Supplemental Information

Supplemental Figures

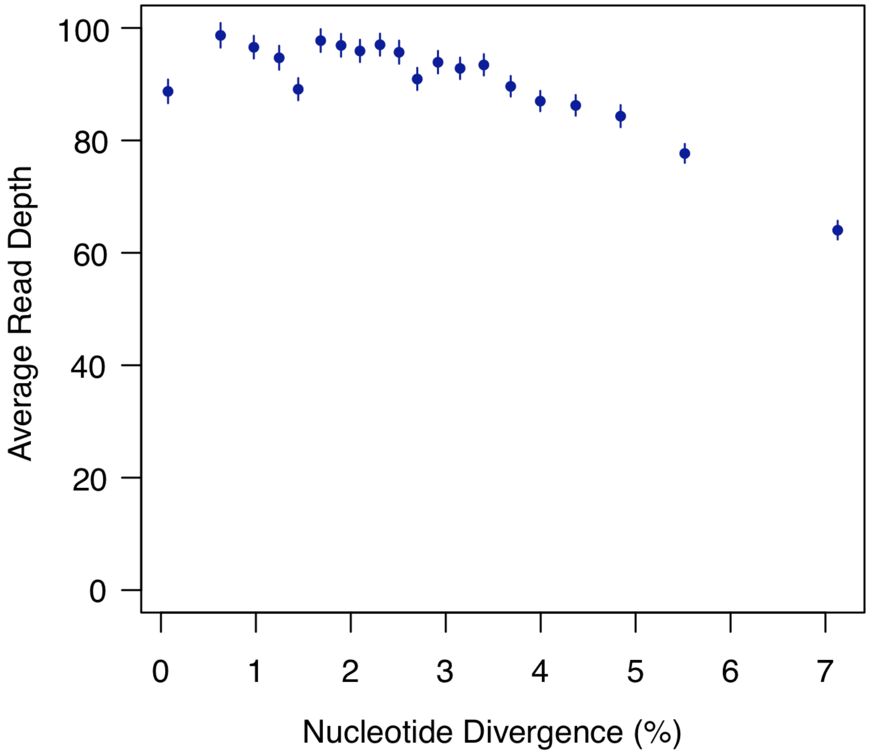

Figure S1. Average macaque read depth of captured targets versus human-macaque sequence divergence. We calculated sequence divergence between the human and macaque reference genome for 134,401 orthologous targets, which contained no indels. We placed targets into 20 bins of equal size based on their human-macaque sequence divergence and then calculated the mean macaque read depth of the targets within each bin.

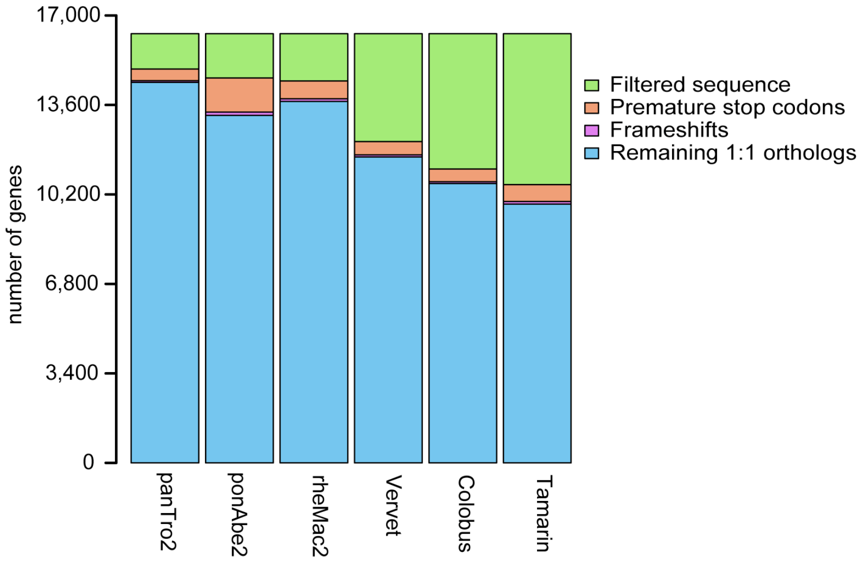

Figure S2. Filtering of orthologous gene alignments. This figure is a summary of the ortholog filtering described in the methods.

