1-1-2014

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

Recommended Citation

Worley, K., Warren, W., Rogers, J., Locke, D., Muzny, D., Mardis, E., Weinstock, G., Tardif, S., Aagaard, K., Archidiacono, N., Arul Rayan, N., Batzer, M., Beal, K., Brejova, B., Capozzi, O., Capuano, S., Casola, C., Chandrabose, M., Cree, A., Diep Dao, M., De Jong, P., Cruz-Herrera del Rosario, R., Delehaunty, K., Dinh, H., Eichler, E., Fitzgerald, S., Flicek, P., Fontenot, C., Fowler, R., Fronick, C., Fulton, L., Fulton, R., Gabisi, R., & Gerlach, D. (2014). The common marmoset genome provides insight into primate biology and evolution. *Nature Genetics, 46* (8), 850-857. https://doi.org/10.1038/ng.3042

nature genetics
OPEN

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[1,2]. 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 primates[3–5]. 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 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[6] 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.

WFIKKN1, 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

*A full list of authors and affiliations appears at the end of the paper.

Received 27 November 2013; accepted 27 June 2014; published online 20 July 2014; doi:10.1038/ng.3042

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

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

[a]GO category number. [b]Positively selected genes (PSGs) identified with a threshold of $P < 0.05$. [c]Total number of genes in the GO category. [d]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.

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[8] and WSSD[9]. 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 genomes[10,12–14], 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**).

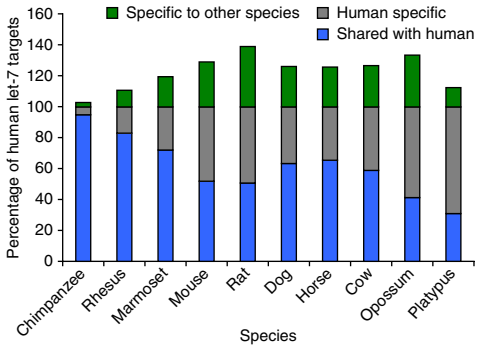
Constrained sequence evolution indicates natural selection and therefore implies conserved function. By extension, lineage-specific constraint indicates lineage-specific function[15,16]. Using the marmoset genome, we detected 23,849 elements constrained in anthropoid primates but not in non-primate mammals[17] (**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 set[18] (**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**).

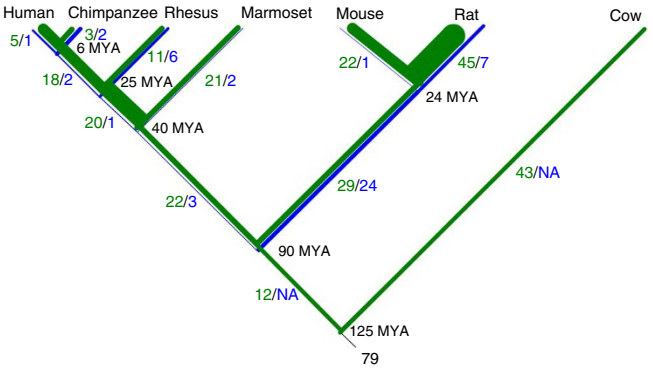
A phylogenetic framework including 4 other primates, 2 rodents and 3 Laurasiatheria showed 429 primate-specific gene families, among which few were present only in marmoset (**Supplementary Fig. 7**, **Supplementary Tables 12–19** and **Supplementary Note**). More than half of these families (221/429) 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.

Figure 1 Predicted let-7–regulated genes (miRNA targets). The numbers of protein-coding genes with predicted targets for let-7 miRNA binding in the 3′ UTR are shown. Only single-copy orthologs are counted, and numbers are relative to the number found in humans (100% on the scale). The number of gene targets shared with humans decreases as the evolutionary distance increases, as expected. However, the proportion of let-7 targets shared with humans is comparable for marmoset, dog, horse and cow, whereas mouse and rat share fewer targets with humans than other non-primate placental mammals.



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 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*, *CLEC4A*, *ANKZF1*, *KRT8* and *CATSPERG*) were previously unrecognized as being positively selected[19]. 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 studies[19], 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 humans[20,21]. 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 2 Gains and losses of let-7–regulated genes. The conserved let-7 miRNA targets variable numbers of genes. We mapped let-7 target gene gains (green) and losses (blue) to the phylogenetic tree of the analyzed species; line thickness indicates the rate of gain or loss. Gains and losses that occurred twice on independent lineages were omitted. Gains exceed losses on each branch of the tree, and the total number gained (196) is 4 times the number lost (49). Primate lineage changes (gains plus losses) exceed non-primate lineage changes (except for the branch leading to rat after divergence from mouse). MYA, million years ago.

