# Positive Selection in the Human Genome Inferred from Human–Chimp–Mouse Orthologous Gene Alignments

A.G. CLARK,* S. GLANOWSKI,† R. NIELSEN,‡ P. THOMAS,¶ A. KEJARIWAL,¶ M.J. TODD,‡ D.M. TANENBAUM,§ D. CIVELLO,** F. LU,§ B. MURPHY,† S. FERRIERA,† G. WANG,† X. ZHENG,¶ T.J. WHITE,** J.J. SNINSKY,** M.D. ADAMS,§,†† AND M. CARGILL**,††

**Molecular Biology & Genetics, Cornell University, Ithaca, New York 14853; †Applied Biosystems, Rockville, Maryland 20850; ‡Biological Statistics & Computational Biology, Cornell University, Ithaca, New York 14853; ¶Protein Informatics, Celera Genomics, Foster City, California 94404; §Celera Genomics, Rockville, Maryland 20850; **Celera Diagnostics, Alameda, California 94502.*

The availability of genomic sequence from diverse organisms allows the opportunity to identify genes that have undergone evolutionary divergence from our most recent common ancestors. By fitting aligned DNA sequences from multiple species to models of sequence divergence it is possible to distinguish divergence due to random drift from that caused by nonneutral processes such as natural selection. The key to this problem is to realize that nucleotide sites can be partitioned a priori according to whether substitutions at these sites change the encoded amino acid or are silent. Under neutrality, these two types of substitutions are expected to be distributed at random, and a variety of tests have been devised to test this null hypothesis. The identification of genes that have undergone positive Darwinian evolution (inferred from an excess of amino acid-changing substitutions) might lead to hypotheses of physiological mechanisms that underlie the specialization of species and their reproductive isolation. Furthermore, discovery of genes that appear to show adaptive evolution in humans may lead to the identification of genes important in human disease.

Although humans and chimpanzees differ by only 1.2% in their coding regions (Chen and Li 2001; Ebersberger et al. 2002), this small level of sequence divergence can still provide clues to adaptive evolution. However, the addition of a third, out-group species provides significantly more information. By adding the orthologous mouse sequence to each gene alignment, it is possible to identify the most likely ancestral allele and therefore information about the direction of DNA substitutions. If human and chimpanzee differ at a nucleotide position, and, say, the chimpanzee and mouse nucleotides are identical at this site, then we can infer that the change occurred most likely on the human lineage after the divergence from our common ancestor with chimpanzee. This logic is formalized and made statistically rigorous by maximum likelihood procedures that are widely available. In this paper, we analyze alignments of 7,645 chimpanzee gene sequences to their unambiguous human and mouse orthologs and identify genes that appear to have undergone adaptive evolution in the human genome.

††Present address: Department of Genetics, Case Western Reserve University, 10900 Euclid Avenue, Cleveland, Ohio 44106.

## IDENTIFICATION OF HUMAN–MOUSE ORTHOLOGS

An accurate and unambiguous identification of human–mouse orthologous gene pairs is necessary for the evolutionary analysis. Incorrectly paired orthologs, paralogous proteins, and inaccurate annotation can all corrupt the evolutionary analysis described here. Our initial set of orthologs was taken from Mural et al. (2002), who had identified 32,598 transcript and 21,638 gene pairs from an analysis of Celera's human and mouse genome assemblies and annotations (Venter et al. 2001; Mural et al. 2002). Each gene pair was scored for four lines of evidence of orthology: sequence identity from tblastx, syntenic anchor or syntenic block evidence, or shared PANTHER protein family classification (Thomas et al. 2003b). This ortholog set provided unambiguous transcript pairs, but there were ambiguities when the transcripts were collapsed into gene pairs. To avoid duplicate evolutionary analysis of the same base, we derived a subset of unambiguous gene orthologs by selecting the gene pair with the most lines of evidence or the pair with the highest tblastx identity. Furthermore, we required that each ortholog be located in a human–mouse syntenic block (Mural et al. 2002), reducing the entire set to 14,104 gene/transcript pairs (85% with tblastx evidence, 90% with syntenic anchor evidence, and 75% with shared PANTHER family evidence; 9,017 pairs had PANTHER assignments). The set of 14,104 mouse–human orthologs was further reduced to 7,645 for methodological reasons (Table 1). The final breakdown of evidence classes was 87% with tblastx evidence, 90% with syntenic anchor evidence, and 80% with shared PANTHER family evidence (5,136 pairs had PANTHER assignments).