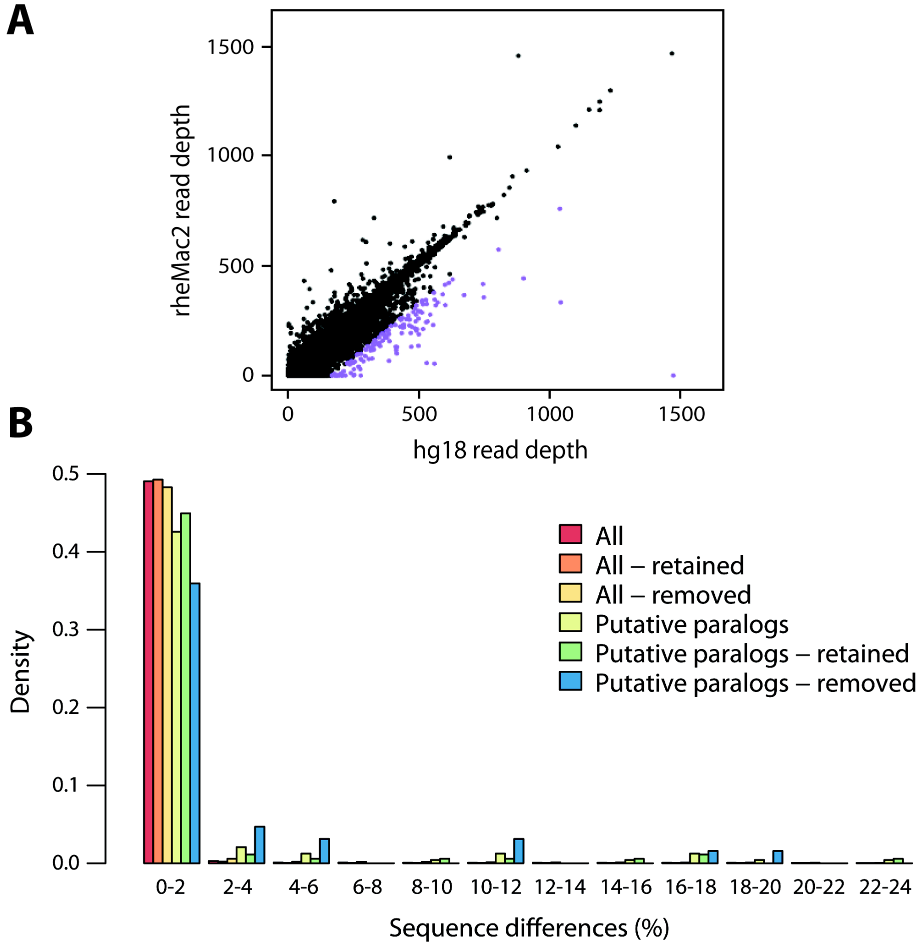

Figure S3. Identification of putative paralogous targeted regions and sequence differences between assembled and reference macaque sequences. (A) We identified putative paralogous captured targets that may be susceptible to mis-assembly by comparing the depth of macaque reads mapped to the macaque reference genome (rheMac2) to the depth of the same reads mapped to the human reference genome (hg18). We consider as putative paralogs (purple), the 137 targets where the hg18 read depth is at least two standard deviations greater than the rheMac2 read depth. (B) Histogram of nucleotide differences between the macaque assembled exome and the macaque reference genome for 153,546 targeted regions that we assembled and uniquely mapped to the macaque reference genome using cross_match (v1.090518, http://www.phrap.org). Targets were further categorized by whether they were putative paralogs and by whether

they are retained or removed following filtering for segmental duplications, extreme heterozygosity and missing sequence (see Methods for more details).

Supplemental Tables

Table S1. Summary of read merging and mapping. Summary of read merging and mapping for each non-human primate exome and two human HapMap exomes (Human 1: NA12878 and Human 2: NA18967). Listed for each exome are the number of paired-end 76 bp reads (PE76) generated, the number of overlapping read pairs merged into single longer reads, the number of read pairs discarded due to gaps in the overlapping portion, the total number of individual reads after read merging, the number of individual reads discarded due to low quality (>10% Ns), the number of reads uniquely mapped to the repeat masked human reference genome with cross_match (v1.090518, http://www.phrap.org) and the number of uniquely mapped reads remaining (and used for assembly of exomes) after filtering for PCR and optical duplicates.

Sample	**PE76 read pairs**	**Merged PE76 read pairs**	**Discarded PE76 read pairs**	**Total individual reads**	**Discarded low quality**	**Uniquely mapped**	**After duplicate filtering**
Human 1	37,520,750	20,945,392	142,933	53,810,242	628,917	47,615,608	40,589,744
Human 2	47,628,042	25,921,661	337,984	68,658,455	2,821,253	52,136,229	47,538,474
Macaque	46,373,786	22,949,875	356,160	69,085,377	929,405	55,206,587	44,741,572
Vervet	47,568,368	20,979,776	301,513	73,553,934	1,300,948	57,365,966	45,683,191
Colobus	46,834,936	22,969,487	357,702	69,984,981	1,732,508	55,369,898	43,484,635
Tamarin	46,044,421	21,469,749	390,156	69,838,781	1,429,809	52,555,753	42,067,032

Table S2. Assembly statistics. Summary of phrap assemblies for each non-human primate exome and two human HapMap exomes (Human 1: NA12878 and Human 2: NA18967). Listed for each exome are the total number of groups of overlapping reads, the number of overlap groups containing more than one read that were assembled using phrap (v1.090518, http://www.phrap.org), the number of assembled contigs, the average length of the assembled contigs, the number of assembled contigs that mapped uniquely to the repeat masked human reference genome with cross_match (v1.090518, http://www.phrap.org) and the number of discarded contigs that mapped to a different location than their individual reads (off location contigs).

Sample	Overlap groups	Groups assembled	Contigs	Avg. contig length (bp)	Mapped contigs	Off location contigs
Human 1	1,988,564	660,218	590,549	214	570,036	13,284
Human 2	2,644,570	981,000	902,670	170	884,566	16,244
Macaque	1,611,560	593,794	592,954	215	570,219	30,016
Vervet	1,631,204	606,930	604,158	224	578,120	28,221
Colobus	1,565,107	637,026	642,079	207	617,462	29,884
Tamarin	1,485,380	569,441	727,059	210	684,602	33,741

Table S3. Linear regression model of macaque read depth. To assess what factors influence captured target read depth, we fit a linear model using observations from 128,914 captured target regions. The response variable, macaque read depth, is in units of reads/base. $R^2 = 0.674$.

β – Slope estimate. Standard errors of the slope estimates are in parentheses.

β^* – Normalized slope estimate. Predictor variables are normalized to have mean 0 and standard deviation 1 so that the slope estimates are comparable. The response variable, macaque read depth, is not normalized. Standard errors of the slope estimates are in parentheses.

p – *p*-value from a two-sided *t*-test with null $\beta = 0$.

Macaque mappability – The number of 76 bp simulated macaque reads uniquely mapped to each base in the human target sequence.

Predictor	$\boldsymbol{\beta}$	$\boldsymbol{\beta^*}$	***p*-value**
(Intercept)	39.7 (1.21)	92.1 (0.13)	$< 2.0 \times 10^{-16}$
Human read depth (reads/base)	0.93 (0.0019)	67.2 (0.14)	$< 2.0 \times 10^{-16}$
Nucleotide differences (%)	-7.72 (0.080)	-13.3 (0.14)	$< 2.0 \times 10^{-16}$
Indels (%)	-4.18 (0.22)	-2.54 (0.13)	$< 2.0 \times 10^{-16}$
Macaque mappability (reads/base)	-0.060 (0.0093)	-0.89 (0.14)	8.46×10^{-11}
GC content (%)	7.10 (1.29)	0.80 (0.14)	3.87×10^{-08}

Table S4. Genomic features of captured and not captured macaque targets. We examined the following genomic features for the 155,707 human targets with best-reciprocal orthologs in the macaque genome: the number of nucleotide differences between human and macaque, the number of indel bases between human and macaque and the GC content. A target was considered "captured" if more than half of the human targeted bases were covered by at least one sequencing read.

