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 (6×) 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 chimerism. 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 primates. 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 alloparental 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 ancestor and have evolved a reproductive and social system in which the dominant male and female monopolize breeding but benefit from alloparental 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), 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 goeldii*), 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, WGAC and WSSD. 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%) but more than in non-redundant sequences (4.7% of the whole genome), slightly less than observed in orangutan (3.8%) where specific collapses in the released assembly diverged. In addition, many families were absent in rhesus macaque, and thus almost half were apparently unique to apes.

Gene content and gene families

The Ensembl gene set 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. 6).

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

| GO category | Description | Genes | PSGs | Total | Excess | $P$ value (MWU) | Adjusted $P$ value (Holm) | Adjusted $P$ value (FET) |
|-------------|-------------|-------|------|-------|--------|----------------|--------------------------|--------------------------|
| 0005576     | Extracellular region | 150   | 1,954| 1.3   | $3.24 \times 10^{-15}$ | $9.80 \times 10^{-12}$ | $3.86 \times 10^{-17}$ |
| 0005615     | Extracellular space   | 63    | 429  | 2.4   | $2.52 \times 10^{-8}$  | $7.61 \times 10^{-5}$  | $1.31 \times 10^{-8}$  |
| 0005747     | Mitochondrial respiratory chain complex I | 8     | 14   | 9.4   | $1.81 \times 10^{-7}$  | $5.47 \times 10^{-4}$  | $2.72 \times 10^{-5}$  |
| 0006952     | Defense response      | 54    | 324  | 2.7   | $2.19 \times 10^{-6}$  | $6.59 \times 10^{-3}$  | $3.38 \times 10^{-9}$  |
| 0004872     | Receptor activity     | 103   | 866  | 2.0   | $3.42 \times 10^{-6}$  | $1.03 \times 10^{-2}$  | $1.05 \times 10^{-8}$  |
| 0007606     | Sensory perception of chemical stimulus | 20    | 136  | 2.4   | $5.82 \times 10^{-6}$  | $1.75 \times 10^{-2}$  | $1.26 \times 10^{-3}$  |
| 0003246     | Carbohydrate binding  | 29    | 203  | 2.3   | $6.81 \times 10^{-6}$  | $2.05 \times 10^{-2}$  | $1.78 \times 10^{-4}$  |
| 0006954     | Inflammatory response | 36    | 181  | 3.3   | $8.39 \times 10^{-6}$  | $2.52 \times 10^{-2}$  | $3.31 \times 10^{-8}$  |
| 0004984     | Olfactory receptor activity | 16    | 107  | 2.5   | $9.88 \times 10^{-6}$  | $2.97 \times 10^{-2}$  | $3.21 \times 10^{-3}$  |
| 0009611     | Response to wounding   | 53    | 332  | 2.6   | $2.93 \times 10^{-5}$  | $8.79 \times 10^{-2}$  | $1.73 \times 10^{-8}$  |
| 0006955     | Immune response       | 41    | 295  | 2.3   | $3.18 \times 10^{-5}$  | $9.53 \times 10^{-2}$  | $1.57 \times 10^{-5}$  |

*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 Platyrhini 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 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-euarchontoglires (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, CLECA4, 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
The genetic basis of twinning has received substantial attention in humans and other animals\(^{25-27}\). Genetic differences drive variation in ovulation number among sheep strains\(^ {25,28}\). 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 Supplementary Note). Potentially functional nonsynonymous substitutions in the EFSHR (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. goeldi (Fig. 4). That species had a reversal of this change to the sequence found in Old World monkeys and other non-twinning New World monkeys. C. goeldi is the only callitrichid that does not regularly twin, 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. The amino acid change encoded in WFIKKN1 is therefore a strong candidate for having a role in the origin of twinning in callitrichids.

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 NLRC4 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.

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 \( a \) 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 SIFT\(^ {22}\) and PolyPhen\(^ {23}\) indicated that these alterations likely affect the function of the corresponding proteins\(^ {24}\) (Supplementary Table 38 and Supplementary Note).

The genetic basis of twinning has received substantial attention in humans and other animals\(^ {25-27}\). Genetic differences drive variation in ovulation number among sheep strains\(^ {25,28}\). 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 SIFT\(^ {22}\) and PolyPhen\(^ {23}\) (Supplementary Note and ref. 24). Potentially functional nonsynonymous substitutions in the EFSHR (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. goeldi (Fig. 4). That species had a reversal of this change to the sequence found in Old World monkeys and other non-twinning New World monkeys. C. goeldi 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. The amino acid change encoded in WFIKKN1 is therefore a strong candidate for having a role in the origin of twinning in callitrichids.

