Article

Integrated analysis of the complete sequence of a macaque genome

https://doi.org/10.1038/s41586-025-08596-w

Received: 7 April 2024

Accepted: 3 January 2025

Published online: 26 February 2025

Check for updates

Shilong Zhang[1,2,25], Ning Xu[3,4,5,6,25], Lianting Fu[1,2,25], Xiangyu Yang[1], Kaiyue Ma[1], Yamei Li[3,4,5], Zikun Yang[1], Zhengtong Li[1], Yu Feng[7], Xinrui Jiang[1], Junmin Han[1], Ruixing Hu[1], Lu Zhang[3,5,8,9], Da Lian[1], Luciana de Gennaro[10], Annalisa Paparella[10], Fedor Ryabov[11], Dan Meng[1], Yaoxi He[5,12,13], Dongya Wu[2,14,15], Chentao Yang[14], Yuxiang Mao[3,4,5,6], Xinyan Bian[3,5], Yong Lu[3,5], Francesca Antonacci[10], Mario Ventura[10], Valery A. Shepelev[16], Karen H. Miga[17], Ivan A. Alexandrov[18], Glennis A. Logsdon[19], Adam M. Phillippy[20], Bing Su[5,12,13,21], Guojie Zhang[2,14,15], Evan E. Eichler[22,23], Qing Lu[1], Yongyong Shi[1,3], Qiang Sun[3,4,5,6] ✉ & Yafei Mao[1,2,24] ✉

The crab-eating macaques (*Macaca fascicularis*) and rhesus macaques (*Macaca mulatta*) are pivotal in biomedical and evolutionary research[1–3]. However, their genomic complexity and interspecies genetic differences remain unclear[4]. Here, we present a complete genome assembly of a crab-eating macaque, revealing 46% fewer segmental duplications and 3.83 times longer centromeres than those of humans[5,6]. We also characterize 93 large-scale genomic differences between macaques and humans at a single-base-pair resolution, highlighting their impact on gene regulation in primate evolution. Using ten long-read macaque genomes, hundreds of short-read macaque genomes and full-length transcriptome data, we identified roughly 2 Mbp of fixed-genetic variants, roughly 240 Mbp of complex loci, 16.76 Mbp genetic differentiation regions and 110 alternative splice events, potentially associated with various phenotypic differences between the two macaque species. In summary, the integrated genetic analysis enhances understanding of lineage-specific phenotypes, adaptation and primate evolution, thereby improving their biomedical applications in human disease research.

Comparative and population genomics substantially improve the understanding of speciation, genetic differences and adaptation in evolutionary biology[3,7–9]. Characterizing the genetic differences, especially major ones between two species, is the basis for revealing lineage-specific genetic changes, and their potential functions in phenotypic differences and adaptation[10–13]. Macaques are one of the naturally widespread non-human primates, well-adapted to various environments and they diverged from humans roughly 25 million years ago (Ma)[1,2,14,15]. Rhesus macaques (*M. mulatta* (MMU)) and crab-eating macaques (*M. fascicularis* (MFA)) are the most commonly used non-human primate models in biomedical research[1–3]. Thus, the characterization of major genetic differences between these two macaque species and between macaques and humans can further optimize the biomedical application of these two important non-human primate models, as well as deepen understanding of the primate evolutionary history.

In this study, we generated a T2T crab-eating macaque genome assembly, 20 haplotype-resolved long-read assemblies from 10 individuals, 151 high-coverage whole-genome sequencing (WGS) and long-read RNA sequencing (RNA-seq) to (1) characterize the T2T macaque genome assembly and its gene annotation, (2) identify the major genetic differences between MFA and MMU with a draft pangenome, and (3) analyse the large-scale genomic differences between humans and macaques (Extended Data Fig. 1).

[1]Bio-X Institutes, Key Laboratory for the Genetics of Developmental and Neuropsychiatric Disorders, Ministry of Education, Shanghai Jiao Tong University, Shanghai, China. [2]Center for Genomic Research, International Institutes of Medicine, Fourth Affiliated Hospital, Zhejiang University, Yiwu, China. [3]Institute of Neuroscience, Center for Excellence in Brain Science and Intelligence Technology, State Key Laboratory of Neuroscience, Chinese Academy of Sciences, Shanghai, China. [4]Shanghai Center for Brain Science and Brain-Inspired Technology, Shanghai, China. [5]National Key Laboratory of Genetic Evolution and Animal Model, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, China. [6]University of Chinese Academy of Sciences, Beijing, China. [7]Chengdu Institute of Biology, Chinese Academy of Sciences, Chengdu, China. [8]School of Life Science and Technology, ShanghaiTech University, Shanghai, China. [9]Lingang Laboratory, Shanghai Center for Brain Science and Brain-Inspired Intelligence Technology, Shanghai, China. [10]Department of Biosciences, Biotechnology and Environment, University of Bari Aldo Moro, Bari, Italy. [11]Masters Program in National Research University Higher School of Economics, Moscow, Russia. [12]State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, China. [13]Yunnan Key Laboratory of Integrative Anthropology, Kunming, China. [14]Center of Evolutionary and Organismal Biology, and Women's Hospital, School of Medicine, Zhejiang University, Hangzhou, China. [15]School of Medicine, Zhejiang University, Hangzhou, China. [16]Institute of Molecular Genetics, Russian Academy of Sciences, Moscow, Russia. [17]University of California Santa Cruz, Santa Cruz, CA, USA. [18]Department of Anatomy and Anthropology and Department of Human Molecular Genetics and Biochemistry, Faculty of Medical and Health Sciences, Tel Aviv University, Tel Aviv, Israel. [19]Department of Genetics, Epigenetics Institute, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. [20]Center for Genomics and Data Science Research, Genome Informatics Section, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA. [21]Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming, China. [22]Department of Genome Sciences, University of Washington School of Medicine, Seattle, WA, USA. [23]Howard Hughes Medical Institute, University of Washington, Seattle, WA, USA. [24]Shanghai Key Laboratory of Embryo Original Diseases, International Peace Maternity and Child Health Hospital, School of Medicine, Shanghai Jiao Tong University, Shanghai, China. [25]These authors contributed equally: Shilong Zhang, Ning Xu, Lianting Fu. ✉e-mail: qsun@ion.ac.cn; yafmao@sjtu.edu.cn

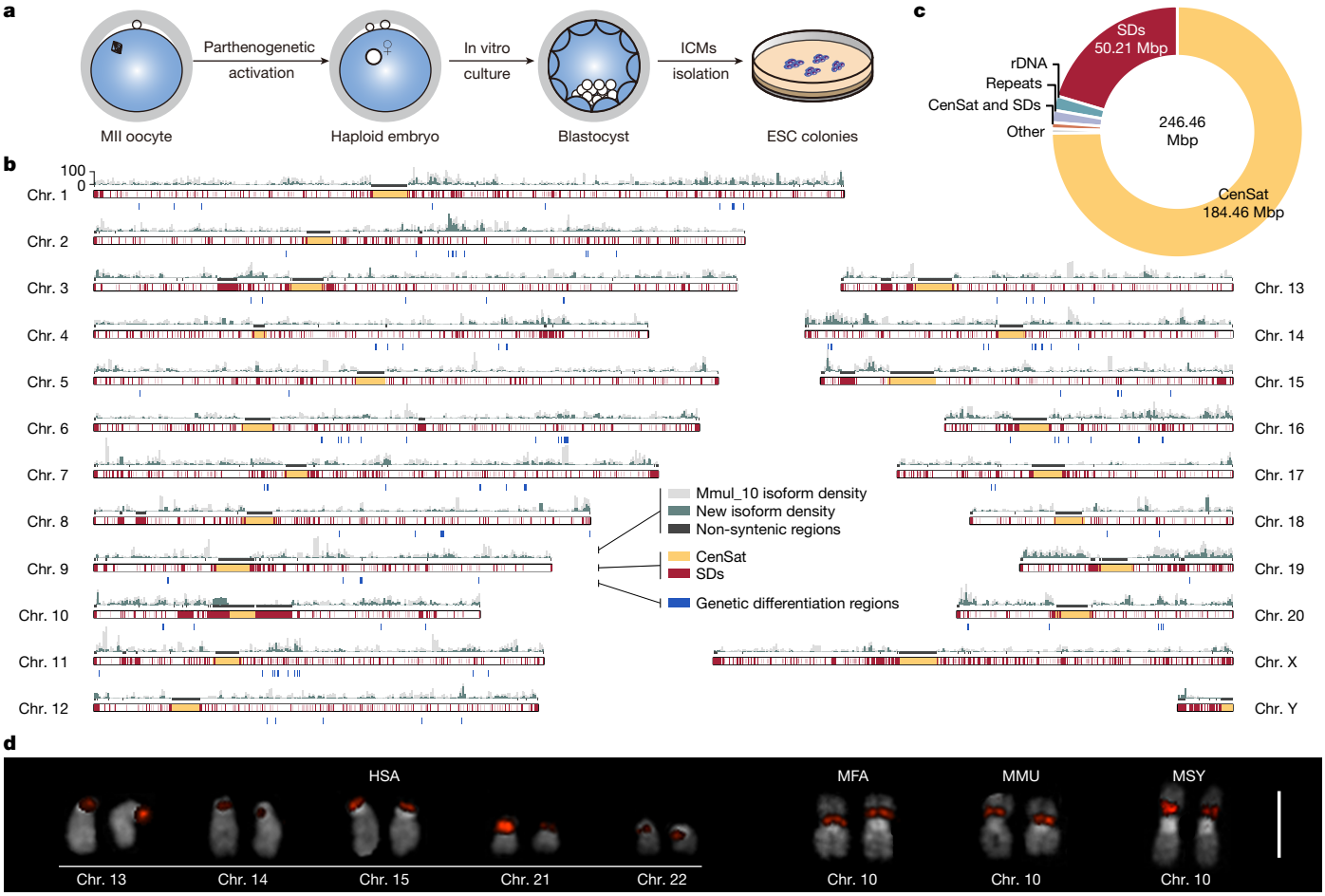


Fig. 1 | Overview of the complete T2T-MFA8 macaque genome. a, Schematic representation of the generation of parthenogenetic embryonic stem cells (ESCs) used for genome assembly. ICMs, inner cell masses. **b**, Ideogram highlighting key features of T2T-MFA8v1.1 assembly. SD, segmental duplication; CenSat, centromeric and pericentromeric satellite. **c**, Pie chart showing the total length and repeat annotation of added sequences. **d**, FISH validation confirming rDNA localization exclusively on macaque chr. 10. Each experiment was repeated three times and ten metaphase spreads with relative fluorochromes were captured for each experiment. Scale bar, 2 µm. HSA, *Homo sapiens*; MSY, *M. sylvanus*.

Genome assembly

To better reconstruct a highly accurate and complete macaque reference genome assembly, we generated an embryo from parthenogenetically activated oocytes of MFA, resulting in a stable diploid state with a 42,XX karyotype that is almost entirely homozygous[16] ('haploid' genome, MFA582-1) (Fig. 1a, Supplementary Figs. 1–3 and Methods). We sequenced MFA582-1 cell line, derived from the embryo's inner cell mass, with several sequencing technologies (Supplementary Fig. 4 and Supplementary Table 1) and used 53× PacBio high-fidelity (HiFi) and 77× Oxford Nanopore Technology (ONT) reads to generate an initial genome assembly with a quality value of 52.63 (ref. 17) (Supplementary Table 2). We then used custom approaches to resolve 18 complex genomic regions and constructed a consensus ribosomal DNA (rDNA) 'morph' (Supplementary Figs. 5 and 6 and Methods). For further refinement, we manually curated 96 large structural errors and 6,629 small errors with Illumina, HiFi and ONT reads[5]. In addition, we assembled a chromosome (chr.) Y without errors from another male MFA (Methods), which includes the pseudoautosomal region (roughly 2.36 Mbp) and alpha satellites (roughly 3.73 Mbp) that were unresolved in the previous macaque assembly (Mmul_10; ref. 1) (Extended Data Fig. 2a). Finally, a complete and high-quality T2T assembly of the macaque genome (T2T-MFA8v1.1, 3.06 gigabase pairs, quality value of 71.27) was reconstructed with an NG50 of 162.13 Mbp (20 + XY) (Fig. 1b,c, Extended Data Table 1, Extended Data Fig. 2b,c and Supplementary Table 3). All these genomic statistics and quality assessment indices were comparable to the current gold standard human genome assembly: T2T-CHM13 (ref. 5) and T2T-CN1 (ref. 18) (Supplementary Note 1 and Supplementary Figs. 7–9).

Previously unexplored genomic regions

rDNA copy number and methylation

The sequence level of rDNA assembly was not available in previous macaque genomes[1,19]. Here, we reconstructed roughly 4.56 Mbp of rDNA regions and match the estimation obtained from the droplet-digital PCR experiments (Supplementary Fig. 6). In contrast to human genomes, there is only one rDNA region in the macaque genome[5]. We used fluorescence in situ hybridization (FISH) to validate this observation in different macaque species, including MFA, MMU and a barbary macaque (*Macaca sylvanus*) (Fig. 1d, Extended Data Fig. 2d, Supplementary Note 1 and Supplementary Fig. 10).

Segmental duplication length and pattern

In the previous macaque assembly (Mmul_10), 58.77% of the segmental duplications (equal to or greater than 1 kilobase pair (kbp) and 90% sequence identity) are unresolved or unlocalized[1] (Supplementary Table 4); here, the T2T-MFA8v1.1 assembly achieves a full resolution of 122.51 Mbp of segmental duplications (roughly 4% of the genome). Of these, 65.23 Mbp are interchromosomal segmental duplications, and 101.97 Mbp are intrachromosomal segmental duplications

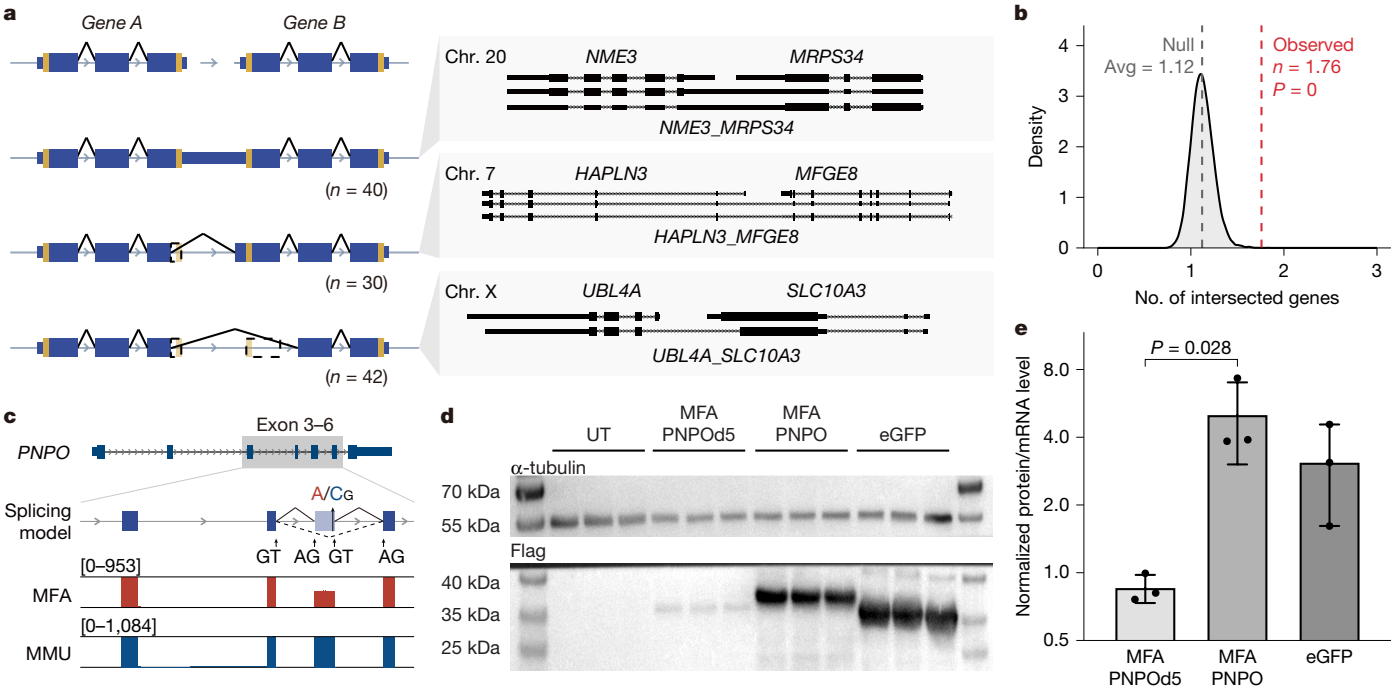


Fig. 2 | Fusion genes and alternative splice sites. a, Schematic illustration of the three types of gene fusion: readthrough ($n = 40$), only stop codon skipping ($n = 30$) and both start and stop codon skipping ($n = 42$). **b**, Gene fusion in a high gene density region. The number of genes adjacent to a fusion gene (red line) is significantly higher than the genome wide average (grey distribution) (one-sided permutation test, empirical $P = 0$). **c**, A fixed-genetic variant between MMU and MFA (CG → AG) influences the splicing pattern of *PNPO*. The bottom two tracks indicate Iso-seq read depth. **d**, Western blot showing reduced protein production of PNPOd5. Each lane is an independent transfection replicate ($n = 3$). UT, untreated. **e**, The mean protein-to-messenger RNA ratio for PNPOd5 is roughly 17% of that of PNPO (one-way analysis of variance with Tukey's multiple comparisons test, $P = 0.028$; error bars, mean ± s.d.). Each dot represents independent transfection replicates ($n = 3$). eGFP, enhanced green fluorescent protein.

