

SAM-GS

Significance Analysis of Microarray for Gene Sets

Qi Liu Irina Dinu Yutaka Yasui

SAM-GS Excel Add-in, version 1.0.2

May 23, 2007

Table of contents

1	Introduction.....	2
2	Obtaining SAM-GS	3
2.1	Changes in version 1.0.2.....	3
3	System requirements.....	3
4	Installation.....	3
5	Document.....	4
6	Example dataset	4
6.1	The gene expression file	4
6.2	The gene set definitions	6
6.3	Using data in multiple sheets	6
7	Running SAM-GS.....	8
8	SAM-GS output	12
9	Troubleshooting	13

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 GSEA (Mootha *et al.*, 2003).

The 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 is binary (e.g., cases and controls). More than two groups or continuous phenotype coding 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 (MS Excel Add-in) made by Yasui Biostatistics Research Group at the University of Alberta. Please register and download the software at <http://www.ualberta.ca/~yyasui/homepage.html>.

2.1 Changes in version 1.0.2

- Time required to perform permutation tests is significantly improved. The method of the analysis remains the same. The default number of permutations is now 1000.

3 System requirements

- Microsoft Excel 2000 or more recent version.
- The latest version of R. This is freely available from the website <http://www.r-project.org/>. Download the windows executable version. The package q-value is needed, which can be downloaded from Prof. John Storey's website at <http://faculty.washington.edu/~jstorey/qvalue/windows.html>.

4 Installation

- Install the latest version of R. We recommend using all the default setting in the installation procedure.
- Install the Q-value library for R. The library can be installed by starting R and selecting the pull-down menu "Packages -> Install package from local zipped

file...", and then selecting the zip file downloaded from the above q-value website (e.g., qvalue_1.1.zip).

- Enable Macros in Excel
 1. Open Excel and Click Tools | Macros | Security
 2. On the Security Level tab click Medium | OK
- Download EdmontonMethods.xla from the website www.sam-gs.org and save it into C:\EdmontonMethods
- Install the add-in in Excel:
 1. Open Excel
 2. Click Tools | Add-ins | Browse to
C:\EdmontonMethods\EdmontonMethods.xla | OK

You should now see an "Edmonton Methods" choice on the Excel menubar.

5 Document

This document is available from <http://www.ualberta.ca/~yyasui/homepage.html>.

6 Example dataset

An example of the use of SAM-GS is available from the SAM-GS website

<http://www.ualberta.ca/~yyasui/homepage.html>, which was taken from Mootha *et al.*

(2003). We downloaded the example dataset from the GSEA web-page:

<http://www.broad.mit.edu/gsea>

6.1 The gene expression file

The Excel file p53.csv (Figure 1) has the gene expression measurements for 10,100 genes (probes) and 50 samples: with 33 being classified as carrying a p53 mutation and 17 as wild type (note that samples in Group 1 should be in adjacent columns and samples in Group 2 should be in adjacent columns). The first row of the spreadsheet has the sample names, one per column, starting at column 3. The first two columns have information about the genes (probes):

Column 1 = Name of the gene (probe)

