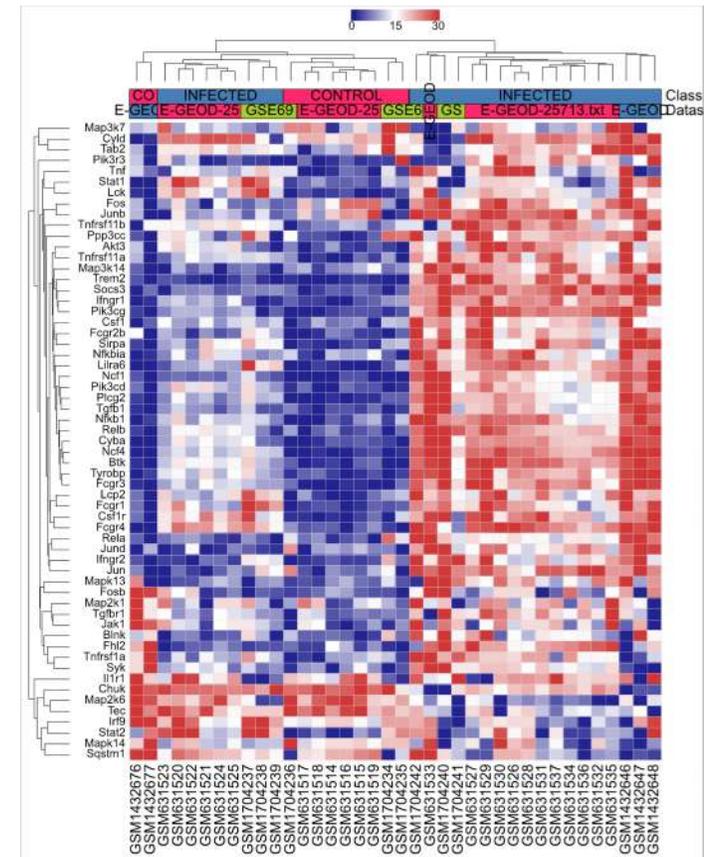
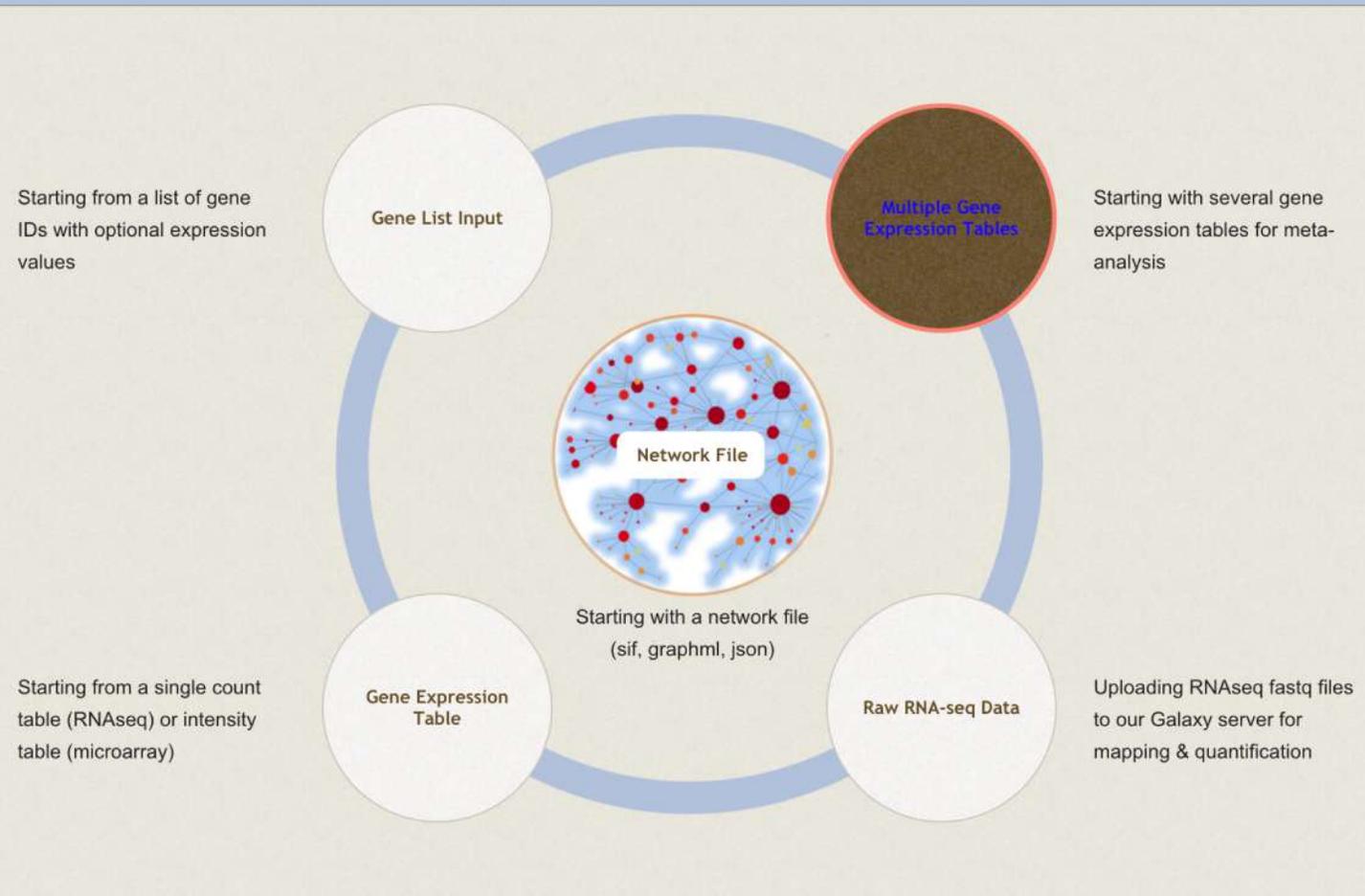