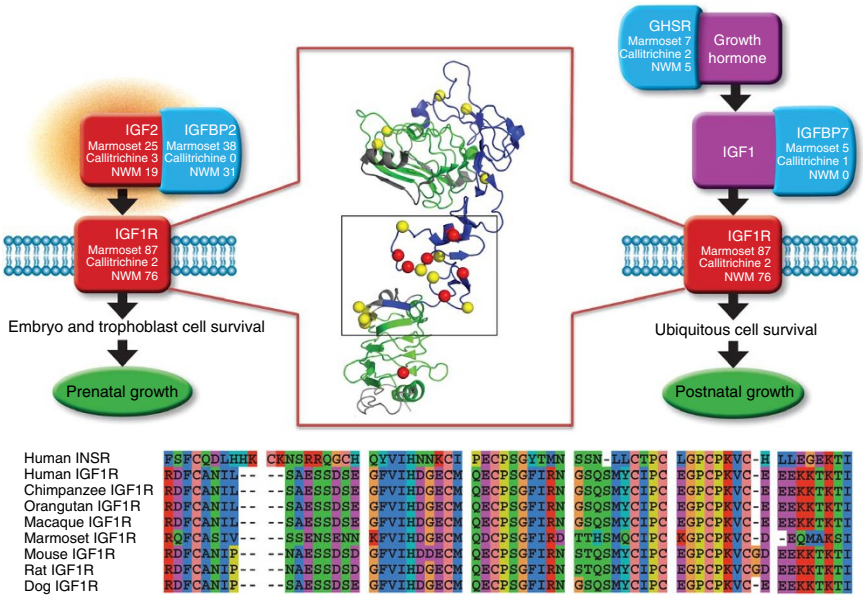


Figure 3 Residues under positive selection in IGF1R. The insulin-like growth factor 1 receptor (IGF1R) interacts with other proteins in growth hormone pathways and has a role in both prenatal (left) and postnatal (right) growth. Proteins encoded by genes in these pathways in marmoset that have residues 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 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 *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. 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[31] and the levels of interleukins IL-5 and IL-12B, involved in T cell development and in allergic responses[32]. 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 tryptase 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

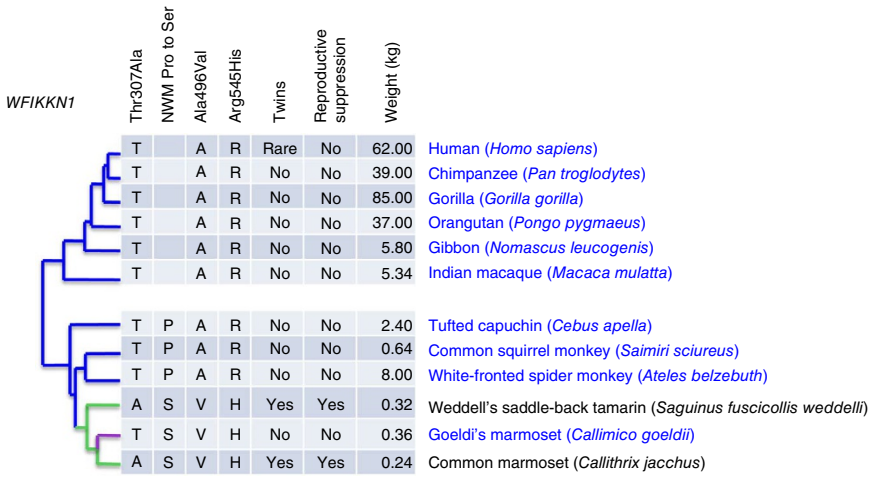


Figure 4 Twinning species and *WFIKKN1* sequence variation. Primate species tree showing species that regularly produce twins in green and those that produce singletons in blue or purple. The phylogeny appears as in ref. 50. In the table, nonsynonymous changes in marmoset *WFIKKN1* are labeled by the encoded amino acid change (p.Thr307Ala, chr. 12: 642,862; NWM Pro to Ser, chr. 12: 642,877, multiple-base insertion within p.Thr310_Ser311insSerSerSerProAla; p.Ala496Val, chr. 12: 643,445; p.Arg545His, chr. 12: 643,592). p.Arg545His is predicted by SIFT[22] to alter protein function and by PolyPhen[23] to be probably damaging. Features related to reproduction, including twin offspring, pair bonding and reproductive suppression in non-breeding females, and adult female weight are shown. Adult female weights are from the International Union for Conservation of Nature (IUCN Red List of Threatened Species, version 2013.2.; see URLs) and the Primate Info Net (apes and marmoset; see URLs). Species on the green branches exhibit phyletic dwarfing, an early period of developmental quiescence and a shared chimeric placenta. Sequence changes in the *WFIKKN1* 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).

