

GeneMerge

Version 1.2, April 2007

Cristian I. Castillo-Davis

Copyright © 2007 by Cristian I. Castillo-Davis. This software package is provided "as is" without warranty of any kind. In no event shall the author be held responsible for any damage resulting from the use of this software. The program package, including source code, executables, example data sets, and this documentation, is distributed free of charge under the terms of the GNU General Public License as published by the Free Software Foundation: Free Software Foundation, Inc., 51 Franklin St, Fifth Floor, Boston, MA 02110-1301 USA

Suggested citation:

Castillo-Davis, C. I. and D. L. Hartl. 2003. GeneMerge-- post-genomic analysis, data-mining and hypothesis testing. *Bioinformatics* 19(7):891-892

<http://www.oeb.harvard.edu/hartl/lab/publications/GeneMerge.html>

Cristian I. Castillo-Davis
Department of Biology
University of Maryland
College Park, MD 20742, USA
Email: castill0@umd.edu

Introduction

GeneMerge is a web-based and stand-alone program written in PERL that returns a range of functional and genomic data for a given set of study genes and provides statistical rank scores for over-representation of particular functions or categories in the dataset. Functional or categorical data of all kinds can be analyzed with GeneMerge. One common use for GeneMerge is the analysis of microarray data.

Getting Started

Macintosh OS X

- (1) Download **GeneMerge1.2-Mac.tar.gz** and double-click it to unpack
- (2) Open the GeneMerge Folder and double click the GeneMerge Icon
- (3) Follow the instructions. That's it!

Linux

- (1) Download and unpack the archive **GeneMerge1.2.tar.gz** by typing
`tar xvzf GeneMerge1.2.tar.gz`
- (2) Check to see where PERL is installed by typing `which perl` in a terminal. If you get the result `/usr/bin/perl` then you are good to go. If not, modify the very first line in the file `GeneMerge1.2.pl`

```
#!/usr/bin/perl -w
```

so that it has the correct path. For example, if perl is installed in `/usr/local/bin/perl` then the line should read: `#!/usr/local/bin/perl -w`

- (3) GeneMerge for Linux is run by the command line. See "Running the Program" below.

Windows 95/98/NT/XP

- (1) Download and install the programming language Perl for Windows. The latest version of *ActivePerl* is recommended. It's free and easy to install.
<http://www.activestate.com/Products/ActivePerl/>
- (2) Download and unzip the archive **GeneMerge1.2.zip** by double clicking on the folder.

(3) GeneMerge for Windows is run on the command line. To open up a "Command Prompt" window, go to *Start>Programs>Accessories>Command Prompt*.

To see what directory you are in, type `dir` and use `cd` to change directories. For example `cd GeneMerge`. Using these commands, move into the GeneMerge directory. You are now ready to run the program. See "Running the Program" below.

Overview

GeneMerge uses 4 input files:

1. Study set gene file
2. Population set gene file
3. Gene-association file
4. Description file

The study set is comprised of genes that are currently under investigation. The population set is comprised of those genes from which the study set was drawn, often all genes on a given microarray. The *gene-association* file links gene names with a particular datum of information using a shorthand identifier (ID). Finally, the *description file* contains human-readable descriptions of gene-association IDs.

Output is a tab-delimited text file that can be opened in most spreadsheet programs. It contains functional or categorical data associated with each gene in the study set and rank scores for over-represented functions/categories, as well as other pertinent data.

Example

Say you perform a microarray experiment and find that 473 genes are up-regulated in a mutant strain of yeast in comparison with the wild type and you'd like to make sense of this finding. The 473 genes comprise your study set. Since there are 6,188 genes on your microarray this is your population set. If you decide that you'd like to see what molecular functions these genes are involved in and if any are statistically over-represented, you would select the GO Molecular Function *gene-association file* for yeast (*S_cerevisiae.MF*) and the complimentary *description file* (*GO.MF.use*). You would then use the following files:

Study set file:	473.genes.txt	- list of up-regulated genes
Population set file:	6188.genes.txt	- list of genes on the array
Gene-association file:	S_cerevisiae.MF	- list of all genes and associated ID
Description file:	GO.MF.use	- IDs and their English descriptions

Running the program

GeneMerge runs on the command-line under Linux and Windows and as an application or on the command-line under Mac OS X. The command-line argument syntax is:

```
./GeneMerge1.2.pl gene-association.file description.file  
population.file study.file output.filename
```

Note: under Windows type `perl GeneMerge1.2.pl` etc. instead of the "dot slash."

