EnrichmentMap Documentation

Release 2.2

Ruth Isserlin, Mike Kucera, Christian Lopes

Feb 21, 2018

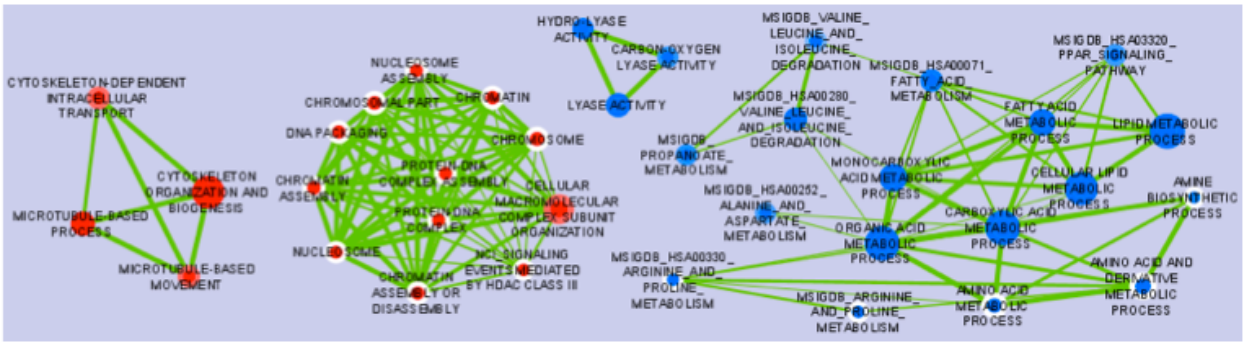
User Guide

The Enrichment Map Cytoscape Plugin allows you to visualize the results of gene-set enrichment as a network. It will operate on any generic enrichment results as well as specifically on Gene Set Enrichment Analysis (GSEA) results. Nodes represent gene-sets and edges represent mutual overlap; in this way, highly redundant gene-sets are grouped together as clusters, dramatically improving the capability to navigate and interpret enrichment results.

Gene-set enrichment is a data analysis technique taking as input:

1. A (ranked) gene list, from a genomic experiment
2. gene-sets, grouping genes on the basis of a-priori knowledge (e.g. Gene Ontology) or experimental data (e.g. co-expression modules)

and generating as output the list of enriched gene-sets, i.e. best sets that summarizing the gene-list. It is common to refer to gene-set enrichment as functional enrichment because functional categories (e.g. Gene Ontology) are commonly used as gene-sets.



CHAPTER 1

Cite EnrichmentMap

- **Enrichment Map: A Network-Based Method for Gene-Set Enrichment Visualization and Interpretation**
  Merico D, Isserlin R, Stueker O, Emili A, Bader GD
  PLoS One. 2010 Nov 15;5(11):e13984.
  PubMed Abstract - PDF

CHAPTER 2

Examples of Use

- **Functional impact of global rare copy number variation in autism spectrum disorders.**
  Pinto D, Pagnamenta AT, Klei L, Anney R, Merico D, Regan R, Conroy J, Magalhaes TR, Correia C, Abrahams BS, Almeida J, Bacchelli E, Bader GD, et al.
  Nature. 2010 Jun 9 (Epub ahead of print)
  PubMed Abstract - PDF
  Nature Blogs
- **Pathway analysis of dilated cardiomyopathy using global proteomic profiling and enrichment maps**
  Isserlin R, Merico D, Alikhani-Koupaei R, Gramolini A, Bader GD, Emili A.
  Proteomics 2010, March 10(6):1316-27
  Pubmed Abstract - PDF

CHAPTER 3

Papers Citing Enrichment Map

- Citations in Pubmed Central
- **Pathway analysis of expression data: deciphering functional building blocks of complex diseases.**
  Emmert-Streib F, Glazko GV.
  PLoS Comput Biol. 2011 May;7(5):e1002053.
  PubMed
- **Inflammasome is a central player in the induction of obesity and insulin resistance.**
  Stienstra R, van Diepen JA, Tack CJ, Zaki MH, van de Veerdonk FL, Perera D, Neale GA, Hooiveld GJ, Hijmans A, Vroegrijk I, van den Berg S, Romijn J, Rensen PC, Joosten LA, Netea MG, Kanneganti TD.
  Proc Natl Acad Sci U S A. 2011 Aug 29.
  PubMed
- **Delineation of Two Clinically and Molecularly Distinct Subgroups of Posterior Fossa Ependymoma**
  Witt H, Mack SC, Ryzhova M, Bender S, Sill M, Isserlin R, Benner A, Hielscher T, Milde T, Remke M, Jones DTW, Northcott PA, Garzia L, Bertrand KC, Wittmann A, Yao Y, Roberts SS, Massimi L, Van Meter T, Weiss WA, Gupta N, Grajkowska W, Lach B, Cho YJ, von Deimling A, Kulozik AE, Witt O, Bader GD, Hawkins CE, Tabori U, Guha A, Rutka JT, Lichter P, Korshunov A, Taylor MD, Pfister SM
  Cancer Cell, Volume 20, Issue 2, 143-157, 16 August 2011
  PubMed Abstract - PDF

CHAPTER 4

Report a Bug or a Problem

- please make sure you don't have formatting issues

  if you are still not sure how to handle formats, or you don't know what's the best suitable analysis for you, please send an email to: daniele[AT]merico[DOT]gmail.com

- please check what's your
  - plugin version and build (from the Cytoscape menu / Plugins / Enrichment Map / About)
  - Cytoscape version (from the Cytoscape menu / Cytoscape)
  - Operating System (e.g. Windows Vista)

and send your bug report to ruth[DOT]isserlin[AT]utoronto.ca

OR

report the bug to the Enrichment map issue tracker:

- go to https://github.com/BaderLab/EnrichmentMapApp
- click on "Issues"
- click on "New Issue"
- write a short description of the issue
- attached session file (.cys) file or example input files if applicable
- Make sure to enter plugin version and build, cytoscape version and operating system.
- click on "Submit new issue"

4.1 Installing

Install Cytoscape

- If you don't have Cytoscape please download and install the latest release from http://www.cytoscape.org/download.php.

Install EnrichmentMap

- Open Cytoscape
- In the main menu select **Apps > App Manager**
- In the App Manager select EnrichmentMap in the list of All Apps and click the Install button.

Alternatively EnrichmentMap can be installed from the App Store.

http://apps.cytoscape.org/apps/enrichmentmap

4.2 Quick Start Guide

4.2.1 Creating an Enrichment Map

You have a few different options:

- Load GSEA Results
- Load Generic Results
- Load David Results
- Load Bingo Results

The only difference between the above modes is the structure of the enrichment table(s). In either case, to use the plugin you will need the following files:

- file.gmt: gene-set to gene ID
- file.txt or .gct: expression matrix [OPTIONAL]
- file.txt or .xls: enrichment table(s)

Note: GSEA saves the enrichment table as a .xls file; however, these are not true Excel files, they are tab-separated text files with a modified extension; Enrichment Map does not work with "true" Excel .xls files.

If your enrichment results were generated from GSEA, you will just have to pick the right files from your results folder. If you have generated the enrichment results using another method, you will have to go to the Full User Guide, File Format section, and make sure that the file format complies with Enrichment Map requirements.

You can use the parameter defaults. For a more careful choice of the parameter settings, please go to *Tips on Parameter Choice*.

4.2.2 Graphical Mapping of Enrichment

- Nodes represent gene-sets.
- Node size represents how many genes are in the gene-set.
- Edges represent mutual overlap.
- Enrichment significance (p-value) is conveyed as node colour intensity.
- The enriched phenotype is conveyed by node colour hue.

Note: In standard two-class designs, where two phenotypes are compared (e.g. treated vs untreated) the colour hue conveys the enriched phenotype; this is equivalent to mapping enrichment in up- and down-regulated genes, if one of the two phenotypes is assumed as reference (e.g. untreated), and the other phenotype is the one of interest; in such a case, enriched in the phenotype of interest means up, and enrichment in the reference phenotype means down.

4.2.3 Exploring the Enrichment Map

- The "Parameters" tab in the "Results Panel" on the right side of the window contains a legend mapping the colours to the phenotypes and displaying the parameters used to create the map (cut-off values and data files).
- The "Network" tab in the "Control Panel" on the left lists all available networks in the current session and at the bottom has a overview of the current network which allows to easily navigate in a network even at higher zoom levels by dragging the blue rectangle (the current view) over the network.
- Clicking on a node (the circle that represents a gene set) will open the "EM Geneset Expression Viewer" tab in the "Data Panel" showing a heatmap of the expression values of all genes in the selected gene set.
- Clicking on an edge (the line between two nodes) will open the "EM Overlap Expression Viewer" tab in the "Data Panel" showing a heatmap of the expression values of all genes both gene sets that are connected by this edge have in common.
- If several nodes and edges are selected (e.g. by dragging a selection box around the desired gene sets) the "EM Geneset Expression Viewer" will show the union of all genes in the selected gene sets and the "EM Overlap Expression Viewer" will show only those genes that all selected gene sets have in common.

4.2.4 Advanced Tips

- With large networks and low zoom-levels Cytoscape automatically reduces the details (such as hiding the node labels and not showing the node borders). To override this mechanism click on "View / Show Graphics Details"
- The VizMapper and the Node- and Edge Attribute Browser open up a lot more visualization options like linking the label size to Enrichment Scores or p-values. Refer to the Cytoscape manual at www.cytoscape.org for more information.
- If you have used Genesets from GSEAs MSigDb, you can access additional informations for each gene set, by adding the a new property:

  (Edit / Preferences / Properties... / Add -> enter property name: nodelinkouturl.MSigDb -> enter property value: http://www.broad.mit.edu/gsea/msigdb/cards/%ID%.html -> [(./)] Make Current Cytoscape Properties default -> (OK)). Now you can right-click on a node and choose LinkOut/MSigDb to open the Database entry of the Geneset represented by that node in your Browser.

- When loading GSEA results there is no need to specify each file. Use the GSEA RPT file to auto-populate all the file fields in the EM interface. Check out: *RPT files*
- You can specify more lax p-value, q-value and coefficient threshold initially and fine tune them after the network is created by adjusting them through the p-value, q-value and coefficient tuners in the results panel. Check out: *The Results Panel*

4.3 File Formats

4.3.1 Gene sets file (GMT file)

- Each row of the geneset file represents one geneset and consists of:

```
geneset name (--tab--) description (--tab--) a list of tab-delimited genes
```

- The geneset names must be unique.
- The gene set file describes the genesets used for the analysis. These files can be obtained...
  1. directly downloading our monthly updated gene-set collections from Baderlab genesets collections. Description of sources and methods used to create collection can be found on the *EnrichmentMap Gene Sets* page.
  2. directly downloading gene-sets collected in the MSigDB
  3. converting gene annotations / pathways from public databases

---

Note: If you use MSigDB Gene Ontology gene-sets, please consider that they do not include all annotations, as an evidence code filter is applied; if you are interested in achieving maximum coverage, download the original annotations.

Note: if you are a R user, Bioconductor offers annotation packages such as `GO.db`, `org.Hs.eg.db`, `KEGG.db`

4.3.2 Expression Data file (GCT, TXT or RNK file) [OPTIONAL]

- The expression data can be loaded in three different formats: gct (GSEA file type), rnk (GSEA file type) or txt.
- The expression data serves two purposes:
  - Expression data is used by the Heatmap when clicking on nodes and edges in the Enrichment map so the expression of subsets of data can be viewed.
  - Gene sets are filtered based on the genes present in the expression file. For example, if Geneset X contains genes {1,2,3,4,5} but the expression file only contain expression value for genes {1,2,3} Geneset X will be represented as {1,2,3} in the Enrichment Map.
- Expression data is not required. In the absence of an expression file Enrichment map will create a dummy expression file to associate with the data set. The dummy expression gives an expression value of 1 for all the genes associated with the enriched genesets in the Enrichment map.

Note: If you are running a two dataset analysis with no expression files the genes for each dataset is calculated based on the enriched genesets. If a geneset is enriched in one dataset and not the other this could create different subsets of genes associated to each datasets and create multiple edges between genesets. To avoid this, create a fake expression file with the set of genes used for both analyses.

GCT (GSEA file type)

- GCT differs from TXT only because of two additional lines that are required at the top of the file.

- The GCT file contains two additional lines at the top of the file.
  - The first line contains `#1.2`.
  - The second line contains `the number of data rows (--tab--) the number of data columns`
  - The third line consists of column headings.

```
name (--tab--) description (--tab--) sample1 name (--tab--) sample2 name
...
```

  - Each line of expression file contains a:

```
name (--tab--) description (--tab--) list of tab delimited expression
values
```

Note: If the GCT file contains Probeset ID's as primary keys (e.g. as you had GSEA collapse your data file to gene symbols) you need to convert the gct file to use the same primary key as used in the gene sets file (GMT file). You have the following options:

- Use the GSEA desktop application: GSEA / Tools / Collapse Dataset
- Run this Python script *collapse_ExpressionMatrix.py* using the Chip platform file that was used by GSEA.

RNK (GSEA file type)

- RNK file is completely different from the GCT or TXT file. It represents a ranked list of genes containing only gene name and a rank or score.
- The first line contains column headings

  For example: `Gene Name (--tab--) Rank Name`
- Each line of RNK file contains:

```
name (--tab--) rank OR score
```

Additional Information on GSEA File Formats

TXT

- Basic file representing expression values for an experiment.
- The first line consists of column headings.

```
name (--tab--) description (--tab--) sample1 name (--tab--) sample2 name
...
```

- Each line of the expression file contains:

```
name (--tab--) description (--tab--) list of tab delimited expression
values
```

4.3.3 Enrichment Results Files

GSEA result files

- For each analysis GSEA produces two output files. One representing the enriched genesets in phenotype A and the other representing the enriched genesets in phenotype B.
- These files are usually named `gsea_report_for_phenotypeA.Gsea.########.xls` and `gsea_report_for_phenotypeB.Gsea.########.xls`
- The files should be loaded in as is and require no pre-processing.
- There is no need to worry about which Enrichment Results Text box to put the two files. The phenotype is specified by the sign of the ES score and is computed internally by the program.

Additional Information on GSEA File Formats

Generic results files

- The generic results file is a tab delimited file with enriched gene-sets and their corresponding p-values (and optionally, FDR corrections)
- The Generic Enrichment Results file needs:
  - gene-set ID (must match the gene-set ID in the GMT file)
  - gene-set name or description
  - p-value
  - FDR correction value
  - Phenotype: +1 or -1, to identify enrichment in up- and down-regulation, or, more in general, in either of the two phenotypes being compared in the two-class analysis
    - +1 maps to red
    - -1 maps to blue
  - gene list separated by commas

Note: Description and FDR columns can have empty or NA values, but the column and the column header must exist.

Note: If no value is provided under phenotype, Enrichment Map will assume there is only one phenotype, and will map enrichment p-values to red.

Examples of Generic Enrichment Result Files

DAVID Enrichment Result File

- Available only in v1.0 or higher
- The DAVID option expects a file as generated by the DAVID web interface.
- When using DAVID as the analysis type there is no requirement to enter either a gmt file or an expression file. Both are options if the user wishes to add them to the analysis.

- The DAVID Enrichment Result File is a file generated by the DAVID Functional Annotation Chart Report and consists of the following fields: **Important**: Make sure you are using CHART Report and NOT a Clustered Report.
  - Category (DAVID category, i.e. Interpro, sp_pir_keywords, . . .)
  - Term - Gene set name
  - Count - number of genes associated with this gene set
  - Percentage (gene associated with this gene set/total number of query genes)
  - P-value - modified Fisher Exact P-value
  - Genes - the list of genes from your query set that are annotated to this gene set.
  - List Total - number of genes in your query list mapped to any gene set in this ontology
  - Pop Hits - number of genes annotated to this gene set on the background list
  - Pop Total - number of genes on the background list mapped to any gene set in this ontology.
  - Fold enrichment
  - Bonferroni
  - Benjamini
  - FDR

Warning: In the absence of a gmt gene sets are constructed based on the field Genes in the DAVID output. This only considers the genes entered in your query set and not the genes in your background set. This will drastically affect the amount of overlap you see in the resulting Enrichment Map.

DAVID Tutorial

BiNGO Enrichment Result File

- Available only in v1.2 or higher
- The BiNGO option expects a file as generated by the BiNGO Cytsocape Plugin.
- When using BiNGO as the analysis type there is no requirement to enter either a gmt file or an expression file. Both are options if the user wishes to add them to the analysis.
- The BiNGO Enrichment Result File is a file generated by the BiNGO cytoscape plugin and consists of the following fields: **Important**: When running BiNGO make sure to check off "Check Box for saving data"
  - The first 20 lines of BiNGO output file list parameters used for the analysis and are ignored by the Enrichment map plugin
  - GO-ID - Gene set name
  - p-value - hypergeometric or binomial Exact P-value
  - corr p-value - corrected p-value
  - x - number of genes in your query list mapped to this gene-set
  - n - number of genes in the background list mapped to this gene-set
  - X - number of genes annotated to this gene set on the background list
  - N - number of genes on the background list mapped to any gene set in this ontology.

- Description - gene list description
- Genes - the list of genes from your query set that are annotated to this gene set.

Warning: In the absence of a gmt gene sets are constructed based on the field Genes in the BiNGO output. This only considers the genes entered in your query set and not the genes in your background set. This will drastically affect the amount of overlap you see in the resulting Enrichment Map.

DAVID Tutorial

RPT files

- A special trick for GSEA results, in any GSEA analysis an rpt file is created that specifies the location of all files (including the gmt, gct, results files, phenotype specification, and rank files).
- Any of the Fields under the dataset tab (Expression, Enrichment Results 1 or Enrichment Results 2) will accept an rpt file and populate GMT, Expression, Enrichment Results 1, Enrichment Results 2, Phenotypes, and Ranks the values for that dataset.
- A second rpt file can be loaded for dataset 2. It will give you a warning if the GMT file specified is different than the one specified in dataset 1. You will have the choice to use the GMT for data set 1, data set 2 or abort the second rpt load.
- An rpt file is a text file with following information (parameters surrounded by " ' ' " are those that EM uses):

```
'''producer_class'''    xtools.gsea.Gsea
'''producer_timestamp'''        1367261057110
param   collapse        false
param   '''cls'''       WHOLE_PATH_TO_FILE/EM_EstrogenMCF7_TestData/ES_NT.cls#ES24_
↪versus_NT24
param   plot_top_x      20
param   norm    meandiv
param   save_rnd_lists  false
param   median  false
param   num     100
param   scoring_scheme  weighted
param   make_sets       true
param   mode    Max_probe
param   '''gmx'''       WHOLE_PATH_TO_FILE/EM_EstrogenMCF7_TestData/Human_GO_
↪AllPathways_no_GO_iea_April_15_2013_symbol.gmt
param   gui     false
param   metric  Signal2Noise
param   '''rpt_label''' ES24vsNT24
param   help    false
param   order   descending
param   '''out'''       WHOLE_PATH_TO_FILE/EM_EstrogenMCF7_TestData
param   permute gene_set
param   rnd_type        no_balance
param   set_min 15
param   include_only_symbols    true
param   sort    real
param   rnd_seed        timestamp
param   nperm   1000
param   zip_report      false
param   set_max 500
param   '''res'''       WHOLE_PATH_TO_FILE/EM_EstrogenMCF7_TestData/MCF7_ExprMx_v2_
↪names.gct
```

```
file    WHOLE_PATH_TO_FILE/EM_EstrogenMCF7_TestData/ES24vsNT24.Gsea.1367261057110/
↪index.html
```

Parameters used by EM and their meaning:

1. producer_class - can be xtools.gsea.Gsea or xtools.gsea.GseaPreranked

- if xtools.gsea.Gsea:
  - get expression file from res parameter in rpt
  - get phenotype information from cls parameter in rot
- if xtools.gsea.GseaPreranked:
  - No expression file
  - use rnk as the expression file from rnk parameter in rot
  - set phenotypes to na_pos and na_neg.
  - NOTE: if you want to make using an rpt file easier for GSEAPreranked there are two additional parameters you can add to your rpt file manually that the rpt function will recognize.
  - To do less manual work while creating Enrichment Maps from pre-ranked GSEA, add the following optional parameters to your rpt file:

```
param(--tab--)phenotypes(--tab--){phenotype1}_versus_{phenotype2}
param(--tab--)expressionMatrix(--tab--){path_to_GCT_or_TXT_formated_
↪expression_matrix}
```

2. producer_timestamp - needed to find the directory with the results files
3. cls - path to class/phenotype file with information regarding the phenotypes:

- path/classfilename.cls#phenotype1_versus_phenotype2
- EM get the path to the class file and also pulls the phenotype1 and phenotype2 from the above field

4. gmx - path to gmt file
5. rpt_label - name of analysis and name of directory that GSEA creates to hold the results. Used when constructing the path to the results directory.
6. out - path to directory where GSEA will put the output directory. Used when constructing the path to the results directory.
7. res - path to expression file.

rpt Searches for the following results files:

```
Enrichment File 1 --> {out}(--File.separator--){rpt_label} + "." + {producer_class} +
↪"." + {producer_timestamp}(--File.separator--) "gsea_report_for_" + phenotype1 + "_
↪" + timestamp + ".xls"
Enrichment File 2 --> {out}(--File.separator--){rpt_label} + "." + {producer_class} +
↪"." + {producer_timestamp}(--File.separator--) "gsea_report_for_" + phenotype2 + "_
↪" + timestamp + ".xls"
Ranks File --> {out}(--File.separator--){rpt_label} + "." + {producer_class} + "." +
↪{producer_timestamp}(--File.separator--) "ranked_gene_list_" + phenotype1 + "_
↪versus_" + phenotype2 +"_" + timestamp + ".xls";
```

- If the enrichments and rank files are not found in the above path then EM replaces the out directory with the path to the given rpt file and tries again.