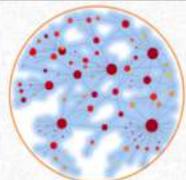
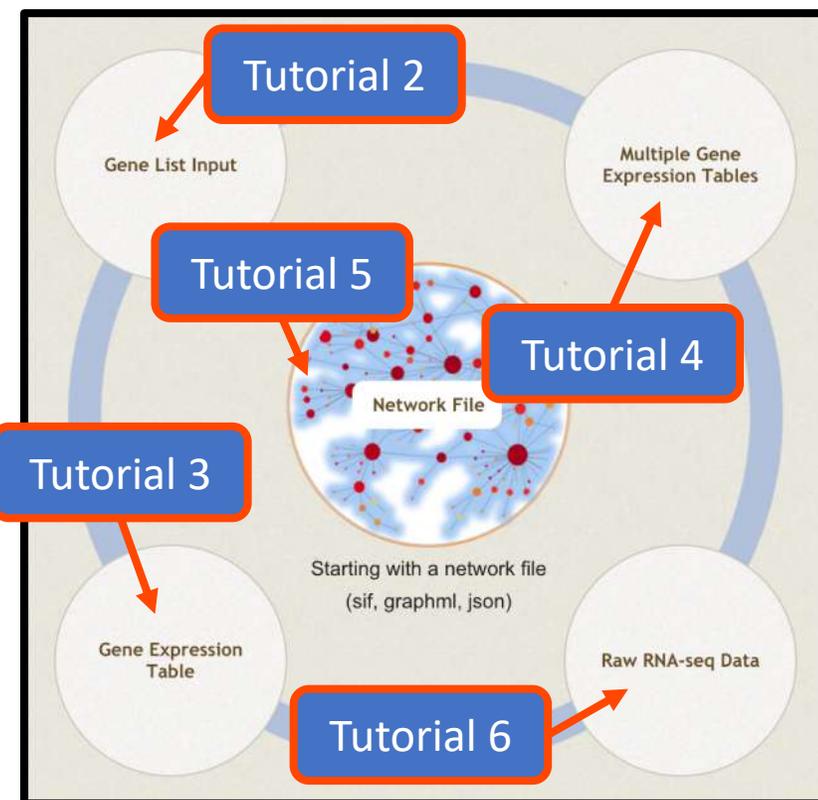
# Tutorial 4: multiple gene expression tables



# 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

- A meta-analysis is a quantitative synthesis of results from multiple studies that test similar hypotheses
- Gene expression meta-analyses aim to identify robust molecular signatures and functional enrichment results to increase understanding of biological processes
- Requires advanced statistics and visualization strategies
- The goal of this tutorial is to complete a meta-analysis of expression profiles from 3 different studies:
  - Perform and combine statistical tests
  - Visualize results in interactive heatmaps, Venn diagrams, and 3D PCA plots

# Appropriate datasets

- The two main steps of a meta-analysis are:
  - Systematic literature review to identify studies that test the same hypothesis
  - Rigorous statistical analysis of the datasets using established methods
- NetworkAnalyst provides a platform for the second step
- For the meta-analysis to be a success, appropriate datasets should be used:
  - Study designs should compare the same experimental factors
  - Gene expression platforms should be comparable (i.e. studies should not be spread over > 10 years)
  - Relative similarity of host factors (i.e. species, tissue, sex, age etc.)

# Upload data

The first step is to upload and process all of your individual datasets. This repeats the steps of a single gene expression table for each dataset - for more details on each step, see tutorial 3.

You can edit meta-data labels here

**Annotate Experiments**

- Choose a condition: Only need
- Group labels: Edit group labels ALL datasets uploaded. To exclude group label to NA.
- Samples with group label: To edit the corresponding whole row.

Choose a condition: Dose

Group labels: Control Treatment

Samples with group label	
QE5-M1-LT	Medium
QE5-M2-LT	Medium
QE5-M3-LT	Medium
QE5-M4-LT	Medium
QE5-M5-LT	Medium
QE1-51-LT	Control
QE1-52-LT	Control
QE1-53-LT	Control

Submit

Use the panel below to upload and prepare each individual data

Click the individual cell to activate each process. Click **Add New** to add a new data set. The maximum total number of samples allowed is **1000**. When all data sets have been processed, Click **Proceed** to proceed. Click the **Try our example** if you want to use example datasets to explore the functions available.

Data Upload	ID Conversion	Annotation	Visualization	Normalization	DE Analysis	Data Summary	Include
Process	Annotate	View	Normalize	Analyze	View		

Add New Try our example

Check QA/QC plots before and after normalization



**Data Formats**

Make sure you have read the [instructions](#) before uploading your datasets. The maximum file size per upload is **50M**. Data can be uploaded as a tab-delimited text file (.txt) or its compressed format (.zip).

Choose File No file chosen Submit

If you don't have supported IDs, ensure the same annotation is used across all datasets and leave the second box "unspecified"

**Data Processing**

The purpose is to convert gene/probe IDs to a common ID format (Entrez ID), so that different data sets can be properly compared. There are 47 built-in microarray probe ID libraries for human, mouse, and rat. For other IDs or organisms, make sure all datasets have the same ID type.

