Transcriptome Annotation using Tandem SAGE Tags

Eric Rivals, Anthony Boureux, Mireille Lejeune, Florence Ottones, Oscar Pecharromàn Pérez, Jorma Tarhio, Fabien Pierrat, Florence Ruffle, Thérèse Commes, Jacques Marti

Published online 20 August 2007 Nucleic Acids Research, 2007, Vol. 35, No. 17 e108
doi:10.1093/nar/gkm495

Transcriptome annotation using tandem SAGE tags

Eric Rivals[1], Anthony Boureux[2], Mireille Lejeune[2], Florence Ottones[2], Oscar Pecharromàn Pérez[3], Jorma Tarhio[3], Fabien Pierrat[4], Florence Ruffle[2], Thérèse Commes[2,*] and Jacques Marti[2]

[1]Laboratoire d'Informatique, de Robotique et de Microélectronique, UMR 5506 CNRS – Université de Montpellier II, 161 rue Ada, 34392 Montpellier 05, [2]Institut de Génétique Humaine, CNRS UPR 1142, 141 rue de la Cardonille, 34396 Montpellier 05, France, [3]Helsinki University of Technology, P.O. Box 5400, FI-02015 HUT, Finland and [4]Skuld-Tech, 134, rue du Curat – Bat. Amarante, 34090 Montpellier, France

Received April 23, 2007; Revised June 5, 2007; Accepted June 6, 2007

ABSTRACT

Analysis of several million expressed gene signatures (tags) revealed an increasing number of different sequences, largely exceeding that of annotated genes in mammalian genomes. Serial analysis of gene expression (SAGE) can reveal new Poly(A) RNAs transcribed from previously unrecognized chromosomal regions. However, conventional SAGE tags are too short to identify unambiguously unique sites in large genomes. Here, we design a novel strategy with tags anchored on two different restrictions sites of cDNAs. New transcripts are then tentatively defined by the two SAGE tags in tandem and by the spanning sequence read on the genome between these tagged sites. Having developed a new algorithm to locate these tag-delimited genomic sequences (TDGS), we first validated its capacity to recognize known genes and its ability to reveal new transcripts with two SAGE libraries built in parallel from a single RNA sample. Our algorithm proves fast enough to experiment this strategy at a large scale. We then collected and processed the complete sets of human SAGE tags to predict yet unknown transcripts. A cross-validation with tiling arrays data shows that 47% of these TDGS overlap transcriptional active regions. Our method provides a new and complementary approach for complex transcriptome annotation.

INTRODUCTION

Mammalian genome-wide analyses are revealing an increasingly complex transcriptome (1). While predictions concerning the number of human protein-coding genes declined from >100 000 to <30 000 since 2001, transcript number estimations followed an opposite trend (2). Attempts to assemble hundreds of ESTs into clusters expected to map on the same locus, as in UniGene (3), did not eliminate the discrepancy between the small number of protein-coding genes and the large number of detected transcripts. Massively parallel hybridization on already known sequence probes, as in classical microarray technologies, cannot explore the whole transcriptome complexity. For this purpose, new generations of high density arrays have been developed using probes which span a genome region at regular intervals, either overlapping or spaced at defined distances (4,5).

Besides these new open strategies, methods based on sequence signatures (tags) such as serial analysis of gene expression (SAGE) also meet the requirements to provide fresh information on unknown transcripts. SAGE tags are extracted from the 3′ most 4-nt 'anchoring site' of cDNAs. The restriction enzyme that cuts cDNA at this topologically defined sites is usually NlaIII (CATG sites), but Sau3A1 (GTAC sites) may be used as well (6). Starting from this site, stretches of 14 or 21 nt (respectively in conventional SAGE and in LongSAGE) are extracted using Bsmf1 or Mme1 as 'tagging' enzymes (7,8). Tags matching known mRNAs are readily identified and the individual frequency of each tag measures the expression level of its cognate mRNA. As the quality of analysis depends on the number of sequenced tags, SAGE was limited up to now by the cost and capacity of the Sanger technique. However, with the advent of new DNA sequencers, the flow rate of tag-based methods may grow by an order of magnitude with a substantial reduction of time and cost of analysis (9–12) and now it becomes realistic to analyze in parallel larger collections of tags.

In addition to the tags of well-annotated mRNAs, SAGE experiments currently reveal tags unmatched to known transcripts. Their high number cannot be explained simply by sequencing errors or genetic diversity,

*To whom correspondence should be addressed. Tel: +33 4 67 14 42 36; Fax: +33 4 67 14 42 36; Email: commes@univ-montp2.fr
Correspondence may also be addressed to Jacques Marti. Tel: +334 67 144241; Email: jmarti@univ-montp2.fr

and many of them are susceptible to reveal new transcripts. The problem is to map these unmatched tags directly on large genomes. For this purpose, we investigated a new strategy, which consists in building two SAGE libraries from the same biological sample, with tags respectively anchored on the two adjacent CATG and GATC sites located at the 3′-end of each cDNA. We developed a new algorithm for assembling these tandem tag pairs on the genome sequence, defining tag-delimited genomic sequences (TDGS). In a small-scale experiment, we checked the rate of success of this strategy on a sample of well-annotated mRNAs, and starting from previously unmatched tags, we evaluated its ability to reveal new transcripts. In a large-scale analysis, we assembled a collection of TDGS based on the whole set of publicly available human SAGE tags. We found that a part of them mapped on transcription sites also indicated by tiling arrays and in addition we detected novel transcribed loci. In conjunction with other high-throughput approaches, this tandem SAGE tags strategy may help to complete the annotation of genomics regions transcribed into polyadenylated [poly(A)] RNAs.

MATERIALS AND METHODS

External datasets

SAGE data were collected from publicly available repositories [http://www.ncbi.nlm.nih.gov/projects/geo/index.cgi: Platforms: GPL4, GPL6 and GPL1485, http://www.prevent.m.u-tokyo.ac.jp/SAGE.html, CAGP project (Sage genie): ftp://ftp1.nci.nih.gov/pub/SAGE/HUMAN/]. The list of SAGE libraries is available (Supplementary Table 1). *Homo sapiens* chromosome sequences (HG17, NCBI build 35) were retrieved from the UCSC Genome Bioinformatics site (http://genome.ucsc.edu/). UniGene cluster-representative sequences were taken from the Hs.seq.uniq. file, retrieved by FTP from the National Center for Biotechnology Information site (ftp://ftp.ncbi.nih.gov/repository/). We used the UniGene built # 162 assembling 4.47 million sequences into 123 995 clusters and providing the same number of cluster-representative sequences. Since SAGE may detect several authentic transcripts from the same locus, we did not use more recent UniGene releases in which transcripts co-locating with known genes have been merged. Alu sequences were taken from RepBase Update (http://www.girinst.org/Repbase_Update.html) (13).

