The common marmoset genome provides insight into primate biology and evolution

Kim C. Worley  
*Baylor College of Medicine*

Wesley C. Warren  
*Washington University in St. Louis*

Jeffrey Rogers  
*Baylor College of Medicine*

Devin Locke  
*Washington University in St. Louis*

Donna M. Muzny  
*Baylor College of Medicine*

See next page for additional authors

Follow this and additional works at: [https://digitalcommons.lsu.edu/biosci_pubs](https://digitalcommons.lsu.edu/biosci_pubs)

**Recommended Citation**

Authors


This article is available at LSU Digital Commons: https://digitalcommons.lsu.edu/biosci_pubs/120
The common marmoset genome provides insight into primate biology and evolution

The Marmoset Genome Sequencing and Analysis Consortium*

We report the whole-genome sequence of the common marmoset (Callithrix jacchus). The 2.26-Gb genome of a female marmoset was assembled using Sanger read data (6x) and a whole-genome shotgun strategy. A first analysis has permitted comparison with the genomes of apes and Old World monkeys and the identification of specific features that might contribute to the unique biology of this diminutive primate, including genetic changes that may influence body size, frequent twinning and chimerism. We observed positive selection in growth hormone/insulin-like growth factor genes (growth pathways), respiratory complex I genes (metabolic pathways), and genes encoding immunobiological factors and proteases (reproductive and immunity pathways). In addition, both protein-coding and microRNA genes related to reproduction exhibited evidence of rapid sequence evolution. This genome sequence for a New World monkey enables increased power for comparative analyses among available primate genomes and facilitates biomedical research application.

Apparently unique among mammals, marmosets routinely produce dizygotic twins that exchange hematopoietic stem cells in utero, a process that leads to lifelong chimerism1–2. As a result of this placental exchange, the blood of adult marmosets normally contains a substantial proportion of leukocytes that are not derived from the inherited germ line of the sampled individual but rather were acquired in utero from its co-twin. In addition, marmosets (subfamily Callitrichinae) and other callitrichines are small in body size as a result of natural selection for miniaturization. This reduced body size might be related to gestation of multiples and to the marmoset social system, also unique among primates3–5. These animals use a cooperative breeding system in which generally only one pair of adults in any social group constitutes active breeders. Other adult group members participate in the care and feeding of infants but do not reproduce. This allopastoral care is rare among anthropoid primates, with the clear exception of humans. The evolutionary appearance of major new groups (for example, superfamilies) of primates has generally been characterized by progressive increases in body size and lifespan, reductions in overall reproductive rate and increases in maternal investment in the rearing of individual offspring. In contrast, marmosets and their callitrichine relatives have undergone a secondary reduction in body size from a larger platyrrhine ancestor6 and have evolved a reproductive and social system in which the dominant male and female monopolize breeding but benefit from allopastoral care provided to their offspring by multiple group members.

Here we report the whole-genome sequencing and assembly of the genome of the marmoset, the first New World monkey to be sequenced (Supplementary Note). Our results include comparisons of this platyrrhine genome with the available catarrhine (human, other hominoid and Old World monkey) genomes, identifying previously undetected aspects of catarrhine genome evolution, including positive selection in specific genes and significant conservation of previously unidentified segments of noncoding DNA. The marmoset genome displays a number of unique features, such as rapid changes in microRNAs (miRNAs) expressed in placenta and nonsynonymous changes in protein-coding genes involved in reproductive physiology, which might be related to the frequent twinning and/or chimerism observed.

WF1KKN1, which encodes a multidomain protease inhibitor that binds growth factors and bone morphogenetic proteins (BMPs)7, has nonsynonymous changes found exclusively in common marmosets and all other tested callitrichine species that twin. In the one callitrichine species that does not produce twins (Callimico goeldi), one change has reverted to the ancestral sequence found in non-twinning primates. GDF9 and BMP15, genes associated with twinning in sheep and humans, also exhibit nonsynonymous changes in callitrichines. We detected positive selection in five growth hormone/insulin-like growth factor (GH–IGF) axis genes with potential roles in diminutive body size and in eight genes in the nuclear-encoded subunits of respiratory complex I that affect metabolic rates and body temperature, adaptations associated with the challenges of a small body size.

Marmosets exhibit a number of unanticipated differences in miRNAs and their targets, including 321 newly identified miRNA loci. Two large clusters of miRNAs expressed in placenta show substantial sequence divergence in comparison to other primates and are potentially involved in marmoset reproductive traits. We identified considerable evolutionary change in the protein-coding genes targeted by the highly conserved let-7 family and notable coevolution of the rapidly evolving chromosome 22 miRNA cluster and the targets of its encoded miRNAs.

The marmoset genome provides unprecedented statistical power to identify sequence constraint among primates, facilitating the
discovery of genomic regions underlying primate phenotypic evolution. The 23,849 regions that exhibit significant sequence constraint among primates but not in non-primate mammals are overwhelmingly noncoding, are disproportionately associated with genes involved in neurodevelopment and retroviral suppression, and frequently overlap transposable elements. For seven genes, we detected positive selection on the branch leading to Catarrhini. Five were newly identified, including genes involved in immunobiology and reproduction (Table 1).

RESULTS

Genome assembly and features

The 2.26-Gb genome of a female marmoset (186/17066) assembled with Sanger read data (6×) and a whole-genome shotgun strategy (Supplementary Fig. 1 and Supplementary Tables 1–4) represents ~90% of the marmoset genome. By all available measures, the chromosomal sequences have high nucleotide and structural accuracy (contig N50 of 29 kb, scaffold N50 of 6.7 Mb; Supplementary Note) and provide a suitable template for initial analysis.

Given the inherent genetic chimerism in this species, blood DNA contained sequences from the germ line of the sampled individual and also from her male co-twin. We took advantage of the sex difference in the co-twins to estimate the proportion of reads originating from the co-twin (Supplementary Fig. 2, Supplementary Tables 5 and 6, and Supplementary Note). These analyses indicated that 10% of the reads in the reference genome data set were derived from the co-twin.

We estimated the amount and size of marmoset segmental duplications using two computational methods, WGAC8 and WSSD3. Assembly-based duplications added a total of 138 Mb of non-redundant sequences (4.7% of the whole genome), slightly less than observed in human or chimpanzee (~5%)-10,12 but more than in orangutan (3.8%)13, where specific collapses in the released assembly version might explain this anomaly (Supplementary Figs. 3 and 4, Supplementary Tables 7–10 and Supplementary Note).

