

Promoter regions of many neural- and nutrition-related genes have experienced positive selection during human evolution

Ralph Haygood^{1,3}, Olivier Fedrigo¹⁻³, Brian Hanson¹, Ken-Daigoro Yokoyama¹ & Gregory A Wray^{1,2}

Surveys of protein-coding sequences for evidence of positive selection in humans or chimpanzees have flagged only a few genes known to function in neural or nutritional processes¹⁻⁵, despite pronounced differences between humans and chimpanzees in behavior, cognition and diet⁶⁻⁸. It may be that most such differences are due to changes in gene regulation rather than protein structure⁹. Here, we present the first survey of promoter (5'-flanking) regions, which are rich in *cis*-regulatory sequences, for evidence of positive selection in humans. Our results indicate that positive selection has targeted the regulation of many genes known to be involved in neural development and function, both in the brain and elsewhere in the nervous system, and in nutrition, particularly in glucose metabolism.

Cognitive, behavioral and dietary differences are among the most conspicuous differences between humans and their closest relatives, the great apes. For example, even in the absence of written language or agriculture, human communications, tools and diets are far more complex and diverse than those of chimpanzees⁶⁻⁸. Such traits are essential to many aspects of human ecology, and it is plausible that many are adaptations. Consistent with this, the protein-coding sequences of several genes known to function in neural or nutritional processes have been shown to bear signatures of positive selection (natural or sexual selection for novel variants) in humans^{10,11}. However, such genes are not prominent in surveys of coding sequences for evidence of positive selection in humans or chimpanzees¹⁻⁵. Instead, these surveys have flagged many genes known to function in immunity, olfaction and spermatogenesis, among other processes.

One possible explanation is that many phenotypic differences between humans and chimpanzees may be due to changes in gene regulation rather than protein structure⁹. In particular, the genetic bases of human neural and nutritional adaptations may reside primarily in *cis*-regulatory sequences (DNA where proteins bind sequence-specifically to regulate transcription), few of which lie within coding sequences¹². Several recent studies point in this direction. First, of the two most thoroughly investigated cases of positive selection on

cis-regulatory sequences in humans, one, *PDYN*, is neural related¹³, and the other, *LCT*, is nutrition related¹⁴. Second, two surveys of linkage disequilibrium among SNPs for signatures of very recent positive selection within human populations, embracing both coding and noncoding sequences, found excesses of signatures in the vicinity of genes in several nutrition- and neural-related categories^{15,16}. Third, two surveys of regions that are highly conserved across vertebrates, except for extensive changes in humans that might be driven by positive selection, found excesses of conserved regions in the vicinity of genes in several neural-related categories^{17,18}. These studies (which were restricted to individual genes, very recent positive selection or highly conserved regions) strengthen the motivation for a systematic assessment of whether *cis*-regulatory sequences of many neural- or nutrition-related genes bear signatures of positive selection in humans. Because *cis*-regulatory sequences are scattered, short and degenerate, most have not been mapped precisely, but several lines of evidence indicate that most are near transcription start sites^{12,19,20}. Accordingly, we surveyed regions immediately upstream (5') of transcription start sites for evidence of positive selection in humans.

Our approach is to compare rates of evolution along the human lineage between a promoter region and nearby intronic sequences (Fig. 1a). We use the term 'promoter region' for the region immediately upstream from a transcription start site, extending (at most) 5 kb or to the next gene upstream. These regions contain many (perhaps most) *cis*-regulatory sequences in the genome^{12,19,20}. The chosen intronic sequences of a gene are the coding-region introns, excluding the first intron (which often contains *cis*-regulatory sequences¹⁹⁻²¹), the ends of each intron (which contain splicing signals²²) and the centers of large introns (which may often contain *cis*-regulatory sequences¹⁹). These sequences are generally among the least constrained in the genome^{3,23,24}, so they constitute a plausible neutral standard accounting for regional variation in mutation and recombination rates. We associated each promoter region with all chosen intronic sequences in a 100-kb window centered on the promoter region. If a promoter region has evolved appreciably faster than the associated intronic sequences, it is likely that *cis*-regulatory sequences within the promoter region have experienced positive selection.

