

GeneMerge

Version 1.5

May, 2021

Cristian I. Castillo-Davis

Copyright © 2021 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. *Bioinformatics* 19(7):891-892

<http://genemerge.net>

Cristian I. Castillo-Davis
Washington, D.C., USA
Email: cristian@castillo-davis.com

- Table of Contents -

Topic	Page
1. Introduction	4
2. Getting Started	4
2.1 MacOS	4
2.2 Linux	5
2.3 Windows	5
3. Overview	6
3.1 Example	6
4. Running the program	6
4.1 Options	7
4.1.1 Custom False Discovery Rate	7
5. Tips on running the program	7
6. Understanding the output	7
7. How to make your own Gene Association files	8
7.1 Example <i>Gene Association</i> file	8
7.2 Example <i>Description</i> file	9
7.3 Using <i>Excel</i> to make Custom Annotation Files	9
8. Troubleshooting	10
9. Appendix	11
9.1 False Discovery Rate (FDR) methodology	11
9.2 Gene names & IDs	13
9.3 Association file naming conventions	13

- Table of Contents (continued) -

Topic	Page
10. References	14
11. Version information.....	14

1. Introduction

Given a set of genes, GeneMerge provides statistical rank scores for over-representation of gene attributes in the set and returns functional or categorical descriptions associated with each gene. It answers the question, "Is there functional or categorical 'enrichment' of some kind among this set of genes?" A common use for GeneMerge is the analysis of microarray and RNA-seq data and data generated from population genomic and evolutionary studies.

GeneMerge uses the hypergeometric distribution to calculate the over-representation of gene attributes and provides Bonferroni and False Discovery Rate (FDR) correction for multiple tests. Output is a tab delimited text file and an HTML file.

To facilitate analysis of gene-related data beyond Gene Ontology functions, GeneMerge uses a generic, easy-to-construct input file format to represent *gene-association* data. You are free to create your own gene-association files to test for enrichment of almost any discrete, gene or locus-based attribute. Examples include, chromosome location, pathway membership, genetic interactions, literature references, mutant phenotypes and more.

Methodology

GeneMerge uses the hypergeometric test statistic to calculate categorical over-representation because it provides exact *P*-values for over-representation (Castillo-Davis and Hartl 2003) and it is the correct test statistic for studies of all sizes.

Tools that use approximations to the hypergeometric distribution (normal, binomial, Poisson), the χ^2 test statistic, Z-scores, or unnormalized hypergeometric tests (using a static population set), do not return statistically correct results under some or all conditions (Huang, Sherman, and Lempicki, 2009).

For further statistical details see Castillo-Davis and Hartl (2003) and Section 9.1 of this manual: 'False Discovery Rate (FDR) Methodology'.

2. Getting Started

2.1 MacOS (Mac OS X)

- (1) Download the archive **GeneMerge1.5.tar.gz** to the Desktop and double-click the archive to unpack it.

- (2) Open the Terminal application on your Mac.
Applications>Utilities>Terminal

(3) To see what directory you are in, type `ls`. Use `cd` to *change to the directory* where GeneMerge is located. To move up a directory use `cd ..` ("dot dot")

```
cd Desktop/GeneMerge
```

(4) GeneMerge for MacOS is run on the command-line. See "Running the Program" below.

2.2 Linux / Unix

(1) Download and unpack the archive **GeneMerge1.5.tar.gz** by typing
`tar xvzf GeneMerge1.5.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 `GeneMerge1.5.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 on the command-line. See "Running the Program" below.

2.3 Windows

(1) Download and install the programming language *Perl* for Windows. *ActivePerl* has been tested, is free, and is recommended:
<https://www.activestate.com/products/perl/downloads/>

(2) Download and unzip the archive **GeneMergeWIN-1.5.zip** by right-clicking on the icon and selecting "Extract All..." from the menu.

(3) GeneMerge for Windows is run on the command-line. To open up a "Command Prompt" window, perform a 'Search' for "command prompt" or "cmd" then click on 'Open'.

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

3. 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 in a given genomic experiment or study. 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 and an HTML file which can be opened in any web browser (*Chrome, Edge, Firefox* etc.). The output 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 important data (see "Understanding the Output" for details).

3.1 Example Case

Say you perform an 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 6,157 genes detected total in your experiment, 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:	6157.genes.txt	- list of all genes in experiment
Gene-association file:	S_cerevisiae.MF	- genes and associated IDs
Description file:	GO.MF.use	- IDs and their English descriptions

4. Running the program

GeneMerge runs on the command-line. The command-line argument syntax is:

```
./GeneMerge1.5.pl gene-association description population  
study outputname
```

Note: in Windows type `perl GeneMergeWIN1.5.pl ...`

You can see a list of the required inputs at any time by typing `./GeneMerge1.5.pl` without any arguments.

