EVOLUTIONARY BIOLOGY

Adaptive functions of structural variants in human brain development

Wanqiu Ding[1]†, Xiangshang Li[1]†, Jie Zhang[1]†, Mingjun Ji[1]†, Mengling Zhang[2], Xiaoming Zhong[1,3], Yong Cao[4], Xiaoge Liu[1], Chunqiong Li[5], Chunfu Xiao[1], Jiaxin Wang[1], Ting Li[1], Qing Yu[1], Fan Mo[6], Boya Zhang[6], Jianhuan Qi[6], Jie-Chun Yang[1], Juntian Qi[1], Lu Tian[1], Xinwei Xu[1], Qi Peng[1], Wei-Zhen Zhou[7], Zhijin Liu[8], Aisi Fu[9], Xiuqin Zhang[1], Jian-Jun Zhang[10], Yujie Sun[2], Baoyang Hu[6], Ni A. An[1,11], Li Zhang[5]*, Chuan-Yun Li[1,5,11,12]*

Quantifying the structural variants (SVs) in nonhuman primates could provide a niche to clarify the genetic backgrounds underlying human-specific traits, but such resource is largely lacking. Here, we report an accurate SV map in a population of 562 rhesus macaques, verified by in-house benchmarks of eight macaque genomes with long-read sequencing and another one with genome assembly. This map indicates stronger selective constrains on inversions at regulatory regions, suggesting a strategy for prioritizing them with the most important functions. Accordingly, we identified 75 human-specific inversions and prioritized them. The top-ranked inversions have substantially shaped the human transcriptome, through their dual effects of reconfiguring the ancestral genomic architecture and introducing regional mutation hotspots at the inverted regions. As a proof of concept, we linked *APCDD1*, located on one of these inversions and down-regulated specifically in humans, to neuronal maturation and cognitive ability. We thus highlight inversions in shaping the human uniqueness in brain development.

INTRODUCTION

Genetic variation is a source of genetic novelty in shaping population structures and species-specific traits (*1*–*8*), which could be divided into categories based on their sizes, ranging from single-nucleotide variants (SNVs) to large-scale structural variants (SVs). SVs can be further classified into two categories: balanced SVs and unbalanced SVs (*9*). Unbalanced SVs are accompanied by gains or losses of DNA fragments, such as deletions and duplications, whereas balanced SVs can cause chromosomal rearrangements, such as inversions and translocations. Considering their larger sizes, SVs are expected to have stronger effects on transcription regulation than SNVs and short insertion/deletion (indel) variants (*10*). However, despite a growing awareness of their significance, the characterization and functional interrogation of SVs have largely lagged behind that of SNVs, partially due to the technical challenges in accurately identifying SVs with short reads obtained via the next-generation sequencing. Moreover, the public benchmarks for the evaluation of the performance of SV detection are limited to a small number of validated SVs in human samples (*11*–*13*), further hindering the studies of complex SVs and those in other species.

Rhesus macaque (*Macaca mulatta*) is a nonhuman primate species closely related to humans in terms of the genome sequences and the physiology (*14*, *15*). Quantification of SVs in macaque populations could thus promote the understanding of their features, turnover, and evolutionary significances, and further provide a niche to clarify the genetic backgrounds underlying human-specific traits. Recent advances in the assembly of macaque reference genomes and the generation of a batch of genome resequencing data have facilitated the profiling of genetic variants in macaque populations (*6*, *15*–*20*). However, the data are scattered throughout the literatures and are largely generated with short-read sequencing, which is error-prone to be used in SV calling with standardized algorithms. In addition, the deep integration of these data from multiple subpopulations of macaques and the issue of kinships among these animals are not well addressed, further confounding the in-depth population genetics studies. Overall, the challenges in the deep integration of these confounded and scattered macaque population genomics data, the difficulty of accurate identification of SVs with short-read sequencing, and the lack of comprehensive benchmarks for macaque SV evaluation have hindered the clarification of the functions and evolutionary significance of SVs in primates.

Here, we provide an accurate macaque SV atlas by integrating whole genome sequencing (WGS) data from 1026 macaques. We addressed the issue of kinships using genome-wide SNV profile, developed an accurate pipeline for SV calling, and established a three-tier benchmark for comprehensive SV evaluation. Furthermore, we explored the evolutionary turnover of these SV events and proposed a practical strategy for prioritizing those with the most important functions in shaping human adaptive evolution. On the basis of this, we identified 75 human-specific inversions and prioritized those inversions which have substantially shaped the human brain transcriptome.

[1] State Key Laboratory of Protein and Plant Gene Research, Laboratory of Bioinformatics and Genomic Medicine, Institute of Molecular Medicine, College of Future Technology, Peking University, Beijing, China. [2] State Key Laboratory of Membrane Biology, Biomedical Pioneer Innovation Center (BIOPIC), School of Life Sciences, Peking University, Beijing, China. [3] Center of Excellence for Leukemia Studies, St. Jude Children's Research Hospital, 262 Danny Thomas Place, Memphis, TN 38105, USA. [4] Department of Neurosurgery, Beijing Tiantan Hospital, Capital Medical University, 119S Fourth Ring Rd W, Fengtai District, Beijing, China. [5] Chinese Institute for Brain Research, Beijing, China. [6] State Key Laboratory of Stem Cell and Reproductive Biology, Institute of Stem Cell and Regeneration, Institute of Zoology, Chinese Academy of Sciences, Beijing, China. [7] State Key Laboratory of Cardiovascular Disease, Fuwai Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China. [8] College of Life Sciences, Capital Normal University, Beijing, China. [9] Wuhan Dgensee Clinical Laboratory, Wuhan, China. [10] Shanxi Key Laboratory of Chinese Medicine Encephalopathy, National International Joint Research Center for Molecular Chinese Medicine, Shanxi University of Chinese Medicine, Jinzhong, China. [11] National Biomedical Imaging Center, College of Future Technology, Peking University, Beijing, China. [12] Southwest United Graduate School, Kunming 650092, China.
*Corresponding author. Email: chuanyunli@pku.edu.cn (C.-Y.L.); zhangli@cibr.ac.cn (L.Z.)
†These authors contributed equally to this work.

RESULTS

Definition of a well-annotated population of independent macaques

To define a clean population of independent macaques for unbiased population genetics analyses of SVs, we first attempted to establish a comprehensive SNV atlas of macaques, which could provide a genetic basis to examine their identity information, such as subpopulations, sexes, and kinship relationships. To this end, we first constructed a better reference macaque genome through the integration of three recent genome assemblies based on the third-generation sequencing. As Mmul_10 (rheMac10) represents the genome assembly with the highest integrity in sequence continuity and base accuracy (fig. S1, A and B), we used this assembly as the template and filled gaps by aligning with it the sequences of two other genome assemblies from an Indian-origin macaque (rheMac8) and a Chinese-origin macaque (rheMacS). A total of 40 gaps in Mmul_10 were filled with the corresponding sequences from the two genomes and further evaluated with the high-coverage Bionano optical map from a macaque and PacBio long-read sequencing from eight additional macaques (Fig. 1A, figs. S1 and S2, tables S1 to S3, and Materials and Methods). The results indicate that these gaps should represent real genomic gaps in Mmul_10 assembly, rather than the assembly errors from other macaque genome assemblies, or individual-specific SVs. Notably, the 40 gap regions that were closed here were distributed across the genome spanning 721,984 base pairs (bp), and some were located in regions with high gene density (Fig. 1A). We named the genome with filled gaps "rheMac10Plus" and performed subsequent analyses with this improved reference genome.

On the basis of the improved macaque genome, we performed whole genome sequencing of 27 captive Chinese-origin macaques and achieved an average coverage of 32-fold (table S4). We also integrated public genome resequencing data of 41 captive Chinese-origin macaques, 76 wild-caught Chinese-origin macaques, and 882 captive Indian-origin macaques (Fig. 1B and table S5) (*16*–*19*). Overall, the genome resequencing data of a total of 1026 macaques were obtained (Fig. 1B), in which 862 of 1026 (84.0%) were sequenced at a depth of more than 30-fold (table S5). The resequencing data of the 1026 macaques were then aligned to the rheMac10Plus genome and further subjected to a two-round variant calling (Fig. 1C and Materials and Methods). A total of 81.3 million SNVs were then identified.

On the basis of this genetic profile, we then carefully examined the identity information of these macaques to remove animals that would confound the conclusions of the subsequent population genetic analyses of SVs (table S5). Notably, through a principal components analysis (PCA) of the SNV genotypes across the 1026 macaques, we found that 12 animals initially identified as having an Indian origin were actually of Chinese origin (fig. S3A). We then examined the reported sex information of these animals by checking the read density along the X chromosome, and found that four female macaques were falsely recorded as male, while one male macaque was reported as a female animal (fig. S3B). We further assigned the information for 74 wild-caught Chinese-origin macaques with missing sex information (fig. S3C). Finally, we carefully examined the kinship of these macaques according to their genetic profiles and subsequently removed 454 macaques that were closely related to other macaques (Fig. 1C and Materials and Methods). Overall, after stringent filtering steps, 572 independent macaques were retained in the following analyses, including 434 captive Indian-origin macaques, 63 captive Chinese-origin macaques, and 75 wild-caught Chinese-origin macaques (table S5).

The PCA of the 572 independent macaques revealed distinct divergence between the Indian-origin and Chinese-origin macaques (Fig. 1D), while the captive Chinese-origin macaques were indistinguishable from the wild-caught Chinese-origin macaques (Fig. 1D), indicating a weaker effect of domestication on the genetic backgrounds of macaques in comparison to the effect of geography. As the wild-caught Chinese-origin macaques with definite habitat information could be divided into five subpopulations widely used in biomedical studies (*18*), we assigned subpopulation information to other Chinese-origin macaques by constructing a neighbor-joining phylogeny combining the two groups of macaques, in which 72 additional Chinese-origin macaques were annotated with the subpopulation information accordingly (Fig. 1E and Materials and Methods).

Overall, we defined a well-annotated population of 572 independent macaques, resulting in a macaque SNV profile with a total of 79.6 million SNVs and 9.1 million indels. Consistent with previous findings (*16*, *19*), we observed a comparable transition-to-transversion ratio, and increased nucleotide diversity in both Indian- and Chinese-origin macaque populations in comparison to that in humans (Fig. 1F). This population thus represents a clean population for unbiased population genetics analyses of SVs.

Construction of a macaque SV map

On the basis of the resequencing data of 572 independent, well-annotated macaques, we next attempted to identify SV events at the population scale. Notably, it is error-prone to use standardized tools to identify SVs with short reads. To identify an accurate SV map, we then developed a pipeline by carefully adjusting the parameters of each tool, performing meta-analyses to integrate the results of different tools, and introducing stringent filters to control for false positives (Fig. 2A and Materials and Methods). In such a case, the requirements to define an SV event are more stringent than previous practices. As a note, the parameters were set and adjusted according to a benchmark of two macaques with both short-read resequencing data and the genome assemblies available (Mmul_10 and rheMacS). In particular, in this pipeline, the signatures of paired-end reads and split reads were adequately integrated to capture candidate regions of inversions (Fig. 2A), a type of balanced SVs that is difficult to detect using short reads. Although most of the reads are not informative in indicating their position and boundaries, the split reads and the poorly paired reads in paired-end sequencing are informative to indicate the inversion event. It is thus practical to pinpoint these events with high-coverage short reads, as long as the two types of informative reads are adequately analyzed. As a proof of concept, we included an example for one inversion we identified to illustrate the principle in calling inversions with both short reads and long reads (Fig. 2B).

After identifying the SVs of each macaque, we then merged the SV calls and genotyped each SV in all of the 572 macaques (Fig. 2A). Briefly, considering the difficulties in defining SV breakpoints in each macaque at a single-nucleotide resolution, which may lead to the underestimation of the SV allele frequency, we defined "SV hotspot" indicating the consensus SV regions shared by different macaques with slightly different boundaries (Fig. 2A and Materials and Methods). Large SVs with lengths of more than 10 Mb were

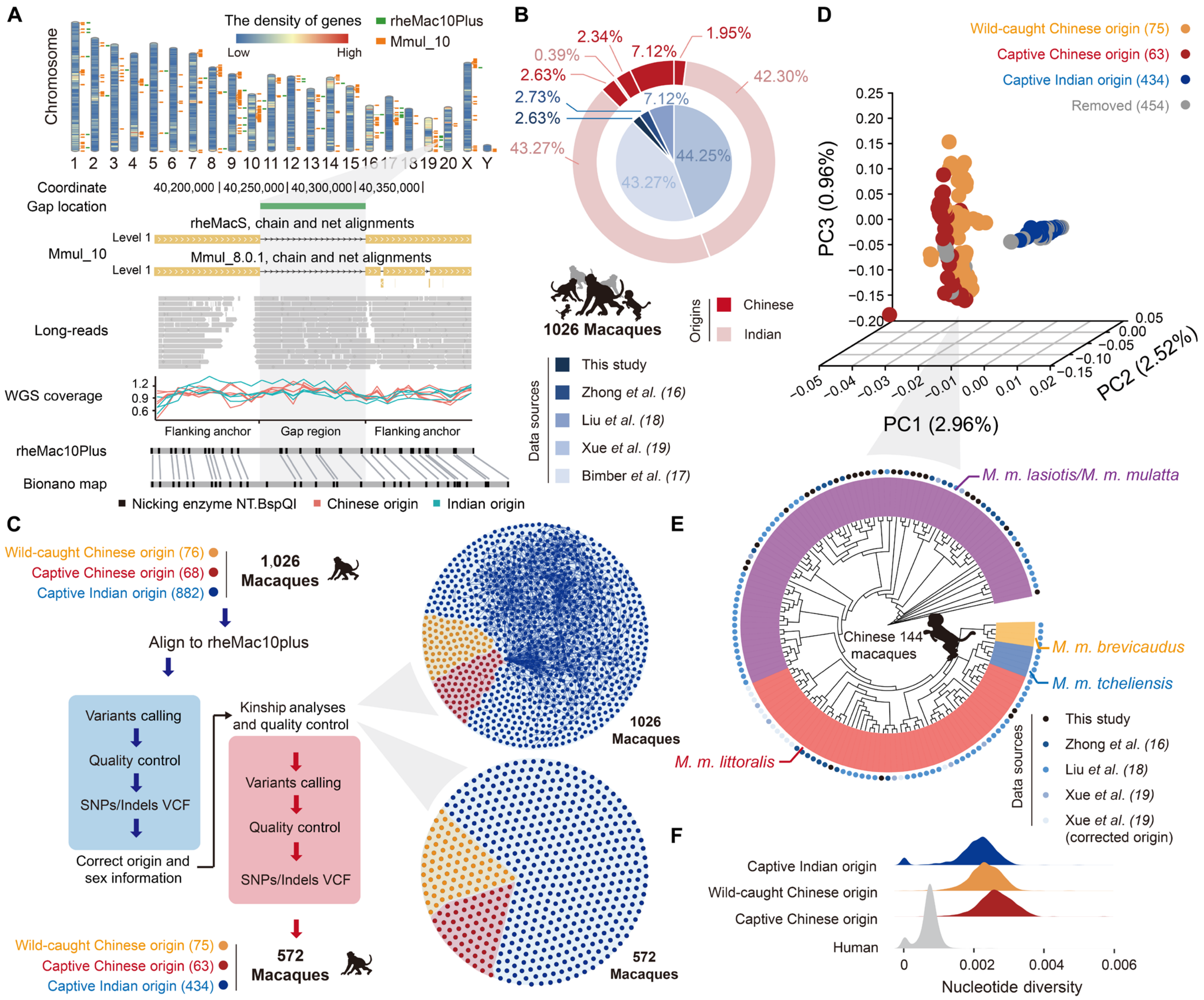

Fig. 1. Population genetic landscape of 1026 macaque genomes. (**A**) Chromosome karyotype showing 40 filled gaps, as indicated by green bars in the rheMac10Plus assembly. The density of genes across the genome is shown in the heatmap. For one of the filled gaps on chromosome 19, the Bionano optical map of one macaque, the long reads of eight macaques, and the coverage of the short reads of 10 macaques (red: Chinese-origin macaques; green: Indian-origin macaques) were aligned and shown accordingly. (**B**) Sources (inner layer) and geographic origins (outer layer) of the 1026 macaques. (**C**) Schematic diagram of the workflow for variant calling with a two-round strategy (blue: first round of calling; red: second round of calling). The original set of macaques (1026 animals) and the set after quality control (572 animals) were partitioned into three clusters (red: captive Chinese-origin macaques; yellow: wild-caught Chinese-origin macaques; blue: captive Indian-origin macaques) based on their genetic profiles. The pairs of macaques with significant kinship relationships are linked by lines. (**D**) Three-dimensional PCA plot showing the relationships of the 572 macaques according to SNV genotypes. (**E**) Neighbor-joining tree showing the genetic distance of the Chinese-origin macaques. Different data sources are indicated by colored dots in the outer layer. Yellow: *M. m. brevicaudatus*; blue: *M. m. tcheliensis*; orange: *M. m. littoralis*; purple: *M. m. lasiotis* or *M. m. mulatta*. (**F**) The genome-wide distribution of nucleotide diversity of captive Indian-origin macaques (blue), captive Chinese-origin macaques (red), wild-caught Chinese-origin macaques (yellow), and humans (gray).

removed to avoid false positives in SV identification with short reads, and 10 macaques with abnormally high numbers of SV hotspots were eliminated from the subsequent analyses (Materials and Methods). Finally, a total of 20,985 SV hotspots were defined for the remaining 562 macaques, including 1335 inversions, 17,267 deletions, and 2383 duplications. As a note, the distance between two adjacent SV hotspots across the genome (median distance: 1,284,360 bp for inversions, 96,398 bp for deletions, and 743,330 bp for duplications) is typically much larger than the size of these SV hotspots (median size: 1025 bp for inversions, 954 bp for deletions, and 330 bp for duplications), indicating that these hotspots could be clearly distinguished from neighboring hotspots. Using this atlas of SV hotspots as a reference, we further genotyped the SVs in each macaque and calculated the allele frequency for each SV hotspot in the population, which substantially increased the sensitivity of SV calling in comparison to the previous round of de novo

