An Empirical Test for Branch-Specific Positive Selection

Gabrielle C. Nickel, David L. Tefft,[1] Karrie Goglin[2] and Mark D. Adams[3]

Department of Genetics, Case Western Reserve University, Cleveland, Ohio 44106

ABSTRACT

The use of phylogenetic analysis to predict positive selection specific to human genes is complicated by the very close evolutionary relationship with our nearest extant primate relatives, chimpanzees. To assess the power and limitations inherent in use of maximum-likelihood (ML) analysis of codon substitution patterns in such recently diverged species, a series of simulations was performed to assess the impact of several parameters of the evolutionary model on prediction of human-specific positive selection, including branch length and d_N/d_S ratio. Parameters were varied across a range of values observed in alignments of 175 transcription factor (TF) genes that were sequenced in 12 primate species. The ML method largely lacks the power to detect positive selection that has occurred since the most recent common ancestor between humans and chimpanzees. An alternative null model was developed on the basis of gene-specific evaluation of the empirical distribution of ML results, using simulated neutrally evolving sequences. This empirical test provides greater sensitivity to detect lineage-specific positive selection in the context of recent evolutionary divergence.

A key question facing modern science is, "What makes humans different from the other great apes?" Sequencing of the human genome has heightened the interest of scientists in understanding the origins of our species and the genetic basis for traits that distinguish us from our closest living relatives, chimpanzees. Sequencing of the chimpanzee genome revealed ~1% fixed single-nucleotide differences with the human genome and ~1.5% lineage-specific DNA in each species (CHENG *et al.* 2005; CHIMPANZEE SEQUENCING AND ANALYSIS CONSORTIUM 2005). Even though the percentage of divergence is small, there are ~35 million single-nucleotide changes and 5 million insertion/deletion differences between human and chimpanzee, most of which are presumed to be evolutionarily neutral (CHIMPANZEE SEQUENCING AND ANALYSIS CONSORTIUM 2005). Thus the complete catalog of human–chimpanzee genome differences is insufficient to reveal the genetic basis for phenotypic divergence between the species.

Several approaches have been presented to the task of identifying alleles that have been subject to positive selection—that is, that have increased in frequency in a population as a result of adaptive evolutionary advantage. These include methods based on allele-frequency differences among human populations (TAJIMA 1993), extended haplotype homozygosity indicative of a recent selective sweep (SABETI *et al.* 2002), a discrepancy between rates of polymorphism and divergence (MCDONALD and KREITMAN 1991), and phylogenetic methods that compare the rates of evolution on different branches of the tree (FELSENSTEIN 1981; GOLDMAN and YANG 1994). Several studies have characterized positive selection in human genes with a goal of identifying the genetic basis of human-specific features such as a larger brain size, bipedalism, and skeletal and dental differences, etc. (VALLENDER and LAHN 2004). Likewise, genomewide surveys have identified genes, gene families, and genome regions with a signature of positive selection (CLARK *et al.* 2003; DORUS *et al.* 2004; BUSTAMANTE *et al.* 2005; CHIMPANZEE; SEQUENCING AND ANALYSIS CONSORTIUM 2005; CLARK and SWANSON 2005; ARBIZA *et al.* 2006; VOIGHT *et al.* 2006; BAKEWELL *et al.* 2007; GIBBS *et al.* 2007). We are most interested in identifying genetic differences associated with fundamental and ancient divergence between humans and the nonhuman primates and so have focused on the phylogenetic approach to identify positive selection that is unique to the human lineage.

The nature of selection acting on a protein-coding sequence is characterized by comparison of the rate of nonsynonymous (d_N) and synonymous (d_S) substitutions. Under neutral evolution, the rate of accumulation of synonymous and nonsynonymous substitutions should be equal and a d_N/d_S value of 1 is expected. A d_N/d_S value >1 is indicative of positive selection, while a value <1 indicates negative selection. About one-third

[1] *Present address:* The Broad Institute, Cambridge, MA 02141.

[2] *Present address:* The Scripps Research Institute, La Jolla, CA 92037.

[3] *Corresponding author:* Department of Genetics, Case Western Reserve University, 10900 Euclid Ave., Cleveland, OH 44106-4955. E-mail: markadams@case.edu

of chimpanzee protein sequences are identical to the sequence of the human ortholog, with the remainder having typically one or two amino acid changes (Chimpanzee Sequencing and Analysis Consortium 2005). Clearly, positive selection cannot be inferred when two protein-coding sequences are identical. But how different must two sequences be for the currently available methods to detect statistically significant positive selection?

Statistical methods based on models of codon substitution patterns have been developed to identify positive selection specific to one branch in a phylogeny, even when only a subset of sites have experienced selection (Yang and Nielsen 2002). Simulation studies using the maximum-likelihood methods implemented in phylogenetic analysis by maximum likelihood (PAML) (Yang 1997) have shown that the approach is conservative with few false positives, but also demonstrated a lack of power, particularly in analyzing closely related sequences (Wong *et al.* 2004; Zhang *et al.* 2005). We were particularly interested in how the maximum-likelihood codon substitution models would perform for analysis of selection on human proteins in comparison to chimpanzee, where there are few divergent positions.

Genomewide studies to date have been based on data from only two (Chimpanzee Sequencing and Analysis Consortium 2005; Nielsen *et al.* 2005; Arbiza *et al.* 2006) or three species (Clark *et al.* 2003; Bakewell *et al.* 2007; Gibbs *et al.* 2007). In the context of human–chimpanzee–macaque gene alignments, only a handful of genes were predicted to be positively selected in humans after multiple-test correction (Bakewell *et al.* 2007; Gibbs *et al.* 2007). Analysis of sequence from a broader set of primate species would place human–chimpanzee differences into a more robust phylogenetic framework and enable more specific analysis of evolutionary rates in each primate lineage. By improving our understanding of the factors that contribute to detection of positive selection, it should be possible to optimize the design of computational experiments to test for positive selection and potentially the selection of additional organisms for genome sequencing. Therefore, we collected sequence data from 12 primate species to provide a more comprehensive view of the phylogenetic breadth that is required to detect positive selection, to explore the patterns of codon substitution that accompany predictions of positive selection, to study the factors that affect a prediction of positive selection, and to examine the evolutionary characteristics of a group of important genes.

Predictions of positive selection in recently diverged species, such as humans and chimpanzees, are often difficult using traditional phylogenetic methods. Simulation studies were performed to gauge the factors that most affect predictions of selection; even under the best circumstances, the extent of divergence found between

human and chimpanzee genes was not sufficient to meet the strict criteria for prediction of selection in the maximum-likelihood (ML) test. Through these studies we developed an empirical test that is more sensitive to the often subtle changes that have occurred along the human lineage and that provides the opportunity to identify additional candidates for positive selection that would not be predicted using currently available methods.

Developmentally regulated transcription factor genes were chosen for study because of the fundamental role that many of these genes play in a broad array of developmental processes, because of the ability of small changes in sequence to cause broad changes in downstream gene expression patterns (*e.g.*, Ting *et al.* 1998; Grenier and Carroll 2000), and because transcription factors (TFs) are overrepresented among genes predicted to be positively selected in previous genomewide studies of selection (Bustamante *et al.* 2005; Chimpanzee Sequencing and Analysis Consortium 2005; Arbiza *et al.* 2006).

