Published online 20 September 2007 Nucleic Acids Research, 2007, Vol. 35, No. 18 e122
doi:10.1093/nar/gkm648

SAGETTARIUS: a program to reduce the number of tags mapped to multiple transcripts and to plan SAGE sequencing stages

Laurent Bianchetti[1,*], Yan Wu[1], Eric Guerin[2], Frédéric Plewniak[1] and Olivier Poch[3]

[1]Plate-forme Bioinformatique de Strasbourg, Institut de Génétique et de Biologie Moléculaire et Cellulaire (CNRS/INSERM/ULP) BP 163, 67404 Illkirch Cedex, [2]Inserm U682, Strasbourg, Laboratoire de Biochimie - Biologie Moléculaire, CHU Strasbourg - Hôpital de Hautepierre and [3]Laboratoire de Bioinformatique et de Génomique Intégratives, Institut de Génétique et de Biologie Moléculaire et Cellulaire (CNRS/INSERM/ULP) BP 163, 67404 Illkirch Cedex, France

Received May 18, 2007; Revised July 20, 2007; Accepted August 6, 2007

ABSTRACT

SAGE (Serial Analysis of Gene Expression) experiments generate short nucleotide sequences called 'tags' which are assumed to map unambiguously to their original transcripts (1 tag to 1 transcript mapping). Nevertheless, many tags are generated that do not map to any transcript or map to multiple transcripts. Current bioinformatics resources, such as SAGEmap and TAGmapper, have focused on reducing the number of unmapped tags. Here, we describe SAGETTARIUS, a new high-throughput program that performs successive precise Nla3 and Sau3A tag to transcript mapping, based on specifically designed Virtual Tag (VT) libraries. First, SAGETTARIUS decreases the number of tags mapped to multiple transcripts. Among the various mapping resources compared, SAGETTARIUS performed the best in this respect by decreasing up to 11% the number of multiply mapped tags. Second, SAGETTARIUS allows the establishment of a guideline for SAGE experiment sequencing efforts through efficient mapping of the CRT (Cytoplasmic Ribosomal protein Transcripts)-specific tags. Using all publicly available human and mouse Nla3 SAGE experiments, we show that sequencing 100 000 tags is sufficient to map almost all CRT-specific tags and that four sequencing stages can be identified when carrying out a human or mouse SAGE project. SAGETTARIUS is web interfaced and freely accessible to academic users.

INTRODUCTION

Genome-wide gene expression profiling is now possible, thanks to the development of complementary high-throughput techniques such as Expressed Sequence Tag (EST) (1), differential display (2), cDNA microarray (3), genome tiling array (4), SAGE (5) and the associated LongSAGE (6) adaptation. cDNA microarrays and genome tiling arrays are based on cDNA target hybridizations with complementary nucleic acid probes immobilized on a surface. The nucleotide sequence of the probes determines which transcripts can be hybridized, and thus which gene expressions can be measured. In contrast to cDNA hybridization strategies, EST, differential display, SAGE and LongSAGE all rely on the cloning and sequencing of cDNA. EST are randomly selected cDNA clones and they have applications in the characterization of gene products and new gene discovery (1), they may although not be effective enough to detect low-abundance transcripts (7). Differential display is a comparative approach which focuses on identifying differentially expressed genes between two cell samples. Finally, SAGE and LongSAGE can provide a measure of all individual gene expressions, including novel genes. Both methods require strong sequencing efforts, but the wealth of gene expression information obtained justifies the investment (8). SAGE has been successfully used to characterize the transcriptomes of yeast (9) and higher eukaryote cell types, both in healthy (10,11) and pathological situations, especially cancer (12,13). Moreover, LongSAGE should significantly improve genome annotation (14,15). Briefly, a SAGE experiment involves the isolation and reverse transcription of all 3′ polyadenylated mRNA expressed in a cell and the cloning,

*To whom correspondence should be addressed. Tel: +33 388653271; Fax: +33 388653201; Email: Laurent.Bianchetti@igbmc.u-strasbg.fr concatenation, sequencing and counting of short stretches of 10 nt called Experimental Tags (ET). The ET are derived from a particular position of the mRNA determined by a restriction enzyme site (anchoring enzyme) (5). The LongSAGE adaptation of SAGE generates longer ET sequences of 17 nt. The observed ET sequence copy frequencies provide the relative expression levels of the transcripts in the cells. A tag is defined by the 10 or 17 nt sequence which is directly downstream of the most 3′ anchoring enzyme restriction site on a 3′ poly-adenylated mRNA. Nla3, which recognizes the CATG motif, is the most widely used anchoring enzyme, but Sau3A, which cleaves DNA at GATC sites, has also been used in SAGE protocols. Once sequenced and counted, the thousands of ET generated by SAGE or LongSAGE experiments must be mapped to their original transcripts in order to identify the genes expressed in the cells.

ET to gene mapping is a bioinformatics task which requires the pre-construction of a specialized database of Virtual Tags (VT). VT are tags which are extracted from transcript sequences recorded in nucleic acid databases, and the VT sequences are associated with the transcripts or genes from which they derive. During the mapping process, ET are compared to VT. When an ET matches a VT, the ET is mapped to the transcript or gene with which the VT is associated, and as a result the identity of the gene which is expressed in the cell is determined. SAGE postulates that ET sequences are random and long enough to be mapped without ambiguity to their original gene. Theoretically, 10 nt can generate 4^{10} (1 048 576) different sequences. For human and mouse genomes, this repertoire of short sequences is largely above the estimated number of 30 000–40 000 different protein coding genes (16,17). Moreover, even the repertoire of human and mouse transcripts which has been estimated at 92 000 distinct mRNA (18) is less than the 4^{10} possibilities of SAGE tag sequences. Consequently, the mapping process is expected to map without ambiguity any ET to a single gene. However, in practice, the mapping process has come up against two major obstacles, namely unmapped and multiply mapped ET. Indeed, many ET do not map to any gene (19–21). Unmapped ET may account for the expression of novel transcripts or they could be the result of an accumulation of sequencing errors (22). Most bioinformatics developments have focused on decreasing the rate of unmapped ET and recently, the TAGmapper tool (21) has been shown to significantly reduce the rate of unmapped ET by means of a 7.1 million specialized VT database. A non-negligible rate of ET has also been reported to map to multiple genes (23,24). Indeed, it has been estimated that a third of human ET are shared by different genes (22), due to the fact that 10 nt long ET sequences might be too short to be specific to a single gene. Therefore, the LongSAGE technique which generates 17 nt long ET may prove to be more useful for resolving multiply mapped ET ambiguities (22).

The reasons why ET may map to multiple genes are debatable. First, ET are sequenced only once and are thus susceptible to sequencing errors. VT are also susceptible to sequencing errors because they derive from transcript records which are of variable primary sequence quality. Due to these sequence errors, false positive matches between ET and VT may occur. Second, transcripts may contain Interspersed Repetitive Elements (IRE) (25,26) which introduce common subsequences. SAGE tags extracted from IRE sequences are likely to lead to multiply mapped ET. Third, multiple gene mapping may also be a consequence of the VT database construction procedure itself. For example, SAGEmap and SAGE Genie mapping resources (27) both propose a VT database which contains internal tags, i.e. VT extracted from 5′ anchoring enzyme sites on an mRNA sequence and TAGmapper also integrates VT extracted from transcripts lacking a 3′ poly-adenylated boundary. These approaches may lead to multiple gene mapping by generating additional matches with VT sequences. Finally, the low complexity of some tag sequences may also lead to multiple gene mapping. For example, the AAAAAAAAAA VT is shared by many different transcript sequences whose most 3′ anchoring enzyme site occurs just before the 3′ poly-adenylated boundary. As a consequence, an AAAAAAAAAA ET will map every transcript associated with an AAAAAAAAAA VT.

The use of Sau3A as an alternative anchoring enzyme to carry out a SAGE experiment may help to resolve tag to transcript mapping ambiguities encountered when using Nla3. As an example, the human Ribosomal Protein 4 (RPL4) transcript (NM_000968) is associated with an Nla3 AAAAAAAAAA tag, whereas its Sau3A tag is CATCGCAGAG. Although Sau3A SAGE experiments are less frequent, the maintenance of an Sau3A mapping resource should be a useful tool for some tag to multiple transcript associations.

