Cairo University

Faculty of Engineering

Systems and Biomedical Department

First Semester - 2024/2025

Cancer Genomics Unplugged:
Investigating Gene Expressions in Depth

Team Members

Yassmin Sayed

Raghda Tarek Neiazy

Under The Supervision Of: Dr. Ibrahim Mohamed

Introduction

Cancer remains a pervasive global health challenge, necessitating innovative approaches for comprehensive understanding and targeted interventions. Advances in molecular biology have unveiled intricate genetic landscapes underlying various cancer types, providing critical insights for personalized therapeutic strategies. Among these, the integration of copy number alterations (CNAs) and gene expressions (GEs) stands as a powerful approach to deciphering the multifaceted genetic architecture of malignancies

CNAs, encompassing chromosomal amplifications and deletions, play a pivotal role in driving oncogenesis by influencing gene dosage and expression levels. Concurrently, GE profiles reflect the dynamic activity of genes within cancerous and normal tissues. The convergence of these two molecular dimensions offers a holistic perspective on the molecular mechanisms governing cancer development and progression.

This research endeavor focuses on two prominent cancer types: Lung Squamous Cell Carcinoma (LUSC) and Kidney Renal Clear Cell Carcinoma (KIRC). By examining these distinct malignancies, we aim to elucidate the interplay between CNAs and GEs, shedding light on the key genetic determinants of tumor behavior and response to treatment.

Medical Background

Genomes and Molecular Structures

Genomes

The genome is the complete set of genetic material within an organism. It encompasses all the DNA, both coding and non-coding, which serves as the blueprint for the structure, function, and regulation of cells. In humans, the genome is composed of approximately 3 billion base pairs of DNA distributed across 23 pairs of chromosomes. These chromosomes, found within the nucleus of each cell, carry the genes necessary for various biological processes.

Molecular Structures

At the molecular level, genetic information is stored in the form of DNA (deoxyribonucleic acid). DNA is a double-stranded helix composed of nucleotides. Each nucleotide consists of a sugar (deoxyribose), a phosphate group, and a nitrogenous base. The four nitrogenous bases - adenine (A), thymine (T), cytosine (C), and guanine (G) - form complementary base pairs (A-T, C-G), stabilizing the DNA structure. This elegant arrangement allows for the faithful replication of DNA during cell division and serves as the basis for genetic inheritance.

Genes

Genes are segments of DNA that contain the instructions for producing specific proteins. Proteins are the workhorses of cellular function, carrying out a myriad of tasks ranging from structural support to enzymatic catalysis. The process of gene expression involves transcription, where a complementary RNA molecule is synthesized from a gene's DNA template, and translation, where the RNA molecule guides the assembly of a corresponding protein. Regulation of gene expression is finely tuned, ensuring that genes are activated or repressed at the right time and in the right cell type.

Understanding the genetic makeup of an organism is foundational to comprehending its biology and health. Genetic variations, such as those occurring in CNAs, can have profound implications for an individual's susceptibility to diseases, including cancer. By investigating the interplay between genes, their regulatory elements, and the broader genomic landscape, researchers can uncover critical insights that pave the way for advancements in diagnosis, treatment, and personalized medicine.

Gene Expressions (GEs)

Gene expression is a dynamic process where genetic information is transcribed and translated into functional proteins. It is finely regulated and varies across different cell types and under different conditions. Dysregulated gene expression is a hallmark of cancer. Oncogenes, which promote cell growth, maybe excessively expressed, while tumor suppressor genes, which inhibit proliferation, may be downregulated. Investigating gene expression patterns provides insights into the underlying molecular mechanisms that drive cancerous behavior.

Lung Squamous Cell Carcinoma (LUSC)

Lung Squamous Cell Carcinoma (LUSC) arises in the epithelial cells lining the airways. These cells are typically flat and scale-like. Understanding the genetic alterations specific to LUSC is crucial, as it helps identify potential targets for therapy and sheds light on the factors contributing to its aggressive nature.

Objectives

In this context, this study leverages state-of-the-art bioinformatics techniques, harnessing the power of Python and scientific packages such as NumPy, SciPy, and Scikit-learn. Statistical methodologies, including hypothesis testing to dissect the intricate GEs.

