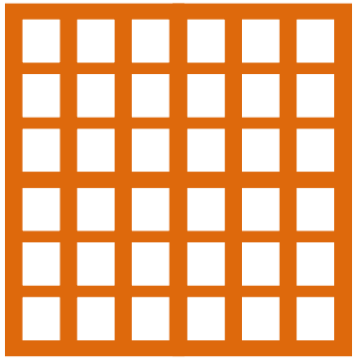
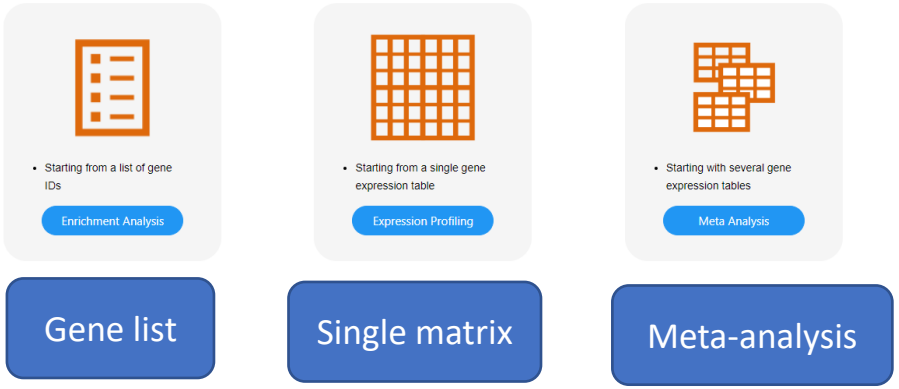
ExpressAnalyst - Tutorial
Starting from a table

-- Comprehensive platform for gene expression and meta-analysis



Intro to ExpressAnalyst

- Web platform for the analysis of gene expression data and meta-analysis
  - Previously part of NetworkAnalyst
- Designed for bench researchers rather than specialized bioinformaticians
- Integrates data processing, statistical analysis and data visualization to support:
  - Data comparisons
  - Biological interpretation
  - Hypothesis generation



Computer and browser requirements

- A modern web browser with JavaScript 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 scatter visualization
- 50MB limit for data upload
    - ~300 samples for gene expression data with 20 000 genes

Goals for this tutorial

- ExpressAnalyst is focused on performing secondary and tertiary analysis of transcriptomics data. It does not deal with raw data processing.

- In this tutorial we are going to go through three main points
  - Show the data format accepted by ExpressAnalyst
  - Go through data upload and processing steps with example dataset
  - Show the different algorithms and contrasts offered for differential expression analysis in ExpressAnalyst

Data format

The data file can be tab delimited (.tab) or comma delimited (.csv)

Sample names

Meta-data

Can be single or two metadata types

Gene/probe ids

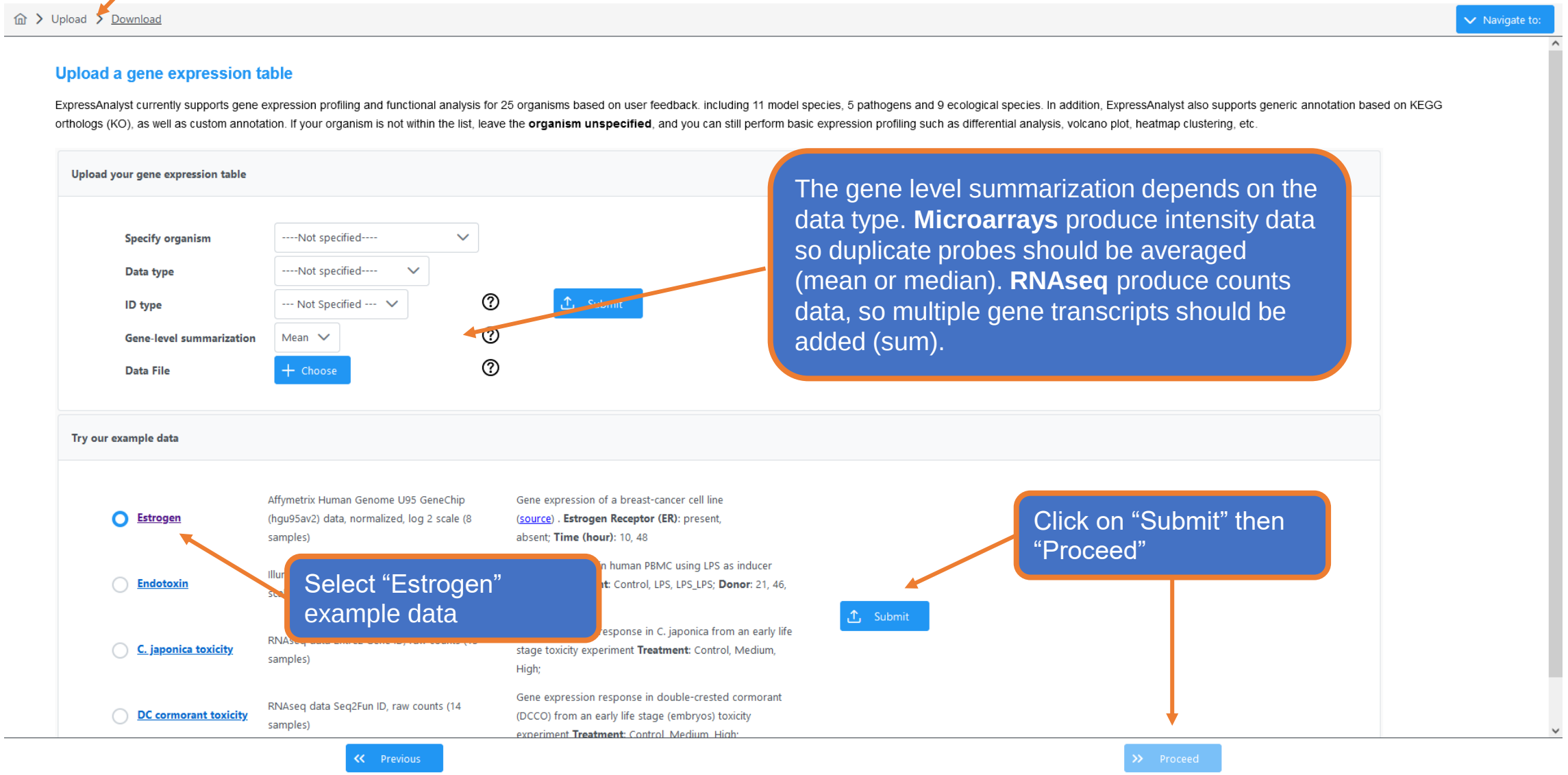
#NAME	low10-1.cel	low10-2.cel	high10-1.cel	high10-2.cel	low48-1.cel	low48-2.cel	high48-1.cel	high48-2.cel
#CLASS:ER	absent	absent	present	present	absent	absent	present	present
#CLASS:TIME	10	10	10	10	48	48	48	48
100_g_at	9.642896152	9.74149593	9.537036294	9.353625042	9.591697198	9.570590003	9.475796234	9.530655159
1000_at	10.39816907	10.25436246	10.00397056	9.903528072	10.3748662	10.03352045	10.34506604	9.86332109
1001_at	5.717613479	5.881007611	5.859563251	5.95402767	5.96054022	6.020889393	5.981080253	6.285192094
1002_f_at	5.512595956	5.801806991	5.571064822	5.608131831	5.390063911	5.494511159	5.508103538	5.630106526
1003_s_at	7.783926552	8.007975311	8.037998859	7.835119841	7.92648674	8.13886965	7.994936847	8.233337701
1004_at	7.289162155	7.603670275	7.488538813	7.771505854	7.521788542	7.599544133	7.456149346	7.675170716
1005_at	9.206737493	8.993802402	8.237894255	8.338003525	9.173196386	9.040470337	7.926105833	8.069686035
1006_at	5.387193668	5.555903141	5.407539579	5.74403733	5.635769674	5.75312714	5.485841919	5.750324626
1007_s_at	11.90333613	11.74451474	11.40879438	11.52722659	11.60691395	11.34409232	11.4218455	11.04993157
1008_f_at	10.11933122	10.98664643	10.83029726	10.02509297	11.044701	11.13829882	10.70570446	11.36951087

...

