

5'-end SAGE for the analysis of transcriptional start sites

Shin-ichi Hashimoto¹, Yutaka Suzuki^{2,4}, Yasuhiro Kasai^{3,4}, Kei Morohoshi¹, Tomoyuki Yamada³, Jun Sese³, Shinichi Morishita³, Sumio Sugano² & Kouji Matsushima¹

Identification of the mRNA start site is essential in establishing the full-length cDNA sequence of a gene and analyzing its promoter region, which regulates gene expression. Here we describe the development of a 5'-end serial analysis of gene expression (5' SAGE) that can be used to globally identify transcriptional start sites and the frequency of individual mRNAs. Of the 25,684 5' SAGE tags in the HEK293 human cell library, 19,893 matched to the human genome. Among 15,448 tags in one locus of the genome, 85.8%–96.1% of the 5' SAGE tags were assigned within –500 to +200 nt of mRNA start sites using the RefSeq, UniGene and DBTSS databases. This technique should facilitate 5'-end transcriptome analysis in a variety of cells and tissues.

Comprehensive analysis of gene expression in different cell sub-populations and microanatomic structures provides insight into human development and physiology. Recently developed functional genomic technologies such as cDNA microarrays¹ and serial analysis of gene expression (SAGE)² allow for the analysis of the expression of thousands of genes. The SAGE method has proved to be a useful tool for quantifying, cataloging and comparing the expression of genes in cells and tissues that are in various physiological, developmental and pathological states^{3–5}.

In the human genome project⁶, genes were predicted based on information obtained from expressed sequence tag (EST) maps, analysis of full-length cDNAs and computational annotation by Genscan, Genie, FGENES and other programs. However, these computational analyses had limitations. For example, they were unable to provide definitive evidence that a hypothetical gene was actually expressed. The Long SAGE method is useful in rapidly identifying not only novel genes and exons but also the frequency of each transcript. However, a remaining challenge in human genome expression profile analysis is the exact identification and annotation of entire expressed genes based on information concerning their 5'-end to 3'-end transcripts. Therefore, a method for precisely and globally identifying the 5' ends of transcripts is needed. Such a 5'-end capping method⁷ would give us the sequence tag to the first base of an mRNA. Using cDNA libraries made from oligo-capped mRNA preparations and SAGE, we developed a method to comprehensively identify the transcriptional start sites of mRNAs.

The method is shown in Figure 1. The mRNA is split into two aliquots and its cap structure is enzymatically replaced with two types of

synthetic oligonucleotides containing *Mme*I, a type-IIIS restriction endonuclease site, and the *Xho*I restriction enzyme site. Oligo-capping mRNA is then converted into first strand cDNA with a random adapter-primer. The second strand is synthesized using biotin-bound 5'-primer and random adaptor-primer by PCR. The double-stranded cDNA is then cleaved with *Mme*I, which cleaves 20 bp away from its recognition site. After the 5'-cDNAs are isolated by binding to streptavidin beads, the two pools of tags are ligated to each other. The subsequent procedure for concatemerization is carried out according to the original SAGE protocol².

Using this method, we characterized 25,684 transcripts expressed in HEK293 cells as a test cell line and compared them to the human genome sequence. A total of 19,893 tags matched perfectly to genomic sequences representing 13,404 different tags (Table 1). Eighty percent (10,706 tags) of 13,404 different tags were assigned to unique positions. Eleven percent (1,483 tags) of the tags matched two loci in the genome, 8.1% (1,090 tags) matched 3–99 loci, and 0.9% (125 tags) matched >100 loci. The tags that mapped to multiple genomic loci mostly corresponded to retrotransposon elements, repetitive sequences, or pseudogenes.

We estimated whether the 5' SAGE tags matched to the mRNA start sites using three databases: the reference sequence database (RefSeq), the Gene Resource Locator (GRL) database, which assembles gene maps that include information on *cis*-elements in regulatory regions and alternatively spliced transcripts, and the database of Transcriptional Start Sites (DBTSS)⁸, which contains systematic 5'-end sequences of human full-length cDNAs. The distance distributions and the number and ratio of small-distance tag occurrences that we obtained are shown in Figure 2a and Table 2. The data indicate that our 5' SAGE tags coincided well with start site information for each database. Most (85.8%–98.2%) tags that mapped to each database were –500 to +200 nt of mRNA start sites. Notably, 23.5%–49.3% of 5' SAGE tags hit the upstream regions of the defined transcription start sites (TSS) in these databases.

