BMC Genomics

BioMed Central

Research article

SAGE is far more sensitive than EST for detecting low-abundance transcripts

Miao Sun[1], Guolin Zhou[1], Sanggyu Lee[1], Jianjun Chen[1], Run Zhang Shi[1] and San Ming Wang*[1,2]

Address: [1]Department of Medicine, University of Chicago, 5841 S. Maryland Avenue, MC2115, Chicago, Illinois 60637, USA and [2]ENH Research Institute, Northwestern University, 1001 University Place, Evanston, IL 60201

Email: Miao Sun - msun@medicine.bsd.uchicago.edu; Guolin Zhou - gzhou@medicine.bsd.uchicago.edu; Sanggyu Lee - salee@medicine.bsd.uchicago.edu; Jianjun Chen - jchen@medicine.bsd.uchicago.edu; Run Zhang Shi - rshi@medicine.bsd.uchicago.edu; San Ming Wang* - swang1@midway.uchicago.edu

* Corresponding author

Published: 05 January 2004

BMC Genomics 2004, **5**:1

Received: 03 September 2003
Accepted: 05 January 2004

This article is available from: http://www.biomedcentral.com/1471-2164/5/1

Abstract

Background: Isolation of low-abundance transcripts expressed in a genome remains a serious challenge in transcriptome studies. The sensitivity of the methods used for analysis has a direct impact on the efficiency of the detection. We compared the EST method and the SAGE method to determine which one is more sensitive and to what extent the sensitivity is great for the detection of low-abundance transcripts.

Results: Using the same low-abundance transcripts detected by both methods as the targeted sequences, we observed that the SAGE method is 26 times more sensitive than the EST method for the detection of low-abundance transcripts.

Conclusions: The SAGE method is more efficient than the EST method in detecting the low-abundance transcripts.

Background

Identification of a complete set of transcripts expressed in a genome is one of the ultimate goals of transcriptome studies. Such information is essential for genome annotation and for further study of the function of each gene. It is well known that three classes of transcripts are expressed from a genome, including high-abundance, intermediate-abundance and low-abundance transcripts [1]. Whereas most of the high- and intermediate-abundance transcripts have been identified, it remains a serious challenge to identify fully the low-abundance transcripts [2-4].

Since the beginning of human genome studies, transcript identification has been performed mainly by the use of EST (expressed sequence tag)-based methods [5]. For identification of low-abundance transcripts, extensive subtraction and normalization have been performed in these EST efforts [4,6]. The number of novel transcripts identified in humans through the EST-based approaches has reached a plateau [2,7]. Recently, the SAGE (series analysis of gene expression) method has been applied for transcriptome analyses, with the collection of large numbers of 10-base SAGE tags from different species [8-10]. Although both the EST and the SAGE method are applied to transcriptome study, they use different approaches. The process of the EST method is that of single transcript-sin-

gle clone-single sequencing; thus, each sequence represents a single transcript. In contrast, the process of SAGE follows the approach of multiple transcripts-multiple tags-single clone-single sequencing; thus, each SAGE sequence represents multiple transcripts. Using the same scale of sequence collection, SAGE should detect far more transcripts than does EST; therefore, SAGE might identify more low-abundance transcripts than does EST. Indeed, it is frequently observed that many SAGE tags have no match among the existing ESTs, and most of these SAGE tags have low copy numbers [11-13]. Our previous analyses indicated that the majority of these unmatched SAGE tags are derived from low-abundance transcripts [7]. To determine whether SAGE is indeed more sensitive than the EST method and, if so, to what extent for the detection of low-abundance transcripts, we used existing EST and SAGE data for analysis, and we report our observations.

