

---

# **EnrichmentMap Documentation**

***Release 3.1***

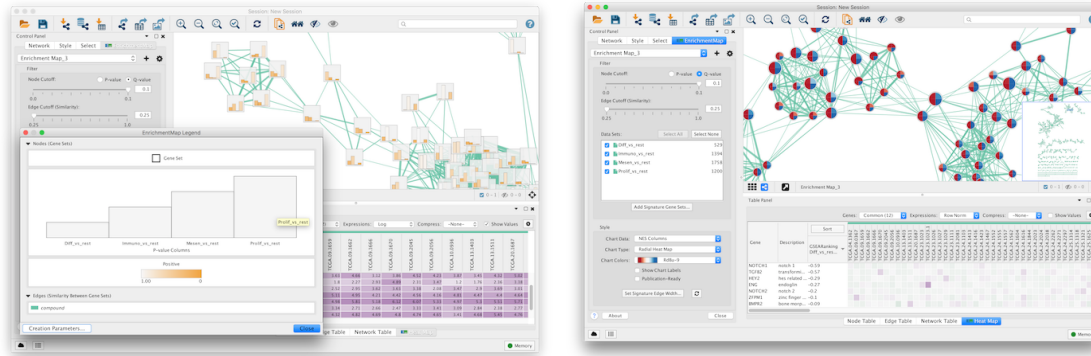
**Ruth Isserlin, Mike Kucera, Christian Lopes**

**Mar 02, 2018**



<b>1</b>	<b>Feature Requests and Reporting Bugs</b>	<b>3</b>
<b>2</b>	<b>Cite EnrichmentMap</b>	<b>5</b>
<b>3</b>	<b>Examples of Use</b>	<b>7</b>
<b>4</b>	<b>Papers Citing Enrichment Map</b>	<b>9</b>
4.1	Installation . . . . .	9
4.2	What's New . . . . .	10
4.3	Quick Tour . . . . .	11
4.4	Creating a Network . . . . .	18
4.5	Network Visualization . . . . .	26
4.6	Columns . . . . .	31
4.7	Main Panel . . . . .	33
4.8	Expression Panel (Heat Map) . . . . .	37
4.9	Legend Dialog . . . . .	42
4.10	Post Analysis (Add Signature Gene Sets) . . . . .	45
4.11	File Formats . . . . .	51
4.12	Tips on Parameter Choice . . . . .	59
4.13	Download Gene Set Files . . . . .	61
4.14	Automating EnrichmentMap . . . . .	67
4.15	Properties . . . . .	69
4.16	collapse_ExpressionMatrix.py . . . . .	71





Enrichment analysis (also known as functional enrichment) is a helpful technique for high-throughput data interpretation. Given a list of genes resulting from an experiment, enrichment analysis enables to identify functional categories that are over-represented. Such functional categories are typically derived from functional annotations (such as the Gene Ontology), or from pathway databases (such as KEGG), or other resources (such as the collection of disease signatures in MSig DB, or protein complexes in MIPS).

However, enrichment results are often characterized by lots of redundancy and inter-dependencies between gene-sets representing functional categories. For instance, *Response to radiation*, *DNA Integrity Checkpoint* and *p53 Pathway* have several genes in common. Since the typical enrichment analysis can output up to 300 hundred different gene-sets, some form of organization is required to navigate results.

To address this, we organize gene-sets into a network, called enrichment map. Two gene-sets are connected in the *enrichment map network* if they have a high overlap, i.e. if they share many genes. Applying automatic layout techniques, groups of inter-related gene-sets tend to cluster together, providing for a much easier and intuitive visualization.

Please also see [The EnrichmentMap Protocol](#) for details on automating EnrichmentMap.



---

## Feature Requests and Reporting Bugs

---

The EnrichmentMap GitHub issue tracker can be used to report a bug or request a feature.

To Report a bug:

- Go to <https://github.com/BaderLab/EnrichmentMapApp/issues>
- Click on *New Issue*
- Write a short description of the issue. It is very helpful to provide a series of steps that can be taken to reproduce the issue.
- If possible attach a session file (.cys) or example input files.
- Enter App version, Cytoscape version and operating system.
- Click on *Submit new issue*





- **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 3

---

### 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](#)



---

## 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](#)

### 4.1 Installation

- Cytoscape minimum version 3.5.1 is required.

Install Cytoscape

Download and install the latest version of Cytoscape at  
<http://www.cytoscape.org/download.php>.

Install EnrichmentMap 3.1

- 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 the **EnrichmentMap Pipeline Collection** can be installed. This will install EnrichmentMap as well as a suite of other apps that work well with EnrichmentMap, including **AutoAnnotate**, **WordCloud** and **clusterMaker2**.

EnrichmentMap can also be installed from the Cytoscape App Store at

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

## 4.2 What's New

### 4.2.1 EnrichmentMap 3.1

- Redesigned Post-Analysis UI
  - Ability to preview signature gene sets before import
  - Ability to import GMT files from the web ([download.baderlab.org](http://download.baderlab.org))
- New Chart: Color by DataSet.
- Support for *g:profiler* 2-phenotype analysis.
- Heat Map: Compress expressions by class.
- Export Legend to PDF.
- Export Heat Map to PDF.
- Ability to specify the name of the EnrichmentMap network.
- Drag-and-Drop folders onto the Create EnrichmentMap Dialog.
- [Several bug fixes](#)
- Requires Cytoscape 3.5.1

### 4.2.2 EnrichmentMap 3.0

EnrichmentMap 3.0 is a major release, with the following new features:

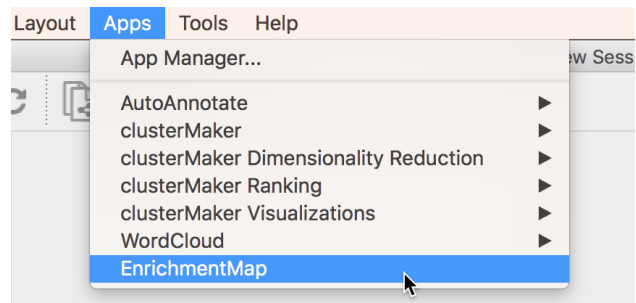
- EnrichmentMap networks can now be created from any number of enrichment data sets (EnrichmentMap 2.0 supported maximum 2 data sets).
- New chart visualizations on nodes for visualizing NES scores, p-values or q-values. Three new charts are available: radial heat-map, linear heat-map and heat-strips.
- New network creation dialog has the ability to scan a folder for files belonging to each dataset. In most cases this removes the need to manually enter the required files.
- New control panel allows the contents and style of the network and charts to be updated dynamically.
- New legend dialog.

- New streamlined HeatMap panel has the ability to summarize expression data.
- Several new commands that allow EnrichmentMap to be automated from external scripts and CyREST.
- [Several bug fixes](#)
- Requires Cytoscape 3.4

## 4.3 Quick Tour

### 4.3.1 Creating the Network


To create an Enrichment Map network go to the Cytoscape main menu and select **Apps > EnrichmentMap**.

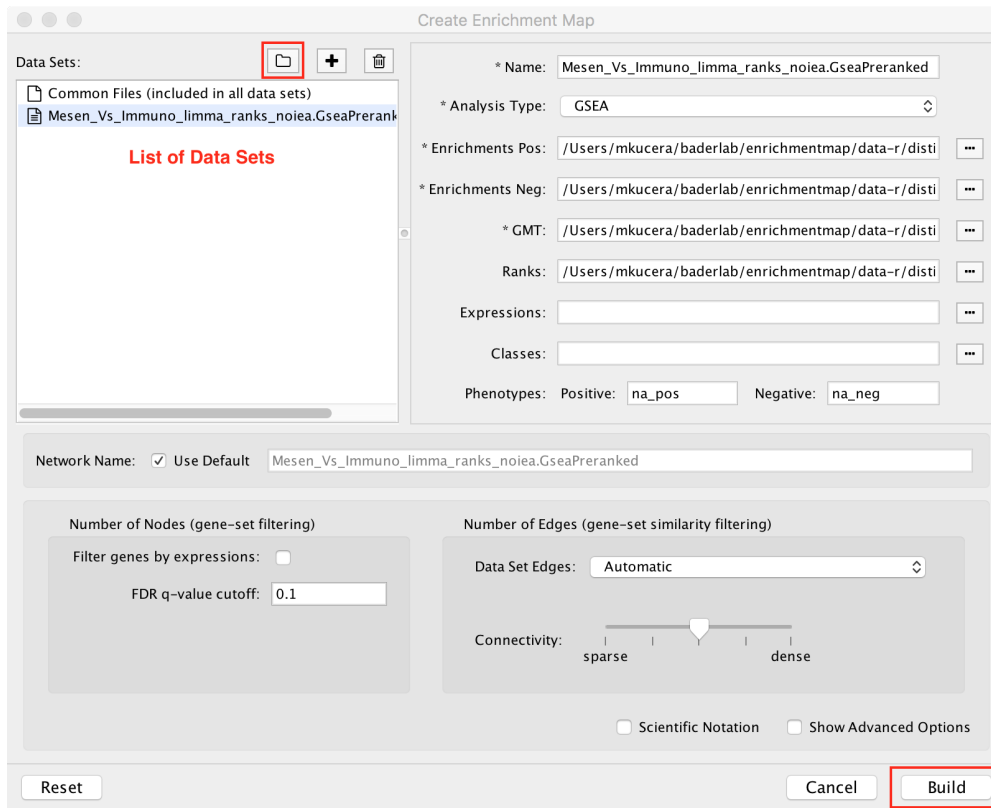


This will show the EnrichmentMap panels and open the **Create EnrichmentMap Dialog**.

**Note:** If the Create EnrichmentMap Dialog does not appear then click the (+) icon at the top of the EnrichmentMap Main Panel.

The quickest way to import data is to scan a folder for enrichment analysis data files (for example a GSEA results folder).

Click the  icon and select a folder. EnrichmentMap will scan the folder for files containing enrichment data, expression data, ranks, classes and gene set definitions. These files will be arranged into a **List of Data Sets**, each of which contains the data for one experiment. Click the **Build** button to create the network.



---

**Note:** New in Cytoscape 3.1: You can initiate a scan by dragging and dropping folders onto the Data Sets List.

---

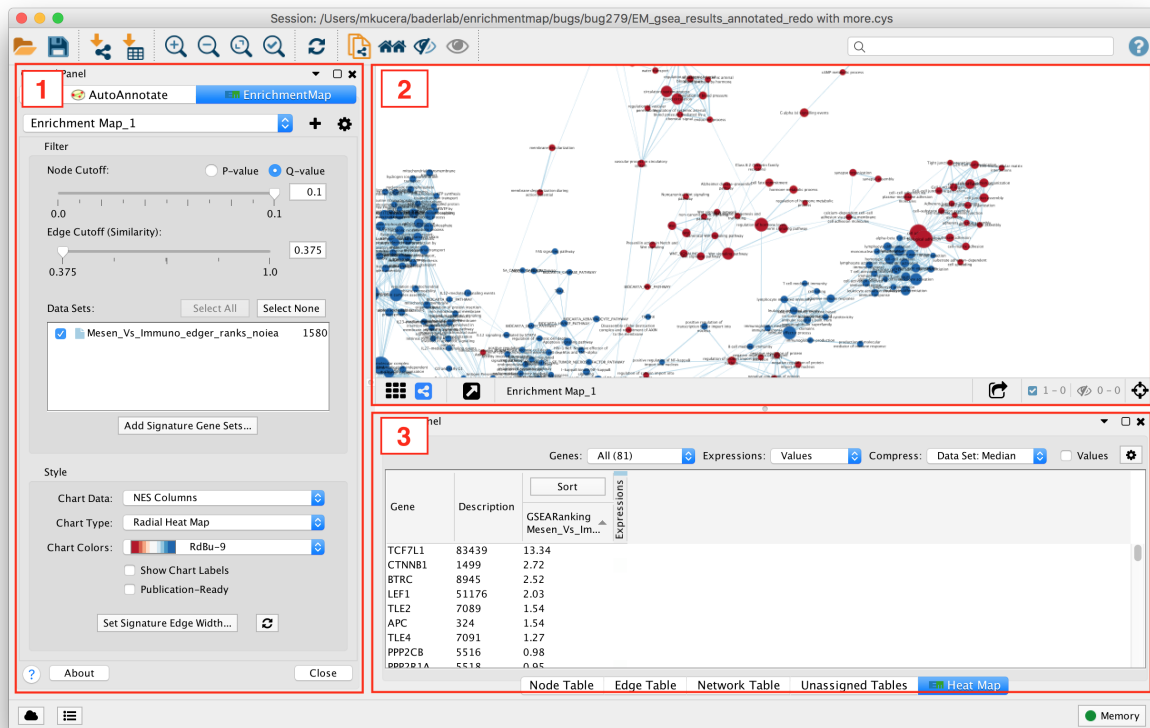
---

**Note:** For more details see [Creating a Network](#)

---



## 4.3.2 Panels



### 1. Main EnrichmentMap Panel

- Used to customize the look of the network in several ways.

### 2. EnrichmentMap Network

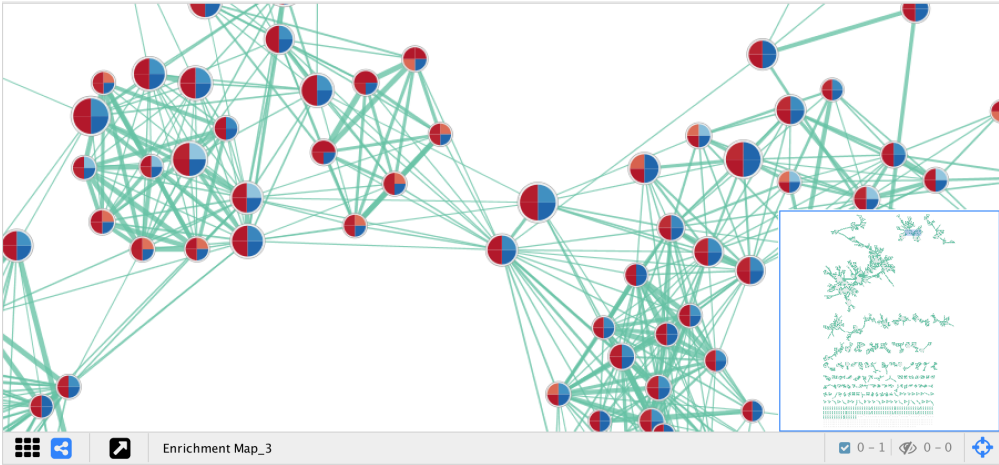
- The Cytoscape network view shows the EnrichmentMap network.

### 3. Expression Panel (Heat Map)

- Shows gene expression data for selected nodes and edges.