To experimentally confirm the 5' ends of known genes identified by 5' SAGE, we compared the directed sequencing data of the 5' ends of captured full-length cDNAs in HEK293 with 5' SAGE data. The representative TSS identified by 5' SAGE and full-length cDNA using the oligo-capping method are shown in Supplementary Figure 1. A part of various 5' SAGE tags corresponded to those of full-length cDNA using the oligo-capping method.

¹Department of Molecular Preventive Medicine, School of Medicine, ²Department of Virology, Institute of Medical Science, ³Department of Computational Biology, Graduate School of Frontier Science, The University of Tokyo, 7-3-1, Hongo, Bunkyo-ku, Tokyo 113-0033, Japan. ⁴These authors contributed equally to this work. Correspondence should be addressed to K.M. (koujim@m.u-tokyo.ac.jp).

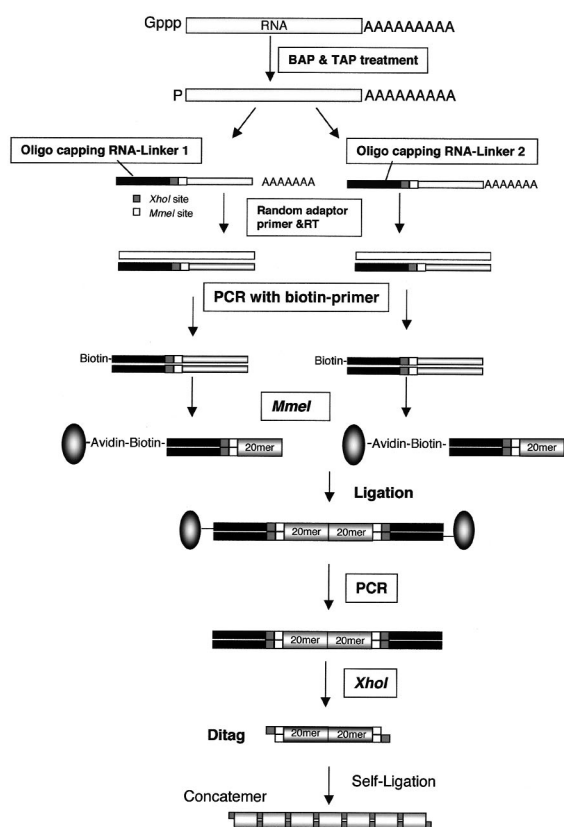


Figure 1 Scheme for construction of a 5'-end SAGE library as detailed in Methods.

We examined the nucleotide preference at the TSS. It has been reported⁹ that the nucleotides of TSS are A (47%), G (28%), C (14%) and T (12%) using 5,880 mRNAs in 276 human genes. Our data showed very similar percentages for the first nucleotide: A (41%), G (32%), C (17%) and T (10%). Taken together, these results suggest that our 5' SAGE method can precisely identify TSS. Thus, the method allows us to obtain both accurate transcriptional start site information and information that could be used to analyze the relationship between promoter and transcription.

Interestingly, no matches to genome tags in the databases were found for 23% of the 5' SAGE tags sequenced in this study. In 39% of these, the first nucleotide was A. Some of the tags without matches to the genome may contain single nucleotide mutations or deletions.

To identify uncharacterized genes, we compared 5' SAGE tags with RefSeq, EST and DBTSS databases. Of the 10,706 unique tags with

a single locus in the genome, 9,376 corresponded to UniGene EST sequences and 7,800 corresponded to RefSeq sequences (Supplementary Table 1). Furthermore, 6,418 unique 5' SAGE tags corresponded to known genes in DBTSS. The remaining tags (12.4%) matched regions within introns (5.4%) of known genes or uncharacterized regions (6.6%). Tags that matched uncharacterized regions primarily hit two sites—completely uncharacterized regions or regions of uncharacterized EST sequences. Evidence of expression of such genes should help to uncover the genes' full-length form by referring to 3' SAGE.