We compared the set of 7,645 human–mouse orthologs analyzed in this paper with orthologs from other sources (Table 2). Specifically, April 2003 downloads of mouse–human orthologs from HomoloGene (ftp.ncbi.nih.gov/pub/HomoloGene/hmlg.ftp), Homologous sequence pairs (ftp.ncbi.hih.gov/refseq/LocusLink/homol_seq_pairs.gz), Homology Map (ftp.ncbi.nlm.nih.gov/Homolgy), and the Mouse Genome Database (ftp.informatics.jax.org/ pub/reports.html) were used for validation. We compared the concordance rate of mouse–hu-

**Table 1.** Derivation of 7645 Human–Mouse Orthologs

| Number of genes | Number of transcripts | Justification for removal |
|---|---|---|
| 21,638 | 32,598 | starting set of human–mouse orthologous pairs |
| –7,534 | –18,494 | removal of ambiguous gene pairs |
| –992 | –992 | removal of genes with an absence of human PCR sequence data |
| –424 | –424 | removal of genes with an absence of chimp PCR sequence data |
| –426 | –426 | removal of genes that failed chimp alignment QC |
| –3,821 | –3,821 | removal of genes that failed mouse alignment QC |
| –796 | –796 | removal of genes with less than 50 amino acids between human, chimp, and mouse |
| = 7,645 | = 7,645 | final set of human–mouse orthologs analyzed |

man orthologs between the different sets and found that the data sets were 97% identical (Table 2). Manual examination of the discordant pairs indicated that the discordances were due to methodological issues rather than evidence supporting the ortholog.

## CONSTRUCTION OF HUMAN–MOUSE–CHIMP ALIGNMENTS

The chimpanzee sequence was produced using 201,805 primer pairs designed to 23,363 human coding sequences based on expertly annotated genes in Celera's human genome assembly (Venter et al. 2001). The primer pairs covered 27.6 Mb of human coding sequence, with the coding sequences of most gene families covered at 92% of the genes. Primer pairs were amplified in one male common chimpanzee (*Pan troglodytes*) (4X0033, Southwest National Primate Research Center) and 39 human females (19 African-Americans and 20 Caucasians, Coriell Cell Repositories), and were sequenced using standard sequence chemistry. Approximately 85% of the human amplicons and 75% of the chimp amplicons resulted in good-quality sequence data that could be analyzed further.

Quality-trimmed chimp traces for a human gene were blasted against human exon sequence (Venter et al. 2001). Matches were ordered by decreasing gap count and placed in exon order to create virtual chimp transcripts. Care was taken to maintain the reading frame by internal trimming of low-quality bases and insertion of unambiguous placeholders for unmatched human exon segments. After assembly, 73% of the human coding sequence was covered with chimp sequence. The gaps in this coverage tended to be spread across genes fairly uniformly. In particular, every gene recovered at least some chimpanzee sequence.

The mouse sequences were originally derived from Celera's mouse genome assembly (Mural et al. 2002). Since the gene annotations were not expertly reviewed, we attempted to improve the automated mouse annotation by generating a mouse transcript based on human–mouse alignments. For human exons that appeared to be "missing" from the orthologous mouse transcript (identified by blasting the human exon set for a gene to the orthologous mouse transcript), we blasted the human exon sequence to surrounding mouse genome sequence. Reconstruction of each mouse transcript based on the human–mouse alignments yielded a "humanized" mouse transcript. Compared to human proteins, humanized mouse transcripts had the same level of protein identity (average 80%) as computationally predicted mouse tran-

**Table 2.** Comparison of Different Mouse–Human Ortholog Sets

| | Number of orthologs | Number of orthologs with NM accessions and unambiguous gene pairs | Number of comparisons possible (either human or mouse gene is present in both data sets) | Percent of data set that can be compared | Percent of orthologs that are the same |
|---|---|---|---|---|---|
| JAX[a] | 8326 | 4116 | 1341 | 33 | 97 |
| HomoloGene[b] | 2692 | 2691 | 1337 | 50 | 97 |
| JAX[a] | 8326 | 4116 | 939 | 23 | 99 |
| Homol seq pairs[c] | 3460 | 1977 | 939 | 47 | 99 |
| HomoloGene[b] | 2692 | 2691 | 715 | 27 | 99 |
| Homol seq pairs[c] | 3460 | 1977 | 715 | 36 | 99 |
| JAX[a] | 8326 | 4116 | 3041 | 74 | 99 |
| Homology Map[d] | 9136 | 5434 | 3041 | 56 | 99 |
| JAX[a] | 8326 | 4116 | 2212 | 54 | 97 |
| This paper | 7645 | 5045 | 2216 | 44 | 97 |
| HomoloGene[b] | 2692 | 2691 | 1527 | 57 | 97 |
| This paper | 7645 | 5045 | 1534 | 30 | 97 |

[a]Mouse Genome Informatics Web Site, The Jackson Laboratory, Bar Harbor, Maine. (ftp.informatics.jax.org/pub/reports.html, 4/2003); at least two lines of evidence.
[b]NCBI Homologene (ftp.ncbi.nih.gov/pub/HomoloGene/hmlg.ftp; 4/2003).
[c]NCBI Homol_seq_pairs (ftp.ncbi.nih/gov/refseq/LocusLink/homol_seq_pairs.gz, 4/2003).
[d]NCBI Homology Map (www.ncbi.nlm.nih.gov/Homology, 4/2003).

scripts, but generated twice as many human–mouse alignments that passed quality control.

