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

Ralph Haygood[1,*,†], Olivier Fedrigo[1,2,†], Brian Hanson[1], Ken-Daigoro Yokoyama[1], and Gregory A. Wray[1,2]

[1]Biology Department, Duke University, Durham, NC 27708

[2]Institute for Genome Sciences and Policy, Duke University, Durham, NC 27708

[*]Corresponding author; Email: rhaygood@duke.edu

[†]These authors contributed equally to this work

Nature Genetics **39**:1140–1144 (2007)

[This version contains approximately 200 words and one figure that were cut from the version appearing in *Nature Genetics*.]

Abstract

Surveys of protein-coding sequences for evidence of positive selection in humans or chimpanzees have flagged surprisingly few genes known to function in neural or nutritional processes, despite pronounced differences between humans and chimpanzees in behavior, cognition, and diet. It may be that most such differences are due to changes in gene regulation rather than protein structure. 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 suggest 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 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 chimpanzees (Johnson-Frey, 2003; Arcadi, 2005; Ungar, 2007). 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 (Vallender and Lahn, 2004; Sabeti et al., 2006). However, such genes are not prominent in surveys of coding sequences for evidence of positive selection in humans or chimpanzees (Clark et al., 2003; Bustamante et al., 2005; Chimpanzee Sequencing and Analysis Consortium, 2005; Nielsen et al., 2005; Yu et al., 2006). Instead, these surveys have flagged many genes known to function in immunity, olfaction, and spermatogenesis, among other processes. Genes known to be neural-related show hardly any sign of positive selection in these surveys (Chimpanzee Sequencing and Analysis Consortium, 2005; Nielsen et al., 2005).

One possible explanation is that many phenotypic differences between humans and chimpanzees may be due to changes in gene regulation rather than protein structure (King and Wilson, 1975; Carroll, 2003). 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 sequences (Wray et al., 2003). 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-related (Rockman et al., 2005), and the other, *LCT*, is nutrition-related (Tishkoff et al., 2007). Second, two surveys of linkage disequilibrium among single-nucleotide polymorphisms 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 (Voight et al., 2006; Wang et al., 2006). Third, two surveys of regions that are highly conserved

across vertebrates except for extensive changes in humans, which might be driven by positive selection, found excesses of regions in the vicinity of genes in several neural-related categories (Pollard et al., 2006; Prabhakar et al., 2006). These studies, 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 (Wray et al., 2003; Blanchette et al., 2006; Crawford et al., 2006). Accordingly, we surveyed regions immediately upstream (5′) from 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 chosen, nearby intronic sequences (Figure 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. This includes some or all of both the so-called core and extended promoters (Cooper et al., 2006). These regions contain many, perhaps most *cis*-regulatory sequences in the genome (Wray et al., 2003; Blanchette et al., 2006; Crawford et al., 2006). The chosen intronic sequences of a gene are the coding-region introns, excluding the first intron, which often contains *cis*-regulatory sequences (Majewsky and Ott, 2002; Blanchette et al., 2006; Crawford et al., 2006), the ends of each intron, which contain splicing signals (Sorek and Ast, 2003), and the centers of large introns, which may often contain *cis*-regulatory sequences (Blanchette et al., 2006). These sequences are generally among the least constrained in the genome (Hellmann et al., 2003; Chimpanzee Sequencing and Analysis Consortium, 2005; Keightley et al., 2005), 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. (The text supplement presents evidence that other possible explanations are unlikely to account for most of our results.) For 16905 genes, we attempted to extract and align promoter regions and chosen 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 6280 genes.

