

Use the bash script to run Kallisto for quantification of abundances :

```
#!/bin/bash

# Directory containing the files

DIRECTORY="rnaseq-mus-musculus-GSE240196"

# Loop over each file in the directory

for FILE in "$DIRECTORY"/*; do

    # Extract the base name of the file (without extension)

    BASENAME=$(basename "$FILE")

    # Create an output directory for each input file

    OUTPUT_DIR="kallisto-output/$BASENAME"

    mkdir -p "$OUTPUT_DIR"

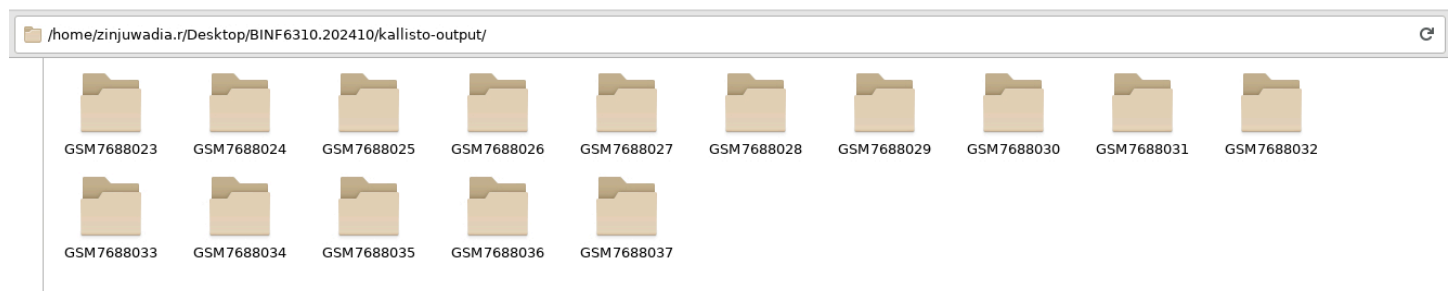
    kallisto quant -i Mus_musculus.idx -l 200 -s 20 -o $OUTPUT_DIR --single $FILE

    mv "$OUTPUT_DIR/abundance.tsv" "$OUTPUT_DIR/$BASENAME.tsv"

    mv "$OUTPUT_DIR/abundance.h5" "$OUTPUT_DIR/$BASENAME.h5"

done
```

This will create different output folder :



Combine Kallisto Outputs :

```
module load R
```

Inside R, install the required packages:

```
install.packages("readr")  
install.packages("dplyr")  
q()
```

Run the R script to join the results:

```
library(readr)  
  
library(dplyr)  
  
# Directory containing the nested kallisto output directories  
  
DIRECTORY <- "kallisto-output"  
  
# List of directories under the main directory  
  
dirs <- list.dirs(path = DIRECTORY, full.names = TRUE, recursive = FALSE)  
  
# Create a list of kallisto .tsv output files based on the nested structure  
  
files <- sapply(dirs, function(d) {  
  file.path(d, paste0(basename(d), ".tsv"))  
})  
  
# Function to read in kallisto abundance.tsv and extract counts  
  
read_kallisto <- function(file) {  
  data <- read_tsv(file)  
  counts <- data$est_counts  
  names(counts) <- data$target_id  
  return(counts)  
}  
  
# Read in data from all files
```

```

data_list <- lapply(files, read_kallisto)

# Sample names

sample_names <- sapply(dirs, function(d) {

  basename(d)

})

# Combine all data into a matrix

count_matrix <- do.call(cbind, data_list)

colnames(count_matrix) <- sample_names

# Write to a CSV file

write.csv(count_matrix, file = paste0(DIRECTORY, "/count_matrix.csv"), row.names = TRUE)

```

Run the R script :

Rscript kallisto-output-join.R

Download the Output File :

count_matrix.csv

Output File look Like this :

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1		GSM7688023	GSM7688024	GSM7688025	GSM7688026	GSM7688027	GSM7688028	GSM7688029	GSM7688030	GSM7688031	GSM7688032	GSM7688033	GSM7688034	GSM7688035	GSM7688036	GSM7688037
2	ENSMUST00000196221.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	ENSMUST00000179664.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	ENSMUST00000177564.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	ENSMUST00000178537.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	ENSMUST00000178862.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	ENSMUST00000179520.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	ENSMUST00000179883.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	ENSMUST00000195858.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	ENSMUST00000179932.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	ENSMUST00000180001.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	ENSMUST00000178815.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	ENSMUST00000177965.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	ENSMUST00000178909.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	ENSMUST00000177646.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Edit the File :

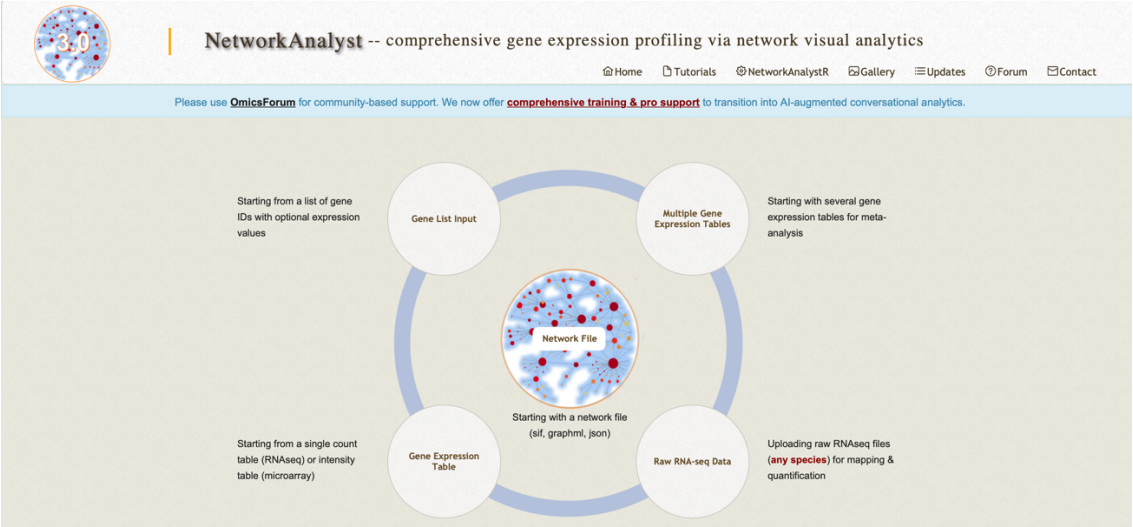
Open the file in a text editor or a program like MS Excel. Manually edit the headers to add annotation for Network Analyst, Save the file as a tab-delimited file.