## 4.3.3 Interpreting the Network

- Nodes represent gene sets.
- Node size represents the number of genes in the gene set.
- Edges represent overlap between gene sets.
- Edge width represents the number of genes that overlap.
- The default layout algorithm causes gene sets with high overlap to cluster together.
- Each node contains a chart that shows the enrichment scores, such as NES (for GSEA), P-value or FDR Q-value. The enriched phenotype is conveyed by a color gradient. The chart data can be changed using the **Style** section of the EnrichmentMap panel.



EnrichmentMap creates several columns in the node and edge tables. They can be seen in the **Node Table** and **Edge Table** panels. Columns created by EnrichmentMap start with “EM”.

Table Panel

Mesen\_Vs\_Immuno\_edger\_ranks\_noiea.GseaPreranked ...

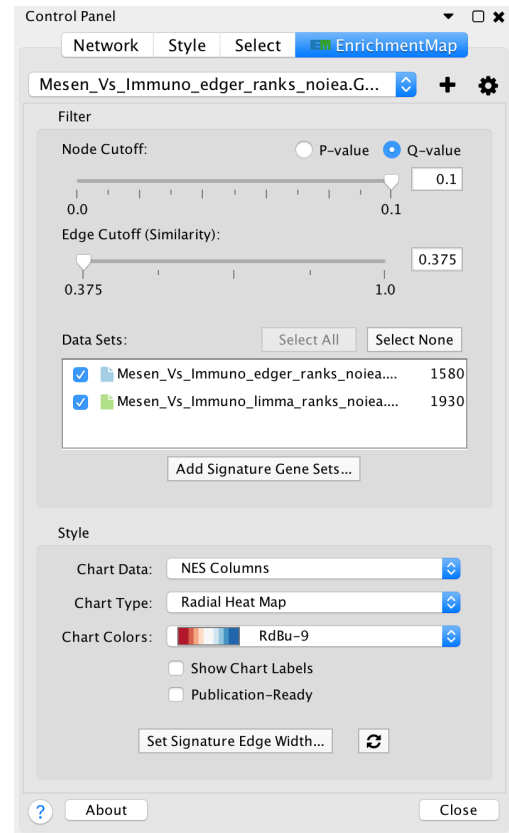
shared name	name	EM5_Name	EM5_GS_DESCR	EM5_GS_Type	EM5_Genes
TELOMERE MAINTEN...	TELOMERE...	TELOMERE MAINT...	telomere maintenance	ENR	[NBN, TERT, SP100, TERF...
TGF-BETA RECEPTO...	TGF-BETA...	TGF-BETA RECEPT...	TGF-beta receptor s...	ENR	[PPP1R15A, UBB, UBC, SM...
NEGATIVE REGULAT...	NEGATIVE ...	NEGATIVE REGUL...	negative regulation ...	ENR	[PHLDB2, RCC2, APOD, P...
RESPIRATORY ELEC...	RESPIRAT...	RESPIRATORY ELE...	Respiratory electron...	ENR	[NDUFA13, UQCRC1, NC...
MULTICELLULAR OR...	MULTICEL...	MULTICELLULAR ...	multicellular organi...	ENR	[RNF207, MYH14, SPTBN...
TRANSLATIONAL EL...	TRANSLA...	TRANSLATIONAL ...	translational elonga...	ENR	[RPL15, RPL19, RPLP2, RI...
DEGRADATION OF T...	DEGRADA...	DEGRADATION O...	Degradation of the e...	ENR	[BMP1, TIMP1, ADAMTS9...
CELL-CELL JUNCTIO...	CELL-CEL...	CELL-CELL JUNCT...	cell-cell junction or...	ENR	[CDH5, CLDN17, TLN2, C...
STEM CELL DIFFERE...	STEM CEL...	STEM CELL DIFFER...	stem cell differentiat...	ENR	[TBX2, ECE2, HGF, LOXL...
BIOCARTA_TGFB_P...	BIOCARTA...	BIOCARTA_TGFB_...	BIOCARTA_TGFB_PA...	ENR	[ZFYVE9, CREBBP, SMAD...
CELL-SUBSTRATE JU...	CELL-SUB...	CELL-SUBSTRATE ...	cell-substrate juncti...	ENR	[ACTN3, RCC2, KRT5, AC...

Node Table Edge Table Network Table Heat Map Unassigned Tables

**Note:** For more details see [Network Visualization](#).

### 4.3.4 Main EnrichmentMap Panel

The Main EnrichmentMap Panel can be used to customize the network in several ways.



- Filter section
  - **Node cutoff slider:** Nodes with a p-value or q-value that do not pass the cutoff are hidden from view.
  - **Edge cutoff slider:** Edges with a similarity score that does not pass the cutoff are hidden from view.
  - **Data set list:** Lists Data Sets that were used to create the network. De-selecting the checkbox next to the name of a data set causes gene set nodes that are not contained in the data set to be hidden from view.
  - **Add signature gene sets button:** Opens the Post Analysis dialog which is used to add more gene sets to the network.
- Style section
  - **Chart data:** Allows to pick which data columns are used by the node charts (e.g. NES, p-value or q-value).
  - **Chart type:** Various chart visualizations are available.
  - **Chart colors:** Various color schemes are available for the charts.

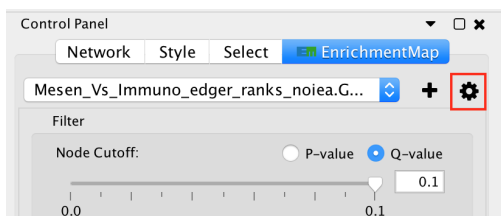
---

**Note:** For more details see [Main Panel](#)

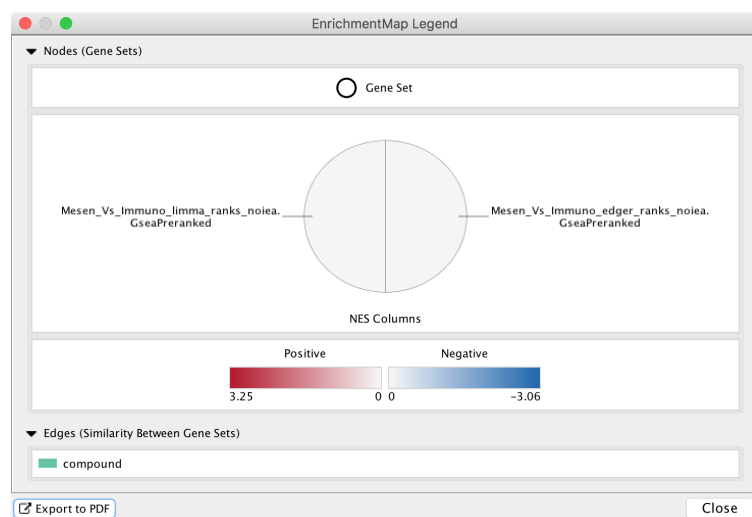
---

### 4.3.5 Legend Dialog

The legend dialog can be opened by clicking on the gear icon at the top of the main panel and selecting **Show Legend**.



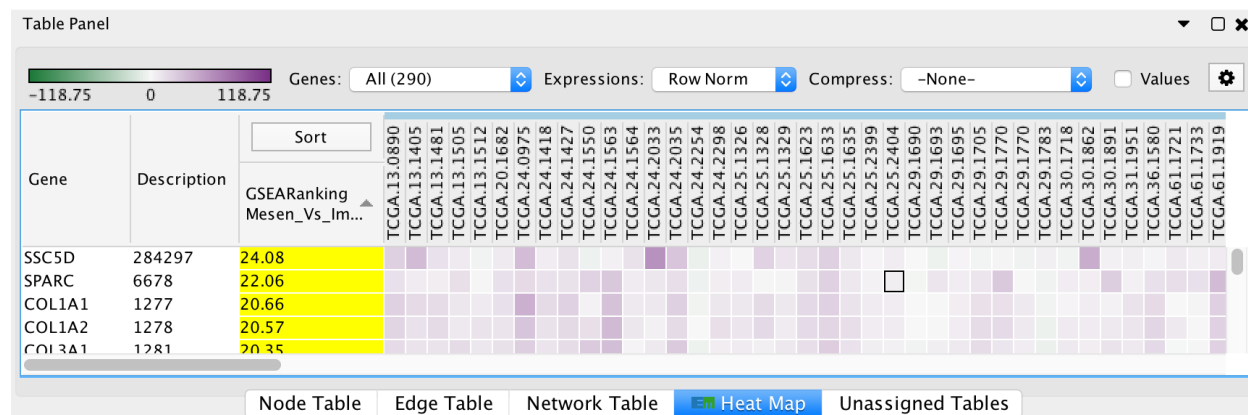
The Legend Dialog shows a visual representation of how the network elements can be interpreted, including what data is visible on the node charts. The legend can be exported as a PDF file.



**Note:** For more details see *Legend Dialog*

#### 4.3.6 Expression Panel (Heat Map)

The HeatMap panel shows gene expression data for selected nodes and edges.



- **Genes:** You may select multiple gene sets (by selecting several nodes or edges). When more than one gene set is selected you may choose to view all of the genes (union) or just the genes that are common (intersection).
- **Expressions:** Allows to show the raw expression data or to normalize the data.

- **Compression:** If there is a large number of expression columns in the HeatMap they can be compressed down to the median, min or max value.
- **Show values:** When selected shows the actual expression values, otherwise just shows the color gradient.
- The contents of the HeatMap can be exported to a TXT or PDF file.

---

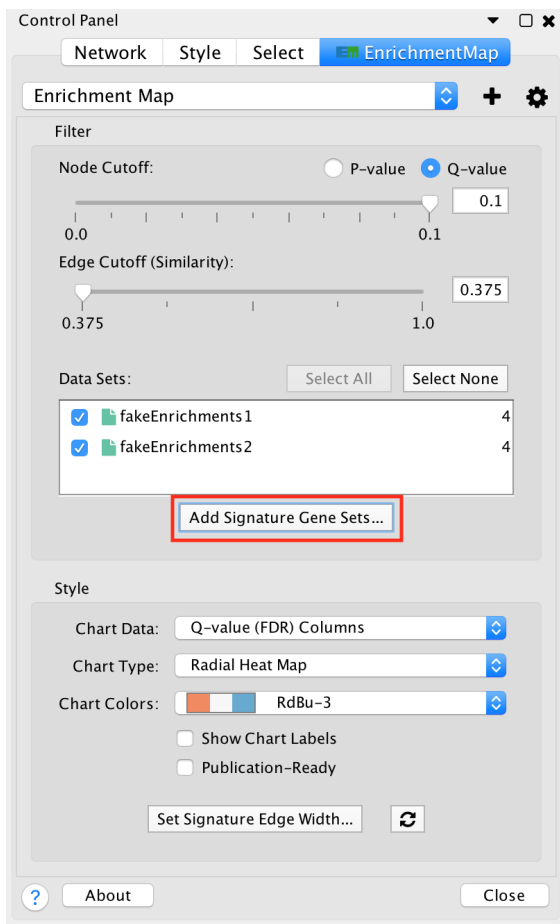
**Note:** For more details see [Expression Panel \(Heat Map\)](#)

---

### 4.3.7 Post Analysis (Add Signature Gene Sets)

The **Add Signature Gene Sets** panel allows you to add more gene sets to an existing network. This is also called **Post Analysis**.

To access the dialog click the **Add Signature Gene Sets...** button on the Main EnrichmentMap panel.



The dialog allows gene set (GMT) files to be loaded from the local file system or downloaded from the web.

Signature Gene Sets

Import	Name	Genes	Largest Overlap	Hypergeometric
<input checked="" type="checkbox"/>	NADH%DRUGBANK%DB00157	144	51	3.0124E-53
<input checked="" type="checkbox"/>	BORTEZOMIB%DRUGBANK%DB00188	16	16	3.695E-38
<input checked="" type="checkbox"/>	(3AR,6R,6AS)-6-((S)-((S)-CYCLOHEX-2-ENYL)(HYDROXY)ME...	14	14	2.2071E-33
<input checked="" type="checkbox"/>	CYT997%DRUGBANK%DB05147	15	12	2.5792E-31
<input checked="" type="checkbox"/>	2-HEXYLOXY-6-HYDROXYMETHYL-TETRAHYDRO-PYRAN-3,...	15	14	1.8447E-25
<input checked="" type="checkbox"/>	EPOTHILONE B%DRUGBANK%DB03010	11	9	1.5067E-22
<input checked="" type="checkbox"/>	EPOTHILONE D%DRUGBANK%DB01873	11	9	1.5067E-22
<input checked="" type="checkbox"/>	REGORAFENIB%DRUGBANK%DB08896	18	11	3.1297E-22
<input checked="" type="checkbox"/>	VITAMIN A%DRUGBANK%DB00162	16	9	4.3858E-21
<input checked="" type="checkbox"/>	CLADRIBINE%DRUGBANK%DB00242	9	9	3.2087E-20
<input checked="" type="checkbox"/>	ANISOMYCIN%DRUGBANK%DB07374	13	10	4.4229E-20
<input checked="" type="checkbox"/>	PUROMYCIN%DRUGBANK%DB08437	12	10	4.4229E-20
<input checked="" type="checkbox"/>	LENVATINIB%DRUGBANK%DB09078	8	8	1.5134E-19

4865 gene sets loaded, 2848 passed cutoff, 2848 selected for import

Edge Weight Parameters

Test: 

Hypergeometric Test

Cutoff: 

0.25

Data Set: 

-- All Data Sets --

Data Set Name: ☒ Use Default 

Human\_DrugBank\_all\_symbol

Hypergeometric Universe

☒ GMT (14067)

☐ Expression Set (20326)

☐ Intersection (12653)

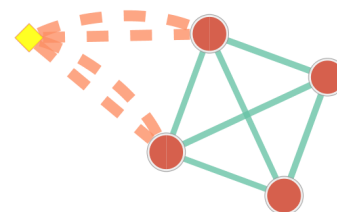
☐ User Defined: 

20,326

Advanced

Cancel

Finish



The result of running Post Analysis is a new node for each signature gene set (yellow diamond) and edges from the signature gene set to each existing gene set when the similarity passes the cutoff test. A new data set for the signature gene sets is added to the data set list on the Main EnrichmentMap panel.

