An empirical test for branch-specific positive selection

Gabrielle C. Nickel, David L. Tefft\textsuperscript{1}, Karrie Goglin\textsuperscript{2},
and Mark D. Adams* 

Department of Genetics

Case Western Reserve University

Cleveland, OH 44106

Current address:
\textsuperscript{1}The Broad Institute
Cambridge, MA 02141

\textsuperscript{2}The Scripts Research Institute
La Jolla, CA 92037

Sequence data from this article have been deposited with the EMBL/GenBank Data Libraries under accession nos. DQ976435- DQ977607 and EF185319- EF185336.
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 dN/dS ratio. Parameters were varied across a range of values observed in alignments of 175 transcription factor (TF) genes that were sequenced in twelve 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 based on 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 about 1% fixed single-nucleotide differences with the human genome, and about 1.5% lineage-specific DNA in each species (CHENG et al. 2005; CSAAC, 2005). Even though the percent divergence is small, there are about 35 million single nucleotide changes and 5 million insertion/deletion differences between human and chimpanzee, most of which are presumed to be evolutionarily neutral (CSAAC, 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
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, genome-wide surveys have identified genes, gene families, and genome regions with a signature of positive selection (CSAAC, ARBIZA et al. 2006; BAKEWELL et al. 2007; BUSTAMANTE et al. 2005; CLARK et al. 2003; 2005; DORUS et al. 2004; GIBBS et al. 2007; VOIGHT et al. 2006). We are most interested in identifying genetic differences associated with fundamental and ancient divergence between humans and the non-human 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 (dN) and synonymous (dS) substitutions. Under neutral evolution, the rate of accumulation of synonymous and nonsynonymous substitutions should be equal and a dN/dS value of one is expected. A dN/dS value greater than one is indicative of positive selection, while a value less than one indicates negative selection. About one-third of chimpanzee protein sequences are identical to
the sequence of the human ortholog, with the remainder having typically one or two
amino acid changes (CSAAC, 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 1997b)) 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.

Genome-wide studies to date have been based on data from only two (ARBIZA et
al. 2006; CSAAC, 2005; NIELSEN et al. 2005) or three species (BAKEWELL et al. 2007;
CLARK et al. 2003; 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 twelve 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 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, the ability of small changes in sequence to cause broad changes in downstream gene expression patterns (e.g., GRENIER and CARROLL 2000; TING et al. 1998), and because TFs are over-represented among genes predicted to be positively selected in previous genome-wide studies of selection (ARBIZA et al. 2006; BUSTAMANTE et al. 2005; CSAAC, 2005).

MATERIALS AND METHODS

Selection of genes and DNA sequencing: Transcription factor genes were selected for sequencing based on a prior prediction of positive selection (BUSTAMANTE et al. 2005; CLARK et al. 2003; 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 NIA Aging Cell Repository Phylogenetic Primate Panel (PRP0001). Human DNA samples were from the NIGMS 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 Dr. 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; GREEN), with each exon from each organism assembled independently. Low quality bases (Q<25) in the consensus were converted to Ns. 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://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 (CSSAC, 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 Supplementary Table S1.

**Phylogenetic analysis:** *Codeml* from the PAML package (v3.14b, YANG 1997b) 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 Supplementary Table S2 and S3. Each gene was first tested by asking whether a model in which the human lineage was allowed to have a different dN/dS ratio than the rest of the tree was a better fit to the data than a model with a single dN/dS ratio throughout the tree (“branch test”). Whether the human-specific dN/dS ratio was significantly greater than 1, was tested by comparison to a model with the dN/dS 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 dN/dS>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 dN/dS 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 (Fig 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 has been submitted to
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 dN/dS, and the magnitude of the foreground dN/dS value were 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 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. 1,000 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 \textit{codeml} with model A and model A-null.

