doi:10.1093/nar/gkm764

Human PAML browser: a database of positive selection on human genes using phylogenetic methods

Gabrielle C. Nickel, David Tefft and Mark D. Adams*

Department of Genetics, Case Western Reserve University, Cleveland, OH, USA

Received August 15, 2007; Revised September 4, 2007; Accepted September 11, 2007

ABSTRACT

With the recent increase in the number of mammalian genomes being sequenced, large-scale genome scans for human-specific positive selection are now possible. Selection can be inferred through phylogenetic analysis by comparing the rates of silent and replacement substitution between related species. Maximum-likelihood (ML) analysis of codon substitution models can be used to identify genes with an accelerated pattern of amino acid substitution on a particular lineage. However, the ML methods are computationally intensive and awkward to configure. We have created a database that contains the results of tests for positive selection along the human lineage in 13 721 genes with orthologs in the UCSC multispecies genome alignments. The Human PAML Browser is a resource through which researchers can search for a gene of interest or groups of genes by Gene Ontology category, and obtain coding sequence alignments for the gene and as well as results from tests of positive selection from the software package Phylogenetic Analysis by Maximum Likelihood. The Human PAML Browser is available at http://mendel.gene.cwru.edu/adamslab/pbrowser.py.

INTRODUCTION

How are humans so genetically similar to the great apes, yet so phenotypically divergent from the other members of this family? This question has intrigued researchers for decades (1,2). One approach to identifying genetic determinants of phenotypic divergence is to examine protein-coding genes for evidence of positive selection (3,4). This type of selection is characterized by a new allele that offers a fitness advantage to an organism, and is rapidly pulled through the population until fixation of the beneficial allele. Genes that have been positively selected along the human lineage, yet remain constrained or selectively neutral in our closest living relatives may offer insight into the biologically significant genetic changes that have occurred since the *Pan–Homo* split.

A variety of methods have been developed for the prediction of positive selection. Some take advantage of population-specific genetic patterns. One such method uses allele frequency differences between populations (5) to uncover loci that have been affected by a genetic hitchhiking event. Similarly, the extended haplotype heterozygosity method (6) measures linkage disequilibrium between two markers with the intent of uncovering a recent selective sweep and is particularly useful in uncovering population-specific selection. Another technique compares the rate of polymorphism within a species to the rate of divergence (fixed difference) between species (7). While these methods are extremely useful for ascertaining genes subject to recent positive selection, they are unable to fully uncover those very ancient and fundamental changes that occurred around the time of human-chimp divergence and are shared by all human populations. For this purpose, analysis of evolutionary rates in a phylogenetic context (3,8) can be used to identify genes that are evolving more rapidly on a particular lineage compared to the rest of the tree (9). The pattern of codon substitution across a phylogenetic tree can be inferred from a multiple sequence alignment using maximum-likelihood methods (10). When the rate of non-synonymous codon changes (d*N*) exceeds the rate of synonymous codon changes (d*S*), positive selection can be inferred. The *codeml* program from the Phylogenetic Analysis by Maximum Likelihood (PAML) package can be used to test different codon substitution models and perform a likelihood ratio test of positive selection along specified lineages based on the d*N*/d*S* ratio (11).

*To whom correspondence should be addressed. Tel: +1 216 368 2791; Fax: +1 216 368 3432; Email: markadams@case.edu or mda13@case.edu
Present address:
David Tefft, The Broad Institute, Cambridge, MA, USA.

Anisimova *et al.* (12,13) have performed large-scale simulation studies to test the effect that parameters such as the number of species, branch lengths, sequence length and sequence divergence has on the tests of positive selection as implemented in PAML. They concluded that predictions of positive selection are unreliable when the sequences being compared are highly similar and when only a small number of species is used. To expand the power and accuracy of the predictions, it is suggested that the number of lineages used in the analysis is increased. Finally, they conclude that multiple models should be used in any analysis of selection in order to ensure robustness in the predictions and to protect against spurious results. In accordance with these findings, our study was designed to include the maximum number of mammalian sequences available to increase the power to detect selection in sequences as similar as human and chimpanzee. Multiple models have also been included in the database to help differentiate between positive selection and relaxation of selective constraint.

Essential to the prediction of selection by phylogenetic analysis is the availability of sequence data from a variety of species. Multispecies alignments for orthologous protein-coding genes are used to infer the ancestral sequence at each internal node within a phylogenetic tree that is necessary in the calculation of the rate and direction of codon substitution. In this way, differences in selective constraint can be predicted on one or many lineages within a phylogeny. Currently, the National Human Genome Research Institute (NHGRI) has approved 43 mammalian species as sequencing targets, many of which are currently underway (http://www.genome.gov/10002154). Five of these have been completed or are in genome refinement (14–20), 21 are slated for draft assembly and the remaining 17 are to be sequenced at low (~2X) coverage. The Genome Browser group at UC Santa-Cruz has produced full-genome multisequence alignments using all of the available vertebrate genome assemblies (21). With the increasing number of genomes being sequenced, phylogenetic analyses can now be performed on a genome-wide scale.

We have collected alignments containing multiple mammalian species for 13 721 orthologous protein-coding genes and examined each for evidence of human-specific positive selection using phylogenetic analysis. The multispecies alignments and the results from the genome-wide selection scan are housed in a web-accessible database named the Human PAML Browser. Users can search by gene or gene family and obtain results from likelihood tests of positive selection on the gene of interest.

These types of analyses can be computationally intensive and time consuming, and the database offers an alternative for many researchers to investigate selection without the difficulty of performing the analysis themselves. The wide variety of species represented in the database also avoids the need to sequence multiple organisms for a comprehensive analysis. The data presented in the Human PAML Browser provides the opportunity for researchers to easily examine their gene(s) of interest for human-specific positive selection and may be a stepping stone for many future studies of selection and its effect on the human genome.

DATA SOURCES AND PROCESSING

Input data

The availability of pre-computed genome alignments for a diverse set of mammalian species represents an excellent starting point for phylogenetic analysis of the pattern of selection operating on individual genes. An assumption of phylogenetic analysis is that the aligned sequences are in fact orthologous. Several groups have constructed sets of orthologous genes (22–25), but the genome-based alignments have certain advantages. Orthology relationships in the multiple alignments are in the context of genomic segments, rather than inferred on the basis of protein alignments, and are thus expected to be reasonably robust. Studies comparing primary reads and finished sequence from a selection of mammals with the human genome suggests that >97% of alignable sequences match at orthologous locations (26). Furthermore, coding sequence information from incomplete (and incompletely annotated) genomes can be used that is not available in protein sequence databases. The disadvantage of using genome alignments is that they do not completely account for lineage-specific duplication and deletion, which can make inference of orthology difficult (27–29).

