Supplementary Methods

Human cis natural antisense transcripts initiated by transposable elements

Andrew B. Conley[1], Wolfgang J. Miller[2] and I. King Jordan[1]

[1]School of Biology, Georgia Institute of Technology, Atlanta, GA 30306, USA
[2]Laboratories of Genome Dynamics, Center of Anatomy and Cell Biology, Medical University of Vienna, Vienna, Austria
Corresponding author: Jordan, I.K. (king.jordan@biology.gatech.edu).

Identification of TE-TSSs from CAGE data

A library of 1,551,672 human CAGE sequence tags [1, 2] was download from the Japanese National Institute of Genetics website (http://genomenetwork.nig.ac.jp/public/download/cage_Database_e.html). The data used in the manuscript correspond to the 2007.3.28 release. Human CAGE sequence tags were mapped to the hg18 version, *i.e.* the National Center for Biotechnology Information release 36, of the reference human genome sequence as previously described [3]. A browser extensible data (BED) format custom track with all CAGE tag-to-genome mapping coordinates, available on request, was generated in order to integrate the CAGE data with the human genome reference sequence annotations available from the UCSC Genome Browser Database [4]. The UCSC Table Browser [5] was used together with a series of custom developed perl scripts, available on request, to identify the intersection between human transcriptional start sites (TSSs) identified by CAGE tags with human transposable elements (TEs). To identify TE-derived TSSs, the human CAGE custom track was intersected with the RepeatMasker [6] (rmsk track) annotation track using 100% overlap and non TE-classes of repetitive DNA were subsequently eliminated from consideration. Specific TE family/class identities of the resulting TE-TSSs were determined by mapping these results back to the rmsk track and parsing the annotation therein. The observed percentages of 1) all, 2) sense and 3) antisense TE-TSS were determined for seven individual classes/families of TEs. The observed values were compared to the expected values that were determined by calculating their relative frequencies in the RepeatMasker annotation for the whole genome. Observed and expected values sum to 100% over all TE categories.

Human genes and TSSs from CAGE data

The UCSC Genome Browser 'Old Known Genes' track annotations (http://www.genome.ucsc.edu/cgi-bin/hgTrackUi?hgsid=99200641&c=chr7&g=knownGeneOld2) were used to define the coordinates of human protein-coding genes on the hg18 reference sequence. These human gene annotations were chosen because they represent a conservative set of gene definitions that are supported by multiple lines of evidence from the SWISS-PROT, TrEMBL and Genbank databases [7, 8]. A custom perl script was used to divide all human genes, from their 5' to 3' ends, into twenty equal sized bins, and the TSSs identified from CAGE data were mapped into gene-specific bins. Where the Old Known Genes track annotates multiple alternative transcript variants transcribed from a single genomic locus in the same direction, the resulting TSSs locations were only counted once. The antisense-versus-sense orientations (ratios) of TSSs were then considered with respect to their location in each bin along the gene lengths. This procedure was repeated for 1) non TE-TSS, 2) all TE-TSS and 3) individual families/classes of TE-TSS.

Relative ages of TE-TSSs

The relative ages of different families/classes of TEs were taken from the RepeatMasker analysis of the human genome reference sequence [9]. Since TE sequences in the human genome are derived from, and related to, copies of once active elements, and have subsequently accumulated mutations after insertion in the genome, the elements can be clustered into phylogenetic trees and grouped into related families/classes. The ensemble of sequences in any given class can be used to compute a consensus

sequence, which is taken to represent the ancient (active) copy of the TE [10]. Such consensus sequences have been extensively constructed from human genome TEs and are available in the Repbase database [11, 12]. Ages of TEs can be then be inferred by comparing the sequence divergence between the extant element sequence identified in the genome and its most closely related consensus sequence [13]. This information is made available as the 'millidiv' output, *i.e.* number of substitutions per 1,000 sites, from the RepeatMasker program. Percent substitution of extant TEs from consensus sequences was used to show that the human genome has experienced successive waves of expansion of different families/classes. Consequently, some families/classes are substantially older (or younger) than others. The most ancient families in the human genome are the L2 and MIR families, while the youngest are L1 and Alu [9]. These are the specific findings that are used to consider the relative ages of TE-TSSs derived from different families/classes.

Divergence (*d*) of extant TE sequences identified in the human genome from their consensus sequences were also used to evaluate the relative ages of TEs within the Alu family of elements. To do this, individual Alu insertion millidiv values were converted to Jukes-Cantor DNA sequence distances [14] using the following formula:

$$d = -3/4 * \ln[1 - 4/3(millidiv/1{,}000)$$

Then, the average and standard deviation *d*-values were computed and compared for Alu elements that donate TSSs versus those that do not using the Student's t-test.

Human-Mouse conservation of TSS

To evaluate the relative human-to-mouse evolutionary conservation, *i.e.* presence/absence of orthologous insertions, of 1) non TE-TSS, 2) all TEs and 3) all TE-TSS, the UCSC Genome Browser 'liftOver' utility was run locally. This program allows for annotation coordinates from one genome, or build, to be directly transferred to a second genome based on where they correspond. In the case of the human-mouse comparison, the coordinate correspondence is based on whole genome sequence alignments [15] represented in the Mouse Chain track (http://www.genome.ucsc.edu/cgi-bin/hgTrackUi?hgsid=99200641&c=chr7&g=chainMm9). To count the number of base pairs conserved between human and mouse for the different categories mentioned above, the 'Base Coverage' utility of the Galaxy Server [16] was used. Relative conservation was measured as the fraction of base pairs conserved for the different categories.

Supplementary Methods References

1 Carninci, P., *et al.* (2006) Genome-wide analysis of mammalian promoter architecture and evolution. *Nat Genet* 38, 626-635

2 Kodzius, R., *et al.* (2006) CAGE: cap analysis of gene expression. *Nat Methods* 3, 211-222

3 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 U S A* 100, 15776-15781

4 Karolchik, D., *et al.* (2003) The UCSC Genome Browser Database. *Nucleic Acids Res* 31, 51-54

5 Karolchik, D., *et al.* (2004) The UCSC Table Browser data retrieval tool. *Nucleic Acids Res* 32, D493-496

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

7 Benson, D.A., *et al.* (2007) GenBank. *Nucleic Acids Res* 35, D21-25

8 Boeckmann, B., *et al.* (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res* 31, 365-370

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

10 Jurka, J., *et al.* (1992) Prototypic sequences for human repetitive DNA. *J Mol Evol* 35, 286-291

11 Jurka, J. (2000) Repbase update: a database and an electronic journal of repetitive elements. *Trends Genet* 16, 418-420

12 Jurka, J., *et al.* (2005) Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res* 110, 462-467

13 Kapitonov, V., and Jurka, J. (1996) The age of Alu subfamilies. *J Mol Evol* 42, 59-65

14 Jukes, T.H., and Cantor, C.R. (1969) Evolution of protein molecules. In *Mammalian protein metabolism* (Munro, H.D., ed), Academic

15 Schwartz, S., *et al.* (2003) Human-mouse alignments with BLASTZ. *Genome Res* 13, 103-107

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