Science Advances
AAAS

Supplementary Materials for

Adaptive functions of structural variants in human brain development

Wanqiu Ding *et al.*

Corresponding author: Chuan-Yun Li, chuanyunli@pku.edu.cn; Li Zhang, zhangli@cibr.ac.cn

Sci. Adv. **10**, eadl4600 (2024)
DOI: 10.1126/sciadv.adl4600

The PDF file includes:

Figs. S1 to S21
Tables S1, S3, S6, S8
Legends for tables S2, S4, S5, S7, S9, S10 to S12

Other Supplementary Material for this manuscript includes the following:

Tables S2, S4, S5, S7, S9, S10 to S12

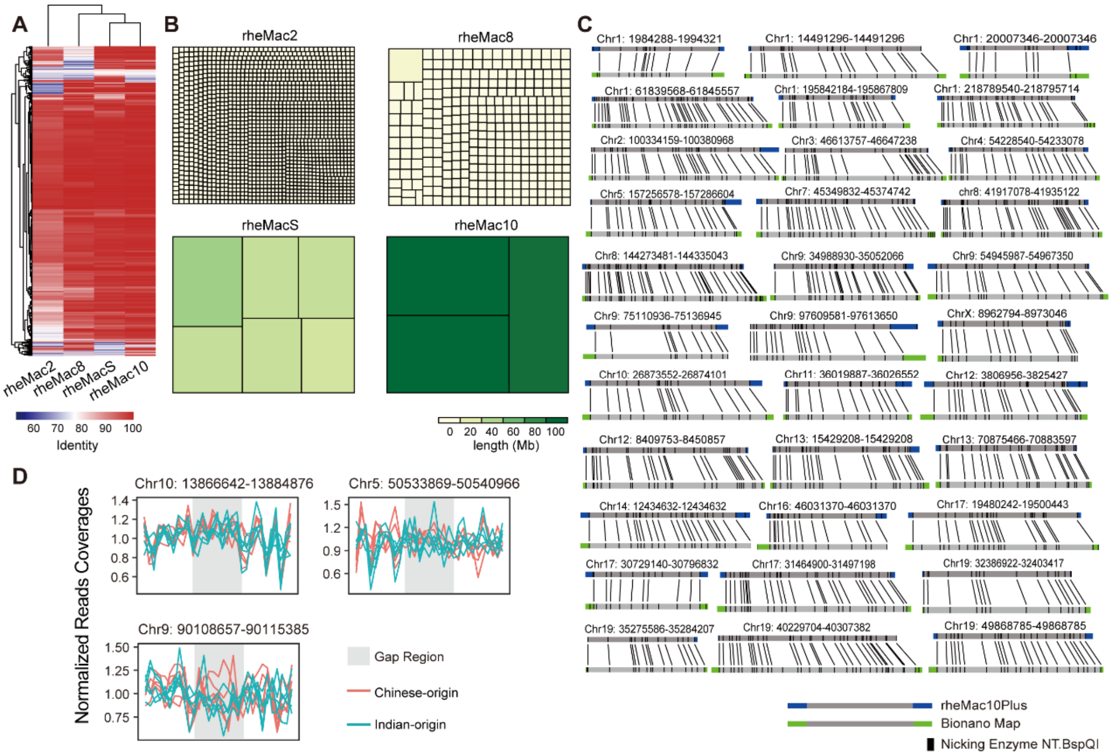

Fig. S1. Gap closure of the macaque reference genome. (**A**) Heatmap showing the base identity of four independent assemblies of the macaque genome (rheMac2, rheMac8, rheMacS and rheMac10), as evaluated by BAC sequences. (**B**) Treemaps showing the different distributions of contig sizes in the four assemblies of the macaque genome. The rectangles in each genome assembly represent the largest contigs accounting for 200 Mb of the assembly. (**C**) Visualization of the alignments between Bionano optical maps and the improved macaque reference genome (rheMac10Plus). For the gap regions and their flanking regions, the enzyme cleavage sites identified from the optical map or defined based on the reference genome sequences are shown with black bars, and the continuous and correctly spaced alignments between them are connected by black lines. (**D**) Reads coverages of the filled gaps and their flanking regions, according to the genome sequencing of ten macaques (orange: Chinese-origin macaques; green: Indian-origin macaques. Gap regions are indicated by gray shaded rectangles).

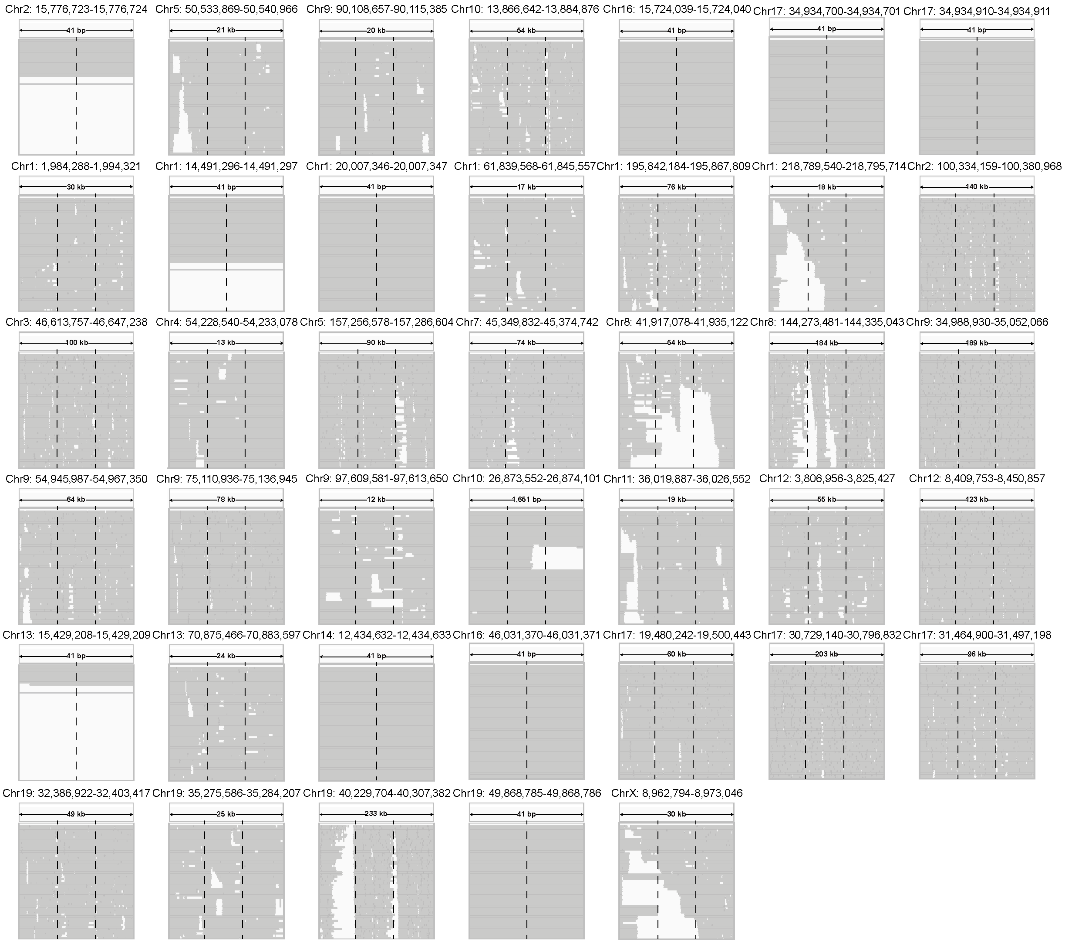

Fig. S2. Verification of 40 gap closures by long HiFi reads of PacBio sequencing. For each gap region (indicated by the black dashed lines) and its length-matched, upstream/downstream flanking regions, the long reads covering these regions were aligned and shown.

