RESEARCH ARTICLE

Evolutionary and Biomedical Insights from the Rhesus Macaque Genome

Rhesus Macaque Genome Sequencing and Analysis Consortium*†

The rhesus macaque (*Macaca mulatta*) is an abundant primate species that diverged from the ancestors of *Homo sapiens* about 25 million years ago. Because they are genetically and physiologically similar to humans, rhesus monkeys are the most widely used nonhuman primate in basic and applied biomedical research. We determined the genome sequence of an Indian-origin *Macaca mulatta* female and compared the data with chimpanzees and humans to reveal the structure of ancestral primate genomes and to identify evidence for positive selection and lineage-specific expansions and contractions of gene families. A comparison of sequences from individual animals was used to investigate their underlying genetic diversity. The complete description of the macaque genome blueprint enhances the utility of this animal model for biomedical research and improves our understanding of the basic biology of the species.

Rhesus macaques (*Macaca mulatta*) (*1*) are one of the most frequently encountered and thoroughly studied of all nonhuman primates (table S1.1). They have a broad geographic distribution that reaches from Afghanistan and India across Asia to the Chinese shore of the Pacific Ocean. As an Old World monkey (superfamily Cercopithecoidea, family Cercopithecidae), this species is closely related to humans and shares a last common ancestor from about 25 million years ago (Mya) (*2*). The two species often live in close association, and macaques exhibit complex and intensely social behavioral repertoires.

The relationship between humans and macaques is even more important because biomedical research has come to depend on these primates as animal models. Compared with rodents, which are separated from humans by more than 70 million years (*2*, *3*), macaques exhibit greater similarity to human physiology, neurobiology, and susceptibility to infectious and metabolic diseases. Critical progress in biomedicine attributed to macaques includes the identification of the "rhesus factor" blood groups and advances in neuroanatomy and neurophysiology. Most important, their response to infectious agents related to human pathogens, including simian immunodeficiency virus and influenza, has made macaques the preferred model for vaccine development. Lesser-known contributions of these animals include their early use in the U.S. space program—a rhesus monkey was launched into space more than a dozen years before any chimpanzee.

The cynomolgus macaque (*M. fascicularis)*, pigtailed macaque (*M. nemestrina*), and Japanese macaque (*M. fuscata*) have all contributed to research, but the rhesus macaque has been used most widely. Taxonomists recognize six *M. mulatta* subspecies (*1*), which differ substantially in their geographical range, body size, and a variety of morphological, physiological, and behavioral characteristics. North American research colonies include animals representing both Indian and Chinese subspecies, although India ended the exportation of these animals in the 1970s.

With the advent of whole-genome sequencing, a highly accurate human genome sequence and a draft of the chimpanzee genome have been generated and compared. The chimpanzee shared a common ancestor with humans approximately 6 Mya (*4*, *5*), and the major impact of the chimpanzee genome sequence data has been in their direct comparison with data from the human genome. However, the chimpanzee data have major limitations. First, because the alignable sequence is only 1 to 2% different from that of the human, there is no informative "signal" to distinguish conserved elements from the overall high background level of conservation. This is exacerbated by the fact that the chimpanzee genome was an incomplete draft, containing sequence errors that could potentially mask true divergence. Second, the differences that are found between humans and chimpanzees are difficult to assign as specific to either the chimpanzee or the human. As a result, the chimpanzee analyses have on their own provided relatively few answers to the fundamental question of the nature of the specific molecular changes that make us human.

By contrast, the genome of the rhesus macaque has diverged farther from our own, with an average human-macaque sequence identity of ~93%. Figure 1 shows the inferred common ancestor for all three species, as well as a common ancestor that predated the human-chimpanzee divergence. A characteristic that is found in humans but not in the chimpanzee can be recognized as a loss in the chimpanzee if it is present in the macaque, or it can be recognized as a gain in the human if it is absent in macaque. In principle, this three-way comparison should make it possible to pinpoint many changes and identify specific underlying mutational mechanisms, which could have been critically important during the past 25 million years in shaping the biology of the three primate species.

We examined the basic elements of the rhesus macaque genome and undertook reconstruction of the major changes in the human-chimpanzee–rhesus macaque (HCR) trio. The regions of the genome that were duplicated in macaque were then identified and correlated with other genome features. Individual macaque genes were studied, and the orthologous genes in the HCR trio were aligned to reveal evidence for the action of selection on individual loci. Additional animals from other populations were also sampled by DNA sequencing to study their genetic diversity. Throughout, complementary methods were applied and the different results combined in order to represent the most complete picture of macaque biology. For a visual representation of some of the insights gained from the genome and more information about the importance of the macaque as a model organism, see the poster in this issue (*6*).

Sequencing the Genome

To generate a draft genome sequence for the rhesus macaque, whole-genome shotgun sequences were assembled. The bulk of the sequencing used DNA from a single *M. mulatta* female, whereas DNA from an unrelated male was used to construct a bacterial artificial chromosome (BAC) library to provide BAC end sequences and to aid in selective finishing. We

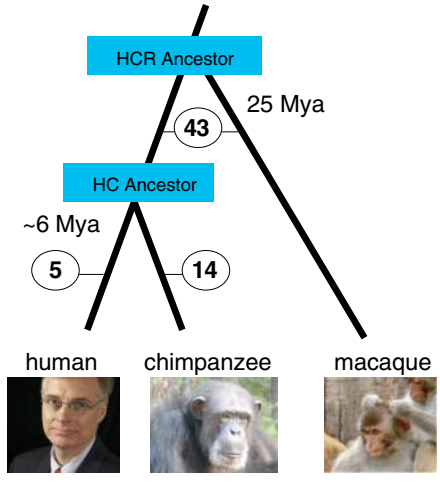

Fig. 1. Evolutionary triangulation in the human, chimpanzee and rhesus macaque lineages (lineage-specific breaks), showing a summary of chromosomal breakpoints on a microscopic scale (Fig. 3) (*7*). Circled numbers indicate numbers of lineage-specific breaks.

*To whom correspondence should be addressed. Richard A. Gibbs, E-mail: agibbs@bcm.edu
†All authors with their contributions and affiliations appear at the end of this paper.

used several whole-genome shotgun libraries with different insert sizes (~3.0, 10, 35, and 180 kb) to generate a total of 18.4 Gb of raw DNA sequence through standard fluorescent Sanger sequencing technologies. Initial assemblies to the intermediate scaffold stage were carried out by the three different assembly methods: Atlas–whole-genome shotgun, parallel contig assembly program (PCAP), and the Celera Assembler (*7*). These were compared by means of more than 200 metrics, including gross sequence statistics, agreement with finished sequence, utility for gene predictions in the Ensembl pipeline, and accuracy of alignment to the human genome. The three unpolished assemblies were found to be largely similar and of high quality, so all were used in combination with other genome data for the subsequent assembly and placement of long sequence segments on the macaque chromosomes (tables S2.1 to S2.4).

To produce an optimal representation of the genome, the three intermediate assemblies were merged (Fig. 2). Melding the assemblies involved mapping the Atlas–whole-genome shotgun and PCAP data to the Celera Assembler output, which had longer contiguity than the other two data sets at this stage of the process. There was little difference between assemblies at the sequence contig level, at which robust sequence alignments guide the reconstructions, so we focused our attention instead on contigs that were joined into scaffolds. Additional pairs of Celera Assembler scaffolds were joined based on their mapping to the other two macaque assemblies. Analysis of the output showed that this composite assembly was superior to any of its components (table S2.4).

During assembly, a comparison with the human genome sequence [National Center for Biotechnology Information (NCBI) accession code bld35] identified a small number (<100) of obvious inconsistencies, such as improper joins of different chromosomes. These scaffolds were therefore split at the misassembly point. The human map was also used to help place large merged scaffolds onto the macaque chromosomes (*8*, *9*) [the chromosome numbering of Rogers *et al.* (*8*) was used] at the highest level of the assembly process. Given that the human data were only used to split scaffolds and that de novo macaque assemblies were always given precedence over the mapping to the human genome in the macaque assembly merging and chromosome assignment process, the final product should not be regarded as a "humanized assembly."

The total length of the combined genome assembly was approximately 2.87 Gb (Table 1). This incorporated ~14.9 Gb of raw sequence, which represents about a 5.2-fold coverage of the macaque genome. Comparison with expressed sequence tag (EST) sequence data and approximately 1.8 Mb of finished sequence (see "Selected sequence finishing," below) indicated that ~98% of the available genome was represented. No misassemblies were identified in that comparison. Contigs showed an N50 (minimum length of contigs representing half of the total length of the assembly) of >25 kb; the N50 for sequence scaffolds was >24 Mb. GenBank accession codes are available online (table S2.5).

Selected sequence finishing. The rhesus macaque genome assembly is a draft DNA sequence, and it contains many gaps. A higher data quality with greater contiguity was desired at several genomic regions that attracted additional interest. In these cases, individual BAC clones were isolated, and data quality was improved by sequence "finishing." Many of these BACs were in regions of pronounced genome duplication, whereas others were gene-rich. All finished BACs, their gene content, and their genome coordinates are listed in table S2.6.

Overview of Genome Features

General organization and content. The macaque genome is organized into 20 autosomes and the XY sex chromosomes. With the exception of 48 breakpoints (Fig. 1)—including three fusions, one fission, and breakpoints induced by inversions that are each detectable through chromosome staining, by radiation hybrid mapping, or by comparative linkage mapping—there is a superficial similarity between the macaque and human chromosomes (*8*–*11*). Several chromosomes in the macaque are also more acrocentric than their human counterparts, but many from the two species are difficult to distinguish.

Nucleotide sequences that aligned between the human and rhesus average 93.54% identity. If, however, small insertions and deletions are included in the calculation, identity is reduced to 90.76%. Considering regions that are difficult to align, such as lineage-specific interspersed repeat elements, would further decrease the level of computed identity. Moreover, evolutionary distances exhibit local fluctuations, as in other mammals (*3*), and less divergence was observed in chromosome X (94.26% identity of aligned bases). The GC-content of the rhesus in aligned bases was not notably lower than that of the human (40.71% versus 40.74%).

Gene content. A human-centric approach was used to generate new macaque gene sets (table S3.1 and fig. S3.1). These sets include (i) Ensembl (*12*) gene models based primarily on the alignment of the human Uniprot and RefSeq resources with the current assembly to define the overall gene model, followed by the introduction of the macaque-specific sequences (mainly as lineage-specific paralogs) in that framework; (ii) Gnomen (NCBI) models that include the consideration of the available (~50,000) macaque ESTs along with the human RefSeq; and (iii) Nscan data that include multiple-species alignments along with cDNA alignments (*13*). Overall, ~20,000 loci were predicted by our methods in which at least one exon was found by two additional predictors. An additional ~5000 loci were each predicted by a single method, but manual inspection of a subset of these loci shows that they are enriched in gene-prediction errors, mainly due to misclassification of evidence (e.g., cDNAs from untranslated regions that were classified as containing protein coding). On average, high-confidence orthologs have 97.5% identity between the human and macaque at both the nu-

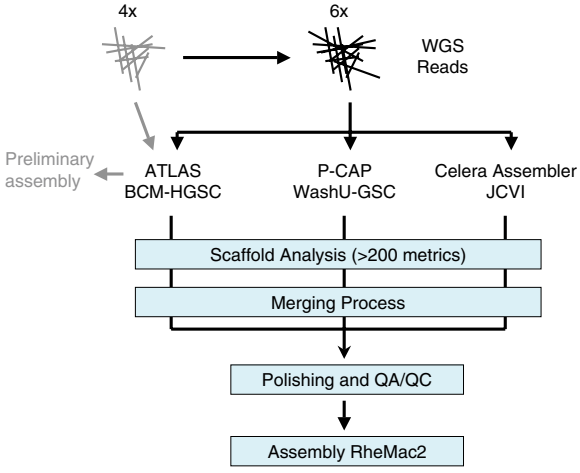

Fig. 2. Assembly by three methods of the rhesus macaque genome. WGS, whole-genome shotgun. BCM-HGSC, Baylor College of Medicine Human Genome Sequencing Center; WashU-GSC, Washington University Genome Sequencing Center; JCVI, J. Craig Venter Institute. QA/QC, quality assurance and quality control.

Table 1. *M. mulatta* assembly statistics. Total bases, excluding gaps, number 2,871,189,834.