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, and applied a maximum-likelihood method to estimate the proportion of chimerism present in the marmoset samples from the sequencing data (**Supplementary Note**). 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 on the basis of the first two principal components (**Supplementary Fig. 11a**), with individual M32784 from Southwest NPRC more similar to individuals from other primate centers. Next, we used ADMIXTURE[44] to assess the ancestry of each individual. With $K = 3$ (**Supplementary Fig. 11b**), three groups corresponding to the colonies were identified. New England RPRC and Wisconsin NPRC individuals formed distinct groups with little admixture. Consistent with the PCA result, two Southwest NPRC individuals (M32783 and M32784) showed appreciable admixture from the other colonies (**Supplementary Fig. 11b**). A neighbor-joining tree using the distance matrix (**Supplementary Fig. 11c**) confirmed that individuals from the same colony were grouped together, with the exception of M32784. The long terminal branch length suggests that most of the diversity exists among individuals.

We identified 107 polymorphic *Alu* insertions in common marmosets (**Supplementary Fig. 10a**). Analysis of these insertions using Structure (version 3.3.2)[45,46] 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[47], 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 (**Supplementary Fig. 13**). Such divergence at multiple levels does indeed underscore the remarkable nature of this platyrrhine monkey species.

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 JACCHUS' 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/map_search.cgi?taxid=9483). The data are also available from the Washington University Genome Institute web site (http://genome.wustl.edu/genomes/view/callithrix_jacchus/), 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_jacchus3.2.1; January 2010). 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. Bandaranaike, 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. Ebong, S. Fernandez, P. Fernando, A. Ferrer, J. Ganer, R. Garcia III, T. Garrett, E. Hawkins, S. Hines, M. Holder, B. Hollins, H. Jiang, B. Johnson, H. Kisamo, L. Lago, M. Lago, C.-Y. Lai, T.-K. Le, F. Legall III, S. Lemon, R. Madu, E. Martinez, I. Mercado, C. Mercado, M. Munidasa, D. Ngo, P. Nguyen, O. Nwaokelemeh, M. Obregon, C. Onwere, A. Parra, H. Paul, A. Perez, Y. Perez, E. Primus, J. Quiroz, B. Schneider, I. Sisson, X.-Z. Song, A. Svatek, T. Taylor, R. Thelus, N. Thomas, R. Thornton, Z. Trejos, K. Usmani, S. Vattathil, 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 J. Steitz for mRNA data used to annotate the 3′ UTRs of genes. The miRNA analysis group acknowledges the contributions of D. Rajapakshe, 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 HG003273 (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.B.), HG002385 (E.E.E.) and P51-OD011133 (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.W., 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.F.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. Ruffier, S. Searle and J.-H.V. annotated the genes. L.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.J.d.J., D.M. and B.Z. prepared the BAC library. S.L.L., L.V.N., I.N., L.P., L.-L.P., C.M.W. and Y.W. prepared the plasmid sequencing libraries. D.G., P.H.G., R.A.H., J.S.M., M. Raveendran, J.R., B.S., J.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., P.F., J. Herrero, E.V.K., A.J.V. 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. performed analysis 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. Sajjadian, 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.F., L.A.F., R.S.F., R.A. Gabis, T.A.G., Y.H., J. Hume, S.N.J., V.J., C.L.K., L.R.L., Y.L., J.L., E.R.M., M.B.M., D.M.M., N.B.N., G.O.O., S.J.R., J.S. and R.A.W. D.R.D. analyzed SNP variation. K.M.A., R.A.H. and S.D.T. analyzed twinning-related genes.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Benirschke, K., Anderson, J.M. & Brownhill, L.E. Marrow chimerism in marmosets. *Science* **138**, 513–515 (1962).
2. Gengozian, N., Batson, J.S. & Eide, P. Hematologic and cytogenetic evidence for hematopoietic chimerism in the marmoset, *Tamarinus nigricollis*. *Cytogenetics* **3**, 384–393 (1964).
3. Goldizen, A.W. Tamarin and marmoset mating systems: unusual flexibility. *Trends Ecol. Evol.* **3**, 36–40 (1988).
4. Leutenegger, W. Maternal-fetal weight relationships in primates. *Folia Primatol. (Basel)* **20**, 280–293 (1973).
5. Tardif, S.D. & Jaquish, C.E. The common marmoset as a model for nutritional impacts upon reproduction. *Ann. NY Acad. Sci.* **709**, 214–215 (1994).
6. Marroig, 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. & Patthy, L. Both WFIKKN1 and WFIKKN2 have high affinity for growth and differentiation factors 8 and 11. *J. Biol. Chem.* **283**, 23677–23684 (2008).

