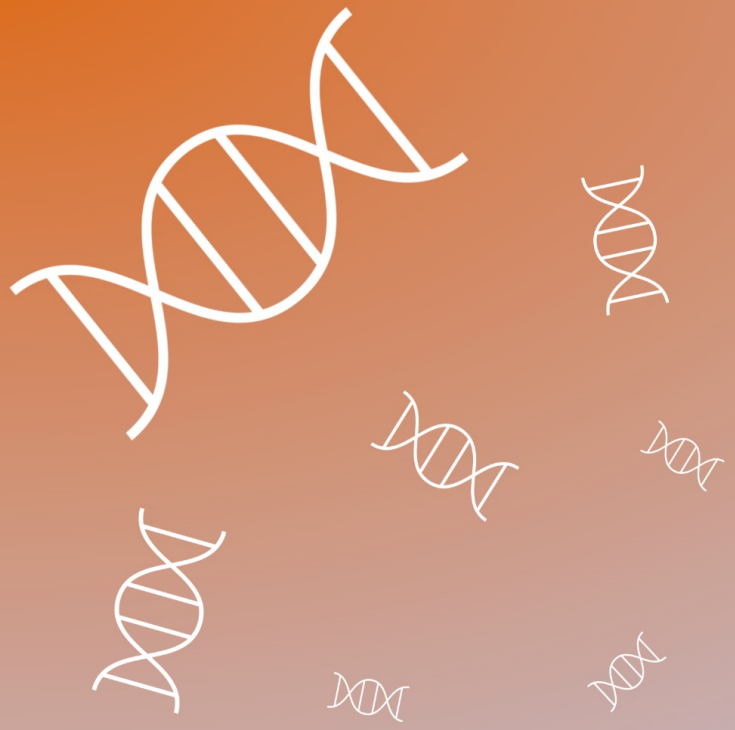


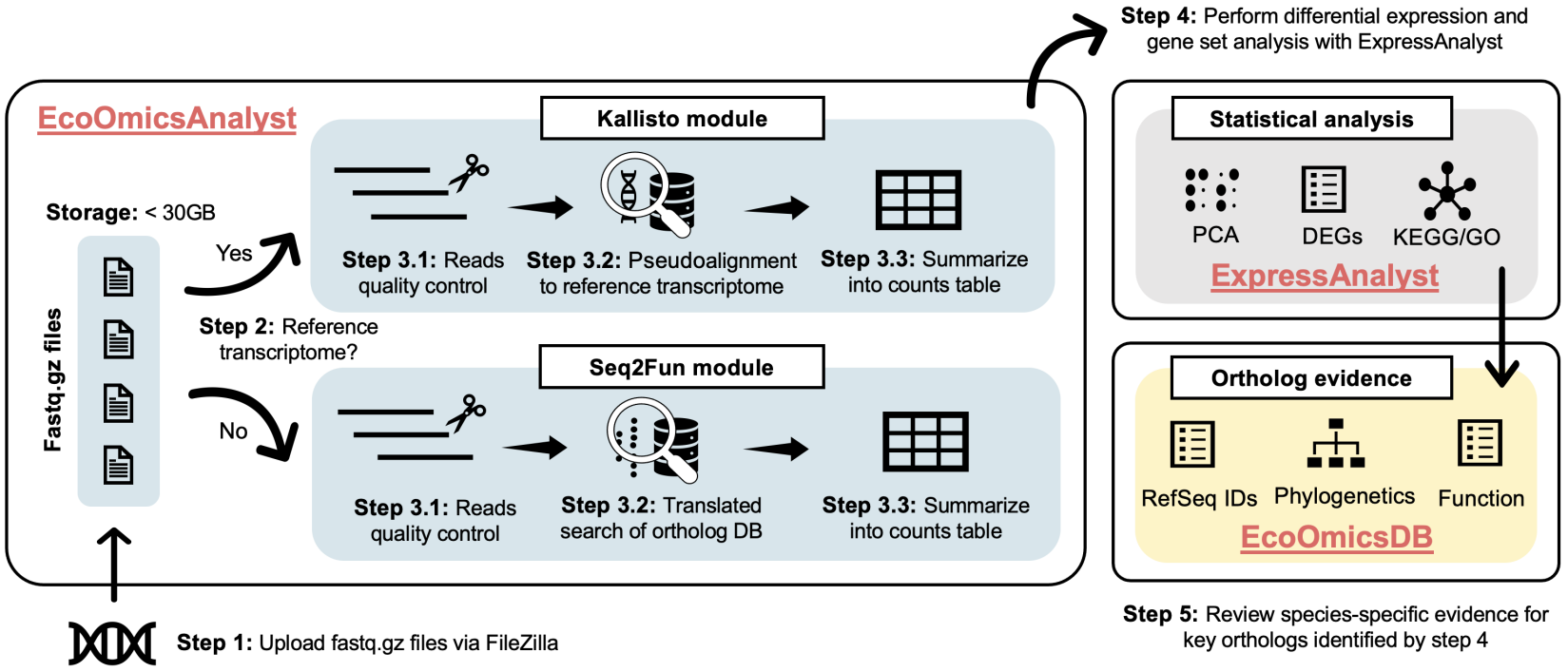
Home Tutorials Forum Updates Contact About Account

ExpressAnalyst for non-model organisms

-- analyzing Seq2Fun counts tables

Introduction

- ExpressAnalyst is part of our comprehensive solution for RNA-seq analysis for species without a reference genome:



Scope of this tutorial

- FASTQ files from three salamander species (2 had no reference genome) were previously processed with Seq2Fun using EcoOmicsAnalyst
  - See the ‘Tutorials’ tab at www.ecoomicsanalyst.ca
- Here, we demonstrate statistical and functional analysis of the resulting Seq2Fun ortholog counts table with ExpressAnalyst
- See tutorials I-IV for other ExpressAnalyst functions

Case study

Comparative transcriptomics of limb regeneration: Identification of conserved expression changes among three species of *Ambystoma*

Varun B. Dwaraka[a,b,*], Jeramiah J. Smith[a], M. Ryan Woodcock[c], S. Randal Voss[b,d,e]

- RNA-seq measured in tissue samples collected after limb amputation
  - Immediately (time0) and after 24 hours (time24)
  - Three species (AND, MAC, MEX)
  - Three replicates per group: (3*time0 + 3*time24)*3 species = 18 samples
  - One outlier removed (MEX, time0)
- Original study quantified MEX with reference transcriptome, AND and MAC with *de novo* transcriptome
  - Each species analyzed independently; results compared afterwards
- Here, data from all species are quantified with Seq2Fun
  - All data analyzed together; species accounted for during statistical analysis

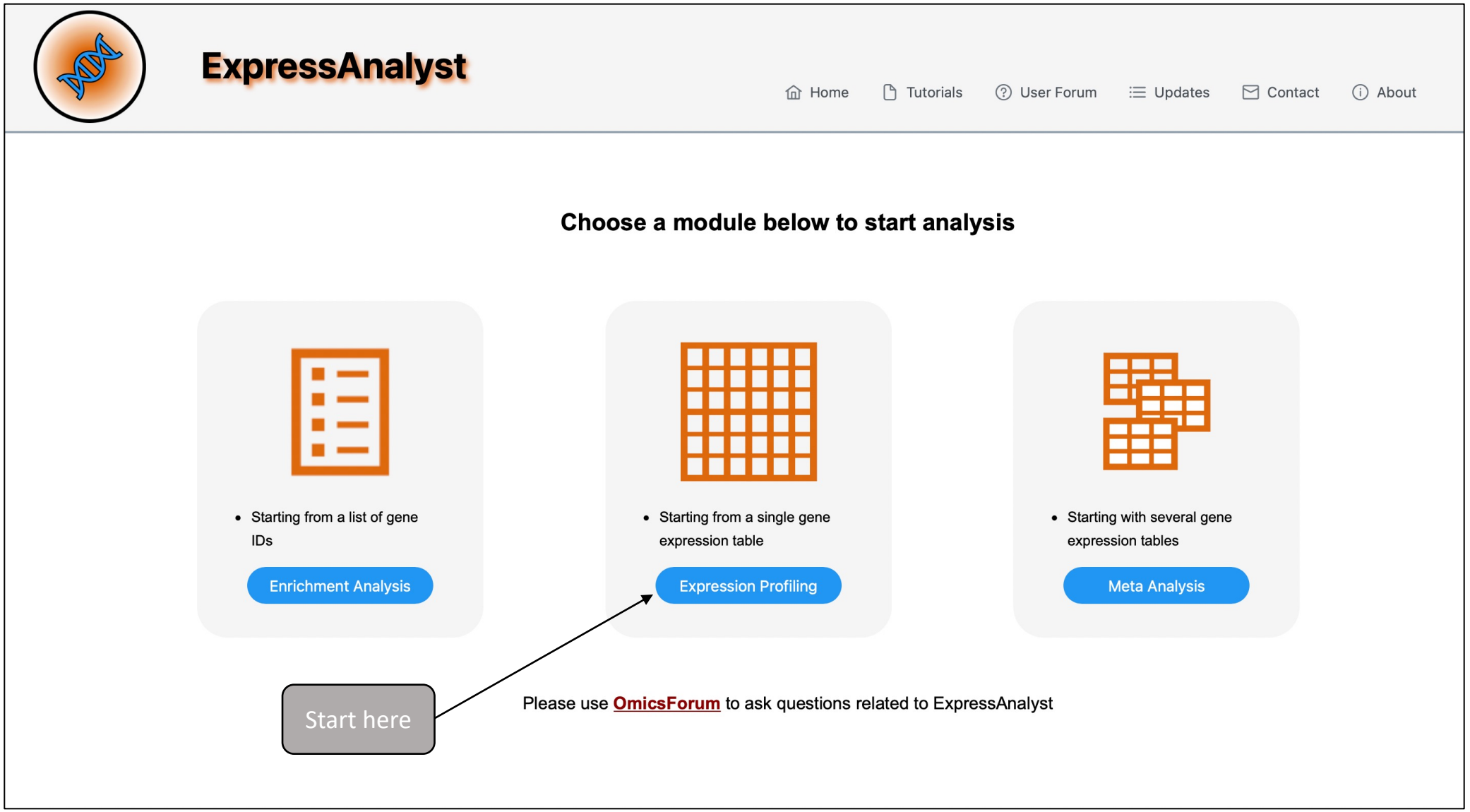
Analysis objectives

- Determine similarities and differences in the transcriptomics of limb regeneration across three salamander species:
  - At whole transcriptome level (PCA plot)
  - In lists of differentially expressed genes
  - At functional level through pathway analysis

