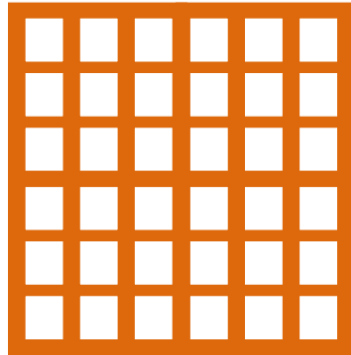


ExpressAnalyst - Tutorial

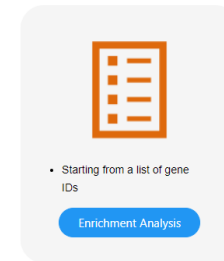
Starting from a table

-- Comprehensive platform for gene expression and meta-analysis

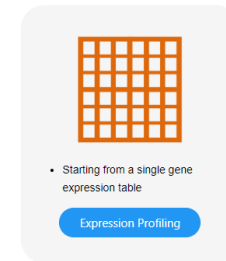


Intro to ExpressAnalyst

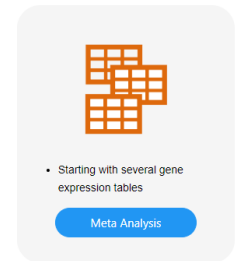
- Web platform for the analysis of gene expression data and meta-analysis
 - Previously part of NetworkAnalyst
- Designed for bench researchers rather than specialized bioinformaticians
- Integrates data processing, statistical analysis and data visualization to support:
 - Data comparisons
 - Biological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



Gene list



Single matrix



Meta-analysis

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

#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

Meta-data

Can be single or two metadata types

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
...								

Gene/probe ids

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

Navigation bar to track analysis progress

Data upload and annotation

Upload Download

Navigate to:

Upload a gene expression table

ExpressAnalyst currently supports gene expression profiling and functional analysis for 25 organisms based on user feedback. including 11 model species, 5 pathogens and 9 ecological species. In addition, ExpressAnalyst also supports generic annotation based on KEGG orthologs (KO), as well as custom annotation. If your organism is not within the list, leave the **organism unspecified**, and you can still perform basic expression profiling such as differential analysis, volcano plot, heatmap clustering, etc.

Upload your gene expression table

Specify organism

----Not specified----

Data type

----Not specified----

ID type

--- Not Specified ---

Gene-level summarization

Mean

Data File

+ Choose



Submit

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).

Try our example data

☒ Estrogen

Affymetrix Human Genome U95 GeneChip (hgu95av2) data, normalized, log 2 scale (8 samples)

Gene expression of a breast-cancer cell line ([source](#)) . **Estrogen Receptor (ER)**: present, absent; **Time (hour)**: 10, 48

☐ Endotoxin

Illu

Select "Estrogen" example data

in human PBMC using LPS as inducer
Treatment: Control, LPS, LPS_LPS; **Donor**: 21, 46,

☐ C. japonica toxicity

RNAseq data (10 samples)

response in C. japonica from an early life stage toxicity experiment **Treatment**: Control, Medium, High;

☐ DC cormorant toxicity

RNAseq data Seq2Fun ID, raw counts (14 samples)

Gene expression response in double-crested cormorant (DCCO) from an early life stage (embryos) toxicity experiment **Treatment**: Control, Medium, High;

Submit

Click on "Submit" then "Proceed"

<< Previous

>> Proceed

View processing results

[Home](#) > [Upload](#) > [Quality Check](#) > [Normalization](#) > [Download](#)

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:	12625
Matched gene number:	11774
Unmatched gene number:	851
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

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

Box plot

Count sum

PCA plot

Density plot



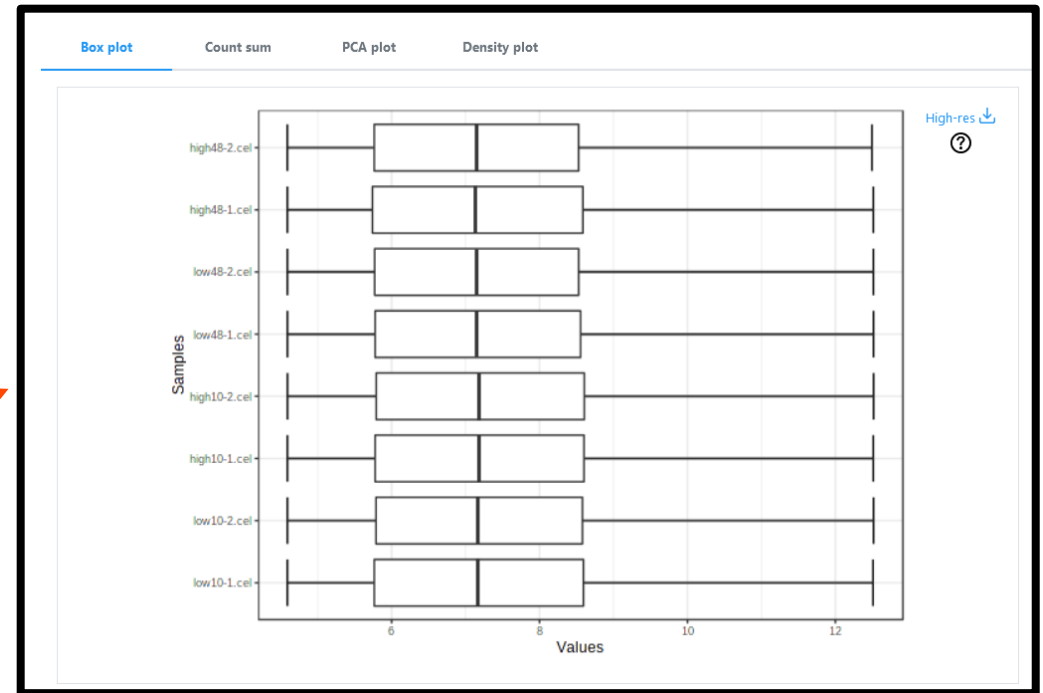
View common QA/QC plots to check the quality of the data

