Sequencing Primate Genomes: What Have We Learned?

Tomas Marques-Bonet,1,2 Oliver A. Ryder,3 and Evan E. Eichler1

1Department of Genome Sciences, University of Washington and the Howard Hughes Medical Institute, Seattle, Washington 98105
2Institut de Biologia Evolutiva (UPF/CSIC), 08003 Barcelona, Catalonia, Spain
3San Diego Zoo’s Institute for Conservation Research, Escondido, CA 92027-70001; email: tmarques@u.washington.edu; oryder@sandiegozoo.org; eee@gs.washington.edu

Key Words
genome, sequencing, variation, gene comparison, speciation, diversity

Abstract

We summarize the progress in whole-genome sequencing and analyses of primate genomes. These emerging genome datasets have broadened our understanding of primate genome evolution revealing unexpected and complex patterns of evolutionary change. This includes the characterization of genome structural variation, episodic changes in the repeat landscape, differences in gene expression, new models regarding speciation, and the ephemeral nature of the recombination landscape. The functional characterization of genomic differences important in primate speciation and adaptation remains a significant challenge. Limited access to biological materials, the lack of detailed phenotypic data and the endangered status of many critical primate species have significantly attenuated research into the genetic basis of primate evolution. Next-generation sequencing technologies promise to greatly expand the number of available primate genome sequences; however, such draft genome sequences will likely miss critical genetic differences within complex genomic regions unless dedicated efforts are put forward to understand the full spectrum of genetic variation.
INTRODUCTION

As new primate genomes become sequenced and are compared within the context of the primate phylogeny (Figure 1), scientists are provided with an unprecedented opportunity to reconstruct the evolutionary history of every basepair of the human genome [see (44, 50, 60) for reviews]. This review focuses on how multiple primate genomes have provided a framework to understand the mode and tempo of primate genome evolution, including an understanding of our gene repertoire and chromosomal organization. We discuss how this information has provided a new understanding of the extent of primate molecular adaptations and stimulated new hypotheses regarding selection in primate genomes. We focus on mounting evidence that gene regulation and expression as opposed to amino acid changes have driven primate adaptations to new environments. Comparisons among primate genomes have begun to reveal more basepairs affected by genome structural variation (IN/DELS, duplications, deletions, insertions, and bursts of retrotransposition events) than single nucleotide changes (16, 28, 165). Such changes occur preferentially within complex regions of the primate genome wherein, paradoxically, newly minted genes and gene families of unknown function have been discovered (77, 79, 124). Comparison of primate genome sequences has provided new insights into the molecular mechanisms that have contributed to chromosomal evolution within our species (51, 86), their influences on recombination (128, 162), and their relationship to natural selection within primates (18). Finally, we discuss the future of primate genomics and the need to strike the right balance between the number of genomes versus the quality of the sequence assemblies. We conclude with a discussion of the challenges of performing functional studies and obtaining material from other primate species that are endangered.

PRIMATE GENOME SEQUENCING: RATIONALE AND PROGRESS

With the initial draft sequence assembly of the human genome in 2001 followed by the mouse genome in 2002 (94, 160), other

Figure 1

Primate genome sequencing status. Primate whole-genome sequencing projects (10/25/08) with respect to the generally accepted primate phylogeny and an outgroup (mouse). Divergence times are estimated based on millions of years according to Goodman et al. 1998.
primate species became high priorities for whole-genome sequencing. A series of proposals and white papers (41) identified less than a dozen key primate species as sequencing targets (Figure 1). From the beginning, species choice was based on two different (sometimes antagonistic and sometimes complementary) rationales: biomedical relevance and the species position within the primate phylogeny with respect to human. The first considered primate species as models of disease and biomedical research. The baboon and macaque were justified because of their use in studying the genetic basis of numerous diseases (e.g., diabetes, cardiovascular disease, hypertension) and as models of transplantation, reproduction, infection, immunity, and pharmacology. Similarly, the squirrel monkey is used as a model for understanding primate neurobiology and infectious disease (in particular, malarial research). Since many of the genes and gene families underlying biological processes such as immunity, reproduction, and drug detoxification change rapidly over short periods of evolutionary time, complete genome sequences would reveal these idiosyncratic aspects and thereby enhance the utility of these models for understanding disease. More importantly, high-quality genomes provide an immediate framework upon which to map genetic traits associated with diseases and to develop transgenic models of disease (164).

The second major rationale was to improve annotation of the human genome. Based on our current understanding of the primate phylogeny (57, 142), there are seven key points of evolutionary transition (Figure 1) with respect to human. Sequencing index species from each of these phylogenetic nodes (i.e., chimpanzee, orangutan, gibbon, etc.) has two advantages. Comparative sequencing will determine the ancestral and derived states of every single basepair within the human genome sequence, thereby assigning mutational events to different time points during human evolution. Second, multiple primate genome comparisons will highlight differences of functional import (bursts of regulatory changes or amino-acid changes within specific lineages) (15). We should emphasize that these fundamental motivations contrast with that of other genome projects. Most mammalian genome projects, for example, aim to identify conserved regulatory elements, exonic sequence and other regions of potential functional significance (102, 152). This focus on the differences demands a unique quality of product to eliminate potential artifacts. High-quality comparative sequencing of primate genomes is necessary in order to provide a balanced view of genome variation (including regions of structural variation, segmental duplications, lineage-specific events and chromosomal variation) and to identify true genetic differences (as opposed to sequencing or assembly artifacts) between the species.

In addition to continued refinement and sequencing of additional primate genomes, several primate species are currently in progress for whole-genome sequencing (Table 1). In most cases, the representative individuals are females to provide adequate coverage of the X chromosome, albeit with the concomitant loss of Y chromosome genetic information. Two working draft assemblies (chimpanzee and macaque) have been published (32, 51) and three additional genome assemblies (orangutan, gibbon, and common marmoset) have been generated and are being analyzed. Although none of the genomes will currently be sequenced to the same standard as the human genome, the need for higher-quality draft assemblies has been recognized owing in part to the frustration of distinguishing artifacts from true sequence differences in the low-coverage three-fold chimpanzee draft assembly (32).

These initial analyses stressed the importance of independent assemblies of nonhuman primate genomes instead of simply mapping reads against the human reference sequence. As a result, typical working draft assemblies now target six- to eight-fold coverage in capillary whole-genome shotgun sequence reads. In five cases, these whole-genome shotgun sequence assemblies will be further supplemented by the sequencing of large-insert BAC clones mapping to structurally complex or biomedically
### Table 1  Primate genome sequencing projects

