Evolutionary and biomedical insights from a marmoset diploid genome assembly

https://doi.org/10.1038/s41586-021-03535-x

Received: 20 September 2020

Accepted: 12 April 2021

Published online: 28 April 2021

Open access

Check for updates

Chentao Yang[1,2,19], Yang Zhou[1,19], Stephanie Marcus[3,19], Giulio Formenti[3,4], Lucie A. Bergeron[2], Zhenzhen Song[5], Xupeng Bi[1], Juraj Bergman[6], Marjolaine Marie C. Rousselle[6], Chengran Zhou[1], Long Zhou[1], Yuan Deng[1,2], Miaoquan Fang[1], Duo Xie[1], Yuanzhen Zhu[1], Shangjin Tan[1], Jacquelyn Mountcastle[4], Bettina Haase[4], Jennifer Balacco[4], Jonathan Wood[7], William Chow[7], Arang Rhie[8], Martin Pippel[9,10], Margaret M. Fabiszak[11], Sergey Koren[8], Olivier Fedrigo[4], Winrich A. Freiwald[11,12], Kerstin Howe[7], Huanming Yang[1,5,13,14], Adam M. Phillippy[8], Mikkel Heide Schierup[6], Erich D. Jarvis[3,4,15] & Guojie Zhang[2,16,17,18] ✉

The accurate and complete assembly of both haplotype sequences of a diploid organism is essential to understanding the role of variation in genome functions, phenotypes and diseases[1]. Here, using a trio-binning approach, we present a high-quality, diploid reference genome, with both haplotypes assembled independently at the chromosome level, for the common marmoset (*Callithrix jacchus*), an primate model system that is widely used in biomedical research[2,3]. The full spectrum of heterozygosity between the two haplotypes involves 1.36% of the genome—much higher than the 0.13% indicated by the standard estimation based on single-nucleotide heterozygosity alone. The de novo mutation rate is 0.43×10^{-8} per site per generation, and the paternal inherited genome acquired twice as many mutations as the maternal. Our diploid assembly enabled us to discover a recent expansion of the sex-differentiation region and unique evolutionary changes in the marmoset Y chromosome. In addition, we identified many genes with signatures of positive selection that might have contributed to the evolution of *Callithrix* biological features. Brain-related genes were highly conserved between marmosets and humans, although several genes experienced lineage-specific copy number variations or diversifying selection, with implications for the use of marmosets as a model system.

A diploid organism carries two haploid genomes with a range of variants, which make substantial contributions to phenotypic variation[4]. Phased haplotype assemblies can help to reveal the *cis*- and *trans*-acting variants on the two homologous genomes. However, most contemporary de novo genome-sequencing efforts produce a single mosaic reference genome derived from parts of both maternal and paternal alleles, with variations between homologous chromosomes normally being disregarded. As a consequence, these methods usually fail to assemble genomic regions with high heterogeneity, resulting in fragmented sequences. A few methods have been developed to produce partial haplotype-phased genome assemblies and showed power in using long sequencing reads to produce long haplotigs (haplotype-specific contigs)[5,6]. However, producing an assembly that is completely phased at the chromosome level for both haplotypes of a diploid genome remains a challenge. Here, as part of the Vertebrate Genomes Project, we used a trio-binning approach[7,8] to produce a chromosome-level, fully haplotype-resolved diploid genome assembly for the common marmoset, *C. jacchus*. This New World primate has been established as an animal model for a broad range of biomedical research such as neuroscience, stem cell biology and regenerative medicine[2,3]. With our high-quality diploid assembly, we discovered new properties of heterozygosity on both autosomes and sex chromosomes of this primate species.

Diploid genome assembly

We generated 63×-coverage PacBio continuous long reads, 55× 10X Genomics Chromium linked-reads, 154× Bionano optical molecules,

[1]BGI-Shenzhen, Shenzhen, China. [2]Villum Centre for Biodiversity Genomics, Section for Ecology and Evolution, Department of Biology, University of Copenhagen, Copenhagen, Denmark. [3]Laboratory of Neurogenetics of Language, The Rockefeller University, New York, NY, USA. [4]Vertebrate Genome Laboratory, The Rockefeller University, New York, NY, USA. [5]University of the Chinese Academy of Sciences, Beijing, China. [6]Bioinformatics Research Centre, Aarhus University, Aarhus, Denmark. [7]Wellcome Sanger Institute, Hinxton, UK. [8]Genome Informatics Section, Computational and Statistical Genomics Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA. [9]Max Planck Institute of Molecular Cell Biology and Genetics, Dresden, Germany. [10]Center for Systems Biology, Dresden, Germany. [11]Laboratory of Neural Systems, The Rockefeller University, New York, NY, USA. [12]Center for Brains, Minds and Machines (CBMM), The Rockefeller University, New York, NY, USA. [13]James D. Watson Institute of Genome Sciences, Hangzhou, China. [14]Guangdong Provincial Academician Workstation of BGI Synthetic Genomics, BGI-Shenzhen, Shenzhen, China. [15]Howard Hughes Medical Institute, Chevy Chase, MD, USA. [16]State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, China. [17]China National GeneBank, BGI-Shenzhen, Shenzhen, China. [18]Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming, China. [19]These authors contributed equally: Chentao Yang, Yang Zhou, Stephanie Marcus. ✉e-mail: guojie.zhang@bio.ku.dk

105× chromosome conformation capture (Hi-C) reads from a captive male F_1 marmoset and 70× short-read sequences from the DNA of both parents (Supplementary Table 1, Supplementary Fig. 1). We used an updated version of TrioCanu[7,8] to bin the PacBio long reads of the F_1 marmoset via *k*-mers of the parental short reads, and assembled each set into haploid-specific contigs, which were independently scaffolded with the 10X, Bionano and Hi-C data[8] (Extended Data Fig. 1, Supplementary Fig. 2, Supplementary Tables 2, 3). The final contig and scaffold NG50 values after manually curation were 7.7. Mb and 146 Mb for the maternal assembly and 12.1 Mb and 136 Mb for the paternal assembly, respectively. *k*-mer assessment indicated that the assemblies were fully phased (Extended Data Fig. 2a, Supplementary Figs. 3, 4). Each haploid genome includes 22 autosomes and each of the two sex chromosomes (X and Y), with 99.45% and 98.94% of the maternal and paternal alleles assigned to chromosomes, respectively. The assembled chromosome lengths showed a clear linear correlation with the estimated marmoset karyotype lengths[8,9] (Extended Data Fig. 2b, Supplementary Note, Supplementary Tables 4, 5, Supplementary Fig. 5). Although marmosets show prevalent genetic chimerism between twins and triplets in utero[10], the chimeric level of the F_1 male muscle sample used in this study was very low, as expected[11] (Extended Data Fig. 1d–g, Supplementary Fig. 6, Supplementary Tables 6, 7, Supplementary Note).

We estimated the single-base-pair accuracy rate to be 99.996% for the maternal assembly and 99.998% for the paternal assembly (Supplementary Note, Supplementary Fig. 7, Supplementary Tables 8, 9). About 93% and 88% of the gaps in the previously published marmoset reference genome cj3.2[12] were closed in our maternal and paternal assemblies, respectively, and both showed an increase of over 290-fold in contig N50, with 95.75% and 93.62% of the contigs being over 1 Mb, respectively (Extended Data Fig. 2c). Iso-Seq full-length transcriptome data also suggest a high completeness of our assembly (Supplementary Note, Supplementary Tables 10, 11). Comparison with two other recently released chromosome-level assemblies (cj1700 and cj2019) showed 16 large intra-chromosome-level structural variants (SVs) (larger than 1 Mb) and 3 inter-chromosomal SVs (Supplementary Tables 12, 13). PacBio long reads and 10X linked-reads confirmed that our assemblies were correct (Supplementary Figs. 8, 9, Supplementary Tables 12–14). However, these differences may also be due to the large structural polymorphisms.

Heterozygosity between parental genomes

In traditional genome-sequencing efforts, heterozygosity is normally estimated by mapping sequencing reads onto a mosaic reference genome, resulting in limited phase information of the heterozygous variants. Our assemblies enable us to directly compare the two parentally inherited genomes and identify the full spectrum of genetic variants between the parental alleles, including single nucleotide variations (SNVs), insertion and deletions (indels) and large SVs (Supplementary Fig. 10). We identified 3.47 million SNVs and around 232,000 short (maximum of 50 base pairs (bp)) indels across the whole genome (Fig. 1a), with 96.5% SNVs confirmed by short-read mapping. PCR experiments validated 99.6% and 95.2% randomly selected SNVs and short indels (Supplementary Note, Supplementary Tables 15–17), indicating that our diploid assembly enabled us to detect allelic variants with considerably high accuracy. We found a correlation between SNV rate and indel rate (Supplementary Fig. 11a), in which both displayed a unimodal distribution across the genomes (Supplementary Figs. 11b, 12). Consistent with laboratory inbreeding, we observed 28 genomic regions with long runs of homozygosity (Fig. 2a), with the longest one spanning more than 10 Mb (Supplementary Fig. 13a). This pattern can also be detected in other marmoset samples with short-read resequencing data[13] (Supplementary Fig. 13b, Supplementary Table 18), suggesting that captive marmosets are suffering a notable reduction of genetic diversity.

Heterozygous variation in regulatory or coding regions could result in allele-specific expression profiles or different products of the same genes from the two alleles[14]. We found that approximately 1.1% of SNVs and 0.58% of indels were located in protein-coding genes or regulatory regions. In particular, 8,144 SNVs caused non-synonymous substitutions and 274 indels caused frame-shifting mutations, which can produce allele-specific transcripts and proteins. This observation was validated by the Iso-Seq data, in which we detected that 2,537 genes produced transcripts with variation in open-reading frames from the parental alleles (Supplementary Fig. 14).

SVs contribute substantial genetic diversity with important evolutionary and medical implications. By comparing the two haploid genomes, we identified 11,663 SVs (larger than 50 bp), including 6,064 large indels, 27 inversions, 34 translocations, 5,514 copy number variations (CNVs) and 24 inverted translocations (Fig. 2a, Supplementary Table 19). We validated 95.7% of the large indels and 74.2% of the SVs with PacBio long reads, as well as 14 of 17 randomly selected large indels by PCR (Supplementary Fig. 15, Supplementary Table 20). By counting all types of variation between the two haploid genomes, we estimate the overall rate of heterozygosity on the autosomes of the sequenced individual to be around 1.36%.

Large heterogeneous SVs could cause a high incidence of chromosomally unbalanced gametes and thus are normally rare[15]. We found that 72% of SVs were shorter than 1.5 kb, with an average length of about 3.5 kb. The longest SV was a 304-kb inversion (Supplementary Fig. 16). We observed a higher density of LINE (L1) elements around the inversions ($P = 0.03752$, one-sided *t*-test). The indel peak at a length of 300 bp were enriched with Alu repeats (Supplementary Fig. 17a; $P = 2.2 \times 10^{-16}$, Chi-squared test, Supplementary Note). About 33% of the inversion variations between haplotypes were located between two inverted repeat sequences (Supplementary Fig. 17b), indicating that they were introduced by a repeat mechanism[16]. We detected and validated 58 genomic translocation events that differed between the two haplotypes, including 50 genes (Fig. 2a, Supplementary Table 21). About half of the affected genes were completely translocated from one allele to a different genomic location in the other allele. The mechanism driving such translocations remains to be elucidated.

