Combining SAGE Tags to Predict Genomic Transcribed Regions

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

Combining SAGE tags to predict genomic transcribed regions

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], Jacques Marti[2] and Thérèse Commes[2]

[1] LIRMM, Univ. Montpellier II, CNRS, Montpellier, France, {rivals}@lirmm.fr
[2] Institut de Génétique Humaine, CNRS Montpellier, France, {marti,commes}@univ-montp2.fr
[3] Helsinki University of Technology, Finland
[4] Skuld-Tech, Montpellier, France

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 RNAs transcribed from previously unrecognized genomic 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, 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 tag-delimited genomic sequences overlap transcriptional active regions. Our method provides a new and complementary approach for complex transcriptome annotation.*

Keywords: transcriptome, antisens RNA, high-throughput, pattern matching, experiments

1 Introduction

Mammalian genome-wide analyses are revealing an increasingly complex transcriptome [1,3]. While predictions concerning the number of human protein-coding genes declined from more than $100,000$ to less than $30,000$ since 2001, transcript number estimations followed an opposite trend [2] . Attempts to assemble ESTs into clusters expected to map on the same locus, as in UniGene, did not eliminate the discrepancy between the small number of protein-coding genes and the large number of detected transcripts. Massively parallel hybridisation on already known sequence probes, such as in microarray technologies, are restricted to the probes chosen at the design stage and cannot thus explore the whole transcriptome. For this purpose, new generations of high density arrays have been developed using probes that span a genome region at regular intervals, either overlapping or spaced at defined distances [1,3].

Besides this new open strategies, methods based on sequence signatures (tags) such as Serial Analysis of Gene Expression (SAGE) [11] also meet the requirements to provide fresh information on unknown transcripts. SAGE tags are extracted from the $3'$ most 4-nucleotide "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 [12]. Starting from this site, stretches of 14 or 21 nucleotides (respectively in conventional SAGE and in LongSAGE) are extracted using Bsmf1 or Mme1 as "tagging" enzymes [11,8]. 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, once limited by sequencing cost, can now analyse millions of tags in a single experiment with the advent of new DNA sequencers [5].

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, 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 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 first checked the rate of success of this strategy on a sample of well-annotated mRNAs, and starting from previously unmatched tags, we then 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 genomic transcribed regions.

2 Computational Method

The virtual SAGE analysis of UniGene cluster-representative sequences [13] was performed using the Preditag software (Skuld-tech) as described in [6]. 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 processed non canonical and antisense tags. We perform 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-pb 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 of the Boyer-Moore-Horspool algorithm [9]. 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. 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 coordinates of this G-C tag pair is then recorded together with the sequence comprised between the two tags, which we call a *Tag-Delimited Genome Sequence (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 largest tag sets (106, 748 and 619, 771 G and C-tags, respectively), the search on the human genome required < 2 hours on a PC. Supplemental data, inclusive detailed material and methods, is available at `http://www.lirmm.fr/~rivals/GENOMICS`.

3 Results

Overall structure of SAGE data. We assembled two sets of 270 and 15 libraries built respectively with NlaIII and Sau3A1 as anchoring enzymes, associating local data and a large number of publicly available human SAGE libraries (see Supplementary Table 1). This collection, assembling 13.7 million C-tags and 0.5 million G-tags, is called UniSAGE hereafter. The total number of distinct C-tags registered in UniSAGE is 619, 770, i.e., 59% of 4^{10} possible 10-nucleotide combinations, largely exceeding the numbers of UniGene clusters and of well-annotated mRNAs. The count distribution reveals that most tags not associated with an RNA are observed at low levels, while tags matching RefSeq sequences are the most abundant ones. Several works report evidence for a large discrepancy between the numbers of tags and that of known genes, and conclude that a large number of tags originate from authentic transcripts [4,7]. Among the set of tags expressed at low levels, distinguishing between biological and artefactual ones prompts for novel approaches. Here, we propose and evaluate a novel approach called Tandem SAGE, on twin macrophage libraries and on the whole UniSAGE.

