RNA-Seq Data Analysis

Anika Tahsin

2 December, 2025

1 Introduction

RNA sequencing (RNA-Seq) is widely used to quantify genome-wide gene expression and to detect transcriptional changes between conditions such as disease and normal tissue. A typical workflow includes data loading and preprocessing, differential expression (DE) analysis, visualization of the DE results, and downstream functional or pathway analysis on the resulting gene sets.

This report analyses the public breast cancer RNA-Seq dataset `GSE183947` from the NCBI Gene Expression Omnibus (GEO) [1]. The dataset contains 30 pairs of primary breast cancer and matched adjacent normal tissues, sequenced and normalized to fragments per kilobase per million mapped reads (FPKM). The goal of this assignment is to:

- Load the RNA-Seq data using an R/Python script.
- Perform differential gene expression analysis using an appropriate statistical method.
- Visualize the results with a volcano plot and an MA plot.
- Perform downstream enrichment and functional annotation using tools such as DAVID or similar resources.

2 Materials and Methods

2.1 Dataset

The analysis is based on the GEO series normalized `GSE183947`, which reports expression profiling by high-throughput sequencing for human breast cancer. In total, 30 pairs of normal and cancerous tissues from the same surgical excision were collected from three hospitals in Guangzhou, China. RNA sequencing was performed and the data were normalized to FPKM values. The downloaded expression matrix contains:

- 20,246 rows corresponding to genes.
- 60 columns of expression values: 30 normal samples and 30 tumor samples.

The column names in the matrix follow the pattern `CA.*` for normal samples and `CAP.*` for cancer samples, which is used to separate the two groups programmatically.

2.2 Data Loading and Preprocessing (Task 1)

All analyses were performed in a Python Jupyter notebook (running on Kaggle). The FPKM expression matrix was loaded using the `pandas` library:

- The first column of the CSV file was treated as the gene identifier.
- Columns whose names start with `CA.` were assigned to the normal group.
- Columns whose names start with `CAP.` were assigned to the cancer group.

To stabilise variance and reduce the impact of extreme expression values, the FPKM data were transformed as:

$$\text{log2FPKM} = \log_2(\text{FPKM} + 1).$$

All subsequent analyses used these log-transformed expression values.

2.3 Differential Expression Analysis (Task 2)

Because this dataset is available as FPKM (normalised expression) and not as raw counts, I used a simple but standard approach based on a two-sample t-test rather than count-based models such as DESeq2. For each gene:

1. The $\log_2$ expression values across the 30 cancer samples and the 30 normal samples were extracted.

2. A Welch two-sample t-test (unequal variances) was applied using the `scipy.stats.ttest_ind` function with `equal_var=False`.

3. The $\log_2$ fold change ($\log_2$FC) was computed as:

   $$\log_2\text{FC} = \text{mean(cancer)} - \text{mean(normal)},$$

   so that positive values represent up-regulation in cancer relative to normal tissue.

4. The resulting p-values were adjusted for multiple testing using the Benjamini-Hochberg procedure to obtain the false discovery rate (FDR).

Genes were classified as significantly differentially expressed if they satisfied both:

- $\text{FDR} < 0.05$, and
- $|\log_2\text{FC}| > 1$ (at least two-fold change on the expression scale).

2.4 Visualization: Volcano and MA Plots (Task 3)

Two standard visualisations were produced using `matplotlib`:

- **Volcano plot:** The x-axis shows $\log_2$FC (cancer vs. normal), and the y-axis shows $-\log_{10}(\text{FDR})$. Non-significant genes are shown in one colour, and significant genes ($\text{FDR} < 0.05$, $|\log_2\text{FC}| > 1$) are highlighted in a contrasting soft colour. A dashed line marks the FDR=0.05 threshold, and dotted vertical lines mark $\log_2\text{FC} = \pm 1$.

- **MA plot:** The x-axis shows the mean $\log_2$ expression across cancer and normal samples,

  $$A = \frac{1}{2}\left(\text{mean}_{\text{cancer}} + \text{mean}_{\text{normal}}\right),$$

  and the y-axis shows the $\log_2$FC. A horizontal line at zero indicates no expression change between conditions. Again, significant and non-significant genes are coloured differently.

2.5 Downstream Functional and Pathway Analysis (Task 4)

For downstream interpretation, I performed two complementary analyses:

1. **Enrichment of up-regulated genes using g:Profiler.**
   The list of significantly up-regulated genes ($\text{FDR} < 0.05$ and $\log_2\text{FC} > 1$) was supplied to the g:Profiler `g:GOSt` tool (organism *Homo sapiens*). Enrichment was computed against Gene Ontology Biological Process (GO:BP) and KEGG pathway databases. Results were downloaded as a CSV file for further inspection.