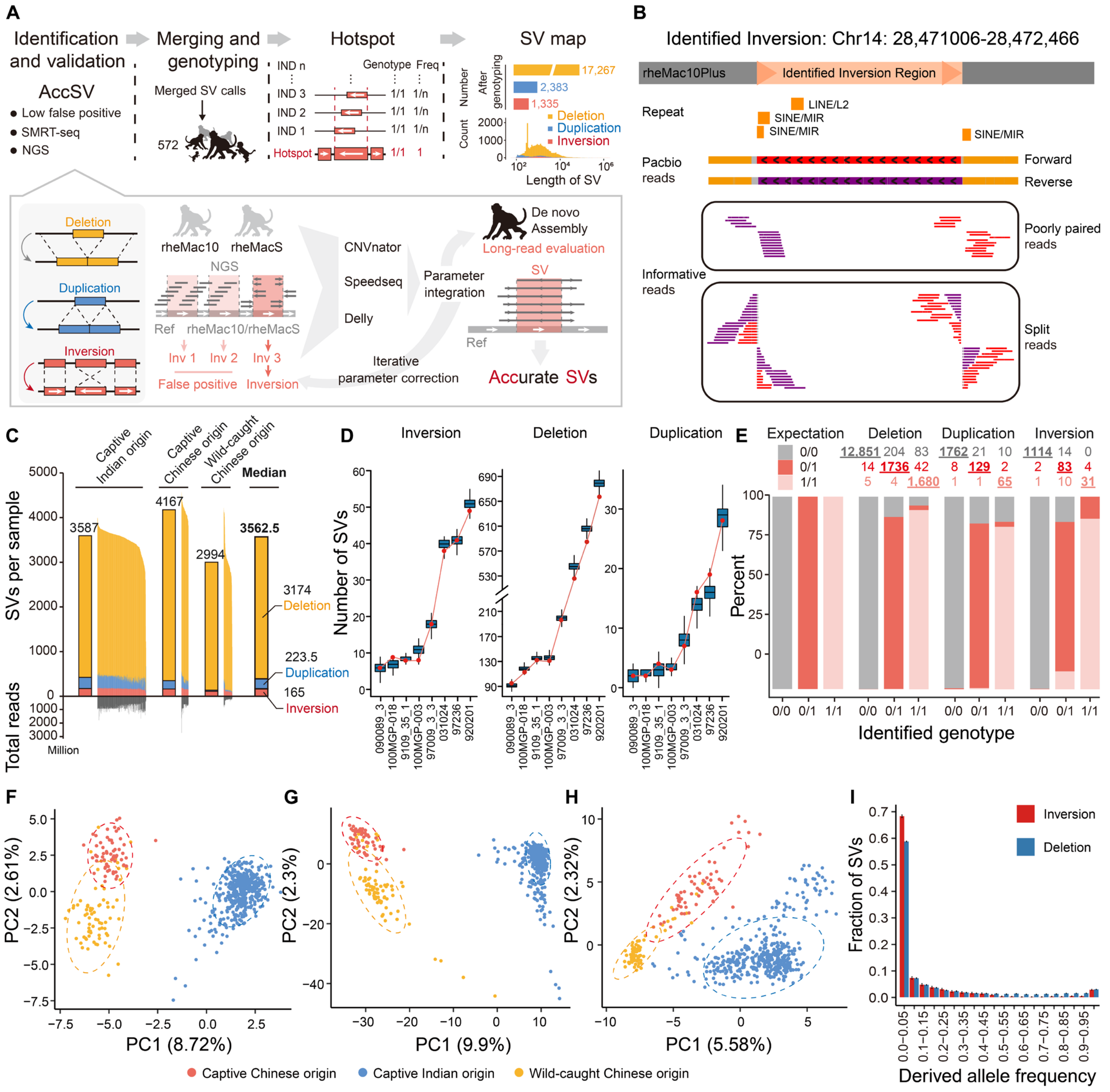

Fig. 2. Construction and characterization of SV map for macaque population. (**A**) The pipeline for SV map construction, including the processes of SV identification, validation, genotyping, SV hotspot definition, and allele frequency calculation. (**B**) The genomic region of one inversion we identified was shown as an example, with the split reads, poorly paired reads, and the long reads supporting its existence aligned and shown accordingly. (**C**) The distribution of the count of SVs per macaque genome for three types of SVs in macaques of different origins. The median number of SVs is shown for each group. The total number of reads of deep sequencing for each macaque is also shown. (**D**) Verification of the SV events with long HiFi reads of eight macaques. Boxplots showing the distribution of the theoretical number of verified SVs at the current sequencing depth of HiFi reads, obtained from 10,000 times of simulations. The detected number of verified SVs in each macaque was indicated by the red dot. (**E**) Validation of the genotypes of SVs in one macaque based on the long-read sequencing and genome assembly in one macaque. For each type of genotype identified with short reads (0/0, 0/1, or 1/1), the percentages of verified SVs are summarized and shown in different colors. The numbers of SVs of each type are shown, and those with verified genotypes are underlined. (**F** to **H**) PCA plots showing the relationships of the 562 macaques according to the genotypes of inversion (F), deletion (G), and duplication (H) variations. Macaques with different origins are labeled with different colors. (**I**) Site frequency spectra of the derived alleles for inversions and deletions in 562 macaques.

calling. Overall, a median of 3562 SVs were detected in each macaque, including 165 inversions, 3174 deletions, and 223 duplications (Fig. 2C).

To confirm that these events represented bona fide SVs, we then developed a three-tier benchmark to evaluate the performance of our pipeline in SV calling. We first evaluated the pipeline with public benchmarks in human. Briefly, we selected the HG002 genome with verified deletion calls as the benchmark callset (*21*), and evaluated the performance of our pipeline in calling these deletion events. Notably, the pipeline achieved a high precision score (97.8%). We then used another benchmark callset of HG001 genome from Pendleton *et al.* (*22*) to evaluate the performance of our pipeline in calling inversions. The pipeline achieved a precision score of 88.9% in calling these inversions. As the parameters were adjusted according to the features of macaque short-read sequencing data, the real precision score in calling SVs in macaque populations should be higher.

Considering the complexity of SVs and the lack of comprehensive, high-quality reference SV standards in macaques, we further developed in-house benchmarks for evaluating the SV calling in macaques, according to the principles of public benchmarking standards, in that the long sequencing reads, especially the de novo genome assembly, could provide a more accurate SV atlas. To this end, from the population of 562 macaques, we selected 8 macaques and sequenced their genomes with different coverages of long HiFi reads (table S3 and Materials and Methods). For each macaque animal, we then evaluated the performance of our pipeline in calling deletions, inversions, and duplications, using SVs called in this macaque by long HiFi reads as a benchmark. As the sequencing of these samples was not saturated, for each SV type in each macaque animal, we performed a simulation strategy to estimate the theoretical number of verified SVs at the current sequencing depth, assuming that the SVs and their genotypes were accurately identified with short reads (Materials and Methods). Overall, the average verification rates (97.1%, 95.2%, and 97.3% for deletions, inversions, and duplications, respectively) indicate the high accuracy of our pipeline in SV calling with short reads (Fig. 2D).

Finally, for another Chinese-origin, male rhesus macaque with normal phenotype from the population of 562 macaques, we sequenced its genome with high coverage long-read sequencing, and then de novo assembled its genome, which was further used to evaluate the performance of our pipeline in SV calling with short reads (Materials and Methods). Specifically, we applied single-molecule real-time (SMRT) long-read sequencing technology and sequenced the genomic DNA. Data (290 Gbp) were generated, with the N50 of the subread length of 14.3 kb and an average genome coverage of 96.9-fold (table S6 and Materials and Methods). We further de novo assembled its genome on the basis of the integration of the short-read sequencing, long-read sequencing and Bionano optical data, resulting in a genome assembly with 2.99 Gbp informative bases, supported by 4742 contigs (N50 = 4.6 Mbp, Materials and Methods). On the basis of the long reads and the assembled genome, we then evaluated the genotype of each SV identified with short reads of the same macaque (Fig. 2E). The verification rate (92.3%, 90.1%, and 86.4% for deletions, inversions, and duplications, respectively) indicates that the SVs we identified with short reads are accurate. For each type of SV events, one verified case is shown in fig. S4.

In this accurate SV atlas of 562 macaques, we found similar densities of SVs across macaques with different origins and sexes, a pattern consistent with that in human subpopulations (*3*) (Fig. 2C and fig. S5A). Furthermore, although the number of SVs located on each chromosome was correlated with the length of the chromosome, a pattern consistent with previous reports (fig. S5, B to D) (*3*, *23*), we observed enrichments of inversions on chromosomes 16 and 19, and a depletion on chromosome 18, as well as an enrichment of deletions on chromosome 5, and depletions on chromosomes 7 and X. These regions with unbalanced distributions of SVs need further investigation. Similar to the SNV profile (Fig. 1D), the profile of these SVs could discriminate macaques from different subpopulations with an even higher discrimination efficiency (Fig. 2, F to H). Of note, the genetic distances between the captive Chinese-origin macaques and the wild-caught Chinese-origin macaques were smaller than those between the captive Chinese-origin macaques and the captive Indian-origin macaques, recapitulating the above conclusion based on SNVs for a weaker effect of domestication on the genetic background in comparison to that of geography (Figs. 1D and 2, F to H).

Selectively constrained inversions in macaque population

To investigate whether SVs are selectively constrained, we inferred the ancestral state of each SV and compared the site frequency spectra of the derived allele for each type of SV events (Materials and Methods). Previous studies have reported that deletions are more deleterious than duplications (*24*). Notably, we found that the allele frequency spectrum of inversions was more left-skewed than that of deletions, indicating even stronger purifying selection on the fixation of inversions than deletions (Fig. 2I). This finding is consistent with previous reports that polymorphic inversions are largely deleterious due to recombination suppression and the subsequent accumulation of deleterious mutations (*25*–*28*). Considering the possibly stronger effects of inversion, we next focused specifically on this type of SV in our subsequent evolutionary and functional genomics analyses (Discussion).

We first investigated whether inversions with different features or genomic locations were shaped by natural selection to different degrees (Fig. 3A). When inspecting the distribution of these inversions, we found that they tended to be depleted in functional regions, such as exons (Permutation test, $P < 0.001$), putative promoters (Permutation test, $P < 0.001$), and enhancers (Permutation test, $P = 0.02$; Fig. 3B), as significantly fewer events were located in these regions than in randomly shuffled, length-matched regions used as a background. In contrast, these inversions were overrepresented in intergenic regions (Permutation test, $P < 0.001$; Fig. 3B). Accordingly, the inversions in exonic regions showed an excess number of low-frequency variants of the derived allele, in comparison to those located on intronic regions (Wilcoxon rank-sum test, $P < 2.2 \times 10^{-16}$) or intergenic regions (Wilcoxon rank-sum test, $P < 2.2 \times 10^{-16}$; Fig. 3C).

When further dividing these inversions into groups according to their sizes and locations on the three-dimensional genome (Fig. 3A), we found that the inversions with larger sizes showed an excess proportion of low-frequency variants compared with short inversions (Wilcoxon rank-sum test, $P < 2.2 \times 10^{-16}$; Fig. 3C). Moreover, inversions disrupting topologically associating domain (TAD) boundaries, as defined in macaque fetal brains (*29*), also showed an excess

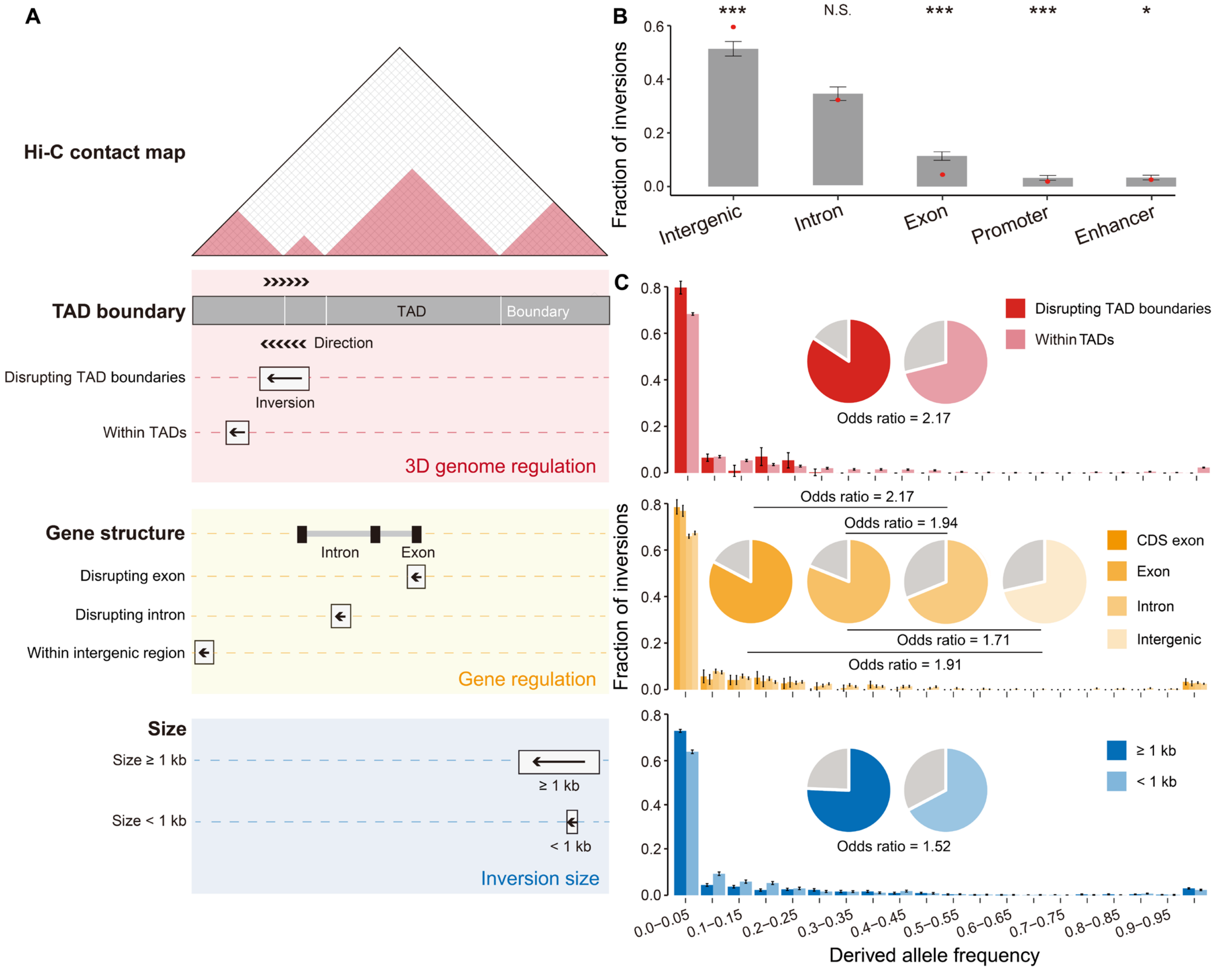

Fig. 3. Inversions in regulatory regions are selectively constrained. (**A**) Classification of inversions by different features and genomic locations, including the sizes of the inversions, their locations on the genes, and their three-dimensional genomic architecture. (**B**) Proportions of inversions at different genomic locations. The background distribution of inversions located in each genomic region, as estimated based on 1000 shuffled regions with matched lengths, is shown in a bar plot, with the error bars representing the standard deviations. For each bar plot, the observed value is indicated as a red dot, with the empirical *P* value calculated as the percentage of the 1000 replicates. **P* < 0.05, ****P* < 0.001, N.S., not significant. (**C**) Site frequency spectra of the derived allele for different classifications of inversions. For each group of inversions, the fraction of inversions with a low frequency of derived alleles (less than 5%) is shown and compared, with the odds ratios shown accordingly.

of low-frequency variants, in that the frequency spectrum of the derived allele was significantly left-skewed relative to that of the inversions located within TADs (Wilcoxon rank-sum test, $P < 2.2 \times 10^{-16}$; Fig. 3C and fig. S6).

We then investigated the distribution of the inversions with extremely low frequencies in various genomic regions. In contrast to inversions with higher frequencies, low-frequency inversions were typically larger in size (odds ratio = 1.52), and located with a higher proportion on gene regions (odds ratio = 2.17) or regulatory regions disrupting TAD boundaries (odds ratio = 2.17; Fig. 3C). These findings thus indicate that the inversions located in regulatory regions are subjected to stronger purifying selection, suggesting a practical strategy for prioritizing them with the most important functions, as the fixed inversions with stronger effects on gene structure or expression regulation were more likely maintained by selective pressure owing to their adaptive functions.

Identification of 75 fixed human-specific inversions

Prompted by the assumption that the list of fixed inversions with stronger effects on gene structure or expression regulation should be enriched with functional inversions driving species-specific traits, we next identified human-specific inversions based on the above macaque SV atlas and comparative genomics analyses in multiple outgroups. To this end, we first identified 1972 species-specific inversions between humans and macaques, ranging in size from 51 bp to 69 Mb, through genome-wide alignment followed by intensive manual curation in the UCSC genome browser (Fig. 4A, table S7, and Materials and Methods). The number of species-specific

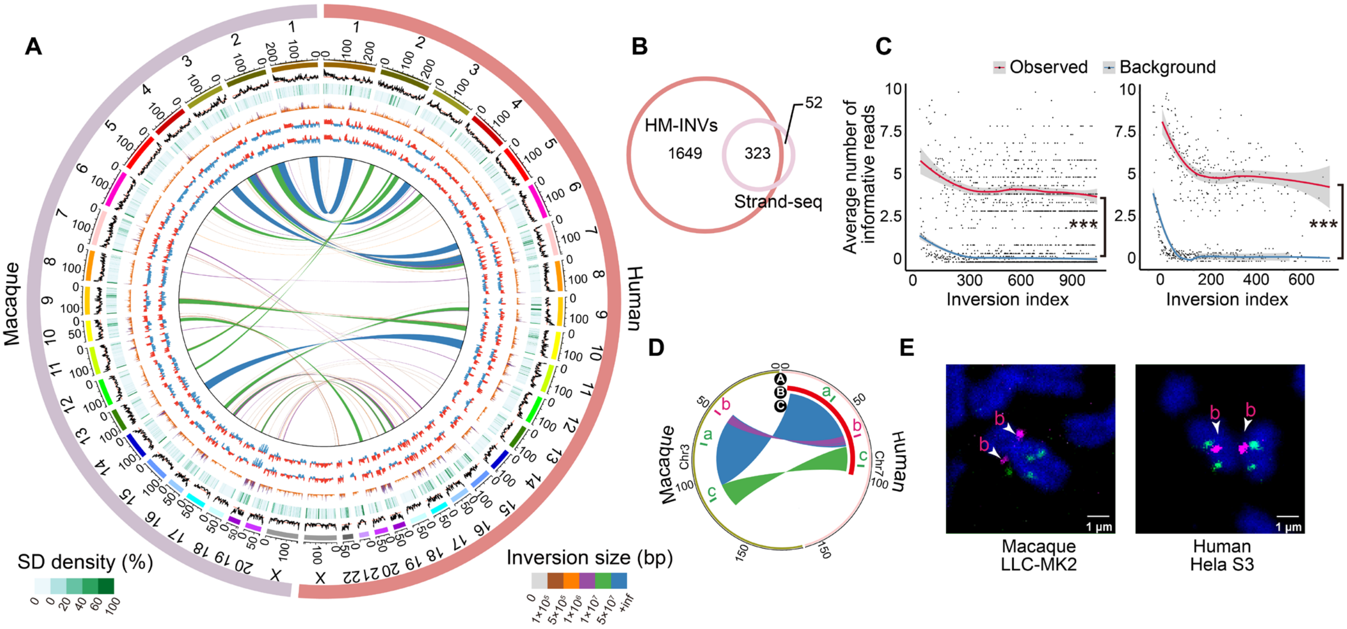

Fig. 4. Identification and verification of species-specific inversions between humans and macaques. (**A**) Circos plot showing the profile of inversions across humans and macaques (HM-INVs), with genomic features aligned according to the coordinates. From the outside to the inside: GC content (%), segmental duplication (SD) density (%), gene density, A/B compartments from fetal cortical plates (CP) and germinal zone (GZ), and the locations of HM-INVs. The average GC contents are indicated by orange lines. Tracks are plotted in 500-kb windows. (**B**) Overlap between HM-INVs in this study and the public list of species-specific inversions between humans and macaques as defined by in Maggiolini *et al.* (*4*) (Strand-seq). (**C**) Validation of species-specific inversions between humans and macaques with Strand-seq data. For candidate inversions with reads coverage ≥3 in the Strand-seq study, the average numbers of Strand-seq informative reads (Observed) were shown and compared with the background (Background, see details in Materials and Methods), for candidates identified specifically in our study (left), or by both studies (right). Inversions were arranged in descending order of their length. Local regression curves for the average numbers of the informative reads (red) and the background (blue) were shown. Wilcoxon rank-sum tests, ****P* < 0.001. (**D**) Circos plot depicting the arrangement of one complex HM-INV chosen for FISH validation. The track A represents the genomic regions where the probes were designed, with the order of colors indicating the expected form of inversions in humans and macaques based on the definition in this study. Tracks B and C display the forms of these inversions identified by Strand-seq (one large inversion) and in this study (three complex inversions with breakpoint reuse), respectively. (**E**) Validation of the complex HM-INV in (D) in the macaque LLC-MK2 cell line (left) and human HeLa S3 cell line (right).

inversions on each chromosome was correlated well with the length of the chromosome, consistent with the observation of the polymorphic inversions in macaque population (fig. S7).

