science.sciencemag.org/content/370/6523/eabc6617/suppl/DC1

Supplementary Materials for

Sequence diversity analyses of an improved rhesus macaque genome enhance its biomedical utility

Wesley C. Warren*, R. Alan Harris, Marina Haukness, Ian T. Fiddes, Shwetha C. Murali, Jason Fernandes, Philip C. Dishuck, Jessica M. Storer, Muthuswamy Raveendran, LaDeana W. Hillier, David Porubsky, Yafei Mao, David Gordon, Mitchell R. Vollger, Alexandra P. Lewis, Katherine M. Munson, Elizabeth DeVogelaere, Joel Armstrong, Mark Diekhans, Jerilyn A. Walker, Chad Tomlinson, Tina A. Graves-Lindsay, Milinn Kremitzki, Sofie R. Salama, Peter A. Audano, Merly Escalona, Nicholas W. Maurer, Francesca Antonacci, Ludovica Mercuri, Flavia A. M. Maggiolini, Claudia Rita Catacchio, Jason G. Underwood, David H. O'Connor, Ashley D. Sanders, Jan O. Korbel, Betsy Ferguson, H. Michael Kubisch, Louis Picker, Ned H. Kalin, Douglas Rosene, Jon Levine, David H. Abbott, Stanton B. Gray, Mar M. Sanchez, Zsofia A. Kovacs-Balint, Joseph W. Kemnitz, Sara M. Thomasy, Jeffrey A. Roberts, Erin L. Kinnally, John P. Capitanio, J. H. Pate Skene, Michael Platt, Shelley A. Cole, Richard E. Green, Mario Ventura, Roger W. Wiseman, Benedict Paten, Mark A. Batzer, Jeffrey Rogers*, Evan E. Eichler*

*Corresponding author. Email: warrenwc@missouri.edu (W.C.W.); jr13@bcm.edu (J.R.); eee@gs.washington.edu (E.E.E.)

Published 18 December 2020, *Science* **370**, eabc6617 (2020)
DOI: 10.1126/science.abc6617

This PDF file includes:

Materials and Methods
Supplementary Text
Figs. S1 to S30
Captions for Tables S1 to S29
Captions for Database S1
Description of Repeat Sequences Database
References (*39*–*80*)

Other Supplementary Material for this manuscript includes the following:
(available at science.sciencemag.org/content/370/6523/eabc6617/suppl/DC1)

Tables S1 to S29 (Excel)
Database S1 (Excel)
Repeat Sequences Database (zipped archive file)
MDAR Reproducibility Checklist (PDF)

Materials and Methods

DNA sequencing and assembly. We isolated high molecular weight DNA (HMW-DNA) from a single female rhesus macaque (Coriell sample AG071017) of Indian origin using the MagAttract HMW-DNA Kit (Qiagen; manufacturer's protocol) and generated 45-fold paired-end whole-genome sequencing (WGS; 150 bp length) using the Illumina HiSeq X instrument. We mapped AG71017 Illumina data to Mmul_8.0.1 along with 133 previously resequenced macaques (*3*) and identified single-nucleotide variants (SNVs) using standard Genome Analysis Toolkit (GATK) (*39*) best practices. We confirmed the Indian-origin identity of AG07107 by using 3D and 2D Principal Components Analysis (PCA) (Figs. 6A and S28). Also, a standard metaphase chromosomal spread of AG07107 fibroblast cells was generated to confirm a normal karyotype (Fig. S29). We generated 66-fold coverage of SMRT sequencing data using the PacBio RS II instrument (V2 chemistry) for a total of 198 Gbp with an average subread length of 13,629 bp. SMRT sequences were error corrected (119 Gbp, 39.7X coverage and average length of 9,801 bp) and assembled with FALCON-integrate version 1.7.5 using the following parameters: max_diff 120, max_cov 120, min_cov 1, min_seed_length 11, followed by contig base error correction with Quiver (*40*). We further corrected indel base errors by aligning AG07107 Illumina WGS using BWA-MEM version 0.7.10 and a custom FreeBayes-based pipeline described in Kronenberg et al. (*5*).

Assembly scaffolding and chromosome assignment. We applied an iterative procedure to scaffold error-corrected contigs and constructed a draft assembly. First, we generated a Bionano Saphyr physical map (Bionano Genomics) of AG07107 and then built a *de novo* hybrid scaffold structure using the Bionano Access software as previously described (*41*). Sequence merging tools (GRC) were used to merge any possible sequence overlaps, order and orient contigs, and identify possible misassemblies, all supported by Bionano data (*42*). Next, we adjusted assembly scaffold structure using AG07107 proximity ligation sequence data (Dovetail Genomics Hi-C) following the HiRise protocol (*43*). Hi-C libraries were constructed as described previously (*44*) to ~30-fold coverage of Illumina paired-end reads (150 bp). We aligned the Hi-C data to the input assembly (hybrid scaffolds) with a modified version of SNAP (*45*) marking PCR duplicates with Novosort (http://www.novocraft.com/products/novosort/) and developed a likelihood model to guide assembly decisions related to breaking misjoins as well as joining and orienting contigs within scaffolds. Final gap closure steps were then applied to this Hi-C altered assembly structure (*18*).

To assign highly contiguous scaffolds to individual chromosomes, we used the nucmer program in MUMmer4 (*46*) and performed genome-wide alignments using these parameters (-l 100 -c 200 run with both -mum and -maxmatch) to the Mmul_8.0.1 reference as a guide. Secondary alignments to human (GRChg38.12) and *Macaca fascicularis* references (macFas5) were also considered. We manually investigated interchromosomal discrepancies and, if sufficient evidence allowed, sequence breaks were made to correct these assembly errors. We applied a series of evidence-based processes to finalize chromosome sequence order and orientation. By convention, chromosomes are typically represented and organized starting with their short arm. In some chromosomes, an inversion or a centromere repositioning event may change the "cytogenetic" orientation of the chromosome. We note that this convention was not followed in the previous Mmul_8.0.1 assembly where chromosomes 1 (1), 2 (3), 4 (6), 10 (20/22), 13 (2p) and 18 (18) (human phylogenetic group in parenthesis) are not represented in this way. In Mmul_10, however, we have corrected this so these chromosomes are in the opposite orientation with respect to Mmul_8.0.1. Whole-genome BLAST alignments using the BLAST parameters: blastn -best_hit_overhang 0.1 -best_hit_score_edge 0.1 -evalue 0.0001 -soft_masking true -task megablast -word_size 28 were used to assess differences between Mmul_10 and Mmul_8.0.1 in completeness. This step is in conjunction with the use of the precomputed WindowMasker masked regions. The final assembled chromosomes are available through the GenBank assembly accession GCF_003339765.1. The Y chromosome was added for completeness and was based on an independent assembly of bacterial artificial chromosome (BAC) clones from a male macaque (*19*).

Contiguity and directionality assessment. Regional changes in assembly directionality were detected using strand sequencing (Strand-seq) (*47*). Strand-seq is a single-cell sequencing technique able to track directionality of individual homologous chromosomes based on mapping short reads, which originate from single-stranded DNA, to a *de novo* assembly. First, we aligned Strand-seq libraries to both Mmul_8.0.1 and Mmul_10 macaque assemblies using BWA (version 0.7.15-r1140) with default settings. Duplicate reads and reads with low mapping quality value (QV < 10) were removed prior to our analysis and we selected 60 informative Strand-seq libraries. We used breakpointR (Bioconductor 3.10) to detect recurrent changes in read directionality by compiling short reads across all informative Strand-seq libraries. A regional change in directionality was detected as a switch in reads mapping to the plus strand ('Crick') or negative strand ('Watson') of a *de novo* assembly. We calculated the percentage of genes that overlap with the detected misassemblies in both Mmul reference assemblies based on UCSC Table Browser annotations (Ensembl genes – Mmul_8.0.1 and RefSeq genes – Mmul_10). We used Strand-seq data to interrogate potentially problematic regions, such as the KIR region where we aligned Strand-seq data specifically to the chromosome 19 scaffold (CM014354.1) and visualized structural differences using breakpointR (Bioconductor 3.10) for all informative Strand-seq libraries (n = 60). All sequence resources used for computational tasks associated with *de novo* assembly, scaffolding, and genome accuracy assessments are available (Table S1).

BAC-end sequence (BES) analysis. We mapped Sanger BES (CHORI-250) to the Mmul_10 assembly using MegaBLAST 2.2.9 and identified 60,015 BES that were concordant by predicted length of insert and orientation to the assembly. We identified 104,238 SNV and 52,423 indel differences based on 51 Mbp of aligned sequences with a minimum PHRED ≥ 40 from the Sanger traces.

Iso-Seq. Full-length cDNA was prepared and sequenced from various tissue sources (Table S12). For prefrontal cortex brain and testes, total RNA was isolated from a rhesus male (Oregon National Primate Research Center) euthanized at seven years of age. The male received no treatments and was healthy at the time of death. RNA was prepared using the Qiagen AllPrep DNA/RNA/miRNA Universal Kit. In addition, mRNA was prepared from a macaque induced pluripotent stem cell (iPSC) line (*48*) and fetal brain material from developmental stage E80 somatosensory cortex, E100 parietal lobe, and E100 anterior cerebellum (California National Primate Research Center). For iPSCs and fetal brain, Iso-Seq library production and sequencing was performed as previously described (*49*) with the following modifications: samples were cDNA amplified using standard non-barcoded primers and optionally barcoded at the SMRTbell library step using barcoded adapter ligation. In lieu of strict size fractionation, size rebalancing was performed using sequential 0.4X/1X AMPure PB bead washes and repooling equal molar amounts of the two elutions. Samples were run on 10 total SMRT Cell 1Ms on the Sequel II platform. Prefrontal cortex brain and testes samples were processed with the Iso-Seq Express protocol (https://www.pacb.com/wp-content/uploads/Procedure-Checklist-Iso-Seq-Express-Template-Preparation-for-Sequel-and-Sequel-II-Systems.pdf) and barcoded using barcoded adapters for pooling and sequencing on one SMRT Cell 8M on the Sequel II platform. Collected data was optionally demultiplexed, then analyzed with circular consensus sequencing (CCS) and the Iso-Seq analysis pipeline to generate full-length non-chimeric (FLNC) reads ensuring each has a poly-A tail, plus a single 3' and 5' primer signal for downstream analysis. Collected data was optionally demultiplexed with lima (demultiplex barcoding), then analyzed with CCS with a requirement of 1 minimum pass and at least 0.9 identity (--minPasses 1 --min-rq 0.9). The Iso-Seq analysis pipeline was used to generate FLNC reads ensuring each has a poly-A tail, plus a single 3' and 5' primer signal for downstream analysis.

Gene annotation. We performed gene annotation with the Comparative Annotation Toolkit (CAT) (*50*). CAT leverages whole-genome alignments to transfer annotations from one source genome to one or more target genomes (*51*). CAT also runs AUGUSTUS (*52*) in both the comparative gene prediction mode (*53*) and in a single-genome mode that utilizes Iso-Seq data to predict alternative isoforms. CAT then combines all of these annotation methods into a final consensus annotation set that represents orthology relationships as well as species-specific information. To confirm the CAT results, additional gene annotation for the Mmul_10 assembly was generated using standard NCBI (*54*) and Ensembl (*55*) pipelines, including masking of repeats prior to *ab initio* gene predictions, for evidence-supported gene-model building. All annotation processes used publicly available RNA-seq and Iso-Seq data from diverse tissue sources (Table S12) (*20*). We identified putative novel exons and splice junctions in the AugustusPB annotation set by using the Cactus alignment to project AugustusPB predictions back to human. Here, we define novel exons as those annotated in Mmul_10 whose orthologous DNA in human is either partially or completely missing. We required that putative novel exon predictions be at least 30 bp long and have at least three supporting Iso-Seq reads.

Segmental duplication analyses. To detect sequence-resolved SDs in the Mmul_10 assembly, we applied the whole-genome analysis comparison (WGAC) method (*36*). This method detects duplications by generating pairwise alignments of ≥1 kbp at ≥90% sequence identity, excluding repeat-masked sequence (RepeatMasker 3.3.0 using the union of primate and mammal libraries). We also used excess read depth to identify duplications not properly resolved in the assembled genomes where allelic and paralogous sequence reads had been inadvertently collapsed during the assembly process. PacBio continuous long reads were uniquely mapped to the assembly with BLASR (blasr $READS $ASM -sa $ASMSA -sdpTupleSize 13 -sdpMaxAnchorsPerPosition 10 -maxMatch 25 -minMapQV 30 -bestn 2 -advanceExactMatches 15 -clipping subread –sam). Assembly collapses were defined as regions >=15 kbp where sequence read depth exceeds the mean coverage by at least three standard deviations, calculated across 100 bp windows excluding the top and bottom five percent of read depth values (*37*). Windows with >75% RepeatMasker content were excluded in the 15 kbp length requirement.

After detection of collapsed regions of the assembly, we applied Segmental Duplication Assembler (SDA) (*26*) to define paralogous sequence variants and correlation clustering to partition reads into groups corresponding to individual paralogs. The resulting SDA contigs were locally assembled with Canu (*56*), assessed for sequence read depth, and error corrected with Arrow. We annotated SDA-resolved contigs by mapping FLNC reads, obtained from rhesus macaque-derived iPSCs, brain, and testis from the Iso-Seq experiments. The higher accuracy of the contigs and FLNC reads allowed for duplicate genes to be ambiguously assigned to highly identical SDA contigs. We aligned FLNC reads separately to both the Mmul_10 assembly and the SDA-resolved contigs with minimap2 (*57*) and compared gap-compressed percent identity between the alignments for each read. We assigned gene names to each Iso-Seq read based on best alignment to GRCh38 RefSeq and CHESS 2.2 (*58*) annotations and performed a two-sided Wilcoxon signed-rank test on the percent identity difference of alignments to SDA and the original assembly, grouped by SDA contig and gene.

FISH validation of duplications. We selected 19 regions from Mmul_10 that showed evidence of collapsed SDs (Table S23); 17 out of 19 of the regions also showed evidence of SDs by WGAC (*36*). We tested each region by interphase and metaphase FISH on a female rhesus macaque lymphoblast cell line (*Macaca mulatta*, MMU1). We selected representative large-insert BACs (CHORI-250) based on mapping of BES data to MmuI_10 and selected only BACs where at least 40 kbp corresponded to the putative duplication. We examined SD distribution (intrachromosomal compared to interchromosomal) and the dispersal pattern (pericentromeric, subtelomeric and interstitial). One probe (CH250-489A14) mapping to chromosome 17p was selected as a negative control mapping to a unique region of the macaque genome and showed the expected single FISH signal.

Repetitive sequence analyses. We analyzed and compared repeat content of macaque genomes using RepeatMasker (RepeatMasker-Open-4.0; accessed Nov 2019) and the Dfam 3.0 repeat library (*59*). We categorized common elements into broad types (DNA transposons, LTR transposons, non-LTR transposons), as well as more specific categories (e.g., LINE/L1, LINE/L2, etc.). We classified full-length mobile element insertions (MEIs) from RepeatMasker output using a customized python script. Next, we defined full-length *Alu* repeats within a start position of no less than 4 bp and an end position not shorter than 267 bp; full-length LINE-1 elements were ≥6000 bp. Then we extracted 600 bp of 5' and 3' flanking sequence and compared to other primate genomes in a sequential BLAT (*60*): human (*Homo sapiens*; hg38) followed by the olive baboon (*Papio anubis*; Panu3.0; from NCBI). We determined lineage specificity by assessing presence or absence in the target genomes.

We assigned lineage specific *Alu* and full-length LINE elements to subfamilies using COSEG (www.repeatmasker.org/COSEGDownload.html) for both Mmul_8.0.1 and Mmul_10 genome assemblies. Briefly, full-length MEIs were aligned via crossmatch (www.phrap.org/phredphrapconsed.html) with the default settings to the *AluY* (*61*) and 3' end of the L1PA5 elements (*59*), respectively. COSEG was then used to group *Alu* subfamilies and determine subfamily structure. The middle A-rich region of the *AluY* subfamily consensus sequence was excluded from analysis when determining *Alu* subfamily assignment, whereas two or three subfamily-specific diagnostic mutations were used to distinguish *Alu* and LINE1 element subfamilies. We considered a distinct *Alu* or LINE1 subfamily as those with at least ten members and performed a network analysis of MEI subfamilies using Gephi (v0.9.1) (*62*) (Tables S24 and S25). For the *Alu* and LINE1 networks, we obtained Old World monkey (OWM) consensus sequences from Dfam 3.0 (*59*) and RepBase (*63*) as well as previously reported rhesus macaque subfamilies (*64*) to create an enhanced RepeatMasker library. We compared MEI content between Mmul_8.0.1 and Mmul_10 focusing on all LTR elements from RepeatMasker and full-length *Alu* and LINE1 elements obtained from the COSEG analysis. We applied the UCSC Genome Browser liftOver tool between the Mmul_10 and Mmul_8.0.1 assemblies, and parsed these using a custom python script with additional follow up for sequences deleted in Mmul_8.0.1 but present in Mmul_10.

Analysis of the 5' UTR OWM LINE1 elements. Primate-specific LINE1 families (L1PA) evolve in lineage-specific waves; old L1PA families are denoted with higher numbers (e.g., L1PA7, L1PA6), while L1PA5 was active in the human–rhesus common ancestor. L1RS (rhesus-specific) families evolved specifically within the OWM lineage after divergence from the humans (and are distinct from L1PA4 and younger L1 elements in apes). We compared L1RS RepeatMasker annotations for a variety of OWM genomes (macFas5, papAnu4, and rhiRox1) to annotations in the two Indian-origin macaque genome assemblies. We also used the youngest human L1PA element, L1HS (human-specific L1s), hg38 annotations as an outgroup. Full-length elements (>6000 bp) of all L1_RS elements were extracted and aligned using BLAT to a consensus version of L1PA5 (ancestrally active at human–OWM divergence) on the UCSC Repeat Browser (*65*). We computed coverage tracks based on L1RS alignments noting drops in coverage most notably in the 5' UTR, demonstrating changes from the L1PA5 consensus (all L1RS elements are originally derived from L1PA5). These alignments can be viewed on the UCSC Repeat Browser: https://genome.ucsc.edu/s/jdf2001/L1RS.

Whole-genome sequencing of rhesus research population. We received biomaterials (DNA, blood or tissue samples) from nine US rhesus macaque research colonies (Table S26) for the purpose of WGS. The samples include the study animal used for version Mmul_8.0.1 and AG07107 used for preparation of Mmul_10. Among those 853 rhesus macaques, there were 810 rhesus macaques of Indian origin, 12 of Chinese origin, and 31 that were initially reported as Indian origin but were shown to be Chinese rhesus based on WGS genotyping and PCA. When collaborating institutions provided EDTA-treated blood or tissue (generally brain, liver or spleen) for a given animal, DNA was purified using Puregene blood or tissue kits (Gentra). WGS was performed over an eight-year period. Consequently, as technology improved, the sequencing platforms used to generate next-generation sequencing reads for this dataset progressed as follows: Illumina HiSeq 2000, HiSeq Rapid 2500, HiSeq X, and NovaSeq platforms, generating 2 X 100 bp or 2 X 150 bp paired-end reads, as is typical for each platform. All underlying sequence data have been deposited into GenBank (BioProject ID: PRJNA251548; Table S1).

We used a compendium of best practices to call sequence variants (Fig. S30). First, BWA-MEM version 0.7.12 (*66*) was used to map all sequences to the Mmul_10 reference, which also included the mitochondria genome (NC_005943.1) and had the pseudoautosomal region of chromosome Y masked. To identify reads potentially originating from a single fragment of DNA and mark them in the BAM files, we used Picard MarkDuplicates version 1.105. SNVs were then called using GATK version 4.1.2.0 and a VCF file was generated. The hard filters suggested by the developers of GATK were applied to the SNVs and all failing SNVs were removed. We then used GATK VariantAnnotator to annotate SNVs applying AlleleBalance. SNVs with an allelic balance for heterozygous calls (ABHet=ref/(ref+alt)) ABHet < 0.2 or ABHet > 0.8 were removed. Ensembl Variant Effect Predictor (VEP) release 98 (*67*) was used to annotate the SNVs in the VCF file based on combined Ensembl and RefSeq gene annotations. The average sequencing coverage used for SNV genotype calls in the high-coverage samples (n = 769) was 33.59X and average sequencing read coverage for moderate coverage samples (n = 84) was 7.99X.

PCA and admixture analyses. We performed PCA on all 853 genomes before and after linkage disequilibrium (LD) pruning of SNVs. Autosomal SNVs, excluding unplaced scaffolds, were filtered with Plink for missing call rates > 0.05 (--geno 0.05) and minor allele frequency < 0.1 (--maf 0.1) resulting in a dataset of 14,128,568 SNVs. Plink --pca was performed on this dataset to generate eigenvectors, and principal components 1-3 were plotted using the MATLAB scatter3 function. We applied ADMIXTURE (V1.23) to assess population structure in the macaque research colonies. For LD pruning, we removed related SNVs where r2 >= 0.2 using Plink (V1.90) reducing the number of segregating SNVs from 14,128,568 to 193,634. We performed cross-validation experiments using ADMIXTURE to find the best fitting model, calculated Fst among groups with VCFtools v. 0.1.17 and organized populations based on a hierarchical clustering of pairwise Fst. Because the number of samples sequenced varied by research center, we randomly downsampled 68 individuals from the four largest sample sets (California National Primate Research Center [NPRC], Oregon NPRC, Tulane NPRC, and Wisconsin NPRC) to assess the effect of sample size differences.

Genome diversity. We calculated diversity statistics for individual macaques and examined them in the context of the populations outlined in the 3D PCA plot of Fig. 6A-B. We calculated genome-wide heterozygosity as heterozygous SNVs divided by ungapped autosomal assembly length (Fig. S17). We also provide estimates of inbreeding coefficients (Fig. S18) and runs of homozygosity (ROH; Fig. S19) for Indian macaques across the defined research populations. We performed Welch two-sample t-tests comparing Cayo macaques to other Indian macaques (Table S27). Cayo samples show lower genetic diversity as measured by heterozygosity and higher levels of homozygosity as measured by the inbreeding coefficient and ROH. This is consistent with the longer period of time that the Cayo population has been reproductively isolated and suggests a genetic bottleneck has occurred.

Linkage disequilibrium (LD). To address the extent of LD decay in macaque populations, we performed a preliminary analysis. We selected 411 unrelated Indian macaques and assessed the pattern of LD. We applied PopLDdecay (*68*) to calculate the correlation coefficient (r2) among pairwise SNVs and used the mean-bin method to plot the distance of LD decay among adjacent SNVs. As a comparison, we repeated the same analysis using human population data (399 African samples and 1 European sample). LD was estimated for pairwise SNVs within a window of 200 kbp. Among the Indian macaques, we estimate an average LD (r2 = 0.038) between SNVs and find that at an r2 = 0.1 the average distance between two SNVs is ~4.2 kbp (Fig. S20). This is in contrast to the human dataset where we observe an average LD (r2 = 0.03) and where we estimate the average distance is nearly ~10.1 kbp between two SNVs (when r2 = 0.1). (Fig. S20). Thus, LD overall appears to be decaying faster in rhesus macaque and extending further compared to a human subsampled population of Africans.

Variant pathogenicity analysis. In order to assess the impact of variants in orthologous human genes, we projected the variants on the human genome (GRCh38). This was done by using the Picard LiftoverVcf tool with LIFTOVER_MIN_MATCH=0.95 and RECOVER_SWAPPED_REF_ALT=true, which means if the ALT allele equals the new REF allele then the REF and ALT alleles will be swapped, otherwise the SNV will be removed. Picard LiftoverVcf was performed reciprocally requiring that human sites must lift over back to original position on rhesus in order to be retained. A total of 67,917,330 SNVs (79.23% of all rhesus SNVs) reciprocally lifted over to human, including 9,876,720 SNVs where the REF and ALT were swapped in human relative to rhesus. For each gene, variants in the protein-coding regions of the gene (excluding UTRs and intronic regions) were extracted and classified as missense or likely gene-disruptive (LGD), adapting previous variant classification criteria (*69*). Missense variants alter the amino acid of the protein while LGD variants result in loss-of-function, and include stop-gain, start-lost, splice-donor or splice-acceptor variants.

We focused on an assessment of macaque genetic variants among 187 genes where rare or *de novo* deleterious variants have been implicated with human neurodevelopmental disorders (NDDs) (*32*), including autism (*33*). Variants were projected onto the human genome (GRCh38) using the Picard-based liftOver procedure outlined earlier and annotated using Ensembl VEP release 99 with the ‘Ensembl homo sapiens 99_GRCh38' database. SNVs were filtered to retain only those found in protein-coding regions of genes as annotated in gene models from NCBI RefSeq Genes (update 2019-12-06). This excludes intronic regions, UTRs, and noncoding genes such as lincRNA. For each variant, transcript choice and variant effects were prioritized based on: 1) recurring biotype, 2) CCDS status, 3) variant rank, and 4) canonical transcripts (default order is: https://uswest.ensembl.org/info/docs/tools/vep/script/vep_other.html#pick_options)

[./vep -i (*53*) -o (*56*) --cache --pick --pick_order biotype,ccds,rank,canonical --force_overwrite --vcf --fork 4]. These SNVs were then annotated for predicted pathogenicity using CADD (v1.4). Genes were binned by mutation tolerance measures (pLI, missenseZ, as per (*69*)) and normalized variant counts were estimated to reveal genes with high mutation intolerance in humans (pLI >= 0.9) harboring naturally occurring variants in macaques, some of which are predicted to be likely pathogenic in humans (CADD score >= 25).

Structural variation. We performed structural variant (SV) comparisons between the Indian and Chinese macaques (*20*). Briefly, we applied PBSV (*70*), Sniffles (*71*), and Smartie-SV (*5*) to map Chinese and Indian macaque SVs to the human genome (GRCh38).

Supplementary Text

Assembly gap closure. For final gap closure we applied "merauder", the gap-closing module of meraculous (*72*), to close a subset of sequence gaps using AG07107 WGS data. Following this step, we also applied a custom-based approach to close remaining gaps using AG07107 SMRT sequences. Specifically, these SMRT sequences were mapped against the HiRise assembly using minimap2 (*57*) and tagged all secondary alignments (option --secondary=yes). Next, we converted the alignment into its binary form (BAM file), sorted it by coordinates, and indexed it with SAMtools v1.7 (using htslib 1.8), (*66*). We filtered a subset of primary alignments with MAPQ > 20 to identify potentially closable gaps. We iterated over the identified gaps and selected a subset of alignments (minimum subset size = 2) that span the gap in flanking regions of 20 bp, both up and downstream. We extract the subreads of the alignments that mapped within these coordinates and we ran a sequence length agreement procedure for determining a consensus length of the subreads that will be used to close a specific gap since the length of these subreads may vary because of possible insertions, deletions, and soft-clippings generated during mapping. We aligned the resulting subset of reads with MUSCLE (*73*) and generated a consensus sequence. Finally, if the length of the consensus sequence (which includes the flanks) was smaller than twice the selected flank size (i.e., 40 bp), we considered the gap closed with no additional sequence and removed the gap. Otherwise, the "filled sequence" was used to update the genome.