[Previous](#)

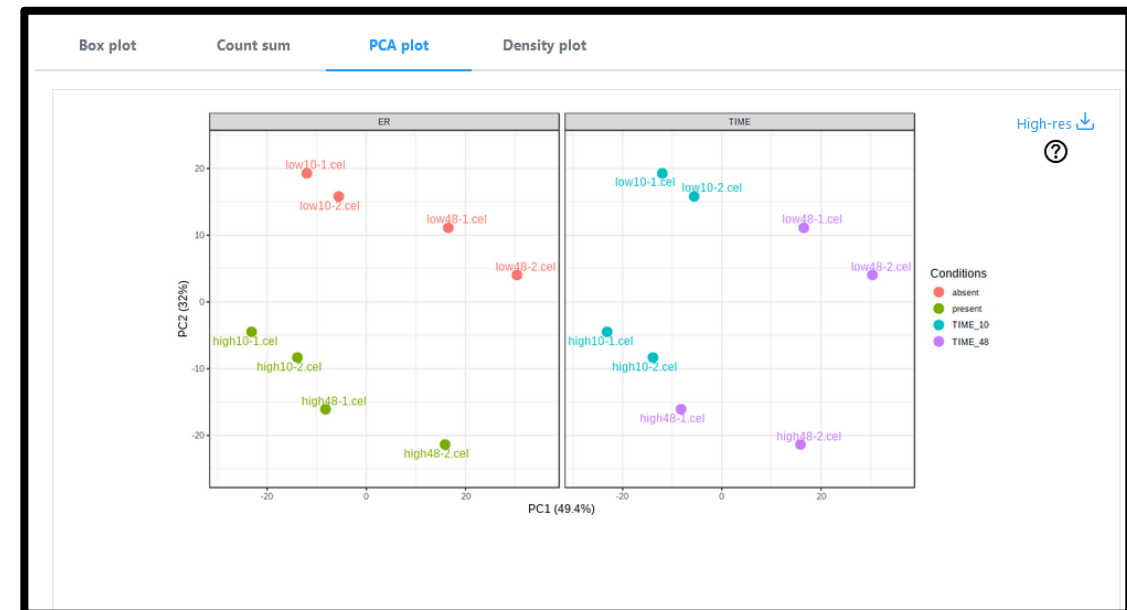
[Proceed](#)

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.



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.

Normalize and filter the data

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

The screenshot shows a web interface for 'Data Filtering & Normalization'. At the top is a breadcrumb trail: [Home](#) > [Upload](#) > [Quality Check](#) > [Normalization](#) > [Download](#). Below this is the title 'Data Filtering & Normalization' and a brief description: 'Filtering serves to remove data that are unlikely to be informative or simply erroneous. Normalization is crucial for a reliable detection of experiment.' The interface is divided into two main sections: 'Filtering' and 'Normalization'. The 'Filtering' section includes a 'Variance filter' slider set to 15, a 'Low abundance' slider set to 5, and a checkbox for 'Filter unannotated genes' which is checked. The 'Normalization' section has four radio button options: 'None' (selected), 'Log2 Transformation', 'Variance Stabilizing Normalization (VSN)', 'Quantile Normalization', and 'VSN followed by Quantile Normalization'. A blue 'Submit' button is located to the right of the 'Filtering' section. Below the settings is a preview area showing a heatmap with a dendrogram on the left and a 'High-res' download link. At the bottom right is a blue 'Proceed' button. Four callout boxes with orange arrows provide additional information: one explains the purpose of filtering, another explains the percentile-based filters, a third points to the 'Submit' button, and a fourth explains the normalization methods.

Filtering serves to remove data that are unlikely to be informative or simply erroneous. Normalization is crucial for a reliable detection of experiment.