To verify the accuracy of the list of species-specific inversions between humans and macaques, we first compared it with a previous study reporting the species-specific inversions between humans and macaques based on Strand-seq data (*4*). Of the 375 inversions reported by Maggiolini *et al.*, 323 (86.1%) were also identified in this study, and we substantially expanded the list by including 1649 additional inversions (Fig. 4B). Notably, the 1649 additional inversions identified only in our study should also represent bona fide, species-specific inversion events. First, the list of species-specific inversions between humans and macaques was identified through comparative analyses of the assembled genomes, rather than de novo identification with short- or long-read sequencing data. In such a case, the boundaries of each species-specific inversion were supported by the original long reads used in the assembly of the human/macaque genomes, and the false positives possibly introduced by the segmental duplications near the inversions could be well controlled. Second, when performing pairwise genome alignments between human and three other old world monkeys, including crab-eating macaque, baboon, and green monkey, 93.6% of these inversions were also supported by the genome assembly of at least one of these monkeys, providing additional verification of these events from the perspective of phylogenetic relationships. Third, we investigated why these 1649 inversions detected in this study were not identified in the original Strand-seq data. To this end, we analyzed the 61 single-cell libraries that were used to detect inversions in the original Strand-seq study. Notably, Strand-seq performs well to identify long inversions, while its sensitivity in identifying short inversions, or inversions with lower sequencing coverage, is relatively low. As expected, for the 1649 inversions identified only in our study, when tracing the original Strand-seq signals, we found that the number of informative reads in these regions were actually significantly higher than that of the background (Materials and Methods, Wilcoxon rank-sum tests, $P < 2.2 \times 10^{-16}$; Fig. 4C), a pattern consistent with that of the shared list of 323 inversions (Wilcoxon rank-sum tests, $P < 2.2 \times 10^{-16}$; Fig. 4C). Moreover, for one discordant inversion between the two studies, we performed interphase fluorescence in situ hybridization (FISH, Fig. 4, D and E, fig. S8, and table S8) with three position-specific oligo pools targeting the inversion region, which clearly supported the result in this study, in that the complex inversion was composed of two tandem inversions with breakpoint reuse, rather than a whole inversion as proposed by Maggiolini *et al.* (*4*) (Fig. 4, D and E). This case study indicates that the method we proposed could even accurately clarify such complex, interlinked inversion events. Overall, the complete, accurate list of species-specific inversions between humans and macaques provides a basis for further delineating human-specific inversions.

We then performed phylogenetic analysis to trace the evolutionary trajectories of these species-specific inversions with the genome sequences from nine other nonhuman primate species (Fig. 5A and Materials and Methods). Notably, marmosets and squirrel monkeys were used as outgroups to infer the ancestral state of each inversion. In total, 1240 of these inversions were adequately assigned the lineage information, including 101 human-specific inversions (table S9), 48 macaque-specific inversions, 283 Hominoidea-specific inversions, and 808 Cercopithecidae-specific inversions (Fig. 5A). Inversions with ambiguous ancestral states were excluded from the subsequent analyses. Notably, we covered 11 of the 12 human-specific inversions as previously reported (*4*) and included an additional 90 human-specific inversions. We found that inversions of Hominoidea origins (including human-specific and Hominoidea-specific inversions) were enriched with relatively larger inversions (Fisher's exact test, $P = 0.046$; Fig. 5A). Considering that the fixed inversions with larger sizes are likely to be maintained because of their adaptive functions, if these long inversions are fixed in humans, they may have contributed substantially to recent human evolution.

To further exclude polymorphic inversions not detected in the macaque reference genome but exist in macaque populations, we examined the states of the 101 candidate, human-specific inversions in the genomes of the 562 macaques. Briefly, the number of improperly aligned, paired-end reads in the region of interest was used as an indication of the existence of an inversion. Generally, in the homologous regions of these human-specific inversions in macaques, significantly fewer improperly aligned reads were identified, in comparison to polymorphic inversions with high frequencies in the macaque population (Wilcoxon rank-sum test, $P < 2.2 \times 10^{-16}$; Fig. 5B),

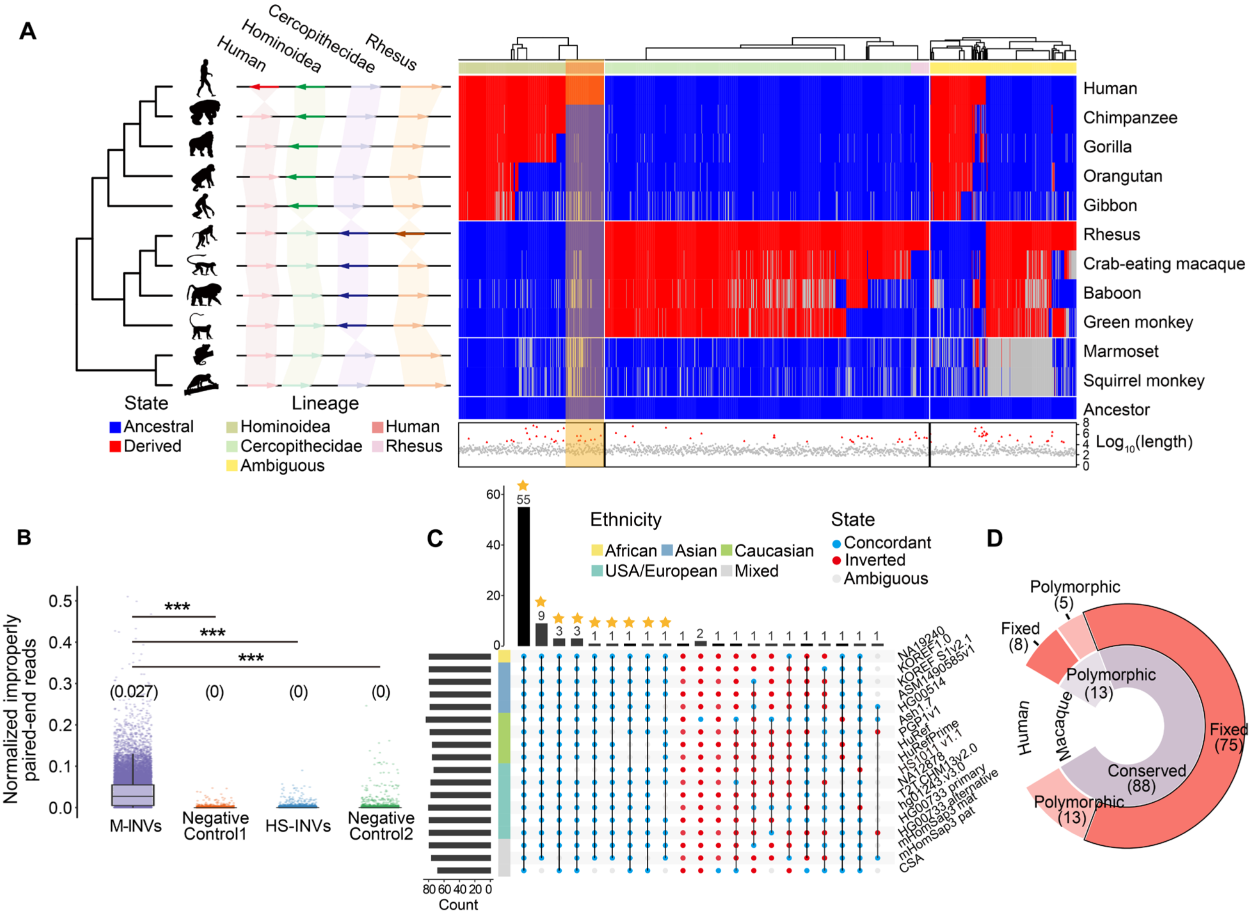

Fig. 5. Identification of human-specific inversions. (**A**) Schematic illustration (left) of four classes of lineage-specific inversions, including human-specific (Human), Hominoidea-specific (Hominoidea), Cercopithecidae-specific (Cercopithecidae), and macaque-specific (Rhesus) inversions. Heatmap (right) showing the arrangement of HM-INVs in comparison to the estimated ancestral states. HM-INVs are ordered in columns, and each row corresponds to a species based on the phylogeny. Blue: the ancestral allele; red: the derived allele; gray: ambiguous state. The hierarchical clustering of HM-INVs and the lineage specificity annotation for HM-INVs are shown at the top. The $\log_{10}$-transformed length for each HM-INV is shown at the bottom, among which the HM-INVs with lengths >95% quantile are indicated with red triangles. (**B**) The distribution of the normalized number of improperly aligned, paired-end reads, for different groups of regions. M-INVs: macaque polymorphic inversions with high frequency; HS-INVs: human-specific inversions; Negative Control1 and Negative Control2: two groups of negative controls of shuffled regions (Materials and Methods). The median read number is shown above each boxplot. (**C**) UpSetR plot showing the number of human-specific inversions that are grouped based on their orientation relative to the reference genome across 18 haplotype assemblies. Blue: alignments concordant with hg38; red: inverted form; gray: ambiguous alignments. Human-specific inversions fixed in the population are highlighted with asterisks. Wilcoxon rank-sum tests, ***$P < 0.001$. (**D**) Donut plot showing the classification of 101 candidate human-specific inversions based on their polymorphic states in human and macaque populations.

indicating that the corresponding regions of most human-specific inversions were not polymorphic in the macaque population. Notably, we identified 13 deep polymorphic inversions that could be detected in both human and macaque populations (table S9 and Materials and Methods), which were removed from the list of candidate human-specific inversions.

Moreover, to further investigate whether the 88 human-specific inversions have been fixed in humans, we examined the state of each inversion in genome assemblies from 15 human individuals of multiple ethnicities (table S10). Three human genomes with diploid assemblies were split into two haplotypes in this analysis. Notably, for 75 of the 88 human-specific inversions, the inverted state could be detected in all of these human genome assemblies (Fig. 5, C and D), which were defined as fixed human-specific inversions (table S9). Together, although we could not fully exclude the possibility that some of these inversions are still not completely fixed, according to their distributions in the 15 human individuals, it is more likely that they have been fixed.

Modulations of the human transcriptome by human-specific inversions

To study the features of these 75 fixed human-specific inversions, we first investigated whether they could introduce human-specific gene regulation through the reconfiguration of ancestral three-dimensional genome architecture. To this end, we first constructed Hi-C libraries from the prefrontal cortex (PFC) tissues of adult humans and macaques, and subsequently generated 1.3 billion valid contact pairs (table S11). The quality of the Hi-C data was validated by distance-dependent interaction frequency decay and the ratio of cis- and trans-interactions (fig. S9A and table S11). Overall, a cross-species Hi-C map was developed to delineate hierarchical chromatin architecture, such as A/B compartments, TADs, and loops (fig. S9), with map resolutions of 2.5 and 5.45 kb for human and macaque, respectively (table S11).

We then compared the chromosome structures between these human-specific inversions and their orthologous regions in macaque, which was used as a proxy for determining their ancestral status, assuming that the chromosome structures of these regions have remained unchanged in the macaque lineage since its divergence from the human lineage. To investigate whether some of these human-specific inversions may be involved in the reconfiguration of gene regulation through chromosomal rearrangement after the divergence of humans and macaques, we identified inversions with coordinate overlapping with specific three-dimensional genome domains, such as compartments, TADs, and loops. Three of these inversions, associated with 47 protein-coding genes, are involved in the switching of TAD structures, and four human-specific inversions in the rewiring of preexisting chromatin loops are putatively involved in the transcriptional regulation of nearby protein-coding genes. Among the four human-specific inversions associated with the chromatin loops, the largest one (chr18_inv1) introduces substantial changes of the three-dimensional genomic architecture by disrupting the ancestral TAD domain and the chromatin loops around the breakpoints, which may further account for the differential expression of the associated genes such as *CLUL1*, *COLEC12*, and *GREB1L* (fig. S10). The other three inversions overlapped with the chromatin loops may have a similar association with the expression of the genes located in these regions.

Second, as previous findings suggest that the regions of inversion tend to accumulate mutations due to recombination suppression (*25*, *26*, *28*), we investigated whether these human-specific inversions could introduce regional mutation hotspots. We found that these regions harbored significantly more divergent sites than their adjacent regions in the human lineage since their divergence from chimpanzees (Wilcoxon signed rank test, $P = 1.3 \times 10^{-8}$ for the upstream regions, $P = 5.6 \times 10^{-9}$ for the downstream regions; Fig. 6A and fig. S11). In contrast, in the orthologous regions of these human-specific inversions in macaques, the numbers of divergent sites were comparable with those in the adjacent regions (Wilcoxon signed rank test, $P = 0.56$ for the upstream regions, $P = 0.68$ for the downstream regions; Fig. 6B). In particular, we found an increased density of genetic divergence on promoter regions of genes located at these human-specific inversions, which may directly contribute to the cross-species differences of gene expression (Fig. 6C).

As the polymorphic data within macaque population indicate that inversions with stronger effects are under stronger purifying selection, the inversions with stronger effects in humans (or not eliminated during the evolution) are thus more likely fixed due to their adaptive functions. Following this strategy, we then prioritized these human-specific inversions according to their sizes, effects on gene regulation, and impacts on the three-dimensional genomic architecture (Fig. 6D). To this end, we ranked these human-specific inversions and classified them into two categories based on the strength of their effects (Fig. 6D and Materials and Methods). Although these human-specific inversions generally showed higher level of divergence than the adjacent regions (Fig. 6A), the inversions with stronger effects showed a subtle increase in divergence, possibly due to a quicker pace of fixation shaped by positive selection (Wilcoxon rank-sum test, $P = 0.011$; Fig. 6E).

Notably, for the fixed human-specific inversions with stronger effects, the resultant changes in transcriptome might be correlated with the direction of human adaptive evolution. The genes located around the inversions with stronger effects showed a greater degree of differential expression in fetal brain between humans and macaques, than genes associated with inversions with weaker effects (one-tailed Wilcoxon rank-sum test, $P = 0.028$; Materials and Methods), or the genome-wide human-macaque ortholog pairs as a background (one-tailed Wilcoxon rank-sum test, $P = 0.021$), indicating the adaptive roles of these inversions in early brain development (Fig. 6F and fig. S12). Our results thus provide a strategy for prioritizing human-specific inversions with adaptive functions in human evolution, especially in human brain development.

Contribution of a human-specific inversion to human uniqueness in brain development

Among these 75 human-specific inversions, we focused on the inversion with the highest rank score (Fig. 6D and table S12). This inversion was found to cover 13 Mb of chromosome 18 and included 62 protein-coding genes (Fig. 6D and fig. S10). This human-specific inversion transferred the outward telomeric segment of the ancestral genome to the vicinity of the neocentromere in humans, and shifted the inward segment to the telomeric side. It disrupted the TAD structure near the inward breakpoint, and led to a loss of contact with previously adjacent region (figs. S10 and S13). Newly formed interactions accompanying the inversion event were not detected, possibly due to the formation of a neocentromere in the human genome (figs. S10 and S13). Moreover, this inversion

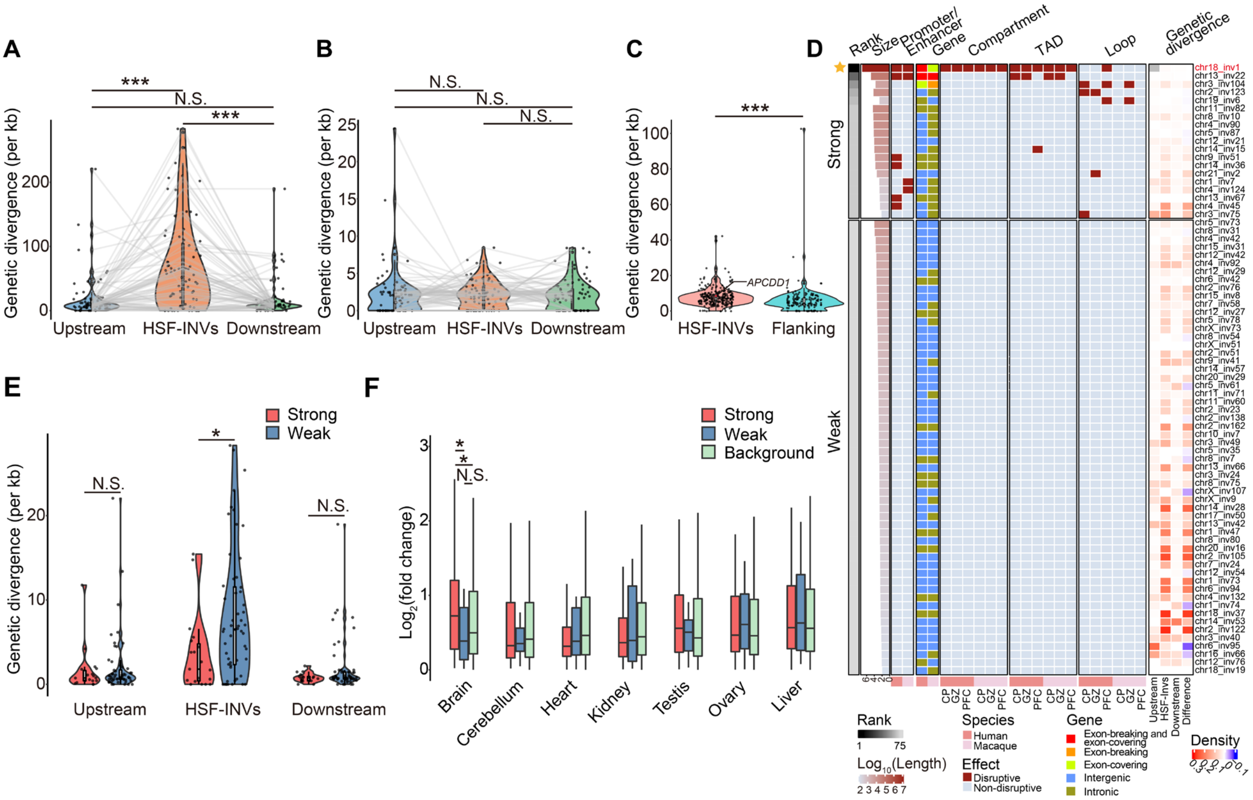

Fig. 6. Characteristics of human-specific inversions. (**A**) Violin plots showing the genetic divergences of fixed human-specific inversions (HSF-INVs) and their length-matched, upstream and downstream genomic regions (Upstream and Downstream regions). Wilcoxon signed rank tests were performed. Wilcoxon signed rank tests, $***P < 0.001$. (**B**) Violin plots showing the genetic divergences of homologous macaque regions of the HSF-INVs and their upstream and downstream regions. Wilcoxon signed rank tests; N.S., not significant. (**C**) Violin plots showing the genetic divergences of promoter regions in fixed human-specific inversions (HSF-INVs) and promoter regions in length-matched flanking genomic regions (Flanking regions). The red dot indicates promoter of *APCDD1*. Wilcoxon rank-sum test, $***P < 0.001$. (**D**) Classification of 75 fixed human-specific inversions into two groups with different degree of regulatory effects (Strong and Weak), based on their sizes and locations in the human and macaque genomes. The genetic divergence relative to the human-chimpanzee common ancestor is also shown for the 75 HSF-INVs and corresponding upstream and downstream regions. The difference in the genetic divergence between each inversion and the average of its upstream and downstream regions is also shown (Difference). (**E**) Violin plots showing the genetic divergence for inversions with strong (Strong) or weak (Weak) effects, and their upstream and downstream regions. One-sided, Wilcoxon rank-sum tests, $*P < 0.05$, N.S., not significant. (**F**) The $\log_2$-transformed fold changes in gene expression in the fetal brain between humans and macaques, for genes located on inversions with strong (Strong) or weak (Weak) effects, as well as for genome-wide orthologs as a background (Background). Wilcoxon rank-sum tests, $*P < 0.05$, N.S., not significant.

introduces mutation hotspots across the inversion region, which harbors more divergent sites than the adjacent region (density of divergent sites: 0.0073 versus 0.0039 per bp).