(Supplementary Figs. 11 and 12 and Supplementary Table 5). The length of segmental duplications in the macaque genome (T2T-MFA8v1.1, 122.51 Mbp) is 104.85 Mbp less than that in the human genome (T2T-CHM13v2.0, ref. 20, 227.36 Mbp). Meanwhile, the macaque genome shows a higher tendency for segmental duplications to be represented as intrachromosomal states when considering the length of segmental duplications (chi-square test, $P < 2.2 \times 10^{-16}$) (Supplementary Table 6). Human segmental duplications are roughly 1.37 times more enriched in pericentric regions than macaque segmental duplications (chi-square test, $P < 2.2 \times 10^{-16}$), whereas macaque segmental duplications are roughly 1.15 times more enriched in subtelomeric regions (chi-square test, $P < 2.2 \times 10^{-16}$) (Supplementary Tables 7 and 8).

Centromere structure

The active centromeres composed of distinct α-satellite suprachromosomal families (SFs) in macaques have not been characterized on a whole-genome level. Here, we showed that macaque centromeres showcase extensive SF7-derived arrays flanked by smaller[6,21,22], older pieces from SF8–SF13, remnants of ancient centromeres shared with ape lineages[6,21,23,24] (length 3.22 Mbp (chr. 4) to 13.88 Mbp (chr. 15); Extended Data Fig. 2e and Supplementary Figs. 13 and 14). We observed that these regions consist of a homogeneous core formed by S1S2a and S1S2b dimers (roughly 81.4%), alongside smaller and more divergent layers featuring 3-mers, 6-mers and X monomers (roughly 17.1%) potentially from extinct macaque ancestors[25] (Supplementary Fig. 15).

Novel genes and gene fusion

We predicted 21,119 protein-coding genes (92,230 transcripts) and 12,077 non-coding genes (17,840 transcripts) in T2T-MFA8v1.1 (Extended Data Table 1, Supplementary Table 9 and Methods). Compared to the gene annotation of Mmul_10, we added 25,113 transcripts in the previously annotated genes in Mmul_10 and found 204 novel protein-coding genes in T2T-MFA8v1.1, including 92 previously unannotated genes and 112 fusion genes (Extended Data Fig. 3a). Of these, *SLC25A15*, absent from the previous macaque genome assembly (Mmul_10), encodes the mitochondrial ornithine transporter 1 (ORNT1 or ORC1) and plays a role in serine transportation[26] (Extended Data Fig. 3b). Pathogenic mutations in *SLC25A15* are associated with hyperornithinaemia–hyperammonaemia–homocitrullinuria syndrome in humans[27]. Therefore, the improved gene annotation could probably enhance the biomedical applications of macaques.

Full-length RNA-seq has enabled the identification of previously overlooked fusion genes (also known as chimeric RNA) resulting from transcriptional readthrough or intergenic splicing[28]. Fusion genes are well-studied in plants and human cancer cells, but their presence in normal tissues remains controversial[29,30]. Here, we observed 112 fusion genes expressed in healthy macaques (greater than or equal to 5 full-length transcripts) and validated them with short-read RNA-seq data from different MFA individuals[31] (Fig. 2a, Supplementary Fig. 16 and Supplementary Tables 10 and 11). In addition, we found that the number of genes located in the 10 kbp flanking regions of the fusion genes is 1.57 times that of the null distribution (permutation test, empirical $P = 0$) (Fig. 2b and Methods). These findings suggest that these fusion genes are probably concentrated in regions of high gene density (Supplementary Fig. 17).

Comparative long-read transcriptomics

We identified 795 alternative splicing differences spanning 577 genes between MFA and MMU (Supplementary Fig. 18 and Supplementary Tables 9 and 12). Among 209 coding exon-skipping events, we validated 110 exon-skipping differences (validation rate 52.63%). Specifically,

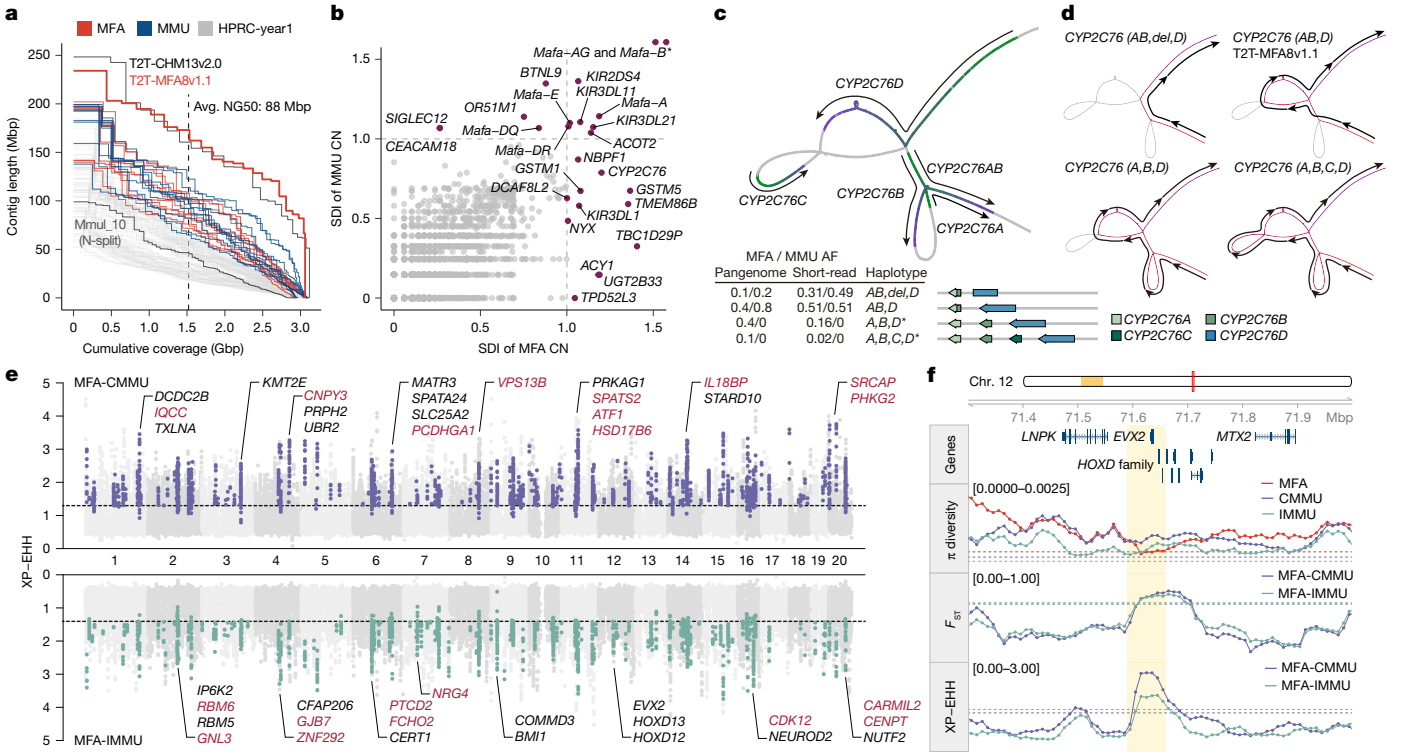


Fig. 3 | A pangenome graph with 20 haplotype-resolved macaque assemblies and genetic differentiation regions between MFA and MMU. **a**, Cumulative genome length distribution of ten haplotype-resolved MFA assemblies (red) and ten MMU assemblies (blue) (average (avg) NG50, 88 Mbp), compared with T2T-MFA8v1.1, T2T-CHM13v2.0, Mmul_10 (split by Ns) and 94 human genome assemblies from HPRC-year1 (light grey). **b**, Copy number (CN) differentiation between MFA and MMU. *Mafa-AG* and *Mafa-B* data points are off the axis. SDI, Shannon diversity index. **c**, Structural haplotypes of *CYP2C76* copies, with green and purple marking the start and end of the gene body, respectively. Frequency statistics for each haplotype are shown below. **d**, Graphical representation of four structural haplotypes of *CYP2C76*, with red and purple representing the start and end of a path, respectively. **e**, Genetic differentiation analysis between MFA and MMU. Manhattan plots show XP-EHH scores for MFA versus CMMU (top) and MFA versus IMMU (bottom) with horizontal dotted lines indicating the top 5% threshold. Differential regions identified as the top 5% XP-EHH, bottom 5% π diversity and top 5% F_{ST} are marked in purple or green. Genes with fixed amino acid changes are marked with deep red. **f**, A genetic differentiation region associated with the *HOXD* gene family.

12 exon-skipping events showed a difference of over 20% in skipping levels between the two species (Methods).

Of particular interest, exon-5 of *PNPO* is alternatively skipped in all tissues of MFA, but not in those of MMU (Fig. 2c), with further evidence from short-read quantification and PCR validation (Extended Data Fig. 3c and Supplementary Fig. 19). A variant with different allele frequencies between MMU and MFA were identified in the *PNPO* exon-5 (allele frequencies for C allele in MMU 99.4% and in MFA 17.7%) (Fig. 2c). We proposed that the C → A alteration probably introduces a new canonical splice acceptor site in *PNPO*, potentially contributing to the observed exon-5 skipping (Fig. 2c and Extended Data Fig. 3d,e). Subsequently, we assessed the relative expression of exon-5 and exon-3 of *PNPO* by RT–qPCR in MFA individuals with the A/A, C/A and C/C genotypes. Our findings showed reduced exon-5 expression in individuals with A/A genotypes compared to those with C/C genotypes (Extended Data Fig. 3d). These results support that the exon-skipping event of *PNPO* is probably related to the C/A difference (T2T-MFA8v1.1, chr. 16:58,789,504-58,789,504). It is noteworthy that the C/C genotype is prevalent at the orthologous site in humans[32] (GRCh38, chr. 17:47,945,984-47,945,984; allele frequencies 0.99999, gnomAD v.4.0).

We assessed the functional impact of exon skipping in MFA PNPO (pyridoxamine 5′-phosphate oxidase) by comparing the relative protein levels (Fig. 2d, Supplementary Fig. 20 and Methods). The exon-5 deletion variant showed a 5.9-fold reduction in relative protein level compared to full-length *PNPO* ($P = 0.028$) (Fig. 2e), indicating that exon-5 deletion may decrease protein stability or reduce expression efficiency. PNPO is a critical enzyme involved in the vitamin B6 metabolism pathway, and recessive mutations in *PNPO* may result in epileptic encephalopathy in humans[33]. Notably, PNPO expression is reduced but not absent in A/A macaques (Extended Data Fig. 3d). The C/A allele frequency difference between MFA and MMU does not suggest a higher epilepsy risk for MFA under normal conditions. However, the potential impact of vitamin B6 deficiency on epilepsy risk between the two species remains uncertain and requires further investigation.

Genetic differences between MFA and MMU

A draft macaque pangenome graph

To better understand the major genetic differences between MFA and MMU, we sequenced and assembled genomes from five genetically unrelated individual of each species (Fig. 3a, Extended Data Fig. 4a, Supplementary Figs. 21 and 22, Supplementary Table 13 and Methods). We next constructed a Minigraph-Cactus[34] pangenome with these data to characterize 50,130,400 single-nucleotide variants (SNVs) (1 bp), 12,654,525 indels (insertion-deletions) (2–49 bp) and 1,063,693 structural variants (greater than or equal to 50 bp) (Extended Data Fig. 4b–f, Supplementary Note 2 and Supplementary Figs. 23 and 24). The number of genetic variants between two macaques is at least three times higher than that between two humans, highlighting a greater level of genetic diversity in macaques[35] (Extended Data Fig. 4c).

SDRs and genetic copy number variation

We identified 370 structurally divergent region (SDR) hotspots (roughly 240 Mbp) on the basis of the bubbles greater than 10 kbp in the pangenome graph (Supplementary Table 14 and Methods). Within these hotspots, we discovered 17 genes, including *Mafa-AG/B*, *CYPs*, *GSTMs*

and *RNASEs*, each represented by at least four gene copies (Extended Data Fig. 4g and Methods). Meanwhile, we genotyped and quantified the extent of gene copy number variation to identify 26 genes that showed significant variability between MFA and MMU (Shannon diversity index equal to or greater than 1) (Fig. 3b). We next analysed the genes that both were contained in the SDR hotspots and showed copy number differences between MFA and MMU. *Mafa-AG* and *Mafa-B* (*HLA-G* and *HLA-B* in humans) in the major histocompatibility complex class I region showed the largest copy number diversity between MFA and MMU (Fig. 3b and Extended Data Fig. 4h). The major histocompatibility complex genomic region is well established for its expansion and diversification in macaque genomes compared to human genomes[36,37] (average length in humans 4.49 Mbp, in macaques 5.96 Mbp; two-sided Mann–Whitney *U*-test, $P = 3.5 \times 10^{-12}$; Supplementary Fig. 25). We identified 12 distinct structural haplotypes (length 5.44 to 6.65 Mbp), including a 407-kbp insertion and a 674-kbp deletion in the maternal haplotype of MFA186ZAI, with respect to T2T-MFA8v1.1, altering 37 genes.

Of particular interest, *CYP2C76* is a monkey-specific gene absent in apes[38,39] (Extended Data Fig. 4i). We newly identified four structural haplotypes (length 21 to 186 kbp) in macaques, wherein the copy number of *CYP2C76* varies from two to four (Fig. 3c,d). Of these, one haplotype with *CYP2C76A*, *CYP2C76B* and *CYP2C76D*, and another with *CYP2C76A*, *CYP2C76B*, *CYP2C76C* and *CYP2C76D*, are both present in MMU but not in MFA (with pangenome allele frequency of 40% and 10%, respectively) (Fig. 3c,d). In addition, the *GSTM* gene cluster, including *GSTM1*, *GSTM2* and *GSTM5*, showed five structural haplotypes in the pangenome graph (length 40 to 87 kbp) (Extended Data Fig. 4j–l). One major haplotype was shared by MFA and MMU, which includes *GSTM1B*, *GSTM2* and *GSTM5A*. The other four minor haplotypes showed a slight allele frequency discrepancy between the two species (Extended Data Fig. 4l, Supplementary Note 3 and Supplementary Tables 15 and 16).

The fixed-genetic variants

To gain a deeper insight into both the fixed and polymorphic genetic variants in MFA and MMU, we used WGS of 94 unrelated MFA, 67 unrelated CMMU (Chinese rhesus macaque) and 88 unrelated IMMU (Indian rhesus macaque) individuals to identify 558,157 fixed SNVs (roughly 558 kbp), 76,997 fixed indels (roughly 284 kbp) and 2,797 fixed structural variants (roughly 1,333 kbp) between MFA and MMU with PanGenie[40] (Extended Data Fig. 5a,b, Supplementary Fig. 26, Supplementary Tables 17–19, Supplementary Note 4 and Methods).

A 1 bp fixed SNV, for example, distinguishes MFA from MMU and apes within *PLA2G3*, which encodes a secretory calcium-dependent phospholipase A2 protein involved in mast cell maturation[41] (Extended Data Fig. 5c). Furthermore, a fixed 72 bp insertion in *EHBP1L1* in MFA, with respect to MMU and apes, led to a 24 amino acid insertion. *EHBP1L1* is involved in apicobasal polarity and cilia length regulation[42] (Extended Data Fig. 5d and Supplementary Fig. 27). These genes are biomedically important: the *PLA2G3* gene, involved in lipid metabolism, affects sperm maturation and fertility when knocked out in mice[43]; *EHBP1L1*, part of the Rab8/10-EHBP1L1-Bin1-dynamin pathway, is essential for cell polarity and its loss leads to lethal anaemia in mice[42]. Further investigation is needed to understand the phenotypic impacts of these genetic differences between the two macaque species.

Genetic differentiation regions

Genetic differentiation regions are often hotspots that harbour fixed-genetic differences between MFA and MMU. Here, we leveraged an integrated approach (π diversity, F_{ST} and XP-EHH) to identify roughly 16.76 Mbp genomic regions appeared to be common to both MFA-CMMU and MFA-IMMU genetic differentiation (Fig. 3e, Supplementary Table 20 and Methods). These genes are also implicated in various biological functions, including glycogen metabolic process ($P = 2.40 \times 10^{-2}$), nervous system development ($P = 3.60 \times 10^{-2}$) and the regulation of stem cell pluripotency[44] ($P = 1.40 \times 10^{-3}$) (Extended Data Fig. 5e–g, Supplementary Table 21 and Supplementary Note 5).

The fixed-genetic variants are enriched by 9.43-fold in 1,248 potential regulatory elements within the genetic differentiation regions, compared to the whole genome (chi-square test, $P < 2.2 \times 10^{-16}$). Of particular interest, we observed that fixed variants in the *HOXD* region were more likely to intersect with regulatory elements (chi-square test, $P < 2.2 \times 10^{-16}$). *HOXD13*, a member of a highly conserved family of transcription factors, is essential in mammalian morphogenesis[45,46]. The recent expression quantitative trait loci analysis in deer mice suggests that the expression difference of *HOXD13* is an important determinant in tail-length variation for environmental adaptation[47]. In our study, we also observed genetic differentiation on the *HOXD13* locus, with six fixed-genetic variants in MFA versus MMU (Fig. 3f and Supplementary Fig. 28). Given the longer tail in MFA than in MMU, we propose that *HOXD13* plays a similar role in macaque tail development. If confirmed, this would represent another example of convergent evolution.

These newly discovered fixed-genetic differences, potentially shaped by selection or genetic drift, are essential resources for future functional assays to illustrate various phenotypic variations between MFA and MMU (for example, tail length, reproduction and others).