<table>
<thead>
<tr>
<th>Species</th>
<th>Scientific name</th>
<th>Individual</th>
<th>Sequencing center(s)</th>
<th>Target</th>
<th>Assembly status</th>
<th>WGS</th>
<th>Query</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td><em>Homo sapiens</em></td>
<td>RPCI-11 (M)</td>
<td>Consortium</td>
<td>Finishing genome</td>
<td>Build37</td>
<td>78,410,449</td>
<td>TRACE_TYPE_CODE = “WGS” and SPECIES_CODE = “HOMO SAPIENS”</td>
</tr>
<tr>
<td>Chimpanzee</td>
<td><em>Pan troglodytes</em></td>
<td>Clint #C0471 (M)</td>
<td>WUGSC and BI/MIT</td>
<td>Draft2</td>
<td>PanTro2</td>
<td>31,339,291</td>
<td>TRACE_TYPE_CODE = “WGS” and SPECIES_CODE = “PAN TROGLODYTES”</td>
</tr>
<tr>
<td>Gorilla</td>
<td><em>Gorilla gorilla</em></td>
<td>Kamillah (F)</td>
<td>Sanger</td>
<td>Hybrid</td>
<td>Not released</td>
<td>8,302,852</td>
<td>TRACE_TYPE_CODE = “WGS” and (SPECIES_CODE = “GORILLA GORILLA” or SPECIES_CODE = “GORILLA GORILLA GORILLA”)</td>
</tr>
<tr>
<td>Orangutan</td>
<td><em>Pongo pygmaeus</em></td>
<td>Susie ISIS#71 (F)</td>
<td>BCM and WUGSC</td>
<td>Draft2</td>
<td>Ponab2</td>
<td>14,726,737</td>
<td>TRACE_TYPE_CODE = “WGS” and (SPECIES_CODE = “PONGO ABELII” or SPECIES_CODE = “PONGO PYGMAEUS” or SPECIES_CODE = “PONGO PYGMAEUS ABELII”)</td>
</tr>
<tr>
<td>Gibbon</td>
<td><em>Nomascus leucogenys</em></td>
<td>Asia #0098 (F)</td>
<td>BCM and WUGSC</td>
<td>Draft2</td>
<td>In progress</td>
<td>28,885,424</td>
<td>TRACE_TYPE_CODE = “WGS” and SPECIES_CODE = “NOMASCUS LEUCOGENYS” or SPECIES_CODE = “NOMASCUS LEUCOGENYS LEUCOGENYS”</td>
</tr>
<tr>
<td>Rhesus macaque</td>
<td><em>Macaca mulatta</em></td>
<td>#25311 (F)</td>
<td>BCM, JCVI, WUGSC</td>
<td>Draft2</td>
<td>Rhemac2</td>
<td>23,017,908</td>
<td>TRACE_TYPE_CODE = “WGS” and SPECIES_CODE = “MACACA MULATTA”</td>
</tr>
<tr>
<td>Cynomolgus macaque</td>
<td><em>Macaca fascicularis</em></td>
<td>ND</td>
<td>WUGSC</td>
<td>Draft1</td>
<td>Not started</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Species</td>
<td>Trace Type Code</td>
<td>Species Code</td>
<td>Center</td>
<td>Status</td>
<td>Coverage</td>
<td>Notes</td>
<td></td>
</tr>
<tr>
<td>--------------------------------</td>
<td>-----------------</td>
<td>------------------------</td>
<td>--------------</td>
<td>-----------------</td>
<td>----------</td>
<td>------------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>Baboon P. hamadryas</td>
<td>Draft1</td>
<td>BCM</td>
<td>ND</td>
<td>In progress</td>
<td>12,717,107</td>
<td>TRACE_TYPE_CODE = “WGS” and SPECIES_CODE = “PAPIO HAMADRYAS”</td>
<td></td>
</tr>
<tr>
<td>Vervet C. aethiops</td>
<td>Draft1</td>
<td>WUGSC</td>
<td>ND</td>
<td>Not started</td>
<td>25,615,089</td>
<td>TRACE_TYPE_CODE = “WGS” and SPECIES_CODE = “CALLITHRIX JACCHUS”</td>
<td></td>
</tr>
<tr>
<td>Marmoset C. jacchus</td>
<td>Draft2</td>
<td>Caljach1</td>
<td>SFPRI17066</td>
<td>Not started</td>
<td>690,374</td>
<td>TRACE_TYPE_CODE = “WGS” and SPECIES_CODE = “SAIMIRI BOLIVIENSIS” or SPECIES_CODE = “SAIMIRI BOLIVIENSIS BOLIVIENSIS”</td>
<td></td>
</tr>
<tr>
<td>Squirrel monkey Saimiri boliviensis*</td>
<td>Draft1</td>
<td>BI/MIT</td>
<td>ND</td>
<td>Not started</td>
<td>13,953,011</td>
<td>TRACE_TYPE_CODE = “WGS” and SPECIES_CODE = “OTOLEMUR GARNETTI”</td>
<td></td>
</tr>
<tr>
<td>Galago/Otolemur garnetti</td>
<td>Draft1 Low coverage 2X</td>
<td>BI/MIT</td>
<td>ND</td>
<td>7,749,934</td>
<td></td>
<td>TRACE_TYPE_CODE = “WGS” and SPECIES_CODE = “MICROCEBUS MURINUS”</td>
<td></td>
</tr>
</tbody>
</table>

1 For each genome project, we identify the specific individual whose genome is being sequenced, the responsible genome sequencing center, and number of available capillary whole-genome shotgun sequences (WGS) in the trace repository. As of 10/1/2008, Draft1 = genomes with 6–8 coverage WGS only; Draft2 = genomes with 6–8X coverage with targeted sequencing of BACs for difficult regions.

2 Gorilla genome was approved by Wellcome Trust 12/2005, 2X capillary (3730) WGS data and 12 X WGS Solexa/Illumina WGS data were generated as of 12/2007; only 2X 3730 has been released as of 10/2008; Goal has shifted from WGS of capillary reads to a hybrid assembly using next-generation technology, NIH approved targeted sequencing of complex regions within BACs.

* Specific squirrel monkey species is not yet definite.
relevant regions of the genome (a procedure known as genome refinement). This is especially crucial with respect to the phylogenetic index species such as chimpanzee, orangutan, gibbon, macaque, and squirrel monkey (indicated as Draft 2 in Table 1) where duplications and structural variation are now thought to account for more genetic variation than single basepair differences. Unfortunately, the gorilla genome assembly is not slated for the same standard. The Sanger Center has opted to generate an experimental hybrid assembly consisting of 12X Solexa sequencing data combined with 2X coverage of whole-genome assembly data. BAC refinement has been proposed but will be carried out as part of an NIH initiative.

FEATURES OF PRIMATE GENOME VARIATION

Although relatively few genome-wide comparisons have been completed to date (32, 51), analyses of the human, macaque, and chimpanzee genomes, as well as targeted resequencing of specific genomic segments, have revealed some important features and general principles of primate genome evolution. The alignment of the majority of genomic sequence from closely related primates is relatively trivial (38, 152) and shows a neutral pattern of single-nucleotide variation consistent with the primate phylogeny (Figure 2), although the rate of single-nucleotide variation has varied by a factor of three-fold within different lineages (42, 96, 143). Notably, the pattern of single-nucleotide variation also varies as a function of chromosome structure and organization (32, 51). Metacentric and acrocentric chromosomes show significant differences in variation, and regions within 10 Mb of telomeres evolve more rapidly perhaps as a consequence of biased gene conversion (Figure 2). On average, 10% of the genomic sequence has proven more elusive in terms of orthologous alignment. This includes segmental duplications, subtelomeric regions, pericentromeric regions, and lineage-specific repeats. Such regions typically cannot be resolved strictly by whole-genome sequence assembly (WGSA), thus larger insert clones (i.e., BAC clones) provide better substrates for resolving these areas.

Comparative sequence data highlight the value of genomic sequence from nonhuman primates to determine the ancestral and derived status of human alleles (27, 81). There have been some surprises. Phylogenetic analysis of resequenced regions among humans and the great apes reveal that as many as 18% of genomic regions are inconsistent with the Homo-Pan clade, and, rather, support a Homo-Gorilla clade (27). This has been taken as evidence of lineage-sorting and/or an ancestral hominid population size greater than five times that of the effective human population size ($n = 10,000$). Another surprise has been the identification of ancestral allelic variants that now occur as disease alleles within the human population, i.e., pyrin and familial Mediterranean fever (51, 136). Such findings suggest that the functional and selective effects of mutations change over time, perhaps as a result of environmental changes or compensatory genetic mutations.

COMPARATIVE GENE ANALYSES

The unique attributes of the primate order, and more specifically that of human, may be explained either as a consequence of key amino-acid changes within the coding sequences of a subset of critical genes or as a result of dramatic changes in how genes are regulated both temporally and spatially. Genome-wide analyses have provided traction for both of these views, although not with equal levels of support. A critical component of these analyses has been the construction of rigorous multiple-sequence alignments among primate genes. Despite the ease at which genomic sequences can be aligned among primate genomes, the number of genes that can be assigned to 1:1:1 orthologous group has changed only slightly with the first two nonhuman primate genomes sequenced. A three-way comparison involving chimp-human-mouse identified 7645 orthologues [by the estimate of (31)] as compared to 10,376 by human-chimp-macaque (51) over
Human SDs

Figure 2

Human and primate sequence similarity. Whole-genome shotgun sequences from a human (blue), chimpanzee (red), and macaque (brown) were aligned against the human reference genome and the nucleotide divergence was computed in 100-kbp windows along (a) chromosome 2 and (b) chromosome 7. Note the increase in sequence divergence within 10 Mb of the telomere.

the total estimated 20,000 genes in the human genome. It follows then that a large fraction of human genes have not been subjected to three-way orthologous comparisons, and the pattern of selection (and its directionality) operating on ~50% of genes has not yet been adequately interrogated. With this caveat, we consider the lessons learned.

Evolution of Coding Sequences

The anthropocentric view of primate evolution, in which humans have acquired multiple idiosyncratic features needed for our success as a species, led to the tacit assumption that differences in a subset of key genes would explain the evolution of our species within the context of the primate order. Most of the first genomewide studies of natural selection have thus been focused on coding sequence and estimates of omega (dN/dS)—the ratio of nonsynonymous versus synonymous substitutions as a metric of the signal of the intensity of such selection. The first three-way primate genome comparisons suggested that human and chimpanzee
genes had similar average omega values but significantly larger values than macaque or rodents (human = 0.169, chimpanzee = 0.175, macaque = 0.124, nonprimate mammals ~0.11) (51). This was explained by a relaxation of purifying selection during hominoid evolution as a consequence of smaller effective population sizes (see below, and see Figure 7).

Earlier studies that used mouse coding sequences as an outgroup (31) reported the first global estimates of genes under positive selection in human and chimpanzee, identifying more than 500 genes under positive selection (especially glycoporphin C, protamines, and a gene family related to nociception). The posterior refinement with the macaque assembly, however, reduced that number to only ~200 genes (32, 51). Looking at pervasive adaptive evolution in genes within the same category led groups to identify gene ontologies enriched for positive selection (51, 115). The common conclusion of most of these studies is an overrepresentation of rapidly evolving genes in biological processes associated with immunity, olfaction, host defense, and reproduction. These enrichments, however, are not particular to human but represent adaptive changes more generally important in evolution of the primate and broadly other mammalian lineages.