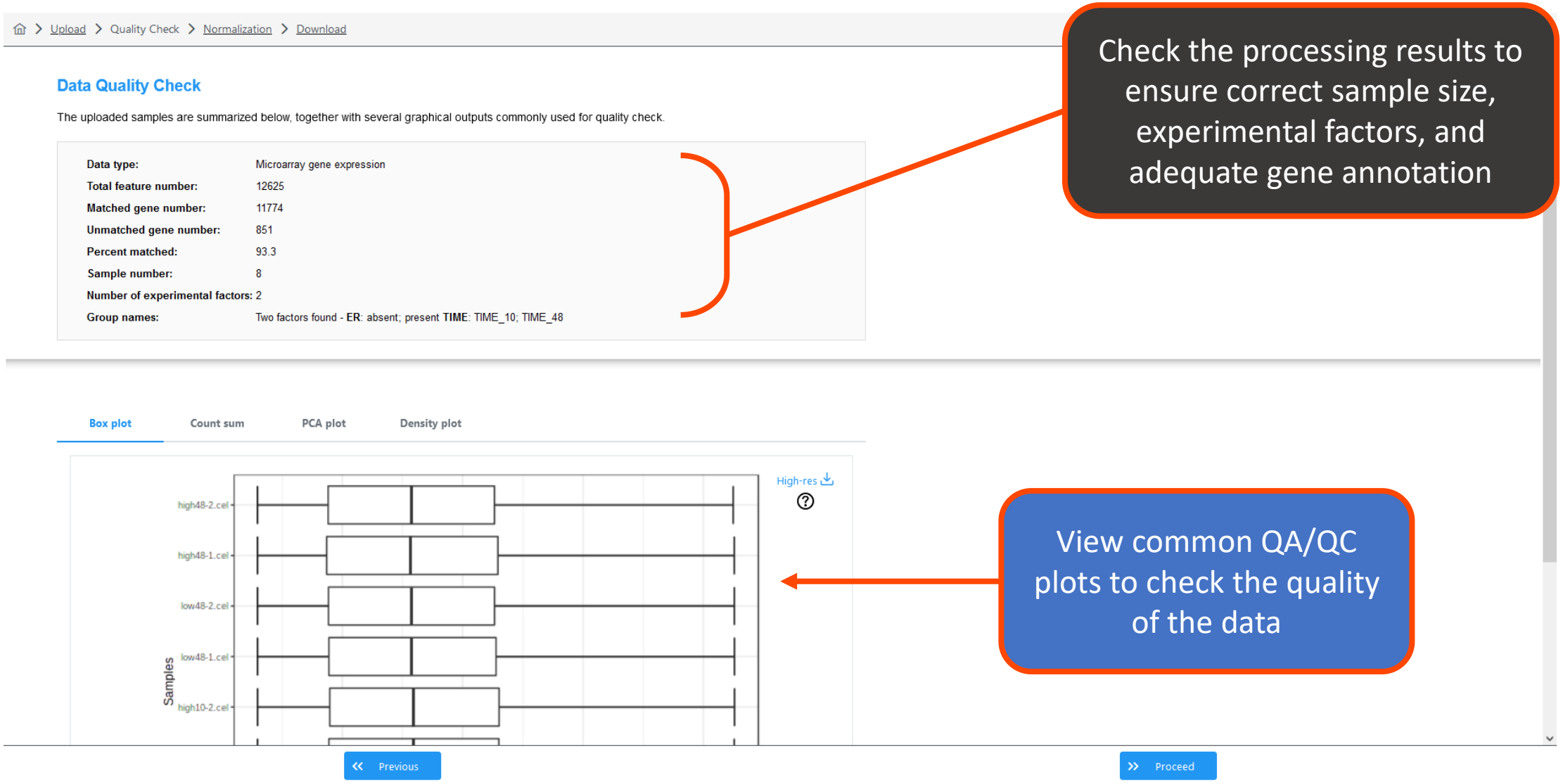
https://www.expressanalyst.ca/ExpressAnalyst/resources/data/test/estrogen.txt

Data upload and annotation

Navigation bar to track analysis progress



View processing results

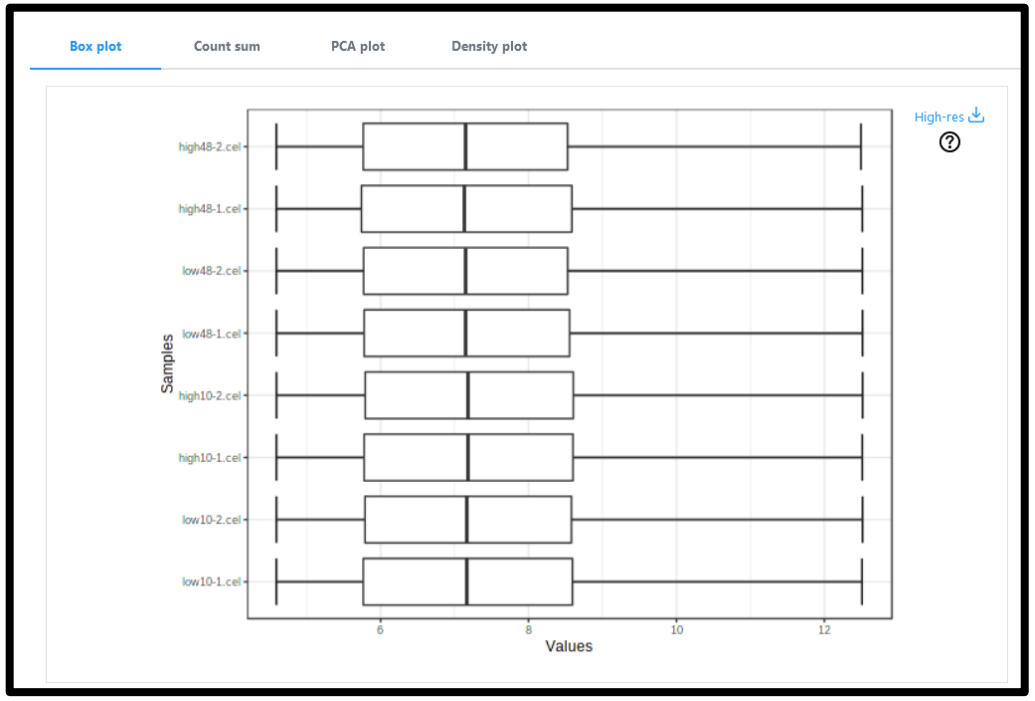


Check the processing results to ensure correct sample size, experimental factors, and adequate gene annotation

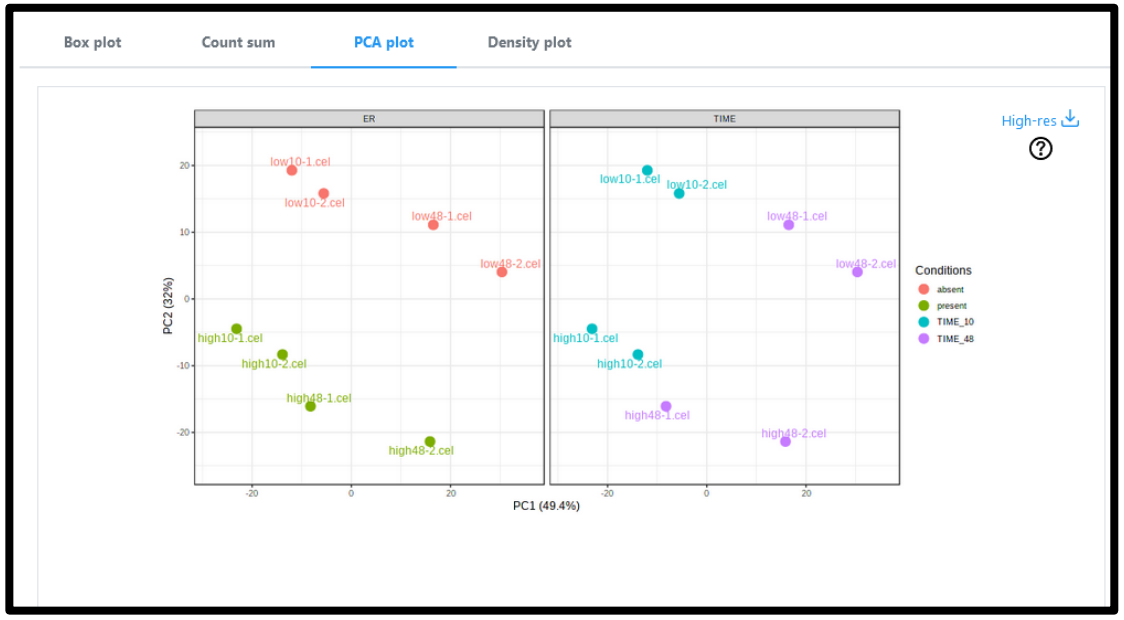
View common QA/QC plots to check the quality of the data

