BIOINFORMATICS DISCOVERY NOTE Vol. 24 no. 14 2008, pages 1563–1567
doi:10.1093/bioinformatics/btn243

Genome analysis

Retroviral promoters in the human genome

Andrew B. Conley, Jittima Piriyapongsa and I. King Jordan*
School of Biology, Georgia Institute of Technology, 310 Ferst Drive, Atlanta, GA 30306, USA

Received on February 7, 2008; revised on May 19, 2008; accepted on May 20, 2008

Associate Editor: John Quackenbush

ABSTRACT

Motivation: Endogenous retrovirus (ERV) elements have been shown to contribute promoter sequences that can initiate transcription of adjacent human genes. However, the extent to which retroviral sequences initiate transcription within the human genome is currently unknown. We analyzed genome sequence and high-throughput expression data to systematically evaluate the presence of retroviral promoters in the human genome.

Results: We report the existence of 51 197 ERV-derived promoter sequences that initiate transcription within the human genome, including 1743 cases where transcription is initiated from ERV sequences that are located in gene proximal promoter or 5′ untranslated regions (UTRs). A total of 114 of the ERV-derived transcription start sites can be demonstrated to drive transcription of 97 human genes, producing chimeric transcripts that are initiated within ERV long terminal repeat (LTR) sequences and read-through into known gene sequences. ERV promoters drive tissue-specific and lineage-specific patterns of gene expression and contribute to expression divergence between paralogs. These data illustrate the potential of retroviral sequences to regulate human transcription on a large scale consistent with a substantial effect of ERVs on the function and evolution of the human genome.

Contact: king.jordan@biology.gatech.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Approximately 5% of the human genome sequence is derived from retroviruses (Lander *et al*., 2001). Retroviral genomic sequences are remnants of past infections that resulted in the integration of provirus genomes into the DNA of germline cells (Bock *et al*., 2000; Bromham, 2002). The abundance of these so-called endogenous retrovirus sequences (ERVs) testifies to the extent that human evolution has been shaped by successive waves of viral invasion (Sverdlov, 2000).

One way that ERVs have affected the function and evolution of the human genome is by donating regulatory sequences that control the expression of nearby genes. The gene regulatory effects of ERVs were first uncovered in a number of anecdotal studies on specific genes (reviewed in Bannert *et al*., 2004; Medstrand *et al*., 2005). For instance, the long terminal repeat (LTR) of a human ERV (HERV-E) was shown to serve as an enhancer element that confers parotid-specific expression on the amylase gene (Samuelson *et al*., 1990). Later, more systematic computational analyses of the human genome sequence revealed that many human genes contained ERV-derived regulatory regions, suggesting an even greater contribution of retroviruses to human gene regulation (Jordan *et al*., 2003; van de Lagemaat *et al*., 2003). Continued efforts to characterize ERV-derived promoters have turned up several new cases in recent years (Dunn *et al*., 2003, 2006; Romanish *et al*., 2007). Nevertheless, the full extent of the contribution of ERV sequences to the initiation of transcription in the human genome has yet to be appreciated.

Initiation of transcription by ERV promoters often results in the production of alternative transcripts that are both tissue-specific and lineage-specific. For instance, testis-specific expression of the human gene encoding the neuronal apoptosis inhibitory protein (NAIP) is driven by an LTR promoter sequence, whereas a distinct LTR promoter in rodents confers constitutive expression of the orthologous gene (Romanish *et al*., 2007). An ERV LTR sequence also serves as an alternative promoter that drives expression of the beta1,3-galactosyltransferase five gene specifically in colorectal tissue (Dunn *et al*., 2003).

The lineage-specific regulatory effects of ERV promoters can be attributed to the fact that ERV sequences result from past germline infections, many of which occurred relatively recently along specific evolutionary lineages. In fact, most of the ERV sequences in the human genome are primate-specific (Sverdlov, 2000), while most human genes are far more ancient and share orthologs with distantly related species (Lander *et al*., 2001). This means that regulatory effects exerted by ERV promoters will often lead to expression differences between primate and non-primate orthologs or between deeper evolutionary lineages for more ancient ERVs. In other words, ERV promoters are likely to drive evolutionary changes in gene expression, long thought to be an important determinant of species divergence (King *et al*., 1975).