Specify organism: C. japonica (japanese quail)

Gene ID or Probe platform: Entrez ID

OK ID Conversion: Total [ 20165 ] Matched [ 16024 ] Unmatched [ 4141 ]

Process Done

# Upload data

The screenshot displays the NetworkAnalyst web interface. At the top, the browser address bar shows the URL: <https://www.networkanalyst.ca/faces/uploads/MetaLoadView.xhtml>. The main header includes the NetworkAnalyst logo and the text "NetworkAnalyst -- a web-based platform for gene expression profiling &". Below the header, there is a navigation bar with "Home" and "FAQs" links. The main content area is titled "panel below to upload and prepare each individual data" and contains instructions: "individual cell to activate each process. Click **Add New** to add a new data set. The maximum total number of samples allowed is **1000**. When all data sets have been processed, **Proceed** to proceed. Click the **Try our example** if you want to use example datasets to explore the functions available."

The interface features a horizontal menu with the following items: Upload, ID Conversion, Annotation, Visualization, Normalization, DE Analysis, Data Summary, Include, and a trash icon. Below this menu are buttons for "Add New", "Try our example", and "Upload merged data".

Three panels are highlighted with orange boxes and arrows:

- DE Analysis Panel:** This panel allows users to perform differential expression analysis. It includes a "Contrasts" section with dropdown menus for "Control" and "Treatment" separated by "versus". Below this is a "Set p value (FDR) cutoff" field with the value "0.05" and a "Submit" button. A pie chart below the form shows "Sig [167]" in blue and "Non-Sig [15856]" in orange.
- Data Normalization Panel:** This panel provides instructions on expression value comparison and normalization. It includes a "Data Type" dropdown set to "RNA-seq data (counts)" and a "Normalization procedure" dropdown set to "Log2-counts per million". There is also a checkbox for "Perform auto-scaling" and a "Submit" button.
- Data Summary Panel:** This panel summarizes the results of previous steps. It includes a "Data Summary" section with a text box explaining that users should decide which two conditions to compare if there are more than two. Below this is a table of statistics: "Number of annotated genes: 20165", "Number of available samples: 10", "Missing values: 0 (0.0%)", "Normalization procedures used: Log2-counts per million", and "Number of DE genes: 167". At the bottom, there is a "Set order of comparison:" section with dropdowns for "Control" and "Treatment" separated by "versus", and a "Done" button.

At the bottom of the page, there is a "Proceed" button and a footer that reads "Xia Lab @ McGill (last updated 2018-12-14)".

Make sure the contrasts compare the same factors for all uploaded datasets

Repeat this process for additional datasets

# Select example data

For the rest of this tutorial, we will use the example data

Click "Try our example" and "Yes"

Click "Proceed"

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

Use the panel below to upload and prepare each individual data

Click the individual cell to activate each process. Click **Add New** to add a new data set. The maximum total number of samples allowed is **1000**. When all data sets have been processed, click **Proceed** to proceed. Click **Try our example** if you want to use example datasets to explore the functions available.

Datasets	Data Type	Description	Phenotype
<a href="#">E-GEOD-25713</a>	Illumina MouseWG-6 v2.0 Bead Array;	Three testing datasets (containing subset of 5000 genes) from a meta-analysis of helminth infections in mouse liver ( <a href="#">details</a> )	CONTROL INFECTED
<a href="#">E-GEOD-59276</a>	Illumina MouseRef-8 v2.0 Bead Array;		
<a href="#">GSE69588</a>	Affymetrix Mouse Gene 1.0 ST Array;		

Buttons: Yes, Cancel, Try our example, Upload merged data, Previous, Proceed

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

# View integrity check results

For a meta-analysis to be done properly, the individual analyses must test contrasts between the same factors. The integrity check ensures that the labels are consistent for all previous analytical steps.

Use the panel below to upload and prepare each individual data

Click the individual cell to activate each process. Click **Add New** to add a new data set. The maximum total number of samples allowed is **1000**. When all data sets have been processed, Click **Proceed** to proceed. Click the **Try our example** if you want to use example datasets to explore the functions available.

Data Upload	ID Conversion	Annotation	Visualization	Normalization	DE Analysis	Data Summary	Include	
✓ E-GEOD-25713	✓ Process	✓ Annotate	View	✓ Normalize	✓ Analyze	✓ View	✓	
✓ E-GEOD-59276	✓ Process	✓ Annotate	View	✓ Normalize	✓ Analyze	✓ View	✓	
✓ GSE69588	✓ Process	✓ Annotate	View	✓ Normalize	✓ Analyze	✓ View	✓	

Integrity Check Result

OK, all datasets passed integrity check. Click **Next** button to next page.

You can download the merged data here: [Download](#)

Cancel Next

1

Click "Next" and "Proceed"

2

Previous Proceed

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

# View raw data and correct batch effect

NetworkAnalyst

https://www.networkanalyst.ca/NetworkAnalyst/faces/Secure/metastat/MetaQCView.x

NetworkAnalyst -- network-based visual analytics for gene expr

Adjust study batch effect (Combat)  Update

PCA plot Density plot

PC2 (1.5%)

PC1 (97.5%)

Conditions

- CONTROL
- INFECTED

Datasets

- E-GEO-25713.txt
- E-GEO-59276.txt
- GSE69588.txt

PC2 (8%)

PC1 (52.1%)

Conditions

- CONTROL
- INFECTED

Datasets

- E-GEO-25713.txt
- E-GEO-59276.txt
- GSE69588.txt

Previous Proceed

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

1

2

3

Use the PCA and density plots to check the quality of the data. Here we see significant batch effect, so select Combat and click "Update".

