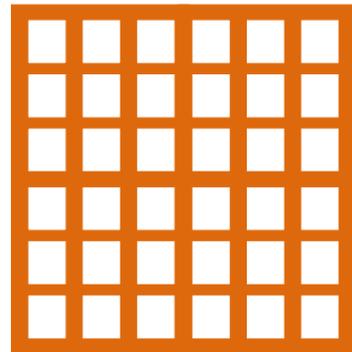


# ExpressAnalyst - Tutorial

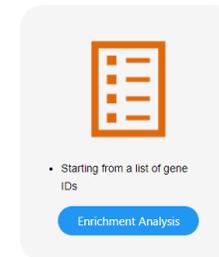
## Starting from a table

-- Comprehensive platform for gene expression and meta-analysis

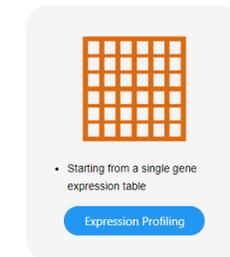


# Intro to ExpressAnalyst

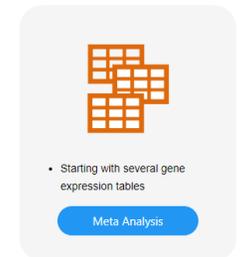
- Web platform for the analysis of gene expression data and meta-analysis
  - Previously part of NetworkAnalyst
- Designed for bench researchers rather than specialized bioinformaticians
- Integrates data processing, statistical analysis and data visualization to support:
  - Data comparisons
  - Biological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

Meta-data

#CLASS:ER absent absent present present absent absent present present

#CLASS:TIME 10 10 10 10 48 48 48 48

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

Home > 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

### 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...  
... in human PBMC using LPS as inducer  
**Treatment**: Control, LPS, LPS\_LPS; **Donor**: 21, 46,
- [C. japonica toxicity](#)  
RNASeq... 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

Previous Proceed

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 on "Submit" then "Proceed"

Select "Estrogen" example data

# 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.

<b>Data type:</b>	Microarray gene expression
<b>Total feature number:</b>	12625
<b>Matched gene number:</b>	11774
<b>Unmatched gene number:</b>	851
<b>Percent matched:</b>	93.3
<b>Sample number:</b>	8
<b>Number of experimental factors:</b>	2
<b>Group names:</b>	Two factors found - ER: absent; present TIME: TIME_10; TIME_48

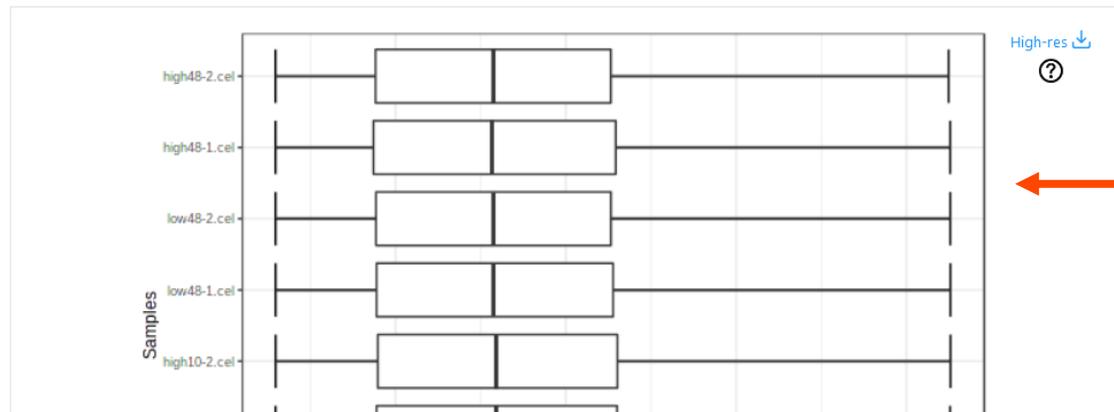
Check the processing results to ensure correct sample size, experimental factors, and adequate gene annotation