Filtering:

- Variance filter: 15
- Low abundance: 5
- Filter unannotated genes: ☒

Normalization:

- ☒ None
- ☐ Log2 Transformation
- ☐ Variance Stabilizing Normalization (VSN)
- ☐ Quantile Normalization
- ☐ VSN followed by Quantile Normalization

Submit

High-res

Proceed

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”.

Navigation: [▼ Navigate to:](#)

Data Filtering & Normalization

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.

Filtering:

- Variance filter: 15 ?
- Low abundance: 5 ?
- Filter unannotated genes: ☒
- 1** ☐ None

Normalization:

- ☐ Log2 Transformation
- ☐ Variance Stabilizing Normalization (VSN)
- ☐ Quantile Normalization
- ☐ VSN followed by Quantile Normalization

2 Click “Submit” and “Proceed”

3

[Box plot](#) [PCA plot](#) [Density plot](#) [MSD plot](#)

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

High-res [?](#)

Conduct differential expression analysis

[Home](#) > [Upload](#) > [Quality Check](#) > [Normalization](#) > [Differential Analysis](#) > [Download](#)

▼ Navigate to:

Differential Expression Analysis

Statistical method

☒ Limma ☐ EdgeR ☐ DESeq2 ?

Study Design

Primary Factor ?

Secondary Factor This is a blocking factor ☐ ?

☒ Specific comparison versus

☐ Against a common control ?

Comparison of Interest

☐ Nested comparisons versus Interaction only ☒ ?

☐ Pairwise comparisons ?

☐ Time series ?

Submit

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

<< Previous

The last two statistical methods are available for RNAseq data

Conduct DE 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.

Home > Upload > Quality Check > Normalization > Differential Analysis > Download

Differential Expression Analysis

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 1

☐ Against a common control: absent ?

☐ Nested comparisons: absent vs. present versus absent vs. present Interaction only ☒ ?

☐ Pairwise comparisons ?

☐ Time series ?

Submit 2

3

Click "Submit" and "Proceed"

<< Previous >> Proceed

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.

total sig. genes: 139

Download

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

The table below shows at most top 1000 genes ranked by p-values. Use the **Download Result** link above to get the whole result table. Significant genes are in orange.

Gene	View Details	logFC	AveExpr	t	P.Value	adj.P.Val	B
PCNA	NCBI	-2.2355	9.1368	-15.016	3.214E-8	1.6251E-4	9.2032
TK1	NCBI	-2.8983	9.8509	-13.954	6.5173E-8	1.6251E-4	8.6124
MYBL2	NCBI	-2.9243	8.5321	-13.661	7.992E-8	1.6251E-4	8.4384
TFF1	NCBI	-3.1988	12.116	-13.083	1.2089E-7		8.0808
GLA	NCBI	-1.5815	8.7099	-12.791	1.4996E-7		7.8924
ID3	NCBI	1.4969	11.000	11.000	7.000E-7		7.8777
BAK1	NCBI	1.7522	11.000	11.000	7.000E-7		7.8353
MCM3	NCBI	-1.5599	11.000	11.000	7.000E-7		7.4118
MCM7	NCBI	-2.1023	11.000	11.000	7.000E-7		7.2125

You can view individual gene expression pattern by clicking on the image icon



<< Previous

Proceed >>


Visual analytics overview

[Home](#) > [Upload](#) > [Quality Check](#) > [Normalization](#) > [Differential Analysis](#) > [Sig. Genes](#) > Analysis Overview > [Download](#)

[Navigate to:](#)


Analysis Overview

Please note some analysis will become inapplicable depending on your input data type, and you will not be able to select these methods.



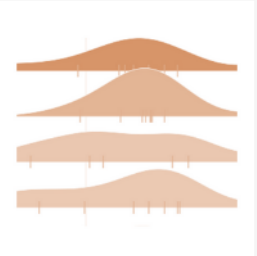
- Interactive volcano plot to display the DE genes.

[Volcano Plot](#)



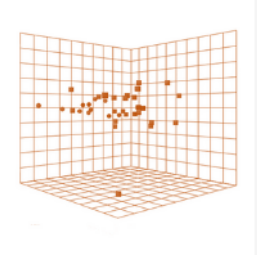
- Visualize functional categories that are enriched in a network.

[Enrichment Network](#)



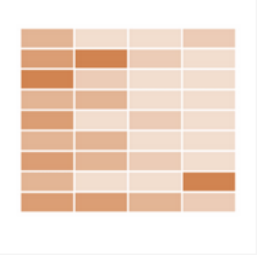
- Visualize fold-change distribution of enriched pathways

[Ridgeline Chart](#)



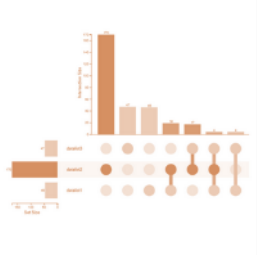
- Explore overall distributions of samples and genes in 3D space

[Dimension Reduction](#)



- Interactive heatmap to explore gene expression pattern

[ORA](#) [GSEA](#)