The SAGE method can be used to obtain quantitative transcript information. In the profile of the 5'-end transcripts in HEK293 cells (Supplementary Table 2), the most frequently expressed genes were identified as neurofilament 3 (NEF3), which had an expression frequency of 1.43%, followed by elongation factor 2. Several genes such as NEF3, heat shock 70 kDa protein 1A, calreticulin and heterogeneous nuclear ribonucleoprotein H1 were found using different tags. It is theorized that certain genes can be transcribed from different TSS. For example, our data suggest that both calreticulin and lactate dehydrogenase A can be transcribed from seven different transcriptional start sites (Fig. 2b).

We performed Long SAGE on the 3' ends of mRNA in HEK293 cells to validate the accuracy of our 5' SAGE results. Using 3' Long SAGE, we characterized 81,212 transcript tags. A total of 54,050 tags matched genomic sequences representing 15,423 different tags (Table 1). Seventy-five percent (11,613 tags) of 15,423 different tags matched one site in the genome. Furthermore, 8,359 3' SAGE tags were found to match known genes in the UniGene EST database, and 5,267 3' SAGE tags matched RefSeq genes (Supplementary Table 1). Nine percent of tags (1,395 tags) matched two loci in the genome, 13.2% (2,039 tags) matched 3–99 loci and 2.4% (376 tags) matched >100 loci. The percentage of tags that matched multiple sites in the genome was very similar for 5' SAGE and 3' SAGE (Table 1). On other hand, 5' SAGE tags were very heterogeneous whereas 3' SAGE tags were not.

It has been shown that tags present at >10 copies per genome are on average more highly expressed than those present at only one copy per genome². Similarly, our data demonstrated that the relative expression level was higher in 3 to ~99 loci/genome than in other fractions in the 5' SAGE and 3' SAGE libraries, because of association between gene expression and gene duplication through retrotransposition.

To estimate the extent of the similarity between the two libraries, we compared the expressed genes of 5' SAGE and 3' Long SAGE. Because 5' and 3' tags were randomly sampled separately from their 5' and 3' ends, the probability that 5' tags could have associated with a particular full-length cDNA sequence was expected to coincide with the probability that 3' tags matched the cDNA. However, owing to incomplete collection of full-length cDNA sequences or alternatively spliced transcripts, it was not easy to determine the exact correspondence between 5' and 3' tags even though they might have originated from the same coding region.

Table 1 Experimental matching of SAGE tag to genome

| Tag loci in genome ³ | 5'-end SAGE tags to genome ¹ | | | 3'-end SAGE tags to genome ² | | |
|---------------------------------|---|----------------------------------|---------------------------|---|----------------------------------|---------------------------|
| | Tags mapped to genome (%) | Unique tags mapped to genome (%) | Relative expression level | Tags mapped to genome (%) | Unique tags mapped to genome (%) | Relative expression level |
| 1 loci/genome | 15,448 (77.7) | 10,706 (79.9) | 1.44 | 34,139 (63.2) | 11,613 (75.3) | 2.94 |
| 2 loci/genome | 2,037 (10.2) | 1,483 (11.1) | 1.37 | 6,739 (12.5) | 1,395 (9.0) | 4.83 |
| 3–99 loci/genome | 2,275 (11.4) | 1,090 (8.1) | 2.09 | 12,265 (22.7) | 2,039 (13.2) | 6.02 |
| >100 loci/genome | 133 (0.7) | 125 (0.9) | 1.06 | 907 (1.7) | 376 (2.4) | 2.42 |
| Total tags | 19,893 (100) | 13,404 (100) | 1.40 | 54,050 (100) | 15,422 (100) | 2.13 |

¹Number of 18-bp 5' SAGE tags getting hits in genome; 5,791 tags of 25,684 tags sequenced did not get hits. Relative expression level was determined by dividing the total number of transcript tags observed in the library by the number of different tags. ²Number of 20-bp 3' SAGE tags getting hits in genome; 27,162 tags of 81,211 sequenced did not get hits.

³Number of genome locations matched by individual SAGE tags.

