www.sciencemag.org/cgi/content/full/316/5822/222/DC1

Supporting Online Material for

Evolutionary and Biomedical Insights from the Rhesus Macaque Genome

Rhesus Macaque Genome Sequencing and Analysis Consortium

Correspondence should be addressed to Richard A. Gibbs. E-mail: agibbs@bcm.edu

Published 13 April, *Science* **316**, 222 (2007)
DOI: 10.1126/science.1139247

This PDF file includes:

Other Supporting Online Material for this manuscript includes the following:
(available at www.sciencemag.org/cgi/content/full/316/5822/222/DC1)

Tables as zipped archives:
Table S2.3. Assembly statistics by chromosome.
Table S2.4. Detailed comparison of three different assemblies.
Table S4.1. Determination of the lineage specificity of the pericentric inversions that distinguish the human and chimpanzee.
Table S5.1. Duplications detected in the rhesus genome by three complementary methods.
Table S5.3. Array CGH data for gene gains in macaque relative to human.
Table S5.4. Array CGH values for HLA Class I–related genes among macaque and hominoid lineages.
Table S8.1. List of 229 gene candidates for ancestral human gene mutations.
Table S8.2. Plasma amino levels in eight macaques.
Table S8.3. Determination of acylcarnitine (nM) in rhesus.
Table S9.1. List of oligonucleotide probes for rhesus mRNA transcripts on Agilent rhesus macaque microarray chip.
Table S9.2. Content of Agilent rhesus macaque microarray and the sequence homology to predicted macaque genes.
Table S9.4. Hybridization of Agilent rhesus macaque microarray to mRNA samples from either infected lung tissue or whole blood, in a macaque model of influenza.

Supplementary Online Materials:
Evolutionary and Biomedical Insights from the Rhesus Macaque Genome,

(Rhesus Macaque Genome Sequencing and Analysis Consortium)

Introduction:

The following online materials support the publication of 'Evolutionary and Biomedical Insights from the Rhesus Macaque Genome' by the Rhesus Macaque Genome Sequencing and Analysis Consortium (RMGSAC). This project follows a 'White Paper' outlining the benefits of determining a draft sequence of this important primate (http://www.genome.gov/Pages/Research/Sequencing/SeqProposals/RhesusMacaqueSEQ021203.pdf) and subsequent approval and funding by the National Human Genome Research Institute (NHGRI).

Tables that are too large for the pdf of this supplementary manuscript are available in the associated file of tables. Below, we refer to that as "the associated rhesus file". The organization of the SOM follows the organization of the accompanying manuscript- e.g Introduction, Sequencing the Genome etc.)

1. Introduction

Table S1.1: Basic Information concerning Rhesus Macaques

Taxonomy	
A) Nomenclature: *Macaca mulatta*, includes six named subspecies (1)	
B) Taxonomic Details:	
Order:	Primates
Infraorder:	Catarrhini
Superfamily:	Cercopithecoidea
Family:	Cercopithecidae
Subfamily:	Cercopithecinae
Tribe:	Cercopithecini
Genus:	*Macaca* (Genus includes 20 species)
Body Size	
Adult body weight: Male 5.6-10.9 kg, Female 4.4-10.9 kg (2)	
Life History	
A) Lifespan: 29 years (4)	
B) Gestation: 164 days (5)	
C) Seasonal breeders producing singleton births	
D) Female sexual maturity approximately 54 months (6)	
Geographic Distribution	
India north of the Krishna River, north and east across eastern Afghanistan, Kashmir, Nepal, Sikkim, Bhutan and northern Myanmar into southern, eastern and northeastern China to the Yangtse River. Also found north of the lower Huang Ho River, with an additional population isolated on Hainan Island.	
Research Use	
A) Rhesus macaques are the most widely used nonhuman primate in biomedical research. There are more than 50,000 rhesus monkeys in US Colonies. The primary research applications of rhesus are in the fields of neuroscience, immunology and infectious diseases especially AIDS research, reproductive biology, stem cell biology, metabolism and obesity, diabetes, behavioral biology and addiction.	
B) From 2002 through 2005, the PubMed database lists 3713 papers published concerning rhesus macaques	

2. Sequencing the Genome

Genome Resources that were available for this project are described in **Table S2.1**, below.

Genome Resources		
Maps		
Rogers map	PMID: 16321502	
Murphy map	PMID: 16039092	
Fingerprint contig map	BCGSC	http://www.bcgsc.ca/downloads/rhesusmap.tar.gz
PGI mapped BACs	PMID: 15687293	http://brl.bcm.tmc.edu/pgi/rhesus/index.rhtml
Gene Lists	**URL**	
NCBI – Gnomen	http://www.ncbi.nlm.nih.gov/mapview/map_search.cgi?taxid=9544	
Ensembl – Ensembl	http://www.ensembl.org/Macaca_mulatta/index.html	
UCSC – Nscan	http://www.genome.ucsc.edu/cgi-bin/hgGateway?hgsid=82024512&clade=vertebrate&org=Rhesus&db=0	
Large Insert Libraries		
BAC library	CHORI-250	http://bacpac.chori.org/rhesus250.htm
Fosmid library	WUGSC	

Table S2.1: Genome Resources for the rhesus macaque.

Genome Sequencing Details: Approximately 1/3 of the DNA sequence reads were generated at each of three sites: The Baylor College of Medicine Human Genome Sequencing Center (BCM-HGSC); the Washington University Genome Sequencing Center (WashU-GSC) and the J. Craig Venter Institute (JCVI). All wgs DNA sequences were from a single *Macaca mulatta* female. BAC ends were from an unrelated male. Standard AB fluorescent Sanger sequencing methods were used. The sequences described in **Table S2.2** were accumulated and used in the assembly:

Insert Size in Assembly	Number of reads
Less than 7kb	21,104,958
7kb to 20kb	2,844,486
20kb to 60kb	781,822
Greater than 60kb	255,882

Table S2.2: Distribution of insert sizes in the assembly

The amalgamated assembly described in the main text was evaluated by more than 200 different statistics, resulting in the final assembly used for the analysis. The initial assembly programs were Atlas-WGS (7), PCAP (8) and the Celera Assembler (9). The characteristics of the merged assembly are described below in **Table S2.3** (Assembly statistics by chromosome) and **Table S2.4** (Detailed comparison of three different assemblies) (***please note these are large tables and so are made available in the associated rhesus file***).

Table S2.3: Assembly statistics by chromosome (see the associated rhesus file).

Table S2.4: Detailed comparison of three different assemblies (see the associated rhesus file).

The accession numbers for the sequence assemblies and other sequence data generated in this study are shown in **Table S2.5**: Sequence Accessions

Table S2.5: Sequence Accessions and Trace Archive (TI) numbers for Indian and Chinese wgs reads for SNP Discovery (table continues)

RheMac2 Assembly Accessions	Wgs reads for SNP discovery (TI Numbers Representing 353 ranges for a total of 33,012 sequence reads from Chinese and Indian Macaques)
Contigs (AANU01000001 to AANU01301039 or AANU01*). Chromosome accessions (CM000288-CM000308). Unplaced scaffolds on Chromosomes (CH666572-CH667783). ChromUn (Ch667784-CH670322).	1377233212..1377233302,1377239536..1377239628, 1377243034..1377243127,1377248115..1377248207, 1377280075..1377280168,1377281375..1377281466, 1377283298..1377283392,1377285702..1377285796, 1377286814..1377286905,1377288384..1377288477, 1377288570..1377288663,1377289779..1377289874, 1377293667..1377293856,1377295203..1377295297, 1377298879..1377298973,1389242995..1389243080, 1389249326..1389249413,1389253533..1389253623, 1389258964..1389259051,1389682265..1389682348, 1389686316..1389686402,1394316411..1394316498, 1394321268..1394321353,1395850659..1395850748, 1395856466..1395856554,1395861827..1395861917, 1395867703..1395867789,1395875471..1395875562, 1395882080..1395882169,1395886582..1395886671, 1395892439..1395892526,1398913507..1398913599, 1399372795..1399372886,1399376148..1399376240, 1399378212..1399378303,1399381591..1399381683, 1399382742..1399382835,1399383022..1399383112, 1399385029..1399385122,1399386045..1399386132, 1399386930..1399387021,1399387358..1399387450, 1399388261..1399388354,1399388620..1399388710, 1399390275..1399390367,1399390547..1399390641, 1399461904..1399461995,1399462084..1399462172, 1399464635..1399464726,1399466382..1399466471, 1399466552..1399466643,1399467555..1399467644, 1399468095..1399468188,1399471068..1399471254, 1399471341..1399471430,1399472184..1399472275, 1399476459..1399476549,1399477004..1399477188, 1399479150..1399479239,1399481708..1399481891, 1399482353..1399482446,1399483163..1399483252, 1399483885..1399483974,1399487330..1399487423, 1399487968..1399488060,1399488422..1399488511, 1399488874..1399488963,1399490833..1399490924, 1399492016..1399492109,1399493296..1399493473,

	1399494937..1399495028,1399499651..1399499742, **1399500109..1399500199,1399549984..1399550074,** **1399554828..1399554919,1405956116..1405956204,** **1405958022..1405958113,1405961307..1405961395,** **1405963587..1405963766,1405965149..1405965324,** **1405968248..1405968338,1405970631..1405970715,** **1405972543..1405972634,1405973364..1405973453,** **1405986609..1405986691,1406924402..1406924493,** **1406928204..1406928295,1406933779..1406933871,** **1406938716..1406938807,1406943833..1406943923,** **1411280386..1411280475,1411281088..1411281180,** **1411284672..1411284764,1411287256..1411287345,** **1411290171..1411290264,1411293548..1411293641,** **1411296117..1411296209,1411297969..1411298062,** **1411302801..1411302892,1411331566..1411331657,** **1411335166..1411335260,1411340115..1411340209,** **1411345507..1411345600,1411349902..1411349994,** **1411356443..1411356535,1411361127..1411361220,** **1424080475..1424080564,1424093696..1424093784,** **1424174861..1424174951,1424175317..1424175410,** **1424186651..1424186742,1424186833..1424186922,** **1424187017..1424187108,1424195163..1424195256,** **1424196827..1424196920,1424199558..1424199650,** **1424205442..1424205534,1424207102..1424207193,** **1424207566..1424207656,1424233874..1424233965,** **1424235794..1424235884,1424243735..1424243827,** **1431302411..1431302500,1431306401..1431306486,** **1431307321..1431307409,1431311972..1431312064,** **1431852885..1431852974,1431853069..1431853161,** **1431856365..1431856454,1431858566..1431858657,** **1431863491..1431863580,1431868816..1431868905,** **1434936859..1434936949,1434937685..1434937767,** **1434964647..1434964733,1434968370..1434968459,** **1434968736..1434968821,1434969920..1434970008,** **1436371670..1436371759,1436377632..1436377723,** **1436404168..1436404260,1437880886..1437880973,** **1437886042..1437886131,1437891207..1437891297,** **1437897091..1437897180,1437897461..1437897552,** **1437901797..1437901886,1437903977..1437904066,** **1437907382..1437907473,1437908112..1437908116,** **1437940473..1437940557,1437946960..1437947053,** **1437947700..1437947793,1437951613..1437951706,** **1437958859..1437958950,1437959508..1437959600,** **1437963010..1437963103,1437963632..1437963721,** **1437967125..1437967217,1437968959..1437969050,** **1437973438..1437973529,1437973805..1437973897,** **1437974628..1437974720,1437979007..1437979099,** **1437983767..1437983860,1443742132..1443742224,** **1443747810..1443747903,1443752375..1443752465,** **1443754301..1443754393,1443754935..1443755029,** **1443756007..1443756098,1443758029..1443758119,** **1443759856..1443760044,1443760928..1443761019,** **1443763048..1443763140,1443764600..1443764692,** **1443765140..1443765231,1443765859..1443765949,** **1443768956..1443769048,1443769929..1443770021,** **1443771283..1443771373,1443773811..1443773903,** **1443785894..1443785984,1443786207..1443786293,** **1443789434..1443789524,1443794954..1443795044,** **1443795956..1443796045,1443799383..1443799471,** **1448546416..1448546507,1448551312..1448551402,** **1448555990..1448556077,1448561229..1448561313,** **1448578462..1448578553,1448583405..1448583496,** **1448588118..1448588208,1448593047..1448593138,**

	1448597666..1448597755,1448600910..1448601000, 1448607142..1448607231,1448611267..1448611357, 1449002700..1449002792,1449007680..1449007772, 1449012581..1449012672,1449016862..1449016954, 1449173833..1449173920,1449178744..1449178833, 1449184497..1449184583,1449189314..1449189399, 1458998314..1458998407,1459302358..1459302447, 1459306775..1459306864,1459311111..1459311198, 1459315094..1459315184,1468358923..1468359015, 1468368049..1468368140,1468372921..1468373011, 1468375312..1468375401,1468381829..1468381918, 1468386514..1468386604,1468389373..1468389466, 1468394073..1468394164,1468394260..1468394352, 1468394910..1468395000,1468397860..1468397949, 1468399240..1468399329,1468399599..1468399689, 1468400607..1468400696,1468404013..1468404104, 1468404468..1468404561,1468405476..1468405565, 1468408702..1468408793,1468408982..1468409074, 1468411236..1468411326,1468413455..1468413549, 1468416886..1468416979,1468422141..1468422234, 1468427114..1468427205,1475813333..1475813423, 1475819436..1475819524,1475825574..1475825665, 1475829909..1475829996,1475836029..1475836121, 1475841286..1475841380,1475848563..1475848655, 1475850128..1475850220,1475851869..1475851962, 1475855922..1475856013,1475858175..1475858267, 1475864038..1475864132,1475868739..1475868830, 1475875105..1475875195,1475879112..1475879204, 1475880859..1475880950,1475884775..1475884866, 1475885235..1475885327,1475890169..1475890263, 1475890536..1475890627,1475894780..1475894871, 1475895981..1475896074,1475901747..1475901839, 1475930144..1475930236,1481850659..1481850751, 1481854613..1481854704,1481856011..1481856103, 1481859846..1481859937,1481860033..1481860126, 1481860217..1481860310,1481861229..1481861321, 1481863491..1481863582,1481863859..1481863951, 1481865249..1481865342,1481866449..1481866539, 1481868254..1481868347,1481869902..1481869992, 1481870262..1481870355,1481873110..1481873203, 1481873387..1481873478,1485354510..1485354603, 1485356368..1485356459,1485358210..1485358303, 1485360052..1485360146,1485361892..1485361985, 1485364195..1485364290,1485364477..1485364569, 1485365505..1485365599,1485368065..1485368160, 1485369347..1485369442,1485369808..1485369900, 1485371200..1485371295,1485371854..1485371949, 1485373781..1485373875,1485376830..1485376923, 1485379974..1485380069,1495155003..1495155097, 1495162520..1495162614,1495165973..1495166066, 1495170960..1495171054,1495178708..1495178801, 1495184306..1495184400,1495187261..1495187356, 1495558477..1495558571,1496239059..1496239154, 1496244234..1496244329,1496246841..1496246935, 1496252489..1496252582,1496273528..1496273621, 1496277386..1496277479,1496299617..1496299708, 1496304871..1496304964,1496386346..1496386434, 1496393664..1496393750,1496398884..1496398967, 1496400699..1496400789,1496403519..1496403603, 1496408006..1496408097,1496408557..1496408641, 1496412699..1496412792,1496417468..1496417549, 1496419032..1496419124,1496421068..1496421149, 1496424017..1496424097,1496872351..1496872432,

	1496874005..1496874094,1496878894..1496878983, 1496883473..1496883562,1496890632..1496890721, 1496895294..1496895384,1496902418..1496902512, 1496907919..1496908012,1496911701..1496911789, 1496917491..1496917582,1496930478..1496930569, 1496930755..1496930847,1496934686..1496934776, 1496939827..1496939917,1496946582..1496946675, 1496950974..1496951066,1496959583..1496959675, 1496965942..1496966033,1496971511..1496971602, 1496975210..1496975301,1496977097..1496977189, 1496980823..1496980913,1496981932..1496982021, 1496986244..1496986332,1496989468..1496989558, 1497013741..1497013834,1497015523..1497015611, 1497018506..1497018597,1497019050..1497019143, 1497025157..1497025249,1497030494..1497030585, 1498466641..1498466733,

In order to obtain higher quality sequence in targeted regions, individual BAC clones were isolated by one or more of several different methods. These included PGI (10), BES mapping and the identification of genome coordinates by WGAC. Many of these BACs were in regions of pronounced genome duplication, whereas others were selected because they were in ENCODE regions (http://www.genome.gov/10005107) or were gene-rich. Finished BACs, their gene content and their genome coordinates are listed in **Table S2.6.** (Summary of finished BACs for rhesus macaque). Note some of these are in progress and status should be checked at the NCBI.

Supplementary Table S2.6. **Summary of finished BACs for rhesus macaque,**

CLONE NAME	NCBI ACCESSION	MACAQUE GENES AND HUMAN ORTHOLOGS
SEGMENTAL DUPLICATION CLONES		
CH250-316D24	AC189801	none
CH250-2C21	AC190280	DUX1,DIP2C,DUX4,DUX4C,FRG2
CH250-169M23	AC189884	DUX1,DIP2C,DUX4,DUX4C,FRG2
CH250-246B15	AC190405	DUX1,DIP2C,DUX4,DUX4C,FRG2
CH250-361O6	AC190406	DUX1,DIP2C,DUX4,DUX4C,FRG2
CH250-18D16	AC189878	DIP2C
CH250-215O17	AC192770	none
CH250-149E8	AC190278	DUX1,DIP2C,DUX4,DUX4C,FRG2
CH250-343K15	AC191820	DUX1,DIP2C,DUX4,DUX4C,FRG2
CH250-320H23	AC191822	DIP2C,LARP5
CH250-368P5	AC191821	CKAP1,LARP5
CH250-438D19	AC192771	WDR37,IDI1,IDI2,GTPBP4,C10orf110
CH250-375C15	AC191818	C10orf108
CH250-397E10	AC191819	DIP2C

CH250-249G9	AC191462	DIP2C,DIP2A
CH250-450C05	AC191853	LIPI
CH250-378O7	AC191859	C10orf108
CH250-230K21	AC191880	none
CH250-411K20	AC191852	none
CH250-135L17	AC191850	DKFZp434I1020,LOC440295,GOLGA8E,GOLGA8G,LOC400464,GOLGA8A,FLJ32679,GOLGA8B
CH250-265I3	AC191883	OR4F3,OR4F15,OR4F16,OR4F6,OR4F21,OR4F29,OR5BF1
CH250-483J10	AC191855	OR2G3,OR4F21,OR4F3,OR4F4,OR4F5,OR4F16,OR4F17,OR4P4,OR4F29
CH250-214L19	AC191884	OR2G3,OR4N5,OR4P4
CH250-405C19	AC191851	OR4M1,OR4M2,OR11H1,OR4N2,OR5BF1,OR4N4,OR4Q3
CH250-392A15	AC191858	NR2F6,OR11G2,OR11H12,OR11H1,OR5BF1,OR11A1,OR11H4,OR11H6,OR4Q3
CH250-88O18	AC191848	TESK1,C20orf22,KIAA0980,PSF1
CH250-318I09	AC191879	ZNF133,HDHD4,ZNF337,KIAA0980,ZNF589,HKR1
CH250-5J1	AC191847	ACTBL1,POTE8,ANKRD21,POTE14,POTE15,POTE2
CH250-176F21	AC191856	ACTBL1,LOC441956,DIP,POTE8,C9orf79,ANKRD21,POTE14,POTE15,POTE2
CH250-102I21	AC191849	LOC400968,LOC441956,DIP,C9orf79
CH250-518D17	AC191885	TUBA2
CH250-311I19	AC191857	TPTE,TPTE2
CH250-238L1	AC191881	TPTE,LOC400927,TPTE2
CH250-243I15	AC191882	PRKCH,C14orf106,LOC441931,VN1R5,GAB4
CH250-483H18	AC191854	PRKCH,LOC441931,VN1R5
PRAME REGION CLONES		
CH250-64G9	AC191983	PRAMEF1,PRAMEF2,C1orf158,PRAMEF4,PRAMEF5,PRAMEF6,PRAMEF7,PRAMEF8,PRAMEF9,LOC343066,PRAMEF10
CH250-384N22	AC191987	PRAMEF1,PRAMEF2,C1orf158,PRAMEF7,PRAMEF8,LOC343066
CH250-216D12	AC191985	PRAMEF1,PRAMEF2,PRAMEF3,PRAMEF4,PRAMEF5,PRAMEF6,PRAMEF7,PRAMEF8,PRAMEF9,PRAMEF10
CH250-122A16	AC191984	PRAMEF1,PRAMEF2,PRAMEF3,PRAMEF4,PRAMEF5,PRAMEF6,PRAMEF8,PRAMEF9,PRAMEF10
CH250-20P10	AC191982	PRAMEF8,BRWD1
CH250-340N14	AC191986	PDPN,PRAMEF8,BRWD1
CCL3 REGION CLONES		
CH250-42A15	AC192296	CCL3L3,CCL4L1,CCL18,CCL4L2,GOLPH2,TBC1D3,CCL3,CCL4,USP6,TBC1D3B,TBC1D3C,AF449272,AF457195,AF449266,AF457196,AF449267,LOC440452,CCL3L1
CH250-63J20	AC192069	AF457196,AF449267
CH250-271K24	AC142898	TBC1D3,USP6,TBC1D3B,TBC1D3C,LOC440452
CH250-352A6	AC189964	CCL16,FLJ43826,CCL18,RDM1,LYZL6,AF449272,AF449276,CCL14,CCL23,CCL15

CH250-325M9	AC192012	CCL16,FLJ43826,RDM1,LYZL6,CCL5,C17orf66,AF457194,AF449268,CCL14,CCL15
CH250-16P11	AC192068	TBC1D3,USP6,TBC1D3B,TBC1D3C,LOC440452
CH250-482K15	AC192014	none
CH250-168M9	AC192070	TBC1D3,USP6,NF1,TBC1D3B,TBC1D3C,LOC440452
CH250-141K20	AC192297	NF1
CH250-419N13	AC192013	NF1,OMG,EVI2B
COMMUNITY SUBMITTED GENE LIST CLONES		
CH250-276H13	AC189957	KIRREL3,KIRREL,FLJ21103,DCPS,ST3GAL4,H17,TIRAP,SRPR
CH250-359J12	AC189965	TM9SF1,RIPK3,MGC5987,GMPR2,DHRS1,ADCY4,IPO4,NEDD8,LTB4R2,C14orf21,CHMP4A,CIDEB,LTB4R,RABGGTA,TGM1,NFATC4,TINF2,GMPR,TSSK4
CH250-374E17	AC171635	GART,IFNGR2,SON,DONSON,C21orf55,TMEM50B
CH250-42A15	AC192296	CCL3L3,CCL4L1,CCL18,CCL4L2,GOLPH2,TBC1D3,CCL3,CCL4,USP6,TBC1D3B,TBC1D3C,AF449272,AF457195,AF457196,AF449266,AF449267,LOC440452,CCL3L1
CH250-431C15	AC170191	IL13,U19848,AY244790,IL4,RAD50,IL5,L26027,AY376144,AF457197
CH250-452F20	AC190047	FLJ10726,IL18,LOC120379,AF303732,DLAT,DIXDC1,SDHD,DQ148040,BCDO2,TEX12
CH250-498B22	AC171341	IL13,U19848,AY244790,RAD50,IL5
CH250-499P8	AC190043	OPRS1,CCL27,CCL19,ARID3C,AY288833,CNTFR,GALT,DCTN3,AF449273,C9orf23,AF449275,IL11RA,AF449278,CCL21
GENE RICH CLONES		
CH250-24J13	AC189934	TM9SF1,NRL,CPNE6,CPNE7,RNF31,DHRS4L2,MGC5987,PSME1,PSME2,MAFA,MAFB,REC8L1,IPO4,DHRS4,CHMP4A,MAF,LOC161247,LRRC16,C14orf121,C14orf122,PCK2,WDR23,ISGF3G,TSSK4
CH250-366P8	AC189961	AJ560720,C20orf112,C6orf21,DUSP19,VARSL,C6orf25,APOM,LY6G5B,C6orf27,LY6G5C,BAT2,LY6G6C,LY6G6D,C6orf47,BAT3,LY6G6E,BAT4,BAT5,NUP62,DDAH2,MSH5,VARS,CLIC1,CSNK2B
CH250-359J12	AC189965	TM9SF1,RIPK3,MGC5987,GMPR2,DHRS1,ADCY4,IPO4,NEDD8,LTB4R2,C14orf21,CHMP4A,CIDEB,LTB4R,RABGGTA,TGM1,NFATC4,TINF2,GMPR,TSSK4
CH250-181E1	AC190286	AY116212,USP21,UFC1,PFDN2,PPOX,NR1I3,APOA2,TOMM40L,FCER1G,USF1,NDUFS2,DEDD,ARHGAP30,NIT1,PVRL4,KARCA1,B4GALT2,B4GALT3,ADAMTS4
CH250-120K7	AC190276	TRPT1,KCNK4,HSPC152,URP2,STIP1,DNAJC4,PLCB3,NUDT22,ESRRA,FKBP2,VEGFB,BAD,C11orf4,FLJ37970,TM7SF1,LIMK2,PPP1R14B,RPS6KA4
CH250-39J23	AC189871	LBX2,GCS1,ZNHIT4,DOK1,DQX1,HTRA2,RTKN,LOC130951,LOXL2,LOXL3,FLJ14397,MRPL53,PCGF1,FLJ12788,WBP1,AUP1,TLX2
CH250-390G5	AC190040	FLJ36268,ENTPD2,UAP1L1,C9orf140,LCN12,ABCA2,NPDC1,FLJ45224,PTGDS,LOC286257,CLIC3,FUT7,DPP7,MAN1B1,EID-3