The sample name must be in the first line, followed by the metadata labels. Each Metadata starts with new line begging with “#CLASS”.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	#NAME	GSM768802	GSM768802	GSM768802	GSM768802	GSM768802	GSM768802	GSM768802	GSM768803	GSM768803	GSM768803	GSM768803	GSM768803	GSM768803	GSM768803	GSM768803
2	#CLASS	Control	Control	Control	Control	Control	Control	Negative	Negative	Negative	Negative	Negative	Negative	Positive	Positive	Positive
3	ENSMUST00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	ENSMUST00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	ENSMUST00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	ENSMUST00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	ENSMUST00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	ENSMUST00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	ENSMUST00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	ENSMUST00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	ENSMUST00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	ENSMUST00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	ENSMUST00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	ENSMUST00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	ENSMUST00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	ENSMUST00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	ENSMUST00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	ENSMUST00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
19	ENSMUST00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	ENSMUST00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21	ENSMUST00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
22	ENSMUST00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Analysis in NetworkAnalyst.ca :

Go to [NetworkAnalyst.ca](#) , Choose the "Gene Expression Table" module.



Choose the "Gene Expression Table" module and fill the options.

Upload a gene expression table

ExpressAnalyst currently supports gene expression profiling and functional analysis for 28 organisms including 11 model species, 5 pathogens and 12 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.

Specify organism

M. musculus (mouse)

Analysis Type

Differential Expression

Data type

Counts (bulk RNA-seq)

ID type

Genbank ID

Data File

+ Choose

count_matrix_latest.txt 6.7 MB

☒ Metadata included

Metadata File

+ Choose

Submit

Different types of visualizations :

Data Quality Check

Omics data overview Metadata overview

The uploaded samples are summarized below, together with several graphical outputs commonly used for quality check.

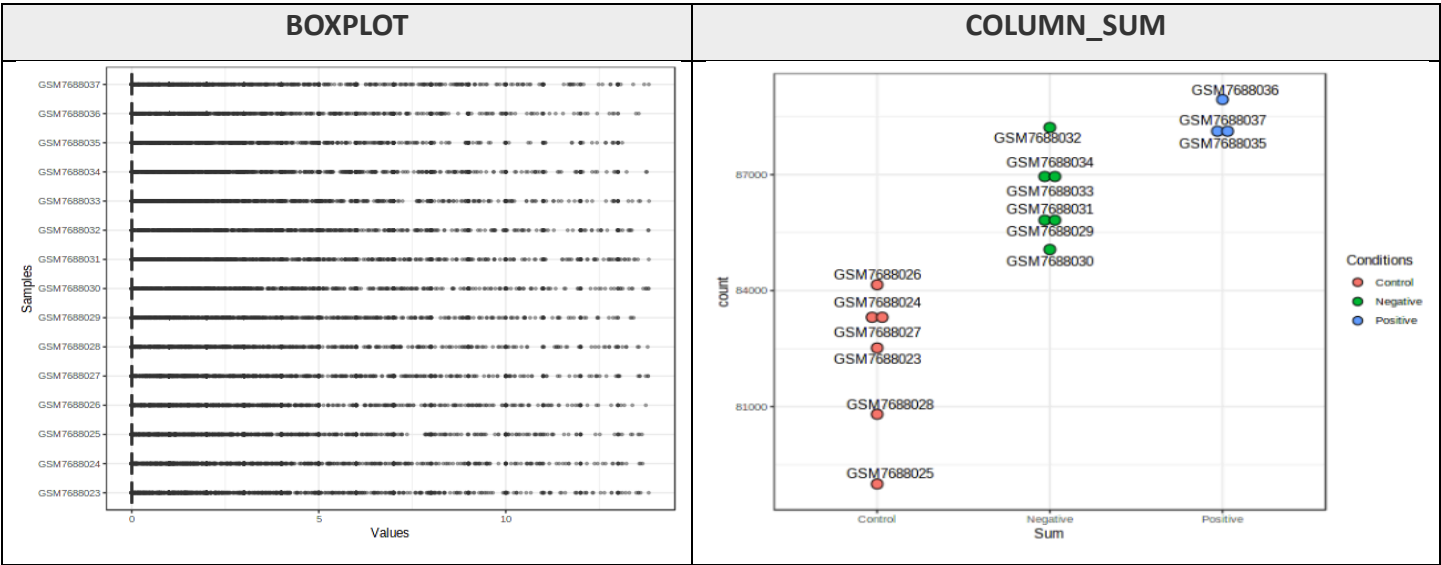
Data type:	RNA count table
Total feature number:	58336
Matched feature number:	20303 (34%)
Sample number:	133
Number of experimental factors:	5
Total read counts:	2.46e+09
Average counts per sample:	1.85e+07
Maximum counts per sample:	3.38e+07
Minimum counts per sample:	3.07e+06

