Fall 12-13-2023

Classifying Different Cancer Types Based on Transcriptomics Data Using Machine Learning Algorithms

Eunice Olorunshola

Classifying Different Cancer Types Based on Transcriptomics Data Using Machine Learning Algorithms

by

Eunice Olorunshola

Under the Direction of Committee Chair's Murray Patterson, PhD

A Thesis submitted in Partial Fulfillment of the Requirements for the Degree of

Master of Science

in the College of Arts and Sciences

Georgia State University

2023

ABSTRACT

Cancer, a complex group of diseases characterized by abnormal cell growth, presents a significant global health challenge. Accurate classification of cancer types is vital for effective treatment and improved patient outcomes. This master's thesis addresses the crucial issue associated with accurate cancer classification. It analyzes transcriptomic data of RNA sequencing, from six cancer subtypes (breast, colorectal, glioblastoma, hepatobiliary, lung, pancreatic) and a healthy control group. This research utilizes several machine learning algorithms to construct accurate cancer classification models using gene expression profiles and gene count data. The study incorporates advanced techniques such as feature selection, data preprocessing, and model optimization. The primary objective is to enhance our understanding of transcriptomic signatures distinguishing one cancer type from another, with potential applications in early diagnosis, treatment selection, and biomarker discovery. Through the power of machine learning, this research contributes to advancing effective cancer classification and management strategies in this ongoing battle.

INDEX WORDS: Classification Transcriptomic, Machine Learning, Gene Expression Profiles, Subtypes, RNA-Sequencing

Classifying Different Cancer Types Based on Transcriptomics Data Using Machine Learning Algorithms

by

Eunice Olorunshola

Committee Chair: Murray Patterson

Committee: Alex Zelikovsky

Jonathan Shihao Ji

Electronic Version Approved:

Office of Graduate Services

College of Arts and Sciences

Georgia State University

December 2023

DEDICATION

To my mom and dad, siblings, and friends that supported me throughout this journey.

ACKNOWLEDGEMENTS

I would like to express my heartfelt gratitude to the following individuals and organizations who have played a pivotal role in the completion of this thesis: I am deeply thankful to my advisor Dr. Murray Patterson for his unwavering support, guidance, and mentorship throughout the entire research process his expertise were valuable in shaping this work. I also want to extend my appreciation to the members of my thesis committee, Dr. Alex Zelikovksy and Dr. Jonathan Shihao Ji for supporting this thesis and providing feedback. My deepest gratitude goes to my family for their prayers, love, and support. I also want to thank my best friends Aicha, Horace and my friends for their advice, encouragement and love and support throughout this journey. I am grateful to my fellow classmates and former colleagues who shared their knowledge, ideas, to help with the research experience. I also appreciate the resources provided by Georgia State University and Computer Science department that facilitated the research process. This journey has been a tremendous learning experience filled with challenges and growth. I am also thankful for this opportunity to take part on this academic endeavor and for the lessons it has imparted. In closing, I want to express my deepest appreciation to everyone mentioned above. Your love, support, belief in my work, advice, and encouragement have been an inspiration and I am truly grateful.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 INTRODUCTION

Cancer, characterized by the uncontrolled growth of abnormal cells, remains a formidable global health challenge. Based on statistics from the WHO, every year, more than 8.2 million people die from cancer, accounting for approximately 13 percent of deaths worldwide, indicating that cancer is one of the most threatening diseases in the world [1].

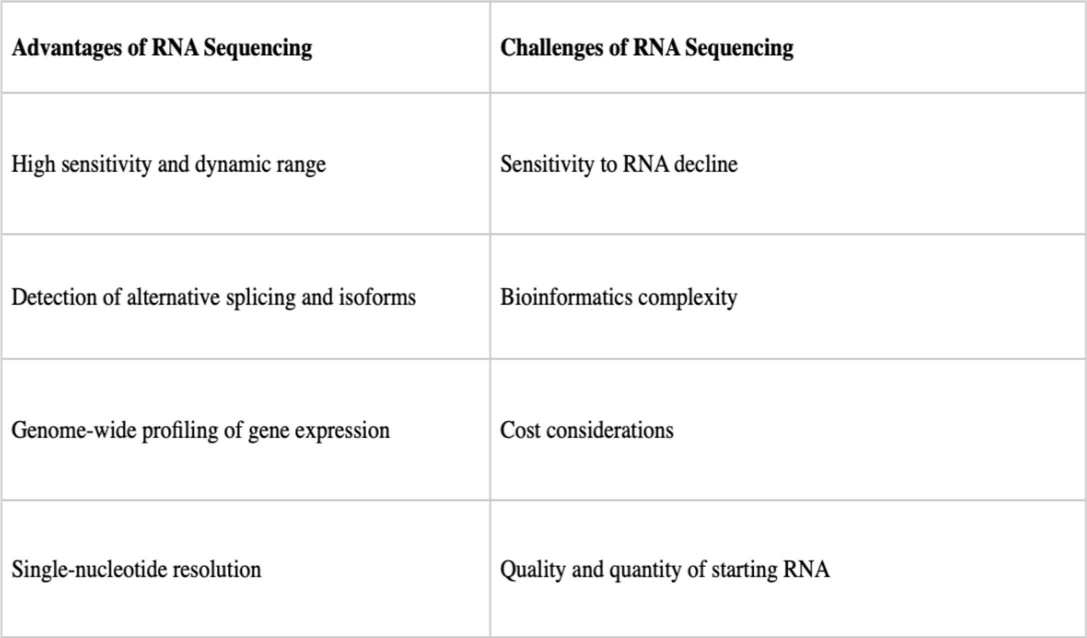
Recent years have witnessed a revolutionary convergence of machine learning algorithms and gene expression data, accompanying in profound insights and innovations in the field of biology and healthcare. These studies have marked a significant paradigm shift in how we understand and utilize gene expression profiles, shaping the landscape of recent research in genomics. The most important function of transcriptome profiling is to determine the differentially expressed genes occurring in a body or detect variations in genes at different levels [2]. However, analyses of RNA gene expression data are quite complex because of their high dimensions, complexity, and the existence of duplications in feature values [3]. Therefore, a need for automatic feature extractions exists, which may be addressed through machine learning algorithms [4]. Machine learning is a branch of artificial intelligence which is used to identify associations among data by finding underlying patterns using experience and learning [6]. Machine learning models have helped in differentiating cancer subtypes based on gene expression patterns, facilitating more accurate diagnosis and personalized treatment strategies. Furthermore, they have accelerated the discovery of noble biomarkers associated with any diseases, including cancer, enhancing early disease detection. Machine learning has also helped with the process of drug discovery by predicting the effectiveness of drug compounds based on their influence on gene expression patterns.

This master’s thesis undertakes the critical issue of accurate cancer classification, building on the foundation of recent studies involving machine learning algorithms and gene expression data. It centers its analysis on transcriptomic data obtained through RNA sequencing, from six cancer subtypes: breast, colorectal, glioblastoma, hepatobiliary, lung, and pancreatic cancer, and a health control group. By using gene expression profiles and gene count data, the research utilizes several arrays of machine learning algorithms to construct highly accurate cancer classification models. The study's methodology uses advanced techniques, including feature selection, data preprocessing, and model optimization which helps with increasing the accuracy of the classification models. The implications of this work extend to early cancer diagnosis, personalized treatment selection and the discovery of potential biomarkers, thereby contributing to the advancement of effective cancer classification and management strategies. Through the power of machine learning, this study embodies hope in the ongoing battle against cancer that continues to be a major impact on global health.

1.1 RNA Sequencing: An Introduction

RNA Sequencing (RNA-Seq) is a new and popular technique that is used to detect new isoforms and transcripts by providing more normalized and less noisy data for prediction and classification purposes [4,5]. It has changed our understanding of gene expression, and helped in elucidating the complexities of gene regulation, uncovering novel insights into the molecular mechanisms of diseases, and allowing the discovery of potential therapeutic targets. Compared to previous Sanger sequencing and microarray-based methods, RNA-Seq provides far higher coverage and greater resolution of the dynamic nature of the transcriptome [5]. The principle of RNA sequencing involves the conversion of RNA molecules into a library of cDNA fragments, which are then sequenced using high throughput sequencing platforms. By mapping the resulting sequence reads to a reference genome or transcriptome, researchers can quantify gene expression levels, identify alternative splicing events, detect novel transcripts, and conduct post transcriptional modifications. In addition to polyadenylated messenger RNA (mRNA) transcripts, RNA-Seq can be applied to investigate different populations of RNA, including total RNA, pre-MRNA, and noncoding RNA, such as microRNA and long ncRNA [5]. Table 1.1 shows a comprehensive overview highlighting the strengths and considerations associated with RNA Sequencing.

Table 1.1 Comparison of Advantages and Challenges in RNA Sequencing



Advantages of RNA Sequencing	**Challenges of RNA Sequencing**
High sensitivity and dynamic range	Sensitivity to RNA decline
Detection of alternative splicing and isoforms	Bioinformatics complexity
Genome-wide profiling of gene expression	Cost considerations
Single-nucleotide resolution	Quality and quantity of starting RNA

1.1.1 Importance and Role in Genomics

Genomic data, such as RNA-Seq have become widely available due to the popularity of high throughput sequencing technology [6]. As an important part of next generation sequencing, RNA sequencing has made great contributions in various fields, especially cancer research, including studies on differential gene expression analysis and cancer biomarkers, cancer heterogeneity and evolution, cancer drug resistance, the cancer microenvironment and immunotherapy, neoantigens, etc [7]. In genomics research, RNA-Seq helps in deciphering how genes are activated under multiple conditions in different cell types. It also helps in showing the roles of non-coding RNA molecules, revealing disease expression patterns, and capturing the dynamics of gene expression over time. By providing a deeper comprehensive view of the transcriptome, RNA-Seq allow researchers to explore the functional elements of the genome, identify novel transcripts, and explore the impact of genetic variations on gene expression. As RNA-Seq techniques continues to advance and providing new insights and innovations it also helps genomics grow by facilitating the identification of disease biomarkers, therapeutic targets, and a deeper understanding of the molecular mechanisms supporting biological processes.

1.1.2 Transcriptomics and Understanding RNA

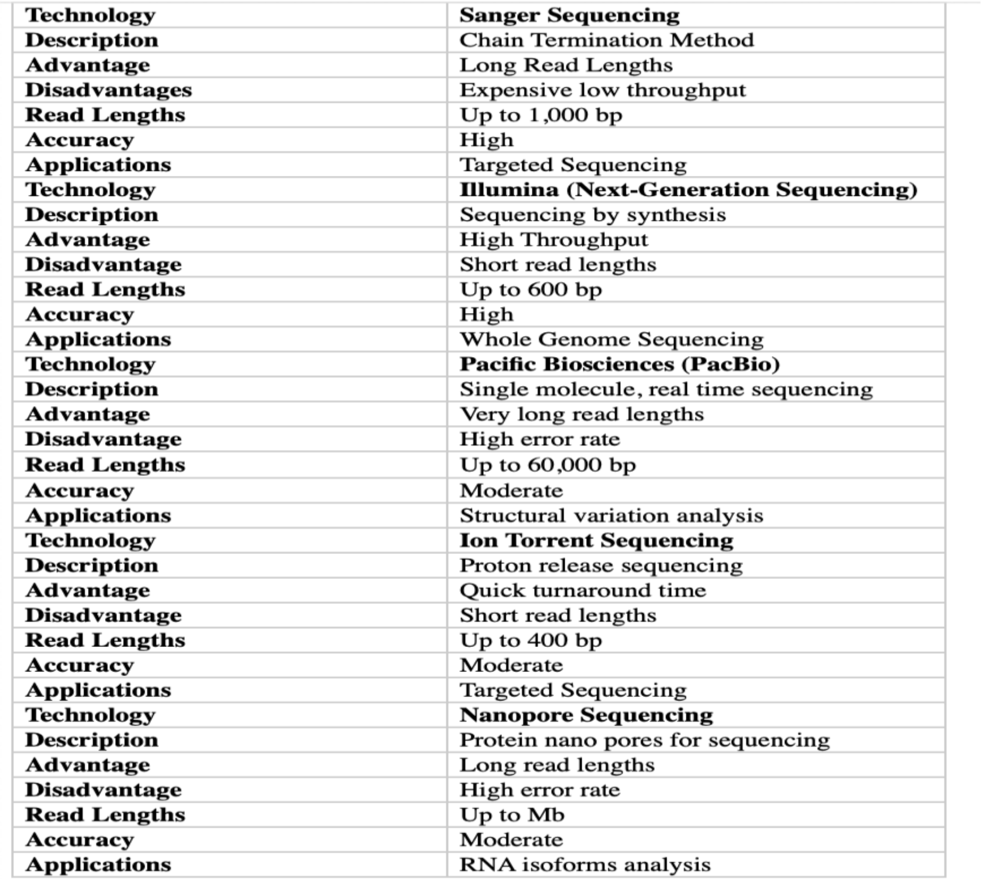
Transcriptomics is a branch of molecular biology and genomics that focuses on the study of RNA transcripts in a cell. It involves the analysis of the complete set of RNA molecules, produced in a specific cell or tissue. Understanding RNA, a fundamental molecule, is the focus of transcriptomics research. RNA molecules are essential for interpreting the functional elements of the genome and understanding development and disease [8]. The transcriptome has a high degree of complexity and encompasses multiple types of coding and noncoding RNA species. Messenger RNA (mRNA) molecules were the most frequently studied RNA species because they encoded proteins via the genetic code. In addition to protein coding mRNA, there is a diverse group of noncoding RNA (ncRNA) molecules that are functional [8]. Previously most known ncRNAs fulfilled basic cellular functions such as ribosomal RNAs and transfer RNAs involved in mRNA translation, small nuclear RNA (snRNAs) involved in splicing and small nucleolar RNAs (snoRNAs) involved in the modification of rRNAs [9]. The first transcriptomics studies were performed using hybridization-based microarray technologies, which provide a high throughput option at relatively low cost [10]. Transcriptomics helps identify alternative splicing events, post transcriptional modifications and non-coding RNAs (microRNAs and long non-coding RNAs) helps in gene regulation.

By researchers studying the transcriptome, they can gain a deeper understanding of the molecular mechanisms allowing several biological processes, development stages, disease states, and responses to environmental changes. RNA sequencing (RNA-Seq), microarray analysis, and quantitative polymerase chain reaction(qPCR) are part of transcriptomics techniques and are used to explore gene expression and transcriptomic profiles. Understanding the several functions and roles of RNA is important in deciphering complex biological processes, revealing the mechanisms of diseases and contributing to the advancement of genetic research.

1.2 RNA Sequencing Technologies

RNA sequencing technologies have allowed scientists to gain a deeper knowledge of gene expression and transcriptomics with precedented accuracy. Next-generation platforms such as llumina and Ion Torrent have been helpful in this process. These platforms generate vast amounts of sequence data, providing a comprehensive view of the transcriptome. In principle, any high-throughput sequencing technology can be used for RNA-Seq [10]. It has also been helpful in revealing the role of non-coding RNA molecules such as microRNAs and long non-coding RNAs, in gene regulation. Single-cell RNA sequencing, an extension of this technology has an added a new dimension by allowing the study of individual cells, revealing cellular heterogeneity, and aiding in the understanding of complex biological processes. Table 1.2 gives a better overview of each RNA Sequencing Technologies and a comparison of each of them with key information.