Assembly breakage and correction. The Macaca mulatta AG07107 assembly was aligned against the human reference (GRChg38.12), Macaca fascicularis (macFas5), and Macaca mulatta (Mmul_8.0.1) references using nucmer (*46*), and then we broke the assembly into 1,000 bp nonoverlapping segments using BLAT. Possible breakpoints, where at least 50 kbp of sequence aligned to a chromosome other than the primary chromosome for the remainder of the scaffold, or where at least 50 kbp of sequence aligned to a discontinuous location (>100 kbp apart from the neighboring segment), were manually reviewed. Order and orientation were defined initially using the alignments to only the *Macaca* genomes. After creation of the chromosomal formatted files, chromosomal sequences were again aligned against GRChg38.12 and careful comparisons were made with the published mapping data (*74*) with any discrepancies subjected to manual review. Scaffolds with at least 100 kbp uniquely aligned were placed along the order/oriented chromosome. For those scaffolds with <100 kbp of sequence uniquely aligned, if at least 25 kbp uniquely aligned to a macaque or human chromosome and at least 80% of the unique placements were from the same macaque/human chromosome, the sequence was assigned to the appropriate chromosome. The remaining 95 Mbp were considered unplaced.

Strand-seq data analysis. High-quality Strand-seq single-cell libraries were obtained from a lymphoblast cell line derived from one macaque (*Macaca mulatta*, MMU1). The cells were maintained using standard culture conditions and 40 uM of BrdU was added to the media for 23 hours prior to sorting. Single cells were deposited into a 96-well plate using the BD FACSMelody cell sorter and Strand-seq library construction was pursued for single cells following the protocol described by Sanders et al. (*75*). Libraries were sequenced on a NextSeq500 (MID-mode, 75 bp paired-end protocol), demultiplexed, and data aligned to GRCh38/hg20 (BWA 0.7.15). Low-quality libraries, such as those with high background reads, were excluded from analysis, and 61 high-quality cells were obtained for inversion analysis.

ncRNA classification. Our reported increase in the number of annotated ncRNAs is the result of improvements in Mmul_10 genome representation, additional RNA-seq and Iso-Seq evidence, and upgrades to the NCBI ncRNA annotation pipeline. In our study, we utilize the extensive NCBI gene annotation reports for each macaque assembly to show the improvements to ncRNA annotation. Unfortunately, we can't distinguish which factor contributed most significantly to the improvements seen in the NCBI ncRNA annotation of Mmul_10. Nonetheless, we show these overall increases in the numbers of ncRNA for Mmul_10 versus Mmul_8.0.1 using the NCBI classifications in Table S8. In addition, we show better Mmul_10 ncRNA representation for the various ncRNA classifications, which suggests these are due to increased assembly quality since we use the alignments of the same human ncRNAs for both assemblies as part of the CAT pipeline (Table S9). Finally, we provide the sequence coordinates of each rhesus macaque ncRNA type by their location in Mmul_8.0.1 and Mmul_10 when aligned to human ncRNAs (Table S28).

Quality assessment of the Mmul_10 MHC and KIR regions. MHC genotyping assays demonstrate that the MHC class I region of the AG07107 fibroblast line is homozygous for the Mamu-A004 haplotype, which contains a pair of functional Mamu-A genes (*76*) (*77*) (*78*). The Mmul_10 assembly provides an exceptionally accurate representation of this Mamu-A region, which is the most common haplotype in both Indian- and Chinese-origin rhesus macaques. Mmul_10 sequences for the Mamu-A and Mamu-A3 genes are nucleotide identical to genomic sequences characterized by multiple independent methods for the *Mamu-A1*004:01:01* and *Mamu-A4*14:03:01:01* alleles, respectively (Table S2). This pair of Mamu-A genes are separated by 117.5 kbp on chromosome 4 and both are contained on a single rhesus macaque BAC (MMU063G23, Accession number AC148670.1). The Mmul_10 assembly is virtually identical over 189,655 bp with this BAC sequence despite the fact that it was derived from a different individual, differing only by four SNVs and two short insertions of simple sequence repeat units (*10*).

In contrast to the Mamu-A region, the Mamu-B region of the Mmul_10 assembly has been collapsed and includes gaps of 87 kbp and 46 kbp. AG07107 fibroblasts are heterozygous for the Mamu-B048 and Mamu-B055 haplotypes based on genotyping assays, each of which are expected to contain tandem arrays of 10 or more Mamu-B-like genes and pseudogenes (*76*) (*77*) (*78*). The Mmul_10 assembly contains six Mamu-B-like genes, including copies of *Mamu-B*041:01:01:01* and *Mamu-B*064:01:01:01* from the Mamu-B048 haplotype that are identical to cDNA and genomic sequences determined by independent methods (Table S2). Another pair of Mamu-B genes from the Mamu-B048 haplotype (*Mamu-B*054:nov01ps* and *Mamu-B*134:04:01:01*) are present in the Mmul_10 assembly but these sequences contain 1 or 10 SNVs, respectively compared to expected genomic sequences from unrelated rhesus macaques (Table S2). The remaining pair of Mamu-B genes in Mmul_10 are pseudogenes (*Mamu-B*061:nov01ps* and *Mamu-B11L:01ps*) that are associated with the alternate Mamu-B055 haplotype of AG07107. The bulk of the Mamu-B gene cluster that is expected for the Mamu-B055 haplotype of AG07107 cells resides on a 737,392 bp chromosome 4 unlocalized scaffold (NW_021160161) and a 172,297 bp unplaced genomic scaffold (NW_021162083). The NW_021160161 scaffold contains at least two functional genes (*Mamu-B*052:01:01:01* and *Mamu-B*058:02:01:01*) that are identical to expected genomic sequences (Table S2). In addition, this NW_021160161 scaffold contains at least seven more Mamu-B genes and pseudogenes with varying degrees of mismatches relative to expected sequences on both the B055 and B048 haplotypes. Likewise, the NW_021162083 unplaced scaffold includes *Mamu-B*055:01:01:01*, which is identical to the expected genomic sequence plus another Mamu-B pseudogene. This unplaced scaffold appears to belong at least in part within the 170 kbp assembly gap of the NW_021160161 scaffold. One additional 32,498 bp unplaced genomic scaffold (NW_021161166) contains the gene *Mamu-B*109:nov02* that also perfectly matches a genomic sequence expected for the Mamu-B048 haplotype. In total, nine of twenty MHC class I gene sequences in Mmul_10 appear to be identical to allelic variants defined by independent methods (Table S2). Sequence variants between the remaining class I genes/pseudogenes and expected sequences resulting from chimeric assemblies as well as large gaps and multiple unplaced scaffolds illustrate the difficult challenges that remain for assembly of complex segmentally duplicated genomic regions such as the MHC even with the long-read PacBio approach used for Mmul_10.

Review of the MHC class II region revealed that the majority of the classical loci in Mmul_10 also suffer from assembly artifacts that result in the collapsing of sequence reads from both alleles of these heterozygous loci into chimeric sequences for the primary chromosome 4 (NC_041757) assembly (Table S2). The Mamu-DRA locus appears to be an exception to this observation since the Mmul_10 coding sequence (CDS) is identical to a cDNA sequence for the *Mamu-DRA*01:04:01* allele. The assembly process was also nearly successful for the Mamu-DP region. In this case, the Mamu-DPA1 CDS and Mamu-DPB1 CDS for NC_041757 only differs by SNVs relative to the coding sequences of the *Mamu-DPA1*02:07:02* and *Mamu-DPB1*06:07* alleles that were predicted by our MiSeq genotyping results with AG07107 DNA. In addition, there is another 85,793 bp unlocalized chromosome 4 genomic scaffold (NW_021160159) that contains *Mamu-DPA1*02:04* and *Mamu-DPB1*08:01* alleles, which are identical to the expected coding sequences (EF204947, EF362434) on the second Mamu-DP haplotype in AG07107 fibroblasts (Table S2).

Observations regarding the quality of KIR genes on chromosome 19 in the Mmul_10 assembly are summarized in Table S3. The expected cluster with five KIR genes lies within the Leukocyte Receptor Complex from NC_041772: 54,503,383-54,606,068 on chromosome 19 (*21*). Unexpectedly, an additional cluster with KIR genes has been placed 3.6 Mbp away at extreme telomeric end of chromosome 19 in the Mmul_10 assembly (NC_041772: 58,197,630-58,312,841. This tandem array of eight KIR genes is immediately distal of a 100 bp assembly gap and lies in an inverted transcriptional orientation relative to the KIR cluster in the Leukocyte Receptor Complex (Table S3). Designating this 118 kbp telomeric KIR cluster as a chromosome 19 unlocalized genomic scaffold would have been more appropriate than the current artifactual fusion onto the end of chromosome 19.

Seven of thirteen annotated Mamu-KIR genes in Mmul_10 have coding sequences that are identical to previously described KIR transcripts of rhesus macaques (*79*). Although few genomic Mamu-KIR sequences are available for comparison, a fosmid sequence (KT332856) from an independent rhesus macaque is completely identical to the *KIR2DL4* gene of Mmul_10 (Table S3). Likewise, a BAC genomic sequence (BX842591.2) characterized by Sambrook and coworkers only differs from the Mmul_10 *KIR2DS4* gene by an SNV in a polyA tract (*21*). This same BAC also contains an allelic variant of the Mmul_10 *KIR3DL2* gene that is 99.73% identical over 13,300 bp. Five of the six remaining Mmul_10 KIR genes differ by only one to ten SNVs over their coding regions compared to previously described Mamu-KIR alleles (Table S3). These may reflect chimeric assemblies that have collapsed closely related allelic variants or they could be novel allelic variants in AG07107 cells. Resolution of these possibilities will require targeted analyses of AG07107 cells such as those described recently by Bruijnsteijn and coworkers (*79*).

Segmental duplication analyses. WGAC detected a total of 111.56 Mbp of assembled SDs in 7,626 nonredundant loci (>1 kbp and >90% sequence identity). There are in total 44,410 pairwise alignments between duplications, but only 9,476 pairs with both loci assigned to chromosomes (Fig. S10). Of the 9,476 SDs assigned to chromosomes, 6,372 were interchromosomal, and an additional 755 were intrachromosomal but separated by at least 1 Mbp. The 111 Mbp of assembled SDs is a >3-fold improvement compared to WGAC analysis of the previous Sanger-based *Macaca mulatta* assembly which identified just 32 Mbp of assembled SDs (*2*).

Sequence orientation. The SDA method, which relies on the alignment of raw sequencing reads to the assembly to find regions of increased alignment depth, detected 276 regions of collapsed assembly, for a total of 9.1 Mbp (N50 = 36.2 kbp). Of those 9.1 Mbp, 5.5 Mbp (60%) are assignable to assembled chromosomes, with the remainder on unplaced or unlocalized contigs (Table 1). Compared to the WGAC-identified assembled SDs, only 5.8 Mbp overlap, corresponding to collapsed assemblies that do not fully represent each locus of the duplications. SDA recovers an additional 19.1 Mbp of resolved sequence (N50 = 37.4 kbp) from the 9.1 Mbp of collapsed assembly. For example, SDA resolved two collapses within the MHC class I region, which is expanded in OWMs. Application of SDA to the two collapses (24.3 and 84.7 kbp) produced six resolved contigs of size 34.5-107.2 kbp. Each SDA contig aligned with higher percent sequence identity (0.2%-1.2% absolute improvement) to a BAC tiling path, considered the standard for the rhesus macaque MHC region, over the locus from another Indian origin *Macaca mulatta* individual (*10*). The *ZNF669* gene family is expanded to approximately 50 copies in *Macaca mulatta*, as measured by qPCR and NanoString (*80*) and SDA identifies nine assembly collapses (20-92 kbp) corresponding to representations of *ZNF669* genes in the Mmul_10 assembly. SDA assembles 53 contigs from these collapses, producing an additional 1.9 Mbp of assembled sequence (N50 = 36.5 kbp). Four of these contigs are better representations of transcribed members of the *ZNF669* family, as demonstrated by improved alignment of full-length cDNA sequences compared to the original Mmul_10 assembly (Fig. 3).

FISH validation. In Fig. S2, we describe reiterative BAC clone mapping experiments to validate inversion locations (Table S29). Our FISH results confirm an increase in intrachromosomal duplications and specifically pericentromeric mapping. At this time, we cannot exclude a possible sampling bias in the clone region selection that increased the call for pericentromeric regions.

Repetitive element analyses. The repeat content of the rhesus macaque assemblies Mmul_8.0.1 and Mmul_10 were analyzed using a local installation of RepeatMasker with the most recent Dfam3 library. The raw output was parsed using Excel into broad categories (DNA transposons, LTR transposons, non-LTR transposons) and more specific categories (e.g., LINE/L1, LINE/L2, etc.; Database S1, worksheet "Repeat content").

Lineage-specific *Alu* elements. As another measure of sequence assembly quality, we computationally compared the rhesus macaque assemblies Mmul_8.0.1 and Mmul_10 for *Alu* element content. The assemblies had comparable numbers of total *Alu* elements (1,312,984 vs. 1,248,216) in Mmul_8.0.1 and Mmul_10 as well as non-truncated, full-length, *Alu* family members (820,192 vs. 818,508) (Database S1, worksheet "Lineage-specific *Alu* counts"). The slightly lower numbers of total *Alu* elements in addition to full-length *Alu* repeats in the Mmul_10 assembly is presumably a direct result of a higher quality assembly with considerably less elements ending up at the end of contigs and effectively being counted multiple times.

We also performed a sequential comparison of the numbers of lineage-specific *Alu* repeats within each assembly by stepwise comparison to previously sequenced primate genomes, first by comparison to human (Homo sapiens; hg38) followed by the olive baboon (Papio anubis; Panu3.0) (Database S1, worksheet "lineage-specific *Alu* counts"). These results were used for the COSEG (https://github.com/rmhubley/coseg/blob/master/README.md) analyses of lineage-specific *Alu* subfamilies. Next, the stepwise comparison for lineage specificity included all OWM genomes currently available from NCBI. Using this approach, once again the Mmul_8.0.1 assembly tended to have slightly higher counts of lineage-specific *Alu* repeats when compared to the new Mmul_10 assembly consistent with the total counts of *Alu* repeats within each assembly (Database S1, worksheet "lineage-specific *Alu* counts").

***Alu* element subfamily analysis.** In order to determine the mode and tempo of *Alu* element expansion within the macaque lineage, we performed a COSEG analysis of the lineage-specific *Alu* repeats identified in each assembly. This type of approach is a measure of the numbers of retrotransposition competent *Alu* elements as well as their duplication efficiencies. The overall numbers of lineage-specific *Alu* elements were comparable with a slight decrease in the Mmul_10 assembly as compared to Mmul_8.0.1, and the number of propagating *Alu* subfamilies was also similar between Mmul_10 (105 subfamilies) and Mmul_8.0.1 (110 subfamilies). A network analysis of all subfamilies of *Alu* elements identified by COSEG was created by uploading the source and target subfamily information into Gephi (v0.9.1) (Table S24). A GEPHI image of the subfamilies in the Mmul_8.0.1 assembly is shown in Fig. S12A-B. There are two notable differences between the *Alu* networks of the two assemblies. The first is the separation of the blue nodes in Mmul_8.0.1. that show a mixture of AluYRd and AluYm1, while in the Mmul_10 assembly there is a clear separation of the AluYRd and AluYm1 derived subfamilies into separate bursts. The second is the difference in connectivity of the two networks. In the Mmul_8.0.1 network AluYRc, AluYm, AluYk and AluYf nodes (green) are connected to the AluMacYa3 nodes (purple), while in the Mmul_10 network there is a connection of the AluYRc, AluYm, AluYk and AluYf nodes to the AluYRd nodes (light blue burst centered around subfamily1). The increase in the number of different *Alu* subfamilies propagating in the newer assembly is another indicator that the sequence, assembly, and repetitive element libraries are all higher quality than Mmul_8.0.1 making the identification of subfamilies considerably less ambiguous and the resultant network better consolidated. The ancestral subfamily root for both sets of lineage-specific *Alu* subfamilies was determined using an in-house RepeatMasker library (Database Repeat Sequences). These results are available in Database S1, worksheet "COSEG Alu subfamilies RM". The *Alu* subfamily consensus sequences identified in this study (110 from Mmul_10 and 105 from Mmul_8.0.1) are available in FASTA format (Database Repeat Sequences).

Full-length LINE1 elements. The final comparison that we made was to determine the overall number of full-length L1 elements contained within each genome assembly. The number of full-length L1 elements provides an upper boundary on the number of germline L1 elements that are potentially retrotransposition competent. Because of their overall size, it is also a good measure of the sequence assembly quality. The number of full-length (≥6000 bp) L1 elements in the Mmul_10 assembly (6,892 L1 insertions) is nearly twice that of the Mmul_8.0.1 assembly (4,380 L1 insertions) (Table S16). Because they are 6 kbp in length, these elements often end up at the end of contigs in short sequence read assemblies and are inadvertently over counted in total and under counted as full length as a result. Such a large difference in full-length L1 element count is a good measure of the higher quality of the new Mmul_10 assembly. This number is also consistent with other previously sequenced primate genomes. Complete sequences for the full-length L1 sequences identified from both assemblies in this study are available in FASTA format (Database Repeat Sequences).

LINE1 element subfamily analysis. In order to determine the mode and tempo of L1 element expansion within the macaque lineage, we performed a COSEG analysis of the full-length L1 repeats identified in each assembly. This type of approach is a measure of the numbers of retrotransposition competent L1 elements as well as their duplication efficiencies. A network analysis of all subfamilies of L1 elements identified by COSEG was created by uploading the source and target subfamily information into GEPHI (v0.9.1) (*62*). A GEPHI image of the L1 subfamilies in the Mmul_8.0.1 and Mmul_10 assemblies appears in Fig. 4A-B. The branching and clustering patterns (as seen in Fig. 4A-B via color-coding) are similar in both assemblies. However, there is a change in distribution of the number of L1 elements in each the colored clusters, with an increase in the amount of younger L1 subfamilies in the Mmul_10 to the Mmul_8.0.1 assembly. The ancestral root for both sets of full-length L1 subfamilies was determined using an in-house RepeatMasker library. These results are available in Database S1, worksheet "COSEG LINE1 subfamilies RM". The L1 Alu subfamily consensus sequences identified in this study (58 from Mmul_10 and 61 from Mmul_8.0.1) are available in FASTA format (Database Repeat Sequences).

Assembly liftOver differences. To determine the differences between the two Indian rhesus assemblies, three categories of repetitive elements (Alu, L1 and LTR/ERV) were analyzed via liftOver. For L1 elements, 4,959 of 6,892 full-length L1 elements were successfully lifted from Mmul_10 to Mmul_8.0.1, while 1,933 failed liftOver. Of the failed L1 insertions, 97 of these were deleted or entirely absent from the Mmul_8.0.1 assembly, while 35 were partially deleted and 1,801 were split in the Mmul_8.0.1 genome. Of the 97 L1 elements absent in the Mmul_8.0.1 assembly, 65% were found on unassembled chromosomes, while 25% were found on chromosomes 3, 4, 14 and 19. The split insertions explain the lower number of full-length elements found in the Mmul_8.0.1 assembly compared to the Mmul_10 assembly (6,892 vs. 4,380). These results are available in Database S1, worksheet "L1 10 to 8 liftOver failed".

Although the majority of the L1 insertions were successfully lifted, there were differences between the location in the Mmul_8.0.1 assembly and Mmul_10 assembly. There were 4,959 full-length L1 elements with a successful liftOver between Mmul_8.0.1 and Mmul_10. Of these, 363 (7.3%) were found on a different chromosome in Mmul_10 compared to their location in the Mmul_8.0.1 assembly (Fig. 4C). Not surprisingly, many that were on unplaced chromosomes in Mmul_8.0.1 were now placed on chromosomes in

Mmul_10 (n = 92), as previously observed for Alu and LTR elements (Fig. 4D; Figs. S12 and S13). The X/Y chromosomes (N = 40/42) were the next largest source of differences, followed by chr11 with 36 L1 elements located on different chromosomes in Mmul_10.

For the *Alu* elements, 761,536 of the 818,508 full-length *Alu* insertions successfully lifted from the Mmul_10 to the Mmul_8.0.1 assembly; 20,970 of the 56,972 that failed were completely deleted from the Mmul_8.0.1 assembly compared to Mmul_10. These deleted *Alu* insertions were relatively evenly distributed among all of the placed chromosomes, with chromosomes 19 and 18 as notable exceptions (10% and 7%, respectively). These results are available in Database S1, worksheet "Alu 10 to 8 liftover failed".

Although the majority of the *Alu* insertions were successfully lifted, there were differences between the location in the Mmul_8.0.1 and Mmul_10 assemblies. The number of full-length *Alu* elements with successful liftOver between Mmul_8.0.1 and Mmul_10 was 761,536. While the chromosomal locations between assemblies were largely congruent, there were some discrepancies. The greatest differences were *Alu* elements on unplaced chromosomes in Mmul_8.0.1 (n = 8,291) that are now placed on chromosomes in Mmul_10, distributed throughout the genome (Fig. S12D). The next largest difference was chr2 with 1,591 *Alu* elements located on different chromosomes in Mmul_10 (Fig. S12C). This is due to most of them (n = 1,299) now being located on chr12 in the Mmul_10 assembly.

All elements from the RepeatMasker output that were identified as LTR/ERV were subject to the liftOver analysis. We found 706,177 of the total 732,024 sequences successfully lifted over from the Mmul_10 to the Mmul_8.0.1 assembly. A total of 17,353 LTR/ERV insertions present in the Mmul_10 assembly were entirely absent from Mmul_8.0.1. Because all LTR/ERV insertions were considered, the deleted insertions that failed to liftOver were further filtered by length. The majority of the deleted liftOver insertions were less than 1,000 bp and/or found on unplaced scaffolds. Only 10 LTR elements were greater than 7,000 bp, with only three of these on identified chromosomes. These results are available in Database S1, worksheet "LTR 10 to 8 liftover failed".

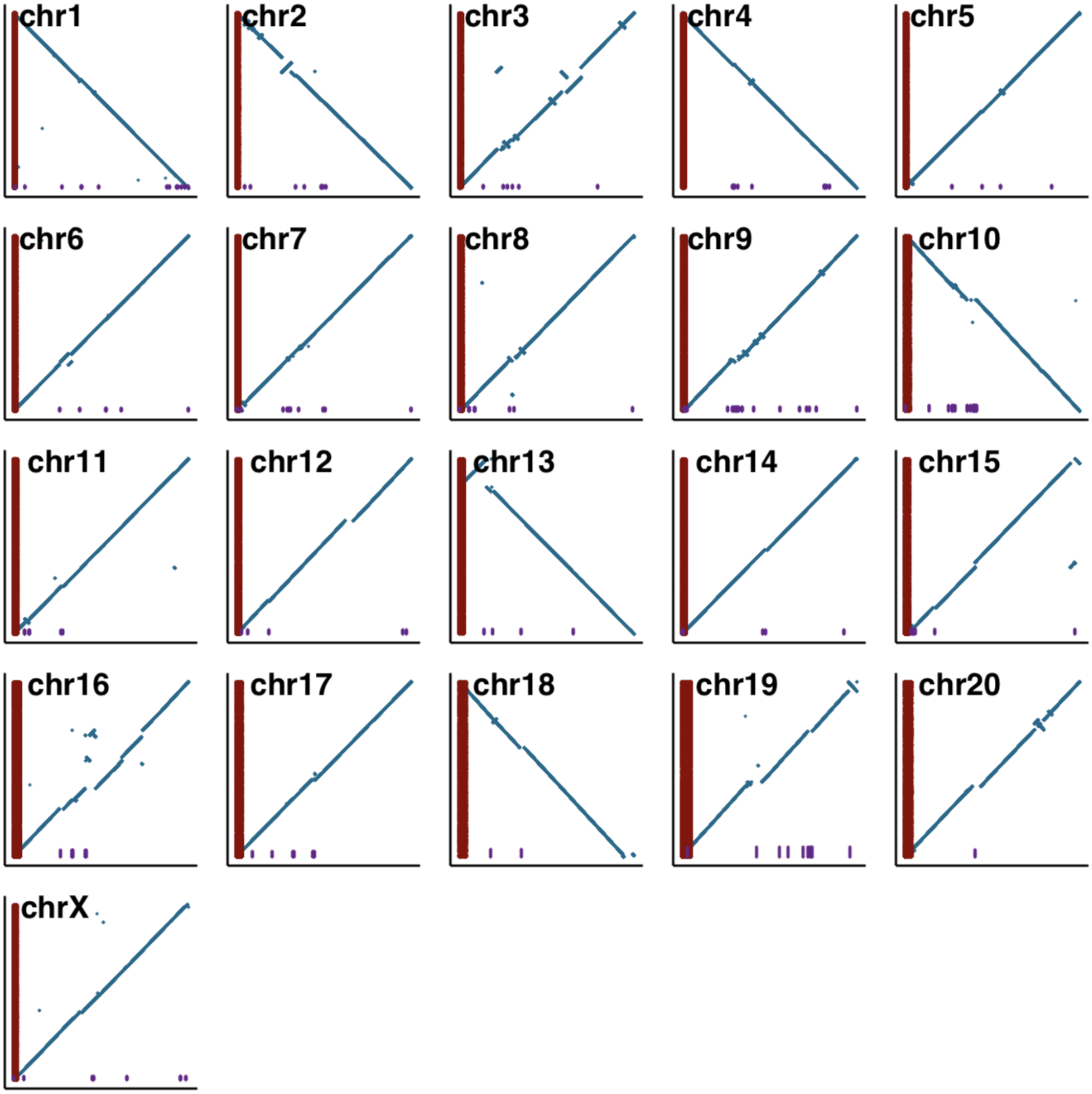
Although the majority of the LTR insertions were successfully lifted, there were differences between the location in the Mmul_8.0.1 assembly and Mmul_10 assembly. The number of LTR elements of all sizes with successful liftOver between Mmul_8.0.1 and Mmul_10 was 706,177. Of these, 11,627 (1.65%) were found on a different chromosome in Mmul_10 compared to their location in the Mmul_8.0.1 assembly (Fig. S13A). Not surprisingly, many that were on unplaced chromosomes in Mmul_8.0.1 were now placed on chromosomes in Mmul_10 (n = 7,430) (Fig. S13B). The next largest source of differences was chr2 with 1,647 LTR elements located on different chromosomes in Mmul_10 (Fig. 13A). This is due to most of them (n = 1,476) now being located on chr12 in Mmul_10 (Fig. S13C). This is the single largest 1:1 chromosome shift and was also observed with the *Alu* elements lifted. The next largest group is from chr4 in Mmul_8.0.1 to chr7 in Mmul_10 (n = 142). This number was n = 68 for *Alu* and not as pronounced.

Only 200 of these LTR elements are >7 kbp long based on Mmul_10 coordinates. The Mmul_8.0.1 chromosome was the same as Mmul_10 in 134 cases (only 67%, compared to over 98% when LTR length is not considered). The unplaced chromosome in Mmul_8.0.1 was the source of only five differences, whereas Mmul_8.0.1 chromosome 5 had n = 17, chromosome 12 had n = 12, and Y chromosome had n = 18 LTR elements >7 kbp placed to different chromosomes in Mmul_10. Many of the 200 over 7 kbp are MacERV-derived elements.