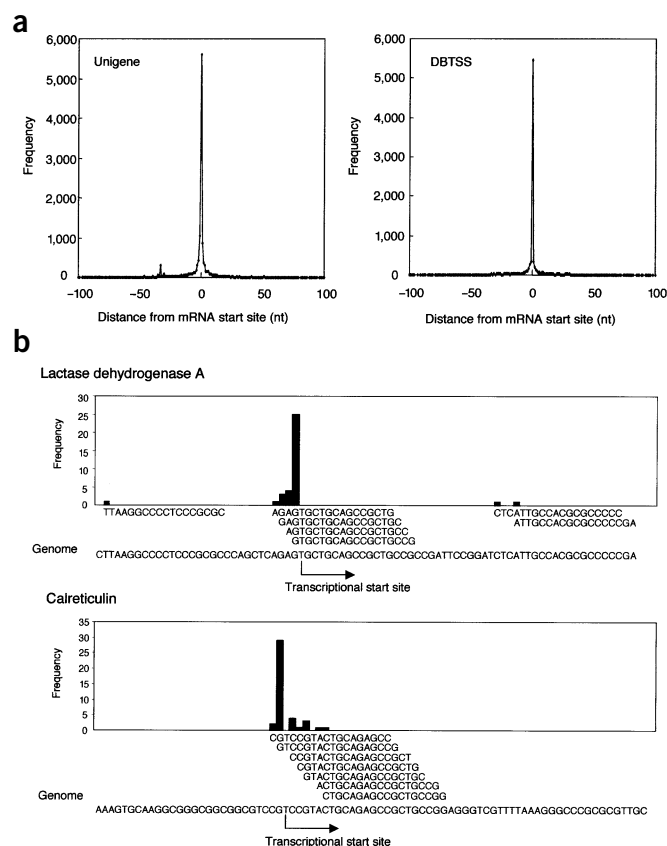


Figure 2 5' SAGE tags hit around the defined transcription start sites. (a) Distance of 5' SAGE tags relative to the mRNA start sites in the UniGene and DBTSS sequences. Distances are shown as the number of upstream (–) and downstream (+) nucleotides (x-axis). The mRNA start site is depicted as 0. The frequency of the 5' SAGE tag is given on the y-axis. A small distance between the aligned positions of each 5' SAGE tag and their corresponding gene implies that the 5' tags are almost consistent with the known 5' transcriptional start site. We used UniGene and DBTSS databases separately to determine their coverage differences. (b) mRNA start sites and frequency.

One possible approach would be to put together EST alignments that shared exons in common, treat such a cluster as a gene coding locus, map 5' and 3' SAGE tags to these clusters and their upstream regions and then uncover correspondences between 5' and 3' SAGE tag expression. Using this approach, we counted pairs of 3' and 5' tags for each gene coding region (Supplementary Fig. 2). Comparison of expression patterns revealed that most genes were expressed at similar levels in both libraries, though several transcripts were expressed at substantially different levels. Pearson correlation coefficients of 5' SAGE and 3' SAGE libraries showed moderate similarity (0.36). This moderate correlation was probably due to dispersion of the frequencies from the 5' SAGE and 3' SAGE libraries. There are several possible reasons for the appearance of these tags, including: (i) PCR amplification errors in 5' SAGE and 3' SAGE; (ii) appropriation of the *Nla*III restriction site in 3' SAGE by a small number of genes; (iii) appropriation of the *Xho*I restriction site in 5' SAGE by a small number of genes; (iv) the presence of unknown splicing variants of mRNA in 5' SAGE and 3' SAGE; and (v) an annotation error for tags hitting multiple genomic loci or an EST-annotation error in the genome.

This study identified only a fraction of the genes expressed in HEK293 cells. A much larger number of tags from a variety of different cell types

Table 2 Distance of 5' SAGE tags relative to mRNA start sites in each database

| Approximate distance from the start site of each database (nt) | Tag number (%) | | |
|--|----------------|---------------|--------------|
| | RefSeq | UniGene (GRL) | DBTSS |
| –500 to –201 | 349 (3.2) | 204 (1.5) | 160 (1.6) |
| –200 to –51 | 887 (8.1) | 335 (2.4) | 253 (2.5) |
| –50 to –1 | 4,179 (38.1) | 3,957 (28.8) | 1,965 (19.5) |
| 0 to +50 | 3,173 (28.9) | 8,673 (63.2) | 7,149 (70.8) |
| +51 to +200 | 837 (7.6) | 311 (2.3) | 209 (2.1) |
| (–500 to +200) | 9,425 (85.8) | 13,480 (98.2) | 9,736 (96.4) |
| Total tags | 10,982 (100) | 13,723 (100) | 10,098 (100) |

The tags that correspond in mapping with the 5' ends of genes from each database were analyzed as described in Figure 2.

cultured under varying environmental conditions will be required to more thoroughly describe the compendium of expressed genes.