8. Bailey, J.A., Yavor, A.M., Massa, H.F., Trask, B.J. & Eichler, E.E. Segmental duplications: organization and impact within the current human genome project assembly. *Genome Res.* **11**, 1005–1017 (2001).
9. Bailey, J.A. *et al.* Recent segmental duplications in the human genome. *Science* **297**, 1003–1007 (2002).
10. Lander, E.S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
11. Cheng, Z. *et al.* A genome-wide comparison of recent chimpanzee and human segmental duplications. *Nature* **437**, 88–93 (2005).
12. Chimpanzee Sequencing and Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**, 69–87 (2005).
13. Locke, D.P. *et al.* Comparative and demographic analysis of orang-utan genomes. *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).
15. Wang, Q.F. *et al.* Detection of weakly conserved ancestral mammalian regulatory sequences by primate comparisons. *Genome Biol.* **8**, R1 (2007).
16. Wang, Q.F. *et al.* Primate-specific evolution of an *LDLR* enhancer. *Genome Biol.* **7**, R68 (2006).
17. del Rosario, R.C.H., Arul Rayan, N. & Prabhakar, S. Noncoding origins of anthropoid traits and a new null model of transposon functionalization. *Genome Res.* (in the press).
18. Potter, S.C. *et al.* The Ensembl analysis pipeline. *Genome Res.* **14**, 934–941 (2004).
19. Kosiol, C. *et al.* Patterns of positive selection in six mammalian genomes. *PLoS Genet.* **4**, e1000144 (2008).
20. Choi, J.H. *et al.* Clinical and functional characteristics of a novel heterozygous mutation of the *IGF1R* gene and IGF1R haploinsufficiency due to terminal 15q26.2->qter deletion in patients with intrauterine growth retardation and postnatal catch-up growth failure. *J. Clin. Endocrinol. Metab.* **96**, E130–E134 (2011).
21. Fang, P. *et al.* Severe short stature caused by novel compound heterozygous mutations of the insulin-like growth factor 1 receptor (*IGF1R*). *J. Clin. Endocrinol. Metab.* **97**, E243–E247 (2012).
22. Kumar, P., Henikoff, S. & Ng, P.C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.* **4**, 1073–1081 (2009).
23. Adzhubei, I.A. *et al.* A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248–249 (2010).
24. Harris, R.A. *et al.* Evolutionary genetics and implications of small size and twinning in callitrichine primates. *Proc. Natl. Acad. Sci. USA* **111**, 1467–1472 (2014).
25. Bodin, L. *et al.* A novel mutation in the bone morphogenetic protein 15 gene causing defective protein secretion is associated with both increased ovulation rate and sterility in Lacaune sheep. *Endocrinology* **148**, 393–400 (2007).
26. Hoekstra, C. *et al.* Dizygotic twinning. *Hum. Reprod. Update* **14**, 37–47 (2008).
27. Palmer, J.S. *et al.* Novel variants in growth differentiation factor 9 in mothers of dizygotic twins. *J. Clin. Endocrinol. Metab.* **91**, 4713–4716 (2006).
28. Galloway, S.M. *et al.* Mutations in an oocyte-derived growth factor gene (BMP15) cause increased ovulation rate and infertility in a dosage-sensitive manner. *Nat. Genet.* **25**, 279–283 (2000).
29. Vinzing, M. *et al.* NAIP and Ipaf control *Legionella pneumophila* replication in human cells. *J. Immunol.* **180**, 6808–6815 (2008).
30. Zhao, Y. *et al.* The NLRC4 inflammasome receptors for bacterial flagellin and type III secretion apparatus. *Nature* **477**, 596–600 (2011).
31. McNerney, M.E., Guzior, D. & Kumar, V. 2B4 (CD244)-CD48 interactions provide a novel MHC class I–independent system for NK-cell self-tolerance in mice. *Blood* **106**, 1337–1340 (2005).
32. Lloyd, C.M. & Hessel, E.M. Functions of T cells in asthma: more than just T_H2 cells. *Nat. Rev. Immunol.* **10**, 838–848 (2010).
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).
34. Schlecht, U. *et al.* Expression profiling of mammalian male meiosis and gametogenesis identifies novel candidate genes for roles in the regulation of fertility. *Mol. Biol. Cell* **15**, 1031–1043 (2004).
35. Sharma, N., Kaur, J., Xu, H., 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).
36. Pavlopoulou, A., Pampalakis, G., Michalopoulos, I. & Sotiropoulou, G. Evolutionary history of tissue kallikreins. *PLoS ONE* **5**, e13781 (2010).
37. Caughey, G.H. Mast cell tryptases and chymases in inflammation and host defense. *Immunol. Rev.* **217**, 141–154 (2007).
38. Trivedi, N.N., Tong, Q., Raman, K., Bhagwandin, V.J. & Caughey, G.H. Mast cell α and β tryptases changed rapidly during primate speciation and evolved from γ-like transmembrane peptidases in ancestral vertebrates. *J. Immunol.* **179**, 6072–6079 (2007).
39. Schwartz, S. *et al.* Human-mouse alignments with BLASTZ. *Genome Res.* **13**, 103–107 (2003).
40. Sadowski, T., Dietrich, S., Koschinsky, F. & Sedlacek, R. Matrix metalloproteinase 19 regulates insulin-like growth factor–mediated proliferation, migration, and adhesion in human keratinocytes through proteolysis of insulin-like growth factor binding protein-3. *Mol. Biol. Cell* **14**, 4569–4580 (2003).
41. Cheung, V.G., Sherman, S.L. & Feingold, E. Genetics. Genetic control of hotspots. *Science* **327**, 791–792 (2010).
42. Myers, S. *et al.* Drive against hotspot motifs in primates implicates the *PRDM9* gene in meiotic recombination. *Science* **327**, 876–879 (2010).
43. Fumasoni, I. *et al.* Family expansion and gene rearrangements contributed to the functional specialization of *PRDM* genes in vertebrates. *BMC Evol. Biol.* **7**, 187 (2007).
44. Alexander, D.H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
45. Falush, D., Stephens, M. & Pritchard, J.K. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* **164**, 1567–1587 (2003).
46. Pritchard, J.K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959 (2000).
47. Pollard, K.S. *et al.* An RNA gene expressed during cortical development evolved rapidly in humans. *Nature* **443**, 167–172 (2006).
48. Enard, W. FOXP2 and the role of cortico-basal ganglia circuits in speech and language evolution. *Curr. Opin. Neurobiol.* **21**, 415–424 (2011).
49. Kingsley, C.B. Identification of causal sequence variants of disease in the next generation sequencing era. *Methods Mol. Biol.* **700**, 37–46 (2011).
50. Perelman, P. *et al.* A molecular phylogeny of living primates. *PLoS Genet.* **7**, e1001342 (2011).

