BINF6310 20953 INTRO COMPU METHODS IN BIOINFORMATICS ASSIGNMENT - MODULE 10 ASSIGNMENT

TASK : RNA-Seq Data Analysis Using Kallisto and Further Data Interpretation Using Networkanalyst.Ca.

INTRODUCTION :

kallisto is a program for quantifying abundances of transcripts from RNA-Seq data, or more generally of target sequences using high-throughput sequencing reads.

ANALYSIS STEPS :

Input data file Path for Kallisto :

```
/courses/ BINF6310.202410/data/rnaseq-mus-musculus-GSE240196
```

Access the Discovery Server

```
ssh zinjuwadia.r@login.discovery.neu.edu
```

Allocate Resources

```
srun --pty --partition=courses --export=ALL --mem=8G -t 02:00:00 bash
```

Establish a working environment :

```
module load miniconda3/23.5.2
source activate binf6310
conda install – c bioconda kallisto
```

Download the Necessary File :

Visit Ensembl Mouse Genome, click on "Download Fasta" under Gene Annotation (right side of the screen), Choose the "cdna" folder and download the transcriptome file.

```
Mus_musculus.GRCm39.cdna.all.fa.
```

Run Kallisto and Build the index:

```
kallisto index -i Mus_musculus.idx Mus_musculus.GRCm39.cdna.all.fa.
```

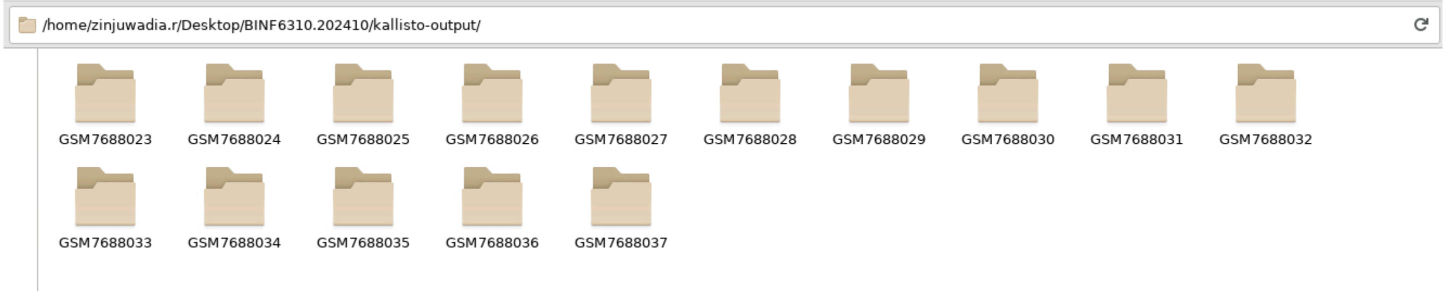
Create an output directory :

```
mkdir kallisto-output
```

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 :



/home/zinjuwadia.r/Desktop/BINF6310.202410/kallisto-output/

GSM7688023 GSM7688024 GSM7688025 GSM7688026 GSM7688027 GSM7688028 GSM7688029 GSM7688030 GSM7688031 GSM7688032
GSM7688033 GSM7688034 GSM7688035 GSM7688036 GSM7688037

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)
```

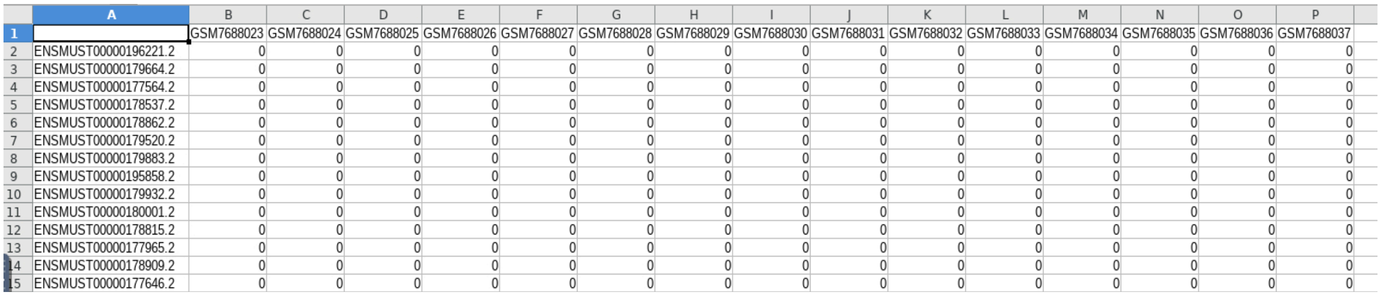
Run the R script :

```
Rscript kallisto-output-join.R
```

Download the Output File :