For segmental duplications of >10 kb in length with >94% sequence identity (Supplementary Table 8), we compared the results from the two independent methods to measure artifactual duplications and mistaken assembly collapses. Both methods identified a total of 18 Mb of duplications, of which 26 Mb represented possible artifactual duplications and 53 Mb represented possible collapses. To validate the methods, we tested 97 clones by FISH mapping to marmoset chromosomes (Supplementary Table 9). Both methods successfully identified segmentally duplicated regions, and, unlike in previous studies, WGAC seemed better suited than WSSD to detect duplication in the marmoset. The degree to which this is due to the chimeric nature of the individual sequenced is not clear, although chimerism is certainly a contributing factor.

The overall repeat composition of the marmoset genome was similar to those of other sequenced primate genomes10,12–14, containing ~1.1 million Alu elements, ~660,000 of which were full length. However, in the recent past, Alu retrotransposition appeared to be somewhat slower in marmoset than in human and rhesus macaque (Supplementary Note).

Constrained sequence evolution indicates natural selection and therefore implies conserved function. By extension, lineage-specific constraint indicates lineage-specific function15,16. Using the marmoset genome, we detected 23,849 elements constrained in anthropoid primates but not in non-primate mammals17 (Supplementary Note). These anthropoid-specific constrained (ASC) sequences potentially drove primate phenotypic evolution and are abundant in noncoding regions (for example, upstream of SNTG1), although coding exons are also represented (for example, in PGBD3) (Supplementary Fig. 5a,b).

Annotated transposable elements contributed 46% of ASC base pairs. We validated the enhancer activity of six elements (of eight tested) in human embryonic stem cells (ESCs) (Supplementary Fig. 5c,d and Supplementary Table 11) and showed that their mouse orthologs had little or no functional activity. This data set highlights specific loci that acquired new functional roles in the primate lineage and suggests molecular mechanisms underlying unique primate traits.

Gene content and gene families

The Ensembl gene set18 (Supplementary Fig. 6 and Supplementary Note) of 21,168 genes (44,973 transcripts) included 219 genes with marmoset protein support and 15,706 genes without marmoset protein evidence but with human protein evidence. The remaining 5,243 genes had transcripts supported by protein data from other sources (Supplementary Fig. 6g,h).

A phylogenetic framework including 4 other primates, 2 rodents and 3 Laurasiatheria showed 429 primate-specific gene families, among which few were present only in marmoset (Supplementary Fig. 7, Supplementary Tables 12–19 and Supplementary Note). More than half of these families (221/429) were indeed absent in marmoset, suggesting that they emerged after catarrhine-platyrrhine divergence. In addition, many families were absent in rhesus macaque, and thus almost half were apparently unique to apes.

Table 1 Gene Ontology (GO) categories enriched for genes positively selected in marmoset

<table>
<thead>
<tr>
<th>GO categorya</th>
<th>Description</th>
<th>Genes</th>
<th>PSGs</th>
<th>Total</th>
<th>Excess</th>
<th>P value (MWU)</th>
<th>Adjusted P value (Holm)</th>
<th>P value (FET)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0005576</td>
<td>Extracellular region</td>
<td>150</td>
<td>1,954</td>
<td>1.3</td>
<td>3.24 × 10⁻¹⁵</td>
<td>9.80 × 10⁻¹²</td>
<td>3.86 × 10⁻¹⁷</td>
<td></td>
</tr>
<tr>
<td>0005615</td>
<td>Extracellular space</td>
<td>63</td>
<td>429</td>
<td>2.4</td>
<td>2.52 × 10⁻⁸</td>
<td>7.61 × 10⁻⁵</td>
<td>1.31 × 10⁻⁸</td>
<td></td>
</tr>
<tr>
<td>0005747</td>
<td>Mitochondrial respiratory chain complex I</td>
<td>8</td>
<td>14</td>
<td>9.4</td>
<td>1.81 × 10⁻⁷</td>
<td>5.47 × 10⁻⁴</td>
<td>2.72 × 10⁻⁵</td>
<td></td>
</tr>
<tr>
<td>0006952</td>
<td>Defense response</td>
<td>54</td>
<td>324</td>
<td>2.7</td>
<td>2.19 × 10⁻⁶</td>
<td>6.59 × 10⁻³</td>
<td>3.38 × 10⁻⁹</td>
<td></td>
</tr>
<tr>
<td>0004872</td>
<td>Receptor activity</td>
<td>103</td>
<td>866</td>
<td>2.0</td>
<td>3.42 × 10⁻⁶</td>
<td>1.03 × 10⁻²</td>
<td>1.05 × 10⁻⁸</td>
<td></td>
</tr>
<tr>
<td>0007606</td>
<td>Sensory perception of chemical stimulus</td>
<td>20</td>
<td>136</td>
<td>2.4</td>
<td>5.82 × 10⁻⁶</td>
<td>1.75 × 10⁻²</td>
<td>1.26 × 10⁻³</td>
<td></td>
</tr>
<tr>
<td>0030246</td>
<td>Carbohydrate binding</td>
<td>29</td>
<td>203</td>
<td>2.3</td>
<td>6.81 × 10⁻⁶</td>
<td>2.05 × 10⁻²</td>
<td>1.78 × 10⁻⁴</td>
<td></td>
</tr>
<tr>
<td>0006954</td>
<td>Inflammatory response</td>
<td>36</td>
<td>181</td>
<td>3.3</td>
<td>8.39 × 10⁻⁶</td>
<td>2.52 × 10⁻²</td>
<td>3.31 × 10⁻⁸</td>
<td></td>
</tr>
<tr>
<td>0004984</td>
<td>Olfactory receptor activity</td>
<td>16</td>
<td>107</td>
<td>2.5</td>
<td>9.88 × 10⁻⁶</td>
<td>2.97 × 10⁻²</td>
<td>3.21 × 10⁻³</td>
<td></td>
</tr>
<tr>
<td>0009611</td>
<td>Response to wounding</td>
<td>53</td>
<td>332</td>
<td>2.6</td>
<td>2.93 × 10⁻⁵</td>
<td>8.79 × 10⁻²</td>
<td>1.73 × 10⁻⁸</td>
<td></td>
</tr>
<tr>
<td>0006955</td>
<td>Immune response</td>
<td>41</td>
<td>295</td>
<td>2.3</td>
<td>3.18 × 10⁻⁵</td>
<td>9.53 × 10⁻²</td>
<td>1.57 × 10⁻⁵</td>
<td></td>
</tr>
</tbody>
</table>

*GO category number. *Positively selected genes (PSGs) identified with a threshold of P < 0.05. *Total number of genes in the GO category. *Fold enrichment in positively selected genes over background.

