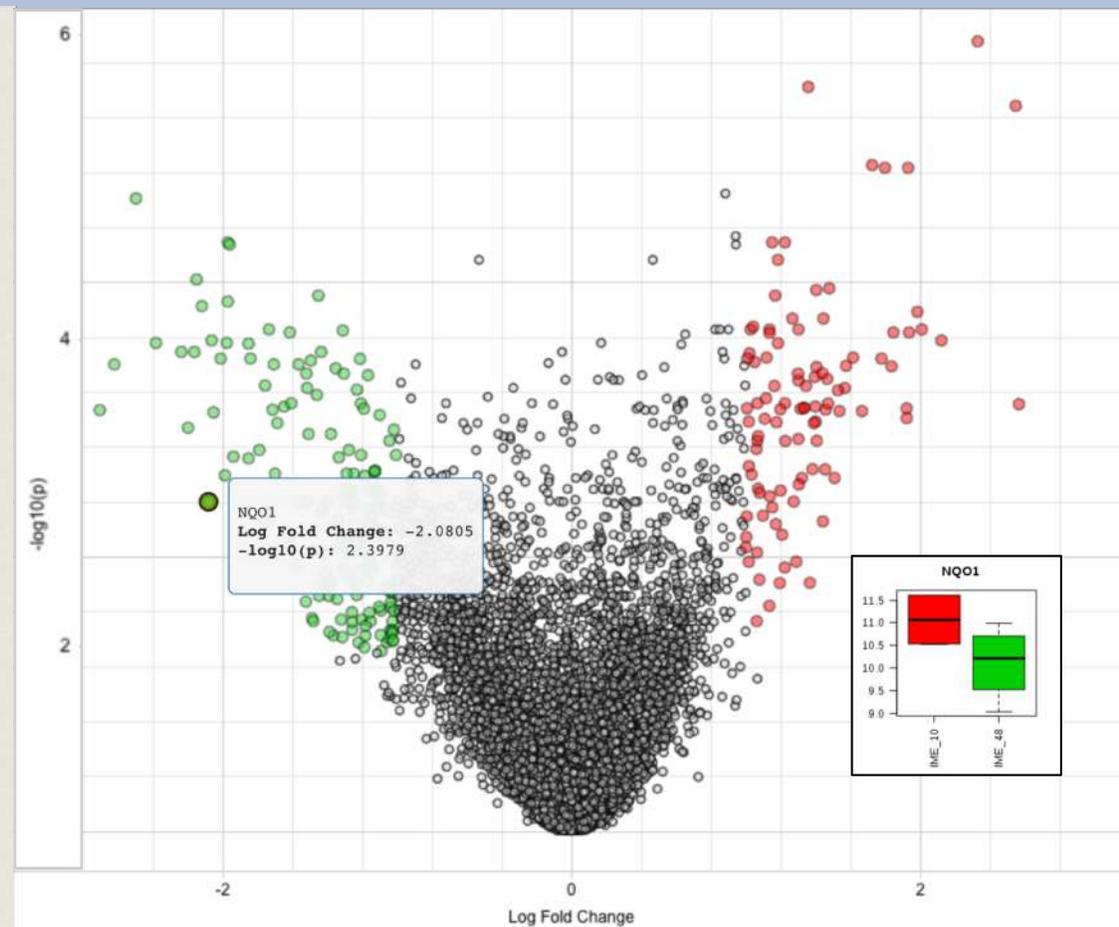
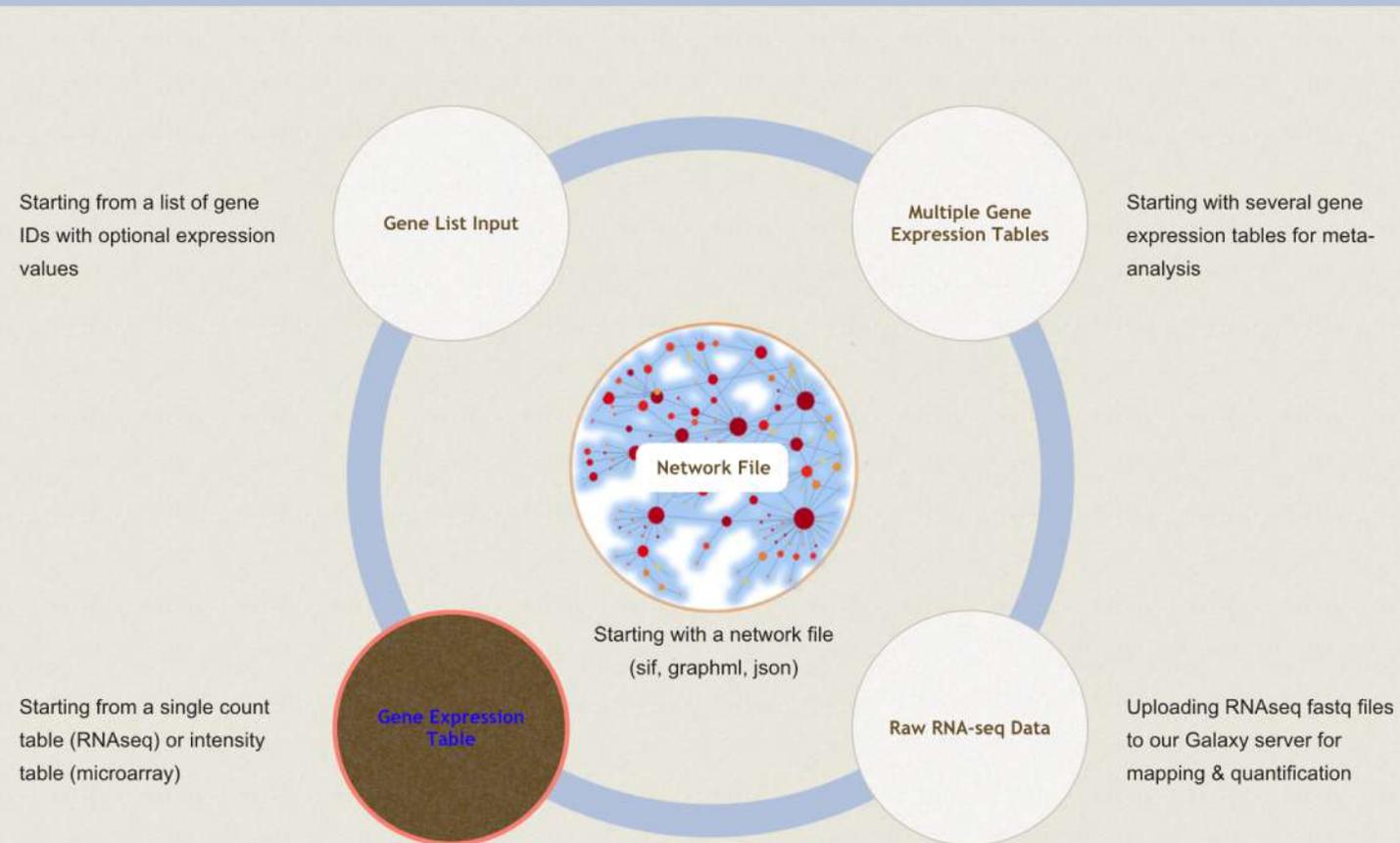