2. **Functional annotation of down-regulated genes using DAVID.**
   The list of significantly down-regulated genes (FDR < 0.05 and $\log_2$FC < -1) was exported as a text file and uploaded to the DAVID Bioinformatics Resources web server as a gene list (official gene symbols, human background). DAVID's "Annotation Summary Results" and functional annotation reports were used to examine:
   - Disease-related annotation databases (e.g. DISGENET, GAD, OMIM).
   - Pathway resources (KEGG, Reactome, WikiPathways, BioCarta, BBID).
   - Functional annotations (Gene Ontology, UniProt Keywords, sequence features).

3 Results

3.1 Differential Expression

The pipeline identified a substantial number of differentially expressed genes (DEGs) between cancer and matched normal tissues:

- Total number of genes tested: 20,246.
- Significantly differentially expressed (FDR < 0.05 and $|\log_2\text{FC}| > 1$): 1,796 genes.
- Up-regulated in cancer: 872 genes.
- Down-regulated in cancer: 924 genes.

Examples of strongly up-regulated genes include *DEFB130*, *LIMS3*, *LCN6*, *CCDC177*, and *KLK9*, with $\log_2$FC values in the range of approximately +3 to +4, corresponding to 8-16 fold increases in expression in tumors relative to normal tissues. Several histone cluster genes and extracellular matrix-related genes, such as *MMP11* and *COL1A1*/*COL1A2*, show strong down-regulation.

3.2 Volcano Plot

Figure 1 shows the volcano plot for all genes. Most points cluster around $\log_2$FC close to zero, reflecting genes with similar expression in both conditions, whereas the significantly differentially expressed genes form the characteristic "wings" at large positive or negative $\log_2$FC values and high $-\log_{10}(\text{FDR})$.

The symmetry of the volcano plot indicates that the number of up- and down-regulated genes is comparable. The right wing corresponds to genes that are overexpressed in cancer, while the left wing captures genes that are more highly expressed in normal tissue.

3.3 MA Plot

Figure 2 shows the MA plot, where each point represents a gene positioned by its mean $\log_2$ expression and $\log_2$FC. The majority of genes lie near the horizontal line at zero, demonstrating similar expression levels in both conditions. Significant DEGs appear as bands above $\log_2\text{FC} = 1$ and below $\log_2\text{FC} = -1$ across a wide range of mean expression values.

3.4 Functional Enrichment of Up-regulated Genes

For Task 4, I performed a downstream enrichment and pathway analysis on the significantly up-regulated genes (FDR < 0.05 and $\log_2$FC > 1).

Method. Functional enrichment was carried out using the g:Profiler `g:GOSt` tool (via the official Python client) with organism *Homo sapiens*. The input gene set consisted of the 872 significantly up-regulated genes. The analysis included:

- Gene Ontology (GO) enrichment: Biological Process (GO:BP).
- Pathway enrichment: KEGG human pathways.