Table 1.2 Comparison of RNA Sequencing Technologies



Technology	**Sanger Sequencing**
Description	Chain Termination Method
Advantage	Long Read Lengths
Disadvantages	Expensive low throughput
Read Lengths	Up to 1,000 bp
Accuracy	High
Applications	Targeted Sequencing
Technology	**Illumina (Next-Generation Sequencing)**
Description	Sequencing by synthesis
Advantage	High Throughput
Disadvantage	Short read lengths
Read Lengths	Up to 600 bp
Accuracy	High
Applications	Whole Genome Sequencing
Technology	**Pacific Biosciences (PacBio)**
Description	Single molecule, real time sequencing
Advantage	Very long read lengths
Disadvantage	High error rate
Read Lengths	Up to 60,000 bp
Accuracy	Moderate
Applications	Structural variation analysis
Technology	**Ion Torrent Sequencing**
Description	Proton release sequencing
Advantage	Quick turnaround time
Disadvantage	Short read lengths
Read Lengths	Up to 400 bp
Accuracy	Moderate
Applications	Targeted Sequencing
Technology	**Nanopore Sequencing**
Description	Protein nano pores for sequencing
Advantage	Long read lengths
Disadvantage	High error rate
Read Lengths	Up to Mb
Accuracy	Moderate
Applications	RNA isoforms analysis

1.2.1 Next-Generation Sequencing (NGS) Platforms

The development of high-throughput next-generation sequencing (NGS) has revolutionized transcriptomics by enabling RNA analysis through the sequencing of complementary DNA (cDNA) [11]. This technological development eliminated many challenges posed by hybridization-based microarrays and Sanger sequencing-based approaches that were previously used for measuring gene expression. It has also transformed our ability to decode and understand the genetic information in DNA and RNA. A typical RNA-Seq experiment consists of isolating, RNA, converting it to complementary DNA (cDNA), preparing the sequencing library, and sequencing it on NGS platform. [11]. NGS platforms such as Illumina, Ion Torrent, and Oxford Nanopore, have several methodologies but share the same common feature of generating massive amounts of sequence data. Ilumina sequencing is known for its accuracy, and it is a popular technology used in RNA-Seq it relies on the generation of clusters of DNA fragments and sequences these clusters. Ion Torrent sequencing technology is valued for its speed and scalability making it possible to target sequencing. It directly measures pH changes as nucleotides are used during DNA synthesis. Oxford Nanopore sequencing gives the advantage of long read sequencing, help with the study of complex genomic regions, and real time sequencing. It also can pass single DNA molecules through nanopores and read the sequence as they pass.

1.2.1.1 Advantages and Challenges

Although RNA-Seq is still a technology under active development, it offers several key advantages over existing technologies. First unlike hybridization-based approaches, RNA-Seq is not limited to detecting transcripts that correspond to existing genome sequence [11]. RNA-Seq can reveal the precise location of transcription boundaries to a single-base resolution [11]. Furthermore, 30-bp short reads from RNA-Seq give information about how two exons are connected, whereas longer reads or pair-end short reads should reveal connectivity between multiple exons [11]. It is also not restricted to known genes, making it acceptable for the discovery of novel and unannotated transcripts. A second advantage of RNA-Seq relative to DNA microarrays is that RNA-Seq has very low, if any, background signal because DNA sequences can be unambiguously mapped to unique regions of the genome [11]. While RNA-Seq provides many advantages it also has some challenges. The cost of RNA-Seq experiments can be higher than technologies like microarrays that have been used in studies for many years. Another challenge is the need for high quality RNA, since contamination can lead to biased results. Although there are only a few steps in RNA-Seq, it does involve several manipulation steps during the production of cDNA libraries, which can complicate it use in profiling all types of transcripts [11].

1.3 Aim and Objectives

The fundamental aim of this research is to gain a better comprehension of the unique transcriptomic signatures defining six specific cancer subtypes, including breast, colorectal, glioblastoma, hepatobiliary, lung, and pancreatic cancers, along with a healthy control group, through the analysis of RNA-Sequencing datasets. The main objective is to assess the classification accuracy of machine learning algorithms in distinguishing these cancer subtypes and the healthy control group based on their gene expression patterns. This study focuses on evaluating the performance of various machine learning models in accurately classifying the specified cancer subtypes and healthy controls. The overall goal is to contribute to a deeper understanding of the molecular foundation of different cancers, enhancing the potential for precise and efficient cancer subtype classification.

2 CONCEPTS AND BACKGROUND

2.1 Differential Gene Expression

Differential gene expression analysis is one of the most common tools of RNA sequencing [12]. Samples from different backgrounds (different species, tissues and periods) can be used for RNA sequencing to identify differentially expressed genes, revealing their function and potential molecular mechanisms [13]. More importantly differential gene expression analysis facilitates the discovery of potential cancer biomarkers [14]. It involves comparing RNA transcripts between two or more experimental groups such as healthy vs. diseased tissues. To detect differential expression, a variety of statistical methods have been designed specifically for RNA-Seq data. A popular tool to detect differential expression is Cuffdiff, which is part of the Tuxedo suite of tools (Bowtie, Tophat, and Cufflinks) developed to analyze RNA-Seq data [15]. Increasing differentially expressed genes are being identified by RNA sequencing and new potential cancer biomarkers are being continuously discovered [16]. By analyzing the sequence data, differentially expressed genes are discovered, making the way of molecular mechanisms underlying various biological processes. This information allows for a better understanding on the genetic basis of diseases, identifying potential biomarkers, and revealing novel therapeutic targets.

2.1.1 Gene Expression Data

Gene expression data provides insights into the dynamic activity of genes in a biological sample. The data from gene expression is generated through advanced techniques like RNA sequencing and microarray analysis, allowing researchers to quantify and compare gene expression levels across different samples. Gene expression analysis is the process of identifying the number of transcripts present in a particular cell or tissue type to estimate the level of expressed genes. Gene expression data quantifies the level of transcripts produced by genes, offering insights into which genes are active or not. Differential expression analysis also plays a role in gene expression data and a brief overview of differential expression analysis is discussed above. Gene expression data also helps in revealing alternative splicing patterns, where a single gene can generate multiple mRNA isoforms. There have been many advances in gene expression analysis but the most recent one allows single cell RNA sequencing, providing an approach to explore gene expression at the individual cell level. It also provides a global view of the transcriptome, from protein coding genes, non-coding RNAs and other RNA species. EST libraries represent short fragments of mRNA obtained from a single sequencing procedure carried out from cDNA libraries [17].

2.1.1.1 Gene Expression Datasets

Gene expression datasets represents a complete collection of data showing the activity of genes in several biological samples. Some of the popular gene expression datasets includes The Cancer Genome Atlas (TCGA), Gene Expression Omnibus (GEO), ArrayExpess, and The Human Protein Atlas. These datasets are generated through techniques like RNA-Seq and microarray analysis. Each dataset contains measurement of gene expression levels, that allows researchers to explore how genes responds to specific stimuli and how they are different between healthy and diseased tissues. The datasets are open-source and easily accessible [18].

2.1.2 Microarray Data

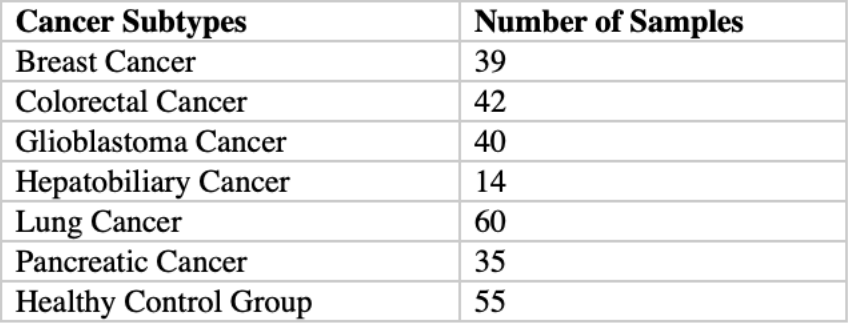
Microarray data is a valuable resource in genomics and molecular biology, providing a comprehensive view of gene expression patterns on a genome wide scale. Microarray technology allows researchers to measure the expression levels of large amounts of genes in a biological sample. These datasets are generated by hybridizing labeled RNA samples to an array of gene specific probes, which can reveal which genes are under expressed or over expressed in various tissues. Microarray data has been helpful in revealing insights into gene regulation, identifying biomarker for diseases, and understanding the molecular mechanisms behind several biological processes. Since microarray data is an older technology and next generation sequencing platforms has gained prominence in recent years. Microarray data remains a valuable archive, particularly for historical gene expression studies, and it continues to contribute to our understanding of genomics.

3 MATERIALS AND METHODS

3.1 Datasets

The dataset includes gene expression profiles of blood from 285 samples of patients who had one of the following cancer subtypes: breast cancer, colorectal cancer, glioblastoma, hepatobiliary cancer, lung cancer, pancreatic cancer and healthy controls. And is accessed from the Gene Expression Omnibus database specifically from the work of Zhang et al., (2017) [19]. And it is important to note that the feature included in all cancer samples is the gene identifiers. Brief details about these cancer types are discussed below. The detailed number of samples in each cancer subtype and healthy control group samples are listed in Table 3.1.

Table 3.1 Dataset Composition



Cancer Subtypes	Number of Samples
Breast Cancer	39
Colorectal Cancer	42
Glioblastoma Cancer	40
Hepatobiliary Cancer	14
Lung Cancer	60
Pancreatic Cancer	35
Healthy Control Group	55

3.1.1 Breast Cancer

Breast cancer is a complex and various disease with distinct subtypes that are set apart by the molecular and genetic features of cancer cells. Luminal A tumors are typically hormone receptor-positive and HER2-negative, known for their slow growth and favorable prognosis. Luminal B breast cancers, also hormone receptor-positive, exhibit higher proliferation markers and a slightly worse outlook. HER2-enriched cancers over express the HER2 gene, demanding HER2-targeted therapies, while triple-negative (basal-like) cancers, lacking key receptors, pose challenges due to limited targeted treatment options. Inflammatory breast cancer, an aggressive subtype, presents with redness and swelling, requiring a multi-pronged approach. Metaplastic breast cancer is rare and complex, while normal-like tumors mirror Luminal A. The gene expression data for breast cancer was sourced from the Gene Expression Omnibus (GEO) under accession number GSE68086. The dataset originally comprised 39 samples; however, it is important to note that none of the samples were successfully downloaded due to issues such as file corruption or download failures.

3.1.2 Colorectal Cancer

Colorectal cancer is the second- and third-most common cancer in women and men [20]. The subtypes of colorectal cancers are categorized into microsatellite stable (MSS) and microsatellite instability-high with MSI-H tumors having better prognosis and responds better to immunotherapy. We utilized gene expression data obtained from the Gene Expression Omnibus (GEO) accession number GSE68086. Comprising 42 samples, the gene expression profiles of colorectal cancer blood tissues are important to our research.

3.1.3 Glioblastoma Cancer

Glioblastoma is the most common primary malignant brain tumor, comprising 16 percent of all primary brain and central nervous system neoplasms [21]. Glioblastoma present at a median age of 64 years but can occur at any age, including childhood [22]. The most common subtype is the classical glioblastomas, marked by EGFR augmentation and chromosome 10 deletion. Proneural glioblastomas, relate to PDGFRA alterations, revealing distinct molecular profile. Mesenchymal glioblastomas are represented by NF1 mutations that are characterized by aggressive growth and a specific immune signature. The datasets were obtained from GSE68086 comprising 40 samples, these samples reflect the gene expression patterns of glioblastoma blood cancer tissues.

3.1.4 Hepatobiliary Cancer

Hepatobiliary cancers are highly lethal. In 2008, approximately 21,370 persons in the United States were estimated to be diagnosed with liver cancer and 9520 with gallbladder cancer. Furthermore, approximately 18,410 deaths from liver cancer and 3340 deaths from gallbladder cancer were estimated to occur [23]. The subtype Hepatocellular carcinoma (HCC), is the most common hepatobiliary cancer that happens in the hepatocytes of the liver and is often related to chronic live diseases such as hepatitis B or C. Intrahepatic cholangiocarcinoma, happens in the bile ducts in the liver, and requires different treatment therapies. Extrahepatic cholangiocarcinoma happens outside of the liver and causes difficulties in treatments such as surgery, radiation, and chemotherapy. Gallbladder cancer is not as common as the other subtypes but can have more promising results through surgical removal of the gallbladder. The datasets are obtained from GSE68086 and consists of 14 samples.

3.1.5 Lung Cancer

Lung cancer remains the leading cause of cancer mortality in men and women in the U.S. and worldwide. About 90 percent of lung cancer cases are caused by smoking and the use of tobacco products. However, other factors such as radon gas, asbestos, air pollution exposures, and chronic infections can contribute to lung carcinogenesis [24]. During 2014, an estimated 224,210 new cases and 159,260 deaths for lung cancer were predicted in the USA [25]. Lung cancer is categorized into non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC) with its own subtypes. Non-small cell lung cancer (NSCLC) subtypes are adenocarcinoma, squamous cell carcinoma, and large cell carcinoma. Adenocarcinoma is the most common subtypes that appears in the lung outer regions, while squamous cell carcinoma happens in the bronchial lining, and large cell carcinoma is the most aggressive subtype. Small cell lung cancer spreads quickly and metastasize earlier then non-small cell lung cancer. Non-small cell lung cancer treatment strategies include surgery, radiation, and therapies while small cell lung cancer treatments include chemotherapy. The datasets obtained from GSE68086 and comprises 60 samples.

3.1.6 Pancreatic Cancer

Pancreatic ductal adenocarcinoma is a relatively uncommon cancer, with approximately 60430 new diagnoses expected in 2021 in the US. The incidence of PDAC is increasing by 0.5 percent to 1.0 percent per year, and it is projected to become the second leading cause of cancer-related mortality by 20230 [26]. Among lifestyle risk factors, current cigarette smoking has the strongest association with PDAC [26]. The median age at diagnosis is the US is 17 years, and PDAC is slightly more common in men than in women (5.5 vs 4.0 per 100000 individuals) [27]. Pancreatic ductal adenocarcinoma (PDAC) is the most common subtype and is known for how quick it becomes aggressive marked by mutations in genes like KRAS and TP53. While Pancreatic neuroendocrine tumors (PNETs) are not as aggressive and better results. The datasets were obtained from GSE68086 and comprises 35 samples.

4 ANALYSIS OF RNA SEQUENCING DATA

Computational analysis tools for RNA sequencing have dramatically increased during the past decade [28]. The choice of a particular tool should be based on the purpose and accuracy of application [29,30,31]. The conventional pipeline for RNA-Seq data includes generating FASTQ-format files contains reads sequenced from an NGS platform, aligning these reads to an annotated reference genome, and quantifying expression of genes. [32]. Although basic sequencing analysis tools are more accessible than ever, RNA-Seq analysis presents unique computational challenges not encountered in other sequencing-based analyses and requires specific consideration to the biases inherent in expression data [32]. A general RNA sequencing data analysis process involves the quality control of raw data, read alignment and transcript assembly, expression quantification and differential expression analysis [32]. (Fig.4.1) gives a better overview of the steps and tools that are used in this process.

