

Decreased Transcription Factor Binding Levels Nearby Primate Pseudogenes Suggest Regulatory Degeneration

Gavin M. Douglas,¹ Michael D. Wilson,^{2,3} and Alan M. Moses^{*,1,4}

¹Department of Ecology and Evolutionary Biology, University of Toronto, Toronto, ON, Canada

²Genetics and Genome Biology Program, SickKids Research Institute, Toronto, ON, Canada

³Department of Molecular Genetics, University of Toronto, Toronto, ON, Canada

⁴Department of Cell and Systems Biology, University of Toronto, Toronto, ON, Canada

*Corresponding author: E-mail: alan.moses@utoronto.ca.

Associate editor: Ilya Ruvinsky

Abstract

Characteristics of pseudogene degeneration at the coding level are well-known, such as a shift toward neutral rates of nonsynonymous substitutions and gain of frameshift mutations. In contrast, degeneration of pseudogene transcriptional regulation is not well understood. Here, we test two predictions of regulatory degeneration along a pseudogenized lineage: 1) Decreased transcription factor (TF) binding and 2) accelerated evolution in putative *cis*-regulatory regions. We find evidence for decreased TF binding levels nearby two primate pseudogenes compared with functional liver genes. However, the majority of TF-bound sequences nearby pseudogenes do not show evidence for lineage-specific accelerated rates of evolution. We conclude that decreases in TF binding level could be a marker for regulatory degeneration, while sequence degeneration in primate *cis*-regulatory modules may be obscured by background rates of TF binding site turnover.

Key words: pseudoenhancer, regulatory evolution, ChIP-seq, neutral evolution, histone modifications.

Regulatory evolution is thought to be a key mechanism underlying the diversity of closely related species (Britten and Davidson 1969; King and Wilson 1975). For example, deletions of conserved regulatory elements have been suggested to underlie human-specific traits (McLean et al. 2011). Despite the importance of regulatory degeneration for human evolution, little is known in general about how regulation evolves at the molecular level once selective constraint has been lifted. In addition, recent studies have found relatively little conservation of TF binding in vertebrates (Schmidt et al. 2010; Ballester et al. 2014), suggesting that most changes in TF binding might be neutral with respect to selection. By understanding how transcriptional regulation degenerates, we can help resolve which aspects of transcriptional regulation natural selection is preserving. Here, we study TF binding and *cis*-regulatory modules (CRMs, which we take to refer to promoters, enhancers, silencers, etc., interchangeably) evolution near human pseudogenes to characterize the process of regulatory degeneration, which we refer to as “pseudoenhancerization” (in the case of enhancers) in analogy with pseudogenization, the process whereby natural selection fails to preserve genes. We focus on regulatory degeneration nearby two unitary pseudogenes, because these represent true losses of function from a species rather than degeneration of a redundant copy (Zhang et al. 2010). This study contrasts with approaches that have looked at a large number of potential pseudogenes in a single species (Pei et al. 2012) in that we consider the evolutionary evidence for regulatory degeneration at a small number of known pseudogenes with clear 1:1 orthologs across mammals.

Two classic examples of unitary pseudogenes in human are urate oxidase, *Uox*, and L-gulonolactone oxidase, *Gulo*, which are functional in the livers of most mammals. *Uox* removes excess nitrogen from the body by metabolizing uric acid (Gustafsson and Unwin 2013) and was pseudogenized independently in both human and gibbon (Zhang et al. 2010). *Gulo* is responsible for encoding the protein required for the final step in the vitamin C (ascorbic acid) synthesis pathway and was lost in the ancestor of human and macaque (Drouin et al. 2011).

There are a number of characteristics of pseudogenization that are used to identify pseudogenes and are all present in *Uox* and *Gulo* (Zhang et al. 2010). First, the former coding sequence of pseudogenes evolves faster than the orthologous coding regions in species where the loci are still functional. Similarly, pseudogenes exhibit a K_a/K_s (ratio of nonsynonymous to synonymous substitution rates) that approaches 1 as evolutionary distance increases. Both of these characteristics are due to relaxed purifying selection acting at the sequence level.

In contrast, it is unknown whether regulatory sequences show similar signatures of relaxed selection. Regulatory elements that regulated the functional ancestors of pseudogenes are expected to degrade following the pseudogenization event. In fact, it is possible that the first steps of pseudogenization actually occur at the regulatory level (Wu et al. 1992; Oda et al. 2002). However, to our knowledge, CRM degeneration outside promoters, or “pseudoenhancerization”, has only been identified once before (Leivonen et al. 1999). Decreased transcription factor (TF) binding levels could be a signal of degeneration, which could be caused by long-range

© The Author 2016. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Open Access

chromatin modifications or evolutionary changes at the sequence level. Also, accelerated rate of evolution (relative to functional CRMs) could be a signal of relaxed selection on CRMs. Furthermore, in analogy to the pattern of K_a/K_s in pseudogenes, there could be an excess of substitutions that decrease transcription factor binding site (TFBS) strength near pseudogenes compared with functional genes. Here, we test these predictions by investigating relaxed constraint in CRMs nearby the recently lost primate pseudogenes *Gulo* and *Uox*. We find evidence for decreased TF binding and H3K4me3 enrichment levels nearby both pseudogenes and lineage-specific relaxation of constraint in two CRMs near *Uox*.