Inversion

The Minigraph-Cactus graph's algorithms fail to fully capture diversity and variation of inversion and centromere. Here, we identified 135 polymorphic inversions (length 10 kbp to 4 Mbp) between MFA and MMU among ten macaques (Extended Data Fig. 5h, Supplementary Table 22 and Methods). To explore the potential biological implications of these inversions, we identified 29 out of 909 protein-coding genes within the 500 kbp flanking regions of inversion breakpoints showing differential expression (Extended Data Fig. 5i and Supplementary Tables 23 and 24).

Centromere

We also investigated whether macaque α-satellite arrays vary between MFA and MMU. Comparing the S1S2 α-satellite arrays in our 20 haplotype-resolved genome assemblies, 139 of 420 centromeric regions (33.10%) were successfully assembled, with lengths varying from 2.56 to 17.05 Mbp (Supplementary Table 25). Both MFA and MMU α-satellite arrays, similar to apes, show sequence similarity block structures with distinct centromere dip regions[24,39] (Extended Data Fig. 6a,b). Furthermore, the α-satellite arrays of chr. 1 are conserved between MFA and MMU, with the ancestral macaque-specific centromere layers preserved in both species (Extended Data Fig. 6a,b). However, the conservation or synteny degree of α-satellite higher-order repeats between two humans (T2T-CHM13 and CHM1) varies considerably[21]. Moreover, we found that the lengths of α-satellite arrays on orthologous chromosomes between MFA and MMU are not significantly different; however, they are, on average, roughly 3.83 times larger than those in humans[6,21] (two-sided Mann–Whitney *U*-test, $P = 4.20 \times 10^{-12}$ for chr. 1 between MFA and *H. sapiens*, and $P = 0.07$ for chr. 1 between MFA and MMU) (Extended Data Fig. 6c). The phylogenetic analysis of 600 randomly chosen MFA and MMU α-satellite monomers from five chromosomes reveals that the S1 and S2 α-satellites cluster into distinct clades whereas S2 splits into separate subclades, S2a and S2b, and in each clade the MFA and MMU monomers are mixed showing no species specificity (Extended Data Fig. 6d). In addition, we observed chromosome-specific monomers[48,49] (Extended Data Fig. 6e) and S1S2 dimers in most MFA centromeres (Supplementary Note 6).

Complete human-macaque genome comparison

We identified 171 syntenic blocks between T2T-CHM13v2.0 and T2T-MFA8v1.1 (more than 100 kbp) with the alignment cluster-based approach LSGvar and manual curation (Supplementary Table 26

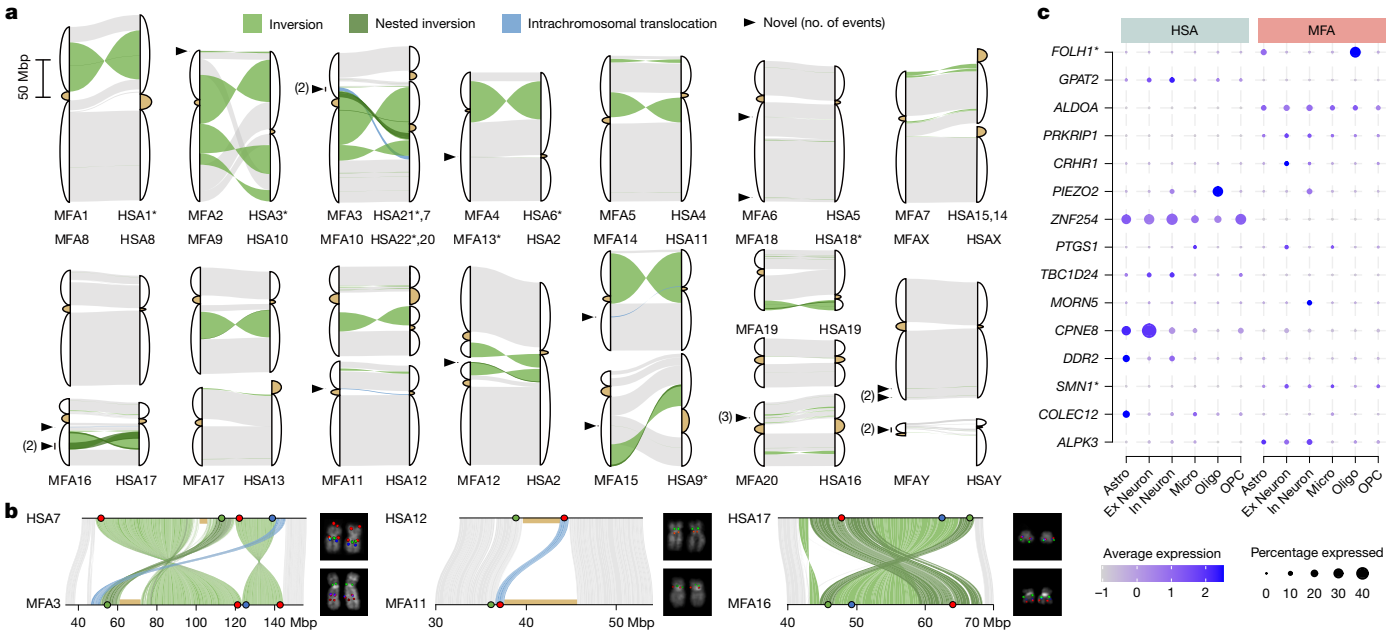


Fig. 4 | Genomic differences between humans and macaques. a, Chromosomal rearrangements between T2T-MFA8v1.1 (MFA) and T2T-CHM13v2.0 (HSA). Macaque and human chromosomes are listed on the left and right, respectively (inversions in green, nested inversions in dark green and intrachromosomal translocations in blue). Newly identified rearrangements ($n = 21$) are marked with triangles, with numbers indicating the count of novel events at each location ($n \geq 2$). An asterisk (*) denotes the inverted orientation of a chromosome strand (q-arm to p-arm). **b**, FISH validation of three newly reported large-scale rearrangements between humans and macaques. Each experiment was repeated three times and ten metaphase spreads with relative fluorochromes were captured for each experiment. **c**, Percentage and expression of genes expressed in different cellular types of the prefrontal cortex in humans and macaques. The genes within segmental duplications are marked by an asterisk. The macaque *FOLH1* and human *FOLH1B* are positional orthologous, indicated as *FOLH1*. Astro, astrocyte; Ex Neuron, excitatory neuron; In Neuron, inhibitory neuron; Micro, microglia; Oligo, oligodendrocyte; OPC, oligodendrocyte precursor cell.

and Methods). These syntenic blocks are separated by 11 centromere repositions, 78 inversions and 4 intrachromosomal translocations in primate evolution (Fig. 4a, Supplementary Figs. 29 and 30 and Supplementary Tables 27 and 28). Among these, 11 centromere repositions, 67 inversions and 4 intrachromosomal translocations represent fixed genomic differences between humans and macaques (Fig. 4a and Supplementary Tables 27 and 28). Notably, 21 large-scale genomic rearrangements (1 centromere reposition, 16 inversions and 4 translocations) are newly reported[50–52] (Supplementary Tables 27 and 28). We selected three regions for FISH validation and the results were consistent with our computational analysis (Fig. 4b and Supplementary Fig. 31).

A total of 1,462 protein-coding genes are included in the 93 large-scale genomic rearrangements with 500 kbp flanking regions of breakpoints. Among these, the expression of 482 genes is significantly different in the brains of humans and macaques[31] (Supplementary Fig. 32 and Supplementary Table 29). In addition to the differentially expressed gene analysis on tissue bulk level, here, incorporating the single-cell RNA-seq data of the dorsolateral prefrontal cortex in humans[53] and macaques[54], we identified 414 genes that show both different expression levels in brain tissues and different expression patterns in cell types, including *FOLH1B*, *APCDD1* and *PIEZO2* (Fig. 4c and Supplementary Table 30). These genes are enriched in DNA repair ($P = 7.29 \times 10^{-4}$), regulation of G protein-coupled receptor signalling pathway ($P = 2.26 \times 10^{-3}$), and intermembrane lipid transfer[44] ($P = 2.75 \times 10^{-2}$) (Supplementary Table 31). Out of 414 genes, 28 are duplicated genes in primate evolution.

Among these genes, *FOLH1* and *FOLH1B*, which encode the glutamate carboxypeptidase II (GCPII) involved in the glutamate regulation in neural activity[55], are of particular interest, as mutations in these genes are linked to intellectual disability[56]. *FOLH1* and *FOLH1B* were duplicated in the ancestor of African great apes at 10.55 Ma (95% CI: 9.30–11.81 Ma). The *FOLH1B* gene in African great apes and *FOLH1* in macaques are orthologs derived from a common ancestor[24,39] (Fig. 5a,b and Supplementary Figs. 33 and 34). Furthermore, *FOLH1* and *FOLH1B* contain the same open reading frame in African great apes, except in humans, in whom a 1,393 bp human-specific deletion results in 57-amino-acid depletion in human *FOLH1B* (allele frequency of 1, haplotype-resolved human genomes from Human Pangenome Reference Consortium (HPRC)-year1; ref. 37) (Supplementary Fig. 35).

Human *FOLH1* showed high and specific expression in oligodendrocytes (roughly 85.3% of 4,005 cells), whereas human *FOLH1B* showed minimal expression in brain cells (observed in only 188 cells) (Fig. 5c–h). Next, we explored whether the roughly 1.4 kbp human-specific deletion plays a crucial role in the non-expression of *FOLH1B* in brain cells. We confirmed the existence of three regulatory elements in the depletion region using single-cell epigenomic data and ENCODE database from humans[57] (Fig. 5i). Thus, we propose that the human-specific deletion alters *FOLH1B* transcription in humans.

In addition, macaque *FOLH1* (syntenic to human *FOLH1B*) was expressed in oligodendrocytes (roughly 64.7% of 4,658 cells), astrocytes (roughly 18.7% of 4,658 cells), and excitatory neurons (roughly 12.6% of 4,658 cells) (Fig. 5h and Supplementary Table 32). The expression patterns of *FOLH1* across different brain cell types between macaques and humans is significantly different (chi-square test, $P < 2.2 \times 10^{-16}$; Fig. 5h and Supplementary Table 32). We then analysed Hi-C data from human and macaque PFCs, revealing potential contacts within two sub-TADs in the macaque *FOLH1* genomic region (Extended Data Fig. 7). By contrast, no obvious sub-TADs were observed in the human *FOLH1* region (Extended Data Fig. 7). These differences suggest that the gene expression variations across different cell types may be driven by the translocated duplication of *FOLH1* between humans and macaques, which in turn altered the chromatin architecture and regulatory landscape (Extended Data Fig. 8a,b). Palindromic sequences flanking the *FOLH1* duplication region were identified in apes but not in marmosets

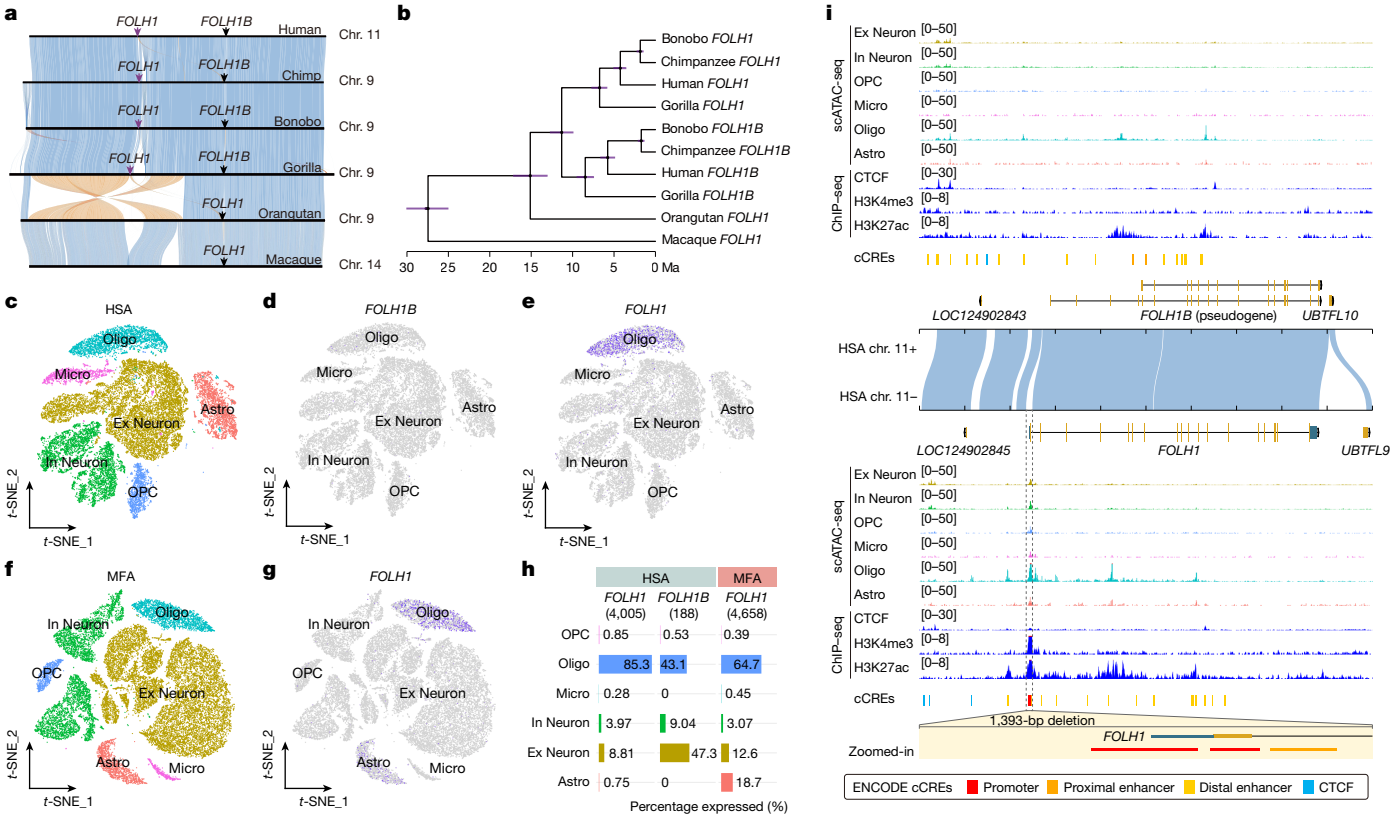


Fig. 5 | Evolutionary divergence in human *FOLH1* and *FOLH1B*. **a**, Syntenic comparison of T2T-MFA8v1.1 chr. 14 (human chr. 11), illustrating the origin of the *FOLH1* gene family. **b**, Phylogenetic tree showing the duplication of *FOLH1* and *FOLH1B* in the ancestor of African great apes (roughly 10.55 Ma). **c**–**g**, *t*-SNE visualization of cell types expressing *FOLH1* and *FOLH1B* in humans and macaques. **h**, Expression proportions of *FOLH1* and *FOLH1B* across cell types, with the total number of expressing cells shown in brackets. **i**, Syntenic comparison and epigenetic profiles of human *FOLH1* and *FOLH1B*, showing a 1,393 bp deletion in *FOLH1B*. A detailed view shows the depletion of *FOLH1* exon-1 and three candidate *cis*-regulatory elements (cCREs) in human *FOLH1B*. CTCF, CCCTC-binding factor.

or macaques[39] (Extended Data Fig. 8b). Given their predisposition to genomic instability and DNA double-strand breaks, we propose that these sequences in the common ancestor of great apes contributed to regional instability, driving the duplication event in primate evolution (Extended Data Fig. 8c). In summary, the evolutionary history and gene expression analysis of *FOLH1* and *FOLH1B* reveal how genomic rearrangements, duplications, and genetic variations alter the gene structure and its regulation in primate evolution. A similar evolutionary mechanism is observed for the inversions associated with *APCDD1* and *PIEZO2* (Extended Data Fig. 9 and Supplementary Note 7).

Discussion

Macaques play crucial roles in biomedical research and the study of primate evolution[1,2,14,50]. The first T2T macaque genome assembly (T2T-MFA8v1.1) shows its unique genomic structure. First, in great apes, segmental duplications are larger and enriched in pericentric regions, whereas in macaques, the pattern is the opposite[39] (Supplementary Tables 5–8). Second, the rDNA array in macaques is only found in chr. 10, in contrast to humans, in whom they are located in five chromosomes[5] (chr. 13, chr. 14, chr. 15, chr. 21 and chr. 22) (Fig. 1d). The chromosome fissions and rDNA dispersals in primate evolution result in smaller acrocentric chromosomes in humans and are suspected to be a genetic basis for human diseases such as Down syndrome. Third, the length and composition of the α-satellite arrays differ between macaques and apes[6,21,22] (Supplementary Note 6). The α-satellite arrays are relatively conserved both within macaque species and between macaque species, whereas they exhibit greater variability in humans[6,21] and nonhuman ape species[24] (Extended Data Fig. 6). We speculate that the α-satellite regions were independently seeded and expanded in macaques, as suggested by the suprachromosomal family comparison and their evolutionary trajectories[21] (Extended Data Fig. 6). Finally, 93 large-scale genomic differences between macaques and humans have been revealed at single-base resolution, most of which are associated with centromere repositions, inversions and translocations[50–52] (Fig. 4). The detailed analysis of *FOLH1* and *FOLH1B* highlights the intricate evolutionary fate of duplicated genes and the regulatory changes driven by both large- and small-scale genomic structural changes (Fig. 5 and Extended Data Figs. 7 and 8). Similarly, in *ARHGAP11* and *ARHGAP11B*, a duplication event coupled with a splice variation has led to their functional divergence[58]. Previous studies primarily focused on how structural variants affect gene expression levels. However, our study, alongside the cases of *APCDD1* and *PIEZO2* (Extended Data Fig. 9), proposes the concept that these genomic differences influence not only gene expression levels but also the cellular expression patterns of genes. These efforts enhance our understanding of gene functional alterations caused by large-scale rearrangements between humans and macaques, which is crucial for comprehending lineage-specific phenotype and human diseases[1,58,59].