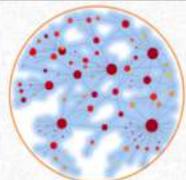
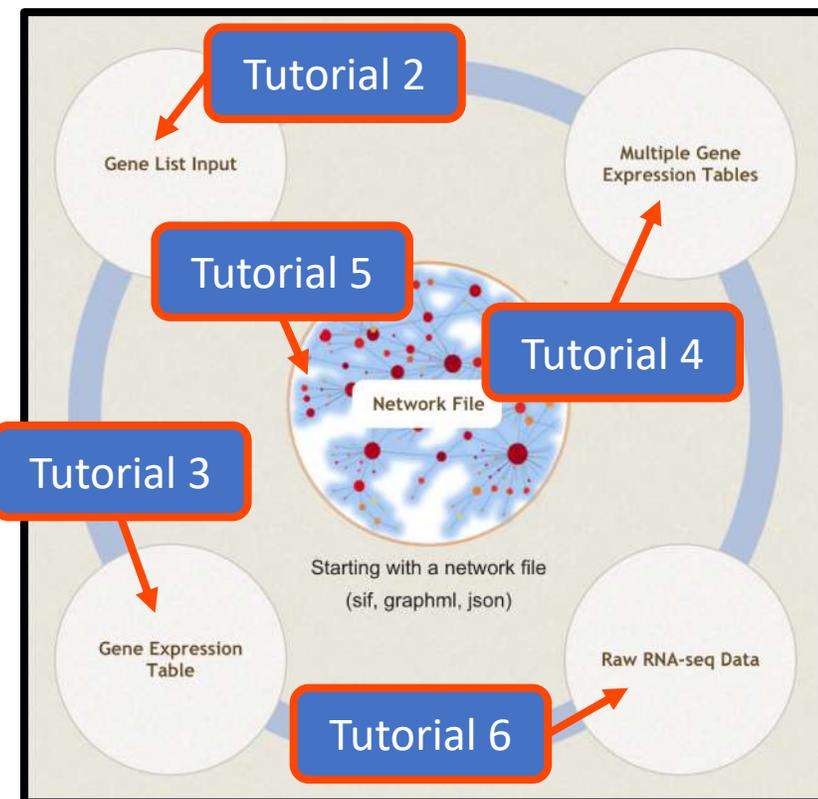
Tutorial 3: gene expression table



Intro to NetworkAnalyst

- Web application that enables complex meta-analysis and visualization
- Designed to be accessible to biologists rather than specialized bioinformaticians
- Integrates advanced statistical methods and innovative data visualization to support:
 - Efficient data comparisons
 - Biological interpretation
 - Hypothesis generation

Tutorial 1: Overview



NetworkAnalyst -- a web-based platform for gene expression profiling & biological network analysis

Computer and browser requirements

- A modern web browser with Java Script enabled
 - Supported browsers include Chrome, Safari, Firefox, and Internet Explorer 9+
- For best performance and visualization, use:
 - Latest version of Google Chrome
 - A computer with at least 4GB of physical RAM
 - A 15-inch screen or bigger (larger is better)
- Browser must be WebGL enabled for 3D network visualization
- 50MB limit for data upload
 - ~300 samples for gene expression data with 20 000 genes

Goals for this tutorial

- Differential expression analysis (DEA) is the foundation of most transcriptomics analysis
- Interpreting and communicating meaning from lists of differentially expressed genes is challenging without high quality visualization
- The goal of this tutorial is to perform DEA on example microarray data:
 - Visualize results of differential expression analysis
 - Perform functional analysis
 - Generate dimension reduction plots

Select example data

Uploaded data should be in matrix form, stored in a text file. See the [FAQs](#) page for details on how to format meta-data and gene IDs. Click on any of the example file names to see how these files are formatted.

The gene level summarization depends on the data type. Microarrays produce intensity data so duplicate probes should be averaged (mean or median). RNAseq produce counts data, so multiple gene transcripts should be added (sum).

Click "Submit" and "Proceed"

Navigation panel to track analysis progress

Select "Estrogen" example data

The screenshot shows a web application interface for gene expression analysis. On the left is a navigation panel with options: Upload Data, Quality Check, Normalization, Differential Analysis, Sig. Genes, Analysis Overview, Network Analysis, Download, and Exit. The main area is titled 'Upload your gene expression table' and contains a form with the following fields: 'Specify organism' (dropdown: ---Not specified---), 'Data type' (dropdown: ---Not specified---), 'ID type' (dropdown: --- Not Specified ---), 'Gene-level summarization' (dropdown: Mean), and 'Data File' (button: Choose File, text: No file chosen). A 'Submit' button is located to the right of the form. Below the form is a section titled 'Try our example data' with two radio buttons: 'Estrogen' (selected) and 'Endotoxin'. The 'Estrogen' example is described as 'Eight Affymetrix Human Genome U95 GeneChip data, normalized, log 2 scale' and 'Gene expression of a breast-cancer cell line (source) . Estrogen Receptor (ER): present, absent; Time (hour): 10, 48'. The 'Endotoxin' example is described as 'Illumina BeadArrays - Refseq normalized, log 2 scale' and 'Gene expression in human PBMC using LPS as inducer (details) Treatment: Control, LPS, LPS_LPS; Donor: 21, 46, 86, 92'. A 'Submit' button is located to the right of the examples. At the bottom right, there is a 'Proceed' button. Arrows and numbered circles (1, 2, 3) point to the 'Estrogen' radio button, the 'Submit' button, and the 'Proceed' button respectively.

View processing results

NetworkAnalyst

https://www.networkanalyst.ca/NetworkAnalyst/faces/Secure/expression/SummaryView.xhtml

NetworkAnalyst -- network-based visual analytics for gene expression profiling, i

Home ? FAQs Tutorials Gallery

Upload Data
Quality Check
Normalization
Differential Analysis
Sig. Genes
Analysis Overview
Network Analysis
Download
Exit

Data Quality Check

The uploaded samples are summarized below, together with several graphical outputs commonly used for quality check.