Alignments of 16 vertebrate species with the human reference sequence were downloaded from the UCSC Genome Browser [http://hgdownload.cse.ucsc.edu/goldenPath/hg18/multiz17way/, (21)]. The genome assembly multiple alignments were constructed from draft genome assemblies from the following organisms: chimpanzee, rhesus macaque, mouse, rat, rabbit, dog, cow, armadillo, elephant, tenrec, opossum, chicken, frog, zebrafish, tetraodon and fugu. Initial results using protein-coding sequence (CDS) coordinates resulted in an unacceptable number of frameshifts, presumably due to small alignment errors near exon/intron junctions. We therefore determined CDS regions by alignment of the longest representative RefSeq mRNA for each gene to the human genome. CDS alignments were then extracted representing all of the mammalian species from the genomic alignment files, representing most mammalian orders including the subclass Metatheria (Figure 1). For *codeml* analysis, alignments were extracted corresponding to each mammalian species with sequence data at a given position. Total 13 721 genes were analyzed; 12 905 of these had at least a portion of the sequence represented in at least eight species. At the time of the analysis, more than 5× coverage of the orangutan genome was available as shotgun sequence reads in the NCBI Trace Archive, but these data had not yet been assembled. We felt that including all available primate sequence was important, so we added inferred protein coding sequence from orangutan to the mammalian CDS alignments by comparing human genomic sequence to the orangutan Trace Archive at NCBI, initially using BLAST, then assembling reads using *phrap* (www.phrap.org) and producing final alignments with the *lalign* algorithm implemented in *matcher* from the EMBOSS package (30,31), and extracting CDS regions based on exon coordinates from the human genomic sequence. Inclusion of orangutan sequence in the alignments had only minimal impact on the human-specific dN/dS, but improved prediction of human-specific substitutions when using the branch + site models described below (data not shown). Columns with gaps in the human sequence were removed from the alignment to facilitate analysis.

Statistical analysis

Markov process models of codon substitution, as implemented in the PAML software package (11), were used to analyze the selective pressures affecting each gene along the human lineage. The PAML software package is available at http://abacus.gene.ucl.ac.uk/software/paml.html. The program *codeml* v3.14b within the PAML package was used to analyze the data. *codeml* allows specification of several different codon substitution models that allow testing of hypotheses related to selection along certain branches of the tree and/or at a subset of codons (sites).

For *codeml* analysis, a directory was made for each gene containing the multiple sequence alignment in PHYLIP format, *codeml* control files, and an appropriate phylogenetic tree file for the species represented in the alignment. Species tree files were constructed from the consensus mammalian tree (Figure 1) (32) essentially by pruning branches that were not represented for a given gene. *codeml* was run on an Apple OSX compute cluster consisting of 20 dual-processor nodes, each equipped with 2 GB of RAM. The compute time for *codeml* analyses under all five models for the 13 721 alignments was 19 days using an average of 30 CPUs concurrently. Following *codeml* analysis, a python script parsed the *codeml* results and alignment files and loaded information to a MySQL database. A web interface was developed using python to provide interactive access to the *codeml* results on a gene by gene basis.

For each gene, *codeml* was run under two test models and three null models, and results from each were then compared as tests of positive selection on the human gene. Each run produced a maximum-likelihood estimate, which is the probability of observing the data under the evolutionary conditions implemented in the model. A likelihood ratio test (LRT) was used to determine whether the test model is a significantly better fit to the data than the null model. A P-value was calculated by comparing two times the difference in log likelihood values to a chi-squared distribution, with the degrees of freedom equal to the difference in number of parameters between the pair of nested tests. The branch model allows the dN/dS ratio on a specified branch of the tree to differ from the average dN/dS ratio across the rest of the tree. The matching null model fixes dN/dS = 1 on the specified branch. If dN/d$S > 1$ in the branch model and the LRT is significant, positive selection can be inferred. For tests of selection on the human branch, we refer to these branch models as model H and model Hnull, respectively. A particular problem in the branch tests, which is due to the short evolutionary time since the human-chimpanzee divergence, occurs when there are only non-synonymous substitutions on the branch between human and the human-chimpanzee ancestor. In this situation, the rate of synonymous substitution dS = 0 and dN/dS is undefined; this is represented in the database as 999. The branch test also requires dN/dS to be elevated across the entire gene, and therefore is considered somewhat unrealistic, particularly for multi-domain proteins where positive selection may be acting on only one domain.

The branch + site models were developed to address positive selection at a subset of sites (codons) on one branch of the tree (33,34). Model A defines four classes of sites, where two of the classes have dN/d$S \leq 1$ on all branches while two additional classes have dN/d$S > 1$ on the lineage of interest (human), but dN/d$S \leq 1$ on the other branches of the tree. Model Anull fixes dN/dS = 1 for the latter two classes and thus comparison of Model A

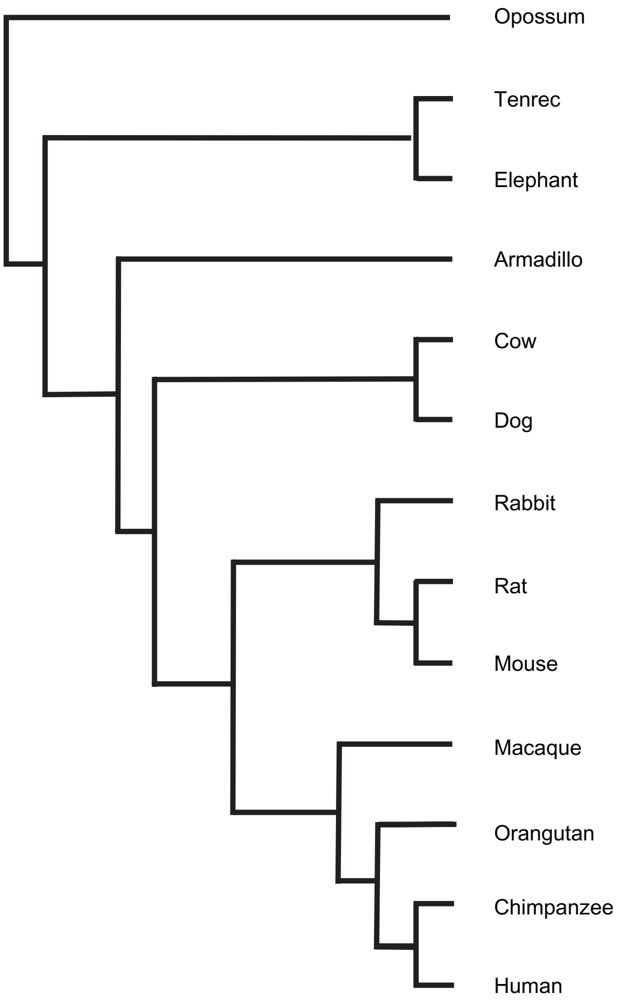


Figure 1. Unrooted mammalian species tree of the organisms used in the construction of the database and the phylogenetic tree used in PAML analysis. Orientation adapted from the UCSC Genome Browser.

