Chapter 12

Visualizing Gene-Set Enrichment Results Using the Cytoscape Plug-in Enrichment Map

Daniele Merico, Ruth Isserlin, and Gary D. Bader

Abstract

Gene-set enrichment analysis finds functionally coherent gene-sets, such as pathways, that are statistically overrepresented in a given gene list. Ideally, the number of resulting sets is smaller than the number of genes in the list, thus simplifying interpretation. However, the increasing number and redundancy of gene-sets used by many current enrichment analysis resources work against this ideal.

"Enrichment Map" is a Cytoscape plug-in that helps overcome gene-set redundancy and aids in the interpretation of enrichment results. Gene-sets are organized in a network, where each set is a node and links represent gene overlap between sets. Automated network layout groups related gene-sets into network clusters, enabling the user to quickly identify the major enriched functional themes and more easily interpret enrichment results.

Key words: Gene-set enrichment, Functional enrichment, Microarray data analysis, Gene ontology, Pathways

1. Introduction

High-throughput genomic experiments often lead to the identification of large gene lists (1). Gene lists are typically defined using statistical methods appropriate to the experimental design. For instance, a frequently applied method is to score genes by their differential expression between two biological states (such as healthy vs. disease). However, these methods for finding interesting genes often do not help the interpretation of the resulting gene lists.

Enrichment analysis is an automated and statistically rigorous technique to analyze and interpret large gene lists using a priori knowledge (2–4). Enrichment analysis assesses the over (or under-) representation of a known set of genes (e.g., a biological pathway)

Gerard Cagney and Andrew Emili (eds.), *Network Biology: Methods and Applications*, Methods in Molecular Biology, vol. 781,
DOI 10.1007/978-1-61779-276-2_12,

within the input gene list. If a statistically significant number of genes from the known set are present in the gene list, it may indicate that the biological pathway plays a role in the biological condition under study. This analysis is repeated for all available known gene-sets, which could number in the thousands.

The growing number of available gene-sets, due to the increased availability of functional annotations, makes enrichment analysis a powerful tool to help researchers gain interesting insights from their high-throughput data. However, this comes at a cost: as gene-set collections get larger and more complex, there may be longer lists of results and increased redundancy between sets. Redundancy is particularly problematic with gene-sets derived from hierarchical functional annotation systems, such as Gene Ontology (GO), as children terms are partially redundant with their parents by definition. Gene-set redundancy constitutes a major barrier for the interpretability of enrichment results, limiting the full exploitation of its analytic power.

Enrichment Map is a visualization tool that organizes gene-sets into a similarity network, where nodes represent gene-sets, links represent the overlap of member genes, and node color encodes the enrichment score (5). Nodes are automatically arranged so that highly similar gene-sets are placed close together; these "clusters" can be easily identified manually and related to biological functions. Enrichment Map is implemented as a freely available and open-source plug-in for the Cytoscape network visualization and analysis software (6, 7).

In the methods section, we clarify fundamental concepts of enrichment analysis and describe how to use Enrichment Map for enrichment visualization. *Protocol 1* describes the basic visualization of enrichment results. *Protocol 2* shows how to visualize and compare two different enrichment results. *Protocol 3* describes how to analyze the relationships between enriched gene-sets and an additional query gene-set.

2. Materials

2.1. Software Installation

Install Cytoscape, the Enrichment Map and the WordCloud plug-ins according to the following instructions. Cytoscape 2.7.0, Enrichment Map 1.0, and WordCloud 1.0 were used for the protocols; later versions are supposed to work as well, but minor changes in the instructions may be required. Please refer to online manuals (http://baderlab.org/Software/EnrichmentMap/UserManual, http://baderlab.org/Software/WordCloudPlugin/UserManual) in case of discrepancies.

2.1.1. Cytoscape Installation

Before proceeding to install Cytoscape, check what version of Java is installed.

Windows (XP or Vista):

1. From *Start button/All programs/Accessories* run: *Command Prompt*
2. Type: java -version

Mac (Mac OS X):

1. From *Applications/Utilities* run: *Terminal*
2. Type: java -version

If a version older than 5.0 is found, please install Java SE 5 (http://www.oracle.com/technetwork/java/javase/downloads/index-jdk5-jsp-142662.html), Java SE 6 (http://www.oracle.com/technetwork/java/javase/downloads/index.html) or later.

To install Cytoscape, download the appropriate installer package for your operating system (Windows, Mac, or Linux) from: http://www.cytoscape.org/download.php.

Make sure the Cytoscape icon is added to the *Start Button* program list (Windows) or to the *Applications* folder (Mac).

2.1.2. Plug-in Installation

To install the Enrichment Map plug-in, first download the plug-in package at: http://baderlab.org/Software/EnrichmentMap#Plugin_Download.

The plug-in file (.jar) will have to be moved into the Cytoscape/plug-in directory. To locate the Cytoscape directory, follow these instructions.

- Windows (XP or Vista): right-click on the Cytoscape icon and select *Properties*.
- Mac (Mac OS X): CTRL-click on the Cytoscape icon and select *Show in Finder*.

To install the WordCloud plug-in, follow the same instructions as above after downloading the .jar from: http://baderlab.org/Software/WordCloudPlugin.

2.2. Sample Data Download

The analysis presented in the protocols can be reproduced downloading the sample data at: http://baderlab.org/Software/EnrichmentMap#Sample_Data_Download.

2.3. Data Formats

Enrichment results can be in two formats:

- The one adopted by *GSEA* (Gene Set Enrichment Analysis)
- The *Generic* format, which is suitable when enrichment results are generated using other software tools.

Table 1
Summary of enrichment map input files

File type	File extension	Separator	Header	# Columns	Requirement
Gene-set	.GMT	Tab	No	Variable	Mandatory
Expression matrix	.txt	Tab	Yes	Constant (any ≥ 3)	Mandatory
Expression matrix	.gct	Tab	Yes	Constant (any ≥ 3)	Mandatory
Enrichment table (GSEA)	.xls	Tab	Yes	Fixed (11)	Mandatory
Enrichment table (Generic)	.txt	Tab	Yes	Fixed (5)	Mandatory
Rank	.RNK	Tab	No	Fixed (2)	Optional
Query gene-set	.GMT	Tab	No	Variable	Optional

The number of columns has been categorized as *Fixed (N)* when every row must have exactly the number of elements in brackets *(N)*, *Constant* when every row must have the same number of elements but this can be different for different files, *Variable* when every row can have a different length

The gene-set file and expression matrix formats are the same for the GSEA and Generic option, only the enrichment table format differs. Data formats are summarized in Table 1.

The Subheading 2.3.8. *Formatting DAVID Data for Enrichment Map* explains step-by-step how to load data from DAVID, a broadly used and cited enrichment tool for *subset gene lists*.

2.3.1. Gene-Set File (.GMT)

The gene-set file contains the gene-set identifiers (IDs), the gene-set names and the member gene IDs. The gene-set .GMT file is tab-separated; each line can have a different length, but must follow the syntax:

<gene-set id><tab><gene-set name><tab><gene-id><tab> <gene-id><…>

No header should be present.

