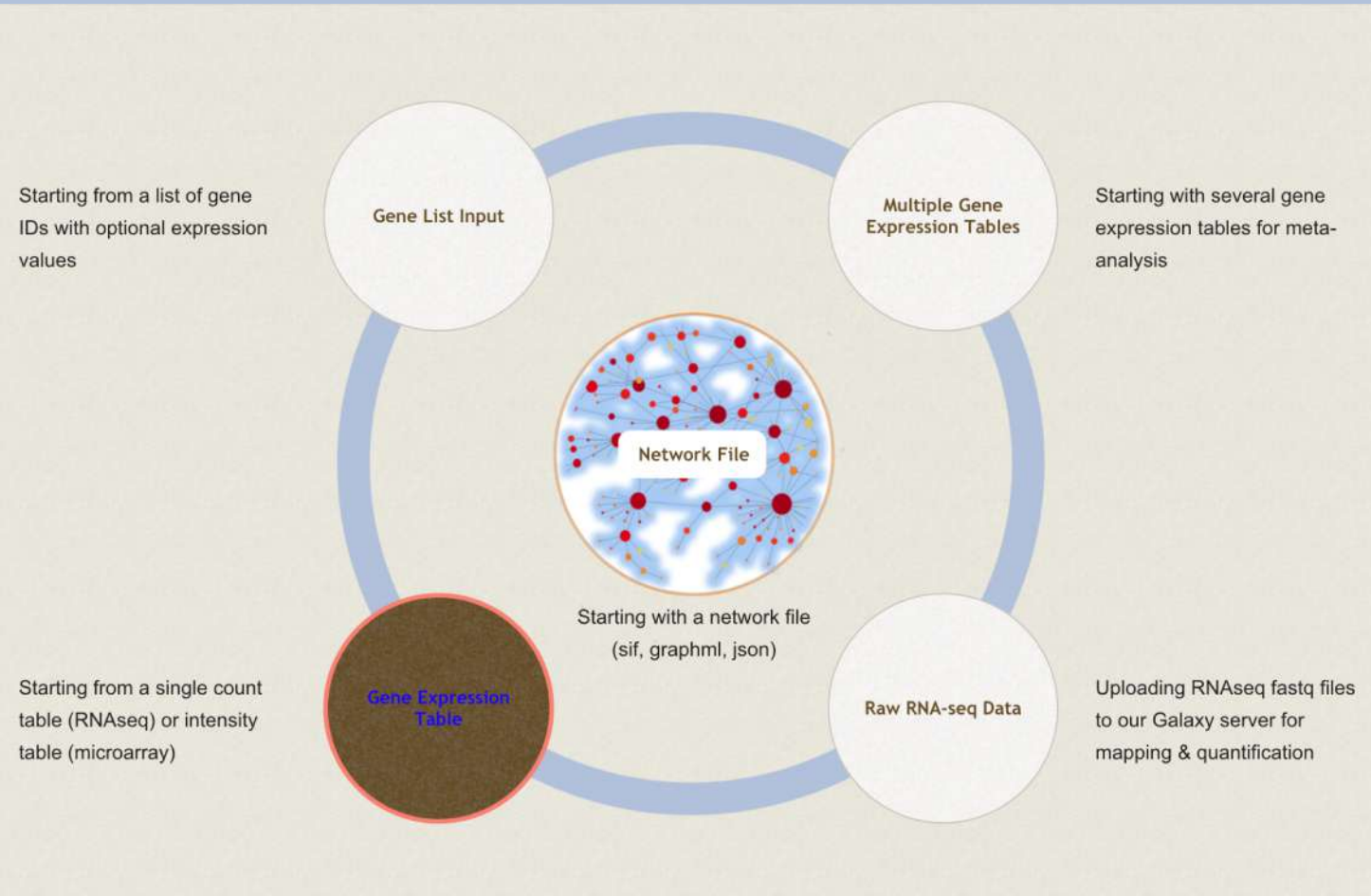
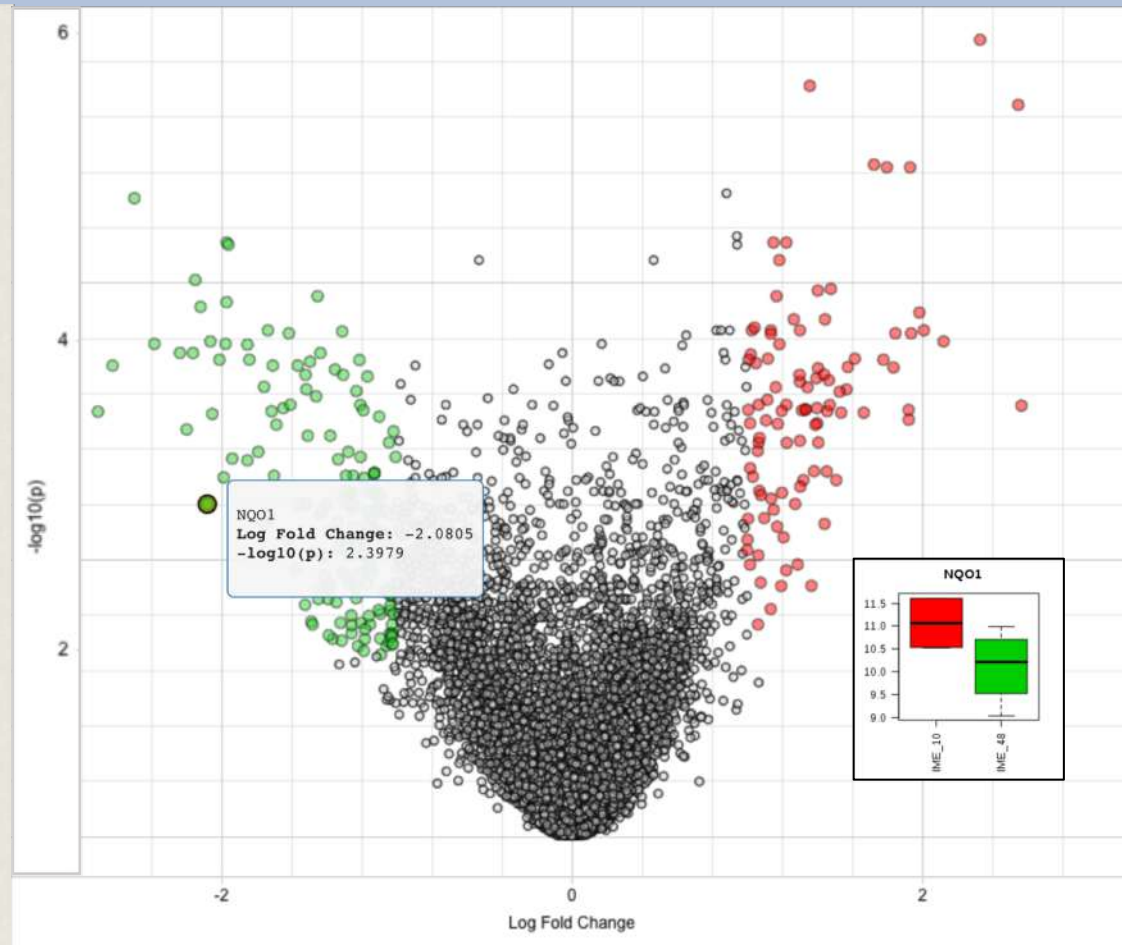
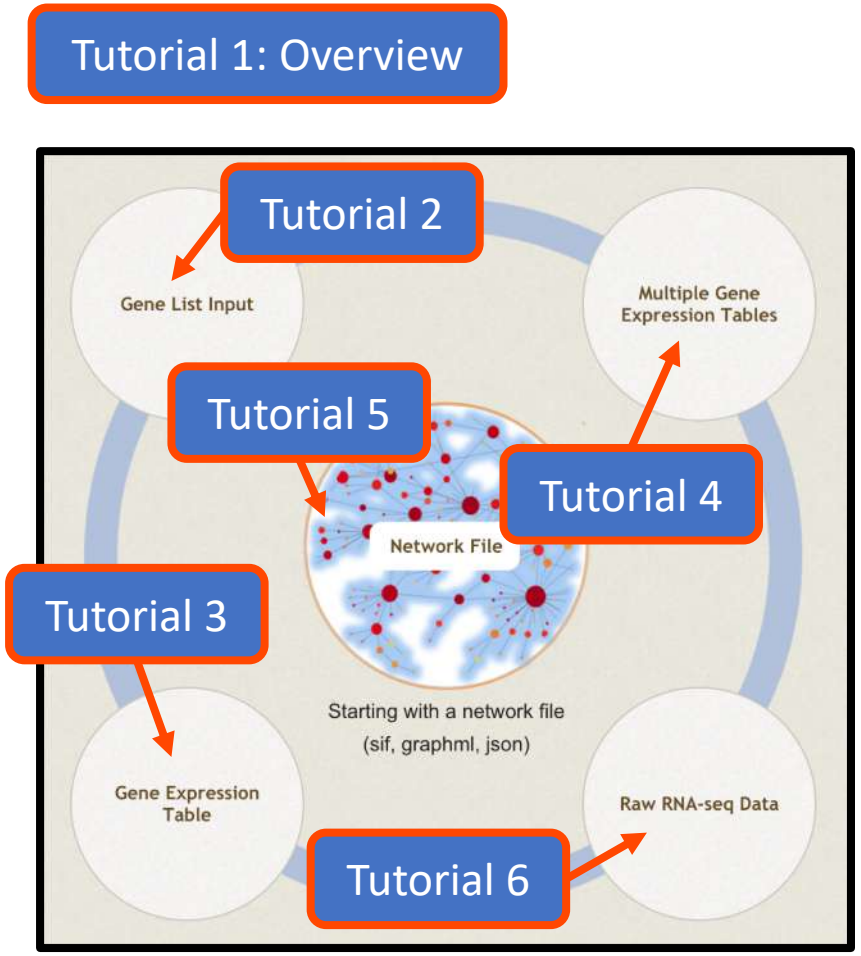
Tutorial 3: gene expression table