Macrophage SAGE libraries

Venous blood from healthy donors was obtained from the Etablissement Français du Sang (Montpellier, France). Monocytes, isolated by adherence to culture flasks, were differentiated into >99% Monocyte Derived Macrophages (MDMs) as previously described (14). Total RNA (50 μg) from 8.10^6 MDMs was extracted with Trizol^{TM} (Invitrogen, Cergy Pontoise, France). Poly(A) mRNA was selected by hybridization to oligo (dT) 25-coated magnetic beads according to manufacturer's instructions (Dynal, Compiegne, France). CATG-tags were prepared using the I-SAGE kit (Invitrogen, Cergy Pontoise, France) and GATC-tags using a modified Sau3A1 SAGE procedure (6). The sequences of 22 387 CATG-tags and 8221 GATC-tags determined by the Centre National de Séquençage (Evry, France) were analyzed for tag detection and counting using the C + tag software (Skuld-Tech, France).

Computational analyses

The virtual SAGE analysis of UniGene cluster-representative sequences was performed using the Preditag software (Skuld-tech, Montpellier, France, http://www.skuldtech.com) as described (15). For each sequence, the tag expected to be observed in a SAGE analysis, i.e., the one originating from the first anchoring site starting from the 3′-end of the sequence, was registered as Rank 1 tag (R1). We also registered tags from upstream anchoring sites (R2, R3, R4) susceptible to reveal technical artifacts or alternative transcripts, and tags read on the opposite strand (AS1 to AS4), which may reveal antisense transcripts (16). We performed this procedure for both CATG and GATC anchoring sites. From the previous set, we selected high quality R1 tags according to the following criteria: RefSeq annotation or mention of a full-length mRNA, known chromosomal location, absence of Alu sequence in the tagged site. Hereafter, a tag will be referred to as a C-tag if anchored on a CATG site (using NlaIII as anchoring enzyme) or as a G-tag if anchored on a GATC site (using Sau3A1).

The algorithm used to assemble tag pairs on the genome is depicted in Figure 1A. It takes as input two sets of experimental tags, one of C-tags and one of G-tags, and retrieves all combinations of successive 5′G-3′C and 5′C-3′G tag pairs on the genome. The algorithm follows three rules. First, each transcript must possess both restriction sites. Among RefSeq mRNAs, we found 4.6% lacking one of them. Second, both sites may be found in any order, implying that two sets of oriented pairs, 5′G-3′C and 5′C-3′G, must be generated. Third, each tag is anchored on the most 3′ restriction site. Therefore, if a G-tag is located in 5′ relatively to the most 3′ C-tag of the transcript, there is no intervening G-tag between them. This assertion holds for the processed transcript but not for genomic DNA, since tags may be located on distinct exons. Because 4-bp restriction sites are frequent, scanning introns will necessarily detect false-positive tags. To alleviate this problem, the genome is scanned using actually observed experimental SAGE tags, so that irrelevant sequences may be skipped over. Intronic 14-bp stretches will be registered only if they are fortuitously identical to real tags (Figure 1C).

For assembling G–C tag pairs, the chromosome sequence is read from the 5′ to the 3′-end. Each occurrence of CATG is searched with a variation (17) of the Boyer–Moore–Horspool algorithm (18). Then it is checked whether CATG with the next 10 symbols matches a tag of the experimental list. This is performed with a hash table holding the variable parts of tags (the 10 nt suffix). Once a C-tag is located, the sequence is scanned again from 5′ toward 3′ and in the same way to find the 3′ most experimental G-tag preceding the C-tag. The chromosomal co-ordinates of this G–C tag pair is

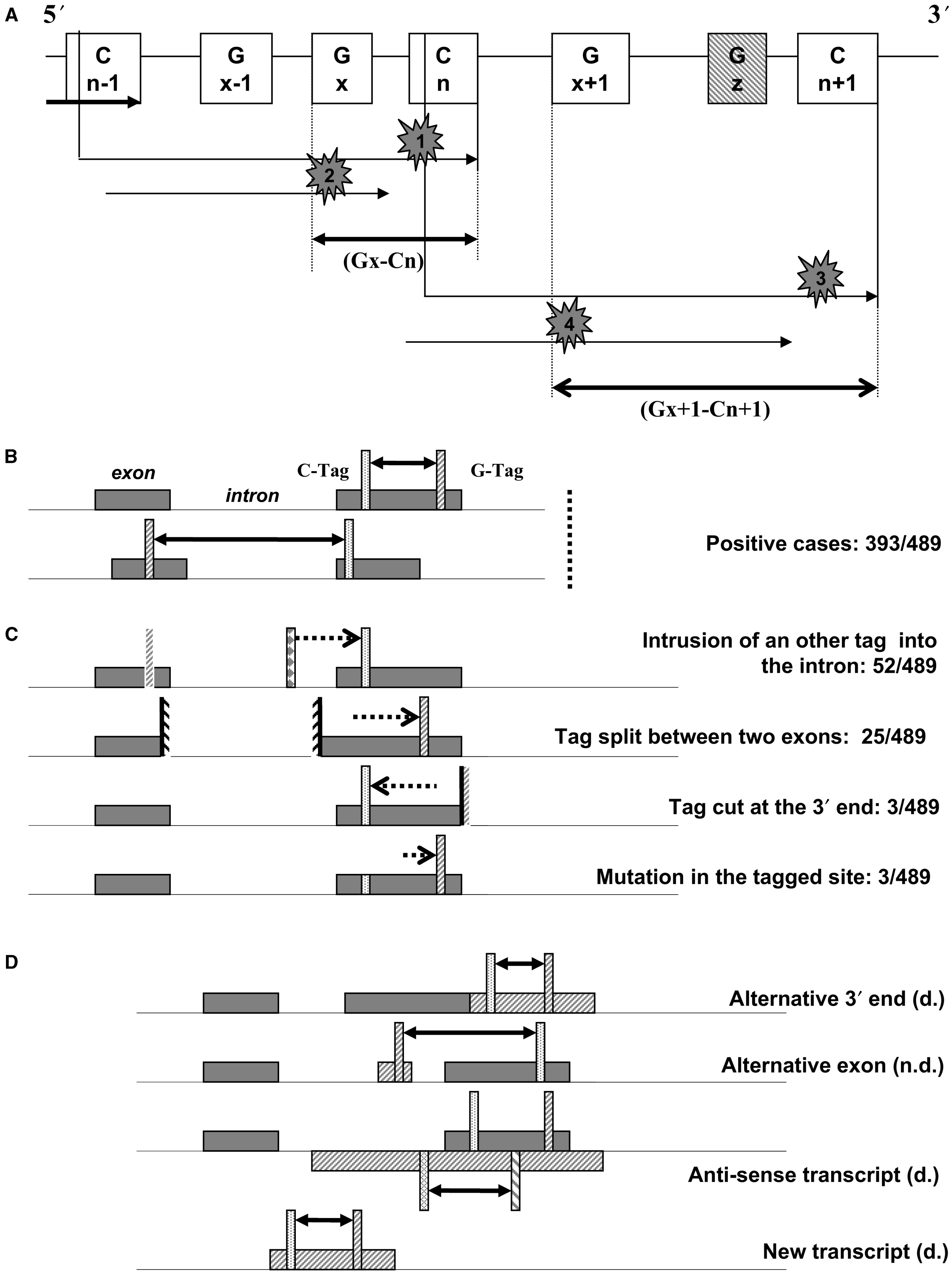
A
5′
3′
C n-1
G x-1
G x
C n
G x+1
G z
C n+1
1
2
3
4
(Gx-Cn)
(Gx+1-Cn+1)
B
exon
intron
C-Tag
G-Tag
Positive cases: 393/489
C
Intrusion of an other tag into the intron: 52/489
Tag split between two exons: 25/489
Tag cut at the 3′ end: 3/489
Mutation in the tagged site: 3/489
D
Alternative 3′ end (d.)
Alternative exon (n.d.)
Anti-sense transcript (d.)
New transcript (d.)