- If you would like to create your own rpt file for your own analysis pipeline you can put your own values for the above used parameters.
- If your analysis only creates one enrichment file you can make a copy of enrichment file 1 in the path of enrichment file 2 with no consequences for EM running.

EDB File (GSEA file type)

- Contained in the GSEA results folder is an edb folder. In the edb folder there are the following files:
  - results.edb
  - gene_sets.gmt
  - classfile.cls [Only in a GSEA analysis. Not in a GSEAPreranked analysis]
  - rankfile.rnk
- If you specify the results.edb file in any of the Fields under the dataset tab (Expression, Enrichment Results 1 or Enrichment Results 2) the gmt and enrichment files fields will be automatically populated.
- If you want to associate an expression file with the analysis it needs to be loaded manually as described here.

Note: The gene_sets.gmt file contained in the edb directory is filtered according to the expression file. If you are doing a two dataset analysis where the expression files are from different platforms or contain different sets of genes the edb gene_sets.gmt file can not be used as genes found in one analysis might be lacking in the other. In this case use the original gmt file (prior to GSEA filtering) and EM will filter each the gene sets separately according to each dataset.

Advanced Settings - Additional Files

- For each dataset there are additional parameters that the user can set but are not required. The advanced parameters include:
  - Ranks file - file specifying the ranks of the genes in the analysis
    - This file has the format specified in the above section - gene (–tab–) rank or score. See *RNK (GSEA file type)* for details.
  - Phenotypes (phenotype1 versus phenotype2)
    - By default the phenotypes are set to Up and Down but in the advanced setting mode the user can change these to any desired text.
- All of these fields are populated when the user loads the input files using the rpt option.

4.3.4 Examples of Generic Enrichment Result Files

Note: For readability the following examples have been formatted in a way, that the content of each column is properly aligned. In the actual files, replace each {tab} and it's surrounding SPACE-characters by one TAB-character. The files can be also easily created with any spreadsheet-program (e.g. Excel) and then saved in the "Tab Delimited Text" format.

Example with all possible columns

```
GO.ID      {tab} Description                      {tab} p.Val {tab} FDR  {tab}␣
↪Phenotype
GO:0000346 {tab} transcription export complex     {tab} 0.01  {tab} 0.02 {tab} +1
GO:0030904 {tab} retromer complex                 {tab} 0.05  {tab} 0.10 {tab} +1
GO:0008623 {tab} chromatin accessibility complex {tab} 0.05  {tab} 0.12 {tab} -1
GO:0046540 {tab} tri-snRNP complex                {tab} 0.01  {tab} 0.03 {tab} -1
...
```

Example without phenotype column

```
GO.ID      {tab} Description                      {tab} p.Val {tab} FDR
GO:0000346 {tab} transcription export complex     {tab} 0.01  {tab} 0.02
GO:0030904 {tab} retromer complex                 {tab} 0.05  {tab} 0.10
GO:0008623 {tab} chromatin accessibility complex {tab} 0.05  {tab} 0.12
GO:0046540 {tab} tri-snRNP complex                {tab} 0.01  {tab} 0.03
...
```

Example without FDR and phenotype

```
GO.ID      {tab} Description                      {tab} p.Val
GO:0000346 {tab} transcription export complex     {tab} 0.01
GO:0030904 {tab} retromer complex                 {tab} 0.05
GO:0008623 {tab} chromatin accessibility complex {tab} 0.05
GO:0046540 {tab} tri-snRNP complex                {tab} 0.01
...
```

Example without FDR but with phenotype

```
GO.ID      {tab} Description                      {tab} p.Val {tab} {tab} Phenotype
GO:0000346 {tab} transcription export complex     {tab} 0.01  {tab} {tab} +1
GO:0030904 {tab} retromer complex                 {tab} 0.05  {tab} {tab} +1
GO:0008623 {tab} chromatin accessibility complex {tab} 0.05  {tab} {tab} -1
GO:0046540 {tab} tri-snRNP complex                {tab} 0.01  {tab} {tab} -1
...
```

Example without Description, FDR and phenotype

```
GO.ID      {tab} {tab} p.Val {tab} {tab} Phenotype
GO:0000346 {tab} {tab} 0.01  {tab} {tab} +1
GO:0030904 {tab} {tab} 0.05  {tab} {tab} +1
GO:0008623 {tab} {tab} 0.05  {tab} {tab} -1
GO:0046540 {tab} {tab} 0.01  {tab} {tab} -1
...
```

Example with dummy-description and without FDR and phenotype

```
GO.ID      {tab} DESCR {tab} p.Val {tab} {tab} Phenotype
GO:0000346 {tab} NA    {tab} 0.01  {tab} {tab} +1
GO:0030904 {tab} NA    {tab} 0.05  {tab} {tab} +1
GO:0008623 {tab} NA    {tab} 0.05  {tab} {tab} -1
GO:0046540 {tab} NA    {tab} 0.01  {tab} {tab} -1
...
```

4.4 Tips on Parameter Choice

4.4.1 Node (Gene Set inclusion) Parameters

- Node specific parameters filter the gene sets included in the enrichment map.
- For a gene set to be included in the enrichment map it needs to pass both p-value and q-value thresholds.

P-value

- All gene sets with a p-value with the specified threshold or below are included in the map.

FDR Q-value

- All gene sets with a q-value with the specified threshold or below are included in the map.
- Depending on the type of analysis the FDR Q-value used for filtering genesets by EM is different
  - For GSEA the FDR Q-value used is 8th column in the gsea_results file and is called "FDR q-val".
  - For Generic the FDR Q-value used is 4th column in the generic results file.
  - For David the FDR Q-value used is 12th column in the david results file and is called "Benjamini".
  - For Bingo the FDR Q-value used is 3rd column in the Bingo results file and is called "core p-value"

4.4.2 Edge (Gene Set relationship) Parameters

- An edge represents the degree of gene overlap that exists between two gene sets, A and B.
- Edge specific parameters control the number of edges that are created in the enrichment map.
- Only one coefficient type can be chosen to filter the edges.

Jaccard Coefficient

```
Jaccard Coefficient = [size of (A intersect B)] / [size of (A union B)]
```

Overlap Coefficient

```
Overlap Coefficient = [size of (A intersect B)] / [size of (minimum( A , B))]
```

Combined Coefficient

- the combined coefficient is a merged version of the jacquard and overlap coefficients.
- the combined constant allows the user to modulate reciprocally the weights associated with the jacquard and overlap coefficients.
- When k = 0.5 the combined coefficient is the average between the jacquard and overlap.

```
Combined Constant = k
Combined Coefficient = (k * Overlap) + ((1-k) * Jaccard)
```

4.4.3 Tips on Parameter Choice

P-value and FDR Thresholds

GSEA can be used with two different significance estimation settings: gene-set permutation and phenotype permutation. Gene-set permutation was used for Enrichment Map application examples.

Gene-set Permutation

Here are different sets of thresholds you may consider for gene-set permutation:

Very permissive:

- p-value < 0.05
- FDR < 0.25

Moderately permissive:

- p-value < 0.01
- FDR < 0.1

Moderately conservative:

- p-value < 0.005
- FDR < 0.075

Conservative:

- p-value < 0.001
- FDR < 0.05

For high quality, high coverage transcriptomic data, the number of enriched terms at the very conservative threshold is usually 100-250 when using gene-set permutation.

Phenotype Permutation

Recommended:

- p-value < 0.05
- FDR < 0.25

In general, we recommend to use permissive thresholds only if your having a hard time finding any enriched terms.

Jaccard vs. Overlap Coefficient

- The Overlap Coefficient is recommended when relations are expected to occur between large-size and small-size gene-sets, as in the case of the Gene Ontology.
- The Jaccard Coefficient is recommended in the opposite case.
- When the gene-sets are about the same size, Jaccard is about the half of the Overlap Coefficient for gene-set pairs with a small intersection, whereas it is about the same as the Overlap Coefficient for gene-sets with large intersections.
- When using the Overlap Coefficient and the generated map has several large gene-sets excessively connected to many other gene-sets, we recommend switching to the Jaccard Coefficient.

Overlap Thresholds

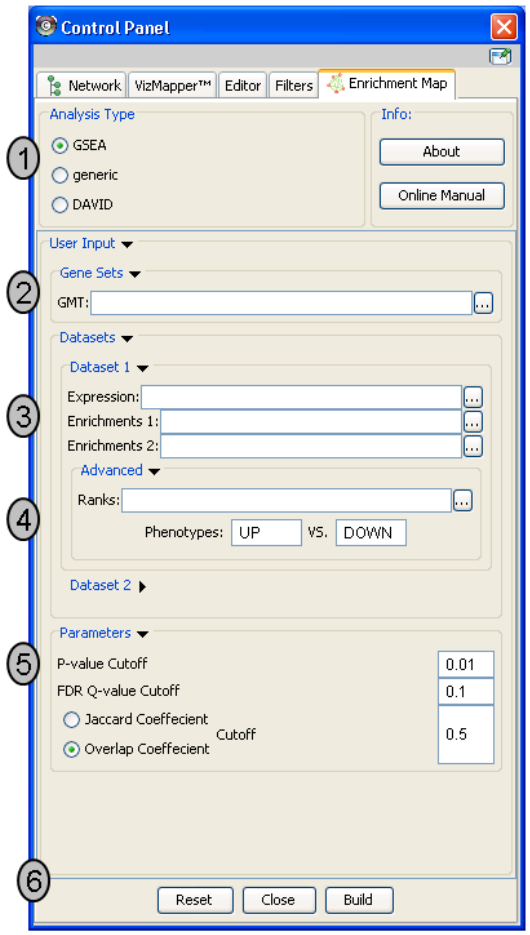
- 0.5 is moderately conservative, and is recommended for most of the analyses.
- 0.3 is permissive, and might result in a messier map.

Jaccard Thresholds

- 0.5 is very conservative
- 0.25 is moderately conservative

4.5 Interfaces

4.5.1 The Input Panel



1. **Analysis Type** - There are two distinct types of Enrichment map analyses, GSEA or Generic.
   - **GSEA** - takes as inputs the output files created in a GSEA analysis. File formats are specific to files created by GSEA. The main difference between this and generic is the number and format of the Enrichment results files. GSEA analysis always has two enrichment results files, one for each of the phenotypes compared.
   - **Generic** - takes as inputs the same file formats as a GSEA analysis except the Enrichment results file is a different format and there is only one enrichment file. Generic File description
   - **DAVID** - (implemented in v1.0 and higher) has no gmt or expression file requirement and takes as input enrichment result file as produced by DAVID David Enrichment Result File description
2. **Genesets** - path to gmt file describing genesets. User can browse hard drive to find file by pressing ... button.
3. **Dataset 1** - User can specify expression and enrichment files or alternatively, an rpt file which will populate all the fields in genesets,dataset # and advanced sections.
4. **Advanced** - Initially collapsed (expand by clicking on arrow head directly next to Advanced), users have the option of modifying the phenotype labels or loading gene rank files.
5. **Parameters** - User can specify p-value, fdr and overlap/jaccard cutoffs. Choosing Optimal parameter values

6. **Actions** - The user has three choices, Reset (clears input panel), Close (closes input panel), and Build Enrichment map (takes all parameters in panel and builds an Enrichment map)

4.5.2 The Data Panel (Expression Viewer)

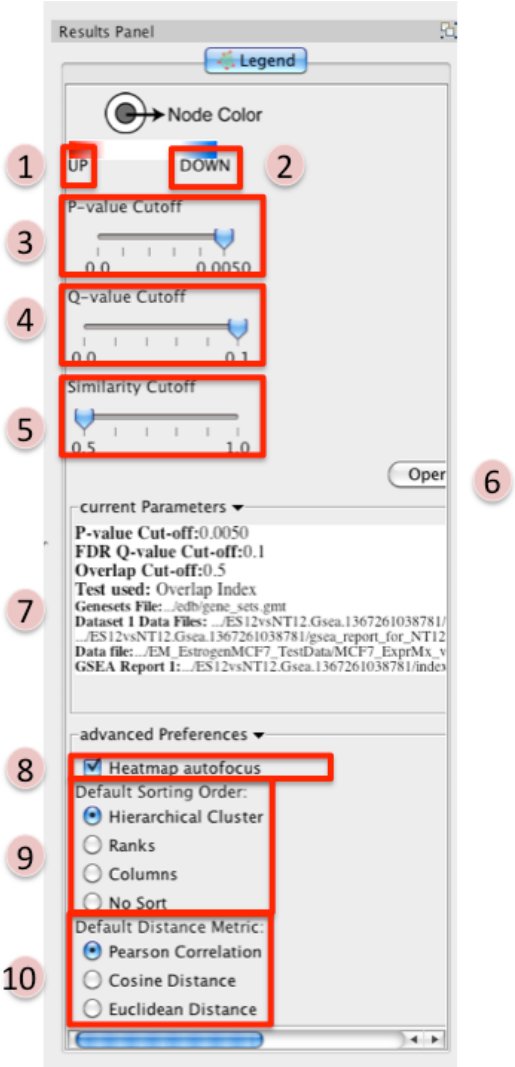
There are two different types of Expression Viewers, each is represented as a separate tab in data panel: EM Overlap and EM Gene set. The only difference between the two expression viewers is the set of gene listed.

1. **EM Overlap Expression Viewer** - shows the expression of genes in the overlap (intersection) of all the genesets selected
2. **EM Geneset Expression Viewer** - shows the expression of genes of the union of all the genesets selected.

- **Normalization**
  - Data as is - represents the data as it was loaded
  - Row Normalize Data - for each value in a row of expression the mean of the row is subtracted followed by division by the row's standard deviation.
  - Log Transform Data - takes the log of each expression value
- **Sorting**
  - Hierarchical cluster - as computed using Pearson correlation of the entire expression set.
  - If rank files for the data sets are provided at input they will show up as 'Dataset 1 Ranking' and 'Dataset 2 Ranking' and by selecting them the user will be able to sort the expression accordingly
    - if an expression value does not have a corresponding rank in the ranking file its expression does not appear in the heatmap.
  - Add Ranking ... - allows user to upload an additional rank file (in the appropriate format, as outlined in Rank file descriptions). There is no limit on the number of rank files that are uploaded. The user is required to give a name to the rank file.
- **Save Expression Set**
  - The user can save the subset of expression values currently being viewed in the expression viewer as txt file.

- **Export Expression Set (PDF)**
  - The user can save the expression heatmap currently being viewed in the expression viewer as pdf file. Unfortunately the pdf version is not perfect. The column titles are at the bottom instead of the top of the heatmap. Due to limitations with the current library being used we are unable to fix this in cityscape 2.8.3 but hope to have a better pdf representation in the EM version for Cytoscape 3.0.

4.5.3 The Results Panel



Reference panel containing legends, slider bars for the user to modify p-value and q-value cut-offs, parameters used for the analysis

1. Phenotype 1
2. Phenotype 2
3. **P-value Cutoff tuner** - Allows you to adjust the p-value threshold used to filter the gene sets.
   - By moving the slider to the left you can decrease the p-value threshold causing nodes (and their edges) to be removed from the network.
   - Moving the slider back to the right will restore the nodes (and their edges).

- You can NOT increase the p-value threshold above what you specified when you built the network.

4. **Q-value Cutoff tuner** - allows you to adjust the q-value threshold used to filter the gene sets.
    - By moving the slider to the left you can decrease the q-value threshold causing nodes (and their edges) to be removed from the network.
    - Moving the slider back to the right will restore the nodes (and their edges).
    - You can NOT increase the q-value threshold above what you specified when you built the network.
5. **Similarity Cutoff tuner** - allows you to adjust the similarity threshold used to filter the gene set overlaps (edges).
    - By moving the slider to the right you can increase the similarity threshold causing edges to be removed from the network.
    - Moving the slider back to the left will restore the edges.
    - You can NOT decrease the similarity threshold below what you specified when you built the network.
6. Button to launch index of GSEA results in a web browser.
7. List of parameters used to create the EM.
8. Heatmap Autofocus
    - selected by default
    - When you click on any node or edge in the network EM automatically updates the expression viewer and makes the focus of the Data panel the overlap expression viewer. When using other plugins in conjunction with EM this feature can get cumbersome.
    - To turn this off unselect “heatmap Autofocus”.
9. **Default Sorting order** - in the expression viewer genes can be sorted by Hierarchical clustering, Ranks, Columns, or No sort. To set the default change selection.
10. **Default Distance Metric** - for hierarchical clustering there are three available distance metrics that can be used to compute distances between genes. By default this is set of pearson correlation. Update this parameter if you which to use one of the other distance metrics.

4.5.4 PostAnalysis Input Panel

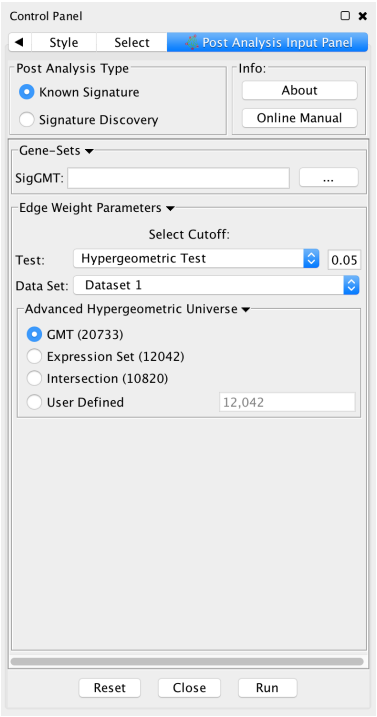
To access Post Analysis go to the menu path: Apps > Enrichment Map > Load Post Analysis Panel.

There are currently two types of Post Analysis Available: Known Signature and Signature Discovery. The contents of the panel will change depending on the type of analysis chosen. Known signature mode calculates post analysis edges for a small subset of known gene-sets. Signature discovery mode allows for filtering of large set of potential signatures to help uncover most likely sets.

The result of running Post Analysis is a new node for each signature gene set (yellow triangle) and edges from the signature gene set to each existing gene set when the similarity passes the cutoff test.

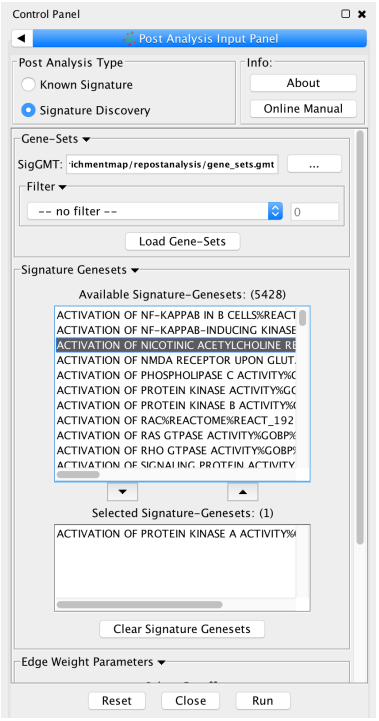
A new visual style is also created and applied to the network after post analysis runs. This visual style is the same as the enrichment map style but with a few additions. Signature edges are pink, signature nodes are yellow triangles, and edge width mapping is calculated differently.