Data type:	Microarray gene expression
Total feature number:	11993
Matched gene number:	11195
Unmatched gene number:	798
Percent matched:	93.3
Sample number:	8
Number of experimental factors:	2
Group names:	Two factors found - ER: absent; present TIME: TIME_10; TIME_48

1

Box plot Count sum PCA plot Density plot

high48-2.cel
high48-1.cel
low48-2.cel

Previous Proceed

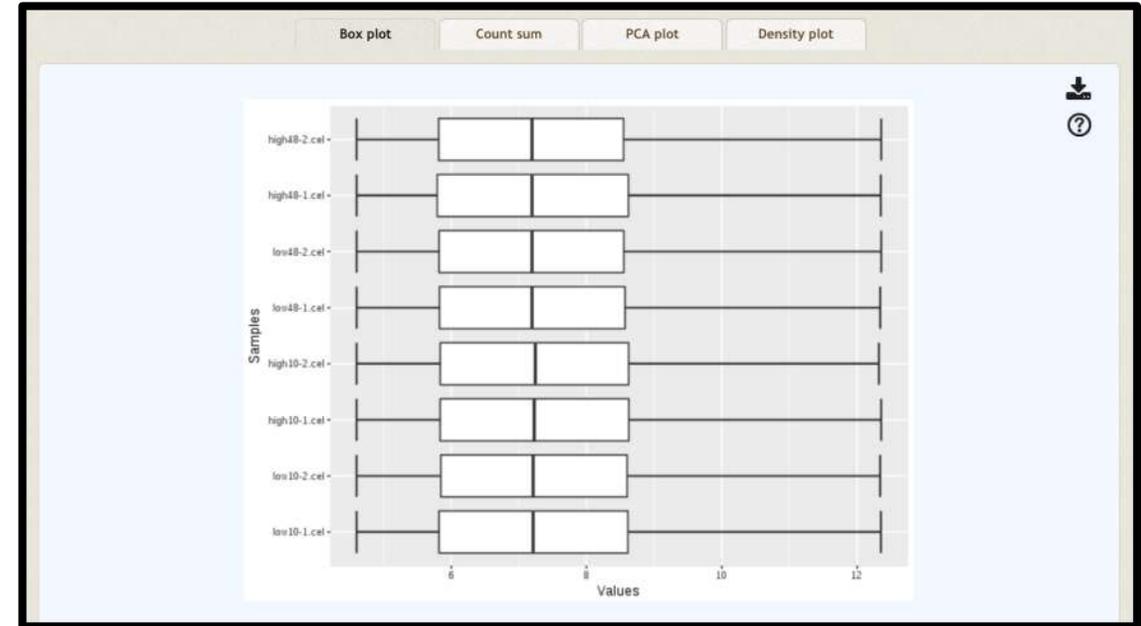
Xia Lab @ McGill (last updated 2019-01-23)

Check the processing results to ensure correct sample size, experimental factors, and adequate gene annotation

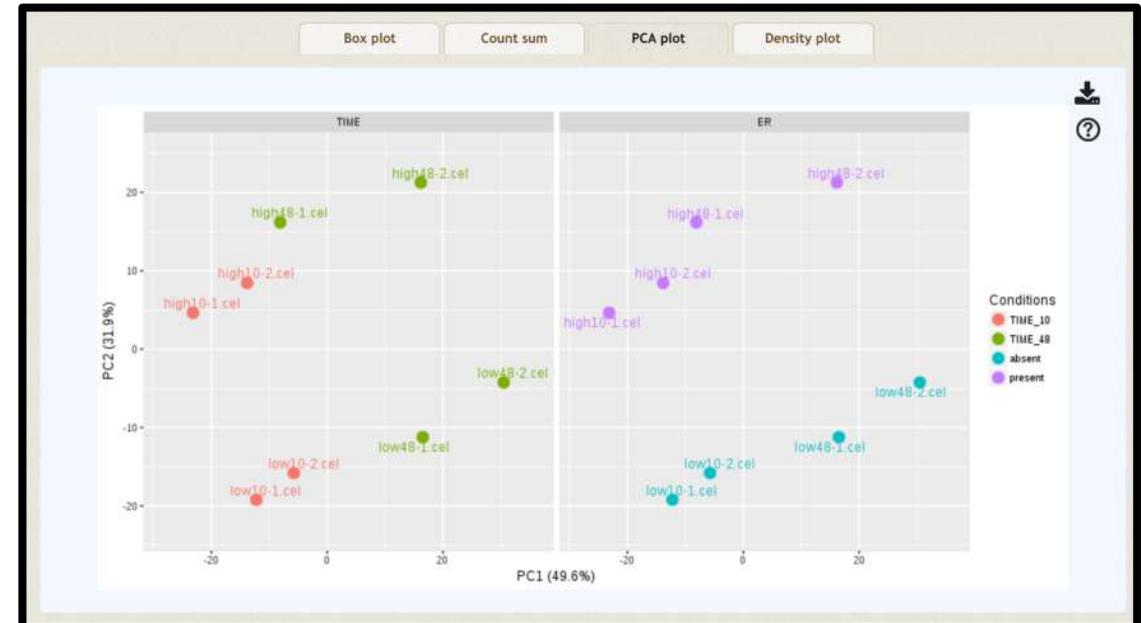
View common QA/QC plots to check the quality of the data

View processing results

Boxplot: since the gene expression intensity is < 20 for all samples, we know they have been log-transformed. Since they all have the same distribution, we know that they have been quantile normalized.



PCA plot: we see that the samples are separated by both time (TIME plot) and by the presence/absence of the estrogen receptor (ER plot). ER seems to be responsible for more variation than TIME.



Normalize and filter the data

Filtering increases statistical power by removing unresponsive genes prior to differential expression analysis (DEA). Proper normalization is essential to draw sound conclusions from the results of DEA.

Adjust the variance and abundance filter to change the number of genes that are excluded from downstream analysis. This number is a percentile – here the 15th percentile of data with the lowest expression will be removed

These are all established, frequently used gene expression normalization methods. DEA results after using different methods should be similar, but not exactly the same.