We first investigated how the regulation of gene expression degenerates following pseudogenization of a coding region by focusing on the TF binding events of the liver-specific TFs CEBPA, FOXA1, HNF4A, and ONECUT1 (Ballester et al. 2014) nearby the two primate pseudogenes *Uox* and *Gulo*. In addition to these four TFs, we also investigated whether enrichment levels of two histone modifications, H3K4me3 and H3K27Ac, have changed nearby these pseudogenes (Villar et al. 2015). H3K4me3 is considered a marker for active promoters (Santos-Rosa et al. 2002), while H3K27Ac is a marker for enhancers (Creighton et al. 2010).

Although *Uox* was pseudogenized in the human-gibbon clade (Wu et al. 1992; Oda et al. 2002), TFs still bind nearby this locus (fig. 1A). However, a clear difference is the loss of binding at the human promoter compared with the other four mammals (fig. 1B). In contrast, there is no reduction in the number of ChIP-seq control (input) reads in this region (shown as gray histograms; fig. 1) indicating that the absence of ChIP-seq signal is not due to lack of read mappability. Similar trends of lost binding can be seen in human and macaque at the *Gulo* locus as well (supplementary fig. S1, Supplementary Material online). Although these results support our prediction that regulatory degeneration would be accompanied by loss of TF binding, evolutionary changes in mammalian TF binding have been observed at the genome-wide scale (Schmidt et al. 2010; Ballester et al. 2014). We therefore sought to compare the patterns observed for pseudogenes with functional liver genes.

We first confirmed that the loss of expression at pseudogene loci (as measured by RNA-seq) stands out relative to gene expression changes in liver genes. Using a Brownian motion (BM) model (see Materials and Methods) we inferred gene expression in the rodent ancestor (because changes since the rodent ancestor can be estimated more reliably than changes since the primate ancestor) and then calculated the change in gene expression from this ancestor for these 2 pseudogenes and 1,373 liver genes. This large set of liver genes was used because the inference of ancestral traits based upon only five extant species is noisy, so we wanted to use a large sample size to estimate the null distribution. The distributions of changes in gene expression along each lineage were converted into standard scores based upon the mean and standard deviation (SD) of the inferred changes in 1,373 liver genes (mean changes of 0 from the rodent ancestor to both human and macaque). The expression level decreases from the rodent ancestor to human for both pseudogenes

relative to the 1,373 liver genes (changes of -3.26 and -4.09 for *Gulo* and *Uox*, respectively; Bonferroni corrected [BF corr.] combined $P < 10^{-6}$; supplementary fig. S2A, Supplementary Material online). In contrast, although *Gulo* expression has significantly decreased in macaque (change: -3.42 ; BF corr. $P < 10^{-4}$), *Uox* expression has not decreased more than expected by chance (change: -1.53 ; BF corr. $P = 0.16$; supplementary fig. S2B, Supplementary Material online). This is expected because *Uox* is still functional and is only expressed moderately lower in macaque compared with other mammals (Oda et al. 2002).

We used a similar strategy to test whether the apparent decreased number of binding events in primate pseudogenes (fig. 1) is statistically significant. We implemented a quantitative measure of overall binding level that takes into account binding intensity, distance from a locus' transcription start site (TSS), and the number of binding events (Wong et al. 2015; see Materials and Methods). Although TFs can regulate genes at a distance, here we are assuming that binding events closer to a gene's TSS are more likely to be regulating that gene. This measure of binding level can be considered a quantitative trait for which we have observations in five species. If relaxation of selection has led to a loss of binding after pseudogenization, we predict that this quantitative trait will decrease in the human lineage. To test this, we computed the difference in human binding level for pseudogenes relative to the rodent ancestor, as above. We find a significant decrease in binding level along the human lineage for both pseudogenes relative to the 1,373 liver genes (-1.13 and -2.22 for *Gulo* and *Uox*, respectively, compared with a mean of -0.147 for liver genes; BF corr. combined < 0.001 ; fig. 2A). The decrease in binding level in macaque is not significant for either *Gulo* or *Uox* (individual locus and combined $P > 0.05$; supplementary fig. S3A, Supplementary Material online).

Additionally, based on the same approach as above, we predicted that the enrichment levels for the histone modification markers H3K4me3 and H3K27Ac would also decrease. There is evidence for a decrease in the H3K4me3 enrichment level along the human lineage (-0.88 and -0.80 for *Gulo* and *Uox*, respectively, compared with a mean of 0.01 for liver genes; BF corr. combined $P < 0.05$; fig. 2B). However, similar to the above analyses on TF binding levels, there is no evidence for decreased H3K4me3 enrichment along the macaque lineage (individual locus and combined $P > 0.05$; supplementary fig. S3B, Supplementary Material online). Decreased H3K4me3 enrichment is consistent with loss of promoter activity, which is expected at these pseudogenes. Interestingly, there was no evidence for decreased enrichment of H3K27Ac nearby the pseudogenes along either primate branch (supplementary fig. S4, Supplementary Material online; individual locus and combined $P > 0.05$), which is consistent with no loss in enhancer activity nearby the pseudogenes.