**Box plot**

Count sum

PCA plot

Density plot



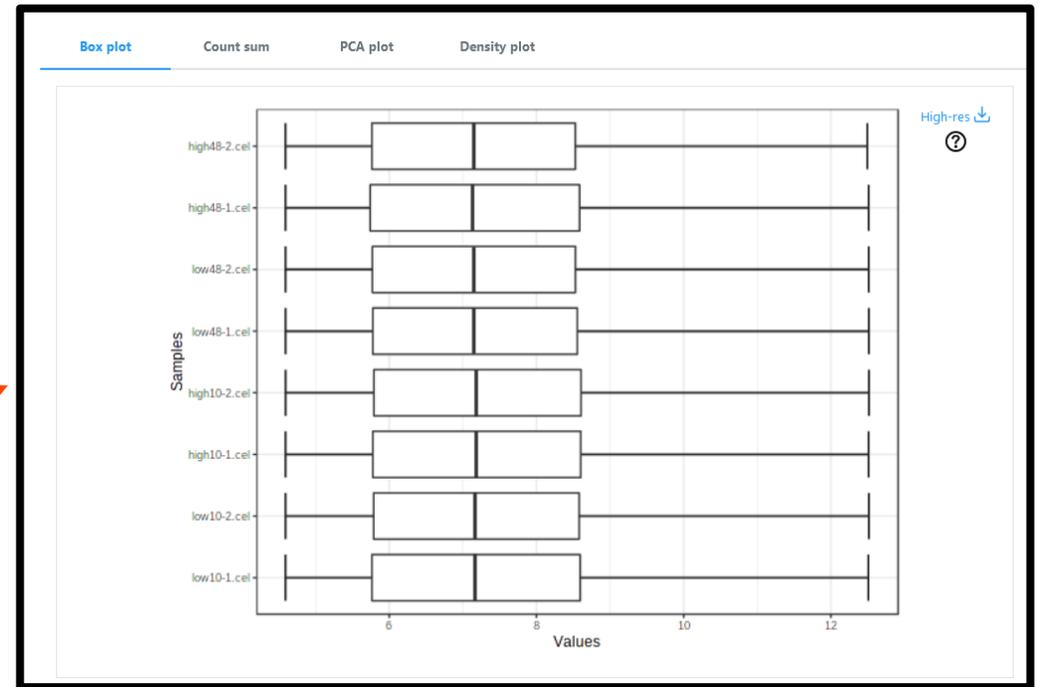
View common QA/QC plots to check the quality of the data