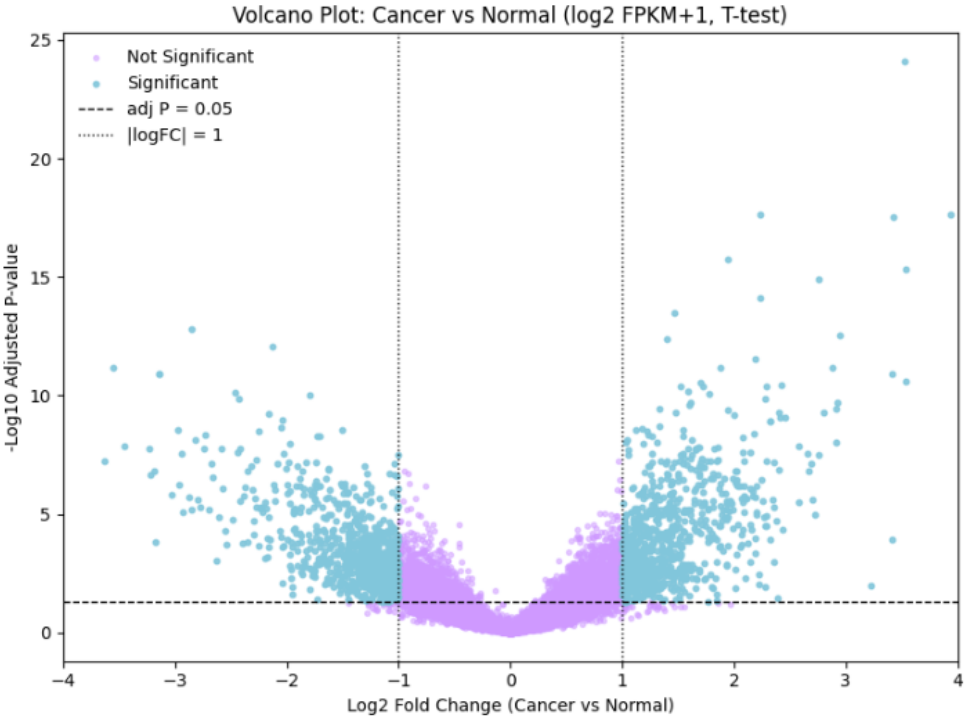


Figure 1: Volcano plot of cancer vs. normal samples. The x-axis shows the $\log_2$ fold change (cancer vs. normal), and the y-axis shows $-\log_{10}(\text{FDR})$. Genes passing the significance thresholds (FDR < 0.05, $|\log_2\text{FC}| > 1$) are highlighted in a contrasting soft colour. Dashed and dotted lines mark the FDR and fold-change thresholds, respectively.

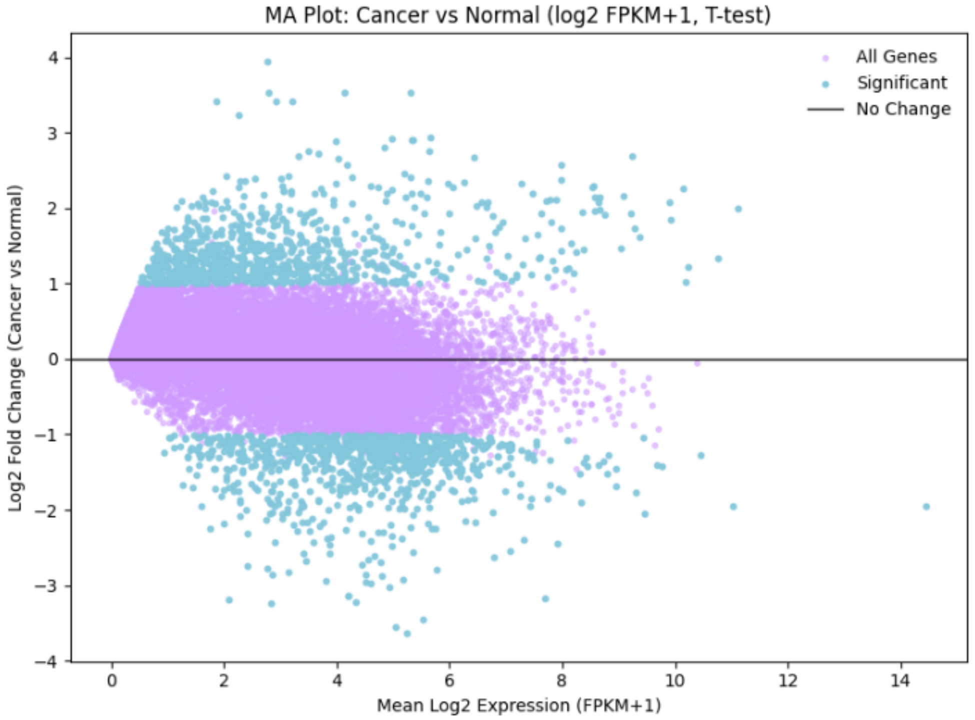


Figure 2: MA plot of cancer vs. normal samples. The x-axis shows the mean $\log_2$ expression (FPKM+1), and the y-axis shows the $\log_2$ fold change (cancer vs. normal). Points in a soft highlight colour correspond to significant differentially expressed genes. The horizontal line marks no change ($\log_2\text{FC} = 0$).

Results. g:Profiler returned in total 380 significantly enriched terms:

- 368 GO:BP terms.
- 12 KEGG pathways.

All reported terms pass g:Profiler's multiple-testing correction (adjusted p-value < 0.05).

The top 10 enriched terms (sorted by adjusted p-value) are:

1. *Animal organ development* (GO:0048513), adj. p $\approx 1.43 \times 10^{-39}$.
2. *Anatomical structure development* (GO:0048856), adj. p $\approx 7.43 \times 10^{-37}$.
3. *Multicellular organismal process* (GO:0032501), adj. p $\approx 6.06 \times 10^{-35}$.
4. *Developmental process* (GO:0032502), adj. p $\approx 2.06 \times 10^{-33}$.
5. *System development* (GO:0048731), adj. p $\approx 1.49 \times 10^{-29}$.
6. *Anatomical structure morphogenesis* (GO:0009653), adj. p $\approx 2.04 \times 10^{-29}$.
7. *Tissue development* (GO:0009888), adj. p $\approx 2.08 \times 10^{-29}$.
8. *Multicellular organism development* (GO:0007275), adj. p $\approx 2.28 \times 10^{-27}$.
9. *Regulation of multicellular organismal process* (GO:0051239), adj. p $\approx 7.17 \times 10^{-26}$.
10. *Cell differentiation* (GO:0030154), adj. p $\approx 4.37 \times 10^{-22}$.