Several groups reported that mRNA start sites⁹ and polyadenylation cleavage sites¹⁰ show heterogeneity. Although TSS differences have been reported for specific genes in tissues¹¹, our data showed that the diversity of TSS already exists on a cellular level. Moreover, our data provide direct evidence of heterogeneity of TSS and 3' end regions using 5' SAGE and 3' SAGE methods. For example, we found that the peroxisome proliferator-activated receptor-binding protein has one TSS and two 3' SAGE tag sites, ribosomal protein S4 has 16 TSS and one 3' SAGE tag site, and calreticulin has seven TSS and one 3' SAGE tag site.

Alternative mRNA splicing contributes in important ways to the complexity of the human proteome. Recent genomic studies suggest that 40%–60% of human genes are alternatively spliced¹². Fifteen percent of point mutations are estimated to cause human genetic diseases by triggering an mRNA splicing defect¹³. It has been reported that 49% of the transcriptional units with multiple splice forms included transcripts in which usage of an alternative TSS was accompanied by alternative splicing of the initial exon¹⁴. We also found that the mRNA start sites of several genes, such as peroxiredoxin 4 (NM_006406), represented not only different splicing variants of mRNA but also different degrees of gene expression, suggesting that alternative transcription may frequently induce alternative splicing.

In conclusion, our 5' SAGE method should facilitate the annotation of genomes. Because this method represents one of the few high-throughput discovery approaches that does not depend on *a priori* knowledge of gene sequences, it can be used for independent validation of *in silico* gene predictions and for the identification of nonannotated regions. In addition, it should be useful for finding single nucleotide polymorphisms in 5'-untranslated or promoter regions. Considering the diversity of 5' ends, it is more appropriate to perform 5' SAGE rather than 3' SAGE when determining the frequency of gene expression. Comprehensive identification of genes transcribed from specific mRNA start sites in different cell types will not only provide insight into the functional complexity of the human genome but may also aid the diagnosis of various disorders such as cancer and immunological and neural diseases.

METHODS

Generation of 3'-Long SAGE library. Total RNA was prepared from HEK293 cells and mRNA was isolated as previously described¹⁵. Long SAGE was performed as previously described². SAGE 2000 software (version 4.12) was used to quantify the abundance of each tag. After elimination of the linker sequences, other potential artifacts and repeated ditags, each tag was analyzed.

Generation of 5' SAGE library. Oligo-capping was done as described by Maruyama and Sugano⁷ with some modifications¹⁶. Briefly, 5–10 μ g of poly(A)+

RNA were treated with 1.2 units of bacterial alkaline phosphatase (BAP; TaKaRa) in 100 μ l of 100 mM Tris-HCl (pH 8.0) containing 5 mM 2-mercaptoethanol and 100 units of RNasin (Promega) at 37 °C for 40 min. After extraction twice with phenol:chloroform (1:1) and ethanol precipitation, the poly(A)⁺ RNA was treated with 20 units of tobacco acid pyrophosphatase (TAP) in 100 μ l of 50 mM sodium acetate (pH 5.5), 1 mM EDTA, 5 mM 2-mercaptoethanol and 100 units of RNasin at 37 °C for 45 min. After phenol:chloroform extraction and ethanol precipitation, 2–4 μ g of the BAP-TAP-treated poly(A)⁺ RNA were divided into two pools, and one of the following RNA linkers containing recognition sites for *XhoI/MmeI* was ligated to each pool—5'-oligo 1 (5'-UUU GGA UUU GCU GGU GCA GUA CAA GGC UUA AUA CUC GAG UCC GACG -3') or 5'-oligo 2 (5'-UUU CTG CUC GAA UUC AAG CUU CUA ACG AUG UAC GCU CGA GUC CGA CG -3')—using 250 units of RNA ligase (TaKaRa) in 100 ml of 50 mM Tris-HCl (pH 7.5), 5 mM MgCl₂, 5 mM 2-mercaptoethanol, 0.5 mM ATP, 25% PEG8000 and 100 units of RNasin at 20 °C for 3–16 h.