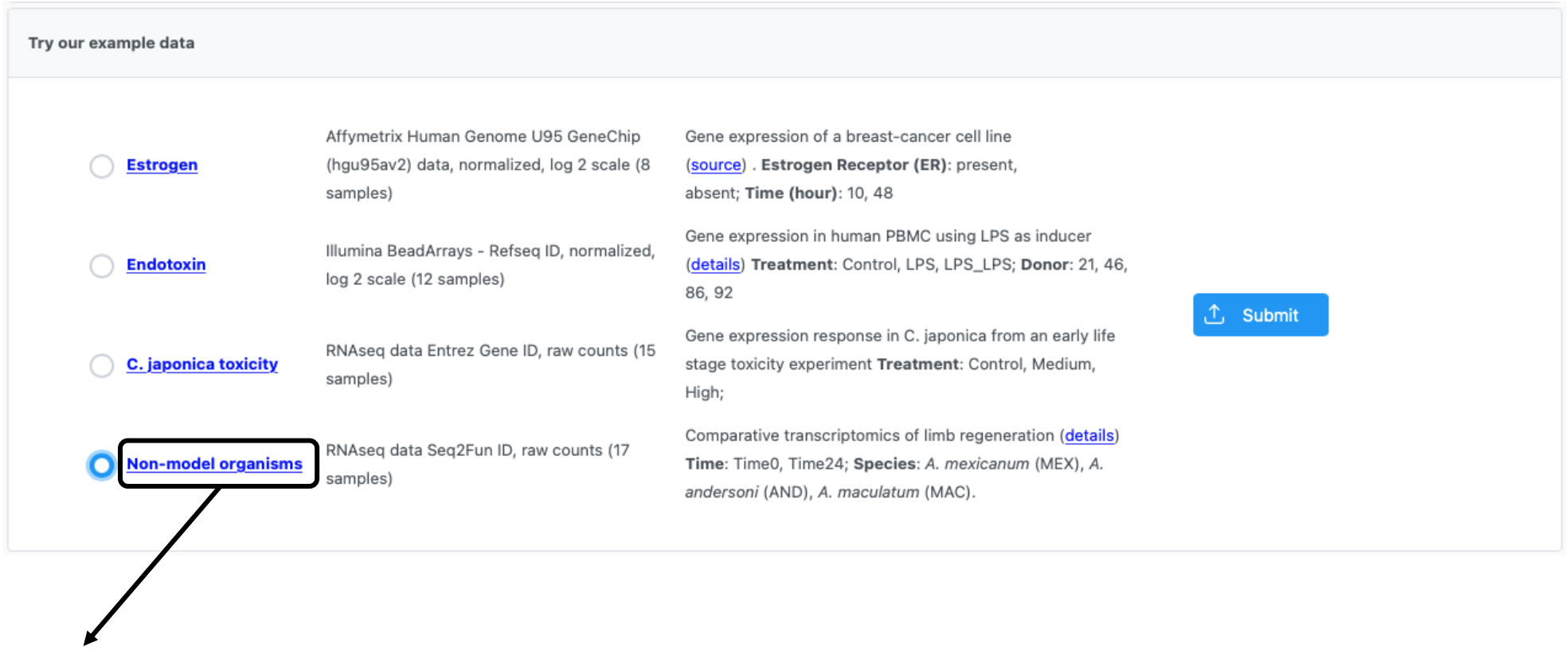
Computer and browser requirements

- 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

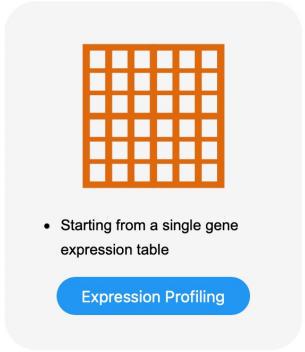
Expression Profiling



Find the data



Click the 4th link to download the dataset



Data Format: Single Matrix

Sample names

Meta-data

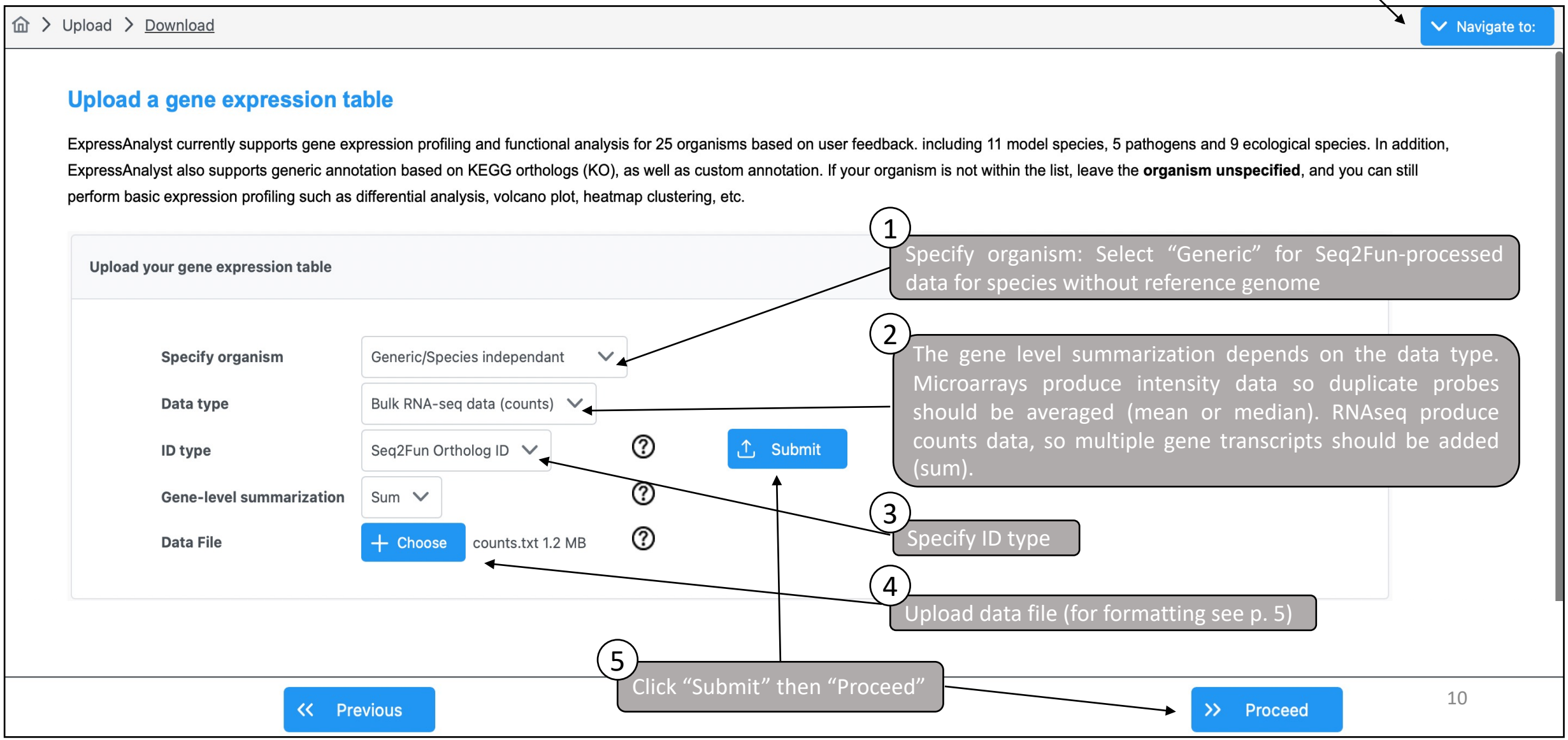
Gene IDs

#NAME	SRR7499348	SRR7499349	SRR7499351	SRR7499352	SRR7499353	SRR7499354	SRR7499355	SRR7499356	SRR7499357	SRR7499358	SRR7499359	SRR7499360	SRR7499361	SRR7499362	SRR7499363	SRR7499364	SRR7499365
#CLASS:Time	time0	time0	time0	time0	time0	time0	time0	time0	time24	time24	time24	time24	time24	time24	time24	time24	time24
#CLASS:Species	MEX	MEX	AND	AND	AND	MAC	MAC	MAC	MEX	MEX	MEX	AND	AND	AND	MAC	MAC	MAC
s2f_0000000100	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
s2f_0000000103	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
s2f_0000000105	13	3	2	1	6	8	6	10	5	7	5	2	6	4	4	0	4
s2f_0000000106	5	8	13	6	5	1	0	3	1	8	3	0	8	5	0	0	1
s2f_0000000107	49	45	41	63	44	22	26	41	49	52	51	40	46	47	3	9	42
s2f_0000000109	3	3	0	0	3	0	0	0	1	0	1	2	0	0	0	0	0
s2f_0000000011	0	0	2	0	1	0	0	0	0	1	0	0	0	1	0	0	1
s2f_0000000110	0	0	0	0	1	0	2	3	0	0	1	1	0	0	0	0	3
s2f_0000000112	2	2	0	0	1	0	0	0	0	1	0	0	0	0	0	0	1
s2f_0000000115	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
s2f_0000000119	2	2	2	2	5	3	1	1	4	3	1	2	3	2	2	1	7
s2f_0000000012	18	9	16	26	10	0	1	0	19	12	15	15	16	10	0	1	0
s2f_0000000120	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0
s2f_0000000123	3	2	0	2	3	1	0	0	1	1	1	1	2	0	0	0	0
s2f_0000000124	2	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
s2f_0000000125	2	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	1
s2f_0000000128	0	1	0	0	0	0	0	1	2	0	0	0	0	0	0	0	0
s2f_0000000126	0	0	0	0	0	1	0	0	0	0	0	1	1	2	0	0	0
s2f_0000000013	2	0	1	1	0	0	0	0	4	0	0	0	1	0	0	0	0
s2f_0000000130	0	0	1	0	1	0	0	0	0	2	3	2	0	0	0	0	1
s2f_0000000131	87	66	89	111	79	37	55	77	100	107	106	97	112	90	16	15	103