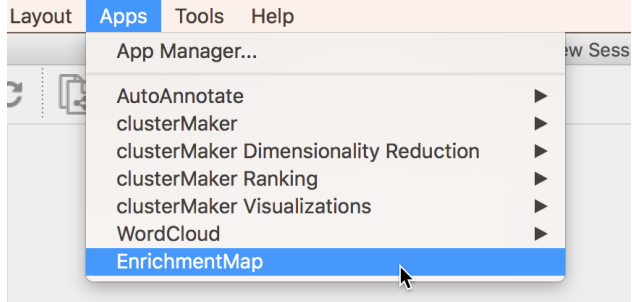
**Note:** For more details see *Post Analysis (Add Signature Gene Sets)*

### 4.3.8 EnrichmentMap Protocol

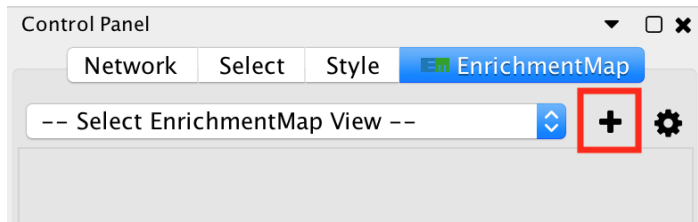
Please see [The EnrichmentMap Protocol](#) for details on automating EnrichmentMap.

## 4.4 Creating a Network

To start using EnrichmentMap go to the Cytoscape main menu and select **Apps > EnrichmentMap**. This will show the EnrichmentMap panels and open the **Create EnrichmentMap Dialog**.

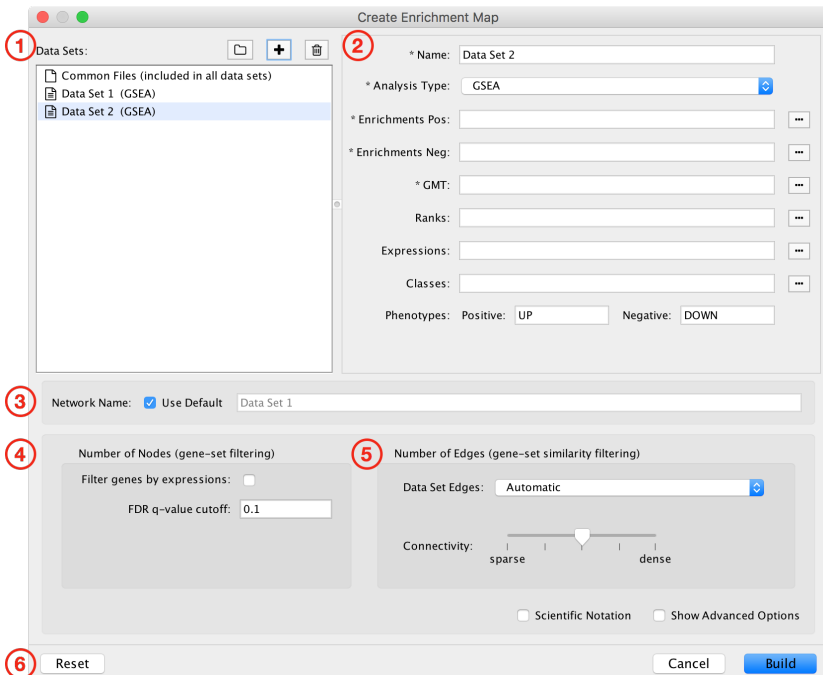


The dialog can also be opened by clicking the  button at the top of the main EnrichmentMap panel.



**Note:** See [File Formats](#) for details on the various file formats accepted by EnrichmentMap.

#### 4.4.1 Create EnrichmentMap Dialog



This dialog is used to enter paths to data files and filtering parameters.

The dialog has the following panels:

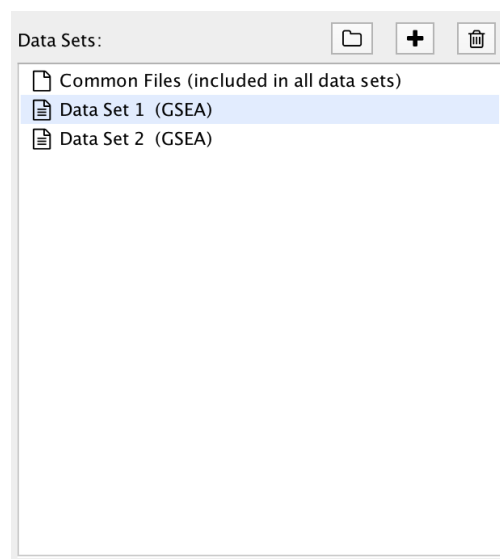
1. List of Data Sets.
2. File entry panel.

3. Network name.
4. Node filtering parameters.
5. Edge filtering parameters.
6. Action buttons.

Each of these panels will be explained in more detail below.

If the dialog is closed and then reopened everything that was previously entered will be saved. You may experiment with creating multiple EnrichmentMap networks with different parameter choices without having to enter all of the information every time. To clear out and reset all fields to their defaults click the **Reset** button at the bottom left of the dialog.

### 4.4.2 1) The Data Set List



A **Data Set** contains the results of one enrichment analysis, along with associated data such as expressions, gene sets and classes.


Selecting an entry in the Data Set List will show the file input fields for that data set.

There is a special entry called **Common Files**. Files entered on this panel will be included in all the data sets.

As of EnrichmentMap 3.0 there is no limit on the number of data sets that can be entered. However in practice adding more data sets increases the size and complexity of the resulting network.

#### Creating Data Sets by Scanning For Files


The first step is to enter the paths to the data set files. This can be a time consuming process if done manually; for that reason EnrichmentMap has the ability to scan a folder and automatically detect enrichment, expression, class and GMT files. These files are automatically assembled into data sets based on naming conventions. This scanning process works well for GSEA results because GSEA outputs a folder of results files.


To scan a folder click the  button, then select a folder. If EnrichmentMap can detect data files it will automatically add one or more data sets to the list. Scanning is based on a heuristic (that may change between versions of EnrichmentMap), so please check that the file entry panel contains the correct files after scanning.



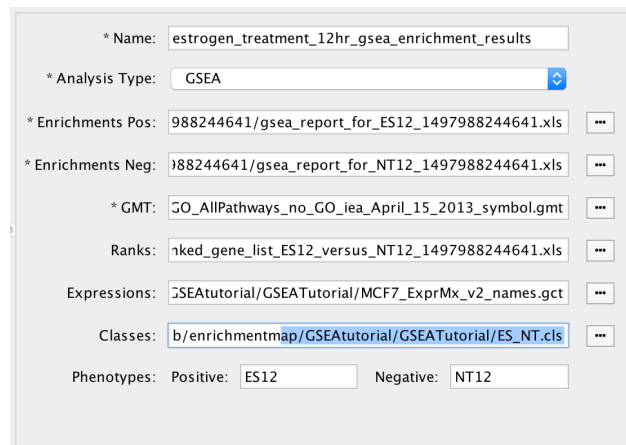
**Note:** You can also Drag-and-Drop folders onto the Data Set List to initiate a scan.

## Creating Data Sets Manually

To manually create a data set click the  button. A new data set will appear in the list and all the file input fields for that data set will be empty.

To delete a data set select it in the list and then click the  button.


### 4.4.3 2) File Entry Panel



The file entry panel has the following fields:

- Data set name
  - The data set can be named anything. However two data sets may not have the same name. When scanning for files a name will be automatically chosen based on the file name of the enrichment file(s).
- Analysis type
  1. GSEA
    - Takes as inputs the output files created in a GSEA analysis. When GSEA is chosen there will be two input fields for enrichment files. GSEA analysis always has two enrichment results files, one for each of the phenotypes compared.
  2. Generic/gProfiler
    - 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.
  3. David/BiNGO/Great
    - Has no GMT or expression file requirement and takes as input enrichment result file as produced by DAVID, BiNGO or GREAT tools.
- File input fields

- There are input fields for Gene Set, Enrichment, Expression, Rank and Class files. Fields with a \* next to their name are required, all other fields are optional. Which fields are required depends on the analysis type.

- Click the  button next to an input field to open a file browser.

- Phenotypes

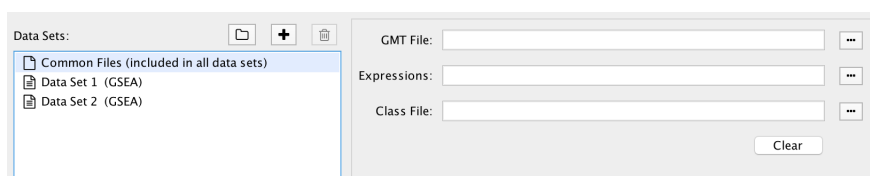
- Enter the names of two classes from the class file that are being compared in the enrichment analysis. When a class file is entered the dialog will parse the class file and automatically fill in these fields.
- These phenotypes will be highlighted in the *Expression Panel (Heat Map)*.

---

**Note:** See *File Formats* for details on the various file formats accepted by EnrichmentMap.

---

### Common Files



Select *Common Files* at the top of the data set list to show a special file entry panel. GMT, expression and class files entered on this panel will be included in all the data sets.

Files on the *Common Files* panel will override files entered in the individual data set panels.

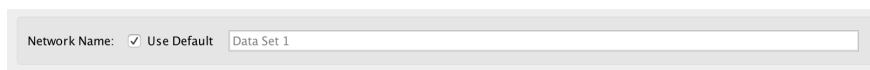
Click the **Clear** button to clear out all three file input fields.

---

**Note:** Even though *Common Files* is located inside the data set list it is not a data set.

---

### 4.4.4 3) Network Name



EnrichmentMap will automatically generate a name for the network based on the name of the first data set in the list.

If you want to choose your own name for the network de-select the **Use Default** checkbox then enter the name.

The name of the network can always be changed after it has been created. Go to the **Network** tab, right click the network name, and select **Rename Network**.

### 4.4.5 4) Gene Set (Node) Filtering

Gene sets must pass the following criteria to be included in the network.

## Basic Options

Number of Nodes (gene-set filtering)

Filter genes by expressions: ☐

FDR q-value cutoff:

- Filter gene sets by expressions
  - If selected genes that are contained in the gene set (GMT) files or the enrichment files, but are not contained in the expression files will not be included in the network.
- FDR q-value cutoff
  - Gene set with a q-value lower than the one entered will not be included in the network.

## Advanced Options

Available when the **Show Advanced Options** checkbox at the bottom right of the dialog is selected.

Number of Nodes (gene-set filtering)

Filter genes by expressions: ☐

FDR q-value cutoff:

p-value cutoff:

NES (GSEA only):

Filter by minimum experiments: ☐

Minimum experiments:

- p-value cutoff
  - Gene sets with a p-value lower than the one entered will not be included in the network.
  - The default value of 1.0 will not cause any gene sets to be removed from the network.
- NES (GSEA only)
  - Positive: Only gene sets from the positive enrichment file will be included.
  - Negative: Only gene sets from the negative enrichment file will be included.
  - All: Both enrichment files will be included
- Filter by minimum experiments
  - Selected this to enable the *Minimum experiments* field.
- Minimum experiments
  - A gene set must be included in this many data sets to be included in the network.

---

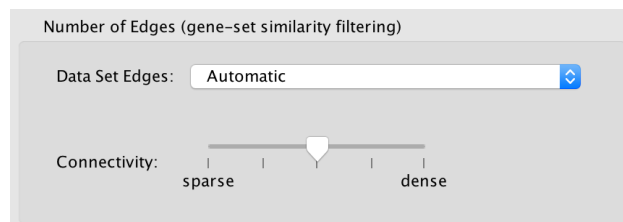
**Note:** See [Tips on Parameter Choice](#) for more details on how to tune gene set filtering.

---

### 4.4.6 5) Gene Set Similarity (Edge) Filtering

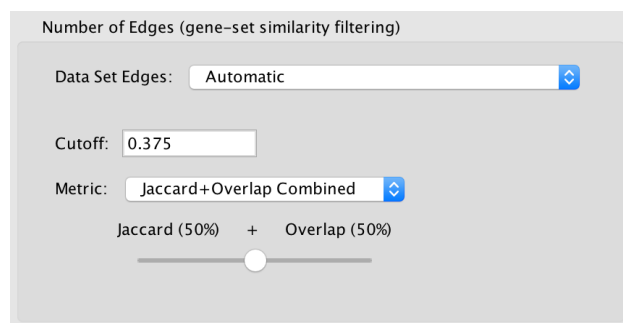
A similarity score is computed for every pair of gene sets based on how many genes they have in common (set intersection). If the similarity score passes the following criteria then an edge will be created between the gene set nodes.

#### Basic Options



- Data set edges (Note: This option has no effect if there is only one data set)
  - Separate edge for each data set
    - \* If a gene set is associated with more than one data set it is possible for the contents of the gene set to be different in each data set. This often happens when the data sets have different expression files and the “*filter gene sets by expressions*” option is enabled. A separate similarity score will be computed for each data set resulting in potentially many more edges and a much denser network.
  - Combine edges across data sets
    - \* Gene sets with the same name are combined (set union) and then the similarity score is calculated.
    - \* There will be at most one edge between a pair of gene set nodes.
  - Automatic (*default*)
    - \* EnrichmentMap decides which of the above options to use.
    - \* If there are exactly two data sets and they have different expression files then *separate edges* is chosen, otherwise *combine edges* is chosen. This is done to be consistent with the behavior of EnrichmentMap 2.0.
- Connectivity
  - Moving the slider towards *sparse* will produce fewer edges, moving it towards *dense* will produce more edges.

#### Advanced Options



When *show advanced options* is enabled the *Connectivity* slider is replaced with options that allow greater control over the number of edges in the network.

- Cutoff
  - Edges with a similarity score lower than the one entered will not be included in the network.
- Metric
  - Used to choose the formula used to calculate the similarity score.
  - Jaccard Coefficient
 

$$\text{Jaccard Coefficient} = [\text{size of (A intersect B)}] / [\text{size of (A union B)}]$$
  - Overlap Coefficient
 

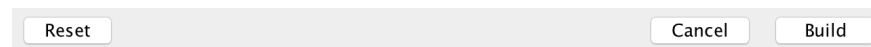
$$\text{Overlap Coefficient} = [\text{size of (A intersect B)}] / [\text{size of (minimum( A , B))}]$$
  - Combined
    - \* Merges the Jaccard and Overlap coefficients.
    - \* When selected a slider appears allowing to adjust the percentage of each coefficient to use.