The reconfiguration of the genomic architecture and the introduction of the regional mutation hotspot should have contributed substantially to the human-specific changes of the transcriptome in this region. For 51 of the 62 human genes located on this inversion, their orthologs in rhesus macaque and mice could be accurately defined. Of the 51 genes, 33 were differentially expressed in brain between humans and macaques, according to the RNA-seq data with corresponding developmental stages across the three species. Of the 33 genes, 23 showed the same trends of expression changes, when comparing their expressions in humans with those in macaques or mice, indicating the consistent regulation of this human-specific inversion on the expression of genes located on it (fig. S14). We then manually curated the functions of these genes from literatures and database annotations, and found that a large portion of them (26 of 33) are involved in key biological processes (e.g., cell proliferation, embryo development, and neural development) associated with human recent evolution, such as *APCDD1* involved in the regulation of neuronal cell fates (*30*, *31*), *TWSG1* in embryogenesis, neural progenitor differentiation and neuronal repair (*32*), as well as *AFG3L2* and *EPB41L3* in the regulation of neuron projection morphogenesis (*33*). It is thus plausible that these differentially expressed genes, possibly regulated by this human-specific inversion, could have contributed substantially to the human-specific features in development.

As a proof of concept, we focused on *APCDD1*, a gene located on this inversion and showed significantly decreased expression in human brains in comparison to macaques and mice (Wilcoxon rank-sum test, $P = 1.6 \times 10^{-3}$; Fig. 7A). Consistent with the findings that inversions could introduce regional mutation hotspots, we found an increased density of genetic divergence on promoter regions of *APCDD1* in humans since the divergence from chimpanzees (Fig. 6C). We then designed dual-luciferase reporter assays to quantify the transcriptional activity of the *APCDD1* promoters in human and macaque, and found that the transcriptional activity of the

APCDD1 promoter sequence in humans is significantly lower than that in macaques (Fig. 7B, Student's *t* test, $P < 1.0 \times 10^{-4}$). Next, we investigated whether the divergent sites accumulated during the evolution of the human lineage contributed directly to the changed activity. To this end, we identified five mutations that occurred specifically in humans after the divergence of human and chimpanzee, which are located on *APCDD1* promoter regions and marked by H3K4me3 signals. When these sites were mutated to their ancestral alleles, three of them could significantly enhance the transcriptional activity of the promoter (Fig. 7B, Student's *t* test, Mutation-1: $P < 1.0 \times 10^{-4}$, Mutation-3: $P = 1.0 \times 10^{-4}$, Mutation-5: $P = 1.6 \times 10^{-3}$). These findings thus support the idea that the accumulated mutations on *APCDD1* promoter, a process accelerated by the formation of this human-specific inversion, contributed to the decreased *APCDD1* expression in humans.

To investigate whether the decreased expression of *APCDD1* is involved in the adaptive evolution of human brain development (Fig. 7C), we first developed a lentiviral-based overexpression assay in human neural progenitor cells (NPCs) to examine the effects of *APCDD1* overexpression on the proliferation and differentiation of neural progenitors (Fig. 7, D and E). Notably, we found a significant reduction of SOX2^{+} progenitor populations [two-way analysis of variance (ANOVA) and Tukey's post hoc test, $P = 1.1 \times 10^{-3}$] and an expansion of SOX2^{-}/TUJ1^{+} neuron populations (two-way ANOVA and Tukey's post hoc test, $P = 9.9 \times 10^{-4}$; Fig. 7, D and E), indicating an accelerated neuronal maturation process in human NPCs with *APCDD1* overexpression, a pattern recapitulating the differences in early brain development between humans and macaques (*34*).

Considering that the varied pace of neuronal maturation has been implicated in different cognitive skills in primates (*35*), and the regulatory roles of *Apcdd1* in neuronal cell fates has been documented (*30*), we further constructed *Apcdd1*$^{+/-}$ mice to investigate whether the partial depletion of *Apcdd1* could contribute directly to the alteration of cell composition in the mouse brain, and further to brain development and functions (Fig. 7, C and F to J, and fig. S15). Specifically, we detected the highest expression of *Apcdd1* at E10.5 mouse brain across different developmental stages (fig. S16). Moreover, the cortex undergoes symmetrical division to expand neural progenitor pool at E10.5, and the early neurogenesis of central nervous system development started before this stage (*36–38*). We thus focused on E10.5 to investigate the potential roles of *Apcdd1* in mouse early neurogenesis. To this end, we performed single-cell RNA sequencing (scRNA-seq) in the brains of wild-type and *Apcdd1*$^{+/-}$ mice at E10.5 and obtained a transcriptional profile of a total of 25,074 cells (Materials and Methods). The scRNA-seq data from each genotype were then integrated to determine the cell type diversity in mouse brains. Cell-type identities were assigned based on known marker genes as reported in literatures (*39–43*), which resulted in nine major classes including radial glia, neuroblast, optic cup, neural crest, surface ectoderm, vascular, blood, immune, and mesenchyme (Fig. 7F and fig. S17). Notably, we observed a significantly decreased proportion of neuroblasts, representing a population of immature neurons, in *Apcdd1*$^{+/-}$ mice (Wald test, $P = 3.9 \times 10^{-5}$; Fig. 7G and Materials and Methods), suggesting the involvement of *Apcdd1* in the regulation of neurogenesis. However, although a trend of increased proportion of radial glial cells could be detected in *Apcdd1*$^{+/-}$ mice, the difference is not significant (Wald test, $P = 0.1$; Materials and Methods). As the radial glial cell is the dominant cell type in E10.5 mouse brains, it is possible that the detection of statistical difference is difficult with the small sample size. Moreover, although we focused on E10.5 mouse brains due to the highest expression of *Apcdd1* at that stage (fig. S16), it is possible that the pattern of cell compositional shift could lag behind the expressions of their key regulators. Together, although single-cell transcriptome studies in additional development stages are required in further studies to fully clarify the regulatory mechanism of *Apcdd1* in mice brain development, the data on E10.5 mouse brains recapitulate the findings in human NPCs for the involvement of *APCDD1* in the regulation of neurogenesis.

To further investigate whether the varied neuronal maturation could induce behavioral change in adult mice, we performed the Morris water maze test to analyze the spatial learning, working memory, and reversal learning abilities of the *Apcdd1*$^{+/-}$ and wild-type mice at 9 to 11 weeks (Fig. 7, H to J). The *Apcdd1*$^{+/-}$ mice showed a comparable swimming ability and vision (Student's *t* test, $P = 0.40$ for acquisition; $P = 0.44$ for reversal phase; fig. S18) and similar performance in the acquisition and reversal phases of learning to the wild-type mice (two-way ANOVA test, $P = 0.83$ for acquisition; $P = 0.47$ for reversal phase; Fig. 7H). However, the *Apcdd1*$^{+/-}$ mice showed a significantly increased ability in the probe trials after the reversal learning, during which the proportion of time they spent in the target quadrant was significantly increased (Student's *t* test, $P = 8.6 \times 10^{-3}$; Fig. 7I). These findings jointly indicate the regulatory effect of *APCDD1*, possibly through a human-specific inversion, on neuronal maturation and subsequent enhanced cognitive skills that are distinct in human.

DISCUSSION

Here, to improve the variant calling in macaque populations, we first attempted to construct a better reference macaque genome, through the integration of three recent genome assemblies based on third-generation sequencing. Notably, the strategy of gap filling by aligning Mmul_10 to the other genome assemblies may introduce assembly errors of other assemblies into Mmul_10. To exclude this possibility, we used a conservative strategy in this step, in which only gaps with continuous sequences spanning the gap regions were filled (Materials and Methods). Moreover, we further sequenced the genome of eight additional macaques with PacBio long HiFi reads, and assessed these regions of gap filling with these HiFi reads. According to the alignments of these long reads on the Mmul_10 genome assembly with gap closures, we found that all of the 40 gap closures could be validated by these long reads (fig. S2), indicating that these gaps should represent real genomic gap regions of Mmul_10, rather than the assembly errors from other macaque genome assemblies, or individual-specific structural variations.

SVs, with larger sizes and expected stronger effects than SNVs, have been reported as one of the major sources of genomic innovations. However, it is error-prone to use standardized algorithms to call SVs with short reads. Notably, despite the inherent advantage of long-read sequencing in pinpointing SVs in a straightforward manner, the cost of this strategy limits its application at a large population scale, and the SVs identified by long-read data in a small number of macaques could only provide limited vision for the complete SV atlas in a macaque population, hindering the in-depth investigations of their global features and evolution. A systematic workflow for reliable SV identification with abundant short-read sequencing data is thus urgently needed. Here, we used an optimized

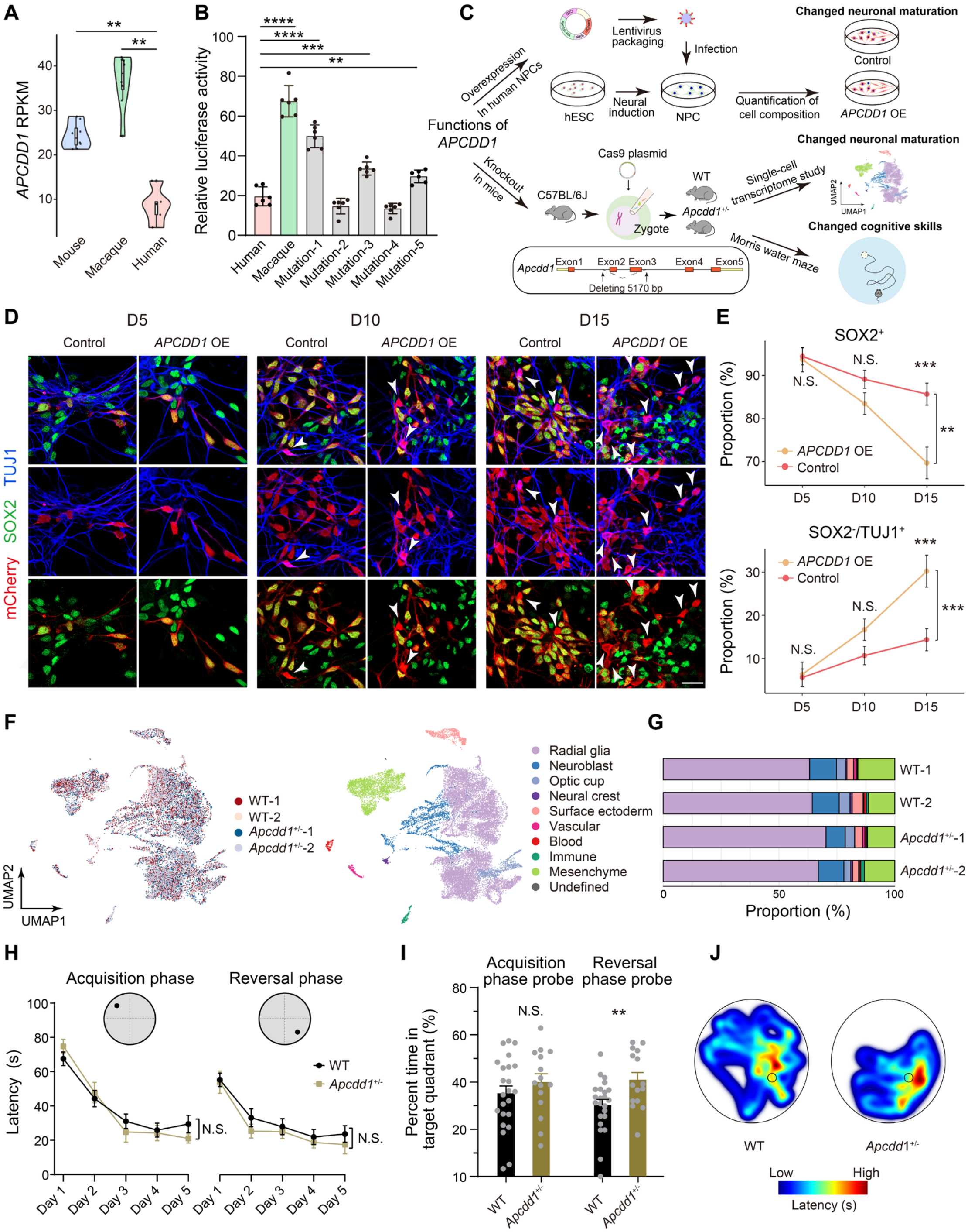

Fig. 7. A human-specific inversion contributes to human uniqueness in brain development. (**A**) Violin plots showing the expressions of *APCDD1* in the brains of humans, macaques, and mice at the mid- to late-fetal stages. Wilcoxon rank-sum tests, $**P < 0.01$. (**B**) Relative luciferase activities of human *APCDD1* promoter (Human), macaque *APCDD1* promoter (Macaque), and human *APCDD1* promoter with mutations (Mutation-1 to Mutation-5). Student's *t* test, $**P < 0.01$, $***P < 0.001$, $****P < 0.0001$. (**C**) The design of experiments for *APCDD1* functions. (**D**) Representative immunostaining of SOX2 (progenitors), TUJ1 (postmitotic neurons), and mCherry (virus-infected cells) in the assays of wild-type NPCs (Control) and NPCs with *APCDD1* overexpression (*APCDD1* OE), at different protocol days (D5, D10, and D15) after lentiviral infection. White arrowheads, $SOX2^-/TUJ1^+$ cells. Scale bars, 20 μm. (**E**) Proportions of $SOX2^+$ progenitors and $SOX2^-/TUJ1^+$ neurons in Control and *APCDD1* OE at different protocol days (D5, D10, and D15). Two-way ANOVA, $**P < 0.01$, $***P < 0.001$. (**F**) UMAP plots of scRNA-seq from brains of wild-type (WT-1 and WT-2) and $Apcdd1^{+/-}$ mice ($Apcdd1^{+/-}$-1 and $Apcdd1^{+/-}$-2) at E10.5, grouped by genotypes (left) or cell types (right). (**G**) Proportion of each cell type in brains of WT-1, WT-2, $Apcdd1^{+/-}$-1, and $Apcdd1^{+/-}$-2. (**H**) The learning curves of the wild-type (WT, $n = 23$) and $Apcdd1^{+/-}$ ($n = 15$) mice in the acquisition and reversal phases. Two-way ANOVA. N.S., not significant. (**I**) Time spent in the target quadrant by WT and $Apcdd1^{+/-}$ mice in the probe trials after the acquisition and reversal phases. Two-sided Student's *t* test, $**P < 0.01$, N.S., not significant. (**J**) Representative plots of escape latency for WT and $Apcdd1^{+/-}$ mice in the probe trials after the reversal phase of learning. Data are represented as the means ± SEMs.

pipeline for accurate SV detection with short reads, in which different SV tools were integrated and cross-validated to reduce false positives. With this pipeline, we present the largest macaque SV landscape to date, based on the deep integration of WGS data of 562 macaques. The high verification rate as indicated by the evaluations with public SV benchmarks, an in-house benchmark of eight macaque genomes with long-read sequencing, and another benchmark of one macaque genome with long-read sequencing and whole genome assembly jointly verified the efficiency of our pipeline in accurately calling SVs in macaque populations. This accurate and quantitative SV map thus provides a basis for clarifying the features, turnover, and evolutionary significance of SVs in primates.

According to the macaque SV map, we found that the number of SVs identified from wild-caught Chinese-origin macaques is significantly lower than the other two populations. Notably, the sequencing depth of wild-caught Chinese-origin macaques (median: 248.2 million reads) was significantly lower than that of the other two macaque populations (median: 700.8 and 893.3 million reads for captive Chinese-origin macaques and captive Indian-origin macaques, respectively; fig. S19). As we introduced stringent filters to control for false positives, it is thus possible that the relatively lower sequencing depths contributed to the lower numbers of SVs identified in wild-caught Chinese-origin macaques. As expected, when using a subset of wild-caught Chinese-origin macaques with comparable sequencing depths with those of captive Chinese-origin macaques, we observed comparable numbers of SVs identified in the two populations (Wilcoxon rank-sum test, $P = 0.32$ for deletions; $P = 0.98$ for duplications; $P = 0.46$ for inversions). Notably, for each SV, the macaques without enough read coverage were not considered in the estimation of the derived allele frequency of the SV event. The varied sequencing depths thus would not introduce biased calculations.