\textbf{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 (\textit{codeml} model A vs. model A\textsubscript{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 \textit{RateAncestor} function of \textit{codeml}. More than 99\% of the codons were inferred with at least 95\% probability, according to \textit{codeml} (Supplemental Fig S1). These ancestral sequences were input as the root sequence in \textit{evolver}, and at least 500 sequences were simulated under a model of neutral evolution (foreground dN/dS = 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 \textit{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, based on averages among genes with dN/dS>1 (see Table S3). The
transition/transversion ratio, $\kappa$, was also held constant at 2.6. The average $\kappa$ for the 175 gene set was 4.1±1.5. A higher $\kappa$ means more transitions, and therefore more synonymous changes. A smaller $\kappa$ is therefore more conservative. Codon substitution frequencies were derived from the appropriate G+C quartile for the specific gene. $\kappa=2.6$ 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+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 $\alpha = 5\%$ (AAGAARD and PHILLIPS 2005; GOLDMAN 1993).
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. 175 transcription factor genes with prior predictions of positive selection (Bustamante et al. 2005; Clark et al. 2003; CSSAC, 2005; Nielsen et al. 2005) or an excess on nonsynonymous substitution in human relative to chimpanzee were selected for phylogenetic sequencing. Exonic sequence data was obtained from human, chimpanzee, bonobo, gorilla, orangutan, gibbon, three Old World monkeys, and three New World monkeys, and lemur (Fig.1). Primary sequence read data from chimpanzee, macaque, and orangutan, were obtained from the NCBI Trace Repository, supplemented the alignments where possible, and annotated mouse and rat orthologs (http://www.informatics.jax.org) were added to the alignments (SI 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.

Multi-species 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 1997a). In each case, results from a null model were compared with results from a test model. The null models assumed neutrality (dN = dS) or negative selection (dN/dS <1) and the test models assumed dN/dS >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 SI Table 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. (SI 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 dN/dS ratio >1 on the human branch, 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 dN>dS and a reasonable sensitivity to detect positive selection across a range of branch lengths and dN/dS 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, dN/dS ratios and codon frequencies found in human genes. These included 'gold standard' sets simulating positive selection with dN/dS>1 at a subset of codons on the human lineage in each simulated sequence and neutrally evolving control sets with dN/dS=1. The complete list of tests, parameters, and results can be found in SI 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, 3, and SI Table S5). Due to the recent divergence of humans and chimpanzees, human-specific branch lengths are quite short (Fig 3). Under a scenario of positive selection, with dN/dS=10 on the foreground (human) lineage and dN/dS<=1 on the other lineages, only a small fraction of simulated sequences had a p-value <0.05 in the strict branch+site test (Fig 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 (Fig 3). For the median branch length of human genes with substitutions, 0.007, less than 10% of sequences 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 dN/dS (Fig. 4A). With weak positive selection (e.g. dN/dS = 2), there is little power to detect positive selection regardless of branch length, while at longer branch lengths, larger dN/dS 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 $dN/dS$ value of the background lineages (Figure 4B). The proportion of significant alignment sets remains consistent over a variety of background $dN/dS$ values, ranging from low ($dN/dS = 0.01$) to nearly neutral ($dN/dS = 0.75$). The proportion of foreground (positively selected) sites also has only modest impact on predictions of positive selection (SI Table S5).

We also examined the performance of *evolver*-generated sequences with $dN/dS=1$ 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 $\leq 3\%$ across a range of branch lengths (Fig 2B), demonstrating that under a neutral model, there are few false predictions of positive selection, confirming results of Zhang , *et al* (ZHANG *et al.* 2005).

In the strict branch+site test, the LRS is compared to a chi-square distribution with one degree of freedom ($X^2_1$) to obtain a $p$-value. In principle, since we are only testing deviation from neutrality in one direction the appropriate comparison is to a 50:50 mixture of $X^2_1$ and point mass 0 (CLARK and SWANSON 2005; SELF and LIANG 1987), and the distribution of LRS values with the neutral models fits the 50:50 mixture
distribution more closely (Fig 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 dN/dS above 1 resulted in a shift in the distribution of LRS values in the comparison of model A and model A-null (Fig. 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 dN/dS=1 on the foreground branch to define an empirical 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 Fig 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 based on the inferred human-chimpanzee ancestor for each tested gene (see 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 $p$-value <0.5 in the strict branch+site test, several with $p>0.5$ (Table 2), and 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, 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 based on 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 twelve 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 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 earlier hominoid and later hominid lineages, preceding the human-chimpanzee split (Evans et al. 2004a; Evans et al. 2004b; 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. (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 improvements 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 dN/dS ratios that are found in human genes in the context of alignments with sequences from
several non-human primates. The two factors with the largest effect were branch length and magnitude of dN/dS on the human branch. Predictions of selection were most efficient on genes with large dN/dS values and on genes with branch lengths >0.03 (Figure 2 and SI 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 fewer than 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 dN/dS 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 dN/dS >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. who used
the available base quality values to restrict analysis to the highest confidence regions of
the chimp genome (BAKEWELL et al. 2007). Two genes predicted to be positively
selected were re-evaluated 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 (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
The 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 proportion of selected sites (Fig 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 1,000 simulated ZFP37 sequences required more than 700 CPU-hours. The empirical test and the 50:50 mixture distribution give similar results at longer branch lengths (Fig 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 over 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 contain 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 (GALANT and CARROLL 2002; HITTINGER et al. 2005; LOHR and PICK 2005; LOHR et al. 2001; RONSHAUGEN et al. 2002) (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.
Acknowledgements:  The authors would like to thank Jason Funt, Dr. Andy Clark and Dr. Tom LaFramboise for helpful discussions.

