Text supplement for “Promoter regions of many neural- and nutrition-related genes have experienced positive selection during human evolution” by R. Haygood, O. Fedrigo, B. Hanson, K.-D. Yokoyama, and G. A. Wray

Materials and methods summary

This section summarizes our materials and methods. The next section further explains our data filtering and statistical techniques.

Detection of positive selection

We downloaded the sequence and chosen annotations of the human genome (hg17 of May, 2004) from the UCSC Genome Bioinformatics web site (Karolchik et al., 2003, http://genome.ucsc.edu) and the Genomic tRNA Database web site (http://lowelab.ucsc.edu/GtRNAdb). We parsed each chromosome into clusters of overlapping transcripts and splices according to the UCSC Known Genes collection, retaining only clusters in which all known transcripts are from the same strand; these are termed “genes” throughout this article. We parsed each gene into regions, intersecting over alternative transcripts and splices, so that what are termed “promoter sites” and “intronic sites” are such sites with respect to all known transcripts and splices. We first excluded 100 bp at each end of each coding-region intron except the first and then included at most 2500 bp at each end of the remainders.

We mapped each gene to the best-matching regions of the chimpanzee and macaque genomes (panTro2 of March, 2006 and rheMac2 of January, 2006) using whole-genome pairwise alignments from UCSC. We discarded any gene whose mapping to either genome violated the dominant syntenies among the three genomes, any gene whose mapping to either genome failed to flank that of either flanking gene, and any genes whose mappings to either genome overlapped, apart from flanking regions. We computed three-species alignments using TBA (Blanchette et al., 2004, http://

www.bx.psu.edu/miller_lab). We masked out bases in chimpanzee and macaque sequences having quality scores less than 40, known non-coding RNA genes in human sequences, and windows of 50 ungapped and unmasked sites containing more than 12 or 17 differences between human and chimpanzee or macaque, respectively. We discarded any promoter region whose alignment contained more than 0.75% such divergence-masked bases or 9% gaps or whose associated intronic alignment contained fewer than 2500 ungapped and unmasked sites. (The data supplement includes results for promoter regions that failed these cutoffs but were otherwise analyzable.)

For each promoter region, we constructed 100 bootstrap replicates over the associated intronic alignment. For each bootstrap replicate, we fitted the null and alternate models to the promoter region and bootstrap replicate using HyPhy (Pond and Muse, 2005, http://www.hyphy.org). For each model, we took the best of 10 fits starting from random points, to guard against local maxima of the likelihood function. We implemented the likelihood ratio test as a χ^2 test with one degree of freedom. We took the median p-value over the bootstrap replicates as the representative p-value for the promoter region. We transformed p-values into q-values using the R package qvalue (Storey and Tibshirani, 2003, http://faculty.washington.edu/∼jstorey/qvalue), under the conservative assumption that the prior estimate of the number of true positives is 0.

Assessment of gene functions

We downloaded PANTHER classifications (HMM Library Version 6.0, http://www.pantherdb.org), obtained Novartis data (GeneAtlas Version 2, http://symatlas.gnf.org/suppl.html#reqdata_geneatlas), and downloaded Khaitovich et al.'s results (http://www.sciencemag.org/cgi/content/full/1108296/DC1). We matched our genes with theirs using HGNC, RefSeq, and UniProt identifiers. For PANTHER categories, we computed p_{MW} using the R function wilcox.test. For Novartis tissues, we took means over multiple arrays per tissue and maxima over multiple probes per gene. The expression levels of a gene in the 73 non-cancerous tissues may be regarded as a vector in 73-dimensional Euclidean space. We defined the specificity score of the gene for a tissue as the square of the cosine of the angle between the vector and the axis corresponding to the tissue. This measure depends on the distribution of expression over tissues but not the magnitude of expression overall. A gene highly specific to one tissue has specificity scores near 1 for this tissue and near 0 for others, even if it is not highly expressed in this tissue. In contrast, a gene maximally expressed in one tissue yet nearly as highly expressed in many others has specificity scores near 0 for all tissues. Among measures having these properties, the one we chose also has the property that for a given gene, the sum over tissues of the specificity scores is 1, which facilitates comparisons among genes. We evaluated the rank correlation between specificity score and p-value for positive selection using the R function cor.test. For Khaitovich et al.'s results, we evaluated the rank correlation between our p-value for positive selection and their ratio of mean-squared expression difference between species to mean-squared expression variability within species using the R function cor.test.

