www.sciencemag.org/content/360/6393/eaar6343/suppl/DC1

Supplementary Materials for

High-resolution comparative analysis of great ape genomes

Zev N. Kronenberg, Ian T. Fiddes*, David Gordon*, Shwetha Murali*, Stuart Cantsilieris*, Olivia S. Meyerson*, Jason G. Underwood*, Bradley J. Nelson*, Mark J. P. Chaisson, Max L. Dougherty, Katherine M. Munson, Alex R. Hastie, Mark Diekhans, Fereydoun Hormozdiari, Nicola Lorusso, Kendra Hoekzema, Ruolan Qiu, Karen Clark, Archana Raja, AnneMarie E. Welch, Melanie Sorensen, Carl Baker, Robert S. Fulton, Joel Armstrong, Tina A. Graves-Lindsay, Ahmet M. Denli, Emma R. Hoppe, PingHsun Hsieh, Christopher M. Hill, Andy Wing Chun Pang, Joyce Lee, Ernest T. Lam, Susan K. Dutcher, Fred H. Gage, Wesley C. Warren, Jay Shendure, David Haussler, Valerie A. Schneider, Han Cao, Mario Ventura, Richard K. Wilson, Benedict Paten, Alex Pollen, Evan E. Eichler†

*These authors contributed equally to this work.
†Corresponding author. Email: eee@gs.washington.edu

Published 8 June 2018, *Science* **360**, eaar6343 (2018)
DOI: 10.1126/science.aar6343

This PDF file includes:

Materials and Methods
Figs. S1 to S53
Captions for Tables S1 to S15
Tables S16 to S58
References

Other Supporting Online Material for this manuscript includes the following:
(available at www.sciencemag.org/content/360/6393/eaar6343/suppl/DC1)

Tables S1 to S15 (Excel)

CONTENTS

Materials and Methods

Sequencing: For single-molecule, real-time (SMRT) sequencing, we generated long-insert genomic libraries using standard protocols (http://www.pacb.com/wp-content/uploads/2015/09/Procedure-Checklist-20-kb-Template-Preparation-Using-BluePippin-Size-Selection.pdf) except we sheared with Megaruptor (Diagenode) and size-selected with BluePippin (Sage Science) at high-pass settings of 15-30 kbp. Orangutan and chimpanzee SMRT genome sequencing were performed using the PacBio RS II platform (University of Washington) and P6v2-C4v2 chemistry (6-hour movies) to an approximate coverage of 107-fold (Clint) and 81-fold (Susie) (reads of insert coverage or ROI). SMRT sequence data for CHM13 and Yoruban (YRI19240) samples were generated similarly (Washington University, St. Louis) as part of the Human Reference Genome Sequencing Consortium. Illumina WGS data were publicly available for Clint, CHM13, and Yoruban samples. For orangutan, we generated a 42-fold sequence coverage dataset from genomic libraries (average insert size 550 bp; Illumina TruSeq DNA PCR-Free library kit) and paired-end 150 bp sequence reads (NextSeq 500). For transcriptome sequencing, both short- and long-read RNA-seq libraries were prepared from polyA+ RNA purified from iPSCs from each of the organisms. For long-read RNA-seq, a modified Iso-Seq library preparation (PacBio) was used and size fractions were isolated with SageELF (Sage Science). For sequencing, 6-hour movies were collected on the PacBio RS II with P6v2-C4v2 chemistry to generate a minimum of 0.5M FLNC reads per organism. For short-read RNA-seq, the RNA was fragmented to approximately 150 nt for stranded RNA-seq libraries (Illumina TruSeq) and sequenced on the HiSeq 2500 (human, gorilla, chimpanzee; paired-end 125 bp) or the NextSeq (orangutan; paired-end 150 bp).

Sequence assembly: We assembled all four genomes (chimpanzee [Clint], Sumatran orangutan [Susie], CHM13 [human], and YRI19240 [human]) using the Falcon genome assembler (git hash 5942bc00; FALCON-integrate git hash cd9e9373 as of 2/1/2016) with read-length cutoff thresholds ranging between 10-15 kbp. Sequence contigs were error-corrected using Quiver (*17*) followed by Pilon (*18*), which applies paired-end Illumina reads to further improve accuracy. We also developed and applied a FreeBayes-based (*65*) indel correction pipeline to fix remaining indel errors. To test how assembly errors might inflate the contig N50, we broke 6,751 Clint_PTRv1 regions of high or low depth and regions with BAC-end discordance. This resulted in a small contig N50 decrease of 0.26 Mbp, since 93.6% (6,322) of the low-confidence regions are contained on contigs smaller than 1 Mbp and 5.9% (399) are less than 100 kbp from contigs ends (contigs greater than 1 Mbp).

AGP construction: We constructed a chromosomal-level AGP (a golden path) of sequence contigs for chimpanzee (Clint_PTRv1) and orangutan (Susie_PABv1) using two primary platforms: 1) optical maps generated using the Bionano Genomics (Bionano) Saphyr platform for scaffold building and 2) bicolor FISH of ~700 large-insert clones used to assign scaffolds to NHP chromosomes. We constructed Bionano optical maps using two different nicking restriction enzymes, Nt.BspQI and NbBssSI, and ran the hybrid scaffold pipeline (Bionano Solve 3.1) to flag and correct chimeric sequence contigs before scaffolding them into 121 and 73 scaffolds with N50 values of 60 Mbp and 102 Mbp for chimpanzee (Clint) and orangutan (Susie), respectively. Scaffold order and orientation were refined using chromosome-specific FISH probes (~1 probe per 5 Mbp) with a higher density of BAC probes mapping to breakpoint regions of known evolutionary rearrangements among the apes (*19*). We further tested and corrected misjoins within and between sequence contigs based on BAC end-sequence alignment to chimpanzee or orangutan, and by measuring SMRT sequence read depth to identify potential collapses in sequence assembly as previously described (*20*). As an another validation approach, we also applied Hi-C chromosomal capture technique and visualized the interaction map using Juicebox (*67*). Juicebox helped identify and resolve four large (>1 Mbp) translocation scaffolding errors in Clint_PTRv1. Comparison of the Clint_PTRv1 AGP to human revealed two of the previously karyotype characterized human–chimpanzee differences on chromosomes 9 and 15 (*19*). Several of these events were not represented in panTro3 (GCA_000001515.3), panTro4 (GCA_000001515.4), or panTro5 (GCF_000001515.7). The finalized assemblies, including indel error-corrected contigs, are released as Clint_PTRv2 and Susie_PABv2.

BAC clone sequencing: To estimate sequence accuracy and establish SV validation rates, we sequenced and assembled the BAC inserts from genomics libraries generated from the same individuals sequenced in this study, with the exception of the gorilla Kamilah. Pooled BAC inserts (5-6 BACs per pool) were prepped as 20 kbp SMRTbell libraries, size-selected, sequenced on the PacBio RS II, and assembled *de novo* using the Canu assembler (*68*) followed by Quiver consensus sequence calling (*17*). Potential sequence misassemblies were identified by BAC end-sequence mapping and read-depth sequence analysis. We reassessed such misassemblies using an alternate assembly approach with Falcon (https://github.com/PacificBiosciences/FALCON) for read overlap and string graph layout, followed by Quiver to generate the consensus sequence. A complete list of all 216 BAC clones and accessions is provided (**Table S15**).

FISH analysis: A subset of larger, subkaryotypic rearrangements were validated by bicolor FISH. Metaphase spreads were obtained from chimpanzee (Clint) and orangutan (Susie) fibroblast cell lines. FISH experiments were performed according to Lichter *et al.* method (*69*) with minor modifications, using BAC clones directly labeled by nick-translation with Cy3-dUTP, Cy5-dUTP and fluorescein-dUTP. Briefly: 1,000 ng of labeled probe were used for FISH experiments; hybridization was performed at 37°C after two minutes of denaturation at 70°C in a hybridization-solution volume of 15 µL consisting of 2xSSC, 50% (v/v) formamide, 10% (w/v) dextran sulphate, 5 µg C0t1 DNA (Roche, Monza, Italy) and 3 µg sonicated salmon sperm DNA. High-stringency post-hybridization washing was performed three times at 60°C in 0.1xSSC solution. Nuclei and metaphases were simultaneously DAPI stained (200 ng/mL in 2xSSC, 5 minutes incubation). DAPI, Cy3, Cy5 and fluorescein fluorescence signals were detected with specific filters using a Leica DMRXA2 epifluorescence microscope equipped with a cooled CCD camera (Princeton Instruments). Digital images were separately recorded as gray-scale pictures, pseudocolored, and merged using Adobe Photoshop software. For interphase FISH experiments, roughly 70-80 nuclei were observed.

SNV and phylogenetic analyses: We examined rates of evolution within the great ape phylogeny, including coding and noncoding regions. We randomly selected 10,000 autosomal CCDS exons adding 500 bp to each segment to increase phylogenetic signal. Overlapping CCDS regions were collapsed. We generated a noncoding set by randomly shuffling the same CCDS exons excluding regions that contain tandem repeats or SDs. Not all of the sampled regions resulted in five-way syntenic alignments; 96.5% (9,674) and 86.0% (8,587) of regions were recovered for the CCDS and random set, respectively. The syntenic regions were aligned with MUSCLE (v3.8.31) and a general time reversible model ("GTR+GAMMA") was fit in the maximum likelihood RAxML (8.2.3) framework. We summarized the 18,233 trees into two maximum clade credibility trees using DendroPy ("mcct").

We initially analyzed tree topologies from 8,586 random MSAs encompassing >11 Mbp of autosomal sequence with an average region size of 1.36 kbp. These regions excluded short tandem repeats (STRs) and segmental duplications (SDs). We found 5,494 trees consistent with the species tree, 17.69% consistent with incomplete lineage sorting (ILS) tree A, and 1,519 with ILS tree B 18.05% (**Fig. 2b**).

To quantify the errors on the fractions of genealogies that support ILS, we performed a permutation test using 100 random replicates, shuffling the 8,586 regions previously mentioned. We estimated the 95% confidence intervals for the fractions of these trees by calculating the 2.5 and 97.5 percentiles of their distributions based on the results of these replicates. Under the simplest model of three species with random mating in the common ancestral species, the two ILS trees are equally likely because the probabilities of the two possible coalescent events leading to ILS in the common ancestor of all three species are the same. Thus, under the null hypothesis of equally likely ILS trees, it is a binomial trial with $p = 0.5$ and $n = 1,519 + 1,550 = 3,069$. The probability of observing 1,519 instances for Tree A and 1,550 instances for Tree B under the null is 0.5881. Therefore, our inference on ILS does not favor either ILS tree.

Gene annotation: We generated two distinct multiple sequence alignment datasets using Progressive Cactus (*70*). The first dataset is a comparison of the human reference genome, GRCh38, and the previous generation of NHP assemblies (UCSC panTro4, gorGor4 and ponAbe2). The second dataset is a comparison between the human reference genome and the SMRT NHP assemblies. Comparative Annotation Toolkit (CAT; https://github.com/ComparativeGenomicsToolkit/Comparative-Annotation-Toolkit) was used to annotate all of the great ape genomes using the human GENCODE V27 as reference with a combination of RNA-seq obtained from SRA as well as iPSC RNA-seq specifically from NHPs. For the new primate assemblies, Iso-Seq reads from iPSCs were also independently used to guide AugustusPB in detecting novel transcripts and isoforms. Two approaches were used to identify novel exons. First, the raw Iso-Seq reads were mapped to GRCh38 and Clint_PTRv1. GRCh38 regions with more than 5× human Iso-Seq depth-of-coverage that did not directly overlap a GENCODE V25 exon or NHP Iso-Seq data were considered novel human exons. Conversely, regions in Clint_PTRv1 that had chimpanzee and/or gorilla coverage, but not human coverage, were considered potential novel NHP exons. Second, novel exons and splice junctions were identified in the AugustusPB annotation set by using homGeneMapping (*71*) to project the coordinates of both comparative annotation sets and Iso-Seq alignments through a Progressive Cactus alignment to establish cross-species support (**Table S1**). After discarding all exons with matches to a comparatively annotated exon and filtering for at least two Iso-Seq supporting reads, we identified 29 novel exons in Clint_PTRv1, 41 novel exons in GSMRT5, and 41 novel exons in Susie_PABv1. Of these, 18, 23 and 7 open-reading frames in Clint_PTRv1 were predicted to be longer in Clint_PTRv1, GSMRT3.2, and Susie_PABv1 increasing protein length an average of 57, 47 and 20 amino acids, respectively. Additionally, we identified 100 novel splice junctions (17 NAGNAG) in Clint_PTRv1, 136 (23 NAGNAG) in GSMRT5, and 110 (19 NAGNAG) in Susie_PABv1.

STR analysis: We tested the hypothesis of differential STR expansion by identifying reciprocal orthologous STR loci. To do so, we first identified STR sequences using RepeatMasker v4.0.1 with primate repeat libraries and Tandem Repeat Finder v4.07b with options "5 5 5 80 40 20 10 -m -ngs -h". Between 370,781 and 385,612 STR sites were identified in each assembly (**Table S2**). Due to occasional dropout of repeat annotation, repeat regions were merged if they were within 25 bp of one another, resulting in 344,354-358,622 STR regions (**Table S2**). To define orthologous STRs with confidently assigned boundaries, we first identified STR sequences that had 250 bp flanking sequences that were no more than 5% repetitive and we mapped the flanking sequences to the other assemblies. Orthologous sequences were defined as the region between the alignments of the two flanking sequences. To avoid artifacts caused from RepeatMasker sensitivity and non-STR SVs at orthologous loci, only orthologous loci that were at least 40 bp and 80% tandem repeat were retained (n = 12,694-16,138) (**Table S3**). We compared the distribution of different lengths of sequences for each pair of genomes and found the average differential expansions were small (1.2-0.02 bp) (**Table S4**). Furthermore, the comparison of distributions of expansion lengths using the Kolmogorov-Smirnov (KS) test showed no consistent pattern of significance (**Table S4**).

PtERV1 analyses: We investigated PtERV1 insertions in the assemblies of chimpanzee (Clint_PTRv1) and gorilla (GSMRT3.2) as well as Illumina WGS data from 72 great apes from five different species of chimpanzee, gorilla, bonobo, human and orangutan (*33*). We used the set of all PtERV1 LTR elements in Clint_PTRv1 (101 full length [>7 kbp in length] and 150 solo-LTR) and GSMRT3.2 (71 full length and 202 LTR) using RepeatMasker and mapped 5 kbp of flanking sequence to identify integration sites with respect to the human reference genome (GRCh38) (**Fig. S31**). We only counted loci as distinct if the flanking primary alignments mapped within 20 kbp of one another (219 loci for Clint_PTRv1 and 175 for GSMRT3.2). As a second approach, we used paired-end sequence data from Illumina WGS where one end anchored within a PtERV1 LTR element and another mapped to a unique location in human GRCh38. We did not observe a single insertion of PtERV1 in human and orangutan genomes but found 508 loci of PtERV1 integration in chimpanzee (18 samples), bonobo (10 samples) and gorilla (21 samples) genomes. PtERV1 integrations were significantly depleted in genic regions (20% of PtERV1 vs. >40% expectation) and biased in the antisense orientation within introns (**Fig. S30**). Overall, PtERV1 insertions are fourfold more likely to map within preexisting ERV elements compared to a random null model.

The intersection of assembly and one-end anchored PtERV1 datasets resulted in a combined set of 540 PtERV1 elements in GRCh38 space (**Fig. S31**). While >98% of gorilla and chimpanzee PtERV1 sites of integration were non-orthologous, we identified one locus mapping to chromosome 19 whose integration breakpoints were identical at the base-pair level between chimpanzee and gorilla assemblies that was also present in all chimpanzees and gorillas (**Fig. 2e**). We aligned the orthologous sequence with MUSCLE 3.8.31 and generated a maximum likelihood tree using RAxML 8.2.9 with a GTR+Gamma substitution model (**Fig. S33**) and created a dated phylogeny using BEAST 2.4.7 with a relaxed clock and an offset exponential prior on human–orangutan divergence with a mean of 1 million years and an offset of 12 million years. This resulted in 100% posterior support for chimpanzee–gorilla monophyly and an estimate of 4.7 mya (95% HPD: [1.9, 7.2]) for human–gorilla divergence (**Fig. S34**), supporting a model of ILS, with a single PtERV1 insertion event prior to human–gorilla divergence. We aligned the full-length PtERV1 sequences in Clint_PTRv1 and GSMRT3.2 (MUSCLE 3.8.31) to generate a maximum likelihood tree (RAxML 8.2.9 with a GTR+Gamma substitution model) (**Fig. S35**).

SV analysis: To call SVs we used smartie-sv pipeline, which aligns, compares, and calls insertions, deletions, and inversions (https://github.com/zeeev/smartie-sv). At the core of the code is a modified version of BLASR, which was designed to align large divergent contigs against a reference genome. We called SNVs, indels, and SVs (50 bp and up) using smartie-sv. We applied two filters to the raw SV calls. First, we required that query contigs cover >50% of the region (1 Mbp reference genome window). Second, we realigned SV flanks using the standard BLASR alignment algorithm to filter our marginal calls. The five call sets were manually merged and then genotyped.

hCONDEL analyses: To enrich for functionally relevant human deletions, we performed a new hCONDEL analysis involving a three-way Progressive Cactus alignment between Clint_PTRv1, rhesus macaque (rheMac8) and mouse (mm10). Because the original hCONDEL analysis did not include chimpanzee or orangutan, we removed this constraint when considering fhDELs, expanding the set from 5,892 to 7,400 candidate regions (*5*). Of these 7,400 regions, only 4,125 chimpanzee regions aligned to mouse. For each of these 4,125 fhDELs, conservation was established by examining 25 bp, 50 bp and 100 bp sliding windows at 90%, 90% and 92% identity, respectively (*5*). We excluded duplicated sequence as part of this analysis. At 25 bp and 92% identity, we identified 930 hCONDEL regions. This set can be further constrained by considering the other great ape assemblies to 795, of which 226 are novel compared to the previous hCONDEL dataset. This is in contrast to the 694 novel hCONDELs we observe when not constraining by other great ape assemblies (**Fig. 3f**). In the comparison of the old and new hCONDELs, we noted that a high number of unique old hCONDELs were not false negatives in the new set, but false positives or previously described as polymorphic in the old set. Indeed, a high proportion of these were polymorphic in CHM13_HSAv1 and YRI_HSAv1. We also detected a number of inversions that were classified as deletions in the old set, a result of the netting and chaining procedure.

Single-cell cortical expression analysis: We investigated gene expression during human and chimpanzee cortical development using single-cell gene expression data from cerebral organoid models (*55*, *56*) and from primary cortex (*57*). We used the alignment program HISAT2 to align human primary cell and organoid reads to GRCh38 and chimpanzee reads to Clint_PTRv1. Counts for each cell were performed using subread-1.5.0 function in featureCounts. After counts were obtained, normalization to counts per million was performed. We removed any cells with <1,000 genes detected or >20% of reads mapping to mitochondrial or ribosomal genes. For each dataset, we independently performed Louvain clustering on single cells based on Jaccard distance in the space of the first 30 principal components of variation (*72*). We then identified cortical radial glia as clusters distinguished by canonical markers, including *GLI3* and *FOXG1*, and excitatory neurons as clusters distinguished by canonical markers, including *NEUROD6* and *FOXG1*. By these metrics, the primary cortical analysis contained 137 radial glia and 183 excitatory neurons, the human organoid dataset contained 123 radial glia and 53 excitatory neurons, and the chimpanzee organoid dataset contained 113 radial glia and 97 excitatory neurons. For the cerebral organoid datasets, we performed differential expression between homologous human and chimpanzee cell types using a likelihood ratio test with a bimodal zero inflated distribution from the Seurat R package, and we applied Bonferroni correction.

Statistical analyses: The significance of SVs overlapping differentially expressed genes was first tested by χ-sq. SVs within 50 kbp upstream or downstream of a gene were counted as a positive association. Not all gene symbols mapped from the expression data to positional data (different annotation sets). In total, we assayed 154 downregulated Excitatory Neuron (EN) genes, 199 upregulated EN, 278 downregulated radial glia (RG) genes, and 355 upregulated RG genes (all significantly differentially expressed). The assayed genes and overlap counts used in the χ-sq test can be found in **Table S13.** To account for gene size and spatial biases, we permuted (1e4) fhSVs, both insertions and deletions, counting the number of fhSV overlap (within 50 kbp) of a differentially expressed gene. The genes and regions used in the permutation tests are listed in **Table S13**. The code used for the permutation test is available on GitHub.

Sequence accuracy and quality assessment: We assessed the sequence accuracy of the four great ape assemblies using three independent metrics. First, alignment of mapped Sanger end-sequence data to estimate a base-pair sequence accuracy of 99.85% (QV 28) for Susie_PABv1 and 99.97% (QV 35) for Clint_PTRv1, within the range of allelic diversity expected for these species.

Second, we generated >7.9 Mbp of high-quality sequence from randomly selected large-insert BAC clones derived from Clint, the chimpanzee [CH251], and Susie, the orangutan [CH276] and human (NA19240 [VMRC64] and CHM13 [VMRC59]) genomes. These raw sequence alignments have an overall sequence identity of >99.7% for diploid and >99.97% for haploid genomes. In order to account for haplotype differences, we identified all heterozygous sites from Illumina whole-genome sequence (WGS) data and assumed that remaining discrepancies are due to errors in our assemblies. This results in genome-wide estimates of one error every 1,250- 2,000 bp for diploid genomes (Clint_PTRv1 QV 33, Susie_PABv1 QV 32, YRI_HSAv1 QV 31) and QV = 36 for haploid CHM13, including indels up to 20 bp in size (**Table 1**).

Third we aligned uniquely mapping protein-coding regions >50 bp from previous primate genome assemblies (2.26/2.40 Mbp of panTro3 and 8.49/8.88 Mbp of ponAbe2). These results indicate an accuracy of 99.95% (QV 33) for Susie_PABv1 and 99.98% (QV 38) for Clint_PTRv1 within genic regions.

I. Genome sequencing and assembly

A. Chimpanzee genome sequence and assembly

A.1 Library preparation and sequencing

DNA source: We isolated and sequenced the genome of a single male chimpanzee (*Pan troglodytes verus*), Clint (Yerkes pedigree number C0471). DNA came from two sources: a commercially available DNA stock (Coriell: NS06006; 90%) and a cultured fibroblast cell line (Coriell: S006007; 10%). Cultured cell pellets were prepared using the Gentra Puregene Cell Kit (P/N: 158767). Cells were lysed, protein precipitated out, and DNA prepared. Eluted DNA was stored at 4°C overnight for two days to resuspend the DNA pellet, with quality control (QC) performed by fluorimetry (Qubit, Life Technologies) and run on a gel to visualize genomic DNA fragmentation.

Genome library preparation and sequencing: DNA fragment libraries (20-40 kbp inserts) were prepared using Megaruptor (Diagenode) shearing. Libraries were sheared at either the 40 kbp setting for size selection at 15 kbp, or the 75 kbp setting for size selection at 30 kbp. Post SMRTbell preparation per the document "Procedure and Checklist - 20 kb Template Preparation Using BluePippin™ Size-Selection System" (Pacific Biosciences [PacBio], http://www.pacb.com/wp-content/uploads/2015/09/Procedure-Checklist-20-kb-Template-Preparation-Using-BluePippin-Size-Selection.pdf). Libraries were size-selected with the BluePippin™ system (Sage Science) at their respective fragment length cutoffs (15 or 30 kbp). Single-molecule, real-time (SMRT) sequence data were generated using the PacBio RS II instrument with P6v2 polymerase binding and C4v2 chemistry kits (P6-C4) and run times of 6-hour movies. Loading concentrations were titrated empirically for each library, averaging 170 picomolar (pM) for >15 kbp size-selected libraries and 280 pM for >30 kbp size-selected libraries. Clint was sequenced to a coverage of 99X (reads of insert (ROI), 3.2 Gbp estimated genome size) or 117X (subread, 3.2 Gbp estimated genome size), on 283 SMRT cells, producing 24 million ROI reads with 33 million subreads (**Fig. S1** and **Table S16**). (Additional sequencing was done bringing totals to a subread depth of coverage of 124X on 306 SMRT cells but these additional reads were too late for the assembly.)

A.2 Genome assembly

We applied Falcon (git hash 5942bc00 and FALCON-integrate git hash cd9e9373 on 2/1/2016) to assemble the chimpanzee genome Clint_PTRv1 from SMRT sequence reads with length cutoff of 15 kbp. The coverage of reads ≥15 kbp is 69X (3.2 Gbp estimated genome size). The assembly contains 2.99 Gbp distributed amongst 4,912 contigs with an N50 of 12.7 Mbp (**Table S17**). There were 957 contigs greater than 100 kbp. The assembly was error-corrected using Quiver (*17*) and then further error-corrected using Pilon (*18*) with 37-fold Illumina paired-end reads (experiment ID: SRX243527). We also applied our own FreeBayes-based (*65*) indel correction pipeline (**Section H**).

A.3 Assembly quality: Concordance and base-pair accuracy

We assessed the quality of Clint_PTRv1 contigs using a variety of methods. In order to identify potential misassemblies and to estimate the sequencing accuracy, we mapped Sanger paired-end sequence data from the Clint large-insert bacterial artificial chromosome (BAC) end sequence (BES) (CHORI-251), and fosmid end sequence (FES) (CHORI-1251) genomic libraries against the Quiver-corrected and Pilon-polished chimpanzee assembly (**Table S18**). For the assayable regions of the assembly (2.79 Gbp with BES/FES mappings), the aligned high-quality BES data (47.3 Mbp of Sanger PHRED>40) from CHORI-271 showed 99.962% sequence identity with the contigs (**Table S19**).

We used BES and FES mapping data to estimate the proportion of the genome that was misassembled; 99.13% of the assembled genome was supported by concordant BES/FES mappings (**Table S19**) suggesting relatively few contig misassemblies. We excluded contigs shorter than 300 kbp in this analysis because of the insert size distribution of the BAC library and the enrichment of these contigs for repetitive DNA. We identified 24.2 Mbp (0.8%) of the genome, corresponding to 1,625 regions, as potentially discordant or lacking a best BES/FES support (i.e., discordant by length/ orientation, or when end sequences mapped to multiple locations or lacked support)—4.9 Mbp (0.1% of the assembled genome length) showed discordant BES/FES mappings and 19.3 Mbp (0.7%) lacked clear BES or FES support. The size distribution of these flagged regions is shown in **Fig. S2**.

We also considered regions with excess sequence read depth as an indication of collapsed repeats or duplications during the assembly process. We mapped all PacBio reads (used for the assembly) to Clint contigs and assessed sequence depth in 1 kbp non-overlapping windows and identified regions with excess or insufficient sequence coverage. Excess read depth was defined as sequence coverage of two standard deviations beyond the mean (84.9 +/- 83.8) or read depth in excess of 253 reads (**Fig. S3**). Regions with insufficient sequence coverage were defined as fewer than five sequence reads. We identified 6,923 high-depth and 22,787 low-depth 1 kbp windows out of ~3 million 1 kbp windows tested. This clustered into 600 and 5,588 high- and low-depth coverage regions respectively; the average cluster size was 4.5 kbp. Of these regions, 76% (4,715/6,186) corresponded to short sequence contigs (<100 kbp in size) while the remainder of these read-depth outliers are positionally enriched at the beginning and end of contigs (**Fig. S4**). As a sanity check, we looked at the distribution of PacBio read depth across GRCh38 autosomes and X chromosomes (**Fig. S5**). The coverage over the autosomes is almost double that of chrX, which is expected because Clint is a male chimpanzee.

We integrated the previously mentioned read-depth and BES/FES mapping outlier regions to define a set of potentially problematic regions. For the BES/FES mappings, we additionally measured the number of concordant and discordant alignments in a 10 kbp sliding window (sliding 2 kbp). Windows with fewer than two concordant BES/FES mappings and greater than five discordant mappings were marked as problematic regions. Similarly, read-depth outlier windows with fewer than five FES/BES concordant mappings were also considered problematic regions. The fraction of bases covered by problematic regions can be seen in **Fig. S6.** Smaller contigs (<100 kbp) tend to be enriched for problematic regions; 6,751 regions were flagged as putative assembly errors. These merged windows consisted of 4,519 aberrant depth, 1,238 poor BES/FES support, and 994 windows with both annotations.

We also analyzed all chimpanzee sequence contigs that did not map to GRCh38 (using Minimap: https://github.com/lh3/minimap and BLASR: https://github.com/PacificBiosciences/blasr). There were 2,228 contigs 100 bp or greater with an average length 39.6 kbp and an average GC content of ~35%. The vast majority (99.3%) consisted of >50% satellite repeat content. None of the Iso-Seq transcript annotations mapped to these contigs suggesting that they were all gene-poor. There were 14 contigs ranging from 179 bp to 23.7 kbp with a satellite content of <50%, and only one contig had more than 10 kbp of unique bases (003092F_1_23876_quiver_pilon) (**Fig. S7**). The average GC content was 43%.

A.4 Comparison with previous chimpanzee reference genomes

We compared our assembly to the chimpanzee genome assembly panTro5 (GenBank: GCA_000001515.7). Normally, a more contiguous assembly adds sequence as it replaces N's with bases and extends and joins contigs. This was the case with our gorilla and orangutan assemblies and with our chimpanzee assembly compared with the 2010 panTro3 reference (UCSC nomenclature). However, the two newer chimpanzee panTro5 assemblies (which are nearly identical to each other) are an exception, having 30 to 60 Mbp of additional sequence relative to our mainly long-read assemblies (**Table S20**).

To determine the amount of sequence added (or removed) by Clint_PTRv1 relative to panTro5, we divided all panTro5 chromosomes into 10 Mbp sequences and aligned using BLASR each 10 Mbp sequence to the corresponding Clint_PTRv1 chromosome and unplaced contigs, yielding a set of alignment blocks for that 10 Mbp sequence. BLASR (*73*) git hash 7cc3379a Nov 2016 version was run with parameters -bestn 1000 -clipping soft -alignContigs -sam -minMapQV 30 -minPctIdentity 50. If each alignment block is C Clint_PTRv1 bases long and P panTro5 bases long and contains N N's within the aligned panTro5 bases, then the number of B bases added is B = C - (P-N), the increase in number of non-N's. This simple calculation is insensitive to panTro5 over or underestimating the number of N's in a gap. To avoid counting the same bases twice, the alignment blocks were sorted by size, the largest analyzed first, and only analyzed if they did not overlap panTro5 bases of any of the previously considered alignment blocks. This resulted in ignoring less than 4% of the 2.6 Gbp of aligned panTro5 bases. Adding all negative B's gives the amount of sequence removed by Clint_PTRv1 with respect to panTro5.0, 27.8 Mbp (not including N's), and adding all the positive B's gives the amount of sequence added, 6.9 Mbp, implying that panTro5.0 has more than three times as much extra sequence than it has missing sequence. This differs from orangutan (see Section B.4) as well as the comparison between Clint_PTRv1 and the 2010 panTro3, which both fit the expected pattern. Furthermore, these numbers should be considered lower limits: if, for example, a single alignment block added 10 kbp of sequence to panTro5 in one place and removed 15 kbp in another place, the net removal of 5 kbp would be used in the sum.

A separate, cruder analysis of insertions ≥1 kbp found ~50 Mbp (including N's) inserted in panTro5 relative to Clint_PTRv1. It was crude in that it counted, for example, a CIGAR string of 1000I1000D1000I1000D as a 2000 bp insertion, ignoring the neighboring deletions. Looking at the base level of insertions >100 bp in alignments of panTro5 relative to Clint_PTRv1 shows that 47% of insertions have exactly one run of N's with extra bases on each side of the run of N's. This indicates that these insertions are due to extra sequence on the ends of contigs prior to being inserted into scaffolds.

We visualized 100 randomly chosen alignment blocks in MUMmer (*74*) (e.g., **Figs. S8-S10**). In this sample there are roughly 10 times as many alignment blocks with more panTro5 bases than Clint_PTRv1 bases.

panTro5 contains 27,797 gaps (runs of N's) with the mean number of N's in a run being 3.4 kbp. We aligned panTro5 to Clint_PTRv1 to identify gap closures. A panTro5 gap is considered closed if a Clint_PTRv1 contig is aligned at that location in panTro5, including both sides of the run of N's. In total, 52% (14,518) of the 27,797 panTro5 gaps were closed in Clint_PTRv1. This is markedly different from the 96.8% gaps closed by our assembly of orangutan. The lower number in chimpanzee is possibly due to incorrect sequence in panTro5 near runs of N's preventing Clint_PTRv1 contigs from aligning on both sides of the run of N's.

We took 2 kbp of sequence in panTro5 on the left and 2 kbp on the right of each run of N's and mapped these to Clint_PTRv1. In cases in which both flanks mapped to the correct contig, 49% of the time they mapped in opposite orientations indicating that one of the flanks was inverted in panTro5. We reported these and many were fixed in a new version of panTro5 (by inverting 4,495 sequences), but some remain. A BAC spanning one of the remaining problem gaps shows an inversion (and additional duplicated sequence) in panTro5 (**Fig. S11**).

A.5 Scaffold construction

The contigs from the assembly were ordered and oriented into scaffolds using Bionano optical maps. The Bionano Genomics (BNG) Instrument Saphyr was used to generate optical molecules using two restriction enzymes, Nb.BssSI and Nt.BspQI (see **Section E**) and assembled into maps. The contigs were aligned to the consensus maps and placed into scaffolds using the HybridScaffolds suite from the Bionano Access software. HybridScaffolds placed 737 contigs of Clint_PTRv1 into 121 scaffolds. Scaffold N50 was 60 Mbp. **Table S21** shows the metrics for optical maps using both enzymes, as well as the combination of the maps to produce scaffolds.

A.6 Chromosome AGP construction

We built chromosomal-level AGPs (a golden path) for Clint_PTRv1 and Susie_PABv2 without the guidance of the human reference genome, integrating Bionano optical maps, BAC-end sequences, Hi-C and FISH. All contigs (>150 kbp) were initially scaffolded with Bionano. Fish probes were then used to order scaffolds into chromosomes. Fully sequenced large-insert BACs were also used to guide this process.

AGP for Clint_PTRv1 was built mainly relying on FISH probes to identify and 'place' scaffolds. FISH is typically used to visualize chromosomes by using colored probes in metaphase spreads of nuclei to identify translocations and chromosomal aberrations. Over the years, there have been multiple BACs (with FISH mappings) sequenced for each chromosome in chimpanzee and human to explore chromosomal similarities and differences between the two. We used these sets of BACs while leveraging the megabase-lengths of our scaffolds to group, order, and orient them (in that order) into whole chromosomes.

In total, 813 BACs with FISH mappings were aligned to Bionano Clint_PTRv1 scaffolds. The BAC order data for each chromosome was obtained from Mario Ventura (see **section F**) and the sequences were obtained from NCBI. BACs for chrX were obtained from Stanyon *et al.*, 2008 (*75*). The number of BACs per chromosome is shown in **Table S22**.

BAC sequences were mapped to the scaffolds using BLASR and best hits in terms of total bases aligned were considered. Based on the mappings, the scaffolds grouped into 24 categories—one for each chromosome and an unplaced group. Each chromosome set contained scaffolds placed into the set by virtue of the BAC mapping. The BAC set for each chromosome contains approximately one mark in every 4-10 Mbp. The long scaffolds are multi-megabases long and contain multiple alignments from the marker sets, thus 'marking' them into their respective chromosome bins. Scaffolds were thus grouped into chromosome groups. This approach successfully placed 100 scaffolds into 23 chromosomal bins.

Within each chromosome bin, the order of mapping of each set of BAC sequences is known. We use this prior knowledge to lay out scaffold sets into meaningful order of sequences. Multiple BAC mappings within each scaffold make it possible to determine the orientation of the scaffolds (increasing or decreasing order of probe mappings). We ordered all chromosomes by using the data from the FISH alignments.

A.7 Quality control and manual curation

Breaking fused scaffolds: FISH analysis correctly identified two instances where large portions of different chromosomes were fused together into a single Bionano scaffold. For instance, the longest scaffold (169 Mbp) consists of a 45 Mbp segment of chr7 erroneously inserted into a chr1 scaffold. In another case, a 33 Mbp scaffold was of chr10 and chr1 (**Fig. S12**). In both cases, the erroneous scaffold was split, and segments placed into their respective correct chromosomal bins manually.

Assessment of completeness and missing genes: We assayed the gene content of the final Clint_PTRv1 AGP. We found only 57 genes that were not in chromosome-level scaffolds (**Table S23**). These genes were on 32 unplaced contigs, many of which had exceptional read depth.

AGP manual finishing using Hi-C data: We generated Hi-C data (see **Section G**) as an orthogonal way to validate contig scaffolding. Hi-C reads were aligned to the Clint_PTRv1 AGP (chromosomes) using BWA-MEM (-5 flag), generating a high-resolution proximity heatmap. Juicebox (*67*), was used to visualize the concordance of hi-c data with the scaffolds (**Fig. S13**). Four large heatmap discrepancies revealed three scaffolding errors involving chromosomes 1, 5, and 6 (**Fig. S13**; **Table S24**). All four cases were erroneous translocation events of single contigs. Discordant BAC-end mappings at the boundaries of the translocated contigs confirm that these contigs were incorrectly scaffolded. In each of these cases, the misplaced contig was manually removed from the scaffolds and added into the rightful chromosome's unlocalized list.

A.8 Cytogenetic evolutionary rearrangements

Chimpanzee and human chromosomes differ by nine pericentric inversions and one chromosomal telomeric fusion (*19*). Clint_PTRv1 chromosomes constructed using Bionano and FISH captured all nine pericentric inversions correctly. **Table S25** shows the relative position of the inversions in GRCh38 space. Two known pericentric inversions in chr15 and chr9 are not seen in the latest chimpanzee genome (UCSC panTro5). Dot plots of Clint_PTRv1 chromosomes against human GRCh38 chromosomes in **Fig. S14** show the inversions. Many of the pericentric inversion boundaries occur in SD regions, which are difficult to assemble and scaffold. We often see contigs/scaffolds break at these boundaries and assemble into different pieces. In Clint_PTRv1, Bionano scaffolds spanned the boundaries of the whole pericentric inversion in two chromosomes (chr4, chr18) and at least one boundary in three chromosomes (chr5, chr12, chr15) as seen in **Fig. S15**.

B. Orangutan genome sequence and assembly

B.1 Library preparation and sequencing

DNA source: Genomic DNA came from a single Coriell cell line (PR01109; EBV transformed lymphoblast) derived from a female Sumatran (*Pongo abelii*) orangutan "Susie" (Studbook #1044; ISIS #71). Susie is deceased and a former resident of the Gladys Porter Zoo, Brownsville, TX. DNA was prepared as previously described (A.1).

Genome library preparation and sequencing: DNA fragment libraries were prepared and sequenced as noted in A.1. Loading concentrations were titrated empirically for each library, averaging 160 pM for >15 kbp size-selected libraries and 340 pM for >30 kbp size-selected libraries. Susie was sequenced to a coverage of 81.3X (ROI, 3.2 Gbp estimated genome size) or 94.9X (subread, 3.2 Gbp estimated genome size) on 296 SMRT cells, producing 21 million ROI reads with 28 million subreads (**Fig. S16** and **Table S27**).

B.2 Genome assembly

The orangutan genome was assembled using Falcon (same settings and version as Clint_PTRv1) with a length cutoff of 10.7 kbp. The coverage of reads ≥10.7 kbp is 68.9 X (3.2 Gbp estimated genome size). The resulting assembly "Susie_PABv1" is 3.04 Gbp in size and consists of 5,771 contigs with a contig N50 of 11.3 Mbp. There are contigs over 100 kbp in size, with 2,898,711,478 total bases (95% of the assembly size) (**Table S28**). The assembly was error-corrected (see above under chimpanzee assembly) using Quiver, then Pilon with 42X Illumina paired-end reads (accession SRR6029680), then our FreeBayes-based correction pipeline (**Section H**) using those same Illumina reads.

B.3 Assembly quality: Concordance and base-pair accuracy

We assessed the quality of our Susie_PABv1 contigs by aligning BES from the CHORI-276 BAC library (Susie reference source). In order to identify potential misassemblies and to estimate the sequencing accuracy, we followed a procedure similar to the one for Clint_PTRv1. In case of Susie_PABv1, 96.8% of the sequence at the contig level is concordant with the BES data (**Table S29**). **Table S30** shows the sequence accuracy and QV, calculated by assessing sequence differences between aligned BES and sequence contigs.

B.4 Comparison with previous orangutan reference genomes

We compared our assembly (Susie_PABv1) to the latest orangutan assembly on NCBI (GCF_000001545.4), "ponAbe2". ponAbe2 has 315,124 gaps (stretches of N's). Following the same procedure as for chimpanzee to determine whether a run of N's in ponAbe2 has been resolved in Susie_PABv1, we found 96.8% (305,069) of all runs of N's were resolved. To determine the amount of sequence added (or removed) by Susie_PABv1 relative to ponAbe2, we followed the same procedure as with chimpanzee of dividing all ponAbe2 chromosomes into 10 Mbp sequences and aligning each 10 Mbp sequence to Susie_PABv1. Each such alignment yields several alignment blocks (shown in **Fig. S17**). In total, Susie_PABv1 added 54.5 Mbp and removed 3.8 Mbp from ponAbe2 (**Table S20**). We determined that 52.51% of the sequence that closed gaps was made up of repeats.

B.5 Scaffold construction

The contigs from the assembly were ordered and oriented into scaffolds using Bionano optical maps. The Bionano Genomics instrument Saphyr was used to generate optical molecules using two nicking enzymes, Nb.BssSI and Nt.BspQI (see Methods), and assembled into maps. The contigs were aligned to the consensus maps and placed into scaffolds using the HybridScaffolds suite from the Bionano Access software. **Table S31** shows the summary metrics for optical maps using both enzymes, as well as the combination of the maps to produce scaffolds. 588 contigs of Susie_PABv1 were scaffolded into 73 scaffolds with a scaffold N50 of 102 Mbp.

B.6 Chromosome AGP construction and manual curation

We built the orangutan AGP using the same methodology outlined in the chimpanzee section. In this case, 490 FISH BACs were used to place 63 scaffolds into 24 chromosomes. For QC of the ordering and orienting, we relied again on BES, FISH, and Hi-C data.

Manual curation: While we found no interchromosomal chimeric contigs, there were intrachromosomal scaffolding errors. Scaffold_5_101m belonging to chr8 had to be manually split and rejoined because FISH and the BAC ends detected an orientation error within the scaffold (**Fig. S18**). Two other discrepancies in the scaffold orientation were detected on chr15 and chr22 when we checked for known inversions. While literature has it that both chr22 and chr15 are collinear with respect to human chromosomes (*11*) (http://www.biologia.uniba.it/orang/), Susie FISH-based chromosomes showed large ~10 Mbp acrocentric inversions. On further investigation, we found that in chr22, the FISH and BAC ends showed unclear evidence for order and orientation of the contigs within the single 8 Mbp scaffold that contained the inversion. Since the signals were not overwhelmingly strong, we did additional bicolor FISH on these locations. Bicolor FISH confirmed the inversion on chr22 (**Fig. S19**), and confirmed a complex rearrangement event in case of chr15 (**Fig. S20**).

Missing genes: As a quality measure, we counted the number of protein-coding genes that were annotated on the contigs and were missing in the chromosomes. We found 66 genes that were left in unplaced scaffolds. **Table S32** shows the contigs and the protein-coding genes that they contain.

Evaluation using chromosome conformation capture methods (Hi-C): We generated Hi-C data and visualized the contact map using Juicebox (**Fig. S21**). Unlike in Clint, no large interchromosomal translocation events were detected. We checked to see if the large acrocentric inversion in chr15 was detected, but there was little evidence in the contact map.

B.7 Cytogenetic evolutionary rearrangements

In the orangutan, all but 12 chromosomes are collinear with human chromosomes (*11*). **Fig. S22** shows dot plots of all the chromosomes that are not collinear (). Blue dotted lines indicate the boundaries of known evolutionary chromosomal rearrangements with human (*11*), while gray solid lines indicate region of SD (segdup regions for Susie_PABv1 [horizontal solid gray lines] here are derived from regions of high read depth). **Fig. S23** contains dot plots of the scaffolds against the human chromosomes, to identify if the breakpoints of the rearrangements are captured inside a scaffold.

II. Human genome sequencing and assembly

As a control for the higher quality of the human reference genome assembly (GRCh38), we also assembled two additional human genomes using the same assembly pipeline and SMRT sequence data generated as part of the Human Reference Genome Sequencing Consortium (*76*).

C. CHM13 assembly (CHM13_HSAv1)

C.1 Sequence and assembly

CHM13 represents a complete hydatidiform mole and, therefore, is devoid of allelic variation (*77*). CHM13 was SMRT sequenced to ~73.1X subread coverage (3.2 Gbp estimated genome size). Falcon (same version as for chimpanzee) was used with a subread length cutoff of 11.4 kbp. The coverage of reads ≥11.4 kbp is 50.3X. The assembly was error-corrected using Quiver, then Pilon with 41X paired-end Illumina data. Assembly summary statistics is shown in **Table S33**. We subsequently applied our FreeBayes-based correction pipeline (**Section H**).

C.2 Assembly quality: Concordance and base-pair accuracy

We assessed the quality of the CHM13 assembly by mapping BES (CHORI-17) generated from CHM1 (another hydatidiform mole) to CHM13. Our analysis showed that 97.11% (2.73 Gbp) of the genome was supported by concordant best-paired BES (**Table S34**). An analysis of aligned high-quality Sanger data (63.45 Mbp of PHRED > 40) from CHORI-17 BES revealed high sequence identity 99.888% (**Table S35**).

C.3 Scaffold construction

We obtained Bionano optical maps (CMAPS) for the CHM13 genome using restriction enzymes Nt.BspQI and Nb.BssSI. Bionano access software from Bionano Genomics was used to integrate the optical map information with the sequence data and build scaffolds. Bionano Access was also used to build the AGP for the scaffolds. This method primarily relies on ‘cut -site’ information as opposed to sequence reads, thus providing us with an orthogonal approach to assess the assembly. Bionano gave a genome assembly comprising of 105 scaffolds resulting in a 2.8 Gbp assembly with an N50 scaffold size of 82.79 Mbp. Details of the assembly statistics are in **Table S36**.

Bionano scaffold QC: In order to assess the contiguity of the Bionano scaffolds, we mapped BES from the CHORI-17 BAC library to the 105 scaffolds generated by Bionano. Overall 98.268% of the assembly was concordant, 57.27 Mbp (2%) had no BES support of any kind, and 20.77 Mbp (0.7%) had discordant BES support.

D. Yoruban (NA19240) assembly (YRI_HSAv1)

D.1 Assembly

DNA from the HapMap Yoruban (NA19240) lymphoblastoid cell line was sequenced using SMRT genome sequencing to ~115X subread coverage (3.2 Gbp estimated genome size). The reads were assembled using Falcon (same version as chimpanzee) with read length cutoff 11.4 kbp. The coverage of reads ≥ 11.4 kbp is 69X. The assembly was error-corrected using Quiver, then Pilon using 40X paired 125 bp Illumina reads (experiment ID SRX1098167) (assembly summary statistics in **Table S37**). Subsequently the FreeBayes-based indel correction pipeline was applied (**Section H**).

D.2 BES concordance and accuracy

We assessed the quality of the assembly by mapping BES (CHORI-17) generated from a haploid hydatidiform mole (CHM1) against the PacBio assembly (YRI_HSAv1). Our analysis showed that 97.73% of the genome was supported by concordant best-paired BES (**Table S38**). An analysis of aligned high-quality Sanger data (63.45 Mbp of PHRED>40) from CHORI-17 BES revealed high sequence identity 99.84% (**Table S39**).

III. Other genomic methods

E. Bionano genomics optical mapping

Bionano DNA labeling: High molecular weight DNA from the CHM13, Clint and Susie samples was labeled following the OptiDNA protocol. Briefly, cells were embedded into a thin layer of low melting point agarose, using a specialized cassette. The cells were then treated in the cassette with Puregene Proteinase K (Qiagen) and RNase A (Qiagen), resulting in purified DNA protected by an agarose matrix. For each nicking enzyme, a separate reaction was performed. For 2 hours at 37°C, DNA was digested with nicking endonucleases Nt.BspQI and Nb.BssSI (New England BioLabs). Nicked DNA was then incubated for one hour at 50°C with Taq polymerase (New England BioLabs) and fluorescently labeled dUTP (Bionano Genomics). Next, the labeled DNA was incubated for 30 minutes at 37°C with Taq ligase (New England BioLabs) and dNTPs. The samples were then removed from the cassettes, melted, and solubilized using Agarase (ThermoFisher Scientific), and the DNA was counterstained with YOYO-1 (ThermoFisher Scientific).

Bionano data collection: Labeled DNA samples were loaded into SaphyrChips (Bionano Genomics) and run on the Saphyr (Bionano Genomics) instrument. Data were collected until 100-fold coverage of long molecules (>150 kbp) was achieved for both Nt.BspQI and Nb.BssSI samples. Bionano software was used to detect linearized DNA using the YOYO-1 counterstain, and to detect the labeled nick sites on the DNA. Sets of single-molecule maps, equivalent to about 100X haploid coverage, for each sample were then used to construct a *de novo* genome assembly.

Bionano *de novo* assembly: *De novo* assembly was performed using Bionano's custom assembler software program (version 5536 and 5566) based on the Overlap-Layout-Consensus paradigm. Pairwise comparison of all DNA molecules was done to create a layout overlap graph, which was then used to create the initial consensus genome maps. By realigning molecules to the genome maps (Refine-B P-Value 10^{-11}) and by using only the best match molecules, a refinement step was done to refine the label positions on the genome maps and to remove chimeric joins. Next, during an extension step, the software aligned molecules to genome maps (Extension P-Value 10^{-11}) and extended the maps based on the molecules aligning past the map ends. Overlapping genome maps were then merged using a Merge P-Value cutoff of 10^{-15}. These extension and merge steps were repeated five times before a final refinement was applied to "finish" all genome maps (Refine Final P-Value 10^{-11}). Two assemblies were constructed per sample—one for each nickase.