We next sought to test for regulatory degeneration at the sequence level for *Uox*. Note that sequence-level analyses were not conducted for *Gulo* because this locus is a pseudogene in macaque and so there is no prediction of accelerated rate of evolution for this locus between the two primate

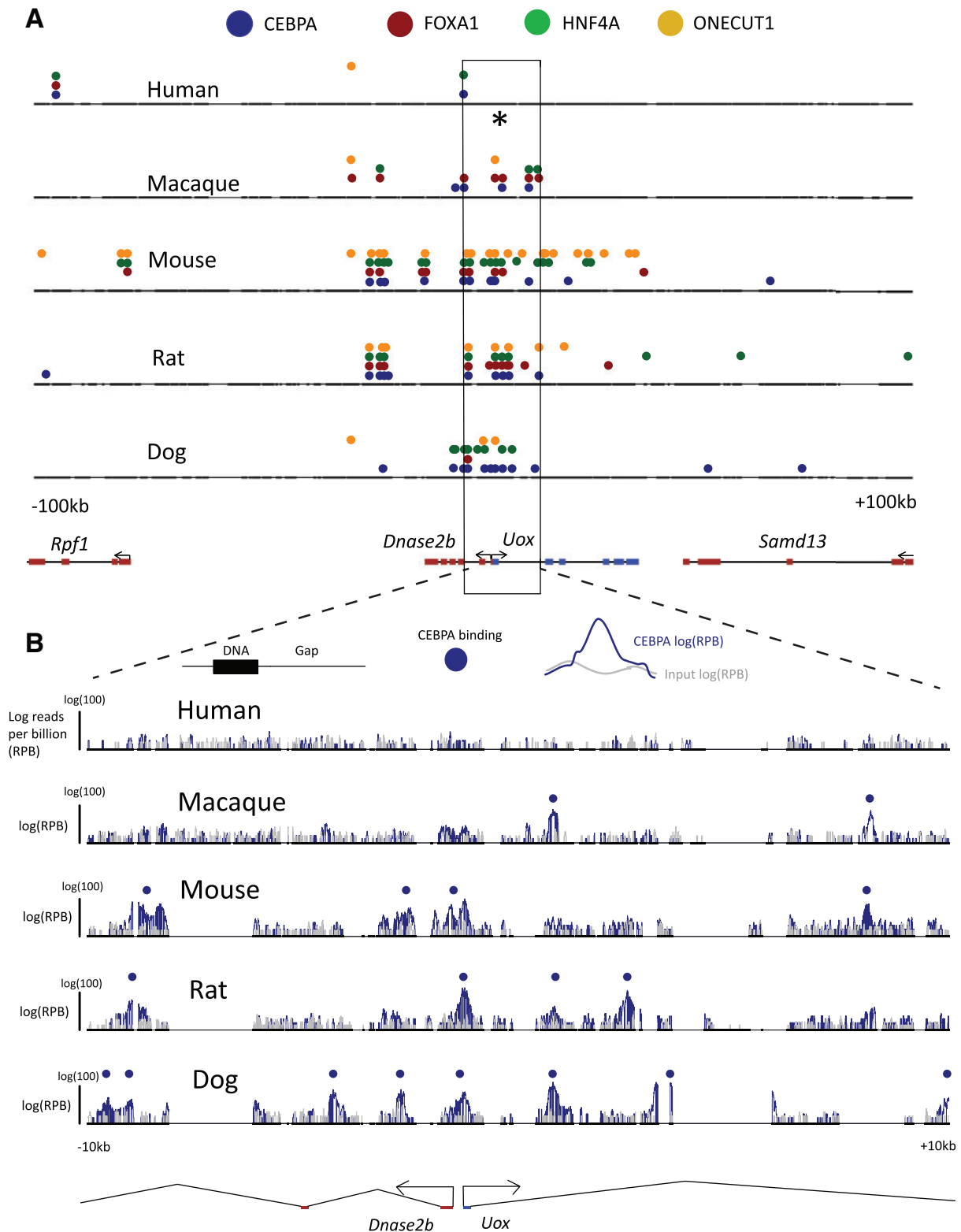


Fig. 1. Multi-species alignment around the *Uox* locus. (A) An alignment corresponding to ± 100 kb around the human *Uox* TSS (chr1: 84,763,514–84,963,514; negative strand). Thick black lines correspond to DNA for each species, while the thinner line indicates gaps. Different colored circles illustrate the binding of the four TFs around this locus (as described in the legend above). * indicates the *Uox* promoter in humans, where there is marked loss in binding compared with the other species. (B) ± 10 kb around the *Uox* TSS (indicated by box in panel A). The gray and blue histograms show the distributions of the input control and CEBPA reads per billion over the region in each species. The vertical black line at the leftmost side of each species name indicates log(100) reads per billion (RPB). Called CEBPA peaks in each species are indicated by blue circles. The *Uox* locus is colored blue at the bottom of the alignment and nearby genes are shown in red.