Table 1. Evolutionary Models used by *codeml*

Model	Description				
Model H	Non-neutral model: human ω^{a} allowed to vary from other branches $\omega_{\mathrm{HUMAN}} \neq \omega_{\mathrm{OTHERS}}$				
Model Hnull	Neutral model: human ω fixed at $\omega = 1$ and allowed to vary from other branches $\omega_{\mathrm{HUMAN}} = 1$ and $\omega_{\mathrm{HUMAN}} \neq \omega_{\mathrm{OTHERS}}$				
Model A	Sites on the human branch allowed to differ				
Background lineages	4 site classes:	$0 < \omega_0 < 1$	$\omega_1 = 1$	$0 < \omega_{2a} < 1$	$\omega_{2b} = 1$
Foreground lineage-human	4 site classes:	$0 < \omega_0 < 1$	$\omega_1 = 1$	$\omega_{2a} \geq 1$	$\omega_{2b} \geq 1$
Model Anull	Sites on the human branch fixed at $\omega = 1$				
Background lineages	4 site classes:	$0 < \omega_0 < 1$	$\omega_1 = 1$	$0 < \omega_{2a} < 1$	$\omega_{2b} = 1$
Foreground lineage-human	4 site classes:	$0 < \omega_0 < 1$	$\omega_1 = 1$	$\omega_{2a} = 1$	$\omega_{2b} = 1$
Model 1a	Sites on all branches nearly neutral				
All lineages	2 site classes: $0 < \omega_0 < 1$, $\omega_1 = 1$				
Compare	Likelihood ratio tests of significance				
Model H, Model Hnull	Branch test: non-neutral evolution				
MA, M1a	Relaxed branch + site test: positive selection or relaxation of selective constraint				
MA, MAnull	Strict branch + site test: positive selection				

[a] $\omega = dN/dS$.

with Model Anull is a strict test of positive selection. Model 1a assumes $dN/dS \leq 1$ at all sites across all branches. A significant LRT in the comparison of Model A with Model 1a can be due to either positive selection or relaxation of selective constraint, since there is no formal test of whether dN/dS is >1 on the human branch, just whether there is a subset of sites on the human branch with a dN/dS ratio that is elevated relative to the rest of the tree. A summary of the models and likelihood ratio tests performed in this analysis are found in Table 1 and more thorough information is available in the PAML documentation. A small number of genes have also been run under site models in which dN/dS is allowed to vary among sites, instead of between lineages. This is helpful in interpreting whether positive selection is occurring specifically along the human lineage or if the gene is rapidly evolving in multiple species.

RESULTS OF TESTS OF SELECTION

Table 2 presents a summary of the results of tests of positive selection for human genes. As expected, the branch test, which requires dN/dS to exceed 1 over the entire coding sequence, returned few significant results. In the strict branch + site test that compares results from Models A and Anull, 244 genes met the nominal threshold for significance of $P < 0.05$. More than twice as many genes met the significance threshold in a comparison of Model A with Model 1a, which can indicate a relaxation of selective constraint or positive selection. Gene Ontology categories were matched to the 244 genes with $P < 0.05$ in the strict branch + site test to assess whether certain categories might be over-represented among genes predicted to be positively selected (Table 3). As has been reported previously (17,35,36), transcription factors and olfactory receptors are over-represented among positively selected genes. Reflecting the abundance of transcription factors, the cellular component category most over-represented is the nucleus.

Table 2. Summary of results of tests of selection on human genes

Test	Significance threshold	Number of genes[a]
Model A versus Model Anull	0.05	244
[strict branch + site test]	0.01	152
	0.001	48
Model A versus Model 1a	0.05	611
[relaxed branch + site test]	0.01	276
	0.001	114
Model H versus Model Hnull	0.05	16
[strict branch test]	0.01	3
	0.001	2

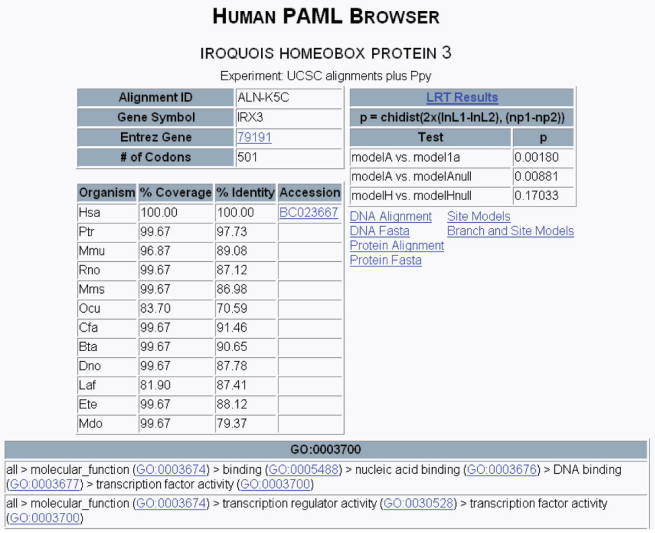
[a] OR5B3 is the only gene with $P < 0.05$ in both branch (Model H versus Model Hnull) and branch + site (Model A versus Model Anull) tests.