<< Previous

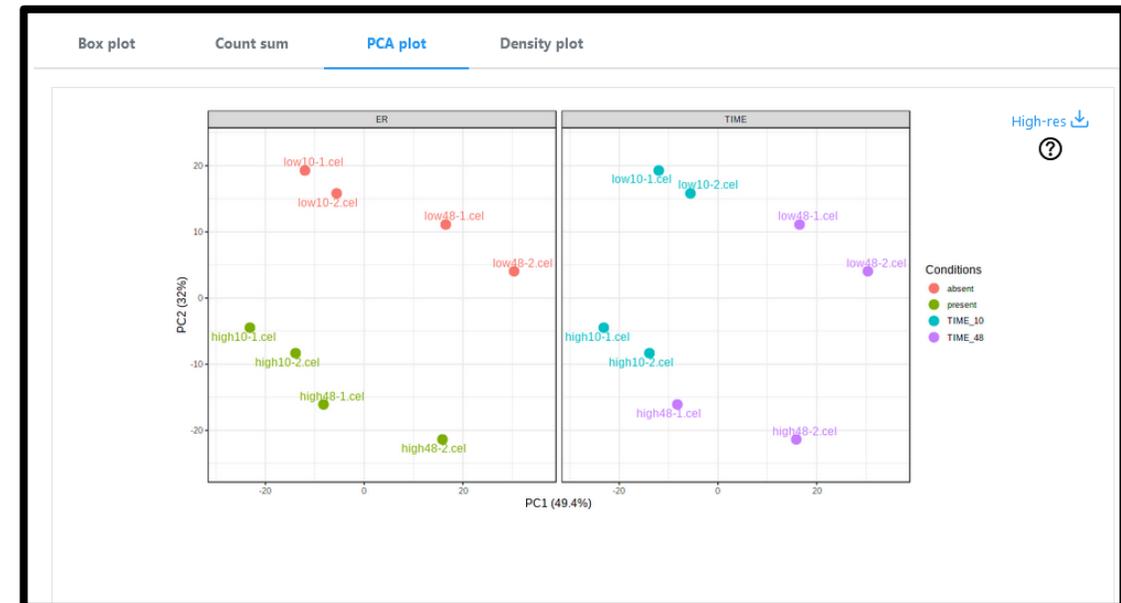
Next >> Proceed

# 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.



# 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 15<sup>th</sup> 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.

The screenshot shows a web interface for data filtering and normalization. The breadcrumb navigation at the top reads: Home > Upload > Quality Check > Normalization > Download. The main heading is "Data Filtering & Normalization". Below this, there is a brief explanation: "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".

**Filtering:**

- Variance filter: A slider set to 15 with a help icon.
- Low abundance: A slider set to 5 with a help icon.
- Filter unannotated genes: A checked checkbox.

**Normalization:**

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

A blue "Submit" button is located to the right of the filtering options. At the bottom right of the interface, there is a "High-res" download icon and a "Proceed" button.

# 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”.

The screenshot displays a web-based interface for data filtering and normalization. At the top right, there is a 'Navigate to:' dropdown menu. The main heading is 'Data Filtering & Normalization'. Below this, a paragraph explains that filtering removes uninformative or erroneous data, and normalization is crucial for reliable detection of transcriptional differences. The interface is divided into two main sections: 'Filtering' and 'Normalization'. In the 'Filtering' section, there are three sliders: 'Variance filter' (set to 15), 'Low abundance' (set to 5), and 'Filter unannotated genes' (checked). In the 'Normalization' section, there are 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 configuration options, there are four tabs: 'Box plot', 'PCA plot', 'Density plot', and 'MSD plot'. The 'Box plot' tab is active, showing a plot with three box plots for samples 'high48-2.cel', 'high48-1.cel', and 'low48-2.cel'. A 'High-res' download icon is visible in the top right of the plot area. At the bottom of the interface, there are two blue buttons: '<< Previous' and '>> Proceed'.

1

2

3

Click “Submit” and “Proceed”

# 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  ?

### Comparison of Interest

Specific comparison  versus

Against a common control  ?

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

# 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.

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

## 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 ?

Set ER as the primary factor

1

2

Click "Submit" and "Proceed"

3

<< Previous

Next >>

# 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.

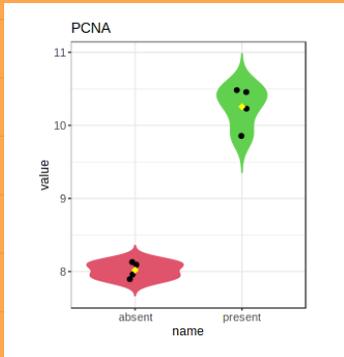
total sig. genes: 139 [Download](#)

1

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 $\updownarrow$	View Details	logFC $\updownarrow$	AveExpr $\updownarrow$	t $\updownarrow$	P.Value $\updownarrow$	adj.P.Val $\updownarrow$	B $\updownarrow$
PCNA	<a href="#">NCBI</a>	-2.2355	9.1368	-15.016	3.214E-8	1.6251E-4	9.2032
TK1	<a href="#">NCBI</a>	-2.8983	9.8509	-13.954	6.5173E-8	1.6251E-4	8.6124
MYBL2	<a href="#">NCBI</a>	-2.9243	8.5321	-13.661	7.992E-8	1.6251E-4	8.4384
TFF1	<a href="#">NCBI</a>	-3.1988	12.116	-13.083	1.2089E-7		8.0808
GLA	<a href="#">NCBI</a>	-1.5815	8.7099	-12.791	1.4996E-7		7.8924
ID3	<a href="#">NCBI</a>	1.4969	11.0000	11.0000	7.0000E-8		7.8777
BAK1	<a href="#">NCBI</a>	1.7522	11.0000	11.0000	7.0000E-8		7.8353
MCM3	<a href="#">NCBI</a>	-1.5599	11.0000	11.0000	7.0000E-8		7.4118
MCM7	<a href="#">NCBI</a>	-2.1023	11.0000	11.0000	7.0000E-8		7.2125