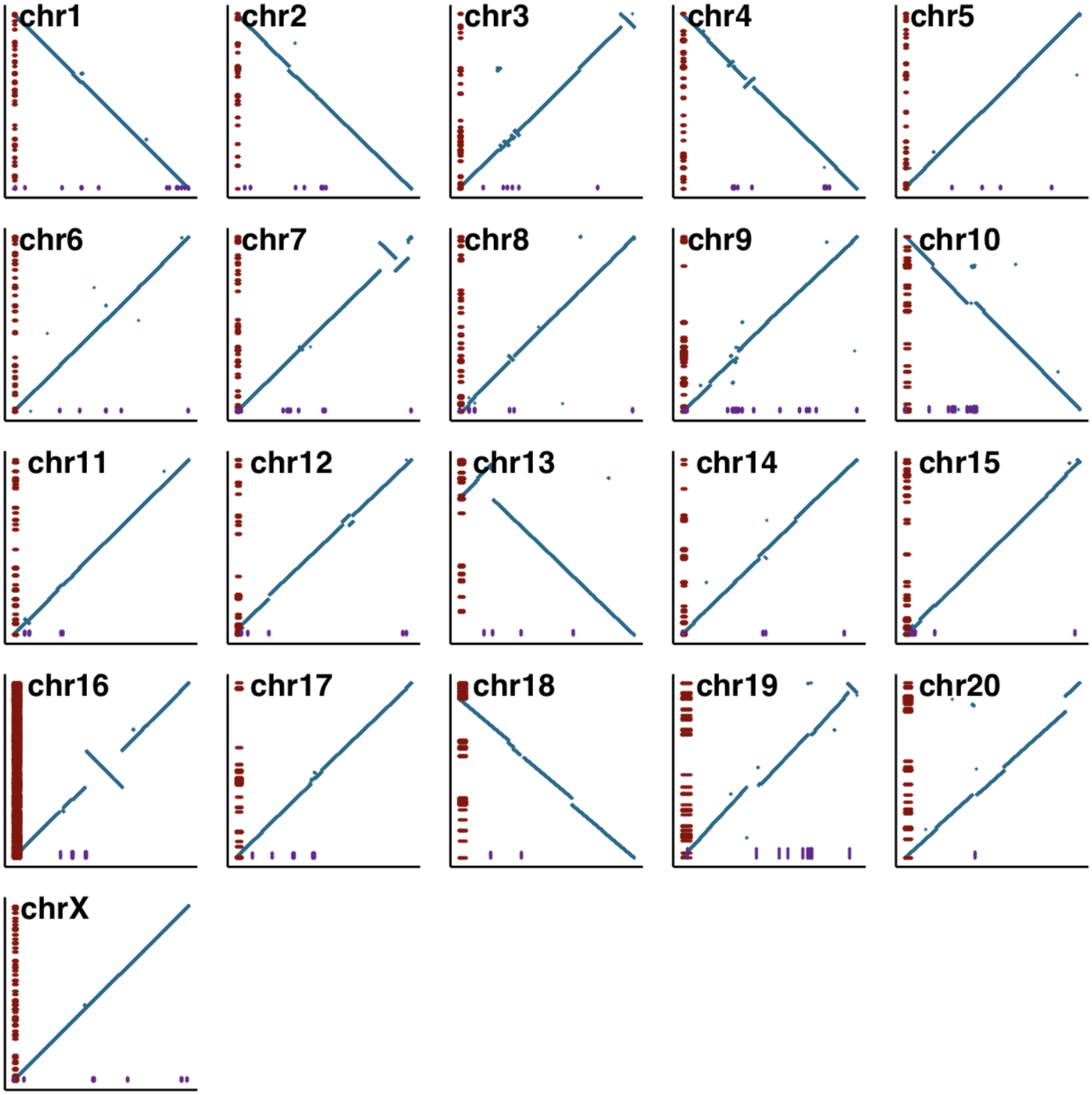
The “split” liftOver failure category for all three repetitive groups analyzed here indicates that Mmul_10 is a less fragmented assembly with the potential to obtain more full-length repetitive elements. In addition, this new assembly allows for the study of previously unobtainable transposable element insertion copies.

Full-length L1RS reveals evidence of an evolutionary arms race. Previous analyses of the rhesus macaque genome have identified rhesus-specific L1 retroelements derived from the primate-specific L1PA family, which are the active LINE1 subfamilies in primates. Classification of these elements is made using sequence similarity of the 3' ends. However, the 5' UTRs of these selfish elements are targeted by host factors that repress their transcription, most notably the KZNF proteins, which have greatly expanded across the primate lineage. By utilizing the UCSC Repeat Browser and mapping full-length rhesus-specific elements to the human L1PA5 consensus sequence (L1PA5 was active at the time of the human–rhesus divergence), we were able to identify rhesus-specific deletion patterns that accrue and persist in these L1 elements. Using these deletion patterns, we propose an order of evolution for the families identified by their 3' sequence. Our model suggests that after the human–rhesus divergence, at least three different regions of the L1RS 5' UTR experienced deletions, possibly to evade the binding of rhesus-specific KZNFs. Analysis of L1RS elements in the genomes of other OWMs supports the model that these sites experience adaptive selection, as all species display coverage drops in younger elements, although the size of the deletion varies suggesting that evasion events occurred independently along the phylogenetic tree. Importantly, two of these three sites overlap deletions also observed in active human-specific L1s, suggesting that KZNF being escaped is shared amongst humans and rhesus macaque. The third site is not deleted in human elements, suggesting that this event was specific to OWMs. In addition, the 3' end of L1RS also experiences continued variation in L1RS elements at a site proximal to an established binding site of the repressive factor BCOR in human cells.

A


chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chr20
chrX

B


chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chr20
chrX

C

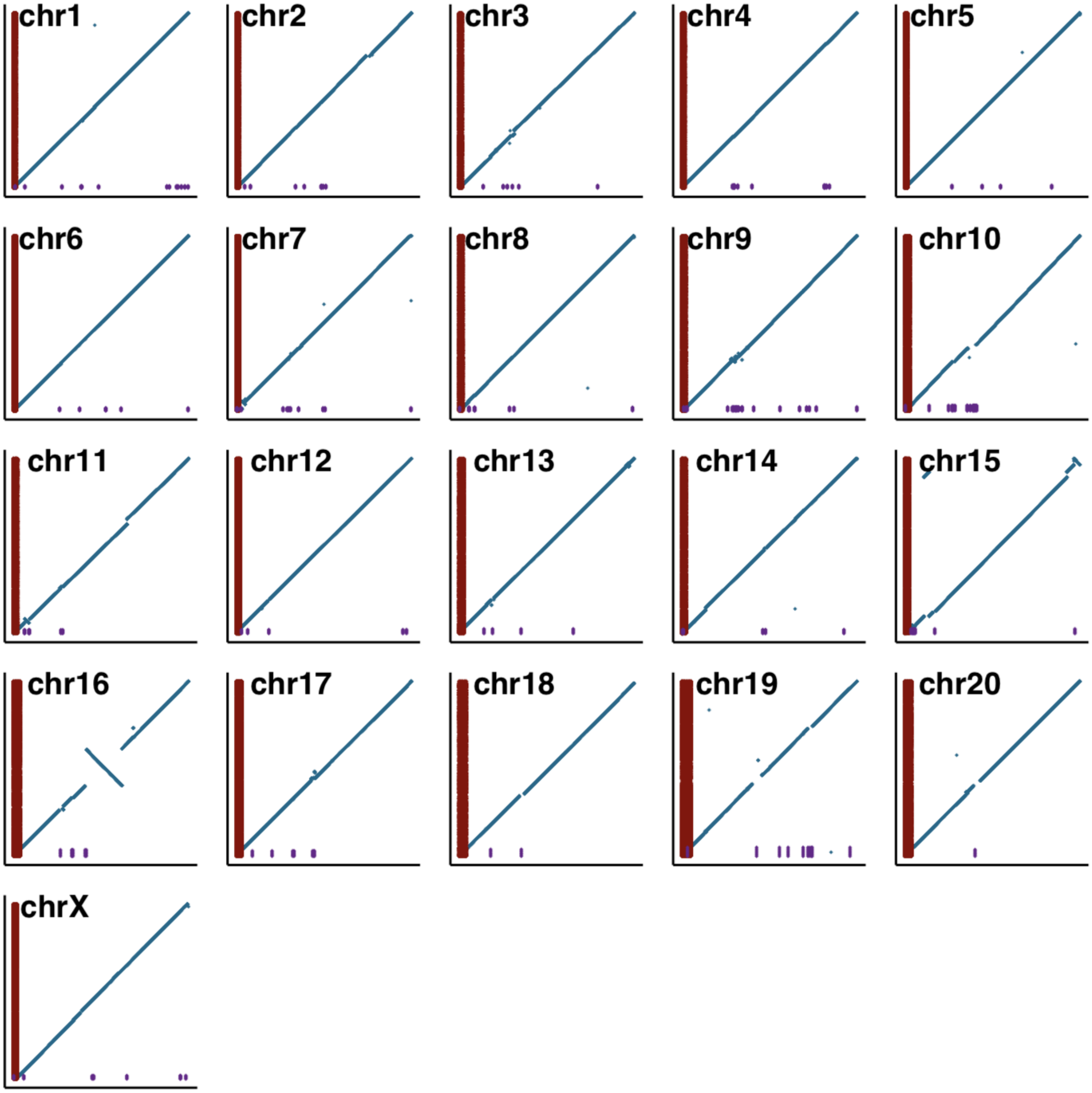


Fig. S1. Synteny comparison of Mmul_10 to the assembled chromosomes of other rhesus macaque assemblies for contig gaps. Alignments (in blue) of (A) Mmul_10 (x-axis) against Mmul_8.0.1 (y-axis), (B) Mmul_10 (x-axis) against rheMacS_1.0 (y-axis), and (C) Mmul_10 (x-axis) against macFas_5.0 (y-axis) assemblies. Contig gaps in each macaque assembly comparison are shown in red and for Mmul_10 are shown in purple.

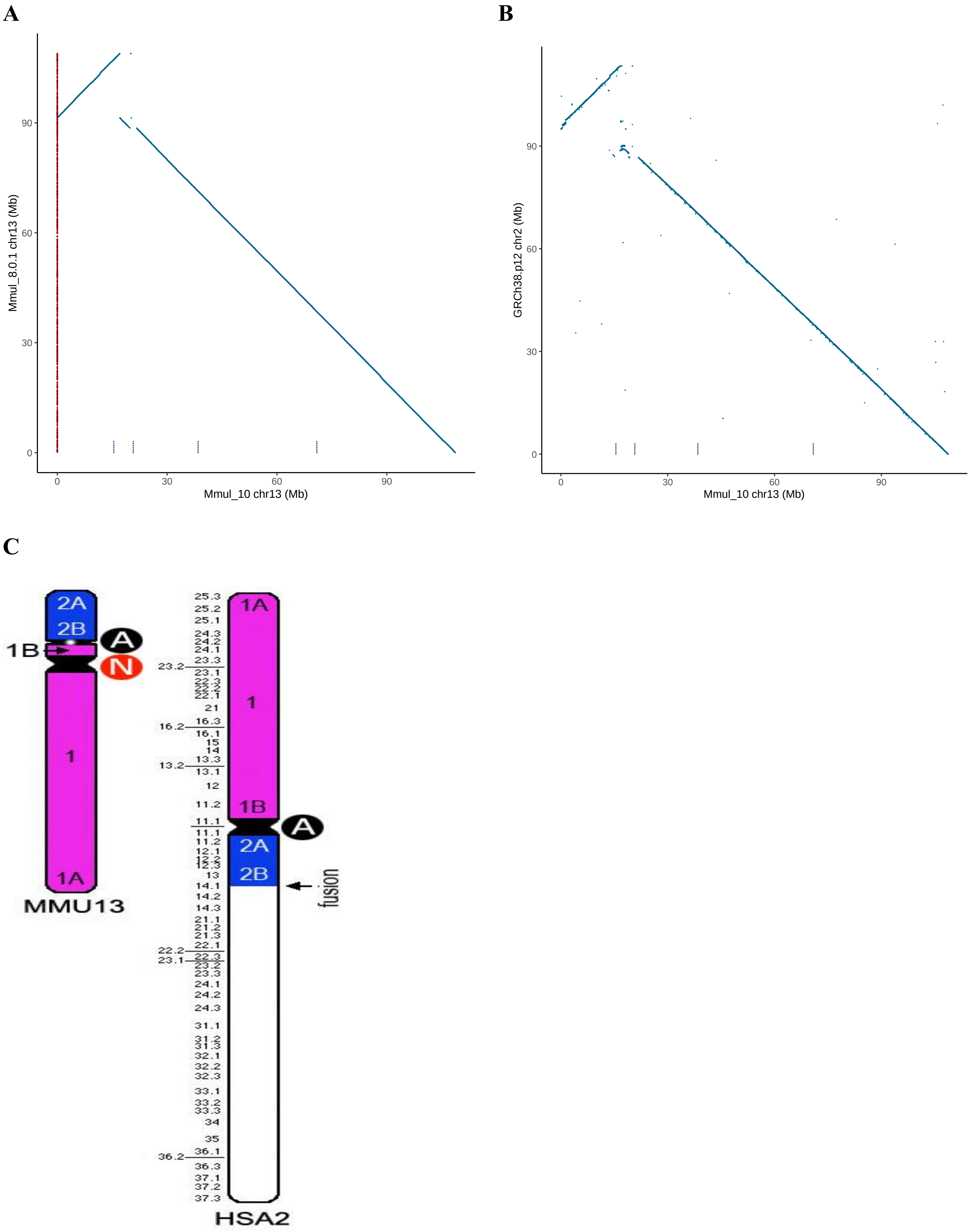

A
B
C
Mmul_8.0.1 chr13 (Mb)
Mmul_10 chr13 (Mb)
GRCh38.p12 chr2 (Mb)
0
30
60
90
2A
2B
1B
1
1A
A
N
MMU13
HSA2
fusion
25.3
25.2
25.1
24.3
24.2
24.1
23.3
23.2
23.1
22.3
22.2
22.1
21
16.3
16.2
16.1
15
14
13.3
13.2
13.1
12
11.2
11.1
11.1
11.2
12.1
12.2
12.3
13
14.1
14.2
14.3
21.1
21.2
21.3
22.1
22.2
22.3
23.1
23.2
23.3
24.1
24.2
24.3
31.1
31.2
31.3
32.1
32.2
32.3
33.1
33.2
33.3
34
35
36.1
36.2
36.3
37.1
37.2
37.3

Fig. S2. Comparative analysis of a rhesus macaque chromosome 13 inversion. (A) Mmul_10 aligned against Mmul_8.0.1 reveals a possible inversion contained within a single scaffold in Mmul_10; contig gaps in Mmul_8.0.1 are shown as red dots along the y-axis and contig gaps in Mmul_10 are shown as purple lines along the x-axis. Each dot in the alignment plot represents 1 kbp of uniquely aligned sequence. (B) The alignment of Mmul_10 to GRChg38.12 also reveals a potential inversion. (C) Inversion revealed by comparative mapping data obtained by reiterative FISH experiments using specific human BAC clones (Table S29). In alphabetical order, (A-B) represent the syntenic orientation of the homologous blocks between human and macaque. Ancestral centromere A and evolutionary new centromere (N) are documented (http://www.biologia.uniba.it/macaque/). The macFas_5.0 alignment (Fig. S1C) reveals the same order between Mmul_10 and macFas_5.0 chromosome 13 showing the existence of the inversion also in macFas_5.0 with respect to the human genome.

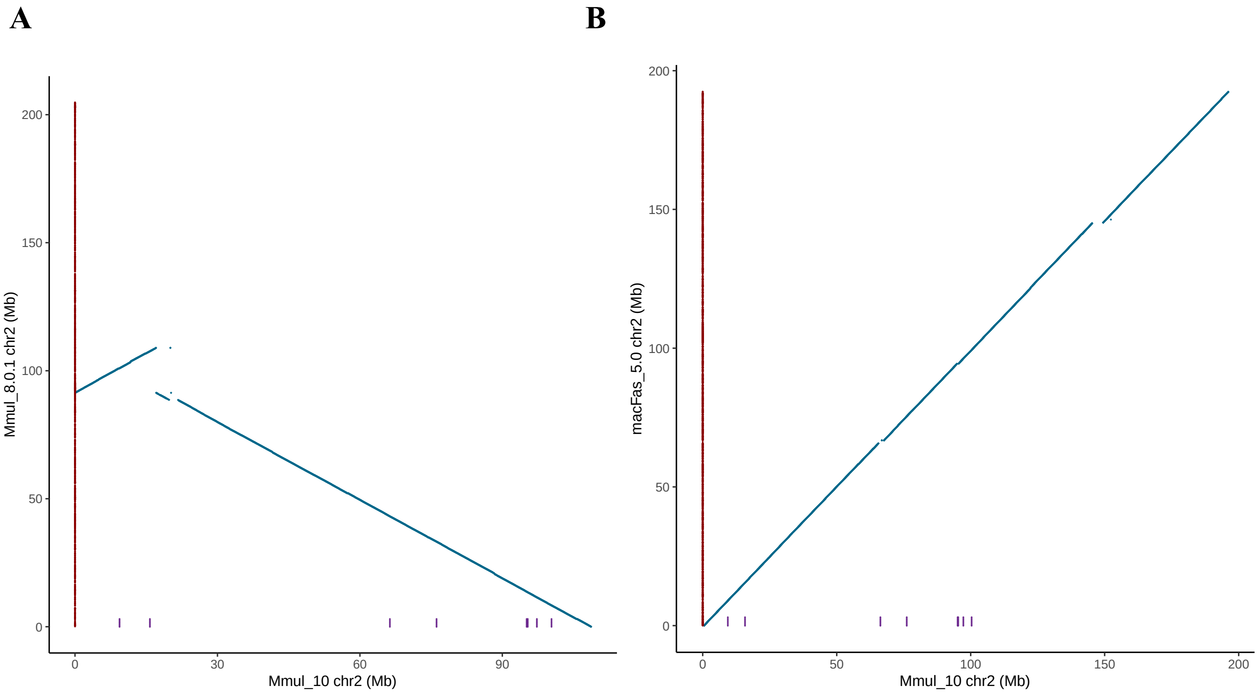


Fig. S3. Chromosome 2 inversions in Mmul_8.0.1 with respect to Mmul_10. (A) Alignment of Mmul_8.0.1 to Mmul_10 showing the inversion. Spanning BACs as well as Strand-seq data confirm the Mmul_10 assembly sequence order and (B) alignment with the macFas_5.0 assembly confirms the Mmul_10 assembly order.

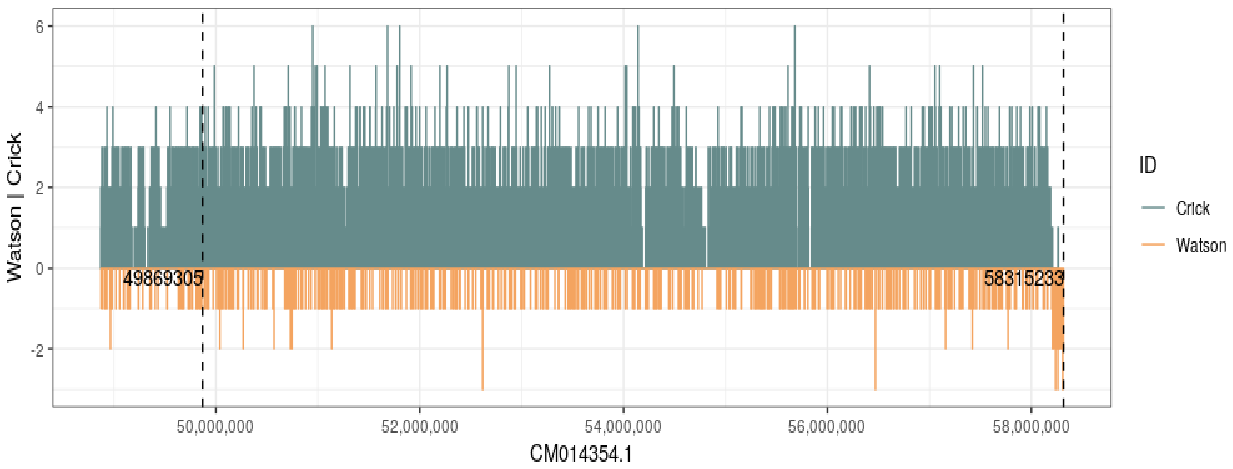


Fig. S4. Strand-seq coverage over KIR region on chromosome 19 (CM014354.1). Along the x-axis we plot cumulative coverage across all Strand-seq libraries (n = 60). Reads mapped to the Crick (plus strand) and Watson (minus strand) direction of the reference genome are shown in teal and orange, respectively. Vertical dashed lines highlight our region of interest (in the middle).

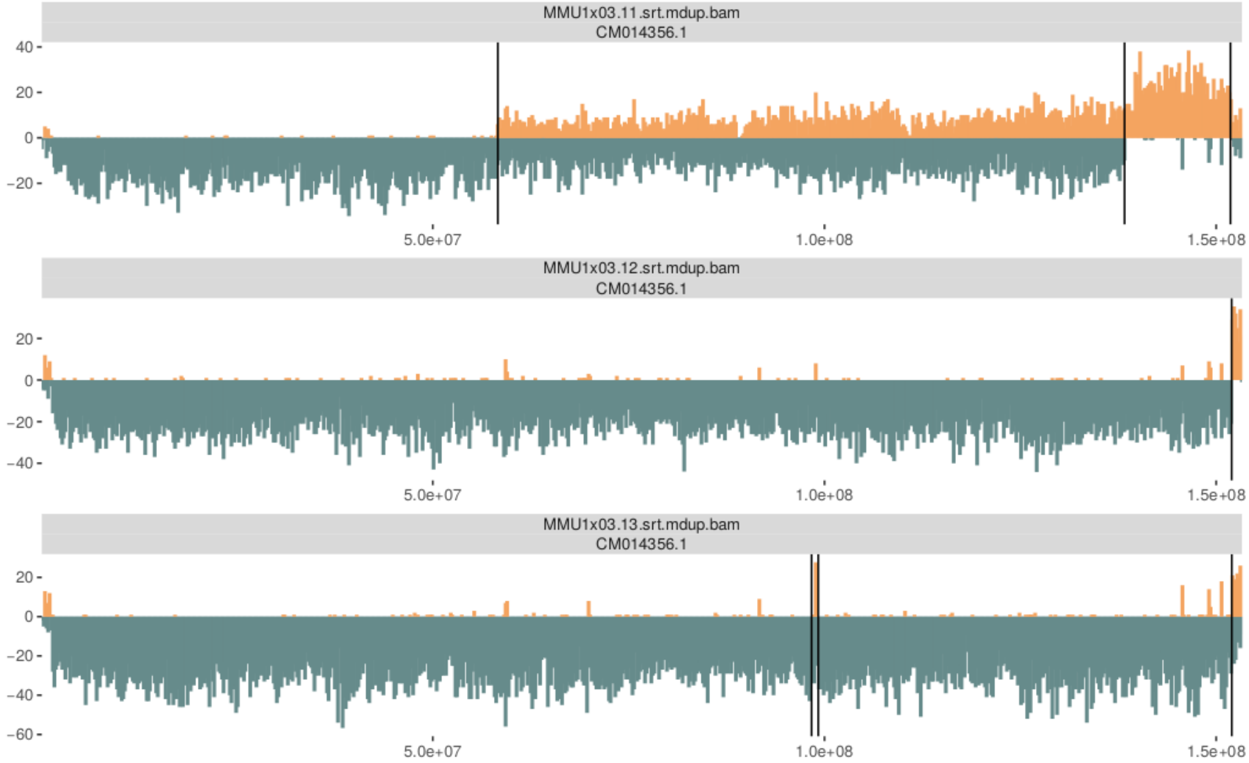


Fig. S5. Strand-seq coverage over KIR region on chromosome 19 (CM014356.1). In this figure three different single cells are shown along with their distribution of Crick (plus - teal) and Watson (minus - orange) reads along the chromosomal scaffold CM014356.1. Each vertical bar represents a number of Watson or Crick reads in a defined genomic bin of size 200 kbp. Black vertical lines denote changes in strand directionality along chromosomal scaffold CM014356.1. At the end of chromosomal scaffold CM014356.1 there is a recurrent change in read directionality. We mark this change as chimerism, as it is unlikely to see such change in directionality over the same genomic region in multiple single cells.

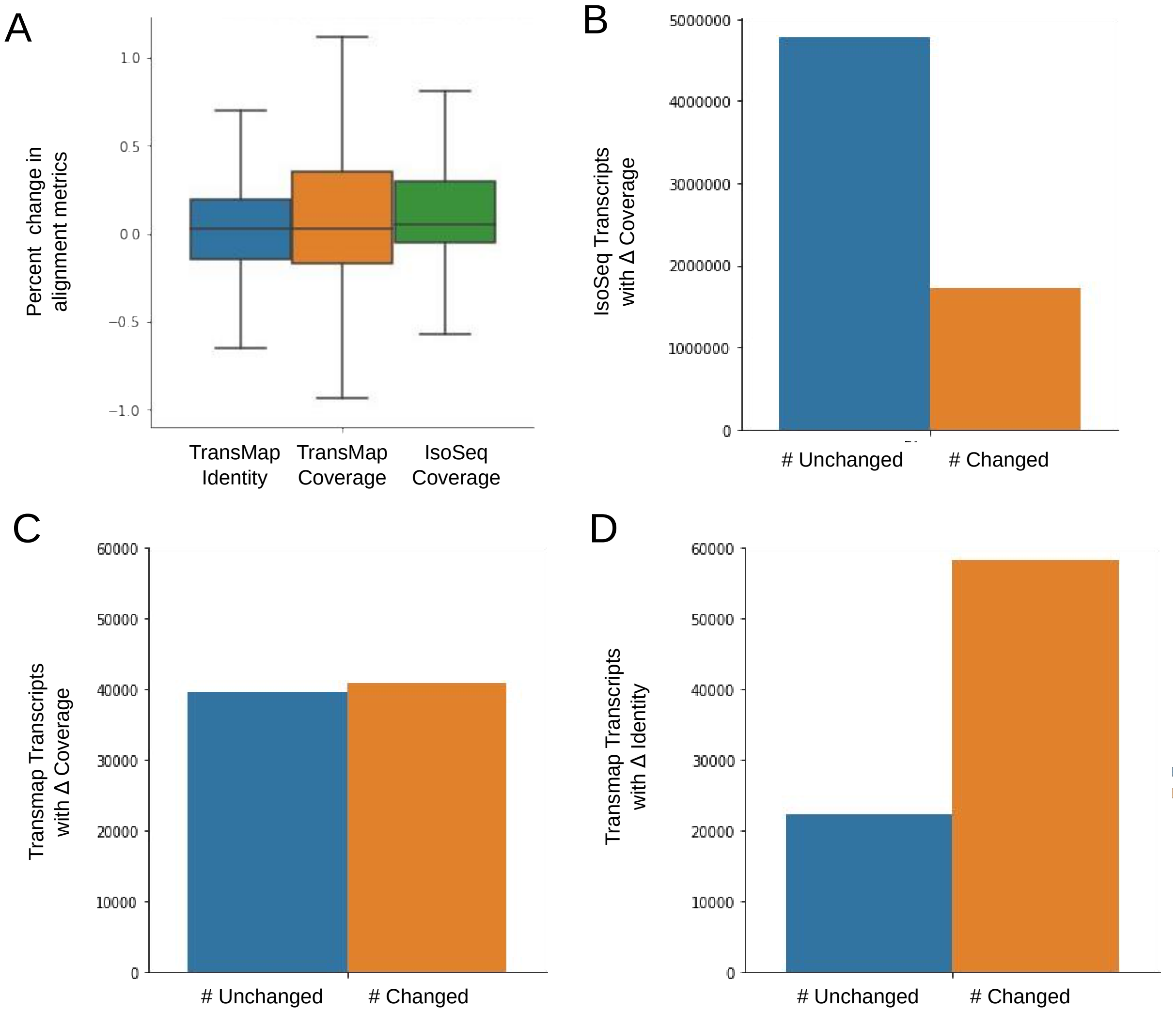


Fig. S6. TransMap and Iso-Seq mappability in Mmul_10 compared to Mmul_8.0.1. Comparative Annotation Toolkit (CAT) was used to project transcripts from GRCh38 to Mmul_10 and Mmul_8.0.1. Alignment coverage and identity were compared for orthologous transcripts found in each assembly pair. Additionally, Iso-Seq transcripts were mapped to both Mmul_10 and Mmul_8.0.1. (A) The box plots show the percentage change in identity and coverage of the TransMap alignments (left, middle), and coverage of Iso-Seq alignments (right) between Mmul_10 and Mmul_8.0.1. Transcripts with unchanged metrics were omitted from the plot. (B) Number of Iso-Seq transcripts that had changes in coverage between the assemblies. (C,D) Number of TransMap transcripts that had changes in coverage and identity between Mmul_10 and Mmul_8.0.1.

