FEBS Letters 580 (2006) 6721–6729

Pan-genome isolation of low abundance transcripts using SAGE tag

Yeong Cheol Kim[a,1], Yong-Chul Jung[a,1], Zhenyu Xuan[b,1], Hui Dong[a], Michael Q. Zhang[b], San Ming Wang[a,c,*]

[a] *Center for Functional Genomics, Division of Medical Genetics, Department of Medicine, ENH Research Institute, Northwestern University, Evanston, IL 60201, United States*
[b] *Cold Spring Harbor Laboratory, New York, NY 11724, United States*
[c] *Robert H. Lurie Comprehensive Cancer Center, Northwestern University, Chicago, IL 60611, United States*

Received 1 September 2006; revised 31 October 2006; accepted 3 November 2006

Available online 14 November 2006

Edited by Gianni Cesareni

Abstract **The SAGE (serial analysis of gene expression) method is sensitive at detecting the lower abundance transcripts. More than a third of human SAGE tags identified are novel representing the low abundance unknown transcripts. Using the GLGI method (generation of longer 3′ EST from SAGE tag for gene identification), we converted 1009 low-copy, human X chromosome-specific SAGE tags into 10210 3′ ESTs. We identified 3418 unique 3′ ESTs, 46% of which are novel and originated from the lower abundance transcripts. However, nearly all 3′ ESTs were mapped to various regions across the genome but not X chromosome. Detailed analysis indicates that those 3′ ESTs were isolated by SAGE tag mis-priming to the non-parent transcripts. Replacing SAGE tags with non-transcribed genomic DNA tags resulted in poor amplification, indicating that the sequence similarity between different transcripts contributed to the amplification. Our study shows the prevalence of novel low abundance transcripts that can be isolated efficiently through SAGE tags mis-priming.**

Keywords: Transcript; Low abundance; SAGE tag; 3′ EST

1. Introduction

Transcripts are the functional carriers of genes. Transcript isolation is essential for gene identification and for functional study of genes. The abundance of different transcripts can vary over a million-fold [1–4]. The higher abundance transcripts tend to be from a limited number of genes with housekeeping functions, while the lower abundance transcripts tend to be from most of the genes with specialized functions. Isolation of full sets of transcripts expressed from a given genome, regardless of abundance, is an ultimate goal in transcriptome studies.

Transcript isolation has been highly successful during the last decade. For example, the large-scale human EST collection has isolated over 7 million ESTs from the human genome [5–7]. (http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html). However, recent evidence shows that novel, low-abundance transcripts are widely present in yeast, fly, mouse, rat, *Arabidopsis*, rice, and human [8–17], indicating that the transcriptome is far more complex than thought [18], and transcript identification is far from complete, even in these extensively characterized model genomes.

The approaches used for large-scale transcript identification include the EST approach that detects transcripts with several hundred bases [5–7] and the SAGE approach that detects transcripts with 10–20 bases [19,20]. Over 7 million copies of human ESTs and 20 million copies of human SAGE tags have been isolated (http://www.ncbi.nlm.nih.gov/projects/SAGE/). While novel transcript identification by the EST approach has decreased dramatically [14], the SAGE approach continues to detect more low abundant transcripts due to its high sensitivity [21]. However, SAGE provides only short sequence information for the detected transcripts. Identification of their original transcripts with longer sequence information will be essential for annotation and functional studies. In this study, we used the GLGI method to convert over a thousand human X chromosome-specific SAGE tags into the 3′ ESTs with the primary aim of identifying the novel transcripts from X chromosome. While nearly half of the isolated 3′ ESTs are novel, most of the 3′ ESTs represent the transcripts from non-X chromosomes. Further analysis reveals that those 3′ ESTs were isolated by SAGE tags through mis-priming mechanism. Here we report the details of the study.

2. Materials and methods

2.1. SAGE data analysis

Experimental 10-base SAGE tags and 17-base long SAGE tags were downloaded from NCBI GEO (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc = gpl4, http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc = gpl1485). The reference 10-base and 17-base long SAGE tags were downloaded from "SAGEmap-full" of the SAGEmap database (ftp://ftp.ncbi.nlm.nih.gov/pub/sage/map/Hs/). Human genome sequences (NCBI 34) were used for extracting 10-base and 17-base genomic tags at the CATG sites. For sense strand, tags were extracted immediately after each CATG site; for anti-sense strand, tags were extracted before each CATG site with reverse/complementary sequences. Human ESTs were downloaded from NCBI dbEST (ftp://ftp.ncbi.nih.gov/repository/dbEST/) and 10-base and 17-base tags were extracted after the last CATG sites in the 3′ ESTs. Different computational programs written in Perl were used for the comparison between the experimental SAGE tags, reference SAGE tags, and the genomic tags.

*Corresponding author. Fax: +1 224 364 5003.
E-mail address: swang1@northwestern.edu (S.M. Wang).

[1]These authors contributed equally to this study.

0014-5793/$32.00
doi:10.1016/j.febslet.2006.11.013

2.2. Selection of X chromosome-specific novel SAGE tags

The 315 human SAGE libraries containing 18966751 SAGE tags were downloaded (http://www.ncbi.nlm.nih.gov/geo/, December 9, 2005). The SAGE tags were matched to the human SAGEmap reference database (SAGEmap full, release 172, October 25, 2005) to identify the non-matched SAGE tags as novel SAGE tag candidates. A chromosome-based human genomic SAGE tag reference database was constructed by extracting 10-bases and 17-bases after all CATG and before CATG (reverse complementary) using the human genome sequences (NCBI 34). All novel SAGE tags were mapped to the genomic tag reference database to identify those specifically mapped to the human X chromosome.

2.3. Conversion of SAGE tags into 3′ ESTs

Each X chromosome-specific novel SAGE tag was used to design a primer with sequences CAGGGACATGxxxxxxxxxx, where CAGGGA is used to increase the length of the primer, CATG in the SAGE tag is the *Nla*III restriction site used for releasing SAGE tags from cDNA templates, and xxxxxxxxxx is the 10-bp SAGE tag sequence. RNA samples from brain and placenta tissues (Stratagene, La Jolla, CA) were treated by using DNase I digestion to eliminate potential genomic DNA contamination, and the purity of digested RNA samples was tested by PCR amplification of the beta-actin gene (GenBank ID NM_001101) using sense primer GGACTTCGAGCAAGAGATGG and antisense primer AGCACTGTGTTGGCGTACAG that span the intron 4 of the beta-actin gene. The amplified genomic DNA will be 329 bp, and the amplified mRNA will be 234 bp.

