

GSECA (Gene set expression coherence analysis)

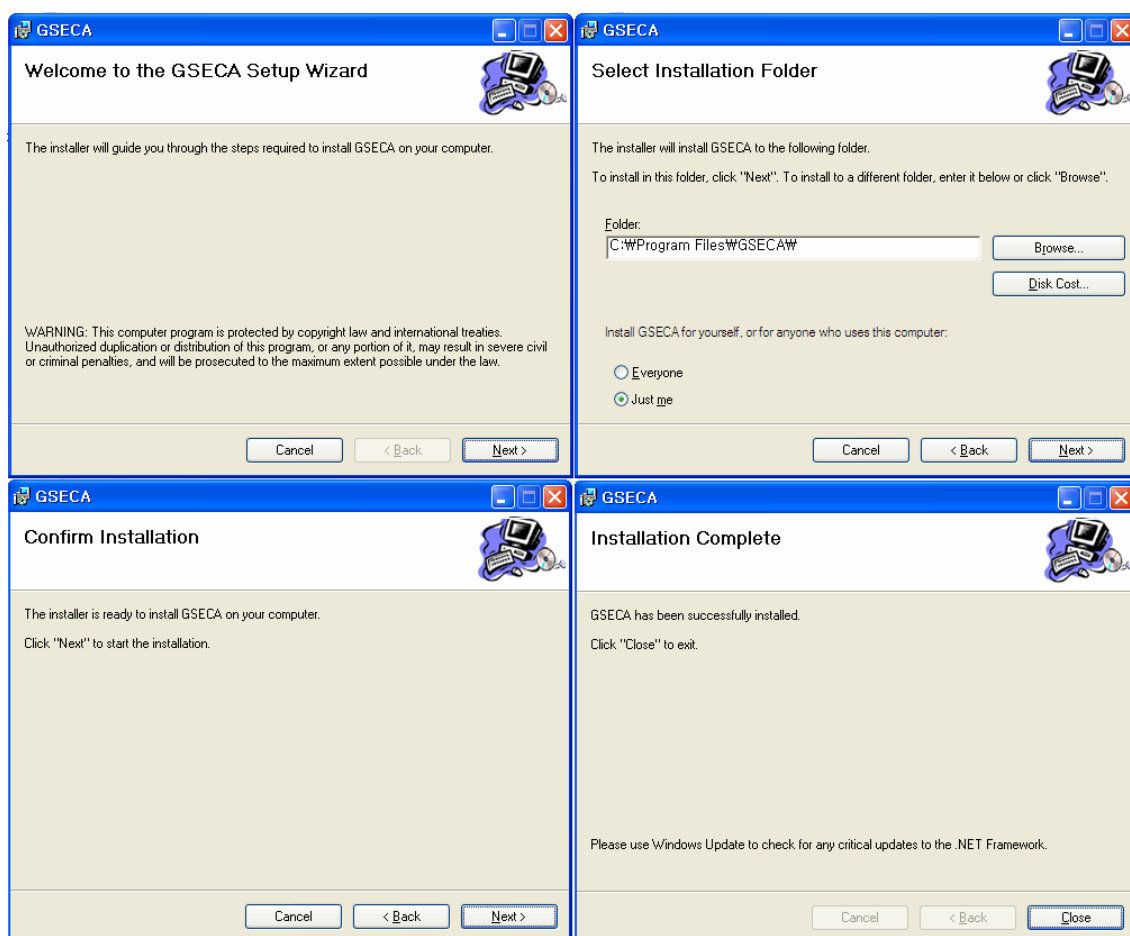
**GSECA (Gene Set Expression Coherence Analysis) for
Integrative and Comprehensive Analysis of Microarray-
based Expression Profiles**

GSECA software package is available with gene set information and testable file in our website: <http://www.systemsbiology.co.kr/GSECA/>

GSECA (Gene set expression coherence analysis)

1. Install

- GSECA is a standalone software running on Microsoft Windows platform. The installation is quite simply done just executing the installation package (compatible to Microsoft Installer. The required files (default gene set files and testable expression profile) will be simultaneously installed with program file (.NET framework might be also installed if the machine does not have an appropriate version .NET).