The application of novel high-throughput techniques for the analysis of gene expression has revolutionized the study of the human transcriptome and revealed far more regulatory complexity than previously imagined. Two techniques in particular, cap analysis of gene expression (CAGE) and paired-end ditag (PET) sequencing, enable the precise genome mapping of many thousands of promoter sequences that initiate transcription. CAGE is a technique that allows for the characterization of short sequence tags from the 5′-most ends of full-length cDNAs (Shiraki *et al*., 2003). Accordingly, mapping CAGE tags to the human genome unambiguously identifies transcription start sites (TSS) and their corresponding promoters. PET sequencing involves the determination of sequences for tags from both the 5′ and 3′ ends of full-length cDNAs (Ng *et al*., 2005). Thus, when PETs are mapped to the genome,

*To whom correspondence should be addressed.

paired transcriptional initiation and termination sites are identified along with the intervening genomic sequences that are transcribed as pre-mRNAs. We used human CAGE and PET data to more thoroughly evaluate the contribution of ERVs to the initiation of transcription in the human genome.

2 METHODS

CAGE tags (n = 1 551 672) were downloaded from the Japanese National Institute of Genetics website(http://genomenetwork.nig. ac.jp/public/ download/ cage_Database_e.html) and mapped to the human genome as previously described (Shiraki *et al*., 2003). The human genome locations of PETs (n = 669 840) were taken from the UCSC Genome Browser (Karolchik *et al*., 2003) annotations (http://www.genome.ucsc. edu/cgi-bin/ hgTrackUi?hgsid=100351785&c=chr9&g=wgEncodeGisRnaPet). The PETs were generated from several cell lines: log phase of MCF7 cells (113 858), MCF7 cells treated with estrogen (4911), HCT116 cells treated with 5-fluorocil (124 770) and log phase of embryonic stem cell hES3 (426 301). Overlapping CAGE tags and overlapping PETs were clustered to identify individual TSS on the human genome. The UCSC Table Browser (Karolchik *et al*., 2004) and the program Galaxy (Giardine *et al*., 2005) were used to compare the locations of CAGE tags and PETs to the locations of human ERVs annotated with the RepeatMasker program (Smit *et al*., 1996–2004). Only ERVs *sensu strictu*, as opposed to more ancient mammalian apparent LTR-retrotransposons (MaLR), were analyzed here. The National Center for Biotechnology (NCBI) Refseq (Pruitt *et al*., 2007) gene model annotations were used to evaluate the production of chimeric transcripts that are initiated by ERVs and read-through into human genes. Transcriptional units (TUs) are defined as genomic regions spanning the 5′ to the 3′ ends of individual Refseq gene models. TUs and 1 kb flanking regions upstream and downstream of TUs were evaluated for the presence of ERV-derived promoters. A series of custom Perl scripts were used to post-process the genome mapping data and to produce browser extendable data (BED) mapping tracks for further analysis with the UCSC Genome Browser. All scripts and mapping data are available upon request.

The genomic presence/absence of ERV insertions across species was evaluated using whole genome sequence alignments of complete mammalian sequences built with the Multiz tool (Blanchette *et al*., 2004). Human genome sequence conservation levels are based on the phastCons tool (Siepel *et al*., 2005). The species distribution of human gene orthologs was assessed using BLASTP (Altschul *et al*., 1997) results from the NCBI Blink utility along with homology annotations from the GeneCards web server (Safran *et al*., 2002). Gene expression analysis was based on the Novartis Gene Expression Atlas version 2 (GNF2) (Su *et al*., 2004).

Detailed information on all methods including PET and CAGE analysis along with gene expression and Gene Ontology (GO) analyses can be found in the Supplementary Material.

3 RESULTS AND DISCUSSION

A total of 49 814 mapped CAGE tag clusters, each corresponding to an individual TSS, were found to map to the ERV LTR sequences (Table 1). The high number of ERV-derived TSS in the human genome identified with CAGE tag mapping underscores the potential of retroviral promoters to drive transcription. However, it is not possible to directly assess whether retroviral promoters identified using CAGE tag mapping actually drive the expression of known human genes. In fact, most of the ERV promoters identified with CAGE map to intergenic regions. This intergenic ERV promoter activity is likely to be a relic of the ERVs' ability to drive transcription of their own genome sequences from LTR promoters and may not necessarily be related to the transcription of human genes. Nevertheless, the presence of widespread ERV promoter activity in the human genome demonstrates that ERV sequences can maintain the ability to promote transcription for millions of years after their initial insertion into the genome.