Click “Submit” to update the QA/QC plots after changing the filtering/normalization

The screenshot shows a web application interface for data filtering and normalization. The main heading is "Data Filtering & Normalization". Below this, there is a paragraph explaining that filtering removes uninformative or erroneous data, and normalization is crucial for reliable expression distributions. The interface includes several controls:

- Filtering:** A section with a "Variance filter" slider set to 15 and a "Low abundance" slider set to 5. There is also a checkbox for "Filter unannotated genes" which is checked.
- Normalization:** A section with radio buttons for different methods: "None" (selected), "Log2 Transformation", "Variance Stabilizing Normalization (VSN)", "Quantile Normalization", and "VSN followed by Quantile Normalization".
- Submit:** A button to update the QA/QC plots.
- Navigation:** "Previous" and "Proceed" buttons at the bottom.
- Plots:** Buttons for "Box plot", "PCA plot", "Density plot", and "MSD plot".
- Footer:** "Xia Lab @ McGill (last updated 2019-01-18)"

Normalize and filter the data

Usually we would normalize our raw data. Since the figures in the previous step showed that the example data was already normalized, select “None”.

The screenshot displays the 'Data Filtering & Normalization' section of the NetworkAnalyst web application. The interface includes a sidebar on the left with navigation options such as 'Upload Data', 'Quality Check', 'Normalization', 'Differential Analysis', 'Sig. Genes', 'Analysis Overview', 'Sig. Genes Analysis', 'Network Analysis', 'Functional Analysis', 'Dimension Reduction', 'Download', and 'Exit'. The main content area features a 'Data Filtering & Normalization' heading and a descriptive paragraph: 'Filtering serves to remove data that are unlikely to be informative or simply erroneous. Normalization is crucial for a reliable detection of transcriptional differences, and to ensure that the expression distributions of each sample are similar across the entire experiment.' Below this, there are two main sections: 'Filtering' and 'Normalization'. The 'Filtering' section includes a 'Variance filter' slider set to 15, a 'Filter unannotated genes' checkbox checked, and a radio button selected for 'None'. The 'Normalization' section includes radio buttons for 'Log2 Transformation', 'Variance Stabilizing Normalization (VSN)', 'Quantile Normalization', and 'VSN followed by Quantile Normalization'. A 'Submit' button is located to the right of the 'Filtering' section. At the bottom of the interface, there are buttons for 'Box plot', 'PCA plot', 'Density plot', and 'MSD plot', along with a plot area showing a scatter plot with a y-axis labeled '1.5'. A 'Previous' button and a 'Proceed' button are located at the bottom of the page. Three numbered callouts (1, 2, 3) are overlaid on the image, pointing to the 'None' radio button, the 'Submit' button, and the 'Proceed' button respectively.

Click “Submit” and “Proceed”

Conduct differential expression analysis

ER	noER		
ER10	ER48	noER10	noER48

One factor

Two factors

Differential Expression Analysis

Differential gene expression analysis using Limma, EdgeR or DESeq2 with support for different study designs.

Statistical method: Limma EdgeR DESeq2

Study Design

Primary Factor: ER

Secondary Factor: --- Not Available --- This is a blocking factor

Comparison of Interest

Specific comparison: absent versus present

Against a common control: absent

Nested comparisons: absent vs. present versus absent vs. present

Pairwise comparisons

Time series

Interaction only

Submit

Previous Proceed

Xia Lab @ McGill (last updated 2019-01-23)

If this was checked, there would only be two defined groups (ER, noER), but downstream statistical comparisons would “control for” differential expression driven by the second factor.

With two factors, you will have more “groups” of samples to compare

The two main steps of DEA are to group samples according to some factors (i.e. treatment vs. control, sex, time), and then specify which groups should be compared with statistical tests. While uploaded data may have more factors, only two can be considered in a single DEA.

Conduct differential expression analysis

We will do a simple, single factor study design. The goal of this analysis is to find the genes that are differentially expressed in cells that have an estrogen receptor (ER), compared to those that do not.

The last two statistical methods are available for RNAseq data

Set ER as the primary factor

Download
Exit

Differential Expression Analysis

Differential gene expression analysis using Limma, EdgeR or DESeq2 with support for different study designs.

Statistical method

Limma EdgeR DESeq2

Study Design

Primary Factor

ER

Secondary Factor

--- Not Available ---

This is a blocking factor

Comparison of Interest

Specific comparison

absent

versus

present

Against a common control

absent

Nested comparisons

absent vs. present

versus

absent vs. present

Interaction only

Pairwise comparisons

Time series

Submit

Click "Submit" and "Proceed"

Previous

Proceed

View differentially expressed genes (DEGs)

Here we see that 33 genes were up-regulated and 110 were down-regulated, according to standard p-value and log2 fold change thresholds. You can change the p-value and FC thresholds and see the effect it has on the # DEGs.

Please use the parameters to identify significant genes

Significance Thresholds
Adjusted p-value: ?
Log2 fold change: ?

Result Summary
Total sig. genes: 123 Up-regulated: 31 Down-regulated: 92

Sort table by: Sorting order:

The table shows at most top 500 genes ranked by p-values. Significant genes are in orange. You can download complete result using the button on the right.

ID	logFC	AveExpr	t	P.Value	adj.P.Val	B	View
PCNA	-2.2355	9.1368	-14.824	2.9131E-8	1.2722E-4	9.3547	
TK1	-2.8983	9.8509	-13.926	5.3778E-8	1.2722E-4	8.8315	
MYBL2	-2.9243	8.5321	-13.643	6.5757E-8	1.2722E-4	8.6571	
TFF1	-3.1988	12.116	-13.099	9.7764E-8	1.3401E-4	8.3091	
GLA	-1.5815	8.7099	-12.503	1.5359E-7	1.3401E-4	7.9006	
BAK1	1.7522	8.9396	12.496	1.545E-7	1.3401E-4	7.9013	
ID3	1.4969	11.529	12.438	1.6162E-7	1.3401E-4	7.8608	

Xia Lab @ McGill (last updated 2019-01-23)

Click "Download Results" for a .csv file of the statistics in the table. Click "Proceed" when finished.

1

2

Analysis overview

The screenshot shows the NetworkAnalyst web interface. The browser address bar displays the URL: <https://www.networkanalyst.ca/NetworkAnalyst/faces/Secure/expression/ExpressVisOverview.xhtml>. The page title is "NetworkAnalyst -- a comprehensive platform for gene expression pro...". A sidebar on the left contains a menu with options: Upload Data, Quality Check, Normalization, Differential Analysis, Sig. Genes, Analysis Overview, Network Analysis, Download, and Exit. The main content area has a heading "Please choose a type of analysis to proceed" and a paragraph of text: "Visual analytics technology aims to integrate interactive visualization with statistical analysis to help navigate complex data. In order to have a good user experience, you need to have a modern web browser with sufficient memory available. We recommend using: a) the latest version of Firefox; b) at least 15-inch display with 1440 x 900 resolution or higher; c) at least 2G available memory with Intel Core i5/i7 or equivalent;". Below this text is a tree diagram of analysis options. The "Sig. Gene Analysis" branch is expanded, showing sub-options: Network Visual Analytics, ORA Enrichment Network, Volcano Plot, ORA Heatmap Clustering, GSEA Enrichment Network, GSEA Heatmap Clustering, PCA 3D, and t-SNE 3D. A blue callout box with the number "1" and an arrow points to the "Volcano Plot" option. A blue callout box at the bottom left contains the text "Select 'Volcano Plot'". A dark grey callout box with an orange border on the right contains the text: "In addition to all of the network analytics, there are seven visualization tools available. In the rest of this tutorial we will explore each one. See tutorials 2 and 5 for examples of network visualization." The footer of the page reads "Xia Lab @ McGill (last updated 2019-01-18)".

