BMI8540 Course Project

Title: Identification of Differentially Expressed Genes in Breast Cancer Using RNA-Sequencing Analysis

Abstract/Purpose: The project analyzes the RNA sequencing data from breast cancer patients to determine which genes are differently expressed between tumor subtypes and normal breast tissue. The automated pipeline is designed as an easy-to-use computational analysis tool to help researchers identify patterns in breast cancer genetics more quickly, such as overexpressed or under-expressed genes, promote the discovery of possible therapeutic targets, and improve current understanding of breast cancer development. Comparing the significant genes in at least two conditions showed 13420 significant genes in triple negative breast cancer versus normal, 11569 significant genes in non-triple negative breast cancer versus normal, and 11596 significant genes for HER-2 positive versus normal. Each plot of the cancer subtypes shows the top fifty genes with upregulated and downregulated genes. In triple negative breast cancer versus the normal, the upregulated genes identified were RN7SL2, SCARNA7, H4C5, while the downregulated genes were IL6, TIPARP, PCBP1, UBC, ZFP36, ZC3H12A, and NXF1. The analysis shows other visualization plots of the different breast cancer subtypes and comparisons for downstream analysis and interpretation of findings.

Project Objectives/Goals:

1. To develop a reproducible bioinformatics pipeline for analyzing RNA-seq data from breast cancer for differentially expressed genes.
2. To integrate a user-friendly database that stores gene expression data and analysis results for efficient data retrieval.
3. To generate visualizations using Python for the interpretation of findings.

Background:

Breast cancer (BC) is among the most common cancers affecting women worldwide, and early detection and treatment significantly impact the survival rate of breast cancer patients [1]. The incidence of breast cancer is rising in South America, Africa, and Asia, probably because of changes in lifestyle and the introduction of screening programmes. Mortality from breast cancer in these regions is also still increasing, partly because of a lack of access to state-of-the-art diagnosis and therapy [1]. The process of gene expression involves the transmission of genetic information from DNA templates into functional proteins [2]. Several advancements have been made in the identification of genetic alterations that occur in cancer. The progression of Ductal carcinoma in situ (DCIS) has been linked to DCIS gene mutations, which include BRCA1/2 deleterious mutation, AKTI, TP53, PIK3CA somatic mutations, SOX, and HOXA5 gene hypermethylation [3]. Neoadjuvant therapy is now a standard treatment for both triple-negative and HER2-positive early breast cancer, and its backbone depends on clinical tumor subtype and includes endocrine therapy, anti-HER2 targeting, and chemotherapy [4]. There is a high recurrence rate of triple negative breast cancer after treatment [5] , while the prognosis for HER-2 enriched tumors remains poor unless treated with targeted therapy, because these tumors are more aggressive [6]. RNA-Sequencing is an approach that measures quantitative information on gene expression, and differentially expressed genes are analyzed using basic bioinformatic tools to determine whether genes in the genome are up- or down-regulated in normal or abnormal conditions [2]. The use of these tools enables the analysis, visualization, and understanding of the differences in genes that are over- or under-expressed as compared to normal tissue in breast cancer, which helps us determine its origin and spread. The pathogenesis of BC needs to be understood while new biomarkers are discovered to aid in prognosis, diagnosis, and treatment [7].

Breast cancer, as one of the leading global cancer types, due to its molecular processes, shows extensive variation between individual patients. BC shows different treatment responses among its various subtypes, and the systematic evaluation of tumor subtypes and normal sample gene expression patterns enables researchers to detect dysregulated genes that could function as diagnostic biomarkers or prognostic indicators, or therapeutic targets. The pipeline, Breast Cancer Gene Expression Analysis (BRCA-GEA) RNA-seq database, was designed due to the

necessity to understand the molecular processes that drive breast cancer development and progression for analysis and interpretation. Current analysis pipelines based on computational methods fail to incorporate structured data storage, which results in fragmented research efforts and makes it difficult to compare results between studies. Also, based on my research, there are no known standardized reproducible workflows that are specifically designed for breast cancer research. Due to this limitation, the BRCA-GEA pipeline functions as a unified system that integrates data processing with differential gene expression analysis and visualization functions, together with database storage capabilities to meet the needs of breast cancer expression studies.

Project Components:

The project components are

1. A working database implementation
2. A reproducible bash shell script
3. A Python script for visualization

Table 1. Files description for the project

Parameters	**Description**
get_data.sh	A bash script used to retrieve raw data for analysis from EBI. The RNA-Seq raw counts data contains 19 samples (3 tumor subtypes and normal) with 40,527 known genes after filtering out the unknowns
run_deseq_analysis.py	A Python script for differential gene expression analysis.
generate_insert_statements.py	Script for generating INSERT statements from raw and analysed results data.
setup_database.py	Script to create tables and insert data from the .sql files.

visualize_results.py	Python script to visualize the differential expression results using different plots.
brca_deseq_pipeline.sh	A bash script to run the pipeline from data acquisition, database population, and visualization of results.