4.1 Steps in Analyzing RNA Sequencing Data

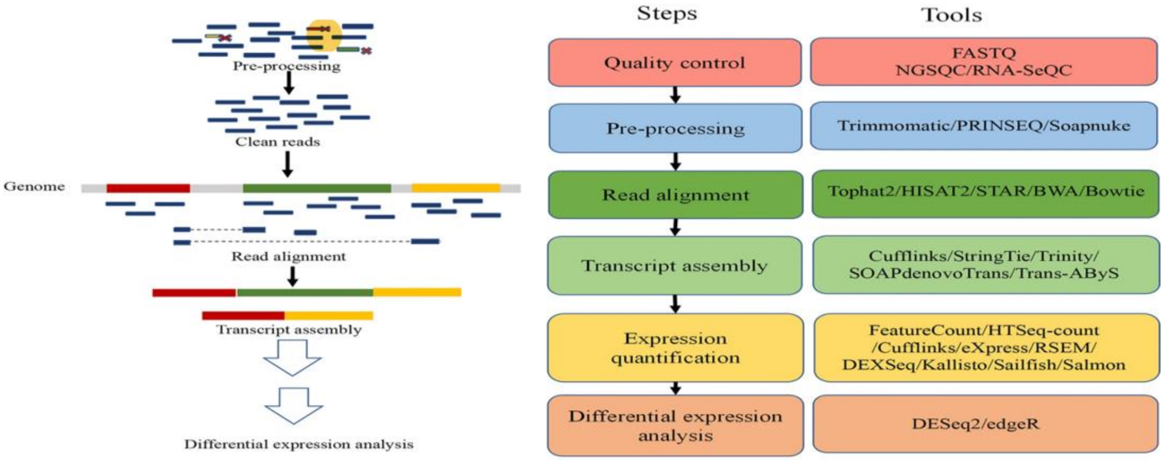

Pre-processing
Clean reads
Genome
Read alignment
Transcript assembly
Differential expression analysis
Steps
Tools
Quality control
FASTQ
NGSQC/RNA-SeQC
Pre-processing
Trimmomatic/PRINSEQ/Soapnuke
Read alignment
Tophat2/HISAT2/STAR/BWA/Bowtie
Transcript assembly
Cufflinks/StringTie/Trinity/
SOAPdenovoTrans/Trans-AByS
Expression quantification
FeatureCount/HTSeq-count
/Cufflinks/eXpress/RSEM/
DEXSeq/Kallisto/Sailfish/Salmon
Differential expression analysis
DESeq2/edgeR

Figure 4.1 RNA Sequencing Data Analysis Process (Adapted from Hong, Mingye et al., 2020 [28])

4.1.1 Raw Data Assessment

Analyzing RNA sequencing data involves multiple step process such as data quality control, preprocessing, read alignment, transcript assembly, and expression quantification that presents the transcriptomic landscape of biological samples. The first journey begins with data quality control, data control begins with an assessment of the raw RNA sequencing data. Assessing the raw data is a fundamental step in the analysis pipeline, making sure that the initial data is of high quality and acceptable for downstream processing. During this assessment stage, factors such as sequence read quality, base call accuracy, and the presence of any contaminants that can disrupt the process are evaluated. Quality scores are examined, potential adapter sequences are identified, and the distribution of read lengths are checked [33]. By following this assessment and addressing the issues in the raw data, researchers can enhance the reliability of subsequent analyses, ultimately leading to more accurate results.

4.1.2 Quality Control Procedures

The preprocessing of RNA Sequencing data includes critical steps to ensure the reliability and accuracy of downstream analyses. Following raw data assessment, the removal of low-quality reads becomes imperative, targeting sequencing errors and adapter contaminations. This careful curation improves the overall dataset quality, enhancing the precision of subsequent alignment and quantification processes. Additionally, handling sequence duplications is addressed, acknowledging their origin in library preparation and sequencing. The identification and removal of duplicates contribute to the mitigation of biases that could impact quantitative measurements and differential expression analyses. These steps collectively form a robust foundation for obtaining meaningful insights from RNA sequencing data.

4.2 Preprocessing Steps for Read Quality

The quality of raw sequencing data significantly impacts downstream analyses, and therefore preprocessing is very important. This section includes crucial steps such as trimming adapter sequences, quality-based trimming, and length filtering. To visually depict this process, Figure 4.2 visualizes the trimming adapter sequences which also includes length filtering highlighting the removal of adapter sequences and length filtering from RNA sequencing reads to enhance accuracy in downstream analysis and improve overall data quality.

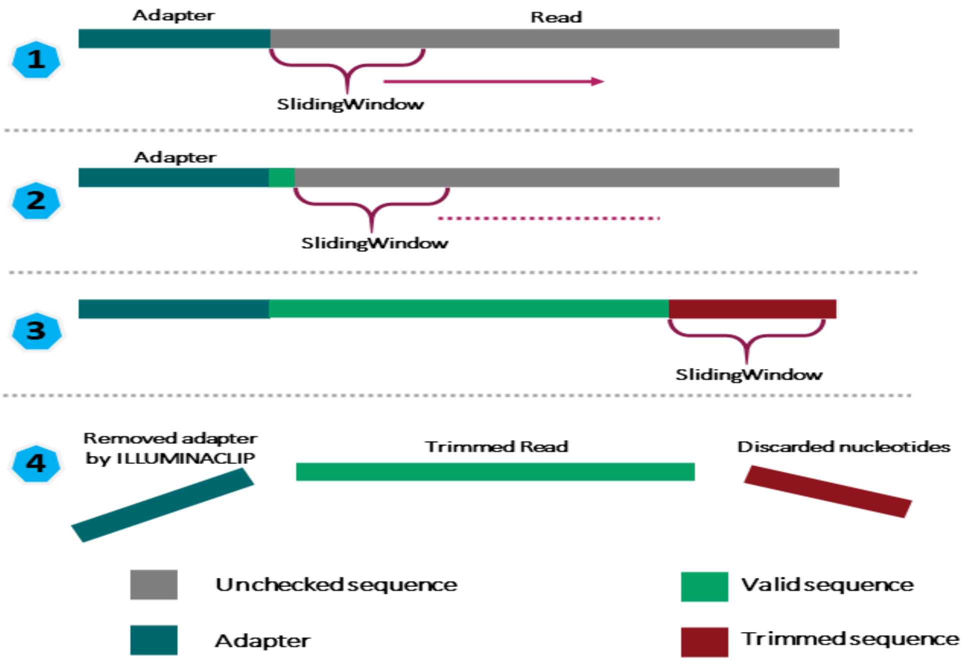


Figure 4.2 Trimming Adapter Sequences and Length Filtering "Title of the Webpage." Trimming and Filtering- Data Processing and Visualization for Metagenomics

4.2.1 Trimming Adapter Sequences

Trimming adapter sequences is helpful in the preprocessing of RNA sequencing data, mainly when dealing with short reads that are generated by high-throughput sequencing platforms. During the process of the library preparation, adapter sequences are applied to the ends of the DNA fragments to facilitate binding to the sequencing flow cell. But these adapters must be removed from the sequencing data to make sure of accurate analysis. Due to this the process of trimming is needed, which involves the identification and removal of these sequences from the reads. Failure to do this step can result in contamination that can disrupt the read alignments, quantification, and downstream analyses. By doing this process correctly and removing the adapter sequences, researchers can improve the accuracy of the data, enhance the accuracy of mapping to the reference genome.

4.2.2 Quality-Based Trimming

Quality-based trimming helps with enhancing the accuracy and reliability of the downstream analyses. The process involves each base in a sequence read is assigned to a quality score, signaling the confidence level of that base accuracy. Low quality bases often indicate unreliable readings, can compromise the accuracy of alignment and gene expression quantification. Quality-based trimming algorithms automatically remove bases with low quality scores, by eliminating unreliable segments but while retaining high-confidence portions of the reads. By applying quality-based trimming it can significantly improve the overall data quality and give me a more accurate insight.

4.2.3 Length Filtering

Length Filtering makes sure that only reads of specific length are retained for downstream analysis in the preprocessing step of RNA sequencing data. During this process, unusual short or long reads that can introduce bias into the data are removed. Short reads can lead to not having enough information for accurate alignment and quantification, while overly long reads can contain sequencing errors which can also lead to inaccurate alignment.

4.3 Read Alignment to Reference Genome

In the process of read alignment to the reference genome, the STAR aligner plays a pivotal role. Spliced Transcripts Alignment to a Reference Genome (STAR) was designed to align the non-contagious sequences directly to the reference genome. STAR algorithm consists of two major steps: seed searching step and clustering/stitching/scoring step [34]. The central idea of the STAR seed finding phase is the sequential search for a Maximal Mappable Prefix (MMP). Figure 4.3 shows the key steps of the Maximal Mappable Prefix strategy, highlighting its role in maximizing the efficiency and accuracy of read alignment to the reference genome. In the first step, the algorithm finds the MMP starting from the first base of the read. In addition to detecting splice junctions, the MMP search, implemented in STAR, enables finding multiple mismatches and indels [34]. In the second phase of the algorithm, STAR builds alignment of the entire read sequence by stitching together all the seeds that were aligned to the genome in the first phase. First, the seeds are clustered together by proximity to a selected set of anchor seeds. The stitching is guided by a local alignment scoring scheme, with user-defined scores (penalties) for matches, mismatches, insertions, deletions, deletions and splice junction gaps, allowing for a quantitative assessment of alignment qualities and ranks [34].

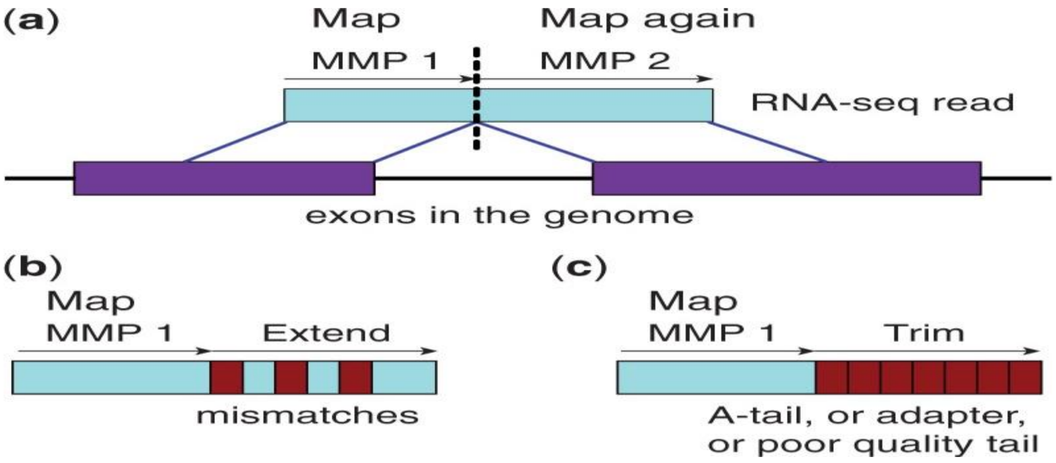


Figure 4.3 Spliced Transcripts Alignment to a Reference (STAR) Aligner Maximal Mappable Prefix (MMP) Strategy Overview (Adapted from Dobin, Alexander et al., 2012 [34])

4.3.1 Selection of Reference Genome

Read alignment to Reference Genome is performed to determine where in the genome the reads come from. The alignment process consists of two steps: 1. Indexing the reference genome 2. Aligning the reads to the reference genome. The selection of an acceptable reference genome is important in the analysis of RNA sequencing data. The reference genome allows for reads alignment. The choice of reference genome highly depends on the species of interest and the availability and quality of reference sequences. When there is a reference genome that has no perfect match, a hybrid reference can be used.

4.3.2 Aligning Reads to the Genome

Aligning reads to the genome, facilitates the accurate mapping of individual sequence reads to their corresponding genomic locations. During this process where each read come from must be identified in the reference genome, which gives an understanding of which genes are expressed and where the expressed regions are located. Accurate read alignment can help in quantifying gene expression levels and identifying splicing events. The outcome of this alignment step is the creation of a Sequence Alignment Map (SAM) file which allows for further analyses.

4.3.3 Generating a Sequence Alignment/Map (SAM File)

Sequence Alignment/Map (SAM) file is a plain text file format used to store the results of sequence read alignments to a reference genome. To generate a SAM file, the process involves aligning individual sequencing reads to a reference genome, determining exactly where the genome is location, and encoding this information in a structured text format. In the SAM file each line corresponds to a single sequence read, containing information details such as the reads name, alignment flags, alignment position and mapping quality. SAM files are also formatted in a way that are readable to humans, it facilitates both manual inspection and the development of custom scripts and algorithms for further data analysis.

4.4 Tools in Analyzing RNA Sequencing Data

The tools used in RNA-Seq data analysis are mainly used in the four general process of RNA-Seq data analysis, including quality control, read alignments, transcript assembly, expression quantification, and differential expression analysis. For the data quality control process the common tools include FASTQ [35], the preprocessing steps uses tools Trimmomatic [36], PRINSEQ [37], and Soapnuke [38]. During the read alignment process the tool used is STAR. During the expression quantification FeatureCounts tool was used for the gene counts data. After normalizing, an expression matrix is generated, and statistical methods can be used to identify differentially expressed genes which is during the differential expression analysis process DESeq2 [39] tool was used to perform this process.

4.4.1 Quantification Tool: FeatureCounts

FeatureCounts is a gene-level quantification approach that utilizes a gene transfer format (GTF) file [40] containing the genome coordinates of exons and genes, and often discard multireads [41]. The data input to FeatureCounts consists of (i) one or more files of aligned reads in either Sequence Alignment/Map (SAM) or Binary Alignment/Map (BAM) format [42]. It also provides other structural elements in the genome such as coding regions, the genome build is essential when obtaining a gene transfer format (GTF) file, because it specifies the reference genome assembly to which the gene annotations in the GTF file correspond. After obtaining the correct GTF file, the next process involves using the GTF file to count the number of reads associated with each feature, providing a fundamental measure of gene expression levels. The FeatureCounts tool also includes ability to accommodate different genome annotations, enabling compatibility with diverse organisms and transcriptome databases. Its efficiency lies in its speed and scalability, making it suited for large studies that are common in genomics research.

4.4.2 Utilizing FeatureCounts for Gene Counts

The process of counts generation using FeatureCounts is a crucial step in RNA-Sequencing data analysis, translating the difficulties of aligned sequencing reads into a quantifiable representation of gene expression. FeatureCounts supports strand-specific read counting if strand- specific information is provided. Read mapping results usually include mapping quality scores for mapped reads [43]. Reads may be paired or unpaired, if paired reads are used then each pair of reads defines a DNA or RNA fragment bookended by the two reads [43]. FeatureCounts excels in this role by systematically parsing through aligned reads and allocating them to specific genomic features, commonly genes. As it crosses the genomic landscape, FeatureCounts accurately tallies the number of reads associated with each gene, producing a comprehensive count for individual genes across all samples. This raw count data forms the backbone of subsequent analyses, offering a snapshot of the excess of each gene in each biological sample. The result of this is a gene-centric counts matrix. The equation (1) below represents the time complexity of the FeatureCounts algorithm. Where f is the number of features, r is the number of reads and k_1 is the number of features included in a genomic bin [43].

$$f \log f + r\sqrt{k_1}$$

equation (1)

4.4.3 Transformation of Individual Counts into an Expression Matrix

The result of generating a gene count using FeatureCounts is a count matrix, this matrix is a tabular representation where genes are in rows and samples are in columns and the number in each cell is the number of reads that mapped to exons in that gene for that sample. The values in the matrix should be counts of sequencing reads or fragments. The transformative step lies in the conversion of these individual counts into a comprehensive expression matrix. Some of the rows can contain only zeros and additionally many rows with only a few fragments total. In this case the raw counts must be normalized, adjusted for library size, pre-filtering is performed to keep only rows that have a count of at least 10 for a minimal number of samples. This normalization process is very important for mitigating technical variations between samples, making sure that the counts matrix accurately reflects the biological refinement of gene expression patterns.

4.5 Metadata Table

Figure 4.5 is a flowchart that shows the crucial steps involved in integrating metadata with RNA Sequencing data. Starting with the DESeq2 package, it details the creation of counts matrix and its integration with metadata.

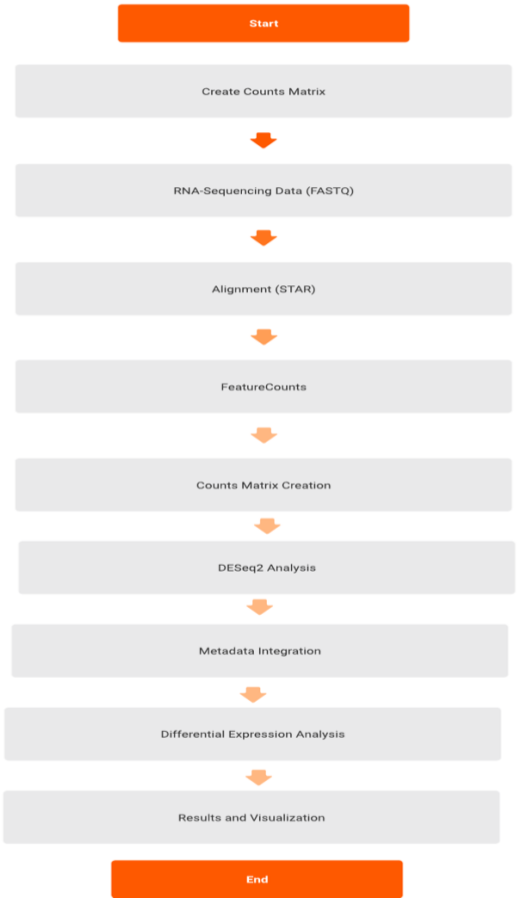


Figure 4.4 Metadata Table Integration Workflow

4.5.1 Creation: Importance and Construction of Metadata Table

The creation of a metadata table is an important step in the orchestration of RNA-Sequencing data analysis. This table is a structured compilation of sample-specific details such as the samples are in rows, experimental conditions, phenotypic characteristics, and any other relevant information that distinguishes one sample from another. The significance of this metadata is its role as a guiding reference, providing important insights into the experimental design and allowing robust statistical analyses. Metadata also makes sure that the biological context of each sample is preserved, allowing for the identification of patterns and correlations between gene expression and experimental variables.