Results and Discussion

Because a SAGE tag is located at the 3' part of a transcript [8], we used 3' ESTs for comparison. We collected 3' ESTs representing low-abundance transcripts by searching UniGene clusters which contained only a single 3' EST (ftp://ftp.ncbi.nih.gov/repository/UniGene/ Hs.seq.all.gz, UniGene Build #161). We identified 42,500 such UniGene clusters and obtained the same number of 3' ESTs. For comparison with SAGE tags, we extracted virtual tags from these ESTs. We identified 32,587 from the 42,500 3' ESTs that have CATG site(s), a pre-condition for release of a SAGE tag from a transcript, and we extracted 32,587 virtual SAGE tags (10 bases downstream of the last CATG) from the 32,587 sequences. We removed virtual tags that were shared by more than one 3' EST. This resulted in a final set of 22,243 virtual tags from 22,243 3' ESTs representing low-abundance transcripts.

To obtain the experimental SAGE tags for the comparison, we downloaded 477,261 SAGE tags containing 6,847,555 copies collected from 154 SAGE libraries http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL4. Comparison of the 22,243 virtual SAGE tags with the experimental SAGE tag set identified 20,575 tags that were present in both sets. By matching the 20,575 tags in the SAGEmap database (http://www.ncbi.nlm.nih.gov/SAGE/), we identified 2,278 tags that represented the same 3' ESTs detected by both the EST method and the SAGE method. We used the 2,278 tags as the final set for quantitative comparison. Whereas each of the 2,278 virtual tags represents a transcript detected only once by the EST method, the copy number in each of the 2,278 experimental SAGE tags represents the frequency of a transcript detected by SAGE. We observed that the total copy number for the 2,278 experimental SAGE tags appeared 59,754 times; 1,424 (63%) of these SAGE tags appeared between two and more than 100 times. On average, SAGE was 26 times more sensitive than the EST method in detecting these transcripts (Table 1). The data clearly show that the SAGE method is much more sensitive than the EST method for the detection of low-abundance transcripts.

What could be the explanations for the difference between the EST and SAGE methods for detecting the low abundant transcripts?

It is unlikely that the difference is due to the depth of sequence collection. The current number of human ESTs reaches to 4.5 millions including 131,229 mRNAs and 1,470,982 3' ESTs, whereas the total human SAGE tags has about 8 millions. Considering that over 20 tags can be detected by a single SAGE sequence, the number of sequences collected from SAGE is far less than that from ESTs. In our previous studies [2], we observed the "loss" effect on EST collection due to the non-specific polydA/dT hybridization during subtraction / normalization widely used in EST library construction [6], as evidenced by the quantitative loss of a group of targeted transcripts, although it will be difficult to give an absolute rate of loss at the whole genome level due to the complexity of the transcriptome. Such a phenomenon can explain in part but other possibilities may also exist for the loss, such as the limitation of cloning efficiency when ligating cDNAs into vector during cDNA library construction, and clonal loss during library transformation etc. In the SAGE process, there is no subtraction / normalization step, and all the cDNA fragments at each step during SAGE library construction have nearly the same length with the same ends till being cloned into vector. Therefore, the repertoire of the total transcripts is well preserved in SAGE libraries for the detection.

It is true that SAGE method has many limitations for transcript detection. For example, a 14-base SAGE tag contains less sequence information for the detected transcript comparing with an EST that has hundred bases; the specificity of a SAGE tag representing a unique transcript is also lower than that of EST, particularly for SAGE tags at higher copies [14-16]; and SAGE can't detect CATG-negative transcripts, although this number is low as shown that only 151 (7.8%) among the 19,399 full-length human cDNAs in the Refseq (NM) database are CATG-negative. Another issue is related with the error SAGE tags. A SAGE tag has 10 bases. In theory, any base within a single tag could be sequencing error leading to the generation of $4 \times 4 \times 4 \times 4 \times 4 \times 4 \times 4 \times 4 \times 4 \times 4 = 4^{10}$ mutated tags. However, such event doesn't happen in the real world [7]. We have converted thousand SAGE tags into their 3' cDNA experimentally using the GLGI method. From these studies, we clearly see that over 70% of the low-copy SAGE tags represent the real transcripts expressed at low level

Table 1: Comparison between EST and SAGE methods for the detection of low-abundance transcripts