Figure 1. (**A**) Schematics of search for tag-delimited genomic sequences (TDGS, double arrows). Upper part: procedure for assembling 5′G- 3′C pairs. Starting from the previously identified C-tag ($n-1$), the program searches the next site on which an experimental C-tag (n) can be positioned (star 1). The genome sequence is then scanned for G-tags ($x-1$, x) and stops (star 2) when the shortest G–C pair is found. In search of the next pair (stars 3 and 4), the G-tag (z) potential tag sequence is skipped because it does not match any G-tag in the experimental dataset. Lower part (**B**, **C** and **D**): illustration of the various causes of success and failure in assembling TDGS. Numerical values in B and C are taken from the study of 489 well-annotated sequences identified in the macrophage SAGE libraries. Cases schematized in D are detailed (d) in Figures 3 and 4.

then recorded together with the sequence comprised between the two tags, which we call a TDGS. The search for the next pair resumes on the nucleotide position following the C-tag anchoring site. G–C pairs are assembled on both DNA strands and the search is iterated for C–G pairs in a similar way. With the larger sets of G-tags and C-tags (106 748 and 619 771 tags, respectively), the search on the full set of human chromosomes required <2 h on a Pentium processor at 1.5 GHz running Linux with 256 megabytes of main memory. The program is available at the following URL: http://www.lirmm.fr/~rivals/GENOMICS.

Proximity between TDGS and tiling arrays data

We retrieved tiling arrays data from the UCSC Genome Bioinformatics site (http://genome.ucsc.edu/). We used transcriptional active regions (TARs) data from Affymetrix Transcriptome Project Phase2, Affymetrix Poly(A)$^+$ RNA transfrags, Yale RNA TARs and Yale Maskless Array synthesizer experiments (5,19). We computed the number of TDGS that either strictly overlap a TAR, or are in a 500-bp vicinity of a TAR.

RESULTS

Overall structure of SAGE data

We assembled two sets of 270 and 15 libraries built with NlaIII and Sau3A1, respectively as anchoring enzymes, associating local data and a large number of publicly available SAGE human libraries (Supplementary Table 1). This collection, assembling 13.7 million C-tags and 0.5 million G-tags, will be called UniSAGE hereafter. The total number of distinct C-tags registered in UniSAGE is 619 770, i.e. 59% of the number of all possible 10-nt combinations (4^{10}), largely exceeding the number of UniGene clusters and the small number of well-annotated mRNAs. A salient feature is that individual tag counts evenly decrease, from a large number of tags observed only once to a small number of highly abundant tags (Figure 2). Similar distributions are observed in individual libraries and in the dataset obtained by summing all UniSAGE tag counts. As illustrated in Figure 2, most unidentified tags are observed at low levels and tags matching RefSeq sequences are the most abundant ones. Among the set of tags expressed at low levels, distinguishing between biological and artifactual ones prompts for novel approaches.

Discrepancy between the numbers of tags and of known genes

Several sources of inflation make rare biological tags indistinguishable from artifacts, including infidelities in reverse transcription and PCR, or inaccuracies in single-run sequencing. To estimate a global error rate, we simulated 1-nt errors on R1 tags of widely expressed genes. We sampled four proteins (EEF1A1 and three ribosomal proteins) for which C-tag variants were uniformly distributed at low level among libraries, thus providing no evidence of them being shared with other abundant transcripts. In each case, all 30 tags differing by 1 nt from the canonical tag were observed more than once in UniSAGE. As a whole, these variants accounted for 6–7% of R1 tag counts (mean 6.6%). Individual variant frequencies varied depending on the substitution position. However, as frequency distributions were similar in the four samples, we could derive an empirical curve fitting well with their mean distribution (R^2 = 0.99). Applying this model to the whole UniSAGE dataset, we estimated the total number of 1-nt variants at nearly 160 000 C-tags. The lack of efficiency of the anchoring enzyme may also generate additional tags generated from upstream sites. R2 tags for widely expressed genes were usually found at low levels. Wide differences from gene to gene made problematic an estimation of their mean frequency. However, tentatively assuming that all transcripts counted at least 100 times in the sum of all libraries generate R2 tags, we may estimate that some 15 000 R2 C-tags are probably registered in UniSAGE. Genetic diversity also increases tag numbers. Single nucleotide polymorphisms (SNPs) either modify tag sequence or create new tags if the anchoring site itself harbors SNPs. This problem has already been investigated elsewhere, the analysis of 54 645 mRNAs from UniGene built #163 revealed 8.6% of SNP-associated alternative tags (20). Altogether, these multiple causes of inflation explain far less than a half of the huge number of different tags registered in UniSAGE. It thus seems unavoidable to conclude that a large number of tags originate from authentic transcripts.

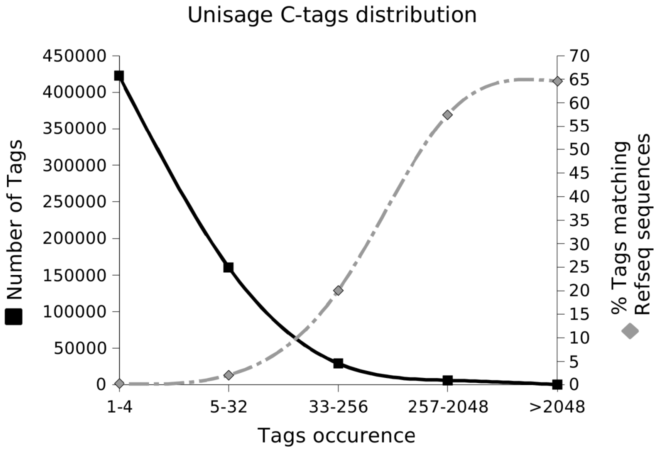

Figure 2. Number of distinct C-tags (left *y*-axis, black square) in five consecutive classes of occurrence, i.e. abundance, (*x*-axis) from 1 to >2048 counts summed up over UniSAGE, and percentage of tags matching RefSeq annotated sequences (right *y*-axis, gray diamonds).