Human and mouse SAGE experiments, which represent 86% of all publicly available SAGE libraries, have generated several millions of ET sequences (22), and as a consequence automatic procedures are now required to map ET to their original genes. Here, we present a new mapping strategy and a new VT database both integrated in a web-interfaced tool, SAGETTARIUS, specifically designed to address the problem of ET to multiple gene mapping. In fact, SAGETTARIUS performs ET to transcript rather than gene mapping. SAGETTARIUS distinguishes three groups of VT to transcript associations, according to mRNA primary sequence quality: individually cloned and verified cDNA (28); High-Throughput cDNA (HTC) (29) which are full-length mRNA, but may be of draft quality (30,31), and Expressed Sequence Tags (EST) which are fragmentary sequences containing a 1% base error rate and thus a VT extracted from an EST has a 10% chance of being false (23). SAGETTARIUS implements a progressive and reductive mapping procedure, during which ET are compared to the three groups of VT to transcript associations.

Furthermore, biologists who plan a SAGE project must decide how many ET to sequence. The number will be a compromise between the gene expression information benefit required and the associated sequencing effort. Indeed, the number of ET required to define a cell transcriptome depends on the confidence level desired for detecting low-abundance mRNA molecules (9).

Furthermore, it has been estimated that a cell generally contains about 300 000–570 000 mRNA molecules (32) which means that at least 300 000 ET should be sequenced in order to cover the smallest transcriptome. Publicly deposited SAGE collections of ET range from a few thousand to more than 300 000 sequenced ET. Currently, most SAGE experiments collect from 50 000 to 100 000 ET sequences (22), but no guidelines exist for the efficient planning of ET sequencing stages. The mapping of ET derived from a reference family of ubiquitously expressed mRNA, ideally containing high-, medium- and low-abundance transcripts (33) would provide a practical framework to plan a SAGE project with successive sequencing stages. In this study, we have investigated whether the CRT can be used as a reference family of mRNA to propose a guideline for SAGE sequencing stages.

MATERIALS AND METHODS

Publicly available SAGE experiments

We downloaded all the publicly available *Homo sapiens* and *Mus musculus* Nla3 SAGE experiments from the Gene Expression Omnibus server (34,35). Thus, 371 human and 123 mouse SAGE experiments were collected. These SAGE experiments originate from various cell types, developmental stages, physiological conditions and pathological situations. The smallest (GSM718, SAGE_HMEC-B41) and largest (GSM14799, SAGE_Brain_fetal_normal_B_S1) human Nla3 SAGE experiments contain respectively, 1430 and 308 589 sequenced ET, whereas the smallest (GSM5050, P10 cerebellum) and largest (GSM75582, SAGE_hypothalamus_adrenalectomized) mouse Nla3 SAGE experiments contain respectively, 464 and 194 345 sequenced ET. The efficiency of the SAGETTARIUS ET to transcript mapping procedure has been evaluated on the GSM14740 experiment (SAGE_Brain_ependymoma_B_R1023 with 122 690 sequenced ET and 40 027 unique ET sequences). A sample of 8 human Nla3 SAGE experiments has been used to evaluate the number of IRE-derived ET mapped to multiple transcripts: GSM23394 (THP-1, cultured THP-1 cells; 2147 unique ET), GSM764 (SAGE_normal_-prostate; 6719 unique ET), GSM14804 (SAGE_Lung_normal_CL_L15; 9078 unique ET), GSM14805 (SAGE_Lung_normal_CL_L16; 11 894 unique ET), GSM762 (SAGE_normal_lung; 24 962 unique ET), GSM14740 (SAGE_Brain_ependymoma_B_R1023; 40 027 unique ET), GSM41378 (SAGE_Embryonic_stem_cell_H9_normal p38_CL_SHES1; 37 097 unique ET) and GSM14799 (SAGE_Brain_fetal_normal_B_S1; 80 125 unique ET).

SAGEmap resources of *H. sapiens* Nla3 VT to Unigene cluster associations

We downloaded from the SAGEmap (NCBI) bioinformatics mapping server (ftp://ftp.ncbi.nlm.nih.gov/pub/sage/map/Hs/Nla3) both 'full' and 'reliable' resources of human Nla3 VT to Unigene cluster associations, built from Genbank version 151 and *H. sapiens* Unigene release 187. These resources propose VT associated with Unigene clusters of transcripts according to the SAGEmap procedure (23). Briefly, the 'full' resource contains significantly more VT to Unigene cluster associations than the 'reliable' one, although the 'reliable' associations are more robust.

Computational servers, operating systems and source code

A Sun Solaris 9 server with 4× ultra-sparc-3 64 bits processors, 800 MHz and 16 GB of RAM hosts the SAGETTARIUS database of VT to transcript associations. The database information is currently stored in a Relational Sybase DBMS (Database Management System). Updates of the SAGETTARIUS database are performed on a 6 × 4 Sun AMD Opteron processors (2.6 GHz) using the Linux Operating System. The SAGETTARIUS database is queried using Perl 5.6.1 scripts integrating the DBI (Database Interface Module). The SAGETTARIUS web interface has been developed in PHP and Perl cgi (common gateway interface) scripts run on a Sun Enterprise 450 (Solaris 9) server with four processors, 1 GB of shared memory and the Unix operating system.

Construction of the VT to transcript association database

VT to transcript associations stored in the SAGETTARIUS database all originate from Genbank (30) (release 151) transcript sequences, except for the VT to CRT associations which originate from the transcript sequences of the RefSeq-RNA database (36). Locally updated copies of Genbank and RefSeq-RNA are maintained on our bioinformatics platform and are both indexed with SRS 7.1.3.1 (Sequence Retrieval Software, Lion Bioscience). The construction of the SAGETTARIUS database of VT to transcript associations relies on a three-step process (Figure 1).

(i) The selection and sorting step exhaustively identifies in Genbank and RefSeq-RNA, the human and mouse 3′ poly-adenylated (at least six adenines) transcripts and classifies the transcripts according to four groups: CRT, verified cDNA, HTC and EST. EST sequences that are 5′ poly-thymidylated instead of being 3′ poly-adenylated are reversed and complemented. We have also obtained from RefSeq-RNA the complete catalog of human and mouse full-length 3′ poly-adenylated CRT and their annotated splicing variants. The Ribosomal Protein Gene (RPG) database (37) provides the complete list of the 80 human and 79 mouse cytoplasmic ribosomal protein coding genes. Using gene name-based queries in the RefSeq-RNA database, we obtained the CRT sequence records. When a CRT sequence record contained a 3′ poly-adenylated boundary and anchoring enzyme sites, the most 3′ VT was extracted according to the SAGE tag definition. When a CRT sequence record did not contain a 3′ poly-adenylated boundary, the VT could not be extracted. Therefore, a blastn search was performed in the HTC-division of Genbank to find a redundant sequence record displaying an additional 3′ poly-adenylated boundary. The best scoring redundant HTC (identity percent >98 and identity number >400)