De novo germline mutations

Germline mutations are the source of genetic diversity and the driving force of both evolution and genetic diseases[17]. However, finding de novo germline mutations is a challenging task, as in traditional assemblies less than half of the mutations can be phased to parental origin[18]. A fully diploid assembly enables us to use each parental haplotype independently as a reference to detect de novo mutations, and validate the loci detected independently from the two references as controls for false-positive calls (Methods, Supplementary Note). We detected nine validated de novo mutations in this trio from the approximately 41% of callable sites in both maternal and paternal genomes (Fig. 1a, Supplementary Table 22). The paternal-to-maternal ratio contribution of de novo mutations to the child was 2:1 (Fig. 1b), which is lower than that in humans (4:1)[18] but similar to the closely related owl monkey (2.1:1)[19]. Our results suggest a mutation rate of 0.43×10^{-8} de novo mutations per site per generation for the marmoset. Using this estimated rate and the evolutionary branch length of marmoset substitutions inferred from whole-genome alignments[20], we estimated a divergence time between New World monkeys and humans at around 48.7 million years ago (Ma), which is close to what was estimated from data for the owl monkey[19].

New sex-differentiation region in the marmoset

On the basis of the sequencing depth of parental short reads on the F_1 male assembly (Methods), we identified X-linked sequences of around 147 Mb, with over 99% in a single X chromosome scaffold

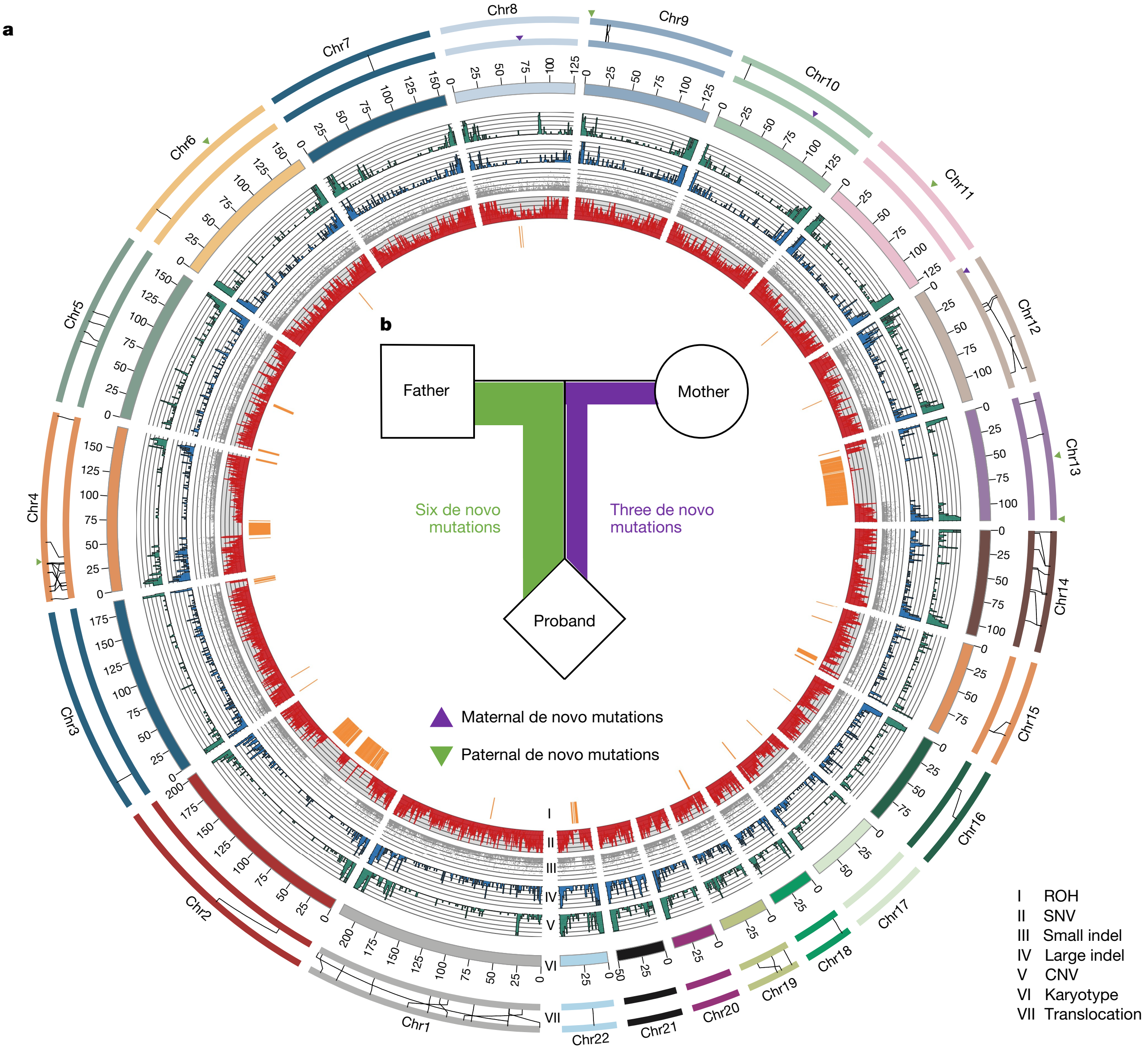


Fig. 1 | Distribution of SNVs, small indels and SVs in a diploid marmoset genome. a, Heterozygosity landscape patterns between the two haploid marmoset genomes. Tracks from inside out (I–VI): distribution of runs of homozygosity (ROH) (>1 Mb), SNV density (window size, 500 kb; range, 0–0.85%), small indel (<50 bp) distribution (*y* axis, indel length), large indel density (≥50 bp; window size, 1 Mb; count, 0–9), CNV density (window size, 1 Mb; count, 0–9) and karyotype. The links in the outermost circles denote differences in translocation events between maternal (inner) and paternal (outer) assemblies (VII). Triangles indicate locations of the de novo mutations in parental alleles. **b**, Schematic showing the proportion of parental sources of the de novo mutations.

(Supplementary Table 23). As the Y chromosome is enriched with repeat elements and segmental duplications, we de-collapsed unplaced and potential Y-linked scaffolds[21] (Supplementary Fig. 18a) then combined read-depth information and Hi-C interactions to identify final Y-linked sequences of 13.85 Mb (Supplementary Fig. 18b, Supplementary Table 24, Methods). This is smaller yet closer to the 20-Mb karyotype estimate[9] and longer than that in other assemblies (Supplementary Table 25).

Our diploid assembly resolved pseudoautosomal regions (PARs) of both the X and the Y chromosome, whereas most other male genomes result in collapsing PARs into one copy with mixed origin. This permits the precise identification of the pseudoautosomal boundary (PAB) in marmosets (Fig. 2a). Marmoset PARs contain nine protein-coding genes, all of which are also found in the human PAR. However, an inversion was found between human and marmoset PARs, and it is likely to occur specifically in the marmoset lineage near its PAB (Fig. 2a, Supplementary Fig. 19). In addition, downstream of this inversion in the X chromosome, we observed a genomic sequence spanning six human PAR orthologues that had become a new sex-differentiation region (SDR) in the marmoset (Fig. 2a). Three genes in the region, *P2RY8Y*, *AKAP17AY* and *ZBEDY*, have been reported to be SDR-linked[22]. We found that they were not collinear with the X chromosome, but were translocated to the middle of the Y chromosome (Fig. 2a, Extended Data Fig. 3, Supplementary Table 26). All of the Y copies accumulated more mutations than their corresponding X copies (Supplementary Fig. 20). Their X–Y genetic divergence was significantly higher than that of the PAR (one-sided *t*-test, $t = 5.7694$, $P = 1.468 \times 10^{-6}$) (Supplementary Table 27), but significantly lower than that of the ancestral SDR

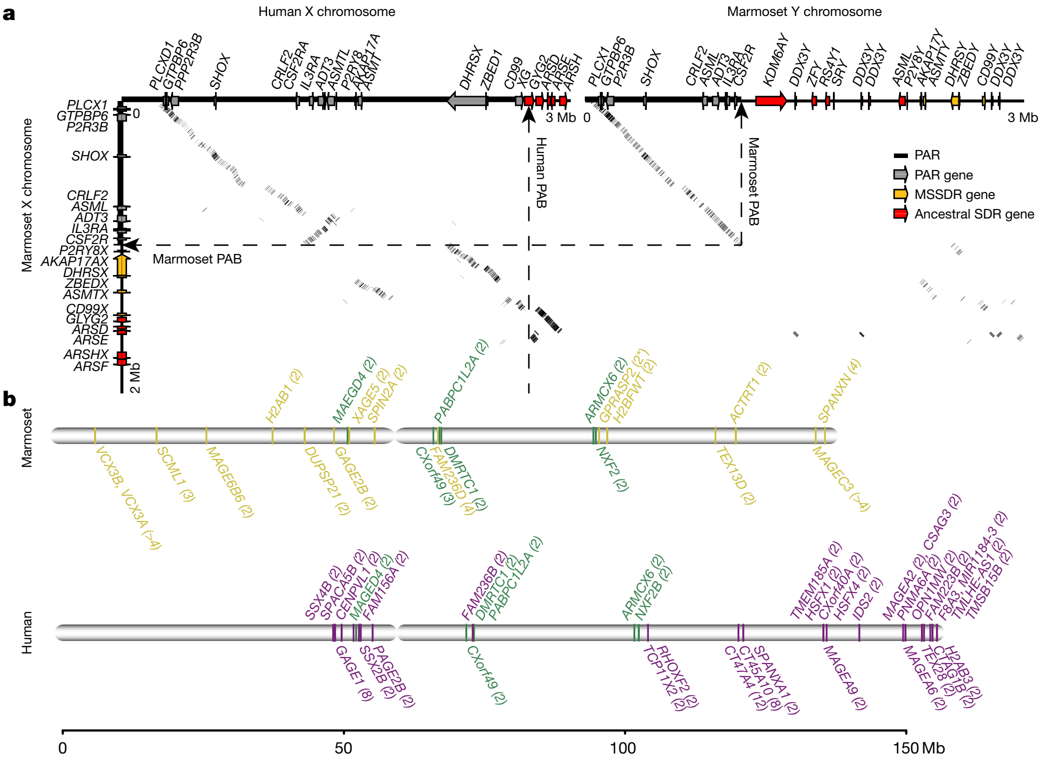


Fig. 2 | Structures of sex chromosomes in marmosets and humans. **a**, Alignment between the marmoset X and Y chromosome reveals a PAR of around 1 Mb in each chromosome. Dashed lines show the boundaries between the PAR and SDR. Alignment between the human and marmoset X chromosome also reveals different PABs between the two species, and an inversion near the marmoset PAB. Grey, PAR genes; orange, MSSDR genes; red, ancestral SDR genes. *ARSE* is also known as *ARSL*. **b**, Distribution of ampliconic genes in the marmoset (yellow) and human (purple) X chromosome. Green, genes that are ampliconic in both species. The copy number for each ampliconic gene is shown in parentheses. Asterisks indicate partial genes. Ampliconic genes with testis-specific expression are shown as the bottom half of the panel for each species. *IDS2* is also known as *IDSP1*.

(one-sided *t*-test, $t = -8.9434$, $P = 3.319 \times 10^{-13}$) (Supplementary Fig. 21), suggesting that its recombination suppression began recently. These new SDR genes also showed a bias in expression in females; however, they were not significantly different from PAR or ancestral SDR genes (Supplementary Fig. 22).

We next applied two divergence-based methods to date the formation of the marmoset-specific SDR (MSSDR) (Supplementary Note, Supplementary Tables 28, 29). On the basis of the marmoset mutation rate estimated above, we inferred that the MSSDR formed at 5.23–9.41 Ma (Supplementary Tables 30, 31). Applying lower mutation rates of the closely related African green monkey (1.11×10^{-9} mutations per position per year (PPPY))[23] and the owl monkey (1.20×10^{-9} PPPY)[24], the formation of the MSSDR was dated at 6.67–12.97 Ma. All of these results indicate that the expansion of the SDR in the marmoset is an evolutionarily young event.

