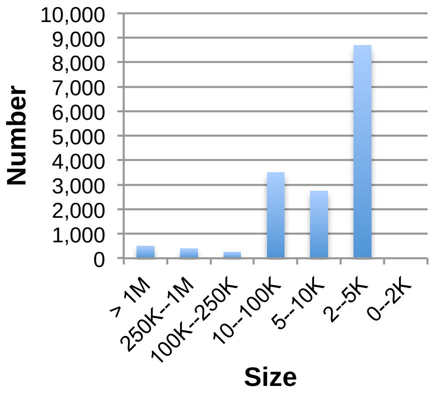
Supplementary Information

The Common Marmoset Genome Provides Insight into Primate Biology and Evolution

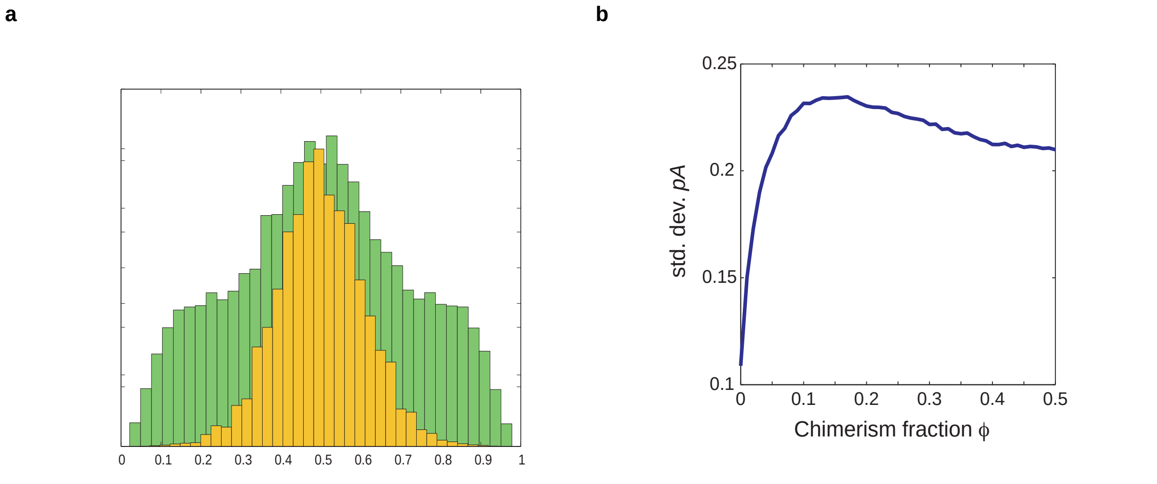
Authors: The Marmoset Genome Sequencing and Analysis Consortium

Table of Contents

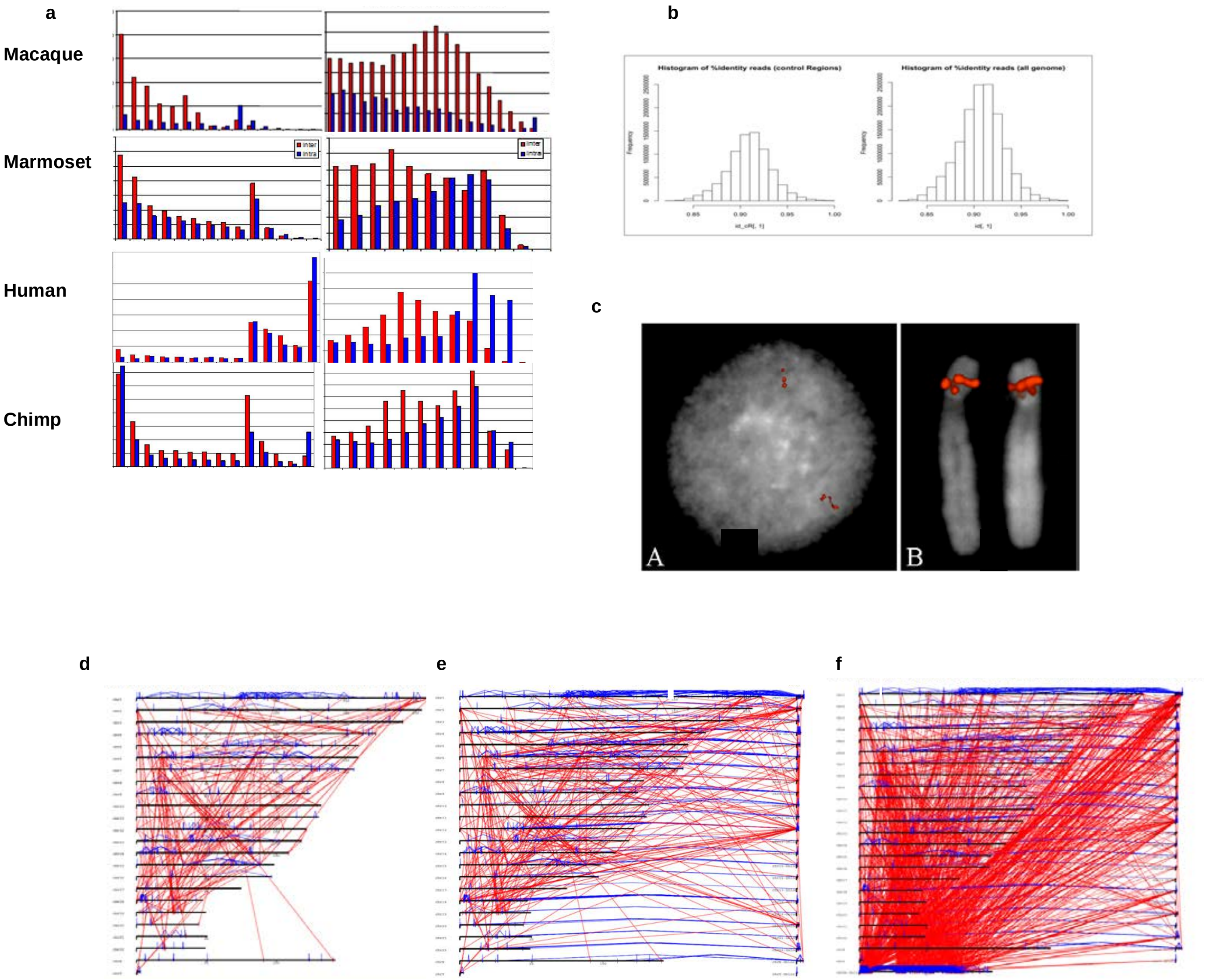
I. Supplementary Figures 1-13.



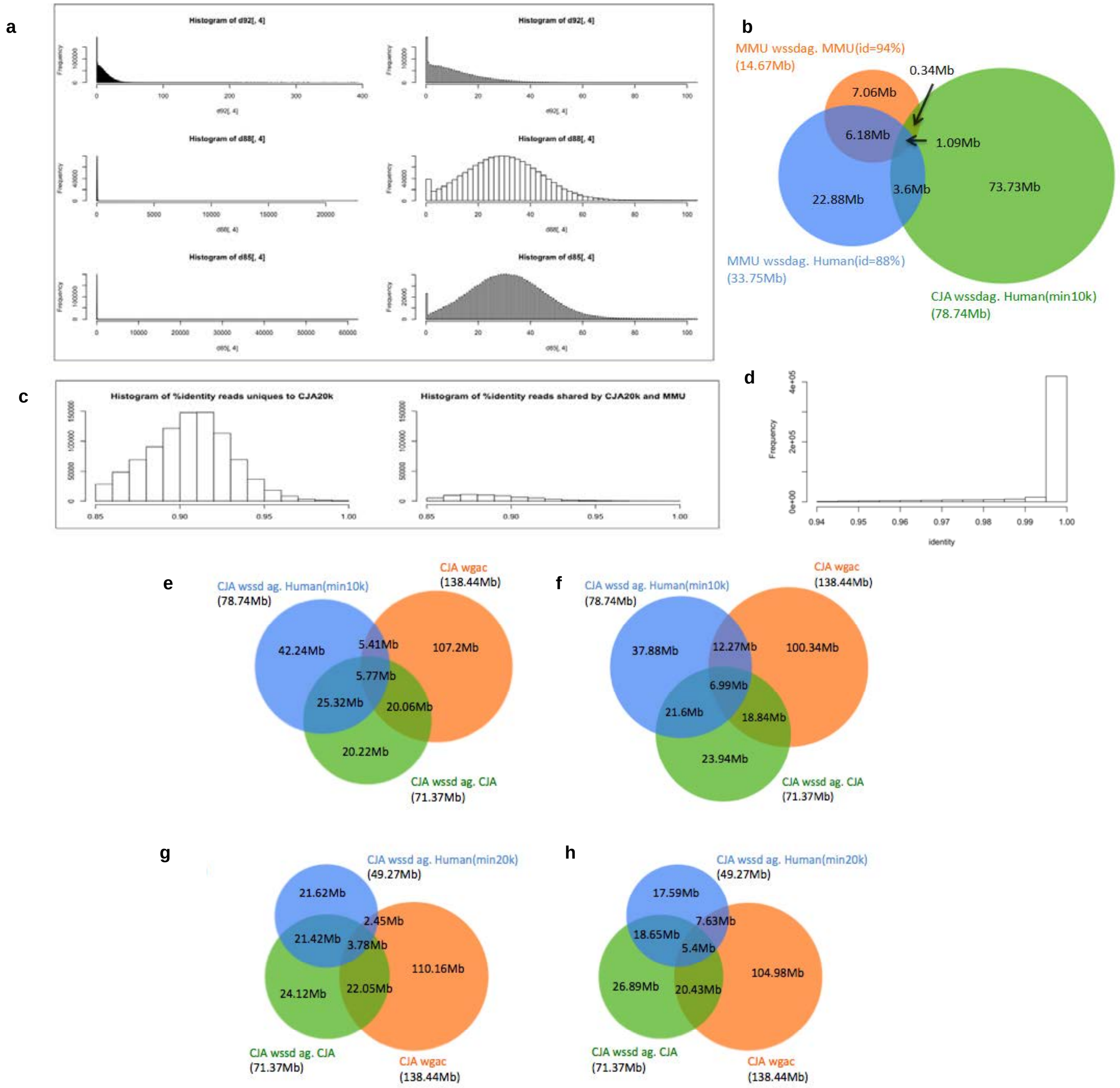
Supplementary Figure 1 | Assembly Scaffolds. Sizes of scaffolds in the assembled genome.



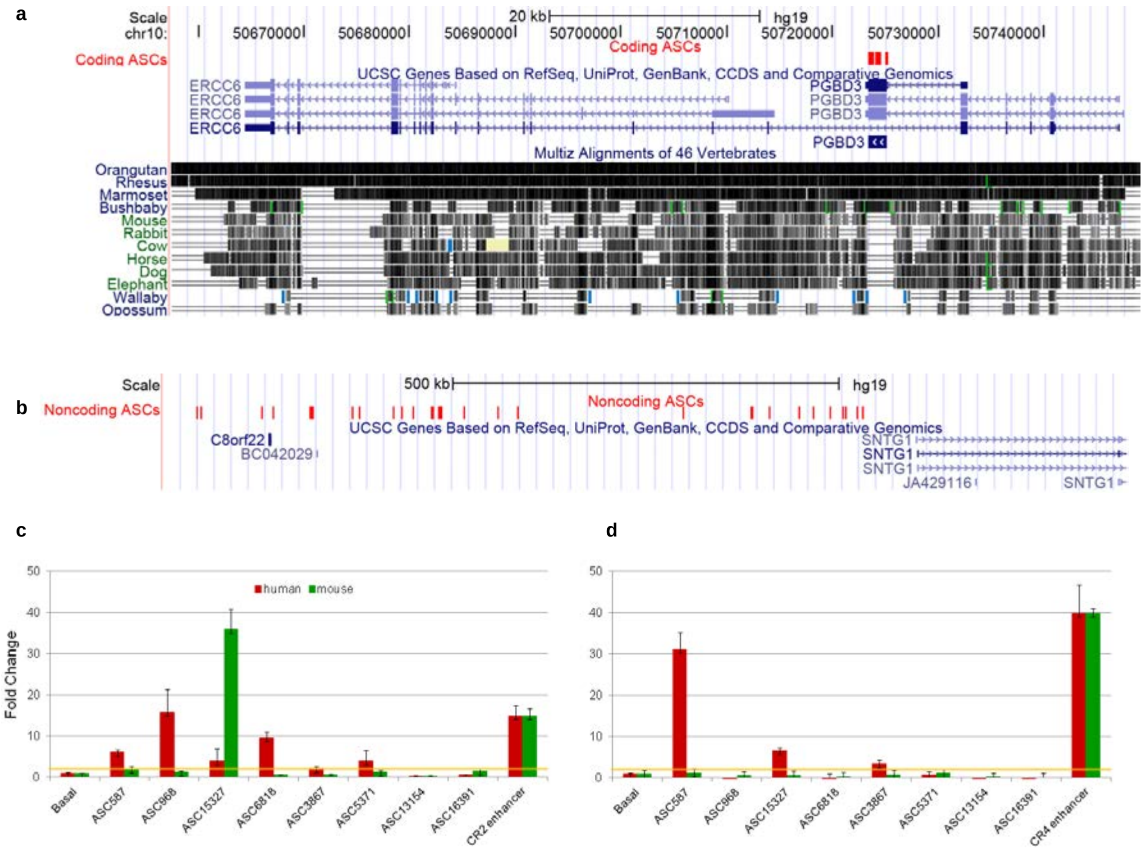
Supplementary Figure 2 | Summary of Chimerism Analysis Data. **a**. Simulated distributions of the proportion of allele A reads out of the total number of reads at a locus for a marmoset with chimeric fraction Φ = 0.33 (green) and Φ = 0 (i.e., not chimeric; yellow). **b.** Standard deviation in the proportion of allele A reads across loci vs. chimerism fraction Φ, by simulation. Excess variance in pA is expected for reads from chimeric samples.



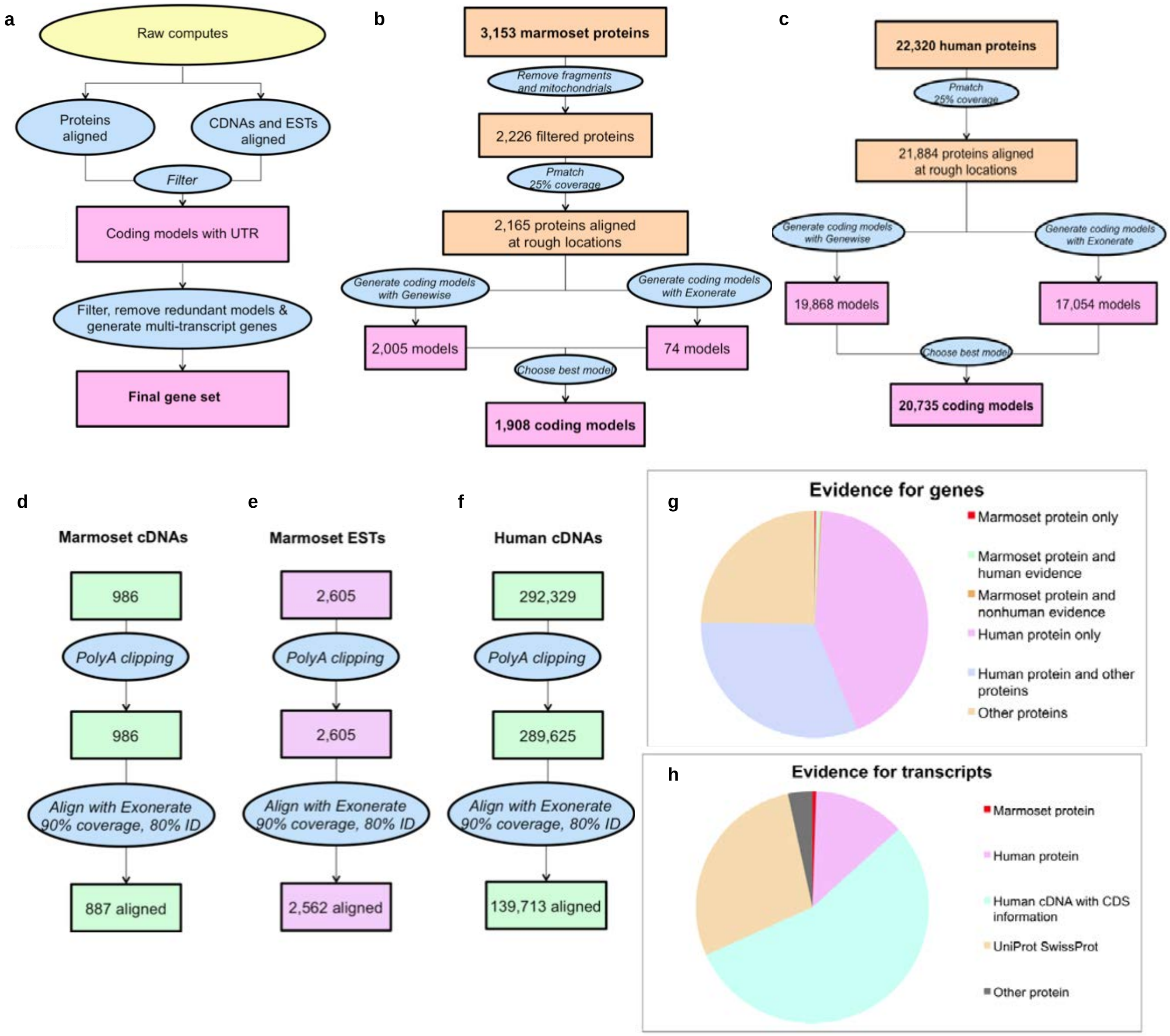
Supplementary Figure 3 | Summary of Structural Variation Data. a. Pairwise length (left) and Identity (right) distribution of marmoset assembly intrachromosomal (blue) and intrachromosomal (red) segmental duplications. Length distributions are partitioned into 1kb to 10kb in 1kb increments and 10 kb to 50 kb in 10 kb increments. Identity distributions are shown for 90 to 100% identity. Note the depletion of duplication with identity >99.5% and excess of short duplication in comparison to other assemblies. **b.** Histograms of the aligned marmoset reads mapped to the human genome. Bins of percent identity between 85% and 100%, showing number of reads in each bin, for reads mapped to regions without a known CNV in any previously analyzed primate (left) and the entire genome (right). In both cases, the mean is 91% and there are relatively few reads mapped with an identity <85%. **c.** Examples of FISH, clones CH259-366A22 (left) and CH259-290F14 (right) selected in WGAC positive regions. **d.** Global view of SDs in the mapped marmoset genome (WGAC). Blue lines are pairs of duplication within the same chromosome, and red lines are the interchromosomal duplications. Each line is a chromosome, chromosome 1 at top and X and Y at the bottom. Only scaffolded chromosomes data are shown. **e.** As in (**d**) with unplaced scaffolds mapped to each chromosome included at the right. **f.** As in (**d**) with unmapped scaffolds (ChrUn) included at the bottom.