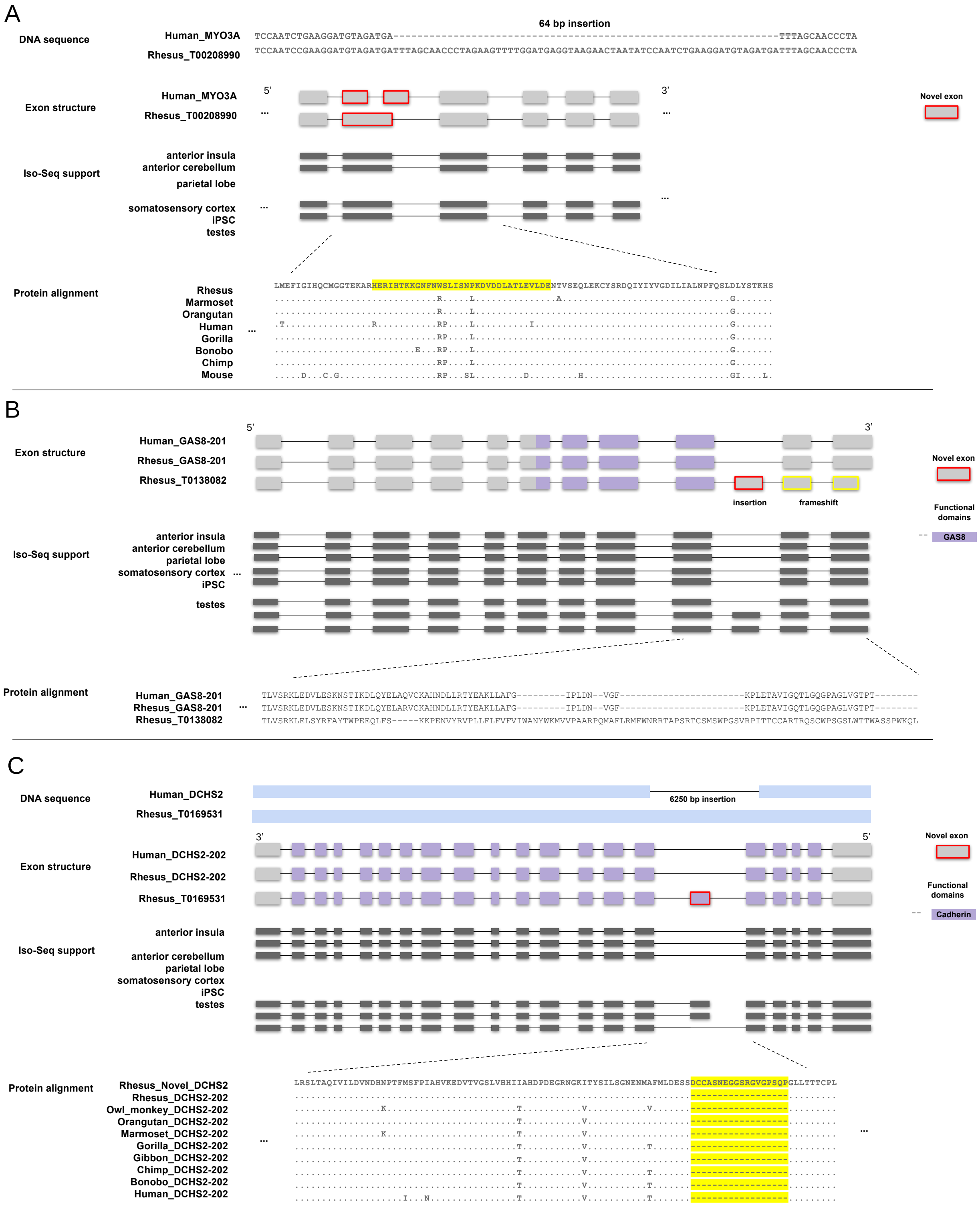


Fig. S7. Novel exons in rhesus macaque. (A) A rhesus-specific insertion of 64 bp in the macaque genome leads to a slightly different exon structure in *MYO3A*. The new isoform is supported by Iso-Seq in various tissue types. Despite the different exon boundaries, the final protein sequence is relatively unchanged, except a couple amino acid substitutions (affected portion of the protein highlighted in yellow). (B) In one isoform of *GAS8*, a

novel exon causes a predicted frameshift in downstream exons, resulting in an early stop codon. This isoform only has evidence of expression in Iso-Seq from testes tissue, where it is alternatively spliced. The alignment of predicted protein sequences demonstrates the frameshift. (C) A rhesus-specific 6,250 bp insertion introduces a novel exon to one of the *DCHS2* isoforms. The new isoform is supported by Iso-Seq in testes tissue, where it is alternatively spliced. However, the Iso-Seq transcripts do not support the exact isoforms predicted in the CAT annotation; rather than an exon skipping event, the novel exon appears to be an alternative starting exon of the gene. A protein alignment with many other primates is shown.

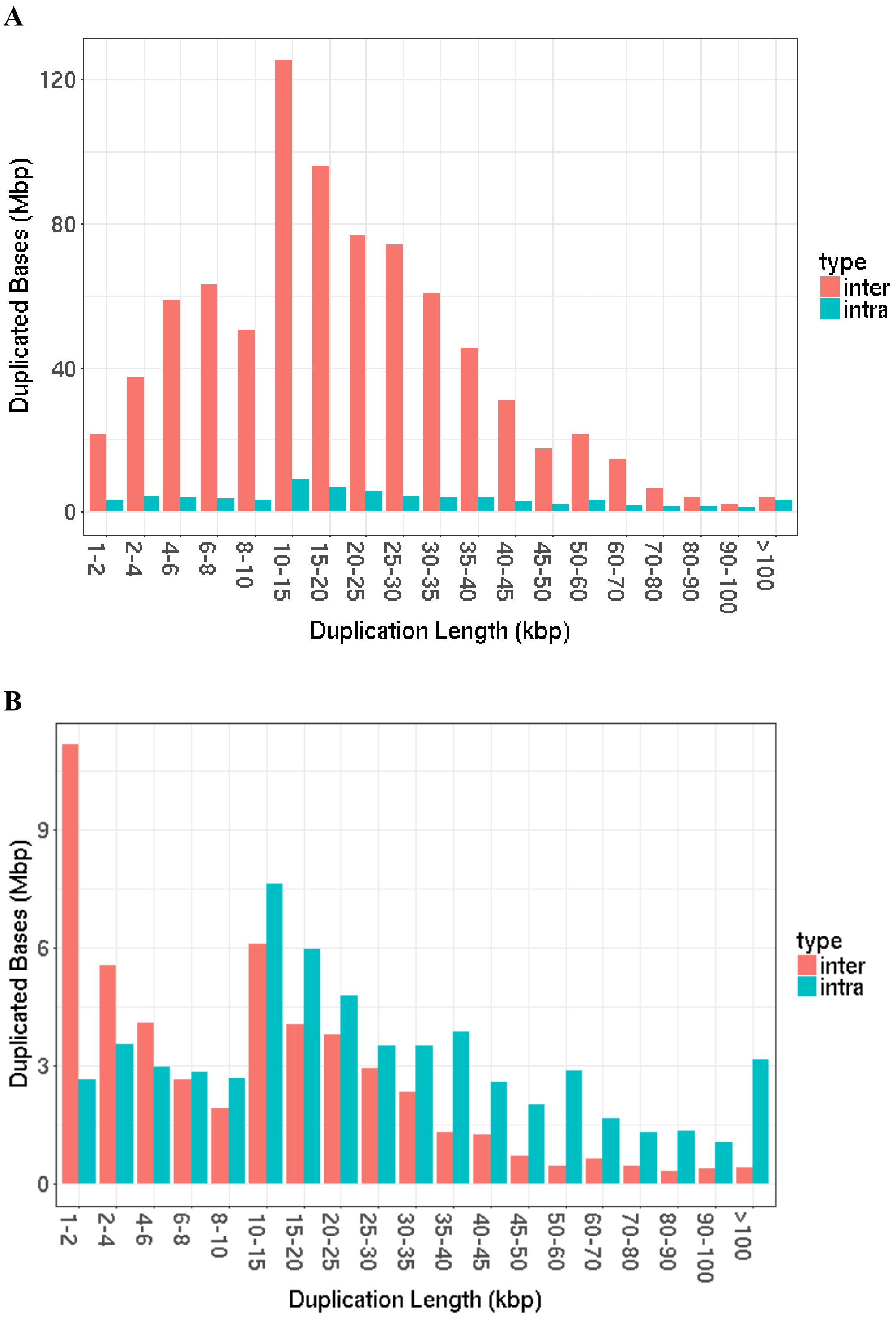


Fig. S8. Macaque segmental duplication length distribution. The number of aligned bases of detected SDs (WGAC) based on the length of alignment: with (A) and without (B) unplaced contigs. Red shows alignments between regions on the same chromosome. Teal shows alignments between regions on different chromosomes (defined as interchromosomal).

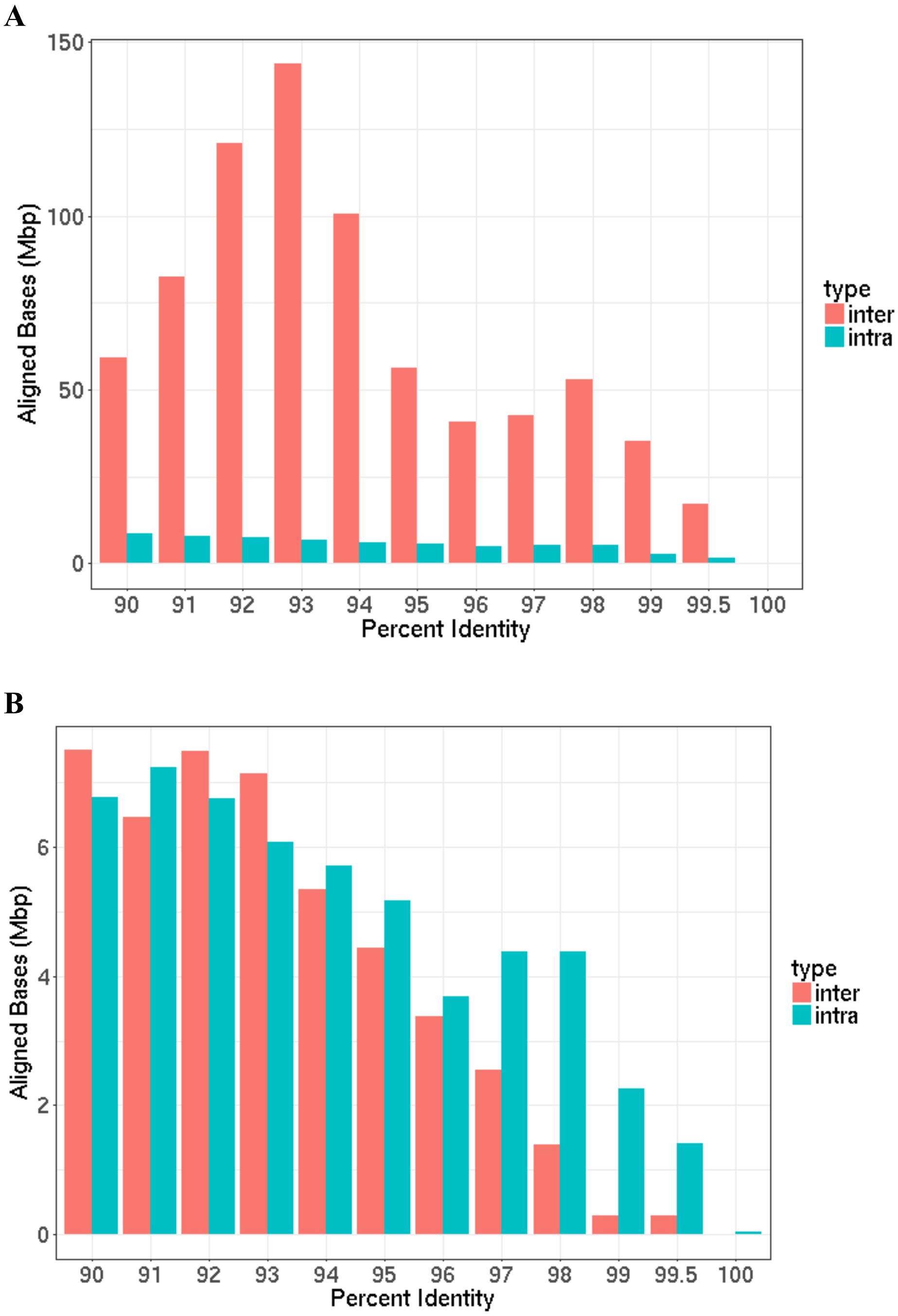


Fig. S9. Macaque segmental duplication percent identity distribution. Aligned bases of duplicated regions by percent identity: with (A) and without (B) unplaced contigs. Duplications between non-homologous (red) and within (teal) homologous chromosomes are shown.

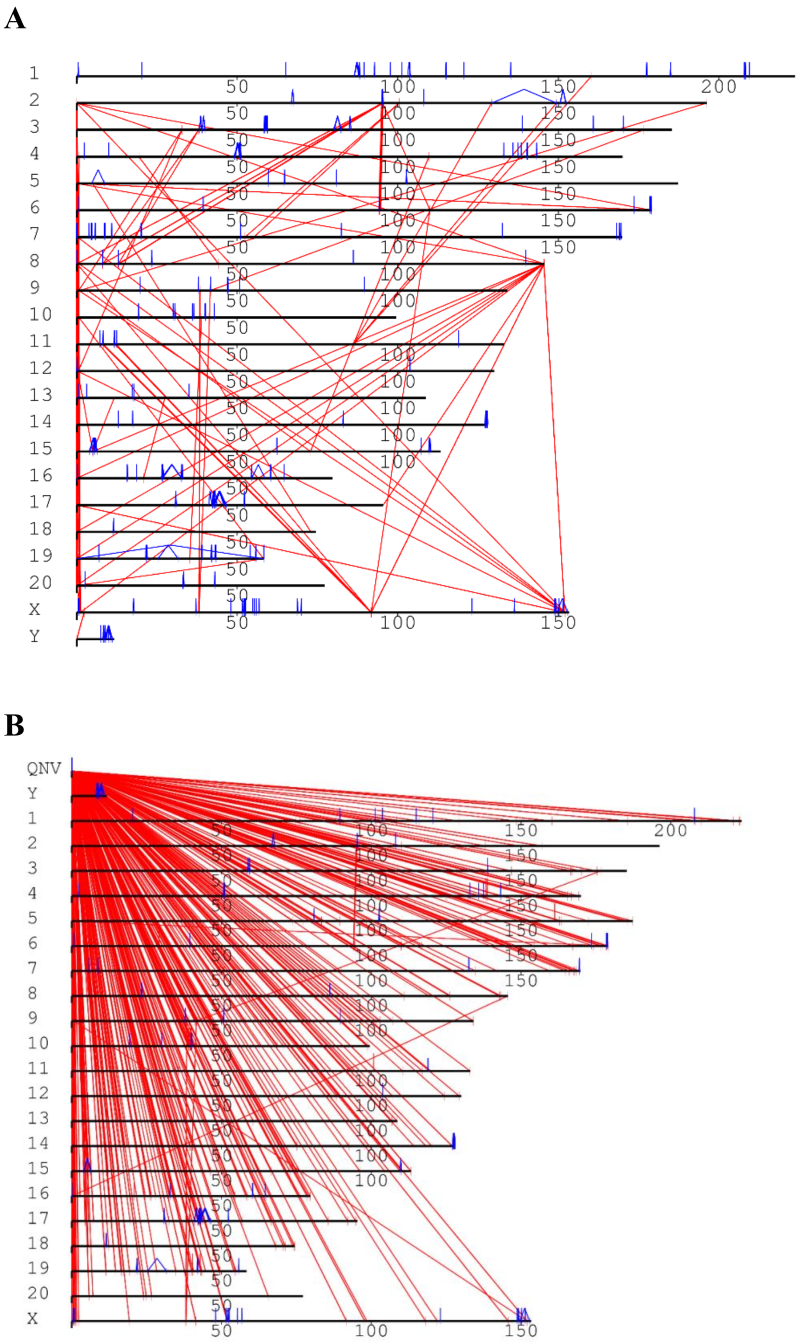


Fig. S10. Genome-wide distribution of segmental duplications. SDs were identified with WGAC. Red lines represent interchromosomal duplications, and blue ticks represent intrachromosomal duplications. (A) Only SDs ≥10 kbp and 95% identical assigned to chromosomes are shown. (B) Only SDs ≥10 kbp and 98% identical are shown. We are not depicting the 2,916 individual unassigned contigs (117 Mbp) but instead project all SDs to a single location labelled here as “QNV”.

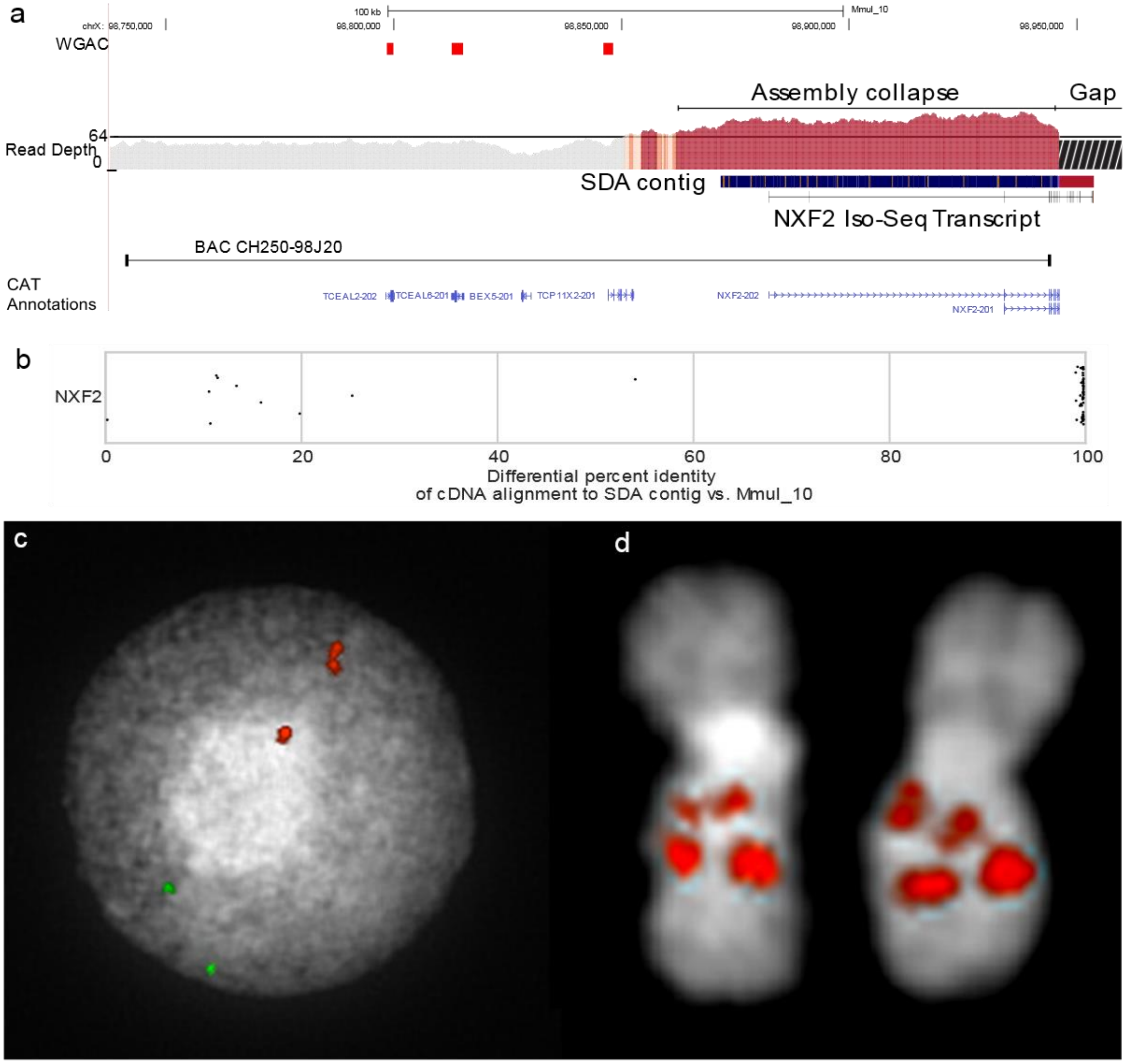


Fig. S11. Resolution and annotation of a collapsed segmental duplication of *NXF2*. (A) The X chromosome of Mmul_10 contains a collapsed duplication corresponding to the Nuclear RNA Export Factor 2 (*NXF2*) locus, as identified by increased read depth. SDA resolves a phased copy of this locus that extends into an assembly gap, better representing the extent of *NXF2* based on mapping of Iso-Seq transcripts. A macaque BAC, CH250-98J20, spans the duplicated *NXF2* locus. (B) Comparison of alignment percent identity between Mmul_10 and the alternate SDA contig demonstrates improved alignment of *NXF2* Iso-Seq transcripts. Transcripts that are <90% aligned to a contig are scored as 0% identity, producing the preponderance of 100% differential percent identity transcripts, as the SDA contig extends into the assembly gap to more fully represent *NXF2*. (C) Interphase FISH image of CH250-98J20 containing the *NXF2* duplication (red) compared to single-copy clone, with single-copy clone CH250-436N5 (green) for comparison. (D) Metaphase FISH image of CH250-98J20 hybridized to chromosome X demonstrates the interstitial intrachromosomal duplication on the q arm of chromosome X.

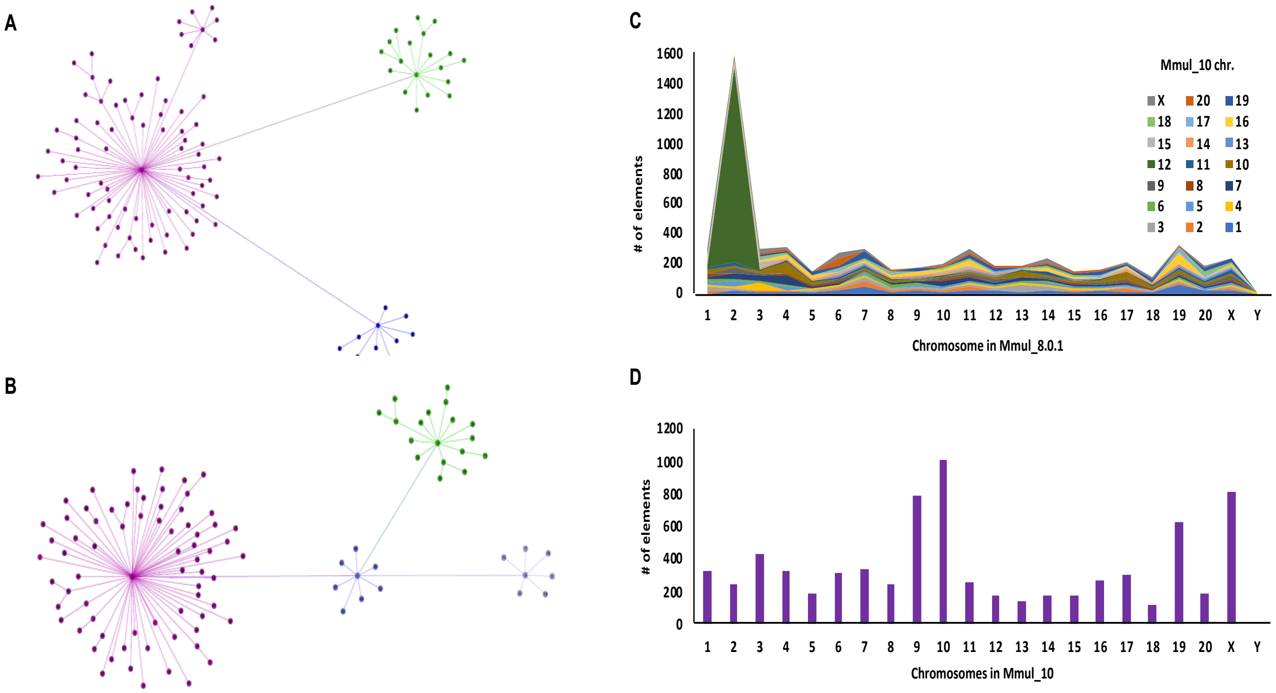


Fig. S12. Full-length *Alu Macaca mulatta* analysis. (A) Mmul_8.0.1 network schematic of the 110 *Alu* subfamilies. (B) Mmul_10 network schematic of the 105 *Alu* subfamilies produced via COSEG and generated in GEPHI. Related subfamilies are clustered together and connected by lines, and all branch out from the central node labeled with subfamily 0 in purple. The two bursts of purple nodes are primarily *Alu*MacYa3-derived subfamilies and blue nodes are *Alu*YRd- or *Alu*Ym1-derived subfamilies. Note: in (B) the light blue nodes centered around subfamily1 are *Alu*Yd2, while the light blue nodes centered around subfamily16 are *Alu*Ym1-derived subfamilies, similar to the Mmul_8.0.1 analysis but split into two separate bursts. The green nodes are a mixture of *Alu*YRc, *Alu*Ym1, *Alu*Yk and *Alu*Yf. Line length between subfamilies is not indicative of number of mutations or evolutionary time between subfamilies. (C) The x-axis shows each chromosome in Mmu_8.0.1 that contained *Alu* elements that post liftOver were found on a different chromosome in Mmul_10. The stacked colors above each Mmu_8.0.1 chromosome represent the liftOver Mmul_10 chromosome in consecutive order: chr1 to chrX from bottom to top. Note the striking redistribution of *Alu* elements from chr2 in Mmul_8.0.1 to chr12 in Mmul_10, as shown in dark green. (D) *Alu* elements on unplaced chromosomes (Un) in Mmu_8.0.1 (n = 8,291) are now on placed chromosomes in Mmul_10, distributed across all chromosomes.