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 NLRC4 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 cells\(^ {21}\) and the levels of interleukins IL-5 and IL-12B, involved in T cell development and in allergic responses\(^ {22}\). 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 fertility\(^ {33,34}\), was lost in marmoset, whereas ISP2, involved in embryo implantation\(^ {35}\), has been duplicated twice. KLK2/3, duplicated in the catarrhine ancestor\(^ {36}\) and involved in reproductive physiology\(^ {33}\), is non-functional in marmoset. Chymase and trypstatin protease changes and CMA1 and MAST duplications potentially affect the immune response\(^ {37,38}\) and mast cell biology, respectively. The duplicated CMA1 gene might be related to the murine-specific mast cell proteases (MCPs) that are absent in hominoids\(^ {39}\). Changes in the C terminus of MMP19, an IGFBP3-processing enzyme\(^ {40}\), 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 meiosis\(^ {41}\) (Supplementary Fig. 10 and Supplementary Note), was duplicated in catarrhine primates. Orthologs encoding all three functional PRDM9 domains have been computationally identified in placental mammals\(^ {42}\); 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 platyrhine 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. Sequence changes in the WFKK1 gene support the phylogenetic tree, with four changes occurring on the branch leading to tamarins and marmosets and a single change in C. goeldii back to the residue found in other primates that produce singletons (purple).

We identified 107 polymorphic Ali insertsions in common marmosets (Supplementary Fig. 10a). Analysis of these insertions using Structure (version 3.3.2) indicated population structure among the marmosets and detected two populations (Supplementary Fig. 12 and Supplementary Table 41). The included marmosets showed varying degrees of admixture, with some individuals mostly assigned to one cluster and others assigned to both clusters (Supplementary Fig. 12). The Structure analysis suggests that the New England RPRC colony is assigned primarily to one cluster and the Wisconsin and Southwest NPRC colonies fall into the other cluster.

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, language (reviewed in ref. 48) or other uniquely human traits. 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.

URLs. NCBI Trace Archive, http://www.ncbi.nlm.nih.gov/Traces/trace.cgi; UCSC Genome Browser, http://genome.ucsc.edu; miRBase, http://www.mirbase.org; Ensembl, http://www.ensembl.org/;
International Union for Conservation of Nature (IUCN) Red List of Threatened Species, http://www.iucnredlist.org/; Primate Info Net, http://pin.primate.wisc.edu/factsheets/; Spanish National Bioinformatics Institute, http://www.inab.org/; Ensembl Genebuild Process Documentation, http://cvs.sanger.ac.uk/cgi-bin/viewvc.cgi/ensembl-doc/pipeline_docs/the_genebuild_process.txt?root=ensembl
&view=co; Ensembl Gene Annotation Pipeline for Marmoset, http://www.ensembl.org/info/docs/genebuild/genome_annotation.html; vertebrate RNA alignments, http://www.ebi.ac.uk/ena/; UniProt, SwissProt/TrEMBL protein sequences, http://www.uniprot.org/; RepeatMasker Open-3.0, http://www.repeatmasker.org/; Washington University (WU)-BLAST package, http://blast.wustl.edu/; miROrtho miRNA annotation database, http://cegg.unige.ch/mirortho; Cluster 3.0 and TreeView software, http://rana.lbl.gov/EisenSoftware.htm; miRmap, http://cegg.unige.ch/mirmap; protease genes, http://degradome.uniovi.es/; Alu PCR conditions and primers, http://batzerlab.lsu.edu/; BAC FISH mapping data exploration, http://www.biologia.uniba.it/marmoset/

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. The sequences are available in the NCBI Trace Archive (see URLs) using the query SPECIES_CODE = ‘CALLITHRIX JACCUS’ together with TRACE_TYPE_CODE = ‘454’ for 454 transcript sequences, ‘WGS’ for plasmid reads, ‘FINISHING’ for BAC finishing reads or ‘CLONEEND’ for fosmid and BAC end sequences. The Illumina sequencing data are available from NCBI under BioProject 13630, and genomic sequences for nine other marmosets are available under BioProject 20401. Data for short RNAs sequenced using Illumina technology are available from miRBase (see URLs). The sequence assembly is accessioned in GenBank (ACFV00000000.1) and is available in NCBI under genome build 1.1 (http://www.ncbi.nlm.nih.gov/mapview/mapview_search.cgi?taxid=9483). The data are also available from the Washington University Genome Institute web site (http://genome.wustl.edu/genomes/view/callithrix_jaccus/), the Baylor College of Medicine Human Genome Sequencing Center web site (https://www.hgsc.bcm.edu/non-human-primates/marmoset-genome-project/), the UCSC Genome Browser (GCA_000004665.1) and Ensembl (C. jaccus3.2.1; January 2011). Cytogenetic data are presented at Campus Universitario Bari, Italy (http://www.biologia.uniba.it/marmoset/).