The Marmoset Genome Sequencing and Analysis Consortium:

Kim C Worley[1,2], Wesley C Warren[3], Jeffrey Rogers[1,2], Devin Locke[3], Donna M Muzny[1,2], Elaine R Mardis[3], George M Weinstock[1–3,38], Suzette D Tardif[4], Kjersti M Aagaard[5], Nicoletta Archidiacono[6], Nirmala Arul Rayan[7], Mark A Batzer[8], Kathryn Beal[9], Brona Brejova[10], Oronzo Capozzi[6], Saverio B Capuano[11], Claudio Casola[12,13,38], Mimi M Chandrabose[1,2], Andrew Cree[1,2], Marvin Diep Dao[1,2], Pieter J de Jong[14,38], Ricardo Cruz-Herrera del Rosario[7], Kim D Delehaunty[3], Huyen H Dinh[1,2], Evan E Eichler[15], Stephen Fitzgerald[9], Paul Flicek[9,16], Catherine C Fontenot[8], R Gerald Fowler[1,2], Catrina Fronick[3], Lucinda A Fulton[3], Robert S Fulton[3], Ramatu Ayiesha Gabisi[1,2], Daniel Gerlach[17,38], Tina A Graves[3], Preethi H Gunaratne[1,2,18,19], Matthew W Hahn[12,13], David Haig[20], Yi Han[1,2], R Alan Harris[1,2,5], Javier Herrero[9,38], LaDeana W Hillier[3], Robert Hubley[21], Jennifer F Hughes[22], Jennifer Hume[1,2], Shalini N Jhangiani[1,2], Lynn B Jorde[23], Vandita Joshi[1,2], Emre Karakor[15], Miriam K Konkel[8], Carolin Kosiol[24], Christie L Kovar[1,2], Evgenia V Kriventseva[16], Sandra L Lee[1,2], Lora R Lewis[1,2], Yih-shin Liu[1,2], John Lopez[1,2], Carlos Lopez-Otin[25], Belen Lorente-Galdos[26,27], Keith G Mansfield[28,38], Tomas Marques-Bonet[27,29], Patrick Minx[3], Doriana Misceo[6,14], J Scott Moncrieff[17], Margaret B Morgan[1,2], Lynne V Nazareth[1,2], Irene Newsham[1,2], Ngoc Bich Nguyen[1,2], Geoffrey O Okwuonu[1,2], Shyam Prabhakar[7], Lora Perales[1,2], Ling-Ling Pu[1,2], Xose S Puente[25], Victor Quesada[25], Megan C Ranck[8], Brian J Raney[30], Muthuswamy Raveendran[1,2], David Rio Deiros[1,2], Mariano Rocchi[6], David Rodriguez[25], Corinna Ross[4], Magali Ruffier[9,16], San Juana Ruiz[1,2],