GSECA (Gene set expression coherence analysis)

2. Installed Files

Along with program files and related information files of genes, (1) gene set files representing different gene annotation categories and (2) a testable expression file will be also installed.

(1) Gene set database (Program Files/GSECA/Geneset/)

FunctionalGeneSet.txt

- 5,025 gene sets, each of which shares a common functional annotation from GO (Gene Ontology), KEGG, GenMAPP database

PromoterGeneSet.txt

- 432 gene sets categorized by the presence of sequence motifs in promoter region (corresponding to TFBS based on sequence information of TRANSFAC)

Note: Detailed descriptions on the construction and source of functional and promoter gene sets, are available in our previously publication (Kim et al, BMC Bioinformatics 8:453, 2007)

miRNAGeneSet.txt

- 162 gene sets corresponding to putative miRNA target genes. Data sets were constructed from the prediction data of Lewis BP et al (Cell, 115:787, 2003)

DrugGeneSet.txt

- Drug-perturbation gene sets representing the over-/under-expressed genes with the treatment of drug in in vitro human cell line model (derived from large-scale expression profiles of Connectivity Map; Justin et al, Science 313:1929, 2006). For each batch (perturbagen-vs-vehicle pair), expression ratio was calculated per gene and the ordered rank list of genes were constructed. Top-ranking 100 genes (up-regulation) and bottom 100 genes (down-regulation) were selected and used as gene sets for the corresponding batch as described previously (Rhodes et al, Neoplasia 9:443, 2007).

(2) Publicly available gene expression sets (Program Files/GSECA/SampledData/)

GDS2431_hErythropoiesis.txt (Kellet et al, Physiol Genomics 28:114, 2006)

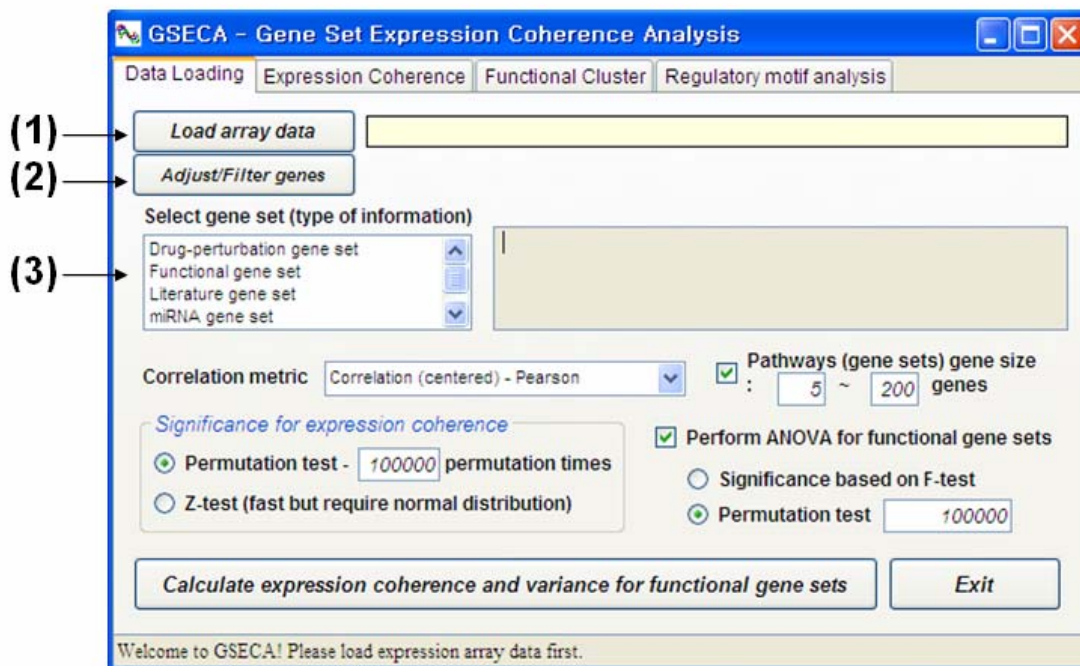
- Time-series expression profile representing erythropoiesis of human bone marrow

