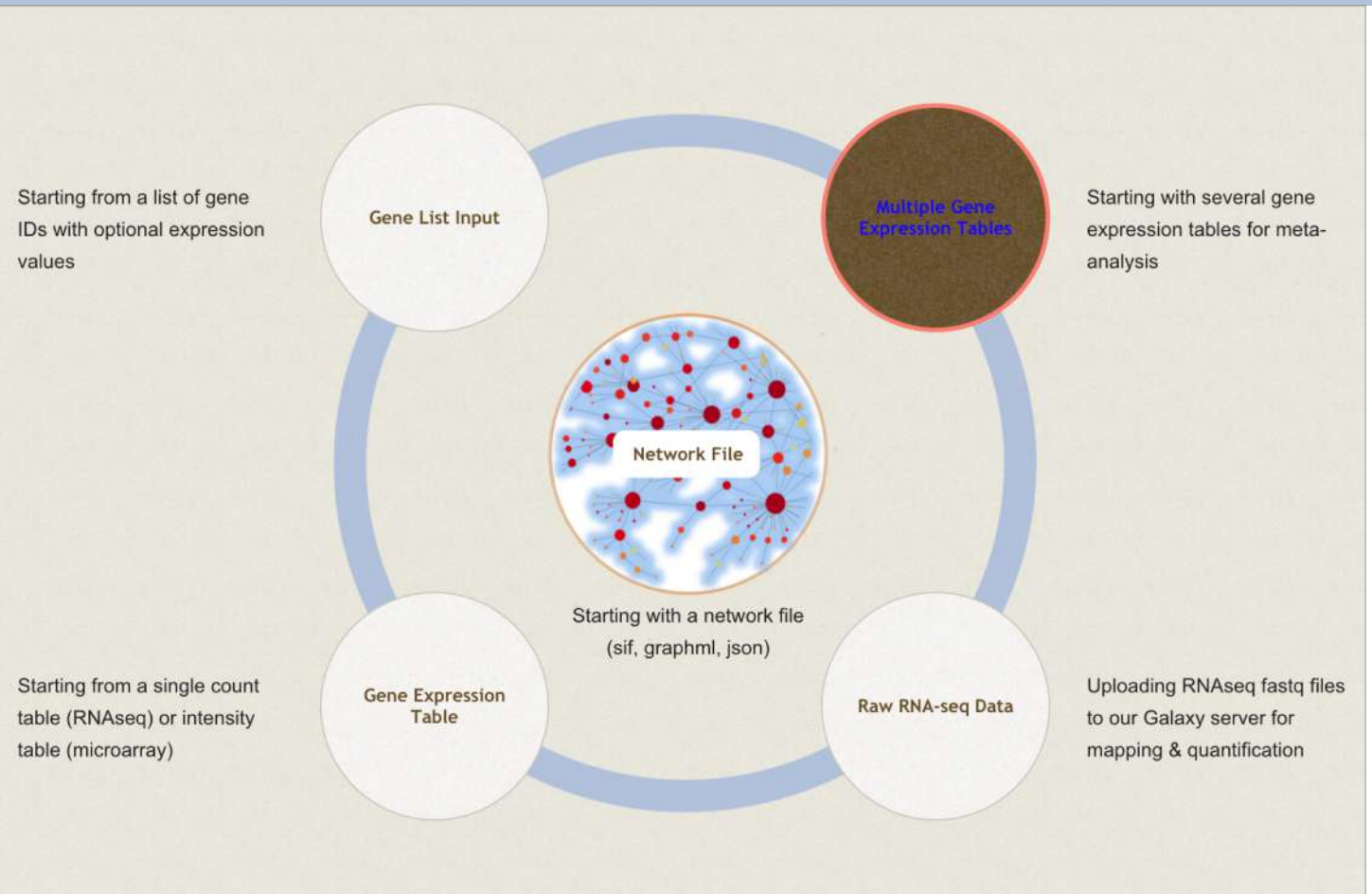
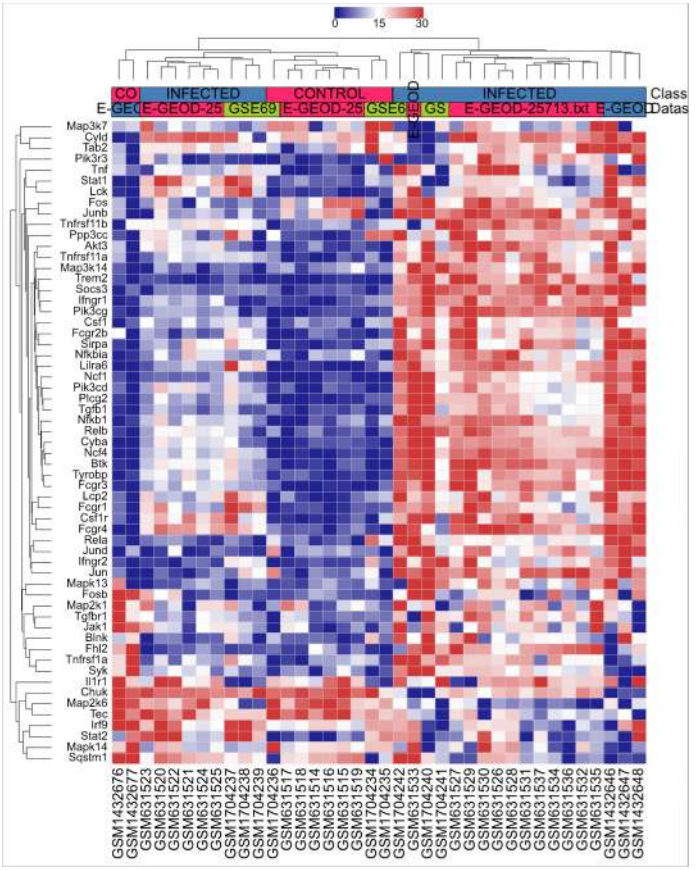
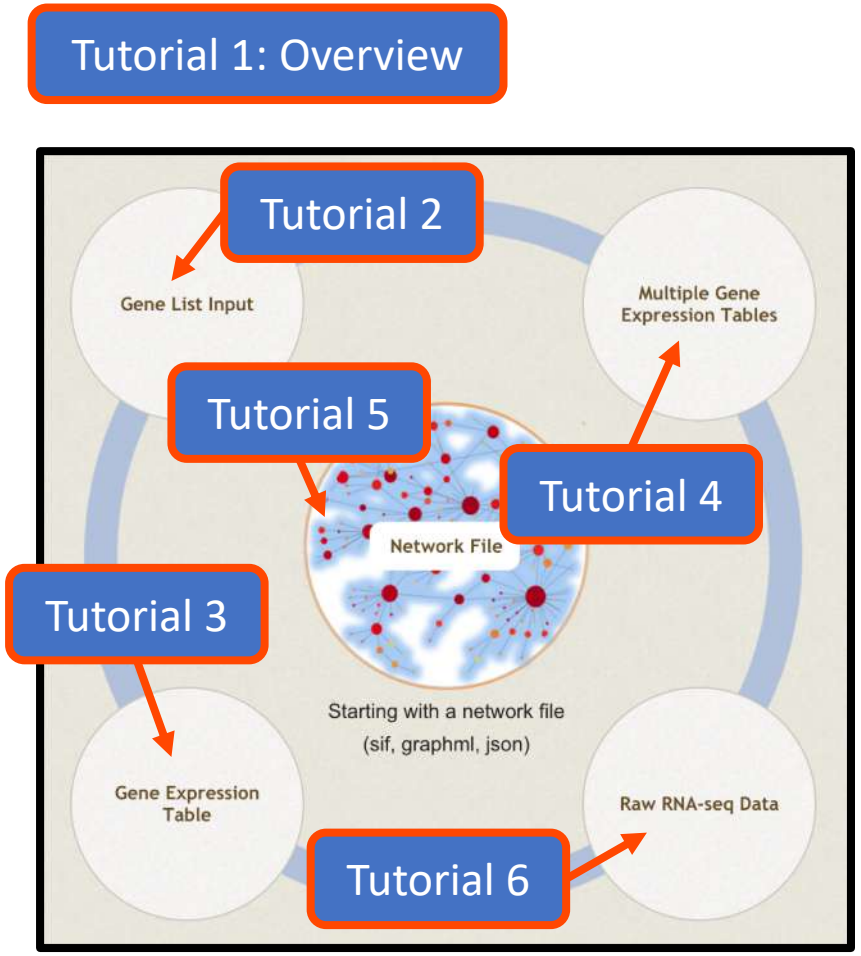
Tutorial 4: multiple gene expression tables





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



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.)