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



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

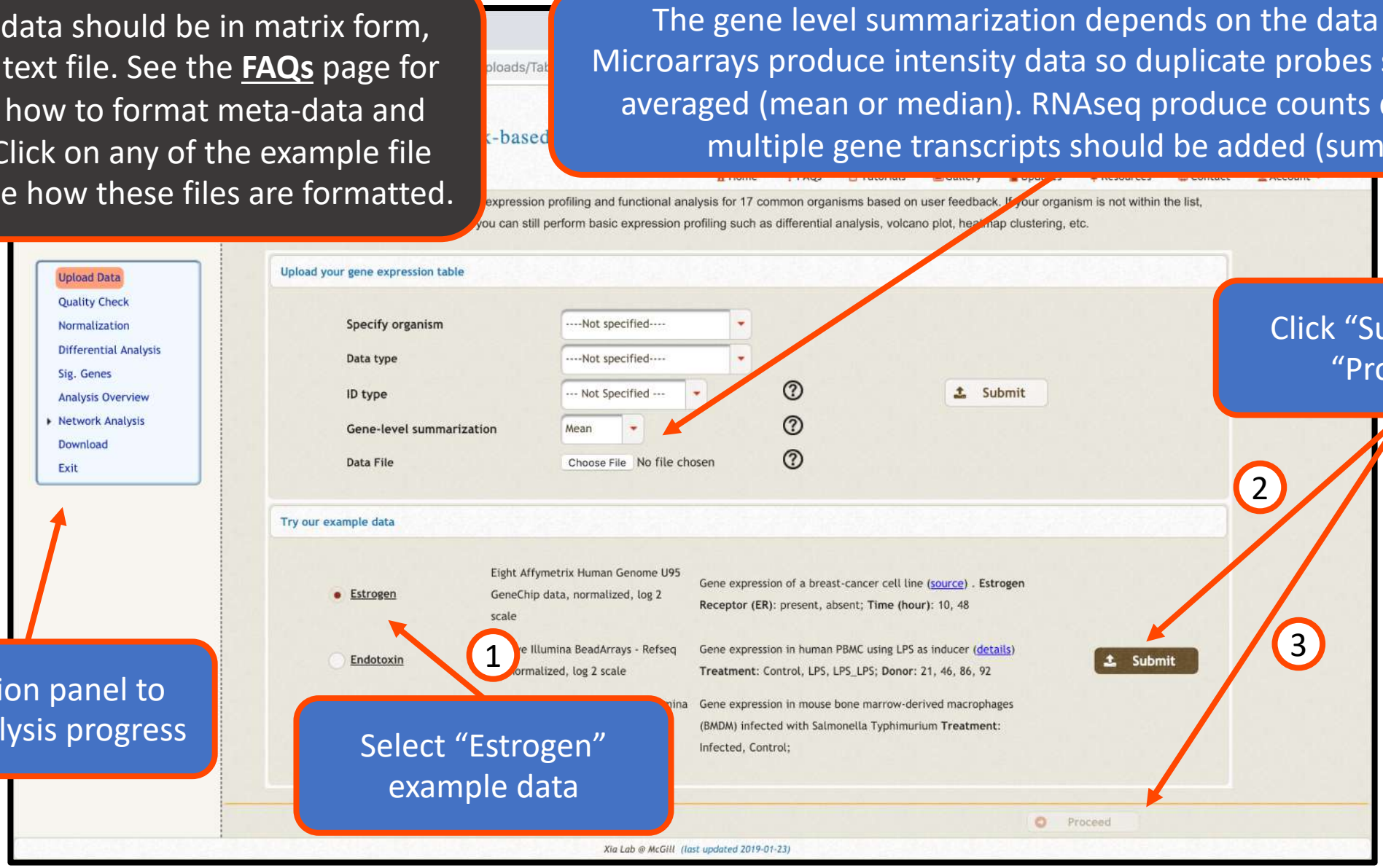
Goals for this tutorial

- 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”

2

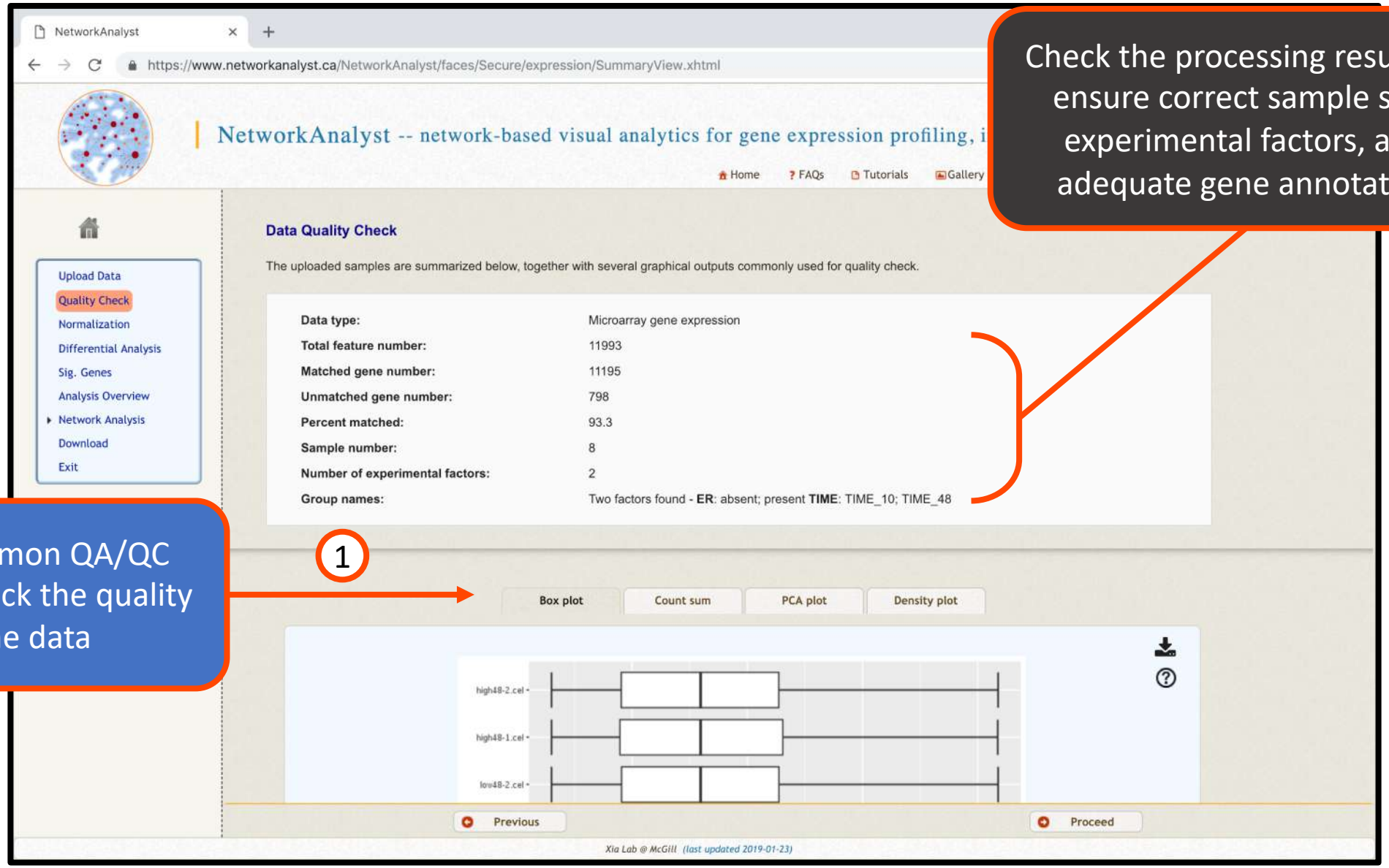
3

1

Navigation panel to track analysis progress

Select “Estrogen” example data

View processing results

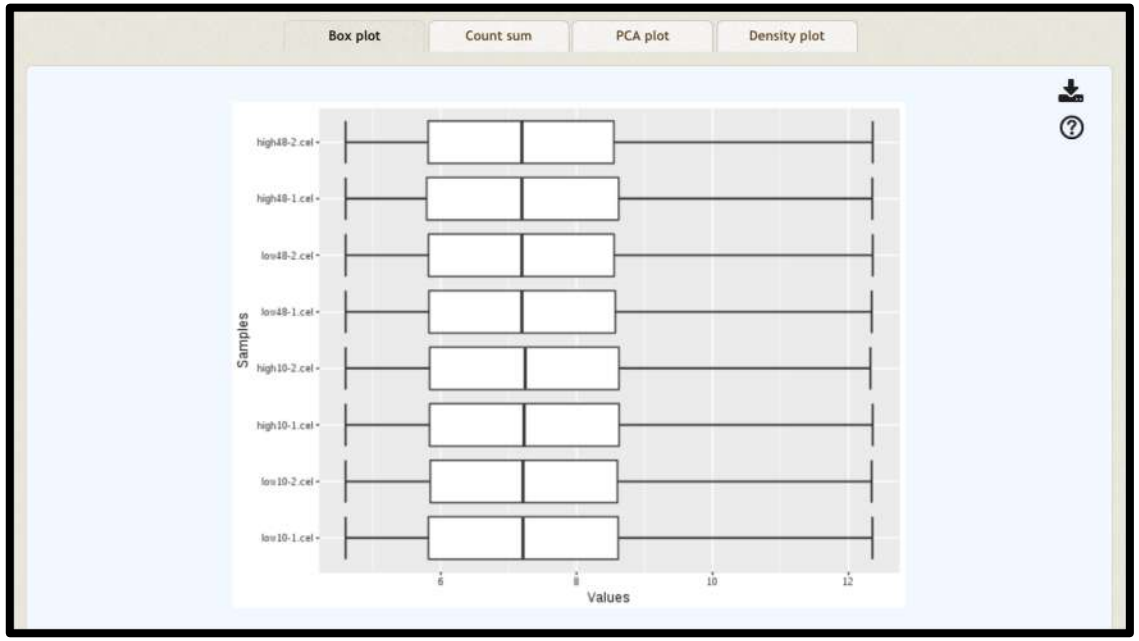


Check the processing results to ensure correct sample size, experimental factors, and adequate gene annotation

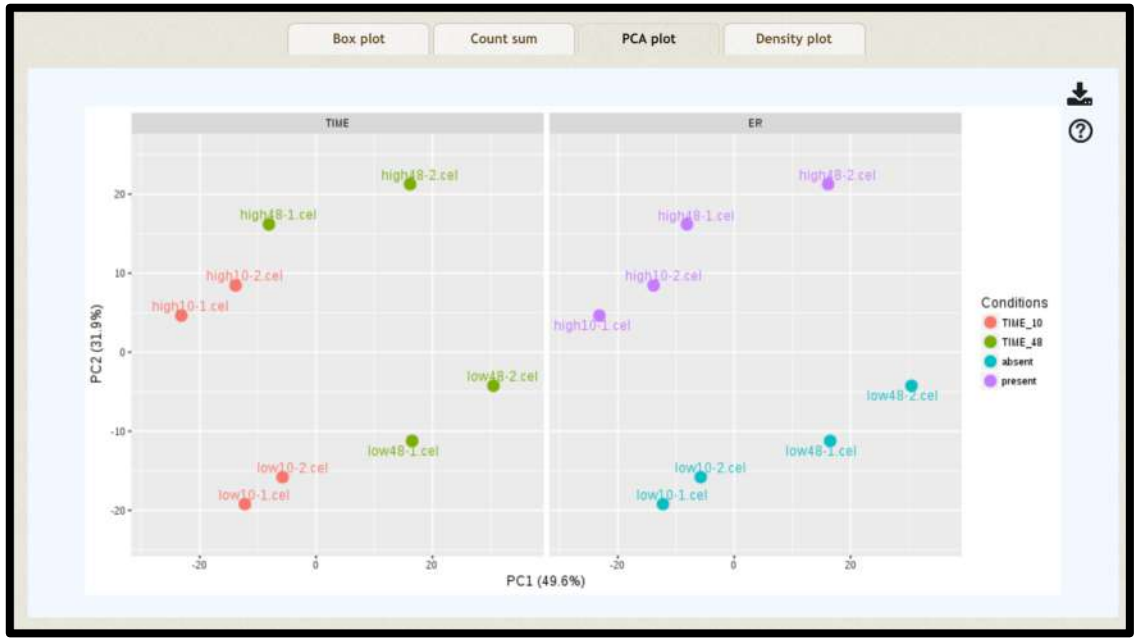
View common QA/QC plots to check the quality of the data

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.