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

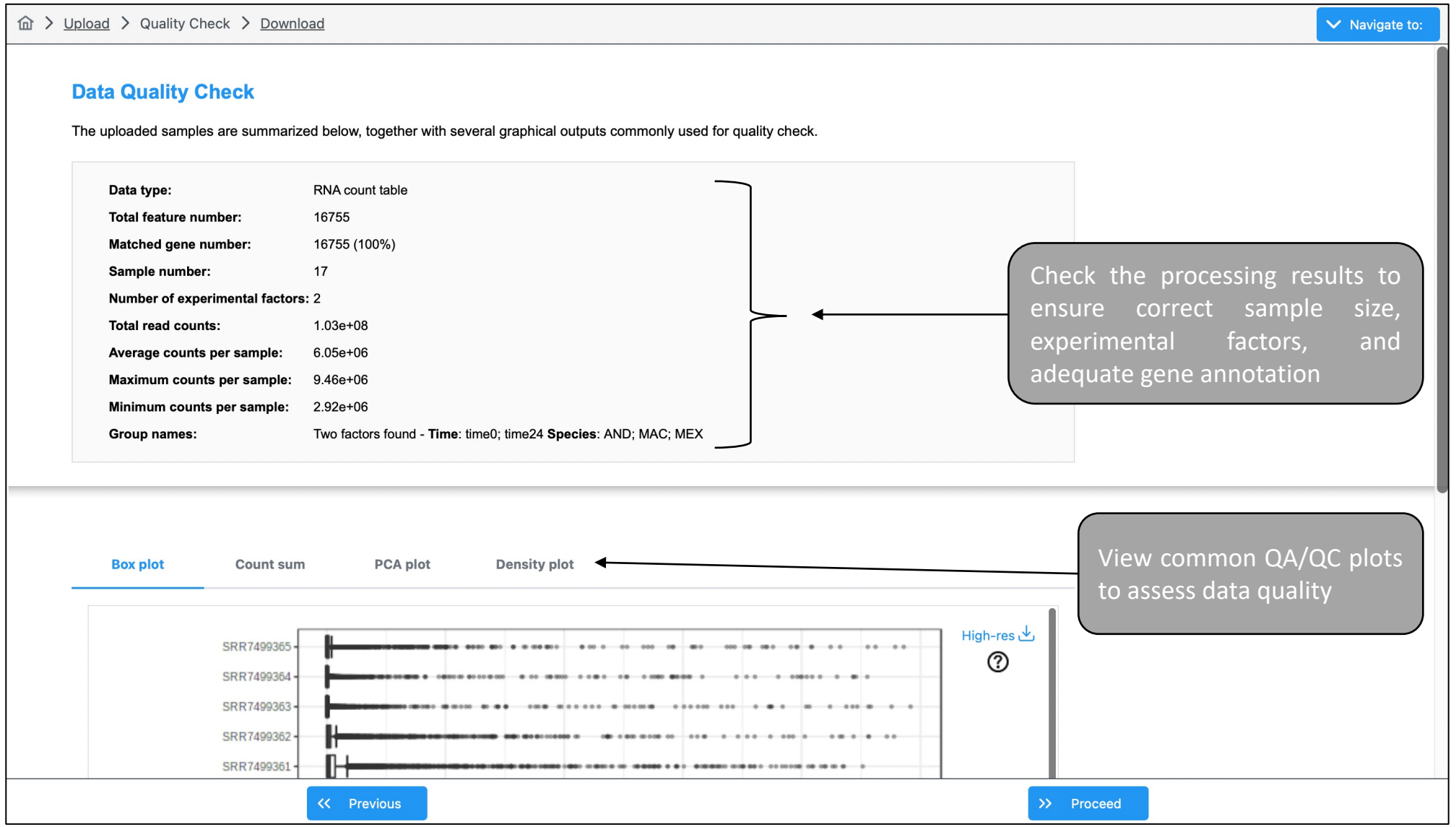
Data Upload and Annotation

Access tutorial, user forum, and updates pages

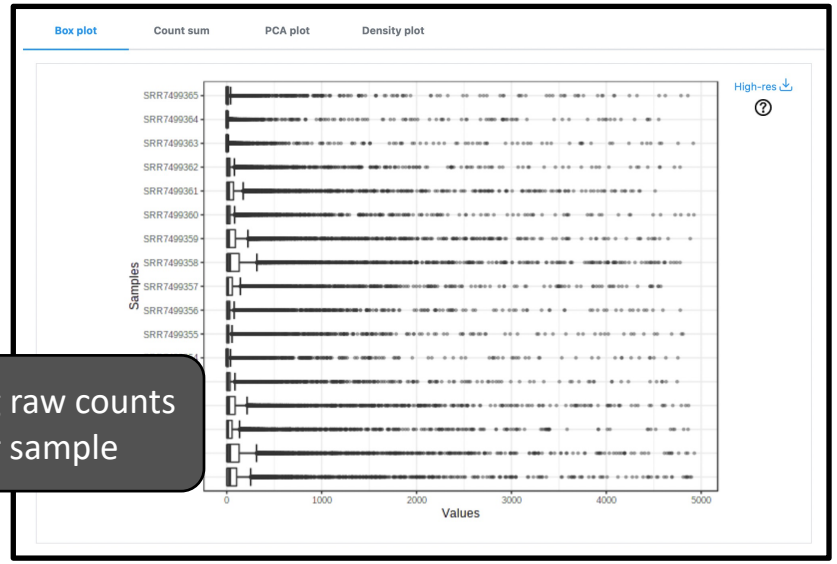


1. Specify organism: Select "Generic" for Seq2Fun-processed data for species without reference genome
2. The gene level summarization depends on the data type. Microarrays produce intensity data so duplicate probes should be averaged (mean or median). RNAseq produce counts data, so multiple gene transcripts should be added (sum).
3. Specify ID type
4. Upload data file (for formatting see p. 5)
5. Click "Submit" then "Proceed"

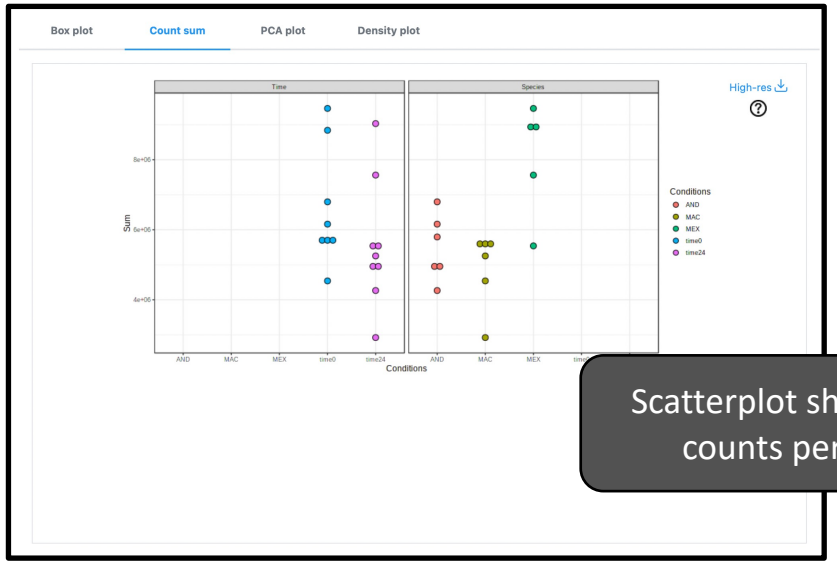
Quality Check: View Processing Results



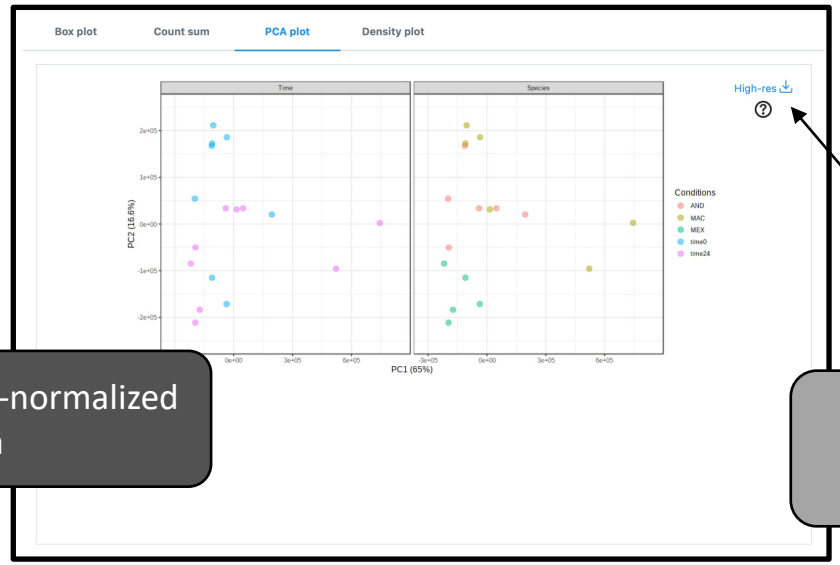
Quality Check: View Quality Control (QC) Plots