Software

Our software is written in Ruby (∼5600 lines), Python (∼850 lines), C (∼300 lines), and HyPhy Batch Language (∼250 lines) and runs under Linux and Mac OS X. It is available upon request.

Potential concerns

This section further explains our data filtering and statistical techniques and presents several auxiliary analyses and other considerations strengthening confidence that many high-scoring genes are genuine cases of positive selection on promoter regions.

Data filtering

Our methods could be misled by several kinds of problem with the data, so we applied several kinds of filtering to the data. We masked out bases in chimpanzee and macaque sequences having quality scores less than 40. (Quality scores were not available for the human genome, but we believe that the coverage of this genome is sufficiently high that this lack is of little concern.) When mapping genes to best-matching regions of the chimpanzee and macaque genomes, we discarded any gene whose mapping to either genome violated the dominant syntenies among the three genomes, any gene whose mapping to either genome failed to flank that of either flanking gene, and any genes whose mappings to either genome overlapped, apart from flanking regions. (As we use the term “gene”, a single gene may include multiple transcripts and splices, and distinct genes do not overlap, apart from flanking regions.) For computing three-species alignments, we used software, TBA, that performs well on simulated neutrally evolving mammalian sequences (Blanchette et al., 2004). We discarded any promoter region whose alignment contained more than 9% gaps (where a gap corresponds to one base), we masked out windows of 50 ungapped and unmasked sites containing more than 12 or 17 differences between human and chimpanzee or macaque, respectively, and we discarded any promoter region whose alignment contained more than 0.75% such divergence-masked bases. (We did not apply the gap and divergence-masked base frequency cutoffs to intronic sequences, because poorly assembled or aligned intronic sequences are unlikely to cause false positives.) Over all analyzed genes, the rank correlations between p-value and promoter gap frequency and between p-value and promoter divergence-masked base frequency do not differ significantly from 0 ($r_S = -0.0084$ and -0.013, two-tailed $p = 0.51$ and 0.33, respectively); this also holds over the high-scoring genes alone ($r_S = 0.023$ and -0.0044, two-tailed $p = 0.59$ and 0.92, respectively). Thus, it seems unlikely that our results are dominated by errors in base calling, genome assembly, ortholog identification, or sequence alignment.

Statistical concerns

Two statistical concerns arise from the fact that we considered only three species, the minimum to which our methods apply. (We could have added mouse and/or rat, but experiments persuaded us that the difficulty of aligning noncoding sequences across primates and rodents would negate the benefit of including more species.) One concern is that although the asymptotic distribution of twice the difference of log-likelihoods between our models in the absence of positive selection is a 50:50 mixture of a point mass at 0 and the χ^2 distribution with one degree of freedom (Pond and Muse, 2005; Zhang et al., 2005), this may not be a satisfactory approximation for small samples. Zhang et al. (2005) showed that even for large samples, the mixture distribution can be liberal compared to the true distribution, whereas the χ^2 distribution without the point mass is conservative. We therefore implemented the likelihood ratio test as a χ^2 test with one degree of freedom.

For each of the 100 genes scoring highest in humans, we used HyPhy to fit the HKY85 model (Hasegawa et al., 1985) to the intronic alignment. Using the observed base frequencies and fitted transition-to-transversion ratio and branch lengths, we simulated evolution under the HKY85 model 100 times. Treating each simulated alignment as a promoter alignment, we paired it with the intronic alignment and analyzed the pair for positive selection. In this way, we obtained 10000 p-values, none of which were smaller than 0.05. Thus, our implementation of the likelihood ratio test is conservative when the promoter region and associated intronic sequences evolve neutrally at the same rates.

The other statistical concern is that fits of evolutionary models to small samples are sensitive to the stochasticity of evolutionary processes, even if the processes are neutral. To mitigate this, we adopted a technique widely used in phylogenetic inference, namely, bootstrapping (Felsenstein, 2004). For each promoter region, we constructed 100 bootstrap replicates over the associated intronic alignment (i.e., we sampled, with replacement, intronic sites until we accumulated a sampled alignment of the same length as the original alignment, and we continued until we accumulated 100 sampled alignments, the bootstrap replicates). (We did not bootstrap over promoter alignments. Doing so would be incoherent with our models, which posit that intronic sequences evolve homogeneously but promoter regions evolve heterogeneously. We aim to detect signals from subsequences of promoter sequences, so it would make little sense to analyze bootstrap replicates, many of which might not contain these subsequences.)

