Nucleic Acids Research, 2007, Vol. 35, Web Server issue ***W163–W168***
doi:10.1093/nar/gkm429

SAGExplore: a web server for unambiguous tag mapping in serial analysis of gene expression oriented to gene discovery and annotation

Tomás Norambuena, Rodrigo Malig and Francisco Melo*

Departamento de Genética Molecular y Microbiología, Facultad de Ciencias Biológicas, Pontificia Universidad Católica de Chile, Alameda 340, Santiago, Chile

Received January 31, 2007; Revised April 29, 2007; Accepted May 13, 2007

ABSTRACT

We describe a web server for the accurate mapping of experimental tags in serial analysis of gene expression (SAGE). The core of the server relies on a database of genomic virtual tags built by a recently described method that attempts to reduce the amount of ambiguous assignments for those tags that are not unique in the genome. The method provides a complete annotation of potential virtual SAGE tags within a genome, along with an estimation of their confidence for experimental observation that ranks tags that present multiple matches in the genome.

The output of the server consists of a table in HTML format that contains links to a graphic representation of the results and to some external servers and databases, facilitating the tasks of analysis of gene expression and gene discovery. Also, a table in tab delimited text format is produced, allowing the user to export the results into custom databases and software for further analysis.

The current server version provides the most accurate and complete SAGE tag mapping source that is available for the yeast organism. In the near future, this server will also allow the accurate mapping of experimental SAGE-tags from other model organisms such as human, mouse, frog and fly. The server is freely available on the web at: http://dna.bio.puc.cl/SAGExplore.html.

INTRODUCTION

A key process in serial analysis of gene expression (SAGE) consists of accurately mapping short sequence tags to known genes. The use of complete genome information, instead of limited and biased transcriptome data, allows the identification and mapping of a larger number of experimental tags, thus facilitating the tasks of gene discovery and annotation. The use of genome information in the tag-to-gene assignment process overcomes the problem of being limited to only those genes for which an EST has been already found. However, this strategy posses new challenge for unambiguous tag mapping because the probability that a short tag sequence will be unique in the genome significantly decreases (1).

In a recent work (2), we have presented a novel and improved method for the tag-to-gene assignment process in SAGE, called hierarchical gene assignment (HGA). The HGA method provides a full annotation of the potential virtual SAGE tags within a genome, along with an estimation of their confidence for experimental observation. We applied this method to the *Saccharomyces cerevisiae* genome, producing the most thorough and accurate annotation of virtual SAGE tags that is available today for this organism.

In this work, we describe the implementation of a web server that can be used to map experimental SAGE tags from yeast against our previously generated annotation of potential genomic tags for this organism using the HGA methodology (2). The server is specifically designed to fully exploit the major benefits of the SAGE technique, which are to assist the processes of gene discovery and annotation (3–5).

SERVER DESCRIPTION

Overview

The server contains three different modules (Figure 1): (i) Genome Explore, (ii) Genome Mapping and (iii) Library Mapping. The first module can be used to explore a genome in the context of a future SAGE experiment, allowing the user to determine before hand if some genes of interest will be accurately measured by

*To whom correspondence should be addressed. Tel: +56 2 686 2279; Fax: +56 2 222 5515; Email: fmelo@bio.puc.cl