Gene-sets can be derived from gene annotation databases, pathway databases and other gene-centered resources. Gene-set files following the GMT format and compiled from public annotation databases can be downloaded from Pathway Commons (http://www.pathwaycommons.org/pc-snapshot/current-release/gsea/), WhichGenes (http://www.whichgenes.org/), MSigDB (http://www.broadinstitute.org/gsea/downloads.jsp) or from the Enrichment Map download page (http://baderlab.org/Software/EnrichmentMap Gene-sets_for_Enrichment_Analysis).

When downloading or generating the gene-set file, it is important to check ID consistency:

(a) Gene IDs must be consistent in the gene-set file and in the expression matrix.

(b) Gene-set IDs must be consistent in the gene-set file and in the enrichment table(s).

Gene-sets should be filtered by size, removing the ones larger or smaller than user-selected thresholds. In particular:

- Large gene-sets, especially when derived from Gene Ontology (GO), have little use when interpreting gene lists because they are often very general. A reasonable maximum size for human or mouse is between 400 and 900 genes. In addition, the accurate exploration of the map can reveal gene-sets connecting very different biological functions (e.g., *GO Protein complex assembly*); these should be removed on the user's discretion, and the layout should be recomputed.
- Small gene-sets (e.g., with less than 5–10 genes) may be more sensitive to small fluctuations in the query gene list, hence their presence may cause fewer gene-sets to be significant according to the FDR.

See Notes 3–4 for additional formatting tips.

2.3.2. Expression Matrix (.txt)

The .txt expression matrix file contains the gene IDs, the gene names (or any other textual annotation) and gene expression signals, or any other quantitative value characterizing genes (e.g., spectral counts, number of mutations). The .txt expression file is tab-separated; every line must have equal length following the syntax:

<gene-id> <tab> <gene-name> <tab> <value> <tab> <value> <…>

The header must be present at the first line. The labels of the first and second columns must be set to "*NAME*" and "*DESCRIPTION*," whereas the value columns can have any label, but repetitions must be avoided.

2.3.3. Expression Matrix (.gct)

The .gct expression matrix file must have two initial lines with this content:

#1.2

<number-of-genes> <tab> <number-of-samples>

followed by the same content as the .txt file (including the header at the first line).

Specifically, <*number-of-genes*> consists of the number of rows having a gene ID, name, and values; <*number-of-samples*> consists of the number of columns having an associated <*value*> (*NAME* and *DESCRIPTION* should not be counted).

2.3.4. Enrichment Table: GSEA (.xls)

GSEA follows the two-condition enrichment logic described in the methods section. Two separate files are generated for enrichment in condition A and B.

The file names have the prefix "*gsea_report_for_*" and extension *.xls* (although they are tab-separated files and not Excel files). Each line must have the same length, and the first line must be the header, with the following labels: "*NAME*", "*GS
follow link to MSigDB*", "*GS DETAILS*", "*SIZE*", "*ES*", "*NES*", "*NOM p-val*", "*FDR q-val*", "*FWER p-val*", "*RANK AT MAX*", and "*LEADING EDGE*". Every column must be present, but only "*NAME*" (the gene-set ID) and "*NES*", "*NOM p-val*", "*FDR q-val*" (gene-set enrichment statistics) are utilized by Enrichment Map.

These files are automatically generated by GSEA and they follow the syntax described above (as of October 2010).

2.3.5. Enrichment Table: Generic (.txt)

The Generic format enrichment table is a tab-separated file (.txt). Each line must have the same length and the first line must be the header. The file must have 5 columns:

1. Gene-set ID
2. Gene-set name
3. Enrichment *p*-value
4. Enrichment FDR (false discovery rate)
5. Phenotype

The header labels can be freely defined by the user, but the columns must be kept in the order defined above. If data are not available for any of the columns, the column should not be skipped, as the file needs to have exactly 5 columns. Any conventional value representing missing data (e.g., "NA") should be used instead.

For one-condition enrichment, all values in the phenotype column must be set to 1. For two-condition enrichment, phenotype values can be set to 1 or –1 to indicate the enrichment in either condition (see Note 5 for a detailed explanation of the numbering conventions).

2.3.6. Rank File (.rnk)

This file is optional. It is tab-separated, without a header, and each line must have the same length. The first column must be gene ID and the second column must be gene scores (e.g., *t statistic* or *log ratio*).

2.3.7. Query Set Gene-Set File (.GMT)

Same format as the main gene-set file.

2.3.8. Formatting DAVID Data for Enrichment Map

DAVID is a Web-based enrichment tool that accepts *subset gene lists* as input. Follow these instructions to import DAVID results using the *Generic* option. An additional format option dedicated to DAVID is currently under development, and it will be available in future releases of the Enrichment Map plug-in.

1. Obtain the gene-sets from DAVID. This currently requires registering at http://david.abcc.ncifcrf.gov/knowledgebase/register.htm. The DAVID Forum (http://david.abcc.ncifcrf.gov/forum) can be consulted for updated information on this topic, using "Download knowledgebase" as search keyword.
2. Format DAVID gene-sets as GMT. This has to be done programmatically. This R script can be easily used by readers who are not confident with scripting: http://baderlab.org/Software/EnrichmentMap/DAVIDgs_R.
3. Generate enrichment results using DAVID and select *Functional Annotation Chart* as output type.
4. Export DAVID results as tab-separated text file by following the *Download File* link of the *Functional Annotation Chart* page. Save the file from the Web browser as a plain text file.
5. Format DAVID enrichment results (tab-separated text) following the *Generic* enrichment table format. This can be done programmatically or using Excel:
   (a) Column 1: label "ID," copy and paste from "Term" column.
   (b) Column 2: label "Name," copy and paste from "Term" column.
   (c) Column 3: copy and paste from "Pvalue" column, original label can be kept.
   (d) Column 4: label "FDR," copy and paste from "Benjamini" column.
   (e) Column 5: label "Phenotype" and add phenotype values accordingly (see note below).

Note: for *one-condition* enrichment, you will generate only one enrichment table from DAVID; therefore, all phenotype values will have to be set to 1. For *two-condition* enrichment, you will generate two enrichment tables from DAVID, each associated to a specific condition and to a different *subset gene list*; phenotype values will have to be set to +1 for one table and −1 for the other table; then the two tables will have to be merged. If any gene-set appears twice with two different phenotype values, enrichment map will neglect the phenotype with higher *p*-value.

3. Methods

3.1. Types of Enrichment Analysis

Before describing how to use Enrichment Map, it is important to clarify a few concepts of enrichment analysis.