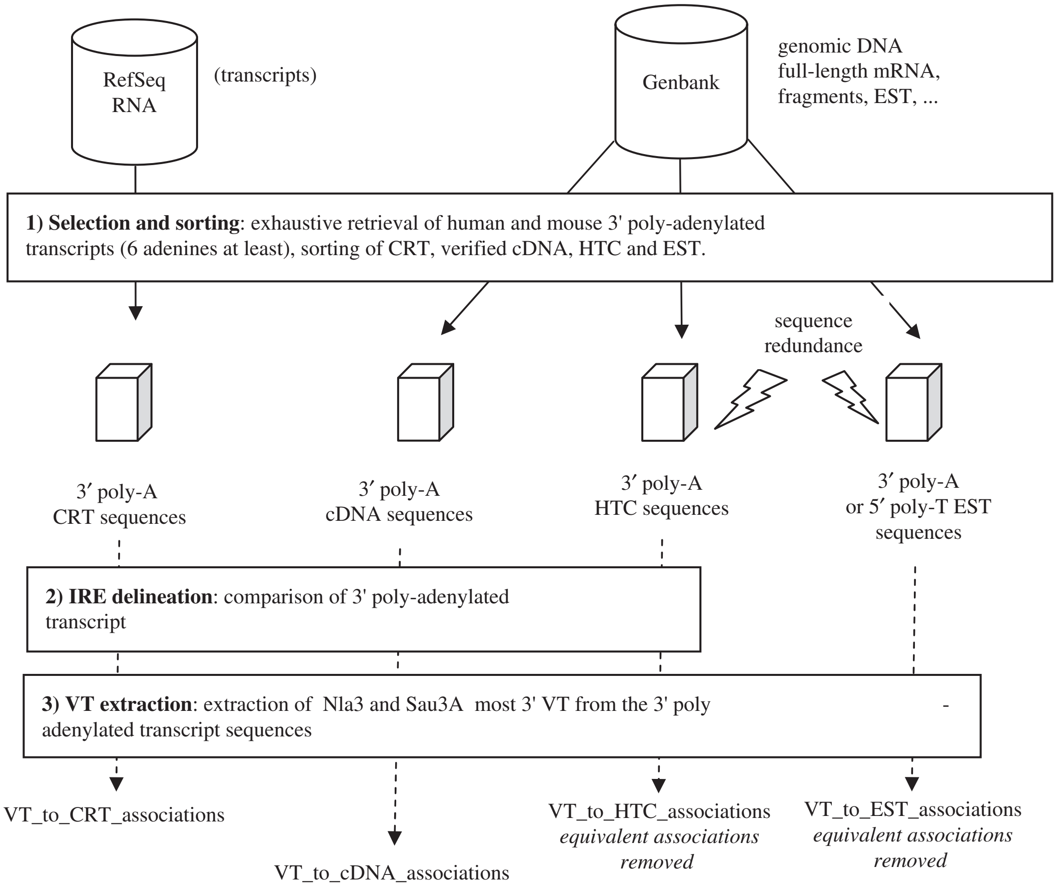

Figure 1. Generation of SAGETTARIUS database information. The final results of the procedure are VT to transcript associations. Bold: procedure steps.

was selected to complement the RefSeq-RNA CRT which lacked the 3′ poly-adenylated boundary.

(ii) Next, the IRE delineation step compares the human and mouse 3′ poly-adenylated transcript sequences to the Genetic Information Research Institute (GIRI) collections of IRE sequences (38) using RepeatMasker V3.1.3 (Smit,A., unpublished data) and Crossmatch (Green,P., unpublished data) and identifies IRE sequences integrated into transcripts and VT derived from IRE sequences.

(iii) Finally in the VT extraction step, VT of both 10 and 17 nt lengths are extracted from transcript sequences to produce VT to transcript associations. Sequence redundancy is especially frequent among HTC and EST sequences and leads to the generation of equivalent VT to transcript associations, i.e. associations displaying the same VT linked to the same transcript description. For HTC sequences, all but one of the equivalent associations are eliminated. SAGETTARIUS uses Unigene cluster titles and EST accession numbers recorded in Unigene. Thus, all EST gathered in the same Unigene cluster and generating the same VT are eliminated except for one representative sequence. However, the accession numbers of the eliminated EST are stored in the SAGETTARIUS database for use in downstream sequence analyses. Finally, VT to transcript associations are stored in 32 specialized libraries (see Supplementary Data at http://bips.u-strasbg.fr/Sage_docs/Supp_Material.html), according to organism (human, mouse), anchoring enzyme (Nla3, Sau3A), tag length (10 or 17 nt) and transcript primary sequence group (CRT, cDNA, HTC, EST).

Calculation of individual CRT detection probability for 10 000, 50 000 and 100 000 sequenced ET

Among all the human and mouse Nla3 SAGE experiments available, we have collected the SAGE experiments with 10 000, 50 000 and 100 000 sequenced ET (±10%). For human, the probabilities of CRT detections were calculated on the available 9, 22 and 8 experiments with 10 000, 50 000 and 100 000 sequenced ET (±10%), respectively. For mouse, the probabilities of CRT detections were calculated on the available 5, 13 and 2 experiments with 10 000, 50 000 and 100 000 sequenced ET (±10%), respectively. For each collected experiment, we have determined whether the 80 human or 79 mouse CRT have been individually detected. The probability of detecting a CRT at 10 000, 50 000 or 100 000 sequenced ET (±10%) is defined as the number of times the CRT has been detected divided by the number of SAGE experiments.

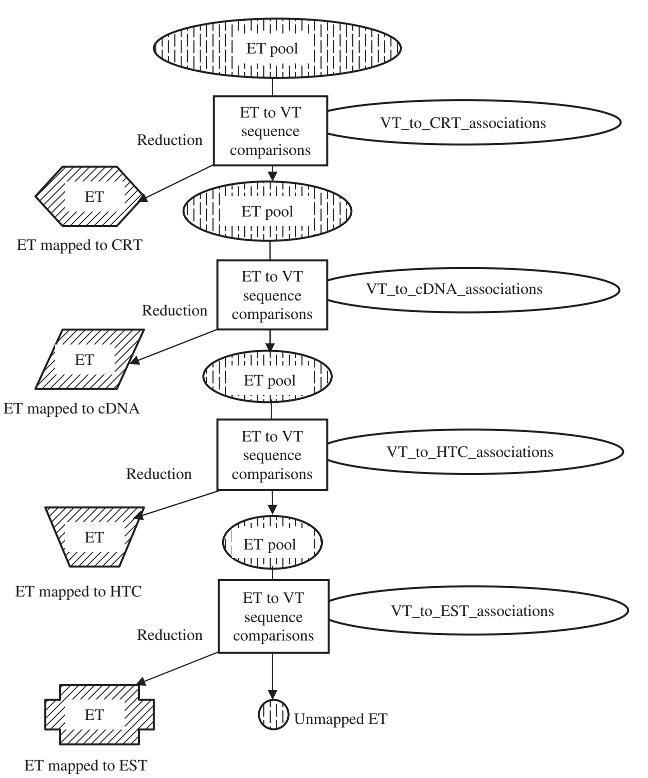

Figure 2. SAGETTARIUS progressive and reductive ET to transcript mapping process.

Progressive and reductive ET to transcript mapping process

SAGETTARIUS ET to transcript mapping is a progressive and reductive process. Unique ET sequences generated by a SAGE experiment are progressively compared to the VT sequences stored in the libraries of VT_to_CRT_associations, VT_to_cDNA_associations, VT_to_HTC_associations and VT_to_EST_associations, according to a four-step process (Figure 2). The process is reductive because unique ET sequences that match VT sequences are removed from the ET pool of the SAGE experiment. At the beginning of the process, unique ET sequences are compared to the VT sequences of the VT_to_CRT_associations library (first step). Unique ET sequences which obtain a match in this library are removed from the pool of ET. Then, the remaining unique ET sequences are compared to the VT sequences of the VT_to_cDNA_associations library (second step). Unique ET sequences which match the VT of this library are removed from the ET pool. Similarly, comparisons of the remaining unique ET sequences and VT sequences are carried out using the VT_to_HTC_associations and VT_to_EST_associations, successively. At the end of the process, unique ET sequences which cannot be matched to any VT sequence in the four libraries remain unmapped.

Web interface

SAGETTARIUS is hosted by the Bioinformatics Platform of Strasbourg (BIPS) and available through a user-friendly web interface at http://bips.u-strasbg.fr/Sage_docs/Sagettarius.php. SAGETTARIUS is interfaced using three main windows. In the submission window, the user provides SAGE experiment parameters (organism, anchoring enzyme, SAGE or LongSAGE protocol) and uploads a list of unique ET sequences. An e-mail address is not required to obtain the mapping results. ET to transcript mapping is displayed in a second window. The assessment of detected CRT is displayed in a third window. Results are stored for 1 month after generation and confidentiality is assured using a user-specific access code (job-ID).