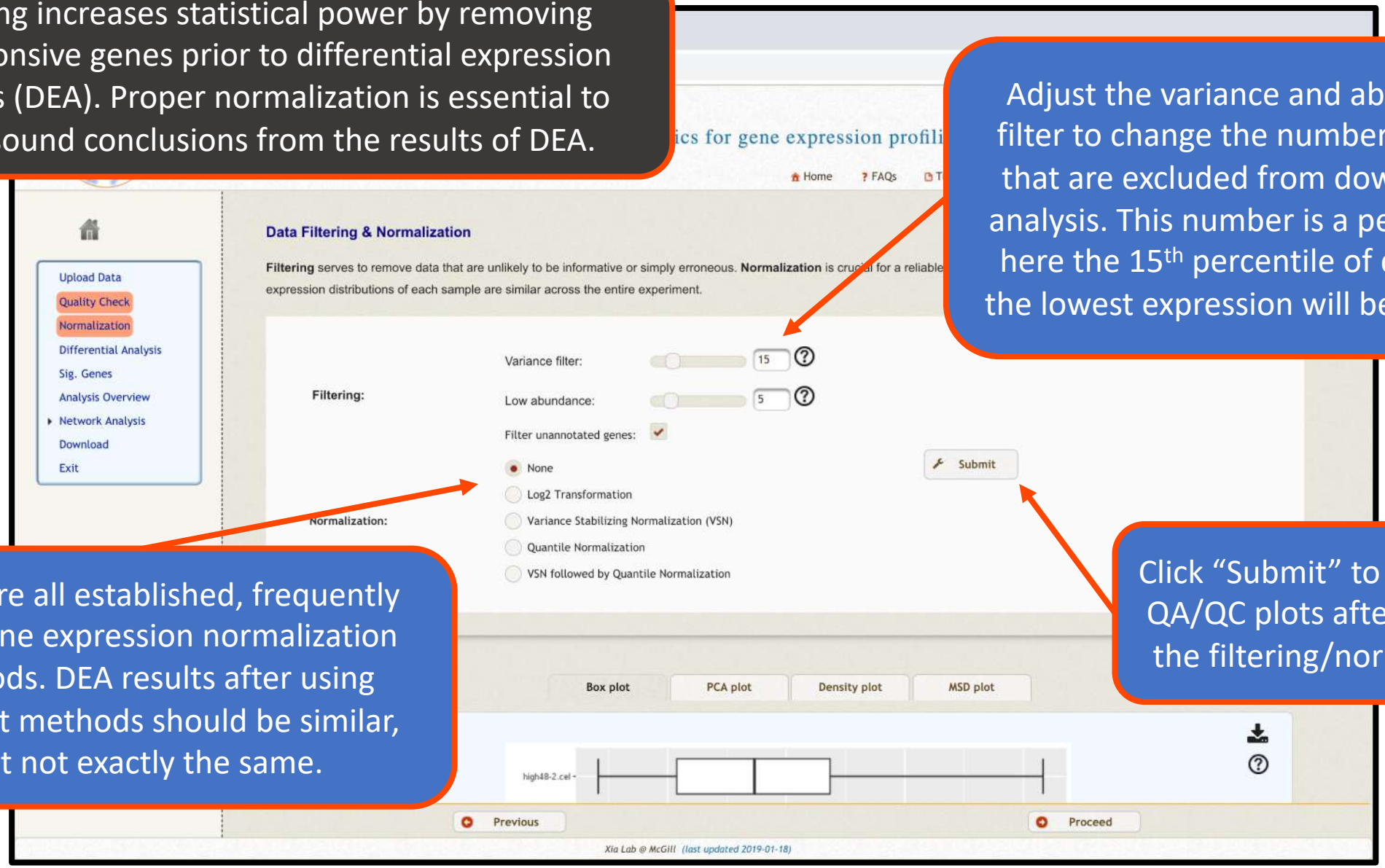


PCA plot: we see that the samples are separated by both time (TIME plot) and by the presence/absence of the estrogen receptor (ER plot). ER seems to be responsible for more variation than TIME.



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.



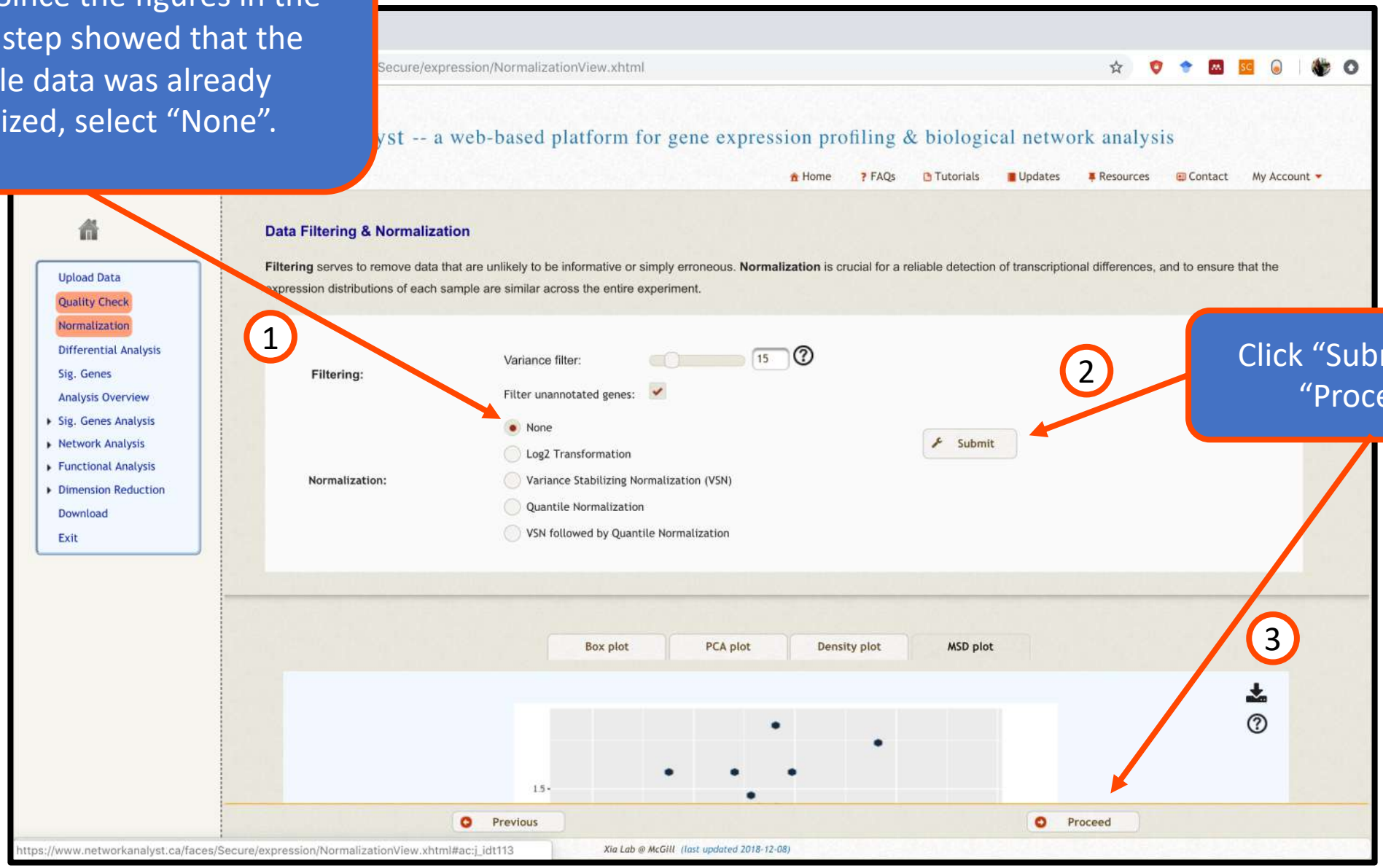
Adjust the variance and abundance filter to change the number of genes that are excluded from downstream analysis. This number is a percentile – here the 15th percentile of data with the lowest expression will be removed