Enrichment analysis typically requires two inputs: the gene-sets based on a-priori knowledge (*known gene-sets*) and the gene list to be functionally characterized using the known gene-sets

(*query gene list*). Known gene-sets usually consist of functional annotations based on controlled vocabularies (e.g., *Gene Ontology*), curated pathways (e.g., *KEGG*), protein families (e.g., proteins carrying the same *PFAM* domain) and other gene-sets based on published experimental data (e.g., tissue-specific gene expression, physical interactors of given disease genes, predicted targets of transcription factors or regulatory RNAs). The type of gene-set used for enrichment analysis should be chosen based on the biological question of interest; for instance, functional annotations and pathways are particularly appropriate to elucidate which biological processes are modulated in relation to cell state changes. The *Gene-set File* Subheading 2 provides more details on how to obtain gene-sets from public resources.

The query gene list is often derived from a high-throughput experiment (e.g., a microarray gene expression study), but it can also be derived from a meta-analysis of the literature or any other source. The query gene list can be a simple set of genes, without any associated quantitative attribute, in which case it will always be a *subset* of a larger gene-set, usually called *universe-set*, composed of all genes in the genome or all genes that could be possibly detected using a given experimental method. The definition of the *universe-set* is highly dependent on how the *subset gene list* was generated. If genes were quantitatively scored for some biological property, the subset gene list can be generated by selecting a score threshold; this is common for microarray analysis, when genes are scored for differentiality following a two-class design (e.g., estrogen-treated vs. untreated cell lines). In these cases, it is also possible to use the scores for all genes (i.e., a *scored gene list*) as the input of enrichment analysis. Unlike the *subset* gene list, all genes of the universe-set will be present in the *scored* gene list. Figure 1 summarizes the relations between query and known gene-sets.

Each known gene-set is tested for enrichment in the query list genes. Different statistical tests are available and have different input requirements. For instance, the broadly used *Fisher's Exact Test* (often referred to as *hypergeometric distribution test*) determines the significance of the overlap between a *subset* gene list and a known gene-set. *GSEA (Gene-Set Enrichment Analysis)* is a popular method, developed by the Broad Institute, that works on *scored* gene lists by testing if known gene-sets are enriched in top-scoring genes from the query list (4).

Depending on the experimental design underlying gene scoring or selection, enrichment analysis can be *one-condition*, *two-condition*, or *multi-condition*. The term "condition" here is synonymous with "class" or "phenotype" in the way these terms are used in enrichment analysis literature and software tools.

In *one-condition* enrichment, the query list genes are associated to a single biological condition. For instance, a ChIP-chip or

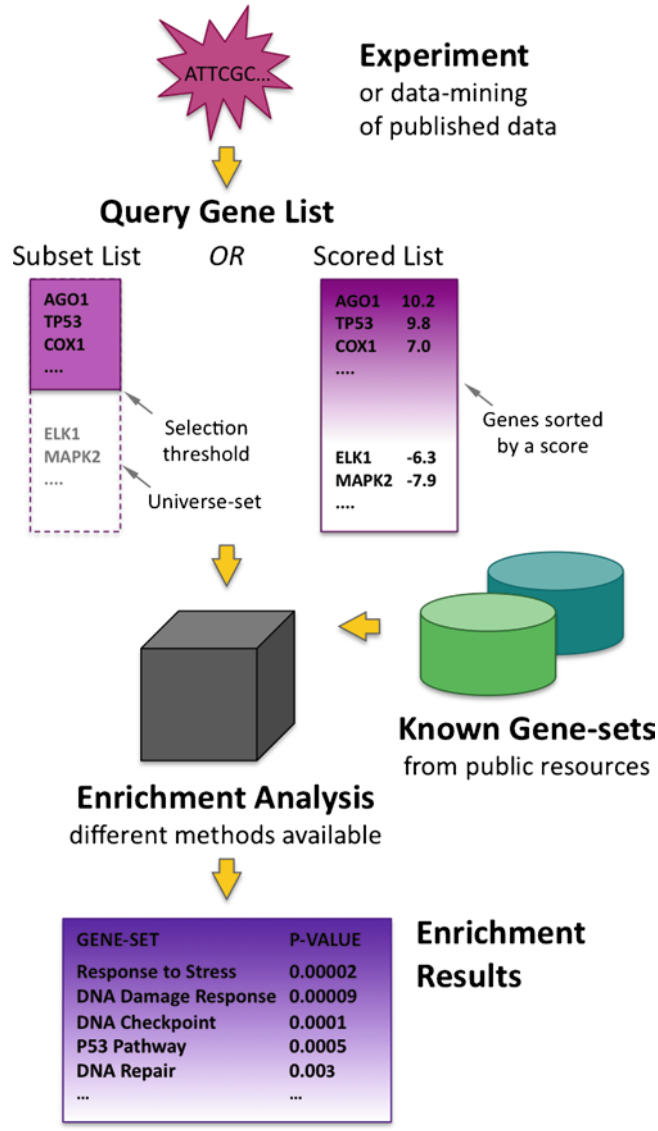

Fig. 1. The information flow from a high-throughput experiment to enrichment results. A high-throughput experiment, or data-mining of published data, typically produces a *query gene list* (of *subset* or *scored* type), which is tested for enrichment in *known gene-sets* utilizing a specific enrichment analysis method.

ChIP-seq experiment can be used to identify which genes are bound by the transcription factor *NF-Y*. Depending on the experimental platform used and the statistical model used to analyze the results, it may be more convenient to select a subset of genes that have a reliable signal for NF-Y binding (*subset gene list*), or to quantitatively score all genes for NF-Y binding (*scored gene list*). The query gene list will then be tested for enrichment in known gene-sets. Assuming that only NF-Y bound genes should be functionally characterized, the enrichment will be *one-condition*.

In *two-condition* enrichment, the query list genes are selected or scored by comparing *two* biological conditions; healthy vs. diseased and treated vs. untreated designs are common examples. Each gene can be more strongly associated to either of the two conditions, thus gene-sets are typically tested for enrichment in condition A or in condition B. For example, in gene expression studies, if condition *A* corresponds to *estrogen-treated cells* and condition *B* corresponds to *not-treated cells*, then the enrichment in condition A and B will be typically referred as enrichment in *upregulation* and *downregulation* after estrogen treatment, respectively.

In *multi-condition* enrichment, gene-sets can be enriched in multiple biological conditions. For instance, a time course experiment can be analyzed using clustering analysis. Each cluster can then be queried for enrichment in known gene-sets. In the end, each known gene-set can be enriched in no cluster, one or more clusters.

Enrichment Map currently supports one-condition and two-condition enrichment results, whereas multi-condition enrichment is an area of future development. The protocols presented in this chapter specifically focus on two-condition enrichment, which is the most common. Readers interested in one-condition enrichment would find instructions in Subheading 2 to appropriately structure their input files. Example enrichment results used in the protocols were derived from microarray gene expression data, as this is a common case; nonetheless, the methods presented here can be applied to any type of genomic data.

