Nucleic Acids Research, 2005, Vol. 33, Web Server issue ***W693–W695***
doi:10.1093/nar/gki444

WEBSAGE: a web tool for visual analysis of differentially expressed human SAGE tags

Jean Pylouster*, Catherine Sénamaud-Beaufort and Tula Ester Saison-Behmoaras

Laboratoire de Biophysique, Muséum National d'Histoire Naturelle, INSERM U565-CNRS UMR 5153, 43 rue Cuvier 75231, Paris Cedex 05, France

Received February 10, 2005; Revised and Accepted March 30, 2005

ABSTRACT

The serial analysis of gene expression (SAGE) is a powerful method to compare gene expression of mRNA populations. To provide quantitative expression levels on a genome-wide scale, the Cancer Genome Anatomy Project (CGAP) uses SAGE. Over 7 million SAGE tags, from 171 human cell types have been assembled. The growing number of laboratories involved in SAGE research necessitates the use of software that provides statistical analysis of raw data, allowing the rapid visualization and interpretation of results. We have created the first simple tool that performs statistical analysis on SAGE data, identifies the tags differentially expressed and shows the results in a scatter plot. It is freely available and accessible at http://bioserv.rpbs.jussieu.fr/websage/index.php.

INTRODUCTION

Serial analysis of gene expression (SAGE) is a powerful method for obtaining comprehensive and quantitative gene expression profiles from cell populations under selected physiological conditions. SAGE counts polyadenylated transcripts by sequencing a short 14 bp tag at the gene's 3′ end, adjacent to the last restriction site. NlaIII is the most commonly used anchoring enzyme. The tag counts are than archived electronically for future analysis and comparisons. Since the first publication introducing SAGE (1), computational tools (2–7) and statistical methods (8–20) have been developed to correctly perform the analysis of a SAGE experiment. Because SAGE determines absolute expression levels, comparisons between different SAGE libraries are relatively easy to perform. For this reason, SAGE has been selected as the major platform technology for the Cancer Genome Anatomy Project (CGAP). For 5 years, a large number of SAGE libraries, generated from diverse cancer and normal tissues in many laboratories, have been accessible via the National Center for Biotechnology Information website (http://www.ncbi.nlm.nih.gov/SAGE). The SAGE Genie website (http://cgap.nci.nih.gov/SAGE) was constructed recently and is equipped with additional search and presentation tools (7). One of these tools allows the comparison of several libraries (SAGEmap xProfiler). SAGE Genie compiles 171 human SAGE libraries containing 6.8 million SAGE tags. These libraries were all constructed by using the same anchoring enzyme, thus all yielding tags likely to be 10 bp downstream from the 3′ most NlaIII site in the transcript. The aim of most SAGE studies is to identify genes of interest by comparing the number of specific tags found in two different SAGE libraries. One of the most attractive applications of transcript profiling is to address the question of expression differences between normal and cancer samples or cancer and metastasis samples in order to define new diagnostic markers and therapeutic targets. Over the past few years, several methods have been reported for determining the statistical significance of gene expression difference provided by the SAGE experiments. Most of these methods have been incorporated into public database systems and analysis programs (2–4,6–8,10–13,18–20).

Several statistical approaches can be used to test for differential expression in SAGE data. If within two types of cells Y and Z, a particular mRNA species has unknown respective concentrations y and z and a total of A tags are sequenced from cell type Y and B tags from cell type Z, and among these, a and b tags, respectively, correspond to the mRNA of interest, the question is: what inference may be made about the relative size of the actual concentrations, y and z? Audic and Claverie (8) have used a Bayesian method for SAGE data analysis. They consider the null hypothesis H_0 that $y = z$, and the alternative that $y \neq z$, and derive formulas based upon the observed data for rejecting H_0 with various degrees of confidence. If a/A and b/B differ significantly, one rejects H_0, concluding that the expression levels y and z are unequal. We have used their statistical approach and developed a useful and flexible tool (WEBSAGE) that analyse and compare a large number of SAGE tags. The novelty of our tool compared with those available online (3,4) consists of the visualization of the results

*To whom correspondence should be addressed. Tel: +33 01 40 79 36 95; Fax: +33 01 40 79 37 05; Email: jpylou@mnhn.fr

of the comparison of two SAGE libraries in a scatter plot. One of the libraries is presented on the *x*-axis and the second one on the *y*-axis. More than two libraries cannot be visualized using our tool unless to pool libraries composed by the sum of all libraries from one group such as cancer group and all libraries from normal group. However, Vêncio *et al*. (20) showed clearly that this summing is an error and proposed a statistical analysis that accounts for the within class variability. In our tool, full SAGE data are very comprehensively represented in plots, which correspond to one or a set of tags, having the same *P*-value and the same apparition in two libraries. Moreover, WEBSAGE gives not only the identification but also the function of the gene.