---

**Note:** See [Tips on Parameter Choice](#) for more details on how to tune gene set filtering.

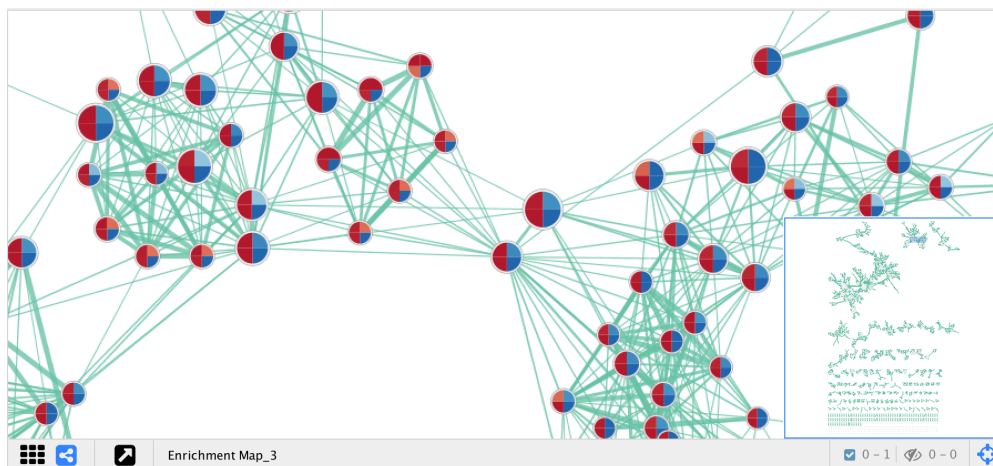
---

#### 4.4.7 6) Action Buttons



- Reset
  - Clears out and resets all fields to their defaults.
- Build
  - Creates the EnrichmentMap network.
  - First runs validation on the inputs. If there are any problems (eg. required fields missing, duplicate data set names) a error dialog is shown. The problems must be fixed before the network can be created.
  - This is a potentially long running task.
- Cancel
  - Close the dialog without creating a network.

## 4.5 Network Visualization



### 4.5.1 Style

EnrichmentMap creates a separate visual style for each network. The visual style has the following characteristics:

- Nodes represent gene sets.
- Node size represents the number of genes in the gene set.
- Edges represent overlap (similarity) between gene sets.
- Edge width represents the number of genes that overlap between a pair of gene sets.
- Node fill represents...
  1. Enrichment scores, such as NES (for GSEA), p-value or q-value.
    - When there is only 1 data set selected the enriched phenotype is conveyed using a color gradient.
    - When there are 2 or more data sets then nodes are overlaid with charts, where each segment of the chart shows enrichment using a color gradient.
  2. Data set membership
    - The *Color by Data Set* chart shows which data sets a gene set is a member of.

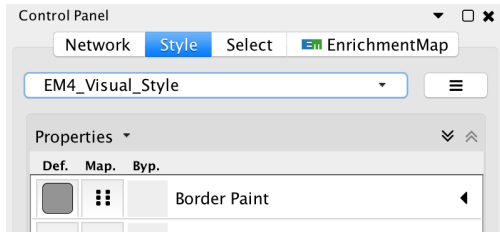
The network is arranged by a force directed layout which causes gene sets with high overlap to cluster together.

Expression data is visualized in a separate panel called the Heat Map panel. For more details see [Expression Panel \(Heat Map\)](#).

### 4.5.2 Visual Properties

**Warning:** EnrichmentMap automatically maintains the visual properties described below. If you manually change these visual properties EnrichmentMap will overwrite your changes. If you want to create a custom visual style you must first create a copy and then make changes to the copy.



Visual properties are available on the **Style** tab of the Control Panel.



### 4.5.3 Node Visual Properties

There are two types of nodes:

1. Enrichment gene set nodes
  - Regular gene set nodes that are created when the network is first created.
  - Enrichment nodes can have many different visualizations depending on the settings in the Style section of the *Main Panel*.
2. Signature gene set nodes
  - Added to an existing network by *Post Analysis (Add Signature Gene Sets)*.
  - Do not have chart visualizations.
  - Edges connected to signature nodes are dashed.

Enrichment	Signature
	

### Visual properties

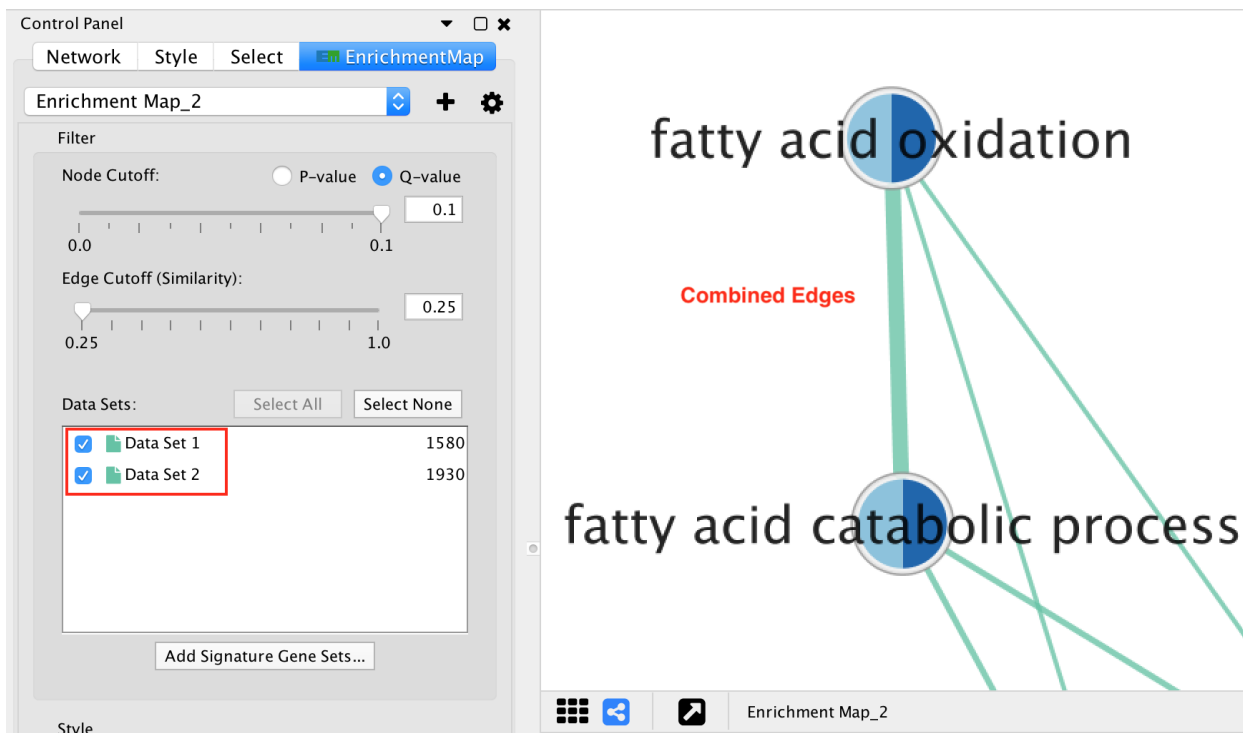
Visual Property	Meaning	Type	Column(s)
Shape	Gene Set Type	Discrete Mapping	EM#_GS_Type (ENR = Square, SIG = Diamond)
Fill Color	NES, p/q-value	Discrete Mapping	EM#_pvalue, EM#_fdr_qvalue, EM_NES
Label	Gene set name	Passthrough Mapping	EM#_GS_DESCR
Size	Size of gene set	Continuous Mapping	EM#_gs_size
Image/Chart 1	NES, p/q-value	Chart	EM#_pvalue, EM#_fdr_qvalue, EM_NES

### 4.5.4 Edge Visual Properties

If there are 2 data sets, and/or the *Separate edge for each data set* option was chosen, then edges will have different colors for each data set. The edge color corresponds to the color of the icon next to the data set name in the main panel.

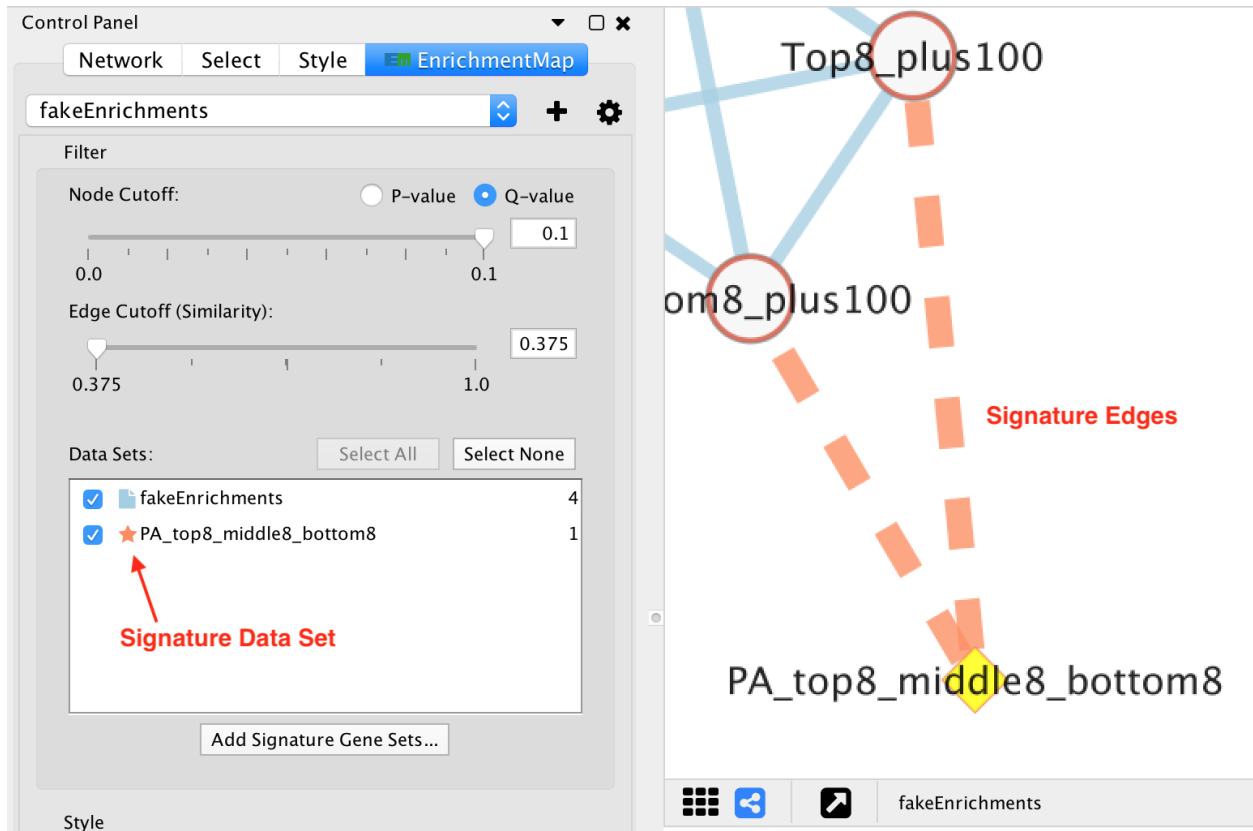


If the network has only one data set, or if the *Combine edges across data sets* option was chosen, then all the edges between enrichment gene sets will be the same color.



Edges connected to signature gene sets have a different color and are dashed.





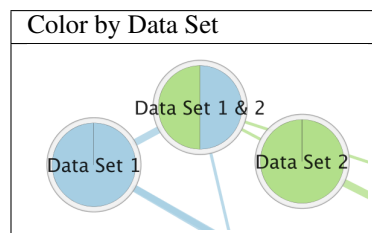
## Visual properties

Visual Property	Meaning	Type	Column(s)
Line Type	Edge Type	Discrete Mapping	interaction (default = solid, sig = dashed)
Stroke Color	Data Set or Signature	Discrete Mapping	EM#_Data Set
Width	Size of gene set overlap	Continuous Mapping	EM#_similarity_coefficient

## 4.5.5 Chart Visualization

There are 6 types of chart visualization available (including *None*).

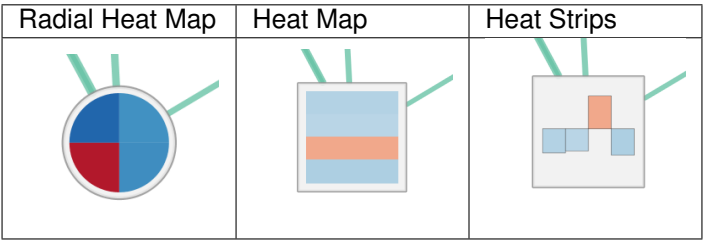
### Color by Data Set



The *Color by Data Set* pie chart shows which data sets each gene set is a member of. The segment color corresponds to the color of the icon next to the data set name in the main panel.

Enrichment

There are 3 charts available for visualizing enrichment values.

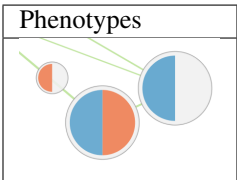


For *Radial Heat Map* and *Heat Map* each segment of the chart is equal size and represents the enrichment value from one data set. For *Heat Strips* the size of each segment also reflects the enrichment value.

The color of each chart segment is a color gradient indicating the enrichment value. The default color scheme shows up-regulated scores in red and down-regulated scores in blue.

*g:Profiler* Phenotypes

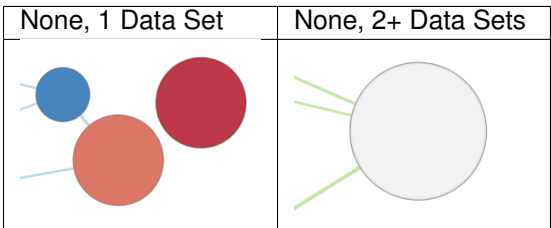
There is a special chart called *Phenotypes* that is only available for *g:Profiler* 2-data set analysis.



The color gradient of each segment indicates the enriched phenotype, calculated using the formula  $1 - p\text{-value}$  multiplied by the sign of the ES score (if using GSEA mode) or the phenotype (if using the Generic mode).

This color gradient is similar to the visualization used by EnrichmentMap 2.0.

None



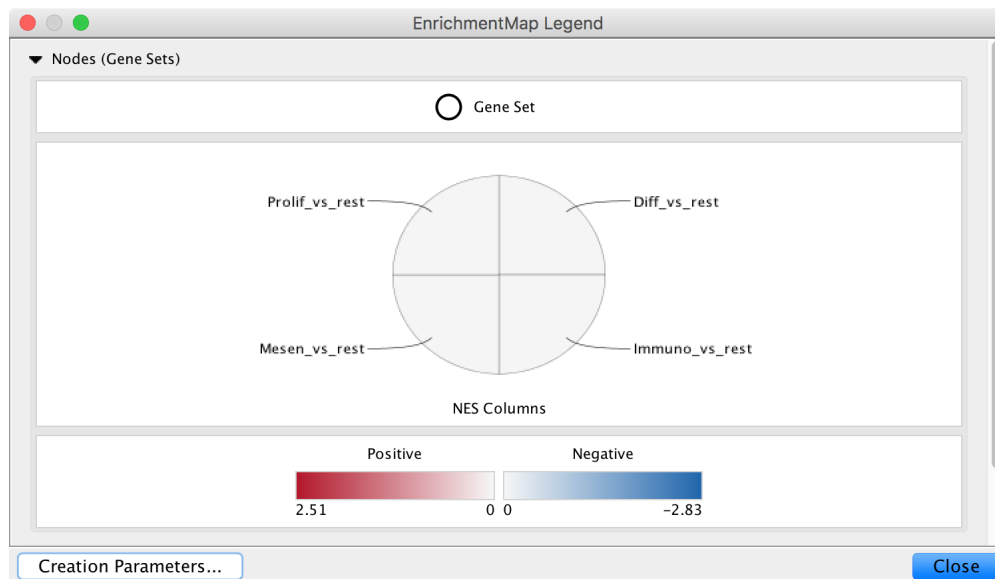
If there are 2 or more data sets then selecting *None* will show a dull gray color on each node that has no meaning.

