

Anushka Sinha

Nov. 1, 2023

02-718 HW 3

**Assume you work as a bioinformatician in a lab that studies aging. Your PI gives you a processed dataset of RNA-seq data and requests that you conduct differential gene expression analysis and gene set enrichment analysis.**

### **1. (6 points) Sequencing techniques**

**You recall from your computational medicine class that gene expression can be measured by microarrays and RNA sequencing. The data you have is from RNA-seq. You decide to research these two different approaches, microarrays vs. RNA-seq. Explain each approach in two or three sentences. What are the pros and cons of each approach?**

RNA-seq (short for RNA sequencing) is a cutting-edge sequencing technique that measures the amount of RNA in a sample and can be used to identify which genes are active or inactive. The process involves extracting RNA from the sample and converting it into cDNA through reverse transcription. The cDNA is then broken down and amplified with adapters to be sequenced. Some of the advantages of RNA-seq include its sensitivity in detecting gene expression, but it is also time-consuming, expensive, requires a lot of computing power, and has challenges in detecting gene expression accurately.

Using hybridization, microarrays can measure the concentration of nucleic acid sequences and are used for gene detection. In order to determine whether or not particular target molecules are present in a biological sample, it immobilizes a large number of molecular probes on a solid surface. In biological research, microarrays have been widely used for tasks like genotyping, gene expression profiling, and determining molecular interactions. Two advantages of microarrays are its cost-effectiveness and standardization. The fact that it depends on hybridization, which may not be specific, and the wide variation found in low-expressed gene samples are some drawbacks.

### **2. Differential Gene Expression Analysis**

**DESeq2 is a widely-used package for analyzing RNA-seq data. It is originally written in the R language**

**(<https://bioconductor.org/packages/devel/bioc/vignettes/DESeq2/inst/doc/DESeq2.html>). However, you can find a Python implementation of the package named**

**PyDESeq2**

**[https://pydeseq2.readthedocs.io/en/latest/auto\\_examples/plot\\_step\\_by\\_step.html](https://pydeseq2.readthedocs.io/en/latest/auto_examples/plot_step_by_step.html).**

You have decided to use either R or Python as your programming language and will conduct the differential expression analysis using the DESeq2 or PyDESeq2 package.

**2.1 (4 points) Data loading. Load the dataset named HW3\_Data.csv which you can download from Canvas. Do you need to transpose this read count table? Why or why not?**

Yes, I need to transpose the read count table because it is more reasonable to visualize the data with rows representing samples and columns representing genes.

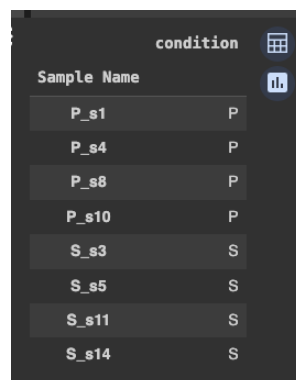
Additionally, to run PyDESeq2, it is necessary to provide a count matrix with rows representing the samples and the columns representing genes, with its dimension being 'number of samples x number of genes'.

**2.2 (5 points) Data filtering. You decide to filter out genes with fewer than 5 total read counts. How many genes have been filtered out? How many remain?**

37763 genes were filtered out. 22214 genes remain.

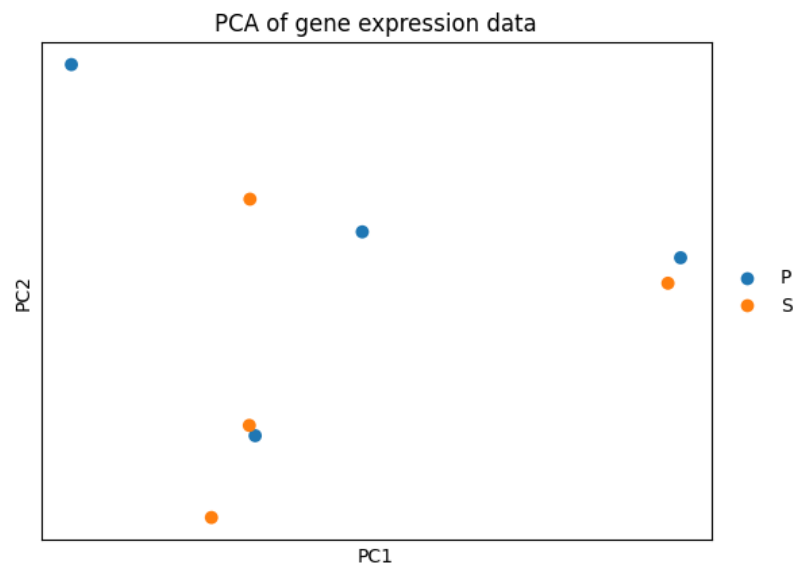
**2.3 (4 points) Metadata. DESeq2 stores virtually all information related to your experiment in a specific object of the class DESeqDataSet. To create a DESeqDataSet object, two mandatory arguments are required:**