In addition to all of the network analytics, there are seven visualization tools available. In the rest of this tutorial we will explore each one. See tutorials 2 and 5 for examples of network visualization.

1

Select "Volcano Plot"

Interactive volcano plot

4

Click on this icon to download high quality SVG format of volcano plot

When finished exploring, click "Analysis Overview" and select "ORA Heatmap Clustering"

Genes that do not pass the logFC or p-value threshold are shaded gray. Upregulated genes are **RED**, Downregulated genes are **GREEN**

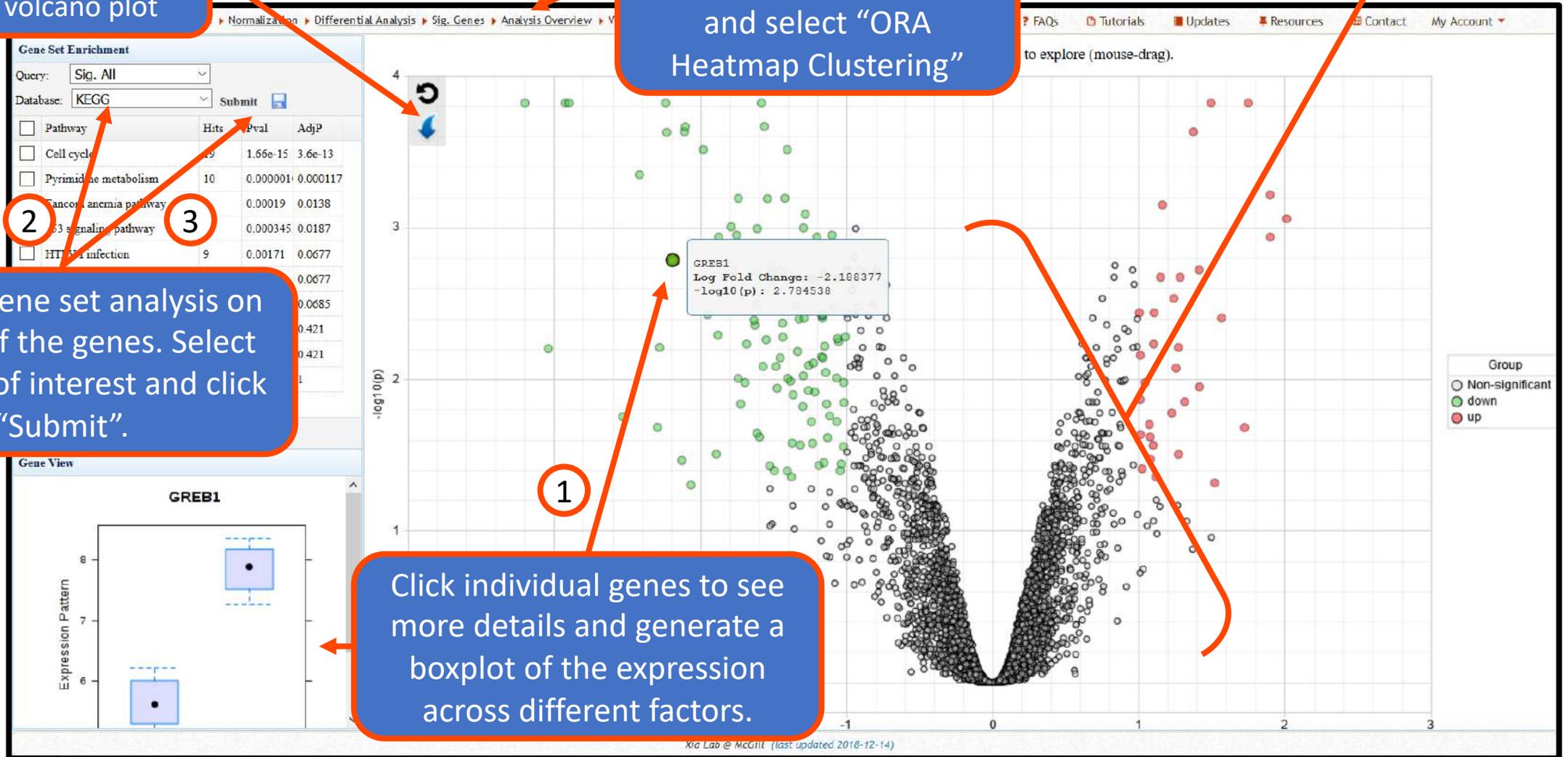
Perform gene set analysis on subsets of the genes. Select database of interest and click "Submit".

