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

February 7, 2007

Surveys of protein-coding sequences for signatures of positive selection in humans and chimpanzees have flagged surprisingly few genes known to be involved in neural or nutritional processes[1–5], despite the pronounced differences between humans and chimpanzees in behavior, cognition, and diet[6–8]. It may be that most such differences are due to changes in gene regulation rather than protein structure[9]. Here, we present the first survey of promoter (5′-flanking) regions, which are rich in *cis*-regulatory sequences, for signatures of positive selection on the human lineage. 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 glucose metabolism.

Cognitive, behavioral, and dietary differences are among the most conspicuous differences between humans and their closest relatives, chimpanzees and other great apes. For example, even in the absence of written language or agriculture, human communication and tools are much more complex than those of chimpanzees[6,7], and humans consume a far wider range of foods than chimpanzees[8]. These traits are essential to many aspects of human ecology, such as the broad range of habitats humans occupy[8], and although assessing the adaptive significance of a trait is often challenging, it is plausible that many human cognitive, behavioral, and dietary traits 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 alleles) in humans[10,11]. Surprisingly, however, such genes are not prominent in surveys of coding sequences for evidence of positive selection in humans and chimpanzees[1–5]. Instead, these surveys have flagged many genes known to function in immunity, olfaction, and spermatogenesis, among other processes. Neural-related genes in particular show little sign of positive selection in these surveys[3,4].

One possible explanation, first suggested by King and Wilson[9], is that many phenotypic differences between humans and chimpanzees may be due to changes in gene regulation rather

than protein structure. 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), very few of which lie within coding sequences[12]. 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[13], and the other, *LCT*, is nutrition-related[14]. 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[15,16]. Third, two surveys of regions that are highly conserved across vertebrates except for extensive changes in humans, which might be due to positive selection, found excesses of such regions in the vicinity of genes in several neural-related categories[17,18]. These studies, limited 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 during human evolution. Because *cis*-regulatory sequences are scattered, short, and imprecise, most have not yet 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′) from transcription start sites and identified associations of functional categories and expression domains with evidence of positive selection on these regions.

Our approach is to compare the rates of evolution along the human lineage between a promoter region and chosen, 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. This includes some or all of both the so-called core and extended promoters. 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 sequences[19–21], the ends of each intron, which

contain splicing signals[22], and the centers of large introns, which may often contain *cis*-regulatory sequences[19]. These sequences are generally among the least constrained in the genome[3,23,24], so they are 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 faster than the associated intronic sequences, it is likely that *cis*-regulatory sequences within the promoter region have been under positive selection. (The **Supplementary Methods** online present evidence that other conceivable 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 publicly available 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 the rates of evolution, we fitted by maximum likelihood two models of single-nucleotide substitutions to each promoter alignment and the associated intronic alignment (**Fig. 1b**). The fitted parameters include ζ (zeta), the ratio of substitution rates in the promoter region to those in the associated intronic sequences[25]; 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 yields a p-value for consistency of the alignments with the null model[26]. A small p-value constitutes a high score for positive selection on the promoter region; 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 the frequencies of the classes being fitted parameters. A high score requires that some

but not all or even most promoter sites have evolved faster than intronic sites. The null model accommodates promoter sites that have evolved under negative selection on the chimpanzee and macaque lineages but neutrally on the human lineage[26]. The contrast between the models is therefore 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[27]. We repeated our analyses allowing ζ to be greater than 1 on the chimpanzee rather than the human lineage. (**Supplementary Tables 1–4** online present complete results.)

Several potential concerns arise regarding these analyses, ranging from data quality to interpretational issues. The **Supplementary Methods** online explain our data filtering and statistical techniques and present several auxiliary analyses and other considerations encouraging confidence that many high-scoring genes are genuine cases of positive selection on promoter regions. In particular, it is unlikely that our results are dominated by errors in base calling, genome assembly, ortholog identification, or sequence alignment, by small-sample fluctuations, by elevated mutation rates or biased gene conversion in promoter regions, or by negative selection on intronic sequences.