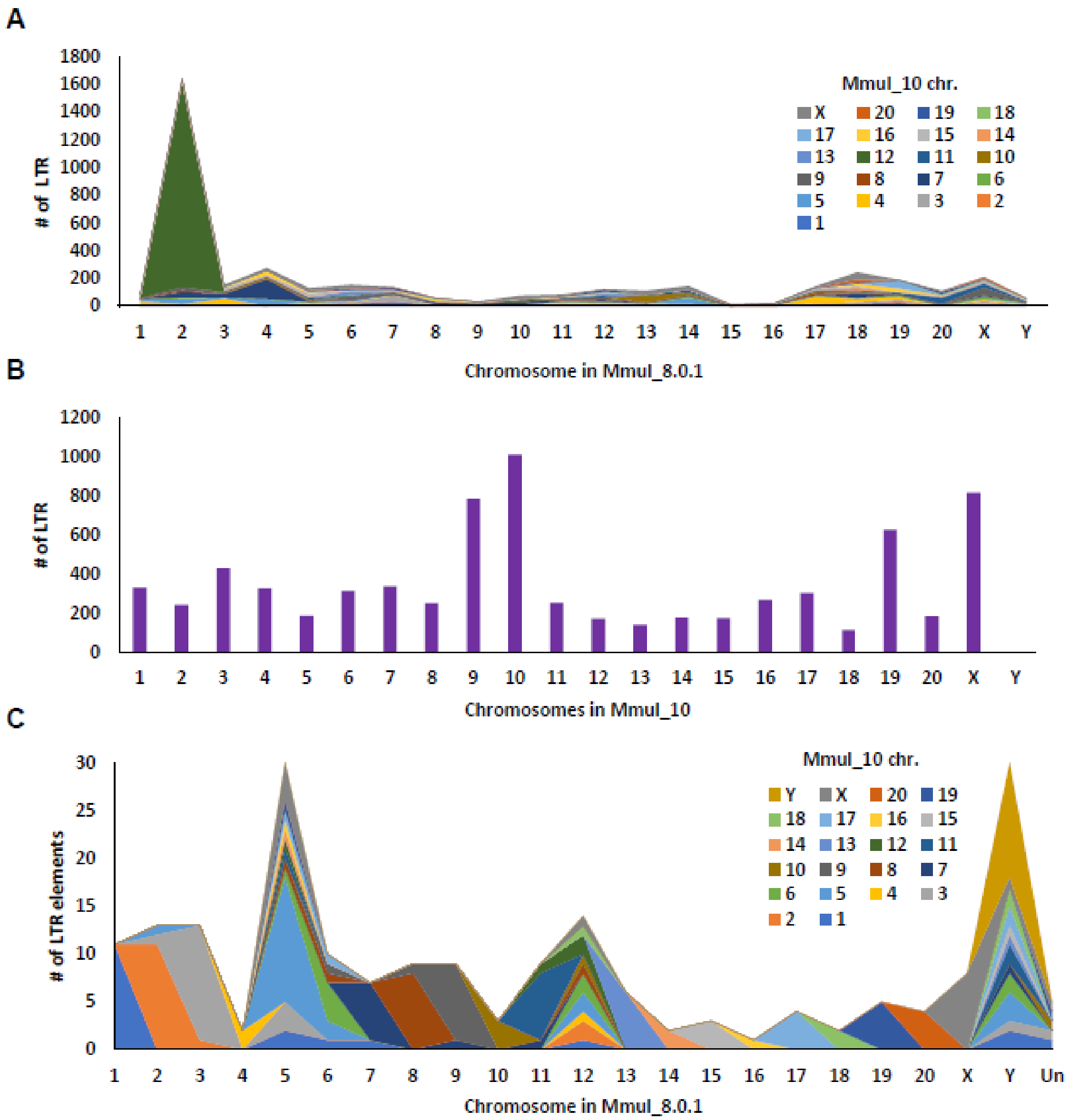


Fig. S13. LTR Mmul_10 to Mmul_8.0.1 liftOver analysis. (A) The x-axis shows each chromosome in Mmul_8.0.1 containing LTR elements that post liftOver were found on a different chromosome in Mmul_10. The stacked colors above each Mmul_8.0.1 chromosome represent the liftOver Mmul_10 chromosome in consecutive order, chr1 to chrX from bottom to top. Note the striking redistribution of LTR elements from chr2 in Mmul_8.0.1 to chr12 in Mmul_10, as shown in dark green. (B) LTR elements on unplaced (Un) chromosomes in Mmul_8.0.1 (n=7,430) are now placed on chromosomes in Mmul_10, distributed across all chromosomes. (C) Potentially full-length LTR elements presenting a differing chromosomal placement between the Mmul_8.0.1 and Mmul_10 assemblies.

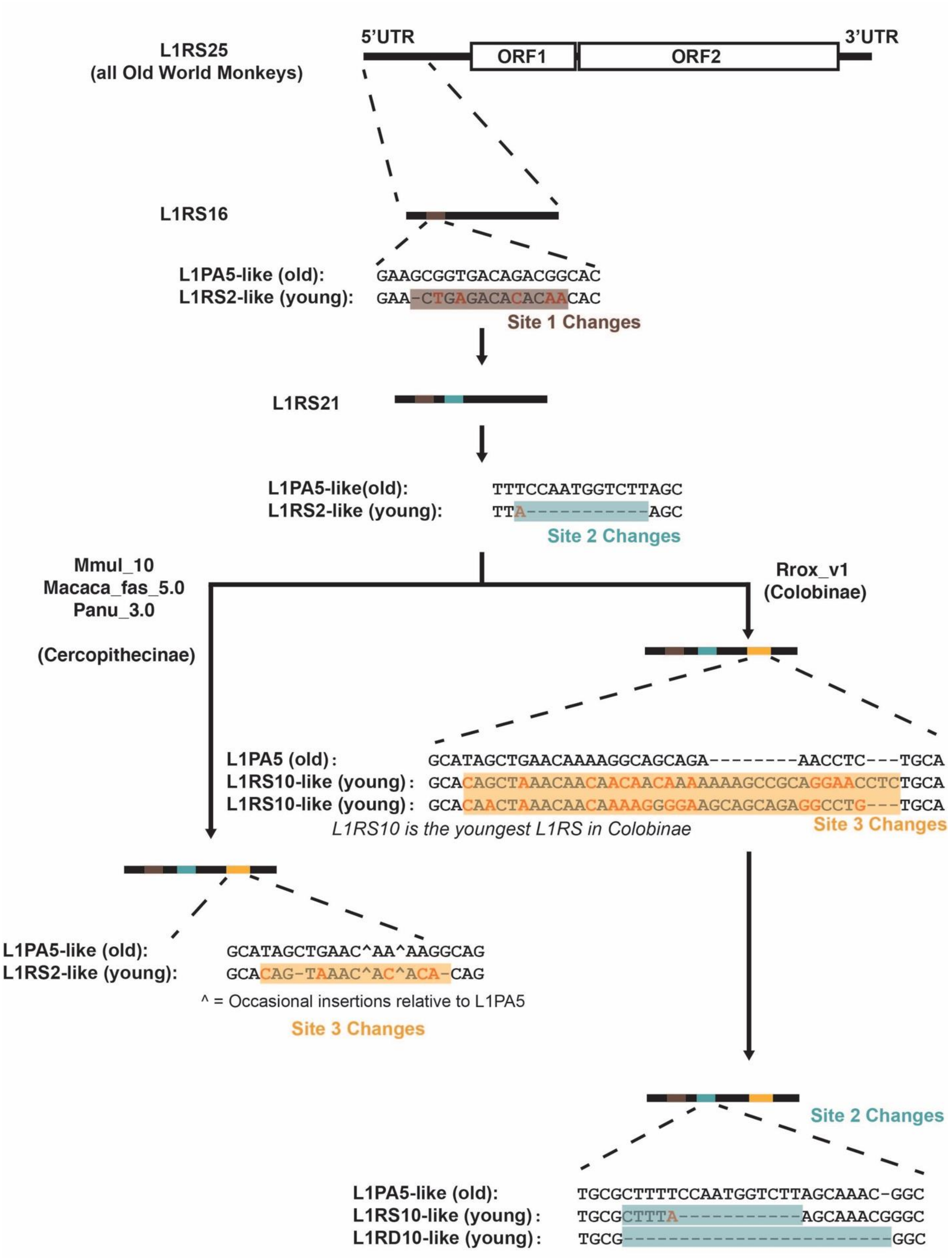

L1RS25
(all Old World Monkeys)
5'UTR
ORF1
ORF2
3'UTR
L1RS16
L1PA5-like (old): GAAGCGGTGACAGACGGCAC
L1RS2-like (young): GAA-CTGAGACACACAACAC
Site 1 Changes
L1RS21
L1PA5-like(old): TTTCCAATGGTCTTAGC
L1RS2-like (young): TTA------------AGC
Site 2 Changes
Mmul_10
Macaca_fas_5.0
Panu_3.0
(Cercopithecinae)
Rrox_v1
(Colobinae)
L1PA5 (old): GCATAGCTGAACAAAAGGCAGCAGA--------AACCTC---TGCA
L1RS10-like (young): GCACAGCTAAACAACAACAACAAAAAAAGCCGCAGGAACCTCTGCA
L1RS10-like (young): GCACAACTAAACAACAAAAGGGGAAGCAGCAGAGGCCTG---TGCA
L1RS10 is the youngest L1RS in Colobinae
Site 3 Changes
L1PA5-like (old): GCATAGCTGAAC^AA^AAGGCAG
L1RS2-like (young): GCACAG-TAAAC^AC^ACA-CAG
^ = Occasional insertions relative to L1PA5
Site 3 Changes
Site 2 Changes
L1PA5-like (old): TGCGCTTTTCCAATGGTCTTAGCAAAC-GGC
L1RS10-like (young): TGCGCTTTA-----------AGCAAACGGGC
L1RD10-like (young): TGCG------------------------GGC

Fig. S14. Schematic of evolutionary changes to the 5' UTR of the L1RS subfamily. Detailed view of representative nucleotide changes that lead to coverage drops in L1RS elements. In order to explore the exact sequence changes at each site, we took a representative random sample of ten L1RS2 instances in Mmul_10 and ten L1RS10 instances in Rrox_v1 (note that the L1RS2 family evolved after the divergence of these two species, therefore there are no L1RS2 instances in Rrox_v1). Since these young families contain the changes that lead to the coverage drops at Sites 1, 2 and 3, we compared the sequences in these instances to that of the ancestral L1PA5 sequence. Our analysis shows L1RS elements in all OWMs experience the same 5' UTR changes at Site 1, which consists of a small deletion and multiple substitutions. All OWMs also show evidence for a small 11 bp deletion at Site 2; however, in Rrox_v1 a larger 24 bp deletion at the same site is more prevalent and appears to be Colobinae-specific. This larger deletion likely occurred after changes at Site 3 as well as the original 11 bp deletion. At Site 3, both Mmul_10 (and other Cercopithecinae) acquire a multitude of changes include insertions and deletions, while Rrox_v1 elements experience changes over a larger region.

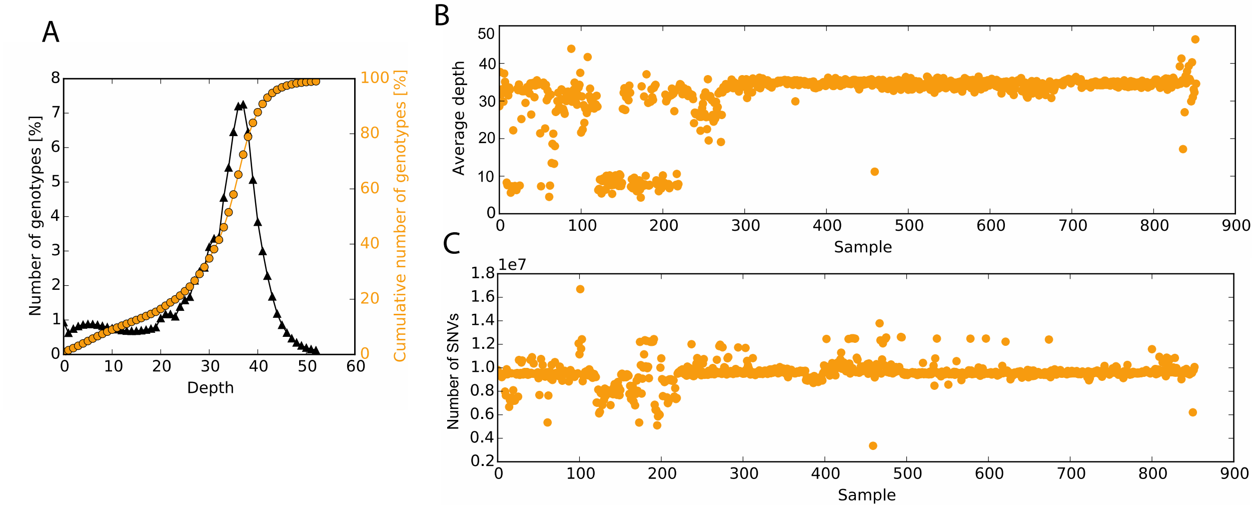


Fig. S15. Macaque SNVs based on WGS and coverage. (A) The percent of genotypes (black) and cumulative percent of genotypes (yellow) covered at the given read coverage depth. (B) The read coverage depth averaged across SNVs for the 853 rhesus samples subjected to WGS. (C) The number of SNVs identified by sample. BCFtools (http://samtools.github.io/bcftools/bcftools.html#stats) were used to calculate statistics based on a GATK-generated VCF file and plot-vcfstats was used to generate graphs.

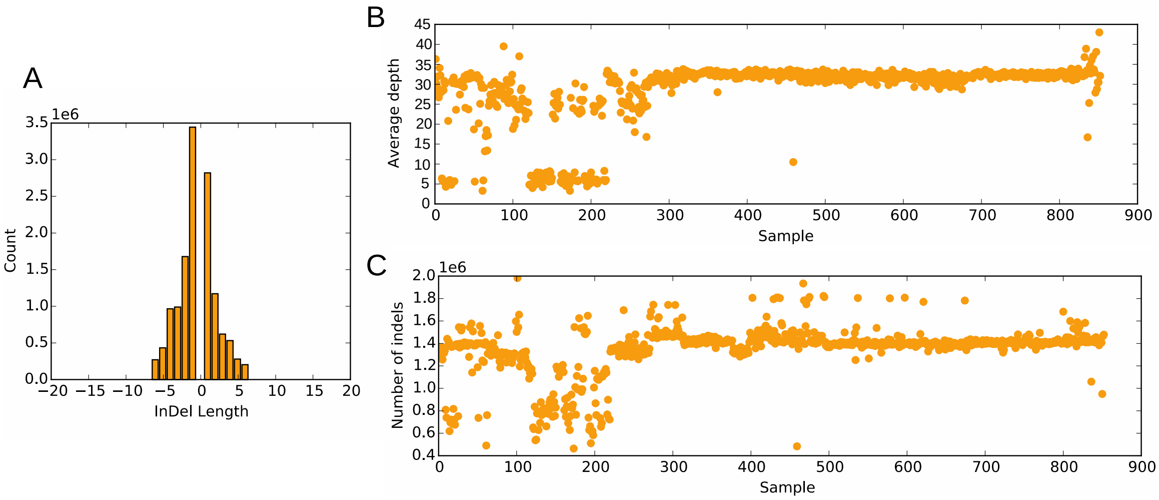


Fig. S16. Macaque indels based on WGS and coverage. (A) The percent of genotypes (black) and cumulative percent of genotypes (yellow) covered at the given read coverage depth. (B) The read coverage depth averaged across indels for the 853 rhesus samples. (C) The number of indels identified by sample. BCFtools stats (http://samtools.github.io/bcftools/bcftools.html#stats) were used to calculate statistics based on a GATK-generated VCF file and plot-vcfstats was used to generate graphs.

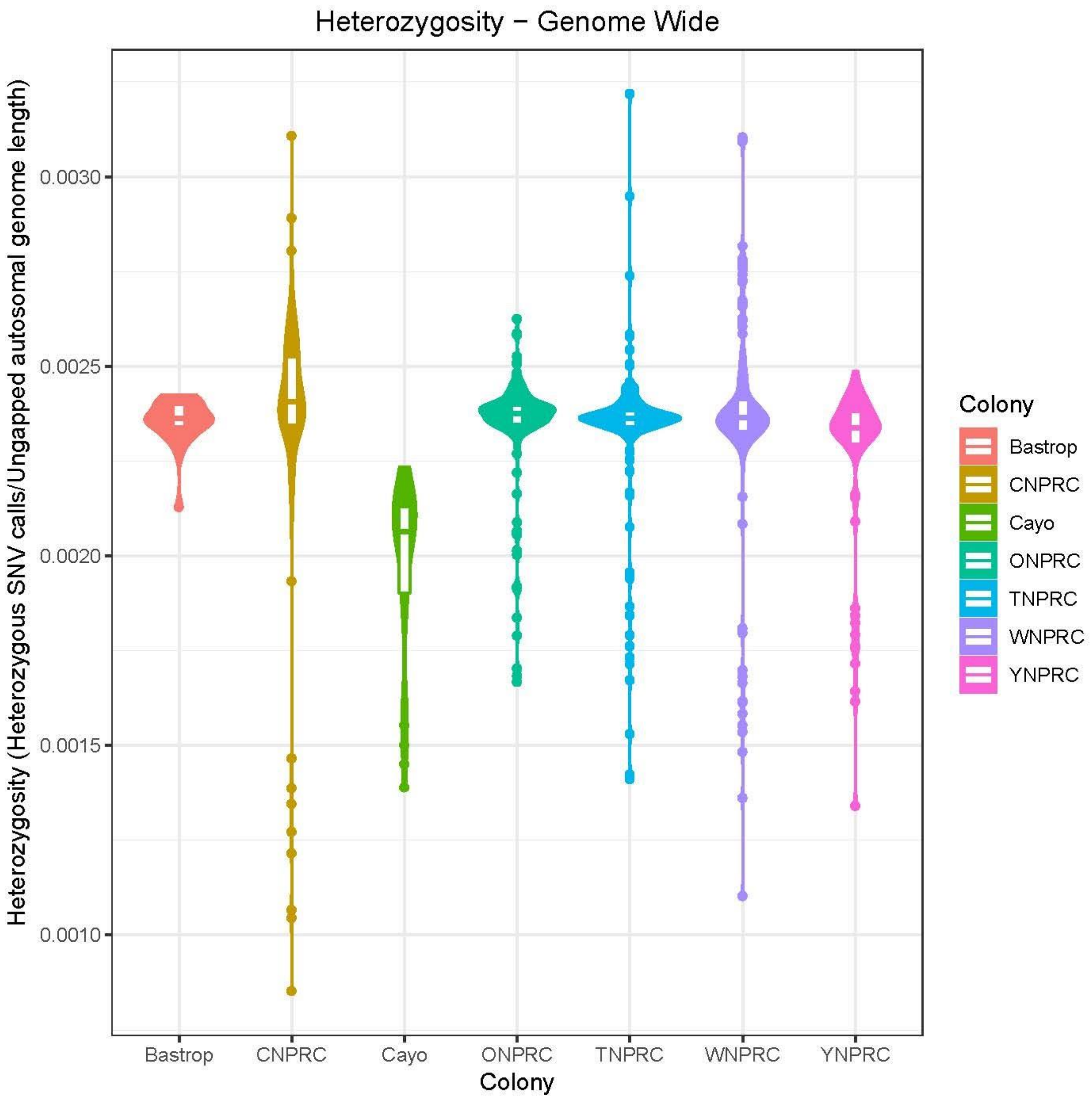


Fig. S17. Genome-wide heterozygosity of US research colonies compared to the isolated Cayo Santiago population. Heterozygosity was calculated as (autosomal heterozygous SNV calls/ungapped autosomal assembly length) for each individual and then plotted separately for each research colony. The Cayo Santiago population shows relatively lower heterozygosity suggesting reduced diversity.

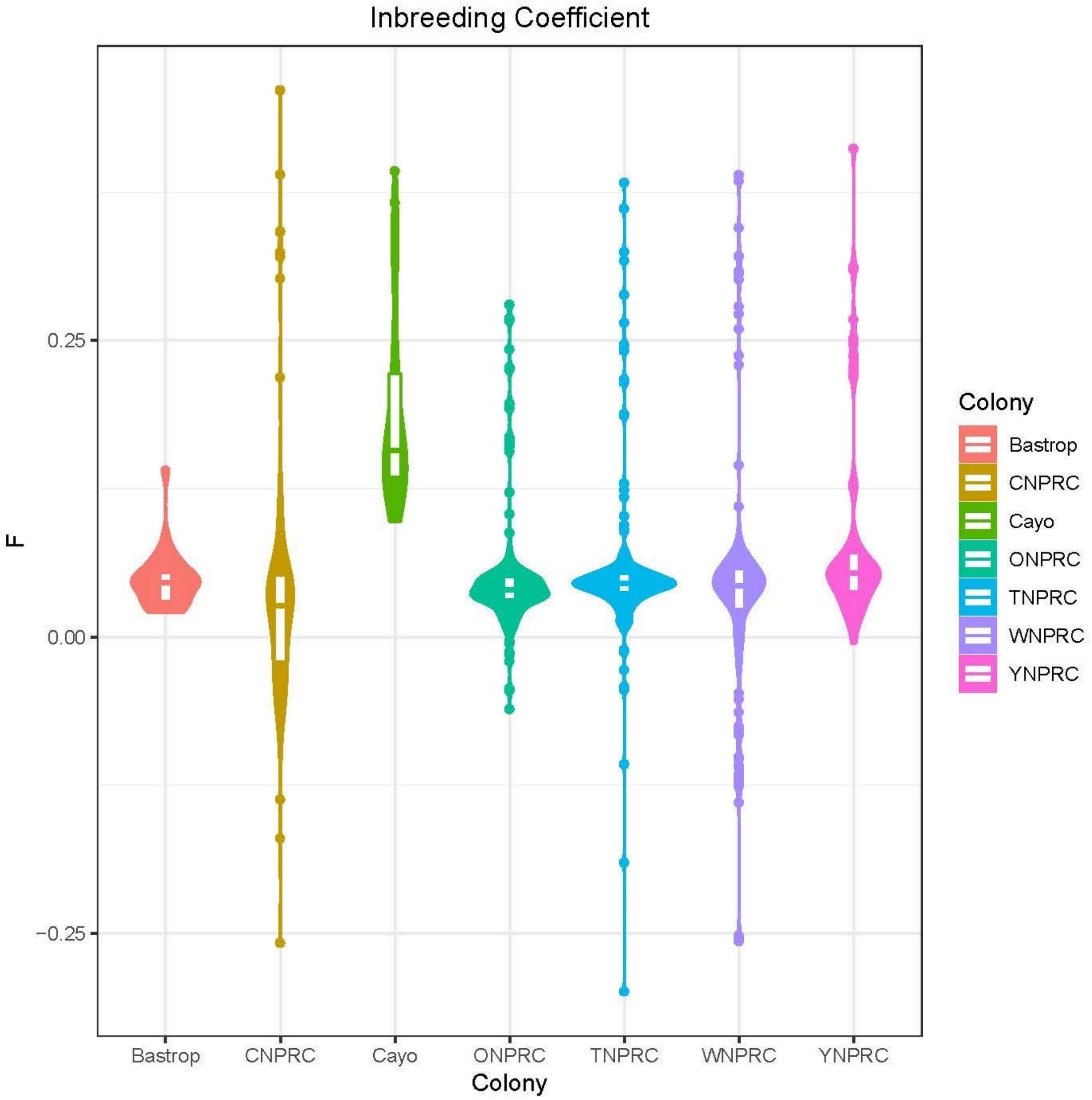


Fig. S18. Inbreeding coefficient estimates of US research colonies compared to the isolated Cayo Santiago population. Method-of-moments F inbreeding coefficient estimates were calculated with PLINK as (observed homozygous SNVs - expected homozygous SNVs) / (total called SNVs - expected homozygous SNVs) for each individual and then plotted separately for each research colony. The Cayo Santiago population shows relatively higher inbreeding coefficients suggesting reduced diversity.

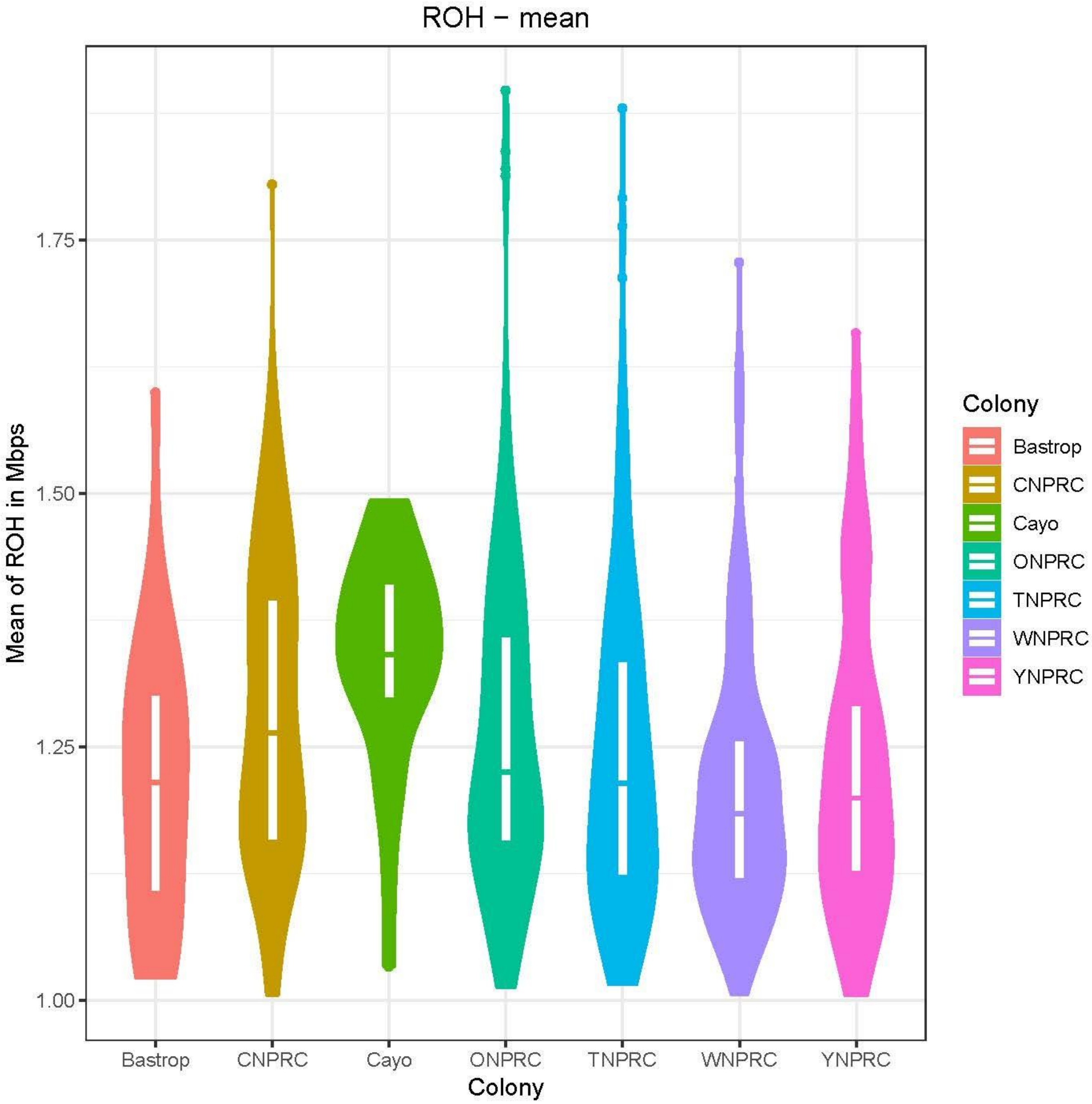


Fig. S19. Estimated runs of homozygosity (ROH) as calculated by PLINK showing the distribution of ROH lengths per US research colony compared to the isolated Cayo Santiago population. The mean of run lengths per individual sample is used to characterize the distribution of ROH. The Cayo Santiago population shows relatively longer ROH mean lengths suggesting reduced diversity.

