Supplementary material for "Combining SAGE tags to predict genomic transcribed regions"

Eric Rivals[1], Anthony Boureux[2], Mireille Lejeune[2], Florence Ottones[2], Oscar Pecharromàn Pérez[3], Jorma Tarhio[3], Fabien Pierrat[4], Florence Ruffle[2], Thérèse Commes[2] and Jacques Marti[2]

[1] LIRMM, Univ. Montpellier II, CNRS, Montpellier, France,{rivals}@lirmm.fr
[2] Institut de Génétique Humaine, CNRS Montpellier, France, {marti,commes}@univ-montp2.fr
[3] Helsinki University of Technology, Finland
[4] Skuld-Tech, Montpellier, France

1 Material and Methods

1.1 External data sets

SAGE data were collected from publicly available repositories (http://www.ncbi.nlm.nih.gov/projects/geo/index.cgi: Platforms: GPL4, GPL6, and GPL1485, http://www.prevent.m.u-tokyo.ac.jp/SAGE.html, CAGP project (Sage genie): ftp://ftp1.nci.nih.gov/pub/SAGE/HUMAN/). The list of SAGE libraries is available (Supplementary Table 1). *Homo sapiens* chromosome sequences (HG17, NCBI build 35) were retrieved from the UCSC Genome Bioinformatics site (http://genome.ucsc.edu). UniGene cluster-representative sequences were taken from the Hs.seq.uniq. file, retrieved by FTP from the National Center for Biotechnology Information site (ftp://ftp.ncbi.nih.gov/repository). We used the UniGene built # 162 assembling $4,47$ million sequences into $123,995$ clusters and providing the same number of cluster-representative sequences. Since SAGE may detect several authentic transcripts from the same locus, we did not use more recent UniGene releases in which transcripts co-locating with known genes have been merged. Alu sequences were taken from RepBase Update (http://www.girinst.org/Repbase_Update.html) [2].

1.2 Macrophage SAGE libraries

Venous blood from healthy donors was obtained from the Etablissement Français du Sang (Montpellier, France). Monocytes, isolated by adherence to culture flasks, were differentiated into $> 99\%$ Monocyte Derived Macrophages (MDMs) as previously described (14). Total RNA (50 micrograms) from 8.106 MDMs was extracted with Trizol (Invitrogen, Cergy Pontoise, France). Polyadenylated mRNA was selected by hybridization to oligo (dT) 25-coated magnetic beads according to manufacturer's instructions (Dynal, Compiegne France). CATG-tags were prepared using the I-SAGE kit (Invitrogen, Cergy Pontoise, France) and GATC-tags using a modified Sau3A1 SAGE procedure [3]. The sequences of $22,387$ CATG-tags and $8,221$ GATC-tags determined by the Centre National de Séquençage (Evry, France) were analyzed for tag detection and counting using the C+tag software (Skuld- Tech, France).

1.3 Proximity between TDGS and Tiling arrays data

We retrieved Tiling arrays data from the UCSC Genome Bioinformatics site (http://genome.ucsc.edu/). We used transcriptional active regions (TARs) data from Affymetrix Transcriptome

Project Phase2, Affymetrix PolyA+ RNA transfrags, Yale RNA TARs and Yale Maskless Array synthesizer experiments [1]. We computed the number of TDGS that either strictly overlap a TAR, or are in a 500 bp vicinity of a TAR.

2 Results

2.1 Cases of novel transcripts

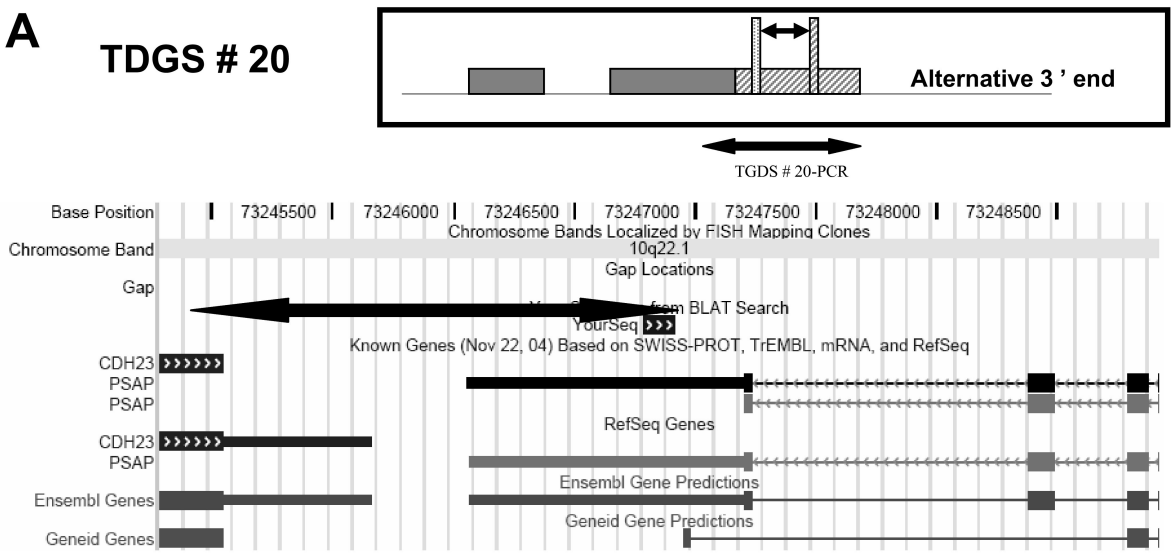

Figure 1. A case of alternative transcript. Alignments of the TDGS# 20 with the UCSC human genome browser. For RT-PCR validation, Macrophage polyA+ RNA were extracted from MDM and the cDNA were synthesized using mRNA and oligo-dT primer. TDGS# 20 corresponds to an example of Class 2 transcript localized near the coding region of CDH23. For PCR, a primer pair was respectively designed in the $3'$ end of CDH23 and in the TDGS # 20. The existence of this new variant transcript was confirmed in macrophage by sequencing.

References

[1] J. Cheng, P. Kapranov, J. Drenkow, S. Dike, S. Brubaker, S. Patel, J. Long, D. Stern, H. Tammana, G. Helt, V. Sementchenko, A. Piccolboni, S. Bekiranov, D. K. Bailey, M. Ganesh, S. Ghosh, I. Bell, D. S. Gerhard, and T. R. Gingeras. Transcriptional Maps of 10 Human Chromosomes at 5-Nucleotide Resolution. *Science*, 308:1149–1154, May 2005.

[2] J. Jurka. Repbase update: a database and an electronic journal of repetitive elements. *Trends Genet*, 16(9):418–20, 2000.

[3] V. Velculescu, L. Zhang, B. Vogelstein, and K. Kinzler. Serial analysis of gene expression. *Science*, 270(5235):484–7, 1995.