Analyzing each promoter region with the associated bootstrap replicates yielded a distribution of p-values, of which the median is a reasonable choice of representative. The 2.5th and 97.5th percentiles of the distribution (estimated by midpoint interpolation) form a 95% confidence interval. The widths of these intervals are not negligible, but they are not inordinate. For example, among the 100 genes scoring highest in humans, the maximum and median 97.5th percentile are only 0.12 and 0.0075, respectively. Thus, our results are probably not dominated by small-sample fluctuations.

A third statistical concern arises from the fact that the lengths of the promoter region and associated intronic sequences vary among genes. Excluding gaps and masked bases, over all analyzed genes, promoter region length varies from 34 bp to 4985 bp with a median of 4294 bp, and intronic sequences length varies from 2502 bp to 53811 bp with a median of 10902 bp; over the high-scoring genes alone, promoter region length varies from 80 bp to 4965 bp with a median of 4376 bp, and intronic sequences length varies from 2514 bp to 50591 bp with a median of 11925 bp. Longer promoter regions might be expected to yield smaller p-values, simply because the more promoter sites are analyzed, the more sites under positive selection may be included. Longer promoter regions or intronic sequences might also be expected to yield more precise estimates of parameters and more power to distinguish the two sequence compartments, although the net effect on p-values is difficult to predict.

These effects appear to be small. Over all analyzed genes, the rank correlation between p-value and promoter region length is -0.072, which differs significantly from 0 (two-tailed $p < 10^{-6}$, $n = 6280$); over the high-scoring genes alone, the rank correlation is -0.077, which does not differ significantly from 0 (two-tailed $p = 0.066$, $n = 575$). As might be expected, these correlations are negative. However, they are small, perhaps because promoter region length varies at the distal ($5'$) end, but functional elements and hence positive selection are concentrated near the proximal ($3'$) end. Over all analyzed genes, the rank correlation between p-value and intronic sequences length is 0.044, which differs significantly from 0 (two-tailed $p = 0.00045$, $n = 6280$); over the high-scoring genes alone, the rank correlation is 0.041, which does not differ significantly from 0 (two-tailed $p = 0.33$, $n = 575$). The sign of these correlations suggests that less positive selection tends to be signaled when more intronic sites are analyzed. However, their magnitudes suggest that this tendency is rather weak. All these correlations weaken upon controlling for the small negative correlation between promoter region length and intronic sequences length. In particular, over all analyzed genes, the partial rank correlation between p-value and intronic sequences length controlling for promoter region length is 0.040; over the high-scoring genes alone, the partial rank correlation is 0.031.

Interpretational issues

Even if a promoter region has evolved appreciably faster than the associated intronic sequences, it might be due to a difference in mutation rates, not selection regime. Although this possibility cannot be wholly excluded, several considerations suggest that it does not predominate. First, surveys of divergence at putatively neutral sites among human, mouse, and rat have found little variation in mutation rates over distances as short as 100 kb (Chuang and Li, 2004; Gaffney and Keightley, 2005), the longest distance spanned by any of our analyses. Second, most of our analyses involve intronic sequences from multiple genes (the median is two), reducing the effect of idiosyncrasies in the evolution of individual genes. Third, differences in mutation rates might engender differences in base frequencies, which might engender correlations between p-value and base frequencies. However, over all analyzed genes, the rank correlations between p-value and GC frequency in promoter region, GC frequency in associated intronic sequences, and difference of GC frequencies between promoter region and associated intronic sequences do not differ significantly from 0 ($r_{\mathrm{S}} = 0.018$, 0.021, and 0.0031, two-tailed $p = 0.17$, 0.090, and 0.81, respectively); this also holds over the high-scoring genes alone ($r_{\mathrm{S}} = 0.0061$, 0.041, and -0.0034, two-tailed $p = 0.89$, 0.33, and 0.94, respectively). Fourth, in our models, sites evolve independently, but in mammalian genomes, hypermutable CpG dinucleotides evolve markedly faster than other dinucleotides (Hellmann et al., 2003; Keightley and Gaffney, 2003; Chimpanzee Sequencing and Analysis Consortium, 2005; Keightley et al., 2005). Genes with higher CpG frequency in the promoter region than in the associated intronic sequences might score high as a consequence. (Many CpG's in promoter regions are not methylated and hence are not hypermutable, but we conservatively suppose that all CpG's in promoter regions are potentially problematic.) Following Keightley and Gaffney (2003), Gaffney and Keightley (2005), and Keightley et al. (2005), we considered not merely CpG's but all CG-susceptible sites (CGSS's, following C or preceding G in any of the three species), which have elevated probabilities of having been parts of CpG's in the course of evolution. We discarded randomly chosen non-CGSS's or CGSS's from each intronic alignment so as to raise or lower, respectively, the CGSS frequency in the intronic alignment until it approximately equaled that in the associated promoter alignment. We repeated our analyses with this approximate equalization, and the rank correlation between p-values with and without it is 0.98.