The translocation of the MSSDR on the Y chromosome makes the PAR of the marmoset the shortest among primates recorded so far[25]. As X–Y recombination during male meiosis is limited to the PAR, this region is known to contain the highest per-site recombination rate in the genome[26] and an increased intensity of GC-biased gene conversion[27]. Consistently, we observed a higher GC content in the marmoset PAR relative to the human PAR (one-sided *t*-test, $t = 3.1327$, $P = 0.0011$) (Supplementary Fig. 21). We also observed a 4.3-fold-higher rate of heterozygosity in the marmoset PAR (0.52%) compared to the average rate in autosomes (0.12%) (Supplementary Fig. 23), suggesting that more-intense recombination in the shorter marmoset PAR causes more mutations.

Ampliconic genes—genes with highly similar adjacent copies—are a notable and enigmatic feature of most sex chromosomes[28]. They are often found specifically expressed in the testes and experience a very rapid turnover of copy number[29], leading to the hypothesis that ampliconic genes are involved in sexual antagonism[29]. We detected 22 ampliconic genes on the marmoset X chromosome (Fig. 2b), of which 12 showed testes-restricted expression, at a proportion close to that in humans (40%). Six of the marmoset X-linked ampliconic genes were also present in the human X chromosome with overall similar duplication patterns, suggesting that they originated from a common ancestor (Fig. 2b, Supplementary Fig. 24). The marmoset Y chromosome also contains five multi-copy genes, of which two (*TSPY* and *RBMY*) are also ampliconic genes in the human Y chromosome[30]. These results suggest that the sex-linked ampliconic genes have evolved under a very dynamic duplication process during primate evolution.

Rapid evolution of the marmoset Y chromosome

In contrast to the X chromosome, which maintained overall conserved synteny during primate evolution (Supplementary Fig. 25), we found that the Y chromosome experienced rapid structural changes. This is probably due to the accumulation of mutations as a consequence of Muller's ratchet effect[31]. We detected at least three large inversions and one large translocation involving genes between the male-specific region of the Y chromosome (MSY) in humans and marmosets. The human MSY contained 48 protein-coding genes and the marmoset MSY contained 46, but with different gene properties (Fig. 3a): Twenty-two human MSY genes were absent in the marmoset; of these, 15 of evolved during the evolution of the Hominoidea and the rest were ancestral gametologues that have become inactive or been lost in marmosets (Fig. 3a). Several MSY genes crucial for spermatogenic functions (for example, *HSFY1* and *VCY*) (Supplementary Note) have been lost in marmosets, or lost function owing to frame-shift mutations (for example,

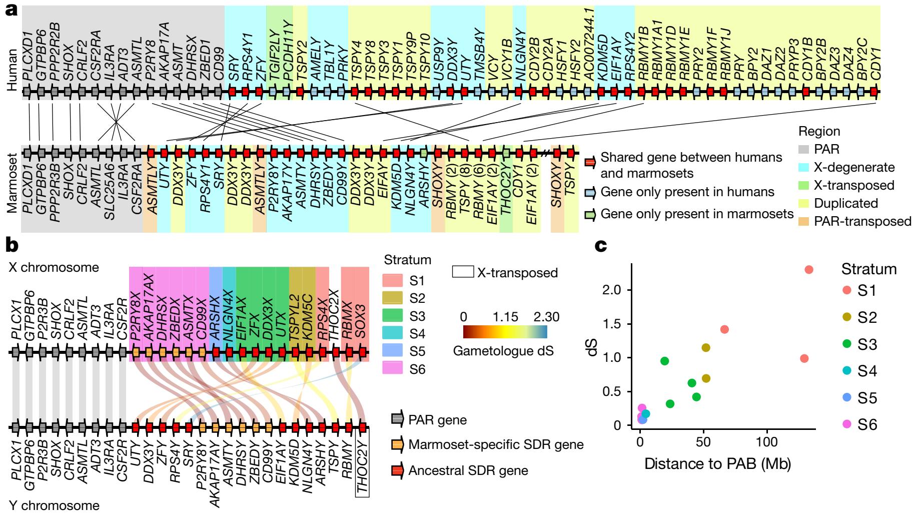


Fig. 3 | Comparison of sex chromosomes across species. a, Y-chromosome gene synteny between humans and marmosets. Lines between human and marmoset indicate one-to-one orthologues. Distance is not drawn to scale. The number of paralogues in unplaced de-collapsed Y-linked scaffolds are marked in parentheses. *ADT3* is also known as *SLC25A6*; *AC007244.1* is ENSG00000286265 under Ensembl release 98. **b**, Six evolutionary strata found in marmoset sex chromosomes. The colour of the links between X and Y gametologues indicates the pairwise dS value. *THOC2X-Y* was not included in any strata because it is a very recently emerged gametologue pair formed via duplication. **c**, Correlation between pairwise dS and X-chromosome position for 14 X–Y SDR gametologues outside the marmoset PAR. Each point represents one gametologue.

USP9Y) (Supplementary Fig. 26). The loss of these genes might be associated with the monogamous social structure of marmosets[32], which potentially alleviates sperm competition. These findings indicate that although it has been claimed that the marmoset has similar patterns of spermatogenesis to humans[33], there are probably some key differences associated with these genes.

By contrast, the marmoset MSY only contains two genes that are absent in humans—*ARSHY* and *THOC2Y*. *THOC2Y* was thought to be lost early in the eutherian common ancestor and exhibits a high rate of synonymous substitutions (dS value) with its gametologue in marsupials[34]. However, we found that the marmoset *THOC2Y* has a very low dS value (dS = 0.0502) with its X-linked gametologue, suggesting that it is not the ancestral gene but a marmoset-specific MSY gene that has recently been duplicated from its X-chromosome counterpart (Supplementary Fig. 27a). In humans, *THOC2* is widely expressed in many tissues and interacts with *XPO4*[35] which mediates the import of SOX2 and SRY proteins. In the marmoset, both *THOC2X* and *THOC2Y* have become testis-specific genes (tissue specificity index (Tau) > 0.8) (Supplementary Fig. 27b). The remaining MSY genes are present in both species, but some show CNVs (Fig. 3a, Supplementary Fig. 28).

Of the 46 marmoset MSY genes, 18 have their gametologues on the X chromosome (Fig. 3b), and their pairwise dS values between X and Y increased with their distance to the PAB on the X chromosome (Pearson's $r = 0.8342$, $P = 0.0002$) (Fig. 3c, Supplementary Table 27), as in humans[36]. According to the sequence divergence as well as the phylogeny, we inferred the presence of six evolutionary strata in marmoset sex chromosomes, which we named from the oldest to the youngest, S1 to S6 (Fig. 3b). S1–S4 are shared with humans[22,36] (Supplementary Fig. 29), suggesting an ancient origin. S5 of the marmoset contained one gametologue pair, *ARSHX-Y*, which has a low pairwise dS value (0.0605) close to that of gametologues in the MSSDR (Supplementary Table 27). In addition, the X copy of the marmoset is clustered with its Y copy instead of the X copies of other primates (Supplementary Fig. 30), suggesting that this stratum formed specifically in New World monkeys. S6 contained six pairs of gametologues, all residing in the MSSDR. The pairwise dS values of S6 gametologues are much lower than those of the ancestral gametologues (Fig. 3b). Notably, three gametologues (*DHRSX-Y*, *ASMTX-Y* and *CD99X-Y*) in S6 display the highest ratio of pairwise non-synonymous to synonymous substitutions rates (dN/dS value) among all gametologues (Supplementary Table 27). Of them, *CD99X* and *CD99Y* show tissue-specific expression in ovary and testis, respectively (Supplementary Table 32). These features imply a strong directional selection link to sex differentiation on these genes once they were translocated from the PAR in the marmoset.

Genetic basis of marmoset biological traits

As a representative species of Callitrichidae, the marmoset has many notable biological traits, such as small body size[37], twinning[12,38], exudate feeding[39] and maintaining bone density during ageing owing to reduced levels of gonadal oestrogen (thus marmosets do not suffer from age-related osteoporosis[40,41]). To further expand our knowledge on the evolution of these biological features, we scanned for and identified 204 positively selected genes (PSGs) in the marmoset genome and 38 PSGs in the common ancestor of New World monkeys (Supplementary Tables 33–35). We have manually checked these PSGs to avoid potential artefacts due to alignment errors or the differences in sequencing and annotation methods across genomes, although we cannot fully rule out the possibility that the differences in quality between the compared assemblies could have affected some of these results. Among these genes, we found two that may be linked to manifesting diminutive size. Mutations of *ZDHHC13* (PSG in marmosets) in mice causes post-translational lipid modification, resulting in weight loss and reduced bone mineral density[42]. *FGFR1* (PSG in New World monkeys) regulates a feedback signal to control the rate of differentiation of osteoblasts[43], and mutations cause autosomal dominant skeletal disorder[44]. (Supplementary Fig. 31).

Marmosets exhibit several unique reproductive adaptations[37], which include sharing a common placental circulation with siblings[45] and the suppression of reproduction in nondominant females[46]. Previous studies have identified several candidate genes that might be related to these traits[12,38]. We found three marmoset PSGs (*PCSK6*, *NR1D1* and *TGIF1*) that might also contribute to their reproductive adaptation. *PCSK6* is expressed in numerous ovarian cell types and *PCSK6*-mutant mice exhibit progressive loss of ovarian function and formation of ovarian pathology[47]. *NR1D1* is a circadian clock gene and might interact with the gonadotropin-releasing hormone signalling pathway[48]. Knockout of this gene in mice reduces fertility[49]. *TGIF1* is a repressor and reversibly modulates members of the TGF-β/SMAD signalling pathway, which has an important role in reproductive processes, including follicular activation, ovarian follicle development and oocyte maturation[50].

We found three marmoset PSGs (*BCL2L14*, *HOMER3* and *CHADL*) involved in osteoclastogenesis and bone metabolism. *BCL2L14* encodes a member of an anti-apoptotic family of proteins, which are known to suppress the functions of osteoclasts[51]. *HOMER3* participates in osteoclastogenesis and bone metabolism. Deletion of this gene markedly decreased tibia bone density, resulting in bone erosion in mice[52]. *CHADL* encodes a collagen-associated small leucine-rich protein and may influence the differentiation of chondrocytes by acting on its cellular microenvironment[53]. Further experiments are needed to investigate the potential roles of the positively selected substitutions in specialized bone metabolism in marmosets.

Captive marmosets in laboratories are intermittently plagued by gastrointestinal disorders[54], which may result from dietary differences in captivity versus the wild[55]. Wild marmosets feed on gums as one of their primary food sources, to acquire energy and minerals[39]. Compared to captive marmosets, the gut microbiome of wild marmosets is more enriched with *Bifidobacterium*[56]. This probiotic bacterium may function to assist the digestion of gum[57]. We found that *PTGS1*, which mediates the gastrointestinal inflammatory reaction, was under positive selection in the marmoset. Expression of this gene is higher in the intestinal mucosa of obese rats than rats of a normal weight[58,59], but its expression is reduced to normal levels when rats are fed with *Bifidobacterium*[59]. It seems that *PTGS1* may have a role in the gastrointestinal function of marmosets, which might be regulated by their exudivore diet through the probiotic bacteria.

Genomic insights for biomedical research

Marmosets are becoming widely used as primate biomedical models in the neurosciences[2]. Here, we compared 2,533 genes related to brain development and neurodegenerative diseases, and found that the majority are highly conserved between marmosets and humans in both sequence and copy numbers (Supplementary Fig. 32). However, we detected 24 genes that show CNVs and 8 genes that are under diversification selection between the two species. These may be associated with differences in the brain between humans and marmosets (Supplementary Fig. 33, Supplementary Tables 36, 37, Supplementary Note).