View QC plots

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



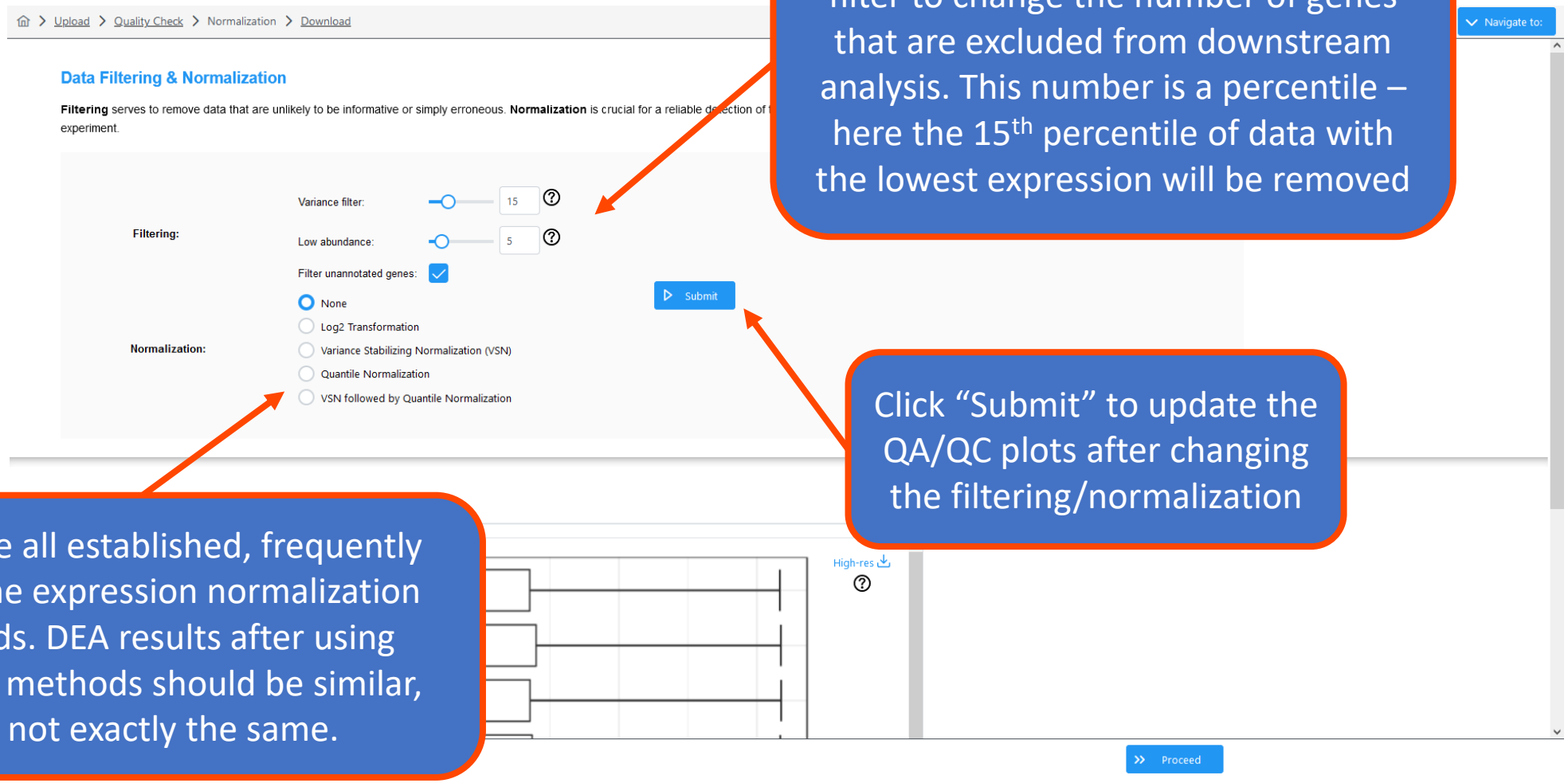
PCA plot: we see that the samples are separated by both metadata, time (TIME plot) and by the presence/absence of the estrogen receptor (ER plot). ER seems to be responsible for more variation than TIME.



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

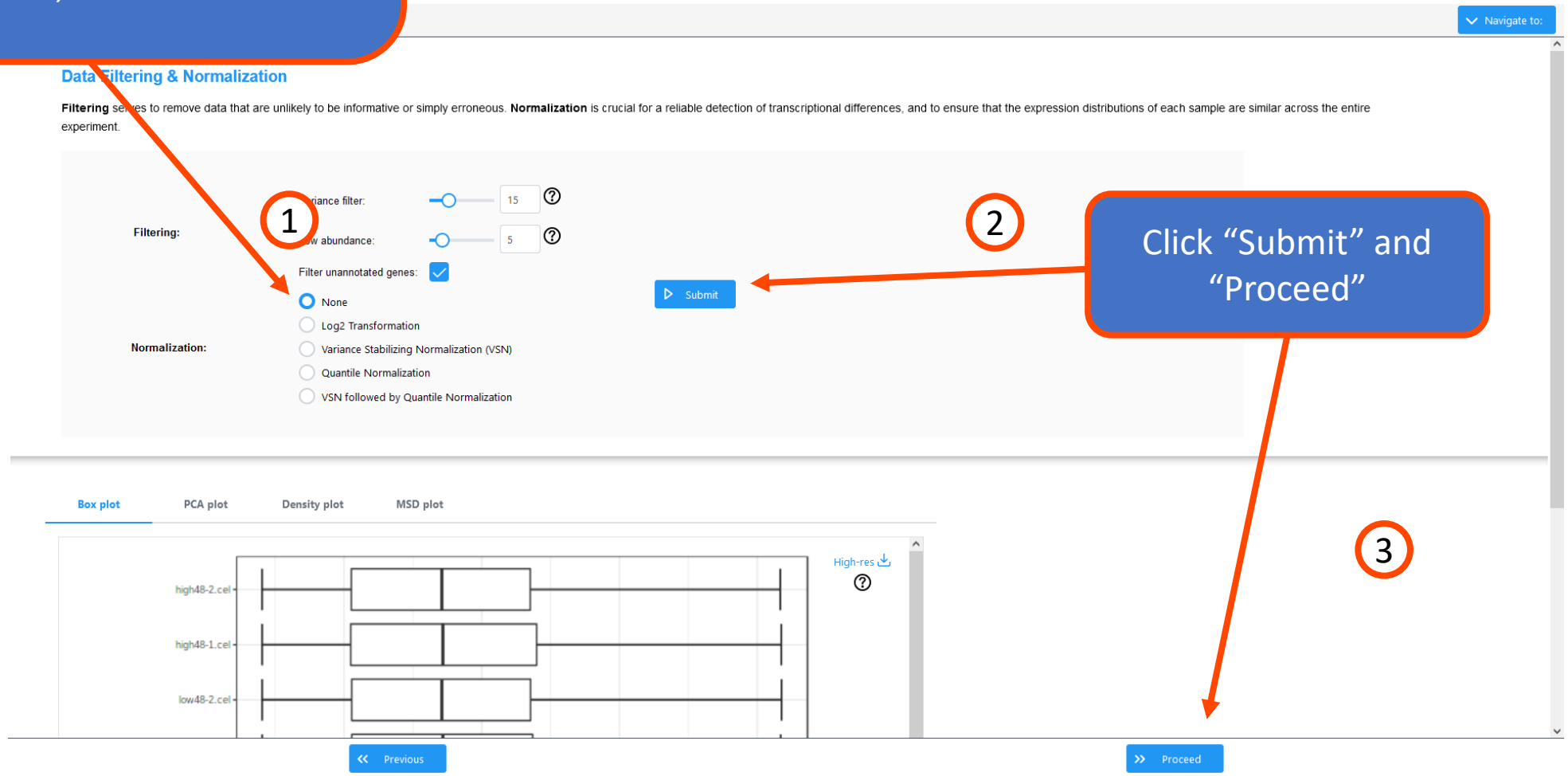


Click “Submit” to update the QA/QC plots after changing the filtering/normalization