2

3

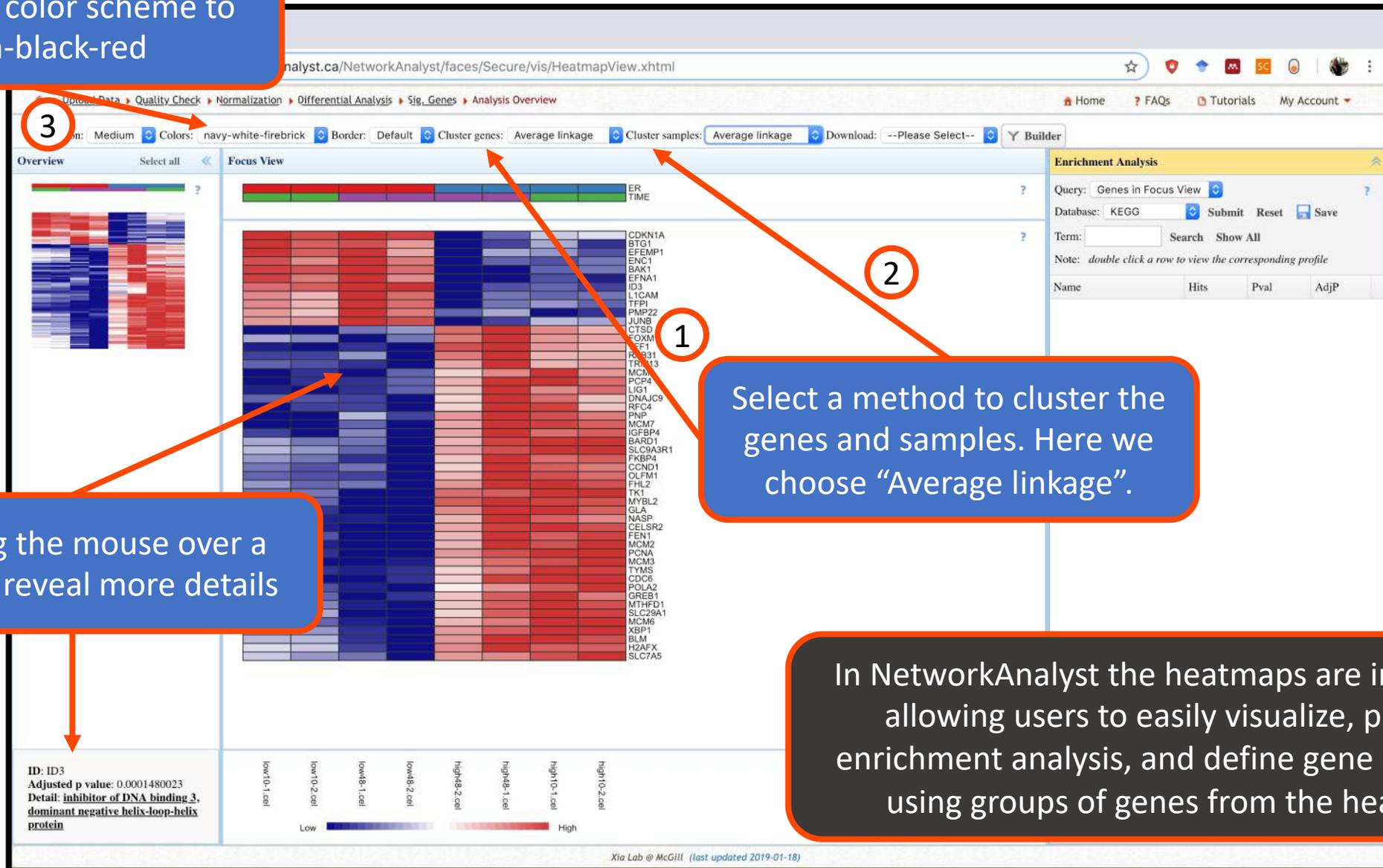
1

Click individual genes to see more details and generate a boxplot of the expression across different factors.



ORA Heatmap clustering and visualization

Change the color scheme to green-black-red



Select a method to cluster the genes and samples. Here we choose "Average linkage".

Hovering the mouse over a gene will reveal more details

In NetworkAnalyst the heatmaps are interactive, allowing users to easily visualize, perform enrichment analysis, and define gene signatures using groups of genes from the heatmap.

Advanced heatmap functions

The focus view is here

Change query to "Genes in Focus View" and click "Submit"

1

2

3

4

When finished exploring, click "Analysis Overview" and select "Gene Set Enrichment Analysis"

Select a group of genes with a distinct expression pattern in the overview by dragging your mouse. They will appear in the focus view.

These genes appear to be enriched in KEGG pathways related to cancer

Adjusted p value: 0.0005121747
Detail: insulin-like growth factor binding protein 4

Name	P-value	Hits
Cell cycle	1.51e-8	9
Pyrimidine metabolism	0.00245	4
One carbon pool by folate	0.00516	2
Bladder cancer	0.0118	2
HTLV-1 infection	0.0258	4
Glioma	0.0534	2
p53 signaling pathway	0.0578	2
Melanoma	0.0578	2
Chronic myeloid leukemia	0.0656	2
Glycosphingolipid biosynthesis - glot	0.0672	1

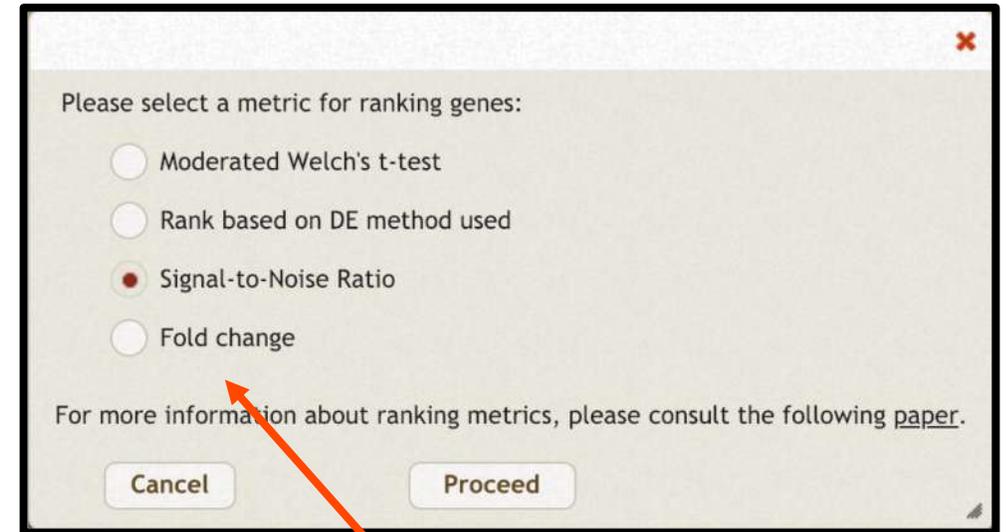
CDKN1A
STG1
BRD1
SLC9A3R1
FKBP4
CCND1
OLF1
FHL2
TK1
MYBL2
GLA
NASP
CELSR2
FEN1
MCM2
PCNA
MCM3
TYMS
CDC6
POLA2
GREG1

Low High

Xia Lab @ McGill (last updated 2018-12-14)

Gene Set Enrichment Analysis (GSEA)

- GSEA is a computational method for determining if the expression of a set of genes (biological pathways, etc.) is correlated with phenotypic differences between sample groups.
- GSEA incorporates actual gene expression data and so it is able to detect more sensitive differences.
- Refer to the original paper for more details on the GSEA:
 - <https://www.pnas.org/content/102/43/15545.short>



The first step in GSEA is to rank genes according to their expression. Try out several different methods – they should give similar results.