During the extension step, the software identified clusters of molecules that aligned to genome maps with end alignment gaps of size >30 kbp (i.e., over 30 kbp of one side of the molecules did not align), selected out these molecules and re-assembled them. In addition, the final refinement step searched for clusters of molecules aligned to genome maps with internal alignment gap of size <50 kbp, in which case the genome maps were converted into two haplotype maps. The extend-and-split function is essential to identify large allelic differences and to assemble across loci with SDs, whereas the refinement haplotype function can find smaller differences.

E.1 Variant calling

Structural variants (SVs) were called based on the alignment profiles between the *de novo* assembled genome maps against the public human reference assembly GRCh38 as well as against the Falcon (PacBio) sequence assemblies. We required an alignment cutoff p-value of 10-12 to identify the best alignments. SV calling was done for the Nt.BspQI and Nb.BssSI assemblies independently. Significant discrepancies in the distance between adjacent labels or the number of unaligned labels between adjacent aligned labels (outlier p-value 3x10-3) indicated the presence of insertion and deletions. Genome maps whose alignments were in opposite orientations indicated the presence of inversion breakpoints. Maps aligning to different chromosomes or aligning over 5 Mbp apart on the same chromosome would suggest inter and intrachromosomal translocations, respectively.

Insertions and deletions captured by each of the single-enzyme assemblies (Nt.BspQI and Nb.BssSI) were compared and merged into a final SV call set. Insertions and deletions that were within 10 kbp and with over 80% reciprocal size similarity were merged together, and the innermost breakpoints were recorded as the merged variant breakpoints. To minimize false positives, we removed calls whose size was less than 500 bp, calls found by single nickase assembly but with a variant confidence score of <0.5, or calls found by both nickases but with a confidence of <0.3.

Similarly, inversions called by each single-enzyme assembly were merged. Subsequently, for inversions less than 5 Mbp in size, we ran a hierarchical clustering method to group inversion events whose start breakpoints were within 20 kbp and whose end breakpoints were within 20 kbp. Then, the outermost coordinates for each cluster were recorded as the final locations. For inversions 5 Mbp or above in the primate samples, we extracted intrachromosomal translocation calls whose alignments were in inverted orientation. We clustered variants whose breakpoints were within 50 kbp. For each cluster, the median breakpoint coordinates were outputted.

Finally, to generate a final translocation call set, we merged translocations from single-enzyme assemblies whose breakpoints were within 20 kbp. The innermost breakpoints were kept.

E.2 Hybrid scaffolding

The inputs for the two-enzyme hybrid scaffolding pipeline were a sequence contig map file, a Bionano Nt.BspQI genome map file, and an Nb.BssSI genome map file. For each enzyme, the sequence contig map file was generated by running an "*in silico* digest" on the assembly contigs. Then, each single-enzyme genome map assembly was aligned to the sequence contigs to identify potential conflicts. Details on how conflicts are identified and resolved in single-enzyme hybrid scaffolding can be found in (*78*). Afterwards, the two-enzyme hybrid scaffold pipeline combined the conflict-resolution decisions to resolve all chimeric joins detected. The pipeline merged (Merge P-Value of 10^{-11}) the conflict-free genome maps and sequence contigs to generate hybrid scaffolds for each enzyme. Then, by identifying sequence contigs common in both Nt.BspQI hybrid scaffolds and Nb.BssSI hybrid scaffolds, the pipeline then merged the single-enzyme scaffolds into two-enzyme hybrid scaffolds, thus further improving contiguity. Subsequently, with higher label information density on the two-enzyme hybrid scaffolds, the pipeline performed a final alignment (Align Final P-Value of 10^{-9}) between the scaffolds and the sequence contigs in order to anchor shorter sequences that were initially missed by the single-enzyme alignments. Finally, AGP and FASTA files for the scaffolds were generated.

E.3 Bionano maps identified sequence structural differences

The comparison of the Bionano data with the Falcon sequence contigs resulted in a number of 'conflict' regions where the optical maps did not agree with the contigs (**Table S40**). In most of these cases, Bionano identified relatively small insertions and deletions with respect to the sequence. These may be heterozygous differences or sequence errors (e.g. tandem repeat collapses). Most larger events represent chimeric contigs and these were also found and resolved by the hybrid scaffold pipeline during scaffolding. While Bionano has single molecules spanning the large size 'conflict' regions, most of these regions have poor PacBio read support, some with poor BAC concordance even, providing additional evidence for sequence misassembly. This result showed that most of the regions with big structural difference were caused by sequence misassemblies; similar to that showed in Jiao W.B. *et al.*, 2017 (*79*).

F. FISH

BAC clones (roughly 1,000) were used as probes in FISH experiments to develop a comparative cytogenetic framework specific to chimpanzee and orangutan lineages as reported in Ventura *et al.*, 2011 (*80*) and Locke *et al.*, 2011 (*11*), respectively. FISH data corresponding to 1,240 human BAC clones were previously hybridized to chimpanzee and orangutan metaphase chromosomal spreads in a series of bicolor FISH experiments in order to establish synteny and conserved genome order; 460 of these clones were used previously during the assembly of the first draft of the orangutan assembly. Additional FISH experiments were used to refine the Yunis and Prakash cytogenetic breakpoints associated with cytogenetic level rearrangements among the great apes.

Metaphase preparations were obtained from two lymphoblast cell lines obtained from chimpanzee (*Pan troglodytes*, PTR) and orangutan (*Pongo abelii*, PAB). Human metaphase spreads were prepared from PHA-stimulated peripheral lymphocytes of normal donors by standard procedures.

All human probes (BAC-FISH experiments), mostly derived from the RP11 human library, were hybridized to both PTR and PAB in reiterative FISH experiments to exactly define inversion breakpoints and probe order using 2- or 3-color FISH experiments. Probe order has been defined in a totally independent way from human chromosome order selecting probes regularly distributed on the human genome. Splitting signals were interpreted as caused by the occurrence of a breakpoint inside the marker. To reject the possible interpretation that splitting signals were caused by the presence of SDs, additional BAC probes, partially overlapping the splitting clone on both sides, were also used.

Hybridizations were carried out following the FISH protocol described previously by Ventura *et al.*, 2007 (*81*). Extraction of total DNA from BACs was performed according to standard methods. Human key BACs, numbered according to their map position on human chromosomes, were directly labeled by nick translation with Cy3-dUTP, Cy5-dUTP (GE Healthcare), or Fluorescein-dUTP (Invitrogen) (*81*). Two hundred nanograms of labeled probe were hybridized on metaphase spreads; hybridization was performed overnight at 37 °C in 2× SSC, 50% (v/v) formamide, 10% (w/v) dextran sulfate, 3 µl C0t-1

DNA (Roche), and 5 µl sonicated salmon sperm DNA, in a volume of 10 µl. Post-hybridization washing was performed at 60 °C in 0.1× SSC (three times, high stringency). Chromosome identification was obtained by DAPI staining, producing a Q-banding pattern. Digital images were obtained using a Leica DMRXA epifluorescence microscope equipped with a cooled CCD camera (Princeton Instruments). Cy3, Cy5, Fluorescein, and DAPI fluorescence signals, detected with specific filters, were recorded separately as grayscale images. Pseudocoloring and merging of images were performed using Adobe Photoshop software.

G. Generation of Hi-C data

DNase Hi-C was performed on 1-2 million cells per species as described in Ramani *et al.*, 2016 (*82*). In brief, cells were crosslinked, digested with DNase I, followed by end-repair, dA-tailing, and ligation of bridge adaptors, and then *in situ* phosphorylation and ligation. After reversing crosslinks, genomic DNA was purified, sheared (Covaris), and subjected to streptavidin-based pulldown to recover biotinylated fragments. These were converted to next-generation sequencing (NGS) libraries by dA-tailing, ligation of sequencing adaptors, PCR amplification, and Ampure (0.8x) based purification. NGS libraries were sequenced on the NextSeq 500 (SN# NS500488) generating paired-end 150 bp reads (**Table S41**).

H. Indel correction

Premature truncating variants (PTVs; indels) remain problematic, even after error-correction software (Quiver and Pilon) remove over 90% of the initial PTV errors. For our Yoruban assembly (YRI_HSAv1) we observed 2,053 PTVs (**Table S42**). But based on human exome data, it is estimated that there are, on average, 120 PTVs (*83*) in an individual. Manual inspection of several hundred PTVs revealed that these variants are heterozygous, have high mapping quality, and often mapped only a few bases from another indel (**Fig. S24**). The assembly sequence should have had either, but instead it had neither or both. This strongly suggests that Quiver and Pilon fail to incorporate one of the correct alleles at such heterozygous loci.

To correct false PTV errors, we developed a pipeline to be used after Quiver and Pilon (available at https://github.com/EichlerLab/indel_correction_pipeline). This pipeline reduced the number of PTVs in YRI_HSA from 2,053 to 111, closer to what we would expect. We applied the false indel-correcting pipeline to YRI_HSAv1, CHM13_HSAv1, Clint_PTRv1, and Susie_PABv1. GSMRT3.2 had previously been corrected with a similar method. The indel-corrected assemblies are part of the V2 assembly release.

The pipeline has seven steps, outlined below. The input is a reference genome and Illumina reads from the same organism, and the output is a corrected assembly.

1. Align Illumina short-read data to the genome assembly, using a matched sample, then call indels "List1" using FreeBayes (requiring a mapping quality of 20).
2. In cases in List1 in which there are two indels ≤6 bp apart, remove the second indel. Also, remove non-indels from List1. In the example of **Fig. S24**, this will keep the 1st column of insertions and remove the second column of insertions.
3. Change the reference to "correctedFreeBayes1" by inserting or deleting bases in List1. This operation is carried out using the VCF-consensus program (*84*).

4. Identify the remaining high-confidence PTVs by looking only at indels in coding sequence that are length not a multiple of three (the ones that change frame) and apply filters that eliminate lower confidence indels (in regions of SD, high-depth regions of the assembly, telomeric, centromeric, small contigs (less than 500K), and mapping within 10 kbp of the ends of contigs). Empirically we found most of these remaining PTVs were heterozygous. To change them to the benign variant that restored frame, we follow the following steps.
5. Align the same Illumina reads from #1 to correctedFreeBayes1.
6. Run FreeBayes (again) using alignments from #5 but just at the sites (±100 bp) of PTV from #4.
7. Change correctedFreeBayes1 to the variants in #6 (both substitutions and indels) using VCF-consensus giving "CorrectedFreeBayes2"—the output of this pipeline.

IV. Comparative analysis

I. Contig N50 vs. Syntenic N50

We sought to compare the synteny of the nonhuman primate (NHP) assemblies to GRCh38, including older versions of the ape reference genomes. For this analysis, we used all contigs, including all placed contigs, all unplaced contigs and, when present, unlocalized contigs. The total number of contigs for each assembly is shown in **Table S43**. We aligned the genomes against GRCh38 using Minimap (https://github.com/lh3/minimap), a tool devised to quickly identify long approximate matches between two genomes. Syntenic blocks were derived from Minimap alignment for every assembly and N50 calculated as the minimum length where 50% of the whole assembly (in this case, 3.2 Gbp) was contained in blocks whose length was equal to or greater than the N50. The scatterplot of contigN50 and syntenic N50 can be found in **Fig. 1**, and the underlying data are shown in **Table S44**.

J. Pairwise alignment between NHPs and GRCh38

We assessed the proportion of the human genome (GRCh38) covered by NHP alignments. First, we aligned each of the great ape PacBio genomes to GRCh38 and counted the number of aligned bases (**Table S45**). It should be noted here that the numbers obtained are aligner-specific and parameter-specific. In this instance, the percentages obtained are generated from BLASR (githash 7cc3379) with parameters -clipping hard -alignContigs –sam -minMapQV 30 -nproc 6 -minPctIdentity 50. Thus, this number is not the total number of assayable GRCh38 bases but rather the bases that align with this aligner and these parameters.

When the individual alignments are grouped together (intersecting alignments), 2.6 Gbp of bases are recovered, making up 86.2% of GRCh38. We then assessed the portion of GRCh38 that was covered by each assembly contig. In case of the chimpanzee, orangutan, gorilla and the Yoruban, less than one percent of the total assayable bases are covered by more than one contig (**Table S45**).

Finally, we measured the intersection of GRCh38 alignment coverage for our SMRT assemblies. In parallel, the same analysis was done for the prior NHP reference genomes (panTro3, P_pygmaeus_2.0.2 and gorGor3). **Table S45** summarizes the results of these studies. In total ~85% of GRCh38 is covered by syntenic alignments of all the primate genomes. In contrast, only ~40% of GRCh38 is covered by the prior nonhuman short-read assemblies.

K. Alignment-based karyotyping

Rapidly evolving regions of the great ape genomes break alignment synteny with the human reference genome. Alignment fragmentation can also be attributed to assembly error in ambiguous, recently duplicated regions of the genome. To untangle evolutionary events from assembly errors, we studied patterns of alignment fragmentation (alignments smaller than 1 Mbp against GRCh38) in the great ape genomes compared to GRCh38. There were 164 regions of alignment fragmentation where the two human assemblies (CHM13_HSAv1 and YRI_HSAv1) and three NHP assemblies all overlapped. As expected, the majority of these events, 128/164, overlapped SDs in GRCh38. We next looked for alignment fragmentation discrepancies between humans and NHPs (**Fig. S25**). There were 70 cases where the human alignments against GRCh38 did not have fragmentation, but all NHPs did (44 were in SDs); alternatively, there were 103 regions where the two human assemblies had alignment fragmentation not seen in NHPs (97 were in SDs). In the unique regions of GRCh38, eight of the NHP-fragmented regions overlapped genes and none of the human fragmented regions did (**Table S46**). One region overlapped with a previously characterized human 1q21 duplication of a transcriptionally active histone cluster (*85*).

We measured the empirical probability of the 70 regions of great ape alignment fragmentation. The permutation test shuffled the small alignments (<1 Mbp) for all five genomes. Out of one million permutations, no test had more than 70 regions (**Fig. S26**) and the highest count observed was 36, resulting in an empirical p-value less than 2.12e-06.

L. Comparative repeat analysis

RepeatMasker (v3.3.0) was used to annotate repeats on all five assemblies (CHM13_HSAv1, Yri_HSAv1, Clint_PTRv1, Susie_PABv1, and GSMRT3). Parameter specifications for RepeatMasker that were used are: -no_is -xsmall -s -e wublast -species primates. **Table S47** shows the repeat summary for the two human genomes.

The gorilla, and to a lesser extent the chimpanzee, show a significant increase in satellite content when compared to human and orangutan. We compared the repeat content of these and previous genome assembles considering both mapped (AGP) and unmapped sequenced contigs (**Table S48**). Repeat content of contigs placed in the AGP is remarkably consistent among the SMRT genome assemblies. The addition of unmapped contigs shows a dramatic increase in satellite repeat content for gorilla consistent with the larger size of the gorilla genome (3.5-4.0 Gbp (*10*)).

We plotted the fraction of repeats against the contig sizes for the four new assemblies (**Fig. S27**) and found that it converges towards 0.5 in larger contigs.

M. STR analysis

Previous studies have indicated differential expansion of short tandem repeat (STR) sequences between humans and other NHPs (*27*, *28*). However, these studies suffer from ascertainment bias due to the methods of STR enrichment, low fidelity sequencing in GC-rich regions, low contiguity in repetitive regions, and genome assemblies guided by the human reference genome. The assemblies in this study, combined with a PacBio-based assembly of the Yoruban individual NA19240, avoided all of these sources of bias. This offers a new dataset to test for differential STR expansion between species. We identified orthologous STRs between species so that the length of STR sequences could be compared and the hypothesis of differential STR expansion tested. To do so, we first identified STR sequences using

RepeatMasker v4.0.1 with primate repeat libraries and Tandem Repeat Finder v4.07b with options "5 5 5 80 40 20 10 -m -ngs -h". Between 370,781 and 389,376 STR sites were identified in each assembly (**Table S3**). Due to occasional dropout of repeat annotation, regions annotated as repeats were merged if they were within 25 bp of one another, resulting in 337,905–353,989 STR regions (**Table S3**). To define orthologous STRs, we first identified STR sequences that had 250 bp flanking sequences that were no more than 50% repetitive and mapped the flanking sequences to the other assemblies, requiring 90% of each flank to align. Orthologous sequences were defined as the region between the alignments of the two flanking sequences. The inference of change of STR length could be affected by non-tandem repeat SVs and other variants mapped within the flanking sequences. To account for this, potentially orthologous STR sequences were additionally filtered by masking with Tandem Repeat Finder (options 5 7 7 80 40 20 20 -m -ngs -h) and excluding all pairs with a sequence annotated less than 80% as tandem repeat. The resulting orthologous STR sequences were subject to ascertainment bias based on the original discovery set found by RepeatMasker. Examining the distribution of lengths of STRs for all pairs of genomes (**Fig. S28**) indicated that the RepeatMasker discovery set was restricted to 20 bp and greater, but the inferred orthologous STRs included smaller STRs, allowing for a shorter bias for inferred STR sequences. This bias was less apparent for STRs over 40 bp.

This identified 60,795–61,251 orthologous STR sequences between genomes (**Table S4**). After filtering for tandem repeat content and limiting to STRs of at least 40 bp, we compared the distribution of differences of STR length from each pair of genomes to test for differential expansion using the Kolmogorov-Smirnov (KS) test; we found that all pairwise combinations of differences were minor (up to an average of 1.5 bp) without a particular trend of significance.

To test if there is a bias in expansion of STRs in coding sequences, gene models were mapped from GRCh38 to the NA19240 genome using BLASR (1.MC.rc46) options -alignContigs -minMapQV 30 and lifted over using samLiftover (https://github.com/mchaisso/mcutils). We found no difference in STR expansions in exons (n = 310, p = 0.856, KS test) and untranslated regions or UTRs (n = 2,794, p = 0.162, KS test) (**Fig. S29**).

N. PtERV1 analysis

We investigated PtERV1 insertions in the assemblies of chimpanzee (Clint_PTRv1) and gorilla (GSMRT3.2) as well as Illumina WGS data from 72 great apes from chimpanzee, gorilla, bonobo, human and orangutan lineages (*33*). We used the set of all PtERV1 LTR elements in Clint_PTRv1 (101 full length (>7 kbp in length) and 150 solo-LTR) and GSMRT3.2 (71 full length and 202 solo-LTR) detected by RepeatMasker and mapped 5 kbp of flanking sequence to identify integration sites with respect to the human reference genome (GRCh38). We only counted loci as distinct if the flanking primary alignments mapped within 20 kbp of one another (219 loci for Clint_PTRv1 and 175 for GSMRT3.2). As a second approach, we used paired-end sequence data from Illumina WGS where one end anchored within a PtERV1 LTR element and another mapped to a unique location in GRCh38. We did not observe a single insertion of PtERV1 in genomes of human and orangutan but found 508 loci of PtERV1 integration in chimpanzee (18 samples), bonobo (10 samples) and gorilla (21 samples) genomes. PtERV1 integrations were depleted in genic regions (20% of PtERV1 vs. >40% expectation) and biased in the antisense orientation within introns (**Fig. S30**). Overall, PtERV1 insertions are fourfold more likely to map within preexisting ERV elements compared to a random null model.

The intersection of assembly and one-end anchored PtERV1 datasets resulted in a combined set of 540 PtERV1 elements in GRCh38 space (**Table S7; Figs. S31 and S32**). While >99% of gorilla and chimpanzee PtERV1 sites of integration were non-orthologous (**Fig. S31**), we identified one locus mapping to chromosome 19 whose integration breakpoints were identical at the base-pair level between chimpanzee and gorilla assemblies that was also present in all chimpanzees and gorillas **(Fig. 2e**). We aligned the orthologous sequence with Muscle 3.8.31 and generated a maximum likelihood tree using RAxML 8.2.9 with a GTR+Gamma substitution model (**Fig. S33**) and created a dated phylogeny using BEAST 2.4.7 with a relaxed clock and an offset exponential prior on human-orangutan divergence with a mean of 1 million years and an offset of 12 million years. This resulted in 100% posterior support for chimpanzee-gorilla monophyly and an estimate of 4.7 million years ago (95% HPD: [1.9, 7.2]) for human-gorilla divergence (**Fig. S34**). This supports a model of ILS, with a single PtERV1 insertion event prior to human-gorilla divergence. We aligned the full-length PtERV1 sequences in Clint_PTRv1 and GSMRT3.2 (Muscle 3.8.31) to generate a maximum likelihood tree using RAxML 8.2.9 with a GTR+Gamma substitution model (**Fig. S35**).

O. Single-nucleotide polymorphism divergence

We characterized primate sequence divergence against the human reference genome, GRCh38. Consistent reports of sequence divergence, chimpanzee has 1.26% divergence, gorilla has 1.60% divergence, and orangutan has 3.12% divergence compared to GRCh38 (**Table S49, Fig. S36**). For the diploid Yoruban (YRI_HSAv1) assembly and the haploid hydatidiform mole (CHM13_HSAv1), the divergence is 0.12% and 0.10%, respectively. Chromosomes 8 and 16 had some of the highest species-level differences between divergence estimates (**Fig. S37**). We also examined patterns of lineage-specific SNVs/indels (**Fig. S38**).

V. Transcript analysis

P. Transcript Analysis

P.1 Iso-Seq full-length cDNA sequencing

Primate induced pluripotent stem cells (iPSC) lines and RNA prep: Human, chimpanzee (*Pan troglodytes*), and gorilla (*Gorilla gorilla*) iPSC lines (derived using retroviral OSKM method) were previously reported in Marchetto *et al.,* 2013 (*59*) and Ramsay *et al.,* 2017 (*87*) (**Table S50**). Orangutan iPSC lines were derived from a fibroblast cell line (11045-4593) obtained from Josephine (DOB: 1960 wild capture), a female Sumatran orangutan (*Pongo abelii*) whose fibroblasts were collected as part of Oliver Ryder's "Frozen Zoo" initiative. Reprogramming was carried out with the CytoTune 2.0 kit (ThermoFisher), a Sendai virus-based, non-integrating method. The orangutan cell line was maintained on E8 Flex media on vitronectin (Gibco/ThermoFisher) feeder-free substrate and reprogrammed after the ninth passage.

Primate iPSCs were grown as described in Marchetto *et al.,* 2013 (*59*) with the following modifications: Cells were cultured in the presence of the 3iL factors described in Chan *et al.*, 2013 (*88*). Upon optimization for primate iPSC culture, the final conditions were: Recombinant Human Leukemia Inhibitory Factor (LIF) 0.4 ng/ml (ProSpec-Tany TechnoGene Ltd: cyt-644); BIO 80 nM, PD0325901 40 nM; Dorsomorphin 80 nM (Cayman Chemical: 13123, 13034, 11967).

Trizol and TURBO DNase (ThermoFisher) were used in RNA isolations. RNA were further purified using Direct-zol RNA MiniPrep and RNA Clean & Concentrator-5 (Zymo Research).

Iso-Seq methodology: Double-stranded cDNA was synthesized by a modified version of the standard Iso-Seq template preparation protocol to incorporate a barcode/molecular identifier at the end of each strand. This helps facilitate deconvolution of PCR duplicate sequences versus unique founder molecules.

Specialized poly-dT oligonucleotides to prime first-strand cDNA synthesis were synthesized (Integrated DNA Technologies) with the following configuration:

5'-AAGCAGTGGTATCAACGCAGAGT(BC 16bp)TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN-3',
where the sequence BC16bp encodes one of 96-16 bp barcodes, V=(A,G,C), and N=(any base)

Oligonucleotides for second-strand synthesis were synthesized (Integrated DNA Technologies) with the following configuration:
5'-AAGCAGTGGTATCAACGCAGAGT(BC16bp)ATACGATTTAGGTGACACTATAGG-3'

where the sequence BC16bp encodes one of 96-16 bp barcodes.

The template switch oligonucleotide (SP6), a chimeric RNA-DNA sequence, was synthesized (Integrated DNA Technologies): AAGCAGTGGTATCAACGCAGAGTACATrGrGrG.

For cDNA amplification, the 5' flanking sequence in the first- and second-strand oligonucleotides was utilized for PCR: /5Phos/AAGCAGTGGTATCAACGCAGAGT.

Total RNA from iPSCs was harvested using TRIzol (ThermoFisher) and polyA RNA was purified with oligo-dT magnetic beads (Dynal, ThermoFisher) by manufacturer's instructions.

PolyA RNA (10 ng) was reverse transcribed in a 10 µL reaction containing 50 mM Tris-HCl (pH 8.3 at 25°C), 75 mM KCl, 3 mM MgCl2, 10 mM DTT, 0.5 mM dNTPs, 100U of Maxima RNaseH- RT (ThermoFisher), 5 µM SP6 template switch oligo and 10 pmols barcoded-oligo-dT primers. An equimolar mix of 96 barcodes was used to add a degree of molecular indexing. Reactions were incubated as following:

45°C for 1 hr
55°C for 30 min
45°C for 30 min
85°C for 10 min

After the heat kill step, the first-strand cDNA was purified by precipitation on magnetic beads (1x AMPure PB; PacBio). The recovered material was subsequently carried into a 50 µL second-stranding reaction in 1x Takara LA Taq HS buffer (Clontech), 200 mM dNTPs, 2.5U of Takara LA Taq HS (Clontech) and 0.5 µM barcoded SP6 second-stranding oligo. This oligo anneals to the 3' ends of the first-strand cDNA at the SP6 sequence added by the strand switch reaction. An equimolar mix of 96 barcodes was used for the second-stranding reactions, incubated as following:

95°C for 1 min

65°C for 10 min

The second-stranding reaction was immediately stopped by depletion of primers by Exonuclease I (New England BioLabs; 10U) and dNTPs by alkaline phosphatase (rSAP: New England BioLabs; 1U) at 37 °C for 20 min. The double-stranded cDNA ("founder molecules") were purified by precipitation onto magnetic beads (0.5X AMPure PB; PacBio).

The double-stranded cDNA (20% of founder molecules) was amplified by a 100 μL PCR reaction in 1x Takara LA Taq HS buffer (Clontech), 250 mM dNTPs, 5U of Takara LA Taq HS (Clontech) and 0.5 μM of the PCR primer. Reactions were incubated as follows:

95°C for 1 minute
95°C for 30 sec |
68°C for 30 sec | *12 cycles*
72°C for 10 min |
72°C for 10 min

Amplified double-stranded cDNA was purified by precipitation onto magnetic beads (0.5X AMPure PB; PacBio). The cDNA was size fractionated by an automated gel electrophoresis and recovery instrument (SageELF, Sage Sciences). Size fractions were then assayed on a Bioanalyzer High Sensitivity Chip and amplified in batches (~1-2 kbp, 2-3 kbp, 3-4 kbp, 4-6 kbp) with the same amplification conditions as the prior PCR. 1-3 kbp fractions were run through 5 cycles, while larger fractions required 8-10 cycles. Final cDNA was purified by precipitation on magnetic beads (0.5X AMPure PB; PacBio) and SMRT sequencing libraries were prepared according to manufacturer guidelines (SMRTbell Template Prep Kit 1.0, PacBio). Final libraries were purified by two sequential precipitations on magnetic beads (2 × 0.5X AMPure PB, PacBio) and assayed both by fluorometer (Qubit, ThermoFisher) for dsDNA concentration and Bioanalyzer (DNA 12000 chip, Agilent) for size.

SMRT sequencing was performed using the P6-C4 chemistry on the PacBio RS II instrument with 6-hour movies. Insert size distributions of putative full-length cDNA were visualized as part of Iso-Seq QC.

P.2 Illumina RNA sequencing methods

~5 ng of A^+ RNA from each source (human iPSC, chimpanzee iPSC, gorilla iPSC, orangutan iPSC) was used as input for the TruSeq Stranded mRNA-seq kit (Illumina) with parameters set for ~150 bp insert size libraries. Final purified libraries were assayed both by fluorometer (Qubit, ThermoFisher) for dsDNA concentration and Bioanalyzer (DNA12000 chip, Agilent) for size. Sequencing-by-synthesis (SBS) was performed on a HiSeq 2500 (human, chimpanzee, gorilla; 2 × 125 bp reads) and NextSeq (orangutan; 2 × 150 bp reads). Reads were demultiplexed using deML (*89*) and trimmed of adapter and low-quality sequence using Trimmomatic (*90*) following QC by FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ accessed 12 July 2016). Reads were mapped to GRCh38 using STAR ((*91*); https://github.com/alexdobin/STAR) and further QC checks were performed using QoRTs (*92*) (**Table S51**).

P.3 Clustering using ICE (iterative clustering algorithm) and transcript analysis

Final full-length Iso-Seq transcripts were generated using ToFU (Transcript isOforms: Full-length and Unassembled; (*93*); https://github.com/EichlerLab/isoseq_pipeline commit 4e4005a) which trims away primer sequences and identifies the transcribed strand orientation based on the location of the polyA tail. To enrich for longer transcripts, size selection (Sage ELF electrophoresis) was performed prior to the sequencing runs for a portion of the SMRT cells.

The raw data were derived from 114 SMRT cells distributed amongst the four species shown (**Table S52**). The cDNA were isolated from various fractionation experiments resulting in an uneven overall distribution. The size distributions of the full-length non-chimeric (FLNC) reads, which correspond to candidate full-length transcripts, are displayed in **Fig. S39**.

Clustering using ICE: In order to obtain high-quality isoform consensus sequence, each sample was run through a custom version of the Iso-Seq bioinformatics pipeline (*93*). Briefly, for each sequencing molecule, an intramolecular circular consensus read (CCS read) was generated from molecules that contain at least one full-pass read. The CCS reads were then classified as full length if the terminal amplification primer sequences were observed on the both ends of the read and a polyA tail was observed. In our modified version (https://github.com/EichlerLab/isoseq_pipeline commit 4e4005a), the full-length (FL) reads were then mapped to their reference with GMAP2 and partitioned by region such that no partition contained reads from multiple resolved copies of an SD. Each partition was then run through an ICE, where each cluster contains FL reads that belong to the same isoform. Each FL read can belong to exactly one cluster. After ICE, non-FL reads were mapped to the ICE consensus sequences; non-FL reads were allowed to map to more than one cluster and partition. Finally, Quiver was used to create a polished consensus sequence for each cluster. Based on the Quiver consensus predicted accuracy, only sequences that have a predicted accuracy of more than 99% were deemed high quality (HQ) and used for the next part of the analysis.

The ICE transcript alignments were used as input to ANGEL (https://github.com/PacificBiosciences/ANGEL). ANGEL is a tool to predict open reading frames in full-length transcript sequences. Predictions are graded as confident, likely and suspicious.

Iso-Seq cross-genome mapping (Iso-Cross): In an effort to identify species-specific exons or transcripts, all Iso-Seq FLNC reads, ICE transcript models, and ANGEL open reading frame predictions were aligned to all of the primate genomes using GMAP2 (version 2015-07-23). Mappings with <20% of bases matching were discarded; mappings with MQ of 0 were retained. The number of Iso-Seq reads mapping to the genomes is in **Table S53**.

Iso-Cross was also used to compare Clint_PTRv1 and Susie_PABv1 to previous assemblies of their respective species. The increased contiguity of the PacBio genome assemblies improved transcript mapping in chimpanzee (**Figs. S40 and S41**) and orangutan (**Fig. S42**). Chimpanzee transcript models, produced by ICE, were mapped with GMAP2 to both Clint_PTRv1 and panTro3. The number of matching bases per transcript was on average 71 bp greater for the Clint_PTRv1 assembly, totaling 4.8 Mbp across all transcripts (including overlapping transcripts). The procedure was repeated for orangutan transcripts, where Susie_PABv1, on average, had 93 more mapped bases per transcript than ponAbe2, with a net gain of 5.1 Mbp.

P.4 Identification of unannotated and species-specific exons

We took the following approach to identify previously unannotated novel exons in the human genome in our data. homGeneMapping was used to project the coordinates of all Iso-Seq intron junctions as well as all transMap-derived junctions between all species. For human, GENCODE V27 was used. For each AugustusPB prediction in a given species, the number of Iso-Seq reads supporting it in all species were counted. A novel prediction must have at least two Iso-Seq reads supporting it in the current reference, and not match an annotated junction in any other species. These were then split into novel splices and novel exons by comparing overlap with transMap in the current reference. A novel exon must have no overlap with a transMap exon to be considered. All novel exons not associated with transMap-derived annotations were then discarded. Finally, the list was filtered by hand-removing all exons associated with noncoding genes, or with coding genes of lower quality, such as those with automated identifiers. Additionally, for novel human exons, we investigated the expression pattern of the identified isoforms that contained those exons using Kallisto (v. 0.42.4) (*94*) with the GTEx dataset (dbGaP version phs000424.v3.p1) (*95*). Expression estimates are displayed in transcripts per million (tpm) for the top three expressing tissues.

Additionally, we identified exons specifically lost or gained in human versus other great apes that may have an effect on protein-coding sequence (**Table S1.2-S1.3**). We selected exons contained in at least five human FLNC reads but without overlap from chimpanzee, gorilla, or orangutan FLNC reads. The inverse experiment was also performed to identify human exon losses. We specifically looked at exons with lengths that were a multiple of three. Exons were inspected manually and removed if more than one junction read was seen that disputed its absence in other species or if there was evidence of a mapping artifact. The remaining hits were annotated with gene names and are shown with the number of supporting reads (in human for exons gained, in chimpanzee for exons lost) as well as their location in both GRCh38 and Clint space. When multiple exons from the same gene or locus were identified, only the first is shown. Note that this method is sensitive to genes and exons whose expression is lower than the detection level in one species but not the other.

P.5 Single-cell cortical expression analysis

We investigated gene expression during human and chimpanzee cortical development using single-cell gene expression data from cerebral organoid models (*55*, *56*) and from primary cortex (*57*, *66*). We used Trim Galore 3.7 to trim adapter sequence and HISAT2 to align human primary cell and organoid reads to GRCh38 and chimpanzee reads to Clint_PTRv1. Counts for each cell were performed using subread-1.5.0 function in featureCounts. After counts were obtained, we normalized to counts per million. We removed any cells with fewer than 1000 genes detected or greater than 20% of reads mapping to mitochondrial or ribosomal genes. For each dataset, we independently performed Louvain clustering on single cells based on Jaccard distance in the space of the first 30 principal components of variation (*72*). We then identified cortical radial glia as clusters distinguished by canonical markers including *GLI3* and *FOXG1*, and excitatory neurons as clusters distinguished by canonical markers including *NEUROD6* and *FOXG1*. By these metrics, the primary cortical analysis contained 137 radial glia and 183 excitatory neurons, the human organoid dataset contained 123 radial glia and 53 excitatory neurons, and the chimpanzee organoid dataset contained 113 radial glia and 97 excitatory neurons. The Nowakowski *et al.* primary cell dataset contained 4,261 cells from across stages of cortical neurogenesis and was used to examine the expression of genes with candidate functional mutations across cortical development. TSNE coordinates were calculated in PCA space (independent of the clusters) using the fast TSNE command in the Seurat R package. For the cerebral organoid datasets, we performed differential expression between homologous human and chimpanzee cell types using a likelihood ratio test with a bimodal zero inflated distribution from the Seurat R package, and we applied Bonferroni correction. In total, we identified 785 genes differentially expressed between human and chimpanzee radial glia, excitatory neurons, or both cell types using this method. Of these 785 genes 224 (28.5%) were not assigned gene expression values in the previous datasets processed against older versions of the chimpanzee genome and 671 (85%) were not previously reported as differentially expressed.

VI. Gene annotation

Q. Quality control using protein-coding exons (CCDS QC)

Exon sequences from the CCDS set downloaded from UCSC table browser (CCDS database release 20,09/08/2016; https://www.ensembl.org/info/genome/genebuild/ccds.html) were mapped to Clint_PTRv1 and Susie_PABv1 as well as their counterparts in the current reference genomes, ponAbe2, panTro5 and ponAbe2, to assess assembly quality. Exons >20 bp were considered for alignment. In total, 175,389 exons representing 16,939 genes and 30,317 CCDS sequences were mapped to the chromosomes of both assemblies using BWA-MEM (-T -1000 -k10 for a very sensitive search). All calculations were done at the exon-mapping level (**Table S1.1**). In chimpanzee, while 94.1% of genes mapped at full length (genes are considered to map at full length if all the exons constituting it map at full length) in both assemblies, Clint_PTRv1 had 193 more genes mapping at full length than panTro5. In the case of orangutan, 3,771 more genes mapped at full length to Susie_PABv1 than in ponAbe2. Missing or unmapped exons are defined as those that contain <10% of total matching bases in the exon. In panTro5, there were 98 missing exons whereas in Clint_PTRv1 this number is reduced to 77 exons. In ponAbe2, there were 797 missing exons, which are reduced to 273 in case of Susie_PABv1. Gaps in the assembly and assembly errors often lead to frameshifts and truncated alignments of exons. We found 1,748 exons with truncated alignments in panTro5 and 1,481 such exons in Clint_PTRv1. In the case of the orangutan, we found 11,846 exons with truncated alignments in ponAbe2 and 4,297 exons with truncating alignments in Susie_PABv1. For the truncated alignments, we checked if they were truncated due to gaps in the sequence. In case of Clint_PTRv1, only 6 exon alignments intersected with gaps whereas in case of panTro5 we found 40 such exons. In case of Susie_PABv1, only 7 exon alignments intersected with gaps, whereas for ponAbe2, 1,979 exon alignments intersected with gaps.

R. Gene annotation of chimpanzee and orangutan

R.1 Comparative Annotation Toolkit (CAT)

CAT (https://github.com/ComparativeGenomicsToolkit/Comparative-Annotation-Toolkit; commit f89a814) was used for genome annotation of Clint_PTRv1 and Susie_PABv1 in addition to re-annotating GSMRT3.2 and the current great ape reference genomes. CAT leverages a reference-free multiple genome alignment produced by Progressive Cactus (*70*) to project the high-quality annotation present on human assembly hg38 (*96*) on to all other genomes in the alignment.

The CAT pipeline produces a filtered annotation set by first using transMap (*97*) to project the GENCODE V27 human annotation on to the great ape assemblies. The transcript projections marked as protein coding in the reference are then used as hints to two parameterizations of AUGUSTUS (*98*, *99*). The first, AugustusTM, strongly believes the transMap projections but cleans them up by enforcing a valid coding gene model. The second, AugustusTMR, also tries to reproduce the transcript but takes in additional RNA-seq and Iso-Seq information.

In addition to these projection-based approaches, two gene-finding parameterizations of AUGUSTUS are used. The first, AugustusCGP, simultaneously predicts transcripts in all genomes in a Progressive Cactus alignment making use of RNA-seq, Iso-Seq and annotation information (*71*). The second, AugustusPB, makes use of Iso-Seq read information to perform transcript predictions on a single genome basis, looking for alternative isoforms.

The resulting five transcript sets are then filtered, scored and combined using a consensus-finding algorithm. This algorithm ranks possible orthologs based on their fidelity to the parent transcript, as measured through projections made by the tool homGeneMapping in addition to transcript-transcript alignments. Transcripts are also ranked based on their extrinsic support. AugustusCGP and AugustusPB predictions are incorporated into the transcript set as either possible novel isoforms of known genes or as novel loci that do not overlap any transMap projections. In this way, the CAT pipeline aims to capture the full landscape of orthologous and paralogous transcripts present in all of the great ape genomes.

For this project, a Progressive Cactus alignment was generated for human hg38, Clint_PTRv1 (chimpanzee), GSMRT3.2 (gorilla), and Susie_PABv1 (orangutan), and the existing gibbon (nomLeu3), rhesus macaque (rheMac8), squirrel monkey (saiBol1), bushbaby (otoGar3) and mouse (mm10) assemblies were used as outgroups. An identical alignment was also generated using the existing gorilla (UCSC gorGor4), chimpanzee (UCSC panTro4) and orangutan (UCSC ponAbe2) reference assemblies. To provide extrinsic hints to AUGUSTUS, RNA-seq reads were obtained from SRA (**Table S54**). RNA-seq as well as Iso-Seq were performed on human, chimpanzee, gorilla and orangutan iPSC lines. Logistic regression was performed to parameterize the AugustusCGP objective function using a subset of the Cactus alignment. Filtering flags were applied to the AugustusCGP and AugustusPB transcript incorporation process—predictions were required to have at least two introns and 80% splice junction support to be considered. An AugustusCGP or AugustusPB transcript must provide at least one splice junction not in the comparative annotation set that is also supported by RNA-seq or Iso-Seq to be included.

The analysis in **Fig. 1d** was performed by looking at the post-filtering transMap transcript sets in each of the SMRT NHP assemblies as well as the existing NHP assemblies. Only protein-coding transcripts that passed filtering in both genomes were considered. The barplots (**Fig. 1d**) on the right report the number of transcript projections with no change in coverage or identity in light colors, and with a change in dark colors. Only the transcripts with a change are measured in the boxplots on the left to prevent the zeros from hiding the distribution. In **Table 2**, the number of novel genes and novel isoforms are reported. Novel genes are defined as loci predicted by AugustusCGP or AugustusPB, have support from RNA-seq or Iso-Seq, and have no overlap with post-filtering transMap projections. Those loci are further differentiated into being either possibly paralogous or putatively novel based on whether they overlapped unfiltered transMap projections or not. Similarly, the novel isoforms are defined by assigning AugustusCGP and/or AugustusPB predictions to a transMap gene based on coordinate and exonic overlap. A transcript prediction is included in the annotation set as a novel isoform if it contains one or more splice junction that is supported by RNA-seq and/or Iso-Seq that is not present in the transMap annotation set. The full set of novel loci and isoform predictions are in **Table S1** (Annotation 1.4-1.6). For novel isoforms, the exact coordinates of all novel splices are included. For novel genes, they are reported as either putatively novel or possibly paralogous.

R.2 Annotation results and validation

CAT identified 56,978 genes (198,909 transcripts) in Clint_PTRv1, 57,067 genes (199,107 transcripts) in GSMRT3.2, and 56,387 genes (197,647 transcripts) in Susie_PABv1. Of these genes, 56,714 (97.3% of total in GENCODE V27) were identified as orthologs to GENCODE V27 for Clint_PTRv1, 56,739 (97.4%) for GSMRT3.2, and 56,147 (96.4%) for Susie_PABv1. See **Table 2** for the full annotation metric list.

To validate the CAT set, the FLNC Iso-Seq reads were given to the ICE pipeline, which is a PacBio software tool for performing isoform clustering and consensus finding. The resulting isoforms were then compared to the CAT set, as well as GENCODE V27 in human as a baseline in both an exact and fuzzy matching scheme. With fuzzy matching, 82.1% of ICE isoforms matched a GENCODE V27 annotation in human, 82.1% matched the CAT set for Clint_PTRv1 and 80.4% matched the CAT set for Susie_PABv1. See the CAT manuscript for more details about this analysis (*22*).

VII. Structural variation

S. Structural variation in great apes

S.1 Methodology and validation

SV calling: SVs in the new assemblies were called against the human reference genome (GRCh38). The SV discovery pipeline (**Fig. S43**) consists of a modified version of BLASR (designed for contig-level alignment) (https://github.com/mchaisso/blasr; commit 2956caed6de86518c48af78c9e0cd80aa307e412) and a CIGAR parsing program "printGaps." The full pipeline can downloaded at https://github.com/zeeev/smartie-sv. For this analysis, we omitted SVs that were smaller than 50 bp, within 0.5 Mbp of the centromere, or within 100 kbp from the telomere. Regions (1 Mbp windows) with more than 50 alignments or less than 0.5 coverage were also excluded from the analysis. PacBio contigs less than 200 kbp were also excluded as many have exceptional PacBio depth of coverage and likely represent high-identity collapsed duplications.

Inversion calling: Inversions were detected as described in Huddleston *et al.*, 2017 (*77*). Whole-genome alignments were used to partition orthologous sequences into 1 Mbp sequences using samSubSeq (https://github.com/mchaisso/mcutils/releases/tag/0.9), and compared using screenInversions (https://github.com/mchaisso/invcheck/releases/tag/0.9.1; options -j 5 -w 2000 -r --noClip). The screenInversions method detects all exact matches between two sequences (wordsize = 11) for both forward and reverse directions, and defines an inversion from a collection of matches from the reverse strand that when reverse-complemented, increase the longest increasing subsequence score of all forward matches.

SV validation: We performed two different SV validation experiments relying on assembled BAC sequences. The first validation experiment targeted large SVs that were in 56 genic regions. At least one NHP BAC was constructed for each region, but many regions had multiple apes represented. In total, 100 BAC clones were analyzed; only five did not support the smartie-sv call (**Table S8**). The second validation experiment measured the accuracy of small events across a randomly chosen set of CHM13_HSAv1 BACs. We sequenced and assembled, 30 CHM13_HSAv1 BACs, no shorter than 50 kbp. Small events on these BACs were called and compared to the smartie-sv calls in the BAC regions. This procedure resulted in 50 BAC-based SV calls, of which 43 were recovered in the smartie-sv CHM13_HSAv1 call set. For small events (<500 bp) we calculated a sensitivity of 0.86 (43TP/50P) and a false discovery rate of 0.17 (9FP/43TP + 9FP).

Lineage assignment: We determined allelic sharing by 50% reciprocal overlap (using BEDTools). For insertions, we used the event length for overlap in GRCh38 space.

After SV calling we screened for contigs that had a high number of inserted and deleted bases. Assembly error and mapping error are two methodological sources of inflated false positives. To find outlier contigs we fit the SV count and contig length data with a linear model, with the expectation of a positive slope and a high correlation within the data. Surprisingly, the human assemblies (CHM13_HSAv1 [r2=0.7] and YRI_HSAv1 [r2=0.4]) had the lowest correlation between contig length and number of SVs (**Fig. S44**). The correlation in the NHPs ranged from 0.87 to 0.71. The YRI_HSAv1 SV call set contig 000287F_1_4175947_quiver_pilon stood out as it was 4.1 Mbp in length and contained 327 insertions and 134 deletions. Similarly, contig 000028F_1_31770352_quiver_pilon was an outlier in the CHM13 SV call set (highest SV count).

Further investigation of the YRI_HSAv1 contig 000287F_1_4175947_quiver_pilon showed a high number of SVs localized around a large complex SV (**Fig. S44**). This SV spans a gap in the GRCh38 reference genome. The BLASR alignment becomes fragmented within the SV resulting in SV count inflation (**Fig. S45A**). The dotplot of the SV shows a large repeat expansion in YRI_HSAv1relative to GRCh38 (**Fig. S45B**). A drop in coverage (**Fig. S45C**) across region of interest suggests an assembly error. However, we see a similar SV call in the CHM13_HSAv1 and Clint_PTRv1 datasets. In the Clint_PTRv1 assembly, there is even coverage and BES support. Taken together, this evidence suggests the gap in GRCh38 is not fully resolved in the YRI_HSAv1 assembly.

S.2 Genotyping SVs

We used whole-genome Illumina sequencing data from the Great Ape Genome Panel (*33*) and Simons Genome Diversity Project (*100*, *101*) to select a diverse subset of great apes to determine which variants were fixed or at high frequency. We selected individuals from each nonhuman subspecies that were free of contamination and had high proportion if copy 2 regions called correctly (>0.91 for non-orangutan; 5/11 orangutan genomes were lower quality but included to increase diversity). In total, this included 11 orangutans, 8 gorillas, 8 chimpanzees, and 2 bonobos. For humans, we selected two individuals from each SGDP continental group and two additional Africans to better account for the group’s higher diversity. All of these 45 genomes were mapped to the GRCh38, CHM13_HSA, Clint_PTRv1, GSMRT3.2 (Susie gorilla), and Susie_PABv1 (Susie orangutan) assemblies with BWA-MEM and whole-genome shotgun sequence detection (WSSD).

We used these mappings to genotype insertions (with respect to GRCh38) on each primate genome. For all events, we applied SVTyper (commit: 7a349c4f53c84642f57ef01193b1fd00b29b5566; v0.1.2) and for larger read-depth-based events (more than 100 unmasked base pairs) we used WSSD to provide more accurate estimates of copy number. VST and FST statistics were applied to diplotype and genotype data respectively to confirm lineage-specific events. An event was considered lineage specific if it was not seen in any other species and had a VST or FST greater than 0.8.

S.3 Comparative SV Analysis

In total, 614,186 SVs were identified across the great ape lineage, including 295,168 deletions, 316,940 insertions and 2,078 inversions. The two human assemblies had between 24,162 and 26,331 SVs, while the NHPs had between 132,669 and 280,580 SVs (**Table S55**).

We identified 2,034 SVs that are common to our assemblies compared to GRCh38. Fixed SVs are either private/rare variation represented only in GRCh38, systematic errors in our assemblies, or errors in the reference genome. Twenty-two of the GRCh38-specific SVs spanned gaps in the reference genome resulting in SV sizes spanning from 500 bp to 50 kbp (**Fig. S46**). Compared to GRCh38, all five assemblies had a more insertions (GRCh38-specific deletions) than deletions (GRCh38-specific insertions), which was most pronounced in CHM13_HSAv1 (1.6 insertion/deletion). The inflated insertion (GRCh38 deletions) count is also seen across lineages (**Fig. S47**). For insertions, GRCh38 has an abundance of LINEs and SINEs containing insertions relative to the other five assemblies. GRCh38 has many simple repeats, low complexity, LINEs and SINEs compared to the PacBio assemblies. Seven reference-specific events intersected were a CDS coding exon (*FOXO6*, *MUC6*, *MUC19*, *ZNF77*, *PLIN4*, *SAMD1*, and *ZNF676*). The 318 bp insertion in *SAMD1* is located within a simple repeat in the first exon and adds 106 amino acids.

SVs were assigned to the ape phylogeny first by comparing our assemblies (**Fig. S48**) and later by genotyping (**Fig. S49**). Within the human lineage, we identify 17,789 (18.5 Mbp) SVs, confirmed to be lineage specific by genotyping (**Fig. S49**). There were 5,892 (10.3 Mbp) fixed human deletions and 11,987 (8.2 Mbp) insertions. Human-specific structural variation was not evenly distributed across the genome, but clustered into hotspots (**Fig. S50**). Large SVs like the human-specific deletion in the *FADS* gene cluster tend to dominate the signal.