4.6 Workflow: Overview of DESeq2 Workflow

The starting point of a DESeq2 analysis is a count matrix k with one row for each gene i and one column for each sample j [44]. The workflow consists of estimating size factors, dispersion, and fitting a negative binomial distribution. Figure 4.6 shows a standard workflow of DESeq2 including the steps with the tools and packages. This process commences with the input of raw RNA-Sequencing data. The sequenced reads, subjected to quality filtering, alignment or mapping to the respective genome using tools such as STAR. The next important step involves quantifying the reads mapped to genes derived from tools like FeatureCounts into the DESeq2 environment. In the DESeq2 framework, the raw counts go through a normalization process, featuring the estimation of size factors and dispersion to make sure of accurate comparisons across samples. The output, a results table that contains genes that are annotated with log2 fold changes and adjusted p-values, quantifying its significance in differential expression.

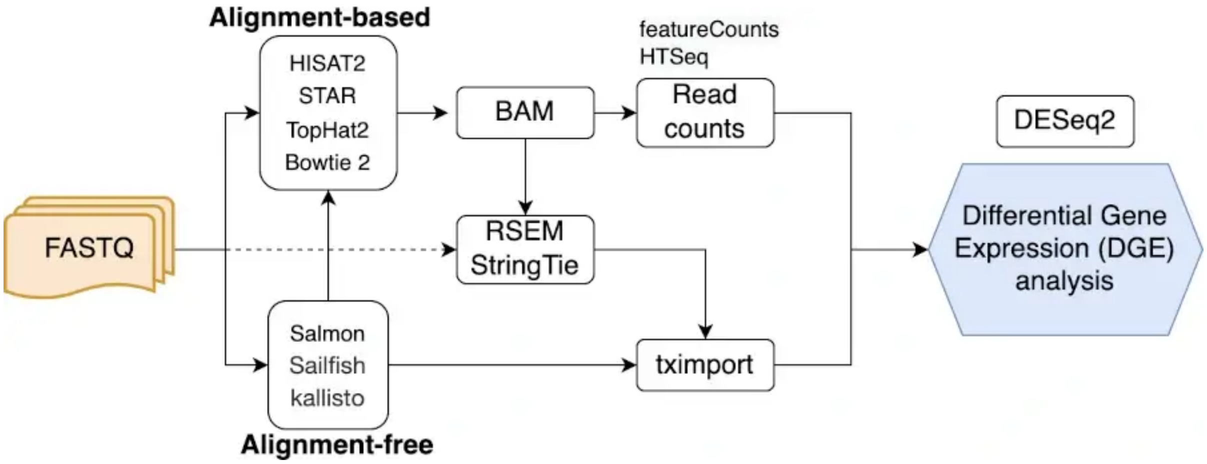


Figure 4.5 Standard workflow of DESeq2 using tools and packages "Title of the Webpage." Differential gene expression analysis using DESeq2 (comphrensive tutorial)

4.6.1 Utilizing DESeq2 for Differential Expression Analysis

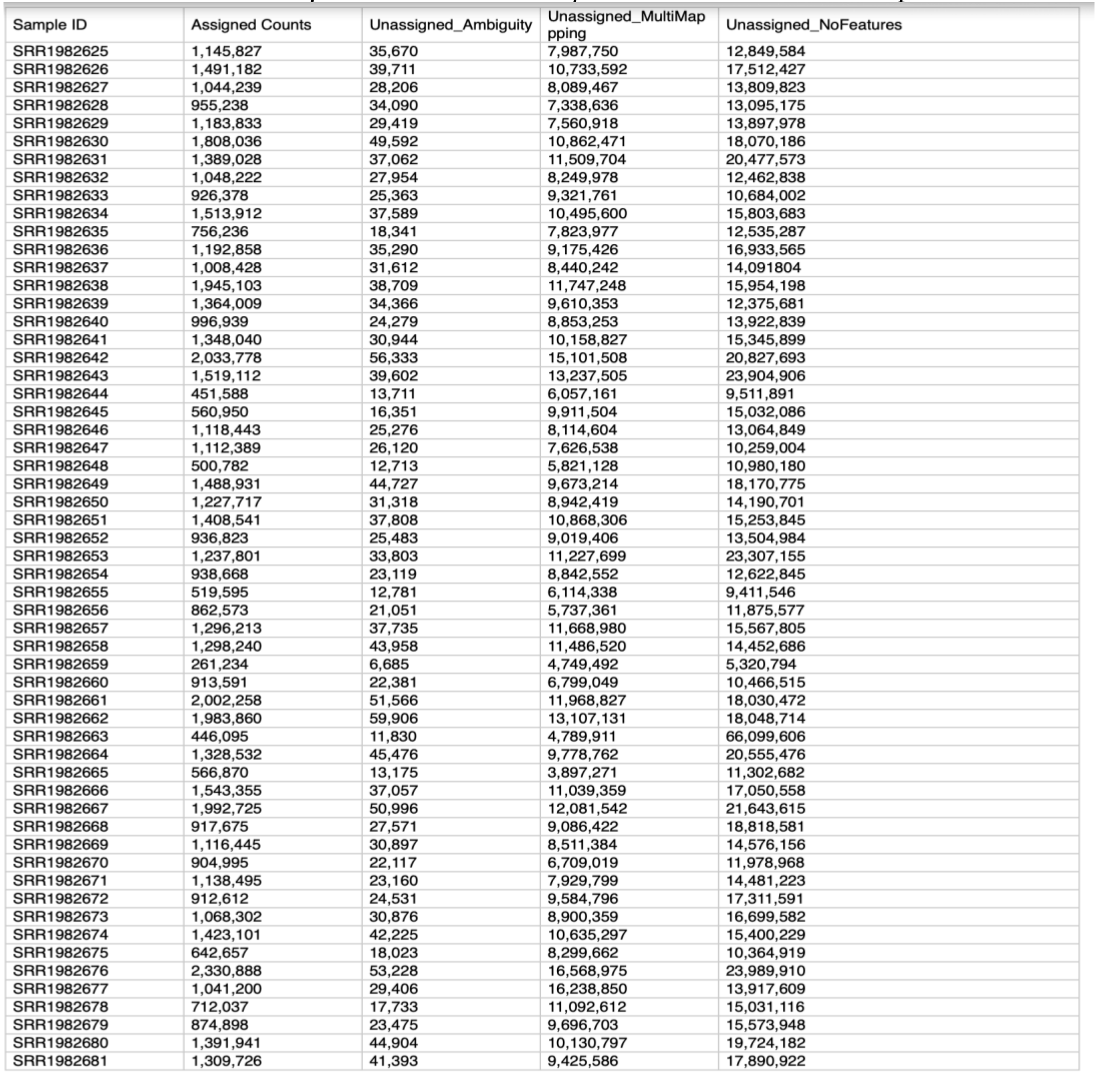
DESeq2 is a subread package in R for analyzing count-based NGS data like RNA-Sequencing and it is available from Bioconductor. There are two tables that contains a csv file, the count matrix is the countData variable and the metadata is the colData. Before preparing the data object in a form that is suitable for analysis, it is very important that the first column of colData (metadata) must match the column names of countData (counts matrix). Once the columns are in the correct order, the DESeqDataSet object can be constructed from the count's matrix and the metadata table. Using the DESeqDataSetFromMatrix function requires the countData (count matrix) to be a matrix or data frame. Either the row names or the first column of the countData must be the identifier that will be used for each gene. The column names of countData are the sample IDs or gene IDs and they must match the row names of colData. There is also a design formula in the data object which specify the model that will be used for testing differential expression. It describes how the counts are expected to change based on experimental factors.

4.7 Performance Results of FeatureCounts on Cancer Samples

The results of FeatureCounts performance were conducted on a comprehensive dataset containing 229 samples across various cancer subtypes, including breast, colorectal, glioblastoma, hepatobiliary, lung, and pancreatic cancer. FeatureCounts effectiveness was assessed by analyzing key assignment statuses such as the counts for Assigned, Unassigned_Ambiguity, Unassigned_MultiMapping, Unassigned_NoFeatures, and Unassigned_Unmapped. It is important to note that the analysis excluded categories with 0 counts, focusing on instances where genes assignments occurred. This performance results provides an understanding of FeatureCounts proficiency in accurately assigning reads to genes in the context of diverse cancer types. The results, detailed in Table 4.7 gives a better overview on

FeatureCounts tool robustness and capabilities, it also highlights its suitability for the subsequent steps in the RNA-Sequencing data analysis pipeline. The average of each category was calculated to gain a better understanding of how FeatureCounts performed on the available samples. The average assigned count is important, representing the number of reads successfully assigned to genomic features. A higher average (106,907,382) indicates a robust capture of relevant information in the genomic data, contributing to the overall success of our analysis.

Table 4.1 Assessment of FeatureCounts Performance on Cancer Samples



Sample ID	Assigned Counts	Unassigned_Ambiguity	Unassigned_MultiMap pping	Unassigned_NoFeatures
SRR1982625	1,145,827	35,670	7,987,750	12,849,584
SRR1982626	1,491,182	39,711	10,733,592	17,512,427
SRR1982627	1,044,239	28,206	8,089,467	13,809,823
SRR1982628	955,238	34,090	7,338,636	13,095,175
SRR1982629	1,183,833	29,419	7,560,918	13,897,978
SRR1982630	1,808,036	49,592	10,862,471	18,070,186
SRR1982631	1,389,028	37,062	11,509,704	20,477,573
SRR1982632	1,048,222	27,954	8,249,978	12,462,838
SRR1982633	926,378	25,363	9,321,761	10,684,002
SRR1982634	1,513,912	37,589	10,495,600	15,803,683
SRR1982635	756,236	18,341	7,823,977	12,535,287
SRR1982636	1,192,858	35,290	9,175,426	16,933,565
SRR1982637	1,008,428	31,612	8,440,242	14,091804
SRR1982638	1,945,103	38,709	11,747,248	15,954,198
SRR1982639	1,364,009	34,366	9,610,353	12,375,681
SRR1982640	996,939	24,279	8,853,253	13,922,839
SRR1982641	1,348,040	30,944	10,158,827	15,345,899
SRR1982642	2,033,778	56,333	15,101,508	20,827,693
SRR1982643	1,519,112	39,602	13,237,505	23,904,906
SRR1982644	451,588	13,711	6,057,161	9,511,891
SRR1982645	560,950	16,351	9,911,504	15,032,086
SRR1982646	1,118,443	25,276	8,114,604	13,064,849
SRR1982647	1,112,389	26,120	7,626,538	10,259,004
SRR1982648	500,782	12,713	5,821,128	10,980,180
SRR1982649	1,488,931	44,727	9,673,214	18,170,775
SRR1982650	1,227,717	31,318	8,942,419	14,190,701
SRR1982651	1,408,541	37,808	10,868,306	15,253,845
SRR1982652	936,823	25,483	9,019,406	13,504,984
SRR1982653	1,237,801	33,803	11,227,699	23,307,155
SRR1982654	938,668	23,119	8,842,552	12,622,845
SRR1982655	519,595	12,781	6,114,338	9,411,546
SRR1982656	862,573	21,051	5,737,361	11,875,577
SRR1982657	1,296,213	37,735	11,668,980	15,567,805
SRR1982658	1,298,240	43,958	11,486,520	14,452,686
SRR1982659	261,234	6,685	4,749,492	5,320,794
SRR1982660	913,591	22,381	6,799,049	10,466,515
SRR1982661	2,002,258	51,566	11,968,827	18,030,472
SRR1982662	1,983,860	59,906	13,107,131	18,048,714
SRR1982663	446,095	11,830	4,789,911	66,099,606
SRR1982664	1,328,532	45,476	9,778,762	20,555,476
SRR1982665	566,870	13,175	3,897,271	11,302,682
SRR1982666	1,543,355	37,057	11,039,359	17,050,558
SRR1982667	1,992,725	50,996	12,081,542	21,643,615
SRR1982668	917,675	27,571	9,086,422	18,818,581
SRR1982669	1,116,445	30,897	8,511,384	14,576,156
SRR1982670	904,995	22,117	6,709,019	11,978,968
SRR1982671	1,138,495	23,160	7,929,799	14,481,223
SRR1982672	912,612	24,531	9,584,796	17,311,591
SRR1982673	1,068,302	30,876	8,900,359	16,699,582
SRR1982674	1,423,101	42,225	10,635,297	15,400,229
SRR1982675	642,657	18,023	8,299,662	10,364,919
SRR1982676	2,330,888	53,228	16,568,975	23,989,910
SRR1982677	1,041,200	29,406	16,238,850	13,917,609
SRR1982678	712,037	17,733	11,092,612	15,031,116
SRR1982679	874,898	23,475	9,696,703	15,573,948
SRR1982680	1,391,941	44,904	10,130,797	19,724,182
SRR1982681	1,309,726	41,393	9,425,586	17,890,922

4.7.1 Averages of FeatureCounts Results

In section 4.7 the performance results of FeatureCounts on cancer samples were presented, detailing key metrics such as Assigned Counts, Unassigned_Ambiguity, Unassigned_MuliMapping, and Unassigned_NoFeatures. To further explain the overall trends, averages were calculated across these categories to provide a comprehensive overview of the distribution of gene expression data for each cancer types. The average (Avg) for each metric was computed using the equation (2) below. The calculated averages for each category are average for Assigned Counts: 106,907,382, Unassigned_Ambiguity: 2,871.389.84, Unassigned_MultiMapping: 86,834,588.696, and Unassigned_NoFeatures: 1,531,655,370.87. These averages offer insights into the typical values observed in each category. The Assigned Counts average represents the average number of reads confidently assigned to specific genes. The Unassigned_Ambiguity average indicates the average number of reads with ambiguous mapping, while Unassigned_MultiMapping signifies the average number of reads mapping to multiple genomic locations. And Unassigned_NoFeatures represents the average number of reads that did not align to any annotated features. Understanding these averages helps in assessing the performance of FeatureCounts in handling distinct aspects of gene expression data.

$$Avg = \frac{Sum\ of\ values\ in\ the\ category}{Number\ of\ samples\ in\ the\ category} \qquad \text{equation (2)}$$