	No. targets	**No. bases**	**Differences (%)**	**Indel bases (%)**	**GC (%)**
Captured targets	154,596	29,479,474	3.07	0.169	49.9
Not captured targets	1,111	162,361	3.92	0.500	51.7

Table S5. High quality coding sequence differences relative to the human reference genome. Coding sequence differences relative to the human reference genome calculated from high quality (≥Q40) sequence for each assembled exome and for the non-human primate reference genomes of chimpanzee (panTro2), orangutan (ponAbe2) and rhesus macaque (rheMac2). Coding sequences are non-overlapping transcripts (longest transcript retained from overlapping transcripts) from the 20080430 version of the CCDS database (Pruitt et al. 2009) totaling 27,583,228 bp. 2,655,850 bp of this sequence was not targeted by our capture method and contributes to the difference between the number of assembled exome bases and the number of bases from the non-human primate reference genomes. Heterozygous sites and sites overlapping known segmental duplications were excluded from all species' sequences. Exons with excess heterozygosity, low read depth or more than half their sequence filtered were removed from the exome assemblies.

Common ≥Q40 sites – 9,106,235 sites that are high quality in all species.

	All sites			**Common ≥Q40 sites**	
Species	**≥Q40 consensus (bp)**	**≥Q40 differences (bp)**	**Difference (%)**	**≥Q40 differences (bp)**	**Difference (%)**
panTro2	23,904,584	128,495	0.54	42,482	0.47
ponAbe2	23,139,970	329,972	1.43	113,938	1.25
rheMac2	21,925,828	574,515	2.62	204,446	2.25
Macaque	17,459,375	439,792	2.52	203,818	2.24
Vervet	18,185,119	465,882	2.56	208,018	2.28
Colobus	16,197,614	428,758	2.65	213,885	2.35
Tamarin	15,346,639	642,281	4.19	353,732	3.88

Table S6. Summary of high quality coding indel lengths. This table summarizes the number of indels with lengths that are multiples of three (3*n*) in gene alignments that contain greater than 75% high quality sequence in all species. Low quality indels were only included if their read depth was sufficiently high (≥4) or they were confirmed in another species. Indels less than 15 bp apart were combined to account for uncertainty in the alignment.

Gaps in human – regions of the alignments causing a gap in the human sequence; appears as an insertion in the other species.

Gaps in other species – regions of the alignments causing a gap in the other sequence; appears as a deletion in the other species.

	Gaps in human			**Gaps in other species**		
Species	**Total**	**3*n***	**3*n* (%)**	**Total**	**3*n***	**3*n* (%)**
panTro2	152	121	79.6%	183	123	67.2%
ponAbe2	442	234	52.9%	471	207	43.9%
rheMac2	452	384	85.0%	487	391	80.3%
Macaque	373	290	77.7%	417	341	81.8%
Vervet	376	309	82.2%	449	357	79.5%
Colobus	359	295	82.2%	480	370	77.1%
Tamarin	584	487	83.4%	904	690	76.3%

Table S7. Numbers of genes showing evidence of positive selection at several FDR thresholds.

1% FDR	5% FDR	10% FDR
52	93	157

Table S8. Complete list of genes tested for evidence of positive selection in primates ranked by significance. Table is provided as a separate supplemental file. For each gene is shown the number of species, the nominal *p*-value from a chi-square approximated likelihood ratio test between CODEML's M7 and M8 models (Yang 2007), the estimated false discovery rate calculated by *q*-values (Storey and Tibshirani 2003) and the average F_{ST} (Tennessen et al. 2010).

dNdS_REF – A '1' indicates a gene previously identified as subject to positive selection from the Rhesus Macaque Sequencing and Analysis Consortium (Rhesus Macaque Genome Sequencing and Analysis Consortium 2007), which used a similar d_N/d_S method and sequences from human, chimpanzee and macaque.

Table S9. **Genes identified by the Rhesus Macaque Genome Sequencing and Analysis Consortium, but not identified by our study.** Shown for each gene is the nominal *p*-value determined from a similar analysis using human, chimpanzee and macaque sequences (Rhesus Macaque Genome Sequencing and Analysis Consortium 2007) (*), the nominal *p*-value and *q*-value (Storey and Tibshirani 2003) from our analysis and the number of species used in our analysis. Genes are ranked by their significance in the other study. The "Note" column indicates genes that were not tested in our analysis because we could not confidently obtain enough sequences (1), the models in CODEML did not converge (2), or they were not targeted by our probe set (3). In total, the other study identified 67 genes under positive selection at an FDR of 10%, 15 of which we also identified at the same FDR threshold.

CCDS	Gene name	Chr	*p*-value*	*p*-value	*q*-value	No. species	Rank	Note
41683.1	*KRTAP5-8*	11	6.20E-16	7.45E-03	3.01E-01	3	361	
42614.1	*LILRB1*	19	7.20E-14			1		1
11896.1	*DSG1*	18	1.10E-10	4.52E-03	2.33E-01	3	280	
14217.1	*MAGEB6*	X	5.30E-08			2		1
7831.1	*MRGPRX4*	11	5.60E-08			2		1
	COL6A5 (FLJ35880)	3	1.70E-07					3
	LOC442247	6	3.80E-07					3
5007.1	*CGA*	6	1.20E-06	1.01E-02	3.52E-01	7	408	
	KRTAP5-4	11	2.70E-06					3
12231.1	*ICAM1*	19	2.70E-06	2.24E-03	1.61E-01	5	201	
	NA_1024667	1	4.50E-06					3
	CCDC45	17	4.90E-06					3
4931.1	*CRISP1*	6	1.60E-05	1.66E-03	1.33E-01	7	184	
7973.1	*FAM111A*	11	2.80E-05					2
12891.1	*LAIR1*	19	3.10E-05			1		1
4854.1	*TREM1*	6	6.30E-05					2
7972.1	*FAM111B*	11	1.30E-04	4.52E-03	2.33E-01	3	285	
3785.1	*DCHS2 (AK123368)*	4	1.30E-04			1		1
31672.1	*C11orf87 (LOC399947)*	11	1.30E-04	1.0	9.87E-01	7	8,821	
31685.1	*CD3E*	11	1.30E-04	2.73E-02	5.79E-01	7	675	
1385.1	*CFH*	1	1.50E-04	6.47E-03	2.87E-01	4	347	
	TCRA	14	1.50E-04					3
	OTTHUMT00000004245	1	1.50E-04					3