We tested for an overlap correlation between lineage-specific SVs and GRCh38 annotations. Consistent with previous studies (*32*), there was a depletion of structural variation within generic features including, introns, CDS and UTRs (**Fig. S51**), as well as the h3k27ac regulatory mark. We also tested for an enrichment of SVs in genes and regulatory regions that have been associated with human brain development, compared to other primates. No overlap enrichment was detected.

Functional annotation clustering of lineage-specific SVs contained within CDS revealed enrichment for the olfaction term for chimpanzee (Benjamini P = 2.1e-3; enrichment score 2.65; high classification stringency). Both gorilla and orangutan did not have any categories that reached significance; however, olfactory categories were highly ranked. Combining gorilla, chimpanzee, and orangutan genes with an SV-containing CDS resulted in a stronger enrichment for olfactory genes (Benjamini P = 2.4e-4; enrichment score 4.94; high classification stringency). One interesting event is the fusion of the human paralogs *OR6B2* and *OR6B3* in gorilla (a 15 kbp deletion).

S.4 hCONDEL analysis

To enrich for functionally relevant human deletions, we applied the same stringent hCONDEL conservation criterion to our dataset (*5*). A three-way alignment between Clint_PTRv1, macaque and mm10 was generated using Progressive Cactus, and conservation was established in sliding windows of varying sizes (100 bp, 50 bp, and 25 bp with 90%, 90%, and 92% identity, respectively). Instead of intersecting conserved regions with fhDELs, we used deletions that were genotyped as fixed in human (gfDELs) because the original hCONDEL dataset did not require conservation in gorilla and orangutan. By removing the constraint that fhDELs must have conservation with all great apes, our fhDELs expanded from 5,892 to 7,400 gfDELs. Then, applying conservation within macaque and mouse shrinks the set to 930 gfDELs that qualify as hCONDELs, of which 694 are novel. This set can be further constrained to 795 by requiring the ancestral region to be fixed in other great ape assemblies, of which 226 are novel compared to the previous hCONDEL dataset (**Fig. S52**).

We confirmed 451/583 (77%) of the total hCONDELs with a liftover tolerance of 3 kbp and by mapping the PanTro2 hCONDELs to Clint_PTRv1 (**Fig. 3f**) (**Table S11**) (90% of the confirmed hCONDELs mapped within ±25 bp of a gfDEL). The hCONDELs that overlapped the gfDELs spanned from >25 kbp down to 1 kbp with an average size of 4.2 kbp. Of the 434 stringent hCONDELs, limited to fixed and verified events, we recovered 89% (386). The size distribution and repeat content of the gfDEL/hCONDEL overlap suggests that LINE-1 elements are overrepresented in the hCONDEL dataset. Indeed, hCONDELs were more repetitive (40%) when compared to the gfDELs we identified.

Investigation into the 23% (132 gfDELs) of the hCONDELs missing from our analysis revealed that the majority could be accounted for as either polymorphic in human or containing more complex structure variation in human (156 were missing if we considered the smaller fhDEL set; **Table S11**). Specifically, there were 72 in one of the human assemblies (YRI or CHM13): 30 were polymorphic in the human population (based on genotyping), another 27 represented a different type of SV or were more complex in organization, 22 were either not assembled in our genomes or represent false negatives in our dataset, and one flanked a gap in the reference genome (**Table S11**). The rest we consider to be false negatives in our dataset. Interestingly, the previously reported hCONDEL affecting an androgen receptor enhancer also shows independent structural changes in gorilla, not correctly assembled/scaffolded in the short-read gorilla (GorGor) genome. Comparing the long-read gorilla assembly to hg38 revealed that the hCONDEL sequence in gorilla involves a complex SV, including an inversion, that preserves the ancestral penile spine and vibrissae enhancer sequence (*35*) but may independently influence AR expression in the gorilla lineage (**Fig. 3g**).

S.5 Sequence resolution of large-scale inversions

In order to identify copy number neutral SVs, such as inversions, we initially implemented three orthogonal approaches for inversion discovery: 1) genome maps generated from Bionano alignments were manually inspected for the presence of inversion breakpoints; 2) chromosomes (composed of ordered and oriented scaffolds) aligned to the human GRCh38 reference assembly were compared using MUMmer; and 3) BES were aligned against the GRCh38 reference assembly. From this manual call set, we identified a total of 29 human-chimpanzee-orangutan inversions, in total validating 62% (18/29) of detected inversions using a combination of FISH, contig assembly and BAC clone-based breakpoint sequencing. In addition, we were also able to confirm a further four previously reported large events that were validated by FISH (**Table S12**). This included a 4.6 Mbp inversion on chromosome 4p16 containing >50 genes that is polymorphic in both human and chimpanzee based on FISH analysis of a limited number of individuals (**Fig. S53**). In order to assess inversions on a more comprehensive scale, we implemented an automated hierarchical clustering approach (see **Section E.1 methods**) based on genome maps generated from Bionano Genomics. We identified a total of 625 (341 chimpanzee and 284 orangutan) human-chimpanzee-orangutan inversions ranging from 9 kbp to 8.4 Mbp in size (**Table S12**). This new set of nonhuman ape inversions will provide the necessary framework for identifying large-scale variation that is still inaccessible to WSSD and assembly.

From our list of human-chimpanzee-orangutan inversions, we selected three regions in chimpanzee for high-quality sequence and assembly using large-insert clones. On chromosome 13q14.3 (chr13:52014388-52688364), we selected five clones (CH251 BAC library) for high-quality sequencing based on BES mapping placements to the GRCh38 reference assembly (**Table S15**). Using a clone-based hierarchical approach, we assembled a ~580 kbp alternate chimpanzee reference haplotype of this region. Comparison with the human reference revealed a ~294 kbp inversion that inverts five genes relative to human. Comparisons between human and mouse revealed that the chimpanzee orientation is the likely ancestral state. In addition to the inversion, the human GRCh38 haplotype differed from the chimpanzee by the presence of an expanded inverted duplication block consisting of three duplications totaling 158.7 kbp. The chimpanzee is in fact missing the large 100.7 kbp centromeric inverted duplication, which likely rendered this region susceptible to inversion among humans (**Fig. 5a**). Next, we sequenced the breakpoints of a large ~2.7 Mbp inversion on chromosome 2q11.2 (chr2:99548000-102250000) containing 17 genes. In order to characterize the inversion breakpoints in more detail, we sequenced 19 BAC clones (**Table S15**), assembling four sequence contigs consisting of 1.65 Mbp of high-quality finished sequence. Comparisons between human and mouse revealed that the chimpanzee orientation is likely the derived state. Sequence analysis revealed that this region in the chimpanzee differed structurally from the human reference by the presence of a large inversion (2.7 Mbp), lineage-specific expansions of the Interlukin gene cluster, a ~100 kbp inverted duplication of *REV1*, an interchromosomal duplication mapping to chromosome 4, and an intrachromosomal duplicative transposition of the RGPD4 core duplicon (**Fig. 5b**). Despite generating a large number of clone inserts mapping to this region (in addition to PacBio WGS sequencing data), the large, complex nature of the duplication blocks flanking the inversion meant that we could not completely resolve this region in the chimpanzee. It is noteworthy, however, that we detected an additional 2.3 Mbp inversion at 2q12-q13 in the chimpanzee that also contains *RGPD4* at the boundaries of the inversion event. This finding is consistent with previous work demonstrating that core duplicons are associated with genomic instability (*103*). Finally, we sequenced and assembled an additional nine BAC clones (**Table S15**) mapping to chromosome 13q14.13, generating a large ~1.2 Mbp alternate reference haplotype in chimpanzee. Sequence analysis revealed the presence of a 1.1 Mbp inversion, including lineage-specific duplications at the boundaries of the event (**Fig. 5c**). Notably, this is one of three large inversions we characterized in apes across chromosome 13 (10.3% of ape inversions), a chromosome which to date has been shown to lack common inversion polymorphisms among humans (*49*).

VIII. Data deposition and accessions

The underlying PacBio sequence data, Illumina sequencing, assembled contigs and assemblies for each of the ape species have been deposited in NCBI under the project accessions PRJNA369439 (chimpanzee, orangutan, CHM13, NA19240), PRJEB10880 (gorilla) (**Table S56**). Clone sequences have been deposited in GenBank under umbrella BioProject ID PRJNA369439 (**Table S15**). Transcriptional data was deposited in NCBI (**Table S57**). The SVs were deposited in DGVa with the accession estd235. The genome assemblies have different names and aliases depending on the institution hosting these genomes (**Table S58**).

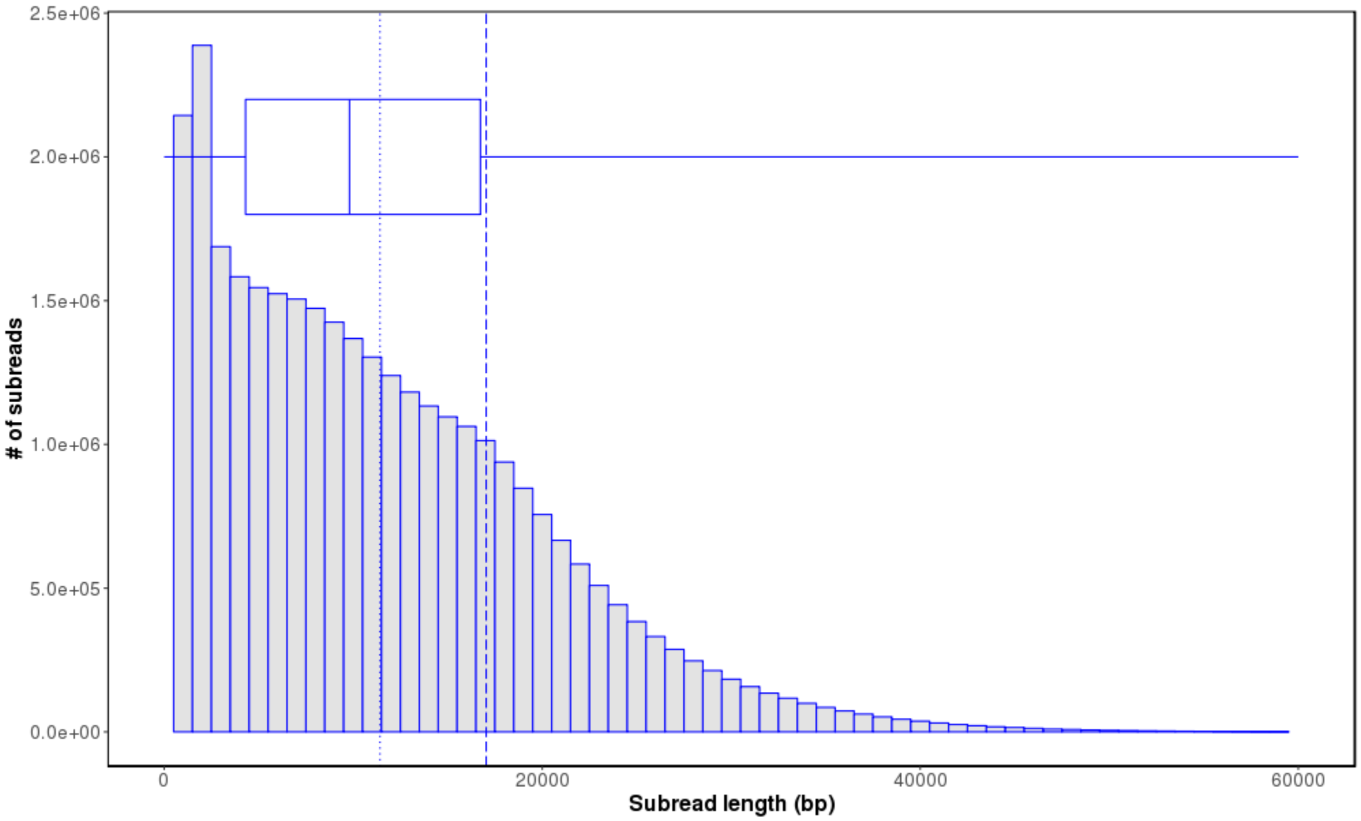


Fig. S1.

Distribution of subread lengths for chimpanzee sequencing data.
Marginal boxplot indicates quartiles. Mean subread length is 11.4 kbp (vertical dotted) and N50 subread length is 17.0 kbp (vertical dashed).

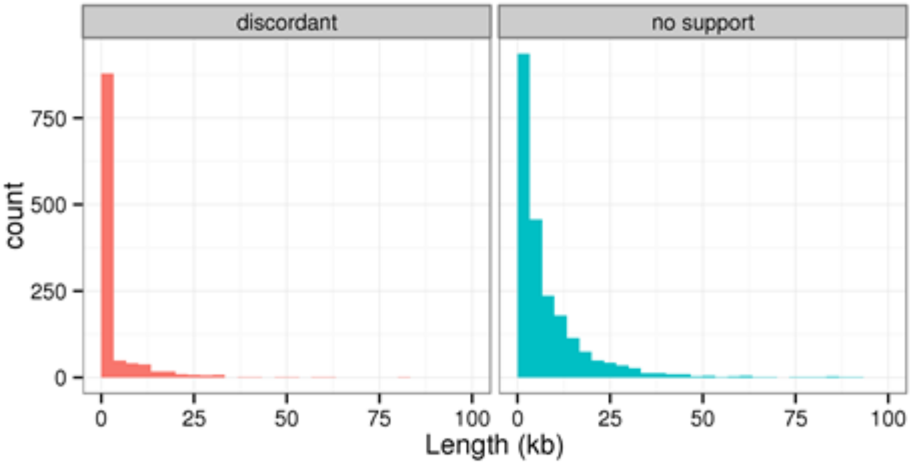


Fig. S2.
Size distribution of discordant regions identified in Clint_PTRv1.
(left) Regions flagged as discordant by virtue of abnormal BES length and/or orientation, or multiple mapping locations. (right) Regions flagged as discordant by virtue of lack of BES/FES support.

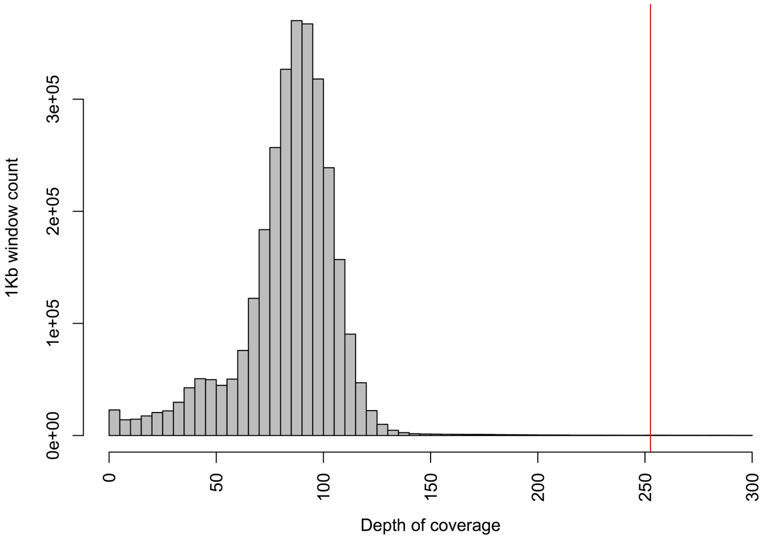


Fig. S3.

PacBio depth of coverage for Clint_PTRv1.

Depth of coverage in the assembly in 1 kbp windows (depth greater than 300 not shown). The red line denotes the high-depth cutoff at 252.6.

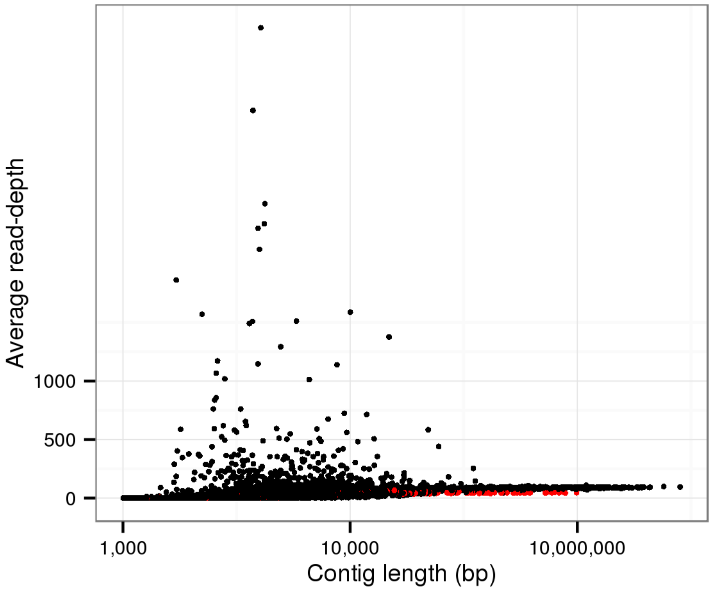


Fig. S4.

Length of contigs versus average PacBio depth of coverage in 1 kbp windows. Contigs mapping to GRCh38 chrX are colored red.

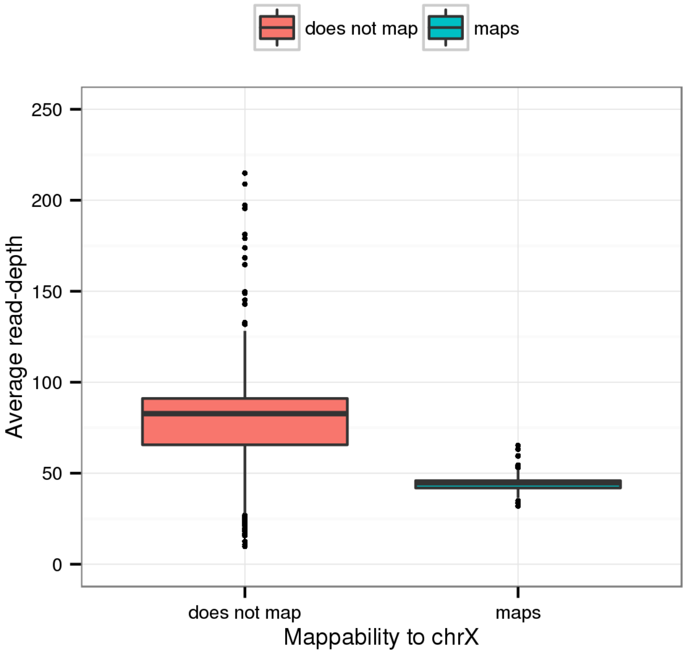


Fig. S5.

Distribution of PacBio depth of coverage across GRCh38 autosomes and chromosome X.
As expected, the diploid genome has double coverage (orange) compared to chromosome X (green) since Clint is a male.

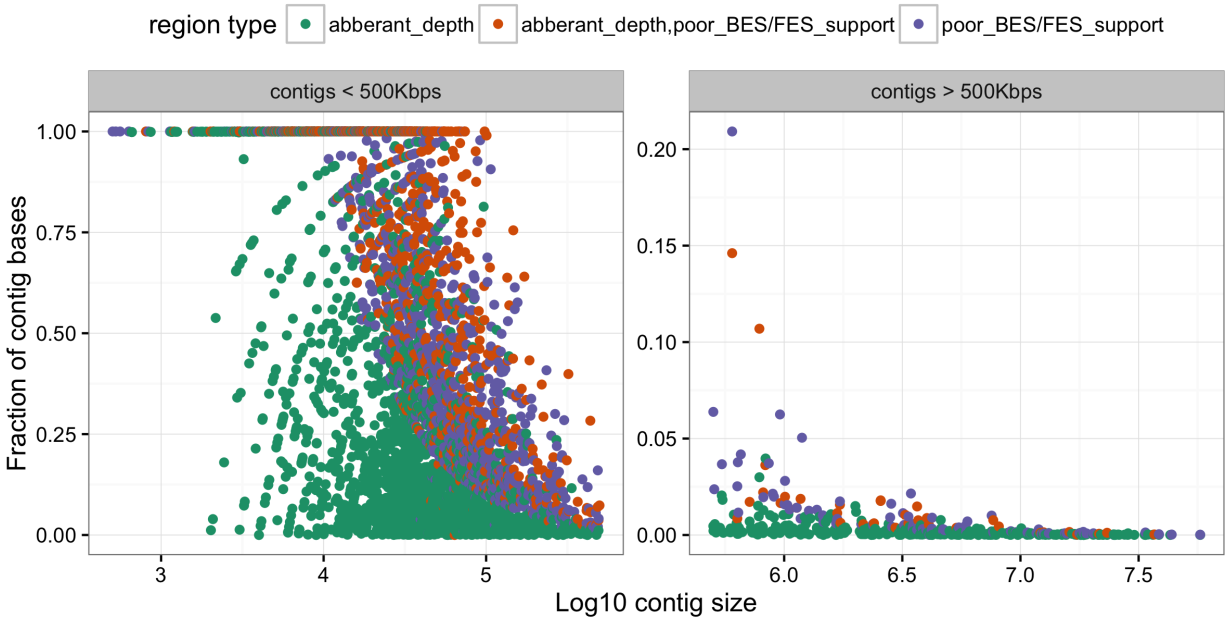


Fig. S6.
Discordant BES/FES and/or abnormal read depth across Clint_PTRv1 contigs.

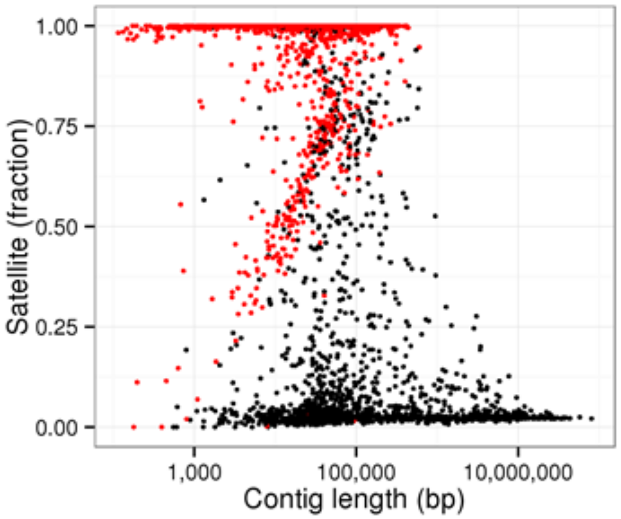


Fig. S7.
Mappability to human (GRCh38) and satellite content of Clint_PTRv1 contigs.
Satellite content defined by use of RepeatMasker and Tandem Repeats Finder. Contigs <100 bp were excluded from the analysis. There were 2,228 unmapped contigs overall. Contigs that do not map to human are colored in red.

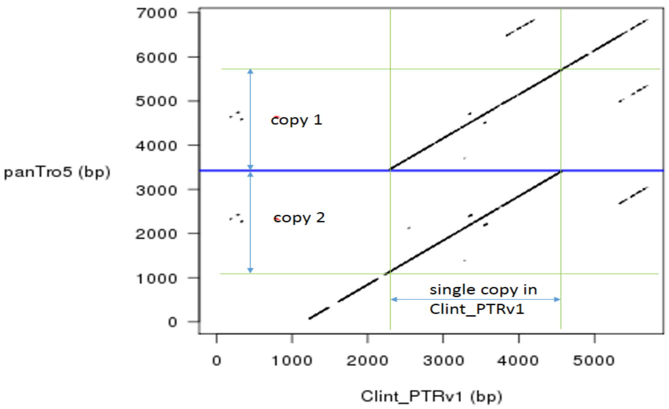


Fig. S8.

Example of duplicated sequence in panTro5.
The size of the panTro5 sequence is ~2,500 bp larger than that of the aligned Clint_PTRv1 sequence due to the duplicated sequence indicated. Even though there are 25 N's in panTro3 (the blue horizontal line), the Clint_PTRv1 assembly shows that there is in fact no missing sequence at all but rather ~2,500 bp is duplicated in panTro5. panTro5's CM000316.3:66521488-66528346 is aligned against Clint_PTRv1's 000001F_1_57587592_quiver_pilon:44071726-44078584.

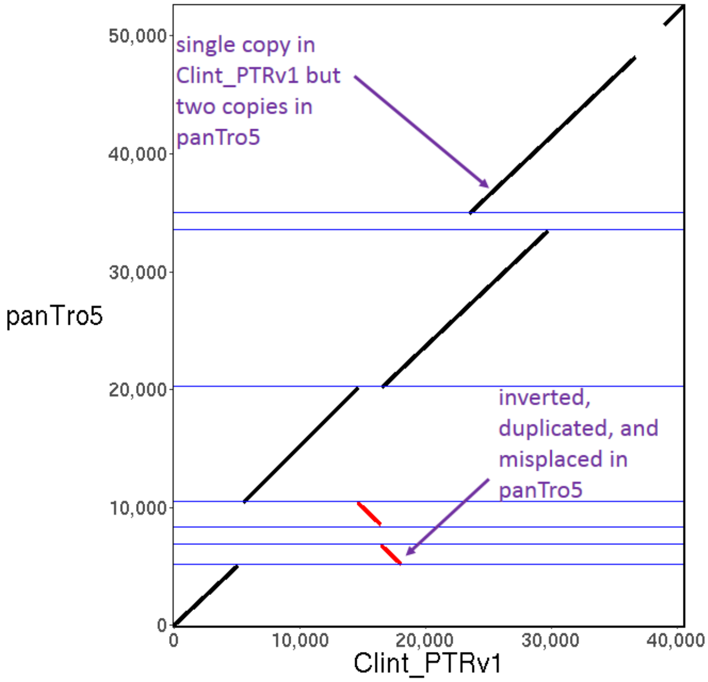


Fig. S9.

Example of misplaced and duplicated sequence in panTro5.
The size of the panTro5 sequence is 11,583 bp larger (not including N's) than the aligned Clint_PTRv1 sequence due to the duplicated sequence (indicated) and due to the duplicated and inverted sequences (red lines, indicated). This region contains 475 N's in seven gaps (blue horizontal lines). panTro5 CM000332.3:28173-80754 is aligned using MUMmer against Clint_PTRv1 001655F_1_57369:11230-51753.

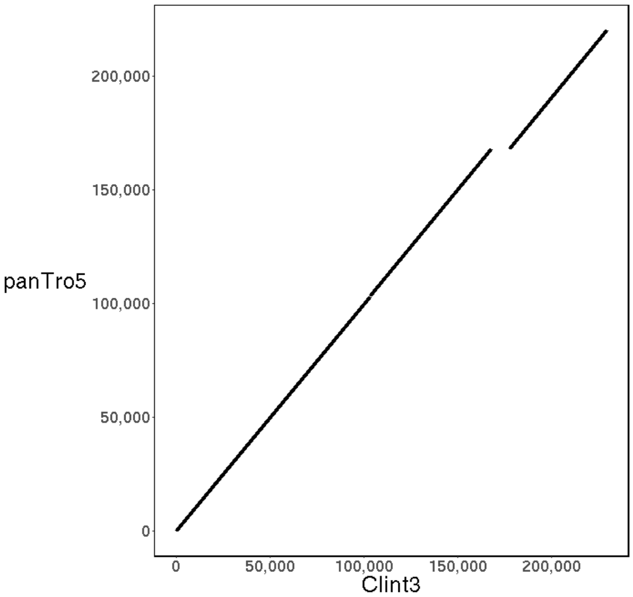


Fig. S10.

Clint_PTRv1 adds bases to panTro5.

The additional sequence is the size of the displacement between the two black lines: 9,298 bp. The panTro5 assembly only had 25 N's at this location. panTro5's CM000314.3 position 5087202-5307586 is aligned by MUMmer against Clint_PTRv1's contig 000193F_1_4106580 position 3878731-4108388.

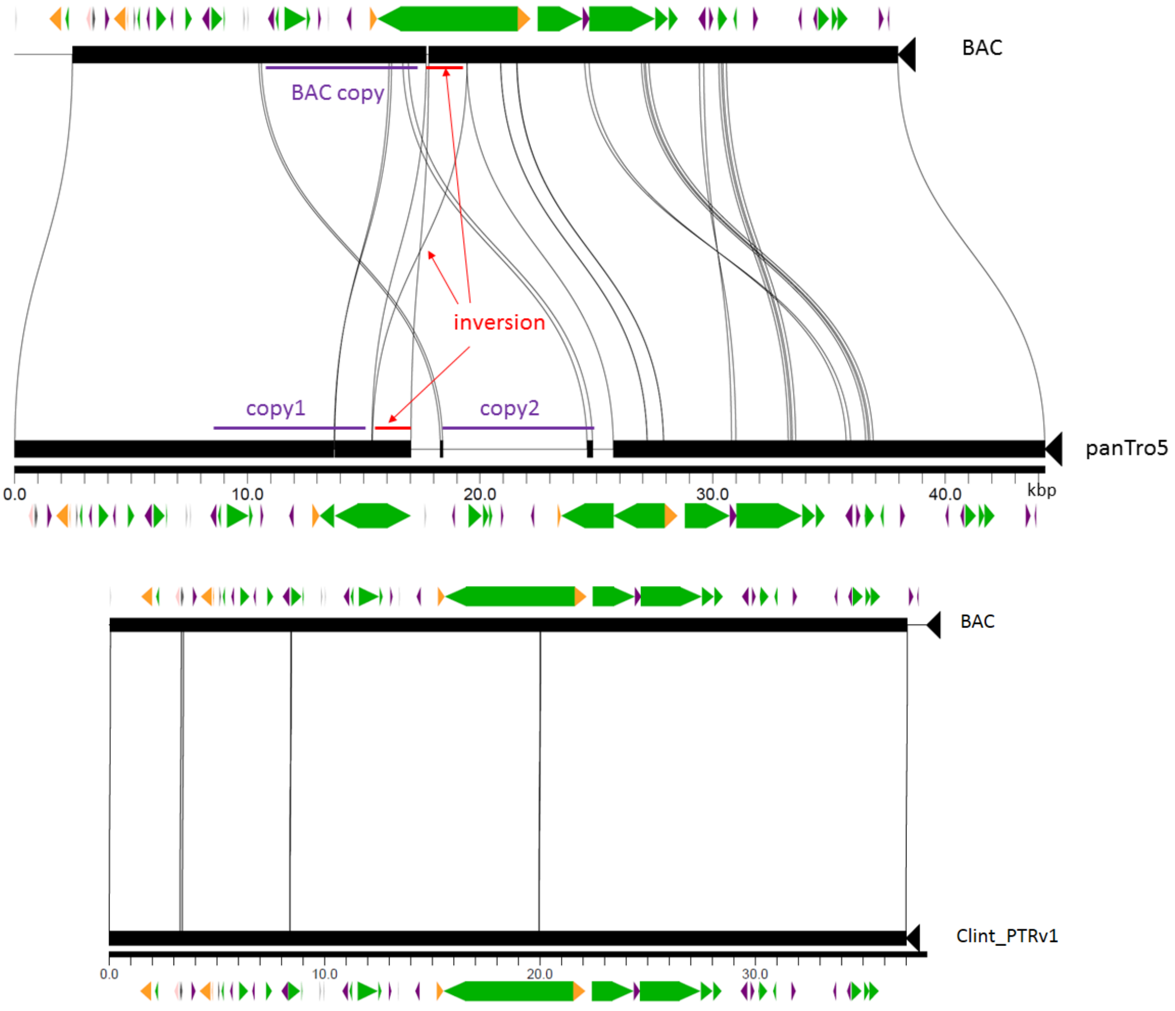


Fig. S11.

Miropeats of a BAC aligned against panTro5 and Clint_PTRv1.
BAC CH251-72H6 was PacBio sequenced to ~1,770X coverage. The top panel, BAC CH251-72H6:150000-187962 aligned against panTro5 CM000329.3:68020000-68064272, shows two types of common problems in panTro5 that are fixed by Clint_PTRv1: inverted sequence and extraneous sequence. The inversion is flanked by N's on both sides suggesting it was a contig that was inserted into the chromosome in the wrong orientation. The purple bar "copy2" is a ~8 kbp insertion in panTro5 that (according to a GenBank blast search against nr/nt) best matches to this very same BAC at location "BAC copy" with 89.6% identity. The purple bar "copy1" (the correct sequence) matches "BAC copy" at 99.6% identity. This suggests that "copy2" is artifactually duplicated sequence. Green arrows indicate LINES, purple SINES, and yellow LTRs. The bottom panel shows the same BAC region aligned very closely against Clint_PTRv1 000109F_1_7982668_quiver_pilon:2067235-2104235. According to cross_match, this alignment is 99.9% identical.

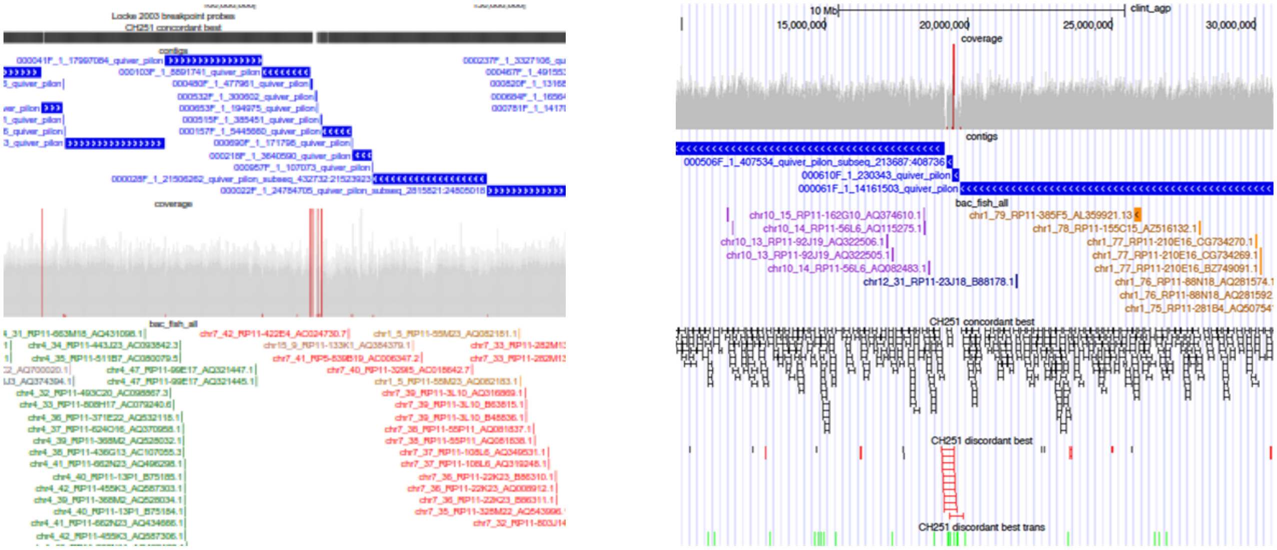

Locke 2003 breakpoint probes
CH251 concordant best
contigs
000041F_1_17997084_quiver_pilon
000103F_1_8891741_quiver_pilon
000480F_1_477961_quiver_pilon
000532F_1_300602_quiver_pilon
000653F_1_194975_quiver_pilon
000515F_1_385451_quiver_pilon
000157F_1_5445660_quiver_pilon
000690F_1_171798_quiver_pilon
000218F_1_3640590_quiver_pilon
000957F_1_107073_quiver_pilon
000028F_1_21506262_quiver_pilon_subseq_432732:21523923
000022F_1_24784705_quiver_pilon_subseq_2815821:24805018
000237F_1_3327106_qu
000467F_1_49155
000820F_1_13168
000684F_1_16564
000781F_1_14171
coverage
bac_fish_all
chr4_34_RP11-443J23_AC093842.3
chr4_35_RP11-511B7_AC080079.5
chr4_47_RP11-99E17_AQ321447.1
chr4_47_RP11-99E17_AQ321445.1
chr4_32_RP11-493C20_AC098867.3
chr4_33_RP11-808H17_AC079240.6
chr4_36_RP11-371E22_AQ532118.1
chr4_37_RP11-624O16_AQ370958.1
chr4_39_RP11-368M2_AQ528032.1
chr4_38_RP11-436G13_AC107055.3
chr4_41_RP11-662N23_AQ496298.1
chr4_40_RP11-13P1_B75185.1
chr4_42_RP11-455K3_AQ587303.1
chr4_39_RP11-368M2_AQ528034.1
chr4_40_RP11-13P1_B75184.1
chr4_41_RP11-662N23_AQ434666.1
chr4_42_RP11-455K3_AQ587306.1
chr7_42_RP11-422E4_AC024730.7
chr15_9_RP11-133K1_AQ384379.1
chr7_41_RP5-839B19_AC006347.2
chr7_40_RP11-32I5_AC018642.7
chr1_5_RP11-55M23_AQ082181.1
chr1_5_RP11-55M23_AQ082183.1
chr7_33_RP11-282M1
chr7_39_RP11-3L10_AQ316869.1
chr7_39_RP11-3L10_B63815.1
chr7_39_RP11-3L10_B48836.1
chr7_38_RP11-55P11_AQ081837.1
chr7_38_RP11-55P11_AQ081836.1
chr7_37_RP11-108L6_AQ349531.1
chr7_37_RP11-108L6_AQ319248.1
chr7_36_RP11-22K23_B86310.1
chr7_36_RP11-22K23_AQ008912.1
chr7_36_RP11-22K23_B86311.1
chr7_35_RP11-328M22_AQ543996
chr7_32_RP11-803J14
10 Mb
clint_agp
15,000,000
20,000,000
25,000,000
30,000,000
coverage
contigs
000506F_1_407534_quiver_pilon_subseq_213687:408736
000610F_1_230343_quiver_pilon
000061F_1_14161503_quiver_pilon
bac_fish_all
chr10_15_RP11-162G10_AQ374610.1
chr10_14_RP11-56L6_AQ115275.1
chr10_13_RP11-92J19_AQ322506.1
chr10_13_RP11-92J19_AQ322505.1
chr10_14_RP11-56L6_AQ082483.1
chr12_31_RP11-23J18_B88178.1
chr1_79_RP11-385F5_AL359921.13
chr1_78_RP11-155C15_AZ516132.1
chr1_77_RP11-210E16_CG734270.1
chr1_77_RP11-210E16_CG734269.1
chr1_77_RP11-210E16_BZ749091.1
chr1_76_RP11-88N18_AQ281574.1
chr1_76_RP11-88N18_AQ281592
chr1_75_RP11-281B4_AQ50754
CH251 concordant best
CH251 discordant best
CH251 discordant best trans

Fig. S12.

Browser images of the two chimeric Bionano scaffolds
(left) Shown is a scaffold (black), contigs (blue) and the BACs mapped to it, colored based on chromosome. Based on the color of the BACs (green and red), there are two different chromosome parts fused together. (right) Similarly, based on the color of BACs (purple and brown-orange), there are two different chromosome pieces fused together.

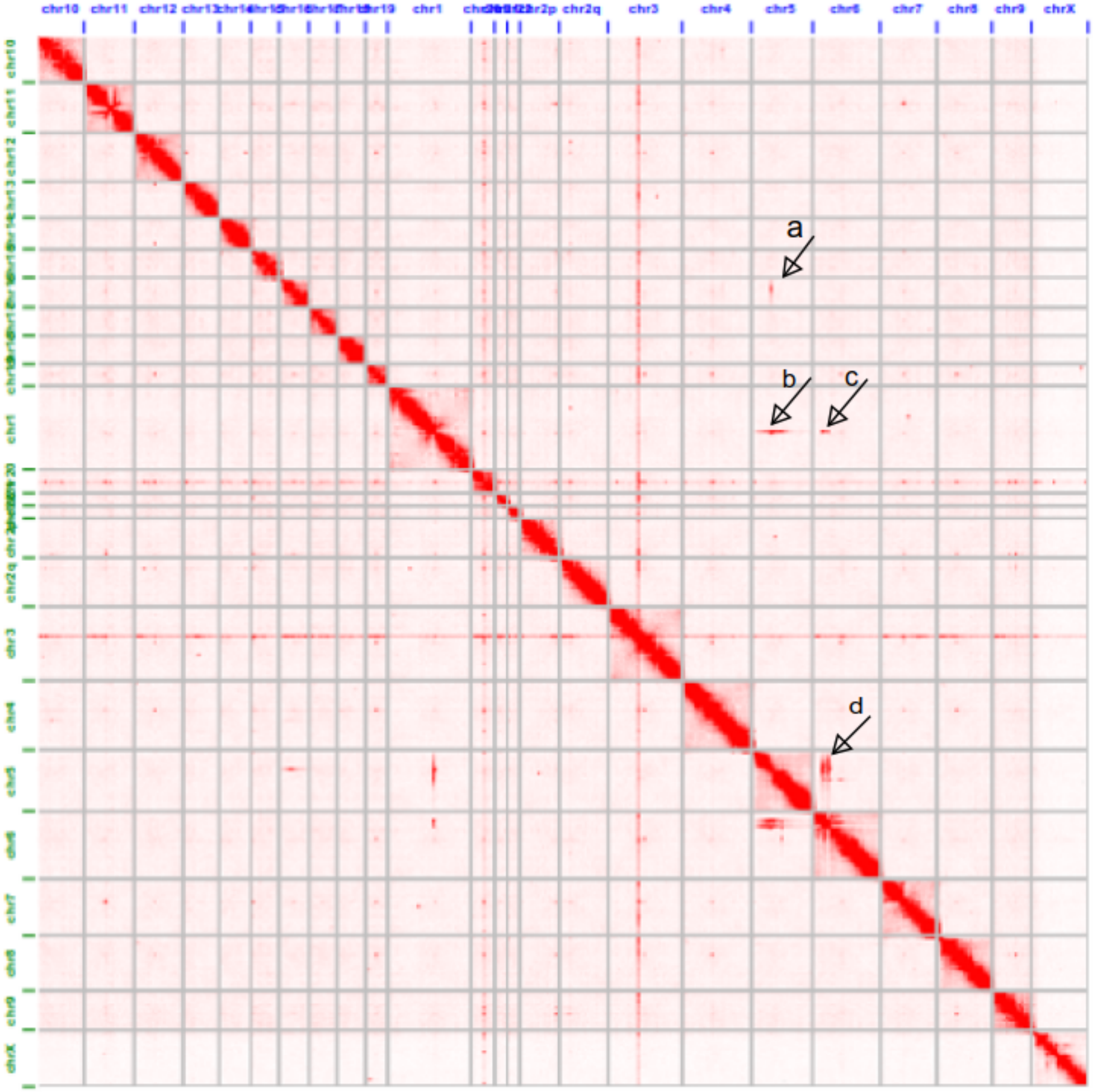
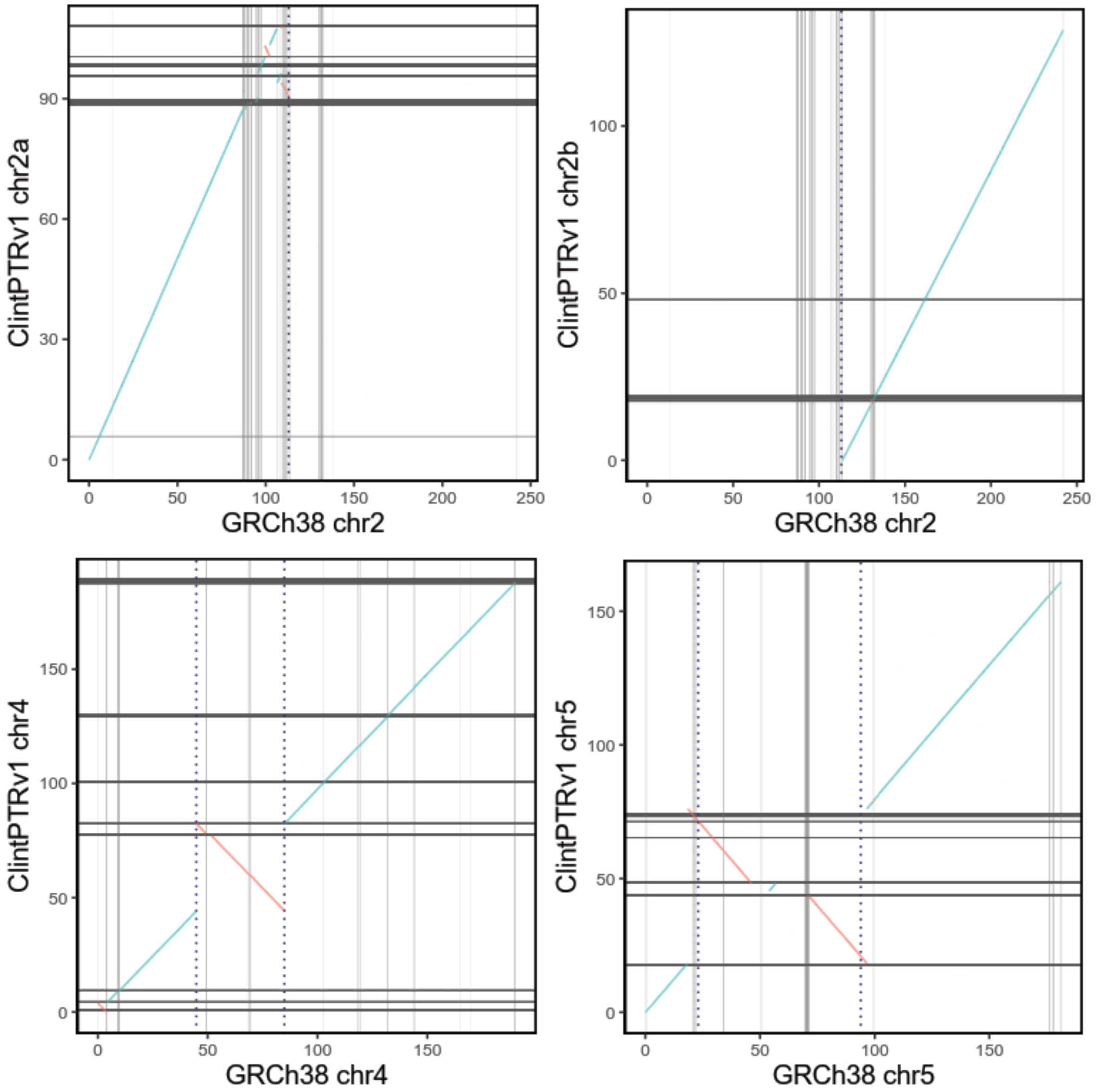
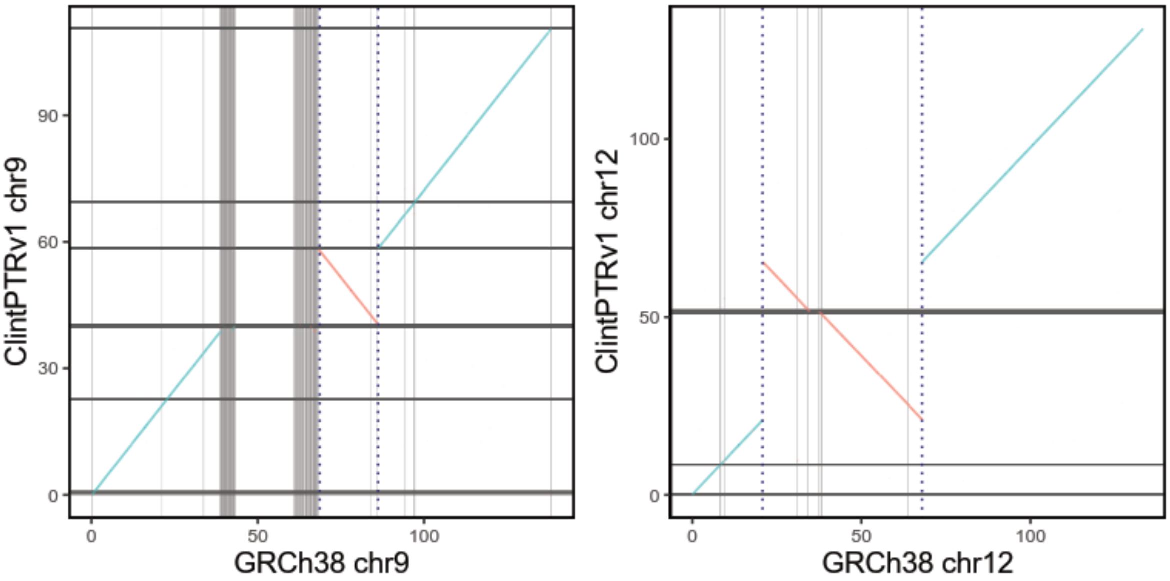


Fig. S13.

All-vs-all chromosome heatmap of the Hi-C data aligned to the Clint_PTRv1 chromosomes.
‘a’, ‘b’, ‘c’ and ‘d’ arrows indicate chimeric scaffolds where a contig was placed into the wrong scaffold, on the wrong chromosome.