CH250-494P20	AC171641	HISPPD2A,CKMT1A,CKMT1B,ELL3,CATSPER2,SERF2,AY919832,SERINC4,SERINC5,MFAP1,STRC,PDIA3,AY680461,DQ148150,HYPK
CH250-361G21	AC190029	PNPLA2,CHID1,PDDC1,POLR2L,MGC45840,SLC25A22,CD151,TALDO1,AF275665,TSPAN4,AP2A2,LRDD,BM88
CH250-139H22	AC190273	TLCD1,FLJ25006,PIGS,RAB34,ALDOC,SPAG5,KIAA0100,RPL23A,SUPT6H,NEK8,SDF2,PROCA1
CH250-293C5	AC190287	EN2,EDF1,C8G,LCN12,ABCA2,FLJ45224,TRAF2,MAMDC4,FBXW5,PTGDS,LOC286257,CLIC3
CH250-45M17	AC189936	YIF1A,C20orf108,MGC33486,CNIH2,KLC2,SLC29A2,KLC4,RAB1B,KNS2,RIN1,OLFML2A,BRMS1,B3GNT6,CD248,PACS1
CH250-370O18	AC189949	H2AFX,VPS11,HMBS,DPAGT1,KCTD6,SLC37A4,MIZF,DLNB14,TRAPPC4,RPS25,TMEM24,HYOU1
CH250-412M23	AC190032	DNTTIP1,PLTP,ACOT8,TNNC2,PPGB,C20orf161,ZSWIM1,ZNF587,ZSWIM3,NEURL2,C20orf165,UBE2C,WFDC3
CH250-65H9	AC189886	BRD2,HLA-DQB2,PPP1R2,PSMB8,PSMB9,TAP1,TAP2,HLA-DMA,HLA-DMB,HLA-DOB,HLA-DRA
CH250-507N8	AC190044	HSPA1A,HSPA2,HSPA1B,C20orf112,AY680559,DUSP19,VARSL,C6orf25,NEU1,C6orf27,HSPA1L,LY6G6C,LY6G6D,C6orf48,NUP62,DDAH2,MSH5,VARS,CLIC1,LSM2
CH250-109K11	AC189889	C16orf53,TAOK2,MVP,MAZ,KIF22,SEZ6L2,KCTD13,LOC124446,PRRT2,LOC253982,CDIPT
CH250-253F13	AC189953	GLT8D1,ITIH1,ITIH3,ITIH4,SPCS1,FTS,NEK4,GNL3,MUSTN1,TMEM110,ZNF610
CH250-340A9	AC190317	C9orf96,SURF1,SURF2,REXO4,C9orf7,SURF4,SLC2A6,SURF5,SURF6,RPL7A,ADAMTS13
CH250-319M4	AC190322	TESK1,CREB3,C9orf100,SIT1,TLN1,CA9,MGC31967,TPM2,GBA2,CD72
CH250-502E5	AC190057	PNPLA2,CHID1,POLR2L,MGC45840,SLC25A22,CD151,AF275665,TSPAN4,AP2A2,LRDD,BM88
CH250-148M8	AC190065	TSPAN31,METTL1,CDK4,TSFM,CYP27B1,CENTG1,AVIL,OS9,DKFZP586D0919,MARCH9,CTDSP2,CCL22
CH250-403D15	AC189962	WFDC9,WFDC11,DNTTIP1,WFDC13,WFDC10A,WFDC10B,ACOT8,TNNC2,C20orf161,ZNF587,UBE2C,WFDC3
CH250-178E24	AC190060	NCOA5,PLTP,PPGB,ZSWIM1,ZSWIM3,NEURL2,C20orf165,C20orf67,ZNF335,SLC12A5,SLC12A7,MMP9
CH250-550M15	AC190310	RPGR,CKS1B,SHC1,ZNF628,TSPAN18,ADAM15,FLAD1,ZBTB7B,EFNA3,EFNA4,DCST1,DCST2,LENEP
CH250-307C21	AC189959	LIX1L,RBM8A,ITGAE,POLR3C,NUDT17,ANKRD35,POLR3GL,AF230105,ITGA10,ITGA11,AF353987,PIAS3,PEX11B,ZNF364
CH250-62N11	AC189937	C14orf8,FLJ20859,SLC39A2,RNASE13,NDRG2,ZNF219,RNASE2,RNASE3,FLJ10357,RNASE7,RNASE8
CH250-311D22	AC189966	FOXQ1,ELK1,AY452560,LOC390511,MTA1,CRIP1,CRIP2,AF045538,MGC4659,AF045539,C14orf80
CH250-273D15	AC189950	NCR3,U19850,AIF1,AY035214,TNF,AY035215,AY035216,AY035217,LST1,BAT1,BAT2,BAT3,LTA,LTB,AF322860,NFKBIL1,HCG9,MCCD1,AF162475,FLJ35429,AJ554301,KIAA1008,ATP6V1G2,AF055388,MICA,AF322859,MICB

CH250-392F18	AC190031	HAGH,NME3,MAPK8IP3,MRPS34,IGFALS,SPAG9,MGC35212,SPSB3,FAHD1,NUBP2,EME2
CH250-463J17	AC190059	YIF1B,KCNK6,DPF1,PSMD8,LOC541469,FLJ44968,C19orf15,C19orf33,GGN,SPINT2,PPP1R14A
CH250-499P8	AC190043	OPRS1,CCL27,CCL19,ARID3C,AY288833,CNTFR,GALT,DCTN3,AF449273,C9orf23,AF449275,IL11RA,AF449278,CCL21
CH250-493K11	AC190064	PIGO,FANCG,DNAJB4,DNAJB5,MGC41945,STOML2,KIAA1539,UNC13B,SYF2,VCP
CH250-271I18	AC143915	ZNF385,GPR84,COPZ1,HNRPA1,NFE2,LOC144983,M84334,ITGA5,NFASC,FLJ32942,NCKAP1L,AY901982,CBX5
CH250-532C21	AC190034	GBA,RPGR,TRIM46,RAG1AP1,KRTCAP2,THBS3,MUC1,DPM3,EFNA1,MTX1,EFNA3
CH250-455E17	AC190058	GPATC4,ISG20L2,HDGF,BCAN,APOA1BP,NES,MRPL24,FLJ44968,C1orf66,HAPLN2,CRABP2,C20orf96,AY680544
CH250-26C5	AC189866	S100A14,S100A16,S100A1,S100A2,C1orf77,S100A3,S100A4,S100A5,S100A6,S100A13
CH250-162K4	AC190067	UBXD5,AY680497,DHDDS,AIM1L,ZNF683,LIN28,CHSY1,SH3BGRL3,CCDC21,CD52
CH250-421D15	AC190321	C10orf62,PGAM1,ZDHHC16,PGAM2,PGAM3,PGAM4,C10orf65,C10orf83,EXOSC1,UBTD1,DQ147953,ANKRD2,MMS19L,DQ148137
CH250-347G1	AC189942	ZIC2,MGC33584,CYHR1,PLXND1,FOXH1,KIFC2,GPT,LRRC14,SAMD11,MGC70857,MFSD3,LRRC24,SALL3,PPP1R16A,RECQL4,PPP1R16B,KIAA1688
CH250-266O12	AC189967	TMEM55B,PTCD2,NP,OSGEP,RNASE10,AF382950,TEP1,APEX1,RNASE9
CH250-155J16	AC190325	IK,WDR55,HARS,NDUFA2,CD14,PRO1580,DND1,ZMAT2,HARSL
CH250-467F12	AC190125	CAPN11,HSP90BB,NFKBIE,SLC35B2,HNRPA1,SLC29A1,SLC29A2,MGC33600,HSPCB,AARSL,DQ147989
CH250-1J22	AC189885	ZNF287,HIST1H1B,HIST2H4,OR2B2,OR2B6,HIST1H3H,HIST1H3I,H4/o,HIST1H3J,HIST1H2AE,HIST1H4J,HIST1H4K,HIST1H2AI,HIST1H4L,ZNF420,HIST1H2AJ,HIST1H2AK,HIST1H2AL,HIST1H2AM,HIST1H2BL,ZNF271,HIST1H2BM,HIST1H2BN,ZNF184,HIST1H2BO
CH250-63D7	AC189938	L10609,ZAN,EPO,TRIP6,ACHE,POP7,EPHB4,AY428851,ARS2,AY428852,LOC402682,SLC12A9
CH250-499G8	AC190042	NOS1,NOS3,ATG9B,C7orf21,CDK5,CENTG3,LOC159090,ACCN3,SLC4A1,SLC4A2,SLC4A3,FASTK,ABCB8
CH250-342K19	AC190300	TAS2R3,CLEC5A,TAS2R4,TAS2R5,OR5M1,SSBP1,LOC136242,TAS2R38,OR9A4
CH250-272B4	AC142582	DPEP3,CTRL,PSKH1,DDX28,PSMB10,LCAT,SLC12A4,DUS2L,DPEP2
CH250-166N7	AC190066	EEF1A2,LSR,USF2,CD22,MAG,HAMP,FLJ25660,GPR40,GPR41,GPR42
CH250-392L15	AC189978	RPL26,ARHGEF15,RANGNRF,C17orf44,ODF4,PFAS,AURKB,C17orf68,LOC124751,SLC25A35,SPTLC1
CH250-355G8	AC189968	PTP4A2,AY680572,TUBG1,TUBG2,HSD17B1,TBPIP,ATP6V0A1,DEAF1,NAGLU,COASY,AY680437,LOC162427,MLX
CH250-1H4	AC189868	LRRC59,XYLT2,FLJ20920,RSAD1,CHAD,MRPL27,MYCBPAP,EME1
CH250-10K12	AC190292	CREB3,NPR2,TLN1,OR2S2,SPAG8,C9orf127,C9orf128,GUCY1B2,HINT2,GBA2

CH250-364N5	AC189969	GLYAT,CANP,SCN4B,GLYATL1,C17orf63,GLYATL2,FLJ22794
CH250-20I5	AC189865	SLC15A3,TMEM109,SLC15A4,ZP1,HSPA5BP1,MGC35295,MGC2574,PRPF19,MS4A10,GPR44
CH250-267C5	AC143583	ZDHHC24,ACTN3,CCS,BBS1,FLJ10786,PELI3,DPP3,CTSF
CH250-310C5	AC190282	TOR1B,DOK1,DQX1,HTRA2,LOC130951,LOXL2,LOXL3,PCGF1,AUP1,SEMA4F,TLX2
CH250-479E19	AC190062	AY680581,EMG1,RERE,PTPN6,ATN1,ENO2,PHB2,GRCC10,ENO3,C3F,B7
CH250-247G15	AC189952	SOAT2,TENC1,FLJ14800,LOC283337,ITGB2,CSAD,AY680514,ITGB7,MFSD5,IGFBP6,RARB,TNS3,DQ148207,DQ148209,RARG
CH250-368D24	AC190039	TMEM56,PTPRA,ProSAPiP1,FLJ13149,AF424826,PTPRE,C20orf116,AVP,MRPS26,UBOX5,AF097356,AF104307,OXT,GNRH2
CH250-194G10	AC189873	DQ148067,MTP18,TBC1D10A,SEC14L2,LOC550631,TBC1D10B,SEC14L3,NR2F1,SEC14L4,SF3A1,SLC35E3,LOC200312,DQ148025,SDC4
CH250-5E18	AC190284	PGC,WDR77,LOC149620,ADORA3,ATP5F1,AY680510,OVGP1,U87259,C1orf88,C1orf162,CHIA,DQ148056
CH250-226C15	AC189935	FDPS,HCN3,CLK2,ASH1L,C1orf2,RUSC1,SCAMP3,C1orf104,PKLR,HCN2
CH250-1F14	AC189940	GPR61,GNAT2,PSMA5,SYPL2,GNAI1,AMPD2,GNAI3,ATXN7L2,CYB561D1,AMIGO1
CH250-135G7	AC190297	BOLA1,LOC440686,H3/o,HIST2H3C,SF3B4,HIST2H2AA,FCGR1A,HIST2H2AB,HIST2H2AC,HIST2H2BE,HIST2H2BF,UBE2D4,SV2A,LOC440607
CH250-120D8	AC190296	KIF4A,RAB41,DGAT2L3,IGBP1,DGAT2L6,ARR3,PDZK11,P2RY4
CH250-61E5	AC189947	SLC22A17,KIAA1443,IL17E,MYH6,MYH7,EFS,PABPN1,BCL2L2,MYH7B,CMTM5
CH250-216H15	AC190290	MESP1,PLIN,MRPL15,ANPEP,KIF27,KIF7,LOC56964,PEX11A,C15orf42
CH250-184G11	AC189872	APXL2,SEPT10,LEAP-2,ANKRD43,GDF9,UQCRQ,DQ148113,AFF4,KIF3A
CH250-494C3	AC190271	PSORS1C1,PSORS1C2,C6orf15,TCF19,POU5F1,CCHCR1,HCG27,CDSN,FLJ25680
CH250-172L10	AC190315	TREML1,NFYA,LOC221442,UNC5CL,APOBEC2,TREM2,C6orf130,BZRPL1
CH250-103G2	AC189948	PSORS1C1,PSORS1C2,C6orf15,TCF19,POU5F1,CCHCR1,HCG27,CDSN,FLJ25680
CH250-403H8	AC189971	DBNL,GCK,FLJ22269,YKT6,POLD2,MYL7,AEBP1,POLM
CH250-294N6	AC189958	MYST1,ZNF646,PRSS36,STX4A,ZNF668,VKORC1,FLJ42291,BCKDK,ZFP260,PRSS8,FLJ32130
CH250-15N7	AC189869	PTK9L,AY525619,PPM1M,WDR51A,GLYCTK,ALAS1,TLR9,DNAH1,TMEM113
CH250-77G6B	AC190126	HOMER3,COPE,BTBD14B,COMP,GDF1,CRTC1,DDX49,LASS1,RENT1
CH250-499D20	AC190061	EN1,ITSN1,ITSN2,ANKRD47,WBSCR23,NDUFA7,CD320,ANGPTL4,RAB11B,LASS4
CH250-425B20	AC190055	TNFSF9,CRB3,GTF2F1,MGC34725,KHSRP,SLC25A23,DENND1C,SLC25A25
CH250-174K8	AC190294	TNFSF9,CRB3,MGC34725,KHSRP,SLC25A23,DENND1C,SLC25A25,TNFSF7

CH250-482A9	AC190063	RAMP2,AF480426,CNTNAP1,CNNM3,TUBG2,VPS25,EZH1,CCR10,FLJ21019,WNK4,XTP7
CH250-352A6	AC189964	CCL16,FLJ43826,CCL18,RDM1,LYZL6,AF449272,AF449276,CCL14,CCL23,CCL15
CH250-329B10	AC190323	MGC71993,CLEC10A,ASGR1,ASGR2,SLC16A11,C17orf49,SLC16A13,BCL6B,FUNDC2
CH250-37N14	AC189870	C9orf96,SURF1,SURF2,SURF4,SURF5,SURF6,RPL7A,ABO
CH250-116I10	AC189939	CRAT,DOLPP1,NUP188,PHYHD1,PPP2R4,FAM73B,TMEM15,SH3GLB2,DQ148047
CH250-452F20	AC190047	FLJ10726,IL18,LOC120379,AF303732,DLAT,DIXDC1,SDHD,DQ148040,BCDO2,TEX12
CH250-413H13	AC190054	DNHD1,APBB1,C11orf47,ARFIP2,HPX,TRIM2,TRIM3,FXC1
CH250-276H13	AC189957	KIRREL3,KIRREL,FLJ21103,DCPS,ST3GAL4,H17,TIRAP,SRPR
CH250-294D15	AC190304	ACTBL1,LOC440905,LOC112714,LOC90557,MGC87631,FLJ20297,ANKRD21,FLJ14346,PCQAP,POTE14,POTE15,POTE2,H2-ALPHA,DKFZp434E2321,POTE8,TUBA2
CH250-185C2	AC189867	DTYMK,GAL3ST2,THAP4,NEU4,ING5,MGC25181,ATG4B
CH250-498E23	AC190033	DQ148067,MTP18,TBC1D10A,SEC14L2,LOC550631,TBC1D10B,SEC14L3,NR2F1,SF3A1,LOC200312,DQ148025,SDC4
CH250-165J10	AC190316	WFDC9,WFDC11,WFDC13,WFDC10A,WFDC10B,WFDC6,SPINLW1,AF346414,WFDC8
CH250-20H7	AC170089	HSPC117,IPMK,RFPL1,LOC150297,RFPL2,RFPL3,SLC5A4
CH250-45C7	AC189945	ARF5,GUK1,GJA12,OBSCN,C1orf35,C1orf69,MRPL55,WNT3A,C1orf145,MS4A10,ARF1
CH250-3H1	AC190306	AY680558,COPA,PEA15,CASQ1,CASQ2,PNMA2,PEX19,WDR42A,ATP1A4
CH250-247G11	AC191453	AY116212,C1orf192,MPZ,NR1I3,APOA2,TOMM40L,FCER1G,SDHC,AY680460
CH250-311H21	AC191823	SLC6A8,PNCK,DUSP9,FLJ43855,PLXNB3,STK23,BCAP31,ABCD1,ABCD2
CH250-64F19	AC191459	BRF2,RAB11FIP1,RAB11FIP2,GPR124,PROSC,FLJ22965,SPFH2,ZNF703
CH250-437O5	AC191842	SFTPC,TLL2,HR,BMP1,PAFAH1B1,EPB49,AF361864,C8orf20,NUDT18,LGI3,U06694,RAI16
CH250-445K15	AC191839	LOC440295,LOC388152,FLJ46079,NGRN,MGC75360,LOC390637,FLJ43276,FLJ40113,FLJ22795
CH250-157K15	AC191448	C14orf92,NFKBIB,METTL3,ACTR3,OR10G3,RAB2B,OR10AG1,SALL2
CH250-3O5	AC191945	ZBTB33,GBGT1,OBP2A,OBP2B,AF071830,ABO
CH250-330M19	AC191828	KIAA0319,TTRAP,THEM2,C6orf62,GMNN
CH250-123P9	AC191243	SLC22A7,SRF,TTBK1,TTBK2,CRIP3,PARC,ZNF318,C6orf108
CH250-103F20	AC191814	HLA-DPB1,COL11A2,COL22A1,COL23A1,VPS52,HSD17B8,COL7A1,RING1,COL18A1,SLC39A7,RXRA,RXRB
CH250-354M14	AC191834	ZNF394,ZNF655,ZNF498,PTCD1,DKFZp727G131,CPSF4,ZNF70,LOC285989,ZFP95,ATP5J2
CH250-193G17	AC192348	NOS1,NOS3,ATG9B,CDK5,ACCN3,SLC4A2,KCNH2,ABCB8

CH250-160H4	AC191437	ZAN,TRIP6,ACHE,EPHB4,AY428851,ARS2,AY428852,LOC402682,SLC12A9
CH250-264K21	AC191438	FAM96B,FLJ37464,PDP2,CBFB,FLJ21736,RRAD,B3Gn-T6,CDH16,CES2
CH250-510B20	AC191965	SYNGR4,KLK4,FLJ46266,FLJ10922,GRIN2C,GRIN2D,KLK14,PSCD2,KCNJ14,KDELR1,GRWD1
CH250-348K7	AC191835	AY680489,CDC26,RNF183,PRPF4,WDR31,SLC31A1,SLC31A2
CH250-205A19	AC191461	BSPRY,HDHD3,ALAD,POLE3,WDR31,C9orf43,RGS3
CH250-171F5	AC191244	PHF19,PSMD5,ZSWIM1,TRAF1,FBXW2,C5,CHAF1A
CH250-440C9	AC191841	FOLR3,FLJ20625,DKFZP564M082,LRRC51,RNF121,IL18BP,NUMA1
CH250-427I15	AC191843	STX3A,TCN1,MRPL16,GIF
CH250-320G22	AC191826	FLJ36198,MS4A2,MS4A3,MS4A4A,MS4A6A
CH250-278C9	AC191451	FLJ12529,FLJ20487,HSPC196,DDB1,CYBASC3,MGC13379,DAK
CH250-402B2	AC172325	MGC13125,KIAA0999,APOA1,APOA4,APOA5,APOC3,ZNF259
CH250-85J15	AC191812	IAPP,GYS2,SLCO1A2,RECQL,GOLT1B,MGC10946,FLJ22028
CH250-374E17	AC171635	GART,IFNGR2,SON,DONSON,C21orf55,TMEM50B
CH250-269F23	AC143177	ARVP6125,TCP10L,C21orf63,C21orf59,C21orf77,SYNJ1,SYNJ2
CH250-34J20	AC169831	CRAT,DOLPP1,NUP188,LOC401233,PPP2R4,FAM73B,SH3GLB2,DQ148047
CH250-114D7	AC169793	SYT8,LSP1,FOXE3,TNNT3,MRPL23,TNNI2
CH250-161F13	AC170206	CLASP2,P4HA2,PDLIM2,PDLIM4,SLC22A4,LYST
CH250-385E15	AC172245	PGC,USP44,FRS2,FRS3,C6orf49,USP49,TFEB
CH250-205A10	AC169389	GART,SON,DONSON,ITSN1,CRYZL1
CH250-272A14	AC144284	EVX1,HOXA3,HOXA4,HOXA5,HOXA6,MEOX2,HOXA7,HOXC5,HOXA9,HOXD9,LRP2BP,HOXA10,HOXA11,HOXA13
CH250-42A15	AC192296	CCL3L3,CCL4L1,CCL18,CCL4L2,GOLPH2,TBC1D3,CCL3,CCL4,USP6,TBC1D3B,TBC1D3C,AF449272,AF457195,AF457196,AF449266,AF449267,LOC440452,CCL3L1
CH250-243I15	AC191882	LOC441931,PRKCH,C14orf106,VN1R5,GAB4
CH250-469F11	AC172118	ITGB4BP,MMP24,C20orf44,C20orf128,LOC116143
CH250-147C22	AC169789	RNPC2,RBM12,NFS1,C20orf52,SPAG4,AF345333,CPNE1
CH250-64G9	AC191983	C1orf158,PRAMEF10,PRAMEF1,PRAMEF2,PRAMEF4,PRAMEF5,PRAMEF6,PRAMEF7,PRAMEF8,PRAMEF9,LOC343066

3. Overview of Genome Features

Gene sets: Human gene sequences were used by the NCBI, Ensembl and UCSC to derive each of the available macaque gene sets as shown in (**Table S3.1**). (See below – section VI – for details of the generation of a set of conservative orthologs.)