6507.1	*IFNA8*	9	1.50E-04	9.07E-02	9.87E-01	6	1,276	
	TGOLN2	16	1.80E-04					3
31168.1	*PDSS1*	10	1.80E-04	3.33E-01	9.87E-01	7	2,811	
7753.1	*HBB*	11	2.00E-04	5.49E-01	9.87E-01	5	3,919	
7854.1	*SLC6A5*	11	2.30E-04			1		1
2717.1	*ZNF197*	3	3.40E-04	1.83E-02	4.83E-01	3	542	
42475.1	*ZNRF4*	19	3.90E-04	8.19E-01	9.87E-01	4	5,813	
33522.1	*MRPL39*	21	4.00E-04	4.0E-04	9.87E-01	6	12,698	
5042.1	*COQ3 (RP11-98I9.1-002)*	6	4.00E-04	1.83E-01	9.87E-01	7	1,922	
32792.1	*CEP192 (AB051446)*	18	4.30E-04			1		1
31080.1	*FAM36A*	1	4.30E-04	1.66E-02	4.55E-01	4	528	
33747.1	*LTF*	3	4.90E-04	3.03E-03	1.89E-01	4	226	
	CCDC129	7	4.90E-04					3
2932.1	*CPOX*	3	4.90E-04	2.74E-03	1.81E-01	7	219	
8840.1	*KRT78*	12	5.00E-04	9.05E-01	9.87E-01	7	6,635	
5806.1	*OPN1SW*	7	5.80E-04	5.50E-02	8.33E-01	7	959	
13983.1	*APOBEC3C*	22	5.90E-04	2.24E-02	5.27E-01	5	623	
	RP1-321E8.3-001	X	6.00E-04					3
	KIAA1731 (AB051518)	11	6.80E-04					3
3419.1	*FGFBP2 (KSP37)*	4	6.80E-04	3.33E-01	9.87E-01	6	2,750	
3278.1	*AHSG*	3	6.90E-04					2
4470.1	*HUS1B*	6	6.90E-04	4.08E-02	7.17E-01	4	831	
434.1	*NDUFS5*	1	7.10E-04	1.11E-02	3.71E-01	5	440	
3143.1	*TM4SF1*	3	7.30E+04	6.74E-03	2.87E-01	7	346	
	ADAM32	8	7.60E-04					3
11799.1	*DCXR*	17	8.10E-04	1.11E-02	3.71E-01	7	433	
5959.1	*DEFB1*	8	8.30E-04	1.11E-03	1.04E-01	7	158	
35001.1	*C9orf11*	9	8.30E-04	2.73E-01	9.87E-01	5	2,479	
10276.1	*C15orf39*	15	8.30E-04	3.68E-01	9.87E-01	6	3,022	

Table S10. GO categories enriched for genes predicted to be under positive selection. 13,838 of 15,027 genes tested for positive selection were assigned to UniProt identifiers and used to identify GO categories enriched for genes predicted to be under positive selection. Shown are the numbers of genes assigned to each biological process category, the numbers of genes in a null distribution consisting of all genes except those assigned to the term being tested, the numbers of tests performed and the nominal *p*-values from a one-sided Mann-Whitney U test. Bolded *p*-values are significant after a conservative Bonferroni correction for multiple testing (*p*-value < 0.05). Only categories with a nominal *p*-value less than 0.05 are reported.

GO ID	Biological process	No. tests	Genes in category	Genes in null	*p*-value
GO:0006952	defense response	1,681	476	10,888	**4.03E-14**
GO:0031424	keratinization	1,526	36	10,412	**3.50E-09**
GO:0007606	sensory perception of chemical stimulus	1,523	259	10,376	**2.40E-08**
GO:0055114	oxidation reduction	1,517	510	10,117	5.00E-05
GO:0019882	antigen processing and presentation	1,445	33	9,607	7.80E-05
GO:0046483	heterocycle metabolic process	1,441	227	9,574	7.96E-04
GO:0015698	inorganic anion transport	1,364	41	9,347	6.76E-04
GO:0030193	regulation of blood coagulation	1,363	24	9,306	8.76E-04
GO:0044243	multicellular organismal catabolic process	1,346	22	9,282	9.95E-04
GO:0007586	digestion	1,337	29	9,260	1.72E-03
GO:0042254	ribosome biogenesis	1,333	23	9,231	1.87E-03
GO:0007283	spermatogenesis	1,331	214	9,208	2.36E-03
GO:0007600	sensory perception	1,291	208	8,994	1.26E-03
GO:0032368	regulation of lipid transport	1,270	24	8,786	2.22E-03
GO:0050731	positive regulation of peptidyl-tyrosine phosphorylation	1,255	34	8,762	1.75E-03
GO:0006974	response to DNA damage stimulus	1,219	247	8,728	3.28E-03
GO:0006955	immune response	1,177	137	8,481	4.95E-03
GO:0008544	epidermis development	1,146	69	8,344	7.06E-03
GO:0009566	fertilization	1,124	25	8,275	1.25E-02
GO:0060249	anatomical structure homeostasis	1,121	45	8,250	1.72E-02
GO:0032886	regulation of microtubule-based process	1,098	30	8,205	8.38E-03
GO:0006869	lipid transport	1,088	70	8,175	1.77E-02
GO:0006725	cellular aromatic compound metabolic process	1,076	23	8,105	1.14E-02
GO:0022610	biological adhesion	1,072	396	8,082	1.54E-02
GO:0030198	extracellular matrix organization	1,028	36	7,686	4.17E-03
GO:0031960	response to corticosteroid stimulus	1,016	45	7,650	9.87E-03