- A count matrix with the shape [number of samples, number of genes], containing the read counts.
- A metadata matrix with the shape [number of samples, number of groups], containing sample annotations used to segregate the data into cohorts. While you have the count data, your PI hasn't provided the metadata. Fortunately, you observe patterns in the sample names. Each sample is named as: group + sample\_id. For example, 'P\_s4' implies that the sample with ID 4 is from group P. Leveraging this pattern, you believe you can generate the metadata matrix. Please print your metadata matrix and paste it here. (Hint: Its dimensions should be 8x2.)



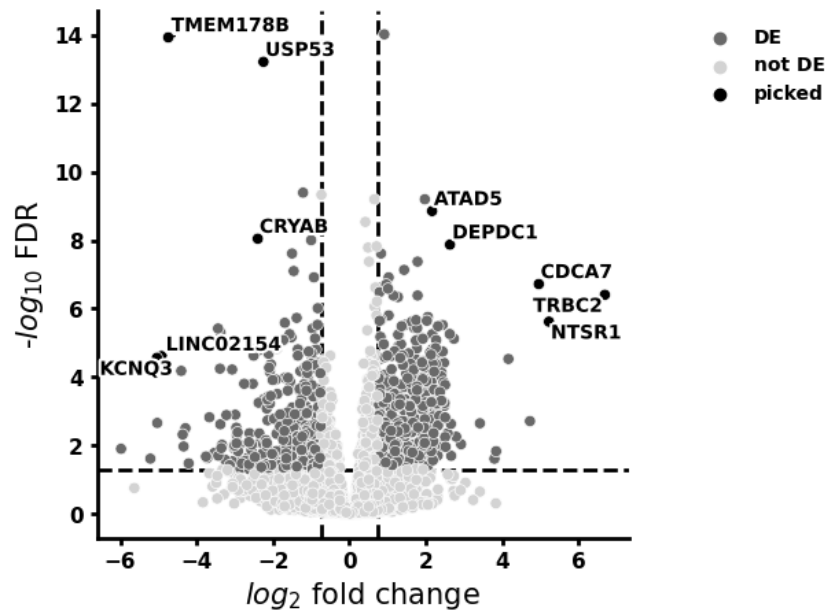
Sample Name	condition
P_s1	P
P_s4	P
P_s8	P
P_s10	P
S_s3	S
S_s5	S
S_s11	S
S_s14	S

**2.4 (6 points) PCA.** An important step before diving into the identification of differentially expressed genes is to check whether expectations about the basic global patterns are met. For example, technical and biological replicates should show similar expression patterns while the expression patterns of, say, two experimental conditions should be more dissimilar. You decide to use the Principal components analysis (PCA) to assess the similarity of expression patterns. Please insert your PCA plot here. What conclusions can you draw from your plot?



According to the PCA plot above, I can conclude that the groups 'P' and 'S' are not linearly separable because a single group cannot divide the gene expression data. There is also significant overlap between the two groups and there is a high amount of noise and randomness.

**2.5 (20 points) Identification of differentially expressed (DE) genes.** After you create a `DeseqDataSet` object, you can run the `deseq2()` method to fit dispersions and log2 fold change (LFC) estimates. Then you can conduct statistical tests to compute p-values and adjusted p-values for differential expression. To determine significant differentially expressed genes, you set this threshold:  $\text{padj} < 0.05$  and the absolute value of  $\text{log2FoldChange} > 1$ . Volcano plots are commonly used to display the results of RNA-seq. Such a plot is a type of scatterplots that show statistical significance (P value) versus the magnitude of change (fold change). It enables a quick visual identification of the genes with large fold changes that are also statistically significant. How many differentially expressed genes did you identify? Draw a volcano plot of your differentially expressed genes and paste it here. How to interpret the volcano plot? (Hint: consider statistical significance, up/down regulation, etc.)



I identified 603 differentially expressed genes.