Table 1. Numbers of ERV-derived TSS in the human genome

Data source	Total TSS[a]	Gene-associated TSS[b]
CAGE	49 814	9292
PET	1513	114

[a] Total number of tag clusters representing individual ERV-derived TSS.
[b] For CAGE data, ERVTSS that map within 1 kb upstream or downstream of Refseq gene annotated 5′ UTR sequences. For PET data, ERV-ditag sequence clusters that start within 1 kb of Refseq gene annotated 5′ UTR sequences, or within 5′ UTRs, and end within Refseq gene TUs, 3′UTRs or 1 kb downstream of 3′ UTRs.

Table 2. Numbers of ERV-human gene associated or chimeric transcripts

CAGE[a]			
Total	Upstream	5′ UTR	TU
9292	193	1550	7549
PET[b]			
PET 3′ ends	PET 5′ ends		
	Upstream	5′ UTR	TU
TU	5	6	34
3′ UTR	12	13	21
Downstream	4	8	11

[a] Counts for ERV-derived CAGE sequence tag clusters that map within human Refseq gene 5′ UTRs or 1 kb upstream or downstream (i.e. within the TU) of the 5′ UTR.
[b] Counts for ERV-derived PET sequence clusters associated with human genes are shown. ERV-PETs with 5′ ends that are 1 kb upstream of human Refseq gene 5′ UTRs, within 5′ UTRs or within TUs are shown in columns. ERV-PETs with 3′ ends that are within TUs, in 3′ UTRs or 1 kb downstream of 3′ UTRs are shown in rows.

In addition to the intergenic ERV promoters, there are 9292 CAGE identified ERV promoters that initiate transcription within 1 kb upstream or downstream of the previously characterized TSS of known human genes (Table 2). 1550 of these ERV CAGE tag clustersmap to 5′ UTRs, consistent with transcription from previously characterized promoters, but the majority (7742) map just upstream of the 5′ UTR, in the proximal promoter region, or downstream within genes' TUs. Therefore, these ERV-derived promoters are likely to be responsible for generating alternative transcripts of human genes.

PET sequence mapping data were also used to search for transcripts that are initiated from ERV promoters, and there are 1513 cases of PET identified ERV promoters in the human genome (Table 1). Because PET sequence tags include both the 5′ and 3′ ends of full-length transcripts, they can be used to identify transcripts that are initiated within ERV sequences and read-through into human gene regions. These cases correspond to chimeric transcripts, composed partially of both ERV and human gene sequences, and demonstrate ERV promoted expression of human genes. This approach identified 114 distinct retroviral TSS that promote transcription of human genes (Table 2 and