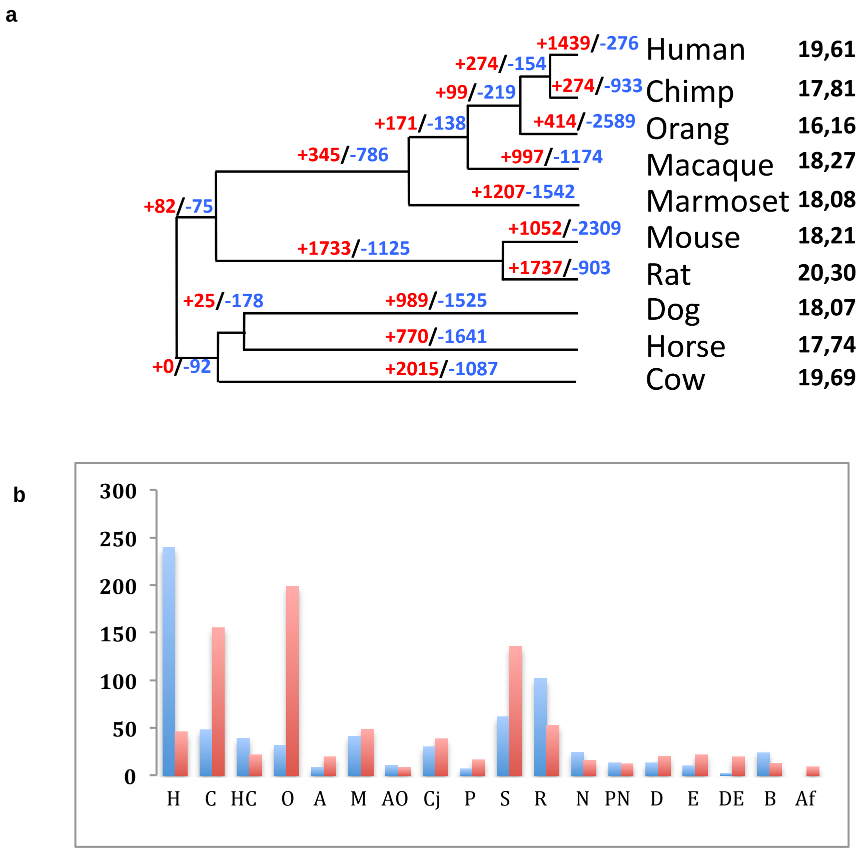
Supplementary Figure 4 | Summary of Structural Variation Data Part Two. **a**. Depth of coverage in 5 kb windows within control regions of single copy sequences in the human genome, shown with identity threshold of 92% (top), 88% (middle) and 85% (bottom), with full distribution on the left and windows with depth of coverage between 0 and 100 on the right. **b.** Shared versus specific duplications between macaque and marmoset genomes. “MMU WSSD against MMU(id=94%)” (orange) and “MMU WSSD against human (id=88%)” (blue) correspond to macaque duplications detected by WSSD using macaque reads aligned with >94% of identity against the human assembly and with >88% of identity against the macaque assembly, respectively. Both datasets are compared with marmoset duplicated sequences determined by WSSD against the human assembly using an identity threshold of 85% and a minimum length of 10 kb “CJA WSSD against human (min 10k)” (green). **c**. Histograms of aligned marmoset read identities between 85% and 100% within marmoset-specific duplications > 20kb (left) and duplications shared (>1 kb) between marmoset and macaque (right). **d.** Identities (94% to 100%) of reads aligned to the assembly in duplications shared with human. **e. through h.** Shared and marmoset-specific duplications detected by three different approaches. “CJA WSSD against CJA” (green) and “CJA WGAC” (orange) correspond to duplications detected in the marmoset assembly by WSSD using marmoset reads aligned with >94% of identity and WGAC respectively. Both datasets are compared with duplicated sequences determined by WSSD using an identity threshold of 85%. **e**. **and f.** Minimum length of 10 kb “CJA WSSD against human (min 10k)” (blue). **g. and h.** Minimum length of 20 kb “CJA WSSD against human (min 20k)” (blue). On the left **(e. and g)** duplicated sequences detected in the human assembly were aligned against the duplications identified on the marmoset assembly. On the right **(f and h)** is the converse, duplicated sequences from marmoset were mapped against duplicated sequences from the human assembly. Notice the consistency of the overlaps.



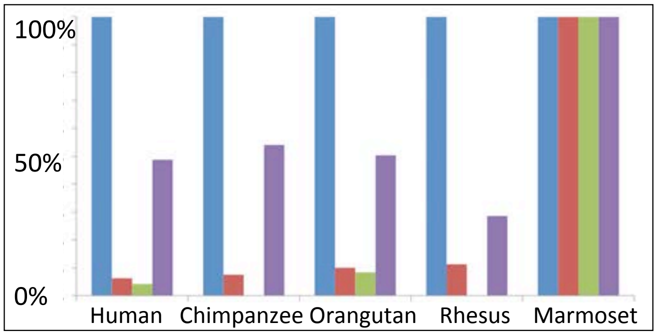
Supplementary Figure 5 | Summary of Anthropoid-Specific Constrained Elements (ASCs). a. Genome browser snapshot of coding anthropoid-specific constrained element. The last exon of PGBD3 is strongly constrained (Pvalue: 8e-15), and derives from a primate-specific insertion of a piggyBac transposon. **b.** Non-coding coding anthropoid-specific constrained element: the gene desert upstream of the brain-specific gene SNTG1 is rich in non-coding ASCs. **c**. Enhancer assays measuring reporter activity of ASCs (red) and their corresponding mouse orthologs (green) in human embryonic stem cells. **d.** Enhancer assays measuring luciferase activity of ASCs (red) and their corresponding mouse orthologs (green) in mouse embryonic stem cells. In (**c**) and (**d**) the fold change values were normalized with respect to the positive controls and elements with fold change of at least 2 (yellow lines) were scored as positive enhancers.



Supplementary Figure 6 | Summary of Gene Annotation and Gene Families. **a**. Summary of Ensembl marmoset gene annotation project. The raw computes 47% repeat masked, 25,582 Uniprot aligned. Proteins aligned from **b.** marmoset and **c**. human generated some models, others were generated with expressed data including **d.** marmoset cDNAs **e.** marmoset ESTs, and **f.** human cDNAs. Evidence for the final gene set **g.** genes and **h.** transcripts. Uniprot protein alignments (99,342 of 112,776) generated 22,897 of 31,863 models with UTRs (not shown).

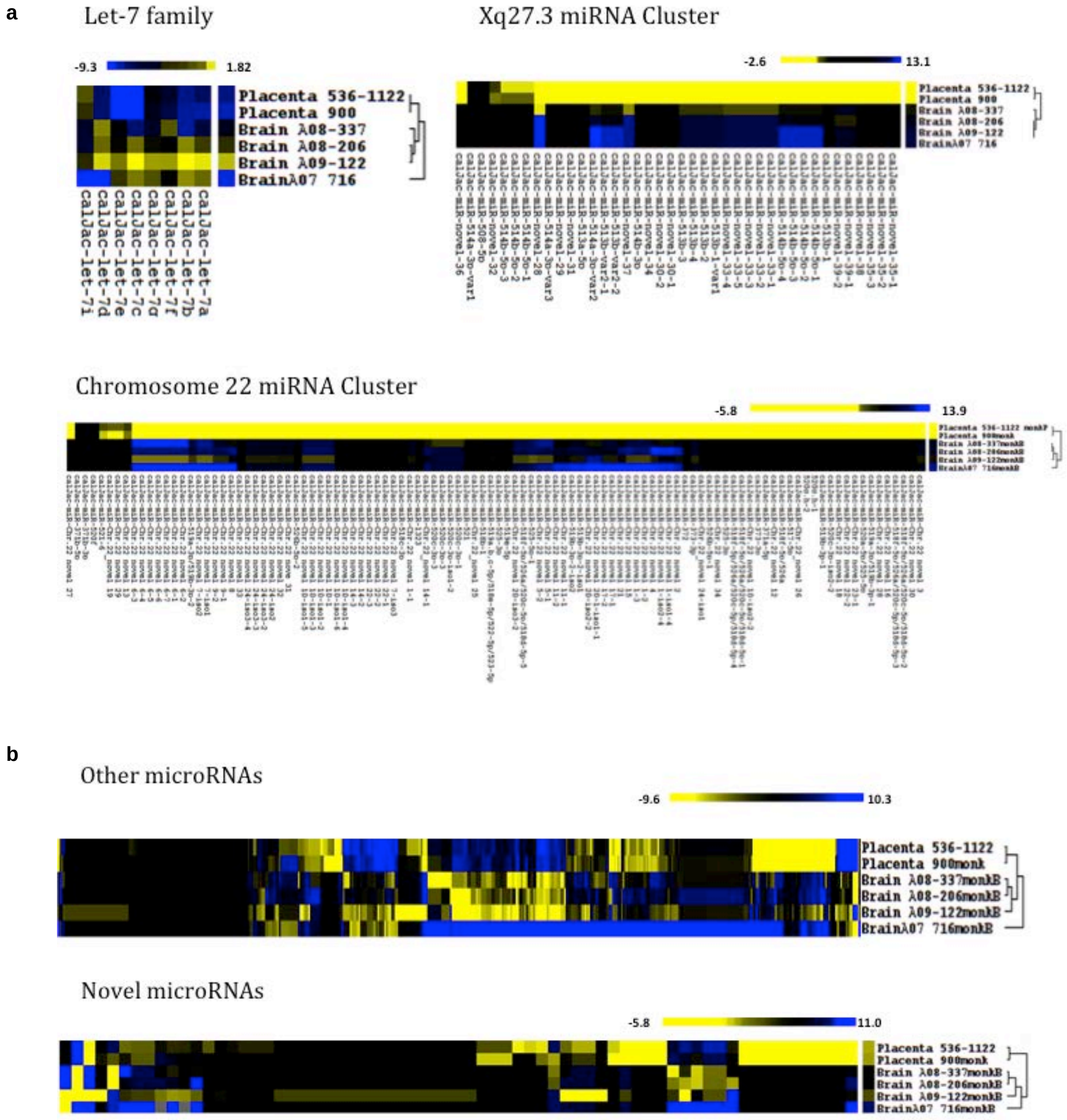


Supplementary Figure 7 | Gene Family Expansion and Contractions. **a**. Inferred gene gains (red) and losses (blue) in 8877 gene families shared by ten mammalian species are reported for each branch of the tree. **b.** Number of gains (red) and losses (blue) per million years for each branch of the tree; human (H), chimpanzee (C), human and chimp (HC), orangutan (O), apes (A), rhesus macaque (M), apes and old world monkeys (AO), marmoset (Cj), primates (P), mouse (S), rat (R), rodents (N), primates and rodents (PN), dog (D), horse (E), dog and horse (DE), cow (B), Afrotheria (Af).

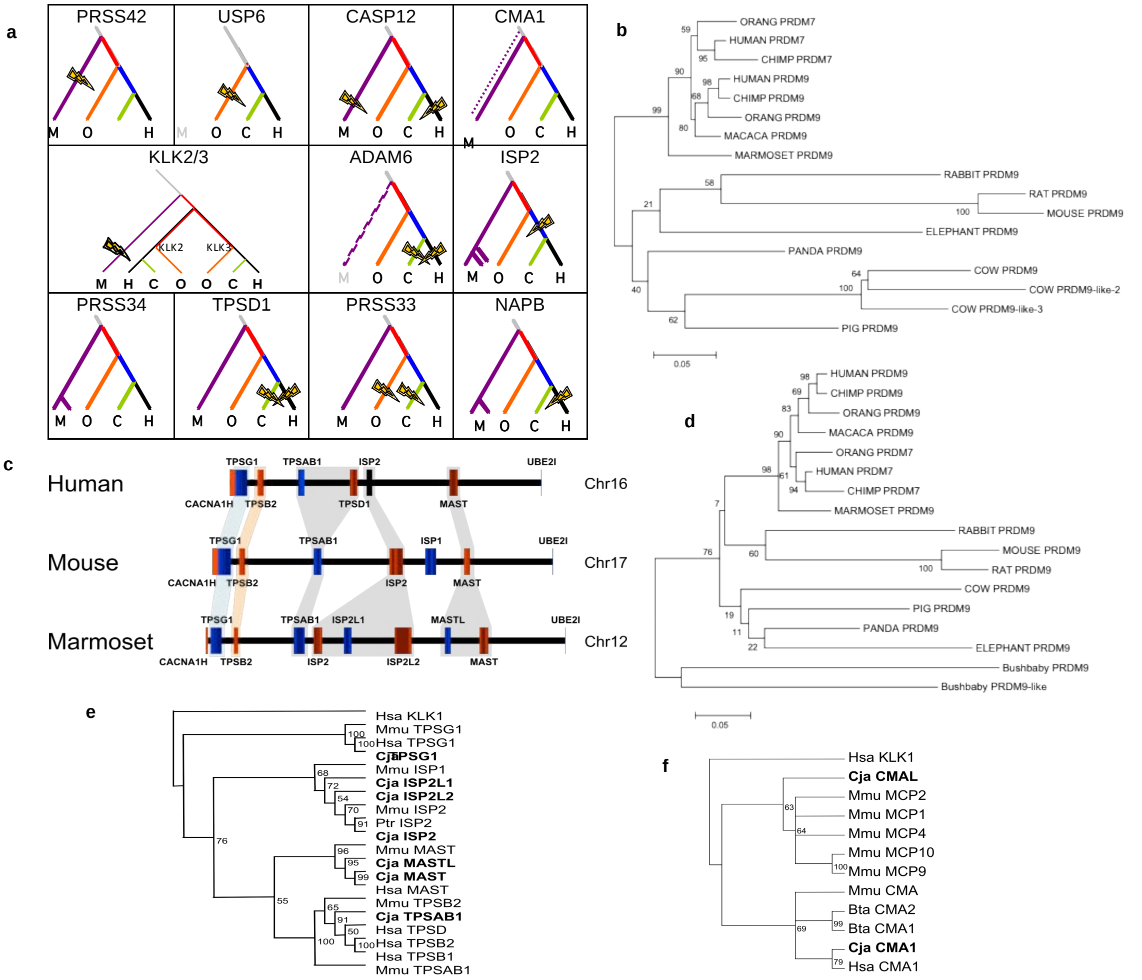


	Key
Let-7 Family	🟦
Chromosome 22 Cluster	🟥
Chromosome X Cluster	🟩
Other known miRNAs	🟪

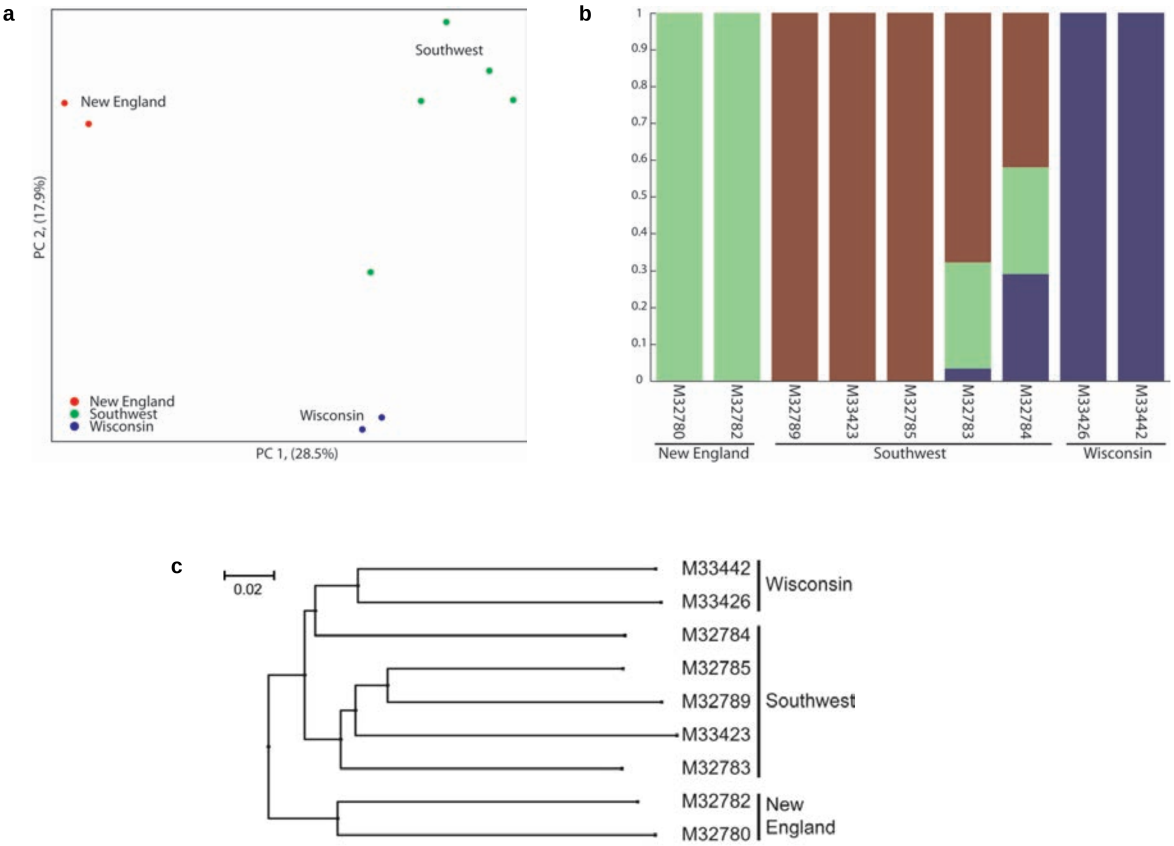
Supplementary Figure 8 | Conservation of miRNAs across four anthropoid primates. Marmoset miRNAs were mapped on the marmoset (calJac3) genome and evaluated against four anthropoid primates including Human (hg18), Chimpanzee (panTro2), Orangutan (ponAbe2) and Rhesus macaque (rhemac2). The fraction of marmoset miRNAs that are perfectly conserved in different primates are plotted. Total number of miRNAs in each group are Let-7 family (8), Chr. 22 cluster (71), Chr. X cluster (22), All others (673).



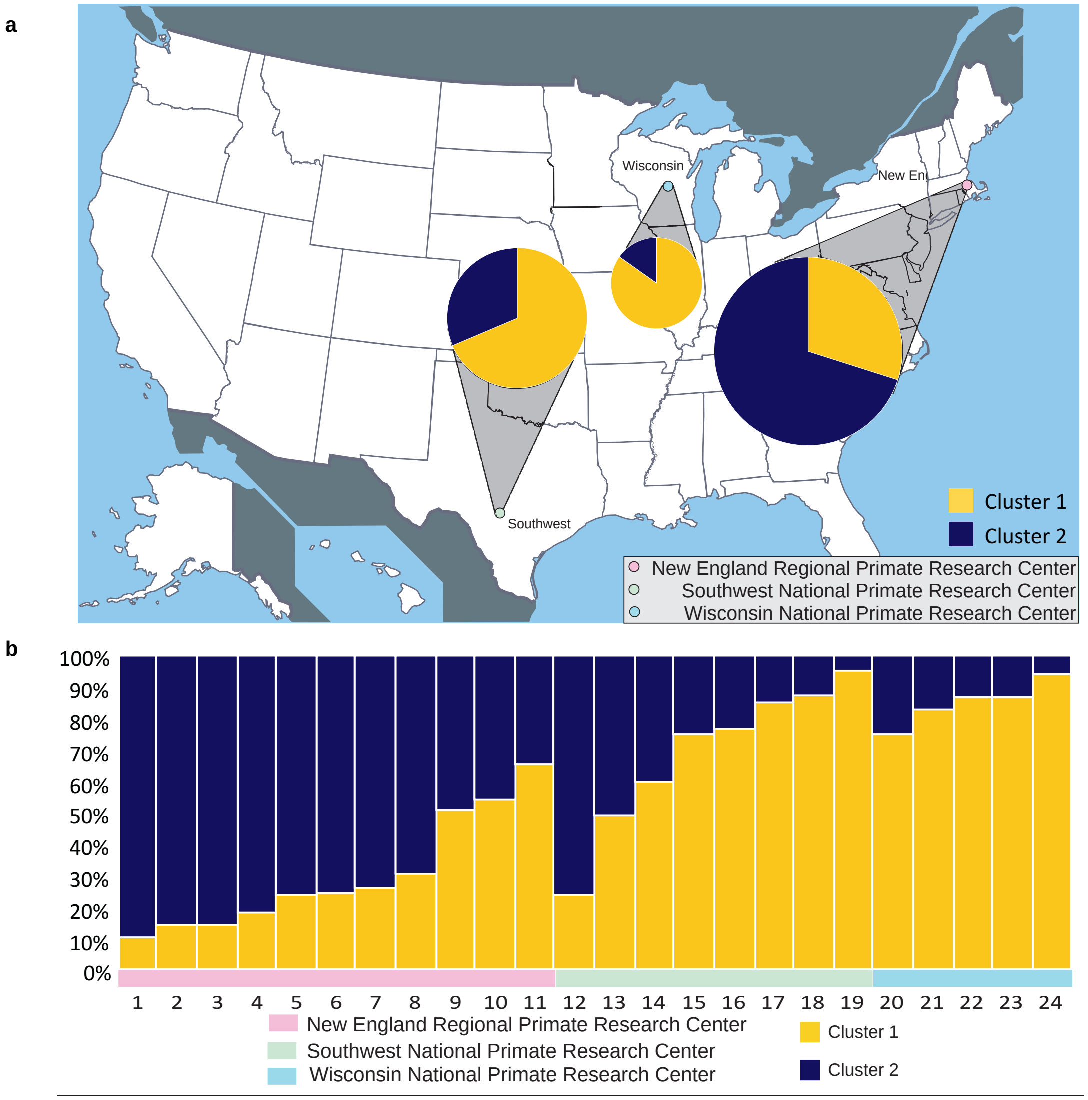
Supplementary Figure 9 | Analysis of miRNAs and their targeted mRNAs. a. Expression in placenta and brain of Let-7, Chr. X, Chr. 22, micro RNAs These heat maps show the relative expression levels of miRNAs in the small RNA fraction isolated from placenta and brain. Brain samples from four different brain samples are on the bottom (A08-337, A08-206, A09-122 and A07-716), samples from two placentas are on the top (S36-1122 and 900). **b**. Expression of novel and all other miRNAs in placenta and brain. This heat map shows the relative expression levels of miRNAs not shown in (**a**): Other microRNAs and microRNAs not found in other species (Novel micro RNAs) samples as in (**a**).