A related issue is the possibility of biased gene conversion in promoter regions (Holmquist, 1992). Pollard et al. (2006a) found that for their Human Acclerated Regions, rate of evolution was strongly correlated with frequency of weak-to-strong substitutions, from A or T to C or G. In contrast, for our promoter regions, the rank correlation between p-value and weak-to-strong substitution frequency over all genes differs significantly from 0 ($r_{\mathrm{S}} = 0.045$, two-tailed $p = 0.00040$), but its magnitude is small, and its sign is opposite that found by Pollard et al.; the correlation over the high-scoring genes alone does not differ significantly from 0 ($r_{\mathrm{S}} = -0.019$, two-tailed $p = 0.66$).

Even if there is a difference in selection regime, it might consist of negative selection on the associated intronic sequences, not positive selection on the promoter region. Again, this possibility cannot be wholly excluded, but several considerations suggest that it does not predominate. First, the chosen intronic sequences are generally among the least constrained in the genome (Hellmann et al., 2003; Keightley and Gaffney, 2003; Chimpanzee Sequencing and Analysis Consortium, 2005; Keightley et al., 2005). Second, analyses involving intronic sequences from multiple genes are less affected by idiosyncrasies in the evolution of individual genes. Third, assuming functional elements in the chosen intronic sequences are fairly sparse, some bootstrap replicates should be largely free of them, and the upper end of the bootstrap p-value distribution should be little distorted by them. As noted above, among the 100 genes scoring highest in humans, the 97.5th percentiles of the bootstrap p-value distribution are not high. Fourth, we analyzed 3763 regions immediately downstream ($3'$) from transcription stop sites, where the next gene downstream is transcribed in the opposite direction. Such doubly downstream regions contain fewer *cis*-regulatory sequences than promoter regions, although they do contain some (Wray et al., 2003; Blanchette et al., 2006; Crawford et al., 2006). We applied the same methods and cutoffs to these regions as to promoter regions, and as expected, the signal from these regions is weaker than from promoter regions. The p-value distribution for doubly downstream regions is significantly higher than for promoter regions (one-tailed Mann–Whitney $p = 0.00055$), and, for example, the percentage of doubly downstream regions with $q < 0.05$ is only half that of promoter regions (0.37% vs. 0.73%).

Even if the promoter region has experienced positive selection, the selection need not have been on *cis*-regulatory sequences. Another possibility is a noncoding RNA gene; such a case was recently discovered (Pollard et al., 2006b). We masked out known noncoding RNA genes, but there may be many unknown ones. This possibility strengthens the motivation for functional analyses of the promoter regions we have flagged.

Rates, frequencies, and sites

Two reviewers requested information about rates of evolution and frequencies of sites under positive selection. Figures S1 and S2 present such information, within the limits of our ability to estimate these quantities. These limits are serious, in that sites under positive selection are uncommon, and we are estimating their rates of evolution from only three species. Even when the alternate model fits much better than the null model, rates and frequencies are generally estimated with low precision and some bias. Low precision is evident from wide variation over bootstrap replicates in estimates of f_3 and ζ_3 (cf. Figure 1 and Table S1 for definitions). The bias is revealed by applying our methods to simulated data featuring positive selection (i.e., parametric bootstrapping), which indicates that f_3 is typically underestimated, ζ_3 is typically overestimated, and their product is typically estimated more accurately than either factor. Accordingly, we recommend caution in interpreting Figure S1 and great caution in interpreting Figure S2.