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

Normalize and filter the data

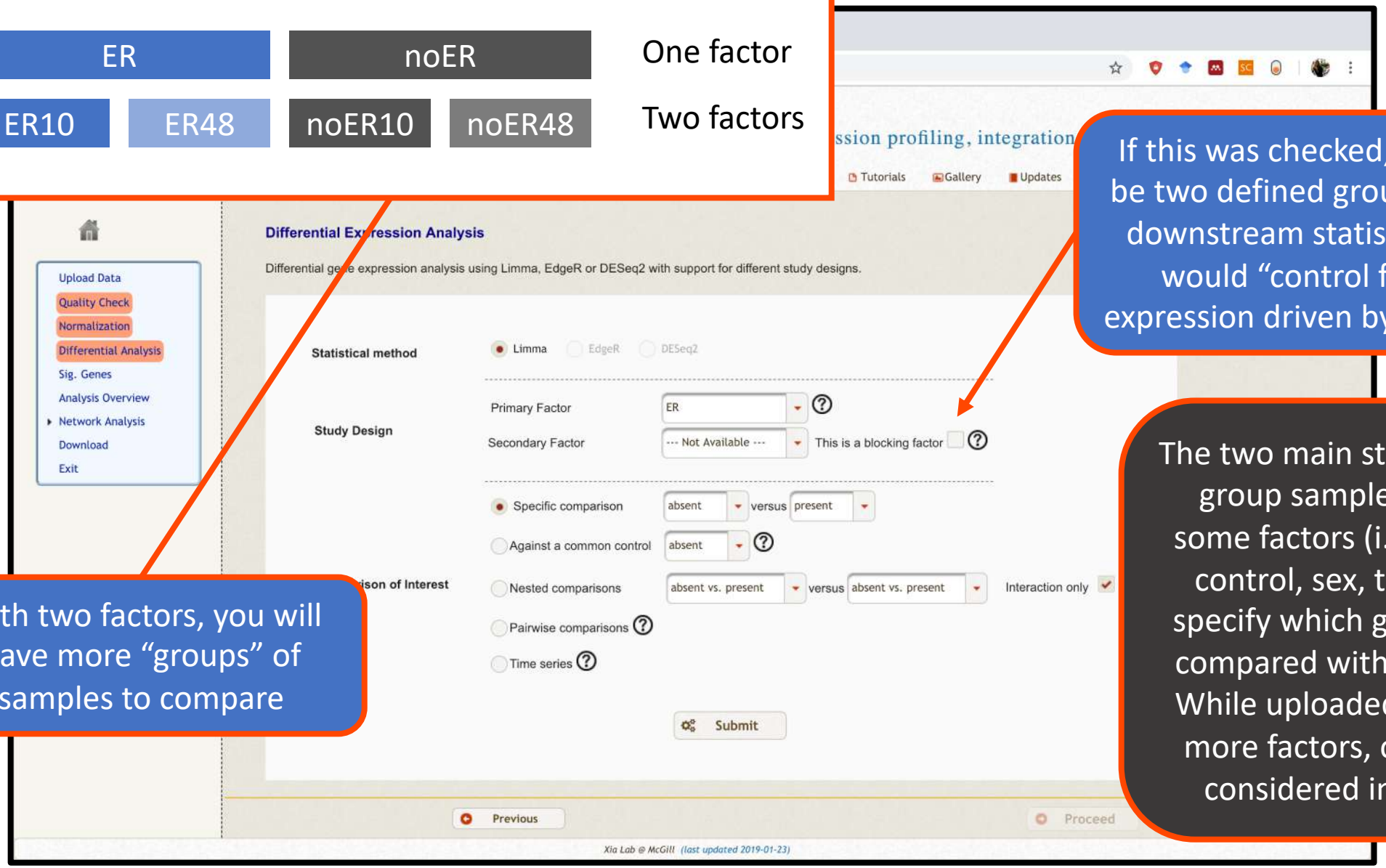
Usually we would normalize our raw data. Since the figures in the previous step showed that the example data was already normalized, select “None”.



Conduct differential expression analysis

ER | noER — One factor

ER10 | ER48 | noER10 | noER48 — Two factors



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.

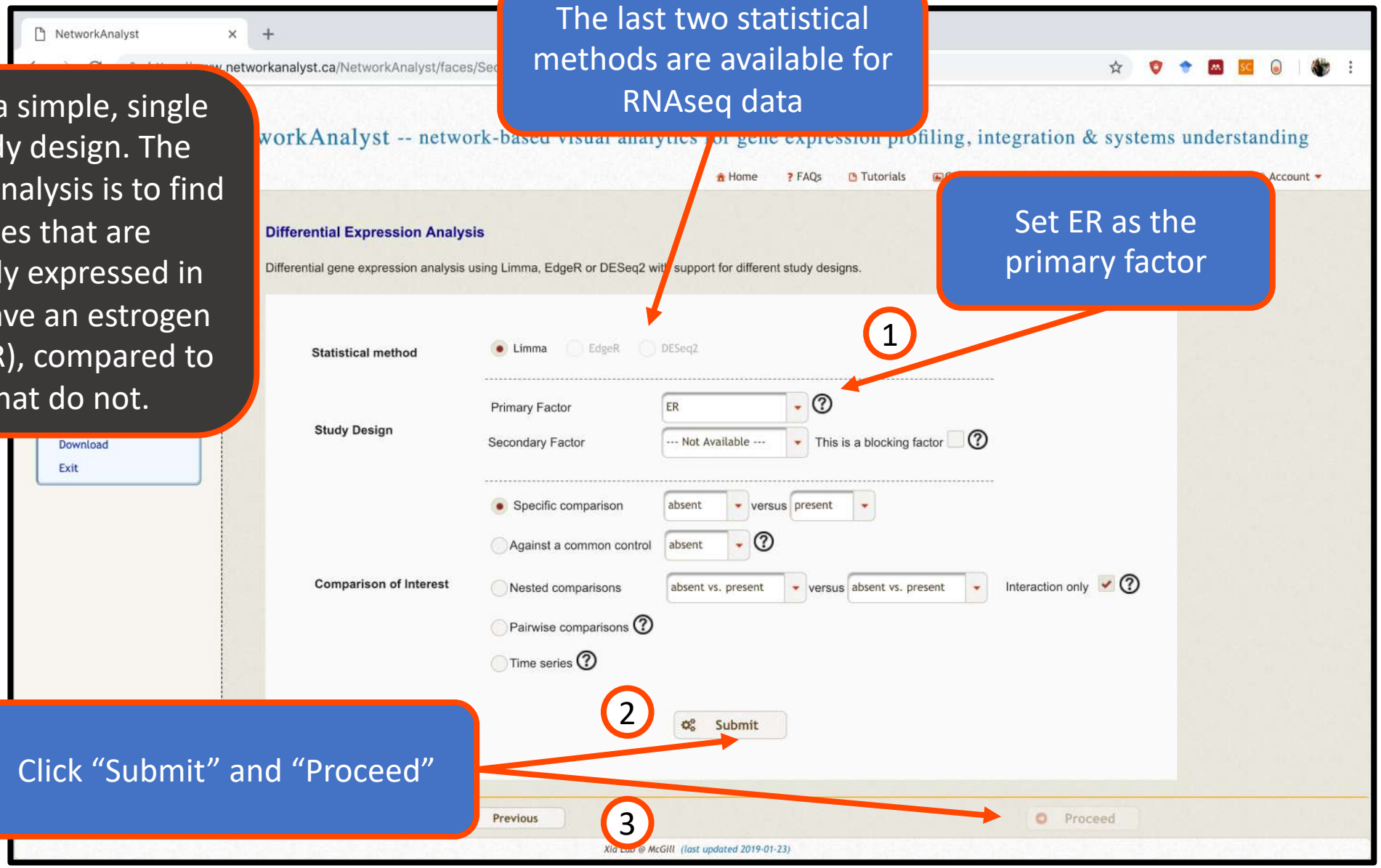
Conduct differential expression analysis

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

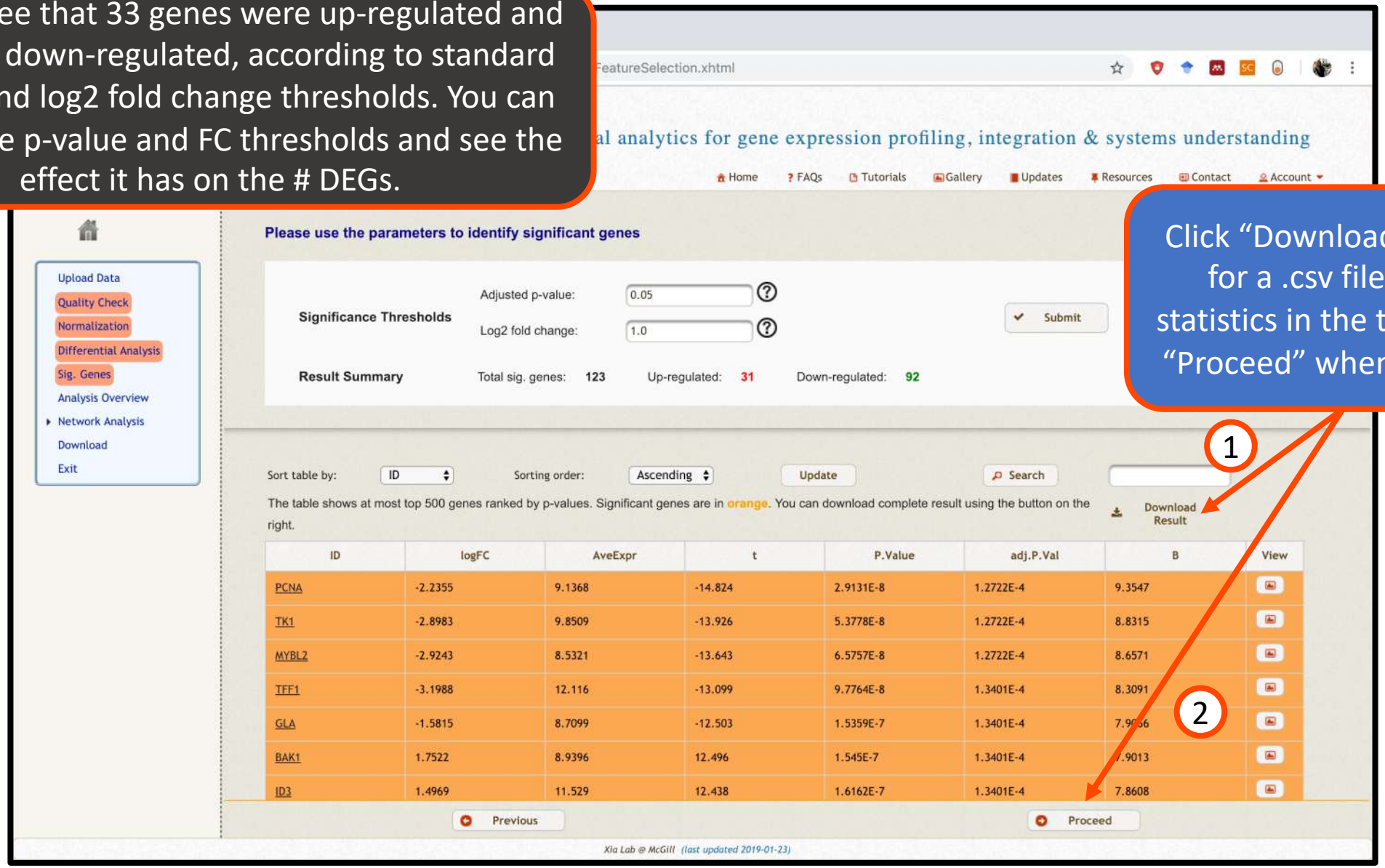
Set ER as the primary factor

Click “Submit” and “Proceed”