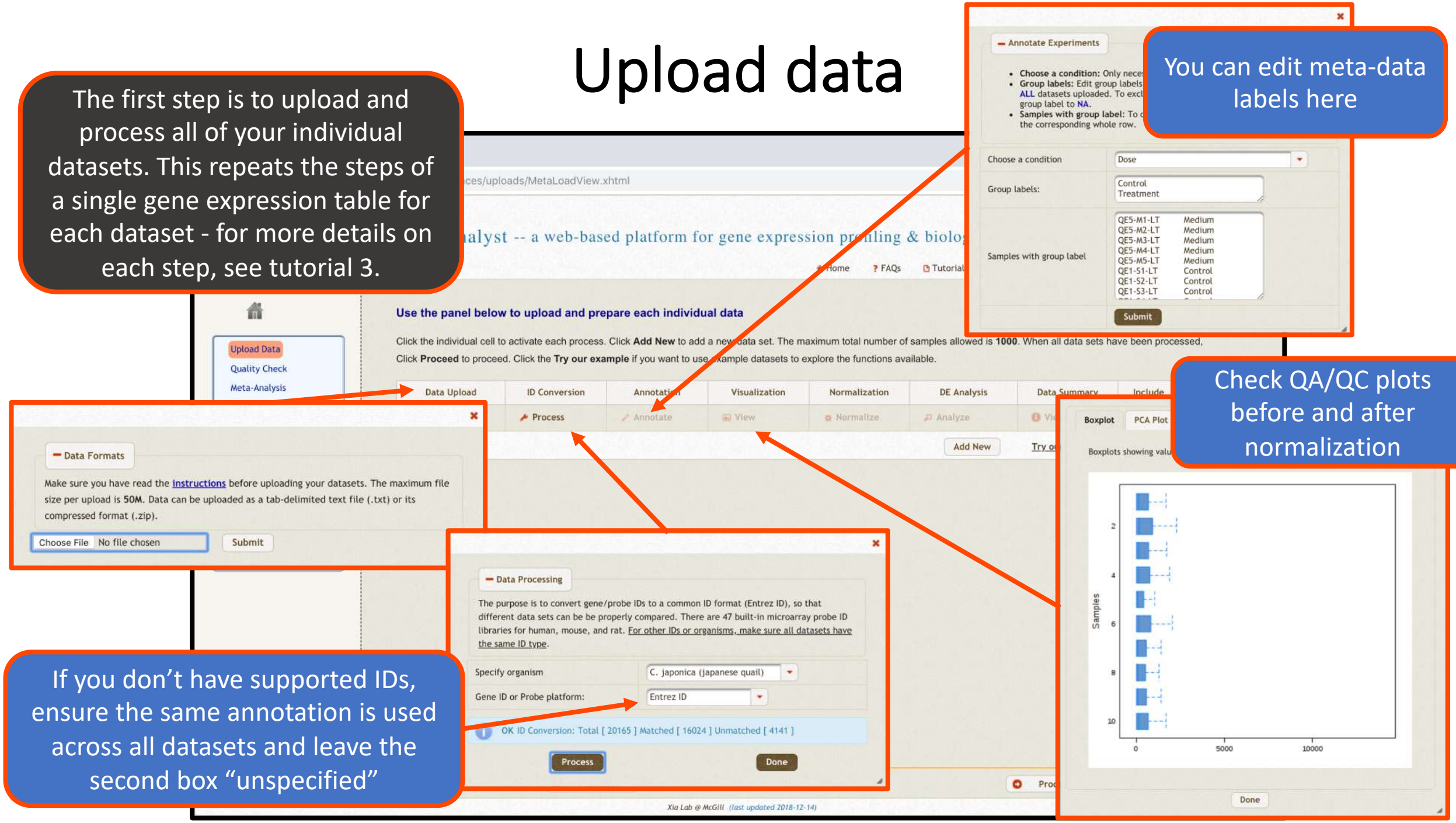
Upload data

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.