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

The authors acknowledge the contributions of the sequence production staff of the Human Genome Sequencing Center: K. Abraham, C. Adams, C. Allen, U. Anosike, T. Attaway, D. Bandaranaik, A. Bell, S.N. Bell, B. Beltran, C. Bickham, J. Chacko, A. Chavez, H.-S. Chu, M. Coyle, M.L. Davila, L. Davy-Carroll, S. Denson, Y. Ding, S. Dugan, V. Ebbing, S. Fernandez, P. Fernandez, A. Ferrer, J. Garner, R. Garcia III, T. Garrett, E. Hawkins, S. Hines, M. Holder, B. Hollins, H. Jiang, B. Johnson, H. Kisambo, L. Lago, M. Lago, C.-Y. Lai, T.-K. Lee, P. Legall III, S. Lemon, R. Medellin, E. Martinez, I. Mercado, C. Mercado, M. Munidasa, D. Ngo, P. Nguyen, O. Nwaokelumen, M. Obregon, C. Onwure, A. Parra, H. Paul, A. Perez, Y. Perez, E. Primus, J. Quirino, B. Schneider, I. Sisson, X.-Z. Song, A. Svatek, T. Taylor, R. Thelus, N. Thomas, R. Thornton, Z. Trejos, K. Usmani, S. Vattalith, D. Villasana, D. Walker, K. Wang, S. Wang, C. White, A. Williams, J. Williams, J. Woodworth and L. Zhang. The Washington University Genome Sequencing Center acknowledges the many people who contributed to the sequencing and analysis in this project who are not named here individually. We thank I. Steitz for miRNA data used to annotate the 3’ UTRs of genes. The miRNA analysis group acknowledges the contributions of D. Jayapakshi, C. Athulathmudali, H. Jiang and A. Moehring. We gratefully acknowledge the assistance of D. Opheim with the figures. 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.J. Wafula), 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-O011133 (Southwest NPRC), and support from the National Science Foundation (NSF-BCS-0751508 to D.E.W.) and the VEGA grant agency: 1/0719/14 (T.V.) 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. J.X. was supported by NHGRI grant K99 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

K.C.W., W.C.R., J.R. and D.L. led the Marmoset Genome Sequencing and Analysis Consortium project. Principal investigators R.A. Gibbs and R.K.W. provided material support. R.A. Gibbs, R.K.W., D.M.M., E.R.M., G.M.W. and W.C.W. led the sequencing project. K.C.W., J.R., R.A.H., K.M.A. and S.D.T. prepared the manuscript. S.D.T. provided samples for genomic sequencing and contributed information on the biology of marmosets. J.E.H., L.B.J., H.S., S.D.T., D.J.W. and J.X. contributed the chimerism estimates. K.B., S.F., P.F., J. Herrero and B.J.R. contributed comparative alignments. M. Ruffner, S. Searle and J.-H.V. annotated the genes. J.W.H. and P.M. assembled the genome sequence. K.M.A., B.B., R.A.H., S.D.T. and T.V. analyzed growth-related genes. D.H. investigated immune-related genes. N.A., O.C. and M. Rocchi performed karyotype analysis. P.D., J.M. and B.Z. prepared the BAC library. S.L.L., L.V.N., J.N., L.P., L.L.P., C.M.W. and W.Y. prepared the plasmid sequencing libraries. D.G., P.F., J.M., J.W.H., R.A.H., J.S.M., M. Raveendran, J.R., B.S., I.B.T., C.E.V., W.X., K.C.W. and E.M.Z. performed the miRNA analysis. M.A.B., R.H., L.B.J., M.K.K., M.C.R., A.F.A.S., S.D.T., B.U., J.A.W., D.J.W. and J.X. analyzed the population genetics. B.B., C.K., L.T. and T.V. analyzed positively selected genes. R.C.-H.D.R., N.A.R. and S.P. defined primate-constrained sequence elements. C.L.-O., X.S.P., V.Q. and D.R. analyzed protease genes. C.C.-F., P.H.G., J. Herrero, E.V.K., A.J.W. and E.M.Z. analyzed protein-coding genes. M.A.B., C.C.F., R.H., M.K.K., A.F.A.S., B.U., J.A.W. and Q.W. analyzed performance of the repeats. S.B.C., K.G.M., C.R. and D.E.W. collected samples. C.C., E.E.E., M.W.H., E.K., B.L.-G., T.M.-B., S. Sajadian, D.R.S. and M.V. analyzed segmental duplications. Sequence was produced by M.M.C., A.C., M.D.D., K.D.D., H.H.D., R.G.F., C.E., L.A.F., R.S.F., R.A. Gibbs, T.A.G., Y.H., J. Hume, S.N.J., C.L.K., D.M.K., M.B.M., D.M.M., N.B.N., G.O.O., S.J.R., J.S. and R.G.F. analyzed segmental segment-related genes.