Known Signature



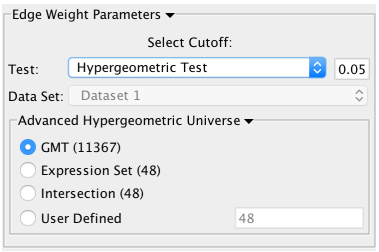
1. Post Analysis Type
    - Known Signature: Calculates the overlap between gene-sets of the current Enrichment Map and all the gene sets contained in the provided signature file.
2. Gene Sets
    - SigGMT: The gmt file with the signature-genesets. These will be compared against the gene-sets from the current Enrichment Map.
3. Edge Weight Parameters
    - Choose a method for generating an edge between a signature-geneset and an enrichment geneset. Described in detail below.
4. Actions:
    - Reset - clears input panel
    - Close - closes input panel
    - Run - takes all parameters in panel and performs the Post-Analysis

Signature Discovery



1. Post Analysis Type
    - Signature Discovery: Calculates the overlap between gene-sets of the current Enrichment Map and the selected genesets.
2. Gene-Sets
    - The gmt file with the signature-genesets.
    - Filter: Genesets from the gmt file that do not pass the filter test will not be loaded.
    - Load Gene-Sets: Press after the gmt file and filter have been chosen to load the signature-genesets.
3. Available Signature Genesets: Once the genesets have been loaded this box will contain a list of all the genesets in the SigGMT file (that passed the filter).
    - To highlight more than one geneset at at time hold the Shift, Command or Ctrl keys while clicking with the mouse.
4. Selected Signature Genesets: The analysis will be performed with all genesets in this list. Use the down- and up-buttons to move highlighted genesets from one list to the other.
5. Edge Weight Parameters: Choose a method for generating an edge between a signature-geneset and an enrichment geneset. Described in detail below.
6. Actions:
    - Reset - clears input panel
    - Close - closes input panel
    - Run - takes all parameters in panel and performs the Post-Analys

Edge Weight Parameters



1. Test: Select the type of statistical test to use for edge width.
2. Cutoff: Edges with a similarity value lower than the cutoff will not be created.
3. Data Set: If the enrichment map contains multiple data sets choose the one to use here.
4. Notes:
   - The results of the calculations will be available in the edge table after post analysis runs.
   - The edge "interaction type" will be sig.
   - The hypergeometric test is always calculated, even if it is not used for the cutoff. The results are made available in the edge table.
5. Available Tests
   - Hypergeometric Test is the probability (p-value) to find an overlap of k or more genes between a signature geneset and an enrichment geneset by chance.

$$P(K \geq k) = \sum_{K=k}^{n} f(K;N,m,n) = \sum_{K=k}^{n} \frac{\binom{m}{K}\binom{N-m}{n-K}}{\binom{N}{n}}$$

with:

k (successes in the sample) : size of the Overlap,
n (size of the sample) : size of the Signature geneset
m (total number of successes) : size of the Enrichment Geneset
N (total number of elements) : size of the union of all Enrichment Genesets

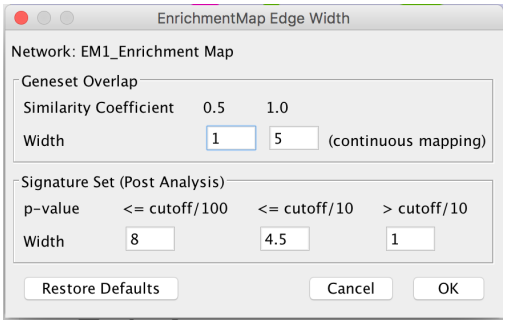
     - Advanced Hypergeometric Universe: Allows to choose the value for N.
       * GMT: all the genes in the original gmt file, Expression Set: number of genes in the expression set,
       * Intersection: number of genes in the intersection of the gmt file and expression set,
       * User Defined: manually enter a value).

- Overlap has at least X genes
  - The number of genes in the overlap between the enrichment map gene set and the signature gene set must be at least X for the edge to be created.

- Overlap is X percent of EM gs
  - The size of the overlap must be at least X percent of the size of the Enrichment Map gene set.
- Overlap is X percent of Sig gs
  - The size of the overlap must be at least X percent of the size of the Signature gene set.
- Mann-Whitney (Two-sided, one-sided greater, one-sided less)
  - Note: The Mann-Whitney test requires ranks. It will not be available if the enrichment map was created without ranks.
  - Calculates the p-value using the Mann-Whitney U test where the first sample is the ranks in the overlap and the second sample is all of the ranks in the expression set.

Edge Width

When you create an Enrichment Map network a visual style is created. The default edge width property is a continuous mapping to the "similarity_coefficient" column. After running post-analysis the rules for calculating edge width become more complicated. Edge width for edges between enrichment sets are still based on the "similarity_coefficient" column, but edges between signature sets and enrichment sets are based on the statistical test used for cutoff. Currently Cytoscape does not provide a visual mapping that is capable of "if-else" logic. In order to work around this limitation, the width of the edges is calculated by EnrichmentMap and put into a new column called "EM1_edge_width_formula". Then the edge width property uses a continuous mapping to that column.



- Edge Width Dialog
  - In order to change the rules used to calculate edge width go to the menu path: Apps > EnrichmentMap > Post Analysis Edge Width.
  - Geneset Overlap: Set the end points of the contiuous mapping for edge width for edges between enrichment sets.
  - Signature Set: Set the edge width value for signature set edges that are less than cutoff/100, <= cutoff/10 and > cutoff/10.
  - Click OK to recalculate the values in the "EM1_edge_width_formula" column.

4.6 Attributes

4.6.1 Node Attributes

For each Enrichment map created the following attributes are created for each node:

EM#_Name The gene set name.

EM#_Formatted_name A wrapped version of the gene set name so it is easy to visualize. Note: This is the default label of the node but some users find it easier to arrange the network when the name is not wrapped. If this is the case in the vizmapper the user can switch the label mapping from EM#_formatted_name to EM#_name.

EM#_GS_DESCR The gene set description (as specified in the second column of the gmt file).

EM#_Genes The list of genes that are part of this gene set.

Additionally there are attributes created for each dataset (a different set for each dataset if using two dataset mode):

EM#_pvalue_dataset(1 or 2) Gene set p-value, as specified in GSEA enrichment result file.

EM#_qvalue_dataset(1 or 2) Gene set q-value, as specified in GSEA enrichment result file.

EM#_Colouring_dataset(1 or 2) Enrichment map parameter calculated using the formula 1-pvalue multiplied by the sign of the ES score (if using GSEA mode) or the phenotype (if using the Generic mode)

GSEA specific attributes (these attributes are not populated when creating an enrichment map using the generic mode).

EM#_ES_dataset(1 or 2) Enrichment score, as specified in GSEA enrichment result file.

EM#_NS_dataset(1 or 2) Normalized Enrichment score, as specified in GSEA enrichment result file.

EM#_fwer_dataset(1 or 2) Family-wise error score, as specified in GSEA enrichment result file.

4.6.2 Edge Attributes

For each Enrichment map created the following attributes are created for each edge:

EM#_Overlap_size The number of genes associated with the overlap of the two genesets that this edge connects.

EM#_Overlap_genes The names of the genes that are associated with the overlap of the two genesets that this edge connects.

EM#_similarity_coefficient The calculated coefficient for this edge.

4.7 Additional Features

4.7.1 Launch Enrichment Map from the command line

Requirements:

- Enrichment Map v1.3 or higher
- Commandtool App - available from Cytsocape App store.
- GSEA results in an edb directory

Command tool can be used from:

- the command line

```
java -Xmx1G -jar "{path_to_cytoscape_dir}\cytoscape.jar" -p "{path_to_plugin_dir}
↪\plugins" -S "{path_to_script_file}"``
```

- cytoscape command window
  - Plugins->commandtool–>Command Window...

- script file
  - Plugins->commandtool–>Run Script...

Command Options:

```
enrichmentmap build: Build an enrichment map from GSEA results (in an edb directory)
  Arguments:
    [edbdir=value] --> REQUIRED
    [expressionfile=value] --> OPTIONAL
    [overlap=value] --> OPTIONAL
    [pvalue=value] --> OPTIONAL
    [qvalue=value] --> OPTIONAL
    [similaritymetric=value] --> OPTIONAL
    [combinedconstant=value] --> OPTIONAL
```

Example Command (for command window):

```
enrichmentmap build edbdir="{path_to_edb_directory}" pvalue=0.01 qvalue=0.1 overlap=0.
↪5
                    similaritymetric="jaccard" expressionfile="{path_to_expression_
↪file}"
```

4.7.2 Calculate Gene set relationships

To analyze the relationships that exists between genesets in the absence of an enrichment analysis an Enrichment map can be built with just the gene set definition file.

In the input panel specify only a gmt file and click on build.

Warning: This task requires a lot of memory. In a normal enrichment analysis we compute similarities only for the gene sets that pass the thresholds in addition to constraining the genes of the gene sets by the given expression set which drastically decreases the computations of similarity required. The smaller the gmt file the less memory required.

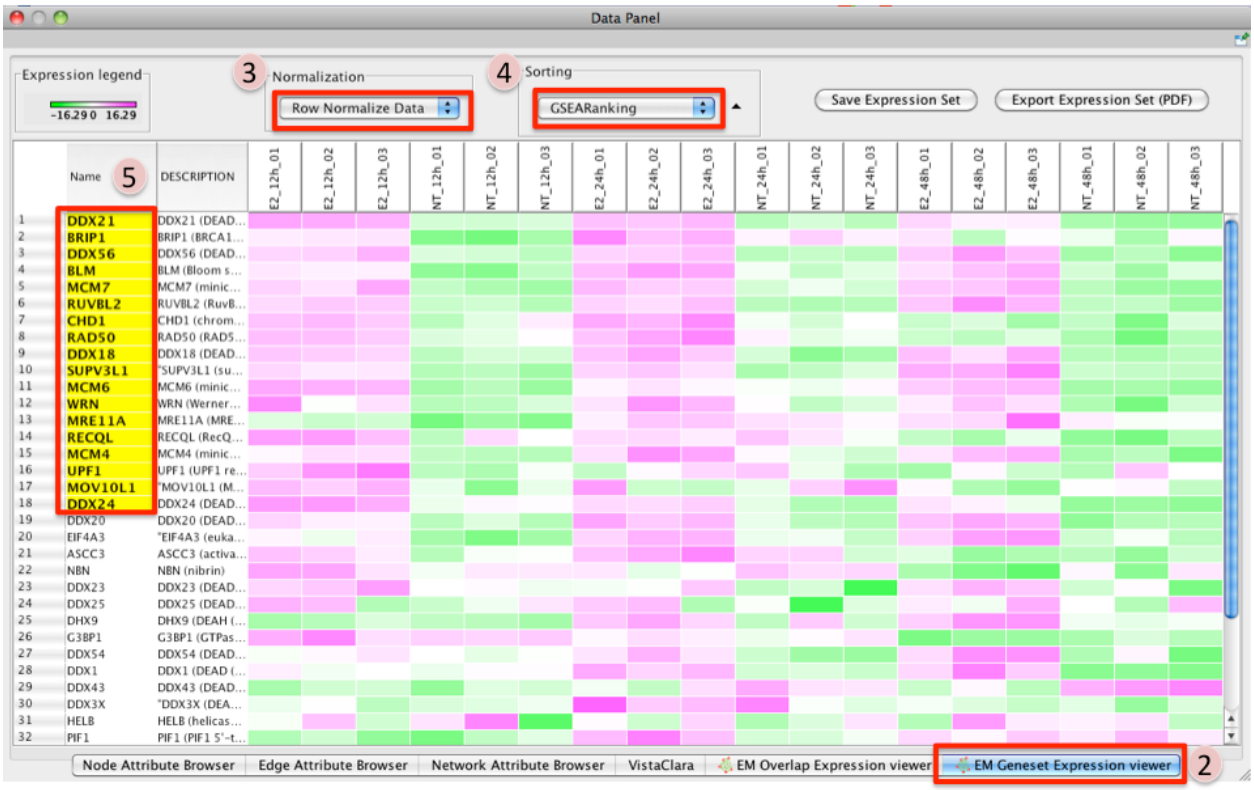
4.7.3 GSEA Leading Edge Functionality

For every gene set that is tested for significance using GSEA there is a set of proteins in that gene set defined as the Leading Edge. According to GSEA the leading edge is:

> "the subset of members that contribute most to the ES. For a positive ES, the leading edge subset is the set of members that appear in the ranked list prior to the peak score. For a negative ES, it is the set of members that appear subsequent to the peak score."

In essence, the leading edge is the set of genes that contribute most to the enrichment of the gene set.

For Enrichment Map, leading edge information is extracted from the gsea enrichment results files from the column denoted as Rank at Max. Rank at max is the rank of the gene where the ES score has the maximal value, i.e. the peak ES score. Everything with a better rank than the rank at max is part of the leading edge set.



1. To access GSEA leading edge information click on an individual Node. Leading edge information is currently only available when looking at a single gene set.
2. In the Data Panel the expression profile for the selected gene set should appear in the EM GenesetExpression viewer tab.
3. Change the Normalization to your desired metric.
4. Change the Sorting method to GSEARanking.
5. Genes part of the leading edge are highlighted in Yellow.

4.7.4 Customizing Defaults with Cytoscape Properties

The Enrichment Map Plugin evaluates a number of Cytoscape Properties with which a user can define some customized default values. These can be added and changed with the Cytoscape Preferences Editor (Edit / Preferences / Properties...) or by directly editing the file `cytoscape.props` within the `.cytoscape` folder in the User's HOME directory.

Supported Cytoscape Properties:

EnrichmentMap.default_pvalue

- Default P-value cutoff for Building Enrichment Maps
- Default Value: 0.05
- valid Values: float >0.0, <1.0

EnrichmentMap.default_qvalue

- Default Q-value cutoff for Building Enrichment Maps
- Default Value: 0.25

- valid Values: float >0.0, <1.0

EnrichmentMap.default_overlap

- Default Overlap coefficient cutoff for Building Enrichment Maps
- Default Value: 0.50
- valid Values: float >0.0, <1.0

EnrichmentMap.default_jaccard

- Default Jaccard coefficient cutoff for Building Enrichment Maps
- Default Value: 0.25
- valid Values: float >0.0, <1.0

EnrichmentMap.default_overlap_metric

- Default choice of similarity metric for Building Enrichment Maps
- Default Value: Jaccard
- valid Values: Jaccard, Overlap

EnrichmentMap.default_sort_method

- Set the default sorting in the legend/parameters panel to Hierarchical Clustering,
- Ranks (default the first rank file, if no ranks then it is no sort), Column (default is the first column) or no sort.
- Default Value: Hierarchical Cluster
- valid Values: Hierarchical Cluster, Ranks, Columns, No Sort

EnrichmentMap.hieracical_clusteting_theshold

- Threshold for the maximum number of Genes before a dialogue opens to confirm if clustering should be performed.
- Default Value: 1000
- valid Values: Integer

nodelinkouturl.MSigDb.GSEA Gene sets

- LinkOut URL for MSigDb.GESA Gene sets.
- Default Value: http://www.broad.mit.edu/gsea/msigdb/cards/%ID%.html
- valid Values: URL

EnrichmentMap.disable_heatmap_autofocus

- Flag to override the automatic focus on the Heatmap once a Node or Edge is selected.
- Default Value: FALSE
- valid Values: TRUE, FALSE

4.8 EnrichmentMap Gene Sets

EnrichmentMap is a Cytoscape plugin developed in the Baderlab to help visualize, navigate and analyze functional enrichment results as generated from programs such as Gene Set Enrichment Analysis(GSEA), BiNGO, or David. Some enrichment programs, such as GSEA, allow the user to search against their own gene set database. As annotation

(gene set) sources are regularly updated as new information is discovered we set up an automated system to update our gene set collections so we are always using the most up-to-date annotations.

If you use these gene sets, please cite our Enrichment Map paper.

Note: (January 2016) With the latest build of pathways we have removed KEGG from the main compilation set of pathways. If you would like to include KEGG in your analysis the sets are located in the *misc/* directory and can be appended to your gmt file.

Note: (April 2012) Genesets files from December 2011, January 2012, Februrary 2012, and March 2012 had an error in the up-propagation of GO. Up-propagation only followed the *is-a* relationship and did not follow the *part-of* relationship which translates into missing annotations. This primarily effects genesets in GO cellular compartment.

4.8.1 Summary

Gene Set Files can be downloaded from: Baderlab genesets collections

Enrichment Map Gene Sets are a set of Gene Set files in GMT format (compatible with GSEA) updated monthly from original source locations available with:

- Entrez gene ids
- UniProt accessions
- Gene symbols

The GMT File format contains one Gene Set per line. Each line contains:

- Name (tab) Description (tab) Gene (tab) Gene (tab) . . .
- In our format:
  - Name = Gene Set Name % Gene Set Source % Gene Set Source identifier
    - Example –> ATP-dependent protein binding%GO%GO:0043008 OR arginine biosynthesis IV%HUMANCYC%ARGININE-SYN4-PWY
  - Description = Gene Set Name
    - Example –> ATP-dependent protein binding OR arginine biosynthesis IV
  - Gene = identified by one of the three possible identifiers (Entrez gene id, UniProt accession or gene symbols)
  - **IMPORTANT NOTE**: Originally we used the "|" to separate information in the Name field but we came across issues with this separator in GSEA so we changed to "%". The "%" was used as of the December 2011 build.

In the main directory (current_release/Human/symbol) there are 5 primary files to choose from:

Human_GO_AllPathways_with_GO_iea_{Date}_{ID}.gmt Contains genesets from all 3 divisions of GO (biological process, molecular function, cellular component) including annotations that have evidence code IEA (inferred from electronic annotation), ND (no biological data available), and RCA (inferred from reviewed computational analysis) and all pathway resources.

Human_GO_AllPathways_no_GO_iea_{Date}_{ID}.gmt Contains genesets from all 3 divisions of GO (biological process, molecular function, cellular component) excluding annotations that have evidence code IEA (inferred from electronic annotation), ND (no biological data available), and RCA (inferred from reviewed computational analysis) and all pathway resources.

Human_GOBP_AllPathways_with_GO_iea_{Date}_{ID}.gmt Contains only genesets from GO biological process including annotations that have evidence code IEA (inferred from electronic annotation), ND (no biological data available), and RCA (inferred from reviewed computational analysis) and all pathway resources.

Human_GOBP_AllPathways_no_GO_iea_{Date}_{ID}.gmt (recommended file) Contains only genesets from GO biological process excluding annotations that have evidence code IEA (inferred from electronic annotation), ND (no biological data available), and RCA (inferred from reviewed computational analysis) and all pathway resources.

Human_AllPathways_{Date}_{ID}.gmt Contains only genesets from all pathways resources.

4.8.2 Current Stats

Human http://download.baderlab.org/EM_Genesets/current_release/Human/Summary_Geneset_Counts.txt

Mouse http://download.baderlab.org/EM_Genesets/current_release/Mouse/Summary_Geneset_Counts.txt

4.8.3 Sources

Human

Source	File Type	ID extracted	Frequency source is updated	Number of pathways	File Origin
KEGG	GMT	Symbol	static as of July 1, 2011	236	KEGG ftp site (July 2011)
MSigDB - C2	GMT	Entrez gene	sporadically	Biocarta - 217, Other - 47	manual download
NCI	BioPAX	Entrez gene	sporadically	219 pathways	scripted download of zipped release
Institute of Bioinformatics (IOB)	BioPAX	Entrez gene	sporadically	35 pathways - 10 are the same as CellMap, 1 is the same as NetPath	received directly from IOB - static (July 2011)
NetPath (IOB)	BioPAX	Entrez gene	static	25 pathways - 12 are cancer pathways (10 are CellMap), 13 are immunity pathways	scripted download of files numbered 1-25
HumanCyc	BioPAX	UniProt	updated periodically	249 Pathways	scripted download of zipped release
Reactome	BioPAX	UniProt	updated release	1117 pathways (release 37)	scripted download of zipped release
GO	GAF	Uniprot	released once a month	13034 no GO IEA, 15181 with GO IEA	scripted download from EBI ftp site
MSigDB - C3	GMT	Entrez gene	sporadically	221 miRs, 616 TFs	manual download
Panther	BioPAX	UniProt	updated periodically	307 Pathways	scripted download of biopax archive

Mouse

Source	File Type	ID extracted	Frequency source is updated	Number of pathways	File Origin
Reactome	BioPAX	UniProt	updated release	946 pathways (release 37)	scripted download of zipped release
GO	GAF	MGI	released once a month	14563 no GO IEA, 15041 with GO IEA	scripted download from MGI ftp site
KEGG	GMT	Entrez gene	static as of July 1, 2011	236	translated from Human using Homologene
MSigDB - C2	GMT	Entrez gene	sporadically	total 880: Kegg - 186, Reactome - 430, Biocarta - 217, Other - 47	translated from Human using Homologene
NCI	GMT	Entrez gene	sporadically	219 pathways	translated from Human using Homologene
Institute of Bioinformatics (IOB)	GMT	Entrez gene	sporadically	35 pathways - 10 are the same as CellMap, 1 is the same as NetPath	translated from Human using Homologene
NetPath (IOB)	GMT	Entrez gene	static	25 pathways - 12 are cancer pathways (10 are CellMap), 13 are immunity pathways	translated from Human using Homologene
HumanCyc	GMT	Entrez gene	updated periodically	249 Pathways	translated from Human using Homologene
Panther	BioPAX	UniProt	updated periodically	307 Pathways	translated from Human using Homologene

4.8.4 Specialty Gene Sets

The bulk of our genesets are groupings from similar biological processes, pathways and functional annotations but there are a few additional collections of sets that we don't group with them. They include:

miRs

- Sets consisting of all the targets for a given microRNA.
- miR genesets are retrieved from Msigdb c3 collection.