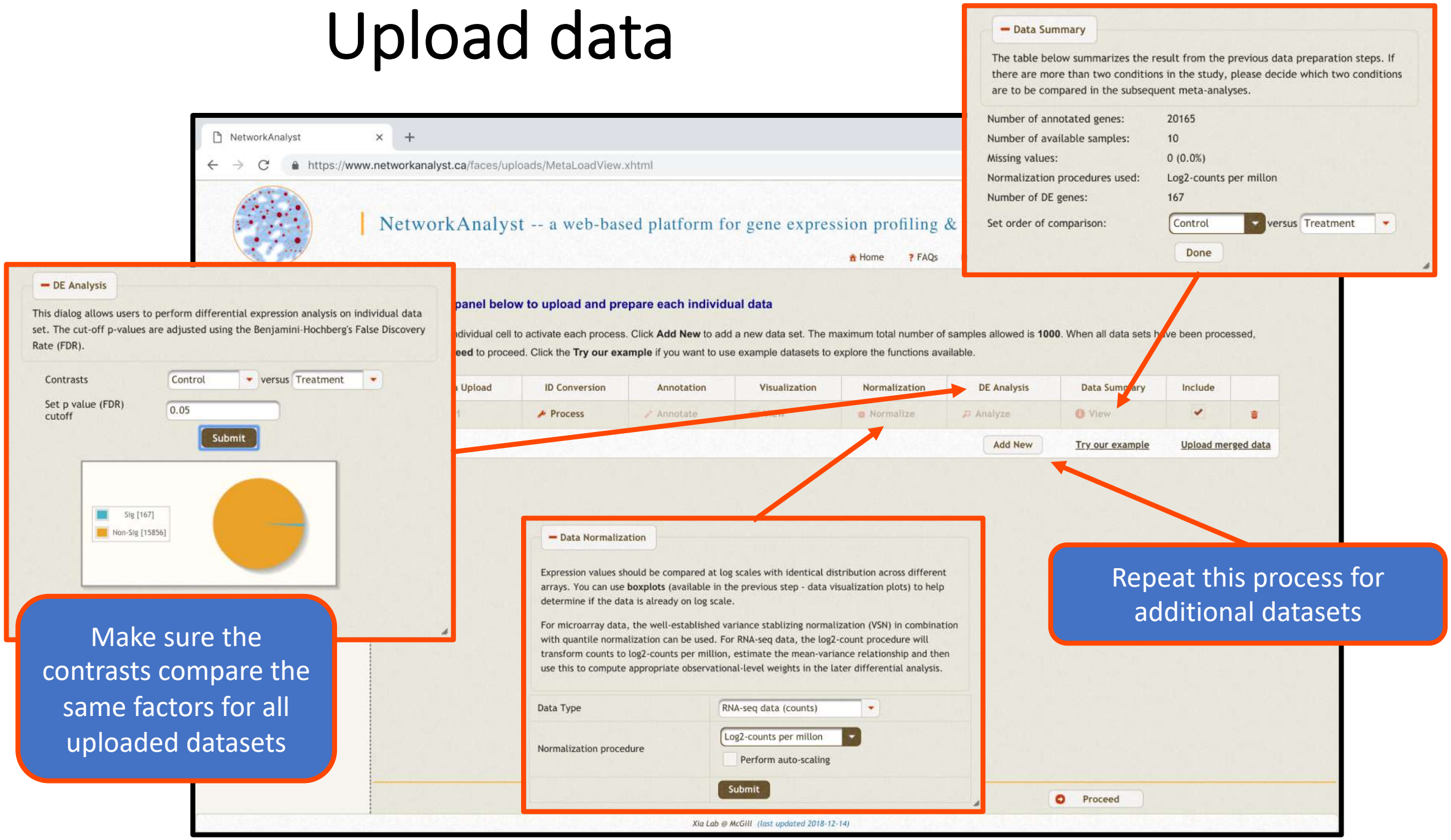
You can edit meta-data labels here

Check QA/QC plots before and after normalization

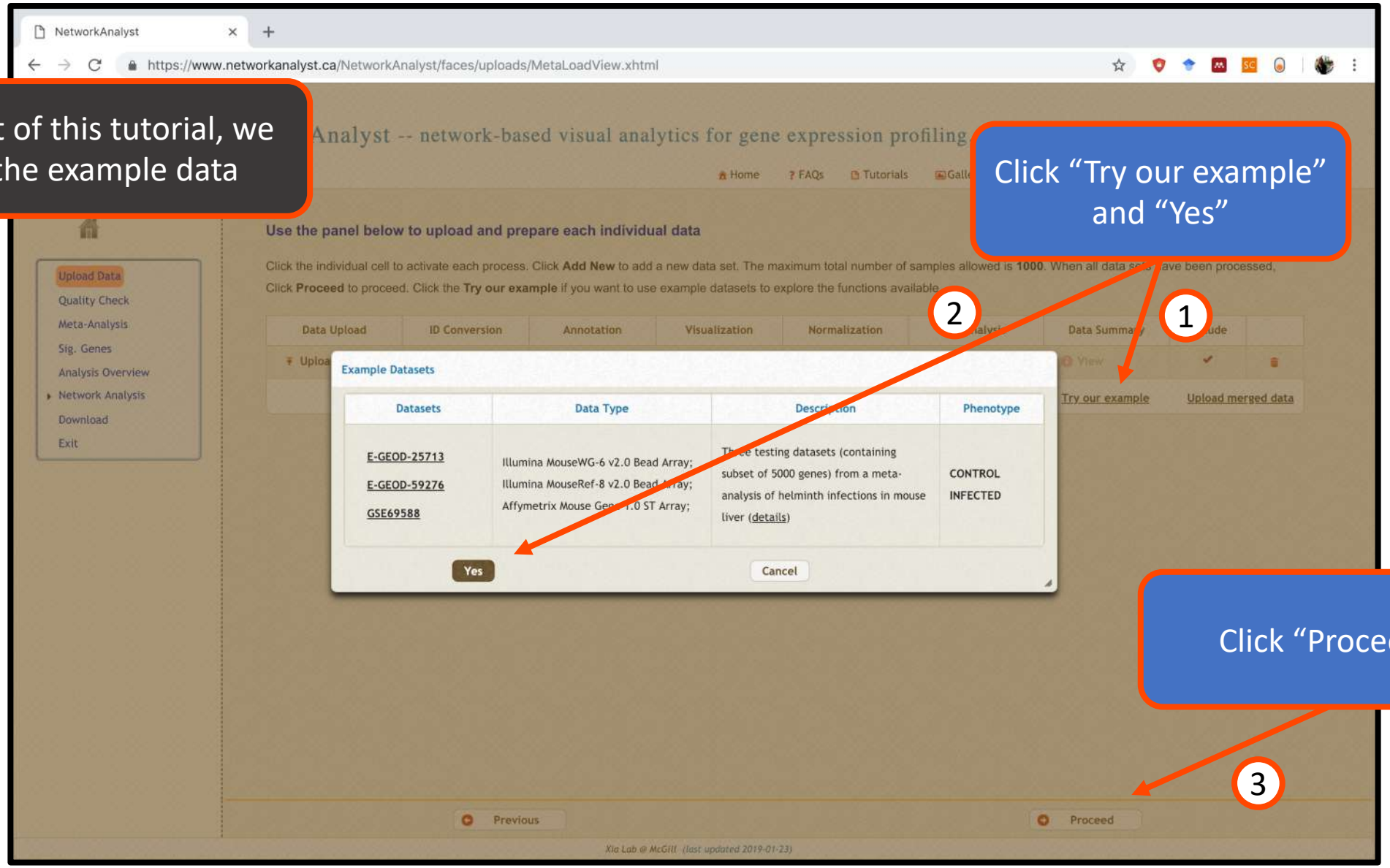
If you don't have supported IDs, ensure the same annotation is used across all datasets and leave the second box "unspecified"



Upload data


NetworkAnalyst
https://www.networkanalyst.ca/faces/uploads/MetaLoadView.xhtml
NetworkAnalyst -- a web-based platform for gene expression profiling &
Home
FAQs
Data Summary
The table below summarizes the result from the previous data preparation steps. If there are more than two conditions in the study, please decide which two conditions are to be compared in the subsequent meta-analyses.
Number of annotated genes: 20165
Number of available samples: 10
Missing values: 0 (0.0%)
Normalization procedures used: Log2-counts per millon
Number of DE genes: 167
Set order of comparison: Control versus Treatment
Done
DE Analysis
This dialog allows users to perform differential expression analysis on individual data set. The cut-off p-values are adjusted using the Benjamini-Hochberg's False Discovery Rate (FDR).
Contrasts Control versus Treatment
Set p value (FDR) cutoff 0.05
Submit
Sig [167]
Non-Sig [15856]
panel below to upload and prepare each individual data
Upload
ID Conversion
Annotation
Visualization
Normalization
DE Analysis
Data Summary
Include
Process
Annotate
Normalize
Analyze
View
Add New
Try our example
Upload merged data
Data Normalization
Expression values should be compared at log scales with identical distribution across different arrays. You can use boxplots (available in the previous step - data visualization plots) to help determine if the data is already on log scale.
For microarray data, the well-established variance stablizing normalization (VSN) in combination with quantile normalization can be used. For RNA-seq data, the log2-count procedure will transform counts to log2-counts per million, estimate the mean-variance relationship and then use this to compute appropriate observational-level weights in the later differential analysis.
Data Type RNA-seq data (counts)
Normalization procedure Log2-counts per millon
Perform auto-scaling
Submit
Proceed
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”

