SAM-GS

Significance Analysis of Microarray for Gene Sets

He Gao, Stacey Fisher, Qi Liu, Sachin Bharadwaj, Shahab Jabbari, Irina Dinu, Yutaka Yasui

Last modified on September 5, 2012

Table of Contents

1 Introduction

SAM-GS (Significance Analysis of Microarray for Gene Sets) is a statistical technique for assessing the associations of gene expression in *a-priori* defined gene sets, or biological pathways with a binary phenotype in microarray experiments. It was proposed by Dinu *et al.* (2007) as an alternative to Gene Set Enrichment Analysis (GSEA) (Mootha *et al*., 2003).

The required inputs to SAM-GS are: (1) gene expression measurements of each sample; (2) a phenotype indicator of each sample; and (3) definitions of gene sets or biological pathways, whose associations with the phenotype are of primary scientific interest. The phenotype must be binary (e.g., cases vs. controls). Continuous phenotype coding (more than two phenotypic groups) may be considered in the future. SAM-GS computes a t-like statistic for each member of a gene set, in the same way as SAM does, and uses the sum of their squares over the gene set as the measure of association between the gene set and the phenotype of interest. Statistical significance of the association is assessed using a permutation test, permuting the phenotype labels. Multiple gene sets can be considered in an analysis, assessing the false discovery rate of each gene set (Storey, 2002; Storey and Tibshirani, 2003; Storey, Taylor and Siegmund, 2004).

This document assumes the basic knowledge of SAM, p-value and q-value and permutation tests.

2 Obtaining SAM-GS

SAM-GS is free software (Microsoft Excel Add-in) created by the Yasui Biostatistics Research Group at the University of Alberta, Canada. You can download the program from the "Software/Programs" section on http://www.ualberta.ca/~yyasui/homepage.html.

3 System requirements

- 32-bit Microsoft Excel 2007 or 2010.
- 32-bit or 64-bit Microsoft Windows XP/Vista/7.

4 Installation

1) Download the "EdmontonMethods.rar" package from the "Software/Programs" section of our website: http://www.ualberta.ca/~yyasui/homepage.html
2) Uncompress the folder to the C:\ directory. You should see a folder "C:\EdmontonMethods" as shown in Figure 1. The program will not run if the folder is located anywhere other than the C:\ directory.

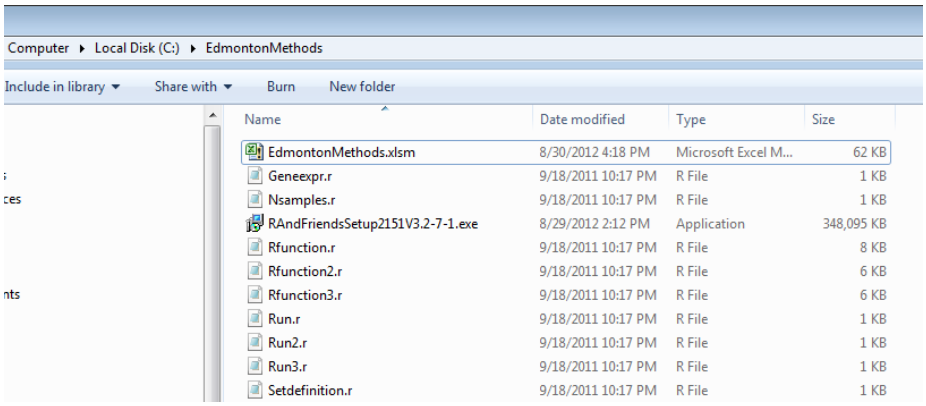


Figure 1

3) Open the folder and execute “RAndFriendSetup215V3.2-7-1.exe”. This package will install the latest version of R and other related software on your computer. We recommend selecting all 3 components in the installation procedure (Figure 2).

