Supplementary Material

Rapidly evolving human promoter regions

Martin S Taylor, Tim Massingham, Yoshihide Hayashizaki, Piero Carninci, Nick Goldman & Colin A M Semple.

email: mst@ebi.ac.uk

Discussion

Several lines of evidence complement the reanalysis of Haygood *et al.'*s data, presented in the main section of this manuscript, to support the conclusion that promoter regions have a higher rate of substitution (and insertion, and deletion) than other regions of the genome. Here we briefly outline those lines of evidence.

Our previous work (Taylor *et al*. 2006; Carninci *et al*. 2006) was not confined to analysis of the primate lineage, but also considered the evolution of experimentally defined promoter regions in the rodent lineage. Only 0.7% of protein coding gene (PCG) promoter regions were evolving faster than the neutral estimate in the rodent lineage (Taylor *et al*. 2006). The vast majority of rodent promoter regions exhibited a robust signal of purifying selection. This presents a stark contrast to the 61% of PCG promoter regions we found to be evolving faster than our neutral rate estimates in the primate lineage (Taylor *et al*. 2006), and the 46% we find by reanalysis of Haygood *et al*.'s data. The difference between these datasets, 61% versus 46%, can be explained by alternate definitions of transcription start sites between studies, and the biased filtering of alignments by Haygood *et al*, where rapidly evolving promoters regions were partially masked or excluded entirely, but similar filtering was not applied to the intronic alignments. This filtering was appropriate and conservative with respect to a screen for positive selection (Haygood *et al*. 2007) but introduces a systematic bias against rapidly evolving promoters for the purposes of establishing genome-wide trends in promoter evolution.

Discordance between the rodent and primate lineages in promoter evolution has previously been noted by Keightley *et al.* (2005). They reported that promoter regions show widespread evidence of purifying selection in the rodent lineage, whereas in the primate lineage the evolutionary rate of promoters was not significantly different from a local neutral rate estimate (although a trend is apparent in their data, suggesting regions upstream of genes do exceed the neutral estimate). They interpreted this as a loss of constraint in hominid promoter regions (Keightley *et al*. 2005). Our analysis of a much larger dataset, shows that the genome-wide primate promoter region rate significantly exceeds the neutral rate estimate, an observation that cannot be explained by a loss of constraint alone. All of these studies (Keightley *et al*. 2005; Taylor *et al*. 2006; Haygood *et al*. 2007) have controlled for and excluded trivial explanations such as sequence error, CpG effects and the megabase-scale variation in mutation rate across mammalian genomes. This leaves two competing, but not necessarily mutually exclusive, theories: (1) that there has been widespread positive selection on promoter sequences in the primate lineage, possibly coupled to a general relaxation of selective constraints, and (2) that primate promoters have a higher background mutation rate than introns and other nearly-neutral control sequences.

The arguments put forward (Keightley *et al*. 2005) to suggest that there has been a relaxation of selective constraint, based on the small effective size of hominid populations, also argue for a reduction in the efficiency of positive selection relative to other mammals. This makes it difficult to reconcile the coupling of relaxed constraint with positive selection as an explanation for the higher than neutral promoter substitution rates. Many of the human promoter regions (9.9%; 619/6280) do show a significant signal of constraint in the branch-only test. Single-nucleotide resolution

substitution profiles (Taylor *et al*. 2006), particularly within core promoter regions, are consistent between the primate and rodent lineages and indicate that purifying selection has continued to act in a consistent manner despite the measured rates exceeding the neutral estimates in the primate lineage. This conclusion is reinforced by the results of Bush and Lahn (2005).

It is also interesting to consider the functional profile of genes whose promoters have been identified by Taylor *et al*. and Haygood *et al*. The 0.7% of rodent promoter regions evolving faster than neutral estimates[3] represent 83 genes that were significantly enriched for immunological- and apoptosis-related annotation. Both of these categories have previously been found to be over-represented in screens for positive selection in protein coding sequences (Nielsen e al. 2005). In contrast, the 61% of rapidly evolving human PCG promoter regions were significantly enriched for terms relating to basal metabolic activity (e.g. "primary metabolism") (Taylor *et al*. 2006) that we considered to represent housekeeping functions. As such, they seem to be unlikely candidates for positive selection in the primate lineage and may indicate a correlation between germline transcriptional activity and elevated substitution rate (Taylor *et al*. 2006). Similar terms (e.g. "glycolysis", "carbohydrate metabolism", "coenzyme metabolism" and "protein folding") are abundant in the over-represented categories presented by Haygood *et al*., who have interpreted this as positive selection driven by dietary changes in the primate lineage.