Of the 6280 analyzed genes, 46 (0.73%) have $q < 0.05$, so the 5% false discovery rate set is nonempty[27]. 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 implies that the promoter regions of at least 750 human genes have experienced positive selection. This is of the same order of magnitude as in surveys of coding sequences; methodological differences complicate more precise comparisons. For promoter regions, positive selection appears 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 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 (http://www.pantherdb.org). 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 score higher than analyzed genes outside the category (by Mann–Whitney testing). **Table 1** lists the most significant results. This assessment is instructive but limited, in that many genes lack PANTHER classifications, many others have classifications that do not encompass all available information about their functions (e.g., the low density lipoprotein receptor *LDLR* is well known to play an important role in cholesterol homeostasis but is classified by PANTHER as being involved only in oogenesis), and some PANTHER categories do not immediately correspond to organismal traits (e.g., protein folding, oncogene, and anion transport). Accordingly, we surveyed the biomedical literature for information about the 100 highest-scoring genes on the human lineage 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 on the human lineage. 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, including *PRSS12*, *NTRK2*, *ME2*, *STX1A*, and *SCN1A*, and in similar functions elsewhere in the nervous system, including *ISL2*, *SLIT2*, *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; *ROBO3*, a receptor of *SLIT2*, scores high in chimpanzees only. At least three genes apparently relevant to neurodegenerative diseases score high in humans, namely, *SCRG1*, which is overexpressed in

Creutzfeldt–Jakob disease; *TMED10*, which inhibits production of amyloid beta peptides, whose accumulation is a critical feature of Alzheimer disease; and *ITM2C*, which directly interacts with beta-secretase, which cleaves amyloid precursor protein. *TMED10* also scores high in chimpanzees. The apparent relevance to Alzheimer disease is intriguing in view of observations that humans are more susceptible than chimpanzees to some pathologies of this disease[28]. The PANTHER neurogenesis and other neuronal activity categories are enriched with high-scoring genes in both species, but only five of these 31 genes score high in both species, 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 on the human lineage, 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 proteins *HK1* encodes 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 gluconeogenesis 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; and *FOXC2*, a transcription factor that is a major regulator of adipocyte metabolism. All of these genes except *GCG* have variants known to be associated with diseases, including a promoter polymorphism in *GCK* associated with type 2 diabetes and coronary artery disease. *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 of 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 with high-scoring genes in both species, but only seven of these 45 genes score high in both species. In one survey of coding sequences[1], the PANTHER amino acid metabolism category is enriched with high-scoring genes in both species. We do not see such categorical enrichment, but the high 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 (http://symatlas.gnf.org), we explored whether positive selection on promoter regions is associated with gene expression in particular tissues or cell types. This kind of analysis is complicated by the fact that most genes are expressed in multiple tissues, and even if a gene is maximally expressed in some tissue, it may be nearly as highly expressed in others, so simply associating genes with their tissues of maximal expression is unsatisfactory. For each of 5049 genes analyzed by both us and Novartis and for each of the 73 non-cancerous tissues analyzed by Novartis, we therefore computed a score between 0 and 1 representing the specificity of the gene to the tissue (cf. Methods); 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 specificity to the tissue with positive selection. On the human lineage, there is one significant correlation, for pancreas (one-tailed $p = 0.044$) (**Fig. 3**). This association is consistent with positive selection on nutritional traits, but perhaps surprisingly, the correlation for pancreatic beta cells separately is not significant, and no gene mentioned above scores high for pancreas specificity. Genes scoring

high for both pancreas specificity and positive selection on the human lineage include *CPB1*, a carboxypeptidase; *SERPINI2*, a protease inhibitor whose disruption causes malnutrition in mice[29]; and *ABCC2*, an anion transporter. On the chimpanzee lineage, there are several significant correlations, for testis seminiferous tubule (one-tailed $p = 0.024$) and seven neural tissues led by olfactory bulb and spinal cord (one-tailed $p = 0.0096$ and 0.012) (**Fig. 3**). The association with testis specificity 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 and hence the potential for detecting an association with positive selection. Moreover, the expression of a gene may be under positive selection in a tissue to which the gene is not specific. A fully satisfactory analysis requires knowledge of how promoter sequence variation relates to expression variation.