Transcription Factors

- Sets consisting of all the targets for a given transcription factor.
- TF genesets are retrieved from Msigdb c3 collection.

Disease Phenotype

- Sets consisting of all known proteins associated with the given disease.
- Disease phenotype genesets are retrieved from the Human phenotype ontology. Genes associated with a particular disease are annotated to it. In addition, in the same style as the Gene Ontology, the relationship between each disease is stored creating an ontology of diseases. Annotations are up-propagated to related disease terms.

Drugs Targets

- Sets consisting of all the known or predicted targets for a given drug.
- Drug target information is retrieved from drugbank. Drugbank is a resource containing 6711 drug entries including 1447 FDA-approved small molecule drugs, 131 FDA-approved biotech (protein/peptide) drugs, 85 nutraceuticals and 5080 experimental drugs. In addition to the compilation of all drugs contained in drugbank geneset files are also created for each of the defined drug categories including approved, experimental, illicit, nutraceutical, and small molecule.

4.8.5 File Structure

< > denotes directory

- <Release> - directory is named according to date sets were updated.
  - <Species>
    - <Identifier> - (either Entrez gene, UniProt, Gene symbol)
      - <GO>
      - BP = biological process
      - MF = molecular function
      - CC = Cellular component
      - All = BP + MF + CC
      - no_GO_IEA - indicates that the file excludes GO annotations with evidence codes - 'IEA' (inferred from electronic annotation), 'ND' (No biological data available), 'RCA' (inferred from reviewed computational analysis)
      - with_GO_IEA - indicates that the file includes GO annotations with evidence codes - 'IEA' (inferred from electronic annotation), 'ND' (No biological data available), 'RCA' (inferred from reviewed computational analysis)
      - <Pathways>
      - <miRs>
      - <TF>
      - <Disease phenotypes>
- In each <identifier> directory There are amalgamated Gene Set files:
  - AllPathways - contains all pathway sources in the Pathways directory
  - GOPathways - contains all GO (MF, BP, CC) and all Pathway sources in the Pathways directory.

4.8.6 Creating customized Gene Sets

Download the desired gene set files you would like to use in your customized set and concatenate the files.

For example, to combine Human_IOB_Entrezgene.gmt Human_NetPath_Entrezgene.gmt, you can use the following linux command:

```
cat Human_IOB_Entrezgene.gmt Human_NetPath_Entrezgene.gmt > MyCustomizedSet.gmt
```

4.8.7 References

1. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M.
   KEGG for integration and interpretation of large-scale molecular data sets.
   Nucleic Acids Res. 2011 Nov 10. PMID: 22080510
   Pubmed.

2. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP.
   Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.
   Proc Natl Acad Sci U S A. 2005 Oct 25;102(43):15545-50. PMID: 16199517
   Pubmed.

3. Schaefer CF, Anthony K, Krupa S, Buchoff J, Day M, Hannay T, Buetow KH.
   PID: the Pathway Interaction Database.
   Nucleic Acids Res. 2009 Jan;37(Database issue):D674-9. PMID: 18832364
   Pubmed.

4. Kandasamy K, et al
   NetPath: a public resource of curated signal transduction pathways.
   Genome Biol. 2010 Jan 12;11(1):R3. PMID: 20067622
   Pubmed.

5. Romero P, Wagg J, Green ML, Kaiser D, Krummenacker M, Karp PD.
   Computational prediction of human metabolic pathways from the complete human genome.
   Genome Biol. 2005;6(1):R2. Epub 2004 Dec 22. PMID: 15642094
   Pubmed.

6. Croft D, O'Kelly G, Wu G, Haw R, Gillespie M, Matthews L, Caudy M, Garapati P, Gopinath G, Jassal B, Jupe S, Kalatskaya I, Mahajan S, May B, Ndegwa N, Schmidt E, Shamovsky V, Yung C, Birney E, Hermjakob H, D'Eustachio P, Stein L.
   Reactome: a database of reactions, pathways and biological processes
   Nucleic Acids Res. 2011 Jan;39(Database issue):D691-7. PMID: 21067998
   Pubmed.

7. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G.
   Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.
   Nat Genet. 2000 May;25(1):25-9. PMID: 10802651
   Pubmed.

8. Mi H, Lazareva-Ulitsky B, Loo R, Kejariwal A, Vandergriff J, Rabkin S, Guo N, Muruganujan A, Doremieux O, Campbell MJ, Kitano H, Thomas PD.
   The PANTHER database of protein families, subfamilies, functions and pathways.

Nucleic Acids Res. 2005 Jan 1;33(Database issue):D284-8. PubMed PMID: 15608197
Pubmed.

4.9 Tutorials

4.9.1 GSEA Tutorial

This quick tutorial will guide you through the generation of an Enrichment Map for an analysis performed using GSEA Gene Set Enrichment Analysis.

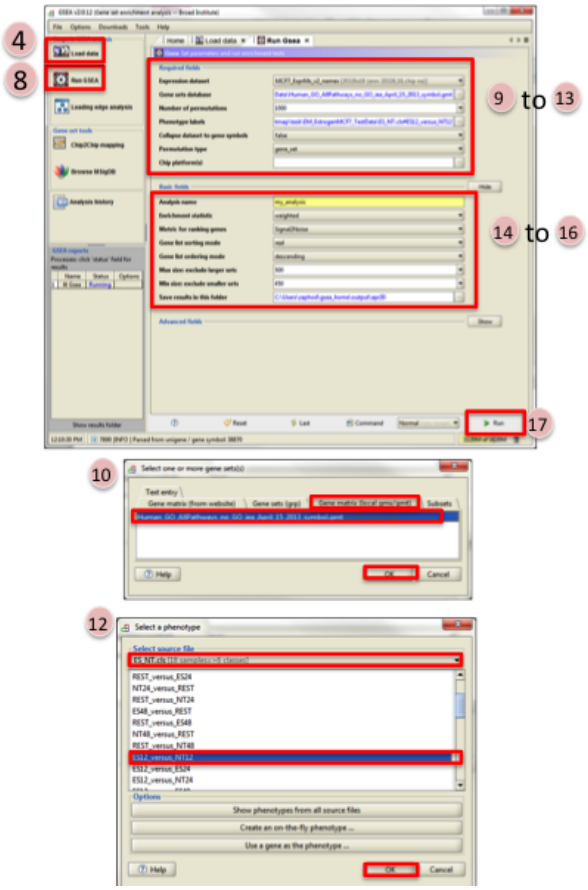
Files

Download the test data: `GSEATutorial.zip`

Description of the files contained in the GSEATutorial folder:

- `ES_NT.cls` Phenotype definition for expression file required by GSEA.
- `MCF_ExpMX_v2_names.gct` Expression File - Estrogen treatment, Official Gene Name as key. - Data for 12hr,24hr and 48hr.
- `Human_GO_AllPathways_no_GO_iea_April_15_2013_symbol.gmt` Gene set definition file.

Step 1: Generate GSEA output files



1. GO to GSEA website
2. Click on Downloads in the page header.
    - From the *javaGSEA Desktop Application* right click on *Launch with 1 Gb memory*.
    - Click on "Save Target as. . . " and save shortcut to your desktop or your folder of choice so you can launch GSEA for your analysis without having to navigate to it through your web browser.
3. Double click on GSEA icon you created.
4. Click on *Load data* in left panel.
5. Click on *Browse for files. . .* in newly opened Load data panel.
6. Navigate to directory where you stored tutorial test set files. Select raw expression (.gct) file, sample class file(.cls) and gene set (.gmt) file. Click on *Open*.
7. Wait until confirmation box appears indicating that all files loaded successfully. Click on *Ok*.
8. Click on *Run GSEA* in left panel.
9. Select the *Expression dataset*:
    - Click on the arrow next to the *Expression dataset* text box.
    - Select the expression set you wish to run the analysis on (MCF7_ExprMx_v2_names.gct).
10. Select the *Gene Set Database*:

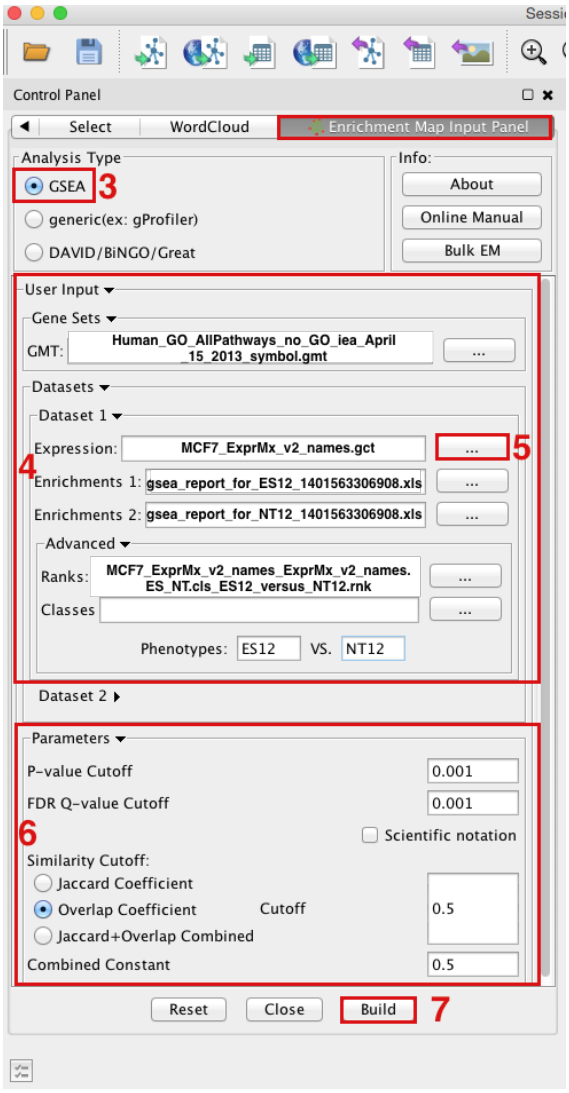
- Click on next to the text box of Gene Set Database.
- Click on *Gene Matrix (local gmx/gmt)* tab.
- Select gmt file Human_GO_AllPathways_no_GO_iea_April_15_2013_symbo.gmt and click on *Ok*.

11. Select the *Phenotype labels* file
    - Click on ... next to the text box of *Phenotype* labels.
    - Make sure *Select source file* is set to ES_NT.cls.
    - Select *ES12_versus_NT12* and click on *Ok*.
12. Click on the down arrow next to the text box for *Collapse dataset to gene symbols*. Select *false*.
13. Click on the down arrow next to the text box for *Permutation type*. Select gene_set.
14. Click on *Show* next to *Basic fields*.
15. Click in text box next to *Analysis name* and rename (example:estrogen_treatment_12hr_gsea_enrichment_results).
16. Click on ... next to *Save results in this folder* text box. Navigate to the folder where you wish to save the results (preferably the same directory where all the input files have been saved).
17. Click on *Run* in the bottom right corner.

Note: Repeat steps for the 24hrs time-point but use ES24_versus_NT24 phenotype labels in step 11 instead and in step 15 change the Analysis name (example:estrogen_treatment_24hr_gsea_enrichment_results).

Step 2: Generate Enrichment Map with GSEA Output

GSEA results produced by Step 1: `EM_EstrogenMCF7_GSEAresults.zip`



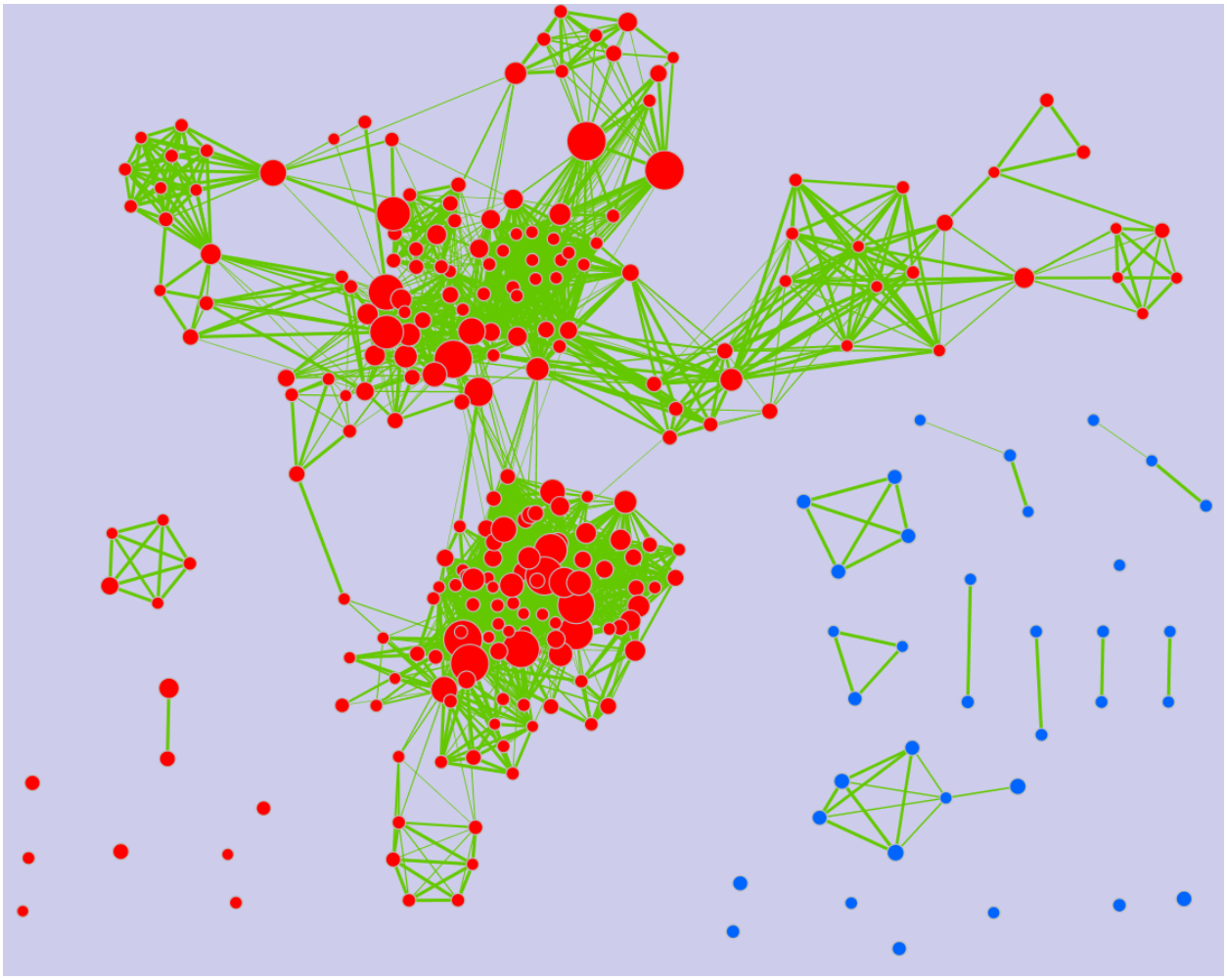
1. Open Cytoscape
2. Locate the Apps tab and select *EnrichmentMap* > *Create Enrichment Map*
3. Make sure the Analysis Type is set to GSEA
4. **OPTION 1** - Manually load all files Please select the following files by clicking on the respective (. . .) button and selecting the file in the Dialog:
    - Gene Sets / GMT:
    - *Human_GO_AllPathways_no_GO_iea _April_15_2013_symbol.gmt* (can be found in directory where you extracted the files downloaded in GSEATutorial.zip)
    - Dataset 1 / Expression: *MCF7_ExprMx_v2_names.gct* (can be found in directory where you extracted the files downloaded in GSEATutorial.zip)
    - Dataset 1 / Enrichments 1: *gsea_report_for_ES12_##############.xls* (can be found in directory where you put the GSEA results specified in Part 1, step 15)
    - Dataset 1 / Enrichments 2: *gsea_report_for_NT12_##############.xls* (can be found in directory where you put the GSEA results specified in Part 1, step 15)
    - Click on *Advanced* to expand the panel

- Dataset 1 / Ranks: *MCF7_ExprMx_v2_names_ExprMx_v2_names.ES_NT.cls_ES12_versus_NT12.rnk* (OPTIONAL) (can be found in directory where you put the GSEA results specified in Part 1, step 15)
- Dataset 1 / Phenotypes 1: *ES12* VS *NT12* (OPTIONAL)

5. **OPTION 2** - Populate all fields using GSEA rpt file
    - Dataset 1 / Expression : *ES12vsNT12.Gsea.##############.rpt* (can be found in directory where you put the GSEA results specified in Part 1, step 15)
    - NOTE: If you are populating the fields using a rpt file and any of the file names appear in red font then the file EM needs was not found. This can happen if you move your GSEA results folders around after they have been created. For the missing file follow step 5 and re-populate the effected fields.
6. Tune Parameters
    - P-value cut-off: *0.001*
    - Q-value cut-off: *0.05*
    - Overlap coefficient cut-off: *0.5*
7. Build Enrichment Map
8. Go to View, and activate Show Graphics Details

Step 3: Examining Results

Example EM session - Estrogen treatment vs no treatment at 12hr `ES12_EM_example.cys`

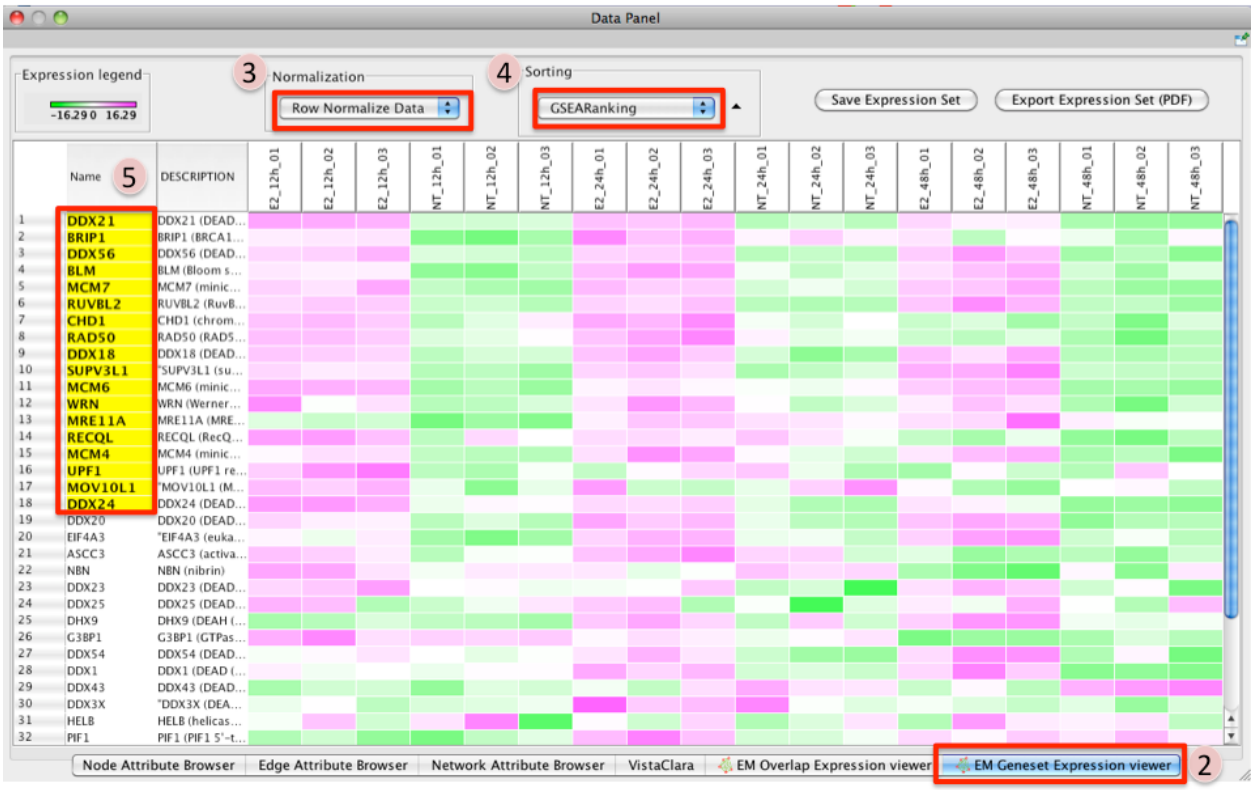


Legend:

1. Node (inner circle) size corresponds to the number of genes in dataset 1 within the geneset
2. Colour of the node (inner circle) corresponds to the significance of the geneset for dataset 1.
3. Edge size corresponds to the number of genes that overlap between the two connected genesets. Green edges correspond to both datasets when it is the only colour edge. When there are two different edge colours, green corresponds to dataset 1 and blue corresponds to dataset 2.