GSECA (Gene set expression coherence analysis)

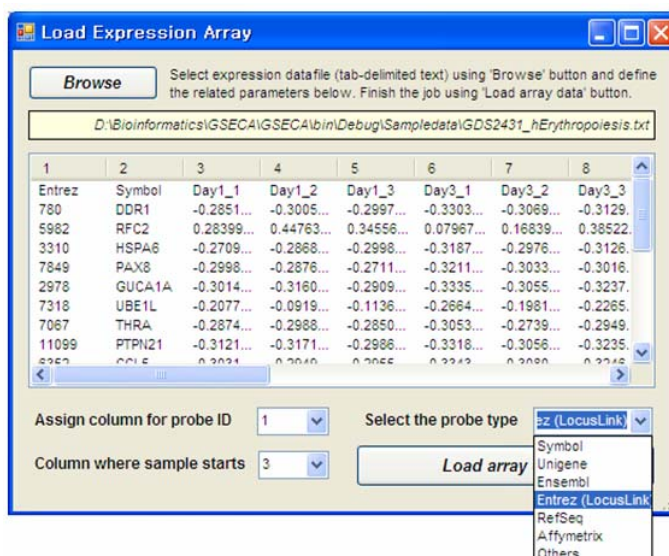
3. Basic Instructions

1. Data load

The main window is composed for 4 functional sections. In the first section of **Data Loading**, the upload/adjusting of expression dataset, the selecting the gene sets and basic parameter tuning can be done.



(1) By pressing Load array data, a new menu-window will show up.



GSECA (Gene set expression coherence analysis)

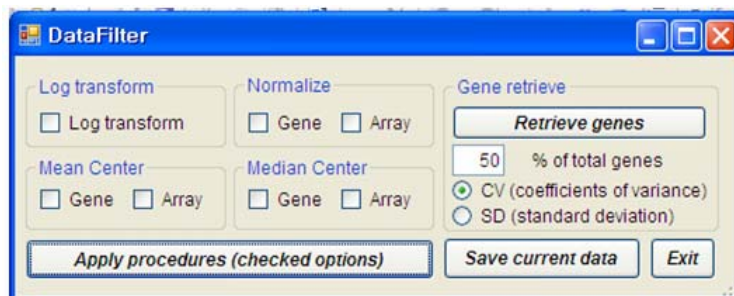
By pressing **Browse** button, you can select the file of expression data to import. After selecting the file, the path and file name will be shown along with the first few example lines of the dataset in a tabular format (in a table). The array datasets comprise the expression values of multiple probes, each of which corresponds to an individual gene. The column representing the probe annotation (**Assign column for probe ID**) and column indicating the start point of data values (**Column where sample starts**) must be designated. The type of probe annotation in selected column for (**Select the probe type**) should be also indicated in pull-down menu.

Note: The gene sets used in GSECA analysis has 'gene symbol-based annotation'. Thus, for proper gene matching between array dataset and gene set, the probe annotations in array data sets are automatically converted into gene symbol annotation. If a column for gene symbols is already present in dataset (i.e., column 2 in the example screenshot above), this column should be selected in (4) as well as 'Symbol' as probe type in (6). If gene symbol is not available (as is often case), one of the columns (Unigene, Ensembl, RefSeq, Entrez or LocusLink, and Affymetrix ID) must be selected both in (4) as column index and in (6) as probe type. If a column with heterogeneous annotation types should be selected, 'Others' is a preferable option for (6). By pressing Load array button, the software reads the expression data and automatically convert the probe annotation into gene symbols. After converting, the matching rates (the probe number successfully matched to gene symbol/total probe number present in array dataset) will be displayed after loading dataset.