	Contigs	Scaffolds
Total number	301,039	122,580
N50 size in bp	25,707	24,345,431
Number to N50	32,114	36
Largest in bp	219,335	98,200,701

cleotide and amino acid sequence levels. (The nucleotide and amino acid percentages agree because roughly one-third of nucleotide differences within coding regions change an amino acid.)

Overall repetitive landscape. Repeat elements account for ~50% of the genomes of all sequenced primates (*14*) (Table 2). Similar to the human, the rhesus macaque contains about 320,000 recognizable copies from more than 100 different families of DNA transposons and more than half a million recognizable copies of endogenous retroviruses (ERVs). In general, the DNA transposons show no new lineages, but the ERVs demonstrate a complex phylogeny and many examples of new and expanded family members, some resulting from horizontal transmission. In addition, we conservatively estimate that ~20,000 L1s [a family of long interspersed elements (LINEs)], and ~110,000 *Alu* elements [a primate-specific family of short interspersed elements (SINEs)], were specifically acquired in the Old World monkey lineage. These two retrotransposon families accounted for most lineage-specific insertions and have played a major role in shaping genomic architecture. Among them, rhesus macaque–specific subsets (derived from the L1PA5 lineage and *Alu*Y) are frequently polymorphic and can be assayed by polymerase chain reaction (PCR) genotyping analyses for genetic studies (*15*).

Determining Ancestral Genome Structure

Cytogenetically visible rearrangements. The most notable genomic differences among the HCR trio are the presence of cytogenetically visible rearrangements. The human and chimpanzee karyotypes are distinguishable by one chromosome fusion and nine cytogenetically visible pericentric inversions (*16*); with the use of the macaque as an outgroup, all of these breakpoints (except those induced by two inversions) have now been characterized at the DNA sequence level (*17*). Analysis of genomic sequence confirms that 14 breakpoints, corresponding to seven inversions, occurred in the chimpanzee lineage, as indicated in Fig. 1. (Five of the inversions are summarized in table S4.1.) The pericentric inversions of human chromosomes 1 and 18 and the fusion creating human chromosome 2 are specific to the human. Comparison of the reconstructed human-chimpanzee ancestral genome and the rhesus genome reveals 43 breakpoints on the microscopic scale (Figs. 1 and 3).

Submicroscopic rearrangements. Previous analyses [reviewed in (*14*)] have indicated that primate genomes harbor more structural differences than visible by cytogenetic staining. Analysis of these events is complicated by two issues: the draft state of the genomes and the presence of extensive segmental duplications. We analyzed these structural rearrangements by using the distance between orthologous blocks in each species to infer the ancestral genome structure and determine where rearrangements occurred on the phylogenetic tree. We excluded events smaller than 10 kilobase pairs (kbp), which are mostly due to retroposon insertions, and focused on cytogenetically undetectable breakpoints induced by insertions, deletions, inversions, and complex rearrangements of sizes between 10 kbp and 4 Mbp. Data were combined from inversion detection and ancestral reconstructions by the contiguous ancestral regions method (*18*) and gap detection by the genomic triangulation method (*19*), which further integrates data from genomic sequence comparisons (*20*) and comparative maps (*8*, *9*, *21*). The analysis revealed more than 1000 rearrangement-induced breakpoints through the HCR lineages, of which 820 occur between rhesus and the reconstructed human-chimpanzee ancestor (Fig. 3 and fig. S4.1). Each chromosome therefore constitutes a complex mosaic, with multiple changes introduced to orthologous counterparts. When rhesus macaque is compared with the human-chimpanzee ancestor, the X chromosome exhibits three times more rearrangements per megabase than the autosomes. This is both statistically significant and consistent with a slightly more than threefold difference observed in the human lineage following the branching off of chimpanzee (*19*). Given that a slower rate of variability at the single-nucleotide level in the X chromosome compared with autosomes has been interpreted as support for speciation models, this difference is worthy of further investigation (*22*).

Table 2. Summary of repeat content of the rhesus macaque genome compared with the human and chimpanzee genomes. hg18, human genome version 18; panTro2, *Pan troglodytes* version 2; rheMac2, rhesus macaque version 2; LTR, long terminal repeat; MIR, mammalian interspersed repeat. SVA is a composite repetitive element named after its main components, SINE, variable number of tandem repeats, and *Alu*; includes SVA precursor elements.

Species	DNA	LTR/ERV	LINE		SINE		SVA
			L1	L2	*Alu*	MIR	
hg18	355,000	506,000	572,000	363,000	1,144,000	584,000	3400
panTro2	305,000	453,000	558,000	315,000	1,111,000	553,000	4400
rheMac2	327,000	432,000	531,000	298,000	1,094,000	539,000	150

Duplications in the Genome and Gene Family Expansions

Genomic Duplications. Segmental duplication of genomic regions and the genes they contain

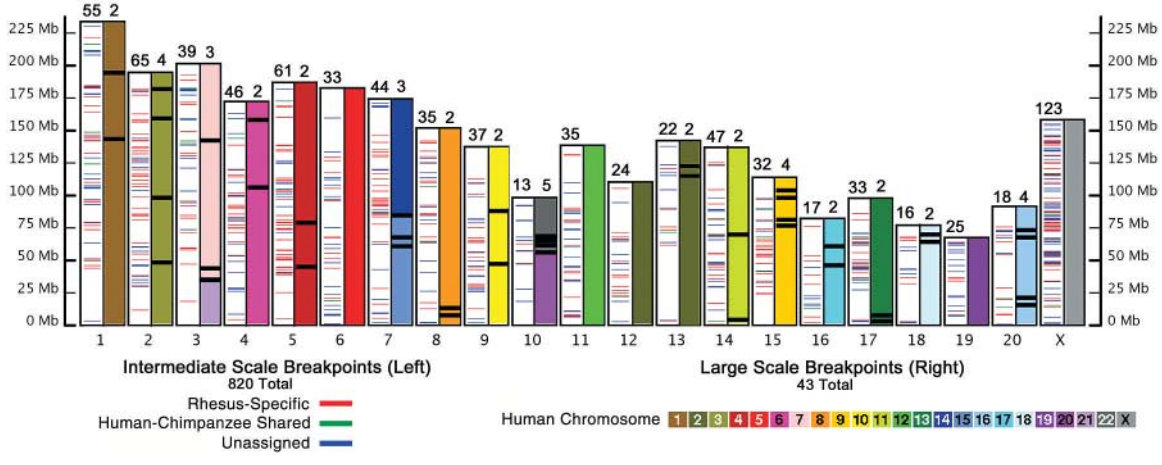

Fig. 3. Chromosomal breakpoints between rhesus macaque and the human-chimpanzee ancestor. Each chromosome is represented by a white bar (left) and a colored bar (right). A total of 820 thin horizontal lines in the white bars represent submicroscopic breakpoints (10-kbp to 4-Mbp range) detected by genomic triangulation (*19*), and 43 thick black lines in the colored bars represent breakpoints on a microscopic scale (>4 Mbp) (*7*). Numbers above each bar show the total lines within the bar.

are well known in mammals and are postulated to drive fundamental processes, including the birth of new genes and the subsequent expansion of gene families (*23*). To discover duplications in the macaque genome, we used a battery of different complementary approaches. Two of these, whole-genome assembly comparison (*24*) and BLASTZ (*25*) analysis of segmental duplications, depended directly on the assembly. We used a third method, whole-genome shotgun sequence detection (*26*), that calculated depth of coverage of the raw shotgun sequence reads relative to the assembly. A fourth procedure was created on the basis of BAC end sequence reads combined with BACs that were directly mapped by means of the pooled genomic indexing method (*21*). The common interspersed repeat families were not considered in any of these analyses.

The first two approaches identified approximately 35.0 Mb of a recently duplicated sequence in the macaque assembly. A further ~15 Mb were collapsed in the assembly and discovered by whole-genome shotgun sequence detection (fig. S5.1 and table S5.1). Adjusting for these collapsed duplications and the overall assembly coverage, we estimate that approximately 66.7 Mb or 2.3% of the macaque genome consists of segmental duplication (Fig. 4)—this proportion is substantially lower than that of either the human or chimpanzee genome (5 to 6%) (*26*, *27*).

The pooled genomic indexing and BAC end sequence read methods suggested slightly higher levels of overall duplication, on the basis of fluorescence in situ hybridization analysis of randomly selected large-insert BAC clones (*28*). However, this estimate was still less than the 4.8% recently estimated for the baboon genome (*28*). Overall, we consider 2.3% to be the lower bound of duplicated genomic DNA in the macaque genome.

As with the human and chimpanzee, the analysis of the macaque assembly revealed an enrichment of segmental duplications near gaps, centromeres, and telomeres (*14*, *29*). The study also identified segmental duplications that contain genes of high biological significance. For example, the *CCL3L1-CCL4* gene region [for which copy-number variation in humans is correlated with susceptibility to HIV infection (*30*)], cytochrome P450 (associated with toxicity response), *KRAB-C2H2* zinc finger (a developmental regulatory transcription factor), olfactory receptor (smell), human leukocyte antigen (HLA), and other immune and autoantigen gene families were all observed in regions of genome duplication.

Expansion of gene families. Two approaches were used to study gene family structure directly within the draft genome sequence: (i) a statistical approach, based on a likelihood model of gene gain and loss across the mammalian tree (*31*) and (ii) hybridization of whole genomic DNA to cDNA arrays [a variation of array-based comparative genomic hybridization (array CGH)] to observe changes in gene content directly (*32*). The results are shown in Tables 3 and 4.

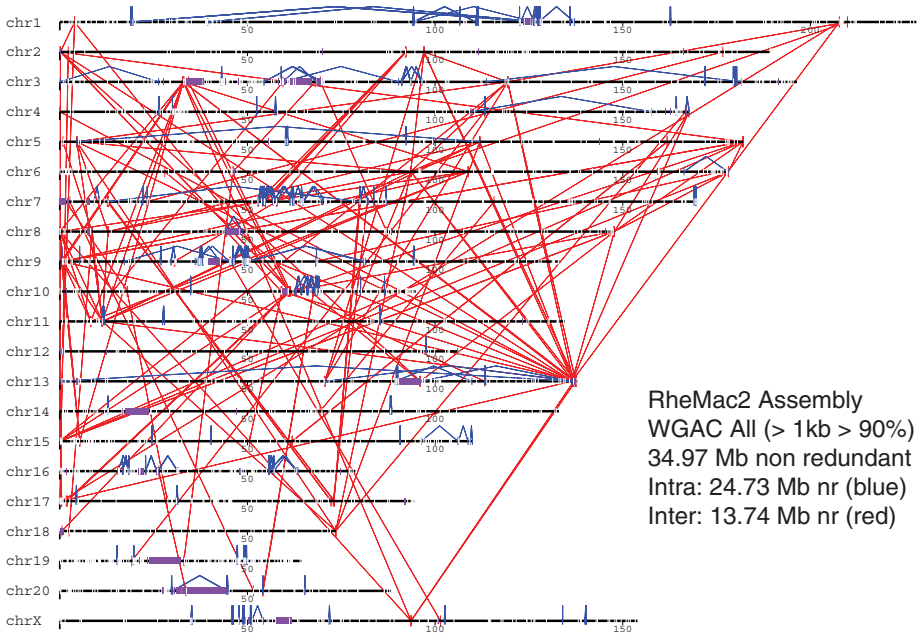

Fig. 4. Global pattern of macaque segmental duplications. The statistics are based on all WGAC duplications (> 90%, >1 kb in length), whereas the figure displays only those between 90 and 95% sequence identity and >10 kb in length for simplicity. Red lines indicate interchromosomal (Inter) duplications, blue ticks show intrachromosomal (Intra) events, and purple bars show centromeric, acrocentric, and/or large-gap regions. WGAC, whole-genome assembly comparison. nr, nonredundant.

The statistical approach revealed that 1358 genes were gained by duplication along the macaque lineage. This method simultaneously estimates rates of change along individual lineages and generates a quantitative assessment of confidence in rate differences among lineages. Iterative modeling revealed higher rates in primates, relative to other mammals. The rates are similar to those obtained by independent methods in both humans (*33*) and rodents (*3*).

We identified 108 gene families, computationally predicted to have changed in size among the primates, evolving at a significantly higher rate than the overall primate rates of gene gain and loss (all $P < 0.0001$, Table 3). More than 60% of the macaque-specific expansions display evidence of positive selection in their coding sequences, supporting the notion that this rate disparity may be driven by natural selection.