GSEA Heatmap Clustering

The screenshot shows the NetworkAnalyst web interface. The browser address bar is <https://www.networkanalyst.ca/NetworkAnalyst/faces/Secure/vis/GSEAView.xhtml>. The navigation menu includes: Upload Data, Quality Check, Normalization, Differential Analysis, Sig. Genes, and Analysis Overview. The Analysis Overview page is active, showing a Dataset dropdown, Color scheme (navy-white-firebrick), Border (Yes), Cluster (Both), Cluster distance (Euclidean), Cluster linkage (Average), and Download options.

Enrichment Plot: A line graph showing enrichment scores. The y-axis is labeled 'enrichment score' and ranges from 0.0 to 0.6. A green curve peaks at approximately 0.6. A red dashed horizontal line is at the top of the plot.

Enriched Pathways: A table with columns: Name, Hits, Pval, and AdjP. The 'Cell cycle' pathway is selected (checked) and highlighted in yellow.

Name	Hits	Pval	AdjP
<input checked="" type="checkbox"/> Cell cycle	43/65	1.43e-4	0.00501
<input type="checkbox"/> Purine metabolism	23/30	1.44e-4	0.00501
<input type="checkbox"/> Oocyte meiosis	22/30	1.51e-4	0.00501
<input type="checkbox"/> Pyrimidine metabolism	22/30	1.53e-4	0.00501
<input type="checkbox"/> Fanconi anemia pathway	25/40	1.7e-4	0.00501

Heatmap: A heatmap titled 'Cell cycle' showing enrichment scores for various genes. The y-axis lists genes: E2F1, PKMYT1, CHEK2, PLK1, ESPL1, CDC20, PTTG1, CCNB1, CCNA2, CCNB2, SMC3, E2F3, CDC7, CDKN2C, BUB1, DBF4, ANAPC5, ORC1, CDC25A, and CCNE2. The x-axis is labeled 'Class Secondary' with categories 'absent' and 'present'. A color scale from blue (low) to red (high) is shown. A 'TIME 48' label is present above the heatmap.

Annotations:

- 1:** Points to the 'Submit' button in the Enriched Pathways table.
- 2:** Points to the 'Cell cycle' pathway in the Enriched Pathways table.
- 3:** Points to the 'Analysis Overview' link in the navigation menu.

Text Box 1 (Left): Select the pathway of interest to generate an enrichment plot and heatmap of enrichment scores. Click "Submit".

Text Box 2 (Right): When finished exploring, click "Analysis Overview" and select "GSEA Enrichment Network".

Xia Lab @ McGill (last updated 2019-01-18)

GSEA Enrichment Network

Choose from 9 different databases to perform GSEA on. Select "GO:BP" and click "Submit"

See tutorial 2 for more details on the visualization tools to manipulate the network appearance

- KEGG
- Reactome
- ✓ GO:BP
- GO:MF
- GO:CC
- PANTHER:BP
- PANTHER:MF
- PANTHER:CC
- Motif

Enrichment analysis

Database: GO:BP

Operation: Union Submit

Data Name	Size
dataSet	5804

Current gene list

Union (1 datalists)

Name	Hits	Pval	AdjP
Cell cycle phase	41/90	1.12e-4	2.02e-3
Mitotic cell cycle	25/50	1.13e-4	2.02e-3
MRNA metabolic process	69/135	1.15e-4	2.02e-3
Response to DNA damage	29/46	1.15e-4	2.02e-3
Regulation of cell cycle	90/140	1.15e-4	2.02e-3
RNA processing	159/374	1.16e-4	2.02e-3
Chromosome organization	258/510	1.16e-4	2.02e-3
M phase	57/128	1.22e-4	2.02e-3

Each significantly enriched gene set from GSEA is represented as a node. Gene sets with overlapping genes are connected with an edge (calculated using the overlap coefficient or Jaccard index). The network visualization simplifies the interpretation of GSEA results by grouping similar gene sets together.

Let's look at this cluster in more detail. Hover your mouse over each node to find the gene set name, and select it in the results table.

GSEA Enrichment Network

Gene set nodes in enrichment networks are “meta-nodes” because clicking them reveals more nodes that correspond to the overlapping genes from that gene set. Expanding meta-nodes for the whole network can result in an extremely dense network, so here we extract a few gene sets of interest first.

Click “Extract” to visualize the selected gene sets on their own

The screenshot shows the GSEA Enrichment Network interface. On the left, there is a sidebar with a search bar and a table of data sets. The main area displays a dense network of nodes (colored red, yellow, and blue) connected by lines. On the right, a 'Result Table' is visible, listing various biological processes with checkboxes for selection. A red circle with the number '1' is placed over the 'Extract' button in the table.

Name	Hits
<input checked="" type="checkbox"/> Pattern specification proc...	69/1
<input type="checkbox"/> Positive regulation of imr...	17/5
<input type="checkbox"/> Regulation of response to	394
<input type="checkbox"/> Transmembrane receptor	42/9
<input checked="" type="checkbox"/> Morphogenesis of an epit...	212
<input type="checkbox"/> Angiogenesis	85/1
<input type="checkbox"/> Regulation of cell migrati...	218
<input type="checkbox"/> Axon guidance	100
<input type="checkbox"/> Regulation of cytokine pr...	28/4
<input type="checkbox"/> Regulation of defense res...	73/1
<input checked="" type="checkbox"/> Tube development	20/5
<input checked="" type="checkbox"/> Heart development	126
<input type="checkbox"/> Negative regulation of ce...	4/1
<input type="checkbox"/> Cytokine production	47/1
<input checked="" type="checkbox"/> Embryonic morphogenesis	62/1
<input type="checkbox"/> Inflammatory response	142
<input type="checkbox"/> Immune effector process	97/2
<input checked="" type="checkbox"/> Tissue morphogenesis	44/8
<input type="checkbox"/> Regulation of MAPK cas...	36/1

This close-up view shows a smaller, less dense network of nodes, all colored yellow, representing the selected gene sets. The nodes are interconnected by lines, forming a clear structure. The background is dark, making the nodes and connections stand out.

GSEA Enrichment Network

Double-click a few meta-nodes to see the nested genes. After, select the “Bipartite network” view to expand all of the meta-nodes.

Change the scope to “Node-neighbours” and drag the meta-nodes apart to decrease the network density.