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/

1. Benirschke, K., Anderson, J.M. & Brownhill, L.E. Marrow chimerism in marmosets. J. Biol. Chem. 23677–23684 (2008).

2. Gengozian, N., Batson, J.S. & Eide, P. Hematologic and cytogenetic evidence for hematopoietic chimerism in the marmoset, Tamarinus naglicicius. Cyto genetics 3, 384–393 (1994).

3. Goldizen, A.W. Tamarin and marmoset mating systems: unusual flexibility. Trop. Ecol. 3, 35–40 (1988).

4. Leutenegger, W. Maternal-fetal weight relationships in primates. Folia Primatol. (Basel) 20, 280–293 (1991).

5. Tardif, S.D. & Jaquish, C.E. The common marmoset as a model for nutritional impacts upon reproduction. Am. NY Acad. Sci. 709, 214–215 (1994).

6. Maroog, G. & Cheverud, J. Size as a line of least resistance II: direct selection on size or correlated response due to constraints? Evolution 64, 1470–1488 (2010).

7. Kondás, K., Szlama, G., Trexler, M. & Pathy, I. Both WFIKKN1 and WFIKKN2 have high affinity for growth and differentiation factors 8 and 11. J. Biol. Chem. 283, 23677–23684 (2008).

855 

ARTICLES

NATURE GENETICS VOLUME 46 | NUMBER 8 | AUGUST 2014
The Marmoset Genome Sequencing and Analysis Consortium:

Kim C Worley1,2, Wesley C Warren3, Jeffrey Rogers1,2, Devin Locke3, Suzette D Tardif4, Kjersti M Aagaard5, Nicoletta Archidiacono5, Nirmala Arul Raya7, Mark A Batzer5, Kathryn Beal9, Brona Brejova10, Oronzo Capozzi6, Saverio B Capuano11, Claudio Casola12,13,38, David Rio Deiros1,2, Mariano Rocchi6, David Rodriguez25, Corinna Ross4, Magali Ruffier9,16, San Juana Ruiz1,2, Mariano Rocchi6, David Rodriguez25, Corinna Ross4, Magali Ruffier9,16, San Juana Ruiz1,2, Mariano Rocchi6, David Rodriguez25, Corinna Ross4, Magali Ruffier9,16, San Juana Ruiz1,2,