Frequency of detection	Virtual tags from 3' EST	Experimental SAGE tags	
		Tags	Copies
1	2,278	854	854
2		482	964
3		313	939
4		190	760
5		97	485
6 to 10		217	1,578
11 to 20		86	1,234
21 to 100		37	1,279
>100		2	51,661
Total	2,278	2,278	59,754

(these are experimentally confirmed. The real rate may be higher considering the limitation of the experimental sensitivity). Although there are certainly error SAGE tags, these error SAGE tags cannot be a significant portion in the total SAGE tag collection, particularly for the SAGE tags with low copies. Regardless these limitations, SAGE does have unique features for transcriptome study. Among these is that the presence of a SAGE tag implies in large the presence of a transcript.

It is worth to indicate that we only focused on the known low-abundance transcripts for the analysis. For the unknown low-abundance transcripts, many of them may not be present in EST libraries therefore not detectable as novel ESTs. However, these unknown low-abundance transcripts may be well preserved in SAGE libraries therefore readily detectable as novel SAGE tags.

Conclusions

The high sensitivity of the SAGE method for transcript detection becomes valuable for the isolation of low-abundance transcripts. Coupling amplification-based high-throughput methods such as the GLGI (generation of longer 3'cDNA from SAGE tag for gene identification) methods [17] for converting SAGE tags into the original transcripts provides an efficient way for isolating low-abundance transcripts.

Methods

Sequences used for the analysis

The ESTs were downloaded from UniGene database (Build #161) (ftp://ftp.ncbi.nih.gov/repository/UniGene/Hs.seq.all.gz). The UniGene clusters containing CATG+ 3' ESTs were identified. Virtual SAGE tags were extracted from these 3' ESTs after their last CATG sites. The virtual SAGE tags were pooled and tags with the same sequences were then combined to generate the final virtual SAGE tag list from the 3' ESTs with quantitative information for each tag.

The experimental SAGE tags were downloaded from GEO database that contained 154 SAGE libraries http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL4. The SAGE tags from different libraries were pooled. The same SAGE tags in the pool were combined with the copy number to generate the final SAGE tags with quantitative information for each SAGE tags.

Computational process

Computational programs were designed using java language for the extraction of virtual SAGE tags from the 3' ESTs, and for the comparison between the experimental SAGE tags and EST-derived virtual SAGE tags. The programs are available upon request.

List of abbreviations

EST – expressed sequence tag

SAGE – serial analysis of gene expression

GLGI – generation of longer 3'cDNA from SAGE tag for gene identification

Authors' contributions

M.S. carried out all computational analyses. G.Z., S.L., J.C., and R.Z.S. generated experimental data for the development of the concept. S.M.W. designed the study. All authors read and approved the final manuscript.

Acknowledgments

The authors thank Dr. Janet D. Rowley for her encouragement for this study. The study was supported by grants from the G. Harold and Lelia Y. Mathers Charitable Foundation and NIH (S.M.W.).