To test run GeneMerge from the command-line type the following (in Windows use backslashes `\` instead of forward slashes and `perl GeneMergeWIN1.5.pl...`):

```
./GeneMerge1.5.pl Example/S_cerevisiae.MF Example/GO.MF.use
Example/pop.txt Example/study.txt mytest.txt
```

You can open the resulting `mytest.html` in a web browser and compare it with `test.out.html` in the *Example* directory. Or, open `mytest.txt` in a spreadsheet program and compare it with `test.out.xlsx` in the *Example* directory.

4.1 Options

4.1.1 Custom False Discovery Rate

By default GeneMerge calculates FDR = 1%, 5%, and 10%. If no custom FDR is specified, GeneMerge will also calculate FDR = 0.5%. To specify a custom FDR, use the syntax below. Remember to include the percent (%) symbol.

```
./GeneMerge1.5.pl gene-association description
population study outputname X%
```

5. Tips on running GeneMerge

Make sure your input files are correctly formatted. In particular, make sure the study set and population set do not contain any duplicate entries, *i.e.*, the same exact gene name on different lines. The former will result in an error. The latter will result in inexact population frequency calculations which will affect the results. Stray whitespace around a gene name or lack of a newline character in one file but not another can cause problems since the names will not 'match', resulting in an error. You can search for and replace hidden space and tab characters with 'nothing' in most text editors. Finally, make sure that all files end with an empty line.

6. Understanding the output

Output is provided in 1) 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 opening/importing "as tab delimited" and 2) an HTML file that can be opened in a web browser.

Both output files list each gene-association term found in the study set along with its human-readable description, frequency in the population set, frequency in the study set,

and statistical enrichment score (*P*-value)— uncorrected and corrected. Below is a breakdown of each column header.

GMRG_Term	GeneMerge term, for example a GO identifier "GO:0001234"
Pop_freq	frequency 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)
P	<i>P</i> -value for over-representation of this term in the study set
Bonf_Cor_P	Bonferroni corrected <i>P</i> -value
FDR_10	if the <i>P</i> -value is accepted at a FDR = 10% (True / False)
FDR_5	if the <i>P</i> -value is accepted at a FDR = 5% (True / False)
FDR_1	if the <i>P</i> -value is accepted at a FDR = 1% (True / False)
FDR_X_perc	if the <i>P</i> -value is accepted at a FDR = X% custom FDR (T / F)
Description	GeneMerge term's human-readable 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 no gene-association data associated with them are listed as well. The threshold *P*-value for each False Discover Rate (FDR) percentage is also reported. Finally the number of population non-singletons, *i.e.*, the number of terms that contribute to the Bonferroni correction is also given.

7. How to make your own Gene Association files

Many structured text files for use with GeneMerge come with the package and more may be available for download at <http://genemerge.net> 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 (' *tab* ' refers to the tab character) and 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
```

7.1 Example of a Gene Association file for *Drosophila melanogaster*:

```
FBgn0000038      GO:0004889;
FBgn0000053      GO:0004637;GO:0004641;
```

```
FBgn0000054      GO:0016252;
FBgn0000064      GO:0004332;
FBgn0000120      GO:0016030;GO:0004641;GO:0004637
...
```

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 they are separated by a semi-colon.

7.2 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.

7.3 Using *Excel* to make Custom Gene-Association and Description Files

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

1. Download a file 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 gene-association 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 "*Paragraph Mark*" with `";^p."` where `^p` is the symbol for *Paragraph Mark*.
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.

8. 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.

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 human-readable 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 human-readable 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 (or vice-versa) some terms might not be found.

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

If multiple copies of the same gene ID are present in your study set (but not population set) then it is possible for the study fraction to be larger than the population fraction. The probability that more genes can be drawn from the population than exist in the population is zero, and GeneMerge reports this result. Take care that your study set and population set do not contain any duplicate entries. Check the study and pop fractions in the output. If there are more terms in the study fraction than in the population fraction, you know you have a duplicate somewhere in the study set. Finally, check for stray whitespace or missing line endings in gene set input files (see Section 5 "Tips on Running the Program").

Another possibility is that the *P*-value is so low it is out of range of your computer (typically $P < \sim 1e-300$) and the value is returned as zero. Take a look at the study versus the population fractions. If you have extreme enrichment and no duplicates, it is likely the *P*-value is too low to calculate. I have yet to see this in real data, but it is possible.

The HTML output file won't open or crashes my web browser when I double-click it

This will happen if the HTML output file is too large (>5-30MB) for the browser or computer to render and can happen when working with well-annotated organisms with large genomes (for example, human). For such output files, you can open the GeneMerge *text output file* in a spreadsheet program like *Excel* using 'tab' as the separator. You can often 'drag-and-drop' this file straight onto the application icon and follow the dialog.

My analysis takes a long time (minutes) to run

This can happen when working with well-annotated species with large genomes where gene-association files are large (>5MB) with GMRG terms numbering in hundreds of thousands or even millions. Depending on your hardware, an analysis using the largest (20MB) 2 million-term gene-association file (H_sapiens.hBP) with 30,000 population genes and 5,000 study genes will take about 4 minutes. However, this is an extreme case. A 'typical' analysis usually takes less than 1 minute. But your study may not be typical. Please be patient!

GeneMerge gives an error message 'gene is not present in the population set', but it looks like it's there!

This can happen when there is whitespace (spaces, tabs, etc.) following the gene name in one file but not the other, for example "GENE_411" vs "GENE_411 ". Solution: remove the whitespace. You can search for and replace hidden space and tab characters with 'nothing' in most any text editor. Another possibility is that one of the gene names is located at the end of a file and does not have a *newline* (return) character after it, for example, "GENE_411\n" vs "GENE_411". Solution: make sure all files end with an empty line after the last gene.

Importing to spreadsheet, I get an error, "The data could not be loaded completely because the maximum number of columns per sheet was exceeded"

This can happen if the number of columns of GM output exceeds the maximum number of columns in your spreadsheet program (typically 1,024 or 16,000). For GeneMerge data, this means that some *Contributing_Genes* entries may not contain *all* the contributing genes for a given term (only the first 1,024 or 16,000 contributing gene names can be stored). However, all other GeneMerge data (terms, *P*-values, corrected *P*-values, FDR tags, descriptions, etc.) are unaffected. Click "OK" and the data will usually load-- just keep in mind not all contributing genes will be listed for all GMRG terms.

I found significant over-representation for the GO term "molecular function", "biological process", or "cellular component". What does this mean?

These GO 'root' terms are used to annotate genes with no known function. Over-representation of such a term represents enrichment for genes with no known function in the corresponding GO category (namespace).

9. Appendix

9.1 False Discovery Rate Methodology

Because GeneMerge assesses over-representation for all functional categories for a given set of genes, a correction is necessary to account for over-representation that will invariably occur by chance when multiple tests are carried out. A strict correction for multiple tests is the Bonferroni correction (Sokal and Rohlf, 1995) which adjusts *P*-values by multiplying each by the total number of tests carried out. This correction is conservative because it adjusts the *P*-value of each test perfectly back to the original

alpha level allowing zero false positives beyond what would be expected by chance.