RESULTS

VT to CRT associations

The ribosome, which acts as a catalyst for protein synthesis, is universal and essential for all organisms (39) and the expression of genes encoding cytoplasmic ribosomal proteins is expected to be ubiquitous in every cell. The ribosomal proteins are coded by 80 genes in human and 79 in mouse. We have constructed curated libraries of VT to CRT associations for both organisms. Most 3′ poly-adenylated mRNA coding for human ribosomal proteins are (82%) available in the RefSeq-RNA database. The remaining 18% are present in the HTC division of Genbank, although the primary sequences may be of lower quality. For the mouse organism, the transcripts coding the 79 cytoplasmic ribosomal proteins are all available in the RefSeq-RNA database, but only 28% of them display the mandatory 3′ poly-adenylated boundary required to derive VT. We therefore complemented the RefSeq-RNA mouse CRT lacking 3′ poly-adenylated boundaries by redundant 3′ HTC sequences showing a 3′ poly-adenylation. For Nla3 and Sau3A anchoring enzymes, we identified the specificities of CRT and their associated VT (Table 1) such as CRT displaying AAAAAAAAAA VT, Nla3 and/or Sau3A undetectable CRT lacking an anchoring enzyme site, distinguishable versus indistinguishable splicing transcript variants and VT derived from IRE subsequences integrated into CRT. Overall, for both human and mouse organisms, Nla3 and Sau3A SAGE protocols respectively detect 76 and 74 CRT which are the products of different ribosomal genes. The human L4 and L13 CRT are associated with the AAAAAAAAAA VT and therefore their expressions are not measurable. Interestingly, for both organisms, the S21 CRT does not contain any Nla3 or Sau3A restriction site, making the expression of the transcript completely undetectable by the current SAGE experimental protocols. Nla3 detects slightly more CRT than Sau3A, since two human CRT are undetectable by Nla3, whereas six are undetectable by Sau3A. For the mouse organism, 3 CRT cannot be detected by Nla3 and 5 by Sau3A. The transcript splicing variants 1 and 2 coding for S29 protein isoforms display different VT and their expression can thus be distinguished. However, most CRT splicing variants cannot be distinguished by SAGE because they have identical VT. Finally, the human L32 CRT is the only sequence found to contain an IRE which overlaps the most 3′ VT. This VT is identical to the ERO1-like transcript (AF081886) VT. The human and

Table 1. Characteristics of Nla3 and Sau3 VT sequences associated with human and mouse CRT. t.v.: splicing Transcript Variant. Asterisk symbol: no CRT presents this characteristic.

	Homo sapiens		*Mus musculus*	
Total number of genes encoding ribosomal proteins	80		79	
Anchoring enzyme	Nla3	Sau3A	Nla3	Sau3A
Detectable CRT (splicing variants excluded)	76	74	76	74
AAAAAAAAAA tag	L4, L13	*	*	*
Undetectable CRT	S21, L7A	S10, S12, S21, S25 L35A, L36	S21, L6, L35A	S21, S25, L31 L37, L37A
Distinguishable CRT splicing variants	S29 t.v. 1 & 2	S24 t.v. 1 & 2 S29 t.v. 1 & 2	*	*
Indistinguishable CRT splicing variants (Variants 1 and 2)	S15A, S24, L3, L6, L8, L9, L14, L17, L34, L38,UBA52, L41, P0	S15A, L3, L6, L8, L9, L14, L17, L34, L38, UBA52, L41, P0, SA	*	*
(Variants 1, 2 and 3)	L32	S14, L32		
VT derives from an IRE integrated in the CRT	*	L32 displays the same VT as the ERO1-like mRNA (F081886)	*	*

mouse catalogs of cytoplasmic ribosomal SAGE VT can be viewed on the web at http://bips.u-strasbg.fr/Sage_bin/RiboTable.cgi?Hs and Mm, respectively.

VT to cDNA, HTC and EST associations

We have exhaustively determined how many VT to transcript associations can be generated from the transcript sequences recorded in Genbank (Table 2). The SAGETTARIUS automatic procedure which generates VT to transcript associations counts the total number of human and mouse 3′ poly-adenylated cDNA, HTC and EST sequences recorded in Genbank. In Genbank release 156, the total number of transcript sequence records reaches 6.9 and 4.2 million for human and mouse, respectively. Most of these transcripts are EST (6.7 million for human and 4 million for mouse). Individually cloned and verified cDNA are the least represented group of transcripts. They account for 31 331 human and 15 153 mouse sequences. Unexpectedly, more than twice as many HTC have been sequenced for mouse (169 332) than for human (75 275). The 3′ poly-adenylation of a transcript sequence is mandatory to derive a robust VT to transcript association, and therefore we have determined how many Genbank transcripts contain a 3′ poly-adenylation of at least six adenines. A total of 6206 sequences out of 31 331 human verified cDNA show a 3′ poly-adenylated boundary (2955 out of 15 153 for mouse). 3′ poly-adenylated HTC represent 18 099 human and 15 703 mouse sequence records. Furthermore, 3′ poly-adenylated HTC are a subset of all HTC. Finally, 3′ poly-adenylated EST (693 858 human and 196 048 mouse sequences) also represent a subset of the total EST sequence records. Overall, 718 161 human and 214 706 mouse 3′ poly-adenylated transcript records were used to derive VT to transcript associations. Nla3 and Sau3A VT to transcript associations have been exhaustively derived from the

Table 2. VT to cDNA, HTC and EST associations in the SAGETTARIUS database. This assessment has been established on the database of VT to transcript associations built from Genbank release 156. *H.s.*: *Homo sapiens*, *M.m.*: *Mus musculus*. tr. seq.: transcript sequence. n.d.: not determined.

	cDNA		HTC		EST	
	H.s.	*M.m.*	*H.s.*	*M.m.*	*H.s.*	*M.m.*
Genbank tr. seq. records	31 331	15 153	75 275	169 332	6 771 069	4 059 938
Genbank 3′ poly-A tr. seq. records	6206	2955	18 099	15 703	693 858	196 048
VT to tr. seq. associations (Nla3)	6155	2936	13 537	12 541	135 793	60 535
VT to tr. seq. associations (Sau3A)	6025	2898	13 099	12 190	105 297	43 985
tr. seq. without Nla3 site	51	19	101	86	n.d.	n.d.
tr. seq. without Sau3A site	181	57	485	325	n.d.	n.d
tr. seq. without Nla3/Sau3 site	4	1	15	10	n.d.	n.d

pool of Genbank 3′ poly-adenylated transcripts. For HTC and EST, equivalent VT to transcript associations were removed. We have observed that Nla3 provides slightly more VT to transcript associations than Sau3A, which means practically that Nla3 can map more ET than Sau3A. Total 6155 and 2936 Nla3 VT to transcript associations were derived from individually cloned and verified cDNA for human and mouse, respectively. 3′ poly-adenylated HTC generated 13 537 human and 12 541 mouse Nla3 non-equivalent VT to transcript associations. In addition, 135 793 human and 60 535 mouse Nla3 non-equivalent VT to transcript associations were created from 3′ poly-adenylated EST. The SAGETTARIUS database (Genbank release 156) includes a total of 155 471 human Nla3 VT to transcript associations and 76 012 for mouse. These VT to transcript associations exceed the estimated repertoire of genes and transcripts coded by the human and mouse genomes. Interestingly, we noticed that some 3′ poly-adenylated transcripts do not generate any VT to transcript associations due to a lack of anchoring enzyme sites on their sequences. Total 41 human-verified cDNA lack the Nla3 site (16 for mouse) and four times as many transcripts were found to lack the Sau3A site. Finally, four human transcripts, supported by verified cDNA records are undetectable by both Nla3 and Sau3A anchoring enzymes: Rad51C truncated protein (AF029670); apoptosis-related protein PNAS-1 (AF229831); PNAS-117 (AF275813) and as previously mentioned the ribosomal protein S21 (NM_001024) encoding mRNA. For mouse, only the S21 mRNA cannot be detected. An exhaustive list of transcripts which escape the expression measure by Nla3, Sau3A or both enzyme SAGE protocols is accessible on the SAGETTARIUS web site. This list will be updated every 2 months in conjunction with Genbank releases.

Unique VT sequences associated with a single versus multiple transcripts

In the SAGETTARIUS database of VT to transcript associations, we have determined the number of unique VT sequences which are associated with a single versus multiple transcripts (Table 3). Almost twice as many unique VT sequences are available for human (120 234) than for mouse (65 497). For human, 84% of unique VT sequences are associated with a single transcript, and 91% for mouse. Thus, unique VT sequences associated with multiple transcripts account for 16 and 9% VT respectively for the two organisms. We have also investigated whether the lengthening of VT from 10 to 17 nt performed by the LongSAGE technique can help to reduce the percentage of unique VT sequences associated with multiple transcripts. We observed that the number of unique VT sequences increases by a factor of 1.2 when long VT are extracted from 3′ poly-adenylated transcripts and that the percentage of unique VT sequences associated with multiple transcripts decreases to 6 and 4% for human and mouse, respectively. Thus, the LongSAGE protocol significantly decreases the percentage of unique VT sequences associated with multiple transcripts. This is consistent with the fact that 17 nt long tags are more specific than 10 nt. However, 8327 and 3019 unique VT sequences remain associated with multiple transcripts for human and mouse, respectively.