Gene Lists	Exons	Transcripts	Genes
NCBI – Gnomen	64,515	43,198	23,088
Ensembl - Ensembl	247,383	37,031	22,804
UCSC – Nscan	199,206	22,003	22,003

Table S3.1: Protein-coding genes on chromosomes available through public portals

A comparison of the identified genes was carried out at the BCM-HGSC using part of the GLEAN pipeline (http://www.bioperl.org/wiki/Aaron_Mackey). Although a consensus gene set was not derived (the usual function of GLEAN), the different gene builds were compared as illustrated in **Figure S3.1** (a comparison of gene predictions for rhesus macaque).

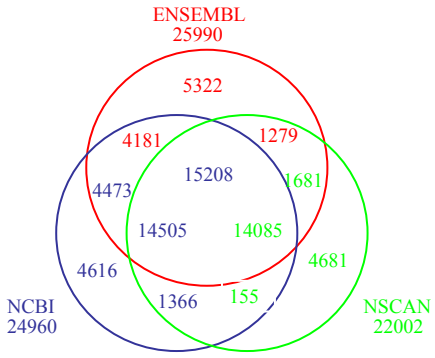


Figure S3.1: A comparison of gene predictions for rhesus macaque**.**

Evaluation and investigation of repeats:

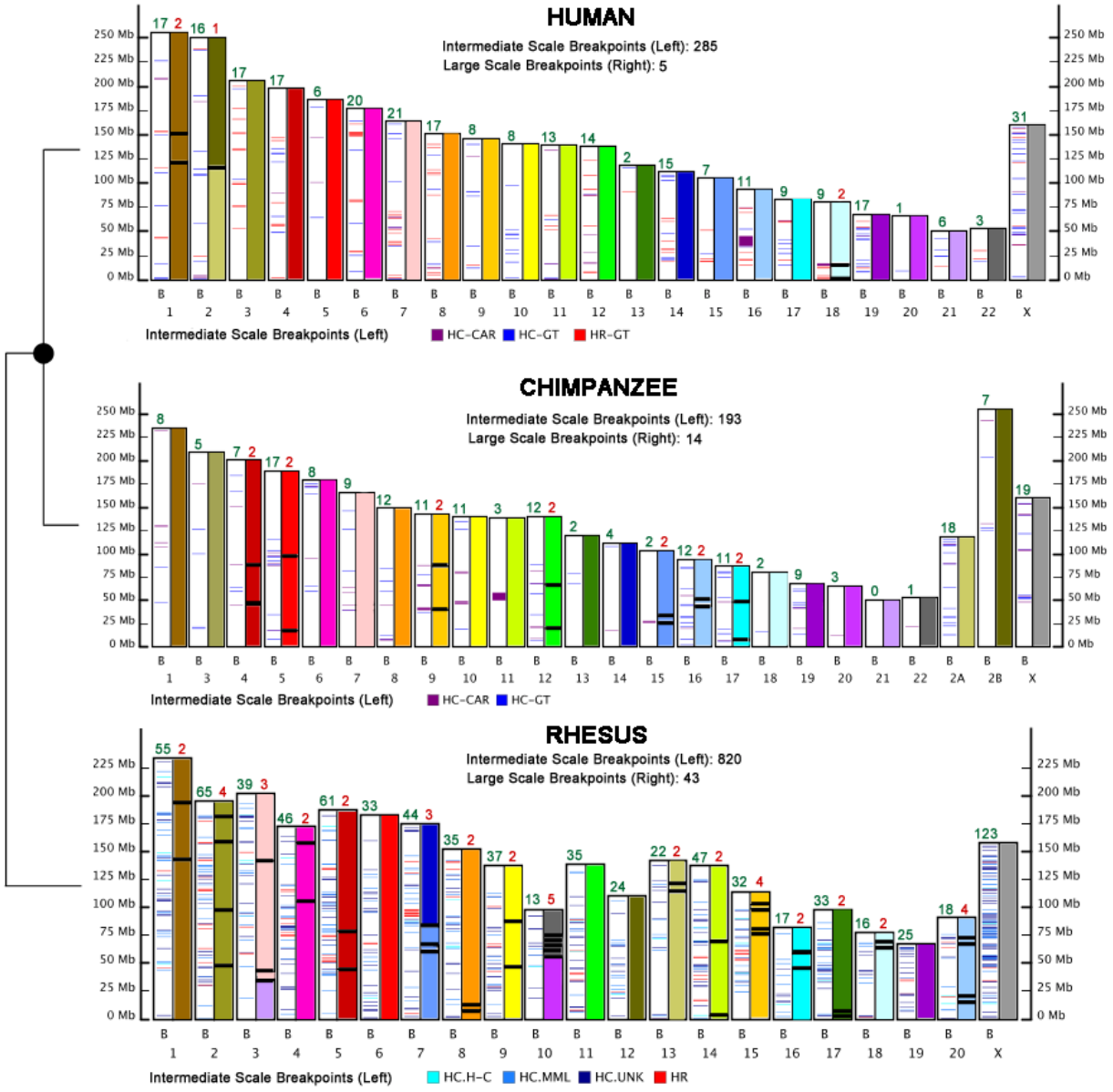
*Computational***:** To conduct a count of all repetitive elements identified in the human (hg18), *Pan Troglodytes* (panTro2), and rhesus macaque (rheMac2) genomes, we used the latest available RepeatMasked annotations (http://genome.ucsc.edu) of the three genomes. With an *in silico*, in-house script, the number of elements for each repetitive element class was counted. To avoid overcounts of fragmented larger elements (e.g. LINEs), those elements with the same ID number were counted as one element.

4 . Determining Ancestral Genome Structure

Cytogenetically visible rearrangements: Sequence alignments were generated in order to detect the lineage of chromosomal rearrangements that have occurred in the lineages (summarized in **Table S4.1**).

Table S4.1: Pericentric inversions: The Table shows determination of the lineage specificity of the pericentric inversions that distinguish the human and chimpanzee (Pan troglodytes, PTR) chromosomes using the macaque genome as an outgroup. (see the associated rhesus file).

In the main text of the manuscript, figure 3 shows a comprehensive analysis of the rhesus specific chromosomal breakpoints. **Figure S4.** Breakpoints occurring in the human, chimpanzee and macaque lineages (full figure) – shows the range of breakpoints for each of the three members of the lineage.


HUMAN
Intermediate Scale Breakpoints (Left): 285
Large Scale Breakpoints (Right): 5
Intermediate Scale Breakpoints (Left)
HC-CAR
HC-GT
HR-GT
CHIMPANZEE
Intermediate Scale Breakpoints (Left): 193
Large Scale Breakpoints (Right): 14
Intermediate Scale Breakpoints (Left)
HC-CAR
HC-GT
RHESUS
Intermediate Scale Breakpoints (Left): 820
Large Scale Breakpoints (Right): 43
Intermediate Scale Breakpoints (Left)
HC.H-C
HC.MML
HC.UNK
HR

5. Duplications in the Genome and Gene Family Expansions

Several different methods were used to assess genome and gene cluster duplications:

Genomic Duplications: The segmental duplication content of the macaque genome was assessed using three different methods; two dependent on the assembly and one based on an assessment of excess depth-of-coverage of whole-genome shotgun sequence data against RheMac2. A BLAST-based whole genome assembly comparison (WGAC) method was used to identify pairwise alignments representing, >1 kb and >90% identity (11). The results were compared to a BLASTZ self-alignment of the macaque genome that was filtered for the following parameters: minimum chained duplication length= 1kb, minimum alignment seed: 100 bp, maximum simultaneous gap size= 100 bp and weighted sequence identity >90%). Macaque high-copy repeat sequences were removed post-analysis using newly constructed macaque library of common repeats. This analysis excluded 22.4% of the alignments due to lineage-specific endogenous retrovirus and L1 expansions.

As larger, high-identity duplications (>94%) are frequently collapsed within working draft sequence assemblies (12), we compared these assembly-based results to whole genome shotgun sequence detection (WSSD) database of macaque segmental duplications. WSSD identifies regions > 10 kb in length with a significant excess of high-quality WGS reads (13) within overlapping 5 kb windows. WSSD analysis was based on a comparison of MMU 22,590,543 WGS reads against 400 kb segments of the rheMac2 assembly. 18,355,056 reads were remapped to the assembly based on the following criteria.(>94% sequence identity; >200bp non-repeat-masked bp and at least 200 bp of PhredQ>30 bp). Duplication intervals that were greater than >94% identity and > 10 kb in size (after chaining across gap regions in the macaque genome) and that were not supported by WSSD, were excluded from the genome-wide calculation of segmental duplications.

The details of the results are:

***WGAC*:** A total of 32.00 Mb (1.4%) of non-redundant sequence was detected by the BLAST-based WGAC method (>1 kb and >90% sequence identity). More non-redundant basepairs mapped to intrachromosomal duplications (22.3 Mb intra vs. 11.4 interchromosomal) (Fig. S5.1), while the number of interchromosomal duplications exceeded intrachromosomal duplications by five-fold (count of alignments=25,593 (intra) vs 5,266 (inter). Duplications were enriched near centromeres and telomeres. There was an apparent deficit (6.0 Mb) of large (>10 kb) and high identity duplications (>94%) within the macaque assembly based on the self genome comparison.

BLASTZ Analysis of Segmental Duplications: A second assembly-based method (based on analysis of BLASTZ alignments) was used to detect an additional (2.97 Mb) segmental duplications > 90% and > 1 kb in length 88.5% (22,927 kb/ 25,901 kb) of BLASTZ duplicated basepairs were shared with those detected by the WGAC analysis. The majority of the additional duplications detected by BLASTZ self-alignment chains captured additional alignments near the length and percent identity thresholds of the WGAC analysis. For example, 65.4% (3513/4820) of BLASTZ-only alignments were between 1-2 kb in length, while 54.1% (2608/4820) were

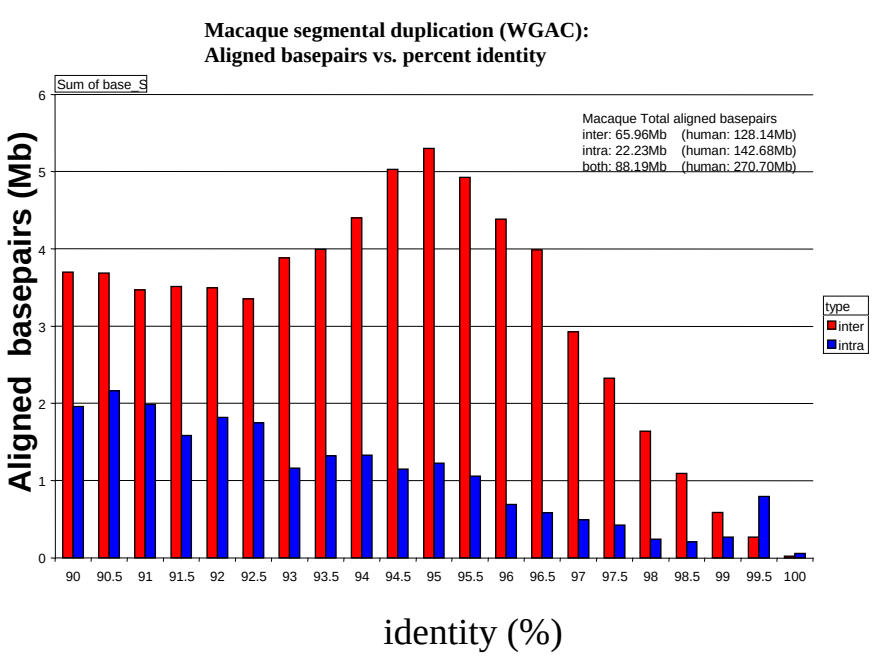
between 90-93% sequence identity. The combined set of WGAC and BLASTZ self-alignments were used to represent assembly-based segmental duplications (34.97 Mb).

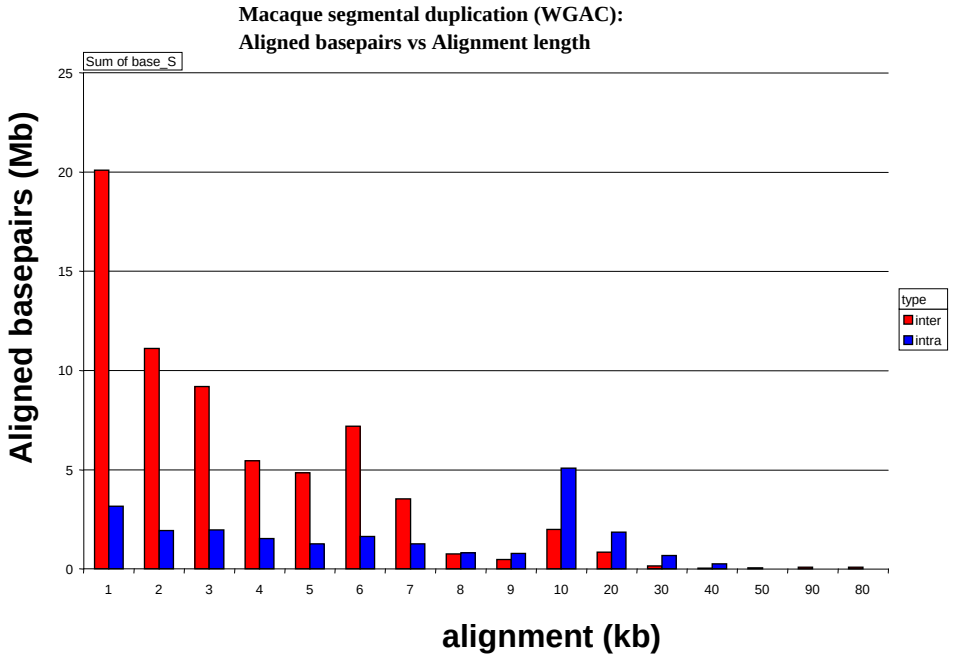
***WSSD*:** A total of 14.96 Mb of duplicated sequences (>94%; >10 kb) were predicted based on WSSD analysis (Table 1). 57% (3.4/6.0 Mb) of duplicated basepairs were shared between WGAC and WSSD at these parameters, leaving 2.5 Mb of sequence that may represent misassembled sequence. We predict that an additional 14.2 Mb of recently duplicated material is underrepresented within the macaque assembly. This includes regions associated with CCL3L, cytochromeP450, KRAB-C2H2 zinc finger, olfactory receptor, HLA and other immune/autoantigen gene families.

Estimate of total duplication: If we assume that all 2.5 Mb of unsupported WGAC duplication is artifactual and that the 14.2 Mb represents collapsed duplication, we can estimate the total duplication of macaque to be ~ 43.7 Mb (not corrected for copy number) or 1.5% of the current assembly. Chromosomes 8, 9 and 19 and show the highest proportion of duplicated basepairs (~%). Not surprising, sequence from the unmapped chromosome is almost entirely duplicated (71.9% by WSSD analysis).

Figure S5.1: Sequence identity and length of Macaque segmental duplications. The total number (WGAC) of aligned bases was calculated and binned based on per cent sequence identity (a) and alignment length (b).

Supplementary Figure S5.1: WGAC Statistics (>90% > 1 kb)





Sequence identity and length of Macaque segmental duplications. The total number (WGAC) of aligned bases was calculated and binned based on per cent sequence identity (a) and alignment length (b).

The results of detection of duplications by three different procedures, in each chromosomal region is shown in **Table S5.1.**

Table S5.1: Duplications detected in the rhesus genome by three complementary methods. (see the associated rhesus file).

A summary of the duplications by chromosome is in **Table S5.2.**

Table S5.2: Summary of Duplications in the Rhesus Macaque Genome

Chrom.	Size (mb)	NR Seg Dup (kb)	Percent Duplicated
1	228	4,350	1.91
2	190	1,180	0.62
3	196	4,748	2.42
4	168	2,702	1.61
5	182	1,344	0.74
6	178	2,032	1.14
7	170	4,778	2.81
8	148	6,101	4.13
9	133	6,747	5.06
10	95	2,854	3.01
11	135	2,035	1.51
12	107	1,147	1.08
13	138	5,214	3.78
14	133	2,705	2.03
15	110	2,691	2.44
16	79	2,281	2.90
17	94	2,403	2.54
18	74	638	0.87
19	64	2,847	4.42
20	88	983	1.11
X	154	4,766	3.10
chrUn	0.4	317	71.89
Total	**2,864**	**66,748**	**2.33**

The methods for experimental detection of gene variation by cDNA array CGH analysis of macaque are as follows: Genomic DNAs from three individual macaques were obtained from Leslie Lyons at the University of California, Davis, California National Primate Research Center, and used for cDNA array CGH analysis under conditions previously described (3). Three separate array CGH pair-wise comparisons were conducted involving rhesus macaque (test) and human (reference) genomic DNAs. To minimize false positive signals, more stringent selection criteria were used here than reported in Fortna et. al (3). Selection criteria used to identify genes that showed array CGH-predicted increases in copy number in macaque relative to human were as follows: for a given cDNA to be selected as "increased in macaque", all macaques tested (i.e. 3 out of 3) had to exhibit an array CGH value ($\log_2$ ratio of test over reference) of >0.5 for that cDNA and for at least one adjacent (based on genome position) cDNA. All human vs human values had to be <0.5 for those cDNAs exhibiting macaque increases. Full array CGH data for the macaque comparisons has been deposited in the Stanford Microarray Database (SMD) (http://genome-www.stanford.edu/microarray).

Table S5.3: array CGH data for gene gains in macaque relative to human. Column A represents the IMAGE Clone number that corresponds to the cDNA sequence spotted on the human cDNA microarray for the gene gains in macaque relative to human. Column B lists chromosomal location and nucleotide position of the cDNA sequences according to hg13 as well as additional descriptors. Columns C-H represent array CGH $\log_2$ fluorescence ratios for individual primates surveyed. An array CGH $\log_2$ fluorescence ratio of greater than 0.5 is indicative of a copy number increase in macaque relative to human. (see the associated rhesus file).

Computational comparison of macaque array CGH results: From the 124 array-based comparative genomic hybridization (array CGH)-predicted IMAGE clones that show macaque increases relative to human, the clones were sorted by chromosome and nucleotide position and grouped according to gene in order to condense the clones into a non-redundant gene list. IMAGE clone IDs were extracted from this "macaque>human" array CGH copy number set and their GenBank accessions were acquired from the SOURCE database (http://source.stanford.edu) for all IMAGE clones. The accessions were then used to retrieve EST sequences via NCBI's "Batch Entrez" application (http://www.ncbi.nlm.nih.gov/entrez/batchentrez.cgi?db=Nucleotide) and a differential BLAT analysis was conducted using these sequences.

RheMac2 and hg18 genomes were downloaded from UCSC and indexed on our in-house BLAT server, running on an OpenSuSe Linux machine. The EST sequences for our 124 array CGH-predicted IMAGE clones were repeat masked using RepeatMasker (http://www.repeatmasker.org) and the October 06, 2006 version of Repbase (http://www.girinst.org) and used as BLAT queries against each genome (rheMac2 and hg18). Those results were filtered for matches scoring above 200 and BLAT hit counts for each IMAGE clone in our set of 124 predicted clones were compared using a custom Perl script to find those predicted by BLAT to have increased copy number. Random gene copy number variants were simulated using Perl by randomly selecting an IMAGE clone and its neighbor from the master data set of all clones on our array. Five random sets were generated and the procedure detailed for our BLAT analysis was run on each set to get a measure of the number of clones that would be expected to show increases through chance. Fisher's Exact Method was computed in R (http://www.R-project.org) and the p-value between

the 5 random datasets (4/275) and the array CGH-predicted macaque increases consistent with BLAT results (28/51) was determined to be <0.0001.

Sequence information for the computationally-predicted genes showing copy number gains (data from the University of Indiana) was obtained using the web-based BioMart at Ensembl (http://www.ensembl.org/biomart/index.html) to download sequences for each Ensembl ID in their dataset. A local BLAST database was built by running formatdb on these sequences and our EST sequences were used as BLAST queries against this database with an expect cutoff of 10^{-10} to obtain significant alignments. The BLAST score report was further screened using BioPerl's BLAST report methods for sequences that had High-scoring segment pairs (HSPs) of more than 50 residues and greater than 75% agreement. All IMAGE clones that passed through these tests were posited to agree with computationally-predicted gains in copy number in the macaque relative to human.

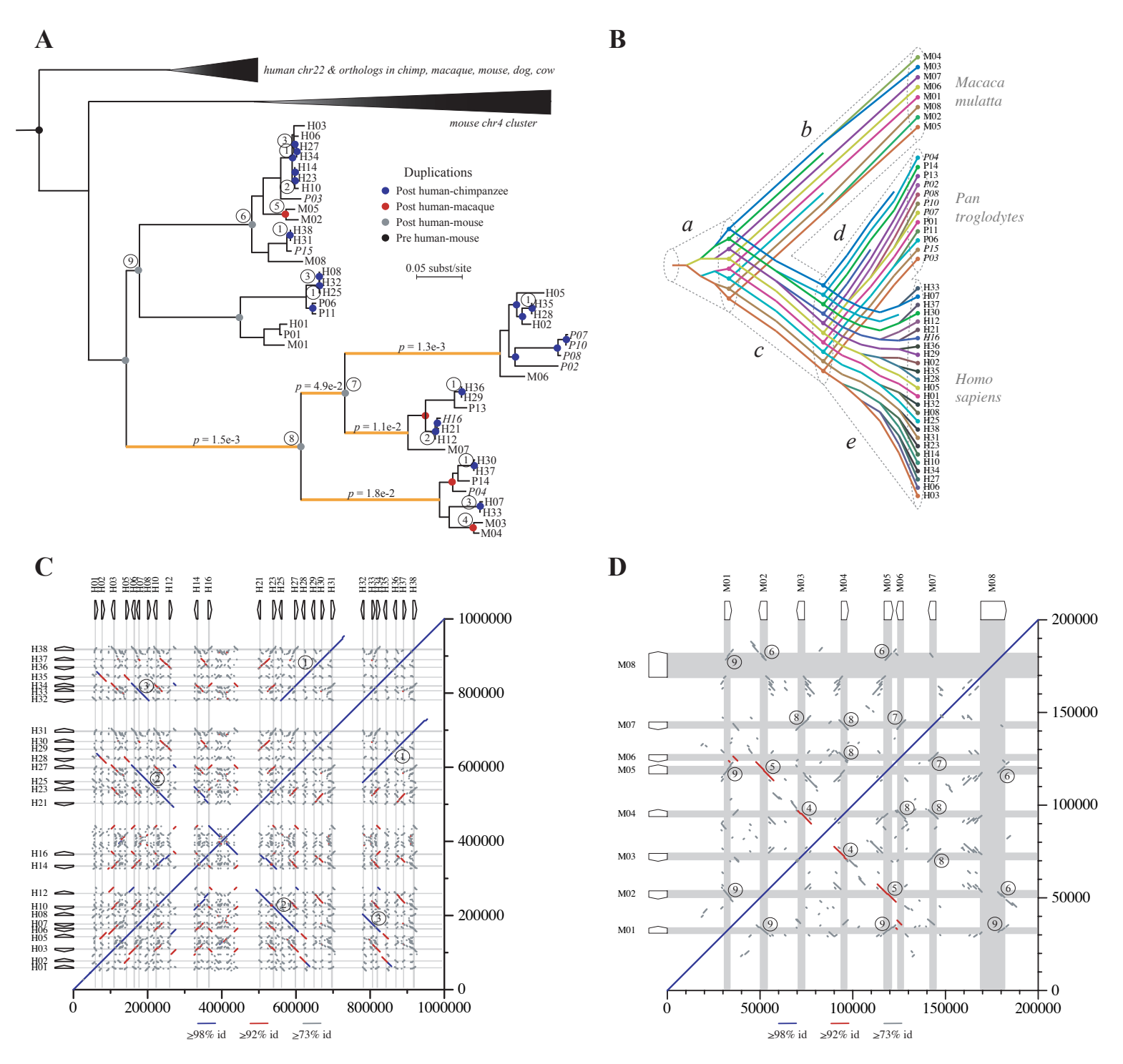
22 HLA-related genes located in the region orthologous to human chromosome 6p21 were discovered – see **Table S5.3**).