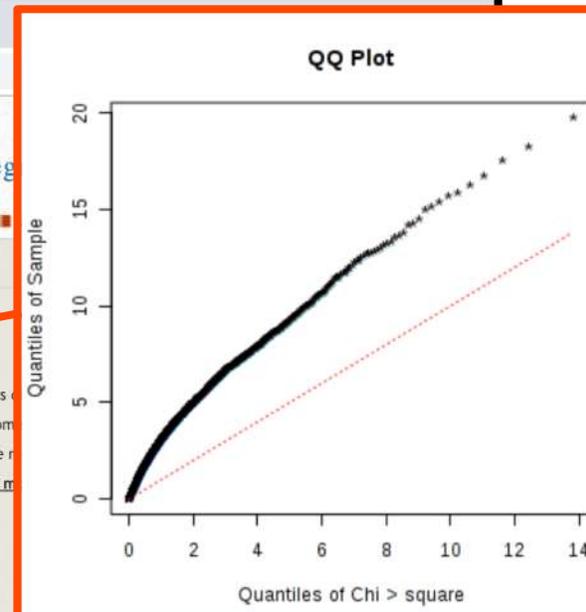
After applying Combat, the study batch effect has been greatly reduced!

Click "Proceed"

# Conduct gene-level meta-analysis

NetworkAnalyst has four approaches for gene-level meta-analysis. The first two are recommended, while the second two (vote counting and direct merging) should be used for exploratory purposes only. Since we have many DEGs, we choose to combine based on effect sizes.

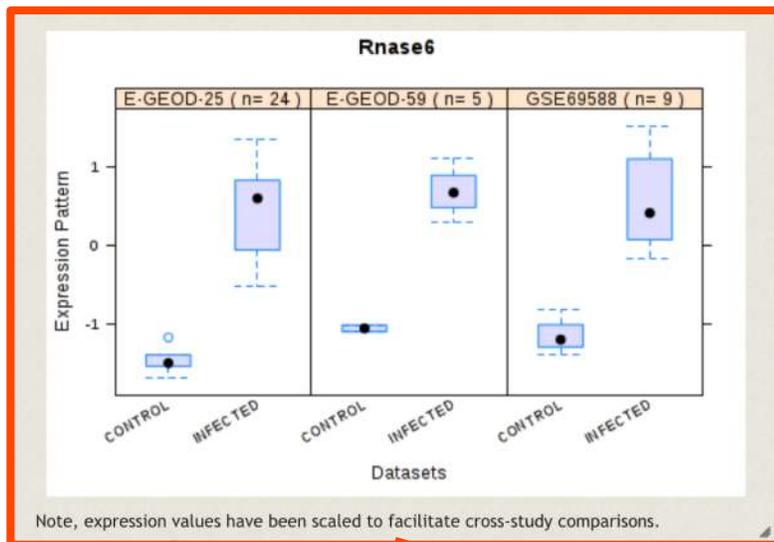
From the Q-Q plot we see that the data deviates substantially from the straight line, so select REM.



Here we will base the meta-analysis on effect sizes. To choose between a FEM and REM, generate a Q-Q plot by clicking "Cochran's Q Tests".

Click "Proceed"

# View results of meta-analysis



Work-based visual analytics for gene expression profiling, integration & systems understanding

Resources Contact Account

Result

Analysis are given in columns with the corresponding gene ID. The complete result is available in the Download button below.

Sort by: Pval Sorting order: Ascending Update Search Download

ID	E-GEOD-25713	E-GEOD-5927	Pval	View		
Rnase6	1.426	1.0526	2.01E-7			
Ccr5	1.6814	1.2098	2.01E-7			
Krt23	-2.5724	-1.5873	2.07E-6			
Gvin1	0.7015	0.80062	4.94E-6			
Tnfrsf18	1.1219	0.88817	2.27E-6			
Csf2rb2	1.9952	0.78834	3.727E-6			
Slc22a5	-0.71141	-0.93643	-1.1026	-2.7748	1.026E-5	
Ccl19	0.94445	0.69382	0.83701	2.6591	1.6905E-5	
Cd44	0.79847	0.43892	0.71901	2.6544	1.6966E-5	
Ikzf1	0.77847	0.48341	0.65178	2.5918	1.9326E-5	
Pira11	1.3282	0.66736	0.59659	2.6213	1.9326E-5	
Fkbp11	0.88198	0.97039	0.71163	2.5663	2.4342E-5	

Previous Proceed

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

The results can be sorted before being downloaded as a .csv file.

Click on the picture icon to see a boxplot of a specific gene across datasets

Click "Proceed"

1

# Analysis overview

We now want to further analyze and visualize the results of the statistical analysis. There are 4 datasets to work with: the 3 individual datasets and their significant genes, and the combined statistics from the meta-analysis. The “Sig. Gene Analysis” tools are based on the 4 lists of significant genes. The “Global Analysis” tools use the matrix of combined statistics from the meta-analysis for GSEA tools and all gene expression data for PCA 3D.

See tutorial 2b for more details on how to use Venn and Chord Diagrams to compare the overlap of multiple gene lists

Upload Data  
Quality Check

#### Data available:

- E-GEOD-25713 (2963)
- E-GEOD-59276 (2878)
- GSE69588 (33)
- Genes from meta-analysis (2022)

Note: you need to perform gene-level meta-analysis to select Genes from meta-analysis option.