If there is exactly 1 data set selected then the node fill color will be a color gradient that indicates the enriched phenotype, calculated using the formula  $1 - p\text{-value}$  multiplied by the sign of the ES score (if using GSEA mode) or the phenotype (if using the Generic mode).

This color gradient is similar to the visualization used by EnrichmentMap 2.0.

### 4.5.6 Legend

The legend dialog can be used to see which chart segment corresponds to which data set and the meaning of the colors.



Use the *Style* section of the *Main Panel* to change the type chart and the color scheme.

Style

Chart Data: NES Columns

Chart Type: Radial Heat Map

Chart Colors: RdBu-9

☒ Show Chart Labels

☐ Publication-Ready

Set Signature Edge Width...

↺

## 4.6 Columns

EnrichmentMap creates several columns in the node and edge tables. They can be seen in the **Node Table** and **Edge Table** panels.

Table Panel

Mesen\_Vs\_Immuno\_edger\_ranks\_noiea.GseaPreranked ...

shared name	name	EM5_Name	EM5_GS_DESCR	EM5_GS_Type	EM5_Genes
TELOMERE MAINTEN...	TELOMERE...	TELOMERE MAINT...	telomere maintenance	ENR	[NBN, TERT, SP100, TERF
TGF-BETA RECEPTO...	TGF-BETA...	TGF-BETA RECEPT...	TGF-beta receptor s...	ENR	[PPP1R15A, UBB, UBC, SA
NEGATIVE REGULAT...	NEGATIVE ...	NEGATIVE REGUL...	negative regulation ...	ENR	[PHLDB2, RCC2, APOD, P
RESPIRATORY ELEC...	RESPIRAT...	RESPIRATORY ELE...	Respiratory electron...	ENR	[NDUFA13, UQCRC1, NC
MULTICELLULAR OR...	MULTICEL...	MULTICELLULAR ...	multicellular organi...	ENR	[RNF207, MYH14, SPTBN
TRANSLATIONAL EL...	TRANSLA...	TRANSLATIONAL ...	translational elonga...	ENR	[RPL15, RPL19, RPLP2, R
DEGRADATION OF T...	DEGRADA...	DEGRADATION O...	Degradation of the e...	ENR	[BMP1, TIMP1, ADAMTS9
CELL-CELL JUNCTIO...	CELL-CEL...	CELL-CELL JUNCT...	cell-cell junction or...	ENR	[CDH5, CLDN17, TLN2, C
STEM CELL DIFFERE...	STEM CEL...	STEM CELL DIFFER...	stem cell differentiat...	ENR	[TBX2, ECE2, HGF, LOXL3
BIOCARTA_TGFB_P...	BIOCARTA...	BIOCARTA_TGFB_...	BIOCARTA_TGFB_PA...	ENR	[ZFYVE9, CREBBP, SMAD3
CELL-SUBSTRATE JU...	CELL-SUB...	CELL-SUBSTRATE ...	cell-substrate juncti...	ENR	[ACTN3, RCC2, KRT5, AC

Node Table Edge Table Network Table Heat Map Unassigned Tables

Columns created by EnrichmentMap have the following pattern:

EM#\_column\_name (data set name)

Where...

- # is a number that is automatically assigned to each network
- *Data set name* is used to differentiate between data sets. There will be one such column for each data set.

## 4.6.1 Node Columns

**EM#\_Name** The gene set name.

**EM#\_Formatted\_name** A wrapped version of the gene set name so it is easy to visualize.

**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.

**EM#\_gs\_size** Number of genes the union of the gene set across all data sets.

**EM#\_GS\_Type** Used by the visual style to discern between regular enrichment nodes and signature gene set nodes.

Additionally there are attributes created for each dataset:

**EM#\_pvalue (...)** Gene set p-value, as specified in GSEA enrichment result file.

**EM#\_fdr\_qvalue (...)** Gene set q-value, as specified in GSEA enrichment result file.

**EM#\_Colouring (...)** 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 (...)** Enrichment score, as specified in GSEA enrichment result file.

**EM#\_NES\_dataset (...)** Normalized Enrichment score, as specified in GSEA enrichment result file.

**EM#\_fwer\_qvalue (...)** Family-wise error score, as specified in GSEA enrichment result file.

## 4.6.2 Edge Columns

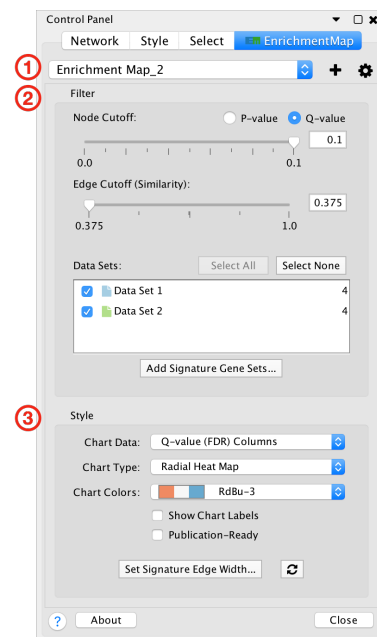
For each Enrichment map created the following attributes are created for each edge:

**EM#\_Data Set** Contains the name of the data set that the edge is associated with, or ‘compound’ if the *Combine edges across data sets* option was selected when the network was created.

**EM#\_Overlap\_size** The number of genes associated with the overlap of the two gene sets that this edge connects.

**EM#\_Overlap\_genes** The names of the genes that are associated with the overlap of the two gene sets that this edge connects.

**EM#\_similarity\_coefficient** The calculated coefficient for this edge.



## 4.7 Main Panel

The main EnrichmentMap panel has the following sections:

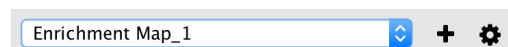
1. Toolbar
2. Filter section
3. Style section

Each of these sections will be explained in more detail below.



If the main panel is not visible go to the Cytoscape main menu and select **Apps > EnrichmentMap**.

To hide the panel click the **Close** button at the bottom right.

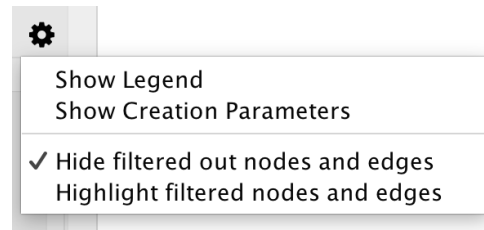
### 4.7.1 1) Toolbar



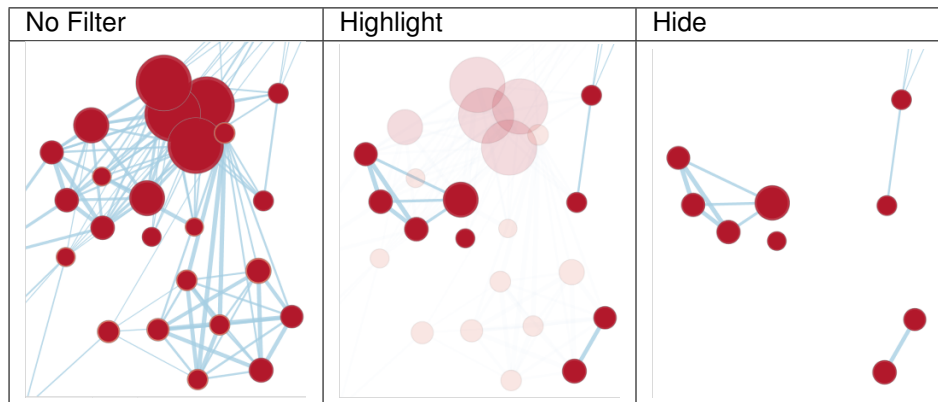
- Network combo box
  - This combo box can be used to quickly switch between EnrichmentMap networks without having to navigate to the *Network* tab. Only networks created by EnrichmentMap are listed.

- Plus button 
  - Opens the *Create EnrichmentMap Dialog*.
- Gear button 
  - Opens the panel menu (explained in more detail below).

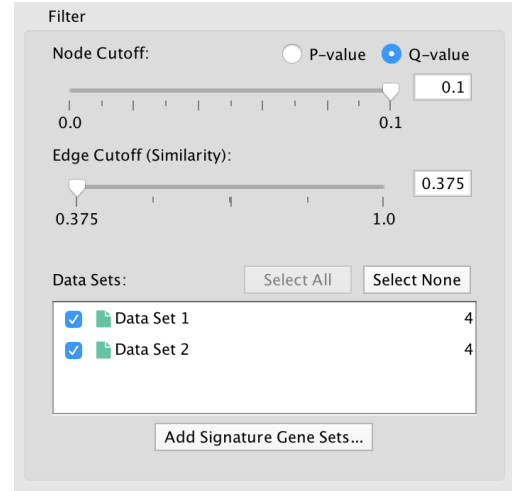
## Panel Menu



- Show Legend
  - Opens the *Legend Dialog*.
- Show Creation Parameters
  - Opens the *Creation Parameters Dialog*.
- Hide/Highlight filtered nodes and edges
  - Changes the appearance of nodes and edges that are filtered out. See [2\) Filter Section](#) below.



## 4.7.2 2) Filter Section



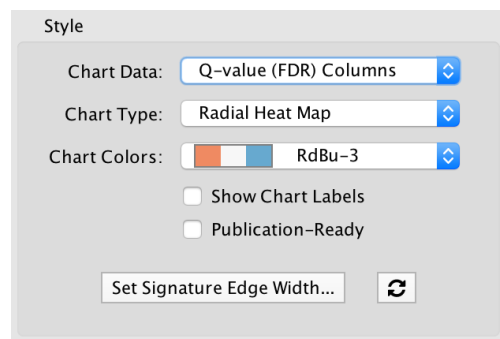
The filter section is used to hide nodes and edges in the network.

- Node cutoff
  - Use the radio buttons to switch between p-value and q-value.
  - The slider is initially set all the way to the right, which corresponds to the value that was entered in the *Create EnrichmentMap Dialog* when the network was created.
  - As the slider is moved to the left nodes with a p-value/q-value greater than the cutoff are hidden. Edges connected to hidden nodes are also hidden.
  - P-values can be found in the *Node Table* in columns that start with *EM#\_pvalue*.
  - Q-values can be found in the *Node Table* in columns that start with *EM#\_fdr\_qvalue*.
- Edge cutoff
  - This slider is initially set all the way to the left, which corresponds to the smallest edge similarity score in the network.
  - As the slider is moved to the right edges with a similarity score less than the cutoff are hidden.
  - Similarity scores can be found in the *Edge Table* in the column named *EM#\_similarity\_coefficient*.
- Data Sets list
  - The data set list shows then names of all the data sets as well as the number of gene sets in each data set.
  - Initially the checkbox next to each data set is selected.
  - De-selecting the checkboxes hides gene set nodes that are only contained in those data sets.
- Add Signature Gene Sets button
  - Click to open the *Post Analysis (Add Signature Gene Sets)* dialog.

The number of hidden nodes and edges can be seen in the status bar under the network view.



### 4.7.3 3) Style Section




The style panel is mainly used to manipulate chart visualizations on nodes.

For more details on chart visualizations see [Chart Visualization](#).

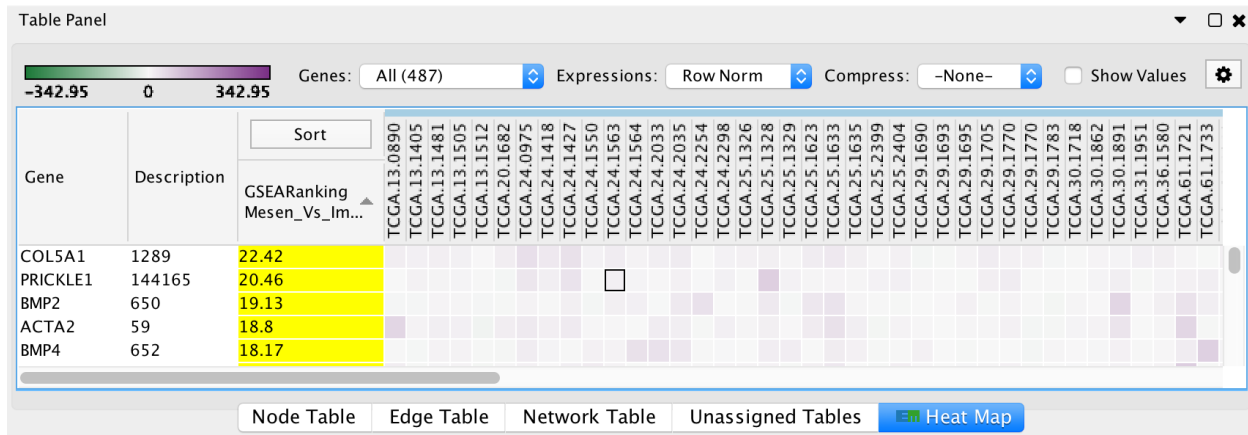
- Chart Data
  - – None –
    - \* If there is 1 data set then node shows a pre-computed color gradient for the p-value. If there are 2 or more data sets then the node color has no meaning and is set to grey.
  - NES Columns
    - \* Enrichment values from the *EM#\_NES* columns are used.
    - \* Only available if the analysis type is GSEA.
  - P-value Columns
    - \* Enrichment values from *EM#\_pvalue* columns are used.
  - Q-value (FDR) Columns
    - \* Enrichment values from *EM#\_fdr\_qvalue* columns are used.
  - Color by Data Set
    - \* Pie chart that shows which data sets each gene set is a member of.
  - Phenotypes
    - \* Only available for *g:Profiler* 2-data set analysis.
    - \* The color gradient of each chart segment indicates the enriched phenotype.
- Chart Type
  - Field is enabled if *Chart Data* is set to *NES Columns*, *P-value Columns* or *Q-value Columns*.
  - Various chart types are available, for more details see [Chart Visualization](#).
- Color Scheme
  - Several [Color Brewer](#) colorblind safe palettes are available.
  - When *NES Columns* is chosen for *Chart Data* then the **RdBu-9** palette will be available. This palette is the same as the standard color gradient used in EnrichmentMap 2.0.
- Show chart labels
  - Enable this option to show enrichment values for each chart segment.
- Publication-Ready



- Makes the network view ready for printing. Removes node labels and sets the network background to white.
- Set Signature Edge Width...
  - Opens a dialog that has several options for how the width of signature edges is calculated. For more details see [Edge Width Dialog](#).
- Refresh Button 
  - Resets the visual style.
  - Sometimes Cytoscape does not update the visual style properly. To fix any inconsistencies between the network view and the Style section of the main panel click this button.