MATERIALS AND METHODS

Selection of genes and DNA sequencing: Transcription factor genes were selected for sequencing on the basis of a prior prediction of positive selection (Clark *et al.* 2003; Bustamante *et al.* 2005; Nielsen *et al.* 2005) or involvement in developmental processes (gene ontology term GO:0032502) having an excess of nonsynonymous substitution relative to synonymous substitution in a comparison of human and chimpanzee genome annotations. Primate DNA samples were obtained from the National Institute on Aging Cell Repository Phylogenetic Primate Panel (PRP0001). Human DNA samples were from the National Institute of General Medical Sciences Human Genetic Cell Repository Human Variation Collection (NA00946, NA00131, NA01814, NA14665, and NA03715). Both panels were obtained through Coriell Cell Repositories. Hylobates and Papio samples were kindly provided by Evan Eichler. Primer design and PCR amplification of exons and DNA sequencing were performed as described (Kidd *et al.* 2005). When alternative transcripts were present, the longest isoform was chosen for sequencing. Sequence was obtained from both strands of each PCR product and trace files were processed by phred and phrap (Ewing and Green 1998; Ewing *et al.* 1998; http://www.phrap.org), with each exon from each organism assembled independently. Low-quality bases ($Q < 25$) in the consensus were converted to N's. Sequence from each species was aligned to the human genomic DNA using *matcher* (an implementation of the Pearson *lalign* algorithm from the EMBOSS package) and the location of exons inferred from coordinates in the reference human sequence. Mouse and rat orthologs were obtained from Mouse Genome Informatics (http://www.informatics.jax.org) and aligned to the human sequence using *transAlign* (Bininda-Emonds 2005). Rhesus macaque and orangutan alignments were supplemented with sequences extracted from the NCBI Trace Archive when our data did not cover the entire coding region. Chimpanzee alignments were supplemented with orthologous segments extracted from the chimpanzee draft genome assembly (Chimpanzee Sequencing and Analysis Consortium 2005), Build 1. Multiple sequence alignments of exonic regions were then constructed for each exon, again in reference to the

human sequence, using *matcher* and *sim4* (FLOREA *et al.* 1998), and exons were assembled to form the entire coding sequence of the gene. The resulting alignments were reformatted to serve as input to the phylogenetic analysis programs. The total sequence from each source (generated for this study and previously publicly available) and the average gene coverage for each organism are given in supplemental Table S1.

Phylogenetic analysis: Codeml from the PAML package (v3.14b, YANG 1997) was run on an Apple OS10.5 computer cluster and results were stored in a custom-built MySQL database. The codeml models used and the tests performed are summarized in supplemental Tables S2 and S3. Each gene was first tested by asking whether a model in which the human lineage was allowed to have a different d_N/d_S ratio than the rest of the tree was a better fit to the data than a model with a single d_N/d_S ratio throughout the tree ("branch test"). Whether the human-specific d_N/d_S ratio was significantly >1 was tested by comparison to a model with the d_N/d_S ratio fixed at 1 on the human branch ("strict branch test"). The optimized branch-site method of Yang and Nielsen (YANG and NIELSEN 2002; YANG *et al.* 2005; ZHANG *et al.* 2005) was used with nested null models to test whether positive selection might be occurring at a subset of sites (codons) on the human lineage by comparison of results from model A *vs.* model A-null ("strict branch-site test") and model A *vs.* model 1a ("relaxed branch-site test"). The strict branch-site test requires $d_N/d_S > 1$ at a subset of sites, while the relaxed branch-site test requires only that a subset of sites on the human lineage have a significantly elevated d_N/d_S ratio compared to those sites on the remainder of the tree and can thus be significant in the presence of either positive selection or a relaxation of selective constraint (ZHANG *et al.* 2005). A consistent species tree was used for each analysis (Figure 1). To estimate significance in the form of a *P*-value, a likelihood-ratio test (LRT) was used. Two times the difference in log-likelihood values between two models was compared to a chi-square distribution with degrees of freedom equal to the difference in number of parameters for the two models being compared. All results with $P < 0.05$ in the LRT were repeated to ensure that the result was robust to variation in the initial ML conditions or failure of the ML estimation to converge to the global optimum. A complete listing of genes with selected results is given as supplemental Table S4. All primary sequence data have been submitted to GenBank; alignments and codeml results are available at http://mendel.gene.cwru.edu/adamslab/pbrowser.py.

Simulations testing the performance of codeml: The sensitivity of the strict branch-site test to branch length, proportion of sites within the positively selected site classes, background d_N/d_S, and the magnitude of the foreground d_N/d_S value was examined by using the NSbranchsites function in the *evolver* program contained in the PAML software package to create simulated sequences under a specified evolutionary model. Codon frequencies were derived from a set of 11,485 human genes with high-quality orthologs in several mammalian species (NICKEL *et al.* 2008). These genes were divided into quartiles of G + C content and codon frequencies were derived for each quartile. The four quartiles were <45.2%, 45.2–52.19%, 52.2–58.79%, and ≥58.8%. It has recently been shown that variations in substitution pattern as a result of variance in the recombination rate can essentially be accounted for by G + C content (BULLAUGHEY *et al.* 2008). The specifications of each simulation can be found in supplemental Table S5. Branch lengths were calculated by concatenating the alignments for the 175 transcription factors and evaluating the alignment using DNAml of the PHYLIP program package (FELSENSTEIN 1993). Evolver was used to create alignment sets of coding sequences with 400 codons

from each G + C quartile separately. One thousand simulated sequence sets were generated for models of neutral evolution, while 200 simulated sequences were created for models reflecting positive selection along the human lineage. The resulting multispecies sequence alignments were analyzed using codeml with model A and model A-null.

Empirical tests of positive selection using sequences simulated under a model of neutral evolution: An alternative null model was developed to assess the likelihood of positive selection by comparing the actual results from the strict branch-site test (codeml model A *vs.* model A-null) to results obtained using sequences simulated under a model of neutral evolution. For each gene tested, the sequence of the human–chimpanzee ancestor was inferred using the RateAncestor function of codeml. More than 99% of the codons were inferred with at least 95% probability, according to codeml (supplemental Figure S1). These ancestral sequences were input as the root sequence in evolver, and at least 500 sequences were simulated under a model of neutral evolution (foreground $d_N/d_S = 1$ in site classes 2a and 2b). The background branch lengths were calculated as described above, and the human branch length used was that reported by codeml using the branch model H. The proportion of sites within site classes 0, 1, 2a, and 2b was held constant at 0.6, 0.1, 0.26, 0.04, respectively, on the basis of averages among genes with $d_N/d_S > 1$ (see supplemental Table S3). The transition/transversion ratio, κ, was also held constant at 2.6. The average κ for the 175-gene set was 4.1 ± 1.5. A higher κ means more transitions and therefore more synonymous changes. A smaller κ is therefore more conservative. Codon substitution frequencies were derived from the appropriate G + C quartile for the specific gene. κ = 2.6 is within one standard deviation below the mean for each G + C quartile based on the 11,485-gene set.