These terms indicate that up-regulated genes in the tumour samples are strongly involved in developmental and morphogenetic programmes, consistent with the notion that cancer cells often re-activate embryonic and organ-development pathways. KEGG pathway results further include signalling routes such as PI3K-Akt and calcium signalling (not shown here in detail), which are well known to be dysregulated in breast cancer.

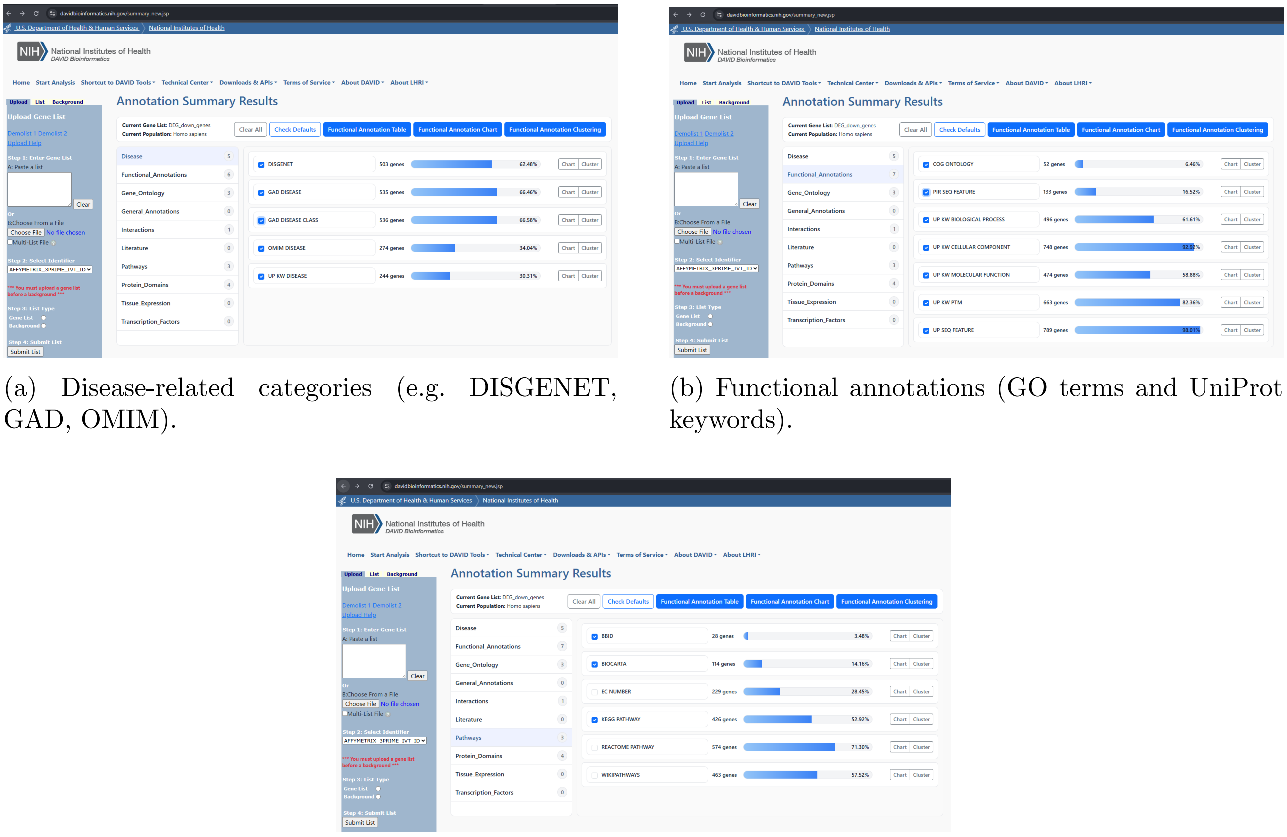
3.5 DAVID Analysis of Down-regulated Genes

The significantly down-regulated genes were submitted to DAVID [2] for further functional annotation. DAVID's Annotation Summary indicated substantial coverage across multiple annotation categories, as illustrated in Figures 3a-3b.

