1 INTRODUCTION

This manual is intended to walk you through the functions of eSAGE and show you the logic behind them. If you find any software bugs or need additional information about eSAGE, please contact Elliott Margulies at elliott@nhgri.nih.gov. Thank you for trying eSAGE.

2 New in version 1.2

- Since version 1.10, eSAGE can calculate the GC content of your SAGE Tags. This function can now be accessed by clicking on File –> Calculate GC Content. This is described in section 3.4 on page 4.

- Improved Tag-to-UniGene cluster matching can now be facilitated with the added option to import flatfiles generated from the ehm-Tag-Mapping algorithm (described in Margulies EH, Kardia SLR, and Innis JW (2001) *Genome Research* 11:1686-1698. For more information about the ehm-Tag-Mapping method, see the links from `http://genome.nhgri.nih.gov/eSAGE`.

3 File

3.1 Open eSAGE Tag Database

Opens an eSAGE Tag Database for input. The selected database must have been created by eSAGE and should *not* have been modified by Access 2000. After selecting the database, the file path of the database and summary statistics are displayed in the upper left and right and corners of the window respectively.

Use the drive and directory tree to move where sequence files are stored. The following file formats are acceptable:

- ASCII text files with a *.seq extension containing any characters from the standard IUPAC code.

- PHD files (*.phd.1) generated from *Phred*-analyzed sequence trace data. *Phred* quality values are used as a more accurate method of excluding low quality sequence data. eSAGE will treat any base call with a quality value < 20 as an N. However, eSAGE keeps a “CATG” if the average quality value for the 4 bases is > 18.

- A useful utility for renaming multiple file extensions at once can be found at:

  `http://www.zdnet.com/pcmag/pctech/content/17/11/ut1711.001.html`

Once the directory containing sequence files has been opened, use the shift and control keys with the mouse to select multiple sequence files. Pressing “Extract Tags” will perform the following tag extraction algorithm on each selected sequence file in the list:

1. If “Limit Tags” is checked, eSAGE will determine if the Total number of Tags in the database has exceeded the number entered in the “Limit Tags” box. If Yes, eSAGE does nothing and moves to the next sequence file. If No, proceed to 2.

2. Extracts each DiTag by searching for “CATG”.

(a) If the DiTag length is between (TagLength value)*2 and the maximum DiTag length value and the DiTag sequence is unique in the DiTag table, it is added as a new record with the Frequency field = 1 and the Extracted field = False. If it is NOT unique, it is noted in the Frequency field of the matching DiTag record.

(b) If the DiTag length is not of an appropriate length to be extracted, it is noted in the SmallDiTags or LargeDiTags field, as appropriate, and not added to the DiTag table.

3. For each DiTag record where the Extracted field = False, eSAGE extracts the two SAGE Tags from the DiTag sequence, then changes the Extracted field to True. If the SAGE Tag sequence contains only unambiguous bases (A, C, G, or T), it is noted in the GoodTags field of the Files table and added to the Tag table. If a SAGE Tag does contain an ambiguous base, it is not added to the Tag table.

4. A new record in the Files table is created containing the information on the extracted sequence file.

5. The file extension is changed from *.seq to *.anl (or *.phd.1 to *.anlp) and eSAGE moves to the next sequence file in the list.

3.2 Create New eSAGE Tag Database

An eSAGE Tag database contains three tables that are used to store Tag data, DiTag data and information on each extracted sequence file. A fourth LibInfo table stores information about the SAGE library obtained from the user when the SAGE library is created. See Table 1 on the next page for an outline of the fields and their descriptions in the four tables.

3.3 Compare 2 eSAGE Tag Databases

You must select two eSAGE Tag databases to compare and a pre-existing database (such as the eSAGE Analysis database with predefined queries and reports) to store the output Compare table. Linker sequences to be removed have been preset for those derived from using the suggested linker sequences in the SAGE protocol v1.00c. Pressing "Compare SAGE Databases" will combine the Tag tables from the two selected eSAGE databases and put the result in the Compare table. Two additional calculated fields in the Compare table are as follows:

1. Fold-difference in abundance of a particular SAGE Tag. 0's are treated as 1's to avoid division by 0. The negative inverse of the fold-difference is presented if the resulting number is less than one, representing SAGE Tags the are more abundant in database #2. These numbers are rounded to one digit after the decimal.