The multiple alignment program ClustalW, run with default parameters from the ClustalX package v1.83 α (Thompson et al. 1997), was used to align human, chimp, and mouse coding sequences. Mouse–human and chimp–human alignments were independently examined for the introduction of alignment gaps that would result in a frame-shifted human protein. Although these alignment gaps could represent real differences between two species, other causes (incorrect base calls, annotation, or ortholog inference) are more likely, especially between human vs. mouse alignments. Only alignments that produced either zero insertion/deletions, or those that had gaps whose length was a multiple of three bases were analyzed further. The alignment failure rate for chimp–human and mouse–human pairs was 3.3% and 31%, respectively (Table 1). The alignments passing quality control were converted into Phylip format (Felsenstein 1981) and are available at http://panther.celera.com/appleraHCM_alignments/index.jsp for download.

## EVOLUTIONARY MODELS

A commonly used measure to identify genes undergoing adaptive protein evolution involves comparing the ratio of nonsynonymous to synonymous substitution rates for each gene ($d_N/d_S$). If selection pressures favor proteins with altered amino acid sequence, then nonsynonymous changes are favored at the nucleotide level, and $d_N/d_S$ will be greater than 1. However, since humans and chimps have a relatively recent common ancestor, the overall sequence divergence is only about 1.2% (Chen and Li 2001), and coding regions show less than half this level of divergence (Shi et al. 2003). This results in wide variation from gene to gene in the absolute count of synonymous nucleotide changes, and low values of $d_S$ in the denominator result in large variance in $d_N/d_S$ ratios. Requiring that the predicted protein sequence must have diverged at more than one residue from the most parsimonious ancestral sequence can partly control this variance. There were 363 human genes (4.7% of the total) that had more than one amino acid difference and a $d_N/d_S > 1$. Formal statistical models to test for significant departure from a neutral evolutionary model were also applied to go beyond this ad hoc description of the genes.

A model specifying sequence divergence with parameters fitted by maximum likelihood (Felsenstein 1981) can be applied to multispecies sequence alignments and an evolutionary tree that describes the ancestral history of those sequences. This approach can be extended by allowing separate parameters specifying rates of synonymous and nonsynonymous substitution (Goldman and Yang 1994; Muse and Gaut 1994), variability among amino acid residues in their degree of constraint, and lineage-specific differences in the divergence rates (Yang and Nielsen 2002).

In the first of two evolutionary models applied to the set of 7,645 alignments, we applied a classical test of the null hypothesis of $d_N/d_S = 1$ in the human lineage (Nielsen and Yang 1998; Yang 2002). This test may be rejected if $d_N/d_S > 1$, showing evidence of positive selection, or $d_N/d_S < 1$, showing strong conservation of the gene in the human lineage. The neutral null hypothesis of model 1 was rejected by 72 genes (0.94%) at $p < 0.001$, 414 (5.4%) at $p < 0.01$, and 1216 genes (15.9%) at $p < 0.05$. There were 6 genes (0.08%) with $p < 0.05$ and $d_N/d_S > 1$.

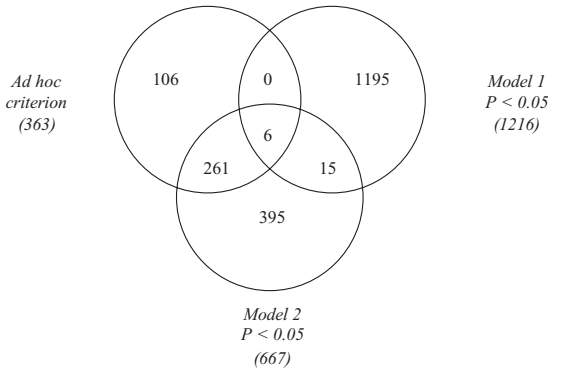


**Figure 1.** Overlap of human genes identified as exhibiting a signature of positive selection by the ad hoc criterion, a model testing for departure of $d_N/d_S$ from 1 (Model 1) and a model testing for excess nonsynonymous substitution within a domain of the protein in the human lineage only (Model 2).

The second formal model applied to test for positive selection is modified from Yang and Nielsen (2002) and allows variation in the $d_N/d_S$ ratio among lineages and among sites at the same time (see also Yang and Swanson 2002). In this method (Model 2), a likelihood ratio test of the hypothesis of neutrality is performed by comparing the likelihood values for two hypotheses. Under the null hypothesis, it is assumed that all sites are either neutral ($d_N/d_S = 1$) or evolve under negative selection ($d_N/d_S < 1$). Under the alternative hypothesis, some of the sites are allowed to evolve by positive selection in the human lineage only. The neutral null hypothesis of Model 2 was rejected by 28 genes (0.37%) at $p < 0.001$, 178 genes (2.3%) at $p < 0.01$, and 667 genes (8.7%) at $p < 0.05$.