Pathogenic effects of mutations are highly dependent on their genomic context[60,61]. We therefore scanned the marmoset genome for human pathogenic sites that cause or increase the risk of nervous system diseases. Notably, four genes in marmosets include substitutions that encode amino acids that are pathogenic in humans: *APOE*^{C130R}, *GBA*^{N227S}, *SNCA*^{A53T} and *PAH*^{R176Q} (Supplementary Figs. 34–36, Supplementary Table 38). All of them are fixed in the 12 marmoset individuals with genomic data[13]. Comparison with other primates suggests that the *GBA* and *PAH* genomic contexts are unique to the marmoset (Supplementary Figs. 35, 36). The presence of these two marmoset genes encoding amino acids that are pathogenic in humans suggests that this species might have evolved specific mechanisms to compensate for their pathogenic effects, and highlights the critical need to consider variation in the genomic context when using marmosets as models in human disease research.

Benefits of a diploid assembly

The ultimate goal of creating a reference genome assembly is to produce a gapless, chromosome-level assembly with all sequences fully phased into haplotypes. Several previous efforts have been made towards this goal using the information of a pedigree and/or long reads[5,6]. Our findings demonstrate the power of using a trio-binning approach, in combination with long-read sequencing[7,8], to produce a diploid genome with the two parental haplotypes assembled independently. This method captures the full range of heterozygous variations at high rates of accuracy between the two alleles, resulting in a rate of heterozygosity that is 10 times higher than that found in most genomic studies that use only heterozygous SNVs. Our diploid assembly includes sequences that are more complete for both sex chromosomes—a particular challenge in the case of the Y chromosome with its densely repetitive elements. Whenever trio samples are available, this sequencing and assembly strategy offers the means to generate high-quality, phased reference genomes for a range of species, especially those with high rates of heterozygosity.

Online content

Any methods, additional references, Nature Research 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/s41586-021-03535-x.

1. Aleman, F. The necessity of diploid genome sequencing to unravel the genetic component of complex phenotypes. *Front. Genet*. **8**, 148 (2017).
2. Okano, H., Hikishima, K., Iriki, A. & Sasaki, E. The common marmoset as a novel animal model system for biomedical and neuroscience research applications. *Semin. Fetal Neonatal Med*. **17**, 336–340 (2012).
3. Kishi, N., Sato, K., Sasaki, E. & Okano, H. Common marmoset as a new model animal for neuroscience research and genome editing technology. *Dev. Growth Differ*. **56**, 53–62 (2014).
4. Wood, A. R. et al. Allelic heterogeneity and more detailed analyses of known loci explain additional phenotypic variation and reveal complex patterns of association. *Hum. Mol. Genet*. **20**, 4082–4092 (2011).
5. Chin, C.-S. et al. Phased diploid genome assembly with single-molecule real-time sequencing. *Nat. Methods* **13**, 1050–1054 (2016).
6. Weisenfeld, N. I., Kumar, V., Shah, P., Church, D. M. & Jaffe, D. B. Direct determination of diploid genome sequences. *Genome Res*. **27**, 757–767 (2017).
7. Koren, S. et al. De novo assembly of haplotype-resolved genomes with trio binning. *Nat. Biotechnol*. **36**, 1174–1182 (2018).
8. Rhie, A. et al. Towards complete and error-free genome assemblies of all vertebrate species. *Nature* https://doi.org/10.1038/s41586-021-03451-0 (2021).
9. Sherlock, J. K., Griffin, D. K., Delhanty, J. D. A. & Parrington, J. M. Homologies between human and marmoset (*Callithrix jacchus*) chromosomes revealed by comparative chromosome painting. *Genomics* **33**, 214–219 (1996).
10. Benirschke, K., Anderson, J. M. & Brownhill, L. E. Marrow chimerism in marmosets. *Science* **138**, 513–515 (1962).
11. Sweeney, C., Ward, J. & Vallender, E. J. Naturally occurring, physiologically normal, primate chimeras. *Chimerism* **3**, 43–44 (2012).
12. The Marmoset Genome Sequencing and Analysis Consortium. The common marmoset genome provides insight into primate biology and evolution. *Nat. Genet*. **46**, 850–857 (2014).
13. Sato, K. et al. Resequencing of the common marmoset genome improves genome assemblies and gene-coding sequence analysis. *Sci. Rep*. **5**, 16894 (2015).
14. Nembaware, V., Wolfe, K. H., Bettoni, F., Kelso, J. & Seoighe, C. Allele-specific transcript isoforms in human. *FEBS Lett*. **577**, 233–238 (2004).
15. Anton, E., Blanco, J. & Vidal, F. Meiotic behavior of three D;G Robertsonian translocations: segregation and interchromosomal effect. *J. Hum. Genet*. **55**, 541–545 (2010).
16. Stankiewicz, P. & Lupski, J. R. Genome architecture, rearrangements and genomic disorders. *Trends Genet*. **18**, 74–82 (2002).
17. Acuna-Hidalgo, R., Veltman, J. A. & Hoischen, A. New insights into the generation and role of de novo mutations in health and disease. *Genome Biol*. **17**, 241 (2016).
18. Goldmann, J. M. et al. Parent-of-origin-specific signatures of de novo mutations. *Nat. Genet*. **48**, 935–939 (2016).
19. Chintalapati, M. & Moorjani, P. Evolution of the mutation rate across primates. *Curr. Opin. Genet. Dev*. **62**, 58–64 (2020).
20. Zoonomia Consortium. A comparative genomics multitool for scientific discovery and conservation. *Nature* **587**, 240–245 (2020).
21. Vollger, M. R. et al. Long-read sequence and assembly of segmental duplications. *Nat. Methods* **16**, 88–94 (2019).
22. Bellott, D. W. et al. Mammalian Y chromosomes retain widely expressed dosage-sensitive regulators. *Nature* **508**, 494–499 (2014).
23. Pfeifer, S. P. Direct estimate of the spontaneous germ line mutation rate in African green monkeys. *Evolution* **71**, 2858–2870 (2017).
24. Thomas, G. W. C. et al. Reproductive longevity predicts mutation rates in primates. *Curr. Biol*. **28**, 3193–3197 (2018).
25. Raudsepp, T. & Chowdhary, B. P. The eutherian pseudoautosomal region. *Cytogenet. Genome Res*. **147**, 81–94 (2015).
26. Hinch, A. G., Altemose, N., Noor, N., Donnelly, P. & Myers, S. R. Recombination in the human pseudoautosomal region PAR1. *PLoS Genet*. **10**, e1004503 (2014).
27. Holmquist, G. P. Chromosome bands, their chromatin flavors, and their functional features. *Am. J. Hum. Genet*. **51**, 17–37 (1992).
28. Hughes, J. F. & Page, D. C. The biology and evolution of mammalian Y chromosomes. *Annu. Rev. Genet*. **49**, 507–527 (2015).

29. Mueller, J. L. et al. Independent specialization of the human and mouse X chromosomes for the male germ line. *Nat. Genet*. **45**, 1083–1087 (2013).
30. Lucotte, E. A. et al. Dynamic copy number evolution of X- and Y-linked ampliconic genes in human populations. *Genetics* **209**, 907–920 (2018).
31. Bachtrog, D. A dynamic view of sex chromosome evolution. *Curr. Opin. Genet. Dev*. **16**, 578–585 (2006).
32. Wahab, F., Drummer, C. & Behr, R. Marmosets. *Curr. Biol*. **25**, R780–R782 (2015).
33. Millar, M. R., Sharpe, R. M., Weinbauer, G. F., Fraser, H. M. & Saunders, P. T. Marmoset spermatogenesis: organizational similarities to the human. *Int. J. Androl*. **23**, 266–277 (2000).
34. Cortez, D. et al. Origins and functional evolution of Y chromosomes across mammals. *Nature* **508**, 488–493 (2014).
35. Havugimana, P. C. et al. A census of human soluble protein complexes. *Cell* **150**, 1068–1081 (2012).
36. Lahn, B. T. & Page, D. C. Four evolutionary strata on the human X chromosome. *Science* **286**, 964–967 (1999).
37. Abbott, D. H., Barnett, D. K., Colman, R. J., Yamamoto, M. E. & Schultz-Darken, N. J. Aspects of common marmoset basic biology and life history important for biomedical research. *Comp. Med*. **53**, 339–350 (2003).
38. Harris, R. A. et al. Evolutionary genetics and implications of small size and twinning in callitrichine primates. *Proc. Natl Acad. Sci. USA* **111**, 1467–1472 (2014).
39. Power, M. L. in *The Evolution of Exudativory in Primates* 25–44 (Springer, 2010).
40. Colman, R. J. Absence of estrogen depletion bone loss in female common marmosets. *J. Bone Miner. Res*. **12**, S342 (1997)
41. Binkley, N. et al. Zoledronate prevents the development of absolute osteopenia following ovariectomy in adult rhesus monkeys. *J. Bone Miner. Res*. **13**, 1775–1782 (1998).
42. Saleem, A. N. et al. Mice with alopecia, osteoporosis, and systemic amyloidosis due to mutation in *Zdhhc13*, a gene coding for palmitoyl acyltransferase. *PLoS Genet*. **6**, e1000985 (2010).
43. Iseki, S., Wilkie, A. O. & Morriss-Kay, G. M. Fgfr1 and Fgfr2 have distinct differentiation- and proliferation-related roles in the developing mouse skull vault. *Development* **126**, 5611–5620 (1999).
44. White, K. E. et al. Mutations that cause osteoglophonic dysplasia define novel roles for FGFR1 in bone elongation. *Am. J. Hum. Genet*. **76**, 361–367 (2005).
45. Moore, H. D. M., Gems, S. & Hearn, J. P. Early implantation stages in the marmoset monkey (*Callithrix jacchus*). *Am. J. Anat*. **172**, 265–278 (1985).
46. Saltzman, W., Schultz-Darken, N. J., Severin, J. M. & Abbott, D. H. Escape from social suppression of sexual behavior and of ovulation in female common marmosets. *Ann. NY Acad. Sci*. **807**, 567–570 (1997).
47. Mujoomdar, M. L., Hogan, L. M., Parlow, A. F. & Nachtigal, M. W. *Pcsk6* mutant mice exhibit progressive loss of ovarian function, altered gene expression, and formation of ovarian pathology. *Reproduction* **141**, 343–355 (2011).
48. Cho, H. et al. Regulation of circadian behaviour and metabolism by REV-ERB-α and REV-ERB-β. *Nature* **485**, 123–127 (2012).
49. Chomez, P. et al. Increased cell death and delayed development in the cerebellum of mice lacking the rev-erbA(alpha) orphan receptor. *Development* **127**, 1489–1498 (2000).
50. Zhang, Z. et al. *TGIF1* and *SF1* polymorphisms are associated with litter size in Small Tail Han sheep. *Reprod. Domest. Anim*. **55**, 1145–1153 (2020).
51. Lagasse, E. & Weissman, I. L. Enforced expression of Bcl-2 in monocytes rescues macrophages and partially reverses osteopetrosis in *op*/*op* mice. *Cell* **89**, 1021–1031 (1997).
52. Son, A. et al. Homer2 and Homer3 modulate RANKL-induced NFATc1 signaling in osteoclastogenesis and bone metabolism. *J. Endocrinol*. **242**, 241–249 (2019).
53. Tillgren, V., Ho, J. C. S., Önnerfjord, P. & Kalamajski, S. The novel small leucine-rich protein chondroadherin-like (CHADL) is expressed in cartilage and modulates chondrocyte differentiation. *J. Biol. Chem*. **290**, 918–925 (2015).
54. Ludlage, E. & Mansfield, K. Clinical care and diseases of the common marmoset (*Callithrix jacchus*). *Comp. Med*. **53**, 369–382 (2003).
55. Bailey, M. T. & Coe, C. L. Intestinal microbial patterns of the common marmoset and rhesus macaque. *Comp. Biochem. Physiol. A*. **133**, 379–388 (2002).
56. Malukiewicz, J. et al. The gut microbiome of exudivorous wild and captive marmosets. Preprint at https://doi.org/10.1101/708255 (2020).
57. Turroni, F. et al. Glycan utilization and cross-feeding activities by Bifidobacteria. *Trends Microbiol*. **26**, 339–350 (2018).
58. Wiśniewski, J. R., Friedrich, A., Keller, T., Mann, M. & Koepsell, H. The impact of high-fat diet on metabolism and immune defense in small intestine mucosa. *J. Proteome Res*. **14**, 353–365 (2015).
59. Plaza-Díaz, J. et al. *Adamdec1*, *Ednrb* and *Ptgs1*/*Cox1*, inflammation genes upregulated in the intestinal mucosa of obese rats, are downregulated by three probiotic strains. *Sci. Rep*. **7**, 1939 (2017).
60. Jordan, D. M. et al. Identification of *cis*-suppression of human disease mutations by comparative genomics. *Nature* **524**, 225–229 (2015).
61. Storz, J. F. Compensatory mutations and epistasis for protein function. *Curr. Opin. Struct. Biol*. **50**, 18–25 (2018).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Sample collection, processing and sequencing