- Visualize intersections of multiple results

[Upset Diagram](#)

[<< Previous](#) [Downloads](#)

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

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

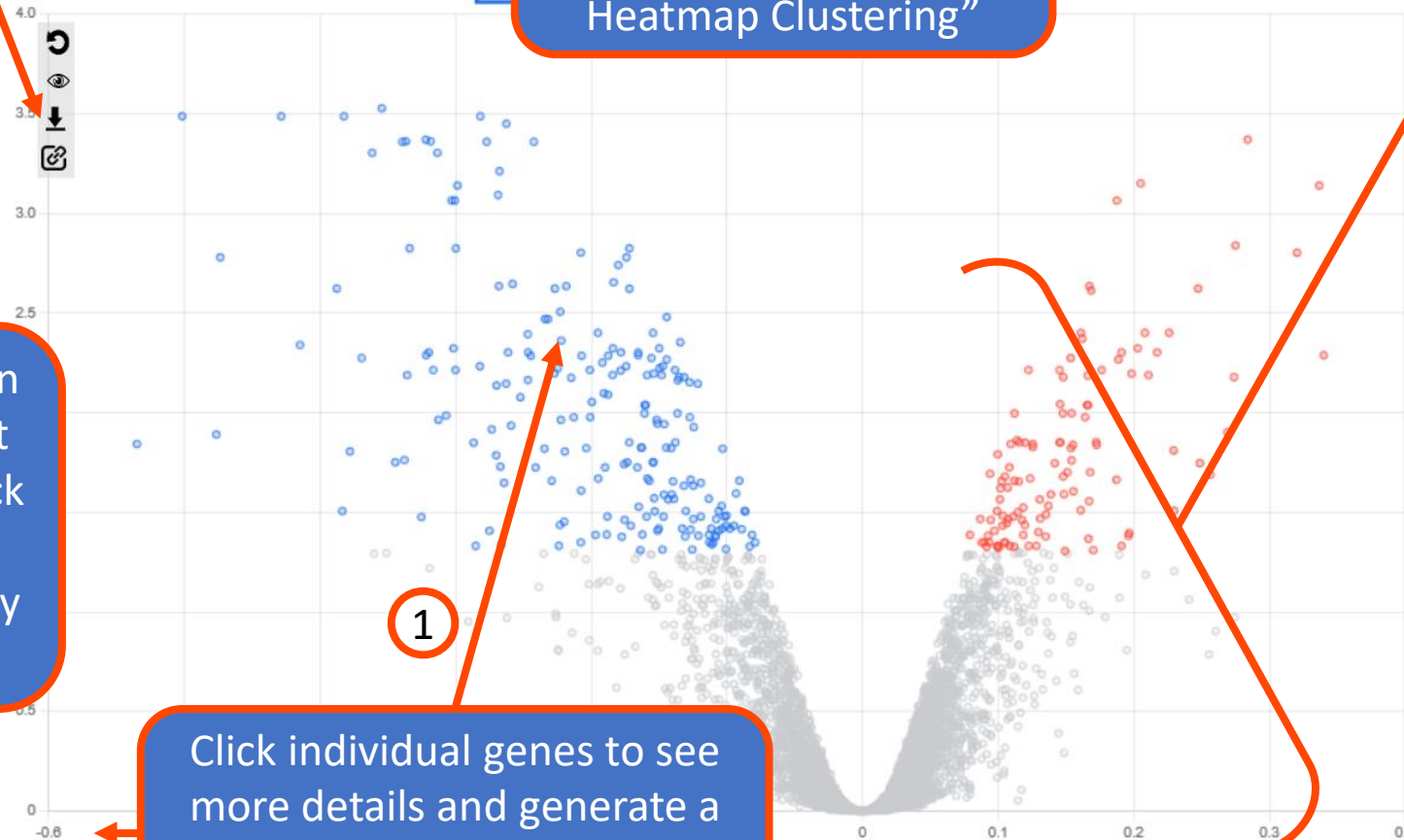
Perform gene set analysis on subsets of the genes. Select database of interest and click "Submit". You can also download the result table by clicking on the "Save" icon

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

Enrichment Analysis (count in gene sets)

Query: Sig. All Database: KEGG Submit

Pathway	Hits	Pval	AdjP
DNA replication	18	1.47e-17	4.67e-15
Cell cycle	27	9.36e-10	1.49e-7
Homologous recombination	11	3.72e-8	0.0000035
Mismatch repair	8	9.55e-7	0.0000035
Nucleotide excision repair	10	0.0000035	0.000177