The ensuing sections of this paper delve into the methodological framework, delineating the steps undertaken to process and analyze the multi-platform molecular data. The research outcomes not only hold implications for advancing our understanding of cancer genomics but also bear potential for the development of targeted therapies tailored to individual patients.

Methods

Data Acquisition and Preprocessing

Obtaining accurate and reliable multi-platform molecular data forms the bedrock of our investigation into Lung Squamous Cell Carcinoma (LUSC). The data, comprising Gene Expressions (GE) was meticulously collected to ensure comprehensive coverage.

Data Acquisition

- GE Data: We sourced gene expression profiles from paired samples of cancerous and healthy tissues for LUSC. These profiles, stored in tab-separated files, provide quantified measures of gene activity across the genome.

Data Preprocessing

In the process of data preparation, ensuring the quality and integrity of the dataset is of paramount importance. Our approach involves meticulously scrutinizing each dataset to identify and rectify any missing or erroneous data points. This is critical because missing data can introduce biases and inaccuracies into downstream analyses.

Identification of Missing Values

Our code iterates through the dataset, examining each data point to identify any instances where information is missing.

Missing values are identified based on criteria defined in the code (e.g., zero values).

Handling Missing Values

For genes with more than or equal to 50% zeros, they are filtered out from further analysis. This is done to ensure that genes with a significant portion of missing data do not unduly influence the results.

Maintaining Data Integrity

By carefully addressing missing values, you are maintaining the integrity of the dataset, which is crucial for producing reliable and accurate analytical results.

Differential Expression Analysis

Differential Expression Analysis (DEA) serves as a pivotal step in discerning the genes that play a significant role in cancer progression. Given the paired nature of our samples, a specialized approach was employed.

Paired Sample Analysis

Considering that the samples were paired (i.e., each cancerous sample corresponds to a specific healthy tissue sample), we implemented a hypothesis testing approach tailored to this design. This ensures that the analysis correctly accounts for the dependency between paired samples. The goal is to identify Differentially Expressed Genes (DEGs) using a robust statistical approach of hypothesis testing.

Independent Sample Analysis

In the alternative scenario where samples are independent (i.e., no specific pairing between cancerous and healthy tissue samples), a different approach is necessary. We will conduct hypothesis testing designed for independent samples. This involves assessing the normality of both the normal and cancerous datasets for lung cancer.

Hypothesis Testing

Now, let's delve into the heart of our analysis - identifying the Differentially Expressed Genes (DEGs). We do this by employing hypothesis testing, which is a robust statistical approach.

Choice of Statistical Test

For this crucial step, we need to accurately identify the DEGs. To achieve this, we turn to hypothesis testing methods. Specifically, we make use of the Shapiro-Wilk normality test. This test serves a vital role in our analysis. It helps us determine whether the gene expression data adheres to a normal distribution. This assumption of normality is pivotal, especially for certain parametric tests.

Paired Sample Analysis

Considering the unique paired nature of our samples - where each cancerous sample is intricately linked to a specific healthy tissue sample - a specialized approach is warranted. This tailored approach ensures that our analysis correctly addresses the dependency between these paired samples.
In our case, we apply the Wilcoxon signed-rank test. This non-parametric test excels at comparing the distribution of differences between paired samples. It's especially adept in situations where the data may not conform to a normal distribution.

Independent Sample Analysis

Recognizing the distinct independence of our samples, where no specific pairing exists between cancerous and healthy tissue samples, we require a specialized approach tailored for independent datasets. This ensures that our analysis appropriately addresses the absence of pairing between samples.

In this scenario, we opt for the Wilcoxon rank-sum test. This non-parametric test is well-suited for comparing the distribution of values between independent samples. Its strength lies in its ability to handle situations where the data may not adhere to a normal distribution, providing a robust alternative for our independent sample analysis.

Significance Assessment

Our commitment to a comprehensive understanding extends beyond mere identification of differences in gene expression between cancerous and healthy tissues. We take an extra step to rigorously evaluate the significance of these observed disparities through meticulous calculations of test statistics and p-values.

To ensure the reliability of our findings, we tailor our approach based on the distribution characteristics of the data. When dealing with datasets exhibiting a normal distribution, the Shapiro-Wilk test assists in confirming the normality assumption. Conversely, for non-normally distributed data, such as in independent sample analyses, we turn to the Wilcoxon rank-sum test.