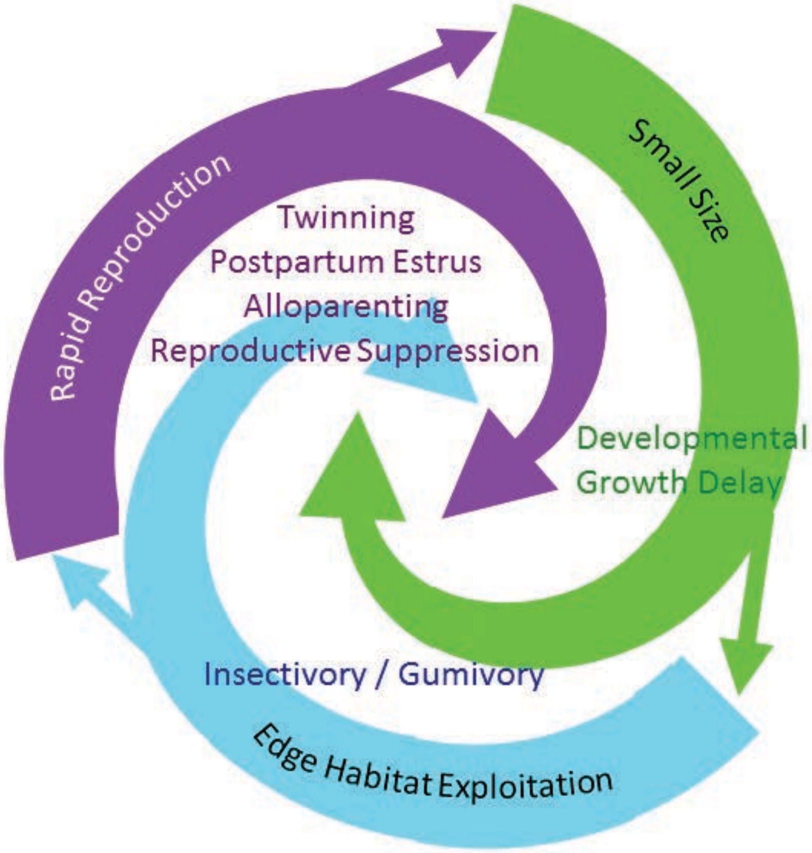
Supplementary Figure 10 | Gene Families – Proteases and PDRM9. **a**. Representative events in the evolution of the primate degradome. Genes are shown on marmoset (M), orangutan (O), chimp (C), and human (H) branches. **b.** Maximum-likelihood tree of PRDM9 and PRDM7 proteins (KRAB and SET domains) from placental mammals. Numbers on nodes represent bootstrap support percentages (1000 replications). **c**. Genomic comparison of tryptases in human, mouse and marmoset.. **d.** Maximum-likelihood tree of PRDM9 and PRDM7 proteins (KRAB and SET domains) from placental mammals, including two sequences from the bushbaby genome. Numbers on nodes represent bootstrap support percentages (1000 replications).**e.** Most parsimonious phylogenetic tree with the tryptases depicted in (**c**). The tree was rooted with human kallikrein 1. Bootstrapping scores are indicated beside each node. Only nodes present in at least half the bootstrapped trees are considered. **f.** Phylogenetic analysis of marmoset, human and murine chymases. Bovine CMA genes were included for comparison. The tree was rooted with human kallikrein 1. Bootstrapping scores are indicated beside each node. Only nodes present in at least half the bootstrapped trees are considered.



Supplementary Figure 11 | Summary of SNP Variation Analyses. **a**. Principal component analysis of SNPs separates 9 animals from the three primate research centers; Southwest (green), New England (red), and Wisconsin (blue). **b.** Admixture among these 9 marmosets in the US. Two animals on left are from New England RPRC, two on right from Wisconsin NPRC and five in middle from Southwest NPRC. **c.** Neighbor joining tree using distance matrix among these 9 US marmosets.



Supplementary Figure 12 | Summary of *Alu* Variation Analyses. a. Pie charts show the overall diversity of the included common marmosets for each Primate Center separately. The size of the pie charts is proportional to the number of individuals included in the analysis. **b**. Shown is the population structure for each common marmoset individually. The numbers below the bars refer to the common marmoset individuals in (**Supplementary Table 41**). Some individuals are primarily assigned to one cluster while others show varying degrees of admixture between the two populations.



Supplementary Figure 13 | Interaction Model. The co-evolution of several unusual traits within marmosets can be explained through a model positing a positive feedback loop triggered initially by external forces of selection. We suggest that early in the evolutionary radiation of New World monkeys, the ancestral callitrichine shifted to an ecological niche consisting of insectivory/gumivory and exploitation of edge habitats. This provided selective advantages to animals that could reproduce rapidly, and hence leave many descendants, whenever new empty habitats were discovered. Selection for rapid reproduction led to twinning and postpartum estrous. Selection for small body size (both in adults and neonates) occurred in parallel with selection for rapid reproduction. These processes led to the origin of callitrichine paternal care and alloparenting, as these behavioral adaptations reduced the energetic demands on breeding females, allowing greater investment in pregnancy and fetal growth. Reproductive suppression of subordinate females facilitated alloparenting and rapid increases in population-level reproductive output as dispersal into newly disturbed edges by subordinates could quickly generate new social groups consisting of reproductively competent individuals that were suppressed in their prior social group. All these processes reinforced this feedback loop, ultimately producing a remarkable suite of behavioral and physiological adaptations that are reflected in the unique molecular traits described in the text.

II. Supplementary Tables 1-41.

Supplementary Table 1 | Reads used as input to whole genome shotgun assembly.

Read Type	Insert Size (kb)	Reads (M)	>Phred 20 Bases (M)	Sequence Coverage	Physical Coverage
Plasmid(WU/BCM)	4	25.4	17.8	6.34	19.0
Fosmid (WU)	40	1	0.7	0.26	7.2
BAC(WU)	170	0.3	0.2	0.07	8.8
Total		26.7	18.7	6.67	35.0

Supplementary Table 2 | Assembly contiguity.

Contiguity	Contig	Scaffold
Number	202,484	16,089
Bases	2,762,785,644	
Q20 bases	2,739,051,791	
Average length (bp)	13,644	171,719
Maximum length (bp)	325,720	39,893,044
N50 length (bp)	29,184	6,770,053
N50 number	27,104	113

Supplementary Table 3 | Assembly statistics by tier.

Tier	Top	Middle	Bottom
	up to 1 Gb	1 to 1.5 Gb	1.5 Gb to end
Contig Number	15,896	15,488	171,100
Average length (bp)	62,910	32,284	7,380
Longest length (bp)	325,720	40,006	26,353
Total bases	1,000,024,638	500,009,227	1,262,751,779
N50 contig length (bp)	62,457	32,582	13,418
N50 contig number	5,732	6,947	33,309
Scaffold Number	66	67	15,956
Average length (bp)	15,222,477	7,475,542	78,794
Longest length (bp)	39,893,044	9,465,787	5,465,816
Total bases	1,004,683,511	500,861,336	1,257,240,797
N50 scaffold length (bp)	15,848,333	7,899,522	1,975,386
N50 scaffold number	25	29	191

Supplementary Table 4 | **ESTdata.** Contigs indicates number of assembled EST contigs longer than 400 bp. Percentages indicate percentage of ESTs aligned to the genome with >=9% or >=20%, >=50%, >=90% of their length aligning.

Tissue	Input Reads	Contigs	Aligned over this % of EST length			
			>=9%	>=20%	>=50%	>=90%
CXAK	269,344	6,652	98.6	89.3	87.9	83.6
brain	531,850	10,880	96.8	82.9	80.7	74.7
kidney	444,008	10,245	97	83.3	81.3	75.5
testis	1,160,754	23,630	70.5	56.2	52.7	45.3
spleen	472,310	10,648	97.3	84.7	82.4	76.8
liver	501,505	8,609	97.3	83.8	81.4	75.2

Supplementary Table 5 | Classes of Chimerias. Classes of chimeras by number of B alleles in each fraternal co-twin, with expected counts (256 total).

Twin 1	Twin 2	Count
0	0	36
0	1	24
0	2	4
1	0	24
1	1	80
1	2	24
2	0	4
2	1	24
2	2	36

Supplementary Table 6 | Estimated chimerism fractions Φ_{max} for nine marmosets. Standard deviation of the proportion of allele A reads (pA) across loci is sd *pA*.

ID	Φ_{max}	# Loci	sd *pA*
32780	0.351	1,628,439	0.188
32782	0.366	1,695,603	0.202
32783	0.355	1,537,826	0.197
32784	0.336	1,368,745	0.201
32785	0.36	1,339,963	0.192
32789	0.192	1,321,960	0.185
33423	0.125	1,208,124	0.118
33426	0.305	1,146,663	0.191
33442	0.286	999,688	0.191

Supplementary Table 7 | WGAC Duplication Analysis. Total and non-redundant (nr) lengths are given.

WGAC Duplications	Marmoset	Human
	(Caljac 3)	(Hg18)
total genome length	2.915Gb	3.108Gb
chrom length	2.770Gb	3.080Gb
number of WGAC pairs	62,863	25,914
number of inter chrom	40,550	15,530
number of intra chrom	22,313	10,384
nr length	138 Mb	159.2 Mb
nr length of inter chrom	97 Mb	74.5 Mb
nr length of intra chrom	75 Mb	114.5 Mb
nr length for chrom only pairs	50 Mb	153.2 Mb

Supplementary Table 8 | Duplication Analysis for Duplications > 10kb.

WGAC-identity	WGAC	WSSD	shared	WGAC only	WSSD only
>=90%	59Mb	71Mb	20Mb	39Mb	51Mb
>=94%	44Mb	71Mb	18Mb	26Mb	53Mb

Supplementary Table 9 | FISH Confirmation of Segmental Duplications. FISH confirmation of segmental duplications, including clones identified as potential to test (Select), analyzed by FISH (Eval.) and confirmed as duplicated (Dup.).

WGAC	WSSD	Select	Eval.	Dup.	Percent Dup.
+	+	43	37	32	86%
-	+	24	18	11	61%
+	-	30	26	22	85%
Total		97	81	65	80%

Supplementary Table 10 | Duplications. Number of Mb. of all, inter- and intra-chromosomal duplications per chromosome. See supplementary data set: SUP_TABLES.xlsx.

Supplementary Table 11 | Tested Anthropoid-Specific Constrained Elements.

	Tested Regions	
	hg19 coordinates	**mm9 coordinates**
Basal	N/A	N/A
ASC587	chr8:30398708-30399596	chr8:34920368-34920954
ASC968	chr22:33458310-33459125	chr10:85963613-85964123
ASC15327	chr5:175030874-175032030	chr13:54260406-54261737
ASC6818	chr9:138231858-138232470	chr2:28228178-28228833
ASC3867	chr10:14226659-14228062	chr2:4063140-4063818
ASC5371	chr20:2215353-2216233	chr2:129783145-129783517
ASC13154	chr15:29741429-29741960	chr7:72190402-72190869
ASC16391	chr5:136426571-136427886	chr13:57638743-57639146
CR2 enhancer	N/A	N/A
CR4 enhancer	N/A	N/A

Supplementary Table 12 | Useable Genes After Each /filtering Step. Species include chimpanzee (chimp), orangutan (orang), rhesus macaque (maq), marmoset (marm), mouse, rat and dog.

TEST	**Chimp**	**Orang**	**Maq**	**Marm**	**Mouse**	**Rat**	**Dog**
Clean in human	21,361	21,361	21,361	21,361	21,361	21,361	21,361
Synteny	19,291	18,520	18,100	17,607	17,142	15,977	17,402
Gaps	19,073	18,509	18,086	17,524	17,141	15,973	17,372
Frameshifts	17,963	16,109	16,660	15,569	15,512	14,257	15,488
Gene structure	17,626	15,665	16,234	15,130	14,761	13,518	14,868
Recent duplications	13,039	11,768	12,295	13,873	11,294	10,224	11,505

Supplementary Table 13 | Gene Turnover in 8,877 Gene Families used in CAFÉ Analysis. Showing number of gene families (Family) and genes in the database (Ensembl58), the analysis, expanded, contracted, and lost.

Species	**Ensembl58**		**Total Analyzed**		**Expanded**		**Contracted**		**Lost**	
	Family	**Genes**	**Family**	**Genes**	**Family**	**Genes**	**Family**	**Genes**	**Family**	**Genes**
Human	9,431	20,705	8,549	19,619	786	1,439	258	276	127	128
Chimp	9,236	18,882	8,366	17,811	227	288	810	933	310	329
Orang	8,877	17,303	7,910	16,161	337	414	2,082	2,589	808	868
Macaque	8,786	19,517	8,224	18,279	673	997	1,057	1,174	550	567
Marmoset	**8,786**	**18,506**	**8,259**	**18,088**	**738**	**1,207**	**1,276**	**1,542**	**541**	**568**
Mouse	8,385	19,007	7,981	18,215	400	1,052	1,757	2,309	673	762
Rat	8,793	21,136	8,348	20,306	1,027	1,737	764	903	306	321
Dog	8,376	18,243	8,232	18,076	613	989	1,133	1,525	438	502
Horse	8,333	18,023	8,197	17,741	419	770	1,333	1,641	473	510
Cow	8,577	20,552	8,346	19,693	1,063	2,015	931	1,087	440	476

Supplementary Table 14 | Gene Turnover in 429 Primate-Specific Gene Families. Showing number gained (+) and lost (-).

Species	Families			Genes		
	Total	+	-	Total	+	-
Human	351	25	89	488	51	95
Chimp	388	17	49	512	33	53
Orang	370	23	74	473	24	83
Macaque	200	29	241	559	307	280
Marmoset	164	6	281	192	6	345

Supplementary Table 15 | Expanded Families. Expanded gene families. See supplementary data set: SUP_TABLES.xlsx.

Supplementary Table 16 | Contracted Families. Contracted gene families. See supplementary data set: SUP_TABLES.xlsx.

Supplementary Table 17 | Validated Expansions. Manually validated gene family expansions in marmoset duplicated regions. See supplementary data set: SUP_TABLES.xlsx.

Supplementary Table 18 | Genes Present in Marmoset and Mouse But Absent in Human. All are single copy in marmoset and mouse except for Olfr550 (two genes in marmoset), Cd24a (three genes in marmoset), Olfr367 (six genes in mouse) and Csn1s2b (two copies in mouse). Branches where the loss occurred are human (H), human and chimp (HC), apes (A), and catarrhines (R).

Gene Symbol	ID	B	Function/expression
Olfr550	2360	H	Odorant receptor
Olfr600	2437	HC	Odorant receptor
Prdxdd1	11400	HC	hypothetical protein LOC67939
Tmem210	13967	HC	Adult male testis cDNA, hypothetical Proline-rich protein
4932418E24Rik	8014	A	hypothetical protein LOC329366
5330437I02Rik	8728	A	hypothetical protein LOC319888
1700013G24Rik	12748	A	hypothetical protein LOC69380
Olfr367	2341	R	Odorant receptor
1100001G20Rik	7942	R	WDNM1-like protein
1700007B14Rik	11275	R	RIKEN cDNA 1700007B14
Cd24a	12564	R	CD24a antigen
Cox8b	12577	R	Cox8b
Csn1s2b	12887	R	casein alpha s2-like A and B

Supplementary Table 19 | Primate-Specific Genes Absent in Marmoset and Single Copy in Human.

Gene	Function/expression
CLECL1 4260	C-type lectin-like, expressed by antigen presenting cells including dendritic cells
DSCR4 16369	Contributes significantly to the pathogenesis of Down syndrome, mainly expressed in placenta
GPR148 2457	G protein-coupled receptor 148
INSL4 8114	Highly expressed in the early placenta. Epil peptides expression in the villous cytotrophoblast differs from syncytiotrophoblast
PRR4 1632	Abundantly expressed in lacrimal gland, found in acinar cells but not in the intralobular ducts

Supplementary Table 20 | Marmoset miRNAs by Category.

MiRNA Category	Group	Mature MiRNA		Hairpin
		#	%	#
100% Conserved 5' seed, 0-3 mismatches, expressed	A	280	36%	406
Novel expressed (1-3 mismatches to miRNAs from other species)	B	61	8%	109
100% Conserved 5' seed, 0-3 mismatches, not expressed	C	108	14%	223
Novel not expressed (1-3 mismatches to other species' miRNAs)	D	66	9%	66
Novel expressed (no matches to miRNAs from other species)	E	246	32%	345
Predictions - machine vector system	F	13	2%	15
Published predictions[135]		0	0%	9
TOTAL miRNAs from Marmoset		**774**	**100%**	**1173**

Supplementary Table 21 | Marmoset miRNAs and Corresponding Hairpins. The example families of Let-7 and the large clusters on Chromosome 22 and the X chromosome are listed individually.

	Mature miRNAs	Hairpins	Key
Let-7 Family	8	10	🟦
Chromosome 22 Cluster	71	112	🟥
Chromosome X Cluster	22	40	🟩
Other known miRNAs	401	671	🟪
Other novel miRNAs	193	259	
Predicted hairpins from SVM	13	15	
Other predictions	66	66	
Total	**774**	**1173**	

Supplementary Table 22 | Marmoset miRNAs. List of all MiRNAs in Marmoset, grouped as in Supplementary Table 20. See supplementary data set: SUP_TABLES.xlsx.

Supplementary Table 23 | Summary. Data for Supplementary Tables 20, 21. See supplementary data set: SUP_TABLES.xlsx.

Supplementary Table 24 | Marmoset_caljac3. Mapping positions of marmoset MiRNAs. See supplementary data set: SUP_TABLES.xlsx.

Supplementary Table 25 | Human_blast. Blast results for marmoset MiRNAs mapped to Human hg18. See supplementary data set: SUP_TABLES.xlsx.

Supplementary Table 26 | Chimp_blast. Blast results for marmoset MiRNAs mapped to Chimp. See supplementary data set: SUP_TABLES.xlsx.

Supplementary Table 27 | Rhesus_blast. Blast results for marmoset MiRNAs mapped to Rhesus. See supplementary data set: SUP_TABLES.xlsx.