GSEA Leading Edge Information:

1. Click on a node (gene set) in the Enrichment map.
2. In the Data Panel, expression profile of all genes included in the selected gene-set should appear in the Heat Map (nodes) viewer tab
3. Change the Normalization to your desired metric.
4. Change the Sorting method to GSEARanking.
5. Genes part of the leading edge are highlighted in yellow.



Note: Leading edge information is currently only available when looking at a single gene set.

More Tutorials

For more detailed tutorials check out:

- Visualizing gene-set enrichment results using the Cytoscape plug-in enrichment map.
  Merico D, Isserlin R, Bader GD.
  Methods Mol Biol. 2011;781:257-77. doi: 10.1007/978-1-61779-276-2_12.
- Global proteomic profiling and enrichment maps of dilated cardiomyopathy.
  Isserlin R, Merico D, Emili A.
  Methods Mol Biol. 2013;1005:53-66. doi: 10.1007/978-1-62703-386-2_5.

4.9.2 GSEA Tutorial - GSEA Interface

This quick tutorial will guide you through the generation of an Enrichment Map for an analysis performed using GSEA Gene Set Enrichment Analysis directly from the GSEA interface.

Files

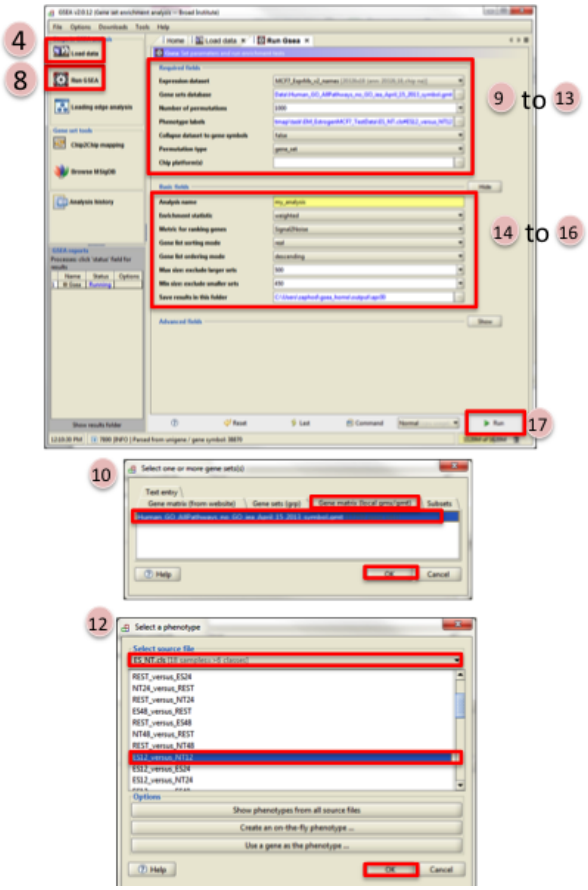
Download the test data: `GSEATutorial.zip`

Description of the files contained in the GSEATutorial folder:

- `ES_NT.cls` Phenotype definition for expression file required by GSEA.

- `MCF_ExpMX_v2_names.gct` Expression File - Estrogen treatment, Official Gene Name as key. - Data for 12hr,24hr and 48hr.
- `Human_GO_AllPathways_no_GO_iea_April_15_2013_symbol.gmt` Gene set definition file.

Step 1: Generate GSEA output files

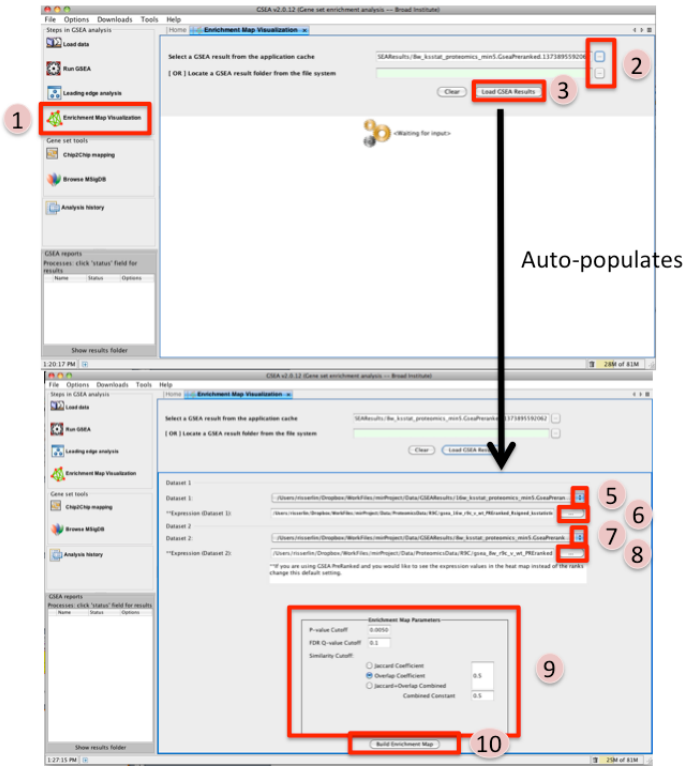


1. GO to GSEA website
2. Click on Downloads in the page header.
   - From the *javaGSEA Desktop Application* right click on *Launch with 1 Gb memory*.
   - Click on "Save Target as…" and save shortcut to your desktop or your folder of choice so you can launch GSEA for your analysis without having to navigate to it through your web browser.
3. Double click on GSEA icon you created.
4. Click on *Load data* in left panel.
5. Click on *Browse for files…* in newly opened Load data panel.
6. Navigate to directory where you stored tutorial test set files. Select raw expression (.gct) file, sample class file(.cls) and gene set (.gmt) file. Click on *Open*.
7. Wait until confirmation box appears indicating that all files loaded successfully. Click on *Ok*.
8. Click on *Run GSEA* in left panel.
9. Select the *Expression dataset*:

- Click on the arrow next to the *Expression dataset* text box.
- Select the expression set you wish to run the analysis on (MCF7_ExprMx_v2_names.gct).

10. Select the *Gene Set Database*:
    - Click on next to the text box of Gene Set Database.
    - Click on *Gene Matrix (local gmx/gmt)* tab.
    - Select gmt file Human_GO_AllPathways_no_GO_iea_April_15_2013_symbo.gmt and click on *Ok*.
11. Select the *Phenotype labels* file
    - Click on . . . next to the text box of *Phenotype* labels.
    - Make sure *Select source file* is set to ES_NT.cls.
    - Select *ES12_versus_NT12* and click on *Ok*.
12. Click on the down arrow next to the text box for *Collapse dataset to gene symbols*. Select *false*.
13. Click on the down arrow next to the text box for *Permutation type*. Select gene_set.
14. Click on *Show* next to *Basic fields*.
15. Click in text box next to *Analysis name* and rename (example:estrogen_treatment_12hr_gsea_enrichment_results).
16. Click on . . . next to *Save results in this folder* text box. Navigate to the folder where you wish to save the results (preferably the same directory where all the input files have been saved).
17. Click on *Run* in the bottom right corner.

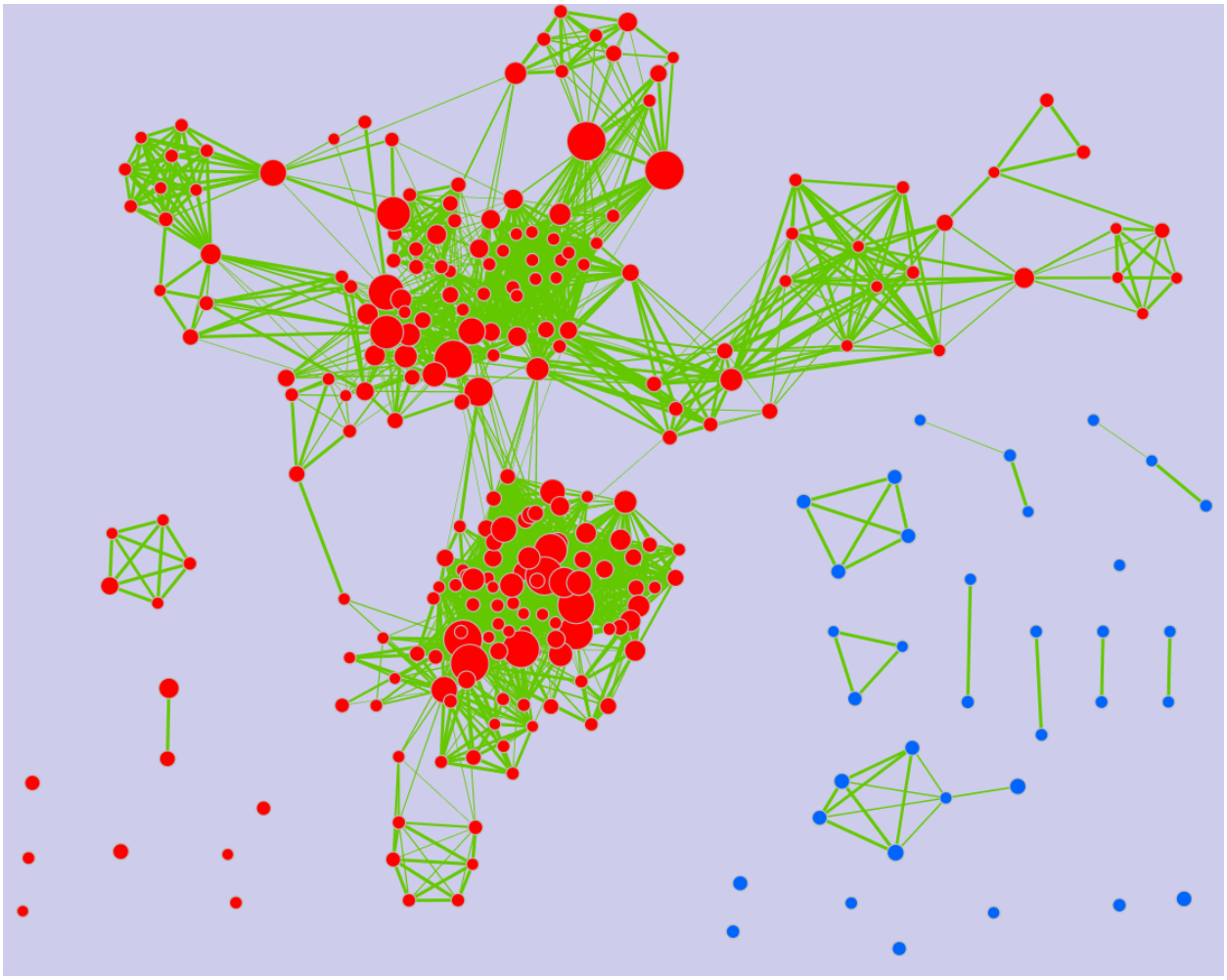
Step 2: Generate Enrichment Map



1. Once GSEA has completed click in the Steps in GSEA analysis panel click on Enrichment Map Visualization. When you click on the Steps in GSEA analysis cytoscape 3.3 or higher should automatically be launched. It will take a few seconds for cytoscape to load. If you try and create a network before it is finished initializing GSEA will not be able to communicate with cytoscape yet. *(In Cytoscape 3.3.0 the Cyotscape splash screen will not disappear until it has finished initializing but in later version you will be able to configure the Enrichment map parameters within GSEA while cytoscape is initializing).*
2. Navigate to the analysis or analyses you wish to create an enrichment map for. There are two ways to do this:
    - click on the ... next to Select a GSEA result from the application cache. From the list select the set of analyses to load. (hold down CTRL or COMMAND key to select multiple analyses). Click on OK. B. click on the ... next to [OR] Locate a GSEA result folder from the file system. Navigate to the GSEA result directory you wish to use. Click on OK.
3. Click on Load GSEA Results. *NOTE: if the GSEA analysis was performed on a dataset that was not collapsed it will take a few seconds for the information to load as it needs to collapse it first.*
4. Bottom frame will appear, auto-populating file fields according to the GSEA results folders specified. Multiple GSEA folders can be specified. If more than one folder is specified bottom frame will contain specifications for two datasets. If only one directory is specified then only one dataset will be accommodated.
5. The user can specify which of the datasets to use as dataset 1 by selecting dataset from drop down list. Selecting a different dataset will automatically populate *Expression(Dataset 1)* with the corresponding expression file.
6. If you have conducted a GSEA analysis on a Preranked list of genes but wish to see the original expression file associated with your enrichment map update the path to the expression file next to Expression (Dataset 1).
7. The user can specify which of the datasets to use as dataset 2 by selecting dataset from drop down list. Selecting a different dataset will automatically populate Expression(Dataset 2) with the corresponding expression file.
8. If you have conducted a GSEA analysis on a Preranked list of genes but wish to see the original expression file associated with your enrichment map update the path to the expression file next to Expression (Dataset 2).
9. Tune Parameters. Check out *Tips on Parameter Choice* (check out tips for choosing parameters)
    - P-value cut-off: *0.001*
    - Q-value cut-off: *0.05*
    - Overlap coefficient cut-off: *0.5*
10. Click on Build Enrichment Map
11. Cytoscape should launch and create your Enrichment map.
12. Go to View, and activate Show Graphics Details

Step 3: Examining Results

Example EM session - Estrogen treatment vs no treatment at 12hr `ES12_EM_example.cys`

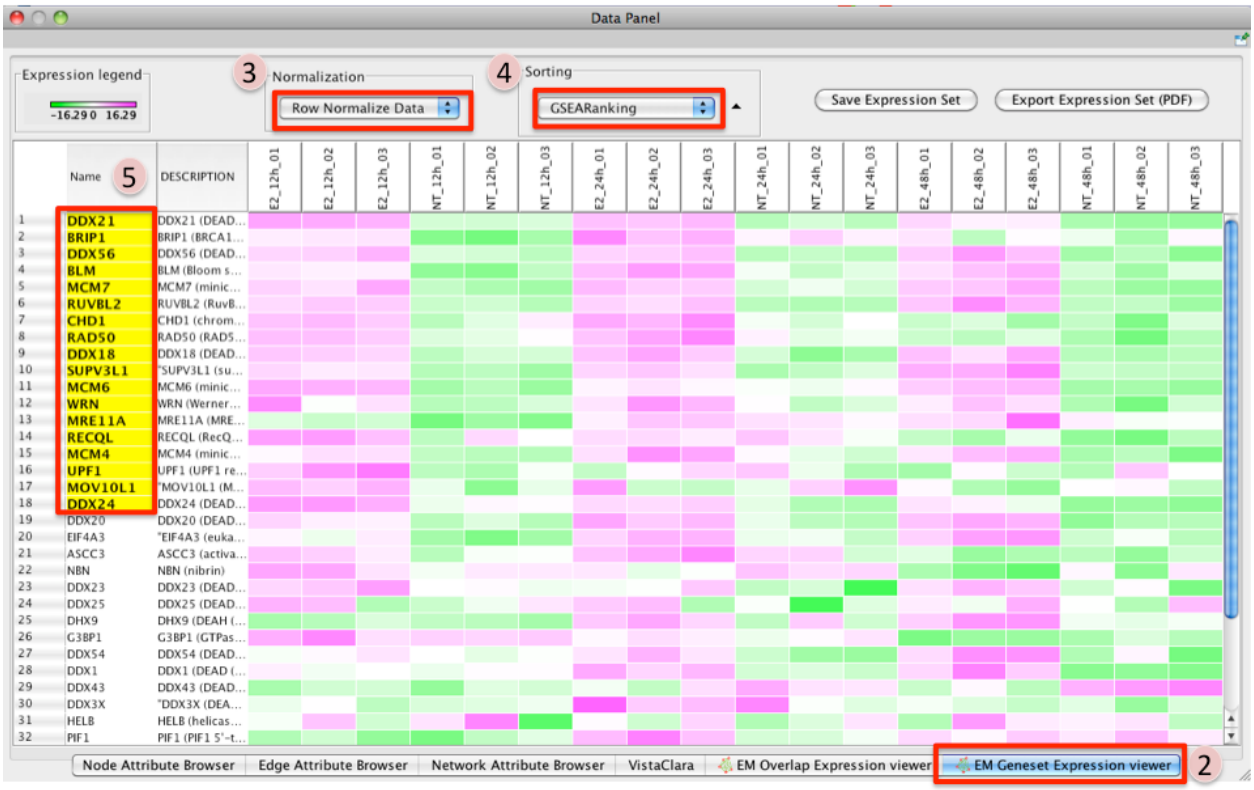


Legend:

1. Node (inner circle) size corresponds to the number of genes in dataset 1 within the geneset
2. Colour of the node (inner circle) corresponds to the significance of the geneset for dataset 1.
3. Edge size corresponds to the number of genes that overlap between the two connected genesets. Green edges correspond to both datasets when it is the only colour edge. When there are two different edge colours, green corresponds to dataset 1 and blue corresponds to dataset 2.

GSEA Leading Edge Information:

1. Click on a node (gene set) in the Enrichment map.
2. In the Data Panel, expression profile of all genes included in the selected gene-set should appear in the Heat Map (nodes) viewer tab
3. Change the Normalization to your desired metric.
4. Change the Sorting method to GSEARanking.
5. Genes part of the leading edge are highlighted in yellow.



Note: Leading edge information is currently only available when looking at a single gene set.

More Tutorials

For more detailed tutorials check out:

- Visualizing gene-set enrichment results using the Cytoscape plug-in enrichment map.
  Merico D, Isserlin R, Bader GD.
  Methods Mol Biol. 2011;781:257-77. doi: 10.1007/978-1-61779-276-2_12.
- Global proteomic profiling and enrichment maps of dilated cardiomyopathy.
  Isserlin R, Merico D, Emili A.
  Methods Mol Biol. 2013;1005:53-66. doi: 10.1007/978-1-62703-386-2_5.

4.9.3 DAVID Tutorial

This quick tutorial will guide you through the generation of an Enrichment Map for an analysis performed using DAVID Functional Annotation Tool,

Files

Download the test data: `DavidTutorial.zip`

Description of the files contained in the DavidTutorial folder:

- `12hr_topgenes.txt` List of top genes expressed in Estrogen dataset at 12hr - Official Gene Symbol.

- `24hr_topgenes.txt` List of top genes expressed in Estrogen dataset at 24hr - Official Gene Symbol.
- `12hr_David_Output.txt` Estrogen treatment - 12hr DAVID result chart - Screen shot of where to get DAVID output chart
- `24hr_David_Output.txt` Estrogen treatment - 24hr DAVID result chart
- `Estrogen_expression_file.txt` Expression File - Estrogen treatment, Official Gene Name as key.