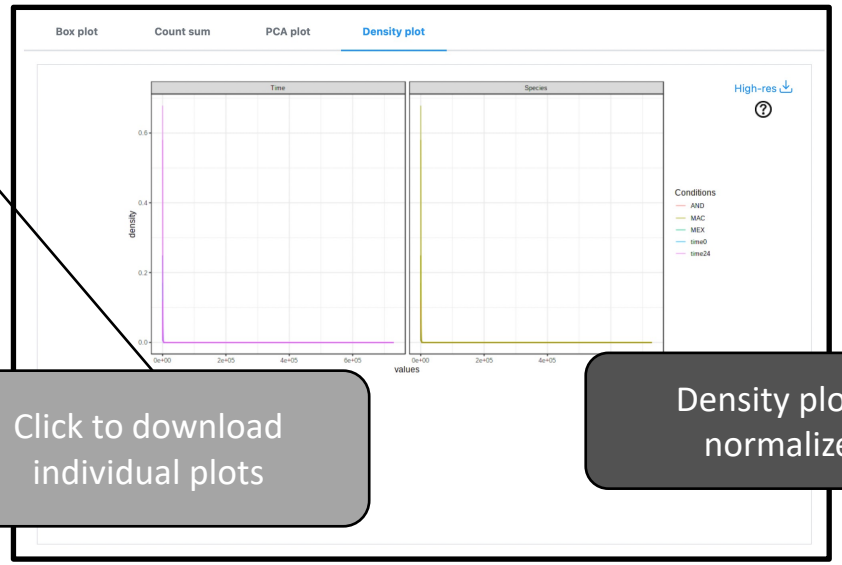
Boxplot showing raw counts per gene per sample



Scatterplot showing total counts per sample



PCA plot of non-normalized data

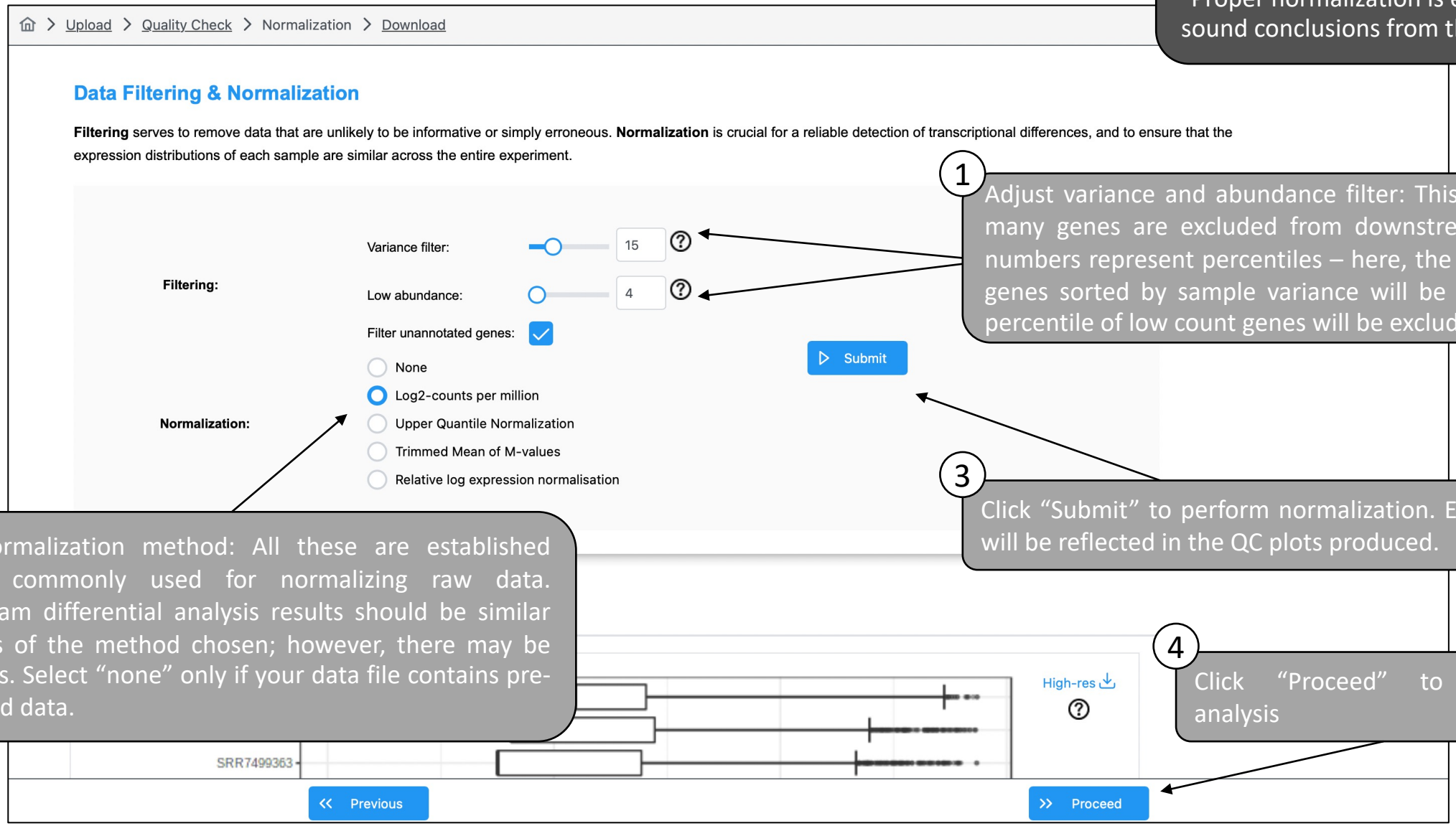


Density plot of non-normalized data

Click to download individual plots

Data Filtering and Normalization

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.



1 Adjust variance and abundance filter: This determines how many genes are excluded from downstream analysis. The numbers represent percentiles – here, the 15th percentile of genes sorted by sample variance will be excluded; the 4th percentile of low count genes will be excluded.

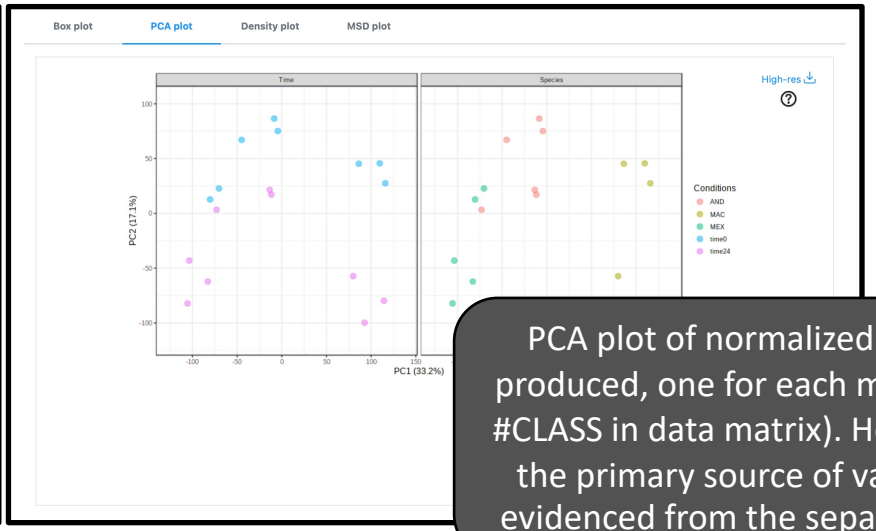
2 Select normalization method: All these are established methods commonly used for normalizing raw data. Downstream differential analysis results should be similar regardless of the method chosen; however, there may be exceptions. Select "none" only if your data file contains pre-normalized data.

3 Click "Submit" to perform normalization. Effects will be reflected in the QC plots produced.

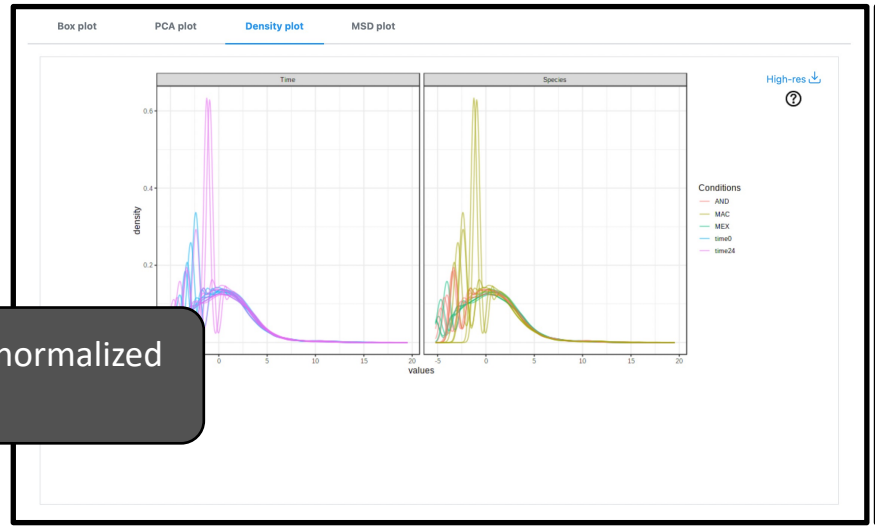
4 Click "Proceed" to continue analysis

Data Filtering and Normalization

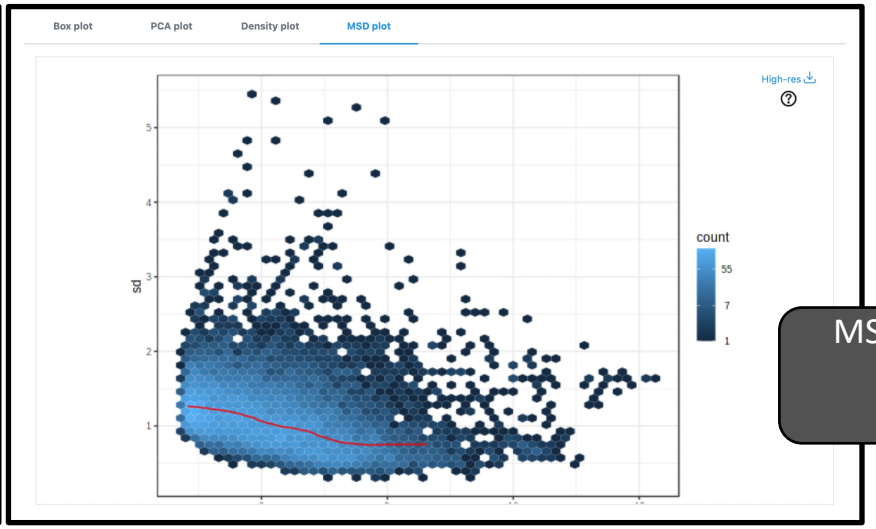
Box plot of normalized counts. Here, Log2-counts per million was used.



