BIOINFORMATICS ORIGINAL PAPER

Vol. 21 no. 8 2005, pages 1389–1392
doi:10.1093/bioinformatics/bti205

Genome analysis

LongSAGE analysis revealed the presence of a large number of novel antisense genes in the mouse genome

Matthias B. Wahl[1,†], Ulrich Heinzmann[2] and Kenji Imai[1,*]

[1]Institute of Developmental Genetics and [2]Institute of Pathology, GSF-National Research Center for Environment and Health, Ingolstädter Landstrasse 1, D-85764 Neuherberg, Germany

Received on September 4, 2004; revised on November 23, 2004; accepted on December 2, 2004

Advance Access publication December 7, 2004

ABSTRACT

Motivation: Despite the increasing notions of the functional importance of antisense transcripts in gene regulation, the genome-wide overview on the ontology of antisense genes has not been obtained. Therefore, we tried to find novel antisense genes genome-wide by using our LongSAGE dataset of 202 015 tags (consisting of 41 718 unique tags), experimentally generated from mouse embryonic tail libraries.

Results: We identified 1260 potential antisense genes, of which 1001 are not annotated in EnsEMBL, thereby being regarded as novel. Interestingly their sense counterparts were co-expressed in the majority of the cases.

Conclusions: The use of LongSAGE transcriptome data is extremely powerful in the identification of thus-far unknown antisense transcripts, even in the case of well-characterized organisms like the mouse.

Contact: imai@gsf.de

INTRODUCTION

Control of gene expression by naturally occurring antisense transcripts has been discovered in prokaryotes more than 20 years ago (reviewed by Simons, 1988). In recent years, global gene expression surveys by microarrays (Yamada *et al*., 2003), as well as computational analyses based on full-length cDNA (Kiyosawa *et al*., 2003) or expressed sequence tag (EST) sequences (Shendure and Church, 2002) suggested that the genomes of higher vertebrates also contain a substantial number of genes that harbor oppositely oriented overlapping transcripts. However, the number of well-characterized antisense genes is still small (Vanhee-Brossollet and Vaquero, 1998). Therefore, the genome-wide overview on the ontology of antisense genes (i.e. how many genes are associated with their antisense genes) remains largely unknown. Furthermore, it cannot be ruled out that at least some of the identified sense–antisense gene pairs in the above-cited papers are artifacts, owing to experimental errors intrinsic to the corresponding approach. For example, hybridization-based approaches are sensitive to cross-hybridization, and the alignment of transcript sequences to the genome can be erroneous, owing to the wrong annotation of the transcript orientation. Furthermore, around half of the antisense transcripts are single-exon genes, and therefore the correct orientation cannot be confirmed by the analysis of canonical splice sites (Kiyosawa *et al*., 2003). Thus, most approaches are only qualitative, and do not evaluate the quantitative aspect of sense/antisense transcript pairs in a certain tissue.

Serial Analysis of Gene Expression (SAGE) (Velculescu *et al*., 1995) is a promising method to evaluate antisense transcripts in a genome-wide scale. For a SAGE analysis, a priori knowledge of transcript sequences is not required, which is an intrinsic and principle advantage of SAGE over microarray approaches. Furthermore, the SAGE method is far more effective than EST sequencing (Sun *et al*., 2004) and supplies quantitative gene expression data as digital counts. In the present study, we analyzed a dataset of 202 015 LongSAGE tags generated from mouse embryonic tails and examined it in comparison with public cDNA and EST datasets to determine and evaluate antisense genes in the mouse genome.

MATERIALS AND METHODS

Tissue dissection

Tails of 268 stage-matched E10.5 mouse embryos of C57BL/6N origin (Charles River) were microdissected by four sagittal cuts into four tissue parts, including the tail bud, the caudal two-third or the rostral one-third of the presomitic mesoderm and two pairs of most recently formed somites. The collected tissues were homogenized in the Binding/Lysis buffer (Dynal) and immediately stored at −80°C until use.