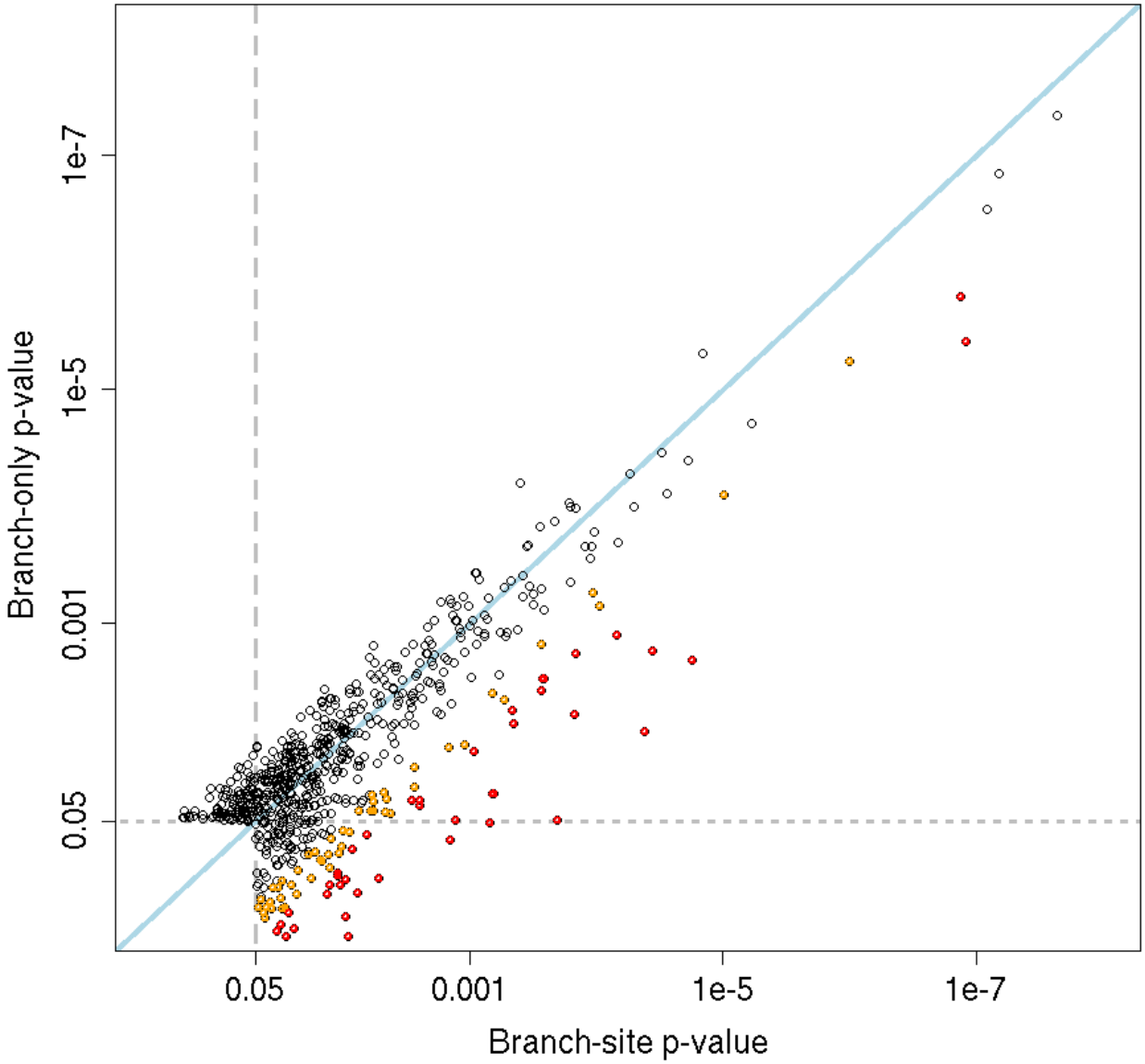
Detailed methods

Our reanalysis of Haygood *et al*.'s data used the entire "fiducial" alignments of both promoter regions and their paired, local intron derived sequences (kindly provided by Ralph Haygood). Branch lengths were estimated for the unrooted tree: (human, chimpanzee, macaque).