Several groups have performed genome scans for positive selection on the human lineage using divergence data (17,35,37,38). One feature that sets this study apart is the large number and wide variety of mammalian species used. When only a small number of species is used (e.g. human–chimpanzee–mouse or human–chimpanzee–macaque) or outgroups that are too divergent (35,38,39) there can be a high degree of uncertainty as to whether a substitution occurred on the human or chimpanzee lineage. This can lead to both an increase in the false positive rate, as well as a lack of power to detect selection if multiple substitutions have occurred at the same codon. The inclusion of a third great ape, the orangutan, the Old World monkey rhesus macaque, and several non-primate mammals adds considerable power to model the substitution pattern. As recent studies have concluded (35,37), there is a relative paucity of human-specific positively selected genes and these

Table 3. Gene ontology categories over-represented among genes with $P < 0.05$ in the strict branch + site test

Gene Ontology ID	Category name	Corrected *P*-value
Biological process		
GO:0006350	Transcription	0.003
GO:0006355	Regulation of transcription, DNA-dependent	0.003
GO:0007608	Sensory perception of smell	0.02
GO:0006814	Sodium ion transport	0.02
GO:0007165	Signal transduction	0.02
Molecular function		
GO:0005515	Protein binding	0
GO:0046872	Metal ion binding	0.00005
GO:0016740	Transferase activity	0.00009
GO:0008270	Zinc ion binding	0.001
GO:0003677	DNA binding	0.001
GO:0000166	Nucleotide binding	0.001
GO:0003676	Nucleic acid binding	0.006
GO:0004984	Olfactory receptor activity	0.006
GO:005524	ATP binding	0.01
GO:0031402	Sodium ion binding	0.02
Cellular component		
GO:0005634	Nucleus	0
GO:0016020	Membrane	0.0004
GO:0005737	Cytoplasm	0.002

Gene Ontology classification was performed using Onto-Express (43,44).



HUMAN PAML BROWSER

IROQUOIS HOMEOBOX PROTEIN 3

Experiment: UCSC alignments plus Ppy

Alignment ID	ALN-K5C
Gene Symbol	IRX3
Entrez Gene	79191
# of Codons	501

Organism	% Coverage	% Identity	Accession
Hsa	100.00	100.00	BC023667
Ptr	99.67	97.73	
Mmu	96.87	89.08	
Rno	99.67	87.12	
Mms	99.67	86.98	
Ocu	83.70	70.59	
Cfa	99.67	91.46	
Bta	99.67	90.65	
Dno	99.67	87.78	
Laf	81.90	87.41	
Ete	99.67	88.12	
Mdo	99.67	79.37	

LRT Results

p = chidist(2x(lnL1-lnL2), (np1-np2))

Test	p
modelA vs. model1a	0.00180
modelA vs. modelAnull	0.00881
modelH vs. modelHnull	0.17033

DNA Alignment
DNA Fasta
Protein Alignment
Protein Fasta
Site Models
Branch and Site Models

GO:0003700

all > molecular_function (GO:0003674) > binding (GO:0005488) > nucleic acid binding (GO:0003676) > DNA binding (GO:0003677) > transcription factor activity (GO:0003700)

all > molecular_function (GO:0003674) > transcription regulator activity (GO:0030528) > transcription factor activity (GO:0003700)

Figure 2. PAML database summary results for Iroquois homeobox 3 (IRX3). The likelihood ratio tests are as follows: the relaxed branch + site test (Model A versus Model 1a), the strict branch + site test (Model A versus Model Anull) and the branch test (Model H versus Model Hnull). The three letter species codes are Hsa (human), Ptr (chimp), Mmu (macaque), Rno (rat), Mms (mouse), Ocu (rabbit), Cfa (dog), Bta (cow), Dno (armadillo), Laf (elephant), Ete (tenrec) and Mdo (opossum).

events appear to be relatively rare in the genome. Bakewell (37) suggests that the small long-term effective population size of humans as they migrated out of Africa may be masking positive selection, as many advantageous alleles have yet to become fixed in the human population.

USER INTERFACE

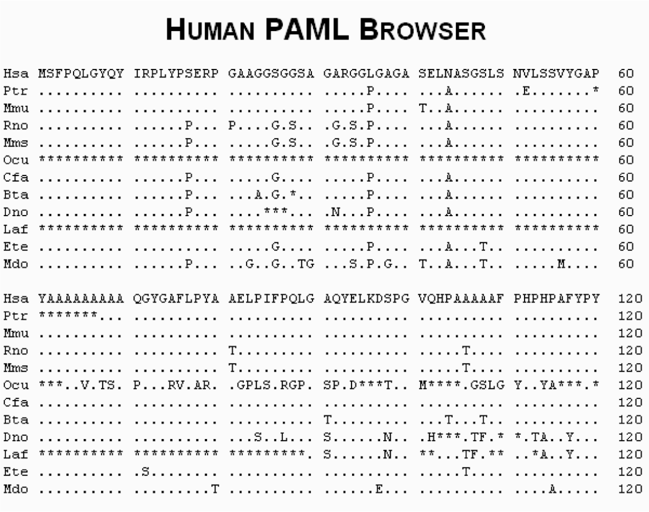
The Human PAML Browser can be accessed at http://mendel.gene.cwru.edu/adamslab/pbrowser.py and there are a number of ways to query the database. A gene of interest can be searched for by gene symbol, Entrez Gene ID, mRNA accession (RefSeq) and gene name. To examine genes by their biological process, molecular function or cellular component, Gene Ontology IDs (http://www.geneontology.org/) may also be used to query the database. For sets of genes with similar names, a wildcard * can be inserted. A more recent addition is the ability to query the database by statistical significance. The user can set the threshold *P*-value for one of the three likelihood ratio tests to extract genes by significance. The return screen from a query contains one or more genes related to the search, and the user has the ability to chose the alignment set used in the analysis, which is primarily the UCSC alignments plus orangutan.

Database organization

Upon gene choice, the user is presented with a summary screen from the PAML analysis (Figure 2). This screen contains direct links to the Entrez ID, the mRNA accession and the GO terms for the gene. The organisms used in this analysis are ordered by evolutionary distance from humans, and information is given on the percent coverage and average percent identity of the aligned sequence from each organism as compared to the orthologous human sequence. The results from the likelihood ratio tests in the form of a *P*-value are displayed for all three tests of positive selection: the relaxed branch + site test, the strict branch + site test and the branch test. Also contained are links to the DNA and protein multispecies alignments (Figure 3) as well as the sequences in FASTA format to facilitate subsequent analysis. It is of note that an asterisk in the protein alignments refer to an unknown amino acid or gap in the sequence, not a stop codon.

The bulk of the data is contained under the Branch and Site Models link (Figure 4). To the right is the table of results from the branch Models H and Hnull (Figure 4A). The numbers under the branch heading refer to the relationship of the branch from an internal node to a terminal node or to another internal node. Terminal branches are labeled by organism. One dN/dS ratio is present across the entire tree with the exception of the terminal branch to human (Hsa). dN/dS on the human branch is calculated in Model H and fixed at 1 in Model Hnull.