Gene-specific empirical *P*-values were calculated by running codeml with model A and model A-null on each of the simulated neutral human sequences, using the original primate plus rodent multiple-sequence alignments for each gene and calculating the likelihood-ratio statistic (LRS). The empirical *P*-value is defined as the fraction of sequences simulated under a model of neutral evolution that had a LRS greater than or equal to the LRS obtained with the actual data. The gene-specific significance threshold is based on the empirical distribution of LRS values, using a parametric bootstrap to specify a significance threshold of α = 5% (GOLDMAN 1993; AAGAARD and PHILLIPS 2005).

RESULTS

Sequencing of transcription factor genes and tests of selection: We sought to identify evidence of adaptive protein evolution specific to the human lineage with the hypothesis that genes under positive selection have contributed to the phenotypic changes that occurred since the most recent common ancestor of humans and chimpanzees. A total of 175 transcription factor genes with prior predictions of positive selection (CLARK *et al.* 2003; BUSTAMANTE *et al.* 2005; CHIMPANZEE SEQUENCING AND ANALYSIS CONSORTIUM 2005; NIELSEN *et al.* 2005) or an excess of nonsynonymous substitution in human relative to chimpanzee were selected for phylogenetic sequencing. Exonic sequence data were obtained from human, chimpanzee, bonobo, gorilla, orangutan, gibbon, three Old World monkeys, three New World monkeys,
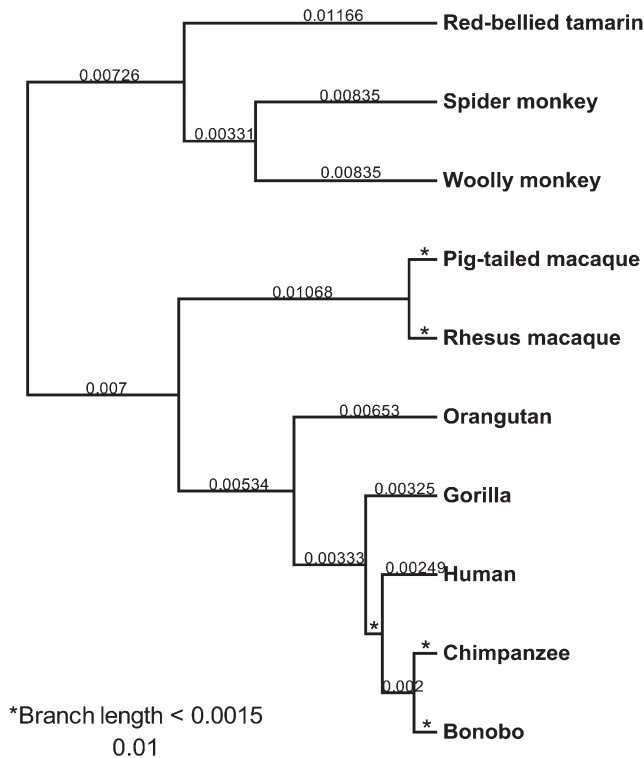
FIGURE 1.—Phylogenetic tree of the primate species used in phylogenetic analyses. Primates were chosen to include a representative from all the major subdivisions of primates. These include Parvorder Platyrrhini (New World monkeys: red-bellied tamarin, spider monkey, woolly monkey), Parvorder Catarrhini (Old World monkeys: pig-tailed macaque, rhesus macaque), Family Hylobatidae (lesser apes: orangutan), and Family Hominidae (great apes: gorilla, chimpanzee, bonobo, and human). Suborder Strepsirrhini (prosimians: lemur) is not depicted in the tree. Mouse and rat were used as an outgroup (not shown). Branch lengths were calculated from concatenated alignments of the 175 transcription factor genes using the DNAML program in PHYLIP (FELSENSTEIN 1993). *Branch length <0.0015.

and lemur (Figure 1). Primary sequence read data from chimpanzee, macaque, and orangutan were obtained from the NCBI Trace Repository, supplementing the alignments where possible, and annotated mouse and rat orthologs (http://www.informatics.jax.org) were added to the alignments (supplemental Table S1). Coverage is excellent for the great apes, rodents, and rhesus macaque, but lower for the monkeys, reflecting a lower rate of PCR success in more divergent species.

Multispecies alignments were constructed of the protein-coding DNA sequence of each gene and the codon substitution pattern in the alignments was evaluated under several evolutionary models as implemented in the codeml program from the PAML package (YANG 1997). In each case, results from a null model were compared with results from a test model. The null models assumed neutrality ($d_N = d_S$) or negative selection ($d_N/d_S < 1$) and the test models assumed $d_N/d_S > 1$

on the human lineage (strict branch test) or at a subset of codon sites in the gene along the human branch (strict branch-site test). A list of the models and tests of selection is presented in supplemental Tables S2 and S3.

Only three genes have $P < 0.05$ in the strict branch-site test: *ISGF3G*, *SOHLH2*, and *ZFP37* (Table 1). Seventeen genes have $P < 0.05$ in the relaxed branch-site test alone, indicating that they may have experienced a relaxation of selective constraint or simply show weak signals of positive selection. (supplemental Table S4). None of the nominal *P*-values survive Bonferroni correction, and a false discovery rate analysis (BENJAMINI and HOCHBERG 1995) also predicts that the likelihood of observing true positives is small.

Simulations to assess the sensitivity of the strict branch-site test: The failure to observe significant positive selection was surprising, given the method for choosing genes and the fact that more than one-third have a d_N/d_S ratio >1 on the human branch, and prompted us to wonder whether the genes we chose were truly not positively selected or whether the ML method lacks power to detect positive selection for these closely related species. We therefore embarked on a series of simulations to study the performance of codeml, using branch lengths and substitution patterns that are typical of the divergence time of humans from non-human primates.

In presenting the strict branch-site test, ZHANG *et al.* (2005) performed an extensive series of simulations of codeml that demonstrated a low false-positive rate in the absence of $d_N > d_S$ and a reasonable sensitivity to detect positive selection across a range of branch lengths and d_N/d_S ratios. However, they primarily focused on selection acting at an internal branch (thereby improving detection of selected sites) and on trees with branch lengths longer than those encountered in human genes.

We used evolver, which generates alignment sets that correspond to a user-defined evolutionary model, to create simulated primate gene alignments for use in evaluating the performance of the strict branch-site test of codeml. Several scenarios were examined by varying parameters that reflect the range of branch lengths, d_N/d_S ratios, and codon frequencies found in human genes. These included "gold standard" sets simulating positive selection with $d_N/d_S > 1$ at a subset of codons on the human lineage in each simulated sequence and neutrally evolving control sets with $d_N/d_S = 1$. The complete list of tests, parameters, and results can be found in supplemental Table S5.