Cancel

OK

Visual analytics technology aims to integrate interactive visualization with statistical analysis to help navigate complex data. To get the best experience, you need to have a modern web browser with sufficient memory available. We recommend using: a) the latest version of Google Chrome; b) Firefox; c) at least 15-inch display with 1440 x 900 resolution or higher; d) at least 2G available memory with Intel Core i5/i7 or equivalent;

Sig. Gene Analysis

Global Analysis

Network Visual Analytics

ORA Enrichment Network

ORA Heatmap Clustering

Venn Diagram

Chord Diagram

GSEA Enrichment Network

GSEA Heatmap Clustering

PCA 3D

Network Visual Analytics can only be performed on a single list at a time. See tutorial 2a and 5 for more details on creating networks.

# ORA Heatmap Clustering

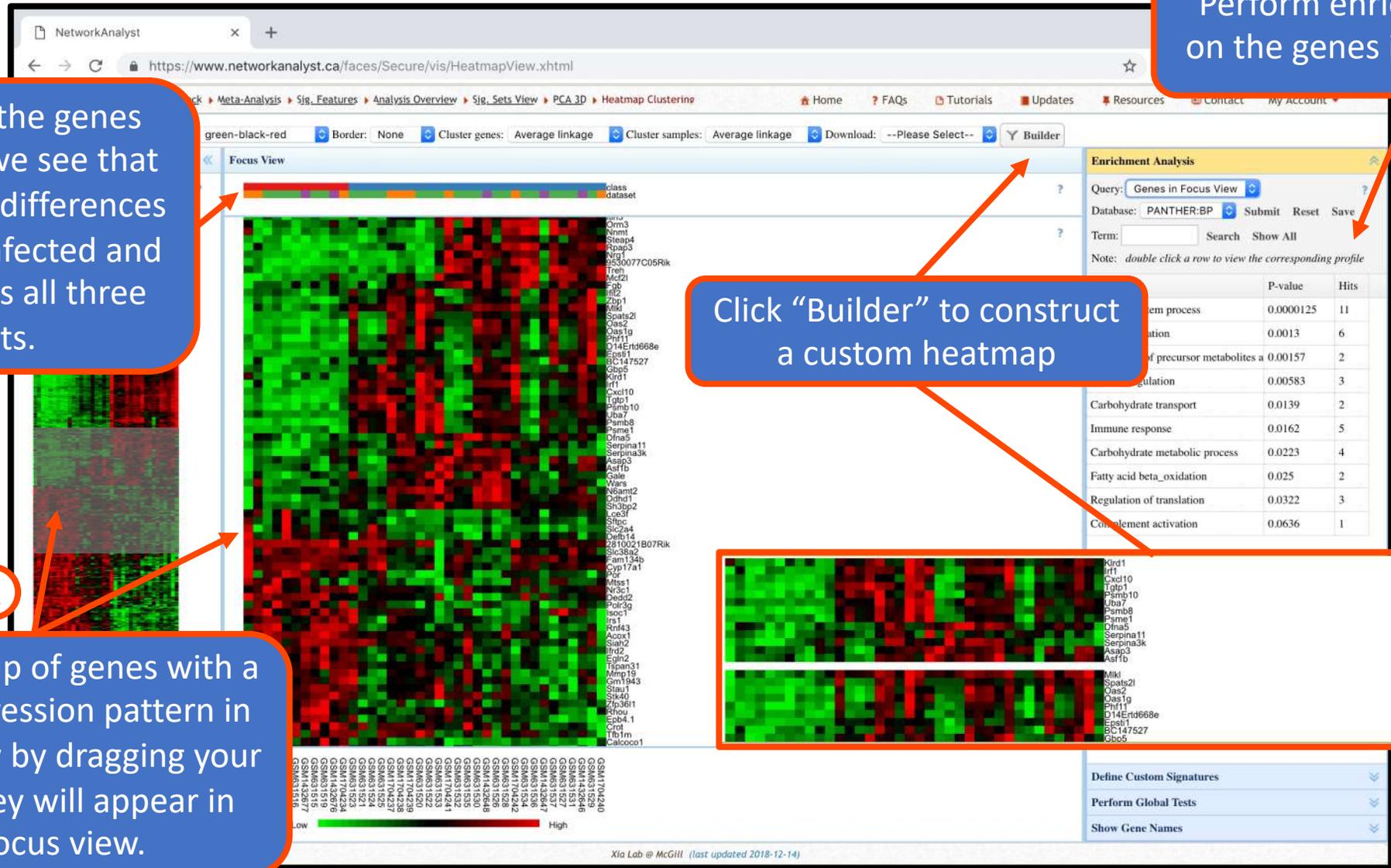
The ORA heatmaps are interactive, allowing users to easily visualize, perform enrichment analysis, and define gene signatures using groups of genes from the heatmap.

Perform enrichment analysis on the genes in the focus view

By clustering the genes and samples, we see that there are clear differences between the infected and controls across all three datasets.

