D550–D552 *Nucleic Acids Research, 2005, Vol. 33, Database issue*
doi:10.1093/nar/gki085

5′SAGE: 5′-end Serial Analysis of Gene Expression database

Yasuhiro Kasai[1,4], Shin-ichi Hashimoto[3], Tomoyuki Yamada[1], Jun Sese[1], Sumio Sugano[2], Kouji Matsushima[3] and Shinichi Morishita[1,*]

[1]Department of Computational Biology and [2]Department of Medical Genome Sciences, Graduate School of Frontier Science, The University of Tokyo, 5-1-5 Kashinoha, Kashiwa City, Chiba 277-8562, Japan, [3]Department of Molecular Preventive Medicine, School of Medicine, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan and [4]Hitachi Software Engineering Co., Ltd, 1-1-45 Suehiro-cho, Tsurumi-ku, Yokohama 230-0045, Japan

Received August 13, 2004; Revised and Accepted October 12, 2004

ABSTRACT

To comprehensively identify transcription start sites and the frequencies of individual mRNAs in human cell libraries, a method of 5′ end Serial Analysis of Gene Expression (SAGE) was developed recently, which makes it possible to collect a large amount of start site information, and subsequently, we have established a related database server called 5′SAGE. This database displays the observed frequencies of individual 5′ end SAGE tags and previously unknown transcription start sites in the promoter regions, introns and intergenic regions of known genes. 5′SAGE will be useful for analyzing promoter regions and start site variation in different tissues, and is freely available at http://5sage.gi.k.u-tokyo.ac.jp/.

INTRODUCTION

The analysis of transcription start sites has attracted considerable attention in the recent years. There is heterogeneity in human mRNA start sites (1); 40–60% of human genes are transcribed alternatively (2), and 49% of multi-exon transcripts are accompanied by alternative splicing of the initial exon (3). Transcription start sites might be altered in a variety of different cell types or affected by environmental conditions, such as methylation. Although an extensive collection of transcription start sites for a large number of human genes is available (4), the frequencies of individual start sites are unclear. There is a need for high-throughput technology to monitor the statistics of start site occurrences for a comprehensive understanding of the start site gene expression mechanism. Microarrays are unsuitable for this purpose because of their inability to detect novel start sites.

The serial analysis of gene expression (SAGE) method (5) has demonstrated its effectiveness at cataloging large quantities of expressed genes in cells or tissues from a variety of physiological, developmental and pathological states (6–11). The original SAGE[5] generates short (10+4 bp) nucleotide sequences, called tags, derived from the 3′ ends of transcripts; however, typical tags are too short to be uniquely identified with their corresponding genes. This shortcoming was resolved using the LongSAGE method (12), a high-throughput means of profiling 21 bp tags, which are sufficiently long to be unambiguously identified with genes in most cases. However, existing SAGE methods are designed to monitor the 3′ ends of transcripts, and the challenge was to extend the SAGE method so that it would be capable of capturing the novel 5′ ends of transcripts and efficiently quantifying individual 5′ end occurrences. Recently, Hashimoto *et al*. (13) developed such a system for human cell lines, while Shiraki *et al*. (14) reported a system for mouse cell lines. The 5′SAGE database stores a collection of data accumulated by using the Hashimoto *et al*.'s system.

METHODS

Hashimoto *et al*. (13) have described the details of the method, and we present a brief summary here. The method first profiles 21 bp tags by using a novel way of combining the oligo-capping method (15), a modification of the oligo-capping method (16) and the LongSAGE method (6). Subsequently, these 5′SAGE tags are aligned with the human genome to locate their positions, to begin a search for neighboring mRNA start sites.

We found that 19 893 of 25 684 5′SAGE tags in a human cell line, HEK293, were matched to the human genome. Of the

*To whom correspondence should be addressed. Tel: +81 47 136 3984; Fax: +81 47 136 3977; Email: moris@k.u-tokyo.ac.jp

15 448 tags that hit a locus within the human genome, 85.8–96.1% of the 5′SAGE tags were assigned to within −500 to +200 nt of the mRNA start sites in the RefSeq, UniGene (17) and DBTSS (4) databases, while 1774 tags were within the introns of known genes or uncharacterized regions, indicating possible novel start sites.

USE OF 5′SAGE

In the 5′SAGE database server, users can browse transcription start sites and frequencies of individual genes by querying on the accession numbers of sequences in RefSeq, cluster identifiers in UniGene or symbol names, such as HDAC. To retrieve all the genes in the server, the word 'ALL' can be input at the query box. The user can impose additional conditions on the number of distinct start sites and the total frequency of 5′SAGE tags monitored for individual genes of interest. For instance, one can look for genes by monitoring five or more distinct start sites with 10 or more 5′SAGE tag occurrences. In response to the query, the system returns the list of qualifying genes. Clicking on each gene displays a window for browsing the transcription start sites (Figure 1).