To compare rates, we fitted by maximum likelihood two models of single-nucleotide substitutions to each promoter alignment and the associated intronic alignment (Figure 1b; cf. Table S1). Obviously, these models do not encompass the full complexity of molecular evolution. However, extensive applications of similar models to coding sequences have shown that such models can yield valuable insights (Bielawski and Yang, 2005). The fitted parameters include ζ (zeta), the ratio of substitution rates in the promoter region to those in the associated intronic sequences (Wong and Nielsen, 2004); 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 less than or equal 1, representing negative or no selection on the promoter region, whereas the alternate model allows ζ to be greater than 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 model (Bielawski and Yang, 2005). 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 lineage (Zhang et al., 2005). 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 significance (Storey and Tibshirani, 2003). We repeated our analyses allowing ζ to be greater than 1 on the chimpanzee instead of the human lineage. (The data supplement presents these basic results.)

Of the 6280 analyzed genes, 46 (0.73%) have $q < 0.05$, so the 5% false discovery rate set is nonempty (Storey and Tibshirani, 2003). 575 (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 a third of all human genes, naive extrapolation suggests that the promoter regions of at least 750 human genes have experienced positive selection. Positive selection appears to be as prevalent on the chimpanzee as on the human lineage (Figure 2); the p-value distributions are not significantly separated (two-tailed Mann–Whitney $p = 0.63$). Positive selection appears 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 (Mi et al., 2005). Of the 6280 analyzed genes, 3850 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 (cf. Table S2). 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 (e.g., protein folding, oncogene, and anion transport). We therefore 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, http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM.)

Neural development and function are prominent themes, especially in humans. Relevant PANTHER categories include neurogenesis, ectoderm development, nerve–nerve synaptic transmis-

sion, 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, including *PRSS12*, *NTRK2*, *UCHL3*, *ME2*, *STX1A*, and *SCN1A*, and similar functions elsewhere in the nervous system, including *ISL2*, *SLIT2*, *CHRNA9*, *ADAM22*, *SCN9A*, and *GLRA1*. Several of these genes have variants known to be associated with diseases, including a coding deletion in *PRSS12* associated with mental retardation and coding polymorphisms in *ME2* and *SCN1A* associated with epilepsy. *NTRK2*, *STX1A*, and *SLIT2* also score high in chimpanzees. Several genes relevant to neurodegenerative diseases score high in humans, including *SCRG1*, which is overexpressed in Creutzfeldt–Jakob disease; *TMED10*, whose protein product inhibits production of amyloid beta peptides, whose accumulation is a critical feature of Alzheimer disease; and *ITM2C*, whose protein products directly interact with beta-secretase, which cleaves amyloid precursor protein. *TMED10* also scores high in chimpanzees. The scores of *TMED10* and *ITM2C* are intriguing in view of evidence that humans are more susceptible than chimpanzees to certain pathologies of Alzheimer disease (Olson and Varki, 2003). 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 appears that positive selection has particularly targeted the regulation of glucose metabolism. Relevant PANTHER categories include carbohydrate metabolism, glycolysis, other polysaccharide metabolism, and anion transport. Glucose metabolism-related genes scoring high in humans include *HK1* (hexokinase 1), which catalyzes the first step in glycolysis (i.e., the protein products of this gene do so); *GCK* (glucokinase), which does likewise and is a major regulator of glucose metabolism; *GPI* (glucose-6-phosphate isomerase), which catalyzes the second step in glycolysis; *PFKFB3*, which indirectly affects the activity of phosphofructokinase, which catalyzes the third step in glycolysis; *GCG* (glucagon), which stimulates gluconeogene-

sis and glycogenolysis; *GALE* (galactose epimerase), which catalyzes the last step in galactose metabolism (from UDP-galactose to UDP-glucose); *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 in *GCK* and *ABCC8* associated with type 2 diabetes and hyperinsulinemic hypoglycemia, respectively. *GCG* and *PFKFB3* also score high in chimpanzees. Other nutrition-related genes scoring high in humans include *LDHA* (lactate dehydrogenase-A), 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. *MMP20*, *HSD17B4*, *USHBP1*, and *SCLY* also score high in chimpanzees. 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. In one survey of coding sequences (Clark et al., 2003), the PANTHER amino acid metabolism category is enriched for positive selection. The scores of genes such as *USHBP1*, *HPD*, and *SCLY* affirm that positive selection has targeted amino acid metabolism, not only through protein structure but also through gene regulation.