References

1. Bishop J, Morton J, Rosbach M, Richardson M: **Three abundance classes in HeLa cell messenger RNA.** *Nature* 1974, **250:**199-204.
2. Wang S, Fears S, Zhang L, Chen J, Rowley J: **Screening poly(dA/dT)- cDNAs for gene identification.** *Proc Natl Acad Sci USA* 2000, **97:**4162-4167.
3. Kapranov P, Cawley SE, Drenkow J, Bekiranov S, Strausberg RL, Fodor SP, Gingeras TR: **Large-scale transcriptional activity in chromosomes 21 and 22.** *Science* 2002, **296:**916-919.
4. Carninci P, Waki K, Shiraki T, Konno H, Shibata K, Itoh M, Aizawa K, Arakawa T, Ishii Y, Sasaki D, Bono H, Kondo S, Sugahara Y, Saito R, Osato N, Fukuda S, Sato K, Watahiki A, Hirozane-Kishikawa T, Nakamura M, Shibata Y, Yasunishi A, Kikuchi N, Yoshiki A, Kusakabe M, Gustincich S, Beisel K, Pavan W, Aidinis V, Nakagawara A, Held WA, Iwata H, Kono T, Nakauchi H, Lyons P, Wells C, Hume DA, Fagiolini M, Hensch TK, Brinkmeier M, Camper S, Hirota J, Mombaerts P, Muramatsu M, Okazaki Y, Kawai J, Hayashizaki Y: **Related Articles, Links Abstract Targeting a complex transcriptome: the construction of the mouse full-length cDNA encyclopedia. Targeting a complex transcriptome: the construction of the mouse full-length cDNA encyclopedia.** *Genome Res* 2003, **13:**1273-1289.
5. Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merril CR, Wu A, Olde B, Moreno RF, Kerlavage AR, McCombie WR, Venter JC: **Complementary DNA sequencing: expressed sequence tags and human genome project.** *Science* 1991, **252:**1651-1656.
6. Bonaldo M, Lennon G, Soares M: **Normalization and subtraction: two approaches to facilitate gene discovery.** *Genome Res* 1996, **6:**791-806.
7. Chen J, Sun M, Lee S, Zhou G, Rowley J, Wang S: **Identifying novel transcripts and novel genes in the human genome by using novel SAGE tags.** *Proc Natl Acad Sci USA* 2002, **99:**12257-12262.
8. Velculescu V, Zhang L, Vogelstein B, Kinzler K: **Serial analysis of gene expression.** *Science* 1995, **270:**484-487.
9. Velculescu VE, Madden SL, Zhang L, Lash AE, Yu J, Rago C, Lal A, Wang CJ, Beaudry GA, Ciriello KM, Cook BP, Dufault MR, Ferguson AT, Gao Y, He TC, Hermeking H, Hiraldo SK, Hwang PM, Lopez MA, Luderer HF, Mathews B, Petroziello JM, Polyak K, Zawel , Zhang W, Zhang X, Zhou W, Haluska FG, Jen J, Sukumar S, Landes GM, Riggins GJ, Vogelstein B, Kinzler KW: **Analysis of human transcriptomes.** *Nat Genet* 1999, **23:**387-388.
10. Boon K, Osorio EC, Greenhut SF, Schaefer CF, Shoemaker J, Polyak K, Morin PJ, Buetow KH, Strausberg RL, De Souza SJ, Riggins GJ: **An anatomy of normal and malignant gene expression.** *Proc Natl Acad Sci USA* 2002, **99:**11287-11292.
11. Zhou G, Chen J, Lee S, Clark T, Rowley JD, Wang SM: **The pattern of gene expression in human CD34+ hematopoietic stem/progenitor cells.** *Proc Natl Acad Sci USA* 2001, **98:**13966-13971.
12. Lee S, Zhou G, Clark T, Chen J, Rowley JD, Wang SM: **The pattern of gene expression in human CD15+ myeloid progenitor cells.** *Proc Natl Acad Sci USA* 2001, **98:**3340-3345.
13. Hashimoto S, Nagai S, Sese J, Suzuki T, Obata A, Sato T, Toyoda N, Dong HY, Kurachi M, Nagahata T, Shizuno K, Morishita S, Matsushima K: **Gene expression profile in human leukocytes.** *Blood* 2003, **101:**3509-3513.
14. Chen J, Rowley J, Wang SM: **Generation of longer cDNA fragments from serial analysis of gene expression tags for gene identification.** *Proc Natl Acad Sci USA* 2000, **97:**349-353.
15. Lee S, Clark T, Chen J, Zhou G, Scott LR, Rowley JD, Wang SM: **Correct identification of genes from SAGE tag sequences.** *Genomics* 2002, **79:**598-599.
16. Clark T, Lee S, Ridgway S, Wang SM: **Computational analysis of gene identification with SAGE.** *J Comput Biol* 2002, **9:**513-526.
17. Chen J, Lee S, Zhou G, Wang SM: **High-throughput GLGI procedure for converting a large number of serial analysis of gene expression tag sequences into 3' complementary DNAs.** *Genes Chromosomes Cancer* 2002, **33:**252-256.