Table 3. Unique VT sequences associated with a single *vs* multiple transcripts in the SAGETTARIUS database (Nla3). *H.s*: *Homo sapiens*, *M.m: Mus musculus*. SAGE: 10 nt VT, LongSAGE: 17 nt VT. tr.: transcript

	VT to transcript (cDNA, HTC, EST) associations			
	H.s.		*M.m.*	
	SAGE	LongSAGE	SAGE	LongSAGE
unique VT seq.	120 234	149 853	65 497	72 800
unique VT seq. associated to a single tr.	101 117	141 526	59 440	69 781
unique VT seq. associated to multiple tr.	19 117	8327	6057	3019

SAGE sequencing stages based on the detection of CRT-specific tags

We have estimated how many ET should be sequenced when planning a human or mouse SAGE project in order to map the 80 human or 79 mouse CRT-specific tags. It is noteworthy that the expression of the complete repertoire of human and mouse CRT cannot be detected by Nla3 or Sau3A SAGE protocols, since several CRT sequences lack the anchoring enzyme sites or are associated with AAAAAAAAAA tags. However, the expression of 76 CRT should be detected by Nla3 (74 CRT for Sau3A) for both organisms. We downloaded the ET collections of all publicly available *H. sapiens* and *M. musculus* Nla3 SAGE experiments from the Gene Expression Omnibus server, with the number of sequenced ET ranging from 1430 to 308 589 and 464 to 194 345, respectively. The unique ET sequences collected in all experiments were automatically mapped to their transcripts by SAGETTARIUS and the detected CRT-specific tags were counted. We observed that the number of detected CRT-specific tags increases with the number of sequenced ET in both organisms (Figure 3). Here, we propose to use the detection of CRT-specific tags to define four major sequencing stages when planning a human or mouse SAGE experiment. The initial sequencing of 10 000 ET corresponds to a rapid increase in CRT-specific tag detection (stage 1). Total 10 000 ET allows the detection of 63 CRT with a probability greater than 0.66 (Table 4) in human and mouse. The sequencing of 40 000 additional ET (stage 2) substantially increases the number of detected CRT-specific tags which reaches 65 in human and 69 in mouse with a probability greater than 0.66. The sequencing of a further 50 000 ET (stage 3) allows the detection of 69 CRT in human and 70 in mouse. Finally, the sequencing of more than 100 000 ET corresponds to a plateau in CRT-specific tag detection (stage 4). We have analyzed the probabilities of individual CRT-specific tag detection in Nla3 human and mouse SAGE experiments with 10 000 ±10%, 50 000 ±10% and 100 00 ±10% sequenced ET (Table 4 and also Supplementary Data at

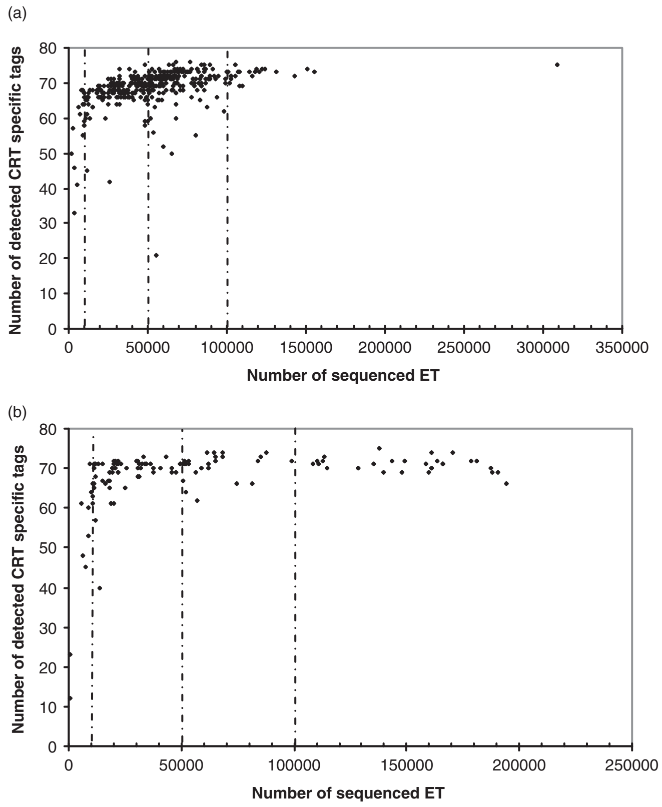

Figure 3. Progressive detection of CRT-specific tags in (**a**) 371 human Nla3 SAGE experiments with the number of sequenced ET ranging from 1430 to 308 589 and (**b**) 123 mouse Nla3 SAGE experiments with the number of sequenced ET ranging from 464 to 194 345. In both human and mouse, SAGE experiments can be divided into four major sequencing stages (- -) based on the detection of CRT-specific tags.

http://bips.u-strasbg.fr/Sage_docs/Supp_Material.html for details). We have derived three classes of CRT detection Probabilities (P): high ($P > 0.66$), medium ($0.66 \geq P > 0.33$) and low ($P \leq 0.33$). In human, a set of 63 CRT display a high detection probability at stages 1, 2 and 3 of ET sequencing. Outside this set, the more ET are sequenced, the greater is the probability to detect most of the remaining CRT (SA, S4Y, S23, S24, L14, L22, L32, L34, L23A and UBA52). The human L10, L28 and L37 CRT invariably remain in the low probability detection class at stages 1, 2 and 3 of ET sequencing. In mouse, a set of 62 CRT display a high detection probability at stages 1, 2 and 3 of ET sequencing. Moreover, 53 CRT of these CRT have human orthologs in the set of the 63 human CRT which display a high detection probability at stages 1, 2 and 3 of ET sequencing. The more ET are sequenced, the greater is the probability to detect most of the remaining mouse CRT (S15A, S8, S16, S17, L7, L14, L27A, L29 and L22). The mouse S25, L28, and L36 CRT invariably remain in the low probability detection class at stages 1, 2 and 3 of ET sequencing. Moreover, the mouse L10 CRT was not detected in any SAGE experiment. A number of discrepancies are observed, e.g. the human SA CRT requires at least the sequencing of 100 000 ET to be part of the medium detection probability class, whereas its mouse ortholog already belongs to the high detection probability class when 10 000 ET are sequenced. In addition, the human L37 CRT invariably belongs to the low detection probability class regardless of the sequencing stage, whereas the mouse ortholog is invariably part of the high detection probability class. Unexpectedly, the mouse L26 CRT is more likely to be detected when less ET are sequenced. Some of these discrepancies might be due to the SAGE experiment sampling.

Table 4. Probabilities (P) of individual CRT-specific tag detections correlated with the number of sequenced ET in human and mouse SAGE experiments. Group A contains 63 human CRT, namely S2, S3, S3A, S4X, S5, S6, S7, S8, S9, S10, S11, S12, S13, S14, S15, S15A, S16, S17, S18, S19, S20, S25, S26, S27, S27A, S28, S29, S30, L3, L5, L6, L7, L8, L9, L10A, L11, L12, L13A, L15, L17, L18, L18A, L19, L21, L23, L24, L26, L27, L27A, L29, L30, L31, L35, L35A, L36, L36A, L37A, L38, L39, L41, P0, P1 and P2. Group B contains 62 mouse CRT, namely SA, S2, S3, S3A, S4, S5, S6, S7, S9, S10, S11, S12, S13, S14, S15, S18, S19, S20, S23, S24, S26, S27, S27A, S28, S29, FBR-MuSV, L3, L4, L5, L7A, L8, L9, L10A, L11, L12, L13, L13A, L15, L17, L18, L18A, L19, L21, L23, L23A, L24, L27, L30, L31, L32, L34, L35, L36A, L37, L37A, L38, L39, UBA52, L41, P0, P1 and P2. Bold: invariably low detectable CRT