```
count_matrix.csv
```

Output File look Like this :

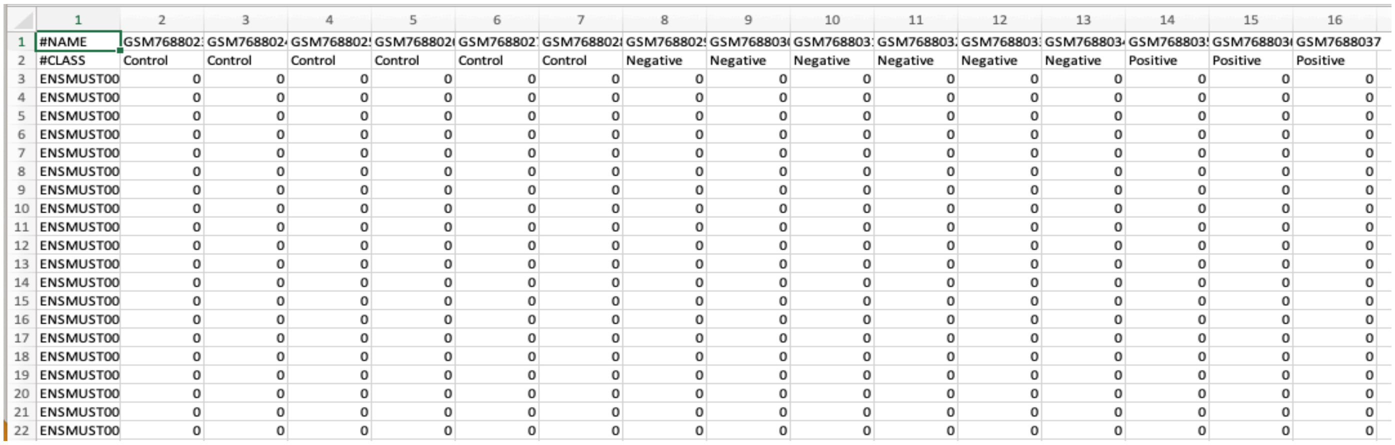


	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1		GSM7688023	GSM7688024	GSM7688025	GSM7688026	GSM7688027	GSM7688028	GSM7688029	GSM7688030	GSM7688031	GSM7688032	GSM7688033	GSM7688034	GSM7688035	GSM7688036	GSM7688037
2	ENSMUST00000196221.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	ENSMUST00000179664.2	0	0	0	0	0	0	0	0	0	0	0	0	0	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.

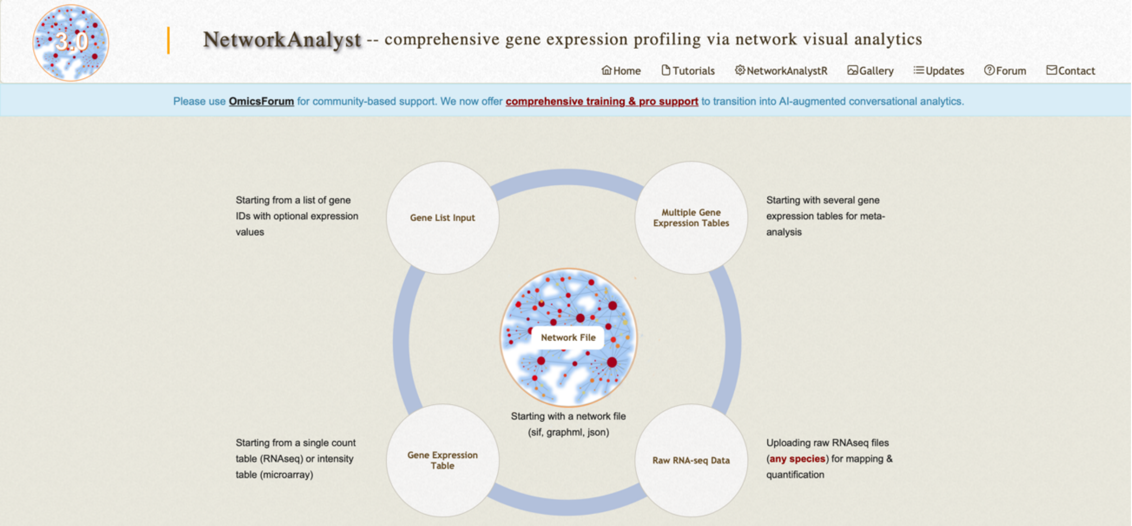
The sample name must be in the first line, followed by the metadata labels. Each Metadata starts with new line begging with "#CLASS".



	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	GSM7688037
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.



Please use OmicsForum for community-based support. We now offer comprehensive training & pro support to transition into AI-augmented conversational analytics.

Starting from a list of gene IDs with optional expression values

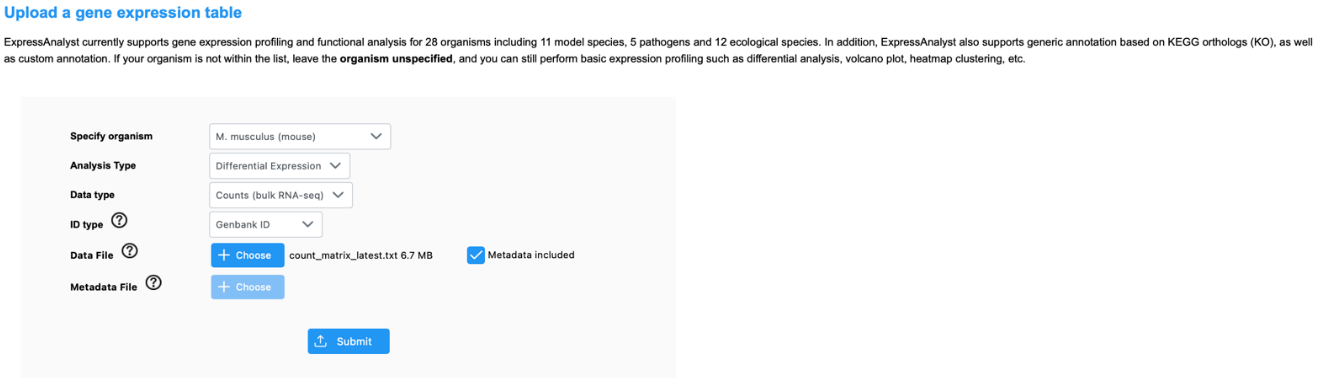
Starting with several gene expression tables for meta-analysis

Starting with a network file (sif, graphml, json)

Starting from a single count table (RNAseq) or intensity table (microarray)

Uploading raw RNAseq files (any species) for mapping & quantification

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

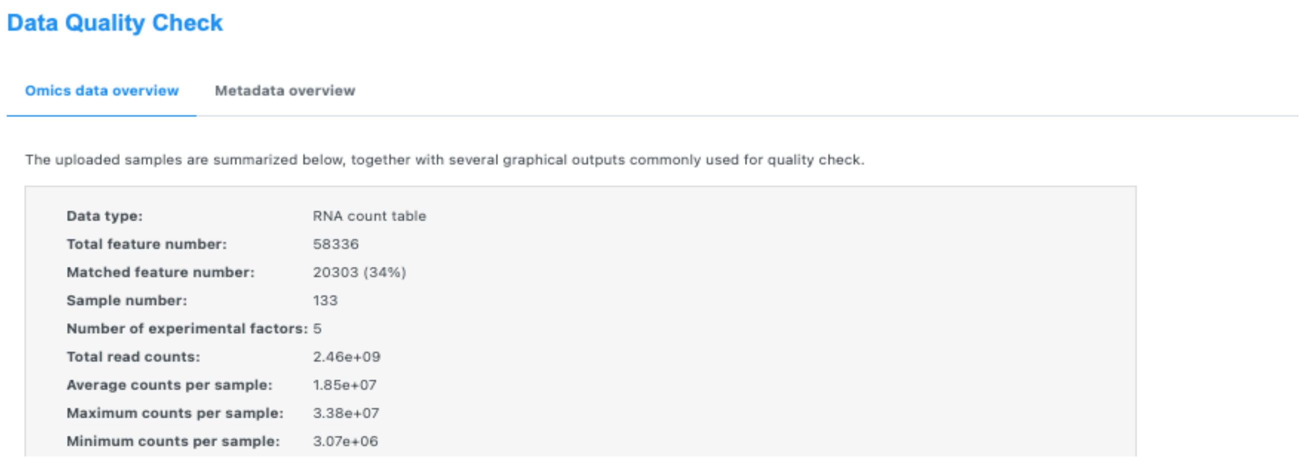


Upload a gene expression table

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

Different types of visualizations :



Data Quality Check

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

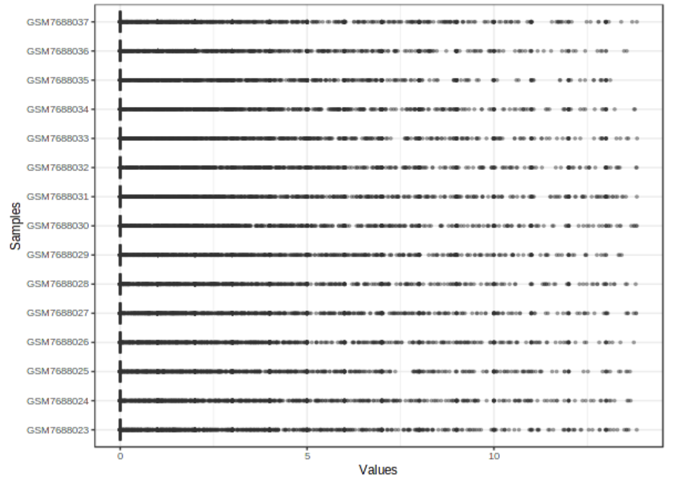