Analysis of twin SAGE libraries

Causes of failure and rate of success in search of already known genes. The algorithm illustrated in Figure 1A was used to assemble a set of 8085 different C-tags and 4217 G-tags obtained from twin libraries built in parallel from a unique macrophage RNA sample. The different C-tags and G-tags were mapped as C–G and G–C tandem pairs on the genome. To test the algorithm efficiency, we sampled 489 gene sequences selected from the UniGene

Table 1. Evaluation of the ability of our algorithm or LongSAGE tags to localize on the genome.

Panel A:

	Total analyzed	Not found	Located		
			Unique	⩾2	Sum
LongSAGE (Number of tags)	98 142	34 659	51 933	11 550	63 483
%	100%	35%	53%	12%	65%
Tandem—UniSAGE tags sets	11 998	3110	8189	699	8888
%	100%	26%	68%	6%	74%
Tandem—Macrophage twin libraries	489	96	336	57	393
%	100%	20%	69%	12%	80%

Panel B:

Tandem—Macrophage twin libraries causes of localization failure for known transcripts	Total	Found elsewhere	Tag intrusion in intron	Tag at the junction of 2 exons	Tag in Poly A tail	SNP in tag
Number of transcripts	96	13	52	25	3	3
%	100%	14%	54%	26%	3%	3%

(Panel **A**) Values for both methods to localize tags or TDGS, using tags recognized as matching well-known gene transcripts (also call R1-tags) on UniSAGE or Macrophage twin libraries data or tags with occurrence more than once for LongSAGE.

(Panel **B**) Evaluation of the ability of our algorithm or LongSAGE tags to localize on the genome. Causes of localization failure on Macrophage twin libraries with the tandem algorithm.

collection for providing high quality R1 tag pairs (Table 1A). We observed 393 positive cases (Figure 1B) and 96 cases of discrepancy (Figure 1C and Table 1B) between this test sample and the set of genomic pairs: 13 tag pairs were not detected on the chromosome indicated by UniGene but elsewhere in the genome on a related pseudo-gene, and 83 pairs were not found on the genome. The major case of failure (52 pairs, 10.6%) was the intrusion of another tag in the intervening intron, causing abortion of the pairing process between tags located on separate exons. For 25 pairs (5.1%), one of the tags was undetectable in the genome because, being created by the junction of two exons, it exists only in the processed transcript. In three cases, one of the anchoring sites was located at the end of the transcript, the full-length tag being created by poly(A) tail extension. In three other cases, the UniGene and genome sequences differed from the genome by the presence of a SNP. The chromosomal locations indicated by UniGene and found by genome scanning were identical for 393 pairs, i.e. 80% of the whole test collection. We found a unique chromosomal location for 336 pairs (85.5% of these successfully assembled tag pairs, 69% of the whole test sample) and additional genome sites for 57 of them (Table 1A and B).

New transcripts detected by pairing unmatched tags. We collected experimental tags considered as unmatched tags because they match neither any R1 to R4 tags, nor their antisense counterparts (AS1 to AS4), nor any Alu sequences. Tags according to these criteria were involved in 251 tag pairs. TDGS showed variable length, from four nucleotides for two overlapping C- and G-tags to 20 000 nt, with 50% of them not exceeding 2800 nt. We used the functions of the UCSC Genome Browser (http://genome.ucsc.edu/) to get a representation of these TDGS in their chromosomal context and investigated in more details their properties. Although Alu-matching tags had been removed, we still found TDGS overlapping repeated elements. One G-tag has generated tag pairs with 56 different C-tags. The annotation revealed that this G-tag maps on a LINE repeat. The 187 remaining sequences resulted from the assembly of 136 C-tags and 118 G-tags, with all tags having at most five matches in the genome. This suggests that requiring a low multiplicity of matches for each tag should help filtering out TDGS generated by abundant genomic repeats. The 187 sequences were sorted by individual inspection into three classes. Class 1 contains 14 sequences matching well-annotated genes for which UniGene did not provide the expected reference sequence. Class 2 (Figure 1D) contains 39 TDGS mapping in close vicinity of a known locus. Class 3 (Figure 1D) collects 134 TDGS (71.7%) mapping in genomic regions lacking indications for transcription sites, or bearing an expressed sequence, but in inverted orientation (e.g. antisense transcript).

Case studies. We successfully confirm by RT–PCR 50% (14 out of 27) of a subset of tested cases (Supplementary Table 2). An example of Class 1 sequence is provided by TDGS # 227. We found it mapping within the locus of Gelsolin (GSN: NM_198252), a well-known component of the macrophage. An example of Class 2 sequence is provided by TDGS # 20 located near the coding region of CDH23 (Figure 3A). We confirmed by RT–PCR the expression of this transcript in the macrophage: we amplified the end of the 3′cDNA using primers designed on the genomic sequence. The analysis of the amplicon sequence validated the existence of an alternative polyadenylation site. Class 3 contains antisense transcripts of known genes. TDGS # 212 was found mapping in a