GLGI reactions were performed to convert SAGE tags into 3′ cDNAs following the procedures described previously [22,23]. Briefly, mRNAs were purified from total RNA samples using oligo $(dT)_{25}$ magnetic beads (Dynal, Brown Deer, WI). Double strand poly dA/dT (-) cDNAs were synthesized using M-MLV Reverse Transcriptase (Invitrogen, Carlsbad, CA) and the 5′ biotinylated, 3′ anchored oligo (dT) primers (5′ biotin-ATCTAGAGCGGCCGC-T16-A/G/CA/CG/CC), and digested by *Nla*III. The 3′ cDNAs after the last CATG were then isolated using streptavidin magnetic beads (Dynal). GLGI reactions were performed in 96-well plates including Hotstart Taq polymerase (Denville, Metuchen, NJ), each sense primer, universal antisense primer ACTATCTAGAGCGGCCGCTT and 3′ cDNAs. The amplified products of each reaction were cloned into the pGEM-T vector (Promega, Madison, WI), transformed into *E. coli* TOP10 (Invitrogen) and plated in a single well of the 48-well Qtrays (Genetix, Hampshire, UK). Twelve clones from each transformation were selected for sequencing collection. Plasmids were purified by using the Montage Plasmid Miniprep96 Kit (Millipore, Billerica, MA). DNA sequencing reactions were performed using the Big-Dye Terminator v3.1 Cycle Sequencing Kit (ABI, Foster City, CA), and sequences were collected in an ABI 3730XL DNA sequencer using Phred20 as the cutoff. Only the sequences containing the SAGE tag-based sense primer in the 5′ end and the polyA tail at the 3′ end were considered as the qualified 3′ ESTs. The same sequences originated from the same sense primer were combined as a unique sequence. Poly A signals were identified in each sequence by searching the poly A signal sequences AATAAA, ATTAAA AATTAA, AATAAC, AATAAT, AATACA, ACTAAA, AGTAAA, CATAAA, GATAAA, and TATAAA upstream 100 bps from the 3′ end of each sequence [24].

2.4. Confirmation of 3′ ESTs by using RT-PCR

Sense primer and antisense primer were designed based on each selected 3′ EST sequence to generate the amplicons between 100 and 300 bases. One hundred nanograms of brain or placenta total RNA were used as the templates for cDNA synthesis. For antisense confirmation, cDNA was synthesized by using each antisense primer, and PCR was followed by adding each sense primer. Conditions for PCR amplification were 35 cycles of 94 °C for 30 s, 60 °C for 30 s (50 °C was set for DR978181), and 72 °C for 1 min, and then extended at 72 °C for 7 min. PCR products were checked on 1% agarose gels.

2.5. Estimation of the abundance of the identified transcripts by using real-time PCR

The same sequences and primers used for RT-PCR confirmation were used for real-time PCR. The reactions were performed following the manufacturer's protocol (Stratagene, La Jolla, CA). Briefly, first-strand cDNA was synthesized by using oligo dT_{12-18} primer and 1 ug of DNase I treated total RNA in a total volume of 20 μl. Two microliters of each primer set and 2 μl of the synthesized cDNA were added to FullVelocity SYBR® Green QPCR master mix (Stratagene) in a total volume of 25 μl. The mixtures were placed in an Mx3000P instrument (Stratagene) and the PCR program was run at 1 cycle at 95 °C for 5 min, 45 cycles at 95 °C for 30 s, 60 °C for 60 s and 72 °C for 60 s. Beta-actin transcripts were used as a control representing the high abundance transcripts. The SAGE tag copies of beta-actin in brain and placenta tissues were identified from brain SAGE library GSM763 and placenta SAGE library GSM 14750 (http://www.ncbi.nlm.nih.gov/projects/geo/).

2.6. Control experiments using genomic tag primers and random primers

Genomic DNA sequences of 14-bps were extracted from the non-repetitive genomic regions in the human X chromosome (NCBI 34) that are free from known genes, mRNAs, and ESTs. The 14-bp sequences were filtered through the SAGE tag databases (GPL4 and SAGEmap 187) to exclude any tag that match existing SAGE tags. Those without matches were used as the sense primer. Six bases (CAGGGA), as used in the SAGE tag-based primers, were added to the 5′ end of each primer to increase the length to 20 bp. The same universal primer, placenta and brain 3′ cDNAs used in the GLGI were used for the reaction.

2.7. Match 3′-ESTs to the known human transcripts

Each 3′-EST was searched in known human mRNAs in RefSeq database (ftp://ftp.ncbi.nih.gov/refseq/H_sapiens/mRNA_Prot/) and EST database (ftp://ftp.ncbi.nih.gov/repository/dbEST/). If the 3′-EST could align with any mRNA or EST with over 95% identity with 90% coverage, the 3′-EST was annotated as "known". Otherwise, it was defined as "novel". Each 3′ EST was also matched to the trEST and trGEN databases that contains the predicted human transcript contigs based on UniGene and the EMBL databases [25,26]. *E* value $1.0e^{-15}$ was used as the cutoff to divide the matched ones from the unmatched ones [27]. When we compared 3′-EST with trEST and trGEN, we separated 3′-ESTs into five classes: (a) known transcript, which the 3′-end of 3′-EST matches less than 10 bp upstream of 3′-end of transcript in database, or the 3′-end of transcript in database matches less than 10 bp upstream of 3′-end of 3′-EST; (b) 3′-end shortened transcript, which the 3′-end of 3′-EST matches more than 10 bp upstream of 3′-end of transcript; (c) 3′-end extended transcript, which the 3′-end of transcript in the database matches more than 10 bp upstream of 3′-end of 3′-EST; (d) other alternative transcript, which are those 3′-ESTs that match transcript in the database with e-value lower than $1.0e^{-15}$ but do not belong to any of the above classes; (e) unmatched 3′-ESTs."