Two complementary views are provided for analyzing transcription start points. The 'Start Site View' initially displays the narrow, 150 bp region surrounding the transcription start site of the representation gene in RefSeq or UniGene, while the 'Global View' presents entire structures of individual transcripts that are helpful in comprehending alternatively spliced transcripts at a glance. Users can change the zoom magnification of each view independently by setting the ruler unit to an alternative base pair length. The thick horizontal blocks in the pictures represent exons. The orange vertical lines depict transcription start points; the depth of each orange line below the axis shows the frequency of the transcription start site. The thick, green, horizontal lines are CpG islands, which are regions of 50 or more bp consisting of at least 50% G or C nucleotides. Nucleotides are displayed when the ruler unit is set to 10 bp.

For instance, Figure 1 shows the transcription start sites of neurofilament 3 (NEF3). Note the large number of start points detected for NEF3; most are novel, and some start at the second or third exon. Genes with many start sites are remarkably common. The 'Transcription Start Site View' also lists 5′SAGE tags, their distances from the representative start site, their frequencies and their nearest expressed sequence tags. We have performed Long SAGE on the 3′ ends of mRNA in HEK293 cells to validate the accuracy of our 5′SAGE results (13). The total frequency of 3′SAGE tags associated with the representative gene is also displayed with the 3′SAGE tag sequences and distances from the start site. 5′SAGE tags are typically more diverse than 3′SAGE tags. As 5′SAGE and 3′SAGE tags are sampled independently at random, the Pearson correlation coefficient between the frequencies of 5′SAGE and 3′SAGE tags indicates moderate similarity (13).

UPDATES AND FUTURE DIRECTIONS

As on October 2004, the 5′SAGE database presents transcription start sites collected from human cell lines, HEK293 and Ramos. Start site information in other human cell lines is being collected for the analysis of start point variation in different tissues, and will be made available at the same website.

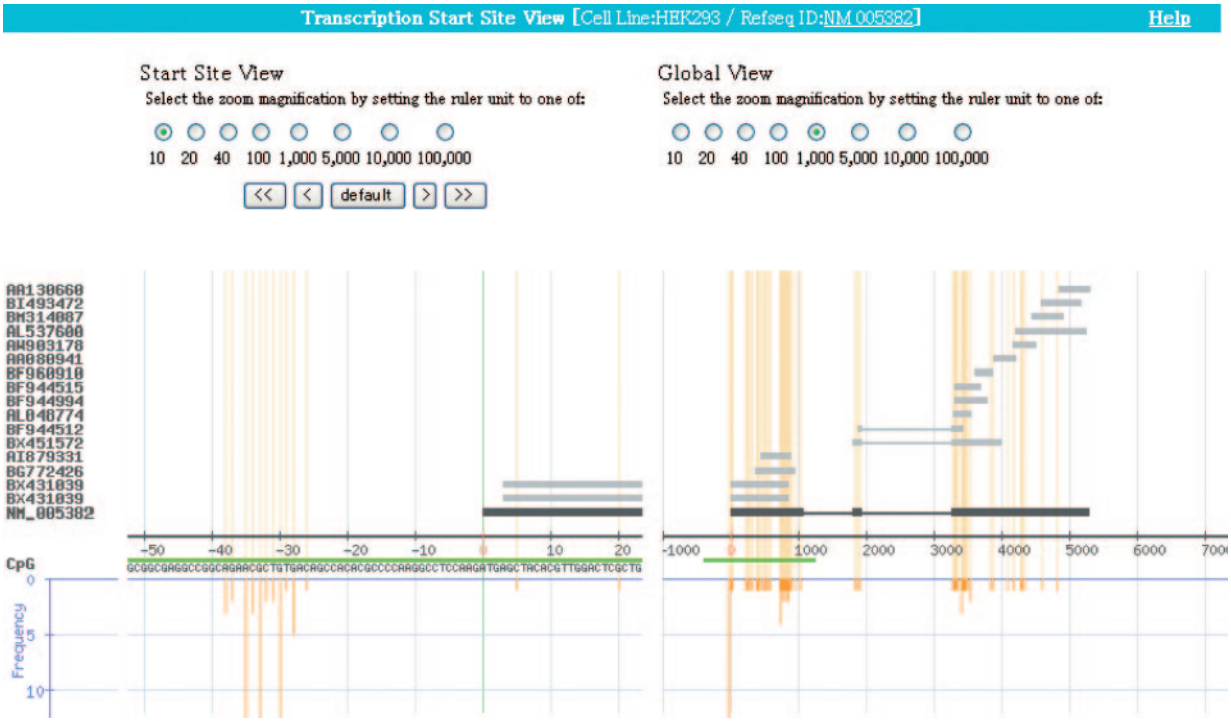

Figure 1. The use of 5′SAGE. The 'Start Site View' indicates the frequencies of start sites using orange lines for the start points of the gene being considered, while the 'Global View' presents the overall structures of individual genes to illustrate alternative splice variants.