Gene copy-number estimates by genomic hybridization (cDNA array CGH) (*32*) identified 51 genes (124 cDNAs) with copy-number increases in the macaque, relative to the human (Table 4 and table S5.2). Of these array CGH-predicted macaque-specific increases, 33% (17 out of 51) were also found by computational analysis of gene family gains and losses. A separate analysis found that 55% (28 out of 51) are increased in copy number as estimated by BLAST-like Alignment Tool (BLAT)–based (*34*) predictions from the rheMac2 assembly. In contrast, when random sets of genes (cDNAs) were chosen for BLAT queries, only 1.45% suggest copy-number increases ($P < 0.0001$).

The genome-wide acceleration identified in primates may be due to an explosion in the number of *Alu* transposable elements in the primate ancestor, which may have allowed an increase in the rates of nonallelic homologous recombination, leading to higher rates of both duplication and deletion (*35*). Alternatively, the rates of duplicate gene fixation may be due to the small population size in primates (*36*) relative to rodents.

Particular expanded gene families. Expansion of individual gene families may help to identify processes that distinguish biological features among organisms. One example in humans is the preferentially expressed antigen of melanoma (*PRAME*) gene family that consists of a single gene on chromosome 22q11.22 and a cluster of several dozen genes on chromosome 1p36.21. *PRAME* and PRAME-like genes are actively expressed in cancers but normally manifest testis-specific expression and may thus have a role in spermatogenesis. The genomic organization is complicated; the cluster on human chromosome 1 exhibits copy-number variation in human populations (*37*, *38*) and, together with a similar orthologous cluster on mouse chromosome 4, apparently arose by translocation not long before the divergence of primates and

rodents, about 85 Mya (*39*) (Fig. 5 and fig S5.2). After that translocation event, the human and mouse gene clusters expanded independently. Evidence for positive selection has been found in these genes, and two segmental duplications postdating human-chimpanzee divergence added about a dozen genes to the human cluster.

To properly resolve evolutionary changes in the PRAME gene family, we further sequenced six macaque BAC clones to achieve a higher data quality, and we assembled them into a single contig (table S2.6). These eight PRAME genes were compared with human and chimpanzee genes identified from the latest assemblies for both species. We estimated a phylogeny for all identified genes, designating the mouse gene cluster and the human *PRAME* gene on chromosome 22 as outgroups. We then reconciled this gene tree with the species tree by maximum parsimony. Our reconstruction reveals extensive duplication early in primate evolution (Fig. 5B, branch *a*), in recent chimpanzee evolution (Fig. 5B, branch *d*), and, most notably, in recent human evolution (Fig. 5B, branch *e*). The PRAME gene cluster appears to have been much less dynamic on the macaque lineage (Fig. 5B, branch *b*) and in early hominins (the human and chimpanzee branch, Fig. 5B, branch *c*). A large inverted tandem duplication occurred on the macaque lineage shortly after divergence from the human lineage, but no additional large-scale rearrangements are evident. The relative quiescence in macaque allows us to identify older duplications that are difficult to discern in the exceedingly complex human self-alignments (*7*).

The inferred PRAME gene tree shows pronounced differences in evolutionary rates across branches, as well as some quite long branches that suggest bursts of adaptive change. Using maximum likelihood methods, we found evidence of positive selection on several of these branches (Fig. 5A). This positive selection, combined with the highly variable pattern of gene duplication and expansion, suggests that the PRAME gene family has played a key role in species evolution.

We identified a second segment of extensive genomic duplications concentrated at the telomere of macaque chromosome 9, orthologous to a human locus at 10p15.3 and observed by multiple approaches to be distributed throughout the macaque genome. The genes phosphofructokinase-platelet form (*PFKP*) and *DIP2C* were expanded in this region and yielded the highest array CGH macaque-to-human ratios in the genome (average $\log_2$ ratios of 3.30 and 2.54, respectively). *DIP2C* is implicated in segmentation patterning, although its relevance to macaque evolution is currently obscure. *PFKP* is important in sugar (fructose) metabolism, raising the possibility that the pronounced copy-number expansion in macaque may be relevant to the

Table 3. Gene families with significant copy-number expansions ($P < 0.0001$) in the human and the identical statistic for the rhesus macaque. Gene family ID, identification numbers from Ensembl version 41. Family size, number of gene copies in the current genome assemblies. Gains and losses, number of genes gained and lost since the human's split with chimpanzee or the macaque's split with human-chimpanzee lineage. IG, immunoglobulin; IGE, immunoglobulin E; Pre, precursor; MHC, major histocompatibility complex; TCR, T cell receptor; ENV, envelope; ATP, adenosine 5′-triphosphate.

Gene family ID	Description	Family size	Gains	Losses
Expanded in human				
ENSF00000000020	IG heavy chain V region	42	10	0
ENSF00000000073	Receptor	56	16	0
ENSF00000000233	Peptidyl prolyl cis trans isomerase	38	9	0
ENSF00000000312	Histone H2b	28	7	0
ENSF00000000597	Golgin subfamily A	49	26	0
ENSF00000000664	Ankyrin repeat domain	33	9	0
ENSF00000000822	Unknown	15	9	0
ENSF00000000841	Tripartite motif	21	7	1
ENSF00000000936	Centaurin gamma	15	9	0
ENSF00000001036	Cold inducible RNA binding	22	8	0
ENSF00000001546	Ubiquitin carboxyl terminal hydrolase	16	13	2
ENSF00000001599	Leucine-rich repeat	14	7	0
ENSF00000001665	DNA mismatch repair PMS2	12	5	0
ENSF00000001738	Unknown	15	7	0
ENSF00000001920	40S ribosomal S26	13	7	1
ENSF00000001974	Unknown	17	3	0
ENSF00000002160	Double homeobox	15	13	0
ENSF00000002570	Keratin associated 5	7	2	0
ENSF00000003683	Unknown	5	3	0
ENSF00000004835	Ambiguous	13	9	0
Expanded in macaque				
ENSF00000000014	HLA class I	17	12	0
ENSF00000000037	HLA class I	16	10	0
ENSF00000000070	Keratin type I	65	30	0
ENSF00000000077	Histone H3	32	11	0
ENSF00000000085	IG kappa chain V region	47	22	2
ENSF00000000138	Keratin type II	39	10	0
ENSF00000000150	Taste receptor type 2	23	9	0
ENSF00000000178	Aldo keto reductase family 1	19	9	0
ENSF00000000397	Ral guanine nucleotide dissoc stim.	19	10	1
ENSF00000000432	Killer cell IG receptor Pre MHC class I	9	3	0
ENSF00000000630	TCR beta chain V region Pre	18	9	0
ENSF00000000705	ENV polyprotein	13	11	0
ENSF00000000766	60S ribosomal l7A	26	17	1
ENSF00000000773	Ribosomal l7	23	12	0
ENSF00000000826	60S ribosomal l23A	20	6	0
ENSF00000001027	60S ribosomal l17	12	3	0
ENSF00000001077	Nucleoplasmin	17	9	0
ENSF00000001211	67-kD laminin	18	10	0
ENSF00000001235	Nonhistone chromosomal HMG 17	24	12	0
ENSF00000001236	60S ribosomal l31	23	11	0
ENSF00000001249	60S ribosomal l12	16	8	0
ENSF00000001359	USP6 N terminal	14	10	0
ENSF00000001460	Prohibitin	7	4	0
ENSF00000001671	60S ribosomal l32	10	6	0
ENSF00000001861	40S ribosomal S10	9	5	0
ENSF00000002239	60S ribosomal l19	8	5	0
ENSF00000002279	40S ribosomal S17	8	4	0
ENSF00000002476	60S ribosomal l18	7	4	0
ENSF00000002633	IGE binding	19	14	0
ENSF00000003321	Argininosuccinate synthase	9	6	0
ENSF00000003395	10-kD heat shock protein	11	8	0
ENSF00000004083	ATP synthase subunit G	4	3	0
ENSF00000007347	Unknown	7	3	0

Table 4. Genes identified as expanded in copy number in the macaque, relative to the human, by the array CGH method. The leftmost column represents IMAGE cDNA clones that show array CGH–predicted copy number increases in the rhesus macaque relative to the human. The middle two columns list corresponding gene names and array CGH $\log_2$ macaque-to-human ratios. The rightmost column presents BLAT-predicted copy numbers based on rheMac2 and hg18 genome assemblies.

IMAGE clone	Gene	Average $\log_2$ array CGH ratio	rheMac2/hg18 BLAT-predicted copy numbers
41109/1900937	*PFKP*	3.30	3/2†
454926/1862434	*DIP2C*	2.54	4/2†
1475421/757369	*EST*	1.74	7/4†
50877/110020	*EST*	1.48	3/4
795258/1574131/191978	*ATP5J2*	1.42	29/10†
824545/278888	*EST*	1.37	0/1
2457916/322067	*DNAJC8*	1.37	9/6†
504421/435036	*ADFP*	1.27	6/4†
769921/146882	*UBE2C*	1.17	8/3†
155620/154809	*IGL*	1.14	8/10
1985794*/1470105	*EST*	1.14	1/6
32083/270786	*FLJ30436*	1.13	2/3
306344/773260	*MAT2B*	1.11	3/2†
884480/194908	*COX7C/PRO2463*	1.09	13/4†
244205*/462961/768172/824776*/123971	*DHFR*	1.05	14/13†
72745/1626871*	*HLA*	1.02	1/5
1493107/1637726	*LTB4DH/EST*	0.97	3/2†
163407/843374	*STOM*	0.96	2/2
258666/428043	*PSMB7/EST*	0.96	3/2†
112498/824894	*EST*	0.95	0/0
32231/770984	*EST/FLJ12442*	0.95	3/2†
981713*/953542*/981925*	*EST*	0.95	0/0
1636233/814459	*C9orf23*	0.93	4/2†
529185/609265	*SELK*	0.93	6/4†
298965/1472754/512003*	*COX6B*	0.93	7/4†
322561/240620*	*EST*	0.92	31/22†
208656/415195	*FLJ20294*	0.92	2/2
840698/39977	*FLJ20254/MAPRE3*	0.90	4/2†
773287/1635681	*NDUFA2*	0.89	4/2†
756763*/725401*	*EST*	0.85	1/1
80742/80694	*EST*	0.85	0/0
1415672*/1558664*	*EST*	0.84	0/0
323806/38029	*EST*	0.83	2/2
595547/997889*	*EST*	0.83	1/1
953654*/953643*	*EST*	0.83	0/0
783035*/783249*	*EST*	0.83	1/1
884272*/1415750*	*H3F3A*	0.83	45/40†
322175/210873	*EST/PPY2*	0.81	1/1
1569731/1569604	*EST*	0.79	4/4
292982/129431	*EST*	0.78	1/2
112785/361565*	*RoXaN/GLUD1*	0.78	7/4†
292452/450327	*SMBP*	0.78	2/2
1606275/1534633	*Corf129/STOM*	0.77	3/2†
212847*/1415750*	*EST*	0.76	22/16†
664121*/745347*	*PIG7/EST*	0.76	4/2†
982122/982113/121546/503715	*EST/FLJ14668*	0.73	9/6†
950688/811603	*EST/ATP6V1G1*	0.73	4/3†
327202/194384	*EST/BTF3*	0.72	1/1
897007/897676*	*EST*	0.71	1/1
301388/825470	*TOP2A*	0.69	6/2†
590390/756469*	*RoXaN*	0.62	3/2†

*Consistent with computational analysis of gene family gains and losses. †BLAT-based copy-number estimates of rheMac2 and hg18 genome assemblies that are consistent with array CGH predictions.

high-fruit diet common among macaques. As with other array CGH copy-number estimates, the functional status of the additional copies is not known. Six of the individual macaque BACs that mapped to the region revealed related duplicated sequences on rhesus chromosome 3, which formed from the fusion of orthologs of human chromosomes 7 and 21, suggesting that these genes may have played a role in this expansion.