Although it is difficult to accurately identify inversions with short reads, according to our evaluations with the three-tier benchmarks, our pipeline actually performs well in calling inversions, with the verification rate comparable with deletions. Moreover, we found stronger signals of purifying selection on the fixation of inversions than other types of SVs (fig. S20), consistent with previous reports that polymorphic inversions are largely deleterious due to recombination suppression and the subsequent accumulation of deleterious mutations. Besides this finding, we also found that the derived allele frequency spectrum of duplications was more left-skewed than that of deletions (fig. S20). However, such a profile is likely a result of biased identification of duplications with lower derived allele frequency, rather than an indication for stronger selective constrains. Briefly, considering the fact that the identification of duplications using short reads is more error-prone than that of deletions and inversions, we introduced more stringent filtering steps to control for false positives, especially for those located in repetitive genomic regions. It is thus plausible that a larger portion of neutral duplication events with a relatively higher level of derived allele frequency, enriched in repetitive genomic regions, were removed from the following analyses. Together, considering the strength of purifying selection and the robustness of identification for different types of SVs, we focused on inversions in this study.

On the basis of the macaque SV map and a comparative genomics study with multiple outgroup species, we identified a list of 75 human-specific inversions apparently fixed in humans. Moreover, using our quantitative SV map in macaque population, we found that the inversions located in regulatory regions, such as genic regions and TAD boundaries, are subjected to enhanced purifying selection, suggesting a practical strategy for prioritizing these human-specific SVs with the most important functions in shaping human adaptive evolution. Accordingly, when we classified these human-specific inversions into two categories on the basis of the strength of their effects, we found that the genes located in inversions with stronger effects showed a higher degree of differential expression in fetal brains between humans and macaques, which was well correlated with the direction of human adaptive evolution in brain development. These fixed, human-specific inversions with stronger effects should thus have substantially shaped the human brain transcriptome during human evolution, through their dual effects of reconfiguring ancestral genomic architecture and introducing regional mutation hotspots. As a special note, in a recent study, Zhou *et al.* focused on insertions and deletions recently evolved in great apes, and identified genes related to these lineage-specific SVs. They subsequently found the association of these genes with brain development and functions (*44*). The two studies focusing on different types of SVs thus jointly highlight the functional significance of young SVs in shaping complex, lineage-specific traits during the human brain evolution.

Our findings suggest that the accumulated mutations on *APCDD1* promoter, a process accelerated by the formation of this human-specific inversion, should have contributed to the decreased expression in humans. However, the reconfiguration of the ancestral chromosome structure may also be involved in this regulation. Notably, along with the emergence of this human-specific inversion, the locus of *APCDD1* was moved to a region near the newly formed centromere in human. Further investigations are thus needed to justify the effects of changed chromosome structure on *APCDD1* expression during the human evolution. Moreover, for the other differentially expressed genes located on this inversion, similar studies are needed to directly connect their changed expressions with this inversion.

The contribution of regulatory changes to human evolution has been proposed to supplement the contribution of sequence alterations in coding regions. The proof-of-concept study of *APCDD1* regulation indicates that the human-specific down-regulation of even one single gene, presumably regulated by a human-specific inversion, could substantially modulate the process of neuronal maturation and subsequently improve cognitive ability, recapitulating the unique features of human brain development. The human-specific inversions identified here could thus act as a previously neglected genetic source underlying the uniqueness of human brain development.

MATERIALS AND METHODS

Ethics statement

The animal samples used in this study were approved by the Animal Care and Use Committee of Peking University (IMM-LiCY-1). The human PFC sample in this study was approved by the IRB of Beijing Tiantan Hospital, Capital Medical University (KY2017–035-02).

Sample collection

Human brain tissue

Human brain tissue was obtained from Beijing Tiantan Hospital. All patients in this study provided written informed consent for sample collection and data analysis.

Rhesus macaque brain tissue

Frozen brain tissues of 26 rhesus macaques were obtained from the animal facility at the Institute of Molecular Medicine, Peking University (accredited by the Association for Assessment and Accreditation of Laboratory Animal Care). Detailed information for these macaques is provided in table S4.

HeLa-S3 cell line

The HeLa-S3 cell line was obtained from the laboratory of Y. Sun, Peking University. HeLa-S3 cells were grown in 10-mm glass-bottom imaging dishes (Cellvis, #D35–10-1-N) with 2 ml of modified medium [high-glucose Dulbecco' s Modified Eagle Medium (DMEM), Thermo Fisher Scientific Gibco, #10569-044] supplemented with 10% (v/v) fetal bovine serum (Thermo Fisher Scientific Gibco, #10091-148) and penicillin-streptomycin antibiotics (100 U/ml; Thermo Fisher Scientific Gibco, #15070-063) under regular cell culture conditions (37°C, 5% CO_2, humidified atmosphere). The cells were passaged at a proportion of 1:8 to 1:10 with trypsin (Life Technologies, #25200-056) every 3 days or when they reached 80% confluence.

LLC-MK2 cell line

Macaque Lilly Laboratories Cell-Monkey Kidney 2 (LLC-MK2) cells were obtained from the National Collection of Authenticated Cell Cultures (CSTR: 19375.09.3101MONGNO6). The LLC-MK2 cells were maintained in vitro in RPMI 1640 (Thermo Fisher Scientific Gibco, #C11875500BT) supplemented with 10% (v/v) fetal bovine serum (Thermo Fisher Scientific Gibco, #10091-148) and penicillin-streptomycin antibiotics (100 U/ml; Thermo Fisher Scientific Gibco, #15070–063) under cell culture conditions (37°C, 5% CO_2, humidified atmosphere). The cells were passaged at a proportion of 1:3 to 1:4 every 3 days.

Deep sequencing

Genomic DNA was extracted from frozen brain samples using Ultra Pure Phenol:Chloroform:Isoamyl Alcohol (Invitrogen, 25:24:1, v/v). For PacBio sequencing, SMRTbell genomic libraries (>20 kb in length) were constructed from high-quality DNA extracted from the brain and muscle tissues of one male macaque, and deep sequencing was performed on the RS II and Sequel platforms with the P6-C4 sequencing reagent. For whole-genome sequencing with short reads, libraries were constructed from DNA extracted from brain or blood tissues of 27 macaques, and deep sequencing was performed on MGISEQ-2000 (MGI) sequencing systems to generate 150-bp, paired-end reads.

Hi-C library preparation

Frozen PFC samples (0.2 g) from a human and a macaque were ground in liquid nitrogen, after which cell suspensions were prepared. The filtered cell suspensions were fixed in 1% formaldehyde for 30 min. The cross-linked DNA was digested with 200 U MboI (NEB, #R0147M), and the digested fragment ends were filled with biotin-14-dATP (Invitrogen, #19524016), dCTP, dGTP, and dTTP by DNA Polymerase I Klenow Fragment (NEB, #M0210L). The resulting blunt-end fragments were religated by T4 DNA ligase (Enzymatics, #L603-HC-L). Then, the ligated cross-linked DNA was reversed using proteinase K (TIANGEN, #RT403). DNA purification was performed by removing biotin from unligated ends using T4 DNA polymerase (Enzymatics, #P7080L). The purified DNA was then sheared to a length of 350 to 500 bp using Covaris M220, and biotin-labeled DNA was pulled down with Dynabeads M-280 Streptavidin (Invitrogen, #11205D). Library preparation was performed using a DNA library preparation kit (Vazyme, #ND607). The libraries were then sequenced on the Illumina NovaSeq 6000 sequencing platform to generate 150-bp paired-end reads.

Gap closure

Gap closure was performed according to the genome assembly of Mmul_10. Genome-wide pairwise alignment was performed using lastz (Parameters: --notransition --step = 20 --ambiguous = n --chain) (*45*) between Mmul_10 and two other reference genomes, rheMac8 (Indian origin) and rheMacS (Chinese origin). A gap was closed when two conditions were met: (i) there was a continuous sequence spanning the gap region in alignment with rheMac8 or rheMacS, and (ii) the flanking anchor sequences along the gap regions were longer than the gaps. The sequences spanning the gap regions were extracted as padding sequences. If a gap in Mmul_10 could be filled based on both of the genome assemblies, the rheMacS genome was preferentially used, considering its higher genome continuity.

To evaluate the quality of gap closure, a Bionano optical map was generated for one macaque (monkey ID: blood-270-zhongxm_HKNGYCCXX_L1). Briefly, to create a Bionano optical map, DNA was extracted from the liver of the same macaque used for PacBio sequencing and digested with the nicking enzyme NT.BspQI. Standard library preparation and optical scanning were then performed. Bionano Solve software was used to assemble the optical map from scratch to obtain scaffolds (version: 3.1) and then the script fa2cmap_multi_color.pl was used to obtain the optical map scaffold with enzyme digestion (enzyme name: BspQ). OMTools (version: 1.4a) (*46*) was used for read mapping and visualization. Only gap sequences with at least one mapped restriction enzyme cleavage site or with a consistent order of restriction enzyme cleavage sites around the gap were retained.

Ten macaque samples with an average sequencing depth of 36-fold (five Chinese-origin macaques and five Indian-origin macaques) were selected to evaluate the quality of gap closure. The whole-genome sequencing data of these samples were aligned to the improved genome assembly with 44 gaps filled using BWA-MEM (*47*) with the default parameters (version: 0.7.17-r1188). Then, the read coverages of the gap regions and their length-matched, upstream or downstream regions were calculated and compared using bamdst with the default parameters (version: 1.0.9).

Another eight Chinese-origin macaques with PacBio long-read sequencing data were further selected to assess these regions of gap closure (table S3). The long reads of these samples were aligned to the improved genome assembly with 40 gaps filled, using pbmm2 with default parameters (version: 1.8.0). The alignments of these regions were then illustrated using IGV (*48*) with default parameters (version: 2.16.0).

Identification and annotation of SNVs in the macaque population

The whole-genome sequencing data of rhesus macaques were aligned to the genome assembly of rheMac10Plus with BWA-MEM using the default parameters (version: 0.7.17-r1188). PCR duplicates were marked for the merged BAM files using GATK MarkDuplicates (version: 4.2.2.0), and only nonduplicate reads were used for the downstream analyses.

The best-practice workflows of the Genome Analysis Toolkit (GATK, version: 4.2.2.0) were used to call SNVs. Indels were realigned with IndelRealignment, and base quality was recalibrated with BaseRecalibrator. The SNVs on each chromosome in each sample were called with HaplotypeCaller. The known SNV datasets were obtained from recent macaque population studies (*16*, *19*), which were used to evaluate the variants with VariantRecalibrator, with a threshold of sensitivity above 99%. To ensure the accuracy of variants, the evaluated variants were then used as the new "known SNV dataset" to call and evaluate the SNVs as illustrated in the above workflow. The PLINK tool (version: v1.90b6.16 64-bit) (*49*, *50*) was further used to filter the samples with frequencies of missing calls greater than 10% and low-quality SNVs with the following exclusion criteria: (i) variants with frequencies of missing calls greater than 10% and (ii) variants with Hardy-Weinberg equilibrium exact test *P* values below 0.001. The SnpEff tool (version: 5.0e) (*51*) was then used to annotate these SNVs.

Kinship analyses

The relationships between individual macaques were estimated using KING software (version: 2.2.5) (*52*). The kinship coefficients for all pairs of macaques were estimated and assigned to one of the following categories: twins, parent-offspring, siblings, second-degree relatives, or third-degree relatives, using the recommended kinship coefficient boundaries (*52*). By considering the family trees, we filtered coefficients for pairs of relatives inferred to be second-degree relatives or closer to identify a list of independent macaque animals (*53*).

Population genetic analyses

For both human and macaque populations, the nucleotide diversity (π) was calculated for genome-wide sliding windows (3 Mb for each window and 3 Mb for each step) using VCFtools (version: 0.1.17) (*54*).

Genetic diversity and population structure analyses

A neighbor-joining phylogeny for 144 Chinese-origin macaques was constructed based on the P distance matrix as calculated by VCF2Dis (version: 1.47), according to the autosomal genetic variations. The FastME tool (version: 2.0) (*55*) was used to visualize the phylogenetic tree. The subpopulation information was assigned for animals with an unknown geographic distribution on the basis of the clustering results and the genetic distance. Subpopulations of *M. m. mulatta* and *M. m. lasiotis* were combined due to their admixture structures. PCA of SNVs on all of the autosomes in the macaque population was performed using GCTA (version: 1.93.0) (*56*). A PCA scatterplot for principal components 1–3 was then created using the plotly package in R (version: 4.1.2). The fractions of the variance explained by the three components were calculated according to the Kaiser-Guttman criterion and the broken-stick model.

Identification of SVs in the macaque population

PCR duplicates were marked for the raw mapped reads using Picard (version: 2.25.4) and then SVs of each animal were identified using DELLY (version: 0.8.2) (*57*) and SpeedSeq (version: 0.1.2) (*58*). Multiple filters and intersection strategies were used to obtain an accurate set based on the evaluation of macaques with genome assemblies. Specifically, only SVs with "QUAL = PASS" were retained from the DELLY results. Among all SV calls from DELLY and SpeedSeq, only SVs with lengths of at least 50 bp on autosomal and X chromosomes were retained. For the results of CNVnator, thresholds were set as follows: (i) both E-val 1 and E-val 2 were less than 0.05; (ii) q0 was less than 0.5 and not equal to −1.

Inversions, deletions, and duplications were then identified separately. Briefly, for candidate inversions, only SpeedSeq calls supported by at least three paired-end reads and three split reads were retained. In addition, only inversions with a consensus region shared by DELLY and SpeedSeq calls (with a coordinate overlap of at least 90%) were retained. For candidate deletions, only DELLY calls supported by at least three paired-end reads and three split reads were retained. Moreover, only deletions with a consensus region shared by DELLY and SpeedSeq calls (with a coordinate overlap of at least 50%) were retained. Notably, we removed the deletion calls showing at least 20% overlap with gap regions in the macaque reference genome and retained only candidate deletions showing significantly lower read depths in the deletion regions in contrast to their adjacent regions. For candidate duplications, only DELLY calls supported by at least one split read were retained. Moreover, only duplications with a consensus region shared by DELLY and SpeedSeq calls (with a coordinate overlap of at least 50%) were retained. Candidate duplications located in repeat regions were retained in the following analyses only when they were also identified by the CNVnator tool (with a coordinate overlap of at least 80%).

For each SV type, the SV calls from the macaque population were then merged into a single callset with the svtools tool (version: 0.0.1) (*59*), which was used as a reference to genotype each macaque animal. Considering the limited resolution in identifying breakpoints, we clustered and merged the intersecting SV calls to obtain consensus SV hotspots with BEDtools merge. Overlapped SV calls were merged into a single interval. Per-sample quality control was then performed according to these SV hotspots. Notably, we removed 10 macaque samples in which the SV counts were >10 median absolute deviations from the median count of the corresponding SV type. Allele frequencies were then recalculated based on the SV hotspots. The genotypes were assigned as "missing" for individual samples if there were ambiguous genotypes on this SV. Samplot (version: 1.3.0) (*60*) was used to visualize the SV calls.

The SV calls in individual macaques were validated with both long-read sequencing data and an intensive manual check. Vapor (version: 1.0) (*61*) was used to validate the SV calls from the macaque with long sequencing reads. The ambiguous calls were visualized with the SV-plaudit tool (version: 2.0.0) (*62*) for a manual check.

Development of a three-tier benchmark

We first evaluated the performance of our pipeline in SV detection with public benchmarks. Briefly, we selected the HG002 genome with verified deletion calls as the benchmark callset (*21*) and evaluated the performance of our methods in calling these deletion events, by processing the short-read sequencing data of this genome (SRR12898337, data generated by Illumina HiSeq X Ten with a coverage of 42-fold) and benchmarking using truvari tool provided by the GIAB (*63*). We selected the SV calls with SVTYPE = DEL and performed truvari bench with default parameters (−-pctsim = 0: to ignore sequence comparison, −-includebed HG002_SVs_Tier1_v0.6.bed file: to benchmark in high-confidence regions, −-passonly: to filter out the highest confidence set of SVs). As this callset does not include inversion events, we then used another benchmark callset from Pendleton *et al.* (*22*) to evaluate the performance of our pipeline in calling inversions, by processing the short-read sequencing data of the HG001 genome (SRR8454588, data generated by Illumina HiSeq 4000 with a coverage of 30 folds) and benchmarking using BEDtools tool.

Second, from the list of 562 macaques, we selected 8 macaques and sequenced their genomes with different coverages of long HiFi reads (table S3). For each macaque animal, we then evaluated the performance of our pipeline in calling deletions, duplications, and inversions, using SVs called in this macaque by long reads as a benchmark. As the sequencing of these samples was not saturated, for each SV type in each macaque animal, we performed a simulation step to estimate the theoretical number of verified SVs at the current sequencing depth, assuming the SVs and their genotypes were accurately identified with short reads (fig. S21). Specifically, for each macaque animal, we first calculated the number of SVs identified by short reads and covered by at least one long read. For each of these SVs, the profile of long reads covering the SV region was then simulated, according to the coverage of the long reads across this region and the genotypes of the SV (homozygous and the heterozygous) as defined by the short-read sequencing. This process was repeated 10,000 times, and 10,000 sets of long reads covering these SV regions were obtained. For each round of simulation, the number of verified SVs was then counted, 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. The verification rate was then calculated by dividing the real number of verified SVs by the theoretical number of verifiable SVs from 10,000 simulations.

Finally, from the population of 562 macaques, we selected the macaque with short-read sequencing data and the Bionano optical map generation (as mentioned in the "Gap closure" section), and further sequenced its genome with high-coverage, long-read sequencing. We further de novo assembled its genome on the basis of the integration of the short-read sequencing, long-read sequencing, and Bionano optical data. Briefly, long-read de novo genome assembly was performed with FALCON software. Contigs were further polished with Arrow using long-read sequencing data and Pilon (*64*) (version 1.2) using short-read sequencing data. De novo genome assembly with Bionano optical map data was performed by BioNano_Solve (version Solve3.1_08232017) with manufacturer-recommended parameters. Hybrid scaffolding of the polished PacBio contigs and Bionano optical map were performed by the hybrid scaffolding module packaged with the BioNano_Solve software. The genome assembly of rheMac8 assembly and the genomic markers were then used to generate chromosome-level assembly by ordering the hybrid scaffolds and contigs. Finally, the ordered hybrid scaffolds and contigs were linked together by filling with 3-Mbp "N" sequences between adjacent scaffolds or contigs. The long reads and the de novo assembled genome were then used as the benchmark to evaluate the performance of our pipeline in SV calling with short reads.

PCA of SVs

PCA was performed with the R command prcomp based on the genotype information for 562 unrelated animals, in which the macaques were classified into three groups: captive Indian-origin macaques, captive Chinese-origin macaques, and wild-caught Chinese-origin macaques.