GO:0006732	coenzyme metabolic process	992	50	7,605	1.15E-02
GO:0001775	cell activation	983	75	7,555	2.01E-02
GO:0010562	positive regulation of phosphorus metabolic process	966	36	7,480	1.33E-02
GO:0008643	carbohydrate transport	937	39	7,444	1.74E-02
GO:0034097	response to cytokine stimulus	931	31	7,405	1.70E-02
GO:0007010	cytoskeleton organization	915	227	7,317	4.39E-02
GO:0016042	lipid catabolic process	857	69	7,090	2.76E-02
GO:0055085	transmembrane transport	849	343	7,021	4.65E-02

Table S11. Genes assigned to the GO biological process category, “keratinization”, ranked by statistical evidence of positive selection. Listed below are 36 of the 41 genes assigned to the “keratinization” category in GO (GO:0031424) that were tested for positive selection. The majority of these genes reside in a cluster on chromosome 1. Shown for each gene are the nominal *p*-value from a chi-square approximated likelihood ratio test between CODEML’s M7 and M8 models (Yang 2007), the corresponding *q*-value (Storey and Tibshirani 2003) and the overall rank. For the top 6 (bolded *q*-values), there is statistical evidence for positive selection at an FDR of 10%.

CCDS	Gene Name	Chr	Description	No. Species	*p*-value	*q*-value	Rank
30866.1	*SPRR2E*	1	small proline-rich protein 2E	7	1.13E-07	**1.28E-04**	13
30867.1	*SPRR2F*	1	small proline-rich protein 2F	6	4.56E-07	**3.98E-04**	17
1030.1	*IVL*	1	involucrin	4	6.81E-07	**5.32E-04**	19
1015.1	*LCE3C*	1	late cornified envelope 3C	6	2.04E-05	**6.73E-03**	44
30865.1	*SPRR2B*	1	small proline-rich protein 2B	6	6.13E-05	**1.62E-02**	56
11737.1	*EVPL*	17	envoplakin	7	3.71E-04	**5.73E-02**	95
1032.1	*SPRR1A*	1	small proline-rich protein 1A	6	1.11E-03	1.04E-01	159
1020.1	*LCE2B*	1	late cornified envelope 2B	4	2.24E-03	1.61E-01	204
1033.1	*SPRR3*	1	small proline-rich protein 3	6	2.48E-03	1.70E-01	213
30870.1	*LOR*	1	loricrin	4	2.47E-02	5.56E-01	638
1031.1	*SPRR4*	1	small proline-rich protein 4	6	2.47E-02	5.56E-01	643
1021.1	*LCE2A*	1	late cornified envelope 2A	3	3.34E-02	6.50E-01	747
1014.1	*LCE3D*	1	late cornified envelope 3D	5	8.21E-02	9.87E-01	1,226
1026.1	*LCE1C*	1	late cornified envelope 1C	4	1.35E-01	9.87E-01	1,573
30864.1	*SPRR2D*	1	small proline-rich protein 2D	6	1.65E-01	9.87E-01	1,825
33435.1	*TGM3*	20	transglutaminase 3	7	2.02E-01	9.87E-01	1,974
1024.1	*LCE1E*	1	late cornified envelope 1E	5	3.01E-01	9.87E-01	2,595
10526.1	*PPL*	16	periplakin	4	3.01E-01	9.87E-01	2,645
1034.1	*SPRR2A*	1	small proline-rich protein 2A	6	4.07E-01	9.87E-01	3,208
1011.1	*LCE5A*	1	late cornified envelope 5A	4	4.49E-01	9.87E-01	3,392
30863.1	*SPRR1B*	1	small proline-rich protein 1B (cornifin)	6	4.49E-01	9.87E-01	3,393
8835.1	*KRT2*	12	keratin 2	5	6.07E-01	9.87E-01	4,160
1025.1	*LCE1D*	1	late cornified envelope 1D	4	6.70E-01	9.87E-01	4,685

1019.1	*LCE2C*	1	late cornified envelope 2C	4	6.70E-01	9.87E-01	4,704
1017.1	*LCE3A*	1	late cornified envelope 3A	5	6.70E-01	9.87E-01	4,726
1016.1	*LCE3B*	1	late cornified envelope 3B	7	8.19E-01	9.87E-01	5,592
1013.1	*LCE3E*	1	late cornified envelope 3E	7	9.05E-01	9.87E-01	6,225
1018.1	*LCE2D*	1	late cornified envelope 2D	4	9.05E-01	9.87E-01	6,289
1027.1	*LCE1B*	1	late cornified envelope 1B	5	9.05E-01	9.87E-01	6,314
43777.1	*SHARPIN*	8	SHANK-associated RH domain interactor	4	9.05E-01	9.87E-01	6,650
9622.1	*TGM1*	14	transglutaminase 1	6	1.0	9.87E-01	7,436
12606.1	*CNFN*	19	cornifelin	6	1.0	9.87E-01	9,232
1023.1	*LCE1F*	1	late cornified envelope 1F	4	1.0	9.87E-01	9,754
30868.1	*SPRR2G*	1	small proline-rich protein 2G	5	1.0	9.87E-01	9,950
31777.1	*PPHLN1*	12	periphilin 1	6	1.0	9.87E-01	10,567
1028.1	*LCE1A*	1	late cornified envelope 1A	3	1.0	9.87E-01	12,504

Table S12. GO categories enriched for genes with increased human population differentiation. The top ten GO categories enriched for genes under positive selection in primates were also tested for higher levels of population differentiation, as measured by F_{ST} between European and African populations (Tennessen et al. 2010). Shown are the number of genes assigned to each biological process category, the number of genes in a null distribution consisting of all genes except those assigned to the term being tested, the average F_{ST} for each group and the nominal *p*-values from a one-sided Mann-Whitney U test. Bolded *p*-values are significant after a Bonferroni correction for multiple testing (*p*-value < 0.05).