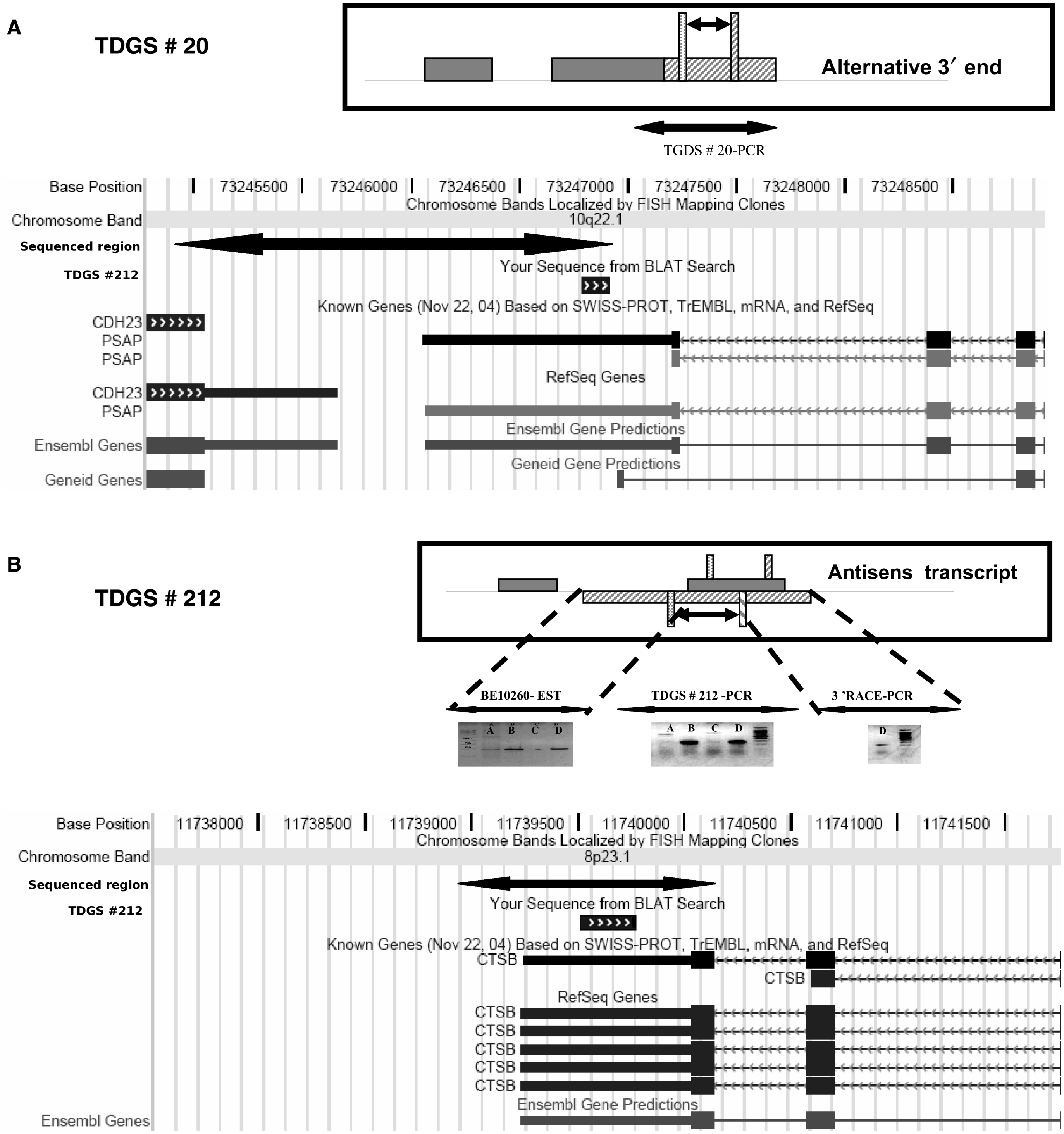

Figure 3. Characterization and annotation of validated TDGS. Alignments of the TDGS# 20 and # 212 to the UCSC human genome browser. For RT–PCR validation, Macrophage poly(A)+ RNA were extracted from MDM and the cDNA were synthesized using mRNA and oligo-dT primer. (**A**) TDGS# 20 corresponds to an example of Class 2 transcript localized near the coding region of CDH23. For PCR, a primer pair was respectively designed in the 3′-end of CDH23 and in the TDGS # 20. The existence of this new variant transcript was confirmed in macrophage by sequencing. (**B**) TDGS # 212 is an example of class 3 transcript. Experiments without reverse transcription (A, C) and with DNAse treatment (C, D) were performed to detect DNA contamination. For transcript validation, a first PCR was realized with primers pairs designed on TDGS # 212 and a second one with primers respectively in the 3′-end of EST EB10260 and TDGS #212. The 3′-end of the transcript was validated by 3′RACE (3′RACER kit, Invitrogen, France). The sequenced PCR products validated the existence of a transcript in inverted orientation relatively to the Cathepsin B gene (CTSB, NM_0019082).

region where the UCSC Browser indicated numerous ESTs mapping in both orientations. RT–PCR and RACE extension of macrophage poly(A)$^+$RNA confirmed the existence of a transcript in inverted orientation relatively to the Cathepsin B gene (CTSB, NM_0019082, Figure 3B). TDGS # 5, 6, and 54 illustrate a more complex situation (Figure 4A). We registered TDGS # 5 in our Class 1 because it maps on the site of a computer-predicted sequence on chromosome 1, corresponding to a full-length cDNA newly registered in GenBank as CR601947. TDGS # 6, also in Class 1, shares its C-tag with TDGS # 5, but its sequence is longer (821 nt instead of 376) because its G-tag is located in 3′ of the TDGS # 5 one. Apparently, TDGS # 5 and 6 correspond to two alternative transcripts from the same locus on chromosome 1 (Figure 4B). In addition, TDGS # 54, registered in Class 3, reveals another site, in a region of chromosome 14. TDGS # 5 and TDGS # 54 are delimited by identical C- and G-tags, have exactly the same length (376 nt) and the BLAST algorithm indicates 93% homology between the two sequences, which differ only by 20 nt. To assess the transcription of this intervening region, we checked it against a tiling array database. We found (Figure 4C) that the sequence of chromosome 14 matched a sequence registered in the Affymetrix dataset harbored at UCSC genome site (Affy Txn Phase2) and annotated as being expressed from the same site (5). Finally having sequenced the PCR product, we could conclude that the transcript originates from the chromosome 14 locus.

In silico experiment with the whole UniSAGE data set

The algorithm was used to assemble the whole set of UniSAGE-registered C- and G-tags. We first tested its efficiency on a sample of 11 998 gene sequences, selected from the UniGene collection for providing high quality R1 tag pairs (Table 1A). The algorithm failed to found a chromosomal tag pair for 3110 of them (26%) and succeeded in identifying 8888 gene sequences (74%). Among them, 8189 (68%) were assigned to a unique genome site. In comparison, when searching for the subset of R1 LongSAGE tags (70 284 tags, 8% of all LongSAGE tags), 71% identify a unique genomic location. With the whole dataset of the LongSAGE tags (98 142 tags observed more than once), this percentage drops to 53% (Table 1A). Experimental tags considered as unmatched were collected according to the same criteria as described above. In UniSAGE, 321 498 C-tags and 49 103 G-tags fall in this category. Working on the subset of tags found at least three times in the sum of all libraries, the algorithm assembled 93 859 potential tag pairs on the genome. We evaluated the TDGS length of unmatched tags pairs and compared them to well-annotated TDGS obtained from high quality R1 tags pairs (Figure 5). Well-annotated TDGS extend up to 6000 nt. Unmatched TDGS span from 4 to 20 000 nt; however, more than 84% do not exceed 6000 bases. Each tag being present several times in the genome can be involved in several pairs. A direct examination being unpractical in this case, we cross-checked the set of TDGS with tiling arrays data. We computed the proximity between each TDGS and transcriptionally-active regions (TARs) from tiling arrays and found 43 813 TDGS (47%) overlap a TAR, and 65 808 (70%) are located <500 bp away from a TAR.