The Minigraph-Cactus pangenome graph not only contributes to our understanding of the fixed-genetic differences between the two macaque species but also helps explain the haplotype diversity of the 240-Mbp complex loci, which are usually associated with immune diversity, in MFA and MMU (Fig. 3 and Extended Data Fig. 4). In addition, the analysis of alternative splicing differences provides new insights into the genetic differences at the transcriptome level between MFA

and MMU (Fig. 2). Moreover, when integrated with our macaque fixed-genetic-difference atlas (Fig. 3), we found that only 30.12% of genes in the genetic differentiation regions contain variants resulting in amino acid changes (chi-square test, $P = 1.25 \times 10^{-27}$) (Supplementary Table 33). This suggests that most phenotypic differences may arise from gene regulation rather than alterations in protein sequences, as has been debated[60]. These genetic differences and diversity provide a roadmap for further understanding the functional biological differences of the two macaques, which is important for future biomedical research.

In summary, this study both establishes a robust genetic foundation for the macaque biomedical models and also improves our understanding of primate evolution, human diseases and lineage-specific adaptation.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-025-08596-w.

1. Warren, W. C. et al. Sequence diversity analyses of an improved rhesus macaque genome enhance its biomedical utility. *Science* https://doi.org/10.1126/science.abc6617 (2020).
2. Rhesus Macaque Genome Sequencing and Analysis Consortium et al. Evolutionary and biomedical insights from the rhesus macaque genome. *Science* **316**, 222–234 (2007).
3. Rogers, J., Gibbs, R. A., Rogers, J. & Gibbs, R. A. Comparative primate genomics: emerging patterns of genome content and dynamics. *Nat. Rev. Genet.* **15**, 347–359 (2014).
4. Haus, T. et al. Genome typing of nonhuman primate models: implications for biomedical research. *Trends Genet.* **30**, 482–487 (2014).
5. Nurk, S. et al. The complete sequence of a human genome. *Science* **376**, 44–53 (2022).
6. Altemose, N. et al. Complete genomic and epigenetic maps of human centromeres. *Science* **376**, 4178 (2022).
7. Eichler, E. E. Genetic variation, comparative genomics, and the diagnosis of disease. *N. Engl. J. Med.* **381**, 64–74 (2019).
8. Seehausen, O. et al. Genomics and the origin of species. *Nat. Rev. Genet.* **15**, 176–192 (2014).
9. Zoonomia Consortium, A comparative genomics multitool for scientific discovery and conservation. *Nature* **587**, 240–245 (2020).
10. Mao, Y. & Zhang, G. A complete, telomere-to-telomere human genome sequence presents new opportunities for evolutionary genomics. *Nat. Methods* https://doi.org/10.1038/s41592-022-01512-4 (2022).
11. Mao, Y. et al. Structurally divergent and recurrently mutated regions of primate genomes. *Cell* **187**, 1547–1562 (2024).
12. Kuhlwilm, M., Boeckx, C., Kuhlwilm, M. & Boeckx, C. A catalog of single nucleotide changes distinguishing modern humans from archaic hominins. *Sci. Rep.* **9**, 8463 (2019).
13. Zeberg, H., Jakobsson, M. & Pääbo, S. The genetic changes that shaped Neandertals, Denisovans, and modern humans. *Cell* **187**, 1047–1058 (2024).
14. He, Y. et al. Long-read assembly of the Chinese rhesus macaque genome and identification of ape-specific structural variants. *Nat. Commun.* **10**, 4233 (2019).
15. Cooper, E. B. et al. The Natural History of Model Organisms: the rhesus macaque as a success story of the Anthropocene. *eLife* https://doi.org/10.7554/eLife.78169 (2022).
16. Yang, H. et al. Generation of haploid embryonic stem cells from *Macaca fascicularis* monkey parthenotes. *Cell Res.* **23**, 1187–1200 (2013).
17. Rautiainen, M. et al. Telomere-to-telomere assembly of diploid chromosomes with Verkko. *Nat. Biotechnol.* **41**, 1474–1482 (2023).
18. Yang, C. et al. The complete and fully-phased diploid genome of a male Han Chinese. *Cell Res.* **33**, 745–761 (2023).
19. Jayakumar, V. et al. Chromosomal-scale de novo genome assemblies of cynomolgus macaque and common marmoset. *Sci. Data* **8**, 159 (2021).
20. Vollger, M. R. et al. Segmental duplications and their variation in a complete human genome. *Science* **376**, 6965 (2022).
21. Logsdon, G. A. et al. The variation and evolution of complete human centromeres. *Nature* **629**, 136–145 (2024).
22. Logsdon, G. A. et al. The structure, function and evolution of a complete human chromosome 8. *Nature* **593**, 101–107 (2021).
23. Shepelev, V. A., Alexandrov, A. A., Yurov, Y. B. & Alexandrov, I. A. The evolutionary origin of man can be traced in the layers of defunct ancestral alpha satellites flanking the active centromeres of human chromosomes. *PLoS Genet.* **5**, 1000641 (2009).
24. Makova, K. D. et al. The complete sequence and comparative analysis of ape sex chromosomes. *Nature* **630**, 401–411 (2024).
25. Alexandrov, I., Kazakov, A., Tumeneva, I., Shepelev, V. & Yurov, Y. Alpha-satellite DNA of primates: old and new families. *Chromosoma* **110**, 253–266 (2001).
26. Papalazarou, V. et al. Phenotypic profiling of solute carriers characterizes serine transport in cancer. *Nat. Metab.* **5**, 2148–2168 (2023).
27. Salvi, S., Dionisi-Vici, C., Bertini, E., Verardo, M. & Santorelli, F. M. Seven novel mutations in the ORNT1 gene (SLC25A15) in patients with hyperornithinemia, hyperammonemia, and homocitrullinuria syndrome. *Human Mutation* **18**, 460 (2001).
28. Wu, H., Li, X. & Li, H. Gene fusions and chimeric RNAs, and their implications in cancer. *Genes Dis.* **6**, 385–390 (2019).
29. Zhang, Y. et al. Readthrough events in plants reveal plasticity of stop codons. *Cell Rep.* **43**, 113723 (2024).
30. Sahoo, S. et al. Identification and functional characterization of mRNAs that exhibit stop codon readthrough in *Arabidopsis thaliana*. *J. Biol. Chem.* **298**, 102173 (2022).
31. Mao, Y.-X. et al. Comparative transcriptome analysis between rhesus macaques (*Macaca mulatta*) and crab-eating macaques (*M. fascicularis*). *Zool. Res.* https://doi.org/10.24272/j.issn.2095-8137.2023.322 (2023).
32. Chen, S. et al. A genomic mutational constraint map using variation in 76,156 human genomes. *Nature* **625**, 92–100 (2023).
33. Barile, A. et al. Molecular characterization of pyridoxine 5′-phosphate oxidase and its pathogenic forms associated with neonatal epileptic encephalopathy. *Sci. Rep.* **10**, 13621 (2020).
34. Hickey, G. et al. Pangenome graph construction from genome alignments with Minigraph-Cactus. *Nat. Biotechnol.* **42**, 663–673 (2023).
35. Xue, C. et al. The population genomics of rhesus macaques (*Macaca mulatta*) based on whole-genome sequences. *Genome Res.* **26**, 1651–1662 (2016).
36. Karl, J. A. et al. Complete sequencing of a cynomolgus macaque major histocompatibility complex haplotype. *Genome Res.* **33**, 448–462 (2023).
37. Liao, W.-W. et al. A draft human pangenome reference. *Nature* **617**, 312–324 (2023).
38. Uno, Y., Fujino, H., Kito, G., Kamataki, T. & Nagata, R. CYP2C76, a novel cytochrome P450 in cynomolgus monkey, is a major CYP2C in liver, metabolizing tolbutamide and testosterone. *Mol. Pharmacol.* **70**, 477–486 (2006).
39. Yoo, D. et al. Complete sequencing of ape genomes. Preprint at *bioRxiv* https://doi.org/10.1101/2024.07.31.605654 (2024).
40. Ebler, J. et al. Pangenome-based genome inference allows efficient and accurate genotyping across a wide spectrum of variant classes. *Nat. Genet.* **54**, 518–525 (2022).
41. Taketomi, Y. et al. Mast cell maturation is driven via a group III phospholipase A2-prostaglandin D2–DP1 receptor paracrine axis. *Nat. Immunol.* **14**, 554–563 (2013).
42. Wu, J. et al. EHBP1L1, an apicobasal polarity regulator, is critical for nuclear polarization during enucleation of erythroblasts. *Blood Adv.* **7**, 3382–3394 (2023).
43. Sato, H. et al. Group III secreted phospholipase A2 regulates epididymal sperm maturation and fertility in mice. *J. Clin. Invest.* **120**, 1400–1414 (2010).
44. Sherman, B. T. et al. DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). *Nucleic Acids Res.* **50**, 216 (2022).
45. Darbellay, F. et al. The constrained architecture of mammalian Hox gene clusters. *Proc. Natl Acad. Sci. USA* **116**, 13424–13433 (2019).
46. Dollé, P. et al. Disruption of the Hoxd-13 gene induces localized heterochrony leading to mice with neotenic limbs. *Cell* **75**, 431–441 (1993).
47. Kingsley, E. P. et al. Adaptive tail-length evolution in deer mice is associated with differential Hoxd13 expression in early development. *Nat. Ecol. Evol.* **8**, 791–805 (2024).
48. Willard, H. F. Chromosome-specific organization of human alpha satellite DNA. *Am. J. Hum. Genet.* **37**, 524–532 (1985).
49. O'Keefe, C. L. & Matera, A. G. Alpha satellite DNA variant-specific oligoprobes differing by a single base can distinguish chromosome 15 homologs. *Genome Res.* https://doi.org/10.1101/gr.10.9.1342 (2000).
50. Maggiolini, F. et al. Single-cell strand sequencing of a macaque genome reveals multiple nested inversions and breakpoint reuse during primate evolution. *Genome Res.* **30**, 1680–1693 (2020).
51. Ventura, M. et al. Evolutionary formation of new centromeres in macaque. *Science* **316**, 243–246 (2007).
52. Rocchi, M., Archidiacono, N., Schempp, W., Capozzi, O. & Stanyon, R. Centromere repositioning in mammals. *Heredity* **108**, 59–67 (2011).
53. Ma, S. et al. Molecular and cellular evolution of the primate dorsolateral prefrontal cortex. *Science* https://doi.org/10.1126/science.abo7257 (2022).
54. Wu, J. et al. Integrating spatial and single-nucleus transcriptomic data elucidates microglial-specific responses in female cynomolgus macaques with depressive-like behaviors. *Nat. Neurosci.* **26**, 1352–1364 (2023).
55. Blakely, R. D., Robinson, M. B., Thompson, R. C. & Coyle, J. T. Hydrolysis of the brain dipeptide *N*-acetyl-L-aspartyl-L-glutamate: subcellular and regional distribution, ontogeny, and the effect of lesions on *N*-acetylated-α-linked acidic dipeptidase activity. *J. Neurochem.* **50**, 1200–1209 (1988).
56. Rahn, K. A. et al. Inhibition of glutamate varboxypeptidase II (GCPII) activity as a treatment for cognitive impairment in multiple sclerosis. *Proc. Natl Acad. Sci. USA* **109**, 20101–20106 (2012).
57. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* https://doi.org/10.1038/nature11247 (2012).
58. Florio, M. et al. Human-specific gene ARHGAP11B promotes basal progenitor amplification and neocortex expansion. *Science* **347**, 1465–1470 (2015).
59. Ma, K., Yang, X. & Mao, Y. Advancing evolutionary medicine with complete primate genomes and advanced biotechnologies. *Trends Genet.* https://doi.org/10.1016/j.tig.2024.11.001 (2024).
60. King, M.-C. & Wilson, A. C. Evolution at two levels in humans and chimpanzees. *Science* **188**, 107–116 (1975).

Methods

The usage of research animals in this study underwent evaluation and approval by the Primate Life Sciences Ethics Committee of the Centre for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences (ION-2019043R03). In line with reduction principles, surgical procedures were conducted with utmost care to minimize pain and discomfort. Painless euthanasia was carried out following appropriate anaesthesia and euthanasia protocols. Before euthanasia by means of exsanguination, monkeys were administered Zoletil 50 injection at a dose of 25 mg kg^{-1} to ensure proper sedation.

We generated a parthenogenesis cell line, MFA582-1, from an MFA, from which we produced 53× HiFi, 77× ONT, 32× Illumina WGS and 156× Hi-C data. Using these data, we assembled a complete genome (T2T-MFA8v1.1) using Hifiasm[61,62], Verkko[17] and our SUNK-based pipeline. The genome assembly was annotated with 15.27 million full-length, non-concatemer reads of MFA across 15 different tissues, including six brain tissues. Then 14.82 million full-length, non-concatemer reads from 15 tissues of MMU were also sequenced, and we used these data to perform a comparative transcriptome analysis with IsoSeq3 and SQANTI3 (ref. 63). Segmental duplication analysis using SEDEF[64] and protein structure prediction using AlphaFold2 (ref. 65) were performed. We sequenced ten macaque individuals using HiFi and ONT and assembled their genomes with Hifiasm[61,62]. A draft pangenome graph was generated using Minigraph-Cactus[34]. Genetic variations, including SNVs, indels and structural variants, were decomposed by vcfwave[66]. Another genetic variation callset was generated from the callers with DeepVariant[67,68], PBSV[69], PAV[70], SVIM[71] and SVIM-asm[72]. PanGenie[40] and fastCN[73] were used to genotype the structural variants and copy number variants from a diversity panel of 249 unrelated macaque individuals. A cluster-based approach (LSGvar) was developed to characterize large-scale rearrangements between macaques and humans. These large-scale genomic differences were genotyped using 94 HPRC genomes[37] and 20 macaque genomes. We used DESeq2 (ref. 74), Limma[75], Cell Ranger[76] and Seurat[77,78] to investigate gene expression and cell type differentiation of genes associated with these large genomic difference regions. Genetic differentiation regions were identified using three independent approaches: π diversity[79], F_{ST} (ref. 80) and XP-EHH[81]. More detailed information can be found in the Supplementary Information.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The genome assemblies used in this study, including T2T-CHM13v2.0 (GCF_009914755.1), Mmul_10 (GCF_003339765.1), MFA1912RKSv2 (GCF_012559485.2), rheMacS_1.0 (GCA_008058575.1) and Macaca_fascicularis_6.0 (GCA_011100615.1), are available from the National Centre for Biotechnology Information (NCBI) Genome. Previously published sequencing data used in this study, including PRJNA1004471 (ref. 31), PRJNA251548 (ref. 1) PRJNA345528 (ref. 82), PRJNA832687 (ref. 54), PRJNA854879 (ref. 53), PRJNA882074 (ref. 83), PRJNA953340 (ref. 84) and PRJCA018217 (ref. 85), are available from the NCBI Sequence Read Archive or National Genomics Data Center Genome Sequence Archive. The chromatin immunoprecipitation sequencing data of human brain used in this study are available from the ENCODE database (ref. 57). The raw Illumina, PacBio HiFi, ONT and Hi-C data of T2T-MFA8 are deposited in NCBI under BioProject accession number PRJNA1037719. The raw PacBio HiFi, ONT and Hi-C data of ten macaque individuals are deposited in NCBI under BioProject accession number PRJNA1041301. The Iso-seq data are deposited under NCBI BioProject accession number PRJNA1041301. The Illumina sequences of 151 WGS macaque genomes are deposited in NCBI under BioProject accession number PRJNA1041301. The T2T-MFA8 genome assembly is deposited in NCBI GenBank under accession number GCF_037993035. T2T-MFA8 assembly, annotations and the UCSC track hub are available at GitHub (https://github.com/zhang-shilong/T2T-MFA8).

Code availability

Custom scripts used in this study are available via Zenodo at https://doi.org/10.5281/zenodo.14220081 (ref. 86).