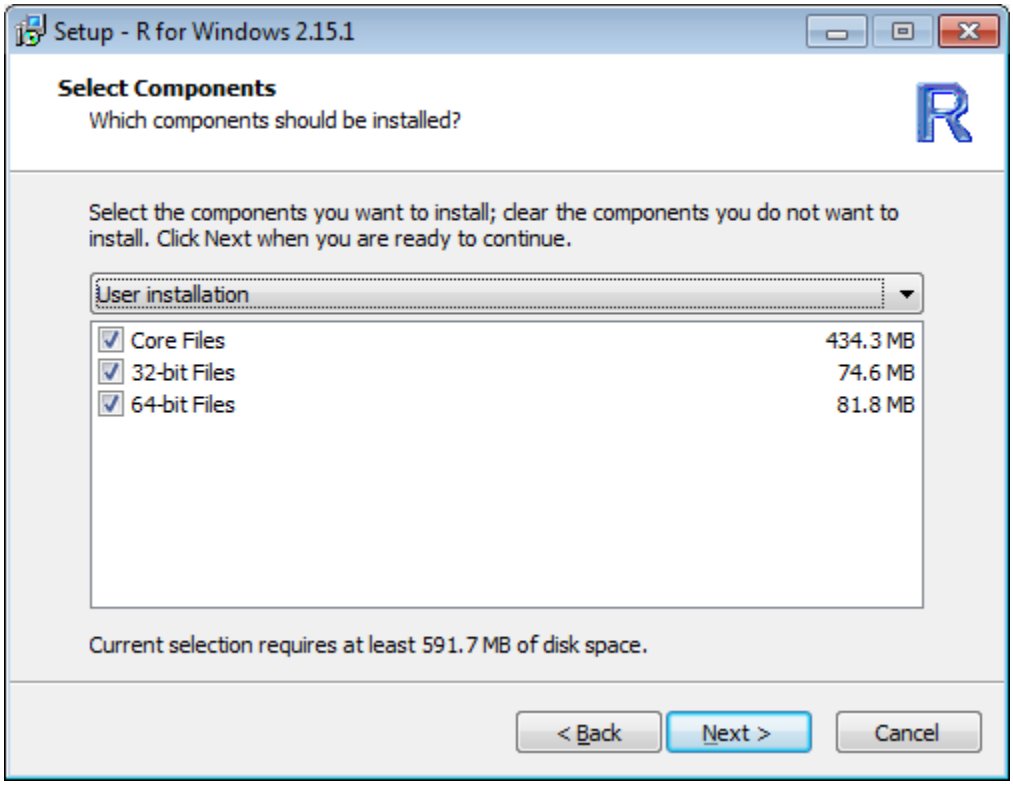


Figure 2

4) Once the installation has finished successfully, a word file will open (Figure 3). You can close this file.

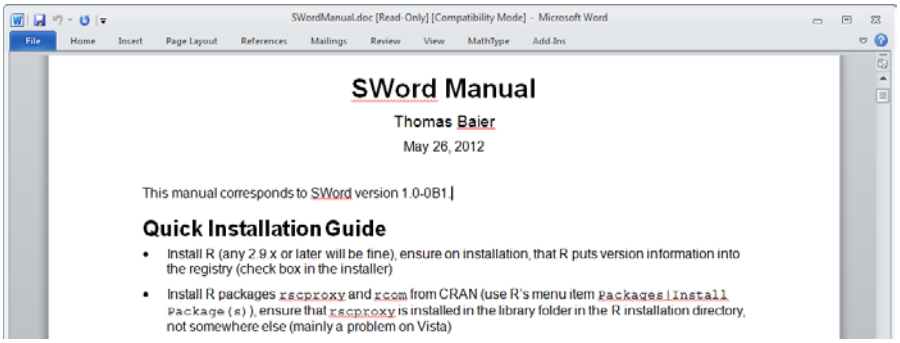


Figure 3

5) Once the installation process has fully completed, execute R (you may do this by double clicking the 'R i386 2.15.1' icon,  , located on your desktop or through Start -> All Programs -> R -> R i386 2.15.1). In to the R interface (Figure 4), type the following command to install an additional package, 'qvalue', to the R program:

```
source("http://bioconductor.org/biocLite.R")
biocLite("qvalue")
```