DISCUSSION

In the present work, we developed a new algorithm associating pairs of gene expression signatures to localize their position on the genome. This work was motivated by studies on a large SAGE dataset (UniSAGE) showing a discrepancy between the number of loci for well-annotated genes and the large number of potential transcripts suggested by the number of tags. Using the whole set of presently available NlaII and Sau3A1 individual SAGE tags (619 000 and 106 748, respectively), this algorithm, efficient enough to process on a standard desktop computer, predicted 93 859 potentially transcribed sites in the human genome. This observation corroborates independent evaluations based on the prediction of functional transcription units and on the experimental results of tiling arrays (21–24).

Genetic polymorphism and technical errors do not account for all unmatched SAGE tags. Apart from a non-specific natural transcription noise, they may reveal genuine transcripts justifying more thorough investigations. As an open method, SAGE may indeed reveal transcripts never observed before because they are expressed in rare physiological conditions or in unique cell types, such as the terminally differentiated macrophage studied in the present work. Allowing retrospective comparisons of large datasets, it enables to distinguish a uniformly distributed transcription noise from the controlled expression of tissue-specific transcripts. Its main limitation is to detect only Poly(A) transcripts. Other wet-lab methods exist for other kinds of RNA molecules (25,26).

The algorithm described here may used for assembling any pair of tags irrespective of their length (14–21 bp) or position (5′ or 3′) on the transcript. For instance, it could be used in the context of the new method developed to form paired-end ditags (PETs) in which the 5′ and 3′ tags defining both ends of cDNAs are physically linked and sequenced together (27); this method is valuable for accurate transcript demarcation but it requires full-length cDNAs, which may be difficult to synthesize. Moreover, our algorithm could use data simultaneously collected from different cDNAs tags technology [SAGE and derivates technologies as Massively Parallel Signature Sequencing (MPSS) (28)] However, the approach investigated here, based on the current SAGE technology, is technically less demanding, particularly with the new DNA sequencers available today (see the Introduction section).

While tags matching well-annotated mRNAs are easily recognized, most tags are in any case difficult to locate directly on the genome. Individual 14-bp tags cannot be mapped unambiguously since large genomes are statistically expected to contain multiple copies of any 14-bp stretch. Theoretically, a 21-bp tag should occur only once per genome if nucleotides were

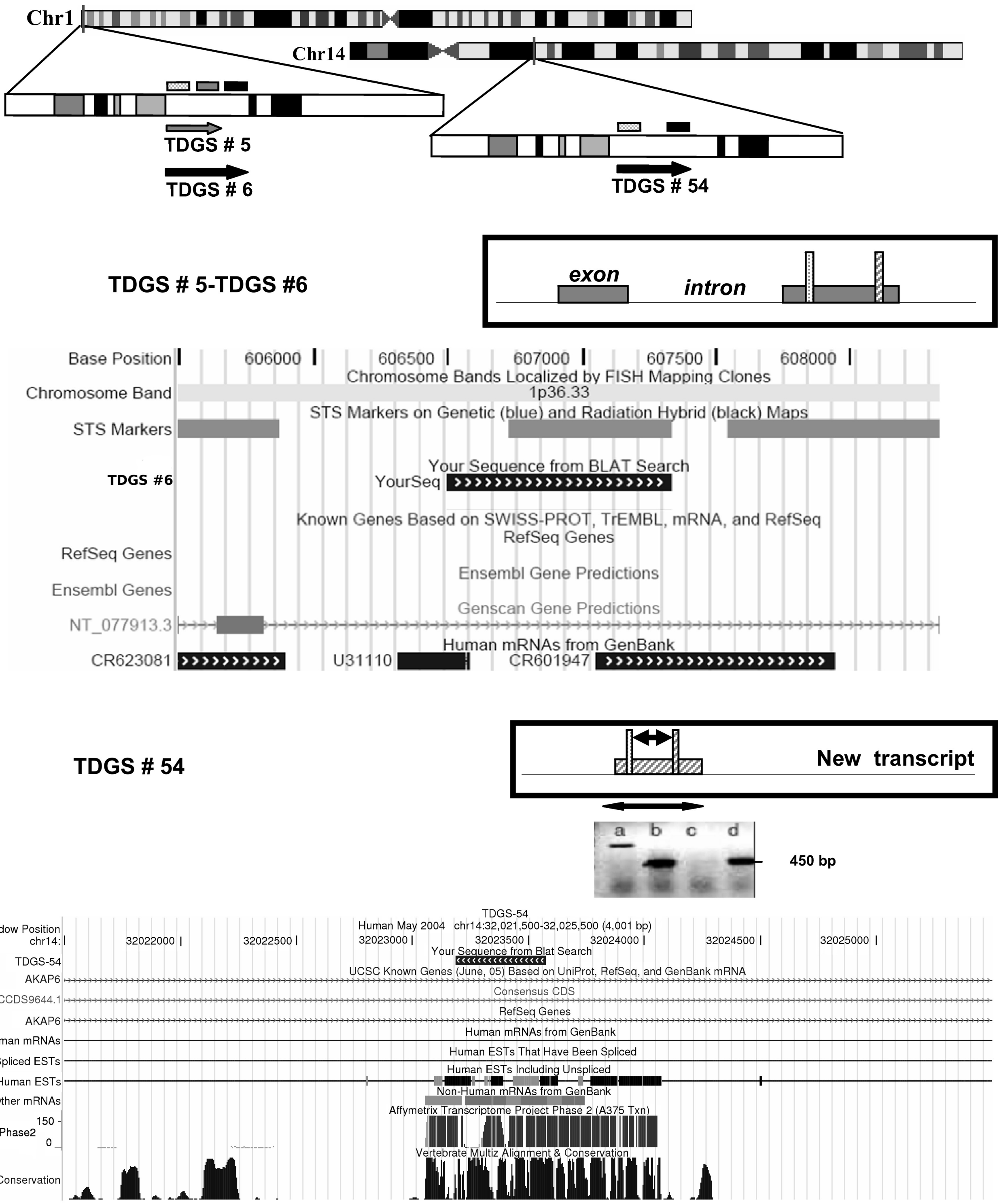

Figure 4. Characterization and annotation of validated TDGS. Alignments of the TDGS# 5, 6, 54 to the UCSC human genome browser. For RT–PCR validation, Macrophage poly(A)$^+$ RNA were extracted from MDM and the cDNA were synthesized using mRNA and oligo-dT primer. (**A** and **B**) TDGS # 5 (376 bp) maps on chomosome 1, corresponding to the sequence of new full-length cDNA registered in GenBank as CR601947. TDGS # 6, shares with TDGS # 5 the same C-tag (dotted boxes) but its sequence is longer (821 bp) because its G-tag is located in 3′ of the TDGS # 5 one. (A and **C**) In addition, TDGS # 54, reveals another site, in a region of chromosome 14. The same conditions as described in Figure 4 were used to PCR validation. RT–PCR analysis followed by sequence checking confirm the existence of this new transcript. The sequence of chromosome 14 matched a sequence registered in the Affymetrix dataset harbored at UCSC (Affy Txn Phase2)