LongSAGE library construction

Poly(A)-positive RNA was isolated using the mRNA DIRECT kit (Dynal), bound to oligo(dT)25 magnetic beads, and immediately proceeded to LongSAGE library construction, according to the standard protocol (Saha *et al*., 2002), with our-own modifications (Wahl *et al*., 2004). For each LongSAGE library construction, distinct linker/primer combinations were used in order to avoid cross-contaminations of LongSAGE tags between the libraries (see Supplementary Table). Sequencing was performed with big-dye terminators on an ABI 3100 DNA Analyzer (Applied Biosystems).

LongSAGE data acquisition

By using Phred (Ewing *et al*., 1998), extracted LongSAGE tags were considered only when all bases had a Phred score of 10 or higher. LongSAGE tag sequences were further assessed by SAGEScreen (Akmaev and Wang, 2004). The complete LongSAGE tag dataset used in this study is deposited at the Gene Expression Omnibus database at NCBI with the accession number GSM26978.

Tag-to-gene mapping

Assignments of obtained LongSAGE tags to corresponding genes were carried out, using information from public databases such as the UniGene, Mouse

*To whom correspondence should be addressed.

†Present address: Stowers Institute for Medical Research, 1000 E. 50th Street, Kansas City, MO 64110, USA

Genome Database (MGD) and EnsEMBL, and described in detail in the accompanying paper (Wahl *et al*., 2005).

Identification of antisense transcripts

All cDNA and EST sequences overlapping with a LongSAGE tag were retrieved from the EnsEMBL databases: i.e. mus_musculus_core_19_30, mus_musculus_estgene_19_30 and mus_musculus_est_19_30, which were derived from the NCBI mouse genome assembly (build 30). Sequences were processed further only when canonical splice donor and/or acceptor sites could be identified immediately flanking its HSP (transcript sequence against the genome), or when the whole sequence is aligned to the genomic sequence to take single-exon genes into account. Next, all transcripts overlapping with a LongSAGE tag were compared against EnsEMBL and EST genes. In the case where a LongSAGE tag was found on the opposite strand within a 10 kb region of a EnsEMBL or EST gene, the LongSAGE tag/transcript sequence pair was considered to represent a antisense gene, according to the following criteria: (1) the transcript sequence(s) also overlapped with an exon of the EnsEMBL gene (following splicing rules), or (2) they shared a minimum percentage identity of 95% over at least 150 bp. Antisense LongSAGE tags assigned to the same sense EnsEMBL gene were considered as being derived from the same antisense gene.

Quantification of gene expression for sense and antisense tags

The expression level of each of the sense or antisense genes listed in Table 3 was determined as the sum of counts of all alternative LongSAGE tags derived from the single gene.

RESULTS

LongSAGE library construction

C57BL/6 mice were used in this study, so that LongSAGE tag sequences could be best compared to mouse genome sequences that were mostly of C57BL/6 origin. The tail of E10.5 mouse embryos was microdissected into four different parts, and mRNA was extracted and used for generating a total of eight LongSAGE libraries: two independent libraries from each of the four tail parts. From each library, excluding linker tags and duplicate ditags, around 25 000 tags were sequenced and a total of 202 015 tags were collected, corresponding to 41 714 unique tags (Table 1). Although a very small amount of starting RNA material required a slightly increased number of PCR cycles for ditag amplification (exactly 36 cycles for each library), the frequency of duplicate ditags (1–4%) was low. Furthermore, GC-content (Margulies *et al*., 2001b) as well as the incidence of linker tags (0.5–4%) is comparable to other good SAGE libraries (Margulies *et al*., 2001a), and a comparison of the pairs of LongSAGE libraries generated from the same tissue part showed a high reproducibility, pointing to the good quality and consistency of the LongSAGE libraries.