1

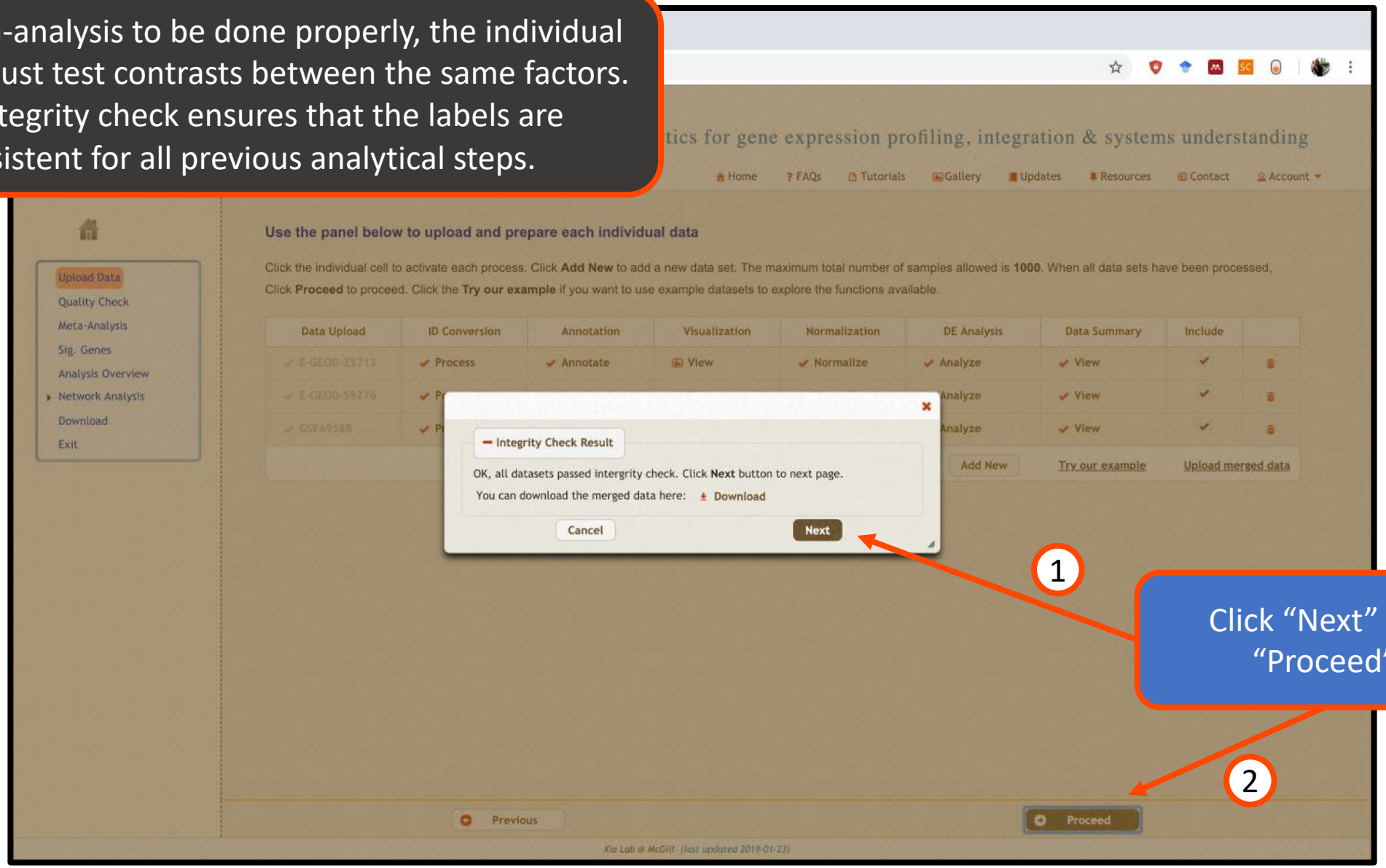
2

Click “Proceed”

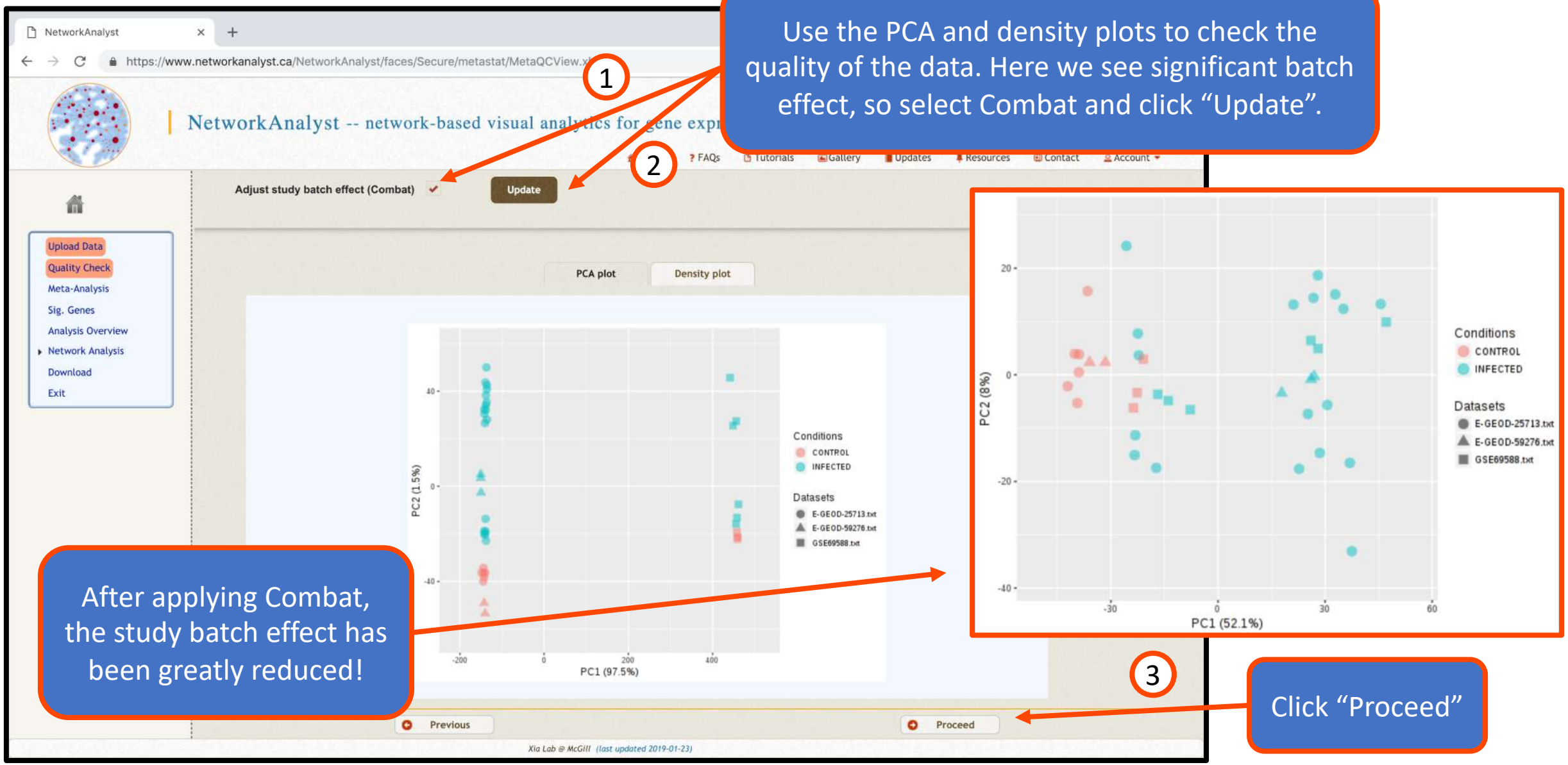
3

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.