Table S5.4: **array CGH values for HLA Class I-related genes among macaque and hominoid lineages**. Column A represents the IMAGE Clone number that corresponds to the cDNA sequence spotted on the human cDNA microarray for the HLA Class-I related genes. Column B lists chromosomal location and nucleotide position of the cDNA sequences according to hg13 as well as additional descriptors. Columns C-W represent array CGH $\log_2$ fluorescence ratios for individual primates surveyed. An array CGH $\log_2$ fluorescence ratio of greater than 0.5 is indicative of a copy number increase in the non-human primate relative to human. The array CGH data for human and great ape lineages was obtained from that reported by ([3]). (see the associated rhesus file).

Prame Gene Cluster Mapping:

In order to characterize the PRAME gene cluster, we examined (panTro2) and the rhesus contig (assembled from 8 finished BACs), by aligning the human genes predicted by Birtle et al. 2005 (14) and requiring valid start codons and splice signals. Not all of the predicted coding regions end with a stop codon, because some of them have longer splice-isoforms. The self-alignments pictured in the main text in Figure S5.4 C and D were computed by the Blastz program (15) with default parameters.

Figure S5.2: The PRAME gene cluster. Panels (A) and (B) are reproduced from the published paper. (C) Dot plot of the PRAME region on human chromosome 1 vs. itself. Blue, red, and gray lines indicate the levels of similarity expected for duplications following the human-chimpanzee, human-macaque, and human-mouse divergences, respectively. (D) Similar dot plot for macaque, based on a contig assembled from six BAC clones.


A
human chr22 & orthologs in chimp, macaque, mouse, dog, cow
mouse chr4 cluster
Duplications
Post human-chimpanzee
Post human-macaque
Post human-mouse
Pre human-mouse
0.05 subst/site
p = 1.3e-3
p = 4.9e-2
p = 1.1e-2
p = 1.5e-3
p = 1.8e-2
B
Macaca mulatta
Pan troglodytes
Homo sapiens
a
b
c
d
e
C
≥98% id
≥92% id
≥73% id
D
≥98% id
≥92% id
≥73% id

Array CGH and statistical analyses also identified duplications and expansion in the HLA loci of the rhesus macaque. These are illustrated in Figure S5.3 - Expansion at the Rhesus Macaque HLA Locus.

Figure S5.3: Expansion at the Rhesus Macaque HLA Locus: A) Human chromosome 6, with Treeview Image of 6p21 containing HLA Class I-related genes. B) Enlarged Treeview Image C) A gene tree for the family of HLA Class I-related genes in macaque, human and chimp (following page).
.

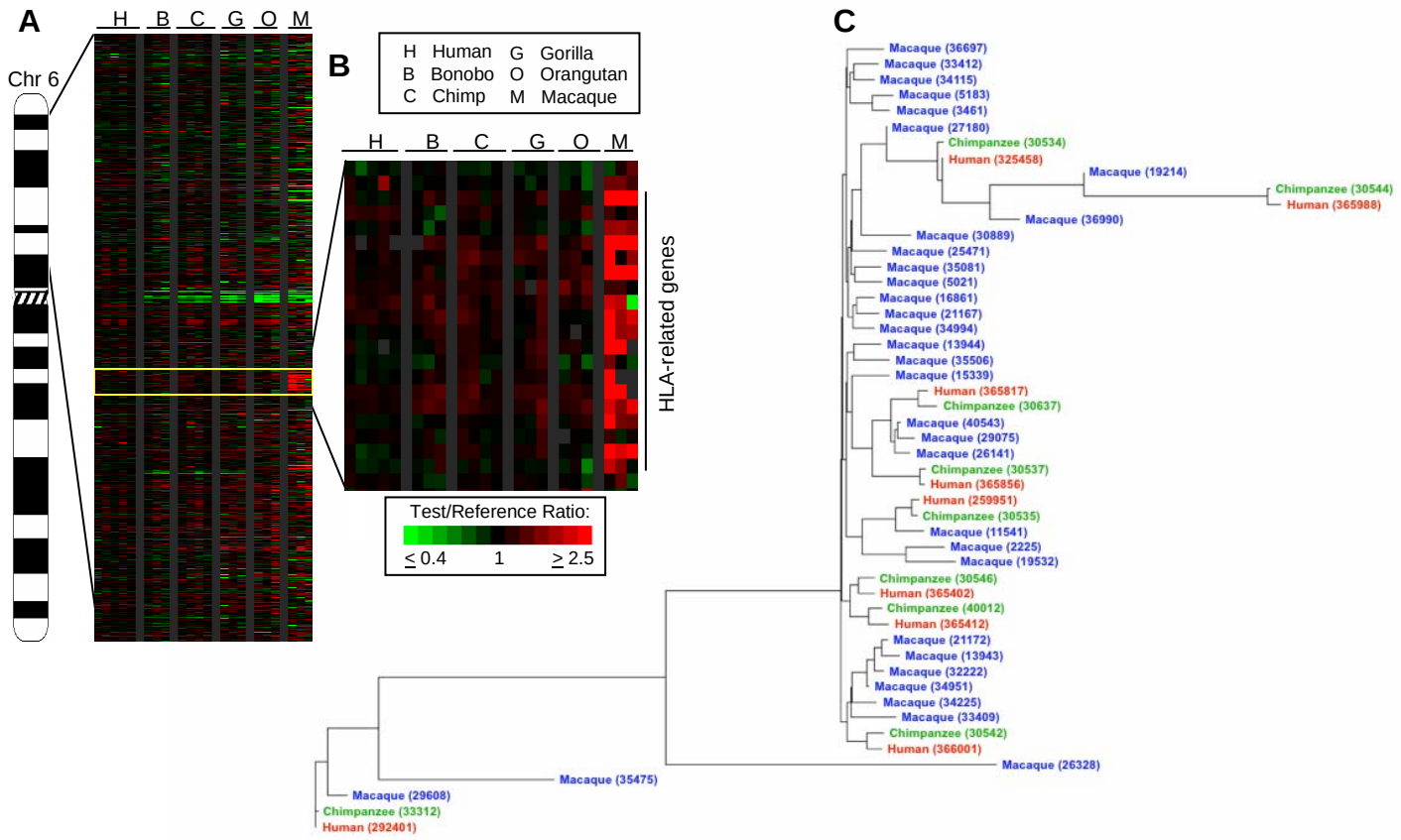


Figure S5.3 Expansion at the Rhesus Macaque HLA Locus

6. Orthologous Relationships

Pipeline for Ortholog Identification

The starting point for our orthology pipeline was the union of all annotated human genes in the RefSeq (downloaded on June 26, 2006) (16), knownGene (KGII) (17) and VEGA (build 30) (18) sets. Incomplete or short (<100bp) coding regions were eliminated, as were gene structures that had incorrect start codons, stop codons, splice sites, or in-frame stop codons with respect to the human genome assembly (hg18). The remaining transcripts were clustered based on overlap of coding sequences, resulting in 21,256 gene clusters.

Using the MULTIZ program (19), whole-genome multiple alignments were constructed for the macaque genome and the most recent human (hg18), chimpanzee (panTro2), mouse (mm8), rat (rn4), and dog (canFam2) genome assemblies from UC Santa Cruz (20), with hg18 as the reference genome. To reduce the likelihood of paralogous alignments, the pairwise syntenic nets (21) with respect to the human genome were used as input to MULTIZ. The human gene annotations were then mapped to these multiple alignments based on their hg18 coordinates.

The candidate gene structures and corresponding alignments were then subjected to a series of filters designed to minimize impact on subsequent analyses of annotation errors, sequencing and assembly errors, pseudogenes, nonorthologous alignments, and nonconserved gene structures. To maximize flexibility, this analysis was done in a pairwise fashion, based on pairwise alignments of hg18 with each non-human genome, as implied by the MULTIZ multiple alignments. Briefly, we performed the following tests for each candidate human gene and each pairwise alignment.

1. *Synteny filter*: The gene must map to the non-human genome (NHG) via a single chain of sequence alignments including at least 80% of its coding sequence (CDS). This alignment chain must meet the length and score thresholds required for inclusion in the UCSC syntenic net.
2. *Completeness filter*: All CDS bases must either be aligned, be accounted for by microindels evident from the MULTIZ alignments, or correspond to known sequencing gaps in the NHG. No more than 10% of CDS bases may fall in sequencing gaps.
3. *Frame-shift filter*: Frame-shift indels in CDSs are prohibited, unless they are compensated for within 15 bases. (In this case, the intervening region is masked with 'N's for downstream analysis.)
4. *Conserved exon-intron structure filter*: According to the alignments and genome assemblies, start codons, stop codons, and splice sites must be conserved in the NHG, and nonsense mutations may not be present. Thus, the human CDS and the aligned region of the NHG must with high probability represent exactly the same complete CDS.
5. *Recent duplication filter:* Many recently duplicated genes fail the synteny filter, but some slip past this filter. Therefore, we also eliminated any gene in any species that was more similar (in terms of CDS identity) to another annotated gene in the same species than would be expected if the divergence of these paralogs had preceded the earliest speciation event in question (e.g., the human/macaque speciation for the primate analysis, or the

human/mouse speciation for the primate/rodent analysis). To estimate the identity threshold, we conservatively used the 5th percentile of percent identities observed across species (e.g., between human and macaque genes, or human and mouse genes) in the CDS alignments of genes that had passed all other filters (i.e., we erred on the side of rejecting too many genes for being too similar to other genes in the same genome). Note that a gene that has recently spawned a pseudogene or that has an unannotated paralog will not be discarded by this filter.

All filters were applied to all gene transcripts, ignoring clustering, and the results were stored in a database. Multi-way sets were obtained by combining the results of pairwise filters. For example, a transcript passed a filter with respect to human/chimp/macaque if and only if it passed that filter with respect to human/chimp and human/macaque. Finally, at most one transcript per cluster was selected for downstream analysis. If multiple transcripts in a single cluster passed all filters, the one with the longest CDS was selected. In the event of a tie, preference was given to RefSeq genes, then Vega genes, then KGII genes. Further ties were resolved arbitrarily.

The numbers of genes that passed various filters are shown in the main manuscript as **Figure 6**. Here, the numbers shown represent clusters, not individual transcripts—i.e., if any transcript in a cluster passes the filter in question, then the cluster (loosely, the gene) is counted as passing the filter.

For each set of 1:1 orthologs that passed these filters, multiple alignments of CDSs of all genes were constructed by extracting and concatenating fragments from the MULTIZ alignments corresponding to CDS exons. When a subset of species was considered (e.g., the three primates), sequences for the other species were removed and columns with all gaps were discarded. The MULTIZ alignments were of generally high quality and no realignment was deemed necessary. All alignments of CDSs were verified to have multiple-of-three lengths and no stop codons.

Contribution of Sequencing and Assembly Error

We reran our pipeline for 294 genes, using 81 finished BAC sequences in place of the macaque assembly. The BACs resulted in a 16% increase in the number of genes passing all filters. In the regions in question, very few genes failed the synteny or duplication filters, so these results reflect the improvement in CDS completeness and apparent gene structure conservation gained by substituting finished BACs for the draft assembly (stages 2–4 of the pipeline; see above). Most of the improvement occurs because gaps are filled or apparent frame-shift indels are revealed to be artifacts of sequencing error.

Of our starting set of 21,256 genes, 18,886 passed the synteny filter with respect to macaque and 5,526 of these were discarded by filters 2–4, leaving 13,360 that passed filters 1–4 for macaque. Extrapolating from the BAC analysis, finished sequence would be expected to improve the number that passed filters 1–4 by 16%, to 15,498. (We ignore filter 5 in this analysis, as it has only a small effect.) Therefore, an expected $13{,}360 \times 16\% = 2{,}138$ genes—or 39% of the 5,526 genes eliminated by filters 2–4—were discarded due to flaws in the macaque assembly. The

remaining 5,526 - 2,138 = 3,388 were discarded for a combination of reasons, including annotation errors[1], alignment errors, and genuine frame-shift indels or changes to exon-intron structures. It is difficult with the data we have to distinguish among these causes.

In addition to the 5,526 genes that fail filters 2–4 with respect to macaque, 1,011 fail because they are syntenic with respect to macaque but not chimpanzee, and another 1,250 because they pass all filters but filters 2–4 in chimpanzee. Thus, 2,261 genes are discarded from our orthologous trios purely due to the chimpanzee genome. Of the 1,250 that fail filters 2–4 in chimpanzee, about a third fail filter 2 (completeness), typically due to assembly gaps in chimpanzee, and about half fail filter 3 (frame-shifts), frequently because of sequencing error (judging by quality scores). In addition, many of the 1,011 that fail filter 1 (synteny) with respect to macaque can be explained by assembly gaps, and many that fail filter 4 (exon-intron structure) also appear to be artifacts of sequencing error. We conservatively estimate that at least half of the 2,261 genes that fail only the chimpanzee filters, or 1,130 genes, can be attributed purely to flaws in the chimpanzee assembly.

Taking this analysis one step further, we can estimate approximately how many genes would be "rescued" by finished genome sequences for both chimpanzee and macaque. As discussed above, at least about 1,130 would be rescued from finished chimpanzee sequence alone, and another 2,138 would no longer fail the macaque filters. However, some of these 2,138 would still fail the chimpanzee filters. Based on the BAC analysis, about 80% of genes rescued in macaque still fail the filters with respect to the draft chimpanzee genome, probably because the locations of assembly flaws in the two genomes are correlated. However, reasoning as above, at least half of these 80% would be recovered with finished chimpanzee sequence. Thus, at most $2{,}138 \times 40\% = 855$ genes would be lost, and the remaining $2{,}138 - 855 = 1{,}283$ would be rescued. Adding the 1,130 rescued from chimpanzee alone, we obtain a total of 2,413 genes. Thus, finished chimpanzee and macaque genomes would allow the total number of genes passing our stringent filters to increase to an estimated $10{,}376 + 2{,}413 = 12{,}789$, an increase of 23%.

Analysis of d_N/d_S

Maximum likelihood estimates of $\omega = d_N/d_S$ for each gene were obtained using the codeml program (22), with F3 × 4 codon frequencies, equal amino acid distances (aaDist=0), a single ω across sites and across branches (model=0, NSsites=0), and separate estimation of κ per gene (fix_kappa=0). The tree topology shown in Figure S6.2 was assumed. The significance of the difference between the primate and rodent ω distributions was evaluated by a one-sided Mann-Whitney U test. To obtain separate estimates of ω for each branch, codeml was run on a concatenation of all CDS alignments, with options as above but with model=1 (the "free-ratio" model).

Likelihood Ratio Tests for Positive Selection

Test T_A is essentially the test of "site models" 2a versus 1a that was introduced by Nielsen and

[1] The set of 5,526 appears to be enriched for weakly supported genes from the UCSC Known Genes set.

Yang (23) and subsequently refined by Wong et al. (24) and Yang et al. (25). To reduce the number of parameters estimated per gene, the complete set of 10,376 genes was divided into eight equally sized classes by G+C content in third codon positions, the branch lengths and κ were estimated separately for each class (under model 1a), and these estimates were subsequently held fixed, in a G+C-dependent way, for the LRTs. Instead of a complete set of branch lengths, a single scale parameter for the branch lengths (μ) was estimated per gene. Thus, only the parameters μ, ω_0, and p_0 for model 1a, and the additional parameters ω_2 and p_1 for model 2a, were estimated per gene. This parameterization is not supported by codeml, so we developed our own software for parameter estimation and likelihood computation.

We computed nominal *P*-values empirically, based on simulation experiments. 10,000 alignments were simulated under the nearly neutral model for each G+C class using the 'evolver' program in the PAML package (22). Alignment lengths and values of μ, ω_0 and p_0 were drawn from the empirical distribution defined by our 10,376 alignments (using estimates obtained under the nearly neutral model), and the other parameters (κ, the codon frequencies, and the branch-length proportions) were fixed at global estimates from the data for each G+C class. Log likelihood ratios were computed for these simulated data sets, exactly as they would be computed for the real data. The nominal *P*-value for a log likelihood ratio of *r* was defined as the fraction of all simulated alignments for which a log likelihood ratio greater than or equal to *r* was observed. If fewer than 10 simulated alignments had log likelihood ratios greater than *r*, then the *P*-value was approximated by assuming $2r$ should have a χ^2 distribution with one degree of freedom under the null hypothesis. (Our simulation experiments suggested that this approximation is quite good for small *P*-values.) The method of Benjamini and Hochberg (1995) (26)was used to estimate the appropriate *P*-value threshold for a false discovery rate of <0.1.

Tests T_H, T_C, and T_M are essentially test 2 of Yang and Nielsen (27) (see also Zhang et al. (28)except that, as above, κ and the branch-length proportions were estimated by pooling genes within the same G+C class, and only a scale parameter μ was estimated per gene. In this case, the *P*-values were computed using a 50:50 mixture of a χ^2_1 distribution and a point mass at zero, as supported by the simulation experiments of Zhang et al. (28).

Post-processing Filters for Low Sequence Quality

In an initial application of the LRTs, we found that the genes identified as being under positive selection were strongly enriched for overlap with low quality sequence (LQS) in the chimpanzee assembly, and slightly enriched for overlap with macaque LQS. On further inspection, many of these genes were false positives resulting from clusters of apparent nonsynonymous substitutions that were actually miscalled bases. To address this problem, we systematically identified all apparent nonsynonymous substitutions between human and chimp or between human and macaque that correspond to regions with quality <20 in the chimpanzee or macaque assemblies, respectively. We then discarded all genes for which >20% of nonsynonymous substitutions with respect to chimp or macaque, coincided with low quality regions of the corresponding genome. Manual inspection of 25 individual cases suggested this was a reasonable criterion. This led to the rejection of 257 genes due to low quality chimp sequence and 59 genes due to low quality

macaque sequence. It reduced the number of genes identified by T_A from 84 to 67, by T_C from 72 to 14, and by T_M from 134 to 131. (The two genes identified by T_H were unaffected.) Thus, chimpanzee LQS had a much larger effect on our results than did macaque LQS, and the most dramatic impact was on the genes identified as being under selection on the branch to chimpanzee.

Comparison of Number of Genes with Previous Studies

As noted in the text, Nielsen et al. (2005) ([29]) reported only 35 genes with nominal $P < 0.05$, and when considering multiple comparisons, were only able to establish that the 5% FDR set was nonempty. In contrast, the use of macaque allows 15 genes in hominins to be identified with FDR <0.1, plus another 163 genes for other branches of the three-species phylogeny. In addition, the macaque genome allows one to distinguish between selection on the branch to human and selection on the branch to chimpanzee.

A comparison with Clark et al. (2003) ([30]) is less straightforward, because this study made use of two different tests and various nominal *P*-value thresholds. The set of genes most comparable to the ones reported here was a set of 28 genes with nominal $P < 0.001$ according to "Model 2," an LRT similar to our T_H. In comparison, T_H identifies 16 genes with nominal $P < 0.001$, only the top two of which meet the FDR < 0.1 criterion. Thus, the use of mouse instead of macaque would seem to improve power slightly, although it is also possible that our stringent filters allowed some spurious cases of positive selection to be avoided. Our simulation results (see below) indicate that macaque is slightly preferable to mouse when positive selection is strong, but mouse is slightly preferable when selection is weak.

Gene Classification Analysis

For each of the 10,376 orthologous gene trios, we searched for class assignments in the Gene Ontology (GO) (31) and PANTHER databases (32), using the identifiers of overlapping transcripts for cross-referencing if necessary. At least one GO category was obtained for 8637 genes (83%) and at least one PANTHER category for 8499 genes (81%). We looked for categories significantly overrepresented in sets of genes predicted to be under positive selection, as compared with the background set of 10,376. For a given category *C* and subset of genes *S*, a 2×2 contingency table was created for the numbers of genes within and outside *S*, and assigned or not assigned to *C*, then a (one-sided) *P*-value for independence of rows and columns was computed by Fisher's exact test. These *P*-values were then corrected for multiple comparisons using the method of Holm (1979) (33) . Each gene was considered to belong to all parent categories of the ones to which it was directly assigned.

The above analysis is only useful for relatively large subsets of genes (e.g., as identified by test T_A), so we performed an alternative analysis based directly on the log likelihood ratios from each test (29). For each gene category *C*, this analysis compares the distributions of log likelihood ratios of genes within and outside *C* using a Mann-Whitney *U* (MWU) test. A one-sided MWU *P*-value is computed, reflecting a shift toward larger log likelihood ratios within category *C*. These

nominal MWU *P*-values are then corrected for multiple comparisons. This method is capable of identifying categories whose genes show signs of positive selection, even if few of those genes meet our stringent thresholds for significance. On the other hand, it may also identify categories of genes that merely show a tendency for relaxation of constraint.

Power Analysis

We repeated the study of positive selection on human-chimp-macaque but omitted the macaque sequence from the 10,376 orthologous gene trios. Left with pairwise alignments and an unrooted tree consisting of one branch, we could only perform the test for positive selection across branches (T_A). We refer to this test applied to the human/chimpanzee data as T_A'. For the data set of 10,376 human-chimpanzee alignments, test T_A' identified 20 genes with FDR <0.1. Therefore, test T_A, with its 67 genes, would appear to increase the power to detect positive selection in primates more than threefold. On the other hand, tests T_H and T_C together identify 15 genes, somewhat fewer than identified by T_A.

In fact, tests T_A, T_H, T_C, and T_A' all detect slightly different classes of genes, so any comparison among them is imperfect. For example T_A may detect genes experiencing moderate positive selection on both the human and chimpanzee branches, while these genes may not be identified by T_H, T_C, or even T_A (if they are not also under selection on the macaque branch); and T_A may identify genes that are experiencing strong positive selection on the branch to macaque, but weak selection or no selection on the branches to human or chimpanzee. Nevertheless, the fact remains that the macaque genome allows a total of 178 genes to be identified as undergoing positive selection on one or more branches of the three-species primate phylogeny, while human--chimpanzee comparisons based on the same methods allow only 20 genes to be identified.

We also evaluated by simulation the power of both types of tests of positive selection—the T_A test detecting selection across branches and the tests detecting selection on one branch of the phylogeny. For the single-branch case, we focused on test T_H. Using 'evolver' (34), we generated 1000 alignments for each of a range of values of ω. The other parameters, including the transition/transversion ratio κ, the codon frequencies and the branch lengths, were fixed at values estimated from the data. In the case of T_A, we simulated alignments of lengths 200 and 500 codons, and assumed constant ω among lineages and among sites (model M0). We then applied test T_A to these simulated data sets exactly as we had applied it to the real data, and for each value of ω recorded the fraction of all data sets that were predicted to be under positive selection. When $\omega \leq 1$, this fraction is an estimate of the false positive rate, and when $\omega > 1$ it is an estimate of the power of the test.

These experiments indicate that test T_A has essentially no power to detect positive selection in genes of lengths of 200 codons when only human and chimpanzee sequences are available (**Figure S6.3**). This is important because more than 20% of the genes in our set of 10,376 have lengths of 200 codons or fewer. (The median length is 370 codons and the mean is 462.9 codons.) Even for longer-than-average genes of 500 codons, the test has no power unless selection is quite

strong ($\omega > 2$). When the macaque genome is included, the power is considerably better. Short genes can still be detected only in the presence of strong selection ($\omega > 2$), but longer genes can be detected with reasonably good sensitivity when $\omega = 2$ and with lower but non-negligible sensitivity when $\omega = 1.5$. For large ω and genes of 500 codons, the macaque genome improves power by two-fold to three-fold. This is consistent with the results discussed above, based on real data. Still, it is striking how little power we have to detect weak to moderate positive selection, even with the macaque genome.