Identification of sense–antisense gene pairs

To minimize false-positive cases, we determined only those antisense transcripts, which were supported by LongSAGE tag as well as cDNA and/or EST sequences. Therefore, ESTs overlapping with the genomic position of a LongSAGE tag were determined and considered to be corresponding to the same transcript as the LongSAGE tag, only when they could be aligned to the genome on the same strand following splicing rules (Fig. 1). The LongSAGE tag/transcript pair was considered to be in an antisense orientation to a EnsEMBL gene, if the LongSAGE tag was located either within an exon of the sense EnsEMBL gene (Fig. 1A), or the associated transcript overlapped outside the LongSAGE tag with the EnsEMBL gene (Fig 1B). As summarized in Table 2, among a total of 18 205 transcript-verified LongSAGE tags, a total of 1468 were in an antisense orientation to an annotated EnsEMBL gene, suggesting that these antisense LongSAGE tags represented potential antisense genes.

Table 1. Overview of combined LongSAGE libraries

Total tags excluding linker sequences and dupl. ditags	202 015
Unique tags	41 714
With count >1	12 508

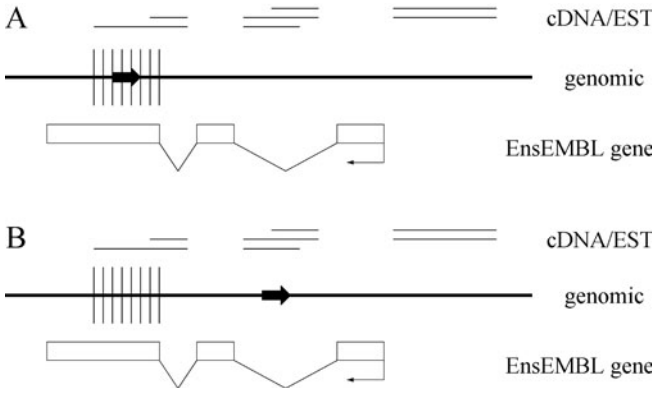

Fig. 1. Schematic drawings to explain how antisense transcripts were identified: only LongSAGE tags supported by cDNA and/or EST sequences were used for determining antisense transcripts. Either (**A**) the LongSAGE tag was found on the reverse complementary strand of an exon of an EnsEMBL gene or (**B**) the LongSAGE tag was located outside an exon of an EnsEMBL gene, but a transcript (cDNA/EST) sequence aligned to the LongSAGE tag overlapped with at least one exon of an EnsEMBL gene on the reverse complementary strand. The transcriptional orientation of an EnsEMBL gene is indicated by the thin arrow. The vertical bars indicate the overlap between the sense and antisense genes. Bold arrows: LongSAGE tags; thin, short lines at the top of each panel: cDNA/EST sequences; bold horizontal lines: genomic sequences; and open boxes: exons of an EnsEMBL gene.

Table 2. Sense—antisense gene pairs

Number of antisense genes[a]	1260
Included in EnsEMBL	259
Novel: not included in EnsEMBL	1001
With two or more counts[b]	594
Average tag counts	4.8
Number of antisense tags	1468
Included in EnsEMBL	296
Novel: not included in EnsEMBL	1172
Sense genes expressed	981
Average tag counts	32.1
Sense gene not expressed	279
With two or more counts[b]	135

[a]When more than one antisense tags were identified in an antisense orientation to the same EnsEMBL gene, they were considered to derive from one antisense gene.
[b]For tag counts, all LongSAGE tags for a sense or antisense gene were summed.