2

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

<< Previous

Proceed >>

# Visual analytics overview

🏠 > Upload > Quality Check > Normalization > Differential Analysis > Sig\_Genes > Analysis Overview > Download

Navigation: [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.



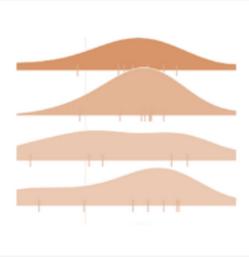
- Interactive volcano plot to display the DE genes.

Volcano Plot



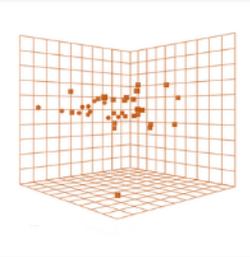
- Visualize functional categories that are enriched in a network.

Enrichment Network



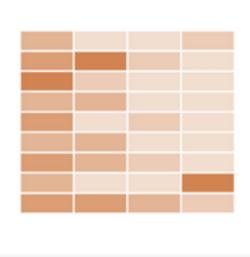
- Visualize fold-change distribution of enriched pathways

Ridgeline Chart



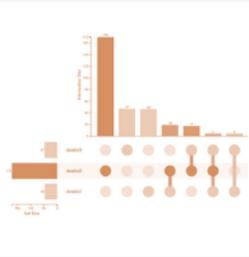
- Explore overall distributions of samples and genes in 3D space

Dimension Reduction



- Interactive heatmap to explore gene expression pattern

ORA   GSEA



- Visualize intersections of multiple results

Upset Diagram

Navigation: [Previous](#)   [Downloads](#)

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**

Enrichment Analysis (query in gene sets)