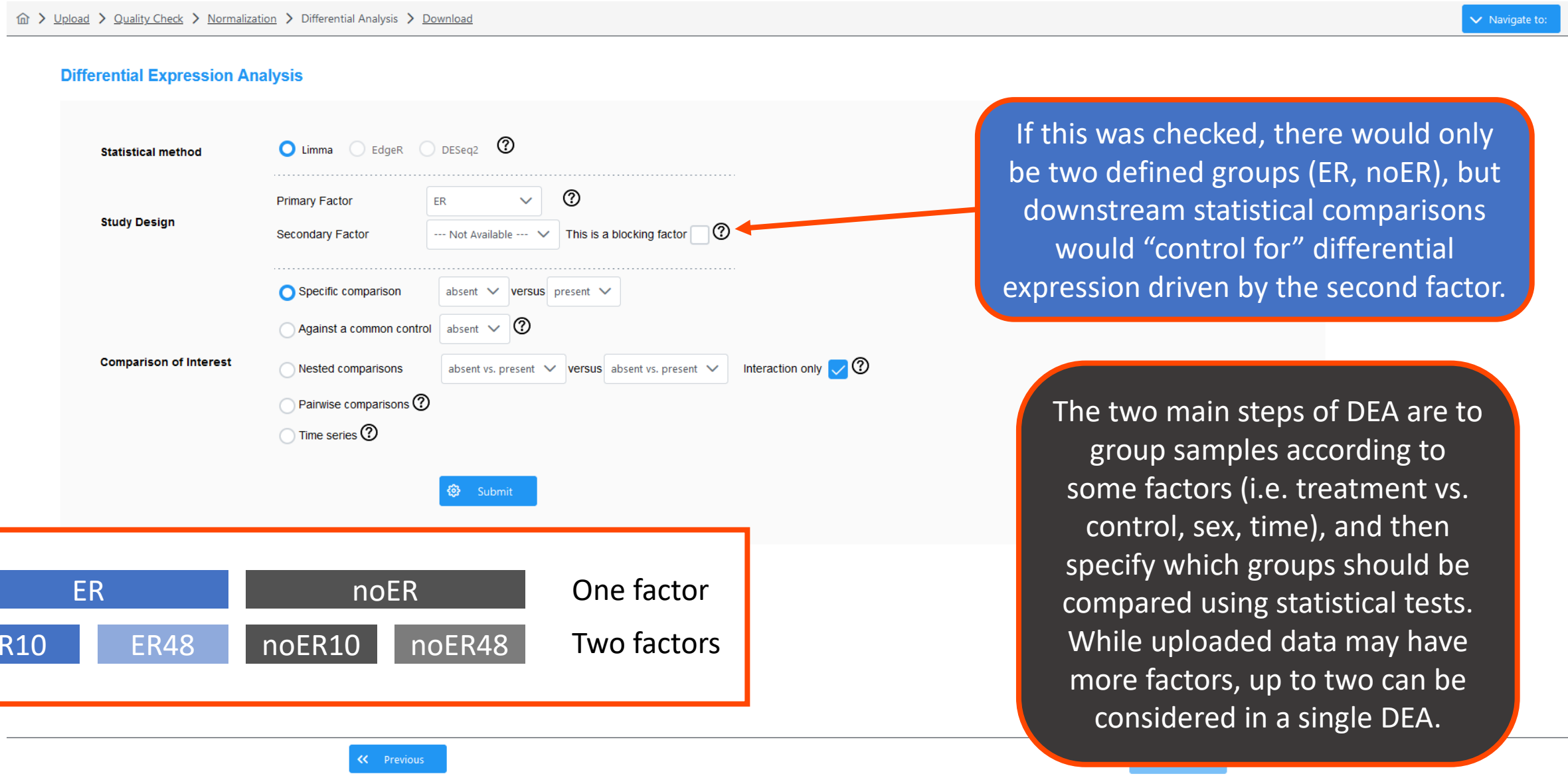
These are all established, frequently used gene expression normalization methods. DEA results after using different methods should be similar, but not exactly the same.

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



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.

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 using statistical tests. While uploaded data may have more factors, up to two can be considered in a single DEA.

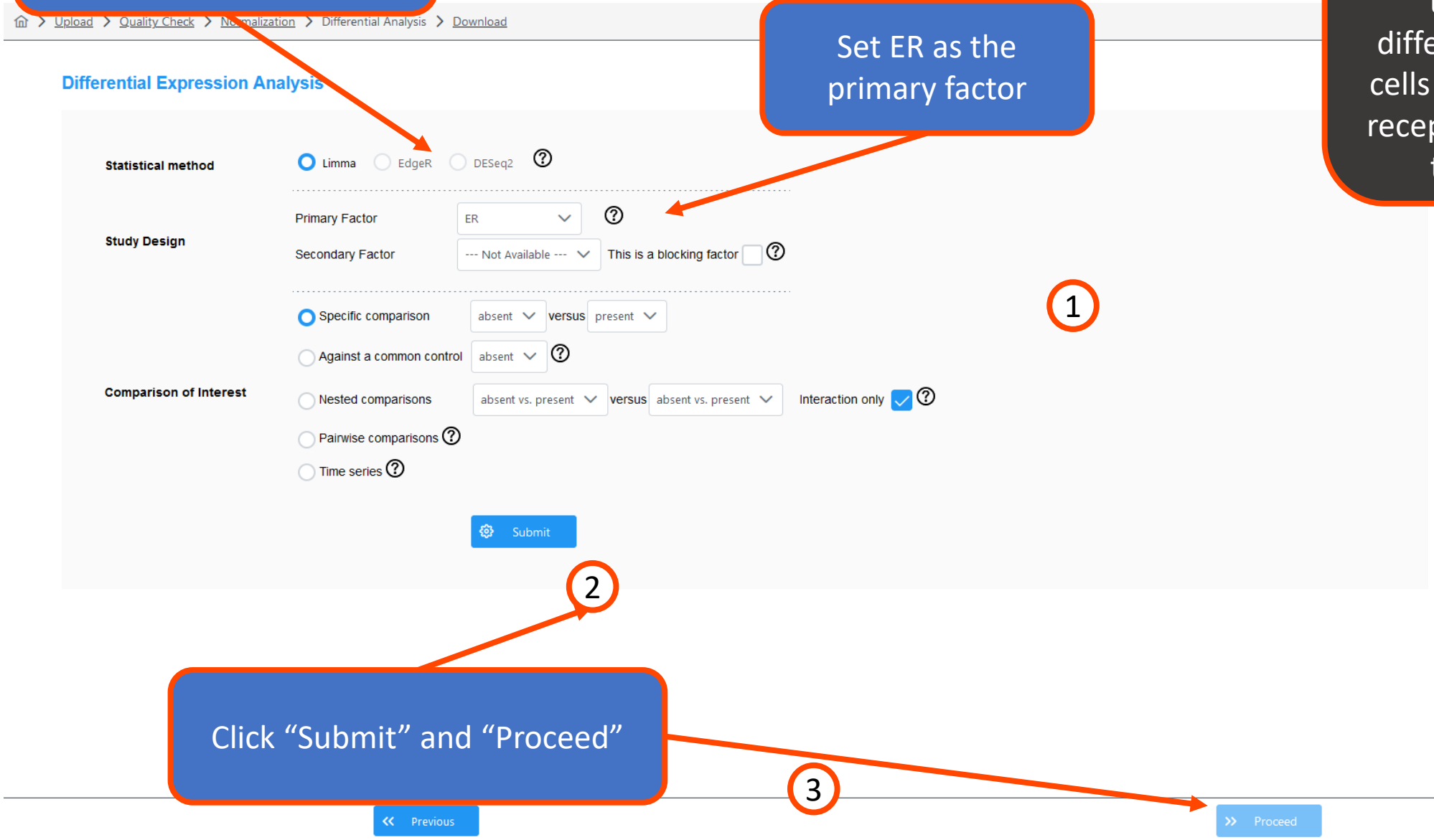
ER	noER	One factor
ER10	ER48	noER10	noER48	Two factors

Conduct DE analysis

The last two statistical methods are available for RNAseq data

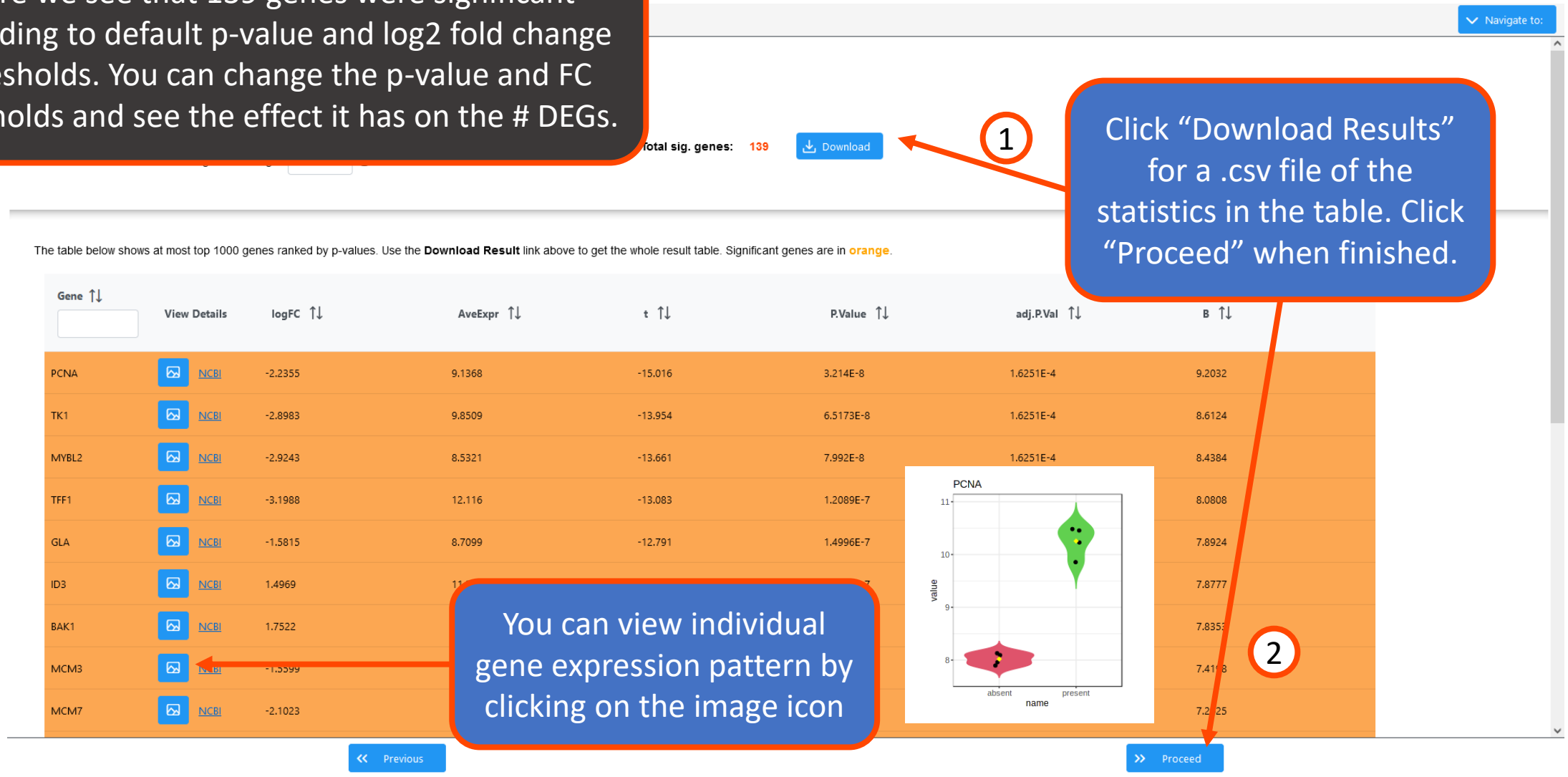
We will do a simple, single factor study design. The goal of this analysis is to find the genes that are differentially expressed in cells that have an estrogen receptor (ER), compared to those that do not.