ClintPTRv1 chr2a
0
30
60
90
0
50
100
150
200
250
GRCh38 chr2
ClintPTRv1 chr2b
0
50
100
0
50
100
150
200
250
GRCh38 chr2
ClintPTRv1 chr4
0
50
100
150
0
50
100
150
GRCh38 chr4
ClintPTRv1 chr5
0
50
100
150
0
50
100
150
GRCh38 chr5


ClintPTRv1 chr9
90
60
30
0
0
50
100
GRCh38 chr9
ClintPTRv1 chr12
100
50
0
0
50
100
GRCh38 chr12

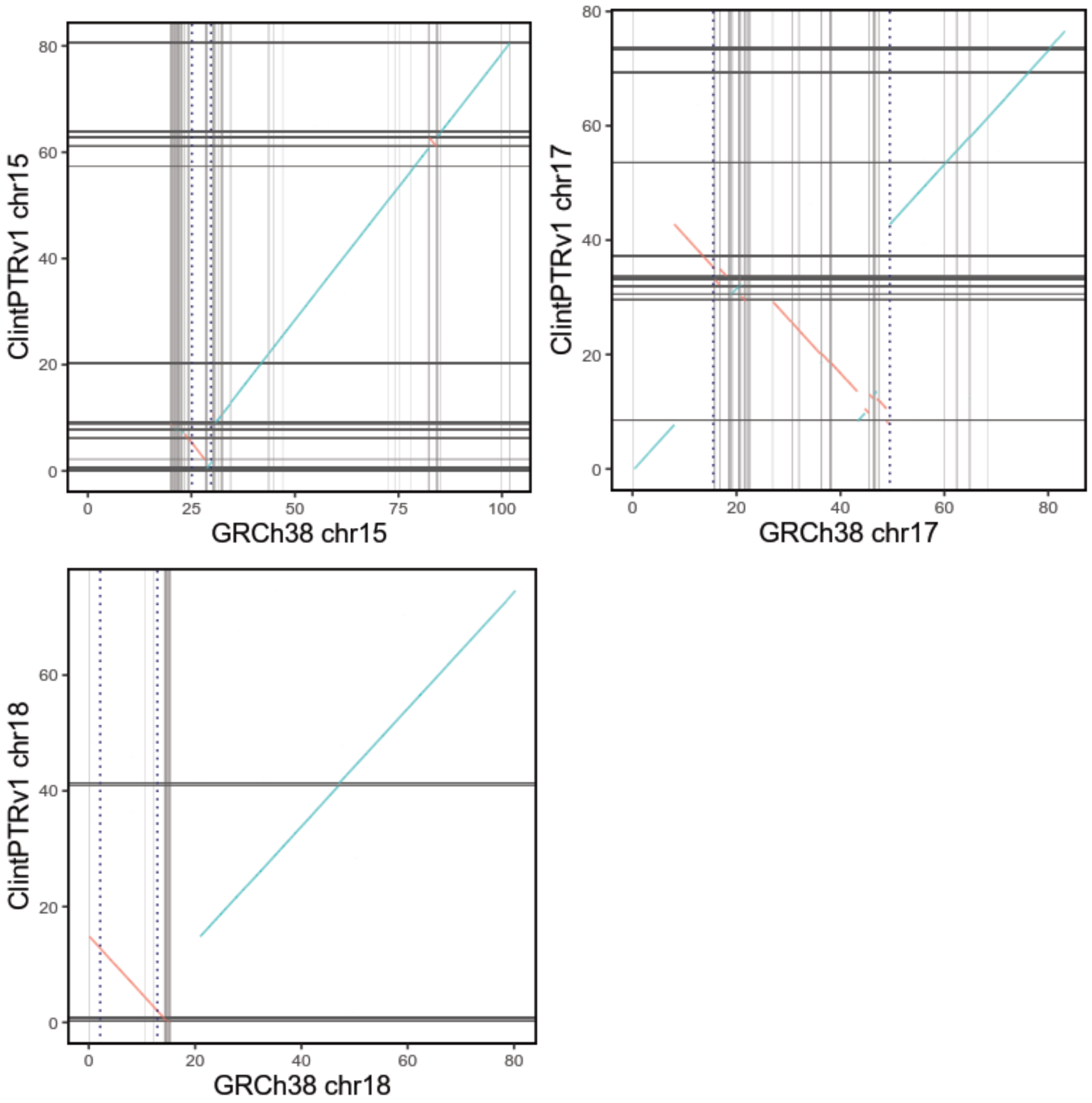
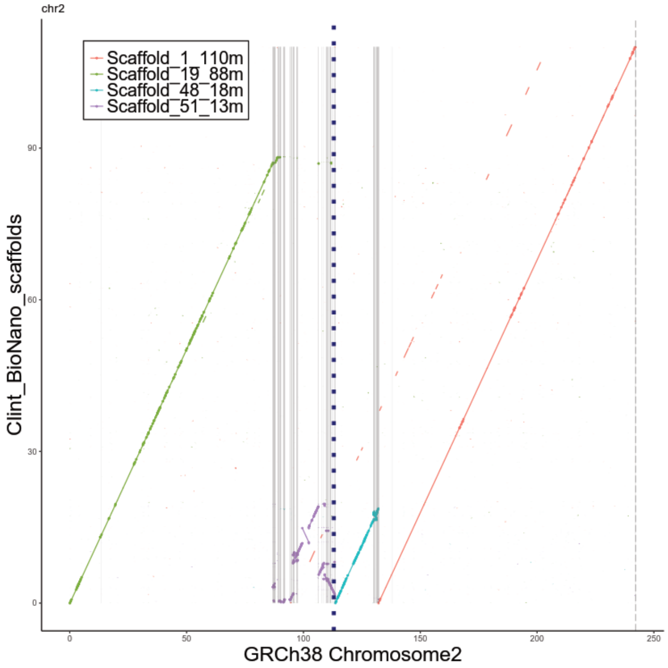
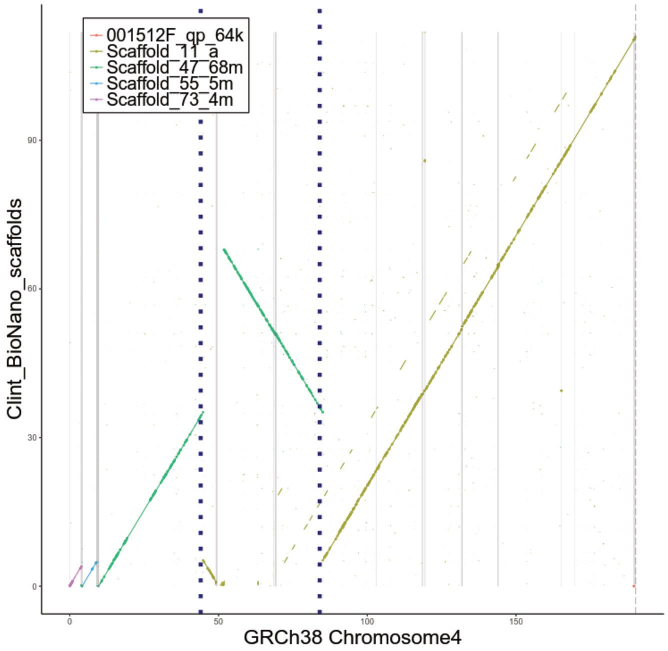
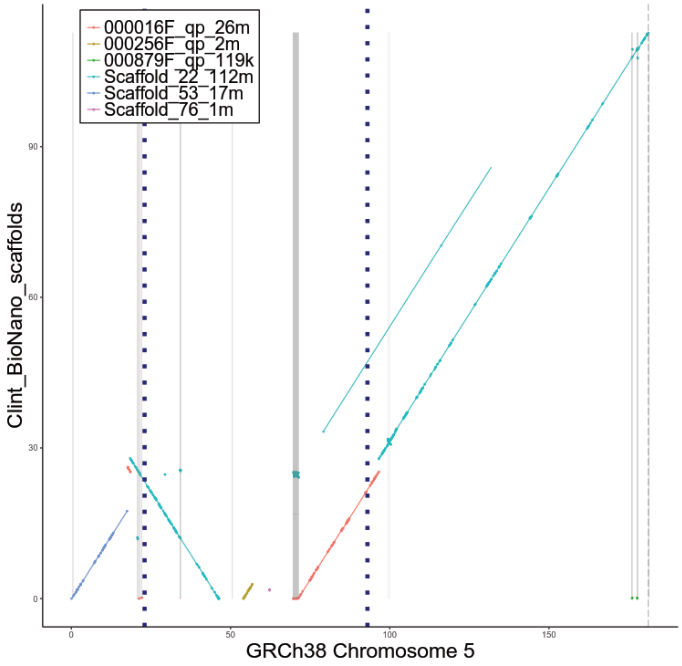


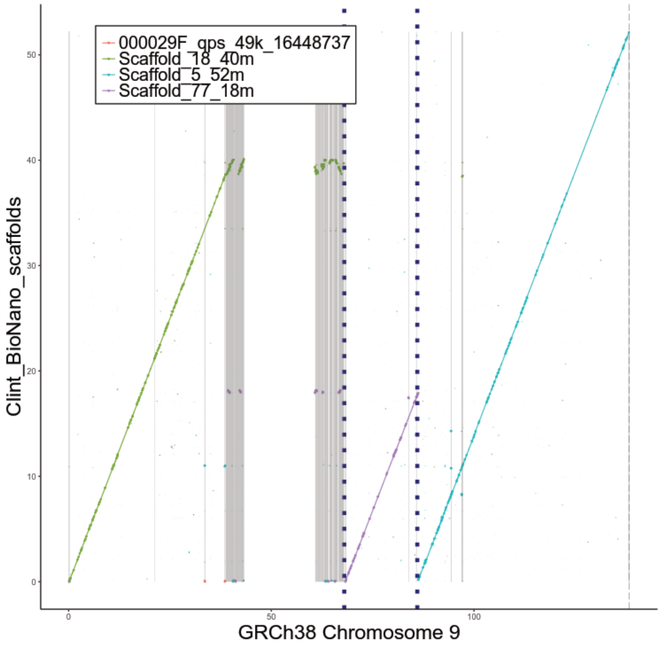
Fig. S14.

Dot plots showing all captured Yunis-Prakash inversions and chromosomal fusions.
The human reference genome is on the x-axis and chimpanzee is on the y-axis. Blue dashed lines indicate the boundaries of the inversion/fusion, while gray solid lines indicate regions containing SDs (segdup regions for Clint_PTRv1 (horizontal solid gray lies) here are indicated by regions of high read depth).


chr2
Scaffold_1_110m
Scaffold_19_88m
Scaffold_48_18m
Scaffold_51_13m
Clint_BioNano_scaffolds
0
30
60
90
0
50
100
150
200
250
GRCh38 Chromosome2


001512F_qp_64k
Scaffold_11_a
Scaffold_47_68m
Scaffold_55_5m
Scaffold_73_4m
Clint_BioNano_scaffolds
0
30
60
90
0
50
100
150
GRCh38 Chromosome4


000016F_qp_26m
000256F_qp_2m
000879F_qp_119k
Scaffold_22_112m
Scaffold_53_17m
Scaffold_76_1m
Clint_BioNano_scaffolds
0
30
60
90
0
50
100
150
GRCh38 Chromosome 5


000029F_qps_49k_16448737
Scaffold_18_40m
Scaffold_5_52m
Scaffold_77_18m
Clint_BioNano_scaffolds
0
10
20
30
40
50
0
50
100
GRCh38 Chromosome 9

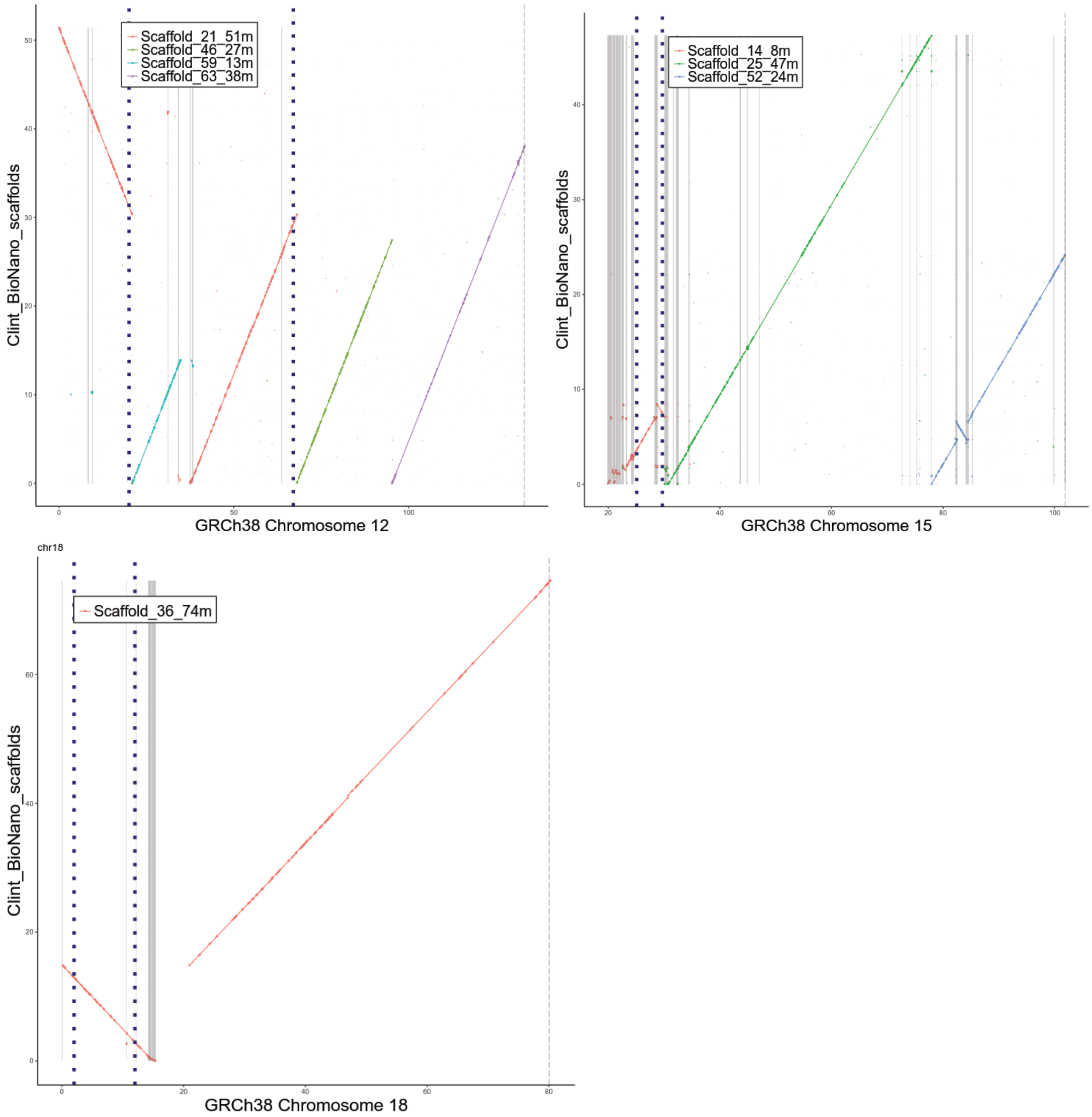


Fig. S15.

Boundaries of known inversions captured within Bionano scaffolds.
Scaffolds are separately mapped to human chromosomes to identify if the boundaries of known inversions (**Table S26**) have been captured within the scaffold.

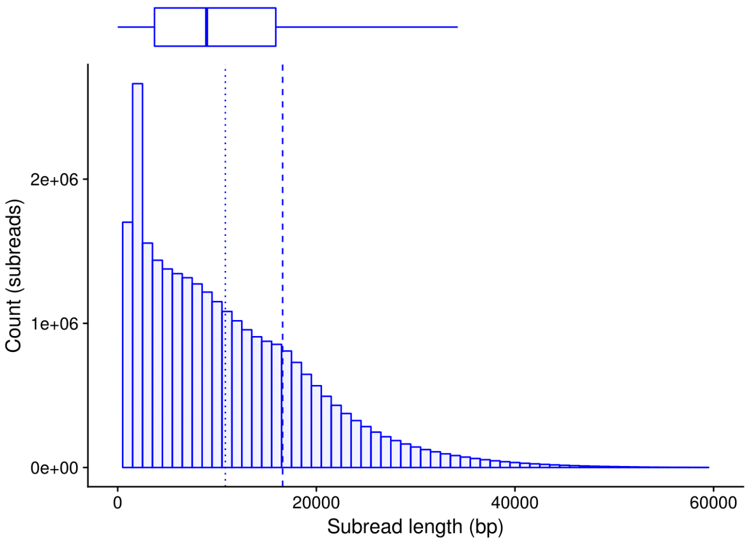


Fig. S16.

Distribution of subread lengths for orangutan sequencing data.
Marginal boxplot indicates quartiles with an average subread length of 10.8 kbp (vertical dotted) and an N50 subread length of 16.6 kbp (vertical dashed).

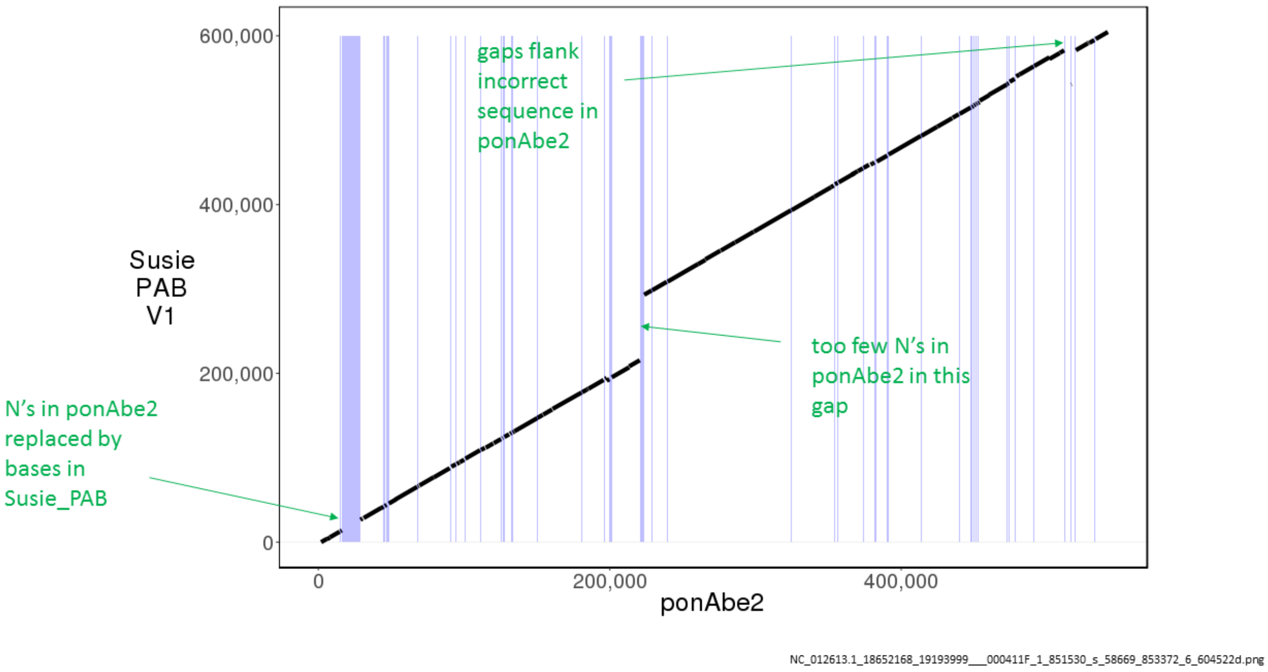


Fig. S17.

Alignment of ponAbe2 to Susie_PABv1.
Light blue vertical lines are the locations of N's in ponAbe2. Notice that Susie_PABv1 has more sequence than ponAbe2 in this alignment, which we found to be true in most alignments. ponAbe2 NC_012613.1:18652168-19193999 is aligned against Susie_PABv1 000411F_1_851530_s_58669_853372:6-604522.

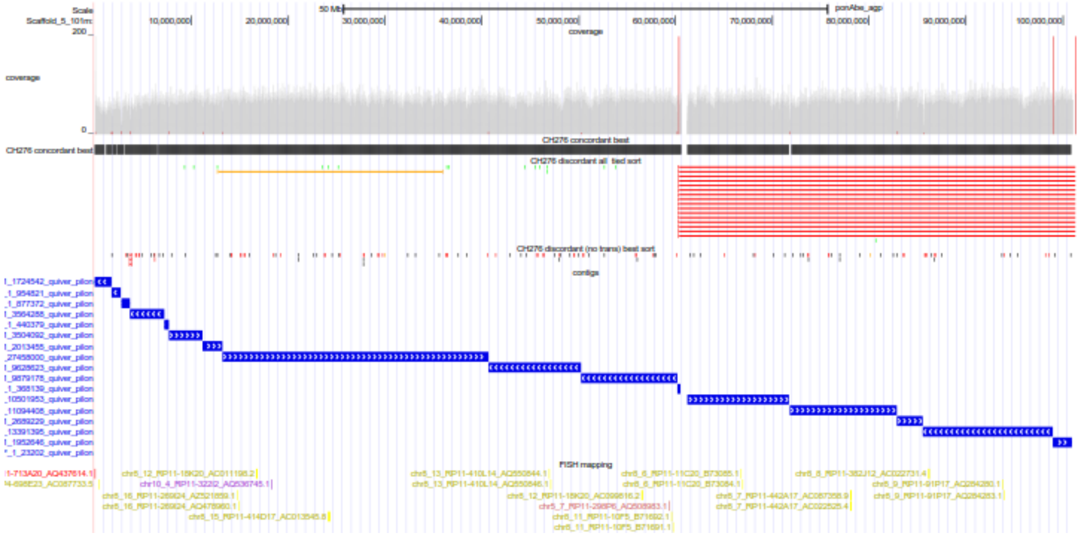


Fig. S18.

Example of a scaffolding error identified by BES mapping in Susie_PABv1.
A 101 Mbp scaffold with an incorrect inversion had to be manually split and rejoined. Scaffold_5_101m clearly showed an incorrectly oriented part, detected by FISH and BAC ends.

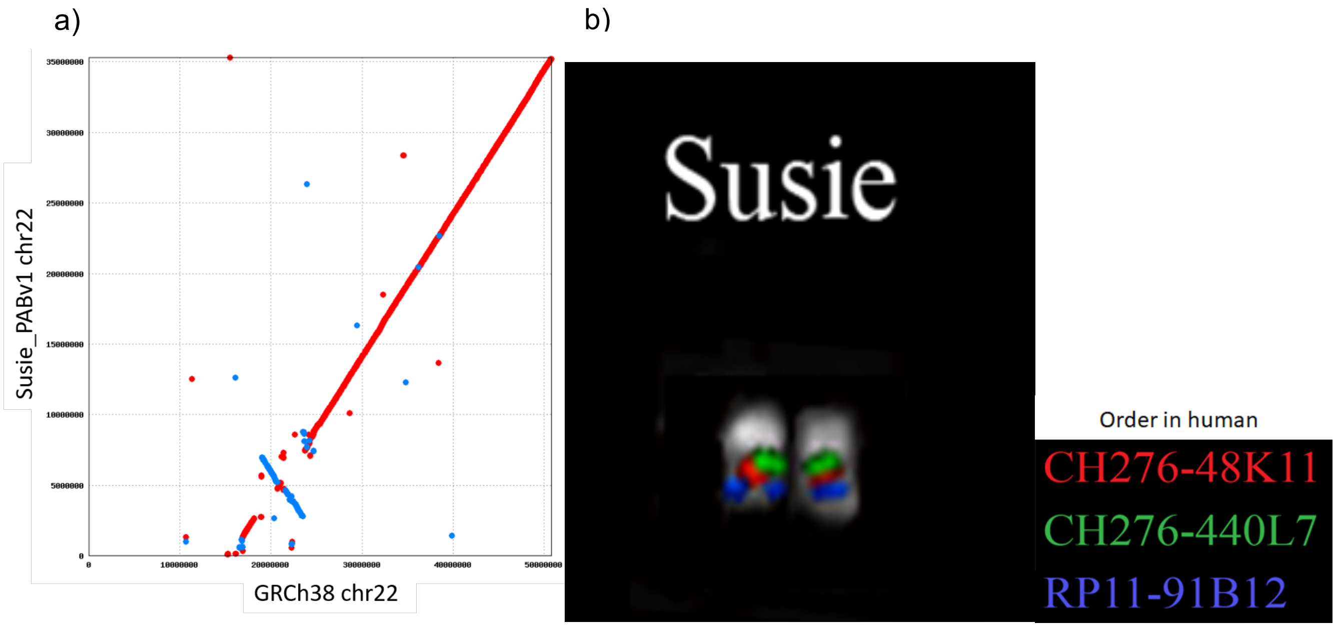


Fig. S19.

An 8 Mbp inversion confirmed by bicolor FISH in orangutan chr22.
(a) When aligned to human chr22, we detected a large ~8 Mbp inversion in Susie_PABv1 (b) Bicolor FISH image confirming the orientation of the scaffold. Right panel shows the order of the probes in human (red-green-blue). The blue probe is a human clone use as anchor outside the inversion.

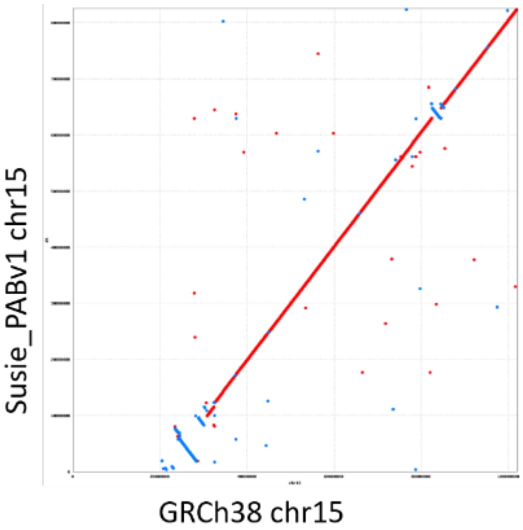


Fig. S20.
A complex inversion detected in chr15 of Susie_PABv1 confirmed by bicolor FISH.

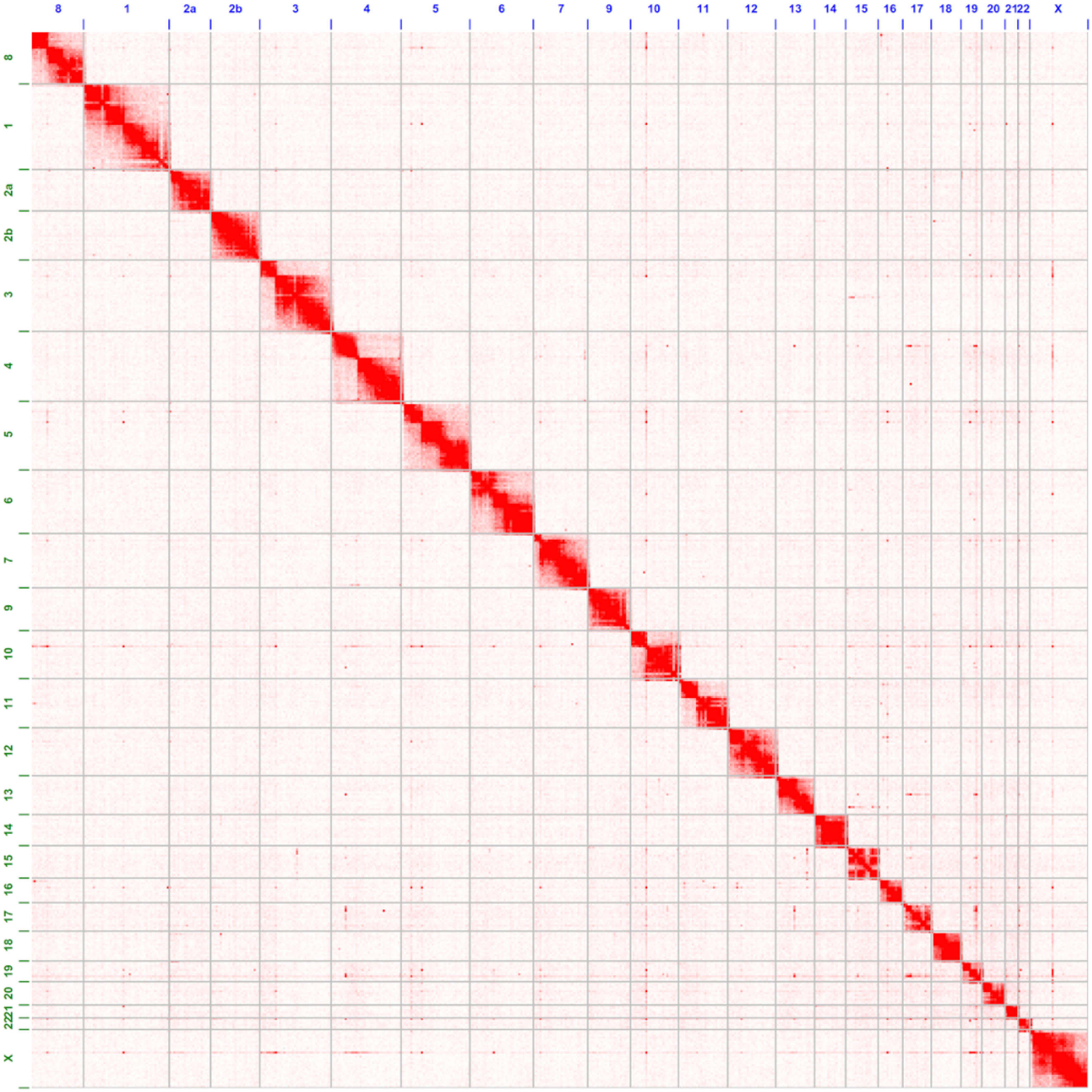
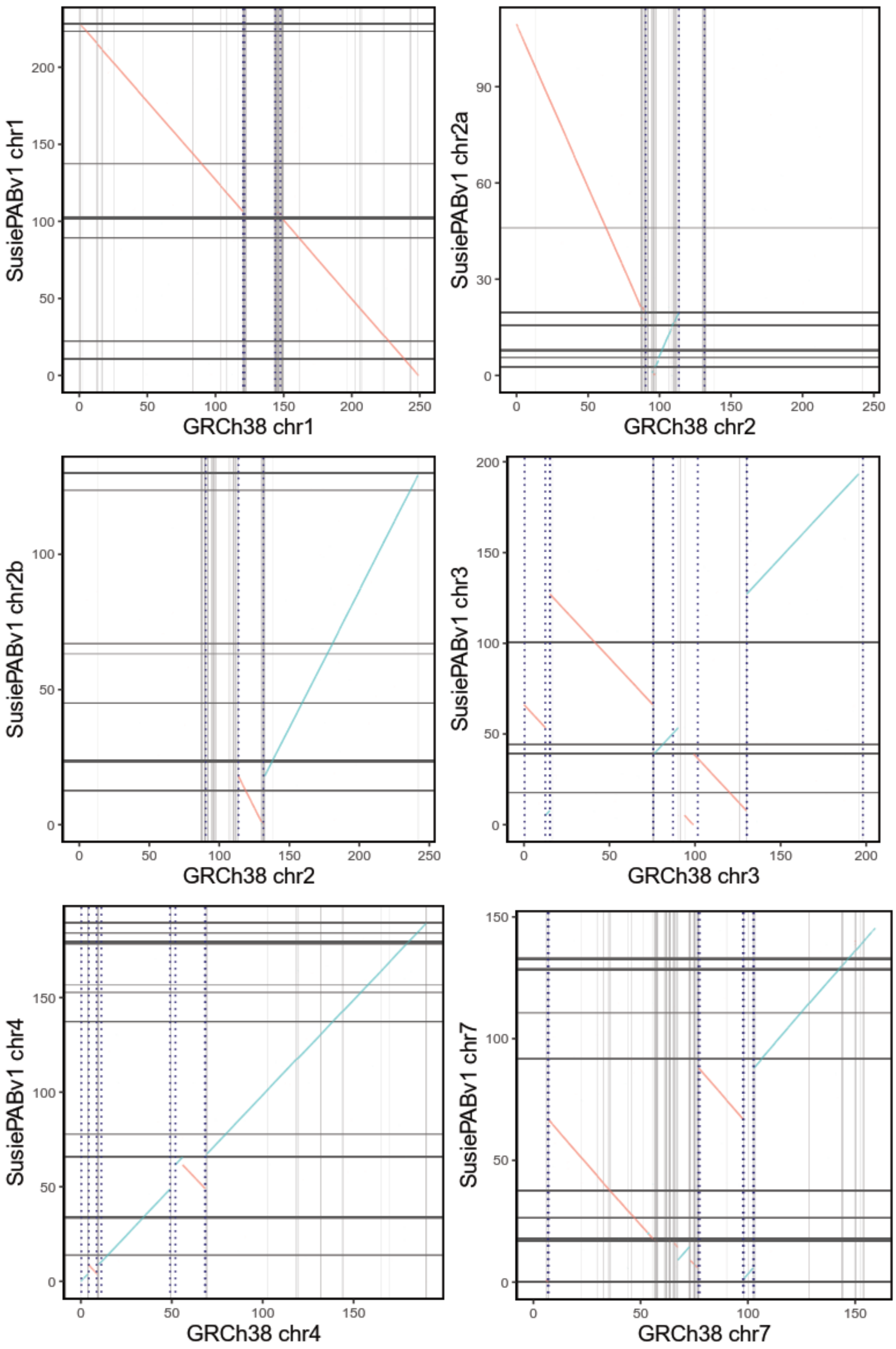


Fig. S21.
All-vs-all chromosome heatmap of the Hi-C data aligned to the Susie_PABv1 chromosomal AGP.


SusiePABv1 chr1
200
150
100
50
0
0
50
100
150
200
250
GRCh38 chr1
SusiePABv1 chr2a
90
60
30
0
0
50
100
150
200
250
GRCh38 chr2
SusiePABv1 chr2b
100
50
0
0
50
100
150
200
250
GRCh38 chr2
SusiePABv1 chr3
200
150
100
50
0
0
50
100
150
200
GRCh38 chr3
SusiePABv1 chr4
150
100
50
0
0
50
100
150
GRCh38 chr4
SusiePABv1 chr7
150
100
50
0
0
50
100
150
GRCh38 chr7

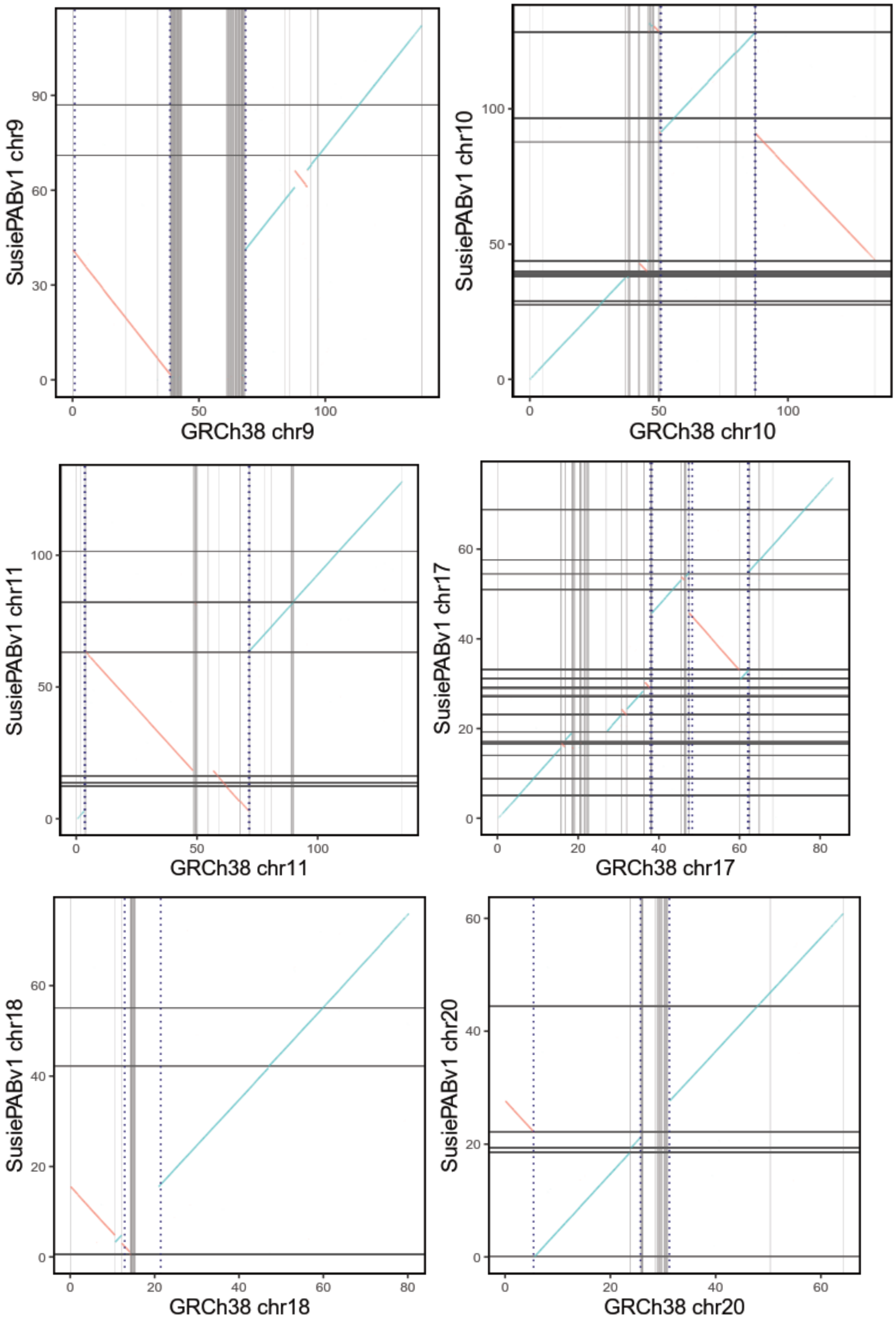
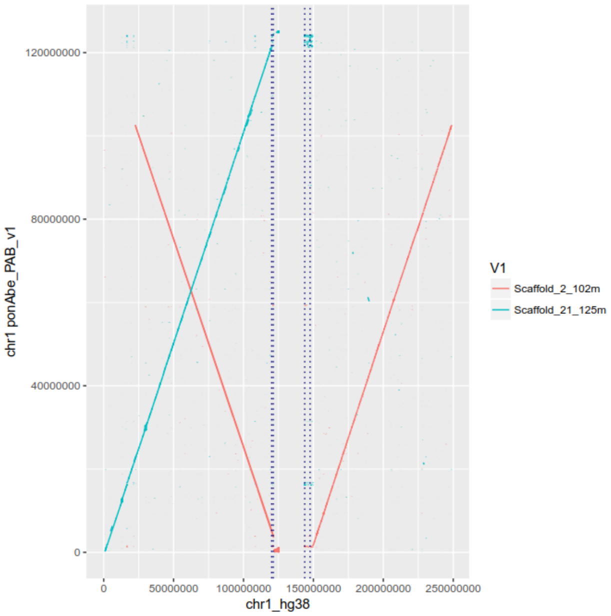
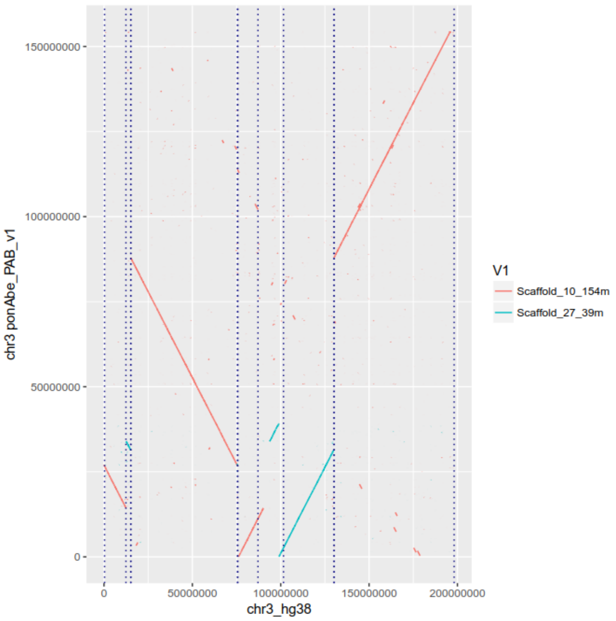
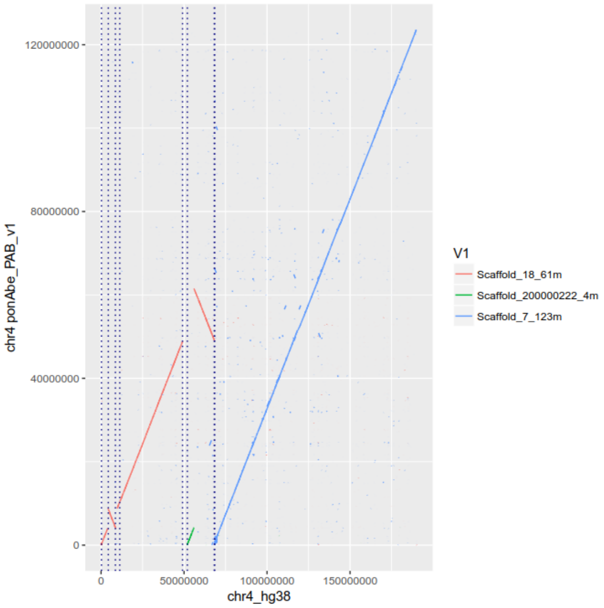


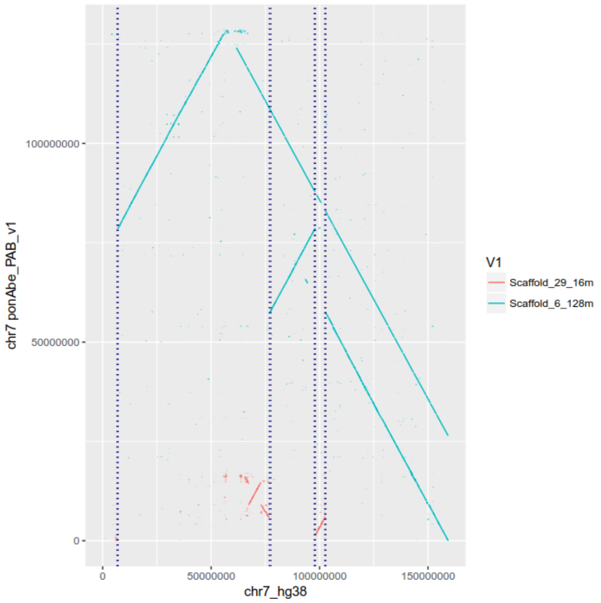
Fig. S22.

Known evolutionary inversions in orangutan relative to human captured in Susie_PABv1 chromosomes.


chr1 ponAbe_PAB_v1
120000000
80000000
40000000
0
0
50000000
100000000
150000000
200000000
250000000
chr1_hg38
V1
Scaffold_2_102m
Scaffold_21_125m


chr3 ponAbe_PAB_v1
150000000
100000000
50000000
0
0
50000000
100000000
150000000
200000000
chr3_hg38
V1
Scaffold_10_154m
Scaffold_27_39m


chr4 ponAbe_PAB_v1
120000000
80000000
40000000
0
0
50000000
100000000
150000000
chr4_hg38
V1
Scaffold_18_61m
Scaffold_200000222_4m
Scaffold_7_123m


chr7 ponAbe_PAB_v1
100000000
50000000
0
0
50000000
100000000
150000000
chr7_hg38
V1
Scaffold_29_16m
Scaffold_6_128m

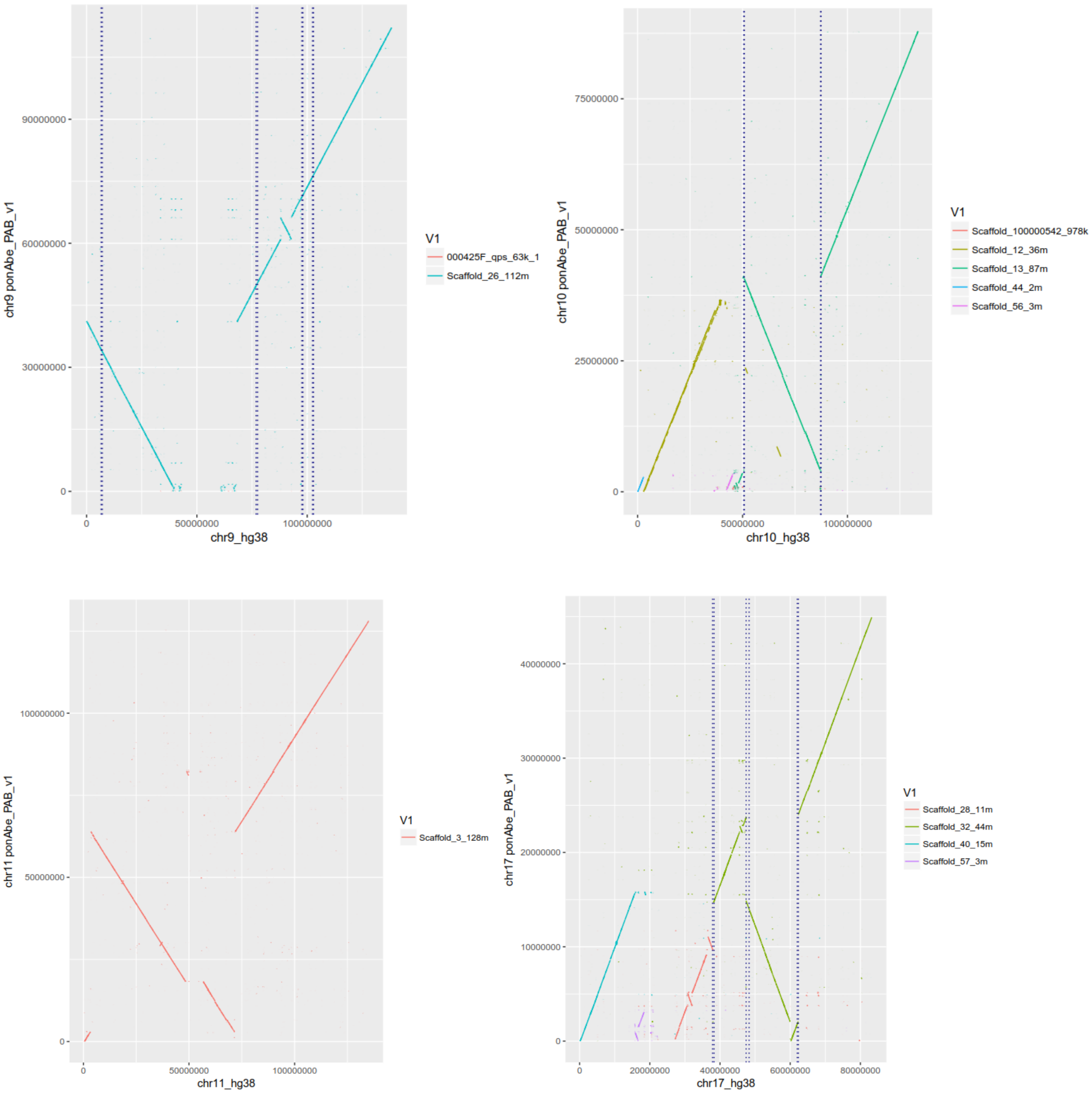


Fig. S23.
Boundaries of known evolutionary rearrangements captured by the Susie_PABv1 scaffolds.

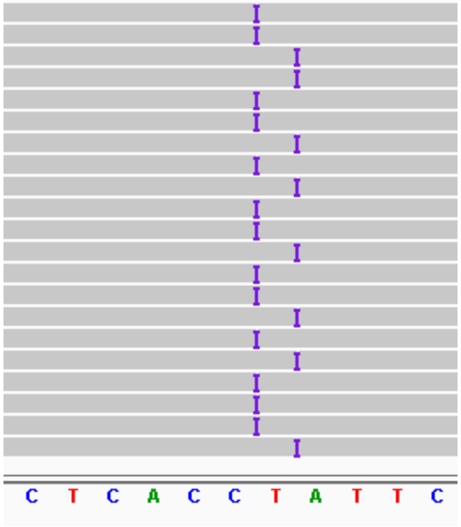


Fig. S24.

Remaining PTVs after Quiver and Pilon.
Illumina reads (gray bars) of both alleles aligned were against the (incorrect) reference (which has already been processed by Quiver and Pilon). Each purple bar indicates an insertion respect to the reference. Each read has one or the other insertion, but not both. Thus, the reference should have one of these two insertions, not both, but it has neither. Shown is YRI_HSAv1 000002F_1_20976909_quiver_pilon:18,556,558-18,556,598.

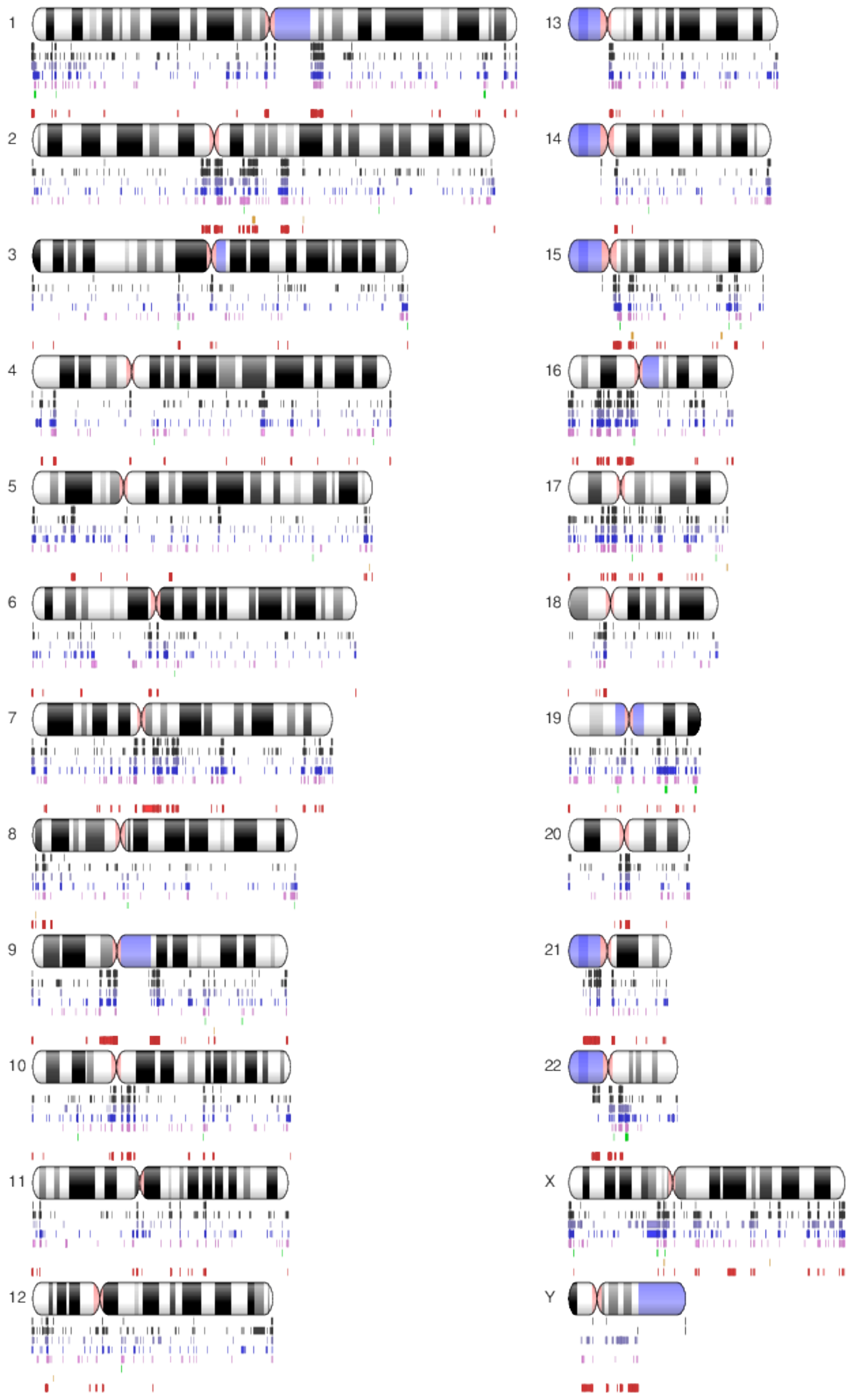


Fig. S25.