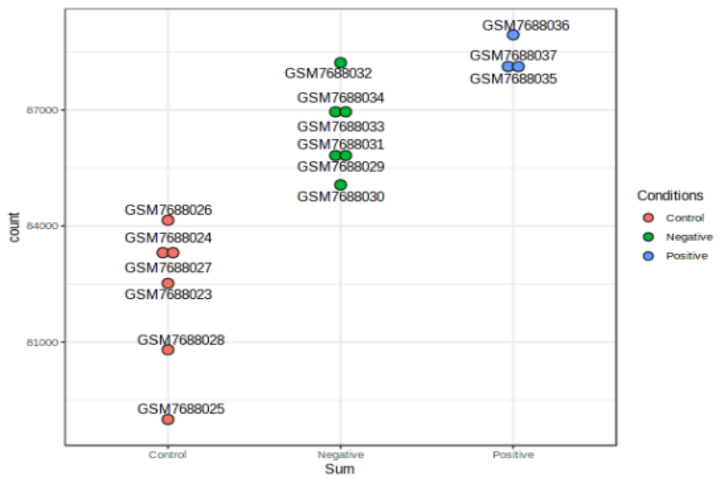
Data type:	RNA count table
Total feature number:	58336
Matched feature number:	20303 (34%)
Sample number:	133
Number of experimental factors:	5
Total read counts:	2.46e+09
Average counts per sample:	1.85e+07
Maximum counts per sample:	3.38e+07
Minimum counts per sample:	3.07e+06

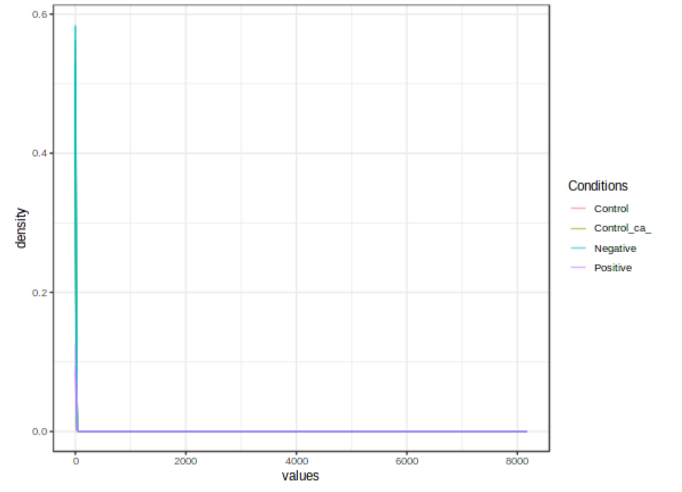
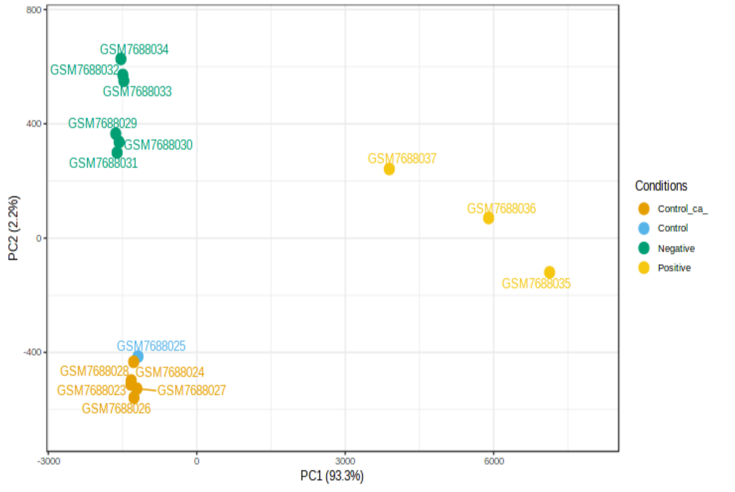
View common QA/QC plots to check the quality of the data.

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.

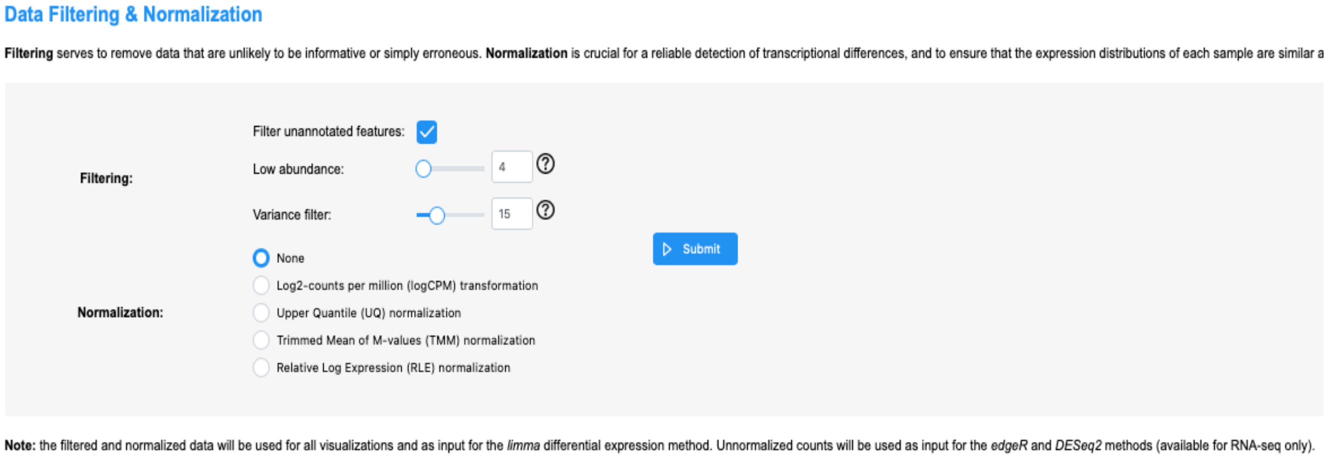
PCA plot : Principal component analysis (PCA) is becoming more popular as a method for extracting significant patterns from complicated biological information. No samples or features (variables) are discarded using PCA. Instead, by building principal components (PCs), it minimizes the overwhelming number 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.

BOXPLOT	COLUMN_SUM
	

Density Plot	PCA plot
	

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: 4 ?

Variance filter: 15 ?

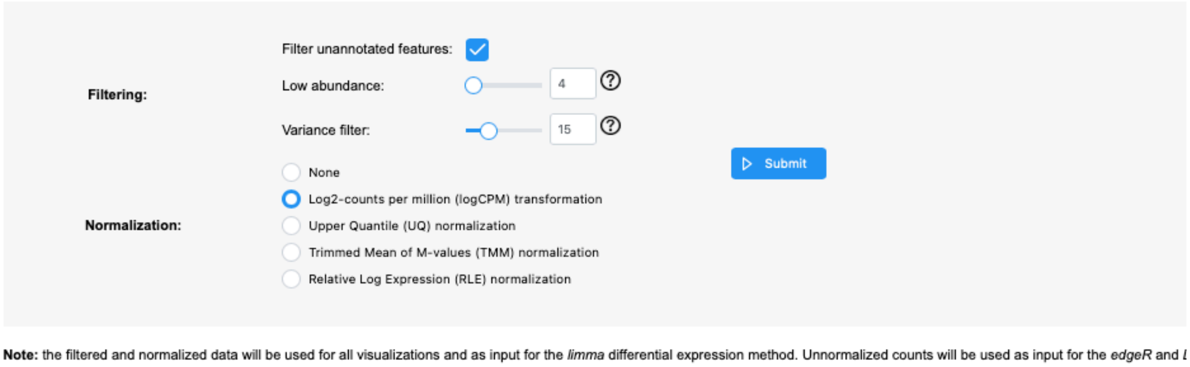
Normalization:

● 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: 4 ?

Variance filter: 15 ?

Normalization:

○ None

● Log2-counts per million (logCPM) transformation

○ Upper Quantile (UQ) normalization

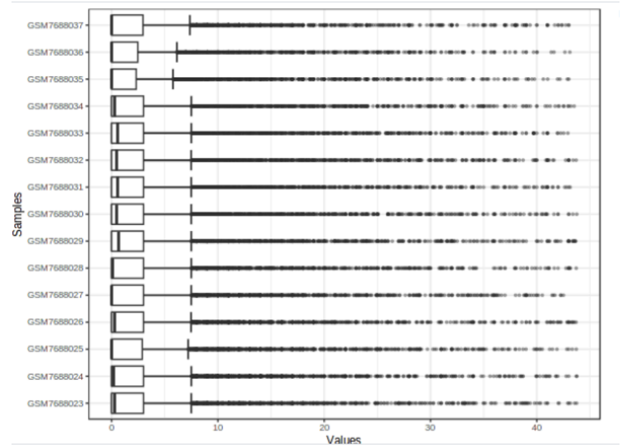
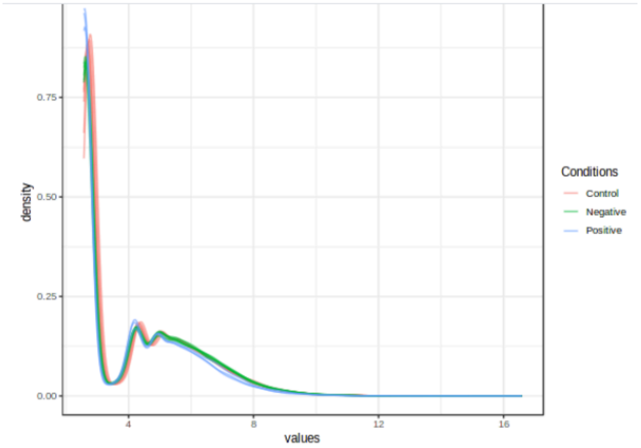
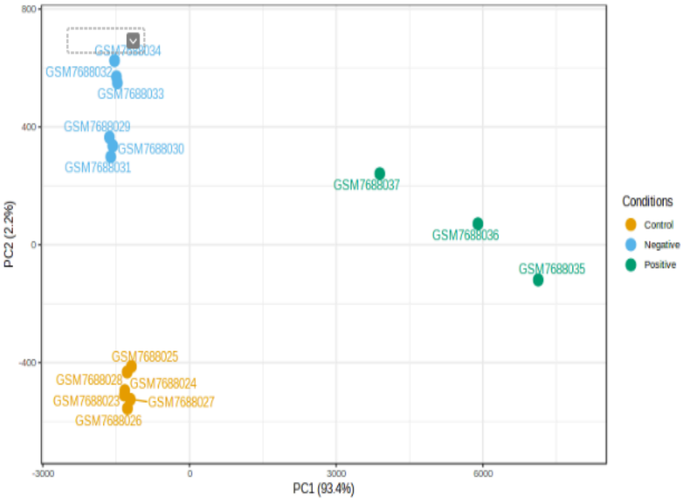
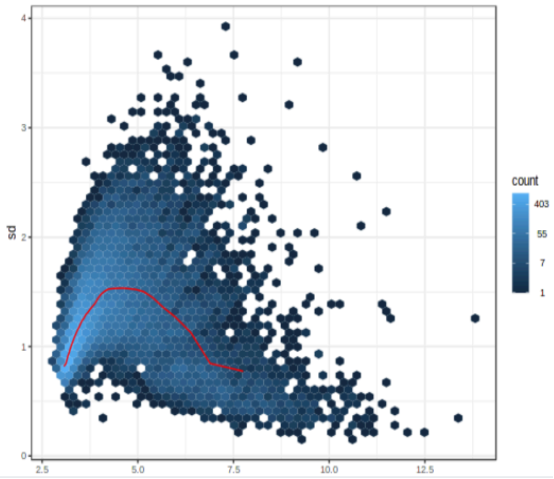
○ Trimmed Mean of M-values (TMM) normalization

○ 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 *l*

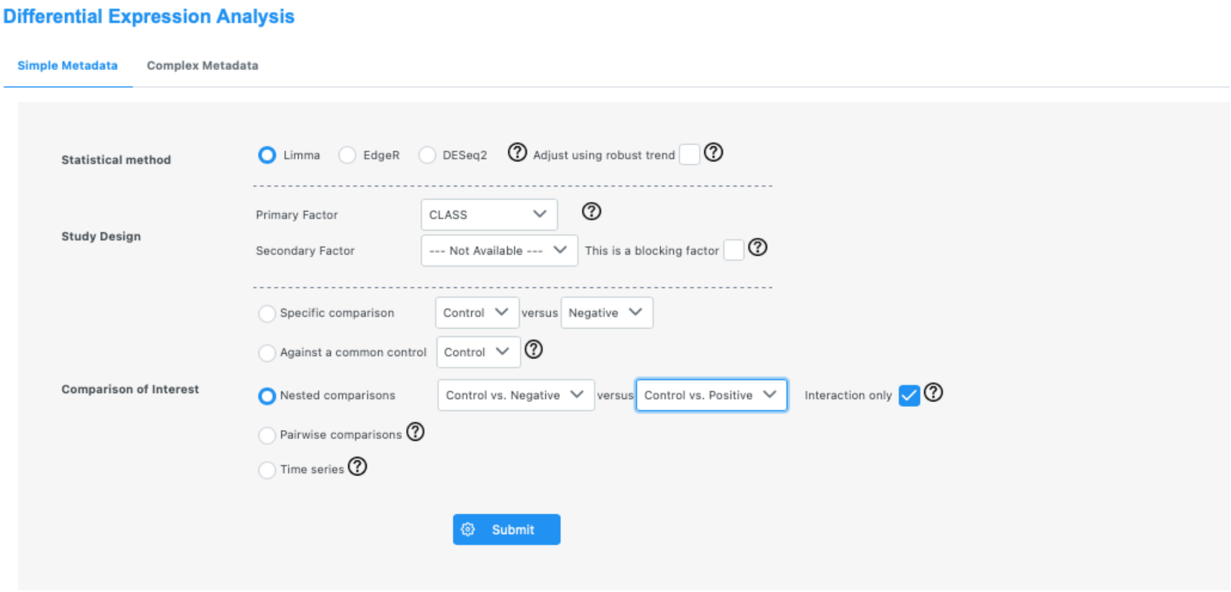