Using the Novartis Gene Expression Atlas (Su et al., 2004), 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 5049 genes analyzed by both us and Novartis and each of 73 non-cancerous tissues assessed by Novartis, we computed a score

between 0 and 1 representing the specificity of the gene to the tissue (cf. text supplement); the specificity score of a gene for its tissue of maximal expression is low if the gene is 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 is negative, indicating an association of tissue specificity with positive selection. In humans, there is one significant correlation, for pancreas (one-tailed $p = 0.044$) (Figure 3a). This association is consonant with positive selection on metabolic traits, but no gene mentioned above scores high for pancreas specificity. Genes scoring high for both pancreas specificity and positive selection in humans include *CPB1*, a carboxypeptidase; *SERPINI2*, a protease inhibitor whose disruption causes malnutrition in mice (Loftus et al., 2005); and *ABCC2*, an anion transporter. In chimpanzees, there are several significant correlations, for testis seminiferous tubule (one-tailed $p = 0.024$) and neural tissues led by olfactory bulb and spinal cord (one-tailed $p = 0.0096$ and 0.012, respectively) (Figure 3b–d). The association with testis specificity is consonant with two surveys of coding sequences (Chimpanzee Sequencing and Analysis Consortium, 2005; Nielsen et al., 2005). It should be noted that tissues vary in the extent to which genes are specific to them and 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 Khaitovich et al. (2005), which constitute the most extensive survey currently available of gene expression differences between humans and chimpanzees. For 3317 genes analyzed by both us and Khaitovich et al. and each of five tissues (brain, heart, kidney, liver, and testis) assessed by Khaitovich et al., we computed 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. In humans, all these correlations are nominally negative, consistent with associations of expression divergence with positive selection, but none is statistically significant; the strongest is for kidney (one-tailed $p = 0.086$). However, this weakness is not surprising. Khaitovich et al. measured 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 roles, or others not yet known, positive selection has targeted is beyond the reach of our analyses. Conversely, the functions of other high-scoring genes are almost unknown. For example, for approximately a quarter of the 100 genes scoring highest in humans, we found almost no information, and for approximately the same number, we found only basic biochemical or expression information. Our results motivate 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.

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 an NSF Postdoctoral Fellowship in Biological Informatics to R. H. (Grant No. 0434655).

References

Arcadi, A. C., 2005. Language evolution: What do chimpanzees have to say? *Current Biology* **15**:R884–R886.

Bielawski, J. P., and Yang, Z., 2005. Maximum likelihood methods for detecting adaptive protein evolution. Nielsen, R. (Ed.), *Statistical methods in molecular evolution*, pp. 103–124. Springer, New York, NY.

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.

Bustamante, C. D., Fledel-Alon, A., Williamson, S., Nielsen, R., Hubisz, M. T., Glanowski, S., Tanenbaum, D. M., White, T. J., Sninsky, J. J., Hernandez, R. D., Civello, D., Adams, M. D., Cargill, M., and Clark, A. G., 2005. Natural selection on protein-coding genes in the human genome. *Nature* **437**:1153–1157.

Carroll, S. B., 2003. Genetics and the making of *Homo sapiens*. *Nature* **422**:849–857.

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

Clark, A. G., Glanowski, S., Nielsen, R., Thomas, P. D., Kejariwal, A., Todd, M. A., Tanenbaum, D. M., Civello, D., Lu, F., Murphy, B., Ferriera, S., Wang, G., Zheng, X., White, T. J., Sninsky, J. J., Adams, M. D., and Cargill, M., 2003. Inferring nonneutral evolution from human–chimp–mouse orthologous gene trios. *Science* **302**:1960–1963.