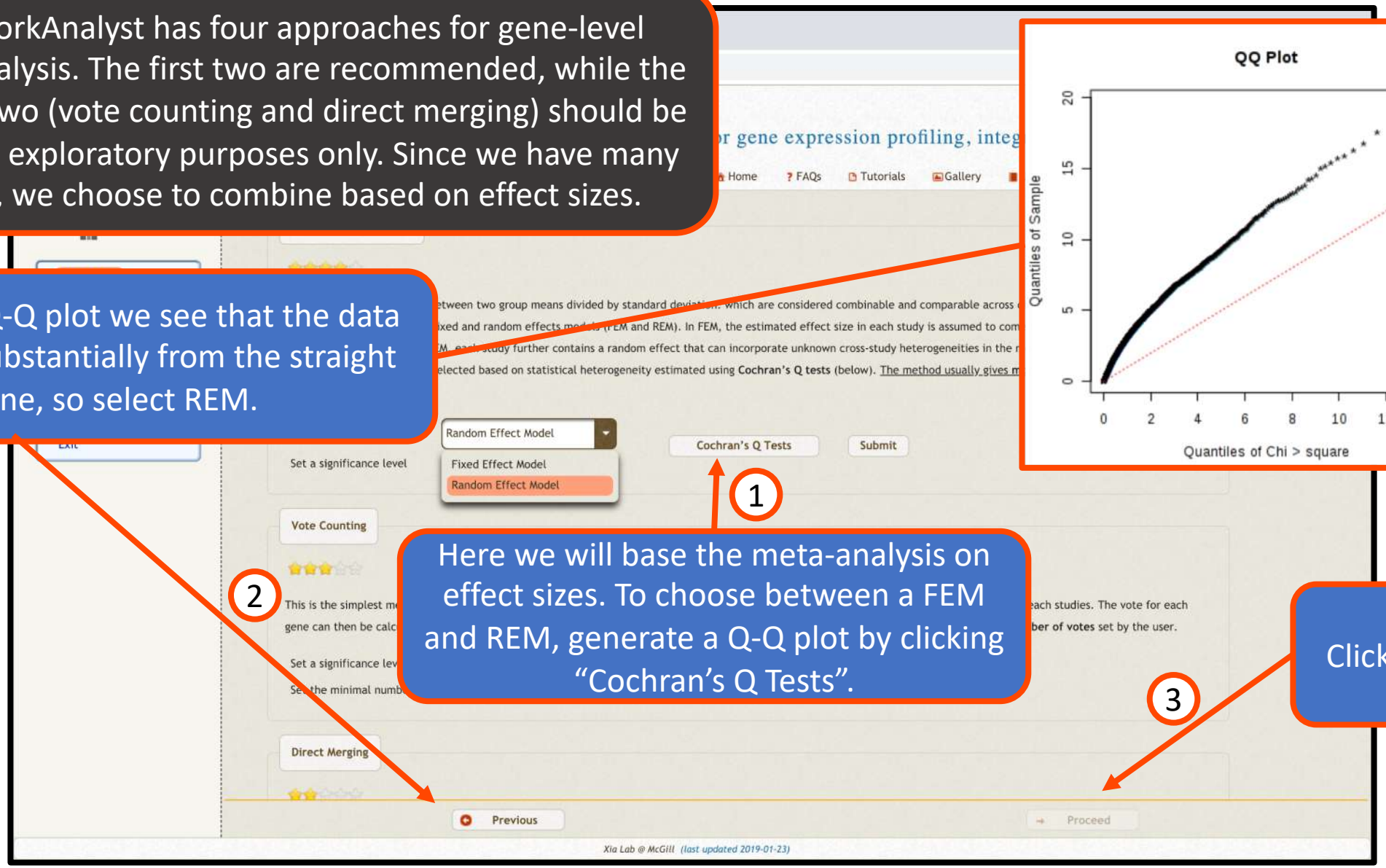
View raw data and correct batch effect



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.



1

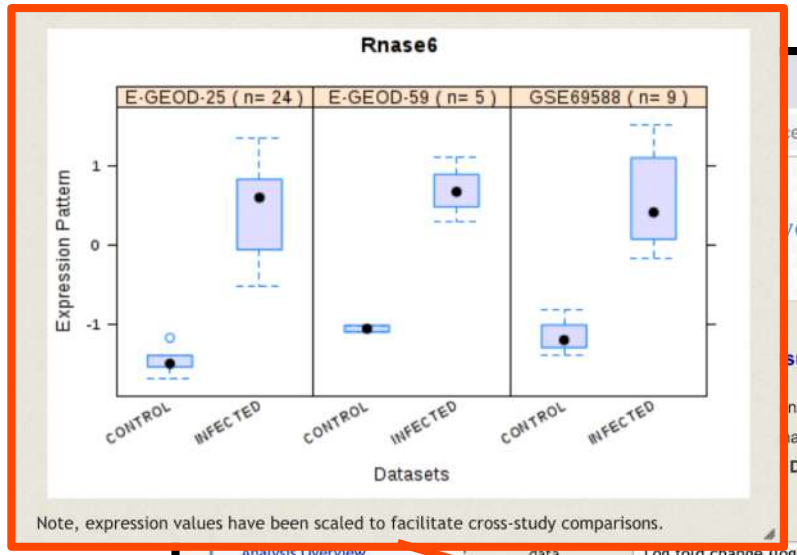
Here we will base the meta-analysis on effect sizes. To choose between a FEM and REM, generate a Q-Q plot by clicking "Cochran's Q Tests".

2

3

Click "Proceed"

View results of meta-analysis



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

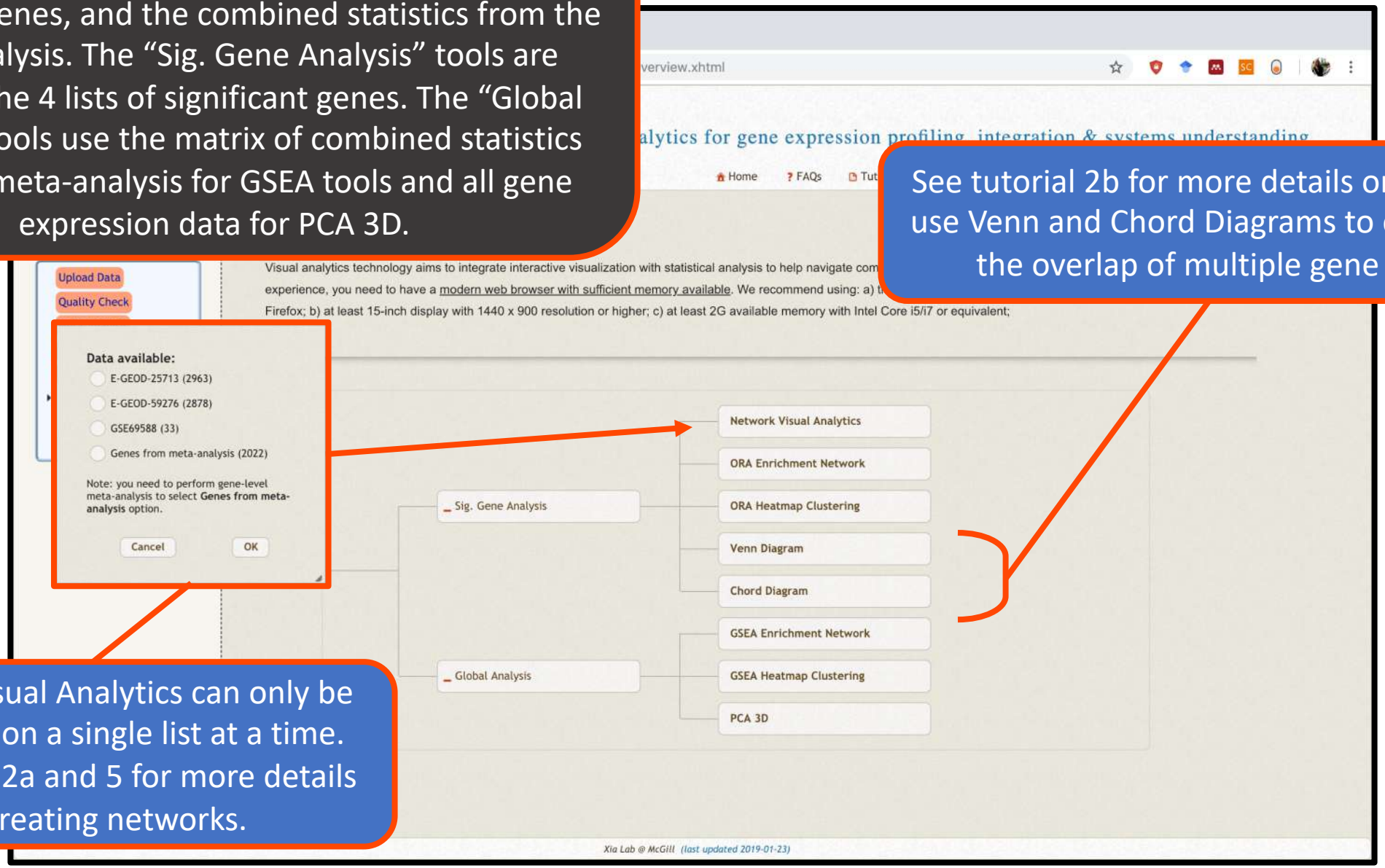
Click on the picture icon to see a boxplot of a specific gene across datasets

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

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.