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

	A	B	C	D	E	F	G	H	I	J	K	L
1	NAME	DESCR	786-0	BT-549	CCRF-CCOLO 2	EKVX	HCC-29	HCT-15	HOP-62	HOP-92	H	H
2	TACC2	na	46.05	82.17	16.87	98.6	141.02	114.32	134.34	44.95	73.96	
3	C14orf1	na	108.34	59.04	25.61	33.11	42.53	9.12	9.36	310.96	101.74	2
4	AGER	na	42.2	25.75	76.01	40.41	32.17	48.28	58.27	42.4	49.68	
5	32385_2	na	7.43	13.94	8.55	21.13	15.09	19.05	16.47	7.6	10.88	
6	RBM17	na	11.4	3	3.16	2.34	4.43	1.56	6.04	6.16	1.41	
7	DYT1	na	148.09	317.17	316.66	147.23	125.78	261.39	268.41	212.51	142.51	1
8	CORO1	na	8.62	9.12	1572.5	5.91	5.31	11.98	128.77	7.51	7.04	
9	WT1	na	206.74	136.71	141.34	129.09	138.01	138.16	130.14	123.8	131.91	1
10	SYCP2	na	7.94	35.68	7.8	1.97	7.75	4.73	7.23	5.72	2.02	
11	SULF1	na	10.45	8.5	4.05	4.77	2.35	3.72	3.34	809.88	2.58	1
12	C19orf2	na	6.22	5.16	3.95	37.56	110.36	208.29	251.67	4.43	3.34	
13	PHYH	na	209.99	253.07	90.36	61.83	360.49	145.01	130.16	165.5	343.02	1
14	31336_2	na	3.35	5.28	2.98	4.82	4.36	1.45	5.57	4.38	3.57	
15	TOP2B	na	349.1	462.16	1348.4	274.07	730.39	350.95	635.21	644.7	516.55	5
16	ATP9A	na	126.28	197.49	117.66	380.3	354.14	84.19	243.24	235.76	252.57	2
17	TGFBR	na	11.73	13.13	16.2	12.97	9.55	8.3	20.65	17.05	19.68	
18	IRF7	na	98.46	246.49	52.48	70.54	33.35	198.34	155.66	144.43	56.44	
19	ZNF189	na	24.28	56.32	45.09	44.11	36.9	37.76	65.64	36.62	24.88	
20	VTI1B	na	184.52	85.33	123.2	128.74	167.21	135.93	165.11	182.51	245.75	
21	32413_2	na	32.17	23.72	14.51	14.68	22.96	18.22	6.37	19.93	8.14	

Figure 1

6.2 The gene set definitions

The Excel files C2part1.csv (Figure 2A) and C2part2.csv (Figure 2B) have the definitions of gene sets, taken from GSEA web-page: <http://www.broad.mit.edu/gsea>

The reason for using two files is that an Excel file can contain only 256 columns which is insufficient for the number of gene sets 308 we have here: please see Section 6.2 for details.

The first row of C2part1.csv has the gene set names, starting from the second column, while the first column in C2part1.csv has the gene names (10,100 genes). For each of the 10,100 genes, if the gene is in the gene set, 1 is assigned to the corresponding cell of C2part1.csv. Otherwise, 0 is assigned to that cell. Missing values are not allowed in the gene expression files or the gene set files.

6.3 Using data in multiple sheets

The maximum number of columns one can have in an Excel worksheet is 256 columns. If you have more than 256 columns, you can arrange the data in multiple sheets before invoking SAM-GS. In the above example, there are 50 samples in the gene expression spreadsheet p53.scv. Plus the first two columns, the total number of columns is 52. One worksheet is enough to cover the gene expression data in this case.

There are 308 gene sets. Plus the first column with the gene names, the total number of columns is 309, which exceeds the 256. The C2part1.csv file has the first column with gene names, plus 255 columns, one for each gene set. The remaining 53 gene sets are arranged in C2part2.csv, one per column.

If you have to use multiple sheets for gene expression, only the first sheet contains the gene name and gene description columns. Similarly, if you have to use multiple sheets for the gene set definitions, only the first sheet contains the gene sets names.