After removing unligated 5'-oligo, cDNA was synthesized with RNaseH-free reverse-transcriptase (Superscript II, Gibco BRL). For the 5'-end-enriched cDNA library, 10 pmol of random adapter-primer (5'-GCG GCT GAA GAC GGC CTA TGT GGC CNN NNC-3') was used and incubation was carried out at 12 °C for 1 h and 42 °C for another hour.

After first-strand synthesis, RNA was degraded in 15 mM NaOH at 65 °C for 1 h. The cDNA that was made from 10 μ g of oligo-capped poly(A)⁺ RNA was amplified in a volume of 100 μ l using an XL PCR kit (Perkin-Elmer) with 16 pmol of 5' (5' biotin- GGA TTT GGT GGT GCA GTA CAA CTA GGC TTA ATA-3' or 5' biotin- CTG CTC GAA TTC AAG CTT CTA ACG ATG TAC G-3') and 3' (5'-GCG GCT GAA GAC GGC CTA TGT-3') PCR primers. Random adapter-primer-primed cDNA was amplified for 10 cycles at 94 °C for 1 min, 58 °C for 1 min and 72 °C for 2 min. PCR products were extracted with phenol:chloroform (1:1) once, ethanol-precipitated and digested with the *MmeI* type-IIS restriction endonuclease (University of Gdansk Center for Technology Transfer). Digestion was performed at 37 °C for 2.5 h using 40 units *MmeI* in 300 μ l of 10 mM HEPES, pH 8.0, 2.5 mM potassium acetate, 5 mM magnesium acetate, 2 mM DTT and 40 μ M S-adenosylmethionine. The digested 5'-terminal cDNA fragments were bound to streptavidin-coated magnetic beads (Dyna). cDNA fragments that bound to the beads were directly ligated together in 16 μ l of reaction mixture containing four units of T4 DNA ligase in the supplied buffer for 2.5 h at 16 °C. The ditags were amplified by PCR using primers 5'-GGA TTT GGT GGT GCA GTA CAA CTA GGC-3' and 5'-CTG CTC GAA TTC AAG CTT CTA ACG ATG-3'. The PCR products were analyzed by PAGE and digested with *XhoI*. The band containing the ditags was excised and self-ligated to produce long concatemers which were then cloned into the *XhoI* site of pZero 1.0 (Invitrogen). Colonies were screened with PCR using M13 forward and reverse primers. PCR products containing inserts of more than 600 bp were sequenced with the Big Dye terminator ver.3 and analyzed using a 3730 ABI automated DNA sequencer (Applied Biosystems). All electropherograms were reanalyzed by visual inspection to check for ambiguous bases and to correct misreads. SAGE 2000 software (version 4.12) was used to quantify the abundance of each tag. In this study, we obtained 18–19 bp tag information. 5' SAGE and 3' SAGE data are available at <http://sage.gi.k.u-tokyo.ac.jp/>.

Generation of captured full-length cDNA library. The captured full-length cDNA library in HEL293 was constructed using the oligo-capping method as described before¹⁶. The 5' ends of 7,903 clones sequenced were analyzed as well as 5' SAGE tags. The raw sequencing data are deposited in DDBJ and are also available as a **Supplementary Note** online and at <http://sage.gi.k.u-tokyo.ac.jp/>.

Association of 5' SAGE tag with corresponding genes. To assess the validity of 5' SAGE tags for identifying transcriptional start sites, we avoided aligning 5' SAGE tags (18 bp) with the current cDNA/EST database because the sequences were not always read from their transcriptional start sites. We instead attempted to align our 5'-tags with the human genome sequence, the NCBI build 34 available from <http://genome.ucsc.edu/>, by using the alignment program ALPS that is publicly available from <http://alps.gi.k.u-tokyo.ac.jp/>. Only tags that matched in the sense orientation were considered for this analysis.

The regions adjacent to the alignment location of each 5'-tag were searched to find their corresponding transcript by using the Gene Resource Locator database¹⁷, a database of sequence alignments from various resources such as UniGene (Build 162)¹⁸. A major problem that we ran into was that, owing to

retro-transposition and genome-duplication, one 5'-tag could be aligned with multiple locations, though many of these were noncoding regions. The issue was resolved by selecting gene-coding locations that were annotated in the UniGene database. Although 3'-tags often fell in the 3'-end exons, 5'-tags did not necessarily hit the first exons. Thus the search was made within 500 bp from the alignment position of each 5'-tag.

