Sequence analysis

A large quantity of novel human antisense transcripts detected by LongSAGE

Xijin Ge1, Qingfa Wu1, Yong-Chul Jung1, Jun Chen1 and San Ming Wang1,2,*

1Center for Functional Genomics, Division of Medical Genetics, Department of Medicine, ENH Research Institute and 2Robert H. Lurie Comprehensive Cancer Center, Northwestern University Feinberg School of Medicine, 1001 University Place, Evanston, IL 60201 USA

Received on April 11, 2006; revised on July 5, 2006; accepted on August 1, 2006
Advance Access publication August 7, 2006
Associate Editor: Nikolaus Rajewsky

ABSTRACT

Motivation: Taking advantage of the high sensitivity and specificity of LongSAGE tag for transcript detection and genome mapping, we analyzed the 632 813 unique human LongSAGE tags deposited in public databases to identify novel human antisense transcripts.

Results: Our study identified 45 321 tags that match the antisense strand of 9804 known mRNA sequences, 6606 of which contain antisense ESTs and 3198 are mapped only by SAGE tags. Quantitative analysis showed that the detected antisense transcripts are present at levels lower than their counterpart sense transcripts. Experimental results confirmed the presence of antisense transcripts detected by the antisense tags. We also constructed an antisense tag database that can be used to identify the antisense SAGE tags originated from the antisense strand of known mRNA sequences included in the RefSeq database.

Conclusions: Our study highlights the benefits of exploring SAGE data for comprehensive identification of human antisense transcripts and demonstrates the prevalence of antisense transcripts in the human genome.

Contact: swang1@northwestern.edu

Supplementary information: Supplementary data are available at Bioinformatics online

1 INTRODUCTION

It is well-known that the genomic loci coding for genes can be transcribed from the opposite direction, resulting in the generation of antisense transcripts (Simons et al., 1988). Antisense transcripts play important roles in many biological processes including imprinting (Lee et al., 1999) and regulation of gene expression (Novina and Sharp, 2994). Comprehensive identification of antisense transcripts will enhance our understanding of biological basics, and provide tools for functional interference of processes. Large-set expression sequence data have been used to identify antisense transcripts expressed in several genomes (Shendure and Church, 2002; Fahey et al., 2002, Lehner et al., 2002; Yelin et al., 2003; Chen et al., 2004; Katayama et al., 2005; Siddiqui et al., 2005; Riken Group, 2005).

Many lower abundant transcripts in the genomes analyzed so far have not been identified (Bishop et al., 1974; Chen et al., 2002; Lee et al., 2005). Therefore, the antisense transcripts identified from the currently known transcripts might not be comprehensive. Exploring various transcript resources could identify more antisense transcripts present at low abundant levels. One of these transcript resources is the SAGE (serial analysis of gene expression; Velculescu et al., 1995) data. SAGE collects short tags near the 3’ end of transcripts. The orientation of the detected transcript is preserved in the SAGE tag sequence; therefore, SAGE tags can distinguish sense and antisense transcripts. Moreover, the high sensitivity of SAGE over other approaches leads to the detection of more transcripts expressed at lower-abundant levels (Saha et al., 2002). Therefore, exploring SAGE data might lead to the identification of more antisense transcripts. Indeed, analyzing mouse SAGE tags has resulted in the identification of mouse antisense transcripts (Quere et al., 2004; Wahl et al., 2005).

In this study, we explored the possibility of using 21 bp LongSAGE tags in the public database to identify novel human antisense transcripts. Compared with the standard 14 bp SAGE tags, the 21 bp LongSAGE tags provide high specificity for mapping in the human genome sequences (Saha et al., 2002). Thus, the genome sequences are used as the hub to compare the existing known transcripts and SAGE tags in order to identify the antisense transcripts detected by SAGE tags. Our analysis identified a substantial number of novel antisense transcripts in the human genome. Here we report the results from this study.

2 METHODS

2.1 Identification of antisense SAGE tags

The LongSAGE tags from 29 human LongSAGE libraries were used for the study (http://www.ncbi.nlm.nih.gov/geo/). All tags were matched through BLAT to the human genome sequences (Kent, 2002; http://genome.ucsc.edu/, HG17, May 2004). The NlaIII cutting site, ‘CATG’, was added to the 5’ end of each 17 bp tag before mapping. For such mapping we used a tile (word) size of 12 and step size of 6. No gap or mismatch was allowed. Tags uniquely mapped to the human genome were identified and used for the study. The information for each mapped tag, including the mapping location, annotation and transcripts of mRNA and EST, was incorporated into a MySQL database. The uniquely mapped SAGE tags were searched against mRNA and EST sequences in the mapped genomic loci. The orientation of mRNA and EST in the genome is based on the mapping information in UCSC genome browser. If a tag matched to an exon or intron of a well-annotated mRNA sequence, but for the opposite direction, the tag was classified as an antisense tag; if a tag matched to an EST aligning to the
well-annotated mRNA but in the opposite orientation, the tag was classified as the known antisense tag. A tag matching in the opposite direction of an mRNA sequence but without matching to a known antisense EST was defined as the novel antisense tag representing a novel antisense transcript detected only by the SAGE tag.