PCA plot of normalized data. Here, two plots are produced, one for each meta-data factor (defined by #CLASS in data matrix). Here we see that "Species" is the primary source of variability in the dataset, as evidenced from the separation of groups along PC1.



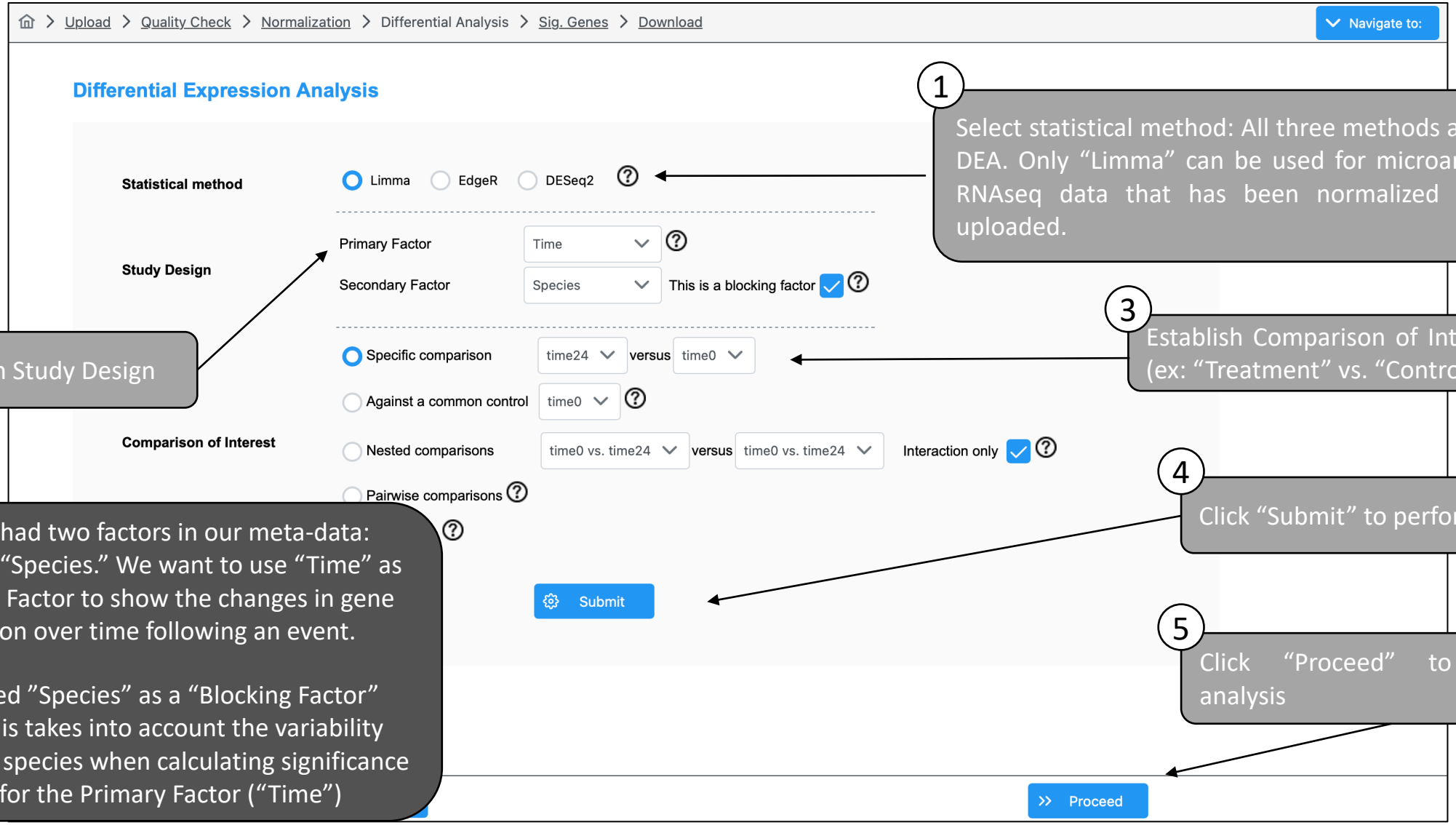
Density plot of normalized data



MSD (Mean vs. Standard Deviation) plot of normalized data

Perform DEA

The purpose of DEA is to group samples by a factor (ex: “time”, “treatment”, “sex”) and then perform statistical tests comparing groups within that factor for significant changes.



1 Select statistical method: All three methods are validated for DEA. Only “Limma” can be used for microarray data or for RNAseq data that has been normalized prior to being uploaded.

2 Establish Study Design

3 Establish Comparison of Interest (ex: “Treatment” vs. “Control“)

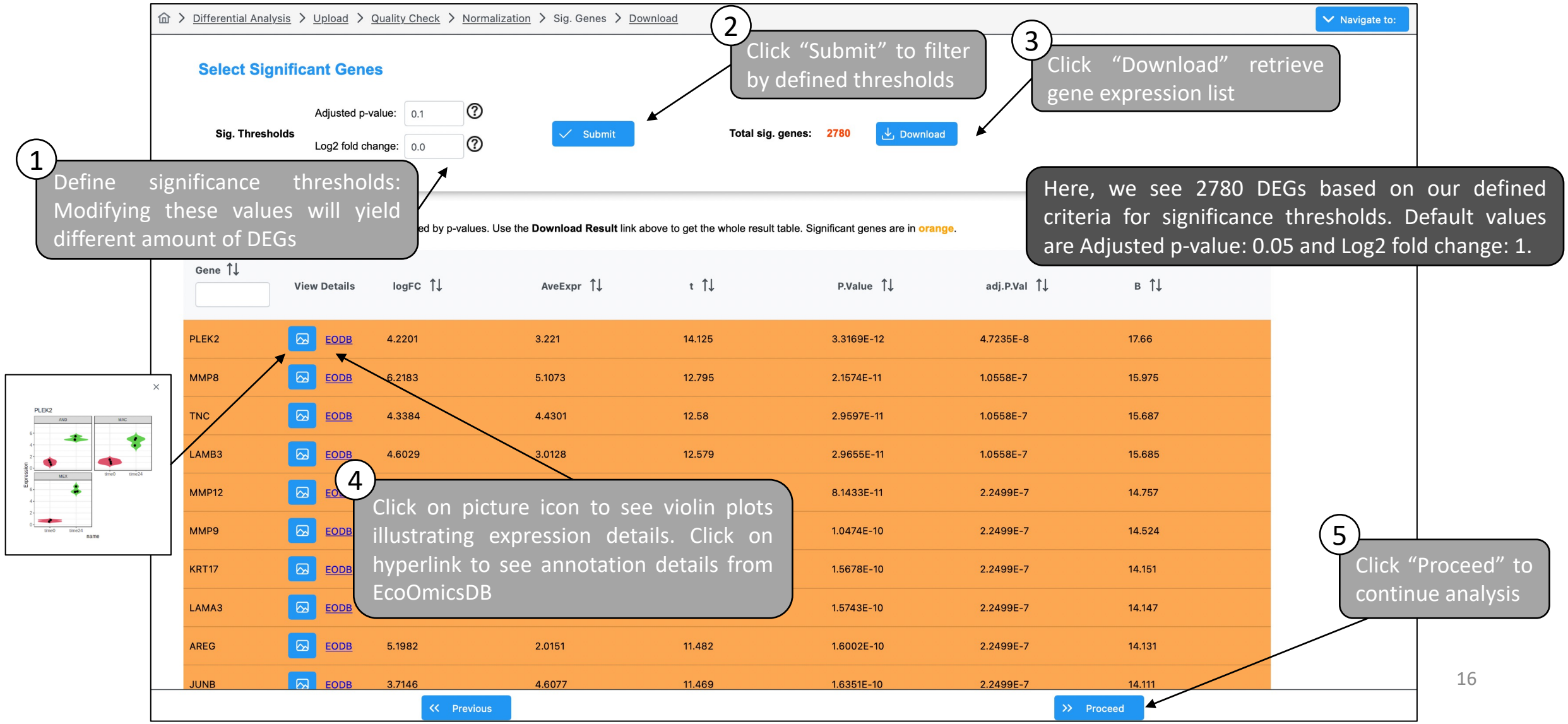
4 Click “Submit” to perform DEA.

5 Click “Proceed” to continue analysis

Here, we had two factors in our meta-data: “Time”, and “Species.” We want to use “Time” as the Primary Factor to show the changes in gene expression over time following an event.

