and X–Y homolog gene pairs to the assembly. We found that 92% of genes from this sex-chromosome set aligned to a single scaffold. On this scaffold, the Y genes fall in the 0-9 Mb range and the X homolog genes fall between 9-135 Mb range. We also observed that PAR genes aligned in the 6-9 Mb range, overlapping with the alignment space of the Y genes (Fig. 2B).

In addition, from the Control-FREEC CN calls on this scaffold, we observed that the scaffold has mostly copy number of 1 (CN = 1; intervals 1 and 3) and it also has a segment (interval 2) with 2 CN = 2 intervals (intervals 2A and 2C) separated by a short CN = 1 interval (interval 2B) (Fig. 2A). We observed the alignments of PAR genes located only in intervals 2A and 2C, while we observed only Y genes in intervals 1 and 2B, and mostly X genes in interval 3. On the basis of this evidence, we introduced 4 breaks in the scaffold. The version HMol_V3 is identical to the previous version HMol_V2, except for these 4 breaks.

Gene annotation and repetitive elements

We submitted the genome assembly version HMol_V2 for gene annotation through NCBI. The evidence used to support the gene predictions came from over 1.3 billion of Illumina RNA-seq reads from 8 *H. lar* samples and 1 sample from *H. moloch* (Supplementary Table 3). The NCBI Hylobates moloch Annotation Release 101 (https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Hylobates_moloch/101/) reports 33,952 genes and pseudogenes (27,291 and 6,031, respectively) with a transcript mappability of 98.91%. We used the annotation as an extra layer of evidence that supported the breaks on version HMol_V2.

Table 2. Summary statistics of the sequencing datasets used and the assembly.

Sequencing data				
Type of reads			**Number of reads**[a]	**Estimated coverage**[b]
Shotgun	Illumina	2 × 150	499,807,768	51.70×
		2 × 100	336,992,689	23.24×
Chicago		2 × 75	184,189,646	9.52×
Hi-C		2 × 100	193,957,854	13.37×
Linked reads		2 × 150	729,161,586	75.43×
Long reads	PacBio CCS		751,217	1.17×
	Oxford Nanopore		2,804,845	4.68×
Genome assembly				
# Contigs	43,502			
Contig N50 (L50)	265,822 (2,985)			
Longest contig	2,599,352			
# Scaffolds	11,396			
Scaffold N50 (L50)	125,196,221 (8)			
Longest scaffold	239,559,583			
Gaps/Gb	8,657			
# Gaps	25,106			
Mean size	2,484			
Size distribution	(10–100)	(100–1k)	(1k–10k)	(10k–100k)
# gaps per range	6,177	11,662	5,280	1,987
BUSCO Scores[c]	C: 94.5% [S:91.8%, D:2.7%], F:2.0%, M:3.5% (n = 9,226)			
K-mer completeness	95.1		Base QV	46

[a]Number of read pairs for Illumina short-read datasets. Number of reads for long reads.
[b]Estimated coverage is calculated with genome size of 2.9 Gb.
[c]BUSCO Scores: C, complete single-copy and duplicated. S, Single-copy. D, Duplicated. F, Fragmented. M, Missing. n, database size, number of genes analyzed.