2.8. Map 3′-ESTs to the human genome

Each 3′-EST was mapped to the human genome by using BLAT with the default parameter settings (NCBI 34). The 3′ ESTs that could not be mapped by BLAT were mapped by BLASTN with the e-value cutoff as $1.0e^{-1}$. For fine genomic mapping, we extracted the mapped genomic DNA sequence, and ran the global alignment program ALIGN (from FASTA package) to align the 3′-EST with the corresponding genomic DNA sequences to calculate the global sequence identity. A mapping was determined if more than 90% of a 3′-EST mapped to the genome, and the identity between genomic DNA and 3′-EST was at least 90%. In case of multiple mapped regions, only those with a BLAT or BLAST score no more than 0.5% lower than the score of the best-mapped region were kept, which is the cutoff value used in the human genome browser (http://genome.ucsc.edu/). The 3′-ESTs with more than five mapped loci in the genome were excluded for further analysis. The immediate downstream region of the mapped genomic region was checked to identify potential genomic poly-A sequences. A 3′-EST was marked as "internal poly A priming" if more than eight continuous As were located within the 20 bp downstream flanking region [28]. Using the annotated RefSeqs and mRNAs in the mapped regions, we classified each mapped 3′ EST into the "Intergenic" or "Intragenic" group. The "Intragenic" group was further classified into "sense" and "antisense" subsets. The "sense" includes "known", "intronic" (completely mapped in a single intron), "extended 5′ end" (extended more than 10 bp

upstream of the annotated 5′-end), "extended 3′ end" (extended more than 10 bp downstream of the annotated 3′ end), "shortened 3′ end" (the 3′-end of a 3′-EST maps more than 10 bp upstream of the annotated 3′ end), "cross-conjunction" (the 3′-EST overlaps the exon–intron conjunction), and "antisense" (the 3′-EST maps to the reversed orientation of annotated gene). Genes mapped to the "intergenic" region were classified into the "known" subset (with EST information) and "novel" subset (with no known transcript information).

3. Results and discussion

3.1. SAGE data analysis

Three sets of data were used for the analysis, including (1) the experimental human SAGE tags that represent the human transcripts *detected* by SAGE; (2) the reference human SAGE tags extracted from well-annotated mRNAs and ESTs that represent the known human transcripts *detectable* by SAGE [29]; and (3) the reference human genomic tags extracted from the human genome sequences that represent the *possibly transcribed loci and detectable* by SAGE in the human genome. The experimental human SAGE data include the standard SAGE tags of 10-bases (664 615 unique tags identified from 16 350 661 tag copies originated from 307 human SAGE libraries), and long SAGE tags of 17-bases (630 837 unique tags identified from 3 616 090 copies originated from 29 human long SAGE libraries); the reference human SAGE tags include reference 10-base standard SAGE tags and reference 17-base long SAGE tags (577 224 10-base tags and 1 291 619 17-base tags, respectively); the reference human genomic tags include standard 10-base and 17-base long SAGE genomic tags extracted from the human genome sequences adjunct to CATG sites in sense and anti-sense orientations (957 056 standard and 19 618 123 long genomic tags from 26 201 271 genomic locations). The genomic tags provide evidence for the genomic origin of SAGE tags and serve as a discriminator to eliminate uncertain SAGE tags. An experimental SAGE tag that matches a reference SAGE tag and a reference genomic tag implies that this SAGE tag is originated from a known transcript expressed from the genome; an experimental SAGE tag that has no match to reference SAGE tag but maps to genomic tag implies that this SAGE tag is likely originated from a novel transcript expressed from the genome; an experimental SAGE tag that has no match to reference SAGE tag nor maps to genomic tag implies that the origin of SAGE tag is uncertain.

Fig. 1 shows the results of the comparison. For the 664 615 experimental standard 10-base SAGE tags, 645 061 map to the reference genomic tags, 447 618 (69.4%) of which match to both the reference SAGE tags of known transcripts and reference genomic tags, and 197 443 (30.6%) have no match to the reference SAGE tags but map to the reference genomic tags; for the 630 837 long SAGE tags, 217 640 map to the reference genomic tags, 109 548 (50.3%) of which match both the reference SAGE tags of known transcripts and genomic tags, and 108 092 (49.7%) have no match to the reference SAGE tags but map to the reference genomic tags. These novel SAGE tags have low copy numbers in the SAGE libraries, implying that most of the transcripts detected by novel SAGE tags are the low-abundance transcripts.

The actual number of novel transcripts should be higher than that of the novel SAGE tags, based upon the following considerations: Many low abundance transcripts are below the threshold of SAGE detection; a third of the standard 10-base SAGE tags are shared by different transcripts [30]; transcripts lacking the *Nla*III site for tag releasing are not detected by SAGE; and certain SAGE tags excluded from this analysis might be from true novel transcripts.

3.2. Conversion of human X chromosome-specific SAGE tags into 3′ ESTs

By searching human SAGE data, we identified 1009 novel SAGE tags that do not match known human transcripts but map solely to the genomic sequences of the human X chromosome. All these novel SAGE tags have low-copies in their original SAGE libraries. Using the GLGI method, we converted these SAGE tags into 3′ ESTs using RNA samples from brain and placenta. By sequencing 12 clones per GLGI reaction, we obtained 13 824 raw sequences. After excluding unqualified sequences and combining redundant sequences, we identified 3418 unique 3′ ESTs, each of which contains the original SAGE tag at its 5′ end and polyA tail at its 3′ end (Table 1A). A total of 945 of the 1009 SAGE tags (94%) contributed these final sequences (Table 1B). The 3418 3′ ESTs have been deposited in GenBank with Accession Numbers from DR977574 to DR980991.

3.3. Confirmation of the isolated transcripts

To verify that the isolated sequences were not from contaminated genomic DNA but rather from transcripts, we performed PCR to test the RNA samples by using the sense and anti-sense primers that span an intron of the beta-actin gene. No genomic DNA signal was detected in the RNA samples treated after RNase A. Therefore, the isolated sequences likely originated from RNA rather than genomic DNA contamination (Fig. 2A). We then used RT-PCR to verify the isolated transcripts. Sense and antisense primers were designed based on each selected 3′ EST (Supplementary Table 1), and RNA samples from brain and placenta were used as the template. To confirm the antisense sequences, cDNA was synthesized by using each antisense primer for PCR amplification. Of the 30 selected 3′ ESTs, 28 were detected, including all five antisense sequences detected in both RNA samples, 18 sense sequences detected in both RNA samples and five sense sequences detected in either RNA sample (Fig. 2B). The positive detection reveals that most of the detected transcripts were indeed present in the RNA sample. Those two negatively detected 3′ ESTs might be related with the poor amplification efficiency of the primers.