SUPPLEMENTARY MATERIAL

Table S1. Sequence coverage of transcription factor genes and data sources

Table S2. Evolutionary models and parameter sets for codeml analysis

Table S3. Branch+site test site classes

Table S4. Codeml results from 175 transcription factors

Table S5. Sensitivity analysis: Factors that affect predictions of positive selection

Table S6. Effect of synonymous and nonsynonymous substitution on predictions of positive selection

Figure S1. Probability of accurate ancestral sequence reconstruction

Alignments and full results at http://mendel.gene.cwru.edu/adamslab/pbrowser.py
FIGURE LEGENDS

Figure 1. Phylogenetic tree of the primate species used in phylogenetic analyses.

Primates were chosen so as 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 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.

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 dN/dS = 10 (A) and under models of neutral evolution with foreground dN/dS = 1 (B, C). In (A) and (B), the LRS corresponding to p=0.05 is shown as a horizontal grey 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 one degree of freedom.
freedom ($X^2_1$, red diamonds) and a 50:50 mixture distribution of the $X^2_1$ 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 dN/dS on the foreground lineage was set at 2 (circles) or 10 (triangles).

Figure 4. Contribution of dN/dS values to predictions of positive selection. (A) The ability to detect positive selection using a range of dN/dS values for the foreground (human) lineage for different branch lengths was evaluated. (B) The effect of the dN/dS on the background lineages was assessed under models of positive selection (dN/dS =10; black bars) and neutral evolution (dN/dS=1; gray bars). The human branch length was 0.01.
Figure 5. Comparison of the strict branch+site test, 50:50 mixture test, and empirical test on simulated sequences. 215 sequences were simulated under a model of positive selection using a dN/dS value for the foreground lineage of 2 (yellow, orange and red bars) or 10 (green, blue, and black bars). The percent correctly classified was calculated using three different methods of assessing significance: 1) the strict branch+site test, which uses a $X^2$ distribution, 2) a 50:50 mixture of a point mass of 0 and the $X^2$, and 3) the empirical test. For the empirical test, the threshold for each branch length was defined based the distribution of LRS values for the paired neutral simulation (foreground dN/dS=1).

Figure 6. Comparison of the strict branch+site test and empirical test on human genes.

An empirical probability of positive selection was determined for 26 human transcription factor genes based on 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.
LITERATURE CITED


CLARK, A. G., S. GLANOWSKI, R. NIELSEN, P. D. THOMAS, A. KEJARIWAL et al., 2003


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.

EVANS, P. D., J. R. ANDERSON, E. J. VALLENDER, S. S. CHOI and B. T. LAHN, 2004a


FELSENSTEIN, J., 1993 PHYLIP (Phylogeny Inference Package) version 3.5c., pp. Department of Genetics, University of Washington, Seattle.


KOUPRINA, N., A. PAVLICEK, G. H. MOCHIDA, G. SOLOMON, W. GERSCH et al., 2004


Table 1. *codeml* results for transcription factor genes with significant results in the strict branch+site test of positive selection