Another macaque-specific increase involves the 22 HLA-related genes located in the region orthologous to human chromosome 6p21 (table S5.4). A previous study found that HLA gene copy number was higher in the macaque than in the human (*40*), and our results confirm and extend this finding, demonstrating that the macaque HLA copy number is greater than that found for the human as well as all four great ape species (fig S5.3). This finding also suggests that, although the macaque has been extensively used to model the human immune response, there may be substantial and previously unappreciated differences in HLA function between these species. Notably, the copy number of another immune system–related gene cluster, immunoglobulin lambda-like (*IGL*) at 22q11.23, is also predicted to be increased in the macaque (table S5.4). Members of the *IGL* locus encode light chain subunits that are part of the Pre–B cell receptor; do not undergo rearrangements; and, when mutated, can result in B cell deficiency and agammaglobulinemia. Additional known genes predicted by array CGH to have markedly increased copy numbers in the macaque relative to the human include *DHFR*, *ATP5J2*, *DNAJC8*, *ADFP*, and *MAT2B*. Overall, the main characteristics of the set of amplified genes were their diversity and the wide variety of genomic regions they occupied.

Orthologous Relationships

The macaque genome has also allowed for a detailed study of more subtle changes that have accumulated within orthologous primate genes. The average human gene differs from its ortholog in the macaque by 12 nonsynonymous and 22 synonymous substitutions, whereas it differs from its ortholog in the chimpanzee by fewer than three nonsynonymous and five synonymous substitutions. Similarly, 89% of human-macaque orthologs differ at the amino acid level, as compared with only 71% of human-chimpanzee orthologs. Thus, the chimpanzee and human genomes are in many ways too similar for characterizing protein-coding evolution in primates, but the added divergence of the macaque helps substantially in clarifying the signatures of natural selection.

General characteristics of orthologous genes. We developed an automatic pipeline to identify 10,376 trios of HCR genes to which we could assign a high confidence of 1:1:1 orthology. For comparison, we also identified 6762 hu-

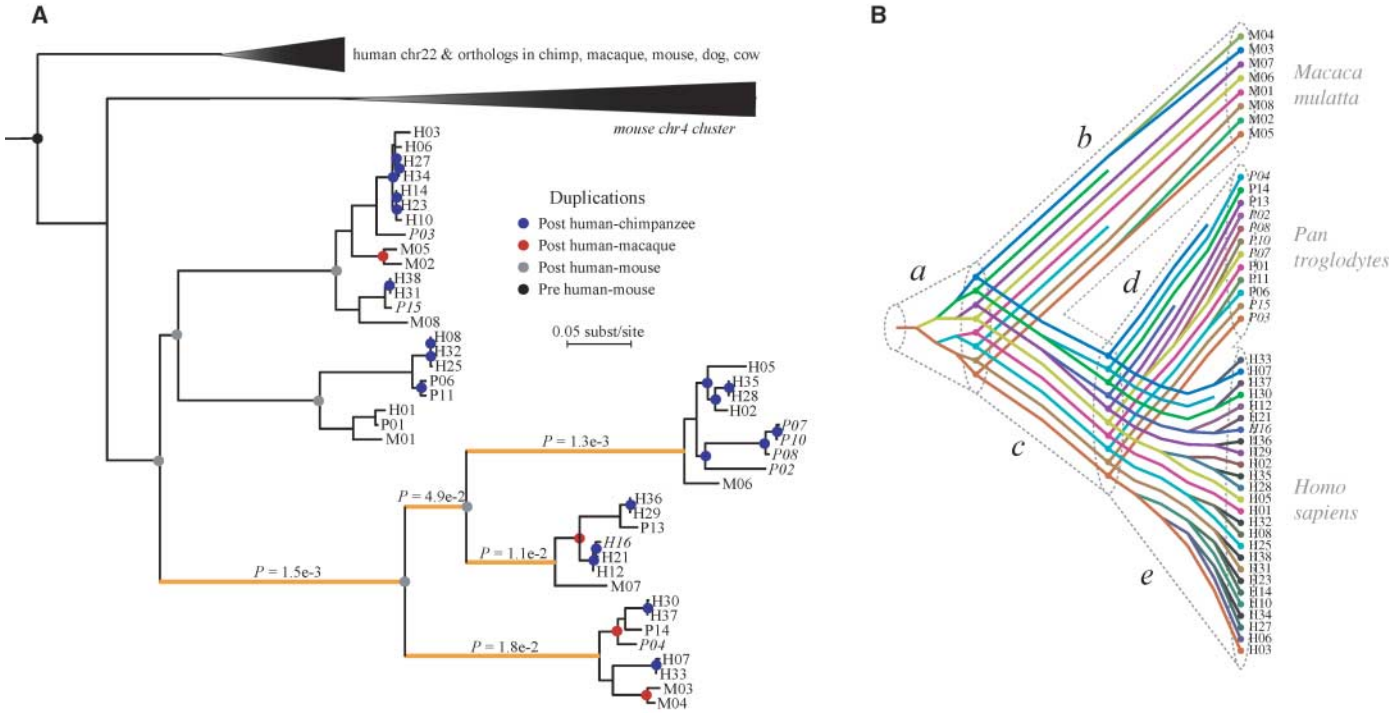

Fig. 5. Organization of the PRAME gene cluster in the HCR lineages. (**A**) Maximum-likelihood phylogeny for PRAME-like genes in the human (H), chimpanzee (P), and rhesus macaque (M) genomes. Colored circles indicate inferred duplication events, partial genes are shown in italics, and branches showing significant evidence of positive selection are colored orange (*P* values are shown above orange lines). Scale bar, 0.05 substitutions per site. (**B**) Another view of the same phylogeny, showing the duplication history in the context of the species tree (*7*).

man, macaque, mouse, and rat quartets; 5641 HCR, mouse, and rat quintets; and 5286 HCR, mouse, and dog quintets. Because the human gene models are by far the best characterized for primates, we first identified a set of 21,256 known human protein-coding genes derived from a union of the RefSeq (*41*), Vega (*42*), and University of California–Santa Cruz Known Genes (*43*) collections. These genes were then mapped to synteny-based genome-wide multiple alignments (*44*, *45*) and subjected to a series of rigorous filters to eliminate spurious annotations, paralogous alignments, genes that have become pseudogenized in one or more species, and genes with incompletely conserved exon-intron structures (*7*). The genes that pass all filters represent 1:1:1 orthologs in which aligned protein-coding bases are highly likely to encode proteins in all species, with identical reading frames.

Despite the draft quality of the chimpanzee and macaque assemblies, the majority of human genes mapped through syntenic alignments to the chimpanzee (93% of genes) and macaque (89%) genomes (Fig. 6) (*7*), and most of these genes were completely alignable in their coding regions. Fairly large fractions of human genes, however, were discarded because of apparent frame-shift insertions and deletions (indels) or nonconserved exon-intron structures with respect to their putative chimpanzee or macaque orthologs. On the basis of 81 finished BACS covering 294 genes, we estimate that, out of 5526 genes failing the filters for alignment completeness, frame-shift indels, and conserved exon-intron structure, 2138 (39%) were discarded completely because of flaws in the macaque assembly; the remaining 3388 (61%) were discarded either because of genuine changes to genes or because of annotation or alignment errors (*7*). Another 2261 genes passed the human-macaque filters but failed the human-chimpanzee filters, and a large majority of these failures were probably due to flaws in the chimpanzee assembly. Altogether, we estimate that finished genomes for the macaque and chimpanzee would allow the number of genes in high-confidence orthologous trios to be increased by at least 23%, to ~12,800 (*7*). Notably, our conservative ortholog sets may create a bias against fast-evolving genes and therefore may lead to underestimates of average levels of divergence and the prevalence of positive selection.

Alignments of the 10,376 orthologous trios were used to estimate the ratio of the rates of nonsynonymous and synonymous substitutions per gene (denoted ω), with continuous-time Markov models of codon evolution and maximum likelihood methods for parameter estimation (*46*–*48*). This yielded a mean estimate of ω = 0.247 (median 0.144), close to the value of 0.23 estimated for human and chimpanzee genes (*29*). About 9.8% of all genes show no nonsynonymous changes in the three species, and 2.8% have ω > 1, suggesting that they are under positive selection. Consistent with previous studies (*49*), certain classes of genes exhibit unusually large or small ω values, such as those assigned to the gene ontology (*50*) category "immune response," which have an ω distribution shifted significantly toward larger values, and those assigned to the "transcription factor activity" category, which have a distribution shifted toward smaller values (fig. S6.1).

Our estimates for ω in primates are considerably larger than previously reported estimates for rodents, which have a median of 0.11 (*3*), and larger than similar estimates from primate-versus-rodent comparisons (*29*) (Fig. 7). To compare the average rates of evolution of protein-coding genes in primates with those in other mammals, we estimated a separate value of ω for each branch of a five-species phylogeny, pooling data from all 5286 one-to-one orthologs for these species (fig. S6.2). We obtained similar estimates of ω for the human (ω = 0.169) and chimpanzee (ω = 0.175) lineages, but substantially smaller estimates for the branches leading to nonprimate mammals (ω = 0.104 to 0.128), suggesting a reduction in purifying selection in hominins (*29*). The estimate of ω for the macaque lineage (ω = 0.124) is substantially smaller than the estimates for the human and chimpanzee and is closer to the estimates for the mouse and dog, perhaps reflecting the larger population size of

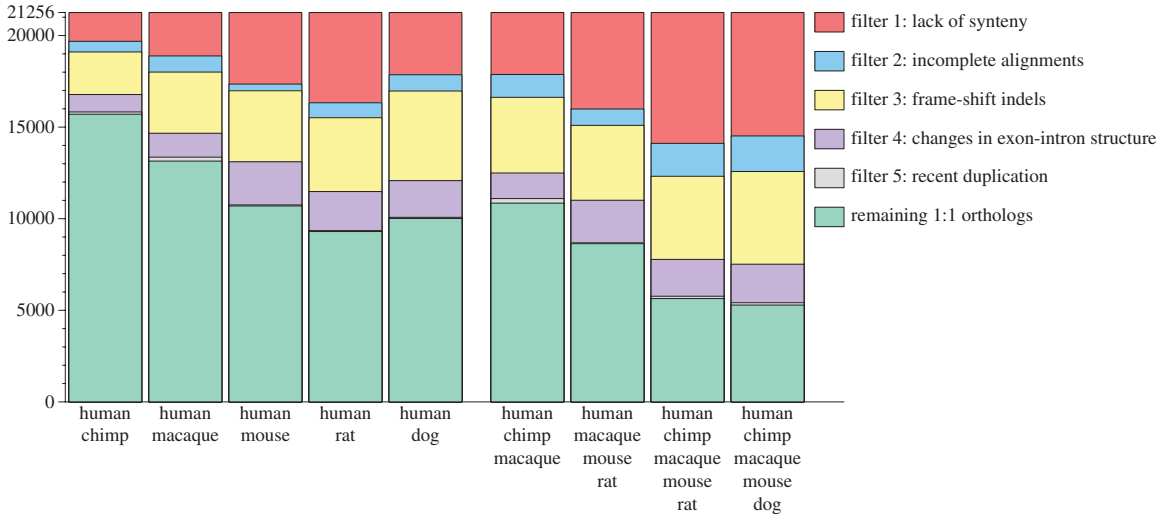

Fig. 6. Numbers of human genes passing successive filters in the orthology analysis pipeline. Genes are required to fall in regions of large-scale synteny between genomes, to have completely aligned coding regions, not to have frame-shift indels or altered gene structures, and not to show signs of recent duplication.

macaques compared with the other primates. The estimates for the internal branches between the most recent common ancestors of the human and mouse and of the human and macaque, as well as the most recent common ancestors of the human and macaque and of the human and chimpanzee, are nearly equal to the macaque estimate. This suggests that protein-coding sequence evolution in macaques may have occurred at a typical primate rate, whereas it is the elevated rates in hominins that may be anomalous.

When primate and rodent ω of individual genes were compared, primate orthologs were found to be evolving more rapidly by a 3:2 ratio. This asymmetry was also evident among genes showing substantial differences in primate ω (ω_p), on the basis of human-macaque alignments, and rodent ω (ω_r), deduced from mouse-rat alignments. According to a strict Bonferroni correction for multiple testing, 22 genes showed statistically significant $\omega_p > \omega_r$, whereas only three genes showed $\omega_r > \omega_p$ (McNemar $P < 0.001$). If multiple testing criteria are relaxed, the bias toward larger ω_p is more notable (144 versus 8; tables S6.1 and S6.2). Cases of $\omega_p > \omega_r$ generally reflect an increase in ω_p, whereas cases of $\omega_r > \omega_p$ result both from an increase in ω_r and a decrease in ω_p. The genes showing statistically significant $\omega_p > \omega_r$ are enriched for functions in sensory perception of smell and taste as well as for regulation of transcription (*7*).