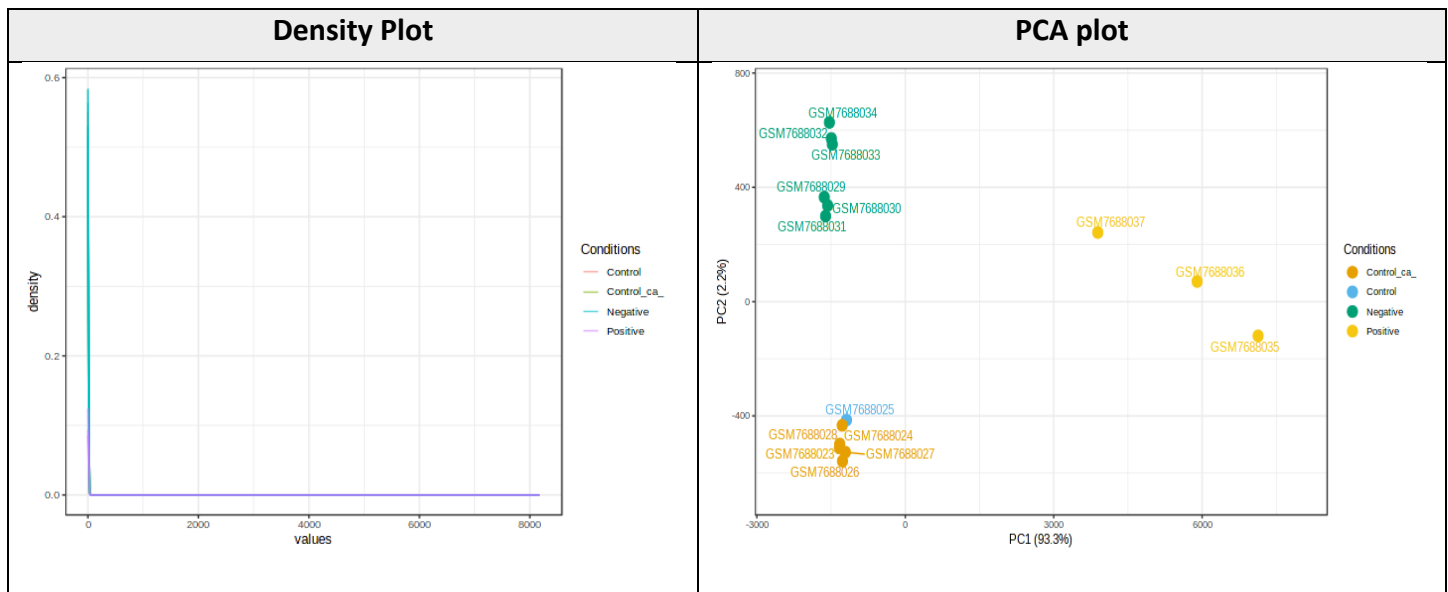
View common QA/QC plots to check the quality of the data.

Density Plot : A density plot is a smooth curve that depicts the data distribution. Rather than the frequency, the curve reflects the proportion of data in each range. This implies that the height of the curve shows the proportion of the data that falls into that range rather than how many times a number appears. Density charts are useful for visualizing data dispersion. It displays proportions and is important for analyzing patterns of data.

PCA plot : Principal component analysis (PCA) is becoming more popular as a method for extracting significant patterns from complicated biological information. No samples or features (variables) are discarded using PCA. Instead, by building principal components (PCs), it minimizes the overwhelming number of dimensions. PCs describe variation and account for the original traits' various impacts. These effects, or loadings, may be traced back from the PCA plot to determine what causes the variations across clusters.

Box Plot : A box plot is made up of two components: a box and a set of whiskers. The lowest point represents the data set's minimum value, while the highest point represents the data set's maximum value (from left to right). The box is drawn from the first to third quartiles (Q1), with a horizontal line in the middle denoting the median. The plot can be oriented horizontally or vertically.





Normalize and filter the data :

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 a

Filtering:

Filter unannotated features: ☒

Low abundance: ?

Variance filter: ?

Normalization:

☒ None

☐ Log2-counts per million (logCPM) transformation

☐ Upper Quantile (UQ) normalization

☐ Trimmed Mean of M-values (TMM) normalization

☐ Relative Log Expression (RLE) normalization

Note: the filtered and normalized data will be used for all visualizations and as input for the *limma* differential expression method. Unnormalized counts will be used as input for the *edgeR* and *DESeq2* methods (available for RNA-seq only).

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

Normalization according to logCPM Transformation :

Filtering:

Filter unannotated features: ☒
Low abundance: ?
Variance filter: ?

Normalization:

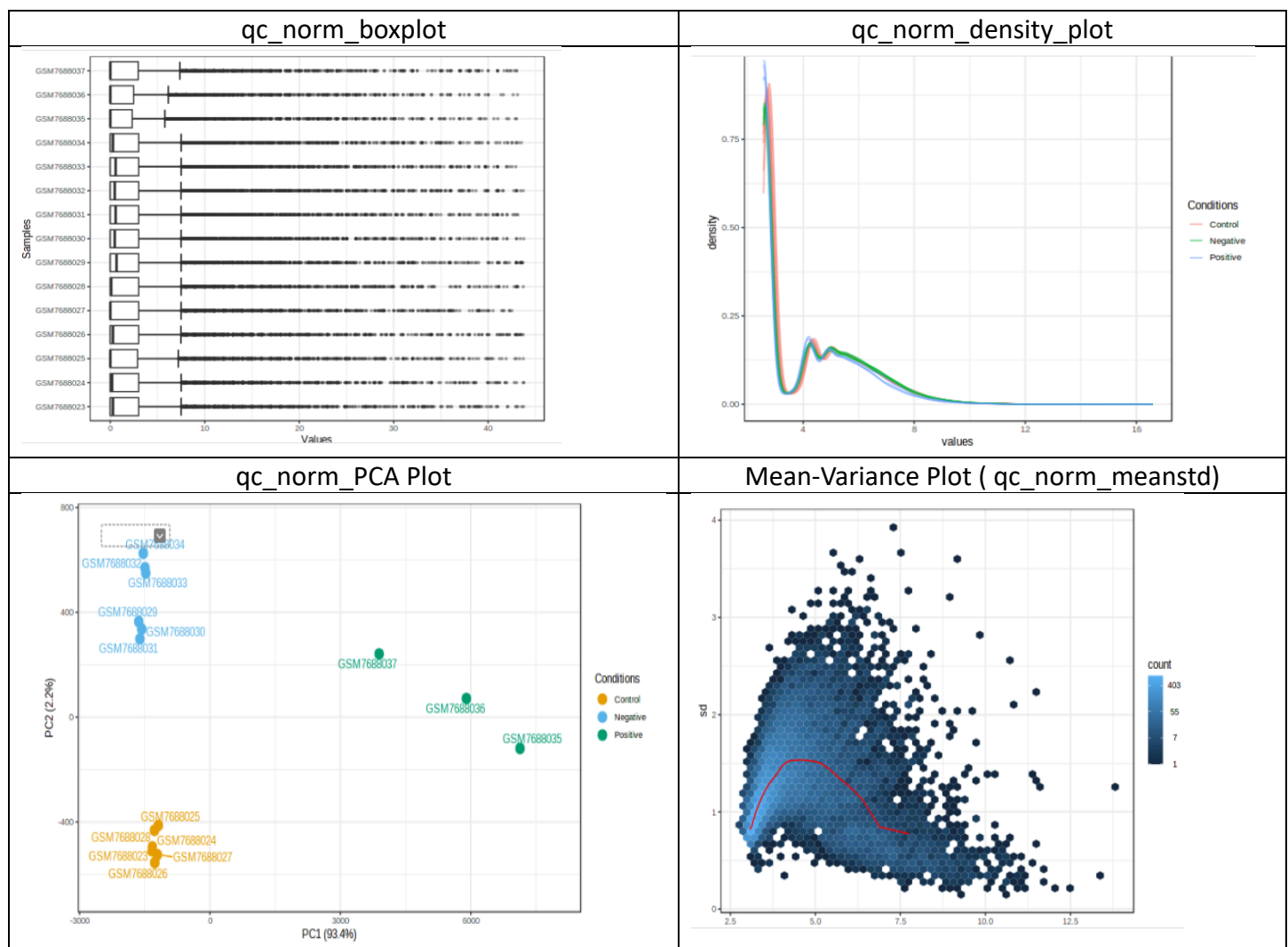
☐ None
☒ Log2-counts per million (logCPM) transformation
☐ Upper Quantile (UQ) normalization
☐ Trimmed Mean of M-values (TMM) normalization
☐ Relative Log Expression (RLE) normalization

Submit

Note: the filtered and normalized data will be used for all visualizations and as input for the *limma* differential expression method. Unnormalized counts will be used as input for the *edgeR* and *I*

→ Diagnostic plot summarizing the standard deviation versus mean measures of read in the sample for each feature. It checks whether there is a dependence between counts and Variance.

→ Plot of density against log2 of read counts. It displays the relative distribution of different counts in each group.



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 compared to those that do not.

Differential Expression Analysis

Simple Metadata

Complex Metadata

Statistical method

☒ Limma

☐ EdgeR

☐ DESeq2

☐ Adjust using robust trend

Study Design

Primary Factor

CLASS

Secondary Factor

--- Not Available ---

This is a blocking factor

Comparison of Interest

☐ Specific comparison

Control

versus

Negative

☐ Against a common control

Control

☒ Nested comparisons

Control vs. Negative

versus

Control vs. Positive

Interaction only

☐ Pairwise comparisons

☐ Time series

Submit

View differentially expressed genes (DEGs)

Log2 fold change: 0.0

Sig. Thresholds

P-value: 0.05

☒ Use adjusted p-value

Submit

Total sig. features: 413

Download

The table below shows at most top 1000 features ranked by p-values. Use the Download Result link above to get the whole result table. Significant features are in orange. For dose/time series analysis, a feature will be highlighted if it passes the fold-change and p-value thresholds for any group.

Name	Detail	logFC	AveExpr	t	P.Value	adj.P.Val	B	Figure
ENSMUST000000000	ENSMUST00000000104	-6.6607	6.5949	-18.1	3.935E-13	8.4004E-9	18.12	
ENSMUST000000005	ENSMUST00000005727	6.3946	5.1752	17.324	8.4115E-13	8.9784E-9	17.591	
ENSMUST000000003	ENSMUST00000003219	-7.4678	7.4871	-16.497	1.957E-12	1.3926E-8	16.987	
ENSMUST000000007	ENSMUST00000007941	-5.9783	6.214	-15.158	8.3298E-12	4.4456E-8	15.91	
ENSMUST000000000	ENSMUST00000000105	-6.0857	6.0924	-14.368	2.0631E-11	8.8085E-8	15.211	
ENSMUST000000001	ENSMUST00000001558	-7.3661	7.2641	-14.103	2.8232E-11	1.0045E-7	14.965	
ENSMUST000000002	ENSMUST00000002790	-5.515	6.2451	-13.147	9.1493E-11	2.7903E-7	14.026	
ENSMUST000000008	ENSMUST00000008784	-4.7861	5.6659	-12.515	2.0679E-10	5.5181E-7	13.358	
ENSMUST000000011	ENSMUST00000011052	4.7476	4.5164	12.328	2.6488E-10	6.2828E-7	13.153	
ENSMUST000000009	ENSMUST00000009604	-5.2359	5.949	-12.211	3.0972E-10	6.6119E-7	13.023	
ENSMUST000000006	ENSMUST00000006851	-5.4857	6.2034	-11.758	5.7396E-10	1.1139E-6	12.505	
ENSMUST000000002	ENSMUST00000002786	4.24	4.3133	11.14	1.3742E-9	2.4446E-6	11.762	
ENSMUST000000003	ENSMUST00000003221	-4.7169	5.9359	-10.704	2.601E-9	4.2713E-6	11.211	
ENSMUST000000002	ENSMUST00000002531	-4.2607	5.101	-10.539	3.3282E-9	5.075E-6	10.996	
ENSMUST000000003	ENSMUST00000003471	-5.4725	6.9879	-10.471	3.6867E-9	5.2469E-6	10.907	
ENSMUST000000008	ENSMUST00000008931	-4.8455	8.1749	-10.365	4.3342E-9	5.4995E-6	10.765	
ENSMUST000000014	ENSMUST00000014648	-4.9783	5.7728	-10.358	4.3794E-9	5.4995E-6	10.756	
ENSMUST000000002	ENSMUST00000002723	-5.12	6.8968	-10.162	5.9121E-9	7.0118E-6	10.492	
ENSMUST000000015	ENSMUST00000015029	-5.0065	5.7664	-9.8508	9.6072E-9	1.0794E-5	10.063	