A recent report using macaque as an outgroup has suggested that it is the chimpanzee lineage with an excess of positive selection compared to humans (7). Bakewell and colleagues found that even after multiple test correction, the chimpanzee has an excess of genes under positive selection (~40% more than human), although these conclusions should be tempered by the incomplete nature of the chimpanzee genome assembly and the potential for sequence errors to confound such estimates asymmetrically. Nevertheless, this view was also supported by the most comprehensive genomic comparison to date including six genomes (human, chimpanzee, macaque, mouse, rat, and dog). Given the increased power of this six-way comparison, it was possible to identify genes involved in biological processes rapidly diverging within the primate lineage in contrast to rodents. Once again, genes related to sensory perception (e.g., pain receptor NPFF2 and color vision genes OPN1SW), immunity (CCR1), and defense were overrepresented among positively selected genes (93). The previous study also found mRNA transcription, stress response, and protein metabolism as the main biological processes with excess of amino-acid replacements in chimpanzee but not in human. Despite some discrepancies, these and other studies, in general, did not find an enrichment of positively selected genes related to brain development or size.

**Gene Regulation and Expression**

After considerable scrutiny of coding sequences searching for traces of adaptive evolution, it has become increasingly apparent that regulatory differences must be playing a key role in specifying primate adaptations (25). This is not a new concept. Thirty years ago Wilson & King suggested that the phenotypic changes between humans and great apes were too dramatic to be explained by the rather limited degree of protein variation (91). There are many layers of complexity ranging from changes in gene expression, chromatin regulation, and patterns of alternative splicing. For example, a whole-genome comparison of alternative splicing between human and chimpanzee found that as much as 6%–8% of the analyzed exons showed evidence of differential alternative splicing (23).

Numerous microarray studies have focused on gene-expression differences between human and chimpanzee (21, 43, 84, 154) or even between male and female individuals of different primates (129). There has been a strong emphasis on the investigation of genes acting in the brain especially in light of human cognitive specializations. In this regard, however, it should be noted that chimps may, in fact, have some mental skills that are enhanced, such as faster hand-eye signaling of character recognition (69) (see http://www.pri.kyoto-u.ac.jp/ai/video/video_library/project/project.html for a library of impressive videos on the project). Significant gene-expression differences

---

**Marques-Bonet • Ryder • Eichler**
between human and chimpanzee brains have been reported, but, surprisingly, other tissues (such as heart or liver) showed a larger number of differences (see [126] for a review on the use of microarrays on primate gene expression). Since the divergence of human and chimpanzee, ∼100 genes have altered their gene expression patterns. Some studies suggest that gene expression differences in the brain have been biased toward the human lineage as a result of upregulation in humans (21, 43) (Table 2), although other studies contradict these results (52). In contrast, sex differences in brain gene expression have been a conserved feature over the course of primate evolution (129).

One interesting example of a differential gene expression between human and chimpanzee is the two- to nearly sixfold increase in gene expression of THBS2 and THBS4 (thrombospondin 2 and 4) within specific parts of the human brain when compared to chimpanzee or macaque (22). In this study, the authors showed that the levels of THBS2 and THBS4 mRNA are highly differentiated in the forebrain (cortex and caudate) but not in the cerebellum or other analyzed tissues. As these genes are involved in synaptogenesis, there was speculation that these changes in gene expression may be underlying some of the human cognitive specializations.

Recent reports suggest that the hypothetical rapid relative rate of change in gene expression in the human brain has occurred in the context of a decreased rate of coding sequence evolution (compared to the genomic average) (159). Genes whose expression is restricted to the brain show greater conservation than genes expressed in the brain along with other tissues, perhaps as a consequence of stronger selective constraint operating within brain biochemical networks. Despite this constraint, there are some interesting examples of adaptive evolution for brain-expressed genes. The glutamate dehydrogenase (GDH) gene (20) shows both evidence of increased copy as a result of duplication and traces of positive selection within the human lineage. It encodes a protein that plays an important role in neurotransmitter recycling.

There has been renewed interest on the role of posttranscriptional regulation via microRNA (miRNA) and transcription factors. The study of miRNA, for example, has become increasingly relevant because of the central role miRNA plays in regulating the transcriptome of plants and animals (8). A whole-genome comparison of brain miRNA of human and chimpanzee found that more than 10% of the human brain miRNA are primate specific whereas only 1% is human specific. Although such differences might affect a large number of potential gene targets, no bias appears to exist within the human lineage (14). For example, the authors reported 14 human- and 15 chimpanzee-specific miRNA expansions. In contrast, the evolution of protein transcription factors may have not have evolved so uniformly (17). Gilad and colleagues (52) analyzed the expression profile of ∼1000 orthologous genes in human, chimpanzee, orangutan, and macaque by cDNA array hybridization and found transcription factors enriched among genes upregulated in humans. Similarly, as a class, transcription factors, in particular C2H2 zinc finger genes, show the greatest excess of amino-acid replacements within the human lineage (32), providing further support for the notion that gene expression changes have been pivotal during human evolution. These data suggest a concerted effect of positive selection and gene expression on transcription factors in altering gene expression.

Concomitant with positive selection and changes in expression of transcription factors, it follows that cis-regulatory regions (promoters, enhancers, etc.) might similarly show evidence of accelerated evolution. Using the background substitution rate of intronic regions, Haygood and colleagues developed a method to detect an excess of single-basepair changes within human promoters (compared to chimpanzee and macaque). Several genes were detected by this method, including an apparent excess of genes related to neural development and nutrition (63). Another approach has been to study highly conserved noncoding DNA that shows a burst
Table 2  List of genes upregulated in humans