Supplementary Table 28 | Orangutan_blast. Blast results for marmoset MiRNAs mapped to Orangutan. See supplementary data set: SUP_TABLES.xlsx.

Supplementary Table 29 | Sequence Changes in Rapidly Evolving miRNA Clusters.

Chromosome	22	X
Total miRNA	71	22
5' seed 100% conserved	17%	22%
1 seed, 1-3 total nt. changes	27%	23%
>3 nt. changes = novel	56%	55%

Supplementary Table 30 | Sequencing of miRNAs.

	Sample Name	Total reads	Usable
B	A07-716monkB	93,712	11,357
B	A09-122monkB	11,664,231	8,446,016
B	A08-206monkB	16,126,169	6,864,450
B	A08-337monkB	7,650,502	4,483,342
P	900monkP	20,051,907	14,079,376
P	536-1122 monkP	21,175,120	13,064,579

Supplementary Table 31 | Placenta. Human MiRNA, Gene ID, Go terms. See supplementary data set: SUP_TABLES.xlsx.

Supplementary Table 32 | Pregnancy. Human MiRNA, Gene ID, Go terms. See supplementary data set: SUP_TABLES.xlsx.

Supplementary Table 33 | Known miRNAs. Gene ID, MiRNA family ID, Species ID, MSA positions, UTR positions, Group number, Site type, etc. See supplementary data set: SUP_TABLES.xlsx.

Supplementary Table 34 | Let7. Data for Let-7 miRNA targets. See supplementary data set: SUP_TABLES.xlsx.

Supplementary Table 35 | ChrX. Data for Chr. X miRNA targets. See supplementary data set: SUP_TABLES.xlsx.

Supplementary Table 36 | Chr22. Data for Chr. 22 miRNA targets. See supplementary data set: SUP_TABLES.xlsx.

Supplementary Table 37 | Gene Counts. Gene count data for miRNAs.. See supplementary data set: SUP_TABLES.xlsx.

Supplementary Table 38 | Positive Selection. Selected genes under positive selection. See supplementary data set: SUP_TABLES.xlsx.

Supplementary Table 39 | Targeted Sequencing. Genes and genomic coordinates for targeted resequencing. See supplementary data set: SUP_TABLES.xlsx.

Supplementary Table 40 | Phylogenetic Panel. Source animals used for the phylogenetic panel. See supplementary data set: SUP_TABLES.xlsx.

Supplementary Table 41 | Common Marmoset Samples in the Population Panel. *Callithrix jacchus* individuals in the population panel shown in Supplementary Figure 12 analyzed using *Alu* insertions.

Sex	Litter	PC	ID	Alternate ID	Other ID
F	NA	N	cj 337-00, A03-168	Cm_NE-10	cj 337-00
F	NA	N	cj 15-98, A02-446	Cm_NE-6	cj 15-98
F	NA	N	cj 65-98, A03-145	Cm_NE-9	cj 65-98
M	NA	N	cj 99-01, A02-485	Cm_NE-3	cj 99-01
M	NA	N	cj 389-97, A02-387	Cm_NE-1	cj 389-97
F	NA	N	cj 12-97, A02-737	Cm_NE-7	cj 12-97
M	NA	N	cj 215-99, A02-677	Cm_NE-5	cj 215-99
F	NA	N	32782	Cm_1-2008	1-2008
M	NA	N	cj 501-98, A02-418	Cm_NE-2	cj 501-98
M	NA	N	cj 72-98, A02-584	Cm_NE-4	cj 72-98
F	NA	N	cj 393-99, A02-738	Cm_NE-8	cj 393-99
F	NA	S	Cm_SW-27537	Cm_SW-27537	27537
M	3 M	S	Cm_SW-19478	Cm_SW-19478	19478
F	NA	S	Cm_SW-27552	Cm_SW-27552	27552
M	2 M 1 F	S	Cm_SW-17461	Cm_SW-17461	17461
M	2 M	S	Cm_SW-17953	Cm_SW-17953	17953
M	3 M	S	M32784	Cm_272	272
M	1 M	S	Cm_SW-19570	Cm_SW-19570	19570
F	2 F	S	Cm_SW-25556	Cm_SW-25556	25556
NA	NA	W	Cm_34040	Cm_34040	CJ0910
NA	NA	W	Cm_34044	Cm_34044	CJ1544
NA	NA	W	33426	Cm_CJ1590	CJ1590
NA	NA	W	Cm_34042	Cm_34042	CJ1660
NA	NA	W	Cm_34043	Cm_34043	CJ1580

III. Supplementary Note.

1. New World monkey (NWM) phylogeny.

Marmosets are New World, or neotropical, primates. An excellent overview of neotropical primate phylogeny and population genetics is provided by Moreira and coauthors[1]. The neotropical primates are proposed to have diverged from the ancestral anthropoid lineage during the Eocene - 38-47 million years ago[2-4]. The question of how these ancestral neotropical primates migrated to the Americas from Africa is unanswered. Initial taxonomies based upon morphology grouped neotropical primates into two families – the Cebidae and the Callitrichidae. The Callitrichidae included the marmosets and tamarins that all share a number of unusual traits, including very small body size; presence of claws instead of nails; production of litters; and the absence of a third molar. Initially these traits were thought to be primitive but later cladistic analyses of morphological characteristics indicated these traits were derived. Analyses of molecular data from a variety of nuclear and mitochondrial genes have generated a revised taxonomy, with three neotropical families: the Atelidae, the Pitheciidae and the Cebidae. In this new taxonomy, the marmosets and tamarins are grouped into a subfamily – Callitrichinae - within the family Cebidae, with that family also including owl monkeys (*Aotus*), cebus monkeys (*Cebus*) and squirrel monkeys (*Saimiri*). The disparate results obtained by analyses of different genes suggest a rapid diversification of the neotropical primates, with the three lineages diverging around 20-26 million years ago. Within the callitrichine primates, the molecular data support the contention that the smallest members of this lineage – i.e. the marmosets – are more derived than the larger tamarins[5].

2. Assembly Quality Assessment.

Comparisons of the initial assembly with 81 finished BAC sequences from the CHORI-259 BAC library indicated high structural accuracy. The BAC library is from Animal #252/17081, the full brother of the source of the WGS data (Animal #186/17066). Some small supercontigs (most <5 kb) were not positioned within larger supercontigs (<1 event per 500kb). While not strictly errors, these do affect overall assembly statistics. There are also small, undetected overlaps (most <1 kb) between consecutive contigs (~1 event per 100 kb), occasional local mis-ordering of small contigs (~1 events per 2 Mb), and small contigs incorrectly inserted within larger supercontigs (<1 event per 100 kb). Overall, the rate of rearrangements with respect to finished BACs was comparable to previous WGS assemblies. Nucleotide-level accuracy is high by several measures. About 98% of the consensus bases in the marmoset sequence have quality scores[53] of at least Q40 corresponding to an error rate of $<=10^{-4}$. Comparison of the WGS sequence to the finished BACs is consistent with this estimate, giving a high quality discrepancy rate of $3x10^{-4}$ substitutions and $2x10^{-4}$ indels (no more than expected given the heterozygosity rate, as 75% of the polymorphic alleles in the WGS sequence assembly will differ from the single-haplotype BACs). Restricting analysis to high-quality bases, the nucleotide-level accuracy of the WGS assembly is essentially equal to that of "finished" sequence.

As with the chimpanzee and other whole genome shotgun-based assemblies, the most problematic regions are those containing segmental duplications[12].

We examined the coverage of other primate genomes. Ignoring chromosome Y, 88% of the human genome is covered by a chained BLASTZ alignment with the marmoset genome. Similarly, 88% of the chimpanzee and 87% of the orangutan genomes are covered by an alignment with marmoset. Conversely, 98.5% of the marmoset genome is covered by a chained BLASTZ[39] alignment with the human, chimpanzee and orangutan genomes.

3. Marmoset Karyotype.

NWM have on average more rearrangements relative to the common primate ancestor than do humans[7]. Marmoset, however, has a relatively well-conserved chromosome complement. The following marmoset (CJA) chromosomes correspond to single human chromosomes (HSA): CJA3 (HSA4), CJA4 (HSA6), CJA9 (HSA12), CJA11 (HSA11), CJA22 (HSA19). But four of these five chromosome pairs contain inversions that occurred in either the lineage leading to human or to marmoset. Only the chromosome pair CJA22-HSA19 retains a marker order, including the centromere, which is perfectly conserved. Interestingly, all four of the mentioned pairs also differ in the position of the centromere. From previous studies we know that HSA6 and HSA11 harbor repositioned centromeres that were seeded in the human lineage after the Catarrhini/Platyrrhini split[8,9], but the available data do not allow us to define with certainty which marmoset centromeres are conserved and which were repositioned during evolution. If marmoset chromosomes are not compared to human, but to the primate ancestor, as defined[7], then CJA14 (HSA2p) also has to be considered "conserved".

It is also worth noting the behavior of the centromeres of marmoset chromosomes CJA15, CJA17, and CJA21. They resulted from the fissions, which occurred in the platyrrhine ancestor, of the ancestral association composed of human HSA3 and HSA21[10]. The marker order of these chromosomes is perfectly conserved in *Lagothrix lagothricha* (woolly monkey, Atelidae) and *Callicebus pallescens* (white-coated titi, Pitheciidae). However, in each of these three chromosomes the centromere appears to be located at one telomere in one species, and at the opposite telomere in the other two species, indicating centromere exchanges between telomeres.

4. Chimerism of reads and assembly

Reports from the published literature indicate that the amount of DNA in circulating blood that is chimeric, or derived from the littermate, varies from individual to individual. To assess the proportion of the DNA in the female reference sample that was from the male twin, we used two different methods. The first method assessed the amount of Y chromosome DNA in the reference sequence and the second used the sequence reads to assess the SNP alleles at non-homozygous SNPs in the genome.

4.1 Estimation of fraction of marmoset WGS sequence derived from male DNA

To calculate the fraction of marmoset WGS sequence derived from male DNA using the Y chromosome, the following variables are used:
Variables:
L = read length
N = total number of reads in assembly
F = size of female genome
M = size of male genome
(F =~ M)
β = fraction of reads in assembly that are male
X = sequence coverage
M_Y = size of Y

Calculated from female genome (assuming 6x coverage of WGS):

$$X = LN/F = 6$$
$$N = 6F/L$$

Number of male reads is βN
Number of female reads is $(1-\beta)N$

Actual coverage for male genome (X_M) is:

$$(\beta LN)/F = \beta(L/F)(6F/L) = 6\beta = X_M$$

Coverage for the Y is:

$$0.5X_M = 3\beta$$

From our BLAST search of the WGS using a 780 kb contig of single-copy Y sequence as a query, we estimate that 26% of the Y is represented in the WGS sequence. Therefore, 74% of the Y is not represented.

The probability not to be represented is taken from the Lander-Waterman Model[11]:

$$(1 – (L/M_Y))\hat{}N_Y =\sim e\hat{}-((LN_Y)/M_Y) = e\hat{}-3\beta = 0.74$$

Therefore,

$$\beta = (-\ln(0.74))/3 = 0.1$$

10% of WGS is male-derived

4.2 Estimation of chimerism in blood samples based upon SNP analysis

Since many marmosets are hematopoetic chimeras, blood samples will often contain DNA from two fraternal twins. Sequencing reads from the sample that cover diallelic SNP loci will show the two alleles present in the genotypes of both twins. The alleles of the twin whose DNA constitutes the smaller fraction of the sample will be less common in sequencing reads than the alleles of the other twin. As an example, consider a non-chimeric diploid individual with genotype AG at a locus. Half of the sequencing reads covering that locus are expected to carry the A allele, half the G allele. By contrast, in a chimera with twin genotypes AA and GG, in which the AA-bearing twin constitutes 20% of the sample, we expect 20% of reads to contain the A allele and 80% to contain the G allele. When taken across millions of SNP loci, such differences in the expected distributions of read numbers allow us to estimate the chimerism fraction of a marmoset blood DNA sample from sequencing reads.

Let Φ be the fraction of the chimeric sample from the twin that contributes the smaller amount of DNA (the ‘minor’ twin). We wish to estimate Φ by the method of maximum likelihood, so we build a probabilistic model of the data as follows.

At any SNP locus, let the number of reads for allele A (arbitrarily chosen) be distributed binomially with parameters *n* equal to the number of total reads covering the locus and *p* equal to the expected proportion of allele A reads. Let m_1 and m_2 be the numbers of A alleles in the genotypes of the first (minor) and the second twin. Possible values of m_1 and m_2 are 0, 1 and 2. Then

E1. $p(\Phi, m_1, m_2) = \Phi m_1 + (1- \Phi)m_2$

and the probability of the observed number of allele A reads, **Z** is

E2. $Pr(\mathbf{Z} \mid \Phi, m_1, m_2) = Bin(\mathbf{Z}, n, p(\Phi, m_1, m_2))/ C$

where *Bin()* is the binomial probability distribution function and $C = 1 - Bin\ (\mathbf{Z}, 0, p(\Phi, m_1, m_2)) – Bin\ (n, n, p(\Phi, m_1, m_2))$. Since we only consider loci with reads representing two alleles, we divide by *C* to model a truncated binomial distribution that excludes $\mathbf{Z} = 0$ and $\mathbf{Z} = n$.

The true but unknown genotypes of the chimeric co-twins enter through the variables m_1 and m_2. In order to compute a likelihood that depends only on Φ, we eliminate m_1 and m_2 by summing over all possible genotype combinations according to their expected frequencies:

E3. $L(\mathbf{Z} \mid \Phi) = \sum_{\text{all m1,m2}} Pr(\mathbf{Z} \mid \Phi, m_1, m_2)\ Pr(m_1, m_2)$

$Pr(m_1,m_2)$ are the frequencies of the 9 distinct combinations of values of m_1 and m_2 that are expected to appear in fraternal twins from unrelated parents, assuming the two alleles are equally frequent. To compute these frequencies, we proceed as follows: (1) label the four parental alleles 1 through 4; (2) construct the 4 possible offspring genotypes using those labeled alleles; (3) list the 16 possible chimeric numeric genotypes that could arise from pairings of those offspring; (4) list the 16 possible assignments of allelic states A or B to each of the four numbered parental alleles; (5) construct the 256 possible chimeric genotypes that result from combining the 16 numeric chimera genotypes with the 16 A/B state assignments; and finally, (6) condense and count the A/B-specified chimera genotypes into classes according to the number of B alleles in each co-twin. The doubly-homozygous genotype classes are ignored, since they cannot yield variable sequence reads. The classes of chimera are shown in Supplementary Table 5.

To find the maximum-likelihood estimate of Φ given the data $\mathbf{Z}$, we sum logarithms of $L(\mathbf{Z} \mid \Phi)$ computed for each locus and find the maximizing value Φ_{max} by numerical optimization (golden section and parabolic interpolation, Matlab 2011a).

E4. $\Phi_{max} = \arg\max_{0.5 \geq \Phi_{max} \geq 0} \sum_{\text{all loci}} ln(L(\mathbf{Z} \mid \Phi))$

Simulation results suggest this estimation procedure works well with average coverage of 24 reads per locus, (typical of the marmoset diversity panel data), using fewer than 100,000 loci.

We used only reads from reliable polymorphic autosomal loci to estimate chimerism fractions Φ_{max} for 9 marmosets. We used only loci at which both alleles were observed in at least two individuals, with minimum SNP quality of 50, minimum read mapping quality of 50, and a minimum of 5 and maximum of 60 reads mapping to the locus in any individual. The results are shown in Supplementary Table 6. Similar results were obtained without the constraint on numbers of reads and when using only loci with at least 24 reads.

Chimerism causes the variance of the proportion of allele A reads out of the total number of reads at a locus (*pA*) to increase relative to what would be expected for a pure diploid individual. Supplementary Figure 2a demonstrates this by comparing simulated distributions of *pA* for a marmoset with chimeric fraction $F = 0.33$ (green) and $F = 0$ (i.e., not chimeric; yellow). The distribution of *pA* for the chimeric individual is wider than the distribution for the non-chimeric individual. This pattern is apparent in Supplementary Table 6, where the standard deviation of *pA* is correlated with F_{max}. Supplementary Figure 2b shows the relationship for $F = 0$ to 0.5, using simulations of 50,000 loci with an average of 24 reads per locus for each point. The expected standard deviation of *pA* for a non-chimeric individual is 0.109 (given an expected 24 reads per locus); for chimerism fractions greater than 0.05, the standard deviation is more than twice that. This excess variance is diagnostic of chimerism.

Simulated counts of allele A and allele B sequencing reads in co-twins and co-triplets of varying composition were generated as follows. For each SNP locus (modeled independently), a minor allele frequency was randomly drawn from a beta distribution and the alleles in the two diploid parental genotypes were generated based on that frequency. To model the ascertainment of SNP loci in the data, loci were accepted according to the probability of observing at least two copies of the least-frequent allele in a sample of 16 alleles. Diploid co-twin or co-triplet genotypes were then randomly simulated based on the parental genotypes. The number of A and B allele read counts were then simulated as binomial variants based on the desired chimerism fraction *f* and the desired expected number of total reads. As with the actual data, SNP loci with reads from only one of the alleles were then dropped.

Marmosets frequently give birth to litters of triplets. We extended the above method to the triplet case, counting the additional chimeric genotype classes and using constrained two-dimensional numerical optimization to estimate the chimerism fractions for the lowest-contributing and next-lowest-contributing fraternal sibling. However, the shape of the likelihood surfaces for data simulated under various combinations of chimerism fractions suggests that the added degrees of freedom prevent accurate estimation of these chimerism fractions in this framework. For samples from triplet litters, the estimation method for twins yields high chimerism fractions, but that simply indicates the presence of substantial chimerism.

5. Segmental duplication analyses

5.1 Duplications on the marmoset genome as a quality control check

We estimated the genome duplications in the marmoset genome using two computational methods. With the first method, WGAC, we find 138 Mb (non-redundant basepairs or 4.7% of the whole genome) as "duplicated" between and within chromosomes. These genomes[M10, M11, M12, M13] used Sanger capillary sequencing technology, so the amount of duplication detected is better than current nextgen human assemblies (< 1%,[12]).

Based on the marmoset genome assembly, we classify most of the duplications as interchromosomal (65%, 22,313 intra and 40,550 inter) (Supplementary Table 7). Without duplications identified in ChrUn, we observe similar proportions, 71% being interchromosomal (7,282 intra and 18,197 inter). Excluding duplications detected in ChrUn, we show an unequal distribution of intra-chromosomal duplicated content along chromosomes. Marmoset chromosomes 1, 18, 22 and Y show an enrichment of intra-chromosomal duplications (>2-fold enrichment compared to the expected duplication content based on the proportion of the genome mapped to each chromosome, assuming duplicated sequence is randomly distributed along the genome excluding ChrUn), whereas chromosomes 3, 8 and 17 (<0.5-fold) appear to be reduced in the fraction of duplication content. For inter-chromosomal duplications, apart from ChrUn, which includes 56% of the inter-chromosomal duplications, we show an excess of duplications in chromosomes 18, 22 and Y (Supplementary Table 10). It is worth mentioning that compared to the distribution of duplications in the human genome assembly, we observe a depletion of recent duplications and an excess of short duplications (Supplementary Figures 3a, 3d, 3e, 3f).

Next we assessed the segmental duplication content by measuring excess read-depth against the marmoset genome, (whole genome shotgun sequence detection, WSSD). WSSD predicts a total of 71 Mb of duplicated sequences (>94%; >10 Kb ad 32.5 > 94% and > 20 Kb) which is less than what was previously reported for the chimpanzee genome (70.59 Mb, >94%; >20 Kb[M11]).