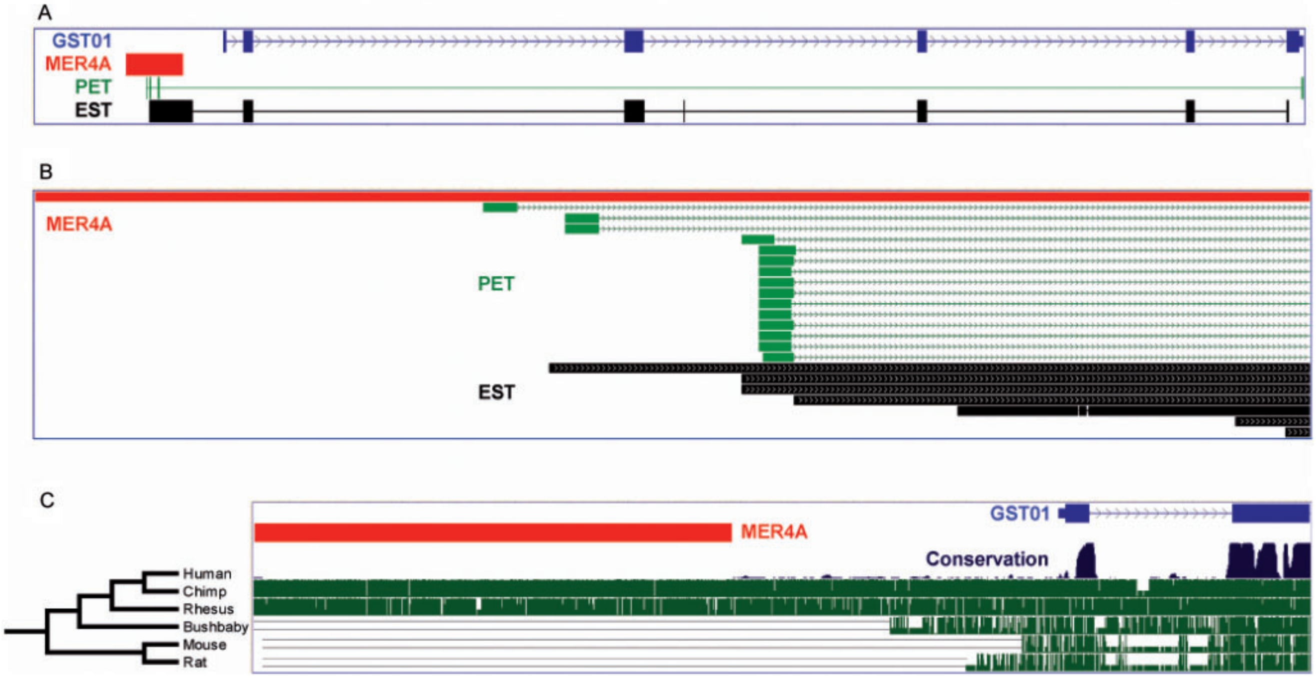

Fig. 1. MER4A alternative promoter of the *GSTO1* gene. (**A**) TheMER4A (red) ERV sequence is located in the proximal promoter region <500 bp from the *GSTO1* 5′ UTR. The locations of PET sequences (green) and spliced ESTs (black) are shown. (**B**) The MER4A (red) sequence region is enlarged and the individual PET sequences (green) and spliced ESTs (black) that support the existence of this promoter are shown. (**C**) Evolutionary conservation of MER4A versus *GSTO1*. MER4A is only found in chimp and rhesus and no other mammals (green bars), i.e. it is not conserved, whereas the adjacent *GSTO1* exons are conserved across mammalian species (green and blue bars).

Supplementary Table 1). Twenty-one of these retroviral promoters have colocated ESTs, which independently support their ability to initiate transcription. These retroviral TSS correspond to 124 Refseq transcripts over 97 distinct gene loci. The positions of TSS for ERV-derived human gene promoters were analyzed to evaluate whether ERVs provide canonical promoters or promote alternative transcripts. While there are a number of ERV TSS that map to 5′ UTRs (Table 2 and Supplementary Figure 1), and are thus taken to promote transcription at (or near) previously characterized TSS, the majority of ERV promoters promote alternative transcription of human genes from upstream regions or from within the TU (Table 2). This further underscores the fact that ERVs promote alternative transcription of human genes.

The ability of ERVs to promote alternative transcripts of human genes is illustrated (Fig. 1) by the case of an alternative promoter of the glutathione-s-transferase omega 1 encoding gene (*GSTO1* Refseq accession NM_004832) found on chromosome 10q25.1. The GSTO1 protein is a member of the theta class glutathione s-transferase-like family, and it has been shown to act as a stress response protein through cellular redox homeostasis (Kodym *et al*., 1999). *GSTO1* nucleotide polymorphisms have been implicated in a number of cerebrovascular diseases including Alzheimer's disease, Parkinson's disease, vascular dementia and stroke (Kolsch *et al*., 2004; Li *et al*., 2003).

There is an ERV LTR sequence from the MER4A subfamily of sequences <500 bp upstream of the Refseq annotated 5′ UTR of *GSTO1* (Fig. 1A). There are 15 individual PET sequences, forming three distinct TSS clusters, that have 5′ ends inside of the MER4A sequence and 3′ ends in the 3′ UTR of *GSTO1* (Fig. 1A and B). All of the MER4A PET sequences were derived from only one of the four PET libraries ($\chi^2 = 8.6$ $P = 0.04$), log phase of embryonic stem cell hES3, indicating that this promoter is tissue- or condition-specific. In addition to the PET sequence-based evidence, there are a number of spliced expressed sequence tags (ESTs) that also indicate the MER4A sequence as an alternative promoter for *GSTO1* (Fig. 1B).