→ Diagnostic plot summarizing the standard deviation versus mean measures of read in the sample foe each feature. It checks whether there is a dependence between counts and Variance.

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

<table>
<tr><th>qc_norm_boxplot</th><th>qc_norm_density_plot</th></tr>
<tr><td>
</td><td>
</td></tr>
<tr><th>qc_norm_PCA Plot</th><th>Mean-Variance Plot (qc_norm_meanstd)</th></tr>
<tr><td>
</td><td>
</td></tr>
</table>

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

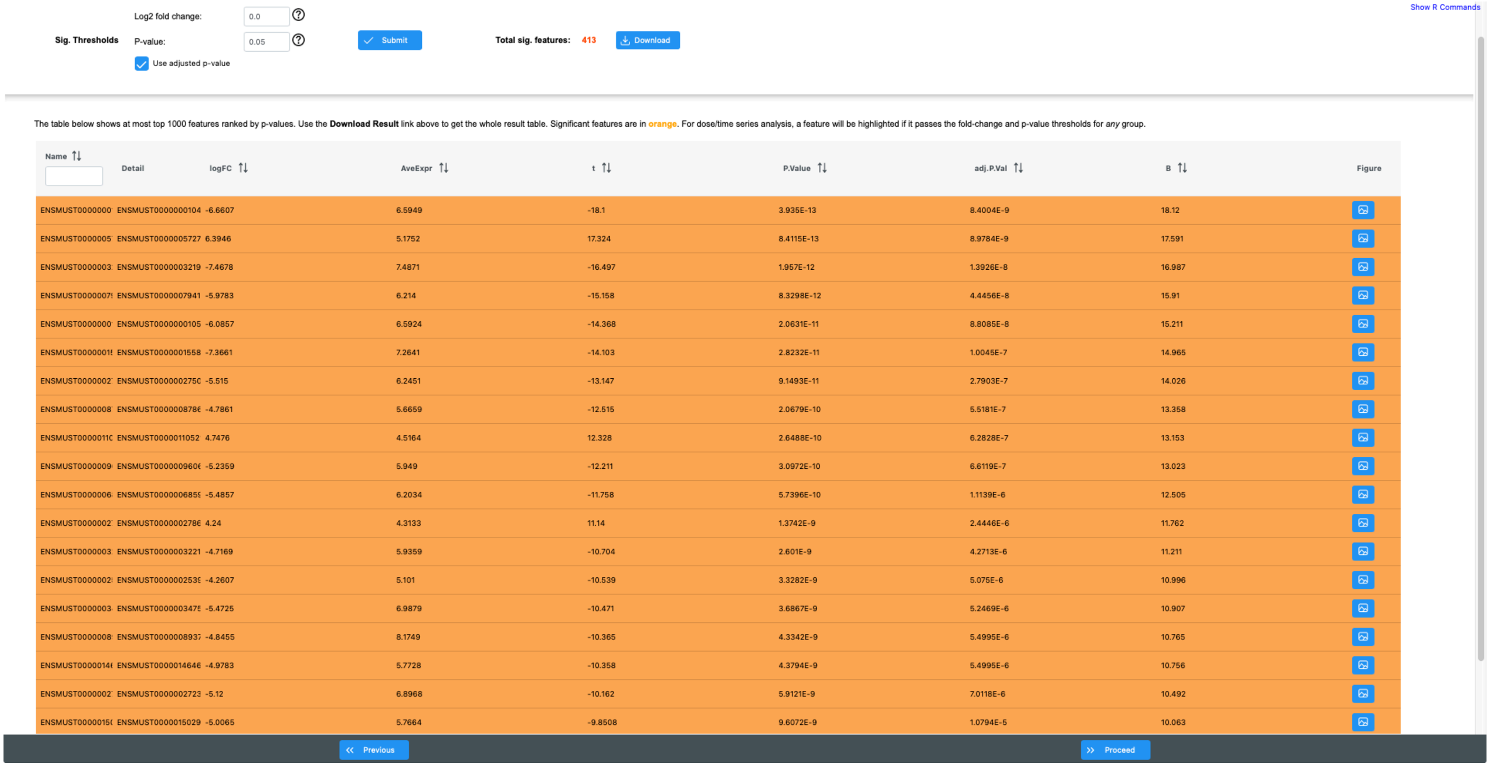
○ Against a common control Control ⓘ

● Nested comparisons Control vs. Negative versus Control vs. Positive Interaction only ☑ ⓘ

○ Pairwise comparisons ⓘ

○ Time series ⓘ

View differentially expressed genes (DEGs)



Sig. Thresholds Log2 fold change: 0.0 ⓘ P-value: 0.05 ⓘ ☑ Use adjusted p-value

Total sig. features: 413

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
ENSMUST0000000	ENSMUST0000000104	-6.6607	6.5949	-18.1	3.935E-13	8.4004E-9	18.12	🖼
ENSMUST0000005	ENSMUST0000005727	6.3946	5.1752	17.324	8.4115E-13	8.9784E-9	17.591	🖼
ENSMUST0000003	ENSMUST0000003219	-7.4678	7.4871	-16.497	1.957E-12	1.3926E-8	16.987	🖼
ENSMUST0000007	ENSMUST0000007941	-5.9783	6.214	-15.158	8.3298E-12	4.4456E-8	15.91	🖼
ENSMUST0000000	ENSMUST0000000105	-6.0857	6.5924	-14.368	2.0631E-11	8.8085E-8	15.211	🖼
ENSMUST0000001	ENSMUST0000001558	-7.3661	7.2641	-14.103	2.8232E-11	1.0045E-7	14.965	🖼
ENSMUST0000002	ENSMUST000000275	-5.515	6.2451	-13.147	9.1493E-11	2.7903E-7	14.026	🖼
ENSMUST0000008	ENSMUST000000878	-4.7861	5.6659	-12.515	2.0679E-10	5.5181E-7	13.358	🖼
ENSMUST000001	ENSMUST0000011052	4.7476	4.5164	12.328	2.6488E-10	6.2828E-7	13.153	🖼