Samples were collected at an AAALAC-accredited facility from an F_1 male marmoset (3 months old) at The Rockefeller University, under USDA- and IACUC-approved protocols. The quadriceps muscle was dissected, collected and flash-frozen in liquid nitrogen immediately after euthasol administration; we extracted genomic DNA from the muscle sample. This DNA was used for Bionano optical mapping, PacBio library preparation and SMRT sequencing, 10X Genomics linked-read sequencing, Arima Hi-C library preparation and Illumina sequencing. We collected blood samples from both parents of the F_1 male (mother, 3 years 10 months; father, 3 years 7 months) for Illumina sequencing by shaving the area (thigh for saphenous vein and tail for lateral tail vein), applying 2% lidocaine jelly, prepping the vein with alcohol and collecting less than 2 ml blood per sample (1× sample for male and female) via intravenous blood draw into EDTA tubes.

For annotation purposes, we collected more than 18 tissues from the brother of the F_1 male. Blood was collected from the saphenous vein pre-mortem using the method described above. All additional tissues were dissected, collected and flash-frozen in liquid nitrogen or powdered dry ice immediately after euthasol administration; the brain and testes were dissected at first and all tissues were dissected and frozen within a 30-min period post-mortem. RNA integrity numbers (RINs) for all tissues used for PacBio SMRT sequencing and Iso-Seq analysis ('Sample processing and sequencing' in Supplementary Note) were high, ranging from 8.2 (lung) to 9.9 (cerebellum). We performed Mashmap quality control analyses of sequencing reads to rule out any potential contamination or poor sequencing before assembling (Supplementary Fig. 1).

Sample size, randomization and blinding

We aim to use parental SNVs to determine and phase the two haplotype genomes of the offspring, thus the sample size for genome sequencing is three. Bioinformatic analyses were performed with all available data. Randomization for genome and transcriptome sequencing is not applied in this study. For SNV and indel PCR validation, variation sites were randomly selected by the Linux command 'sort −R'. Blinding was not necessary for genome and transcriptome sequencing or PCR validation of genetic variation. The study aims to identify the genetic differences inherited from parental genomes, so only the DNA sample of the F_1 individual was used for PCR validation.

Genome assembly

We combined the previously developed trio-binning approach[7] and further advanced the Vertebrate Genomes Project (VGP) assembly pipeline[8] for scaffolding, to generate the haplotype-phased marmoset assembly (Supplementary Fig. 2). In the first step, we used TrioCanu (v.1.8+287) to bin PacBio long reads of the F_1 male into maternal and paternal haplotypes using haplotype-specific 21-mer markers generated from the Illumina short reads of the mother and father. After binning, TrioCanu independently generated contigs for each haplotype (haplotigs). From here on, the maternal and paternal haplotigs underwent the same steps independently. Separately, we assembled the mitochondrial genome with the mitoVGP pipeline (v.2.2)[62] and added it to the haplotigs to keep any raw mitochondrial reads from being mapped to nuclear sequences, which would result in lower sequence quality after polishing. We used Arrow from SMRT Link (v.6.0.0.47841) to improve base-calling accuracy and purge_dups (v.1.0.0)[63] in an adapted trio mode to remove overlaps at the ends of contigs. The resulting polished, purged haplotigs were scaffolded in three stages: first, we used the 10X linked-reads in two rounds of Scaff10X (v.4.1.0) (https://github.com/wtsi-hpag/Scaff10X) to generate the primary scaffolds; second, we generated Bionano cmaps and used Bionano Solve (v.3.2.1_04122018)[64] for hybrid scaffolding and to break mis-assemblies; third, we used Salsa2 (v2.2)[65] to generate chromosome-level scaffolds using the molecular contact information from Hi-C linked reads. Finally, we performed a second round of Arrow polishing on the maternal and paternal scaffolds with the binned long reads. During this round of polishing, gaps between contigs were closed by the gap-filling function of Arrow. The parental haplotypes were then combined in a single assembly and underwent two rounds of short-read polishing using Long Ranger (v.2.2.2)[66] for short-read alignment and freebayes (v.1.3.1)[67] for polishing (Supplementary Note). After splitting the scaffolds by haplotype and removing the mitochondrial genome from each assembly, the two assemblies (named mCalJac1.mat and mCalJac1.pat) underwent manual curation using the gEVAL tool[68], in particular to correct structural assembly errors. In the abbreviated name, m is mammal; CalJac is the abbreviated Latin species name; 1 is the first VGP assembly of this species; and mat and pat are maternal and paternal haplotypes, respectively.

Identification of sex-linked sequences and additional Y-chromosome assembly

To identify X-linked and Y-linked sequences in mCalJac1 (GCA_011100555.1), we mapped parental short reads to the assembly with BWA ALN (v.0.7.12)[69]. Coverage was extracted with SAMTools (v.1.2) and normalized by the peak coverage. In the identification of X-linked sequences, the normalized female-versus-male (F/M) coverage ratio was calculated and plotted in a 5-kb window, and scaffolds with a F/M coverage ratio within the range 1.5 to 2.5 were identified as X-linked. In Y-linked sequence identification, the normalized F/M coverage ratio was calculated and plotted in a 2-kb window and scaffolds with a F/M coverage ratio within a 0.0 to 0.3 range were identified as Y-linked. We further manually examined large scaffolds in the maternal and paternal assemblies and included the Y chromosome Super_scaffold_pat_24. This scaffold was missing in the 0.3 cut-off condition because the first 1-Mb sequence shows an equal pattern of female and male coverage as the PAR.

In these previous steps, only Y-linked sequences of around 6 Mb were identified, about 14 Mb smaller than the expected 20-Mb size based on karyotyping. As sex chromosomes are notoriously difficult to assembly, and no primate has had a complete Y chromosome sequenced, to determine whether we missed any Y-chromosomal sequences, we performed additional assembly steps. We used Hi-C interaction information to call back potential Y-linked contigs that were filtered by our strict filtering on the basis of low female read depths. Arima Hi-C reads were mapped to mCalJac1 and the Hi-C interaction matrix was generated by HiCPro (v.2.10.0)[70]. At 10-kb resolution, we extracted the interaction strength of every unplaced scaffold to each autosome, X or Y chromosome. Unplaced scaffolds with more than five interaction strength values to both autosomes/X and Super_scaffold_pat_24 were selected, and the interaction strength with the autosomes/X and the interaction strength with Y was compared for each scaffold by two-sided Wilcoxon rank-sum test. With a false discovery rate (FDR)-corrected *P* value cut-off of 0.01, we further identified 17 scaffolds that show a significantly higher interaction with Super_scaffold_pat_24 than with other chromosomes, and considered them putative Y-linked scaffolds. To validate this result, we collected sequences of bacterial artificial chromosome mapped to the marmoset Y chromosome from NCBI and mapped them to mCalJac1 with minimap2. Almost all BAC sequences mapped to the eight Y-linked scaffolds were identified by the sequencing depth method. One, BAC AC279170.1, was previously missed, but can now be mapped to pat_scaffold_39_arrow_ctg1, which was identified by the Hi-C method. Thus, the dataset identified by the Hi-C method complements the dataset identified by the sequencing depth method. Combining these two datasets, a total of 25 potential Y-linked scaffolds (around 14.13 Mb) were identified from mCalJac1 (Supplementary Table 39).

Next, we mapped the PacBio raw reads to the assembly and found that some of the potentially Y-linked scaffolds had regions of considerably high coverage compared to autosomes and X chromosomes, indicative of collapsed sequences, which would cause the artificially high level of Hi-C interaction and introduce false-positive Y-linked sequences. To de-collapse these regions, we used the Segmental Duplication Assembler (SDA)[21] and mapped the SDA-assembled contigs to their original scaffolds with minimap2 to remove potential assembly artefacts. To replace the original collapsed sequence in the assembly with the most plausible candidate de-collapsed sequence, we applied 'the longest rule': start with the de-collapsed sequence in the SDA output that has the longest stretch mapping back to the original scaffold, then select the second sequence with the longest match that does not overlap the previous one, and so on. Once all the non-overlapping de-collapsed sequences with the longest matches were selected, we filled in the gaps using the original scaffold as a backbone, and left 1,000 'N's (gap indicating unknown nucleotides in the assembly) between each contig.

To further exclude false positives from the de-collapsed Y dataset, we refiltered the sequences with the sex-differential depth ratio and the Hi-C interaction criteria as mentioned above (Supplementary Table 24). However, as only the uniquely mapped reads were used in calculating the Hi-C interaction between unplaced scaffolds and autosomes/X/Y, our results underestimate Y-chromosomal DNA, including many de-collapsed Y scaffolds with multiple copies that might still be missed.

Detection of SNPs, indels and SVs using whole-haplotype genome alignment

To call heterozygous sites between the two haploid sequences, independent of the GenomeScope calculation, we first performed a Mummer (v.3.23) alignment with the parameters of 'nucmer -maxmatch -l 100 -c 500'. Because our assemblies span most repetitive sequences, repeat-masking treatment was not necessary before conducting the Mummer alignment. A series of custom scripts (https://github.com/comery/marmoset) identified and sorted our SNPs and indels in the alignments. We used svmu (v.0.4-alpha)[71], Assemblytics (v.1.2)[72], and SyRi (v.1.0)[73], to detect SVs from Mummer alignment. After several test rounds, we found that svmu reported more accurate large indels, and Assemblytics detected CNVs, particularly tandem repeats, whereas SyRi detected other SVs well. We used these three methods and combined the results as confident SVs. We used default parameters for svmu, Assemblytics, and recommended nucmer alignment for SyRi (https://schneebergerlab.github.io/syri/).

To generate a high-quality SV dataset, we manually checked all inversions and translocations with the following steps: (1) clip 300 bp of upstream/downstream flanking sequence of each break point between the two haplotypes, blast against local PacBio reads with threshold identity >96% and aligned length >550 bp, and require the SV region where the maternal and paternal sequences aligned to have high similarity (>90%); (2) if (1) failed, then check the 10X linked-read count between a 5-kb flanking region; (3) if any break point is not supported by 10X linked-reads, check the Hi-C heat map of this region; if it shows an inversion or translocation pattern on heat map or an ambiguous situation, then remove it.

To evaluate the accuracy of SV detection, we searched the binned PacBio reads around the break points of both maternal and paternal assemblies for all indels in chromosome 1. We looked for one of the following three features to determine the indel as accurate: (1) at least one single PacBio long read from each haplotype that spans the entire indel region with the variation found in each haplotype; (2) overlapping PacBio reads that span the two break points; or (3) manually validated PacBio read alignment by the Integrative Genomics Viewer (IGV)[74]. Finally, we found that 95.7% of indels are correct when considering the breakage location; however, 74.2% are accurate when considering both boundary and location.