Query: Sig. All

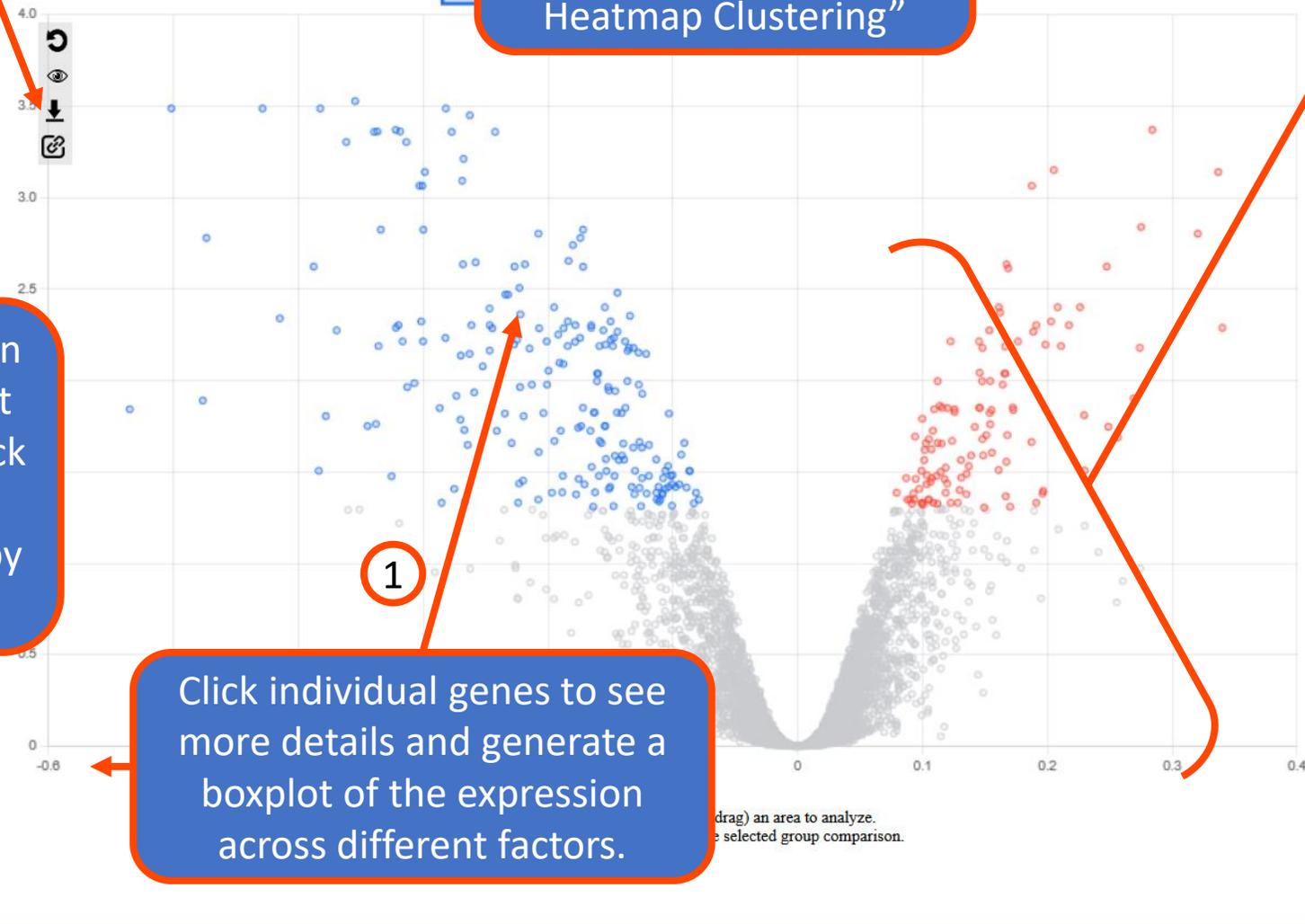
Database: KEGG

Pathway	Hits	Pval	AdjP
DNA replication	18	1.47e-17	4.67e-15
Cell cycle	27	9.36e-10	1.49e-7
Homologous recombination	1	3.72e-8	0.0000035
Mismatch repair	8	9.55e-7	0.0000035
Nucleotide excision repair	10	0.0000035	0.000177

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.