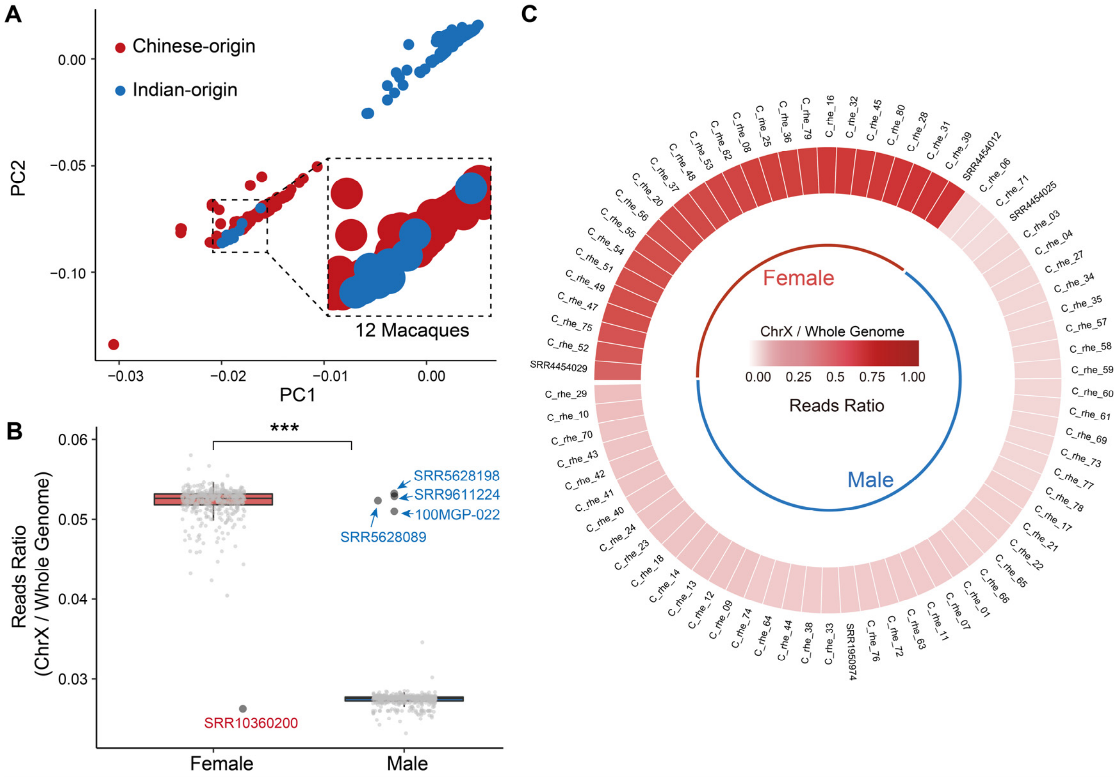

Fig. S3. Revision of the source and sex information of macaques. (**A**) PCA plot showing the genetic distances and relationships of the 1,026 macaques, which were clustered into two major groups (red: Chinese-origin macaques; blue: Indian-origin macaques). The 12 macaques with confusing resource information in the original report are highlighted in a zoomed-in view. (**B**) The distribution of reads densities on the X chromosomes of 1,026 macaques, normalized with that in the whole genome for the corresponding macaque individual. The IDs of macaques with confusing sex information are indicated. The Wilcoxon rank-sum test was performed. ***P value < 0.001. (**C**) Heatmap showing the reads densities on the X chromosomes of 74 wild-caught Chinese-origin macaques.

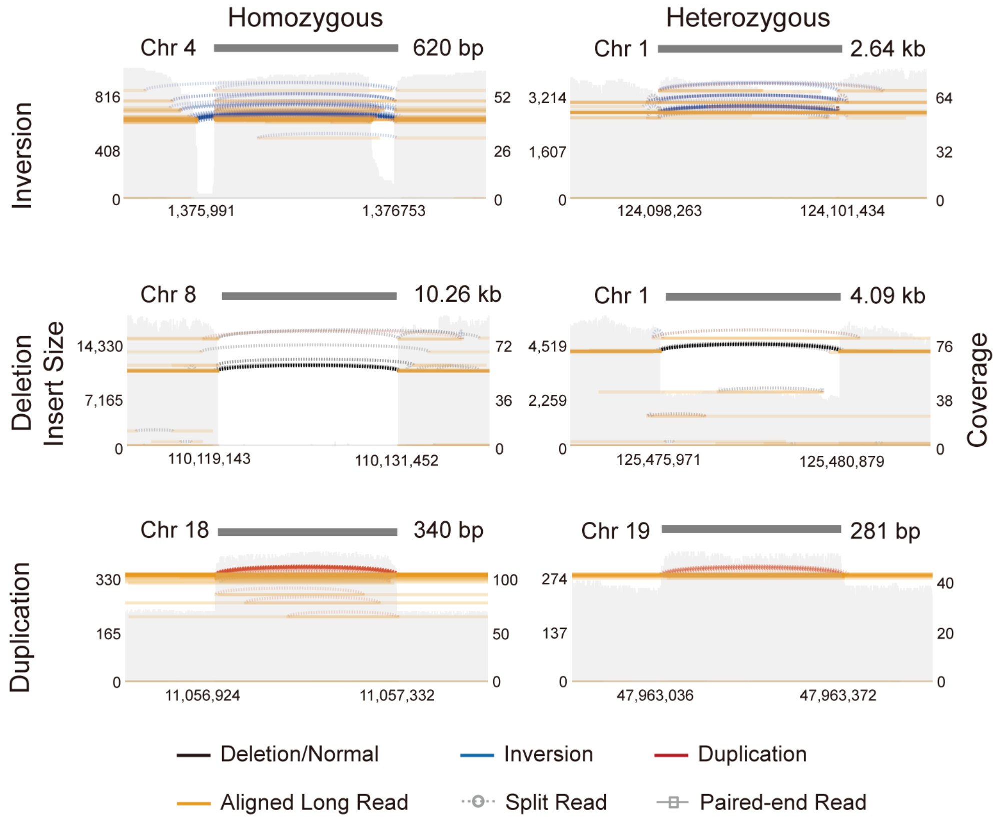

Fig. S4. Samplots of representative SVs verified by long-reads sequencing. The aligned long-reads are color-coded according to the type of alignments (concordant/discordant insert size, pair order, split alignment, or long read). The coverage of each region is shown with a gray-filled background.