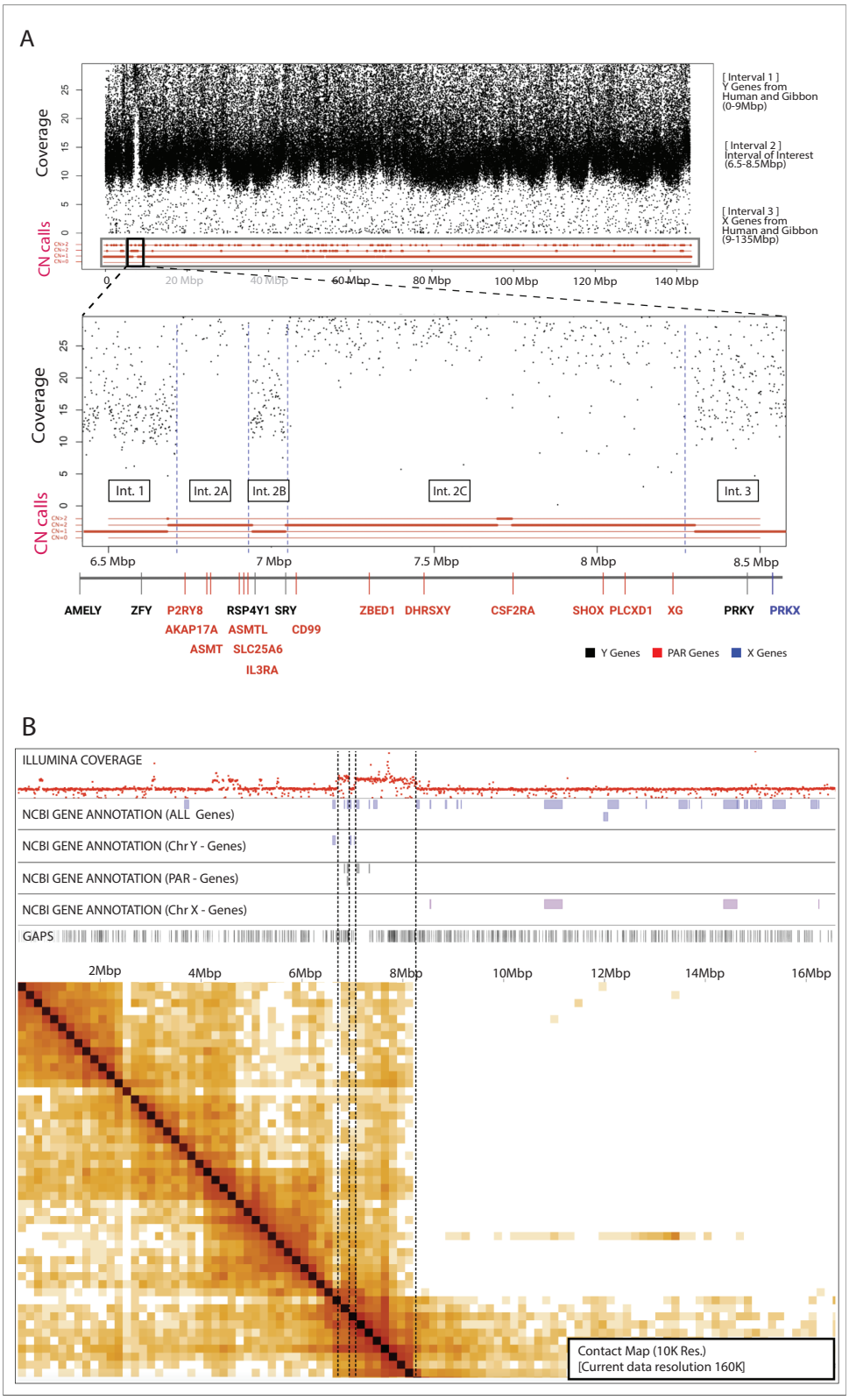

Fig. 2. Analysis of sex-chromosome scaffolds. (A) Coverage depth and copy number calls along scaffold WKKJ02000007. Inset expands the region between coordinates 6.5–8.5 Mbp. Alongside we show the coordinates of PAR genes and X–Y homolog gene pairs. (B) Hi-C contact map zoomed into first 16 Mb of scaffold WKKJ02000007, jointly with coverage and the NCBI gene annotations that show further evidence of a mis-assembly. Vertical dashed lines represent breaks in the scaffold that were applied to solve the mis-assembly.

We estimated the total repeat content of the final Javan gibbon genome (HMol_V3) to be ~49.1% (Supplementary Table 4). The majority of the identified repeats consist of retrotransposon elements (46.87%), divided into short interspersed nuclear elements (SINEs; 13.44% of repeats), long interspersed nuclear elements (LINEs; 21.11% of repeats), long terminal repeats (LTRs) retrotransposons (8.72% of repeats), and DNA elements (3.6% of repeats). Gibbon genomes are characterized by a lineage-specific composite retrotransposon, called LAVA (Carbone et al. 2012). In the Javan gibbon genome we retrieved 2,909 repeats annotated as LAVA. An additional 1,380 repeats were annotated as SVA (SVA_A) by Repeat Masker, however we know that gibbon genomes only include a handful of SVAs (Wang et al. 2005). Therefore, we are confident that the majority of repeats annotated as SVA represent mis-annotated LAVA elements, likely due the high similarity between these 2 elements. Finally, ~2.05% of all repeats were predicted to be small RNAs, satellites, or simple or low-complexity repeats (Supplementary Table 2).