3.4. Estimation of the abundance of isolated transcripts

We used real-time RT-PCR to estimate the abundance of the same 30 isolated transcripts used for RT-PCR confirmation (Supplementary Table 1). The beta-actin transcript was used as the reference that was expressed at high abundance levels in both brain and placenta tissues (721 out of the total 63 208 SAGE tag copies in the brain SAGE library and 1141 out of the total 118 083 SAGE tag copies in the placenta SAGE library). Of the 30 selected sequences, 29 (except DR978181) were detected in both placenta and brain RNA samples. All 29 in the placenta and 28 in the brain (except DR979502) were detected after the beta-actin signal (Fig. 3A). The abundance of these transcripts covered several orders of magnitude and most were lower than that of beta-actin; DR979332 had the

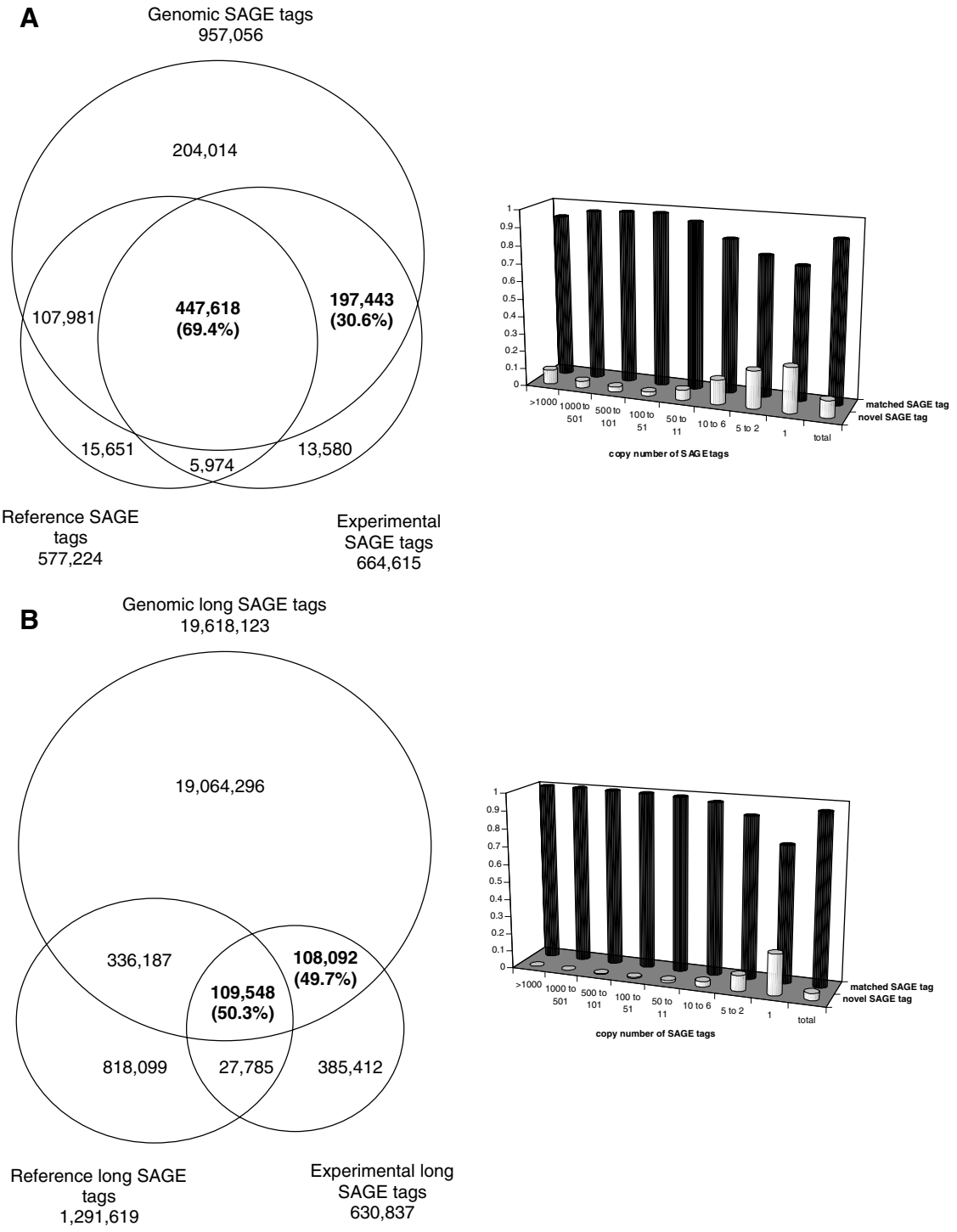


Fig. 1. The comparison between SAGE tags, reference SAGE tags of known human transcripts and genomic tags of the human genome sequences. A shows the analysis of the standard 10-bp tags. B shows the analysis of the 17-bp long tags. Of the experimentally collected SAGE tags that mapped to the genome, 30.6% of the 10-base tags and 49.7% of the 17-base tags have no match to the corresponding reference SAGE tags. The histograms show the copy number distribution of the matched SAGE tags and the novel SAGE tags in their original SAGE libraries. In both types of SAGE tags, the novel SAGE tags have lower copies than the matched SAGE tags.

lowest abundance (Fig. 3B). Transcripts not visible in gel by regular RT-PCR at 35 PCR cycles (Fig. 2B) were detected by real-time PCR at 45 cycles, suggesting very low abundance levels for no. 6, 16, 17, 22, 23 in both placenta and brain, 28 in placenta, 19, 24 and 30 in brain.