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

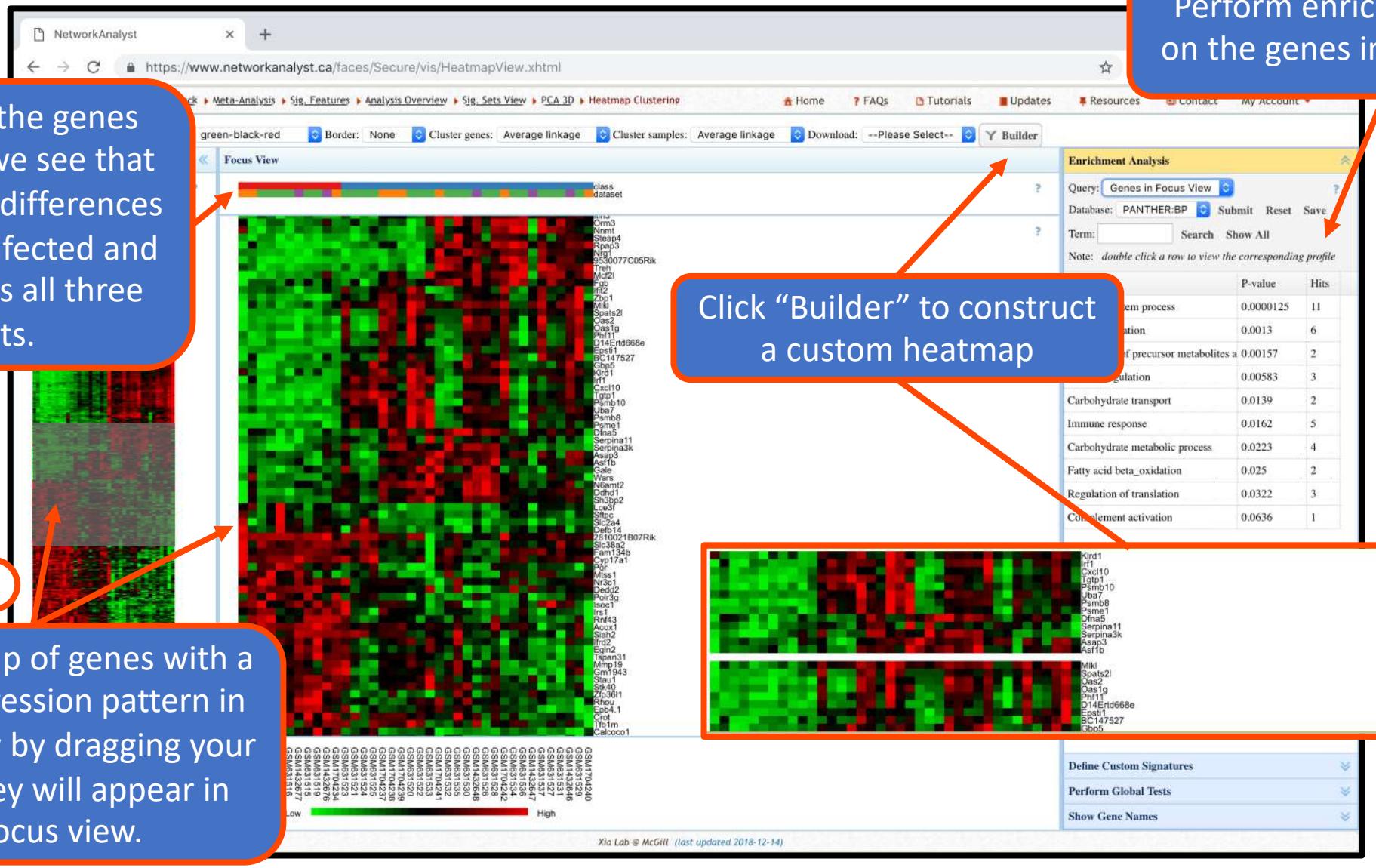
Perform enrichment analysis on the genes in the focus view

2

Click “Builder” to construct a custom heatmap

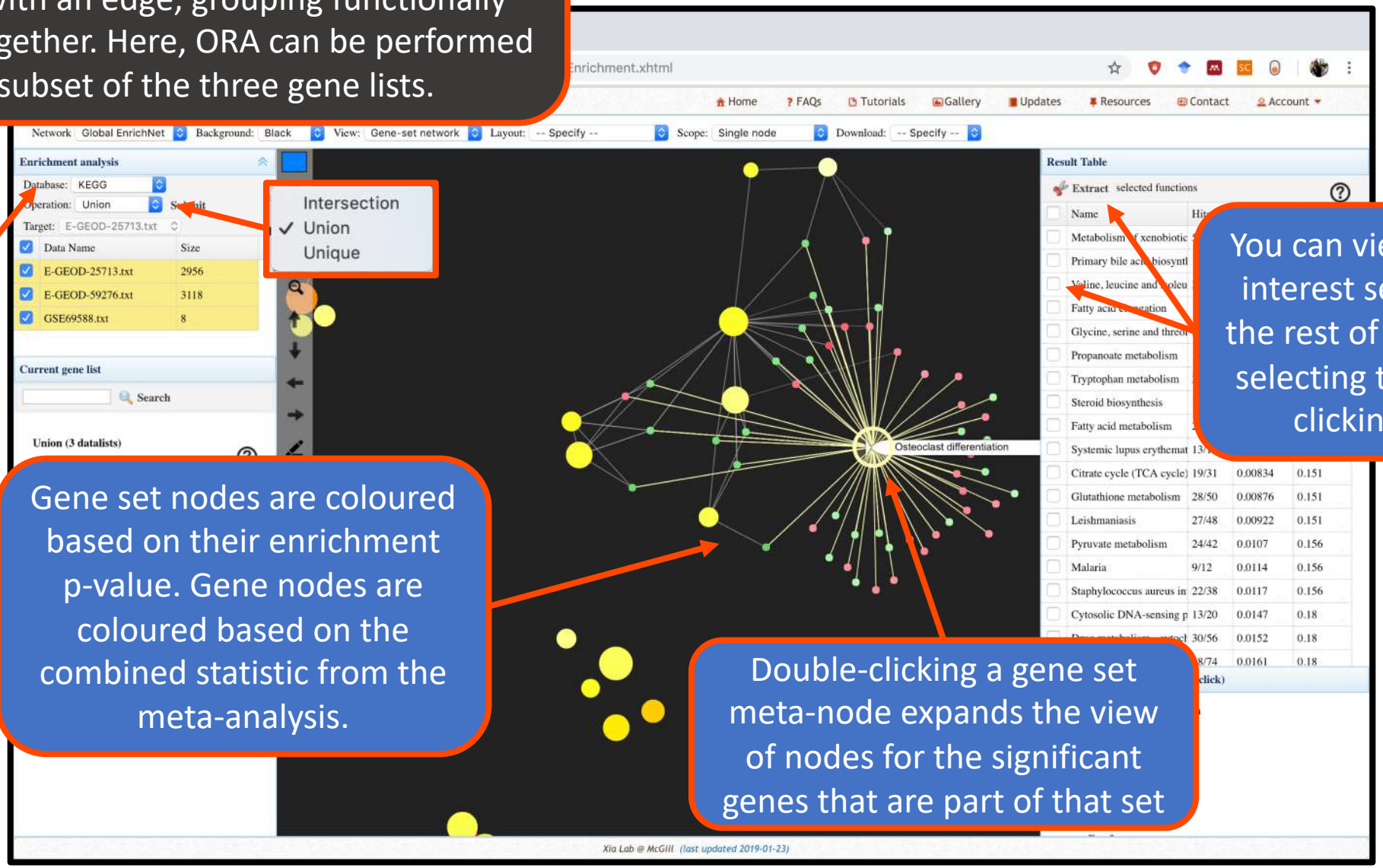
1

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



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.