15. Wang, Q.F. (2005).
12. Chimpanzee Sequencing and Analysis Consortium. Initial sequence of the chimpanzee genome. Nature 437, 99–105 (2005).
20. Choi, J.H. et al. Expression profiling of mammalian male meiosis and gametogenesis identifies novel candidate genes for roles in the regulation of fertility. Mol. Cell. Biol. 15, 1031–1043 (2004).
13. Locke, D.P. et al. Comparative and demographic analysis of orang-utan genomics. Nature 469, 529–533 (2011).
14. Gibbs, R.A. et al. Evolutionary and biomedical insights from the rhesus macaque genome. Science 316, 222–234 (2007).
12. Palmer, J.S. et al. Detection of weakly conserved ancestral mammalian regulatory sequences by primate comparisons. Genome Biol. 8, R1 (2007).
13. Pavlopoulou, A. et al. Comparative demographic history of the vervet monkey (Cercopithecus aethiops). PLoS Genet. 5, e1000144 (2009).
12. Adzhubei, I.A. et al. A method and server for predicting damaging missense mutations. Nat. Methods 7, 1013–1018 (2009).
10. Poter, S.C. et al. The Ensembl analysis pipeline. Genome Res. 14, 934–941 (2004).
22. Yonehara, S. et al. A model of human primate speciation based on multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164, 1567–1587 (2003).
33. Dorus, S., Evans, P.D., Wyckoff, G.J., Choi, S.S. & Lahn, B.T. Rate of molecular evolution of the seminal protein gene SEMG2 correlates with levels of female promiscuity. Nat. Genet. 36, 1326–1329 (2004).
18. Alexander, D.H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 19, 1695–1664 (2009).
19. Falush, D., Stephens, M. & Pritchard, J.K. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 165, 1567–1587 (2003).
34. Pavlopoulou, A., Pamalakas, G., Michalopoulos, I. & Sotiropoulou, G. Evolutionary history of tissue kallikreins. PLoS ONE 5, e13781 (2010).
27. Palmer, J.S. et al. Expression profiling of mammalian male meiosis and gametogenesis identifies novel candidate genes for roles in the regulation of fertility. Mol. Cell. Biol. 15, 1031–1043 (2004).
13. Sharma, N., Kaur, J., Xu, X., Zur Nieden, N. & Rancourt, D. Characterization of secretory leukocyte protease inhibitor as an inhibitor of implantation serine proteinases. Mol. Reprod. Dev. 75, 1136–1142 (2008).
12. Pavlopoulou, A., Pamalakas, G., Michalopoulos, I. & Sotiropoulou, G. Evolutionary history of tissue kallikreins. PLoS ONE 5, e13781 (2010).
27. Palmer, J.S. et al. Expression profiling of mammalian male meiosis and gametogenesis identifies novel candidate genes for roles in the regulation of fertility. Mol. Cell. Biol. 15, 1031–1043 (2004).
Saba Sajjadian1, Jirneh Santibanez1,2, Daniel R Schrider12,13, Steve Searle16, Helen Skaltsky22,31, Benjamin Soibam18, Arian F A Smit21, Jayantha B Tennakoon18, Lubomir Tomaska32, Brygg Ullmer33,34, Charles E Vejar17, Mario Ventura15, Albert J Vilella9, Tomas Vinar10, Jan-Hinnerk Vogel16, Jerilyn A Walker8, Qing Wang8, Crystal M Warner1,2, Derek E Wildman35, David J Witherspoon23, Rita A Wright1,2, Yuanqing Wu1,2, Weimin Xiao18, Jinchuan Xing23,38, Evgeny M Zdobnov17,36,37, Baoli Zhu14,38, Richard A Gibbs1,2 & Richard K Wilson3

1Human Genome Sequencing Center, Baylor College of Medicine, Houston, Texas, USA. 2Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas, USA. 3The Genome Institute, Washington University St. Louis, St. Louis, Missouri, USA. 4Barshop Institute for Longevity and Aging Studies, University of Texas Health Science Center at San Antonio, San Antonio, Texas, USA. 5Department of Obstetrics and Gynecology, Baylor College of Medicine, Houston, Texas, USA. 6Department of Biology, Bari University, Bari, Italy. 7Genome Institute of Singapore, Singapore. 8Department of Biological Sciences, Louisiana State University, Baton Rouge, Louisiana, USA. 9European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, UK. 10Faculty of Mathematics, Physics and Informatics, Comenius University in Bratislava, Bratislava, Slovakia. 11Wisconsin National Primate Research Center, Madison, Wisconsin, USA. 12Department of Biology, Indiana University, Bloomington, Indiana, USA. 13School of Informatics and Computing, Indiana University, Bloomington, Indiana, USA. 14Children’s Hospital Oakland Research Institute, Oakland, California, USA. 15Department of Genome Sciences, University of Washington, Seattle, Washington, USA. 16Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, UK. 17Genetic Medicine and Development, University of Geneva Medical School, Geneva, Switzerland. 18Department of Biology and Biochemistry, University of Houston, Houston, Texas, USA. 19Department of Pathology, Baylor College of Medicine, Houston, Texas, USA. 20Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, Massachusetts, USA. 21Institute for Systems Biology, Seattle, Washington, USA. 22Whitehead Institute, Cambridge, Massachusetts, USA. 23Department of Human Genetics, University of Utah School of Medicine, Salt Lake City, Utah, USA. 24Institut für Populationsgenetik, Vetmeduni Vienna, Vienna, Austria. 25Departamento de Bioquímica y Biología Molecular, Instituto Universitario de Oncología, Barcelona, Spain. 26Spanish National Bioinformatics Institute (INB), Barcelona, Spain. 27Institució Catalana de Recerca i Estudis Avançats, Institut de Biologia Evolutiva Consejo Superior de Investigaciones Científicas–Universitat Pompeu Fabra, Barcelona, Spain. 28Division of Comparative Pathology, New England Primate Research Center, Harvard University Medical School, Southborough, Massachusetts, USA. 29CNAG (Centro Nacional de Análisis Genómico), Barcelona, Spain. 30Center for Biomolecular Science and Engineering, School of Engineering, University of California, Santa Cruz, Santa Cruz, California, USA. 31Howard Hughes Medical Institute, Whitehead Institute, Cambridge, Massachusetts, USA. 32Department of Genetics, Faculty of Natural Sciences, Comenius University in Bratislava, Bratislava, Slovakia. 33Center for Computation and Technology, Louisiana State University, Baton Rouge, Louisiana, USA. 34Department of Computer Sciences, Louisiana State University, Baton Rouge, Louisiana, USA. 35Center for Molecular Medicine and Genetics, Wayne State University School of Medicine, Detroit, Michigan, USA. 36Swiss Institute of Bioinformatics, Geneva, Switzerland. 37Division of Molecular Biosciences, Imperial College London, London, UK. 38Present 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, chimerism 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 (Phred^51 quality of ≥20), were assembled using PCAP^52. 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 BLASTZ^53. Orthologous and paralogous alignments^54 were differentiated, and only ‘reciprocal best’ alignments were retained (totaling 15,576,643 bases) were merged into the final chromosomal files.