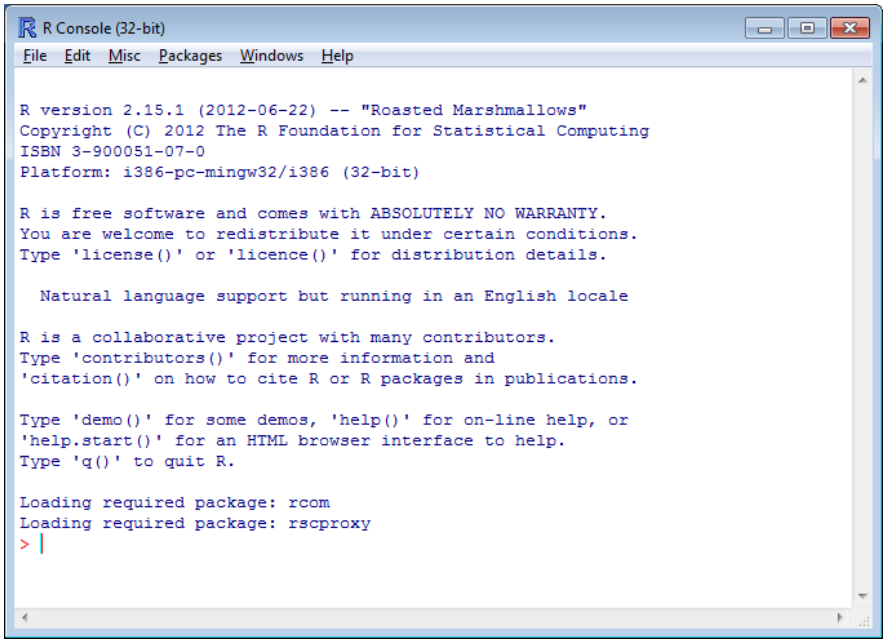


Figure 4

6) Close the R program and open the Excel file "EdmontonMethods.xlsm" included in our package by double clicking it. In order to execute our program, you must enable the Macro functions in the Excel software. To do this, please select the options of "Trust access to the VBA project object model" and "Enable all macros (not recommended; potentially dangerous code can run)" listed under "Macro Settings" (Office Button -> Excel Options -> Trust Center -> Trust Center Settings -> Macro Settings) and then click OK (Figure 5).

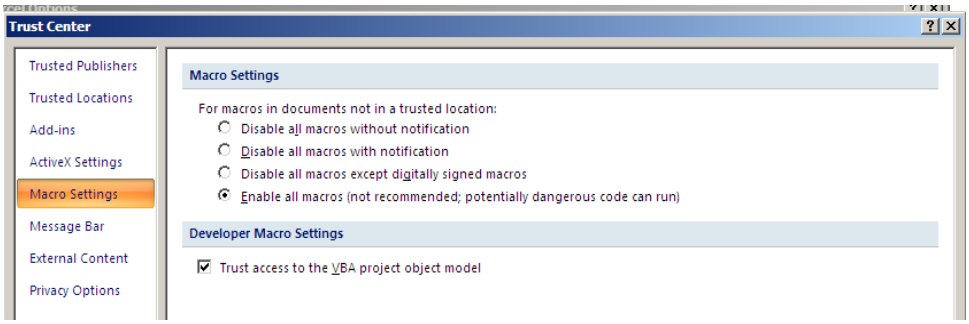


Figure 5

7) Once you have completed the above procedures, browse to "Add-Ins." You will now be able to see the "Edmonton Methods" option appearing on the ribbon and 2 sheet tabs, "SAM-GS" and "Preview" available at the bottom of the worksheet (Figure 6) (Caution: do not delete or alter the worksheet names. Doing so will prevent our program from performing its designated tasks).

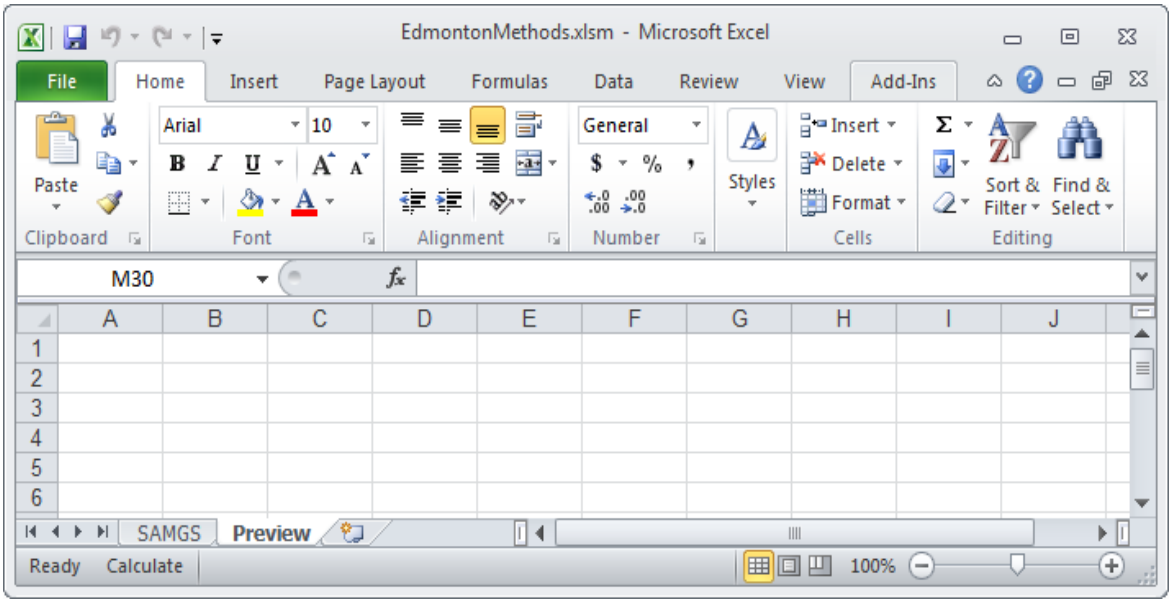


Figure 6

5 Document

This document is available from the "Software/Programs" section of our webpage http://www.ualberta.ca/~yyasui/homepage.html.