The authors wish if to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

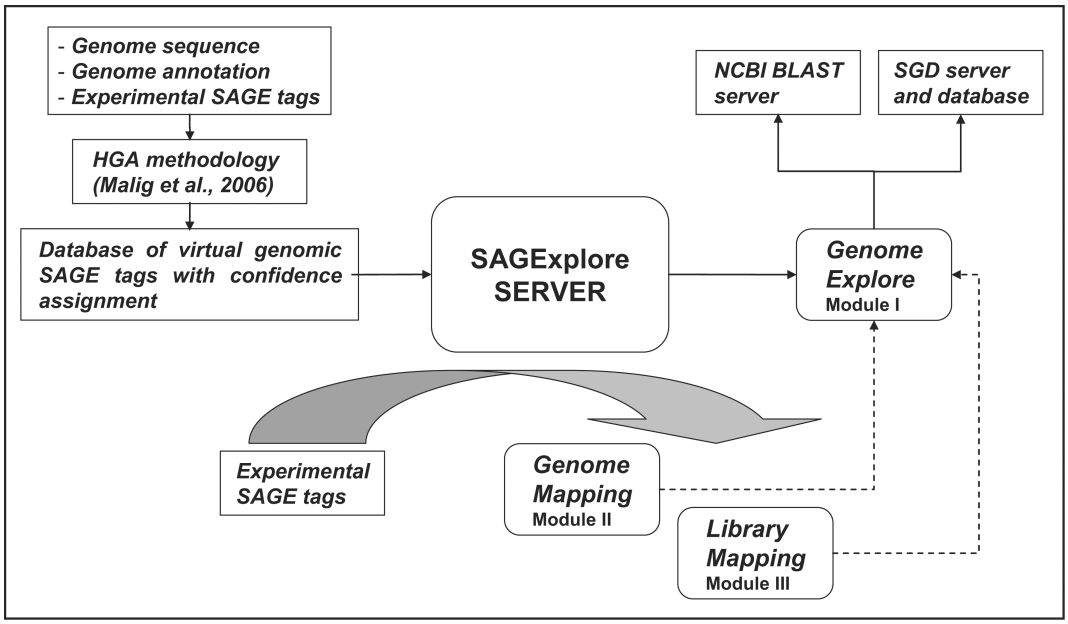

Figure 1. Layout and flowchart of the SAGExplore server. The core of the server is a MySQL database of virtual genomic SAGE tags with confidence assignments that was generated by the HGA methodology and it has been recently described in detail (2). The server has three modules: (i) Genome Explore, (ii) Genome Mapping and (iii) Library Mapping. The module I is linked to NCBI BLAST server and to the Saccharomyces Genome Database (SGD) and does not require any input from the user. The modules II and III require experimental tag sequences and their counts (optional) as input, are linked to the module I and, through this, to the external servers and databases. For more details about the functioning of these modules see the main text. This server has been programmed in MySQL and PHP languages and uses the JpGraph graphics library.

SAGE (i.e. systems biology, study of gene regulatory networks or specific metabolic pathways, etc). This is useful for the planning of SAGE experiments and it can also be used for education purposes when teaching about the SAGE technique. This module provides a friendly graphical interface, links to external servers and databases and it is also invoked by the other two modules. The second module can be used to map experimental SAGE tags against the existing annotation of potential genomic tags. The results are clearly presented in a table that contains dynamic links to the graphical interface of the genome explore module and to external servers and databases. Graphical expression maps of specific genes, specific genomic regions, full chromosomes or the complete genome can also be produced on the flight. The third module can be used to map experimental SAGE tags against all existing libraries of experimental SAGE tags produced by others. This allows the user to quickly compare its own SAGE results against those from previous SAGE experiments performed upon different conditions. This is useful to identify new SAGE tags and also to easily and simultaneously compare two or more full gene expression profiles.

Input data and parameters

Though the different modules of the SAGExplore server have distinct functionalities and independent forms for submitting a query, some of their input parameters are common. Table 1 summarizes the complete list of user-specified input parameters and user-provided input data that each module requires. Common parameters to the three modules include the specification of the organism name and the anchoring-tagging enzyme pair used in SAGE, as well as the selection of several options for displaying the results of a query.

When exploring a genome, various tag features can be specified by the user, which include: frequency of occurrence of a given tag sequence in the genome, annotated elements in the genome where a tag maps, the tag confidence assignment (high, low or undefined), location within a gene element (ORF, 3′-UTR, 5′-UTR, exon or intron), and if the tag is near and downstream to an internal poly(A) region within a gene (Figure 2A; Supplementary Figure 2). This allows the user to study in detail the reliability of any potential virtual tag in the genome. Additionally, the user can also provide to the server any input data specifying different type of genomic regions to explore such as: one or more genes, one or more genome fragments, one or more chromosomes, or the complete genome (Figure 2A; Supplementary Figure 2). This allows the user to evaluate the expected reliability of SAGE results for a specific set of genes or genomic regions of interest. Furthermore, the combination of these user-provided input parameters and user-specified input data provides a powerful tool to perform almost any possible query. On-line help about the required format to submit input data is available on each query form of the server. In addition to this, some simple examples of properly formatted input data are also provided.

When mapping experimental tags against the database of genomic virtual tags or against the known experimental SAGE libraries, a list of experimental tag sequences should be provided by the user. The observed counts for

Table 1. Input options and requirements of the different modules of the SAGExplore server

Input	Description	Genome explore	Genome mapping	Library mapping
User-specified input parameters	Organism specification	Yes	Yes	Yes
	Anchoring-Tagging enzyme pair	Yes	Yes	Yes
	Odds ratio for tag confidence assignment	Yes	Yes	No
	Genomic mapping context and tag categories	Yes	No	No
	Output display options	Yes	Yes	Yes
User-provided input data	List of genomic regions	Yes	No	No
	List of experimental SAGE tags	No	Yes	Yes

each tag from multiple experimental points can also be included. In the case of the genome mapping module, this allows the building of graphic genome expression maps. In the case of the library mapping module, this allows a simple and fast comparison against existing gene expression profiles under different experimental conditions previously reported by others (6–8).

