D482–D483 *Nucleic Acids Research, 2004, Vol. 32, Database issue*
DOI: 10.1093/nar/gkh058

The Mouse SAGE Site: database of public mouse SAGE libraries

Petr Divina and Jiří Forejt*

Centre for Integrated Genomics, Institute of Molecular Genetics, Academy of Sciences of the Czech Republic,
Vídeňská 1083, CZ-142 20, Prague 4, Czech Republic

Received August 14, 2003; Revised and Accepted September 24, 2003

ABSTRACT

The Mouse SAGE Site is a web-based database of all available public libraries generated by the Serial Analysis of Gene Expression (SAGE) from various mouse tissues and cell lines. The database contains mouse SAGE libraries organized in a uniform way and provides web-based tools for browsing, comparing and searching SAGE data with reliable tag-to-gene identification. A modified approach based on the SAGEmap database is used for reliable tag identification. The Mouse SAGE Site is maintained on an ongoing basis at the Institute of Molecular Genetics, Academy of Sciences of the Czech Republic and is accessible at the internet address http://mouse.biomed.cas.cz/sage/.

INTRODUCTION

Serial analysis of gene expression (SAGE) is a well-established technique for gene expression profiling (1). SAGE uses short nucleotide tags (10 bp) from the defined position in the transcripts for the identification of expressed genes. The ligation of the tags into long concatemers and their sequencing results in the qualitative and quantitative gene expression profile of a particular tissue.

The main benefits of SAGE include the digital output and the identification of novel genes. The digital output allows direct comparisons of SAGE libraries constructed in different laboratories as long as the anchoring enzyme used in construction of the library is the same. Various tests have been proposed to distinguish significantly different frequencies of tags between SAGE libraries (2–5). The identification of SAGE tags is dependent on the information stored in sequence databases. The SAGEmap database (6) is the commonly used resource for the assigment of tags to transcriptional clusters in the UniGene database (7). Tags without associations to known genes can be further analysed to discover new genes (8).

The SAGE data are usually presented on the web pages of individual laboratories or shared in the Gene Expression Omnibus (GEO) public repository (9), which serves as a central distribution hub of public expression data generated by high-throughput techniques such as microarrays and SAGE.

An excellent website known as SAGE Genie (10) was created as part of the Cancer Genome Anatomy Project. This database contains data from more than 150 human SAGE libraries, predominantly from normal and cancer tissues. Several web-based tools are available for visualization, searching and analysis of human SAGE data. SAGE Genie uses a sophisticated approach for tag-to-gene identification based on the confident tag list and the ranking of sequence databases.

Here we present the Mouse SAGE Site—the database of SAGE libraries generated from various mouse tissues and cell lines that have been publicly available to date and were constructed using the NlaIII anchoring enzyme.

ORGANIZATION OF THE DATABASE

Database construction

The database collection currently consists of 56 publicly available mouse SAGE libraries and is continuously updated. Forty-one libraries were obtained from the GEO repository (9); an additional 15 libraries were added from individual laboratories that published their libraries on the Internet or in their publications. The total of 2 150 000 tags are stored in the database at present. An up-to-date list of the assembled SAGE libraries with a reference to their source is available on the web page http://mouse.biomed.cas.cz/sage/content.

All the SAGE libraries were organized and data processed in a uniform way. The libraries were annotated with information about the tissue origin, tissue histology or pathology status, source type (bulk, cell line, cell culture) and with further information about their construction. Each library was labelled with a unique name best describing the origin and status of the tissue. The preparation of actual SAGE data included removal of the linker-derived tags and all potential 1 bp linker variations. The SAGE library size was then constituted as the total number of tags excluding linker impurities.

A modified approach based on the SAGEmap database (6) was used for reliable tag-to-gene identification. The full list of tags extracted from mRNA and EST sequences is provided as part of the SAGEmap database and is available from the internet address ftp://ftp.ncbi.nlm.nih.gov/pub/sage/map/. This list includes tags extracted from the sequences in the Reference Sequence Project (RefSeq), the Mammalian Gene Collection (MGC), and the GenBank and dbEST databases

*To whom correspondence should be addressed. Tel: +420 24447 2273; Fax: +420 29644 2154; Email: jforejt@biomed.cas.cz

(11–14). In SAGEmap, according to this list, tag-to-gene associations are classified by a reliability score and the tag-to-gene associations with the top two reliability scores are considered reliable. In the Mouse SAGE Site, tag-to-gene associations supported by at least one mRNA sequence from RefSeq, MGC, GenBank or at least three ESTs with a poly(A) signal or eight ESTs with no poly(A) signal were considered as reliable and used for tag identification (see Supplementary Material for detailed information). The tags with reliable associations to 12 or more UniGene clusters were labelled as 'repetitive/low-complexity' to be easily distinguished. All possible tags with associations to the mitochondrial genome were extracted from the mouse mitochondrion genome sequence, accession no. J01420, and labelled as 'mitochondrial'.