6 Preparing your datasets

Sample SAM-GS datasets (Mootha *et al.*, 2003, obtained from the GSEA webpage: http://www.broad.mit.edu/gsea) are available from the SAM-GS website: http://www.ualberta.ca/~yyasui/homepage.html. These files will be used for a demonstrative purpose in the subsequent texts.

6.1 The gene expression file

Your gene expression file must be formatted in a precise way for SAM-GS to analyze it. Below, we describe this formatting, using the p53.csv gene-expression dataset as a model.

The Excel file p53.csv (Figure 7) contains the gene expression measurements for 10,100 genes (probes) and 50 samples, 33 of which are classified as carrying a p53 mutation, while the other 17 are classified as wild type (note that samples must be in adjacent columns with no spaces between them). The first row of the spreadsheet must list the sample classifiers, one per column, starting with column 3. The first two columns contain information about the genes (probes):

Column 1 = Name of the gene (probe).

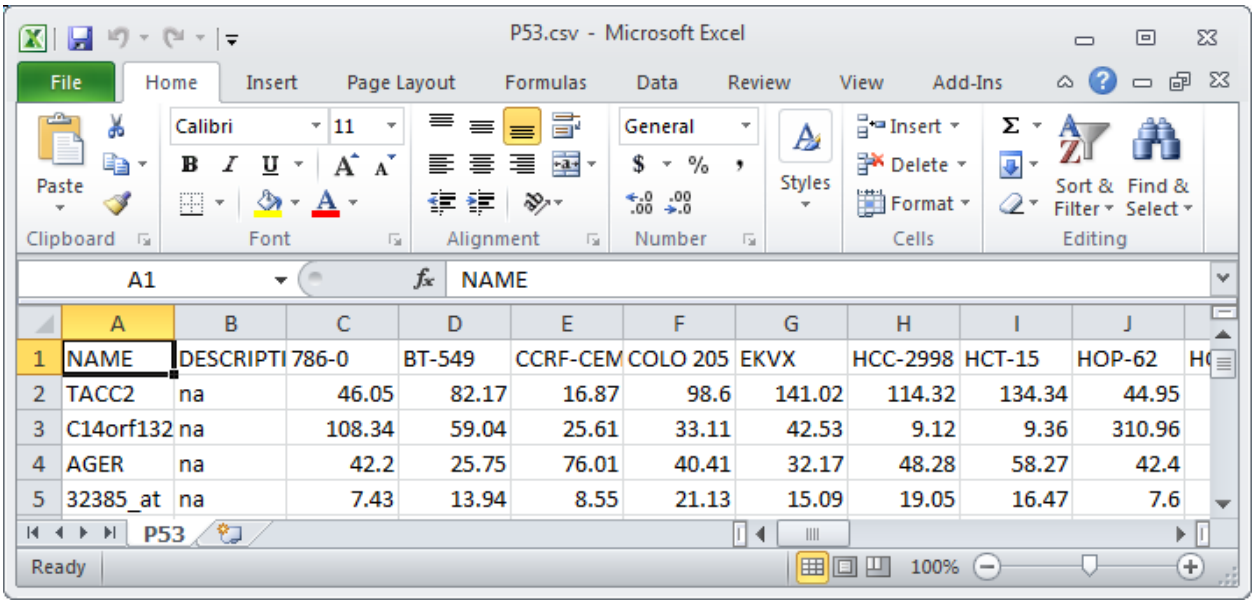
Column 2 = Description of the gene (probe) for users' reference.

Gene expression measurements corresponding to each gene-sample pair fills in the dataset.

6.2 The gene set definition file

Your gene set definition file is an Excel file which contains specifics regarding your gene sets. A sample gene set definition file C2.csv (Figure 7) can be obtained from our website (originally obtained from the GSEA web-page: http://www.broad.mit.edu/gsea). It contains 308 gene sets. Below, we will describe the formatting of gene set definition files, using our sample file to illustrate.

The first row of a gene set definition file contains the gene set names, starting from the second column, while the first column contains the gene names, beginning in the second row. This is illustrated in C2.csv (Figure 7). For each of the 10,100 genes in this example, a 1 is assigned if the gene is present in the gene set, while a 0 is assigned if the gene is not present in the gene set. Missing values are not accepted in the gene expression files or the gene set files.



	A	B	C	D	E	F	G	H	I	J
1	NAME	DESCRIPTI	786-0	BT-549	CCRF-CEM	COLO 205	EKVX	HCC-2998	HCT-15	HOP-62
2	TACC2	na	46.05	82.17	16.87	98.6	141.02	114.32	134.34	44.95
3	C14orf132	na	108.34	59.04	25.61	33.11	42.53	9.12	9.36	310.96
4	AGER	na	42.2	25.75	76.01	40.41	32.17	48.28	58.27	42.4
5	32385_at	na	7.43	13.94	8.55	21.13	15.09	19.05	16.47	7.6

Figure 7

In some situations, user may prefer to have multiple set definition files, for example, C2part1.csv and C2part2.csv from our website. They contain 254 gene sets and 54 gene sets, respectively. This is also acceptable.