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

[1]Human Genome Sequencing Center, Baylor College of Medicine, Houston, Texas, USA. [2]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas, USA. [3]The Genome Institute, Washington University St. Louis, St. Louis, Missouri, USA. [4]Barshop Institute for Longevity and Aging Studies, University of Texas Health Science Center at San Antonio, San Antonio, Texas, USA. [5]Department of Obstetrics and Gynecology, Baylor College of Medicine, Houston, Texas, USA. [6]Department of Biology, Bari University, Bari, Italy. [7]Genome Institute of Singapore, Singapore. [8]Department of Biological Sciences, Louisiana State University, Baton Rouge, Louisiana, USA. [9]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, UK. [10]Faculty of Mathematics, Physics and Informatics, Comenius University in Bratislava, Bratislava, Slovakia. [11]Wisconsin National Primate Research Center, Madison, Wisconsin, USA. [12]Department of Biology, Indiana University, Bloomington, Indiana, USA. [13]School of Informatics and Computing, Indiana University, Bloomington, Indiana, USA. [14]Children's Hospital Oakland Research Institute, Oakland, California, USA. [15]Department of Genome Sciences, University of Washington, Seattle, Washington, USA. [16]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, UK. [17]Genetic Medicine and Development, University of Geneva Medical School, Geneva, Switzerland. [18]Department of Biology and Biochemistry, University of Houston, Houston, Texas, USA. [19]Department of Pathology, Baylor College of Medicine, Houston, Texas, USA. [20]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, Massachusetts, USA. [21]Institute for Systems Biology, Seattle, Washington, USA. [22]Whitehead Institute, Cambridge, Massachusetts, USA. [23]Department of Human Genetics, University of Utah School of Medicine, Salt Lake City, Utah, USA. [24]Institut für Populationsgenetik, Vetmeduni Vienna, Vienna, Austria. [25]Departamento de Bioquímica y Biología Molecular, Instituto Universitario de Oncología, Universidad de Oviedo, Oviedo, Spain. [26]Spanish National Bioinformatics Institute (INB), Barcelona, Spain. [27]Institució Catalana de Recerca i Estudis Avançats, Institut de Biologia Evolutiva Consejo Superior de Investigaciones Científicas–Universitat Pompeu Fabra, Barcelona, Spain. [28]Division of Comparative Pathology, New England Primate Research Center, Harvard University Medical School, Southborough, Massachusetts, USA. [29]CNAG (Centro Nacional de Análisis Genómico), Barcelona, Spain. [30]Center for Biomolecular Science and Engineering, School of Engineering, University of California, Santa Cruz, Santa Cruz, California, USA. [31]Howard Hughes Medical Institute, Whitehead Institute, Cambridge, Massachusetts, USA. [32]Department of Genetics, Faculty of Natural Sciences, Comenius University in Bratislava, Bratislava, Slovakia. [33]Center for Computation and Technology, Louisiana State University, Baton Rouge, Louisiana, USA. [34]Department of Computer Sciences, Louisiana State University, Baton Rouge, Louisiana, USA. [35]Center for Molecular Medicine and Genetics, Wayne State University School of Medicine, Detroit, Michigan, USA. [36]Swiss Institute of Bioinformatics, Geneva, Switzerland. [37]Division of Molecular Biosciences, Imperial College London, London, UK. [38]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, 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 calJac3) 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[39]. Orthologous and paralogous alignments[53] were differentiated, and only 'reciprocal best' alignments were retained and used to generate the marmoset AGP files, as in previously described methods[12]. 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 and methods[54] and assembled using Newbler[55] and BLAT[56] were aligned against the marmoset genome.

Using >700 human BAC clones, we established the synteny block organization of the marmoset chromosomes and disambiguated inconsistencies and uncertainties in the genome assembly.

Gene feature annotation. Annotations with RefSeq[57] and Ensembl[18,58] used the general methods described (see URLs). The raw compute stage of Ensembl annotation (**Supplementary Fig. 6a**) screened genomic sequence using RepeatMasker[59] (version 3.2.5; 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 TRF[61].

Predicted features included transcription start sites (Eponine-scan[62] and FirstEF[63]), CpG islands (described in ref. 64) and tRNAs[65]. Genscan results on repeat-masked sequence were input for UniProt[66], UniGene[67] and Vertebrate RNA (see URLs) by WU-BLAST[68,69] alignments, resulting in 252,582 UniProt, 316,384 UniGene and 317,679 Vertebrate RNA sequences aligning.

Genewise[70] and Exonerate[71] produced coding sequence models using marmoset and human UniProt, SwissProt/TrEMBL (see URLs) and RefSeq[72] 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 Exonerate[72] (**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 humans[73]) 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 Apollo[74] 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 process[19,20] 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)[8], and the second assessed excess depth of coverage of whole-genome sequencing data mapped to the assembly (WSSD)[9]. All scaffolds were repeat masked (RepeatMasker; see URLs) and window masked[75] 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 genomes[76], 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 DNase I 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 SalI 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; A08-337monkB, 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[78,79], omitting reads with <4 copies, <10 nt or >10 repetitive nucleotides and reads that matched *Escherichia coli* sequences using

doi:10.1038/ng.3042

WU-BLAST[69] (**Supplementary Table 2**). Expressed miRNAs that were 100% conserved (group A; 291 miRNAs) or had 1–3 mismatches (group B; 240 miRNAs) relative to at least one other species in miRBase 17.0 (ref. 80) were identified. Known miRNAs in miRBase 17.0 that mapped to the marmoset genome identified miRNAs that were conserved (100% match, group C; 119 miRNAs) or novel (with 1–3 mismatches, group D; 120 miRNAs). Sequences in groups A–D (~22 nt in length) aligned with BLAT (*-stepSize = 5 repMatch = 100000 -minScore = 0 -minIdentity = 0 -fine*) and their flanking sequences (±200 bp) extracted from UCSC were folded twice using Vienna RNAfold[78] to confirm hairpin structures with the mapped sequence in the mature miRNA location. Group E contained the 91 novel miRNAs identified (20 passed high-stringency filters), which were trimmed to include only the hairpin bases (60–150 nt) (**Supplementary Table 2**).