The branch length of the human (foreground) lineage had the greatest influence on the power to detect positive selection, which falls as branch length decreases (Figures 2 and 3 and supplemental Table S5). Due to the recent divergence of humans and chimpanzees, human-specific branch lengths are quite short

TABLE 1

Codeml results for transcription factor genes with significant results in the strict branch-site test of positive selection

| | Branch-site tests | | | | Branch tests | | | | | | Empirical test: |
| | Strict (MA *vs.* MA-null) | | Relaxed (MA *vs.* M1a) | | Strict (MH *vs.* MH-null) | | Relaxed (MH *vs.* M0) | | d_N/d_S | | |
Gene	$2 \times (\ln L_1 - \ln L_2)^a$	P	$2 \times (\ln L_1 - \ln L_2)$	P	$2 \times (\ln L_1 - \ln L_2)$	P	$2 \times (\ln L_1 - \ln L_2)$	P	Human	Other	P
ZFP37	4.62	*0.032*	8.88	*0.012*	4.70	*0.030*	14.4	*<0.001*	999[b]	0.24	*<0.001*
ISGF3G	6.24	*0.013*	6.50	*0.039*	0.18	0.67	0.48	0.48	0.69	0.38	*<0.001*
SOHLH2	5.16	*0.023*	9.08	*0.011*	5.18	*0.023*	9.62	*0.002*	999	0.48	*<0.001*

[a] Likelihood-ratio statistic (LRS) $= 2 \times (\ln L_1 - \ln L_2)$, where $\ln L_1$ is the maximum-likelihood estimate for the test model and $\ln L_2$ is the maximum-likelihood estimate for the null model.
[b] Values of 999 for d_N/d_S indicate $d_S = 0$, so d_N/d_S is undefined. Values in italics reflect $P < 0.05$.

(Figure 3). Under a scenario of positive selection, with $d_N/d_S = 10$ on the foreground (human) lineage and $d_N/d_S \leq 1$ on the other lineages, only a small fraction of simulated sequences had a *P*-value <0.05 in the strict branch-site test (Figure 2A). Across most of the range of

observed human branch lengths, few sequences simulated under a model of positive selection met the threshold for significance in the strict branch-site test (Figure 3). For the median branch length of human genes with substitutions, 0.007, <10% of sequences
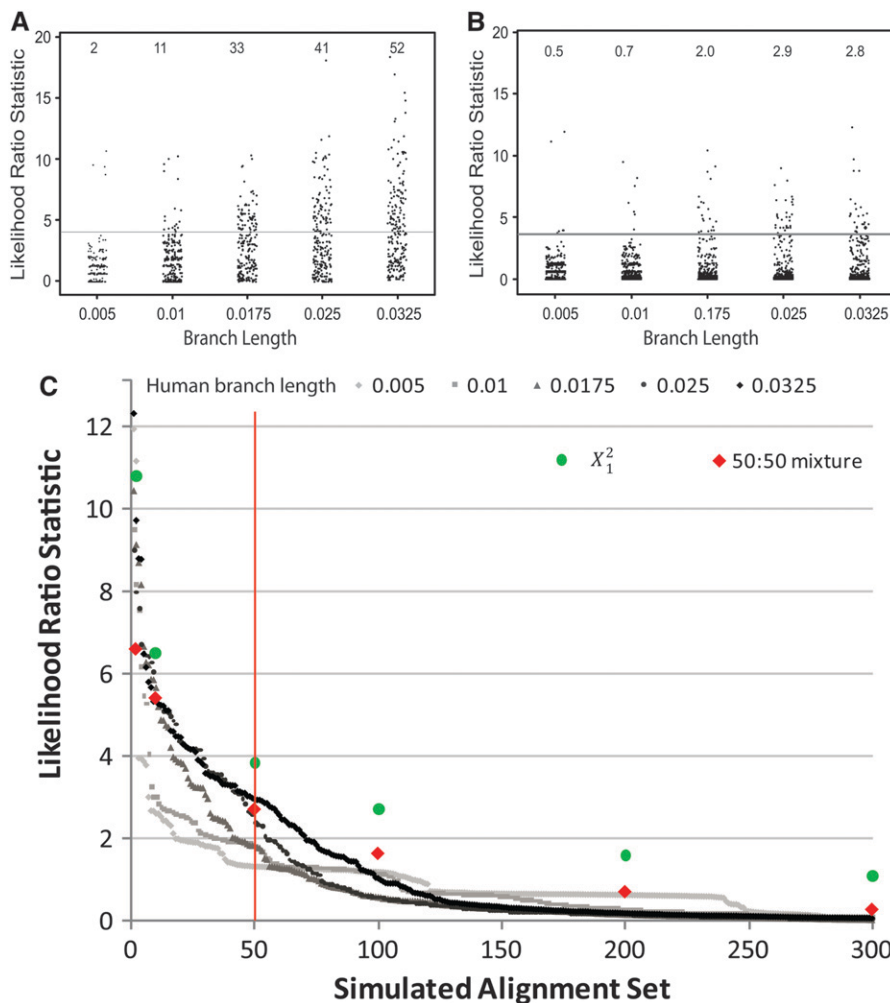

FIGURE 2.—Distribution of likelihood-ratio statistic (LRS) values from the strict branch-site test of simulated gene sets. Sequences were simulated with different branch lengths for the human branch under models of positive selection with foreground $d_N/d_S = 10$ (A) and under models of neutral evolution with foreground $d_N/d_S = 1$ (B and C). (A and B) the LRS corresponding to $P = 0.05$ is shown as a horizontal gray line. The percentage of simulated sequences with $P < 0.05$ is shown at the top of each data series. The expected distribution of LRS values according to the chi-square distribution with 1 d.f. (χ^2_1, red diamonds) and a 50:50 mixture distribution of the χ^2_1-distribution and a point mass of zero (green circles) is shown in C with a subset of the distribution of LRS values from neutral simulations from B. The red line marks the 5% empirical significance threshold.
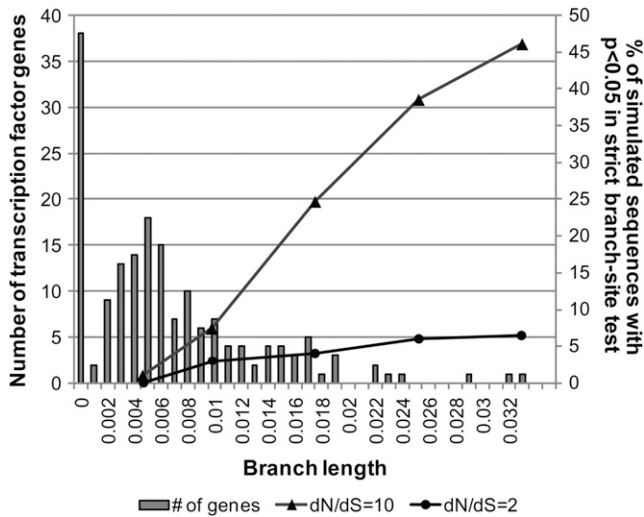
FIGURE 3.—Prediction of positive selection across a range of branch lengths representative of human genes. The distribution of the human branch length for the 175 transcription factor genes is shown (columns). The secondary axis represents the percentage of simulated alignment sets that were correctly classified as positively selected in the strict branch-site test when d_N/d_S on the foreground lineage was set at 2 (circles) or 10 (triangles).

simulated under the model of positive selection are correctly classified. At longer branch lengths, more simulated positively selected sequences are correctly classified; however, <1% of human genes have branch lengths >0.035, and positive selection is not completely predicted even at this relatively long branch length.