Ideogram showing short alignments (<1 Mbp) for human and NHP against GRCh38.
From top to bottom: CHM13, Yoruban, chimpanzee, orangutan, the intersection between NHPs complemented against human assemblies (56 regions), and human SDs ≥50 kbp.

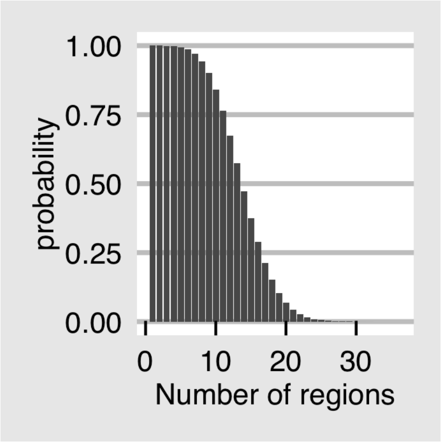


Fig. S26.

The cumulative density function from the permutation test.
Each bar shows the probability of finding N or more fragmented alignment regions in the NHPs.

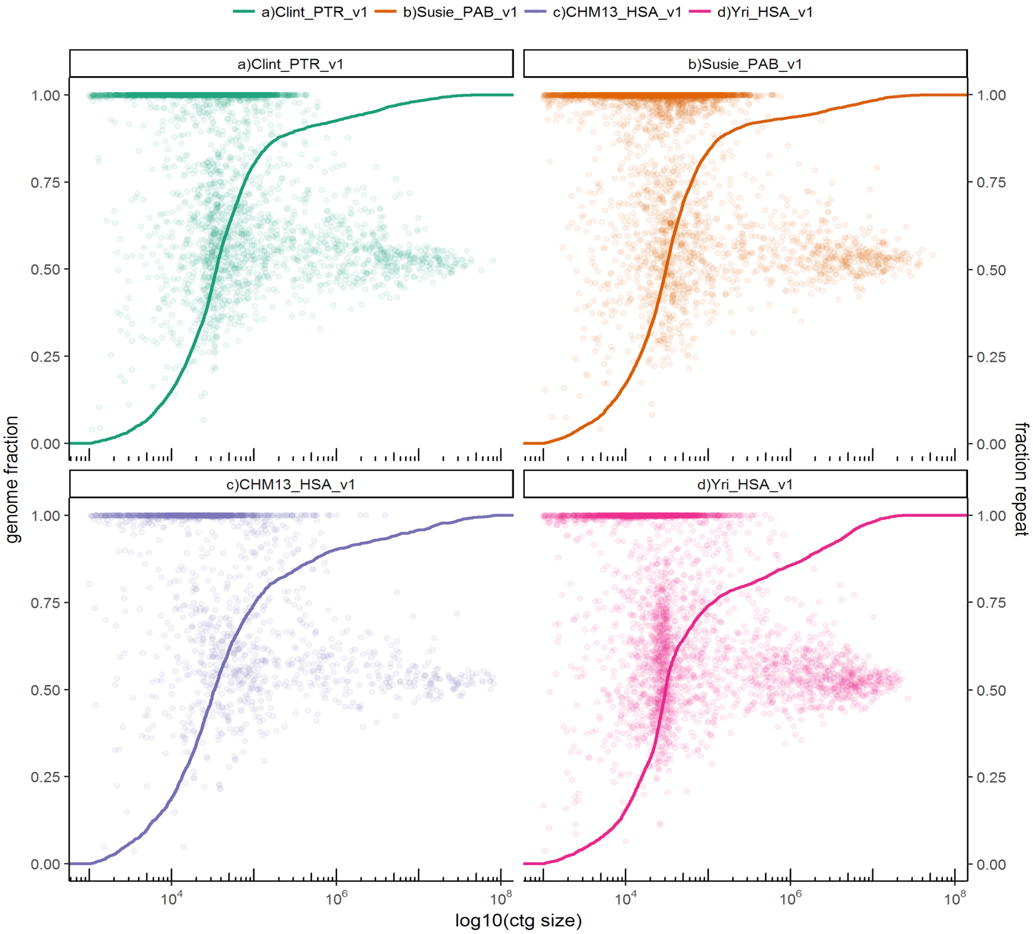


Fig. S27.

Genome representation by contig sizes and repeat content.
The curve represents the contig sizes, and the points represent the fraction of repeats in the contigs. Repeat content is determined using RepeatMasker and Tandem Repeats Finder. The fraction of repeats converges towards 0.5 in larger contigs. Contigs are sorted from smallest to largest, from left to right, and summed until all bases in the assembly are represented (empirical cumulative distribution).

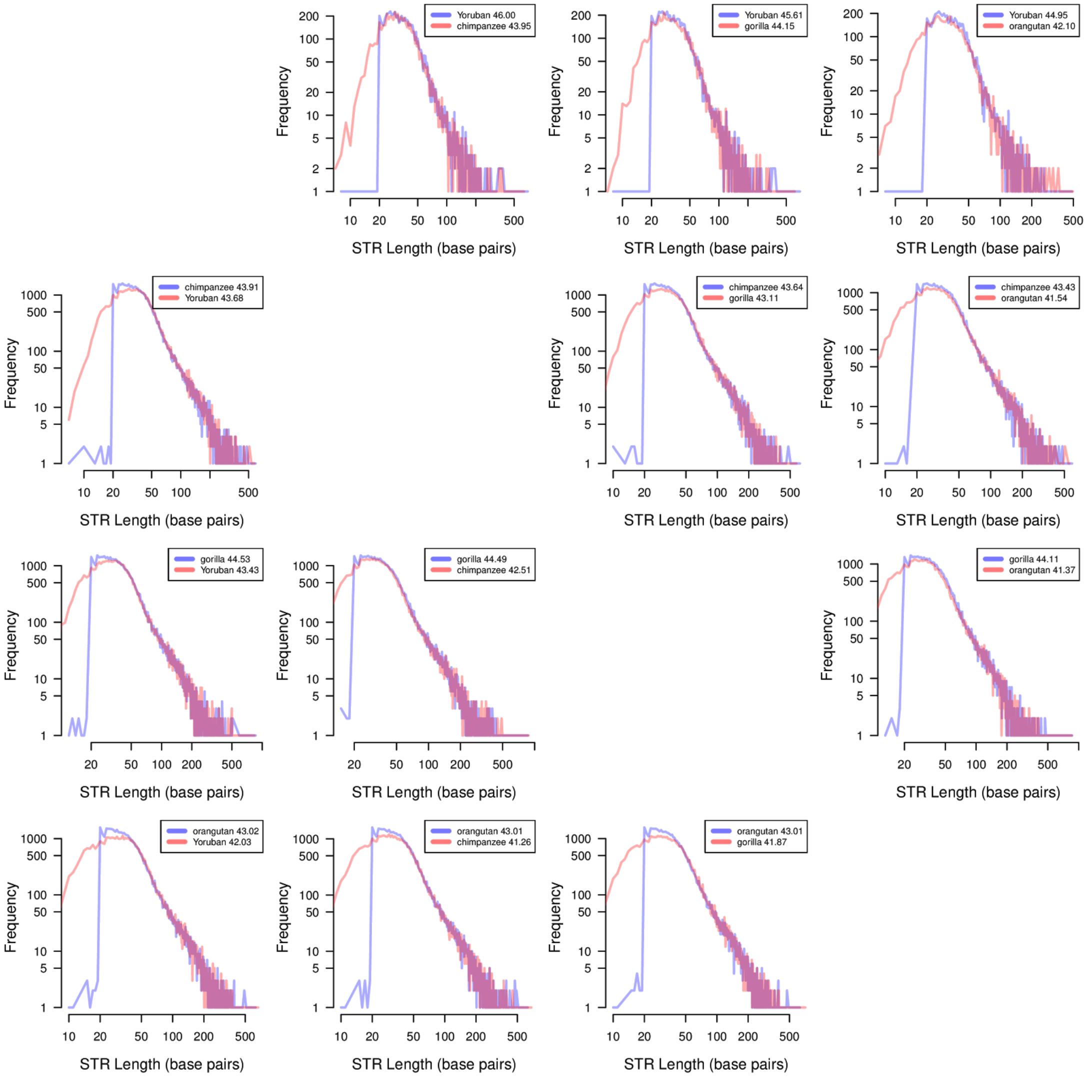
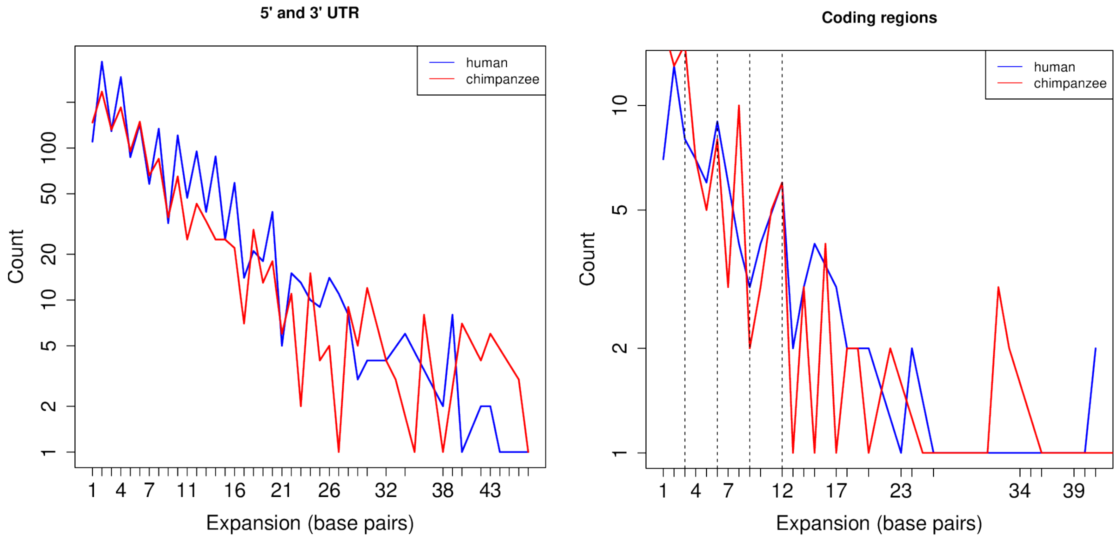


Fig. S28.

STR length distribution.

The distribution of lengths of STRs from RepeatMasker discovery sets (blue) and inferred orthologous STRs (red). The discovery sets are limited to RepeatMasker annotations starting at 20 bp while the inferred sets are of any length, generating a discovery set dependent bias for STRs below 40 bp.



*

Fig. S29.

Exonic STR expansions.

The counts of expansions of STRs in human and chimpanzee coding sequences (left) and UTR sequences (right). The difference in the distributions is insignificant (n = 310, p = 0.856, KS test). STR expansions in UTR sequences were similarly not significant (n = 2,794, p = 0.162, KS test).

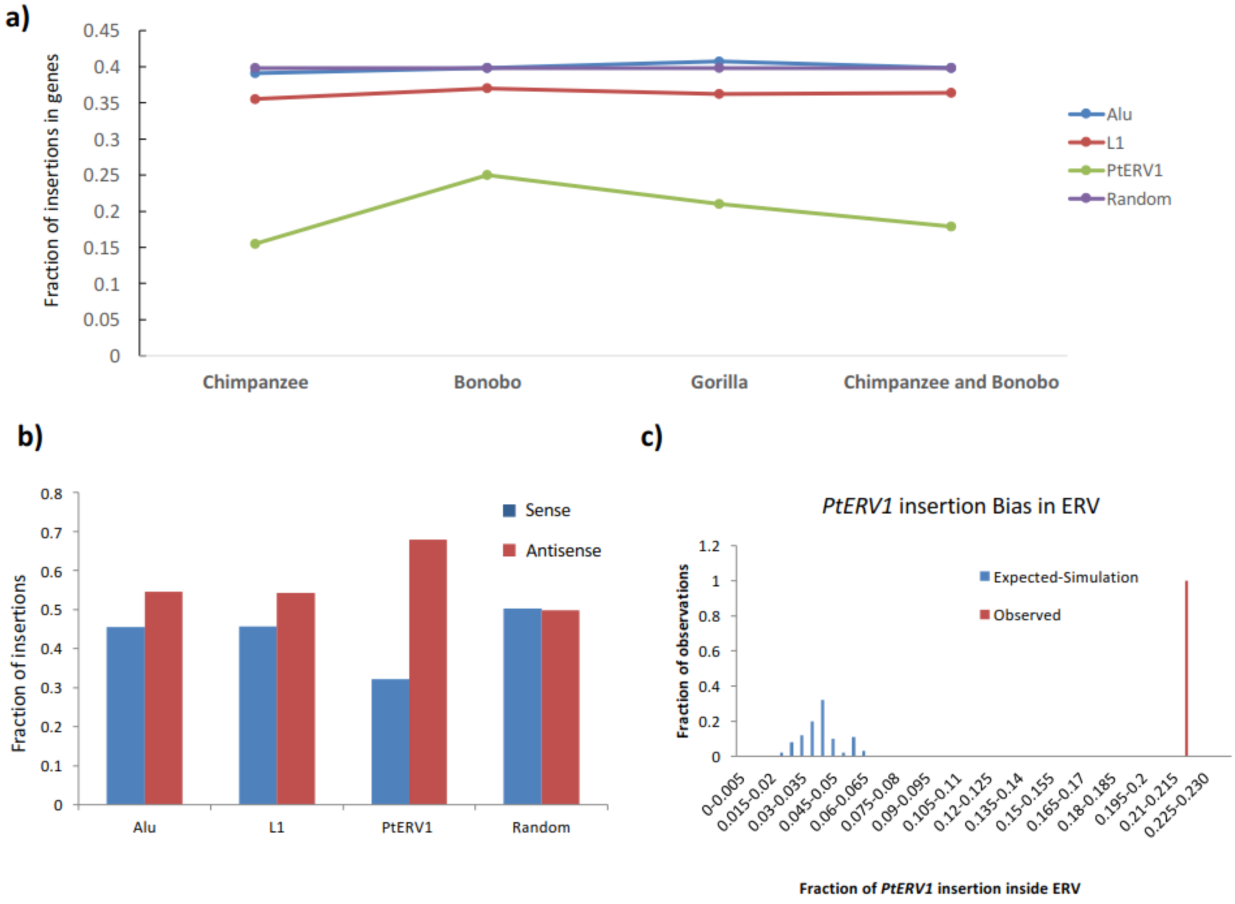


Fig. S30.

PtERV1 integration biases.

Pattern of PtERV1 insertion in genome. (a) A comparison of the fraction of lineage-specific retrotransposons (Alu, PtERV1, L1) mapping within genic regions (in chimpanzee, gorilla, bonobo, and ancestrally shared events in chimpanzee and bonobo). Integration sites were defined based on mapping of Illumina WGS OEA to the human reference (*86*). The data show a depletion of PtERV1 in genic regions across all species compared to other retrotransposons and random expectation. (b) Reduction of PtERV1 insertions in sense orientation in comparison to antisense orientation within genes. Retrotransposon orientation was defined based on 5' to 3' synthesis and classified with respect to sense/antisense orientation of RefSeq gene annotation. (c) Significant enrichment of PtERV1 insertion in annotated endogenous retroviral repeats. 21% of all PtERV1 map within ancestral ERVs compared to null distribution of 5% based on human genome organization (GRCh36).

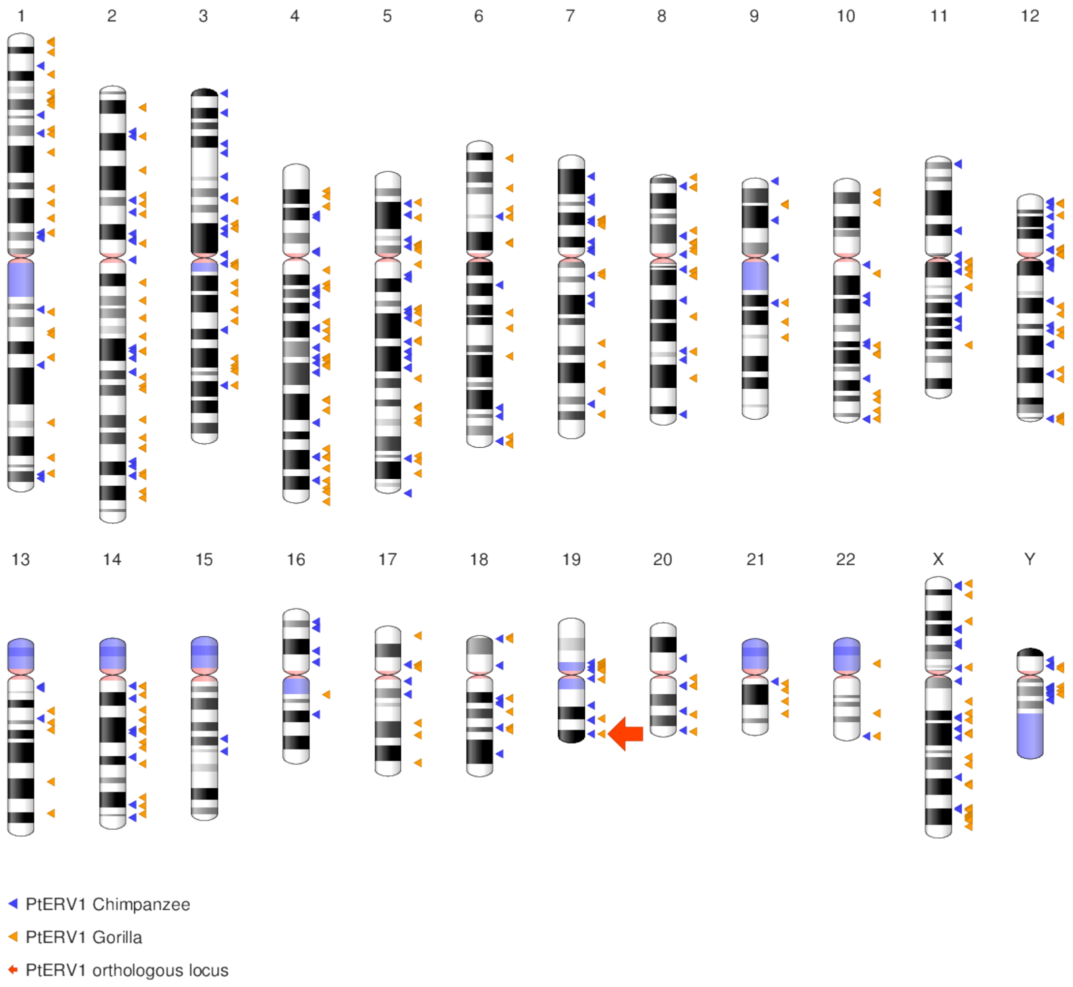


Fig. S31.

Ideogram of PtERV1 loci.

These events were identified in chimpanzee and gorilla from assembly-based and OEA (one-end anchored) methods. The PtERV1 orthologous locus on chr19 is indicated with a red arrow.

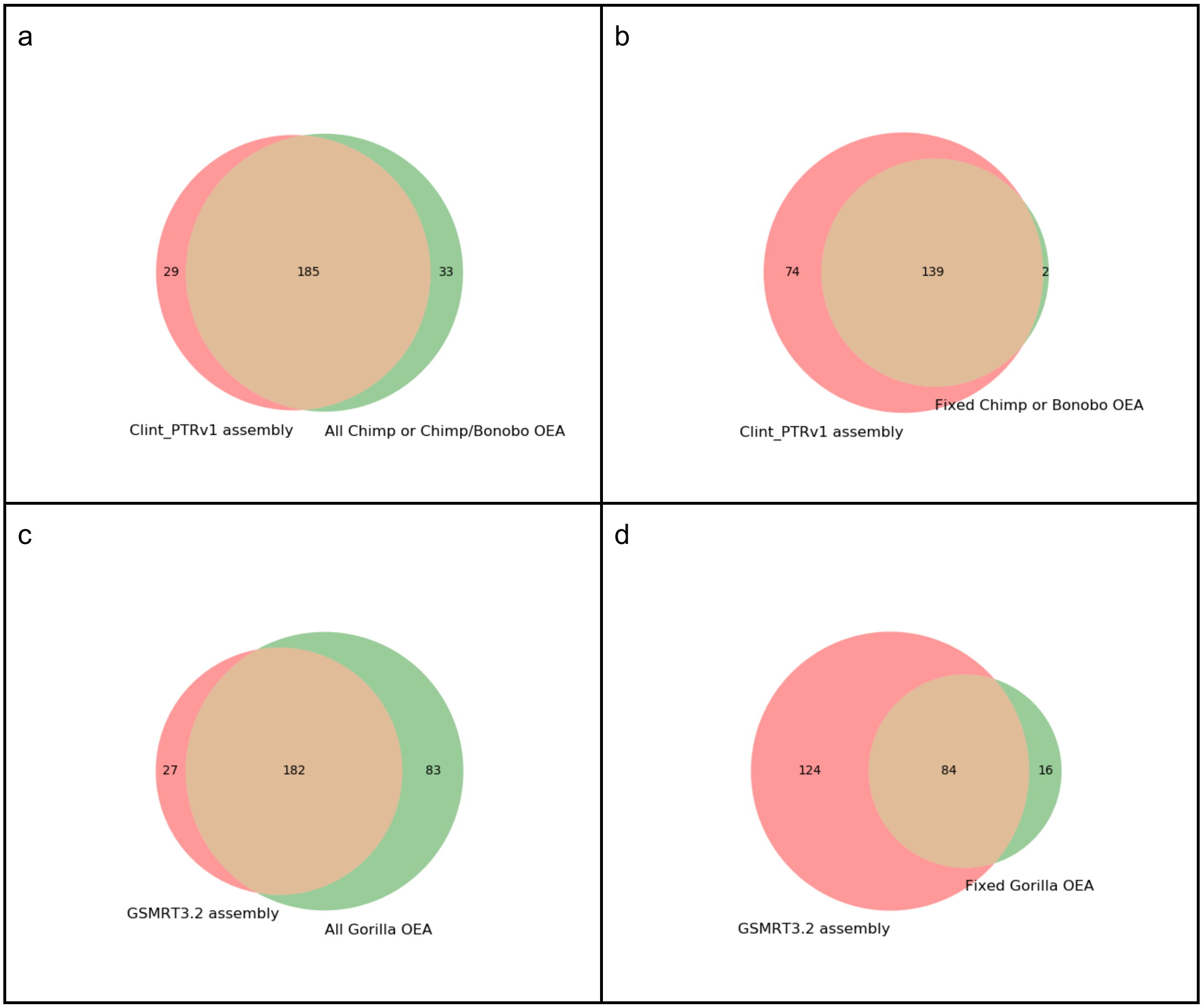


Fig. S32.

PtERV1 locus intersections between assembly and OEA methods.
(a) All chimpanzee insertions, (b) chimpanzee insertions detected by assembly or classified as fixed in chimpanzee by OEA, (c) all gorilla insertions, and (d) gorilla insertions detected by assembly or classified as fixed in chimpanzee by OEA.

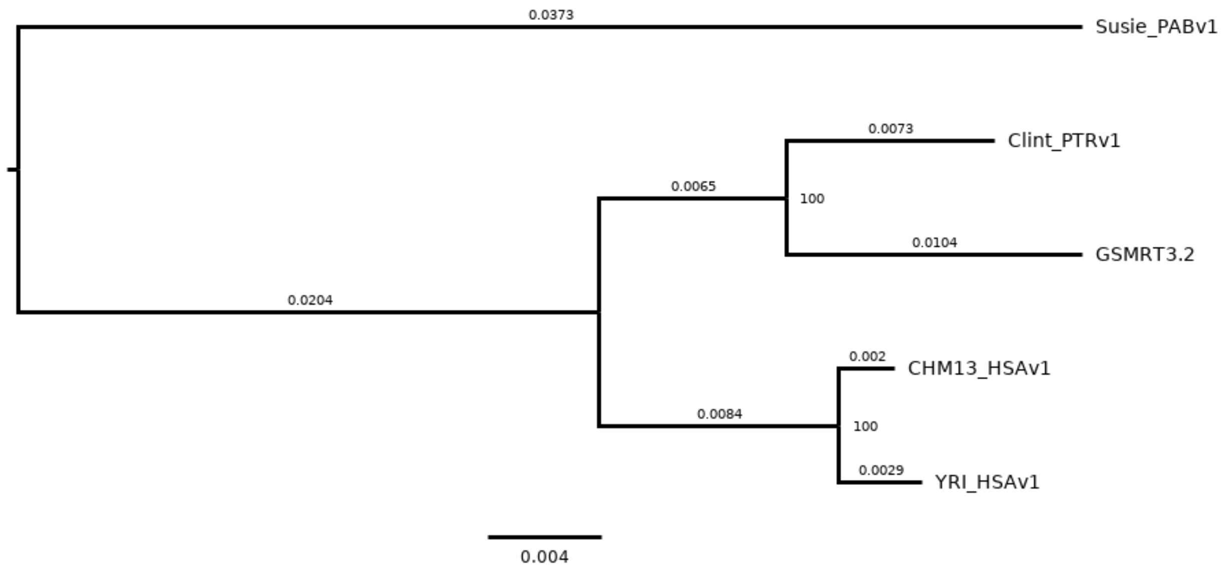


Fig. S33.

Orthologous PtERV1 tree supporting ILS.
Branch lengths (substitutions per site) are shown above the lineages and node bootstrap support label internal nodes (percentage of replicates supporting split; 1,000 replicates). The phylogeny was generated with RAxML 8.2.9 using a GTR+Gamma model from a 12,108 bp multiple-sequence-alignment.

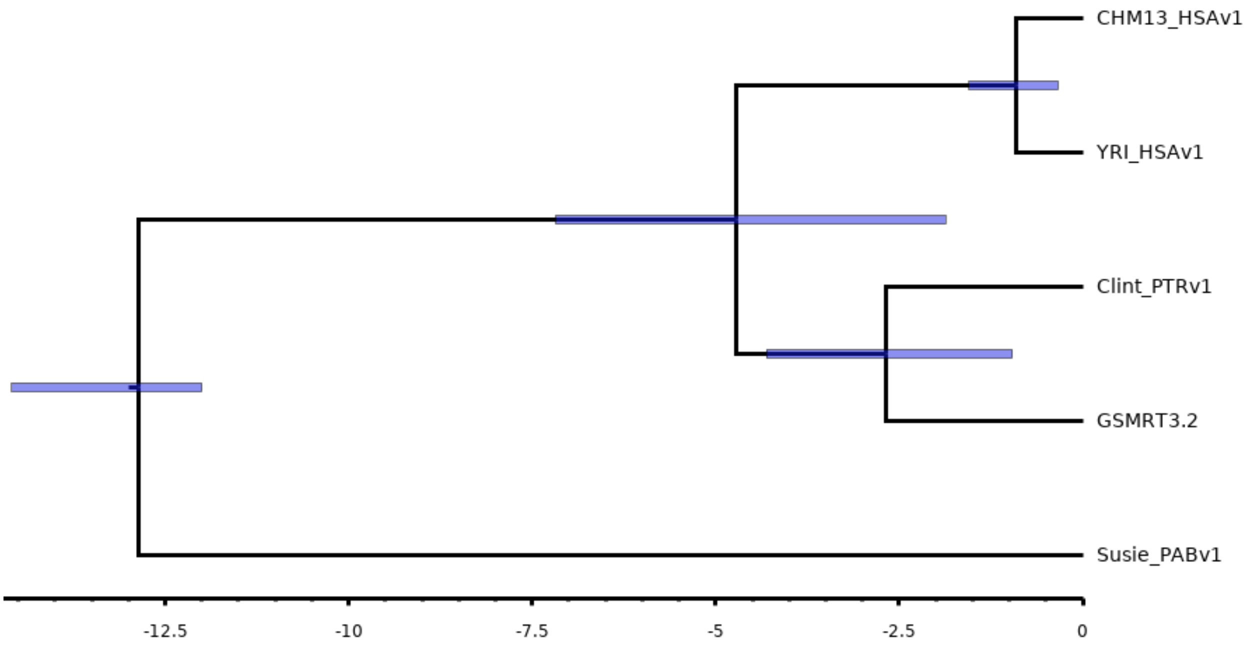


Fig. S34.

Maximum clade credibility tree of PtERV1 chr19 orthologous locus.
Scale bar indicates divergence time in millions of years before present. Blue bars indicate 95% highest posterior density of node age.

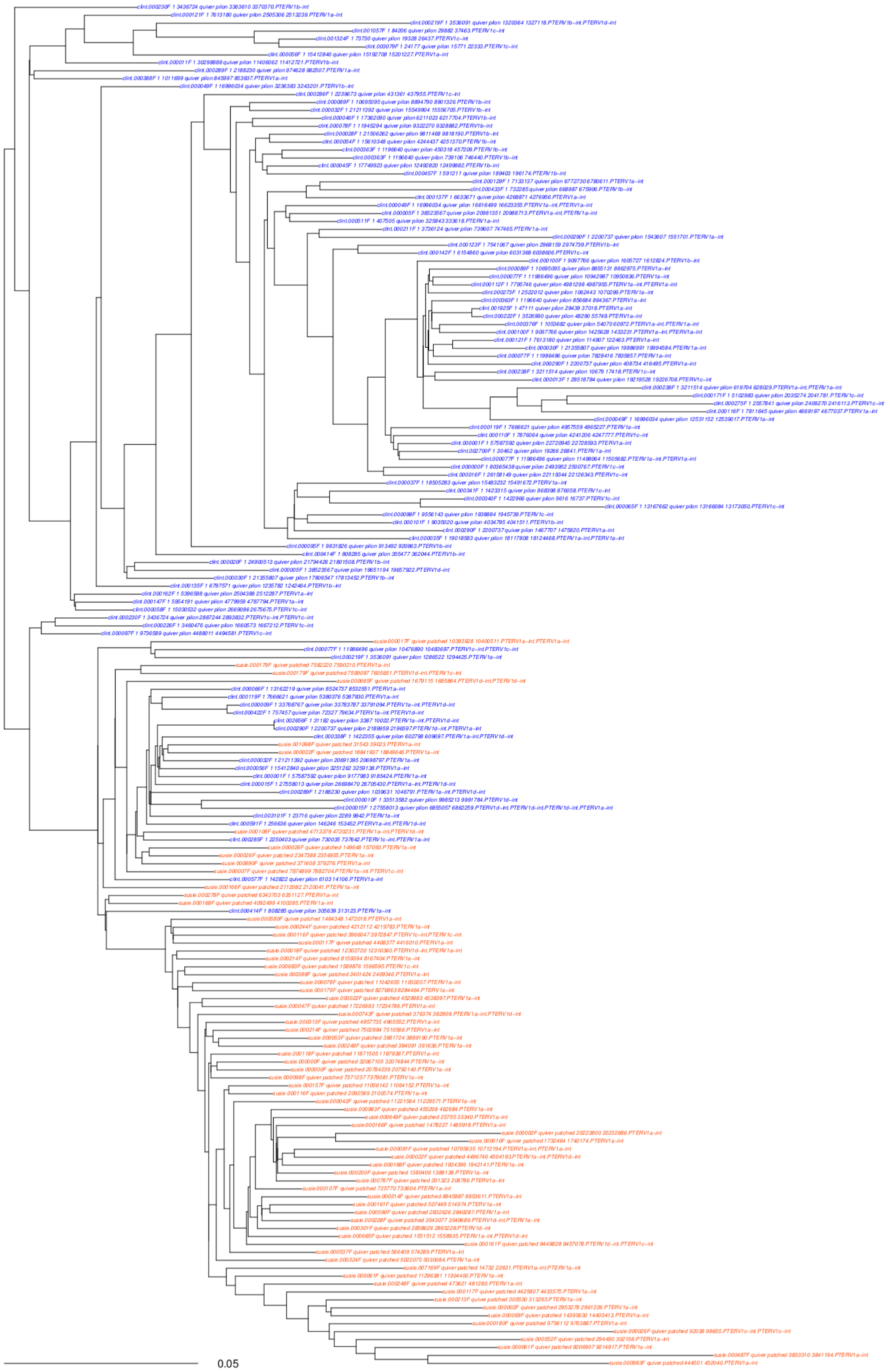

clint.000230F 1 3436724 quiver pilon 3363610 3370370.PTERV1b-int
clint.000121F 1 7613180 quiver pilon 2505306 2513239.PTERV1a-int
clint.000219F 1 3536091 quiver pilon 1320364 1327118.PTERV1b-int.PTERV1d-int
clint.001057F 1 84206 quiver pilon 29882 37463.PTERV1c-int
clint.001324F 1 73730 quiver pilon 19328 26437.PTERV1c-int
clint.003079F 1 24177 quiver pilon 15771 22333.PTERV1c-int
clint.000056F 1 15412840 quiver pilon 15192708 15201227.PTERV1a-int
clint.000011F 1 30298888 quiver pilon 11406062 11412721.PTERV1b-int
clint.000289F 1 2188230 quiver pilon 974628 982507.PTERV1a-int
clint.000388F 1 1011699 quiver pilon 845997 853937.PTERV1a-int
clint.000049F 1 16996034 quiver pilon 3236383 3243201.PTERV1b-int
clint.000286F 1 2239673 quiver pilon 431361 437955.PTERV1c-int
clint.000089F 1 10695095 quiver pilon 8894790 8901326.PTERV1b-int
clint.000032F 1 21211392 quiver pilon 15549904 15556705.PTERV1b-int
clint.000046F 1 17362090 quiver pilon 6211023 6217704.PTERV1b-int
clint.000078F 1 11945294 quiver pilon 9322270 9328882.PTERV1b-int
clint.000028F 1 21506262 quiver pilon 9811468 9818190.PTERV1b-int
clint.000054F 1 15610348 quiver pilon 4244437 4251370.PTERV1b-int
clint.000363F 1 1196640 quiver pilon 450318 457209.PTERV1b-int
clint.000363F 1 1196640 quiver pilon 739106 746440.PTERV1b-int
clint.000045F 1 17749923 quiver pilon 12492820 12499882.PTERV1b-int
clint.000457F 1 591211 quiver pilon 189403 196174.PTERV1b-int
clint.000129F 1 7133137 quiver pilon 6772730 6780611.PTERV1a-int
clint.000433F 1 732285 quiver pilon 668987 675906.PTERV1b-int
clint.000137F 1 6633671 quiver pilon 4268871 4276956.PTERV1a-int
clint.000049F 1 16996034 quiver pilon 16616499 16623355.PTERV1a-int.PTERV1a-int
clint.000005F 1 38523567 quiver pilon 20981351 20988713.PTERV1a-int.PTERV1a-int
clint.000511F 1 407505 quiver pilon 325843 333618.PTERV1a-int
clint.000211F 1 3736124 quiver pilon 739607 747465.PTERV1a-int
clint.000290F 1 2200737 quiver pilon 1543607 1551701.PTERV1a-int
clint.000123F 1 7541067 quiver pilon 2968159 2974739.PTERV1b-int
clint.000142F 1 6154860 quiver pilon 6031368 6038606.PTERV1c-int
clint.000100F 1 9097766 quiver pilon 1605727 1612824.PTERV1b-int
clint.000089F 1 10695095 quiver pilon 8855131 8862975.PTERV1a-int
clint.000077F 1 11986496 quiver pilon 10942987 10950836.PTERV1a-int
clint.000112F 1 7795746 quiver pilon 4981298 4987955.PTERV1a-int.PTERV1a-int
clint.000273F 1 2522012 quiver pilon 1062443 1070299.PTERV1a-int
clint.000363F 1 1196640 quiver pilon 856684 864367.PTERV1a-int
clint.001925F 1 47111 quiver pilon 29439 37018.PTERV1a-int
clint.000222F 1 3526990 quiver pilon 48290 55749.PTERV1a-int
clint.000376F 1 1053682 quiver pilon 54070 60972.PTERV1a-int.PTERV1a-int
clint.000100F 1 9097766 quiver pilon 1425628 1433231.PTERV1a-int.PTERV1a-int
clint.000121F 1 7613180 quiver pilon 114907 122463.PTERV1a-int
clint.000030F 1 21355807 quiver pilon 19986991 19994584.PTERV1a-int
clint.000077F 1 11986496 quiver pilon 7828416 7835857.PTERV1a-int
clint.000290F 1 2200737 quiver pilon 408734 416495.PTERV1a-int
clint.000238F 1 3211514 quiver pilon 10679 17418.PTERV1c-int
clint.000013F 1 28518784 quiver pilon 19219528 19226708.PTERV1c-int
clint.000238F 1 3211514 quiver pilon 619704 628029.PTERV1a-int.PTERV1a-int
clint.000171F 1 5102983 quiver pilon 2035274 2041781.PTERV1c-int
clint.000275F 1 2557841 quiver pilon 2409270 2416113.PTERV1c-int
clint.000116F 1 7811645 quiver pilon 4669197 4677037.PTERV1a-int
clint.000049F 1 16996034 quiver pilon 12531152 12539017.PTERV1a-int
clint.000119F 1 7666621 quiver pilon 4957559 4965227.PTERV1a-int
clint.000110F 1 7876064 quiver pilon 4241206 4247777.PTERV1c-int
clint.000001F 1 57587592 quiver pilon 22720945 22728593.PTERV1a-int
clint.002700F 1 30462 quiver pilon 19266 26841.PTERV1a-int
clint.000077F 1 11986496 quiver pilon 11498064 11505682.PTERV1a-int.PTERV1a-int
clint.000000F 1 80365438 quiver pilon 2493952 2500767.PTERV1c-int
clint.000016F 1 26158149 quiver pilon 22119344 22126343.PTERV1c-int
clint.000037F 1 18505283 quiver pilon 15483232 15491672.PTERV1a-int
clint.000341F 1 1423315 quiver pilon 868398 876058.PTERV1c-int
clint.000340F 1 1422966 quiver pilon 9616 16737.PTERV1c-int
clint.000065F 1 13167662 quiver pilon 13166084 13173050.PTERV1c-int
clint.000098F 1 9556143 quiver pilon 1938884 1945739.PTERV1c-int
clint.000101F 1 9035020 quiver pilon 4034795 4041511.PTERV1b-int
clint.000290F 1 2200737 quiver pilon 1467707 1475820.PTERV1a-int
clint.000035F 1 19018583 quiver pilon 18117808 18124468.PTERV1a-int.PTERV1a-int
clint.000095F 1 9831826 quiver pilon 913492 920863.PTERV1b-int
clint.000414F 1 808285 quiver pilon 355477 362044.PTERV1b-int
clint.000020F 1 24900513 quiver pilon 21794426 21801508.PTERV1b-int
clint.000005F 1 38523567 quiver pilon 19651194 19657922.PTERV1d-int
clint.000030F 1 21355807 quiver pilon 17806547 17813452.PTERV1b-int
clint.000135F 1 6797571 quiver pilon 1235782 1242464.PTERV1b-int
clint.000162F 1 5396588 quiver pilon 2504388 2512287.PTERV1a-int
clint.000147F 1 5954191 quiver pilon 4779959 4787794.PTERV1a-int
clint.000058F 1 15030532 quiver pilon 2669086 2675675.PTERV1c-int
clint.000230F 1 3436724 quiver pilon 2887244 2893832.PTERV1c-int.PTERV1c-int
clint.000226F 1 3460476 quiver pilon 1660573 1667212.PTERV1c-int
clint.000097F 1 9736589 quiver pilon 4488011 4494581.PTERV1c-int
susie.000017F quiver patched 10393928 10400511.PTERV1a-int.PTERV1a-int
clint.000077F 1 11986496 quiver pilon 10476890 10483697.PTERV1c-int.PTERV1c-int
clint.000219F 1 3536091 quiver pilon 1286522 1294425.PTERV1a-int
susie.000179F quiver patched 7582220 7590210.PTERV1a-int
susie.000179F quiver patched 7599097 7605651.PTERV1d-int.PTERV1c-int
susie.000665F quiver patched 1679115 1685864.PTERV1d-int.PTERV1d-int
clint.000066F 1 13162219 quiver pilon 8524737 8532551.PTERV1a-int
clint.000119F 1 7666621 quiver pilon 5380376 5387930.PTERV1a-int
clint.000009F 1 33768767 quiver pilon 33783787 33791094.PTERV1a-int.PTERV1d-int
clint.000422F 1 757457 quiver pilon 72327 79634.PTERV1a-int.PTERV1d-int
clint.002656F 1 31182 quiver pilon 3387 10022.PTERV1a-int.PTERV1d-int
clint.000290F 1 2200737 quiver pilon 2189959 2196597.PTERV1d-int.PTERV1a-int
clint.000338F 1 1422355 quiver pilon 602798 609697.PTERV1a-int.PTERV1d-int
susie.001098F quiver patched 31543 39023.PTERV1a-int
susie.000022F quiver patched 18841937 18849646.PTERV1a-int
clint.000032F 1 21211392 quiver pilon 20691395 20698797.PTERV1a-int
clint.000056F 1 15412840 quiver pilon 3251262 3259138.PTERV1a-int
clint.000001F 1 57587592 quiver pilon 9177983 9185424.PTERV1a-int
clint.000015F 1 27558013 quiver pilon 26698470 26705430.PTERV1a-int.PTERV1d-int
clint.000289F 1 2188230 quiver pilon 1039631 1046791.PTERV1a-int.PTERV1d-int
clint.000010F 1 33513582 quiver pilon 9985213 9991784.PTERV1d-int
clint.000015F 1 27558013 quiver pilon 6855057 6862259.PTERV1d-int.PTERV1d-int.PTERV1d-int.PTERV1a-int
clint.003101F 1 23716 quiver pilon 2289 9842.PTERV1a-int
clint.000591F 1 256636 quiver pilon 146246 153452.PTERV1a-int.PTERV1d-int
susie.000108F quiver patched 4713379 4720231.PTERV1a-int.PTERV1d-int
clint.000285F 1 2250403 quiver pilon 730035 737642.PTERV1c-int.PTERV1a-int
susie.000026F quiver patched 149648 157093.PTERV1a-int
susie.000026F quiver patched 2347388 2354955.PTERV1a-int
susie.000890F quiver patched 371608 379276.PTERV1a-int
susie.000007F quiver patched 7874899 7882704.PTERV1a-int.PTERV1c-int
clint.000577F 1 142822 quiver pilon 6103 14106.PTERV1a-int
susie.000166F quiver patched 2112082 2120041.PTERV1a-int
susie.000278F quiver patched 6343703 6351127.PTERV1a-int
susie.000168F quiver patched 4092499 4100285.PTERV1a-int
clint.000414F 1 808285 quiver pilon 305639 313123.PTERV1a-int
susie.000580F quiver patched 1464348 1472018.PTERV1a-int
susie.000244F quiver patched 4212112 4219783.PTERV1a-int
susie.000116F quiver patched 3966047 3972847.PTERV1c-int.PTERV1c-int
susie.000117F quiver patched 4408377 4416010.PTERV1a-int
susie.000018F quiver patched 12302720 12310360.PTERV1d-int.PTERV1a-int
susie.000214F quiver patched 8159394 8167404.PTERV1a-int
susie.000683F quiver patched 1589876 1596595.PTERV1c-int
susie.000388F quiver patched 2401424 2409346.PTERV1a-int
susie.000078F quiver patched 11042655 11050207.PTERV1a-int
susie.000179F quiver patched 8276863 8284464.PTERV1a-int
susie.000022F quiver patched 4529983 4538397.PTERV1a-int
susie.000047F quiver patched 17226993 17234786.PTERV1a-int
susie.000743F quiver patched 376374 382909.PTERV1a-int.PTERV1d-int
susie.000313F quiver patched 4957735 4965552.PTERV1a-int
susie.000214F quiver patched 7502894 7510566.PTERV1a-int
susie.000053F quiver patched 3881724 3889190.PTERV1a-int
susie.000248F quiver patched 384091 391636.PTERV1a-int
susie.000118F quiver patched 11871505 11879387.PTERV1a-int
susie.000000F quiver patched 32067105 32074844.PTERV1a-int
susie.000000F quiver patched 20784239 20792140.PTERV1a-int
susie.000098F quiver patched 7371237 7379081.PTERV1a-int
susie.000157F quiver patched 11056142 11064152.PTERV1a-int
susie.000116F quiver patched 2092569 2100574.PTERV1a-int
susie.000042F quiver patched 11221564 11229571.PTERV1a-int
susie.000983F quiver patched 455208 462684.PTERV1a-int
susie.000649F quiver patched 25755 33340.PTERV1a-int
susie.000166F quiver patched 1478227 1485918.PTERV1a-int
susie.000002F quiver patched 20223800 20232686.PTERV1a-int
susie.000010F quiver patched 1732464 1740174.PTERV1a-int
susie.000091F quiver patched 10705635 10712194.PTERV1a-int.PTERV1a-int
susie.000022F quiver patched 4496746 4504193.PTERV1a-int.PTERV1d-int
susie.000188F quiver patched 1934396 1942141.PTERV1a-int
susie.000200F quiver patched 1380406 1388138.PTERV1a-int
susie.000787F quiver patched 201323 208786.PTERV1a-int
susie.000107F quiver patched 725770 733604.PTERV1a-int
susie.000214F quiver patched 8845887 8853611.PTERV1a-int
susie.000161F quiver patched 507449 514974.PTERV1a-int
susie.000590F quiver patched 2832626 2840287.PTERV1a-int
susie.000228F quiver patched 3543077 3549689.PTERV1d-int.PTERV1a-int
susie.000301F quiver patched 2858626 2865228.PTERV1d-int
susie.000665F quiver patched 1551512 1558935.PTERV1a-int.PTERV1d-int
susie.000161F quiver patched 9449828 9457078.PTERV1d-int.PTERV1c-int
susie.000537F quiver patched 566409 574289.PTERV1a-int
susie.000324F quiver patched 5022075 5030084.PTERV1a-int
susie.007169F quiver patched 14732 22631.PTERV1a-int.PTERV1a-int
susie.000061F quiver patched 11296381 11304400.PTERV1a-int
susie.000248F quiver patched 473621 481280.PTERV1a-int
susie.000117F quiver patched 4425807 4433575.PTERV1a-int
susie.000213F quiver patched 305530 313265.PTERV1a-int
susie.000002F quiver patched 2953278 2961226.PTERV1a-int
susie.000069F quiver patched 14395630 14403413.PTERV1a-int
susie.000180F quiver patched 9756112 9763887.PTERV1a-int
susie.000026F quiver patched 92038 98655.PTERV1c-int.PTERV1c-int
susie.000552F quiver patched 294490 302158.PTERV1a-int
susie.000061F quiver patched 9206907 9214917.PTERV1a-int
susie.000487F quiver patched 3833310 3841194.PTERV1a-int
susie.000983F quiver patched 444501 452040.PTERV1a-int
0.05

Fig. S35.

Maximum likelihood tree of full-length PtERV1 elements.
There are 101 chimpanzee loci shown in blue and 71 gorilla loci in red.

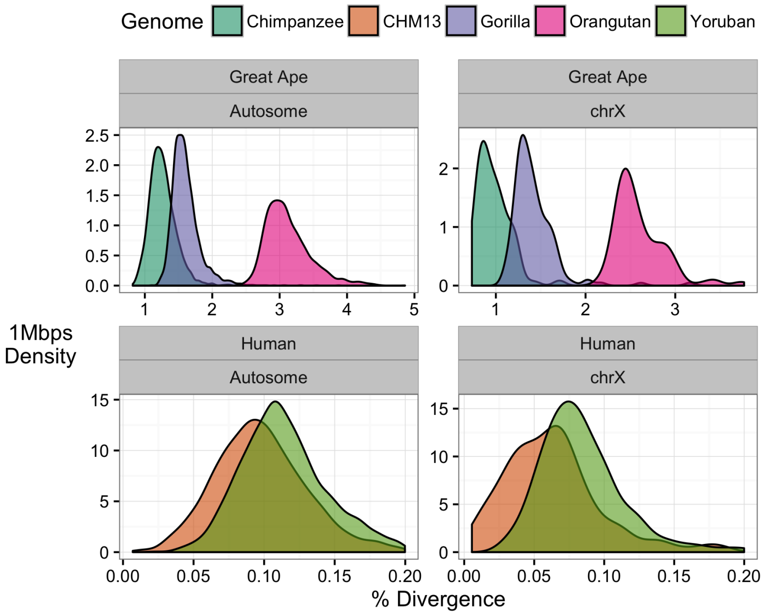


Fig. S36.

A density plot showing the percent divergence in non-overlapping 1 Mbp windows.
Each area is colored by genome and broken down by autosomes and chromosome X. Extreme values, >5% for great apes and >0.2% for human, are excluded for visualization purposes.