Previous

Proceed

Total 413 Differential expressed gene found. You can learn more about this gene by the id. Some example are below:

Transcript: ENSMUST00000001040.7 Icam4-201

Description : intercellular adhesion molecule 4, Landsteiner-Wiener blood group [Source:MGI Symbol;Acc:MGI:1925619]

Gene Synonyms : 1810015M19Rik, Cd242

Location : Chromosome 9: 20,940,669-20,941,891 forward strands.

About this transcript : This transcript has 3 exons, is annotated with 14 domains and features, is associated with 513 variant alleles and maps to 163 oligo probes.

Gene : This transcript is a product of gene ENSMUSG00000001014.7

Transcript: ENSMUST00000057279.6 Olfm12a-201

Description : olfactomedin-like 2A [Source:MGI Symbol;Acc:MGI:2444741]

Gene Synonyms : 4932431K08Rik, photomedin-1

Location : Chromosome 2: 38,821,990-38,853,765 forward strands.

About this transcript : This transcript has 8 exons, is annotated with 18 domains and features, is associated with 2059 variant alleles and maps to 186 oligo probes.

Gene : This transcript is a product of gene ENSMUSG00000046618.8

Analysis overview :



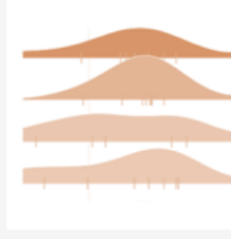
- Interactive volcano plot to display the DE features.

Volcano Plot



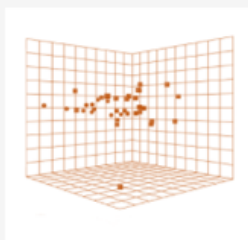
- Visualize functional categories that are enriched in a network.

Enrichment Network



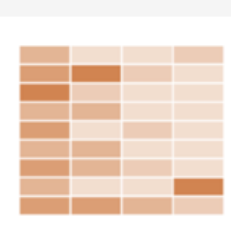
- Visualize fold-change distribution of enriched pathways