Sequenced ET	*Homo sapiens*			*Mus musculus*		
	High $P > 0.66$	**Medium** $0.66 > P > 0.33$	**Low** $P < 0.33$	**High** $P > 0.66$	**Medium** $0.66 > P > 0.33$	**Low** $P < 0.33$
10 000 ±10%	Group A	S24, L14	SA, S4Y, S23, **L10**, L22, L23A, **L28**, L32, L34, **L37**, UBA52	Group B, L26	S8, S16, S17, L7, L14, L27A, L29	S15A, S25, **L10**, L22, **L28, L36**
50 000 ±10%	Group A, S24, L14	S4Y, S23, L22, L23A	SA, **L10**, **L28**, L32, L34, **L37**, UBA52	Group B, S8, S16, S17, L7, L14, L27A, L29	S15A, L26,	S25, **L10**, L22, **L28**, **L36**
100,000 ±10%	Group A, S23, S24, L14, L22, L23A, UBA52	SA, S4Y, L32, L34	**L10, L28, L37**	Group B, S8, S16, S17, L7, L14, L22, L27A, L29	S15A	S25, **L10**, L26, **L28**, **L36**

Comparison of ET mappings performed by SAGETTARIUS and other tools

We mapped the 40 027 unique ET sequences generated by the GSM14740 (SAGE_Brain_ependymoma_B_R1023) SAGE experiment with SAGETTARIUS, TAGmapper, SAGEmap-reliable and SAGEmap-full. SAGETTARIUS maps ET sequences to transcripts, whereas the other tools (TAGmapper, SAGEmap-reliable and SAGEmap-full) map ET sequences to Unigene clusters. For each tool output, the number of ET sequences mapped to a single, multiple and no transcript or Unigene clusters respectively were counted (Figure 4). Important differences were observed between the outputs of the 4 ET mapping resources. SAGETTARIUS performed the best for the reduction of ETs to multiple transcript mappings: only 4357 unique ET sequences out of 40 027 were mapped to multiple transcripts. In contrast, SAGEmap-full resulted in a high rate (70%) of unique ET sequences mapped to multiple Unigene clusters. TAGmapper was the most successful at mapping ET sequences to single Unigene clusters, with three unique ET sequences out of four mapped by TAGmapper to a single Unigene cluster. TAGmapper and SAGEmap-full resulted in the smallest numbers of unmapped ET sequences, whereas SAGETTARIUS produced the most. Indeed, half of the ET sequences did not obtain any match to VT when using SAGETTARIUS. It is noteworthy that 86% of the GSM14740 unique ET that remained unmatched to any VT when using SAGETTARIUS corresponded to single copy ET sequences.

Comparison of ET to CRT mapping produced by different mapping resources

The 40 027 unique ET sequences generated by the GSM14740 (SAGE_Brain_ependymoma_B_R1023) experiment were mapped by SAGETTARIUS, TAGmapper, SAGEmap-reliable and SAGEmap-full mapping resources. In each tool output, we searched for ET sequences that were mapped to CRT or cytoplasmic ribosomal-specific Unigene clusters. In the SAGETTARIUS database, curated specialized libraries of VT to CRT associations were built from 3′ poly-adenylated CRT sequences and their annotated isoforms. Important differences in ET to CRT or Unigene cluster associations were observed between the four mapping tools. Of the 40 027 different ET sequences in the GSM14740 SAGE experiment, SAGETTARIUS mapped 72 unique ET sequences to CRT. These CRT are the products of 72 ribosomal protein coding genes. Both TAGmapper and SAGEmap-reliable tools mapped more than 660 ET sequences of the GSM14740 experiment to cytoplasmic ribosome-specific Unigene clusters and SAGEmap-full associated more than 2400 ET with these Unigene clusters. Our curated library of VT to CRT associations shows that the S21, L7A, L4 and L13 CRT cannot be detected by the Nla3 SAGE protocol. In contrast, TAGmapper proposed a mapping of ET sequences to these four cytoplasmic ribosome protein coding genes. Interestingly, the SAGEmap-reliable mapping resource did not map any ET sequence to S21, whose transcript product (NM_001024) lacks an Nla3 anchoring enzyme site. In addition, no ET sequence was mapped by SAGEmap-reliable and SAGEmap-full for the L4 protein coding gene, which is consistent with the SAGETTARIUS result.

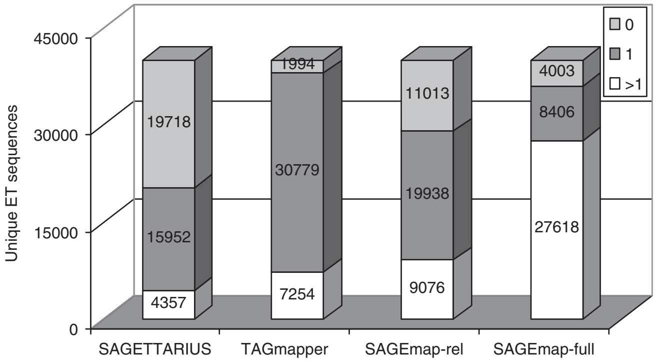

Figure 4. Number of unmapped, multiply mapped and single-mapped ET sequences from the GSM14740 SAGE experiment (40 027 unique ET sequences). ET mappings have been carried out by SAGETTARIUS, TAGmapper, SAGEmap-reliable and SAGEmap-full resources.

Unique ET sequences mapped to multiple transcripts due to IRE subsequences

In a sample of 8 SAGE experiments with the number of unique ET sequences ranging from 2147 (GSM23394) to 80 125 (GSM14799), we measured the rate of unique ET sequences which matched VT derived from IRE subsequences integrated into transcripts. The unique ET sequences of all eight SAGE experiments were mapped to transcripts by means of the SAGETTARIUS tool. For each experiment, we counted the number of unique ET sequences matching IRE-derived VT. In these examined SAGE experiments, the IRE-derived ET sequences/unique ET sequence ratio varies from 3×10^{-3} to 13×10^{-3}, depending on the mRNA expressed in a cell type. This number is however negligible, showing that the power of the SAGE protocol is not significantly weakened by the IRE-derived tag problem.

DISCUSSION

SAGE and LongSAGE are elegant and powerful molecular biology methods designed to measure genome-wide gene expression profiles. They rely on cloning and sequencing techniques carried out on a high-throughput scale. SAGE and LongSAGE have been applied to various organisms and cell types and have generated millions of ET, containing gene expression level information. However, once the ET of an experiment have been sequenced and counted, a major problem must be addressed, determining which genes are expressed using nucleotide sequences as short as 10 or 17 nt. Bioinformatics has taken up this challenge and two problems have been identified. First, ET sequences may map to several genes, and second, ET sequences may not map to any gene. The reasons why ET sequences map to multiple genes or do not map to any gene have been discussed recently (22). Briefly, ET to multiple gene mapping may be due to the sequence match between ET originating from different genes, while unmapped ET may result from new gene expression discovery. In both cases, sequencing errors, VT databases and the mapping procedures themselves can strongly influence the rate of multiply versus unmapped ET. We have developed a new program and VT database, SAGETTARIUS, both specifically designed to address the problem of ET to multiple gene mapping. When developing this resource, we also wished to provide the biologist with a practical means of planning ET sequencing stages during a SAGE project.Among the current mapping tools, SAGETTARIUS is the most successful in decreasing the rate of ET mapping to multiple transcripts. SAGETTARIUS progressively maps ET sequences to libraries of VT to transcript associations, which are hierarchized according to primary sequence reliability. Thanks to this progressive and reductive mapping procedure, SAGETTARIUS avoids ET to multiple transcript mappings which are likely to result from sequencing errors. The number of ET which map to multiple transcripts due to non-random IRE subsequences seems to be negligible. Indeed, it concerns only 3×10^{-3} to 13×10^{-3} unique ET sequences in a complete collection of SAGE tags. Moreover, the LongSAGE protocol efficiently reduces tag to multiple transcript associations.

SAGETTARIUS has been developed to reduce the number of VT associated with multiple transcripts. In order to reliably associate VT to transcripts, we decided to consider only mRNA sequences displaying a 3′ poly-adenylated boundary. Indeed, the absence of a 3′ poly-adenylated boundary might indicate an incomplete mRNA sequence. Consequently, the most 3′ anchoring enzyme site cannot be efficiently determined and no robust VT to transcript association can be derived. In mammals, cleavage of the mRNA on the 3′ boundary and its subsequent poly-adenylation to the newly formed 3′ end occurs 10–30 nt downstream of a specific nucleotide hexamer, generally the AAUAAA pattern (41,42). However, in a recent study carried out on 13 942 human and 11 155 mouse genes, it appeared that only 53.2% of the poly-adenylation signals correspond to the canonical AAUAAA hexamer (43). Moreover, current methods for poly-adenylation signal prediction achieve moderate sensitivity and specificity (44). Therefore, we have chosen not to use mRNA sequences lacking a 3′ poly-adenylated boundary to derive VT to transcript associations even in the case where a poly-adenylation signal could be predicted.