Cooper, S. J., Trinklein, N. D., Anton, E. D., Nguyen, L., and Myers, R. M., 2006. Comprehensive analysis of transcriptional promoter structure and function in 1% of the human genome. *Genome Research* **16**:1–10.

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.

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.

Johnson-Frey, S. H., 2003. What's so special about human tool use? *Neuron* **39**:201–204.

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.

Khaitovich, P., Hellmann, I., Enard, W., Nowick, K., Leinweber, M., Franz, H., Weiss, G., Lachmann, M., and Pääbo, S., 2005. Toward a neutral evolutionary model of gene expression. *Science* **309**:1850–1854.

King, M.-C., and Wilson, A. C., 1975. Evolution at two levels in humans and chimpanzees. *Science* **188**:107–116.

Loftus, S. K., Cannons, J. L., Incao, A., Pak, E., Chen, A., Zerfas, P. M., Bryant, M. A., Biesecker, L. G., Schwartzberg, P. L., and Pavan, W. J., 2005. Acinar cell apoptosis in *Serpini2*-deficient mice models pancreatic insufficiency. *PLoS Genetics* **1**:0369–0379.

Majewsky, J., and Ott, J., 2002. Distribution and characterization of regulatory elements in the human genome. *Genome Research* **12**:1827–1836.

Mi, H., Lazareva-Ulitsky, B., Loo, R., Kejariwal, A., Vandergriff, J., Rabkin, S., Guo, N., Muruganujan, A., Doremieux, O., Campbell, M. J., Kitano, H., and Thomas, P. D., 2005. The PANTHER database of protein families, subfamilies, functions and pathways. *Nucleic Acids Research* **33**:D284–D288.

Nielsen, R., Bustamante, C., Clark, A. G., Glanowski, S., Sackton, T. B., Hubisz, M. J., Fledel-Alon, A., Tanenbaum, D. M., Civello, D., White, T. J., Sninsky, J. J., Adams, M. D., and Cargill, M., 2005. A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biology* **3**:0976–0985.

Olson, M. V., and Varki, A., 2003. Sequencing the chimpanzee genome: Insights into human evolution and disease. *Nature Reviews Genetics* **4**:20–28.

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., 2006. Forces shaping the fastest evolving regions in the human genome. *PLoS Genetics* **2**:1599–1611.

Prabhakar, S., Noonan, J. P., Pääbo, S., and Rubin, E. M., 2006. Accelerated evolution of conserved noncoding sequences in humans. *Science* **314**:786–786.

Rockman, M. V., Hahn, M. W., Soranzo, N., Zimprich, F., Goldstein, D. B., and Wray, G. A., 2005. Ancient and recent positive selection transformed opioid *cis*-regulation in humans. *PLoS Biology* **3**:2208–2219.

Sabeti, P. C., Schaffner, S. F., Fry, B., Lohmueller, J., Varilly, P., Shamovsky, O., Palma, A., Mikkelsen, T. S., Altshuler, D., and Lander, E. S., 2006. Positive natural selection in the human lineage. *Science* **312**:1614–1620.

Sorek, R., and Ast, G., 2003. Intronic sequences flanking alternatively spliced exons are conserved between human and mouse. *Genome Research* **13**:1631–1637.

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.

Su, A. I., Wiltshire, T., S.Batalov, Lapp, H., Ching, K. A., Block, D., Zhang, J., Soden, R., Hayakawa, M., Kreiman, G., Cooke, M. P., Walker, J. R., and Hogenesch, J. B., 2004. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proceedings of the National Academy of Sciences of the United States of America* **101**:6062–6067.

Tishkoff, S. A., Reed, F. A., Ranciaro, A., Voight, B. F., Babbitt, C. C., Silverman, J. S., Powell, K., Mortensen, H. M., Hirbo, J. B., Osman, M., Ibrahim, M., Omar, S. A., Lema, G., Nyambo, T. B., Ghori, J., Bumpstead, S., Pritchard, J. K., Wray, G. A., and Deloukas, P., 2007. Convergent adaptation of human lactase persistence in Africa and Europe. *Nature Genetics* **39**:31–40.