Figure S1 is analogous to Figure 2, but with $\overline{\zeta} = f_1\zeta_1 + f_2\zeta_2 + f_3\zeta_3$ instead of p-value. Like p-value in Figure 2, $\overline{\zeta}$ in Figure S1 is the median over bootstrap replicates. It should be noted that although $\overline{\zeta}$ and p-value are correlated, their relationship is not determinate or monotonic. The vertical and horizontal dashed blue lines in Figure S1 correspond to neutral evolution on the average in humans and chimpanzees, respectively. The clumping of points around these lines suggests that most promoter region evolution is neutral, and positive selection is more likely in one species or the other than in both. The “spur” at the lower left represents promoter regions that are the same in human and chimpanzee but different in macaque.

Figures S2a and S2b plot ζ_3 vs. f_3 in humans and chimpanzees, respectively, for the 575 and 636 genes scoring high in each species. For each gene, the pair of bootstrap replicates out of 100 giving the median estimated $f_3\zeta_3$ by midpoint interpolation were identified, and the midpoint interpolations of their estimated f_3 and ζ_3 are plotted. At face value, points at the upper left represent promoter regions where a few sites are under strong positive selection, whereas points at the lower right represent regions where more sites are under weaker selection. Both situations, and everything in between, seem common, with some bias toward the former. However, the distributions of points in these plots may predominantly reflect the vagaries of parameter estimation.

Ultimately, we would like to identify the positively selected sites within positively selected regions. This goal may be partially attainable using so-called Bayes empirical Bayes methods (Wong et al., 2004; Zhang et al., 2005) and catalogs of known or predicted *cis*-regulatory sequences or modules (Blanchette et al., 2006; Crawford et al., 2006). However, identifying sites is more difficult than discerning their presence, because the collective signal from multiple sites may be detectable even when individual signals from single sites are not (Zhang et al., 2005). This difficulty is acute with only three species, so additional primate genome sequences will be welcome.

References

Blanchette, M., Bataille, A. R., Chen, X., Poitras, C., Laganière, J., Lefèbvre, C., Deblois, G., Giguère, V., Ferretti, V., Bergeron, D., Coulombe, B., and Robert, F., 2006. Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression. *Genome Research* **16**:656–668.

Blanchette, M., Kent, W. J., Riemer, C., Elnitski, L., Smit, A. F. A., Roskin, K. M., Baertsch, R., Rosenbloom, K., Clawson, H., Green, E. D., Haussler, D., and Miller, W., 2004. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Research* **14**:708–715.

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

Chuang, J. H., and Li, H., 2004. Functional bias and spatial organization of genes in mutational hot and cold regions in the human genome. *PLoS Biology* **2**:253–263.

Crawford, G. E., Holt, I. E., Whittle, J., Webb, B. D., Tai, D., Davis, S., Margulies, E. H., Chen, Y., Bernat, J. A., Ginsburg, D., Zhou, D., Luo, S., Vasicek, T. J., Daly, M. J., Wolfsberg, T. G., and Collins, F. S., 2006. Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). *Genome Research* **16**:123–131.

Felsenstein, J., 2004. *Inferring phylogenies*. Sinauer Associates, Sunderland, MA.

Gaffney, D. J., and Keightley, P. D., 2005. The scale of mutational variation in the murid genome. *Genome Research* **15**:1086–1094.

Hasegawa, M., Kishino, H., and Yano, T., 1985. Dating of the human–ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution* **22**:160–174.

Hellmann, I., Zöllner, S., Enard, W., Ebersberger, I., Nickel, B., and Pääbo, S., 2003. Selection on human genes as revealed by comparisons to chimpanzee cDNA. *Genome Research* **13**:831–837.

Holmquist, G. P., 1992. Chromosome bands, their chromatin flavors, and their functional features. *American Journal of Human Genetics* **51**:17–27.

Karolchik, D., Baertsch, R., Diekhans, M., Furey, T. S., Hinrichs, A., Lu, Y. T., Roskin, K. M., Schwartz, M., Sugnet, C. W., Thomas, D. J., Weber, R. J., Haussler, D., and Kent, W. J., 2003. The UCSC genome browser database. *Nucleic Acids Research* **31**:51–54.