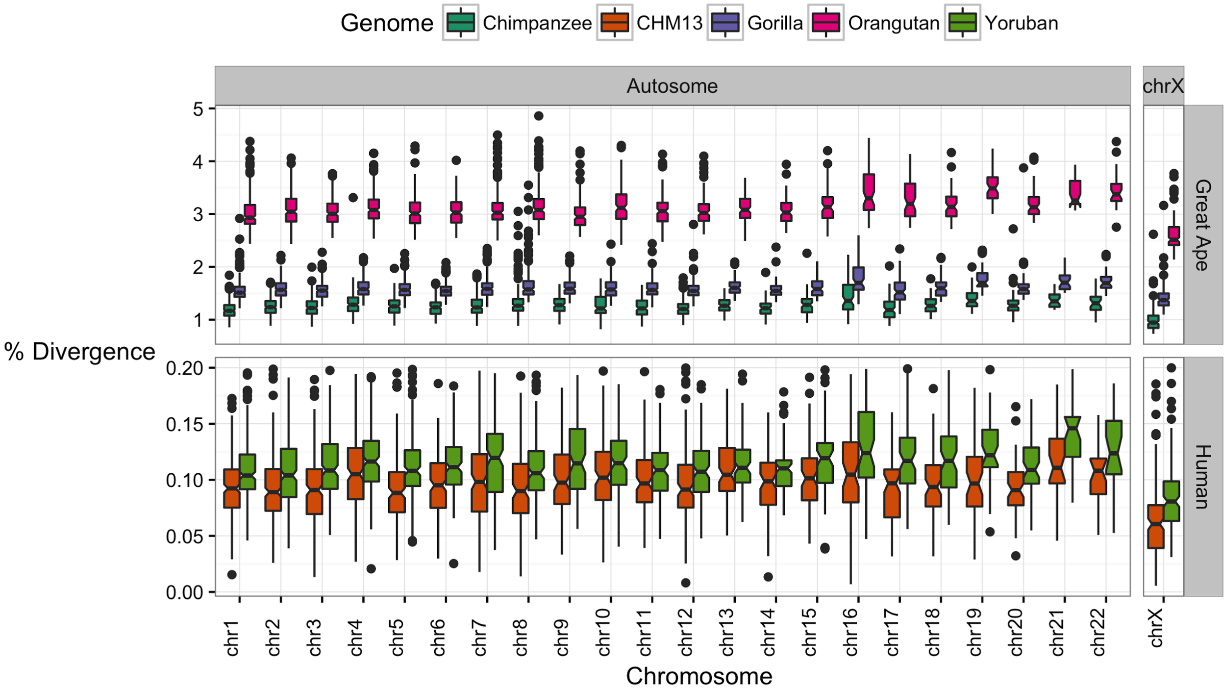
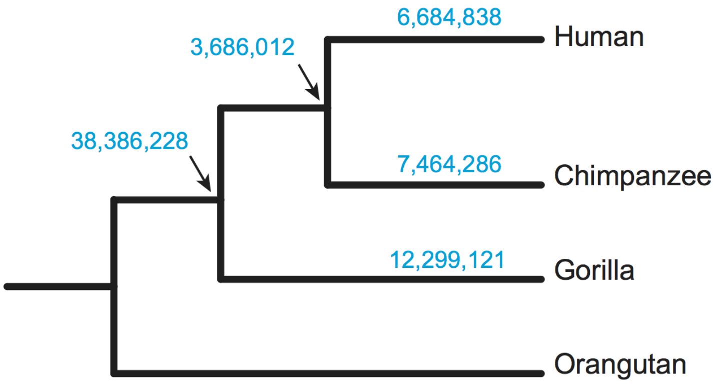


Fig. S37.

A boxplot showing divergence by chromosome.
Extreme values, greater than >5% for great apes and >0.2% for human, are excluded for visualization purposes.



Relative to GRCh38	
# SNVs	# INDELs
3,227,944*	1,602,838*
33,624,524	3,686,012
41,832,271	7,578,904
78,190,671	14,073,859

Fig. S38.
SNV and indel counts (relative to GRCh38) along the great ape ultra-metric cladogram.

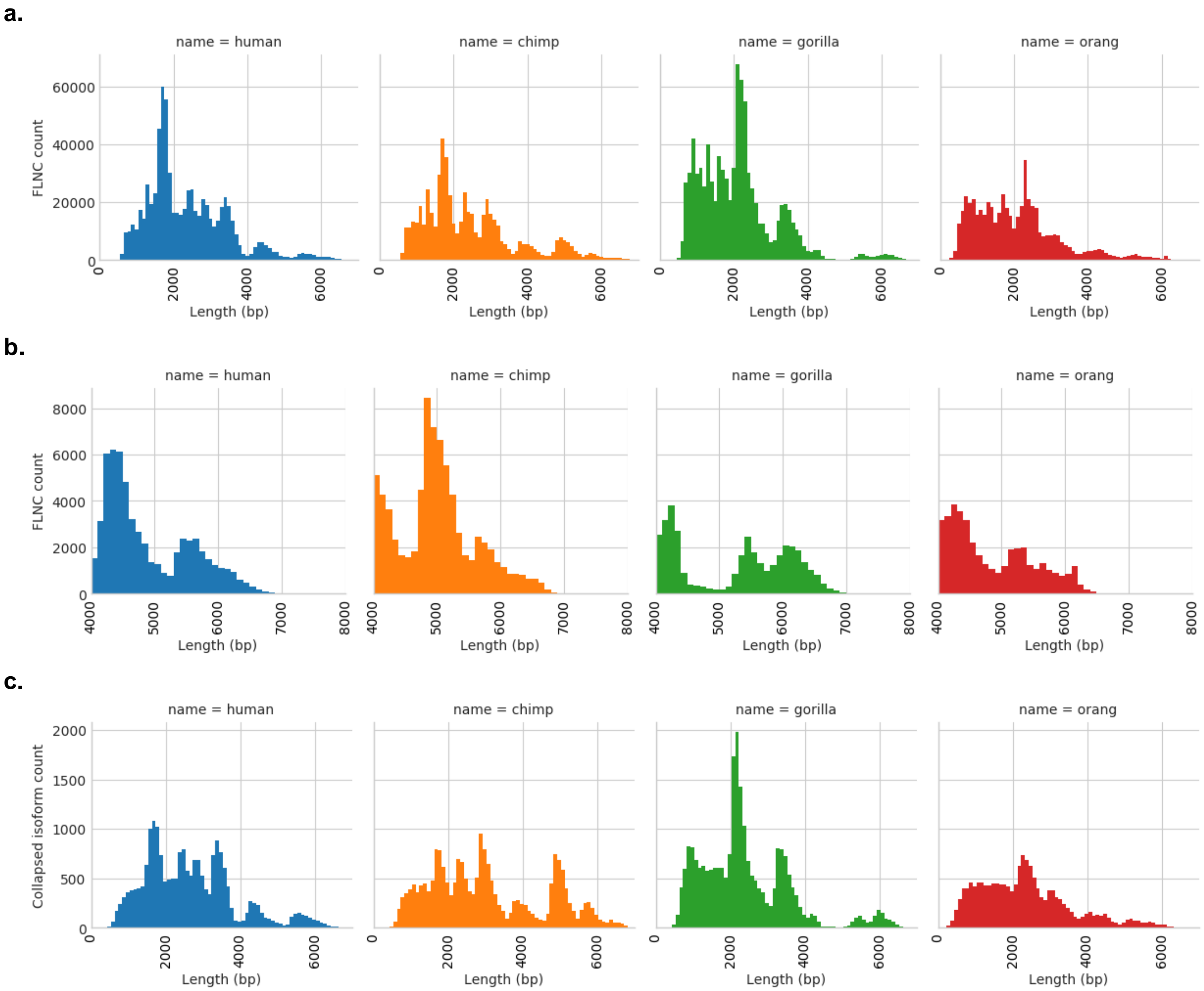


Fig. S39.

Size distributions of FLNCs.
(a) Total and (b) reads above 4 kbp. Size distributions of polished isoforms from ICE are shown in (c).

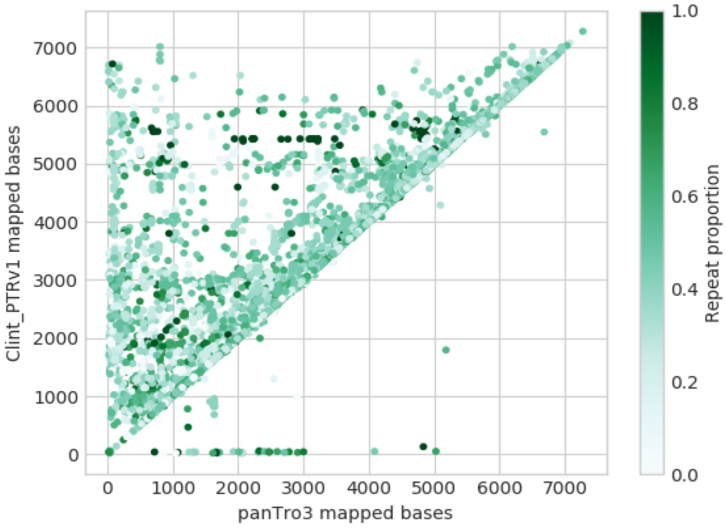


Fig. S40.

Improved transcript mapping in chimpanzee.
The number of matching bases for each chimpanzee transcript when aligned to Clint_PTRv1 or panTro3. 8,607/67,271 transcripts mapped better to Clint_PTRv1 (using a minimum difference of 5 bp), whereas 1,887 mapped better to panTro3, with a net gain of 4,805,356 Mbp (71.43 bp/transcript).

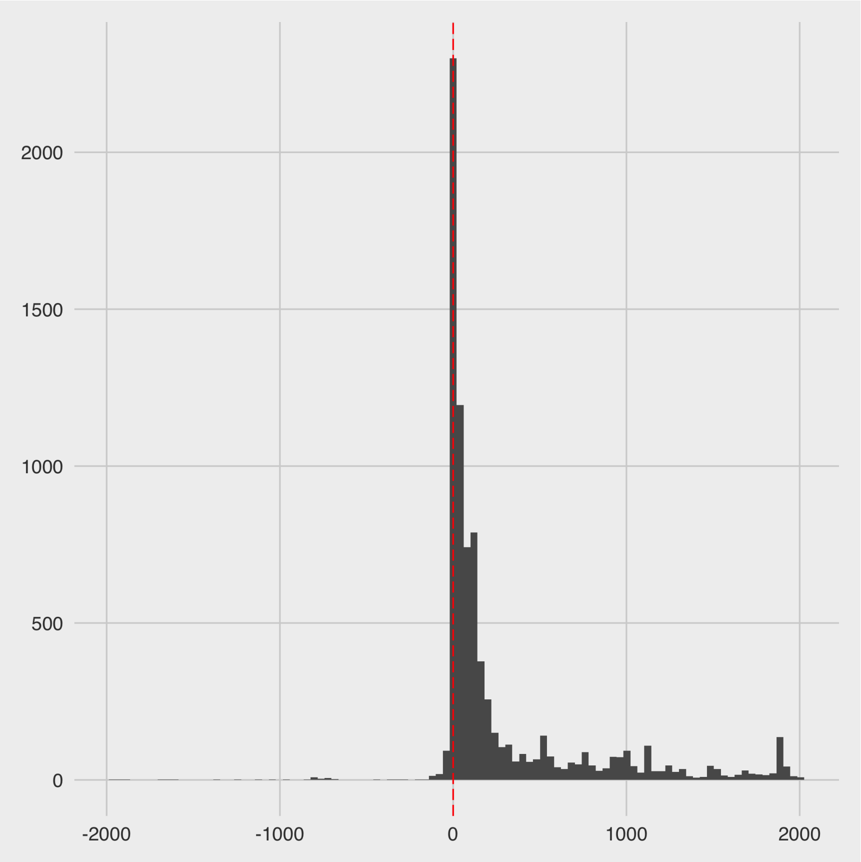


Fig. S41.

The difference in bases mapped between Clint_PTRv1 and panTro3 for each ICE transcript. For visualization purposes, zero values and absolute values greater than 2,000 were excluded.

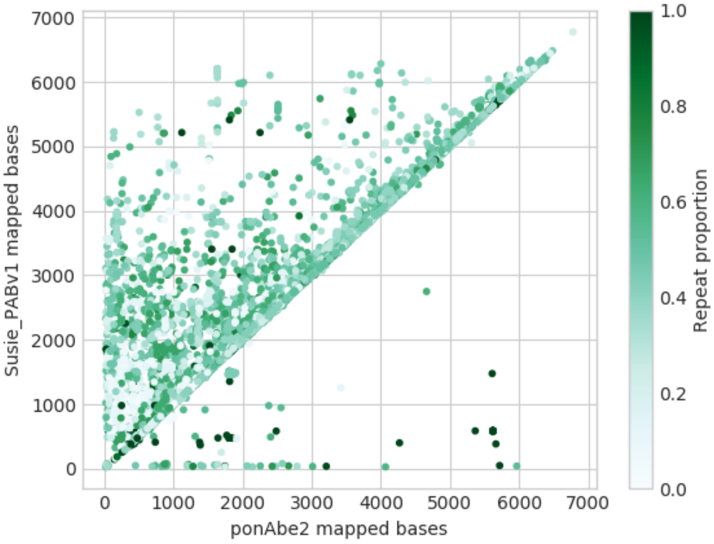


Fig. S42.

Improved transcript mapping in orangutan.
The number of matching bases for each transcript when aligned to Susie_PABv1 and ponAbe2. 10,711/55,012 transcripts mapped better to Susie_PABv1 (using a minimum difference of 5 bp), whereas only 943 mapped better to ponAbe2, with a net gain of 5,135,132 Mbp (93.35 bp/transcript).

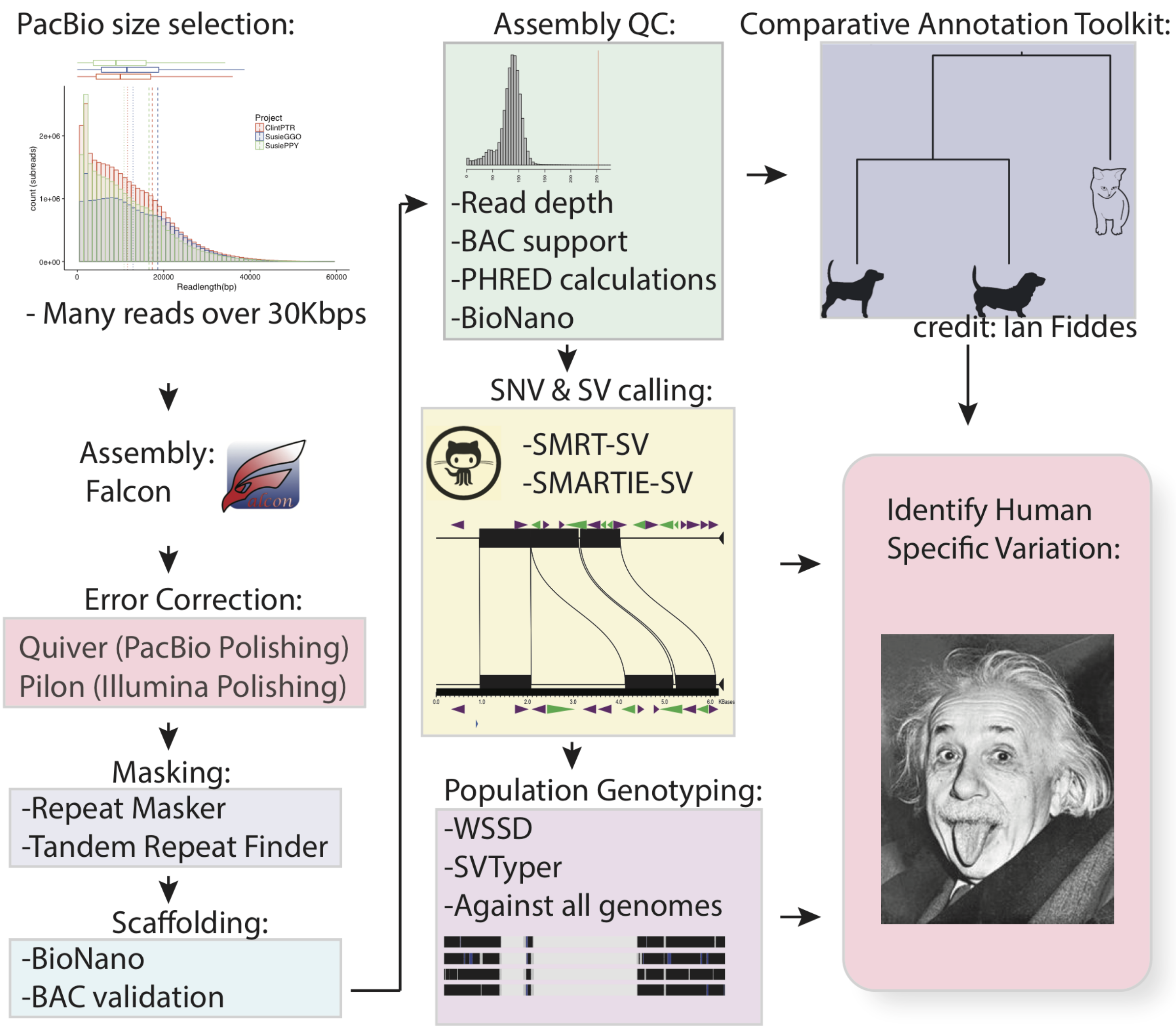


Fig. S43.

Strategy for genome assembly and identifying human-specific variants.
For structural variation detection, we used smartie-sv followed by genotyping with independent methods, read-depth CNV genotyping and paired end genotyping.

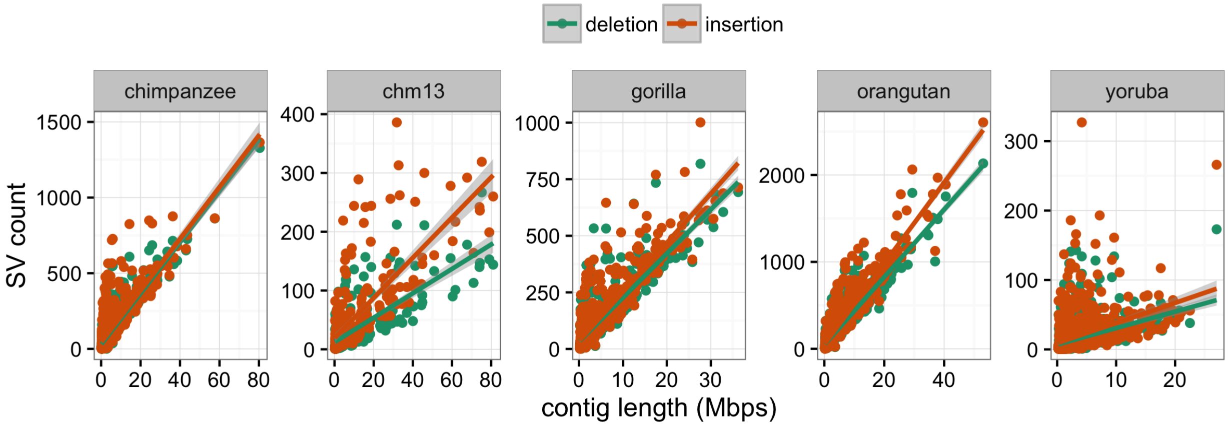


Fig. S44.

Correlation between contig length and the number of SVs on each contig.

The x-axis is contig length and the y-axis is the number of deletions or insertions. A linear model was fit for each SV call set using ggplot2.

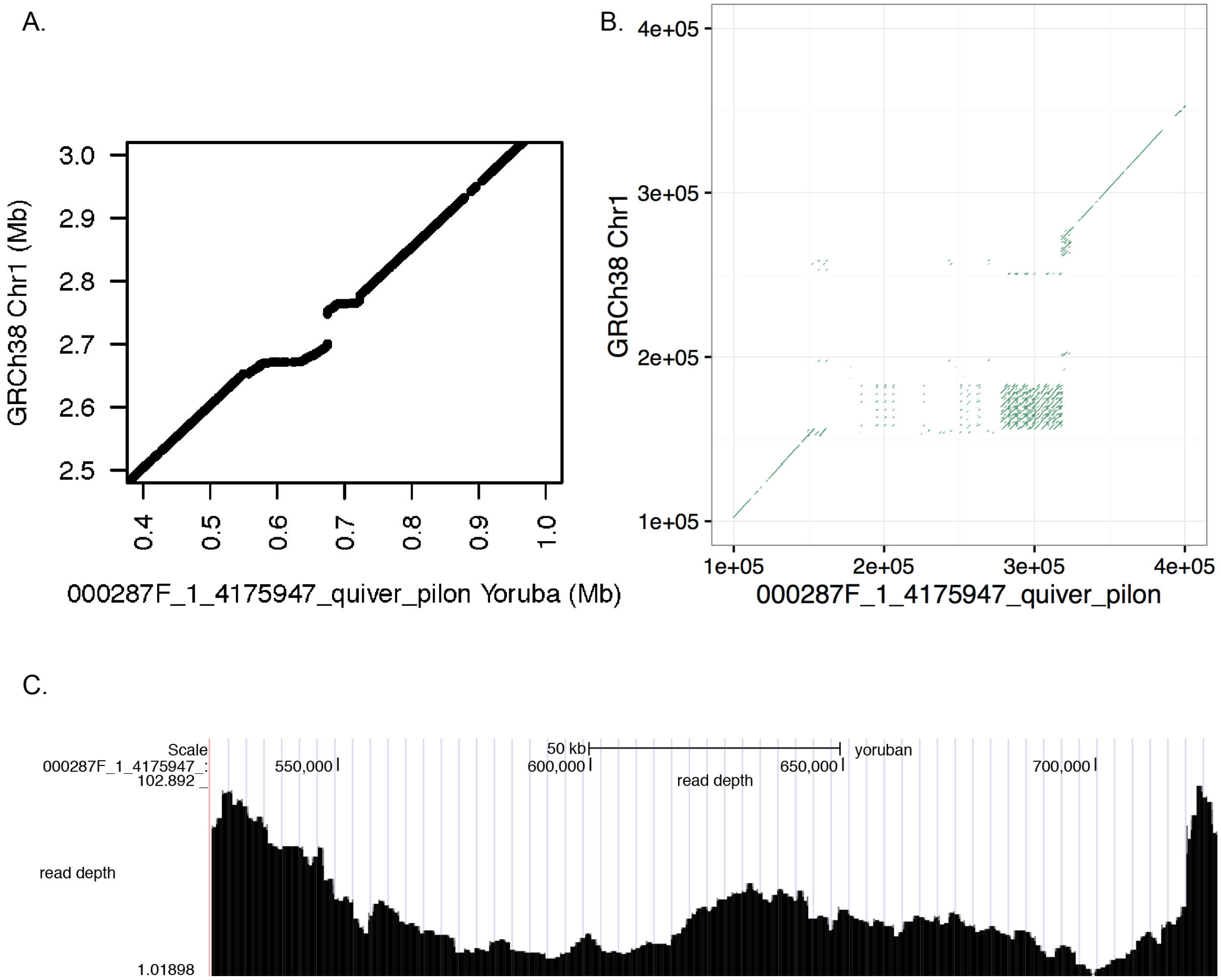


Fig. S45.

A Yoruban SV spanning a GRCh38 gap.

(A) Visual representation of the BLASR alignments. (B) Dotplot of the region spanning the gap in GRCh38. The coordinates are relative to the sub-sequence offset, not the genomic position. (C) Depth in the YRI_HSAv1 assembly across the region of interest. The depth dips in the regions compared to the genomic average ~100X.

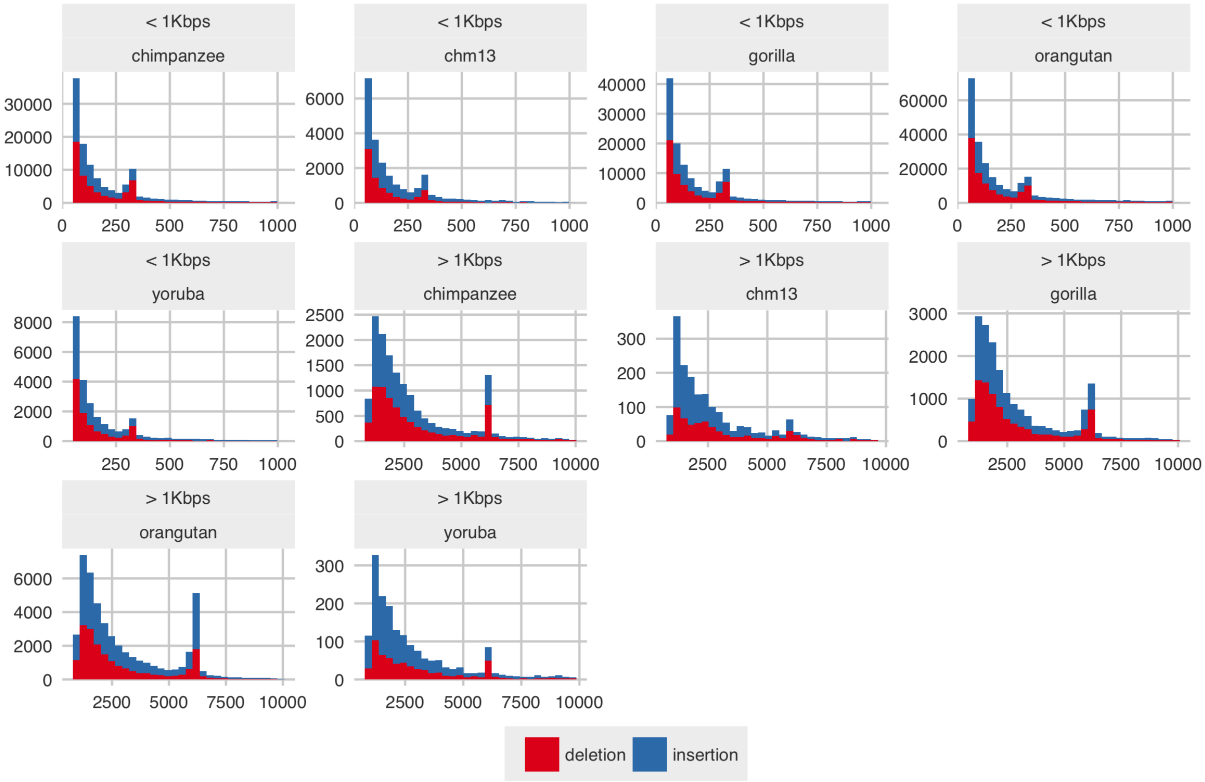


Fig. S46.

Histograms of SV sizes.
Length in base pairs is on x-axis; count is on y-axis. SVs >10 kbp were excluded for visualization purposes.

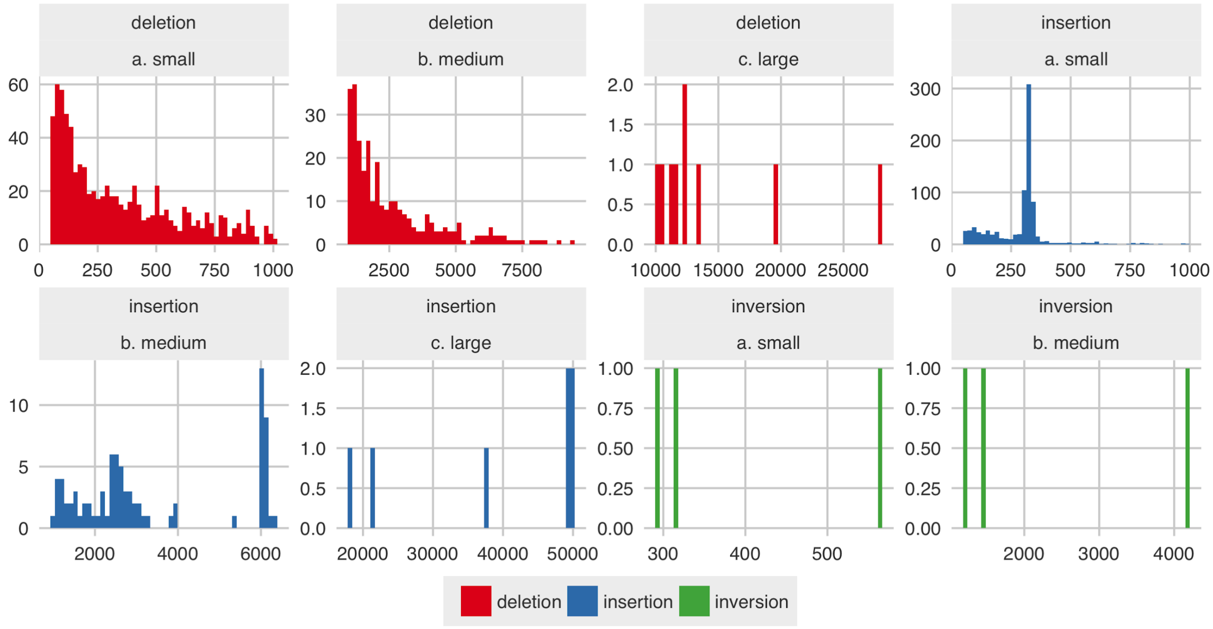


Fig. S47.

The size distribution for GRCh38-specific SVs.
Length in base pairs is on x-axis; count is on y-axis. These are SVs shared amongst the great ape assemblies. Insertions and deletions have been flipped to represent the nature of the SV in GRCh38 relative to the great ape assembles. The spikes at 300 bp and 6 kbp correspond to Alu and LINE elements, respectively. The enrichment at 50 kbp corresponds to SV calls spanning GRCh38 gaps.

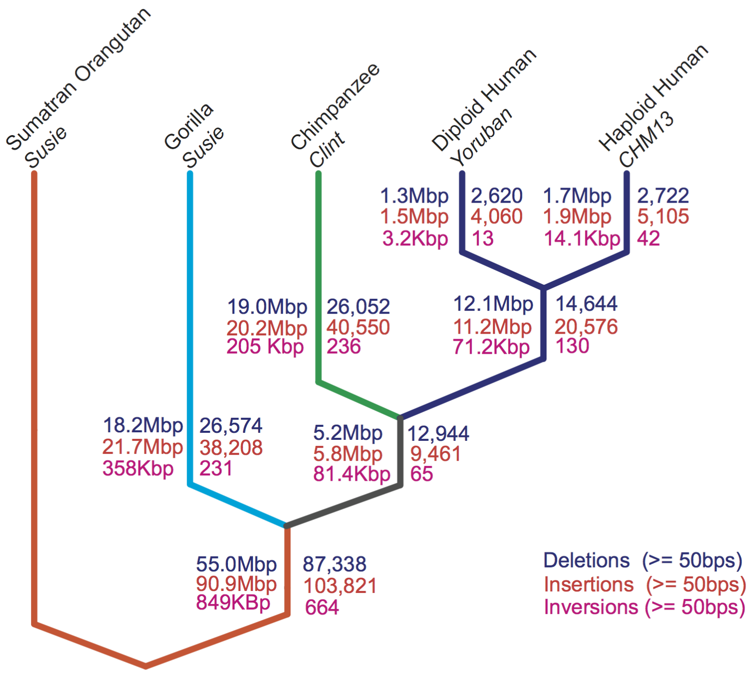


Fig. S48.

Insertion and deletion counts by comparison to assembly, along the great ape cladogram. Deletions are shown in blue and insertions are shown in red. The number of bases affected and the number of events can be found to the left and right of the branches, respectively. Shared SVs were calculated by 50% reciprocal overlap. An outgroup is required to unambiguously assign orangutan-specific variants vs. human-chimpanzee-gorilla shared events.

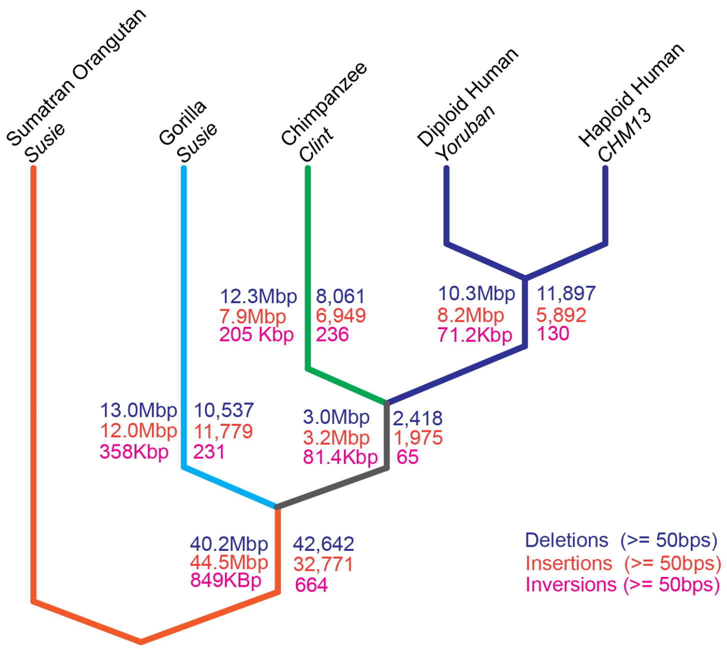


Fig. S49.

Insertion and deletion counts by genotyping, along the great ape cladogram.
Deletions are shown in blue and insertions in red. The number of bases affected and the number of events can be found to the left and right of the branches, respectively. Shared SVs were calculated by 50% reciprocal overlap. Events were determined to be lineage specific if they have either a high VST or FST score (0.8) determined from SVTyper and WSSD. Inversions were not genotyped; therefore, they were assigned by comparing the assemblies. An outgroup is required to unambiguously assign orangutan-specific variants vs. human-chimpanzee-gorilla shared events.

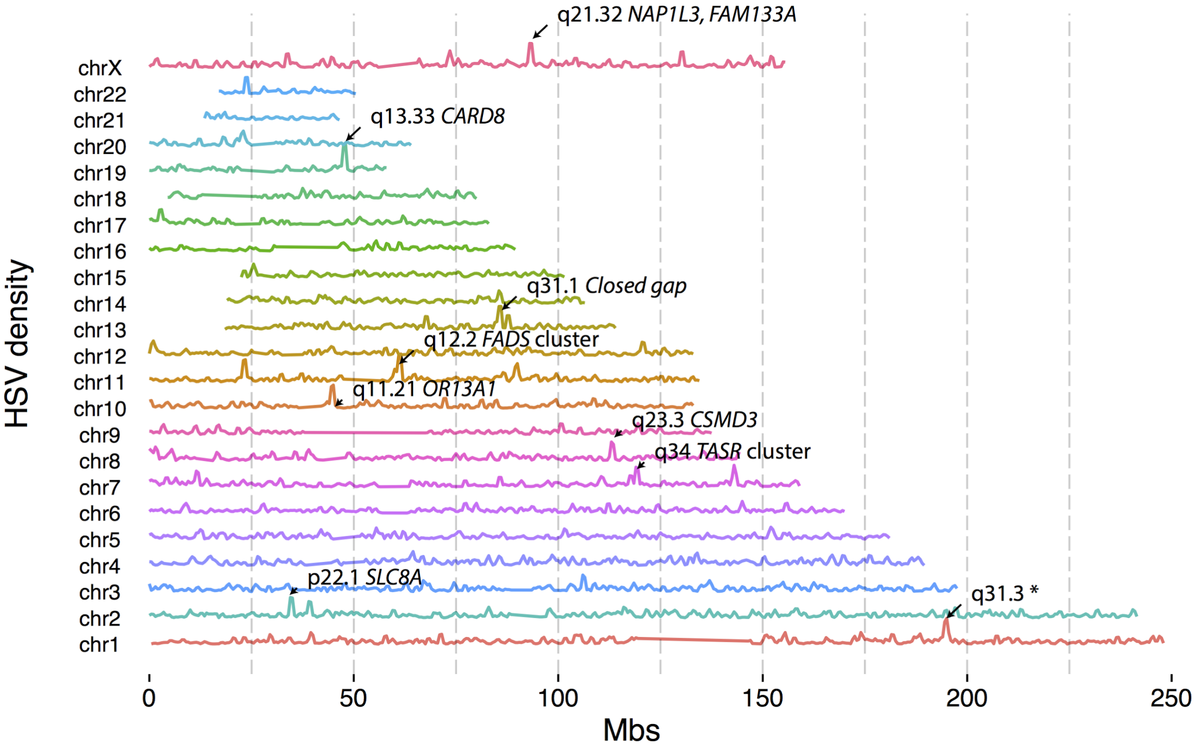


Fig. S50.

Distribution map of human-specific structural variation.
The y-axis is the number of bases contained within a human-specific SV (insertion or deletion). The number of human-specific bases was calculated in a 1 Mbp window with a 250 kbp set.

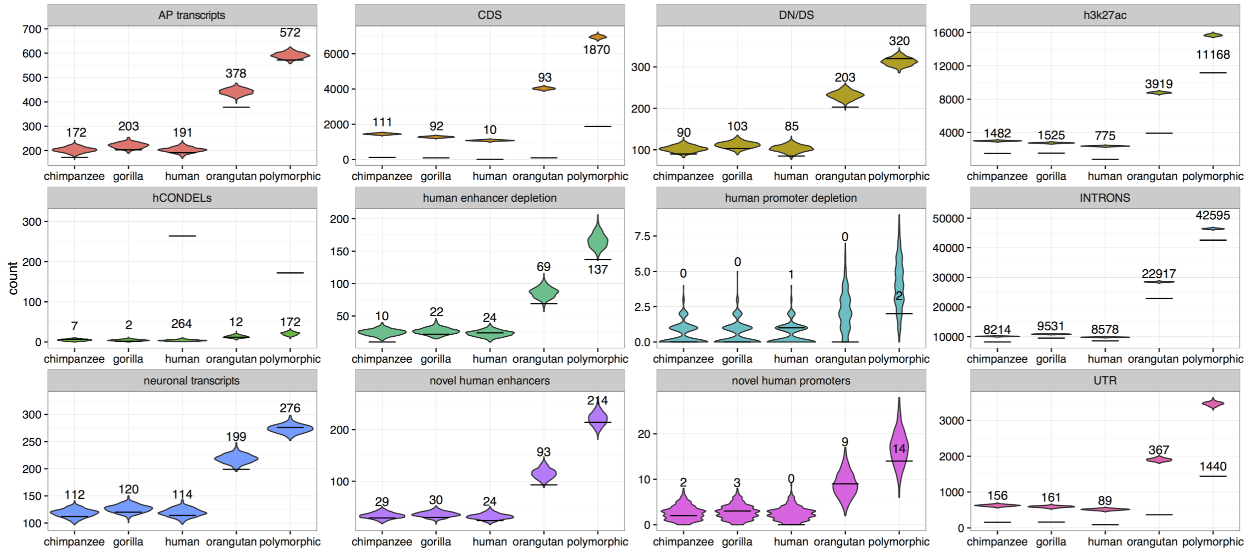


Fig. S51.

SV overlap with GRCh38 annotations.

The number of SVs overlapping each feature class is shown as a horizontal bar; the violins represent the expected feature overlap based on 500 permutations. AP transcripts are Apical Progenitors transcripts that are differentially expressed (up or down) between chimpanzee and human brain organoids (*56*). CDS, introns, and UTR features are from RefSeq GRCh38. DN/DS transcripts are coding sequences showing elevated levels of nonsynonymous to synonymous substitutions, consistent with selection (*102*). hCONDELs are conserved sequences that are not present in the human lineage(*5*). Novel human promoters and enhancers, as well as depleted human promoters and enhancers, were annotated in primate brains using immunoprecipitation and chipSeq (*56*). Almost all annotations show a significant depletion in SV overlap.

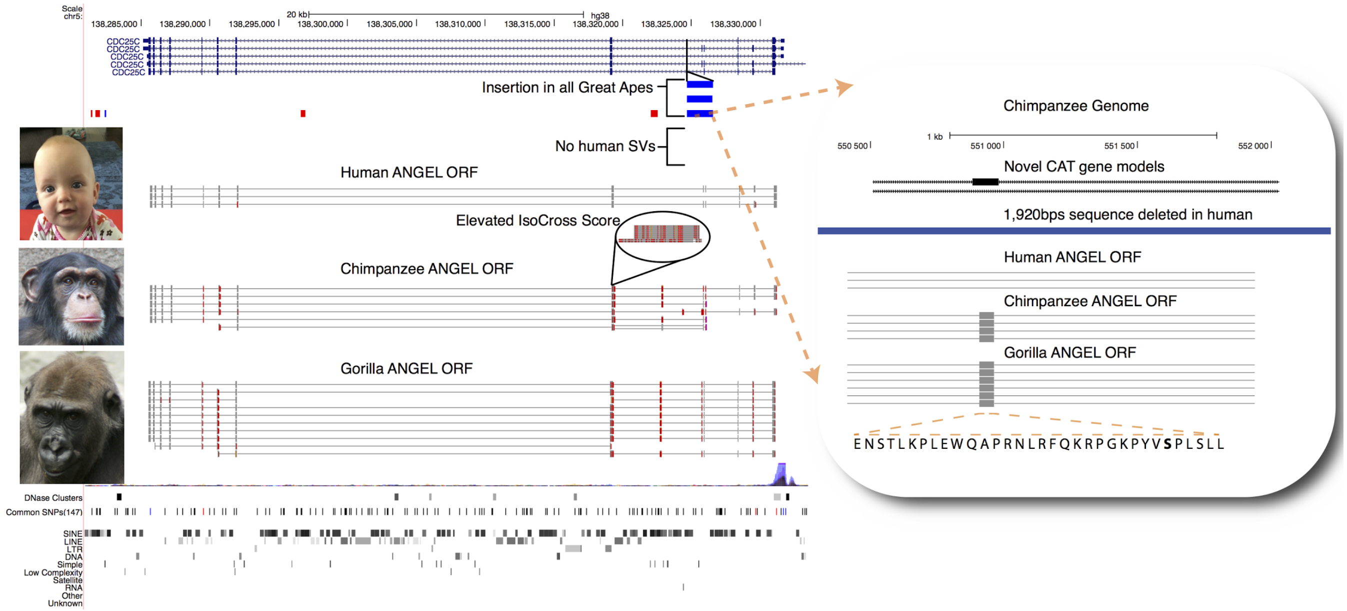


Fig. S52.

Fixed human-specific deletion (fhDEL) of a CDC25C exon.

The blue line indicates an insertion that is shared in all the NHPs relative to GRCh38, i.e., an fhDEL. The CDC25C RefSeq transcript models, ANGEL open reading frame (ORF) predictions, and repeat content are shown in the left panel. The right panel shows the lost exon in the Clint_PTRv1 genome. The 33 amino acid sequence lost has several phosphorylation sites.

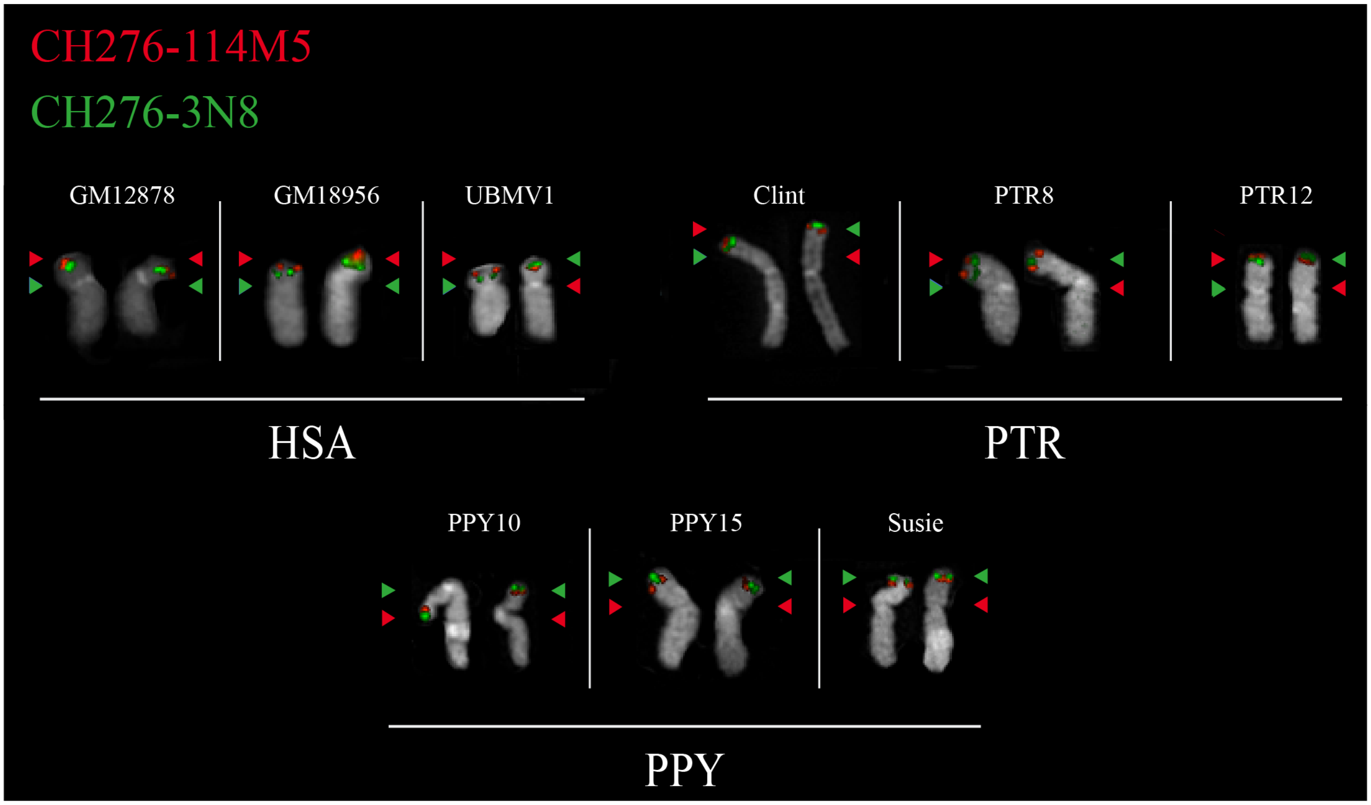


Fig. S53.

Chromosome inversions, originally detected by optical mapping and BAC end sequencing, confirmed by metaphase analysis and interphase FISH experiments.

A 4.6 Mbp inversion on 4p16 is analyzed among 3 humans (HSA), 3 chimpanzees (PTR) and 3 orangutans (Pongo) by high-resolution FISH. The inversion was identified in 1/3 human individuals (UBMV1) by clones CH276-114M5 in red and CH276-3N8 in green. Among chimpanzee, all three individuals are heterozygous for the inversion while all three orangutans are homozygously inverted. The inverted orientation is syntenic to the orientation found in mouse suggesting it is likely the ancestral state.

Table S1.

Comparative genome annotation analysis.
Table S1 provided as separate Excel file.

Table S1.1. Assessing assembly quality using human CCDS. CCDS exons for Homo sapiens were downloaded from UCSC table browser (CCDS database release 20, 09/08/2016). Alignments of exons to the assemblies are used to assess the quality and degree of completeness of the assembly. ‘Full’, ‘partial’ and ‘missing’ are defined by the percentage of aligned bases (see supplementary section VI (Q)). An exon is missing when the aligned length is less than 10%. In S1.1, we compare Clint_PTRv1 and Susie_PABv1 with the current reference for chimpanzee and orangutan. ‘Numbers’, ‘percentages’, are counts and percentage of counts respectively. Values in ‘genes’ and ‘CCDS’ tables are extrapolated from the exon table.

Table S1.2-S1.3. Candidate novel and species-specific exons. A) We identified candidate novel unannotated exons in our gene annotation set, finding 9, 29, 16, and 16 exons in human, chimpanzee, gorilla, and orangutan, respectively. Shown are coordinates in human space (GRCh38) and with respect to the relevant great ape assembly. N/A indicates that lack of a homologous locus. Also shown is the number of supporting PacBio reads for the exon from each species, gene identifiers, and column indicating whether the exon is primarily derived from a common repeat. For human novel exons, the isoforms that contained them were evaluated for their expression pattern with use of the GTEx dataset. B) We identify candidate exons gained or lost specifically between humans and other great apes that have the potential to impact coding sequence. Exons of lengths a multiple of 3 and with more than 5 supporting PacBio reads in human but with no overlap from other great apes were selected to identify human exon gains, and those with 5 supporting reads in chimpanzee but no overlap in human were selected to identify human exon losses. Coordinates are shown in human and in Clint space. Exons were removed if any short-read RNA-seq data disputed the exon’s absence in chimpanzee or human, respectively, or if there was evidence of mapping error. Note that this method is also sensitive to expression differences between the cells assayed. The dearth of exons lost (13) vs. exons gained (57) likely reflects a combination of more stringent filtering enabled by the more completely annotated GRCh38 as well as the higher depth of short-read RNA-seq obtained from human vs. chimpanzee iPSCs.

Table S1.4-S1.6. Candidate novel loci and isoforms. The exact positions of the novel genes and transcripts are reported for each NHP. Novel genes are classified as either putatively novel or possibly paralogous. Putatively novel loci are predicted loci that do not overlap any transMap projection of transcripts from human, while possibly paralogous loci are those that only overlap transMap projections that were filtered out during paralog resolution and represent candidates for gene family expansion. For more detail, see the CAT manuscript (*22*). Novel isoforms have a comma-separated list of the coordinates of the splice junctions present in the transcript, which are supported by RNA-seq or Iso-Seq and not supported by liftover of GENCODE V27. All of the genes and transcripts in this table correspond to the numbers reported in Table 2.

Table S2.

SNV density of each primate against GRCh38. SNV density was calculated in a one megabase-sliding window without overlap. SNVs were called from the contig to genome alignments.

Table S2 provided as separate Excel file.

Table S3.

Count of identified, merged, and sites with unique flanking regions in each genome. The counts of STR sequences used to find orthologous STR sequences between genomes. Raw STR Counts: STR sequences identified by RepeatMasker and Tandem Repeats Finder; Merged STRs: the number of sequences after merging tandem repeat sequences within 25 bp; <5% repeat flank: the number of STR sequences with 250 bp flanking sequences passing a filter requiring <50% repeat.

Table S3 provided as separate Excel file.

Table S4.

Orthologous STR sequences between genomes. The filtered 250 bp sequences flanking STRs were mapped between genomes and further filtered if the mapping QV of either flank was less than 30, if the sequence between the two aligned flanks was less than 80% tandem repeat, or if the minimum STR length was under 40 bp in either species.

Table S4 provided as separate Excel file.

Table S5.