1



# Interactive Heatmap

“Overview” heatmap contains all genes from data matrix



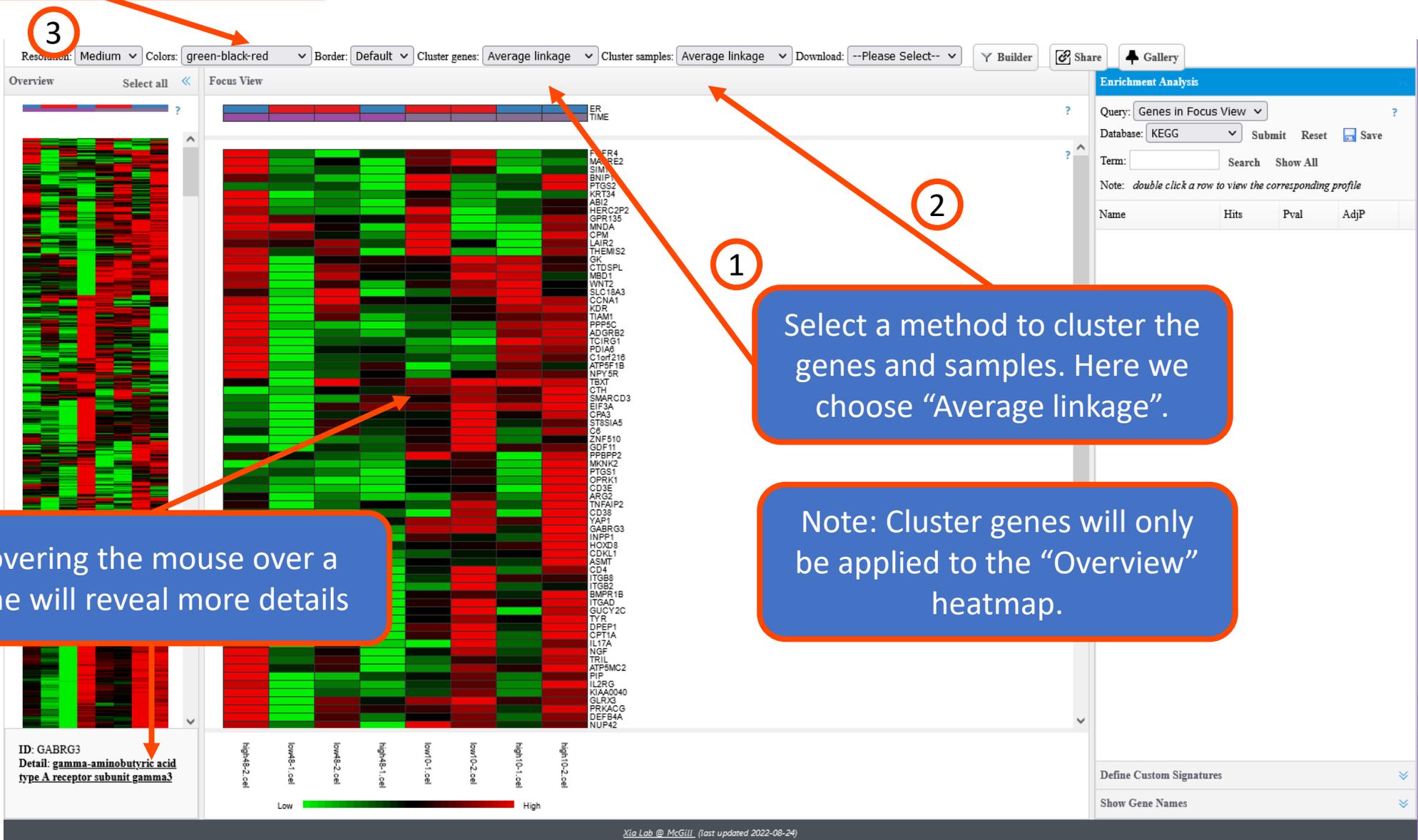
By default, all the significant genes identified from DE analysis are in the “Focus View” heatmap

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



Hovering the mouse over a gene will reveal more details

1

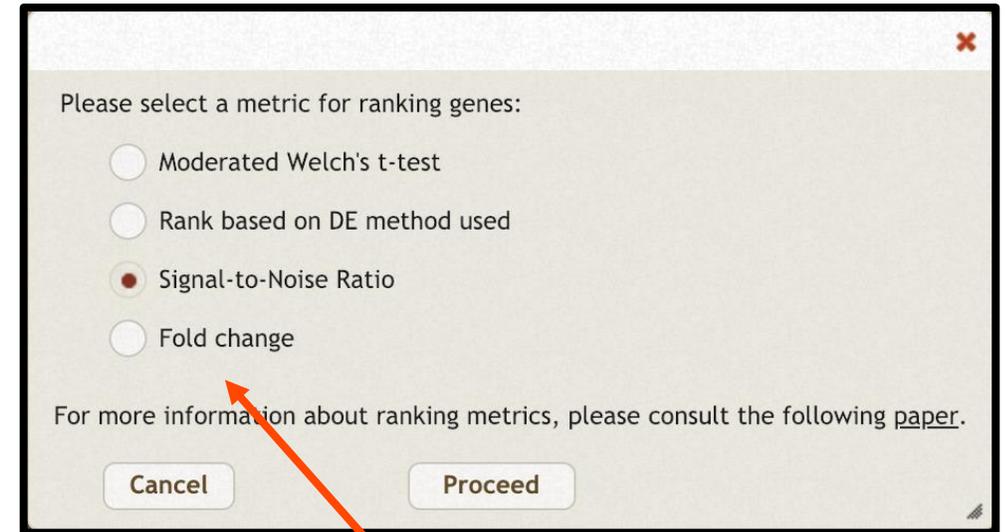
Select a method to cluster the genes and samples. Here we choose "Average linkage".