Estimation of sequencing error and polishing error

To calculate sequencing errors and polishing errors, we established a confident SNP set as a criterion. We used three individual approaches to detect SNPs between two haplotypes: (1) retrieved heterozygous sites from the Mummer alignment between the maternal and paternal haplotypes excluding the sex chromosomes (setA, containing 3.48 million SNVs); (2) GATK pipeline based on mapping of 10X linked-reads from the F_1 offspring (setB); and (3) SAMTools (v.1.8) mpileup followed by bcftools also based on 10X linked-reads mapping (setC). Then, a raw SNP dataset was generated by a two-step procedure: first taking the intersection of setB and setC to generate Set1 (3.72 million SNVs), followed by taking the union of setA and Set1 to get Set2 (3.77 million SNVs). We then took these two sets and selected among them to a high-quality 3.58-million SNP Set3 (Supplementary Fig. 10) with the following criteria applied: (1) 10X linked-read depth lower than 10; (2) filter out sites that do not align to the two haplotype assemblies; (3) filter out sites that we could not call a typical haplotype on the basis of much less than 50% nucleotide distribution ($\pi > 0.4$ and the third highest depth >5, in which π is calculated as: $\pi = 2 \times (AT + AC + AG + TC + TG + CG)/(\mathrm{Totaldepth} \times (\mathrm{Totaldepth} - 1))$

and A, T, C and G represent the sequencing depth of base A, T, C and G for each site. For example, a distribution of 'A:20; T:20; C:14; G:0' indicates a complex condition. We also collected the mapping information from raw PacBio reads and corrected PacBio reads. This allowed us to establish an evidence chain of how the bases in each haplotype changed during assembling and polishing, which allowed us to classify different error types. We classified 195,751 sequencing error sites and 180,712 polishing error sites. The sequencing and polishing error rates were estimated to be 3.41×10^{-5} and 3.66×10^{-5}, respectively. We further validated the variants with PCR experiments (Supplementary Note).

Mutation rate analysis

The 10X linked-reads of the F_1 offspring and the parents' short reads were mapped to each genome assembly independently (paternal and maternal assemblies). Duplicate reads and reads that mapped to more than one region were removed. Variants were called using GATK4 HaplotypeCaller in base-pair resolution mode, calling each single site of the genome. Two independent joint genotypes were produced: one for the three individuals (mother, father and F_1 offspring) mapped to the maternal assembly and one for the three individuals mapped to the paternal assembly. We identified a maternal candidate de novo mutation as a site for which the parents were homozygous for the reference (0/0) and the offspring was heterozygous (0/1) when mapped to the paternal genome. For validation, such a candidate site would be expected to have the parents homozygous for the alternative (1/1), and the offspring heterozygous (0/1) when mapped to the maternal genome. Similarly, a paternal candidate de novo mutation was identified as a site for which the parents were homozygous for the reference (0/0), and the offspring was heterozygous (0/1) when mapped to the maternal genome. Here, again, those candidates were validated if they also appeared in the parents as homozygous for the alternative (1/1), and in the offspring heterozygous (0/1) when mapped to the paternal genome. Additional filters were applied for sites, genotype quality, read depth and number of alternative alleles in the parents and allelic balance in the offspring (Supplementary Note). Finally, we removed any potential sites with sequencing errors, polishing errors or assigning errors, as well as sites that failed the PCR validation. To calculate a rate, we computed the number of callable sites in each genome as the number of sites for which both parents were homozygous for the reference and all individuals passed the depth coverage between half and two times the average depth for each individual, number of alternative alleles allowed, and genotype quality filters. We corrected those callable sites by a negative rate factor, alpha (α), which is the percentage of callable sites that would be filtered away by our site filters (following a known distribution) and the allelic balance filter (which corresponds to the number of sites for which one parent was homozygous for the reference allele, the other parent was homozygous for the alternative allele, and the offspring would be heterozygous, but the reads supporting each allele would be outside our allelic balance filter). The mutation rate was calculated as:

$$\mu = \frac{\text{Mutations}_{\text{maternal}} + \text{Mutations}_{\text{paternal}}}{\text{Callability}_{\text{maternal}} \times (1 - \alpha_{\text{maternal}}) + \text{Callability}_{\text{paternal}} \times (1 - \alpha_{\text{paternal}})} \,.$$

Confirmation of the order of Y-linked sequences

Marmoset Y-chromosome-specific BAC end reads[22] were obtained from the NCBI trace archive and mapped to Y-linked sequences with BWA MEM. Only the primary alignment was kept for each read. BAC location on the Y chromosome from a previous report[22] was also obtained and visualized in a dot plot to confirm the order of the Y-linked sequences in mCalJac1. To confirm the MSSDR translocation in the Y chromosome, we further checked PacBio and 10X linked-reads support at the flanking break point of the MSSDR of the Y chromosome.

Detection of PSGs

We used the BLAST reciprocal best hits (RBH) method (Supplementary Note) to identify high-confidence one-to-one orthologous genes among species, including three other New World monkeys (white-faced capuchin (*Cebus capucinus*), Ma's night monkey (*Aotus nancymaae*) and black-capped squirrel monkey (*Saimiri boliviensis*)); three old world primates (human (*Homo sapiens*), macaque (*Macaca mulatta*) and chimpanzee (*Pan troglodytes*)); and three outgroups (treeshrew (*Tupaia glis*), mouse (*Mus musculus*) and cow (*Bos taurus*)). The marmoset was set as foreground when detecting marmoset-specific PSGs, whereas the New World monkeys were set as foreground when detecting New World monkey-specific PSGs. A total of 13,995 one-to-one orthologous genes were identified. To minimize the effect of gene annotation, we retrieved the corresponding coding sequences that shared the same isoform with human. These genes were used as an input dataset to conduct multiple sequence alignment using PRANK (v.170427)[75] and guidance (v.2.02)[76] to improve the alignment. The positive selection sites within a specific lineage were detected by branch-site model in PAML (v.4.9i)[77]. Genes with an FDR-adjusted *P* value of less than 0.05 were treated as candidates for positive selection. To minimize effects of assembly and alignment, we filtered candidate PSGs if (1) the positively selected site has gaps in more than two species; (2) the positively selected sites had more than two non-synonymous substitution forms (ignoring outgroup), and (3) the flanking region (±10 amino acids) showed over-alignment across species. We also performed a manual check for all individual PSGs to avoid any other false-positive caused by annotation or alignment. Finally, we used read mapping to check the PSG sites to avoid sequencing errors. After filtering, the numbers of PSGs with high confidence detected in marmosets and New World monkeys were 204 and 38, respectively.

Scan for pathogenic or risky mutations in marmosets

Mutation information was obtained from ClinVar (https://ftp.ncbi.nlm.nih.gov/pub/clinvar/tab_delimited/variant_summary.txt.gz, on 30 June 2020) and mutations that were designated to be pathogenic or risky were extracted. Nervous-system-related mutations were extracted with the following keywords: adrenoleukodystrophy, Alzheimer, amyotrophic lateral sclerosis, Angelman, ataxia telangiectasia, Charcot-Marie-Tooth, Cockayne, deafness, Duchenne muscular dystrophy, epilepsy, fragile X syndrome, Friedreich ataxia, Gaucher, Huntington, Lesch-Nyhan syndrome, maple syrup urine disease, Menkes syndrome, myotonic dystrophy, narcolepsy, neurofibromatosis, Niemann-Pick disease, Parkinson disease, phenylketonuria, Refsum disease, Rett syndrome, spinal muscular, spinocerebellar ataxia, Tangier disease, Tay-Sachs disease, tuberous sclerosis, Von Hippel-Lindau syndrome, Wilson disease. Related protein sequences of humans and marmosets were extracted and aligned with PRANK and targeted amino acid sites were scanned to determine whether the human pathogenic or risky mutation is in the marmoset. The genomic coordinates of related codons were extracted to check the alignment of the 12 marmoset individuals with whole-genome-sequencing data. Alignment was visualized and manually examined with Jalview (v.2.11.1.0)[78].

Reporting summary

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

Data availability

Raw sequencing data for the marmoset trio is available under the GenomeArk github (https://vgp.github.io/genomeark/Callithrix_jacchus/). Curatorial information and data mappings to maternal and paternal assemblies are available on the genome evaluation browser, gEVAL (https://vgp-geval.sanger.ac.uk/all_genomes.html). The maternal, paternal, and combined (paternal autosomes and Y chromosome + maternal X chromosome + mitochondrial) assemblies, as well as PacBio Iso-Seq data for annotation, are available under the NCBI BioProject PRJNA560230. The genome assemblies have also been deposited at the CNGB Sequence Archive (CNSA) of the China National GeneBank Database (CNGBdb) with accession numbers CNP0001310 and CNP0001311.

Code availability

The assembly pipeline is available at https://github.com/VGP/vgp-assembly; see Supplementary Tables 2, 3 for the full list of tools used, versions and availability. Workflows and applets built for the VGP are available at DNAnexus (https://www.dnanexus.com/). Custom scripts are available at https://github.com/comery/marmoset and https://github.com/gf777/misc/tree/master/marmoset%20Y.

62. Formenti, G. et al. Complete vertebrate mitogenomes reveal widespread gene duplications and repeats. Preprint at https://doi.org/10.1101/2020.06.30.177956 (2020).
63. Guan, D. et al. Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics* **36**, 2896–2898 (2020).
64. Lam, E. T. et al. Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. *Nat. Biotechnol.* **30**, 771–776 (2012).
65. Ghurye, J. et al. Integrating Hi-C links with assembly graphs for chromosome-scale assembly. *PLOS Comput. Biol.* **15**, e1007273 (2019).
66. Bishara, A. et al. Read clouds uncover variation in complex regions of the human genome. *Genome Res.* **25**, 1570–1580 (2015).
67. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. Preprint at https://arxiv.org/abs/1207.3907 (2012).
68. Chow, W. et al. gEVAL - a web-based browser for evaluating genome assemblies. *Bioinformatics* **32**, 2508–2510 (2016).
69. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
70. Servant, N. et al. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol.* **16**, 259 (2015).
71. Chakraborty, M., Emerson, J. J., Macdonald, S. J. & Long, A. D. Structural variants exhibit widespread allelic heterogeneity and shape variation in complex traits. *Nat. Commun.* **10**, 4872 (2019).
72. Nattestad, M. & Schatz, M. C. Assemblytics: a web analytics tool for the detection of variants from an assembly. *Bioinformatics* **32**, 3021–3023 (2016).
73. Goel, M., Sun, H., Jiao, W. B. & Schneeberger, K. SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biol.* **20**, 277 (2019).
74. Robinson, J. T. et al. Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).
75. Löytynoja, A. Phylogeny-aware alignment with PRANK. *Methods Mol. Biol.* **1079**, 155–170 (2014).
76. Sela, I., Ashkenazy, H., Katoh, K. & Pupko, T. GUIDANCE2: accurate detection of unreliable alignment regions accounting for the uncertainty of multiple parameters. *Nucleic Acids Res.* **43**), W7–W14 (2015).
77. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).
78. Waterhouse, A. M., Procter, J. B., Martin, D. M., Clamp, M. & Barton, G. J. Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**, 1189–1191 (2009).

Acknowledgements This work was supported by the Strategic Priority Research Program of the Chinese Academy of Sciences (XDB31020000); the National Key R&D Program of China (MOST) grant 2018YFC1406901; the International Partnership Program of the Chinese Academy of Sciences (no. 152453KYSB20170002); the Carlsberg Foundation (CF16-0663); the Villum Foundation (no. 25900) to G.Z.; Howard Hughes Medical Institute and Rockefeller

University start-up funds to E.D.J.; the Intramural Research Program of the National Human Genome Research Institute, the National Institutes of Health (NIH) (A.R., A.M.P. and S.K.); the Korea Health Technology R&D Project through the Korea Health Industry Development Institute HI17C2098 (A.R.); the NIH National Institute of General Medical Sciences (no. T32GM007739) and an NIH National Institute of Mental Health F30 (no. MH112351) (M.M.F.); and the Guangdong Provincial Academician Workstation of BGI Synthetic Genomics (no. 2017B090904014) (H.Y.).