Enriched GO categories were identified by Mann-Whitney U test (MWU), nominal P value adjusted for multiple testing by Holm correction (Holm) and Fisher's exact test (FET) using all genes with nominal P < 0.05 in the marmoset lineage likelihood ratio test. Note that the results of the Mann-Whitney U test may also be affected by the relaxation of constraint, whereas Fisher's exact test considers only genes identified as being under positive selection.
Our comparative analysis found surprising changes in the miRNA repertoire and the mRNA targets that they regulate. We identified 777 mature miRNAs (mapping to 1,165 hairpin precursor miRNAs) (Supplementary Tables 20–37). Most were confirmed through expression studies (582; 75%) (Supplementary Note) and were conserved in primates (~55–58%). Many (321 miRNAs mapping to 477 hairpins) were novel (not found in any other species analyzed). These could include miRNAs exclusive to marmoset, miRNAs exclusive to Platyrrhini and conserved miRNAs that are yet to be discovered in other species. The two largest marmoset miRNA clusters (on chromosome 22 and the X chromosome) were expanded in number compared to in humans (112 marmoset versus 49 human chromosome 22 hairpins and 40 marmoset versus 15 human X-chromosome hairpins) (Supplementary Table 22) and showed divergent sequence. Less than 3% of the chromosome 22 and 8% of the X-chromosome miRNAs were conserved across primates (Supplementary Table 22), and most exhibited at least one nucleotide modification in the 5′ seed region (83% of chromosome 22 miRNAs and 78% of X-chromosome miRNAs) compared to their human counterparts (Supplementary Tables 20, 22, 23 and 29). The rapidly evolving chromosome 22 and X-chromosome clusters dominated miRNA expression in marmoset placenta, whereas marmoset brain exhibited a more diverse miRNA expression pattern (Supplementary Fig. 8 and Supplementary Tables 30–32). In contrast, some miRNA families (for example, let-7) were conserved across primates (pins and 40 marmoset versus 15 human X-chromosome hairpins) to in humans (112 marmoset versus 49 human chromosome 22 hairpins and 40 marmoset versus 15 human X-chromosome hairpins) (Supplementary Table 22) and showed divergent sequence. Less than 3% of the chromosome 22 and 8% of the X-chromosome miRNAs were conserved across primates (Supplementary Table 22), and most exhibited at least one nucleotide modification in the 5′ seed region (83% of chromosome 22 miRNAs and 78% of X-chromosome miRNAs) compared to their human counterparts (Supplementary Tables 20, 22, 23 and 29). The rapidly evolving chromosome 22 and X-chromosome clusters dominated miRNA expression in marmoset placenta, whereas marmoset brain exhibited a more diverse miRNA expression pattern (Supplementary Fig. 8 and Supplementary Tables 30–32). In contrast, some miRNA families (for example, let-7) were completely conserved in all five primates (Supplementary Fig. 9).

Changes in the miRNA seed region are expected to correspond with changes in the genes they regulate, unless the miRNAs and their mRNA targets have coevolved. Comparing the annotated genes containing predicted let-7 target sequences (Fig. 1 and Supplementary Note), we found 165 common to human and marmoset, 44 unique to marmoset and 64 unique to human. Despite caveats related to differences in assembly and annotation quality, it is striking that less than half of the targets for this highly conserved family were shared by marmoset and human (Supplementary Table 34), a number similar to that in non-eutherian primates (dog, horse and cow). A phylogenetic analysis of these changes showed that let-7 targets have evolved rapidly in primates in comparison to other species (Fig. 2). The pattern of miRNA-mRNA target evolution differed among the three described miRNA families and even between the two rapidly evolving families (Supplementary Tables 33–37). In the X-chromosome cluster, as expected, fewer than 50% of the target sequences were shared by marmoset and human (Supplementary Table 35). In contrast, in the chromosome 22 cluster, 84% of the targets were shared (Supplementary Table 36), implying considerable coevolution of miRNAs and their targets in the chromosome 22 cluster but not in the X-chromosome cluster.

Small marmosets are believed to have evolved from a larger ancestor; we therefore looked for positively selected genes that might explain the change in size. We identified 37 positively selected genes on the marmoset lineage and 7 on the branch to Catarrhini (false discovery rate (FDR) < 0.01) (Supplementary Table 38). Five of these seven genes (SAMHD1, CLEC4A, ANKZF1, KRT8 and CATSPERG) were previously unrecognized as being positively selected19. An additional 91 positively selected genes could not be traced to a particular branch owing to a lack of identifiable outgroup orthologs. Following trends observed in previous studies19, Gene Ontology (GO) categories related to immunity, physiological defense response and sensory perception were enriched (Table 1). In addition, the ATP synthesis and transport and NADH dehydrogenase activity categories showed enrichment (Mann-Whitney U test, P < 0.05). The latter group contained eight positively selected nuclear genes encoding subunits of respiratory complex I. Resulting differences in complex I regulatory and kinetic properties could affect metabolic rates and body temperature, challenges posed by small body size.