We selected ”Species” as a “Blocking Factor” because this takes into account the variability attributed to species when calculating significance values for the Primary Factor (“Time”)

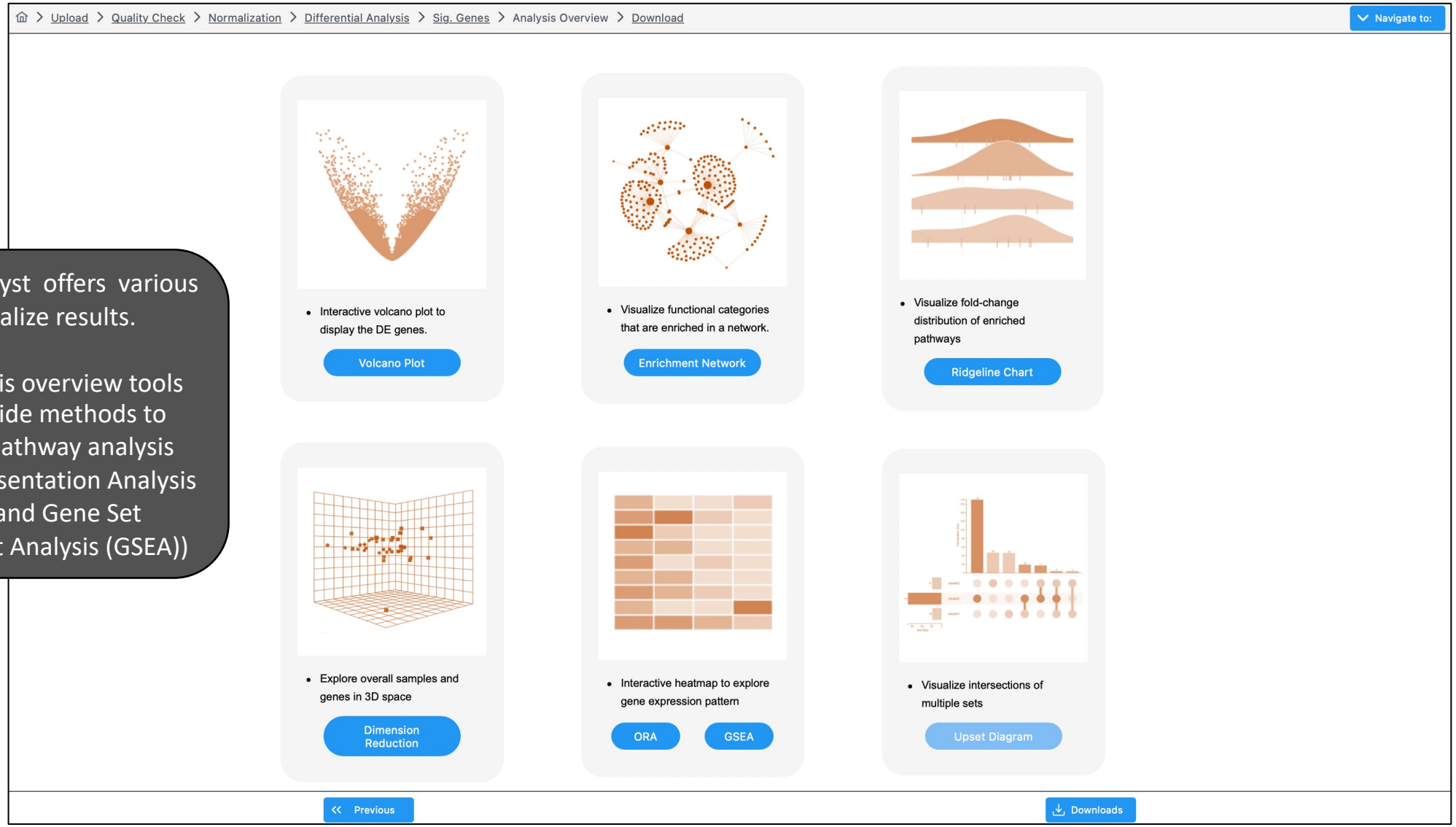
View Differentially Expressed Genes (DEGs)



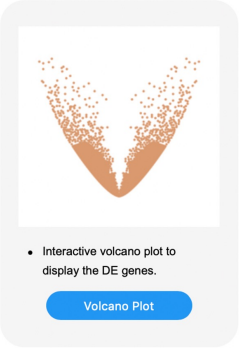
Analysis Overview

ExpressAnalyst offers various tools to visualize results.

The analysis overview tools also provide methods to perform pathway analysis (Over-representation Analysis (ORA) and Gene Set Enrichment Analysis (GSEA))



Analysis Overview: Volcano Plot



Upload > Quality Check > Normalization > Differential Analysis > Sig. Genes > Analysis Overview > Volcano Plot > Download

Navigate to:

Enrichment Analysis (built-in gene sets)

Query: Sig. All

Database: KEGG Submit

Pathway	Hits	Pval	AdjP
Glycolysis / Gluconeogenesis	25	1.53e-8	0.0000063
Type II diabetes mellitus	16	0.0000266	0.00194
Neomycin, kanamycin and gentam	5	0.0000286	0.00194
Cardiac muscle contraction	26	0.000037	0.00194
MicroRNAs in cancer	40	0.0000371	0.00194
Carbon fixation in photosynthetic	11	0.000083	0.0035
Adipocytokine signaling pathway	19	0.0000841	0.0035

Page 1 of 3

Enrichment Analysis (a custom gene set)

Current selection

Gene **PLEK2**
P-val 4.7235e-8
Log-FC 4.2201

Sig.Down [664] Sig.Up [1080] Unsig [12497]

-log10(p)

log2(FC)

2 Perform pathways analysis (ORA) using this pane.

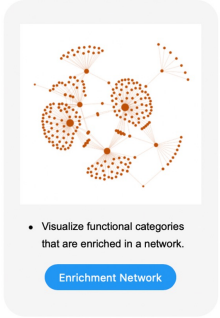
3 Download volcano plot figure (.svg)

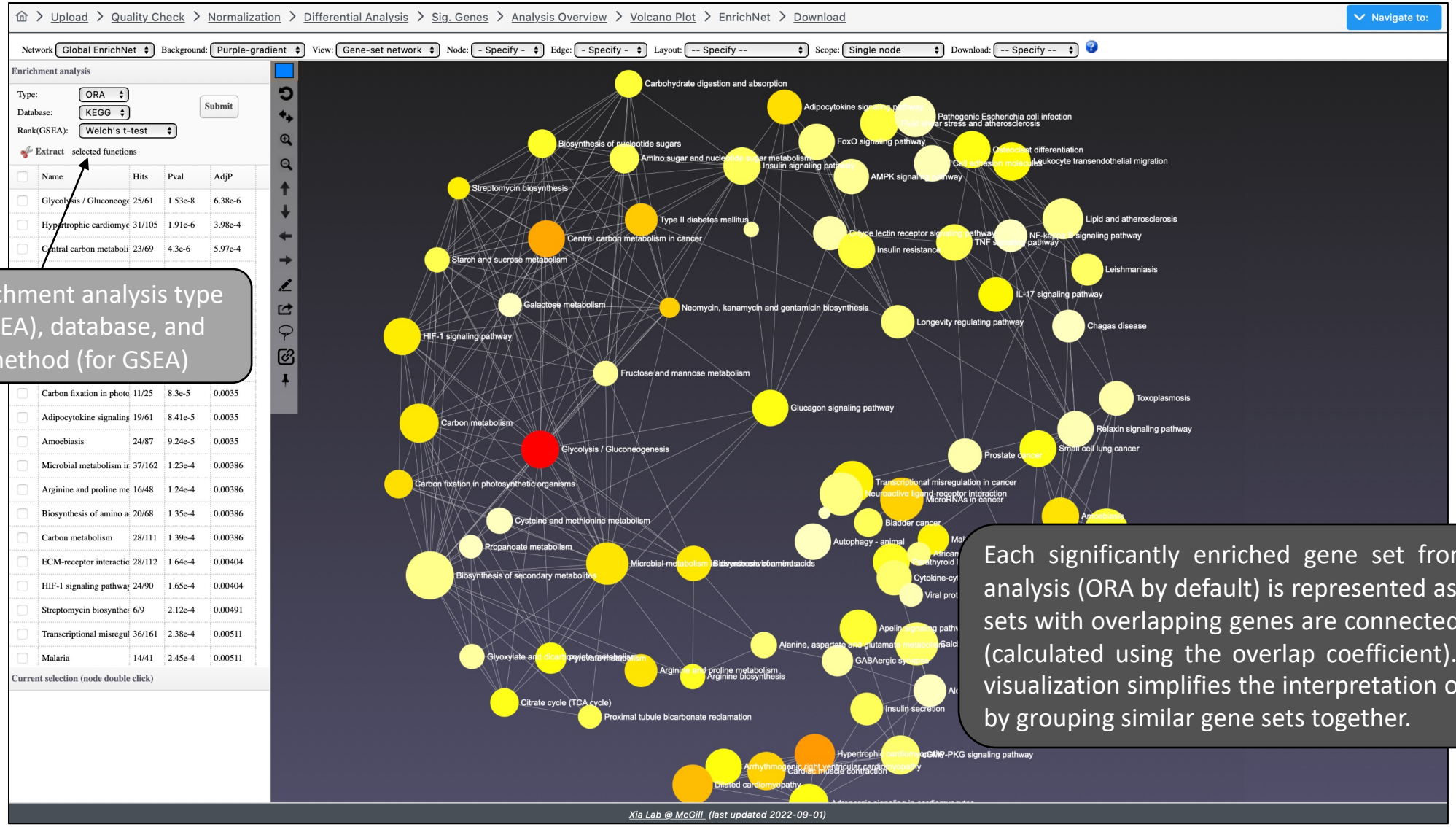
1 Click individual genes to see more details (violin plot). Highlighted genes will appear in the "Current selection" pane

Volcano plots display the -log10(p-value) vs. log2FC from the DEG list.
Significantly upregulated genes (based on user-defined cutoffs) will appear in RED; downregulated genes will appear in BLUE; non-significant genes in the list will remain GREY.