Output of the server

The output of the server for the three modules is presented as a table in HTML format, which can also be exported as a compressed text file (tab delimited text). Therefore, the output data can be easily imported into other software or database applications such as Excel or MySQL for further analysis. Each column header in the output table is linked to a popup window that contains online help explaining its content and/or functionality.

The output data that is displayed varies depending on the particular module. Table 2 shows the complete description of the output data given by each module of the server. As an example, a typical output of the genome mapping module is shown (Figure 2B; Supplementary Figure 3). Some columns in the output table contain a dynamic link that allows the user to retrieve more information about a particular tag by invoking additional queries to this server or to external servers and databases. One of these features consists of the analysis of the genomic context where a given tag maps, which can be graphically explored (Figure 2C; Supplementary Figure 4). This allows the user to see the mapping position of a tag in the genome, along with all the surrounding annotated elements such as genes and their structures (i.e. coding regions and UTRs). Another important feature available is the graphical display of genome expression maps. The server can generate these maps for the complete genome, for a single chromosome (Figure 2D; Supplementary Figure 5) or for a given genomic region, in case the user wants to analyze some specific regions in more detail. The graphic expression maps facilitate the analysis of SAGE data and allow the easy and fast identification of transcriptionally active regions or co-regulated gene clusters under certain experimental conditions.

The complete nucleotide sequence of a gene where a tag maps, along with the detailed representation of the annotated gene structure, can also be automatically extracted and analyzed (Figure 2E; Supplementary Figure 6). In the case that a tag maps onto an intergenic region, a flanking genomic sequence is extracted by the server (500 nts downstream and 500 nts upstream from the tag mapping position). In either case, the extracted sequences that contain the tag can be automatically aligned against the known sequence databases through the BLAST server. These server features are very powerful because they allow a fast and detailed analysis of those interesting tags that could be coming from currently unknown genes (i.e. assisting the processes of gene

Table 2. Output data given by the different modules of the SAGExplore server

Column description	Genome explore	Genome mapping	Library mapping
Sequential tag number	Yes (1)	Yes (1)	Yes (1)
Tag confidence assignment	Yes (2)	Yes (2)	No
Tag sequence	Yes (3)	Yes (3)	Yes (2)[a]
Tag frequency of occurrence in the genome	Yes (4)	Yes (4)	No
Tag odds ratio	Yes (5)	Yes (5)	No
Tag class	Yes (6)	Yes (6)	No
Tag genomic location description	Yes (7)	Yes (7)	No
Tag genomic location type	Yes (8)	Yes (8)	No
Tag position within a transcript	Yes (9)	Yes (9)	No
Chromosome number	Yes (10)	Yes (10)	No
Initial position of the tag in the chromosome	Yes (11)	Yes (11)	No
Chromosome strand	Yes (12)	Yes (12)	No
Standard gene name	Yes (13)[a]	Yes (13)[a]	No
Systematic gene name	Yes (14)[a]	Yes (14)[a]	No
Genomic context	Yes (15)[a]	Yes (15)[a]	No
Tag details	Yes (16)[a]	Yes (16)[a]	No
Display sequence	Yes (17)[a]	Yes (17)[a]	No
BLAST	Yes (18)[a]	Yes (18)[a]	No
Tag counts on each experimental library	No	No	Yes (3–10)
Tag counts	No	Yes (19)	Yes (11)
Tag user-defined information	No	Yes (20)	Yes (12)

[a]This field in the output table has additional information dynamically linked. Some of these links currently point to external servers and databases such as BLAST server and SGD database.
The description of the columns displayed for each SAGE tag by the output tables of the server as a result of a particular query issued to each of the independent modules is specified. The numbers between parenthesis represent the sequential column number of each output table displayed by the server on its three different modules as a result of an issued query.