The construction of the SAGETTARIUS database of VT to transcript associations revealed that 3′ poly-adenylated transcript records are unexpectedly scarce in Genbank. Thousands of human verified cDNA are available in Genbank but most of them lack the crucial 3′ poly-adenylated boundary (80%). Multi-pass sequencing of transcripts including the 3′ poly-adenylated boundary are required to enrich VT databases with high-quality sequences that can be used to derive robust tag to transcript associations. The relative paucity of 3′ poly-adenylated transcripts in Genbank could be one of the reasons why SAGETTARIUS fails to map a significant percentage of ET sequences (49% of GSM14740 unique ET sequences). This is surprising since the SAGETTARIUS database contains more VT to transcript associations than the number of estimated genes and transcripts coded by the human or mouse genomes. The lack of 3′ poly-adenylated transcripts in nucleic acid databases may also have encouraged other mapping resources to overexceed the *sensu stricto* SAGE tag definition by integrating internal tags into VT databases, i.e. VT that are extracted from 5′ anchoring enzyme sites on 3′ poly-adenylated transcript or VT originating from transcripts lacking a 3′ poly-adenylated boundary, thus generating new VT possibilities and decreasing the rate of unmapped tags. Some of these internal tags might be the result of alternative 3′ poly-adenylation or mRNA alternative splicing but in the absence of any validation, these VT remain putative and their associations should be considered with caution. TAGmapper and SAGEmap resources map most ET but the risk of generating multiple mapping increases when the VT database includes internal tags. Another possibility that might explain the high rate of unmapped ET when using SAGETTARIUS could be that the unmapped ET represent novel transcripts whose sequences are not yet available in the nucleic acid transcript databases. Using SAGETTARIUS to map the 40 027 GSM14740 unique ET sequences to transcripts, we observed that 86% of the unmapped ET are single-copy sequences and thus probably correspond to low-abundant transcripts. Indeed, three abundance classes of mRNA have been previously defined: high-, medium- and low-abundance (32). The sequencing of EST (the major source of database transcript records) seems to have technically reached a plateau in new transcript discovery (20). Thus, high- and medium-abundant transcripts may have been identified and made available in databases, but low-abundant transcripts are less well characterized (7). In contrast, SAGE performs significantly better than the EST approach to detect low-abundant transcripts. Consistent with our results, it has been shown in *Drosophila melanogaster* that 55% of the unique ET sequences generated by a SAGE experiment could not be mapped to any known *Drosophila* transcripts. Furthermore most of these ET sequences displayed low-copy numbers (45). In addition, in the mouse, 66% of unique ET sequences generated by a LongSAGE experiment could not be mapped to any known gene (14). Finally, it is unlikely that all unmapped single-copy ET are the result of sequencing errors. Since an ET sequence has a 10% chance of being false, only 10% of all single copy ET are expected to be false and thus unmapped.

The mapping of ET sequences to CRT implemented in SAGETTARIUS allows us to propose four major ET sequencing stages during a human or mouse SAGE project. We found that 63 CRT-specific tags could be detected with a probability greater than 0.66 with a first run of 10 000 sequenced ET (stage 1). This first sequencing stage provides the best information benefit to sequencing effort ratio for CRT detection whereas the subsequent three stages require increased sequencing efforts in order to detect additional CRT. Due to the rapid increase of CRT-specific tag detection during stage 1, we recommend that biologists who are planning a SAGE project should sequence at least 10 000 ET. Moreover, 100 000 sequenced ET (stage 3) represents a major sequencing stage since it allows the detection of almost all CRT. Sequencing more than 100 000 ET provides the lowest information benefit to sequencing effort ratio for CRT detection. However, it may be crucial to sequence more than 100 000 ET in order to detect other transcripts. Discrepancies in the detection of expression have been observed between some human and mouse CRT (SA, L37 and L26). These results should thus be taken with caution. Indeed, we have observed that RefSeq-RNA entries for these specific CRT are derived from HTC sequences. Therefore, their VT to transcript associations must be verified. We are aware that our analysis does not provide an accurate quantitative differential CRT expression study between the SAGE experiments. However, in this study our aims basically were (i) to present a tool to map human and mouse ET (ii) to verify whether the available CRT sequence records were of sufficient quality (3′ sequence poly-adenylation, verified cDNA sequences) to allow an accurate mapping of the 80 human and 79 mouse CRT-specific tags and (iii) to determine the minimal number of ET required to detect all the CRT-specific tags for both organisms. For human, the quality of SAGETTARIUS VT to CRT associations is satisfying since 82% of the CRT records are supported by verified cDNA and display a 3′ poly-adenylated boundary. For the mouse, the library is less satisfying since only 20% of CRT are verified cDNA and the remaining are derived from HTC sequences of more variable quality. A comparative analysis of CRT expression levels between the different SAGE experiments is now conceivable for the human organism. In contrast, for the mouse, the VT to CRT association quality must first be improved. Recently, using micro-array data, it has been shown in 30 different human cell types, that most CRT are coordinately expressed with higher signals in some specific tissues, and that 17 are expressed in a tissue-specific manner (46). Our human VT to CRT association library represents a valuable resource to cross-validate the microarray CRT data with transcript expression information from SAGE experiments.

In conclusion, SAGETTARIUS is a new high-throughput ET to transcript mapper integrating its own database of VT to transcript associations. It performs high quality ET to transcript mapping for human and mouse organisms, Nla3 and Sau3A anchoring enzymes and SAGE and LongSAGE protocols. SAGETTARIUS has the advantage of significantly reducing the rate of multiply mapped tags. However, because of the high rate of unmapped ET, it is advisable to use SAGETTARIUS in combination with other tools.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online and also available at http://bips.u-strasbg.fr/Sage_docs/Supp_Material.html

ACKNOWLEDGEMENTS

We are grateful to Dr J.D. Thompson, Dr H. Koutnikova and Dr Joëlle Henry-Berger for helpful comments, Véronique Geoffroy, Serge Uge and Michel Offner for bioinformatics resource maintenance and Sandra Bour for technical help. We also thank Prof. Pierre Oudet, Natacha Entz-Werle and Marie-Pierre Gaub (UMR7104, CNRS/Inserm/ULP, IGBMC, Illkirch) for their interest in the project. This work was supported by institute funds from the Institut National de la Santé et de la Recherche Médicale (INSERM), the Centre National de la Recherche Scientifique (CNRS) and the Réseau National des Génopôles (RNG). Funding to pay the Open Access publication charges for this article was provided by the RNG.

Conflict of interest statement. None declared.

REFERENCES

1. Adams,M.D., Kelley,J.M., Gocayne,J.D., Dubnick,M., Polymeropoulos,M.H., Xiao,H., Merril,C.R., Wu,A., Olde,B. *et al.* (1991) Complementary DNA sequencing: expressed sequence tags and human genome project. *Science*, **252**, 1651–1656.
2. Liang,P. and Pardee,A.B. (1992) Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. *Science*, **257**, 967–971.
3. Schena,M., Shalon,D., Davis,R.W. and Brown,P.O. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, **270**, 467–470.
4. Bertone,P., Stolc,V., Royce,T.E., Rozowsky,J.S., Urban,A.E., Zhu,X.Z., Rinn,J.L., Tongprasit,W., Samanta,M. *et al.* (2004) Global identification of human transcribed sequences with genome tiling array. *Science*, **306**, 2242–2246.
5. Velculescu,V.E., Zhang,L., Vogelstein,B. and Kinzler,K.W. (1995) Serial analysis of gene expression. *Science*, **270**, 484–486.
6. Saha,S., Sparks,A.B., Rago,C., Akmaev,V., Wang,C.J., Vogelstein,B., Kinzler,K.W. and Veculescu,V.E. (2002) Using the transcriptome to annotate the genome. *Nat. Biotechnol.*, **19**, 508–512.
7. Sun,M., Zhou,G., Lee,S., Chen,J., Shi,R.Z. and Wang,S.M. (2004) SAGE is far more sensitive than EST for detecting low-abundance transcripts. *BMC genomics*, **5**, 1.
8. Pérez-Plasencia,C., Riggins,G., Vasquez-Ortiz,G., Moreno,J., Arreola,H., Hidalgo,A., Pina-Sanchez,P. and Salcedo,M. (2005) Characterization of the global profile of genes expressed in cervical pithelium by Serial Analysis of Gene Expression (SAGE). *BMC Genomics*, **6**, 130.
9. Velculescu,V.E., Zhang,L., Zhou,W., Vogelstein,J., Basrai,M.A., Bassett,D.E., Hieter,P., Vogelstein,B. and Kinzler,W.K. (1997) Characterization of the yeast transcriptome. *Cell*, **88**, 243–251.
10. 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.
11. Sharon,D., Blackshaw,S., Cepko,C.L. and Dryja,T.P. (2002) Profile of the genes expressed in the human peripheral retina,