¹Biology Department, Duke University, Durham, North Carolina 27708, USA. ²Institute for Genome Sciences and Policy, Duke University, Durham, North Carolina 27708, USA. ³These authors contributed equally to this work. Correspondence should be addressed to R.H. (rhaygood@duke.edu).

Received 7 February; accepted 20 June; published online 12 August 2007; doi:10.1038/ng2104

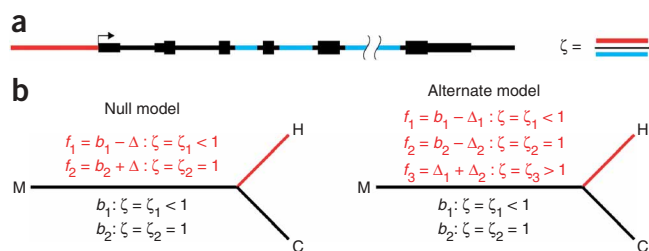


Figure 1 Genes and models. (a) A typical gene. The arrow is the transcription start site, boxes of middling height are UTR exons, and boxes of greater height are coding-region exons. Red indicates the promoter region, and blue indicates the chosen intronic sequences for our analyses. The fitted parameter ζ is the ratio of substitution rates in the promoter region to those in the associated intronic sequences. (b) Our models (see **Supplementary Table 1** for a fuller presentation). H, C and M label the human, chimpanzee and macaque lineages, respectively. Red and black indicate the foreground and background lineages, respectively. On the background lineages, an estimated proportion $b_1 \geq 0$ of promoter sites have an estimated $\zeta = \zeta_1 < 1$, and the remaining proportion $b_2 = 1 - b_1$ have $\zeta = \zeta_2 = 1$ in both models. On the foreground lineage, an estimated proportion $\Delta \geq 0$ of promoter sites change from $\zeta = \zeta_1 < 1$ to $\zeta = \zeta_2 = 1$ in the null model, and estimated proportions $\Delta_1 \geq 0$ and $\Delta_2 \geq 0$ change from $\zeta = \zeta_1 < 1$ and $\zeta = \zeta_2 = 1$ to an estimated $\zeta = \zeta_3 > 1$ in the alternate model.

(The **Supplementary Discussion** online presents evidence that other possible explanations are unlikely to account for most of our results.) For 16,905 genes, we attempted to extract and align promoter regions and intronic sequences from the published human (*Homo sapiens*), common chimpanzee (*Pan troglodytes*) and rhesus macaque (*Macaca mulatta*) genome sequences, macaque being a suitable outgroup for apportioning substitutions between the human and chimpanzee lineages. Missing or questionable data precluded the analysis of many promoter regions, but we were able to analyze the promoter regions of 6,280 genes.

To compare rates, we fitted by maximum likelihood two models of single-nucleotide substitutions to each promoter alignment and the associated intronic alignment (**Fig. 1b** and **Supplementary Table 1** online). The fitted parameters include ζ (zeta), the ratio of substitution rates in the promoter region to those in the associated intronic sequences²⁵; ζ is analogous to the ratio of substitution rates at nonsynonymous sites to those at synonymous sites in coding sequences. The null model constrains ζ to be ≤ 1 , representing negative or no selection on the promoter region, whereas the alternate model allows ζ to be > 1 on the human lineage, representing positive selection on the promoter region. A likelihood ratio test gives a P value for consistency of the data with the null model²⁶. A small P value constitutes a high score for positive selection. We use the term ‘high-scoring genes’ for genes with $P < 0.05$. The models posit different values of ζ for different classes of promoter site, the values of ζ and frequencies of the classes being fitted parameters. A high score requires that some but not all or even most promoter sites have evolved appreciably faster than the average intronic site. The null model accommodates promoter sites that have evolved under negative selection on the chimpanzee and macaque lineages but neutrally on the human lineage²⁶. Thus, the contrast between the models is sensitive to positive selection rather than mere relaxation of negative selection. We transformed P values into Q values, a false discovery rate-based measure of significance²⁷. We repeated our analyses, allowing ζ to be > 1 on the chimpanzee instead of the human lineage. (**Supplementary Tables 2–5** online present these basic results.)