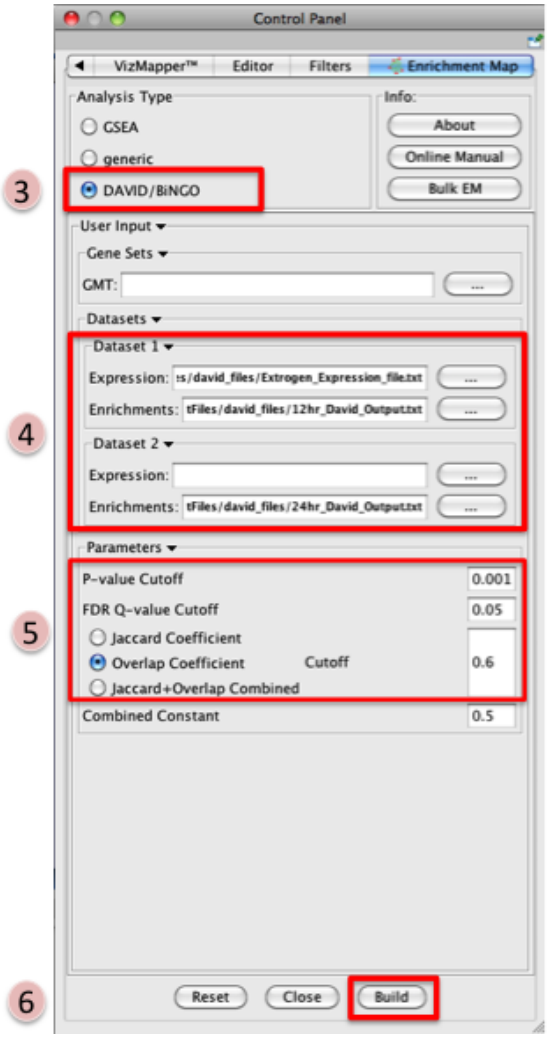
Step 1: Generate DAVID output files

- GO to DAVID website - http://david.abcc.ncifcrf.gov/
- Select and copy all genes in the tutorial file 12hr_topgenes.txt
- In Upload tab of DAVID interface Paste genes in text box marked Step 1: Enter Gene list
- Select Official Gene Symbol in Step 2: Select Identifier
- Select Gene list in Step 3: Select List Type
- Click Submit list
- Select species: Homo sapiens
- Click Functional Annotation Chart - Screen shot of where to get DAVID output chart
- Download file - This is the file you can use in Enrichment Map (Dataset 1 or 2:Enrichment Results)

Note: Repeat these steps for the 24hrs time-point and the file 24hr_topgenes.txt

---

Step 2: Generate Enrichment Map with DAVID Output

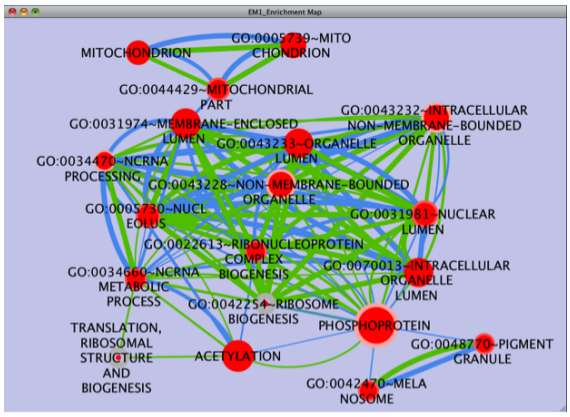


- Open Cytoscape
- Click on Plugins / Enrichment Maps / Load Enrichment Results
- Make sure the Analysis Type is set to DAVID/BiNGO
- Please select the following files by clicking on the respective (. . .) button and selecting the file in the Dialog:
  - NO GMT file is required for DAVID Analysis
  - Dataset 1 / Expression: Estrogen_expression_file.txt (OPTIONAL)
  - Dataset 1 / Enrichments: 12hr_David_Output.txt
  - Click on “Dataset 2 arrow_collapsed.gif” to expand the panel
  - Dataset 2 / Expression: leave empty
  - Dataset 2 / Enrichments 1: 24hr_David_Output.txt (OPTIONAL)
- Tune Parameters
  - P-value cut-off 0.001
  - Q-value cut-off 0.05

    - Check Overlap Coefficient
        * Overlap coefficient cut-off 0.6
- Build Enrichment Map
- Go to View, and activate Show Graphics Details

Note: There are multiple values in DAVID that can be used for the Q-value in EM including Bonferroni, Benjamini, and FDR. In EM we use the Benjamini as the Q-value.

Step 3: Examining Results



Legend:

- Node (inner circle) size corresponds to the number of genes in dataset 1 within the geneset
- Node border (outer circle) size corresponds to the number of genes in dataset 2 within the geneset
- Colour of the node (inner circle) and border(outer circle) corresponds to the significance of the geneset for dataset 1 and dataset 2, respectively.
- Edge size corresponds to the number of genes that overlap between the two connected genesets. Green edges correspond to both datasets when it is the only colour edge. When there are two different edge colours, green corresponds to dataset 1 and blue corresponds to dataset 2.

Note: If you are using two enrichment sets you will see two different colours of edges in the enrichment map. When the set of genes in the two datasets are different (for example, when you are comparing two different species or when you are comparing results from two different platforms) the overlaps are computed for each dataset separately as there is a different set of genes that the enrichments were calculated on. In this case, since the enrichments were reduced to only a subset of most differentially expressed at each time point the set of genes the enrichments are calculated on are different and overlap are calculated for each set separately.

4.9.4 Generic Tutorial

Files

Download the test data: EM_EstrogenMCF7_TestData_Generic.zip

Note: The results should be the same as GSEA.

4.9.5 BiNGO Tutorial

This quick tutorial will guide you through the generation of an Enrichment Map for an analysis performed using the Cytoscape Plugin BiNGO A Biological Network Gene Ontology Tool.

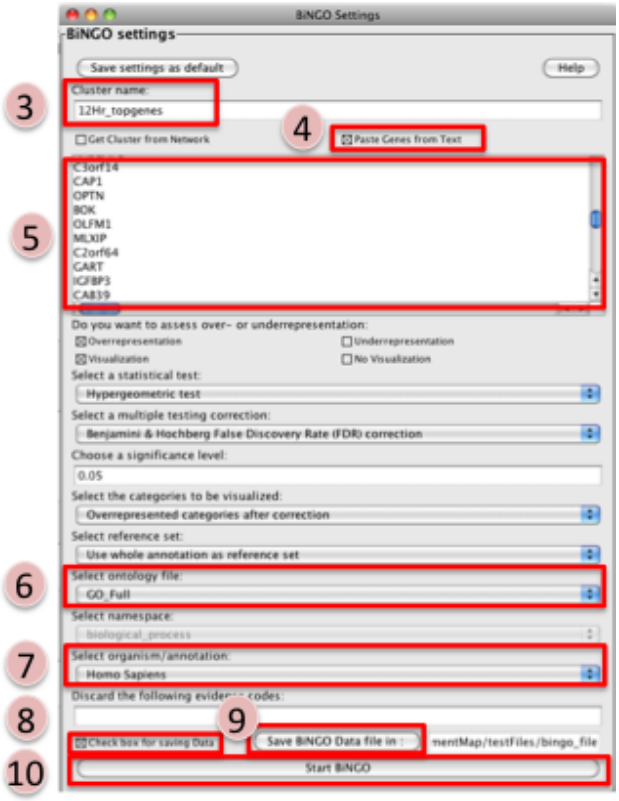
Files

Download the test data: `BingoTutorial.zip`

Description of the tutorial files contained in the BingoTutorial folder:

- `12hr_topgenes.txt` List of top genes expressed in Estrogen dataset at 12hr - Official Gene Symbol.
- `24hr_topgenes.txt` List of top genes expressed in Estrogen dataset at 24hr - Official Gene Symbol.
- `12hr_Bingo_Output.bgo` : Estrogen treatment - 12hr BiNGO result chart
- `24hr_Bingo_Output.bgo` : Estrogen treatment - 24hr BiNGO result chart
- `Estrogen_expression_file.txt`: Expression File - Estrogen treatment, Official Gene Name as key.

Step 1: Generate BiNGO output files

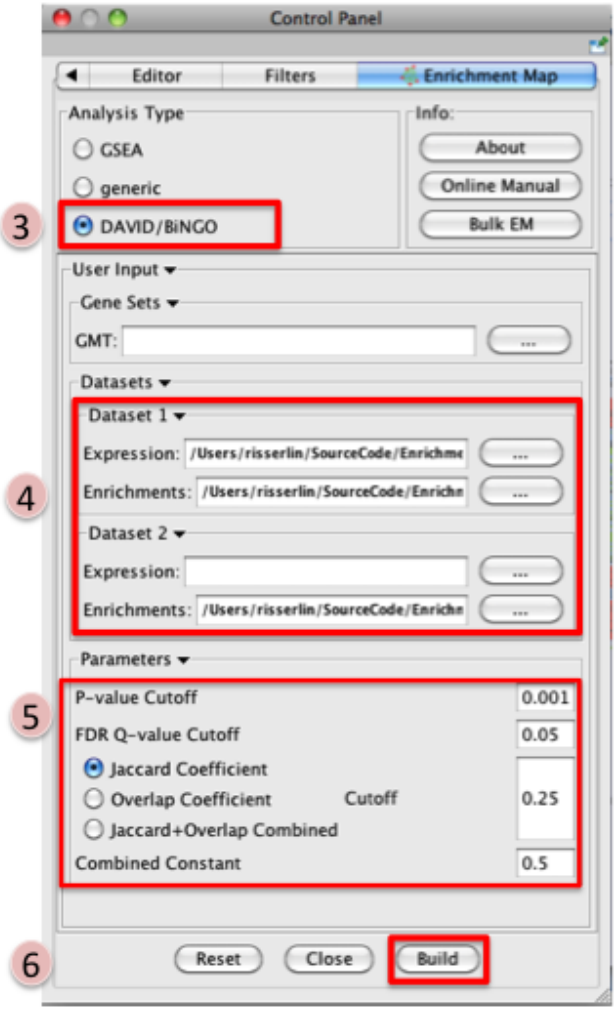


1. Open Cytoscape
2. Click on Plugins / Start Bingo v###
3. Enter the name “12hr_topgenes” in the text box marked Cluster name
4. Select the box Paste Genes from Text
5. Select and copy all genes in the tutorial file 12hr_topgenes.txt. Paste in large text box.
6. Change Select ontolgy file to GO_full
7. Change Select organism/annotation to Homo sapiens
8. Select the box Check box for saving Data
9. Click on Save BiNGO Data file in:. Navigate to desired folder and Click Save
10. Click on Start BiNGO

Note: Repeat these steps for the 24hrs time-point and the file 24hr_topgenes.txt

---

Step 2: Generate Enrichment Map with BiNGO Output

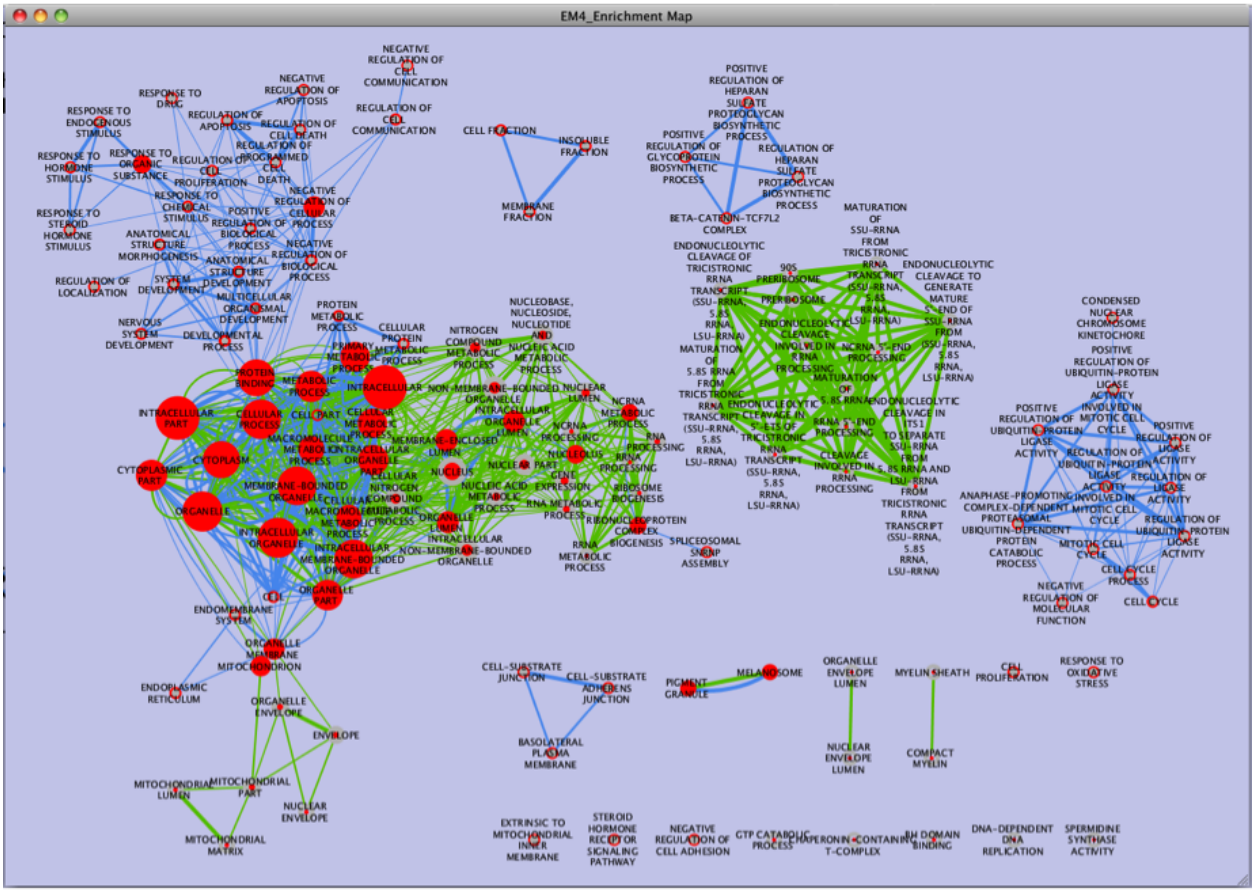


1. Open Cytoscape
2. Click on Plugins / Enrichment Maps / Load Enrichment Results
3. Make sure the Analysis Type is set to DAVID/BiNGO
4. Please select the following files by clicking on the respective (. . .) button and selecting the file in the Dialog:
    - NO GMT file is required for BiNGO Analysis
    - Dataset 1 / Expression: *Estrogen_expression_file.txt* (OPTIONAL)
    - Dataset 1 / Enrichments: *12hr_Bingo_Output.bgo*
    - Click on "Dataset 2" to expand the panel
    - Dataset 2 / Expression: leave empty
    - Dataset 2 / Enrichments 1: *24hr_Bingo_Output.bgo* (OPTIONAL)

5. Tune Parameters
    - P-value cut-off: *0.001*
    - Q-value cut-off: *0.05*
    - Overlap coefficient cut-off: *0.25*
6. Build Enrichment Map
7. Go to View, and activate Show Graphics Details

Note: BiNGO accepts both Entrez Gene IDs [e.g.6672] or gene symbols [STAT1] as input. If Entrez Gene IDs have been used as input, the first column of the expression file should contain Entrez Gene IDs. Similarly, if gene symbols have been used as input, the fist column of the expression file should contain gene symbols.

Step 3: Examining Results



Legend:

1. Node (inner circle) size corresponds to the number of genes in dataset 1 within the geneset
2. Node border (outer circle) size corresponds to the number of genes in dataset 2 within the geneset

3. Colour of the node (inner circle) and border(outer circle) corresponds to the significance based on the BiNGO p-value of the geneset for dataset 1 and dataset 2, respectively.
4. Edge size corresponds to the number of genes that overlap between the two connected genesets. Green edges correspond to both datasets when it is the only colour edge. When there are two different edge colours, green corresponds to dataset 1 and blue corresponds to dataset 2.

Note: If you are using two enrichment sets you will see two different colours of edges in the enrichment map. When the set of genes in the two datasets are different (for example, when you are comparing two different species or when you are comparing results from two different platforms) the overlaps are computed for each dataset separately as there is a different set of genes that the enrichments were calculated on. In this case, since the enrichments were reduced to only a subset of most differentially expressed at each time point the set of genes the enrichments are calculated on are different and overlap are calculated for each set separately.

4.9.6 g:Profiler Tutorial

This quick tutorial will guide you through the generation of an Enrichment Map for an analysis performed using g:Profiler (Functional Profiling of Gene List from large-scale experiments).

Files

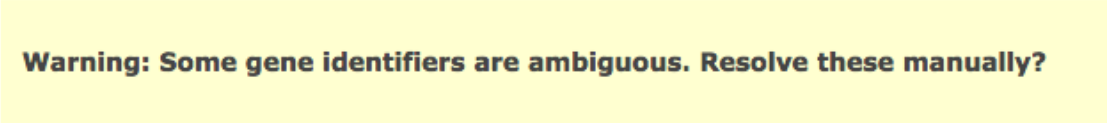
Download the test data: `gProfilerTutorial.zip`

Description of the tutorial files contained in the gProfilerTutorial folder:

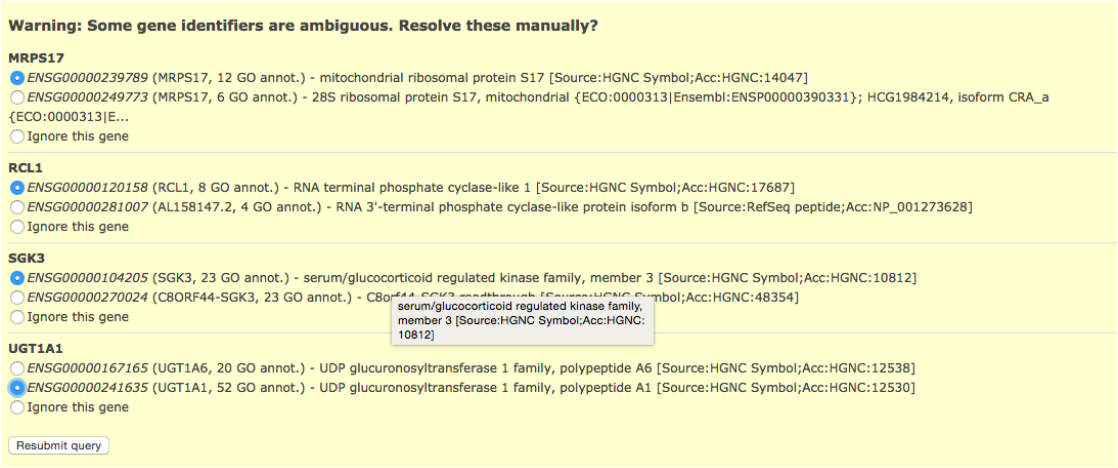
- `12hr_topgenes.txt` List of top genes expressed in Estrogen dataset at 12hr - Official Gene Symbol.
- `24hr_topgenes.txt` List of top genes expressed in Estrogen dataset at 24hr - Official Gene Symbol.

Step 1: Generate g:Profiler output files

1. Go to g:Profiler website
2. Select and copy all genes in the tutorial file 12hr_topgenes.txt in the *Query* box. Make sure that your list contains only official gene symbol (HUGO).
3. In Options, check Significant only, No electronic GO annotations
4. Set the Output type to Generic EnrichmentMap
5. Show advanced options
6. Set Min and Max size of functional category to 3 and 500 respectively.
7. Select 2 for Size of Q&T
8. On the right panel, choose the Gene Ontology Biological process and Reactome
9. Set Significance threshold to Benjamini-Hochberg FDR
10. Click on g:Profile! to run the analysis
    - [Note] - if some of your identifiers in your query have multiple mappings in g:Profiler! by default they get excluded. If this happens you will see the following above the g:Profiler! results:



- Click on above link to manually map each gene to its correct annotation



- Click on Resubmit query to update your results with the specified mappings.
- If the identifier discrepancy warning is ignored there might be differences between the number of genes g:Profiler attributes to a particular gene set and those associated with it in the Enrichment Map.