2

Note: Cluster genes will only be applied to the "Overview" 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

Resolution: Medium Colors: navy-white-firebrick Border: Default Cluster genes: Adjusted P value Cluster samples: --Please Select-- Download: --Please Select-- Share Gallery

Enrichment Plot

Expression pattern

ER TIME

PCNA  
MCM3  
FEN1  
MCM2  
MCM7  
MCM8  
MCM5  
RFC4  
LIG1  
POLA2  
POLD1  
MCM4  
PRIM1  
RNASEH2A  
POLE2  
POLD3  
RFC8  
RFC2  
RFC3  
RPA1  
POLA1  
POLE  
POLD2  
RNASEH1  
POLE3  
PSM2  
SSBP1  
RPA2  
POLD4  
RNASEH2B  
DNA2

Enriched Pathways

Database: KEGG Submit Save

<input type="checkbox"/>	Name	Hits	ES	AdjP
<input checked="" type="checkbox"/>	DNA replication	32/36	-0.915	1.88e-15
<input type="checkbox"/>	Cell cycle	99/124	-0.689	1.1e-11
<input type="checkbox"/>	Spliceosome	85/13	-0.658	4.80e-8
<input type="checkbox"/>	Mismatch repair	19/23	-0.887	2.81e-7
<input type="checkbox"/>	Homologous recombination	24/41	-0.840	1.76e-6
<input type="checkbox"/>	Base excision repair	22/33	-0.836	7.76e-6
<input type="checkbox"/>	Oocyte meiosis	73/12	-0.617	1.9e-5
<input type="checkbox"/>	Fanconi anemia pathway	23/54	-0.814	2.58e-5

high 10.2 cell  
low 1.0 cell  
low 1.2 cell  
high 1.1 cell  
high 1.2 cell

High

Xia Lab @ McGill (last updated 2022-08-24)

# Enrichment Network

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

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

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.

The screenshot shows the 'Enrichment analysis' interface. On the left, a table lists enriched gene sets with columns for Name, Hits, Pval, and AdjP. The 'DNA replication' gene set is highlighted in blue. On the right, a network graph visualizes these gene sets as nodes connected by edges. The 'DNA replication' node is highlighted in red, and an arrow points from a text box to it. Other nodes include 'Cell cycle', 'Pyrimidine metabolism', 'Base excision repair', 'Nucleotide excision repair', 'Mismatch repair', 'Homologous recombination', and 'Fanconi anemia pathway'.

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
Base excision repair	8/21	3.9e-6	1.77e-4
Fanconi anemia pathway	8/21	3.9e-6	1.77e-4
Pyrimidine metabolism	8/29	5.81e-5	0.00231
p53 signaling pathway	9/47	4.19e-4	0.0148
One carbon pool by folate	4/11	0.00151	0.048
Antifolate resistance	5/22	0.00387	0.112
Apoptosis	11/101	0.0111	0.277
Fatty acid elongation	3/10	0.0113	0.277
Biosynthesis of unsaturated fatty acids	3/12	0.0193	0.439
Fatty acid metabolism	5/33	0.0224	0.474

# Enrichment Network

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

Differential Analysis > Sig\_Genes > Analysis Overview > GSEA Heatmap > EnrichNet > Download

View: Bipartite network Node: - Specify - Edge: - Specify - Layout: -- Specify -- Scope: Single node Download: -- Specify --