Hi-C data analyses

For the Hi-C data of the adult PFC and fetal brain (CP and GZ zones) tissues of humans and macaques, we first performed preprocessing with HiC-Pro (version: 2.11.1) (*65*). Paired-end reads in FASTQ format were first aligned against the reference genome (hg38 for human and rheMac10Plus for macaque) and were subjected to rigorous filtering criteria to generate valid contact pairs. We then calculated the resolution of the maps as described in Rao *et al.* (*66*). The valid pairs were then binned to contact maps at different resolutions, which were subjected to ICE (Iterative Correction and Eigenvector Decomposition) normalization. The Juicer toolkit (version: 1.9.9) (*67*) was used to convert the contact maps to .hic files for visualization in the UCSC Genome Browser. Knight-Ruiz (KR) normalization is used for configuring Hi-C tracks. The analysis of distance-dependent contact strength decay was performed using FAN-C1 (version: 0.9.21) (*68*). We then detected basic chromosome organizations in a hierarchical manner. First, Cooltools (version: 0.3.2) (*45*) was used to identify A/B compartments at a 100-kb resolution, with the genome divided into two distinct groups based on the sign of the first principal component (PC1). The group with higher gene density was defined as compartment A, and the other was compartment B. Next, TAD boundaries were detected using the insulation score method at a resolution of 40 kb. Consequently, TAD bodies were defined as genomic regions between two adjacent boundaries. Finally, chromatin loops were called by HiCCUPS (version: 1.9.9) (*66*) at 10- and 25-kb resolutions, and significant interaction pairs were detected using HOMER (version: 4.1.1) (*69*) at 25-kb resolution. Aggregated peak analysis (APA) was performed using the Juicer toolkit to measure the enrichment of the detected chromatin loops. Notably, as a consequence of the discrepancy in the sequencing depth achieved for adult PFC tissues between humans and macaques (approximately 3:1), we down-sampled the contact pairs for human data in proportion to the genome assembly size in the cross-species comparisons.

Prediction of promoter and enhancer regions in the macaque brain

ChIP-seq data of eight distinct anatomical regions of the macaque brain were downloaded from GEO (GSE67978) (*70*). Sequencing reads were aligned to the rheMac10Plus reference genome with Bowtie (version: 1.3.0) (*71*). Peak calling was performed with MACS2 (version: 2.2.7.1, parameters: $P = 10^{-5}$ extsize = 400 local lambda = 100,000) (*72*) for the ChIP-seq data of H3K27ac and H3K4me3 marks. To match the peak resolution of histone marks, peaks smaller than 2000 bp were extended to 2000 bp (*73*, *74*). For H3K27ac modification, the enriched regions identified in at least two biological replicates were defined as reproducible enriched regions. For H3K4me3, as only one biological sample was sequenced, all peaks called were retained. The regions marked by peaks of both H3K27ac and H3K4me3 and located within 1000 bp around the transcription start sites (TSSs) were defined as putative promoter regions. The regions marked by peaks of H3K27ac and located 1000 bp away from TSSs were defined as putative enhancer regions. The peaks were annotated with ChIPseeker tool (version: 1.22.1) (*75*). The final list of predicted promoter and enhancer regions in the macaque brain was obtained by merging the identifications from all of these brain samples.

Inference of the ancestral state for inversions and deletions in the macaque population

Genome assemblies of human (hg38), marmoset (calJac4), and rhesus macaque (rheMac10Plus) were used to trace the ancestral state of inversions and deletions in the macaque population. For inversions, we downloaded the chain/net alignment across humans and marmosets/macaques from the UCSC Genome Browser. The regions labeled with "INV" in the net alignment file were extracted as potential inverted regions. The inversions called within the macaque population were then compared with the annotations in these species, and the ancestral state of each inversion was then defined. Inversions with no authentic alignments or ambiguous ancestral states were excluded from the subsequent analyses. For deletions, we manually generated multiple alignment format (MAF) files between rhesus macaques and humans or marmosets using netToAxt. To decide whether each deletion variant represented a deletion within the population or an insertion in the reference genome, we then partitioned the genomic context of each deletion into two parts: the deletion region and the anchor regions (the upstream and downstream 200-bp regions). Basewise alignments of these two parts were extracted from the MAF files, and the alignment state was then determined if the alignments in both upstream and downstream regions could be anchored. The state (deletion or not) was then determined based on the alignment length of the deletion regions. The ancestral state was thus defined according to the alignments in these species. Deletions with no anchor alignments, inaccurate alignment lengths, or ambiguous ancestral states were excluded from the subsequent analyses.

Classification of population inversions based on regulatory regions

For the annotations of inversions from the perspective of genome architecture, we analyzed public Hi-C data from the fetal brain (CP and GZ zones) (*29*) and defined TAD boundaries as the 10-kb intervals (±5 kb) centered on the start and end coordinates of each TAD. Inversions showing overlap with TAD boundaries were defined as inversions disrupting TAD boundaries. For the annotation of inversions according to gene structures or functional regions, we extracted the gene structure annotations from the rheMac10Plus assembly and defined the sequences outside of all annotated genes as intergenic regions. Promoter and enhancer regions were predicted from the ChIP-seq data of macaque brains. Inversions overlapping with exons were defined as inversions disrupting exonic regions. Inversions overlapping with predicted promoter and enhancer regions were defined as inversions disrupting functional regions.

Detection of inversions between humans and macaques

We used the DoBlastzChainNet.pl script deposited in Kent's utilities of the UCSC Genome Bioinformatics Group to construct a chain/net alignment across the human (GRCh38/hg38) and macaque (rheMac10Plus) genomes. We first set up the alignment with a chain/net of hg38 as the target and rheMac10Plus as the query (-chainMinScore = 5000 -chainLinearGap = medium -syntenicNet), and the alignment in the reverse direction was then generated in "swap" mode (−swap -syntenicNet). The main method implemented with this script was pairwise whole-genome alignment using lastz (version: 1.04.00), with primate-specific parameters specified, which can be found at http://genomewiki.ucsc.edu/index.php/DoBlastzChainNet.pl#lastz_parameter_file. Then, the genomic regions that were labeled "INV" in the net alignment file and reciprocally intersected in both directions were extracted as candidate inversions, and those with more than 80% of the regions falling within SDs, as annotated in hg38, were filtered out. We further performed manual curation of each net inversion by revising the breakpoints of inversions, rescuing nested inversions showing breakpoint reuse, and removing ambiguous cases. Finally, 1972 species-specific inversions across humans and rhesus macaques were retained (table S7).

Validation of species-specific inversions between humans and macaques

Comparison with Strand-seq inversions

As Strand-seq studies typically report candidate inversions with broad boundaries, we used a relatively loose threshold (more than 10% coordinate intersection in the two studies) in determining the overlaps, which were then subjected to a manual check on the UCSC Genome Browser to confirm that these inversions were indeed identified by both methods. Notably, when calculating the overlap for the body regions of these inversions identified separately by the two methods, the mean percentage of the overlapped regions reaches up to 95.8%.

Validations with phylogenetic relationship

We performed pairwise genome alignments between human and other three old world monkeys, including crab-eating macaque, baboon, and green monkey, and checked whether these inversions were supported by the genome assembly of these closely related monkeys.

Strand-seq validation

The bam files for 61 single-cell Strand-seq libraries derived from one macaque lymphoblastoid cell line (LCL) were downloaded from NCBI BioProject under accession number PRJNA625922. We discarded supplementary alignments and only retained properly paired reads. To validate inversions with Strand-seq data, we first defined the template states (CC: plus-plus-strand, WW: minus-minus-strand, or WC: plus-minus-strand) for the flanking regions of these inversions, as described in Porubsky *et al.* (*76*). If the template states of the upstream and the downstream regions of the candidate inversion were identical, then the reads in the opposite direction in the inversion body region were defined as informative reads. Notably, because the removal of the newly synthesized strand may not be perfect, low abundance reads in the opposite direction are expected even for WW or CC chromosomes, typically at an average noise level of 5%. Accordingly, we estimated the background for the number of falsely assigned informative reads by assuming a 5% error rate probability. Only the cases where the reads coverage ≥3 at the inversion loci in the Strand-seq data were included in the evaluations.

FISH validation

To determine the state of discordant inversions between this study and a previous report (*4*), we chose three conserved genomic regions in human and macaque genomes to design probes. Through the order of these distinct probes in the FISH assay, we could distinguish between the two models (Fig. 4D). Oligonucleotide probes were designed using Oligominer (version: 1.7) following the online instructions and the method described in Beliveau *et al.* (*77*). The template oligo pools designed for all six genomic regions were synthesized by Synbio Technologies (Suzhou, China), and the flanking primer binding sequences used to amplify the probes are shown in table S8. The probes were amplified from the oligo pools as described in Li *et al.* (*78*). For the FISH experiments, HeLa-S3 cells and macaque LLC-MK2 cells were first synchronized in metaphase through 16 hours of nocodazole treatment, with final concentrations of 50 and 20 ng/ml, respectively. Subsequently, cells collected under trypsin (Life Technologies #25200-056) treatment were exposed to a hypotonic solution (10 mM KCl for HeLa-S3 cells, 30 mM KCl for macaque LLC-MK2 cells) for 15 min and fixed in 3:1 methanol:acetic acid. Metaphase spreads were then prepared and stained with DAPI (4′,6-diamidino-2-phenylindole) according to the methods of Li *et al.* (*79*). Finally, FISH experiments were conducted, and imaging was performed on a Nikon Live SR CSU-W1 Spinning Disk Confocal microscope as described in Li *et al.* (*78*).

Identification of human-specific inversions

To trace the evolutionary trajectory of each species-specific inversion between humans and macaques, we first selected all of the nonhuman primate species included in the multiple alignments of 99 vertebrate genomes against the human genome (GRCh38/hg38) in the UCSC Genome Browser. Bushbabies were excluded due to an incomplete assembly. Hence, 10 nonhuman primate species were retained for downstream analyses, including chimpanzees (panTro6), gorillas (gorGor6), orangutans (ponAbe3), gibbons (nomLeu3), crab-eating macaques (macFas5), baboons (papAnu4), green monkeys (chlSab2), marmosets (calJac4), squirrel monkeys (saiBol1), and rhesus macaques (rheMac10Plus).

We downloaded the MAF file derived from syntenic net alignment for each nonhuman primate species from the UCSC Genome Browser. For crab-eating macaques, the MAF file was not available in the UCSC Genome Browser, so we manually generated it using netToAxt. For rhesus macaques, we used the MAF file against hg38, which was constructed from scratch as described in the "Detection of inversions between humans and macaques" section. We partitioned the genomic context of each inversion into three parts: the inversion body itself and the upstream and downstream flanking regions, which were twice as long as the inversion. For each inversion, basewise alignments of these three parts were then extracted from the MAF file, and the alignment status of each part was determined if the following criteria were met: (i) 50% of bases fell on the same chromosome, and (ii) 80% of them were oriented in the same direction. The regions were defined as inverted if the following criteria were met: (i) at least one flanking region fell on the same chromosome as the inversion body region, and (ii) these two regions were aligned in the opposite direction with respect to the human genome. Notably, inversions were removed from downstream ancestor state reconstruction if they were not detected as inverted under these criteria between humans and rhesus macaques, as mentioned above.

Next, we used Phytools (version: 1.0–1) (*80*) to infer the ancestral states of these inversions, combining the inversion state across humans and other species and the primate phylogeny information. The phylogenetic trees of the primates were extracted under the primate node in the phylogenetic tree used in the MULTIZ multiple alignment against GRCh/hg38. We performed 1000 stochastic character mappings under the ER model (equal rates for all permitted transitions) for each inversion. For regions with ambiguous inversions, we set equal prior probabilities of 0.5 for both states. Then, we summarized these stochastic maps by determining the posterior probability at each node and uncertain tip, and the state with a posterior probability greater than 0.6 was defined as the estimated state.

With the estimated states at nodes and tips, we then categorized these inversions into two main classes of lineage specificity based on their common ancestors, including Hominoidea-origin and Cercopithecidae-origin sets. Human-specific and macaque-specific inversions were further highlighted from these two sets of distinct origins. For instance, an inversion was classified into a Hominoidea-origin lineage if it was derived in the ape branch and absent from any other nodes or tips. Notably, those inversions that failed in ancestral inference due to uncertainty at some nodes or tips were labeled as ambiguous and were excluded from the subsequent analyses.

Polymorphic status of human-specific inversions

To assess the polymorphic status of candidate human-specific inversions, we first gathered the genome sequences of 15 public human genome assemblies, including individuals of multiple distinct ethnicities or regions. Most of them were de novo assembled using long-read sequencing strategies. Notably, three assemblies (mHomSap3, HG00733, and HuRef) were diploid and were split into maternal and paternal haplotypes for downstream analyses. The metadata of these assemblies can be found in table S10. Similar to the method for the detection of inversions between humans and macaques, we performed pairwise, whole-genome alignment for each assembly against hg38 using DoBlastzChainNet.pl. The output MAF alignment files were used to determine the form of each candidate human-specific inversion in these human assemblies using the same criteria indicated in the previous section. An inversion was defined as fixed if all individuals exhibited concordant alignment with hg38 in this region.

We then evaluated the polymorphic status of the orthologous regions of these candidate human-specific inversions in the population of 572 macaques. Specifically, high-quality (mapping quality >30), improperly aligned, paired-end reads were extracted around the breakpoints of each candidate inversion, which were normalized according to the length of the regions and the total sequencing depth. Inversions identified with a high allele frequency (frequency > 0.5) in the macaque population were used as the positive control. Two sets of randomly shuffled genomic regions with the same length and chromosome distribution were used as the negative controls. The normalized, improperly aligned, paired-end reads were then calculated to assess the status of each region in each macaque animal. An inversion was defined as polymorphic in the macaque population when at least one animal showed inverted alignment according to the rheMac10Plus genome assembly.

To calculate the genetic divergence accumulated in the human or macaque lineage after the divergence of the species, the ancestral sequences were inferred from the EPO multiple alignments of 10 primates (bonobo, chimpanzee, crab-eating macaque, gibbon, gorilla, human, macaque, mouse lemur, Sumatran orangutan, vervet-AGM; fig. S11). Regions downstream and upstream of the inverted regions with matched lengths were used as the controls.

Annotation and classification of fixed human-specific inversions

The features of a "strong effect inversion" include larger size (>10 kb), changing gene structures, overlap with regulatory elements (promoters/enhancers), and changing chromatin architectures of compartments, TADs, and loops in brain (fetal CP, GZ and adult PFC). An inversion was defined as a strong effect inversion if it showed at least one of these features; otherwise, it was classified as a weak-effect inversion. The inversions were ranked according to their combined effects on gene structure or expression regulation.

Gene expression analyses

The RNA-seq reads of human and rhesus macaque across organs and developmental stages were downloaded from Cardoso-Moreira *et al.*

(*81*). The FASTQ files were aligned against GRCh38/hg38 genome for human samples and rheMac10Plus genome for rhesus macaque samples using HISAT2 (version: 2.2.1) (*82*). PCR duplicates were removed using Picard (version: 2.25.4). To minimize the discrepancy of gene annotation across species, we used XSAnno (version: 1.0) (*83*) to generate the comparative gene annotations between human and rhesus macaque based on GENCODE v33. The read numbers mapped to each gene were counted using featureCounts (version: 2.0.2) (*84*). DESeq2 (version: 1.26.0) (*85*) was used to perform differential expression analysis and the $\log_2$-transformed fold change (human versus macaque comparison) derived from it for orthologous genes was used to measure the expression divergence. Genes resided in inversions and 50% flanking regions were defined as putatively inversion-associated genes. We used processed RPKM (reads per kilobase per million mapped reads) tables of human, rhesus macaque, and mouse samples for quantification of gene expression levels, which were also obtained from Cardoso-Moreira *et al.* (*81*).

Enrichment analyses of contact losses

A permutation test was performed to quantify the degree of contact loss associated with an inversion. Briefly, we first counted the significant interactions in the 1-Mb flanking regions bridging the inward breakpoint (BP2) in rhesus macaques. Then, the conserved portions of flanking regions were retrieved from the human genome using LiftOver, with the minimum match percent set as 0.7. Next, the significant interactions were counted. We used the extent of loss to quantify the degree of significant interaction loss, which was defined as $(\text{Count}_m - \text{Count}_h)/\text{Count}_m$, where Count_m and Count_h are the significant interactions of flanking regions in macaques and humans, respectively. Next, BEDtools (version: 2.26.0) (*86*) was used to randomly shuffle intervals 10,000 times for BP2, and the extent of loss was calculated for each permutation following the same approach described above. The empirical *P* value was then calculated to assess the degree of contact loss.

Dual-luciferase reporter assay

The *APCDD1* promoter region of human and rhesus macaque was amplified and inserted into the pGL3-Basic plasmid, respectively. The mutant plasmids were constructed by introduction of point mutations using TaKaRa MutanBEST Kit. Before transfection, 293T cells were plated to 80% confluence in 12-well plates. Cells were then cotransfected with the *APCDD1*-promoter-constructs and pRL-TK plasmid expressing *Renilla* luciferase as the internal control. After 24 hours, the luciferase assay was performed using the dual luciferase reporter assay system (Promega) according to the manufacturer's instructions. *Firefly* and *Renilla* luciferase activities were measured by luminometers.

Human embryonic stem cell culture and neural induction

Human embryonic stem cells (hESCs; H9 cells) were cultured in Essential 8 Medium (Thermo Fisher Scientific, A1517001) on Matrigel (Corning) in six-well plates. The cells were regularly tested for mycoplasma. Neural induction was performed using small molecules as previously described (*87*). In brief, hESCs were detached with dispase to form embryoid bodies (EBs) in neural induction medium (NIM) with 100 nM LDN193189 and 10 μM SB431542. The medium was changed every day. After 4 days, the EBs were plated in neural progenitor medium (NPM) with 100 nM LDN193189 and 10 μM SB431542 onto plastic plates coated with growth factor-reduced (GFR) Matrigel. After 8 days of induction, the hESCs were differentiated into human NPCs.

The NPCs were maintained using NPM with basic fibroblast growth factor (10 ng/ml; bFGF) on GFR Matrigel. The culture medium was changed every day, and the cells were passaged every 6 days with Accutase.

Plasmid construction and lentivirus production

The full-length *APCDD1* sequence was amplified from the cDNA of 293T cells and cloned into the PCDH-CAG-IRES-mCherry backbone (from the laboratory of B. Hu). Lentiviruses were produced by transfecting 293T cells using the PCDH-*APCDD1*-IRES-mCherry plasmid or PCDH-IRES-mCherry plasmid combined with the packaging plasmids psPAX2 and pMD2.G. Then, the supernatant was collected for 3 days after transfection and concentrated with a filter device (Amicon Ultra15 Centrifuge Filters, Merck, Cat# UFC910008).

Lentiviral infection of NPCs

Human NPCs were plated on Matrigel-coated coverslips in 24-well plates with NPM plus 5 μM ROCK inhibitor 2 days before viral infection. Then, a concentrated lentivirus solution was added to the NPM with 5 μM ROCK inhibitor to infect cells. The lentivirus medium was removed on the second day. After infection, the cells were cultured with NPM without bFGF, and the medium was changed every 2 days for differentiation. Five days, 10 days, and 15 days after viral infection, cells grown on coverslips were collected and fixed with 4% paraformaldehyde (PFA) for 30 min at room temperature. Then, the PFA was washed with phosphate-buffered saline (PBS) three times. The fixed cells were conserved in PBS at 4°C for immunofluorescence staining.