Significance of STR length distributions. For each pair of genomes, the distribution of lengths of STR expansions were computed in each genome, and the two distributions were compared using the two-sided Kolmogorov-Smirnov test.

Table S5 provided as separate Excel file.

Table S6.

Human STR expansions and contractions. Human-specific insertions and deletions were determined by 50% reciprocal overlap in GRCh38 space. NHP insertions were given length for overlap purposes (start + SV length). STR-associated SVs were required to have at least 80% of bases annotated as an STR or VNTR in either GRCh38 or Clint_PTRv1. Annotations in the table include the percent/number of bases that are STR/VNTR, the closest exon in GRCh38 space, and the distance to the closest exon in GRCh38 space. The discrepancy between the number of expansions and contractions can be explained by lower overlap of NHP insertions (relative to GRCh38) in GRCh38 space.

Table S6 provided as separate Excel file.

Table S6.1. Human-specific STR expansions (4,921). These events were determined by SMARTIE-SV and Repeat Masker annotations. The proportion of STR and VNTR intersected with the SV call is shown.

Table S6.2. Human-specific STR contractions (1,465). These events were determined by SMARTIE-SV and Repeat Masker annotations. The proportion of STR and VNTR intersected with the SV call is shown.

Table S6.3. Human-specific STRs that are also expanded in both CHM13_HSAv1 and YRI_HSAv1.

Table S7.

Summary of chimpanzee, gorilla, and bonobo PtERV1 insertions.
Table S7 provided as separate Excel file.

Table S8.

SV validation BAC sequencing. SV calls were independently validated by SMRT sequencing of large-insert clones. Columns 1-3 list coordinates of the SV with respect to the GRCh38 reference assembly. Column 4 depicts the nearest gene annotation for the SV. Column 5 lists the specific primate the SV event was detected in. Column 6 lists the BAC clone used for validation. The type of SV (insertion, deletion) and SV size (in bp) are listed in columns 7 and 8. Where possible, lineage specificity (column 9) is determined by copy number heatmaps annotated by NHP WSSD or a 5-way comparison of PacBio assembled contigs.
Table S8 provided as separate Excel file.

Table S9.

Fixed human-specific deletions (fhDELs; Table S9.1) and insertions (fhINSs; Table S9.2). These events are based on two criteria. First, they must be fixed based on the genome-genome alignment comparisons. Second, they must be genotype fixed (VST > 0.8 or FST > 0.8). For additional information on the contents of the table, see the master SV VCF call set and supplemental methods.

Table S9 provided as separate Excel file.

Table S10.

Human-specific SVs of potential functional impact. These data are the Ensembl Variant Effect Predictor (VEP) annotations, limited to the "HIGH" impact classification. For a full description of column reference, visit the VEP webpage.

Table S10 provided as separate Excel file.

Table S11.

Human deletion of conserved regions (hCONDEL) analysis. These contain the novel hCONDELs discovered in this study, the overlap with previous hCONDELs (*5*), and the discrepancies between the two datasets.

Table S11 provided as separate Excel file.

Table S11.1. List of human-specific deletions (7,400). Annotations include mouse conservation, overlap with previous hCONDELs (based on liftover and mapping between panTro2 and Clint_PTRv1), and novelty status.

Table S11.2. Summary of hCONDEL overlap split by the original hCONDEL categories. This list was manually curated and not directly correlated with Table S10.

Table S11.3. A list of 156 previous hCONDELs not found in the current dataset. This comparison was done by liftover and intersection with our 5,892 fixed human-specific SVs. These include events we found in either CHM13_HSAv1 or YRI_HSAv1 (polymorphic in humans), genotyped as not fixed, complex events that are not simply deletions and hCONDELs that we found to be inversions (netting and chaining errors).

Table S12.

Large-scale inversions detected among great apes.

Large-scale inversions detected and validated by optical mapping, large-insert clone-based SMRT sequencing and FISH are listed. Column 2 lists the specific primate the inversion event was detected in. GRCh38 coordinates for the inversions detected by optical mapping are listed in columns 5-7. Column 8 includes the estimated size of inversion events listed in kbp. SD-mediated events are designated in column 9 based on WGAC or NHP WSSD annotations in the GRCh38 reference assembly. Validation of inversion events either by SMRT sequencing of large-insert clones (column 11) or by FISH (column 12) are listed. Human polymorphic inversions using data published previously (*49*, *50*) is included in columns 13 and 14. Citations for previously published inversion events among NHPs are included in column 15. Concordance between human GRCh38 and the mouse reference GRCm38 with respect to orientation is listed in column 16. Lineage specificity is inferred based on a combination of Strand-seq data, BES discordancy and FISH are included in column 17. Lineage-specific duplications identified at the boundaries of the inversion events are inferred through copy number heatmaps annotated by NHP WSSD (columns 18 and 19). Annotation and identification of core duplicons identified previously (*104*) are also included in columns 20 and 21.

Table S12 provided as separate Excel file.

Table S12.1. Large-scale inversions.

Table S12.2. Smartie-sv inversion calls.

Table S12.3. Bionano automated inversion calls.

Table S13.

Expression data.

Differential expression analysis was performed in excitatory neurons and radial glia from human and chimpanzee organoids. p-values are from ‘bimod’ likelihood ratio differential expression test in Seurat. The average difference is log(mean(A))-log(mean(B)) where A is mean of gene.i expression (non-log scale) of human, and B is mean of gene.i expression (non-log scale) of chimpanzee. pct.human and pct.chimp refer to the percent of cells in human or chimpanzee expressing over 0 cpm of a given gene. SV associated denotes whether a gene is within 50 kbp of SVs. Sig denotes whether the gene is significantly differentially expressed (considered significant if adjusted p-value is <0.05 and avg.diff >0.2 where adjusted p-value is Bonferroni corrected. Direction denotes whether the gene was expressed higher or lower in human when compared to chimpanzee expression. The permutation overlap tab (S13.1) describes the SVs assayed in the permutation test for overlap with differentially expressed genes, including the coordinates of the 100 kbp window around an SV, additional information about the type of SV can be found in the info column. The genes assayed in the permutation tab (S13.2) show the GRCh38 coordinates and genes used for the permutation test. The duplicated regions tab (WSSD; S13.3) shows the GRCh38 coordinates for examined for the permutation test.

Table S13 provided as separate Excel file.

Table S14.

Genes-containing noteworthy SVs.

The HUGO gene symbols listed in the first column. The Keg Brite gene description is in the second column, when available, otherwise manually inferred. The SV allele is listed in the third column. Human-specific events are polarized for derived or ancestral, but events in other lineages are relative to GRCh38. Human-specific events are noted with an ‘H’, chimpanzee with a ‘C’, gorilla with a ‘G’, and orangutan with an ‘O’; SVs assigned to a lineage without genotyping are denoted with a ‘U’. VEP functional annotations are in the 5th column, based on hg38 and Clint_PTRv1 annotation datasets. The 6th column contains the single-cell chimpanzee-human differential expression of organoid excitatory neurons (EN) and radial glia (RG) cells. Human upregulated genes, relative to chimpanzee, are marked with “+”, “-” for downregulated and “NC” for no change. Genes with statistically significant fold change are bolded and the p-value is listed.

Table S14 provided as separate Excel file.

Table S15.

Accessions and resources.

Sequenced clones (Table S15.1) and genome assembly and transcriptome-related accessions (Table S15.2).

Table S15 provided as separate Excel file.

Table S16.

Statistics for the PacBio sequencing of chimpanzee (Clint).

Title	Statistic
SMRT cells	283
SMRT cells (>15 kbp)	271
SMRT cells (>30 kbp)	12
Total data (ROI) (Mbp)	318,147
Total coverage (ROI) (X)	99
Total data (subread) (Mbp)	374,503
Total coverage (subread) (X)	117
ROI reads	24,180,297
High-quality subreads	32,804,539
Mean subread length (bp)	11,416
Median subread length (bp)	9,805
Subread N50 (bp)	17,036
Coverage in subreads >20 kbp (X)	43.3

Table S17.

Clint_PTRv1 assembly statistics.
Statistics after applying Quiver and Pilon but before FreeBayes-based indel correction and Bionano contig breaking.

Feature	Statistic
Number of contigs	4,912
Number of contigs not in scaffolds	4,912
Total size of contigs	2,992,670,130 bp
Longest contig	80,428,132 bp (chr6q)
Shortest contig	6 bp
Bases in Contigs > 1 kbp	2,992,624,476 (100.0%)
Bases in Contigs > 10 kbp	2,988,642,396 (99.9%)
Bases in Contigs > 100 kbp	2,867,621,436 (95.8%)
Bases in Contigs > 1 Mbp	2,713,712,200 (90.7%)
Bases in Contigs > 10 Mbp	1,724,914,903 (57.6%)
Mean contig size	609,257 bp
Median contig size	34,261 bp
N50 contig length	12,759,992 bp
L50 contig count	64 bp
contig %A	29.65
contig %C	20.36
contig %G	20.35
contig %T	29.64

Table S18.

Clint_PTRv1 contig assembly concordance based on BES and FES mappings.

Assembly feature	Statistic
Total bases assessed for concordance*	2,798,141,019
Bases spanned by concordant best*	2,773,922,784
Bases spanned by discordant best and not concordant best*	3,324,228
Bases spanned by discordant best, tied, concordant tied and not concordant best	24,218,235
Bases spanned by both concordant best and discordant best	975,288,899
Sanger BES accuracy	99.962%
Proportion of bases spanned by concordant best	99.13%

*Contigs greater than 300 kbp.

Table S19.

Clint_PTRv1 accuracy statistics.

Base-pair accuracy of the assemblies based on BES alignment from the Clint BAC library (CHORI-251). Total variations in the table are defined as the sum of all transitions, transversions, deletions and insertions.

BES/FES Comparison	**Quiver**	**Quiver + Pilon**
Transitions	9,279	10,663
Transversions	4,940	5,552
Deletions (1 bp)	14,665	9,176
Deletions (>1 bp)	5,139	4,580
Insertions (1 bp)	2,783	2,627
Insertions (>1 bp)	2,817	2,845
Total variations	39,623	35,443
High-quality bases (PHRED ≥ 40)	47,377,472	47,259,661
Expected Sanger errors	17,530	17,486
PacBio - Sanger errors	22,093	17,957
Accuracy	0.999534	0.999620
Estimated QV (Quality value)*	33.31	34.20
Estimated QV w/o indels	35.23	34.65
Ti/Tv	1.88	1.92

*Quality Value (QV) is the total probability that the base call is an insertion or substitution or is preceded by a deletion. QV = -10 * log10(p) where p is the error probability.

Table S20.

Clint_PTRv1 assembly comparisons with previous versions.

These numbers do not include N's so inaccuracies in the number of N's have no effect. For example, if the old assembly had 100 N's, and our assembly replaced those 100 N's by 150 bases, 150 bases would be added to the "Bases added to previous assembly" number.

Previous assembly	SMRT Genome Assembly	Bases added to previous assembly (Mbp)	Bases removed from previous assembly (Mbp)
panTro5.1	Clint_PTRv1	3.5	33.3
panTro5.0	Clint_PTRv1	6.9	27.2
panTro3 (Oct 2010)	Clint_PTRv1	45.1	2.6
ponAbe2	Susie_PABv1	54.5	3.8

Table S21.

Scaffolding of Clint_PTRv1 by Bionano Saphyr.
Two nicking restriction enzymes, Nt.BspQI and Nb.BssSI, were used to scaffold Clint_PTRv1 into 121 scaffolds.

	Info	Bionano	Seq	Seq in Hybrid A	Hybrid A	HybridA + not scaffolded seq	fold change*
One enzyme Nt. Nt.BspQI	Count	3135	4912	624	190	4630	3.0x
	N50 (Mbp)	3.5	12.91	13.77	42.04	38.85	
	Total length (Mbp)	5818.21	2992.67	2805.73 (93.75%)	2827.75	3014.69	
	Info	Bionano	Seq	Seq in Hybrid B	Hybrid B	HybridA + not scaffolded seq	fold change*
One enzyme Nb.BssSI	Count	4087	4912	641	226	4649	2.8x
	N50 (Mbp)	2.47	12.91	13.77	38.34	36.58	
	Total length (Mbp)	5907.18	2992.67	2807.64 (93.82%)	2839.06	3024.09	
	Info	All Hybrid***	Seq in All Hybrid ***	All Hybrid + not scaffolded seq	fold change**		
Two enzyme	Count	121	737	4448	4.5x		
	N50 (Mbp)	59.55	13.77	57.99			
	Total length (Mbp)	2801.45	2768.37 (98.82%)	3025.75			
68 sequences were cut during chimeric detection							
bp adjusted							
* Fold change represents the N50 increase between single-color hybrid + not scaffolded sequence relative to the original sequence							
** Fold change represents the N50 increase between All Hybrid + not scaffolded sequence relative to the original sequence							
*** All Hybrid includes merged Hybrid Scaffold A+B & leftover Hybrid Scaffold A & leftover Hybrid Scaffold B							

Table S22.

BAC clones with FISH mappings used for building AGP.

Chromosome	#BAC	Chromosome	#BAC
chr1	55	chr12	51
chr2	47	chr13	33
chr3	64	chr14	25
chr4	51	chr15	32
chr5	50	chr16	18
chr6	44	chr17	44
chr7	43	chr18	8
chr8	46	chr19	20
chr9	43	chr20	20
chr10	35	chr21	14
chr11	24	chr22	14
chrX	32		

Table S23.

Clint_PTRv1 missing genes.

57 protein-coding gene annotations are missing in the chromosomes but present in 31 unplaced contigs.

001351F_1_72220_quiver_pilon	*RHD*
001729F_1_54194_quiver_pilon	*IL3RA*
001913F_1_47401_quiver_pilon	*CES4A*
002272F_1_37823_quiver_pilon	*REP15*
002629F_1_31565_quiver_pilon	*AGAP9*
000151F_1_35110_quiver_pilon	*ANXA2R*
001556F_1_61810_quiver_pilon	*CT45A9*
002473F_1_34352_quiver_pilon	*FAM25C*
004031F_1_9760_quiver_pilon	*USP17L3*
004613F_1_2389_quiver_pilon	*FAM106A*
001881F_1_48643_quiver_pilon	*ZNF286B*
002841F_1_28160_quiver_pilon	*USP17L4*
001321F_1_64302_quiver_pilon	*CATSPER2*
000426F_1_722043_quiver_pilon	*MTRNR2L1*
002329F_1_35279_quiver_pilon	*SPATA31A3*
002873F_1_27568_quiver_pilon	*AL356585.2*
003442F_1_17755_quiver_pilon	*ACAP3,PUSL1*
001574F_1_60847_quiver_pilon	*ZNF157,ZNF41*
001150F_1_85703_quiver_pilon	*HIC2,TMEM191B*
001577F_1_60724_quiver_pilon	*RP11-812E19.9*
002375F_1_35925_quiver_pilon	*MT-CO1,MT-ND2*
003308F_1_19948_quiver_pilon	*RP11-435I10.4*
001375F_1_70920_quiver_pilon	*GJB1,NONO,ZMYM3*
001731F_1_51205_quiver_pilon	*PRAMEF7,PRAMEF8*
002770F_1_29229_quiver_pilon	*FAM231A,FAM231C*
001774F_1_52584_quiver_pilon	*BAGE5,CU104787.1*
002197F_1_39570_quiver_pilon	*HNRNPCL1,PRAMEF2*
000492F_1_430388_quiver_pilon	*LAT,NFATC2IP,SPNS1*
000807F_1_136766_quiver_pilon	*AC009133.22,QPRT,SPN*
001926F_1_47102_quiver_pilon	*HNRNPCL2,HNRNPCL3,HNRNPCL4*
002554F_1_195764_quiver_pilon	*ATP2A1,ATXN2L,CD19,RABEP2,SH2B1,TUFM*
001355F_1_72616_quiver_pilon	*USP17L1,USP17L11,USP17L12,USP17L2,USP17L30*

Table S24.

Interchromosomal translocation errors.

Erroneous contig found	Contig rightful placement	Contig size (Mbp)
chr1	chr5	4.1
chr6	chr5	7.4
chr6	chr5	5.4
chr5	chr16	1.8

Table S25.

Nine pericentric inversions (hg38) seen in Clint_PTRv1.

chr	start	end	size	Clint_PTRv1	UCSC panTro5
chr1	-	-	-	yes	yes - partial
chr4	44,509,907	86,039,028	41,529,121	yes	yes
chr5	23,056,186	93,288,262	70,232,076	yes	yes
chr9	70,447,920	87,988,837	17,540,917	yes	no
chr12	20,826,991	66,590,630	45,763,639	yes	yes
chr15	22,905,050	27,830,650	4,925,600	yes	no
chr16	-	-	-	yes	yes
chr17	15,367,740	44,918,039	29,550,299	yes	yes
chr18	2,136,811	12,904,782	10,767,971	yes	yes

Table S26.

Pericentric inversion breakpoints captured by Bionano scaffolds.

chr	start	end	size	Scaffold spanning breakpoint (L)	(R)
chr2 fusion	113,000,000	113,000,000		na	na
chr4	44,813,133	84,898,851	40,085,718	yes	yes
chr5	23,020,320	93,926,801	70,906,481	yes	no
chr9	68,643,184	86,184,102	17,540,918	no	no
chr12	20,782,790	67,910,583	47,127,793	no	yes
chr15	25,108,810	29,751,155	4,642,345	yes	no
chr17	15,523,701	49,485,678	33,961,977	no	no
chr18	2,146,810	12,914,783	10,767,973	yes	yes

Table S27.

Statistics for the PacBio sequencing of orangutan (Susie).

Title	Statistic
SMRT cells	296
SMRT cells (>15 kbp)	243
SMRT cells (>30 kbp)	53
Total data (ROI) (Mbp)	260,142
Total coverage (ROI) (X)	81.3
Total data (subread) (Mbp)	303,825
Total coverage (subread) (X)	94.9
ROI reads	21,101,271
High-quality subreads	28,037,820
Mean subread length (bp)	10,836
Median subread length (bp)	8,953
Subread N50 (bp)	16,607
Coverage in subreads >20 kbp (X)	33.9

Table S28.

Susie_PABv1 assembly statistics.
These statistics are after applying Quiver and Pilon but not our FreeBayes-based indel correction or the Bionano contig breaking.

Title	Statistic
Number of contigs	5,771
Number of contigs not in scaffolds	5,771
Total size of contigs	3,042,567,509 bp
Longest contig	53,047,495 bp
Shortest contig	3 bp
Bases in contigs >1 kbp	3,042,527,595 (100.0%)
Bases in contigs >10 kbp	3,037,352,996 (99.8%)
Bases in contigs >100 kbp	2,898,711,478 (95.3%)
Bases in contigs >1 Mbp	2,767,083,842 (90.9%)
Bases in contigs >10 Mbp	1,640,595,012 (53.9%)
Mean contig size	527,217 bp
Median contig size	30,439 bp
N50 contig length	11,273,411 bp
L50 contig count	82
contig %A	29.60
contig %C	20.39
contig %G	20.41
contig %T	29.59

Table S29.

Susie_PABv1 BES concordance: Assembly concordance based on BES mappings.

Title	Statistic
Total bases assessed for concordance*	2,829,115,991
Bases spanned by concordant best*	2,737,256,459
Bases spanned by discordant best and not concordant best*	2,185,050
Bases spanned by discordant best, tied, concordant tied and not concordant best	4,069,414
Bases spanned by concordant best and discordant best (intersection; bases in the reference with both concordant and discordant best support)*	74,208,072
Proportion of bases spanned by concordant best	96.8%

*Contigs greater than 300 kbp.

Table S30.

Susie_PABv1 assembly contig accuracy.
Base-pair accuracy of the Quiver polished and Pilon-polished assemblies based on BES alignment from Susie BAC library (CHORI-276). Total variation is defined as the sum of all transitions, transversions, deletions and insertions.

Type	Quiver	Quiver + Pilon
Transitions	39,124	44,072
Transversions	18,357	20,096
Deletions (1 bp)	73,450	37,675
Deletions (>1 bp)	13,263	10,844
Insertions (1 bp)	6,628	5,322
Insertions (>1 bp)	6,605	6,378
Total variations	157,427	124,387
High-quality bases (PHRED ≥ 40)	60,305,413	60,175,532
Expected Sanger errors	22,313	22,265
PacBio - Sanger errors	135,114	102,122
Accuracy	0.997760	0.998303
Estimated QV	26.50	27.70
Estimated QV w/o indels	30.21	29.72
Ti/Tv	2.13	2.19

Table S31.

Scaffolding of Susie_PABv1 by Bionano Saphyr.
Two restriction enzymes, Nt.BspQI and Nb.BssSI, were used to scaffold Susie_PABv1 into 73 scaffolds resulting in a 2.86 Gbp assembly.

	Info	Bionano	Seq	Seq in hybrid A	Hybrid A	HybridA + not scaffolded seq	fold change*
One enzyme Nt. BspQI	Count	2977	5771	511	103	5412	5.1x
	N50 (Mbp)	4.16	11.28	11.80	61.53	57.77	
	Total length (Mbp)	6057.27	3042.57	2806.27 (92.23%)	2829.13	3065.43	
	Info	Bionano	Seq	Seq in Hybrid B	Hybrid B	HybridA + not scaffolded seq	fold change*
One enzyme Nb. BssSI	Count	5425	5771	513	156	5463	4.0x
	N50 (Mbp)	1.81	11.28	11.75	49.52	44.91	
	Total length (Mbp)	5873.22	3042.57	2820.19 (92.69%)	2834.50	3056.88	
	Info	All Hybrid***	Seq in All Hybrid ***	HybridA + not scaffolded seq	fold change**		
Two enzyme	Count	73	588	5305	8.98x		
	N50 (Mbp)	101.87	11.75	101.33			
	Total length (Mbp)	2855.46	2832.81 (99.21%)	3065.22			
32 sequences were cut during chimeric detection							
bp adjusted							
* Fold change represents the N50 increase between single-color hybrid + not scaffolded sequence relative to the original sequence							
** Fold change represents the N50 increase between All Hybrid + not scaffolded sequence relative to the original sequence							
*** All Hybrid includes merged Hybrid Scaffold A+B & leftover Hybrid Scaffold A & leftover Hybrid Scaffold B							

Table S32.

Susie_PABv1 missing gene annotations in the chromosomes.
66 protein-coding genes are missing in the chromosomes but present in 36 contigs.

000981F_1_96892_quiver_pilon	*CA4*
004770F_1_9128_quiver_pilon	*IL3RA*
001341F_1_71650_quiver_pilon	*LAMA5*
002698F_1_34091_quiver_pilon	*GNG13*
002298F_1_40580_quiver_pilon	*PLCXD1*
004099F_1_17077_quiver_pilon	*OR11H2*
001115F_1_86005_quiver_pilon	*FAM106A*
001816F_1_51941_quiver_pilon	*FAM182B*
004032F_1_17898_quiver_pilon	*DEFB115*
000536F_1_259510_quiver_pilon	*TUBA3C*
001047F_1_64889_quiver_pilon	*MTRNR2L1*
002526F_1_36465_quiver_pilon	*PRAMEF19*
002574F_1_35802_quiver_pilon	*PRAMEF17*
002957F_1_30692_quiver_pilon	*CDRT15L2*
000422F_1_720877_quiver_pilon	*ATP6V0E2*
001635F_1_59138_quiver_pilon	*AL645922.1*
000208F_1_4617805_quiver_pilon	*HSFX2*
000088F_1_11264819_quiver_pilon	*CNTN5*
000088F_1_11264819_quiver_pilon	*CNTN5*
000088F_1_11264819_quiver_pilon	*CNTN5*
000088F_1_11264819_quiver_pilon	*CNTN5*
000088F_1_11264819_quiver_pilon	*CNTN5*
000208F_1_4617805_quiver_pilon	*TMEM185A*
001406F_1_68685_quiver_pilon	*LYG1,LYG2*
001510F_1_64069_quiver_pilon	*CST1,CST4*
5691_1_18979_quiver_pilon	*DEFA1B,DEFA3*
004189F_1_15991_quiver_pilon	*FRG2B,FRG2C*
001709F_1_55791_quiver_pilon	*DMRTC1,DMRTC1B*
002487F_1_37013_quiver_pilon	*OR11H1,OR11H12*

000940F_1_105479_quiver_pilon	*HSFX1,MAGEA11*
002089F_1_44958_quiver_pilon	*OR4M2,RP11-294C11.1*
000208F_1_4617805_quiver_pilon	*HSFX2,TMEM185A*
003142F_1_28354_quiver_pilon	*NPIPA2,NPIPB8,NPIPB9*
000900F_1_117400_quiver_pilon	*OR4M1,OR4Q3,RP11-294C11.3*
000675F_1_166976_quiver_pilon	*CCL3L3,CCL4,CCL4L2,TBC1D3,TBC1D3D,TBC1D3F,TBC1D3G,TBC1D3K*
000395F_1_1011912_quiver_pilon	*ANO9,ATHL1,B4GALNT4,BET1L,IFITM1,IFITM5,NLRP6,ODF3,PKP3,PSMD13,PTDSS2,RIC8A,RNH1,SCGB1C1,SCGB1C2,SIGIRR,SIRT3*

Table S33.

CHM13_HSAv1 Falcon assembly statistics post Quiver and Pilon error-correction.

Title	Statistic
Number of contigs	1,923
Number of contigs not in scaffolds	1,923
Total size of contigs	2,875,999,956 bp
Longest contig	81,018,890 bp
Shortest contig	12 bp
Bases in contigs > 1 kbp	2,875,982,803 (100.0%)
Bases in contigs > 10 kbp	2,874,197,652 (99.9%)
Bases in contigs > 100 kbp	2,836,038,054 (98.6%)
Bases in contigs > 1 Mbp	2,743,109,111 (95.4%)
Bases in contigs > 10 Mbp	2,320,174,640 (80.7%)
Mean contig size	1,495,580 bp
Median contig size	31,852 bp
N50 contig length	29,260,714 bp
L50 contig count	30
contig %A	29.55
contig %C	20.47
contig %G	20.42
contig %T	29.57

Table S34.

Assembly quality: Concordance and base-pair accuracy based on BES mappings.

Title	Statistic
Total bases assessed for concordance*	2,806,276,601
Bases spanned by concordant best*	2,725,455,296
Bases spanned by discordant best and not concordant best*	5,559,303
Bases spanned by discordant best, tied, concordant tied and not concordant best	10,241,711
Bases spanned by concordant best and discordant best (intersection; bases in the reference with both concordant and discordant best support)*	256,467,602
Proportion of bases spanned by concordant best	97.11%

*Contigs greater than 300 kbp.

Table S35.

Base accuracy in the CHM13_HSAv1 assembly contigs.

Base-pair accuracy of CHM13 Quiver and Pilon assemblies based on BES alignment from the human haploid hydatidiform mole BAC library (CHORI-17). Total variations in the table are defined as the sum of all transitions, transversions, deletions and insertions.

Type	Quiver	Quiver + Pilon
Transitions	36,516	36,513
Transversions	19,569	19,586
Deletions (1 bp)	23,384	21,340
Deletions (>1 bp)	9,253	8,901
Insertions (1 bp)	7,719	7,749
Insertions (>1 bp)	8,412	8,436
Total variations	104,853	102,525
High-quality bases (PHRED $\leqslant$ 40)	69,886,280	69,872,777
Expected Sanger errors	25,858	25,853
PacBio - Sanger errors	78,995	76,672
Accuracy	0.998870	0.998903
Estimated QV	29.47	29.60
Estimated QV w/o indels	30.96	30.95
Ti/Tv	1.87	1.86

Table S36.

Scaffolding of CHM13 by Bionano.

Two restriction enzymes, Nt.BspQI and Nb.BssSI, were used to scaffold CHM13 into 105 scaffolds resulting in a 2.8 Gbp assembly.

	Info	Bionano	Seq	Seq in hybrid A	Hybrid A	HybridA + not scaffolded seq	fold change*
One enzyme Nt. BspQI	Count	1206	1923	356	127	1816	2.0x
	N50 (Mbp)	4.48	29.49	29.03	59.27	58.17	
	Total length (Mbp)	2897.70	2876.00	2792.81 (97.11%)	2818.34	2901.53	
	Info	Bionano	Seq	Seq in Hybrid B	Hybrid B	HybridA + not scaffolded seq	fold change*
One enzyme Nb. BssSI	Count	2165	1923	360	162	1847	1.9x
	N50 (Mbp)	2.18	29.49	29.03	57.31	57.31	
	Total length (Mbp)	2874.50	2876.00	2795.10 (97.19%)	2816.35	2897.24	
	Info	All Hybrid***	Seq in All Hybrid ***	HybridA + not scaffolded seq	fold change**		
Two enzyme	Count	105	490	1684	2.8X		
	N50 (Mbp)	82.79	29.26	82.79			
	Total length (Mbp)	2839.81	2817.78 (99.22%)	2905.72			
49 sequences were cut during chimeric detection							
bp adjusted							
* Fold change represents the N50 increase between single-color hybrid + not scaffolded sequence relative to the original sequence							
** Fold change represents the N50 increase between All Hybrid + not scaffolded sequence relative to the original sequence							
*** All Hybrid includes merged Hybrid Scaffold A+B & leftover Hybrid Scaffold A & leftover Hybrid Scaffold B							

Table S37.

YRI_HSAv1 Falcon assembly statistics.
Statistics are calculated post Quiver and Pilon error-correction.

Title	Statistic
Number of contigs	3,645
Number of contigs not in scaffolds	3,645
Total size of contigs	2,878,063,856 bp
Longest contig	27,037,623 bp
Shortest contig	57 bp
Bases in contigs > 1K nt	2,878,038,952 (100.0%)
Bases in contigs > 10K nt	2,875,099,275 (99.9%)
Bases in contigs > 100K nt	2,805,937,728 (97.5%)
Bases in contigs > 1M nt	2,653,091,368 (92.2%)
Bases in contigs > 10M nt	946,864,873 (32.9%)
Mean contig size	789,592 bp
Median contig size	30,223 bp
N50 contig length	6,605,884 bp
L50 contig count	129
contig %A	29.56
contig %C	20.45
contig %G	20.45
contig %T	29.54

Table S38.

YRI_HSAv1 assembly concordance based on BES mappings.

Title	Statistic
Total bases assessed for concordance*	2,744,350,098
Bases spanned by concordant best*	2,682,124,629
Bases spanned by discordant best and not concordant best*	4,471,754
Bases spanned by discordant best, tied, concordant tied and not concordant best	5,626,073
Bases spanned by concordant best and discordant best (intersection; bases in the reference with both concordant and discordant best support)*	1,49,488,465
Proportion of bases spanned by concordant best	97.73%

*BES data not from the same source.

Table S39.

YRI_HSAv1 base accuracy of sequence contigs.

Base-pair accuracy of the assembly with Pilon polishing based on BES alignment from the human haploid hydatidiform mole BAC library (CHORI-17). Total variation in the table is defined as the sum of all transitions, transversions, deletions and insertions.

Type	**Quiver + Pilon**
Transitions	42,724
Transversions	22,462
Deletions (1 bp)	32,977
Deletions (>1 bp)	12,071
Insertions (1 bp)	7,376
Insertions (>1 bp)	8,851
Total variations	126,461
High-quality bases (PHRED ≥ 40)	65,910,571
Expected Sanger errors	24,387
PacBio - Sanger errors	102,074
Accuracy	0.998451
Estimated QV	28.10
Estimated QV w/o indels	30.05
Ti/Tv	1.90

Table S40.

Bionano optical map conflicts with the Falcon-based sequence contigs.

Assembly	Susie_PABv1	Clint_PTRv1
Total number of events	4,443	1,103
Deletions	2,109	520
Insertions	2,285	581

Table S41.
Hi-C data generation.

	Gorilla*	**Chimpanzee (Clint)**	**Orangutan (Susie)**
Cell line	AG05251	S006007	PR01109
Number of QC-passing reads	518 Million	383 Million	428 Million

*Cell line is not from the one used for assembly WGS.

Table S42.

Number of PTVs at various indel error-correction stages compared to RefSeq annotation for GRCh38. Notice that only 5% of PTV indels remained in Yoruban after our correction. Our FreeBayes-based indel-correction pipeline has little effect on PTVs in CHM13_HSAv1 since it is haploid and our pipeline mainly fixes a problem due to heterozygosity. The large number of Clint and Susie PTVs on the "Post FreeBayes-based" line probably are not remaining errors but reflect divergence from human.

	CHM13_HSAv1	YRI_HSAv1	Clint_PTRv1	Susie_PABv1
Falcon assembly	46,469	53,935	47,615	65,618
Post Quiver	1,811	7,972	11,699	43,066
Post Pilon	1,069	4,760	9,408	26,568
Filtered Post Pilon	100	2,053	2,289	5,825
Post FreeBayes-based indel-correction pipeline	91	111	849	2,413
Post FreeBayes-based indel-correction pipeline PTVs as percent of Filtered Post Pilon PTVs	91	5	37	41

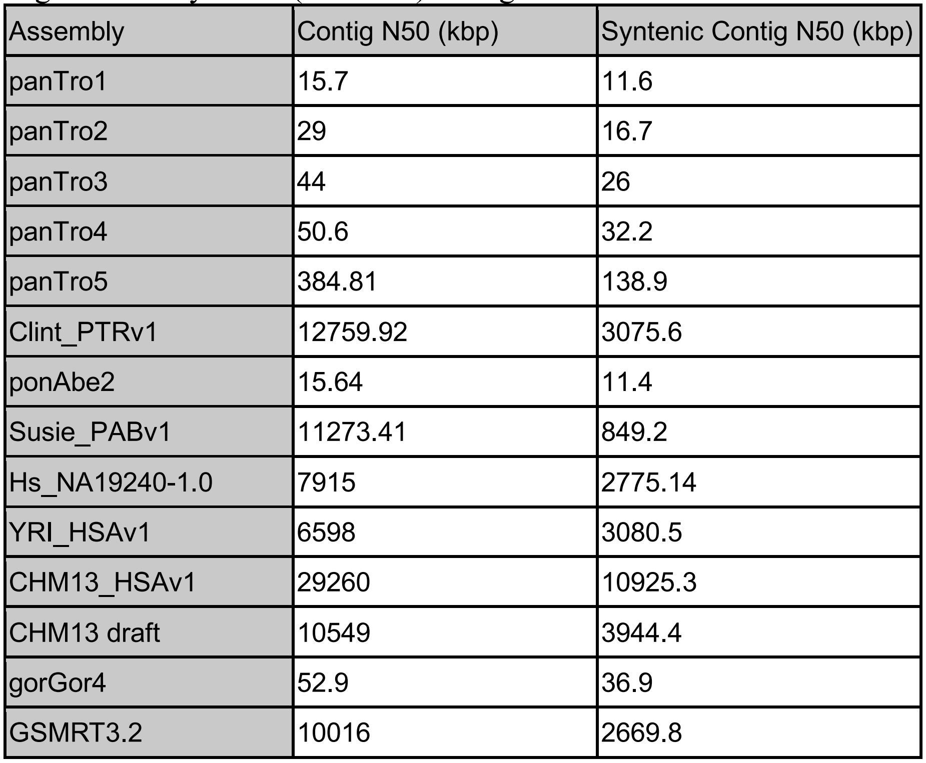
Table S43.

Total number of contigs and assembled length of all public reference assemblies and newer PacBio great ape genomes.

Assembly	Num contigs	Assembled length		Assembly	Num contigs	Assembled length
gorGor4	170,086	2,917,333,143		CHM_draft_assembly	4,961	2,941,135,618
ponAbe2	408,241	3,076,006,523		Hs_NA19240	3,603	2,658,743,407
panTro1	361,864	2,733,948,177		GSMRT3.2	15,997	3,081,475,504
panTro2	183,098	2,478,013,735		Susie_PABv1	5,820	3,042,567,509
panTro3	183,688	2,690,832,212		Clint_PTRv1	5,037	2,992,604,800
panTro4	207,177	2,902,338,967		CHM13_Draft	1,923	2,875,999,956
panTro5	72,784	2,778,536,048		Yri_HSAv1	3,645	2,878,063,856

Table S44.

Contig N50 and syntenic (GRCh38) contig N50.



Assembly	Contig N50 (kbp)	Syntenic Contig N50 (kbp)
panTro1	15.7	11.6
panTro2	29	16.7
panTro3	44	26
panTro4	50.6	32.2
panTro5	384.81	138.9
Clint_PTRv1	12759.92	3075.6
ponAbe2	15.64	11.4
Susie_PABv1	11273.41	849.2
Hs_NA19240-1.0	7915	2775.14
YRI_HSAv1	6598	3080.5
CHM13_HSAv1	29260	10925.3
CHM13 draft	10549	3944.4
gorGor4	52.9	36.9
GSMRT3.2	10016	2669.8

Table S45.

Proportion of GRCh38 assayable by the great ape genomes.
Alignments were generated using BLASR (githash 7cc3379) with parameters -clipping hard -alignContigs –sam -minMapQV 30 -nproc 6 -minPctIdentity 50.

Genome assembly	[A] %GRCh38 assayable bases	[B] % of (A) covered by multiple contigs
Clint_PTRv1	89.60 (2,766,966,581)	0.80 (22,148,790)
Susie_PABv1	88.2 8(2,726,255,894)	0.83 (22,857,911)
GSMRT3.2	89.11 (2,751,913,860)	0.95 (26,333,547)
Yri_HSAv1	90.06 (2,782,865,371)	0.86 (24,080,097)
Clint_PTRv1, Susie_PAB_v2, GSMRT3.2 and Yri_HSA_v1 taken together	85.82 (2,651,790,865)	-
panTro3	1,683,278,492 (54.48)	3.58 (60,272,966)
P_pygmaeus_2.0.2	2,405,306,035 (69.38)	0.0 (0)
gorGor3	2,143,892,483 (77.84)	0.03 (908,738)
panTro3, P_pygmaeus_2.0.2 and gorGor3 taken together	1,247,841,390 (40.38)	-

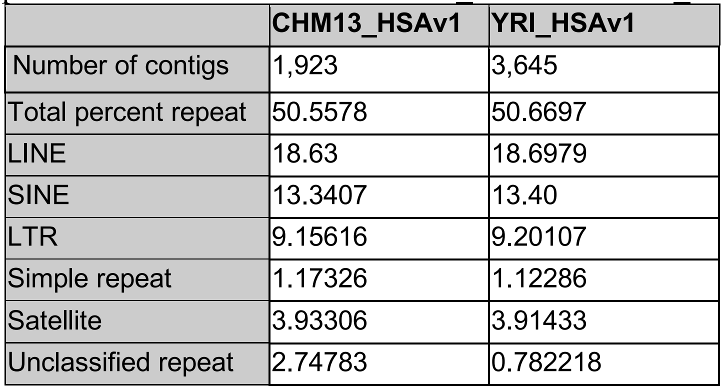
Table S46.

NHP alignment fragmentation regions.

Chromosome	Start	End	RefSeq gene ID
chr1	1,296,550	1,313,196	*ACAP3,CPSF3L,MIR6727,PUSL1*
chr1	227,679,868	227,714,046	*ZNF847P*
chr1	227,715,207	227,817,443	*JMJD4,LOC100130093,PRSS38,SNAP47*
chr1	227,857,373	227,959,986	*MIR5008,WNT9A*
chr1	227,986,033	228,502,931	*ARF1,C1orf145,C1orf35,GJC2,GUK1,HIST3H2A,HIST3H2BB,HIST3H3,IBA57,IBA57-AS1,MIR3620,MIR4666A,MIR6742,MRPL55,OBSCN,RNF187,TRIM11,TRIM17,WNT3A*
chr10	19,126,009	19,129,207	*MALRD1*
chr10	19,132,341	19,134,386	*MALRD1*
chr19	21,031,249	21,073,580	*ZNF430*

Table S47.

RepeatMasker summaries for CHM13_HSAv1 and Yri_HSAv1 de novo PacBio assemblies.



	CHM13_HSAv1	YRI_HSAv1
Number of contigs	1,923	3,645
Total percent repeat	50.5578	50.6697
LINE	18.63	18.6979
SINE	13.3407	13.40
LTR	9.15616	9.20107
Simple repeat	1.17326	1.12286
Satellite	3.93306	3.91433
Unclassified repeat	2.74783	0.782218

Table S48.

RepeatMasker summaries for comparing the previous reference NHP assemblies with the de novo PacBio genomes.

RepeatMasker results for current reference genomes and all new (PacBio) NHP genomes showing total percent repeat in satellites, simple repeats, LINE, SINE and LTR.

FALCON: chromosomes only				Current reference: chromosomes only			
	GSMRT3.2	Clint_PTRv1	Susie_PABv1		gorGor4	panTro5	ponAbe2
Number of contigs	792	690	526	Number of contigs	46,823	44,448	8,571
Total percent repeat	48.132	50.56	50.97	Total percent repeat	47.51	50.49%	45.51
LINE	18.081	18.98	19.6	LINE	16.89	17.86	17.23
SINE	12.87	13.56	13.33	SINE	10.98	12.82	11.68
LTR	8.89	9.36	9.31	LTR	8.45	8.81	8.28
Simple repeat	0.79	0.821	0.81	Simple repeat	0.8	0.94	0.68
Satellite	**3.31**	**3.44**	**3.58**	Satellite	6.44	6	3.8
FALCON: all contigs				Current reference: all contigs			
	GSMRT3.2	Clint_PTRv1	Susie_PABv1		gorGor4	panTro5	ponAbe2
Number of contigs	15,997	4,912	5,771	Number of contigs	166,061	71,660	401,949
Total percent repeat	53.64	52.6	50.67	Total percent repeat	51.49	51.93	51.24
LINE	17.51	18.05	18.5	LINE	18.22	18.24	19.13
SINE	12.51	12.9	12.68	SINE	13.01	13.11	12.93
LTR	8.62	8.94	8.76	LTR	9.07	8.98	9.3
Simple repeat	1.19	0.91	1.8	Simple repeat	1	1	1.08
Satellite	**9.65**	7.64	4.74	Satellite	**5.94**	6.39	4.52

Table S49.

Divergence against GRCh38 measured in 1 Mbp non-overlapping windows.
Divergence was calculated as the number SNVs within a window divided by the number of aligned bases within the window. Since multiple contigs can align to the same region of GRCh38, we counted each aligned base, i.e., some 1 Mbp windows have more than 1 Mbp of sequence in the divergence calculation. Listed in the table are the percent divergence and standard deviation broken into the autosomes and chromosome X.

	Yoruban	CHM13	Chimpanzee	Gorilla	Orangutan
Autosome	0.12 (0.07)	0.1 (0.08)	1.27 (0.20)	1.61 (0.21)	3.12 (0.33)
chrX	0.09 (0.05)	0.07 (0.44)	0.98 (0.22)	1.42 (0.23)	2.6 (0.28)

Table S50.

Primate iPSC lines used for cDNA sequencing.

Species	Name	Sex
Homo sapiens	WT-33	F
Pan troglodytes	818	F
Gorilla gorilla	053	M
Pongo abelii	Jos3C1	F

Table S51.

Paired-end RNA-seq short-read counts for each organism post-demultiplexing.
(iPSC lines are not derived from the same individuals used for genome sequencing and assembly.)

Sample	Read Pairs
human iPSC	91,330,785
chimpanzee iPSC	61,417,249
gorilla iPSC	63,030,948
orangutan iPSC	128,459,588

Table S52.

Summary of full-length Iso-Seq reads.

Sample	SMRT cells	FLNC count	Median FLNC length (bp)	FLNC length IQR (bp)
Human	27	710,974	2067	[1623, 3016]
Chimpanzee	25	565,691	2108	[1596, 3035]
Gorilla	32	881,801	2069	[1299, 2514]
Orangutan	30	528,145	1918	[1169, 2574]

Table S53.

The number of FLNC reads mapping to each genome assembly.
The row labels contain the number of source sample reads in brackets. Each cell contains the number and proportion of the row's reads that map to the column's genome assembly.

Sample [*n* FLNC reads]	**ALIGNED to:** human (fraction)	**ALIGNED to:** chimpanzee (fraction)	**ALIGNED to:** gorilla (fraction)	**ALIGNED to:** orangutan (fraction)
Human cDNA [710,974]	710,899 (0.9998945)	710,478 (0.9993023)	710,501 (0.9993347)	710,521 (0.9993628)
Chimpanzee cDNA [565,691]	565,581 (0.9998055)	565,224 (0.9991745)	565,433 (0.9995439)	565,352 (0.9994007)
Gorilla cDNA [881,801]	881,393 (0.9995373)	880,885 (0.9989612)	881,166 (0.9992799)	880,956 (0.9990417)
Orangutan cDNA [528,145]	527,913 (0.9995607)	527,770 (0.9992900)	527,704 (0.9991650)	527,946 (0.9996232)

Table S54.

Publically available primate RNA-seq obtained via SRA for great ape annotation.

Species	SRA Accessions	Tissues
Orangutan	SRR306792, SRR2176206, SRR2176207	Brain, testis
Gorilla	SRR832925, SRR3053573, SRR306801	Brain, 20 tissue pool
Chimpanzee	SRR2040584, SRR2040585, SRR2040586, SRR2040587, SRR2040588, SRR2040589, SRR2040590, SRR2040591, SRR3711187, SRR3711188, SRR873622, SRR873623, SRR873624, SRR873625	brain, heart, liver, testis, 8-week-old iPSC-derived neurons, undifferentiated iPSC
Human	ERR579132, ERR579133, ERR579134, ERR579135, ERR579136, ERR579137, ERR579138, ERR579139, ERR579140, ERR579141, ERR579142, ERR579143, ERR579144, ERR579145, ERR579146, ERR579147, ERR579148, ERR579149, ERR579150, ERR579151, ERR579152, ERR579153, ERR579154, ERR579155	Ovary, tonsil, fallopian tube, placenta, endometrium, rectum, skeletal muscle, liver, fat, colon, smooth muscle, lung

Table S55.

Filtered SV counts for each assembly against GRCh38.

	CHM13	Yoruban	Clint_PTRv1	Gorilla	Susie_PABv1
Deletions	9,126	11,747	63,634	73,681	136,980
Insertions	14,962	14,528	68,589	76,230	142,631
Inversions	74	55	446	533	969
Total	24,162	26,330	132,669	150,444	280,580

Table S56.

NCBI accessions for assembly and WGS data.

Species	Chimpanzee	Orangutan	Human	Human	Gorilla
Assembly name	Clint_PTRv1	Susie_PABv1	CHM13_HSAv1	Yri_HSAv1	GSMRT3.2
BioProject	PRJNA369439	PRJNA369439	PRJNA369439	PRJNA369439	PRJEB10880
Assembly accession	NBAG00000000	NDHI00000000	NTIA00000000	NTIB00000000	GCA_900006655.1
BioSample WGS (PacBio h5)	SAMN06272697	SAMN06275555	SAMN03255769	SAMN03838746 (external), SRS988474	SAMEA3541598
Illumina	SRX243527 (external)	SRR6029680	ERP014751 (external)	SRX1098167 (external)	SRP018689
Hi-C	SRR5977046	SRR6026886	na	na	SRR6318338
Bionano	SUPPF_0000001269, SUPPF_0000001270	SUPPF_0000001271, SUPPF_0000001272	SUPPF_0000001346, SUPPF_0000001345	na	na
Structural Variation Data	EBI estd235[*] for all 5 organisms	estd235	estd235	estd235	estd235

* download at ftp://ftp.ebi.ac.uk/pub/databases/dgva/estd235_Kronenberg_et_al_2017/vcf/

Table S57.

NCBI accessions for transcriptome data.

RNA-seq	Chimpanzee	Orangutan	Human	Gorilla
BioSample ID	SAMN07611970	SAMN07611972	SAMN07611993	SAMN07611971
SRA/SUB RNA-seq	SRR6025894	SRR6026509	SRR6026510	SRR6025931
SRA/SUB Iso-Seq	SRR6039150 - SRR6039174 (25 runs)	SRR6077502 - SRR6077473 (30 runs)	SRR6051611 - SRR6051585 (27 runs)	SRR6077537- SRR6077506 (32 runs)

Table S58.
Assembly name mappings.