Of the 6,280 analyzed genes, 46 (0.73%) have $Q < 0.05$, so the 5% false discovery rate set is nonempty²⁷. Five hundred seventy-five (9.2%) have $P < 0.05$, corresponding to $Q = 0.55$, which suggests that the promoter regions of at least 250 ($\approx (1 - 0.55) \times 575$) analyzed genes have experienced positive selection. Given that the analyzed genes amount to roughly one-third of all human genes, naive extrapolation suggests that the promoter regions of at least 750 human genes have experienced positive selection. Positive selection seems to be as prevalent on the chimpanzee as on the human lineage (**Fig. 2**); the P value distributions are not significantly separated (two-tailed Mann-Whitney $P = 0.63$). Positive selection seems to be weakly correlated between the two lineages; the rank (Spearman) correlation between P values is 0.27.

We began exploring the biological implications of our results using the PANTHER biological process categories (see Methods). Of the 6,280 analyzed genes, 3,850 are in at least one PANTHER category. For each category containing at least 20 analyzed genes, we evaluated whether analyzed genes within the category tend to score higher than analyzed genes outside the category. **Table 1** lists the most significant results (see also **Supplementary Table 6** online). These results are instructive but limited, in that many genes lack PANTHER categories, many others have categories that do not encompass all available information about their functions and some PANTHER categories do not immediately correspond to organismal traits. Therefore, we surveyed the biomedical literature for information about the 100 genes scoring highest in humans and the other high-scoring genes in the categories listed in **Table 1a**. (Unless otherwise noted, information about gene functions in what follows is available from OMIM.)

A conspicuous proportion of genes scoring high for positive selection in our analyses, especially in humans, are known to be involved in neural development and function. Relevant PANTHER categories include neurogenesis, ectoderm development, nerve–nerve synaptic transmission, neuronal activities, other neuronal activity and anion transport. Genes scoring high in humans are involved in axon guidance, synapse formation and neurotransmission in the brain (such as *PRSS12*, *NTRK2*, *UCHL3*, *ME2*, *STX1A* and *SCN1A*) and similar functions elsewhere in the nervous system (such as *ISL2*, *SLIT2*, *CHRNA9*, *ADAM22*, *SCN9A* and *GLRA1*). Several of these

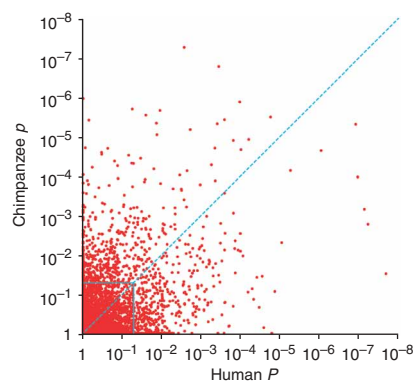


Figure 2 Positive selection in chimpanzees versus humans. Each point represents one gene; the x -axis represents P values on the human lineage, and the y -axis represents P values on the chimpanzee lineage. The solid blue lines correspond to P values of 0.05, and the dashed blue line corresponds to equal P values on the two lineages. Thus, genes scoring high in chimpanzees are plotted toward the upper left and genes scoring high in humans are plotted toward the lower right, genes scoring high in both species are plotted toward the center. (Several genes have $P < 10^{-8}$ on one lineage or the other and hence are not plotted.)

Table 1 PANTHER biological process categories enriched for positive selection^a

| (a) Humans | | | |
|--|--------------------------|----------------------------------|----------------------------------|
| Category ^b | Number of genes analyzed | Human P_{MW} ^c | Chimpanzee P_{MW} ^c |
| Protein folding | 70 | 0.0067 | 0.77 |
| Other neuronal activity ^d | 31 | 0.013 | 0.039 |
| Neurogenesis ^e | 133 | 0.013 | 0.032 |
| Glycolysis ^f | 21 | 0.014 | 0.72 |
| Neuronal activities ^d | 137 | 0.020 | 0.22 |
| Carbohydrate metabolism ^g | 210 | 0.020 | 0.017 |
| Ectoderm development ^h | 169 | 0.020 | 0.11 |
| Mesoderm development | 161 | 0.024 | 0.17 |
| Nerve–nerve synaptic transmission ^d | 25 | 0.025 | 0.34 |
| Vision | 64 | 0.025 | 0.15 |
| Oncogene | 23 | 0.045 | 0.46 |
| Anion transport | 31 | 0.049 | 0.17 |
| (b) Chimpanzees | | | |
| Category ^b | Number of genes analyzed | Chimpanzee P_{MW} ^c | Human P_{MW} ^c |
| DNA replication | 34 | 0.013 | 0.41 |
| Carbohydrate metabolism ^g | 210 | 0.017 | 0.020 |
| Transport | 414 | 0.029 | 0.50 |
| Neurogenesis | 133 | 0.032 | 0.013 |
| Other neuronal activity | 31 | 0.039 | 0.013 |
| Other polysaccharide metabolism ^g | 44 | 0.041 | 0.43 |
| Blood clotting | 32 | 0.049 | 0.47 |