Click "Builder" to construct a custom heatmap

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



1

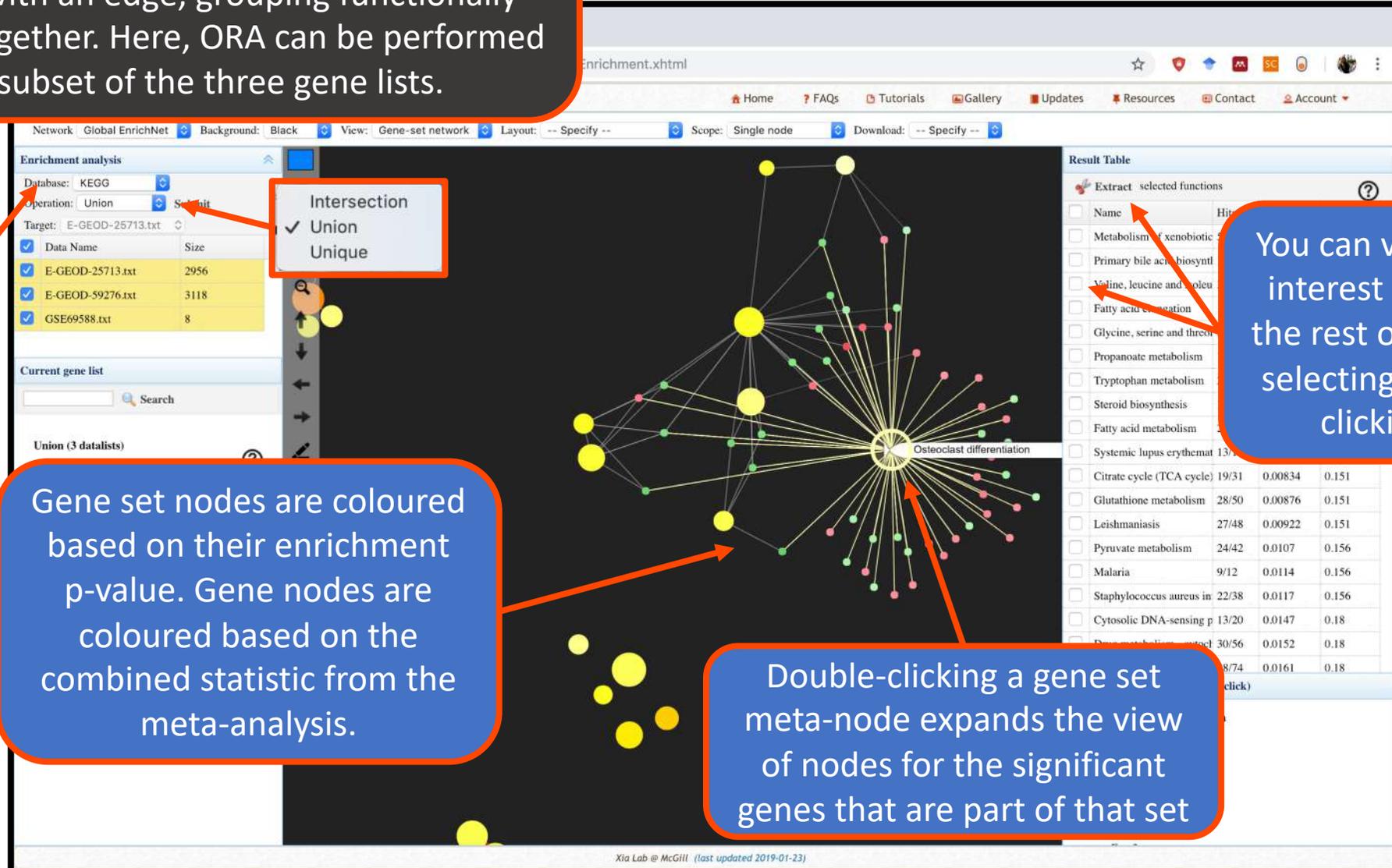
2

# ORA Enrichment Network

Enrichment networks help interpretation of enrichment analysis results since sets with a significant number of overlapping genes are connected with an edge, grouping functionally similar sets together. Here, ORA can be performed on any subset of the three gene lists.

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

- Intersection
- ✓ Union
- Unique



Gene set nodes are coloured based on their enrichment p-value. Gene nodes are coloured based on the combined statistic from the meta-analysis.

Double-clicking a gene set meta-node expands the view of nodes for the significant genes that are part of that set

You can view gene sets of interest separately from the rest of the network by selecting their name and clicking "Extract"

# GSEA for meta-analysis

- A computational method for determining if the expression of a set of genes (biological pathways, etc.) is correlated with phenotypic differences between sample groups
- Incorporates actual gene expression data and so it is able to detect more sensitive differences than simple ORA
- Always requires input genes to be ranked somehow – here the meta-analysis results are used as the ranking metric
- GSEA results using meta-analysis statistics can be thought of as a “gene set-level meta-analysis”
- Refer to the original paper for more details on GSEA:
  - <https://www.pnas.org/content/102/43/15545.short>