Positive selection. Taking advantage of the additional phylogenetic information provided by the macaque genome, we performed a genome-wide scan for positive selection, using our 10,376 HCR orthologous trios and likelihood ratio tests (LRTs) (*51*–*53*). Four different LRTs were performed: test TA, for positive selection across all branches of the phylogeny, and tests TH, TC, and TM for positive selection on the individual branches to human, chimpanzee, and macaque, respectively. Our methods use an unrooted tree and cannot distinguish between the branches to macaque and the human-chimpanzee ancestor; for convenience, we refer to the combined branch as the macaque branch. In all cases, variation among sites in ω was allowed and, to reduce the number of parameters to estimate per gene, the branch-length proportions and transition-transversion ratio (κ) were estimated by pooling data from genes of similar G+C content (*7*). Test TA identified 67 genes, and tests TH, TC, and TM identified 2, 14, and 131 genes (false-discovery rate (FDR) < 0.1 in all cases), respectively. The large number of genes identified for the macaque branch is partly a reflection of its greater length compared with the chimpanzee and human branches (*7*).

These four sets of genes overlap considerably, particularly among their highest scoring predictions (Table 5 and table S6.3). Their union contains 178 genes, or 1.7% of all genes tested. The two genes identified by TH—those encoding the leukocyte immunoglobulin-like receptor LILRB1 and hypothetical protein LOC399947—were also identified by TA, and the gene for LILRB1 was identified by TC as well, indicating evidence of positive selection on multiple branches. However, 12 out of 14 genes identified by TC were not identified by the other tests, indicating possible lineage-specific selection in the chimpanzee. These include sex comb on midleg-like 1 (*SCML1*) and protamine 1 (*PRM1*), which were previously identified in an analysis that could not distinguish between selection on the human and chimpanzee branches (*52*). In addition, 99 genes were identified by TM but not the other tests. These genes may be under lineage-specific selection in the macaque and/or may have experienced positive selection on the branch leading to the most recent common ancestor of the human and chimpanzee.

The genes identified by our tests for positive selection are enriched for several categories from the gene ontology (*50*) and Protein Analysis Through Evolutionary Relationships (PANTHER) (*54*) classification systems that are similar to those observed in previous genome-wide scans for positive selection (*52*, *53*). These include defense response, immune response, T cell–mediated immunity, signal transduction,

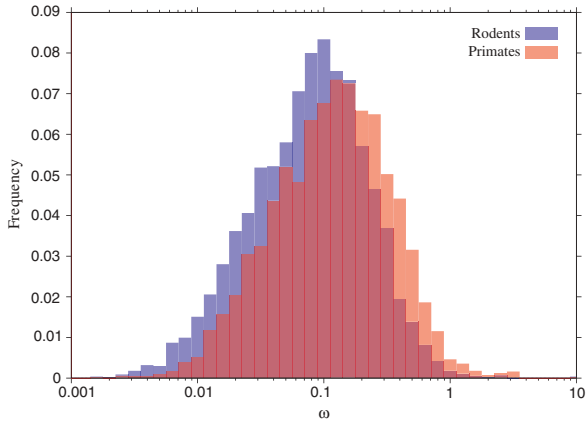

Fig. 7. Distributions of ω in primates versus rodents. Histogram of estimates of $\omega = dN/dS$ for human, chimpanzee, and macaque versus estimates for mouse and rat in 5641 orthologous quintets, showing a pronounced shift toward larger values in primates ($P = 2.2 \times 10^{-16}$, Mann Whitney test). Genes with $dN = 0$ or $dS = 0$ are counted in the relative frequencies but not shown.

and cell adhesion (tables S6.4 to S6.7). Among the genes in these categories are several immunoglobulin-like genes, including those that encode the leukocyte-associated inhibitory receptors LILRB1 and LAIR1 (located in a cluster on chromosome 19), the T cell surface glycoprotein CD3 epsilon chain precursor CD3E, and the intercellular adhesion molecule 1 precursor ICAM1. Other identified genes associated with cell adhesion and/or signal transduction include those that encode DSG1, a calcium-binding transmembrane component of desmosomes, and the transmembrane protein TSPAN8 (which has gained an exon by duplication in the macaque genome). Genes encoding membrane proteins in general are strongly overrepresented; other examples include the genes that encode connexin 40.1, active in cell communication, and *OPN1SW*, the gene encoding blue-sensitive opsin.

In addition, we observed strong enrichments for new categories such as iron ion binding [e.g., the beta globin (*HBB*), lactotransferrin (*LTF*), and cytochrome B-245 heavy chain genes (*CYBB*)] and oxidoreductase activity (e.g., *KRTAP*5-8 and *KRTAP*5-4, which encode keratin-associated proteins, and *NDUFS*5, which encodes a subunit of the nicotinamide adenine dinucleotide ubiquinone oxidoreductase). Two keratin genes, which are important for hair-shaft formation, are present among the top-scoring genes; these genes could conceivably have come under positive selection as a result of mate selection or climate change. Genes classified as part of the extracellular region, which include the keratin genes, are in general overrepresented. Many of the identified genes from this category encode secreted proteins, such as the interferon alpha 8 precursor IFNA8, which exhibits antiviral activity; the interleukin 8 precursor IL8, a mediator of inflammatory response; and CRISP1, which is expressed in the epididymis and plays a role at fertilization in sperm-egg fusion.

We found only weak enrichments for genes involved in apoptosis and spermatogenesis (*52*), but we did see a significant excess of high likelihood ratios among genes involved in fertilization. Other categories that show an excess of high likelihood ratios but that are not enriched for genes identified by our tests include blood coagulation, response to wounding, and related categories; epidermis morphogenesis; KRAB-box transcription factor; and olfactory receptor activity (tables S6.6 and S6.7). Their elevated likelihood ratios may reflect either weak positive selection or relaxation of constraint.

The inclusion of the macaque genome substantially improves statistical power to detect positive selection in primates, compared with previous scans that used only the human and chimpanzee genomes (*29*, *52*). By examining about 8000 human-chimpanzee alignments with a similar LRT, Nielsen *et al.* (*52*) were able to identify only 35 genes with nominal $P < 0.05$, and when considering multiple comparisons, they were able to establish only that a 5% false discovery rate set was nonempty. By contrast, the use of the macaque genome allows the identification of 15 genes under positive selection in hominins and an additional 163 under selection on one or more other branches of the phylogeny, with FDR < 0.1. We estimate that including the macaque genome makes test TA about three times as powerful. However, including macaque rather than mouse (*53*) as an outgroup improves the power of test TH only marginally (*7*).

The genes identified by the LRTs are generally randomly distributed in the genome, and no significant clustering was observed when tested ($P = 0.24$), although small clusters were found on human chromosomes 11 and 19 (*7*). Chromosome 11, with 10 genes identified by test TA, has more than twice the expected number of genes under positive selection, but this enrichment is not significant after correcting for multiple comparisons [$P = 0.10$, Fisher's exact test and Holm correction (*7*)]. However, a significant enrichment was observed for genes overlapping segmental duplications that oc-

Table 5. Selected genes from top 40 showing evidence of positive selection in primates. Accession, the number of the reference transcript for each gene (human). Chr, human chromosome on which reference gene resides. *P* value, nominal *P* value for test TA (*7*). Genes shown have FDR < 0.04. Test, the test (other than test TA) that detected the given gene. The Dup column has a checkmark if a gene overlaps a segmental duplication preceding the human/macaque divergence.

Accession	Gene name	Chr	Description	*P* value	Test	Dup
AB126077	*KRTAP5-8*	11	Keratin-associated protein 5-8	6.20×10^{-16}	TM	✓
NM_006669	*LILRB1*	19	Leukocyte immunoglobulin-like receptor	7.20×10^{-14}	TH, TC	✓
NM_001942	*DSG1*	18	Desmoglein 1 preproprotein	1.10×10^{-10}	–	
NM_173523	*MAGEB6*	X	Melanoma antigen family B, 6	5.30×10^{-8}	TC	✓
NM_054032	*MRGPRX4*	11	G protein–coupled receptor MRGX4	5.60×10^{-8}	TM	✓
NM_000397	*CYBB*	X	Cytochrome b-245, beta polypeptide	1.50×10^{-7}	TM	
NM_001911	*CTSG*	14	Cathepsin G preproprotein	1.50×10^{-7}	TM	
NM_000735	*CGA*	6	Glycoprotein hormones, alpha polypeptide	1.20×10^{-6}	TM	
NM_001012709	*KRTAP5-4*	11	Keratin-associated protein 5-4	2.70×10^{-6}	TM	✓
NM_000201	*ICAM1*	19	Intercellular adhesion molecule 1 precursor	2.70×10^{-6}	TM	
NM_001131	*CRISP1*	6	Acidic epididymal glycoprotein-like 1 isoform 1	1.60×10^{-5}	TM	
NM_002287	*LAIR1*	19	Leukocyte-associated immunoglobulin-like	3.10×10^{-5}	TM	✓
NM_153368	*CX40.1*	10	Connexin40.1	4.90×10^{-5}	–	
NM_018643	*TREM1*	6	Triggering receptor expressed on myeloid cells	6.30×10^{-5}	TM	
NM_000300	*PLA2G2A*	1	Phospholipase A2, group IIA	1.30×10^{-4}	–	
BC020840	*TCRA*	14	T cell receptor alpha chain C region	1.50×10^{-4}	–	
NM_000733	*CD3E*	11	CD3E antigen, epsilon polypeptide	1.50×10^{-4}	TM	
NM_001014975	*CFH*	1	Complement factor H isoform b precursor	1.50×10^{-4}	–	
NM_001423	*EMP1*	12	Epithelial membrane protein 1	1.50×10^{-4}	TM	
NM_001424	*EMP2*	16	Epithelial membrane protein 2	1.50×10^{-4}	TM	
NM_002170	*IFNA8*	9	Interferon, alpha 8	1.50×10^{-4}	–	
NM_030766	*BCL2L14*	12	BCL2-like 14 isoform 2	1.50×10^{-4}	–	
NM_006464	*TGOLN2*	2	Trans-golgi network protein 2	1.80×10^{-4}	TM	
NM_014317	*PDSS1*	10	Prenyl diphosphate synthase, subunit 1	1.80×10^{-4}	–	
NM_000518	*HBB*	11	Beta globin	2.00×10^{-4}	TM	

curred before the human-macaque divergence (P = 0.006, Fisher's exact test), suggesting an increased likelihood of adaptive evolution following gene duplication. Four of the top five genes identified by test TA overlap segmental duplications that predate the human-macaque divergence (Table 5).

Genetic Variation in Macaques

The use of rhesus macaques as animal models of human physiology can be greatly enhanced by an improved understanding of their underlying genetic variation. To explore rhesus genetic diversity and to create resources for further genetic studies, we generated a total of 26.2 Mb of whole-genome shotgun sequence from 16 unrelated individuals (eight of Chinese origin and eight of Indian origin, table S7.1). We next identified 26,479 single-base differences [putative single-nucleotide polymorphisms (SNPs)] through comparison with the reference genome. Overall, we found approximately one SNP per kilobase, which is on average close to that found in similar human studies. There was a surprising difference of 50% in overall diversity between the autosomes and the X chromosome (Fig. 8A); we expected a value of 75%. This expectation was based on differences in effective chromosome population sizes, given that females have two X chromosomes and males carry only one. The reduction in diversity could be due to recent selective sweeps of positively selected recessive mutations on the X chromosome (*55*).

We also found that the frequency of the whole-genome shotgun SNPs differed substantially among the animals from the different populations (0.95/kb in Indian rhesus and 1.06/kb in Chinese rhesus), and there was suggestive variation in SNP density within their subpopulations (SD = 0.0275/kb for Chinese macaques; SD = 0.0527/kb for Indian macaques). Together with complementary data from PCR analysis of polymorphic L1 and *Alu* element insertions (figs. S7.1 and S7.2) that showed population substructure, this prompted additional experiments in which 48 animals from the two populations were surveyed by PCR-direct DNA sequencing. Details and most conclusions from that study have been reported by Hernandez *et al*. (*56*), including a demonstration that >67% of SNPs discovered by direct sequencing are private to each subpopulation. The strong population differentiation is reflected in fixation index (FST) values (a measure of population differentiation) and a marked difference in Watterson's (*57*) estimate of the population mutation rate between the two groups. Here, we observed that the population differences are also reflected in differential distribution of Tajima's D statistic and in linkage disequilibrium across sampled regions (Fig. 8, B and C). Each of these statistics further reflects the possibilities of sweeps of natural selection or major differences in population histories that must be factored into ongoing genetic studies. These initial insights into the underlying patterns of variation within individual animals will therefore provide the basis for future genetic analyses. In addition to their utility for identification of individual animals, the SNP markers will be invaluable for larger-scale population studies.