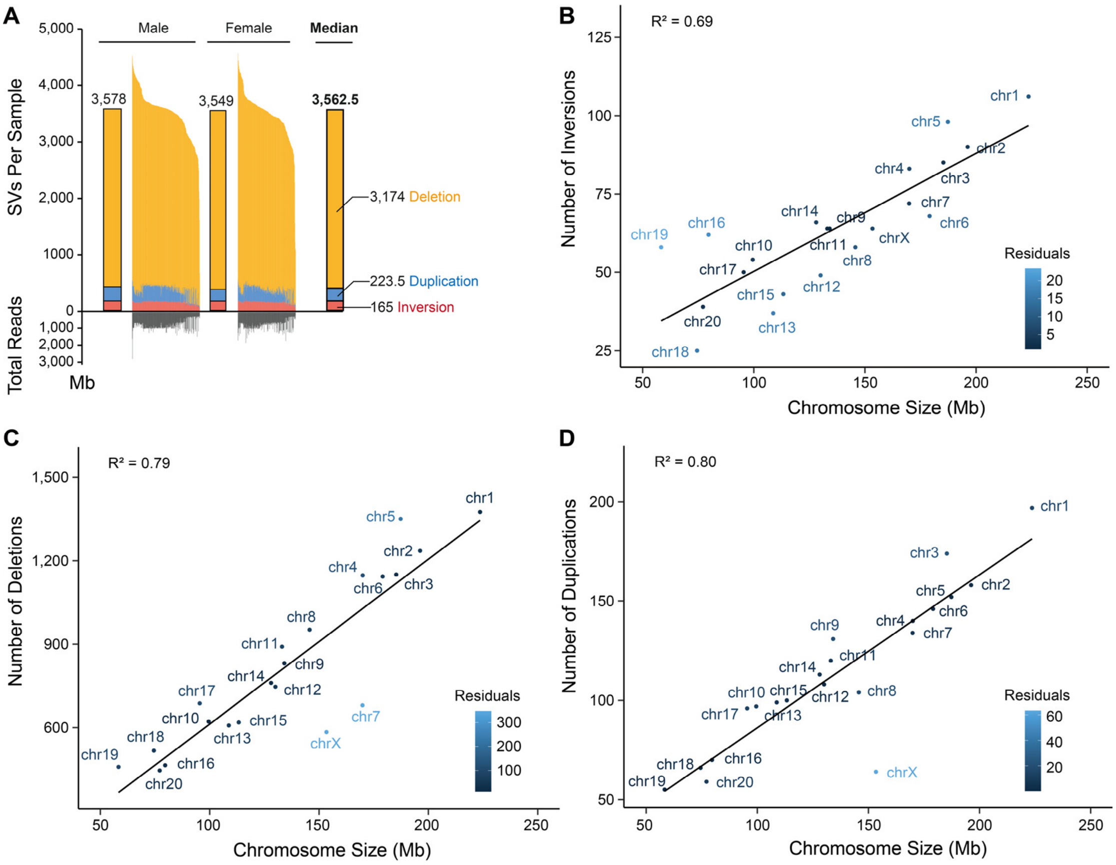

Fig. S5. Characteristics of the SV map in the macaque population. (**A**) The distribution of the count of SVs *per* macaque genome for three types of SVs in macaques of different sexes. The median number of SVs is shown for each group. The total number of deep sequencing reads for each macaque is also shown. (**B-D**) Scatter plots showing the correlation between chromosome size (in Mb) and the numbers of inversions (**B**), deletions (**C**) and duplications (**D**) located on each chromosome.

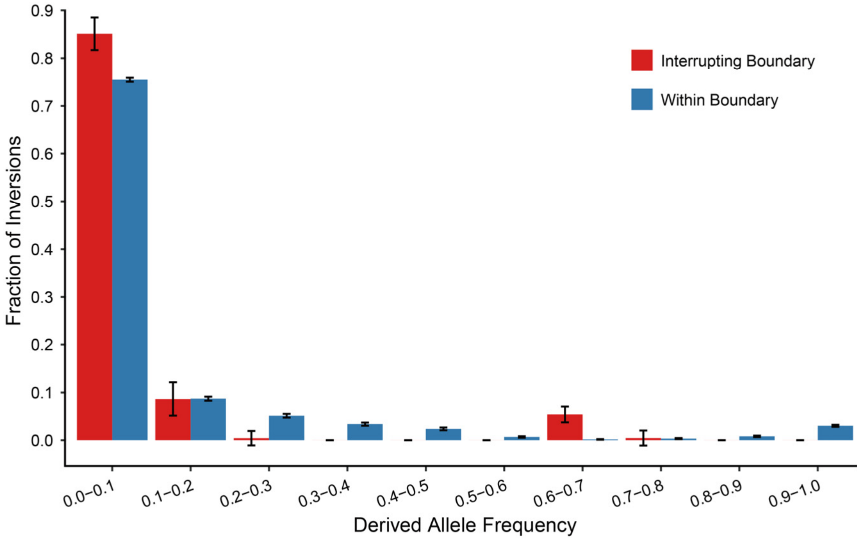

Fig. S6. Site frequency spectra of inversions with different locations. Site frequency spectra of the derived alleles for inversions disrupting TAD boundaries (red) and inversions within TADs (blue), indicated by Hi-C data of fetal CP.

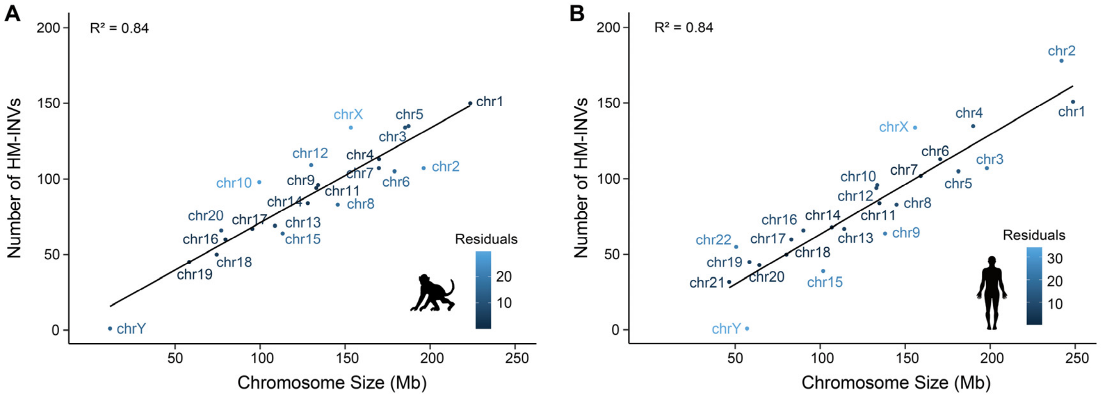

Fig. S7. Correlations between the length of the chromosome and the number of species-specific inversions between humans and macaques. (**A-B**) Scatter plots showing the correlation between chromosome size (in Mb) and the number of HM-INVs located on each chromosome in rhesus macaque (**A**) and human (**B**).