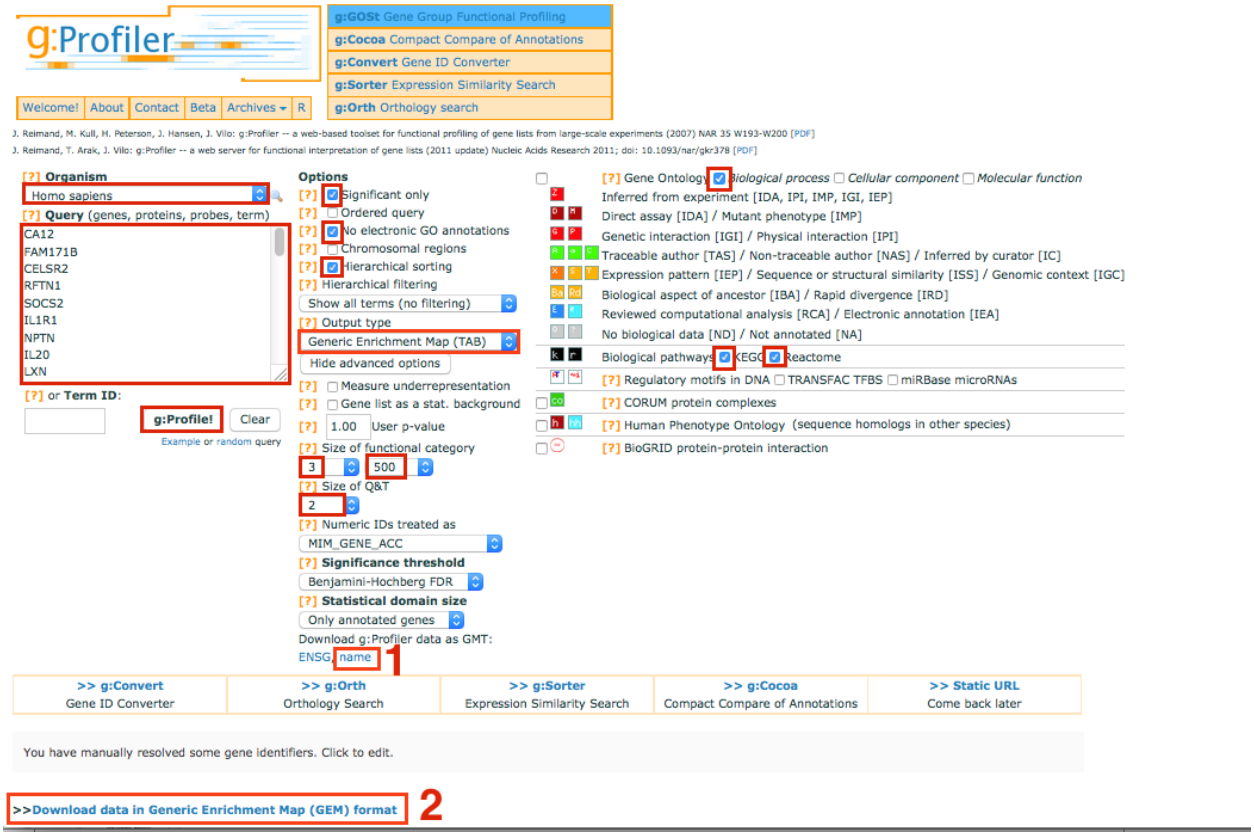
11. Download g:Profiler data as gmt name Note, you will have to unzip the folder
12. Download the result file: Download data in Generic Enrichment Map (GEM) format

---

Note: Repeat these steps for the 24hrs time-point and the file 24hr_topgenes.txt

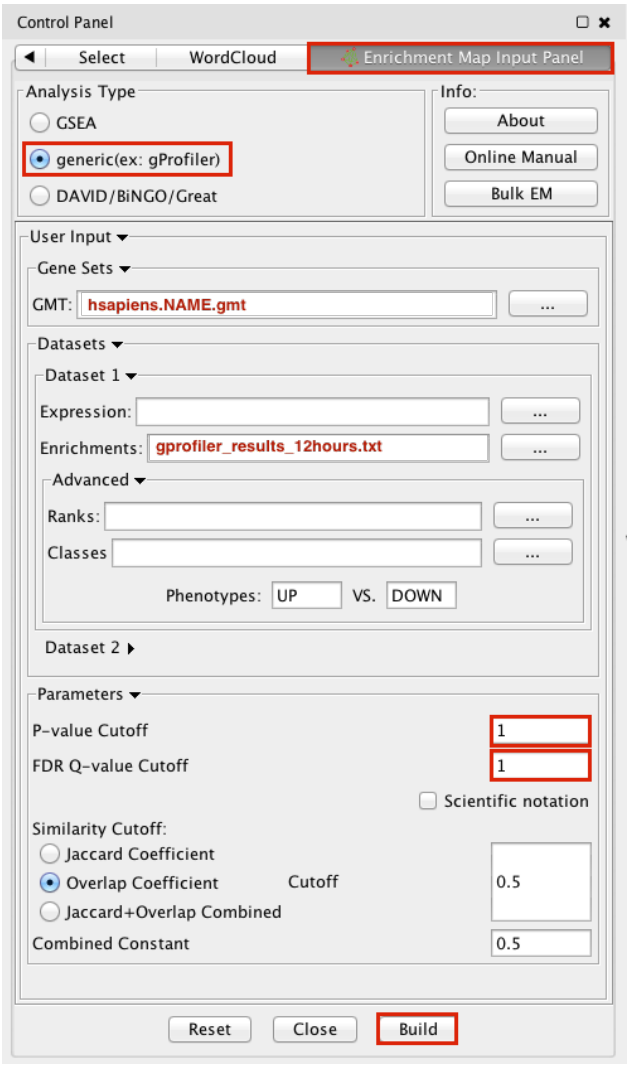
---

Link to a step by step tutorial: `gProfiler_step_by_step.pdf`



Step 2: Generate Enrichment Map with g:Profiler Output

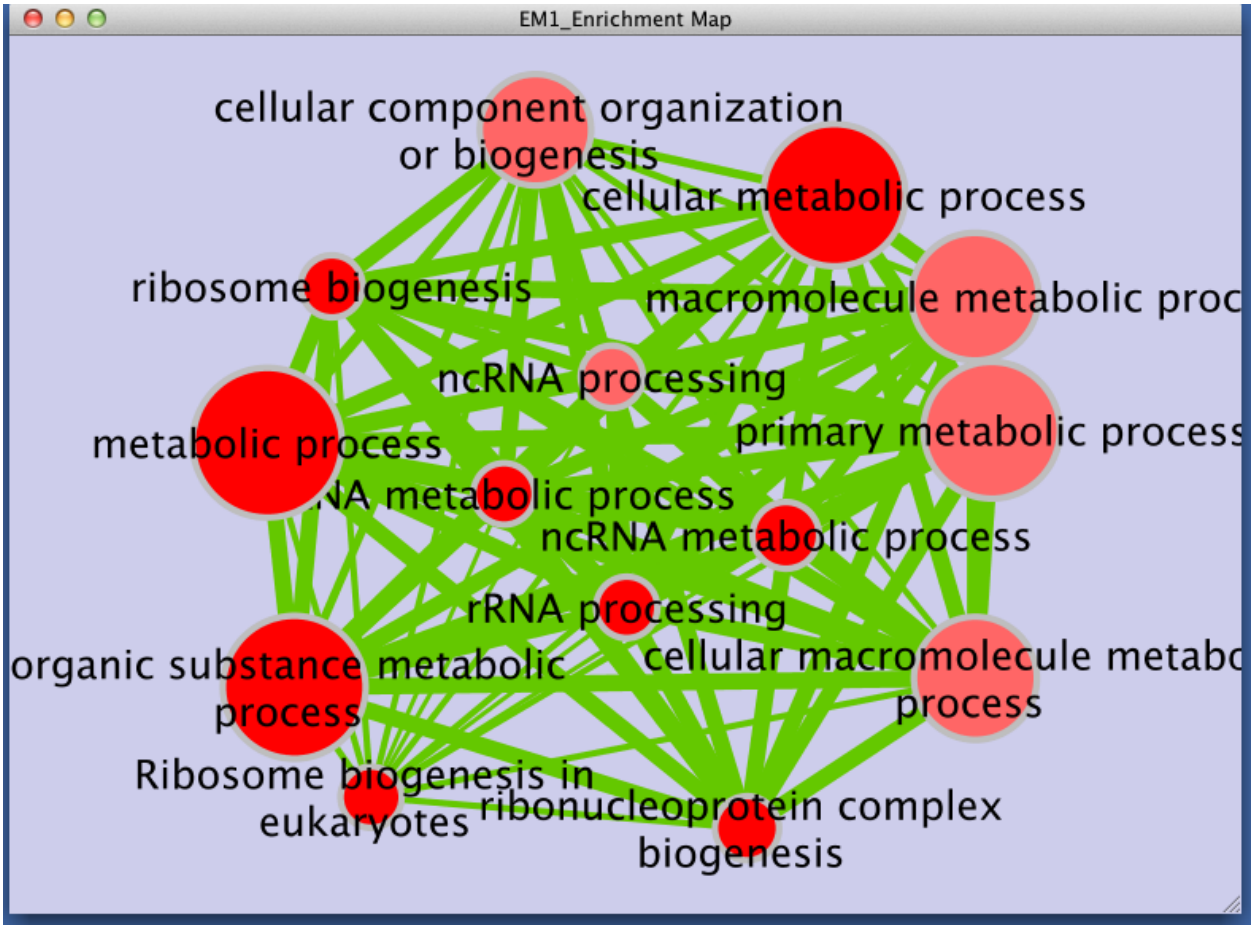
g:Profiler output files from Step 1: `gProfiler_EM.zip`



1. Open Cytoscape
2. In the menu bar, locate the App tab and then select > EnrichmentMap > Create Enrichment Map
3. Make sure the Analysis Type is set to generic(ex:gProfiler)
4. Please select the following files by clicking on the respective (. . .) button and selecting the file in the Dialog:
    - GMT / hsapiens.pathways.NAME.gmt
    - Dataset 1 / Enrichments: gProfiler_results_12hr.txt
5. Tune Parameters
    - P-value cut-off: 1
    - Q-value cut-off: 1
    - Overlap Coefficient cut-off: 0.5
6. Click on the Build radio button at the bottom of the panel to create the Enrichment Map
7. In the menu bar, Go to View, and activate Show Graphics Details
8. In the control panel, go to Style, click on Label and select EM1_GS_DESCR in the Column dropdown. This will label nodes with names rather thsn GO IDs. The selected value may be EM2_GS_DESCR or other if you

have more than one Enrichment Map open in Cytoscape.

Step 3: Examining Results



Legend:

- Node size corresponds to the number of genes in dataset 1 within the geneset
- Colour of the node corresponds to the significance of the geneset for dataset 1.
- Edge size corresponds to the number of genes that overlap between two connected genesets.

4.9.7 GREAT Tutorial

This quick tutorial will guide you through the generation of an Enrichment Map for an analysis performed using Genomic Region Enrichment Annotation Tool (GREAT),

Files

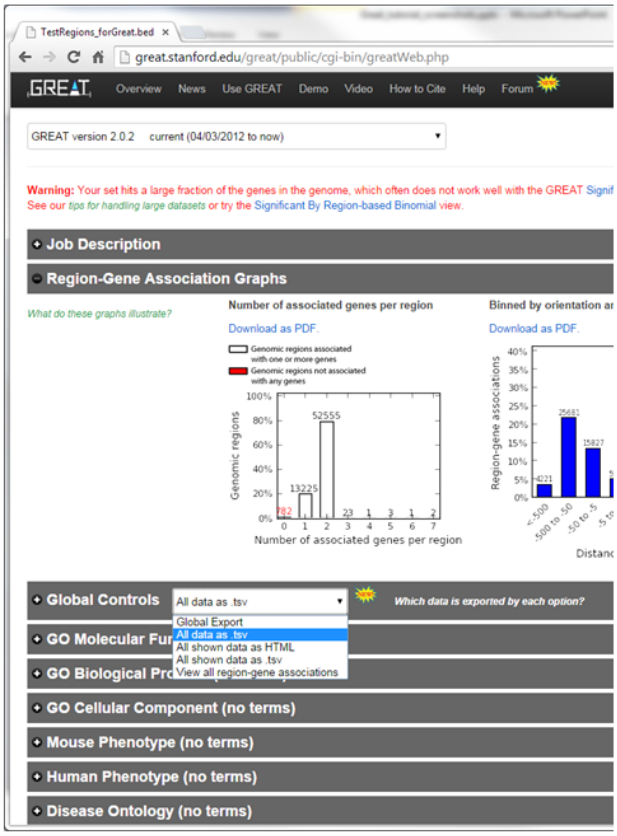
Download the test data: `GREATTutorial.zip`

Description of the tutorial files contained in the GREATTutorial folder:

- `TestRegions_ForGREAT.bed` Example GREAT genomic region input file.

- `GreatExprotAll.tsv` Example of download GREAT output file.
- `20140919-public-2.0.2-3vD5MB-hg19-all-gene.txt` Example downloaded GREAT gene to region association file.
- `geneToRegionExpressionFile.txt` Transformed gene to region association file downloaded from GREAT.

Step 1: Generate GREAT output files



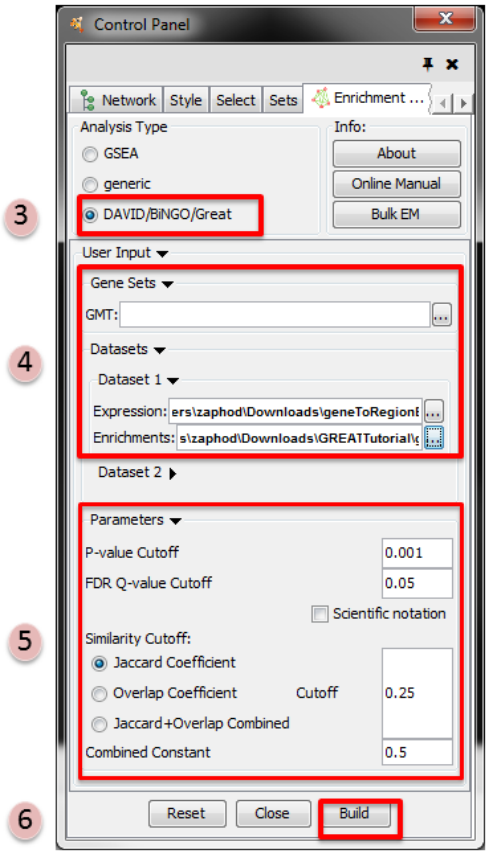
1. GO to GREAT website
2. *Select Species Assembly* associated with your data. For this tutorial select *Human: GRCh37*
3. In *Test regions* click on *Choose File*
4. Navigate to files provided and select *TestRegions_forGreat.bed*
5. Click on *Submit*
6. Once the results page has loaded download all the results - in the *Global controls* heading click on the down arrow next to *Global Export*
7. *Select All data as tsv* - greatExportAll.tsv will automatically be downloaded to your default Downloads directory. This is the file you can use in Enrichment Map (Dataset 1 or 2:Enrichment Results)

Step 1B (Optional): Generate Gene to region association file

Optional - Download the Gene-to-region used by GREAT and modify it to be used in EM as an expression file.

1. In the *Global controls* heading click on the down arrow next to *Global Export*
2. Select *view all region-gene associations*
3. Next to *Gene > genomic region association table* [The table on the right hand side of the page] click on *Download table as text*.
4. File will automatically downloaded into your default Downloads directory (file name is similar to DATE-public-2.0.2-3vD5MB-hg19-all-gene.txt where DATE is the date of download. Name will also change depending on the version of GREAT and genome selected).
5. Open the downloaded file in Excel.
6. Add a row to the top of the file.
7. In the first column enter "Name", and in the second column enter "Description"

Step 2: Generate Enrichment Map with GREAT Output

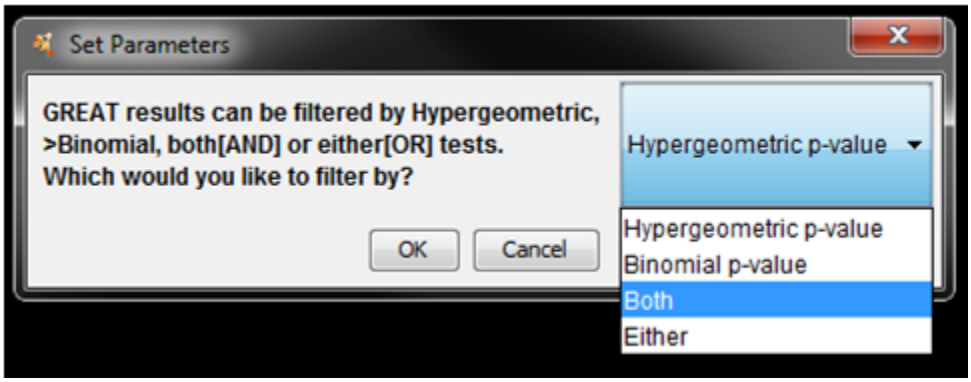


1. Open Cytoscape
2. Click on Apps / Enrichment Maps / Load Enrichment Results
3. Make sure the Analysis Type is set to DAVID/BiNGO/GREAT
4. Please select the following files by clicking on the respective (...) button and selecting the file in the Dialog:
   - *NO GMT file is required for GREAT Analysis*
   - Dataset 1 / Expression: `!geneToRegionExpressionFile.txt` (OPTIONAL)

- Dataset 1 / Enrichments: `!GreatExportAll.tsv`

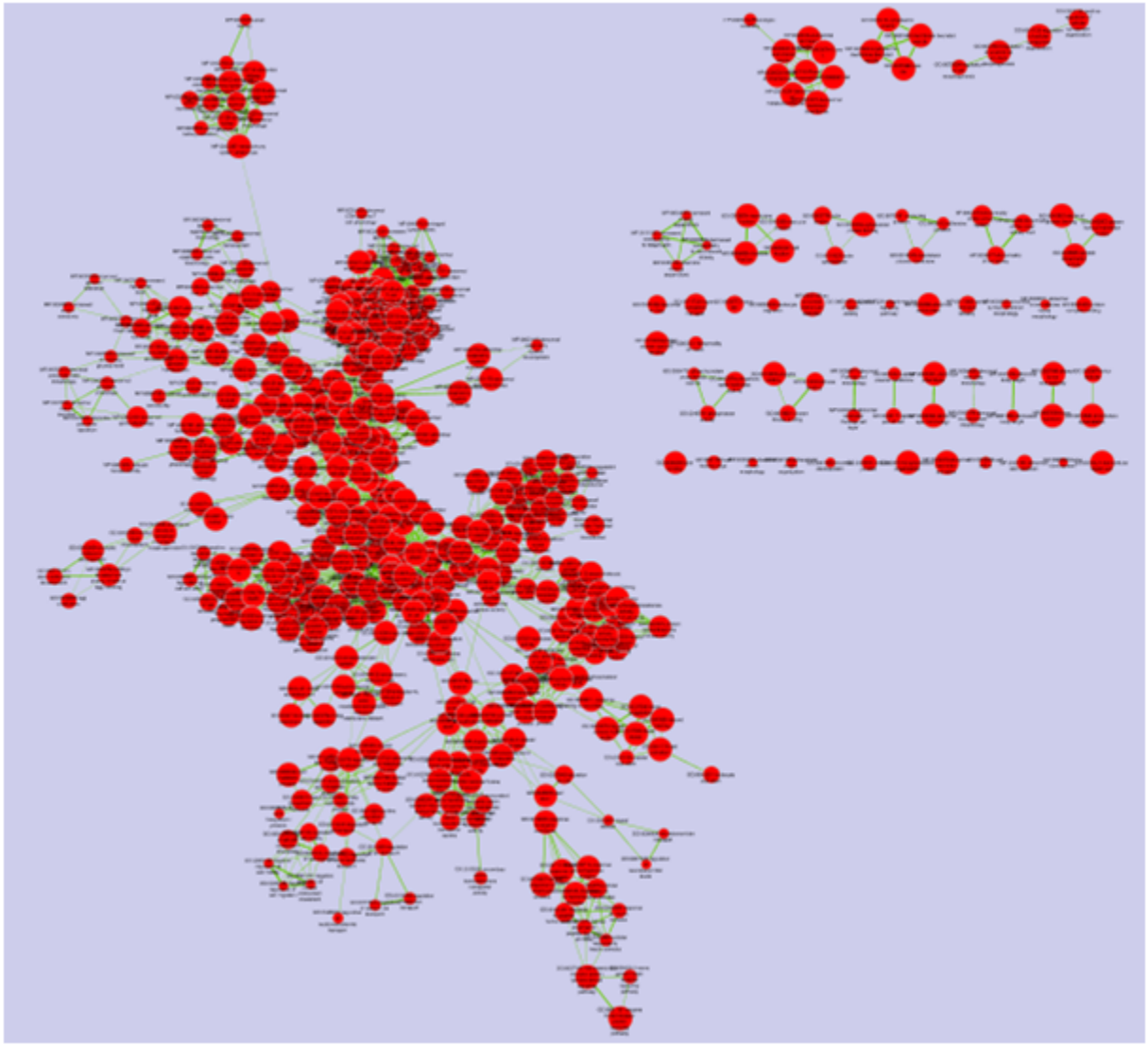
5. Tune Parameters
    - P-value cut-off *0.001*
    - Q-value cut-off *0.05*
    - Jaccard coefficient cut-off *0.25*
6. Build Enrichment Map

Step 3: Filtering GREAT results



- Once the network starts to build a dialog will pop up asking you how you would like to filter the GREAT results. There are four options:
    1. Use Hypergeometric test p-values and FDR only –> **Hypergeometric**
    2. Use Binomial test p-values and FDR only. –> **Binomial**
    3. Use both hypergeometric and binomial test p-values and FDR. Enrichment result must pass threshold for both tests. –> **Both**
    4. Enrichment result must pass one of the above tests to be included in the results –> **Either**
- Select Both

Step 4: Examining Results



Legend:

- Node (inner circle) size corresponds to the number of genes in dataset 1 within the geneset
- Colour of the node (inner circle) corresponds to the significance of the geneset for dataset 1.
- Edge size corresponds to the number of genes that overlap between the two connected genesets.