The overlap between these three sets of genes is high (Fig. 1), but differences reflect the different attributes of the data that the tests consider. For example, small genes or genes with few substitutions may be flagged by the ad hoc criterion, but not attain statistical significance by the evolutionary models. Importantly, Model 2 can detect cases where a portion of the protein (perhaps a protein domain) is undergoing positive selection, but the overall $d_N/d_S$ may not be elevated, resulting in those genes being missed by the ad hoc criterion and by Model 1. For this reason, the remainder of the analysis considers only Model 2 test results.

## THE IMPACT OF LOCAL SEQUENCE COMPOSITION

Genome sequence composition such as GC content, gene density, repeat density, and local recombination rate can influence patterns and rates of sequence divergence (Hellmann et al. 2003; Webster et al. 2003). Before we go

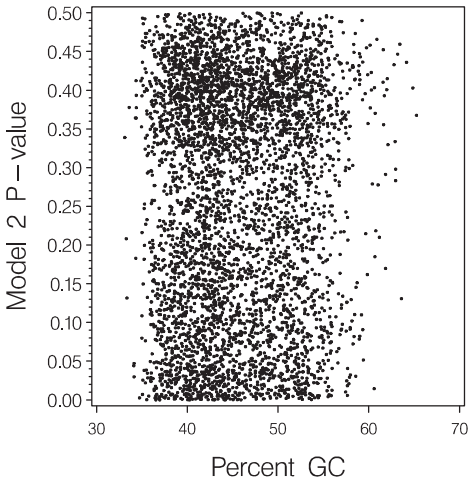


**Figure 2.** The relationship between local GC content for each of the 7,645 genes in this study and the Model 2 *p*-value, expressing the probability that the gene displays a signature of accelerated protein evolution in humans. The lack of correlation is one indication that base composition variation is not spuriously driving the test results.

into the details of which genes appear to exhibit unusual evolutionary patterns in the human lineage, it is important to test whether our model is particularly sensitive to local base composition changes. Although the maximum likelihood estimation should, in principle, take into account variation in base composition, it is easy to directly examine whether any residual correlation exists between GC content and the test results. For the 7,645 set the synonymous substitution rate was significantly correlated with: GC content (0.164, $p < 0.0001$), local recombination rate (Kong et al. 2002) (0.100, $p < 0.001$), and LINE element density (–0.091, $p < 0.0001$). None of these factors is significantly correlated with the nonsynonymous substitution rate or with Model 2 *p*-values (Fig. 2).

Segmental duplications (Bailey et al. 2002) do not appear to cause distortions in our analysis, since genes with close duplicates were underrepresented in our set due to the requirement of strict human–mouse orthology. This is confirmed by the observation that 10% of the genes in our set have at least one coding base pair in a segmental duplication (http://humanparalogy.gene.cwru.edu/SDD, UCSC Aug 2001 release) compared to 13% of genes in the entire human genome (Bailey et al. 2002). There is not an enrichment of segmental duplicated genes in the tail of the Model 2 *p*-value distribution.

## CLASSES OF GENES WITH ATYPICAL SEQUENCE DIVERGENCE

Given the large number of genes that exhibit a signature of natural selection along the human lineage, it becomes important to identify common features of these genes. A powerful approach to organizing such genetic information is to classify genes based on the inferred biological process in which they function or based on the molecular attributes of the gene product. In this analysis, we employ the assignment of the 7,645 genes into classes

**Table 3.** Biological Processes Showing Significant ($p < 0.01$) Positive Selection in the Human Lineage

| Biological Process | Number of genes[a] | P value[a] |
|---|---|---|
| Olfaction | 48 | 0 |
| Sensory perception | 146 (98) | 0 (0.026) |
| Cell surface receptor-mediated signal transduction | 505 (464) | 0 (0.0386) |
| Chemosensory perception | 54 (6) | 0 (0.1157) |
| Nuclear transport | 26 | 0.0003 |
| G-protein mediated signaling | 252 (211) | 0.0003 (0.1205) |
| Signal transduction | 1030 (989) | 0.0004 (0.0255) |
| Amino acid catabolism | 16 | 0.0041 |

[a]Excluding olfactory receptor genes.

based on biological processes and molecular functions using the PANTHER classification system (Thomas et al. 2003b; http://panther.celera.com), which is similar to categories in the Gene Ontology classification (http://www.geneontology.org). Functional classifications of genes were used only if the human protein sequence had a significant score to a PANTHER Hidden Markov Model (NLL-NULL score < –0.50). The accuracy of gene-function associations is shown to be comparable for well-curated model organism databases (Thomas et al. 2003a). For each functional category, a cumulative distribution of Model 2 *p*-values was compared to the cumulative distribution of all genes using the Mann-Whitney U test. Categories of biological processes and molecular functions with $p < 0.01$ under this Mann-Whitney U test were considered significant.