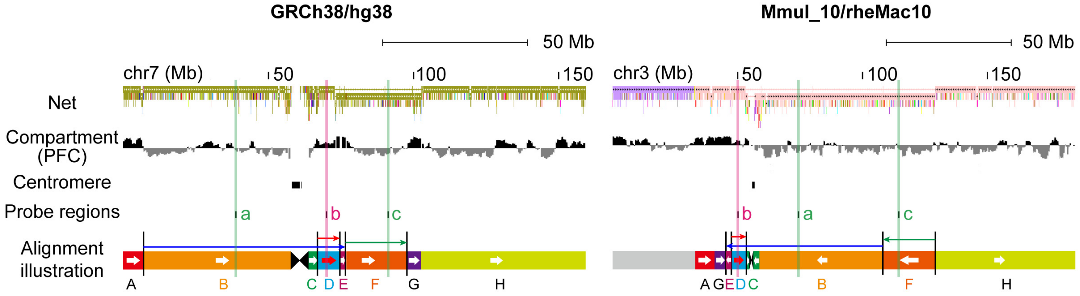

Fig. S8. Probe design for FISH validation. Schematic diagram showing the probe design for FISH assays performed to clarify the structures of one complex HM-INV in humans (left) and macaques (right). The Net track shows the alignment between the human and macaque genomes, followed by the information for A/B compartments and the positions of the centromeres. The locations of the three probes designed (a, b and c) are shown in the two panels, with the same colors used in **Fig. 4D**. The pairwise alignment between the two genome assemblies is illustrated in color-coded blocks from A to H, with the inversions indicated by the arrows.

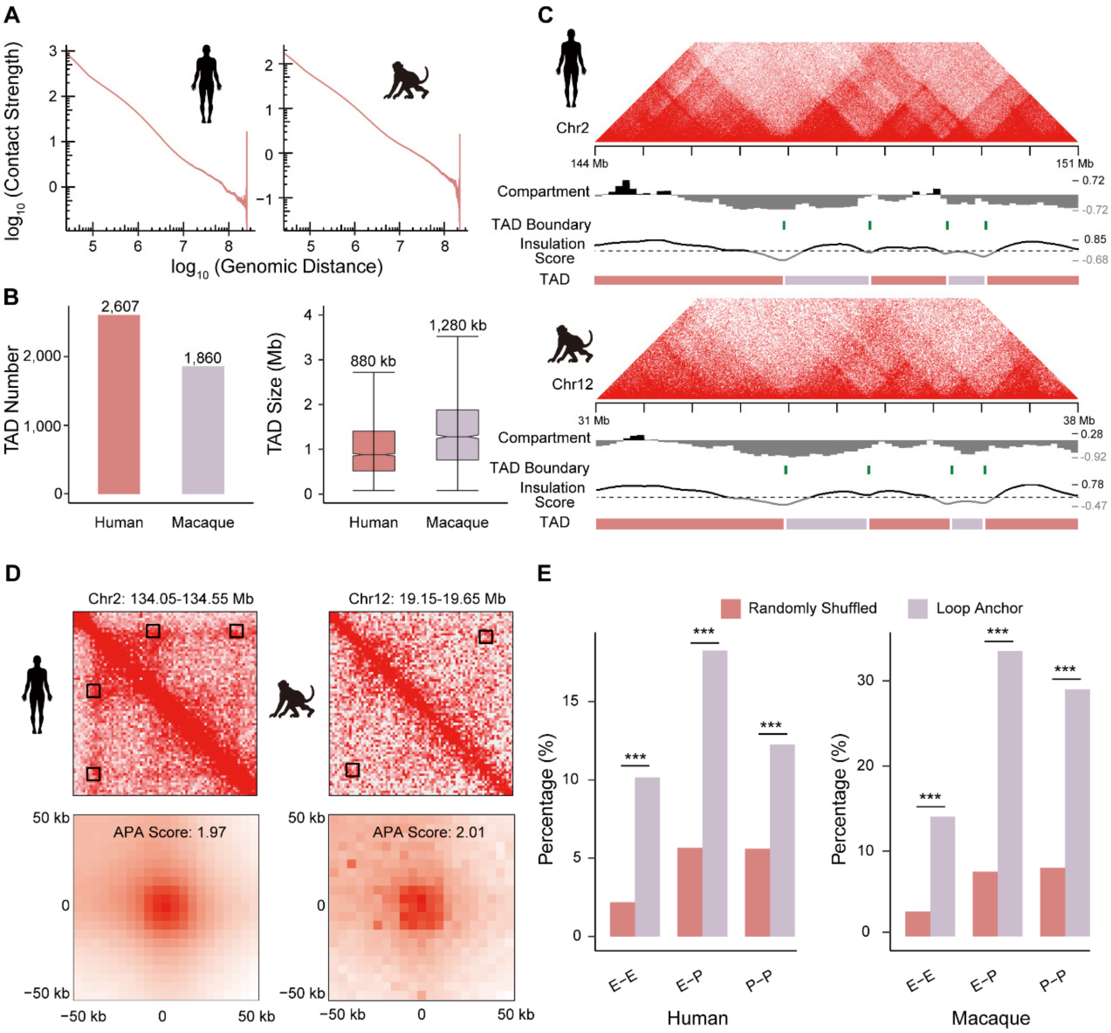

Fig. S9. High-resolution Hi-C contact maps in human and macaque. (**A**) Contact strength as a function of the log10-transformed genomic distance in human and macaque genomes. (**B**) The numbers (left) and sizes (right) of the TADs identified in human and macaque. (**C**) Representative Hi-C contact maps in human (top) and macaque (bottom). For each region, the information of the A/B compartments is shown. The TADs are shown in red and purple blocks, and the positions of the TAD boundaries are highlighted with green bars. Genomic regions (in 40 kb windows) with positive insulation scores are represented by black lines, whereas the regions with negative scores are represented by gray lines. (**D**) Representative loop views (top) for one position in human (left) and another position in rhesus macaque (right), with loops shown in dashed boxes. Aggregate peak analysis plots (APA, bottom) showing the combined signals for all of the loops, as detected in human (left) and rhesus macaque (right). (**E**) The percentage of enhancer-enhancer (E-E), enhancer-promoter (E-P) and promoter-promoter (P-P) interactions for loop anchors and randomly shuffled regions in human and rhesus macaque. Fisher's exact test was performed. ***P value < 0.001.