We compared WGAC and WSSD estimates by focusing on SDs > 10 Kb and >94% sequence identity (Supplementary Table 8). The resulting intersection is a good measure of the quality of an assembly and the ability to detect artifact duplications (WGAC positive and WSSD negative) and general collapse (WSSD positive and WGAC negative). In this case, only 18 Mb of duplications were common to both, while 26 Mb were predicted by WGAC only (possible artifact duplications) and 53 Mb were predicted by WSSD methods alone (possible collapse).

Conclusions from the segmental duplication analysis with respect to quality control of the genome assembly. We see an increase of both (a) potential artifact duplication and (b) a general collapse of duplications in the marmoset assembly. It is unclear whether, in this case, this is due to chimerism in the individual sequenced (around 10%). In any case we confirmed that WGAC and WSSD both successfully identify segmentally duplicated regions and unlike in previous studies with other species, WGAC seems better suited than WSSD to detect duplication in marmoset.

5.2 Comparison with great-ape evolution of segmental duplications

The difficulties in non-primate assembly construction impede, in general, a cross-comparison of duplications based on the assemblies[M76]. To overcome this problem, we previously mapped all human and non-human ape whole genome shotgun (WGS) reads to the human reference genome to reconstruct the evolution of great-ape segmental duplications[M77].

In order to evaluate whether we could apply a similar method with marmoset WGS, we aligned reads from marmoset (library with 27,615,086 reads) against a RepeatMasked version of the human assembly (build35, excluding random sequences). Repeat content with less than 20% sequence divergence from their consensus sequences (determined by RepeatMasker either in the human or the marmoset assemblies) was masked. We selected aligned reads with the following criteria: >200bp of high quality (Phred score threshold of 27), >300bp of alignment length, >0.4 alignment length relative to the read length, <200bp of repeat content.

First, we wanted to determine an appropriate sequence identity threshold for the marmoset alignments to detect duplications. Based on neutral estimates of sequence divergence, all duplications that arose since the divergence of OWM and hominid lineages should be captured using a 94% threshold[M59]. This threshold was applied for comparing human, chimpanzee and orangutan duplication content[M77]. For rhesus macaque, we previously lowered the threshold to 88% taking into account the greater evolutionary distance between human and macaque (93.54% identity in aligned nucleotides in non-repeated unique sequence) and the accelerated rates of single basepair substitution in duplicated sequences.

For the marmoset analysis we initially allowed alignments with >82% of sequence identity. (Supplementary Figure 3b shows the frequency of the identity of mapped reads both in a control set of unique regions of the genome (regions without a known CNV in any previously analyzed primate) and the whole genome. Most of the alignments were found around 91% and almost none was lower than 85%.

We then checked the depth of coverage of 5 Kb length windows within the control regions of unique human sequence, considering reads at different thresholds of identity (85%, 88% and 92%) (Supplementary Figures 4a). Lowering the identity threshold, the distribution becomes more normalized. Moreover, with lower identity we find fewer windows without any coverage. However, there are some windows with high coverage (around 400 reads for 92%, 20,000 reads for 88% and 60,000 for 85%), most likely due to ancient repeat sequences.

All together, we applied an identity threshold of 85% that would be similar to the criteria applied in the macaque analysis.

We chose to determine duplicated regions by identifying reads that mapped with less than 85% identity and showing an excess of depth of coverage (more than the mean plus 3 standard deviations) relative to the distribution in these control regions and excluding the outliers. Using these criteria, 78.74Mb of duplicated sequence belonging to blocks of length greater than 10 kb (CJA WSSD against human (min10k)), of which 49.27 Mb correspond to blocks >20 kb (CJA WSSD against human (min20k)).

Comparison of macaque and marmoset duplications

We wanted to determine the marmoset-specific duplication content after the separation from macaque (Supplementary Figures 4b). Two datasets of macaque duplications were considered: one obtained by mapping macaque reads against the human assembly with more than 88% sequence identity within duplications (MMU WSSD against human); and another obtained after lifting over the duplications identified mapping the reads on the macaque assembly, a self-self WSSD based approach (MMU WSSD against MMU)[32]. Surprisingly, only 5 Mb are shared (out of the total 78.74 Mb of marmoset, and 41.15 Mb of macaque – adding all duplicated sequences detected by WSSD against both human and macaque assemblies). Approximately half of it (2.27Mb) corresponds to duplicated blocks longer than 20 Kb.

In an effort to elucidate the antiquity of shared versus marmoset-specific duplications we plotted the distributions of the alignment similarity of marmoset reads against the human assembly for both, duplicated sequences shared with macaque and unique to marmoset (Supplementary Figures 4c, 4d). Duplicated regions that are common in marmoset and macaque have marmoset divergence around 88% (> 20 Kb). On the other hand, marmoset-specific duplications that are longer than 20 Kb have a median of 90%, implying a younger creation.

Comparison with marmoset duplications detected in the assembly

Finally, we wanted to compare whether the duplications detected in the marmoset assembly (more likely to be real) can be detected with this approach. For this we considered duplications detected on the marmoset assembly by a self-self WSSD approach with 94% identity threshold (CJA WSSD against CJA) and by WGAC approach (CJA WGAC).

To count the overlapping bases as duplicated blocks in the different assemblies, we aligned the duplicated sequences from the marmoset assembly to the duplicated sequences determined in the human assembly using BLAST (with >85% of identity). Then, for each duplicated block we merged sequences that map in the other set of duplicated sequences and are less than 1 kb from each other and count the number of bases, without redundancy. For each duplication, overlaps are considered if the length of the overlapping sequence is >1 kb. Moreover, we also did the reverse analysis, which is that we aligned duplicated marmoset sequences detected on the human assembly to the duplications from the marmoset assembly.

Supplementary Figures 4e, 4f, 4g, 4h show the number of base pairs overlapping on duplications detected by the three different procedures. The main result is that the majority of the duplications are only detectable by just one method, exposing the difficulty of unraveling segmental duplication beyond apes with this approach.

Considering duplications detected by WSSD against both marmoset and human assemblies, we focused on identifying the ones that are marmoset-specific. 25.89 Mb overlap when considering both directions of Blast alignment between CJA WSSD against human (>10 kb) and CJA WSSD against CJA (Supplementary Figures 4e, 4f, 4g, 4h). From these blocks, we excluded overlaps with any other duplications detected by WSSD against the human assembly from human (Celera and NA18507), chimpanzee, gorilla (Kwan), orangutan and macaque reads. For macaque we also excluded duplications detected by WSSD against macaque assembly, which were lifted over to the human assembly. Surprisingly, 23.5 Mb still remained (in blocks >1 kb), which would correspond to marmoset-specific duplications. Almost all reads within these blocks that aligned to the marmoset assembly are >99.5% similar to the assembly (Supplementary Figure 4d). Therefore, this data suggest that we can only determine duplications on the human assembly that correspond to the youngest duplication events in the marmoset branch, and we are missing most of the old and shared duplications between human and NWM.

In conclusion, the great evolutionary distance between human and marmoset complicates the detection of marmoset segmental duplications mapping the reads on the human genome. The great amount of differences based on comparing the duplications in the marmoset genome to the duplications detected mapping marmoset reads on the human genome precludes any further analyses.

6. Sequence elements constrained in anthropoid primates

Functional DNA sequences tend to evolve more slowly than surrounding neutral regions, so long as their function is preserved by natural selection. Constrained sequence evolution is therefore a vital and unbiased indicator of function in unannotated regions of the human genome. So far, this comparative genomics strategy has only been used on a genome-wide scale to detect relatively ancient, broadly shared functional elements. With the sequencing and assembly of the marmoset genome, however, it is now possible to systematically detect functional elements shared only among simian primates[13, M15, M16]. In order to identify candidate genomic loci that potentially drove the evolution of unique primate phenotypes, we designed a whole-genome screen for primate-specific sequence constraint. The strategy was to identify DNA elements that showed strong sequence constraint in multiply aligned anthropoid primate genomes, but little or no constraint at the orthologous locus in non-primate mammals. Full details of the method are given in[M17]. We defined anthropoid-specific constrained (ASC) elements as sequences that were strongly constrained among anthropoids but with no more than 10% intersection (of its length) with elements loosely constrained in non-primate mammals. The ASC set contains 23,849 sequences (9% of all anthropoid-constrained elements)[M17]. Below, we highlight two exceptional loci, one marked by strong anthropoid-specific constraint in a protein-coding exon, and the other by an excess of non-exonic gain of function in the ancestral lineage of anthropoid primates.

A top-ranked coding ASC (Pvalue: 8e-15, genome-wide rank: 38) lies in the last exon of the piggyBac transposable element derived gene, *PGBD3* (Supplementary Figure 5a). *PGBD3* shares its first 5 exons with the CSB gene (Cockayne syndrome group B, also known as *ERCC6*) and terminates in an exon derived from a primate-specific piggyBac transposon insertion[14]. Mutations in *CSB* cause Cockayne syndrome B, which is characterized by post-natal growth failure and progressive neurological dysfunction[15]. The alternatively spliced transcript yields a fusion protein consisting of exons 1-5 of *CSB* joined in frame to the piggyBac transposase. The strong constraint in the transposase sequence has previously been noted, and presumably indicates gain of a new primate-specific function[16].

We performed a gene enrichment analysis on non-coding ASCs by assigning each of them to the nearest gene (5' end). Strikingly, the brain-specific gene *SNTG1* (syntrophin, gamma1) was strongly enriched in ASCs constraint in its upstream gene desert; 40% of anthropoid-constrained elements in this region were represented in the anthropoid-specific list (17/42, Fisher's exact test P-value: 4e-8). *SNTG1* is strongly expressed in the hippocampus, dentate granule cells and pyramidal cell layers, and is highly expressed in neurons of the cerebral cortex[17,18]. The ASCs upstream of *SNTG1* (Supplementary Figure 5b) are potential cis-regulatory regions that may confer unique lineage-specific expression patterns upon this neuronal gene.

Of the eight elements, six exhibited robust enhancer activity in hESCs (Supplementary Figure 5c and Table 11) whereas their mouse orthologs showed zero or low functional activity, with one exception (ASC15327). To address the issue of a bias possibly arising from testing the activity of the mouse elements in a heterologous host system, we tested the same elements in mES cells (Supplementary Figure 5d and Table 11). The human and the mouse elements both showed a lower activity overall in the mouse system, perhaps reflecting developmental stage differences between human and mouse ESCs[19,20]. However even in this assay the ASCs were more likely to function as enhancers than their mouse orthologs (3/8 vs. 0/8). These results further confirm that ASCs can act as newly-evolved primate-specific enhancers[13, M15, M16, M17].

7. Gene and feature annotation

7.1 Mobile elements

The overall repeat composition of the marmoset genome [assembly CalJac 3.2] is similar to other sequenced and analyzed primate genomes (human[M10], chimpanzee[M12], rhesus macaque[M14], orangutan[M13]). As expected, we did not find evidence of DNA transposon mobilization. Altogether, non-LTR (long terminal repeat) retrotransposons – that propagate in genomes through a copy and paste mechanism – are the major contributors to genomic variation caused by insertional mutagenesis.

L1, a long interspersed element (LINE), and *Alu*, a short interspersed element (SINE), are the two retrotransposon families that have been expanding throughout primate radiation[21-23]. Both families have been the major drivers of repeat expansion in the lineage leading to the common marmoset. As seen in other primate lineages, L1 and *Alu* evolved uniquely in the NWM lineage. A full-length L1 is about 6 kb in length and contains two open reading frames (ORF1 and ORF2). L1s are considered autonomous elements as they provide their own enzymatic machinery for retrotransposition[22]. However, the majority of L1 insertions are not capable of retrotransposition due to truncation and/or inactivating point mutations. In contrast, *Alu* elements are about 300 bp in length and do not contain coding sequence. They rely on the enzymatic machinery of L1 for their retrotransposition[24].

Omitting information regarding the length of *Alu* elements and not accounting for the presence of interrupted repeats, the marmoset genome contains about 1.1 million *Alu* elements. Roughly 660,000 of these are full-length (with start position within the first 4 bp, and end position 269 bp or larger of the *Alu* consensus sequence). We reconstructed the *Alu* subfamily evolution in the NWM-lineage leading to the common marmoset. As seen in other primate genomes, *Alu* subfamilies evolved along several branches with several *Alu* subfamilies being active in parallel. The youngest *Alu* subfamilies seem to be derivatives of *Alu*Ta15, a previously identified NWM-specific subfamily[25]. Moreover, about half of the NWM-specific *Alu* elements appear to be derived from *Alu*Ta15 or its derivatives. Along with the rise of *Alu*Ta15, the NWM lineage experienced a small burst of *Alu* retrotransposition. Intriguingly, *Alu*Ta15 has an unconventional origin in that it descended from an *Alu* subfamily generated by a gene conversion event between two ancestral *Alu*S subfamilies, with the resulting hybrid element becoming an active source element[26]. The small burst of retrotransposition associated with *Alu*Ta15 and its derivatives may have been triggered by this gene conversion event in that the sudden creation of several diagnostic mutations simultaneously may have temporarily overwhelmed certain host factors that inhibit *Alu* retrotransposition. In the most recent history, *Alu* retrotransposition appears to be somewhat slower in common marmosets compared to the observed rate in humans and rhesus macaques.

The evolution of L1 in the lineage leading to the common marmoset generally followed a typical linear evolution pattern with one subfamily being succeeded by a younger subfamily. The NWM-specific L1 lineage intercepts with the human-derived L1 tree between L1PA6f and L1PA7. The youngest L1 subfamily has an average divergence of 0.65% and appears to continue to propagate in the marmoset genome.

7.2 Ensembl annotation

The final gene set of 21,168 genes included 219 genes with at least one transcript supported by marmoset protein, a further 15,706 genes without marmoset evidence but with at least one transcript supported by human evidence. The remaining 5,243 genes had transcripts supported by proteins from other sources (Supplementary Figure 6g).

The final transcript set of 44,973 transcripts included 232 transcripts with support from marmoset proteins, 5,731 transcripts with support from human proteins, 24,718 transcripts with support from human cDNA with CDS information, 12,770 transcripts with support from UniProt SwissProt, and 1,522 transcripts with support from other protein sequences (Supplementary Figure 6h).

7.3 Functional RNAs – miRNAs, siRNAs

MicroRNAs (miRNAs) are small non-coding RNAs of ~22 nucleotides in length that have been revealed in the last decade to be important agents of posttranscriptional gene regulation (see Ensembl gene annotation URLs for review). MiRNAs have been postulated to play a role in changes in morphology and body plan of living organisms[27-29] since the number of miRNAs in a genome increases in proportion to the increasing complexity of the organism. A miRNA is initially transcribed as a long primary miRNA (pri-miRNA) which is processed into a ~70 nucleotide stem loop precursor that is further processed into a small ~22 nucleotide double stranded RNA by RNaseIII enzymes Drosha and Dicer respectively[30,31]. Following this, the strand with the least stable 5’-end is selected and loaded into the RNA Induced Silencing Complex (RISC) as the active mature miRNA. The remaining strand (passenger strand or miRNA* strand) is typically degraded.

Once incorporated into the RISC, the mature miRNA binds target mRNAs with complimentary sequences. The captured mRNA targets are silenced through mRNA cleavage and/or translational repression[30]. The minimum requirement for a productive miRNA-mRNA interaction is at least a 6 consecutive nucleotide pairing between the miRNA (including the 5' seed region, nucleotides 2-7) and mRNA target sequences that are largely found in the 3'-UTRs of the mRNAs[30,32-34]. As a result of this minimal base-pairing requirement, a single miRNA can bind, capture and silence hundreds of genes within and across diverse signaling pathways[32].

MiRNA-targetRNA relationships are typically highly conserved during evolution with the 5' seed region of the miRNA showing very little change across species[27] in whole genome alignments of 18 placental mammals and 10 other vertebrates[35]. The complete set of miRNAs discovered to date in human has been determined to contain binding sites in 60% of the genes in the genome and therefore, miRNAs have widespread impact on gene regulation[32]. Here, we present our findings on the evolution, expression and miRNA-target gene functional pairs of miRNAs in the marmoset. In a study that looked at evolutionary conservation of miRNAs cloned from human and chimpanzee brain across 17 genomes 75% of known human miRNAs cloned at the time were found to be conserved in vertebrates and mammals and 14% were conserved in invertebrates, only 10% were primate specific and 1% human specific[36]. New miRNAs are largely identified through deep sequencing approaches and are in general found in low abundance. New human miRNAs reported by Berezikov and collegues[37] were found to have very different conservation profiles with more than 50% of novel human miRNAs conserved only in primates, 30% in mammals, 9% in nonmammalian vertebrates or invertebrates and 8% were specific to humans.

We have identified a total of 777 mature miRNAs that are encoded in the marmoset genome mapping to a total of 1,165 precursor hairpin miRNAs (Supplementary Tables 20-37). Supplementary Table 22 lists all of the marmoset miRNAs, with unique identifiers, sequences and expression data. The first letter of the miRNA IDs indicate the source groups (A through F) described below. Most of these mature miRNAs (582, 75%) have been confirmed through expression in either the marmoset brain or placenta. A total of 261 miRNAs mapping to 375 hairpins have 5' seed sequences that are 100% conserved in at least one other species and are expressed in the marmoset placenta and/or brain (Group A). Another 82 expressed miRNAs mapping to 143 hairpins match a miRNA from another species with 1 to 3 mismatches and contain at least 1 mismatch in the 5' seed sequence (Group B). A third set of 239 expressed miRNAs mapping to 334 hairpins are novel and have not been found in any other species to date (Group E). We anticipate this group of putative novel miRNAs contains a mixture of miRNAs that are exclusive to marmoset, miRNAs that are exclusive to NWM, and conserved miRNAs that are yet to be discovered in other species. Since only two tissues were subjected to small RNA sequencing, we used all known miRNAs from all species in miRBase[M86] version 17.0 to search the marmoset genome. From this we found 107 additional miRNAs mapping to 223 hairpins that have been perfectly conserved in at least one other species (Group C). To account for those miRNAs that are found in miRBase 17.0 that may have diverged during marmoset evolution we also considered miRNAs that matched the marmoset genome with 1-3 mismatches and found an additional 66 miRNAs (Group D). Finally we also found an additional 15 hairpins using comparisons to human miRNAs evaluated with a support vector machine (Group E) and recorded 9 hairpins mapping to the Chr. 22 miRNA cluster (see below, Group F)[38]. Half of the putatitve miRNAs (388) are confirmed in that they have a 5' seed sequence with a 100% match to a miRNA in miRBase 17.0 from another species.

Some miRNA clusters show rapid evolution in marmoset lineage. Comparing marmoset miRNAs to four other anthropoid species including human, rhesus macaque, chimpanzee and orangutan we found 55% to 58% of marmoset miRNAs are conserved. Some families, such as the let-7 family of miRNAs are 100% conserved in all five species (Supplementary Tables 22-28). By contrast, two of the largest clusters of miRNAs in the marmoset genome show considerable expansions in the number of miRNAs and substantial sequence divergence compared to human. The Chr. X cluster has a much smaller fraction of the marmoset miRNAs conserved in human (4%), chimpanzee (0%), orangutan (8%) and rhesus (0%). This Chr. X cluster has expanded considerably more in marmoset (Supplementary Table 22). Human Xq27.3 has a 7-member cluster of miRNAs (including miR-506, miR-507, miR-508, miR-509, miR-510, miR-513 and miR-514) mapping to 15 hairpins. Two members, miR-513 with 3 haripins in human and miR-514 with 4 hairpins in human account for 49% of the expansion of this family into 40 hairpins in marmoset. Another 50% have diverged beyond our ability to recognize a common ancestor into entirely novel miRNAs. The primate-specific cluster on human 19q13.42 with 49 members has expanded in marmoset to generate 112 hairpins on marmoset chromosome 22, for which we have evidence for 71 mature miRNAs (Supplementary Table 22) Only 0% to 3% of the miRNAs from this marmoset Chr.22 cluster are conserved in human, chimpanzee, orangutan and rhesus. Interestingly, there is a distinct boundary to this expansion. Three adjacent miRNAs (caljac-miR-371, caljac-miR-372 and caljac-miR-373) located on the 3'-end of the cluster do not exhibit this rapid expansion in marmoset or other anthropoids.