Ungar, P. S. (Ed.), 2007. *Evolution of the human diet: The known, the unknown, and the unknowable*. Oxford University Press, Oxford, UK.

Vallender, E. J., and Lahn, B. T., 2004. Positive selection on the human genome. *Human Molecular Genetics* **13**:R245–R254.

Voight, B. F., Kudaravalli, S., Wen, X., and Pritchard, J. K., 2006. A map of recent positive selection in the human genome. *PLoS Biology* **4**:446–458.

Wang, E. T., Kodama, G., Baldi, P., and Moyzis, R. K., 2006. Global landscape of recent inferred Darwinian selection for *Homo sapiens*. *Proceedings of the National Academy of Sciences of the United States of America* **103**:135–140.

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

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.

Yu, X.-J., Zheng, H.-K., Wang, J., Wang, W., and Su, B., 2006. Detecting lineage-specific adaptive evolution of brain-expressed genes in human using rhesus macaque as outgroup. *Genomics* **88**:745–751.

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 legends

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 (cf. Table S1 for a fuller presentation). H, C, and M label the human, chimpanzee, and macaque lineages. Red and black indicate the foreground and background lineages. 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.

Figure 2. Positive selection in chimpanzees vs. humans. Each point represents one gene, and the horizontal (vertical) axis represents p-value on the human (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 humans (chimpanzees) only are plotted toward the lower right (upper left), and 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.)

Figure 3. Positive selection and specificity to **(a)** pancreas, **(b)** testis seminiferous tubule, **(c)** olfactory bulb, or **(d)** spinal cord. Each plot is isomorphic to **Figure 2**, with each point color-coded to indicate the specificity of the gene it represents: brighter red indicates higher specificity. Many (few) genes are highly specific to testis seminiferous tubule (olfactory bulb)—there are many (few) bright points. Specificity to pancreas (testis seminiferous tubule, olfactory bulb, or spinal cord) is associated with positive selection on the human (chimpanzee) lineage—most bright points lie below (above) the dashed blue line.

Figure 1

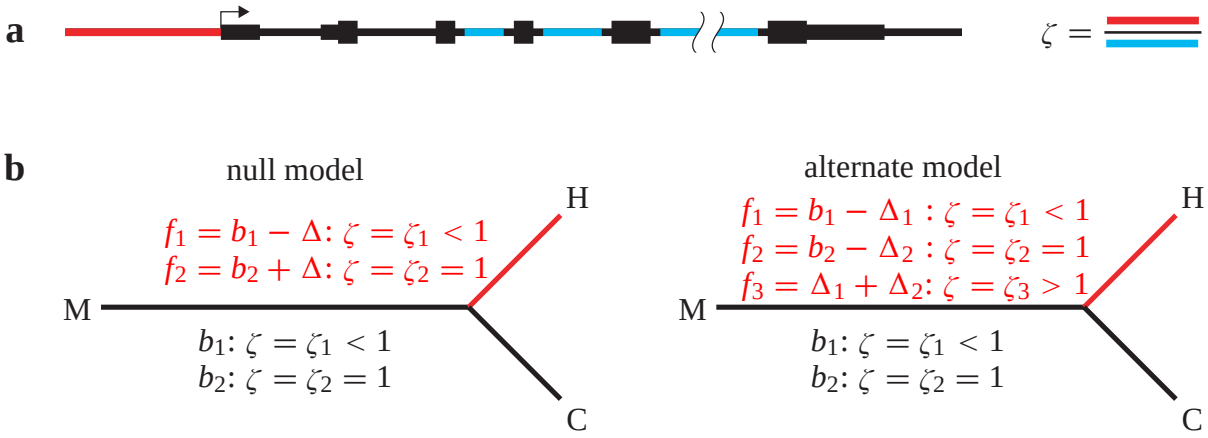
a
ζ =
b
null model
$f_1 = b_1 - \Delta$: $\zeta = \zeta_1 < 1$
$f_2 = b_2 + \Delta$: $\zeta = \zeta_2 = 1$
H
M
C
b_1: $\zeta = \zeta_1 < 1$
b_2: $\zeta = \zeta_2 = 1$
alternate model
$f_1 = b_1 - \Delta_1$: $\zeta = \zeta_1 < 1$
$f_2 = b_2 - \Delta_2$: $\zeta = \zeta_2 = 1$
$f_3 = \Delta_1 + \Delta_2$: $\zeta = \zeta_3 > 1$
H
M
C
b_1: $\zeta = \zeta_1 < 1$
b_2: $\zeta = \zeta_2 = 1$