## 4.8 Expression Panel (Heat Map)

The Heat Map panel shows expression values for genes that are associated with selected nodes and edges.



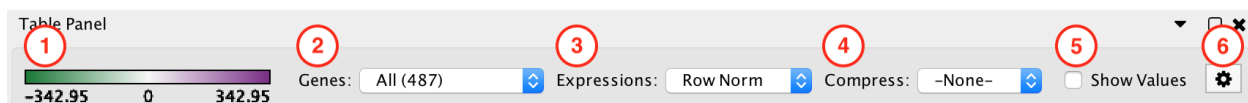
You may select any number of nodes and/or edges in the network. Selecting an edge is equivalent to selecting the two gene set nodes that are connected to the edge.

**Note:** In EnrichmentMap 2.0 there were two separate Heat Map panels for nodes and edges. In EnrichmentMap 3.0 they have been combined into a single panel.

**Note:** “Where are my expression values?”

If there are more than 50 expression values per gene then the *Compress: Data Set Median* option will be automatically enabled. Set *Compress* to *None* to see the original values. See below for more details.

### 4.8.1 Toolbar



- ## 1. Expression Legend

- Click on an expression value cell in the table to show the Expression Legend. It will not be visible until a value is selected.
- Shows the value range and the color gradient for the data set associated with the selected expression value. Note the value range may be different for different data sets.

### 2. Genes

- All
  - Shows all of the genes from all of the selected gene sets (union).
- Common
  - Shows only genes that are common to all selected gene sets (intersection).

### 3. Expressions

- Values
  - Shows the raw values from the expression file(s). Expression values are rounded to two decimal places.
- Row Norm
  - Row normalizes the expression values. 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
  - Takes the log of each expression value.

### 4. Compress

- -None-
  - Shows all of the expression values.
- Class: Median, Min, Max
  - Only available if a class file was provided when the network was created.
  - Shows a single column for each class where the value is the median, min or max of the values in the class.
- Data Set: Median, Min, Max
  - Shows a single column for each data set where the value is the median, min or max of all the values.
  - If the number of expressions per gene is greater than 50 then *Compress: Median* will be automatically enabled.

### 5. Show Values

- When disabled only the color gradients are shown. When enabled the expression values are shown.
- Expression values are rounded to two decimal places.

### 6. Gear button

- Opens the panel options menu.

## 4.8.2 Table

Gene	Description	GSEARanking estrogen_tr...	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
HEY2	HEY2 (hairy...	2.34																		
DYNLT3	"DYNLT3 (d...	1.25																		
HELLS	"HELLS (heli...	1.23																		
RRS1	RRS1 (RRS1...	1.18																		
HIST1H4C	"HIST1H4C ...	1.16																		
TIPIN	TIPIN (TIME...	1.15																		
ATR	ATR (ataxia...	1.04																		

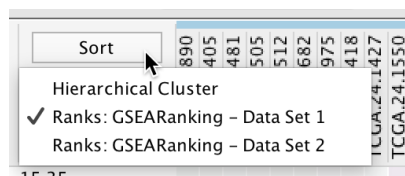
Click on any of the column headers to sort the table by that column.

1. Gene Column

2. Description Column

3. Sort Column

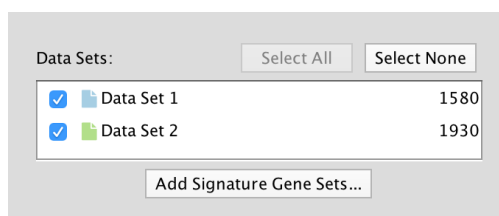
- This column is used to sort by ranks or by hierarchical clustering.
- Click the **Sort** button to show a menu of ranking options.

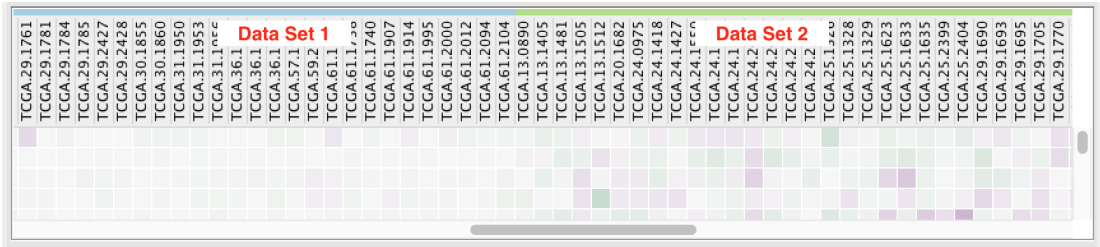


- If a data set has a rank file then the ranks will be listed in the menu.
- See [Panel Menu](#) below for details on how to load additional rank files.
- Hierarchical Clustering: Genes are clustered using a hierarchical clustering algorithm based on their expression values, the resulting hierarchy is then used to sort the genes.

4. Expression Columns

- Shows expression values for each experiment.
- If there is more than one data set and each data set has common expression values then the values will only be shown once.
- If there are two or more data sets and they have different expression values then all the expression values are shown.
  - A colored bar that runs along the top of the expression column headers can be used to differentiate between the data sets. The color of the bar corresponds to the color shown next to the data set name in the main panel.





- Genes that do not have expression data are shown in gray.

Gene	Description	Sort	GSEARanking									
			Mesen_Vs_Im...									
GUSB	2990	-4.65	TCGA.13.0890	TCGA.13.1405	TCGA.13.1481	TCGA.13.1505	TCGA.13.1512	TCGA.20.1682	TCGA.24.0975	TCGA.24.1418		
NUP85	79902	-5.4										
PCK2	5106	-5.9										
GALK1	2584	-6.25										
ALDOB												
AMY1B												
AMY1C												
AMY2A												

5. Phenotype Highlight

- The phenotypes that were entered in the *Create EnrichmentMap Dialog* are highlighted.

6. Leading Edge

- Genes that are part of the leading edge are highlighted in yellow.
- Available for GSEA results when a single gene set is selected.
- See below for more details.

4.8.3 GSEA Leading Edge

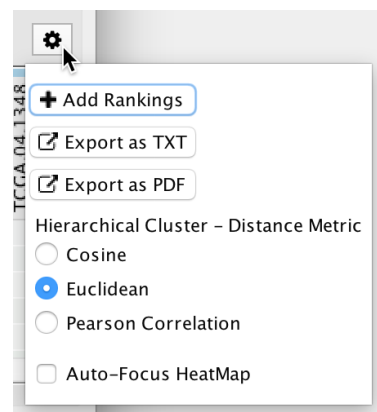
For every gene set that is tested for significance using GSEA there is a set of proteins in that gene set defined as the Leading Edge. According to GSEA the leading edge is:

*“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.

#### 4.8.4 Panel Menu



- Add Rankings
  - Opens a pop-up dialog that allows you to load an additional rank file. See [Add Ranks Dialog](#) below for more details.
- Export as TXT
  - Export the expressions currently being viewed in the heat map table as a tab-separated text file. The first line of the file contains the table headers.
  - If the heat map is showing the leading edge then you will be prompted to save just the genes that are part of the leading edge or all the genes.
- Export as PDF
  - Export the the expressions currently being viewed in the heat map table as a PDF file.
  - The visual state of the table is reflected in the PDF file. For example to show the expression values in the PDF file enable the *show values* option in the toolbar.
- Hierarchical Cluster - Distance Metric
  - Allows to select the distance metric used by the hierarchical cluster algorithm.
- Auto-Focus HeatMap
  - If enabled then every time a node/edge is selected the HeatMap panel will be brought to the front.

---

**Note:** Prior to EnrichmentMap 3.1 the *Export as PDF* option created a PDF file that just contained an image of the table data. As of EnrichmentMap 3.1 the PDF file now contains a proper PDF table that can be copy-pasted and searched.

---

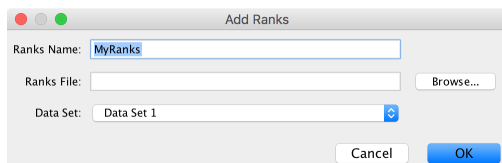


---

**Note:** Auto-Focus HeatMap was enabled by default in EnrichmentMap 2.0. It is now disabled by default as of EnrichmentMap 3.0.

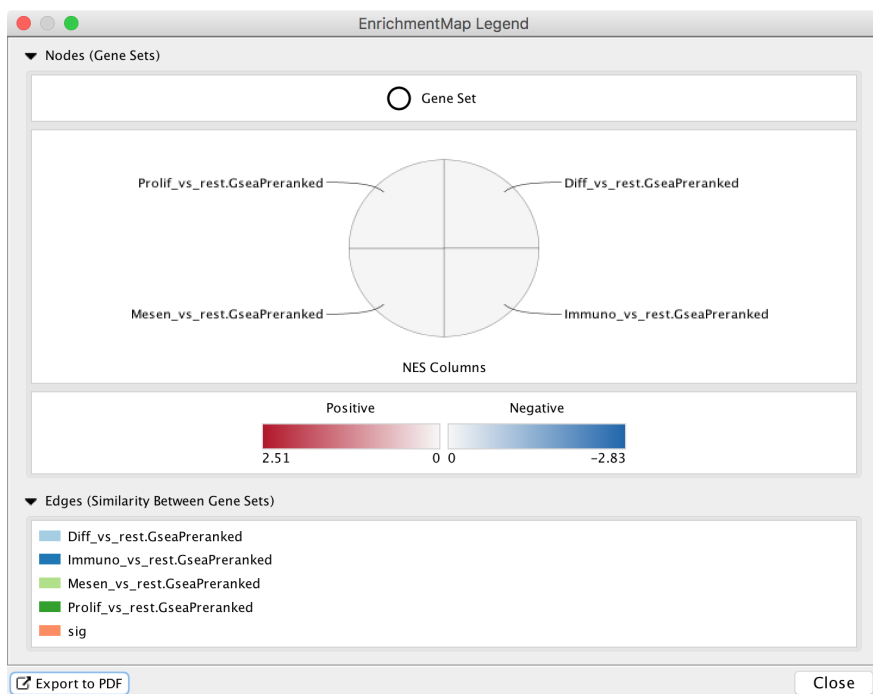
---

## 4.8.5 Add Ranks Dialog



Used to load additional ranks files into an existing data set.

## 4.9 Legend Dialog



The legend dialog provides a legend for the current network visualization.

The legend dialog can be left open. It will update automatically when changes are made to the network visualization using the [3\) Style Section](#) of the main panel.

### 4.9.1 Sections

#### Node Color



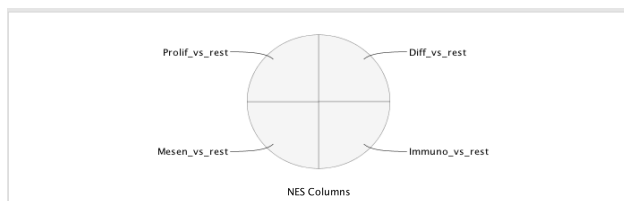
- Enabled when there is one data set and chart data is set to none.
- Shows the range for the node fill color gradient.

## Node Shape



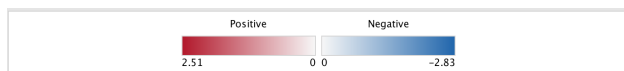
- Shows the meaning of node shape.
- Shows signature nodes after post analysis has been run.

## Node Chart



- Shows the meaning of the node charts.
- Shows how the chart segments map to the data sets.

## Node Chart Colors



- Shows the color range for the chart segment color gradients.

## Data Set Colors



- Shows the color for each data set when the *Color by Data Set* option is chosen.

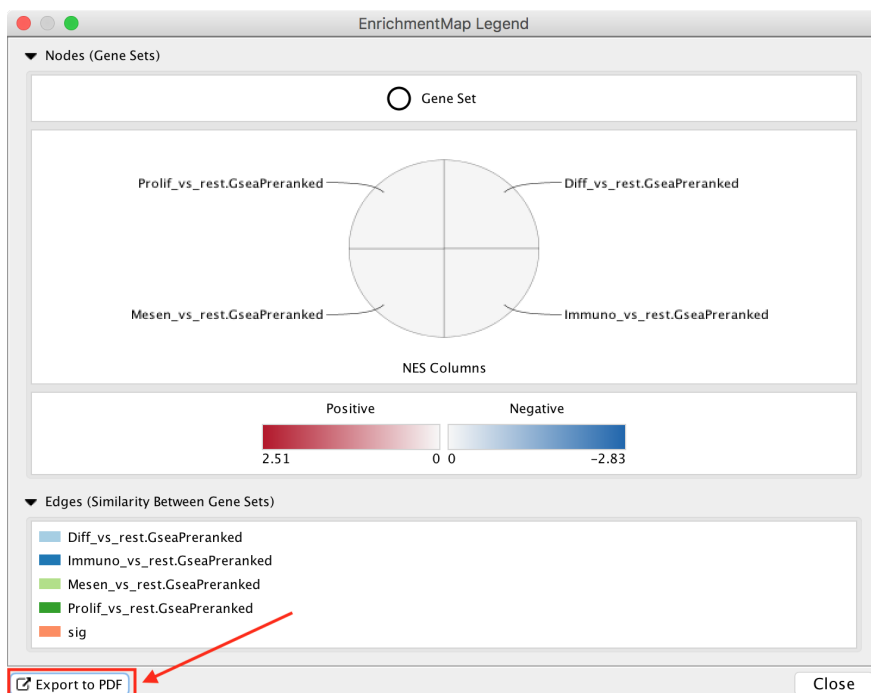
## Edge Color



- Shows how edge color maps to data sets.

## 4.9.2 Export to PDF

The *Export to PDF* button at the bottom left of the dialog can be used to create a PDF file of the legend.



### 4.9.3 Creation Parameters Dialog

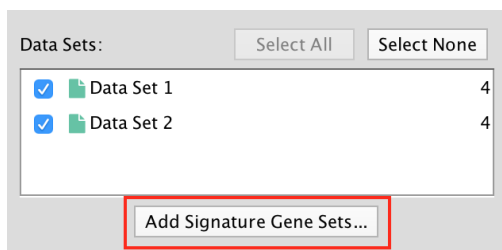
Click the **Creation Parameters** button at the bottom left of the dialog to open the Creation Parameters dialog. This dialog shows the parameters that were originally entered in the *Create EnrichmentMap Dialog* when the network was originally created.