For test T_H, we simulated data sets in a similar way, but this time varied ω only for the branch to human. In this case, power is quite poor even in the presence of fairly strong selection (**Figure S6.4**). These results are generally consistent with those of Zhang et al. (2005) (28). However, the use of macaque rather than mouse as an outgroup does appear to result in a modest increase in power for large ω. When ω is smaller (weak positive selection), mouse is a slightly better outgroup than macaque.

Clustering Analysis

Using a randomization experiment, we tested whether the 67 genes identified by test T_A showed significant signs of clustering in the genome. Taking the 10,376 orthologous trios as a background set, we randomly selected subsets of 67, and for each of these subsets computed the pairwise distances between adjacent genes. (The human coordinate system was used). This experiment was repeated 100 times, yielding 4671 samples of intergenic distances under the null hypothesis of no clustering. We then compared the 46 equivalently defined intergenic distances for the identified genes with these 4671 using a onesided Mann-Whitney *U* test, and found no significant evidence of clustering ($P = 0.24$). Using the same random samples to define an empirical null distribution, we also computed a one-sided *P*-value for each of the 46 observed intergenic distances, and found no distances significantly smaller than expected after adjusting for multiple comparisons. The closest pair of genes, located on chromosome 11 (NM 198947 and NM 022074, both unannotated; 22.6 kb apart) had nominal $P = 0.0032$, and adjusted $P = 0.15$, using the method of Holm (1979) (33).

In a second experiment, we arbitrarily divided the genome into nonoverlapping intervals of 1Mb, 5Mb, 10Mb, and 25Mb in size (again using the human coordinate system), and looked for enrichments for positively selected genes within intervals. This was done using Fisher's exact test, exactly as described above for the gene classification analysis, but treating intervals as gene categories. (In this case, each gene belongs to exactly one category). None of the observed enrichments was significant after a multiple comparison correction. However, this analysis did turn up a 5Mb interval on chromosome 19 (chr19:55,000,000-60,000,000) in which three of 65 genes from the set of 10,376 are predicted to be under positive selection (nominal $P = 0.008$). Two of these genes—*LILRB1* and *LAIR1*—are members of the leukocyte receptor cluster at 19q13.4, which contains more than two dozen immunoglobulin-like leukocyte-expressed receptors. The third, *SIGLEC6*, encodes another immunoglobulin-like protein.

Finally, we looked for enrichments across chromosomes by treating each chromosome as an interval and using Fisher's exact test to identify significant enrichments. In this analysis, (human) chromosome 11 shows a moderate excess of positively selected genes, with 10 out of 593 genes

falling within our set of 67 compared with an expected 3.8, but this enrichment is not significant after adjusting for multiple comparisons ($P = 0.10$).

Comparison of Primates and Rodents

The orthology pipeline was also used to identify rodent orthologs as above. A total of 6,733 genes pass all fiters for ortholog identification for the human, macaque, mouse, rat orthology group. These genes were then subjected to d_N/d_S analysis as above, with estimates of ω made for the primate and rodent branches separately. Fishers Exact test was used with the corresponding Nd_N and Sd_S values (representing multiple hit corrected numbers of amino acid and synonymous substitutions) to assess significance values.

As is perhaps expected, the distribution of ω in primates is more dispersed than in rodents ($P < 0.001$, Wilcoxon signed-rank test) and generally skewed towards larger values. When primate and rodent ω of individual genes were compared, primate orthologs were more rapidly evolving at a 3:2 ratio. This disparity was also evident when genes showing significant differences between the two species pairs were considered separately. This may be the result of increased positive selection in the primate lineage or a relaxation of constraint in the primate lineage. Either of these possibilities is hypothesized to be exaggerated by demographic differences between the two species including smaller effective population sizes in primates. These demographic differences may also account for a decrease in efficacy of negative selection in primates.

Following strict Bonferroni correction for multiple testing, only 3 genes were significantly faster in rodents compared to 22 significantly faster in primates. If multiple testing criteria is relaxed (due to constraints from gene size and evolutionary distance, significance values are often capped short of the multiple testing corrected cutoff) and only raw significance values less than 0.001 are considered, the bias towards primates is even more dramatic (144 vs. 8). In the case of the rodents, significant differences are obtained by both an increase in rodent and a decrease in primate as observed by comparing these values to overall averages. In the case of primates, however, the significant disparities are caused almost entirely by increases in the primate ω as rodent values are statistically indistinguishable from rodent values as a whole.

Several expected categories are overrepresented among genes evolving at a significantly more rapid rate in primates including genes involved in sensory perception of smell and taste. These genes may have undergone relaxed selective constraint in primates relative to rodents or ancestral mammals. Interestingly, genes involved in regulating transcription are also overrepresented among those evolving at a significantly increased rate in primates relative to rodents.

Table S6.1: Genes evolving more rapidly in primates than in rodents

			Primate			Rodent			
Symbol		**Accession**	**ω**	**A**	**S**	**Ω**	**A**	**S**	**Description**
SAC	*	NM_018417	0.8594	130	65	0.1216	103	324	testicular soluble adenylyl cyclase
TMEM20	*	NM_153226	2.1601	39	8	0.0817	19	102	transmembrane protein 20
TSPAN8	*	NM_004616	2.0566	52	10	0.2058	42	74	tetraspanin 8
CNGB1	*	U58837	0.2425	31	36	0.0354	15	135	cyclic nucleotide gated channel beta 1
FBXL21	*	BC106753	2.8751	17	3	0.0678	8	45	F-box and leucine-rich repeat protein 21
FBXO39	*	NM_153230	0.5152	34	21	0.0476	9	56	F-box protein 39
SCN5A	*	AB158470	0.0815	36	114	0.0168	22	334	sodium channel, voltage-gated, type V, alpha subunit
SLCO4C1	*	NM_180991	0.4011	43	40	0.0903	33	146	solute carrier organic anion transporter family, member 4C1
TRIM42	*	NM_152616	0.2229	28	33	0.0317	16	135	tripartite motif-containing 42
ATP8B1	*	NM_005603	0.2006	31	61	0.0375	26	252	ATPase, Class I, type 8B, member 1
MT1G	*	OTTHUMT00000155623	0.4159	8	3	0.0119	19	251	metallothionein 1G
CGA	*	NM_000735	4.0388	28	3	0.1203	6	19	glycoprotein hormones, alpha polypeptide
PGAP1	*	NM_024989	0.3629	30	27	0.0933	35	165	GPI deacylase
LAMB2	*	NM_002292	0.2599	65	111	0.0924	72	341	laminin, beta 2 (laminin S)
ST7L	*	NM_138729	0.3333	15	18	0.0157	2	59	suppression of tumorigenicity 7 like
TNN	*	NM_022093	0.1930	67	105	0.0716	45	213	tenascin N
PCDH11X	*	AY861432	0.4410	90	85	0.1494	67	173	protocadherin 11 X-linked
DCHS1	*	NM_003737	0.2039	72	181	0.0861	81	492	dachsous 1 (Drosophila)
KIF2B	*	NM_032559	0.4052	55	45	0.1243	39	113	kinesin family member 2B
TAAR2	*	NM_014626	0.5719	15	12	0.0355	6	61	trace amine associated receptor 2
OR6N2	*	OTTHUMT00000059068	0.4695	19	17	0.0542	7	60	olfactory receptor, family 6, subfamily N, member 2
ABCB1	*	NM_000927	0.3928	53	56	0.1259	73	223	ATP-binding cassette, sub-family B (MDR/TAP), member 1
PIGA		NM_002641	0.8344	23	12	0.0913	12	50	phosphatidylinositol glycan anchor biosynthesis, class A (paroxysmal nocturnal hemoglobinuria)
LOC129530		NM_174898	1.3221	16	5	0.0866	10	42	hypothetical protein LOC129530
WDR19		AK026780	0.6928	15	9	0.0716	13	75	WD repeat domain 19
OR2AG1		NM_001004489	0.7844	43	22	0.2355	94	173	olfactory receptor, family 2, subfamily AG, member 1
IKIP		NM_153687	0.7973	25	13	0.1321	33	97	IKK interacting protein
ABCG5		NM_022436	0.4687	43	38	0.1221	51	152	ATP-binding cassette, sub-family G (WHITE), member 5 (sterolin 1)
MCCC2		NM_022132	0.3449	22	29	0.0585	17	120	methylcrotonoyl-Coenzyme A carboxylase 2 (beta)
PAPD1		NM_018109	0.8799	71	35	0.2596	59	91	PAP associated domain containing 1
FBXW9		OTTHUMT00000151939	0.4855	46	27	0.1574	26	64	F-box and WD-40 domain protein 9
ABCB4		NM_000443	0.2159	30	59	0.0560	31	224	ATP-binding cassette, sub-family B (MDR/TAP), member 4
CDON		NM_016952	0.3602	64	74	0.1453	80	234	cell adhesion molecule-related/down-regulated by oncogenes
DSG4		NM_177986	0.2715	41	60	0.0862	34	165	desmoglein 4
CPB2		NM_001872	0.5785	32	21	0.1170	25	78	carboxypeptidase B2 (plasma)
PFKM		NM_000289	0.1305	10	31	0.0061	2	118	phosphofructokinase, muscle
OR10G8		NM_001004464	0.4323	35	29	0.0837	12	54	olfactory receptor, family 10, subfamily G, member 8
OTUD6A		NM_207320	0.2847	23	12	0.0884	22	70	OTU domain containing 6A
HR		NM_005144	0.4313	75	72	0.1845	62	153	hairless homolog (mouse)
ZMYND15		NM_032265	0.3482	45	52	0.1081	28	109	zinc finger, MYND-type containing 15
GFM1		NM_024996	0.2754	22	29	0.0744	37	202	G elongation factor, mitochondrial 1
C1QBP		NM_001212	0.5474	18	12	0.0588	6	39	complement component 1, q subcomponent binding protein
PDE1C		NM_005020	0.2374	16	28	0.0267	7	91	phosphodiesterase 1C, calmodulin-dependent 70kDa
ATF5		NM_012068	0.4891	19	18	0.0430	3	35	activating transcription factor 5
ITGA6		NM_000210	0.1684	21	44	0.0415	17	162	integrin, alpha 6
MRPS9		NM_182640	0.5315	43	29	0.1675	36	86	mitochondrial ribosomal protein S9
PRPS1L1		BC062797	0.6260	9	5	0.0352	4	44	phosphoribosyl pyrophosphate synthetase 1-like 1
FRMD7		NM_194277	0.4348	33	29	0.1104	25	87	FERM domain containing 7
CBFA2T3		NM_175931	0.0839	22	59	0.0212	6	101	core-binding factor, runt domain, alpha subunit 2; translocated to, 3
PPEF2		NM_006239	0.3251	36	46	0.0836	31	133	protein phosphatase, EF-hand calcium binding domain 2
CD37		NM_001040031	0.4513	15	7	0.0703	6	32	CD37 molecule
IQSEC3		NM_015232	0.1187	39	65	0.0552	25	135	IQ motif and Sec7 domain 3
YBX2		NM_015982	0.7156	19	13	0.0431	2	24	Y box binding protein 2
NDUFS5		NM_004552	0.8765	16	7	0.0998	12	45	NADH dehydrogenase (ubiquinone) Fe-S protein 5, 15kDa (NADH-coenzyme Q reductase)
PTCD3		NM_017952	0.8138	77	37	0.3092	79	103	Pentatricopeptide repeat domain 3
SLC6A11		NM_014229	0.1208	11	27	0.0102	3	89	solute carrier family 6 (neurotransmitter transporter, GABA), member 11
RXFP3		NM_016568	0.2398	34	31	0.0930	27	91	relaxin/insulin-like family peptide receptor 3
RANBP17		NM_022897	0.2270	26	49	0.0570	23	160	RAN binding protein 17
ALDH6A1		NM_005589	0.4629	11	11	0.0528	12	98	aldehyde dehydrogenase 6 family, member A1
ELOVL6		NM_024090	2.7637	5	1	0.0103	1	29	ELOVL family member 6, elongation of long chain fatty acids (FEN1/Elo2, SUR4/Elo3-like, yeast)
IFT140		BC035577	0.0684	28	70	0.0220	8	97	intraflagellar transport 140 homolog (Chlamydomonas)
OR13J1		OTTHUMT00000052381	0.2613	42	38	0.0730	15	56	olfactory receptor, family 13, subfamily J, member 1
SLC11A2		NM_000617	0.3204	16	24	0.0522	16	125	solute carrier family 11 (proton-coupled divalent metal ion transporters), member 2
LTA4H		NM_000895	0.2977	10	15	0.0362	12	142	leukotriene A4 hydrolase

			Primate			Rodent			
Symbol		**Accession**	**ω**	**A**	**S**	**Ω**	**A**	**S**	**Description**
GRIN2A		NM_000833	0.1077	26	75	0.0342	19	197	glutamate receptor, ionotropic, N-methyl D-aspartate 2A
PDCD4		NM_014456	0.1722	8	16	0.0142	4	105	programmed cell death 4 (neoplastic transformation inhibitor)
ANG		NM_001145	1.3890	42	11	0.2865	27	35	angiogenin, ribonuclease, RNase A family, 5
DCDC2		NM_016356	0.4488	23	19	0.1045	24	89	doublecortin domain containing 2
PARP3		NM_005485	0.3372	41	34	0.1187	42	107	poly (ADP-ribose) polymerase family, member 3
FLJ10357		NM_018071	0.3464	59	82	0.1467	74	237	hypothetical protein FLJ10357
SYT3		NM_032298	0.1330	15	36	0.0203	4	82	synaptotagmin III
GOLGB1		NM_004487	0.4496	136	122	0.2559	450	684	golgi autoantigen, golgin subfamily b, macrogolgin (with transmembrane signal), 1
CHCHD8		NM_016565	1.9576	6	1	0.0314	3	29	coiled-coil-helix-coiled-coil-helix domain containing 8
PPP1R1C		AF494535	0.7986	10	5	0.0398	2	23	protein phosphatase 1, regulatory (inhibitor) subunit 1C
ETNK2		BC010082	0.7029	12	5	0.0971	18	65	ethanolamine kinase 2
C6orf163		NM_001010868	3.1859	14	2	0.1825	35	62	chromosome 6 open reading frame 163
ADAM12		NM_003474	0.4131	50	51	0.1428	51	137	ADAM metallopeptidase domain 12 (meltrin alpha)
MYH2		BC093082	0.1193	18	54	0.0155	6	107	myosin, heavy chain 2, skeletal muscle, adult
CCDC42		NM_144681	0.1769	22	25	0.0485	9	57	coiled-coil domain containing 42
RGAG1		NM_020769	0.9194	100	53	0.4367	161	179	retrotransposon gag domain containing 1
GJB6		NM_006783	0.1046	9	28	0.0001	0	51	gap junction protein, beta 6
LY6G5C		NM_025262	2.1282	20	3	0.2146	30	40	lymphocyte antigen 6 complex, locus G5C
POSTN		BC106709	0.1827	21	45	0.0460	15	135	periostin, osteoblast specific factor
SLC38A2		BC040342	0.5858	7	5	0.0438	13	113	solute carrier family 38, member 2
NDUFA10		NM_004544	0.5547	49	32	0.1589	24	53	NADH dehydrogenase (ubiquinone) 1 alpha subcomplex, 10, 42kDa
TUFM		BC001633	0.1246	10	29	0.0070	1	61	Tu translation elongation factor, mitochondrial
MCM8		NM_032485	0.2841	26	34	0.0839	34	147	MCM8 minichromosome maintenance deficient 8 (S. cerevisiae)
FLJ21062		NM_001039706	0.3073	31	33	0.1146	48	154	hypothetical protein FLJ21062
PPP1R3A		NM_002711	0.7563	102	43	0.3797	181	164	protein phosphatase 1, regulatory (inhibitor) subunit 3A (glycogen and sarcoplasmic reticulum binding subunit, skeletal muscle)
B3GALT5		NM_033170	0.3328	27	32	0.0518	14	69	UDP-Gal:betaGlcNAc beta 1,3-galactosyltransferase, polypeptide 5
ATP2A3		NM_005173	0.0837	28	76	0.0405	16	150	ATPase, Ca++ transporting, ubiquitous
PLA2G6		NM_003560	0.1324	21	46	0.0395	18	154	phospholipase A2, group VI (cytosolic, calcium-independent)
CWF19L1		NM_018294	0.4686	18	17	0.0910	24	102	CWF19-like 1, cell cycle control (S. pombe)
NRXN2		NM_015080	0.0455	14	73	0.0103	7	197	neurexin 2
KRT32		NM_002278	0.1327	23	42	0.0350	7	66	keratin 32
MARCH8		NM_145021	0.3624	7	9	0.0001	0	27	membrane-associated ring finger (C3HC4) 8
TAF5		NM_006951	0.1745	10	26	0.0252	7	128	TAF5 RNA polymerase II, TATA box binding protein (TBP)-associated factor, 100kDa
TMCO5		NM_152453	1.0023	17	6	0.1756	18	44	transmembrane and coiled-coil domains 5
RNF6		NM_005977	0.4364	27	20	0.1765	74	175	ring finger protein (C3H2C3 type) 6
HCN3		NM_020897	0.1249	17	47	0.0236	6	93	hyperpolarization activated cyclic nucleotide-gated potassium channel 3
KIAA1639		AB046859	0.2284	84	109	0.1588	80	210	KIAA1639 protein
RAB37		NM_001006637	0.2553	10	11	0.0245	2	33	RAB37, member RAS oncogene family
DHX57		NM_198963	0.1945	27	57	0.0636	53	310	DEAH (Asp-Glu-Ala-Asp/His) box polypeptide 57
EPHA1		NM_005232	0.2119	32	54	0.0784	24	125	EPH receptor A1
OR6B1		NM_001005281	0.3324	14	16	0.0607	7	53	olfactory receptor, family 6, subfamily B, member 1
OR6N1		OTTHUMT00000059067	0.4153	19	20	0.0776	14	68	olfactory receptor, family 6, subfamily N, member 1
TDRD7		NM_014290	0.2690	28	44	0.0857	48	213	tudor domain containing 7
CNDP2		NM_018235	0.1196	15	36	0.0258	6	84	CNDP dipeptidase 2 (metallopeptidase M20 family)
PTPN3		NM_002829	0.3039	20	31	0.0609	21	126	protein tyrosine phosphatase, non-receptor type 3
ARHGAP24		NM_001025616	0.2066	21	39	0.0470	20	133	Rho GTPase activating protein 24
TBRG4		NM_199122	0.4781	51	41	0.1885	50	105	transforming growth factor beta regulator 4
NDUFA4		OTTHUMT00000060201	0.6977	7	3	0.0001	0	13	NADH dehydrogenase (ubiquinone) 1 alpha subcomplex, 4, 9kDa
ZIC3		NM_003413	0.1447	6	10	0.0001	0	34	Zic family member 3 heterotaxy 1 (odd-paired homolog, Drosophila)
MGST1		NM_145764	1.3192	11	3	0.0915	6	24	microsomal glutathione S-transferase 1
OR8B12		NM_001005195	0.3339	20	23	0.0791	12	63	olfactory receptor, family 8, subfamily B, member 12
ACPP		BC007460	0.9025	28	14	0.2052	40	74	acid phosphatase, prostate
SACS		NM_014363	0.0890	53	208	0.0489	88	684	spastic ataxia of Charlevoix-Saguenay (sacsin)
LBX1		NM_006562	0.1607	5	6	0.0001	0	31	ladybird homeobox 1
AARSD1		NM_025267	0.3529	22	23	0.0795	19	80	alanyl-tRNA synthetase domain containing 1
PLA2G2D		NM_012400	0.6726	18	6	0.0964	11	28	phospholipase A2, group IID
ARHGEF4		AF249745	0.0880	12	28	0.0206	6	88	Rho guanine nucleotide exchange factor (GEF) 4
SLC34A1		NM_003052	0.1002	19	49	0.0324	10	109	solute carrier family 34 (sodium phosphate), member 1
SLC6A12		NM_003044	0.2002	33	47	0.0663	23	102	solute carrier family 6 (neurotransmitter transporter, betaine/GABA), member 12
LMBR1L		BC031550	0.1529	9	25	0.0126	2	67	limb region 1 homolog (mouse)-like
MTHFD1		NM_005956	0.2126	21	40	0.0650	27	167	methylenetetrahydrofolate dehydrogenase (NADP+ dependent) 1, methenyltetrahydrofolate cyclohydrolase, formyltetrahydrofolate synthetase
CA13		NM_198584	0.3000	10	15	0.0214	2	38	carbonic anhydrase XIII
SLC9A2		NM_003048	0.1974	29	57	0.0591	19	119	solute carrier family 9 (sodium/hydrogen exchanger), member 2
SYT12		NM_177963	0.1253	11	21	0.0238	4	60	synaptotagmin XII

			Primate			Rodent			
Symbol		**Accession**	**ω**	**A**	**S**	**Ω**	**A**	**S**	**Description**
CHRM5		NM_012125	0.5355	25	16	0.1341	22	57	cholinergic receptor, muscarinic 5
PIF1		NM_025049	0.2681	26	33	0.1063	27	110	PIF1 5'-to-3' DNA helicase homolog (S. cerevisiae)
SCD		NM_005063	0.2756	18	22	0.0525	13	70	stearoyl-CoA desaturase (delta-9-desaturase)
DCK		NM_000788	0.5912	9	5	0.0884	12	57	deoxycytidine kinase
ZNF467		NM_207336	0.1415	24	42	0.0619	16	97	zinc finger protein 467
CHRNA1		NM_000079	0.2134	13	20	0.0391	8	73	cholinergic receptor, nicotinic, alpha 1 (muscle)
DALRD3		NM_001009996	0.3812	29	32	0.1362	35	117	DALR anticodon binding domain containing 3
MRPS5		AK058160	0.7523	39	22	0.2047	24	47	mitochondrial ribosomal protein S5
ALPL		NM_000478	0.1333	14	27	0.0346	13	116	alkaline phosphatase, liver/bone/kidney
TAT		NM_000353	0.2810	14	21	0.0540	15	104	tyrosine aminotransferas
CACNA1E		NM_000721	0.0656	22	120	0.0210	14	250	calcium channel, voltage-dependent, R type, alpha 1E subunit
TBC1D23		NM_018309	0.2141	12	19	0.0499	16	120	TBC1 domain family, member 23
TMEM26		NM_178505	0.3347	18	22	0.0782	16	81	transmembrane protein 26
OASL		NM_003733	0.4412	40	28	0.1514	67	124	2'-5'-oligoadenylate synthetase-like
TSSK4		NM_174944	0.3617	22	18	0.0788	10	39	testis-specific serine kinase 4
PCM1		AK091406	0.2706	27	29	0.0999	31	104	pericentriolar material 1

* Significant after correction for multiple testing

Table S6.2: Genes evolving more rapidly in rodents than in primates

			Primate			Rodent			
Symbol		**Accession**	**ω**	**A**	**S**	**ω**	**A**	**S**	**Description**
KRTAP19-6	*	NM_181612	0.1470	70	251	0.8069	28	20	keratin associated protein 19-6
HAO2	*	NM_001005783	0.1043	7	30	0.7914	79	42	hydroxyacid oxidase 2 (long chain)
CHI3L1		NM_001276	0.1563	13	27	0.5654	122	61	chitinase 3-like 1 (cartilage glycoprotein-39)
HsG2239		OTTHUMT00000147613	0.0396	1	11	0.9057	55	30	novel protein coding gene
CHRDL1		BC002909	0.0163	1	23	0.2429	33	49	chordin-like 1
KRT19		NM_002276	0.0466	7	39	0.2109	60	78	keratin 19
DNAJB2		NM_006736	0.0001	0	18	0.2677	22	33	DnaJ (Hsp40) homolog, subfamily B, member 2

* Significant after correction for multiple testing

The following tables show further statistics and gene lists associated with the above calculations: A list of all genes identified by log-likelihood ratio tests is shown in **Table S6.3** 'Complete list of genes identified by likelihood ratio tests'; **Table S6.4**. shows 'Gene Ontology categories overrepresented among genes predicted to be under positive selection'; **Table S6.5** shows 'PANTHER categories overrepresented among genes predicted to be under positive selection'; **Table S6.6**. shows 'Gene Ontology categories showing and excess of high likelihood ratios'; and **Table S6.7**shows **'**.PANTHER categories showing and excess of high likelihood ratios.