3.2. Files Required by Enrichment Map

Enrichment Map loads the following files:

1. The *gene-set file (.GMT)* defines available gene sets as a list of gene-set IDs, names or descriptions and member genes. Depending on the enrichment tool used, this file may be automatically generated by the enrichment tool by querying existing databases, or the user may be in charge of providing it. The Subheading 2 includes instruction to generate or retrieve this type of file.
2. The *expression matrix (.txt or .gct)* provides the expression values, or any other quantitative value at the gene level, that were used to score or select genes for gene-set enrichment. Within Enrichment Map, the expression matrix data are visualized using a heat map. These data should be directly provided by the user.
3. The *enrichment tables* consist of the gene-set enrichment results, i.e., the gene-set identifiers and enrichment statistics. These files are generated by the enrichment tool utilized for the analysis. Depending on the type of enrichment tool used, enrichment results can be loaded in the *GSEA* or *Generic*

format. Since *GSEA* is a broadly used tool with very clearly defined input and output files, its data format is directly supported by Enrichment Map. The *Generic* format was designed to be flexible enough to support any type of enrichment method with minimal user effort. The Subheading 2 includes a data reformatting protocol to load enrichment results from *DAVID* (2, 3), another popular enrichment tool developed by the NIAID/NIH.

4. The *gene rank file (.rnk)* is optional and consists of the scored gene list used for the enrichment analysis. It can be used to sort the expression matrix heat map. This file can be generated by the user, or by the enrichment tool.

The specifics of the data formats and the links to download the example enrichment data are in Subheading 2.

3.3. The Gene-Set Overlap Network

Enrichment Map arranges enriched gene-sets as a weighted similarity network. Nodes represent gene-sets passing a user-selected threshold of enrichment significance. Weighted links (i.e., edges) between the nodes represent an "overlap" score depending on the number of genes two gene-sets share. Nodes are automatically arranged so that highly similar gene-sets are placed close together; these clusters can be easily identified manually and related to biological functions. Gene-set enrichment results are graphically mapped to the Enrichment Map: node size represents the number of genes in the gene-set (see Note 1); edge thickness is proportional to the overlap between gene-sets, calculated using the Jaccard or overlap coefficients (see Protocol 1). The enrichment score (specifically, the enrichment *p*-value) is mapped to the node color as a color gradient. For one-condition enrichment results, node color ranges from white (no enrichment) to red (high enrichment). For two-condition enrichment results, node color ranges from red (high enrichment in the first condition, e.g., case) to white (no enrichment) to blue (high enrichment in the second condition, e.g., control). Figure 2 summarizes the Enrichment Map visualization of one-condition and two-condition enrichments.

3.4. Protocol 1: Analysis of One Enrichment

This protocol describes how to load results from a single enrichment analysis (one-condition or two-condition). Enrichment results of the transcriptional response to estrogen stimulation are used as an example in this protocol and in the following one. Microarray data profiling the response to estrogen treatment of MCF7 breast cancer cell lines were downloaded from Gene Expression Omnibus (GSE11352); genes were scored using the t statistic, comparing estrogen-treated vs. untreated cells; enrichment results were generated using GSEA. The example data in this protocol refer to the comparison of treated vs. untreated cells at 24 h of culture, thus a *two-condition* enrichment.

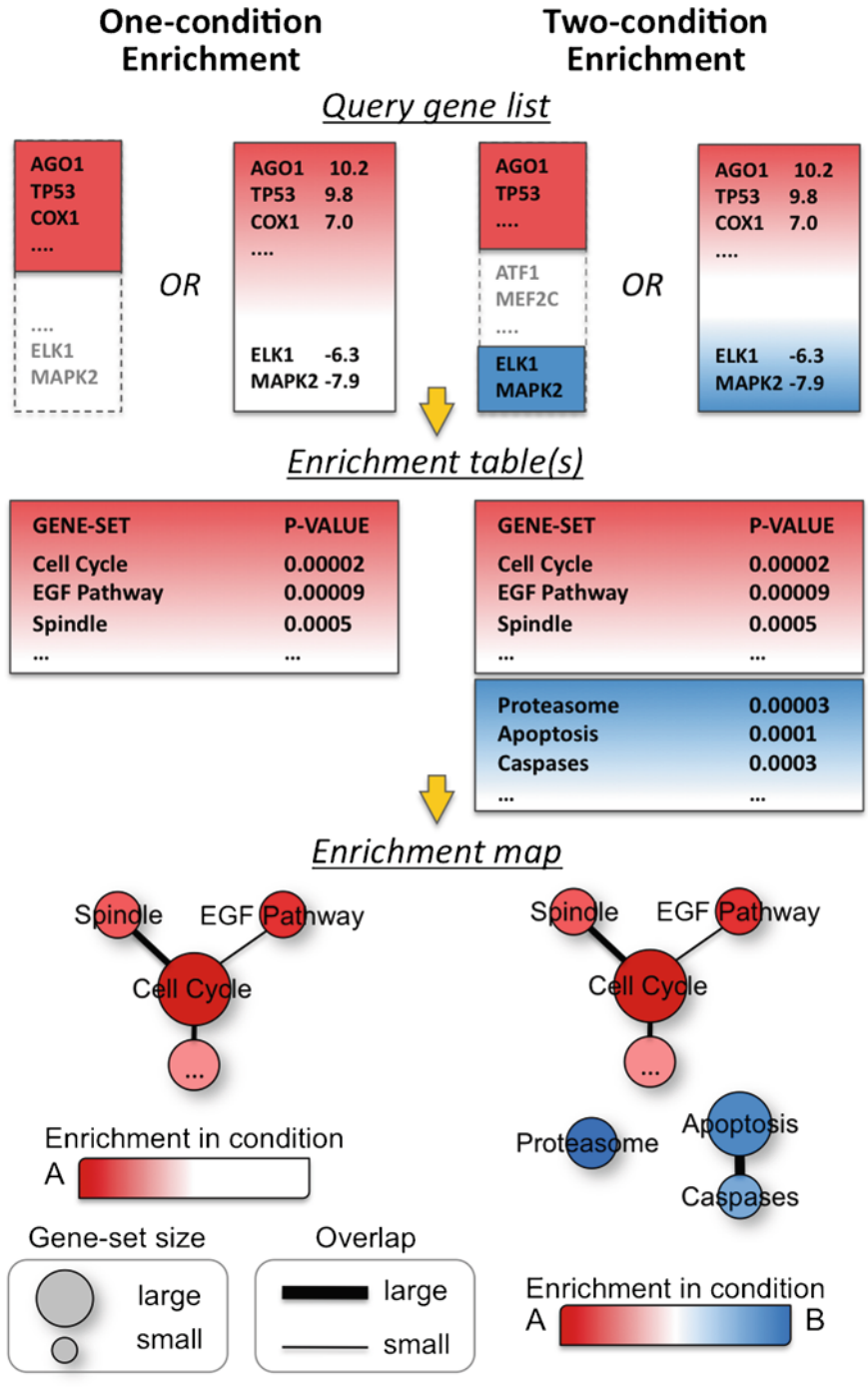

Fig. 2. One-condition and two-condition enrichment can be visualized using Enrichment Map. In one-condition enrichment, enriched gene-sets are associated to a single biological condition. In two-condition enrichment, enriched gene-sets are associated to one of two conditions. Enrichment Map arranges enriched gene-sets as a similarity network, where nodes correspond to gene-sets and links correspond to overlap of member genes.