Methods12. Documented inversions based on FISH data (see URLs) and inversions 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 Dnase1 hypersensitivity in human ESCs^77. The eight human sequences and their mouse orthologs (identified using liftOver; Supplementary Table 11) were amplified from their respective genomic DNA, cloned into the SaI site downstream of luciferase in the pGL3-Pou5f1 vector using the Gateway Cloning System (Invitrogen) and transfected with the reporter constructs into human ESCs (H1- WA-01, WiCell Research Institute) and mouse ESCs (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; A08-332monkB, 13.0 years, female) and two placenta samples, using 36-bp reads on the Illumina 1G Genome Analyzer^78. Usable reads were identified as described^79, omitting reads with <4 copies, <10 nt or >10 repetitive nucleotides and reads that matched Escherichia coli sequences using
MicroRNAs predicted using SVM (group F). Human precursor miRNAs (miRBase 14.0; ref. 81) with WU-BLASTN46,69 (see URLs) matches of >20 bp in length to calljac3.2 (-M 1 -N -1 -Q 3 -R 2 -W 9 --filter dust --mformat 2 -hspsepSmax 40 -e 1e-3) were extended to match their entire length and realigned using MAFFT82 (maxiterate 1000 –localpair –quiet). Matches were identified with (i) length of >40 bp, (ii) a completely conserved seed region (mature miRNA nucleotides 2–8), (iii) >90% mature miRNA sequence identity, (iv) total precursor conservation over >50% of the length, (v) at most two gaps in mature miRNA, (vi) minimum free folding energy (MFE) of <15 kcal/mol, (vii) >40% of bases paired, (viii) mature regions not overlapping a multiple-loop region and (ix) probability of <5% for a randomly shuffled hit sequence to have a lower MFE than the native sequences for <95% of conserved matches. The hit with the lowest MFE for overlapping loci was supported by a Subject Vector Machine (SVM) model trained to distinguish miRNAs from unspecific genomic stem-loop sequences or other noncoding RNAs. Developed for the mirOrtho annotation database83 (see URLs), the model incorporates the thermodynamic, structural and sequence features found in known miRNA genes. Using an initial BLAST e-value cutoff of 1 x 10^-6, an SVM score of greater than 0.5 and 100% mature miRNA sequence conservation to any known miRBase miRNA, we identified 859 genes (group F). Expression profiles were estimated by counting filtered small RNA sequences mapping within 4 bp on the same chromosome as the miRNA, normalized by total number of usable reads. Euclidean hierarchical clustering of genes and arrays with Cluster 3.0 and TreeView84 (see URLs) used the log2 transformation of miRNAs per 10 million usable reads with the median expression value across the 6 samples set to zero. MiRmap85 identified miRNAs with 3’ UTR matches to miRNA bases 2–8 and predicted repression strength with a model encompassing thermodynamic, conservation, probabilistic and sequence-based approaches. We computed the total energy of the miRNA-mRNA duplex (similar to in ref. 86) and the branch length score87, implemented the SPH test in PhyloP88 and computed the statistical significance of the seed match on the basis of 3’ UTR sequence composition. The 3 features of the TargetScan context score89 were included in miRmap for a total of 11 features, of which 3 were novel (see URLs). These data were generated by mapping all human RefSeq genes to marmoset on the basis of the UCSC ‘Other RefSeq’ track, and multiple mapping locations in marmoset were retained and were represented by [refseqAccession],1, [refseqAccession]2, etc. Where the 3’ UTR differs between mapped locations, this difference could reflect true paralogs or assembly errors. The extracted marmoset 3’ UTRs were aligned using MAFFT82 to the TargetScan 5.1 23-way UTR alignments, and marmoset target genes were identified with 3’ UTR binding sites for the mature marmoset chromosome 22 family miRNAs.