SYSTEM AND METHODS

To gain insights into the molecular basis for metastasis, we compared the global gene expression profile of metastatic pancreatic cancer with that of primary cancer. The two cell lines were established from a 60-year-old woman with surgically removed pancreatic carcinoma that presented simultaneously metastatic liver lesion (21). We isolated total cellular RNA from pancreatic cancer cells and metastatic liver cells and constructed each SAGE library, using SAGE protocol version 1.0c (1). Afterwards, tags were extracted from the raw sequence data and repeated ditags were excluded using the SAGE software (version 1) developed by Kinzler and co-workers (1). From the two SAGE libraries 12 090 tags were obtained, enabling us to compare the expression levels of 4807 unique transcripts. At this step, the question of interest is whether one sample has a significant change in expression relative to the other sample for each transcript. Over the past few years, several methods have been developed for determining the statistical significance of gene expression difference derived from the SAGE experiments. Zhang *et al*. (17) used a simulation approach to determine the probability of obtaining the observed difference and more extreme ones. Other statistical approaches of the number of specific tags found in SAGE can be envisioned. Madden *et al*. (19) used a test for differences in tag numbers found in two experimental conditions, seemingly based on Poisson distribution statistics. Their approach suffers from the disadvantage that it can only be used reliably on tag libraries of similar size. The second approach is to look at the number of copies of a specific mRNA per cell as a fraction or proportion of the total number of mRNA molecules in that cell. The same proportion (p) of specific tags should be present in the SAGE library of all sequenced tags. In this approach, the number of tags found