Ridgeline Chart



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

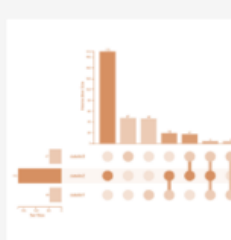
Dimension Reduction



- Interactive heatmap to explore feature abundance pattern

ORA

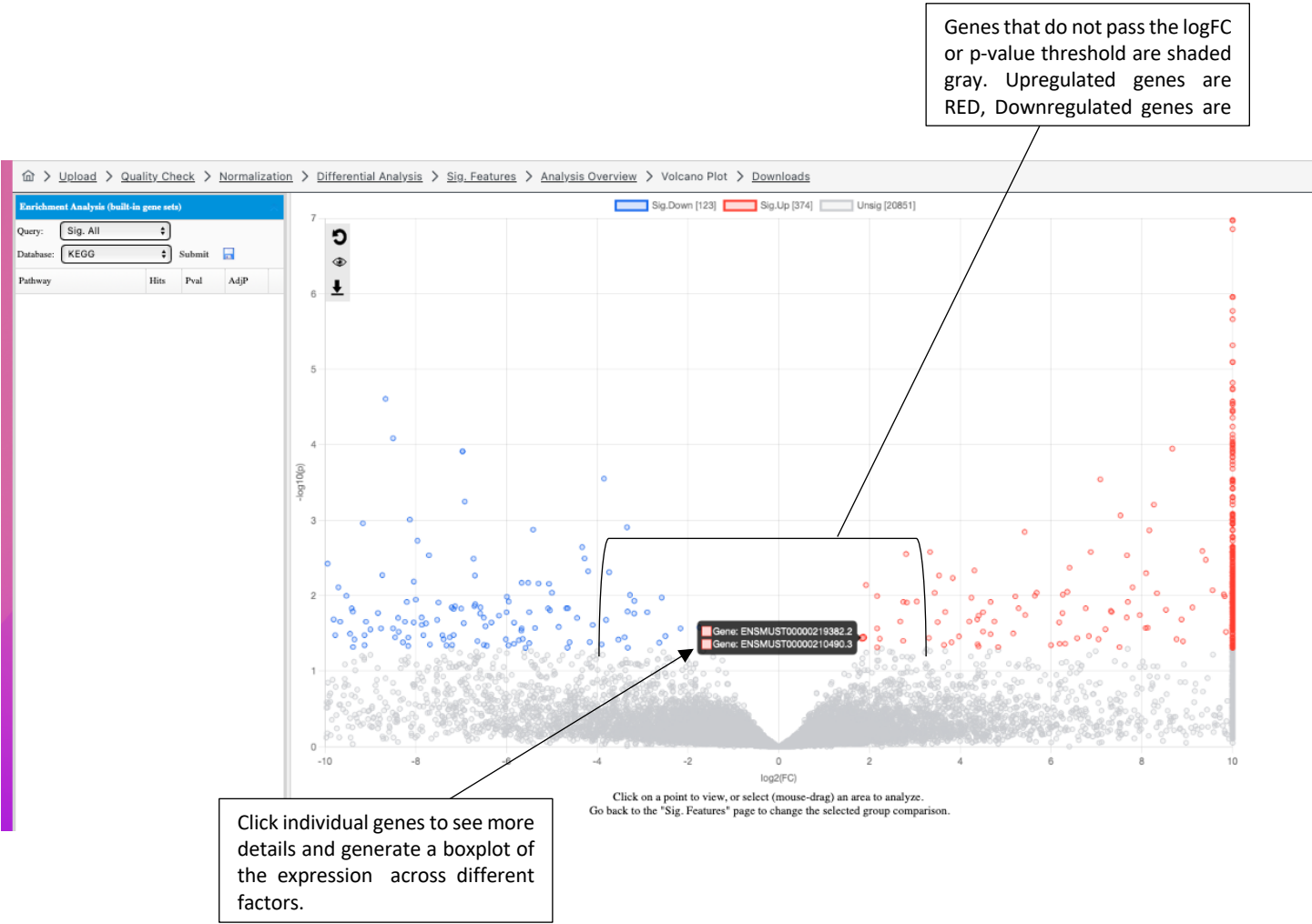
GSEA



- Visualize intersections of multiple results

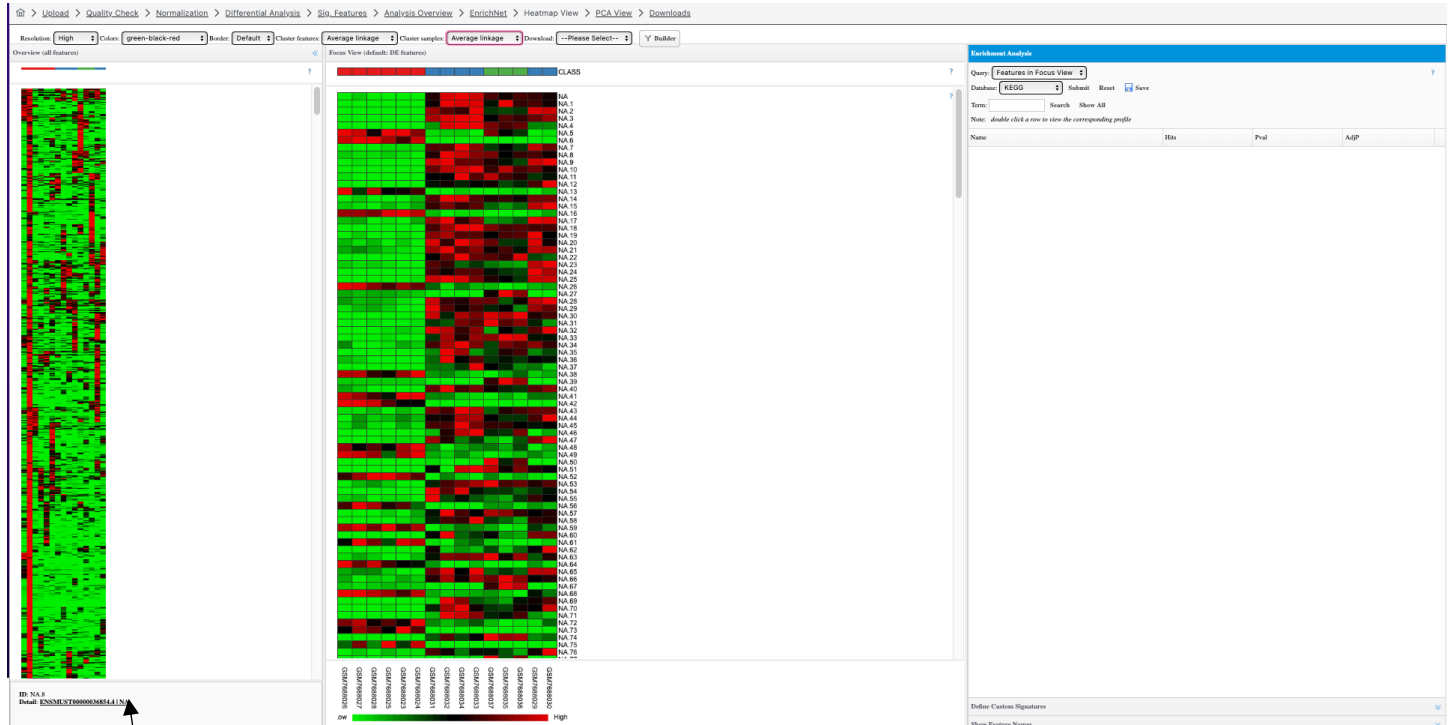
Upset Diagram

Interactive volcano plot :



ORA Heatmap clustering and visualization and Advanced heatmap functions

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.



Select a group of genes with a distinct expression pattern in the overview by dragging your mouse. They will appear in the focus view.

Detailed information about gene :