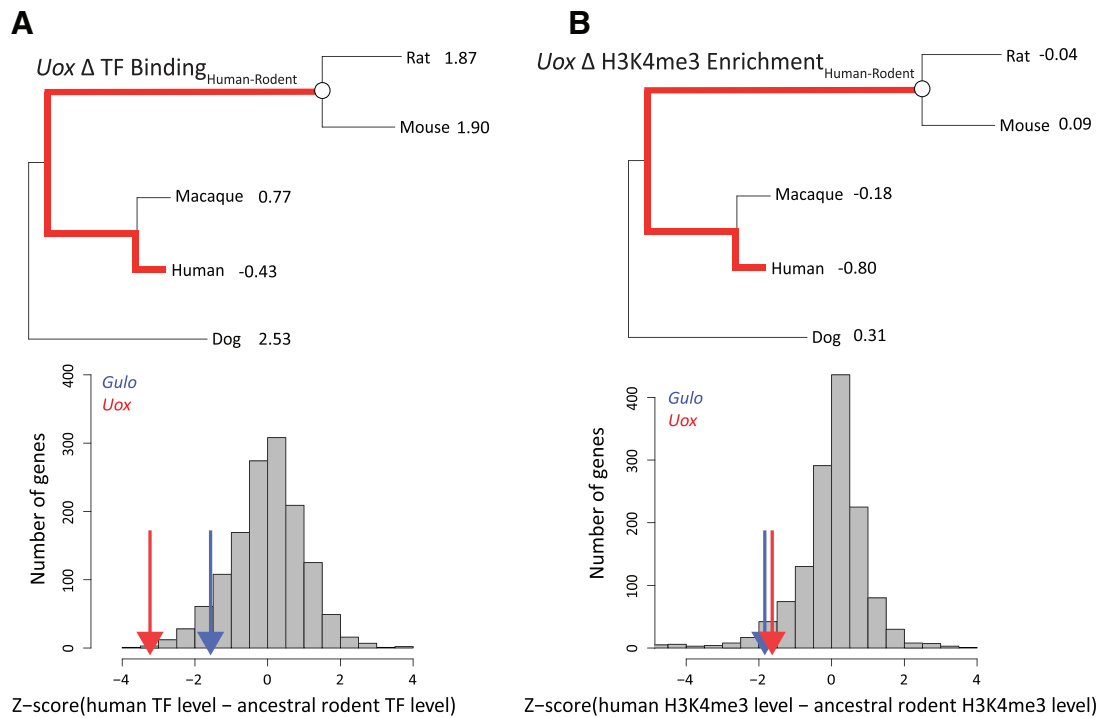


FIG. 2. Both human TF binding levels and H3K4me3 enrichment have decreased nearby primate pseudogenes. (A) The difference in TF binding level between the rodent ancestor and human. (B) The analogous difference in H3K4me3 enrichment (a marker for active promoters) between the rodent ancestor and human. The species trees show the inferred values of the quantitative traits. The white circles indicate the ancestral rodent and the red lines indicate the lineages being compared. Gray histograms correspond to the distribution of changes for 1,373 genes that are liver expressed in mouse, rat, and dog with one-to-one orthologs across the five mammals (in standardized units). *Gulo* and *Uox* are indicated in terms of their standard scores by the blue and red arrows, respectively. In human, where both *Gulo* and *Uox* are pseudogenized, there is a significant decrease in both TF binding and H3K4me3 enrichment levels for the pseudogenes (combined $P < 0.01$).

lineages; noncoding DNA from more distantly related mammals was not sufficiently alignable to confidently conduct sequence-based analysis. Specifically, we tested for lineage-specific accelerated evolution in CRMs nearby *Uox*, which we expected in analogy to coding region degeneration, where the nonsynonymous substitution rate (K_a) along pseudogenized lineages accelerates relative to the synonymous (near-neutral) substitution rate (K_s) (Torrents et al. 2003). We first confirmed that the pseudogene lineage-specific estimate of K_a/K_s for *Uox* was an outlier among the liver genes (fig. 3A). Along the pseudogenized lineage *Uox* has $K_a/K_s = 1.32$ while the $K_a/K_s = 0.649$ in macaque where *Uox* is still functional. The distribution of the differences in K_a/K_s between the two primate sets for each gene clearly shows that this difference for *Uox* is an extreme outlier (fig. 3A; 3.12 SDs). This is consistent with relaxed constraint on the *Uox* amino acid sequence.

Analogous tests for human-gibbon accelerated substitution rates were used on putative regulatory sequences (CRMs) within 100 kb of the *Uox* TSS. Here, regions bound by TFs were taken to be the sequences putatively under selection, while unbound, noncoding flanking sequences were taken to be the neutral reference. We defined the ratio of substitutions in these regions to be K_b/K_u ($b = \text{bound}$, $u = \text{unbound}$) in analogy to the K_a/K_s ratio above (Hahn 2007). Peak sequences nearby liver genes evolve more slowly than

unbound flanking regions, both along the human-gibbon lineage (supplementary fig. S5A, Supplementary Material online; median of 0.008 and 0.007 substitution per site in human flank and peak sequences; Wilcoxon test $P < 10^{-6}$) and in outgroup primates (supplementary fig. S5B, Supplementary Material online; median of 0.035 and 0.030 in macaque flank and peak sequences, respectively; Wilcoxon test BF corr. $P = 0.02$). This slower rate of evolution in peak sequences compared with flanks is consistent with purifying selection acting upon peak sequences and suggests that relaxed purifying selection should be detectable.

We inferred the number of substitutions per site along the human-gibbon lineage and macaque branch in peaks nearby the liver genes and plotted the difference in substitution rates inferred for each gene for all macaque peaks (fig. 3B). The difference in rates for *Uox* falls within these distributions (0.01 SD). Therefore, there is no evidence for relaxed purifying selection on peak sequences nearby *Uox* compared with those nearby functional liver genes. We also investigated whether there is a difference in the pattern of changes in binding strength in TFBSs (Moses 2009) nearby these two gene sets, but there was insufficient power at the single-gene level (supplementary fig. S6, Supplementary Material online).