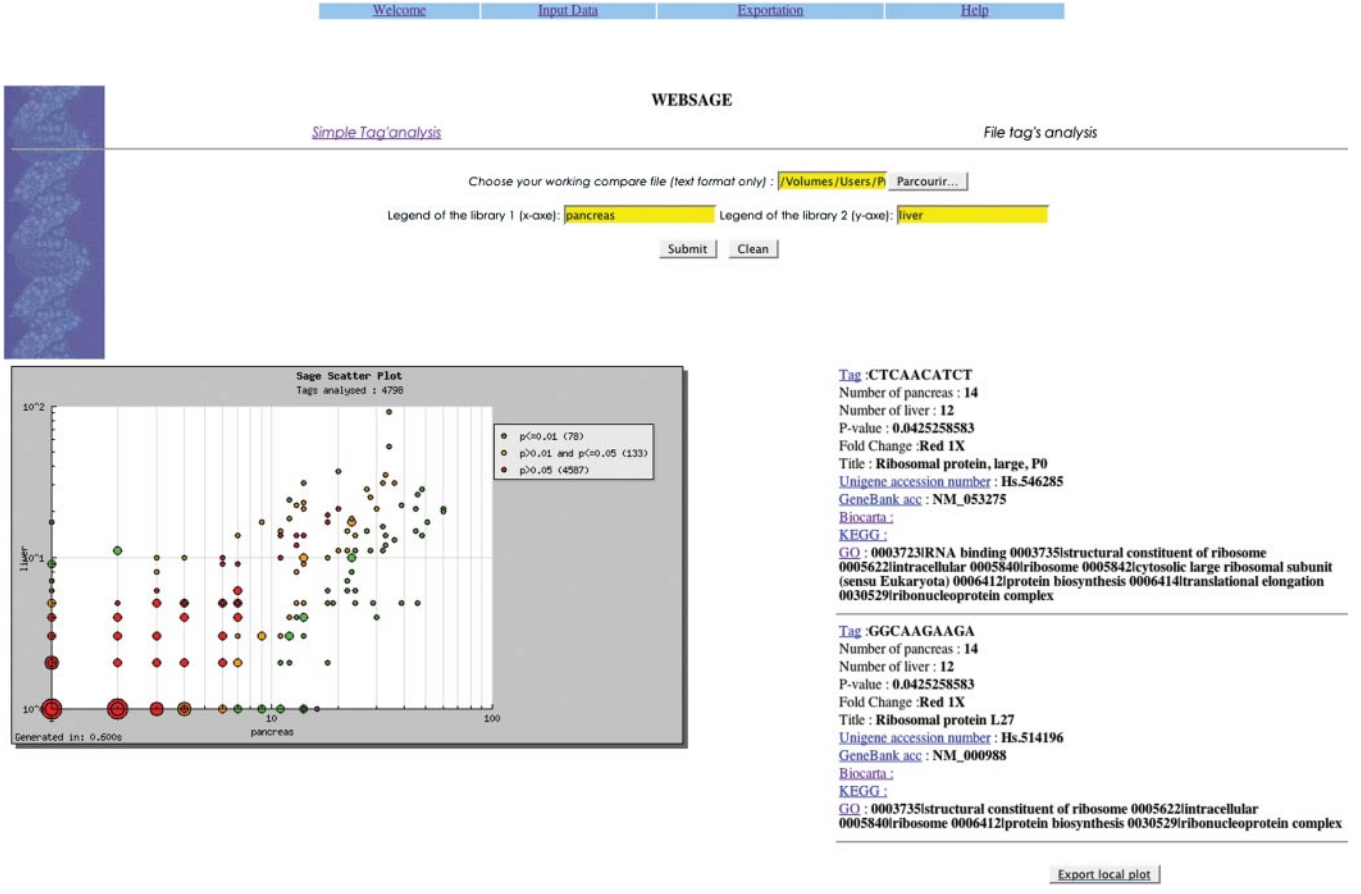

Figure 1. An example of WEBSAGE analysis of differently expressed tags. A total of 4809 tags from pancreatic cancer lines and an isogenic liver metastasis cell line were analysed. A scatter plot is automatically computed from the file to visualize the results. Each plot corresponds to one tag or a set of tags, having the same *P*-value and the same apparition on the two libraries. The size of the plot is proportional to the number of tags contained within it. Once clicked, a web page is displayed with a list of tags and their identification. The identification consists of the name of the tag and the corresponding gene, the *P*-value and the Unigene and GenBank accession number.

is binomially distributed because it is the result of the *p* of each tag to be identified as being the specific mRNA or not. A Bayesian method has been used by Audic and Claverie (8) and Vêncio *et al*. (20) for examining the SAGE data. Comparison of this test and the test based on Poisson-distributed tag counts used by Madden *et al*. (19) shows a difference in sensitivity, the test of Madden *et al*. being more conservative (10). To quantify the transcripts of the two cell line, we have used the significance test (http://igs-server.cnrs-mrs.fr/~audic / significance.html) of Audic and Claverie (8) and developed a tool to visualizing the analysis of differentially expressed tags in a scatter-plot shape (log/log coordinates). WEBSAGE is the software that enables a rapid and comprehensive analysis of the SAGE data and integrates statistical data analysis methods using a database system. WEBSAGE enables simple and rapid analysis of a file of tags and the determination of differentially expressed tags (*P*-values). To provide tag-to-gene links, three files from the NCBI ftp site (SAGE Genie) were stored in database (MySQL) and used at each analysis, to get the tag-to-gene assignment, gene accession number (UniGene and GenBank) and gene function information (Kegg, Biocarta and Gene Ontology databank), respectively. Database is updated manually every 3 months. SAGE Genie provides the best match between gene name and tag, although there are options for manually viewing an alternate tag's expression data. Details concerning the 'Best tag' selection method can be obtained at http://cgap.nci.nih.gov/SAGE/SAGEHelp. Furthermore, the information is stored in a table and visualized in a scatter plot. Figure 1 shows an example of WEBSAGE analysis of our file of tags. The total number of tags analysed included 5257 tags from pancreatic cancer cell line and 6833 from cell line derived from liver metastasis. WEBSAGE computed a *P*-value and classified the tags into three groups: green, $P \leqslant 0.01$ (significant); red, $P > 0.05$ (not significant); and yellow, *P* in]0.01; 0.05] (require further analysis). The information computed for the query is stored in a temporary table during the session. Next, the data are exported in the CSV format (compatibility with all spreadsheets). All the results or only plot's results can be exported. Our application is a web tool, developed in the PHP language using jpgraph library (http://www.aditus.nu/jpgraph/index.php) that has been modified. It runs actually on Apache web server and MySQL SGBD (http://www.mysql.com). A help online is accessible on the site.

In conclusion, WEBSAGE is a useful web service that performs statistical analysis on the SAGE data, identifies the tags differentially expressed and shows the results in a scatter plot. Moreover, the user can query with a specific tag and get a tag-to-gene assignment and gene function from Kegg, Biocarta and Gene Ontology databank.

ACKNOWLEDGEMENTS

We thank Mr David Polverari for helpful discussions and RPBS (Réseau Parisien de Biologie Structurale) for the technical support. The authors are also indebted to two anonymous reviewers who provided very constructive critiques and excellent suggestions. Funding to pay the Open Access publication charges for this article was provided by INSERM.

Conflict of interest statement. None declared.