ENSMUST0000009	ENSMUST000000960	-5.2359	5.949	-12.211	3.0972E-10	6.6119E-7	13.023	🖼
ENSMUST0000006	ENSMUST000000685	-5.4857	6.2034	-11.758	5.7396E-10	1.1139E-6	12.505	🖼
ENSMUST0000002	ENSMUST000000278	4.24	4.3133	11.14	1.3742E-9	2.4446E-6	11.762	🖼
ENSMUST0000003	ENSMUST0000003221	-4.7169	5.9359	-10.704	2.601E-9	4.2713E-6	11.211	🖼
ENSMUST0000002	ENSMUST000000253	-4.2607	5.101	-10.539	3.3282E-9	5.075E-6	10.996	🖼
ENSMUST0000003	ENSMUST000000347	-5.4725	6.9879	-10.471	3.6867E-9	5.2469E-6	10.907	🖼
ENSMUST0000008	ENSMUST000000893	-4.8455	8.1749	-10.365	4.3342E-9	5.4995E-6	10.765	🖼
ENSMUST000001	ENSMUST0000014646	-4.9783	5.7728	-10.358	4.3794E-9	5.4995E-6	10.756	🖼
ENSMUST0000002	ENSMUST0000002723	-5.12	6.8968	-10.162	5.9121E-9	7.0118E-6	10.492	🖼
ENSMUST000001	ENSMUST0000015029	-5.0065	5.7664	-9.8508	9.6072E-9	1.0794E-5	10.063	🖼

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

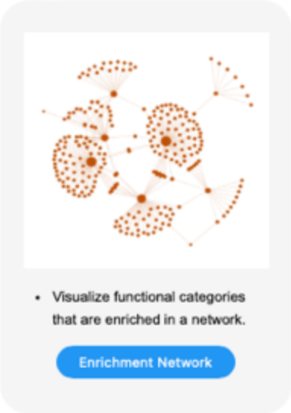
Description : intercellular adhesion molecule 4, Landsteiner-Wiener blood group [Source:MGI Symbol;Acc:MGI:1925619]
Gene Synonyms : 1810015M19Rik, Cd242
Location : Chromosome 9: 20,940,669-20,941,891 forward strands.
About this transcript : This transcript has 3 exons, is annotated with 14 domains and 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 Olfml2a-201

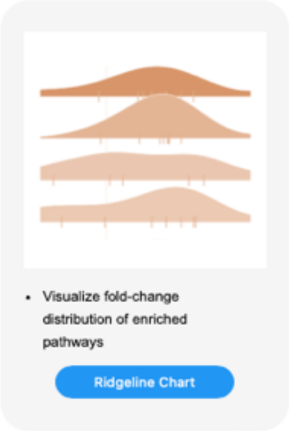
Description : olfactomedin-like 2A [Source:MGI Symbol;Acc:MGI:2444741]
Gene Synonyms : 4932431K08Rik, photomedin-1
Location : Chromosome 2: 38,821,990-38,853,765 forward strands.
About this transcript : This transcript has 8 exons, is annotated with 18 domains and features, is associated with 2059 variant alleles and maps to 186 oligo probes.