3.5. Annotation of the isolated transcripts

We compared the 3′ ESTs to the known human transcripts. The results show that 46% of the 3418 unique 3′ ESTs represent novel transcripts not identified so far (Table 2). This rate is significantly higher than the less than 5% novelty in large-

Table 1
Summary of the resulting 3′ ESTs converted from SAGE tags

A. 3′ ESTs isolated from X chromosome-specific SAGE tags

Items	Numbers	
X chromosome-specific SAGE tags	1009	
Raw sequences generated from SAGE tags	13824	
Sequences not qualified	3614	
Sequences qualified	10210	
Placenta		6561
Brain		3649
Final set of unique 3′ ESTs	3418	
Placenta		2328
Brain		1090
Containing poly A signal	1374	
Length distribution (bp)	30–688	

B. Number of 3′ ESTs contributed by SAGE tags

Number of SAGE Tags (%)	Contributed 3′ ESTs	Number of 3′ ESTs (%)
134 (13)	1	134 (4)
180 (18)	2	360 (11)
185 (18)	3	561 (16)
176 (17)	4	692 (20)
123 (12)	5	645 (19)
77 (8)	6	474 (14)
31 (3)	7	196 (6)
27 (3)	8	248 (7)
12 (1)	9	108 (3)
Total 945 (100)		Total 3418 (100)

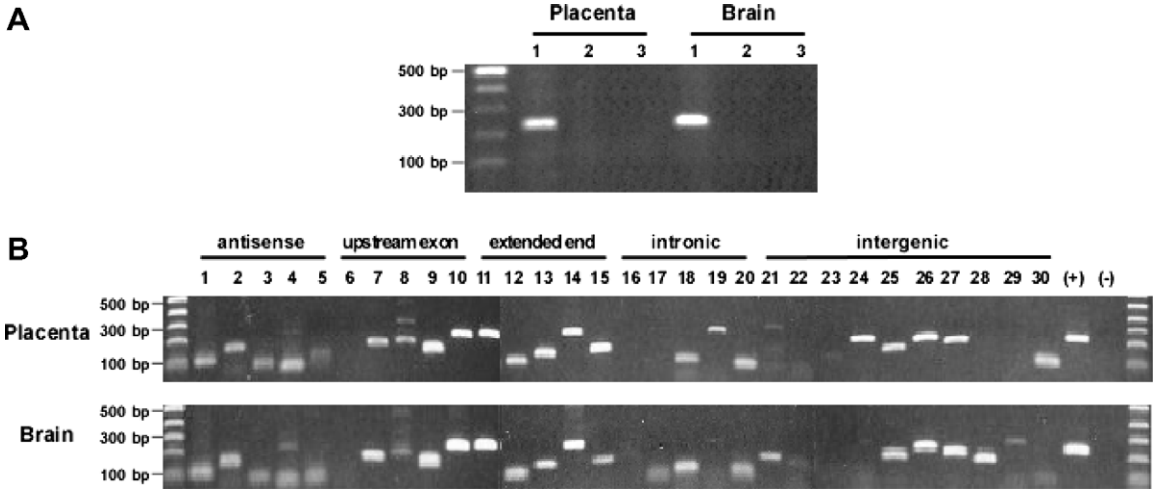
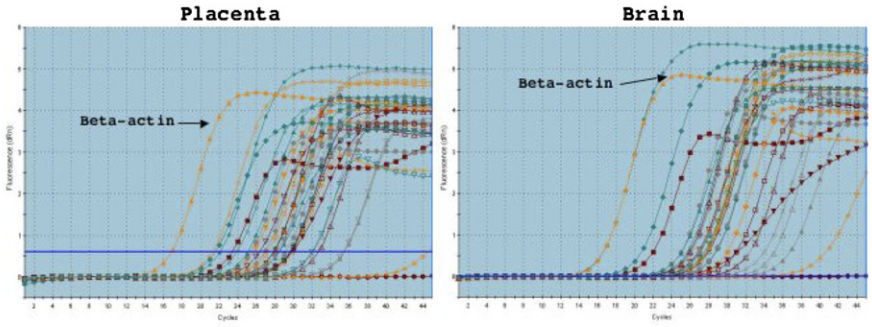


Fig. 2. Confirmation of the origin of the isolated sequences. (A) Determination of genomic DNA contamination. DNase I digested RNA samples were tested by direct PCR amplification of beta-actin genomic DNA with primers spanning an intron. The amplicons from genomic DNA will be 329 bps, whereas that from mRNA will be 234 bp. (1) RT-PCR with RNA samples. (2) PCR with RNase A digested RNA samples. (3) Negative control without RNA. (B) RT-PCR confirmation of detected sequences. A total of 30 isolated sequences were selected for the confirmation. Sense primer and antisense primer were designed based on each sequence. For antisense confirmation, antisense primer was used for cDNA synthesis; for other types of confirmation, oligo dT was used for cDNA synthesis. The order of the tested sequences is the same as listed in the Supplementary Table 1. (+), positive control with beta-actin transcripts; (−), negative control with RNase A digested RNA samples. Most of the sequences, except for a few cases, were detected in both RNA samples with similar size distribution.

scale EST collection [14]. We mapped the 3′ ESTs to the human genome to determine their genomic origins (Table 3). Of the 3418 3′ ESTs, 2503 are directly mapped to 2630 loci in the genome, including 2408 mapped to non-X chromosomes, 89 mapped to X-chromosome and 6 mapped to both X and non-X chromosomes. Of the mapped loci, 28% are at the intergenic regions without annotated genes. Although 72% map to the intragenic regions, 24% are in the antisense strand, 7% of which are novel antisense transcripts. For the 47% mapped to the sense strand of the intragenic regions, 21% map precisely to the annotated 3′ end of the known genes, while the rest map as "intronic," and different spliced variations. In total, 60% of the 3′ ESTs provide various degrees of novel transcriptional information for these mapped loci.

We also compared 915 3′ ESTs which cannot be matched to genome with the predicted transcript contigs in the trEST database and whole genome transcript prediction in trGEN

A. Real-Time PCR quantitation of the detected transcripts



B. Abandance of the detected transcripts related to beta-actin transcripts*

3' EST	Abundance relative to beta-actin	
	Placenta	Brain
Beta-actin	1.00000000	1.0000000
DR978919	0.04976404	0.0006009
DR978736	0.04859671	0.0833557
DR979502	0.02673751	1.1585132
DR978861	0.01136353	0.0526823
DR980876	0.00682757	0.0030340
DR977917	0.00360148	0.0048998
DR978587	0.00286511	0.0050359
DR978950	0.00242350	0.0006717
DR980720	0.00147706	0.0014133
DR980985	0.00099498	0.0015236
DR980934	0.00075838	0.0011697
DY633389	0.00050136	0.0022374
DR980429	0.00048372	0.0008532
DR980291	0.00047004	0.0001490
DR980629	0.00042093	0.0009625
DR980060	0.00035848	0.0000116
DR977597	0.00033308	0.0003278
DR980161	0.00022253	0.0009948
DR978161	0.00019625	0.0011532
DR979161	0.00015206	0.0006955
DR978688	0.00014486	0.0000953
DR977977	0.00013204	0.0002942
DR979782	0.00011486	0.0003199
DR977682	0.00002555	0.0011480
DR980739	0.00002267	0.0000035
DR977754	0.00002019	0.0000351
DR979141	0.00000290	0.0000336
DR979498	0.00000214	0.0000021
DR979332**	0.00000001	0.0000002
DR978181	negative	negative

*The numbers were sorted in descenting order based on those in placenta RNA.
**This is the transcripts detected after 45 cycles.