Immunofluorescence staining

The coverslips were incubated for 1 hour in blocking buffer [1× PBS, 5% (w/v) bovine serum albumin (BSA), and 3% (v/v) Triton X-100] and incubated with primary antibodies diluted in 1% (w/v) BSA and 1% (v/v) Triton X-100 in PBS at 4°C overnight. Antibodies including anti-mCherry, anti-SOX2, and anti-TUJ1 were used. Then, the coverslips were washed with PBS three times and incubated with Alexa Fluor secondary antibodies diluted in the same solution used for the primary antibodies at room temperature for 1.5 hours. After three PBS washes, the coverslips were stained with Hoechst diluted 1/1000 in PBS for 15 min. Then, the coverslips were washed with PBS and mounted on glass slides with mounting reagent. Imaging was then performed using a Zeiss LSM980 confocal microscope, and the images were processed with ZEN software. For statistical analyses, cells were manually counted using ImageJ (*88*). At least 20 images were used for counting in every experimental condition.

Construction of *Apcdd1*-deficient mice

According to the structure of the *Apcdd1* gene, the region from exon 2 to exon 3 of the *Apcdd1*-203 (ENSMUST00000236135.1) transcript was defined as the knockout region. We used CRISPR-Cas9 technology to modify the *Apcdd1* gene. Briefly, sgRNA was transcribed in vitro. Cas9 and sgRNA were microinjected into fertilized eggs of C57BL/6J mice, and the fertilized eggs were transplanted to obtain positive F0 mice. A stable F1 generation mouse model was obtained by mating positive F0-generation mice with C57BL/6J mice. *Apcdd1*$^{+/+}$ (WT), *Apcdd1*$^{+/-}$ (Het), and *Apcdd1*$^{-/-}$ (DKO) mice were then generated using three different breeding schemes: Het × Het, Het × WT, and Het × DKO.

Western blottings

Total protein was extracted from the mouse brain using RIPA lysis buffer (Solarbio R0020) supplemented with 1 mM PMSF and complete protease inhibitor cocktail. Protein concentrations were determined using a Pierce BCA Protein Assay (Thermo Fisher Scientific 23227). The primary antibodies were anti-Apcdd1 (dilution: 1:1000; NOVUS; NB110–92756), anti-Apcdd1 (dilution: 1:1000; R&D; MAB10501-100), and anti–glyceraldehyde-3-phosphate dehydrogenase (GAPDH) (dilution: 1:1000; Abcam; ab8245). The secondary antibodies were horseradish peroxidase (HRP)–conjugated goat anti-rabbit immunoglobulin (IgG; EARTH, E030120–01) and HRP-conjugated goat anti-mouse IgG (EARTH, E030110–01).

scRNA-seq in mouse embryonic brains

The E10.5 mouse embryos were dissected from pregnant mice euthanized by anesthesia and decapitation. The head part of embryo was dissected out by fine tweezers and stored in ShbioTissue Preservation Solution (21903-10) overnight, which were then dissociated using the ShbioTissue Dissociation Kit (219517-10). Briefly, tissues were incubated in 1 ml of enzyme mix on a metal rotor at 37°C for 15 min, during which specimens were gently mixed 20 times using 1000-ml tips every 7 min. Cell dissociations were terminated by adding 2 ml of complete medium and then suspensions were filtered through a 100-μm strainer and centrifuged for 10 min at 500*g*. The supernatant was discarded and the cell pellet was resuspended with 300 μl of suspension buffer (1× Dulbecco's Phosphate-Buffered Saline (DPBS) containing 2% fetal bovine serum). Dyed by DAPI (50 ng/ml) for 3 min and refiltered with a 40-μm strainer, cell suspensions were then sorted with FACS BD Aria III sorter to remove dead cells, fragments, and lumps. Cell numbers of suspensions were measured with fluorescence cell counter (Luna-FLTM, Logos Biosystems) and hemocytometer using trypan blue staining. Cell suspensions were then adjusted to a proper cell concentration for library construction of scRNA-seq.

scRNA-seq was then performed using the DNBelab C high-throughput single-cell RNA library prep kit (MGI) following the manufacturer's protocol. The QubitTM Flex Fluorometer (Thermo Fisher Scientific) and Agilent Fragment Analyzer were used to determine the concentration and assess the quality of the libraries, which were then subjected to DNBSEQ-2000RS for sequencing.

Cell barcode and unique molecular identifiers (UMI) sequences were parsed using DNBelab_C_Series_HT_scRNA-analysis-software. Raw reads were then aligned against GRCm38/mm10 genome and GENCODE M25 gene annotation using *DNBC4tools run* with default parameters to generate sparse gene count matrices. Scanpy (version: 1.9.1) (*89*) was used for downstream analyses. Low-quality cells were filtered out considering the abnormal gene count and high mitochondrial read percentage. We normalized the gene expression by scaling total UMIs to 10,000 in each cell followed by log transformation. The BBKNN algorithm (*90*) was used to remove batch effects across four samples. Then, PCA was performed and the top 50 PCs were used for uniform manifold approximation and projection (UMAP) and Leiden clustering (*91*). We then determined differentially expressed genes using Wilcoxon rank-sum test method for each cluster. The known marker genes (*39*–*43*) was used to assign cell type identities. Finally, for each cell type, the compositional shift between *Apcdd1*$^{+/-}$ mice and the wild type was investigated using Poisson regression analysis with Wald test (*92*).

Behavioral tests

Morris water maze

We performed behavioral tests of 23 wild-type mice and 15 *Apcdd1*$^{+/-}$ mice at 9 to 11 weeks, following previous experience in behavior test with sufficient statistical power. Visual cues were placed on the wall of the testing room approximately 1 m from the edge of the pool. The 120-cm pool was surrounded by a deep-colored curtain with visual cues left inside. A circular escape platform (10 cm in diameter) was fixed and submerged 1.5 cm under the water. On each of the five consecutive training days, the pool was divided into four quadrants, filled with water, and the water was turned white with nontoxic white paint. Each mouse was trained in three trials each day. In each trial, the mouse was released from a pseudo-randomly assigned starting location, which was changed every day but remained consistent for all mice tested. A mouse succeeded in a trial when the mouse mounted and stayed on the platform for 5 s. A mouse failed in a trial when the mouse spent 90 s in the pool and failed to mount the platform. When the mouse failed a trial, the mouse was assisted in mounting the platform and allowed an additional 30 s to remain on the platform. The time the mice spent mounting the platform in each trial was recorded and analyzed. Single probe trials were conducted on day 6. During the probe trial, the platform was removed, and each mouse was given 90 s to navigate in the pool. The number of times the mice swam across the platform location and the time the mice spent in each quadrant were recorded and analyzed. After a week of probe trials, the mice were subjected to a reversal learning test. During reversal learning, the platform was placed in the opposite quadrant of the pool. The mice were trained for five consecutive days, as before. On day 6, probe trials were conducted as described above, and the number of times the mice swam across the platform location and the time the mice spent in each quadrant were recorded and analyzed. The investigators were blinded to the wild-type and *Apcdd1*$^{+/-}$ mice group in the downstream statistical analyses.

Supplementary Materials

This 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

REFERENCES AND NOTES

1. H. J. Abel, D. E. Larson, A. A. Regier, C. Chiang, I. Das, K. L. Kanchi, R. M. Layer, B. M. Neale, W. J. Salerno, C. Reeves, S. Buyske; NHGRI Centers for Common Disease Genomics, T. C. Matise, D. M. Muzny, M. C. Zody, E. S. Lander, S. K. Dutcher, N. O. Stitziel, I. M. Hall, Mapping and characterization of structural variation in 17,795 human genomes. *Nature* **583**, 83–89 (2020).
2. M. A. Almarri, A. Bergström, J. Prado-Martinez, F. Yang, B. Fu, A. S. Dunham, Y. Chen, M. E. Hurles, C. Tyler-Smith, Y. Xue, Population structure, stratification, and introgression of human structural variation. *Cell* **182**, 189–199.e15 (2020).
3. R. L. Collins, A structural variation reference for medical and population genetics. *Nature* **581**, 444–451 (2020).
4. F. A. M. Maggiolini, A. D. Sanders, C. J. Shew, A. Sulovari, Y. Mao, M. Puig, C. R. Catacchio, M. Dellino, D. Palmisano, L. Mercuri, M. Bitonto, D. Porubský, M. Cáceres, E. E. Eichler,

M. Ventura, M. Y. Dennis, J. O. Korbel, F. Antonacci, Single-cell strand sequencing of a macaque genome reveals multiple nested inversions and breakpoint reuse during primate evolution. *Genome Res.* **30**, 1680–1693 (2020).

5. G. H. Perry, F. Yang, T. Marques-Bonet, C. Murphy, T. Fitzgerald, A. S. Lee, C. Hyland, A. C. Stone, M. E. Hurles, C. Tyler-Smith, E. E. Eichler, N. P. Carter, C. Lee, R. Redon, Copy number variation and evolution in humans and chimpanzees. *Genome Res.* **18**, 1698–1710 (2008).
6. W. C. Warren, R. A. Harris, M. Haukness, I. T. Fiddes, S. C. Murali, J. Fernandes, P. C. Dishuck, J. M. Storer, M. Raveendran, L. D. W. Hillier, D. Porubsky, Y. Mao, D. Gordon, M. R. Vollger, A. P. Lewis, K. M. Munson, E. DeVogelaere, J. Armstrong, M. Diekhans, J. A. Walker, C. Tomlinson, T. A. Graves-Lindsay, M. Kremitzki, S. R. Salama, P. A. Audano, M. Escalona, N. W. Maurer, F. Antonacci, L. Mercuri, F. A. M. Maggiolini, C. R. Catacchio, J. G. Underwood, D. H. O'Connor, A. D. Sanders, J. O. Korbel, B. Ferguson, H. M. Kubisch, L. Picker, N. H. Kalin, D. Rosene, J. Levine, D. H. Abbott, S. B. Gray, M. M. Sanchez, Z. A. Kovacs-Balint, J. W. Kemnitz, S. M. Thomasy, J. A. Roberts, E. L. Kinnally, J. P. Capitanio, J. H. P. Skene, M. Platt, S. A. Cole, R. E. Green, M. Ventura, R. W. Wiseman, B. Paten, M. A. Batzer, J. Rogers, E. E. Eichler, Sequence diversity analyses of an improved rhesus macaque genome enhance its biomedical utility. *Science* **370**, (2020).
7. O. Gokcumen, V. Tischler, J. Tica, Q. Zhu, R. C. Iskow, E. Lee, M. H. Y. Fritz, A. Langdon, A. M. Stütz, P. Pavlidis, V. Benes, R. E. Mills, P. J. Park, C. Lee, J. O. Korbel, Primate genome architecture influences structural variation mechanisms and functional consequences. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 15764–15769 (2013).
8. 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. M. E. Welch, M. Sorensen, C. Baker, R. S. Fulton, J. Armstrong, T. A. Graves-Lindsay, A. M. Denli, E. R. Hoppe, P. H. 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**, (2018).
9. J. Weischenfeldt, O. Symmons, F. Spitz, J. O. Korbel, Phenotypic impact of genomic structural variation: Insights from and for human disease. *Nat. Rev. Genet.* **14**, 125–138 (2013).
10. P. H. Sudmant, T. Rausch, E. J. Gardner, R. E. Handsaker, A. Abyzov, J. Huddleston, Y. Zhang, K. Ye, G. Jun, M. Hsi-Yang Fritz, M. K. Konkel, A. Malhotra, A. M. Stütz, X. Shi, F. Paolo Casale, J. Chen, F. Hormozdiari, G. Dayama, K. Chen, M. Malig, M. J. P. Chaisson, K. Walter, S. Meiers, S. Kashin, E. Garrison, A. Auton, H. Y. K. Lam, X. Jasmine Mu, C. Alkan, D. Antaki, T. Bae, E. Cerveira, P. Chines, Z. Chong, L. Clarke, E. Dal, L. Ding, S. Emery, X. Fan, M. Gujral, F. Kahveci, J. M. Kidd, Y. Kong, E. W. Lameijer, S. McCarthy, P. Flicek, R. A. Gibbs, G. Marth, C. E. Mason, A. Menelaou, D. M. Muzny, B. J. Nelson, A. Noor, N. F. Parrish, M. Pendleton, A. Quitadamo, B. Raeder, E. E. Schadt, M. Romanovitch, A. Schlattl, R. Sebra, A. A. Shabalin, A. Untergasser, J. A. Walker, M. Wang, F. Yu, C. Zhang, J. Zhang, X. Zheng-Bradley, W. Zhou, T. Zichner, J. Sebat, M. A. Batzer, S. A. McCarroll; The 1000 Genomes Project Consortium, R. E. Mills, M. B. Gerstein, A. Bashir, O. Stegle, S. E. Devine, C. Lee, E. E. Eichler, J. O. Korbel, An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–81 (2015).
11. Z. Liu, R. Roberts, T. R. Mercer, J. Xu, F. J. Sedlazeck, W. Tong, Towards accurate and reliable resolution of structural variants for clinical diagnosis. *Genome Biol.* **23**, 68 (2022).
12. J. Wagner, N. D. Olson, L. Harris, J. McDaniel, H. Cheng, A. Fungtammasan, Y. C. Hwang, R. Gupta, A. M. Wenger, W. J. Rowell, Z. M. Khan, J. Farek, Y. Zhu, A. Pisupati, M. Mahmoud, C. Xiao, B. Yoo, S. M. E. Sahraeian, D. E. Miller, D. Jáspez, J. M. Lorenzo-Salazar, A. Muñoz-Barrera, L. A. Rubio-Rodríguez, C. Flores, G. Narzisi, U. S. Evani, W. E. Clarke, J. Lee, C. E. Mason, S. E. Lincoln, K. H. Miga, M. T. W. Ebbert, A. Shumate, H. Li, C. S. Chin, J. M. Zook, F. J. Sedlazeck, Curated variation benchmarks for challenging medically relevant autosomal genes. *Nat. Biotechnol.* **40**, 672–680 (2022).
13. M. Mahmoud, H. Doddapaneni, W. Timp, F. J. Sedlazeck, PRINCESS: Comprehensive detection of haplotype resolved SNVs, SVs, and methylation. *Genome Biol.* **22**, 268 (2021).
14. J. Rogers, R. A. Gibbs, Comparative primate genomics: Emerging patterns of genome content and dynamics. *Nat. Rev. Genet.* **15**, 347–359 (2014).
15. Rhesus Macaque Genome Sequencing and Analysis Consortium, 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. Joshi, 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.-Z. Song, T. Wylie, L. Zhang, E. Birney, K. Han, M. K. Konkel, J. Lee, A. F. A. 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, Evolutionary and biomedical insights from the rhesus macaque genome. *Science* **316**, 222–234 (2007).
16. X. Zhong, J. Peng, Q. S. Shen, J. Y. Chen, H. Gao, X. Luan, S. Yan, X. Huang, S. J. Zhang, L. Xu, X. Zhang, B. C. M. Tan, C. Y. Li, RhesusBase PopGateway: Genome-wide population genetics atlas in rhesus macaque. *Mol. Biol. Evol.* **33**, 1370–1375 (2016).
17. B. N. Bimber, M. Y. Yan, S. M. Peterson, B. Ferguson, mGAP: The macaque genotype and phenotype resource, a framework for accessing and interpreting macaque variant data, and identifying new models of human disease. *BMC Genomics* **20**, 176 (2019).
18. Z. Liu, X. Tan, P. Orozco-terWengel, X. Zhou, L. Zhang, S. Tian, Z. Yan, H. Xu, B. Ren, P. Zhang, Z. Xiang, B. Sun, C. Roos, M. W. Bruford, M. Li, Population genomics of wild Chinese rhesus macaques reveals a dynamic demographic history and local adaptation, with implications for biomedical research. *Gigascience* **7**, (2018).
19. 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).
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).
21. J. M. Zook, N. F. Hansen, N. D. Olson, L. Chapman, J. C. Mullikin, C. Xiao, S. Sherry, S. Koren, A. M. Phillippy, P. C. Boutros, S. M. E. Sahraeian, V. Huang, A. Rouette, N. Alexander, C. E. Mason, I. Hajirasouliha, C. Ricketts, J. Lee, R. Tearle, I. T. Fiddes, A. M. Barrio, J. Wala, A. Carroll, N. Ghaffari, O. L. Rodriguez, A. Bashir, S. Jackman, J. J. Farrell, A. M. Wenger, C. Alkan, A. Soylev, M. C. Schatz, S. Garg, G. Church, T. Marschall, K. Chen, X. Fan, A. C. English, J. A. Rosenfeld, W. Zhou, R. E. Mills, J. M. Sage, J. R. Davis, M. D. Kaiser, J. S. Oliver, A. P. Catalano, M. J. P. Chaisson, N. Spies, F. J. Sedlazeck, M. Salit, A robust benchmark for detection of germline large deletions and insertions. *Nat. Biotechnol.* **38**, 1347–1355 (2020).
22. M. Pendleton, R. Sebra, A. W. C. Pang, A. Ummat, O. Franzen, T. Rausch, A. M. Stütz, W. Stedman, T. Anantharaman, A. Hastie, H. Dai, M. H. Y. Fritz, H. Cao, A. Cohain, G. Deikus, R. E. Durrett, S. C. Blanchard, R. Altman, C. S. Chin, Y. Guo, E. E. Paxinos, J. O. Korbel, R. B. Darnell, W. R. McCombie, P. Y. Kwok, C. E. Mason, E. E. Schadt, A. Bashir, Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nat. Methods* **12**, 780–786 (2015).
23. D. Porubsky, A. D. Sanders, W. Höps, P. H. Hsieh, A. Sulovari, R. Li, L. Mercuri, M. Sorensen, S. C. Murali, D. Gordon, S. Cantsilieris, A. A. Pollen, M. Ventura, F. Antonacci, T. Marschall, J. O. Korbel, E. E. Eichler, Recurrent inversion toggling and great ape genome evolution. *Nat. Genet.* **52**, 849–858 (2020).
24. P. H. Sudmant, S. Mallick, B. J. Nelson, F. Hormozdiari, N. Krumm, J. Huddleston, B. P. Coe, C. Baker, S. Nordenfelt, M. Bamshad, L. B. Jorde, O. L. Posukh, H. Sahakyan, W. S. Watkins, L. Yepiskoposyan, M. S. Abdullah, C. M. Bravi, C. Capelli, T. Hervig, J. T. S. Wee, C. Tyler-Smith, G. van Driem, I. G. Romero, A. R. Jha, S. Karachanak-Yankova, D. Toncheva, D. Comas, B. Henn, T. Kivisild, A. Ruiz-Linares, A. Sajantila, E. Metspalu, J. Parik, R. Villems, E. B. Starikovskaya, G. Ayodo, C. M. Beall, A. di Rienzo, M. F. Hammer, R. Khusainova, E. Khusnutdinova, W. Klitz, C. Winkler, D. Labuda, M. Metspalu, S. A. Tishkoff, S. Dryomov, R. Sukernik, N. Patterson, D. Reich, E. E. Eichler, Global diversity, population stratification, and selection of human copy-number variation. *Science* **349**, aab3761 (2015).
25. M. Kirkpatrick, N. Barton, Chromosome inversions, local adaptation and speciation. *Genetics* **173**, 419–434 (2006).
26. M. Kirkpatrick, How and why chromosome inversions evolve. *PLOS Biol.* **8**, e1000501 (2010).
27. E. L. Berdan, A. Blanckaert, R. K. Butlin, C. Bank, Deleterious mutation accumulation and the long-term fate of chromosomal inversions. *PLOS Genet.* **17**, e1009411 (2021).
28. R. Faria, K. Johannesson, R. K. Butlin, A. M. Westram, Evolving inversions. *Trends Ecol. Evol.* **34**, 239–248 (2019).
29. X. Luo, Y. Liu, D. Dang, T. Hu, Y. Hou, X. Meng, F. Zhang, T. Li, C. Wang, M. Li, H. Wu, Q. Shen, Y. Hu, X. Zeng, X. He, L. Yan, S. Zhang, C. Li, B. Su, 3D Genome of macaque fetal brain reveals evolutionary innovations during primate corticogenesis. *Cell* **184**, 723–740.e21 (2021).