In conventional array platform, some genes are represented more than once in a platform. To deal with this redundancy, GSECA gives each member of a set of n replicates a weight $1/n$ in calculation of average, assigning the mean expression values for a gene (a symbol). Thus, the final gene/symbol number after data loading will be often smaller than the actual row number of initial array dataset. For example, the test dataset (GDS2431) based on Affymetrix U133A array platform composed of 54,675 probes (rows) is reduced into the profile of 18,946 non-redundant gene/symbols after loading.

GSECA (Gene set expression coherence analysis)

2. Gene filtering and selection of highly variable genes



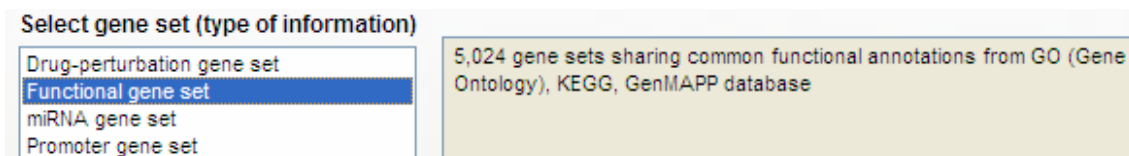
- Preprocessing of 'normalization' and 'variability-based gene filtering' is available with Adjust/Filter genes button in main window. In a new window, 4 kinds of data processing can be performed in a familiar interface with Eisen's Cluster program (Eisen et al, PNAS 95:14863, 1998). The order of operation is also the same with the Eisen's Cluster program (1. log transform → 2. mean center rows → 3. median center rows → 4. normalize rows → 5. mean center columns → 6. median center columns → 7. normalize columns) while the operations are not associative with each other. The selected operations are performed in successive steps by pressing the button of **Apply procedures (checked above)**. In brief, Log transform will replace the data values into \log_2 while cares must be taken since negative values will be neglected for this log transform step. Mean/Median center will set the mean/median of the rows/columns to zero, respectively. Normalize will set the standard deviation of the rows/columns to 1.0 (more detailed descriptions and related issues for operations are available in the manual of Eisen's Cluster and TreeView program).

For conventional clustering, it is generally recommended to use a subset of genes (i.e., highly variable 10% of genes) rather than to use the entire genes in the dataset. For GSECA analysis, it is possible to reduce the entire genes into 'manageable' number of genes. Since variability of genes is usually used for filtering criteria, GSECA provides two options to measure the expression variability - CV (coefficient of variance) or SD (standard deviation). For example, if user selected **CV option/ 50% of total genes** and press **Retrieve genes**, 50% of genes with highest CV will be selected and used for the subsequent analysis. Reducing the gene number affects the subsequent steps and if 50% filtering of genes were performed twice, only a quarter of genes in the initial datasets will remain. To save the adjusted (or reduced) expression dataset into tab-delimited text file, use the **Save current data** button (the saved expression profile is often useful for future analysis with other methods). After adjusting the data, you can return to the main menu with **Exit** button.

GSECA (*Gene set expression coherence analysis*)

3. Selecting gene set data

For GSECA analysis, the gene sets to be analyzed must be selected in the main window.



Gene set information must be provided in a separate, GSECA-compatible text files, which must be located within a designated folder (**/GSECA/Geneset/**). At startup, GSECA automatically reads the compatible gene set files in the folder and include the individual gene set IDs of proper files in gene set reference in a box (left). If a gene set ID is selected in the listbox, the description of the annotation type for corresponding gene set will be shown (right) and selected file will be used for the gene set-based clustering. The use of single gene set file or combined use of multiple gene set files, are both allowed. For example, if users are to investigate the function-versus-drug relationships, both the Drug-perturbation gene set and Functional gene set should be selected in the listbox.