Interactive Heatmap

“Overview”
heatmap contains
all genes from data
matrix



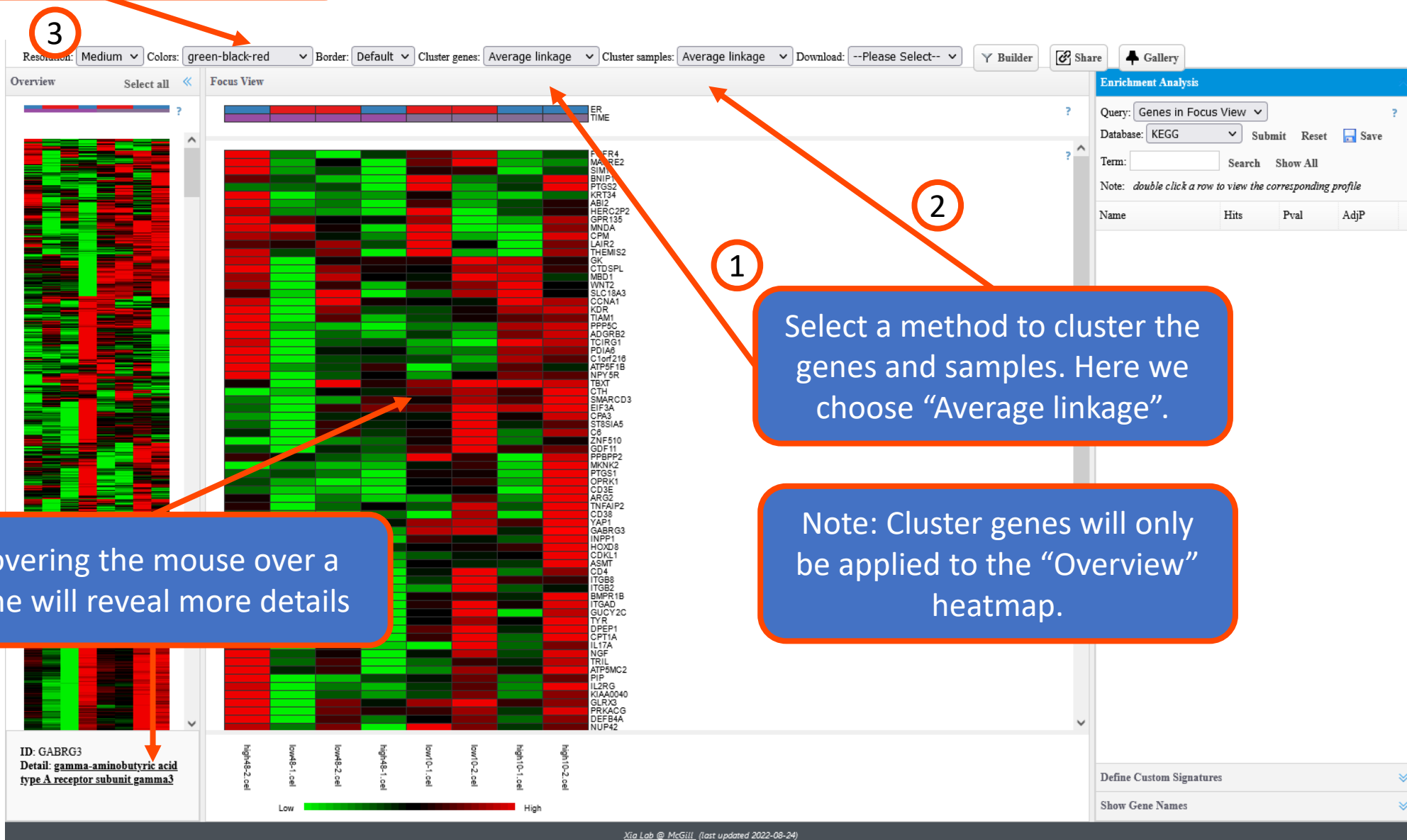
By default, all the
significant genes identified
from DE analysis are in the
“Focus View” heatmap

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.

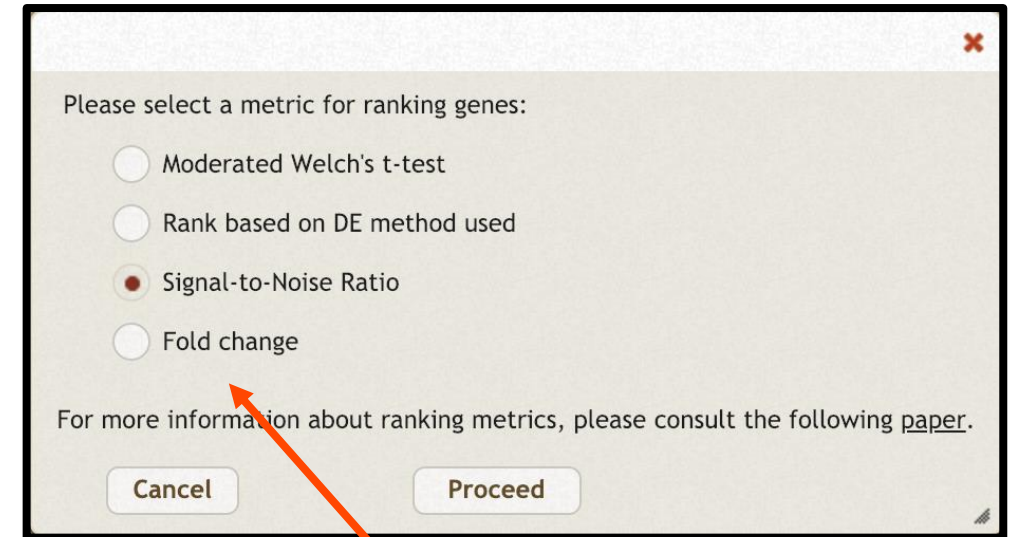
Change the color scheme to green-black-red