Organism	Species	Submitter (UW) provided assembly name	NCBI GenBank accession	RefSeq accession	RefSeq annotation set	UCSC genome database name
Chimpanzee	Pan troglodytes	Clint_PTRv2	GCA_002880755.3	GCF_002880755.1	Pan troglodytes Annotation Release 105	panTro6
Orangutan	Pongo abelii	Susie_PABv2	GCA_002880775.3	GCF_002880775.1	Pongo abelii Annotation Release 103	ponAbe3
Gorilla	Gorilla gorilla	GSMRT3	GCA_900006655.1	NA	NA	gorGor5

References and Notes

1. A. Varki, D. H. Geschwind, E. E. Eichler, Human uniqueness: Genome interactions with environment, behaviour and culture. *Nat. Rev. Genet.* **9**, 749–763 (2008). doi:10.1038/nrg2428 Medline

2. M. C. King, A. C. Wilson, Evolution at two levels in humans and chimpanzees. *Science* **188**, 107–116 (1975). doi:10.1126/science.1090005 Medline

3. A. Fortna, Y. Kim, E. MacLaren, K. Marshall, G. Hahn, L. Meltesen, M. Brenton, R. Hink, S. Burgers, T. Hernandez-Boussard, A. Karimpour-Fard, D. Glueck, L. McGavran, R. Berry, J. Pollack, J. M. Sikela, Lineage-specific gene duplication and loss in human and great ape evolution. *PLOS Biol.* **2**, e207 (2004). doi:10.1371/journal.pbio.0020207 Medline

4. J. L. Boyd, S. L. Skove, J. P. Rouanet, L.-J. Pilaz, T. Bepler, R. Gordân, G. A. Wray, D. L. Silver, Human-chimpanzee differences in a FZD8 enhancer alter cell-cycle dynamics in the developing neocortex. *Curr. Biol.* **25**, 772–779 (2015). doi:10.1016/j.cub.2015.01.041 Medline

5. C. Y. McLean, P. L. Reno, A. A. Pollen, A. I. Bassan, T. D. Capellini, C. Guenther, V. B. Indjeian, X. Lim, D. B. Menke, B. T. Schaar, A. M. Wenger, G. Bejerano, D. M. Kingsley, Human-specific loss of regulatory DNA and the evolution of human-specific traits. *Nature* **471**, 216–219 (2011). doi:10.1038/nature09774 Medline

6. M. Y. Dennis, X. Nuttle, P. H. Sudmant, F. Antonacci, T. A. Graves, M. Nefedov, J. A. Rosenfeld, S. Sajjadian, M. Malig, H. Kotkiewicz, C. J. Curry, S. Shafer, L. G. Shaffer, P. J. de Jong, R. K. Wilson, E. E. Eichler, Evolution of human-specific neural SRGAP2 genes by incomplete segmental duplication. *Cell* **149**, 912–922 (2012). doi:10.1016/j.cell.2012.03.033 Medline

7. C. Charrier, K. Joshi, J. Coutinho-Budd, J.-E. Kim, N. Lambert, J. de Marchena, W.-L. Jin, P. Vanderhaeghen, A. Ghosh, T. Sassa, F. Polleux, Inhibition of SRGAP2 function by its human-specific paralogs induces neoteny during spine maturation. *Cell* **149**, 923–935 (2012). doi:10.1016/j.cell.2012.03.034 Medline

8. M. Florio, M. Albert, E. Taverna, T. Namba, H. Brandl, E. Lewitus, C. Haffner, A. Sykes, F. K. Wong, J. Peters, E. Guhr, S. Klemroth, K. Prüfer, J. Kelso, R. Naumann, I. Nüsslein, A. Dahl, R. Lachmann, S. Pääbo, W. B. Huttner, Human-specific gene ARHGAP11B promotes basal progenitor amplification and neocortex expansion. *Science* **347**, 1465–1470 (2015). doi:10.1126/science.aaa1975 Medline

9. X.-C. Ju, Q.-Q. Hou, A.-L. Sheng, K.-Y. Wu, Y. Zhou, Y. Jin, T. Wen, Z. Yang, X. Wang, Z.-G. Luo, The hominoid-specific gene TBC1D3 promotes generation of basal neural progenitors and induces cortical folding in mice. *eLife* **5**, 206 (2016). doi:10.7554/eLife.18197 Medline

10. A. Scally, J. Y. Dutheil, L. W. Hillier, G. E. Jordan, I. Goodhead, J. Herrero, A. Hobolth, T. Lappalainen, T. Mailund, T. Marques-Bonet, S. McCarthy, S. H. Montgomery, P. C. Schwalie, Y. A. Tang, M. C. Ward, Y. Xue, B. Yngvadottir, C. Alkan, L. N. Andersen, Q. Ayub, E. V. Ball, K. Beal, B. J. Bradley, Y. Chen, C. M. Clee, S. Fitzgerald, T. A. Graves, Y. Gu, P. Heath, A. Heger, E. Karakoc, A. Kolb-Kokocinski, G. K. Laird, G.

Lunter, S. Meader, M. Mort, J. C. Mullikin, K. Munch, T. D. O'Connor, A. D. Phillips, J. Prado-Martinez, A. S. Rogers, S. Sajjadian, D. Schmidt, K. Shaw, J. T. Simpson, P. D. Stenson, D. J. Turner, L. Vigilant, A. J. Vilella, W. Whitener, B. Zhu, D. N. Cooper, P. de Jong, E. T. Dermitzakis, E. E. Eichler, P. Flicek, N. Goldman, N. I. Mundy, Z. Ning, D. T. Odom, C. P. Ponting, M. A. Quail, O. A. Ryder, S. M. Searle, W. C. Warren, R. K. Wilson, M. H. Schierup, J. Rogers, C. Tyler-Smith, R. Durbin, Insights into hominid evolution from the gorilla genome sequence. *Nature* **483**, 169–175 (2012). doi:10.1038/nature10842 Medline

11. D. P. Locke, L. W. Hillier, W. C. Warren, K. C. Worley, L. V. Nazareth, D. M. Muzny, S.-P. Yang, Z. Wang, A. T. Chinwalla, P. Minx, M. Mitreva, L. Cook, K. D. Delehaunty, C. Fronick, H. Schmidt, L. A. Fulton, R. S. Fulton, J. O. Nelson, V. Magrini, C. Pohl, T. A. Graves, C. Markovic, A. Cree, H. H. Dinh, J. Hume, C. L. Kovar, G. R. Fowler, G. Lunter, S. Meader, A. Heger, C. P. Ponting, T. Marques-Bonet, C. Alkan, L. Chen, Z. Cheng, J. M. Kidd, E. E. Eichler, S. White, S. Searle, A. J. Vilella, Y. Chen, P. Flicek, J. Ma, B. Raney, B. Suh, R. Burhans, J. Herrero, D. Haussler, R. Faria, O. Fernando, F. Darré, D. Farré, E. Gazave, M. Oliva, A. Navarro, R. Roberto, O. Capozzi, N. Archidiacono, G. Della Valle, S. Purgato, M. Rocchi, M. K. Konkel, J. A. Walker, B. Ullmer, M. A. Batzer, A. F. A. Smit, R. Hubley, C. Casola, D. R. Schrider, M. W. Hahn, V. Quesada, X. S. Puente, G. R. Ordoñez, C. López-Otín, T. Vinar, B. Brejova, A. Ratan, R. S. Harris, W. Miller, C. Kosiol, H. A. Lawson, V. Taliwal, A. L. Martins, A. Siepel, A. Roychoudhury, X. Ma, J. Degenhardt, C. D. Bustamante, R. N. Gutenkunst, T. Mailund, J. Y. Dutheil, A. Hobolth, M. H. Schierup, O. A. Ryder, Y. Yoshinaga, P. J. de Jong, G. M. Weinstock, J. Rogers, E. R. Mardis, R. A. Gibbs, R. K. Wilson, Comparative and demographic analysis of orangutan genomes. *Nature* **469**, 529–533 (2011). doi:10.1038/nature09687 Medline

12. Chimpanzee Sequencing and Analysis Consortium, Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**, 69–87 (2005). doi:10.1038/nature04072 Medline

13. E. T. Lam, A. Hastie, C. Lin, D. Ehrlich, S. K. Das, M. D. Austin, P. Deshpande, H. Cao, N. Nagarajan, M. Xiao, P.-Y. Kwok, Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. *Nat. Biotechnol.* **30**, 771–776 (2012). doi:10.1038/nbt.2303 Medline

14. J. N. Burton, A. Adey, R. P. Patwardhan, R. Qiu, J. O. Kitzman, J. Shendure, Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nat. Biotechnol.* **31**, 1119–1125 (2013). doi:10.1038/nbt.2727 Medline

15. J. Eid, A. Fehr, J. Gray, K. Luong, J. Lyle, G. Otto, P. Peluso, D. Rank, P. Baybayan, B. Bettman, A. Bibillo, K. Bjornson, B. Chaudhuri, F. Christians, R. Cicero, S. Clark, R. Dalal, A. Dewinter, J. Dixon, M. Foquet, A. Gaertner, P. Hardenbol, C. Heiner, K. Hester, D. Holden, G. Kearns, X. Kong, R. Kuse, Y. Lacroix, S. Lin, P. Lundquist, C. Ma, P. Marks, M. Maxham, D. Murphy, I. Park, T. Pham, M. Phillips, J. Roy, R. Sebra, G. Shen, J. Sorenson, A. Tomaney, K. Travers, M. Trulson, J. Vieceli, J. Wegener, D. Wu, A. Yang, D. Zaccarin, P. Zhao, F. Zhong, J. Korlach, S. Turner, Real-time DNA sequencing from single polymerase molecules. *Science* **323**, 133–138 (2009). doi:10.1126/science.1162986 Medline

16. Materials and methods are available as supplementary materials.

17. C.-S. Chin, D. H. Alexander, P. Marks, A. A. Klammer, J. Drake, C. Heiner, A. Clum, A. Copeland, J. Huddleston, E. E. Eichler, S. W. Turner, J. Korlach, Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods* **10**, 563–569 (2013). doi:10.1038/nmeth.2474 Medline

18. B. J. Walker, T. Abeel, T. Shea, M. Priest, A. Abouelliel, S. Sakthikumar, C. A. Cuomo, Q. Zeng, J. Wortman, S. K. Young, A. M. Earl, Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLOS ONE* **9**, e112963 (2014). doi:10.1371/journal.pone.0112963 Medline

19. J. J. Yunis, O. Prakash, The origin of man: A chromosomal pictorial legacy. *Science* **215**, 1525–1530 (1982). doi:10.1126/science.7063861 Medline

20. D. Gordon, J. Huddleston, M. J. P. Chaisson, C. M. Hill, Z. N. Kronenberg, K. M. Munson, M. Malig, A. Raja, I. Fiddes, L. W. Hillier, C. Dunn, C. Baker, J. Armstrong, M. Diekhans, B. Paten, J. Shendure, R. K. Wilson, D. Haussler, C.-S. Chin, E. E. Eichler, Long-read sequence assembly of the gorilla genome. *Science* **352**, aae0344 (2016). doi:10.1126/science.aae0344 Medline

21. L. F. K. Kuderna, C. Tomlinson, L. W. Hillier, A. Tran, I. T. Fiddes, J. Armstrong, H. Laayouni, D. Gordon, J. Huddleston, R. Garcia Perez, I. Povolotskaya, A. Serres Armero, J. Gómez Garrido, D. Ho, P. Ribeca, T. Alioto, R. E. Green, B. Paten, A. Navarro, J. Betranpetit, J. Herrero, E. E. Eichler, A. J. Sharp, L. Feuk, W. C. Warren, T. Marques-Bonet, A 3-way hybrid approach to generate a new high-quality chimpanzee reference genome (Pan_tro_3.0). *Gigascience* **6**, 1–6 (2017). doi:10.1093/gigascience/gix098 Medline

22. I. T. Fiddes, J. Armstrong, M. Diekhans, S. Nachtweide, Z. N. Kronenberg, J. G. Underwood, D. Gordon, D. Earl, T. Keane, E. E. Eichler, D. Haussler, M. Stanke, B. Paten, Comparative Annotation Toolkit (CAT)–simultaneous clade and personal genome annotation. bioRxiv 231118 [Preprint]. 8 December 2017. https://doi.org/10.1101/231118.

23. N. Elango, J. W. Thomas, S. V. Yi; NISC Comparative Sequencing Program, Variable molecular clocks in hominoids. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 1370–1375 (2006). doi:10.1073/pnas.0510716103 Medline

24. P. Moorjani, C. E. G. Amorim, P. F. Arndt, M. Przeworski, Variation in the molecular clock of primates. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 10607–10612 (2016). doi:10.1073/pnas.1600374113 Medline

25. W. H. Li, M. Tanimura, P. M. Sharp, An evaluation of the molecular clock hypothesis using mammalian DNA sequences. *J. Mol. Evol.* **25**, 330–342 (1987). doi:10.1007/BF02603118 Medline

26. D. M. Bickhart, B. D. Rosen, S. Koren, B. L. Sayre, A. R. Hastie, S. Chan, J. Lee, E. T. Lam, I. Liachko, S. T. Sullivan, J. N. Burton, H. J. Huson, J. C. Nystrom, C. M. Kelley, J. L. Hutchison, Y. Zhou, J. Sun, A. Crisà, F. A. Ponce de León, J. C. Schwartz, J. A. Hammond, G. C. Waldbieser, S. G. Schroeder, G. E. Liu, M. J. Dunham, J. Shendure, T. S. Sonstegard, A. M. Phillippy, C. P. Van Tassell, T. P. L. Smith, Single-molecule

sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. *Nat. Genet.* **49**, 643–650 (2017). doi:10.1038/ng.3802 Medline

27. D. C. Rubinsztein, W. Amos, J. Leggo, S. Goodburn, S. Jain, S.-H. Li, R. L. Margolis, C. A. Ross, M. A. Ferguson-Smith, Microsatellite evolution—evidence for directionality and variation in rate between species. *Nat. Genet.* **10**, 337–343 (1995). doi:10.1038/ng0795-337 Medline

28. M. T. Webster, N. G. C. Smith, H. Ellegren, Microsatellite evolution inferred from human-chimpanzee genomic sequence alignments. *Proc. Natl. Acad. Sci. U.S.A.* **99**, 8748–8753 (2002). doi:10.1073/pnas.122067599 Medline

29. C. T. Yohn, Z. Jiang, S. D. McGrath, K. E. Hayden, P. Khaitovich, M. E. Johnson, M. Y. Eichler, J. D. McPherson, S. Zhao, S. Pääbo, E. E. Eichler, Lineage-specific expansions of retroviral insertions within the genomes of African great apes but not humans and orangutans. *PLOS Biol.* **3**, e110 (2005). doi:10.1371/journal.pbio.0030110 Medline

30. N. Polavarapu, N. J. Bowen, J. F. McDonald, Identification, characterization and comparative genomics of chimpanzee endogenous retroviruses. *Genome Biol.* **7**, R51 (2006). doi:10.1186/gb-2006-7-6-r51 Medline

31. S. M. Kaiser, H. S. Malik, M. Emerman, Restriction of an extinct retrovirus by the human TRIM5alpha antiviral protein. *Science* **316**, 1756–1758 (2007). doi:10.1126/science.1140579 Medline

32. P. H. Sudmant, T. Rausch, E. J. Gardner, R. E. Handsaker, A. Abyzov, J. Huddleston, Y. Zhang, K. Ye, G. Jun, M. H. Fritz, M. K. Konkel, A. Malhotra, A. M. Stütz, X. Shi, F. P. Casale, J. Chen, F. Hormozdiari, G. Dayama, K. Chen, M. Malig, M. J. P. Chaisson, K. Walter, S. Meiers, S. Kashin, E. Garrison, A. Auton, H. Y. K. Lam, X. J. Mu, C. Alkan, D. Antaki, T. Bae, E. Cerveira, P. Chines, Z. Chong, L. Clarke, E. Dal, L. Ding, S. Emery, X. Fan, M. Gujral, F. Kahveci, J. M. Kidd, Y. Kong, E.-W. Lameijer, S. McCarthy, P. Flicek, R. A. Gibbs, G. Marth, C. E. Mason, A. Menelaou, D. M. Muzny, B. J. Nelson, A. Noor, N. F. Parrish, M. Pendleton, A. Quitadamo, B. Raeder, E. E. Schadt, M. Romanovitch, A. Schlattl, R. Sebra, A. A. Shabalin, A. Untergasser, J. A. Walker, M. Wang, F. Yu, C. Zhang, J. Zhang, X. Zheng-Bradley, W. Zhou, T. Zichner, J. Sebat, M. A. Batzer, S. A. McCarroll, R. E. Mills, M. B. Gerstein, A. Bashir, O. Stegle, S. E. Devine, C. Lee, E. E. Eichler, J. O. Korbel; 1000 Genomes Project Consortium, An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–81 (2015). doi:10.1038/nature15394 Medline

33. J. Prado-Martinez, P. H. Sudmant, J. M. Kidd, H. Li, J. L. Kelley, B. Lorente-Galdos, K. R. Veeramah, A. E. Woerner, T. D. O’Connor, G. Santpere, A. Cagan, C. Theunert, F. Casals, H. Laayouni, K. Munch, A. Hobolth, A. E. Halager, M. Malig, J. Hernandez-Rodriguez, I. Hernando-Herraez, K. Prüfer, M. Pybus, L. Johnstone, M. Lachmann, C. Alkan, D. Twigg, N. Petit, C. Baker, F. Hormozdiari, M. Fernandez-Callejo, M. Dabad, M. L. Wilson, L. Stevison, C. Camprubí, T. Carvalho, A. Ruiz-Herrera, L. Vives, M. Mele, T. Abello, I. Kondova, R. E. Bontrop, A. Pusey, F. Lankester, J. A. Kiyang, R. A. Bergl, E. Lonsdorf, S. Myers, M. Ventura, P. Gagneux, D. Comas, H. Siegismund, J. Blanc, L. Agueda-Calpena, M. Gut, L. Fulton, S. A. Tishkoff, J. C. Mullikin, R. K.

Wilson, I. G. Gut, M. K. Gonder, O. A. Ryder, B. H. Hahn, A. Navarro, J. M. Akey, J. Bertranpetit, D. Reich, T. Mailund, M. H. Schierup, C. Hvilsom, A. M. Andrés, J. D. Wall, C. D. Bustamante, M. F. Hammer, E. E. Eichler, T. Marques-Bonet, Great ape genetic diversity and population history. *Nature* **499**, 471–475 (2013). doi:10.1038/nature12228 Medline

34. Y. A. Pérez-Rico, V. Boeva, A. C. Mallory, A. Bitetti, S. Majello, E. Barillot, A. Shkumatava, Comparative analyses of super-enhancers reveal conserved elements in vertebrate genomes. *Genome Res.* **27**, 259–268 (2017). doi:10.1101/gr.203679.115 Medline

35. P. L. Reno, C. Y. McLean, J. E. Hines, T. D. Capellini, G. Bejerano, D. M. Kingsley, A penile spine/vibrissa enhancer sequence is missing in modern and extinct humans but is retained in multiple primates with penile spines and sensory vibrissae. *PLOS ONE* **8**, e84258 (2013). doi:10.1371/journal.pone.0084258 Medline

36. A. Ameur, S. Enroth, A. Johansson, G. Zaboli, W. Igl, A. C. V. Johansson, M. A. Rivas, M. J. Daly, G. Schmitz, A. A. Hicks, T. Meitinger, L. Feuk, C. van Duijn, B. Oostra, P. P. Pramstaller, I. Rudan, A. F. Wright, J. F. Wilson, H. Campbell, U. Gyllensten, Genetic adaptation of fatty-acid metabolism: A human-specific haplotype increasing the biosynthesis of long-chain omega-3 and omega-6 fatty acids. *Am. J. Hum. Genet.* **90**, 809–820 (2012). doi:10.1016/j.ajhg.2012.03.014 Medline

37. K. Ye, F. Gao, D. Wang, O. Bar-Yosef, A. Keinan, Dietary adaptation of FADS genes in Europe varied across time and geography. *Nat. Ecol. Evol.* **1**, 0167 (2017). doi:10.1038/s41559-017-0167 Medline

38. M. T. Buckley, F. Racimo, M. E. Allentoft, M. K. Jensen, A. Jonsson, H. Huang, F. Hormozdiari, M. Sikora, D. Marnetto, E. Eskin, M. E. Jørgensen, N. Grarup, O. Pedersen, T. Hansen, P. Kraft, E. Willerslev, R. Nielsen, Selection in Europeans on fatty acid desaturases associated with dietary changes. *Mol. Biol. Evol.* **34**, 1307–1318 (2017). doi:10.1093/molbev/msx103 Medline

39. T. N. Petersen, S. Brunak, G. von Heijne, H. Nielsen, SignalP 4.0: Discriminating signal peptides from transmembrane regions. *Nat. Methods* **8**, 785–786 (2011). doi:10.1038/nmeth.1701 Medline

40. N. B. Trunnell, A. C. Poon, S. Y. Kim, J. E. Ferrell Jr., Ultrasensitivity in the Regulation of Cdc25C by Cdk1. *Mol. Cell* **41**, 263–274 (2011). doi:10.1016/j.molcel.2011.01.012 Medline

41. P. Rakic, A small step for the cell, a giant leap for mankind: A hypothesis of neocortical expansion during evolution. *Trends Neurosci.* **18**, 383–388 (1995). doi:10.1016/0166-2236(95)93934-P Medline

42. M. Pendleton, R. Sebra, A. W. C. Pang, A. Ummat, O. Franzen, T. Rausch, A. M. Stütz, W. Stedman, T. Anantharaman, A. Hastie, H. Dai, M. H.-Y. Fritz, H. Cao, A. Cohain, G. Deikus, R. E. Durrett, S. C. Blanchard, R. Altman, C.-S. Chin, Y. Guo, E. E. Paxinos, J. O. Korbel, R. B. Darnell, W. R. McCombie, P.-Y. Kwok, C. E. Mason, E. E. Schadt, A. Bashir, Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nat. Methods* **12**, 780–786 (2015). doi:10.1038/nmeth.3454 Medline

43. A. C. Y. Mak, Y. Y. Y. Lai, E. T. Lam, T.-P. Kwok, A. K. Y. Leung, A. Poon, Y. Mostovoy, A. R. Hastie, W. Stedman, T. Anantharaman, W. Andrews, X. Zhou, A. W. C. Pang, H. Dai, C. Chu, C. Lin, J. J. K. Wu, C. M. L. Li, J.-W. Li, A. K. Y. Yim, S. Chan, J. Sibert, Ž. Džakula, H. Cao, S. M. Yiu, T. F. Chan, K. Y. Yip, M. Xiao, P. Y. Kwok, Genome-wide structural variation detection by genome mapping on nanochannel arrays. *Genetics* **202**, 351–362 (2016). doi:10.1534/genetics.115.183483 Medline

44. L. Feuk, J. R. MacDonald, T. Tang, A. R. Carson, M. Li, G. Rao, R. Khaja, S. W. Scherer, Discovery of human inversion polymorphisms by comparative analysis of human and chimpanzee DNA sequence assemblies. *PLOS Genet.* **1**, e56 (2005). doi:10.1371/journal.pgen.0010056 Medline

45. T. L. Newman, E. Tuzun, V. A. Morrison, K. E. Hayden, M. Ventura, S. D. McGrath, M. Rocchi, E. E. Eichler, A genome-wide survey of structural variation between human and chimpanzee. *Genome Res.* **15**, 1344–1356 (2005). doi:10.1101/gr.4338005 Medline

46. J. M. Szamalek, D. N. Cooper, W. Schempp, P. Minich, M. Kohn, J. Hoegel, V. Goidts, H. Hameister, H. Kehrer-Sawatzki, Polymorphic micro-inversions contribute to the genomic variability of humans and chimpanzees. *Hum. Genet.* **119**, 103–112 (2006). doi:10.1007/s00439-005-0117-6 Medline

47. M. F. Cardone, Z. Jiang, P. D’Addabbo, N. Archidiacono, M. Rocchi, E. E. Eichler, M. Ventura, Hominoid chromosomal rearrangements on 17q map to complex regions of segmental duplication. *Genome Biol.* **9**, R28 (2008). doi:10.1186/gb-2008-9-2-r28 Medline

48. M. C. Zody, Z. Jiang, H.-C. Fung, F. Antonacci, L. W. Hillier, M. F. Cardone, T. A. Graves, J. M. Kidd, Z. Cheng, A. Abouelleil, L. Chen, J. Wallis, J. Glasscock, R. K. Wilson, A. D. Reily, J. Duckworth, M. Ventura, J. Hardy, W. C. Warren, E. E. Eichler, Evolutionary toggling of the MAPT 17q21.31 inversion region. *Nat. Genet.* **40**, 1076–1083 (2008). doi:10.1038/ng.193 Medline

49. A. D. Sanders, M. Hills, D. Porubský, V. Guryev, E. Falconer, P. M. Lansdorp, Characterizing polymorphic inversions in human genomes by single-cell sequencing. *Genome Res.* **26**, 1575–1587 (2016). doi:10.1101/gr.201160.115 Medline

50. M. J. P. Chaisson, A. D. Sanders, X. Zhao, A. Malhotra, D. Porubsky, T. Rausch, E. J. Gardner, O. Rodriguez, L. Guo, R. L. Collins, X. Fan, J. Wen, R. E. Handsaker, S. Fairley, Z. N. Kronenberg, X. Kong, F. Hormozdiari, D. Lee, A. M. Wenger, A. Hastie, D. Antaki, P. Audano, H. Brand, S. Cantsilieris, H. Cao, E. Cerveira, C. Chen, X. Chen, C.-S. Chin, Z. Chong, N. T. Chuang, D. M. Church, L. Clarke, A. Farrell, J. Flores, T. Galeev, G. David, M. Gujral, V. Guryev, W. Haynes-Heaton, J. Korlach, S. Kumar, J. Y. Kwon, J. E. Lee, J. Lee, W.-P. Lee, S. P. Lee, P. Marks, K. Valud-Martinez, S. Meiers, K. M. Munson, F. Navarro, B. J. Nelson, C. Nodzak, A. Noor, S. Kyriazopoulou-Panagiotopoulou, A. Pang, Y. Qiu, G. Rosanio, M. Ryan, A. Stutz, D. C. J. Spierings, A. Ward, A. E. Welsch, M. Xiao, W. Xu, C. Zhang, Q. Zhu, X. Zheng-Bradley, G. Jun, L. Ding, C. L. Koh, B. Ren, P. Flicek, K. Chen, M. B. Gerstein, P.-Y. Kwok, P. M. Lansdorp, G. Marth, J. Sebat, X. Shi, A. Bashir, K. Ye, S. E. Devine, M. Talkowski, R. E. Mills, T. Marschall, J. Korbel, E. E. Eichler, C. Lee, Multi-platform discovery of haplotype-resolved structural variation in human genomes. bioRxiv 193144 [Preprint]. 23

September 2017. https://doi.org/10.1101/193144.

51. B. P. Coe, K. Witherspoon, J. A. Rosenfeld, B. W. M. van Bon, A. T. Vulto-van Silfhout, P. Bosco, K. L. Friend, C. Baker, S. Buono, L. E. L. M. Vissers, J. H. Schuurs-Hoeijmakers, A. Hoischen, R. Pfundt, N. Krumm, G. L. Carvill, D. Li, D. Amaral, N. Brown, P. J. Lockhart, I. E. Scheffer, A. Alberti, M. Shaw, R. Pettinato, R. Tervo, N. de Leeuw, M. R. F. Reijnders, B. S. Torchia, H. Peeters, B. J. O'Roak, M. Fichera, J. Y. Hehir-Kwa, J. Shendure, H. C. Mefford, E. Haan, J. Gécz, B. B. de Vries, C. Romano, E. E. Eichler, Refining analyses of copy number variation identifies specific genes associated with developmental delay. *Nat. Genet.* **46**, 1063–1071 (2014). doi:10.1038/ng.3092 Medline

52. E. R. Sturgill, K. Aoki, P. H. H. Lopez, D. Colacurcio, K. Vajn, I. Lorenzini, S. Majić, W. H. Yang, M. Heffer, M. Tiemeyer, J. D. Marth, R. L. Schnaar, Biosynthesis of the major brain gangliosides GD1a and GT1b. *Glycobiology* **22**, 1289–1301 (2012). doi:10.1093/glycob/cws103 Medline

53. S. Herculano-Houzel, The human brain in numbers: A linearly scaled-up primate brain. *Front. Hum. Neurosci.* **3**, 31 (2009). doi:10.3389/neuro.09.031.2009 Medline

54. M. Y. Dennis, L. Harshman, B. J. Nelson, O. Penn, S. Cantsilieris, J. Huddleston, F. Antonacci, K. Penewit, L. Denman, A. Raja, C. Baker, K. Mark, M. Malig, N. Janke, C. Espinoza, H. A. F. Stessman, X. Nuttle, K. Hoekzema, T. A. Lindsay-Graves, R. K. Wilson, E. E. Eichler, The evolution and population diversity of human-specific segmental duplications. *Nat. Ecol. Evol.* **1**, 0069 (2017). doi:10.1038/s41559-016-0069 Medline

55. J. G. Camp, F. Badsha, M. Florio, S. Kanton, T. Gerber, M. Wilsch-Bräuninger, E. Lewitus, A. Sykes, W. Hevers, M. Lancaster, J. A. Knoblich, R. Lachmann, S. Pääbo, W. B. Huttner, B. Treutlein, Human cerebral organoids recapitulate gene expression programs of fetal neocortex development. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 15672–15677 (2015). Medline

56. F. Mora-Bermúdez, F. Badsha, S. Kanton, J. G. Camp, B. Vernot, K. Köhler, B. Voigt, K. Okita, T. Maricic, Z. He, R. Lachmann, S. Pääbo, B. Treutlein, W. B. Huttner, Differences and similarities between human and chimpanzee neural progenitors during cerebral cortex development. *eLife* **5**, 166 (2016). doi:10.7554/eLife.18683 Medline

57. A. A. Pollen, T. J. Nowakowski, J. Chen, H. Retallack, C. Sandoval-Espinosa, C. R. Nicholas, J. Shuga, S. J. Liu, M. C. Oldham, A. Diaz, D. A. Lim, A. A. Leyrat, J. A. West, A. R. Kriegstein, Molecular identity of human outer radial glia during cortical development. *Cell* **163**, 55–67 (2015). doi:10.1016/j.cell.2015.09.004 Medline

58. Z. He, D. Han, O. Efimova, P. Guijarro, Q. Yu, A. Oleksiak, S. Jiang, K. Anokhin, B. Velichkovsky, S. Grünewald, P. Khaitovich, Comprehensive transcriptome analysis of neocortical layers in humans, chimpanzees and macaques. *Nat. Neurosci.* **20**, 886–895 (2017). doi:10.1038/nn.4548 Medline

59. M. C. N. Marchetto, I. Narvaiza, A. M. Denli, C. Benner, T. A. Lazzarini, J. L. Nathanson, A. C. M. Paquola, K. N. Desai, R. H. Herai, M. D. Weitzman, G. W. Yeo, A. R. Muotri, F. H. Gage, Differential L1 regulation in pluripotent stem cells of humans and apes. *Nature* **503**, 525–529 (2013). doi:10.1038/nature12686 Medline

60. J. Korlach, G. Gedman, S. B. Kingan, C.-S. Chin, J. T. Howard, J.-N. Audet, L. Cantin, E. D. Jarvis, De novo PacBio long-read and phased avian genome assemblies correct and add to reference genes generated with intermediate and short reads. *Gigascience* **6**, 1–16 (2017). doi:10.1093/gigascience/gix085 Medline

61. M. V. Olson, When less is more: Gene loss as an engine of evolutionary change. *Am. J. Hum. Genet.* **64**, 18–23 (1999). doi:10.1086/302219 Medline

62. T. Marques-Bonet, J. M. Kidd, M. Ventura, T. A. Graves, Z. Cheng, L. W. Hillier, Z. Jiang, C. Baker, R. Malfavon-Borja, L. A. Fulton, C. Alkan, G. Aksay, S. Girirajan, P. Siswara, L. Chen, M. F. Cardone, A. Navarro, E. R. Mardis, R. K. Wilson, E. E. Eichler, A burst of segmental duplications in the genome of the African great ape ancestor. *Nature* **457**, 877–881 (2009). doi:10.1038/nature07744 Medline

63. X. Nuttle, G. Giannuzzi, M. H. Duyzend, J. G. Schraiber, I. Narvaiza, P. H. Sudmant, O. Penn, G. Chiatante, M. Malig, J. Huddleston, C. Benner, F. Camponeschi, S. Ciofi-Baffoni, H. A. F. Stessman, M. C. N. Marchetto, L. Denman, L. Harshman, C. Baker, A. Raja, K. Penewit, N. Janke, W. J. Tang, M. Ventura, L. Banci, F. Antonacci, J. M. Akey, C. T. Amemiya, F. H. Gage, A. Reymond, E. E. Eichler, Emergence of a *Homo sapiens*-specific gene family and chromosome 16p11.2 CNV susceptibility. *Nature* **536**, 205–209 (2016). doi:10.1038/nature19075 Medline

64. M. Jain, H. E. Olsen, B. Paten, M. Akeson, The Oxford Nanopore MinION: Delivery of nanopore sequencing to the genomics community. *Genome Biol.* **17**, 239 (2016). doi:10.1186/s13059-016-1103-0 Medline

65. E. Garrison, G. Marth, Haplotype-based variant detection from short-read sequencing. arXiv:1207.3907 [q-bio.GN] (17 July 2012).

66. T. J. Nowakowski, A. Bhaduri, A. A. Pollen, B. Alvarado, M. A. Mostajo-Radji, E. Di Lullo, M. Haeussler, C. Sandoval-Espinosa, S. J. Liu, D. Velmeshev, J. R. Ounadjela, J. Shuga, X. Wang, D. A. Lim, J. A. West, A. A. Leyrat, W. J. Kent, A. R. Kriegstein, Spatiotemporal gene expression trajectories reveal developmental hierarchies of the human cortex. *Science* **358**, 1318–1323 (2017). doi:10.1126/science.aap8809 Medline

67. N. C. Durand, J. T. Robinson, M. S. Shamim, I. Machol, J. P. Mesirov, E. S. Lander, E. L. Aiden, Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Syst.* **3**, 99–101 (2016). doi:10.1016/j.cels.2015.07.012 Medline

68. S. Koren, B. P. Walenz, K. Berlin, J. R. Miller, N. H. Bergman, A. M. Phillippy, Canu: Scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation. *Genome Res.* **27**, 722–736 (2017). doi:10.1101/gr.215087.116 Medline

69. P. Lichter, C. J. Tang, K. Call, G. Hermanson, G. A. Evans, D. Housman, D. C. Ward, High-resolution mapping of human chromosome 11 by in situ hybridization with cosmid clones. *Science* **247**, 64–69 (1990). doi:10.1126/science.2294592 Medline

70. B. Paten, D. Earl, N. Nguyen, M. Diekhans, D. Zerbino, D. Haussler, Cactus: Algorithms for genome multiple sequence alignment. *Genome Res.* **21**, 1512–1528 (2011). doi:10.1101/gr.123356.111 Medline

71. S. König, L. W. Romoth, L. Gerischer, M. Stanke, Simultaneous gene finding in multiple genomes. *Bioinformatics* **32**, 3388–3395 (2016). Medline

72. K. Shekhar, S. W. Lapan, I. E. Whitney, N. M. Tran, E. Z. Macosko, M. Kowalczyk, X. Adiconis, J. Z. Levin, J. Nemesh, M. Goldman, S. A. McCarroll, C. L. Cepko, A. Regev, J. R. Sanes, Comprehensive classification of retinal bipolar neurons by single-cell transcriptomics. *Cell* **166**, 1308–1323.e30 (2016). doi:10.1016/j.cell.2016.07.054 Medline

73. M. J. Chaisson, G. Tesler, Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): Application and theory. *BMC Bioinformatics* **13**, 238 (2012). doi:10.1186/1471-2105-13-238 Medline

74. K. Khelik, K. Lagesen, G. K. Sandve, T. Rognes, A. J. Nederbragt, NucDiff: In-depth characterization and annotation of differences between two sets of DNA sequences. *BMC Bioinformatics* **18**, 338 (2017). doi:10.1186/s12859-017-1748-z Medline

75. R. Stanyon, M. Rocchi, O. Capozzi, R. Roberto, D. Misceo, M. Ventura, M. F. Cardone, F. Bigoni, N. Archidiacono, Primate chromosome evolution: Ancestral karyotypes, marker order and neocentromeres. *Chromosome Res.* **16**, 17–39 (2008). doi:10.1007/s10577-007-1209-z Medline

76. M. J. P. Chaisson, R. K. Wilson, E. E. Eichler, Genetic variation and the de novo assembly of human genomes. *Nat. Rev. Genet.* **16**, 627–640 (2015). doi:10.1038/nrg3933 Medline

77. J. Huddleston, M. J. P. Chaisson, K. M. Steinberg, W. Warren, K. Hoekzema, D. Gordon, T. A. Graves-Lindsay, K. M. Munson, Z. N. Kronenberg, L. Vives, P. Peluso, M. Boitano, C.-S. Chin, J. Korlach, R. K. Wilson, E. E. Eichler, Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome Res.* **27**, 677–685 (2017). doi:10.1101/gr.214007.116 Medline

78. M. H. Weissensteiner, A. W. C. Pang, I. Bunikis, I. Höijer, O. Vinnere-Petterson, A. Suh, J. B. W. Wolf, Combination of short-read, long-read, and optical mapping assemblies reveals large-scale tandem repeat arrays with population genetic implications. *Genome Res.* **27**, 697–708 (2017). doi:10.1101/gr.215095.116 Medline

79. W.-B. Jiao, G. G. Accinelli, B. Hartwig, C. Kiefer, D. Baker, E. Severing, E.-M. Willing, M. Piednoel, S. Woetzel, E. Madrid-Herrero, B. Huettel, U. Hümann, R. Reinhard, M. A. Koch, D. Swan, B. Clavijo, G. Coupland, K. Schneeberger, Improving and correcting the contiguity of long-read genome assemblies of three plant species using optical mapping and chromosome conformation capture data. *Genome Res.* **27**, 778–786 (2017). doi:10.1101/gr.213652.116 Medline

80. M. Ventura, C. R. Catacchio, C. Alkan, T. Marques-Bonet, S. Sajjadian, T. A. Graves, F. Hormozdiari, A. Navarro, M. Malig, C. Baker, C. Lee, E. H. Turner, L. Chen, J. M. Kidd, N. Archidiacono, J. Shendure, R. K. Wilson, E. E. Eichler, Gorilla genome structural variation reveals evolutionary parallelisms with chimpanzee. *Genome Res.* **21**, 1640–1649 (2011). doi:10.1101/gr.124461.111 Medline

81. M. Ventura, F. Antonacci, M. F. Cardone, R. Stanyon, P. D'Addabbo, A. Cellamare, L. J. Sprague, E. E. Eichler, N. Archidiacono, M. Rocchi, Evolutionary formation of new centromeres in macaque. *Science* **316**, 243–246 (2007). doi:10.1126/science.1140615 Medline

82. V. Ramani, D. A. Cusanovich, R. J. Hause, W. Ma, R. Qiu, X. Deng, C. A. Blau, C. M.

Disteche, W. S. Noble, J. Shendure, Z. Duan, Mapping 3D genome architecture through in situ DNase Hi-C. *Nat. Protoc.* **11**, 2104–2121 (2016). doi:10.1038/nprot.2016.126 Medline

83. M. Lek, K. J. Karczewski, E. V. Minikel, K. E. Samocha, E. Banks, T. Fennell, A. H. O'Donnell-Luria, J. S. Ware, A. J. Hill, B. B. Cummings, T. Tukiainen, D. P. Birnbaum, J. A. Kosmicki, L. E. Duncan, K. Estrada, F. Zhao, J. Zou, E. Pierce-Hoffman, J. Berghout, D. N. Cooper, N. Deflaux, M. DePristo, R. Do, J. Flannick, M. Fromer, L. Gauthier, J. Goldstein, N. Gupta, D. Howrigan, A. Kiezun, M. I. Kurki, A. L. Moonshine, P. Natarajan, L. Orozco, G. M. Peloso, R. Poplin, M. A. Rivas, V. Ruano-Rubio, S. A. Rose, D. M. Ruderfer, K. Shakir, P. D. Stenson, C. Stevens, B. P. Thomas, G. Tiao, M. T. Tusie-Luna, B. Weisburd, H.-H. Won, D. Yu, D. M. Altshuler, D. Ardissino, M. Boehnke, J. Danesh, S. Donnelly, R. Elosua, J. C. Florez, S. B. Gabriel, G. Getz, S. J. Glatt, C. M. Hultman, S. Kathiresan, M. Laakso, S. McCarroll, M. I. McCarthy, D. McGovern, R. McPherson, B. M. Neale, A. Palotie, S. M. Purcell, D. Saleheen, J. M. Scharf, P. Sklar, P. F. Sullivan, J. Tuomilehto, M. T. Tsuang, H. C. Watkins, J. G. Wilson, M. J. Daly, D. G. MacArthur; Exome Aggregation Consortium, Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016). doi:10.1038/nature19057 Medline

84. P. Danecek, A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth, S. T. Sherry, G. McVean, R. Durbin; 1000 Genomes Project Analysis Group, The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011). doi:10.1093/bioinformatics/btr330 Medline

85. C. D. Braastad, H. Hovhannisyan, A. J. van Wijnen, J. L. Stein, G. S. Stein, Functional characterization of a human histone gene cluster duplication. *Gene* **342**, 35–40 (2004). doi:10.1016/j.gene.2004.07.036 Medline

86. F. Hormozdiari, M. K. Konkel, J. Prado-Martinez, G. Chiatante, I. H. Herraez, J. A. Walker, B. Nelson, C. Alkan, P. H. Sudmant, J. Huddleston, C. R. Catacchio, A. Ko, M. Malig, C. Baker, T. Marques-Bonet, M. Ventura, M. A. Batzer, E. E. Eichler; Great Ape Genome Project, Rates and patterns of great ape retrotransposition. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 13457–13462 (2013). doi:10.1073/pnas.1310914110 Medline

87. L. Ramsay, M. C. Marchetto, M. Caron, S.-H. Chen, S. Busche, T. Kwan, T. Pastinen, F. H. Gage, G. Bourque, Conserved expression of transposon-derived non-coding transcripts in primate stem cells. *BMC Genomics* **18**, 214 (2017). doi:10.1186/s12864-017-3568-y Medline

88. Y.-S. Chan, J. Göke, J.-H. Ng, X. Lu, K. A. U. Gonzales, C.-P. Tan, W.-Q. Tng, Z.-Z. Hong, Y.-S. Lim, H.-H. Ng, Induction of a human pluripotent state with distinct regulatory circuitry that resembles preimplantation epiblast. *Cell Stem Cell* **13**, 663–675 (2013). doi:10.1016/j.stem.2013.11.015 Medline

89. G. Renaud, U. Stenzel, T. Maricic, V. Wiebe, J. Kelso, deML: Robust demultiplexing of Illumina sequences using a likelihood-based approach. *Bioinformatics* **31**, 770–772 (2015). doi:10.1093/bioinformatics/btu719 Medline

90. A. M. Bolger, M. Lohse, B. Usadel, Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014). doi:10.1093/bioinformatics/btu170 Medline

91. A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, T. R. Gingeras, STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013). doi:10.1093/bioinformatics/bts635 Medline

92. S. W. Hartley, J. C. Mullikin, QoRTs: A comprehensive toolset for quality control and data processing of RNA-Seq experiments. *BMC Bioinformatics* **16**, 224 (2015). doi:10.1186/s12859-015-0670-5 Medline

93. S. P. Gordon, E. Tseng, A. Salamov, J. Zhang, X. Meng, Z. Zhao, D. Kang, J. Underwood, I. V. Grigoriev, M. Figueroa, J. S. Schilling, F. Chen, Z. Wang, Widespread polycistronic transcripts in fungi revealed by single-molecule mRNA sequencing. *PLOS ONE* **10**, e0132628 (2015). doi:10.1371/journal.pone.0132628 Medline

94. N. L. Bray, H. Pimentel, P. Melsted, L. Pachter, Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016). doi:10.1038/nbt.3519 Medline

95. J. Lonsdale, J. Thomas, M. Salvatore, R. Phillips, E. Lo, S. Shad, R. Hasz, G. Walters, F. Garcia, N. Young, B. Foster, M. Moser, E. Karasik, B. Gillard, K. Ramsey, S. Sullivan, J. Bridge, H. Magazine, J. Syron, J. Fleming, L. Siminoff, H. Traino, M. Mosavel, L. Barker, S. Jewell, D. Rohrer, D. Maxim, D. Filkins, P. Harbach, E. Cortadillo, B. Berghuis, L. Turner, E. Hudson, K. Feenstra, L. Sobin, J. Robb, P. Branton, G. Korzeniewski, C. Shive, D. Tabor, L. Qi, K. Groch, S. Nampally, S. Buia, A. Zimmerman, A. Smith, R. Burges, K. Robinson, K. Valentino, D. Bradbury, M. Cosentino, N. Diaz-Mayoral, M. Kennedy, T. Engel, P. Williams, K. Erickson, K. Ardlie, W. Winckler, G. Getz, D. DeLuca, D. MacArthur, M. Kellis, A. Thomson, T. Young, E. Gelfand, M. Donovan, Y. Meng, G. Grant, D. Mash, Y. Marcus, M. Basile, J. Liu, J. Zhu, Z. Tu, N. J. Cox, D. L. Nicolae, E. R. Gamazon, H. K. Im, A. Konkashbaev, J. Pritchard, M. Stevens, T. Flutre, X. Wen, E. T. Dermitzakis, T. Lappalainen, R. Guigo, J. Monlong, M. Sammeth, D. Koller, A. Battle, S. Mostafavi, M. McCarthy, M. Rivas, J. Maller, I. Rusyn, A. Nobel, F. Wright, A. Shabalin, M. Feolo, N. Sharopova, A. Sturcke, J. Paschal, J. M. Anderson, E. L. Wilder, L. K. Derr, E. D. Green, J. P. Struewing, G. Temple, S. Volpi, J. T. Boyer, E. J. Thomson, M. S. Guyer, C. Ng, A. Abdallah, D. Colantuoni, T. R. Insel, S. E. Koester, A. R. Little, P. K. Bender, T. Lehner, Y. Yao, C. C. Compton, J. B. Vaught, S. Sawyer, N. C. Lockhart, J. Demchok, H. F. Moore; GTEx Consortium, The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013). doi:10.1038/ng.2653 Medline

96. J. Harrow, A. Frankish, J. M. Gonzalez, E. Tapanari, M. Diekhans, F. Kokocinski, B. L. Aken, D. Barrell, A. Zadissa, S. Searle, I. Barnes, A. Bignell, V. Boychenko, T. Hunt, M. Kay, G. Mukherjee, J. Rajan, G. Despacio-Reyes, G. Saunders, C. Steward, R. Harte, M. Lin, C. Howald, A. Tanzer, T. Derrien, J. Chrast, N. Walters, S. Balasubramanian, B. Pei, M. Tress, J. M. Rodriguez, I. Ezkurdia, J. van Baren, M. Brent, D. Haussler, M. Kellis, A. Valencia, A. Reymond, M. Gerstein, R. Guigó, T. J. Hubbard, GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Res.* **22**, 1760–1774 (2012). doi:10.1101/gr.135350.111 Medline

97. M. Stanke, M. Diekhans, R. Baertsch, D. Haussler, Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* **24**, 637–644 (2008). doi:10.1093/bioinformatics/btn013 Medline

98. M. Stanke, O. Keller, I. Gunduz, A. Hayes, S. Waack, B. Morgenstern, AUGUSTUS: Ab initio prediction of alternative transcripts. *Nucleic Acids Res.* **34**, W435–W439 (2006). doi:10.1093/nar/gkl200 Medline

99. M. Stanke, O. Schöffmann, B. Morgenstern, S. Waack, Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics* **7**, 62 (2006). doi:10.1186/1471-2105-7-62 Medline

100. S. Mallick, H. Li, M. Lipson, I. Mathieson, M. Gymrek, F. Racimo, M. Zhao, N. Chennagiri, S. Nordenfelt, A. Tandon, P. Skoglund, I. Lazaridis, S. Sankararaman, Q. Fu, N. Rohland, G. Renaud, Y. Erlich, T. Willems, C. Gallo, J. P. Spence, Y. S. Song, G. Poletti, F. Balloux, G. van Driem, P. de Knijff, I. G. Romero, A. R. Jha, D. M. Behar, C. M. Bravi, C. Capelli, T. Hervig, A. Moreno-Estrada, O. L. Posukh, E. Balanovska, O. Balanovsky, S. Karachanak-Yankova, H. Sahakyan, D. Toncheva, L. Yepiskoposyan, C. Tyler-Smith, Y. Xue, M. S. Abdullah, A. Ruiz-Linares, C. M. Beall, A. Di Rienzo, C. Jeong, E. B. Starikovskaya, E. Metspalu, J. Parik, R. Villems, B. M. Henn, U. Hodoglugil, R. Mahley, A. Sajantila, G. Stamatoyannopoulos, J. T. S. Wee, R. Khusainova, E. Khusnutdinova, S. Litvinov, G. Ayodo, D. Comas, M. F. Hammer, T. Kivisild, W. Klitz, C. A. Winkler, D. Labuda, M. Bamshad, L. B. Jorde, S. A. Tishkoff, W. S. Watkins, M. Metspalu, S. Dryomov, R. Sukernik, L. Singh, K. Thangaraj, S. Pääbo, J. Kelso, N. Patterson, D. Reich, The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* **538**, 201–206 (2016). doi:10.1038/nature18964 Medline

101. P. H. Sudmant, S. Mallick, B. J. Nelson, F. Hormozdiari, N. Krumm, J. Huddleston, B. P. Coe, C. Baker, S. Nordenfelt, M. Bamshad, L. B. Jorde, O. L. Posukh, H. Sahakyan, W. S. Watkins, L. Yepiskoposyan, M. S. Abdullah, C. M. Bravi, C. Capelli, T. Hervig, J. T. S. Wee, C. Tyler-Smith, G. van Driem, I. G. Romero, A. R. Jha, S. Karachanak-Yankova, D. Toncheva, D. Comas, B. Henn, T. Kivisild, A. Ruiz-Linares, A. Sajantila, E. Metspalu, J. Parik, R. Villems, E. B. Starikovskaya, G. Ayodo, C. M. Beall, A. Di Rienzo, M. F. Hammer, R. Khusainova, E. Khusnutdinova, W. Klitz, C. Winkler, D. Labuda, M. Metspalu, S. A. Tishkoff, S. Dryomov, R. Sukernik, N. Patterson, D. Reich, E. E. Eichler, Global diversity, population stratification, and selection of human copy-number variation. *Science* **349**, aab3761 (2015). doi:10.1126/science.aab3761 Medline

102. R. D. George, G. McVicker, R. Diederich, S. B. Ng, A. P. MacKenzie, W. J. Swanson, J. Shendure, J. H. Thomas, Trans genomic capture and sequencing of primate exomes reveals new targets of positive selection. *Genome Res.* **21**, 1686–1694 (2011). doi:10.1101/gr.121327.111 Medline

103. K. Mohajeri, S. Cantsilieris, J. Huddleston, B. J. Nelson, B. P. Coe, C. D. Campbell, C. Baker, L. Harshman, K. M. Munson, Z. N. Kronenberg, M. Kremitzki, A. Raja, C. R. Catacchio, T. A. Graves, R. K. Wilson, M. Ventura, E. E. Eichler, Interchromosomal core duplicons drive both evolutionary instability and disease susceptibility of the Chromosome 8p23.1 region. *Genome Res.* **26**, 1453–1467 (2016). doi:10.1101/gr.211284.116 Medline

104. Z. Jiang, H. Tang, M. Ventura, M. F. Cardone, T. Marques-Bonet, X. She, P. A. Pevzner, E. E. Eichler, Ancestral reconstruction of segmental duplications reveals punctuated cores of human genome evolution. *Nat. Genet.* **39**, 1361–1368 (2007). doi:10.1038/ng.2007.9 Medline