2. A calculated p-value for differential expression based on the test statistic developed by Audic and Claverie (1997) *Genome Research* 7:986-95.

3.3.1 GC Content command buttons:

These buttons are no longer valid. Please use the function described in Section 3.4 on page 4.

Field Name	Description
Tags	
Frequency:	Number of times a SAGE Tag sequence has been extracted.
Tag:	10 bp SAGE Tag sequence extracted from the DiTag.
DiTags	
Frequency:	Number of times a DiTag sequence has been extracted.
DiTag:	20 to 26 bp DNA sequence found between two *Nla*III sites (CATG).
Extracted:	Boolean field indicating whether SAGE Tags have been extracted from the DiTag.
Files	
FileName:	Name of extracted sequence file.
SeqLength:	Length in bp of sequence file.
NumOfDiTags:	Number of extracted DiTags of appropriate length.
SmallDiTags:	Number of DiTags smaller than (TagLength value)×2.
LargeDiTags:	Number of DiTags greater than the maximum DiTag length value.
DuplicateDiTags:	Number of DiTags matching a DiTag sequence already present in the DiTag table.
GoodTags:	Number of extracted SAGE Tags containing only unambiguous bases.
NewTags:	Proportion of extracted SAGE Tags not previously entered into the Tag table.
LibInfo	
AnchorEnzyme:	4 bp sequence (Usually *Nla*III) used to mark DiTag boundries.
TagLength:	Length of Tag sequence to be extracted.
MaxDiTagLength:	Maximum allowable DiTag length to be extracted.

Bold names denote the four different tables in an eSAGE Tag database. The fields that comprise each table are listed below the name. A description of the field is also presented to the right of the field name.

Table 1: Outline of data structure in eSAGE

3.4 Calculate GC Content

1. Use the 'Select Database' button to select the eSAGE Tag database for which you would like to calculate GC content.

2. After appropriately selecting a valid database, the path of the database will be displayed to the right of the button.

3. Click on the 'Calculate GC Content' button. As the GC content is calculated for each unique SAGE Tag sequence, the display will count through the total number on unique SAGE Tags in the database. A new table called 'GCcontent' is created in the database.

4. eSAGE also creates a query ('TagsSortedByGC') that sorts all Tags by GC content. Open the database in Microsoft Access and select the Queries tab to view this query. The data in this query can be pasted into Microsoft Excel (or your favorite graphing program) to plot Tags *vs.* GC content. Use the Excel spreadsheet supplied with eSAGE (`AverageGCcalculation.xls`) to determine the average GC content of the SAGE Tags in your library.

4 UniGene

4.1 Import ehm-Tag-Mapping flatfile

Select this option if you have generated or obtained an ehm-Tag-Mapping flatfile. The ehm-Tag-Mapping method is described in Margulies EH, Kardia SLR, and Innis JW (2001) *Genome Research* 11:1686-1698. Information about obtaining and implementing the ehm-Tag-Mapping method can be found at `http://genome.nhgri.nih.gov/eSAGE`.

1. Make sure the flatfile is in DOS text. If your FTP program does not convert from UNIX to DOS text, the converter found at the following web address is free and has been used for this purpose:
   `http://www.blunt1.demon.co.uk/dostou/index.html`

2. If a database to store the imported table does not exist, choose to create a new database. You can always copy the imported table into an existing database at a later time. Otherwise, select the eSAGE Analysis database or another database that is used to store your Compare table.

3. Select the mapping flatfile, which is typically named `ehmMappingFIT.txt`.

4. Press "Begin Importing Flatfile" to create a table named 'ehmMapping'. You can watch the progress of the import function by the number of records being entered into the UniGene table.

.

4.2 Link to ehm-Tag-Mapping

To link a Compare table to an imported ehm-Tag-Mapping, you must first import the mapping flatfile into a Microsoft Access 97 database. See Import UniGene SAGE Tag Mapping to do this (the previous section). Also, the Compare table and ehmMapping table must be in the same database.