Fig. 3. Quantitative measurement of the abundance of the detected sequences by real time PCR. The same sequences, primers and RNA samples used in Fig. 2 were used for this analysis. See Supplementary Table 1 for detailed sequence information. (No. 2 sequence DR978181 was not used for this analysis, as its melting temperature is only 53 °C, far lower than the 60 °C recommended for real-time PCR). The beta-actin transcripts were used as positive control. (A) Histogram of real-time PCR showing the amplification dynamics for each detected sequence. (B) Relative abundance of each sequence normalized to beta-actin. Note that the weakly amplified templates by regular PCR in this figure were well reflected by their lower abundance detected by real time PCR.

database [25,26]. The results show that 449 3′-ESTs have matches in trEST or trGEN under *E*-value cutoff of $1.0e^{-15}$, and 466 3′-ESTs still cannot find significant matches in these databases. Based on the definition described in method (Section 2.7), we found 173 out of those 449 matched 3′-ESTs belong to known transcripts, while 224, 21 and 81 3′-ESTs belong to "3′-end shortened transcript", "3′-end extended transcript", and "other alternative transcript", respectively. Of those 915 genome-unmapped 3′-ESTs, only 173 were related to the predicted known transcripts while 742 (742/915 = 81%) remain as "novel".

We also analyzed the sequences filtered due to their multiple mappings. We set a higher than 95% of the maximum score as a tolerable cutoff for the analysis. There are a total of 1455 loci mapped by 157 3′-ESTs with the scores between 95% and 99.5%. Using 99.5% of maximum match score as cutoff, we

Table 2
Comparison of 3′ ESTs with known transcripts

Class	Number	
Match to known transcripts	1857 (54)	
Known mRNA[a]		1175
EST[b]		1857
No match	1561 (46)	
Total	3418 (100)	

[a]RefSeq (ftp://ftp.ncbi.nih.gov/refseq/H_sapiens/mRNA_Prot/).
[b]dbEST (ftp://ftp.ncbi.nih.gov/repository/dbEST/).

Table 3
Mapping 3′ ESTs to the human genome (NCBI 34)

Classification	No. of mapped 3′ ESTs (%)			No. of mapped loci[a] (%)		
Intergenic region	691 (27)			743 (28)		
Intragenic region	1855 (73)			1887 (71)		
Sense		1234 (49)			1248 (47)	
Known			555 (22)			565 (21)
Intronic			382 (15)			383 (14)
Shortened 3′-end			250 (10)			253 (9)
Cross conjunction			40 (2)			40 (1)
Extended 5′-end			4 (0)			4 (0)
Extended 3′-end			3 (0)			3 (0)
Antisense		621 (25)			639 (24)	
Known			390 (16)			404 (15)
Novel			231 (9)			235 (8)
Total	2546 (100)			2630 (100)		

[a]The number of subset may not equal to the sum, due to multiple mapped loci for some 3′-EST.

found only 231 loci in the genome for those 157 3′-ESTs. The classification for those mapped loci shows that 933 are the intergenic region, and 522 are the intragenic region of which are the antisense orientation. For the 203 with sense orientation, only 25 are known exon, and the rest are intronic and variations. Those results show that there are more novel transcribed loci detected by those 157 sequences although their precise loci cannot be assigned precisely due to the multiple mapping.

3.6. Identification of the mismatched bases between the SAGE tag part of 3′ ESTs and the mapped genomic sequences

All SAGE tags used for the experiment map only to the X chromosome. However, the majority of isolated 3′ ESTs are the transcripts originated from non-X chromosomes. Using the 2417 3′ ESTs that map to single loci in the genome, we compared the 14-bp SAGE tag sequences of these 3′ ESTs and their mapped genomic sequences. Interestingly, we observed widely spread mismatches/gaps between these 14-bp

Table 4
Mis-matched bases between the SAGE tag part of 3′ ESTs and their mapped genomic regions[a]

No. mismatched bases	Location of mismatches/gaps in SAGE tag part of 3′ ESTs														Total No. 3′ ESTs (%)
	C	A	T	G	x	x	x	x	x	x	x	x	x	x	
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	139 (6)
1	8	2	6	1	156	229	137	114	75	87	49	39	17	3	923 (38)
2	74	33	27	6	238	236	180	160	95	72	31	25	12	5	597 (25)
3	135	67	44	24	92	100	79	54	40	26	47	14	3	4	243 (10)
4	149	134	118	80	61	46	39	32	24	25	23	15	10	4	190 (8)
5	125	138	140	119	85	69	28	30	25	18	14	6	7	6	162 (7)
6	77	82	75	68	63	59	31	17	16	18	6	5	8	3	88 (4)
7	19	20	22	19	20	17	19	11	7	4	5	5	4	3	25 (1)
8	9	9	6	8	7	9	8	4	7	9	4	6	5	5	12 (0)
9	19	17	11	16	18	13	17	12	10	11	13	14	10	8	21 (1)
10	8	8	8	8	7	10	7	8	4	6	6	4	8	8	10 (0)
11	5	3	3	4	5	4	5	4	4	5	5	3	3	2	5 (0)
12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0 (0)
13	2	2	1	2	1	2	2	2	2	2	2	2	2	2	2 (0)
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0 (0)
Total	630	515	461	355	753	794	552	448	309	283	205	138	89	53	2417 (100)

[a]Those include 92 3′ ESTs mapped to X chromosome of which 23 mapped to the locations expected by the original SAGE tags.