A prominent example of positive selection in the marmoset lineage could be found in IGF1R (P = 0.0014), which is associated with short stature in humans20,21. The encoded protein had multiple alterations in crucial binding domains (Fig. 3), which likely affect ligand-receptor binding affinity. Other growth hormone–related positively selected genes possibly related to small stature include GHSR (encoding growth hormone secretagogue receptor), IGF2 (encoding insulin-like growth factor 2), IGFBP2 (encoding insulin-like growth factor binding protein 2), IGFBP7 (encoding insulin-like growth factor binding protein 7) and EGF (encoding epidermal growth factor) (marmoset lineage, P < 0.05). Targeted exon sequencing of multiple species identified several callitrichid–specific nonsynonymous substitutions in genes that were strong candidates for influencing
Figure 3 Residues under positive selection in IGF1R. The insulin-like growth factor 1 receptor (IGF1R) interacts with other proteins in growth hormone pathways and has a role in both prenatal (left) and postnatal (right) growth. Proteins encoded by genes in these pathways in marmoset that have residues under positive selection are tallied; the number of changes that can be assigned to either the marmoset or callitrichine New World monkey (NWM) lineages is also shown. In the middle, the first three domains of the IGF1R α chain are shown, with positively selected residues in red (Bayes empirical Bayes analysis posterior probability (PP) > 0.95) and yellow (PP > 0.5). Leucine-rich repeat domains L1 and L2 are shown in green with L1 on top, and the cysteine-rich region CR is shown in blue. An alignment of the IGF1R proteins from several mammalian species (bottom) identifies several marmoset changes in a short region corresponding to the part of the structure enclosed in the black rectangle.

diminutive body size (GDF9, BMP15 and BMP4). Analysis of these mutations by SIFT22 and PolyPhen23 indicated that these alterations likely affect the function of the corresponding proteins24 (Supplementary Table 38 and Supplementary Note).

The genetic basis of twinning has received substantial attention in humans and other animals25–27. Genetic differences drive variation in ovulation number among sheep strains25,28. There is also clear evidence for genetic influence on human twinning, but the specific genes involved have not been identified. We studied 63 candidate genes previously implicated in the control of either body size, number of ova produced in a single estrous cycle or both. Of these, 41 genes with putative marmoset-specific nonsynonymous variants were examined further (Supplementary Tables 39 and 40). Three genes with a role in ovulation (BMP4, FSTL4 and WFIKKN1) encoded likely function-altering amino acid changes as scored by both SIFT22 and PolyPhen23 (Supplementary Note and ref. 24). Potentially functional nonsynonymous substitutions in the FSHR (follicle-stimulating hormone receptor), BMP10, BMP15, GDF9 and GDF15 genes were also found. Notably, a single nonsynonymous substitution in WFIKKN1 was common to all callitrichids we tested, with the exception of C. goeldii (Fig. 4). That species had a reversal of the change to the sequence found in Old World monkeys and other non-twinning New World monkeys. C. goeldii is the only callitrichid that does not regularly twin, and, given its phylogenetic position, it is highly likely to have reverted to singleton births from an ancestral state that exhibited twinning.

Hematopoietic chimerism of marmosets was expected to correlate with marked changes in immune system function. We found positively selected genes related to the immune response significantly enriched in marmoset (threshold of P < 0.05; Table 1). NAIP and NLR4 homologs, conserved in mammals, were absent in marmoset (Supplementary Table 38). These proteins form the NAIP inflammasome in macrophages, a cytoplasmic complex that triggers macrophage inflammatory death through activation of caspase-1 (refs. 29,30) and could affect reproduction, as human NAIP is expressed in the placenta.

Other positively selected genes potentially involved in circumventing unwanted chimerism-associated responses included CD48, encoding a ligand for CD244 (2B4), which is found on the surface of hematopoietic cells and regulates natural killer cells21 and the levels of interleukins IL-5 and IL-12B, involved in T cell development and in allergic responses32. Finally, in contrast to the extensive family of KIR genes that are integral to immune system function in humans and other catarrhine primates, the marmoset genome contained only two KIR genes, one of which was partial.

Most differences in protease gene families observed between marmoset and other primates occurred in genes related to the reproductive and immune systems (Supplementary Note). For example, ADAM6, with a role in fertility33,34, was lost in marmoset, whereas ISP2, involved in embryo implantation35, has been duplicated twice. KLK2/3, duplicated in the catarrhine ancestor36 and involved in reproductive physiology33, is non-functional in marmoset. Chymase and tryptase protease changes and CMA1 and MAST duplications potentially affect the immune response37,38 and mast cell biology, respectively. The duplicated CMA1 gene might be related to the murine-specific mast cell proteases (MCPS) that are absent in hominoids39. Changes in the C terminus of MMP19, an IGFBP3-processing enzyme40, might be related to growth characteristics. Consistent with retrogene analysis (Supplementary Note), there were multiple non-functional single-exon protease-like pseudogenes. Seven of these had complete ORFs without identified transcripts, indicating that they arose from recent retrotranscription events.

PRDM9, which encodes a protein that binds DNA in recombination hot spots and affects recombination activity during meiosis41 (Supplementary Fig. 10 and Supplementary Note), was duplicated in catarrhine primates. Orthologs encoding all three functional PRDM9 domains have been computationally identified in placental mammals42; however, these genes are often not in syntenic locations. In primates (including in human and marmoset), panda, pig and elephant, there is a PRDM9-like gene flanked by a conserved syntenic block including the genes URAH and GAS8. This gene, located near the 16q telomere in human, is labeled PRDM7 in catarrhine primates but PRDM9 in marmoset and non-primates. Another gene (labeled PRDM9 in catarrhine primates) is located between the cadherin genes CDH12 and CDH10 at human 5p14 (ref. 43). This gene is present in chimpanzee, orangutan and rhesus macaque but is absent in marmoset and non-primates. The marmoset genome sequence provided
two types of evidence that support the occurrence of a duplication in the catarrhine lineage after its divergence from platyrrhine primates: the phylogeny of *PRDM9*-like genes (Supplementary Fig. 10b) and their genomic locations.

### Population genetics and polymorphism

Genome sequence diversity was examined in nine marmosets (two from the New England Regional Primate Research Center (RPRC), two from the Wisconsin National Primate Research Center (NPRC) and five from the Southwest NPRC) (Supplementary Fig. 11). This sample size is sufficient to identify common polymorphisms in this species but will not be sufficient to detect a large proportion of low-frequency or rare variants. Chimerism does not interfere with the identification of SNPs that are polymorphic in the species as a whole but does complicate the assignment of genotypes for specific SNPs to specific individuals. We investigated this effect by quantifying read balance (the proportion of reads supporting each allele in apparent heterozygotes) and found different distributions in marmosets in comparison to a human control: more SNPs with read balance fractions between 5% and 25% were observed in marmosets. Simulations indicated that this flattened read balance distribution resulted from bases that were not polymorphic in the sampled individual but were either heterozygous or differently homozygous in the co-twin, with the low level of alternative reads representing the chimeric cells introduced during development (Supplementary Fig. 2a and Supplementary Note).

We also explicitly modeled the expected number of sequencing reads covering a dimorphic SNP locus with one allele or the other, given a known fraction of chimerism. Chimerism fractions ranged from 12% to 37% (Supplementary Table 6 and Supplementary Note).