WU-BLAST comparison identified marmoset miRNAs that were conserved in four anthropoid primates (*-nogaps -N -1000 -mformat = 2 -warning -kap -hspmax = 10*) (marmoset, calJac3; human, hg18; rhesus, rhemac2; orangutan, ponAbe2; chimpanzee, panTro2; from UCSC). BLAT mapping (*-stepSize = 5 repMatch = 100000 -minScore = 0 -minIdentity = 0 –fine*) of the precursor miRNA hairpins encoded on marmoset chromosome 22 to rhesus, orangutan and chimpanzee identified the best matches, which were realigned to marmoset miRNA hairpins, using Smith-Waterman to identify nucleotide changes in the mature miRNA sequences. Human chromosome 19 hairpins were mapped to calJac3 using Galaxy liftOver and BLAT alignment and were realigned as above (see conservation in **Supplementary Tables 3–8**).

MicroRNAs predicted using SVM (group F). Human precursor miRNAs (miRBase 14.0; ref. 81) with WU-BLASTN[68,69] (see URLs) matches of >20 bp in length to calJac3.2 (-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 MAFFT[82] (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 *e* value for overlapping loci was subjected to a Support Vector Machine (SVM) model trained to distinguish miRNAs from unspecific genomic stem-loop sequences or other noncoding RNAs. Developed for the miROrtho annotation database[83] (see URLs), the model incorporates the thermodynamic, structural and sequence features found in known miRNA genes. Using an initial BLAST *e*-value cutoff of 1×10^{-6}, an SVM score of greater than 0.5 and 100% mature miRNA sequence conservation to any known miRBase miRNA, we identified 589 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 TreeView[84] (see URLs) used the $\log_2$ transformation of miRNAs per 10 million usable reads with the median expression value across the 6 samples set to zero.

MiRmap[85] identified mRNAs 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 score[87], implemented the SPH test in PhyloP[88] and computed the statistical significance of the seed match on the basis of 3′ UTR sequence composition. The 3 features of the TargetScan context score[89] 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 MAFFT[82] 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[90] 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 (118,345), 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 CAFE[91] analysis using the following ultrametric tree built according to ref. 92: ((((((chimp:6,human:6):7, orang:13):11, macaca:24):16, marmoset: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**).

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 PAML[93]. Further downstream analysis such as enrichment analysis for GO categories was performed as described[19]. 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 alignments[18] 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 SIFT[94], 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 martinsi**, *Saguinus imperator imperator*, *Saguinus midas niger**, *Saguinus fuscicollis 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 (Velvet[95]), mapped to the genome (BLAT[51]) and aligned (MAFFT[82]). 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 tbex 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 SAMtools[96] (version 0.1.14 (r933:176); command '$ SAMtools

pileup -Bvcf $ref_genome $bam'; filtered q>20, D<100), with monomorphic, multi-allelic and singleton sites removed. Pairwise allele-sharing genetic distance was calculated[97], 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[44] 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[98], leaving 411,924 autosomal SNPs.

***Alu* genetic analysis.** Best matching loci from CalJac3.2 for each *Alu* subfamily were identified using BLAT[51] 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 BLAT[51] 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 PCR[51] 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 $MgCl_2$, 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 analysis[45,46] 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 1,000,000 iterations and 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 burn-in and run length settings. Structure analyses were run on a desktop machine with four CPUs.