To investigate whether any particular CRMs show evidence for degeneration nearby *Uox* (putative “pseudoenhancers”), we applied a likelihood ratio test (LRT) for

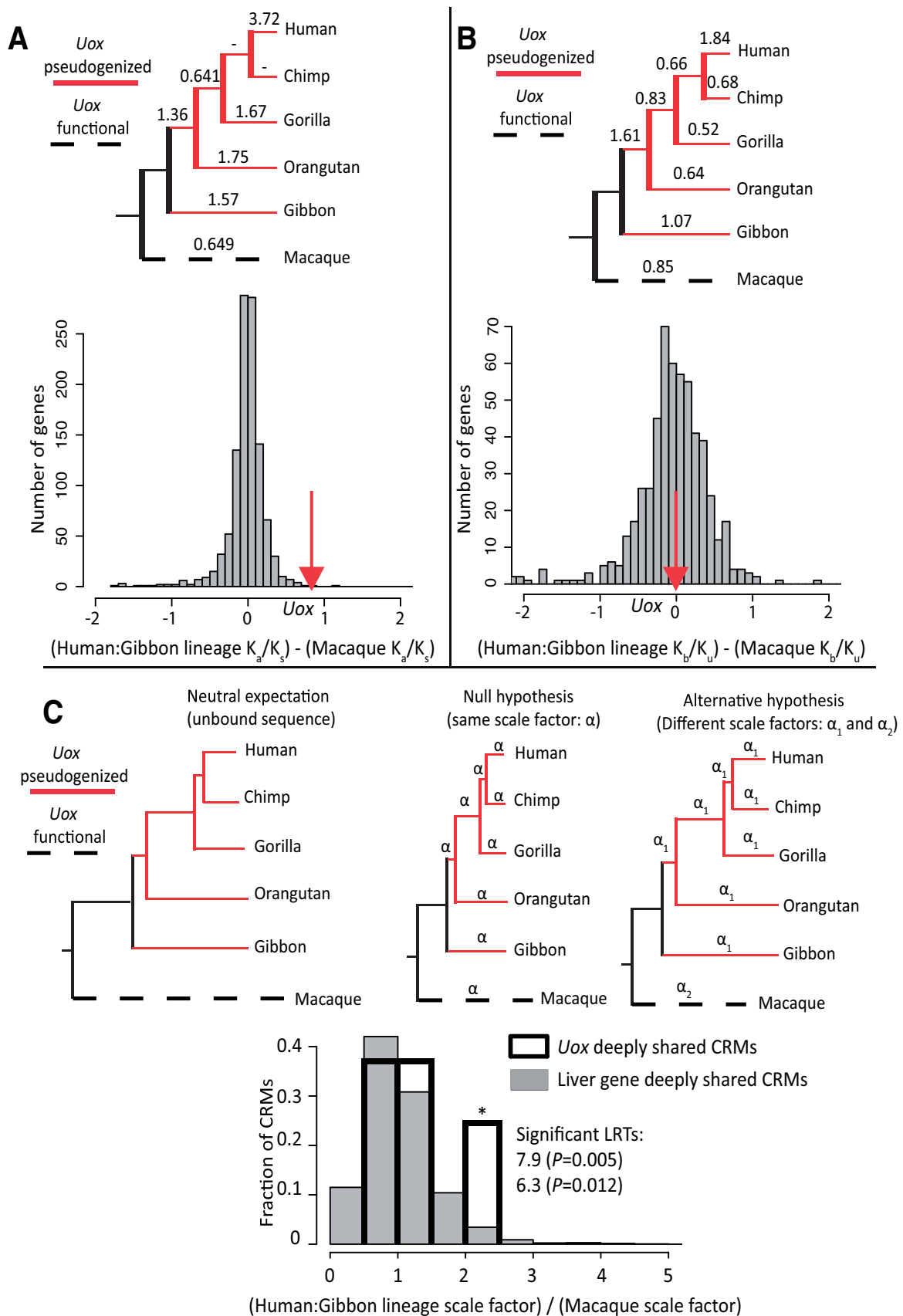


Fig. 3. Tests for sequence degeneration. Red branches and internodes on phylogenies indicate pseudogenized lineages, while the dotted lines indicate that *Uox* is functional in macaque. (A) Comparisons of the nonsynonymous substitution rate (K_a), normalized by the synonymous (near-neutral) substitution rate (K_s) along the human–gibbon lineage compared with macaque. The K_a/K_s is indicated along each considered primate lineage (- indicates insufficient data). The gray histogram shows the background distribution for the difference in K_a/K_s between the two sets of

(continued)

Table 1. The Eight Macaque-Bound CRMs Are Listed along with the Scale Factor Ratios (human-gibbon scale factor/macaque scale factor) and LRTs.

Macaque Coordinates	TF	Scale Factor Ratio	LRT	Raw P Value
chr1:87191849–87192049	FOXA1	2.205	7.935	0.005
chr1:87180804–87181004	ONECUT1	2.002	6.348	0.012
chr1:87187397–87187597	FOXA1	1.207	0.358	0.549
chr1:87199929–87200129	FOXA1	0.847	0.31	0.578
chr1:87219961–87220161	FOXA1	1.131	0.155	0.694
chr1:87204120–87204320	HNF4A	1.135	0.134	0.714
chr1:87189207–87189407	FOXA1	0.944	0.037	0.847
chr1:87271247–87271447	FOXA1	0.994	0	0.985