genes	41bbPa	actinYP	PaktPath	alkPath	amiPatf	arapPat	at1rPatl	atmPatf	atrbrca	badPatf	b
TACC2	0	0	0	0	0	0	0	0	0	0	0
C14orf1	0	0	0	0	0	0	0	0	0	0	0
AGER	0	0	0	0	0	0	0	0	0	0	0
32385_2	0	0	0	0	0	0	0	0	0	0	0
RBM17	0	0	0	0	0	0	0	0	0	0	0
DYT1	0	0	0	0	0	0	0	0	0	0	0
CORO1	0	0	0	0	0	0	0	0	0	0	0
WT1	0	0	0	0	0	0	0	0	0	0	0
SYCP2	0	0	0	0	0	0	0	0	0	0	0
SULF1	0	0	0	0	0	0	0	0	0	0	0
C19orf2	0	0	0	0	0	0	0	0	0	0	0
PHYH	0	0	0	0	0	0	0	0	0	0	0
31336_2	0	0	0	0	0	0	0	0	0	0	0
TOP2B	0	0	0	0	0	0	0	0	0	0	0
ATP9A	0	0	0	0	0	0	0	0	0	0	0
TGFBFR	0	0	0	1	0	0	0	0	0	0	0
IRF7	0	0	0	0	0	0	0	0	0	0	0
ZNF189	0	0	0	0	0	0	0	0	0	0	0
VT11B	0	0	0	0	0	0	0	0	0	0	0
32413_2	0	0	0	0	0	0	0	0	0	0	0

Figure 2A

	A	B	C	D	E	F	G	H	I	J	K
1	GLYCO	GLYCO	GO_000	GO_RO	human	INS	mitoch	PGC	shh_lis	INSULIN	INSULIN
2	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	1	0	0	0
14	0	0	0	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0
16	0	0	0	0	0	0	0	0	0	0	0
17	0	0	0	0	0	0	0	0	0	0	0
18	0	0	0	0	0	0	0	0	0	0	0
19	0	0	0	0	0	0	0	0	0	0	0
20	0	0	0	0	0	0	0	0	0	0	0
21	0	0	0	0	0	0	0	0	0	0	0

Figure 2B

7 Running SAM-GS

Open the gene expression file(s) and the gene set definitions file(s). In any of the opened file, click "Edmonton Methods" on the Excel menubar, a dialog form shown in Figure 3 now pops up.