Set ER as the primary factor



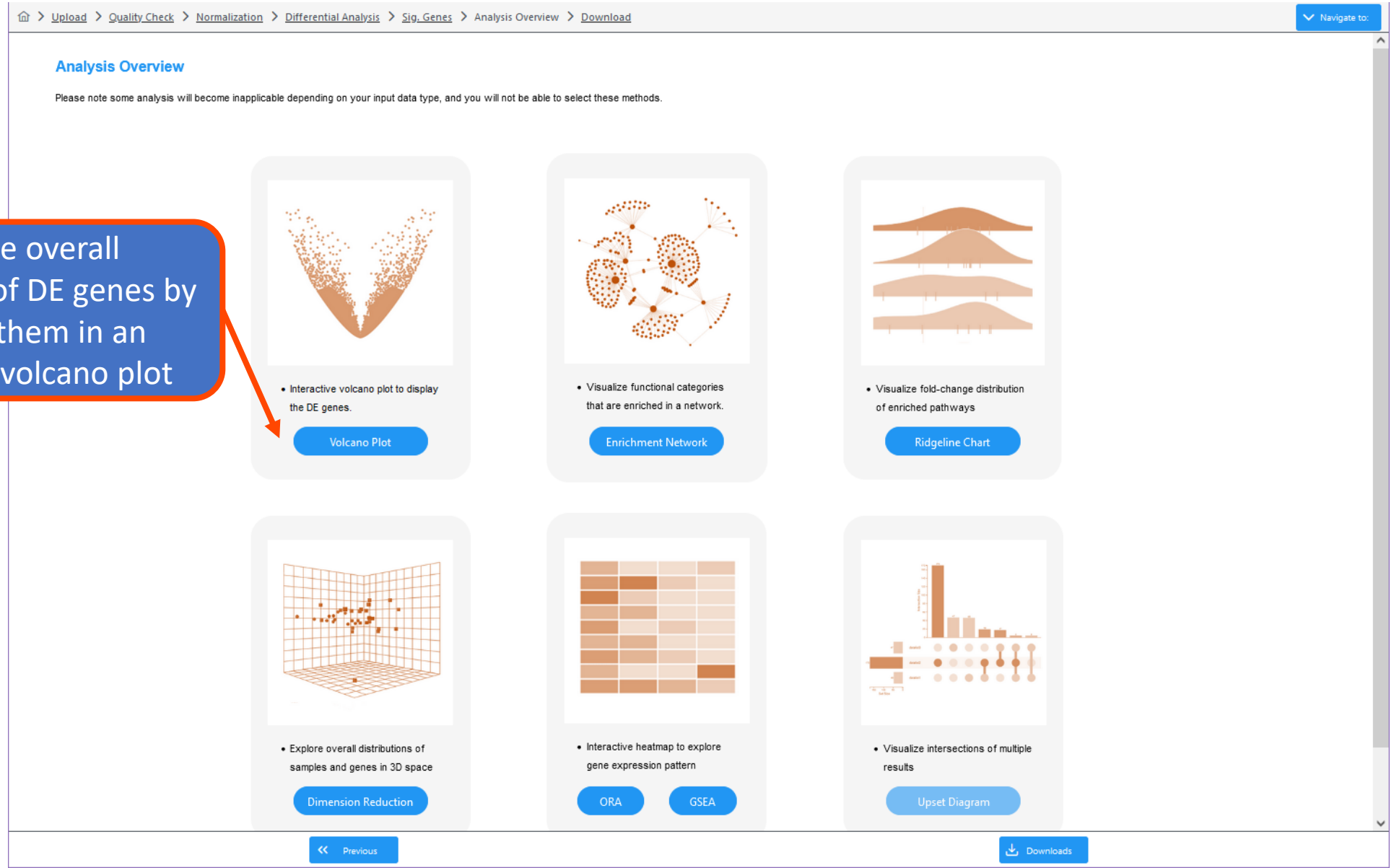
View differentially expressed genes (DEGs)

Here we see that 139 genes were significant according to default 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.



Visual analytics overview

Visualize overall distribution of DE genes by visualize them in an interactive volcano plot