ACKNOWLEDGEMENTS

This work was supported in part by a Grant-in-Aid for Scientific Research on Priority Areas (C) from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

REFERENCES

1. Suzuki,Y., Taira,H., Tsunoda,T., Mizushima-Sugano,J., Sese,J., Hata,H., Ota,T., Isogai,T., Tanaka,T., Morishita,S. *et al*. (2001) Diverse transcriptional initiation revealed by fine, large-scale mapping of mRNA start sites. *EMBO Rep*., **2**, 388–393.
2. Modrek,B. and Lee,C. (2002) A genomic view of alternative splicing. *Nature Genet*., **30**, 13–19.
3. Zavolan,M., Kondo,S., Schonbach,C., Adachi,J., Hume,D.A., Hayashizaki,Y. and Gaasterland,T. (2003) Impact of alternative initiation, splicing, and termination on the diversity of the mRNA transcripts encoded by the mouse transcriptome. *Genome Res*., **13**, 1290–1300.
4. Suzuki,Y., Yamashita,R., Sugano,S. and Nakai,K. (2004) DBTSS: Database of Transcriptional Start Sites: progress report 2004. *Nucleic Acids Res*., **32**, D78–D81.
5. Velculescu,V.E., Zhang,L., Vogelstein,B. and Kinzler,K.W. (1995) Serial analysis of gene expression. *Science*, **270**, 484–487.
6. Madden,S.L., Galella,E.A., Zhu,J., Bertelsen,A.H. and Beaudry,G.A. (1997) SAGE transcript profiles for p53-dependent growth regulation. *Oncogene*, **15**, 1079–1085.
7. Velculescu,V.E., Madden,S.L., Zhang,L., Lash,A.E., Yu,J., Rago,C., Lal,A., Wang,C.J., Beaudry,G.A., Ciriello,K.M. *et al*. (1999) Analysis of human transcriptomes. *Nature Genet*., **23**, 387–388.
8. Hashimoto,S.-I., Suzuki,T., Dong,H.-Y., Yamazaki,N. and Matsushima,K. (1999) Serial analysis of gene expression in human monocytes and macrophages. *Blood*, **94**, 837–844.
9. Hashimoto,S., Nagai,S., Sese,J., Suzuki,T., Obata,A., Sato,T., Toyoda,N., Dong,H.Y., Kurachi,M., Nagahata,T. *et al*. (2003) Gene expression profile in human leukocytes. *Blood*, **101**, 3509–3513.
10. Boon,K., Osorio,E.C., Greenhut,S.F., Schaefer,C.F., Shoemaker,J., Polyak,K., Morin,P.J., Buetow,K.H., Strausberg,R.L., De Souza,S.J. and Riggins,G.J. (2003) An anatomy of normal and malignant gene expression. *Proc. Natl Acad. Sci. USA*, **99**, 11287–11292.
11. Divina,P. and Forejt,J. (2004) The Mouse SAGE Site: database of public mouse SAGE libraries. *Nucleic Acids Res*., **32**, D482–D483.
12. Saha,S., Sparks,A.B., Rago,C., Akmaev,V., Wang,C.J., Vogelstein,B., Kinzler,K.W. and Velculescu,V.E. (2002) Using the transcriptome to annotate the genome. *Nat. Biotechnol*., **20**, 508–512.
13. Hashimoto,S., Suzuki,Y., Kasai,Y., Morohoshi,K., Yamada,T., Sese,J., Morishita,S., Sugano,S. and Matsushima,K. (2004) 5′-end SAGE for the analysis of transcriptional start sites. *Nat. Biotechnol*., **22**, 1146–1149.
14. Shiraki,T., Kondo,S., Katayama,S., Waki,K., Kasukawa,T., Kawaji,H., Kodzius,R., Watahiki,A., Nakamura,M., Arakawa,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.
15. Maruyama,K. and Sugano,S. (1994) Oligo-capping: a simple method to replace the cap structure of eukaryotic mRNAs with oligoribonucleotides. *Gene*, **138**, 171–174.
16. Suzuki,Y., Yoshitomo-Nakagawa,K., Maruyama,K., Suyama,A. and Sugano,S. (1997) Construction and characterization of a full length-enriched and a 5′-end-enriched cDNA library. *Gene*, **200**, 149–156.
17. Wheeler,D.L. (2003) Database Resources of the National Center for Biotechnology. *Nucleic Acids Res*., **31**, 28–33.