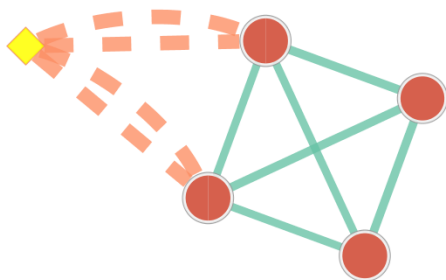
## 4.10 Post Analysis (Add Signature Gene Sets)

The **Add Signature Gene Sets** panel allows you to add more gene sets to an existing network (also called **Post Analysis**). This is done by calculating the overlap between gene sets of the current EnrichmentMap network and all the gene sets contained in the provided signature gene set file.

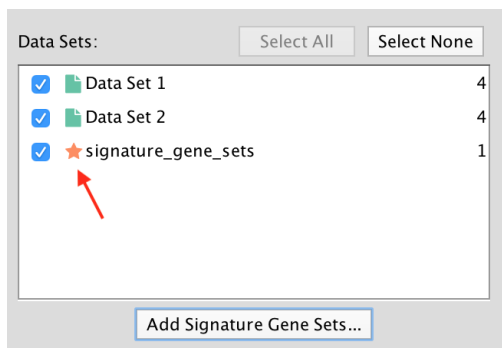
To open the *Add Signature Gene Sets Dialog* click the **Add Signature Gene Sets...** button on the main panel.



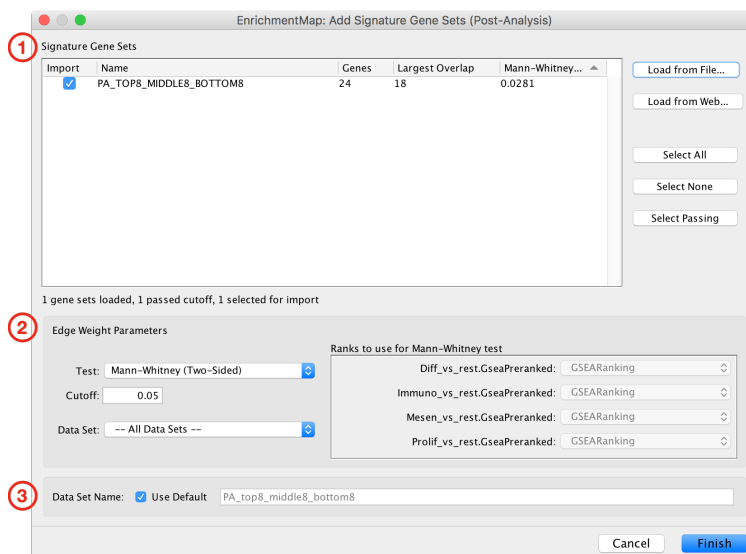
The result of running Post Analysis is a new node for each signature gene set (yellow triangle) and edges from the signature gene set to each existing gene set when the similarity passes the cutoff test.



A new data set is added to the data set list on the Main panel. Signature data sets have a ★ next to their name.



**Note:** The *Add Signature Gene Sets Dialog* has been redesigned for Cytoscape 3.1. There is no longer a distinction between *Known Signature* and *Signature Discovery*.



#### 4.10.1 Add Signature Gene Sets Dialog

The *Add Signature Gene Sets Dialog* has the following sections:

1. Signature Gene Sets List
2. Edge Weight/Cutoff Parameters
3. Data Set Name

To close the dialog click the **Cancel** button. To import signature gene sets click the **Finish** button.

#### 4.10.2 1) Signature Gene Sets List

Import	Name	Genes	Largest Overlap	Mann-Whitney (Two-Si...
<input checked="" type="checkbox"/>	PENTOSAN POLYSULFATE<DRUGBANKIDB00686	3	2	0.0312
<input checked="" type="checkbox"/>	YOHIMBIN<DRUGBANKIDB01392	12	4	0.0321
<input checked="" type="checkbox"/>	VALPROIC ACID<DRUGBANKIDB00313	7	4	0.0329
<input checked="" type="checkbox"/>	TRAZODONE<DRUGBANKIDB00656	7	3	0.0355
<input checked="" type="checkbox"/>	NISOLDIPINE<DRUGBANKIDB00401	5	5	0.0385
<input checked="" type="checkbox"/>	PRAMLINTIDE<DRUGBANKIDB01278	4	4	0.0459
<input checked="" type="checkbox"/>	HISTAMINE PHOSPHATE<DRUGBANKIDB00667	4	2	0.0484
<input type="checkbox"/>	IMIPRAMINE<DRUGBANKIDB00456	22	10	0.0514
<input type="checkbox"/>	CAFFEINE<DRUGBANKIDB00201	11	6	0.0542
<input type="checkbox"/>	TRIMETHOPRIM<DRUGBANKIDB00440	2	2	0.0548
<input type="checkbox"/>	DESIPRAMINE<DRUGBANKIDB01151	17	7	0.0556
<input type="checkbox"/>	CARIPRAZINE<DRUGBANKIDB06016	6	4	0.0564
<input type="checkbox"/>	CINCHOCANINE<DRUGBANKIDB00527	3	3	0.0566
<input type="checkbox"/>	FLUDARABINE<DRUGBANKIDB01073	3	2	0.0622

179 gene sets loaded, 19 passed cutoff, 19 selected for import

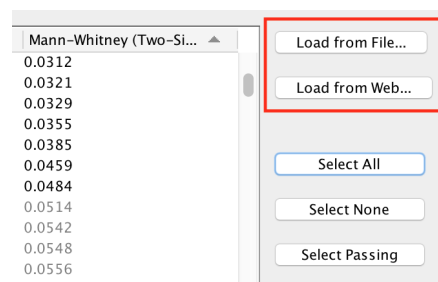
The Signature Gene Sets List allows you to preview and select the gene sets to import into the network. The *Edge Weight* test is pre-computed for each gene set.

The table has the following columns:

- Import
  - Gene sets that are selected in this column will be imported into the network. The checkboxes in this column can be used to select or de-select individual gene sets.
  - You can also use the *Select All*, *Select None*, and *Select Passing* buttons to select multiple gene sets at once.
- Name

- The name of the signature gene set.
- Genes
  - Number of genes in the signature gene set.
- Largest Overlap
  - The signature gene set is compared to every gene set in the network and the size of the largest overlap is displayed.
- Edge Weight (can be Mann-Whitney, Hypergeometric or Overlap)
  - The *Edge Weight* test is computed between the signature gene set and every gene set in the network and the largest result is displayed.
  - Can be Mann-Whitney, Hypergeometric or Overlap depending on what is currently selected in the *Edge Weight Parameters* panel below.
  - Making changes in the *Edge Weight Parameters* panel causes the values in this column to be automatically updated.

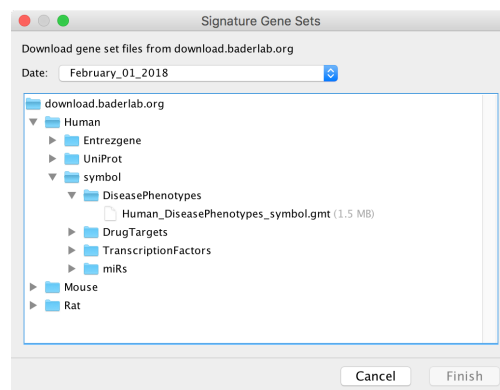
## Loading Gene Sets



The **Load from File...** and **Load from Web...** buttons can be used to load a signature gene set file.

- Load from File...
  - Opens a file browser to select a GMT file on the local file system.
- Load from Web...
  - Opens the **Load from Web** dialog. See below for more details.

## Load from Web

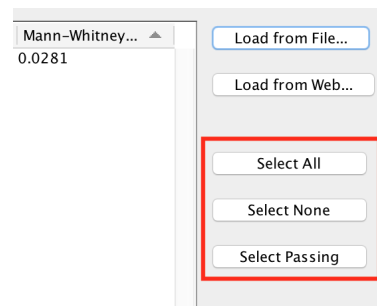


This dialog is used to select a GMT file on a remote server to download. Currently the EnrichmentMap GMT files from [download.baderlab.org](http://download.baderlab.org) are available. See [Download Gene Set Files](#) for more details on the contents of these GMT files.

Select a file and click **Finish** to start downloading. If the download is taking too long it may be canceled at any time.

### Selecting Gene Sets to Import

After loading the gene sets the ones that pass the *Edge Weight* cutoff are selected by default. Gene sets that do not pass the cutoff are de-selected and grayed out.



- Select All
  - Select all the gene sets in the list.
  - Note: Gene sets that do not pass the cutoff will not have any edges when added to the network.
- Select None
  - None of the gene sets in the list will be selected. This is useful if you want to de-select the gene sets that were selected by default and then manually select individual gene sets by clicking on the checkboxes in the *Import* column.
- Select Passing
  - Selects just the gene sets that pass the current cutoff. This is the default.

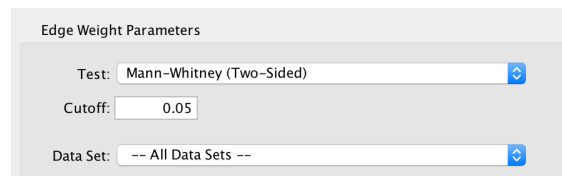
---

**Note:** Any changes made in the *Edge Weight Parameters* panel may cause the current selection to change.

---

### 4.10.3 2) Edge Weight Parameters

Any changes made to the *Edge Weight Parameters* will cause the *Signature Gene Sets List* to be automatically updated.



- Test
  - Select the type of statistical test to use for edge weight.
  - Each type of test is explained in more detail below.
- Cutoff

- Edges with a similarity score lower than the one entered will not be included in the network.
- Data Set
  - Select the data set to run post analysis against, or *All Data Sets*.
  - This field is disabled if the EnrichmentMap network contains only one data set.

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.

## Mann-Whitney

- 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.
- The Mann-Whitney test requires ranks. It will not be available if the enrichment map was created without ranks.
- If a data set has more than one ranks file then it may be selected in the adjacent panel.

**Note:** Additional ranks files can be loaded from the Heat Map [Panel Menu](#).

## Hypergeometric

- Hypergeometric Test is the probability (p-value) to find an overlap of k or more genes between a signature gene set and an enrichment gene set 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 gene set

**m** (total number of successes) : size of the Enrichment gene set

**N** (total number of elements) : size of the union of all Enrichment gene sets

- Hypergeometric Universe: Allows to choose the value for **N**.
  - GMT: All the genes in the original GMT file, before filtering.
  - 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.
- Advanced - Signature Genes To Use: Allows to choose the value for **n**
  - This panel is hidden by default, click the arrow next to *Advanced* to expand the panel.
  - Filtered Signature Gene Sets: Size of the signature gene set intersected with the genes in the data set.
  - Signature Gene Sets: Size of the signature gene set.

## Overlap

- 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.

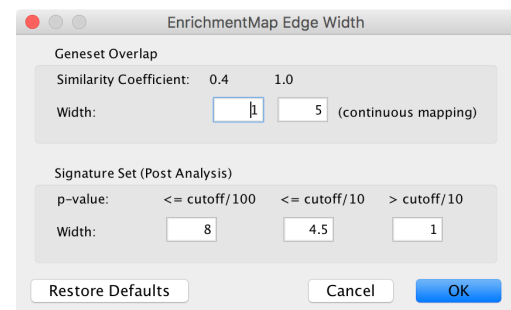
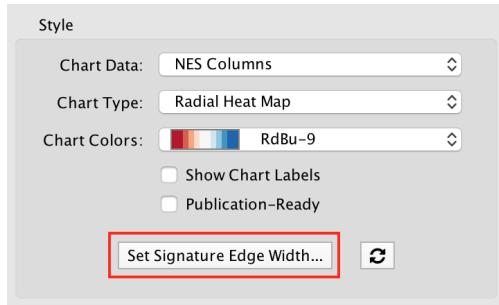
### 4.10.4 3) Data Set Name

A name for the signature data set will be automatically generated based on the name of the GMT file that was loaded. To enter a name for the data set first de-select the **Use Default** checkbox then enter the name.

### 4.10.5 Edge Width Dialog

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.

To open the dialog click the **Set Signature Edge Width..** button in the style section of the main panel.



- Edge Width Dialog
  - Gene set Overlap: Set the end points of the continuous 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,  $\leq$  cutoff/10 and  $>$  cutoff/10.
  - Click OK to recalculate the values in the “EM1\_edge\_width\_formula” column.

## 4.11 File Formats

### 4.11.1 Gene sets file (GMT file)

- Each row of the gene set file represents one gene set and consists of:

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

- The gene set names must be unique.
- The gene set file describes the gene sets 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 [Download Gene Set Files](#) 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`

---

### 4.11.2 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 Heat Map when clicking on nodes and edges in the Enrichment Map.
  - Gene sets can be 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} then 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 EnrichmentMap will create a dummy expression file to associate with the data set. The dummy expression gives an expression value of 0.25 for all the genes associated with the enriched genesets in the Enrichment map.

#### 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

## 4.11.3 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.

## BiNGO Enrichment Result File

- The BiNGO option expects a file as generated by the BiNGO Cytoscape 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.

## 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 rpt

- if xtools.gsea.GseaPreranked:

- No expression file
- use rnk as the expression file from rnk parameter in rpt
- 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_formatted_
↳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.