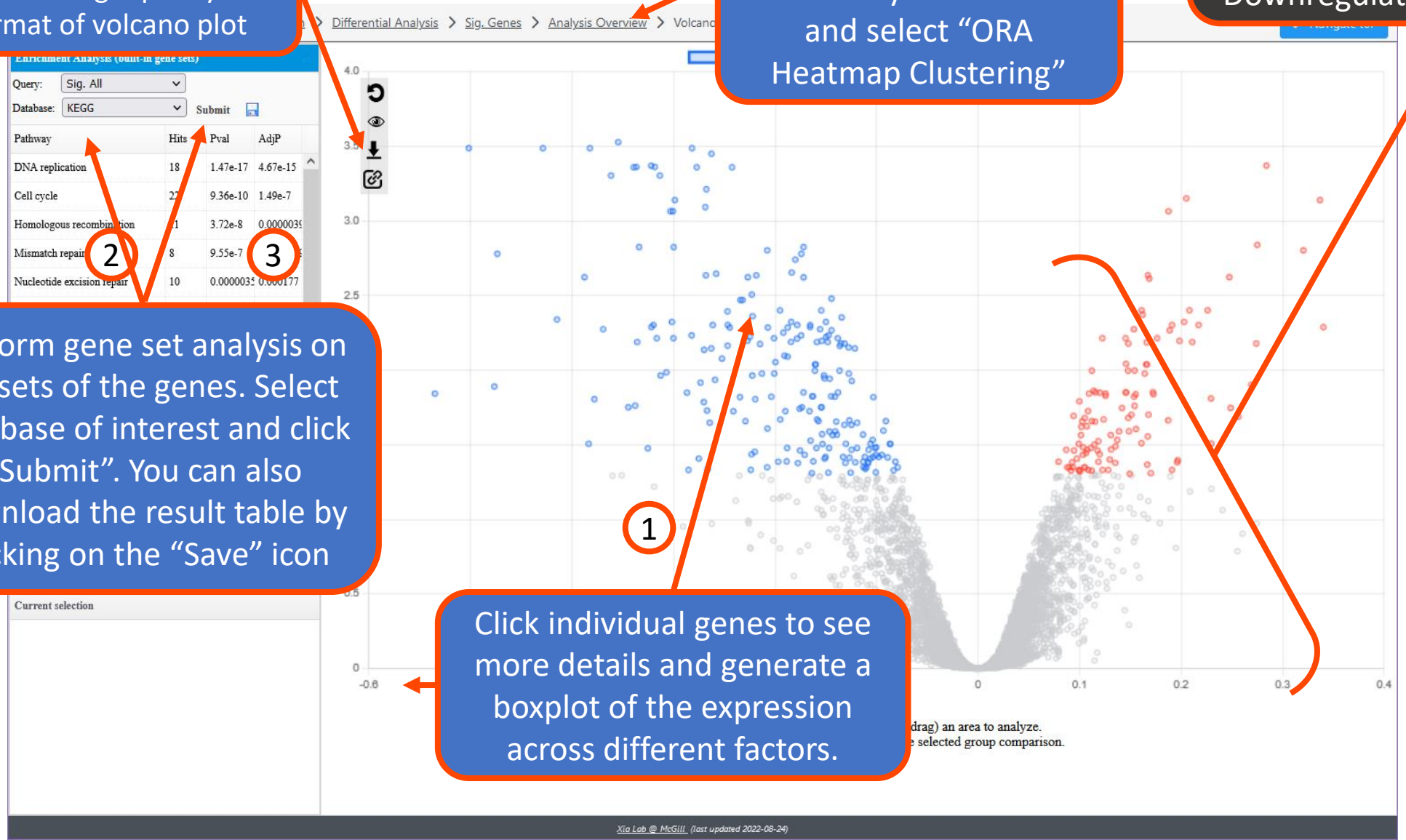
Interactive volcano plot

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



2

3

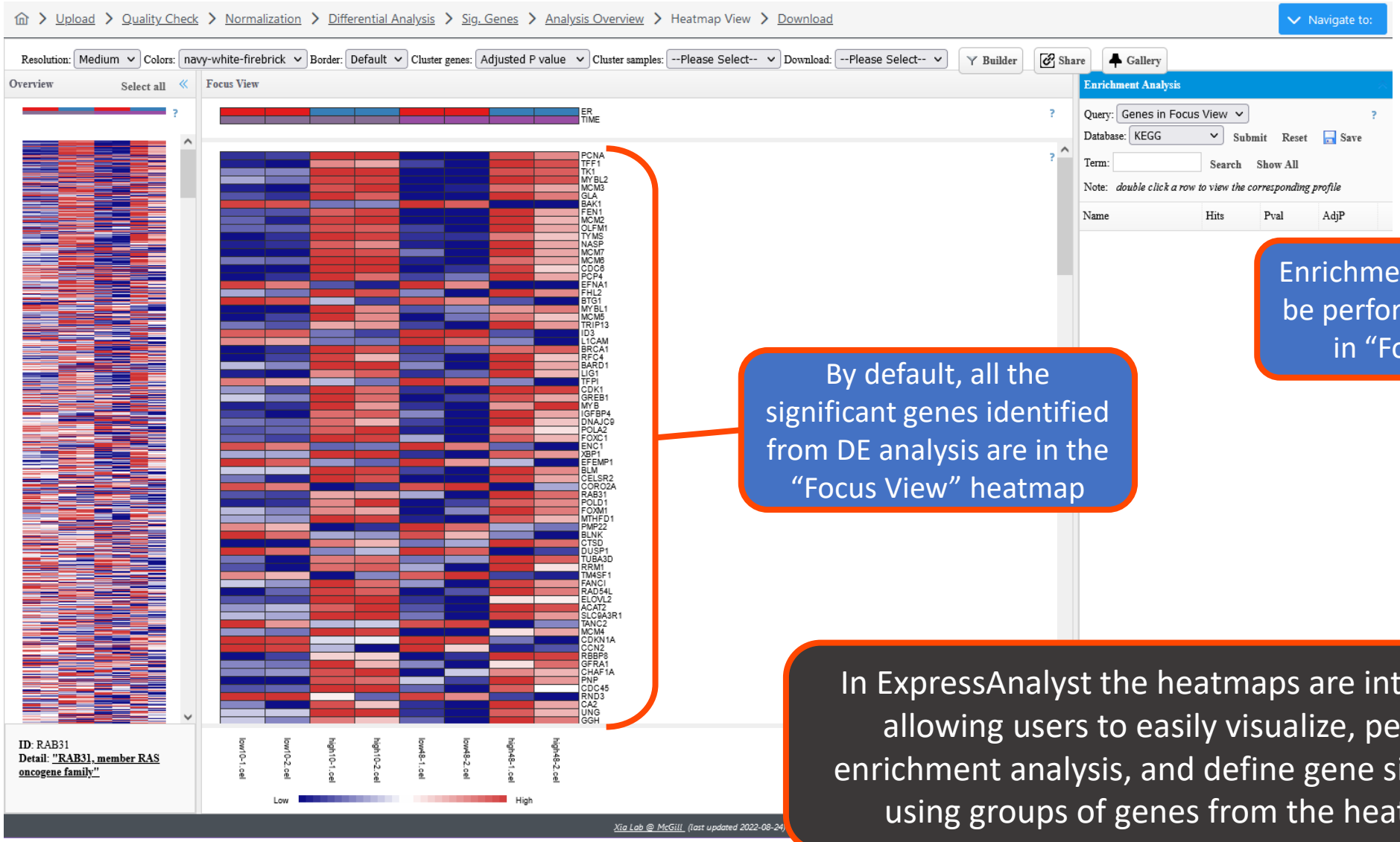
Perform gene set analysis on subsets of the genes. Select database of interest and click “Submit”. You can also download the result table by clicking on the “Save” icon

1

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

Interactive Heatmap

"Overview" heatmap contains all genes from data matrix



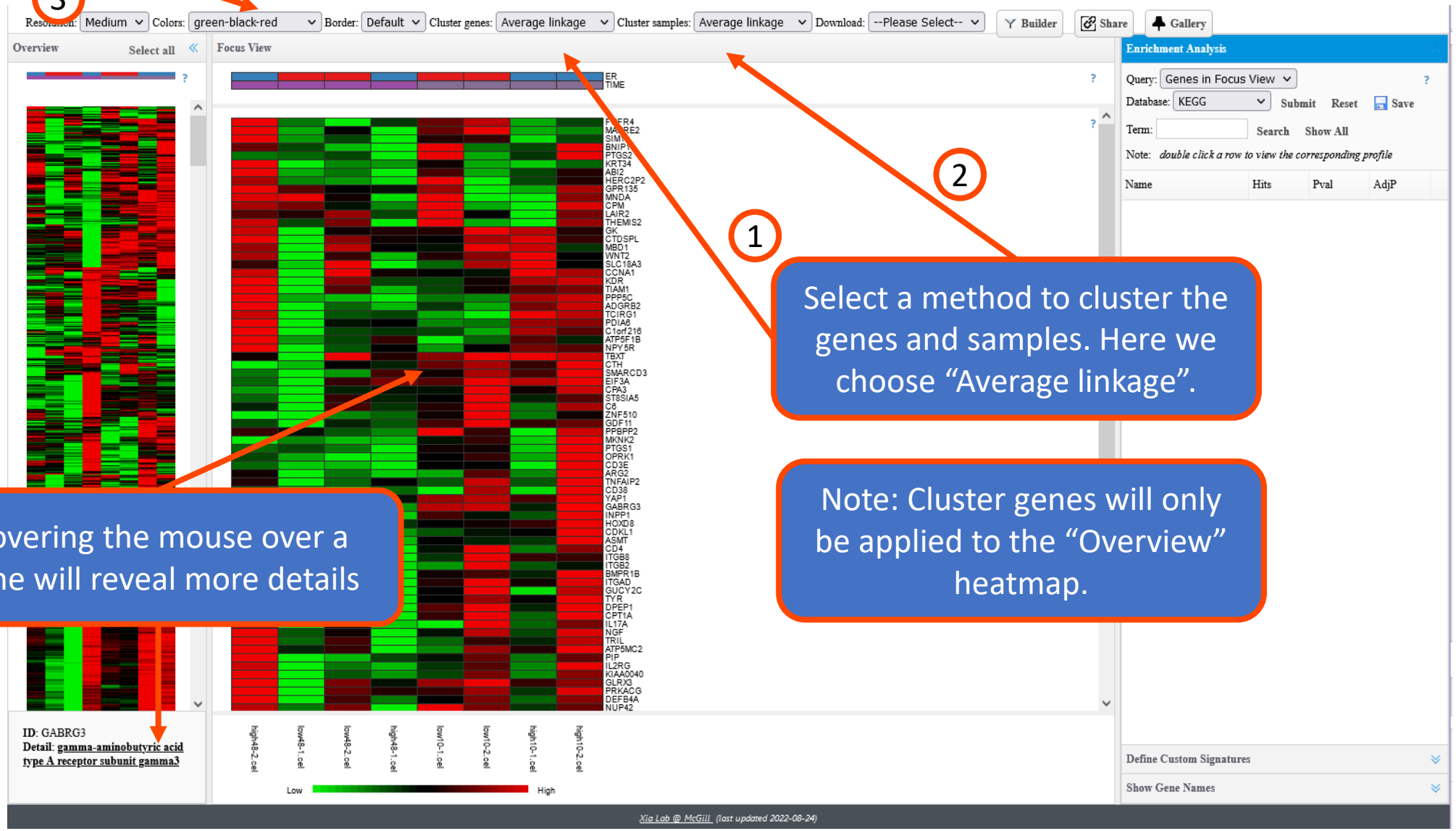
By default, all the significant genes identified from DE analysis are in the "Focus View" heatmap

Enrichment analysis can be performed on genes in "Focus view"

