

eSAGE: managing and analysing data generated with Serial Analysis of Gene Expression (SAGE)

Elliott H. Margulies¹ and Jeffrey W. Innis^{1,2}

¹Departments of Human Genetics and ²Pediatrics & Communicable Diseases, University of Michigan Medical School Ann Arbor, Michigan, 48109-0618, USA

Received on December 13, 1999; revised on March 7, 2000; accepted on March 13, 2000

Abstract

Summary: eSAGE is a comprehensive set of software tools for managing and analysing data generated with Serial Analysis of Gene Expression (SAGE).

Availability: eSAGE is freely available for non-commercial use.

Contact: ehm@umich.edu

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 (Velculescu *et al.*, 1995). Computer software is required for efficient management and analysis of SAGE-generated sequence data, which can be quite substantial. Currently available software (SAGE300 v3.07, available from K.W.Kinzler by request) is suitable for parsing raw SAGE data. However, SAGE300 cannot be used with sequence files containing letters of the IUPAC code beyond A, C, G, T, or N. In addition, SAGE300 does not take advantage of the UniGene SAGE Tag mapping flatfiles for automatic gene identification of SAGE Tags. Furthermore, SAGE300 stores SAGE Tag data in a proprietary format that must be exported before any further analysis can be performed. Finally, there are no tools available for readily following the efficiency of SAGE sequencing or confirming SAGE Tag identities. 'eSAGE' was written to circumvent these problems and to streamline the analysis of SAGE data.

Description

SAGE Tag extraction

A flowchart showing how eSAGE manages SAGE data is presented in Figure 1. Initially, the software is used to create an eSAGE Tag database for each sequenced SAGE library. The eSAGE Tag database contains three data tables that are used to store Tag data, DiTag data, and information on each extracted sequence file. An additional SAGE library information table contains the user's choice of anchoring enzyme sequence, SAGE Tag length and

maximum allowable DiTag length selected for the eSAGE Tag database.

Input DNA sequence files can contain any characters from the standard IUPAC code. In addition to this standard ASCII text file format (*.seq), eSAGE reads PHD files (*.phd.1) generated from *phred*-analyzed sequence trace data (Ewing *et al.*, 1998). eSAGE uses the *phred* quality values for each base as a more accurate method of excluding low quality sequence data from the analysis. After selecting input sequence files, DiTags of the appropriate length are extracted into the DiTag table. Then, SAGE Tags containing only unambiguous bases (A, C, G or T) are extracted from each unique DiTag and added to the Tag table, which stores all SAGE Tag sequences and their corresponding frequency.

Data management

All databases generated and used by eSAGE are formatted for use with Access 97 (Microsoft Corp.), which is required to view, sort and output the data. In an effort to determine the efficiency of SAGE and simplify the task of managing sequences produced from hundreds of clones, we designed eSAGE to store several facts about each sequence file entered into the database including but not limited to the sequence filename and number of extracted SAGE Tags. In addition, information from the DiTag table is used to prevent extracting SAGE Tags from duplicate DiTags, which occurs from either the joining of high abundance SAGE Tags (Welle *et al.*, 1999), PCR bias (Velculescu *et al.*, 1995) or from entering a duplicate sequence into the database. eSAGE displays a warning message box if a clone is found to contain more than three duplicate DiTags, alerting the user to a potential duplicate sequence file.

Comparing two SAGE Tag databases

A Compare table is generated with eSAGE by joining two Tag tables together from user-selected eSAGE Tag databases. The software removes linker sequences from the Compare table, and calculates summary statistics useful for assessing the quality of the SAGE library. In

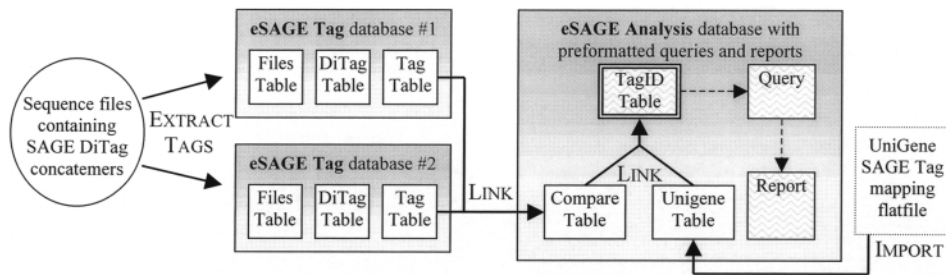


Fig. 1. Outline of eSAGE Data Management.