## 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.

### 4.11.4 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
...
```

## 4.12 Tips on Parameter Choice

### 4.12.1 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”

### 4.12.2 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**

$$\text{Jaccard Coefficient} = [\text{size of (A intersect B)}] / [\text{size of (A union B)}]$$

### Overlap Coefficient

$$\text{Overlap Coefficient} = [\text{size of (A intersect B)}] / [\text{size of (minimum( A , B))}]$$

### 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.

$$\begin{aligned} \text{Combined Constant} &= k \\ \text{Combined Coefficient} &= (k * \text{Overlap}) + ((1-k) * \text{Jaccard}) \end{aligned}$$

## 4.12.3 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:**

- $p\text{-value} < 0.05$
- $\text{FDR} < 0.25$

#### **Moderately permissive:**

- $p\text{-value} < 0.01$
- $\text{FDR} < 0.1$

#### **Moderately conservative:**

- $p\text{-value} < 0.005$
- $\text{FDR} < 0.075$

#### **Conservative:**

- $p\text{-value} < 0.001$
- $\text{FDR} < 0.05$

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:**

- $p\text{-value} < 0.05$
- $\text{FDR} < 0.25$



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

## 4.13 Download Gene Set Files

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

Gene Set Files can be downloaded from: [Baderlab genesets collections](#)

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, February 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.

---

### 4.13.1 Summary

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\_ia\_{Date}\_{ID}.gmt** Contains gene sets 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\_ia\_{Date}\_{ID}.gmt** Contains gene sets 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\_ia\_{Date}\_{ID}.gmt** Contains only gene sets 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\_ia\_{Date}\_{ID}.gmt (recommended file)** Contains only gene sets 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 gene sets from all pathways resources.

### 4.13.2 Current Stats

**Human** [http://download.baderlab.org/EM\\_Genesets/current\\_release/Human/Summary\\_Geneset\\_Counts.txt](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](http://download.baderlab.org/EM_Genesets/current_release/Mouse/Summary_Geneset_Counts.txt)

### 4.13.3 Sources

#### Human

Source	File Type	ID extracted	Frequency source is updated	Number of pathways	File Origin
KEGG	GMT	Sym- bol	static as of July 1, 2011	236	KEGG ftp site (July 2011)
MSigDB - C2	GMT	En- trez gene	sporadically	Biocarta - 217, Other - 47	manual download
NCI	BioPAX	En- trez gene	sporadically	219 pathways	scripted download of zipped release
Institute of Bioinformatics (IOB)	BioPAX	En- trez 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	En- trez 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 peri- odically	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	En- trez gene	sporadically	221 miRs, 616 TFs	manual download
Panther	BioPAX	UniProt	updated peri- odically	307 Pathways	scripted download of biopax archive

## Mouse

Source	File Type	ID ex-tracted	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	En-trez gene	static as of July 1, 2011	236	translated from Human using Homologene
MSigDB - C2	GMT	En-trez gene	sporadically	total 880: Kegg - 186, Reactome - 430, Biocarta - 217, Other - 47	translated from Human using Homologene
NCI	GMT	En-trez gene	sporadically	219 pathways	translated from Human using Homologene
Institute of Bioinformatics (IOB)	GMT	En-trez gene	sporadically	35 pathways - 10 are the same as CellMap, 1 is the same as NetPath	translated from Human using Homologene
NetPath (IOB)	GMT	En-trez gene	static	25 pathways - 12 are cancer pathways (10 are CellMap), 13 are immunity pathways	translated from Human using Homologene
HumanCyc	GMT	En-trez gene	updated periodically	249 Pathways	translated from Human using Homologene
Panther	BioPAX	UniProt	updated periodically	307 Pathways	translated from Human using Homologene

### 4.13.4 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.

### 4.13.5 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.

### 4.13.6 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
```

### 4.13.7 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, Doremioux 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](#).

## 4.14 Automating EnrichmentMap

EnrichmentMap provides several commands which allow basic features to be automated via scripts, the command line or REST.

### 4.14.1 CyREST App

To call commands via REST the CyREST App is required. CyREST is normally installed by default.

CyRest is updated often. There are two ways to install the latest version:

1. App Manager

Open the **App Manager** dialog from the main menu at **Apps > App Manager**. Then go to the **Check For Updates** tab. If there is a newer version available it will be listed.

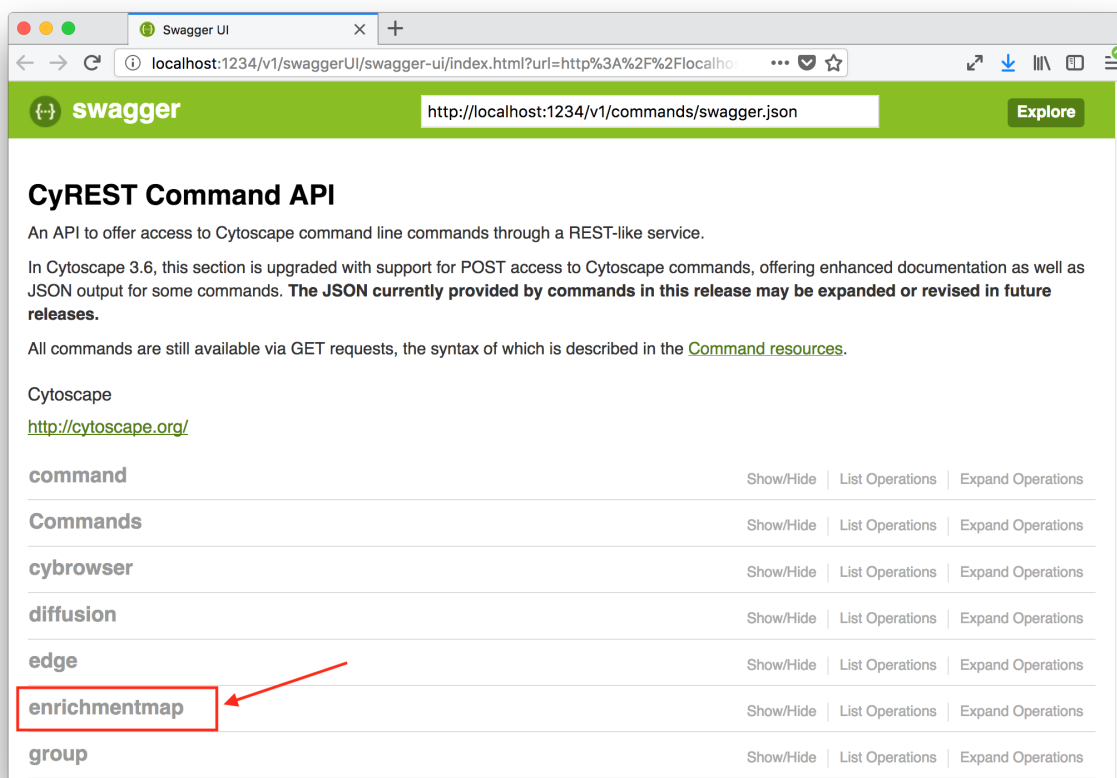
2. App Store

CyRest can also be installed or updated from the [App Store Website](#)

### 4.14.2 Command Documentation

#### CyREST Documentation

On-line documentation for EnrichmentMap commands can be accessed from the main menu at **Help > Automation > CyREST Command API**. This will open a web browser with documentation for all the commands that are available through CyREST. Navigate to the **enrichmentmap** entry and expand it for detailed documentation on each command.



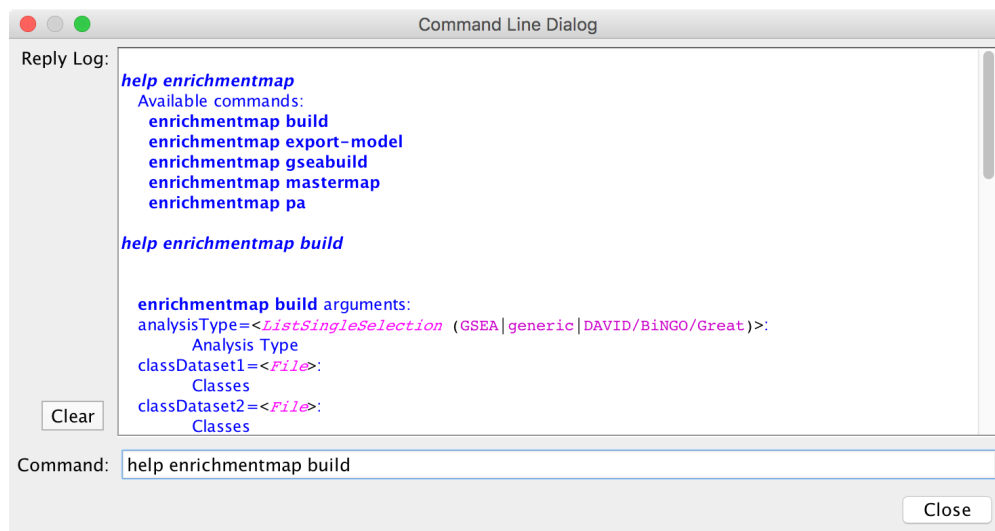
**Note:** For more details on using CyRest see the [CyREST Documentation](#)

## Command Line Dialog Documentation

Open the Command Dialog from the main menu at **Tools > Command Line Dialog**.

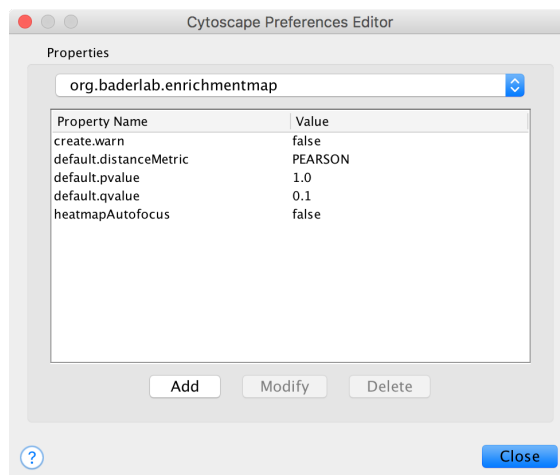
Type `help enrichmentmap` to list the available commands provided by EnrichmentMap. To get help on a particular command type, for example the `build` command, type `help enrichmentmap build`.





**Note:** For more details on the Command Line Dialog see the [Cytoscape User Docs](#)

## 4.15 Properties



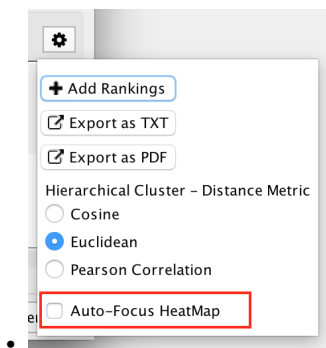
EnrichmentMap has some semi-hidden properties than can be used to customize the behavior of the App.

To manually edit these properties go to the Cytoscape main menu and select **Edit > Preferences > Properties...**. Then select **org.baderlab.enrichmentmap** in the *Cytoscape Preferences Editor*.

Supported properties:

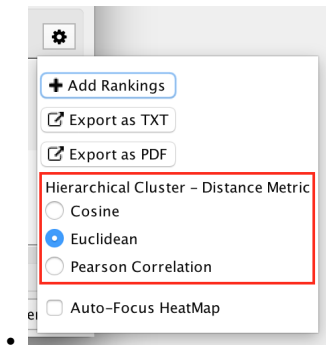
### heatmapAutofocus

- If true then selecting a node/edge in the network will automatically bring the Heat Map panel to the front.
- This property can also be changed from the Heat Map using the *Panel Menu*.
- Default Value: *false*
- Allowed Values: *true, false*



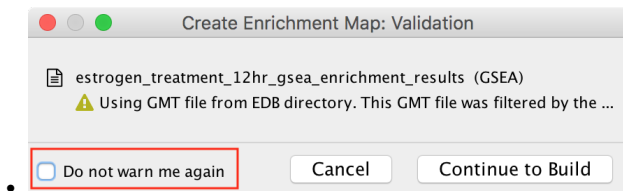
### default.distanceMetric

- Specifies the how the Hierarchical Clustering algorithm calculates distance.
- This property can also be changed from the Heat Map using the [Panel Menu](#).
- Default Value: *PEARSON*
- Allowed Values: *PEARSON*, *EUCLIDEAN*, *COSINE*



### create.warn

- When creating a network with the *Create EnrichmentMap Dialog* sometimes a validation warning dialog will be shown if there are any issues that need to be brought to the user's attention. If this property is set to *false* then the warning dialog will never be shown even if there are issues. Note: The dialog will still be shown if there are errors that prevent the network from being created.
- Default Value: *true*
- Allowed Values: *true*, *false*



### default.pvalue

- Default p-value in the *Create EnrichmentMap Dialog*
- Default Value: *1.0*
- Allowed Values: *> 0.0*, *< 1.0*

Number of Nodes (gene-set filtering)

Filter genes by expressions: ☐

FDR q-value cutoff: 0.1

p-value cutoff: 1.0

NES (GSEA only): All

Filter by minimum experiments: ☐

Minimum experiments: 3

#### default.qvalue

- Default q-value in the *Create EnrichmentMap Dialog*
- Default Value: 0.1
- Allowed Values:  $> 0.0$ ,  $< 1.0$

Number of Nodes (gene-set filtering)

Filter genes by expressions: ☐

FDR q-value cutoff: 0.1

p-value cutoff: 1.0

NES (GSEA only): All

Filter by minimum experiments: ☐

Minimum experiments: 3

## 4.16 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 be done 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 were 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).

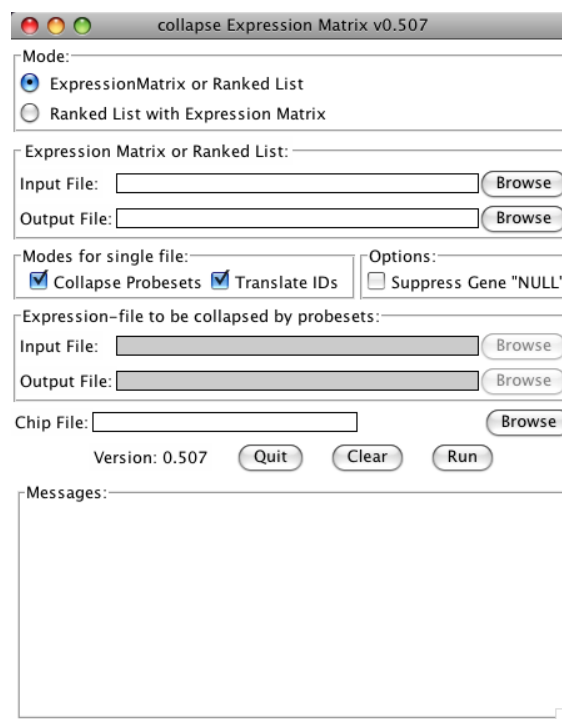
### 4.16.1 Requirements

- 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” or 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)

### 4.16.2 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

### 4.16.3 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](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:

<code>--version</code>	show program's version number and exit
<code>-h, --help</code>	show this help message and exit
<code>-i FILE, --input=FILE</code>	input expression table or ranked list
<code>-o FILE, --output=FILE</code>	output expression table or ranked list
<code>-c FILE, --chip=FILE</code>	Chip File This implies that the Identifiers are to be replaced.
<code>-e FILE, --ei=FILE</code>	(optional) additional input Expression-table, to be restricted to the same probe-sets as the RNK file
<code>-x FILE, --xo=FILE</code>	(optional) corresponding output file for <code>-i/--ei</code> option
<code>--collapse</code>	Collapse multiple probe sets for the same gene symbol (max_probe)
<code>--no-collapse</code>	Don't collapse multiple probesets [default]
<code>--null</code>	suppress Gene with Symbol NULL
<code>-g, --gui</code>	Open a Window to choose the files and options.
<code>-q, --quiet</code>	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 than 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 Control 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`