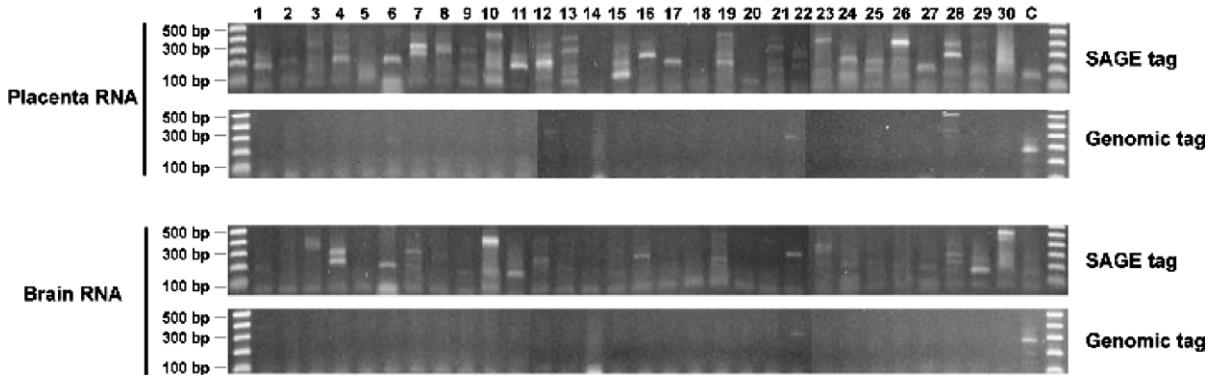


Fig. 4. Comparison between SAGE tag-based primer, genomic sequence-based primer and random primer for transcript detection. Reactions were performed by using 30 SAGE tag-based primers (Supplementary Table 1), 30 genomic sequence-based primers (Supplementary Table 2), and random primers. Placenta and brain samples were used as the templates. The last lanes of the two gels marked by "SAGE tag" were the amplicons from random primers, the last lanes of the two gels marked by "Genomic tag" were the positive control using the SAGE tag primer (No. 16 in Supplementary Table 1). Note that in the genomic sequence-based reaction, only Nos. 22 and 28 in placenta RNA and 22 in brain RNA show positive amplification.

counterpart sequences (Table 4). Overall, 98% of the mappings contain mismatches/gaps, of which mismatches account for 76% and gaps account for 24%. The numbers of mismatched/gapped bases are predominated by the 1- and 2-bases that are located in the middle of the SAGE tag sequences. Because the 14-bp SAGE tag sequences are from the synthesized sense primer, the mismatches/gaps between the 14-bp of the isolated 3′ ESTs and the genomic sequences cannot be related to the issue of fidelity for the PCR amplification and the pattern of mismatches/gap distribution doesn't correlate with possible sequencing errors. Therefore, these transcripts must have been isolated by the SAGE tags through mis-priming.

3.7. Tags from non-transcribed genomic sequences provide poor amplification

To investigate if genomic sequence-based primers could also result in the amplification as seen in the SAGE tag-based primers, we used the non-transcribed genomic DNA sequences in the human X chromosome as the primers for the reaction (Supplementary Table 2). The results show that, of the 30 reactions tested, 28 reactions in placenta and 29 in brain generated negative results (Fig. 4). Those results show that genomic DNA sequences do not provide amplification as efficiently as SAGE tag sequences. The results also imply that the sequence similarity represented by the SAGE tag sequences between different transcripts might account for their highly efficient amplification.

Our study shows that many identified transcripts are at low abundance levels from either the unannotated regions in the genome, or from the annotated genes but with complex sequence variations or from antisense of the annotated genes. The current definition of "higher" or "lower" abundance of transcripts may largely reflect our ability for transcript detection rather than biological significance of the detected genes. The quantitative range covering over six orders of magnitudes between different transcripts determines that the depth of transcript isolation depends largely upon the sensitivity of techniques used. The conventional EST approach has limited power to isolate low abundance transcripts due largely to the issue of cost-efficiency. The PCR approach is very sensitive at detecting the low abundance transcripts, but the required sequence information for two-primer design restricts its use in detecting only the known transcripts. Random primer has been applied successfully for large-scale EST isolation, but it mainly detects the middle region of the targeted transcript population without the 5′ or 3′ end sequence information [31]. The PCR-based GLGI technique requires only one primer (a SAGE tag) to generate 3′ ESTs. The observed mis-priming between a SAGE tag and multiple transcripts makes it possible to use SAGE tags for pan-genome detection of low abundance transcripts.

Acknowledgements: The study was supported by National Institutes of Health (HG002600), the Daniel F. and Ada L. Rice Foundation, and Mazza Foundation.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.febslet.2006.11.013.

References

[1] Bishop, J.O., Morton, J.G., Rosbach, M. and Richardson, M. (1974) Three abundance classes in HeLa cell messenger RNA. Nature 250, 199–204.
[2] Holland, M.J. (2002) Transcript abundance in yeast varies over six orders of magnitude. J. Biol. Chem. 277, 14363–14366.
[3] Czechowski, T., Bari, R.P., Stitt, M., Scheible, W.R. and Udvardi, M.K. (2004) Real-time RT-PCR profiling of over 1400 *Arabidopsis* transcription factors: unprecedented sensitivity reveals novel root- and shoot-specific genes. Plant J. 38, 366–379.
[4] Carter, M.G., Sharov, A.A., VanBuren, V., Dudekula, D.B., Carmack, C.E., Nelson, C. and Ko, M.S. (2005) Transcript copy number estimation using a mouse whole-genome oligonucleotide microarray. Genome Biol. 6, R61.
[5] Adams, M.D., Dubnick, M., Kerlavage, A.R., Moreno, R., Kelley, J.M., Utterback, T.R., Nagle, J.W., Fields, C. and Venter, J.C. (1992) Sequence identification of 2,375 human brain genes. Nature 355, 632–634.
[6] Bonaldo, M.F., Lennon, G. and Soares, M.B. (1996) Normalization and subtraction: two approaches to facilitate gene discovery. Genome Res. 6, 791–806.
[7] Strausberg, R.L., Dahl, C.A. and Klausner, R.D. (1997) New opportunities for uncovering the molecular basis of cancer. Nat. Genet., Spec. No., 415–416.