Analysis of twin SAGE libraries A : 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 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 separated exons. For 31 pairs, one of the tags was undetectable in the genome because it was created either by the junction of two exons, or by the polyA tail extension, or it contains a SNP. Similar observations are reported in [4]. Moreover, we found a unique chromosomal location for 69% of tag pairs (Table 1A and B).

Analysis of twin SAGE libraries, B: new transcripts detected by pairing unmatched tags. We collected experimental tags considered as unmatched tags and filtered out tags issued from genomic repeats. The selected 136 C- tags and 118 G-tags had at most 5 matches in the genome and were involved 187 tag pairs. 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%) mapping in genomic regions lacking

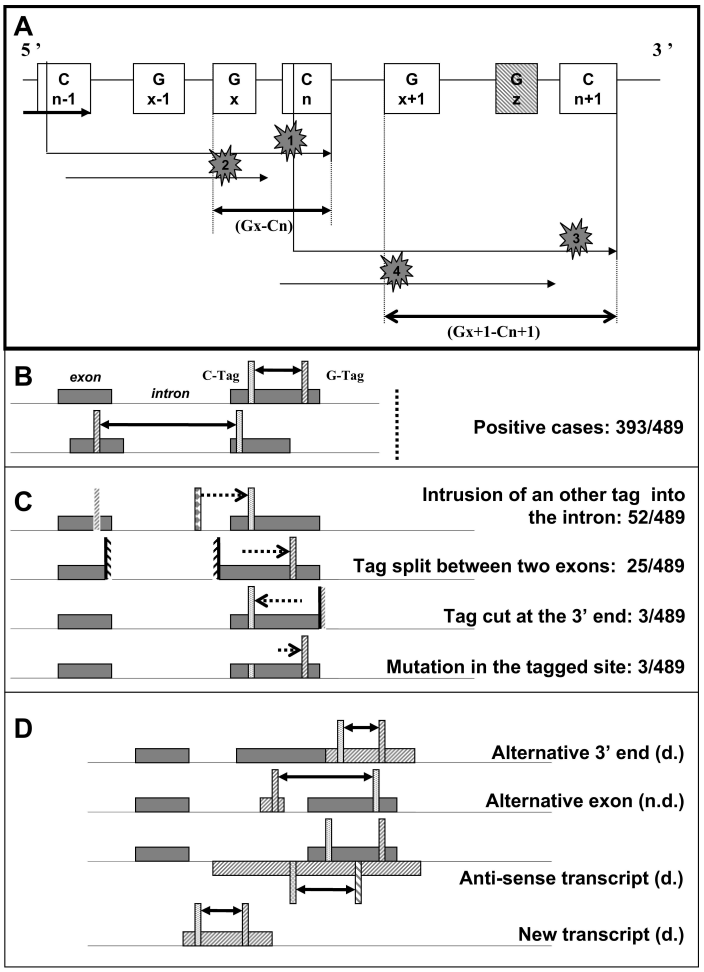

Figure 1. A: Illustration of the Tandem SAGE approach: the procedure for assembling 5′G- 3′C pairs. B,C,D: various causes of success and failure in assembling TDGS.

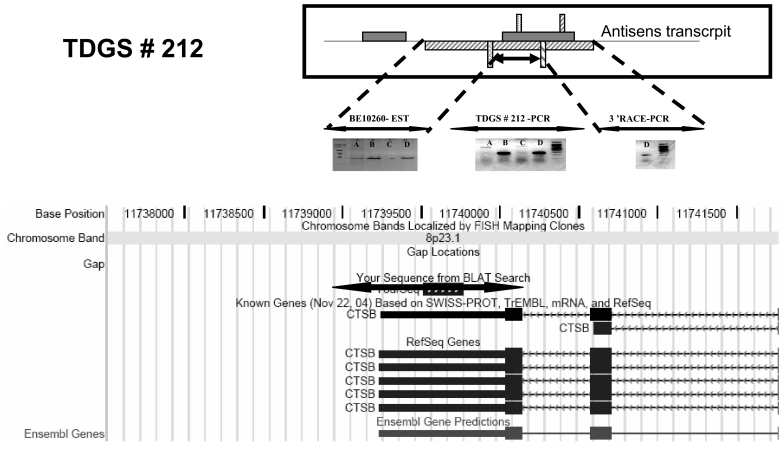

Figure 2. Annotation for class 3 transcript # 212 as visualised with the UCSC Genome browser. The sequence is in antisense orientation of Cathepsin B gene.