Keightley, P. D., and Gaffney, D. J., 2003. Functional constraints and frequency of deleterious mutations in non-coding DNA of rodents. *Proceedings of the National Academy of Sciences of the United States of America* **100**:13402–13406.

Keightley, P. D., Lercher, M. J., and Eyre-Walker, A., 2005. Evidence for widespread degradation of gene control regions in hominid genomes. *PLoS Biology* **3**:0282–0288.

Pollard, K. S., Salama, S. R., King, B., Kern, A. D., Dreszer, T., Katzman, S., Siepel, A., Pedersen, J. S., Bejerano, G., Baertsch, R., Rosenbloom, K. R., Kent, J., and Haussler, D., 2006a. Forces shaping the fastest evolving regions in the human genome. *PLoS Genetics* **2**:1599–1611.

Pollard, K. S., Salama, S. R., Lambert, N., Lambot, M.-A., Coppens, S., Pedersen, J. S., Katzman, S., King, B., Onodera, C., Siepel, A., Kern, A. D., Dehay, C., Igel, H., Ares, Jr., M., Vanderhaeghen, P., and Haussler, D., 2006b. An RNA gene expressed during cortical development evolved rapidly in humans. *Nature* **443**:167–172.

Pond, S. L. K., and Muse, S. V., 2005. HyPhy: Hypothesis testing using phylogenies. Nielsen, R. (Ed.), *Statistical methods in molecular evolution*, pp. 125–181. Springer, New York, NY.

Storey, J. D., and Tibshirani, R., 2003. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America* **100**:9440–9445.

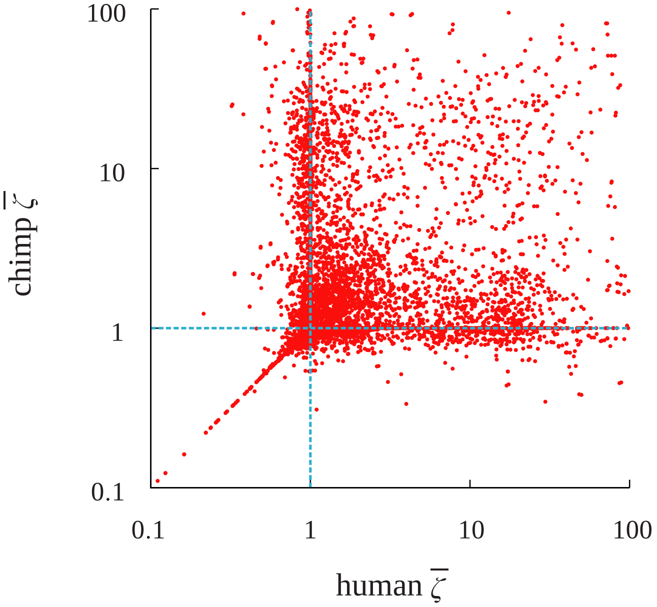
Wong, W. S. W., and Nielsen, R., 2004. Detecting selection in noncoding regions of nucleotide sequences. *Genetics* **167**:949–958.

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

Wray, G. A., Hahn, M. W., Abouheif, E., Balhoff, J. P., Pizer, M., Rockman, M. V., and Romano, L. A., 2003. The evolution of transcriptional regulation in eukaryotes. *Molecular Biology and Evolution* **20**:1377–1419.