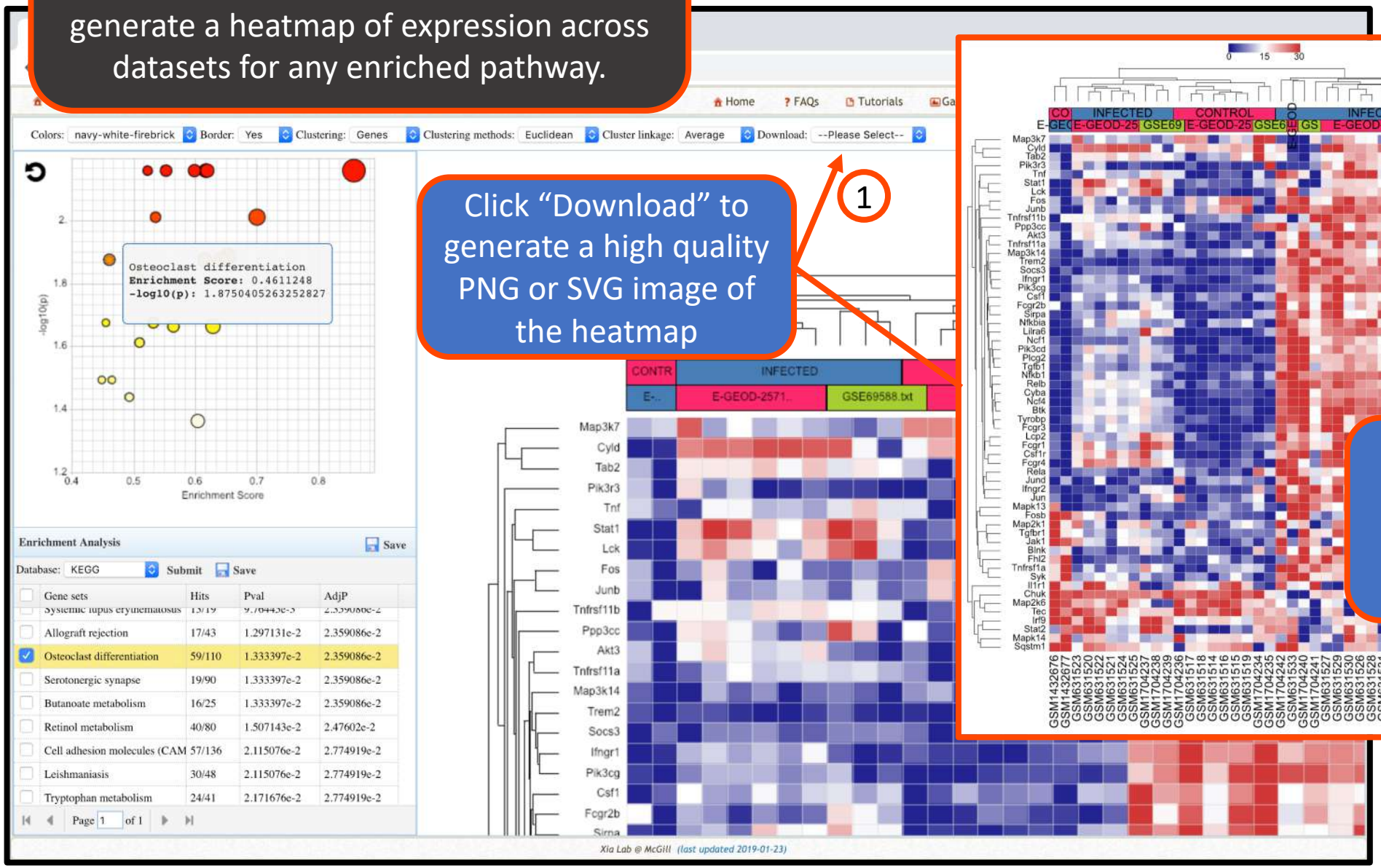


GSEA for meta-analysis

- 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

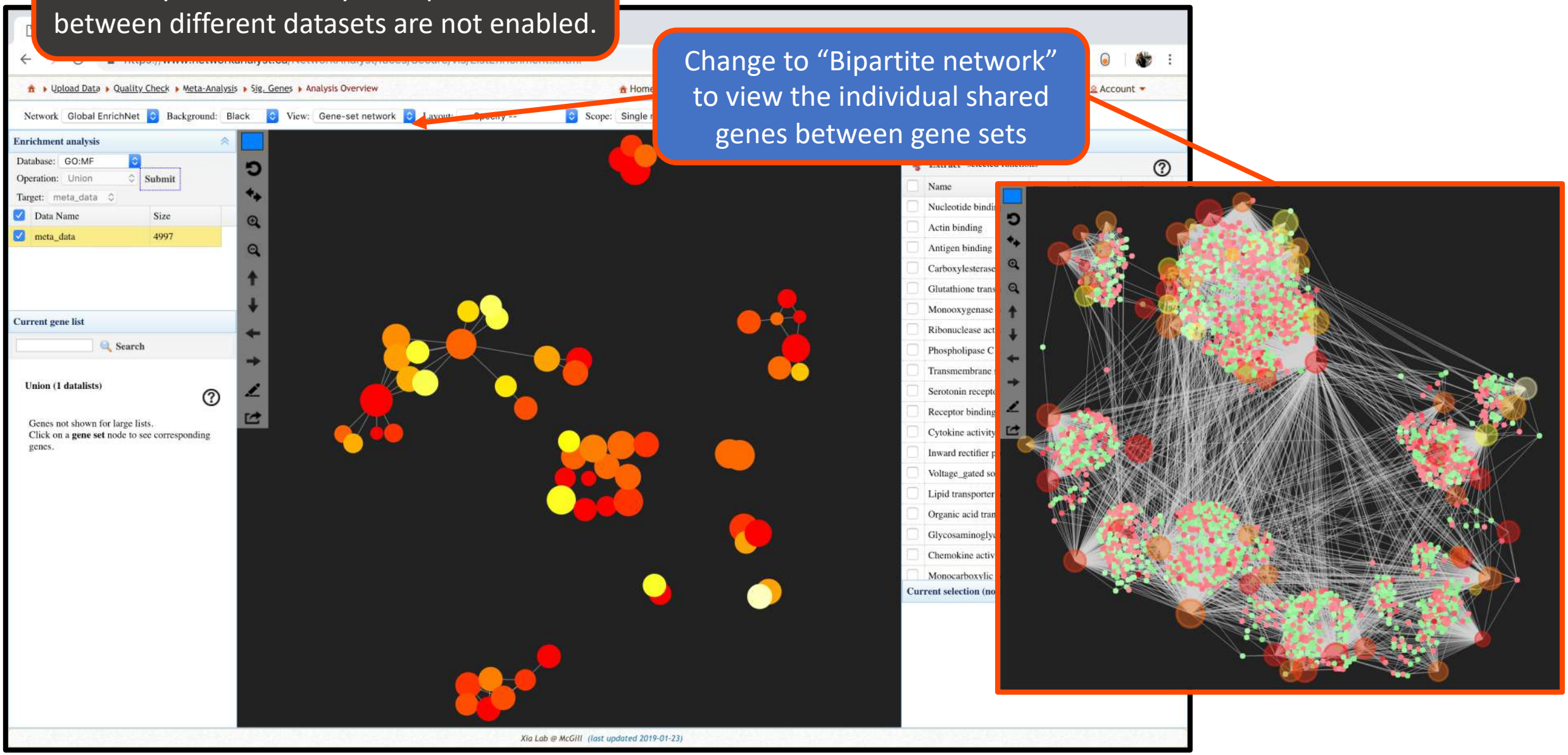
1

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

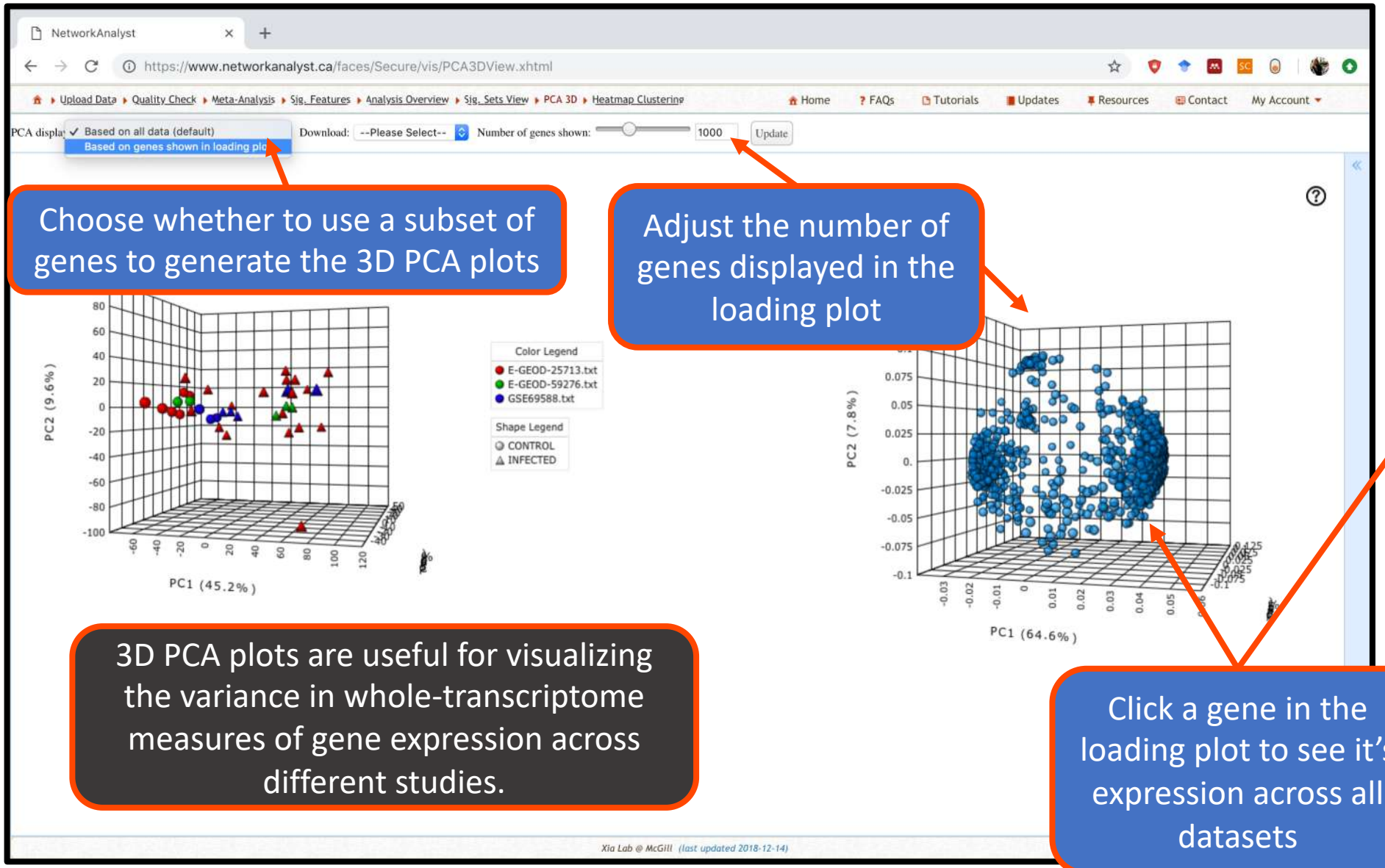
GSEA Enrichment Network

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