Table 3. Existing tools for the analysis and mapping of SAGE tags

Name	Database	Type	Tag counts	Tag mapping	Graphical interface	Organism	Web address
TAGmapper (9)	RefSeq, ESTs	Server	No	Yes	No	Several	http://tagmapper.ibioinformatics.org/
WebSAGE (10)	RefSeq, ESTs	Server	No	Yes	No	Human	http://www2.mnhn.fr/websage/
SAGEmap (11)	SAGE libraries	Server	Yes	Yes	Yes	Several	http://www.ncbi.nlm.nih.gov/projects/SAGE
SAGEnet (N.A.)	SAGE libraries	Database	No	No	No	Human, Mouse, Yeast	http://www.sagenet.org/
SAGE genie (12)	SAGE libraries	Database	No	No	Yes	Human, Mouse	http://cgap.nci.nih.gov/SAGE
ACTG (13)	RefSeq, ESTs	Server	Yes	Yes	No	Human, Mouse	http://retina.med.harvard.edu/ACTG/
5′SAGE (14)	Genome, ESTs, SAGE libraries	Server	No	Yes	Yes	Human	http://5sage.gi.k.u-tokyo.ac.jp/
Mouse SAGE Site (15)	RefSeq, Genome, ESTs, SAGE libraries	Server	No	Yes	No	Mouse	http://mouse.biomed.cas.cz/sage/
Discovery Space (16)	RefSeq, Genome, SAGE libraries	Standalone	Yes	Yes	Yes	Human	http://www.bcgsc.ca/discoveryspace/
Identitag (17)	ESTs, cDNA, SAGE libraries (user provided)	Standalone	No	Yes	No	Several	http://pbil.univ-lyon1.fr/software/identitag/
USAGE (18)	RefSeq, Genome, SAGE libraries	Server	Yes	Yes	No	Several	http://www.cmbi.kun.nl/usage/
SAGExplore[a]	Genome, SAGE libraries	Server	Yes	Yes	Yes	Yeast[b]	http://dna.bio.puc.cl/SAGExplore.html

[a]This work.
[b]Other organisms will be soon available for tag mapping on this server. N.A., 'Not Available'.
The available publications describing the listed tools are cited between parenthesis next to their names. RefSeq and ESTs stand for databases of reference sequences and expressed sequence tags, respectively. Both vary on clustering parameters and definitions. Tag counts column reflects if the tool takes into consideration in some way the observed counts of experimental tags. Some tools perform statistical analysis based on those counts, others only allow to display that information along with the results of the query. Tag mapping column reflects if the tool is able to map experimental tags against a specific built-in database.

discovery and annotation). In addition to this, the rapid design of oligonucleotide primers for experimental validation of some SAGE results by RT-PCR is also greatly facilitated. The ranking of tags for experimental validation is also greatly facilitated by accessing all the annotated tag details (Figure 2F; Supplementary Figure 7).

Other existing SAGE tools

Several other tools have been described for the analysis and mapping of experimental SAGE tags (Table 3). The current release of the SAGExplore server presents several drawbacks as compared to some other tools, most of which are considered as future improvements of this server and are detailed in the next section subsequently. On the other hand, the SAGExplore server has several advantages and some unique features as compared to the other tools, which include: (i) the database of virtual SAGE tags that the server uses has been built by the recently described HGA methodology that assigns a confidence level based on experimental data to those tags that present multiple matches in the genome; (ii) its particular orientation towards facilitating the tasks of gene discovery and annotation; (iii) its graphical interface and the genome explore module, which can also be used for educational purposes and not only for advanced research and (iv) its genomic tag context sequence extraction and tag details display capabilities, which are very useful to speed up the experimental validation of SAGE results.

FUTURE IMPROVEMENTS

The current release of the SAGExplore server only allows the exploring and mapping of SAGE tags to the yeast genome. Also, the virtual tag database was built for a single combination of anchoring-tagging enzymes (NlaIII-BsmFI). Though this enzyme pair is the most frequently used in SAGE experiments, it is expected that other enzyme pairs could be useful to some experimentalists (e.g. long-SAGE uses a different tagging enzyme). Therefore, obvious improvements to the server involve the incorporation of additional organisms and enzyme pairs generally used in SAGE and long-SAGE. In addition to this, the odds ratio used to assign the tag confidences for those cases where a tag is found multiple times in the genome has been specifically tuned for yeast, based on experimental data (2). It is expected that other organisms will have a different optimal odds ratio threshold to define tag confidences according to the HGA methodology. Therefore, flexibility about this parameter needs to be also added when new organisms are included. In addition to this, genome annotation is improved very often by the experimental characterization of new genes. This information is key in the HGA methodology and the database of virtual SAGE-tags should be also updated frequently. Finally, a large amount of SAGE experiments are underway, and thus several experimental libraries are released every year. It will be then necessary to update frequently the database containing the experimental libraries, which are used by the Library Mapping module of this server.