The majority of the miRNAs in the rapidly evolving Chr. 22 and Chr. X clusters exhibit at least one nucleotide modification in the 5'-seed region. Most of the miRNAs in these large clusters (83% of Chr. 22 and 78% of Chr. X miRNAs) showed nucleotide changes within the 5' seed sequence when compared to human (Supplementary Table 22 and Supplementary Table 29). MiRNAs from both clusters showed striking divergence in the 4 anthropoid species as well (Supplementary Table 20). The change in the number as well as nucleotide sequence in the cluster members across the 5 anthropoid species is shown in Supplementary Tables 23-28. These seed region changes in marmoset are expected to substantially change the target repertoire of these miRNAs unless the complimentary 3'-UTR binding sites have co-evolved in the target mRNAs.

The rapidly evolving Chr. 22 and Chr. X clusters dominate the miRNA expression in marmoset placenta whereas the marmoset brain exhibits diverse miRNA expression. We examined the expression characteristics of the miRNAs, sequencing small RNAs from total RNA isolated from placenta and brain. Copy numbers of the full set of marmoset miRNAs expressed in brain and placenta are shown in Supplementary Table 22. We found that a total of 587 miRNAs (76%) are expressed in marmoset brain and/or placenta. The relative expression patterns of the miRNAs in these clusters in marmoset placenta and brain are shown in (Supplementary Figure 9a). The Chr. 22 and Chr. X cluster miRNAs are among the most abundantly expressed miRNAs in the placenta. By contrast, the marmoset brain showed a diversity of miRNA expression. The human 19q13.42 cluster (corresponding to the marmoset Chr. 22 cluster) consists of 53 mature miRNAs, of which 42 are expressed in the human placenta[39,40]. We found 38 of 53 human 19q13.42 miRNA orthologs are expressed in the marmoset placenta as well. A heat map depicting the relative expression of other miRNAs in brain and placenta samples are shown in (Supplementary Figure 9b).

Expression profiles of Marmoset microRNAs. Sequence reads from the small RNA sequencing that passed the quality control filters were used to estimate the expression profiles for marmoset microRNAs. Reads that mapped to the same chromosome as the Marmoset microRNA and lie in the region flanking 4 bases on either side of the microRNA contributed to the total expression profile of the microRNA. The copy numbers of each microRNA was normalized to the total number of usable reads. For profile clustering, the read of each microRNA per 10 million usable reads in each sample was log2 transformed and the median expression value across the six samples were set to zero. Cluster 3.0 and Tree View software was used for cluster analysis and representation[M90] (see URLs). The Euclidean hierarchical clustering was performed on both genes and the arrays at different samples. The samples are two placentas (900 and 536-1122) and brain (A07-716, A09-122, A08-206, A08-337).

7.4 Functional RNAs – miRNA targets

Since most of the miRNAs in the two large clusters in marmoset have nucleotide differences within the 5'-seed region compared with their human homologs, these miRNAs may have evolved different target repertoires in those two species. It is possible that some of these differences are related to the propensity of marmosets to routinely generate multiple infants per pregnancy, whereas other anthropoid primates generally produce only singleton births. Supplementary Tables 31, 32 show the predicted targets of the chromosome 22 cluster that have been linked to pregnancy-related processes and/or found to be expressed in the human placenta. See Supplementary Table 36 for all of the chromosome 22 cluster predicted targets.

In order to study the evolution of miRNA targets in marmosets, and across primates and other mammals, we identified 1564 protein coding genes with clear 1-to-1 orthologs among human, chimpanzee, rhesus macaque, marmoset, mouse, rat, dog, horse, cow, opossum and platypus. TargetScan predicted at least one let-7 target site in the 3'-UTR of at least one of these species in 545 genes. Comparing the genes predicted to contain targets for let-7 across these species, we find that the number of gene targets shared with human declines with greater evolutionary distance (i.e. chimpanzee to rhesus to marmoset to non-primate mammals). However, the proportion of let-7 targets shared with human is roughly the same in marmoset, dog, horse and cow. The proportion of let-7 targets shared with human is lower in rat and mouse than in other non-primate mammals. In parallel, the number of let-7 targets found in the other mammalian species that are not shared with human increases with evolutionary distance from human, with the largest fraction of non-shared targets found in mouse and rat.

Comparing the predicted 3'-UTR targets in human and marmoset for the perfectly conserved let-7 family of miRNAs (see Supplementary Table 34), we found 165 predicted targets common to both species, 44 unique to marmoset and 64 unique to human. The difference in the number of let-7 targets unique to each species does not have an obvious biological explanation but there are technical factors that may contribute, including the different quality of the genome sequences used for predictions (i.e. a completed finished quality human genome versus the draft marmoset genome) and the quality of 3'-UTR annotations (a newly annotated marmoset genome with little expression data versus the well-curated annotations based on extensive transcriptome data in human).

We next plotted the number of let-7 target gains and target losses on the phylogenetic tree for seven species. Note that for this analysis, gains and losses of targets that occurred twice on independent evolutionary lineages, and hence may represent either multiple changes in the same target or sequencing or annotation errors, were not counted, (n=55). Among the target gains and losses that can be mapped unambiguously onto the tree, we see two interesting patterns. First, the number of target gains exceeds the number of target losses in every segment of the tree, with the total number of gains four times the losses (196:49). Among the primates alone, gains of let-7 targets exceed losses by more than five-fold (100:17). Second, the total rate of change (gains plus losses) is greater in the primate lineages than in the non-primate branches of the tree (with the one exception of the branch leading to rat, after its divergence from the mouse lineage). The observation that the number of let-7 target gains exceeds losses in every branch of the tree suggests that the origin of new let-7 targets occurs consistently over evolutionary time, and that once a target sequence appears, selection to retain the new target sequence is more effective than is either selection or genetic drift in eliminating older target sequences shared with other mammalian lineages. This is true despite the fact that the rate of origin of new targets may be expected to be lower than the rate of random mutational hits that eliminate existing targets.

We draw three conclusions regarding let-7 target sequences: a) despite the extreme conservation of let-7 miRNA 5'-seed sequences, among mammals and especially among primates and rodents, there is a considerable rate of evolutionary change in the protein coding genes targeted for regulation by let-7, b) the rate of origin of new let-7 target sequences in 3'-UTRs, and hence the rate of increase in the number of protein coding genes targeted for regulation by let-7, exceeds the rate of loss of let-7 target sequences, at least among mammals, and c) the rate of change in let-7 target evolution may be somewhat higher in primates and especially in catarrhine primates, than in non-primate mammals overall.

The observed pattern of evolutionary change is quite different among the three described marmoset miRNA families: chromosome 22, chromosome X and let-7 (Supplementary Tables 34-36). First, while the let-7 miRNA family shows extreme evolutionary conservation of 5'-seed sequences across vertebrates and invertebrates, both the chromosome 22 and X clusters show greater divergence in 5'-seed sequences. Comparing the chromosome 22 and X clusters to each other, we find an unexpected contrast in the pattern of miRNA:miRNA target evolution between these two miRNA families. There are substantial differences in the X chromosome 5'-seed sequences in marmoset compared with human and other primates, and this is paralleled by divergence of targets, as more than 50% of targeted 3'-UTR sequences are not shared between human and marmoset. On the other hand, for the chromosome 22 family, 5'-seed sequences have also changed significantly between human and marmoset, but in contrast to the chromosome X family, only 16% of targets are not shared between human and marmoset. This implies significant co-evolution of 5'-seed regions and their targets in the chromosome 22 family, but not in the chromosome X family. For let-7, a family of miRNAs that demonstrates remarkable conservation of seed regions across vertebrates and invertebrates, we find a surprisingly high rate of change in 3'UTR targets, with about 40% of targets not shared between marmoset and human.

Taken together, we see a de-coupling of the evolution of miRNA 5'-seed sequences and their targeted 3'UTR sequences in the X chromosome cluster, but much less change in miRNA:target relationships (i.e. more co-evolution of targets and seed regions) in the chromosome 22 cluster.

The specific protein-coding genes that gain or lose 3'-UTR let-7 target sequences include a number of loci of interest. Seven such loci are involved in neural function or neurodevelopment. All primates tested exhibit gains of targets in *GABRP*, *MAP2* and *GOPC*. Catarrhines gained a target in *TP63* and hominoids gained additional targets in *STX1A* and *SRGAP2*. This latter gene has undergone two human-specific partial duplications which, through dimerization with full length SRGAP2, inhibits its function[41]. Evidence suggests that this results in human-specific extension of an early developmental process (neoteny) in the neocortex by slowing dendritic spine maturation and increasing the density of longer spines, and these changes may have played a role in the origin of human neocortical function. The down regulation of SRGAP2 by *de novo* let-7 targeting may have contributed to earlier stages of these evolutionary changes that are shared among hominoid primates. Finally, we note that the single gene showing a loss of let-7 targeting in human, neuropilin 2 (*NRP2*), is associated with axonogenesis, cell morphogenesis during neuron differentiation, and morphogenesis of neuronal projections.

8. Orthologs and positive selection

8.1 Gene family expansions/contractions compared to other primates

A birth-and-death parameter λ in CAFE represents the rate of gene duplication and loss inferred from the extant distribution and size of the investigated gene families. CAFE allows different values of λ to be assigned to different branches on the ultrametric tree, and to test nested models of gene family evolution with different combinations of λ values. Several nested models of gene family evolution were used in the CAFE analysis. Together with the analysis of genes mostly shared by most of these ten mammalian species, we used CAFE to reconstruct the evolution of 429 gene families only present in primates (Supplementary Table 14).

Gene family changes in each branch of the ten species tree, including the marmoset lineage, were inferred by comparison with ancestral gene family size reconstructed using CAFE. Contractions exceed expansions of gene families in most mammalian lineages, with the exception of the branches leading to human, rat and cow and a some inner branches (Supplementary Table 13 and Figure 7a). The lineage-specific patterns of gene duplications and losses within gene families are more apparent when measured in terms of changes per million years (Supplementary Figure 7b). By comparing inferred gene family sizes of the marmoset and the reconstructed primates ancestor genome, we estimate that the evolutionary lineage leading to marmoset has experienced 1207 gene duplications and 1,542 gene losses in 738 and 1276 gene families, respectively (Supplementary Table 13). Among gene families with decreased size in marmoset, 541 represented family extinction. 102 and 77 families showed significant expansion or contraction in marmoset, respectively (P value < 0.05; Supplementary Tables 15, 16).

Gene duplications and gene family expansions in marmoset

Gene prediction in newly sequenced genomes is a challenging task even when comparative resources are available, such as in primates and more generally in placental mammals. Gene annotation in marmoset, and especially the annotation of gene duplicates absent from other sequenced mammalian genomes, could be strongly affected by the large number of retroposed gene copies observed in this genome, which in most cases form non-functional genes that might retain relatively long coding regions. Furthermore, because of the lack of introns, these non-functional gene copies could produce better alignment scores with cDNA sequences used to infer gene structure and expression pattern of genes. The comparison of the genomic location of predicted genes with the coordinates of segmental duplications (SDs) can help validating lineage-specific gene duplicates. We used SD coordinates derived from both WGAC and WSSD analyses (SOM 4) to validate the 3,565 genes from the 738 gene families that appeared to have expanded in the marmoset. We observed only 189 genes (5% of 3,565) overlapping with both WGAC and WSSD regions, and 803 genes (23%) overlapping with either WGAC or WSSD regions. This corresponds to 126 families with genes overlapping both WGAC and WSSD regions, and 415 more families overlapping either WGAC or WSSD regions.

We manually inspected 94 gene families with marmoset genes overlapping with WGAC, WSSD or both WGAC and WSSD regions, and having more genes than human or the ancestral primate nodes of the tree, i.e. families that are likely to be expanded in marmoset. More specifically, we considered true gene duplicates to be all genes predicted by the Ensembl and/or N-SCAN pipeline that are not found in the genome of other primates according to the multiz alignment available at the UCSC Genome Browser[M78].

Among the eleven families of this data set that overlap with both WGAC and WSSD, eight experienced marmoset-specific duplications and two were not further considered because of difficulties in gene number estimates, i.e. they contained many predicted genes on unmapped scaffolds, which could represent different alleles of the same gene instead of paralogous genes. Thus, as expected, duplicated regions predicted by both WGAC and WSSD methods are highly reliable, as also suggested by *in situ* validation of 37 BAC clones (Supplementary Table 19).

Among the 83 families with genes overlapping with either WGAC or WSSD regions, 29 had a high number of unmapped genes and were not further examined; of the 54 remaining families, 16 had gene duplicates specific to marmoset, indicating that ~30% of families (16/54) with genes overlapping with either WGAC or WSSD regions are indeed expanded in marmoset. These estimates are much lower than the ones from *in situ* experiments (~60% for WSSD only and 85% for WGAC only clones). The most relevant examples of gene family expansions in marmoset are described in Supplementary Table 17.

We also manually validated 35 gene families from the 270 families with putative expansion in marmoset but no overlap of their genes with WGAC or WSSD regions. Among the 22 families without unmapped ambiguous genes, 8 showed marmoset-specific duplicates, similar to the proportion found for families with genes overlapping either WGAC or WSSD regions. Interestingly, six of these expanded families are formed by genes encoding subunits of the mitochondrial ATP synthase or subunits of the mitochondrial NADH:ubiquinone oxidoreductase. Members of two of these expanded families (*NDUFA3*, *NDUFA4*) also showed evidence of positive selection, suggesting that selection on these genes might have acted both on gene copies and sequence evolution levels. Furthermore, these six families contain many independently retroposed gene copies in human and marmoset. While they all appear to be pseudogenes in human, several of these gene copies have been maintained in marmoset as functional new genes (they have intact coding sequences and their divergence from the parent genes is >5%). Most of these gene duplicates are in the range of 1kb in length and could have simply been missed by both WGAC and WSSD methods because the minimal size of detectable SD is 1 kb. Indeed, we observe that among the 2,573 marmoset gene duplicates that do not overlap either WSSD or WGAC regions, 32% are <1 kb in length, compared to 26% of duplicates overlapping with either WSSD or WGAC regions, and only 19% overlapping with both WSSD and WGAC regions.

The gene family analysis also highlighted 118 families that are present in both marmoset and mouse but apparently are absent in human. Manual validation showed that in most cases the corresponding genes were either present in human, absent in human and marmoset, absent in mouse, or were unmapped in marmoset. We were able to determine that 13 such families are indeed absent in human and present in marmoset/mouse. A few of them are functionally characterized in mouse, including the Cd24a antigen gene, the casein alpha s2-like B gene (Csn1s2b), and the cytochrome c oxidase, subunit VIIIb gene (Cox8b). Most of the losses occurred before the human-chimpanzee split (Supplementary Table 18).

Gene losses and gene family contractions

As for gene duplications, apparent gene losses need to be validated as they could represent artifacts derived from assembly issues, in particular large assembly gaps and regions of low sequence quality, as well as missing calls of gene prediction software. By combining information from whole-genome[M39] and multiple[42] alignments to other primate genomes, expression data and manual annotation of genetic regions we were able to confirm only 5% of gene losses in the analyzed gene families. Of a total of 142 genes, we confirmed 7 (5%) were lost, 74 (52%) were missed gene predictions, 46 (32%) were assembly issues the source of the artifacts causing the remaining 15 (11%) were ambiguous. The confirmed examples of gene losses in marmoset include these genes (family ID shown in parentheses): *NAIP* (5369), *FSCB* (227), *ASPG* (1892), *HSPB9* (3894), *THAP10* (7326), *LDLRAD2* (7631), *WFDC13* (12647). All seven are single copy in human and do not have expressed sequences from the cDNA libraries (SOM 2.6). Only ASPG has marmoset sequence that aligns with at least 50% of the human coding sequence. These genes are involved in a number of functions (*NAIP* prevents motor-neuron apoptosis, *FSCB* may be involved in the later stages of fibrous sheath biogenesis of the sperm's flagella, *ASPG* is L-asparagine amidohydrolase, *HAPBP9* is a alpha-crystallin-related, testis specific heat shock protein, *THAP10* may play a role in breast cancer, *LDLRAD2* is a low density lipoprotein receptor class A domain, and *WFDC13* is a protease inhibitor)

The most relevant examples of gene family contractions in marmoset are listed by a gene name with family ID, number of copies in human, and number of copies in marmoset, and function listed in parentheses: *ACSM1* (7089, 11, 5, acyl-CoA synthetase medium-chain gene family. GTP-dependent lipoate- activating enzyme that generates the substrate for lipoyltransferase), *FOXA1* (17935, 18, 5, hepatocyte nuclear factors that represent transcriptional activators for liver-specific transcripts such as albumin and transthyretin and they interact with chromatin), *SSX1* (8400, 10, 1, could act as a modulator of transcription, expressed at high level in the testis), *PRAMEF1* (17830, 23, 1, preferentially expressed antigen in melanoma), AGAP1 (18342, 10, 3, Centaurins a protein family involved in membrane traffic and actin cytoskeleton dynamics), CEACAM1 (456, 20, 9, carcinoembryonic antigen-related cell adhesion proteins that play roles in the differentiation and arrangement of tissue three-dimensional structure, angiogenesis, apoptosis, tumor suppression, metastasis, and the modulation of innate and adaptive immune responses), NCR1 (18142, 24, 4, cytotoxicity-activating receptor that may contribute to the increased efficiency of activated natural killer cells to mediate tumor cell lysis), GOLGA2 (450, 38, 7, golgi auto-antigen that are probably involved in maintaining cis-golgi structure), NPIPL3 (456, 20, 2, nuclear pore complex interacting protein-like genes, Morpheus family), SPANXA1 (18142, 24, 3, testis-specific genes required to initiate molecular and morphological changes necessary for the formation of mature spermatozoa), and HTN3 (450, 38, 1, Histatin family which exhibit non-immunological, anti-microbial activity in the oral cavity). AGAP1 and GOLGA2 are expanded in catarrhines or apes. NPIPL3, SPANXA1, and HTN3 are present only in primates.

Evolution of primate-specific gene families

Among the 429 gene families present only in primates, marmoset is the species with fewer genes in these gene families (Supplementary Table 8). Therefore, many of these families are likely to be Catarrhini-specific, or they mostly expanded in the Catarrhini clade. More than half of these families (221/429) are indeed absent in marmoset, indicating that they emerged after the Catarrhini-Platyrrhini divergence. In addition, many families are absent in rhesus macaque, suggesting that almost half of these primates-specific families are unique to apes. The most relevant examples of primate-specific genes absent in marmoset are described in (Supplementary Table 19).

8.2 Positively selected genes in marmoset and other primates

The likelihood ratio test on all branches identified 403 genes that show signs of positive selection (FDR<0.01). We also discovered 37 positively selected genes on the marmoset lineage (FDR<0.01), and 7 Positively selected genes (FDR<0.01) on the branch to Catarrhini. Additional 91 genes are positively selected on at least one of these two branches (FDR<0.01), but could not be traced to a particular branch due to unidentified outgroup orthologs.

We have tested for enrichment of Positively selected genes in functional categories using both the Mann-Whitney U-test (MWU) and the Fisher's exact test (FET), and we have considered all genes with P<0.05 as positives. Functional categories enriched for Positively selected genes on the marmoset lineage are mostly related to immunity, defense, and sensory perception, but also include several atypical categories related to mitochondrial ATP synthesis and transport and NADH dehydrogenase activity (Mann-Whittney U-test, P<0.05). In particular, eight nuclear genes encoding subunits of the respiratory Complex I were subject to positive selection. The small size of marmosets represents a big challenge for its thermoregulatory system, underlined by the fact that the temperature of its body changes by up to 4°C. This must have resulted in thermoregulatory and endocrine adaptation that may be explained by adaptation of these genes. In particular, six proteins (NDUFA3, NDUFA4, NDUFA8, NDUFA10, NDUFB2, and NDUFB9) are accessory subunits thought not to be involved in catalysis, and two remaining subunits (NDUFS2, NDUFS3) are present within the enzymatic core believed to be the minimal assembly required for catalysis[43]. Changes in amino acid sequences may have resulted in different regulatory (for accessory subunits) and kinetic (for catalytic subunits) properties of the complex I, thus affecting metabolic rates and body temperature.