2.2 Identification of sense and antisense pairs for the known mRNAs

The presence of the same UniGene cluster indicated the presence of sense–antisense pairs. A UniGene ID for each mRNA sequence matched by antisense tag was collected from the UniGene database (UniGene release 185). A UniGene ID for each antisense tag was collected from the SAGEmap database (Lai et al., 1999). The version used for this analysis was based on the UniGene release 185. Only the SAGE tag that linked to a single UniGene cluster was used for the analysis. The copy numbers of the sense tags and antisense tags were used for quantitative comparisons.

2.3 Confirmation of antisense transcripts detected by antisense tags

Fetal brain total RNA (Clontech) was treated with RNAse-free DNase I to eliminate genomic DNA contamination (Biolabs, USA). The antisense tag matched mRNA sequences were used for primer design, and strand-specific RT–PCR was used to detect the sense or antisense transcripts (Lee et al., 1999). Primers were designed with the Primer3 software (Rozen and Skaltsky, 2000) to amplify ~100 bp sequences spanning the LongSAGE tag. To detect the sense transcripts, the antisense primer was used for reverse transcription; to detect the antisense transcripts, the sense primer was used for reverse transcription. Reverse transcriptions were performed using Superscript Reverse Transcriptase (Invitrogen). Upon reverse transcription, the corresponding antisense or sense primers were added to the reaction for PCR amplification. A positive control set for each reaction used the oligo dT primer for generating cDNA and sense–antisense primers for PCR amplification. A negative control for each reaction was the same as the positive control except no oligo dT and sense–antisense primers were included for the reactions. PCR products were visualized on 2% agarose gels.

2.4 Construction of antisense SAGE tag database

Sequences in the RefSeq database were downloaded (Pruitt and Tatusova, 2005; http://www.ncbi.nlm.nih.gov/RefSeq/). Each sequence was converted to its virtual antisense transcript by reverse complementary. A 21 bp tag including CATG was extracted after each CATG site in the virtual antisense sequence. The virtual tags were matched to known transcripts in the SAGEmap database to determine whether the tags matched known EST or whether the tag was a novel antisense tag. The virtual tags were also mapped to the genome sequences to eliminate the tags mapped to more than one location in the genome. These virtual antisense tags were then used to form a virtual antisense SAGE tag database to represent the antisense transcripts corresponding to each sequence included in the RefSeq database. The database provides related information for each virtual antisense tag, including the GenBank ID and the chromosome and genome sequence location for each mRNA sequence.

2.5 Homologous antisense pairs in human and mouse

The HomoloGene database was downloaded from NCBI (ftp://ftp.ncbi.nih.gov/pub/HomoloGene/; October 19, 2005 version). This database was searched for mouse homolog genes of our 9804 RefSeq mRNAs that are the target of antisense SAGE tags according to our analysis. This led to 8459 mouse genes identified by RefSeq IDs. The 1260 mouse antisense genes identified by Wahl et al. (2005), given in ENSEMBL gene IDs, were converted to 834 RefSeq IDs. These IDs were then searched against the 8459 genes whose human homologs were found to be targets of antisense tags in our analysis. Similarly, the list of 9713 mouse antisense transcripts identified by the RIKEN Group (2005) was downloaded from the RIKEN website (http://fantom31p.gsc.riken.jp/s_as/). Conversion of the cDNA clone IDs and mRNA IDs led to 3166 RefSeq transcript IDs. The overlap of this list with our list of 8459 genes was calculated.

3 RESULTS

3.1 Identification of SAGE tags representing antisense transcripts

The 3.6 million copies of LongSAGE tags were used for the study. These tags were collected from 29 LongSAGE libraries containing 632 813 unique SAGE tags (Supplementary Table 5). From these unique SAGE tags, 45 321 tags were identified to match the antisense strand of the mRNA sequences. Figure 1 shows a typical example of the sense–antisense tags detected for the gene ‘WD repeat domain 48’. Of these mappings, 48% of tags mapped to exons and 52% mapped to introns of the mRNA sequences, indicating no obvious preference of antisense tags for the exon or intron of the mapped genes. For the 45 321 antisense tags, 20 316 tags matched ESTs that map to the mRNAs but in the opposite orientation. These tags represent the antisense transcripts already detected by the ESTs. The remaining 25 005 tags represent novel antisense transcripts for the mRNA sequences. There were 277 642 mRNAs matched by the 45 321 antisense tags. These ‘sense’ transcripts are distributed in 13 207 UniGene clusters, of which 8 036 clusters contain the antisense ESTs and 5171 clusters contain the novel antisense SAGE tags (Table 1, Supplementary Tables 1 and 2). When matched to RefSeq database, the 45 321 tags are found to map as antisense to 9804 sequences; 6606 of these contain antisense ESTs and 3198 are mapped only by SAGE tags.