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 International (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., Aluru, S., Yang, S.P. & Hillier, L. PCAP: a whole-genome assembly program. *Genome Res.* **13**, 2164–2170 (2003).
53. Kent, W.J., Baertsch, R., Hinrichs, A., Miller, W. & Haussler, D. Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proc. Natl. Acad. Sci. USA* **100**, 11484–11489 (2003).
54. Shin, H. *et al.* Transcriptome analysis for *Caenorhabditis elegans* based on novel expressed sequence tags. *BMC Biol.* **6**, 30 (2008).
55. Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376–380 (2005).
56. Kent, W.J. BLAT—the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).
57. International Applied Genomics Consortium. Genome sequence of the pea aphid *Acyrthosiphon pisum*. *PLoS Biol.* **8**, e1000313 (2010).
58. Curwen, V. *et al.* The Ensembl automatic gene annotation system. *Genome Res.* **14**, 942–950 (2004).
59. Bailey, J.A. & Eichler, E.E. Primate segmental duplications: crucibles of evolution, diversity and disease. *Nat. Rev. Genet.* **7**, 552–564 (2006).
60. Morgulis, 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).
61. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).
62. Down, T.A. & Hubbard, T.J. Computational detection and location of transcription start sites in mammalian genomic DNA. *Genome Res.* **12**, 458–461 (2002).
63. 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).
64. Gardiner-Garden, M. & Frommer, M. CpG islands in vertebrate genomes. *J. Mol. Biol.* **196**, 261–282 (1987).
65. 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).
66. Goujon, M. *et al.* A new bioinformatics analysis tools framework at EMBL-EBI. *Nucleic Acids Res.* **38**, W695–W699 (2010).
67. Sayers, E.W. *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **38**, D5–D16 (2010).
68. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
69. Lopez, R., Silventoinen, V., Robinson, S., Kibria, A. & Gish, W. WU-Blast2 server at the European Bioinformatics Institute. *Nucleic Acids Res.* **31**, 3795–3798 (2003).
70. Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. *Genome Res.* **14**, 988–995 (2004).
71. Slater, G.S. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**, 31 (2005).
72. Pruitt, K.D., Tatusova, T., Klimke, W. & Maglott, D.R. NCBI Reference Sequences: current status, policy and new initiatives. *Nucleic Acids Res.* **37**, D32–D36 (2009).
73. Eyras, E., Caccamo, M., Curwen, V. & Clamp, M. ESTGenes: alternative splicing from ESTs in Ensembl. *Genome Res.* **14**, 976–987 (2004).
74. Lewis, S.E. *et al.* Apollo: a sequence annotation editor. *Genome Biol.* **3**, RESEARCH0082 (2002).
75. Morgulis, A., Gertz, E.M., Schaffer, A.A. & Agarwala, R. WindowMasker: window-based masker for sequenced genomes. *Bioinformatics* **22**, 134–141 (2006).
76. Marques-Bonet, T. *et al.* A burst of segmental duplications in the genome of the African great ape ancestor. *Nature* **457**, 877–881 (2009).
77. Thomas, D.J. *et al.* The ENCODE Project at UC Santa Cruz. *Nucleic Acids Res.* **35**, D663–D667 (2007).
78. Creighton, C.J., Reid, J.G. & Gunaratne, P.H. Expression profiling of microRNAs by deep sequencing. *Brief. Bioinform.* **10**, 490–497 (2009).
79. Creighton, C.J. *et al.* Discovery of novel microRNAs in female reproductive tract using next generation sequencing. *PLoS ONE* **5**, e9637 (2010).
80. Kozomara, A. & Griffiths-Jones, S. miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res.* **39**, D152–D157 (2011).
81. Griffiths-Jones, S., Saini, H.K., van Dongen, S. & Enright, A.J. miRBase: tools for microRNA genomics. *Nucleic Acids Res.* **36**, D154–D158 (2008).
82. Katoh, 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**, 3059–3066 (2002).
83. Gerlach, D., Kriventseva, E.V., Rahman, N., Vejnar, C.E. & Zdobnov, E.M. miROrtho: computational survey of microRNA genes. *Nucleic Acids Res.* **37**, D111–D117 (2009).
84. 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).
85. Vejnar, C.E. & Zdobnov, E.M. MiRmap: comprehensive prediction of microRNA target repression strength. *Nucleic Acids Res.* **40**, 11673–11683 (2012).
86. Kertesz, M., Iovino, N., Unnerstall, U., Gaul, U. & Segal, E. The role of site accessibility in microRNA target recognition. *Nat. Genet.* **39**, 1278–1284 (2007).
87. Stark, A. *et al.* Discovery of functional elements in 12 *Drosophila* genomes using evolutionary signatures. *Nature* **450**, 219–232 (2007).
88. Pollard, K.S., Hubisz, M.J., Rosenbloom, K.R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* **20**, 110–121 (2010).
89. Grimson, A. *et al.* MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Mol. Cell* **27**, 91–105 (2007).
90. Rhead, B. *et al.* The UCSC Genome Browser database: update 2010. *Nucleic Acids Res.* **38**, D613–D619 (2010).
91. Hahn, M.W., De Bie, T., Stajich, J.E., Nguyen, C. & Cristianini, N. Estimating the tempo and mode of gene family evolution from comparative genomic data. *Genome Res.* **15**, 1153–1160 (2005).
92. Springer, M.S. *et al.* The adequacy of morphology for reconstructing the early history of placental mammals. *Syst. Biol.* **56**, 673–684 (2007).
93. Yang, Z. & Nielsen, R. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol. Biol. Evol.* **19**, 908–917 (2002).
94. Ng, P.C. & Henikoff, S. Predicting deleterious amino acid substitutions. *Genome Res.* **11**, 863–874 (2001).
95. Zerbino, D.R. & Birney, E. Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res.* **18**, 821–829 (2008).
96. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993 (2011).
97. Xing, J. *et al.* Fine-scaled human genetic structure revealed by SNP microarrays. *Genome Res.* **19**, 815–825 (2009).
98. 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).
99. Rozen, S. & Skaletsky, H. Primer3 on the WWW for general users and for biologist programmers. *Methods Mol. Biol.* **132**, 365–386 (2000).