Figure 4B contains the results from branch + site Model A. Model A is run under the assumption of four site classes: class 0, sites under negative selection in all branches; class 1, sites evolving neutrally in all branches; class 2a, sites positively selected on the human branch,



HUMAN PAML BROWSER

```
Hsa MSFPQLGYQY IRPLYPSERP GAAGGSGGSA GARGGLGAGA SELNASGSLS NVLSSVYGAP  60
Ptr .......... .......... .......... .....P.... ...A...... .E.......*  60
Mmu .......... .......... .......... .....P.... T..A...... ..........  60
Rno .......... ......P... P....G.S.. .G.S.P.... ...A...... ..........  60
Mms .......... ......P... .....G.S.. .G.S.P.... ...A...... ..........  60
Ocu ********** ********** ********** ********** ********** **********  60
Cfa .......... ......P... .....G.... .....P.... ...A...... ..........  60
Bta .......... ......P... ...A.G.*.. .....P.... ...A...... ..........  60
Dno .......... ......P... ....***... .N...P.... ...A...... ..........  60
Laf ********** ********** ********** ********** ********** **********  60
Ete .......... .......... .....G.... .....P.... ...A...T.. ..........  60
Mdo .......... ......P... ..G..G..TG ...S.P.G.. T..A...T.. .....M....  60

Hsa YAAAAAAAAA QGYGAFLPYA AELPIFPQLG AQYELKDSPG VQHPAAAAAF PHPHPAFYPY  120
Ptr *******... .......... .......... .......... .......... ..........  120
Mmu .......... .......... .......... .......... .......... ..........  120
Rno .......... .......... T......... .......... .....T.... ..........  120
Mms .......... .......... T......... .......... .....T.... ..........  120
Ocu ***..V.TS. P...RV.AR. .GPLS.RGP. SP.D***T.. M****.GSLG Y..YA***.*  120
Cfa .......... .......... .......... .......... .......... ..........  120
Bta .......... .......... .......... T......... ...T...T.. ..........  120
Dno .......... .......... ...S..L... S......N.. .H***.TF.* *.TA..Y...  120
Laf ********** ********** *********. S......N.. **...TF.** ..*A..Y...  120
Ete .......... .S........ .......... .......... .....T.... ..........  120
Mdo .......... .........T .......... ......E... .......... ....A.....  120
```

Figure 3. Multispecies protein alignment for IRX3. Coding sequence was extracted from whole-genome sequence alignments available from the UCSC Genome Browser and trimmed to include only coding regions based on matches with the human CDS. Alignment columns with gaps in the human sequence were excluded. Residues 1–180 of 501 are shown. Residues 36 and 44 were predicted to be positively selected in human with Bayes Empirical Bayes posterior probability of 0.886 and 0.997, respectively (Figure 4B).

but negatively selected on the other branches and class 2b, sites positively selected on the human branch, but neutrally evolving on the other branches. The proportion refers to the fraction of sites within the protein that fall into each of the four site classes. Foreground and background ω is the dN/dS ratio for the human lineage and all the other lineages, respectively. The version of *codeml* implemented uses a Bayes Empirical Bayes (BEB) method for calculating the posterior probability that each site is from a particular site class; sites with a high posterior probability from site class 2a and 2b are inferred to be positively selected (34). A summary table shows the total number of sites predicted to be positively selected as well as the number with a BEB posterior probability $\geq$95 and $\geq$99%. Finally, all sites predicted to be positively selected are listed along with the residue number, the affected amino acid and the BEB probability specific to that residue. The results from the null Model 1a can be retrieved through the Site Models link on the original summary page (Figure 2).