GO ID	Biological process	Genes in cat.	Genes in null	F_{ST} in cat.	F_{ST} in null	*p*-value
GO:0006952	defense response	448	13,809	0.0778	0.0710	0.0161
GO:0031424	keratinization	36	14,221	0.128	0.0711	0.0550
GO:0007606	sensory perception of chemical stimulus	253	14,004	0.0961	0.0708	**2.97E-13**
GO:0055114	oxidation reduction	496	13,761	0.0840	0.0707	**8.24E-04**
GO:0019882	antigen processing and presentation	37	14,220	0.103	0.0711	0.0574
GO:0015698	inorganic anion transport	41	14,216	0.0733	0.0712	0.394
GO:0046483	heterocycle metabolic process	259	13,998	0.0619	0.0714	0.858
GO:0030193	regulation of blood coagulation	35	14,222	0.0664	0.712	0.567
GO:0044243	multicellular organismal catabolic process	23	14,234	0.0791	0.0712	0.0975
GO:0007600	sensory perception	491	13,766	0.0837	0.0708	**5.80E-07**

Table S13. Numbers of genes showing evidence for positive selection on lineages of the phylogeny. Shown for each branch of the primate phylogeny (Figure 1A) are the sequences required to test that branch, the number of genes tested and the numbers of genes showing evidence for positive selection at several FDR thresholds and at a nominal *p*-value < 0.05. *P*-values were computed from a 50:50 mixture of a chi-square distribution with 1 degree of freedom and a point mass at 0 (Zhang et al. 2005). Note that the branch labeled as "tamarin" combines both the branch leading to tamarin as well as the branch leading to the common ancestor of Apes and Old World moneys.

OWM – sequence required from any single Old World monkey: macaque (rheMac2), vervet or colobus.

APE – sequence required from any single Great Ape: human (hg18), chimpanzee (panTro2) or orangutan (ponAbe2).

			No. significant genes			
Branch	**Sequences required**	**No. genes tested**	**1% FDR**	**5% FDR**	**10% FDR**	***p*-value < 0.05**
panTro2	panTro2, hg18, 1 other	14,250	2	3	6	203
hg18	hg18, panTro2, 1 other	14,230	0	0	0	254
panTro2, hg18	panTro2, hg18, ponAbe2	12,316	0	5	6	256
ponAbe2	ponAbe2, hg18 or panTro2, 1 other	12,906	2	4	8	338
panTro2, hg18, ponAbe2	hg18 or panTro2, ponAbe2, 1 OWM, 1 other	8,863	3	3	8	223
rheMac2	rheMac2, vervet, 1 other	11,005	3	4	4	173
vervet	vervet, rheMac2, 1 other	11,003	1	3	7	168
rheMac2, vervet	rheMac2, vervet, colobus	9,609	0	0	0	126
colobus	colobus, rheMac2 or vervet, 1 other	10,695	2	4	8	245
rheMac2, vervet, colobus	rheMac2 or vervet, colobus, 1 APE, tamarin	9,018	5	5	14	264
tamarin	tamarin, 1 OWM, 1 APE	9,906	6	12	84	708

Table S14. Complete lists of genes tested for evidence of positive selection acting on individual lineages ranked by significance. Tables are provided as a separate supplemental file with a worksheet for each individual branch. For each gene are shown the number of species, the length of the sequence, the nominal *p*-value computed from a 50:50 mixture of a chi-square distribution with 1 degree of freedom and a point mass at 0 (Zhang et al. 2005) and the estimated false discovery rate calculated by *q*-values (Storey and Tibshirani 2003).

Table S15. Top 5 GO categories enriched for genes predicted to be under lineage specific positive selection for each lineage. Genes were assigned to UniProt identifiers and used to identify GO categories enriched for genes predicted to be under positive selection along specific lineages. Shown are the number of genes assigned to each biological process category, the number of genes in a null distribution consisting of all genes except those assigned to the term being tested, the number of tests performed and the nominal *p*-values from a one-sided Mann-Whitney U test. Bolded *p*-values are significant after a conservative Bonferroni correction for multiple testing (*p*-value < 0.05).

GO ID	Biological process	No. tests	Genes in category	Genes in null	*p*-value
panTro2					
GO:0007606	sensory perception of chemical stimulus	1,651	189	10,348	**1.45E-14**
GO:0050900	leukocyte migration	1,644	50	10,159	**8.42E-06**
GO:0031424	keratinization	1,606	35	10,109	5.74E-05
GO:0006323	DNA packaging	1,603	23	10,074	8.18E-05
GO:0044242	cellular lipid catabolic process	1,602	70	10,051	5.38E-04
hg18					
GO:0007606	sensory perception of chemical stimulus	1,648	188	10,329	**1.59E-12**
GO:0000723	telomere maintenance	1,640	24	10,141	1.00E-03
GO:0022411	cellular component disassembly	1,632	32	10,117	1.13E-03
GO:0031424	keratinization	1,628	34	10,085	2.39E-03
GO:0007126	meiosis	1,625	54	10,051	5.14E-03
panTro2, hg18					
GO:0006952	defense response	1,538	383	8,968	**7.22E-08**
GO:0007276	gamete generation	1,405	203	8,585	**2.11E-07**
GO:0031424	keratinization	1,366	31	8,382	6.60E-05
GO:0006120	mitochondrial electron transport, NADH to ubiquinone	1,363	31	8,351	1.05E-04
GO:0007606	sensory perception of chemical stimulus	1,361	97	8,320	8.06E-04
ponAbe2					
GO:0007608	sensory perception of smell	1,566	103	9,374	**1.43E-13**
GO:0031424	keratinization	1,558	32	9,271	**1.10E-07**
GO:0042742	defense response to bacterium	1,554	82	9,239	5.27E-05
GO:0031214	biomineral tissue development	1,500	23	9,157	7.43E-03
GO:0010562	positive regulation of phosphorus metabolic process	1,496	107	9,134	1.14E-02
panTro2, hg18, ponAbe2					
GO:0006952	defense response	1,313	263	6,645	**1.47E-06**
GO:0006818	hydrogen transport	1,212	35	6,382	1.08E-03
GO:0050866	negative regulation of cell activation	1,204	28	6,347	1.60E-03
GO:0051606	detection of stimulus	1,172	51	6,319	1.50E-03