Gene : This transcript is a product of gene ENSMUSG00000046618.8

Analysis overview :

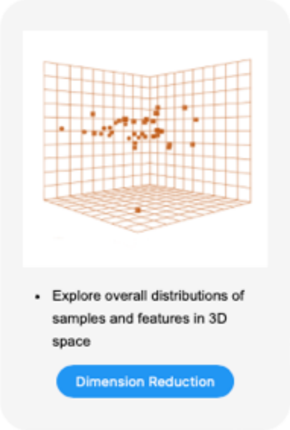
- Interactive volcano plot to display the DE features.



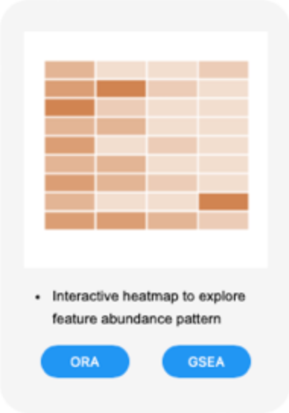
- Visualize functional categories that are enriched in a network.



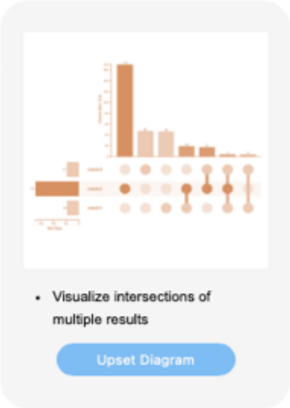
- Visualize fold-change distribution of enriched pathways



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



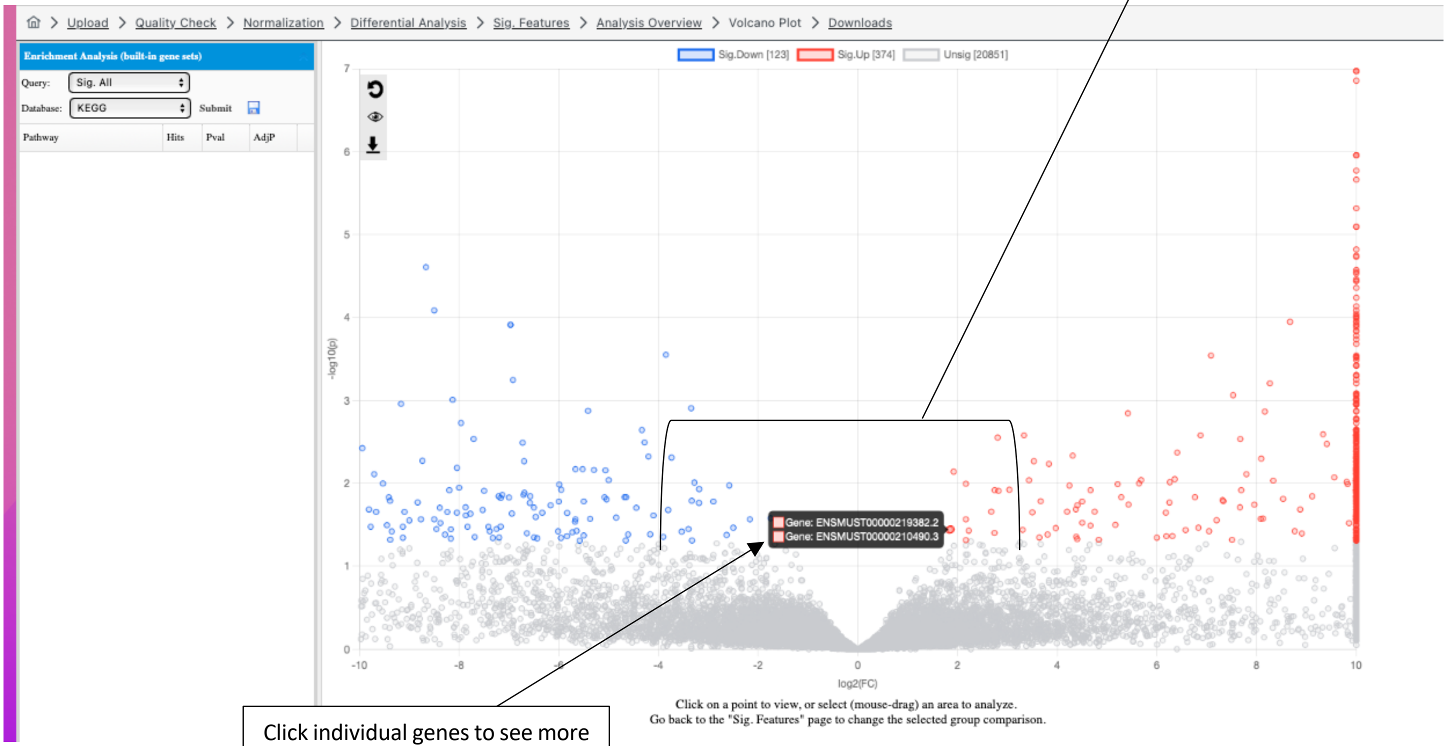
- Interactive heatmap to explore feature abundance pattern



- Visualize intersections of multiple results

Interactive volcano plot :

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

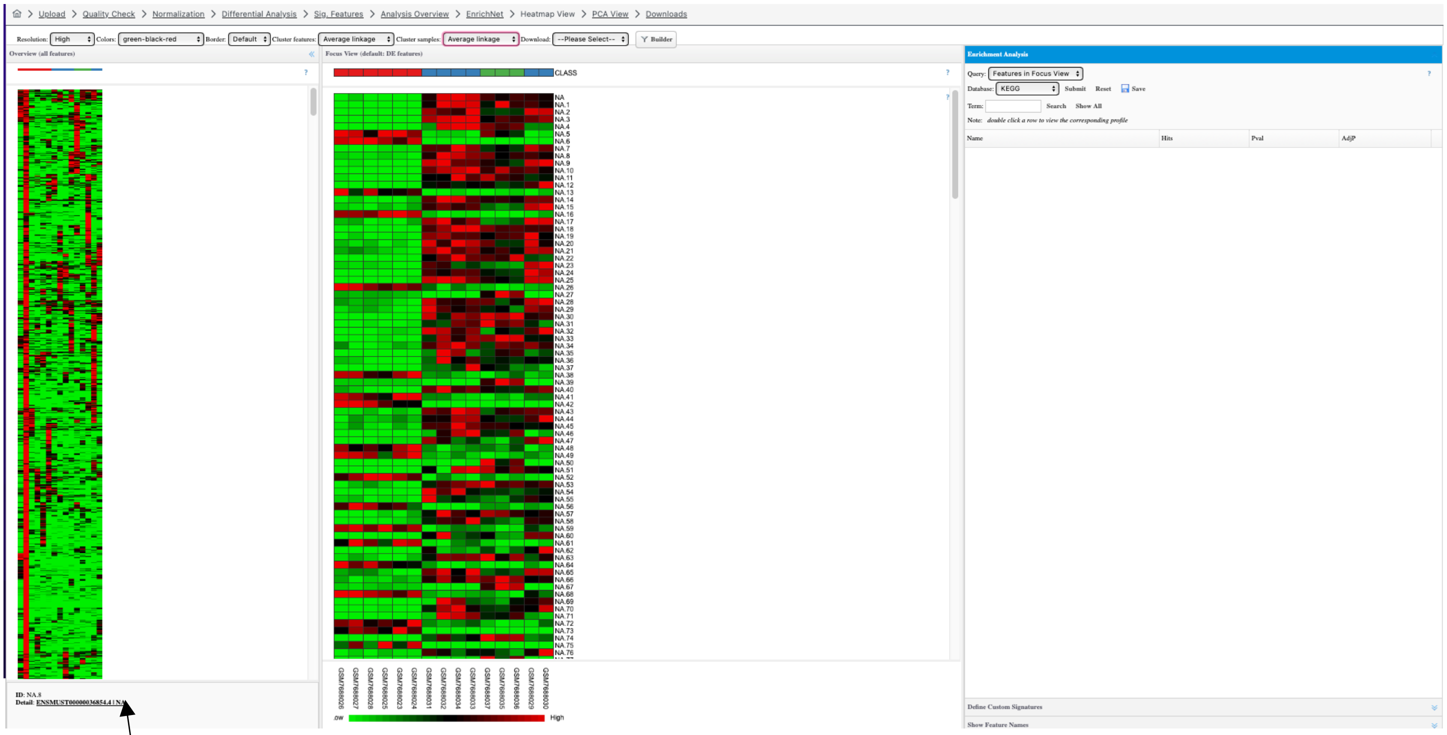

Upload > Quality Check > Normalization > Differential Analysis > Sig. Features > Analysis Overview > Volcano Plot > Downloads
Enrichment Analysis (built-in gene sets)
Query: Sig. All
Database: KEGG
Submit
Pathway
Hits
Pval
AdjP
Sig.Down [123]
Sig.Up [374]
Unsig [20851]
Gene: ENSMUST00000219382.2
Gene: ENSMUST00000210490.3
-log10(p)
0
1
2
3
4
5
6
7
log2(FC)
-10
-8
-6
-4
-2
0
2
4
6
8
10

Click on a point to view, or select (mouse-drag) an area to analyze.
Go back to the "Sig. Features" page to change the selected group comparison.

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

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.



Resolution: High Colors: green-black-red Border: Default Cluster features: Average linkage Cluster samples: Average linkage Download: --Please Select--

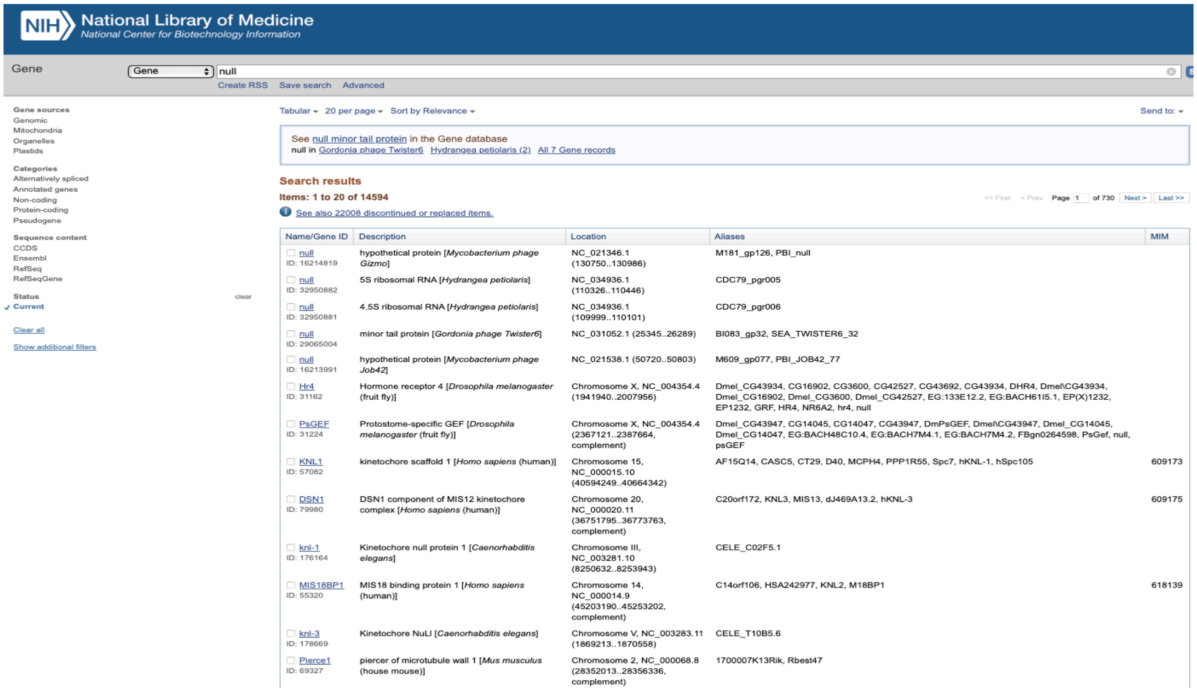
ID: NA.8
Detail: ENSMUST00000036854.4 | NA

Query: Features in Focus View
Database: KEGG
Term:
Note: *double click a row to view the corresponding profile*

Name	Hits	Pval	AdjP

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 :



Gene: Gene null

See null minor tail protein in the Gene database
null in Gordonia phage Twister6 Hydrangea petiolaris (2) All 7 Gene records

Search results

Items: 1 to 20 of 14594

Page 1 of 730

See also 22008 discontinued or replaced items.

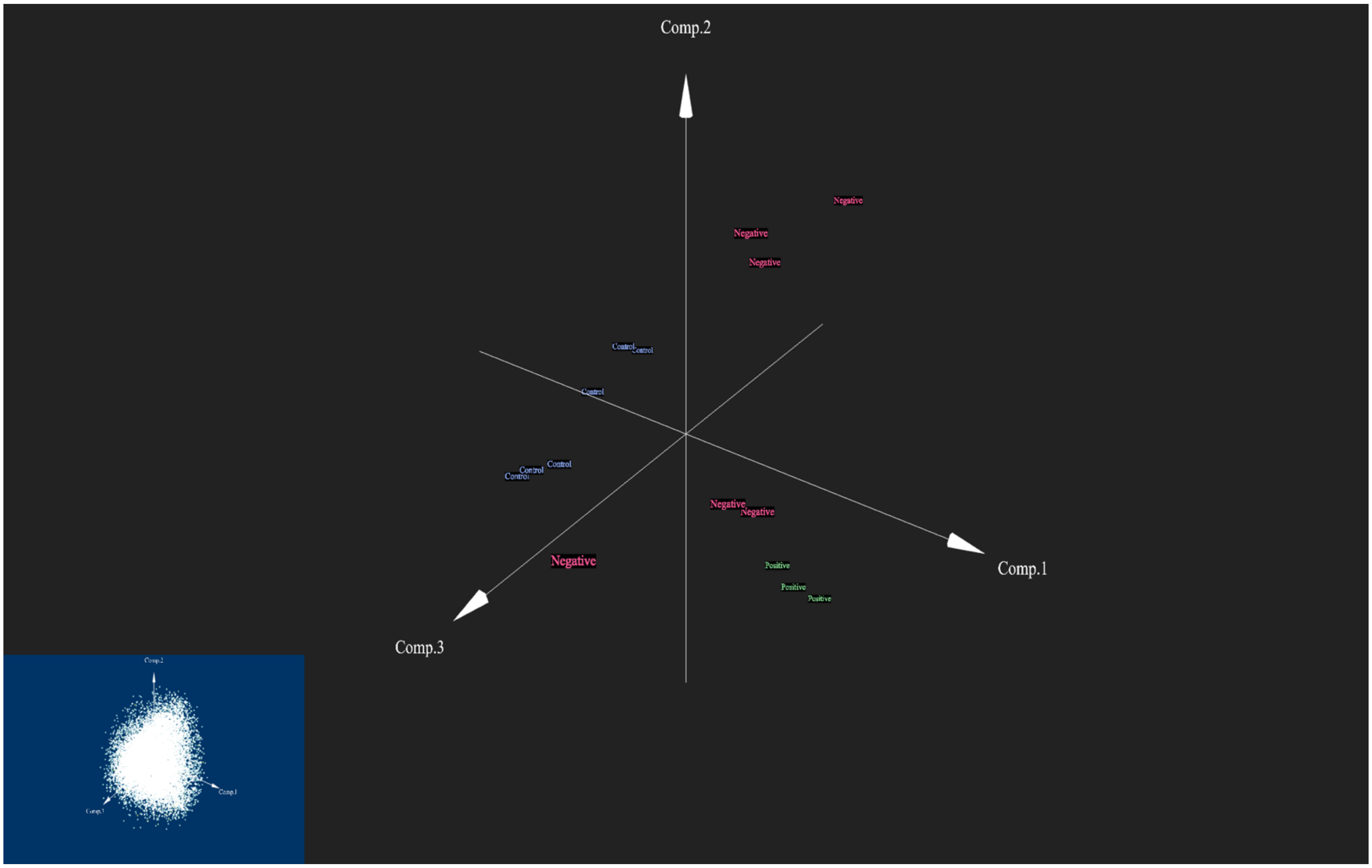
Name/Gene ID	Description	Location	Aliases	MIM
□ null ID: 16214819	hypothetical protein [*Mycobacterium phage Gizmo*]	NC_021346.1 (130750..130986)	M181_gp126, PBI_null	
□ null ID: 32950882	5S ribosomal RNA [*Hydrangea petiolaris*]	NC_034936.1 (110326..110446)	CDC79_pgr005	
□ null ID: 32950881	4.5S ribosomal RNA [*Hydrangea petiolaris*]	NC_034936.1 (109999..110101)	CDC79_pgr006	
□ null ID: 29065004	minor tail protein [*Gordonia phage Twister6*]	NC_031052.1 (25345..26289)	BI083_gp32, SEA_TWISTER6_32	
□ null ID: 16213991	hypothetical protein [*Mycobacterium phage Job42*]	NC_021538.1 (50720..50803)	M609_gp077, PBI_JOB42_77	
□ Hr4 ID: 31162	Hormone receptor 4 [*Drosophila melanogaster* (fruit fly)]	Chromosome X, NC_004354.4 (1941940..2007956)	Dmel_CG43934, CG16902, CG3600, CG42527, CG43692, CG43934, DHR4, Dmel\CG43934, Dmel_CG16902, Dmel_CG3600, Dmel_CG42527, EG:133E12.2, EG:BACH61I5.1, EP(X)1232, EP1232, GRF, HR4, NR6A2, hr4, null	
□ PsGEF ID: 31224	Protostome-specific GEF [*Drosophila melanogaster* (fruit fly)]	Chromosome X, NC_004354.4 (2367121..2387664, complement)	Dmel_CG43947, CG14045, CG14047, CG43947, DmPsGEF, Dmel\CG43947, Dmel_CG14045, Dmel_CG14047, EG:BACH48C10.4, EG:BACH7M4.1, EG:BACH7M4.2, FBgn0264598, PsGef, null, psGEF	