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 UCSC’s 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) and that were present in <3 species were removed, and we analyzed separately families with genes in only one lineage (Euarchonta, 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:6:6)7, orang:13:11, maca:24:16, marm:40:47, (mouse:17,rat:17):70:6, (dog:74,horse:74):9:cow:83):10), where numbers correspond to millions of years (Supplementary Note).

Possitively 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 PAML93. Further downstream analysis such as enrichment analysis for GO categories was performed as described93. 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 SiFT94, and some candidates were omitted owing to conflicting evidence. Genes and coordinates are listed in Supplementary Table 39. The species used for alignment included Saginus bicolor maritimus, Saginus imperator imperator, Saginus midas niger, Saginus fuscicolis weddelli, Callithrix cebuella pygmaea, Leontopithecus rosalia, Cebus apella, Callimico goeldii, Ateles belzebuth and Saimiri sciureus (species with an asterisk were also selected for miRNA sequencing). Sanger sequencing reads were assembled (Velvet95), mapped to the genome (BLAT91) and aligned (MAFFT92). In 49 of the 82 exons sequenced, data were insufficient to determine whether the marmoset nonsynonymous substitutions were callitrichine 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 tex script, and the locations of marmoset protease genes were predicted with bnaver. Putative novel proteases were predicted with hgmix (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 SAMTools96 (version 0.1.14 (+933:176); command ‘$ SAMtools

© 2014 Nature America, Inc. All rights reserved.
pilup -Bcf Sref_genome Sbam; filtered-q>20, D<100), with monomorphic, multi-allelic and singleton sites removed. Pairwise allele-sharing genetic distance was calculated, and the resulting matrix was used for PCA and neighbor-joining tree construction (MATLAB ver. r2010b). Genetic ancestry for each individual was determined with ADMIXTURE 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 PLINK, leaving 411,924 autosomal SNPs.

**Afu genetic analysis.** Best matching loci from CalJac3.2 for each Afu subfamily were identified using BLAT or retrieved from a local RepeatMasker analysis as expressed sequence tags. The number of identifiable population clusters \((K\)\) was determined using initial values of \(K\) for each individual was determined with ADMIXTURE 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 PLINK, leaving 411,924 autosomal SNPs.