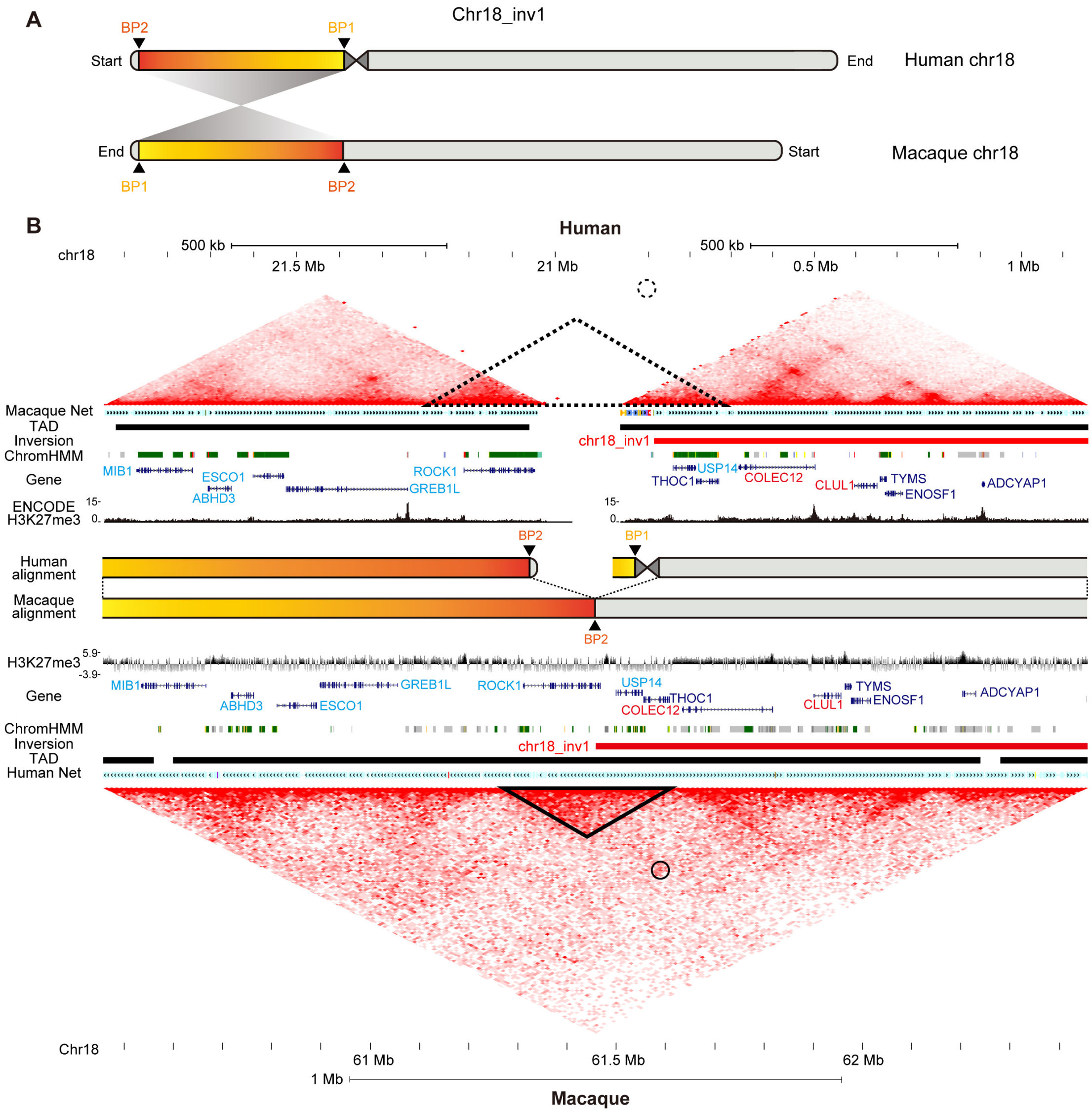

Fig. S10. Hi-C contact maps in human and macaque for a human-specific inversion. (**A**) Schematic illustration of a human-specific inversion on chromosome 18. (**B**) The contact map for the human-specific inversion in human (top) and the orthologous region in macaque (bottom). The interactions predicted to be lost as a consequence of the inversion are indicated by black triangles and circles. The human-macaque net tracks are shown to indicate the inverted alignment. The TADs are represented by black blocks. The chromatin states are indicated in the ChromHMM and H3K27me3 tracks. In comparison with the expression in macaque brains, the genes downregulated in humans were shown in blue, while the upregulated genes were shown in red. A schematic diagram of the detailed alignment between human and macaque genomes is also shown.

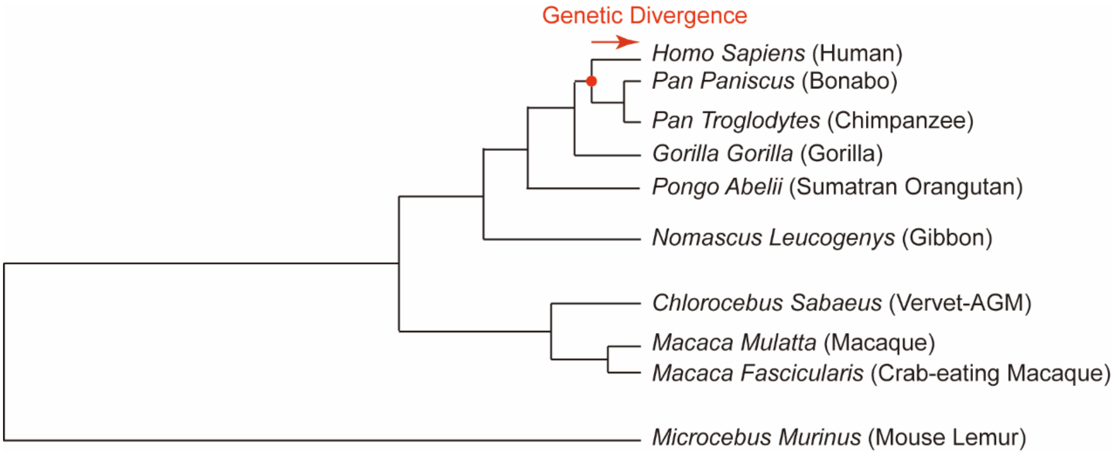
Genetic Divergence
Homo Sapiens (Human)
Pan Paniscus (Bonabo)
Pan Troglodytes (Chimpanzee)
Gorilla Gorilla (Gorilla)
Pongo Abelii (Sumatran Orangutan)
Nomascus Leucogenys (Gibbon)
Chlorocebus Sabaeus (Vervet-AGM)
Macaca Mulatta (Macaque)
Macaca Fascicularis (Crab-eating Macaque)
Microcebus Murinus (Mouse Lemur)

Fig. S11. Tracing the ancestral state with the EPO pipeline. Schematic diagram of the evolutionary tree of ten primate species in the EPO pipeline to trace the ancestral state of variants. The node of the common ancestor of humans and chimpanzees is highlighted with a red dot, and the branch used to calculate genetic divergence in **Fig. 5** is indicated with a red arrow.