These statistical metrics play a crucial role in quantifying the likelihood of observed differences occurring purely by chance, providing a robust assessment of their true significance in both paired and independent sample analyses. In our paired sample analysis, the Wilcoxon signed-rank test is applied, excelling at comparing the distribution of differences between paired samples, especially in situations where the data may not conform to a normal distribution. For independent sample analyses, the Wilcoxon rank-sum test serves a similar purpose, ensuring the assessment of significance is adapted to the unique characteristics of the data.

Fold Change

In the realm of bioinformatics, the concept of fold change serves as a pivotal metric, shedding light on the magnitude of differences in gene expression levels between experimental conditions. As researchers, our pursuit goes beyond mere identification of gene differentials; we seek a nuanced understanding of the scale of these variations, and this is precisely where fold change plays a crucial role.

Fold change, in essence, quantifies the ratio of gene expression levels between two conditions – often the comparison of cancerous and healthy tissues in our study. It provides a clear and interpretable measure of the magnitude of change, allowing us to discern not just whether a gene is differentially expressed, but the extent to which it is so.

The utilization of fold change is strategic, particularly in the identification of Differentially Expressed Genes (DEGs). When combined with statistical tests like the Wilcoxon signed-rank test for paired samples or the Wilcoxon rank-sum test for independent samples, fold change empowers us to distinguish genes with not only statistically significant differences but also biologically meaningful variations.

The addition of epsilon (ε) to both normal and cancerous expression levels before calculating the log2 fold change is critical to prevent infinite or undefined values in the log2 transformation.

By integrating fold change into our analysis, we gain insights into the biological relevance of gene expression alterations. Small but consistent changes across a multitude of genes may collectively contribute to significant physiological impacts. Fold change, therefore, acts as a valuable ally in unraveling the intricacies of molecular dynamics, ultimately enhancing the depth and precision of our bioinformatics investigations.

Volcano Plots

Volcano plots, a fundamental tool in bioinformatics, offer a comprehensive portrayal of gene expression patterns by plotting statistical significance against fold change. Each point on the graph represents an individual gene, with extreme values resembling the peaks of a volcano indicating genes of particular interest.

These plots play a crucial role in the identification of Differentially Expressed Genes (DEGs) by simultaneously considering statistical significance and the magnitude of biological changes. Genes located at the far ends of the plot signify substantial fold changes with high statistical reliability. This dual consideration allows researchers to focus on genes that not only display significant alterations but also possess considerable biological impact.

The significance of volcano plots lies in their ability to guide targeted investigations into the molecular underpinnings of specific conditions. By pinpointing genes with both statistical and biological significance, researchers can efficiently prioritize candidates for further exploration. This prioritization ensures a more streamlined and insightful approach to understanding the molecular intricacies that drive changes in gene expression.

In essence, volcano plots serve as navigational tools, steering researchers towards the most relevant and impactful genes within complex datasets. Their role extends beyond mere visualization, actively aiding in the prioritization and interpretation of DEGs, contributing to a deeper understanding of the biological processes under investigation.

Software Packages Utilized

To implement our analyses and computations effectively, we employed various software packages, each selected for its specific capabilities:

Python: As a versatile and widely used programming language, Python formed the backbone of our data processing and analysis. Its extensive libraries and tools were invaluable in handling multi-platform molecular data.

Pandas: This powerful data manipulation library in Python facilitated the manipulation and analysis of structured data. It provided essential functionality for organizing, cleaning, and transforming the molecular datasets.

NumPy: NumPy was instrumental in performing array operations and mathematical computations. It underpinned many aspects of data handling and calculations within our code.

SciPy: Leveraging the scientific functions and algorithms offered by SciPy, we conducted advanced scientific and technical computations. This library played a crucial role in various statistical tests and analyses.

Matplotlib and Seaborn: These visualization libraries were employed to create clear and informative plots and figures, aiding in the visualization of our results and enhancing their interpretability.

Tabulate: Tabulate facilitated the generation of formatted tables, allowing us to present data in a structured and visually appealing manner. This was especially useful for showcasing specific data points and results.

These software packages were chosen for their robustness, extensive functionality, and suitability for the specific tasks they were employed for. They collectively formed a powerful toolkit that enabled us to conduct a comprehensive analysis of the multi-platform molecular data.