Four kinds of gene sets (and corresponding four gene set files) are provided with GSECA package (**FunctionalGeneSet.txt**, **PromoterGeneSet.txt**, **miRNAGeneSet.txt**, **DrugGeneSet.txt**) and ready-to-use for the analysis. Additional gene set information can be readily included in this reference set. The example **FunctionGeneSet.txt** contents show how to make the GSECA-compatible gene set file.

GSECA (Gene set expression coherence analysis)

	‘Compatible’	Gene set type	Gene set No	Description
1	PathCluster-Compatible	Functional gene set	5025	5,024 gene sets sharing common
2	GO/hypoxanthine phosphoribosyltransferase activity		HPRI1	PRIFDC1
3	GO/calcitonin receptor activity	RCP9	CALCRL	CALCR
4	GO/COPI-coated vesicle	AP3B2		
5	GO/protein polymerization	K-ALPHA-1	TUBG1	TUBB3 TUBG2 TUBB2 TUBA4
6	GO/queuosine biosynthesis	QTRTD1	QTRI1	
7	GO/ATP metabolism	AK1	NADK	FLJ13052 AK5
8	GO/RNA processing	TRNT1	MGC23401	EXOSC6 BICD1 CUGBP2 RBM8A HNRPDL
9	GO/glycine N-methyltransferase activity	GNMT		
10	GO/growth hormone secretion	LITBP4	GAL	
11	GO/zeta DNA polymerase complex	REV3L		
12	GO/glucose 1-dehydrogenase activity	H6PD		

.....

The header line should include tab-separated strings (1) ‘Compatible’ (null string indicating that this file is GSECA-compatible) (2) gene set annotation type (ID to be listed in the listbox of main window) (3) gene set number (gene set number included in the file) (4) description (description to be displayed when the corresponding ID is selected). Below the header line, the actual gene set information should be listed in a line (tab-delimited); gene set ID – gene member 1 (symbol) – gene member 2 (symbol) - ... - gene member n (symbol). It must be also noted that current version of GSECA uses ‘gene symbol’ as common like between expression profiles and gene sets; the gene set file to be included in the dataset must have symbol-based annotation.

Note: MSigDB (as well as other kinds of public databases such as BIND, HPRD, etc.) provides well-defined gene set information useful for gene set-based analysis like GSECA. For example, MSigDB Ver2.5 c2 gene set include the gene set information collected from public literatures. To use this gene set, first download the file at the designated folder (/GSECA/Geneset/) and simply include a following header line at the top of text file (‘/’ represents ‘tab’): GSECA-Compatible/Literature gene set/1892/1,892 gene signatures from public literatures from MSigDB V2.5 public database (c2.v2.5.symbols.gmt). By restarting the program, the included gene set file will appear in the gene set list. Similarly, user-defined custom query gene sets can be easily included in the gene set reference of GSECA and used for extended biological insights.

GSECA (Gene set expression coherence analysis)

4. Expression coherence (EC) and expression variation (EV)

The screenshot shows the GSECA software interface with the following settings:

- Correlation metric:** Correlation (centered) - Pearson
- Pathways (gene sets) gene size:** : 5 ~ 200 genes
- Significance for expression coherence:**
 - Permutation test - 100000 permutation times
 - Z-test (fast but require normal distribution)
- Perform ANOVA for functional gene sets:**
 - Perform ANOVA for functional gene sets
 - Significance based on F-test
 - Permutation test 100000

Buttons: Calculate expression coherence and variance for functional gene sets, Exit

After loading expression dataset and selecting gene sets to be analyzed, expression coherence can be calculated.

- For **correlation metric**, four kinds of metrics can be selected for gene-to-gene distance or correlation measure (centered/uncentered, absolute/not correlation)
- It must be also determined for the **gene size of gene sets**. The default gene size is 5 – 200, thus; GSECA will process only the gene sets having 5 – 200 highly variable genes (this will reduce the actual gene sets to be analyzed). This limitation is necessary in that too small size of functional sets might lead to selection bias, and the functional annotation of large gene sets is indicative of general house-keeping function such as ‘transcription’ (as you expected, too large gene number also seriously increases the calculation time considering pairwise-combination of genes in calculating ‘Expression Coherence’).