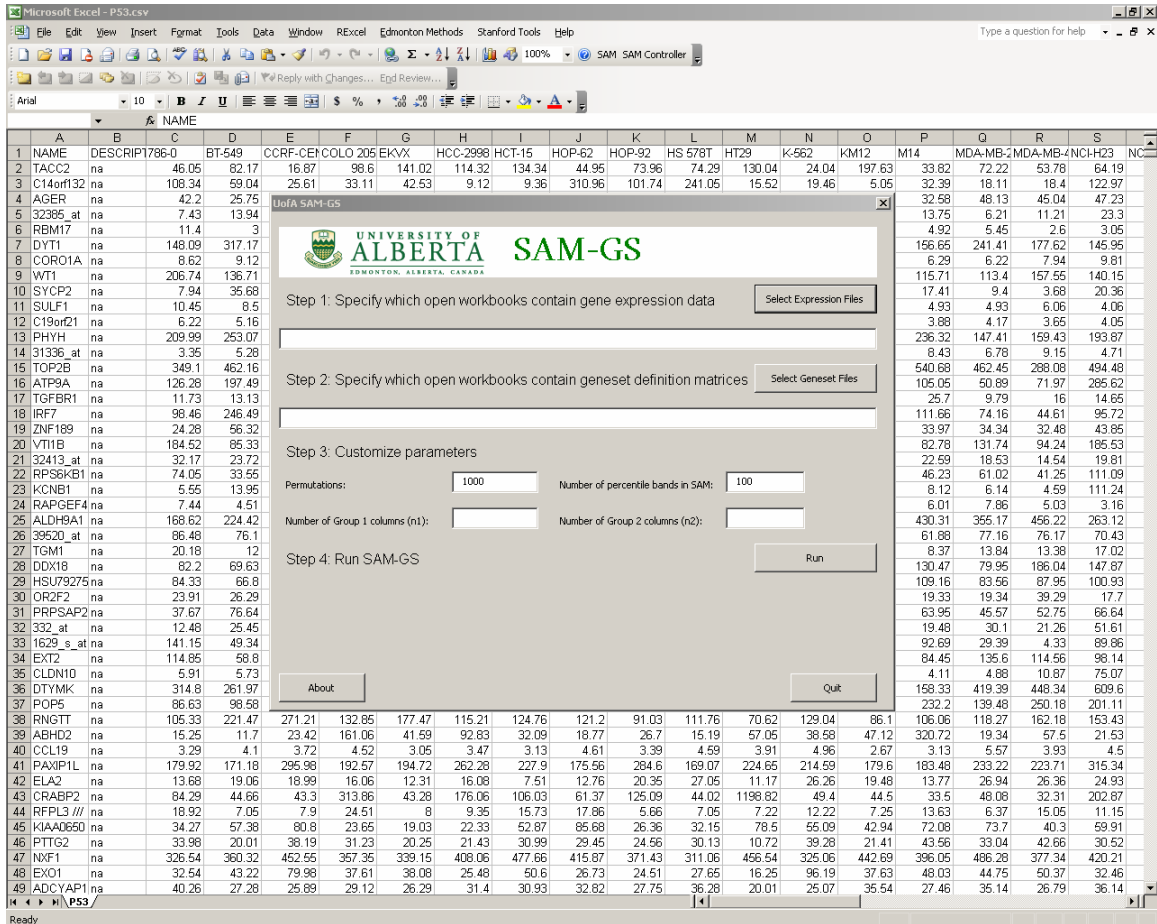


Figure 3

Step 1: Select the gene expression files by clicking the button “Select Expression Files”.

Step 2: Select the gene sets files by clicking the button “Select Geneset Files”.

Step 3: Specify the number of samples in Group1 and Group2, and if desired, change any of the values of the default parameters (the number of permutations and the number of percentile bands in SAM) (Figure 4).