GO:0046942	carboxylic acid transport	1,159	81	6,268	1.48E-03
rheMac2					
GO:0022904	respiratory electron transport chain	1,464	32	8,182	**5.40E-10**
GO:2000021	regulation of ion homeostasis	1,457	47	8,150	**1.55E-05**
GO:0006952	defense response	1,395	355	8,103	5.94E-05
GO:0007606	sensory perception of chemical stimulus	1,296	102	7,748	3.89E-05
GO:0006399	tRNA metabolic process	1,292	79	7,646	9.01E-04
vervet					
GO:0007608	sensory perception of smell	1,463	84	8,180	**1.63E-06**
GO:0006120	mitochondrial electron transport, NADH to ubiquinone	1,459	22	8,096	**1.75E-05**
GO:0006952	defense response	1,457	368	8,074	6.12E-04
GO:0007205	activation of protein kinase C activity by G-protein coupled receptor protein signaling pathway	1,337	21	7,706	5.72E-04
GO:0000075	cell cycle checkpoint	1,322	57	7,685	1.75E-03
rheMac2, vervet					
GO:0006952	defense response	1,366	321	7,167	**4.52E-06**
GO:0015698	inorganic anion transport	1,264	35	6,846	1.37E-03
GO:0006935	chemotaxis	1,259	55	6,811	2.26E-03
GO:0018130	heterocycle biosynthetic process	1,230	37	6,756	2.84E-03
GO:0050818	regulation of coagulation	1,212	20	6,719	5.20E-03
colobus					
GO:0007606	sensory perception of chemical stimulus	1,426	98	7,958	**3.50E-06**
GO:0003018	vascular process in circulatory system	1,419	29	7,860	1.15E-03
GO:0006968	cellular defense response	1,401	32	7,831	1.48E-03
GO:0043038	amino acid activation	1,397	32	7,799	1.45E-03
GO:0048609	reproductive process in a multicellular organism	1,394	230	7,767	1.61E-03
rheMac2, vervet, colobus					
GO:0006952	defense response	1,326	265	6,774	**9.5E-08**
GO:0042129	regulation of T cell proliferation	1,223	22	6,509	2.30E-04
GO:0007586	digestion	1,197	20	6,487	2.78E-03
GO:0071103	DNA conformation change	1,187	28	6,467	3.11E-03
GO:0009451	RNA modification	1,183	27	6,439	2.96E-03
tamarin					
GO:0006955	immune response	1,380	245	7,436	**6.32E-10**
GO:0006631	fatty acid metabolic process	1,291	133	7,191	**9.10E-07**
GO:0007155	cell adhesion	1,232	349	7,058	**8.72E-06**
GO:0031099	regeneration	1,175	40	6,709	1.80E-04
GO:0001775	cell activation	1,151	80	6,669	1.25E-04

Table S16. Sequence coverage of captured miRNAs. Summary of targeted miRNA sequence coverage for each non-human primate exome and two human HapMap exomes (Human 1: NA12879 and Human 2: NA18967). The total size of the captured miRNA target is 48,075 bp and includes ~550 miRNAs. Listed for each exome are the number of bases in the target covered by at least one read, the number of bases assembled and the number of bases assembled with Phred consensus quality score ≥ 40 (Q40; 10^{-4} error rate).

Sample	≥1X coverage (bp)	≥1X coverage (%)	Consensus called (bp)	Consensus called (%)	≥Q40 consensus (bp)	≥Q40 consensus (%)	Avg. coverage
Human 1	43,530	90.6	42,995	89.4	42,336	88.1	91X
Human 2	43,585	90.7	43,312	90.1	42,297	88.0	102X
Macaque	42,086	87.5	40,815	84.9	39,732	82.6	94X
Vervet	42,579	88.6	41,022	85.3	40,063	83.3	93X
Colobus	41,530	86.4	39,955	83.1	38,730	80.6	88X
Tamarin	41,278	85.9	39,070	81.3	36,860	76.7	84X

Table S17. Nucleotide differences and indels between the assembled macaque exome and the macaque reference genome. We calculated the number of nucleotide differences and indels between our assembled macaque targeted sequences and the macaque reference genome for the 153,546 assembled targets that uniquely mapped to the macaque genome using cross_match (v1.090518, http://www.phrap.org). Targets are further categorized by whether they are retained or removed following filtering for segmental duplications, low read depth, extreme heterozygosity and missing sequence (see Methods for more details), and by whether they were putative paralogous targets (see Figure S3). Note that the percent difference for all targets is higher than that reported in the main text (0.253% *vs.* 0.10%) because we used all captured targets which include flanking intronic and miRNA sequences in addition to coding sequences.

	No. targets	**No. bases**	**Differences (%)**	**Indels (%)**
All targets	153,546	27,746,320	0.253	0.0252
All retained targets	122,411	22,921,623	0.228	0.0187
All removed targets	31,135	4,824,697	0.376	0.0560
Putative paralogs	121	21,789	1.55	0.142
Putative retained paralogs	89	16,100	1.07	0.118
Putative removed paralogs	32	5,689	2.90	0.211

Table S18. GO categories enriched for genes that are absent from the macaque exome assembly. 15,827 of 16,707 genes were assigned to UniProt identifiers and used to identify GO categories enriched for genes that are absent from the macaque exome assembly. Shown for each category is the number of genes absent and present in that category, the number of genes absent and present excluding genes in that category and the *p*-value and odds ratio (OR) from a one-sided Fisher's exact test. Bolded *p*-values are significant after a conservative Bonferroni correction for multiple testing (*p*-value < 0.05). Only categories with a nominal *p*-value less than 0.05 are reported.