The x-axis of the volcano plot is log<sub>2</sub> fold change which represents the magnitude of the change and the y-axis is log<sub>10</sub> FDR which represents the statistical significance of different genes. The volcano plot shows the genes with the most significant differential expression at the top and the ones with the least significant differential expression at the bottom. The left side of the volcano plot shows downregulated (negative log<sub>2</sub> fold change) genes and the right side (positive log<sub>2</sub> fold change) shows the upregulated genes. Additionally, the points closer to the center of the log<sub>2</sub> fold change axis represent smaller fold change which indicates that they were not as differentially expressed whereas the points farther from the center of the log<sub>2</sub> fold change axis represent larger fold change and thereby indicating a comparatively larger differential expression.

### 3. Enrichment analysis of genes

**3.1 (5 points) You provided your PI with the differential gene expression results. S/he asked you to continue with functional enrichment analysis of genes, such as Gene Ontology (GO) enrichment analysis and KEGG pathway enrichment analysis. Why does your PI want you to combine DE analysis with functional enrichment analysis of the genes?**

Combining DE analysis with the functional enrichment analysis of the genes allows for a deeper understanding of the biological processes, particularly in the context of researching with high-throughput omics data such as RNA-seq. Using DE analysis, we can determine which genes are being up or down regulated under different experimental conditions. Although this information is valuable in understanding the functions of various genes, it still doesn't provide immediate insight into the biological pathways. This is where the functional enrichment analysis comes in and

helps prioritize the genes that are pertinent to relevant biological pathways. It can also help validate the results of DE analysis by demonstrating where the observed gene expression changes are significant.

**3.2 (5 points) GSEA Preranked.** As described in the GSEA paper (Lec 9), “the goal of GSEA is to determine whether members of a gene set S tend to occur toward the top (or bottom) of the list L, in which case the gene set is correlated with the phenotypic class distinction.” The GSEA requires a ranked list of genes that you supply. Rank the genes in the provided dataset based on statistics of your choice. Which ranking statistics did you choose? Why?

I used ‘stat’ as the ranking statistics. stat is Wald statistic which is used to confirm whether a set of independent variables are significant for the model or not. A greater difference between the groups being compared is represented by a larger absolute value of the test statistic. In this case, it evaluates whether the gene expression value has a significant impact on the outcome (condition: P or S). Therefore, ranking the members of the significant genes set using the wald statistic ‘stat’ helped me determine the order in which these gene expression values are significant for the outcome.

**3.3 (4 points) Libraries.** In order to use the GSEA, you need to provide libraries for gene set enrichment analysis. There are many gene set libraries/databases available, such as the Gene Ontology (GO) and The Kyoto Encyclopedia of Genes and Genomes (KEGG). Briefly introduce each database (GO and KEGG).

Gene Ontology (GO) is a comprehensive database representing gene and gene products across all species. It aims to maintain and develop a standardized vocabulary of gene and gene products across all species and also enable functional interpretation of the experimental data. It is organized into three domains: Biological Process, Cellular Component, and Molecular Function.

[https://en.wikipedia.org/wiki/Gene\\_Ontology](https://en.wikipedia.org/wiki/Gene_Ontology)

The Kyoto Encyclopedia of Genes and Genomes (KEGG) is a resource for systematic analysis of gene functions, linking genomic information with higher order functional information. It is composed of various databases such as GENES, PATHWAY, and KEGG. GENES contains gene catalogs for every fully sequenced genomes. PATHWAY contains graphical depictions of several biological functions such as metabolism, membrane transport, signal transduction, and cell cycle. KEGG offers JAVA graphical tools for altering expression maps, and comparing two genome maps. It also provides computational tools for sequence comparison, graph comparison, and path computation.