The presence of lineage-specific rate acceleration was detected by comparing two probabilistic models of sequence evolution: a null model in which the length of the branch leading to human was constrained to be equal in both the promoter and intronic alignments (nine free parameters), and a more general alternative model where the length of the branch may differ between the alignments (ten free parameters). The nucleotide substitution process was described using separate HKY models with categorical (ncatG=5) gamma rate variation over sites (Hasegawa *et al*, 1985; Yang 1994) for each of the two alignments, capturing the effects of differences in rates of transition and transversion substitution, base composition and rate heterogeneity.

These models were implemented using the baseml component of PAML (version 4.0; Yang, 2007). The null model was a modification of the Mgene=4 model, instead of the standard implementation with a single tree-scaling parameter, the human branch was constrained to be equal for both the promoter and intronic alignments and separate scaling parameters were introduced for the macaque and chimpanzee branches. The modified source code is available on request. The alternate model was the standard implementation of Mgene=0, with lnL scores summed over paired promoter and intronic alignments. To avoid local maxima in the likelihood optimisation, we took the best of 100 fits with random, valid, parameter initialisation. Statistical significance was assessed by one-tailed comparison of the likelihood-ratio test statistic ($\ln L_{alternate} - \ln L_{null}$) to a chi-squared distribution with one degree of freedom. *Q*-value false discovery rates (FDR) were calculated using the method of Story and Tibshirani (2003) implemented in the Qvalue package (version 1.1) of R. A *Q*-value adjusted significance threshold of $p < 0.047$ was used for the branch-only test ($p = 0.047$ has a *Q*-value FDR = 0.545, equal to that reported by Haygood *et al*. for $p = 0.05$ in the branch-site test). All branch length values reported are from the best fit of the alternate model as described above (unconstrained branch lengths with independent estimation of all parameters for promoter and paired intronic alignments).

Supplementary Figure 1



Supplementary Figure 1; Comparison of p-value distributions for promoters containing sites putatively subject to positive section: branch-site p-values from Haygood et al. (2007) (x-axis); plotted against promoters with higher average substitution rates than paired intronic alignments: branch-only p-values (y-axis). Axes show p-values on a $\log_{10}$ scale, more significant values to the top and right. The light-blue solid line indicates equality of p-value estimates. Thresholds for nominal significance ($p < 0.05$) are shown as dotted (branch-only) and dashed (branch-site) grey lines. Instances where the branch-site p-value exceeds an arbitrary threshold of x5 more significant than the branch-only p-value are highlighted in orange, and least x10 highlighted in red.

References

Bush, E.C. & Lahn, B.T. *PLoS Comput Biol*. **1**, e73 (2005).
Carninci, P. *et al. Nat. Genet.* 38, 626-635 (2006).
Hasegawa, M., Kishino, H. & Yano, T. J. *Mol. Evol*. **22**, 160-174 (1985).
Haygood, R., Fedrigo, O., Hanson, B., Yokoyama, K.D. & Wray, G.A. *Nat. Genet*. **39**, 1140-1144 (2007).
Keightley, P.D., Lercher, M.J. & Eyre-Walker, A. *PLoS Biol*. **1**, e42 (2005).
Nielsen, R. *et al*. *PLoS Biol*. **3**, e170 (2005).
Story, J.D. and Tibshirani, R. *Proc. Natl. Acad. Sci. USA* **100**, 9440-9445 (2003).
Taylor, M.S. *et al*. *PLoS Genet*. **2**, e30 (2006).
Yang, Z. *J. Mol. Evol*. **39**, 306-314 (1994).
Yang, Z. *Mol. Biol. Evol.* **24**, 1586-1591 (2007).

Acknowledgements

We thank Ziheng Yang for informative discussions regarding the implementation of the null model in the likelihood ratio test.