Examination of the detailed DAVID functional annotation tables revealed that many down-regulated genes participate in:

- **Cell cycle and mitosis:** terms such as *cell division*, *mitotic cell cycle*, *DNA replication initiation*, and *mitotic cytokinesis*, as well as the KEGG *Cell cycle* pathway.
- **Immune and infection-related pathways:** e.g. KEGG terms for *phagosome*, *antigen processing and presentation*, and infection-related pathways (Herpes simplex virus 1, Epstein-Barr virus, HTLV-1).
- **Extracellular matrix and tissue organisation:** terms such as *collagen fibril organisation* and *extracellular matrix organisation*.

These annotations suggest that down-regulated genes are enriched for core cell-cycle machinery, immune-presentation processes, and extracellular matrix components, which together influence tumour proliferation, immune recognition, and tissue architecture.



Annotation Summary Results

Current Gene List: DEG_down_genes
Current Population: Homo sapiens

Disease	5
Functional_Annotations	6
Gene_Ontology	3
General_Annotations	0
Interactions	1
Literature	0
Pathways	3
Protein_Domains	4
Tissue_Expression	0
Transcription_Factors	0

☑	DISGENET	503 genes	62.48%
☑	GAD DISEASE	535 genes	66.46%
☑	GAD DISEASE CLASS	536 genes	66.58%
☑	OMIM DISEASE	274 genes	34.04%
☑	UP KW DISEASE	244 genes	30.31%

(a) Disease-related categories (e.g. DISGENET, GAD, OMIM).

Annotation Summary Results

Current Gene List: DEG_down_genes
Current Population: Homo sapiens

Disease	5
Functional_Annotations	7
Gene_Ontology	3
General_Annotations	0
Interactions	1
Literature	0
Pathways	3
Protein_Domains	4
Tissue_Expression	0
Transcription_Factors	0

☑	COG ONTOLOGY	52 genes	6.46%
☑	PIR SEQ FEATURE	133 genes	16.52%
☑	UP KW BIOLOGICAL PROCESS	496 genes	61.61%
☑	UP KW CELLULAR COMPONENT	748 genes	92.92%
☑	UP KW MOLECULAR FUNCTION	474 genes	58.88%
☑	UP KW PTM	663 genes	82.36%
☑	UP SEQ FEATURE	789 genes	98.01%

(b) Functional annotations (GO terms and UniProt keywords).

Annotation Summary Results

Current Gene List: DEG_down_genes
Current Population: Homo sapiens

Disease	5
Functional_Annotations	7
Gene_Ontology	3
General_Annotations	0
Interactions	1
Literature	0
Pathways	3
Protein_Domains	4
Tissue_Expression	0
Transcription_Factors	0

☑	BBID	28 genes	3.48%
☑	BIOCARTA	114 genes	14.16%
☐	EC NUMBER	229 genes	28.45%
☑	KEGG PATHWAY	426 genes	52.92%
☐	REACTOME PATHWAY	574 genes	71.30%
☐	WIKIPATHWAYS	463 genes	57.52%

(c) Pathway resources (KEGG, Reactome, WikiPathways, BioCarta, BBID).

Figure 3: DAVID Annotation Summary for the significantly down-regulated gene set.

4 Discussion and Conclusion

This assignment demonstrates a complete RNA-Seq analysis workflow on the breast cancer dataset GSE183947, from data loading through differential expression and visualisation to downstream functional interpretation.

Using $\log_2$(FPKM+1) expression values and Welch's t-test, I identified 1,796 genes that are differentially expressed between cancer and matched normal breast tissue at FDR < 0.05 and at least two-fold change. The volcano and MA plots provide intuitive summaries of the magnitude and significance of these changes across the transcriptome.

Enrichment analysis of up-regulated genes using g:Profiler highlights developmental, morphogenetic, and cancer-associated signalling pathways, including PI3K-Akt and calcium signalling. DAVID analysis of down-regulated genes further reveals strong associations with cell-cycle regulation, immune processes, and extracellular matrix organisation. Together, these results are consistent with known hallmarks of breast cancer and show that the identified DEGs reflect biologically coherent programmes rather than random variation.

Overall, the analysis includes data loading, differential expression analysis, visualisation with volcano and MA plots, and downstream functional enrichment using DAVID and related tools.

Code Availability

Code file and the analysis reports are available in the following GitHub repository:

`https://github.com/Anika-Tahsin-S/Bioinformatics-RNA-Seq-Data-Analysis.git`

References

[1] NCBI. GEO Accession viewer.

[2] NIH. DAVID Functional Annotation Bioinformatics Microarray Analysis — davidbioinformatics.nih.gov.