Xia Lab @ McGill (last updated 2022-09-01)

Analysis Overview: Enrichment Network

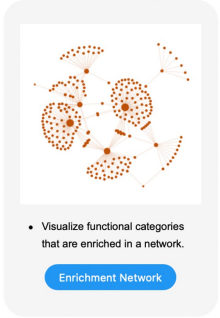




Choose enrichment analysis type (ORA or GSEA), database, and ranking method (for GSEA)

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.

Analysis Overview: Enrichment Network

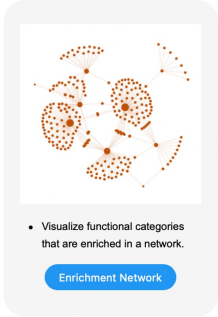


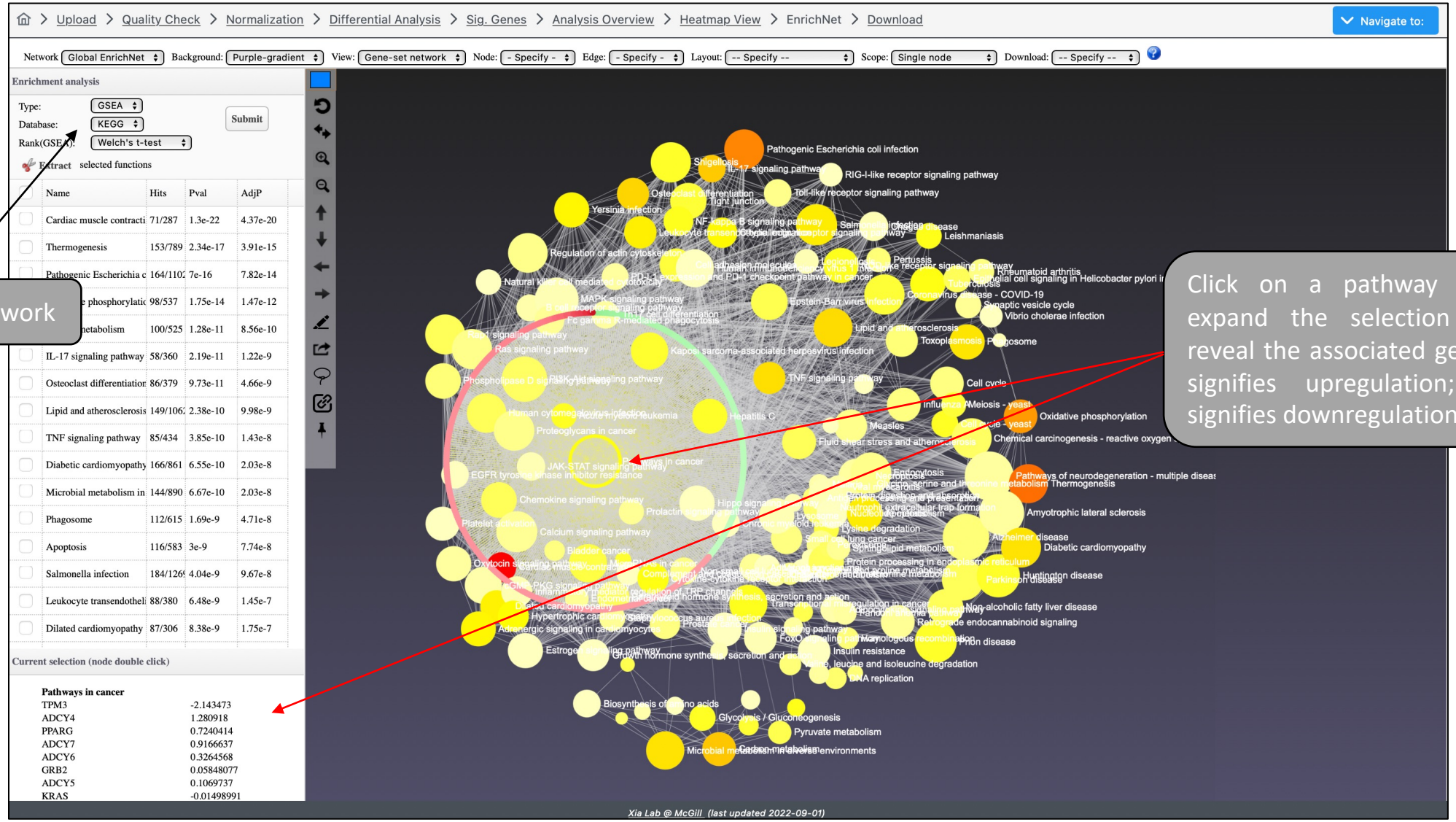
- Visualize functional categories that are enriched in a network.

Enrichment Network

Select ”Bipartite network” in “View” to view all connected genes in pathways

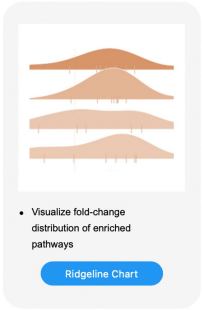
Analysis Overview: Enrichment Network





View GSEA network

Click on a pathway node to expand the selection and to reveal the associated genes. RED signifies upregulation; GREEN signifies downregulation.



Analysis Overview: Ridgeline Plot

Upload > Quality Check > Normalization > Differential Analysis > Sig. Genes > Analysis Overview > Ridgeline > Download

Navigate to:

Settings

Top pathway No. 20
Raw Pval cutoff: 0.05
Sort by: P-value
Ridge color: Teal
Update
Ridge height:
Data point color: Black
Data point shape: Circle
Text size:

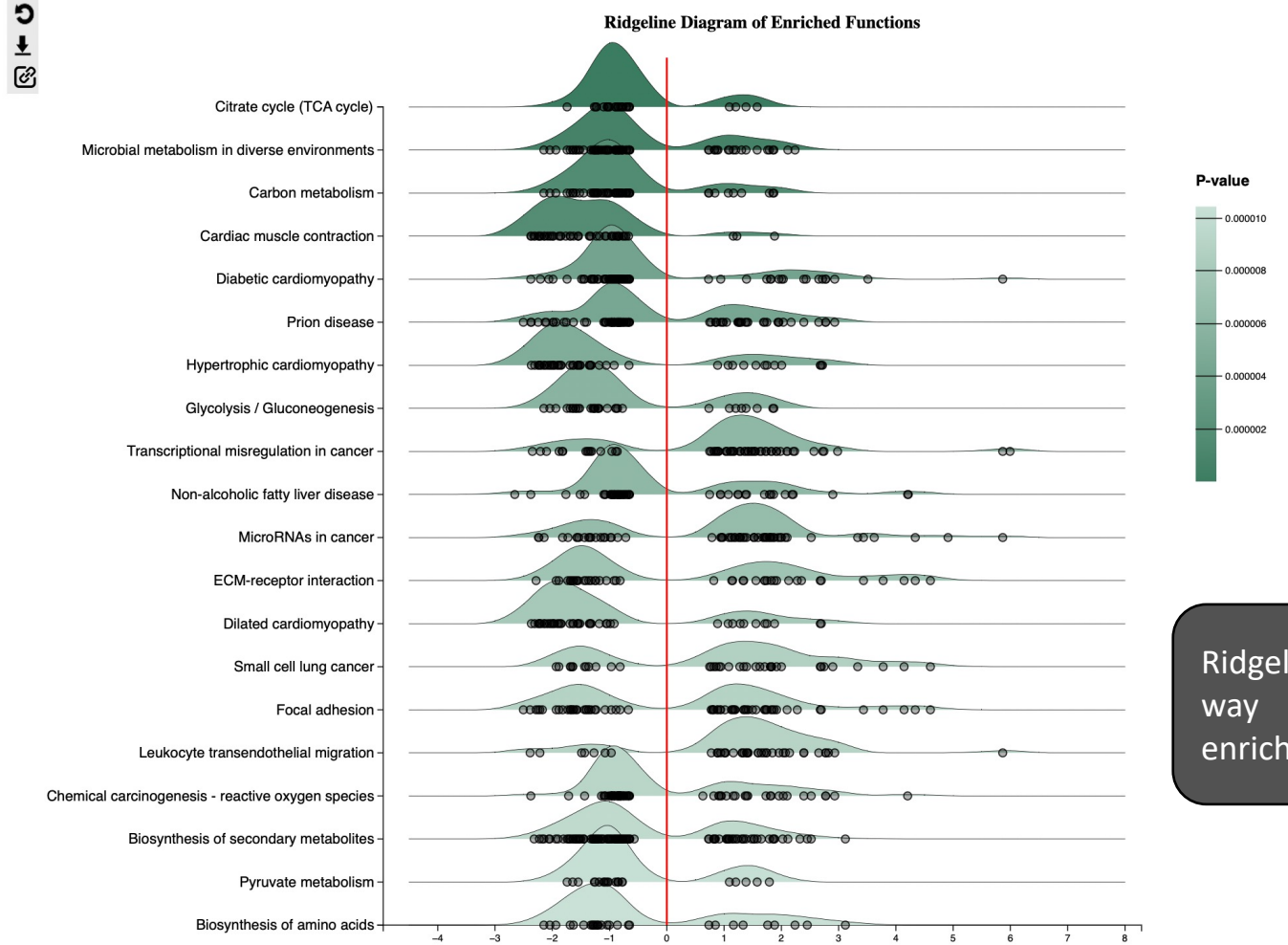
Enrichment Analysis (built-in gene sets)

