

Problem statement for Mini-project

The goal of this project is to identify genes that show differential expression between disease and healthy samples or between two different environmental conditions or drug treatment vs control. You can choose any data set from NCBI GEO and perform a differential gene expression analysis between two different conditions or two different samples.

Each of you will work with a unique dataset (with different dataset ID) and do your data analysis. So, before you start, you need to fill up an excel form with **your dataset ID**. This should be done by 12th April, 2025 in the following google sheet : <https://docs.google.com/spreadsheets/d/1vePCUZTFbk1934Uep0VgtGaH46YIDI-LOPeIa7ZsSUM/edit?usp=sharing>

Specifically, you would have to do the following analysis:

- a) Do a differential gene expression analysis using DESeq2
- b) Identify genes showing differential expression with FDR cut-off of 10% along with log2 fold change (Make a list)
- c) Provide your plots and inferences.
- d) Submit a report along with your results and codes by 20th April, 2025 midnight by uploading results in google drive:
<https://drive.google.com/drive/folders/1TVBg5gY2RDSVF8wgIKukiXDj8D3nU6Lw?usp=sharing>

Here are some stepwise instructions for your help:

1. Take any dataset from NCBI GEO datasets section <https://www.ncbi.nlm.nih.gov/gds>
2. Search name of any organism or disease that you are interested in
3. VERY IMPORTANT: Choose a dataset for your analysis only if the ‘**Type:**’ field shows ‘**Expression profiling by high throughput sequencing**’
4. Download raw count data, because this is required for DESeq2 analysis

Examples

Searching the term "yeast"

Dataset ID: GSE133214

<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE133214>

RNA sequencing of *S. Cerevisiae* treated with the Vacuolar H⁺-ATPase inhibitor concanamycin A (concA)

(Submitter supplied) RNA sequencing results identifying transcript abundance changes that occur after treating budding yeast with V-ATPase inhibitors

Organism: Saccharomyces cerevisiae

Type: Expression profiling by high throughput sequencing

Platform: GPL17342 12 Samples

Searching the term "cancer"

Dataset ID: GSE106775

<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE106775>



Glioblastoma Hijacks Microglial Gene Expression to Support Tumor Growth

(Submitter supplied) Background Glioblastomas are the most common and lethal primary brain tumors. Microglia, the resident immune cells of the brain survey their environment and respond to pathogens, toxins, and tumors. Glioblastoma cells communicate with microglia, in part by releasing extracellular vesicles (EVs). Despite the presence of large numbers of microglia in glioblastoma, the tumors continue to grow, and these neuroimmune cells appear incapable of keeping the tumor in check. more...

Organism: Mus musculus

Type: Expression profiling by high throughput sequencing

Download the **raw count file**. This will contain your data. You can also look at other files.

Download family	Format
SOFT formatted family file(s)	SOFT 
MINiML formatted family file(s)	MINiML 
Series Matrix File(s)	TXT

Supplementary file	Size	Download	File type/resource
GSE106775_WT_Trem2_KO_Apoe_KO.txt.gz	440.5 Kb	(ftp) (http)	TXT
GSE106775_star_genes_erc.counts_microglia.txt.gz	353.5 Kb	(ftp) (http)	TXT

The report should be a pdf file and should be named as
“<Rollno>_BioinformaticsLab_report2025.pdf”

Report format:

Objective – The question that you are answering through the analysis

Dataset – Briefly describe the dataset as you understand the data (Do not copy from website)

Results – Results from your analysis and the plots

Inferences – Inferences drawn from your analysis

Annexure – Should contain your codes

Note: Copying codes and reports from each other will lead to zero marks in the evaluation of the mini-project.