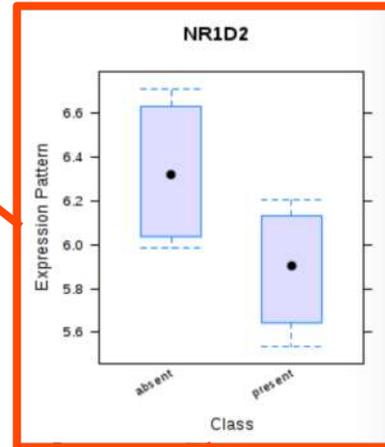
The screenshot displays the NetworkAnalyst web interface for GSEA Enrichment Network analysis. The interface is divided into several sections:

- Enrichment analysis sidebar:** Shows the database set to GO:BP, operation as Union, and target as dataSet. A table lists the data sets: Data Name and Size (dataSet, 5804).
- Current gene list:** Includes a search bar and a section for Union (1 datalists) with a note: "Genes not shown for large lists. Click on a gene set node to see corresponding genes."
- Network visualization:** A central network graph with nodes and edges. A callout box labeled '1' points to the 'View: Bipartite network' dropdown menu. A callout box labeled '2' points to the 'Scope: Node-neighbours' dropdown menu.
- Result Table:** A table of selected functions with checkboxes and hit counts. The current selection is for 'Heart development', which includes a list of genes: DKK1, EFNA1, ID3, SIK1, MB, and CITED2.

At the bottom of the page, the text "Xia Lab @ McGill (last updated 2019-01-23)" is visible.

Dimension reduction plots

Click on a gene in the loading plot to see a boxplot across treatments



PCA and tSNE are both popular methods of capturing whole-transcriptome changes in expression in a few variables. tSNE is a stochastic method and so plots will vary slightly each time they are generated.

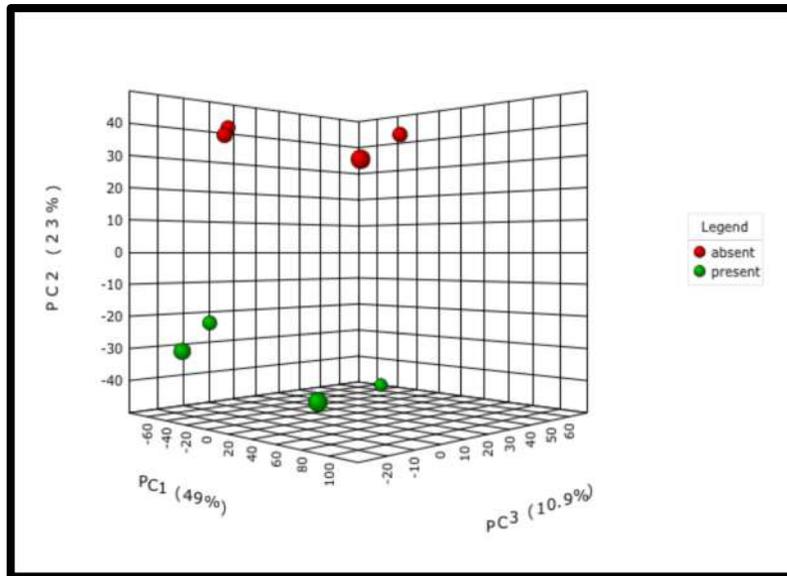


Figure 1: 3D PCA

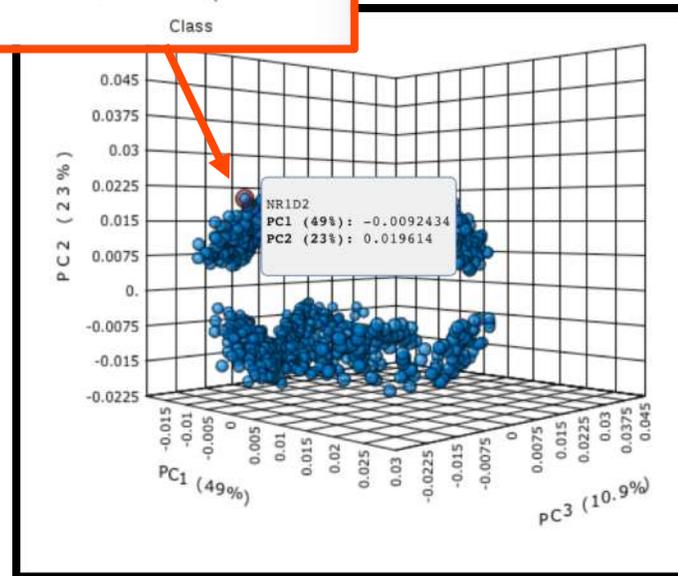


Figure 2: 3D PCA loading plot

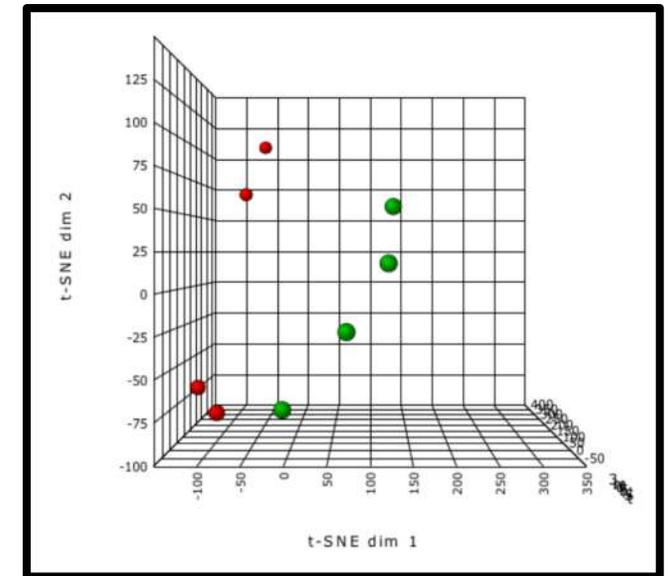
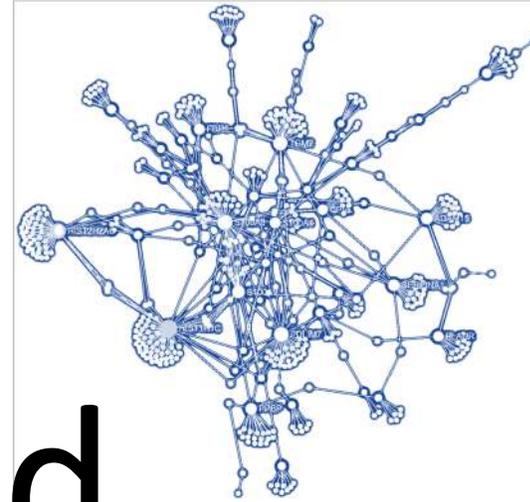


Figure 3: 3D tSNE



The End

*For more information, visit the **FAQs**, **Tutorials**, **Resources**,
and **Contact** pages on www.networkanalyst.ca*