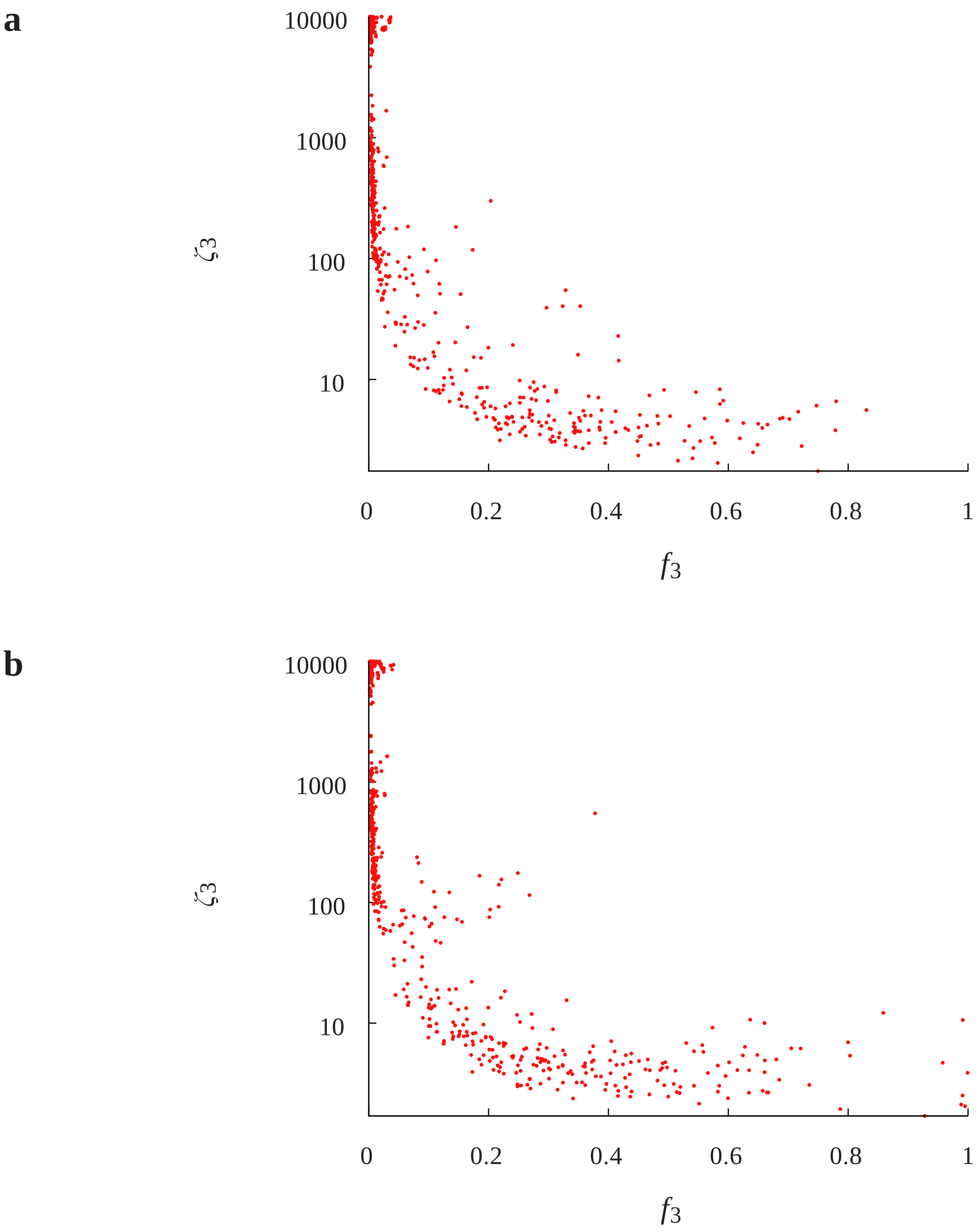
Zhang, J., Nielsen, R., and Yang, Z., 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Molecular Biology and Evolution* **22**:2472–2479.

Figure S1



Rate of evolution in chimpanzees vs. humans. Each point represents one gene, and the horizontal (vertical) axis represents $\overline{\zeta} = f_1\zeta_1 + f_2\zeta_2 + f_3\zeta_3$ on the human (chimpanzee) lineage. The dashed blue lines correspond to $\overline{\zeta} = 1$ on one lineage or the other. $f_3\zeta_3$ may be estimated poorly, so this figure should be interpreted cautiously; see "Rates, frequencies, and sites" for discussion. (Several genes have $\overline{\zeta} < 0.1$ or $\overline{\zeta} > 100$ on one lineage or the other and hence are not plotted.)

Figure S2



Rate of evolution vs. frequency of sites under positive selection in **(a)** humans or **(b)** chimpanzees. Each point represents one high-scoring gene, and the horizontal (vertical) axis represents f_3 (ζ_3) on the designated lineage. f_3 and ζ_3 may be estimated very poorly, so this figure should be interpreted very cautiously; see "Rates, frequencies, and sites" for discussion.