3.2 Expression levels of antisense transcripts are lower than that of sense transcripts

In a given SAGE library, each SAGE tag has a defined copy number representing the expression levels of the corresponding transcripts. Comparison of the copy number between the sense tags and the antisense tags in the same SAGE library provides a quantitative measure of the expression levels of the sense transcripts and antisense transcripts. The fetal brain LongSAGE library was used for the analysis (GSM31935, http://www.ncbi.nlm.nih.gov/projects/geo/). This library has 2259 UniGene clusters that contain 7132 ‘sense’ SAGE tags with 55 824 copies that match mRNA sequences in the same orientation, with an average of 3.2 copies per UniGene cluster or 7.8 copies per tag; in comparison, there are 2929 ‘antisense’ SAGE tags with 4280 copies that match mRNA sequences in the opposite orientation with an average of 1.5 copies per UniGene cluster or 1.5 copies per tag. The antisense tag accounts for 41% of total unique tags, but only contributes 8% of the total copies (Table 2, Supplementary Table 3). This indicates that the antisense transcripts are expressed at levels lower than the sense transcripts. Further comparison of the copy numbers between the antisense tags with matches to ESTs and novel antisense tags shows that the antisense tags with matches to ESTs have 1.8 copies per tag whereas the novel tags have 1.2 copies per tag. This difference reflects the fact that the novel antisense tags represent the lower abundant antisense transcripts not detected by ESTs but detected by SAGE.
3.3 Experimental confirmation of antisense transcripts detected by antisense tags

Thirty antisense tags were selected for experimental confirmation, including 20 novel antisense tags and 10 antisense tags matched to antisense ESTs. The fetal brain LongSAGE library collected >300,000 copies of SAGE tags. The novel antisense tags were selected from those in this library with 2–4 copies to represent the low-abundant antisense transcripts. For the novel antisense tags, 15 out of 20 were detected; for the antisense tags with EST matches, all were detected (Figure 2, Supplementary Table 4). The negative detection for some reactions in both types of candidate tags likely reflects the low-efficiency of specific primers for reverse transcription, as compared with amplification in all positive controls that used cDNA from oligo dT priming. Another factor could be that the RNA sample used for RT–PCR amplification is different from the RNA samples used for long SAGE tag collection.

3.4 Homologous antisense genes in human and mouse

Two previous studies identified mouse genes with antisense transcripts (Wahl et al., 2005; Riken Group, 2005). We performed analysis to determine how many of these genes are homologous to our list of 9804 human transcripts that are found to be target of antisense SAGE tags. We found that 517 (62%) of the 834 genes identified by Wahl et al. (2005) and 1588 (50%) of the 3166 genes identified by RIKEN are homologous to genes in our list (Table 3. Supplementary Tables 1 and 2). Many functionally important genes were identified in the homolog gene list, such as the breast cancer 2, early onset (BRCA2) gene. Germ line mutations in BRCA2 have been linked to an elevated risk of young onset breast cancer (Wooster et al., 1994). RIKEN cDNA data include two antisense transcripts (AK133262 and Ak138160) targeting the 20th intron and last exon, respectively. LongSAGE tags are identified as antisense tags (CTGACTTAAGGATGAAG and CACACACACATTGATTC) mapped to the same regions (Fig. 3). Conservation of antisense transcripts to the BRCA2 gene targeting the same region suggests that these antisense transcripts might play roles in regulating BRCA2 gene function.
Two RIKEN full-length cDNA sequences, AK133262 and AK138160, are antisense to the mouse BRCA2 gene. The location of two antisense SAGE tags (Tag 1 and Tag 2) in the human BRCA2 gene. One antisense EST sequence (BG562419) overlaps with Tag 2. Two RIKEN full-length cDNA sequences, AK133262 and AK138160, are antisense to the mouse BRCA2 gene.

3.5 Database for annotating antisense tags

In a typical SAGE library, a portion of SAGE tags may be derived from antisense transcripts. To facilitate the annotation of the SAGE tags representing antisense transcripts, an antisense tag database was constructed that contains 185 017 virtual SAGE tags matched to antisense of well-annotated mRNA sequences included in the RefSeq database. Each virtual tag has a single mapping location in the human genome. An experimental SAGE tag that matches a virtual tag in the database will indicate that this tag represents an antisense transcript to a sequence in the RefSeq database. The related information for the identified transcripts is also included in the database, including the chromosome origin and sequence location. Testing the database with the 45 321 antisense tags identified in this study shows that 46% of the antisense tags are included in this database, including 31% of the 25 005 novel antisense tags and 64% of the 20 316 antisense tags with EST matches. Table 4 shows a group of representative experimental SAGE tags identified as the antisense tags by this database. The full database is included in the Supplementary Table 6.