<table>
<thead>
<tr>
<th>Gene</th>
<th>Branch+Site Tests</th>
<th>Branch Tests</th>
<th>Empirical Test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Strict (MA vs MA-null)</td>
<td>Relaxed (MA vs M1a)</td>
<td>Strict (MH vs MH-null)</td>
</tr>
<tr>
<td></td>
<td>$2^<em>$(lnL₁ - lnL₂)$^</em>$</td>
<td>$2^*$(lnL₁)</td>
<td>$2^<em>$(lnL₁ - lnL₂)$^</em>$</td>
</tr>
<tr>
<td>ZFP37</td>
<td>4.62 (0.032)</td>
<td>8.88 (0.012)</td>
<td>4.70 (0.030)</td>
</tr>
<tr>
<td>ISGF3G</td>
<td>6.24 (0.013)</td>
<td>6.50 (0.039)</td>
<td>0.18 (0.67)</td>
</tr>
<tr>
<td>SOHLH2</td>
<td>5.16 (0.023)</td>
<td>9.08 (0.011)</td>
<td>5.18 (0.023)</td>
</tr>
</tbody>
</table>

$^*$Values of 999 for dNdS indicate dS=0, so dN/dS is undefined. Values in bold reflect p<0.05. $^1$Likelihood ratio statistic (LRS) = $2^*$(lnL₁-lnL₂), where lnL₁ = maximum likelihood estimate for test model and lnL₂ = maximum likelihood estimate for null model.
Table 2: Comparison of the results from the strict branch+site and empirical tests