SNP calling and heterozygosity

After SNP calling and filtering, we found 5,760,474 high-quality SNPs, from which 5,666,083 were mapped to autosomes and 94,391 to potential sex chromosomes. We obtained autosomal heterozygosity estimates with GenomeScope2.0 (k-mer based) within the 0.3757%–0.3917% range and 0.2103% based on SNP calling only. These levels of heterozygosity fall within the Hylobatidae family range (0.19%–0.41%) although it is lower than what was previously calculated for *H. moloch* (0.31%) (Kim et al. 2011). Transition SNPs (68.36%, 3,873,356 SNPs) were more frequent than transversions (31.57%, 1,789,342 SNPs) with a ratio of 2.16.

Discussion

The genome assembly for *H. moloch* provides a new genomic resource for the study of the gibbons, the endangered small apes. Gibbons represent an important lineage within the primates as they were the first apes to split from the Old World monkeys. Moreover, they have experienced accelerated karyotype evolution, which makes them an ideal model to study evolutionary chromosomal genetics and population genetics. Finally, given their critically endangered status, providing more genetic resources will be key to implementing more targeted conservation strategies.

This is the second reference genome generated for a gibbon species, placing us closer to the goal of sequencing genomes from all 4 extant gibbon genera. We took advantage of the strengths of multiple sequencing technologies to generate a genome assembly of quality Q46 for base accuracy and 94.5% BUSCO score for complete ortholog genes. As gibbon genomes are structurally very divergent from each other, comparisons are expected to yield important insights on how these rapid karyotype changes occur. For instance, by aligning HMol_V3 to the *Nomascus leucogenys* assembly (Fig. 1A), we can identify breaks of synteny between these 2 species' genomes, and their impact on epigenetic architecture, similar to what was done between *Nomascus* and human (Carbone et al. 2006, 2009). The completion of this genome assembly paves the way for generating other high-quality genomes from additional gibbon species, enriching the resources available for this group of endangered primates.

Supplementary material

Supplementary material is available at *Journal of Heredity* online.

Funding

This work was supported by the National Human Genome Research Institute (Grant ID 5U24HG00908103 to R.E.G and R01HG010333 to L.C); the National Science Foundation (Grant ID 1613856 to L.C and 1453527 to P.M.), and the National Institute of General Medical Sciences (Grant ID R01GM130691 to K.D.M.).

Conflict of Interest: R.E.G. is co-founder and paid consultant of Dovetail Genomics LLC.

Acknowledgments

Authors would like to express their gratitude to the staff at the Gibbon Conservation Center (Santa Clarita, CA), especially the director, Ms. Gabriella Skollar, who provided us with the opportunistic gibbon sample. This work used the Vincent J. Coates Genomics Sequencing Laboratory at UC Berkeley, supported by NIH S10 OD018174 Instrumentation Grant.

Data Availability

We have deposited the primary data underlying these analyses under NCBI BioProject PRJNA575281. Raw sequencing data for Lionel sample (NCBI BioSample SAMN12851060) are deposited in the NCBI Short-Read Archive (SRA) under: SRR13326559, SRR13326560, SRR1332661 for shotgun Illumina data; SRR13326557 for Chicago; SRR13326558 for Hi-C; SRR13326555 PacBio long reads; SRR13326556 for 10X Chromium linked reads and SRR13326554 for Oxford Nanopore long reads. RNA-seq data are deposited on GEO under the GSE161191 series, sample GSM4891325. Assembly DNA sequences are on GenBank with accession numbers WKKJ02000000 for HMol_V2 and WKKJ03000000 for HMol_V3.

References

Abdennur N, Mirny LA. 2020. Cooler: scalable storage for Hi-C data and other genomically labeled arrays. *Bioinformatics*. 36:311–316.