Using polymorphic autosomal biallelic SNPs (~7.7 million), we calculated pairwise allele-sharing genetic distances. To test whether the genetic variation among individuals could be explained by their primate colony of origin, we performed principal-component analysis (PCA) based on pairwise distance. PCA separated the three colonies showing that the genetic variation among individuals could be explained by their primate colony of origin, we performed principal-component analysis (PCA) based on pairwise distance. PCA separated the three colonies into two different groups (Supplementary Fig. 12). The Structure analysis suggests that this species but will not be sufficient to detect a large proportion of low-frequency or rare variants. Chimerism does not interfere with the identification of SNPs that are polymorphic in the species as a whole but does complicate the assignment of genotypes for specific SNPs to specific individuals. We investigated this effect by quantifying read balance (the proportion of reads supporting each allele in apparent heterozygotes) and found different distributions in marmosets in comparison to a human control: more SNPs with read balance fractions between 5% and 25% were observed in marmosets. Simulations indicated that this flattened read balance distribution resulted from bases that were not polymorphic in the sampled individual but were either heterozygous or differently homozygous in the catarrhine lineage after its divergence from platyrrhine primates: the phylogeny of *PRDM9*-like genes (Supplementary Fig. 10b) and their genomic locations.

### DISCUSSION

Previous analyses of primate genomes have identified few specific changes that account for phenotypic differences among species, with the exception of genes that influence human brain size,27, language (reviewed in ref. 48) or other uniquely human traits.49 In contrast, our analysis presents a number of specific differences in gene content, miRNA number and sequence, and protein-coding gene sequences in genes known to influence growth, reproduction and twinning propensity, all potentially related to marmoset phenotypic adaptations. Such divergence at multiple levels does indeed underscore the remarkable nature of this platyrrhine monkey species.

Acknowledgments


The marmoset genome project was funded by the National Human Genome Research Institute (NHGRI), including from grants U54 HG003272 (R.A. Gibbs) and U54 HG003079 (R.K.W.), with additional support from the US National Institutes of Health (NIH), including from grants R01 DK077639 (S.D.T.), R01 GM59290 (L.B.J. and M.A.R.), HG002385 (E.E.) and PS1-001113 (Southwest NPRC), and support from the National Science Foundation (NSF BCS-0751808 to D.E.W.) and the VEGA grant agency: 1/0719/14 (TV) and 1/1085/12 (B.B.). C.C.F. and M.C.R. were supported in part by a Howard Hughes Medical Institute grant to Louisiana State University through the Undergraduate Biological Sciences Education program. I.X. was supported by NHGRI grant K9 HG005846. P.H.G. was supported by the Cullen Foundation. T.M.-B. was supported by European Research Council Starting Grant (260372) and MICINN (Spain) grant BFU2011-28549. B.L.-G. was supported by the Spanish National Institute of Bioinformatics (see URLs). E.E.E. is an investigator of the Howard Hughes Medical Institute.

Author Contributions


Competing Financial Interests

The authors declare no competing financial interests.

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

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-sa/3.0/.

Saba Sajjadi\textsuperscript{15}, Jirch Santibanez\textsuperscript{1,2}, Daniel R Schrider\textsuperscript{12,13}, Steve Searle\textsuperscript{16}, Helen Skaletsky\textsuperscript{22,31}, Benjamin Soibam\textsuperscript{18}, Arian F A Smit\textsuperscript{21}, Janyan B Tennakoon\textsuperscript{18}, Lubomir Tomaska\textsuperscript{32}, Brygg Ullmer\textsuperscript{33,34}, Charles E Vejar\textsuperscript{17}, Mario Ventura\textsuperscript{15}, Albert J Vilella\textsuperscript{9}, Tomas Vinar\textsuperscript{10}, Jan-Hinerner Vogel\textsuperscript{16}, Jerilyn A Walker\textsuperscript{8}, Qing Wang\textsuperscript{8}, Crystal M Warner\textsuperscript{1,2}, Derek E Wildman\textsuperscript{35}, David J Witherspoon\textsuperscript{23}, Rita A Wright\textsuperscript{1,2}, Yuanqing Wu\textsuperscript{1,2}, Weimin Xiao\textsuperscript{18}, Jinchuan Xing\textsuperscript{23,38}, Evgeny M Zdobnov\textsuperscript{17,36,37}, Baoli Zhu\textsuperscript{14,38}, Richard A Gibbs\textsuperscript{1,2} & Richard K Wilson\textsuperscript{3}