30. A. Vonica, N. Bhat, K. Phan, J. Guo, L. Iancu, J. A. Weber, A. Karger, J. W. Cain, E. C. E. Wang, G. M. De Stefano, A. H. O'Donnell-Luria, A. M. Christiano, B. Riley, S. J. Butler, V. Luria, Apcdd1 is a dual BMP/Wnt inhibitor in the developing nervous system and skin. *Dev. Biol.* **464**, 72–88 (2020).
31. H. K. Lee, D. Laug, W. Zhu, J. M. Patel, K. Ung, B. R. Arenkiel, S. P. J. Fancy, C. Mohila, B. Deneen, Apcdd1 stimulates oligodendrocyte differentiation after white matter injury. *Glia* **63**, 1840–1849 (2015).
32. E. Suzuki, T. Fukuda, Multifaceted functions of TWSG1: From embryogenesis to cancer development. *Int. J. Mol. Sci.* **23**, (2022).
33. V. Brivio, C. Faivre-Sarrailh, E. Peles, D. L. Sherman, P. J. Brophy, Assembly of CNS nodes of ranvier in myelinated nerves is promoted by the axon cytoskeleton. *Curr. Biol.* **27**, 1068–1073 (2017).
34. S. Kanton, M. J. Boyle, Z. He, M. Santel, A. Weigert, F. Sanchís-Calleja, P. Guijarro, L. Sidow, J. S. Fleck, D. Han, Z. Qian, M. Heide, W. B. Huttner, P. Khaitovich, S. Pääbo, B. Treutlein, J. G. Camp, Organoid single-cell genomic atlas uncovers human-specific features of brain development. *Nature* **574**, 418–422 (2019).
35. L. Shi, X. Luo, J. Jiang, Y. Chen, C. Liu, T. Hu, M. Li, Q. Lin, Y. Li, J. Huang, H. Wang, Y. Niu, Y. Shi, M. Styner, J. Wang, Y. Lu, X. Sun, H. Yu, W. Ji, B. Su, Transgenic rhesus monkeys carrying the human MCPH1 gene copies show human-like neoteny of brain development. *Natl. Sci. Rev.* **6**, 480–493 (2019).
36. S. Lodato, P. Arlotta, Generating neuronal diversity in the mammalian cerebral cortex. *Annu. Rev. Cell Dev. Bi.* **31**, 699–720 (2015).
37. D. J. Di Bella, E. Habibi, R. R. Stickels, G. Scalia, J. Brown, P. Yadollahpour, S. M. Yang, C. Abbate, T. Biancalani, E. Z. Macosko, F. Chen, A. Regev, P. Arlotta, Molecular logic of cellular diversification in the mouse cerebral cortex. *Nature* **595**, 554-+ (2021).
38. F. Stagni, A. Giacomini, S. Guidi, E. Ciani, R. Bartesaghi, Timing of therapies for Down syndrome: The sooner, the better. *Front Behav. Neurosci.* **9**, (2015).
39. G. La Manno, K. Siletti, A. Furlan, D. Gyllborg, E. Vinsland, A. M. Albiach, C. M. Langseth, I. Khven, A. R. Lederer, L. M. Dratva, A. Johnsson, M. Nilsson, P. Lönnerberg, S. Linnarsson, Molecular architecture of the developing mouse brain. *Nature* **596**, 92–96 (2021).
40. O. Franzen, L. M. Gan, J. L. M. Bjorkegren, PanglaoDB: A web server for exploration of mouse and human single-cell RNA sequencing data. *Database (Oxford)* **2019**, (2019).
41. S. Jansky, A. K. Sharma, V. Körber, A. Quintero, U. H. Toprak, E. M. Wecht, M. Gartlgruber, A. Greco, E. Chomsky, T. G. P. Grünewald, K. O. Henrich, A. Tanay, C. Herrmann, T. Höfer, F. Westermann, Single-cell transcriptomic analyses provide insights into the developmental origins of neuroblastoma. *Nat. Genet.* **53**, 683–693 (2021).
42. Z. Gao, K. Ure, J. L. Ables, D. C. Lagace, K. A. Nave, S. Goebbels, A. J. Eisch, J. Hsieh, Neurod1 is essential for the survival and maturation of adult-born neurons. *Nat. Neurosci.* **12**, 1090–1092 (2009).
43. A. M. Tadeu, V. Horsley, Notch signaling represses p63 expression in the developing surface ectoderm. *Development* **140**, 3777–3786 (2013).
44. B. Zhou, Y. He, Y. Chen, B. Su, Comparative genomic analysis identifies great-ape-specific structural variants and their evolutionary relevance. *Mol. Biol. Evol.* **40**, msad184 (2023).
45. R. S. Harris, "Improved pairwise alignment of genomic DNA," thesis, The Pennsylvania State University (2007).
46. A. K. Leung, N. Jin, K. Y. Yip, T. F. Chan, OMTools: A software package for visualizing and processing optical mapping data. *Bioinformatics* **33**, 2933–2935 (2017).
47. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
48. J. T. Robinson, H. Thorvaldsdottir, D. Turner, J. P. Mesirov, igv.js: An embeddable JavaScript implementation of the Integrative Genomics Viewer (IGV). *Bioinformatics* **39**, (2023).
49. S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. W. de Bakker, M. J. Daly, P. C. Sham, PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
50. S. Purcell, PLINK (v1.90b6.16 64-bit), http://pngu.mgh.harvard.edu/purcell/plink/.
51. P. Cingolani, A. Platts, L. L. Wang, M. Coon, T. Nguyen, L. Wang, S. J. Land, X. Lu, D. M. Ruden, A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly* **6**, 80–92 (2012).
52. A. Manichaikul, J. C. Mychaleckyj, S. S. Rich, K. Daly, M. Sale, W. M. Chen, Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010).
53. C. Bycroft, C. Freeman, D. Petkova, G. Band, L. T. Elliott, K. Sharp, A. Motyer, D. Vukcevic, O. Delaneau, J. O'Connell, A. Cortes, S. Welsh, A. Young, M. Effingham, G. McVean, S. Leslie, N. Allen, P. Donnelly, J. Marchini, The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
54. P. Danecek, A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth, S. T. Sherry, G. McVean, R. Durbin; 1000 Genomes Project Analysis Group, The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
55. V. Lefort, R. Desper, O. Gascuel, FastME 2.0: A comprehensive, accurate, and fast distance-based phylogeny inference program. *Mol. Biol. Evol.* **32**, 2798–2800 (2015).
56. J. Yang, S. H. Lee, M. E. Goddard, P. M. Visscher, GCTA: A tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).
57. T. Rausch, T. Zichner, A. Schlattl, A. M. Stütz, V. Benes, J. O. Korbel, DELLY: Structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**, i333–i339 (2012).
58. C. Chiang, R. M. Layer, G. G. Faust, M. R. Lindberg, D. B. Rose, E. P. Garrison, G. T. Marth, A. R. Quinlan, I. M. Hall, SpeedSeq: Ultra-fast personal genome analysis and interpretation. *Nat. Methods* **12**, 966–968 (2015).
59. D. E. Larson, H. J. Abel, C. Chiang, A. Badve, I. das, J. M. Eldred, R. M. Layer, I. M. Hall, svtools: Population-scale analysis of structural variation. *Bioinformatics* **35**, 4782–4787 (2019).
60. J. R. Belyeu, M. Chowdhury, J. Brown, B. S. Pedersen, M. J. Cormier, A. R. Quinlan, R. M. Layer, Samplot: A platform for structural variant visual validation and automated filtering. *Genome Biol.* **22**, 161 (2021).
61. X. Zhao, A. M. Weber, R. E. Mills, A recurrence-based approach for validating structural variation using long-read sequencing technology. *Gigascience* **6**, 1–9 (2017).
62. J. R. Belyeu, T. J. Nicholas, B. S. Pedersen, T. A. Sasani, J. M. Havrilla, S. N. Kravitz, M. E. Conway, B. K. Lohman, A. R. Quinlan, R. M. Layer, SV-plaudit: A cloud-based framework for manually curating thousands of structural variants. *Gigascience* **7**, (2018).
63. A. C. English, V. K. Menon, R. A. Gibbs, G. A. Metcalf, F. J. Sedlazeck, Truvari: Refined structural variant comparison preserves allelic diversity. *Genome Biol.* **23**, 271 (2022).
64. B. J. Walker, T. Abeel, T. Shea, M. Priest, A. Abouelliel, S. Sakthikumar, C. A. Cuomo, Q. Zeng, J. Wortman, S. K. Young, A. M. Earl, Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLOS ONE* **9**, e112963 (2014).
65. N. Servant, N. Varoquaux, B. R. Lajoie, E. Viara, C. J. Chen, J. P. Vert, E. Heard, J. Dekker, E. Barillot, HiC-Pro: An optimized and flexible pipeline for Hi-C data processing. *Genome Biol.* **16**, 259 (2015).
66. S. S. Rao, M. H. Huntley, N. C. Durand, E. K. Stamenova, I. D. Bochkov, J. T. Robinson, A. L. Sanborn, I. Machol, A. D. Omer, E. S. Lander, E. L. Aiden, A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).
67. N. C. Durand, M. S. Shamim, I. Machol, S. S. P. Rao, M. H. Huntley, E. S. Lander, E. L. Aiden, Juicer provides a one-click system for analyzing loop-resolution hi-c experiments. *Cell Syst.* **3**, 95–98 (2016).
68. K. Kruse, C. B. Hug, J. M. Vaquerizas, FAN-C: A feature-rich framework for the analysis and visualisation of chromosome conformation capture data. *Genome Biol.* **21**, 303 (2020).
69. S. H. Duttke, M. W. Chang, S. Heinz, C. Benner, Identification and dynamic quantification of regulatory elements using total RNA. *Genome Res.* **29**, 1836–1846 (2019).
70. M. W. Vermunt, S. C. Tan, B. Castelijns, G. Geeven, P. Reinink, E. de Bruijn, I. Kondova, S. Persengiev; Netherlands Brain Bank, R. Bontrop, E. Cuppen, W. de Laat, M. P. Creyghton, Epigenomic annotation of gene regulatory alterations during evolution of the primate brain. *Nat. Neurosci.* **19**, 494–503 (2016).
71. B. Langmead, C. Trapnell, M. Pop, S. L. Salzberg, Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
72. Y. Zhang, T. Liu, C. A. Meyer, J. Eeckhoute, D. S. Johnson, B. E. Bernstein, C. Nusbaum, R. M. Myers, M. Brown, W. Li, X. S. Liu, Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
73. M. W. Vermunt, P. Reinink, J. Korving, E. de Bruijn, P. M. Creyghton, O. Basak, G. Geeven, P. W. Toonen, N. Lansu, C. Meunier, S. van Heesch; Netherlands Brain Bank, H. Clevers, W. de Laat, E. Cuppen, M. P. Creyghton, Large-scale identification of coregulated enhancer networks in the adult human brain. *Cell Rep.* **9**, 767–779 (2014).
74. D. Villar, C. Berthelot, S. Aldridge, T. F. Rayner, M. Lukk, M. Pignatelli, T. J. Park, R. Deaville, J. T. Erichsen, A. J. Jasinska, J. M. A. Turner, M. F. Bertelsen, E. P. Murchison, P. Flicek, D. T. Odom, Enhancer evolution across 20 mammalian species. *Cell* **160**, 554–566 (2015).
75. G. Yu, L. G. Wang, Q. Y. He, ChIPseeker: An R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics* **31**, 2382–2383 (2015).
76. D. Porubsky, A. D. Sanders, A. Taudt, M. Colomé-Tatché, P. M. Lansdorp, V. Guryev, breakpointR: An R/Bioconductor package to localize strand state changes in Strand-seq data. *Bioinformatics* **36**, 1260–1261 (2020).
77. B. J. Beliveau, J. Y. Kishi, G. Nir, H. M. Sasaki, S. K. Saka, S. C. Nguyen, C. T. Wu, P. Yin, OligoMiner provides a rapid, flexible environment for the design of genome-scale oligonucleotide in situ hybridization probes. *Proc. Natl. Acad. Sci. U.S.A.* **115**, E2183–E2192 (2018).
78. Y. Li, B. Xue, M. Zhang, L. Zhang, Y. Hou, Y. Qin, H. Long, Q. P. Su, Y. Wang, X. Guan, Y. Jin, Y. Cao, G. Li, Y. Sun, Transcription-coupled structural dynamics of topologically associating domains regulate replication origin efficiency. *Genome Biol.* **22**, 206 (2021).
79. W. Li, X. Bai, J. Li, Y. Zhao, J. Liu, H. Zhao, L. Liu, M. Ding, Q. Wang, F. Y. Shi, M. Hou, J. Ji, G. Gao, R. Guo, Y. Sun, Y. Liu, D. Xu, The nucleoskeleton protein IFFO1 immobilizes broken DNA and suppresses chromosome translocation during tumorigenesis. *Nat. Cell Biol.* **21**, 1273–1285 (2019).
80. L. J. Revell, R. Graham Reynolds, A new Bayesian method for fitting evolutionary models to comparative data with intraspecific variation. *Evolution* **66**, 2697–2707 (2012).

81. M. Cardoso-Moreira, J. Halbert, D. Valloton, B. Velten, C. Chen, Y. Shao, A. Liechti, K. Ascenção, C. Rummel, S. Ovchinnikova, P. V. Mazin, I. Xenarios, K. Harshman, M. Mort, D. N. Cooper, C. Sandi, M. J. Soares, P. G. Ferreira, S. Afonso, M. Carneiro, J. M. A. Turner, J. L. VandeBerg, A. Fallahshahroudi, P. Jensen, R. Behr, S. Lisgo, S. Lindsay, P. Khaitovich, W. Huber, J. Baker, S. Anders, Y. E. Zhang, H. Kaessmann, Gene expression across mammalian organ development. *Nature* **571**, 505–509 (2019).
82. D. Kim, J. M. Paggi, C. Park, C. Bennett, S. L. Salzberg, Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).
83. Y. Zhu, M. Li, A. M. Sousa, N. Sestan, XSAnno: A framework for building ortholog models in cross-species transcriptome comparisons. *BMC Genomics* **15**, 343 (2014).
84. Y. Liao, G. K. Smyth, W. Shi, featureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
85. M. I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
86. A. R. Quinlan, I. M. Hall, BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
87. M. Wu, D. Zhang, C. Bi, T. Mi, W. Zhu, L. Xia, Z. Teng, B. Hu, Y. Wu, A chemical recipe for generation of clinical-grade striatal neurons from hESCs. *Stem. Cell Reports* **11**, 635–650 (2018).
88. C. A. Schneider, W. S. Rasband, K. W. Eliceiri, NIH Image to ImageJ: 25 years of image analysis. *Nat. Methods* **9**, 671–675 (2012).
89. F. A. Wolf, P. Angerer, F. J. Theis, SCANPY: Large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
90. K. Polanski, M. D. Young, Z. Miao, K. B. Meyer, S. A. Teichmann, J.-E. Park, BBKNN: Fast batch alignment of single cell transcriptomes. *Bioinformatics* **36**, 964–965 (2020).
91. V. A. Traag, L. Waltman, N. J. van Eck, From Louvain to Leiden: Guaranteeing well-connected communities. *Sci. Rep.* **9**, 5233 (2019).
92. A. L. Haber, M. Biton, N. Rogel, R. H. Herbst, K. Shekhar, C. Smillie, G. Burgin, T. M. Delorey, M. R. Howitt, Y. Katz, I. Tirosh, S. Beyaz, D. Dionne, M. Zhang, R. Raychowdhury, W. S. Garrett, O. Rozenblatt-Rosen, H. N. Shi, O. Yilmaz, R. J. Xavier, A. Regev, A single-cell survey of the small intestinal epithelium. *Nature* **551**, 333–339 (2017).

Acknowledgments: We thank Y. E. Zhang at the Institute of Zoology, Chinese Academy of Sciences for insightful suggestions. We thank S. Gao at MGI Tech for technical support. We thank the National Center for Protein Sciences at Peking University in Beijing, China, for assistance with the DNA FISH experiment. **Funding:** This work was supported by grants from the Ministry of Science and Technology of China (National Key Research and Development Program of China, 2019YFA0801801 and 2018YFA0801405), the National Natural Science Foundation of China (32300510 and 31871272), the Chinese Institute for Brain Research (2020-NKX-XM-11), and the Clinical Medicine Plus X–Young Scholars Project, Peking University (PKU2022LCXQ015). **Author contributions:** C.-Y.L. and L.Z. conceived and designed the study. W.D., X.L., and J.Z. analyzed most of the data and performed statistical analysis. X.M.Z., J.W., C.X., T.L., Q.Y., J.-C.Y., J.T.Q., W.-Z.Z., L.T., and X.X. analyzed part of the data. M.J. performed the NPC overexpression experiments. M.J., C.L., X.L., and Q.Y. performed the mouse experiments. X.L. and M.J. performed the dual-luciferase reporter assays. M.Z. performed the FISH experiments. Y.C., F.M., B.Z., J.H.Q., Q.P., Z.L., A.F., X.Q.Z., J.-J.Z., Y.S., B.H., and N.A.A. contributed to data interpretation. C.-Y.L., L.Z., W.D., X.L., and J.Z. wrote the paper. All authors read and approved the final manuscript. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. The PacBio sequence data, Bionano optical map, Illumina whole-genome sequencing reads, Hi-C sequencing data, and scRNA-seq have been deposited in NCBI under the project accession number PRJNA748883 and GEO website under accession number GSE213693 and are publicly available. Genome assembly rheMac10Plus, SNV, and SV datasets in the rhesus macaque population have been posted to the website of RhesusBase (URL: https://rhesusbase.com/download.jsp). The code in this study can be found at Zenodo via URL: https://doi.org/10.5281/zenodo.10695454.

Submitted 19 October 2023
Accepted 1 March 2024
Published 5 April 2024
10.1126/sciadv.adl4600