<table>
<thead>
<tr>
<th>#ID</th>
<th>Gene symbol</th>
<th>RefSeq ID</th>
<th>Gene description</th>
<th>Chr (Build 35)</th>
<th>Start</th>
<th>End</th>
<th>Direction</th>
<th>Tissue</th>
<th>Source</th>
<th>dN/dS Chimp</th>
</tr>
</thead>
<tbody>
<tr>
<td>BAT8</td>
<td>NM_025256</td>
<td>LA-B associated transcript 8 BAT8 isoform b</td>
<td>chr6</td>
<td>31,955,518</td>
<td>31,973,443</td>
<td>Up</td>
<td>Liver</td>
<td>Gilad et al. 2006</td>
<td>0.0862</td>
<td></td>
</tr>
<tr>
<td>BTAF1</td>
<td>NM_003972</td>
<td>BTAF1 RNA polymerase II, B-TFIID transcription</td>
<td>chr10</td>
<td>93,673,715</td>
<td>93,780,062</td>
<td>Up</td>
<td>Liver</td>
<td>Gilad et al. 2006</td>
<td>0.0000</td>
<td></td>
</tr>
<tr>
<td>C21orf33</td>
<td>NM_004649</td>
<td>es1 protein isoform 1a precursor</td>
<td>chr21</td>
<td>44,382,191</td>
<td>44,384,261</td>
<td>Up</td>
<td>Brain</td>
<td>Enard et al. and Caceres et al.</td>
<td>0.1464</td>
<td></td>
</tr>
<tr>
<td>CA2</td>
<td>NM_000067</td>
<td>Carbonic anhydrase II</td>
<td>chr8</td>
<td>86,578,696</td>
<td>86,580,973</td>
<td>Up</td>
<td>Liver</td>
<td>Gilad et al. 2006</td>
<td>0.1059</td>
<td></td>
</tr>
<tr>
<td>COL6A1</td>
<td>NM_001848</td>
<td>Collagen, type VI, alpha 1 precursor</td>
<td>chr21</td>
<td>46,226,090</td>
<td>46,249,391</td>
<td>Up</td>
<td>Brain</td>
<td>Enard et al. and Caceres et al.</td>
<td>0.1162</td>
<td></td>
</tr>
<tr>
<td>CROC4</td>
<td>NM_006365</td>
<td>Transcriptional activator of the c-fos promoter</td>
<td>chr1</td>
<td>153,187,128</td>
<td>153,212,257</td>
<td>Up</td>
<td>Brain</td>
<td>Enard et al. and Caceres et al.</td>
<td>0.9643</td>
<td></td>
</tr>
<tr>
<td>CTCF</td>
<td>NM_006565</td>
<td>CCCTC-binding factor</td>
<td>chr16</td>
<td>66,153,964</td>
<td>66,230,589</td>
<td>Up</td>
<td>Liver</td>
<td>Gilad et al. 2006</td>
<td>0.0000</td>
<td></td>
</tr>
<tr>
<td>DUSP6</td>
<td>NM_022652</td>
<td>Dual specificity phosphatase 6 isoform b</td>
<td>chr12</td>
<td>88,265,972</td>
<td>88,270,394</td>
<td>Up</td>
<td>Liver</td>
<td>Gilad et al. 2006</td>
<td>0.0683</td>
<td></td>
</tr>
<tr>
<td>ENTPD6</td>
<td>NM_001247</td>
<td>Ectomesodermal triphosphatase diphosphohydrolase 6</td>
<td>chr20</td>
<td>25,157,785</td>
<td>25,160,596</td>
<td>Up</td>
<td>Brain</td>
<td>Enard et al. and Caceres et al.</td>
<td>0.4350</td>
<td></td>
</tr>
<tr>
<td>FCN3</td>
<td>NM_173452</td>
<td>Ficolin 3 isoform 2 precursor</td>
<td>chr1</td>
<td>27,568,218</td>
<td>27,573,883</td>
<td>Up</td>
<td>Liver</td>
<td>Gilad et al. 2006</td>
<td>#N/A</td>
<td></td>
</tr>
<tr>
<td>FLJ10326</td>
<td>#N/A</td>
<td>#N/A</td>
<td>chr1</td>
<td>216,656,117</td>
<td>216,709,777</td>
<td>Up</td>
<td>Liver</td>
<td>Gilad et al. 2006</td>
<td>0.0826</td>
<td></td>
</tr>
<tr>
<td>FOXO1A</td>
<td>NM_000015</td>
<td>Forkhead box O1</td>
<td>chr13</td>
<td>40,027,817</td>
<td>40,138,734</td>
<td>Up</td>
<td>Liver</td>
<td>Gilad et al. 2006</td>
<td>0.0000</td>
<td></td>
</tr>
<tr>
<td>GM2A</td>
<td>NM_000405</td>
<td>GM2 ganglioside activator precursor</td>
<td>chr5</td>
<td>150,628,953</td>
<td>150,631,513</td>
<td>Up</td>
<td>Brain</td>
<td>Enard et al. and Caceres et al.</td>
<td>0.4397</td>
<td></td>
</tr>
<tr>
<td>GOSR1</td>
<td>NM_00100724.1</td>
<td>Golgi SNAP receptor complex member 1 isoform 3</td>
<td>chr17</td>
<td>25,828,552</td>
<td>25,877,957</td>
<td>Up</td>
<td>Liver</td>
<td>Gilad et al. 2006</td>
<td>0.2787</td>
<td></td>
</tr>
<tr>
<td>GOSR1</td>
<td>NM_00100724.1</td>
<td>Golgi SNAP receptor complex member 1 isoform 3</td>
<td>chr17</td>
<td>25,828,552</td>
<td>25,877,957</td>
<td>Up</td>
<td>Brain</td>
<td>Enard et al. and Caceres et al.</td>
<td>0.2787</td>
<td></td>
</tr>
<tr>
<td>GTF2I</td>
<td>NM_032999</td>
<td>General transcription factor II, i isoform 1</td>
<td>chr7</td>
<td>73,516,681</td>
<td>73,619,673</td>
<td>Up</td>
<td>Brain</td>
<td>Enard et al. and Caceres et al.</td>
<td>0.0410</td>
<td></td>
</tr>
<tr>
<td>Gene</td>
<td>Symbol</td>
<td>Description</td>
<td>Chromosome</td>
<td>Start (Human Build35)</td>
<td>End (Human Build35)</td>
<td>Tissue Studied</td>
<td>Study Reference(s)</td>
<td>( \frac{dN}{dS} )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>------------</td>
<td>--------------</td>
<td>------------------------------------------------------------------------------</td>
<td>------------</td>
<td>-----------------------</td>
<td>---------------------</td>
<td>---------------</td>
<td>---------------------</td>
<td>------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HSPA2</td>
<td>NM_021979</td>
<td>Heat shock 70-kDa protein 2</td>
<td>chr14</td>
<td>64,077,212</td>
<td>64,079,708</td>
<td>Up</td>
<td>Brain</td>
<td>0.1539</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LGALS4</td>
<td>NM_006149</td>
<td>Galectin-4</td>
<td>chr19</td>
<td>43,984,158</td>
<td>43,995,409</td>
<td>Up</td>
<td>Liver</td>
<td>0.2471</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NAGPA</td>
<td>NM_016256</td>
<td>N-acetylglucosamine-1-phosphodiester</td>
<td>chr16</td>
<td>5,014,872</td>
<td>5,023,920</td>
<td>Up</td>
<td>Liver</td>
<td>0.0677</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NPD009</td>
<td>#N/A</td>
<td></td>
<td>chr16</td>
<td>8,784,670</td>
<td>8,785,932</td>
<td>Up</td>
<td>Liver</td>
<td>0.4530</td>
<td></td>
<td></td>
</tr>
<tr>
<td>OSBPL8</td>
<td>NM_001003712.1</td>
<td>Oxysterol-binding protein-like protein 8 isoform</td>
<td>chr12</td>
<td>75,269,707</td>
<td>75,477,720</td>
<td>Up</td>
<td>Brain</td>
<td>0.0000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PDE4DIP</td>
<td>NM_014644</td>
<td>Phosphodiesterase 4D interacting protein isoform</td>
<td>chr1</td>
<td>143,562,782</td>
<td>143,706,379</td>
<td>Up</td>
<td>Brain</td>
<td>0.4444</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PMS2L5</td>
<td>NM_174930</td>
<td>Postmeiotic segregation increased 2-like 5</td>
<td>chr7</td>
<td>73,751,552</td>
<td>73,766,505</td>
<td>Up</td>
<td>Brain</td>
<td>#N/A</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PRDX6</td>
<td>NM_004905</td>
<td>Peroxiredoxin 6</td>
<td>chr1</td>
<td>171,713,108</td>
<td>171,724,569</td>
<td>Up</td>
<td>Brain</td>
<td>0.0000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RB1</td>
<td>NM_000321</td>
<td>Retinoblastoma 1</td>
<td>chr13</td>
<td>47,775,911</td>
<td>47,954,123</td>
<td>Up</td>
<td>Liver</td>
<td>0.1163</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RGL1</td>
<td>NM_015149</td>
<td>Ral guanine nucleotide dissociation</td>
<td>chr1</td>
<td>181,871,830</td>
<td>182,164,289</td>
<td>Up</td>
<td>Brain</td>
<td>0.1225</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SF3A3</td>
<td>NM_006802</td>
<td>Splicing factor 3a, subunit 3</td>
<td>chr1</td>
<td>38,195,785</td>
<td>38,229,180</td>
<td>Up</td>
<td>Brain</td>
<td>0.0000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SMAD1</td>
<td>NM_001003688</td>
<td>Sma- and Mad-related protein 1</td>
<td>chr4</td>
<td>146,760,556</td>
<td>146,837,928</td>
<td>Up</td>
<td>Brain</td>
<td>0.1754</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SPTLC1</td>
<td>NM_006415</td>
<td>Serine palmitoyltransferase subunit 1 isoform a</td>
<td>chr9</td>
<td>93,852,933</td>
<td>93,917,459</td>
<td>Up</td>
<td>Brain</td>
<td>0.1333</td>
<td></td>
<td></td>
</tr>
<tr>
<td>STAT6</td>
<td>NM_003153</td>
<td>Signal transducer and activator of transcription</td>
<td>chr12</td>
<td>55,775,457</td>
<td>55,791,428</td>
<td>Up</td>
<td>Liver</td>
<td>0.4141</td>
<td></td>
<td></td>
</tr>
<tr>
<td>THBS4</td>
<td>NM_003248</td>
<td>Thrombospondin 4 precursor</td>
<td>chr17</td>
<td>35,629,100</td>
<td>35,691,965</td>
<td>Up</td>
<td>Brain</td>
<td>#N/A</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TIMP3</td>
<td>NM_000362</td>
<td>Tissue inhibitor of metalloproteinase 3</td>
<td>chr22</td>
<td>31,521,362</td>
<td>31,583,581</td>
<td>Up</td>
<td>Liver</td>
<td>0.0000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>WIRE</td>
<td>NM_133264</td>
<td>WIRE protein</td>
<td>chr2</td>
<td>43,361,192</td>
<td>43,365,396</td>
<td>Up</td>
<td>Brain</td>
<td>#N/A</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

1Examples of genes showing increased gene expression in human compared to chimpanzee. Data were taken from References 52, 126 and references therein. Only one gene is found consistently upregulated in both studies (GOSR1). The Gene Symbol, RefSeq ID, chromosomal location (Human Build35), and the tissue studied in every study are also reported. \( \frac{dN}{dS} \) (a measure of the selection pressure acting upon the sequence) was retrieved from ENSEMBL. Only one gene could be a candidate to have been under positive selection (CA2, \( \omega = 2.13 \)).
A human-accelerated region of single-basepair substitution, HAR1F. (a) A highly conserved noncoding DNA segment, HAR1F, shows an excess of substitutions within the human lineage based on a multiple sequence alignment (123). (b) HAR1F is transcribed into a noncoding RNA. Many of the single-basepair substitutions correspond to compensatory changes within the context of RNA secondary structure (yellow–blue denotes compensatory mutations, purple denotes substitutions in unpaired regions, and red denotes noncompensatory changes).