Male mutation bias. A comparison of human-rhesus substitution rates (calculated at interspersed repetitive elements) between the X chromosome and the autosomes yielded an estimate of the male-to-female mutation rate ratio (α) of 2.87 (95% CI = 2.37 to 3.81; table S7.2). This value is lower than $\alpha = 6$ estimated for the human and chimpanzee (*58*) but higher than $\alpha = 2$ estimated for the mouse and rat (*3*, *59*). Thus, this argues against a uniform magnitude of male mutation bias in mammals (*5*) and supports a correlation between male mutation bias and generation time (*60*, *61*).

Human Disease Orthologs in the Macaque

While the general morphological and physiological similarities between humans and macaques greatly enhance the utility of the latter as a model organism, specific differences in their underlying coding sequences can also provide biological insights. By comparing human disease genes with their macaque equivalents, we identified numerous instances in which the allele observed in the macaque corresponds to the disease allele in the human. These occurrences suggest that the human disease variants could be either persistent (i.e., ancestral) or recurring sequences that represent the recapitulation of ancestral states that may once have been protective, but which now result in adverse consequences for human health (*62*).

To identify the ancestral disease-associated alleles in human, we screened the macaque and chimpanzee assemblies for the presence of any of the 64,251 different disease-causing or disease-associated mutations collected in the Human Gene Mutation Database (*63*, *64*). A total of 229 substitutions were identified for which the amino acid considered to be mutant in human corresponded to the wild-type amino acid present in macaque, chimpanzee, and/or a reconstructed ancestral genome (Table 6) (*65*) (see table S8.1 for a full list).

One surprising result of the analysis was the identification of several human loci that, when mutated, give rise to profound clinical phenotypes, including severe mental retardation. For example, the macaque data revealed deleterious alleles in the ornithine transcarbamylase (OTC) and phenylalanine hydroxylase (PAH) genes, which are associated in human with OTC deficiency and phenylketonuria. In humans, these mutations greatly perturb the normal serum amino acid levels. Direct examination of macaque blood revealed lower concentrations of cystine and cysteine than in the human and slightly higher concentrations of glycine than in the human, but no increase in phenylalanine or ammonia, which might have been a predicted result of these changes (tables S8.2 and S8.3). Although the effect of the observed alleles might be greatly influenced by compensatory mutations (*66*) or other environ-

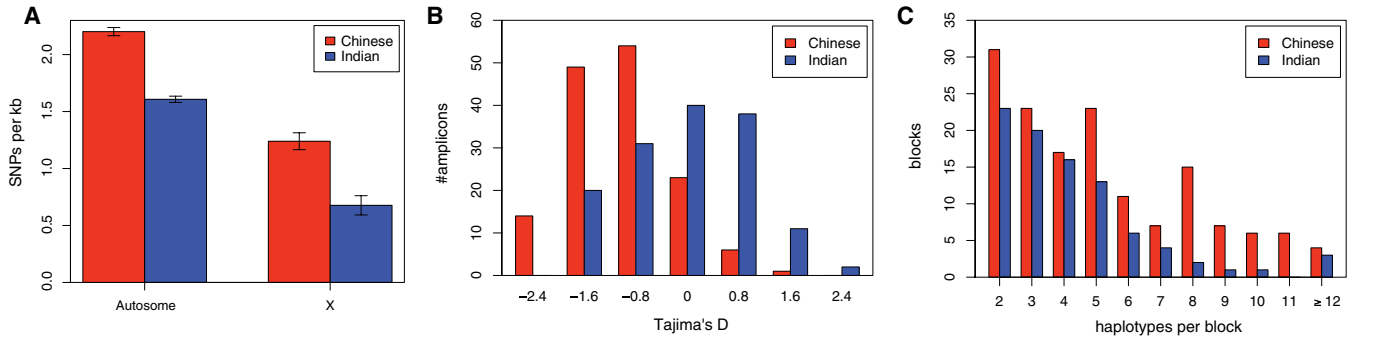

Fig. 8. SNP within rhesus macaques. (**A**) SNP densities per kilobase for eight Chinese (blue) and eight Indian (red) individuals in autosomes and the X chromosome. Error bars indicate standard error with variance calculated across individual-chromosome replicates. (**B**) Distribution of Tajima's D statistic across 166 amplicons for each population (n = 38 for Indian and n = 9 for Chinese individuals). (**C**) The distribution of the number of haplotypes per haplotype block (determined using the four-gamete test) across five regions.

mental factors, it remains a possibility that the basic metabolic machinery of the macaque may exhibit functionally important differences with respect to our own (Fig. 9).

Ancestral mutations were also identified in the *N*-alpha-acetylglucosaminidase (NAGLU) gene that gives rise to mucopolysaccharidoses (Sanfillipo syndrome), which is also characterized by profound mental retardation. Their occurrence invites further investigation of the contribution of this and related genes to the phenotypic differences between macaques and humans, and the potential for further exploration of these monkeys as models for this disorder.

We also identified a human mutation associated with Stargardt disease and macular dystrophy that matches an ancestral allele by replacing lysine with glutamine at position 223 of the human ABCA4 protein (Fig. 9). Umeda *et al*. (*67*) reported the presence of the glutamine in a cynomolgus monkey, and all other eutherian mammals as well as the predicted boreoeutherian ancestral sequence have glutamine at this position. Furthermore, glutamine is present at this residue in *Xenopus*, thereby implying conservation through some 300 million years of vertebrate evolution. Thus, it may be inferred that the ancestral glutamine has been replaced by lysine in humans. Similarly, one *CFTR* mutation [$Phe^{87}\rightarrow Leu^{87}$ (Phe87Leu)] is present not only in most mammals (Fig. 9) but also in *Fugu*, also implying extensive conservation through vertebrate evolution.

Impact of a Genomic Sequence on Biological Studies

In addition to its impact on comparative and genetic studies, the genome sequence reported here heralds a new era in laboratory studies of macaque biology. The full potential for more precise definition of this animal model and its gene content is not yet realized, but the value of the new sequence in guiding DNA microarrays for studying macaque gene expression has already become clear (*68*). Previously, human or macaque EST-based arrays had been used for expression studies (*69*). The most recently released microarray now adds probes designed by alignment of the 3′ untranslated regions of 23,000 human RefSeq genes to the sequences from the initial macaque genome release (January 2005, Mmul0.1, approximate genome coverage of 3.5-fold). The vast majority of the probes on this array (98.5%) now match the current macaque genome release with high confidence and represent 18,690 unique genomic loci. These provide a representation of recognized functional pathways with an enhancement about three times that of the previous data, and overall more uniform and robust hybridization signals compared with those of previous microarrays (*69*) (tables S9.1 to S9.3).

The power of global transcriptional profiling with advanced macaque-specific reagents has been demonstrated in studies of virulence and pathogenicity of influenza from historic pandemic strains, as well as from emerging agents of zoonotic origin. We infected macaques with the human influenza strain A/Texas/36/9 (*70*) and compared the expression changes observed in lung tissues to those seen in whole blood during the course of infection. Figure 10 shows a differential time course of expression between interferon-induced genes and genes in the inflammation pathway, in different tissues (table S9.4). The increased expression in lung tissue shortly after infection reflects the early innate response, whereas genes associated with the reemergence of the inflammation pattern at day 7 implicate a transition to an adaptive immune response. These kinds of studies will be crucial for elucidating all of the transitions from innate to adaptive immune responses and are fully enabled by the macaque-specific microarrays developed from the genome sequences.

We expect many more immediate examples of the impact of other tools developed from the finished macaque genome. For example, the requirement for improvements in PCR-based methods is shown by a recent report on the large-scale cloning of terminal exons for macaque genes, in which the use of human primers was successful, on average, in 67% of cases (*71*). Only a native sequence can allow sufficient precision for these types of highly specific assays. A similar increase of activity in studies of the macaque proteome can be predicted, given that early efforts in macaque proteomics have had to rely on human reference sequences for analyzing liquid chromatography and tandem mass spectrometry data (*70*).

Discussion

The draft genomic sequence reported here has already moved the macaque from a model that has been much studied at the level of physiology, behavior, and ecology to a whole-organism system that can be interrogated at the level of the single DNA base. This transformation is evident in the literature as well as in this special section (*15*, *19*, *57*, *72*).

Additional general conclusions emerged from this study. First, the data make it conceivable to define completely all of the operational components of the pathways underlying the individual biological systems that together constitute the functioning adult macaque. For example, a complete description of all the different macaque immune function components will enable an even more thoughtful use of rhesus macaques in areas such as AIDS research and for vaccine production.

Second, we were struck by the high value of adding regions of genome finishing to the draft sequence for the comparative analyses of genes and duplicated structures. This provides an argument for future finished primate genomes.

Third, the data now provide new opportunities to explore the basic biology of this highly successful species. Rhesus macaques retain a broad geographic distribution with reasonably healthy population numbers and widely studied ecology and ethology. The genetic resources generated in this study will

Table 6. Examples of human mutations that cause inherited disease and match an ancestral or nonhuman primate state. Chr:start-stop shows the address in the March 2006 human assembly. Name is the name used by the Human Gene Mutation Database (*64*). The notation "N>A:CHMT" means that N is the consensus human amino acid, A is the disease-associated form, C is in the current chimp assembly, H is in the inferred human-chimp ancestor, M is in rhesus, and T is in the inferred human-rhesus ancestor (the mouse and dog were used as outgroup species) (*73*).

Chr:start-stop	Strand	Name	Replacement N>A:CHMT	Gene	Disease
chr1:94270150-94270152	-	CM014300	R>Q:RRQR	*ABCA4*	Stargardt disease
chr1:94316821-94316823	-	CM015072	H>R:RRRR	*ABCA4*	Stargardt disease
chr1:94337037-94337039	-	CM042258	K>Q:QQQQ	*ABCA4*	Stargardt disease
chr6:26201158-26201160	+	HM030028	V>A:VVAA	*HFE*	Hemochromatosis
chr7:116936418-116936420	+	CM940237	F>L:FFLL	*CFTR*	Cystic fibrosis
chr7:117054872-117054874	+	CM941984	K>R:KKRK	*CFTR*	Cystic fibrosis
chr12:101761685-101761687	-	CM962547	Y>H:YYHY	*PAH*	Phenylketonuria
chr12:101784521-101784523	-	CM941128	I>T:IITI	*PAH*	Phenylketonuria
chr13:51413354-51413356	-	CM044579	V>A:AAAA	*ATP7B*	Wilson disease
chr13:112843266-112843268	+	CM021094	D>E:DDED	*F10*	Factor X deficiency
chr17:37948991-37948993	+	CM040465	R>Q:RRQQ	*NAGLU*	Sanfilippo syndrome B
chr19:43656115-43656117	+	CM064230	S>G:GGGG	*RYR1*	Malignant hyperthermia
chrX:38111528-38111530	+	CM941115	R>H:RRHH	*OTC*	Ornithine hyperammonemia
chrX:38125613-38125615	+	CM961052	T>M:MTTT	*OTC*	Ornithine hyperammonemia
chrX:138458220-138458222	+	CM045148	E>K:EEKK	*F9*	Hemophilia B

undoubtedly form the basis of many analyses of population variability and inter-population diversity.

Finally, the genomic rearrangements, duplications, gene-specific expansions, and measurements of the impact of natural selection presented here have revealed the rich and heterogeneous genomic changes that have occurred during the evolution of the human, chimpanzee, and macaque. The marked diversity of the types of change that have occurred demonstrate a major feature of primate evolution: The aggregation of changes that we see, even in closely related species, does not reflect smooth, progressive, and orderly genomic divergence. Models of abrupt or punctuated evolution already acknowledge that smooth and continuous change is difficult to achieve on an evolutionary time scale, but this study provides a notable example of the operation of this principle in our close relatives.