Interactive Heatmap



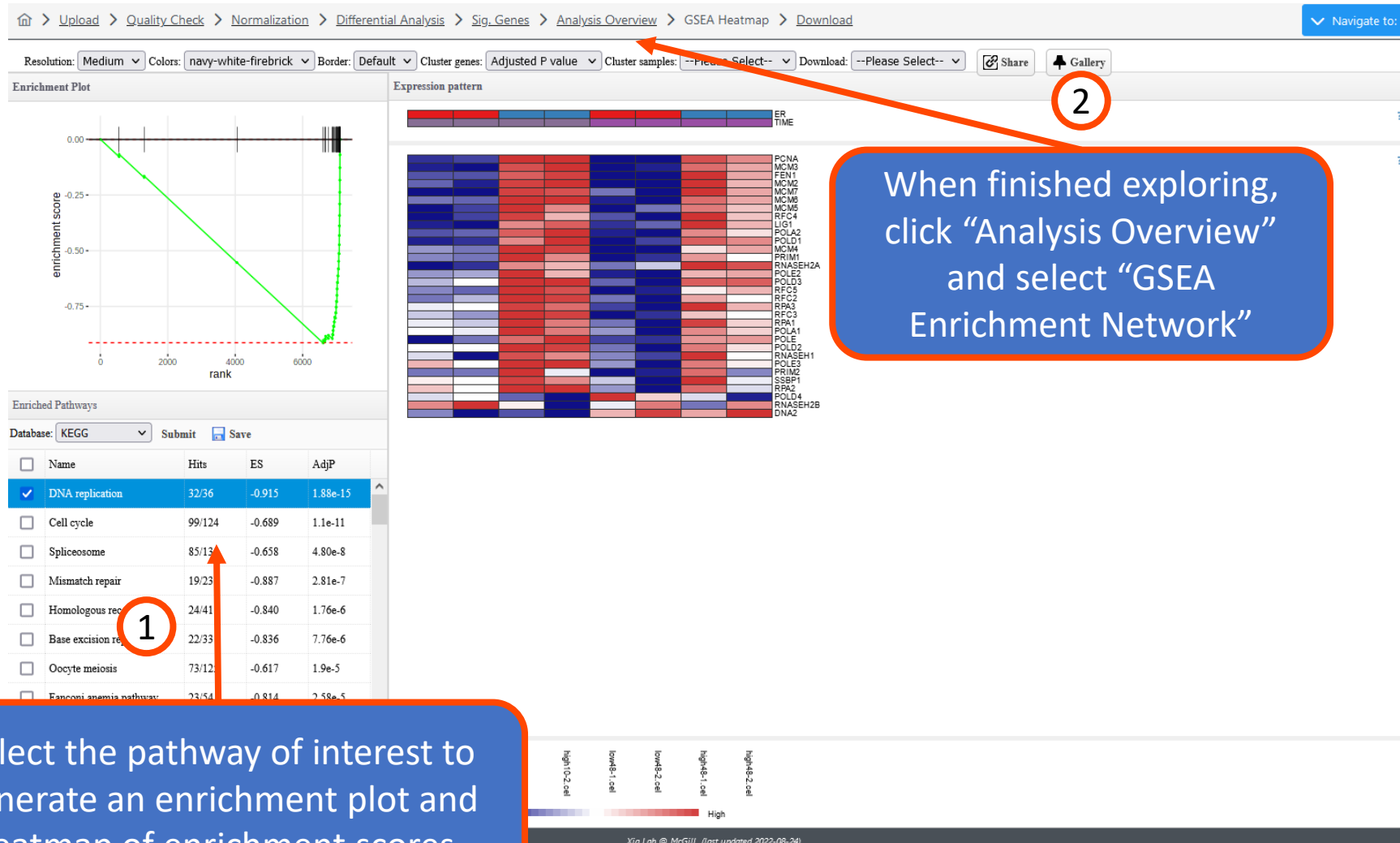
Gene Set Enrichment Analysis (GSEA)

- 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"