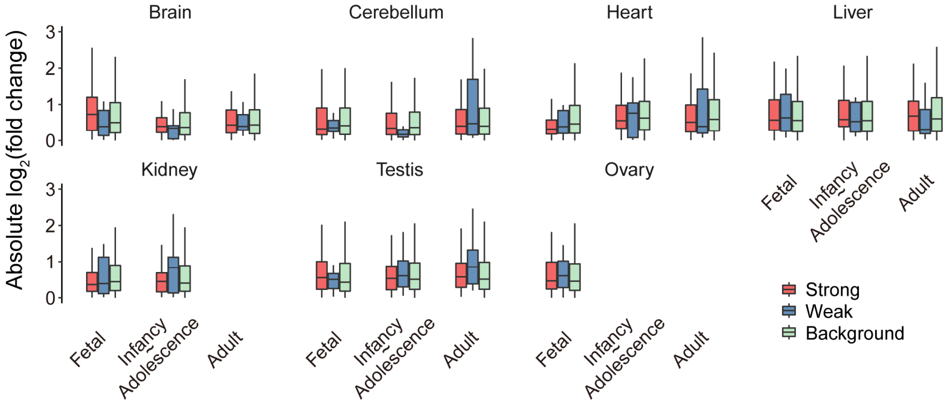

Fig. S12. Comparisons of gene expression in human and macaque tissues. Log_2-transformed fold changes in gene expression between human and macaque tissues across various developmental stages for strong effect inversions (Strong), weak effect inversions (Weak) and genome-wide ortholog pairs as a background (Background).

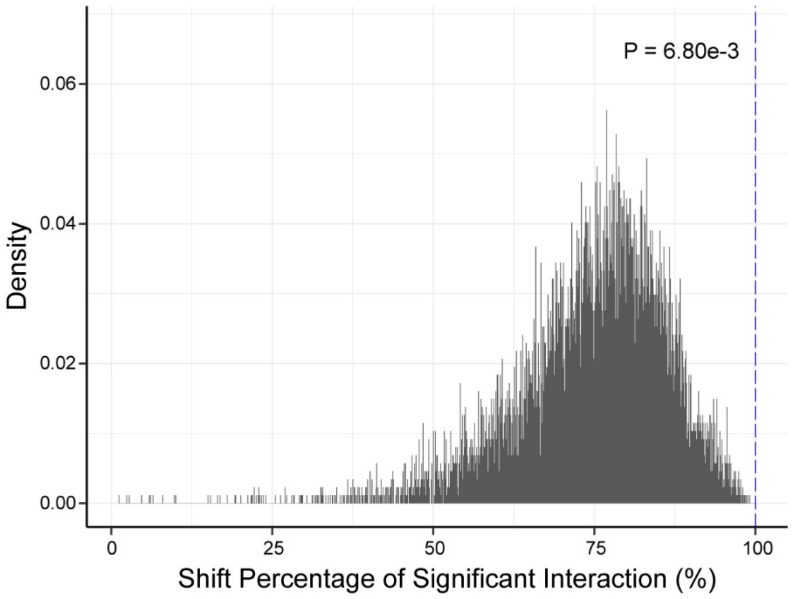

Fig. S13. Contact loss at chr18_inv1. Histogram showing the distribution of the "shift of significant interactions" in 10,000 permutations of randomly shuffled breakpoints, with the blue dotted line indicating the value of the breakpoint downstream of this inversion (chr18_inv1, BP2). A permutation test was performed to calculate the P value.

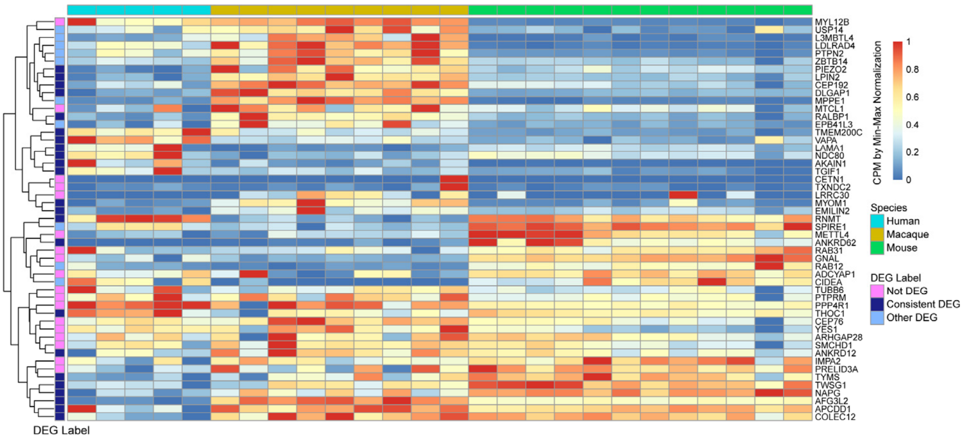

Fig. S14. Comparisons of gene expression levels across humans, macaques and mice. Heatmap showing the normalized expression levels of genes located on the highest ranked human-specific inversion. The expression levels of these genes (in CPMs) were estimated using the RNA-seq data of brain samples with corresponding developmental stages across the three species, which were further normalized and shown according to the color scales. DEG: differentially expressed gene.

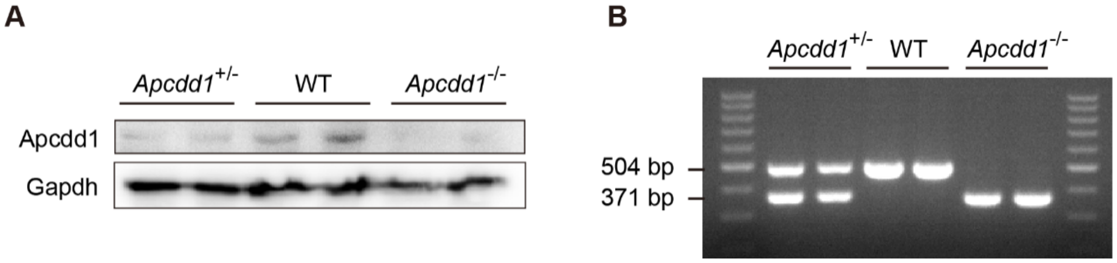

Fig. S15. Evaluation of knockout efficiencies in *Apcdd1*-deficient mice. (**A**) Western blots showing the protein expression of Apcdd1 and Gapdh in the brains of wild-type (WT) and *Apcdd1*-deficient mice (*Apcdd1*$^{+/-}$ and *Apcdd1*$^{-/-}$). (**B**) PCR-amplified products from wild-type (504 bp) and *Apcdd1*-deficient mice (*Apcdd1*$^{+/-}$ and *Apcdd1*$^{-/-}$). The bands of 504 bp and 371 bp indicate the PCR products encoded by wild-type *Apcdd1* and mutant *Apcdd1*, respectively.