5 DESEQ2 ANALYSIS RESULT AND VISUALIZATIONS

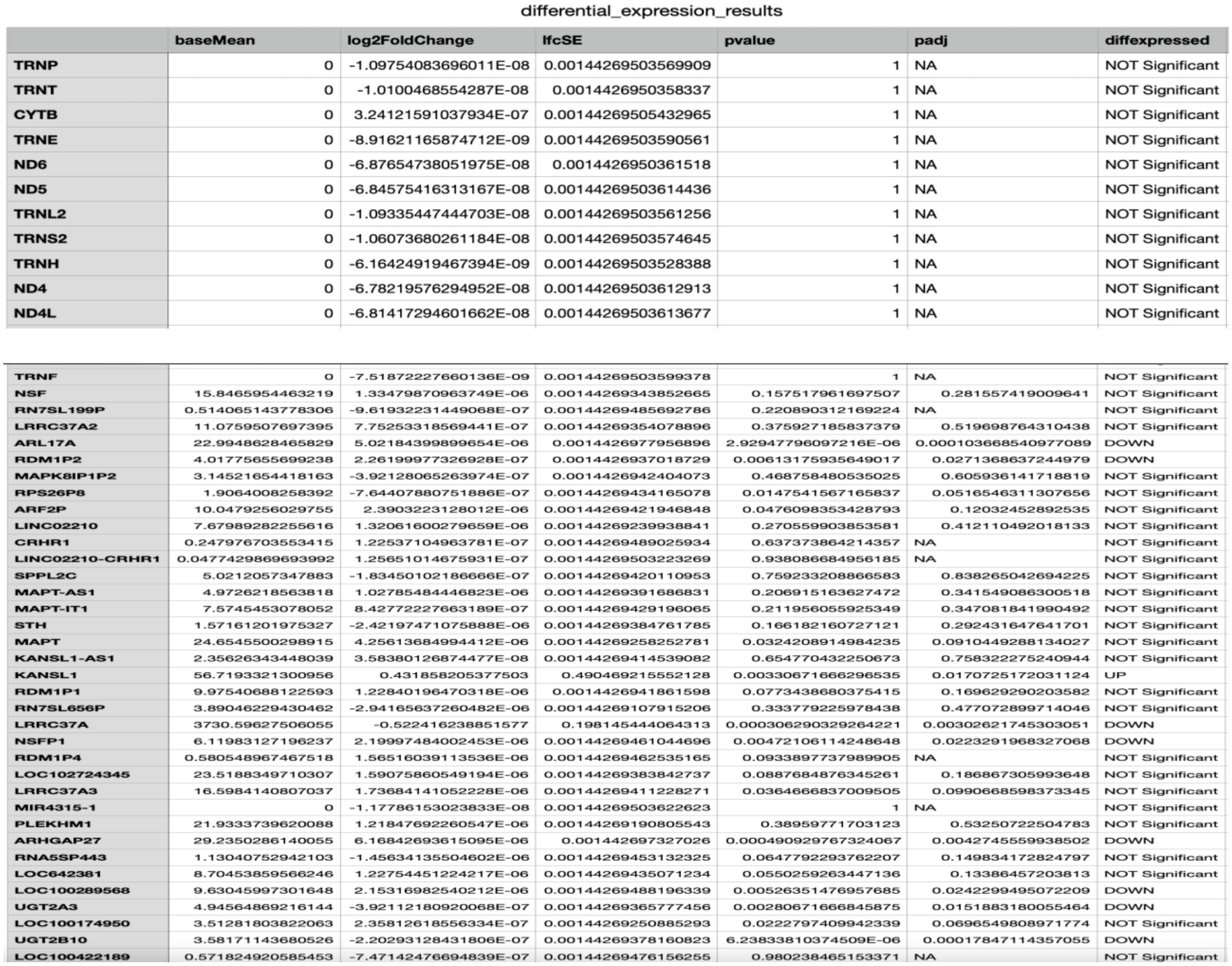
5.1 Introduction

This section provides a comprehensive introduction to the results obtained through the DESeq2 analysis. DESeq2 is a tool used for differential gene expression analysis, allowing the identification of genes with significant upregulation and downregulation across different conditions. The following visualizations present an exploration of these differentially expressed genes. Notable features include a results table detailing genes with significant expression changes, MA plots showcasing the distribution of log-fold changes against mean expression, Principal Component Analysis (PCA) visualizations providing insights into sample relationships, and volcano and dispersion plots displaying the statistical significance of gene expression alterations. Additionally, counts plot visualizations present a detailed view of gene expression counts, while histograms and heatmaps offer an overview of the data distribution and relationships.

5.2 Genes with significant upregulation and downregulation result

In this section, we present a detailed examination of genes exhibiting significant upregulation and downregulation across the cancer types. The results table summarize a comprehensive overview of these genes, highlighting the individual expression changes and statistical significance. Specific genes are identified as the top in the observed transcriptional alterations. The genes in the results table are sorted by the log2 fold change estimate to get the significant genes with the strongest up-regulation and strongest down-regulations. Figure 5.2 gives a better overview providing a curated list of genes important for understanding the differentially expressed level of each cancer subtype. Some values in the table can be set to NA because if in a row all samples have zero counts, the baseMean column will be zero, and the log2 fold change estimates, p-value and adjusted p value all will be set to NA. Another reason is if a row contains a sample with an extreme count outlier, then the p value and adjusted p value will be set to NA. Based on my findings, Gene KANSL1: had a remarkable upregulation of 56.72 log fold. Gene ARL17A: had a significant down-regulation, Gene LRRC37A: had a huge upregulation of 3730.60 log fold. Gene ARHGAP27: had a substantial downregulation this can influence cancer progress, Gene NSFP1: had a significant downregulation and Gene UGT2B10: had a significant downregulation. I also observed that not all genes show significant changes, which is a common observation in large-scale genomic studies.

Table 5.1 Significantly Upregulated and Downregulated Genes across Caner Subtypes



differential_expression_results

	baseMean	log2FoldChange	lfcSE	pvalue	padj	diffexpressed
TRNP	0	-1.09754083696011E-08	0.00144269503569909	1	NA	NOT Significant
TRNT	0	-1.0100468554287E-08	0.0014426950358337	1	NA	NOT Significant
CYTB	0	3.24121591037934E-07	0.00144269505432965	1	NA	NOT Significant
TRNE	0	-8.91621165874712E-09	0.00144269503590561	1	NA	NOT Significant
ND6	0	-6.87654738051975E-08	0.0014426950361518	1	NA	NOT Significant
ND5	0	-6.84575416313167E-08	0.00144269503614436	1	NA	NOT Significant
TRNL2	0	-1.09335447444703E-08	0.00144269503561256	1	NA	NOT Significant
TRNS2	0	-1.06073680261184E-08	0.00144269503574645	1	NA	NOT Significant
TRNH	0	-6.16424919467394E-09	0.00144269503528388	1	NA	NOT Significant
ND4	0	-6.78219576294952E-08	0.00144269503612913	1	NA	NOT Significant
ND4L	0	-6.81417294601662E-08	0.00144269503613677	1	NA	NOT Significant
TRNF	0	-7.51872227660136E-09	0.00144269503599378	1	NA	NOT Significant
NSF	15.8465954463219	1.33479870963749E-06	0.00144269343852665	0.157517961697507	0.281557419009641	NOT Significant
RN7SL199P	0.514065143778306	-9.61932231449068E-07	0.00144269485692786	0.220890312169224	NA	NOT Significant
LRRC37A2	11.0759507697395	7.75253318569441E-07	0.00144269354078896	0.375927185837379	0.519698764310438	NOT Significant
ARL17A	22.9948628465829	5.02184399899654E-06	0.0014426977956896	2.92947796097216E-06	0.000103668540977089	DOWN
RDM1P2	4.01775655699238	2.26199977326928E-07	0.0014426937018729	0.00613175935649017	0.0271368637244979	DOWN
MAPK8IP1P2	3.14521654418163	-3.92128065263974E-07	0.0014426942404073	0.468758480535025	0.605936141718819	NOT Significant
RPS26P8	1.9064008258392	-7.64407880751886E-07	0.00144269434165078	0.0147541567165837	0.0516546311307656	NOT Significant
ARF2P	10.0479256029755	2.3903223128012E-06	0.00144269421946848	0.0476098353428793	0.12032452892535	NOT Significant
LINC02210	7.67989282255616	1.32061600279659E-06	0.00144269239938841	0.270559903853581	0.412110492018133	NOT Significant
CRHR1	0.247976703553415	1.22537104963781E-07	0.00144269489025934	0.637373864214357	NA	NOT Significant
LINC02210-CRHR1	0.0477429869693992	1.25651014675931E-07	0.00144269503223269	0.938086684956185	NA	NOT Significant
SPPL2C	5.0212057347883	-1.83450102186666E-07	0.00144269420110953	0.759233208866583	0.838265042694225	NOT Significant
MAPT-AS1	4.9726218563818	1.02785484446823E-06	0.00144269391686831	0.206915163627472	0.341549086300518	NOT Significant
MAPT-IT1	7.5745453078052	8.42772227663189E-07	0.00144269429196065	0.211956055925349	0.347081841990492	NOT Significant
STH	1.57161201975327	-2.42197471075888E-06	0.00144269384761785	0.166182160727121	0.292431647641701	NOT Significant
MAPT	24.6545500298915	4.25613684994412E-06	0.00144269258252781	0.0324208914984235	0.0910449288134027	NOT Significant
KANSL1-AS1	2.35626343448039	3.58380126874477E-08	0.00144269414539082	0.654770432250673	0.758322275240944	NOT Significant
KANSL1	56.7193321300956	0.431858205377503	0.490469215552128	0.00330671666296535	0.0170725172031124	UP
RDM1P1	9.97540688122593	1.22840196470318E-06	0.0014426941861598	0.0773438680375415	0.169629290203582	NOT Significant
RN7SL656P	3.89046229430462	-2.94165637260482E-06	0.00144269107915206	0.333779225978438	0.477072899714046	NOT Significant
LRRC37A	3730.59627506055	-0.522416238851577	0.198145444064313	0.000306290329264221	0.00302621745303051	DOWN
NSFP1	6.11983127196237	2.19997484002453E-06	0.00144269461044696	0.00472106114248648	0.0223291968327068	DOWN
RDM1P4	0.580548967467518	1.56516039113536E-06	0.00144269462535165	0.0933897737989905	NA	NOT Significant
LOC102724345	23.5188349710307	1.59075860549194E-06	0.00144269383842737	0.0887684876345261	0.186867305993648	NOT Significant
LRRC37A3	16.5984140807037	1.73684141052228E-06	0.00144269411228271	0.0364666837009505	0.0990668598373345	NOT Significant
MIR4315-1	0	-1.17786153023833E-08	0.00144269503622623	1	NA	NOT Significant
PLEKHM1	21.9333739620088	1.21847692260547E-06	0.00144269190805543	0.38959771703123	0.53250722504783	NOT Significant
ARHGAP27	29.2350286140055	6.16842693615095E-06	0.001442697327026	0.000490929767324067	0.0042745559938502	DOWN
RNA5SP443	1.13040752942103	-1.45634135504602E-06	0.00144269453132325	0.0647792293762207	0.149834172824797	NOT Significant
LOC642381	8.70453859566246	1.22754451224217E-06	0.00144269435071234	0.0550259263447136	0.13386457203813	NOT Significant
LOC100289568	9.63045997301648	2.15316982540212E-06	0.00144269488196339	0.00526351476957685	0.0242299495072209	DOWN
UGT2A3	4.94564869216144	-3.92112180920068E-07	0.00144269365777456	0.00280671666845875	0.0151883180055464	DOWN
LOC100174950	3.51281803822063	2.35812618556334E-07	0.00144269250885293	0.0222797409942339	0.0696549808971774	NOT Significant
UGT2B10	3.58171143680526	-2.20293128431806E-07	0.00144269378160823	6.23833810374509E-06	0.00017847114357055	DOWN
LOC100422189	0.571824920585453	-7.47142476694839E-07	0.00144269476156255	0.980238465153371	NA	NOT Significant

5.3 In-depth Exploration of MA Plots: Unraveling Differential Gene Expression with Normal, Ashr, and Apeglm methods

DESeq2 MA plots offer a comprehensive visual representation of the differential expression analysis results. It provides a useful overview for the distribution of the estimated coefficients in the model, the comparisons of interest across all genes. On the x-axis is the average of the counts normalized by size factor and on the y-axis is the log2 fold change for a particular comparison and each gene is represented with a dot. Figure 5.3 shows the 3 methods and gives a comprehensive overview of the differential expression analysis results. The normal MA plot is the original DESeq2 shrinkage estimator and is centered on zero and with a scale that is fit to the data. There are two alternative adaptive shrinkage estimator Apeglm and Ashr. Apeglm (Approximate Posterior Estimates) is the adaptive t prior shrinkage estimator, this method is used for shrinking coefficients which is good for shrinking the noisy LFC estimates while giving low bias LFC estimates for true large differences. Ashr (Adaptive SHrinkage) is the adaptive shrinkage estimator that shrinks log fold changes with very low counts and highly variable counts. The genes with an adjusted p value below a threshold are shown in blue.

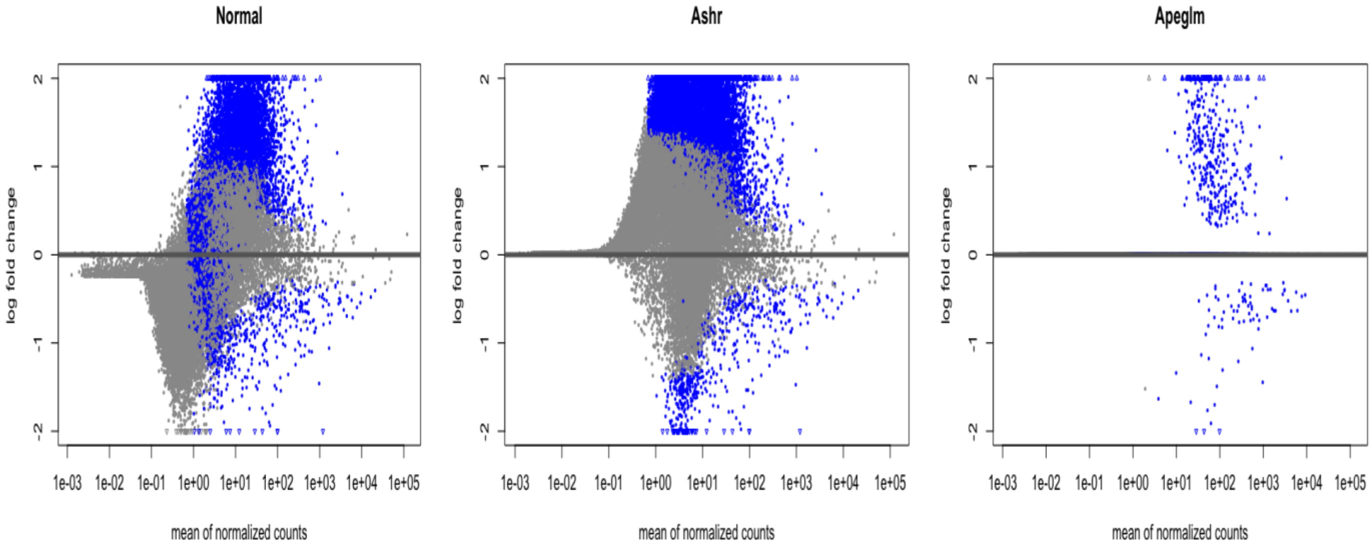


Figure 5.1 MA Plots: Normal, Ashr, and Apeglm Methods

5.4 Exploring Transcriptional Diversity: PCA Visualization

In section 5.4, we examine into the landscape of gene expression through Principal Component Analysis (PCA). PCA is a technique for dimensionality reduction, allows us to explore the underlying patterns in our gene expression data. Through this section we present visual representation that capture the variability and relationships included in samples. Figure 5.4 shows sample to sample distances through the PCA. The samples are projected onto the 2D plane such that they spread out in the two directions that explain most of the differences, on the x axis is the direction that separate the data points the most. PC1 represents the values of the samples and on the y- axis is a direction that must be orthogonal to the first direction, and it separates the data the second most. PC2 represents the values of the samples in that direction, the percentage of PC1 and PC2 variance does not add to 100 percent, because the distances have more dimensions that contain the remaining variance. Based on my observation Pancreas subtype wt had the most overlap because the distance was too large and spread across the dimensions.