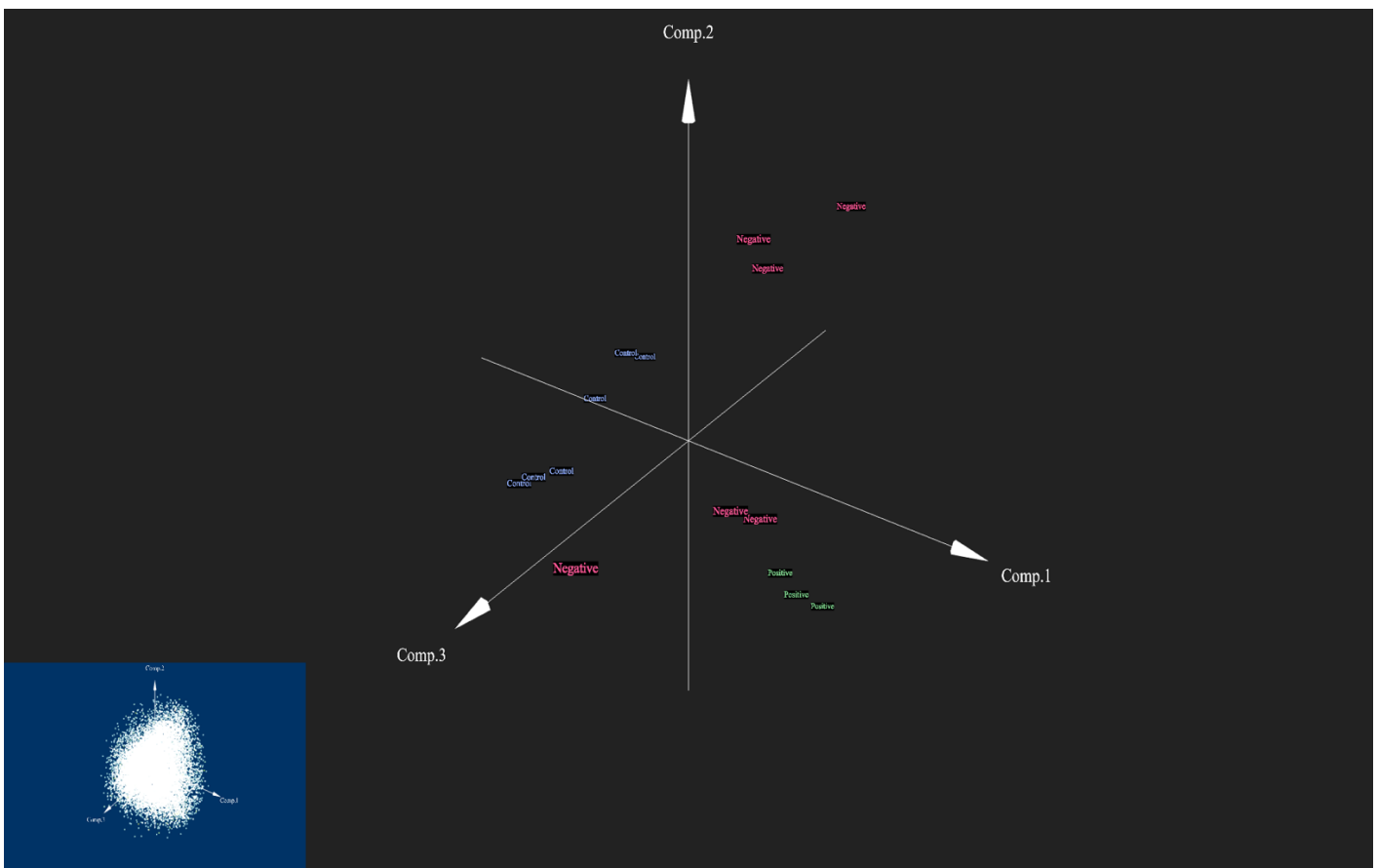
NIH National Library of Medicine National Center for Biotechnology Information				
Gene <input type="text" value="null"/> Create RSS Save search Advanced				
Gene sources: Genomic, Mitochondria, Organellas, Plasmids				
Categories: Alternatively spliced, Annotated genes, Non-coding, Protein-coding, Pseudogene				
Sequence content: CCDS, Ensembl, RefSeq, RefSeqGene				
Status: <input checked="" type="checkbox"/> Current				
Clear all Show additional filters				
Tabular 20 per page Sort by Relevance Send to				
See null minor tail protein in the Gene database null in <i>Gordonia</i> phase Twisteri <i>Hydrangea petiolaris</i> (2) All 7 Gene records				
Search results Items: 1 to 20 of 14594 See also 22008 discontinued or relocated items.				
Name/Gene ID	Description	Location	Aliases	MIM
<input type="checkbox"/> null ID: 16214819	hypothetical protein [Mycobacterium phage Gzm0]	NC_021346.1 (130750..130866)	M181_gp126, PBI_null	
<input type="checkbox"/> null ID: 32950882	5S ribosomal RNA [<i>Hydrangea petiolaris</i>]	NC_034936.1 (110326..110446)	CDC79_pgr005	
<input type="checkbox"/> null ID: 32950881	4.5S ribosomal RNA [<i>Hydrangea petiolaris</i>]	NC_034936.1 (109999..110101)	CDC79_pgr006	
<input type="checkbox"/> null ID: 29065004	minor tail protein [Gordonia phage Twisteri]	NC_031052.1 (25345..26289)	BI083_gp32, SEA_TWISTER6_32	
<input type="checkbox"/> null ID: 16213991	hypothetical protein [Mycobacterium phage Job42]	NC_021538.1 (50720..50803)	M609_gp077, PBI_JOB42_77	
<input type="checkbox"/> Htd ID: 31162	Hormone receptor 4 [<i>Drosophila melanogaster</i> (fruit fly)]	Chromosome X, NC_004354.4 (1941940..2007956)	Dmel_CG43934, CG16902, CG3600, CG42527, CG43692, CG43934, DHRA, Dmel_CG43934, Dmel_CG16902, Dmel_CG3600, Dmel_CG42527, EG-133E12.2, EG-BACH6115.1, EP(X)1232, EP1232, GRF, HRA, NRIK2, h4, null	
<input type="checkbox"/> PaGEF ID: 31224	Prototoxone-specific GEF [<i>Drosophila melanogaster</i> (fruit fly)]	Chromosome X, NC_004354.4 (2367121..2387664, complement)	Dmel_CG43947, CG14048, CG14047, CG43947, DmPaGEF, Dmel_CG43947, Dmel_CG14048, Dmel_CG14047, EG-BACH48C10.4, EG-BACH7M4.1, EG-BACH7M4.2, FBgn0264596, PaGEF, null, paGEF	
<input type="checkbox"/> KNL1 ID: 57082	kinetochore scaffold 1 [<i>Homo sapiens</i> (human)]	Chromosome 15, NC_000015.10 (40584249..40584342)	AF15Q14, CASC5, CT25, D40, MCPH4, PPP1R55, Spc7, hKNL-1, hSpC105	609173
<input type="checkbox"/> DSN1 ID: 79860	DSN1 component of MIS12 kinetochore complex [<i>Homo sapiens</i> (human)]	Chromosome 20, NC_000020.11 (36751795..36773763, complement)	C20orf172, KNL3, MIS13, dJ469A13.2, hKNL-3	609175
<input type="checkbox"/> knl-1 ID: 176164	Kinetochore null protein 1 [<i>Caenorhabditis elegans</i>]	Chromosome III, NC_003281.10 (8250632..8253943)	CELE_C02F5.1	
<input type="checkbox"/> MIS18BP1 ID: 59320	MIS18 binding protein 1 [<i>Homo sapiens</i> (human)]	Chromosome 14, NC_000014.9 (45203190..45253202, complement)	C14orf106, HSA242977, KNL2, M18BP1	618139
<input type="checkbox"/> knl-3 ID: 17669	Kinetochore NulJ [<i>Caenorhabditis elegans</i>]	Chromosome V, NC_003283.11 (11660213..1870556)	CELE_T10B5.6	
<input type="checkbox"/> Pierce1 ID: 69327	pierce1 of microtubule wall 1 [<i>Mus musculus</i> (house mouse)]	Chromosome 2, NC_000068.8 (28352013..28356336, complement)	1700007K13Rik, Rbest47	