Inspection of multiple sequence alignments of complete mammalian genomes reveals that the *GSTO1* adjacent MER4A insertion is present in the human, chimp (*Pan troglodytes*) and rhesus (*Macaca mulatta*) genome sequences but absent in the bushbaby (*Otolemur garnetti*), mouse (*Mus musculus*), rat (*Rattus norvegicus*) and all other placental mammal sequences (Fig. 1C). In other words, that particular MER4A sequence inserted after the primate radiation began, and it is specific to the Haplorrhini suborder, which includes both new world and old world monkeys. On the other hand, the adjacent exonic sequences of *GSTO1* show marked conservation compared to MER4A (Fig. 1C). Comparative sequence analysis with BLASTP indicates that *GSTO1* is far more ancient than the MER4A insertion, having well-conserved orthologs among mammals and other vertebrates along with *Drosophila melanogaster*, *Caenorhabditis elegans* and a number of other more distantly related species.

The comparative sequence analysis suggests the possibility that the MER4A insertion may confer lineage-specific expression pattern on *GSTO1*. Furthermore, there are two human paralogs of *GSTO1*, *GSTO2* and the pseudogene *GSTO3P1*, neither of which has the upstream MER4A insertion. So the specific regulatory effects of the ERV may not only be tissue- and lineage-specific but could also be involved in driving functional differentiation of paralogs via expression differences.

In order to test for potential diversifying regulatory effects of the MER4A insertion on *GSTO1*, we compared tissue-specific expression patterns between human and mouse *GSTO1* and *GSTO2* orthologs as well as between human *GSTO1* and *GSTO2* paralogs using microarray data from the Novartis Gene Expression Atlas version 2 (GNF2) (Su *et al*., 2004). The human–mouse *GSTO1* orthologous pair has a low ($r = -0.06$), which is not significantly different from 0 ($P = 0.77$) and correlation of expression levels across tissues as does the human *GSTO1–GSTO2* paralogous pair ($r = 0.006$; $P = 0.98$) (Supplementary Fig. 2). On the other hand, the human and mouse *GSTO2* orthologous genes, which lack the alternative MER4A promoter, have significantly correlated expression patterns ($r = 0.76$; $P = 2.2\text{e-}6$). These patterns of expression divergence and conservation are consistent with variation in expression introduced by the MER4A-TSS. In all, 37 out of 40 evaluated cases of human genes with ERV-TSS have expression patterns that are not significantly correlated with their mouse orthologs that lack the upstream ERV (Supplementary Fig. 3).

We further evaluated the potential regulatory effects of human ERV-derived promoters by comparing the expression patterns of all human genes with ERV promoters versus genes without ERV promoters using the GNF2 data. Human genes with ERV-TSS have greater tissue specificity than genes lacking ERV promoters, consistent with a diversifying regulatory effect of ERV-TSS (Supplementary Table 2). In particular, ERV-TSS containing genes have anomalously high levels of expression, on average, in brain and testis (Supplementary Figs 4 and 5). A similar pattern of significantly elevated expression in brain and testis was found for ERV CAGE tags (Supplementary Fig. 6). Consistent with the brain-specific expression pattern of ERV-TSS genes, GO functional analysis indicated that these genes are enriched for metabolic and signaling processes active in the brain (Supplementary Table 3 and Supplementary Fig. 7).

Our analysis revealed that retroviral sequences in the human genome encode tens-of-thousands of active promoters; transcribed ERV sequences correspond to 1.16% of the human genome sequence and PET tags that capture transcripts initiated from ERVs cover 22.4% of the genome. These data suggest that ERVs may regulate human transcription on a large scale. However, it is a formal possibility that many of the ERV derived promoters identified here represent leaky transcription, i.e. noise, which is not functionally significant. Definitive proof of biological activity for individual ERV-TSS may have to await experimental confirmation via knock-out data or promoter swapping. However, it will soon be possible to validate ERV-TSS on a genome-scale owing to the accumulation of high-throughput data from tiling array experiments based on ChIP-chip and/or chromatin structure assays. Such data, which are being generated by the ENCODE Project Consortium (2007), measure the distributions of regulatory signatures across genomic sequence. The presence and density of regulatory signals, such as transcription factor binding sites and open or specifically modified chromatin, have been shown to discriminate between biologically active and artifactual TSS and thus could be used to validate ERV-TSS.

Our analysis uncovered more than 100 cases of novel ERV-derived promoters that initiate chimeric ERV-human gene transcripts and several thousand more that are likely to do so. ERV-derived promoters are characterized by their ability to promote alternative transcripts that are expressed in a way that is tissue-specific, lineage-specific and distinct from related paralogous genes. These data underscore the extent to which retrovirus activity has shaped the human transcriptome.