A	Total	Not found	Located		
Nb and % of tags pairs		%	Unique	Nb Locations ≥ 2	Sum
Tandem - Macrophage	489	96 20%	336 69%	57 12%	393 80%
Tandem - UniSAGE	11998	3110 26%	8189 68%	699 6%	8888 74%

B	Total	Found elsewhere	Tag intrusion in intron	Junction of exons	Poly-A tail	SNP in tag
Nb and % of tags	96	13 14%	52 54%	25 26%	3 3%	3 3%

Table 1. A: Capacity of TandemSAGE to identify to correct genomic TDGS for known transcripts, both for macrophage libraries and for UniSAGE. B: Tandem - Macrophage twin libraries, distribution of tag pairs among causes of localisation failure for known transcripts.

indications for transcription sites, or bearing an expressed sequence, but in inverted orientation (e.g., antisense transcript).

We successfully confirm by RT-PCR[5] the expression in the macrophage of 50% (14 out of 27) of a subset of tested cases (Supplementary Table 2). Examples for transcripts of classes 1 and 2 are illustrated in [7]. Class 3 contains antisense transcripts of known genes. For instance, TDGS # 212 maps to a region where the UCSC Browser indicated numerous ESTs mapping in both orientations. RT-PCR and RACE[6] extension of macrophage polyA+RNA confirmed the existence of a transcript in inverted orientation relatively to the Cathepsin B gene (CTSB, NM_0019082, Figure 2).

In silico experiment with the whole UniSAGE data. The algorithm was used to assemble the UniSAGE sets of C- and G-tags. We first tested its efficiency on $11,998$ well-annotated RNAs from Unigene with high quality R1 tag pairs (Table 1A). The algorithm found a chromosomal tag pair for $8,888$ genes (74%). Among them, $8,189$ (68%) were assigned to a unique genome site, which compares favorably to LongSAGE data [4]. Experimental tags considered as unmatched were collected according to the same criteria as described above for the twin libraries. In UniSAGE, $321,498$ C-tags and $49,103$ G-tags fall in this category. Working on the subset of tags found at least 3 times in the sum of all libraries, the algorithm assembled $93,859$ potential tag pairs on the genome. We cross-checked the set of TDGS with tiling arrays data (taken from the UCSC genome site). 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 less than 500 bp away from a TAR.

4 Discussion

In the present work, we developed a new algorithm associating pairs of gene expression signatures to localise 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-annoted genes and the large number of potential transcripts suggested by the number of tags. Using the whole set of presently available NlaIII and Sau3A1 SAGE tags, this efficient algorithm predicted $93,859$ potentially transcribed sites in the human genome. This observation corroborates independent evaluations based on tiling arrays [3]. Our algorithm assembles pairs of tags irrespective of their length or position ($5'$ or $3'$) on the transcript and could be used with other types signatures (paired-end ditags, MPSS, PMAGE [5]) For various reasons, SAGE methods fail to predict a unique genomic location for a transcript,

[5] RT-PCR: Reverse Transcriptase - Polymerase Chain Reaction
[6] RACE: rapid amplification of cDNA ends

which hinders genome annotation. Indeed, among available LongSAGE tags (21 bp) only 15% have a unique location in the genome, while 36% cannot even be found on the genome [4]. Tandem SAGE, which combines two 14 bp tags, indicates a unique location in 68% of the times.