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

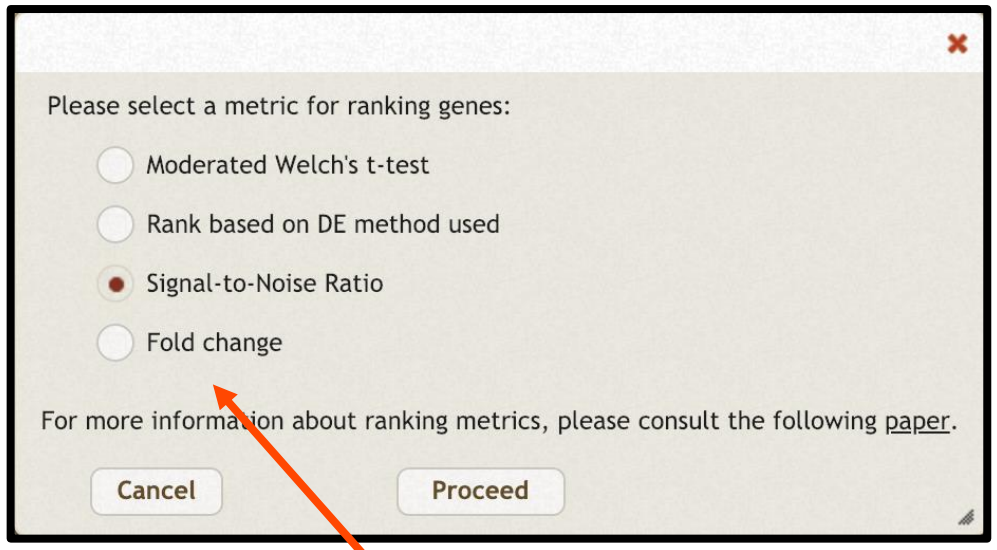
Interactive Heatmap

Change the color scheme to green-black-red



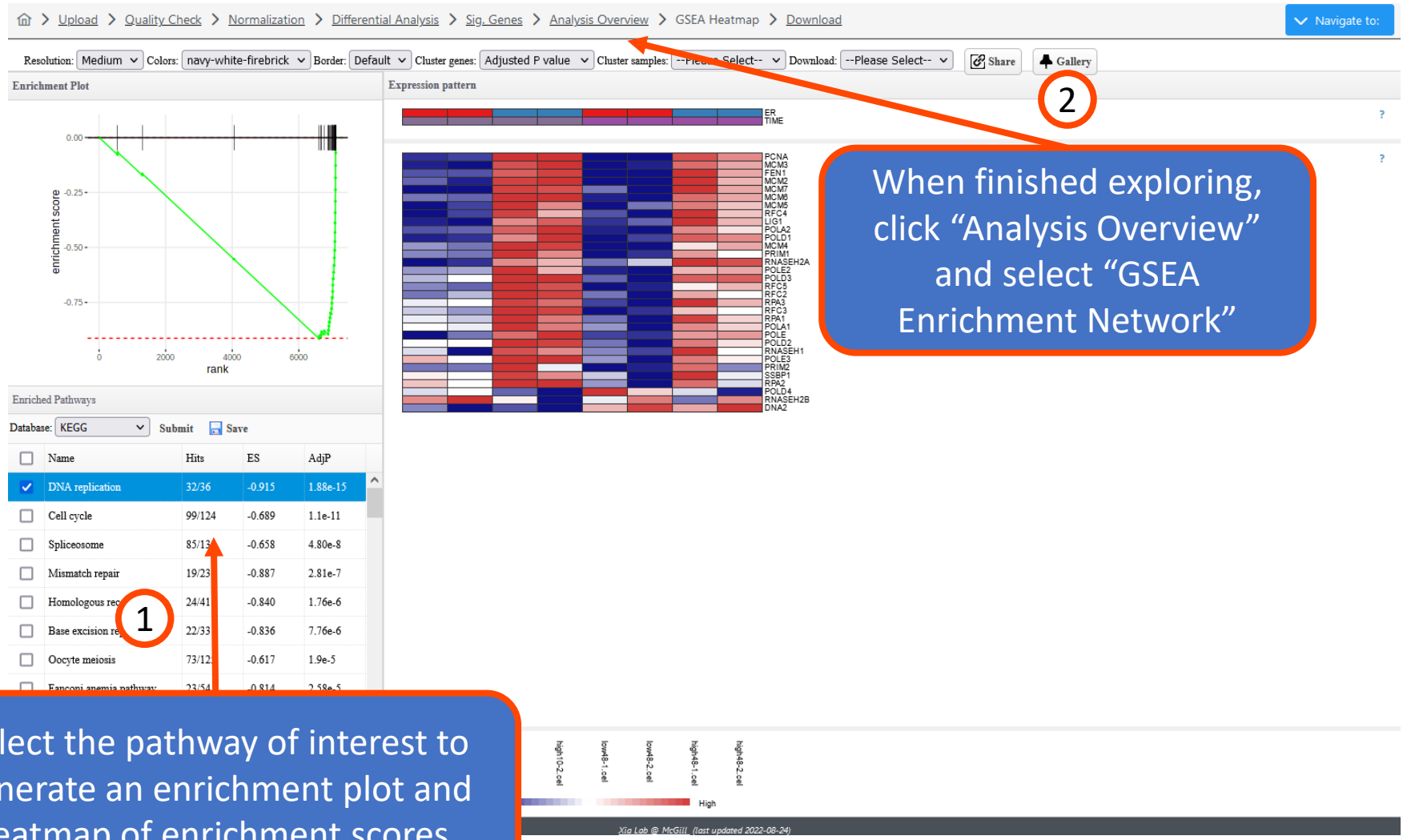
Gene Set Enrichment Analysis (GSEA)

- GSEA is a computational method for determining if the expression of a set of genes (biological pathways, etc.) is correlated with metadata.
- GSEA incorporates the gene expression data (as opposed to significant genes in the case of ORA) and so it can detect more sensitive differences.
- Refer to the original paper for more details on the GSEA:
  - https://www.pnas.org/content/102/43/15545.short



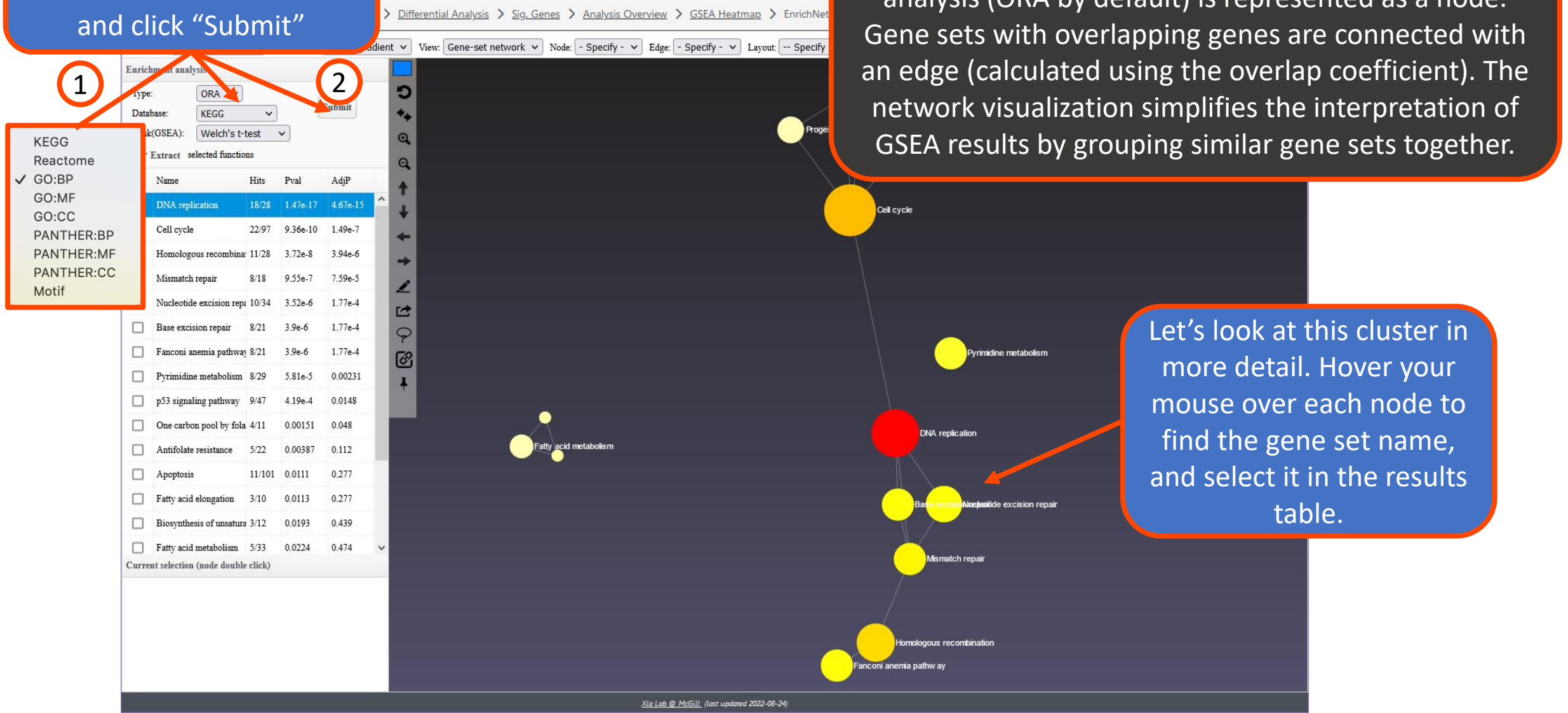
The first step in GSEA is to rank genes according to their expression. Try out several different methods – they should give similar results.