□ KNL1 ID: 57082	kinetochore scaffold 1 [*Homo sapiens* (human)]	Chromosome 15, NC_000015.10 (40594249..40664342)	AF15Q14, CASC5, CT29, D40, MCPH4, PPP1R55, Spc7, hKNL-1, hSpc105	609173
□ DSN1 ID: 79980	DSN1 component of MIS12 kinetochore complex [*Homo sapiens* (human)]	Chromosome 20, NC_000020.11 (36751795..36773763, complement)	C20orf172, KNL3, MIS13, dJ469A13.2, hKNL-3	609175
□ knl-1 ID: 176164	Kinetochore null protein 1 [*Caenorhabditis elegans*]	Chromosome III, NC_003281.10 (8250632..8253943)	CELE_C02F5.1	
□ MIS18BP1 ID: 55320	MIS18 binding protein 1 [*Homo sapiens* (human)]	Chromosome 14, NC_000014.9 (45203190..45253202, complement)	C14orf106, HSA242977, KNL2, M18BP1	618139
□ knl-3 ID: 178669	Kinetochore NuLl [*Caenorhabditis elegans*]	Chromosome V, NC_003283.11 (1869213..1870558)	CELE_T10B5.6	
□ Pierce1 ID: 69327	piercer of microtubule wall 1 [*Mus musculus* (house mouse)]	Chromosome 2, NC_000068.8 (28352013..28356336, complement)	1700007K13Rik, Rbest47	

Dimension reduction plots / PCA View :

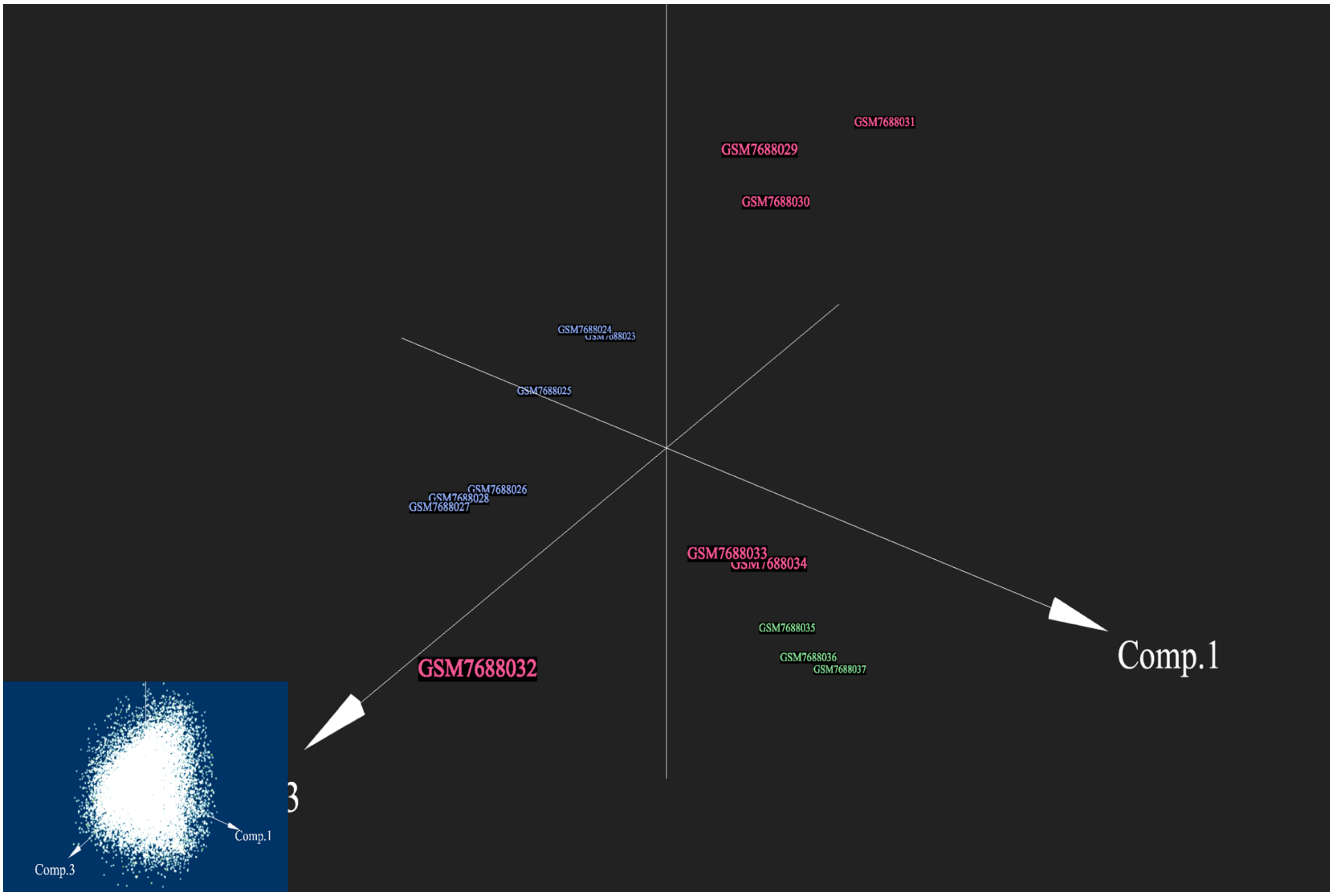
PCA and tSNE are both popular methods of capturing whole-transcriptome changes in expression in a few variables. tSNE is a stochastic method and so plots will vary slightly each time they are generated.

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.

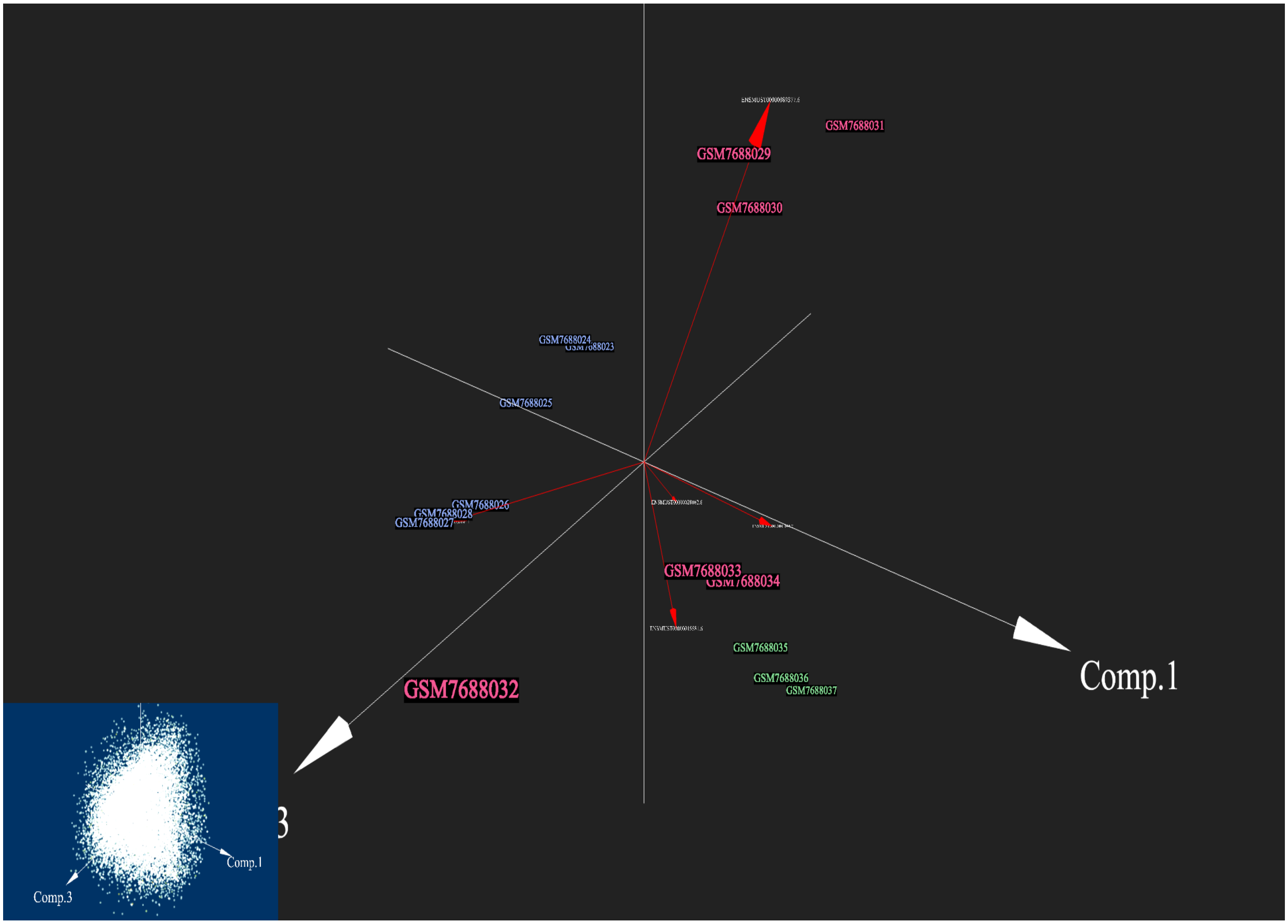


Bio Plot Merge :

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.


ENSMUST00000089377.6
GSM7688031
GSM7688029
GSM7688030
GSM7688024
GSM7688023
GSM7688025
GSM7688026
GSM7688028
GSM7688027
ENSMUST00000028062.8
GSM7688033
GSM7688034
ENSMUST00000015581.6
GSM7688035
GSM7688036
GSM7688037
Comp.1
GSM7688032
Comp.1
Comp.3