4.9.8 Post Analysis Tutorial

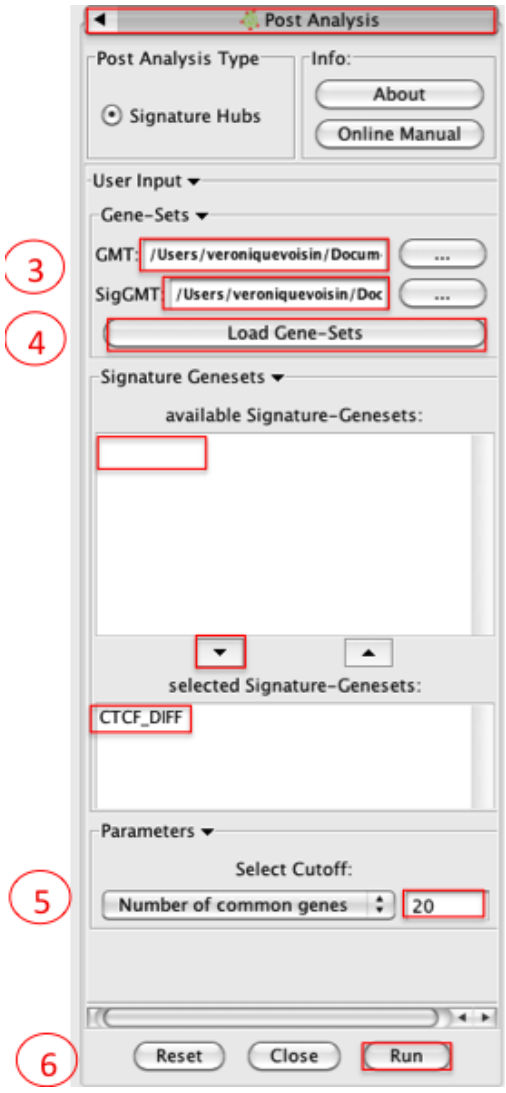
Outline

This quick tutorial will guide you through the creation of an additional gene-set on a pre-existing Enrichment Map. It can for example help to localize microRNA, transcription factors or drug targets in enriched pathways displayed on the map. The new gene-set that we want to add to the network is called the signature gene-set.

Download the test data: `PostAnalysisTutorial.zip`

Description of the tutorial files contained in the PostAnalysisTutorial folder:

- `CTCF_DIFF.gmt` Signature gene-set.
- `Human_GO_AllPathways_no_GO_iea_April_15_2013_symbol.gmt` Gene-set file used to create the original Enrichment Map
- `ES12_EM_example.cys` The Enrichment Map on which we want to add the signature gene-set



Instructions

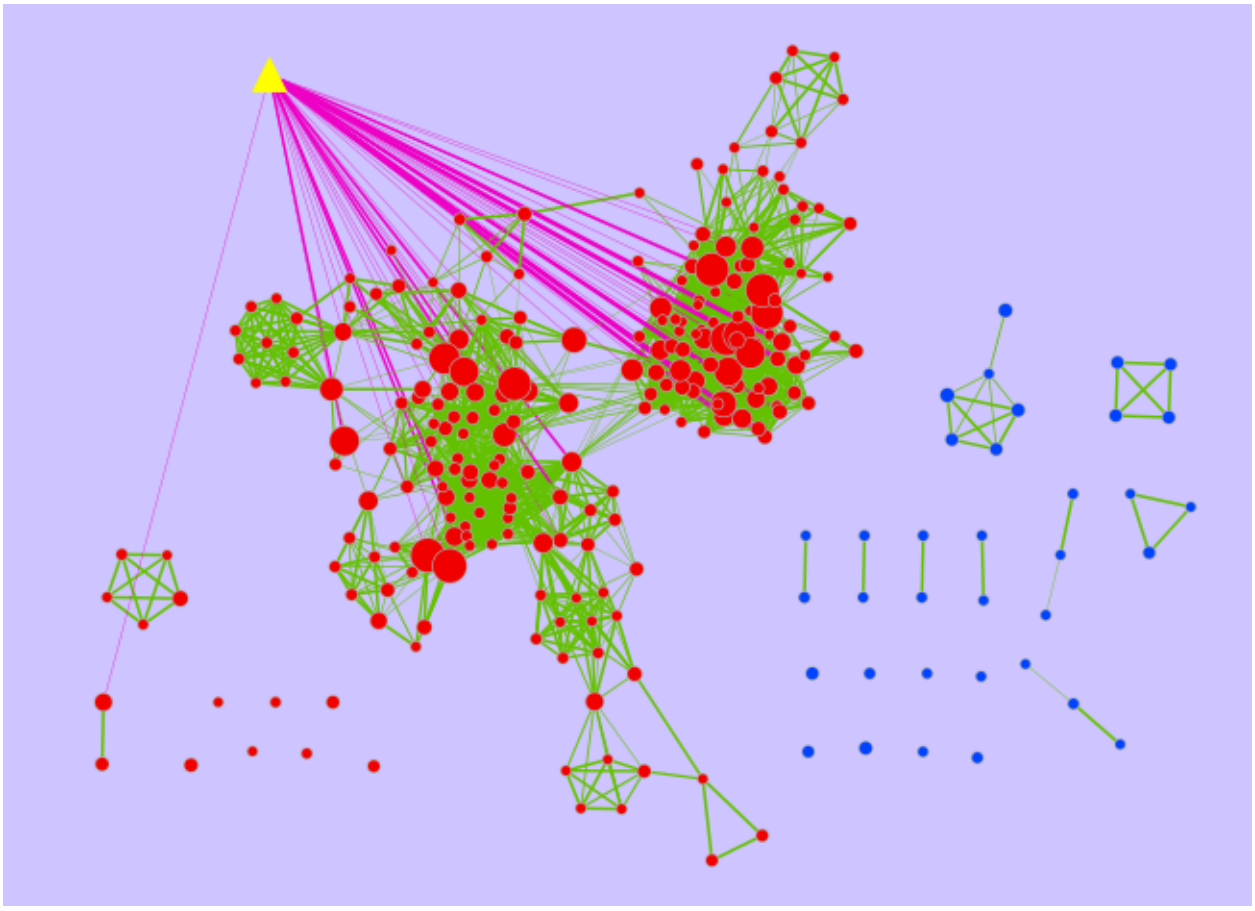
1. Open Cytoscape and Open ES12_EM_example.cys
2. Click on Plugins / Enrichment Map/ Post Analysis
3. Please select the following files by clicking on the respective (. . .) button and selecting the file in the Dialog:
   - GMT / 'Human_GO_AllPathways_no_GO _iea_April_15_2013_symbol.gmt'
   - SigGMT / 'CTCF_DIFF.gmt'
4. Click on Load Gene-Sets

- In the Signature-Genesets box: click on CTCF_DIFF as available Signature-Genesets
- Click on the down arrow to move CTCF_DIFF in the lower box

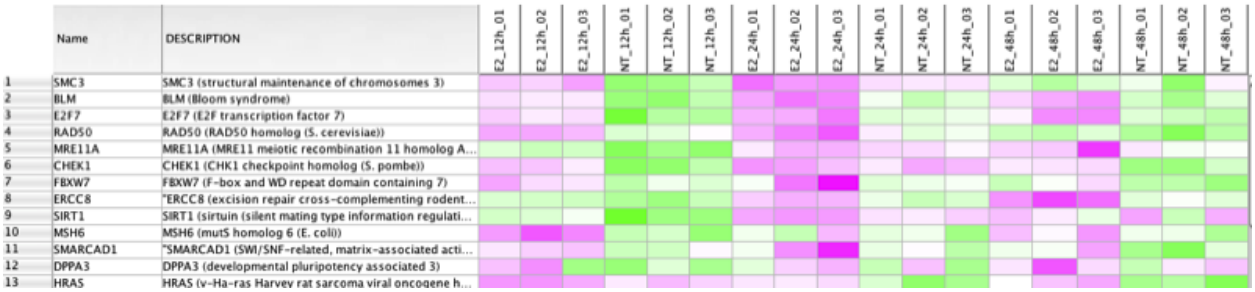
5. Tune parameters
    - Set Number of common genes to 20
6. Click on Run

Examining Results

`ES12_EM_example_PA_CTCF_Diff.cys`



- The yellow triangle is the signature gene-set and the pink edges represent overlap of 20 genes or more between the signature gene-set and a given gene-set from the Enrichment Map (red node). The width of the pink edges is proportional to the number of genes in the overlap.
- This signature gene-set represents the CTCF (transcription factor) target genes that are different in MCF7 cells treated or not with Estrogen.
- Clicking on a pink edge will show the genes contained in this overlap in the gene expression panel:



	Name	DESCRIPTION	E2_12h_01	E2_12h_02	E2_12h_03	NT_12h_01	NT_12h_02	NT_12h_03	E2_24h_01	E2_24h_02	E2_24h_03	NT_24h_01	NT_24h_02	NT_24h_03	E2_48h_01	E2_48h_02	E2_48h_03	NT_48h_01	NT_48h_02	NT_48h_03
1	SMC3	SMC3 (structural maintenance of chromosomes 3)																		
2	BLM	BLM (Bloom syndrome)																		
3	E2F7	E2F7 (E2F transcription factor 7)																		
4	RAD50	RAD50 (RAD50 homolog (S. cerevisiae))																		
5	MRE11A	MRE11A (MRE11 meiotic recombination 11 homolog A...																		
6	CHEK1	CHEK1 (CHK1 checkpoint homolog (S. pombe))																		
7	FBXW7	FBXW7 (F-box and WD repeat domain containing 7)																		
8	ERCC8	"ERCC8 (excision repair cross-complementing rodent...																		
9	SIRT1	SIRT1 (sirtuin (silent mating type information regulati...																		
10	MSH6	MSH6 (mutS homolog 6 (E. coli))																		
11	SMARCAD1	"SMARCAD1 (SWI/SNF-related, matrix-associated acti...																		
12	DPPA3	DPPA3 (developmental pluripotency associated 3)																		
13	HRAS	HRAS (v-Ha-ras Harvey rat sarcoma viral oncogene h...																		

Note: The gene-signature gene-set corresponds to ENCODE CHIP-seq data for the cell line MCF-7 treated or not with estrogen and for the CTCF factor using the tool CSCAN/Browse data (http://159.149.160.51/cscan/). The signature gene-set includes the genes that were different between the two conditions (treated or not with estrogen)

4.10 collapse_ExpressionMatrix.py

Download `collapse_ExpressionMatrix.py`.

This tool can process a gene expression matrix (in GCT or TXT format) ranked list (RNK format) and:

- convert the Identifier based on a Chip Annotation file (e.g. AffyID -> Gene Symbol)
- collapse the expression values or rank-scores for Genes from more than one probe set.

Converting and collapsing can be done either individually or both at the same time.

In case you are collapsing a ranked list (RNK format) to perform a "preRanked GSEA" that you later on want to analyze with EnrichmentMap and want to see an expression heatmap for the genesets, you need to generate an expression matrix that contains the expression values from the same probesets that were chosen to represent the gene in the ranked list. This can by selecting the ranked List (RNK) as the primary input file (-i) and the expression Matrix (GCT or TXT) as additional input Expression-table (-e). When using the GUI this can be done by selecting the mode "Ranked List with Expression Matrix".

In this use-case ID-conversion and collapsing have to be done in the same step. The DESCRIPTION column of the collapsed expression matrix will for every given gene then contain the Probeset-ID of the Probeset with the highest absolute Score in the RNK file and in brackets followed by a list of Probeset-IDs that where omitted due to lower absolute rank-scores.

The option 'Suppress gene "NULL"' (–null) will drop all Probeset ID's assigned to the Gene Symbol NULL, as this is used for probesets that are not linked to any Gene in several Chip-Annotation files available from the Broad Institute's FTP server. (These will be dropped by GSEA anyway).

4.10.1 Requirements

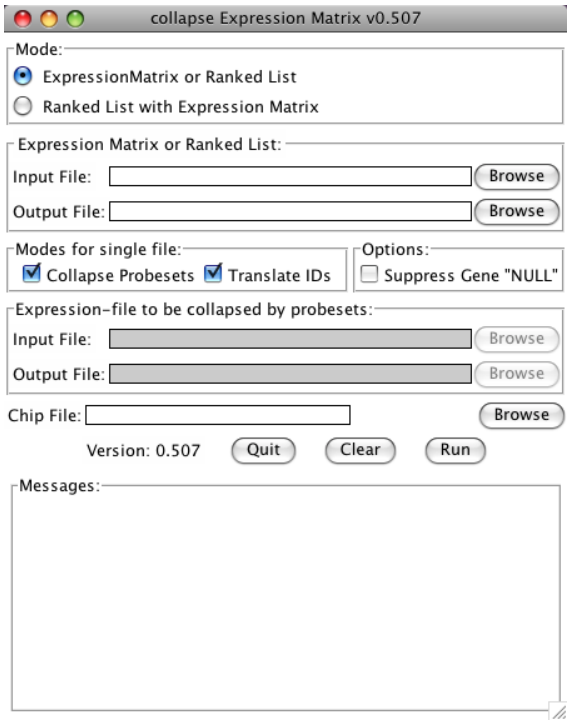
- Python 2.3 or newer (but not Python 3.x!)
- the Tkinter Library (comes with most Python installations) for the GUI

Supported Operating Systems:

- MacOS X 10.5 "Leopard" ore newer (probably also MacOS X 10.4 "Tiger")
- Windows (download and install the most recent version of Python 2.x from: http://www.python.org/download/ or http://www.activestate.com/activepython/downloads/

- Linux (Python and Tcl/Tk are probably already installed out of the box, otherwise install the packages with your Distribution's package manager)

4.10.2 GUI Mode



collapse_ExpressionMatrix.py now has a Tk-based Graphical User Interface (GUI). To use the GUI, just start the program without any arguments. This can be done:

- on Windows: double-click on the collapse_ExpressionMatrix.py-file
- on MacOS 10.5 or newer with installed "Developer Tools":
  - Control-click (or right-click) on the collapse_ExpressionMatrix.py-file in the finder and choose "Open With/Build Applet.app"
  - This will create an MacOS Application collapse_ExpressionMatrix.app which can be started by double clicking.
- on MacOS, Linux or other Unix-like Systems in a Terminal/Shell: see in section "Command Line Mode" how to make the program executable.

After starting the GUI:

- for collapsing either an expression matrix or Ranked gene list:
  1. select mode "Expression Matrix or Ranked List"
  2. use the first Browse-Button to select an Expression Matrix or Ranked gene list as an input file.
  3. use the second Browse-Button to choose a name and location of the output file (the program will suggest to use the same name as the input file with an inserted "_collapsed" before the extension)
  4. choose if the Identifiers should be converted or the file should be collapsed by checking the check-boxes
  5. choose if the Gene Symbol "NULL" should be dropped

6. if Identifiers are to be converted, choose a matching chip file
7. start the conversion by clicking the Run button

- for collapsing a Ranked gene list and generating an expression matrix containing the same probesets:
  1. select Mode "Ranked List with Expression Matrix"
  2. use the first Browse-Button to select the Ranked gene list as an input file.
  3. use the second Browse-Button to choose a name and location of the Ranked gene list output file (the program will suggest to use the same name as the input file with an inserted "_collapsed" before the extension)
  4. choose if the Gene Symbol "NULL" should be dropped
  5. use the third Browse-Button to select an Expression Matrix input file
  6. use the fourth Browse-Button to choose a name and location of the Expression Matrix output file
  7. choose a matching chip file
  8. start the conversion by clicking the Run button

4.10.3 Command Line Mode

If you are familiar with command line tools under Unix/Linux, `collapse_ExpressionMatrix.py -h` gives you all the information you need (if not, see below):

```
$ collapse_ExpressionMatrix.py -h
Usage: collapse_ExpressionMatrix.py [options] -i input.gct -o output.gct [-c platform.
↪chip] [--collapse]

This tool can process a gene expression matrix (in GCT or TXT format) or
ranked list (RNK format) and either replace the Identifier based on a Chip
Annotation file (e.g. AffyID -> Gene Symbol), or collapse the expression
values or rank-scores for Genes from more than one probe set. Both can be done
in one step by using both '-c platform.chip' and '--collapse' at the same
time. If a ranked list is to be collapsed, an additional expression matrix can
be supplied by the -e/-x parameters and will be filtered to contain the same
probe-sets as selected from the RNK file. If however the file supplied by -i
is not recognized as a RNK file, these options have no effect.  For detailed
descriptions of the file formats, please refer to:
http://www.broadinstitute.org/cancer/software/gsea/wiki/index.php/Data_formats
Call without any parameters to select the files and options with a GUI
(Graphical User Interface)

Options:
  --version             show program's version number and exit
  -h, --help            show this help message and exit
  -i FILE, --input=FILE
                        input expression table or ranked list
  -o FILE, --output=FILE
                        output expression table or ranked list
  -c FILE, --chip=FILE  Chip File This implies that the Identifiers are to be
                        replaced.
  -e FILE, --ei=FILE    (optional) additional input Expression-table, to be
                        restricted to the same probe-sets as the RNK file
  -x FILE, --xo=FILE    (optional) corresponding output file for -i/--ei
                        option
  --collapse            Collapse multiple probe sets for the same gene symbol
```

```
                        (max_probe)
  --no-collapse         Don't collapse multiple probesets [default]
  --null                suppress Gene with Symbol NULL
  -g, --gui             Open a Window to choose the files and options.
  -q, --quiet           be quiet
```

MacOS and Linux

On MacOS and Linux you need to make the program executable. Therefore:

- copy the file to a directory, e.g. `${HOME}/bin`
- open a Terminal
- set the executable flag:

  ```
  chmod a+x ${HOME}/bin/collapse_ExpressionMatrix.py
  ```

- if the `${HOME}/bin` directory is not in your search Path (test by running `collapse_ExpressionMatrix.py` from a terminal) add it by adding the line `export PATH=$HOME/bin:$PATH` to your `${HOME}/.bash_profile` using your favourite text editor (pico, vi, emacs, gedit, TextWrangler, etc.) or with the command

  ```
  echo export PATH=\$HOME/bin:\$PATH >> ${HOME}/.bash_profile
  ```

  or refer to your local SysAdmin for any other shell that bash.
- open a new terminal or run source `${HOME}/.bash_profile`
- test with `collapse_ExpressionMatrix.py -h`

Windows

- copy the file to a directory, e.g. `C:\bin`
- open the Conrtol Panel
- open System
- go to Advanced System Settings (on Vista and 7 only)
- go to the Advanced Tab
- Click on Environment-button
- if in the section "User variables for {USERNAME}" there is already an entry called "PATH":
  - click on Edit...
  - append `;C:\bin` at the very end
- otherwise click on New...
  - Variable Name: `PATH`
  - Variable Value: `%PATH%;C:\bin`
- open a Command Prompt (Programs/Accessories)
- test with `collapse_ExpressionMatrix.py -h`