Table S6.3: Complete List of Genes Identified by Likelihood Ratio Tests

No.	Accession	Gene Name	Chr	Description	P_A	P_H	P_C	P_M	Dup
1	AB126077	*KRTAP5-8*	11	keratin associated protein 5-8	**6.2e-16**	1.0e+00	5.0e-01	**9.7e-21**	✓
2	NM_006669	*LILRB1*	19	leukocyte immunoglobulin-like receptor	**7.2e-14**	**4.7e-14**	**1.9e-05**	1.4e-01	✓
3	NM_001942	*DSG1*	18	desmoglein 1 preproprotein	**1.1e-10**	8.5e-02	1.0e+00	1.5e-03	
4	NM_173523	*MAGEB6*	X	melanoma antigen family B, 6	**5.3e-08**	1.5e-04	**7.8e-18**	1.0e+00	✓
5	NM_054032	*MRGPRX4*	11	G protein-coupled receptor MRGX4	**5.6e-08**	5.0e-01	1.0e+00	**3.9e-09**	✓
6	NM_000397	*CYBB*	X	cytochrome b-245, beta polypeptide	**1.5e-07**	1.0e+00	4.6e-01	**1.6e-09**	
7	NM_001911	*CTSG*	14	cathepsin G preproprotein	**1.5e-07**	4.9e-01	1.0e+00	**2.5e-07**	
8	NM_153264	*FLJ35880*	3	hypothetical protein LOC256076	**1.7e-07**	1.1e-02	5.0e-01	**1.6e-04**	
9	NM_001013734	*LOC442247*	6	hypothetical protein LOC442247	**3.8e-07**	5.6e-02	5.4e-03	6.7e-03	
10	NM_000735	*CGA*	6	glycoprotein hormones, alpha polypeptide	**1.2e-06**	1.0e+00	1.8e-03	**8.8e-06**	
11	NM_001012709	*KRTAP5-4*	11	keratin associated protein 5-4	**2.7e-06**	5.0e-01	1.2e-01	**7.5e-07**	✓
12	NM_000201	*ICAM1*	19	intercellular adhesion molecule 1 precursor	**2.7e-06**	1.2e-03	4.5e-01	**1.2e-03**	
13	NM_001024667	-	1	-	**4.5e-06**	1.0e+00	3.8e-02	**9.3e-07**	
14	NM_138363	*CCDC45*	17	coiled-coil domain containing 45	**4.9e-06**	3.8e-01	1.0e+00	7.2e-03	
15	NM_001131	*CRISP1*	6	acidic epididymal glycoprotein-like 1 isoform 1	**1.6e-05**	5.0e-01	1.0e+00	**5.7e-07**	
16	NM_022074	*FAM111A*	11	hypothetical protein LOC63901	**2.8e-05**	2.9e-01	2.6e-01	1.9e-03	
17	NM_002287	*LAIR1*	19	leukocyte-associated immunoglobulin-like	**3.1e-05**	1.0e+00	1.0e+00	**3.5e-06**	✓
18	NM_153368	*CX40.1*	10	connexin40.1	**4.9e-05**	1.0e+00	1.0e+00	5.0e-01	
19	NM_153269	*C20orf96*	20	hypothetical protein LOC140680	**5.5e-05**	2.1e-03	3.9e-01	6.4e-02	
20	NM_004616	*TSPAN8*	12	transmembrane 4 superfamily member 3	**5.6e-05**	1.0e+00	1.0e+00	**3.8e-05**	
21	NM_018643	*TREM1*	6	triggering receptor expressed on myeloid cells	**6.3e-05**	1.0e+00	5.5e-02	**6.7e-04**	
22	NM_144682	*SLFN13*	17	schlafen family member 13	**6.6e-05**	5.0e-01	5.0e-01	**6.3e-05**	✓
23	NM_198947	*FAM111B*	11	hypothetical protein LOC374393	**1.3e-04**	1.0e+00	1.0e+00	**3.9e-06**	
24	AK123368	*AK123368*	4	Hypothetical protein FLJ41374.	**1.3e-04**	1.0e+00	1.0e+00	**2.0e-05**	
25	NM_000300	*PLA2G2A*	1	phospholipase A2, group IIA	**1.3e-04**	4.1e-01	1.0e+00	1.9e-03	
26	NM_207645	*LOC399947*	11	hypothetical protein LOC399947	**1.3e-04**	**8.1e-06**	1.0e+00	4.8e-01	
27	NM_000733	*CD3E*	11	CD3E antigen, epsilon polypeptide	**1.5e-04**	4.6e-01	4.6e-01	**1.9e-06**	
28	NM_001424	*EMP2*	16	epithelial membrane protein 2	**1.5e-04**	1.0e+00	5.0e-01	**1.8e-04**	
29	NM_001423	*EMP1*	12	epithelial membrane protein 1	**1.5e-04**	1.0e+00	1.0e+00	**1.9e-04**	
30	NM_001014975	*CFH*	1	complement factor H isoform b precursor	**1.5e-04**	1.0e+00	2.5e-01	1.3e-02	
31	NM_030766	*BCL2L14*	12	BCL2-like 14 isoform 2	**1.5e-04**	4.7e-01	4.6e-01	1.5e-02	
32	BC020840	*TCRA*	14	T-cell receptor alpha chain C region	**1.5e-04**	3.5e-02	5.0e-01	6.6e-02	
33	OTTHUMT00000004245	*RP11-558F24.1-001*	1	novel protein	**1.5e-04**	3.2e-02	1.1e-01	9.7e-02	
34	NM_002170	*IFNA8*	9	interferon, alpha 8	**1.5e-04**	7.3e-03	3.9e-03	1.2e-01	
35	NM_175900	*C16orf54*	16	hypothetical protein LOC283897	**1.8e-04**	5.0e-01	4.0e-01	**3.6e-05**	
36	NM_006464	*TGOLN2*	2	trans-golgi network protein 2	**1.8e-04**	4.9e-01	5.0e-01	**5.3e-05**	
37	NM_014317	*PDSS1*	10	prenyl diphosphate synthase, subunit 1	**1.8e-04**	3.6e-01	1.0e+00	1.7e-03	
38	NM_000518	*HBB*	11	beta globin	**2.0e-04**	5.0e-01	1.0e+00	**4.8e-06**	
39	NM_001647	*APOD*	3	apolipoprotein D precursor	**2.0e-04**	1.0e+00	1.0e+00	**6.4e-04**	
40	NM_004211	*SLC6A5*	11	solute carrier family 6	**2.3e-04**	3.1e-01	4.2e-01	5.0e-01	
41	NM_001024855	*ZNF197*	3	zinc finger protein 197 isoform 2	**3.4e-04**	1.8e-01	3.3e-01	1.0e+00	
42	BC017592	*ZNRF4*	19	R31343_1.	**3.9e-04**	1.8e-01	3.1e-02	7.0e-03	
43	NM_017446	*MRPL39*	21	mitochondrial ribosomal protein L39 isoform a	**4.0e-04**	6.1e-03	1.0e+00	**1.3e-04**	
44	OTTHUMT00000041603	*RP11-98I9.1-002*	6	-	**4.0e-04**	1.3e-02	1.0e+00	1.6e-01	
45	AB051446	*AB051446*	22	KIAA1659 protein (Fragment).	**4.3e-04**	2.4e-01	2.3e-01	2.4e-02	
46	NM_198076	*FAM36A*	1	family with sequence similarity 36, member A	**4.3e-04**	2.7e-01	1.7e-01	9.4e-02	
47	NM_002343	*LTF*	3	lactotransferrin	**4.9e-04**	1.0e+00	4.3e-01	**1.9e-05**	
48	NM_194300	*CCDC129*	7	coiled-coil domain containing 129	**4.9e-04**	4.8e-01	4.8e-01	**3.0e-05**	
49	NM_000097	*CPOX*	3	coproporphyrinogen oxidase	**4.9e-04**	4.0e-01	4.0e-01	2.2e-02	
50	NM_173352	*KRT78*	12	keratin 5b	**5.0e-04**	1.0e+00	4.9e-01	**1.9e-06**	
51	NM_001708	*OPN1SW*	7	opsin 1 (cone pigments), short-wave-sensitive	**5.8e-04**	1.0e+00	4.9e-01	**2.3e-05**	
52	NM_014508	*APOBEC3C*	22	apolipoprotein B mRNA editing enzyme, catalytic	**5.9e-04**	1.0e+00	5.0e-01	**7.7e-04**	
53	OTTHUMT00000058121	*RP1-321E8.3-001*	X	novel protein similar to LOC347458	**6.0e-04**	1.0e+00	2.1e-01	2.2e-02	✓
54	NM_018224	*C7orf44*	7	hypothetical protein LOC55744	**6.1e-04**	3.8e-01	4.9e-01	2.7e-03	
55	NM_002078	*GOLGA4*	3	golgi autoantigen, golgin subfamily a, 4	**6.6e-04**	5.3e-03	1.0e+00	1.0e+00	
56	AB051518	*AB051518*	11	Hypothetical protein FLJ37899.	**6.8e-04**	3.7e-01	4.3e-01	1.5e-02	
57	NM_031950	*KSP37*	4	Ksp37 protein	**6.8e-04**	1.0e+00	1.0e+00	1.0e+00	
58	NM_001622	*AHSG*	3	alpha-2-HS-glycoprotein	**6.9e-04**	4.8e-01	1.0e+00	**2.7e-04**	
59	NM_148959	*HUS1B*	6	HUS1 checkpoint protein B	**6.9e-04**	3.6e-03	8.6e-02	5.0e-01	
60	NM_004552	*NDUFS5*	1	NADH dehydrogenase (ubiquinone) Fe-S protein 5	**7.1e-04**	1.0e+00	1.0e+00	**1.1e-04**	
61	NM_014220	*TM4SF1*	3	transmembrane 4 superfamily member 1	**7.3e-04**	1.5e-01	2.1e-01	1.0e+00	
62	AY172952	*ADAM32*	8	Metalloproteinase 12-like protein.	**7.6e-04**	1.0e+00	1.0e+00	**1.4e-05**	
63	NM_016286	*DCXR*	17	dicarbonyl/L-xylulose reductase	**8.1e-04**	2.9e-03	1.0e+00	2.6e-02	
64	NM_001245	*SIGLEC6*	19	sialic acid binding Ig-like lectin 6 isoform 1	**8.3e-04**	5.0e-01	1.0e+00	3.4e-03	
65	NM_005218	*DEFB1*	8	defensin, beta 1 preproprotein	**8.3e-04**	5.0e-01	5.0e-01	2.6e-02	
66	OTTHUMT00000055500	*C9orf11-002*	9	chromosome 9 open reading frame 11	**8.3e-04**	3.0e-01	4.7e-01	2.8e-01	
67	NM_015492	*C15orf39*	15	hypothetical protein LOC56905	**8.3e-04**	3.9e-01	3.9e-01	1.0e+00	
68	OTTHUMT00000096701	*RP11-523K4.1-001*	1	novel protein	8.8e-04	5.0e-01	1.0e+00	**3.3e-04**	
69	NM_032338	*C12orf31*	12	hypothetical protein LOC84298	8.9e-04	1.0e+00	1.0e+00	**1.3e-04**	
70	NM_004363	*CEACAM5*	19	carcinoembryonic antigen-related cell adhesion	1.0e-03	5.0e-01	1.0e+00	**2.5e-06**	✓
71	NM_001002254	*DGAT2L4*	X	diacylglycerol O-acyltransferase 2-like 4	1.1e-03	1.0e+00	1.0e+00	**2.8e-04**	
72	NM_032149	*C4orf17*	4	hypothetical protein LOC84103	1.1e-03	3.4e-01	1.0e+00	**3.4e-04**	
73	NM_174901	*FAM9C*	X	family with sequence similarity 9, member C	1.1e-03	4.8e-01	3.5e-01	**5.3e-04**	✓
74	NM_001803	*CD52*	1	CD52 antigen	1.1e-03	1.0e+00	1.0e+00	**2.1e-04**	
75	NM_022782	*MPHOSPH9*	12	M-phase phosphoprotein 9	1.2e-03	5.0e-01	1.0e+00	**8.8e-04**	
76	NM_022366	*TFB2M*	1	transcription factor B2, mitochondrial	1.3e-03	4.9e-01	1.0e+00	**3.1e-04**	
77	NM_001295	*CCR1*	3	chemokine (C-C motif) receptor 1	1.5e-03	1.0e+00	1.0e+00	**1.6e-05**	
78	NM_002652	*PIP*	7	prolactin-induced protein	1.5e-03	1.0e+00	1.0e+00	**6.9e-05**	
79	NM_015324	*KIAA0409*	11	hypothetical protein LOC23378	1.6e-03	1.0e+00	1.0e+00	**5.8e-05**	
80	NM_000638	*VTN*	17	vitronectin precursor	2.1e-03	1.0e+00	1.0e+00	**1.2e-04**	
81	NM_002761	*PRM1*	16	protamine 1	2.3e-03	8.7e-03	**2.5e-04**	1.0e+00	
82	NM_005697	*SCAMP2*	15	secretory carrier membrane protein 2	2.3e-03	1.0e+00	1.0e+00	**2.0e-04**	
83	NM_030763	*NSBP1*	X	nucleosomal binding protein 1	2.4e-03	4.4e-01	1.0e+00	**1.7e-04**	
84	NM_004131	*GZMB*	14	granzyme B precursor	2.9e-03	1.0e+00	1.0e+00	**1.2e-03**	
85	NM_145246	*C10orf4*	10	FRA10AC1 protein isoform FRA10AC1-1	2.9e-03	2.6e-01	1.0e+00	**4.4e-04**	
86	NM_018473	*THEM2*	6	thioesterase superfamily member 2	3.1e-03	5.0e-01	1.0e+00	**4.0e-04**	
87	NM_199289	*NEK5*	13	NIMA (never in mitosis gene a)-related kinase 5	3.1e-03	1.0e+00	1.0e+00	**1.9e-04**	
88	NM_003016	*SFRS2*	17	splicing factor, arginine/serine-rich 2	3.2e-03	1.0e+00	5.0e-01	**6.6e-04**	
89	NM_001025778	*VRK3*	19	vaccinia related kinase 3 isoform 2	3.2e-03	1.0e+00	1.0e+00	**3.1e-04**	
90	NM_002483	*CEACAM6*	19	carcinoembryonic antigen-related cell adhesion	3.3e-03	5.0e-01	5.0e-01	**3.6e-04**	
91	NM_002389	*CD46*	1	CD46 antigen, complement regulatory protein	3.6e-03	5.0e-01	1.0e+00	**8.7e-05**	✓
92	NM_014860	*SUPT7L*	2	SPTF-associated factor 65 gamma	3.7e-03	5.0e-01	5.0e-01	**4.7e-04**	

Table S6.3: Complete List of Genes Identified by Likelihood Ratio Tests

No.	Accession	Gene Name	Chr	Description	P_A	P_H	P_C	P_M	Dup
93	OTTHUMT00000147963	*LMAN1-001*	18	-	3.7e-03	1.0e+00	1.0e+00	**2.9e-04**	
94	NM_000927	*ABCB1*	7	ATP-binding cassette sub-family B member 1	3.8e-03	5.0e-01	5.0e-01	**8.4e-04**	
95	NM_004891	*MRPL33*	2	mitochondrial ribosomal protein L33 isoform a	3.8e-03	1.0e+00	1.0e+00	**2.6e-04**	
96	BC032347	*BC032347*	8	C8orf59 protein.	3.8e-03	1.0e+00	1.0e+00	**9.5e-04**	
97	NM_138371	*FAM113B*	12	hypothetical protein LOC91523	3.9e-03	5.0e-01	1.0e+00	**3.6e-04**	
98	NM_000396	*CTSK*	1	cathepsin K preproprotein	4.1e-03	1.0e+00	3.4e-01	**5.1e-04**	
99	NM_012128	*CLCA4*	1	calcium activated chloride channel 4	4.3e-03	4.8e-01	1.0e+00	**1.4e-04**	
100	NM_001285	*CLCA1*	1	chloride channel, calcium activated, family	4.4e-03	1.0e+00	1.0e+00	**2.2e-05**	
101	AK056484	*AK056484*	7	Hypothetical protein FLJ31922.	4.5e-03	1.0e+00	1.0e+00	**1.2e-04**	
102	NM_000783	*CYP26A1*	10	cytochrome P450, family 26, subfamily A	4.6e-03	1.0e+00	3.2e-01	**2.9e-04**	
103	NM_174941	*CD163L1*	12	scavenger receptor cysteine-rich type 1 protein	4.7e-03	4.9e-01	1.0e+00	**6.4e-04**	
104	NM_033049	*MUC13*	3	mucin 13, epithelial transmembrane	4.9e-03	1.0e+00	1.0e+00	**5.1e-06**	
105	NM_000584	*IL8*	4	interleukin 8 precursor	5.2e-03	1.0e+00	1.0e+00	**1.0e-03**	
106	OTTHUMT00000154594	*BUCS1-002*	16	butyryl Coenzyme A synthetase 1	5.4e-03	1.0e+00	1.0e+00	**1.1e-03**	
107	AK058196	*CCDC13*	3	Hypothetical protein FLJ25467.	5.6e-03	1.0e+00	4.9e-01	**4.0e-04**	
108	AK130385	*AK130385*	20	Hypothetical protein FLJ26875.	6.2e-03	1.0e+00	1.0e+00	**1.4e-05**	
109	NM_175625	*RAB3IP*	12	RAB3A interacting protein isoform beta 2	6.2e-03	5.0e-01	1.0e+00	**1.3e-03**	
110	NM_145276	*ZNF563*	19	zinc finger protein 563	6.5e-03	5.0e-01	1.0e+00	**3.3e-04**	
111	AF279900	*MCM7*	7	PNAS-146.	6.5e-03	1.0e+00	1.0e+00	**7.5e-04**	
112	NM_205838	*LST1*	6	leukocyte specific transcript 1 isoform 3	6.7e-03	4.9e-01	4.9e-01	**9.6e-05**	
113	NM_022377	*ICAM4*	19	intercellular adhesion molecule 4 isoform 2	7.1e-03	1.0e+00	5.0e-01	**2.5e-04**	
114	NM_030588	*DHX9*	1	-	7.6e-03	1.0e+00	1.0e+00	**1.1e-03**	
115	NM_018374	*TMEM106B*	7	hypothetical protein LOC54664	8.0e-03	1.0e+00	1.0e+00	**3.9e-04**	
116	NM_153226	*TMEM20*	10	transmembrane protein 20	8.0e-03	1.0e+00	1.0e+00	**1.3e-05**	
117	NM_024021	*MS4A4A*	11	membrane-spanning 4-domains, subfamily A, member	8.0e-03	1.0e+00	1.0e+00	**4.2e-04**	
118	NM_024576	*OGFRL1*	6	opioid growth factor receptor-like 1	8.2e-03	3.1e-01	**7.3e-04**	5.0e-01	
119	NM_031264	*MUCDHL*	11	mucin and cadherin-like isoform 3	9.0e-03	1.0e+00	1.0e+00	**1.5e-05**	
120	NM_214711	*LOC401137*	4	hypothetical protein LOC401137	9.4e-03	1.0e+00	1.0e+00	**3.6e-04**	
121	AK126014	*AK126014*	4	Hypothetical protein FLJ44026.	9.7e-03	5.0e-01	5.0e-01	**4.9e-05**	
122	AY358798	*AY358798*	13	FRSS1829.	9.9e-03	1.0e+00	1.0e+00	**7.4e-04**	
123	NM_013263	*BRD7*	16	bromodomain containing 7	1.0e-02	5.0e-01	1.0e+00	**1.3e-03**	
124	NM_005603	*ATP8B1*	18	ATPase, Class I, type 8B, member 1	1.0e-02	5.0e-01	4.9e-01	**1.0e-03**	
125	NM_021114	*SPINK2*	4	serine protease inhibitor, Kazal type 2	1.0e-02	1.0e+00	1.0e+00	**1.0e-04**	
126	NM_000574	*CD55*	1	decay accelerating factor for complement	1.1e-02	1.0e+00	1.0e+00	**1.3e-03**	
127	NM_000361	*THBD*	20	thrombomodulin precursor	1.1e-02	4.7e-01	5.0e-01	**1.2e-03**	
128	BC110910	*RBM21*	11	Hypothetical protein FLJ22267.	1.2e-02	1.0e+00	1.0e+00	**1.2e-03**	
129	NM_001008784	*CD200R2*	3	CD200 cell surface glycoprotein receptor isoform	1.2e-02	5.0e-01	4.9e-01	**9.4e-04**	
130	NM_032040	*CCDC8*	19	coiled-coil domain containing 8	1.2e-02	1.0e+00	4.8e-01	**3.7e-04**	
131	NM_178812	*MTDH*	8	LYRIC/3D3	1.3e-02	1.0e+00	1.0e+00	**4.8e-05**	
132	NM_021185	*C19orf15*	19	hypothetical protein LOC57828	1.3e-02	1.0e+00	4.9e-01	**7.6e-04**	
133	NM_058173	*SBEM*	12	small breast epithelial mucin precursor	1.4e-02	1.0e+00	1.0e+00	**1.3e-04**	
134	BX641066	*KLF8*	X	Kruppel-like factor 8	1.5e-02	1.0e+00	1.0e+00	**1.3e-03**	
135	BC110814	*SFI1*	22	SFI1 protein.	1.5e-02	1.0e+00	1.0e+00	**1.3e-03**	
136	NM_001011548	*MAGEA4*	X	melanoma antigen family A, 4	1.6e-02	1.0e+00	4.8e-01	**3.7e-04**	
137	NM_002711	*PPP1R3A*	7	protein phosphatase 1 glycogen-binding	1.7e-02	4.9e-01	4.9e-01	**4.7e-04**	
138	NM_133274	*FCAR*	19	Fc alpha receptor isoform f	1.8e-02	1.0e+00	1.0e+00	**4.2e-04**	
139	AF076494	*IRF7*	11	Putative collagen homolog protein-a.	2.0e-02	1.0e+00	**5.2e-04**	4.6e-01	
140	NM_018322	*C6orf64*	6	hypothetical protein LOC55776	2.0e-02	4.0e-01	4.3e-01	**1.2e-04**	
141	NM_078476	*BTN2A1*	6	butyrophilin, subfamily 2, member A1 isoform 2	2.0e-02	1.0e+00	1.0e+00	**3.2e-05**	✓
142	NM_002029	*FPR1*	19	formyl peptide receptor 1	2.1e-02	1.0e+00	1.0e+00	**4.1e-04**	
143	AK097725	*PLCZ1*	12	Hypothetical protein FLJ40406.	2.1e-02	1.0e+00	1.0e+00	**5.2e-04**	
144	NM_172241	*CTAGE1*	18	cutaneous T-cell lymphoma-associated antigen 1	2.1e-02	1.0e+00	1.0e+00	**1.4e-04**	✓
145	NM_024077	*SECISBP2*	9	SECIS binding protein 2	2.2e-02	1.0e+00	1.0e+00	**9.1e-04**	
146	OTTHUMT00000150637	*CEACAM1-009*	19	carcinoembryonic antigen-related cell adhesion molecule 1 (biliary glycoprotein)	2.4e-02	5.0e-01	1.0e+00	**1.1e-03**	
147	BC003094	*TBCD*	17	TBCD protein.	2.5e-02	1.0e+00	1.0e+00	**6.4e-04**	
148	NM_174939	*MGC39681*	11	-	2.6e-02	1.0e+00	1.0e+00	**2.8e-04**	
149	NM_000873	*ICAM2*	17	intercellular adhesion molecule 2 precursor	2.6e-02	5.0e-01	1.0e+00	**2.9e-04**	
150	NM_004547	*NDUFB4*	3	NADH dehydrogenase (ubiquinone) 1 beta	2.7e-02	1.0e+00	1.0e+00	**2.8e-05**	
151	NM_019606	*BCDIN3*	7	bin3, bicoid-interacting 3	2.8e-02	1.0e+00	**4.7e-04**	1.0e+00	
152	NM_031271	*TEX15*	8	testis expressed sequence 15	2.9e-02	4.8e-01	5.0e-01	**7.6e-04**	
153	NM_032317	*WBSCR18*	7	Williams Beuren syndrome chromosome region 18	2.9e-02	1.0e+00	1.0e+00	**6.5e-04**	
154	NM_152358	*C19orf41*	19	hypothetical protein LOC126123	3.0e-02	1.0e+00	1.0e+00	**7.9e-04**	
155	NM_003771	*KRT36*	17	keratin 36	3.0e-02	3.9e-01	**2.4e-04**	1.0e+00	
156	NM_002994	*CXCL5*	4	chemokine (C-X-C motif) ligand 5 precursor	3.1e-02	1.0e+00	3.9e-01	**3.5e-04**	✓
157	OTTHUMT00000059995	*AC073647.2-001*	7	-	3.2e-02	1.0e+00	1.0e+00	**1.0e-03**	
158	NM_152453	*TMCO5*	15	transmembrane and coiled-coil domains 5	3.3e-02	6.5e-03	**4.4e-04**	1.0e+00	
159	NM_003064	*SLPI*	20	secretory leukocyte peptidase inhibitor	3.8e-02	1.0e+00	1.0e+00	**1.1e-03**	
160	NM_003708	*RDH16*	12	retinol dehydrogenase 16	4.5e-02	1.0e+00	1.0e+00	**1.5e-04**	✓
161	NM_003104	*SORD*	15	sorbitol dehydrogenase	4.5e-02	1.0e+00	1.0e+00	**9.0e-04**	✓
162	NM_152243	*CDC42EP1*	22	CDC42 effector protein 1 isoform a	4.7e-02	1.0e+00	5.0e-01	**1.1e-03**	
163	NM_015245	*ANKS1A*	6	ankyrin repeat and sterile alpha motif domain	5.2e-02	1.0e+00	4.6e-01	**1.0e-03**	
164	NM_007335	*DLEC1*	3	deleted in lung and esophageal cancer 1 isoform	5.6e-02	1.0e+00	**7.9e-05**	5.0e-01	
165	NM_138360	*C14orf121*	14	hypothetical protein LOC90668	6.0e-02	4.1e-01	**7.7e-05**	1.0e+00	
166	NM_003226	*TFF3*	21	trefoil factor 3 precursor	6.5e-02	1.0e+00	4.9e-01	**1.1e-03**	
167	NM_000073	*CD3G*	11	CD3G gamma precursor	7.4e-02	1.0e+00	1.0e+00	**1.1e-05**	
168	NM_006746	*SCML1*	X	sex comb on midleg-like 1 isoform b	1.1e-01	2.9e-01	**5.6e-09**	1.0e+00	
169	NM_002030	*FPRL2*	19	formyl peptide receptor-like 2	1.1e-01	1.0e+00	1.0e+00	**4.0e-04**	
170	NM_053036	*NPFFR2*	4	G protein-coupled receptor 74 isoform 2	1.1e-01	1.0e+00	**5.3e-04**	5.0e-01	
171	NM_016519	*AMBN*	4	ameloblastin precursor	1.2e-01	1.0e+00	1.0e+00	**1.0e-04**	
172	NM_024516	*C16orf53*	16	hypothetical protein LOC79447	1.3e-01	1.0e+00	1.0e+00	**2.5e-04**	
173	NM_001145	*ANG*	14	angiogenin, ribonuclease, RNase A family, 5	1.3e-01	1.0e+00	1.0e+00	**3.5e-04**	
174	NM_152912	*MTIF3*	13	mitochondrial translational initiation factor 3	1.4e-01	1.0e+00	1.0e+00	**6.1e-04**	
175	NM_012089	*ABCB10*	1	ATP-binding cassette, sub-family B, member 10	1.5e-01	1.0e+00	1.0e+00	**1.4e-04**	
176	NM_001004315	*FLJ46210*	3	hypothetical protein LOC389152	1.7e-01	1.0e+00	**5.2e-04**	1.0e+00	✓
177	OTTHUMT00000152323	*PYPAF6-003*	19	-	2.4e-01	1.0e+00	1.0e+00	**3.1e-06**	
178	OTTHUMT00000047532	*PARD3-007*	10	par-3 partitioning defective 3 homolog (C.elegans)	1.0e+00	1.0e+00	**4.7e-04**	5.0e-01	

