Implementing the ehm-Tag-Mapping-Method

Elliott H. Margulies, Ph.D.
Genome Technology Branch
National Human Genome Research Institute
National Institutes of Health
50 South Dr., MSC 8004
Bldg. 50, Rm. 5527
Bethesda, MD 20892-8004

elliott@nhgri.nih.gov

Introduction

The ehm-Tag-Mapping method was developed as part of my doctoral thesis work in the Department of Human Genetics at the University of Michigan. A description of the method can be found in Margulies EH, Kardia SLR, and Innis JW (2001) *Genome Research* 11:1686-1698. Similar information, pertaining to the ehm-Tag-Mapping method only, is provided as an excerpt from my thesis in `MethodDetails.pdf`. The document you are currently reading deals with how to generate SAGE Tag to UniGene cluster matches using the ehm-Tag-Mapping method.

Please note that I am currently working in the Genome Technology Branch at the National Human Genome Research Institute and can be contacted by email.

Instructions for a Windows-based PC

1. Make sure you have a recent version of Perl installed on your PC. Perl is a freely available programming language that has been implemented in a variety of operating system platforms. A version of Perl for the Windows environment can be downloaded at:

   `http://aspn.activestate.com/ASPN/Downloads/ActivePerl/`

   Follow the instructions on that web site to properly install Perl on your computer.

2. Create a subdirectory to store the following files, which implement the ehm-Tag-Mapping method. They can be downloaded at:

   `http://research.nhgri.nih.gov/eSAGE/`

   (a) `ehmTagMappingPC.pl`[1] — the main Tag-Mapping program

   (b) `FindTags.pl` — a program called by `ehmTagMappingPC.pl` that must be in the same directory

   (c) `GetIDesc.pl` — a program that creates a data file needed by the `ehmTagMappingPC.pl` program

[1] NOTE: I have recently modified the Perl scripts to run in a Windows environment. If you would prefer to run these scripts in a Linux environment, use the `ehmTagMappingLINUX.pl` script instead. The other Perl scripts will run on both platforms.

3. In the same subdirectory, download the following UniGene flatfiles (which can be found at `ftp://ftp.ncbi.nih.gov/repository/UniGene/`

   (a) Xx.seq.all.gz[2]

   (b) Xx.data.gz[2]

4. Uncompress the UniGene flatfiles with an appropriate program such at WinZip (available at `http://www.winzip.com/download.htm`).

5. Double-click on the `GetIDesc.pl` script. You will be prompted to type the Xx.data filename. Hit the <Enter> key. After several minutes (depending on the speed of your CPU), you will be notified that the script has created a file called `UniGeneIDs.txt`. This file is needed for the next Perl script.

6. Double-click on the `ehmTagMappingPC.pl` script. You will be prompted to type the Xx.seq.all filename. Hit the <Enter> key and now after several hours (depending on the speed of your CPU), three files will be created:

   (a) **ehmMapping.log** This file is identical to the screen output during the run - useful for further investigating the logic behind any given Tag-UniGene cluster mapping).

   (b) **ehmMapping.txt** This is a tab-delimited file containing the resulting ehm-Tag-Mapping without FIT values (most people will not find this file useful).

   (c) **ehmMappingFIT.txt** This is a comma-delimited file containing the resulting ehm-Tag-Mapping with FIT values, which can be imported with eSAGE v1.2 and higher by selecting UniGene –> Import ehm-Tag-Mapping flatfile. A new ehmTagID table should then be generated with the UniGene –> link to ehm-Tag-Mapping command.

   The order of the fields in this file are as follows:

   SAGE Tag Sequence, UniGene cluster number, UniGene cluster description, NUM[3], SUM[3], COUNT[3], FIT[3]

[2]Xx denotes the particular species of interest (ie. Hs for human, Mm for mouse).

[3]For a description of these values, see Margulies EH, Kardia SLR, and Innis JW (2001) *Genome Research* 11:1686-1698. Also see MethodDescription.pdf, which is an excerpt of my thesis and also contains the detailed relevant information.