GSEA Heatmap Clustering



Enrichment Network

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

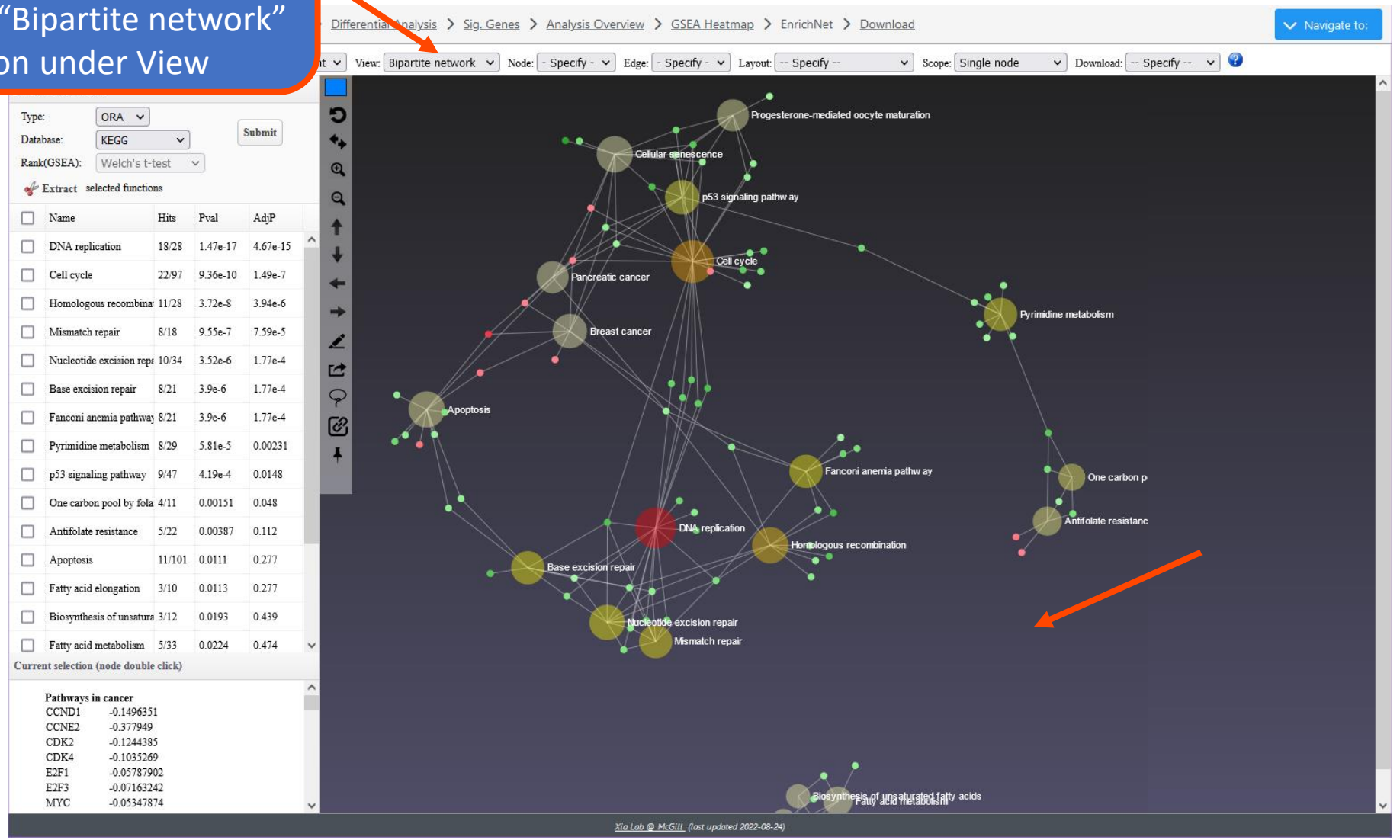


Each significantly enriched gene set from enrichment analysis (ORA by default) is represented as a node. Gene sets with overlapping genes are connected with an edge (calculated using the overlap coefficient). The network visualization simplifies the interpretation of GSEA results by grouping similar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

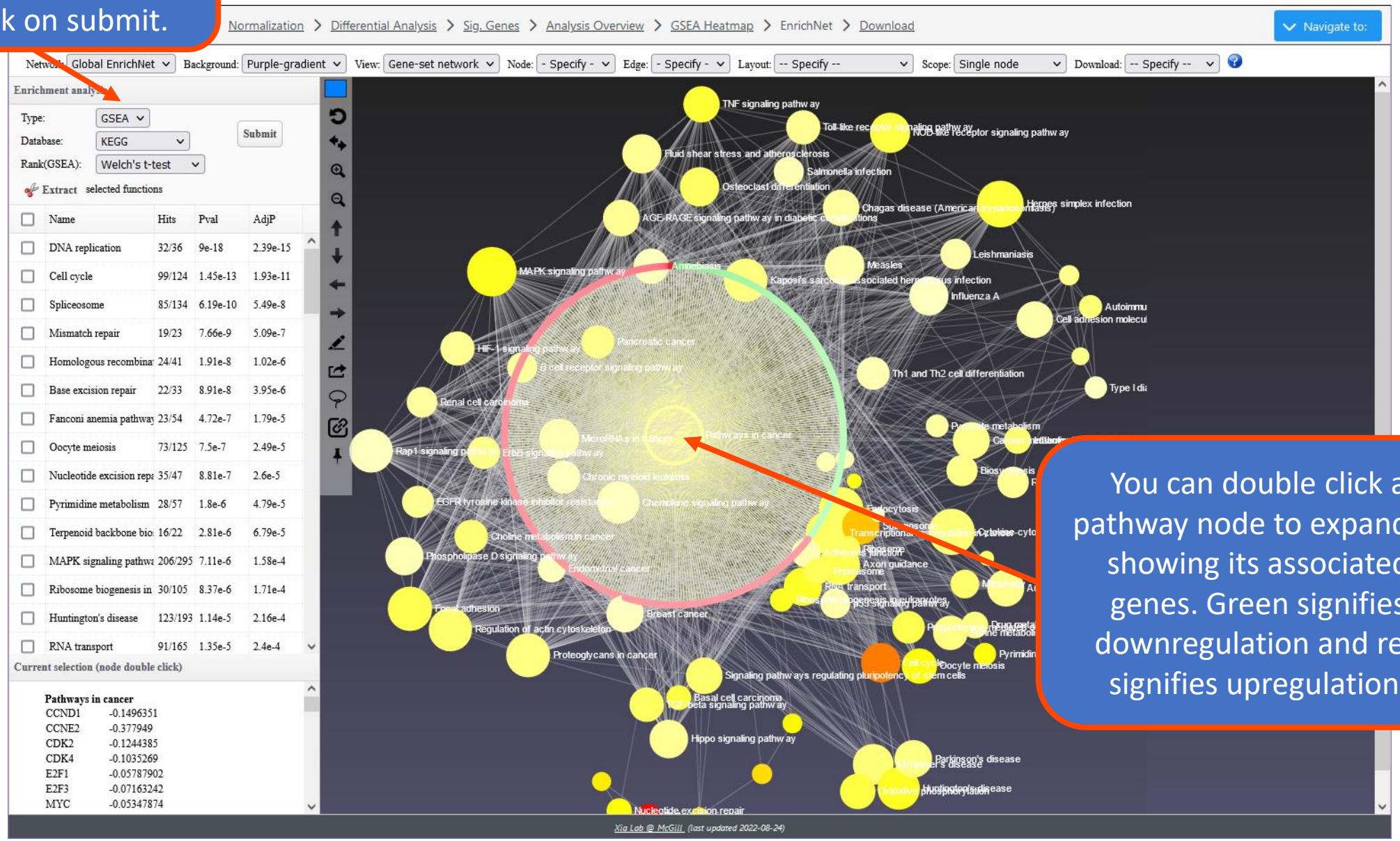
Enrichment Network

You can view all pathway nodes expanded, showing their related genes, by selecting “Bipartite network” option under View



Enrichment Network

Switch to GSEA enrichment network by selecting “GSEA” option and click on submit.



You can double click a pathway node to expand it, showing its associated genes. Green signifies downregulation and red signifies upregulation.

The End

For more information, visit Tutorials, Resources

and Contact pages on www.expressanalyst.ca

Also visit our forum for FAQs on www.omicsforum.ca