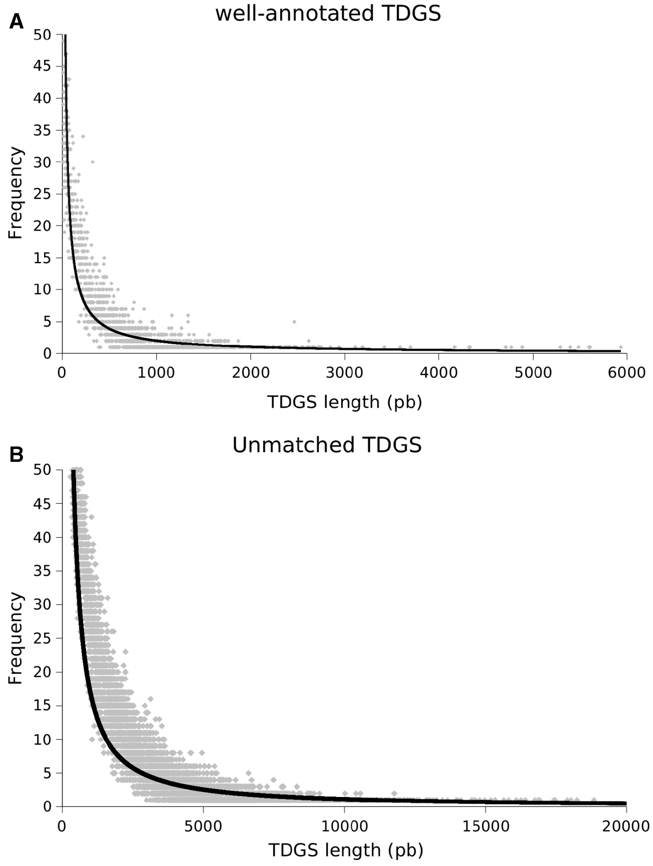

Figure 5. Length of the TDGS assembled from the whole UniSAGE data. Frequency of TDGS length (in base pairs) on well-known annotated TDGS (R1 TDGS) (**A**) and on unkown TDGS (unmatched TDGS) (**B**). Each point represents a TDGS (in gray) (~99 and 92.5% TDGS are shown, respectively), and the curve (in black) is a power regression curve.

randomly distributed. However, the benefit of LongSAGE is limited because extending sequence length increases the rate of technical errors, the number of mismatches due to genetic polymorphism and the risk of tags being split in two exons. It must be stressed that among all 21-bp tags sequenced up to now, 89% have been observed only once and that only 8% of all tags correspond to well-known genes. Even in the case of well-annotated mRNAs, 15% of 21-bp tags cannot be mapped directly on the genome. As a whole, we found 66% of LongSAGE tags unassigned to the published human genome sequence, in agreement with other groups who found 70% on a preliminary draft version and 64.5% on a more recent release, respectively (8,29).

The presence of repeated elements inserted in the genome limits the possibility to locate tags on a unique position. In the human genome, 71% of 21-bp tags are observed only once instead of 99.83% if nucleotides were distributed at random (8). In the present work, we tested the algorithm with pairs of 14-bp classical SAGE tags, thus requiring perfect matches on 28 positions. For simplification, we put aside individual tags matching Alu sequences inserted in the 3′ non-coding region of multiple human mRNAs. Nevertheless, we still observed tags involved in multiple pairs. These repeats inflate the number of TDGS but cannot be rejected as erroneous since insertion elements and pseudogenes may actually be transcribed. Finally, we found 68% of TDGS matching a unique site, close to the 71% observed for individual 21-bp tags. Whatever the method, either based on physical hybridization (as in tiling arrays) or *in silico* searches (in the present case), it is obvious that insertion elements complicate the interpretation of transcriptome data (30).

Using a test sample of experimental tags obtained from twin macrophage libraries, we detected 187 potentially new transcripts. Among them, 39 appeared as alternative transcripts of known genes, while 134 potential sites were found in intergenic regions or in antisense orientation of known genes. The hypothesis that a newly detected TDGS identifies a novel site is initially based on the presence of the two experimental tags at both ends of the sequence. At this stage, false-positive cases are unavoidable and additional data are needed to confirm the expression of the intervening sequence. The classical solution is to design primers in the region defined by a TDGS, perform RT–PCR, and sequence the resulting amplicon, as in GLGI method (31). We confirmed by this the existence of five new transcription variants of known genes and 10 novel transcripts, i.e., 50% of candidates. In other cases, we did not detect amplicons or found complex patterns. Among technical causes of failure, natural catabolic products may inhibit amplification and primers may be captured by irrelevant transcripts. Nevertheless, it must be stressed that a positive rate of 50% offers the possibility to identify thousands of new biologically relevant transcription sites at the whole genome scale.

Although individual validations provide definitive evidence, the interest of high-throughput strategies is lost in this time-consuming approach. Other strategies may help to select rapidly the best candidates. The size of TDGS is by itself informative and may be used to classify them, assuming that the shorter ones have a higher probability to match genuine transcripts. Another round of selection can be based on comparisons with independent datasets. The ENCODE project plans to use tiling arrays as a major tool for human genome annotation (32). Here, we showed the possibility to connect efficiently both kinds of data, a task very difficult with classical microarray and conventional SAGE data. We found a 47% overlap between our TDGS collection and TARs. This result shows that the tandem SAGE strategy corroborates for a part the results of tiling arrays and enables to reveal new transcripts having escaped from other detection systems. As a whole, these results emphasize the importance to combine independent and complementary methods for thoroughly exploring the transcribed part of the genome.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors wish to thank Laurent Manchon (Skuld-Tech) for his help in processing SAGE data. This work was supported by Sidaction and the ANRS. M.L. was the recipient of a scholarship from the Ligue Régionale contre le Cancer Languedoc-Roussillon. The work of J.T. and E.R. was supported by the Academy of Finland. E.R. thanks the Department of Computer Science of the University of Helsinki for a travel fellowship. We acknowledge the support of the Languedoc-Roussillon Génopole®. Funding to pay the Open Access publication charges for this article was provided by ACI IMPBIO Repevol (CNRS).

Conflict of interest statement. None declared.