^aSee **Supplementary Table 6** for further analyses. ^bOrdered by human (a) or chimpanzee (b) P_{MW} . Each listed category contains at least 20 analyzed genes. There are 127 such categories, with extensive overlap. ^cNominal one-tailed Mann–Whitney P value: the probability that analyzed genes within the category have P values for positive selection no lower than analyzed genes outside the category. ^dThe nerve–nerve synaptic transmission and other neuronal activity categories are contained in the neuronal activities category. For the remainder of the neuronal activities category, human $P_{MW} = 0.46$, and chimpanzee $P_{MW} = 0.62$.

^eThe neurogenesis category is contained in the ectoderm development category. For the remainder of the ectoderm development category, human $P_{MW} = 0.44$ and chimpanzee $P_{MW} = 0.81$. ^fThe glycolysis category is contained in the carbohydrate metabolism category. For the remainder of the carbohydrate metabolism category, human $P_{MW} = 0.080$ and chimpanzee $P_{MW} = 0.0078$.

^gThe other polysaccharide metabolism category is contained in the carbohydrate metabolism category. For the remainder of the carbohydrate metabolism category, chimpanzee $P_{MW} = 0.073$ and human $P_{MW} = 0.014$.

genes have variants known to be associated with diseases, including coding polymorphisms associated with mental retardation and epilepsy. Several genes relevant to neurodegenerative diseases score high in humans, including *SCRG1*, which is overexpressed in Creutzfeldt–Jakob disease; *TMED10*, the product of which inhibits production of amyloid beta peptides, whose accumulation is a critical feature of Alzheimer’s disease; and *ITM2C*, the product of which directly interacts with beta-secretase, which cleaves amyloid precursor protein. We were intrigued by the scores of *TMED10* and *ITM2C* in view of evidence that humans are more susceptible than chimpanzees to certain pathologies of Alzheimer’s disease²⁸. The PANTHER neurogenesis and other neuronal activity categories are enriched for positive selection in both humans and chimpanzees, but of the 31 genes scoring high in one species or the other, only five score high in both, suggesting that positive selection has targeted different neural traits in the two species.

Nutrition, including ingestion, digestion and metabolism, is also a prominent theme, especially in humans, where it seems that positive selection has targeted the regulation of glucose metabolism in particular. Relevant PANTHER categories include carbohydrate metabolism, glycolysis, other polysaccharide metabolism and anion transport. Glucose metabolism–related genes scoring high in humans include *HK1*, the product of which catalyzes the first step in glycolysis; *GCK*, which does the same and is a major regulator of glucose metabolism; *GPI*, which catalyzes the second step in glycolysis;

PFKFB3, which indirectly affects the activity of phosphofructokinase, which catalyzes the third step in glycolysis; *GCG*, which stimulates gluconeogenesis and glycogenolysis; *GALE*, which catalyzes the last step in galactose metabolism; *KLF11*, a glucose-inducible transcription factor whose targets include insulin; *ABCC8*, a potassium channel component modulating insulin release from pancreatic beta cells; and *FOXC2*, a transcription factor that is a major regulator of adipocyte metabolism. All these genes except *GCG* have variants known to be associated with diseases, including promoter polymorphisms associated with type 2 diabetes and hyperinsulinemic hypoglycemia. Other nutrition-related genes with high scores in humans include *LDHA*, which catalyzes the interconversion of lactate and pyruvate; *MMP20*, a catalyst of tooth enamel formation; *KRT4*, an upper digestive tract keratin; *HSD17B4*, a catalyst of fatty acid catabolism and bile acid formation; *MCEE*, a catalyst of fatty and amino acid catabolism; *USHBP1*, *HPD* and *SCLY*, catalysts of leucine, tyrosine and selenocysteine catabolism, respectively; and *LDLR*, which mediates the endocytosis of low-density lipoprotein particles. All these genes except *SCLY* have variants known to be associated with diseases. The PANTHER carbohydrate metabolism category is enriched for positive selection in both humans and chimpanzees, but of the 45 genes scoring high in one species or the other, only seven score high in both.