<table>
<thead>
<tr>
<th>Gene</th>
<th>Length (codons)</th>
<th>N'</th>
<th>S'</th>
<th>Background dN/dS</th>
<th>Foreground dN/dS (Human)</th>
<th>Relaxed branch+site test p</th>
<th>Strict branch+site test p</th>
<th>empirical p</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASPM</td>
<td>3477</td>
<td>1</td>
<td>7</td>
<td>0.109</td>
<td>5.455</td>
<td>0.553</td>
<td>0.473</td>
<td><strong>0.027</strong></td>
</tr>
<tr>
<td>BLZF1</td>
<td>400</td>
<td>3</td>
<td>0</td>
<td>0.146</td>
<td>999†</td>
<td>0.084</td>
<td>0.260</td>
<td>0.105</td>
</tr>
<tr>
<td>CDX1</td>
<td>265</td>
<td>2</td>
<td>0</td>
<td>0.104</td>
<td>999</td>
<td>0.292</td>
<td>0.307</td>
<td>0.125</td>
</tr>
<tr>
<td>CDX4</td>
<td>284</td>
<td>4</td>
<td>0</td>
<td>0.358</td>
<td>999</td>
<td><strong>0.027</strong></td>
<td>0.120</td>
<td><strong>0.031</strong></td>
</tr>
<tr>
<td>CITED2</td>
<td>270</td>
<td>2</td>
<td>0</td>
<td>0.038</td>
<td>999</td>
<td><strong>0.028</strong></td>
<td>0.426</td>
<td>0.124</td>
</tr>
<tr>
<td>DMRT3</td>
<td>472</td>
<td>1</td>
<td>3</td>
<td>0.084</td>
<td>0.859</td>
<td><strong>0.005</strong></td>
<td>0.489</td>
<td><strong>0.037</strong></td>
</tr>
<tr>
<td>EHF</td>
<td>300</td>
<td>1</td>
<td>0</td>
<td>0.048</td>
<td>313</td>
<td>0.139</td>
<td>0.471</td>
<td>0.102</td>
</tr>
<tr>
<td>ELK4</td>
<td>431</td>
<td>2</td>
<td>0</td>
<td>0.270</td>
<td>999</td>
<td>0.092</td>
<td>0.212</td>
<td><strong>0.042</strong></td>
</tr>
<tr>
<td>FOXI1</td>
<td>378</td>
<td>5</td>
<td>3</td>
<td>0.083</td>
<td>0.332</td>
<td>0.481</td>
<td>0.811</td>
<td>0.277</td>
</tr>
<tr>
<td>FOXJ2</td>
<td>574</td>
<td>2</td>
<td>0</td>
<td>0.099</td>
<td>75.8</td>
<td><strong>0.025</strong></td>
<td>0.239</td>
<td>0.085</td>
</tr>
<tr>
<td>GBX2</td>
<td>348</td>
<td>1</td>
<td>0</td>
<td>0.014</td>
<td>999</td>
<td>0.050</td>
<td>0.531</td>
<td>0.255</td>
</tr>
<tr>
<td>HESX1</td>
<td>185</td>
<td>1</td>
<td>0</td>
<td>0.269</td>
<td>999</td>
<td>0.186</td>
<td>0.429</td>
<td>0.180</td>
</tr>
<tr>
<td>HOXA5</td>
<td>270</td>
<td>2</td>
<td>1</td>
<td>0.029</td>
<td>0.445</td>
<td>0.095</td>
<td>1.000</td>
<td>0.833</td>
</tr>
<tr>
<td>HOXD4</td>
<td>255</td>
<td>2</td>
<td>0</td>
<td>0.082</td>
<td>999</td>
<td><strong>0.021</strong></td>
<td>0.288</td>
<td>0.076</td>
</tr>
<tr>
<td>KLF10</td>
<td>390</td>
<td>2</td>
<td>0</td>
<td>0.164</td>
<td>999</td>
<td><strong>0.039</strong></td>
<td>0.229</td>
<td>0.090</td>
</tr>
<tr>
<td>MCPH1</td>
<td>740</td>
<td>9</td>
<td>4</td>
<td>0.112</td>
<td>1.000</td>
<td>0.315</td>
<td>1.000</td>
<td>0.748</td>
</tr>
<tr>
<td>MEOX1</td>
<td>254</td>
<td>2</td>
<td>0</td>
<td>0.127</td>
<td>999</td>
<td>0.075</td>
<td>0.298</td>
<td>0.185</td>
</tr>
<tr>
<td>NR2C1</td>
<td>603</td>
<td>4</td>
<td>0</td>
<td>0.089</td>
<td>999</td>
<td><strong>0.008</strong></td>
<td>0.192</td>
<td>0.309</td>
</tr>
<tr>
<td>Gene</td>
<td>n</td>
<td>dS</td>
<td>dN</td>
<td>dN/dS</td>
<td>p</td>
<td>N</td>
<td>S</td>
<td>p</td>
</tr>
<tr>
<td>--------</td>
<td>---</td>
<td>----</td>
<td>----</td>
<td>-------</td>
<td>-----------</td>
<td>----</td>
<td>----</td>
<td>-------</td>
</tr>
<tr>
<td>NR3C2</td>
<td>984</td>
<td>8</td>
<td>3</td>
<td>0.102</td>
<td>0.770</td>
<td>0.026</td>
<td>0.318</td>
<td>0.069</td>
</tr>
<tr>
<td>PEG3</td>
<td>1590</td>
<td>2</td>
<td>0</td>
<td>0.232</td>
<td>0.944</td>
<td>0.067</td>
<td>0.803</td>
<td>0.177</td>
</tr>
<tr>
<td>PGR</td>
<td>933</td>
<td>2</td>
<td>0</td>
<td>0.197</td>
<td>0.893</td>
<td>0.072</td>
<td>0.469</td>
<td>0.029</td>
</tr>
<tr>
<td>RBM9</td>
<td>367</td>
<td>2</td>
<td>0</td>
<td>0.400</td>
<td>999</td>
<td>0.410</td>
<td>0.184</td>
<td>0.041</td>
</tr>
<tr>
<td>SCML1</td>
<td>208</td>
<td>2</td>
<td>0</td>
<td>1.279</td>
<td>999</td>
<td>0.446</td>
<td>0.231</td>
<td>0.075</td>
</tr>
<tr>
<td>SOHLH</td>
<td>425</td>
<td>1</td>
<td>0</td>
<td>0.484</td>
<td>999</td>
<td>0.011</td>
<td>0.023</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>T</td>
<td>435</td>
<td>2</td>
<td>0</td>
<td>0.052</td>
<td>999</td>
<td>0.068</td>
<td>0.318</td>
<td>0.124</td>
</tr>
<tr>
<td>TAF11</td>
<td>211</td>
<td>2</td>
<td>0</td>
<td>0.138</td>
<td>999</td>
<td>0.031</td>
<td>0.234</td>
<td>0.034</td>
</tr>
<tr>
<td>ZFP37</td>
<td>630</td>
<td>9</td>
<td>0</td>
<td>0.237</td>
<td>999</td>
<td>0.011</td>
<td>0.032</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>ZNF192</td>
<td>578</td>
<td>4</td>
<td>1</td>
<td>0.210</td>
<td>1.605</td>
<td>0.270</td>
<td>0.411</td>
<td>0.101</td>
</tr>
</tbody>
</table>

*Substitutions between the inferred human-chimpanzee ancestor and the human sequence.  N=nonsynonymous, S=synonymous.  †Values of 999 for dNdS indicate dS=0, so dN/dS is undefined.  Values in bold reflect p<0.05.