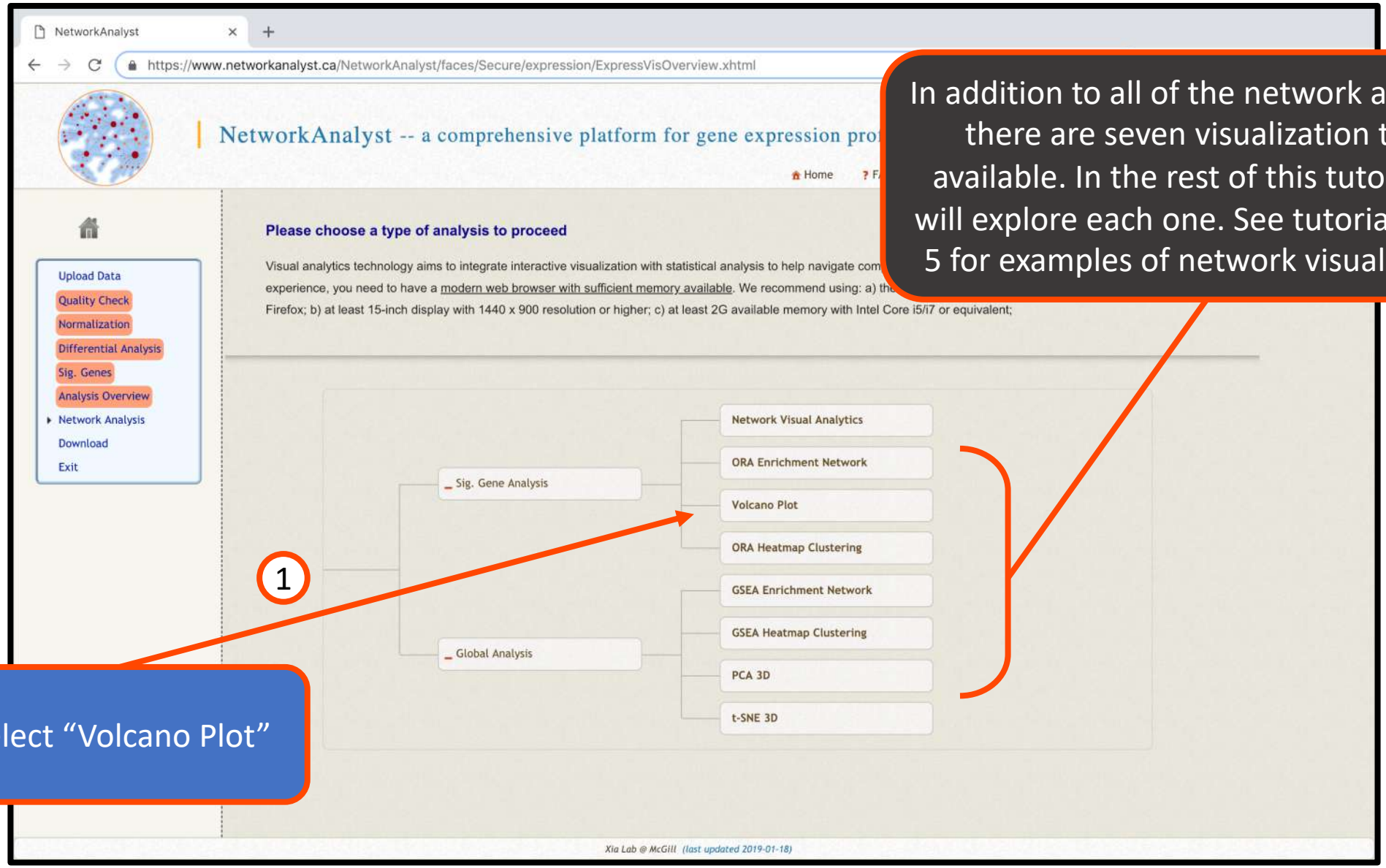
View differentially expressed genes (DEGs)

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.



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

Analysis overview



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.

Select “Volcano Plot”

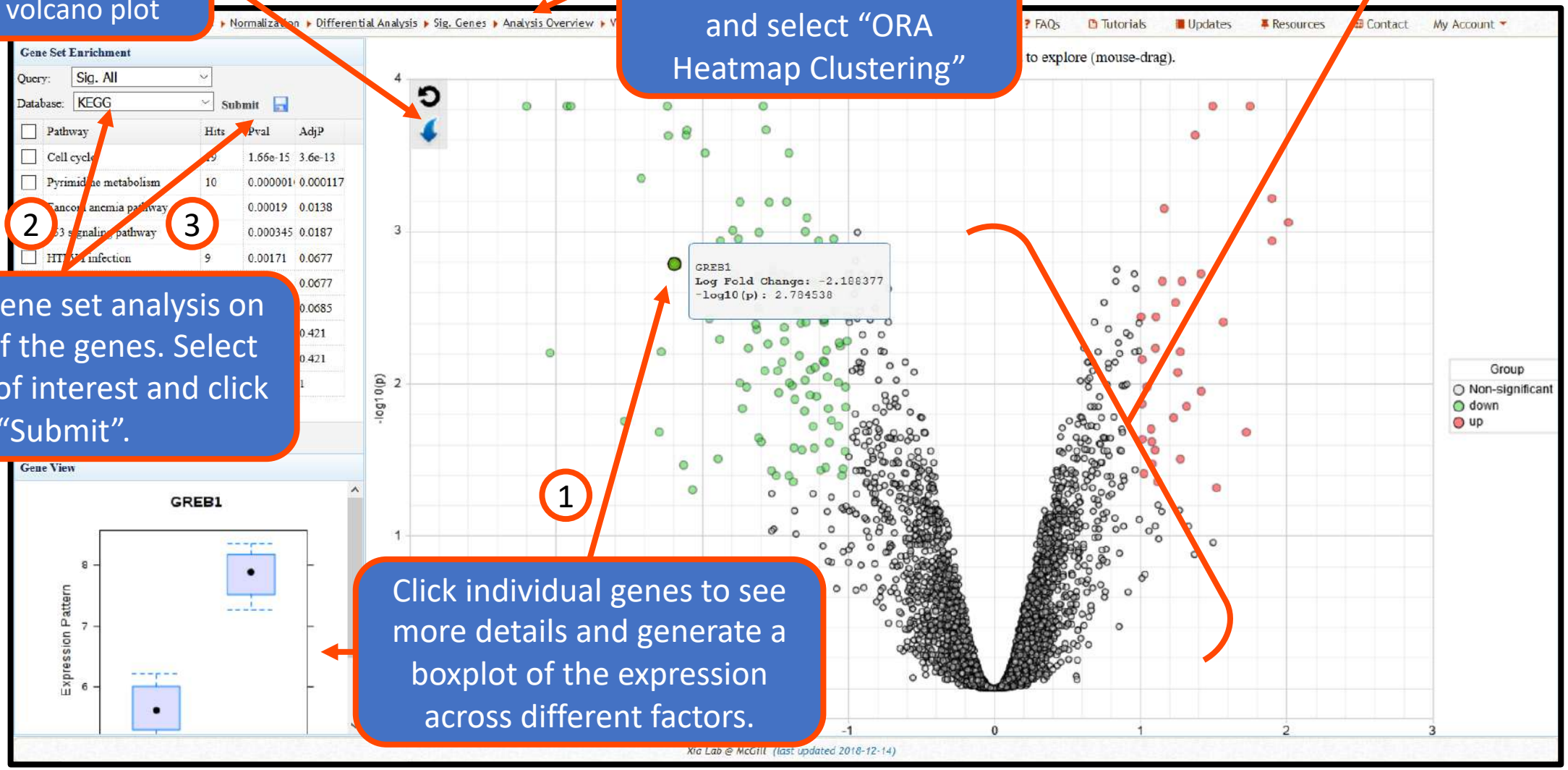
Interactive volcano plot

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

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

4

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



2

3

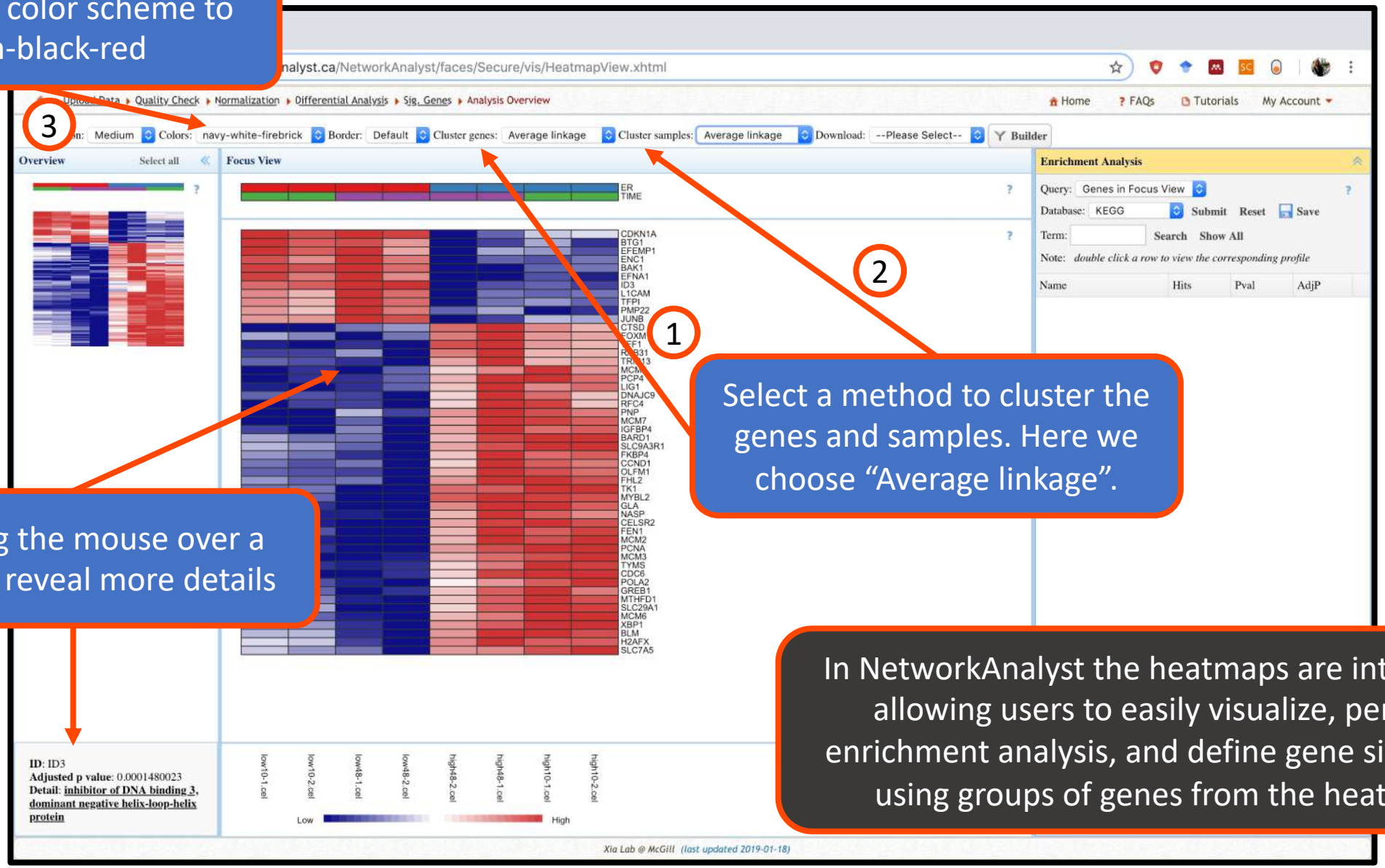
Perform gene set analysis on subsets of the genes. Select database of interest and click “Submit”.