Figure 3

A human-accelerated region of single-basepair substitution, HAR1F (a) A highly conserved noncoding DNA segment, HAR1F, shows an excess of substitutions within the human lineage based on a multiple sequence alignment (123). (b) HAR1F is transcribed into a noncoding RNA. Many of the single-basepair substitutions correspond to compensatory changes within the context of RNA secondary structure (yellow–blue denotes compensatory mutations, purple denotes substitutions in unpaired regions, and red denotes noncompensatory changes).

of evolutionary changes within the human lineage (123). Among such regions, HAR1 (human accelerated region 1) has the distinction of being one of the most accelerated unique regions of the genome and is itself part of a novel RNA gene (HAR1F) that is highly expressed within the cortex of the human brain (see Figure 3). Similarly, Prabhakar and colleagues identified a 546-basepair element (HACNS1) conserved among terrestrial vertebrate genomes that had acquired 16 specific changes within the human lineage. In vivo transgenic experiments showed that a subset of these human-specific changes confers patterns of expression strongly associated with limb development relative to chimpanzees or macaques. Their results implicate changes within the conserved noncoding sequences in creating a de novo enhancer associated with anterior wrist and proximal thumb development. Although the role of this enhancer with respect to nearby genes CENTG2 and GBX2 is unknown, the authors speculate that this burst of nucleotide changes may have contributed to the dexterity of the human hand or opposability of the thumb (125) (Figure 4).

In total, several lines of evidence strongly suggest that gene-expression differences have been a catalyst of primate evolution and adaptive specializations. These findings raise the possibility that there has been an overemphasis on coding sequence differences and, perhaps, too much attention paid to the brain (as opposed to other morphological traits). Proving that these gene regulatory differences played a major role in primate adaptation remains a significant challenge, especially in the absence of “relevant” model organisms where these mutational effects can be directly tested. However, the discovery of previously unrecognized evolutionary targets potentially important in neurodevelopment and hand dexterity predicts that mutations in these may underlie neurocognitive disease in humans. The discovery of de novo mutations in association with human disease for these transcripts or cis-acting elements would provide strong evidence of their functional import.
HACNS1 human-accelerated mutations create a limb enhancer. (a) Multiple sequence alignment of the putative enhancer HACNS1 with other vertebrate genomes shows an excess of 13 human-specific substitutions (red). (b) Transgenic mouse lacZ expression pattern using an enhancer in which the 13 human-specific substitutions were introduced against the orthologous chimpanzee sequence background. (c) Expression pattern using an enhancer reverting the substitutions in the human sequence to the nucleotide states in chimpanzee and rhesus. These results show that those substitutions drive strong gene expression within the limbs.

**PRIMATE GENOME EVOLUTION**

Genomes are highly dynamic entities that evolve rapidly and nonuniformly through time; the genomes of primates are no exception. As genome-wide data and new technologies have become available, structural variation and copy-number differences have emerged as another important aspect of primate genetic variation (16, 28). With limited exceptions such as the gibbon (108), there are relatively few cytological differences among primate chromosomes. Most of the cytogenetic differences between humans and apes (167) have now been well characterized and most resolved at the molecular level (24, 55, 59, 87, 89, 98, 133, 149–151). However, the majority of structural changes smaller than a few megabasepairs preclude detection by standard cytogenetic approaches. With the sequencing of the non-human primate genomes, the extent of this submicroscopic variation became more evident. A comparison of human and chimpanzee genomes estimated more than $\sim 90$ Mb of DNA affected by insertion, deletion, duplication, and inversion (being $\sim 40–45$ Mb in each lineage) (28, 107). Although single-basepair changes are far more numerous, structural variants have been estimated to affect 3–4 times the number of basepairs between human and great ape (i.e., 90 million basepairs of structural variation versus 30 million basepairs of single nucleotide difference between human and chimpanzee).

Due to the limitations of the early draft assemblies, these estimates likely represent a lower bound. Newman and colleagues (114), for example, mapped chimpanzee fosmid end-pairs against the human genome assembly and detected more than 500 insertion, deletion, and inversion events, ranging in length from 12 to 40 kbp. Most of these were previously unknown. In another study comparing the human and chimpanzee genome, a total of 1576 putative regions of inversion were detected (47). Using the gorilla genome as an outgroup,
many of the validated sites were found to be relatively young events and some were polymorphic within the human lineage. Similarly, Szamalek and colleagues (148) performed gene-order comparisons for more than 10,000 orthologous genes between human and chimpanzees and identified 71 putative microrearrangements. The importance of structural variation and copy-number polymorphism in human disease and disease susceptibility and its abundance suggest that such variation may have had profound repercussions in the evolution of human and great ape primates. The role of these events in altering the pattern of gene expression and chromatin organization has not yet been addressed.

Two features of primate genomes may account for the abundance of large-scale structural variation observed in primate genomes, namely segmental duplications and retrotransposons. Segmental duplications (SDs) are segments of DNA greater than 1 kbp in length with high sequence identity (>90%) that typically map to two or more locations in the genome. Recent comparative work among humans and non-human great-apes has shown that the human and great-ape lineage are particularly enriched for interspersed duplications with a suggested burst occurring in the common ancestor of the human and African great-apes (105). Breakpoints of large-scale structural variation between and within species preferentially map to regions of segmental duplications (1, 4, 121). When compared to other sequenced mammalian genomes, primate (especially human and great ape) segmental duplications tend to be larger, more complex, and more interspersed [(5, 29, 77, 137, 138; for a detailed description of the organization and distribution of primate SDs see (6)]. These peculiar features promote further genomic instability leading to extensive copy-number variation both within and between species enhanced by long-distance nonallelic homologous recombination and other less well understood mechanisms of genetic exchange (78). As a result of this constant genomic turnover, evolutionarily shared duplications often show as much copy-number variation as lineage-specific duplication events (120, 121).

New insights regarding the evolution of human SDs have recently come to light (77, 109). Using a computational graph-theory and phylogenetic approach, Jiang et al. revealed that human SDs are frequently organized around “core” elements that show greater EST and exon density when compared to flanking segmental duplications (77). Detailed comparative sequencing of one of these core elements on chromosome 16 in human and ape showed that the complexity of these regions has emerged as a result of the serial accretion of duplicated segments centered around the core duplication segment (Figure 5). In different primate lineages, the core segments have moved or been copied to new locations or entirely different chromosomes leading to a completely different suite of segmental duplications all centered around the same core duplication (78). This feature helps to explain why unique genes mapping in close proximity to these duplication blocks have a ten-fold increased probability of being duplicated when compared to a random model of genome duplication (28, 40, 140)—a phenomenon known as duplication shadowing (28).

In addition to being hotspots of genomic structural variation, segmental duplications are substrates for the emergence of new genes and gene families (39, 79). Primate segmental duplications are enriched for exons when compared to mouse segmental duplications (137). Most of the primate gene expansions correspond to regions of segmental duplication (28, 36). This includes the olfactory receptor gene family (153), which has been subjected to sudden decline in catarrhine primates (53, 54), although it may still play a critical role in kin recognition (71).