\textsuperscript{1}Human Genome Sequencing Center, Baylor College of Medicine, Houston, Texas, USA. \textsuperscript{2}Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas, USA. \textsuperscript{3}The Genome Institute, Washington University St. Louis, St. Louis, Missouri, USA. \textsuperscript{4}Barshop Institute for Longevity and Aging Studies, University of Texas Health Science Center at San Antonio, San Antonio, Texas, USA. \textsuperscript{5}Department of Obstetrics and Gynecology, Baylor College of Medicine, Houston, Texas, USA. \textsuperscript{6}Department of Biology, Bari University, Bari, Italy. \textsuperscript{7}Genome Institute of Singapore, Singapore. \textsuperscript{8}Department of Biological Sciences, Louisiana State University, Baton Rouge, Louisiana, USA. \textsuperscript{9}European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, UK. \textsuperscript{10}Faculty of Mathematics, Physics and Informatics, Comenius University in Bratislava, Bratislava, Slovakia. \textsuperscript{11}Wisconsin National Primate Research Center, Madison, Wisconsin, USA. \textsuperscript{12}Department of Biology, Indiana University, Bloomington, Indiana, USA. \textsuperscript{13}School of Informatics and Computing, Indiana University, Bloomington, Indiana, USA. \textsuperscript{14}Children's Hospital Oakland Research Institute, Oakland, California, USA. \textsuperscript{15}Department of Genome Sciences, University of Washington, Seattle, Washington, USA. \textsuperscript{16}Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, UK. \textsuperscript{17}Genetic Medicine and Development, University of Geneva Medical School, Geneva, Switzerland. \textsuperscript{18}Department of Biology and Biochemistry, University of Houston, Houston, Texas, USA. \textsuperscript{19}Department of Pathology, Baylor College of Medicine, Houston, Texas, USA. \textsuperscript{20}Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, Massachusetts, USA. \textsuperscript{21}Institute for Systems Biology, Seattle, Washington, USA. \textsuperscript{22}Whitehead Institute, Cambridge, Massachusetts, USA. \textsuperscript{23}Department of Human Genetics, University of Utah School of Medicine, Salt Lake City, Utah, USA. \textsuperscript{24}Institut für Populationsgenetik, Vetmeduni Vienna, Vienna, Austria. \textsuperscript{25}Department de Bioquímica y Biología Molecular, Institut Universitari de Recerca i Estudis Avançats, Institut de Biologia Evolutiva Consejo Superior de Investigaciones Científicas–Universitat Pompeu Fabra, Barcelona, Spain. \textsuperscript{26}Division of Comparative Pathology, New England Primate Research Center, Harvard University Medical School, Southborough, Massachusetts, USA. \textsuperscript{27}CNAG (Centro Nacional de Análisis Genómico), Barcelona, Spain. \textsuperscript{28}Center for Biomedical Science and Engineering, School of Engineering, University of California, Santa Cruz, Santa Cruz, California, USA. \textsuperscript{29}Howard Hughes Medical Institute, Whitehead Institute, Cambridge, Massachusetts, USA. \textsuperscript{30}Department of Genetics, Faculty of Natural Sciences, Comenius University in Bratislava, Bratislava, Slovakia. \textsuperscript{31}Center for Computation and Technology, Louisiana State University, Baton Rouge, Louisiana, USA. \textsuperscript{32}Department of Computer Sciences, Louisiana State University, Baton Rouge, Louisiana, USA. \textsuperscript{33}Center for Molecular Medicine and Genetics, Wayne State University School of Medicine, Detroit, Michigan, USA. \textsuperscript{34}Swiss Institute of Bioinformatics, Geneva, Switzerland. \textsuperscript{35}Division of Molecular Biosciences, Imperial College London, London, UK. \textsuperscript{36}Present addresses: The Jackson Laboratory for Genomic Medicine, Farmington, Connecticut, USA (G.M.W.), Department of Bioengineering, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA (P.J.d.J.), Department of Biology, Saint Louis University, St. Louis, Missouri, USA (C.C.), Boehringer Ingelheim, Vienna, Austria (D.G.), Bill Lyons Informatics Centre, University College London Cancer Institute, University College London, London, UK (J. Herrero). Translational Sciences, Novartis Institutes for Biomedical Research, Cambridge, Massachusetts, USA (K.G.M.), Department of Genetics, Human Genetics Institute of New Jersey, Rutgers, State University of New Jersey, Piscataway, New Jersey, USA (J.X.), and Chinese Academy of Sciences Key Laboratory of Pathogenic Microbiology and Immunology, Institute of Microbiology, Chinese Academy of Sciences, Beijing, China (B.Z.). Correspondence should be addressed to K.C.W. (kworley@bcm.edu).
ONLINE METHODS

Additional information describing New World monkey phylogeny, genome sequencing, assembly and quality assessment, chromerism assessment, analysis of segmental duplications, sequence constraint, gene annotation, orthologs and sequence variation is available in the Supplementary Note.

Genome sequencing and assembly. The 26.7 million sequence reads, generated on ABI3730 instruments (Supplementary Table 1) with an average read length of 700 bases (Phred51 quality of ≥20), were assembled using PCAP25. The assembly was filtered to remove known non-marmoset sequence contaminants, and singleton contigs and supercontigs <2 kb in length. The final assembly included 99.98% of the input reads and had 59% AT content. WUGCCallithrix jacchus-3.2 was submitted to GenBank (UCSC version callJac3) and used by Ensembl to build gene models. Statistics (Supplementary Table 2) are for the initial assembly, before integrating in finished BACs and adding interscaffold gaps and gaps representing centromeres and telomeres. The final assembly spans 2.91 Gb, with 2.77 Gb ordered and oriented along specific chromosomes. The assembly represents an arbitrary consensus of the individual marmoset’s alleles.

Non-repetitive assembly data were aligned against the repeat-masked human genome at UCSC using BLASTZ23. Orthologous and paralogous alignments26 were differentiated, and only ‘reciprocal best’ alignments were retained and used to generate the marmoset AGP files, as in previously described methods12. Documented inversions based on FISH data (see URLs) and inversions suggested by the assembly and supported by additional mapping data (for example, fosmid and BAC end pairs) were also introduced. Centromeres were placed on the basis of their positions identified from cytogenetic data (Supplementary Note). A total of 81 finished CHORI-259 marmoset BACs (totaling 15,576,643 bases) were merged into the final chromosomal files.

Marmoset cDNAs (Supplementary Table 4) generated at the Genome Institute at Washington University with Roche 454 Life Sciences instruments were placed on the basis of their positions identified from cytogenetic data (for example, fosmid and BAC end pairs) were also introduced. Centromeres were placed on the basis of their positions identified from cytogenetic data (Supplementary Note). A total of 81 finished CHORI-259 marmoset BACs (totaling 15,576,643 bases) were merged into the final chromosomal files.

Gene feature annotation. Annotations with RefSeq57 and Ensembl18,58 used the general methods described (see URLs). The raw compute stage of Ensembl annotation (Supplementary Fig. 6a) screened genomic sequence using RepeatMasker29 (version 3.2.9; parameters ‘-nolow -species homo -s’) and Dust (J. Kuzio, R. Tatusov and D.J. Lipman, personal communication, briefly described in ref. 60) (together masking 47%) and TRF61.

Predicted features included transcription start sites (Eponine-scand62 and FirstEF63), CpG islands (described in ref. 64) and tRNAs65. Genscan results on repeat-masked sequence were input for UniProt66, UniGene67 and Vertebrate RNA (see URLs) by WU-BLAST68,69 alignments, resulting in 252,582 UniProt, 316,384 UniGene and 317,679 Vertebrate RNA sequence alignments.

Genewise70 and Exonerate71 produced coding sequence models using marmoset and human UniProt, SwissProt, TrEMBL (see URLs) and RefSeq72 proteins mapped to the genome (Pmatch; R. Durbin, unpublished data) (Supplementary Fig. 6b,c). One model per locus was selected using the BestTargeted module. Species-specific data (here, for marmoset and human) generated 1,908 (of 3,153) marmoset protein and 20,735 (of 22,320) human protein ‘targeted stage’ models with UTRs.