61. Cheng, H. et al. Haplotype-resolved assembly of diploid genomes without parental data. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-022-01261-x (2022).
62. Cheng, H. et al. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods* https://doi.org/10.1038/s41592-020-01056-5 (2021).
63. Pardo-Palacios, F. J. et al. SQANTI3: curation of long-read transcriptomes for accurate identification of known and novel isoforms. *Nat. Methods* https://doi.org/10.1038/s41592-024-02229-2 (2024).
64. Numanagić, I. et al. Fast characterization of segmental duplications in genome assemblies. *Bioinformatics* https://doi.org/10.1093/bioinformatics/bty586 (2018).
65. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* https://doi.org/10.1038/s41586-021-03819-2 (2021).
66. Garrison, E., Kronenberg, Z. N., Dawson, E. T., Pedersen, B. S. & Prins, P. A spectrum of free software tools for processing the VCF variant call format: vcflib, bio-vcf, cyvcf2, hts-nim and slivar. *PLOS Comput. Biol.* https://doi.org/10.1371/journal.pcbi.1009123 (2022).
67. Yun, T. et al. Accurate, scalable cohort variant calls using DeepVariant and GLnexus. *Bioinformatics* https://doi.org/10.1093/bioinformatics/btaa1081 (2021).
68. Poplin, R. et al. A universal SNP and small-indel variant caller using deep neural networks. *Nat. Biotechnol.* https://doi.org/10.1038/nbt.4235 (2018).
69. Wenger, A. M. et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-019-0217-9 (2019).
70. Ebert, P. et al. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* https://doi.org/10.1126/science.abf7117 (2021).
71. Heller, D. & Vingron, M. SVIM: structural variant identification using mapped long reads. *Bioinformatics* https://doi.org/10.1093/bioinformatics/btz041 (2019).
72. Heller, D. & Vingron, M. SVIM-asm: structural variant detection from haploid and diploid genome assemblies. *Bioinformatics* https://doi.org/10.1093/bioinformatics/btaa1034 (2021).
73. Vollger, M. R. & Harvey, W. T. mrvollger/fastCN-smk: v0.2. *Zenodo* https://doi.org/10.5281/zenodo.8136270 (2023).
74. Love, M. I. et al. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* https://doi.org/10.1186/s13059-014-0550-8 (2014).
75. Ritchie, M. E. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* https://doi.org/10.1093/nar/gkv007 (2015).
76. Zheng, G. X. Y. et al. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* https://doi.org/10.1038/ncomms14049 (2017).
77. Hao, Y. et al. Integrated analysis of multimodal single-cell data. *Cell* https://doi.org/10.1016/j.cell.2021.04.048 (2021).
78. Hao, Y. et al. Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-023-01767-y (2023).
79. Nei, M., Li, W. H., Nei, M. & Li, W. H. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc. Natl Acad. Sci. USA* https://doi.org/10.1073/pnas.76.10.5269 (1979).
80. Bhatia, G., Patterson, N., Sankararaman, S. & Price, A. L. Estimating and interpreting FST: the impact of rare variants. *Genome Res.* https://doi.org/10.1101/gr.154831.113 (2013).
81. Sabeti, P. C. et al. Genome-wide detection and characterization of positive selection in human populations. *Nature* https://doi.org/10.1038/nature06250 (2007).
82. Liu, Z. et al. Population genomics of wild Chinese rhesus macaques reveals a dynamic demographic history and local adaptation, with implications for biomedical research. *Gigascience* https://doi.org/10.1093/gigascience/giy106 (2018).
83. Ding, W. et al. Adaptive functions of structural variants in human brain development. *Sci. Adv.* https://doi.org/10.1126/sciadv.adl4600 (2024).
84. Zemke, N. R. et al. Conserved and divergent gene regulatory programs of the mammalian neocortex. *Nature* https://doi.org/10.1038/s41586-023-06819-6 (2023).
85. Ning, C. et al. Epigenomic landscapes during prefrontal cortex development and aging in rhesus. *Natl Sci. Rev.* https://doi.org/10.1093/nsr/nwae213 (2024).
86. Zhang, S. Integrated analysis of the complete sequence of a macaque genome (code archive). *Zenodo* https://doi.org/10.5281/zenodo.14220081 (2024).

Acknowledgements We thank the staff of the Non-human Primate Facility of the Centre for Excellence in Brain Science and Intelligence Technology for their assistance in animal care. We thank the HPRC and Primate T2T Consortium for providing the long-read human and great ape genome assemblies. We thank T. Brown for proofreading and editing the manuscript. We thank K. Makova for providing valuable comments. The computations in this study were run on the Siyuan-1 supported by the Centre for High Performance Computing at Shanghai Jiao Tong University. This work was supported, in part, by National Natural Science Foundation of China (grant nos. 32370658 to Yafei Mao, 82021001 to Q.S., 32300490 to D.W. and 82001372 to X.Y.); by Shanghai Pujiang Programme (grant no. 22PJ1407300), Shanghai Jiao Tong University 2030 Initiative (grant no. WH510363003/016) and Natural Science Foundation of Chongqing, China (grant no. CSTB2024NSCQ-JQX0004) to Yafei Mao; by the National Key Research and Development Programme of China (grant no. 2022YFF0710901), the Biological Resources

Programme of the Chinese Academy of Sciences (KFJ-BRP-005) and the National Science and Technology Innovation 2030 Major Programme (grant no. 2021ZD0200900) to Q.S.; by the Shanghai Rising-Star Programme (grant no. 24YF2721800 to K.M.); by the China Postdoctoral Science Foundation (grant no. 2022M713072 to Y.F.); by the National Institutes of Health (NIH) (grant nos. HG002385 and HG010169 to E.E.E., R00GM147352 to G.A.L. and R01HG011274-01 to K.H.M.); by the Intramural Research Programme of the National Human Genome Research Institute, NIH to A.M.P. and by the Centre for Integration in Science of the Ministry of Aliyah, Israel to I.A.A. We also acknowledge financial support to M.V. under the National Recovery and Resilience Plan, Mission 4, Component 2, Investment 1.1, Call for tender No. 104 published on 2 February 2022 by the Italian Ministry of University and Research (MUR), financed by the European Union—NextGenerationEU—Project Title Telomere-to-telomere sequencing: the new era of centromere and neocentromere evolution (CenVolution)—grant no. CUP H53D23003260006—Grant Assignment Decree No. 1015 adopted on 7 July 2023 by the Italian MUR. E.E.E. is an investigator of the Howard Hughes Medical Institute.

Author contributions Yafei Mao and Q.S. conceived the project. S.Z., N.X., X.Y., Q.S. and Yafei Mao generated sequencing data. N.X., Yamei Li, X.B., Yong Lu, L.Z., Yuxiang Mao and Q.S. contributed the macaque samples, maintained the 582-1 cell line and performed rDNA validation. S.Z., L.F., X.Y., Y.H., D.M., K.M., C.Y., D.W., G.Z., B.S., Y.S., Q.S. and Yafei Mao assembled genomes, analysed the data and performed quality control analyses. A.M.P. and E.E.E. contributed the non-human great ape genomes and analysed the data. S.Z., L.F., Z.Y., E.E.E. and Yafei Mao performed the performed structural variant analyses. S.Z., L.F., Z.Y., J.H. and Yafei Mao performed large-scale genomic difference analyses. L.d.G., A.P., F.A. and M.V. performed the FISH analyses. S.Z., X.J., F.R., G.A.L., V.A.S., K.H.M., I.A.A. and Yafei Mao performed centromere analyses. S.Z., L.F., Z.Y., X.J., Yuxiang Mao, Q.S. and Yafei Mao performed single-cell RNA-seq and *FOLH1* gene family and *PIEZO2* analyses. S.Z., Y.F., D.W. and Yafei Mao performed the genetic differentiation analyses. L.F., R.H. and Q.L. performed the protein structure prediction analyses. S.Z., Z.L., X.Y., K.M., D.L. and Yafei Mao performed the functional experiments. S.Z., L.F., Q.S. and Yafei Mao drafted the manuscript. V.A.S. is retired from the Institute of Molecular Genetics. All authors read and approved the manuscript.

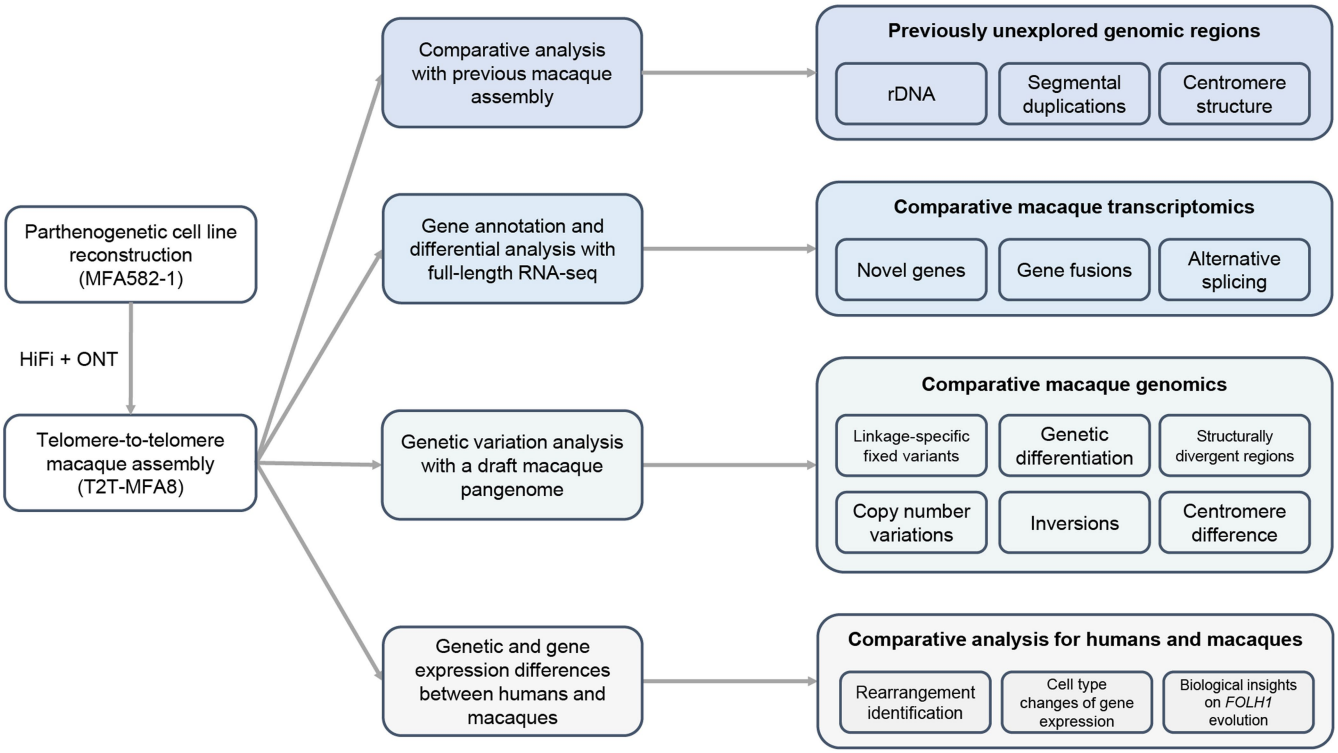
Competing interests E.E.E. is a scientific advisory board member of Variant Bio. The other authors declare no competing interests.

Additional information

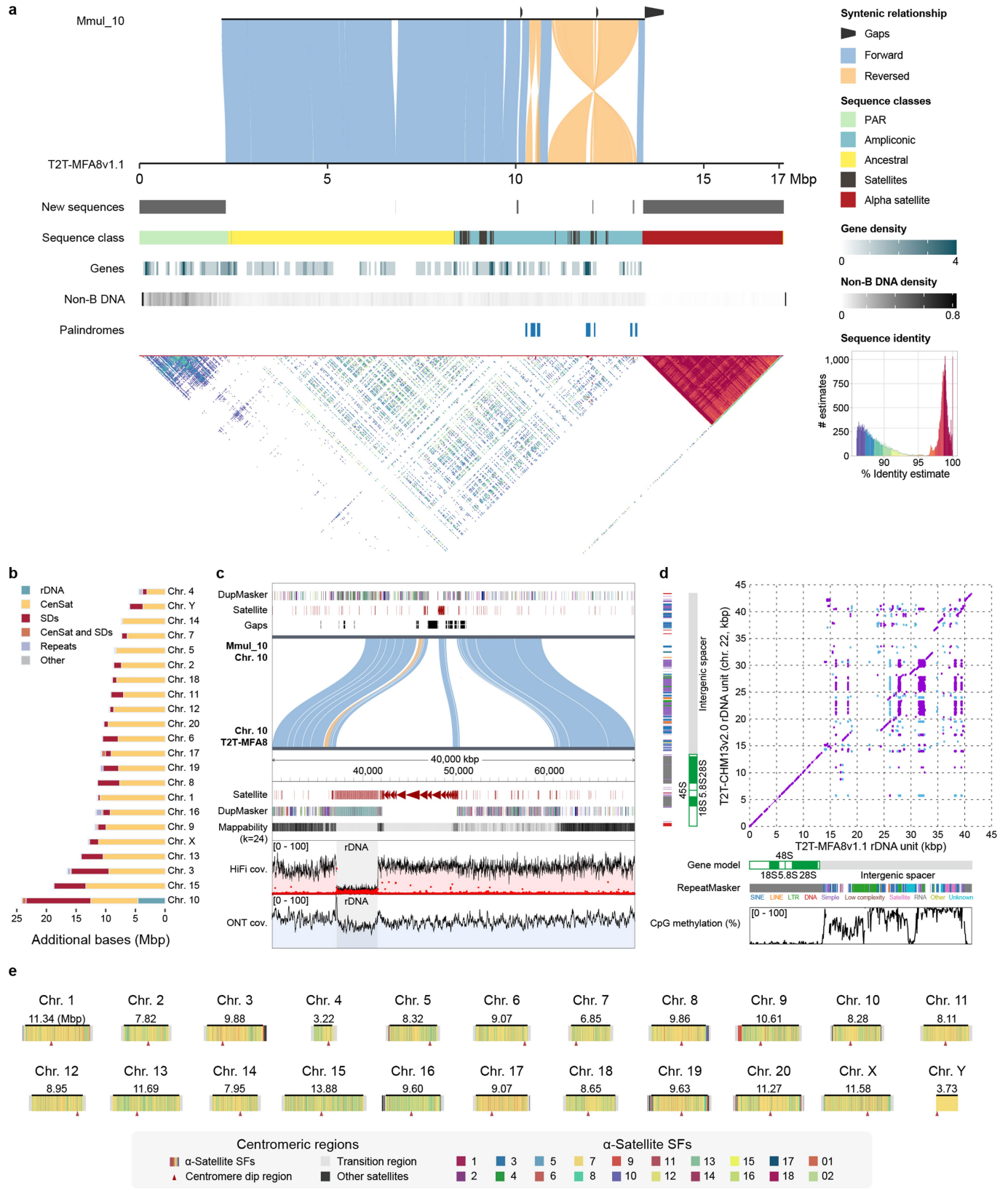
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-025-08596-w.

Correspondence and requests for materials should be addressed to Qiang Sun or Yafei Mao.

Peer review information *Nature* thanks Martin Kuhlwilm, Katerina Guschanski and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

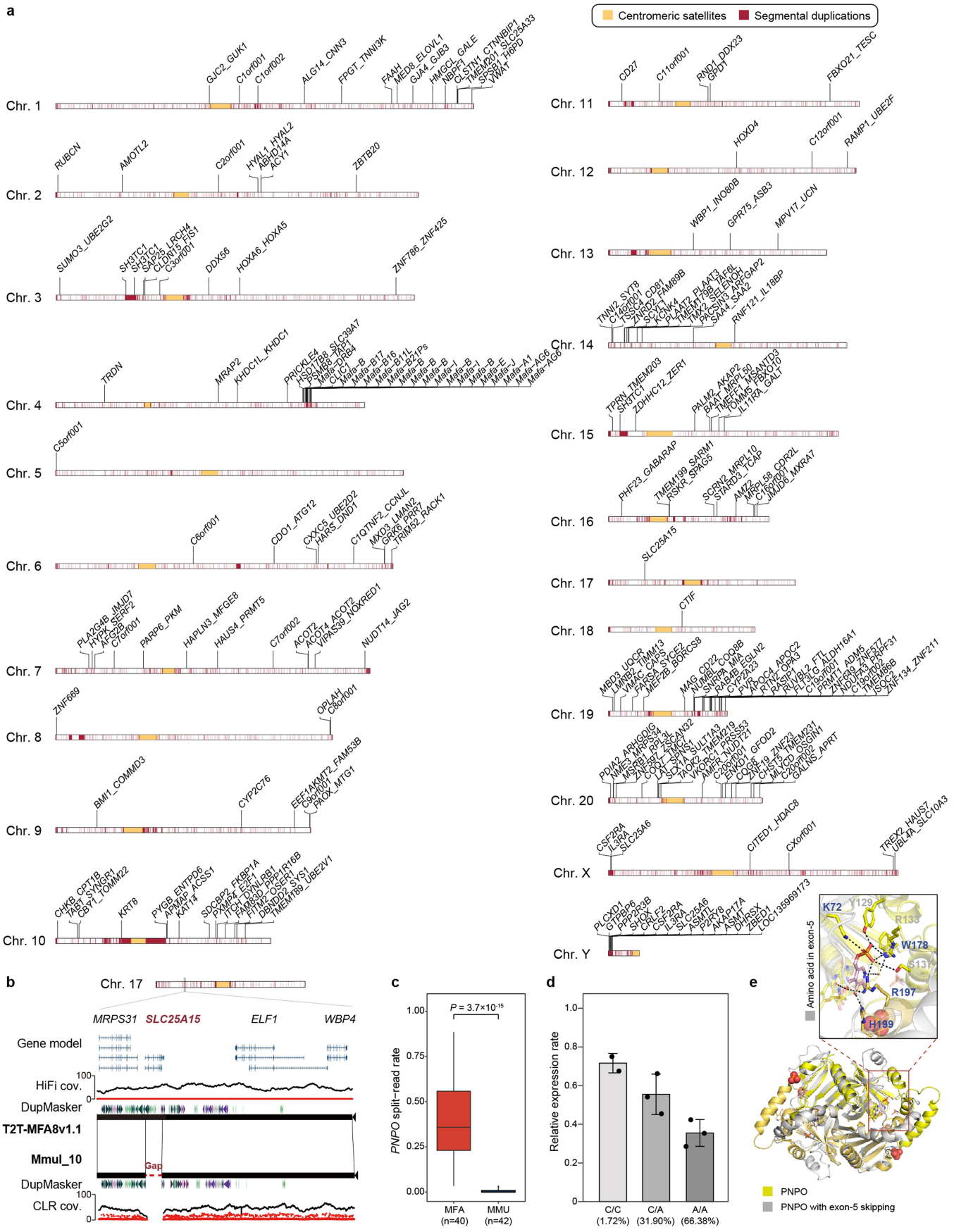


Extended Data Fig. 1 | The conceptual workflow of this study. This diagram illustrates the research strategy in this study.