Dimension reduction plots / PCA View :

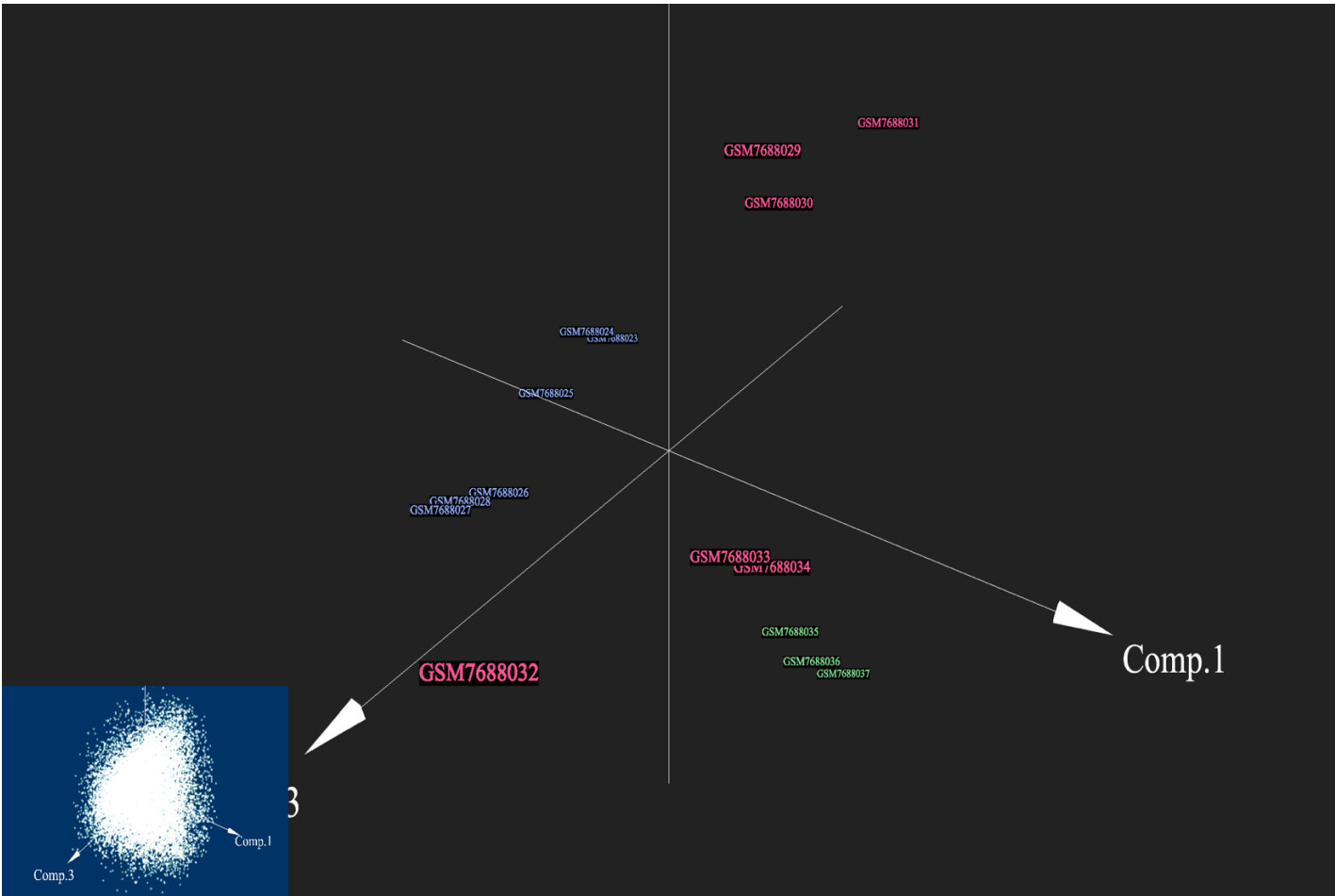
PCA and tSNE are both popular methods of capturing whole-transcriptome changes in expression in a few variables. tSNE is a stochastic method and so plots will vary slightly each time they are generated.

Score Plot (Sample Meta) :

The biplot overlays the scores of samples from the first two principal components onto a scatter plot. Each point on the biplot represents an individual sample in the dataset. The position of a sample on the biplot indicates its relative location in the reduced-dimensional space defined by the first two principal components.



Score Plot (Sample Text) : it shows the sample id text across axis.



Bio Plot Merge :

A biplot is a graphical representation that combines information from both the score plot and the loading plot in a PCA analysis. It provides a way to visualize the relationships between samples (observations) and variables (features) in a single plot.

The position of samples on the biplot can reveal patterns, clusters, or trends in the data.

Proximity of samples on the biplot suggests similarity, while samples that are farther apart are more dissimilar in terms of the first two principal components. The direction and length of the arrows (loadings) indicate which variables contribute the most to the separation observed in the score plot. Variables that are near the tips of the arrows have a higher influence on the principal components.