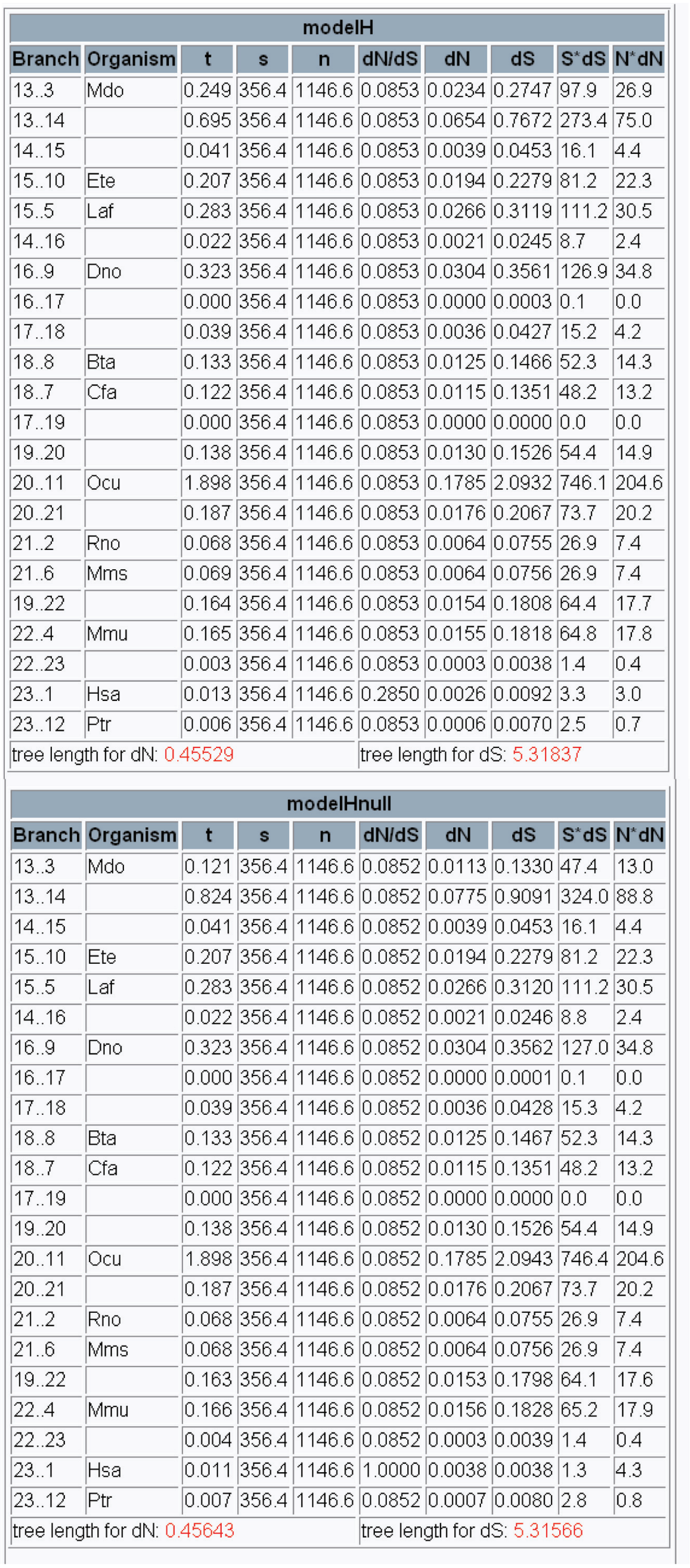
IMPORTANT LIMITATIONS

There are several limitations of the data and analyses presented in the Human PAML Browser that the user should consider. The results presented in the browser were obtained using an automated process at multiple steps from alignment creation to extraction of coding sequences, to maximum-likelihood analysis, to database loading and web display. Most alignments and result sets have not been examined manually or curated in any way. Bad alignments lead to bad phylogenetic inferences. In particular, some genes with very low P-values in tests of selection appear to have alignment problems with many human-specific substitutions clustered at adjacent residues. Incorrect assignment of orthology could also result in misleading P-values. The whole-genome alignments used here have the advantage of aligning coding sequences in a larger (often much larger) syntenic framework, but a gene family approach that accounts for gene duplication and deletion events would be a useful adjunct to interpretation of the information presented in the Human PAML Browser. Differential gene loss following duplication can lead to 1:1 paralogs being mistaken for 1:1 orthologs. This problem is particularly acute in organisms, such as yeast, which have experienced whole-genome duplication events (40), but could also affect vertebrate alignments, particularly given the extent of lineage-specific segmental duplication in primates (41). Gene trees have not been constructed for the alignments used here, but reconciliation of the gene tree with the species tree using a program such as NOTUNG (42) would be a useful exercise in the context of follow-up study.

The P-values in the Human PAML Browser have not been corrected for multiple tests, and so care should be taken in interpretation of results. Another factor that potentially impacts the P-values is that in some cases, the ML methods fail to converge to a global optimum, resulting in an inappropriately low P-value. Finally, genome assemblies and assembly alignments are improving over time. It is strongly recommended that the user validate the results presented here for any gene of interest by re-extracting gene sequences and repeating the *codeml* analysis. This will serve the dual purposes of incorporating new and improved genome assembly data and ensuring that the ML analysis is stable.

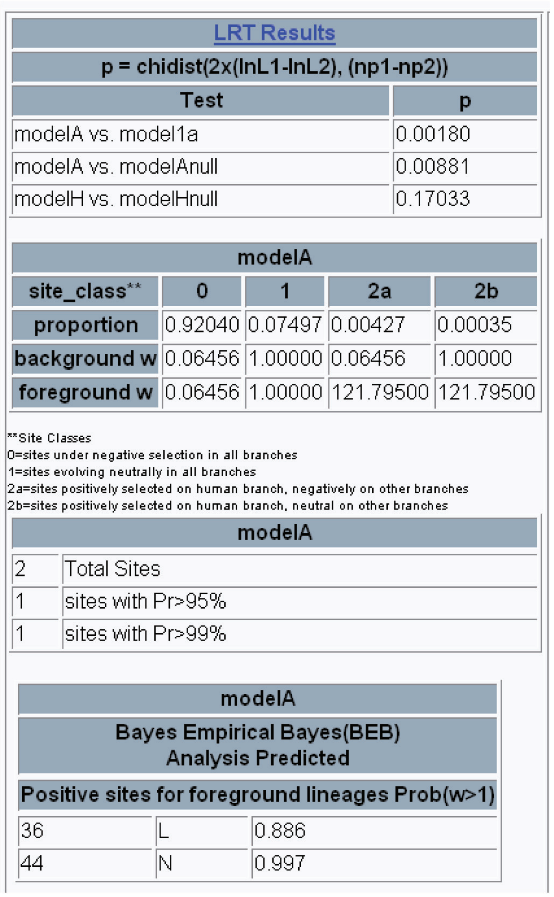
CONCLUSION

We have generated a database containing the PAML results for tests of human-specific positive selection. The simple web interface makes access to PAML results readily available, compared to the tasks of preparing properly formatted alignment, phylogenetic tree, and *codeml* control files and parsing five different *codeml* output files. The multispecies alignments used in each analysis are readily available in FASTA file format for further analysis using other methods or to examine selection among other orders/families of mammals or in individual species. The Human PAML Browser is intended to aide other researchers as they search for the selective pressures that have affected their gene or gene family of interest, and can be a stepping stone for many future studies of positive selection and the human genome.



modelH									
Branch	Organism	t	s	n	dN/dS	dN	dS	S*dS	N*dN
13..3	Mdo	0.249	356.4	1146.6	0.0853	0.0234	0.2747	97.9	26.9
13..14		0.695	356.4	1146.6	0.0853	0.0654	0.7672	273.4	75.0
14..15		0.041	356.4	1146.6	0.0853	0.0039	0.0453	16.1	4.4
15..10	Ete	0.207	356.4	1146.6	0.0853	0.0194	0.2279	81.2	22.3
15..5	Laf	0.283	356.4	1146.6	0.0853	0.0266	0.3119	111.2	30.5
14..16		0.022	356.4	1146.6	0.0853	0.0021	0.0245	8.7	2.4
16..9	Dno	0.323	356.4	1146.6	0.0853	0.0304	0.3561	126.9	34.8
16..17		0.000	356.4	1146.6	0.0853	0.0000	0.0003	0.1	0.0
17..18		0.039	356.4	1146.6	0.0853	0.0036	0.0427	15.2	4.2
18..8	Bta	0.133	356.4	1146.6	0.0853	0.0125	0.1466	52.3	14.3
18..7	Cfa	0.122	356.4	1146.6	0.0853	0.0115	0.1351	48.2	13.2
17..19		0.000	356.4	1146.6	0.0853	0.0000	0.0000	0.0	0.0
19..20		0.138	356.4	1146.6	0.0853	0.0130	0.1526	54.4	14.9
20..11	Ocu	1.898	356.4	1146.6	0.0853	0.1785	2.0932	746.1	204.6
20..21		0.187	356.4	1146.6	0.0853	0.0176	0.2067	73.7	20.2
21..2	Rno	0.068	356.4	1146.6	0.0853	0.0064	0.0755	26.9	7.4
21..6	Mms	0.069	356.4	1146.6	0.0853	0.0064	0.0756	26.9	7.4
19..22		0.164	356.4	1146.6	0.0853	0.0154	0.1808	64.4	17.7
22..4	Mmu	0.165	356.4	1146.6	0.0853	0.0155	0.1818	64.8	17.8
22..23		0.003	356.4	1146.6	0.0853	0.0003	0.0038	1.4	0.4
23..1	Hsa	0.013	356.4	1146.6	0.2850	0.0026	0.0092	3.3	3.0
23..12	Ptr	0.006	356.4	1146.6	0.0853	0.0006	0.0070	2.5	0.7
tree length for dN: 0.45529					tree length for dS: 5.31837				