3.4.1. Load Data and Set Parameters

1. Open Cytoscape.
2. From the top menu, select *Plugins/Enrichment Map/Load Enrichment Results*.
3. Select the format of enrichment analysis results, *Analysis Type*: *GSEA* or *Generic*. Example data: select *GSEA*.
4. If the selected analysis type is *GSEA* and the enrichment results generated by GSEA were not moved from the original directory, the *report file* (*.rpt*) can be used to automatically load all the other required files; load the report file in any of the *Dataset 1* boxes. Otherwise, follow instructions 5–8.
5. Load the *gene-set file (.GMT)*. Example data: load *GO_Hs_EG_f_hgu133p2_v2.gmt*.
6. Load the *expression matrix (.txt or .gct)*. Example data: load *MCF7_ExprMx_v2.txt*.
7. Load the *enrichment tables*. If you have selected the *GSEA* analysis type, you will have to load two files, each associated to a condition (*Enrichments 1*, *Enrichments 2* boxes); if you have otherwise selected the *Generic* analysis type, you will have to select only one file. Example data: load *EnrTable_24h_E2_v2.xls* and *EnrTable_24h_NT_v2.xls*.
8. If available, load rank data. Click on *Advanced* to display the loading text box and button. Example data: load *tTest_24h.rnk* and set *Phenotypes* to *ES_24h* and *NT_24h*.
9. Set the *p*-value and FDR (False Discovery Rate) parameters. The *P-value Cutoff* and the *FDR Q-value Cutoff* can be used to control the stringency of the analysis: only gene-sets with enrichment statistics satisfying these thresholds will be displayed by Enrichment Map. Example data: set the *p*-value cutoff to 0.001 and FDR cutoff to 0.05 (more stringent than defaults).
10. Select the similarity coefficient and its cutoff. Select the *Jaccard Coefficient* only if the gene-sets have comparable sizes (e.g., 200–300 genes). If Gene Ontology derived sets are present, select the *Overlap Coefficient*. The similarity coefficient cutoff should be tuned to optimize network connectivity. If many gene-sets are disconnected, try decreasing the cutoff value. If biologically unrelated gene-sets are connected, or if the network is close to being fully connected, increase the cutoff value. Example data: select the *Overlap Coefficient*, keep the default cutoff value of 0.5.
11. Generate the enrichment map by clicking on the *Build* button. A view of the network will be generated.

3.4.2. Globally Explore the Network

1. To explore different areas of the network, move to the *Network* tab in the Cytoscape *Control Panel* (left) and move the blue selection area in the network minimap. Use zoom icons on the Cytoscape icon bar (top) to zoom in and zoom out.
2. Automatic layout can be recalculated from the menu: *Layout/ Cytoscape Layouts/Force Directed Layout/EM1_similarity_coefficient.*
3. Node color mapping and other graphical features can be tuned using the *VizMapper* tab in the Cytoscape *Control Panel* (left). For instance, the nodes can be colored according to the *GSEA NES score* (normalized enrichment score) instead of the *p*-value. From the *Visual Mapping Browser*, select *Node Color* and change its value from *EM1_Colouring_dataset1* to *EM1_NES_ dataset1*; click on the color gradient, then click on the *Min/ Max* button in the dialog box; set the NES min to −3 and the NES max to 3 (see Note 2 for details); finally, arrange the color markers in the dialogue box (top triangles) to obtain the preferred color mapping intensity. This is useful to better distinguish the differences in enrichment strength.
4. The slider bars in the *Results Panel* (right) can be used to interactively visualize how different *p*-value and FDR cutoff values affect the network.

3.4.3. Explore Specific Gene-Sets

1. By clicking on a gene-set, the *Data Panel* will display a heat map that visualizes gene expression signals (or any other quantitative value in the expression matrix file) on a magenta-green gradient. If signals have not been already normalized by-row, i.e., if different rows can have signals at different orders of magnitude, change *Normalization* from *Data As Is* to *Row Normalize Data*. This normalization works best with gene expression from one-dye arrays (e.g., Affymetrix), whereas for absolute count data (e.g., proteomics) *Log Transform Data* may work better. The heat map can be sorted by hierarchical clustering, using rank files or following the order of a single column. The first two options are available in the *Sorting* dropdown menu. To sort according to a column, just click on a column header in the heat map.
2. Gene-set attributes can be visualized by selecting the *Node Attribute Browser* tab (*Data Panel*, bottom); if certain attributes do not display, click on the first icon from the left of the *Data Panel* and modify the attribute selection.
3. By default, clicking on a node will use the Data Panel to display the heat map. If you want to change this setting, modify the *Advanced Preferences* in the *Results Panel* (bottom). If you do so, you will have to select the *EM Geneset Expression viewer* tab (*Data Panel*) to display the heat map.

3.4.4. Explore Gene-Set Clusters

1. Gene-set clusters can be usually spotted by eye throughout the network. They consist of highly overlapping gene-sets with slightly different biological meanings. A cluster typically corresponds to a biological function or "theme."
2. To investigate the biological meaning of a cluster, select all nodes and look through their name using the *Node Attribute Browser* as explained before.
3. The *WordCloud* Cytoscape plug-in can also be used to produce summaries of the gene-set names. Select the nodes of interest (their color will turn to yellow), then select from the menu *Plugins/WordCloud/Create Cloud*. The word cloud will be visualized in the *Data Panel*. In the *Cloud Parameters/ Attribute Choice* frame, click on the *Edit* button to set the *Current Values* to *EM1_GS_DESCR* and then click on the *update* button in the bottom; in this way, a summary of gene-set names will be displayed.

Figure 3 displays the enrichment map for estrogen treatment at 24 h, and shows how the WordCloud plug-in can help summarizing the content of a gene-set cluster.

3.5. Protocol 2: Analysis of Two Enrichments

Enrichment results of the transcriptional response to estrogen stimulation are used as an example in this protocol. Microarray data were obtained and analyzed as described previously. In this protocol, two enrichments are loaded in the same map to compare the estrogen response at different time points (at 12 and 24 h of culture). Note that each enrichment follows a two-condition logic, as it results from the comparison of treated vs. untreated cells at a specific time point.

3.5.1. Load Data and Set Parameters

1. Follow the instructions in Protocol 1. In addition, load the enrichment tables for *Dataset 2* (click on the frame name or the triangle to display the boxes and buttons). Example data: load *EnrTable_12h_E2_v2.xls* and *EnrTable_12h_NT_v2.xls* as *Dataset 1*, whereas *EnrTable_24h_E2_v2.xls* and *EnrTable_24h_ NT_v2.xls* as *Dataset 2*.
2. The node center will be colored according to *Dataset 1* and the node border according to *Dataset 2*.