We compared our results to those of Khaitovich et al.[30], 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 for each of the five tissues (brain, heart, kidney, liver, and testis) analyzed 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. All the 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 $p = 0.086$). 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 undoubtedly 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 highest-scoring genes, we found almost no information, and for approximately the same number, we found only basic biochemical or expression information. Our results motivate functional characterization 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. However, much further work is needed to confirm and elaborate this suggestion, in part 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; this 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 assays to map the *cis*-regulatory sequences of neural- and nutrition-related genes and probe the consequences of their changes during human evolution. Similar assays on 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 chosen annotations of the human genome (hg17 of May, 2004) from the UCSC Genome Bioinformatics web site (http://genome.ucsc.edu) and the Genomic tRNA Database web site (http://lowelab.ucsc.edu/GtRNAdb). We started from the UCSC Known Genes collection, which distinguishes alternative transcripts and splices. We parsed each chromosome into clusters of overlapping transcripts and splices, retaining only those in which all transcripts were from the same strand; these are termed “genes” throughout this article. We parsed each gene into regions, taking intersections over alternative transcripts and splices, hence what are termed “promoter sites” and “intronic sites” are such sites

with respect to all 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 departed from 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 (http://www.bx.psu.edu/miller_lab). We masked out bases in chimpanzee and macaque sequences having quality scores under 40, known non-coding RNA genes in human sequences, and bases in 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 over 0.75% such divergence-masked bases or 9% gaps or whose associated intronic alignment contained fewer than 2500 ungapped and unmasked sites. (**Supplementary Tables 1–4** online include results for promoter regions that failed these cutoffs but were otherwise analyzable.) See the **Supplementary Methods** online 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 (http://www.hyphy.org). Our models amount to the HKY85 model modulated by ζ in the promoter region in the same way that Zhang et al.'s[26] preferred models are modulated by ω at nonsynonymous sites (cf. Ref. 25). For each model, we took the best of 10 fits starting from random points, thus guarding 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

(http://faculty.washington.edu/∼jstorey/qvalue). See the **Supplementary Methods** online for further explanation of our statistical techniques.

Assessment of gene functions. We downloaded PANTHER data (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 of expression levels and the axis corresponding to the tissue. 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.

Acknowledgments

We thank J. Pavisic and T. Severson for contributions to the analyses, 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 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

1. Clark, A. G. *et al.* Inferring nonneutral evolution from human–chimp–mouse orthologous gene trios. *Sci.* **302**, 1960–1963 (2003).

2. Bustamante, C. D. *et al.* Natural selection on protein-coding genes in the human genome. *Nat.* **437**, 1153–1157 (2005).

3. Chimpanzee Sequencing and Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nat.* **437**, 69–87 (2005).

4. Nielsen, R. *et al.* A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biol.* **3**, 0976–0985 (2005).

5. 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).

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

7. Arcadi, A. C. Language evolution: What do chimpanzees have to say? *Curr. Biol.* **15**, R884–R886 (2005).

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

9. King, M.-C. & Wilson, A. C. Evolution at two levels in humans and chimpanzees. *Sci.* **188**, 107–116 (1975).

10. Vallender, E. J. & Lahn, B. T. Positive selection on the human genome. *Hum. Mol. Genet.* **13**, R245–R254 (2004).

11. Sabeti, P. C. *et al.* Positive natural selection in the human lineage. *Sci.* **312**, 1614–1620 (2006).

12. Wray, G. A. *et al.* The evolution of transcriptional regulation in eukaryotes. *Mol. Biol. Evol.* **20**, 1377–1419 (2003).

13. Rockman, M. V. *et al.* Ancient and recent positive selection transformed opioid *cis*-regulation in humans. *PLoS Biol.* **3**, 2208–2219 (2005).