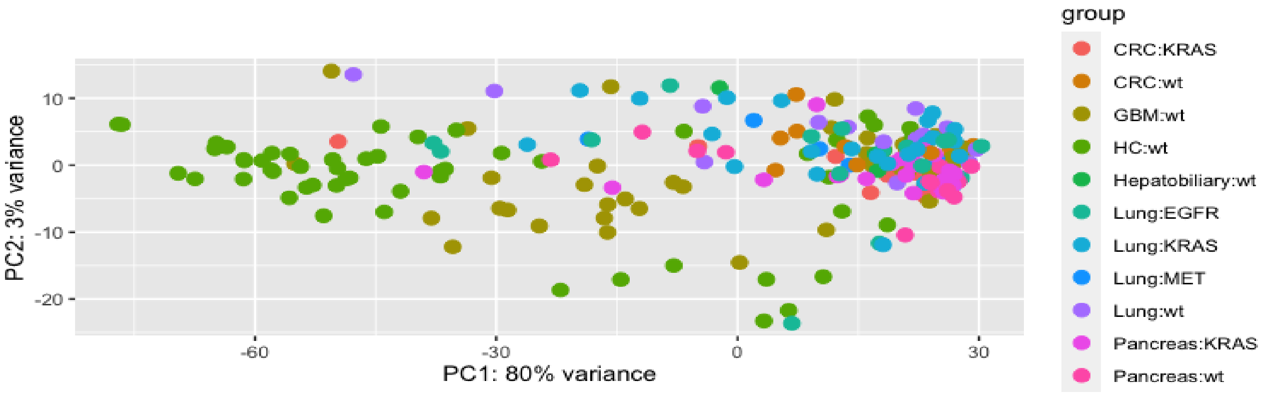


Figure 5.2 Principal Component Analysis (PCA) provides insights into the variance and patterns in the gene expression data across different cancer subtypes

5.5 Exploring Differential Gene Expression with Volcano Plots

The volcano plot visualization examines into the differential gene expression landscape and is a type of scatterplot that shows statistical significance p value vs magnitude of change (fold change). It allows for a visual identification of genes with large fold changes that are also statistically significant. In a volcano plot, the most upregulated genes are towards the right, the most downregulated genes are towards the left, and the most statistically significant genes are towards the top. On the x-axis represents the fold change in gene expression between two conditions and the y-axis represents the statistical significance of the change, often expressed as -log10(p-value). Figure 5.5 shows a volcano plot that allows us to understand the molecular distinctions across the cancer analyzed subtypes by highlighting all significant genes that surpass a significance threshold adjusted p-value < 0.05 and highlight genes with a considerable fold change log2(fold change) > 1. Based on the observation of the plot, there is not a densely populated symmetrical V shape because the observations are reduced or the variation in response is not so evenly distributed.

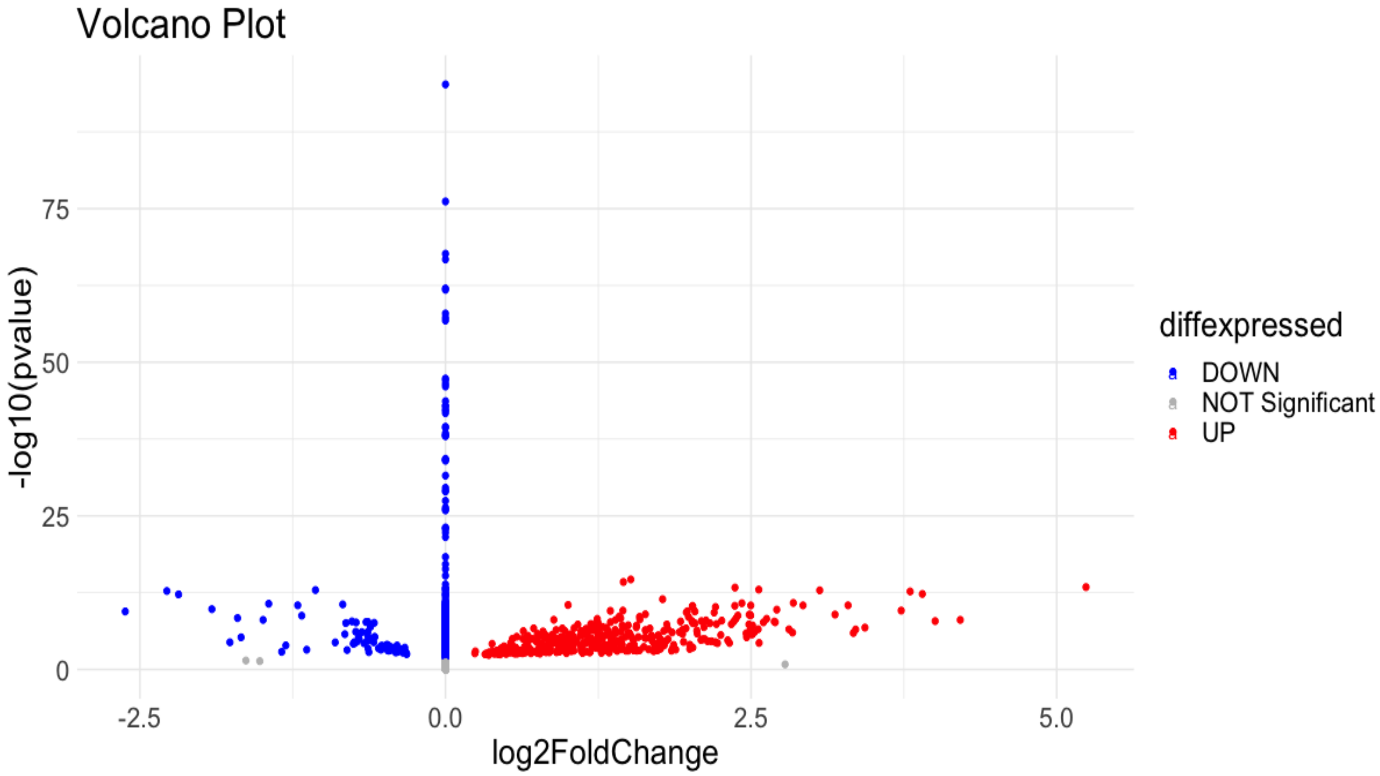


Figure 5.3 Volcano Plot visually depict gene expression changes

5.6 Exploring Gene Expression Distribution with Counts Plot

In this section, it goes into counts plots offering a detailed exploration of gene expression patterns across samples. These plots visualize the distribution of read counts for each gene across the groups. On the x-axis represents the mutation subtypes or groups and, on the y-axis, depicts the normalized counts or expression levels of the gene. Normalized counts are expression levels of the gene that are normalized to account for variations in library size and other technical factors. Figure 5.6 shows 3 counts plot gives us a better interpretation of RNA-Sequencing data by examining these plots, patterns and trends in gene expression can be identified. The plots show the gene which had the smallest p value from the results. This gene was represented in two other plots through its normalized counts with lines connecting cell lines. Based on my observation the Gene with the smallest p value is RNA5SP500.

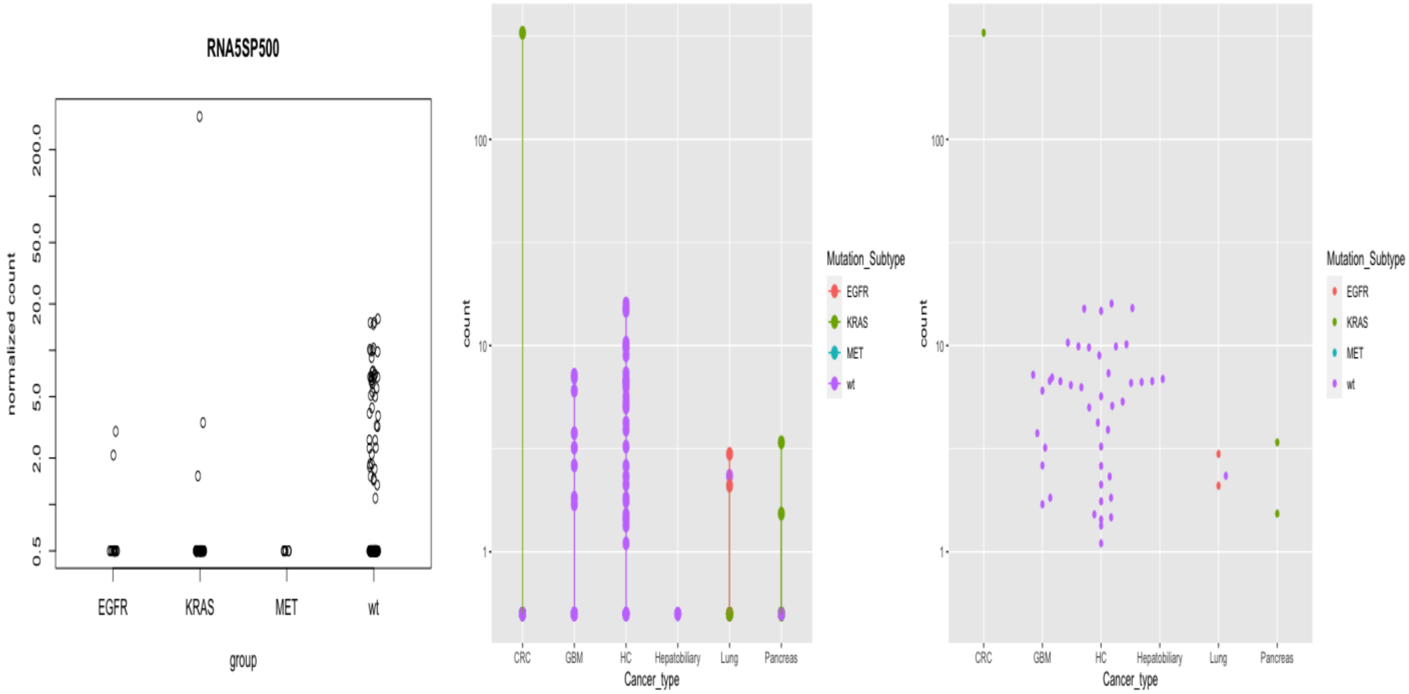


Figure 5.4 Visualizing gene expression distribution through Counts Plots

5.7 Exploring Gene Expression Variability: Heatmap Visualization of Top 20 Genes

In this section, we delve into a comprehensive exploration of gene expression variability through heatmap visualization. This technique allows us to recognize patterns in the expression of the top 20 genes with the highest variance across the samples. Figure 5.7 provides a visually representation, allowing the identification of trends in gene expression patterns. Based on my observation of the heatmap, Gene TTN has been regarded as an important marker for the distinction of six cancer subtypes and healthy controls group. Another gene ENPP7P4 is suitable for it to act as a liquid biopsy marker and regulation of the cell cycle.

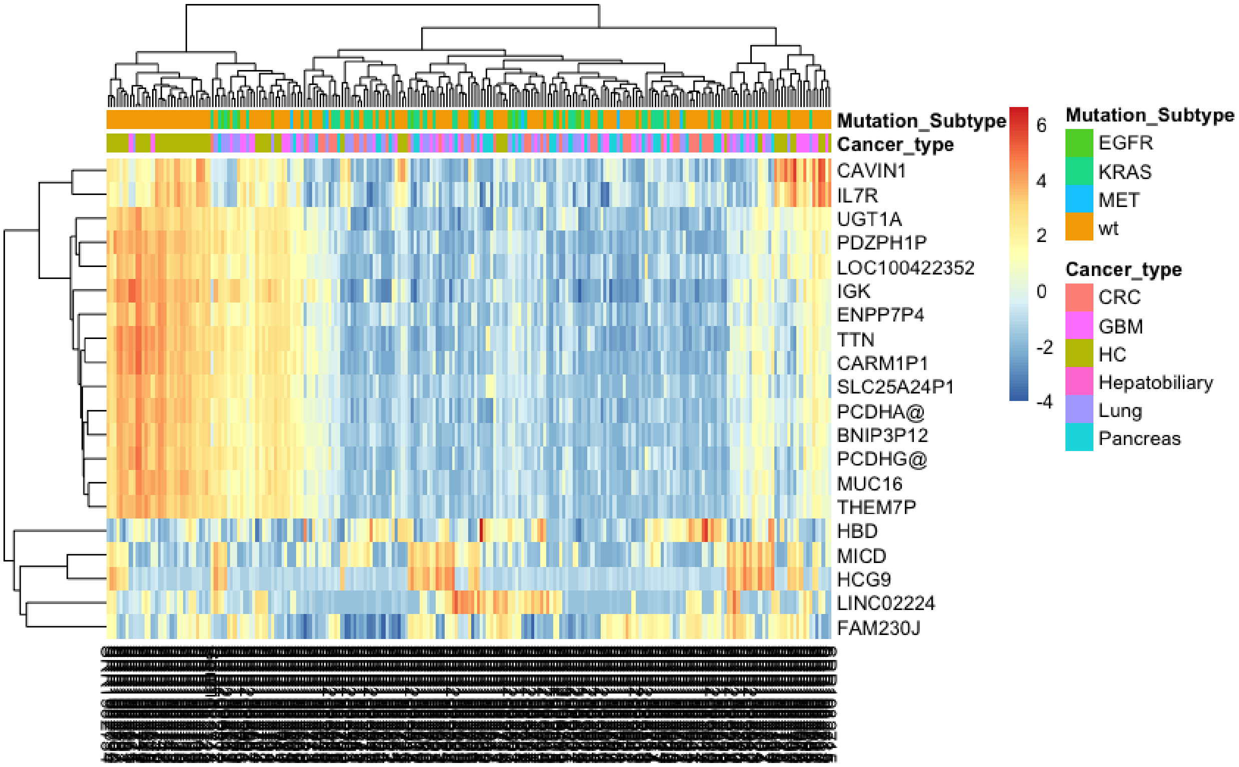


Figure 5.5 Heatmap of top 20 genes with high variance across samples

5.8 Exploring Gene Expression Distributions: Histogram Analysis across Cancer Subtypes

In this section, histograms visualizations provide a comprehensive view of the distribution patterns of gene expression values in and across cancer subtypes and the healthy control group. The histograms offer insights into the spread and frequency of expression levels, aiding in the identification of distinctive expression profiles. Figure 5.8 shows two histograms, the histogram on the left is a histogram that displays the distribution of p values for genes with a mean normalized count larger than 1. It is formed by excluding genes with very small counts, which can generate spikes in the histogram if not removed. The histogram on the right is a histogram that presents the ratio of small p values, binned by normalized count. The p values are from a test of log2 fold change greater than 1 or less than -1. This histogram is formed by demonstrating that genes with very low mean counts have little influence and are best if removed from testing. By removing the low count genes from the input to the FDR procedure, we can find more genes to be significant along the genes that we keep by applying independent filtering.

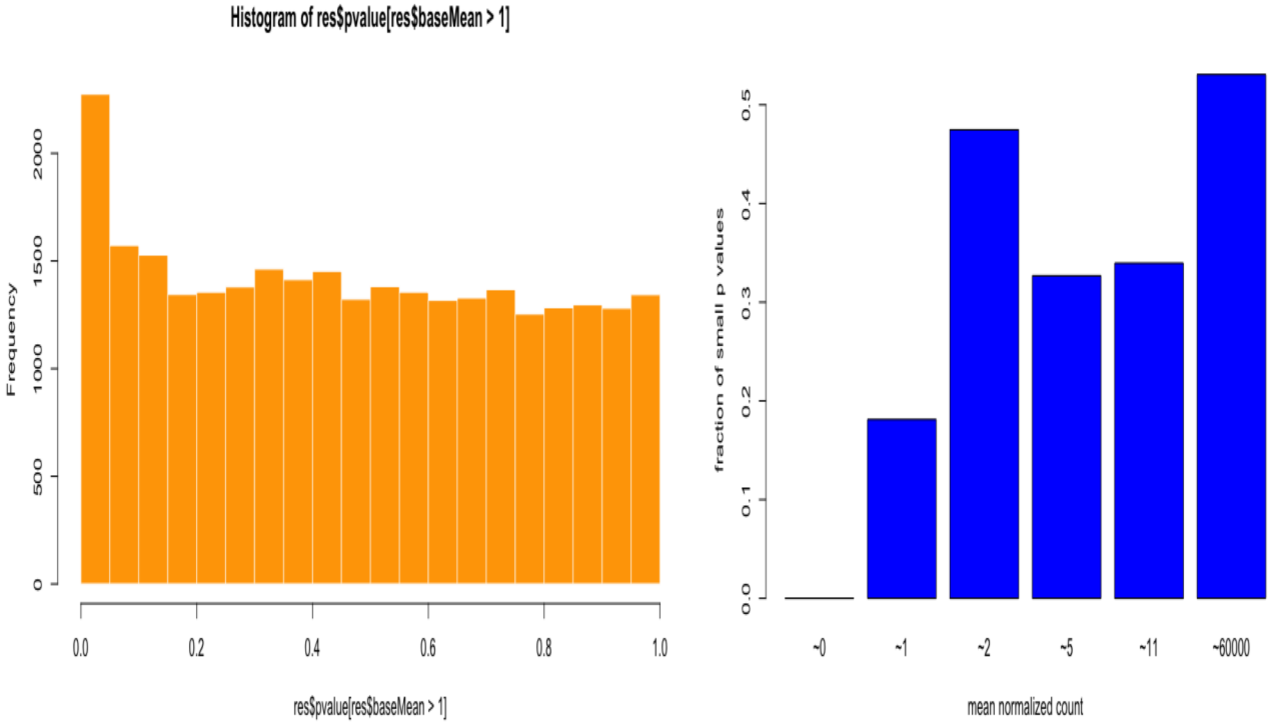


Figure 5.6 Capturing the diversity in gene expression through histogram analysis across cancer subtypes

6 MACHINE LEARNING ANALYSIS

6.1 Introduction

Machine learning, which falls under the umbrella of artificial intelligence and computer science, involves the development of algorithms and models that enable computers to learn and make predictions or decisions autonomously without the need for explicit programming instructions [45]. Cancer classification is the process of categorizing different types of cancers based on their characteristics, such as the site of origin, histological features, genetic mutations, and clinical behavior [45]. Accurate classification of cancer plays a significant part in ensuring precise diagnosis, treatment planning, and predicting patient outcomes [46]. Classification constitutes a fundamental undertaking in supervised learning, where the objective is to train a model to forecast the class designation of a given input by considering its distinctive attributes [46]. Accurate cancer classification holds importance of personalized treatment strategies. The unique genetic signatures identified through precise subtype classification allows clinicians to get treatment regimens based on the specific molecular characteristics of each cancer subtype. Furthermore, accurate classification contributes to the identification of novel biomarkers associated with distinct cancer subtypes, aiding advancements in early detection and targeted therapies. Figure 6.1 shows an overview of the machine learning applications for cancer classification, and it gives a simpler understanding.