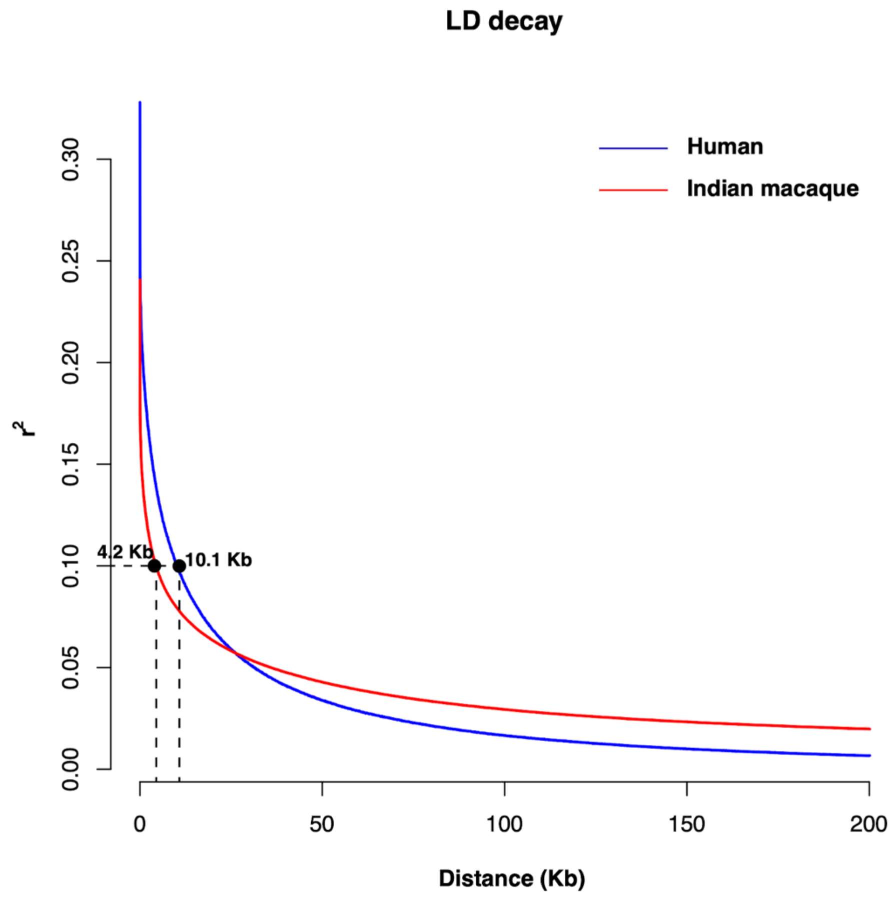


Fig. S20. Linkage disequilibrium (LD) decay of Indian macaques compared to human. We selected 411 unrelated Indian macaques and assessed the pattern of LD. We applied PopLDdecay (*68*) to calculate the correlation coefficient (R2) among pairwise SNVs and used the mean-bin method to plot the distance of LD decay among adjacent SNVs. As a comparison, we repeated the same analysis using human population data (399 African samples and 1 European sample). LD was estimated for pairwise SNVs within a window of 200 kbp. Among the Indian macaques, we estimate an average LD (r2 = 0.038) between SNVs and find that at an r2 = 0.1 the average distance between two SNVs is ~4.2 kbp.

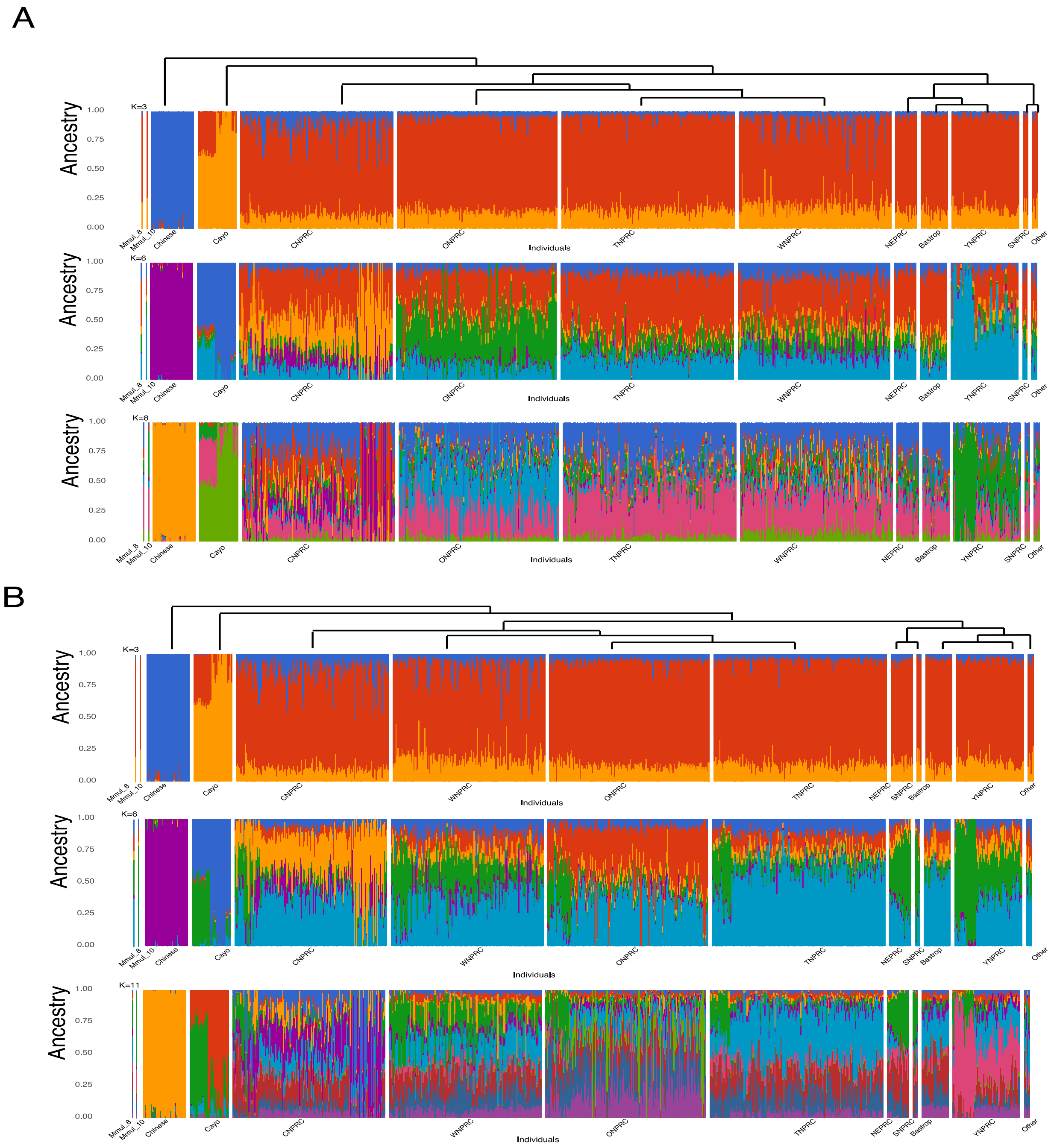


Fig. S21. ADMIXTURE analysis of rhesus macaque populations. (A) ADMIXTURE analysis results based on SNVs identified from WGS data for K = 3, 6, and 8 (n = 853, SNVs = 14,128,568). K = 8 is the best model fitting to our data based on cross-validation experiments. Populations were clustered (dendogram) based on pairwise Fst matrix. The y-axis shows the proportions of the evolutionary clustering components.
(B) ADMIXTURE analysis based on a LD pruning dataset for K=3, 6, 11 (n = 853, SNVs = 193,684). The K = 11 is the best model fitting to our data based on cross-validation.

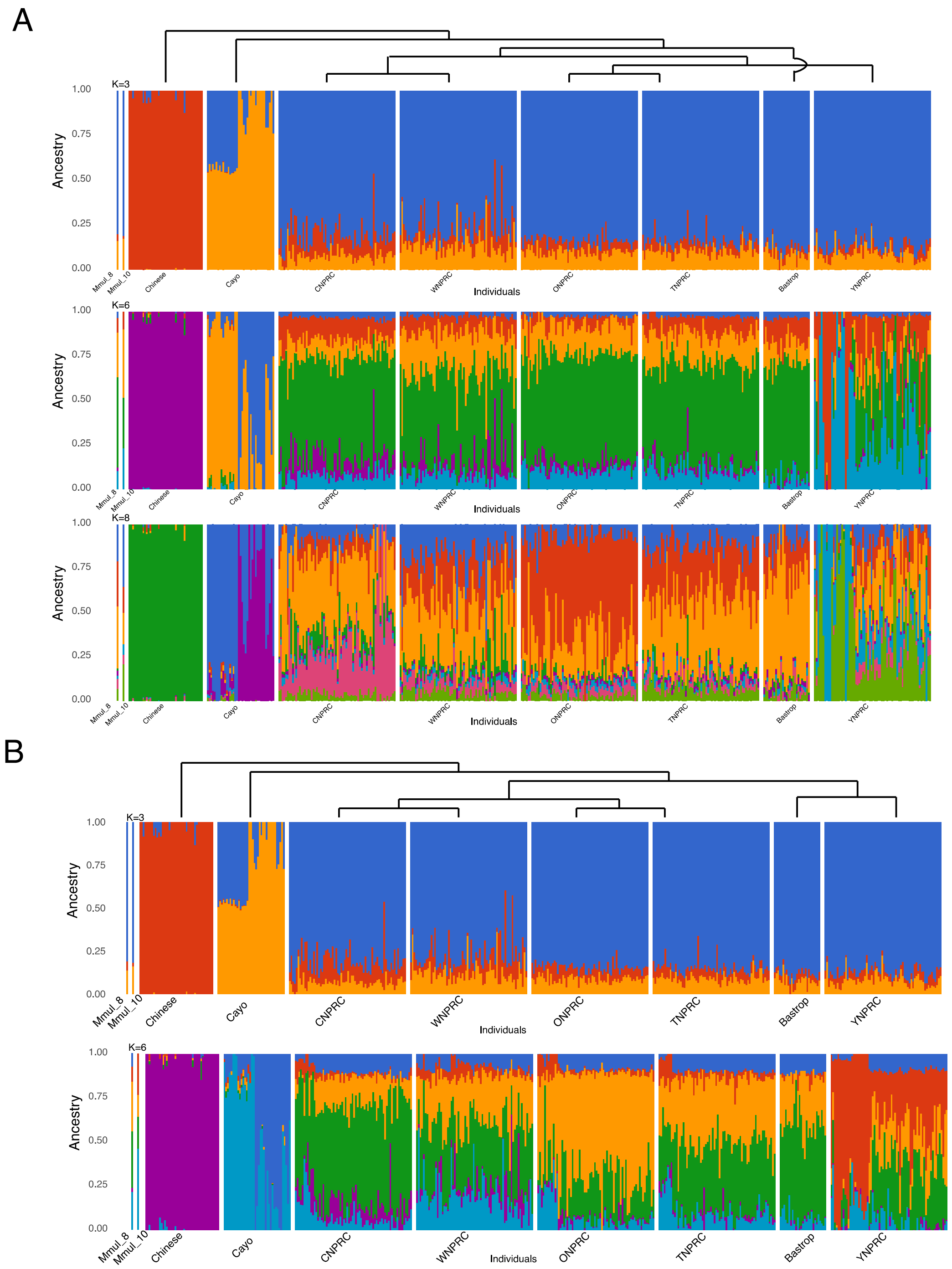


Fig. S22. ADMIXTURE analysis based on down sampled rhesus macaque populations. Analysis is the same as Fig. S23 except populations from different centers were down sampled to be the same size. (A) ADMIXTURE analysis results for down sampled size dataset from K = 3, 6, 8 (n = 484, SNVs = 14,128,568). The K = 8 is the best model fitting to our data by cross-validation. (B) ADMIXTURE analysis results for the down sampled LD pruning dataset from K = 3, 6 (n = 484, SNVs = 193,684). The K = 6 is the best model fitting to our data by cross-validation. Note: SNPRC, NEPRC, and other groups are not represented.

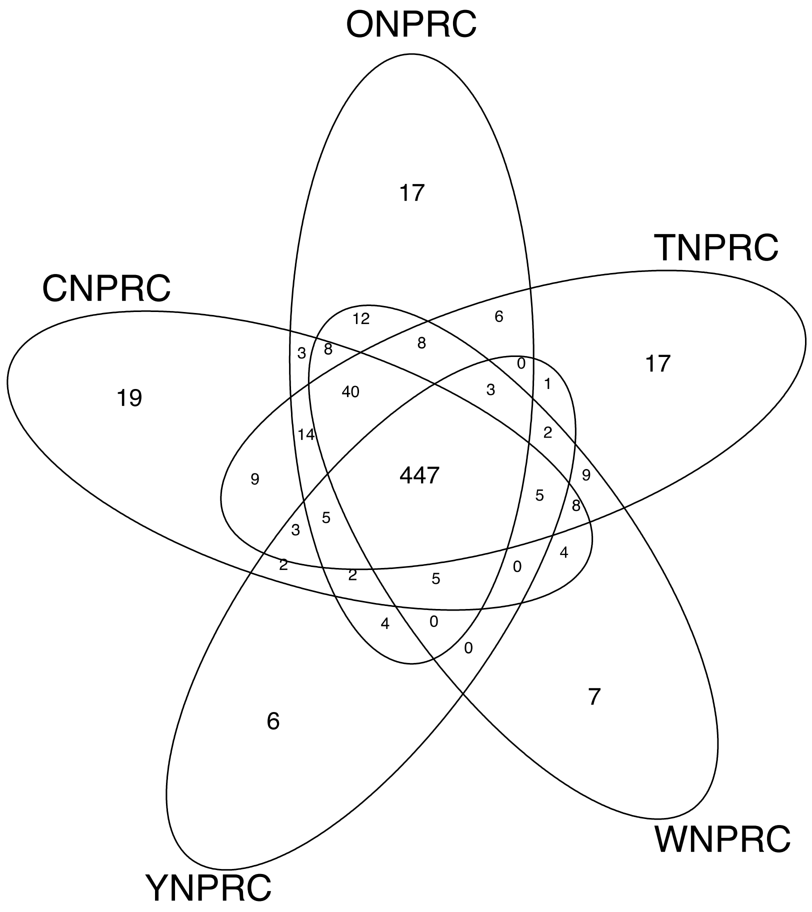


Fig. S23. Venn diagram of homozygous likely gene-disruptive (LGD) variants among rhesus macaque research centers. Singleton LGD variants from each of the five primate centers previously described in the main text were removed.

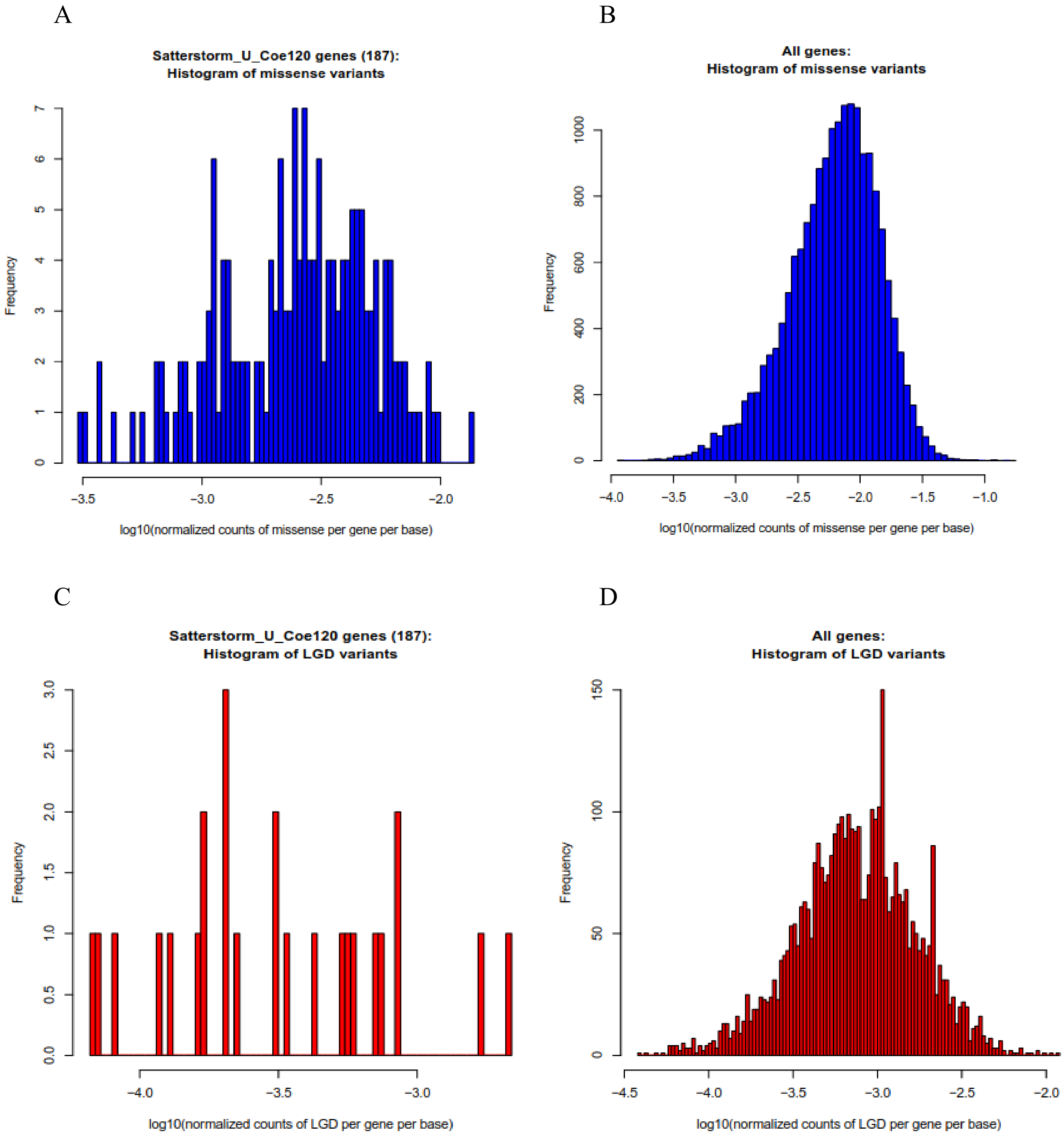
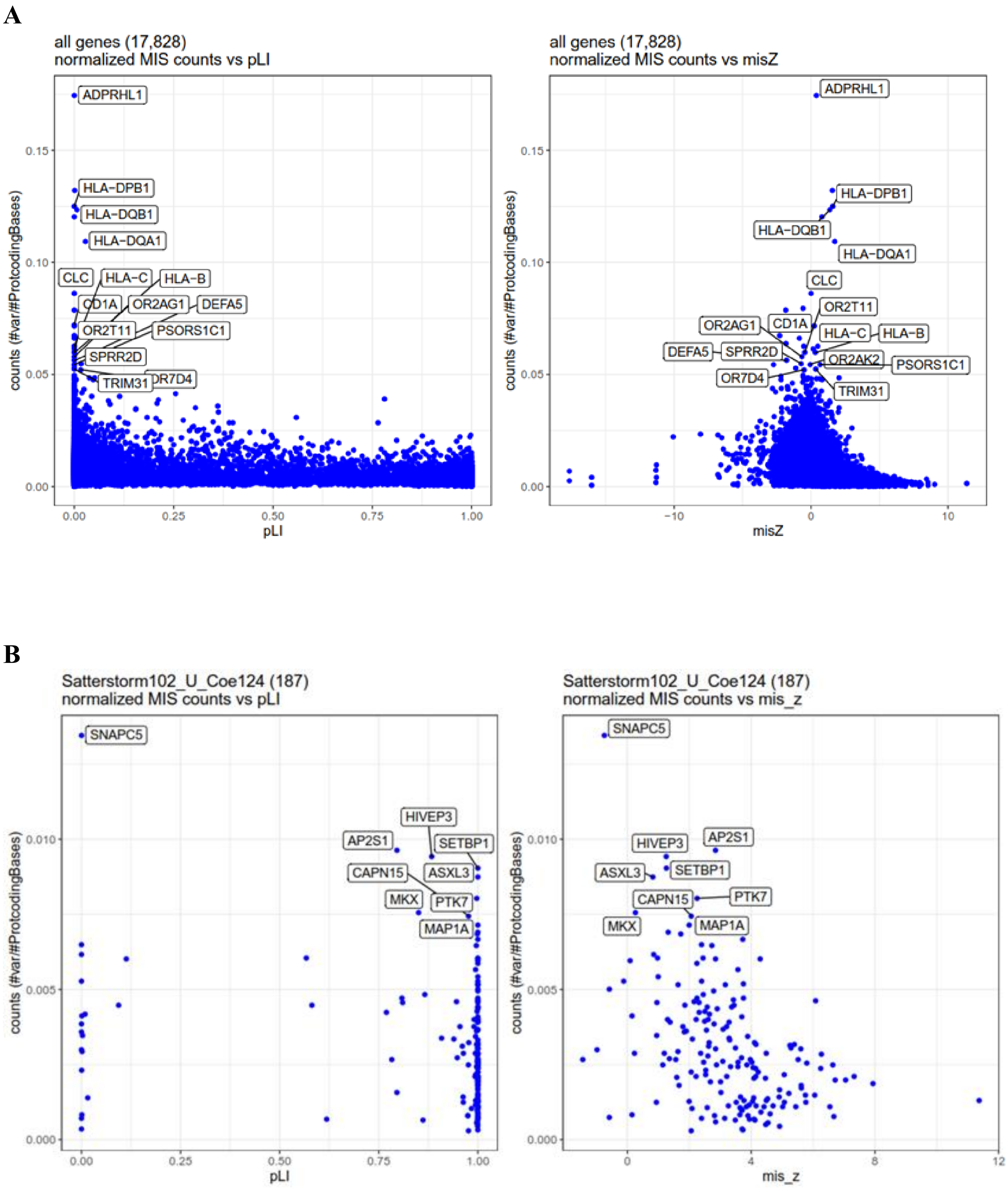


Fig. S24. The frequency of predicted protein-altering variants by gene. Histogram of SNV counts per gene per base for (A) missense variants in all orthologous genes (n = 17,828) and (B) missense variants in genes associated with NDDs (n = 187), (C) LGD variants for all genes, and (D) LGD variants in NDD genes.


A
all genes (17,828)
normalized MIS counts vs pLI
counts (#var/#ProtcodingBases)
pLI
0.00
0.05
0.10
0.15
0.00
0.25
0.50
0.75
1.00
ADPRHL1
HLA-DPB1
HLA-DQB1
HLA-DQA1
CLC
HLA-C
HLA-B
CD1A
OR2AG1
DEFA5
OR2T11
PSORS1C1
SPRR2D
TRIM31
OR7D4
all genes (17,828)
normalized MIS counts vs misZ
counts (#var/#ProtcodingBases)
misZ
−10
0
10
ADPRHL1
HLA-DPB1
HLA-DQB1
HLA-DQA1
CLC
OR2T11
OR2AG1
CD1A
HLA-C
HLA-B
DEFA5
SPRR2D
OR2AK2
PSORS1C1
OR7D4
TRIM31
B
Satterstorm102_U_Coe124 (187)
normalized MIS counts vs pLI
counts (#var/#ProtcodingBases)
pLI
0.000
0.005
0.010
0.00
0.25
0.50
0.75
1.00
SNAPC5
HIVEP3
AP2S1
SETBP1
CAPN15
ASXL3
MKX
PTK7
MAP1A
Satterstorm102_U_Coe124 (187)
normalized MIS counts vs mis_z
counts (#var/#ProtcodingBases)
mis_z
0
4
8
12
SNAPC5
HIVEP3
AP2S1
ASXL3
SETBP1
CAPN15
PTK7
MKX
MAP1A

C

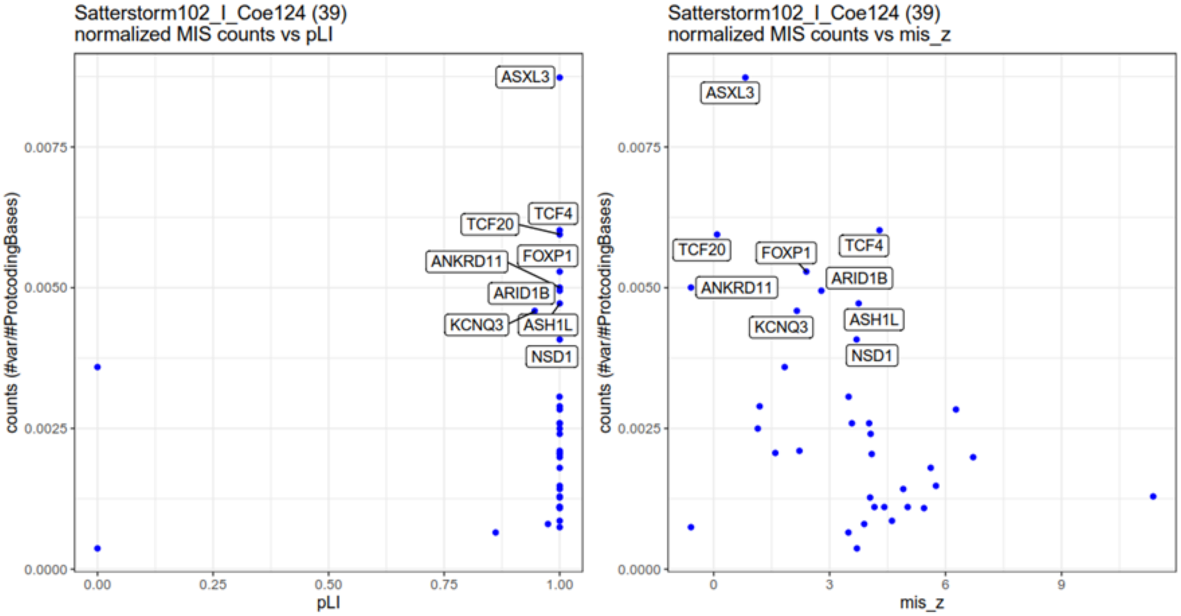
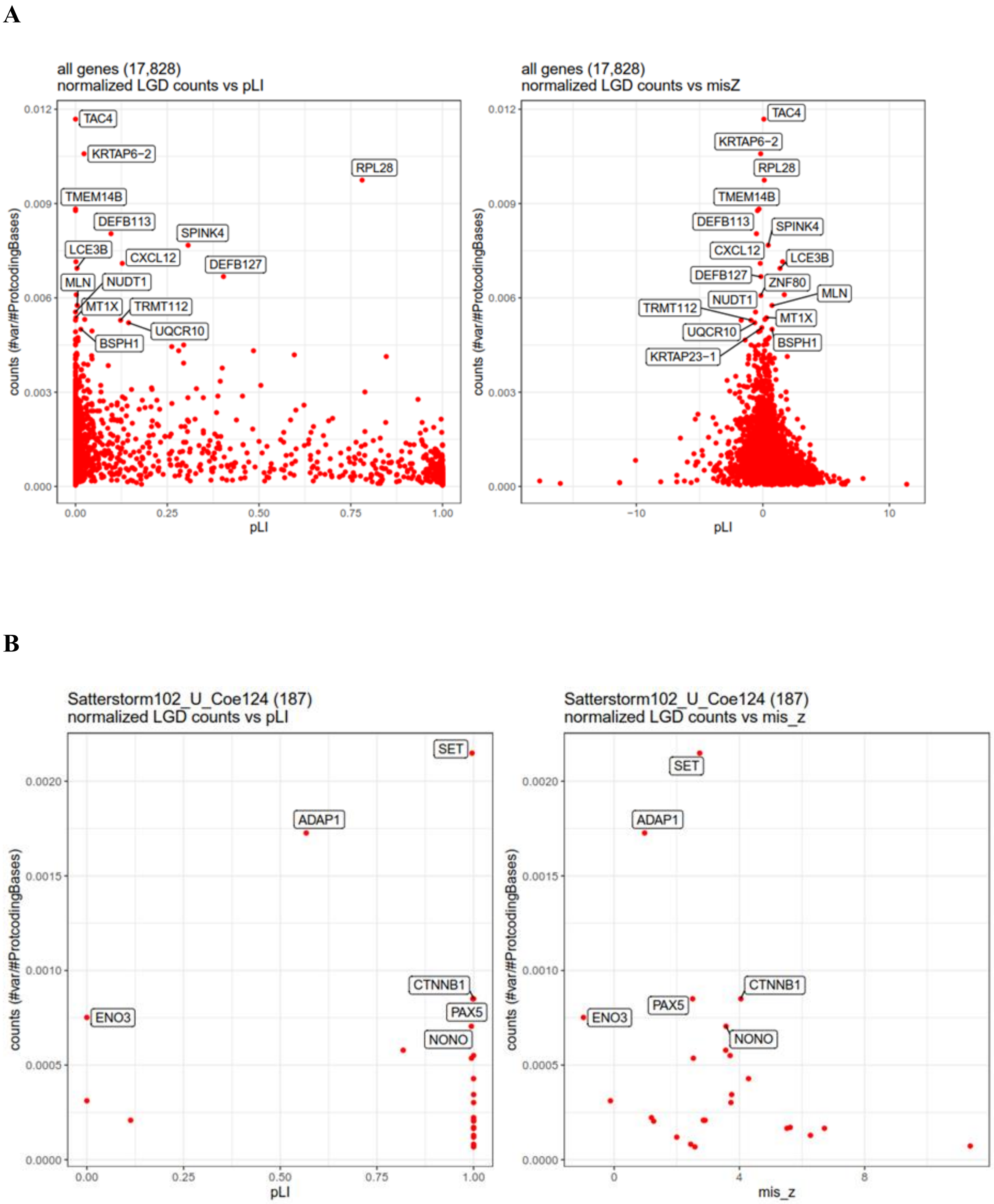


Fig. S25. Distribution of macaque missense variants based on intolerance to mutation. Missense counts per base plotted by pLI and missense mutations (mis_z) in humans: (A) All genes, (B) NDD gene set is based on the union of (*32*) and (*33*), and (C) NDD gene set is based on the intersection of (*32*) and (*33*). Duplicated genes are excluded. Genes with normalized counts >= 0.005 are highlighted.