The ‘**Expression Coherence (EC)**’ is used as a measure the extent of ‘how the expression of genes in a set are correlated with each other’. As distance measure, GSECA calculated the correlation for all possible pairs of genes, omitting self-comparisons. The mean value of all PCC was used as the “expression coherence” of the functional gene set (Pavlidis, P. et al (2002) Pac.Symp.Biocomput., 474-485).

- Significance level of expression coherence can be calculated by two strategies; permutation-based tests or Z-test. For permutation-based test, we randomly selected n genes (n is the number of genes for functional gene set under consideration) and calculated the expression coherences. This is iterated for predefined permutation times (default; 100,000), and the number of tests that acquired higher expression coherence than expression coherence of functional gene set under consideration, is determined as P value. As you expected, this might be considerably time-consuming; thus, GSECA

GSECA (Gene set expression coherence analysis)

provides another option, Z-test. For each functional gene set, Z values is determined considering the expression coherence (E_c), gene size of functional gene set (m), the mean (μ) and standard deviation (σ) of all PCC values in the genes in the array: $Z = (E_c - \mu) \times m^{1/2} / \sigma$. The Z value can be converted into P value. This simple method is must faster than permutation tests, providing an initial candidate lists of functional gene sets to be considered. However, Z-test works under the assumption of Gaussian distribution of PCC for possible pairs of genes, and it must be cautious in using Z-based significance tests.

<p>The 'Expression Variance (EV)' is used as a measure for the variation of gene members in individual gene sets using conventional ANOVA tests. The significance for EV can be directly calculated from F value or permutation-based tests.</p>

5. Functional Clustering

Annotated function	Size	EC score	Sig (EC)	EV score	Sig (EV)
GO/MHC class I protein complex	6	0.927145727097272	0	4.34617...	0.00059...
GO/antigen presentation	6	0.927145727097272	0	4.34617...	0.00039...
GO/ribosome biogenesis	17	0.599034480748813	0	2.26958...	0.00419...
KEGG/Porphyrin and chlorophyll metabol...	12	0.629894484250417	0	6.78369...	0
GO/RNA processing	31	0.283361590951543	0	2.18165...	0.00669...
GO/humoral immune response	10	0.654162227721269	0	2.65566...	0.00179...
GO/hemoglobin complex	6	0.747266377758757	0	2.48626...	0.00449...
GO/antigen presentation, exogenous ant...	8	0.885836995088591	0	12.7141...	0
GO/antigen processing, endogenous ant...	6	0.849220320909282	0	4.51224...	0.00019...
GenMAPP/Heme_Biosynthesis	8	0.923307824104252	0	6.94931...	0
GO/structural constituent of ribosome	68	0.273986775723646	0	6.43934...	0
GO/MHC class I receptor activity	6	0.927145727097272	0	4.34617...	0.00049...
GO/antigen presentation, endogenous a...	6	0.966097294877194	0	6.32841...	0
GO/MHC class II receptor activity	9	0.893673669584518	0	15.9615...	0
GO/antigen processing, exogenous anti...	9	0.56091836994103	9.99900009999E-05	8.79159...	0
GO/heme biosynthesis	10	0.593774661698444	0.0001999800019998	5.45262...	9.99900...
GO/methyltransferase activity	29	0.247565447791911	0.0001999800019998	3.45747...	0.00029...
GO/nucleosome assembly	36	0.247113787202311	0.0001999800019998	2.55317...	0.00199...
GO/spindle	6	0.734054589445166	0.0003999600039996	1.83699...	0.01644...
GO/translation factor activity, nucleic acid...	7	0.577664598495076	0.0004999500049995	2.87566...	0.00199...
GO/cytokinesis	6	0.632172616243764	0.0007999200079992	1.66684...	0.02662...
GO/DNA-directed RNA polymerase activity	18	0.300374472042276	0.0007999200079992	1.43866...	0.02926...
GO/tricarboxylic acid cycle	8	0.543046056589803	0.000999900009999	1.04189...	0.08844...