To test run GeneMerge from the command line type the following:

```
./GeneMerge1.2.pl AssociationFiles/S_cerevisiae.MF  
DescriptionFiles/GO.MF.use Example/pop.txt  
Example/study.txt test.out
```

You can open the `test.out` file in a text editor (See Understanding the Output) and compare it with the `test.out` file in the Examples directory.

If you are using GeneMerge for OS X you can select gene-association categories from the pull-down menu and enter study and population genes either by pasting or by specifying local files. The appropriate description file is automatically detected. If you make your own gene-association and description files make sure they follow the recommended format so they are auto-detected (See below).

Understanding the output

Output is a tab-delimited text file that can be opened in a spreadsheet program like Excel either by cutting and pasting from a text editor or importing "as tab delimited." The output file lists each gene-association term found in the study set along with its English description, frequency in the population set, frequency in the study set, and statistical enrichment score-- uncorrected and corrected. Below is a breakdown of each column header.

GMRG_Term	GeneMerge term, for example a GO identifier "GO:0001234"
Pop_freq	fraction of genes in the population with this term
Pop_frac	fraction of genes in the population with this term (whole numbers)
Study_frac	fraction of genes in the study set with this term (whole numbers)
Raw_es	<i>P</i> -value
e-score	Bonferroni corrected <i>P</i> -value
Description	GeneMerge term's English description
Contributing_genes	All the genes that are associated with this term in the study set

The output file also lists the total number of population and study genes, the total number

of GeneMerge terms examined, and the number of genes that have terms associated with them. Genes with that have no gene-association data associated with them are listed as well. Finally the number of population non-singletons, *i.e.* the number of terms that contribute to the Bonferroni correction is also given.

How to make your own Gene Association files

Many structured text files for use with GeneMerge come with the package and more are available for download at:

<http://www.oeb.harvard.edu/hartl/lab/publications/GeneMerge.html>

However, it's easy to make your own gene-association files for use with GeneMerge. Use a text editor to make two files with the following formats (make sure you save as plain "text"):

Gene-association file format

```
genename tab functionID;  
genename tab functionID;  
genename tab functionID;functionID;
```

Description file format

```
functionID tab description_of_function  
functionID tab description_of_function  
functionID tab description_of_function
```

Here's an example of a Gene Association file for *Drosophila melanogaster*:

```
FBgn0000038      GO:0004889;  
FBgn0000039      GO:0004889;  
FBgn0000053      GO:0004637;GO:0004641;  
FBgn0000054      GO:0016252;  
FBgn0000055      GO:0004022;  
FBgn0000064      GO:0004332;  
FBgn0000120      GO:0016030;  
...
```

The FBgn numbers are *Flybase* gene names and the GO:XXXXXXXX terms are *Gene Ontology Consortium* (2000) IDs for specific functions. The white-space is a single tab. Each ID is followed by a semi-colon and if more than one ID is associated with a gene then these are separated by a semi-colon.

Here's an example of a Description file:

```
GO:0016505      apoptotic protease activator  
GO:0016504      protease activator  
GO:0008189      apoptosis inhibitor
```

GO:0005194 cell adhesion molecule
GO:0008014 calcium-dependent cell adhesion molecule
...

The ID terms here are *Gene Ontology* IDs for specific functions. The human-readable functional descriptions follow after a single tab. Note these lines do not have to end in semi-colons.

You can use a text editor and spreadsheet program to make these files. The following are instructions using Word and Excel on a Mac but similar steps should work on other platforms.