Table 3. Twenty most abundant novel antisense transcripts not included in EnsEMBL

Antisense tag[a]	No.[b]	Count[c]	Sense gene tag[a]	No.[b]	Count[c]	EnsEMBL ID of sense gene
CAGATTTAGGTGCTTTC	1	100	TGTGCCAAGTGTGTCCG	2	148	ENSMUSG00000003387
CAGGCTTCCATACCACC	3	58	CCTCCATCCTTTATACT	3	64	ENSMUSG00000020849
TGCAGAAAGGAGGCATA	4	57	TGTCTATGATGCCCTCT	11	2601	ENSMUSESTG00000018151
TTGAACACTTTATGATG	1	53	GCCAACTCTGCCTGACC	3	224	ENSMUSG00000000031
GTGTTGAGGGGTCGGTG	1	52	ATAGTAAGCTTTGAACT	3	53	ENSMUSG00000029581
TCAGGTCATTTTACCGG	4	45	AATGTAAAGGGGACAGC	4	49	ENSMUSG00000020176
CAAGTTCTTTTCCCTCT	1	33	TTGGTGAAGGAAAAAGC	2	322	ENSMUSG00000049775
TTGCTCTCAGCTTCGGT	4	29	GCATCGACCCCACCGGT	4	124	ENSMUSG00000001525
GGAGCAGAGGAGCCCGA	2	28	CTCGCTGTCTCGGATTC	2	13	ENSMUSG00000030274
CACTGGCCTTCCCTCCT	1	27	CTCTGACTTACTGTTGG	1	9	ENSMUSG00000023175
GAGCAAGTTAATTCTCC	1	26	CTGTCTCTAGATCCAGC	4	33	ENSMUSG00000019960
CAACCATCATCTTCCAC	1	22	CTTCTCATTTGTACGTT	4	37	ENSMUSG00000020585
TCACTATAGCAACATCC	3	18	CAGTTCTATATTTTGTT	4	13	ENSMUSG00000051391
TTGGCTACTAGCCCGCG	3	18	CTAATAAAGCCACTGTG	3	190	ENSMUSG00000050299
CCAAAATTAGGAAAAAC	1	18	GCCTGCCCCCTGTGGCC	2	20	ENSMUSG00000037373
CCTCGCACAGTGCGCCA	1	17	AGGTCGGGTGGAAGTAC	2	323	ENSMUSG00000051030
TGCACTGCTGAAAACTC	1	17	GGTGACTGGAGCGCCTT	5	128	ENSMUSG00000050953
GCAATTTGGTGTCTTGC	1	17	ND[d]		0	ENSMUSESTG00000016244
GTGCTGGGATTGACTGG	1	16	GCCGTACACCCACCCTC	2	206	ENSMUSG00000006498
GACATTTAAAACAGCAG	2	15	GTTCAGCAAAGGTTTCC	3	61	ENSMUSG00000027620

[a]Representative antisense and sense tags are shown for a sense/antisense gene pair.
[b]The number of alternative tags for a gene including the listed representative tag is indicated.
[c]Total tag count from all alternative tags observed in this study is shown.
[d]No reliable LongSAGE tag for the indicated EnsEMBL gene was defined. Online Supplementary data for the complete list of the 1260 sense/antisense gene pairs as a tab-delimited text file is available for downloading at the journal website.

Evaluation of sense—antisense gene pairs

We further analyzed the observed antisense LongSAGE tags supported by transcript sequences more in detail. As depicted in Table 2, these 1468 antisense tags correspond to 1260 genes, by considering that different LongSAGE tags antisense to the same sense gene were derived from the same gene. For 78% of the antisense genes, at least one transcript for the sense counterpart was detected in our LongSAGE dataset. In general, the sense genes were expressed at a higher level (average tag count: 32.1) than the antisense genes (average tag count: 4.8). The 20 most abundant antisense transcripts not included in the EnsEMBL gene set are listed in Table 3. Furthermore, a complete list of 1260 sense-antisense gene pairs as an online Supplementary material in a tab-delimited text format is available for downloading from the OUP server.