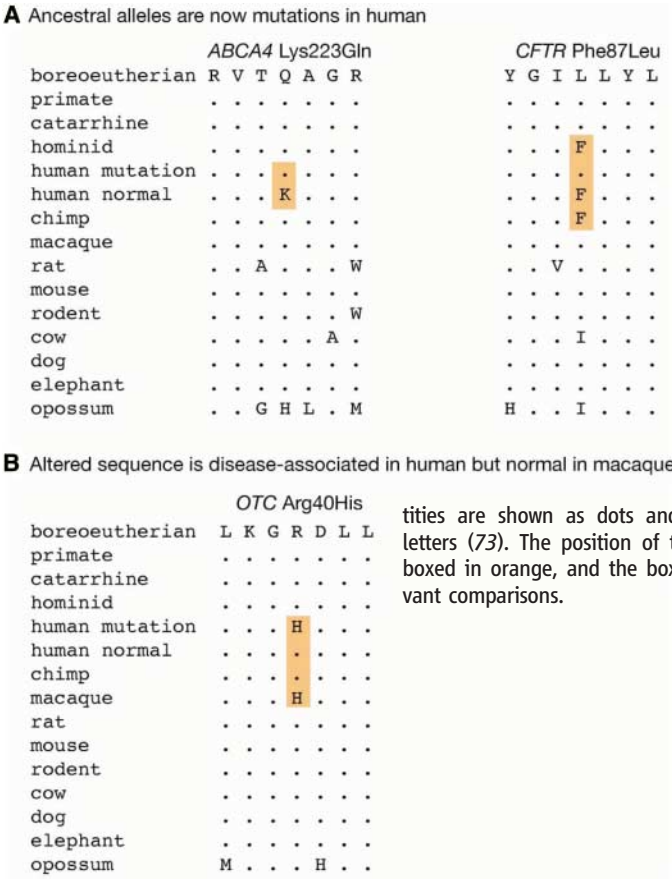

Fig. 9. Ancestral disease mutations. Examples of human mutations that match the sequences of chimp and/or macaque are shown. (**A**) Genes in which the ancestral allele is now the disease-associated allele in humans. (**B**) An instance in which the mutant allele in humans is the normal allele in macaque. The amino acid sequences predicted for the boreoeutherian ancestor (*65*) are given on the top row of each alignment block. Identities are shown as dots and differences are given as letters (*73*). The position of the mutation in humans is boxed in orange, and the box extends through the relevant comparisons.

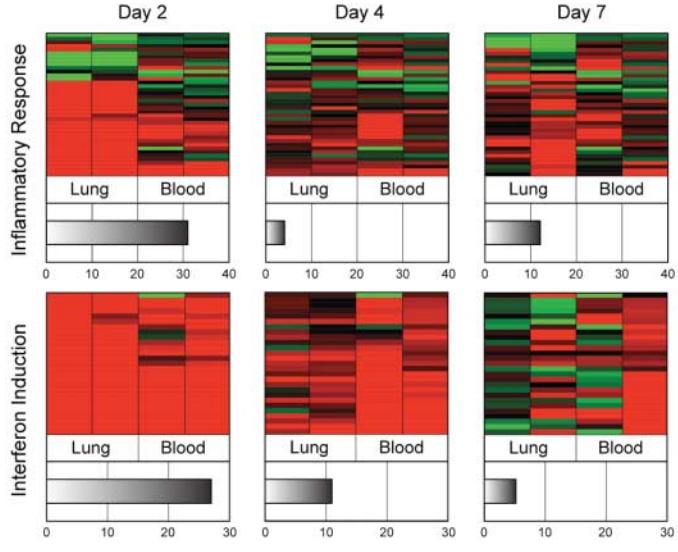

Fig. 10. Application of rhesus-specific microarrays. A microarray based on the rhesus macaque draft genome was used to analyze gene expression in a macaque model of human influenza infection. Gray bars measure an overall response for indicated functional categories, based on corresponding heat maps, and reveal a significant rebound in expression at day 7 for genes associated with the inflammatory response, when compared to interferon induction. Red, increased expression; green, reduced expression. Details are given in (*7*, *70*).

References and Notes

1. C. Groves, *Primate Taxonomy* (Smithsonian Institution Press, Washington, DC, 2001).
2. S. Kumar, S. B. Hedges, *Nature* **392**, 917 (1998).
3. R. A. Gibbs *et al.*, *Nature* **428**, 493 (2004).
4. F. C. Chen, W. H. Li, *Am. J. Hum. Genet.* **68**, 444 (2001).
5. N. Patterson, D. J. Richter, S. Gnerre, E. S. Lander, D. Reich, *Nature* **441**, 1103 (2006).
6. L. M. Zahn, B. R. Jasny, Eds., poster from the special issue on the Macaque Genome, *Science* **316**, following p. 246 (13 April 2007); interactive online (www.sciencemag.org/sciext/macaqueposter/).
7. Materials, methods, and additional discussion are available on *Science* Online.
8. J. Rogers *et al.*, *Genomics* **87**, 30 (2006).
9. W. J. Murphy *et al.*, *Genomics* **86**, 383 (2005).
10. B. Dutrillaux, *Hum. Genet.* **48**, 251 (1979).
11. J. Wienberg, R. Stanyon, A. Jauch, T. Cremer, *Chromosoma* **101**, 265 (1992).
12. T. J. Hubbard *et al.*, *Nucleic Acids Res.* **35**, D610 (2007).
13. M. J. van Baren, M. R. Brent, *Genome Res.* **16**, 678 (2006).
14. International Human Genome Sequencing Consortium, *Nature* **409**, 860 (2001).
15. K. Han *et al.*, *Science* **316**, 238 (2007).
16. J. J. Yunis, O. Prakash, *Science* **215**, 1525 (1982).
17. H. Kehrer-Sawatzki, D. N. Cooper, *Hum. Genet.* **120**, 759 (2007).
18. J. Ma *et al.*, *Genome Res.* **16**, 1557 (2006).
19. R. A. Harris *et al.*, *Science* **316**, 235 (2007).
20. K. J. Kalafus, A. R. Jackson, A. Milosavljevic, *Genome Res.* **14**, 672 (2004).
21. A. Milosavljevic *et al.*, *Genome Res.* **15**, 292 (2005).
22. N. H. Barton, *Curr. Biol.* **16**, R647 (2006).
23. J. A. Bailey, E. E. Eichler, *Nat. Rev. Genet.* **7**, 552 (2006).
24. J. A. Bailey, A. M. Yavor, H. F. Massa, B. J. Trask, E. E. Eichler, *Genome Res.* **11**, 1005 (2001).
25. S. Schwartz *et al.*, *Genome Res.* **13**, 103 (2003).
26. J. A. Bailey *et al.*, *Science* **297**, 1003 (2002).
27. Z. Cheng *et al.*, *Nature* **437**, 88 (2005).
28. X. She *et al.*, *Genome Res.* **16**, 576 (2006).
29. The Chimpanzee Sequencing and Analysis Consortium, *Nature* **437**, 69 (2005).
30. E. Gonzalez *et al.*, *Science* **307**, 1434 (2005).
31. M. W. Hahn, B. T. De, J. E. Stajich, C. Nguyen, N. Cristianini, *Genome Res.* **15**, 1153 (2005).
32. A. Fortna *et al.*, *PLoS Biol.* **2**, E207 (2004).
33. M. Lynch, J. S. Conery, *J. Struct. Funct. Genomics* **3**, 35 (2003).
34. W. J. Kent, *Genome Res.* **12**, 656 (2002).
35. J. A. Bailey, E. E. Eichler, *Cold Spring Harb. Symp. Quant. Biol.* **68**, 115 (2003).
36. M. Lynch, M. O'Hely, B. Walsh, A. Force, *Genetics* **159**, 1789 (2001).
37. A. J. Iafrate *et al.*, *Nat. Genet.* **36**, 949 (2004).

38. J. Sebat *et al.*, *Science* **305**, 525 (2004).
39. Z. Birtle, L. Goodstadt, C. Ponting, *BMC Genomics* **6**, 120 (2005).
40. R. Daza-Vamenta, G. Glusman, L. Rowen, B. Guthrie, D. E. Geraghty, *Genome Res.* **14**, 1501 (2004).
41. K. D. Pruitt, T. Tatusova, D. R. Maglott, *Nucleic Acids Res.* **35**, D61 (2007).
42. J. L. Ashurst *et al.*, *Nucleic Acids Res.* **33**, D459 (2005).
43. F. Hsu *et al.*, *Bioinformatics* **22**, 1036 (2006).
44. M. Blanchette *et al.*, *Genome Res.* **14**, 708 (2004).
45. W. J. Kent, R. Baertsch, A. Hinrichs, W. Miller, D. Haussler, *Proc. Natl. Acad. Sci. U.S.A.* **100**, 11484 (2003).
46. N. Goldman, Z. Yang, *Mol. Biol. Evol.* **11**, 725 (1994).
47. R. Nielsen, Z. Yang, *Genetics* **148**, 929 (1998).
48. Z. Yang, *Comput. Appl. Biosci.* **13**, 555 (1997).
49. D. Graur, L. Li, *Fundamentals of Molecular Evolution* (Sinauer Associates, Tel Aviv, Israel, ed. 2, 2000).
50. M. Ashburner *et al.*, *Nat. Genet.* **25**, 25 (2000).
51. Z. Yang, R. Nielsen, *Mol. Biol. Evol.* **19**, 908 (2002).
52. R. Nielsen *et al.*, *PLoS Biol.* **3**, e170 (2005).
53. A. G. Clark *et al.*, *Science* **302**, 1960 (2003).
54. H. Mi *et al.*, *Nucleic Acids Res.* **33**, D284 (2005).
55. A. J. Betancourt, Y. Kim, H. A. Orr, *Genetics* **168**, 2261 (2004).
56. R. D. Hernandez *et al.*, *Science* **316**, XXXX (2007).
57. G. A. Watterson, *Theor. Popul. Biol.* **7**, 256 (1975).
58. J. Taylor, S. Tyekucheva, M. Zody, F. Chiaromonte, K. D. Makova, *Mol. Biol. Evol.* **23**, 565 (2006).
59. K. D. Makova, S. Yang, F. Chiaromonte, *Genome Res.* **14**, 567 (2004).
60. A. Bartosch-Harlid, S. Berlin, N. G. Smith, A. P. Moller, H. Ellegren, *Evolution Int. J. Org. Evolution* **57**, 2398 (2003).
61. M. P. Goetting-Minesky, K. D. Makova, *J. Mol. Evol.* **63**, 537 (2006).
62. J. V. Neel, *Am. J. Hum. Genet.* **14**, 353 (1962).
63. P. D. Stenson *et al.*, *Hum. Mutat.* **21**, 577 (2003).
64. *Human Gene Mutation Database* at the Institute of Medical Genetics in Cardiff (www.hgmd.org; October 2006 release).
65. M. Blanchette, E. D. Green, W. Miller, D. Haussler, *Genome Res.* **14**, 2412 (2004).
66. L. Gao, J. Zhang, *Trends Genet.* **19**, 678 (2003).
67. S. Umeda *et al.*, *Invest. Ophthalmol. Vis. Sci.* **46**, 683 (2005).
68. J. D. Chismar *et al.*, *Biotechniques* **33**, 516 (2002).
69. J. C. Wallace *et al.*, *BMC Genomics* **8**, 28 (2007).
70. T. Baas *et al.*, *J. Virol.* **80**, 10813 (2006).
71. E. R. Spindel *et al.*, *BMC Genomics* **6**, 160 (2005).
72. M. Ventura *et al.*, *Science* **316**, 243 (2007).

73. Single-letter abbreviations for the amino acid residues are as follows: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; and Y, Tyr.

74. This project was supported by National Human Genome Research Institute grants to the Baylor College of Medicine Human Genome Sequencing Center (U54 HG003273), Washington University Genome Sequencing Center (U54 HG003079), and the J. Craig Venter Institute (U54 HG003068). We thank members of the NHGRI staff for their ongoing efforts: A. Felsenfeld, J. Peterson, M. Guyer, and W. Lu. Additional acknowledgments of support are available online (7). We thank the California National Primate Research Center (NPRC), Oregon NPRC, Southwest NPRC, and Yerkes NPRC for contributing biological samples used in this study.