1. *Download a spreadsheet with the genomic data you are interested in*
2. *Open it in Excel*
3. *Organize the data into categories of your choosing if it's not already categorized. For example, you'll have to split continuous data into chunks.*
4. *Organize the data into categories so that there are two columns, one with gene names, the adjacent column with IDs*
5. *Copy and paste the two columns into Word using Paste Special --> "unformatted text"*
6. *Do a search and replace for the line ending to add semi-colons. Replace ^p with ;^p.*
7. *Save As plain "text"*

Description files can be made along the same lines, just skip step 6. If there are no IDs for your genomic data just make them up in Excel. A list of numbers works just fine, just make sure that each function/category gets a unique ID.

File Naming Conventions

You can name files whatever you'd like but the following convention insures that other people will be able to understand and use your files if you choose to make them available. Note if you are using the Mac OS X application the naming convention for description files should be followed if you want them to be auto-detected and appear in the pull-down menu.

Gene-association files are named using the first letter of the organism's genus, an underscore, the species name, a dot, and an abbreviation for the function/category. For example for a file listing the chromosomal location of each gene in human, the gene-association file could be called:

`H_sapiens.CHR`

The matching description file would then be called:

`H_sapiens.CHR.use`

* The one exception to this rule is if you work with gene-association descriptions that have been standardized. For example, GO terms are the same for all organisms and consequently, every organism need not have it's own description file. Thus, `S_cerevisiae.MF` and `D_melanogaster.MF` both use `GO.MF.use`.

Troubleshooting

No GMRG terms are found for any of my genes!

It is possible that there isn't any information for the particular genes you analyzed (especially if it is a small number) but more likely is that you are using gene names that are not the same as the ones in the gene-association file. Make sure you translate your gene names to those that are used in the gene-association data you want to use. Synonym tables of gene names are usually available from genome databases.

Some functions/categories have a *P*-value of "0" for over-representation! How is this possible?

If there are genes in your study set that do not appear in your population set then the probability that they can be drawn from the population is zero, and GeneMerge correctly reports this fact. Make sure that your study set is actually a subset of the population set. Stray spaces at the end of gene names in one file can sometimes cause this to happen even if the names look the same. A less likely scenario is that the *P*-value is so low it is out of range of the computer you are using and the value is truncated. I haven't seen this happen yet though.

Every description reads "couldn't find description for this term in description_file.use"

This will happen if you select the wrong description file (.use file) for a particular gene-association file. GeneMerge will look up the English descriptions for each ID and won't find any!

A few descriptions say "couldn't find description for this term in description_file.use"

This will happen if the English descriptions of some IDs in the gene-association file you are using were not found. This happens commonly when you update either a gene-association file or a description file without updating the other and there are slight mismatches. For example, the *Gene Ontology Consortium* constantly revises their IDs and some of them may be no longer used. If you use an older gene-association file with a new description file some terms might not be found. I try to post up-to-date files on the website: <http://www.oeb.harvard.edu/hartl/lab/publications/GeneMerge.html>

Release Notes

April 6, 2007

No changes to the software but new Gene Ontology gene-association files are provided for *A. thaliana*, *B. taurus*, *C. elegans*, *D. melanogaster*, *Danio rerio*, *G. gallus*, *H. sapiens*, *M. musculus*, *O. sativa*, *P. falciparum*, *R. norvegicus* and *S. cerevisiae*. Updated GO description files are also provided (GO.MF.use, GO.BP.use, GO.CC.use). Thus, if you are analyzing organisms using GO that haven't been updated, be sure to use the older GO.XX.use files in the folder "OLD" within the DescriptionFiles folder.

March 27, 2005

No changes to the software but new Gene Ontology gene-association files are provided for *A. thaliana*, *C. elegans*, *D. melanogaster*, *H. sapiens*, *M. musculus*, and *S. cerevisiae*. Updated GO description files are also provided (GO.MF.use, GO.BP.use, GO.CC.use). Thus, if you are analyzing organisms using GO that haven't been updated, be sure to use the older GO.XX.use files in the folder "OLD" within the DescriptionFiles folder.

References

Castillo-Davis, C. I. and D. L. Hartl. 2003. GeneMerge-- post-genomic analysis, data-mining and hypothesis testing. *Bioinformatics* 19(7):891-892

The Gene Ontology Consortium. Gene Ontology: tool for the unification of biology. 2000. *Nat. Genet.* 25: 25-29