6.1.1 Role in Research

The goals of this research revolve around utilizing machine learning techniques to gain meaningful insights from high-dimensional gene expression data for accurate cancer classification. The specific objectives enclose the identification of distinctive transcriptomic signatures for six major cancer subtypes: breast, colorectal, glioblastoma, hepatobiliary, lung, and pancreatic cancer, including a healthy control group. Through the application of machine learning models, the study aims to generate robust classification algorithms that are capable of differentiation between these cancer types based on their unique gene expression profiles. Artificial learning techniques have a pivotal role to play in cancer classification by analyzing complex and high – dimensional datasets. By identifying hidden patterns and relationships in the data, machine learning algorithms can discover subtle associations between generic or molecular markers and different cancer types, leading to improved classification accuracy [47].

6.2 Data Overview

In this section, delve into a comprehensive understanding of the dataset. The dataset employed in this study is derived from the gene expression omnibus (GEO) accession number GSE68086. It contains samples from six distinct cancer subtypes breast, colorectal, glioblastoma, hepatobiliary, lung, and pancreatic cancer and including a healthy control group. The dataset comprises a total of 285 but only 229 samples was successful, with each cancer subtype contributing varying sample sizes: breast (39 samples), colorectal (42 samples), glioblastoma (40 samples), hepatobiliary (14 samples), lung (60 samples), pancreatic (35 samples), and the healthy control group (55 samples). A characteristic of the genomic data is its high-dimensional containing 16,383 genes. Furthermore, the gene count data increases the dimensionality, culminating in a total of 43,682 features. The high-dimensional dataset leads to challenges in the analysis and interpretation of gene expression patterns.

6.3 Data Preprocessing

This section delves into the importance of data preprocessing particularly feature scaling in the gene expression data. Feature scaling, is an important preprocessing step that makes sure that all features contribute equally to the model performance, preventing dominance by variables with larger magnitudes. By normalizing features, feature scaling enhances the comparability and interpretability of different attributes, allowing a more accurate representation of the biological patterns. This is mainly vital in the context of cancer subtype classification where variations in gene expression helps with distinguishing between different tumor types. The feature method used in my analysis is StandardScaler for SVM and MLP classifier. StandardScaler scales the features to have a mean of 0 and a standard deviation of 1. Equation (3) shows the formula how the Z-score normalization is computed. StandardScaler was used because the datasets ranges were greatly different from each other, therefore this method is used to standardize the range of functionality. Considerations I observed during feature scaling is to reduce the features for the performance of the model and to increase the sample size.

$$X_{Normalized} = X - \frac{\min(X)}{\max(X)} - \min\,(X) \qquad \text{equation (3)}$$

6.3.1 Labeling

In the process of classifying cancer subtypes based on transcriptomic data, the labeling process plays an important role in shaping the structure of the machine learning models. The classes or labels applied in this classification are Cancer_types and Mutation_Subtypes. These labels were assigned to samples using the DESeq2 analysis process, the metadata file contains the columns of Cancer_types and Mutation_Subtypes. The counts matrix and metadata file were integrated into the machine learning analysis.

6.4 Model Training

Machine learning and artificial intelligence algorithms can be trained using large datasets to develop predictive models for cancer classification. These models can incorporate various data types, such as clinical information, imaging data, and molecular profiles, to classify tumors and assist in diagnosis and treatment decisions [48]. Seven classifiers were used: SVM, KNN, LR, DT, RF, NB, and MLP. SVM is a powerful and widely used method because it can handle data that cannot be linearly divided by translating it into a higher-dimensional space with a linear boundary to separate the classes. In SVM, the purpose is to select the optimal hyperplane for classifying the data [49]. Machine learning techniques such as K-Nearest Neighbors (KNN) are used for both classification and regression problems. Its operation is based on selecting the k closet neighbors to an object in the training dataset, where k is a positive integer of the user's choosing [49]. Logistic regression is a widely adopted statistical technique employed to analyze datasets that encompass one or more independent variables, which have the potential to influence the outcome [49]. The Decision Tree is a widely using algorithm utilized in machine learning and artificial intelligence to address classification and regression tasks. Within a decision tree, an internal node signifies and attribute test, a branch signifies the result of the test, and a leaf node represents a prediction or class label [49]. A classification and regression ensemble learning system is called Random Forest. It is a kind of decision tree method that generates numerous decision trees and combines their prediction to obtain a more reliable and accurate outcome [49]. Naïve Bayes is a probability algorithm based on Bayes theorem which calculates the probability of a hypothesis given observed evidence. In the context of classification, Naïve Bayes assumes that features are conditionally independent, given the class label. One advantage of a naïve Bayes classifier is that it only needs to estimate the necessary parameters (mean and variance of variables) based on a small amount of training data [50]. The equation (4) shows how the probability is computed. P(y|x) is the posterior probability of class y given features x, P(x|y) is the likelihood of observing features x given class y, P(y) is the prior probability of class y, and P(x) is the probability of observing features x. Multi-Layer Perceptron (MLP) is a neural network architecture with fully connected layers where each neuron in a hidden layer is connected to all other neurons in the neighboring layers [49].

$$P(y|x) = \frac{P(x|y) \cdot P(y)}{P(x)} \qquad \text{equation (4)}$$

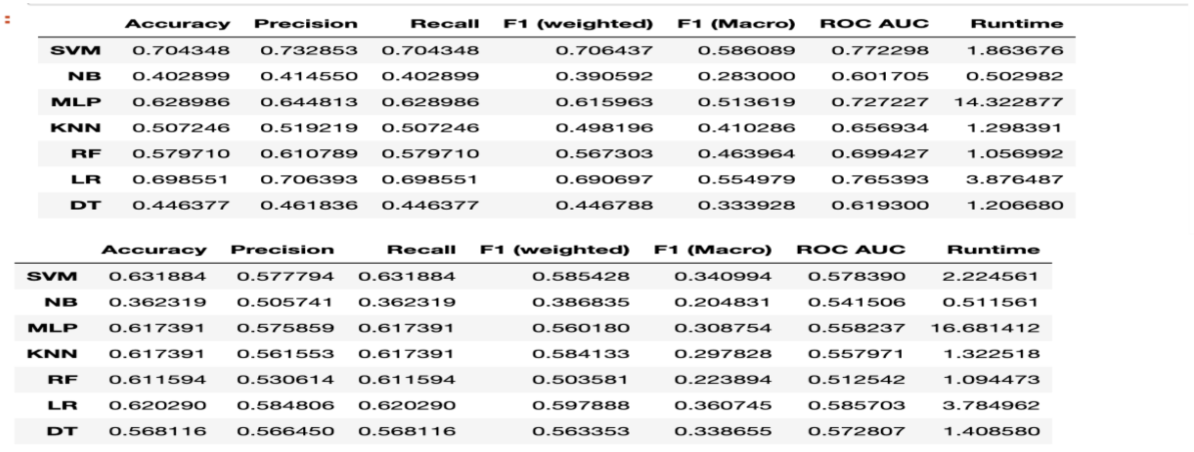
6.4.1 Evaluation

The evaluation process in this study is secured in the utilization of cross-validation, a technique used to assess the performance of a model and reduce the challenges related to overfitting or underfitting. Specifically, ShuffleSplit with five folds, was applied to split the datasets into subsets for training and testing the classifiers. ShuffleSplit generates a user defined number of independent train or test dataset splits. It shuffles the data before splitting it into train and test sets. The choice of five folds aligns with the standard usage in cross-validation balancing computational efficiency and detailed model evaluation. In each iteration, one of the folds is used as the test set, and the remaining four folds are used as the training set. This process is repeated five times with a different fold as the test set in each iteration. The decision to split the test size to 30% in each fold reflects a calculated balance making sure an ample amount of data for testing while preserving a substantial portion for training. Following this approach provides a reliable estimate of the model performance on new unseen data and helps reduce overfitting. The challenges that were faced through this process is that since there was not enough sample size and a high dimensional dataset the performance of the classifiers was greatly affected.

6.5 Results

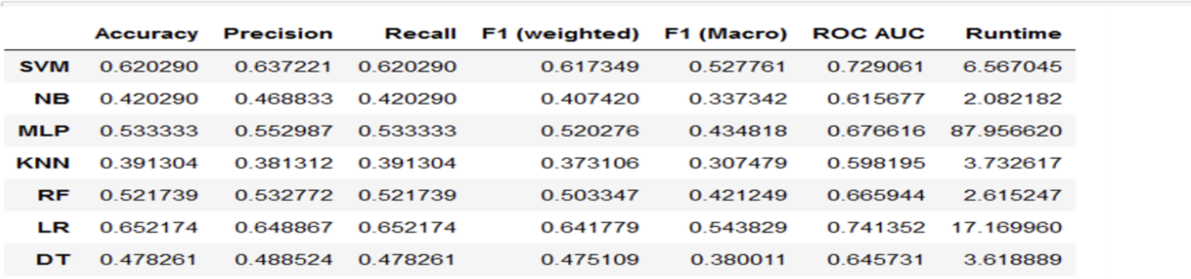
In this section, the focus is on a detailed examinations of the results from the seven classifiers are presented, clarifying their performance in the classification of both cancer_types and mutation _subtypes. The classifiers, including SVM, KNN, Random Forest, Logistic Regression, Decision Tree, MLP, and Naïve Bayes were systematically evaluated across two datasets: one with 229 samples and 16,383 features and the other with 229 samples and the other with 229 samples and 43,682 features. Table 6.5 highlights the results and provides a comparative overview of the classifiers performance. The discussion encompasses a detailed analysis of how well each model performed in capturing the patterns and variations in the genomic data. Metrics such as accuracy, precision, recall, F1 score (weighted, macro, and micro), and ROC AUC are examined carefully to provide a comprehensive understanding of the classifier's efficacy. For the first dataset with 229 samples, 16383 features in Cancer_type SVM accuracy is 70.4% and Mutation_subtype accuracy is 64.3%. Observation: SVM performed better in cancer type classification. KNN accuracy in Cancer_type is 50.7% and accuracy in Mutation_subtype is 61.2%, observation: KNN performs unexpectedly better in mutation subtypes classification. Random forest accuracy in Cancer_type is 57.9% and accuracy is Mutation_subtype is 61.1%, logistic regression accuracy in Cancer_type is 69.8% and accuracy in Mutation_subtype is 62%. Decision Tree accuracy in Cancer_type is 44% and accuracy in Mutation_subtype is 56.8%, Naïve Bayes had the lowest performance across the metrics and the lowest accuracy of 40% and 36% for both classifications. Based on the observations for the second dataset of 229 samples and 43,682 features SVM still had the highest performance and highest accuracy of 64.3% for both classifications. I also observed that SVM went down for this dataset because this was the original dataset that did not have normalization performed. Logistic regression performance was still consistent with the second highest accuracy of 65% and 61% for both classifications.

Table 6.1 Performance Results of Seven Classifiers for Cancer types and Mutation Subtypes Classification

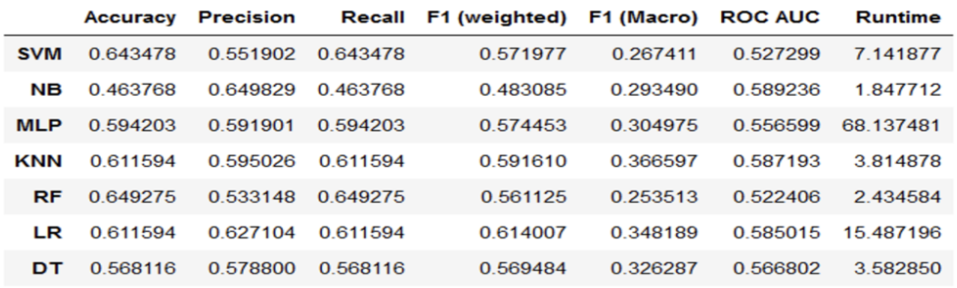


	Accuracy	Precision	Recall	F1 (weighted)	F1 (Macro)	ROC AUC	Runtime
SVM	0.704348	0.732853	0.704348	0.706437	0.586089	0.772298	1.863676
NB	0.402899	0.414550	0.402899	0.390592	0.283000	0.601705	0.502982
MLP	0.628986	0.644813	0.628986	0.615963	0.513619	0.727227	14.322877
KNN	0.507246	0.519219	0.507246	0.498196	0.410286	0.656934	1.298391
RF	0.579710	0.610789	0.579710	0.567303	0.463964	0.699427	1.056992
LR	0.698551	0.706393	0.698551	0.690697	0.554979	0.765393	3.876487
DT	0.446377	0.461836	0.446377	0.446788	0.333928	0.619300	1.206680

	Accuracy	Precision	Recall	F1 (weighted)	F1 (Macro)	ROC AUC	Runtime
SVM	0.631884	0.577794	0.631884	0.585428	0.340994	0.578390	2.224561
NB	0.362319	0.505741	0.362319	0.386835	0.204831	0.541506	0.511561
MLP	0.617391	0.575859	0.617391	0.560180	0.308754	0.558237	16.681412
KNN	0.617391	0.561553	0.617391	0.584133	0.297828	0.557971	1.322518
RF	0.611594	0.530614	0.611594	0.503581	0.223894	0.512542	1.094473
LR	0.620290	0.584806	0.620290	0.597888	0.360745	0.585703	3.784962
DT	0.568116	0.566450	0.568116	0.563353	0.338655	0.572807	1.408580



	Accuracy	Precision	Recall	F1 (weighted)	F1 (Macro)	ROC AUC	Runtime
SVM	0.620290	0.637221	0.620290	0.617349	0.527761	0.729061	6.567045
NB	0.420290	0.468833	0.420290	0.407420	0.337342	0.615677	2.082182
MLP	0.533333	0.552987	0.533333	0.520276	0.434818	0.676616	87.956620
KNN	0.391304	0.381312	0.391304	0.373106	0.307479	0.598195	3.732617
RF	0.521739	0.532772	0.521739	0.503347	0.421249	0.665944	2.615247
LR	0.652174	0.648867	0.652174	0.641779	0.543829	0.741352	17.169960
DT	0.478261	0.488524	0.478261	0.475109	0.380011	0.645731	3.618889