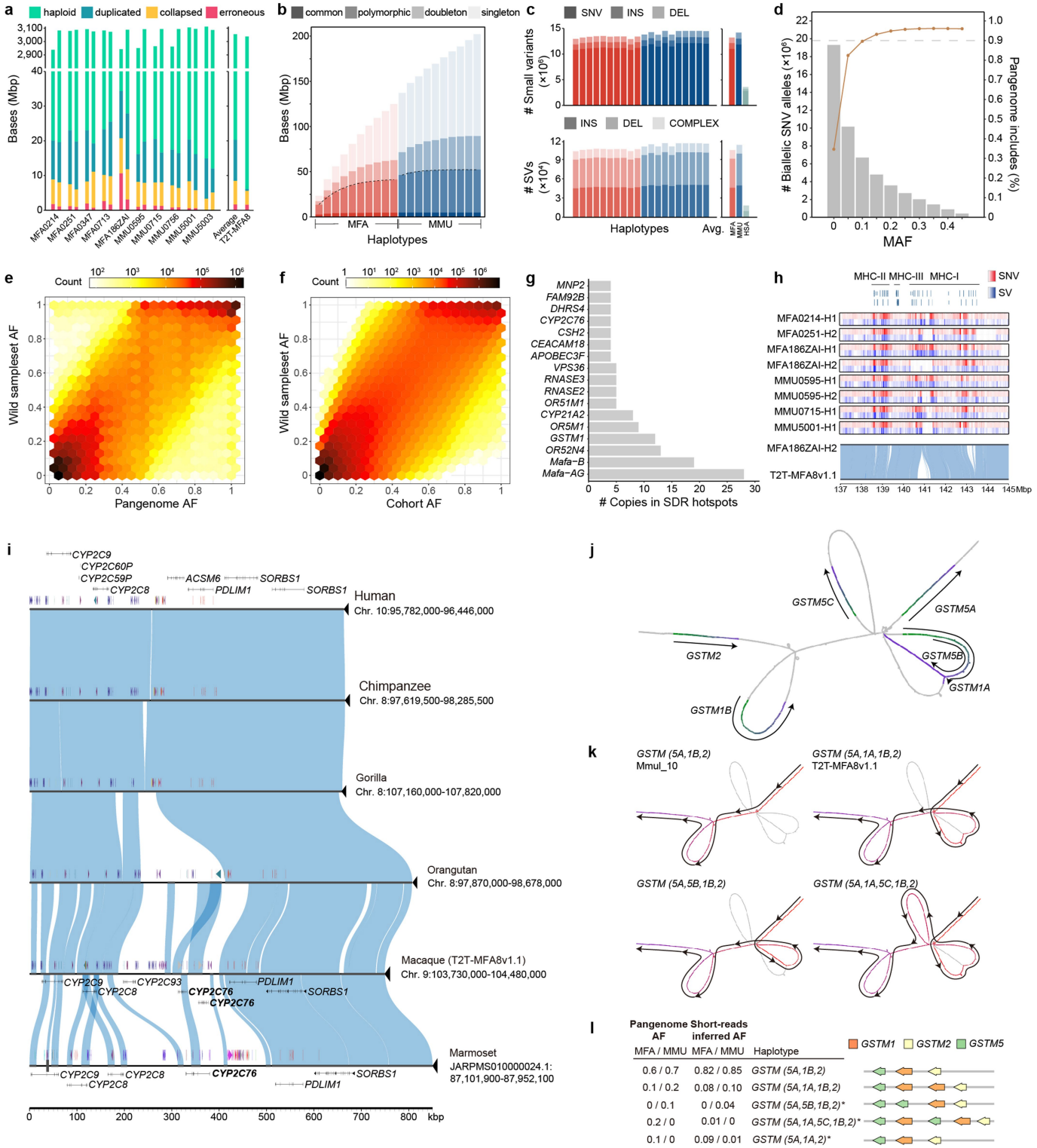
Extended Data Fig. 2 | See next page for caption.

Extended Data Fig. 2 | Previously unresolved regions. (a) A synteny plot (top) displays the alignment of the newly assembled chr. Y (T2T-MFA8v1.1) against the previous macaque assembly (Mmul_10). Blue and yellow blocks represent forward and reversed alignments, respectively. The tracks (bottom) show the newly assembled sequences (compared to Mmul_10), sequence classes, gene density, non-B DNA density, palindromes, and intrachromosomal sequence identity, respectively. **(b)** The bar plot illustrates the repeat annotation of newly added sequences. **(c)** The syntenic comparison highlights the rDNA and centromere regions on chr. 10 between T2T-MFA8v1.1 and Mmul_10. The upper panel illustrates the syntenic relationship between these assemblies, alongside their repeat annotations and mappability. In the lower panel, the HiFi and ONT coverage for T2T-MFA8v1.1 is depicted, with black and red dots marking the primary and secondary alleles, respectively. **(d)** Syntenic comparison of rDNA units between T2T-MFA8v1.1 (chr. 10) and T2T-CHM13v2.0 (chr. 22). The dot plot demonstrates a conserved synteny in the rDNA coding regions between humans and macaques. The common repeat annotation and methylation patterns are listed along the axes. **(e)** The complete centromere assemblies of T2T-MFA8v1.1. Colors represent the suprachromosomal families (SF) of α-satellites, with the lengths of the α-satellite arrays indicated. The centromere dip regions are marked with triangles, as obtained by methylation calling.



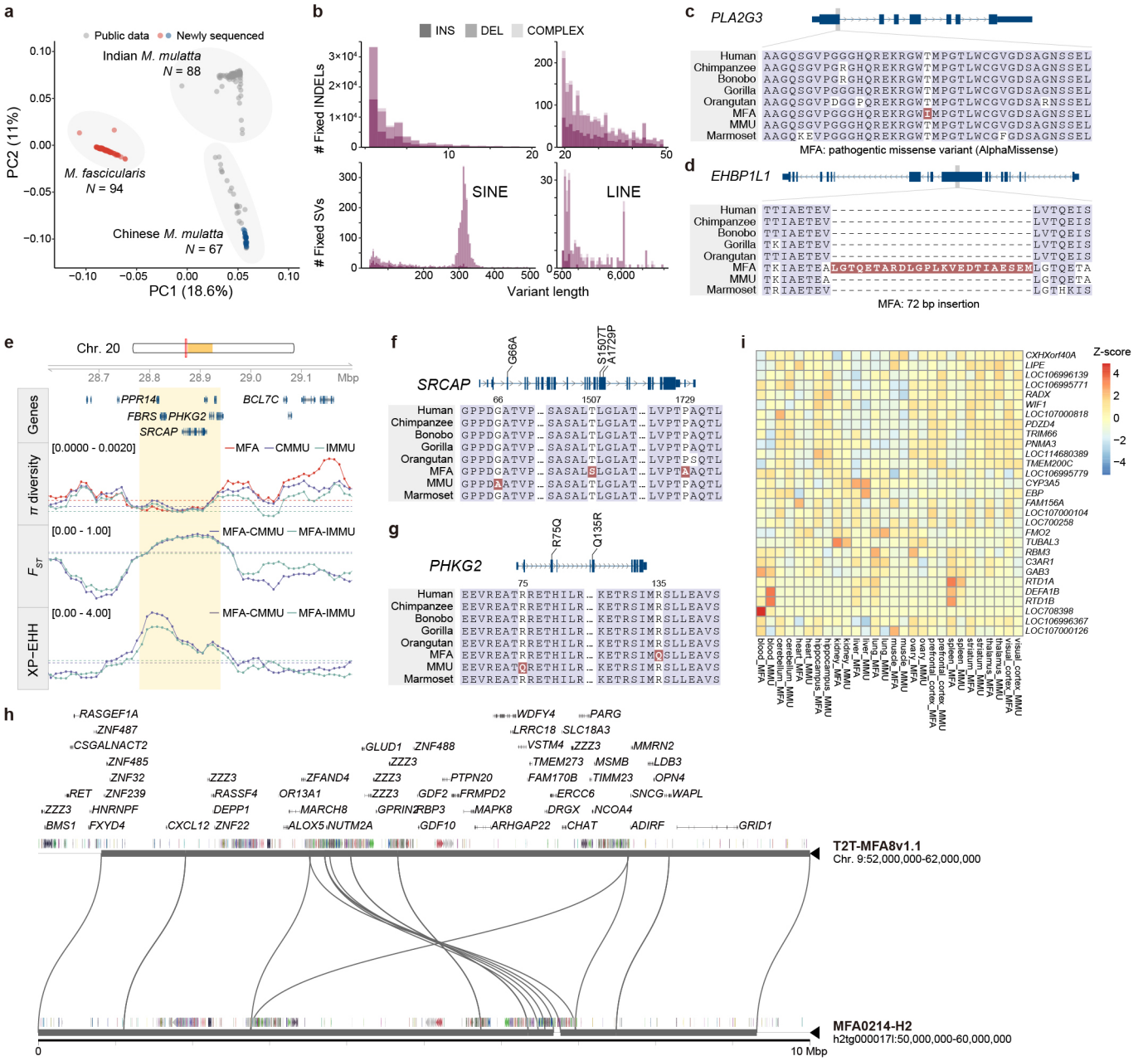
Extended Data Fig. 3 | See next page for caption.

Extended Data Fig. 3 | The comprehensive gene annotation set of T2T-MFA8v1.1 and *PNPO* analysis. (a) The ideogram track shows the centromeric satellites (yellow) and segmental duplications (red), with newly added protein-coding genes labeled above. Genes that are not available in NCBI are marked with "CXorfXXX". **(b)** The red dashed line represents a 21 kbp unassembled region in Mmul_10. Gene models are shown on the top with read-depth validation below. CLR: continuous long reads. **(c)** The short-read RNA-seq confirms the exon-skipping event in MFA (two-sided Mann-Whitney U test). The y-axis refers to the split-read rate of exon-5 on *PNPO*. Box plots denote median and interquartile range (IQR), with whiskers 1.5×IQR. The number of biological replicates is indicated in parentheses below each plot. **(d)** The qPCR validation supports that the genotypes (C/C, C/A, and A/A) are potentially associated with exon-5 skipping in MFA. The genotype frequencies are listed in the parentheses below each plot. Each dot represents different biological replicates (error bars, mean ± s.d.). **(e)** The predicted protein structures of PNPO with and without exon-5 suggest the potential loss of enzyme activity due to disrupted interactions. The zoomed-in panel highlights key amino acids (K72, Y129, R133, S137, W178, R197, and H199) within the active site, with those specific to exon-5 (Y129, R133, and S137) shown in gray.

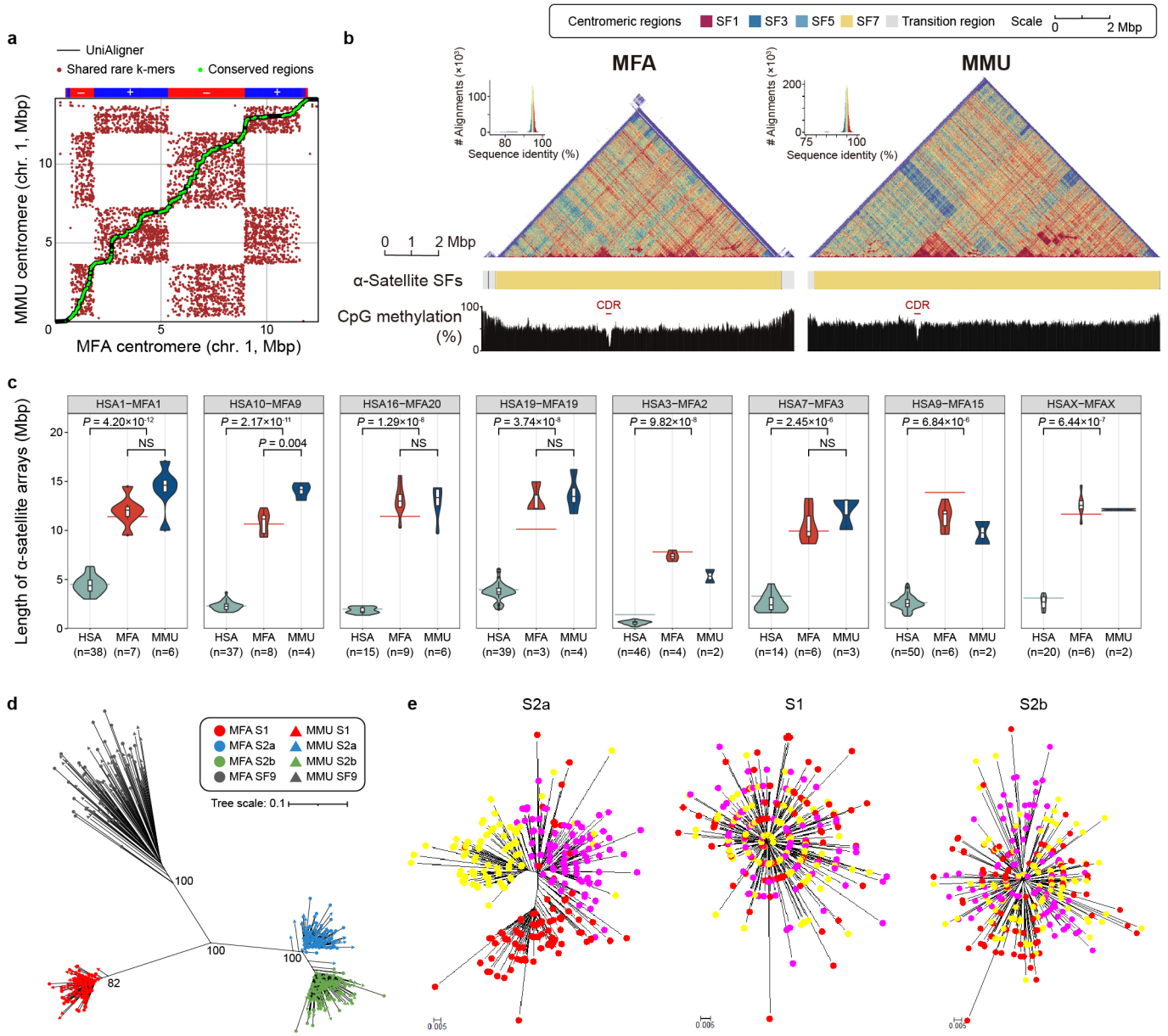


Extended Data Fig. 4 | See next page for caption.