14. Tishkoff, S. A. *et al.* Convergent adaptation of human lactase persistence in africa and europe. *Nat. Genet.* **39**, 31–40 (2007).

15. Voight, B. F., Kudaravalli, S., Wen, X. & Pritchard, J. K. A map of recent positive selection in the human genome. *PLoS Biol.* **4**, 446–458 (2006).

16. 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).

17. Pollard, K. S. *et al.* Forces shaping the fastest evolving regions in the human genome. *PLoS Genet.* **2**, 1599–1611 (2006).

18. Prabhakar, S., Noonan, J. P., Pääbo, S. & Rubin, E. M. Accelerated evolution of conserved non-coding sequences in humans. *Sci.* **314**, 786–786 (2006).

19. 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).

20. Crawford, G. E. *et al.* Genome-wide mapping of dnase hypersensitive sites using massively parallel signature sequencing (mpss). *Genome Res.* **16**, 123–131 (2006).

21. Majewsky, J. & Ott, J. Distribution and characterization of regulatory elements in the human genome. *Genome Res.* **12**, 1827–1836 (2002).

22. Sorek, R. & Ast, G. Intronic sequences flanking alternatively spliced exons are conserved between human and mouse. *Genome Res.* **13**, 1631–1637 (2003).

23. Hellmann, I. *et al.* Selection on human genes as revealed by comparisons to chimpanzee cdna. *Genome Res.* **13**, 831–837 (2003).

24. Keightley, P. D., Lercher, M. J. & Eyre-Walker, A. Evidence for widespread degradation of gene control regions in hominid genomes. *PLoS Biol.* **3**, 0282–0288 (2005).

25. Wong, W. S. W. & Nielsen, R. Detecting selection in noncoding regions of nucleotide sequences. *Genet.* **167**, 949–958 (2004).

26. 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).

27. Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. USA* **100**, 9440–9445 (2003).

28. Olson, M. V. & Varki, A. Sequencing the chimpanzee genome: Insights into human evolution and disease. *Nat. Rev. Genet.* **4**, 20–28 (2003).

29. Loftus, S. K. *et al.* Acinar cell apoptosis in *Serpini2*-deficient mice models pancreatic insufficiency. *PLoS Genet.* **1**, 0369–0379 (2005).

30. Khaitovich, P. *et al.* Toward a neutral evolutionary model of gene expression. *Sci.* **309**, 1850–1854 (2005).

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 intronic sequences chosen for our analyses. The fitted parameter ζ is the ratio of substitution rates in the promoter region to those in the intronic sequences. The promoter region is analyzed with all chosen intronic sequences in a window centered on the promoter region, which usually includes intronic sequences from multiple genes. **(b)** Our models. 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 $\zeta < 1$, and the remaining proportion $b_2 = 1 - b_1$ have $\zeta = 1$ in both models. On the foreground lineage, an estimated proportion $\Delta \geq 0$ of promoter sites change from $\zeta < 1$ to $\zeta = 1$ in the null model, and estimated proportions $\Delta_1 \geq 0$ and $\Delta_2 \geq 0$ change from $\zeta < 1$ and $\zeta = 1$ to an estimated $\zeta > 1$ 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 on the human (chimpanzee) lineage only are plotted toward the lower right (upper left), and genes scoring high on both lineages 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 tissue specificity. Each plot is isomorphic to **Figure 2**, with each point color coded to indicate the specificity of the gene it represents to a particular tissue: darker red indicates higher specificity. Many (few) genes are highly specific to testis seminiferous tubule (olfactory bulb)—there are many (few) dark points. Specificity to pancreas (testis seminiferous tubule, olfactory bulb, or spinal cord) is associated with positive selection on the human (chimpanzee) lineage—most dark points lie below (above) the dashed blue line.