REFERENCES

1. Velculescu,V.E., Zhang,L., Vogelstein,B. and Kinzler,K.W. (1995) Serial analysis of gene-expression. *Science*, **270**, 484–487.
2. 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.
3. Margulis,E.H. and Innis,J.W. (2000) eSage: managing and analysing data generated with serial analysis of gene expression (SAGE). *Bioinformatics*, **16**, 650–651.
4. van Kampen,A.H., van Schaik,B.D., Pauws,E., Michiels,E.M., Ruijter,J.M., Caron,H.N., Versteeg,R., Heisterkamp,S.H., Leunissen,J.A., Baas,F. *et al*. (2000) USAGE: a web-based approach towards the analysis of SAGE data. *Bioinformatics*, **16**, 899–905.
5. Claverie,J.M. (1999) Computational methods for the identification of differential and coordinated gene expression. *Hum. Mol. Genet*., **8**, 1821–1832.
6. Buetow,K.H., Klausner,R.D., Fine,H., Kaplan,R., Singer,D.S. and Strausberg,R.L. (2002) Cancer Molecular Analysis Project: weawing a rich cancer research tapestry. *Cancer Cell*, **1**, 315–318.
7. Boon,K., Osorio,E.C., Greenhut,S.F., Schaefer,C.F., Shoemaker,J., Polyak,K., Morin,P.J., Buetow,K.H., Strausberg,R.L., De Souza,S.J. *et al*. (2002) An anatomy of normal and malignant gene expression. *Proc. Natl Acad. Sci. USA*, **99**, 11287–11292.
8. Audic,S. and Claverie,J.-M. (1997) The significance of digital gene expression profiles. *Genome Res*., **7**, 986–995.
9. Romualdi,C., Bortoluzzi,S. and Danieli,G.A. (2001) Detecting differentially expressed genes in multiple tag sampling experiments: comparative evaluation of statistical tests. *Hum. Mol. Genet*., **10**, 2133–2141.
10. Ruijter,J.M., Kampen,A.H.C. and Baas,F. (2002) Statistical evaluation of SAGE libraries: consequences for experimental design. *Physiol. Genomics*, **11**, 37–44.
11. Stollberg,J., Urschitz,J., Urban,Z. and Boyd,C.D. (2000) A quantitative evaluation of SAGE. *Genome Res*., **10**, 1241–1248.
12. Kal,A.J., van Zonneveld,A.J., Benes,V., van den Berg,M., Koerkamp,M.G., Alberman,K., Starck,N., Ruijter,J.M., Richter,A., Dujon,B., Ansorge,W. and Tabak,H.F. (1999) Dynamics of gene expression revealed by comparison of serial analysis of gene expression transcript profiles from yeast grown on two different carbon sources. *Mol. Biol. Cell*, **10**, 1859–1872.
13. Man,M.Z., Wang,X. and Wang,Y. (2000) POWER_SAGE: comparing statistical tests for SAGE experiments. *Bioinformatics*, **16**, 953–959.
14. Michiels,E.M.C., Oussoren,E., Van Grenigen,M., Pauws,E., Bossuyt,P.M.M., Voûte,P.A. and Baas,F. (1999) Genes differentially expressed in medulloblastoma and fetal brain. *Physiol. Genomics*, **1**, 83–91.
15. Chen,H., Centola,M., Altschul,S.F. and Metzger,H. (1998) Characterization of gene expression in resting and activated mast cells. *J. Exp. Med*., **188**, 1657–1668.
16. Lal,A., Lash,A.E., Altschul,S.F., Velculescu,V., Zhang,L., McLendon,R.E., Marra,M.A., Prange,C., Morin,P.J., Polyak,K. *et al*. (1999) A public database for gene expression in human cancers. *Cancer Res*., **59**, 5403–5407.
17. Zhang,L., Zhou,W., Velculescu,V.E., Kern,S.E., Hruban,R.H., Hamilton,S.R., Vogelstein,B. and Kinzler,K.W. (1997) Gene expression profiles of normal and cancer cells. *Science*, **276**, 1268–1272.
18. Kenzelman,M. and Mühlemann,K. (2000) Transcriptome analysis of fibroblast cells immediate-early after human cytomegalovirus infection. *J. Mol. Biol*., **304**, 741–751.
19. Madden,S.L., Galella,E.A., Zhu,J.S., Bertelsen,A.H. and Beaudry,G.A. (1997) Sage transcript profiles for P53-dependent growth regulation. *Oncogene*, **15**, 1079–1085.
20. Vêncio,R., Brentani,H., Patrão,D. and Pereira,C. (2004) Bayesian model accounting for within-class biological variability in Serial Analysis of Gene Expression (SAGE). *BMC Bioinformatics*, **5**, 119.
21. Curtis,L.J., Li,Y., Gerbault-Seureau,M., Kuick,R., Dutrillaux,A-M., Goubin,G., Fawcett,J., Cram,S., Dutrillaux,B., Hanash,S. and Muleris,M. (1998) Amplification of DNA Sequences from chromosome 19q13.1 in human pancreatic cell lines. *Genomics*, **53**, 42–55.