<https://pubmed.ncbi.nlm.nih.gov/10592173/>

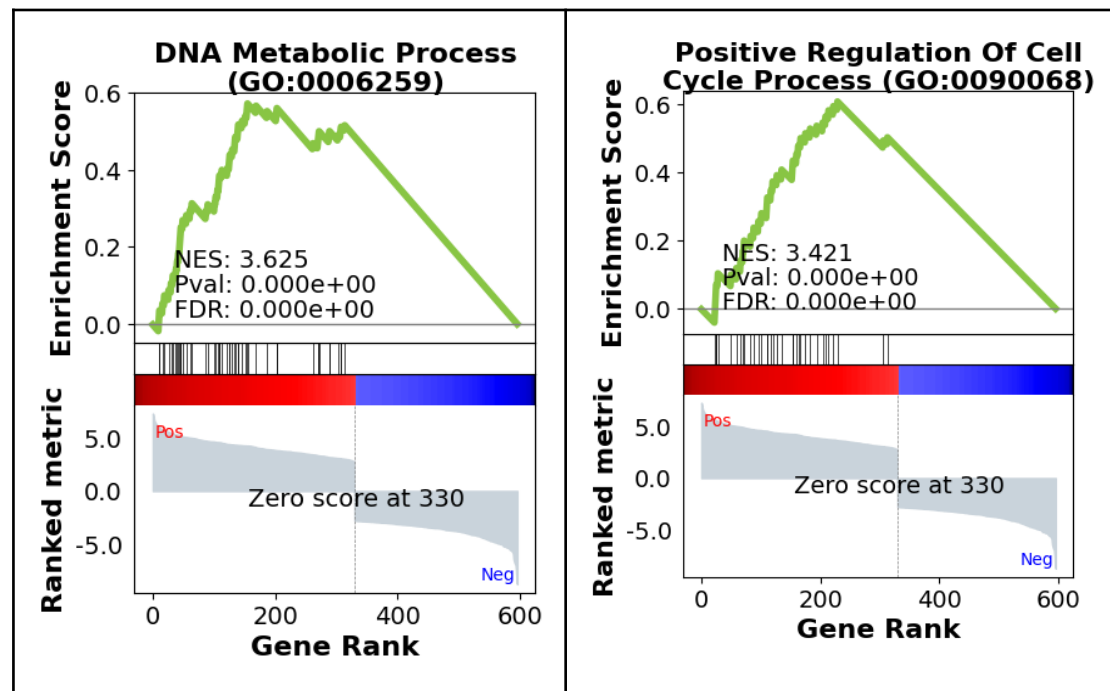
**3.4 (6 points) Gene Ontology.** After researching various libraries, you choose the Gene Ontology database. You learned that GO assigns the following three categories to genes: biological process, molecular function, and cellular component. Briefly explain each category.

Biological Process describes gene function in the context of various operations or sets of molecular events. Cellular Component describes gene products in the context of different cellular locations and extracellular environments. Molecular Function describes biochemical activities of a gene product at the molecular level such as catalytic activity.

**3.5 (15 points) GSEA.** You perform GSEA based on the 'GO\_Biological\_Process\_2023' library. List the top two GO terms with the largest GSEA enrichment scores and their corresponding GSEA plots (which display the running enrichment score for a gene set as the analysis walks down the ranked gene list) here.

The top two GO terms with the largest GSEA enrichment scores are: DNA Metabolic Process (GO:0006259) and Positive Regulation Of Cell Cycle Process (GO:0090068).

GSEA plots:



**3.6 (10 points) Interpretation.** Your PI told you that the objective of this RNA-seq experiment is to compare senescent (aging or aged cells that have lost the power to divide) and proliferative (actively dividing) human cell line

**experiment. There are 4 senescent groups and 4 proliferative groups. Based on this information, can you justify why the top 2 GO terms you identified in Question 3.5 are biologically relevant to this project?**

First GO term: DNA metabolic process

DNA metabolism includes DNA synthesis and degradation reactions involved in DNA replication and repair. When DNA metabolism is not regulated correctly, DNA damage starts to accumulate, triggering DNA damage response (DDR). DNA damage leads to mutations or chromosomal abnormalities ultimately causing genome instability. Cellular senescence is triggered by severely shortened telomeres which activates DDR. DDR impacts intracellular communication and dampened growth signaling. In this way, DNA metabolic processes are biologically relevant to this project.

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9844150/#:~:text=DNA%20damage%20leads%20to%20mutations,damage%20susceptibility%20and%20repair%20access.>

Second GO term: Positive regulation of cell cycle process

Positive regulators of cell cycle processes include two protein groups that allow cells to pass through regulatory checkpoints: cyclins and cyclin-dependent kinases (CDKs). CDK20 plays a positive regulation on cell cycles by activating CDK2. In this way, positive regulation of the cell cycle process is biologically relevant to this project since it directly affects cell proliferation.

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7698114/>

**3.7 (10 points) GO term similarity network** The enrichment results (e.g., from GSEA) may contain a long list of significantly enriched terms which have highly redundant information and are difficult to summarize. The enrichment results can be simplified by clustering the functional terms into groups where the terms in the same group provide similar information. The similarities between terms are important for clustering and can be calculated as Jaccard coefficient or overlap coefficient. Plot a GO term similarity network using the top 15 GO terms from your GSEA results. What are the nodes and edges in the network?

Pathways are shown as circles (nodes) that are connected with lines (edges) if the pathways share many genes. The thickness of the lines (edges) depends on the number of genes they share.