In the human lineage, genes involved in two different biological process, olfaction and amino acid catabolism, have a significant tendency to show human-specific accelerations of protein evolution (Table 3, Fig. 3). The olfaction biological process contains mostly olfactory receptors (OR), and it is reasonable to hypothesize that the different lifestyles of human, mouse, and chimp might lead to selective pressure on these genes. Since there has been a rapid rate of loss of function of OR genes in hu-

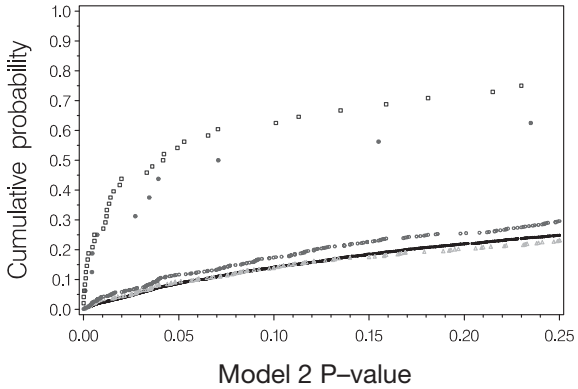


**Figure 3.** Model 2 *p*-value distributions of selected groups of genes. The plot gives the cumulative fraction of selected biological processes showing the excess of cases of significant positive selection in genes for olfaction (*open squares*), amino acid catabolism (*closed circles*), and Mendelian disease genes (*open circles*) relative to the overall distribution of genes (*dots*, fused into one line). The distribution of developmental genes (*open triangles*) that do not show significant excess is shown for comparison.

mans (Gilad et al. 2003), which would be expected to show increased nonsynonymous substitution, we verified that most of the OR genes in our set are bona fide genes (http://bioinformatics.weizmann.ac.il/HORDE). Our results, suggesting that many of the still active OR genes display a signature of positive selection, are supported by the observation that there is a discordance between levels of human polymorphisms in many OR genes from interspecific divergence (Gilad et al. 2003).

Genes involved in amino acid catabolism also show evidence of adaptive evolution. It is possible that the radical change in diet between human and chimps might be partly responsible for this pattern of divergence. Of the eight protein catabolism genes (GSTZ1, HGD, PAH, BCKDHA, PCCB, HAL, ALDH6A1, AMT) with the lowest Model 2 *p*-values (http://panther.celera.com/appleraHCM_alignments/index.jsp), all have been implicated in human metabolic disorders. In fact, there is a significant tendency ($p < 0.001$, Kolmogorov-Smirnov test) for genes implicated in human Mendelian disorders (http://www.ncbi.nlm.nih.gov/htbin-post/Omim/getmorbid) to exhibit significant ($p < 0.01$) Model 2 *p*-values.

## GENES WITH ATYPICAL SEQUENCE DIVERGENCE

Of the 667 genes showing evidence of adaptive evolution under Model 2 ($p < 0.05$), there are several categories of genes, such as developmental, hearing, and reproductive, that are particularly interesting considering the physiological differences between humans and chimps (Table 4). Considering the importance and general conservation across evolution of these traits, it is not surprising that the classes as a whole do not show adaptive evolution (Fig. 3). However, given the important functions of genes in these classes, each one may account disproportionately for specific phenotypic differences.