DISCUSSION

Our analysis has identified 1260 potential antisense genes. The absence of annotated EnsEMBL genes for most of the antisense tags detected in our dataset points out that the EnsEMBL gene annotation pipeline omits most of the genes in antisense orientation to protein-coding genes. Indeed, the number of sense–antisense gene pairs within the EnsEMBL dataset, 217 in human (Shendure and Church, 2002), is dramatically lower than those 2418 pairs observed in mouse full-length cDNA sequences (Kiyosawa *et al*., 2003). Interestingly, the number of potential antisense genes in the LongSAGE libraries is more than half compared to the numbers found in the Riken Fantom2 set. Since only limited types of tissues were analyzed in this study, it is conceivable that the number of antisense genes represented in the Riken Fantom2 dataset still underestimates the quantity of existing antisense genes. This is in accordance with the notion that only half of known sense/antisense gene pairs were detected in the Riken Fantom2 dataset (Kiyosawa *et al*., 2003). Furthermore, a recent global gene expression study in Arabidopsis detected that more than one-third of all genes have an expressed antisense counterpart, of which ∼10% are even co-expressed in the same tissue (Yamada *et al*., 2003). This indicates that a large fraction of eukaryotic genes have an antisense counterpart. It is also interesting that, in the majority of the cases, expression of an antisense gene is associated with the expression of its sense gene, and that the expression level of sense genes is much higher than that of corresponding antisense genes. These observations are consistent with the notions that antisense genes are expressed often at low levels and that their transcripts are often unstable (Storz, 2002).

The biological function of those antisense transcripts might be of great interest. Recently, a major focus in the community is on micro RNAs (miRNAs) that are generated by the successive processing of RNAs to 21–23 base long RNAs, which inhibit the transcription of its targets (reviewed by Bartel, 2004). Yet computational predictions (Lim *et al*., 2003a,b) and experimental cloning (Lagos-Quintana *et al*., 2002) led to the identification of less than 900 miRNAs in all species (miRNA registry) (Griffiths-Jones, 2004), and are therefore outnumbered by the antisense genes mentioned above. However, owing to its short length (∼22 bases) and since precursors of miRNAs do not have to be bidirectionally located on the opposite strand of the genome, many miRNAs will not be captured with the strategy applied. Because of its different structure the identified antisense transcripts might function by a different

mechanism other than that of miRNAs. Studies over the last few years suggest that antisense transcripts function in various different ways, finally regulating or antagonizing its sense counterpart. Cases have been reported in which antisense transcripts affect alternative splicing (Munroe and Lazar, 1991), RNA editing (Kumar and Carmichael, 1997), X-inactivation (Lee and Lu, 1999), translational regulation (Li and Murphy, 2000), imprinting (Sleutels *et al*., 2002) and transcriptional repression by methylation (Tufarelli *et al*., 2003). Experimental validations are required in future studies to define the nature and functions of the potential antisense genes identified in this study.

ACKNOWLEDGEMENTS

We thank Rudi Balling (GBF) for valuable comments to this work. We also thank Ken Kinzler and Victor Velculescu (Johns Hopkins University) for the SAGE protocol and software, and Jerzy Adamski and Peter Lichtner (Genome Analysis Center, GSF) for providing the sequencing facility. This work was supported by the GSF.

REFERENCES

Akmaev,V.R. and Wang,C.J. (2004) Correction of sequence-based artifacts in serial analysis of gene expression. *Bioinformatics*, **20**, 1254–1263.

Bartel,D.P. (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, **116**, 281–297.

Ewing,B., Hillier,L., Wendl,M.C. and Green,P. (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.*, **8**, 175–185.

Griffiths-Jones,S. (2004) The microRNA Registry. *Nucleic Acids Res.*, **32**, 109–111.

Kiyosawa,H., Yamanaka,I., Osato,N., Kondo,S. and Hayashizaki,Y. (2003) Antisense transcripts with FANTOM2 clone set and their implications for gene regulation. *Genome Res.*, **13**, 1324–1334.

Kumar,M. and Carmichael,G.G. (1997) Nuclear antisense RNA induces extensive adenosine modifications and nuclear retention of target transcripts. *Proc. Natl Acad. Sci., USA*, **94**, 3542–3547.

Lagos-Quintana,M., Rauhut,R., Yalcin,A., Meyer,J., Lendeckel,W. and Tuschl,T. (2002) Identification of tissue-specific microRNAs from mouse. *Curr. Biol.*, **12**, 735–739.