The Z-test algorithm described previously (4) was implemented for pairwise comparisons of tag frequencies between SAGE libraries. Performing a lot of pairwise comparisons leads to an accumulation of Type I errors and increases the chance of detecting false positives in significantly different tags. To resolve this issue, the Benjamini–Hochberg correction of false discovery rate (15) was applied.

The database was constructed to allow easy updating of supporting databases and the addition of new public mouse SAGE libraries.

Database description

The Mouse SAGE Site is accessible without restrictions via the world wide web at the address http://mouse.biomed.cas.cz/sage/. The database aims to provide mouse geneticists with easy-to-use web-based tools for exploiting mouse SAGE data.

The tools Browse, Compare and Search are currently available for the SAGE data. Users can browse the content of each SAGE library with reliable tag identification to UniGene clusters and filter the list by several criteria including tag sequence, UniGene cluster, gene symbol, chromosomal location, LocusLink and MGI accession. Separate lists of tags with matches to the mitochondrial genome, repetitive tags and tags with unreliable matches are provided for each SAGE library. The Compare tool allows users to set up two pools of SAGE libraries and display differentially expressed genes at the selected significance level and specified fold factor. The data of all SAGE libraries can be searched by similar criteria to those for the Browse tool. The Search output shows the normalized tag count distribution (tags per million) across all SAGE libraries. The results from the Compare and Search tools can be exported into tab-delimited text format for further analysis by the user. All these tools use the modified approach for reliable tag identification described above and provide direct links from gene identifiers to external databases—UniGene, LocusLink and the Mouse Genome Database (16). Online documentation explains the features of each tool in more detail.

The Mouse SAGE Site is updated as soon as new builds of SAGEmap and UniGene databases are released and new public SAGE libraries from the mouse are available. The site will be improved in accordance with the progress of tag-to-gene identification and requests from the scientific community.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online. The first part of the supplement explains the reliable tag identification approach used in the Mouse SAGE Site. Subsequent parts show sample outputs from the Compare and Search tools.

ACKNOWLEDGEMENTS

We thank J. Novotney and P. Strnad for comments on the manuscript. This work is supported by the project of the Czech Ministry of Education, Youth and Sports No. LN00A079, Centre for Integrated Genomics. J.F. is supported as an International Scholar of Howard Hughes Medical Institute.

REFERENCES

1. Velculescu,V.E., Zhang,L., Vogelstein,B. and Kinzler,K.W. (1995) Serial analysis of gene expression. *Science*, **270**, 484–487.
2. 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 in normal and cancer cells. *Science*, **276**, 1268–1272.
3. Ruijter,J.M., Van Kampen,A.H. and Baas,F. (2002) Statistical evaluation of SAGE libraries: consequences for experimental design. *Physiol. Genomics*, **11**, 37–44.
4. 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., Dujon,B. *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.
5. Audic,S. and Claverie,J.M. (1997) The significance of digital gene expression profiles. *Genome Res.*, **7**, 986–995.
6. 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.
7. Wheeler,D.L., Church,D.M., Federhen,S., Lash,A.E., Madden,T.L., Pontius,J.U., Schuler,G.D., Schriml,L.M., Sequeira,E., Tatusova,T.A. *et al.* (2003) Database resources of the National Center for Biotechnology. *Nucleic Acids Res.*, **31**, 28–33.
8. 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.
9. Edgar,R., Domrachev,M. and Lash,A.E. (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.*, **30**, 207–210.
10. 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.
11. Boguski,M.S., Lowe,T.M. and Tolstoshev,C.M. (1993) dbEST—database for 'expressed sequence tags'. *Nature Genet.*, **4**, 332–333.
12. Strausberg,R.L., Feingold,E.A., Grouse,L.H., Derge,J.G., Klausner,R.D., Collins,F.S., Wagner,L., Shenmen,C.M., Schuler,G.D., Altschul,S.F. *et al.* (2002) Generation and initial analysis of more than 15 000 full-length human and mouse cDNA sequences. *Proc. Natl Acad. Sci. USA*, **99**, 16899–16903.
13. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2003) GenBank. *Nucleic Acids Res.*, **31**, 23–27.
14. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2003) NCBI Reference Sequence Project: update and current status. *Nucleic Acids Res.*, **31**, 34–37.
15. Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate—a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B*, **57**, 289–300.
16. Blake,J.A., Richardson,J.E., Bult,C.J., Kadin,J.A. and Eppig,J.T. (2003) MGD: the Mouse Genome Database. *Nucleic Acids Res.*, **31**, 193–195.