6.3 Using data in multiple sheets

While our SAM-GS software works only with Excel 2007/2010, data prepared using Excel 2003 may still be loaded, however, the user will encounter problems if their file contains more than 256 columns. To remedy this problem, one must split the file into multiple files. In this example (C2.csv) there are 308 gene sets. When you include the first column with the gene names, the total number of columns is 309, which exceeds the Excel 2003 maximum of 256. Therefore this data must be split in to two files. The first file can include the first 255 gene sets, and the first column which contains the gene names, totaling 256 columns. The remaining 53 gene sets are arranged in a second file without the gene set name column (refer to our example C2part1.csv and C2part2.csv).

7 Running SAM-GS

Step 1: Select "Edmonton Methods" from the ribbon under "Add-Ins." A dialog box will appear (Figure 8).

Step 2: Select "SAM-GS", under "Algorithm" and fill out the four parameters required for the execution of the program:

> **Number of Permutations**: SAM-GS uses permutations to obtain p-values. The more permutations, the more accurate the resulting p-values are. However, more permutations will require more time to run. The default number of permutations is 1000.

Number of percentile bands in SAM: This parameter is used for computation of s_0. For details, please see Tusher *et al.*, 2001. The default number is 100.

Number of Group 1 columns (n_1): The number of samples in Group 1. The first n_1 columns (samples) of the gene expression file belong to group 1 (either the case or control group).

Number of Group 2 columns (n_2): The number of samples in Group 2. The remaining columns (samples) belong to group 2 (the complement of group 1). The sum of n_1 and n_2 equals to total number of samples in your study.

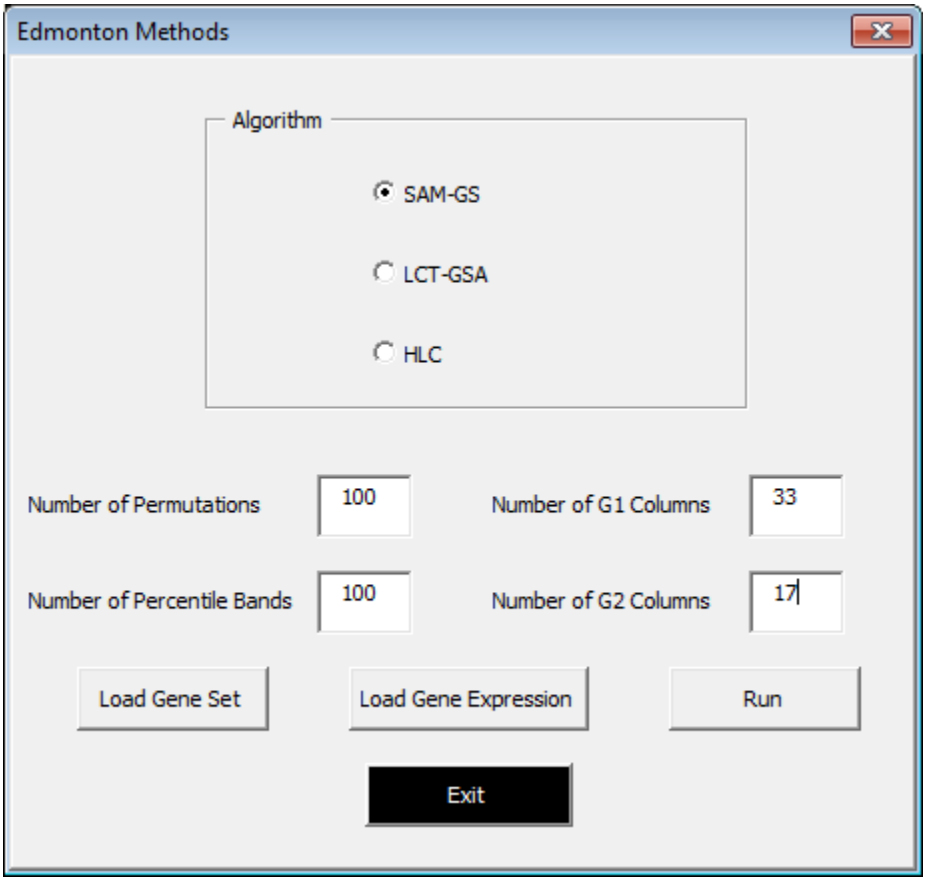


Figure 8

Step 3: Load the gene set data by selecting "Load Gene Set" and selecting your gene set definition file(s). You have the option of choosing gene set definition files as binary files (e.g. C2.csv from our website), or as the GSEA format (c2.v2.symbols.gmt from our website). If you have multiple binary set definitions files (e.g. C2part1.csv and C2part2.csv from our website), you can choose the number of sheets you have. In the case you have 2 multiple files (e.g. C2part1.csv and C2part2.csv), therefore choose option 2. Click "OK" to load the files (Figure 9).