Author contributions G.Z. and E.D.J. initiated and designed the project. S.M., J.M., B.H., J. Balacco, M.M.F., O.F., W.A.F. and H.Y. coordinated and performed sample collection and sequencing. S.M., G.F., J.W., W.C., K.H., A.R., M.P., A.M.P., S.K., Y. Zhou, X.B., Z.S. and G.Z. performed genome assembling, curation and evaluation. C.Y., Y. Zhou, S.M. and G.F. performed the chimeric analysis. C.Y., L.A.B., X.B., C.Z. and G.Z. performed genetic diversity analysis. L.A.B. and G.Z. calculated mutation rates. C.Y. and S.T. performed experimental validation. Y. Zhou, L.Z., J. Bergman, M.M.C.R., G.Z. and M.H.S. performed analysis of sex chromosomes. C.Y., Y. Zhou, Y.D., M.F., C.Z., D.X. and Y. Zhu. performed positive-selection analysis. Y. Zhou, Z.S. and G.Z. performed brain- and disease-related analysis. G.Z., E.D.J., M.H.S, C.Y., Y. Zhou, S.M., L.A.B., J. Bergman, M.M.C.R., G.F., X.B. and Z.S. wrote the manuscript.

Competing interests The authors declare no competing interests.

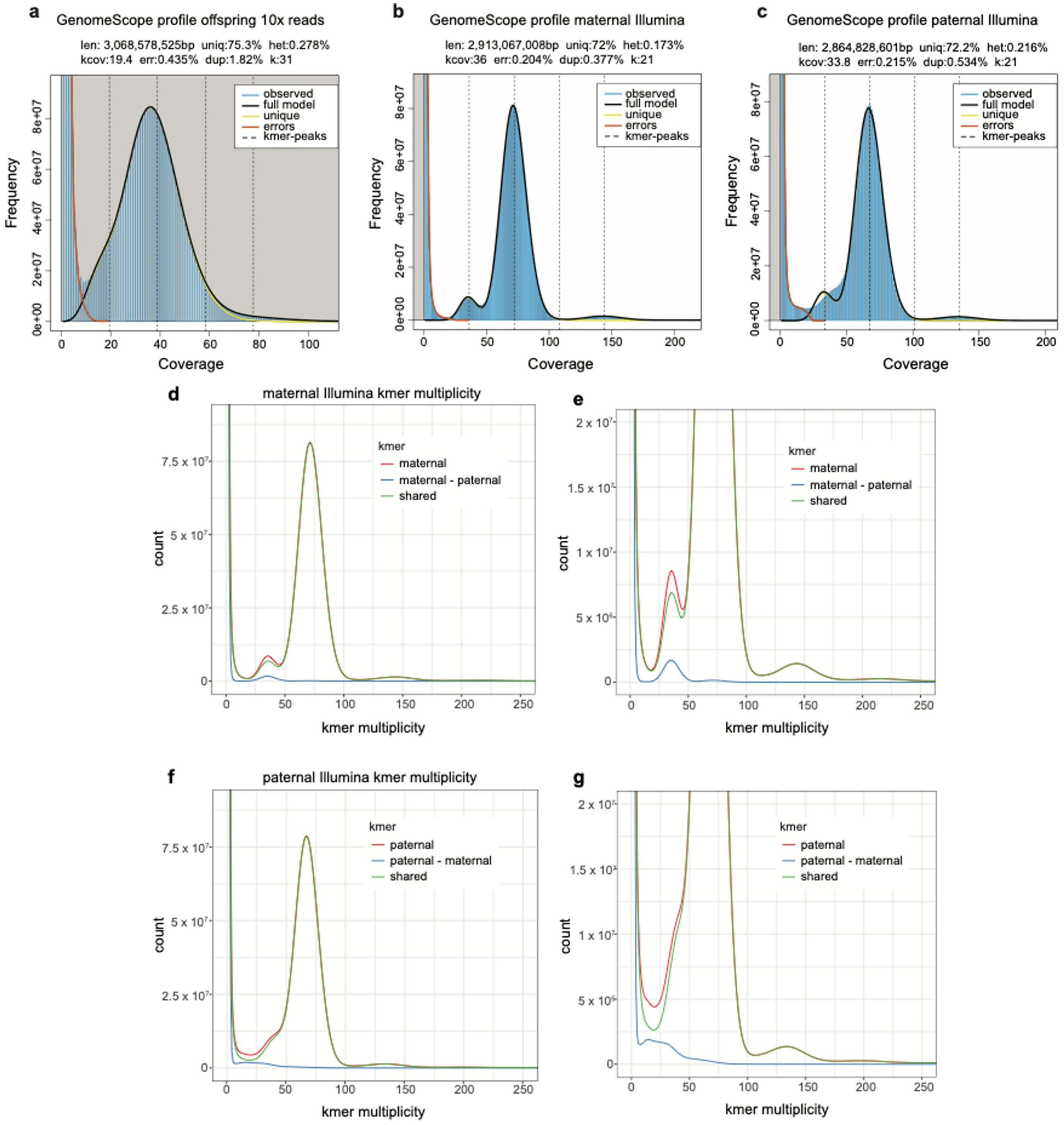
Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-021-03535-x.

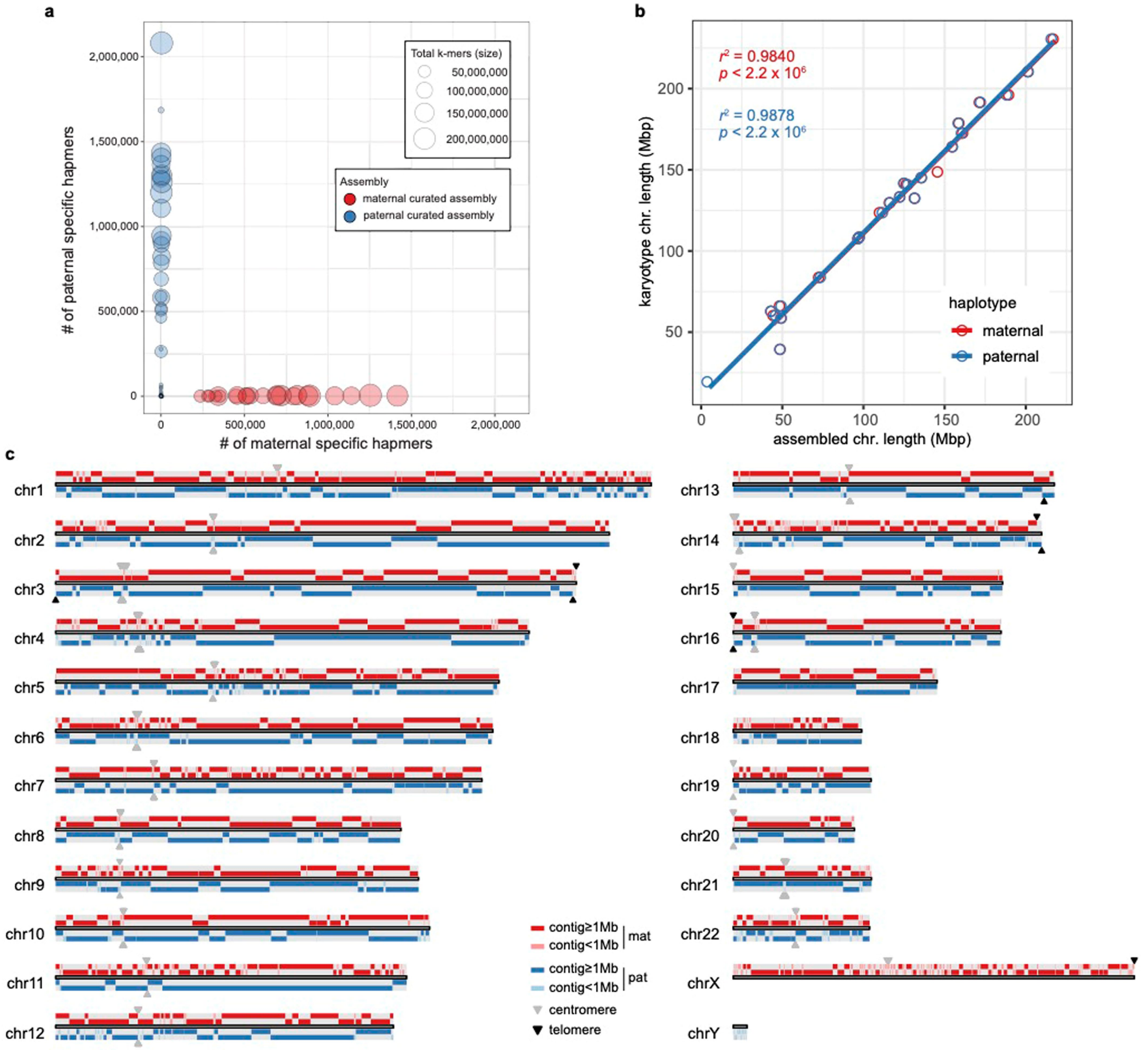
Correspondence and requests for materials should be addressed to G.Z.

Peer review information *Nature* thanks Joanna Malukiewicz, Vagheesh Narasimhan and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

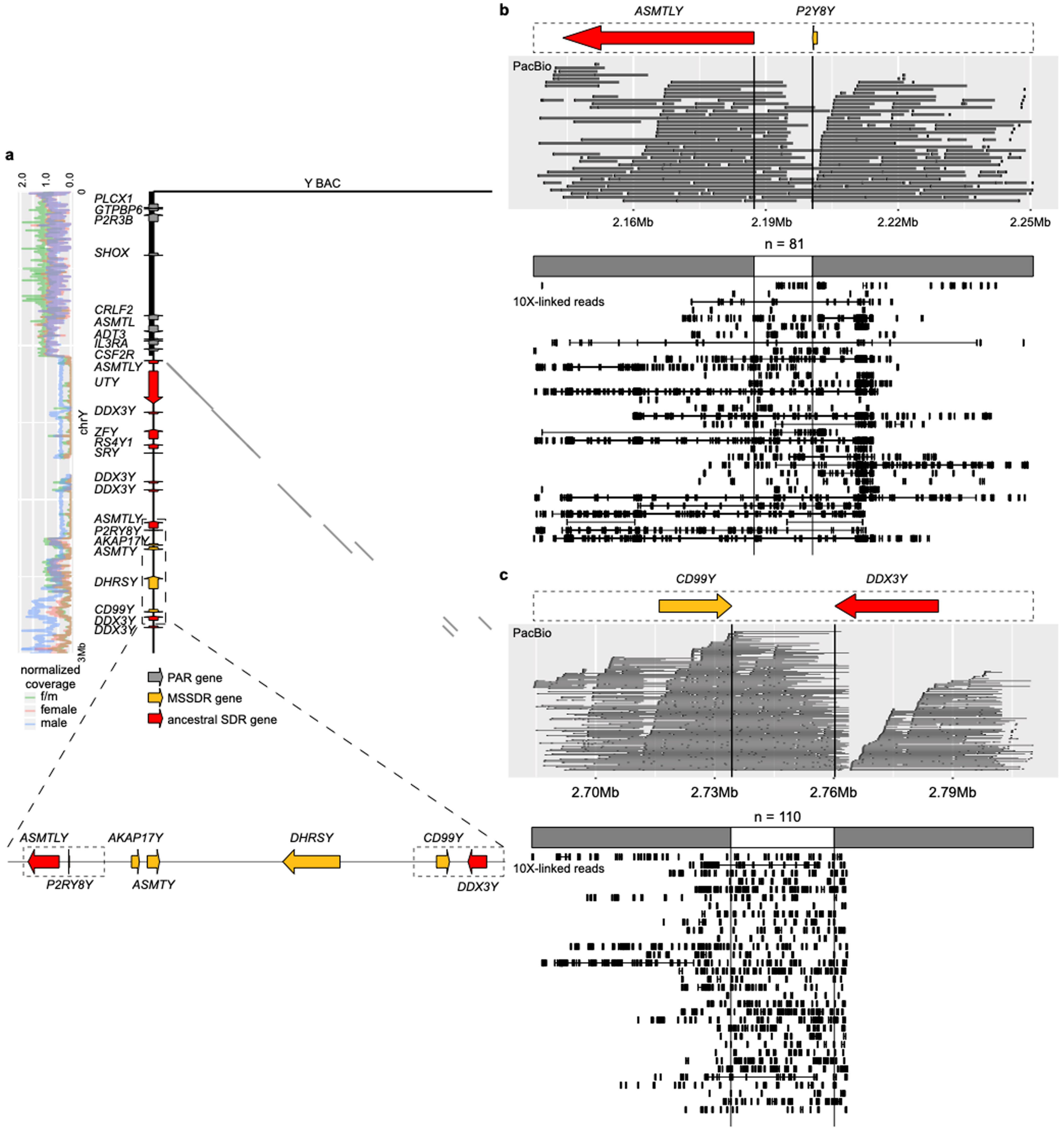
Reprints and permissions information is available at http://www.nature.com/reprints.