**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.

51. Ewing, B. & Green, P. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res. 8, 186–194 (1998).
52. Huang, X., Wang, J., Arul, S., Yang, S.P. & Hillier, L. PCAP: a whole-genome assembly program. Genome Res. 13, 2164–2170 (2003).
53. Kent, W., Baertsch, R., Hinrichs, A., Miller, W. & Haussler, D. Evolution’s cauldron: genomic data on the African great ape ancestor. Nature Genet. 39, 125–125 (2011).
54. Lewis, S. et al. Apollo: a sequence annotation editor. Genome Biol. 3, R1002 (2002).
55. Margulis, A., Gertz, E.M., Schaffer, A.A. & Agarwala, R. A fast and symmetric DUST implementation to mask low-complexity DNA sequences. J. Comput. Biol. 13, 1028–1040 (2006).
56. Olson, G. Tandem repeat resolution: a program to analyze DNA sequences. Nucleic Acids Res. 27, 573–580 (1999).
57. Down, T. & Hubbard, T. J. Computational detection and location of transcription factor binding sites in mammalian genomic DNA. Genome Res. 12, 458–461 (2002).
58. Davuluri, R.V., Grosse, I. & Zhang, M.Q. Computational identification of promoters and first exons in the human genome. Nat. Genet. 29, 412–417 (2001).
59. Gardiner-Garden, M. & Frommer, M. CpG islands in vertebrate genomes. J. Mol. Biol. 196, 261–282 (1987).
60. Lowe, T.M. & Eddy, S.R. TRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 25, 955–964 (1997).
61. Goujon, M. et al. A new bioinformatics analysis tools framework at EMBL-EBI. Nucleic Acids Res. 38, W695–W699 (2010).
62. Sayers, E.W. et al. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 38, D5–D16 (2010).
63. Aittschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. Basic local alignment search tool. J. Mol. Biol. 215, 404–410 (1990).
64. Lopez, S., Silventoinen, V., Robinson, S., Kibria, A. & Gish, W. WU-BLAST2 server at the European Bioinformatics Institute. Nucleic Acids Res. 31, 3795–3798 (2003).
65. Birney, E., Clamp, M. & Durbin, R. GeneWise and Genowise. Genome Res. 14, 1087–1090 (2004).
66. Slater, G.S. & Birney, E. Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics 6, 3 (2005).
67. Pruitt, K.K., Tatusova, T., Klimke, W. & Maglott, D.R. NCBI Reference Sequence: a curated non-redundant sequence database of genomes, cDNAs, and proteins. Nucleic Acids Res. 37, D226–D232 (2009).
68. Eyras, E., Caccamo, M., Curwen, V. & Clamp, M. ESTGene: alternative splicing from ESTs in Ensembl. Genome Res. 14, 976–987 (2004).
69. Lewis, S.E. et al. Apollo: a sequence annotation editor. Genome Biol. 3, R1002 (2002).
70. Margulis, A., Gertz, E.M., Schaffer, A.A. & Agarwala, R. A fast and symmetric DUST implementation to mask low-complexity DNA sequences. J. Comput. Biol. 13, 1028–1040 (2006).
71. Kent, W.J. BLAT—the BLAST-like alignment tool. Genome Res. 12, 656–664 (2002).
72. Deininger, P.C. & Heinloth, S. Predicting deleterious amino acid substitutions. Genome Res. 11, 863–874 (2001).
73. Kraft, K., Misawa, K., Kuma, K. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 30, 3660–3668 (2002).
74. Gerlach, D., Krintsenva, E.V., Rahman, N., Vejnar, C.E. & Zdobnov, E.M. miRTho: computational survey of microRNA genes. Nucleic Acids Res. 37, D111–D117 (2009).
75. Eisen, M.B., Spellman, P.T., Brown, P.O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. USA 95, 14863–14868 (1998).
76. Vejnar, C.E. & Zdobnov, E.M. MiRnap: comprehensive predictive microRNA target repression strength. Nucleic Acids Res. 40, 11673–11681 (2012).
77. Kertesz, M., Iovino, N., Urenstein, U., Gaul, U. & Siepel, A. The role of site accessibility in microRNA target recognition. Nat. Genet. 39, 1278–1281 (2007).
78. Stark, A. et al. Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures. Nature 450, 219–232 (2007).
79. Polard, K.S., Hubisz, M.J., Rosenbloom, K.R. & Siepel, A. Detection of nonfunctional substitution rates on mammalian phylogenies. Genome Res. 20, 110–120 (2010).
80. Grimson, A. et al. MicroRNA targeting specificity in mammals: determinants beyond seed pairing. Mol. Cell 27, 91–105 (2007).
81. Rhead, B. et al. The UCSC Genome Browser database: update 2010. Nucleic Acids Res. 38, D153–D161 (2010).
82. Hahn, M.W., De Bie, T., Stajich, J.E., Nguyen, C. & Cristianini, N. Estimating the substitution rates on mammalian phylogenies. Genome Res. 15, 1153–1160 (2005).
83. Stajich, J.E. et al. The additivity of morphology for reconstructing the early history of placental mammals. Syst. Biol. 56, 673–684 (2007).
84. Yang, Z. & Nielsen, R. Codon-substitution models for detecting molecular adaptation based on novel deep-sequencing data. Nucleic Acids Res. 38, 2011–2021 (2010).
85. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81, 559–575 (2007).
86. Rozen, S. & Skaletsky, H. Primer3 on the WWW for general users and for biologist programmers. Methods Mol. Biol. 132, 365–386 (2000).