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



1

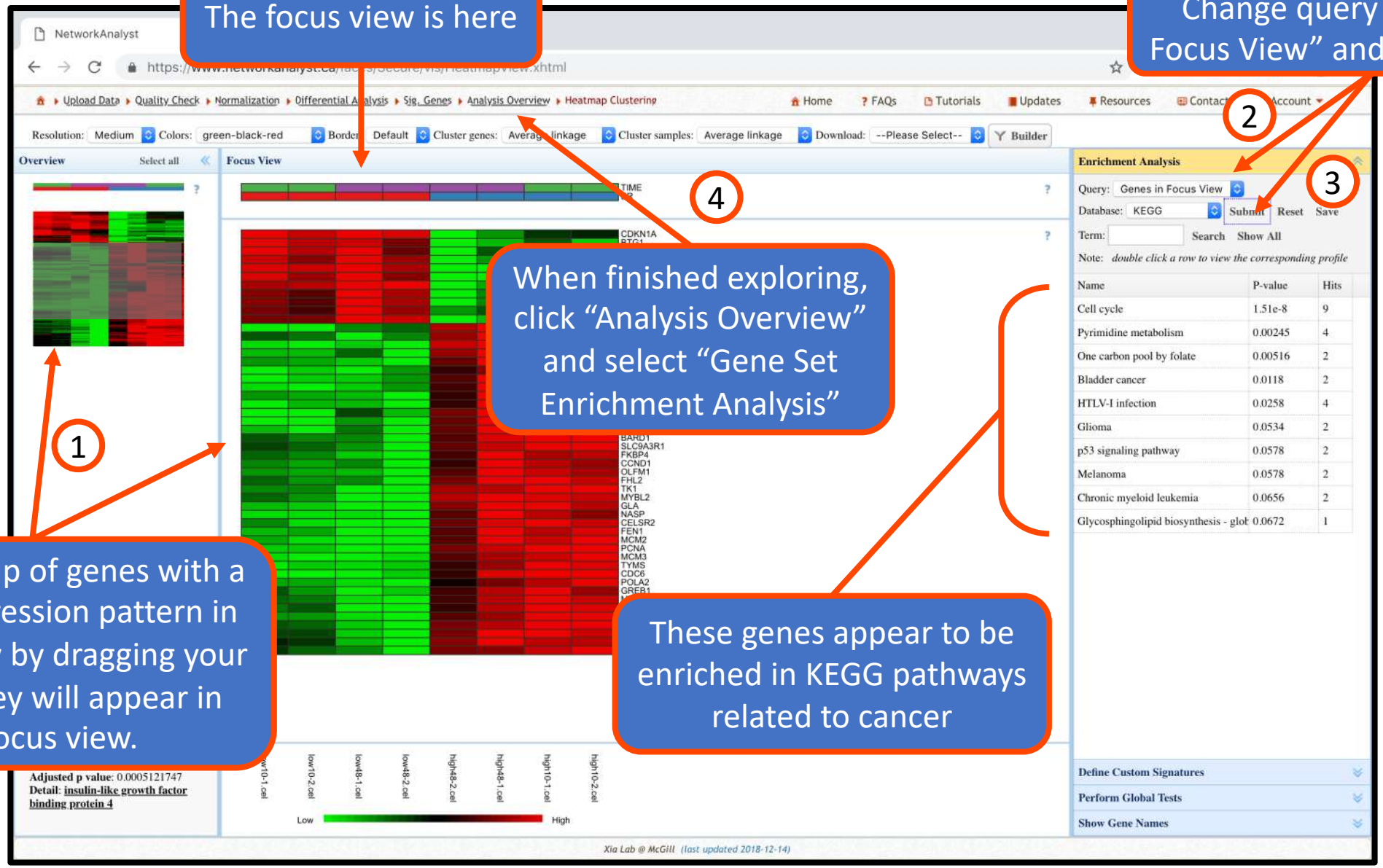
2

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

Hovering the mouse over a gene will reveal more details

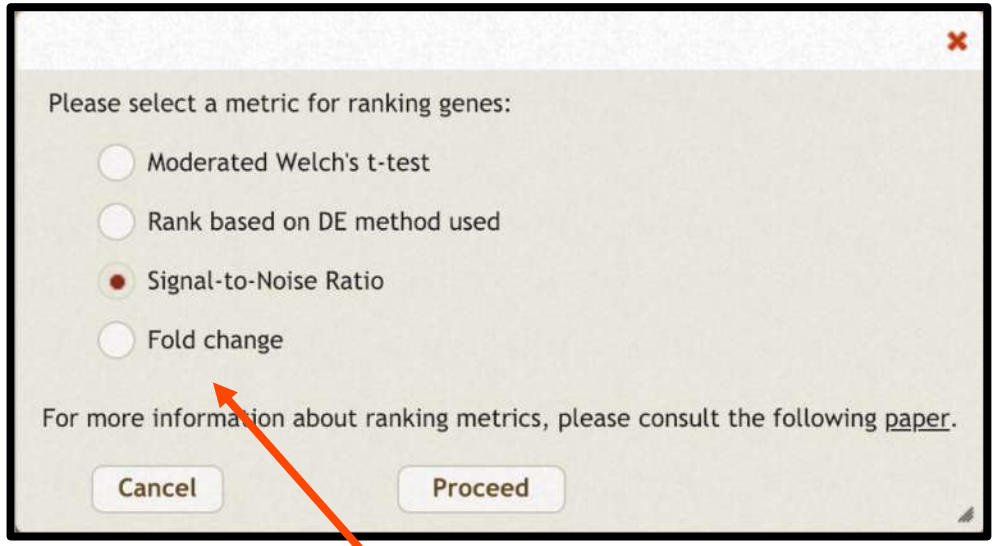
In NetworkAnalyst the heatmaps are interactive, allowing users to easily visualize, perform enrichment analysis, and define gene signatures using groups of genes from the heatmap.

Advanced heatmap functions



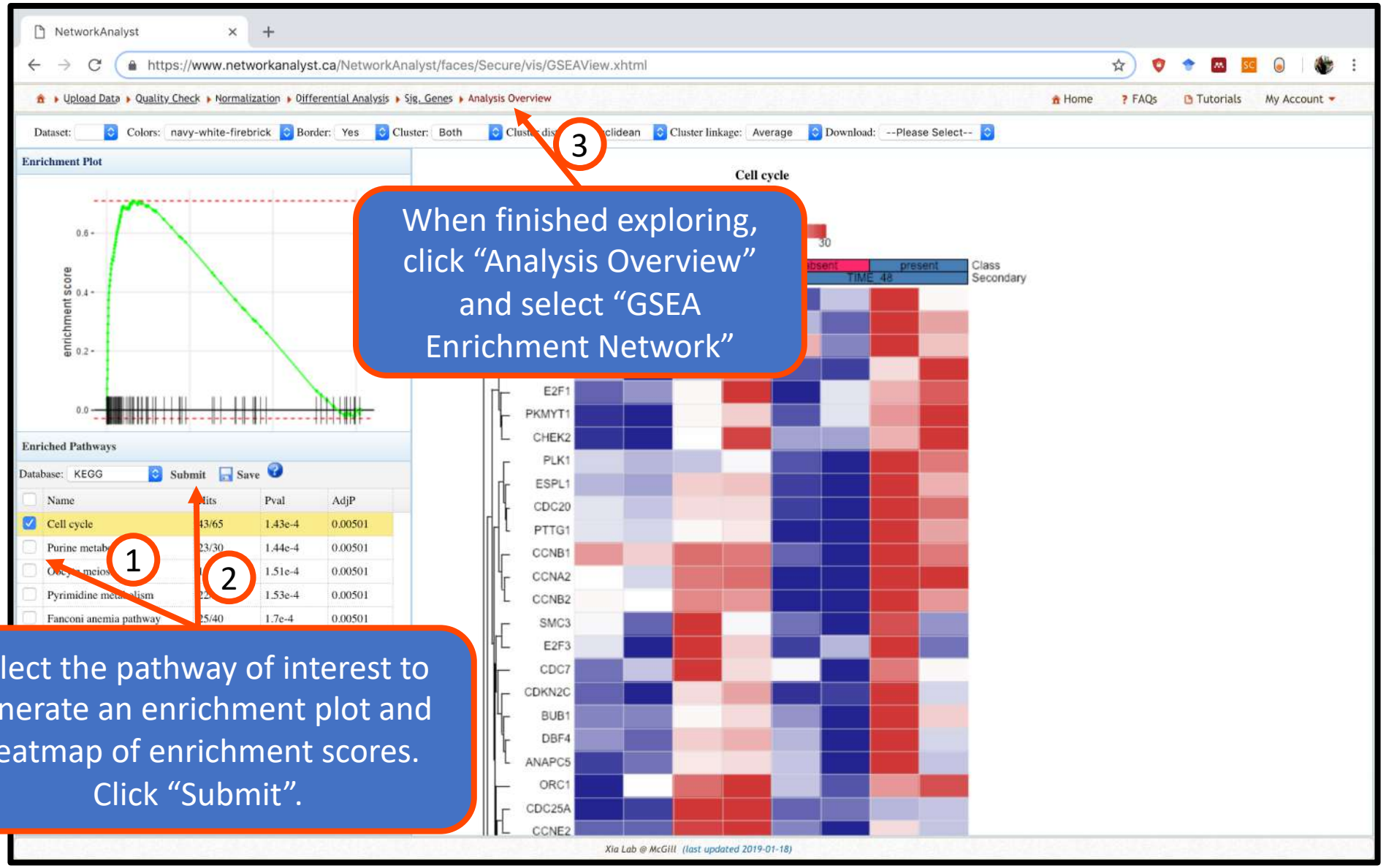
Gene Set Enrichment Analysis (GSEA)

- 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.