The power to predict positive selection is also related to the magnitude of d_N/d_S (Figure 4A). With weak positive selection (*e.g.*, $d_N/d_S = 2$), there is little power to detect positive selection regardless of branch length, while at longer branch lengths, larger d_N/d_S results in a higher fraction of simulated sequences that meet the threshold for significance in the strict branch-site test, although this is insufficient to overcome the limitations of shorter branch lengths. Of much less influence is the d_N/d_S value of the background lineages (Figure 4B). The proportion of significant alignment sets remains consistent over a variety of background d_N/d_S values, ranging from low ($d_N/d_S = 0.01$) to nearly neutral ($d_N/d_S = 0.75$). The proportion of foreground (positively selected) sites also has only a modest impact on predictions of positive selection (supplemental Table S5).

We also examined the performance of evolver-generated sequences with $d_N/d_S = 1$ on the foreground lineage to assess how well these sequences match the neutral expectation. The fraction of simulated sequences with $P < 0.05$ in the strict branch-site test was ≤3% across a range of branch lengths (Figure 2B), demonstrating that under a neutral model, there are few false predictions of positive selection, confirming results of ZHANG *et al.* (2005).
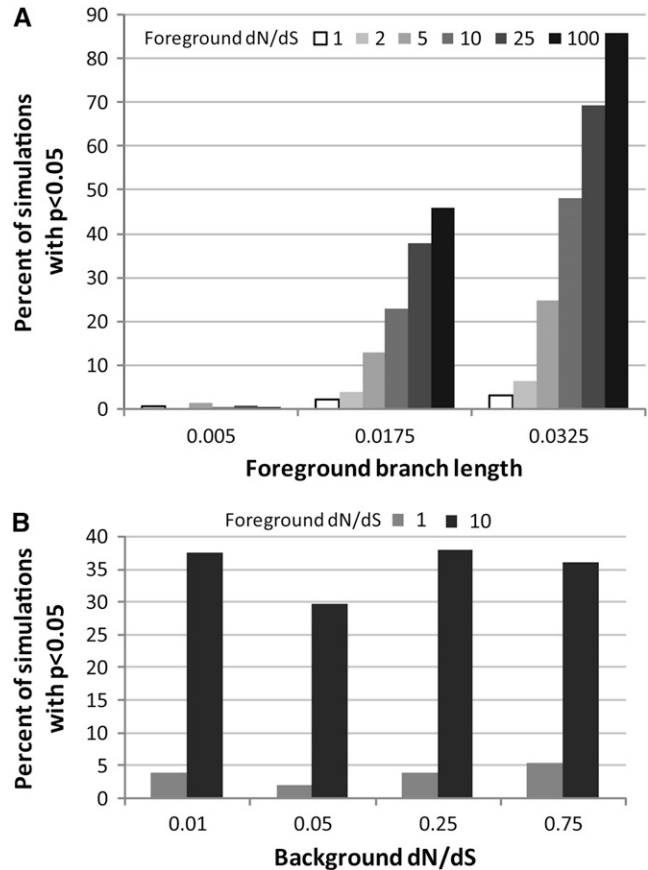

FIGURE 4.—Contribution of d_N/d_S values to predictions of positive selection. (A) The ability to detect positive selection using a range of d_N/d_S values for the foreground (human) lineage for different branch lengths was evaluated. (B) The effect of d_N/d_S on the background lineages was assessed under models of positive selection ($d_N/d_S = 10$; solid bars) and neutral evolution ($d_N/d_S = 1$; shaded bars). The human branch length was 0.01.

In the strict branch-site test, the LRS is compared to a chi-square distribution with 1 d.f. (χ_1^2) to obtain a P-value. In principle, since we are testing deviation from neutrality only in one direction, the appropriate comparison is to a 50:50 mixture of χ_1^2 and point mass 0 (SELF and LIANG 1987; CLARK and SWANSON 2005), and the distribution of LRS values with the neutral models fits the 50:50 mixture distribution more closely (Figure 2C), suggesting that evolver-simulated sequences do in fact represent a reasonable approximation of the null distribution.

Alternative null model using an empirical test: The simulations showed that for any given set of parameters, increasing d_N/d_S above 1 resulted in a shift in the distribution of LRS values in the comparison of model A and model A-null (Figure 2). As an alternative to determining significance using a likelihood-ratio test as in the strict branch-site test, we evaluated use of the distribution of LRS values from simulations with $d_N/d_S = 1$ on the foreground branch to define an empirical
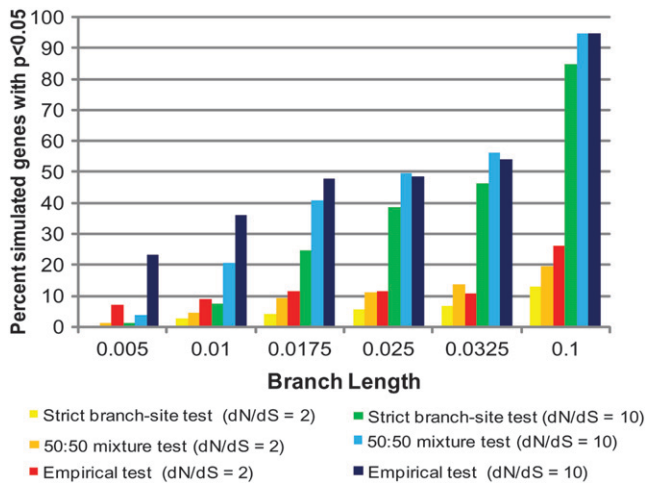
Figure 5.—Comparison of the strict branch-site test, the 50:50 mixture test, and the empirical test on simulated sequences. A total of 215 sequences were simulated under a model of positive selection using a d_N/d_S value for the foreground lineage of 2 (yellow, orange, and red bars) or 10 (green, blue, and black bars). The percentage correctly classified was calculated using three different methods of assessing significance: (1) the strict branch-site test, which uses a χ^2_1-distribution, (2) a 50:50 mixture of a point mass of 0 and the χ^2_1, and (3) the empirical test. For the empirical test, the threshold for each branch length was defined on the basis of the distribution of LRS values for the paired neutral simulation (foreground $d_N/d_S = 1$).

threshold of significance. In this method, the 5% significance threshold is defined as the LRS value such that 5% of the simulated neutral sequences have a LRS value above the threshold. In evaluating sequences simulated under a model of positive selection, if the LRS exceeds the 5% threshold defined in the matched neutral simulation, we would consider that there is less than a 5% probability that the pattern of selection can be accounted for by the neutral model. As shown in Figure 5, more sequences simulated under models of positive selection were correctly classified when using the empirically defined 5% significance threshold, compared to the LRT. The increase in sensitivity was particularly evident at shorter branch lengths.