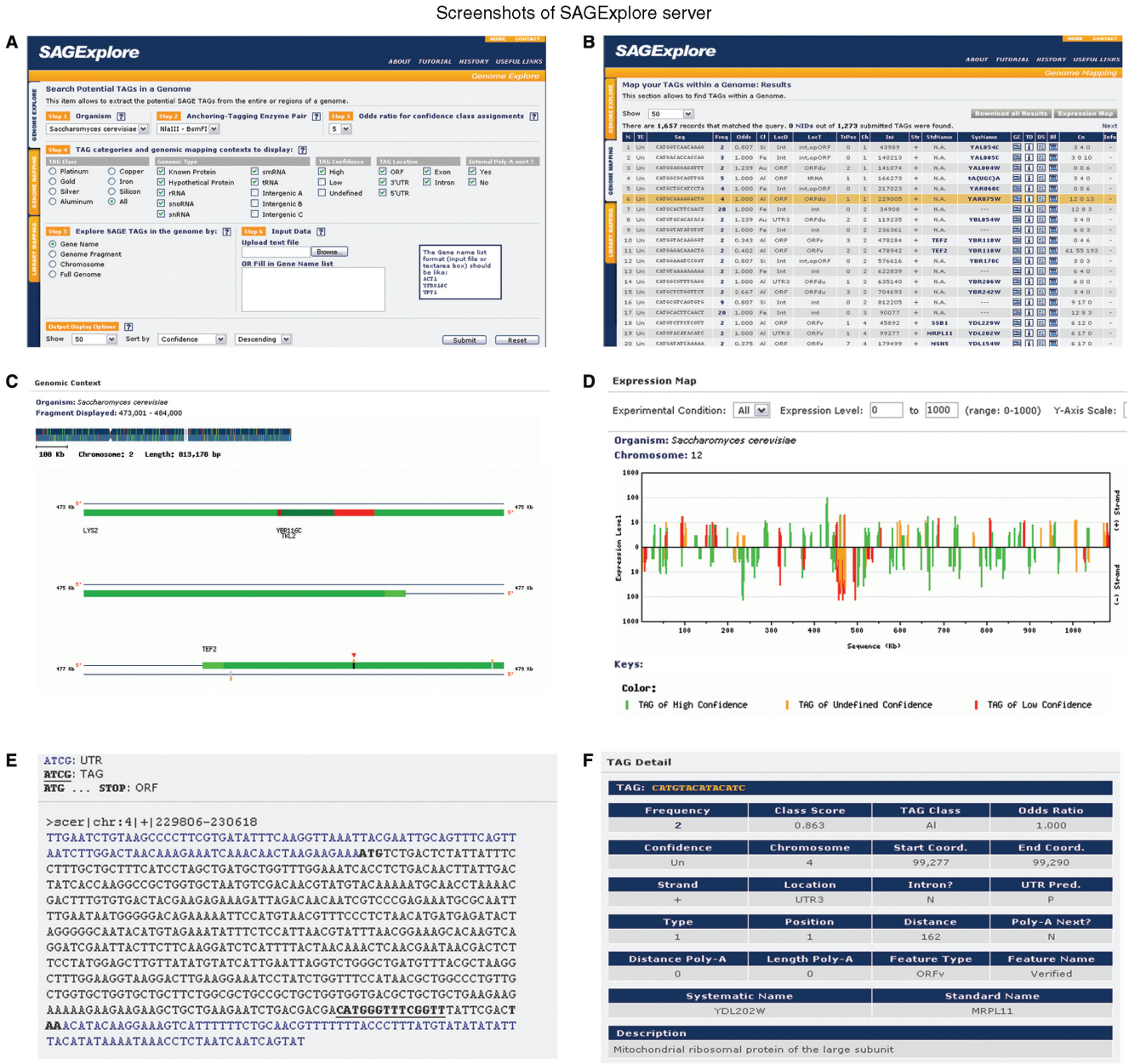

Figure 2. (A) Input form of genome explore module. (B) Output table of genome mapping module. (C) Graphic display of the genomic context for a selected tag. (D) Graphical genome expression map. (E) An example of the sequence details that are displayed for a particular gene where a given tag maps. (F) An example of the genomic details available for each tag.