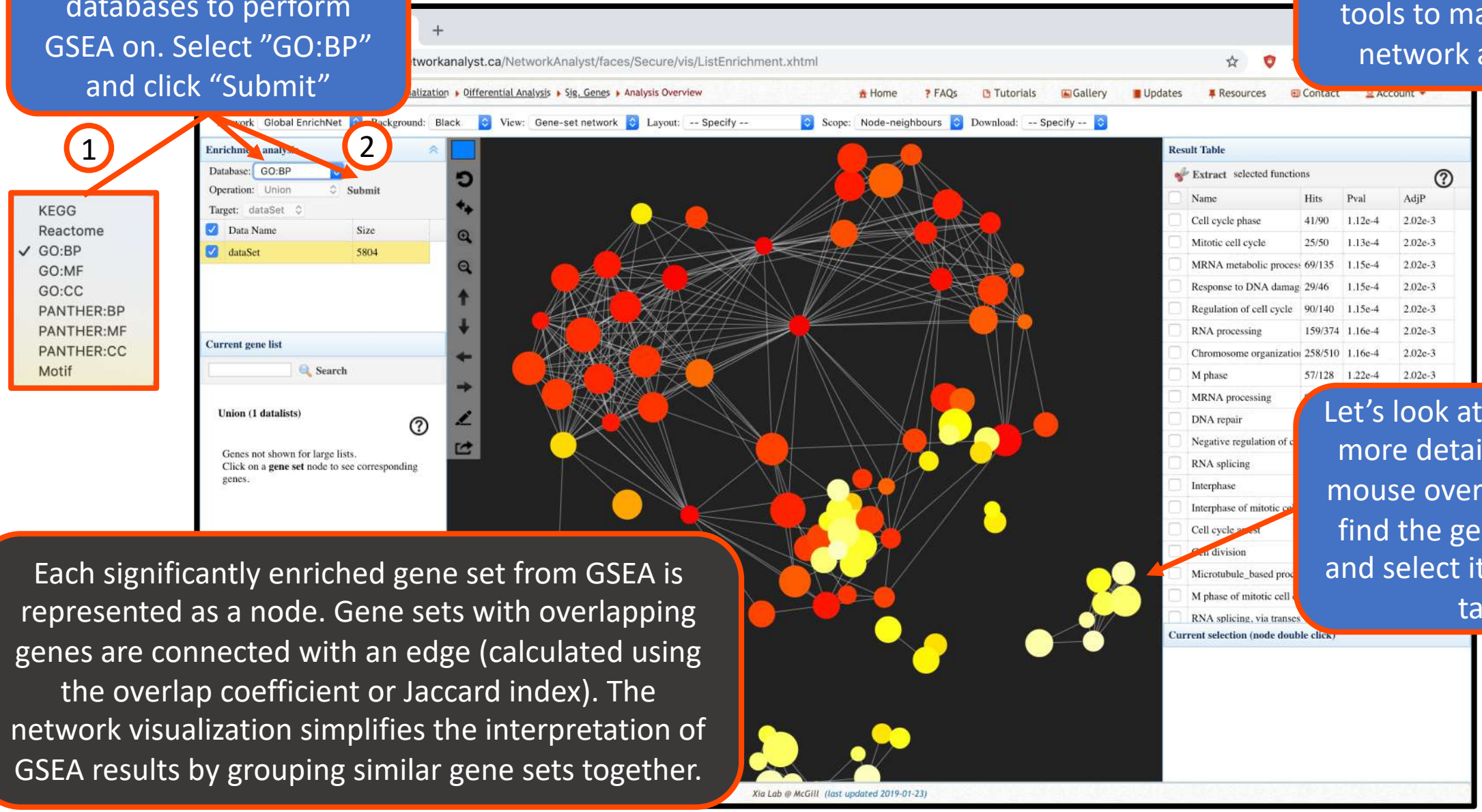
GSEA Heatmap Clustering



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



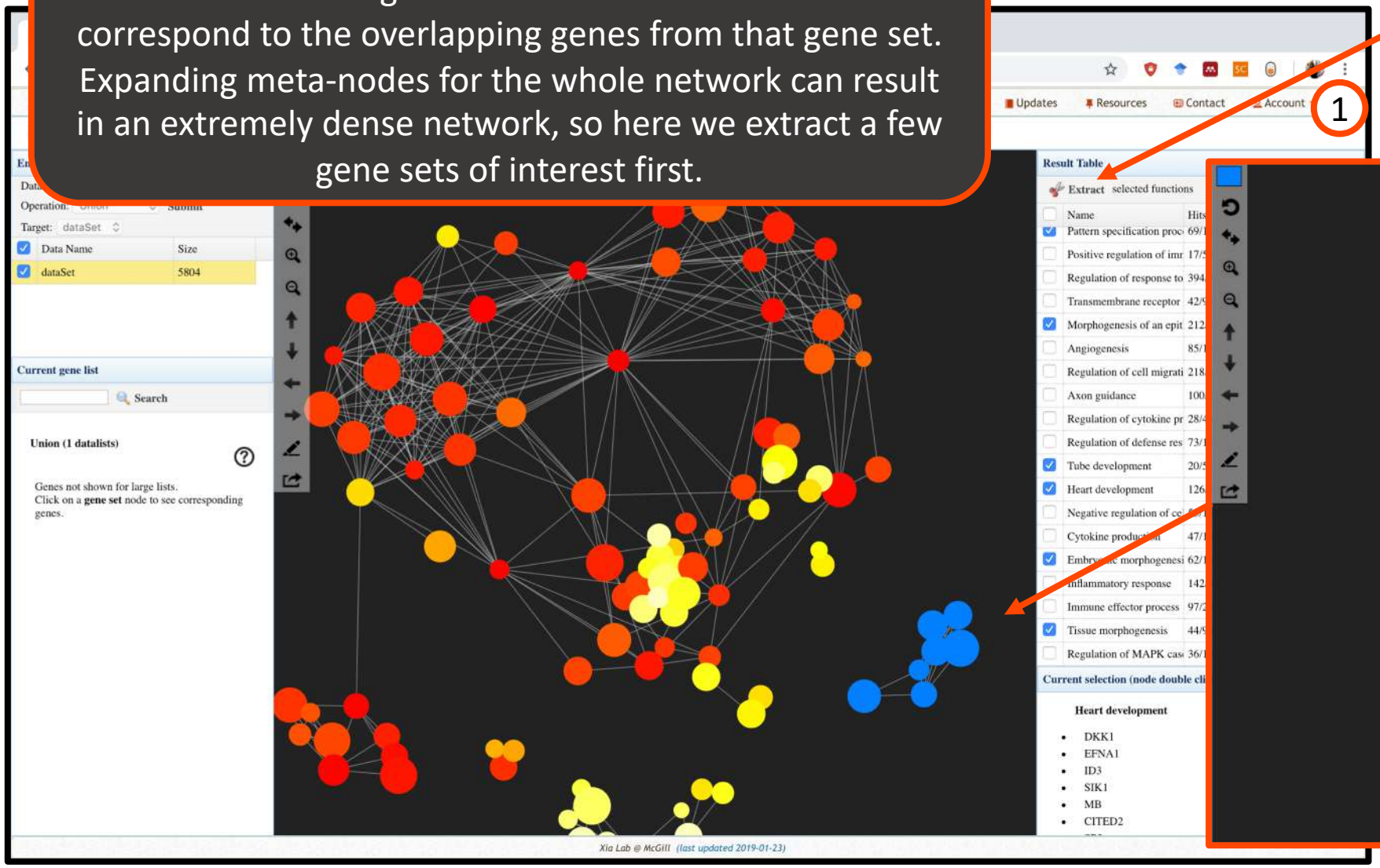
Let’s look at this cluster in more detail. Hover your mouse over each node to find the gene set name, and select it in the results table.

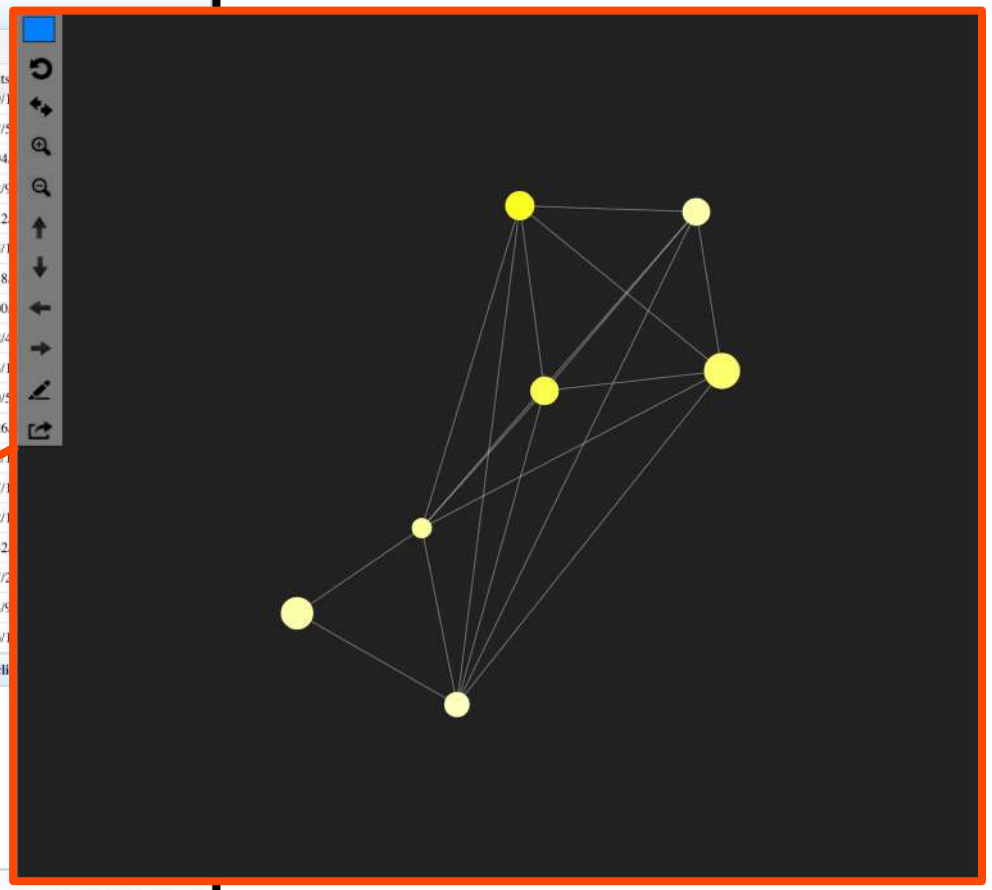
Each significantly enriched gene set from GSEA is represented as a node. Gene sets with overlapping genes are connected with an edge (calculated using the overlap coefficient or Jaccard index). The network visualization simplifies the interpretation of GSEA results by grouping similar gene sets together.