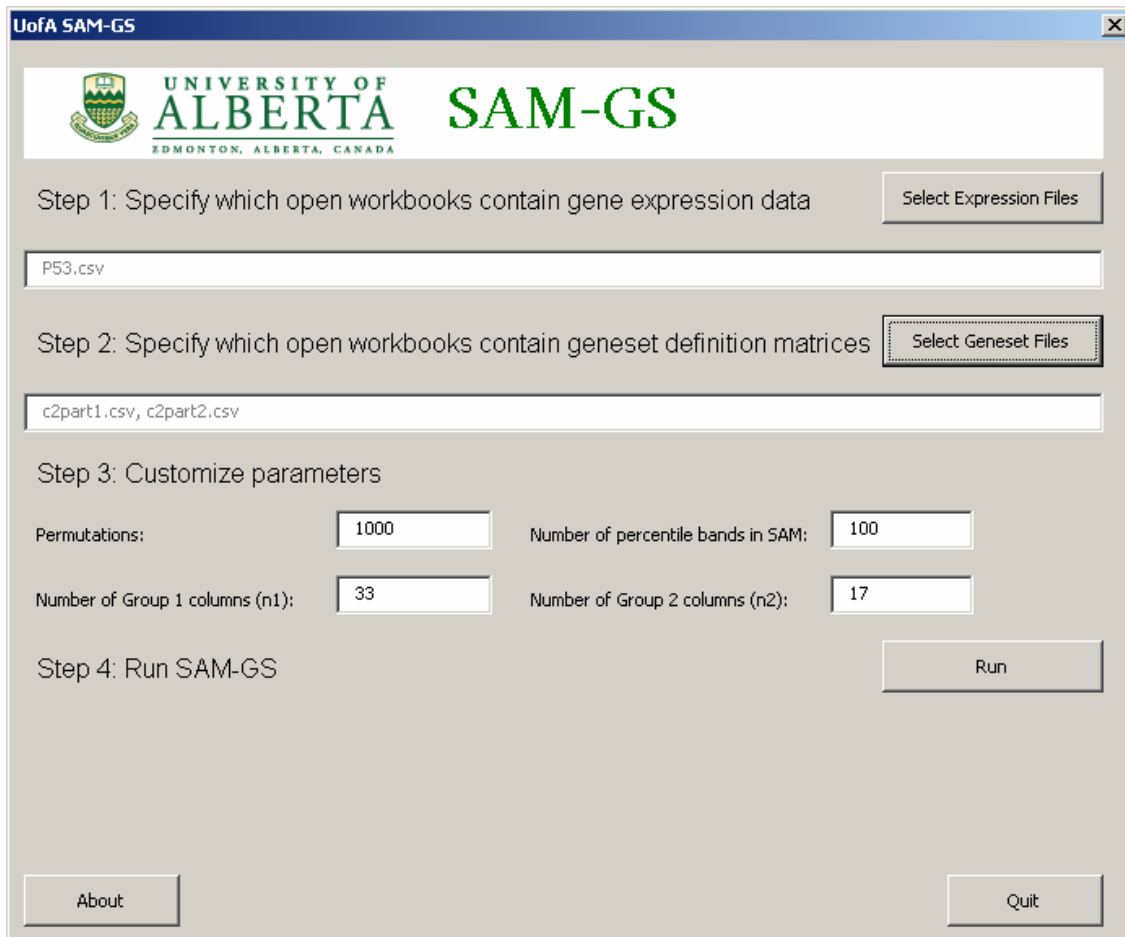


Figure 4

Step 4: Click the Run button to run the analysis. Wait till a message box with “Done” pops out, then click “OK” (Figure 4B). (Run time depends on the datasets, the number of permutations and the PC system. In our system, the example took 5 minutes.)