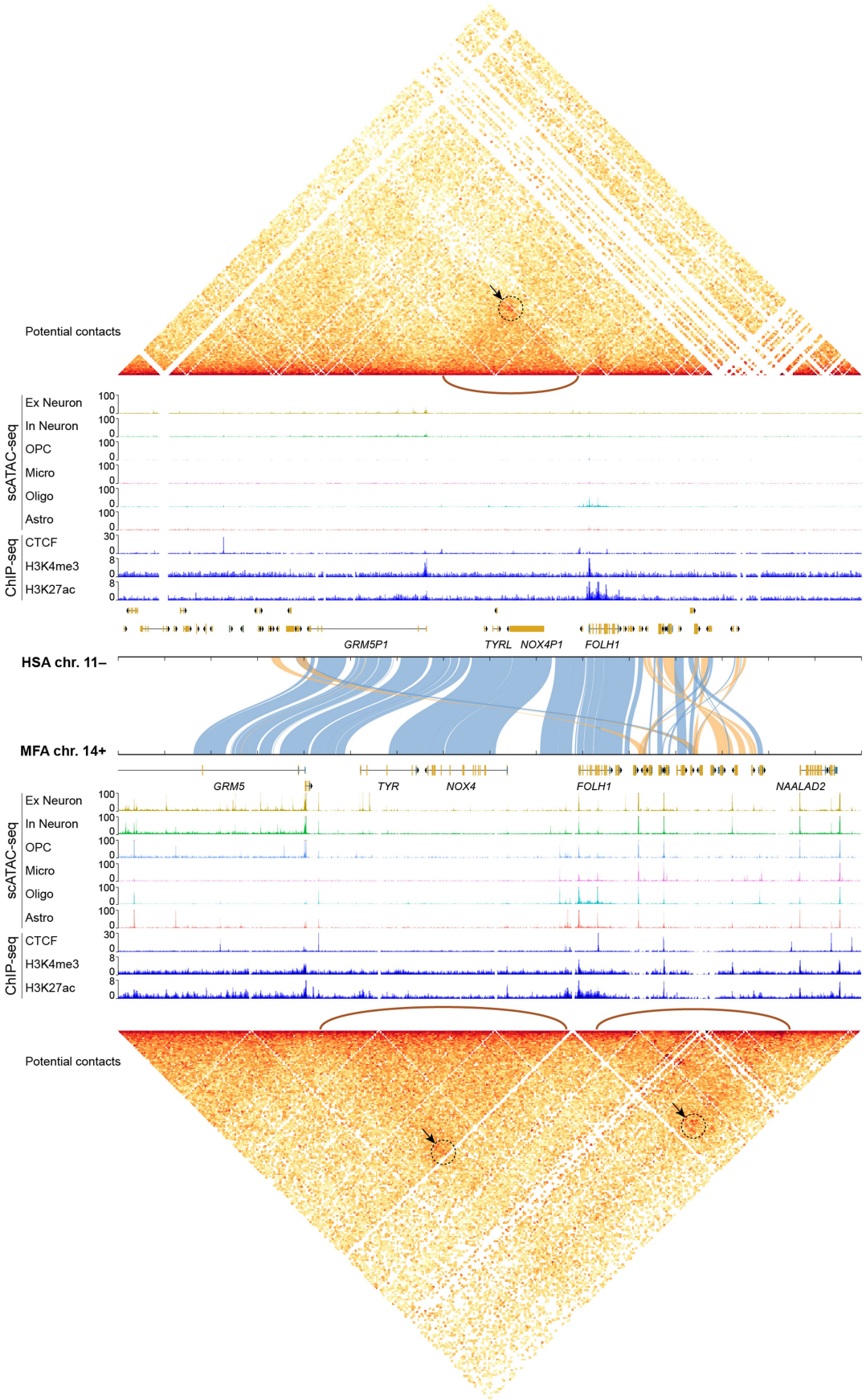
Extended Data Fig. 4 | The quality control, variant discovery, and structural haplotype analysis of the macaque pangenome. (a) Flagger evaluation of 20 haplotype-resolved assemblies is shown on the left panel, while the right panel shows the average across 20 assemblies and the evaluation of T2T-MFA8 (no chr. Y). **(b)** The cumulative number of added bases when adding assemblies one by one is illustrated, with red representing MFA and blue representing MMU. The total of added polymorphic sequences shows slow growth after the seventh MFA or MMU assembly. The species switch (MFA → MMU) increases the yield of added sequences. Transparent colors indicate singleton (AF < 5%), doubleton (5% ≤ AF < 10%), polymorphic (10% ≤ AF < 50%), and common (AF ≥ 50%) alleles. **(c)** The left panel shows the number of small variants (top) and SVs (bottom) per haplotype in the pangenome graph. The right panel shows the average number of small variants (top) and SVs (bottom) of MFA, MMU, and humans (from the HPRC-year1 MC pangenome graph). **(d)** The biallelic SNV comparison between the pangenome graph and the macaque whole-genome sequencing (WGS) cohort (289 macaques). The gray histogram illustrates the count of SNVs from the macaque cohort at MAF cutoffs (x-axis, e.g., MAF > 0.05 includes the SNV count with MAF greater than 0.05), while the line chart represents the fraction of these SNVs covered by the pangenome. This panel shows that the pangenome graph covers 80% of genetic variation with MAF ≥ 5% in the macaque cohort. **(e, f)** These panels show the correlation of AFs between the pangenome and 79 wild samples (e) and between the macaque cohort and the same wild samples (f). **(g)** The bar plot illustrates the most common copy number (CN) variable genes in SDR hotspots of macaques. The x-axis represents the number of gene copies that can be mapped to a bubble in the pangenome graph, while the y-axis shows the 17 most CN variable genes. **(h)** This panel demonstrates the complexity of major histocompatibility complex (MHC) in macaques. SNV and SV densities for eight structural haplotypes with gene models are shown above (top). The syntenic relationship between T2T-MFA8v1.1 and MFA186ZAI-H2 (bottom) shows a ~ 1 Mbp deletion in MFA186ZAI-H2 with respect to T2T-MFA8v1.1. **(i)** This panel displays the syntenic relationship of the *CYP2C76* region in primates. In each assembly, the syntenic regions are represented as blocks, while non-syntenic regions are represented as thin lines, along with their DupMasker and gene annotation attached to each genome segment. **(j)** The structural representation of the *GSTM* family is shown, with the gene annotation. Green and purple refer to the start and end of *GSTM* gene bodies, respectively. **(k)** The graphical representation of four structural haplotypes of *GSTM* follows different paths in the pangenome, with red and purple representing the start and end of a path, respectively. The haplotype of T2T-MFA8v1.1 is *GSTM (5 A, 1 A, 1B, 2)*. **(l)** The table illustrates the frequency statistics of *GSTM* haplotypes and their schematic graph. The frequency of structural haplotypes in the pangenome graph is displayed in the first column, while the inferred frequency from the population with short-read genotyping is shown in the second column.



Extended Data Fig. 5 | The fixed variants, genetic differentiation regions, and inversions between MFA and MMU. (a) Principal component analysis (PCA) of three macaque populations. The first component (18.6%, x-axis) separates MFA (red) and MMU, while the second component (11%, y-axis) distinguishes CMMU (Chinese rhesus macaque) and IMMU (Indian rhesus macaque). The macaque individuals are clustered according to each population. Newly sequenced samples in this study are marked in color, while the samples from the previous study are marked in gray. **(b)** Lineage-specific fixed genetic variation. The length distribution of fixed INDELs and SVs are shown in the left panel (INDEL: 2-20 bp (top), SV: 50-500 bp (bottom)) and right (INDEL: 20-50 bp (top), SV: 500-10000 bp (bottom)). Notable peaks for *Alu* and *L1* are at 300 bp and 6000 bp. A fixed SNV in *PLA2G3* **(c)** and a fixed SV in *EHBP1L1* **(d)** result in amino acid differences between MFA and MMU. **(e)** A genetic differentiation region associated with *SRCAP* and *PHKG2*. The gene models, π diversity, F_{ST}, and XP-EHH across the genomic region are shown from top to bottom. The dotted lines indicate the bottom 5% threshold from π diversity, the top 5% from F_{ST}, and the top 5% from XP-EHH, respectively. **(f, g)** Fixed missense variants of *SRCAP* (f) and *PHKG2* (g) result in amino acid differences between MFA and MMU. **(h)** The syntenic relationship of the inversion with the longest length (4 Mbp) within macaques, with the gene annotation above. **(i)** The heatmap shows the DEGs within the 500 kbp flanking regions of macaque inversion (≥10 kbp) breakpoints (Z-score of rlog-transformed counts). Each row represents a gene and each column represents a tissue.

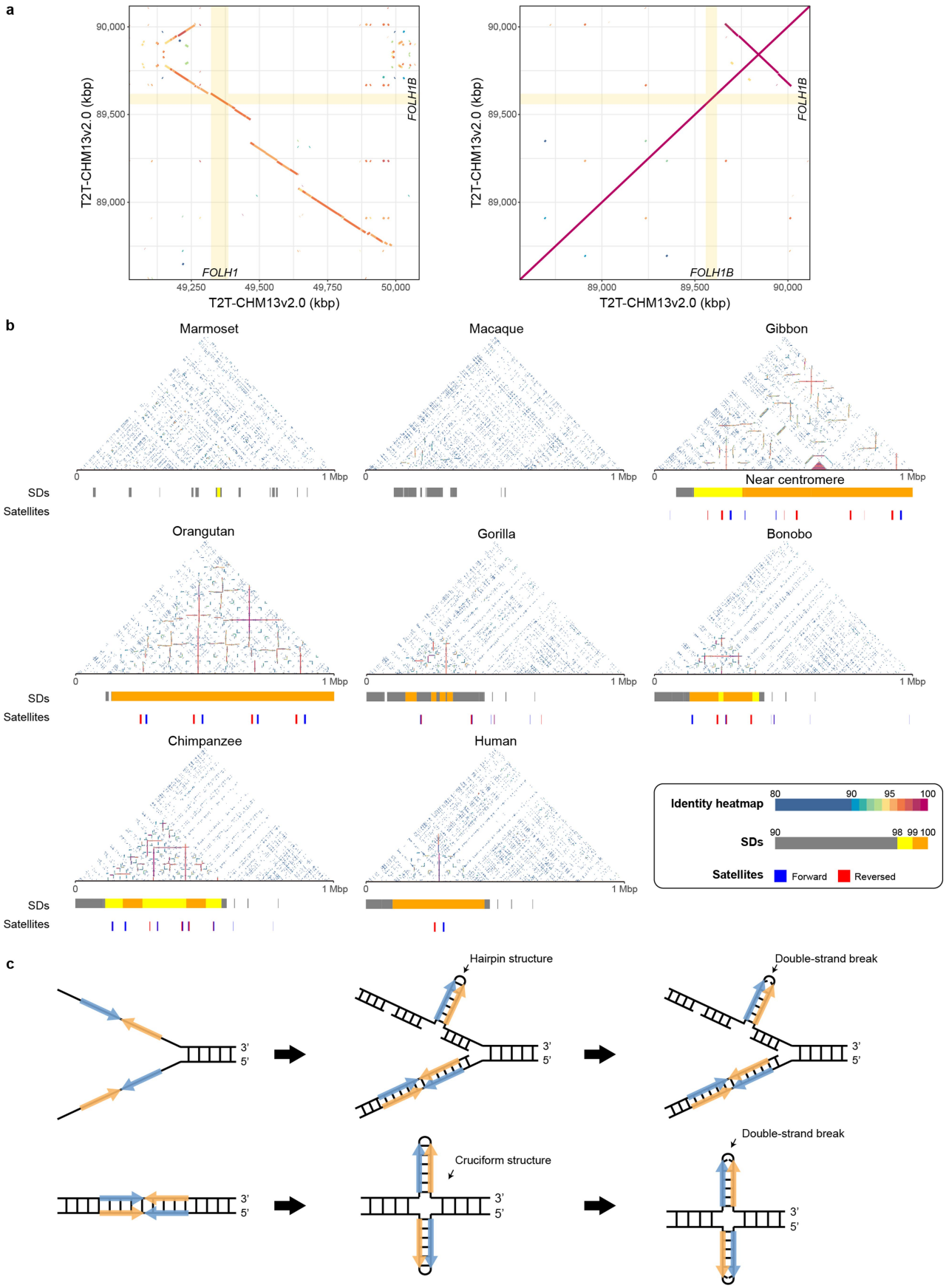


Extended Data Fig. 6 | The comparative analysis on macaque centromeres. **(a)** The dot plot shows the chr. 1 α-satellite arrays between MFA and MMU, generated with UniAligner. The red dots refer to the common rare *k*-mers ($k \geq 80$) and the green dots refer to the conserved regions between two centromeres. The black line indicates the optimal rare alignment path. The α-satellite array strand track is shown above the dot plot (blue for forward strand (+) and red for reverse strand (−)). **(b)** The SF and methylation patterns of α-satellite arrays on chr. 1 for both MFA and MMU are depicted. Sequence similarity within the 5 kb block is visualized using ModDotPlot, with the CDRs highlighted in red by corresponding methylation levels. **(c)** The green, red, and blue violin plots represent the length distribution of α-satellite arrays for HSA, MFA, and MMU, respectively. The horizontal lines indicate the length of reference genomes (green for T2T-CHM13v2.0 and red for T2T-MFA8v1.1). Box plots show median and IQR, with whiskers 1.5×IQR. The *P* values are calculated with the two-sided Mann-Whitney U test, and the number of assembled centromeres is indicated in parentheses below each plot. NS: not significant. **(d)** The phylogenetic tree shows that the S1 (red), S2a (blue), S2b (green), and SF9 α-satellites (dark gray) of MFA (round) and MMU (triangle) mixed in their respective separate clades. **(e)** The phylogeny trees for monomers of S1S2 dimers from MFA chr. 8 (yellow), chr. 11 (red) and chr. 17 (lilac). S2a has chromosome-specific variants while S1 and S2b do not.



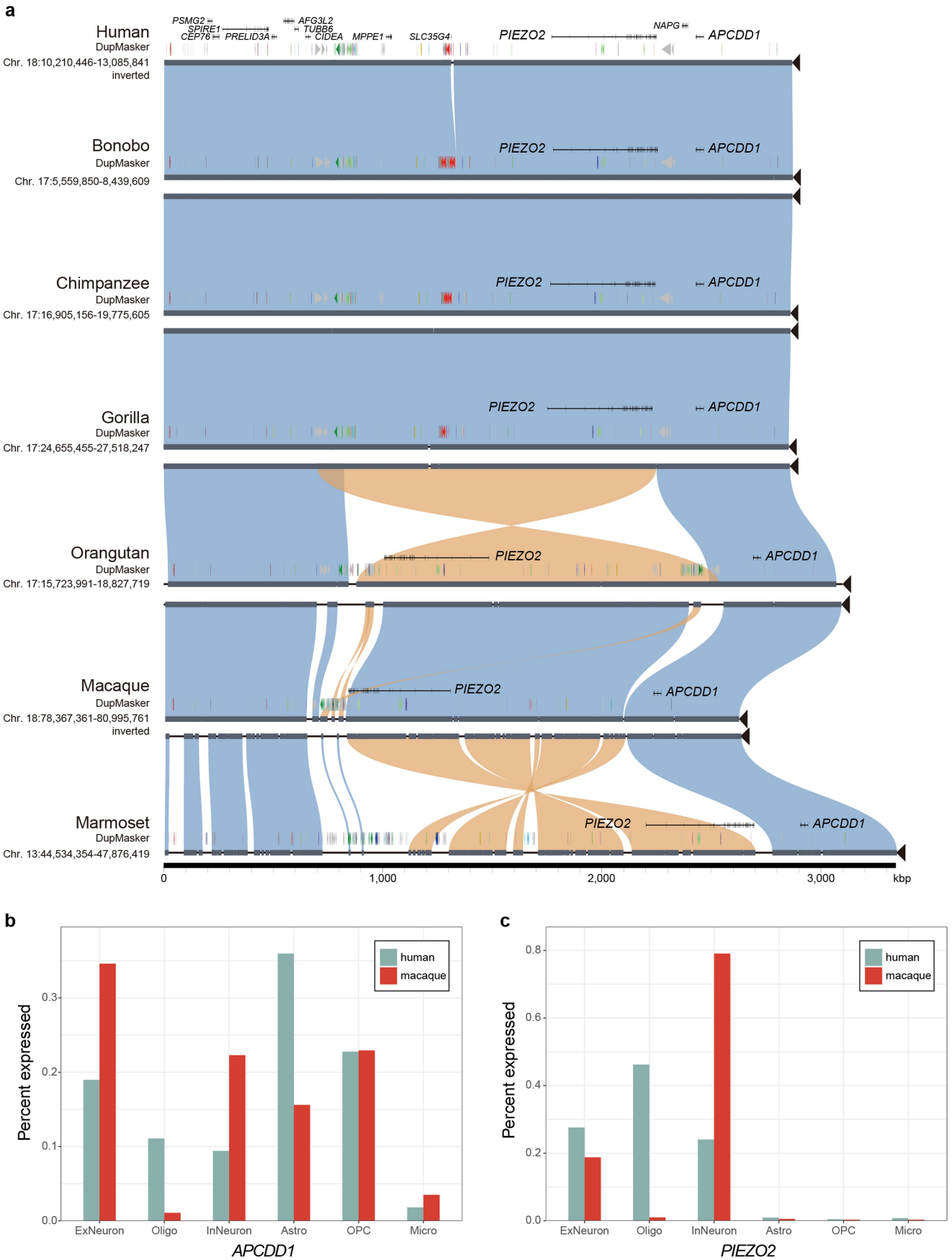
Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | The multi-omics profiles between human *FOLH1* and macaque *FOLH1*. The top panel illustrates the multi-omics profiles at human *FOLH1* locus (T2T-CHM13v2.0 chr. 11, reversed strand), while the bottom panel shows the corresponding profiles in macaque *FOLH1* locus (T2T-MFA8v1.1 chr. 14, forward strand). For the syntenic plot in the middle, blue and yellow blocks represent forward and reversed alignments, respectively. The potential contacts are depicted as loops alongside the Hi-C contact maps, with arrows marking these interactions within the maps. The scATAC-seq tracks are normalized with transcription start site enrichment score, the ChIP-seq tracks are normalized with bins per million mapped reads, and the contact maps are normalized with ICE (iterative correction and eigenvector decomposition).



Extended Data Fig. 8 | See next page for caption.

Extended Data Fig. 8 | The genetic mechanisms of the palindrome-mediated translocation. **(a)** The dot plots illustrate the syntenic relationship between the ancestral and duplicated copies (left panel), as well as the self-syntenic relationship of the ancestral copy (right panel). The positions of human *FOLH1* and *FOLH1B* are highlighted with a yellow background. **(b)** The panel displays sequence identity heatmaps for NHPs, with the 1 Mbp flanking region of the *FOLH1* q-arm, including segmental duplications (SDs) and satellite sequences shown below. Vertical lines in the identity heatmaps indicate palindromic sequences. **(c)** The schematic diagram describes the potential, reported DNA double-strand break repair mechanism underlying palindrome-mediated translocation. Palindromic sequences and their directions are indicated with arrows.



Extended Data Fig. 9 | See next page for caption.

Extended Data Fig. 9 | The evolutionary history of *APCDD1* and *PIEZO2* and their expression patterns. (a) The syntenic relationship of *APCDD1* and *PIEZO2* in primates is shown with minimiro, with gene annotations and DupMasker attached to each genome segment. *PIEZO2* is located inside an inversion in the primate evolution, while *APCDD1* is located near the inversion. **(b, c)** The bar plot shows the proportion of cell types for expressed cells on *APCDD1* (b) and *PIEZO2* (c). The proportion differences in expressed cell type are observed in *APCDD1* and *PIEZO2* between humans and macaques. Ex Neuron, excitatory neuron; Oligo, oligodendrocyte; In Neuron, inhibitory neuron; Astro, astrocyte; OPC, oligodendrocyte precursor cell; Micro: microglia.