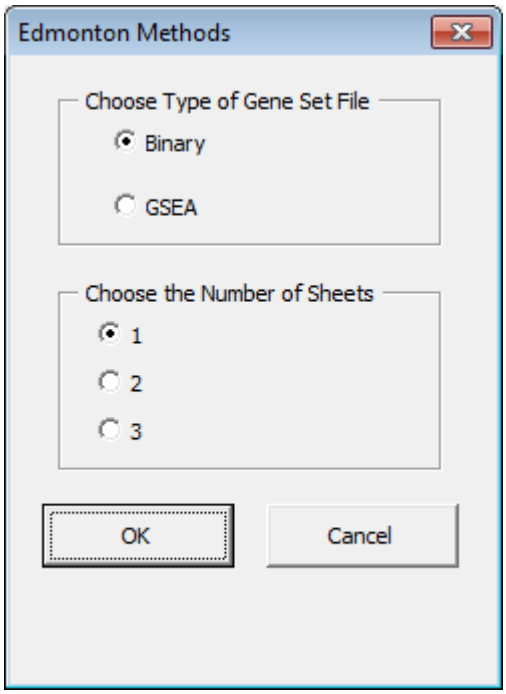


Figure 9

Step 4: Load the gene expression data by clicking "Load Gene Expression" and selecting your gene expression data file(s).

Step 5: Select "Run" to start the computation. You will see the dialog shown as Figure 10. Click "OK" and the program will start to run. This process may take some time.

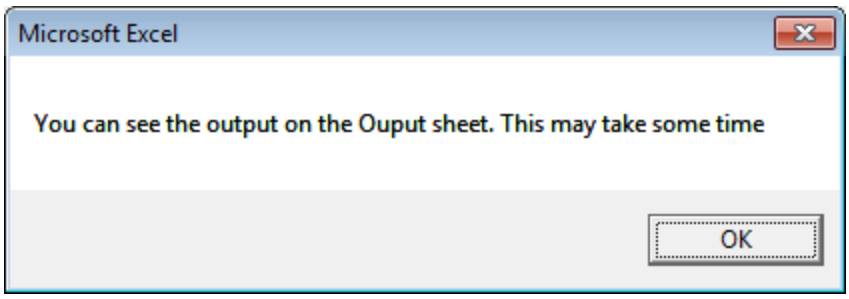


Figure 10

Once a “Done” message appears on the screen (Figure 11), the program has completed the task and outputted the results to the “Output” worksheet (please note that the time length of execution is dependent on the size of datasets, number of permutations requested and configuration of your computer systems. On our computer, the program took about 2 minutes to finish analyzingthe sample datasets with 1000 permutations, and about 20 seconds with 100 permutations.