Most of the human developmental genes that appear to

**Table 4.** Selected Genes Showing Evidence of Adaptive Evolution in the Human Lineage

| Biological process | Gene symbol | Gene name | Model 2 *p*-value |
|---|---|---|---|
| Development | ALPL | alkaline phosphatase, liver/bone/kidney | 3.69E-03 |
| | BMP4 | bone morphogenetic protein 4 | 2.50E-02 |
| | CDX4 | caudal type homeobox transcription factor 4 | 1.57E-03 |
| | DIAPH1 | diaphanous homolog 1 (*Drosophila*) | 2.01E-02 |
| | EPHB6 | EphB6 | 3.53E-02 |
| | EYA1 | eyes absent homolog 1 (*Drosophila*) | 6.49E-03 |
| | EYA4 | eyes absent homolog 4 (*Drosophila*) | 3.48E-03 |
| | FOXI1 | forkhead box I1 | 2.87E-03 |
| | FOXP2 | forkhead box P2 | 2.67E-03 |
| | HOXA5 | homeobox A5 | 2.16E-02 |
| | HOXD4 | homeobox D4 | 4.55E-03 |
| | MEOX2 | mesenchyme homeobox 2 | 3.45E-02 |
| | MGP | matrix Gla protein | 3.94E-02 |
| | MIXL1 | Mix1 homeobox-like 1 (*Xenopus laevis*) | 4.61E-02 |
| | MMP20 | matrix metalloproteinase 20 (enamelysin) | 3.18E-02 |
| | NEUROG1 | neurogenin 1 | 3.57E-02 |
| | NLGN3 | neuroligin 3 | 2.80E-02 |
| | NTF3 | neurotrophin 3 | 8.92E-03 |
| | OTOR | otoraplin | 3.46E-02 |
| | PHTF | putative homeodomain transcription factor 1 | 5.24E-02 |
| | PLXNC1 | plexin C1 | 5.93E-03 |
| | POU2F3 | POU domain, class 2, transcription factor 3 | 3.34E-02 |
| Development | SEMA3B | Semaphorin 3B | 3.57E-03 |
| | SIM2 | single-minded homolog 2 (*Drosophila*) | 4.57E-02 |
| | SNAI1 | snail homolog 1 (*Drosophila*) | 6.74E-03 |
| | TECTA | tectorin alpha | 1.57E-05 |
| | TLL2 | tolloid-like 2 | 3.28E-03 |
| | TRAF5 | TNF receptor-associated factor 5 | 6.35E-04 |
| | WHN | winged-helix nude | 5.57E-04 |
| | WIF1 | WNT inhibitory factor 1 | 2.71E-02 |
| | WNT2 | wingless-type MMTV integration site family member 2 | 2.99E-02 |
| Amino acid catabolism | ALDH6A1 | aldehyde dehydrogenase 6 family, member A1 | 2.72E-02 |
| | AMT | aminomethyltransferase (glycine cleavage system protein T) | 3.45E-02 |
| | BCKDHA | branched chain keto acid dehydrogenase E1α polypeptide | 4.64E-03 |
| | GSTZ1 | glutathione transferase ζ 1 (maleylacetoacetate isomerase) | 9.04E-04 |
| | HAL | histidine ammonia-lyase | 7.18E-03 |
| | HGD | homogentisate 1,2-dioxygenase (homogentisate oxidase) | 4.39E-03 |
| | PAH | phenylalanine hydroxylase | 7.08E-02 |
| | PCCB | propionyl Coenzyme A carboxylase, beta polypeptide | 3.93E-02 |
| Reproduction | GNRHR | gonadotropin-releasing hormone receptor | 8.48E-03 |
| | MTNR1A | melatonin receptor 1A | 3.97E-02 |
| | PAPPA | pregnancy-associated plasma protein A | 8.18E-04 |
| | PGR | progesterone receptor | 1.05E-03 |

be under adaptive evolution fall into four main categories: skeletal development, neurogenesis, reproduction, and homeotic transcription factor genes (Table 4). For example, the homeotic transcript factor genes CDX4, HOXA5, HOXD4, MEOX2, POU2F3, MIXL1, and PHTF play key roles in early development and have Model 2 *p*-values less than 0.05. TRAF5 plays a key role in osteoclast proliferation and may be implicated in accelerated growth of the long bones in the leg (Kanazawa et al. 2003). TRAF5 shows adaptive evolution along with six other skeletal developmental genes. At least 10 genes involved in neurogenesis processes, including axonal guidance and synapse remodeling, have low Model 2 *p*-values. For example, the SIM2 transcription factor has been implicated in human Down syndrome and memory defects in mice (Chrast et al. 2000). FOXN1, or winged helix nude, encodes a transcription factor involved in keratin gene expression. Mutations in this gene cause athymia, resulting in a severely compromised immune system. Developmental defects in *Drosophila* and *Caenorhabditis elegans* are also observed when this gene is mutated. A plausible hypothesis is that the relative hairlessness of humans compared to chimps is in part determined by FOXNI. Hypotheses like this are generated in abundance by studies such as this, and an exciting aspect of the work is that such hypotheses are amenable to future testing.

The anatomy and physiology of reproduction are strikingly different between humans and chimpanzees. Several genes involved in pregnancy appear to exhibit nonneutral evolution (Table 4). For example, the progesterone receptor (PGR) is involved with maintenance of the uterus and may be involved in the acrosome reaction (Gadkar et al. 2003). The reproductive hormone receptors GNRHR and MTNR1A also have significant Model 2 *p*-values.

Several genes associated with the development of hearing appear to have undergone adaptive evolution (Table 4). α-Tectorin, which shows the most significant Model 2 *p*-value, plays an important role in the tectorial membrane of the inner ear. When it is mutated, humans show high-frequency hearing loss (Mustapha et al. 1999) and mouse knockout mice are deaf. Other genes under human-specific selection, DIAPH1, FOXI1, and EYA4, cause hearing loss in humans when mutated.