Use EC cutoff EC value P value
 Use EV cutoff EV value P value

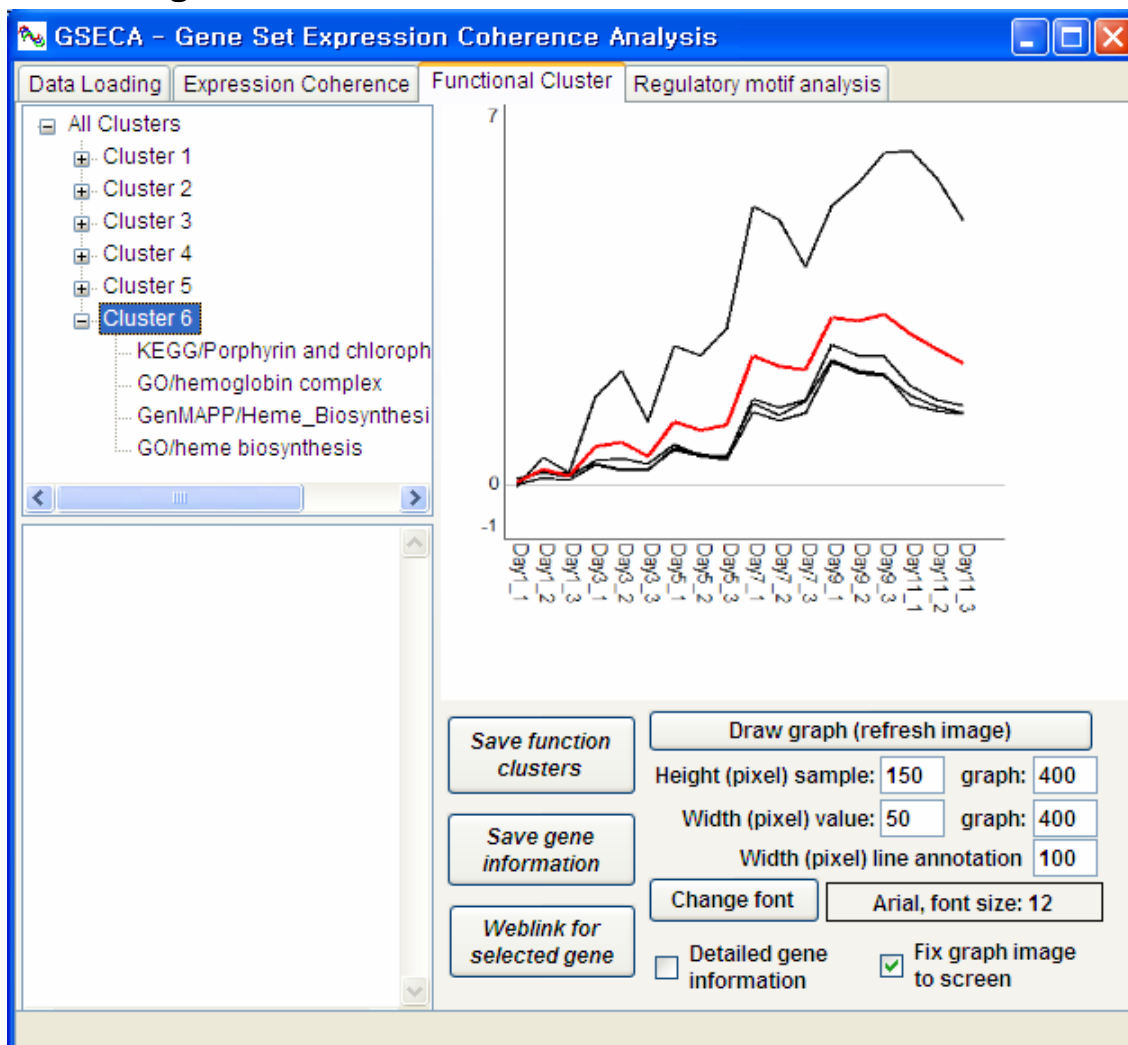
K-means cluster number:
 K-means clustering time:

Buttons: Perform gene set-based K-means clustering, Save current table, Save average expression of gene sets

Calculating the scores of EC/EV is finished. Check the results in Expression Coherence section.