We then developed an approach, based on the empirical distribution of LRS values of neutrally evolving sequences, to defining the probability of positive selection for real human genes. We used evolver to generate neutrally evolved sequences along the human branch on the basis of the inferred human–chimpanzee ancestor for each tested gene (see MATERIALS AND METHODS). Each neutrally simulated "human" sequence was then analyzed by codeml, using model A and model A-null in the context of the primate plus rodent alignment. In this method, an empirically determined significance threshold (at $\alpha = 0.05$) was used; if the LRS of the real alignment was greater than that for

at least 95% of the neutrally evolved sequences, positive selection was inferred.

Empirical P-values were obtained for each gene that had a P-value <0.5 in the strict branch-site test, for several with $P > 0.5$ (Table 2), and for two additional control genes (see below). Figure 6 is a comparison of the nominal P-values from the strict branch-site test with P-values obtained from the empirical test using the alternative null model. Eight genes had $P < 0.05$ in the empirical test: CDX4, DMRT3, ELK4, PGR, RBM9, SOHLH2, TAF11, and ZFP37. The control gene ASPM was also significant in the empirical test. These genes have a nonsynonymous substitution rate greater than can be accounted for on the basis of a model of neutral evolution. There is only a modest overlap (Table 2) in genes with $P < 0.05$ in the empirical test and $P < 0.05$ in the relaxed branch-site test of selection, indicating that the empirical test is not simply identifying genes undergoing a relaxation of selective constraint.

ISGF3G was significant in the strict branch-site test, but not in the empirical test. This gene has a rare "double-hit" codon substitution in which two of the three nucleotides at one codon have changed since the human–chimpanzee ancestor. An ancestral serine codon (TCA) is changed to valine (GTA) at position 129 in human. Apparently codeml weights this highly unlikely substitution quite strongly. Introducing a single double-hit substitution into other genes resulted in a shift to $P < 0.05$ in the strict branch-site test (data not shown). The selection status of ISGF3G is thus unclear.

To guard against the possibility that some genes with low P-values scored well due to the presence of rare nonsynonymous polymorphisms in the human sequence used for the alignment, the SNP content of each was examined in dbSNP and by comparison to sequences in dbEST. CDX4 and ZFP37 were also examined by complete exon sequencing in 12 humans. In ZFP37, one high-frequency nonsynonymous polymorphism was found at codon 7 (rs2282076). Use of the ancestral allele at this position had a negligible impact on the empirical test. A nonsynonymous substitution in BNC1 at residue 274 (rs1130108) was determined to be a rare polymorphism. Use of the major allele at this position resulted in loss of significance; results presented here are for the sequence with the major allele.

Comparison with previous predictions of positive selection: Three genes that have been previously suggested to demonstrate positive selection in humans were also examined: FOXP2 (ENARD et al. 2002), ASPM (EVANS et al. 2004b), and MCPH1 (EVANS et al. 2004a). FOXP2 is significant in the relaxed branch-site test, but not in the strict branch-site test, in accord with the results of Enard et al., whereas neither ASPM nor MCPH1 reached significance in either traditional test of human-specific positive selection. This agrees with the results of several other studies that indicate that both proteins experienced rapid adaptive evolution along both the

TABLE 2

Comparison of the results from the strict branch-site and empirical tests

Gene	Length (codons)	N^a	S^a	d_N/d_S Background	d_N/d_S Foreground (human)	Relaxed branch-site test P	Strict branch-site test P	Empirical P
ASPM	3477	17	7	0.109	5.455	0.553	0.473	*0.027*
BLZF1	400	3	0	0.146	999^b	0.084	0.260	0.105
CDX1	265	2	0	0.104	999	0.292	0.307	0.125
CDX4	284	4	0	0.358	999	*0.027*	0.120	*0.031*
CITED2	270	2	0	0.038	999	*0.028*	0.426	0.124
DMRT3	472	11	3	0.084	0.859	*0.005*	0.489	*0.037*
EHF	300	1	0	0.048	313	0.139	0.471	0.102
ELK4	431	2	0	0.270	999	0.092	0.212	*0.042*
FOXI1	378	5	3	0.083	0.332	0.481	0.811	0.277
FOXJ2	574	2	0	0.099	75.8	*0.025*	0.239	0.085
GBX2	348	1	0	0.014	999	0.050	0.531	0.255
HESX1	185	1	0	0.269	999	0.186	0.429	0.180
HOXA5	270	2	1	0.029	0.445	0.095	1.000	0.833
HOXD4	255	2	0	0.082	999	*0.021*	0.288	0.076
KLF10	390	2	0	0.164	999	*0.039*	0.229	0.090
MCPH1	740	9	4	0.112	1.000	0.315	1.000	0.748
MEOX1	254	2	0	0.127	999	0.075	0.298	0.185
NR2C1	603	4	0	0.089	999	*0.008*	0.192	0.309
NR3C2	984	5	1	0.102	0.770	*0.026*	0.318	0.069
PEG3	1590	8	3	0.232	0.944	0.067	0.803	0.177
PGR	933	9	4	0.197	0.893	0.072	0.469	*0.029*
RBM9	367	2	0	0.400	999	0.410	0.184	*0.041*
SCML1	208	2	0	1.279	999	0.446	0.231	0.075
SOHLH2	425	10	0	0.484	999	*0.011*	*0.023*	*<0.001*
T	435	2	0	0.052	999	0.068	0.318	0.124
TAF11	211	2	0	0.138	999	*0.031*	0.234	*0.034*
ZFP37	630	9	0	0.237	999	*0.011*	*0.032*	*<0.001*
ZNF192	578	4	1	0.210	1.605	0.270	0.411	0.101

^{a} Substitutions between the inferred human–chimpanzee ancestor and the human sequence. *N*, nonsynonymous; *S*, synonymous.

^{b} Values of 999 for d_N/d_S indicate $d_S = 0$, so d_N/d_S is undefined. Values in italics reflect $P < 0.05$.

earlier hominoid and later hominid lineages, preceding the human–chimpanzee split (Evans *et al.* 2004a,b; Kouprina *et al.* 2004; Wang and Su 2004). In the empirical tests, only *ASPM* reached significance.

Among the 175 genes analyzed by phylogenetic sequencing here, 73 overlap with genes analyzed by Clark *et al.* (2003), using alignments of human, chimpanzee, and mouse only. Fifteen of the overlapping genes were predicted to be positively selected in the Clark *et al.* study ($P \leq 0.05$ using the M2 method). Of these 15, none had nominally significant *P*-values in the strict branch-site test using the more extensive phylogenetic sequence set described here and one (CDX4) had an empirical *P*-value <0.05. Several factors may explain the higher number of predictions in the Clark analysis, including a lack of accuracy in reconstruction of the ancestral state in alignments involving only human, chimpanzee, and mouse sequences and differences in the details of the ML methods used. There is a better concordance between M2 *P*-values and *P*-values from the relaxed branch-site test, suggesting that improve-

ments in codeml and use of the more stringent test are primarily responsible for fewer predictions of positive selection here (data not shown).