Using the Novartis Gene Expression Atlas (see Methods), we explored whether positive

selection on promoter regions is associated with gene expression in particular tissues or cell types. Most genes are expressed in multiple tissues, and, even if a gene is maximally expressed in one tissue, it may be nearly as highly expressed in others, so associating genes with their tissues of maximal expression is unsatisfactory. Accordingly, for each of 5,049 genes analyzed by both us and Novartis and for each of 73 noncancerous tissues assessed by Novartis, we computed a score between 0 and 1 representing the specificity of the gene to the tissue (see Methods); the specificity score of a gene for its tissue of maximal expression was low if the gene was nearly as highly expressed in other tissues. For each tissue, we evaluated whether the rank correlation between specificity score and P value for positive selection was negative, indicating an association of tissue specificity with positive selection. In humans, there was one significant correlation, for pancreas (one-tailed $P = 0.044$), which is consonant with positive selection on metabolic traits, but no gene mentioned above scored high for pancreas specificity. Genes scoring high for both pancreas specificity and positive selection in humans included *CPB1*, a carboxypeptidase; *SERPINI2*, a protease inhibitor whose disruption causes malnutrition in mice²⁹; and *ABCC2*, an anion transporter. In chimpanzees, there were several significant correlations, including testis seminiferous tubule (one-tailed $P = 0.024$), which is consonant with two surveys of coding sequences^{3,4}. It should be noted that tissues vary in the extent to which genes are specific to them—hence the potential for detecting an association with positive selection. Moreover, the

regulation of a gene may be under positive selection in a tissue to which the gene is not specific.

We compared our results to those of ref. 30, which constitutes the most extensive survey thus far of gene expression differences between humans and chimpanzees. For 3,317 genes analyzed both by us and in ref. 30 and for each of five tissues assessed in ref. 30, we computed the rank correlation between our P value for positive selection and their ratio of expression difference between species to expression variability within species. In humans, all these correlations were nominally negative, consistent with associations of expression divergence with positive selection, but none was statistically significant; the strongest was for kidney (one-tailed $P = 0.086$). However, we were not surprised by this weakness. Ref. 30 measures expression in recently deceased adults, whereas many promoter regions have presumably experienced positive selection with respect to expression during development or under particular physiological conditions. Moreover, many expression differences presumably arise from *trans*- rather than *cis*-regulatory changes.

Some high-scoring genes, including several mentioned above, are known to have multiple, distinct organismal roles. For example, in addition to catalyzing the second step in glycolysis, *GPI* serves as a lymphokine in the formation of antibody-secreting cells. Discerning which of these (or other yet unknown) roles has been targeted by positive selection is beyond the reach of our analyses. Conversely, the functions of other high-scoring genes are almost unknown. For example, for approximately half of the 100 genes scoring highest in humans, we found at most basic biochemical or expression information. Our results provide motivation for functional analyses of these genes.

In conjunction with previous surveys of coding sequences, the present survey of promoter regions suggests that human cognitive, behavioral and dietary adaptations have arisen primarily through changes in *cis*-regulatory sequences. Much further work is needed to confirm and elaborate this suggestion, partly because such adaptations are probably numerous and diverse. Complementary approaches to sequence analysis, incorporating human polymorphisms or focusing on gains and losses of genetic material, will yield further information about positive selection on promoter regions. Approaches such as ours will gain power by incorporating sequences from additional primates, which is already possible for individual genes and will be an important avenue of research in the near future. More important in the long run are functional analyses to map the *cis*-regulatory sequences of neural- and nutrition-related genes and probe the consequences of changes in them during human evolution. Similar analyses of segregating variants of these sequences and statistical tests for associations between segregating variants and organismal traits are also important. Our work provides attractive candidates for such research.