Type: ORA Database: KEGG Rank(GSEA): Welch's t-test

Extract selected functions

<input type="checkbox"/>	Name	Hits	Pval	AdjP
<input type="checkbox"/>	DNA replication	18/28	1.47e-17	4.67e-15
<input type="checkbox"/>	Cell cycle	22/97	9.36e-10	1.49e-7
<input type="checkbox"/>	Homologous recombination	11/28	3.72e-8	3.94e-6
<input type="checkbox"/>	Mismatch repair	8/18	9.55e-7	7.59e-5
<input type="checkbox"/>	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)

**Pathways in cancer**

CCND1	-0.1496351
CCNE2	-0.377949
CDK2	-0.1244385
CDK4	-0.1035269
E2F1	-0.05787902
E2F3	-0.07163242
MYC	-0.05347874

Xia Lab @ McGill (last updated 2022-08-24)

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

# Enrichment Network

Normalization > Differential Analysis > Sig\_Genes > Analysis Overview > GSEA Heatmap > EnrichNet > Download

Network: Global EnrichNet Background: Purple-gradient View: Gene-set network Node: - Specify - Edge: - Specify - Layout: -- Specify -- Scope: Single node Download: -- Specify --

Enrichment analysis Type: GSEA Database: KEGG Rank(GSEA): Welch's t-test Submit

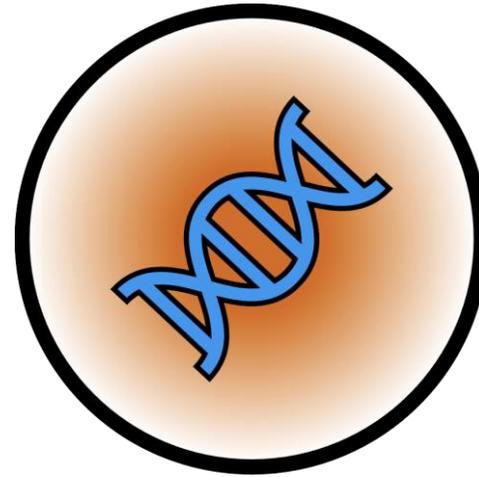
Name	Hits	Pval	AdjP
<input type="checkbox"/> DNA replication	32/36	9e-18	2.39e-15
<input type="checkbox"/> Cell cycle	99/124	1.45e-13	1.93e-11
<input type="checkbox"/> Spliceosome	85/134	6.19e-10	5.49e-8
<input type="checkbox"/> Mismatch repair	19/23	7.66e-9	5.09e-7
<input type="checkbox"/> Homologous recombination	24/41	1.91e-8	1.02e-6
<input type="checkbox"/> Base excision repair	22/33	8.91e-8	3.95e-6
<input type="checkbox"/> Fanconi anemia pathway	23/54	4.72e-7	1.79e-5
<input type="checkbox"/> Oocyte meiosis	73/125	7.5e-7	2.49e-5
<input type="checkbox"/> Nucleotide excision repair	35/47	8.81e-7	2.6e-5
<input type="checkbox"/> Pyrimidine metabolism	28/57	1.8e-6	4.79e-5
<input type="checkbox"/> Terpenoid backbone biosynthesis	16/22	2.81e-6	6.79e-5
<input type="checkbox"/> MAPK signaling pathway	206/295	7.11e-6	1.58e-4
<input type="checkbox"/> Ribosome biogenesis in eukaryotes	30/105	8.37e-6	1.71e-4
<input type="checkbox"/> Huntington's disease	123/193	1.14e-5	2.16e-4
<input type="checkbox"/> RNA transport	91/165	1.35e-5	2.4e-4

Current selection (node double click)

Pathways in cancer	
CCND1	-0.1496351
CCNE2	-0.377949
CDK2	-0.1244385
CDK4	-0.1035269
E2F1	-0.05787902
E2F3	-0.07163242
MYC	-0.05347874

Xia Lab @ McGill (last updated 2022-08-24)

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](http://www.expressanalyst.ca)  
Also visit our forum for FAQs on [www.omicsforum.ca](http://www.omicsforum.ca)*