After selecting the database containing the Compare and UniGene tables, you will see nothing, but eSAGE is working. A message box will be displayed when the link is complete. The resulting table is named “ehmTagID”.

4.3 Import UniGene SAGE Tag Mapping

Select this function to import a UniGene SAGE Tag mapping flatfile. These data files were developed by Alex Lash at NCBI. This import function must be performed before you link a Compare table to a UniGene mapping.

1. Download the desired UniGene SAGE Tag mapping flatfile. For Human mappings, go directly to `http://www.ncbi.nlm.nih.gov/SAGE/` for downloading instructions. For mouse and rat mappings, download by anonymous FTP at `ftp://ncbi.nlm.nih.gov`, initial remote host directory is `pub/alash`. You can only use the Tag to Gene mappings. Gene to Tag mappings cannot be imported at this time.

2. Make sure the flatfile is in DOS text. If your FTP program does not convert from UNIX to DOS text, the converter found at the following web address is free and has been used for this purpose:
   `http://www.blunt1.demon.co.uk/dostou/index.html`

3. You are now ready to import the UniGene mapping.

4. If a database to store the imported table does not exist, choose to create a new database. You can always copy the imported table into an existing database at a later time. Otherwise, select the eSAGE Analysis database or another database that is used to store your Compare table.

5. Select the mapping flatfile.

6. Press “Import” to create a table named UniGene. You can watch the progress of the import function by the number of records being entered into the UniGene table.

4.4 Link to UniGene Mapping

To link a Compare table to a UniGene SAGE Tag mapping, you must first import the mapping flatfile into a Microsoft Access 97 database. See Import UniGene SAGE Tag Mapping to do this. Also, the Compare table and UniGene table must be in the same database. After selecting the database containing the Compare and UniGene tables, you will see nothing, but eSAGE is working. A message box will be displayed when the link is complete. The resulting table is named “TagID”.

When the Compare and TagID tables are in the eSAGE Analysis database, the information will be automatically linked to the predefined queries and reports. You can also create customized queries and reports, or modify the preexisting ones, all of which will be automatically updated when a TagID table is generated with new data.

5 Tools

5.1 Find DiTags containing Tag

This allows you to rapidly identify additional bases from “unidentified” SAGE Tags in order to generate synthetic primers for RT-PCR.

1. Select the eSAGE database containing the DiTag table you want to search.
2. Enter the SAGE Tag sequence.
3. Press ″Find DiTags″. This will display a list of all DiTags containing the entered sequence. They will all be arranged with the entered sequence reading left to right. In other words, if your sequence matches the right SAGE Tag in a DiTag, the reverse complement of the DiTag is displayed.
4. The DiTags can be written to a text file by pressing ″Save DiTags to file″.

5.2 Search for Tag in Sequence Files

1. Select files (*.anl) as you would in Open eSAGE Tag Database.
2. Enter the Tag sequence to be searched.
3. Press “Find Files” and a list of sequence files containing the entered sequence or its reverse complement, will be displayed.

5.3 Find Tag in GenBank Entry

Use this function to automate the search for poly(A) signals and SAGE Tag sequences in GenBank entries. This is useful to find GenBank entries in the UniGene cluster containing the SAGE Tag sequence. You can also analyze a file containing a list of FASTA formatted sequences.

1. Save the GenBank entry (in GenBank view) as PC text to your computer.
2. Select this function and press ″Load Sequence″ to load the sequence into view.
3. If necessary press “Reverse comp” To search the reverse complement of this sequence. This is useful for many 3′ EST reads.
4. Press “Find poly(A)” to find all poly(A) signals in the sequence (`AATAAA` or `ATTAAA`). This is useful if the gene has alternative transcripts that can potentially produce several different SAGE Tags.
5. Select the poly(A) signal you want to use by clicking on that number with the mouse.
6. Press “Find Tag” to find the SAGE Tag representing the 11 bp sequence 3′ to the last *Nla*III site before the selected poly(A) signal (or before the end of the sequence if no poly(A) signal is selected).