modelHnull									
Branch	Organism	t	s	n	dN/dS	dN	dS	S*dS	N*dN
13..3	Mdo	0.121	356.4	1146.6	0.0852	0.0113	0.1330	47.4	13.0
13..14		0.824	356.4	1146.6	0.0852	0.0775	0.9091	324.0	88.8
14..15		0.041	356.4	1146.6	0.0852	0.0039	0.0453	16.1	4.4
15..10	Ete	0.207	356.4	1146.6	0.0852	0.0194	0.2279	81.2	22.3
15..5	Laf	0.283	356.4	1146.6	0.0852	0.0266	0.3120	111.2	30.5
14..16		0.022	356.4	1146.6	0.0852	0.0021	0.0246	8.8	2.4
16..9	Dno	0.323	356.4	1146.6	0.0852	0.0304	0.3562	127.0	34.8
16..17		0.000	356.4	1146.6	0.0852	0.0000	0.0001	0.1	0.0
17..18		0.039	356.4	1146.6	0.0852	0.0036	0.0428	15.3	4.2
18..8	Bta	0.133	356.4	1146.6	0.0852	0.0125	0.1467	52.3	14.3
18..7	Cfa	0.122	356.4	1146.6	0.0852	0.0115	0.1351	48.2	13.2
17..19		0.000	356.4	1146.6	0.0852	0.0000	0.0000	0.0	0.0
19..20		0.138	356.4	1146.6	0.0852	0.0130	0.1526	54.4	14.9
20..11	Ocu	1.898	356.4	1146.6	0.0852	0.1785	2.0943	746.4	204.6
20..21		0.187	356.4	1146.6	0.0852	0.0176	0.2067	73.7	20.2
21..2	Rno	0.068	356.4	1146.6	0.0852	0.0064	0.0755	26.9	7.4
21..6	Mms	0.068	356.4	1146.6	0.0852	0.0064	0.0756	26.9	7.4
19..22		0.163	356.4	1146.6	0.0852	0.0153	0.1798	64.1	17.6
22..4	Mmu	0.166	356.4	1146.6	0.0852	0.0156	0.1828	65.2	17.9
22..23		0.004	356.4	1146.6	0.0852	0.0003	0.0039	1.4	0.4
23..1	Hsa	0.011	356.4	1146.6	1.0000	0.0038	0.0038	1.3	4.3
23..12	Ptr	0.007	356.4	1146.6	0.0852	0.0007	0.0080	2.8	0.8
tree length for dN: 0.45643					tree length for dS: 5.31566				



LRT Results	
p = chidist(2x(lnL1-lnL2), (np1-np2))	
Test	p
modelA vs. model1a	0.00180
modelA vs. modelAnull	0.00881
modelH vs. modelHnull	0.17033

modelA				
site_class**	0	1	2a	2b
proportion	0.92040	0.07497	0.00427	0.00035
background w	0.06456	1.00000	0.06456	1.00000
foreground w	0.06456	1.00000	121.79500	121.79500

**Site Classes
0=sites under negative selection in all branches
1=sites evolving neutrally in all branches
2a=sites positively selected on human branch, negatively on other branches
2b=sites positively selected on human branch, neutral on other branches

modelA	
2	Total Sites
1	sites with Pr>95%
1	sites with Pr>99%

modelA		
Bayes Empirical Bayes(BEB) Analysis Predicted		
Positive sites for foreground lineages Prob(w>1)		
36	L	0.886
44	N	0.997

Figure 4. Likelihood test data and results for IRX3. (**A**) Branch tests for selection, Model H (test model) and Model Hnull (neutral model). The variables in the table are as follows: *t*, the length of the branch; *s* and *n*, the number of synonymous and non-synonymous sites, respectively; d*N*/d*S*, the ratio of the rate of non-synonymous and synonymous substitution for the branch; d*N* and d*S*, the rate of synonymous and non-synonymous substitution on the branch; *S**d*S* and *N**d*N*, a rough estimate of the absolute number of synonymous and non-synonymous substitutions. (**B**) Branch + site tests for selection, including the proportion of codons in each site class for the background and foreground (human) lineage, and the posterior probability that a given site is positively selected.

ACKNOWLEDGEMENTS

We thank Neil Molyneaux for additional contributions to the Human PAML Browser and Ulrich Neuss for maintaining the compute cluster. Funds for this work were provided by a start-up award to M.D.A. from Case Western Reserve University. Funding to pay Open Access publication charges for this article was provided by Case Western Reserve University.

Conflict of interest statement. None declared.