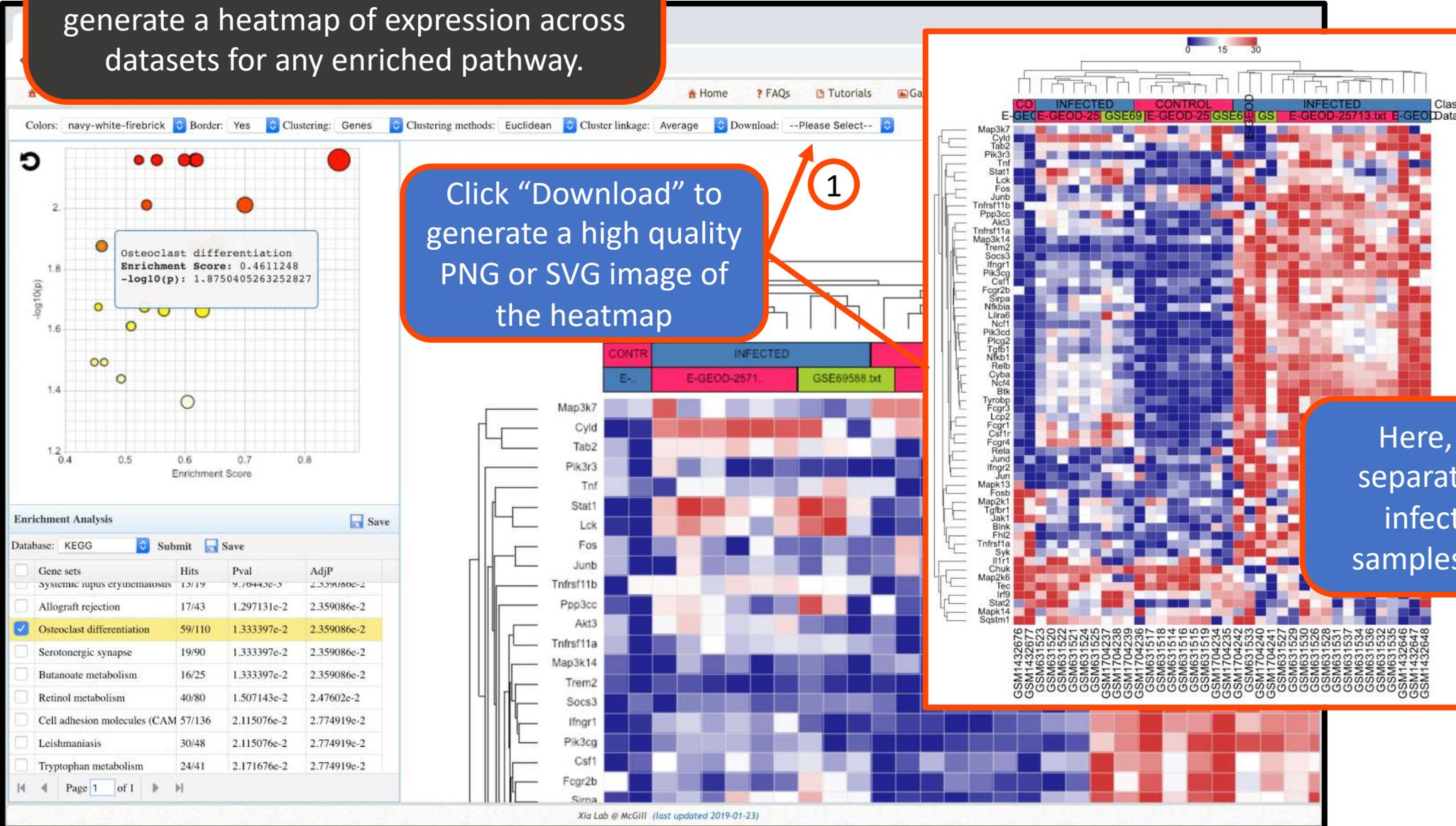
# GSEA Heatmap Clustering

GSEA is performed using the meta-analysis results to rank the genes. The GSEA heatmap tool allows users to generate a heatmap of expression across datasets for any enriched pathway.

Click "Download" to generate a high quality PNG or SVG image of the heatmap

1

Here, we see general separation between the infected and control samples for this gene set



# GSEA Enrichment Network

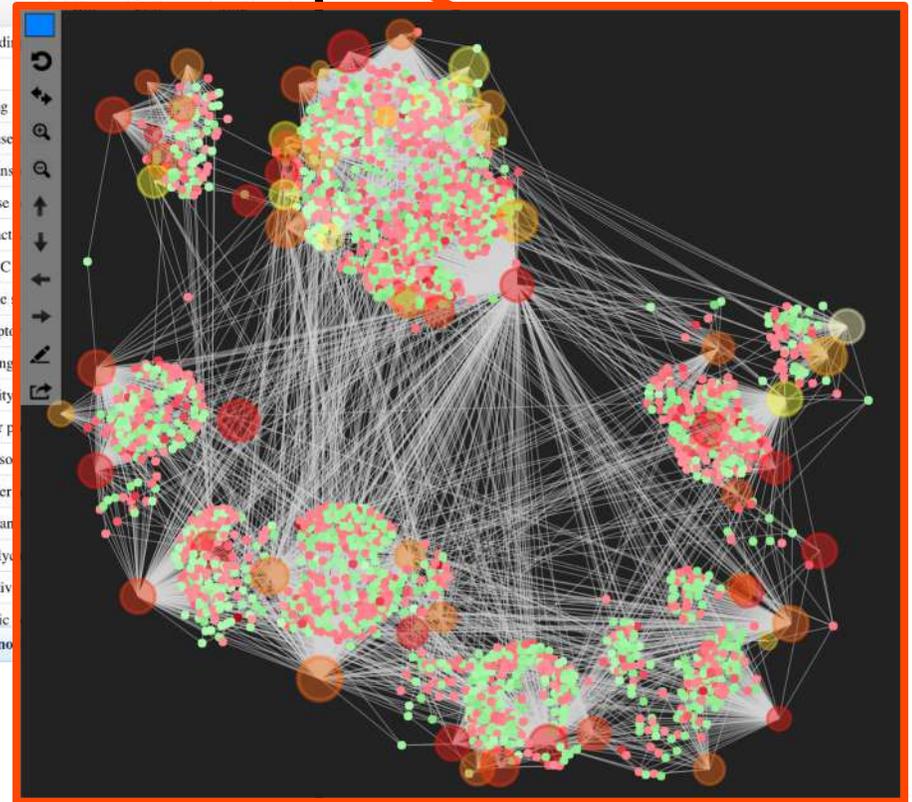
Since GSEA is performed using the meta-analysis results only, set operations between different datasets are not enabled.