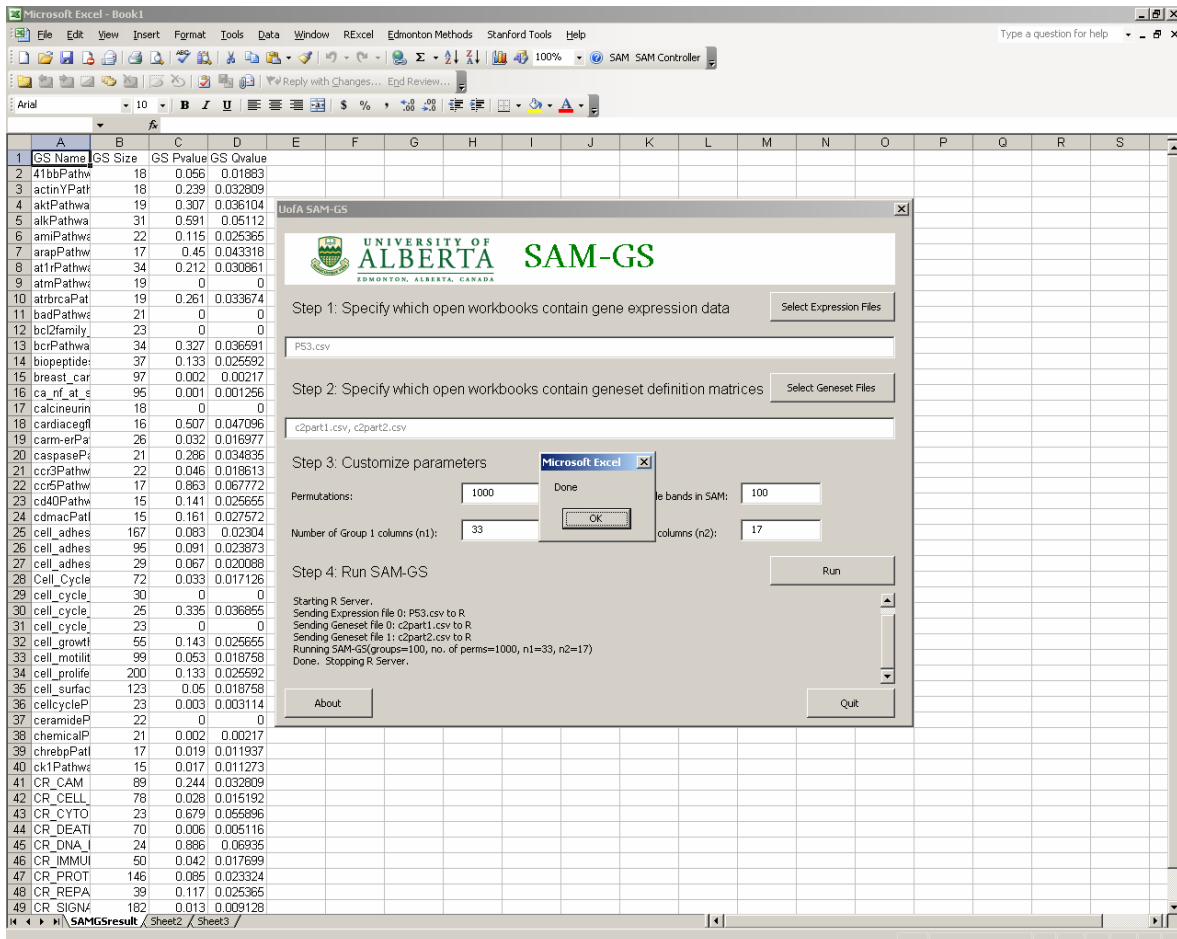


Figure 4B

The software creates a separate Excel file, named “book1”, which contains the results, including the gene set name, the gene set size, the p-value and q-value for each gene set.

In Step 3, you can specify the values of the following parameters:

Permutations: SAM-GS uses permutations to get p-values. The bigger the number of permutations is, the more accurate the resulting p-values are. But 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.

Number of Group 2 columns(n_2): The number of samples in Group 2.

8 SAM-GS output

The sheet of the analysis results (Figure 5, SAMGSresult) shows 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.

GS Name	GS Size	GS Pvalue	GS Qvalue
41bbPathway	18	0.056	0.01882961
actinYPathway	18	0.239	0.03280906
aktPathway	19	0.307	0.03610389
alkPathway	31	0.591	0.05111991
amiPathway	22	0.115	0.02536534
arapPathway	17	0.45	0.04331841
at1rPathway	34	0.212	0.03086055
atmPathway	19	0	0
atrbrcPathway	19	0.261	0.03367386
badPathway	21	0	0
bcl2family_and_reg_network	23	0	0
bcrPathway	34	0.327	0.03659079
biopeptidesPathway	37	0.133	0.02559213
breast_cancer_estrogen_signalling	97	0.002	0.0021703
ca_nf_at_signalling	95	0.001	0.00125649
calcineurinPathway	18	0	0
cardiacegfPathway	16	0.507	0.04709627
carm-erPathway	26	0.032	0.01697654
caspasePathway	21	0.286	0.03483547
ccr3Pathway	22	0.046	0.01861305
ccr5Pathway	17	0.863	0.06777178
cd40Pathway	15	0.141	0.02565484
cdmacPathway	15	0.161	0.02757193
cell_adhesion	167	0.083	0.02304047
cell_adhesion_molecule_activity	95	0.091	0.02387326
cell_adhesion_receptor_activity	29	0.067	0.02008847
Cell_Cycle	72	0.033	0.01712647

Figure 5

9 Troubleshooting

When you run the software and get the error message “Run-time error ‘13’ Type mismatch”, then it may be due to the following problems:

- In the gene set files, each column contains values 1 and 0, indicating a gene in the gene set or not. If a column contains only 0's, i.e., no gene is in the gene set, you will get the above error message. Please ensure that each gene set (column) has at least one gene (one entry of 1 in the column).
- If you have appropriate data files and correct parameter numbers, but you still get the above error message, it may be due to a Microsoft's software update issue. The MS Excel on your machine may not be using the most updated service pack. You can go to <http://office.microsoft.com/en-us/downloads/default.aspx> and click the "Office Update" on the left box to update your MS Office.

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.