1

2

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

Enrichment analysis

Type: **ORA**

Database: **KEGG**

Method (GSEA): **Welch's t-test**

Submit

Extract selected functions

Name	Hits	Pval	AdjP
DNA replication	18/28	1.47e-17	4.67e-15
Cell cycle	22/97	9.36e-10	1.49e-7
Homologous recombination	11/28	3.72e-8	3.94e-6
Mismatch repair	8/18	9.55e-7	7.59e-5
Nucleotide excision repair	10/34	3.52e-6	1.77e-4
<input type="checkbox"/> Base excision repair	8/21	3.9e-6	1.77e-4
<input type="checkbox"/> Fanconi anemia pathway	8/21	3.9e-6	1.77e-4
<input type="checkbox"/> Pyrimidine metabolism	8/29	5.81e-5	0.00231
<input type="checkbox"/> p53 signaling pathway	9/47	4.19e-4	0.0148
<input type="checkbox"/> One carbon pool by folate	4/11	0.00151	0.048
<input type="checkbox"/> Antifolate resistance	5/22	0.00387	0.112
<input type="checkbox"/> Apoptosis	11/101	0.0111	0.277
<input type="checkbox"/> Fatty acid elongation	3/10	0.0113	0.277
<input type="checkbox"/> Biosynthesis of unsaturated fatty acids	3/12	0.0193	0.439
<input type="checkbox"/> Fatty acid metabolism	5/33	0.0224	0.474

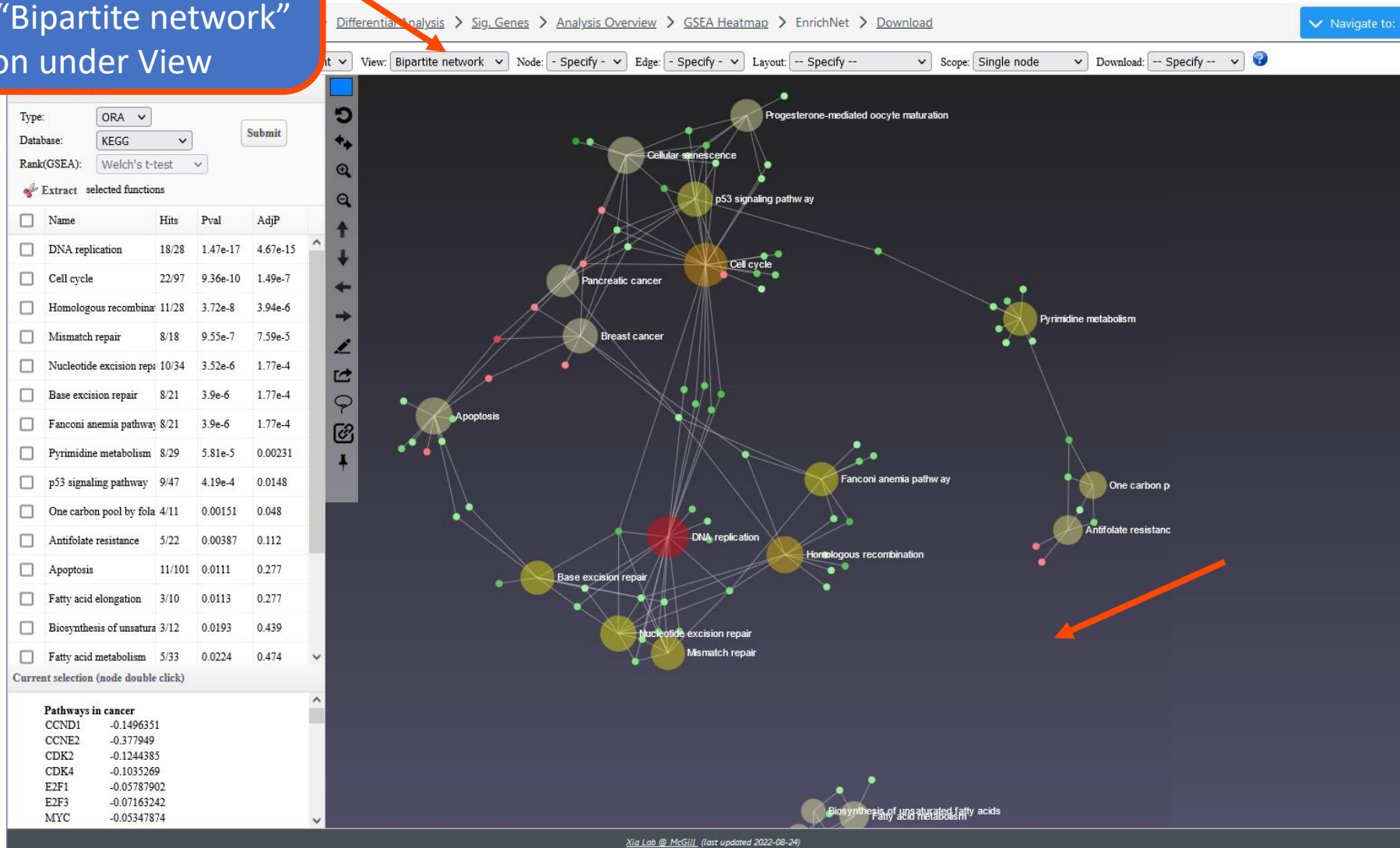
Current selection (node double click)

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.