Andayani N, Morales JC, Forstner MRJ, Supriatna J, Melnick DJ. 2001. Genetic variability in mtDNA of the silvery gibbon: implications for the conservation of a critically endangered species. *Conserv Biol*. 15:770–775.

Audebert JB. 1798. *Histoire naturelle des singes et des makis*. Paris, France: L'an VII (1797).

Boeva V, Popova T, Bleakley K, Chiche P, Cappo J, Schleiermacher G, Janoueix-Lerosey I, Delattre O, Barillot E. 2012. Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. *Bioinformatics*. 28:423–425.

Cabanettes F, Klopp C. 2018. D-GENIES: dot plot large genomes in an interactive, efficient and simple way. *PeerJ*. 6:e4958.

Carbone L, Harris RA, Gnerre S, Veeramah KR, Lorente-Galdos B, Huddleston J, Meyer TJ, Herrero J, Roos C, Aken B, et al. 2014. Gibbon genome and the fast karyotype evolution of small apes. *Nature*. 513:195–201.

Carbone L, Harris RA, Mootnick AR, Milosavljevic A, Martin DIK, Rocchi M, Capozzi O, Archidiacono N, Konkel MK, Walker JA, et al. 2012. Centromere remodeling in Hoolock leuconedys

(Hylobatidae) by a new transposable element unique to the gibbons. *Genome Biol Evol.* 4:648–658.
Carbone L, Harris RA, Vessere GM, Mootnick AR, Humphray S, Rogers J, Kim SK, Wall JD, Martin D, Jurka J, et al. 2009. Evolutionary breakpoints in the gibbon suggest association between cytosine methylation and karyotype evolution. *PLoS Genet.* 5:e1000538.
Carbone L, Vessere GM, ten Hallers BFH, Zhu B, Osoegawa K, Mootnick A, Kofler A, Wienberg J, Rogers J, Humphray S, et al. 2006. A high-resolution map of synteny disruptions in gibbon and human genomes. *PLoS Genet.* 2:e223.
Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, et al; 1000 Genomes Project Analysis Group. 2011. The variant call format and VCFtools. *Bioinformatics.* 27:2156–2158.
DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet.* 43:491–498.
Dutrillaux B, Rethore MO, Aurias A, Goustard M. 1975. Karyotype analysis of 2 species of gibbons (*Hylobates lar* and *H. concolor*) with different banding species. *Cytogenet Cell Genet.* 15:81–91.
Godfrey AK, Naqvi S, Chmátal L, Chick JM, Mitchell RN, Gygi SP, Skaletsky H, Page DC. 2020. Quantitative analysis of Y-Chromosome gene expression across 36 human tissues. *Genome Res.* 30:860–873.
Goloborodko A, Abdennur N, Venev S; hbbrandao, gfudenberg. 2018. pairtools. Available from: https://zenodo.org/record/1490831.
Green RE, Malaspinas A-S, Krause J, Briggs AW, Johnson PLF, Uhler C, Meyer M, Good JM, Maricic T, Stenzel U, et al. 2008. A complete Neandertal mitochondrial genome sequence determined by high-throughput sequencing. *Cell.* 134:416–426.
Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. QUAST: quality assessment tool for genome assemblies. *Bioinformatics.* 29:1072–1075.
Hartley GA, Okhovat M, O'Neill RJ, Carbone L. 2021. Comparative analyses of gibbon centromeres reveal dynamic genus-specific shifts in repeat composition. *Mol Biol Evol.* 38:3972–3992.
Hubley R, Finn RD, Clements J, Eddy SR, Jones TA, Bao W, Smit AFA, Wheeler TJ. 2016. The Dfam database of repetitive DNA families. *Nucleic Acids Res.* 44:D81–D89.
Kent WJ. 2002. BLAT--the BLAST-like alignment tool. *Genome Res.* 12:656–664.
Kerpedjiev P, Abdennur N, Lekschas F, McCallum C, Dinkla K, Strobelt H, Luber JM, Ouellette SB, Azhir A, Kumar N, et al. 2018. HiGlass: web-based visual exploration and analysis of genome interaction maps. *Genome Biol.* 19:125.
Kim SK, Carbone L, Becquet C, Mootnick AR, Li DJ, de Jong PJ, Wall JD. 2011. Patterns of genetic variation within and between Gibbon species. *Mol Biol Evol.* 28:2211–2218.