METHODS

Detection of positive selection. We downloaded the sequence and annotations of the human genome (hg17 from May 2004) from the University of California Santa Cruz (UCSC) Genome Bioinformatics website and the Genomic tRNA Database website (see URLs section below). 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 intron within a coding region, omitting the first intron, and then included at most 2,500 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 mapped location to either genome violated the dominant syntenies among the three genomes, any gene whose mapped location in either genome did not flank that of either flanking gene, and any genes whose mapped locations in either genome overlapped, apart from flanking regions. We computed three-species alignments using TBA (see URLs section below). We masked bases in chimpanzee and macaque sequences having quality scores less than 40, known noncoding 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 human and macaque, respectively. We discarded any promoter region whose alignment contained either more than 0.75% such divergence-masked bases or more than 9% gaps or whose associated intronic alignment contained fewer than 2,500 ungapped and unmasked sites. (Supplementary Tables 2–5 include results for promoter regions that failed these cutoffs but were otherwise analyzable.) See the **Supplementary Discussion** for further explanation of our data filtering.

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 (see URLs section below). For each model, we took the best of ten 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* (see URLs section below), under the conservative assumption that the prior estimate of the number of true positives is 0. See the **Supplementary Discussion** for further explanation of our statistical techniques.

Assessment of gene functions. We downloaded PANTHER classifications (HMM Library Version 6.0), obtained Novartis data (GeneAtlas Version 2), and downloaded the results of ref. 30 (see URLs section below). 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 noncancerous 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 the results of ref. 30, we evaluated the rank correlation between our P value for positive selection and their ratio of expression difference between species to expression variability within species using the R function *cor.test*.

Software. Our software is written in Ruby (~5,600 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.

URLS. PANTHER: <http://www.pantherdb.org>; OMIM: <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM>; Novartis Gene Expression Atlas: <http://symatlas.gnf.org>; UCSC Genome Bioinformatics: <http://genome.ucsc.edu>; Genomic tRNA Database: <http://lowelab.ucsc.edu/GtRNAdb>; TBA: http://www.bx.psu.edu/miller_lab; HYPHY: <http://www.hyphy.org>; *qvalue*: <http://faculty.washington.edu/~jstorey/qvalue>; Novartis data: http://symatlas.gnf.org/suppl.html#reqdata_geneatlas; results of ref. 30: <http://www.sciencemag.org/cgi/content/full/1108296/DC1>.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

We thank J. Pavisic and T. Severson for assistance with gene annotations; G. Barber, M. Diekhans, W. Kent, S. Kosakovsky Pond and W. Miller for advice

about their software; F. Hsu, K. Rosenbloom and A. Zweig for advice about UCSC resources and J. Horvath, J. Pritchard, M. Turelli, H. Willard and members of the G. Wray laboratory for comments on the manuscript. Most of the computations were performed on the Duke Shared Cluster Resource, which is maintained by the Duke Center for Computational Science, Engineering and Medicine. This research was supported by the Duke Institute for Genome Sciences and Policy and a US National Science Foundation Postdoctoral Fellowship in Biological Informatics to R.H. (grant number 0434655).

COMPETING INTERESTS STATEMENT

The authors declare no competing financial interests.

Published online at <http://www.nature.com/naturegenetics>

Reprints and permissions information is available online at <http://npg.nature.com/reprintsandpermissions>

