Evolutionary turnover of mammalian transcription start sites, Supplementary text

Martin C Frith[1,3], Jasmina Ponjavic[1,5], David Fredman[4], Chikatoshi Kai[1], Jun Kawai[1], Piero Carninci[1,2], Yoshihide Hayshizaki[1,2], Albin Sandelin[1]*

1 Genome Exploration Research Group, RIKEN Genomic Sciences Centre (GSC), RIKEN Yokohama Institute, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa, 230-0045, Japan.

2 Genome Science Laboratory, Discovery and Research Institute, RIKEN Wako Institute, 2-1 Hirosawa, Wako, Saitama, 351-0198, Japan

3 Institute for Molecular Bioscience, University of Queensland, Brisbane, Qld 4072, Australia.

4 Computational Biology Unit, Bergen Center for Computational Science, University of Bergen, HIB, Thormøhlensgate 55, N-5008 Bergen, Norway

5 Present address: MRC Functional Genetics Unit, Department of Physiology, Anatomy and Genetics, University of Oxford, Oxford OX1 3QX, United Kingdom.

* Corresponding author. Email: rgscerg@gsc.riken.jp

Running title: Evolutionary turnover of transcription start sites

This Supplementary text contains supporting methodology and a summary of information on the reproducibility and accuracy of the CAGE technology.

SUPPLEMENTARY METHODS

SNP density and derived allele frequencies

To examine if TSS turnover is associated with recent adaptive evolution in humans, we analyzed human polymorphism data in the turnover and reference TSS sets. First, we compared their densities of human dbSNP(Sherry et al. 2001) polymorphisms and, as a control for any SNP ascertainment bias, SNPs from a a single unbiased discovery effort (see below). The densities were similar (3.66×10^{-3} versus 3.58×10^{-3} and 3.38×10^{-4} versus 2.79×10^{-4} respectively). Next, we examined the derived allele frequencies of SNPs genotyped in the HapMap project (release 20)(Altshuler et al. 2005). No significant differences between the derived allele frequency (DAF) distributions of 12 SNPs residing in turnover set, and 115 SNPs within the reference set were found in any of the populations (see below). Although the number of observations is too low for any definitive conclusions, these results are consistent with the SNP density analysis in that they show no evidence of different selection pressures on the two TSS sets in human populations.

SNPs were obtained from dbSNP build 124 (http://www.ncbi.nlm.nih.gov/SNP/) (dbSNPs)(Sherry et al. 2001). A subset of dbSNPs from a single unbiased discovery effort (The International SNP Map Working Group. 2001) was extracted by the dbSNP submitter id "TSC-CSHL". Overlap determined by human genome coordinates (NCBI b35) gave 63 SNPs in the turnover set (toSNPs) and 877 SNPs in reference set (rfSNPs). SNP densities were calculated by dividing SNP count by TSS region length, and compared using Fisher's exact test. An unambiguous ancestral allele could be determined for 53 toSNPs and 609 rfSNPs by

BLAT (Kent 2002) of the SNP alleles and 30 bp flanking sequence to each side against the chimpanzee genome draft (Chimpanzee Sequencing and Analysis Consortium 2005), requiring at least 58 matching bases and that the chimp allele matched one of the human alleles. HapMap (release 20) (The International HapMap Consortium. 2005.) genotypes polymorphic in at least one population were available for 11 toSNPs and 114 rfSNPs. The distributions of derived allele frequencies (DAF) for toSNPs and rfSNPs were compared in each population by a Kolmogorov-Smirnov test and a Fisher's exact test on SNPs partitioned by having a DAF above or below 10%.

TATA-box occurrence in promoters with transcription start site turnover

TATA-boxes were predicted using the position-specific weight matrix defined by Bucher(Bucher 1990) deposited in the JASPAR database(Vlieghe et al. 2006) in the -50 to -1 region relative to each mouse cDNA start site in the turnover and reference set. The TFBS scientific programming module(Lenhard and Wasserman 2002) was used for scanning, using a 75% score cutoff. The choice of genomic window is motivated by the preferred TATA-TSS distance (30-31 bp)(Carninci et al. 2006; Hahn 2004). A promoter with one or more predicted TATA sites was considered “TATA positive”. The turnover and background set have no significant difference in TATA-boxes (P=0.30, Fisher’s exact test). Other choices of cutoffs and genomic windows did not change the results significantly (data not shown). This is also true if the number of TATA sites is used instead of the number of positive promoters.