## CONCLUSIONS

There has been considerable interest in obtaining the genome sequence of the chimpanzee, our closest relative, because of the notion that, by comparing our two genomes, it might be possible to infer which genetic differences are responsible for the morphological, physiological, and behavioral factors that differentiate us. At 1% sequence divergence, however, we expect there to be roughly 3 million base pairs of sequence difference, and the discrimination between substitutions that are totally unimportant and substitutions that are causal to our biological differences appears to be a steep challenge. Fortunately, the phylogenetic approach offers a promise to make progress on this problem. With multiple related species arranged on a phylogenetic tree, models of molecular evolution can place the mutations on particular lineages of the tree. This information can be used to infer what DNA sequence changes have occurred specifically along the lineage subsequent to the node representing our common ancestry with chimpanzee, and reflecting changes that occurred in our line of descent since that time. The challenge that remains is that many of these changes will have arisen purely because our population size is finite, and because random mutations, provided they are not too deleterious, may go to fixation by random drift. It is humbling to consider that potentially a large portion of the genomic differences between humans and chimps have arisen by such a purely neutral process. If one asks, "What are the genes that make us human?", these random changes may surely be an important class of genes that carry this label.

In this paper, we apply methods that have been used by many others to infer which genes have been undergoing positive or adaptive evolution. The idea is based on the relative rates of substitution at silent (synonymous) and at replacement (nonsynonymous) nucleotide positions in the gene. Strictly neutral genes are expected to have equal rates of substitution for these two classes of sites, while most genes have some selective constraint and show considerable deficit of replacement changes. Formal statistical approaches allow us to test the null hypothesis that the changes are compatible with neutrality, and to make quite incisive tests of alternative hypotheses about the way that selection has acted (Yang and Nielsen 2002).

The application of this inferential approach identified a long list of genes for which it is all too easy to tell an evolutionary story about how these genes are important for human–chimp differences. However, it should be emphasized that these approaches are strictly exploratory and that they really only highlight hypotheses to be tested by additional data collection at several levels. The finding of positive selection in genes such as α-tectorin suggests that there may be differences in the hearing acuity of humans and chimpanzees, and the data available on the subject are too sparse to properly address the issue. This motivates specific hearing tests of chimpanzees, with the idea that aspects of vocal speech may place additional requirements on hearing not faced by speechless chimpanzees. For every gene cited in this paper, there are additional experiments that must be done to solidify the evidence that these genes may be involved in human–chimpanzee differences. In the case of significant differences in biological processes, the case based only on DNA sequences becomes more compelling, mostly because each test involves many genes showing aberrant (nonneutral) behavior. That amino acid catabolism should be a biological process showing rapid adaptive evolution suggests further research into the physiology of digestion of low- versus high-protein diets, and consideration of the differences among primates in diet. Dietary changes are not the only thing that might be driving this difference. Demands on protein synthesis during brain development might be the driver of this signal of past natural selection. Despite all these uncertainties, and the

worry that this approach only raises possibilities rather than proving anything, it seems nearly a certainty that comparative genomic methods like this will serve as a powerful generator of hypotheses that will admit further analysis of the differences in all levels of biological function between humans and our near relatives.

## ACKNOWLEDGMENTS

We thank the employees of the Celera Genomics sequencing center for their excellent technical participation; J. Duff, C. Gire, M.A. Rydland, C. Forbes, and B. Small for development and maintenance of software systems, laboratory information management systems, and analysis programs. We also thank S. Hannenhalli and S. Levy for helpful discussions.

## REFERENCES

Bailey J.A., Gu Z., Clark R.A., Reinert K., Samonte R.V., Schwartz S., Adams M.D., Myers E.W., Li P.W., and Eichler E.E. 2002. Recent segmental duplications in the human genome. *Science* **297:** 1003.

Chen F.C. and Li W.-H. 2001. Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. *Am. J. Hum. Genet.* **68:** 444.

Chrast R., Scott H.S., Madani R., Huber L., Wolfer D.P., Prinz M., Aguzzi A., Lipp H.P., and Antonarakis S.E. 2000. Mice trisomic for a bacterial artificial chromosome with the single-minded 2 gene (Sim2) show phenotypes similar to some of those present in the partial trisomy 16 mouse models of Down syndrome. *Hum. Mol. Genet*. **9:** 1853.

Ebersberger I., Metzler D., Schwarz C., and Pääbo S. 2002. Genomewide comparison of DNA sequences between humans and chimpanzees. *Am. J. Hum. Genet*. **70:** 1490.

Felsenstein J. 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. *J. Mol. Evol.* **17:** 368.

Frazer K.A., Chen X., Hinds D.A., Pant P.V., Patil N., and Cox D.R. 2003. Genomic DNA insertions and deletions occur frequently between humans and nonhuman primates. *Genome Res.* **13:** 341.

Gadkar S., Shah C.A., Sachdeva G., Samant U., and Puri C.P. 2003. Progesterone receptor as an indicator of sperm function. *Biol. Reprod*. **67:** 1327.

Gilad Y., Man O., Pääbo S., and Lancet D. 2003. Human specific loss of olfactory receptor genes. *Proc. Natl. Acad. Sci.* **100:** 3324.

Goldman N. and Yang Z. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.* **11:** 725.