Consistency with known 5' transcriptional start sites. A small distance between the aligned positions of each 5' SAGE tag and its corresponding gene implied that the 5' tag was nearly consistent with the known 5' transcriptional start site. To calculate the distance, however, we had to bear in mind that, near the 5' tag, multiple cDNA/EST sequence alignments might be observed because of alternative splicing. To resolve this problem and assign a unique value to the distance, we selected the alignment that was closest to the 5' tag. The distance was defined as being negative if the 5' tag was located in the upstream region of its corresponding cDNA. Otherwise, the value was deemed positive or zero; the zero distance indicated the perfect coincidence. To observe the overall distance distribution, we calculated the total number of 5' SAGE tag occurrences between -500 to +200 nt of the mRNA start sites. We used RefSeq, UniGene(GRL) and DBTSS databases separately to determine differences in their coverage of the transcriptional start sites.

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS

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

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Received 20 January; accepted 7 June 2004

Published online at <http://www.nature.com/naturebiotechnology/>

- Duggan, D.J., Bittner, M., Chen, Y., Meltzer, P. & Trent, J.M. Expression profiling using cDNA microarrays. *Nat. Genet.* **21**, 10–14 (1999).
- Saha, S. *et al.* Using the transcriptome to annotate the genome. *Nat. Biotechnol.* **20**, 508–512 (2002).
- Madden, S.L., Galella, E.A., Zhu, J., Bertelsen, A.H. & Beaudry, G.A. SAGE transcript profiles for p53-dependent growth regulation. *Oncogene* **15**, 1079–1085 (1997).
- Velculescu, V.E. *et al.* Analysis of human transcriptomes. *Nat. Genet.* **23**, 387–388 (1999).
- Hashimoto, S. *et al.* Gene expression profile in human leukocytes. *Blood* **101**, 3509–3513 (2003).
- Lander, E.S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
- Maruyama, K. & Sugano, S. Oligo-capping: a simple method to replace the cap structure of eukaryotic mRNAs with oligoribonucleotides. *Gene* **138**, 171–174 (1994).
- Suzuki, Y. *et al.* DBTSS: DataBase of human Transcriptional Start Sites and full-length cDNAs. *Nucleic Acids Res.* **30**, 328–331 (2002).
- Suzuki, Y. *et al.* Diverse transcriptional initiation revealed by fine, large-scale mapping of mRNA start sites. *EMBO Rep.* **2**, 388–393 (2001).
- Pauws, E., van Kampen, A.H., van de Graaf, S.A., de Vijlder, J.J. & Ris-Stalpers, C. Heterogeneity in polyadenylation cleavage sites in mammalian mRNA sequences: implications for SAGE analysis. *Nucleic Acids Res.* **29**, 1690–1694 (2001).
- Shiraki, T. *et al.* 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 (2003).
- Modrek, B. & Lee, C. A genomic view of alternative splicing. *Nat. Genet.* **30**, 13–19 (2002).
- Krawczak, M., Reiss, J. & Cooper, D.N. The mutational spectrum of single base-pair substitutions in mRNA splice junctions of human genes: causes and consequences. *Hum. Genet.* **90**, 41–54 (1992).
- Zavolan, M. *et al.* Impact of alternative initiation, splicing, and termination on the diversity of the mRNA transcripts encoded by the mouse transcriptome. *Genome Res.* **13**, 1290–1300 (2003).
- Hashimoto, S.-I., Suzuki, T., Dong, H.-Y., Yamazaki, N. & Matsushima, K. Serial analysis of gene expression in human monocytes and macrophages. *Blood* **94**, 837–844 (1999).
- Suzuki, Y., Yoshitomo-Nakagawa, K., Maruyama, K., Suyama, A. & Sugano, S. Construction and characterization of a full length-enriched and a 5'-end-enriched cDNA library. *Gene* **200**, 149–156 (1997).
- Honkura, T., Ogasawara, J., Yamada, T. & Morishita, S. The Gene Resource Locator: gene locus maps for transcriptome analysis. *Nucleic Acids Res.* **30**, 221–225 (2002).
- Wheeler, D.L. Database Resources of the National Center for Biotechnology. *Nucleic Acids Res.* **31**, 28–33 (2003).