accelerated evolution along the human–gibbon lineage. This tests whether a scale factor (relative to a neutral reference) for the rate of evolution within the two primate sets is significantly higher along the human–gibbon lineage than over the rest of the primate phylogeny (fig. 3C; Cooper et al. 2005; Siepel et al. 2006). We identified candidate highly conserved CRMs nearby *Uox* by requiring macaque binding events to be shared with at least two other nonhuman mammals. Within 100 kb of the *Uox* TSS there are eight independent candidates for CRM degeneration based upon these criteria. Although these candidate CRMs were short (201 bp), simulations indicated that we had sufficient power to detect accelerated evolution along these lineages using the LRTs given reasonable rate increases (supplementary fig. S7 and supplementary text, Supplementary Material online). We applied the LRT to test for accelerated evolution along the human–gibbon lineage. We found that 2/8 of these candidate degenerative CRMs are evolving significantly faster along the human–gibbon lineage (fig. 3C; raw $P < 0.05$). These include multiple deeply shared binding events in the *Uox* promoter and a deeply shared ONECUT1 binding event in an intergenic region downstream of a neighboring gene (the coordinates and LRTs are provided in table 1, alignments are shown in supplementary figs. S8 and S9, Supplementary Material online). This result is consistent with relaxed selection specifically along the human–gibbon lineage in the *Uox* promoter and a putative enhancer nearby the *Uox* TSS. These significant scale factor ratios (fig. 3C) are within the range of values observed when the same tests were performed on peaks nearby the set of 1,373 liver genes.

We also tried to localize peaks that lost binding along pseudogenized lineages using the binding data alone. However, we were unable to do so because peaks with lost binding along pseudogenized lineages can be identified near functionally conserved liver genes. Thus, the multi-species ChIP-seq data are too variable to be used alone to identify

candidate pseudoenhancers. On the other hand, the quantitative decrease in TF binding (fig. 2A) suggests that at least some portion of TF binding events are preserved by purifying selection and that decreased binding levels reflect an increased proportion of TF binding events near pseudogenes being nonfunctional. We attempted to demonstrate the generality of our results on unitary pseudogenes by analyzing three other primate unitary pseudogenes (Zhang et al. 2010), but this resulted in mixed results, likely because functional orthologs of these pseudogenes are not highly expressed in the livers of all the other mammals considered (supplementary table S1, Supplementary Material online). Future work focusing on a larger set of pseudogenes could reveal whether decreased binding level is a general signature of regulatory degeneration. Although we did identify a candidate pseudoenhancer using sequence-based analyses (fig. 3C), the overall decrease in binding level is likely related to decreases in H3K4me3 enrichment levels (fig. 2B), and not changes in H3K27Ac enrichment levels (supplementary fig. S4, Supplementary Material online). This finding is consistent with degeneration at the pseudogene promoter, while the patterns in distal CRMs are more difficult to identify. Similar evidence for low levels of TF binding and histone modification levels has previously been shown nearby pseudogenes in human (Pei et al. 2012), but our work indicates that these observations are due to decreases in binding levels along the human lineage. Importantly, it is unclear how the decreases in TF binding, expression, and histone modification levels are linked.

Our evidence for accelerated evolution along the human–gibbon lineage in two regulatory elements (near the primate pseudogene *Uox*) is consistent with degeneration of regulatory elements in analogy to coding pseudogenization. However, we do not consider this strong evidence for pseudoenhancerization at the sequence level, because only one putative enhancer and the *Uox* promoter (as reported in Oda

Fig. 3. Continued

primates based upon the 1,373 liver genes. *Uox* is a clear outlier of this distribution (red arrow). (B) Comparisons of the bound sequence substitution rate (K_b) normalized by the unbound sequence substitution rate (K_u) along the human–gibbon lineage compared with macaque. The primate phylogeny indicates the K_b/K_u along each considered primate lineage. The histogram is analogous to panel A, except that it corresponds to K_b/K_u rather than K_a/K_s . *Uox* is not an outlier on this distribution of background differences. (C) LRT of accelerated evolution in CRMs specifically along the human–gibbon lineage. This is a test of whether the scale factor for the rate of evolution in a CRM is the same or different between the human–gibbon lineage and macaque. Scale factors are relative to the branch lengths of a neutral reference phylogeny based upon unbound, flanking DNA. The three phylogenies represent this neutral reference phylogeny, the null (same scaling factors) model, and the alternative (different scaling factor factor) models.

et al. 2002) were identified. In addition, there was no signal of relaxed selection along the human–gibbon lineage when all bound sequences nearby *Uox* were combined (fig. 3B). This result could suggest that a large proportion of bound sequences nearby gene sets are nonfunctional (Cuvanovich et al. 2014) or under very weak selection. However, it is also possible that CRM degeneration is masked by the background turnover in TFBSs (Stergachis et al. 2014; Vierstra et al. 2014). We believe that TFBS turnover could result in decreased power to detect accelerated evolution due to the rapid gain and loss of functional TFBSs, although we have not ruled out the possibility that there are too few sites considered to detect degeneration of enhancers. Additionally, the structure of CRMs, particularly enhancers, is more complex than has been assumed in the above analyses (Arnosti and Kulkarni 2005). It is not obvious that an accelerated rate of evolution necessarily occurs within degenerating enhancers in general. Alternatively, many regions of animal genomes have been identifiably bound by many different TFs simultaneously even though the known TFBSs of these TFs are not found (Moorman et al. 2006; ENCODE Project Consortium 2012), so it is not clear that regulatory degeneration could be identified at the sequence level in these cases. Taken together our analysis suggests that although pseudoenhancers may be identifiable in a small number of cases, decreased TF binding level over an entire locus could be a more reliable signature of regulatory degeneration.

Materials and Methods

Methods are described in detail in the [Supplementary Material](#) online, below is a brief summary.