A
all genes (17,828)
normalized LGD counts vs pLI
counts (#var/#ProtcodingBases)
0.012
0.009
0.006
0.003
0.000
TAC4
KRTAP6−2
RPL28
TMEM14B
DEFB113
SPINK4
LCE3B
CXCL12
DEFB127
MLN
NUDT1
MT1X
TRMT112
UQCR10
BSPH1
0.00
0.25
0.50
0.75
1.00
pLI
all genes (17,828)
normalized LGD counts vs misZ
counts (#var/#ProtcodingBases)
0.012
0.009
0.006
0.003
0.000
TAC4
KRTAP6−2
RPL28
TMEM14B
DEFB113
SPINK4
CXCL12
LCE3B
DEFB127
ZNF80
MLN
TRMT112
NUDT1
MT1X
UQCR10
BSPH1
KRTAP23−1
−10
0
10
pLI
B
Satterstorm102_U_Coe124 (187)
normalized LGD counts vs pLI
counts (#var/#ProtcodingBases)
0.0020
0.0015
0.0010
0.0005
0.0000
SET
ADAP1
CTNNB1
PAX5
NONO
ENO3
0.00
0.25
0.50
0.75
1.00
pLI
Satterstorm102_U_Coe124 (187)
normalized LGD counts vs mis_z
counts (#var/#ProtcodingBases)
0.0020
0.0015
0.0010
0.0005
0.0000
SET
ADAP1
CTNNB1
PAX5
NONO
ENO3
0
4
8
mis_z

C

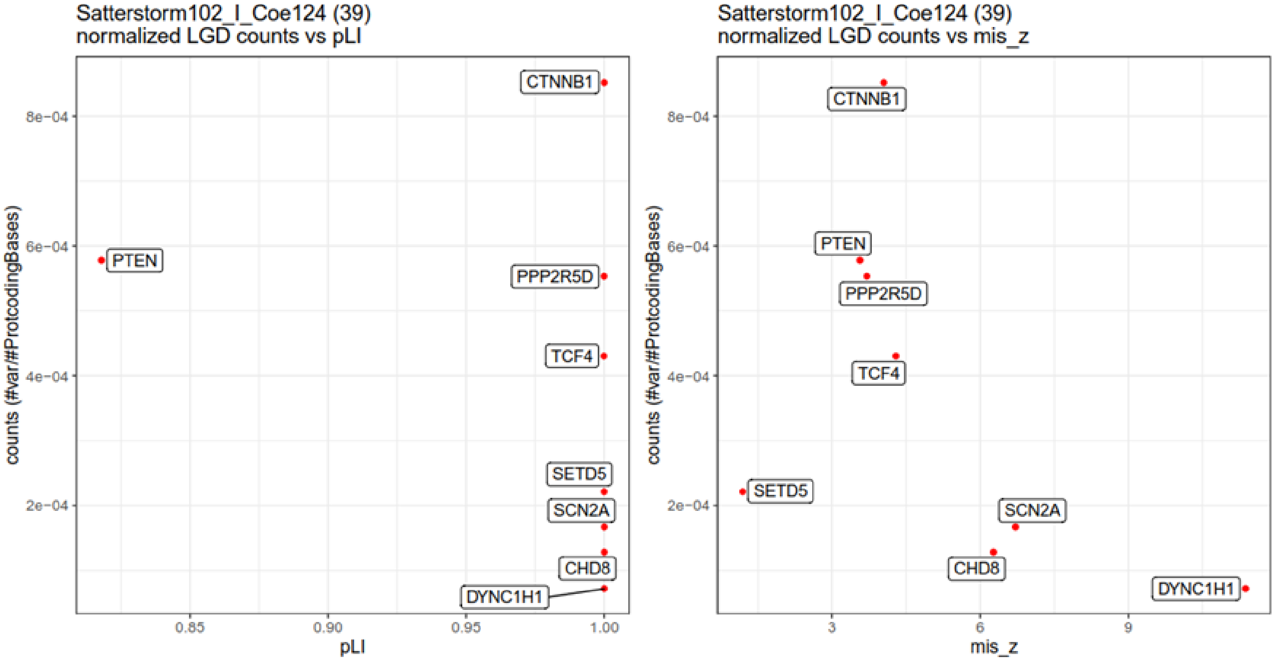


Fig. S26. Distribution of macaque LGD variants based on intolerance to mutation. LGD counts per base plotted by pLI and missense mutations (mis_z) in humans: (A) All genes, (B) NDD gene set is based on the union of (*32*) and (*33*), and (C) NDD gene set is based on the intersection of (*32*) and (*33*). Duplicated genes are excluded. Genes with normalized counts >= 0.005 are highlighted.

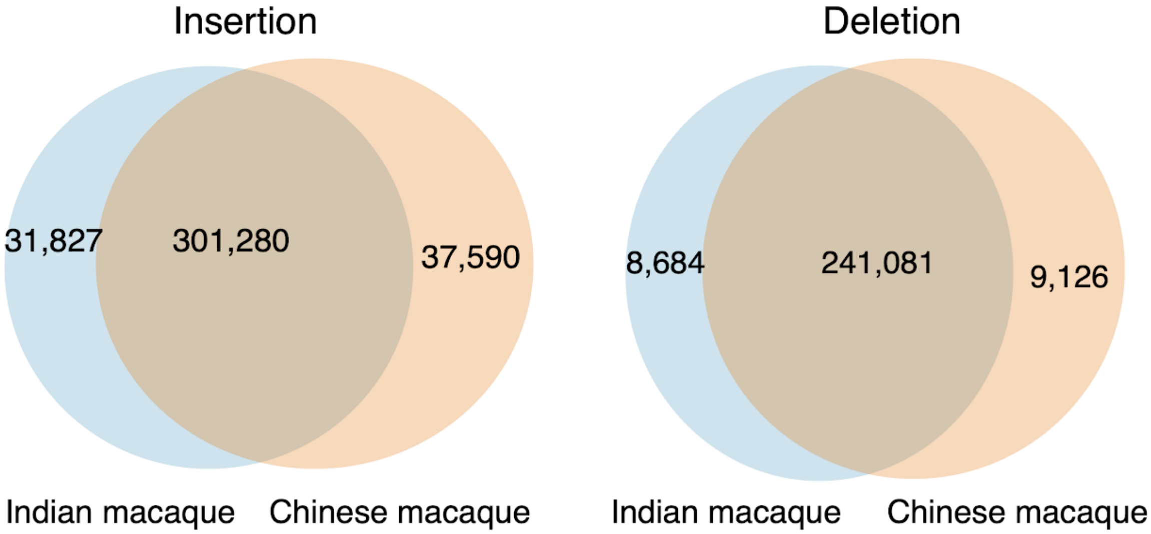


Figure S27. Shared and unique SVs in the reference genomes of Indian (Mmul_10) and Chinese (rheMacS) rhesus macaques.

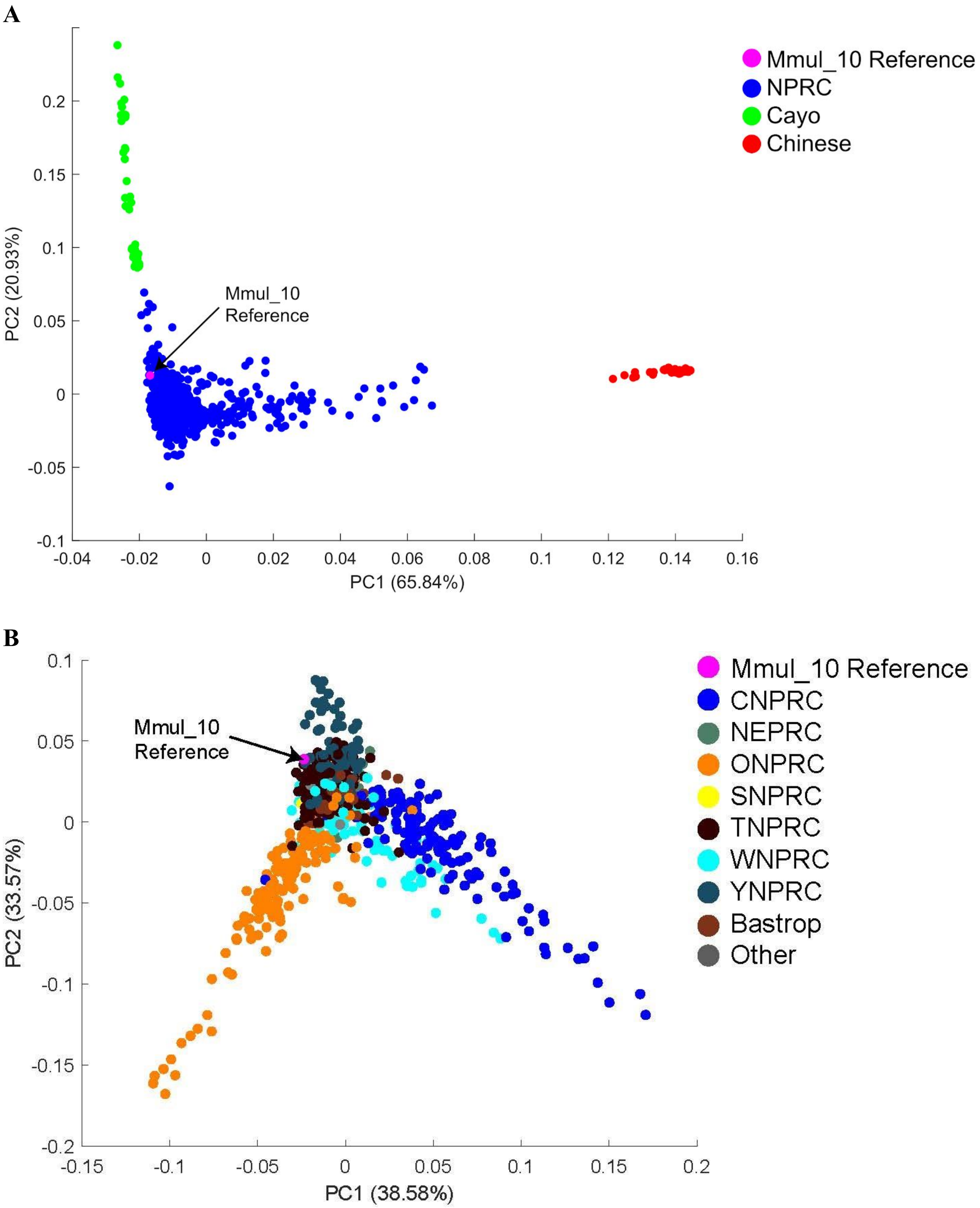


Fig. S28. Principal Components Analysis (PCA) of the Mmul_10 genome and other rhesus macaque samples. The PCA plot was generated by Primus depicting the genetic relationship of the new reference (Mmul_10) to other macaques. AG07107 DNA

corresponding to Mmul_10 was used to construct an Illumina short-read library. The library was sequenced to ~30X depth on an Illumina X10 instrument. All sequences were aligned to Mmul_10 and GATK was used to call SNVs. The gVCF file includes AG07107, Mmul_10 reference animal, and 853 rhesus macaques sequenced in this study. All individuals are known to be either Indian or Chinese origin, and they are associated with eight different research colonies. We selected out the SNVs with GATK score >10,000, which produced >14.2 million SNVs for PCA. In panels A and B we show genetic relatedness values for all macaques and only those in US research colonies, respectively.

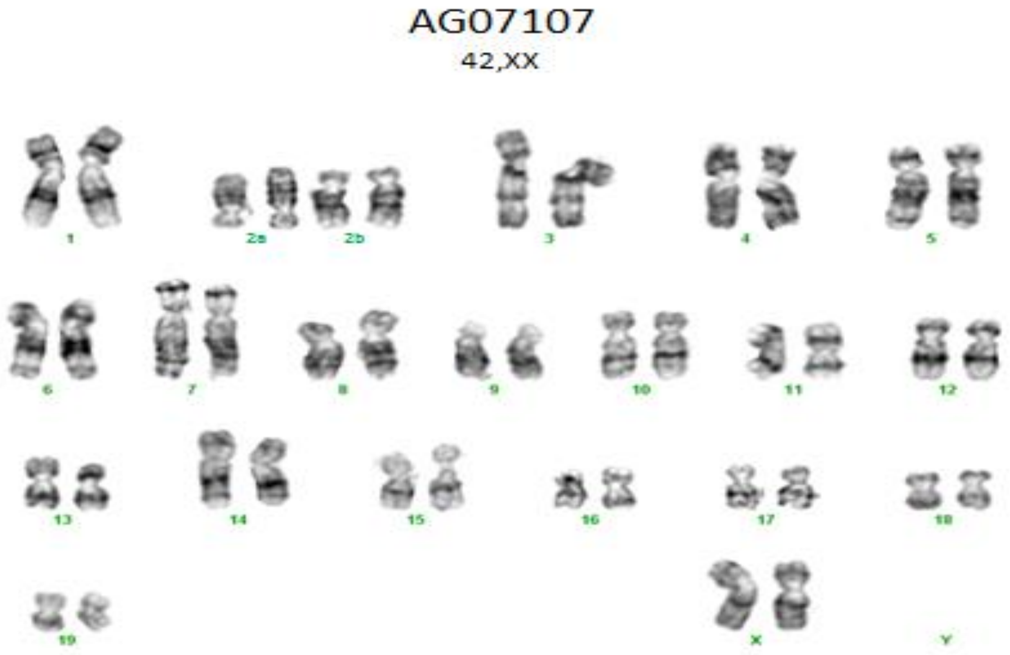


Fig. S29. Chromosome karyotype of Mmul_10 reference macaque. We evaluated a female rhesus macaque skin fibroblast cell line (AG07107; MMU) for karyotype structure. This cell line was not transformed, taken at two years of age, and was received at passage 4. AG07107 confluent cells were karyotyped at the Oregon Health and Science University karyotyping lab. Twenty metaphase cells were examined and ten metaphases were karyotyped. All metaphases appeared normal (female 42, XX).

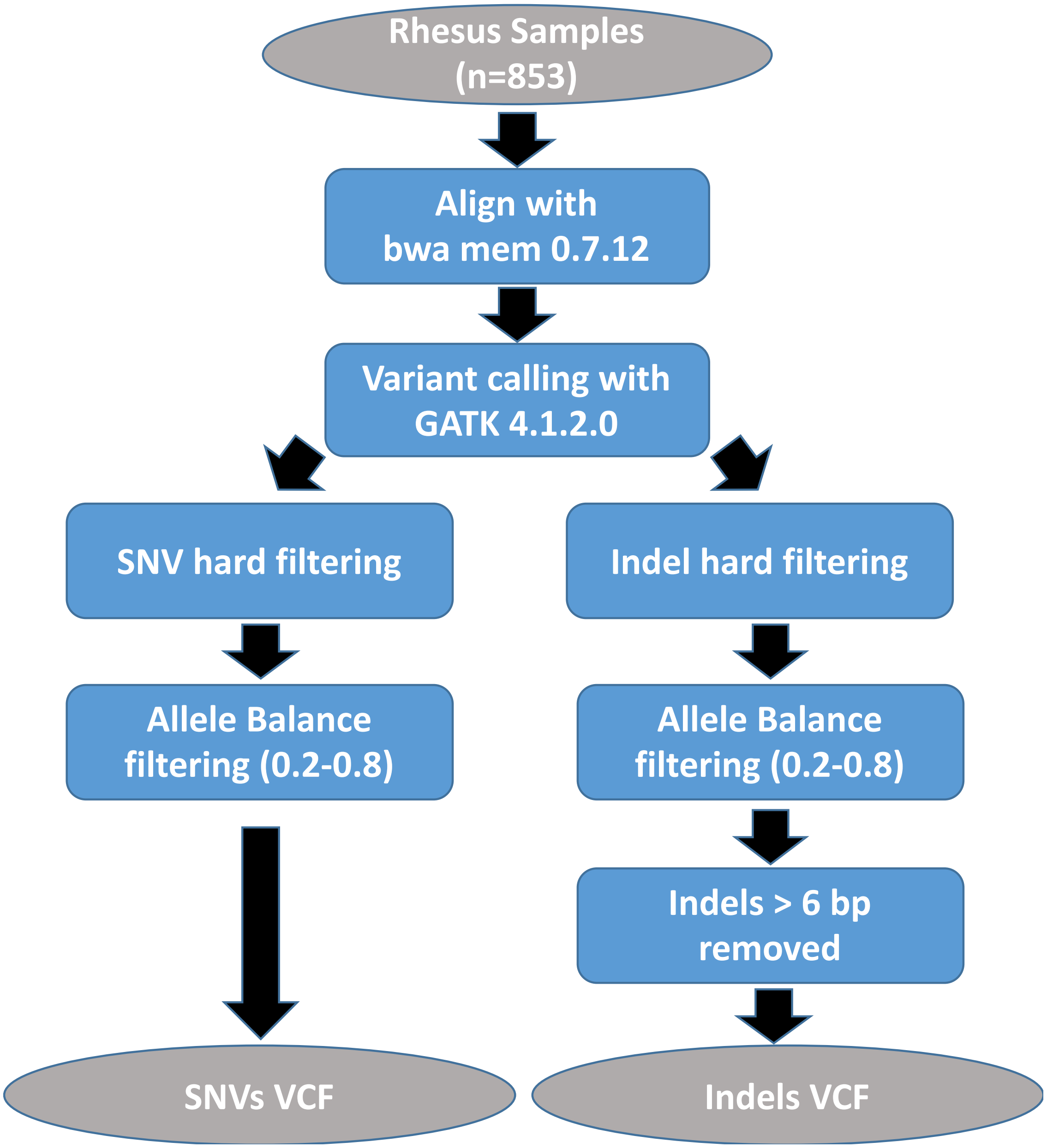


Fig. S30. Schematic of the sequence variant calling pipeline for rhesus macaque research populations. The variant calling pipeline was applied to samples from 853 rhesus macaques to generate gVCF files. The variant hard filter parameters applied are described at https://software.broadinstitute.org/gatk/documentation/article?id=11097. Allele balance was calculated as ref/(ref+alt) across samples with heterozygous genotypes and variants with values between 0.2-0.8 retained.

Captions for Tables S1 to S29

Table S1. Sequence sources for *de novo* assembly and population analyses of the rhesus macaque genome.

Table S2. Summary of MHC Class I and II annotated genes in Mmul_10.

Table S3. Summary of assembled KIR genes as annotated in Mmul_10.

Table S4. A comparison of possible misorientation events detected by Strand-seq in Mmul_10 compared to Mmul_8.0.1.

Table S5. Summary of Mmul_10 inversions that remain in the assembly.

Table S6. Summary of the BUSCO gene completeness analysis.

Table S7. Representative gene annotation metrics for sequenced OWM genomes and human.

Table S8. The total counts of annotated ncRNAs by type for assembled rhesus macaque genomes using the NCBI pipeline.

Table S9. The number of annotated ncRNAs missing by type for assembled rhesus macaque genomes using human ncRNA alignments and the CAT workflow.

Table S10. Ensembl gene annotations expanded in humans/collapsed in the rhesus macaque genome.

Table S11. Ensembl gene annotations expanded in rhesus macaque compared to human.

Table S12. Rhesus macaque Iso-Seq data used for characterization of gene annotation.

Table S13. Summary of split gene mappings using CAT processing of Ensembl gene annotation.

Table S14. Novel exon discovery in rhesus macaque using Iso-Seq mapping and CAT gene annotation.

Table S15. Summary of detected Mmul_10 assembly collapses using SDA.

Table S16. Summary of RepeatMasker characterization of the rhesus macaque assemblies Mmul_8.0.1 and Mmul_10.

Table S17. Total variant counts by type for the reference and population animals.

Table S18. Most severe indel VEP consequences based on merged Ensembl and RefSeq gene models. The consequences are ordered by severity as estimated by Ensembl.

Table S19. Summary of SNV characterization among genes associated with neurodevelopment.

Table S20. A summary of all LGD homozygous SNVs across 853 macaques.

Table S21. Total counts of SVs in the rheMacS Chinese and Mmul_10 Indian macaque genomes using independent callers.

Table S22. The number of annotated SVs among Chinese and Indian macaque populations that are predicted to alter gene structure.

Table S23. Summary of FISH validation of rhesus macaque segmental duplications.

Table S24. Rhesus macaque assembly comparisons for GEPHI Alu output.

Table S25. Rhesus macaque assembly comparisons for GEPHI LINE output.

Table S26. Summary of rhesus macaque populations sampled by research center.

Table S27. A Welch two-sample t-test comparing Cayo Santiago island to other Indian macaques.

Table S28. ncRNA classification by rhesus macaque assembly position with coverage and identity based on human ncRNA alignments.

Table S29. List of human BAC clones used to validate detected assembly inversions in Mmul_10 compared to Mmul_8.0.1.

Captions for Database S1

Database S1 contains various results associated with the specific analysis of repeat types and their sequence coordinate transitions between rhesus macaque assemblies.

Worksheet repeat content length provides the total estimated length of each repeat type for both Mmul_10 and Mmul_8.0.1 assemblies.

Worksheet Lineage-specific Alu counts provides numbers of total *Alu* elements in Mmul_8.0.1 and Mmul_10 as well as non-truncated, full-length, *Alu* family members.

Worksheet COSEG Alu subfamilies RM provides RepeatMasker output for ascertained Alu subfamilies from Mmul_8.0.1 and Mmul_10.

Worksheet COSEG LINE1 subfamilies RM provides RepeatMasker output for ascertained LINE1 subfamilies from Mmul_8.0.1 and Mmul_10.

Worksheet L1 10 to 8 liftover failed provides liftover coordinates of full-length L1s that are present in Mmul_10 but absent in Mmul_8.0.1.

Worksheet Alu 10 to 8 liftover failed provides liftover coordinates of full-length Alu repeants that are present in Mmul_10 but absent in Mmul_8.0.1.

Worksheet LTR 10 to 8 liftover failed provides liftover coordinates of full-length Alu repeants that are present in Mmul_10 but absent in Mmul_8.0.1.

Description of Repeat Sequences Database

Database Repeat Sequences are a collection of annotated repeat sequences classified in the rhesus macaque genome assemblies and is available for download online.

References and notes

1. J. A. Bailey, E. E. Eichler, Primate segmental duplications: Crucibles of evolution, diversity and disease. *Nat. Rev. Genet.* **7**, 552–564 (2006). doi:10.1038/nrg1895 Medline

2. R. A. Gibbs, J. Rogers, M. G. Katze, R. Bumgarner, G. M. Weinstock, E. R. Mardis, K. A. Remington, R. L. Strausberg, J. C. Venter, R. K. Wilson, M. A. Batzer, C. D. Bustamante, E. E. Eichler, M. W. Hahn, R. C. Hardison, K. D. Makova, W. Miller, A. Milosavljevic, R. E. Palermo, A. Siepel, J. M. Sikela, T. Attaway, S. Bell, K. E. Bernard, C. J. Buhay, M. N. Chandrabose, M. Dao, C. Davis, K. D. Delehaunty, Y. Ding, H. H. Dinh, S. Dugan-Rocha, L. A. Fulton, R. A. Gabisi, T. T. Garner, J. Godfrey, A. C. Hawes, J. Hernandez, S. Hines, M. Holder, J. Hume, S. N. Jhangiani, V. Jshi, Z. M. Khan, E. F. Kirkness, A. Cree, R. G. Fowler, S. Lee, L. R. Lewis, Z. Li, Y. S. Liu, S. M. Moore, D. Muzny, L. V. Nazareth, D. N. Ngo, G. O. Okwuonu, G. Pai, D. Parker, H. A. Paul, C. Pfannkoch, C. S. Pohl, Y. H. Rogers, S. J. Ruiz, A. Sabo, J. Santibanez, B. W. Schneider, S. M. Smith, E. Sodergren, A. F. Svatek, T. R. Utterback, S. Vattathil, W. Warren, C. S. White, A. T. Chinwalla, Y. Feng, A. L. Halpern, L. W. Hillier, X. Huang, P. Minx, J. O. Nelson, K. H. Pepin, X. Qin, G. G. Sutton, E. Venter, B. P. Walenz, J. W. Wallis, K. C. Worley, S. P. Yang, S. M. Jones, M. A. Marra, M. Rocchi, J. E. Schein, R. Baertsch, L. Clarke, M. Csürös, J. Glasscock, R. A. Harris, P. Havlak, A. R. Jackson, H. Jiang, Y. Liu, D. N. Messina, Y. Shen, H. X. Song, T. Wylie, L. Zhang, E. Birney, K. Han, M. K. Konkel, J. Lee, A. F. Smit, B. Ullmer, H. Wang, J. Xing, R. Burhans, Z. Cheng, J. E. Karro, J. Ma, B. Raney, X. She, M. J. Cox, J. P. Demuth, L. J. Dumas, S. G. Han, J. Hopkins, A. Karimpour-Fard, Y. H. Kim, J. R. Pollack, T. Vinar, C. Addo-Quaye, J. Degenhardt, A. Denby, M. J. Hubisz, A. Indap, C. Kosiol, B. T. Lahn, H. A. Lawson, A. Marklein, R. Nielsen, E. J. Vallender, A. G. Clark, B. Ferguson, R. D. Hernandez, K. Hirani, H. Kehrer-Sawatzki, J. Kolb, S. Patil, L. L. Pu, Y. Ren, D. G. Smith, D. A. Wheeler, I. Schenck, E. V. Ball, R. Chen, D. N. Cooper, B. Giardine, F. Hsu, W. J. Kent, A. Lesk, D. L. Nelson, W. E. O'brien, K. Prüfer, P. D. Stenson, J. C. Wallace, H. Ke, X. M. Liu, P. Wang, A. P. Xiang, F. Yang, G. P. Barber, D. Haussler, D. Karolchik, A. D. Kern, R. M. Kuhn, K. E. Smith, A. S. Zwieg; Rhesus Macaque Genome Sequencing and Analysis Consortium, Evolutionary and biomedical insights from the rhesus macaque genome. *Science* **316**, 222–234 (2007). doi:10.1126/science.1139247 Medline

3. C. Xue, M. Raveendran, R. A. Harris, G. L. Fawcett, X. Liu, S. White, M. Dahdouli, D. Rio Deiros, J. E. Below, W. Salerno, L. Cox, G. Fan, B. Ferguson, J. Horvath, Z. Johnson, S. Kanthaswamy, H. M. Kubisch, D. Liu, M. Platt, D. G. Smith, B. Sun, E. J. Vallender, F. Wang, R. W. Wiseman, R. Chen, D. M. Muzny, R. A. Gibbs, F. Yu, J. Rogers, The population genomics of rhesus macaques (*Macaca mulatta*) based on whole-genome sequences. *Genome Res.* **26**, 1651–1662 (2016). doi:10.1101/gr.204255.116 Medline

4. B. N. Bimber, R. Ramakrishnan, R. Cervera-Juanes, R. Madhira, S. M. Peterson, R. B. Norgren Jr., B. Ferguson, Whole genome sequencing predicts novel human disease models in rhesus macaques. *Genomics* **109**, 214–220 (2017). doi:10.1016/j.ygeno.2017.04.001 Medline

5. Z. N. Kronenberg, I. T. Fiddes, D. Gordon, S. Murali, S. Cantsilieris, O. S. Meyerson, J. G. Underwood, B. J. Nelson, M. J. P. Chaisson, M. L. Dougherty, K. M. Munson, A. R. Hastie, M. Diekhans, F. Hormozdiari, N. Lorusso, K. Hoekzema, R. Qiu, K. Clark, A.