Change to "Bipartite network" to view the individual shared genes between gene sets

The screenshot displays the GSEA Enrichment Network interface. The top navigation bar includes 'Upload Data', 'Quality Check', 'Meta-Analysis', 'Sig\_Genes', and 'Analysis Overview'. The main interface is divided into several sections:

- Enrichment analysis:** Database: GO:MF, Operation: Union, Target: meta\_data. A 'Submit' button is visible.
- Current gene list:** A search bar and a table with columns 'Data Name' and 'Size'. The table shows 'meta\_data' with a size of 4997.
- Network visualization:** A large network graph with nodes of various colors (red, orange, yellow, green) and sizes, connected by edges. The nodes represent gene sets, and the edges represent shared genes between them.
- Gene set list:** A list of gene sets with checkboxes for selection. The 'Current selection (no)' is visible at the bottom.

The interface also includes a 'View' dropdown menu set to 'Gene-set network' and a 'Scope' dropdown set to 'Single'. The footer text reads 'Xia Lab @ McGill (last updated 2019-01-23)'.



# Dimension reduction

The screenshot shows the NetworkAnalyst web interface. At the top, there are navigation tabs: Upload Data, Quality Check, Meta-Analysis, Sig. Features, Analysis Overview, Sig. Sets View, PCA 3D, and Heatmap Clustering. Below the navigation, there are controls for the PCA display: a dropdown menu set to 'Based on all data (default)', a 'Download' dropdown set to '--Please Select--', and a slider for 'Number of genes shown' set to 1000. Two callout boxes point to these controls: one pointing to the dropdown menu and another pointing to the slider. The main area contains two 3D PCA plots. The left plot shows data points colored by dataset (E-GEOD-25713.txt in red, E-GEOD-59276.txt in green, GSE69588.txt in blue) and shaped by condition (CONTROL as circles, INFECTED as triangles). The right plot shows a dense cluster of blue circles. A callout box points to a point in this plot. At the bottom, a dark grey box contains text about the utility of 3D PCA plots. The footer of the page reads 'Xia Lab @ McGill (last updated 2018-12-14)'.

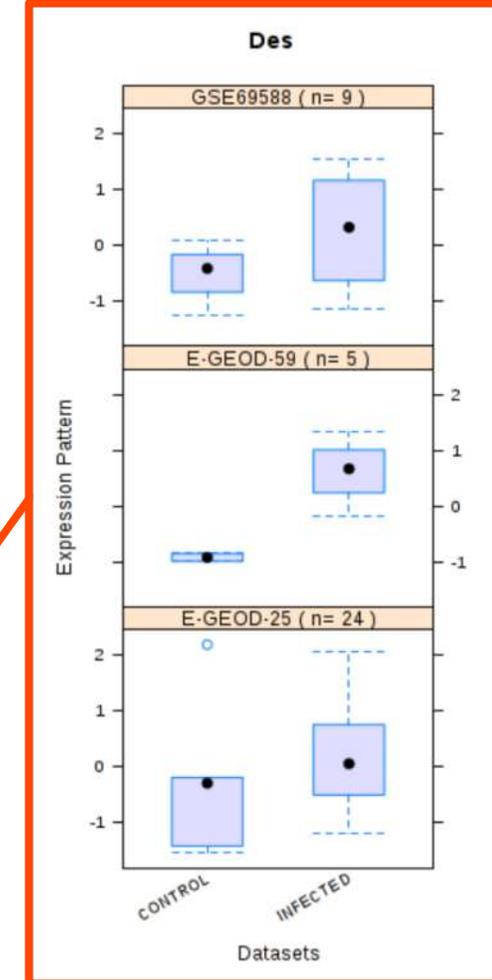
Choose whether to use a subset of genes to generate the 3D PCA plots

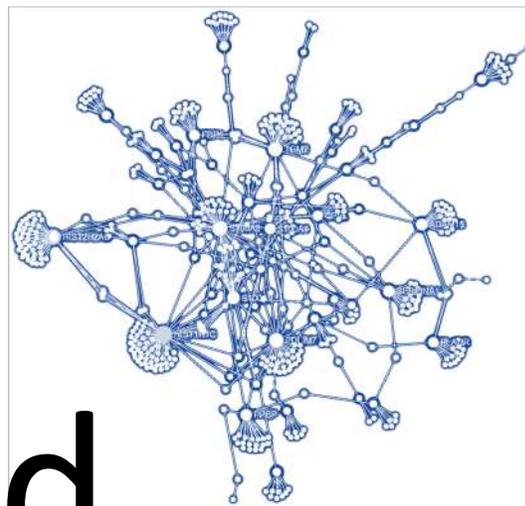
Adjust the number of genes displayed in the loading plot

3D PCA plots are useful for visualizing the variance in whole-transcriptome measures of gene expression across different studies.

Click a gene in the loading plot to see its expression across all datasets

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





# The End

*For more information, visit the **FAQs**, **Tutorials**, **Resources**,  
and **Contact** pages on [www.networkanalyst.ca](http://www.networkanalyst.ca)*