Koehler U, Bigoni F, Wienberg J, Stanyon R. 1995. Genomic reorganization in the concolor gibbon (*Hylobates concolor*) revealed by chromosome painting. *Genomics.* 30:287–292.
Li H. 2011a. Improving SNP discovery by base alignment quality. *Bioinformatics.* 27:1157–1158.
Li H. 2011b. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics.* 27:2987–2993.
Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics.* 34:3094–3100.
Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics.* 25:1754–1760.
Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics.* 25:2078–2079.
Mangs AH, Morris BJ. 2007. The Human Pseudoautosomal Region (PAR): origin, function and future. *Curr Genomics.* 8:129–136.
Mckenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20:1297–1303.
Mrasek K, Heller A, Rubtsov N, Trifonov V, Starke H, Claussen U, Liehr T. 2003. Detailed *Hylobates* lar karyotype defined by 25-color FISH and multicolor banding. *Int J Mol Med.* 12:139–146.
Nijman V. 2015. IUCN Red List of Threatened Species: *Hylobates moloch*. Available from: https://www.iucnredlist.org/species/10550/17966495.
Okhovat M, Nevonen KA, Davis BA, Michener P, Ward S, Milhaven M, Harshman L, Sohota A, Fernandes JD, Salama SR, et al. 2020. Co-option of the lineage-specific LAVA retrotransposon in the gibbon genome. *Proc Natl Acad Sci USA.* 117:19328–19338. doi:10.1073/pnas.2006038117
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.* 26:342–350.
Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics.* 26:841–842.
R Core Team. 2020. *R: a language and environment for statistical computing*. Vienna (Austria): R Foundation for Statistical Computing. Available from: https://www.R-project.org/
Ranallo-Benavidez TR, Jaron KS, Schatz MC. 2020. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat Commun.* 11:1432.
Reichard UH, Barelli C, Hirai H, Nowak MG. 2016. The evolution of gibbons and Siamang. In Louise Barrett (ed.). *Evolution of gibbons and Siamang: phylogeny, morphology, and cognition*. New York (NY): Springer New York. p. 3–41.
Rhie A, Walenz BP, Koren S, Phillippy AM. 2020. Merqury: reference-free quality and phasing assessment for genome assemblies. *Genome Biol.* 21:245.
Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics.* 31:3210–3212.
Smit AFA, Hubley RR, Green PR. 2013. RepeatMasker Open-4.0. Available from: http://repeatmasker.org/.
Thibaud-Nissen F, Souvorov A, Murphy T, DiCuccio M, Kitts P. 2013. *Eukaryotic Genome Annotation Pipeline. The NCBI Handbook [Internet]*. 2nd ed. National Center for Biotechnology Information (US). Available from: https://www.ncbi.nlm.nih.gov/books/NBK169439/.
Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, et al. 2013. From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. *Curr Protoc Bioinformatics.* 43:11.10.1–11.10.33.
Veeramah KR, Woerner AE, Johnstone L, Gut I, Gut M, Marques-Bonet T, Carbone L, Wall JD, Hammer MF. 2015. Examining phylogenetic relationships among gibbon genera using whole genome sequence data using an approximate Bayesian computation approach. *Genetics.* 200:295–308.
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.
Wang H, Xing J, Grover D, Hedges DJ, Han K, Walker JA, Batzer MA. 2005. SVA elements: a hominid-specific retroposon family. *J Mol Biol.* 354:994–1007.
Weisenfeld NI, Kumar V, Shah P, Church DM, Jaffe DB. 2017. Direct determination of diploid genome sequences. *Genome Res.* 27:757–767.
Wickham H. 2016. ggplot2: Elegant Graphics for Data Analysis. Available from: https://ggplot2.tidyverse.org.
Xu C, Li Q, Efimova O, He L, Tatsumoto S, Stepanova V, Oishi T, Udono T, Yamaguchi K, Shigenobu S, et al. 2018. Human-specific features of spatial gene expression and regulation in eight brain regions. *Genome Res.* 28:1097–1110.