A more flexible approach is provided by the False Discovery Rate (FDR) (Benjamani and Hochberg 1995), which allows an investigator to accept a certain percentage of false positives when the outcome of individual tests is of less interest than the overall test pattern, as is the case in many genomic applications. When carrying out multiple tests using the same hypothesis and test statistic, we increase the risk of false positives (Type I Error). Thus we are concerned with the number of tests that reject the null hypothesis correctly (true positives) relative to those that do so spuriously (false positives). Following Benjamani and Hochberg (1995), if V = number false positive tests and S = number of true positive tests, we can define Q as the proportion of false positives relative to the total number of positives,

$$Q = \frac{V}{V + S} \quad (1)$$

For example, if we can accept 20 false positive tests out of 200 positive results total, then $Q = 20/180+20 = 0.10$, or an FDR of 10%. Once an acceptable FDR has been chosen, it is applied to the data as follows: 1) order all test P -values from smallest to largest, indexed by $i = 1, 2, 3... m$, where m is the total number of tests, 2) find the largest i for which,

$$P_{(i)} \leq \frac{i}{m} q^* \quad (2)$$

where $P_{(i)}$ is the P -value associated with the i th ranking test and q^* is the FDR. P -values equal to or smaller than those of test i are flagged as being reliable at that particular FDR (Figure 1).

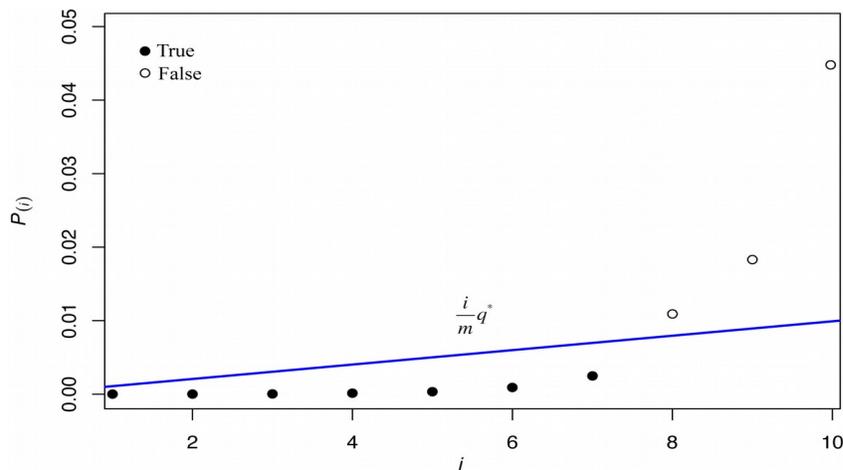


Figure 1.

This is the procedure implemented in GeneMerge v1.4 and v1.5. Each observed *P*-value in the output is tagged with a logical flag (True/False) for a given FDR threshold using this method. By default, GeneMerge returns FDR tagged output for FDR = 1%, 5%, 10%, and a custom FDR, if specified. The threshold *P*-value for each FDR percentage is also reported.

9.2 Gene Names and IDs

Gene-association data for each species use a particular gene name. An attempt is made to use the most stable and ubiquitous gene identifiers for each species for pre-packaged gene-association files. Of course, if you are making your own gene-association files you can use whatever gene name type you would like.

Note: Gene names used by particular databases/communities may change over time. Please check your gene-association data and gene list identifiers to make sure they are of the same type. Below are examples of gene identifiers that have been used in GO gene-association files for different species in the past. Note, sometimes there can be mixed types!

Species	Example	Gene Name Type	Notes
<i>A. thaliana</i>	Locus:2204237	TAIR Accession	-
<i>B. taurus</i>	F1MT32, Q0VBW6	UniProt ID	-
<i>C. elegans</i>	Q8IG65, WBGene00006868	UniProt ID, WormBase ID	Mixed ID types.
<i>C. lupus familiaris</i>	F1PG78	UniProt ID	-
<i>D. melanogaster</i>	FBgn0264560	FlyBase ID	-
<i>D. discoideum</i>	DDB_G0289041	DictyBase Gene ID	-
<i>D. rerio</i>	ZDB-GENE-071218-6	ZFIN ID	-
<i>E. coli</i>	P0CE47	UniProt ID	-
<i>G. gallus</i>	F1NXU9	UniProt ID	-
<i>H. sapiens</i>	Q9H9D4	UniProt ID	-
<i>M. musculus</i>	MGI:1099438	MGI ID	-
<i>O. sativa</i>	Q8S5I0	UniProt ID	-
<i>P. falciparum</i>	PF3D7_1117800	GeneDB Systematic Name	?
<i>R. norvegicus</i>	1561333, F1LW88	RDG ID, UniProt	Mixed ID types.
<i>S. cerevisiae</i>	S000000331	SGD ID	-
<i>S. pombe</i>	SPAC22H12.05c	PomBase Systematic ID	-
<i>S. scrofa</i>	F1RVT9	UniProt ID	-

Example gene names used by GO

9.3 Gene-Association & Description File Names

In order to easily identify gene-association and description files we use the convention that files indicate the genus and species plus an abbreviated description of the type of association in all capital letters. For example, for *Homo sapiens* and the gene-association "chromosome location" we write:

H_sapiens.CHR (gene-association file)

Corresponding description files are appended with a ".use":

H_sapiens.CHR.use (description file)

In the case where multiple gene-association files use the same *description file*, as is the case with GO annotations, we do not include a species name but still use the .use suffix. For example: **GO.MF.use**

For *multi-species* gene-association data, we use an **X** and the genus name to indicate it is a 'cross-species' gene-association file. For example, the GO 'Molecular Function' gene-association file for the *Aspergillus Genome Database* which includes multiple species, we write:

X_Aspergillus.MF (a multi-species gene-association file)

10. References

Benjamini, Y. and Hochberg, Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B. (Methodological)*. 57: 289-300.

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

Huang, D. W., Sherman, B. T. and Lempicki, R. A. 2009. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research*. 37(1): 1-13.

Sokal, R.R. and Rohlf, F.J. 1995. *Biometry: the principles and practice of statistics in biological research*. 3rd edition. W. H. Freeman and Co., New York.

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

11. Version notes

Version	Date	Notes
0.9	2001	development version
1.0	2002	used in Ranz, Castillo-Davis, Meiklejohn, & Hartl. 2003. <i>Science</i>
1.1	2002	unreleased, \b bugfix
1.2	2003	official release, Castillo-Davis & Hartl 2003. <i>Bioinformatics</i>
1.3	2009	unreleased, added FDR, HTML output, sprintf numbers, 'leapfrog' function, input file error checking
1.4	2015	official release, added threshold FDR P -values, pop frequency optimization, sprintf and contrib. genes bugfix, check on P -value underflow, removed leapfrog, more error checking, error help
1.5	2021	official release, term count optimization