addition to importing the SAGE Tag sequence and its corresponding frequency from each library, the Compare table also contains two calculated fields that our laboratory has found useful in ranking differentially expressed SAGE Tags for further analysis: (i) fold-difference in abundance of a particular SAGE Tag; (ii) the calculated *p*-value based on the test statistic developed by Audic and Claverie (1997).

Tag identification

The software can be used to import SAGE Tag to UniGene mapping flatfiles generated from human, mouse, or rat UniGene clusters into an Access 97 table. These data sets are a sophisticated way of determining which UniGene cluster corresponds to a given SAGE Tag sequence and are freely available by anonymous FTP (<http://www.ncbi.nlm.nih.gov/SAGE/>). Once imported, eSAGE links the UniGene table to the user's Compare table and creates a single, informative TagID table. eSAGE includes a utility for verifying SAGE Tag to UniGene cluster mapping. This utility searches either a GenBank sequence file or a file containing a list of FASTA-formatted sequences downloaded from the relevant UniGene cluster for poly(A) signals and corresponding SAGE Tag sequences in both orientations.

Additional utilities

eSAGE provides several other useful data analysis tools. An eSAGE Analysis database containing several predefined queries and reports can be automatically linked to the user's data when the eSAGE Analysis database is used to store the Compare, UniGene and TagID tables generated by the software. The predefined queries and reports display SAGE Tags with greater than 5-fold difference in abundance and/or a *p*-value ≤ 0.05 (Audic and Claverie, 1997). Users can also modify pre-existing queries and reports, or create their own, all of which are automatically updated when a TagID table is generated with new data. There also is a utility that finds all DiTags containing a specific SAGE Tag sequence. This allows users to rapidly

identify additional bases from 'unidentified' SAGE Tags in order to generate synthetic primers for RT-PCR (van den Berg *et al.*, 1999; Chen *et al.*, 2000). Finally, there is a utility that can list all sequence files that contain a particular SAGE Tag sequence, useful for analysing Tag-specific chromatographs.

Summary

eSAGE was written in Visual Basic v6.0 (Microsoft Corp.) and is compatible with the Windows 95/98 and NT v4.0 operating systems (Microsoft Corp.). This software has enhanced our ability to analyze and present SAGE data and should also be useful to other laboratories performing SAGE.

Acknowledgements

E.H.M. is supported by the Institutional Training Program in Genomic Science (T32 HG00040). We thank Dr Kinzler and Dr Velculescu for providing us with their SAGE Analysis software and helpful discussions about the SAGE protocol and data analysis. We also thank Dr Lash for the SAGE Tag to UniGene cluster mapping flatfiles. This work was supported in part by a grant from NIH (HD34059).

References

- Audic, S. and Claverie, J.M. (1997) The significance of digital gene expression profiles. *Genome Res.*, **7**, 986–995.
- Chen, J.J., Rowley, J.D. and Wang, S.M. (2000) Generation of longer cDNA fragments from serial analysis of gene expression tags for gene identification. *Proc. Natl. Acad. Sci. USA*, **97**, 349–353.
- Ewing, B., Hiller, L., Wendl, M.C. and Green, P. (1998) Base-calling of automated sequencer traces using *phred*. I. Accuracy assessment. *Genome Res.*, **8**, 175–185.
- Velculescu, V.E., Zhang, L., Vogelstein, B. and Kinzler, K.W. (1995) Serial analysis of gene-expression. *Science*, **270**, 484–487.
- Welle, S., Bhatt, K. and Thornton, C.A. (1999) Inventory of high-abundance mRNAs in skeletal muscle of normal men. *Genome Res.*, **9**, 506–513.
- van den Berg, A., van der Leij, J. and Poppema, S. (1999) Serial analysis of gene expression: rapid RT-PCR analysis of unknown SAGE tags. *Nucleic Acids Res.*, **27**, e17.