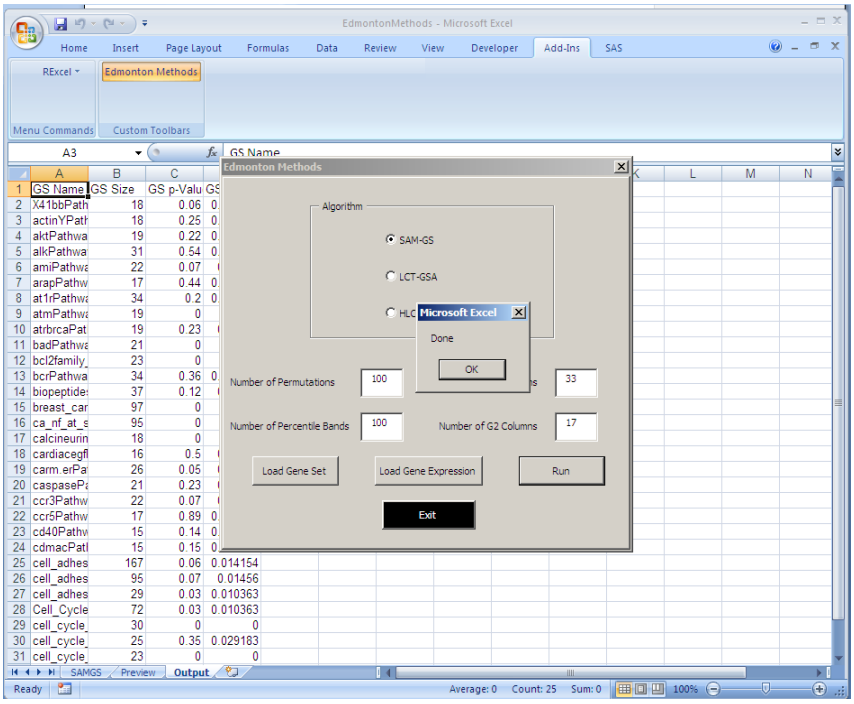
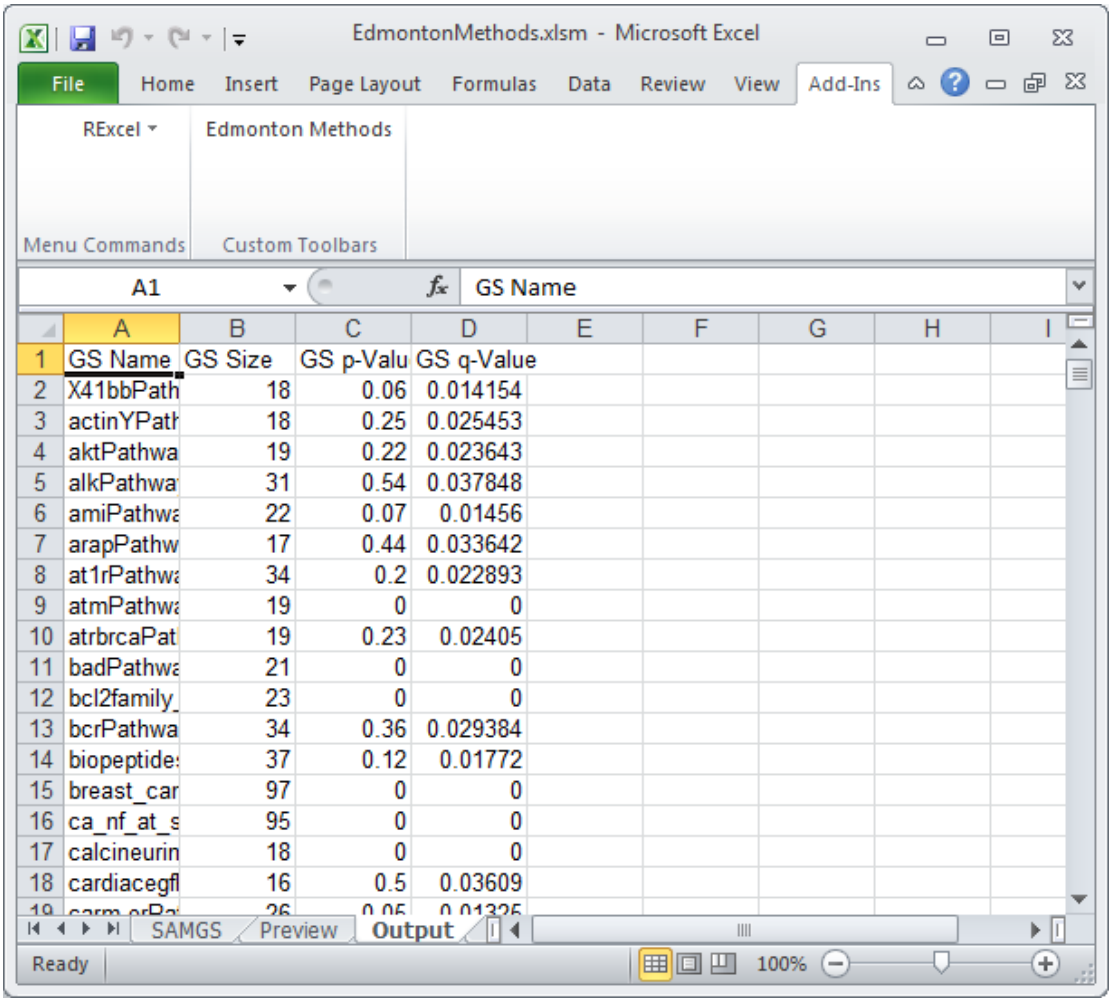


Figure 11

8 SAM-GS output

The analysis results display the p-value and q-value of each gene set based on the permutation test for no association between the gene expression of the gene set and the binary phenotype (Figure 12).



	A	B	C	D
1	GS Name	GS Size	GS p-Valu	GS q-Value
2	X41bbPath	18	0.06	0.014154
3	actinYPath	18	0.25	0.025453
4	aktPathwa	19	0.22	0.023643
5	alkPathwa	31	0.54	0.037848
6	amiPathwa	22	0.07	0.01456
7	arapPathw	17	0.44	0.033642
8	at1rPathwa	34	0.2	0.022893
9	atmPathwa	19	0	0
10	atrbrcaPat	19	0.23	0.02405
11	badPathwa	21	0	0
12	bcl2family_	23	0	0
13	bcrPathwa	34	0.36	0.029384
14	biopeptide	37	0.12	0.01772
15	breast_car	97	0	0
16	ca_nf_at_s	95	0	0
17	calcineurin	18	0	0
18	cardiacegfl	16	0.5	0.03609

Figure 12

9 Troubleshooting

- If, after installing the SAM-GS program, you open Excel by double clicking "EdmontonMethods.xlsm" and you do not see "Add-Ins" on the menu bar as below (Figure 13), RExcel was likely not installed successfully. To solve this, please re-run

"RAndFriendsSetup2151V3.2-7-1.exe" and follow the step-by-step setup dialogs until you see the word file shown in Figure 3 automatically open.