- After calculating EC or EV (with significance), the results can be viewed in **Expression coherence** section. Main table shows the functional annotation of gene sets, size (gene number), EC score and significance of EC score (also EV score and EV significance). By **clicking a column header**, the results will be sorted with respect to the selected item values. The results can be saved using **Save current table** or **Save average expression of gene sets** (this will save the individual gene sets with mean expression values across the time- or condition-scales).
- For functions clustering, a subset of gene set must be selected. Four kinds of criteria can be used (absolute EC value or significance of EC +/- absolute EV value or significance of EV). For example, when **USE EC cutoff** and **P value** options are checked with the value of **0.0005**, the gene set whose EC-significance is less than 0.0005 will be selected and used for K-means clustering.
- After determining the cutoff, press the button **Perform gene set-based K-means clustering** to perform functional clustering. The k-means cluster number and iteration number can be defined before clustering.

6. Investigation of Functional Clusters



- Functional clusters are listed in a table. If **All clusters** are selected in the table, the mean expression patterns of individual functional clusters are shown as a graph in right panel. If a functional cluster is selected, the belonging gene sets are shown as well as a graph in the panel (red line; mean expression of the functional cluster, black line; mean expression patterns of individual gene sets). Switch the option **Fix graph image to screen** also to fit the graph image in the panel.

- In right panel, the image size level can be adjusted, determining the width/height of the graph and width/height of label annotations. The type and size of font for graph can be also determined.

- If a gene set belonging to a functional cluster is selected in functional cluster table, the gene members belonging to the selected gene set will be shown in the table below. If **Detailed gene information** option is on, it automatically reads an reference file and display a more detailed information on the gene members.

GSECA (Gene set expression coherence analysis)

- To save the results, use **Save function cluster** and **Save gene information** button. Weblink via external gene database of PubMed will be enabled if the **Weblink for selected gene** is pressed with a gene in the table is selected.

7. Identification of Enriched Motifs

Motif annotation	Significance
Functional cluster 3	
M00221[SREBP-1]	0.0844007759934286
M00222[Hand1/E47]	0.00513669835944566
M00804[E2A]	0.0246230604670376
M00807[EGR]	0.0526125007638779
M00672[TEF]	0.0731187734391761
M00257[RREB-1]	0.0806931238543219
M00761[p53_decamer]	0.0594826348968819
M00406[MEF-2]	0.0731187734391761
M00260[HLF]	0.00989057942485785
M00912[C_EBP]	0.0583110408414279

- The motif gene sets must be selected, each of which will be measured for the enriched with functional cluster(s) in terms of expression similarities. The functional clusters can be selected in **Target functional cluster**. The significance level under which, the enrichment and motif annotations will be displayed, can be also determined in **Show sets under significance P**. To determine the enrichment, two measures are provided either by using non-parametric GSEA method (genes in array are ordered according to the similarities or correlation with the seed expression values of functional cluster and enrichment score/significance is calculated by Kolmogorov-Smirnov tests and permutations tests) or by using hypergeometric distribution (predefined gene subset highly correlated with the seed values of functional cluster is a priori defined and enrichment/significance is defined using binomial tests).

- The pairwise enrichment test for motif gene sets (to identify motif synergy representing the combinatorial action between regulatory motifs) can be also identified using hypergeometric distribution. The used metric is based on previous literature (Elkon R et al, Genome Res 13:773, 2003). This is a highly challenging issue (related to the construction of higher-order of transcriptional regulatory network) and I am continuously looking for an advanced method or algorithm.