Figure 2

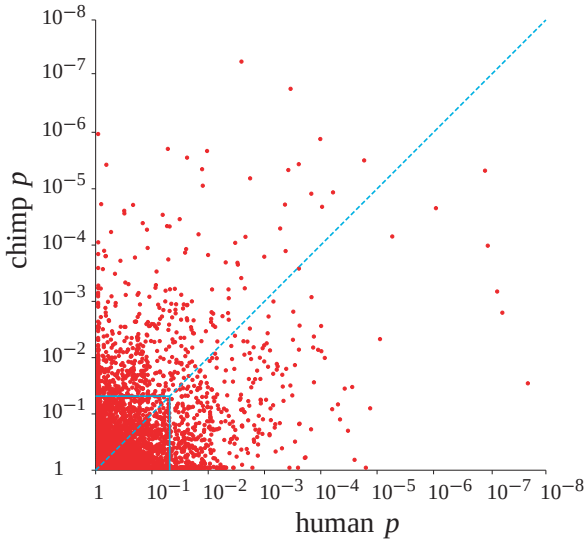

Figure 3

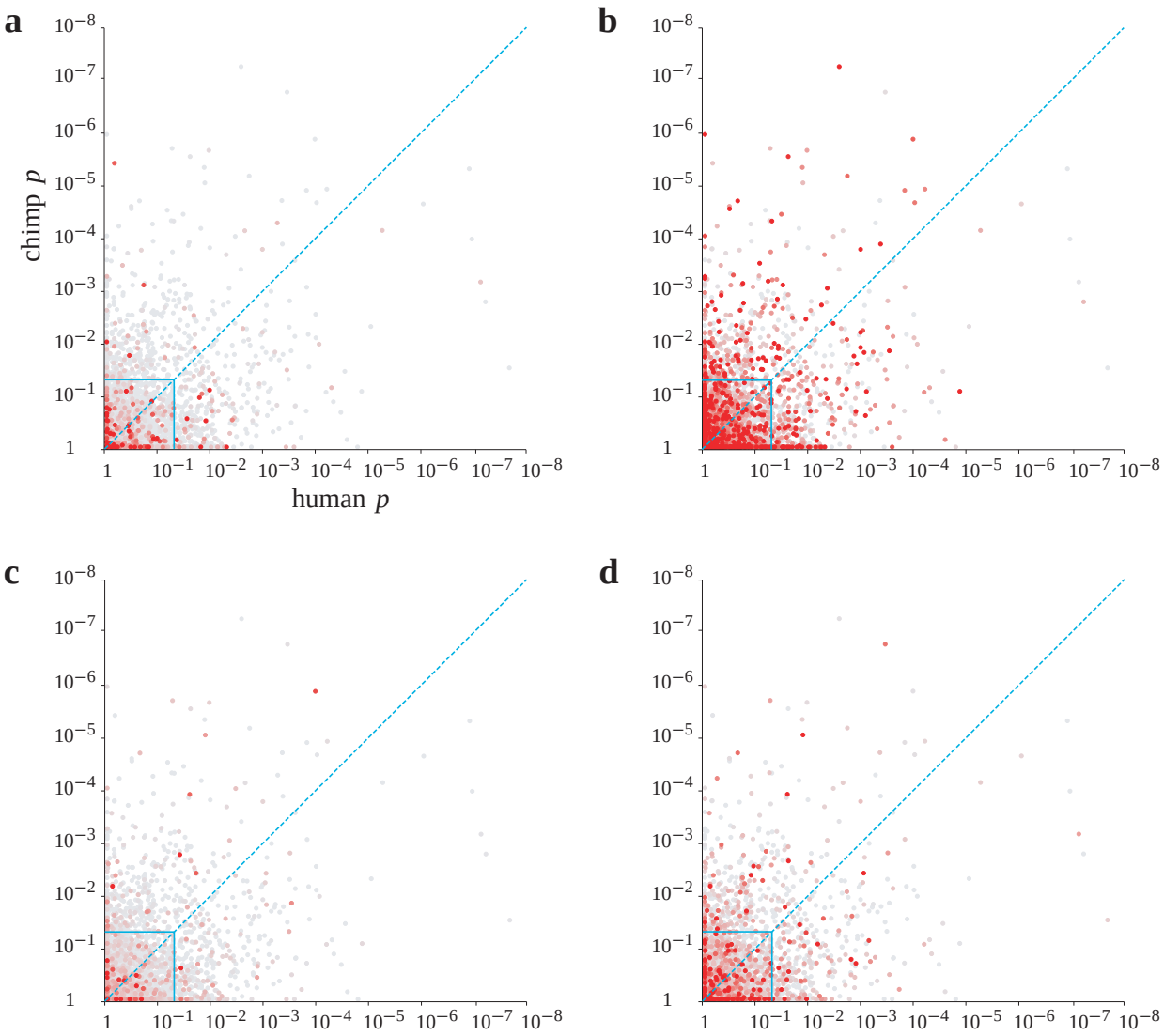

Table 1: PANTHER biological process categories enriched for positive selection[1]

a: In humans

category[2]	analyzed genes	human p_{MW}[3]	chimp p_{MW}[3]
protein folding	70	0.0067	0.77
other neuronal activity[4]	31	0.013	0.039
neurogenesis[5]	133	0.013	0.032
glycolysis[6]	21	0.014	0.72
neuronal activities[4]	137	0.020	0.22
carbohydrate metabolism[6]	210	0.020	0.017
ectoderm development[5]	169	0.020	0.11
mesoderm development	161	0.024	0.17
nerve–nerve synaptic transmission[4]	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: In chimpanzees

category[2]	analyzed genes	chimp p_{MW}[3]	human p_{MW}[3]
DNA replication	34	0.013	0.41
carbohydrate metabolism[7]	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[7]	44	0.041	0.43
blood clotting	32	0.049	0.47

[1]Cf. Table S2 for further analyses.

[2]Ordered 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.

[3]Nominal 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.

[4]The 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 $p_{\mathrm{MW}} = 0.46$ and chimp $p_{\mathrm{MW}} = 0.62$.

[5]The neurogenesis category is contained in the ectoderm development category. For the remainder of the ectoderm development category, human $p_{\mathrm{MW}} = 0.44$ and chimp $p_{\mathrm{MW}} = 0.81$.

[6]The glycolysis category is contained in the carbohydrate metabolism category. For the remainder of the carbohydrate metabolism category, human $p_{\mathrm{MW}} = 0.080$ and chimp $p_{\mathrm{MW}} = 0.0078$.

[7]The other polysaccharide metabolism category is contained in the carbohydrate metabolism category. For the remainder of the carbohydrate metabolism category, chimp $p_{\mathrm{MW}} = 0.073$ and human $p_{\mathrm{MW}} = 0.014$.