[8] Wyers, F., Rougemaille, M., Badis, G., Rousselle, J.C., Dufour, M.E., Boulay, J., Regnault, B., Devaux, F., Namane, A. and Seraphin, B., et al. (2005) Cryptic pol II transcripts are degraded by a nuclear quality control pathway involving a new poly (A) polymerase. Cell 121, 725–737.
[9] Lee, S., Bao, J., Zhou, G., Shapiro, J., Xu, J., Shi, R.Z., Lu, X., Clark, T., Johnson, D. and Kim, Y.C., et al. (2005) Detecting novel low-abundance transcripts in Drosophila. RNA 11, 939–946.
[10] Okazaki, Y., Furuno, M., Kasukawa, T., Adachi, J., Bono, H., Kondo, S., Nikaido, I., Osato, N., Saito, R. and Suzuki, H., et al. (2002) Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. Nature 420, 563–573.
[11] Scheetz, T.E., Laffin, J.J., Berger, B., Holte, S., Baumes, S.A., Brown 2nd., R., Chang, S., Coco, J., Conklin, J. and Crouch, K., et al. (2004) High-throughput gene discovery in the rat. Genome Res. 14, 733–741.
[12] Seki, M., Narusaka, M., Kamiy, A., Ishida, J., Satou, M., Sakurai, T., Nakajima, M., Enju, A., Akiyama, K. and Oono, Y., et al. (2002) Functional annotation of a full-length *Arabidopsis* cDNA collection. Science 296, 141–145.
[13] Bao, J., Lee, S., Chen, C., Zhang, X., Zhang, Y., Liu, S., Clark, T., Wang, J., Cao, M. and Yang, H., et al. (2005) Serial analysis of gene expression study of a hybrid rice strain (LYP9) and its parental cultivars. Plant Physiol. 138, 1216–1231.
[14] Wang, S.M., Fears, S.C., Zhang, L., Chen, J.J. and Rowley, J.D. (2000) Screening poly (dA/dT)-cDNAs for gene identification. Proc. Natl. Acad. Sci. USA 97, 4162–4167.
[15] Bertone, P., Stole, V., Royce, T.E., Rozowsky, J.S., Urban, A.E., Zhu, X., Rinn, J.L., Tongprasit, W., Samanta, M. and Weissman, S., et al. (2000) Global identification of human transcribed sequences with genome tiling arrays. Science 306, 2242–2246.
[16] Imanishi, T., Itoh, T., Suzuki, Y., O'Donovan, C., Fukuchi, S., Koyanagi, K.O., Barrero, R.A., Tamura, T., Yamaguchi-Kabata, Y. and Tanino, M., et al. (2004) Integrative annotation of 21,037 human genes validated by full-length cDNA clones. PLoS Biol. 2, 856–875.
[17] Cheng, J., Kapranov, P., Drenkow, J., Dike, S., Brubaker, S., Patel, S., Long, J., Stern, D., Tammana, H. and Helt, G., et al. (2005) Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. Science 20, 1149–1154.
[18] Carninci, P. (2006) Tagging mammalian transcription complexity. Trends Genet., doi:10.1016/j.tig.2006.07.003.
[19] Velculescu, V.E., Zhang, L., Vogelstein, B. and Kinzler, K.W. (1995) Serial analysis of gene expression. Science 270, 484–487.
[20] 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.
[21] Chen, J., Sun, M., Lee, S., Zhou, G., Rowley, J.D. and Wang, S.M. (2002) Identifying novel transcripts and novel genes in the human genome by using novel SAGE tags. Proc. Natl. Acad. Sci. USA 99, 12257–12262.
[22] Chen, J.J., Rowley, J.D. and Wang, S.M. (2000) Generation of longer cDNA fragments from serial analysis of gene expression tags for gene identification. Proc. Natl. Acad. Sci. USA 97, 349–453.
[23] Chen, J., Lee, S., Zhou, G. and Wang, S.M. (2002) High-throughput GLGI procedure for converting a large number of serial analysis of gene expression tag sequences into 3′ complementary DNAs. Genes, Chromos. Cancer 33, 252–261.
[24] Caron, H., van Schaik, B., van der Mee, M., Baas, F., Riggins, G., van Sluis, P., Hermus, M.C., van Asperen, R., Boon, K., Voute, P.A. and Heisterkamp, S., et al. (2001) The human transcriptome map: clustering of highly expressed genes in chromosomal domains. Science 291, 1289–1292.
[25] Pagni, M., Iseli, C., Junier, T., Falquet, L., Jongeneel, V. and Bucher, P. (2001) trEST, trGEN and Hits: access to databases of predicted protein sequences. Nucl. Acids Res. 29, 148–151.
[26] Sperisen, P., Iseli, C., Pagni, M., Stevenson, B.J., Bucher, P. and Jongeneel, C.V. (2004) trome, trEST and trGEN: databases of predicted protein sequences. Nucl. Acids Res. 32, D509–D511.
[27] Dias, N.E., Correa, R.G., Verjovski-Almeida, S., Briones, M.R., Nagai, M.A., da Silva Jr., W., Zago, M.A., Bordin, S., Costa, F.F. and Goldman, G.H., et al. (2000) Shotgun sequencing of the human transcriptome with ORF expressed sequence tags. Proc. Natl. Acad. Sci. USA 97, 3491–3496.
[28] Nam, D.K., Lee, S., Zhou, G., Cao, X., Wang, C., Clark, T., Chen, J., Rowley, J.D. and Wang, S.M. (2002) Oligo(dT) primer generates a high frequency of truncated cDNAs through internal poly(A) priming during reverse transcription. Proc. Natl. Acad. Sci. USA 99, 6152–6156.
[29] Lai, A., Lash, A.E., Altschul, S.F., Velculescu, V., Zhang, L., McLendon, R.E., Marra, M.A., Prange, C., Morin, P.J. and Polyak, K., et al. (1999) A public database for gene expression in human cancers. Cancer Res. 59, 5403–5407.
[30] Ge, X., Jung, Y.C., Wu, Q., Kibbe, W.A. and Wang, S.M. (2006) Annotating nonspecific SAGE tags with microarray data. Genomics 87, 173–180.
[31] Camargo, A.A., Samaia, H.P., Dias-Neto, E., Simao, D.F., Migotto, I.A., Briones, M.R., Costa, F.F., Nagai, M.A., Verjovski-Almeida, S. and Zago, M.A., et al. (2001) The contribution of 700,000 ORF sequence tags to the definition of the human transcriptome. Proc. Natl. Acad. Sci. USA 98, 12103–12108.