GO ID	Biological process	No. tests	Cat. absent genes	Cat. present genes	Absent genes	Present genes	*p*-value	OR
GO:0007608	sensory perception of smell	2,130	146	213	2,775	9,470	**5.3E-14**	2.34
GO:0006355	regulation of transcription, DNA dependent	2,121	681	1632	2,629	9,257	**5.2E-14**	1.47
GO:0031424	keratinization	1,688	21	19	1,948	7,625	**7.1E-06**	4.33
GO:0034340	respose to type I interferon	1,683	22	22	1,927	7,606	**1.1E-05**	3.95
GO:0016339	calcium-dependent cell-cell adhesion	1,674	11	9	1,905	7,584	5.9E-04	4.86
GO:0007586	digestion	1,669	20	31	1,894	7,575	1.3E-03	2.58
GO:0006952	defense response	1,662	120	344	1,874	7,544	1.4E-03	1.40
GO:0048609	multicellular organismal reproductive process	1,522	77	204	1,754	7,200	1.1E-03	1.55
GO:0034728	nucleosome organization	1,462	23	40	1,677	6,996	1.2E-03	2.40
GO:0031023	microtubule organizing center organization	1,457	12	16	1,654	6,956	3.6E-03	3.15
GO:0051258	protein polymerization	1,455	15	23	1,642	6,940	3.0E-03	2.76
GO:0006275	regulation of DNA replication	1,445	15	27	1,627	6,917	8.5E-03	2.36
GO:0070507	regulation of microtubule cytoskeleton organization	1,434	12	19	1,612	6,890	8.4E-03	2.70
GO:0042384	cilium assembly	1,423	9	13	1,600	6,871	1.4E-02	2.97
GO:0016579	protein deubiquitination	1,417	10	16	1,591	6,858	1.6E-02	2.69
GO:0006351	transcription, DNA-dependent	1,414	32	85	1,581	6,842	1.6E-02	1.63
GO:0003013	circulatory system process	1,399	17	37	1,549	6,757	1.7E-02	2.00
GO:0034470	ncRNA processing	1,374	14	122	1,532	6,720	1.7E-02	1.51
GO:0050909	sensory perception of taste	1,362	11	21	1,490	6,598	2.4E-02	2.32
GO:0007275	multicellular organismal development	1,357	73	254	1,479	6,577	4.4E-02	1.28

GO:0007565	female pregnancy	1,308	11	24	1,406	6,323	4.2E-02	2.06

Supplemental Text

Text S1. Identification of paralogous sequences and evaluation of their impact on exome assemblies.

Genes that have duplicated to become paralogs in other lineages are susceptible to mis-assembly because reads from multiple genomic locations may map to a single location in the human genome. To identify problematic paralogous targets we mapped macaque reads to both the human (hg18) and macaque (rheMac2) reference genomes. We then compared the depth of human target sequences to that of 155,707 orthologous target sequences in the macaque genome. As putative paralogs, we identified 137 targets with substantially higher read depth in human compared to macaque (Figure S3A).

To evaluate the impact these putative paralogous targets have on the macaque exome assembly, we compared assembled target sequences to the macaque reference genome sequence. Targets that are mis-assembled will have more nucleotide differences and indels compared to correctly assembled targets (which will have some differences due to polymorphisms in macaque). The assembled putative paralogous targets have a ~6-fold increase in the number of nucleotide differences (1.6% *vs*. 0.25%) and indels (0.14% *vs*. 0.025%) compared to the entire set of assembled targets (Table S17 and Figure S3B). This indicates that the putative paralogous targets are enriched for mis-assembled sequences.

We performed several post-assembly filtering steps to reduce the amount of mis-assembled sequences from paralogs, such as removing targets that overlap known segmental duplications or that have high levels of heterozygosity. The targets that are removed by filtering have a higher proportion of nucleotide differences (0.38% *vs*. 0.23%) and indels (0.056% *vs*. 0.019%) compared to those that are retained. For the subset of targets that are putative paralogs, filtering reduces the proportion of nucleotide differences (from 2.9% to 1.1%) and indels (from 0.21% to 0.12%) (Table S17 and Figure S2B). This demonstrates that our filtering steps remove targets that are more likely to have assembly errors. The nucleotide differences for the retained putative paralogs remain high, however, suggesting that a small fraction of our final exome assemblies contain paralogous assembly errors.

Text S2. Types of genes absent from positive selection analyses.

To better understand the types of genes that failed to capture, assemble or pass filters, we labeled human genes with macaque orthologs as "absent" if they had no macaque sequence following filtering. We then identified gene ontology (GO) terms that were enriched for absent genes (Table S18). The GO terms with the most significant enrichments are "sensory perception of smell" and "regulation of transcription, DNA-dependent". This may indicate that genes in large families (such as olfactory receptors or zinc-finger transcription factors) are difficult to assemble or are preferentially removed by our filtering procedure. Another of the most significant categories, "keratinization", is also enriched for genes predicted to be under positive selection (Table S9), suggesting that some rapidly evolving genes may be excluded from our analysis.

We checked whether any of the GO terms enriched for absent genes are related to immune response or reproduction, because genes involved in these processes are known to be rapidly evolving. The terms "response to type I interferon", "defense response" and "multicellular organismal reproductive success" are enriched for absent genes, but only the first category is significant after Bonferroni correction for the number of tests performed (Table S18). These results indicate some fraction of rapidly evolving genes may be excluded from our positive selection analysis, but most are likely to be retained.

Supplemental References

Pruitt KD, Harrow J, Harte RA, Wallin C, Diekhans M, Maglott DR, Searle S, Farrell CM, Loveland JE, Ruef BJ et al. 2009. The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. *Genome Res* **19:** 1316-1323.

Rhesus Macaque Genome Sequencing and Analysis Consortium. 2007. Evolutionary and biomedical insights from the rhesus macaque genome. *Science* **316:** 222-234.

Storey JD, Tibshirani R. 2003. Statistical significance for genomewide studies. *Proc Natl Acad Sci USA* **100:** 9440-9445.

Tennessen JA, Madeoy J, Akey JM. 2010. Signatures of positive selection apparent in a small sample of human exomes. *Genome Research* **20:** 1327-1334.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24:** 1586-1591.

Zhang J, Nielsen R, Yang Z. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol Biol Evol* **22:** 2472-2479.