In general, gene expansions seem to have been common in primate evolution (36, 61), and many notable examples have been found [such as PRAME (expressed antigen of melanoma) within the human lineage and PFKP (sugar metabolism) and DIP2C (segmentation patterning) within the macaque lineage]. Examples of reported human-specific expansions...
Core duplication hypothesis of human segmental duplications. Duplicative transposition or biased gene conversion duplicates copies to new locations with different unique sequences flanking the duplicated copy of the core; a secondary duplication moves into a third location but also carries flanking duplicons from the second locus to the third; the process repeats itself during the course of evolution, leading to the formation of large complex blocks of intrachromosomal segmental duplication along the chromosome that can now promote nonallelic homologous recombination. These events occur before and after speciation, leading to both lineage-specific and shared duplication blocks between closely related species. Segmental duplications located at the periphery are more likely to be evolutionary young or lineage specific (e.g., duplication D), whereas the core is common to all the duplication blocks. Paralogous sequence exchanges can occur between loci changing the sequence and erasing evolutionary history. Data based on primate comparative sequencing (78) and evolutionary reconstruction of human segmental duplications (77).
Table 3  Great ape and human gene family expansions

<table>
<thead>
<tr>
<th>Gene family name</th>
<th>Representative gene</th>
<th>Gene info</th>
<th>Evidence of positive selection</th>
<th>Gene expression</th>
<th>Estimated human copy number</th>
<th>Chromosome</th>
<th>Comment</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Morpheus</td>
<td>NPIP</td>
<td>Nuclear pore complex interacting protein</td>
<td>++</td>
<td>Widely expressed</td>
<td>18</td>
<td>16</td>
<td>50X dN/dS</td>
<td>(79)</td>
</tr>
<tr>
<td></td>
<td>DUF1220 (NBPF)</td>
<td>NBPF11</td>
<td>+ (but see 155)</td>
<td>Neurons (among other tissues)</td>
<td>11</td>
<td>1</td>
<td>Expressed in neurons</td>
<td>(124, 155)</td>
</tr>
<tr>
<td></td>
<td>RGP</td>
<td>RANBP2</td>
<td>+</td>
<td>Testis/cortex</td>
<td>8</td>
<td>2</td>
<td></td>
<td>(30)</td>
</tr>
<tr>
<td></td>
<td>LRRC37</td>
<td>LRRC37</td>
<td>Unknown</td>
<td>Ubiquitous</td>
<td>3</td>
<td>17</td>
<td></td>
<td>(77)</td>
</tr>
<tr>
<td></td>
<td>TBC1D3</td>
<td>TRE2/USP6</td>
<td>Unknown</td>
<td>Testis</td>
<td>4</td>
<td>17</td>
<td></td>
<td>(119)</td>
</tr>
</tbody>
</table>

1Examples of gene families embedded in “core” dupion (77) that have been greatly duplicated specifically in great apes (NPIP, RANBP2, TRE2, NBPF11) or in primates (LRRC37). Evidence of positive selection as well as information on the gene function and expression or the estimated copy number in human (according to the number of WGS reads mapping to the region) is also provided.
Figure 6

A burst of PTERV1 retroviral insertions in African great apes and Old World monkey species but not orangutan or human. (a) Whole-genome landscape of PTERV1 integration sites is shown using the human chromosomes as a reference (build35). (b) A detailed view on one chromosome (chromosome 2) is shown. While some species has been massively plagued by the retrovirus that integrated into the germline (chimpanzee, gorilla, macaque, and baboon), other species (human and orangutan) are unusually devoid. Furthermore, ∼95% of these sites were found to be in nonorthologous locations when compared between species, suggesting episodic events early in the history of each of these species potentially from an exogenous source.

While some species has been massively plagued by the retrovirus that integrated into the germline (chimpanzee, gorilla, macaque, and baboon), other species (human and orangutan) are unusually devoid. Furthermore, ∼95% of these sites were found to be in nonorthologous locations when compared between species, suggesting episodic events early in the history of each of these species potentially from an exogenous source.

Promote speciation events between contiguous (parapatric) or partly overlapping (sympatric) populations. Among the several models of chromosomal speciation (see 161), the so-called suppressed-recombination model of speciation suggests that chromosomal rearrangements serve as a genetic barrier between populations and, hence, substitutions linked with the rearranged chromosomes cannot be freely exchanged among karyotypically distinct subpopulations, eventually leading to incompatibilities and thus to speciation (112). Studies within a variety of different lineages such as Drosophila, Anopheles, murids, shrews, or sunflowers (3, 10, 116, 131, 132) have provided some support for this model of speciation.

In the first study to test predictions of suppressed-recombination chromosomal speciation models in primates, Navarro & Barton reported an association between chromosomal...
rearrangements and higher evolutionary rates based on an analysis of a limited set of 115 autosomal orthologous genes from humans and chimpanzees (113). This and other observations (such as a lower level of polymorphism in rearranged chromosomes) were consistent with the predictions of the model. Several interpretations for the results were given, the most controversial being the suggestion that under the hypothesis tested, the results would hold only if the chromosomal rearrangements had been barriers in parapatry for no less than half of the time of divergence between humans and chimpanzees (130). This conclusion was striking because it ran counter to both anthropological data and the molecular dating of rearrangement events (151).

Subsequent analyses have questioned the results and inferences from the initial paper. For example, Lu and colleagues found that rapidly evolving genes may not have a homogeneous distribution among chromosomes and that rearranged chromosomes may have been linked to rapidly evolving genes due to factors unrelated to speciation. An alternative explanation was that the GenBank sequences used by the initial study and by Lu et al. (99) were biased and not representative of the rest of the genome. This latter explanation seemed to be the main reason for the initial result since subsequent genome-wide studies (32, 51, 64) found evolutionary rates threefold smaller than the original dataset. Similarly, studies (mainly from noncoding DNA) found that the average nucleotide divergence was in fact lower in rearranged chromosomes when compared to colinear chromosomes (103). This observation was replicated by Khaitovich and colleagues (90) but not by Zhang and colleagues (168), although the latter considered only a subset of genes (genes with statistical different expression pattern in human and chimpanzee).

Overall, it appears that there is little evidence in support of human-chimpanzee speciation via suppressed-recombination. Chromosomal speciation in primates, however, cannot be definitively ruled because (a) chromosomal speciation does not have to involve all rearranged chromosomes, and (b) speciation might have involved other noncoding functional elements, such as genes that do not encode proteins (microRNAs, for example) or other regulatory elements (such as transcription factor binding sites). The discrepancy between gene expression and positive selection data may provide some support for this view.

Ancient Hybrid Models

Coalescent models have recently been applied to the question of human and chimpanzee speciation. Based on the hypothesis that allopatric models predict similar divergence times among different regions of the genome, Osada & Wu found that coding regions shared deeper genealogies than intergenic regions as suggested by parapatric models, because coding regions are more likely to have been involved in hybrid incompatibilities or adaptive evolution (117). The analysis supported a complex parapatric view of speciation between human and chimps. Patterson and colleagues (118) addressed the question by analyzing the heterogeneity of human chimpanzee divergencies in DNA sequence based on analysis of genomic sequence from human, chimpanzee, gorilla, and...
other related species (orangutan or macaque). Human-chimpanzee genetic divergences varied from $\sim 84\%$ of the average to up to $147\%$, a range of more than 4 million years. Notably, they observed a lower divergence between X-chromosomal sequences than for autosomal sequences, a circumstance they suggest could be explained if human and chimpanzee initially diverged, and then later exchanged genes before separating permanently.

They proposed that humans and chimpanzees would be more closely related through the X chromosome, by the following process: If human and chimpanzee ancestors initially speciated and then interbred, male hybrids might be partly sterile, as Haldane’s rule on heterogametic sex suggested (62). A viable population could then only have arisen if the fertile females mated back to one of the ancestral populations, producing fertile male hybrids that then transmitted X chromosomes derived almost entirely from the initial population. Several concerns, however, have risen with respect to the interpretation of these results. Barton (9) noted that the scenario proposed by Patterson and colleagues is not the only compatible explanation. In fact, the heterogeneity in divergence observed between human and chimpanzee genomes is consistent with quick speciation (allopatric) from a large ancestral population ($N_e \sim 45,000$), a view supported by others (66). Moreover, a similar level of X-chromosome divergence could also be explained by the amount of time that copies of the X chromosomes spent in the male/female lineages, which of course depends on the male/female ratio ($\alpha$) (100, 157). Additionally, Burgess & Yang (19) analyzed 7.4 Mb aligned for human, chimp, gorilla, orangutan, and macaque showing that the peculiar reduction of X-chromosome diversity may have arisen as a result of specific selective sweeps on the X chromosome prior to the human/chimp speciation (19).

In summary, applications of genome-wide datasets to develop models of primate speciation have provided no clear consensus on the underlying mechanisms. The controversy and confusion are largely a reflection of the fact that sequence divergence data are much less uniform than anticipated. As population genetic models evolve to more adequately incorporate these data from additional primate genomes, greater insight into speciation should emerge.

**PRIMATE DIVERSITY AND RECOMBINATION**

**Primate Diversity**

Analysis of genetic variation within and between primate populations is a powerful tool to understand the evolutionary history, demography, and effective ancestral population size of primate populations. High-throughput analyses of SNPs, haplotypes, and CNVs, for example, have significantly resolved the population ancestry and geographic origin of humans (72, 95, 127). Although studies of nonhuman primates are less developed, the availability of reference genomes and the sampling of genetic diversity (largely SNPs) from multiple individuals of different species have begun to provide other contrasting models of primate demography. From the first restriction fragment length polymorphism (RFLP) studies of ape mitochondrial DNA (46) suggesting that chimpanzees and gorillas possessed two to three times more genetic diversity than humans, it has been evident that humans have less genetic diversity than that of our closest great-ape relatives. Later analysis of nuclear loci confirmed the existence of fundamentally different levels of genetic diversity, although at a lower level—approximately 50% more than humans (76, 80).