3.5.2. Interpretation (Example Data)

1. The majority of nodes have the same color, meaning that the estrogen response is similar at 12 and 24 h. Several nodes have more intense border colors, suggesting they are more induced at 24 h, whereas a few have more intense center color, suggesting they are more induced at 12 h.

Figure 4 shows the 12–24 h estrogen response enrichment map after some manual node repositioning to improve clarity;

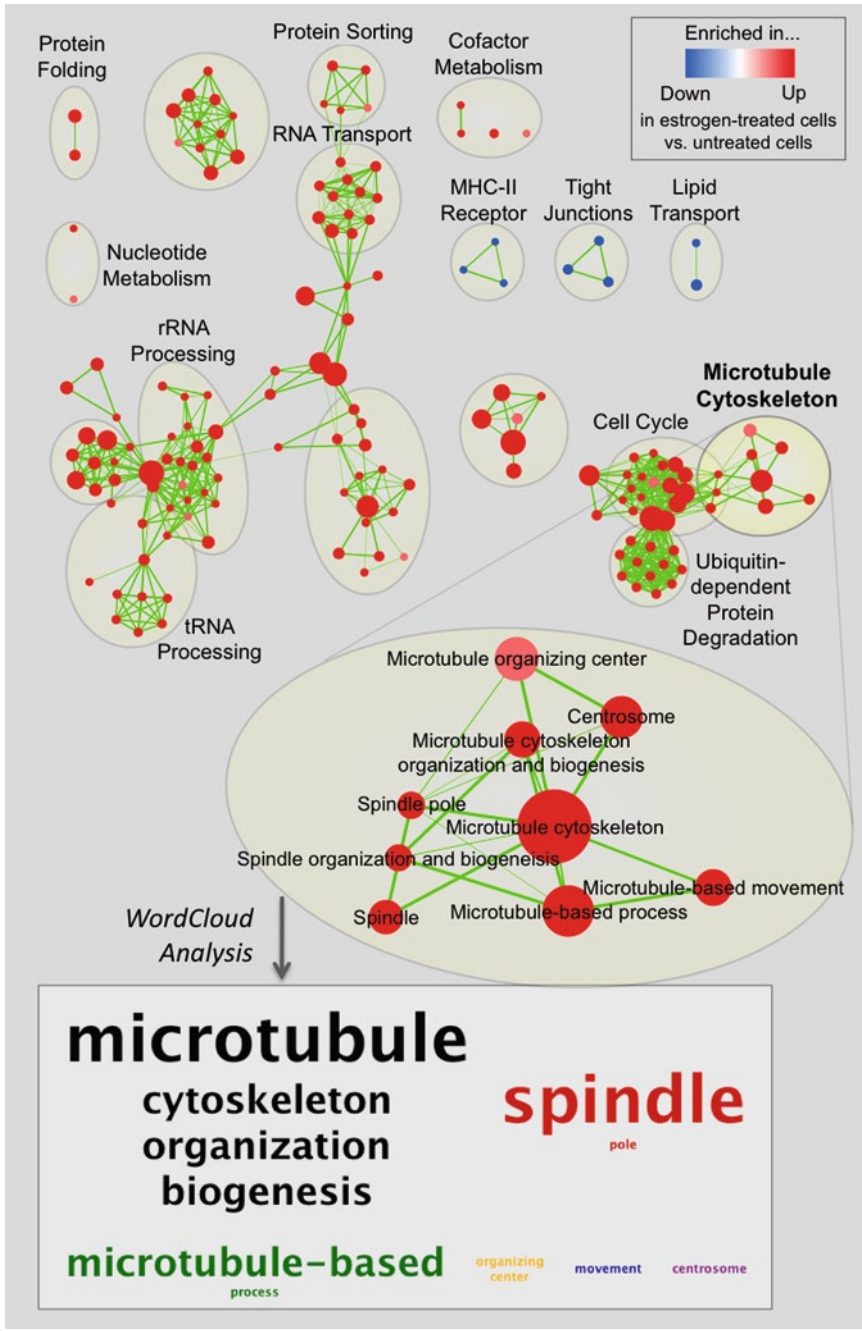

Fig. 3. Enrichment Map of the 24-h estrogen response. The network was manually rearranged to improve layout, and major clusters were manually labeled. The WordCloud plug-in can be used to automatically process gene-set names and help with cluster labeling. This is exemplified for the *Microtubule Cytoskeleton* cluster; the clustered word cloud summarizing its gene-set names is displayed in the bottom.

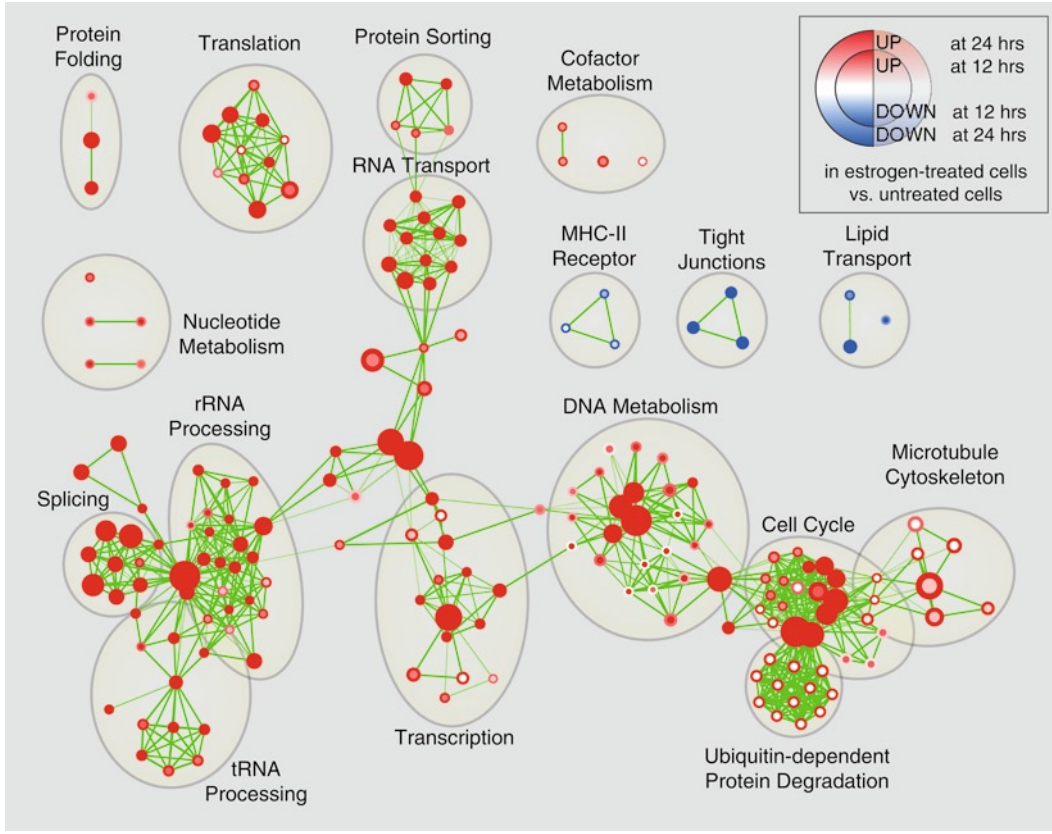

Fig. 4. Enrichment Map of the 12–24-h estrogen response enrichments. The network was manually rearranged to improve layout, and major clusters were manually labeled as in Fig. 3.

clusters of gene-sets belonging to the same functional theme were manually identified and annotated. Figure 5 shows the heat-map visualization of gene expression patterns for selected gene-sets.

