Bioinformatics Summer School Report

Name: Aisha Soliman Alsomiry
Internship Organization: University of Toronto
Supervisor's Name: Dr. Hagar Labouta
Project title: Multi-Omics Data Analysis and Visualization to Investigate the Molecular Profile of Breast Cancer Cells in an Organ-on-a-Chip Model

1. Introduction

Key Projects and Objectives:

During my summer internship, I focused on analyzing RNA bulk data obtained from wet lab experiments conducted by the lab (PhD student). My primary objective was to process and analyze RNA-seq data to identify differentially expressed genes (DEGs) and understand their biological significance.

Specific Tasks and Outcomes:

RNA-Seq Data Analysis: I used R to identify differentially expressed genes in breast cancer cells compared to controls. I also experimented with different criteria, and fold change (FC) thresholds to ensure the results were meaningful and biologically relevant.

Gene Annotation: Using the Biomart R tool, I converted Ensembl IDs into gene symbols, allowing for more intuitive data interpretation.

Visualization: I created a volcano plot using R's ggplot2 library to visualize the distribution of DEGs, highlighting significant genes with biological relevance.

Pathway Enrichment Analysis: I conducted pathway enrichment analysis using Database for Annotation, Visualization, and Integrated Discovery (DAVID).

2. Materials and Methodology

I used different libraries in R; ggplot2, Biomart, and hgnc_symbol database. I used the Database for Annotation, Visualization and Integrated Discovery (DAVID) to categorize these DEGs into biological processes, molecular functions, and cellular components. Concurrently, I use KEGG pathway analysis to map the DEGs to known signaling pathways.

3. Results and Discussion

In this study, we identified 3,532 genes that were differentially expressed in breast cancer cells compared to the control. Functional enrichment analysis using DAVID highlighted several key pathways that were notably enriched among the up-regulated genes, including those related to the exon-exon junction complex and DNA replication factor A complex. Its upregulation may help cancer cells prevent apoptosis by stabilizing mRNAs that promote oncogenic gene expression.

Additionally, the upregulation of the DNA replication factor A complex plays a crucial role in the initiation and elongation phases of DNA replication, and its dysregulation can lead to uncontrolled cell proliferation, a hallmark of cancer.

4. Skills and Knowledge Gained

I acquired several technical skills that have strengthened my foundation in bioinformatics and data analysis. I gained hands-on experience with RNA-seq data processing using tools like DESeq2, where we applied statistical methods to identify differentially expressed genes (DEGs). I also developed my coding skills using R, utilizing packages such as biomaRt for gene annotation and ggplot2 for data visualization. Additionally, I performed pathway enrichment analysis using DAVID, enhancing my ability to interpret large-scale biological data.

Regarding soft skills, I significantly developed my communication and teamwork abilities through collaboration with lab members and mentors. Regular discussions about data interpretation and troubleshooting improved my problem-solving skills and presenting my findings in regular lab meetings refined my ability to share complex scientific concepts. Moreover, I had the opportunity to be part of the discussion of various projects, including those in biomedical engineering, with other lab members, which broadened my interdisciplinary understanding. These experiences have prepared me to contribute effectively to interdisciplinary research teams and further my understanding of bioinformatics.

5. Challenges and Solutions

one of the most significant challenges I faced was mastering the integration of multi-omics data, which involved processing and analyzing large datasets from various sources. The complexity of aligning data from different omics layers required both technical precision and a deep understanding of bioinformatics tools.

To overcome this challenge, I focused on strengthening my coding skills in R and improving my data analysis techniques. I invested time in learning advanced statistical methods and data visualization techniques, which allowed me to interpret the data more effectively. Additionally, I sought guidance from my mentors, Yasmin Abdelgadeer and Dr. Hagar Labouta, who provided valuable insights and encouraged me to think critically about my approach. Through persistence, continuous learning, and collaboration, I was able to successfully overcome the challenges and produce meaningful results that contributed to the overall objectives of my project.

6. Conclusion

My experience working on this integrated omics data analysis pipeline has been transformative in shaping my career aspirations. By applying multi-omics technologies to breast cancer research, I gained a deep understanding of the complex molecular mechanisms driving disease progression, which reinforced my passion for bioinformatics and precision medicine. This project

has solidified my decision to pursue a PhD in this field, where I aim to further explore the potential of integrative approaches to uncover novel therapeutic targets.

Throughout the project, I developed several critical skills. I mastered my coding abilities in R, allowing me to efficiently analyze large-scale biological datasets. Additionally, I mastered data visualization techniques, which were essential for interpreting and presenting complex results. The interdisciplinary nature of the project also enhanced my ability to integrate diverse data types, such as transcriptomics and proteomics, to draw meaningful biological insights. Beyond the technical skills, I developed strong problem-solving abilities, learned to manage and analyze vast amounts of data, and improved my communication skills through collaborative teamwork.

7. Acknowledgments

I would like to express my sincere gratitude to the KAUST Academy for their financial support, which made this internship possible. I am also deeply thankful to Yasmin Abdelkader, PhD student, and Dr. Hagar Labouta, principal investigator, for their invaluable mentorship and guidance throughout my research. Additionally, I would like to thank the Data Science program at the University of Toronto for their professional development sessions, which greatly enhanced my skills and knowledge. Most importantly, I want to extend my heartfelt thanks to my family for believing in me and providing endless support. Their encouragement has been the foundation of my journey.

8. References (if applicable)

List any references or sources cited in the report.

1- Christopher Hogue. An automated method for finding molecular complexes in large protein interaction networks. 2003

2- Rossi C, Cicalini I. Breast cancer in the era of integrating "Omics" approaches. Oncogenesis. 2022

3- Aurelie Tomczak, Jonathan M. Interpretation of biological experiments changes with evolution of the Gene Ontology and its annotations. Nature. 2018