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

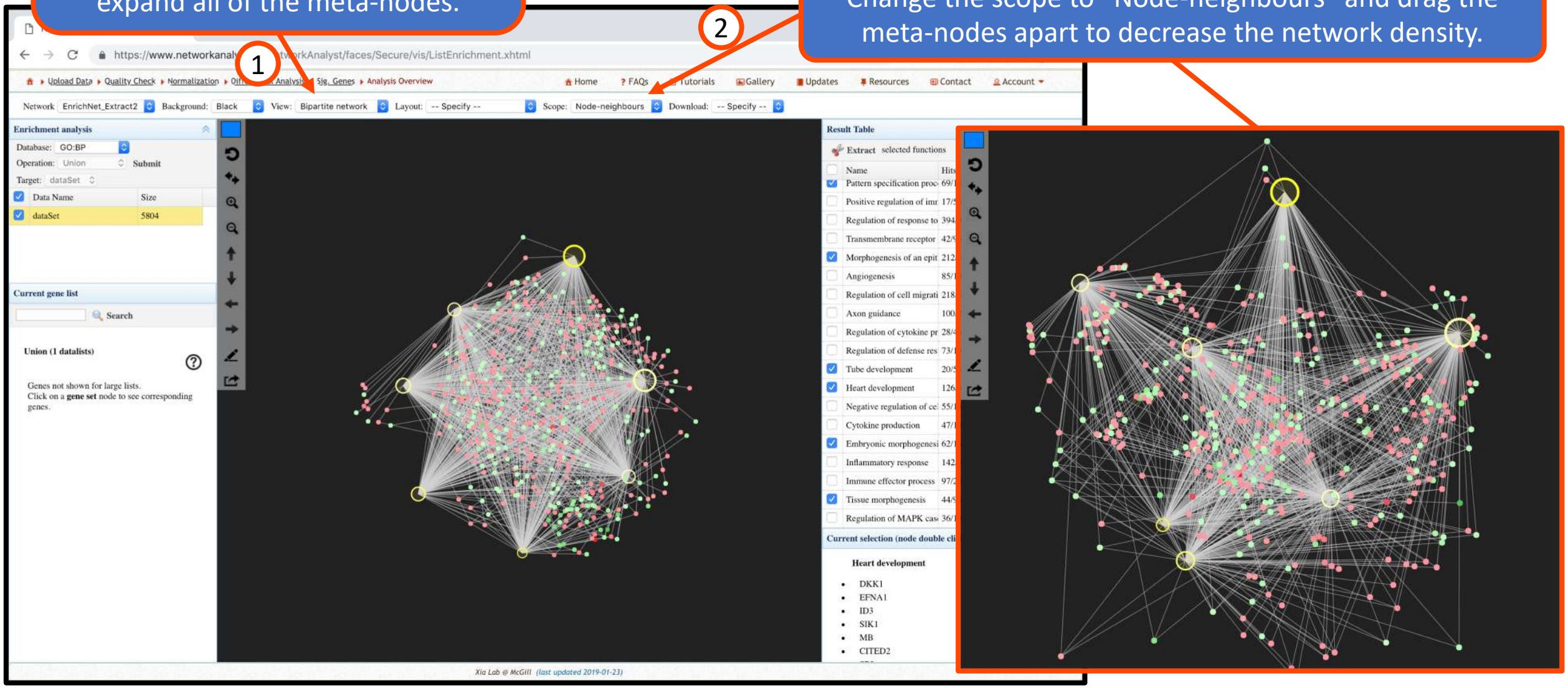




GSEA Enrichment Network

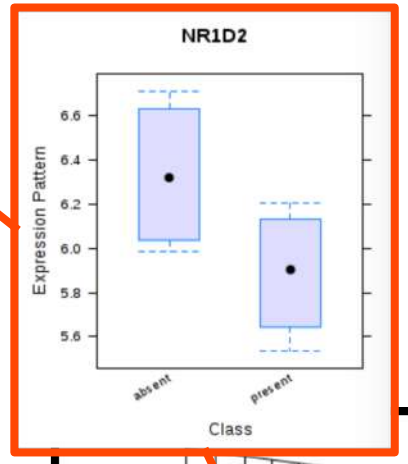
Double-click a few meta-nodes to see the nested genes. After, select the "Bipartite network" view to expand all of the meta-nodes.

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



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.

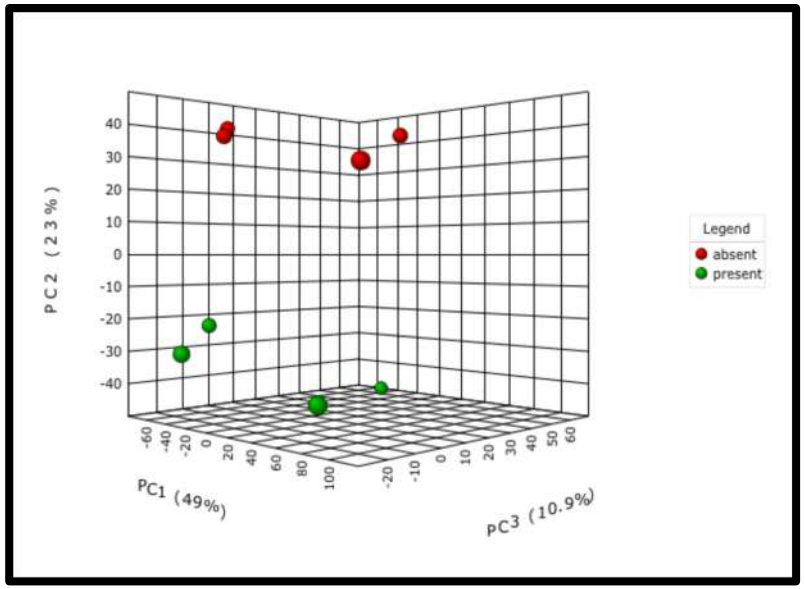


Figure 1: 3D PCA

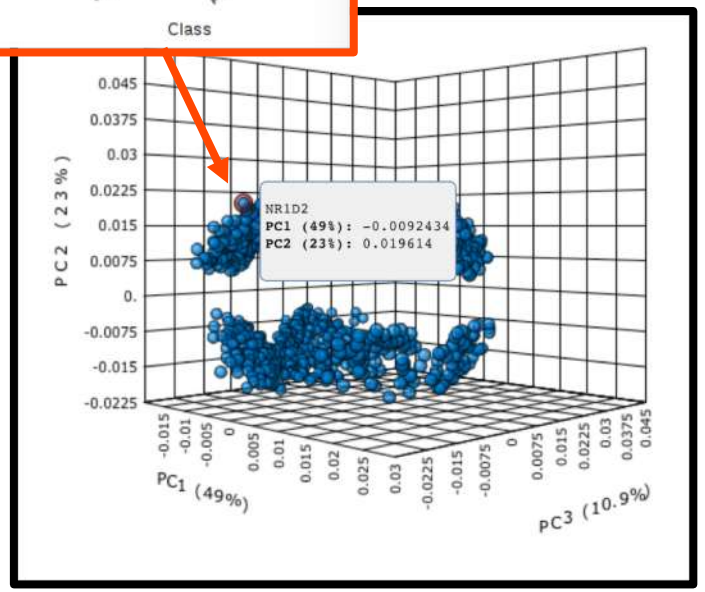


Figure 2: 3D PCA loading plot

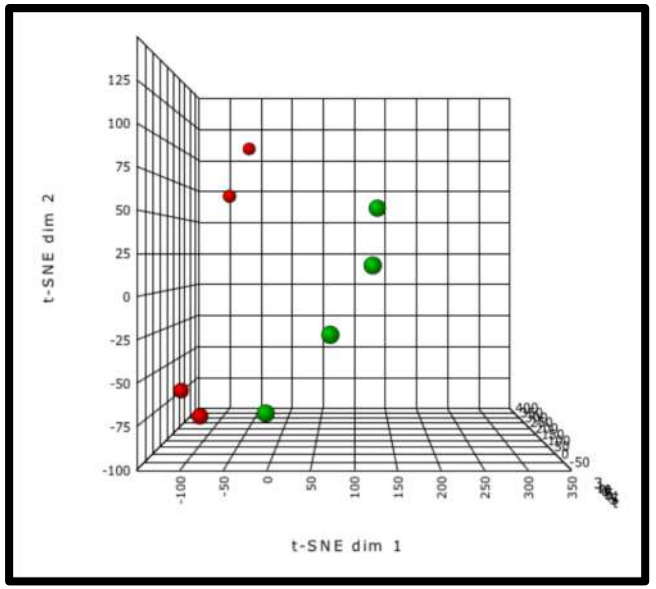
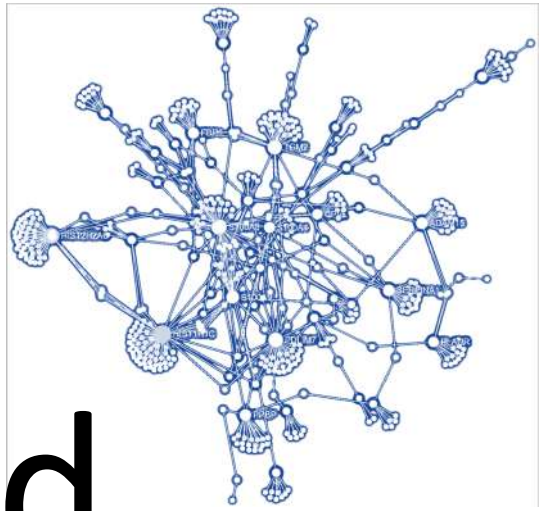


Figure 3: 3D tSNE



The End

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