Documentation:

All codes were written with descriptive comments and are reproducible. The README page of GitHub gives a full description of the project, its usage, and license. The link to the data used for the project is also linked in the GitHub page for easy access.

Users:

1. Graduate Students – Both for bioinformatics and non-bioinformatics graduate students to perform differential gene expression analysis.
2. Cancer Researchers – To analyze differential gene expression with BRCA data with a quick, reproducible framework.
3. Instructors/Faculty Members – To demonstrate to students/learners a reproducible pipeline for differential gene expression analysis using RNA-seq data.

Implementation constraints:

- My original project design was to implement the project using FASTQ files but was limited by the available space resources on Odin, so I resorted to using already processed raw counts data.
- To run the Python implementation of DESeq2, I require version 3.8 and above, and the working environment on Odin offers version 3.7. I couldn't reinstall another version, so I used the base Conda environment on Odin, which I found to have a Python 3.8 version, and then installed pydeseq2 for my analysis.

- Another limitation was accessing MySQL client from the base environment, I resolved this by installing PyMySQL, which helped me to implement the database part of the project.

Privacy:

The database is private, therefore, users should use their database to implement the whole pipeline.

References:

[1] N. Harbeck and M. Gnant, "Breast cancer," *The Lancet*, vol. 389, no. 10074, pp. 1134–1150, Mar. 2017, doi: 10.1016/S0140-6736(16)31891-8.

[2] M. Griffith, J. R. Walker, N. C. Spies, B. J. Ainscough, and O. L. Griffith, "Informatics for RNA Sequencing: A Web Resource for Analysis on the Cloud," *PLOS Comput. Biol.*, vol. 11, no. 8, p. e1004393, Aug. 2015, doi: 10.1371/journal.pcbi.1004393.

[3] C. Zhu, H. Hu, J. Li, J. Wang, K. Wang, and J. Sun, "Identification of key differentially expressed genes and gene mutations in breast ductal carcinoma in situ using RNA-seq analysis," *World J. Surg. Oncol.*, vol. 18, no. 1, p. 52, Mar. 2020, doi: 10.1186/s12957-020-01820-z.

[4] N. Harbeck and M. Gnant, "Breast cancer," *Lancet Lond. Engl.*, vol. 389, no. 10074, pp. 1134–1150, Mar. 2017, doi: 10.1016/S0140-6736(16)31891-8.

[5] "Triple Negative Breast Cancer | | General Surgery Dublin | Breast Surgery Dublin." Accessed: Apr. 28, 2025. [Online]. Available: https://www.breastsurgeryireland.com/triple-negative-breast-cancer-general-reconstructive-aesthetic-breast-surgery-dublin.html

[6] G. Menon, F. M. Alkabban, and T. Ferguson, "Breast Cancer," in *StatPearls*, Treasure Island (FL): StatPearls Publishing, 2025. Accessed: Apr. 29, 2025. [Online]. Available: http://www.ncbi.nlm.nih.gov/books/NBK482286/

[7] L. Fang, Q. Liu, H. Cui, Y. Zheng, and C. Wu, "Bioinformatics Analysis Highlight Differentially Expressed CCNB1 and PLK1 Genes as Potential Anti-Breast Cancer Drug Targets and Prognostic Markers," *Genes*, vol. 13, no. 4, p. 654, Apr. 2022, doi: 10.3390/genes13040654.

[8] "Step-by-step PyDESeq2 workflow — PyDESeq2 0.5.0 documentation." Accessed: Apr. 26, 2025. [Online]. Available: https://pydeseq2.readthedocs.io/en/stable/auto_examples/plot_step_by_step.html#sphx-glr-auto-examples-plot-step-by-step-py

[9] “Python 3 Script for Generating SQL Insert Statements from CSV Data,” SQLServerCentral. Accessed: Apr. 26, 2025. [Online]. Available: https://www.sqlservercentral.com/scripts/python-3-script-for-generating-sql-insert-statements-from-csv-data

[10] “Experiment < Expression Atlas < EMBL-EBI.” Accessed: May 03, 2025. [Online]. Available: https://www.ebi.ac.uk/gxa/experiments/E-GEOD-52194/Downloads

[11] “Differential expression in Python with pyDESeq2 - YouTube.” Accessed: Apr. 29, 2025. [Online]. Available: https://www.youtube.com/watch?v=wIvxFEMQVwg&t=851s

[12] bioinfokit: Bioinformatics data analysis and visualization toolkit. Python. Accessed: May 03, 2025. [MacOS, Microsoft :: Windows :: Windows 10, Unix]. Available: https://github.com/reneshbedre/bioinfokit

[13] “📙 Emojipedia — 😃 Home of Emoji Meanings 💁👌🛕😍,” Emojipedia. Accessed: May 1, 2025. [Online]. Available: https://emojipedia.org