- Clark, A.G. *et al.* Inferring nonneutral evolution from human–chimp–mouse orthologous gene trios. *Science* **302**, 1960–1963 (2003).
- Bustamante, C.D. *et al.* Natural selection on protein-coding genes in the human genome. *Nature* **437**, 1153–1157 (2005).
- Chimpanzee Sequencing and Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**, 69–87 (2005).
- Nielsen, R. *et al.* A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biol.* **3**, e170 (2005) (doi:10.1371/journal.pbio.0030170).
- Yu, X.-J., Zheng, H.-K., Wang, J., Wang, W. & Su, B. Detecting lineage-specific adaptive evolution of brain-expressed genes in human using rhesus macaque as outgroup. *Genomics* **88**, 745–751 (2006).
- Johnson-Frey, S.H. What's so special about human tool use? *Neuron* **39**, 201–204 (2003).
- Arcadi, A.C. Language evolution: What do chimpanzees have to say? *Curr. Biol.* **15**, R884–R886 (2005).
- Ungar, P.S. (ed.). *Evolution of the Human Diet: the Known, the Unknown, and the Unknowable* (Oxford Univ. Press, Oxford, 2007).
- King, M.-C. & Wilson, A.C. Evolution at two levels in humans and chimpanzees. *Science* **188**, 107–116 (1975).
- Vallender, E.J. & Lahn, B.T. Positive selection on the human genome. *Hum. Mol. Genet.* **13**, R245–R254 (2004).
- Sabeti, P.C. *et al.* Positive natural selection in the human lineage. *Science* **312**, 1614–1620 (2006).
- Wray, G.A. *et al.* The evolution of transcriptional regulation in eukaryotes. *Mol. Biol. Evol.* **20**, 1377–1419 (2003).
- Rockman, M.V. *et al.* Ancient and recent positive selection transformed opioid *cis*-regulation in humans. *PLoS Biol.* **3**, e387 (2005) (doi:10.1371/journal.pbio.0030387).
- Tishkoff, S.A. *et al.* Convergent adaptation of human lactase persistence in Africa and Europe. *Nat. Genet.* **39**, 31–40 (2007).
- Voight, B.F., Kudaravalli, S., Wen, X. & Pritchard, J.K. A map of recent positive selection in the human genome. *PLoS Biol.* **4**, e72 (2006) (doi:10.1371/journal.pbio.0040072).
- Wang, E.T., Kodama, G., Baldi, P. & Moyzis, R.K. Global landscape of recent inferred Darwinian selection for *Homo sapiens*. *Proc. Natl. Acad. Sci. USA* **103**, 135–140 (2006).
- Pollard, K.S. *et al.* Forces shaping the fastest evolving regions in the human genome. *PLoS Genet.* **2**, e168 (2006) (doi:10.1371/journal.pgen.0020168).
- Prabhakar, S., Noonan, J.P., Pääbo, S. & Rubin, E.M. Accelerated evolution of conserved noncoding sequences in humans. *Science* **314**, 786 (2006).
- Blanchette, M. *et al.* Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression. *Genome Res.* **16**, 656–668 (2006).
- Crawford, G.E. *et al.* Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). *Genome Res.* **16**, 123–131 (2006).
- Majewski, J. & Ott, J. Distribution and characterization of regulatory elements in the human genome. *Genome Res.* **12**, 1827–1836 (2002).
- Sorek, R. & Ast, G. Intronic sequences flanking alternatively spliced exons are conserved between human and mouse. *Genome Res.* **13**, 1631–1637 (2003).
- Hellmann, I. *et al.* Selection on human genes as revealed by comparisons to chimpanzee cDNA. *Genome Res.* **13**, 831–837 (2003).
- Keightley, P.D., Lercher, M.J. & Eyre-Walker, A. Evidence for widespread degradation of gene control regions in hominid genomes. *PLoS Biol.* **3**, e42 (2005) (doi:10.1371/journal.pbio.0030042).
- Wong, W.S.W. & Nielsen, R. Detecting selection in noncoding regions of nucleotide sequences. *Genetics* **167**, 949–958 (2004).
- Zhang, J., Nielsen, R. & Yang, Z. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol. Biol. Evol.* **22**, 2472–2479 (2005).
- Storey, J.D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. USA* **100**, 9440–9445 (2003).
- Olson, M.V. & Varki, A. Sequencing the chimpanzee genome: insights into human evolution and disease. *Nat. Rev. Genet.* **4**, 20–28 (2003).
- Loftus, S.K. *et al.* Acinar cell apoptosis in *Serpini2*-deficient mice models pancreatic insufficiency. *PLoS Genet.* **1**, e38 (2005) (doi:10.1371/journal.pgen.0010038).
- Khaitovich, P. *et al.* Toward a neutral evolutionary model of gene expression. *Science* **309**, 1850–1854 (2005).