macula, and retinal pigment epithelium determined through serial analysis of gene expression (SAGE). *Proc. Natl Acad. Sci. USA*, **99**, 315–320.
12. Riggins,G.J. and Strausberg,R.L. (2001) Genome and genetic resources from the Cancer Genome Anatomy Project. *Hum. Mol. Genet*., **10**, 663–667.
13. Zhang,L., Zhou,W., Velculescu,V.E., Kern,S.E., Hruban,R.H., Hamilton,R.S., Vogelstein,B. and Kinzler,K.W. (1997) Gene expression profiles in normal and cancer cells. *Science*, **276**, 1268–1272.
14. Wahl,M.B., Heinzmann,U. and Kenji,I. (2005) LongSAGE analysis significantly improves genome annotation: identifications of novel genes and alternative transcripts in the mouse. *Bioinformatics*, **12**, 1393–1400.
15. Wahl,M.B., Heinzmann,U. and Kenji,I. (2005) LongSAGE analysis revealed the presence of a large number of novel antisense genes in the mouse genome. *Bioinformatics*, **21**, 1389–1392.
16. Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C., Baldwin,J., Devon,K., Dewar,K., Doyle,M. *et al*. (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
17. Waterston,R.H., Lindblad-Toh,K., Birney,E., Rogers,J., Abril,J.F., Agarwal,P., Agarwala,R., Ainscough,R., Alexandersson,M. *et al*. (2002) Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520–562.
18. Fields,C., Adams,M.D., White,O. and Venter,J.C. (1994) How many genes in the human genome? *Nat. Genet*., **7**, 345–346.
19. Waghray,A., Schober,M., Feroze,F., Yao,F., Virgin,J. and Chen,Y.Q. (2001) Identification of differentially expressed genes by serial analysis of gene expression in human prostate cancer. *Cancer Res*., **61**, 4283–4286.
20. 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.
21. Bala,P., Georgantas,R.W., Suhdir,D., Suresh,M., Shanker,K., Vrushabendra,B.M., Civin,C.I. and Pandey,A. (2005) TAGmapper: a web-based tool for mapping SAGE tags. *Gene*, **364**, 123–129.
22. Wang,S.M. (2006) Understanding SAGE data. *Trends Genet*., **23**, 42–50.
23. Lash,A.E., Tolstoshev,C.M., Wagner,L., Schuler,G.D., Strausberg,R.L., Riggins,G.J. and Altschul,S.F. (2000) SAGEmap: a public gene expression resource. *Genome Res*., **10**, 1051–1060.
24. Stollberg,J., Urshitz,J., Urban,Z. and Boyd,C.D. (2006) A quantitative evaluation of SAGE. *Genome Res*., **10**, 1241–1248.
25. Yulug,I.G, Yulug,A. and Fisher,E.M.C. (1995) The frequency and position of Alu repeats in cDNAs, as determined by database searching. *Genomics*, **27**, 544–548.
26. Dagan,T., Sorek,R., Sharon,E., Ast,G. and Graur,D. (2004) AluGene: a database of Alu elements incorporated within protein-coding genes. *Nucleic Acids Res*., **32**, 489–492.
27. Boon,K., Osorio,E.C., Greenhut,S.F., Schaefer,C.F., Shoemaker,J., Polyak,K., Morin,P.J., Buetow,K.H., Strausberg,R.L. *et al*. (2002) An anatomy of normal and malignant gene expression. *Proc. Natl Acad. Sci. USA*, **99**, 11287–11292.
28. Boguski,M.S. (1999) Biosequence exegesis. *Science*, **286**, 453–455.
29. Gerhard,D.S., Wagner,L., Feingold,E.A., Shenmen,C.M., Grouse,L.H., Schuler,G., Klein,S.L., Old,S., Rasooly,R. *et al*. (2004) The MGC Project Team. The status, quality, and expansion of the NIH full-length cDNA project:the Mammalian Gene Collection (MGC). *Genome Res*., **14**, 2121–2127.
30. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2007) Genbank. *Nucleic Acids Res*., **35**, 21–25.
31. Bianchetti,L., Thompson,J.D., Lecompte,O., Plewniak,F. and Poch,O. (2005) vALId: validation of protein sequence quality based on multiple alignment data. *J. Bioinform. Computat. Biol*., **3**, 929–947.
32. Hastie,N.C. and Bishop,J.O. (1976) The expression of three abundance classes of messenger RNA in mouse tissues. *Cell*, **9**, 761–774.
33. Bishop,J.O., Morton,J.G., Rosbash,M. and Richardson,M. (1974) Three abundance classes in HeLa cell messenger RNA. *Nature*, **250**, 199–204.
34. Edgar,R., Domrachev,M. and Lash,A.E. (2002) Gene Expression Omnibus: NCBI gene expression hybridization array data repository. *Nucleic Acids Res*., **30**, 207–210.
35. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.A., Canese,K., Chetvernin,V., Church,D.M., DiCuccio,M., Edgar,R. *et al*. (2007) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*., **35**, 5–12.
36. Pruitt,K.D. and Maglott,D.R. (2001) RefSeq and LocusLink:NCBI gene-centered resources. *Nucleic Acids Res*., **29**, 137–140.
37. Nakao,A., Yoshihama,M. and Kenmochi,N. (2004) RPG: the Ribosomal Protein Gene database. *Nucleic Acids Res*., **32**, 168–70.
38. Jurka,J., Kapitanov,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.
39. Yoshihama,M., Uechi,T., Asakawa,S., Kawasaki,K., Kato,S., Higa,S., Maeda,N., Minoshima,S., Tanaka,T. *et al*. (2002) The human ribosomal protein genes: sequencing and comparative analysis of 73 genes. *Genome Res*., **12**, 379–390.
40. Zinn,A.R., Alagappan,R.K., Brown,L.G., Wool,I. and Page,D.C. (1994) Structure and function of ribosomal protein S4 genes on the human and mouse sex chromosomes. *Mol. Cell. Biol*., **14**, 2485–2492.
41. Colgan,D.F. and Manley,J.L. (1997) Mechanism and regulation of mRNA polyadenylation. *Genes. Dev*., **11**, 2755–2766.
42. Beaudoing,E., Freier,S., Wyatt,J.R., Claverie,J.M. and Gautheret,D. (2000) Patterns of variant polyadenylation signal usage in human genes. *Genome Res*., **10**, 1001–1010.
43. Tian,B., Hu,J., Zhang,H. and Lutz,C.S. (2005) A large-scale analysis of mRNA polyadenylation of human and mouse genes. *Nucleic Acids Res*., **33**, 201–212.
44. Cheng,Y., Miura,R.M. and Tian,B. (2006) Prediction of mRNA polyadenylation sites by support vector machine. *Bioinformatics*, **22**, 2320–2325.
45. Lee,S., Bao,J., Zhou,G., Shapiro,J., Xu,J., Shi,R.Z., Lu,X., Clark,T., Johnson,D. *et al*. (2005) Detecting novel low-abundant transcripts in *Drosophila*. *RNA*, **11**, 939–946.
46. Ishii,K., Washio,T., Uechi,T., Yoshihama,M., Kenmochi,N. and Tomita,M. (2006) Characteristics and clustering of human ribosomal protein genes. *BMC Genomics*, **7**, 37.