REFERENCES

1. Carninci,P., Kasukawa,T., Katayama,S., Gough,J., Frith,M.C., Maeda,N., Oyama,R., Ravasi,T., Lenhard,B. *et al.* (2005) The transcriptional landscape of the mammalian genome. *Science*, **309**, 1559–1563.
2. Claverie,J.M. (2005) Fewer genes, more noncoding RNA. *Science*, **309**, 1529–1530.
3. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., DiCuccio,M., Edgar,R. *et al.* (2006) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **34**, D173–D180.
4. Bertone,P., Stolc,V., Royce,T.E., Rozowsky,J.S., Urban,A.E., Zhu,X., Rinn,J.L., Tongprasit,W., Samanta,M. *et al.* (2004) Global identification of human transcribed sequences with genome tiling arrays. *Science*, **306**, 2242–2246.
5. Cheng,J., Kapranov,P., Drenkow,J., Dike,S., Brubaker,S., Patel,S., Long,J., Stern,D., Tammana,H. *et al.* (2005) Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science*, **308**, 1149–1154.
6. Virlon,B., Cheval,L., Buhler,J.M., Billon,E., Doucet,A. and Elalouf,J.M. (1999) Serial microanalysis of renal transcriptomes. *Proc. Natl Acad. Sci. USA*, **96**, 15286–15291.
7. Velculescu,V.E., Zhang,L., Vogelstein,B. and Kinzler,K.W. (1995) Serial analysis of gene expression. *Science*, **270**, 484–487.
8. 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.
9. Margulies,M., Egholm,M., Altman,W.E., Attiya,S., Bader,J.S., Bemben,L.A., Berka,J., Braverman,M.S., Chen,Y.J. *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376–380.
10. Shendure,J., Mitra,R.D., Varma,C. and Church,G.M. (2004) Advanced sequencing technologies: methods and goals. *Nat. Rev. Genet.*, **5**, 335–344.
11. Nielsen,K.L., Hogh,A.L. and Emmersen,J. (2006) DeepSAGE – digital transcriptomics with high sensitivity, simple experimental protocol and multiplexing of samples. *Nucleic Acids Res.*, **34**, e133.
12. Quéré,R., Manchon,L., Pierrat,F., Ludewig,U., Nesch,G., Frey,B., Commes,T., Marti,J. and Piquemal,D. (2007) Rapid and Accurate Pyrosequencing of Serial Analysis of Gene Expression Ditags. *Roche Application Note*, **4**, 2–8.
13. Jurka,J., Kapitonov,V.V., Pavlicek,A., Klonowski,P., Kohany,O. and Walichiewicz,J. (2005) Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.*, **110**, 462–467.
14. Woelk,C.H., Ottones,F., Plotkin,C.R., Du,P., Royer,C.D., Rought,S.E., Lozach,J., Sasik,R., Kornbluth,R.S. *et al.* (2004) Interferon gene expression following HIV type 1 infection of monocyte-derived macrophages. *AIDS Res. Hum. Retroviruses*, **20**, 1210–1222.
15. Piquemal,D., Commes,T., Manchon,L., Lejeune,M., Ferraz,C., Pugnère,D., Demaille,J., Elalouf,J.M. and Marti,J. (2002) Transcriptome analysis of monocytic leukemia cell differentiation. *Genomics*, **80**, 361–371.
16. Quéré,R., Manchon,L., Lejeune,M., Clément,O., Pierrat,F., Bonafoux,B., Commes,T., Piquemal,D. and Marti,J. (2004) Mining SAGE data allows large-scale, sensitive screening of antisense transcript expression. *Nucleic Acids Res.*, **32**, e163.
17. Tarhio,J. and Peltola,H. (1997) String matching in the DNA alphabet. *Software: Practice and Experience*, **27**, 851–861.
18. Horspool,R.N. (1980) Practical fast searching in strings. *Software: Practice and Experience*, **10**, 501–506.
19. Rinn,J.L., Euskirchen,G., Bertone,P., Martone,R., Luscombe,N.M., Hartman,S., Harrison,P.M., Nelson,F.K., Miller,P. *et al.* (2003) The transcriptional activity of human Chromosome 22. *Genes Dev.*, **17**, 529–540.
20. Silva,A.P., De Souza,J.E., Galante,P.A., Riggins,G.J., De Souza,S.J. and Camargo,A.A. (2004) The impact of SNPs on the interpretation of SAGE and MPSS experimental data. *Nucleic Acids Res.*, **32**, 6104–6110.
21. Semon,M. and Duret,L. (2004) Evidence that functional transcription units cover at least half of the human genome. *Trends Genet.*, **20**, 229–232.
22. Johnson,J.M., Edwards,S., Shoemaker,D. and Schadt,E.E. (2005) Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments. *Trends Genet.*, **21**, 93–102.
23. Bertone,P., Gerstein,M. and Snyder,M. (2005) Applications of DNA tiling arrays to experimental genome annotation and regulatory pathway discovery. *Chromosome Res.*, **13**, 259–274.
24. Mockler,T.C., Chan,S., Sundaresan,A., Chen,H., Jacobsen,S.E. and Ecker,J.R. (2005) Applications of DNA tiling arrays for whole-genome analysis. *Genomics*, **85**, 1–15.
25. Cummins,J.M., He,Y., Leary,R.J., Pagliarini,R., Diaz,L.A.Jr, Sjoblom,T., Barad,O., Bentwich,Z., Szafranska,A.E. *et al.* (2006) The colorectal microRNAome. *Proc. Natl Acad. Sci. USA*, **103**, 3687–3692.
26. Huttenhofer,A. and Vogel,J. (2006) Experimental approaches to identify non-coding RNAs. *Nucleic Acids Res.*, **34**, 635–646.
27. Ng,P., Wei,C.L., Sung,W.K., Chiu,K.P., Lipovich,L., Ang,C.C., Gupta,S., Shahab,A., Ridwan,A. *et al.* (2005) Gene identification signature (GIS) analysis for transcriptome characterization and genome annotation. *Nat. Methods*, **2**, 105–111.
28. Brenner,S., Johnson,M., Bridgham,J., Golda,G., Lloyd,D.H., Johnson,D., Luo,S., McCurdy,S., Foy,M. *et al.* (2000) Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat. Biotechnol.*, **18**, 630–634.
29. Ge,X., Wu,Q., Jung,Y.C., Chen,J. and Wang,S.M. (2006) A large quantity of novel human antisense transcripts detected by LongSAGE. *Bioinformatics*, **22**, 2475–2479.
30. Bertone,P., Trifonov,V., Rozowsky,J.S., Schubert,F., Emanuelsson,O., Karro,J., Kao,M.Y., Snyder,M. and Gerstein,M. (2006) Design optimization methods for genomic DNA tiling arrays. *Genome Res.*, **16**, 271–281.
31. 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 Chromosomes Cancer*, **33**, 252–261.
32. ENCODE. (2004) The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science*, **306**, 636–640.