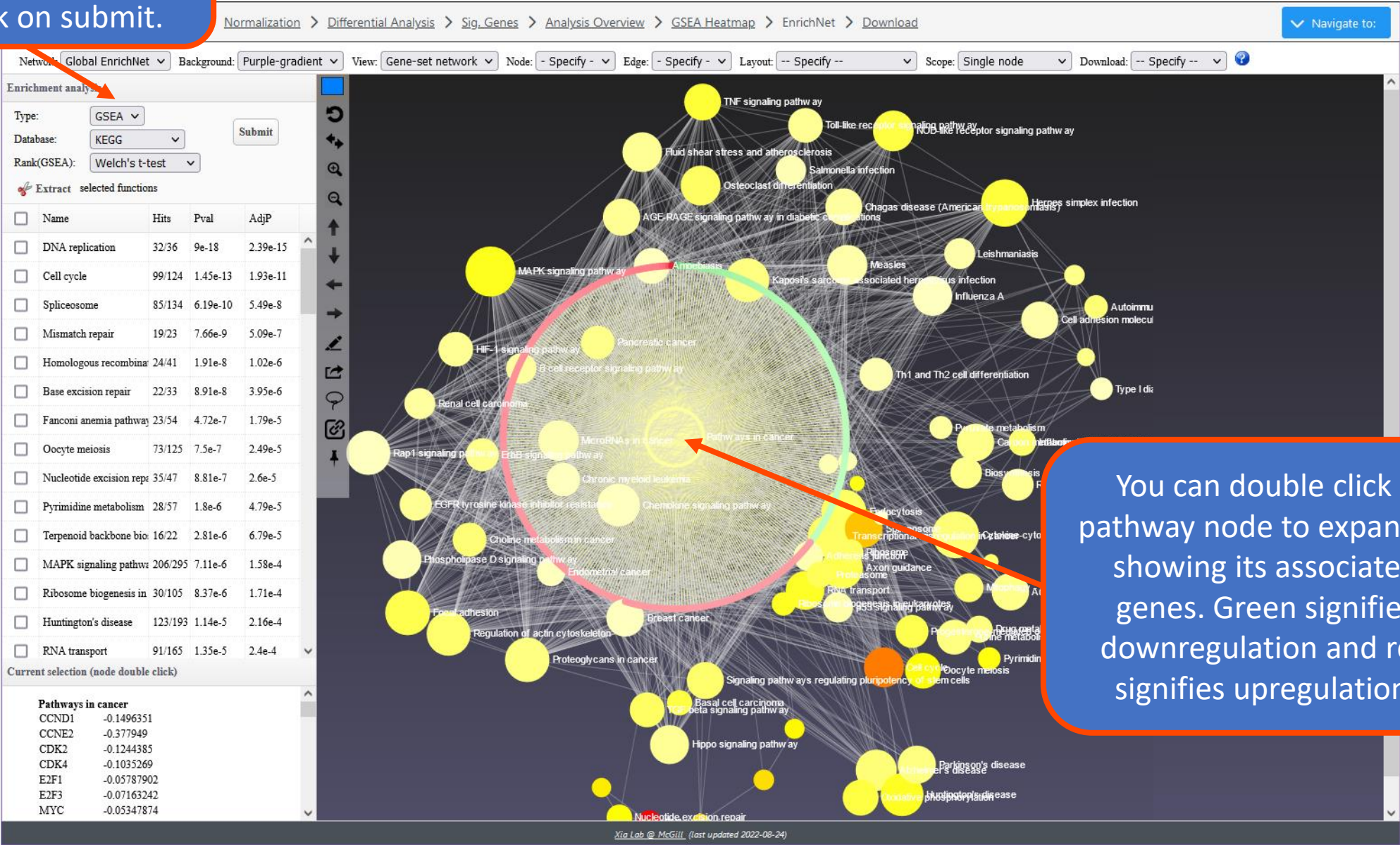
Enrichment Network

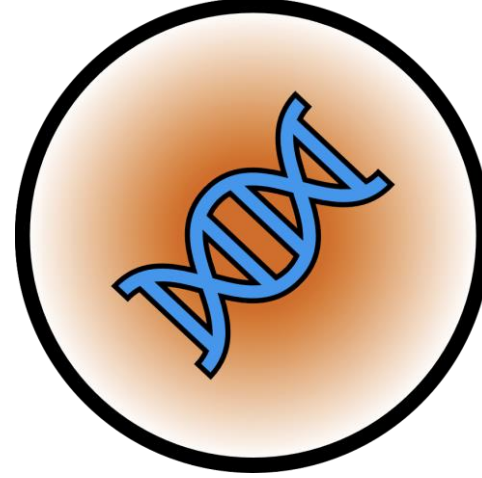
You can view all pathway nodes expanded, showing their related genes, by selecting “Bipartite network” option under View



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

Enrichment Network





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*