REFERENCES

1. King,M.-C. and Wilson,A.C. (1975) Evolution at two levels in humans and chimpanzees. *Science*, **188**, 107–116.
2. Carroll,S.B. (2003) Genetics and the making of *Homo sapiens*. *Nature*, **422**, 849–857.
3. Li,W.H., Wu,C.I. and Luo,C.C. (1985) A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Mol. Biol. Evol.*, **2**, 150–174.
4. Kimura,M. (1983) *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge.
5. Tajima,F. (1993) Statistical analysis of DNA polymorphism. *Jpn. J. Genet.*, **68**, 567–595.
6. Sabeti,P.C., Reich,D.E., Higgins,J.M., Levine,H.Z.P., Richter,D.J., Schaffner,S.F., Gabriel,S.B., Platko,J.V., Patterson,N.J. *et al.* (2002) Detecting recent positive selection in the human genome from haplotype structure. *Nature*, **419**, 832–837.
7. McDonald,J.H. and Kreitman,M. (1991) Adaptive protein evolution at the Adh locus in Drosophila. *Nature*, **351**, 652–654.
8. Felsenstein,J. (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.*, **17**, 368–376.
9. Yang,Z. (2002) Inference of selection from multiple species alignments. *Curr. Opin. Genet. Dev.*, **12**, 688–694.
10. Goldman,N. and Yang,Z. (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.*, **11**, 725–736.
11. Yang,Z. (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.*, **13**, 555–556.
12. Anisimova,M., Bielawski,J.P. and Yang,Z. (2001) Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. *Mol. Biol. Evol.*, **18**, 1585–1592.
13. Anisimova,M., Bielawski,J.P. and Yang,Z. (2002) Accuracy and power of bayes prediction of amino acid sites under positive selection. *Mol. Biol. Evol.*, **19**, 950–958.
14. International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
15. Venter,J.C., Adams,M.D., Myers,E.W., Li,P.W., Mural,R.J., Sutton,G.G., Smith,H.O., Yandell,M., Evans,C.A. *et al.* (2001) The sequence of the human genome. *Science*, **291**, 1304–1351.
16. International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. *Nature*, **431**, 931–945.
17. Chimpanzee Sequencing and Analysis Consortium (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, **437**, 69–87.
18. Rat Genome Sequencing Project Consortium (2004) Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature*, **428**, 493–521.
19. Lindblad-Toh,K., Wade,C.M., Mikkelsen,T.S., Karlsson,E.K., Jaffe,D.B., Kamal,M., Clamp,M., Chang,J.L., Kulbokas,E.J. III *et al.* (2005) Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature*, **438**, 803–819.
20. Mouse Genome Sequencing Consortium (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520–562.
21. Blanchette,M., Kent,W.J., Riemer,C., Elnitski,L., Smit,A.F., Roskin,K.M., Baertsch,R., Rosenbloom,K., Clawson,H. *et al.* (2004) Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.*, **14**, 708–715.
22. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., DiCuccio,M., Edgar,R. *et al.* (2007) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **35**, D5–D12.
23. Eppig,J.T., Bult,C.J., Kadin,J.A., Richardson,J.E., Blake,J.A., Anagnostopoulos,A., Baldarelli,R.M., Baya,M., Beal,J.S. *et al.* (2005) The Mouse Genome Database (MGD): from genes to mice – a community resource for mouse biology. *Nucleic Acids Res.*, **33**, D471–D475.
24. Hubbard,T.J., Aken,B.L., Beal,K., Ballester,B., Caccamo,M., Chen,Y., Clarke,L., Coates,G., Cunningham,F. *et al.* (2007) Ensembl 2007. *Nucleic Acids Res.*, **35**, D610–D617.
25. Lee,Y., Sultana,R., Pertea,G., Cho,J., Karamycheva,S., Tsai,J., Parvizi,B., Cheung,F., Antonescu,V. *et al.* (2002) Cross-referencing eukaryotic genomes: TIGR Orthologous Gene Alignments (TOGA). *Genome Res.*, **12**, 493–502.
26. Margulies,E.H., Vinson,J.P., Miller,W., Jaffe,D.B., Lindblad-Toh,K., Chang,J.L., Green,E.D., Lander,E.S., Mullikin,J.C. *et al.* (2005) An initial strategy for the systematic identification of functional elements in the human genome by low-redundancy comparative sequencing. *Proc. Natl Acad. Sci. USA*, **102**, 4795–4800.
27. Ohno,S. (1970) *Evolution by Gene Duplication*. Springer-Verlag, Heidelberg.
28. Prince,V.E. and Pickett,F.B. (2002) Splitting pairs: the diverging fates of duplicated genes. *Nat. Rev. Genet.*, **3**, 827–837.
29. Roth,C., Rastogi,S., Arvestad,L., Dittmar,K., Light,S., Ekman,D. and Liberles,D.A. (2007) Evolution after gene duplication: models, mechanisms, sequences, systems, and organisms. *J. Exp. Zoolog. B Mol. Dev. Evol.*, **308**, 58–73.
30. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
31. Pearson,W.R. and Lipman,D.J. (1988) Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.
32. Murphy,W.J., Eizirik,E., O'Brien,S.J., Madsen,O., Scally,M., Douady,C.J., Teeling,E., Ryder,O.A., Stanhope,M.J. *et al.* (2001) Resolution of the early placental mammal radiation using Bayesian phylogenetics. *Science*, **294**, 2348–2351.
33. Yang,Z. and Nielsen,R. (2002) Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol. Biol. Evol.*, **19**, 908–917.
34. Yang,Z., Wong,W.S. and Nielsen,R. (2005) Bayes empirical Bayes inference of amino acid sites under positive selection. *Mol. Biol. Evol.*, **22**, 1107–1118.
35. Arbiza,L., Dopazo,J. and Dopazo,H. (2006) Positive selection, relaxation, and acceleration in the evolution of the human and chimp genome. *PLoS Comput. Biol.*, **2**, e38.
36. Bustamante, C.D., Fledel-Alon, A., Williamson, S., Nielsen, R., Todd Hubisz, M., Glanowski, S., Tanenbaum, D.M., White, T.J., Sninsky, J.J., Hernandez, R.D. *et al.*. (2005) Natural selection on protein-coding genes in the human genome. *Nature*, **437**, 1153–1157.
37. Bakewell,M.A., Shi,P. and Zhang,J. (2007) More genes underwent positive selection in chimpanzee evolution than in human evolution. *Proc. Natl Acad. Sci. USA*, **104**, 7489–7494.
38. Clark, A.G., Glanowski, S., Nielsen, R., Thomas, P.D., Kejariwal, A., Todd, M.A., Tanenbaum, D.M., Civello, D., Lu, F., Murphy, B. *et al.* (2003) Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. *Science*, **302**, 1960–1963.
39. Gibbs,R.A., Rogers,J., Katze,M.G., Bumgarner,R., Weinstock,G.M., Mardis,E.R., Remington,K.A., Strausberg,R.L., Venter,J.C. *et al.* (2007) Evolutionary and biomedical insights from the rhesus macaque genome. *Science*, **316**, 222–234.
40. Scannell,D.R., Frank,A.C., Conant,G.C., Byrne,K.P., Woolfit,M. and Wolfe,K.H. (2007) Independent sorting-out of thousands of duplicated gene pairs in two yeast species descended from a whole-genome duplication. *Proc. Natl Acad. Sci. USA*, **104**, 8397–8402.

41. Bailey,J.A. and Eichler,E.E. (2006) Primate segmental duplications: crucibles of evolution, diversity and disease. *Nat. Rev. Genet.*, **7**, 552–564.
42. Chen,K., Durand,D. and Farach-Colton,M. (2000) NOTUNG: a program for dating gene duplications and optimizing gene family trees. *J. Comput. Biol.*, **7**, 429–447.
43. Draghici,S., Khatri,P., Bhavsar,P., Shah,A., Krawetz,S.A. and Tainsky,M.A. (2003) Onto-Tools, the toolkit of the modern biologist: Onto-Express, Onto-Compare, Onto-Design and Onto-Translate. *Nucleic Acids Res.*, **31**, 3775–3781.
44. Khatri,P., Bhavsar,P., Bawa,G. and Draghici,S. (2004) Onto-Tools: an ensemble of web-accessible, ontology-based tools for the functional design and interpretation of high-throughput gene expression experiments. *Nucleic Acids Res.*, **32**, W449–W456.