Wild primate populations are notoriously difficult to census, suggesting that genetic approaches to estimating effective population size can help understand the evolution of current populations and their demographic history. Previous studies of chimpanzee genetic diversity suggested that chimpanzees can be differentiated in three main subpopulations: Eastern (*Pan troglodytes schweinfurthii*), Central (*Pan troglodytes troglodytes*), and Western (*Pan troglodytes verus*). From all three...
subpopulations, Central chimpanzees show the greatest genetic diversity and the largest effective population size (12, 13, 26, 48, 49, 80, 144, 163, 166). The effective population size of Central chimpanzees has been estimated to be much higher than human and Western chimpanzees perhaps as a result of three- to four-fold population expansion after separation from the Western chimpanzees (Figure 7). Other studies suggest an even larger effective population size (~100,000) (26). Given that the estimated effective population size for that common ancestor of bonobo and chimp is approximately 25,000 (13,000 to 30,000), which in turn is smaller than

<table>
<thead>
<tr>
<th></th>
<th>Bornean orangutan</th>
<th>Sumatran orangutan</th>
<th>Western gorilla</th>
<th>Eastern gorilla</th>
<th>Western chimpanzee</th>
<th>Central chimpanzee</th>
<th>Eastern chimpanzee</th>
<th>Bonobo</th>
<th>Human</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Actual population sizes (IUCN red list)</strong></td>
<td>~45-69 K</td>
<td>7.3 K</td>
<td>~300 K</td>
<td>~10 K</td>
<td>21.3-55.6 K</td>
<td>70-116.5 K</td>
<td>6.4-9.6 K</td>
<td>29.5-50 K</td>
<td>&gt; 6,000 M</td>
</tr>
<tr>
<td><strong>Effective population sizes</strong></td>
<td>~10 K</td>
<td>~17 K</td>
<td>~15 K</td>
<td>~7 K</td>
<td>~9 K</td>
<td>~25-100 K</td>
<td>~20 K</td>
<td>~10 K</td>
<td>~10 K</td>
</tr>
</tbody>
</table>

Figure 7
Demographic parameters and speciation times of humans and great apes. Current population sizes (red; IUCN red list of threatened species at http://www.iucnredlist.org/), effective population sizes, and estimated ancestral population sizes (blue) are shown in thousands of individuals for each hominid species and subspecies. Divergence times in either millions of years (Mya) or thousands of years (Kya) are indicated. Population genetic parameters are based on diversity data presented in the following papers: Wall (158), Won & Hey (163), Becquet & Przeworski (13), Hobolth et al. (66), Burgess & Yang (19), Steiper (141), and Caswell et al. (26). * correspond to most of the studies but Burgess & Yang (19) suggested a lower bond of 14 Mya to 22 Mya.
that estimated from human and chimpanzee (∼100,000) (13, 19, 26, 66, 158, 163), the data suggest that human and chimpanzees have experienced a drastic reduction in their effective population sizes.

Similarly, the genetic diversity of Western gorillas and Eastern gorillas and orangutan subspecies also suggests lower effective population sizes than their ancestral population (∼40,000 and ∼87,000, respectively) (13, 19, 66) and lower than the common ancestral human, chimp, gorilla, and orangutan population sizes (70,000 and up to 127,000) (19). On a much smaller scale, and using an approach that allows the ratio of past and present population effective size to be estimated (145), Goossens and colleagues (58) analyzed microsatellite allelic variation in endangered Kinabatangan (Bornean) orangutans and inferred a drastic reduction in population size over the past several decades. In this instance, the reduction in population size has been attributed to direct effects of recent human activities, especially habitat reduction through logging and agricultural activities. Taken together, a consensus emerges that the ancestral effective population size in hominoid lineage was five- to tenfold higher than almost all primate effective population sizes, suggesting a continuous reduction of effective individuals during great-ape evolution.

The recent sequence of the macaque genome has been accompanied by a detailed diversity analysis of a few Kb (∼150 Kb) in 47 macaque individuals from two populations (9 from Chinese and 38 from Indian populations) (51, 65). This survey showed that Chinese macaques have an excess of rare variants when compared to the Indian population where intermediate frequency variants predominate. These findings contrasted with earlier studies based on mtDNA and a much smaller set of SNPs that showed much more modest levels of differentiation (45, 139). Demographic inferences of these findings indicate that the Chinese population was first expanded and that later there was a drastic reduction of population size in the Indian population (population size in the Chinese population is now estimated to be one order of magnitude higher than the Indian population). These findings have immediate practical application in mapping of genetic traits. The patterns of LD (linkage disequilibrium) decay suggest that Indian macaques are especially useful for disease association studies since their LD patterns extend even farther than humans, hence fewer markers would be required to discover significant associations. On the other hand, the Chinese rhesus would be more useful in winnowing the interval associated with any phenotypic trait (65) as a consequence of greater decay in the patterns of LD disequilibrium.

Recombination

Among the forces shaping primate genomes are adaptations of cellular mechanisms that facilitate reproduction, including the structural components of recombination pathways and impacts of genome organization. Recombination hotspots are defined as narrow regions of the genome (1–2 Kb) in which recombination occurs to a higher degree than the genome average. Recombination has been assessed from direct genotyping of recombination events in sperm (2, 74) or inferred by population genetic analysis of genome sequence data (33). The former method, although time-consuming, identified the existence of recombination hotspots in human males (2, 74). An alternative approach is to make use of the patterns of LD generated to estimate the recombination rate required to produce the population distribution of observed haplotypes. Importantly, the majority of hotspots detected by sperm typing have been also detected by LD-based approaches (75; but see 85). Myers et al. (110) proposed that the human genome contains ∼25,000 hotspots (approximately one hotspot every 50 Kb) and that there are sequence motifs associated with those hotspots that could be experimentally demonstrated to modulate the intensity of recombination (110). These results suggested that recombination rates could be regulated (at least in part) by
Recombination hotspots vary between chimpanzee and human. (a) Distribution of shared hotspots under two statistical models. In the first model (blue) all the hotspots are independent and in the second (red) all hotspots are shared. The observed proportion of shared hotspots (8%) suggests that allelic hotspots of recombination evolve rapidly and change over short periods of evolutionary time. (b) On a broader scale, the recombination rates (measured in 50-Kb windows) correlate, suggesting that the landscape of recombination remains relatively constant.

cis-acting sequences, and motifs have been identified associated with both allelic and non-allelic recombination (111). Genetic linkage maps in baboon, green monkey, and macaque have shown that the recombination maps are significantly shorter than the human (34, 73, 134). The available results suggest that humans and, perhaps, hominoids more generally have evolved higher rates of recombination with implications of greater genetic diversity in a background of a reduced rates of single nucleotide substitution (42, 56, 143).

One of the most interesting findings that has emerged from comparative maps of fine-scale recombination has been that recombination hotspots are not necessarily conserved between human and chimpanzee (128, 162) (Figure 8). These findings establish that positional recombination rates may change rapidly over time, a concept reinforced by the finding that recombination rates are highly variable among different humans (92). However, when analyzed on a broader scale (∼50 Kb), a weak correlation in rates of recombination exists between humans and chimpanzees (128, 162) and among other mammals (37). This has opened the possibility that the regional background rate of recombination remains relatively constant but that the actual hotspots are transient, perhaps as a result of competition. Yet this difference in recombination must occur in a background where the sequence (and thus the sequence motifs) are nearly identical, suggesting that other structural changes in trans-regulation, epigenetic factors, or selection against alternative alleles in the motifs.
affect the specificity of recombination and ultimately rates of recombination in different primate lineages.

**THE FUTURE**

A complete framework for understanding of the human genome requires knowledge of its evolution, necessitating comparative studies with other species, including closely related primates. With the emergence of next-generation sequencing technology (101) and the concomitant reduction in sequencing costs, it is no longer outside the realm of possibilities that "complete" genome sequencing of all species of the primate order will be achieved within the next ten years. The prospects of a “500 Primate Genomes Sequencing” project should be tempered with the need to sequence the genomes to a standard of high quality. In light of the complex organization of the human genome and the finding that so much of the genetic difference maps to these complex regions, simply aligning sequencing reads to a reference genome and cataloguing the single-nucleotide differences between and within species should not be the end goal. As evidenced by the sequencing of the human and chimpanzee genomes, sequencing primate genomes well is nontrivial. Even whole-genome sequence assembly of primate genomes using "Sanger" capillary sequences leads to the significant loss of information, including the absence of entire genes and gene families (Figure 9). The excitement of sequencing more primate genomes thus should be balanced by insisting that the most dynamic regions are not excluded as part of the process. This requires a greater investment of resources and careful attention to details, which in this era of “slash-and-burn” genomics is becoming a dying art.