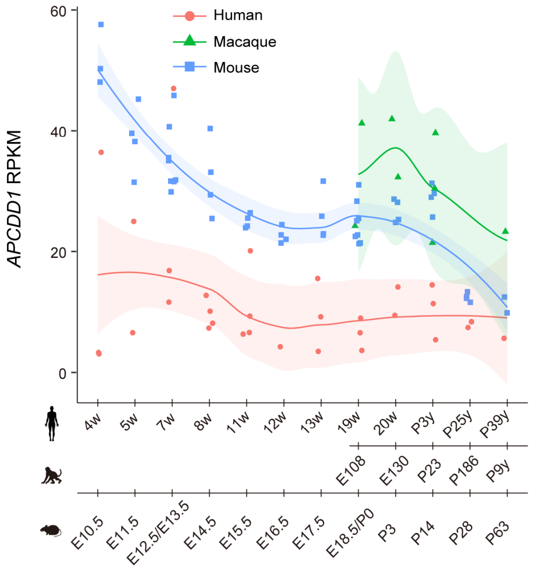

Fig. S16. Expression levels of *APCDD1* in the forebrains of humans, macaques and mice across developmental stages, with developmental correspondence based on this study (81).

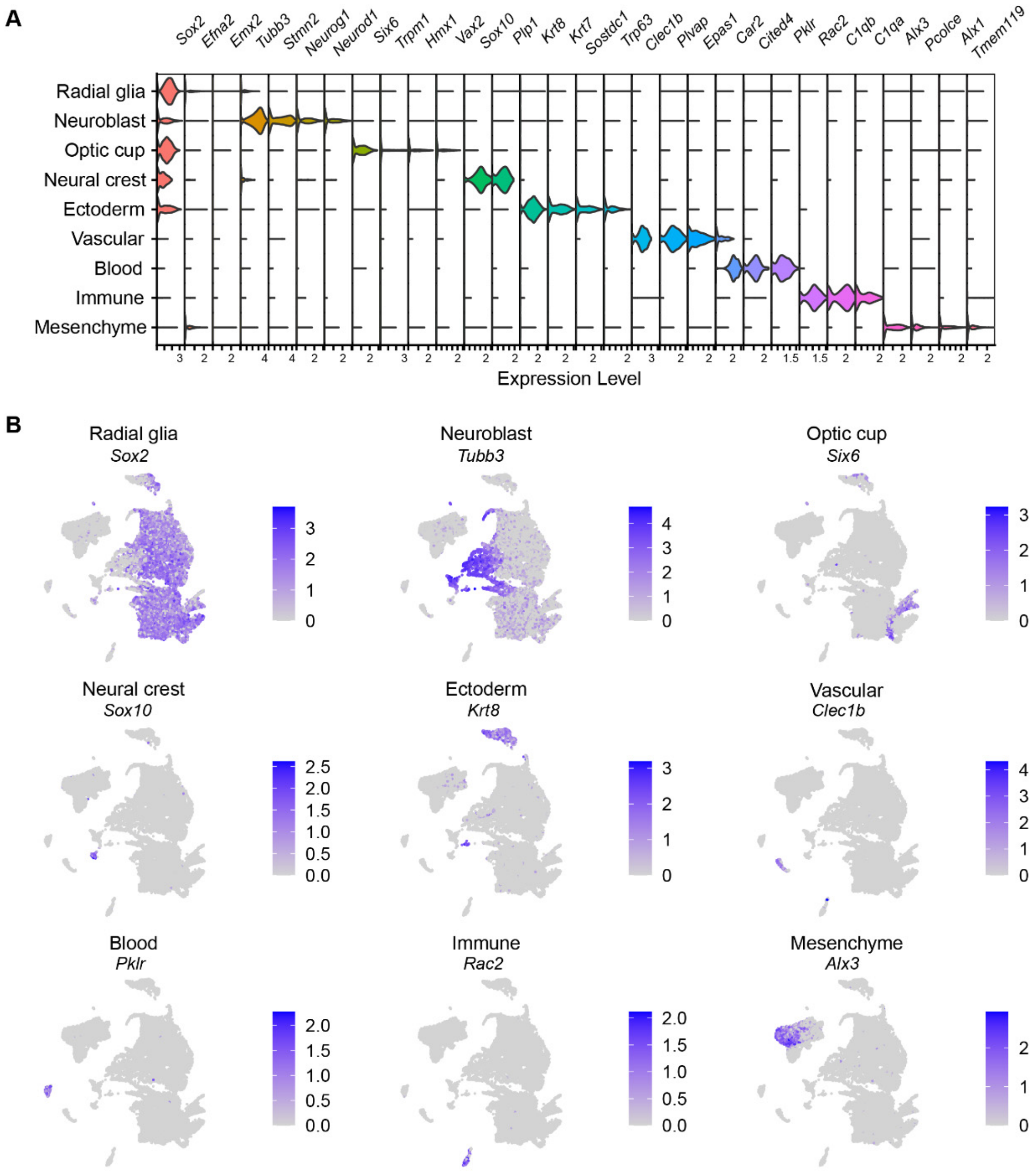

Fig. S17. Representative marker genes of scRNA-seq from E10.5 mice brains. (**A**) Violin plot of gene expression levels of select marker genes (columns) for each cell cluster (row). (**B**) Representative markers of scRNA-seq clustering results for each cell type.

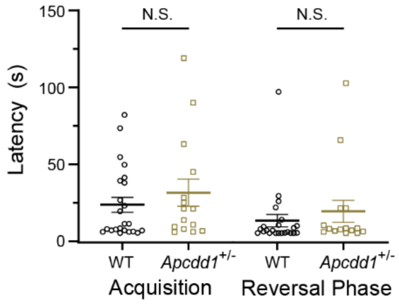

Fig. S18. Escape latency of visible platform trials. Distribution of the escape latency of visible platform trials for wild-type (WT) and *Apcdd1*-deficient mice (*Apcdd1*$^{+/-}$) in the acquisition and reversal phases. n = 38 mice (WT, n = 23; *Apcdd1*$^{+/-}$, n = 15). Two-tailed, unpaired Student's t test was performed. N.S., not significant.

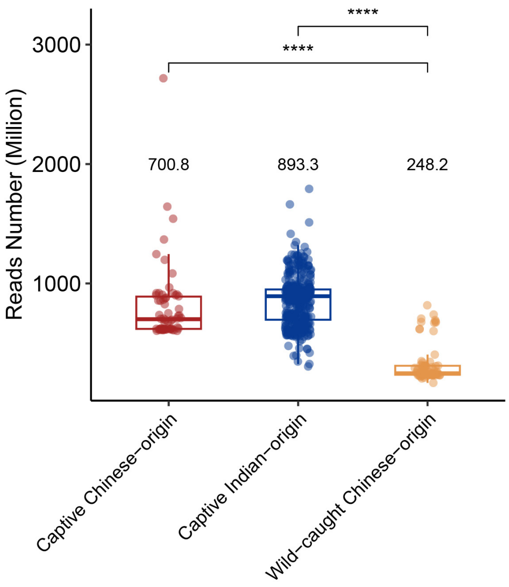

Fig. S19. Comparisons of the sequencing depth among three macaque populations. The distributions of the sequencing depth were shown, for wild-caught Chinese-origin macaques, captive Chinese-origin macaques, and captive Indian-origin macaques. Wilcoxon rank-sum test, ****P value < 0.0001.