Rhesus Macaque Genome Sequencing and Analysis Consortium

Project Leader: Richard A. Gibbs[1,2]

White paper: Jeffrey Rogers,[3] Michael G. Katze,[4] Roger Bumgarner,[4] Richard A. Gibbs,[1,2] George M. Weinstock[1,2]

Principal investigators: Richard A. Gibbs,[1,2] Elaine R. Mardis,[5] Karin A. Remington,[6] Robert L. Strausberg,[6] J. Craig Venter,[6] George M. Weinstock,[1,2] Richard K. Wilson[5]

Analysis leaders: Mark A. Batzer,[7] Carlos D. Bustamante,[8] Evan E. Eichler,[9] Richard A. Gibbs,[1,2] Matthew W. Hahn,[10] Ross C. Hardison,[11] Kateryna D. Makova,[11] Webb Miller,[11] Aleksandar Milosavljevic,[1,2] Robert E. Palermo,[4] Adam Siepel,[8] James M. Sikela,[12] George M. Weinstock[1,2]

Genome sequencing: Tony Attaway,[1,2] Stephanie Bell,[1,2] Kelly E. Bernard,[5] Christian J. Buhay,[1,2] Mimi N. Chandrabose,[1,2] Marvin Dao,[1,2] Clay Davis,[1,2] Kimberly D. Delehaunty,[5] Yan Ding,[1,2] Huyen H. Dinh,[1,2] Shannon Dugan-Rocha,[1,2] Lucinda A. Fulton,[5] Ramatu Ayiesha Gabisi,[1,2] Toni T. Garner,[1,2] Richard A. Gibbs,[1,2] Jennifer Godfrey,[5] Alicia C. Hawes,[1,2] Judith Hernandez,[1,2] Sandra Hines,[1,2] Michael Holder,[1,2] Jennifer Hume,[1,2] Shalini N. Jhangiani,[1,2] Vandita Joshi,[1,2] Ziad Mohid Khan,[1,2] Ewen F. Kirkness[6] (leader), Andrew Cree,[1,2] R. Gerald Fowler,[1,2] Sandra Lee,[1,2] Lora R. Lewis,[1,2] Zhangwan Li,[1,2] Yih-shin Liu,[1,2] Stephanie M. Moore,[1,2] Donna Muzny[1,2] (leader), Lynne V. Nazareth[1,2] (leader), Dinh Ngoc Ngo,[1,2] Geoffrey O. Okwuonu,[1,2] Grace Pai,[6] David Parker,[1,2] Heidie A. Paul,[1,2] Cynthia Pfannkoch,[6] Craig S. Pohl,[5] Yu-Hui Rogers,[6] San Juana Ruiz,[1,2] Aniko Sabo,[1,2] Jireh Santibanez,[1,2] Brian W. Schneider,[1,2] Scott M. Smith,[5] Erica Sodergren,[1,2] Amanda F. Svatek,[1,2] Teresa R. Utterback,[1,2] Selina Vattathil,[1,2] Wesley Warren[5] (leader), George M. Weinstock,[1,2] Courtney Sherell White[1,2]

Genome assembly: Asif T. Chinwalla[5] (leader), Yucheng Feng,[5] Aaron L. Halpern,[6] LaDeana W. Hillier,[5] Xiaoqiu Huang,[13] Ewen F. Kirkness,[6] Pat Minx,[5] Joanne O. Nelson,[5] Kymberlie H. Pepin,[5] Xiang Qin,[1,2] Karin A. Remington,[6] Granger G. Sutton[6] (leader), Eli Venter,[6] Brian P. Walenz,[6] John W. Wallis,[5] George M. Weinstock,[1,2] Kim C. Worley[1,2] (leader), Shiaw-Pyng Yang[5]

Mapping: LaDeana W. Hillier,[5] Steven M. Jones,[14] Marco A. Marra,[14] Mariano Rocchi,[15] Jacqueline E. Schein,[14] John W. Wallis[5]

Sequence finishing: Christian J. Buhay,[1,2] Yan Ding,[1,2] Shannon Dugan-Rocha,[1,2] Alicia C. Hawes,[1,2] Judith Hernandez,[1,2] Michael Holder,[1,2] Jennifer Hume,[1,2] Ziad Mohid Khan,[1,2] Zhangwan Li,[1,2] Dinh Ngoc Ngo,[1,2] Aniko Sabo[1,2]

Assembly comparison: Robert Baertsch,[16] Asif T. Chinwalla,[5] Laura Clarke,[17] Miklós Csürös,[18] Jarret Glasscock,[5] R. Alan Harris,[1,2] Paul Havlak,[1,2] LaDeana W. Hillier,[5] Andrew R. Jackson,[1,2] Huaiyang Jiang,[1,2] Yue Liu,[1,2] David N. Messina,[5] Xiang Qin,[1,2] Yufeng Shen,[1,2] Henry Xing-Zhi Song,[1,2] George M. Weinstock[1,2] (leader), Kim C. Worley[1,2] (leader), Todd Wylie,[5] Lan Zhang[1,2]

Gene prediction: Ewan Birney,[17] Laura Clarke[17]

Repetitive elements: Mark A. Batzer[7] (leader), Kyudong Han,[7] Miriam K. Konkel,[7] Jungnam Lee,[7] Webb Miller,[11] Arian F. A. Smit,[19] Brygg Ullmer,[20] Hui Wang,[7] Jinchuan Xing[7,21]

Ancestral genomes and segmental duplications: Richard Burhans,[11] Ze Cheng,[9] Miklós Csürös,[18] Evan E. Eichler,[9] R. Alan Harris,[1,2] Andrew R. Jackson,[1,2] John E. Karro,[11] Jian Ma,[22] Aleksandar Milosavljevic[1,2] (leader), Brian Raney,[22] Xinwei She[9]

Gene duplication/gene families: Michael J. Cox,[12] Jeffery P. Demuth,[10] Laura J. Dumas,[12] Matthew W. Hahn[10] (leader), Sang-Gook Han,[10] Janet Hopkins,[12] Anis Karimpour-Fard,[23] Young H. Kim,[24] Jonathan R. Pollack,[24] James M. Sikela[12] (leader)

PRAME Gene Family Analysis: Webb Miller[11] (leader), Donna Muzny,[1,2] Brian Raney,[22] Aniko Sabo,[1,2] Adam Siepel,[8] Tomas Vinar[8]

Orthologous genes: Charles Addo-Quaye,[11] Jeremiah Degenhardt,[8] Alexandra Denby,[8] Melissa J. Hubisz,[25] Amit Indap,[8] Carolin Kosiol,[8] Bruce T. Lahn,[25,26] Heather A. Lawson,[11] Alison Marklein,[8] Rasmus Nielsen,[27] Adam Siepel[8] (leader), Eric J. Vallender,[25,26] Tomas Vinar[8]

Population genetics: Mark A. Batzer[7] (leader), Carlos D. Bustamante[8] (leader), Andrew G. Clark,[28] Jeremiah Degenhardt,[8] Betsy Ferguson,[29] Richard A. Gibbs,[1,2] Matthew W. Hahn,[10] Kyudong Han,[7] Ryan D. Hernandez,[8] Kashif Hirani,[1,2] Amit Indap,[8] Hildegard Kehrer-Sawatzki,[30] Jessica Kolb,[30] Miriam K. Konkel,[7] Jungnam Lee,[7] Lynne V. Nazareth,[1,2] Shobha Patil,[1,2] Ling-Ling Pu,[1,2] Jeffrey Rogers,[3] Yanru Ren,[1,2] David Glenn Smith,[3] Brygg Ullmer,[20] Hui Wang,[7] David A. Wheeler,[1,2] Jinchuan Xing[7,21]

Sex chromosome evolution: Kateryna D. Makova,[11] Ian Schenck[11]

Human disease orthologs: Edward V. Ball,[31] Rui Chen,[1,2] David N. Cooper,[31] Belinda Giardine,[11] Richard A. Gibbs,[1,2] Ross C. Hardison[11] (leader), Fan Hsu,[22] W. James Kent,[22] Arthur Lesk,[11] Webb Miller,[11] David L. Nelson,[2] William E. O'Brien,[2] Kay Prüfer,[32] Peter D. Stenson[31]

Additional biological impact of genomic sequence: Michael G. Katze,[4] Robert E. Palermo[4] (leader), James C. Wallace[4]

Macaque sample collection: Hui Ke,[33] Xiao-Ming Liu,[34] Peng Wang,[33] Andy Peng Xiang,[33] Fan Yang[33]

Genome browser: Robert Baertsch,[16] Galt P. Barber,[22] David Haussler[35,16] (leader), Donna Karolchik,[22] Andy D. Kern,[22] Robert M. Kuhn,[22] Kayla E. Smith,[22] Ann S. Zwieg[22]

[1]Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX 77030, USA. [2]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX 77030, USA. [3]Department of Genetics, Southwest Foundation for Biomedical Research, San Antonio, TX 78227, USA. [4]Department of Microbiology, University of Washington, Seattle, WA 98195, USA. [5]Genome Sequencing Center, Washington University, St. Louis, MO 63108, USA. [6]J. Craig Venter Institute, 9704 Medical Center Drive, Rockville, MD 20850, USA. [7]Department of Biological Sciences, Biological Computation and Visualization Center, Center for BioModular Multi-scale Systems, Louisiana State University, Baton Rouge, LA 70803, USA. [8]Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, NY 14853, USA. [9]Department of Genome Sciences, University of Washington, Seattle, WA 98195, USA. [10]Department of Biology and School of Informatics, Indiana University, Bloomington, IN 47405, USA. [11]Center for Comparative Genomics and Bioinformatics, Pennsylvania State University, University Park, PA 16802, USA. [12]Human Medical Genetics and Neuroscience Programs, Department of Pharmacology, University of Colorado at Denver and Health Sciences Center, Aurora, CO 80045, USA. [13]Department of Computer Science, Iowa State University, Ames, IA 50011, USA. [14]Genome Sciences Centre, British Columbia Cancer Agency, 570 West 7th Avenue, Vancouver, BC, Canada. [15]Department of Genetics and Microbiology, University of Bari, Bari, Italy. [16]Department of Bioinformatics, University of California Santa Cruz, Santa Cruz, CA 95060, USA. [17]The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire, CB10 1SA, UK. [18]Département d'Informatique et de Recherche Opérationnelle, Université de Montréal, Montréal, QC H3C 3J7, Canada. [19]Institute for Systems Biology, 1441 North 34th Street, Seattle, WA 98103–8904, USA. [20]Center for Computation and Technology, Department of Computer Sciences, Louisiana State University, Baton Rouge, LA 70803, USA. [21]Eccles Institute of Human Genetics, University of Utah, Salt Lake City, UT 84112, USA. [22]Center for Biomolecular Science and Engineering, University of California Santa Cruz, Santa Cruz, CA 95064, USA. [23]Department of Preventative Medicine and Biometrics, University of Colorado at Denver and Health Sciences Center, Aurora, CO 80045, USA. [24]Department of Pathology, Stanford University, Stanford, CA 94305, USA. [25]Department of Human Genetics, University of Chicago, Chicago, IL 60637, USA. [26]Howard Hughes Medical Institute, Department of Human Genetics, University of Chicago, Chicago, IL 60637, USA. [27]Institute of Biology, University of Copenhagen, Copenhagen DK-1017, Denmark. [28]Department of Molecular Biology and Genetics, Cornell University, Ithaca, NY 14853, USA. [29]Genetics Research and Informatics Program, Oregon National Primate Research Center, Beaverton, OR 97006, USA. [30]Institute of Human Genetics, University of Ulm, Ulm, 89081, Germany. [31]Institute of Medical Genetics, Cardiff University, Heath Park, Cardiff, CF14 4XN, UK. [32]Department Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Leipzig, 04103, Germany. [33]Centre for Stem Cell Biology and Tissue Engineering, Sun Yat-sen University, Guangzhou 510080, China. [34]South-China Primate Research and Development Center, Guangzhou 510080, China. [35]Howard Hughes Medical Institute, Santa Cruz, CA 95060, USA.

Supporting Online Material

www.sciencemag.org/cgi/content/full/316/5822/222/DC1
Materials and Methods
SOM Text
Figs. S1.1 to S7.2
Tables S1.1 to S9.4
References and Notes

22 December 2006; accepted 16 March 2007
10.1126/science.1139247