Figure 1

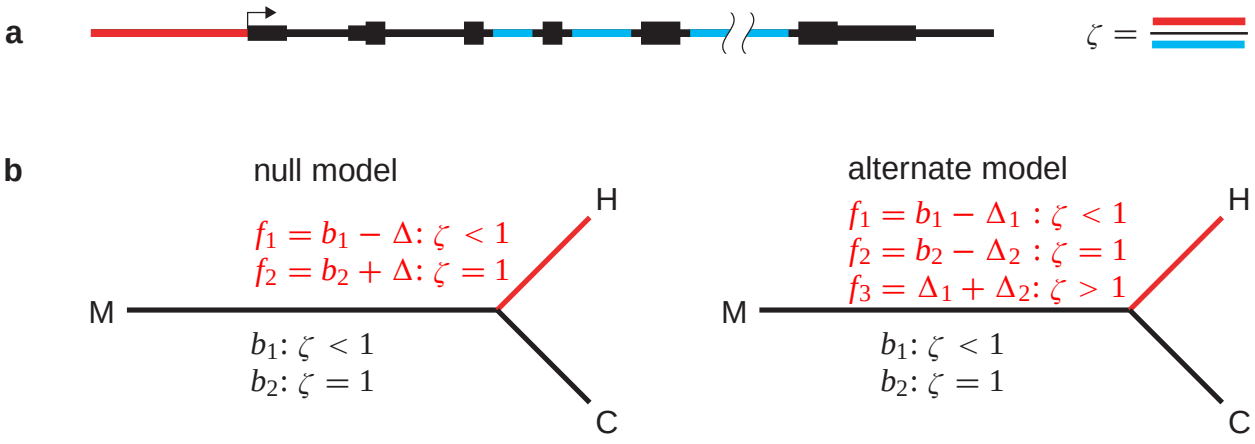

Figure 2

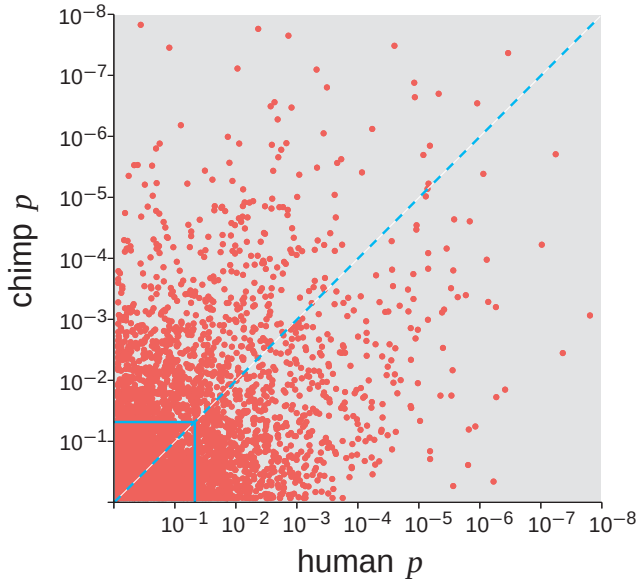

Figure 3

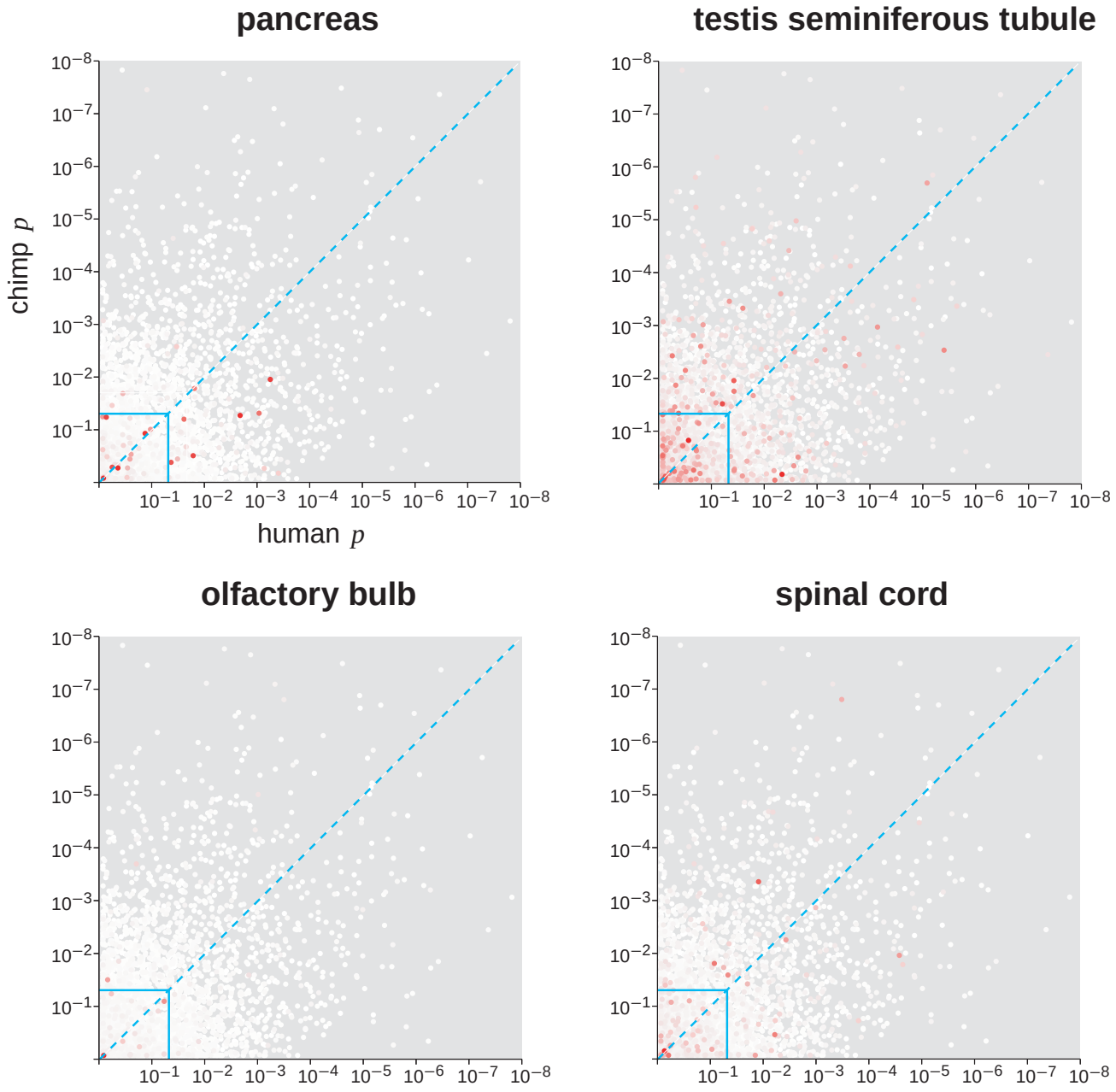

Table 1: PANTHER biological process categories enriched with high-scoring genes

a: On the human lineage

category[1]	analyzed genes	human p_{MW}[2]	chimp p_{MW}[2]
protein folding	70	0.0067	0.77
other neuronal activity[3]	31	0.013	0.039
neurogenesis[4]	133	0.013	0.032
glycolysis[5]	21	0.014	0.72
neuronal activities[3]	137	0.020	0.22
carbohydrate metabolism[5]	210	0.020	0.017
ectoderm development[4]	169	0.020	0.11
mesoderm development	161	0.024	0.17
nerve–nerve synaptic transmission[3]	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: On the chimpanzee lineage

category[1]	analyzed genes	chimp p_{MW}[2]	human p_{MW}[2]
DNA replication	34	0.013	0.41
carbohydrate metabolism[6]	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[6]	44	0.041	0.43
blood clotting	32	0.049	0.47

[1]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.

[2]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.

[3]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 $p_{\mathrm{MW}} = 0.62$.

[4]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 $p_{\mathrm{MW}} = 0.81$.

[5]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 $p_{\mathrm{MW}} = 0.0078$.

[6]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 $p_{\mathrm{MW}} = 0.014$.