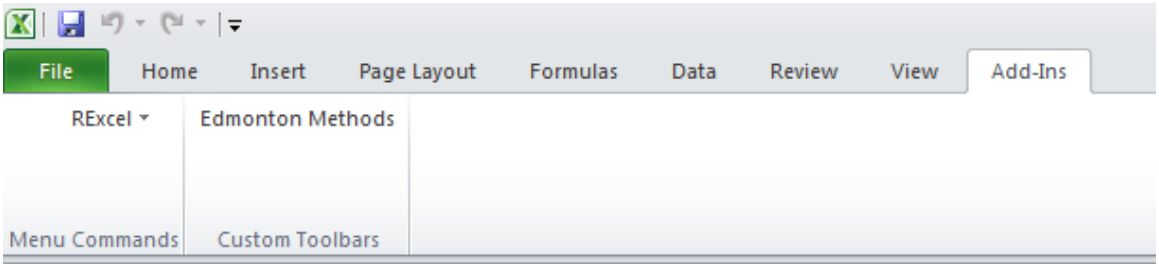


Figure 13

- If you encounter problems while loading datasets, this may be because the "Preview" worksheet is missing in the program. To fix this, you need to insert a blank worksheet beside the 'SAMGS' tab and re-name it "Preview", or re-open the "EdmontonMethods.xlsm" which, by default, has the "Preview" worksheet.
- If you see the error message shown in Figure 14, it is possible that you are missing files in the "C:\EdmontonMethods" folder, or you have not used the folder "C:\EdmontonMethods". Please make sure to uncompress the "EdmontonMethods.rar" folder to "C:\" as shown in Figure 1.

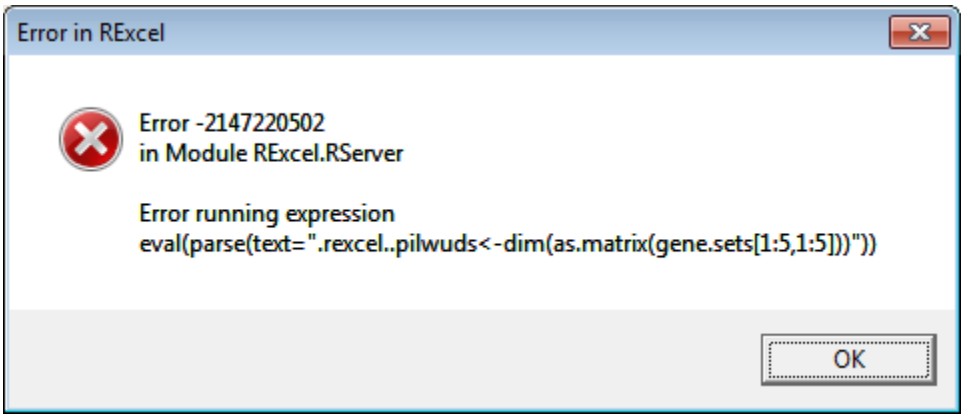


Figure 14

References

1. Dinu I., Potter, J. D., Mueller, T., Liu, Q., Adewale, A. J., Jhangri, G. S., Einecke**,** G., Famulski, K. S., Halloran, P., and Yasui, Y. (January 2007). Improving GSEA for Analysis of Biologic Pathways for Differential Gene Expression across a Binary Phenotype. *COBRA Preprint Series*, Article 16.

2. Mootha, V. K., Lindgren, C. M., Eriksson, K. F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstrale, M., Laurila, E., Houstis, N., Daly, M. J., Patterson, N., Mesirov, J. P., Golub, T. R., Tamayo, P., Spiegelman, B., Lander, E. S., Hirschhorn, J. N., Altshuler, D. & Groop, L. C. (2003) *Nat Genet* **34,** 267-73.

3. Tusher, V. G., Tibshirani, R. & Chu, G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A* **98,** 5116-21.

4. Storey JD. (2002) A direct approach to false discovery rates. *Journal of the Royal Statistical Society, Series B*, **64**: 479-98.

5. Storey JD and Tibshirani R. (2003) Statistical significance for genome-wide experiments. *Proceeding of the National Academy of Sciences*, **100**: 9440-5.

6. Storey JD, Taylor JE, and Siegmund D. (2004) Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: A unified approach. *Journal of the Royal Statistical Society, Series B*, **66**: 187-205.