3.6. Protocol 3: Query Set Analysis

Once an enrichment map has been generated, the *query set analysis* can be used to investigate the overlap between a *query gene-set*, not present in the map, and the enriched gene-sets in the map. In this protocol, we investigate the overlap between gene-sets enriched in the estrogen response (Protocol 2 example) and experimentally determined direct targets of the estrogen receptor (8). We are specifically interested in evaluating whether the transcriptional response to estrogen is dominated by indirect targets. The query set analysis can also be used to identify the relations between known disease genes and the gene expression or genetic alterations observed in that disease.

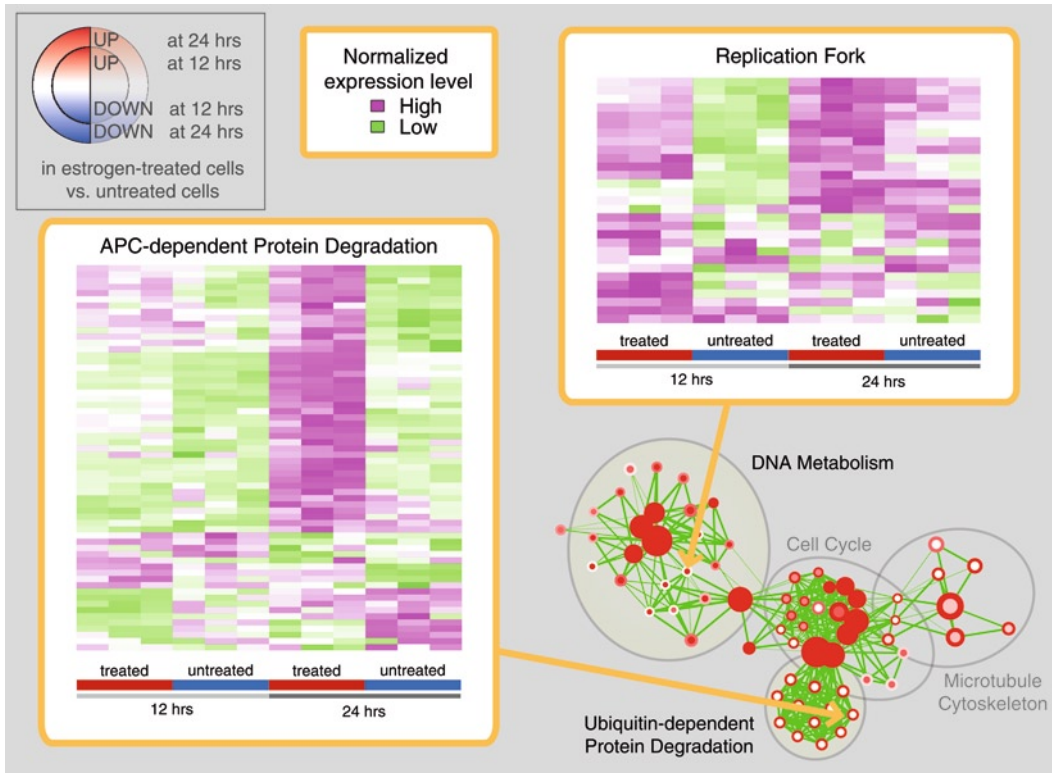

Fig. 5. Investigating gene expression patterns within two clusters of the 12–24-h estrogen response enrichment map. Enrichment of the *Replication Fork* gene-set is significant and associated to estrogen treatment (dark node center) at 12 h but not at 24 h (white node border); accordingly, the heat-map displays globally higher expression levels in estrogen-treated cells, although with a smaller difference at 24 h due to increased levels in the untreated cells, which is responsible for the absence of enrichment at that time point. Enrichment of the *APC-dependent Protein Degradation* gene-set is significant and associated to estrogen treatment at 24 h (dark node border) but not at 12 h (white node center); accordingly, the heat-map displays globally higher expression in estrogen-treated cells, although this pattern is much stronger at 24 h, explaining why a enrichment significance is observed only at that time point.

3.6.1. Load Data and Set Parameters

1. Select from the menu: *Plugins/Enrichment Map/Post Analysis.*
2. Make sure the gene-set file used to generate the current enrichment map is correctly loaded in the *GMT* box.
3. Load the query gene-set(s) in *SigGMT* box. Example data: load *estrogenTargetsLin2007.GMT*
4. Press the *Load Gene-sets* button.
5. Select the query gene-set(s) to use in the analysis and press the down arrow to add the selected gene-set(s) to the active selection. Example data: select *E2_TARGETS.*
6. Set the *Hypergeometric Test cutoff*. The overlap between the query gene-set(s) and preexisting gene-sets will be tested using the hypergeometric test (also known as Fisher's Exact Test) and only gene-set pairs with overlap *p*-value smaller than the