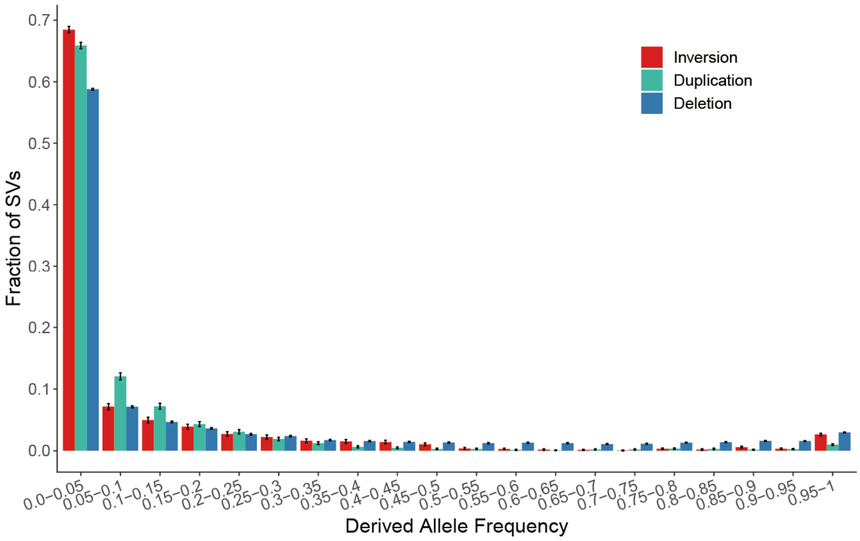

Fig. S20. Site frequency spectra of the derived alleles for inversions, duplications and deletions. In a population of 562 macaques, the site frequency spectra of the derived alleles for inversions (red), duplications (green) and deletions (blue) were shown and compared.

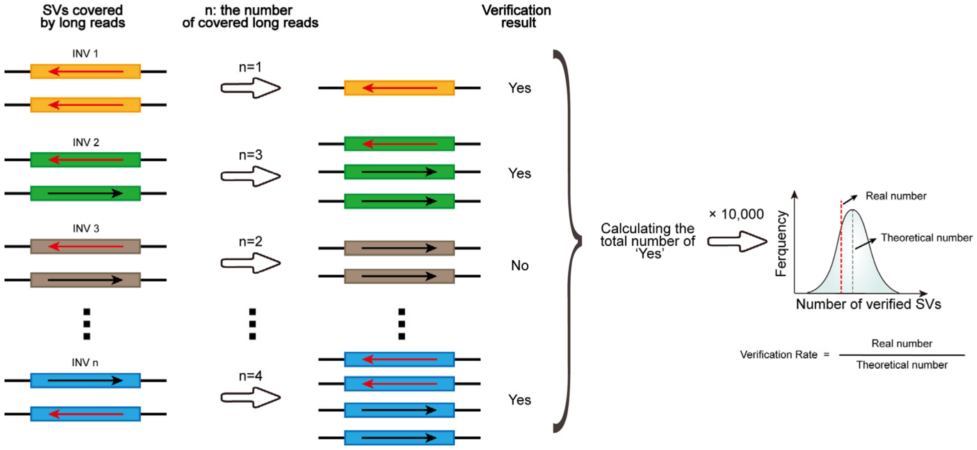
SVs covered by long reads
n: the number of covered long reads
Verification result
INV 1
n=1
Yes
INV 2
n=3
Yes
INV 3
n=2
No
INV n
n=4
Yes
Calculating the total number of 'Yes'
× 10,000
Real number
Theoretical number
Ferquency
Number of verified SVs
Verification Rate = Real number / Theoretical number

Fig. S21. Estimation of the theoretical number of SVs for long-read sequencing. For each macaque with both short-read and long-read sequencing data, the candidate SVs identified by short reads and covered by long reads were identified. For each of these SVs, the profile of long reads was simulated according to their coverage across the region and the genotype of the SV (homozygous/heterozygous) as defined by the short-read sequencing. This process was repeated for 10,000 times. The number of verified SVs was counted for each round of simulation, and the average number of verified SVs from the 10,000 simulations was defined as the theoretical number of verifiable SVs at the current sequencing depth.

Table S1. Statistics of the Bionano optical map data for one captive Chinese-origin macaque.

Length Bin	Molecule Count	Quantity (Gb)	Average Length (kb)	Molecule N50 (kb)	Labels (/100 kb)
>150 kb	1,606,066	477.2	297.11	317.43	9

Table S3. Statistics of PacBio HiFi reads for eight rhesus macaque samples.

Sample ID	HiFi Reads	HiFi Yield (bp)	HiFi Read Quality (median)	HiFi Read N50
920201	496,251	9,047,694,354	31	18,532
031024	423,194	8,305,659,930	31	20,601
97236	359,142	6,765,613,234	31	19,266
	63,001	1,303,569,644	30	20,544
97009_3_3	116,056	1,825,651,416	32	16,095
9109_35_1	61,779	1,044,755,305	31	17,093
100MGP-018	72,040	1,215,420,459	31	17,505
090089_3	35,080	635,061,116	31	18,604
100MGP-003	27,375	495,108,397	31	18,197

Table S6. Statistics for PacBio genome sequencing of one captive Chinese-origin macaque.

Type	Total Bases	Total Bases (Subreads)
RSII	110,826,954,922	101,952,187,859
Sequel	195,530,217,571	188,638,334,863
Total	306,357,172,493	290,590,522,722

Type	Min Read Length	Max Read Length	Mean Read Length	Read N50 Length	Read Count
Total	50	193,083	8,700	14,339	33,399,602

Table S8. Genomic coordinates of the probes and flanking primers used in the FISH assays.

Species	Genomic Region Index	Genomic Coordinates	Flanking Primers
Human (GRCh38)	a	chr7:38,505,934-38,966,169	CATCGAAGCGTGTGGCTACC
	b	chr7:69,818,995-70,273,729	CATCGAAGCGTGTGGCTACC
	c	chr7:91,071,663-91,570,490	TCGTTCCGCATTGACCAATC
Rhesus macaque (Mmul_10)	a	chr3:74,238,416-74,693,339	CCTGTGCGGAAATCGCGAGA
	b	chr3:49,980,073-50,434,877	TCGTTCCGCATTGACCAATC
	c	chr3:114,399,067-114,898,598	GGATTGCCGCATGGTTTCCG

Table S2 (separate file). Statistics of the whole-genome sequencing data of five Indian- and five Chinese-origin macaques used in the evaluation of gap closure.

Table S4 (separate file). Information for 27 captive Chinese-origin macaques in whole genome sequencing.

Table S5 (separate file). Statistics for 1,026 rhesus macaques with whole genome sequencing data.

Table S7 (separate file). 1,972 inversions between human and rhesus macaque.

Table S9 (separate file). 101 candidate human-specific inversions in human and macaque populations.

Table S10 (separate file). Metadata of 15 human genome assemblies used in this study.

Table S11 (separate file). Statistics of genome-wide contact maps for adult human and macaque.

Table S12 (separate file). Classification of 75 fixed human-specific inversions based on putative effects.