Using a test sample of experimental tags obtained from twin macrophage libraries, we detected 187 potentially new transcripts: 39 appeared as alternative transcripts of known genes, while 134 were found in intergenic regions or in antisense orientation. We confirmed experimentally the existence of an RNA for 50% of our candidates. Such a success rate offers the possibility to identify thousands of new, biologically relevant transcribed regions at genome scale. To further select the best candidates among the large set of potential novel transcripts, we must use other criteria: for instance the length of TDGS [7] or evidence of their expression based on orthogonal and independent data sets. The ENCODE project plans to use tiling arrays as a major tool for genome annotation [10]. Here, we showed the possibility to connect efficiently both kinds of data, a difficult task with classical microarray and SAGE data. We found a 47% overlap between our TDGS collection and TARs. Hence, the Tandem SAGE strategy corroborates in part tiling arrays results and also reveals new transcripts having escaped from other detection systems. This emphasises the importance to combine independent and complementary methods for thoroughly exploring the transcribed part of the genome.

Acknowledgements This work was supported by Sidaction, the ANRS, the Ligue Régionale contre le Cancer Languedoc-Roussillon, and the Academy of Finland.

References

[1] P. Bertone, V. Stolc, T. Royce, J. Rozowsky, A. Urban, X. Zhu, J. Rinn, W. Tongprasit, M. Samanta, S. Weissman, M. Gerstein, and M. Snyder. Global identification of human transcribed sequences with genome tiling arrays. *Science*, 306(5705):2242–6, 2004.

[2] J. Claverie. Fewer genes, more noncoding rna. *Science*, 309:1529–30, 2005.

[3] P. Kapranov, J. Cheng, S. Dike, D. A. Nix, R. Duttagupta, A. T. Willingham, P. F. Stadler, J. Hertel, J. Hackermuller, I. L. Hofacker, I. Bell, E. Cheung, J. Drenkow, E. Dumais, S. Patel, G. Helt, M. Ganesh, S. Ghosh, A. Piccolboni, V. Sementchenko, H. Tammana, and T. R. Gingeras. RNA Maps Reveal New RNA Classes and a Possible Function for Pervasive Transcription. *Science*, 316(5830):1484–1488, 2007.

[4] C. Keime, M. Semon, D. Mouchiroud, L. Duret, and O. Gandrillon. Unexpected observations after mapping longsage tags to the human genome. *BMC Bioinformatics*, 8(1):154, 2007.

[5] J. Kim, G. Porreca, L. Song, S. Greenway, J. Gorham, G. Church, C. Seidman, and J. Seidman. Polony Multiplex Analysis of Gene Expression (PMAGE) in Mouse Hypertrophic Cardiomyopathy. *Science*, 316(5830):1481–1484, 2007.

[6] D. Piquemal, T. Commes, L. Manchon, M. Lejeune, C. Ferraz, D. Pugnère, J. Demaille, J.-M. Elalouf, and J. Marti. Transcriptome analysis of monocytic leukemia cell differentiation. *Genomics*, 80:361–71, 2002.

[7] E. Rivals, A. Boureux, M. Lejeune, F. Ottones, O. Pérez, J. Tarhio, F. Pierrat, F. Ruffle, T. Commes, and J. Marti. Transcriptome Annotation using Tandem SAGE Tags. *Nucleic. Acids Res.*, 35(17):e108, 2007.

[8] A. S. S Saha, C. Rago, V. Akmaev, C. Wang, B. Vogelstein, K. Kinzler, and V. Velculescu. Using the transcriptome to annotate the genome. *NatBio*, 20(5):508–12, 2002.

[9] J. Tarhio and H. Peltola. String matching in the dna alphabet. *Softw., Pract. Exper.*, 27(7):851–861, 1997.

[10] The ENCODE Project Consortium. Identification and analysis of functional elements in 1% of the human genome by the encode pilot project. *Nature*, 447:799–816, 2007.

[11] V. Velculescu, L. Zhang, B. Vogelstein, and K. Kinzler. Serial analysis of gene expression. *Science*, 270(5235):484–7, 1995.

[12] B. Virlon, L. Cheval, J. Buhler, E. Billon, A. Doucet, and J. Elalouf. Serial microanalysis of renal transcriptomes. *Proc. Natl. Acad. Sci.*, 96(26):15286–15291, 1999.

[13] D. Wheeler, T. Barrett, D. Benson, S. Bryant, K. Canese, V. Chetvernin, D. Church, M. DiCuccio, R. Edgar, and S. F. et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 34:D173–180, 2006.