Mutations within type-1 insulin-like growth receptor IGF1R (marmoset lineage PSG, P=0.0014) were previously shown to cause short stature in different organisms[44]. Using Bayes empirical Bayes method[45] included in PAML, we have identified amino acid sites under positive selection in IGF1R and mapped those sites onto the known crystal structure of the first three domains L1-Cys-rich-L2 (PDB accession 1igr, [46]). The result is shown in Figure 3. IGF1R shows multiple mutations within the L1 and L2 domains of the α chain crucial for binding insulin-like molecules. The marmoset receptor protein also shows a striking sequence of mutations within the Cys-rich loop essential for binding specificity of the ligand. Such extensive changes likely affect ligand-receptor binding affinity.

Twinning in marmosets (and tamarins) is associated with an unusual feature: adult marmosets are chimeras of cells derived from two (or more) products of conception. Cells from the products of conception colonize the bone marrow of both twins. As a result, blood samples from adult marmosets with a twin of the opposite sex contain lymphocytes with both XX and XY karyotypes[47]. This type of chimerism is a result of placental fusion during development and it represents a challenge for the immune system that might be reflected by changes in the cell surface proteins. Positively selected CD48 is a ligand for CD244 (2B4) and is broadly expressed on the surface of hematopoietic cells and it also participates in regulation of natural killer cells[M31]. The marmoset variant of CD48 may have been adapted to the generation of germ chimeras. Other positively selected proteins that may be involved in circumventing unwanted responses associated with the chimerism include interleukins IL5 and IL12B, involved in T cell development and in allergic responses[M32]. Finally, GMCL1 (germ cell-less 1) protein is involved in embryogenesis in *Drosophila melanogaster*, in particular in producing the germ cell line[48], though its precise role in mammalian development remains obscure. GMCL1 is positively selected in marmosets and it could have a similar role in the production of chimeric twins.

Marmoset has been used as a model of various infectious diseases[49]. In this regard it is interesting to note that one of the positively selected genes encodes a cell-surface antigen, complement regulatory protein CD46, which was found to be a major receptor for a special class of adenoviruses[50].

Several genes encoding ligands (e.g., IGF2, EGF, CSF1, BMP5) for various types of receptors were also subjects of positive selection, although this does not seem to be accompanied by co-evolution of the corresponding receptors. It is possible that the changes in amino acid sequences affected the affinity of the ligand-receptor binding and thus the sensitivity of the signal transduction pathway.

9. Specific gene family studies

9.1 Genes involved in growth pathways and twinning

Marmosets display positive selection for many suites of genes that are also under positive selection in other mammalian groups, including genes associated with immunity and sensory systems. Marmosets also display positive selection for a number of genes in the Growth Hormone axis and that is an unusual finding. This observation is of interest given the proposed selection for secondary reduction in body size that occurred in the callitrichid lineage[51,52]. Changes in function of these genes may be associated with altered pre- and post-natal growth in marmosets, when compared to other primates. Genes undergoing positive selection include those for the Growth Hormone Secretagogue Receptor, isoform 1a (GHSR), Insulin-like Growth Factor II (IGF-II), Insulin-like Growth Factor 1 Receptor (IGF-IR), and IGF binding protein x. Wallis[53] has proposed that NWM, as a group, have undergone periods of rapid change in insulin-related peptides, including insulin and IGF-I. The information for insulin is convincing in that there are multiple NWM that display identical differences from human in genome sequence. In the case of IGF-I, the only NWM data presented are from the marmoset, leaving open the question as to whether changes in IGF are general NWM or may be marmoset-specific.

Mutations of GHSR and IGF-IR are associated with short stature in humans. Over a dozen nonsense and missense mutations in the GHSR gene have been identified in humans with short stature[M21, M21, 54-58]. Some of these mutations affect the ability of GHSR to bind its ligand, ghrelin, but many affect the constitutive activity of GHSR. Such changes in constitutive activity are associated with short stature in humans while the occurrence of other deleterious phenotypes is variable depending upon the specific mutation. Further exploration of the nature of positive selection for GHSR in marmosets is warranted as a possible molecular mechanism resulting in the hypothesized secondary reduction in body size in the callitrichid primate lineage.

Numerous IGF-IR missense and nonsense mutations are also associated with growth retardation in humans. In this case, the growth retardation is both pre- and post-natal[59,60]. In the marmoset, the IGF1-R gene shows modifications that can be reasonably proposed to result in changes in binding specificity. There are multiple marmoset-specific mutations within the L1 and L2 domains of the α chain crucial for binding insulin-like molecules, and a striking sequence of mutations within the Cys-rich region loop essential for binding specificity of the ligand. As with GHSR, changes in IGF-IR activity may underlie adaptive changes in post-natal growth in marmosets. They may also be related to unusual features of placentation and prenatal growth in callitrichid primates. This scenario as a possible route to secondary size reduction is of particular interest given proposals that miniaturization in the callitrichid group is related to deceleration of pre-natal as opposed to post-natal growth[52,61].

Marriog & Cheverud[M6] propose that callitrichid primates, in comparison with other NWM, exhibit a gestation length that is to be expected based upon their body size but that they display significantly slower pre-natal growth rates. However, their analysis is based upon the assumption that pre-natal growth patterns are similar between callitrichid primates and other NWM. In fact, marmosets and other callitrichid primates display an unusual pattern of placental and embryonic development that suggests timing of these early events may be a critical feature in ultimate miniaturization.

Callitrichid primates also display a suite of reproductive characteristics that are unusual among anthropoid primates. They typically ovulate 2 ova at a time, producing twins. The gestation of the two conceptuses takes place in a simple (i.e. not bicornuate) uterus and is supported by a bi-discoid placenta. A simple uterus and a bidiscoid placenta are characteristic of most anthropoid primates. However, the gestation of multiple offspring in a simple uterus is rare in mammals and often associated with pathologies, such as freemartism in cattle and twin-twin transfusion syndrome in humans. Such pathologies do not occur in marmosets. Placentation and gestation in callitrichid primates display unusual features of both timing and location of developmental events that seem likely to be the (adaptive?) result of this habitual, unusual species biology. The bi-discoid placenta of marmosets is produced from trophoblasts contributed by both conceptuses. Within a week after implantation, the blastocysts rapidly expand, filling the uterine cavity. The touching chorionic walls of each embryo fuse, such that the conceptuses are held in a common chorion. As the placenta develops there are extensive vascular connections within each disk and between the disks, such that they function, in terms of exchange, as one unit. Around day 61, hemtopoietic foci begin to develop within the placenta, peaking in mass at around day 100 then declining so that few are present at delivery around day 143. These hematopoietic foci within this chimeric placenta are the source of hematopoietic cells for both embryos, resulting in hemaotopoic chimerism.

The timing of the development of the placenta and organogenesis is unusual, with the embryo being largely quiescent until around day 40, such that organogenesis lags behind that observed in other primates by about three weeks. It has been suggested that this delayed organogenesis has its origins in the selection pressures stemming from the common occurrence of energetically expensive pregnancy and lactation overlapping in marmosets - another rare feature of this group being the lack of a post-partum anovulatory period. However, it is worth exploring the possibility that the developmental lag and timing of hematopoietic foci is tied to protection of the fetuses against common pathologies known to be associated with gestating litters in a simplex uterus. This delayed developmental pattern may also be the ultimate source of small birth size leading to small adult size in this group of primates.

This altered placental and embryonic development process may be associated with positive selection for both IGF-II and IGFIR, given that IGF-II as well as IGF-I are ligands for IGF-IR. IGF-II is expressed primarily in tissues arising from the blastocyst, that being the embryo and the portions of the placenta that are embryonic in origin. The expression pattern of IGF-II in marmosets has not been examined. Studies[62] demonstrated that IGF-II plays a critical role in the turnover and renewal of trophoblasts in the human placenta and studies of mice with genetic modifications of IGF-II indicate that reduced IGF-II activity decreases the size of the placenta.

It is plausible that the notable difference in timing of placental development in marmosets when compared to other primates is related to differences in IGF-II and IGF-IR function, given the role that IGF-II plays in differentiation, cell turnover and cell renewal as the trophoblasts differentiate into sycytiotrophoblasts playing different roles in the different parts of the placenta. Smith and Moore[63], for example, documented invasion of the syncytiotrophoblasts into the maternal blood vessels at around day 60 of pregnancy in the marmoset, an event that occurs at around 11 days after ovulation in humans.

The most common scenario proposed for the evolution of callitrichid primates is that small body size preceded increased ovulation number and production of litters. However, the possibility that delayed early placental and embryonic development might actually be a mechanism through which miniaturization is achieved raises the question of whether increased litter size might, itself, have been a driving force for producing smaller monkeys.

9.2 Protease Genes

The degradome can be defined as the complete repertoire of proteases in an organism[64]. From a genetic point of view, the degradome is highly attractive for several reasons. First, the degradome is composed of a large number of diverse genes. Thus, the human degradome includes more than 560 protease genes, encompassing five different catalytic classes and 67 families, some arranged in genomic tandems and some dispersed throughout the genome[65]. On the other hand, the number of known proteases allows the genomic study of the degradome with computer-assisted manual methods which avoid some of the noise and biases caused by purely automatic methods[66]. We have previously used this approach to discover copy number variations, gene gains or losses, and inactivating mutations in proteases from different animals including mouse, rat, platypus or zebra finch[67-70]. We have also compared the human degradome with those of other primates such as chimpanzee[71] and orangutan[M13]. Additionally, and because proteases have been related to a wealth of biological and pathological processes[72], we have used this accumulated knowledge on protease functions to raise hypotheses that link the genomic sequence to biological traits in a given organism. In this regard, it is remarkable that comparative genomic analyses of diverse degradomes have singled out the reproductive and immunological systems as the main protease evolution drivers[M12,73-75]. As expected, most of the differences between the degradomes of human and marmoset occur at genes related to the reproductive and immune systems (Supplementary Figure 10a).

Reproductive system

Proteases play multiple and diverse roles in reproduction, from spermatogenesis to embryo implantation. We found several events in the marmoset genome that might affect some of these genes.

- **KLK3,** or prostate-specific antigen, is a serine protease that has been shown to degrade semenogelins and change the physical properties of semen, which, in turn, relates to semen competition[76]. Our analysis shows no marmoset KLK3, which is in agreement with a previous biochemical study[77] and is consistent with the idea that KLK2 and KLK 3 arose from a common ancestor after the divergence of platyrrhini and catarrhini. Interestingly, the KLK2/KLK3 gene that is present in marmoset features a premature stop codon. The lack of this protein would be expected to increase the viscosity of semen, thereby decreasing the probability of female fertilization by subsequent mating males. In contrast, humans, chimpanzees, and orangutans contain one copy of each gene.
- **ADAM6**, is a metalloprotease specifically expressed in meiotic germ cells and may play a role in regulation of fertility[M34]. While this gene is functional in mouse, rat, and orangutan, ADAM6 was independently pseudogenized in humans and chimpanzees. We have found no orthologue of this protease gene in marmoset, which suggests that ADAM6 was independently lost.
- The serine proteases **ISP1** and **ISP2** are believed to play a role in embryo implantation in mice[M35], although ISP2 has been pseudogenized in a common

ancestor to humans and chimpanzees and ISP1 is not present in humans, chimpanzees or orangutans. We have previously described a complete and putatively functional ISP2 gene in orangutan. Strikingly, the present analysis shows that the marmoset genome contains three ISP2-like genes in tandem inside the **tryptase cluster** of serine proteases (Supplementary Figure 10b). All three genes display the features associated with serine protease activity, including three conserved His, Asp and Ser residues that form the catalytic triad. A phylogenetic analysis showed that one of these genes is a *bona fide* orthologue of mouse ISP2, whereas the other two genes, that we have called ISP2L1 and ISP2L2, show evidence for divergent evolution (Supplementary Figure 10e). Interestingly, ISP2L2 has lost the fourth characteristic disulfide bridge through mutation of two Cys residues. While a different origin for marmoset ISP2L1 and ISP2L2 genes cannot be ruled out, none of them clusters phylogenetically with murine ISP1 when using maximum parsimony or maximum likelihood methods. This is consistent with tandem duplication of ISP2 followed by fast mutation rates in two of the resulting genes. Thus, the marmoset orthologue of murine ISP2 seems to have been subjected to evolutionary pressure to retain its function, whereas the other two genes may have acquired novel biological functions.

Immune system

Several proteases known to be involved in the immune system show evidences of the strong evolutionary pressure on this system.

- The genomic **tryptase cluster** contains several serine proteases that play a role in mast cell biology[M37], and, in mammals, their genes have frequently evolved by tandem duplication events[M38]. In marmoset, the tryptase cluster seems to have undergone important specific rearrangements. In addition to the tandem duplications in ISPs, we have found that mastin (**PRSS34**) has been duplicated (Supplementary Figure 10b). A preliminary analysis of the rhesus monkey genome suggests that this primate has orthologs for both mastins. Therefore, mastin seems to have been duplicated in a primate ancestor and then one of the copies has been lost in hominoids. Notably, the genomes of human, chimpanzee and orangutan contain one pseudogenized copy of delta-tryptase (**TPSD**), which is not present in marmoset.
- The **chymase cluster** includes several serine protease genes whose products are also stored and secreted by mast cells[M37]. In marmoset, there are two copies of **CMA1** (chymase-1) in tandem. While one of the copies is an orthologue of human CMA1, the other copy, named **CMAL**, does not cluster phylogenetically with human or bovine chymases (Supplementary Figure 10f). The result of this phylogenetic analysis suggests that **CMAL** may have arisen from the ancestral serine protease that mouse chymases stemmed from. However, a loss of functional constraints in CMAL after CMA duplication followed by fast mutation cannot be ruled out. Moreover, the conservation of all of the catalytic residues suggests that CMAL is an active serine protease which may be fulfilling novel roles. Notably, the cow genome also contains two tandem copies of chymase-1. However, both copies cluster together in the phylogenetic analysis (Supplementary Figure 10f), and therefore this duplication seems to be independent in origin and consequences from the duplication found in marmoset.

- The haptoglobin cluster in marmoset contains two genes (**HP** and **HPR**) and lacks the third one (**HPP**). Only the orangutan genome contains all three active haptoglobins, whereas the human genome lacks HPP and the chimpanzee genome shows a truncated HPR copy.
- The serine protease **PRSS33** is a macrophage-specific serine-protease whose expression is up-regulated in activated macrophages[78]. While humans and marmosets show a fully functional copy of this gene, chimpanzees, orangutans and rhesus monkeys have independently lost PRSS33 through an Alu-mediated recombination mechanism and two different premature stop codons, respectively.
- The cysteine protease **CASP12** transiently inhibits the activity of caspase-1, which slows down inflammatory cytokine processing in response to septic infections[79]. This gene has been pseudogenized in marmoset through premature stop codons. Human CASP12 is a pseudogene in most of the population, whereas chimpanzees and orangutans contain a functional copy of this gene.
- The aspartyl protease **NAPB**, which is specifically expressed in spleen, thymus and lymphoid and myeloid cells, may have been duplicated in marmoset. This gene has been pseudogenized in human and is functional in orangutan and chimpanzee.
- The hominoid-specific cysteine protease **USP6** arose from the fusion of duplicates from USP32 and TBC1D3. The link of human USP6 to the immune system has been discovered when a large functional genomic screen showed that the expression of this gene is necessary for HIV infection[80]. As expected, the marmoset genome contains no USP6 orthologue, consistent with the proposed hominoid origin of this gene.

The analysis of the marmoset degradome has also shown interesting traits related to different marmoset features. Thus, we have found that marmoset **MMP19** contains a frameshift mutation as compared to human MMP19 at its C-terminus. In human, this matrix metalloproteinase (MMP) is unique in that it contains a C-terminal extension after the hemopexin domain. Moreover, it is the only MMP that is expressed at significant levels in most tissues under quiescent conditions. Interestingly, MMP19-deficient mice develop diet-induced obesity due to adipocyte hypertrophy[81]. Furthermore, this protease is a candidate IGFBP3-processing enzyme, and therefore its activity might influence growth and development[M40]. Since the C-terminal extension of MMP19 is poorly characterized, the functional consequences of this marmoset-specific mutation are not clear. One possible scenario is that the distinct marmoset C-terminal extension affects the binding of MMP19 to the cell surface[82].

Finally, we have found an abnormally high number of **single-exon protease-like open reading frames**. These putative genes usually arise from retrotranscribed mRNAs inserted into the genome. Most of these single-exon ORFs lack a promoter and therefore are not transcribed and quickly accumulate inactivating and pseudogenizing mutations. Consistent with this, we have found no specific mRNA reads for any of these ORFs. On the other hand, the fact that we have found seven of these single-exon pseudogenes with a complete conserved ORF (**UCHL1**, **UQCRC2**, **EIF3S3**, **POH1**, **PSMA4** and two instances of **DJ1**) suggests either that some of these genes are functional or that these retrotranscription events took place recently.

9.3 PRDM9/PRDM7

The PRDM9 protein binds DNA motifs present in many human and mouse recombination hot spots and is known to affect recombination activity during meiosis[M42, 83-85]. We sought to understand the relation between PRDM9 and related genes. PRDM9 and PRDM7 protein sequences were retrieved using the human proteins as queries in BLAT searches on the UCSC Genome Browser and BLAST searches on the NCBI nr and wgs databases. The MAFFT software[86] was used to generate multiple alignments of the N-terminus region of PRDM9 and PRDM7 (amino acids 1-367 of human PRDM9), which contains a KRAB domain and a SET domain, from multiple primate and other mammalian genomes. Phylogenetic trees were obtained with parsimony and maximum-likelihood algorithms (options: JTT method, with invariant sites, complete deletion of gaps/missing data, NNI heuristic method, 1000 bootstrap replicates, automatically made initial tree) implemented in MEGA 5.0[87].

After extensive research, we did not find any evidence of a second *PRDM9*-like gene in the marmoset genome, which supports our conclusion that this duplication event is catarrhine-specific. To rule out the possibility that a second gene has been deleted in the marmoset and was originally present in all primates, we explored the genomes of the tarsier, bushbaby and mouse lemur. Both tarsier and mouse lemur had short sequences corresponding only to the PRMD9 SET domain, which were not further analyzed. The bushbaby genome showed three *PRDM9*-like sequences, one of which is neighbored by *URAH* and *GAS8*. A phylogenetic tree of the KRAB and SET domains from mammalian PRDM7/PRDM9 proteins shows that the bushbaby paralogs branch outside of the other species' sequences, probably because of their fast evolution resulting in a long-branch attraction effect on the tree (Supplementary Figure 10d). The many sequence changes in bushbaby *PRDM9*-like genes compared to other mammals could also derive from the low coverage of this assembly. One of the bushbaby *PRDM9*-like sequences has been excluded from this tree because it shows a truncated sequence, although it appeared to group with the two other bushbaby sequences in another phylogenetic reconstruction (data not shown).

A comprehensive analysis of the complex phylogeny of *PRDM9* and *PRDM7* genes has also been performed by Dr. Thomas Pringle (see URLs).