4 DISCUSSION

Taking advantage of the high sensitivity of SAGE for transcript detection and the high specificity of LongSAGE tags mapping to the genome, our study identified 45 321 antisense tags for 9804 mRNAs. Comparison with those identified in previous studies shows that 96 (67%) of the 144 pairs identified by Shendure and Church (2002), 1931 (72%) of the 2667 identified by Yelin et al. (2003) and 2248 (82%) of the 2736 identified by Chen et al. (2004) are included in our list. The number identified in this study is twice the number of antisense transcripts identified by previous studies, indicating that antisense is far more prevalent in the human genome than previously considered.

Considering the complexity of SAGE library construction, some of the antisense tags might be owing to artifacts, and short 21 bp tags are less specific than longer sequences such as EST and cDNA sequences. Situation may also exist that a LongSAGE tag named as novel may not be real novel, as it could map to a different location of a gene that already have antisense EST evidence. Currently, there is no consensus in defining the sense/antisense pairs. Some previously identified sense/antisense pairs, such as BC005542 and BC062929 identified by Katayama et al. (2005), consist of two long genes that are transcribed from opposite strands with small overlapping in their untranslated regions (UTRs). In our analysis, a tag is classified as an ordinary sense tag if it matches the sense strand of any mRNA. So our list of antisense tags should not include tags that mapped to such regions. Nonetheless, we cannot rule out the possibility that some antisense tags could map just outside such regions. Therefore, experimental verification will be needed to confirm the existence of each antisense tag. However, considering the consistency with the existing antisense human transcripts, homology to mouse genes with antisense transcripts and experimental detection of antisense transcripts for the candidate tags, it is likely that most of the antisense transcripts detected by SAGE tags exist.

As shown in Table 1, a significant portion of total unique tags (about two-thirds of 632 813) could not be mapped to the genome. Further investigation shows that most of these non-mapped tags were detected only once (singleton). For example, 64% of total unique tags in the fetal brain library are singleton, of which only 30% can be mapped to the genome. In contrast, 80.5% non-singleton tags can be mapped to the genome. One of the causes could be related with DNA sequencing error. SAGE tags are collected by single-pass DNA sequencing reaction that is prone to errors. With the same error rate per base pair, Long SAGE tags will suffer a higher error rate than conventional 10 bp SAGE tags owing to its longer sequences. The proportion of erroneous tags further increases when multiple SAGE libraries are pooled together, as multiple genuine tags representing the same transcripts in different libraries will overlap while the erroneous tags do not. In addition, the origins of the reference human genome sequences and the SAGE tags were from different individual genomes. The differences of individual genomes, such as structural variation and SNPs, in human population could also affect the mapping result (Feuk et al., 2006).

Most genes in the genome are expressed at low-abundant levels. This could have significant implications with regard to the antisense transcripts because, as shown in our study, the antisense tags are largely present at low copy levels. Similar observation was also made in mouse antisense SAGE tags (Wahl et al., 2005). Because of its high sensitivity, SAGE detects more low-abundant transcripts, and therefore more antisense transcripts. SAGE data provide a useful resource for antisense transcript identification. With the increasing collection of LongSAGE tags, more potential antisense transcripts might be identified to provide more comprehensive coverage of antisense transcripts expressed in the genome.
quantitative differences among different transcripts vary over six orders (Holland, 2002), and SAGE cannot detect all transcripts. Many low copy SAGE tags are clearly not the collection of sequencing errors but represent the true transcripts. With the improvement of sequencing technologies, more low-abundant transcripts would be detected, and more potential antisense transcripts could be detected. The identification of antisense SAGE tags provides evidence for the presence of antisense transcripts. Using a combination of experimental and bioinformatics approaches, we need to identify the original antisense transcripts in the form of 3' ESTs or full-length cDNAs for genome annotation and functional study.

ACKNOWLEDGEMENTS

The authors appreciate the availability of the LongSAGE data deposited in the Gene Expression Omnibus for this study. The authors thank Yeong Cheol Kim and Wendy S. Rubinstein for discussions, and W. James Kent for suggestions on the usage of BLAT program. This study was supported by National Institute of Health, US Department of Defense, Daniel F. and Ada L. Rice Foundation, and Evanston Northwestern Healthcare Breast/Ovarian Research Program.

Conflict of Interest: none declared.

REFERENCES

Bishop,J.O. et al. (1994) Three abundance classes in HeLa cell messenger RNA. Nature, 368, 199–204.