ACKNOWLEDGEMENTS

Funding: A.B.C. and I.K.J. are supported by the School of Biology, Georgia Institute of Technology. J.P. is supported by the Ministry of Science and Technology of Thailand.

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. *et al*. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*., **25**, 3389–3402.

Bannert,N. *et al*. (2004) Retroelements and the human genome: new perspectives on an old relation. *Proc. Natl Acad. Sci. USA*, **101** (Suppl. 2), 14572–14579.

Blanchette,M. *et al*. (2004) Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res*., **14**, 708–715.

Bock,M. *et al*. (2000) Endogenous retroviruses and the human germline. *Curr. Opin. Genet. Dev*., **10**, 651–655.

Bromham,L. (2002) The human zoo: endogenous retroviruses in the human genome. *Trends Ecol. Evol*., **17**, 91–97.

Dunn,C.A. *et al*. (2003) An endogenous retroviral long terminal repeat is the dominant promoter for human beta1,3-galactosyltransferase 5 in the colon. *Proc. Natl Acad. Sci. USA*, **100**, 12841–12846.

Dunn,C.A. *et al*. (2006) Transcription of two human genes from a bidirectional endogenous retrovirus promoter. *Gene*, **366**, 335–342.

ENCODE Project Consortium (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE project. *Nature*, **447**, 799–816.

Giardine,B. *et al*. (2005) Galaxy: a platform for interactive large-scale genome analysis. *Genome Res*., **15**, 1451–1455.

Jordan,I.K. *et al*. (2003) Origin of a substantial fraction of human regulatory sequences from transposable elements. *Trends Genet*., **19**, 68–72.

Karolchik,D. *et al*. (2003) The UCSC genome browser database. *Nucleic Acids Res*., **31**, 51–54.

Karolchik,D. *et al*. (2004) The UCSC table browser data retrieval tool. *Nucleic Acids Res*., **32**, D493–D496.

King,M.C. *et al*. (1975) Evolution at two levels in humans and chimpanzees. *Science*, **188**, 107–116.

Kodym,R. *et al*. (1999) The cloning and characterization of a new stress response protein. A mammalian member of a family of theta class glutathione s-transferase-like proteins. *J. Biol. Chem*., **274**, 5131–5137.

Kolsch,H. *et al*. (2004) Polymorphisms in glutathione S-transferase omega-1 and AD, vascular dementia, and stroke. *Neurology*, **63**, 2255–2260.

Lander,E.S. *et al*. (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.

Li,Y.J. *et al*. (2003) Glutathione S-transferase omega-1 modifies age-at-onset of Alzheimer disease and Parkinson disease. *Hum. Mol. Genet*., **12**, 3259–3267.

Medstrand,P. *et al*. (2005) Impact of transposable elements on the evolution of mammalian gene regulation. *Cytogenet. Genome Res*., **110**, 342–352.

Ng,P. *et al*. (2005) Gene identification signature (GIS) analysis for transcriptome characterization and genome annotation. *Nat. Methods*, **2**, 105–111.

Pruitt,K.D. *et al*. (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res*., **35**, D61–D65.

Romanish,M.T. *et al*. (2007) Repeated recruitment of LTR retrotransposons as promoters by the anti-apoptotic locus NAIP during mammalian evolution. *PLoS Genet*., **3**, e10.

Safran,M. *et al*. (2002) GeneCards 2002: towards a complete, object-oriented, human gene compendium. *Bioinformatics*, **18**, 1542–1543.

Samuelson,L.C. *et al*. (1990) Retroviral and pseudogene insertion sites reveal the lineage of human salivary and pancreatic amylase genes from a single gene during primate evolution. *Mol. Cell. Biol.*, **10**, 2513–2520.

Shiraki,T. *et al*. (2003) Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc. Natl Acad. Sci. USA*, **100**, 15776–15781.

Siepel,A. *et al*. (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res*., **15**, 1034–1050.

Smit,A.F.A. *et al*. (1996–2004) RepeatMasker Open-3.0.

Su,A.I. *et al*. (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl Acad. Sci. USA*, **101**, 6062–6067.

Sverdlov,E.D. (2000) Retroviruses and primate evolution. *Bioessays*, **22**, 161–171.

van de Lagemaat,L.N. *et al*. (2003) Transposable elements in mammals promote regulatory variation and diversification of genes with specialized functions. *Trends Genet*., **19**, 530–536.