Extended Data Fig. 1 | GenomeScope analyses. a, GenomeScope (v.1.0) profile for 31-mers collected from the F_1 10X linked-reads using Meryl (https://github.com/marbl/meryl) (following GEM (gel-bead in emulsion) barcode trimming). Heterozygosity estimated at a maximum of 0.287%. Read error rate estimated at a maximum of 0.435%. Genome haploid length estimated at a maximum of 3,068,578,525 bp, repeat length estimated at a maximum of 757,852,942 bp and unique length estimated at a maximum of 2,310,725,582 bp. **b**, **c**, Genomescope profiles of the maternal (**b**) and paternal (**c**) 21-mers collected from the raw Illumina data. The observed paternal data do not fit GenomeScope's robust model (black line) for a diploid organism and exhibit higher overall heterozygosity than the maternal data (0.216% compared to 0.173%). This supports a premise that the father's sequencing reads contain a level of chimerism, whereas the mother's reads contain negligible representation of alternative alleles, at most. Further analysis of the parental Illumina data shows that the *k*-mer multiplicity distribution varies greatly between the maternal and paternal sets. **d**–**g**, The maternal *k*-mers (**d**, **e** (**e** shows a magnified version of **d**)) show clear distributions with a distinct haploid peak at half coverage (around 35×), whereas the paternal *k*-mers (**f**, **g** (**g** shows a magnified version of **f**)) show an irregular distribution with no clearly defined haploid peak. This provides further evidence that the paternal data exhibit a level of chimerism.



Extended Data Fig. 2 | Trio-based diploid genome assembly. a, Hapmer (haplotype-specific *k*-mer) blob plot of the curated marmoset assemblies. Red, maternal haplotype; blue, paternal haplotype. The size of each blob indicates the total number of *k*-mers counted in an individual scaffold and the position of each blob is plotted according to the number of contained maternal and paternal hapmers. We see that maternal and paternal hapmers are highly phased, with some slight representation of paternal hapmers in several maternal scaffolds (those that do not lie directly on the *x* axis). We can also see a higher representation of paternal hapmers identified within scaffolds of the paternal assembly than maternal hapmers identified in scaffolds of the maternal assembly. **b**, Correlation between the assembled chromosome sizes and the chromosome lengths estimated by karyotype image data. A total of 23 chromosomes are plotted and the coefficient of determination is calculated for each assembly. **c**, Schematic plot mapping the assembled maternal and paternal assigned contigs onto marmoset assembled chromosomes. Top, maternal alleles; bottom, paternal alleles. Contig sizes, centromeres and telomeres are indicated.



Extended Data Fig. 3 | Confirmation of the MSSDR translocation in the marmoset Y chromosome. **a**, Marmoset Y-chromosome-specific BAC reads were obtained from the NCBI trace archive and constructed into a pseudo-Y chromosome according to their position from a previous study[20]. The linear alignment between mCalJac1's Y chromosome and marmoset bacterial artificial chromosome mapped to the Y chromosome confirms the MSSDR translocation. The MSSDR translocation on the Y chromosome is highlighted in yellow and the two regions that span the break points and its flanking 50 kb are highlighted in dashed boxes. **b**, The region spanning *ASMTLY* and *P2RY8Y* is supported by PacBio reads and 10X linked-reads (only a proportion of them were shown). In the 10X linked-reads panel, each rectangle represents a read and each line represents a 10X DNA molecule. A total of 81 10X linked-read DNA molecules support the linkage of *ASMTLY* and *P2RY8Y*. **c**, The region spanning *CD99Y* and *DDX3Y* is supported by PacBio reads and 10X linked-reads (only a proportion of them shown). A total of 110 10X linked-read DNA molecules support the linkage of *CD99Y* and *DDX3Y*.

nature research

Corresponding author(s): Guojie Zhang

Last updated by author(s): Mar 31, 2021

Reporting Summary

Nature Research 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 Research 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	Data collection did not involve any software or code.
Data analysis	Common bioinformatic and statistical analysis software packages were used, including: TrioCanu (v1.8+287), smrtlink (v6.0.0.47841), purge_dups (v1.0.0), scaff10x (v4.1.0), Bionano Solve (v3.2.1_04122018), Salsa2 (v2.2), mitoVGP pipeline (v2.2), longranger (v2.2.2), freebayes (v1.3.1), gEVAL (https://vgp-geval.sanger.ac.uk), Mummer (v3.23), minimap2 (v2.13), bwa (v0.7.17-r1188), refaligner (7437.7523rel), hicpro (v2.10.0), svmu (v0.4-alpha), Assemblytics (v1.2), SyRi (v1.0), Integrative Genomics Viewer (v2.8.6), GATK (v4.1.4.1), samtools (v1.8 & v1.2), NGMLR (v0.2.7), BCFtools (v1.8, v1.9-102-g958180e), ggplot2 (v3.3.2), circos (v0.69-8), BatchPrimer3 (v1.0), BLAST+ (v2.9.0+), cdhit (v4.8.1), BLAST (v2.7.1, v2.2.26), GeneWise (v2.4.1), exonerate (v2.2.0), LASTZ (v1.04.00), PRANK (v150803, v170427), Gblocks (v0.91b), PAML (v4.8 & v4.9i),HISTA2 (v2.0.5), DESeq (v1.9.12), RaxML (v8.2.9), orthoMCL (v1.4), TreeBest (v1.9.2), Jalview (v2.11.1.0), Mashmap (v2.0), proc10xG (v0.0.2), meryl (v1.0), Merqury (v1.0), genoPlotR (v0.8.9), MGRA2 (v2.2), SDA (git commit 4ca0c07), guidance (v2.02) Custom scripts are open source and available on GitHub at https://github.com/comery/marmoset and https://github.com/gf777/misc/tree/master/marmoset%20Y.

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 Research 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 list of figures that have associated raw data
- A description of any restrictions on data availability

Raw sequencing data for the marmoset trio is available under the GenomeArk github (https://vgp.github.io/genomeark/Callithrix_jacchus/). Curatorial information and data mappings to maternal and paternal assemblies are available on the genome evaluation browser, gEVAL (https://vgp-geval.sanger.ac.uk/all_genomes.html). The maternal, paternal, and combined (paternal autosomes and Y chromosome + maternal X chromosome + mitochondrial) assemblies, as well as PacBio Iso-Seq data for annotation, are available under the NCBI BioProject PRJNA560230 (http://www.ncbi.nlm.nih.gov/bioproject/PRJNA560230). The genome assemblies have also been deposited at CNSA of CNGBdb with accession CNP0001310 and CNP0001311. Chimpanzee NGS reads are obtained from ERP002376. The human SNV data of HG00096 was obtained from https://www.internationalgenome.org/. Published marmoset genomes are obtained with accession code GCA_000004665.1, GCA_001269965.1, GCA_002754865.1, GCA_009663435.1, GCA_009811775.1. Genomes used in brain related genes study include: human (hg38), marmoset (mCalJac1), chimpanzee (Clint_PTRv2), rhesus macaque (rheMacS), Ma's night monkey (Anan_2.0), and Chinese tree shrew (TS_2.0). Genomes used in positive selection section include: cow, human, chimpanzee, mouse from Ensembl 98 and Chinese tree shrew (TS_2.0), Cebus capucinus (GCF_001604975.1), Saimiri boliviensis (GCF_000235385.1), Aotus nancymaae (GCF_000952055.2) from NCBI.

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

We aim to use parental SNV to determine and phase the two offspring haplotype genome, thus the sample size for genome sequencing is three. Bioinformatic analyses were performed with all available data.

Data exclusions

Sex chromosomes are excluded in genetic variation analysis.
In PCR validation, we excluded SNPs located in repeat elements.
Variations in chimeric regions were excluded. Various filters were applied at the potential Mendelian violation to reduce false-positive calls, especially at chimerism sites. The first filter was on the site and applied as follows: QD < 2.0, FS > 20.0, MQ < 40.0, MQRankSum < -2.0, MQRankSum > 4.0, ReadPosRankSum < -3.0, ReadPosRankSum > 3.0, SOR > 3.0. The second set of filters were applied to each individual:
- a depth filter DP < 0.5 × individual average depth and DP > 2 × individual average (average depth offspring: 40.5X, father: 72.6X, and mother: 76.9X). This filter would remove any high coverage caused by mapping problems and low coverage sites that are more sensitive to false-positive calls.
- a genotype quality filter GQ < 99 for at least one individual. This filter was set particularly high (generally GQ < 40 to 60 in other de novo studies) to avoid a maximum of chimerism sites in the father, as those sites tend to have a lower genotype quality due to the presence of multiple alleles.
- an alternative allele filter AD > 0 allowed in the homozygous parents. Again, this filter was set stringent with no alternative allele allowed in any parents as most of the chimerism sites would present at least a few alternative alleles in the variant calling files.
- an allelic balance filter AB < 0.3 and AB > 0.7 on the reads supporting the alternative allele in the heterozygous offspring. This filter would remove any potential sequencing errors in the offspring or chimerism cells as those should present a lower allelic balance (~10-20 %) than the real de novo mutations (~50 %).
In positive selection gene analysis, to minimize effects of alignment, we filtered genes based on the condition of its positively selected sites following these criterions, 1) sites with gap number more than 2 were excluded; 2) sites with nonsynonymous substitutions larger than 2 were excluded; and 3) more complicated cases found manual checks. If one gene had no confident site, the gene would be removed.

Replication

Experiments performed in this study aim to validate the variation between the two alleles of the offspring, thus the experiments were performed based on the offspring DNA sample and replication is not applied in this study.

Randomization

Randomization for genome and transcriptome sequencing is not applied in this study. For SNV and indel PCR validation, variation sites were randomly selected by Linux command "sort -R".

Blinding

Blinding was not necessary for genome and transcriptome sequencing, as well as genetic variation PCR validation. The study aim to study the genetic difference inherent from parental genome, so only the F1 individual DNA sample is used for PCR validation.

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
☒	□ Human research participants
☒	□ Clinical data
☒	□ Dual use research of concern

Methods

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

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	Species: Callithrix jacchus. No unique strain. Male and female animals used. Ages: mCalJac1 (M) = 3 months, mCalJac2 (M) = 3 years, mCalJac3 (F) = 3 years, mCalJac4 (M) = 1.5 years.
Wild animals	Study did not involve wild animals.
Field-collected samples	Study did not involve field-collected samples.
Ethics oversight	USDA, AAALAC, and The Rockefeller University IACUC

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