DISCUSSION

Inference of positive selection relies on observation of an excess rate of nonsynonymous codon substitutions relative to synonymous substitutions. In closely related species, with few total substitutions, these rates are low and the uncertainty in defining them is high, limiting the ability to confidently predict positive selection. We have shown that the strict branch-site test of codeml is largely unable to detect positive selection when it exists across a broad range of parameters of branch length and d_N/d_S ratios that are found in human genes in the context of alignments with sequences from several nonhuman primates. The two factors with the largest effect were branch length and magnitude of d_N/d_S on the human branch. Predictions of selection were most efficient on genes with large d_N/d_S values and
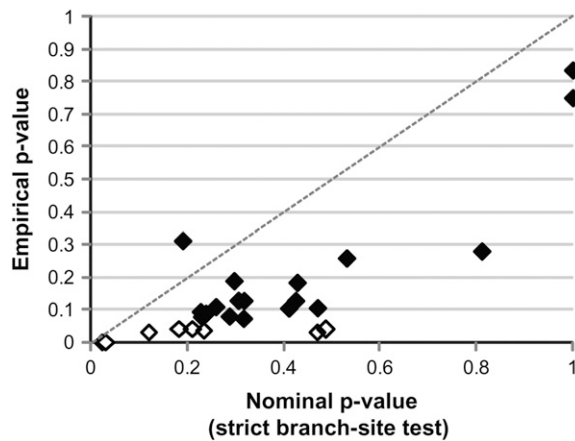
FIGURE 6.—Comparison of the strict branch-site test and the empirical test on human genes. An empirical probability of positive selection was determined for 26 human transcription factor genes on the basis of codeml analysis of at least 500 alignments with the human sequence substituted with a simulated sequence derived from the human–chimpanzee ancestor, using a model of neutral evolution. The empirical probability represents the proportion of simulated neutral sequences that gave an LRS greater than that observed for the real gene in the strict branch-site test.

on genes with branch lengths >0.03 (Figure 2 and supplemental Table S5). The 175 transcription factor genes used in this study have a median branch length of 0.007, which is well below the optimum for prediction of positive selection. In fact, a branch length of >0.035 is necessary to predict positive selection in a 400-codon gene 50% of the time, and <1% of human genes match these criteria. The standard likelihood-ratio test (strict branch-site test) used with codeml model A and model A-null therefore lacks the power to detect positive selection given the phylogenetic history of most human genes.

The tests for positive selection used here were designed to detect an alteration in evolutionary rate that is specific to the human lineage. An implicit assumption is that the rate is constant (or less variable) along other branches. The site models M8 and M8a can be used to assess whether there are subsets of codons that are evolving rapidly across the tree (YANG *et al.* 2000). None of the eight genes with significant predictions of positive selection in the empirical test are significant with the site models. Codeml can also be used with a "free-ratio" model, in which d_N/d_S is estimated individually on each branch of the tree. This model is very parameter rich and cannot be effectively compared to a suitable null model and so branch-specific rates must be considered approximations. Each branch with $d_N/d_S > 1$ in the free-ratio model was tested individually and no branches other than human exhibited a significant LRT in the branch or branch-site tests for the three genes in Table 1.

After collection of additional sequence data, 65 of the 175 human TF genes encoded proteins that are identical

in chimpanzee, so sequence errors may have contributed to previous indications of positive selection. The importance of comparing accurate sequence is highlighted by the impact of single substitutions on the codeml results. A concern about accuracy was also highlighted by BAKEWELL *et al.* (2007), who used the available base quality values to restrict analysis to the highest confidence regions of the chimp genome. Two genes predicted to be positively selected were reevaluated by codeml after the addition of a single synonymous substitution and the removal of individual nonsynonymous substitutions. These small changes within the alignments resulted in a loss of significance in the strict branch-site test, but not in the empirical test (supplemental Table S6).

The empirical test of selection described here is an alternative null test that seems to have greater sensitivity to detect positive selection than the strict branch-site test at short branch lengths. The test essentially assesses whether the observed likelihood of positive selection can be accounted for by a neutral model of evolution. By modeling neutral evolution specifically on the human branch, while keeping the rest of the tree constant, the test allows us to assess the potential for positive selection on a gene-specific basis. This new method may therefore serve as a useful adjunct to the phylogenetic methods employed in codeml. Evolver-generated neutral sequences seem to perform quite consistently across a range of branch lengths, codon frequencies, and proportions of selected sites (Figure 2B). Nonetheless, the neutral model cannot fully describe the true evolutionary history and so caution is advised in interpretation of results of the empirical test as for any computational inference of positive selection. The empirical test is also impractical for large-scale use due to the computational cost: codeml analysis for 1000 simulated *ZFP37* sequences required >700 CPU hours. The empirical test and the 50:50 mixture distribution give similar results at longer branch lengths (Figure 5). We therefore recommend use of the empirical test for genes with branch length <0.025.

The genes analyzed here are not a random or representative sampling of TF genes, but were selected specifically to be biased toward those more likely to show evidence of positive selection. Despite this emphasis, none of the 175 genes analyzed met the strict criteria for positive selection in the branch-site tests after multiple-test correction, and, in the empirical test, 8 genes were predicted to be positively selected. A similar dearth of evidence for human-specific positive selection was seen in analyses of >13,000 human–chimpanzee–macaque gene alignments (BAKEWELL *et al.* 2007; GIBBS *et al.* 2007), although both studies reported a somewhat larger number of genes with a prediction of positive selection on the chimpanzee branch.

A subset of transcription factor genes was identified that contains a significant excess of nonsynonymous

substitution that has been shown to be unlikely to occur by neutral evolution of the gene. There are several well-described examples of evolution of transcription factors that contribute to morphological divergence (Lohr *et al.* 2001; Galant and Carroll 2002; Ronshaugen *et al.* 2002; Hittinger *et al.* 2005; Lohr and Pick 2005) (see Hsia and McGinnis 2003 for a review). For example, the Hox protein Ultrabithorax (Ubx) represses limb development in insects, but not in crustaceans (Ronshaugen *et al.* 2002), and this difference has been mapped to an interaction domain with pleiotropic functions (Hittinger *et al.* 2005). Further work will be required to determine whether the predictions of positive selection in the genes presented here are indicative of functional divergence in the encoded TF proteins and to link that divergence to phenotypic distinctions between humans and chimpanzees.

LITERATURE CITED

Aagaard, J. E., and P. Phillips, 2005 Accuracy and power of the likelihood ratio test for comparing evolutionary rates among genes. J. Mol. Evol. **60:** 426–433.

Arbiza, L., J. Dopazo and H. Dopazo, 2006 Positive selection, relaxation, and acceleration in the evolution of the human and chimp genome. PLoS Comput. Biol. **2:** e38.

Bakewell, M. A., P. Shi and J. Zhang, 2007 More genes underwent positive selection in chimpanzee evolution than in human evolution. Proc. Natl. Acad. Sci. USA **104:** 7489–7494.

Benjamini, Y., and Y. Hochberg, 1995 Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. B **57:** 289–300.

Bininda-Emonds, O. R., 2005 TransAlign: using amino acids to facilitate the multiple alignment of protein-coding DNA sequences. BMC Bioinform. **6:** 156.

Bullaughey, K. L., M. Przeworski and G. Coop, 2008 No effect of recombination on the efficacy of natural selection in primates. Genome Res. **18:** 544–554.