Raw compute UniProt alignments were filtered, sequences with UniProt Protein Existence (PE) classifications of level 1 or 2 were mapped with WU-BLAST, and coding models were built with Genewise in regions outside of targeted stage models, generating an additional 57,019 mammalian and 42,323 non-mammalian ‘similarity stage’ models.

Marmoset cDNAs and ESTs and human cDNAs from the European Nucleotide Archive (ENA), GenBank and the DNA Data Bank of Japan (DDBJ) with their polyA tails removed were aligned to the genome using Exonerate72 (Supplementary Fig. 6d-f). With cutoffs of 90% coverage and 80% identity, 139,713 (of 292,329) human cDNAs, 887 (of 986) marmoset cDNAs and 2,562 (of 2,605) marmoset ESTs aligned. EST-based gene models (similar to those for humans73) are displayed in a separate website track from the Ensembl gene set. Similarity stage coding models were filtered to remove models with little cDNA or EST support, visualized using Apollo74 and extended using human cDNA and marmoset expressed sequences, resulting in 1,501 (of 2,119) marmoset, 13,150 (of 20,735) human and 22,897 (of 31,863) UniProt coding models with UTRs. Redundant transcript models were removed, and remaining models were clustered wherever any coding exons from two transcripts overlapped. More information on the Ensembl automatic gene annotation process75,20 is available in the references and the Supplementary Note.

Segmental duplications. Segmental duplications in Callithrix jacchus-3.2 were estimated using two computational methods; one compares assembly segments using BLAST (Whole-Genome Assembly Comparison, WGAC)46, and the second assessed excess depth of coverage of whole-genome sequencing data mapped to the assembly (WSSD)53. All scaffolds were repeat masked (RepeatMasker; see URLs) and window masked25 using the specific marmoset repeat library (Supplementary Note) composed of retrotransposons and other low-complexity sequences. WGAC identifies pairwise alignments of >1 kb in length and >90% identity; WSSD identifies segmental duplications of >10 kb in length and >94% identity. For WSSD, we mapped reads using Megablast with >94% sequence identity, >200 bp non-repeat-masked sequence length and at least 200 bp of Phred Q of >30 bp.

We assessed 97 clones using FISH on lymphoblast cell line nuclei and metaphase chromosomes from a marmoset unrelated to individual 186/17066. Duplicated probes had >2 signals in 95–98% of >60 observed nuclei (Supplementary Fig. 3c). Sixteen clones showing strong hybridization background were tested three times without a clear pattern emerging and were removed from further analysis. This unusual background might be due to incomplete masking by RepeatMasker and/or competitive hybridization conditions during FISH. Nine (of 16) of these clones belonged to the category that were absent in WGAC and present in WSSD, consistent with them corresponding to collapsed repeats.

As in the assessment of ape genomes76, we aligned 27,615,086 marmoset reads to the human genome (Build 35; excluding random sequences) with repeat content masked (<20% divergent from the consensus; RepeatMasker in either human or marmoset). Aligned reads had >200 bp of high-quality sequence (Phred score >27), >300 bp of aligned sequence, >40% read length aligned and >200 bp repeat content. After evaluation, we applied an identity threshold of 85%, similar to the criteria applied in the macaque analysis. See the Supplementary Note for details.

Sequence elements constrained in anthropoid primates. ASCs were defined using the pipeline briefly outlined in the Supplementary Note and described in detail in ref. 17. To validate the functional role of the bioinformatically defined elements as transcriptional enhancers, we tested eight noncoding ASCs in ESC enhancer assays. Candidates were selected on the basis of DNaese1 hypersensitivity in human ESCs77. The eight human sequences and their mouse orthologs (identified using liftOver; Supplementary Table 11) were amplified from their respective genomic DNA, cloned into the Sall site downstream of luciferase in the pGL3-Pout51 vector using the Gateway Cloning System (Invitrogen) and transfected with the reporter constructs into human ESCs (H1-VA-01, WiCell Research Institute) and mouse ES (E14TG2A, American Type Culture Collection, CRL-1821) using FuGENE HD (Roche) or Lipofectamine 2000 (Invitrogen), respectively. Both cell lines are routinely tested for mycoplasma contamination (Lonza Detection kit, LT07-318). A Renilla luciferase plasmid (pRL-SV40, Promega) was cotransfected into cells as an internal control. Cells were collected 48 h after transfection, and the luciferase activities of the cell lysates were measured using the Stop-Glow Dual-Luciferase Reporter Assay System (Promega) (Supplementary Note).

MicroRNAs. MiRNAs (877; Supplementary Table 2) were identified as being expressed or predicted on the basis of cross-species conservation of mature miRNA or hairpin sequences. Small RNAs were sequenced from total RNA isolated from prefrontal cortex brain samples (A07-716monkB, 3.2 years, male; A09-122monkB, 12.8 years, female; A08-206monkB, 13.4 years, male; A08-337monkB, 13.0 years, female) and two placenta samples, using 36-bp reads to the human genome (Build 35; excluding random sequences) with repeat sequence (Phred score >30 bp) masked, >40% read length aligned and >200 bp repeat content. After evaluation, we applied an identity threshold of 85%, similar to the criteria applied in the macaque analysis. See the Supplementary Note for details.

© 2014 Nature America, Inc. All rights reserved.
Identification of one-to-one orthologs. Conservative one-to-one orthologs for marmoset and human, chimpanzee, rhesus macaque, orangutan, mouse, rat and dog were identified using UCSC76 whole-genome alignments and genes (July 2010), including partial transcripts missing 10% of the sequence on both ends. Transcripts on chromosomes of >100 nucleotides in length in RefSeq (58,126), knownGene (116,345), Ensembl (128,193) and VEGA (73,873) clustered into 21,694 genes on the basis of location.

Each transcript was transferred to other species and subjected to testing designed to exclude genes that have undergone large-scale changes other than point mutations (as in ref. 19) and testing for breaks in synteny, significant assembly gaps overlapping the transcript, frameshift and nonsense mutations, conservation of gene structure elements (splice sites, start codons and stop codons) and recent duplications causing misassignment of one-to-one orthology. Clean transcripts passed all tests. We chose a representative clean transcript for each locus, preferring longer transcripts that were clean in more species (summarized in Supplementary Table 12). This conservative set (13,717 one-to-one orthologs for human and marmoset) included 41% covering all 8 species, 27% missing in 1 species, 15% missing in 2 species, 10% missing in 3 species and less than 7% missing in more than 3 species.