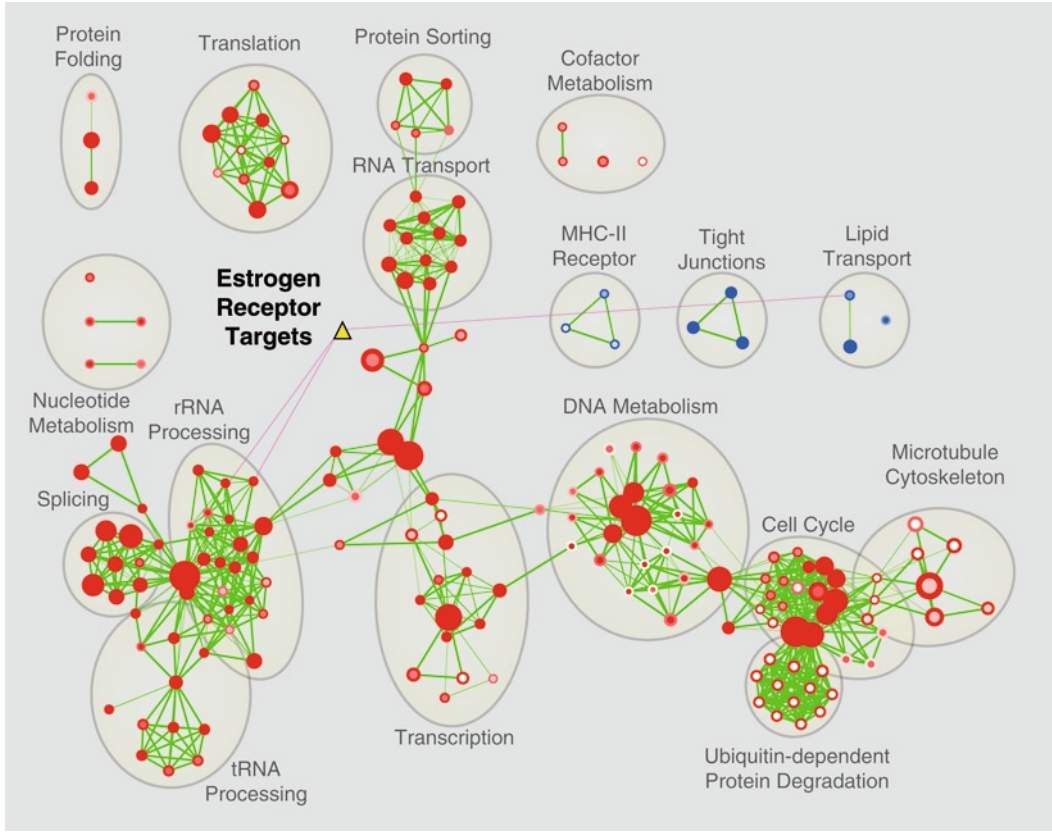

Fig. 6. Query set analysis investigating the relations between gene-sets enriched in the transcriptional response to estrogen treatment (12–24 h) and direct targets of estrogen receptor (query gene-set). The small triangle represents the query gene-set, and magenta links represent overlaps with uncorrected *p*-value <0.05 according to Fisher's Exact Test. Only a few links are present, even at a liberal *p*-value threshold. This suggests that either there are many false negatives in this target list, or the transcriptional response to estrogen is dominated by indirect targets.

cutoff will be connected by an edge. Example data: set to 0.05 (less stringent than default).

7. Press the *Run* button. The query gene-set will be displayed as a yellow triangle, connected to preexisting gene-sets by magenta edges. If no edges are present, the cutoff parameter may be too stringent.

3.6.2. Interpretation (Example Data)

1. The direct targets of the estrogen receptor and the gene-sets enriched by estrogen response have very little overlap: only three edges are displayed, even at a liberal Fisher's Exact Test *p*-value cutoff. Either the list of estrogen receptor targets is incomplete due to false negatives, or the transcriptional response observed after estrogen treatment is dominated by indirect targets.

Figure 6 displays the results of the query set analysis.

4. Notes

1. Gene-sets defined in the gene-set file are automatically intersected with the genes in the expression matrix before generating the gene-set network. Thus, genes that are not in the expression data set are not included in the enrichment map. This should be kept in mind when interpreting the results, especially if the expression matrix has a limited or biased coverage of the genome: the overlaps between gene-sets may be different than when using all genes in the gene-sets and/or they may be supported by a limited number of genes.
2. Unlike the FDR or *p*-value, it is harder to define general thresholds for the NES score color mapping. We suggest the user to: (a) define an initial mapping based on the NES value at typical FDR thresholds (e.g., associate the NES at FDR 0% or 1% to the highest color intensity and the NES at FDR 5% or 10% to white); for one-condition enrichment, only positive NES should be considered; for two-condition enrichment, the NES thresholds should be equal in absolute value for the positive and negative range (positive NES is associated to condition A, whereas negative NES is associated to condition B); (b) refine the initial mapping depending on the intended use of the Enrichment Map analysis; for instance, if only the most enriched gene-sets have intense color, it is very easy to prioritize the most enriched gene-sets within clusters; however, the other gene-sets may be less visible.
3. We recommend using Entrez-Gene as a gene identification system. Pathway Commons gene-sets are available with Entrez-Gene or official gene symbols. Several ID systems can be selected in WhichGenes, including Entrez-Gene. MSigDB gene-sets are available only with official symbols as gene IDs. The gene-sets at the Enrichment Map download page are available only with Entrez-Gene as gene IDs.
4. We recommend using capitalized alphanumeric IDs to identify gene-sets, rather than text labels (e.g., *GO:0007049* for GO *Cell cycle*). Capitalization is recommended as certain enrichment software packages, such as GSEA, use capitalized gene-set IDs in their results files.
5. The phenotype values represent the condition of gene-set enrichment. For one-condition enrichments, only one value will have to be used ("1"). For two-condition enrichments, two values should be used ("1" and "-1"). The choice of "-1", instead of other values, follows the sign convention of the NES score in the GSEA format (i.e., positive NES is associated to condition A and negative NES is associated to condition B).

References

1. Allison, D. B., Cui, X., Page, G. P., and Sabripour, M. (2006) Microarray data analysis: from disarray to consolidation and consensus., *Nature reviews. Genetics 7*, 55–65.
2. Huang da, W., Sherman, B. T., Tan, Q., Collins, J. R., Alvord, W. G., Roayaei, J., Stephens, R., Baseler, M. W., Lane, H. C., and Lempicki, R. A. (2007) The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists, *Genome Biol 8*, R183.
3. Huang da, W., Sherman, B. T., and Lempicki, R. A. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources, *Nat Protoc 4*, 44–57.
4. Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., and Mesirov, J. P. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles., *Proceedings of the National Academy of Sciences of the United States of America 102*, 15545–15550.
5. Isserlin, R., Merico, D., Alikhani-Koupaei, R., Gramolini, A., Bader, G. D., and Emili, A. (2010) Pathway analysis of dilated cardiomyopathy using global proteomic profiling and enrichment maps. *Proteomics* 1316–1327.
6. Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks, *Genome Res 13*, 2498–2504.
7. Cline, M. S., Smoot, M., Cerami, E., Kuchinsky, A., Landys, N., Workman, C., Christmas, R., Avila-Campilo, I., Creech, M., Gross, B., Hanspers, K., Isserlin, R., Kelley, R., Killcoyne, S., Lotia, S., Maere, S., Morris, J., Ono, K., Pavlovic, V., Pico, A. R., Vailaya, A., Wang, P.-L., Adler, A., Conklin, B. R., Hood, L., Kuiper, M., Sander, C., Schmulevich, I., Schwikowski, B., Warner, G. J., Ideker, T., and Bader, G. D. (2007) Integration of biological networks and gene expression data using Cytoscape., *Nature protocols 2*, 2366–2382.
8. Lin, C.-Y., Vega, V. B., Thomsen, J. S., Zhang, T., Kong, S. L., Xie, M., Chiu, K. P., Lipovich, L., Barnett, D. H., Stossi, F., Yeo, A., George, J., Kuznetsov, V. A., Lee, Y. K., Charn, T. H., Palanisamy, N., Miller, L. D., Cheung, E., Katzenellenbogen, B. S., Ruan, Y., Bourque, G., Wei, C.-L., and Liu, E. T. (2007) Whole-genome cartography of estrogen receptor alpha binding sites., *PLoS genetics 3*, e87.