We used publicly available ChIP-seq data produced from the livers of human, macaque, mouse, rat, and dog for four liver-enriched TFs: CEBPA, FOXA1, HNF4A, ONECUT1 (Schmidt et al. 2010; Funnell et al. 2013; Stefflova et al. 2013; Ballester et al. 2014) and the histone modification markers H3K4me3 and H3K27Ac (Villar et al. 2015). Reads were mapped to each reference genome with Bowtie2 (Langmead and Salzberg 2012) with default parameters. Numbers of mapped reads and quality metrics are shown in [supplementary table S2, Supplementary Material](#) online, for the liver-enriched TFs. Peak calling was performed with SWEMBL (v3.3.1; <https://www.ebi.ac.uk/~swilder/SWEMBL/>) with the parameter “-R 0.005.” TF peaks were called as reproducible in both replicates if they overlapped by $\geq 75\%$ reciprocally ([supplementary table S3, Supplementary Material](#) online).

Primate unitary pseudogenes were acquired from Zhang et al. (2010) and were screened by eye for whether they have liver expression in mouse, rat, and dog (based upon previously published RNA-seq data, see [Supplementary Material](#) online), as well as whether they have clear 1:1 orthologs based on Ensembl annotations and/or based on sequence similarity. The set of 1,373 liver expressed genes that are one-to-one orthologs across the 5 mammals was constructed based on a cut-off of ≥ 5 read fragments per kilobase of exon per million reads within the livers of mouse, rat, and dog.

The binding level for orthologs across species was based upon a previous approach (Wong et al. 2015). Before running these analyses, for each TF and species, read alignment files were downsampled to the number of reads mapped in the replicate with the fewest number of reads. This approach combines the number of binding events, their binding intensity, and their distance from the TSS into a single measure. Specifically, binding level (a_{is}), where i is an orthologous gene in species s , is given by

$$a_{is} = \log \left(\sum_j \sum_k \frac{g_{ijks}}{d_{ijks} + 0.1} \right)$$

where k is a peak within 100 kb of the TSS of gene i , g_{ijks} is the intensity of peak k for TF j in species s and d_{ijks} is the distance (in bp) from the TSS to the summit of peak k . The region size used is 100 kb and is discussed further elsewhere (Wong et al. 2015). A pseudocount of 0.1 is added to the denominator to ensure this value is never zero. The mean and variance in binding level was calculated for all TFs and species was determined for all liver expressed genes. Because the signal-to-noise ratio differs greatly across ChIP-seq experiments, binding levels were converted into standard scores (for each species and TF) so that they could be better compared across species. The standard scores of binding levels for each TF were then summed to obtain a single measure of binding level per gene. Maximum-likelihood ancestral binding levels were reconstructed under a BM model in R using the “ace” function of the ape package (Paradis et al. 2004). Enrichment levels for the histone modification markers were calculated using the same steps as for binding level.

All sequence analyses reported were performed on peak and flank coordinates sliced from the 38 eutherian EPO alignment (Release 74; Cunningham et al. 2015). Peak sequences were taken as ± 100 bp from peak summits, while flanking regions were taken as ± 2 kb from peak ends that did not intersect other peaks or exons. These alignments were realigned with MAFFT (L-INS-i method; v6.882b; Katoh and Toh 2008). Shared peaks across species were called based upon overlapping summits (taken as the peak center) within 150 bp of each other within the alignment. Primate sequences within the human–tarsier lineage were parsed from these alignments. Ancestral sequences at each node in the primate phylogeny were reconstructed using the “prequel” program of PHAST (v1.3; Hubisz et al. 2011), which were used to call substitutions. LRTs were performed with the “phyloP” program of PHAST to test for accelerated evolution within the human–gibbon lineage relative to the macaque branch. Alignments are available for download at <http://www.mose.slab.csb.utoronto.ca/gavin/alignments/>.

Supplementary Material

[Supplementary tables S1–S3, figures S1–S9](#), and text are available at *Molecular Biology and Evolution* online (<http://www.mbe.oxfordjournals.org/>).

Acknowledgments

This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) through grants to A.M.M. and a fellowship to G.M.D. M.D.W. was funded by NSERC (436194-2013) and a Canada Research Chair. We thank Christopher Brown and an anonymous reviewer for suggesting the analysis of chromatin modification data.