Table S6.4: GO Categories Overrepresented Among Genes Predicted to be Under Positive Selection

Category	Description	N^a	${n_A}^b$	$E[n_A]^c$	${P_A}^d$	${n_M}^e$	$E[n_M]^f$	${P_M}^{dg}$
GO:0006955	immune response	335	8	2.2	**1.35e-03**	13	4.2	**3.10e-04**
GO:0051707	response to other organism	129	5	0.8	**1.43e-03**	–	–	–
GO:0006952	defense response	268	7	1.7	1.64e-03	10	3.4	2.07e-03
GO:0005506	iron ion binding	155	5	1.0	3.20e-03	6	2.0	1.37e-02
GO:0016491	oxidoreductase activity	386	8	2.5	3.26e-03	10	4.9	2.42e-02
GO:0009607	response to biotic stimulus	158	5	1.0	3.47e-03	–	–	–
GO:0005576	extracellular region	586	10	3.8	4.11e-03	14	7.4	1.58e-02
GO:0005615	extracellular space	244	6	1.6	4.79e-03	8	3.1	1.21e-02
GO:0006118	electron transport	172	5	1.1	4.98e-03	7	2.2	6.08e-03
GO:0051869	physiological response to stimulus	706	11	4.6	5.19e-03	20	8.9	**5.20e-04**
GO:0044459	plasma membrane part	837	11	5.4	1.75e-02	21	10.6	1.78e-03
GO:0005886	plasma membrane	990	12	6.4	2.34e-02	26	12.5	**2.30e-04**
GO:0005887	integral to plasma membrane	684	9	4.4	3.09e-02	20	8.6	**3.50e-04**
GO:0031226	intrinsic to plasma membrane	689	9	4.4	3.22e-02	20	8.7	**3.80e-04**
GO:0050874	organismal physiological process	1043	12	6.7	3.34e-02	27	13.2	**2.20e-04**
GO:0044421	extracellular region part	380	6	2.5	3.55e-02	10	4.8	2.20e-02
GO:0006091	generation of precursor metabolites and energy	297	5	1.9	4.25e-02	9	3.7	1.29e-02
GO:0004872	receptor activity	767	9	5.0	5.71e-02	19	9.7	3.44e-03
GO:0000267	cell fraction	440	6	2.8	6.39e-02	12	5.6	9.63e-03
GO:0004871	signal transducer activity	1008	10	6.5	1.12e-01	20	12.7	2.76e-02
GO:0016021	integral to membrane	2248	19	14.5	1.20e-01	49	28.4	**3.00e-05**
GO:0031224	intrinsic to membrane	2255	19	14.6	1.22e-01	49	28.5	**3.00e-05**
GO:0050896	response to stimulus	901	9	5.8	1.24e-01	19	11.4	1.80e-02
GO:0044425	membrane part	2392	19	15.4	1.85e-01	50	30.2	**7.00e-05**
GO:0016020	membrane	2880	21	18.6	2.96e-01	53	36.4	1.08e-03
GO:0007166	cell surface receptor linked signal transduction	682	5	4.4	4.52e-01	14	8.6	4.86e-02

[a]Number of genes from set of 10,376 that were classified in category or one of its descendant categories.

[b]Number of the 84 genes identified by test T_A (any branch) in category or descendant. Categories with $n_A < 5$ are excluded.

[c]Number of genes identified by T_A expected to belong to category if categories were randomly assigned to genes.

[d]Nominal one-sided P-value by Fisher's exact test. Bold indicates significance at 0.1 level after a conservative correction for FWER (Holm). All categories with nominal $P < 0.05$ for either T_A or T_M are shown.

[e]Number of the 134 genes identified by test T_M (macaque branch) in category or descendant. Categories with $n_M < 5$ are excluded.

[f]Number of genes identified by T_M expected to belong to category if categories were randomly assigned to genes.

[g]Too few genes were identified by test T_H (human branch) to allow this analysis to be performed. The results for T_C (chimpanzee branch) were generally similar to those for T_M but fewer categories showed significant enrichments.

Table S6.5: PANTHER Categories Overrepresented Among Genes Predicted to be Under Positive Selection

Category	Description	N^a	${n_A}^b$	$E[n_A]^c$	${P_A}^d$	${n_M}^e$	$E[n_M]^f$	${P_M}^{dg}$
MF00004	Immunoglobulin receptor family member	46	6	0.3	**4.41e-07**	5	0.6	**2.69e-04**
MF00173	Defense/immunity protein	145	8	0.9	**3.90e-06**	7	1.8	**2.38e-03**
BP00148	Immunity and defense	622	11	4.0	**1.96e-03**	19	7.9	**2.87e-04**
BP00274	Cell communication	595	10	3.8	**4.58e-03**	16	7.5	**3.31e-03**
MF00001	Receptor	651	10	4.2	8.51e-03	17	8.2	**3.33e-03**
BP00122	Ligand-mediated signaling	221	5	1.4	1.39e-02	5	2.8	1.48e-01
BP00179	Apoptosis	246	5	1.6	2.11e-02	5	3.1	2.00e-01
BP00124	Cell adhesion	271	5	1.7	3.04e-02	10	3.4	**2.25e-03**
BP00102	Signal transduction	1644	16	10.6	5.60e-02	30	20.8	2.12e-02

[a]Number of genes from set of 10,376 that were classified in category or one of its descendant categories.

[b]Number of the 84 genes identified by test T_A (any branch) in category or descendant. Categories with $n_A < 5$ are excluded.

[c]Number of genes identified by T_A expected to belong to category if categories were randomly assigned to genes.

[d]Nominal one-sided P-value by Fisher's exact test. Bold indicates significance at 0.1 level after a conservative correction for FWER (Holm). All categories with nominal $P < 0.05$ for either T_A or T_M are shown.

[e]Number of the 134 genes identified by test T_M (macaque branch) in category or descendant. Categories with $n_M < 5$ are excluded.

[f]Number of genes identified by T_M expected to belong to category if categories were randomly assigned to genes.

[g]Too few genes were identified by test T_H (human branch) to allow this analysis to be performed. The results for T_C (chimpanzee branch) were generally similar to those for T_M but fewer categories showed significant enrichments.

Table S6.6: GO Categories Showing an Excess of High Likelihood Ratios

Category	Description	N^a	$P_A{}^b$	$P_H{}^c$	$P_C{}^d$	$P_M{}^e$
GO:0051869	physiological response to stimulus	677	**7.98e-13**	2.94e-02	3.03e-03	**2.66e-09**
GO:0006955	immune response	321	**1.42e-08**	9.80e-02	8.61e-03	**9.32e-08**
GO:0002376	immune system process	378	**5.84e-08**	6.59e-02	1.85e-02	**8.40e-08**
GO:0002245	physiological response to wounding	185	**2.27e-07**	7.62e-01	1.72e-01	**4.03e-10**
GO:0007606	sensory perception of chemical stimulus	52	**3.18e-07**	7.53e-04	1.85e-02	4.68e-04
GO:0006952	defense response	256	**3.57e-07**	6.45e-01	3.07e-01	**8.05e-08**
GO:0050874	organismal physiological process	1002	**3.89e-07**	2.93e-02	1.92e-02	**5.99e-07**
GO:0009611	response to wounding	194	**8.21e-07**	7.49e-01	2.34e-01	**3.20e-09**
GO:0050909	sensory perception of taste	10	**9.18e-07**	5.95e-01	5.55e-01	**5.09e-06**
GO:0002217	physiological defense response	214	**3.79e-06**	5.66e-01	3.69e-01	**2.08e-07**
GO:0005576	extracellular region	567	**4.82e-06**	1.28e-01	8.13e-02	**9.38e-11**
GO:0048730	epidermis morphogenesis	13	**2.90e-05**	3.14e-03	6.56e-03	1.64e-01
GO:0009913	epidermal cell differentiation	13	**2.90e-05**	3.14e-03	6.56e-03	1.64e-01
GO:0007596	blood coagulation	40	8.97e-05	4.22e-01	8.78e-03	**1.29e-05**
GO:0009605	response to external stimulus	264	1.06e-04	6.98e-01	3.80e-01	**1.00e-06**
GO:0006954	inflammatory response	145	1.06e-04	7.85e-01	6.14e-01	**6.82e-07**
GO:0042060	wound healing	43	1.09e-04	5.64e-01	9.88e-03	**2.34e-05**
GO:0004984	olfactory receptor activity	40	2.28e-04	**1.20e-05**	6.89e-03	3.79e-02
GO:0005882	intermediate filament	56	1.69e-03	2.53e-02	8.18e-03	**3.96e-05**
GO:0045111	intermediate filament cytoskeleton	56	1.69e-03	2.53e-02	8.18e-03	**3.96e-05**
GO:0005615	extracellular space	241	1.88e-03	7.23e-02	9.53e-02	**6.35e-06**
GO:0045087	innate immune response	38	3.37e-03	6.73e-01	1.23e-01	**2.93e-05**
GO:0002541	activation of plasma proteins during acute inflammatory response	19	5.71e-03	7.32e-01	2.29e-01	**1.37e-05**
GO:0006956	complement activation	19	5.71e-03	7.32e-01	2.29e-01	**1.37e-05**
GO:0002526	acute inflammatory response	30	6.66e-03	6.77e-01	8.23e-01	**3.32e-05**
GO:0006118	electron transport	167	8.36e-03	4.72e-01	5.88e-01	**3.49e-05**
GO:0044421	extracellular region part	369	1.24e-02	3.40e-01	2.49e-01	**6.78e-07**

[a]Number of genes from set of 10,376 that were classified in each category or one of its descendant categories. Categories with $N < 10$ are excluded.

[b]Nominal one-sided P-value from MWU test of log likelihood ratios from test T_A (any branch). A small P-value indicates a significant shift toward larger T_A-based log likelihood ratios among genes within a category relative to genes not in the category. Bold indicates significance at 0.1 level after a conservative correction for FWER (Holm). Only categories significant in at least one test are shown.

[c]Same as (b) but for test T_H (human branch)

[d]Same as (b) but for test T_C (chimpanzee branch)

[e]Same as (b) but for test T_M (macaque branch)

Table S6.7: PANTHER Categories Showing an Excess of High Likelihood Ratios

Category	Description	N^a	$P_A{}^b$	$P_H{}^c$	$P_C{}^d$	$P_M{}^e$
MF00173	Defense/immunity protein	141	**1.99e-16**	2.08e-01	3.51e-01	**6.63e-19**
MF00004	Immunoglobulin receptor family member	44	**2.33e-11**	1.47e-02	2.04e-01	**2.91e-10**
BP00148	Immunity and defense	608	**2.59e-06**	6.20e-01	1.74e-01	**1.21e-10**
BP00157	Natural killer cell mediated immunity	28	**3.20e-06**	4.19e-01	1.31e-01	**2.64e-06**
BP00240	Fertilization	16	**2.64e-05**	4.06e-01	7.30e-01	**2.64e-04**
MF00224	KRAB box transcription factor	263	**3.83e-05**	1.06e-01	5.57e-01	1.61e-02
MF00256	Intermediate filament	41	**6.01e-05**	4.07e-02	2.08e-02	**4.06e-06**
MF00015	Other receptor	95	**1.70e-04**	5.37e-01	4.57e-01	1.68e-02
MF00198	Structural protein	88	**2.66e-04**	1.48e-02	1.99e-03	1.21e-03
BP00153	Complement-mediated immunity	27	4.80e-04	5.34e-01	2.77e-01	**1.48e-04**
MF00174	Complement component	23	9.65e-04	4.00e-01	1.91e-01	**6.40e-05**
BP00176	Blood clotting	35	3.18e-03	2.04e-01	2.88e-02	**5.33e-06**
BP00155	Macrophage-mediated immunity	61	1.21e-02	9.90e-02	2.81e-01	**4.88e-05**

[a] Number of genes from set of 10,376 that were classified in each category or one of its descendant categories. Categories with $N < 10$ are excluded.

[b] Nominal one-sided P-value from MWU test of log likelihood ratios from test T_A (any branch). A small P-value indicates a significant shift toward larger T_A-based log likelihood ratios among genes within a category relative to genes not in the category. Bold indicates significance at 0.1 level after a conservative correction for FWER (Holm). Only categories significant in at least one test are shown.

[c] Same as (b) but for test T_H (human branch)

[d] Same as (b) but for test T_C (chimpanzee branch)

[e] Same as (b) but for test T_M (macaque branch)

Figures 6.1-6.4 (following pages):

Figure S6.1 Shift in ! = d_N/d_S in genes belonging to the GO categories "immune response" and "transcription\factor activity."

Figure S6.2 An estimate for ω for each branch of a five-¬species phylogeny,

Figure S6.3 Power of test T_A as a function of ! = d_N/d_S for simulated human/chimpanzee/macaque and human/macaque/mouse alignments of 500 codons. Note the logarithmic scale on the x-axis.

Figure S6.4 Power of test T_M as a function of ! = d_N/d_S for simulated human/chimpanzee/macaque and human/macaque/mouse alignments of 500 codons. Note the logarithmic scale on the x-axis.

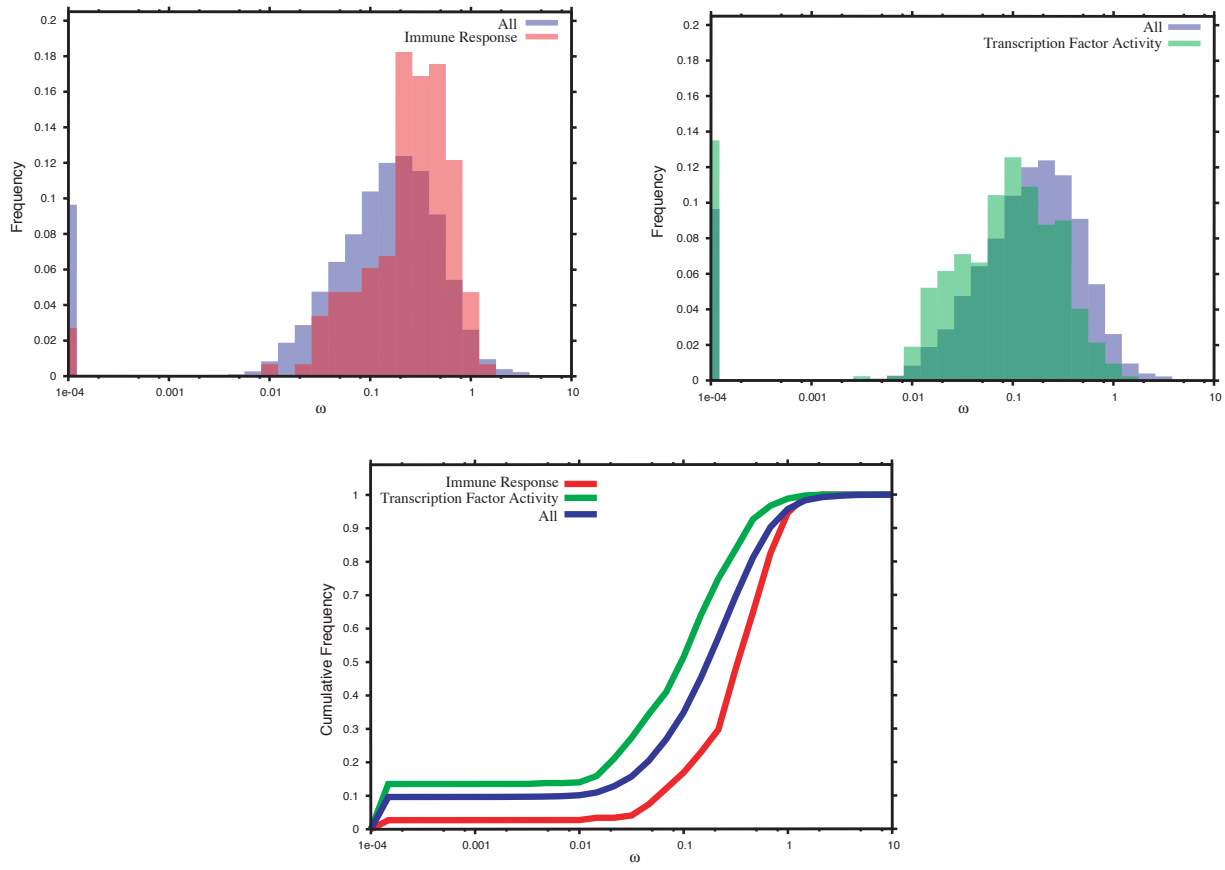


Figure S6.1: Shift in $\omega = d_N/d_S$ in genes belonging to the GO categories "immune response" and "transcription factor activity."

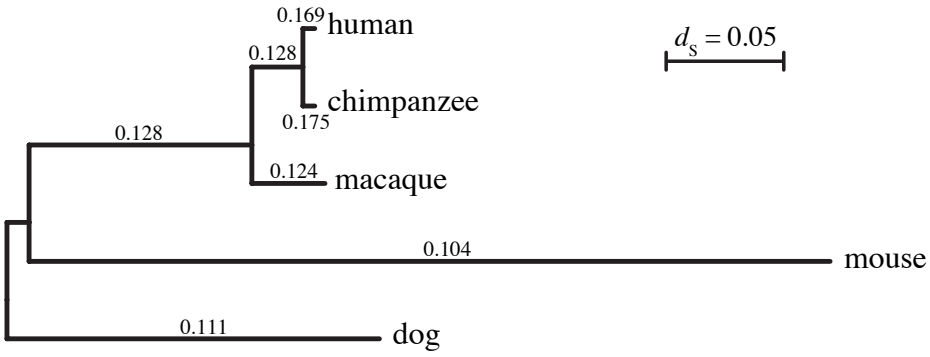


Figure S6.2: An estimate of ω for each branch of a five-species phylogeny. Show is the maximum-likelihood phylogeny for 5286 orthologous quintents, with branch lengths drawn in proportion to the estimated number of synonymous substitutions per synonymous site (d_S). Each branch is labeled with the corresponding estimate of ω.