Currently, we are building a database of virtual SAGE-tags for *Xenopus tropicalis* organism, which genome has been recently sequenced. This will constitute a larger challenge than that faced for yeast when implementing this server with the HGA methodology (2), since the 12 megabytes of yeast are not comparable to the 1.5 gigabytes for Xenopus (150 times larger). This constitutes an intermediate point with human, which should be the next genome to be incorporated into this server. We also plan in the near future to include the genomes of other model organisms such as *Mus musculus*, *Drosophila melanogaster* and *Arabidopsis thaliana*. The order or priority will depend on the user requests and feedback, but we are willing to help the SAGE community by providing useful tools to get the most of information out of these expensive large-scale experiments.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

This work was funded by grants 1050688, 1070357 and 1051112 from FONDECYT. Funding to pay the Open Access publication charges for this article was provided by grant 1051112 from FONDECYT.

Conflict of interest statement. None declared.

REFERENCES

1. Wahl,M., Heinzmann,U. and Imai,K. (2004) LongSAGE analysis significantly improves genome annotation: identifications of novel genes and alternative transcripts in the mouse. *Bioinformatics*, **21**, 1393–1400.
2. Malig,R., Varela,C., Agosin,E. and Melo,F. (2006) Accurate and unambiguous tag-to-gene mapping in SAGE by a hierarchical gene assignment procedure. *BMC Bioinformatics*, **7**, 487–507.
3. Boheler,K.R. and Stern,M.D. (2003) The new role of SAGE in gene discovery. *Trends Biotechnol*, **21**, 55–57.
4. Harbers,M. and Carninci,P. (2005) Tag-based approaches for transcriptome research and genome annotation. *Nat. Methods*, **7**, 495–502.
5. 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**, 1–4.
6. Velculescu,V.E., Zhang,L., Zhou,W., Vogelstein,J., Basrai,M.A., Bassett,, D.E.,Jr, Hieter,P., Vogelstein,B. and Kinzler,K.W. (1997) Characterization of the yeast transcriptome. *Cell*, **88**, 243–251.
7. Kal,A.J., van Zonneveld,A.J., Benes,V., van den Berg,M., Koerkamp,M.G., Albermann,K., Strack,N., Ruijter,J.M., Richter,A. *et al*. (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.
8. Varela,C., Cardenas,J., Melo,F. and Agosin,E. (2005) Quantitative analysis of wine yeast gene expression profiles under winemaking conditions. *Yeast*, **22**, 369–383.
9. Bala,P., Georgantas,R.W., Sudhir,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.
10. Pylouster,J., Sénamaud-Beaufort,C. and Saison-Behmoaras,T.E. (2005) WEBSAGE: a web tool for visual analysis of differentially expressed human SAGE tags. *Nucleic Acids Res.*, **33**, 693–695.
11. 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.
12. Liang,P. (2002) SAGE Genie: a suite with panoramic view of gene expression. *Proc. Natl Acad. Sci. USA*, **99**, 11547–11548.
13. Galante,P.A.F., Trimarchi,J., Cepko,C.L., de Souza,S.J., Ohno-Machado,L. and Kuo,W.P. (2007) Automatic Correspondence of Tags and Genes (ACTG): a tool for the analysis of SAGE, MPSS, and SBS Data. *Bioinformatics*, **23**, 903–905.
14. Kasai,Y., Hashimoto,S., Yamada,T., Sese,J., Sugano,S., Matsushima,K. and Morishita,S. (2005) 5'SAGE: 5'-end Serial Analysis of Gene Expression database. *Nucleic Acids Res.*, **33**, 550–552.
15. Divina,P. and Forejt,J. (2004) The Mouse SAGE Site: database of public mouse SAGE libraries. *Nucleic Acids Res.*, **32**, 482–483.
16. Robertson,N., Oveisi-Fordorei,M., Zuyderduyn,S.D., Varhol,R.J., Fjell,C., Marra,M., Jones,S. and Siddiqui,A. (2007) DiscoverySpace: an interactive data analysis application. *Genome Biol.*, **8**, 6–18.
17. Keime,C., Damiola,F., Mouchiroud,D., Duret,L. and Gandrillon,O. (2004) Identitag, a relational database for SAGE tag identification and interspecies comparison of SAGE libraries. *BMC Bioinformatics*, **5**, 143–154.
18. 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. *et al*. (2000) USAGE: a web-based approach towards the analysis of SAGE data. Serial Analysis of Gene Expression. *Bioinformatics*, **16**, 899–905.