Extended Data Table 1 | Genome statistics of T2T-MFA8v1.1 and comparison of macaque genome assemblies

Statistics	Mmul_10	T2T-MFA8v1.1	Difference (±%)
Summary			
Assembled bases (Gbp)	2.97	3.06	+2.99
Unplaced bases (Mbp)	117.34	0	-100.00
Gap bases (Mbp)	34.44	0	-100.00
Number of Contigs	3,185	22	-99.31
Contig NG50 (Mbp)	41.83	162.13	+287.56
Gene annotation			
Number of genes	32,591	33,196	+1.86
Number of protein-coding genes	20,965	21,119	+0.73
Number of transcripts	86,980	110,070	+26.55
Number of protein-coding transcripts	67,410	92,230	+36.82
Segmental duplications			
Percentage of segmental duplications (%)	2.50	4.00	not applicable
Segmental duplication bases (Mbp)	71.38	122.51	+71.62
Number of segmental duplications	33,440	71,564	+114.01
RepeatMasker			
Percentage of repeats (%)	50.53	54.57	not applicable
Repeat bases (Mbp)	1,442.16	1,669.83	+15.79
Long interspersed nuclear elements (Mbp)	601.31	607.99	+1.11
Short interspersed nuclear elements (Mbp)	396.31	406.77	+2.64
Long terminal repeats (Mbp)	259.61	264.82	+2.01
Satellite (Mbp)	25.69	205.16	+698.49
DNA (Mbp)	103.48	106.07	+2.51
Simple repeat (Mbp)	38.46	42.69	+11.02
Low complexity (Mbp)	6.60	6.84	+3.67
Retroposon (Mbp)	0.05	0.05	+8.23
rRNA (Mbp)	0.88	2.38	+169.66

This table presents the assembled bases, gaps, gene annotations, segmental duplications, and repeat annotations for both T2T-MFA8v1.1 and the previous macaque assembly (Mmul_10).

nature portfolio

Corresponding author(s): Yafei Mao & Qiang Sun

Last updated by author(s): Feb 2, 2025

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	SRA Toolkit v3.0.5 https://github.com/ncbi/sra-tools
Data analysis	Custom scripts used in this study are available at Zenodo (https://doi.org/10.5281/zenodo.14220081). In this study, the used softwares/packages are listed below: Another Gtf/Gff Analysis Toolkit (AGAT) v1.2.0 https://github.com/NBISweden/AGAT Assembled Genomes Compressor (AGC) v1.1 https://github.com/refresh-bio/agc Bandage v0.9.0 https://github.com/rrwick/Bandage BandageNG v2022.09 https://github.com/asl/BandageNG BCFtools v1.17 https://github.com/samtools/bcftools BEAST2 v2.7.5 https://github.com/CompEvol/beast2 BEDtools v2.31.0 https://github.com/arq5x/bedtools2 BioPython v1.81 https://github.com/biopython/biopython Bowtie2 v2.5.4 https://github.com/BenLangmead/bowtie2 BUSCO v5.4.6 https://gitlab.com/ezlab/busco BWA v0.7.17 https://github.com/lh3/bwa Canu v2.2 https://github.com/marbl/canu CCS v6.4.0 https://github.com/PacificBiosciences/ccs Cell Ranger v7.2.0 https://github.com/10XGenomics/cellranger Cell Ranger ARC v2.0.2 https://www.10xgenomics.com/support/software/cell-ranger-arc/latest circlize v0.4.15 https://github.com/jokergoo/circlize compleasm v0.2.1 https://github.com/huangnengCSU/compleasm

cooler v0.10.2 https://github.com/open2c/cooler
cooltools v0.7.0 https://github.com/open2c/cooltools
cuteSV v2.0.2 https://github.com/tjiangHIT/cuteSV
Cytoscape v3.9.1 https://github.com/cytoscape/cytoscape
cyvcf2 v0.30.22 https://github.com/brentp/cyvcf2
dbscan v1.1-11 https://github.com/mhahsler/dbscan
deepTools v3.5.5 https://github.com/deeptools/deepTools
DeepVariant v1.5.0 https://github.com/google/deepvariant
DESeq2 v1.38.3 https://github.com/thelovelab/DESeq2
Dfam v3.6 https://dfam.org/
DoubletFinder v2.0.3 https://github.com/chris-mcginnis-ucsf/DoubletFinder
dplyr v1.1.4 https://github.com/tidyverse/dplyr
DupMasker v4.1.2 https://github.com/rmhubley/RepeatMasker/blob/master/DupMasker
Exonerate v2.2.0 https://www.ebi.ac.uk/about/vertebrate-genomics/software/exonerate
fastCN v0.2 https://github.com/mrvollger/fastCN-smk
fastp v0.22.0 https://github.com/OpenGene/fastp
FastQC v0.12.1 https://github.com/s-andrews/FastQC
fastsimcoal2 v2.8 http://cmpg.unibe.ch/software/fastsimcoal2/
Flagger v0.2 https://github.com/mobinasri/flagger
flye v2.9.3-b1797 https://github.com/mikolmogorov/Flye
GATK v3.7 https://gatk.broadinstitute.org/
GenMap v1.3.0 https://github.com/cpockrandt/genmap
GenomeScope2.0 commit eca7b88 https://github.com/tbenavi1/genomescope2.0
GenomicRanges v1.50.2 https://github.com/Bioconductor/GenomicRanges
gfabase v0.6.0 https://github.com/mlin/gfabase
gfa-wp commit a8238c0 https://github.com/pangenome/gfa-wp
GffRead 0.12.7 https://github.com/gpertea/gffread
gffutils v0.12 https://github.com/daler/gffutils
ggbreak v0.1.2 https://github.com/YuLab-SMU/ggbreak
ggforce v0.4.1 https://github.com/thomasp85/ggforce
gggenes v0.5.1 https://github.com/wilkox/gggenes
ggplot2 v3.4.3 https://github.com/tidyverse/ggplot2
ggpubr v0.6.0 https://github.com/kassambara/ggpubr
ggridges v0.5.4 https://github.com/wilkelab/ggridges
ggsignif v0.6.4 https://github.com/const-ae/ggsignif
GLnexus v1.2.7 https://github.com/dnanexus-rnd/GLnexus
GraphAligner v1.0.17 https://github.com/maickrau/GraphAligner
Gviz v1.42.1 https://github.com/ivanek/Gviz
harmony v1.2.0 https://github.com/immunogenomics/harmony
HiC-Pro v3.1.0 https://github.com/nservant/HiC-Pro
Hifiasm v0.18.2, v0.19.5, v0.19.8 https://github.com/chhylp123/hifiasm
HISAT2 v2.2.1 https://github.com/DaehwanKimLab/hisat2
HMMER v3.4 https://github.com/EddyRivasLab/hmmer
HTSeq v2.0.4 https://github.com/htseq/htseq
Hum-AS-HMMER https://github.com/fedorrik/HumAS-HMMER_for_AnVIL
Integrative Genome Browser (IGV) v2.8.10 https://github.com/igvteam/igv
IQTREE v2.1.4 http://www.iqtree.org/
IsoSeq v3.8.1 https://github.com/PacificBiosciences/IsoSeq
iTOL https://itol.embl.de/
Jalview v2.11.3.2 https://www.jalview.org/
Jasmine v1.1.5 https://github.com/mkirsche/Jasmine
Jbrowse v2.13.1 https://jbrowse.org/jb2/
Jellyfish v2.3.0 https://github.com/gmarcais/Jellyfish
Juicebox v2.17.00 https://github.com/aidenlab/Juicebox
JuicerTools v1.22.01 https://github.com/aidenlab/juicertools
karyoploteR v1.24.0 https://github.com/bernatgel/karyoploteR
Liftoff v1.6.3 https://github.com/agshumate/Liftoff
lima v2.7.1 https://github.com/pacificbiosciences/barcoding/
limma v3.54.2 https://github.com/cran/limma
LSGvar https://github.com/YafeiMaoLab/LSGvar
MACS2 v2.2.9.1 https://pypi.org/project/MACS2/
MAFFT v7.515, v7.520 https://github.com/GSLBiotech/mafft
Merfin v1.1 https://github.com/arangrhie/merfin
Merqury v1.3 https://github.com/marbl/merqury
meryl v1.4 https://github.com/marbl/meryl
Minigraph-Cactus v2.6.6 https://github.com/ComparativeGenomicsToolkit/cactus
minimap2 v2.24, v2.26 https://github.com/lh3/minimap2
minimiro commit 8a77b25 https://github.com/mrvollger/minimiro
mrsFAST v3.4.2 https://github.com/sfu-compbio/mrsfast
MultiQC https://github.com/MultiQC/MultiQC
MUMmer v3.23 https://mummer.sourceforge.net/
nanopolish v0.14.0 https://github.com/jts/nanopolish
NCBI Foreign Contamination Screen (FCS) v0.4.0 https://github.com/ncbi/fcs
NCBI RMBLAST v2.11.0 https://www.repeatmasker.org/rmblast/
NucFreq v0.1 https://github.com/mrvollger/NucFreq
panacus v0.2.2 https://github.com/marschall-lab/panacus

PanGenie v3.0.0 https://github.com/eblerjana/pangenie
PAV v2.3.4 https://github.com/EichlerLab/pav
PBSV v2.9.0 https://github.com/PacificBiosciences/pbsv
PEPPER-DeepVariant v0.8.0 https://github.com/kishwarshafin/pepper
pheatmap v1.0.12 https://github.com/raivokolde/pheatmap
Picard v3.2.0 https://github.com/broadinstitute/picard
PLINK v1.90b6.26 https://github.com/chrchang/plink-ng
plotly v5.15.0 https://github.com/plotly/plotly.py
PopGenome commit ab39ee8 https://github.com/pievos101/PopGenome
portion v2.4.1 https://github.com/AlexandreDecan/portion
primatR v0.1.0 https://github.com/daewoooo/primatR
Python v3.8.18, v3.9.16, v3.10.8 https://www.python.org/
qqman v0.1.9 https://github.com/stephenturner/qqman
R v4.2.0, v4.3.0, v4.3.1 https://www.r-project.org/
RepeatMasker v4.1.2 https://github.com/rmhubley/RepeatMasker
RepeatModeler v2.0.4 https://github.com/Dfam-consortium/RepeatModeler
ribotin commit 3d33151 https://github.com/maickrau/ribotin
rustybam v0.1.33 https://github.com/mrvollger/rustybam
SafFire https://github.com/mrvollger/SafFire
SAMtools v1.16, v1.17 https://github.com/samtools/samtools
sedef v1.1 https://github.com/vpc-ccg/sedef
selscan v1.2.0 https://github.com/szpiech/selscan
seqkit v2.4.0 https://github.com/shenwei356/seqkit
seqtk v1.4 https://github.com/lh3/seqtk
Seurat v4.4.0, v5.0.1 https://github.com/satijalab/seurat
sgkit v0.7 https://github.com/sgkit-dev/sgkit
Signac v1.13.0 https://github.com/stuart-lab/signac
singularity v1.2.5-1.el8 https://github.com/sylabs/singularity
Snakemake v7.28.3 https://github.com/snakemake/snakemake
SQANTI3 v5.1.2 https://github.com/ConesaLab/SQANTI3
SRA Toolkit v3.0.5 https://github.com/ncbi/sra-tools
StainedGlass v0.6 https://github.com/mrvollger/StainedGlass
StringTie v2.2.1 https://github.com/gpertea/stringtie
SUPPA2 v2.3 https://github.com/comprna/SUPPA
SVIM v1.4.2 https://github.com/eldariont/svim
SVIM-asm v1.0.3 https://github.com/eldariont/svim-asm
SV-pop https://github.com/EichlerLab/svpop
Tandem Repeat Finder (TRF) v4.09 https://github.com/Benson-Genomics-Lab/TRF
The Database for Annotation, Visualization and Integrated Discovery (DAVID) https://david.ncifcrf.gov/
trimAl v1.4.1 https://github.com/inab/trimal
TrimGalore! v0.6.10 https://github.com/FelixKrueger/TrimGalore
UpSetR v1.4.0 https://github.com/hms-dbmi/UpSetR
Variant Effect Predictor (VEP) v109.3 https://github.com/Ensembl/ensembl-vep
vcfbub v0.1.0 https://github.com/pangenome/vcfbub
vcflib v1.0.9 https://github.com/vcflib/vcflib
VCFtools v0.1.16 https://github.com/vcftools/vcftools
vegan v2.6-4 https://github.com/vegandevs/vegan
VennDiagram v1.7.3 https://github.com/cran/VennDiagram
Verkko v1.2 https://github.com/marbl/verkko
WindowMasker v1.0.0 https://github.com/goeckslab/WindowMasker
winnowmap v2.03 https://github.com/marbl/Winnowmap
Yak v0.1 https://github.com/lh3/yak

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The genome assemblies used in this study, including T2T-CHM13v2.0 (GCF_009914755.1), Mmul_10 (GCF_003339765.1), MFA1912RKSv2 (GCF_012559485.2), rheMacS_1.0 (GCA_008058575.1) and Macaca_fascicularis_6.0 (GCA_011100615.1), are available from the NCBI Genome. Previously published sequencing data used in this study, including PRJNA1004471, PRJNA251548, PRJNA345528, PRJNA832687, PRJNA854879, PRJNA882074, PRJNA953340 and PRJCA018217, are available from the NCBI Sequence Read Archive or NGDC Genome Sequence Archive. The ChIP-seq data of human brain used in this study is available from the ENCODE database. The raw Illumina, PacBio HiFi, ONT, and Hi-C data of T2T-MFA8 are deposited in NCBI under BioProject accession number PRJNA1037719. The raw PacBio HiFi, ONT, and Hi-C data of 10 macaque individuals are deposited in NCBI under BioProject accession number PRJNA1041301. The Iso-Seq data are deposited under NCBI BioProject accession number PRJNA1041301. The Illumina sequences of 151 WGS macaque genomes are deposited in NCBI under BioProject accession number PRJNA1041301. The T2T-MFA8 genome assembly is deposited in NCBI GenBank under accession number GCF_037993035. T2T-MFA8 assembly, annotations, and the UCSC track hub are available at GitHub (https://github.com/zhang-shilong/T2T-MFA8).

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	NA
Reporting on race, ethnicity, or other socially relevant groupings	NA
Population characteristics	NA
Recruitment	NA
Ethics oversight	NA

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No sample-size calculation was performed. We determined the sample size based on its adequacy for capturing the majority of genetic variants among macaques.
Data exclusions	No data was excluded in the analyses.
Replication	RT-PCR experiments were conducted on crab-eating macaques, with two or three replications per group (depending on the genotype), and all replication attempts were successful. Each FISH experiment was repeated 3 times and 10 metaphase spreads with relative fluorochromes were captured for each experimentm, and all attempts were successful.
Randomization	For pangenome sample selection, we first performed Illumina whole-genome sequencing on a macaque cohort and assessed the relationships between the individuals. Based on these results, we then selected 10 unrelated macaque individuals for long-read sequencing.
Blinding	Blinding is not applicable to this study, as we need to investigate the genetic background of each macaque individually in order to capture the majority of genetic variants.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☐	☒ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used	Mouse anti-FLAG (Sigma-Aldrich, F1804), mouse anti-α-tubulin (Sigma-Aldrich, T9026), rabbit anti-PNPO (Invitrogen, PA5-55687), Anti-Mouse IgG, HRP (Invitrogen, A16006), Anti-Rabbit IgG, HRP (Invitrogen, A16104), and Rabbit Anti-Monkey IgG Antibody (H+L) (Bioss, bs-0335R-HRP).
Validation	The antibody information and validation statements are provided in the manufacturer's websites: mouse anti-FLAG (Sigma-Aldrich, F1804): https://www.sigmaaldrich.cn/CN/en/product/sigma/f1804; mouse anti-α-tubulin (Sigma-Aldrich, T9026): https://www.sigmaaldrich.cn/CN/en/product/sigma/t9026; rabbit anti-PNPO (Invitrogen, PA5-55687): https://www.thermofisher.cn/cn/en/antibody/product/PNPO-Antibody-Polyclonal/PA5-55687; Anti-Mouse IgG, HRP (Invitrogen, A16006): https://www.thermofisher.cn/cn/en/antibody/product/Donkey-anti-Goat-IgG-H-L-Cross-Adsorbed-Secondary-Antibody-Polyclonal/A16006; Anti-Rabbit IgG, HRP (Invitrogen, A16104): https://www.thermofisher.cn/cn/en/antibody/product/Goat-anti-Rabbit-IgG-H-L-Cross-Adsorbed-Secondary-Antibody-Polyclonal/A16104; Rabbit Anti-Monkey IgG Antibody (H+L) (Bioss, bs-0335R-HRP): https://www.biossusa.com/products/bs-0335r-hrp.

Eukaryotic cell lines

Policy information about cell lines and Sex and Gender in Research

Cell line source(s)	We generated a female macaque embryonic stem cell (ESC) line.
Authentication	Short Tandem Repeats (STRs) genotyping was performed for the cell line authentication.
Mycoplasma contamination	We confirm that the cell line is free from mycoplasma contamination.
Commonly misidentified lines (See ICLAC register)	No commonly misidentified lines were used in this study.

Animals and other research organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research, and Sex and Gender in Research

Laboratory animals	We used 10 macaques for long-read whole-genome sequencing (5 females and 5 males), tissues from 5 macaques for long-read transcriptome sequencing (3 females and 2 males, including infant and adult), and 151 macaques for short-read whole-genome sequencing. Detailed information (species, sex, age and development stage) for each individual is provided in Supplementary Table.
Wild animals	No wild animals were used in this study.
Reporting on sex	Detailed information for each individual is provided in Supplementary Table.
Field-collected samples	No field-collected samples were used in this study.
Ethics oversight	The usage of research animals in this study underwent evaluation and approval by Primate Life Sciences Ethics Committee of the Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences (ION-2019043R03).

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Plants

Seed stocks	NA
Novel plant genotypes	NA
Authentication	NA