Raja, A. E. Welch, M. Sorensen, C. Baker, R. S. Fulton, J. Armstrong, T. A. Graves-Lindsay, A. M. Denli, E. R. Hoppe, P. Hsieh, C. M. Hill, A. W. C. Pang, J. Lee, E. T. Lam, S. K. Dutcher, F. H. Gage, W. C. Warren, J. Shendure, D. Haussler, V. A. Schneider, H. Cao, M. Ventura, R. K. Wilson, B. Paten, A. Pollen, E. E. Eichler, High-resolution comparative analysis of great ape genomes. *Science* **360**, eaar634 (2018). doi:10.1126/science.aar6343 Medline

6. J. Rogers, R. A. Gibbs, Comparative primate genomics: Emerging patterns of genome content and dynamics. *Nat. Rev. Genet.* **15**, 347–359 (2014). doi:10.1038/nrg3707 Medline

7. K. K. A. Van Rompay, Tackling HIV and AIDS: Contributions by non-human primate models. *Lab Anim. (NY)* **46**, 259–270 (2017). doi:10.1038/laban.1279 Medline

8. H. Feldmann, F. Feldmann, A. Marzi, Ebola: Lessons on vaccine development. *Annu. Rev. Microbiol.* **72**, 423–446 (2018). doi:10.1146/annurev-micro-090817-062414 Medline

9. Y. Zhou, J. Sharma, Q. Ke, R. Landman, J. Yuan, H. Chen, D. S. Hayden, J. W. Fisher 3rd, M. Jiang, W. Menegas, T. Aida, T. Yan, Y. Zou, D. Xu, S. Parmar, J. B. Hyman, A. Fanucci-Kiss, O. Meisner, D. Wang, Y. Huang, Y. Li, Y. Bai, W. Ji, X. Lai, W. Li, L. Huang, Z. Lu, L. Wang, S. A. Anteraper, M. Sur, H. Zhou, A. P. Xiang, R. Desimone, G. Feng, S. Yang, Atypical behaviour and connectivity in SHANK3-mutant macaques. *Nature* **570**, 326–331 (2019). doi:10.1038/s41586-019-1278-0 Medline

10. R. Daza-Vamenta, G. Glusman, L. Rowen, B. Guthrie, D. E. Geraghty, Genetic divergence of the rhesus macaque major histocompatibility complex. *Genome Res.* **14**, 1501–1515 (2004). doi:10.1101/gr.2134504 Medline

11. B. K. Dray, M. Raveendran, R. A. Harris, F. Benavides, S. B. Gray, C. J. Perez, M. J. McArthur, L. E. Williams, W. B. Baze, H. Doddapaneni, D. M. Muzny, C. R. Abee, J. Rogers, Mismatch repair gene mutations lead to lynch syndrome colorectal cancer in rhesus macaques. *Genes Cancer* **9**, 142–152 (2018). doi:10.18632/genesandcancer.170 Medline

12. B. Y. Liao, J. Zhang, Null mutations in human and mouse orthologs frequently result in different phenotypes. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 6987–6992 (2008). doi:10.1073/pnas.0800387105 Medline

13. J. Seok, H. S. Warren, A. G. Cuenca, M. N. Mindrinos, H. V. Baker, W. Xu, D. R. Richards, G. P. McDonald-Smith, H. Gao, L. Hennessy, C. C. Finnerty, C. M. López, S. Honari, E. E. Moore, J. P. Minei, J. Cuschieri, P. E. Bankey, J. L. Johnson, J. Sperry, A. B. Nathens, T. R. Billiar, M. A. West, M. G. Jeschke, M. B. Klein, R. L. Gamelli, N. S. Gibran, B. H. Brownstein, C. Miller-Graziano, S. E. Calvano, P. H. Mason, J. P. Cobb, L. G. Rahme, S. F. Lowry, R. V. Maier, L. L. Moldawer, D. N. Herndon, R. W. Davis, W. Xiao, R. G. Tompkins; Inflammation and Host Response to Injury, Large Scale Collaborative Research Program, Genomic responses in mouse models poorly mimic human inflammatory diseases. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 3507–3512 (2013). doi:10.1073/pnas.1222878110 Medline

14. S. M. Peterson, T. J. McGill, T. Puthussery, J. Stoddard, L. Renner, A. D. Lewis, L. M. A. Colgin, J. Gayet, X. Wang, K. Prongay, C. Cullin, B. L. Dozier, B. Ferguson, M. Neuringer, Bardet-Biedl Syndrome in rhesus macaques: A nonhuman primate model of

retinitis pigmentosa. *Exp. Eye Res.* **189**, 107825 (2019). doi:10.1016/j.exer.2019.107825 Medline

15. A. Moshiri, R. Chen, S. Kim, R. A. Harris, Y. Li, M. Raveendran, S. Davis, Q. Liang, O. Pomerantz, J. Wang, L. Garzel, A. Cameron, G. Yiu, J. T. Stout, Y. Huang, C. J. Murphy, J. Roberts, K. N. Gopalakrishna, K. Boyd, N. O. Artemyev, J. Rogers, S. M. Thomasy, A nonhuman primate model of inherited retinal disease. *J. Clin. Invest.* **129**, 863–874 (2019). doi:10.1172/JCI123980 Medline

16. J. Rogers, M. Raveendran, G. L. Fawcett, A. S. Fox, S. E. Shelton, J. A. Oler, J. Cheverud, D. M. Muzny, R. A. Gibbs, R. J. Davidson, N. H. Kalin, CRHR1 genotypes, neural circuits and the diathesis for anxiety and depression. *Mol. Psychiatry* **18**, 700–707 (2013). doi:10.1038/mp.2012.152 Medline

17. D. H. Abbott, J. Rogers, D. A. Dumesic, J. E. Levine, Naturally occurring and experimentally induced rhesus macaque models for polycystic ovary syndrome: Translational gateways to clinical application. *Med. Sci. (Basel)* **7**, 107 (2019). doi:10.3390/medsci7120107 Medline

18. See the supplementary materials.

19. J. F. Hughes, H. Skaletsky, L. G. Brown, T. Pyntikova, T. Graves, R. S. Fulton, S. Dugan, Y. Ding, C. J. Buhay, C. Kremitzki, Q. Wang, H. Shen, M. Holder, D. Villasana, L. V. Nazareth, A. Cree, L. Courtney, J. Veizer, H. Kotkiewicz, T.-J. Cho, N. Koutseva, S. Rozen, D. M. Muzny, W. C. Warren, R. A. Gibbs, R. K. Wilson, D. C. Page, Strict evolutionary conservation followed rapid gene loss on human and rhesus Y chromosomes. *Nature* **483**, 82–86 (2012). doi:10.1038/nature10843 Medline

20. Y. He, X. Luo, B. Zhou, T. Hu, X. Meng, P. A. Audano, Z. N. Kronenberg, E. E. Eichler, J. Jin, Y. Guo, Y. Yang, X. Qi, B. Su, Long-read assembly of the Chinese rhesus macaque genome and identification of ape-specific structural variants. *Nat. Commun.* **10**, 4233 (2019). doi:10.1038/s41467-019-12174-w Medline

21. J. G. Sambrook, A. Bashirova, S. Palmer, S. Sims, J. Trowsdale, L. Abi-Rached, P. Parham, M. Carrington, S. Beck, Single haplotype analysis demonstrates rapid evolution of the killer immunoglobulin-like receptor (KIR) loci in primates. *Genome Res.* **15**, 25–35 (2005). doi:10.1101/gr.2381205 Medline

22. C. R. Catacchio, F. A. M. Maggiolini, P. D'Addabbo, M. Bitonto, O. Capozzi, M. Lepore Signorile, M. Miroballo, N. Archidiacono, E. E. Eichler, M. Ventura, F. Antonacci, Inversion variants in human and primate genomes. *Genome Res.* **28**, 910–920 (2018). doi:10.1101/gr.234831.118 Medline

23. M. Seppey, M. Manni, E. M. Zdobnov, BUSCO: Assessing genome assembly and annotation completeness. *Methods Mol. Biol.* **1962**, 227–245 (2019). doi:10.1007/978-1-4939-9173-0_14 Medline

24. 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. *Genome Res.* **28**, 1029–1038 (2018). doi:10.1101/gr.233460.117 Medline

25. R. D. Finn, P. Coggill, R. Y. Eberhardt, S. R. Eddy, J. Mistry, A. L. Mitchell, S. C. Potter, M. Punta, M. Qureshi, A. Sangrador-Vegas, G. A. Salazar, J. Tate, A. Bateman, The Pfam protein families database: Towards a more sustainable future. *Nucleic Acids Res.* **44** (D1), D279–D285 (2016). doi:10.1093/nar/gkv1344 Medline

26. M. R. Vollger, P. C. Dishuck, M. Sorensen, A. E. Welch, V. Dang, M. L. Dougherty, T. A. Graves-Lindsay, R. K. Wilson, M. J. P. Chaisson, E. E. Eichler, Long-read sequence and assembly of segmental duplications. *Nat. Methods* **16**, 88–94 (2019). doi:10.1038/s41592-018-0236-3 Medline

27. M. J. Chaisson, S. Mukherjee, S. Kannan, E. E. Eichler, Resolving multicopy duplications *de novo* using polyploid phasing. *Res Comput Mol Biol* **10229**, 117–133 (2017). doi:10.1007/978-3-319-56970-3_8 Medline

28. J. D. Fernandes, A. Zamudio-Hurtado, H. Clawson, W. J. Kent, D. Haussler, S. R. Salama, M. Haeussler, The UCSC repeat browser allows discovery and visualization of evolutionary conflict across repeat families. *Mob. DNA* **11**, 13 (2020). doi:10.1186/s13100-020-00208-w Medline

29. M. Imbeault, P. Y. Helleboid, D. Trono, KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks. *Nature* **543**, 550–554 (2017). doi:10.1038/nature21683 Medline

30. A. Bergström, S. A. McCarthy, R. Hui, M. A. Almarri, Q. Ayub, P. Danecek, Y. Chen, S. Felkel, P. Hallast, J. Kamm, H. Blanché, J.-F. Deleuze, H. Cann, S. Mallick, D. Reich, M. S. Sandhu, P. Skoglund, A. Scally, Y. Xue, R. Durbin, C. Tyler-Smith, Insights into human genetic variation and population history from 929 diverse genomes. *Science* **367**, eaay5012 (2020). doi:10.1126/science.aay5012 Medline

31. A. Widdig, L. Muniz, M. Minkner, Y. Barth, S. Bley, A. Ruiz-Lambides, O. Junge, R. Mundry, L. Kulik, Low incidence of inbreeding in a long-lived primate population isolated for 75 years. *Behav. Ecol. Sociobiol.* **71**, 18 (2017). doi:10.1007/s00265-016-2236-6 Medline

32. B. P. Coe, H. A. F. Stessman, A. Sulovari, M. R. Geisheker, T. E. Bakken, A. M. Lake, J. D. Dougherty, E. S. Lein, F. Hormozdiari, R. A. Bernier, E. E. Eichler, Neurodevelopmental disease genes implicated by de novo mutation and copy number variation morbidity. *Nat. Genet.* **51**, 106–116 (2019). doi:10.1038/s41588-018-0288-4 Medline

33. F. K. Satterstrom, J. A. Kosmicki, J. Wang, M. S. Breen, S. De Rubeis, J.-Y. An, M. Peng, R. Collins, J. Grove, L. Klei, C. Stevens, J. Reichert, M. S. Mulhern, M. Artomov, S. Gerges, B. Sheppard, X. Xu, A. Bhaduri, U. Norman, H. Brand, G. Schwartz, R. Nguyen, E. E. Guerrero, C. Dias, C. Betancur, E. H. Cook, L. Gallagher, M. Gill, J. S. Sutcliffe, A. Thurm, M. E. Zwick, A. D. Børglum, M. W. State, A. E. Cicek, M. E. Talkowski, D. J. Cutler, B. Devlin, S. J. Sanders, K. Roeder, M. J. Daly, J. D. Buxbaum, Large-scale exome sequencing study implicates both developmental and functional changes in the neurobiology of autism. *Cell* **180**, 568–584.e23 (2020). doi:10.1016/j.cell.2019.12.036 Medline

34. D. M. Church, L. Goodstadt, L. W. Hillier, M. C. Zody, S. Goldstein, X. She, C. J. Bult, R. Agarwala, J. L. Cherry, M. DiCuccio, W. Hlavina, Y. Kapustin, P. Meric, D. Maglott, Z. Birtle, A. C. Marques, T. Graves, S. Zhou, B. Teague, K. Potamousis, C. Churas, M.

Place, J. Herschleb, R. Runnheim, D. Forrest, J. Amos-Landgraf, D. C. Schwartz, Z. Cheng, K. Lindblad-Toh, E. E. Eichler, C. P. Ponting; Mouse Genome Sequencing Consortium, Lineage-specific biology revealed by a finished genome assembly of the mouse. *PLOS Biol.* **7**, e1000112 (2009). doi:10.1371/journal.pbio.1000112 Medline

35. International Human Genome Sequencing Consortium, Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945 (2004). doi:10.1038/nature03001 Medline

36. J. A. Bailey, A. M. Yavor, H. F. Massa, B. J. Trask, E. E. Eichler, Segmental duplications: Organization and impact within the current human genome project assembly. *Genome Res.* **11**, 1005–1017 (2001). doi:10.1101/gr.GR-1871R Medline

37. J. A. Bailey, Z. Gu, R. A. Clark, K. Reinert, R. V. Samonte, S. Schwartz, M. D. Adams, E. W. Myers, P. W. Li, E. E. Eichler, Recent segmental duplications in the human genome. *Science* **297**, 1003–1007 (2002). doi:10.1126/science.1072047 Medline

38. J. D. Fernandes, M. Haeussler, J. Armstrong, K. Tigyi, J. Gu, N. Filippi, J. Pierce, T. Thisner, P. Angulo, S. Katzman, B. Paten, D. Haussler, S. R. Salama, KRAB zinc finger proteins coordinate across evolutionary time scales to battle retroelements. bioRxiv 429563 [Preprint]. 27 September 2018. https://doi.org/10.1101/429563.

39. B. L. Browning, S. R. Browning, A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am. J. Hum. Genet.* **84**, 210–223 (2009). doi:10.1016/j.ajhg.2009.01.005 Medline

40. J. Chin, "FALCON: experimental PacBiol. diploid assembler" (2014); https://github.com/PacificBiosciences/falcon/tree/v0.1.3.

41. 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

42. S. Deschamps, Y. Zhang, V. Llaca, L. Ye, A. Sanyal, M. King, G. May, H. Lin, A chromosome-scale assembly of the sorghum genome using nanopore sequencing and optical mapping. *Nat. Commun.* **9**, 4844 (2018). doi:10.1038/s41467-018-07271-1 Medline

43. N. H. Putnam, B. L. O'Connell, J. C. Stites, B. J. Rice, M. Blanchette, R. Calef, C. J. Troll, A. Fields, P. D. Hartley, C. W. Sugnet, D. Haussler, D. S. Rokhsar, R. E. Green, Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *Genome Res.* **26**, 342–350 (2016). doi:10.1101/gr.193474.115 Medline

44. N. H. Lazar, K. A. Nevonen, B. O'Connell, C. McCann, R. J. O'Neill, R. E. Green, T. J. Meyer, M. Okhovat, L. Carbone, Epigenetic maintenance of topological domains in the highly rearranged gibbon genome. *Genome Res.* **28**, 983–997 (2018). doi:10.1101/gr.233874.117 Medline

45. M. Zaharia, W. J. Bolosky, K. Curtis, A. Fox, D. Patterson, S. Shenker, I. Stoica, R. M. Karp, T. Sittler, Faster and more accurate sequence alignment with SNAP. arXiv:1111.5572v1 [cs.DS] (23 November 2011).

46. G. Marçais, A. L. Delcher, A. M. Phillippy, R. Coston, S. L. Salzberg, A. Zimin, MUMmer4: A fast and versatile genome alignment system. *PLOS Comput. Biol.* **14**, e1005944 (2018). doi:10.1371/journal.pcbi.1005944 Medline

47. E. Falconer, M. Hills, U. Naumann, S. S. S. Poon, E. A. Chavez, A. D. Sanders, Y. Zhao, M. Hirst, P. M. Lansdorp, DNA template strand sequencing of single-cells maps genomic rearrangements at high resolution. *Nat. Methods* **9**, 1107–1112 (2012). doi:10.1038/nmeth.2206 Medline

48. 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

49. M. L. Dougherty, J. G. Underwood, B. J. Nelson, E. Tseng, K. M. Munson, O. Penn, T. J. Nowakowski, A. A. Pollen, E. E. Eichler, Transcriptional fates of human-specific segmental duplications in brain. *Genome Res.* **28**, 1566–1576 (2018). doi:10.1101/gr.237610.118 Medline

50. J. Armstrong, I. T. Fiddes, M. Diekhans, B. Paten, Whole-Genome Alignment and Comparative Annotation. *Annu. Rev. Anim. Biosci.* **7**, 41–64 (2019). doi:10.1146/annurev-animal-020518-115005 Medline

51 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

52. M. Stanke, R. Steinkamp, S. Waack, B. Morgenstern, AUGUSTUS: A web server for gene finding in eukaryotes. *Nucleic Acids Res.* **32**, W309–W312 (2004). doi:10.1093/nar/gkh379 Medline

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

54. K. D. Pruitt, T. Tatusova, G. R. Brown, D. R. Maglott, NCBI Reference Sequences (RefSeq): Current status, new features and genome annotation policy. *Nucleic Acids Res.* **40** (D1), D130–D135 (2012). doi:10.1093/nar/gkr1079 Medline

55. D. R. Zerbino, P. Achuthan, W. Akanni, M. R. Amode, D. Barrell, J. Bhai, K. Billis, C. Cummins, A. Gall, C. G. Girón, L. Gil, L. Gordon, L. Haggerty, E. Haskell, T. Hourlier, O. G. Izuogu, S. H. Janacek, T. Juettemann, J. K. To, M. R. Laird, I. Lavidas, Z. Liu, J. E. Loveland, T. Maurel, W. McLaren, B. Moore, J. Mudge, D. N. Murphy, V. Newman, M. Nuhn, D. Ogeh, C. K. Ong, A. Parker, M. Patricio, H. S. Riat, H. Schuilenburg, D. Sheppard, H. Sparrow, K. Taylor, A. Thormann, A. Vullo, B. Walts, A. Zadissa, A. Frankish, S. E. Hunt, M. Kostadima, N. Langridge, F. J. Martin, M. Muffato, E. Perry, M. Ruffier, D. M. Staines, S. J. Trevanion, B. L. Aken, F. Cunningham, A. Yates, P. Flicek, Ensembl 2018. *Nucleic Acids Res.* **46** (D1), D754–D761 (2018). doi:10.1093/nar/gkx1098 Medline

56. 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

57. H. Li, Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018). doi:10.1093/bioinformatics/bty191 Medline

58. M. Pertea, A. Shumate, G. Pertea, A. Varabyou, F. P. Breitwieser, Y.-C. Chang, A. K. Madugundu, A. Pandey, S. L. Salzberg, CHESS: A new human gene catalog curated from thousands of large-scale RNA sequencing experiments reveals extensive transcriptional noise. *Genome Biol.* **19**, 208 (2018). doi:10.1186/s13059-018-1590-2 Medline

59. T. J. Wheeler, J. Clements, S. R. Eddy, R. Hubley, T. A. Jones, J. Jurka, A. F. A. Smit, R. D. Finn, Dfam: A database of repetitive DNA based on profile hidden Markov models. *Nucleic Acids Res.* **41** (D1), D70–D82 (2013). doi:10.1093/nar/gks1265 Medline

60. W. J. Kent, BLAT—The BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002). doi:10.1101/gr.229202 Medline

61. M. A. Batzer, P. L. Deininger, U. Hellmann-Blumberg, J. Jurka, D. Labuda, C. M. Rubin, C. W. Schmid, E. Zietkiewicz, E. Zuckerkandl, Standardized nomenclature for Alu repeats. *J. Mol. Evol.* **42**, 3–6 (1996). doi:10.1007/BF00163204 Medline

62. J. M. Bastian M. H., paper presented at the AAAI Conference on Weblogs and Social Media, 2009.

63. J. Jurka, V. V. Kapitonov, A. Pavlicek, P. Klonowski, O. Kohany, J. Walichiewicz, Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* **110**, 462–467 (2005). doi:10.1159/000084979 Medline

64. K. Han, M. K. Konkel, J. Xing, H. Wang, J. Lee, T. J. Meyer, C. T. Huang, E. Sandifer, K. Hebert, E. W. Barnes, R. Hubley, W. Miller, A. F. A. Smit, B. Ullmer, M. A. Batzer, Mobile DNA in Old World monkeys: A glimpse through the rhesus macaque genome. *Science* **316**, 238–240 (2007). doi:10.1126/science.1139462 Medline

65. J. D. Fernandes, W. A. Zamudio-Hurtado, J. Kent, D. Haussler, S. R. Salama, M. Haeussler, The UCSC Repeat Browser allows discovery and visualization of evolutionary conflict across repeat families. bioRxiv 429613 [Preprint]. 27 November 2019. https://doi.org/10.1101/429613.

66. H. Li, R. Durbin, Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595 (2010). doi:10.1093/bioinformatics/btp698 Medline

67. W. McLaren, L. Gil, S. E. Hunt, H. S. Riat, G. R. S. Ritchie, A. Thormann, P. Flicek, F. Cunningham, The Ensembl variant effect predictor. *Genome Biol.* **17**, 122 (2016). doi:10.1186/s13059-016-0974-4 Medline

68. C. Zhang, S. S. Dong, J. Y. Xu, W. M. He, T. L. Yang, PopLDdecay: A fast and effective tool for linkage disequilibrium decay analysis based on variant call format files. *Bioinformatics* **35**, 1786–1788 (2019). doi:10.1093/bioinformatics/bty875 Medline

69. 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

70. Pacific Biosciences, "pbsv - PacBio structural variant (SV) calling and analysis tools" (Pacific Biosciences, 2018); https://github.com/PacificBiosciences/pbsv.

71. F. J. Sedlazeck, P. Rescheneder, M. Smolka, H. Fang, M. Nattestad, A. von Haeseler, M. C. Schatz, Accurate detection of complex structural variations using single-molecule sequencing. *Nat. Methods* **15**, 461–468 (2018). doi:10.1038/s41592-018-0001-7 Medline

72. J. A. Chapman, I. Ho, S. Sunkara, S. Luo, G. P. Schroth, D. S. Rokhsar, Meraculous: De novo genome assembly with short paired-end reads. *PLOS ONE* **6**, e23501 (2011). doi:10.1371/journal.pone.0023501 Medline

73. R. C. Edgar, MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004). doi:10.1093/nar/gkh340 Medline

74. 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

75. A. D. Sanders, E. Falconer, M. Hills, D. C. J. Spierings, P. M. Lansdorp, Single-cell template strand sequencing by Strand-seq enables the characterization of individual homologs. *Nat. Protoc.* **12**, 1151–1176 (2017). doi:10.1038/nprot.2017.029 Medline

76. J. A. Karl, P. S. Bohn, R. W. Wiseman, F. A. Nimityongskul, S. M. Lank, G. J. Starrett, D. H. O'Connor, Major histocompatibility complex class I haplotype diversity in Chinese rhesus macaques. *G3 (Bethesda)* **3**, 1195–1201 (2013). doi:10.1534/g3.113.006254 Medline

77. C. G. Shortreed, R. W. Wiseman, J. A. Karl, H. E. Bussan, D. A. Baker, T. M. Prall, A. K. Haj, G. K. Moreno, M. C. T. Penedo, D. H. O'Connor, Characterization of 100 extended major histocompatibility complex haplotypes in Indonesian cynomolgus macaques. *Immunogenetics* **72**, 225–239 (2020). doi:10.1007/s00251-020-01159-5 Medline

78. J. R. Caskey, R. W. Wiseman, J. A. Karl, D. A. Baker, T. Lee, R. J. Maddox, M. Raveendran, R. A. Harris, J. Hu, D. M. Muzny, J. Rogers, D. H. O'Connor, MHC genotyping from rhesus macaque exome sequences. *Immunogenetics* **71**, 531–544 (2019). doi:10.1007/s00251-019-01125-w Medline

79. J. Bruijnesteijn, N. G. de Groot, N. Otting, G. Maccari, L. A. Guethlein, J. Robinson, S. G. E. Marsh, L. Walter, D. H. O'Connor, J. A. Hammond, P. Parham, R. E. Bontrop, Nomenclature report for killer-cell immunoglobulin-like receptors (KIR) in macaque

species: New genes/alleles, renaming recombinant entities and IPD-NHKIR updates. *Immunogenetics* **72**, 37–47 (2020). doi:10.1007/s00251-019-01135-8 Medline

80. R. C. Iskow, O. Gokcumen, A. Abyzov, J. Malukiewicz, Q. Zhu, A. T. Sukumar, A. A. Pai, R. E. Mills, L. Habegger, D. A. Cusanovich, M. A. Rubel, G. H. Perry, M. Gerstein, A. C. Stone, Y. Gilad, C. Lee, Regulatory element copy number differences shape primate expression profiles. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 12656–12661 (2012). doi:10.1073/pnas.1205199109 Medline