Hellmann I., Zöllner S., Enard W., Ebersberger I., Nickel B., and Pääbo S. 2003. Selection on human genes as revealed by comparisons to chimpanzee cDNA. *Genome Res.* **13:** 831.

Kanazawa K., Azuma Y., Nakano H., and Kudo A. 2003. TRAF5 functions in both RANKL- and TNFalpha-induced osteoclastogenesis. *J. Bone Miner. Res*. **18:** 443.

King M.C. and Wilson A.C. 1975. Evolution at two levels in humans and chimpanzees. *Science* **188:** 107.

Kong A., Gudbjartsson D.F., Sainz J., Jonsdottir G.M., Gudjonsson S.A., Richardsson, B., Sigurdardottir S., Barnard J., Hallbeck B., Masson G., Shlien A., Palsson S.T., Frigge M.L., Thorgeirsson T.E., Gulcher J.R., and Stefansson K. 2002. A high-resolution recombination map of the human genome. *Nat. Genet*. **31:** 241.

Mural R.J., Adams M.D., Myers E.W., Smith H.O., Miklos G.L., Wides R., Halpern A., Li P.W., Sutton G.G., Nadeau J., Salzberg S.L., Holt R.A., Kodira C.D., Lu F., Chen L., Deng Z., Evangelista C.C., Gan W., Heiman T.J., Li J., Li Z., Merkulov G.V., Milshina N.V., Naik A.K., and Qi R., et al. 2002. A comparison of whole-genome shotgun-derived mouse chromosome 16 and the human genome. *Science* **296:** 1661.

Muse S.V. and Gaut B.S. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Mol. Biol. Evol.* **11:** 715.

Mustapha M., Weil D., Chardenoux S., Elias S., El-Zir E., Beckmann J.S., Loiselet J., and Petit C. 1999. An alpha-tectorin gene defect causes a newly identified autosomal recessive form of sensorineural pre-lingual non-syndromic deafness, DFNB21. *Hum. Mol. Genet*. **8:** 409.

Nielsen R. and Yang Z. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* **148:** 929.

Shi J., Xi H., Wang Y., Zhang C., Jiang Z., Zhang K., Shen Y., Jin L., Zhang K., Yuan W., Wang Y., Lin J., Hua Q., Wang F., Xu S., Ren S., Xu S., Zhao G., Chen Z., Jin L., and Huang W. 2003. Divergence of the genes on human chromosome 21 between human and other hominoids and variation of substitution rates among transcription units. *Proc. Natl. Acad. Sci.* **100:** 8331.

Thomas P.D., Campbell M.J., Kejariwal A., Mi H., Karlak B., Daverman R., Diemer K., Muruganujan A., and Narechania A. 2003a. PANTHER: A library of protein families and subfamilies indexed by function. *Genome Res.* **13:** 2129.

Thomas P.D., Kejariwal A., Campbell M.J., Mi H., Diemer K., Guo N., Ladunga I., Ulitsky-Lazareva B., Muruganujan A., Rabkin S., Vandergriff J.A., and Doremieux O. 2003b. PANTHER: A browsable database of gene products organized by biological function, using curated protein family and subfamily classification. *Nucleic Acids Res*. **31:** 334.

Thompson J.D., Gibson T.J., Plewniak F., Jeanmougin F., and Higgins D.G. 1997. The ClustalX windows interface: Flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* **24:** 4876.

Venter J.C., Adams M.D., Myers E.W., Li P.W., Mural R.J., Sutton G.G., Smith H.O., Yandell M., Evans C.A., Holt R.A., Gocayne J.D., Amanatides P., Ballew R.M., Huson D.H., Wortman J.R., Zhang Q., Kodira C.D., Zheng X.H., Chen L., Skupski M., Subramanian G., Thomas P.D., Zhang J., Gabor Miklos G.L., and Nelson C., et al. 2001. The sequence of the human genome. *Science* **291:** 1304.

Webster M.T., Smith N.G., and Ellegren H. 2003. Compositional evolution of noncoding DNA in the human and chimpanzee genomes. *Mol. Biol. Evol.* **20:** 278.

Yang Z. 2002. Inference of selection from multiple species alignments. *Curr. Opin. Genet. Dev.* **12:** 688.

Yang Z. and Nielsen R. 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol. Biol. Evol.* **19:** 908.

Yang Z. and Swanson W.J. 2002. Codon-substitution models to detect adaptive evolution that account for heterogeneous selective pressures among site classes. *Mol. Biol. Evol.* **19:** 49.


Cold Spring Harbor Symposia on Quantitative Biology

# Positive Selection in the Human Genome Inferred from Human–Chimp–Mouse Orthologous Gene Alignments

A.G. CLARK, S. GLANOWSKI, R. NIELSEN, et al.

*Cold Spring Harb Symp Quant Biol* 2003 68: 479-486
Access the most recent version at doi:10.1101/sqb.2003.68.479