References

- Arnosti DN, Kulkarni MM. 2005. Transcriptional enhancers: intelligent enhanceosomes or flexible billboards? *J Cell Biochem*. 94:890–898.
- Ballester B, Medina-Rivera A, Schmidt D, González-Porta M, Carlucci M, Chen X, Chessman K, Faure AJ, Funnell AP, Goncalves A, et al. 2014. Multi-species, multi-transcription factor binding highlights conserved control of tissue-specific biological pathways. *Elife* 3:e02626.
- Britten RJ, Davidson EH. 1969. Gene regulation for higher cells: a theory. *Science* 165:349–357.
- Cooper GM, Stone EA, Asimenos G, Green ED, Batzoglou S, Sidow A. 2005. Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res*. 15:901–913.
- Creyghton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, Steine EJ, Hanna J, Lodato MA, Frampton GM, Sharp PA, et al. 2010. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci U S A*. 107:21931–21936.
- Cunningham F, Amode MR, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fitzgerald S, et al. 2015. Ensembl 2015. *Nucleic Acids Res*. 43:D662–D669.
- Cusanovich DA, Pavlovic B, Pritchard JK, Gilad Y. 2014. The functional consequences of variation in transcription factor binding. *PLoS Genet* 10:e1004226.
- Drouin G, Godin JR, Pagé B. 2011. The genetics of vitamin C loss in vertebrates. *Curr Genomics*. 12:371–378.
- ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* 489:57–74.
- Funnell APW, Wilson MD, Ballester B, Mak KS, Burdach J, Magan N, Pearson RCM, Lemaigre FP, Stowell KM, Odom DT, et al. 2013. A CpG mutational hotspot in a ONECUT binding site accounts for the prevalent variant of hemophilia B Leyden. *Am J Hum Genet*. 92:460–467.
- Gustafsson D, Unwin R. 2013. The pathophysiology of hyperuricaemia and its possible relationship to cardiovascular disease, morbidity and mortality. *BMC Nephrol*. 14:164.
- Hahn MW. 2007. Detecting natural selection on *cis*-regulatory DNA. *Genetica* 129:7–18.
- Hubisz MJ, Pollard KS, Siepel A. 2011. Phast and Rphast: phylogenetic analysis with space/time models. *Brief Bioinform*. 12:41–51.
- Katoh K, Toh H. 2008. Recent developments in the MAFFT multiple sequence alignment program. *Brief Bioinform*. 9:286–298.
- King M, Wilson AC. 1975. Evolution at two levels in humans and chimpanzees. *Science* 188:107–116.
- Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods*. 9:357–359.
- Leivonen S, Piao YS, Peltoketo H, Numchaisrika P, Vihko R, Vihko P. 1999. Identification of essential subelements in the *hHSD17B1* enhancer: difference in function of the enhancer and that of the *hHSD17BP1* analog is due to -480C and -486G. *Endocrinology* 140:3478–3487.
- McLean CY, Reno PL, Pollen AA, Bassan AI, Capellini TD, Guenther C, Indjeian VB, Lim X, Menke DB, Schaar BT, et al. 2011. Human-specific loss of regulatory DNA and the evolution of human-specific traits. *Nature* 471:216–219.
- Moorman C, Sun LV, Wang J, de Wit E, Talhout W, Ward LD, Greil F, Lu XJ, White KP, Bussemaker HJ, et al. 2006. Hotspots of transcription factor colocalization in the genome of *Drosophila melanogaster*. *Proc Natl Acad Sci USA* 103:12027–12032.
- Moses AM. 2009. Statistical tests for natural selection on regulatory regions based on the strength of transcription factor binding sites. *BMC Evol Biol*. 9:286.
- Oda M, Satta Y, Takenaka O, Takahata N. 2002. Loss of urate oxidase activity in hominoids and its evolutionary implications. *Mol Biol Evol*. 19:640–653.
- Paradis E, Claude J, Strimmer K. 2004. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* 20:289–290.
- Pei B, Sisu C, Frankish A, Howald C, Habegger L, Mu XJ, Harte R, Balasubramanian S, Tanzer A, Diekhans M, et al. 2012. The GENCODE pseudogene resource. *Genome Biol*. 13:R51.
- Santos-Rosa H, Schneider R, Bannister AJ, Sherriff J, Bernstein BE, Emre NC, Schreiber SL, Mellor J, Kouzarides T. 2002. Active genes are trimethylated at K4 of histone H3. *Nature* 419:407–411.
- Schmidt D, Wilson MD, Ballester B, Schwalie PC, Brown GD, Marshall A, Kutter C, Watt S, Martinez-jimenez CP, Mackay S, et al. 2010. Five-vertebrate ChIP-seq reveals transcription factor binding. *Science* 328:1036–1040.
- Siepel A, Pollard KS, Haussler D. 2006. New methods for detecting lineage-specific selection. *Res Comput Mol Biol*. 3909:190–205.
- Stefflova K, Thybert D, Wilson MD, Streeter I, Aleksic J, Karagianni P, Brazma A, Adams DJ, Talianidis I, Marioni JC, et al. 2013. Cooperativity and rapid evolution of cobound transcription factors in closely related mammals. *Cell* 154:530–540.
- Stergachis AB, Neph S, Sandstrom R, Haugen E, Reynolds AP, Zhang M, Byron R, Canfield T, Stelling-Sun S, Lee K, et al. 2014. Conservation of *trans*-acting circuitry during mammalian regulatory evolution. *Nature* 515:365–370.
- Torrents D, Suyama M, Zdobnov E, Bork P. 2003. A genome-wide survey of human pseudogenes. *Genome Res*. 7:2559–2567.
- Vierstra J, Rynes E, Sandstrom R, Zhang M, Canfield T, Hansen RS, Stelling-Sun S, Sabo PJ, Byron R, Humbert R, et al. 2014. Mouse regulatory DNA landscapes reveal global principles of *cis*-regulatory evolution. *Science* 346:1007–1012.
- Villar D, Berthelot C, Flicek P, Odom DT, Villar D, Berthelot C, Aldridge S, Rayner TF, Lukk M, Pignatelli M. 2015. Enhancer evolution across 20 mammalian species. *Cell* 160:554–566.
- Wong ES, Thybert D, Schmitt BM, Stefflova K, Odom DT, Flicek P. 2015. Decoupling of evolutionary changes in transcription factor binding and gene expression in mammals. *Genome Res*. 25:167–178.
- Wu XW, Muzny DM, Lee CC, Caskey CT. 1992. Two independent mutational events in the loss of urate oxidase during hominoid evolution. *J Mol Evol*. 34:78–84.
- Zhang ZD, Frankish A, Hunt T, Harrow J, Gerstein M. 2010. Identification and analysis of unitary pseudogenes: historic and contemporary gene losses in humans and other primates. *Genome Biol*. 11:R26.