Bustamante, C. D., A. Fledel-Alon, S. Williamson, R. Nielsen, M. T. Hubisz *et al.*, 2005 Natural selection on protein-coding genes in the human genome. Nature **437:** 1153–1157.

Cheng, Z., M. Ventura, X. She, P. Khaitovich, T. Graves *et al.*, 2005 A genome-wide comparison of recent chimpanzee and human segmental duplications. Nature **437:** 88–93.

Clark, N. L., and W. J. Swanson, 2005 Pervasive adaptive evolution in primate seminal proteins. PLoS Genet. **1:** e35.

Clark, A. G., S. Glanowski, R. Nielsen, P. D. Thomas, A. Kejariwal *et al.*, 2003 Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. Science **302:** 1960–1963.

Chimpanzee Sequencing and Analysis Consortium, 2005 Initial sequence of the chimpanzee genome and comparison with the human genome. Nature **437:** 69–87.

Dorus, S., E. J. Vallender, P. D. Evans, J. R. Anderson, S. L. Gilbert *et al.*, 2004 Accelerated evolution of nervous system genes in the origin of Homo sapiens. Cell **119:** 1027–1040.

Enard, W., M. Przeworski, S. E. Fisher, C. S. Lai, V. Wiebe *et al.*, 2002 Molecular evolution of FOXP2, a gene involved in speech and language. Nature **418:** 869–872.

Evans, P. D., J. R. Anderson, E. J. Vallender, S. S. Choi and B. T. Lahn, 2004a Reconstructing the evolutionary history of microcephalin, a gene controlling human brain size. Hum. Mol. Genet. **13:** 1139–1145.

Evans, P. D., J. R. Anderson, E. J. Vallender, S. L. Gilbert, C. M. Malcom *et al.*, 2004b Adaptive evolution of ASPM, a major determinant of cerebral cortical size in humans. Hum. Mol. Genet. **13:** 489–494.

Ewing, B., and P. Green, 1998 Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res. **8:** 186–194.

Ewing, B., L. Hillier, M. C. Wendl and P. Green, 1998 Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res. **8:** 175–185.

Felsenstein, J., 1981 Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. **17:** 368–376.

Felsenstein, J., 1993 *PHYLIP (Phylogeny Inference Package)*, Version 3.5c. University of Washington, Seattle.

Florea, L., G. Hartzell, Z. Zhang, G. M. Rubin and W. Miller, 1998 A computer program for aligning a cDNA sequence with a genomic DNA sequence. Genome Res. **8:** 967–974.

Galant, R., and S. B. Carroll, 2002 Evolution of a transcriptional repression domain in an insect Hox protein. Nature **415:** 910–913.

Gibbs, R. A., J. Rogers, M. G. Katze, R. Bumgarner, G. M. Weinstock *et al.*, 2007 Evolutionary and biomedical insights from the rhesus macaque genome. Science **316:** 222–234.

Goldman, N., 1993 Statistical tests of models of DNA substitution. J. Mol. Evol. **36:** 182–198.

Goldman, N., and Z. Yang, 1994 A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol. Biol. Evol. **11:** 725–736.

Grenier, J. K., and S. B. Carroll, 2000 Functional evolution of the Ultrabithorax protein. Proc. Natl. Acad. Sci. USA **97:** 704–709.

Hittinger, C. T., D. L. Stern and S. B. Carroll, 2005 Pleiotropic functions of a conserved insect-specific Hox peptide motif. Development **132:** 5261–5270.

Hsia, C. C., and W. McGinnis, 2003 Evolution of transcription factor function. Curr. Opin. Genet. Dev. **13:** 199–206.

Kidd, J. M., K. C. Trevarthen, D. L. Tefft, Z. Cheng, M. Mooney *et al.*, 2005 A catalog of nonsynonymous polymorphism on mouse chromosome 16. Mamm. Genome **16:** 925–933.

Kouprina, N., A. Pavlicek, G. H. Mochida, G. Solomon, W. Gersch *et al.*, 2004 Accelerated evolution of the ASPM gene controlling brain size begins prior to human brain expansion. PLoS Biol. **2:** E126.

Lohr, U., and L. Pick, 2005 Cofactor-interaction motifs and the cooption of a homeotic Hox protein into the segmentation pathway of Drosophila melanogaster. Curr. Biol. **15:** 643–649.

Lohr, U., M. Yussa and L. Pick, 2001 Drosophila fushi tarazu. A gene on the border of homeotic function. Curr. Biol. **11:** 1403–1412.

McDonald, J. H., and M. Kreitman, 1991 Adaptive protein evolution at the Adh locus in Drosophila. Nature **351:** 652–654.

Nickel, G. C., D. Tefft and M. D. Adams, 2008 Human PAML browser: a database of positive selection on human genes using phylogenetic methods. Nucleic Acids Res. **36:** D800–D808.

Nielsen, R., C. Bustamante, A. G. Clark, S. Glanowski, T. B. Sackton *et al.*, 2005 A scan for positively selected genes in the genomes of humans and chimpanzees. PLoS Biol. **3:** e170.

Ronshaugen, M., N. McGinnis and W. McGinnis, 2002 Hox protein mutation and macroevolution of the insect body plan. Nature **415:** 914–917.

Sabeti, P. C., D. E. Reich, J. M. Higgins, H. Z. Levine, D. J. Richter *et al.*, 2002 Detecting recent positive selection in the human genome from haplotype structure. Nature **419:** 832–837.

Self, S. G., and K. Liang, 1987 Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. J. Am. Stat. Assoc. **82:** 605–610.

Tajima, F., 1993 Statistical analysis of DNA polymorphism. Jpn. J. Genet. **68:** 567–595.

Ting, C. T., S. C. Tsaur, M. L. Wu and C. I. Wu, 1998 A rapidly evolving homeobox at the site of a hybrid sterility gene. Science **282:** 1501–1504.

Vallender, E. J., and B. T. Lahn, 2004 Positive selection on the human genome. Hum. Mol. Genet. **13:** R245–R254.

Voight, B. F., S. Kudaravalli, X. Wen and J. K. Pritchard, 2006 A map of recent positive selection in the human genome. PLoS Biol. **4:** e72.

Wang, Y. Q., and B. Su, 2004 Molecular evolution of microcephalin, a gene determining human brain size. Hum. Mol. Genet. **13:** 1131–1137.

Wong, W. S., Z. Yang, N. Goldman and R. Nielsen, 2004 Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. Genetics **168:** 1041–1051.

YANG, Z., 1997 PAML: a program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. **13:** 555–556.

YANG, Z., and R. NIELSEN, 2002 Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Mol. Biol. Evol. **19:** 908–917.

YANG, Z., R. NIELSEN, N. GOLDMAN and A. M. PEDERSEN, 2000 Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics **155:** 431–449.

YANG, Z., W. S. WONG and R. NIELSEN, 2005 Bayes empirical Bayes inference of amino acid sites under positive selection. Mol. Biol. Evol. **22:** 1107–1118.

ZHANG, J., R. NIELSEN and Z. YANG, 2005 Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol. Biol. Evol. **22:** 2472–2479.

Communicating editor: M. NORDBORG