9.4 HLA, KIR and immunogenetics

BAC clones from the MHC gene cluster were sequenced to determine the gene structure in the region. A number of clones were selected to represent a tiling path across the MHC region, these include: CH259-467M4 (AC242519), CH259-463N5 (AC242709), CH259-499E13 (AC242520), CH259-5M19 (AC242849), CH259-474J24 (AC242654), CH259-233C2 (AC242517), CH259-32E2 (AC242518), CH259-269H10 (AC242532), CH259-484C6 (AC242710), CH259-217M17 (AC243176), CH259-510K19 (AC242711), CH259-148J2 (AC243944), CH259-318L2 (AC242653), CH259-36P11 (AC242497), CH259-178F11 (AC243724), CH259-297P6 (AC243001), CH259-49P2 (AC242576), CH259-77F15 (AC242730), CH259-13G8 (AC242643), CH259-86F8 (AC242577), CH259-273L7 (AC242714), CH259-113D14 (AC243719), CH259-370E12 (AC242575), CH259-116F24 (AC243192), CH259-127E13 (AC243262), CH259-127H5 (AC243194), CH259-15O7 (AC243457), CH259-18G14 (AC243409), CH259-285C16 (AC243298), CH259-334C4 (AC243273), CH259-337O2 (AC243263), CH259-353B4 (AC244390), CH259-357A10 (AC244391), CH259-387J11 (AC243265), CH259-420A8 (AC243266), CH259-485P14 (AC244392), CH259-515I8 (AC243897), CH259-528M3 (AC242610). In addition, clone CH259-528M03 was selected to sequence as it contains the genomic region that corresponds to the KIR gene family expansion in human.

URLs

Ensembl gene build, http://www.ensembl.org/info/docs/genebuild/genome_annotation.html; Cluster 3.0 and Tree View software, http://rana.lbl.gov/EisenSoftware.htm; Phylogeny of *PRDM9* and *PRDM7* by Dr. Thomas Pringle, http://genomewiki.ucsc.edu/index.php/PRDM9:_meiosis_and_recombination#Comparative_genomics_in_placental_mammals.

References

Note: References from the manuscript are preceded by the letter M to distinguish them from the additional references listed here.

1 Moreira, M. A., Bonvicino, C. R., Soares, M. A. & Seuanez, H. N. Genetic diversity of neotropical primates: phylogeny, population genetics, and animal models for infectious diseases. *Cytogenet Genome Res* **128**, 88-98, doi:10.1159/000291485 (2010).

2 Marivaux, L. *et al.* Anthropoid primates from the Oligocene of Pakistan (Bugti Hills): data on early anthropoid evolution and biogeography. *Proc Natl Acad Sci U S A* **102**, 8436-8441, doi:10.1073/pnas.0503469102 (2005).

3 Miller, E. R. & Simons, E. L. Dentition of Proteopithecus sylviae, an archaic anthropoid from the Fayum, Egypt. *Proc Natl Acad Sci U S A* **94**, 13760-13764 (1997).

4 Seiffert, E. R. *et al.* Basal anthropoids from Egypt and the antiquity of Africa's higher primate radiation. *Science* **310**, 300-304, doi:10.1126/science.1116569 (2005).

5 Schneider, H. *et al.* Can molecular data place each neotropical monkey in its own branch? *Chromosoma* **109**, 515-523 (2001).

6 Huang, X., Wang, J., Aluru, S., Yang, S. P. & Hillier, L. PCAP: a whole-genome assembly program. *Genome Res* **13**, 2164-2170, doi:10.1101/gr.1390403 (2003).

7 Stanyon, R. *et al.* Primate chromosome evolution: ancestral karyotypes, marker order and neocentromeres. *Chromosome Res* **16**, 17-39, doi:10.1007/s10577-007-1209-z (2008).

8 Capozzi, O. *et al.* Evolutionary descent of a human chromosome 6 neocentromere: a jump back to 17 million years ago. *Genome Res* **19**, 778-784, doi:10.1101/gr.085688.108 (2009).

9 Cardone, M. F. *et al.* Evolutionary history of chromosome 11 featuring four distinct centromere repositioning events in Catarrhini. *Genomics* **90**, 35-43, doi:10.1016/j.ygeno.2007.01.007 (2007).

10 Ventura, M. *et al.* Localization of beta-defensin genes in non human primates. *Eur J Histochem* **48**, 185-190 (2004).

11 Lander, E. S. & Waterman, M. S. Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics* **2**, 231-239, doi:0888-7543(88)90007-9 [pii] (1988).

12 Alkan, C., Sajjadian, S. & Eichler, E. E. Limitations of next-generation genome sequence assembly. *Nat Methods* **8**, 61-65, doi:10.1038/nmeth.1527 (2011).

13 Boffelli, D. *et al.* Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science* **299**, 1391-1394, doi:10.1126/science.1081331 (2003).

14 Newman, J. C., Bailey, A. D., Fan, H. Y., Pavelitz, T. & Weiner, A. M. An abundant evolutionarily conserved CSB-PiggyBac fusion protein expressed in Cockayne syndrome. *PLoS Genet* **4**, e1000031, doi:10.1371/journal.pgen.1000031 (2008).

15 Nance, M. A. & Berry, S. A. Cockayne syndrome: review of 140 cases. *Am J Med Genet* **42**, 68-84, doi:10.1002/ajmg.1320420115 (1992).

16 Glazko, G. V. & Nei, M. Estimation of divergence times for major lineages of primate species. *Mol Biol Evol* **20**, 424-434 (2003).

17 Piluso, G. *et al.* Gamma1- and gamma2-syntrophins, two novel dystrophin-binding proteins localized in neuronal cells. *J Biol Chem* **275**, 15851-15860, doi:10.1074/jbc.M000439200 (2000).

18 Bashiardes, S. *et al.* SNTG1, the gene encoding gamma1-syntrophin: a candidate gene for idiopathic scoliosis. *Hum Genet* **115**, 81-89, doi:10.1007/s00439-004-1121-y (2004).

19 Hanna, J. *et al.* Human embryonic stem cells with biological and epigenetic characteristics similar to those of mouse ESCs. *Proc Natl Acad Sci U S A* **107**, 9222-9227, doi:10.1073/pnas.1004584107 (2010).

20 Tesar, P. J. *et al*. New cell lines from mouse epiblast share defining features with human embryonic stem cells. *Nature* **448**, 196-199 (2007).

21 Konkel, M. K., Walker, J. A. & Batzer, M. A. LINEs and SINEs of primate evolution. *Evolutionary Antrhopology* **19**, 236-249 (2010).

22 Cordaux, R. & Batzer, M. A. The impact of retrotransposons on human genome evolution. *Nat Rev Genet* **10**, 691-703, doi:10.1038/nrg2640 (2009).

23 Belancio, V. P., Hedges, D. J. & Deininger, P. Mammalian non-LTR retrotransposons: for better or worse, in sickness and in health. *Genome Res* **18**, 343-358, doi:10.1101/gr.5558208 (2008).

24 Dewannieux, M., Esnault, C. & Heidmann, T. LINE-mediated retrotransposition of marked Alu sequences. *Nat Genet* **35**, 41-48, doi:10.1038/ng1223 (2003).

25 Ray, D. A. *et al.* Alu insertion loci and platyrrhine primate phylogeny. *Mol Phylogenet Evol* **35**, 117-126, doi:10.1016/j.ympev.2004.10.023 (2005).

26 Ray, D. A. & Batzer, M. A. Tracking Alu evolution in New World primates. *BMC Evol Biol* **5**, 51, doi:10.1186/1471-2148-5-51 (2005).

27 Bartel, D. P. MicroRNAs: target recognition and regulatory functions. *Cell* **136**, 215-233, doi:10.1016/j.cell.2009.01.002 (2009).

28 Heimberg, A. M., Sempere, L. F., Moy, V. N., Donoghue, P. C. & Peterson, K. J. MicroRNAs and the advent of vertebrate morphological complexity. *Proc Natl Acad Sci U S A* **105**, 2946-2950, doi:10.1073/pnas.0712259105 (2008).

29 Prochnik, S. E., Rokhsar, D. S. & Aboobaker, A. A. Evidence for a microRNA expansion in the bilaterian ancestor. *Dev Genes Evol* **217**, 73-77, doi:10.1007/s00427-006-0116-1 (2007).

30 Sempere, L. F., Cole, C. N., McPeek, M. A. & Peterson, K. J. The phylogenetic distribution of metazoan microRNAs: insights into evolutionary complexity and constraint. *J Exp Zool B Mol Dev Evol* **306**, 575-588, doi:10.1002/jez.b.21118 (2006).

31 Bartel, D. P. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* **116**, 281-297, doi:S0092867404000455 [pii] (2004).

32 Lee, R. C., Feinbaum, R. L. & Ambros, V. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell* **75**, 843-854, doi:0092-8674(93)90529-Y [pii] (1993).

33 Friedman, R. C., Farh, K. K., Burge, C. B. & Bartel, D. P. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res* **19**, 92-105, doi:10.1101/gr.082701.108 (2009).

34 Grimson, A. *et al.* MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Mol Cell* **27**, 91-105, doi:10.1016/j.molcel.2007.06.017 (2007).

35 Lewis, B. P., Shih, I. H., Jones-Rhoades, M. W., Bartel, D. P. & Burge, C. B. Prediction of mammalian microRNA targets. *Cell* **115**, 787-798, doi:S0092867403010183 [pii] (2003).

36 Miller, W. *et al.* 28-way vertebrate alignment and conservation track in the UCSC Genome Browser. *Genome Res* **17**, 1797-1808, doi:10.1101/gr.6761107 (2007).

37 Berezikov, E. *et al.* Diversity of microRNAs in human and chimpanzee brain. *Nat Genet* **38**, 1375-1377, doi:10.1038/ng1914 (2006).

38 Zhang, R., Wang, Y. Q. & Su, B. Molecular evolution of a primate-specific microRNA family. *Mol Biol Evol* **25**, 1493-1502, doi:10.1093/molbev/msn094 (2008).

39 Barrett, T. *et al.* NCBI GEO: archive for functional genomics data sets--10 years on. *Nucleic Acids Res* **39**, D1005-1010, doi:10.1093/nar/gkq1184 (2011).

40 Edgar, R., Domrachev, M. & Lash, A. E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* **30**, 207-210 (2002).

41 Charrier, C. *et al.* Inhibition of SRGAP2 function by its human-specific paralogs induces neoteny during spine maturation. *Cell* **149**, 923-935, doi:10.1016/j.cell.2012.03.034 (2012).

42 Paten, B., Herrero, J., Beal, K., Fitzgerald, S. & Birney, E. Enredo and Pecan: genome-wide mammalian consistency-based multiple alignment with paralogs. *Genome Res* **18**, 1814-1828, doi:10.1101/gr.076554.108 (2008).

43 Clason, T. *et al.* The structure of eukaryotic and prokaryotic complex I. *J Struct Biol* **169**, 81-88, doi:10.1016/j.jsb.2009.08.017 (2010).

44 Forbes, B. E. Molecular mechanisms underlying insulin-like growth factor action: How mutations in the GH: IGF axis lead to short stature. *Pediatr Endocrinol Rev* **8**, 374-381 (2011).

45 Yang, Z., Wong, W. S. & Nielsen, R. Bayes empirical bayes inference of amino acid sites under positive selection. *Mol Biol Evol* **22**, 1107-1118, doi:10.1093/molbev/msi097 (2005).

46 Garrett, T. P. *et al.* Crystal structure of the first three domains of the type-1 insulin-like growth factor receptor. *Nature* **394**, 395-399, doi:10.1038/28668 (1998).

47 Haig, D. What is a marmoset? *American Journal of Primatology* **49**, 285-296 (1999).

48 Jongens, T. A., Hay, B., Jan, L. Y. & Jan, Y. N. The germ cell-less gene product: a posteriorly localized component necessary for germ cell development in Drosophila. *Cell* **70**, 569-584, doi:0092-8674(92)90427-E [pii] (1992).

49 Mansfield, K. Marmoset models commonly used in biomedical research. *Comp Med* **53**, 383-392 (2003).

50 Gaggar, A., Shayakhmetov, D. M. & Lieber, A. CD46 is a cellular receptor for group B adenoviruses. *Nat Med* **9**, 1408-1412, doi:10.1038/nm952 (2003).

51 Ford, S. M. Callitrichids as phyletic dwarfs and the place of the Callitrichidae in Platyrrhini. *Primates* **21**, 31-43 (1980).

52 Marroig, G. & Cheverud, J. M. Size as a line of least evolutionary resistance: diet and adaptive morphological radiation in New World monkeys. *Evolution* **59**, 1128-1142 (2005).

53 Wallis, M. New insulin-like growth factor (IGF)-precursor sequences from mammalian genomes: the molecular evolution of IGFs and associated peptides in primates. *Growth Horm IGF Res* **19**, 12-23, doi:10.1016/j.ghir.2008.05.001 (2009).

54 Howard, A. D. *et al.* A receptor in pituitary and hypothalamus that functions in growth hormone release. *Science* **273**, 974-977 (1996).

55 Liu, G., Fortin, J. P., Beinborn, M. & Kopin, A. S. Four missense mutations in the ghrelin receptor result in distinct pharmacological abnormalities. *J Pharmacol Exp Ther* **322**, 1036-1043, doi:10.1124/jpet.107.123141 (2007).

56 Pantel, J. *et al.* Loss of constitutive activity of the growth hormone secretagogue receptor in familial short stature. *J Clin Invest* **116**, 760-768, doi:10.1172/JCI25303 (2006).

57 Pantel, J. *et al.* Recessive isolated growth hormone deficiency and mutations in the ghrelin receptor. *The Journal of clinical endocrinology and metabolism* **94**, 4334-4341, doi:10.1210/jc.2009-1327 (2009).

58 Wang, H. J. *et al.* Ghrelin receptor gene: identification of several sequence variants in extremely obese children and adolescents, healthy normal-weight and underweight students, and children with short normal stature. *The Journal of clinical endocrinology and metabolism* **89**, 157-162 (2004).

59 Abuzzahab, M. J. *et al.* IGF-I receptor mutations resulting in intrauterine and postnatal growth retardation. *N Engl J Med* **349**, 2211-2222, doi:10.1056/NEJMoa010107 (2003).

60 Inagaki, K. *et al.* A familial insulin-like growth factor-I receptor mutant leads to short stature: clinical and biochemical characterization. *The Journal of clinical endocrinology and metabolism* **92**, 1542-1548, doi:10.1210/jc.2006-2354 (2007).

61 Plavcan, J. M. & Gomez, A. M. Relative tooth size and dwarfing in callitrichines. *J. Hum. Evol.* **25**, 241-245 (1993).

62 Forbes, K., Westwood, M., Baker, P. N. & Aplin, J. D. Insulin-like growth factor I and II regulate the life cycle of trophoblast in the developing human placenta. *Am J Physiol Cell Physiol* **294**, C1313-1322, doi:10.1152/ajpcell.00035.2008 (2008).

63 Smith, C. A. & Moore, H. D. An ultrastructural study of early chorionic villus formation in the marmoset monkey (Callithrix jacchus). *Anat Embryol (Berl)* **181**, 59-66 (1990).

64 Lopez-Otin, C. & Overall, C. M. Protease degradomics: a new challenge for proteomics. *Nat Rev Mol Cell Biol* **3**, 509-519, doi:10.1038/nrm858 (2002).

65 Quesada, V., Ordonez, G. R., Sanchez, L. M., Puente, X. S. & Lopez-Otin, C. The Degradome database: mammalian proteases and diseases of proteolysis. *Nucleic acids research* **37**, D239-243, doi:10.1093/nar/gkn570 (2009).

66 Ordonez, G. R., Puente, X. S., Quesada, V. & Lopez-Otin, C. Proteolytic systems: constructing degradomes. *Methods in molecular biology* **539**, 33-47, doi:10.1007/978-1-60327-003-8_2 (2009).

67 Ordonez, G. R. *et al.* Loss of genes implicated in gastric function during platypus evolution. *Genome biology* **9**, R81, doi:10.1186/gb-2008-9-5-r81 (2008).

68 Puente, X. S. & Lopez-Otin, C. A genomic analysis of rat proteases and protease inhibitors. *Genome research* **14**, 609-622, doi:10.1101/gr.1946304 (2004).

69 Puente, X. S., Sanchez, L. M., Overall, C. M. & Lopez-Otin, C. Human and mouse proteases: a comparative genomic approach. *Nat Rev Genet* **4**, 544-558, doi:10.1038/nrg1111 (2003).

70 Quesada, V., Velasco, G., Puente, X. S., Warren, W. C. & Lopez-Otin, C. Comparative genomic analysis of the zebra finch degradome provides new insights into evolution of proteases in birds and mammals. *BMC genomics* **11**, 220, doi:10.1186/1471-2164-11-220 (2010).

71 Puente, X. S., Gutierrez-Fernandez, A., Ordonez, G. R., Hillier, L. W. & Lopez-Otin, C. Comparative genomic analysis of human and chimpanzee proteases. *Genomics* **86**, 638-647, doi:10.1016/j.ygeno.2005.07.009 (2005).

72 Lopez-Otin, C. & Bond, J. S. Proteases: multifunctional enzymes in life and disease. *The Journal of biological chemistry* **283**, 30433-30437, doi:10.1074/jbc.R800035200 (2008).

73 Gibbs, R. A. *et al.* Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428**, 493-521, doi:10.1038/nature02426 (2004).

74 Warren, W. C. *et al.* The genome of a songbird. *Nature* **464**, 757-762, doi:10.1038/nature08819 (2010).

75 Warren, W. C. *et al.* Genome analysis of the platypus reveals unique signatures of evolution. *Nature* **453**, 175-183, doi:10.1038/nature06936 (2008).

76 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. *Nature genetics* **36**, 1326-1329, doi:10.1038/ng1471 (2004).

77 Valtonen-Andre, C., Olsson, A. Y., Nayudu, P. L. & Lundwall, A. Ejaculates from the common marmoset (Callithrix jacchus) contain semenogelin and beta-microseminoprotein but not prostate-specific antigen. *Mol Reprod Dev* **71**, 247-255, doi:10.1002/mrd.20257 (2005).

78 Chen, C., Darrow, A. L., Qi, J. S., D'Andrea, M. R. & Andrade-Gordon, P. A novel serine protease predominately expressed in macrophages. *Biochem J* **374**, 97-107, doi:10.1042/BJ20030242 (2003).

79 Saleh, M. *et al.* Differential modulation of endotoxin responsiveness by human caspase-12 polymorphisms. *Nature* **429**, 75-79, doi:10.1038/nature02451 (2004).

80 Brass, A. L. *et al.* Identification of host proteins required for HIV infection through a functional genomic screen. *Science* **319**, 921-926, doi:10.1126/science.1152725 (2008).

81 Pendas, A. M. *et al.* Diet-induced obesity and reduced skin cancer susceptibility in matrix metalloproteinase 19-deficient mice. *Mol Cell Biol* **24**, 5304-5313, doi:10.1128/MCB.24.12.5304-5313.2004 (2004).

82 Mauch, S., Kolb, C., Kolb, B., Sadowski, T. & Sedlacek, R. Matrix metalloproteinase-19 is expressed in myeloid cells in an adhesion-dependent manner and associates with the cell surface. *Journal of immunology* **168**, 1244-1251 (2002).

83 Baudat, F. *et al.* PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. *Science* **327**, 836-840, doi:10.1126/science.1183439 (2010).

84 Cheung, V. G., Sherman, S. L. & Feingold, E. Genetics. Genetic control of hotspots. *Science* **327**, 791-792, doi:10.1126/science.1187155 (2010).

85 Parvanov, E. D., Petkov, P. M. & Paigen, K. Prdm9 controls activation of mammalian recombination hotspots. *Science* **327**, 835, doi:10.1126/science.1181495 (2010).

86 Katoh, K., Asimenos, G. & Toh, H. Multiple alignment of DNA sequences with MAFFT. *Methods Mol Biol* **537**, 39-64, doi:10.1007/978-1-59745-251-9_3 (2009).

87 Tamura, K. *et al.* MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol* **28**, 2731-2739, doi:10.1093/molbev/msr121 (2011).