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

Pathway	Hits	Pval	AdjP
Citrate cycle (TCA cycle)	25	1.335161e	4.753172e
Microbial metabolism in diverse e	70	3.914975e	6.968655e
Carbon metabolism	53	1.602741e	1.613461e
Cardiac muscle contraction	47	1.812877e	1.613461e
Diabetic cardiomyopathy	75	7.153082e	5.092994e
Prion disease	83	1.20767e-5	7.16551e-
Hypertrophic cardiomyopathy	47	3.263409e	1.659676e
Glycolysis / Gluconeogenesis	32	9.196049e	4.092242e
Transcriptional misregulation in ca	62	1.582627e	5.785553e
Non-alcoholic fatty liver disease	56	1.625155e	5.785553e
MicroRNAs in cancer	64	3.366362e	0.0000010
ECM-receptor interaction	46	1.196538e	0.0000035
Dilated cardiomyopathy	46	2.233609e	0.0000061
Small cell lung cancer	39	7.875149e	0.0000200
Focal adhesion	70	0.0000011	0.0000269
Leukocyte transendothelial migrat	42	0.0000013	0.0000292
Chemical carcinogenesis - reactive	66	0.0000018	0.0000385

30 | Page 1 of 12

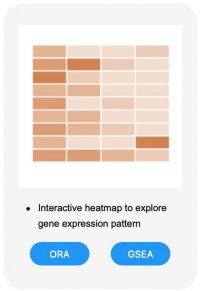
Displaying 1 to 30 of 348 items

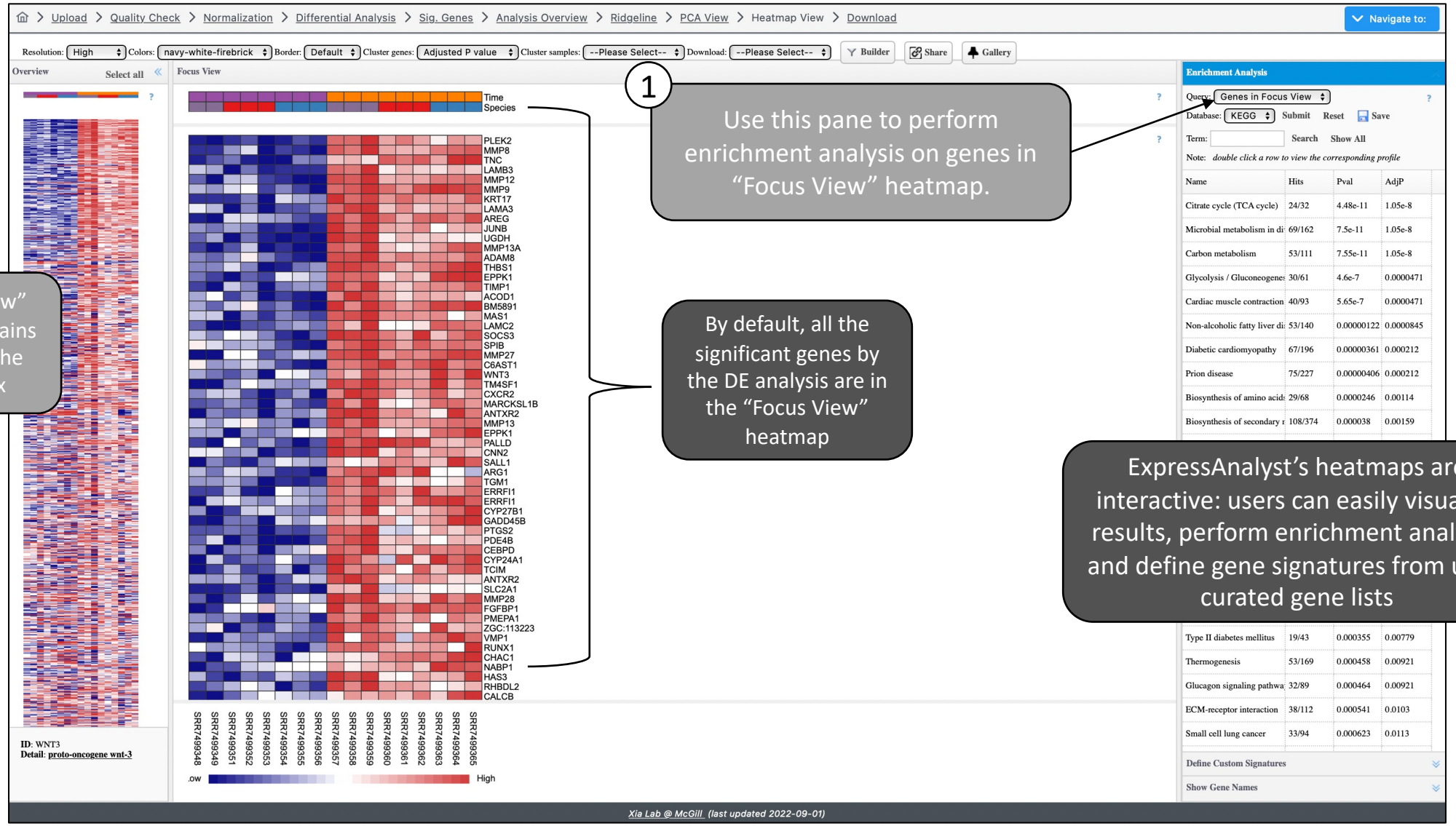


Xia Lab @ McGill (last updated 2022-09-01)

Ridgeline plot is an alternative way to visualize pathways enrichment results

Analysis Overview: ORA Heatmap





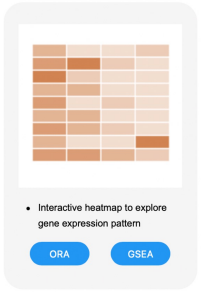
1

Use this pane to perform enrichment analysis on genes in "Focus View" heatmap.

The "Overview" heatmap contains all genes in the data matrix

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

ExpressAnalyst's heatmaps are interactive: users can easily visualize results, perform enrichment analysis, and define gene signatures from user-curated gene lists



Analysis Overview: ORA Heatmap

Change the color scheme

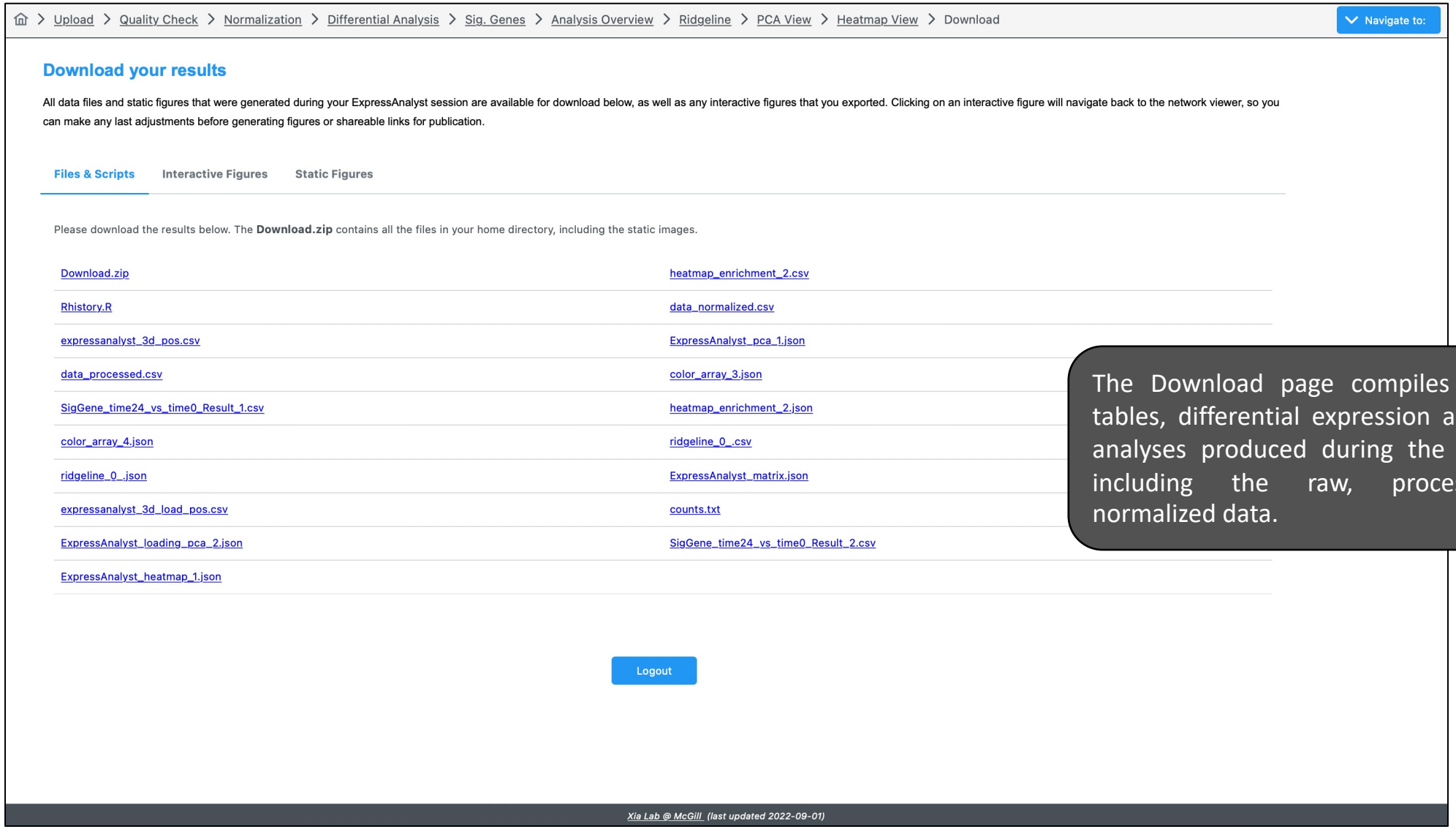
Hover mouse over gene to reveal more details

Define clustering method of genes and samples

ExpressAnalyst's interactive heatmap visualization are easily customizable

Click here to customize gene list

Download Results



The Download page compiles all figures, tables, differential expression and pathway analyses produced during the user's visit, including the raw, processed, and normalized data.

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