Lee,J.T. and Lu,N. (1999) Targeted mutagenesis of Tsix leads to nonrandom X inactivation. *Cell*, **99**, 47–57.

Li,A.W. and Murphy,P.R. (2000) Expression of alternatively spliced FGF-2 antisense RNA transcripts in the central nervous system: regulation of FGF-2 mRNA translation. *Mol. Cell. Endocrinol.*, **170**, 233–242.

Lim,L.P., Glasner,M.E., Yekta,S., Burge,C.B. and Bartel,D.P. (2003a) Vertebrate microRNA genes. *Science*, **299**, 1540.

Lim,L.P., Lau,N.C., Weinstein,E.G., Abdelhakim,A., Yekta,S., Rhoades,M.W., Burge,C.B. and Bartel,D.P. (2003b) The microRNAs of *Caenorhabditis elegans*. *Genes Dev.*, **17**, 991–1008.

Margulies,E.H., Kardia,S.L. and Innis,J.W. (2001a) A comparative molecular analysis of developing mouse forelimbs and hindlimbs using serial analysis of gene expression (SAGE). *Genome Res.*, **11**, 1686–1698.

Margulies,E.H., Kardia,S.L. and Innis,J.W. (2001b) Identification and prevention of a GC content bias in SAGE libraries. *Nucleic Acids Res.*, **29**, E60.

Munroe,S.H. and Lazar,M.A. (1991) Inhibition of c-erbA mRNA splicing by a naturally occurring antisense RNA. *J. Biol. Chem.*, **266**, 22083–22086.

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.

Shendure,J. and Church,G.M. (2002) Computational discovery of sense-antisense transcription in the human and mouse genomes. *Genome Biol.*, **3**, RESEARCH0044.

Simons,R.W. (1988) Naturally occurring antisense RNA control—a brief review. *Gene*, **72**, 35–44.

Sleutels,F., Zwart,R. and Barlow,D.P. (2002) The non-coding Air RNA is required for silencing autosomal imprinted genes. *Nature*, **415**, 810–813.

Storz,G. (2002) An expanding universe of noncoding RNAs. *Science*, **296**, 1260–1263.

Sun,M., Zhou,G., Lee,S., Chen,J., Shi,R.Z. and Wang,S.M. (2004) SAGE is far more sensitive than EST for detecting low-abundance transcripts. *BMC Genomics*, **5**, 1.

Tufarelli,C., Stanley,J.A., Garrick,D., Sharpe,J.A., Ayyub,H., Wood,W.G. and Higgs,D.R. (2003) Transcription of antisense RNA leading to gene silencing and methylation as a novel cause of human genetic disease. *Nat. Genet.*, **34**, 157–165.

Vanhee-Brossollet,C. and Vaquero,C. (1998) Do natural antisense transcripts make sense in eukaryotes? *Gene*, **211**, 1–9.

Velculescu,V.E., Zhang,L., Vogelstein,B. and Kinzler,K.W. (1995) Serial analysis of gene expression. *Science*, **270**, 484–487.

Wahl,M.B., Heinzmann,U. and Imai,K. (2005) LongSAGE analysis significantly improves genome annotation: identifications of novel genes and alternative transcripts in the mouse. *Bioinformatics*, **21**, 1393–1400.

Wahl,M., Shukunami,C., Heinzmann,U., Hamajima,K., Hiraki,Y. and Imai,K. (2004) Transcriptome analysis of early chondrogenesis in ATDC5 cells induced by bone morphogenetic protein 4. *Genomics*, **83**, 45–58.

Yamada,K., Lim,J., Dale,J.M., Chen,H., Shinn,P., Palm,C.J., Southwick,A.M., Wu,H.C., Kim,C., Nguyen,M. *et al*. (2003) Empirical analysis of transcriptional activity in the Arabidopsis genome. *Science*, **302**, 842–846.