Dimension reduction

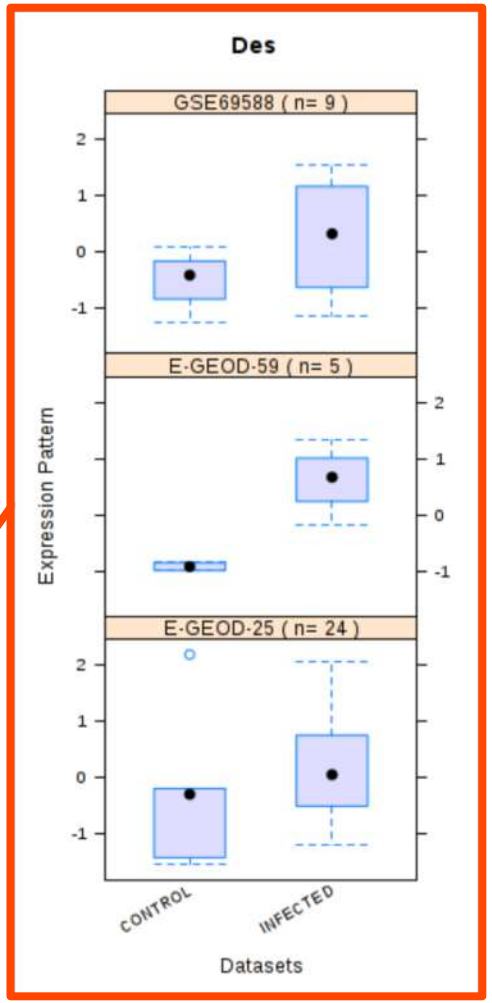


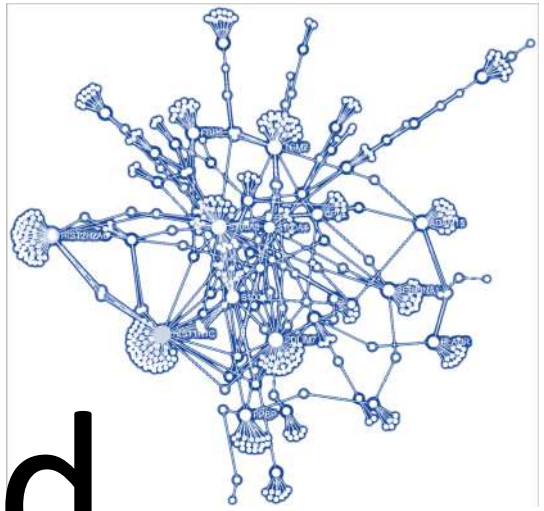
Choose whether to use a subset of genes to generate the 3D PCA plots

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 it's expression across all datasets





The End

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