Gene family evolution. Gene family evolution was investigated in four other primates, two rodents and three Laurasiatheria with fully sequenced genomes (human, chimpanzee, orangutan, rhesus macaque, marmoset, mouse, rat, dog, horse and cow). Gene families, including gene and protein names and genome coordinates, were retrieved from Ensembl gene trees, version 58 (see URLs). Genes with multiple short introns (<50 bp) or short coding regions (<100 bp) that were present in <3 species were removed, and we analyzed separately families with genes in only one lineage (Euarchoonta, Glires and Laurasiatheria). The final set included most genes and families from the original Ensembl annotations (Supplementary Table 13) and was used to infer ancestral family size with maximum-likelihood CAFFE91 analysis using the following ultrametric tree built according to ref. 92: ((((((chimp;human:6:7); orang:13:11; macaca:24:16; marmoset:40:47; (mouse:17,rat:17):70:16, (dog:74,horse:74):9);cow:83:10), where numbers correspond to millions of years (Supplementary Note).

Positively selected genes. Positively selected genes among the one-to-one orthologs were identified using Markov models of codon evolution and maximum-likelihood methods similar to PAML92. Further downstream analysis such as enrichment analysis for GO categories was performed as described92. The Supplementary Note details the genes identified using FDR < 0.01.

Genes involved in growth pathways and twinning. Candidate genes identified using 33-way EPO alignments18 containing marmoset nonsynonymous substitutions (compared to human) conserved in haplorhine primates (human, chimpanzee, gorilla, orangutan, rhesus macaque and tarsier) were sequenced. The NS effect was defined using SIFT84, and some candidates were omitted owing to conflicting evidence. Genes and coordinates are listed in Supplementary Table 39. The species used for alignment included Saguinus bicolor marinus*, Saguinus imperator imperator, Saguinus midas niger*, Saguinus fuscicolis weddelli, Callithrix cebula pygmaea*, Leontopithecus rosalia*, Cebus apella, Callimico goeldii, Atelus belzebuth and Saimiri sciureus* (species with an asterisk were also selected for mRNA sequencing). Sanger sequencing reads were assembled (Velvet95), mapped to the genome (BLAT51) and aligned (MAFFT82). In 49 of the 82 exons sequenced, data were insufficient to determine whether the marmoset nonsynonymous substitutions were callitchrine or New World monkey specific (Supplementary Note).

Protease genes. We mined the marmoset genome for protease genes (see URLs) using BATI (Blast, Annotate, Tune, Iterate). Curated human proteases were compared to the marmoset genome with the TBlastN algorithm using the tblx script, and the locations of marmoset protease genes were predicted with bsniffer. Putative novel proteases were predicted with bgmix (Supplementary Note) and were visually inspected.

Variation analysis. SNPs (7,697,538) in reads aligned to the genome using the Burrows-Wheeler Aligner (BWA, version 0.5.9–r16; default parameters) were called using SAMtools86 (version 0.1.14 (r933:176); command ‘S AMtools
pulp - Bcfl Sref_genome Sbam; filtered q<20, D<100, with monomorphic, multi-allelic and singleton sites removed. Pairwise allele-sharing genetic distance was calculated55, and the resulting matrix was used for PCA and neighbor-joining tree construction (MATLAB ver. r2010b). Genetic ancestry for each individual was determined with ADMIXTURE44 in a given number of populations without using population designation. We filtered out SNPs with linkage disequilibrium ($r^2$ < 0.2 within each 100-SNP window using PLINK98, leaving 411,924 autosomal SNPs.

**Alu genetic analysis.** Best matching loci from CallA3.2 for each Alu subfamily were identified using BLAST51 or retrieved from a local RepeatMasker analysis using a custom library. Subfamilies with evidence of recent mobilization (divergence of up to 1%) from the consensus sequence were used for population genetics analyses. For phylogenetic analyses, Alu insertions of subfamilies were selected with varying divergence from the consensus sequence.

We retrieved marmoset Alu elements with ~500 bp of flanking sequence, identified orthologous loci using BLAST51 and retrieved the sequences if the flanking sequence matched unambiguously in the other genome and the Alu insertion was absent. We did this for human, chimpanzee, orangutan and rhesus macaque. We aligned the flanking sequence (BioLign/BioEdit) and selected primers (manually or using Primer3; ref. 99) to minimize nucleotide substitutions and other Alu insertions. Primers were tested using UCSC In-Silico PCR151 and were synthesized by Sigma-Aldrich.

PCR amplifications (96-well format) were performed using a Perkin Elmer GeneAmp 9700 or Bio-Rad i-cycler thermocycler in a 25-µl volume containing 15–25 ng of template DNA, 200 nM of each primer, 1.5–2 mM MgCl2, 1× PCR buffer (50 mM KCl, 10 mM Tris-HCl, pH 8.3), 0.2 mM dNTPs and 1–2 U Taq DNA polymerase. PCR conditions included an initial denaturation step at 94 °C for 90 s followed by 32 cycles of denaturation at 94 °C for 20 s, annealing at 57 °C for 20 s (see URLs for exceptions) and extension at 72 °C for 30–70 s, depending on the amplicon size, with a final extension step at 72 °C for 2 min. If necessary, we used a temperature gradient with HeLa DNA to determine the optimal annealing temperature. We fractionated 20 µl of each reaction in a 2% agarose gel containing 0.1 µg/ml ethidium bromide at 175 V for 50–60 min and visualized the amplicons with UV fluorescence.

Using genotype data from unlinked markers we inferred population structure, omitting information on the origin of the samples, with a model-based clustering analysis54–56 under the admixture model that assumes that individuals might have mixed ancestry.

The number of identifiable population clusters (K) with the highest likelihood was determined using initial values of K of 1 to 5, a burn-in period of 10,000,000 steps, a run length of 1,000,000 steps repeated at least 5 times. After determining K to be 2, 25 replications were run under identical conditions, omitting information on the origin of the samples, with a model-based clustering analysis54–56 under the admixture model that assumes that individuals might have mixed ancestry.

**Marmoset samples.** The marmoset samples used in this study were obtained under protocols approved by the relevant institutional animal care and use committees from animals maintained in Association for Assessment and Accreditation of Laboratory Animal Care (AAALAC)-accredited animal care programs.