GO analysis

We compared the gene ontology (GO) annotation between 74 GO-annotated TUs (transcriptional unit as defined in(Carninci et al. 2005; Carninci et al. 2006)) from the turnover

set and 1,120 GO-annotated TUs from the reference set. We used the GO annotation for each TU provided by FANTOM3 (Carninci et al. 2005; Carninci et al. 2006) and excluded the TUs without any GO term assignments from the analysis. For each of the 966 testable GO terms that were associated with at least 3 TUs from the turnover or reference set, a two-sided Fisher's exact test was performed to test if the GO term is significantly associated with the TU from the turnover set compared to the reference set or *vice versa*. The Bonferroni method(Westfall and Wolfinger 1997) was then applied to correct the resulting p-values for multiple testing.

REPRODUCIBILITY AND ACCURACY OF CAGE TAGS

Here is a summary of evidence for the validity of CAGE tags, which is described in more detail elsewhere (Carninci et al. 2006).

CAGE technology relies on two independent biochemical events: the extension of reverse transcriptase to the 5-‘ end of the transcript, and the CAP-dependent second strand synthesis, capture and cloning of a cDNA. Instances of truncated products require a failure of both reactions: a failure of the reverse transcriptase to generate a full-length product combined with an erroneous capture of the uncapped product by the CAP trapping procedure. The cap selection is highly efficient: when measuring the enrichment for RNApolII-derived mRNAs to uncapped ribosomal RNAs: a 330-fold enrichment was observed.

Multiple lines of evidence show that even single CAGE tags are reliable indicators of TSS locations. In a case study, 86% of single CAGE tags in the OPRM locus were confirmed by the independent RACE method. Comparison with randomly selected reported experiments on single

promoters (in which the TSS had been determined by nuclease protection, RNase protection and primer extension) show that in the majority of cases the CAGE data is wholly consistent with the previous results.

TSS reproducibility over multiple libraries (each library is an independent experiment) is high: if an exact TSS starting nucleotide position is defined by just two tags, these are from different libraries in 77% of cases. Further evidence of the non-randomness of the CAGE tags was obtained by analyzing the distribution of initiation site [-1,+1] dinucleotides for TSS having different levels of CAGE tag support separately (note that the -1 position is not part of the tag). Regardless of the level of support, each dinucleotide distribution is significantly different from a distribution of randomly sampled genomic dinucleotides. Lastly, there is a clear correlation between TSS defined by CAGE and active promoters identified by chromatin immuno-precipitation of transcription complexes in humans(Carninci et al. 2006).

SUPPLEMENTARY REFERENCES

Altshuler, D., L.D. Brooks, A. Chakravarti, F.S. Collins, M.J. Daly, and P. Donnelly. 2005. A haplotype map of the human genome. *Nature* **437:** 1299-1320.

Bucher, P. 1990. Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. *J Mol Biol* **212:** 563-578.

Carninci, P. T. Kasukawa S. Katayama J. Gough M.C. Frith N. Maeda R. Oyama T. Ravasi B. Lenhard C. Wells R. Kodzius K. Shimokawa V.B. Bajic S.E. Brenner S. Batalov A.R. Forrest M. Zavolan M.J. Davis L.G. Wilming V. Aidinis J.E. Allen A. Ambesi-Impiombato R. Apweiler R.N. Aturaliya T.L. Bailey M. Bansal L. Baxter K.W. Beisel T. Bersano H. Bono A.M. Chalk K.P. Chiu V. Choudhary A. Christoffels D.R. Clutterbuck M.L. Crowe E. Dalla B.P. Dalrymple B. de Bono G. Della Gatta D. di Bernardo T. Down P. Engstrom M. Fagiolini G. Faulkner C.F. Fletcher T. Fukushima M. Furuno S. Futaki M. Gariboldi P. Georgii-Hemming T.R. Gingeras T. Gojobori R.E. Green S. Gustincich M. Harbers Y. Hayashi T.K. Hensch N. Hirokawa D. Hill L. Huminiecki M. Iacono K. Ikeo A. Iwama T. Ishikawa M. Jakt A. Kanapin M. Katoh Y. Kawasawa J. Kelso H. Kitamura H. Kitano G. Kollias S.P. Krishnan A. Kruger S.K. Kummerfeld I.V. Kurochkin L.F. Lareau D. Lazarevic L. Lipovich J. Liu S. Liuni S. McWilliam M. Madan