The opportunity, however, to sequence the true extent of primate diversity is time-limited. It is a commentary on the success of human beings that the rapid population expansion that has taken place in our...
species over the past approximately 10,000 years has led to dramatic reductions in other primate species. Fully 48% of all primate species are threatened; 70% of Asian primates face extinction and two dozen are critically endangered (http://www.primate-sg.org/redlist08.htm; 68). All of the great apes, to which humans are most closely related (Figure 1)—comprising chimpanzees, bonobos, gorillas, and orangutans—are endangered or critically endangered. Declining populations of primates are the object of conservation efforts to secure habitat and reduce threats to population viability. Knowledge of the genetic structure of primate populations, their demographic history, and significant life-history attributes that can be inferred from assessments of genetic diversity can contribute to efforts to conserve self-sustaining wild primate populations. Thus, there is clear opportunity of comparative genomics studies to contribute to primate conservation actions (135).

Additional whole-genome sequencing projects for primates serve societal interests in identifying the functional elements of the human genome and provide a basis for evaluating genetic diversity and nucleotide sequence evolution not only for species whose genomes have been sequenced, but for closely related species as well. The practical consideration of obtaining appropriate samples for preliminary studies, partial or complete genome sequencing, and postassembly resequencing has received relatively little attention, yet represents a significant community need. The U.S. National Science Foundation recently established the Integrated Non-human Primate Biomaterials and Information Resource (www.ipbir.org) to assemble, characterize and distribute high-quality DNA samples of known provenance with accompanying demographic, geographic, and behavioral information in order to stimulate and facilitate research in primate genetic diversity and evolution, comparative genomics, and population genetics. Many samples in the IPBIR were derived from primates in zoos or U.S. National Primate Centers, and a large proportion of the cell cultures serving as sources of DNA for distribution came from the Frozen Zoo® of the Zoological Society of San Diego.

In addition to access to DNA from index individuals, nucleic acid preparations from additional unrelated individuals are essential for evaluating genetic diversity, SNP discovery, copy-number variation, recombination parameters, effective population size, and parameters of selection, including selective sweeps. SNP identification and analysis is of direct interest as SNP assays are a basic currency of genetic variation, relevant to population genetics, disease association, and phylogeographic studies alike. The potential utility of high-throughput SNP assays is manifested in the aforementioned studies of geographic population of rhesus macaques and the effort to identify factors associated with variability in responses to SIV infection that distinguish rhesus macaques of Indian and Chinese origin (45, 65). Recently, comparison of SNPs in cynomolgus (Macaca fascicularis) and rhesus macaque (M. mulatta) identified shared polymorphisms, raising the possibility that these two well-recognized species may be more closely related than suggested by mitochondrial DNA or morphological comparisons (146). This unexpected finding serves as an additional example of the significance of genomic studies for the understanding of evolutionary processes in primate speciation and phenotypic diversification, relevant to medical primatology and conservation of wild primate populations alike.

Among the primate order of mammals, comparative genomic studies have advanced more rapidly for taxa closely related to humans, chimpanzees, macaques, and baboons. As complete genome-sequencing projects advance for other primate families, including the New World monkeys (Cebidae) and strepsirhine primates (lemurs, lorises, aye-aye, pottos, and galagos), new insights are anticipated as, particularly for a lemur genome project, new information about primate adaptations and evolution can be anticipated (67).

The availability of complete genome sequences from additional primate species and
the elucidation of intraspecific and population diversity at the nucleotide sequence and cytogenetic levels can be applied to conservation efforts for wild populations of threatened primates. Constraints in sample collection may require exploration of new technological approaches as, for example, current studies of intraspecific variation in wild populations of great apes depend on noninvasive samples such as feces. Genome sequencing projects for primates should typically include components evaluating genetic diversity of geographically distinct populations. Ideally, new partnerships between field researchers, governments, conservation organizations, and genome biologists can advance both the understanding of genomic mechanisms of primate adaptations and evolution, while enriching the understanding of primatology and contributing to primate conservation.

DISCLOSURE STATEMENT
The authors are not aware of any biases that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS
We are indebted to J. Pedersen, K. Pollard, S. Salama, M. Caceres, T. Preuss, J. Noonan, M. Oldham, D. Geschwind, Z. Cheng, and S. Ptak for providing access to their data and figures. We would like to thank T. Brown for comments during the preparation of the manuscript. T. M-B is funded by a Marie Curie Fellowship. O.R. acknowledges support from NSF BCS-0094993, the Caesar Kleberg Wildlife Foundation, and the John and Beverly Stauffer Foundation. E.E.E. is an investigator of Howard Hughes Medical Institute. E.E.E. acknowledges support from NIH grant GM058815.

LITERATURE CITED


83. Kapitonov VV, Jurka J. 2005. RAG1 core and V(D)J recombination signal sequences were derived from Transib transposons. *PLoS Biol.* 3:e181
146. Street SL, Kyes RC, Grant R, Ferguson B. 2007. Single nucleotide polymorphisms (SNPs) are highly conserved in rhesus (Macaca mulatta) and cynomolgus (Macaca fascicularis) macaques. BMC Genomics 8:480

www.annualreviews.org • Primate Comparative Genomics


Contents

Chromosomes in Leukemia and Beyond: From Irrelevant to Central Players
Janet D. Rowley ................................................................. 1

Unraveling a Multifactorial Late-Onset Disease: From Genetic Susceptibility to Disease Mechanisms for Age-Related Macular Degeneration
Anand Swaroop, Emily Y. Chew, Catherine Bowes Rickman, and Gonsalo R. Abecasis ... 19

Syndromes of Telomere Shortening
Mary Armanios .................................................................. 45

Chronic Pancreatitis: Genetics and Pathogenesis
Jian-Min Chen and Claude Ferec ......................................... 63

The Genetics of Crohn's Disease
Johan Van Limbergen, David C. Wilson, and Jack Satsangi ........... 89

Genotyping Technologies for Genetic Research
Jiannis Ragoussis ............................................................... 117

Applications of New Sequencing Technologies for Transcriptome Analysis
Olena Morozova, Martin Hirst, and Marco A. Marra .................. 135

The Posttranslational Processing of Prelamin A and Disease
Brandon S.J. Davies, Loren G. Fong, Shao H. Yang, Catherine Coffinier, and Stephen G. Young ................................................................. 153

Genetic Testing in Israel: An Overview
Guy Rosner, Serena Rosner, and Avi Orr-Urtreger ..................... 175

Stewardship of Human Biospecimens, DNA, Genotype, and Clinical Data in the GWAS Era
Stephen J. O'Brien .............................................................. 193

Schistosoma Genetics: New Perspectives on Schistosome Biology and Host-Parasite Interaction
Ze-Guang Han, Paul J. Brindley, Sheng-Yue Wang, and Zhu Chen ....................... 211

Evolution of Genomic Imprinting: Insights from Marsupials and Monotremes
Marilyn B. Renfree, Timothy A. Hore, Geoffrey Shaw, Jennifer A. Marshall Graves, and Andrew J. Pask ................................................................. 241
Methods for Genomic Partitioning
Emily H. Turner, Sarah B. Ng, Deborah A. Nickerson, and Jay Shendure .................. 263

Biased Gene Conversion and the Evolution of Mammalian Genomic Landscapes
Laurent Duret and Nicolas Galtier ............................................................. 285

Inherited Variation in Gene Expression
Daniel A. Skelly, James Ronald, and Joshua M. Akey ...................................... 313

Genomic Analyses of Sex Chromosome Evolution
Melissa A. Wilson and Kateryna D. Makova ................................................. 333

Sequencing Primate Genomes: What Have We Learned?
Tomas Marques-Bonnet, Oliver A. Ryder, and Evan E. Eichler ......................... 355

Genotype Imputation
Yan Li, Cristen Willer, Serena Sanna, and Gonçalo Abecasis ............................ 387

Genetics of Athletic Performance
Elaine A. Ostrander, Heather J. Huson, and Gary K. Ostrander ...................... 407

Genetic Screening for Low-Penetrance Variants in Protein-Coding Diseases
Jill Waalen and Ernest Beutler ..................................................................... 431

Copy Number Variation in Human Health, Disease, and Evolution
Feng Zhang, Wenli Gu, Matthew E. Hurles, and James R. Lupski ................... 451

The Toxicogenomic Multiverse: Convergent Recruitment of Proteins
Into Animal Venoms
Bryan G. Fry, Kim Roelants, Donald E. Champagne, Holger Scheib,
Joel D.A. Tyndall, Glenn F. King, Timo J. Nenculainen, Janette A. Norman,
Richard J. Lewis, Raymond S. Norton, Camila Renjifo,
and Ricardo C. Rodríguez de la Vega ......................................................... 483

Genetics, Medicine, and the Plain People
Kevin A. Strauss and Erik G. Puffenberger .............................................. 513

Indexes

Cumulative Index of Contributing Authors, Volumes 1–10 .............................. 537

Cumulative Index of Chapter Titles, Volumes 1–10 .......................................... 541

Errata

An online log of corrections to *Annual Review of Genomics and Human Genetics* articles may be found at http://genom.annualreviews.org/