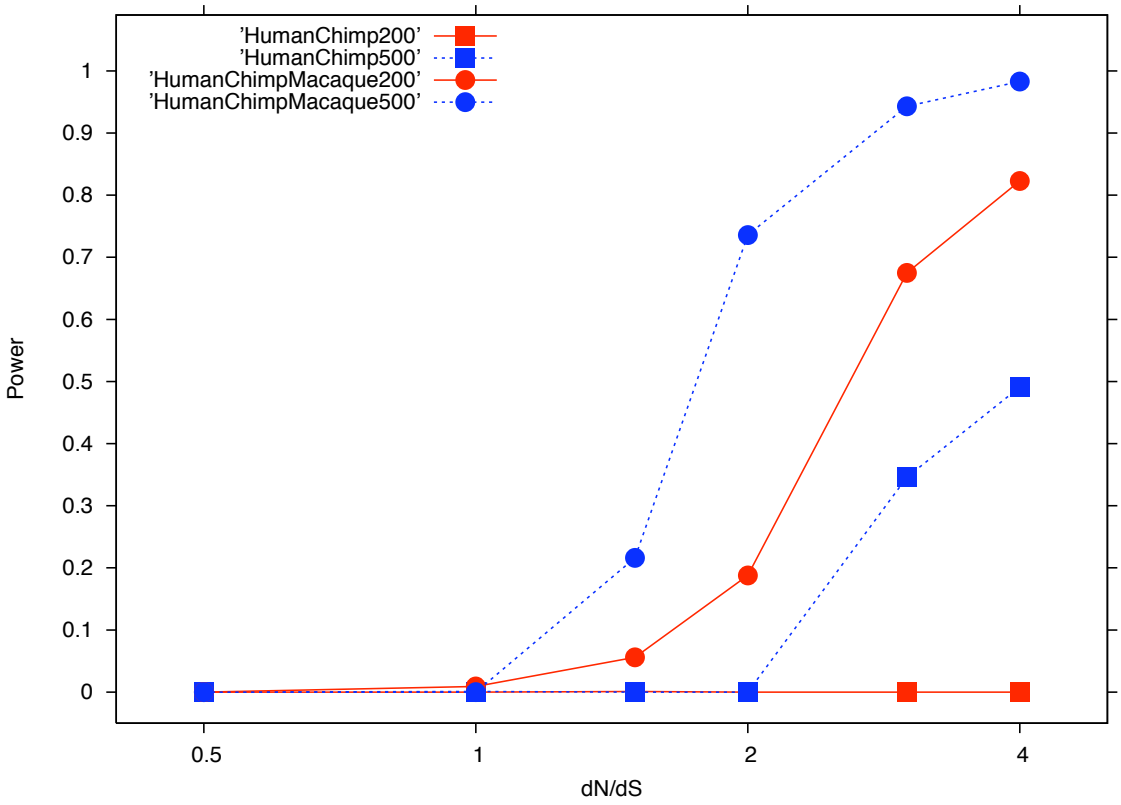


Figure S6.3: Power of test T_A as a function of $\omega = d_N/d_S$ for simulated human/chimpanzee and human/chimpanze/macaque alignments of 200 and 500 codons. Note the logarithmic scale on the x-axis.

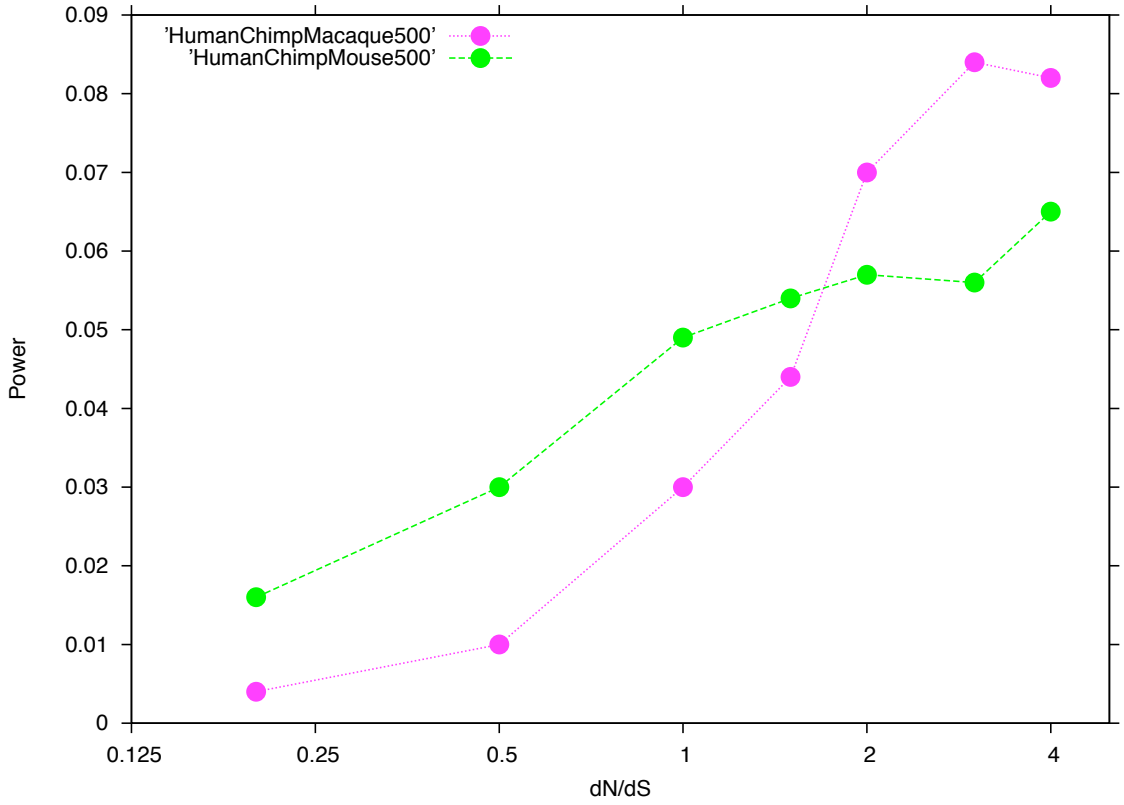


Figure S6.4: Power of test T_M as a function of $\omega = d_{\mathrm{N}}/d_{\mathrm{S}}$ for simulated human/chimpanzee/macaque and human/macaque/mouse alignments of 500 codons. Note the logarithmic scale on the x-axis.

7. Genetic Variation within Macaques

To study genetic variation within macaques we first obtained animal samples from collaborators. The long term aim is to establish a resource of a standard DNA sample set for widespread use in genotyping studies. **Table S7.1** shows the IDs for samples that were accumulated and used for the studies reported here. (Please note that 8 Chinese macaques and 8 Indian macaques were used for the wgs comparison. Initially 24 Chinese macaques and 24 Indian macaques were to be shared for the subsequent re-sequencing studies and retrotransposon genotyping. A subset was initially mis-labeled (clerical error at the point of origin) and this resulted in an un-even number of animals from each population being further analyzed. Moreover, DNA from one addition animal at one site where PCR was performed was depleted and DNA from one animal of the other population at another site was refractory to PCR. Hence there are subtle differences in the listed numbers of animals used, in different parts of the manuscript e.g 37 vs. 38 Indian macaques; 9 vs. 10 Chinese macaques).

WGS Libraries

8 Indian Macaque IDs
17719
17722
17753
17757
18402
18403
18409
18415

8 Chinese Macaque IDs
21328
21368
21616
21693
22125
22894
23511
23524

Table S7.1: ID's of animals used in wgs SNP discovery.

SNPs from wgs reads

Whole genome shotgun sequencing of 8 Chinese and 8 Indian origin rhesus macaques was performed using standard methods. The reads were compared to rhemac2 by BLAST and the results parsed to select single base discrepancies in regions where the neighboring bases had a contiguous high quality threshold (Q20 >20) spanning more than 30 contiguous bases on either side of the base difference. All sequence data are submitted to the NCBI trace archive (see **Table S2.6**, above for Genbank accessions).

Genetic Studies using retrotransposon insertion polymorphism

Using the marker set and the strategy described in the main text the allelic frequencies of 177 insertion loci were obtained for samples from the Indian and the Chinese macaque samples.

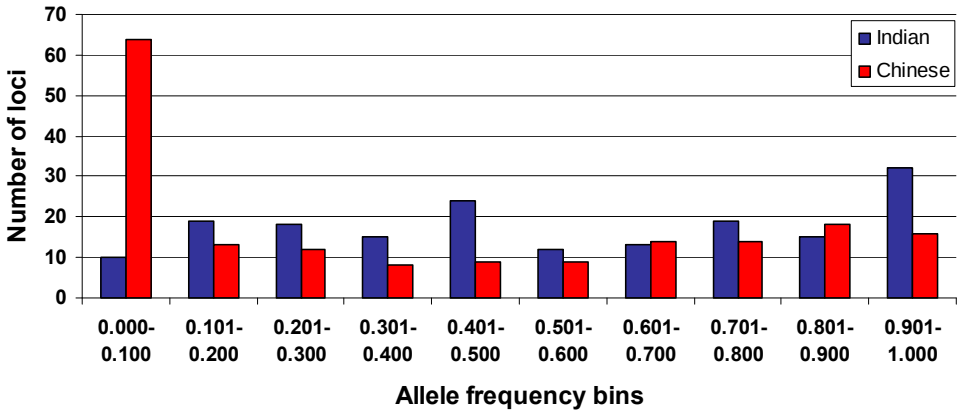


Figure S7.1: Allele frequency distribution of 177 polymorphic retrotransposon insertions in the two rhesus populations.

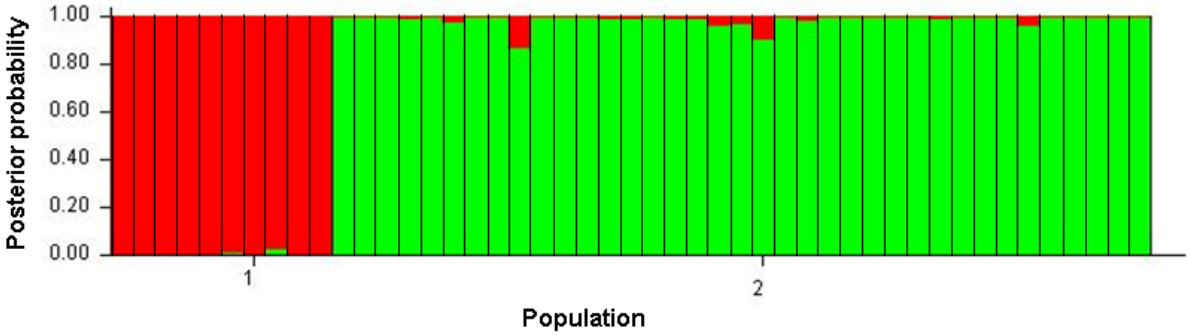


Figure S7.2: Population structure of 10 Chinese and 37 Indian rhesus macaque individuals. Estimates of individual ancestry proportion for 177 polymorphic *Alu*/L1 genetic markers when two ancestral populations are assumed. Each individual's genome is represented by a vertical bar, where each color (red/green) represents the proportion of an individual's ancestry derived from each of the two populations.

Structure 2.0 analysis

To conduct a population genetic study with rhesus macaques from two different geographic origins, we genotyped all samples by using a PCR analysis. Three different genotypes were possible: homozygote insertion present, heterozygote insertion present/absent, and homozygote insertion absent. Allele frequency distribution of 177 recently integrated retrotransposon polymorphisms was calculated (**Figure S7.1**) and put in a format compatible with Structure 2.0 (35,35). This software performs model-based clustering on genotypic data to infer a population structure. For each of the 177 rhesus macaques, Structure 2.0 estimated the proportion of heritage from each of K population clusters (**Figure S7.2**). As the population structure of the rhesus macaques on the population panel has not previously been predicted, we ran several simulations with different Ks (number of predicted populations). Also, a variety of data settings were used to investigate the best model. Finally, we used the recommended settings with a burn-in of 10,000 iterations and a run of 10,000 replications. Each run was replicated several times (>3) using a desktop computer.

Population structure:

Our structure analysis resulted in a highest likelihood value with K (populations) =2. All individuals were assigned to their population origin with a posterior probability of at least 86% (all but two were assigned with probability >96%, **Figure S7.2**). This implies that the Indian and Chinese populations are genetically separate groups with almost no indication of ongoing admixture. Indian rhesus macaques possess higher and more evenly distributed filled allele (insertion present) frequencies than Chinese rhesus macaques (**Figure S7.1**), consistent with a relatively recent bottleneck in the Indian rhesus macaque population (36).

Population-specific SNP content and calculations of linkage disequilibrium

The studies of population structure, size estimates and linkage disequilbrium as described in the main text and (37).

Male mutation bias methods

The substitution rates were estimated from the human-macaque pairwise alignments at interspersed repeats inserted prior to human-macaque divergence with the REV substitution model as implemented in PAML (34). Sites with phred quality scores below 20 in macaque were removed prior to calculations. Additionally, we removed pseudoautosomal regions and stratum 5 from chromosome X. Autosomal substitution rate was calculated as the average among all autosomes weighted by their lengths. Alpha was calculated from the X-to-autosomal substitution rate ratio using the formula derived by Miyata et al. (38). The 95% confidence interval for alpha was estimated using the bootstrap method. Namely, we divided the alignments into 100-kb windows and removed chromosomal labels from them. Next, we randomly selected the alignments 1,000 times with replacement, each time allocating a certain number of windows to chromosome X (and this number is equal to the original number of X chromosomal windows) and the rest to autosomes, and estimated alpha from each of such pseudosamples. See **Table S7.2** below for details of the calculated substitution rate.

Chromosome	Substitution rate
1	0.0709
2	0.0706
3	0.0699
4	0.0726
5	0.0705
6	0.0708
7	0.0736
8	0.0721
9	0.0726
10	0.0719
11	0.0713
12	0.0714
13	0.0730
14	0.0714
15	0.0728
16	0.0793
17	0.0787
18	0.0721
19	0.0981
20	0.0715
21	0.0768
22	0.0824
Autosomes[1]	0.0727
X	0.0610

Table S7.2: Chromosome-specific human-macaque substitution rates.
[1]Autosomal rate was calculated as average rate among autosomes weighted by their lengths**.**

8. Human Disease Orthologs in Macaque

Human Gene Mutation Database (39) (http://www.hgmd.org; October 2006 release). A total of 236 substitutions were identified where the amino acid considered to be mutant in human corresponded to the wild-type amino acid present in macaque, chimpanzee and/or an ancestor (40). These are listed in **Table S8.1**.

Table S8.1: List of 229 gene candidates for ancestral human gene mutations: (see the associated rhesus file).

As some of the mutations reflected alterations in genes involved in human intermediatary metabolism, we elected to directly measure macaque blood to determine steady state metabolic levels. Eight animals were sampled and plasma from each was analyzed at the Baylor College of Medicine Biochemistry Diagnostics Clinic using an automated amino acid analyzer. The study revealed lower concentrations of cyst(e)ine than in human and slightly higher concentrations of glycine than in human, but no increase in phenylalanine or ammonia which might have been predicted from the observed changes (see **Tables S8.2 and S8.3)**.

Table S8.2: Plasma amino levels in eight macaques: . (see the associated rhesus file).

Table S8.3: Determination of acylcarnitine (nM) in rhesus: . (see the associated rhesus file).

9. Leveraging the Impact of a Genomic Sequence on Biological Studies

The rhesus microarray was develop by the Katze group (University of Washington) in collaboration with Agilent Technologies, Inc (41). The array uses 60-mer oligonucleotide probes to measure mRNA transcript levels, cf. **Table S9.1**: (List of oligonucleotide probes for rhesus mRNA transcripts, on Agilent rhesus macaque microarray chip (FASTA format)); comparison to predicted macaque genes is shown in **Table S9.2**: (Content of Agilent rhesus macaque microarray and the sequence homology to predicted macaque genes):

Table S9.1: List of oligonucleotide probes for rhesus mRNA transcripts on Agilent rhesus macaque microarray chip (FASTA format): (see the associated rhesus file).

Table S9.2 Content of Agilent rhesus macaque microarray and the sequence homology to predicted macaque genes: (see the associated rhesus file).

Finally**,** a summary of the attributes of the available microarray, as determined during the course of this study, is given in **Table S9.3 (**Attributes of an oligonucleotide microarray derived from the rhesus macaque genome sequence.)

Table S9.3: Attributes of an oligonucleotide microarray derived from the rhesus macaque genome sequence.

Probe Origins	# of Probes	# of Unique Genes or Loci	%-Identity
Alignment of human RefSeq mRNA entries and Mmul_0.1	18382		
> MegaBlast hits to RefSeq Predicted		13575	99.9
> MegaBlast hits to Genomic Contigs[a]		4427	99.8
> BLAT hits to chromosome assemblies[b]		182	97.3[c]
Derived from Rhesus ESTs	1036		
> MegaBlast hits to RefSeq Predicted		153[d]	99.9
> MegaBlast hits to Genomic Contigs[a]		329[d]	99.8
> BLAT hits to chromosome assemblies[b]		24	97.9[c]
Totals	**19419**[e]	**18690**	

Notes for table S9.3:

[a]MegaBlast searches using 60-mer probe sequences as queries were performed against either the collection of predicted mRNA transcripts produced in NCBI genome assembly Build 1.1, which

is based on sequence release Mmul_051212, or the RefSeq collection of chromosomal contigs used in the reference assembly. [b]BLAT searches were conducted with the UCSC, Jan. 2006 genome assembly (rheMac2, also derived from Mmul_051212). [c]Note that %-Identity scoring is not strictly comparable for BLAT alignments vs. MegaBlast alignments. [d]These results are those not redundant with prior hits in the table; 89% of the EST probes were matched by MegaBlast. [e]A total 19120 probes (98.5%) of the probe sequences were matched in this manner.

Influenza Model:
Macaques infected with the human influenza strain A/Texas/36/91 ([42]) were compared for expression changes in lung tissues, to those seen in whole blood during the course of infection. Gene lists for the functional categories (GO Bioprocess Level 6) of Interferon Induction (**Figure 10**), Inflammatory Response (**Figure 10**), and Apoptosis are given in **Table S9.4** (Hybridization of Agilent rhesus macaque microarray to mRNA samples from either infected lung tissue or whole blood, in a macaque model of influenza). This table includes the numerical values used in the construction of **Figure 10**. For each ontological category, genes of interest were chosen as those where the ratio of transcript vs. control was 2-fold or greater ($p < 0.01$) in at least two of the twelve microarray experiments. The ordering of the genes in the heat maps was based on the clustering for Day 2

Table S9.4: Hybridization of Agilent rhesus macaque microarray to mRNA samples from either infected lung tissue or whole blood, in a macaque model of influenza: (see the associated rhesus file).

List of Supplementary Figures

LIST OF TABLES:

Table S7.1 ID's of animals used in wgs SNP discovery.

Table S7.2 **C**hromosome-specific human-macaque substitution rates.

Table S8.1 List of 229 gene candidates for ancestral human gene mutations.

Table S8.2 Plasma amino levels in eight macaques.

Table S8.3 Determination of acylcarnitine (nM) in rhesus.

Table S9.1 List of oligonucleotide probes for rhesus mRNA transcripts on Agilent rhesus macaque microarray chip.

Table S9.2 Content of Agilent rhesus macaque microarray and the sequence homology to predicted macaque genes.

Table S9.3 Attributes of an oligonucleotide microarray derived from the rhesus macaque genome sequence.

Table S9.4 Hybridization of Agilent rhesus macaque microarray to mRNA samples from either infected lung tissue or whole blood, in a macaque model of influenza.

Reference List

1. S. Kumar and S. B. Hedges, *Nature* 392, 917-920 (1998).
2. J. E. Fa, *Mammal Rev* 19, 45-81 (1989).
3. A. Fortna et al., *PLoS.Biol.* 2, E207 (2004).
4. C. Ross, *Primates* 33, 207-215 (1992).
5. S. Hartwig-Scherer and R. D. Martin, *Am.J.Phys.Anthropol.* 88, 37-57 (1992).
6. D. J. a. M. C. P. Melnick, in ***Primate Societies***, B.B.Smuts, Ed. (**Univ. of Chicago Press**, 1986).
7. E. Sodergren et al., *Science* 314, 941-952 (2006).
8. X. Huang et al., *Nucleic Acids Res.* 34, 201-205 (2006).
9. E. W. Myers et al., *Science* 287, 2196-2204 (2000).
10. A. Milosavljevic et al., *Genome Res.* 15, 292-301 (2005).
11. J. A. Bailey, A. M. Yavor, H. F. Massa, B. J. Trask, E. E. Eichler, *Genome Res.* 11, 1005-1017 (2001).
12. X. She et al., *Nature* 431, 927-930 (2004).
13. J. A. Bailey et al., *Science* 297, 1003-1007 (2002).
14. Z. Birtle, L. Goodstadt, C. Ponting, *BMC.Genomics* 6, 120 (2005).
15. S. Schwartz et al., *Genome Res.* 13, 103-107 (2003).
16. K. D. Pruitt, T. Tatusova, D. R. Maglott, *Nucleic Acids Res.* (2006).
17. F. Hsu et al., *Bioinformatics.* 22, 1036-1046 (2006).
18. J. L. Ashurst et al., *Nucleic Acids Res.* 33, D459-D465 (2005).
19. M. Blanchette et al., *Genome Res.* 14, 708-715 (2004).
20. R. M. Kuhn et al., *Nucleic Acids Res.* (2006).
21. W. J. Kent, R. Baertsch, A. Hinrichs, W. Miller, D. Haussler, *Proc.Natl.Acad.Sci.U.S.A* 100, 11484-11489 (2003).
22. Z. Yang, *Genet.Res.* 69, 111-116 (1997).

23. R. Nielsen and Z. Yang, *Genetics* 148, 929-936 (1998).

24. A. Wong et al., *Genomics* 84, 239-247 (2004).

25. Z. Yang, W. S. Wong, R. Nielsen, *Mol.Biol.Evol.* 22, 1107-1118 (2005).

26. Y. Benjamini and Y. .Hochberg, *Journal of the Royal Statistical Society* 57, 289-300 (1995).

27. Z. Yang and R. Nielsen, *Mol.Biol.Evol.* 19, 908-917 (2002).

28. J. Zhang, R. Nielsen, Z. Yang, *Mol.Biol.Evol.* 22, 2472-2479 (2005).

29. R. Nielsen et al., *PLoS.Biol.* 3, e170 (2005).

30. A. G. Clark et al., *Science* 302, 1960-1963 (2003).

31. M. Ashburner et al., *Nat.Genet.* 25, 25-29 (2000).

32. H. Mi et al., *Nucleic Acids Res.* 33, D284-D288 (2005).

33. S. Holm, *Scandinavian Journal of Statistics* 6, 65-70 (1979).

34. Z. Yang, *Comput.Appl.Biosci.* 13, 555-556 (1997).

35. D. Falush, M. Stephens, J. K. Pritchard, *Genetics* 164, 1567-1587 (2003).

36. D. G. Smith and J. McDonough, *Am.J.Primatol.* 65, 1-25 (2005).

37. S. H. Williamson et al., *Proc.Natl.Acad.Sci.U.S.A* 102, 7882-7887 (2005).

38. T. Miyata, H. Hayashida, K. Kuma, K. Mitsuyasu, T. Yasunaga, *Cold Spring Harb.Symp.Quant.Biol.* 52, 863-867 (1987).

39. P. D. Stenson et al., *Hum.Mutat.* 21, 577-581 (2003).

40. M. Blanchette, E. D. Green, W. Miller, D. Haussler, *Genome Res.* 14, 2412-2423 (2004).

41. J. C. Wallace et al., *BMC.Genomics* 8, 28 (2007).

42. T. Baas et al., *J.Virol.* 80, 10813-10828 (2006).

43: **Acknowledgements**:
W.M. acknowledges support from NHGRI grant HG002238; R. C. H. from NIDDK grant DK65806; A.M. from NIHNHGRI grants R01 02583-01 and R01 004009-1, and NIH-NCRR grant U01 RR 18464. M. H. was supported by NSF grant: DBI-0543586; M. A. B. by NSF grants BCS-0218338, NIH GM59290, EPS-0346411 and the State of Louisiana Board of Regents Support Fund; K. P. was supported by a Max Planck Society and Marie Curie Fellowship. HGMD acknowledges its appreciation of financial support from BIOBASE GmbH,

Wolfenbuettel, Germany. We thank Jack Harding of the NCRR for his ongoing active support of development of resources for primate genomics. We wish to thank the following institutions that generously contributed biological samples used in this study: California National Primate Research Center, Oregon NPRC, Southwest NPRC and Yerkes NPRC. The individual rhesus macaque sample used for whole genome shotgun sequencing was contributed by the Southwest NPRC, and the BAC library was generated from a sample from the California NPRC. The other organizations contributed samples used in the analyses of population variability.