Babu M. Madera L. Marchionni H. Matsuda S. Matsuzawa H. Miki F. Mignone S. Miyake K. Morris S. Mottagui-Tabar N. Mulder N. Nakano H. Nakauchi P. Ng R. Nilsson S. Nishiguchi S. Nishikawa F. Nori O. Ohara Y. Okazaki V. Orlando K.C. Pang W.J. Pavan G. Pavesi G. Pesole N. Petrovsky S. Piazza J. Reed J.F. Reid B.Z. Ring M. Ringwald B. Rost Y. Ruan S.L. Salzberg A. Sandelin C. Schneider C. Schonbach K. Sekiguchi C.A. Semple S. Seno L. Sessa Y. Sheng Y. Shibata H. Shimada K. Shimada D. Silva B. Sinclair S. Sperling E. Stupka K. Sugiura R. Sultana Y. Takenaka K. Taki K. Tammoja S.L. Tan S. Tang M.S. Taylor J. Tegner S.A. Teichmann H.R. Ueda E. van Nimwegen R. Verardo C.L. Wei K. Yagi H. Yamanishi E. Zabarovsky S. Zhu A. Zimmer W. Hide C. Bult S.M. Grimmond R.D. Teasdale E.T. Liu V. Brusic J. Quackenbush C. Wahlestedt J.S. Mattick D.A. Hume C. Kai D. Sasaki Y. Tomaru S. Fukuda M. Kanamori-Katayama M. Suzuki J. Aoki T. Arakawa J. Iida K. Imamura M. Itoh T. Kato H. Kawaji N. Kawagashira T. Kawashima M. Kojima S. Kondo H. Konno K. Nakano N. Ninomiya T. Nishio M. Okada C. Plessy K. Shibata T. Shiraki S. Suzuki M. Tagami K. Waki A. Watahiki Y. Okamura-Oho H. Suzuki J. Kawai and Y. Hayashizaki. 2005. The transcriptional landscape of the mammalian genome. *Science* **309:** 1559-1563.

Carninci, P., A. Sandelin, B. Lenhard, S. Katayama, K. Shimokawa, J. Ponjavic, C.A.M. Semple, M.S. Taylor, P. Engstrom, M. Frith, A.R. Forrest, W. Alkema, S.L. Tan, C. Plessy, R. Kodzius, T. Ravasi, T. Kasukawa, S. Fukuda, M. Kanamori-Katayama, Y. Kitazume, H. Kawai, C. Kai, H. Konno, K. Nakano, S. Mottagui-Tabar, P. Arner, A. Chesi, S. Gustincich, F. Persichetti, H. Suzuki, S.M. Grimmond, C. Wells, V. Orlando, C. Wahlestedt, E.T. Liu, M. Harbers, J. Kawai, V.B. Bajic, D.A. Hume, and Y. Hayashizaki. 2006. Genome-wide analysis of mammalian promoter architecture and evolution based upon human and mouse CAGE data. *Nat Genet* **In press**.

Hahn, S. 2004. Structure and mechanism of the RNA polymerase II transcription machinery. *Nat Struct Mol Biol* **11:** 394-403.

Lenhard, B. and W.W. Wasserman. 2002. TFBS: Computational framework for transcription factor binding site analysis. *Bioinformatics* **18:** 1135-1136.

Sherry, S.T., M.H. Ward, M. Kholodov, J. Baker, L. Phan, E.M. Smigielski, and K. Sirotkin. 2001. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* **29:** 308-311.

Vlieghe, D., A. Sandelin, P.J. De Bleser, K. Vleminckx, W.W. Wasserman, F. van Roy, and B. Lenhard. 2006. A new generation of JASPAR, the open-access repository for transcription factor binding site profiles. *Nucleic Acids Res* **34:** D95-97.

Westfall, P.H. and R.D. Wolfinger, 3 -8. 1997. Multiple Tests with Discrete Distributions. *The American Statistician* **51:** 3-8.