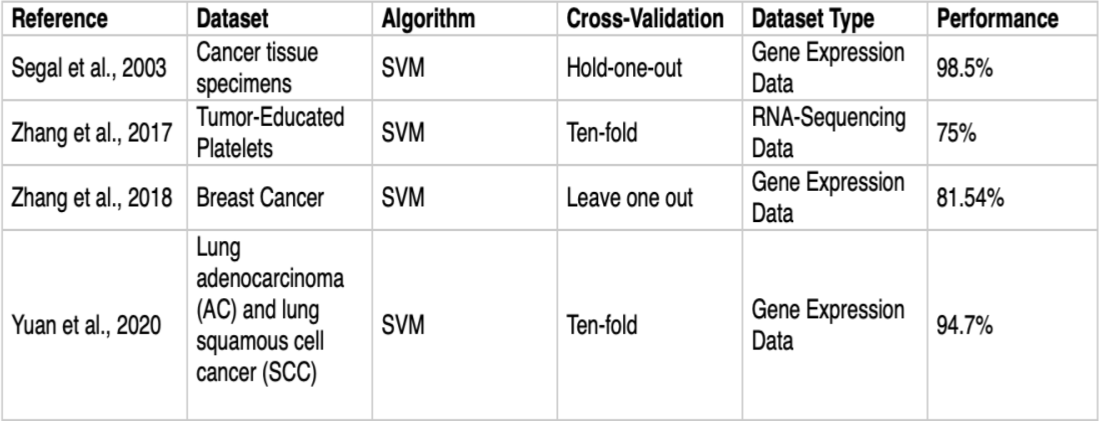


	Accuracy	Precision	Recall	F1 (weighted)	F1 (Macro)	ROC AUC	Runtime
SVM	0.643478	0.551902	0.643478	0.571977	0.267411	0.527299	7.141877
NB	0.463768	0.649829	0.463768	0.483085	0.293490	0.589236	1.847712
MLP	0.594203	0.591901	0.594203	0.574453	0.304975	0.556599	68.137481
KNN	0.611594	0.595026	0.611594	0.591610	0.366597	0.587193	3.814878
RF	0.649275	0.533148	0.649275	0.561125	0.253513	0.522406	2.434584
LR	0.611594	0.627104	0.611594	0.614007	0.348189	0.585015	15.487196
DT	0.568116	0.578800	0.568116	0.569484	0.326287	0.566802	3.582850

6.6 Comparison with Existing Studies

This section compares our results with the reference paper and other relevant studies, in comparing our results with the reference paper by Zhang YH et al., 2017 several main observations and comparisons appear. First, both studies emphasize the significance of machine learning, and the choice of SVM as a classifier is shared between the studies, contributing to the reliability and accuracy of cancer subtype classifications. In terms of performance evaluation, both studies utilize ten -fold-cross-validation, making sure of robust and comparable assessments. However, there was differences such as the referenced paper dataset had 285 samples and 13,445 features by discarding the low counts. In our study there was two datasets with 229 samples 16,383 features and 229 samples 43,682 features. Referenced paper applied an mRMR method as the feature selection method, in our study the only additional process we followed was normalizing the counts. Both studies had SVM as the highest accuracy, referenced paper SVM accuracy was 74% and in our study the accuracy was 70% without any fine tuning or feature selection method. Normalization was performed for various reasons such as it allows for more accurate identification of differentially expressed genes, it also makes sure that the expression values are on a common scale, allowing valid comparisons between samples. Since there are some longer genes with more counts, normalization corrects for this bias allowing for a fair comparison between the genes of different lengths. The similarities in applying liquid biopsy for noninvasive detection and leveraging quantitative gene expression profiles highlights the shared recognition of these methodologies across studies. Table 6.6 highlights the comparison of our study, referenced paper, and other relevant studies SVM accuracy. In conclusion, the comparison of our results with the reference paper and additional studies in the literatures reveals both similarities differences. While shared methodologies and approaches provide a structure for understanding cancer subtypes, variations show the challenges and context specific of genomic data.

Table 6.2 A comparative overview of SVM Accuracy across diverse studies



Reference	Dataset	Algorithm	Cross-Validation	Dataset Type	Performance
Segal et al., 2003	Cancer tissue specimens	SVM	Hold-one-out	Gene Expression Data	98.5%
Zhang et al., 2017	Tumor-Educated Platelets	SVM	Ten-fold	RNA-Sequencing Data	75%
Zhang et al., 2018	Breast Cancer	SVM	Leave one out	Gene Expression Data	81.54%
Yuan et al., 2020	Lung adenocarcinoma (AC) and lung squamous cell cancer (SCC)	SVM	Ten-fold	Gene Expression Data	94.7%

6.7 Future Work

This section discusses future work that can address the challenges that occurred due to the high dimensional dataset which can also improve the performance in cancer subtype classification. Enhancing feature selection methodologies is an important aspect in improving the performance of the classifiers. Feature selections plays a role in refining the set of attributes used for classification leading to improved model interpretability and performance. In section 6.5 the table shows existing literature studies and there SVM accuracy, but each study also used a feature selection method. Segal et al., 2003 used gene ranking for feature selection including the fisher score method, A standard Student's t test was used to compare the expression in one tumor type with that in the remaining samples. The resulting p values were then used to rank the genes, and the desired number of genes was then selected for use. Finally, a statistic t test, as determined for all samples was used to provide an overall ranking of the genes in order of relevance for each tumor classification [45]. By following this feature selection method, SVM performance had an accuracy of 98.5%. Zhang et al., 2017 used mRMR feature selection method that extracted the relevance between features and targets that can be essential biomarkers for the classification of cancer subtypes and healthy control group. After following this feature selection method, SVM performance had an accuracy of 75%. The feature selection method that we will follow is the reads2vec. Reads2vec is a method that transforms raw sequencing data into distributed representations, capturing patterns and relationships in the data. It allows the extraction of meaningful features from RNA-seq data. Applying reads2vec for feature selection can enhance the classification accuracy of cancer subtypes by identifying discriminative genomic patterns. The process of using this method first involves converting the RNA-Sequencing reads into vector representations. This allows the algorithms to learn and represent complex relationships between genes more effectively. Including reads2vec in future studies could be the getaway for more accurate and cancer subtype classification, contributing to advancements in precision medicine.

7 CONCLUSION

In this thesis, we delved into the landscape of cancer subtypes of classification using machine learning algorithms, with a particular focus on RNA-Sequencing data from blood platelets. Drawing inspiration from the work of Zhang et al., (2017), we navigated through the complexities of liquid biopsy and machine learning applications in cancer classification. Our research spanned various cancer subtypes, including breast, colorectal glioblastoma, hepatobiliary, lung, and pancreatic cancer. Leveraging robust machine learning algorithms such as SVM, Random Forest, Naïve Bayes, Logistic Regression, Multi-Perceptron, KNN, and Decision Tree. Our study aimed to enhance the accuracy and precision of cancer subtypes classification. Also, a comparative analysis with existing studies was conducted to get an overview on the varying SVM accuracies in different contexts. As we move forward, the use of feature selection method reads2vec emerges as a crucial aspect to improve the accuracy of the seven classifiers in cancer subtypes classification. This study details the importance of innovative approaches such as liquid biopsy and advanced machine learning methodologies in the pursuit of accurate and personalized cancer diagnostics.

REFERENCES

1. Parsons, Heather M., et al. “Who Treats Adolescents and Young Adults with Cancer? A Report from the AYA HOPE Study.” *Journal of Adolescent and Young Adult Oncology*, vol. 4, 2015, pp. 141-150.

2. McGuire, Shaun. *World Cancer Report 2014*. World Health Organization, International Agency for Research on Cancer, WHO Press, 2016.

3. Wesolowski, Sergiusz, et al. “A Comparison of Methods for RNA-Seq Differential Expression Analysis and a New Empirical Bayes Approach.” vol. 3, 2013, pp. 238-258.

4. Goksuluk, Dincer, et al. “MLSeq: Machine Learning Interface to RNA-Seq Data.” *Computer methods and programs in biomedicine*, vol. 175, 2019, pp. 223-231, doi:10.1016/j.cmpb.2019.04.007.

5. Waseem, Quadri, et al. “Future Technology: Software-Defined Network (SDN) Forensic.” *Symmetry*, vol. 13, no. 5, 2021, p. 767. *Crossref*, https://doi.org/10.3390/sym13050767.

6. Angra, Sheena, and Sachin Ahuja. *Machine learning and its applications: A review*. 2017 International Conference on Big Data Analytics and Computational Intelligence (ICBDAC). 2017, Chirala, Andhra Pradesh, India.

7. Robinson, Mark D., et al. "edgeR: A Bioconductor package for differential expression analysis of digital gene expression data." *Bioinformatics (Oxford, England),* vol. 26, 2010, pp. 139-140, doi:10.1093/bioinformatics/btp616.

8. Harbers, Matthias, and Piero Carninci. "Tag-based approaches for transcriptome research and genome annotation." *Nature Methods*, vol. 2, no. 7, 2005, pp. 495-502, doi:10.1038/nmeth768.

9. Trapnell, Cole, et al. "Differential analysis of gene regulation at transcript resolution with RNA-seq." *Nature biotechnology*, vol. 31, no. 1, 2013, pp. 46-53.

10. Glaus, Peter, et al. "Identifying differentially expressed transcripts from RNA-seq data with biological variation." *Bioinformatics*, vol. 28, 2012, pp. 1721-1728.

11. Wang, Zhong, et al. "RNA-Seq: a revolutionary tool for transcriptomics." *Nature Review Genetics*, vol. 10, 2007, pp. 57-63.

12. Zubovic, Lorena, et al. "The altered transcriptome of pediatric myelodysplastic syndrome revealed by RNA sequencing." *Journal of Hematology & Oncology*, vol. 13, 2020, p. 135, DOI: 10.1186/s13045-020-00974-3.

13. Oshlack, Alicia, et al. “From RNA-seq reads to differential expression results.” *Genome Biology*, vol. 11, 2010, p. 220, DOI: 10.1186/gb-2010-11-12-220

14. Govindarajan, Meinusha, et al. “High-throughput approaches for precision medicine in high-grade serous ovarian cancer.” *ournal of Hematology & Oncology*, vol. 13, 2020, p. 134, DOI: 10.1186/s13045-020-00971-6.

15. Trapnell, Cole, et al. “Differential analysis of gene regulation at transcript resolution with RNA-seq.” *Nature Biotechnology*, vol. 31, no. 1, 2013, pp. 46-53, DOI: 10.1038/nbt.2450

16. Hong, Mingye, et al. “RNA sequencing: new technologies and applications in cancer research.” *Journal of Hematology & Oncology*, vol. 13, 2020, p. 166, DOI: 10.1186/s13045-020-01005-x.

17. Parkhomuchuk, Dmitri, et al. “Transcriptome analysis by strand-specific sequencing of complementary DNA.” *Nucleic Acids Research*, vol. 37, no. 18, 2009, p. e123, DOI: 10.1093/nar/gkp596.

18. Alharbi, Fadi, and Aleksandar Vakanski. “Machine Learning Methods for Cancer Classification Using Gene Expression Data: A Review.” *Bioengineering (Basel)*, vol. 10, no. 2, 2023, p. 173, DOI: 10.3390/bioengineering10020173.

19. Zhang, Yu-Huang, et al. “Identifying and analyzing different cancer subtypes using RNA-seq data of blood platelets.” *Oncotarget*, vol. 8, no. 50, 2017, pp. 87494-87511, DOI: 10.18632/oncotarget.20903.

20. “Globocan.” 2012, http://globocan.iarc.fr/Default.aspx.

21. Davis, Mary Elizabeth. “Glioblastoma: Overview of Disease and Treatment.” *Clinical Journal of Oncology Nursing*, vol. 20, 2016, pp. S2-8.

22. Thakkar, Jigisha P., et al. “Epidemiologic and molecular prognostic review of glioblastoma.” *Cancer Epidemiology, Biomarkers and Prevention*, vol. 23, 2014, pp. 1985-1996.

23. Benson, Al B., et al. “Hepatobiliary Cancers, Version 2.2021, NCCN Clinical Practice Guidelines in Oncology.” *Journal of the National Comprehensive Cancer Network*, vol. 7, 2009, pp. 350-91.

24. Lemjabbar-Alaoui, Hassan, et al. “Lung cancer: biology and treatment options.” *Biochimica et Biophysica Acta*, vol. 1856, 2015, pp. 189-210, doi:10.1016/j.bbcan.2015.08.002.

25. *Cancer Facts & Figures 2014*. Atlanta: American Cancer Society, 2014.

26. Park, Wungki, et al. “Pancreatic Cancer: A Review.” *JAMA (Journal of the American Medical Association)*, vol. 326, no. 9, 2021, pp. 851-862, DOI: 10.1001/jama.2021.13027.

27. Rawla, Prashanth, et al. “Epidemiology of Pancreatic Cancer: Global Trends, Etiology and Risk Factors.” *World Journal of Oncology*, vol. 10, no. 1, 2019, pp. 10-27, DOI: 10.14740/wjon1166.

28. Hong, Mingye, et al. “RNA sequencing: new technologies and applications in cancer research.” *Journal of Hematology & Oncology*, vol. 13, 2020, p. 166, DOI: 10.1186/s13045-020-01005-x.

29. Soverini, Simona, et al. “Next-generation sequencing for BCR-ABL1 kinase domain mutation testing in patients with chronic myeloid leukemia: a position paper.” *Journal of Hematology & Oncology*, vol. 12, no. 1, 2019, p. 131, DOI: 10.1186/s13045-019-0815-5.

30. Chatterjee, Aniruddha, et al. *A Guide for Designing and Analyzing RNA-Seq Data*. vol. 1783, Methods in Molecular Biology, 2018.

31. Conesa, Ana, et al. “A survey of best practices for RNA-seq data analysis.” *Genome Biology*, vol. 17, 2016, p. 13.

32. Kukurba, Kimberly R., and Stephen B. Montgomery. “RNA Sequencing and Analysis.” *Cold Spring Harbor Protocols*, vol. 2015, no. 11, 2015, pp. 951-969, DOI: 10.1101/pdb.top084970.

33. Alcalde, Fernado Garcia, et al. “Qualimap: evaluating next-generation sequencing alignment data.” *Bioinformatics*, vol. 28, no. 20, 2012, pp. 2678-2679, DOI: 10.1093/bioinformatics/bts503

34. Dobin, Alexander, et al. “STAR: ultrafast universal RNA-seq aligner.” *Bioinformatics*, vol. 29, no. 1, 2012, pp. 15-21, DOI: 10.1093/bioinformatics/bts635.

35. Chen, Shifu, et al. “fastp: an ultra-fast all-in-one FASTQ preprocessor.” *Bioinformatics*, vol. 34, no. 17, 2018, pp. i1884-i890, DOI: 10.1093/bioinformatics/bty560.

36. Bolger, Anthony M., et al. “Trimmomatic: a flexible trimmer for Illumina sequence data.” *Bioinformatics*, vol. 30, no. 15, 2014, pp. 2114-2120, DOI: 10.1093/bioinformatics/btu170.

37. Schmieder, Robert, and Robert Edwards. “Quality control and preprocessing of metagenomic datasets.” *Bioinformatics*, vol. 27, no. 6, 2011, pp. 863-864, DOI: 10.1093/bioinformatics/btr026.

38. Chen, Yuxin, et al. “SOAPnuke: a MapReduce acceleration-supported software for integrated quality control and preprocessing of high-throughput sequencing data.” *Gigascience*, vol. 7, no. 1, 2018, pp. 1-6, DOI: 10.1093/gigascience/gix120.

39. Love, Michael I., et al. “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.” *Genome Biology*, vol. 15, no. 12, 2014, p. 550, DOI: 10.1186/s13059-014-0550-8.

40. Hiller, David, and Wing Hung Wong. “Simultaneous Isoform Discovery and Quantification from RNA-Seq.” *Statistical Biosciences*, vol. 5, 2013, pp. 100-118, DOI: 10.1007/s12561-012-9069-2.

41. Conesa, Ana, et al. “A survey of best practices for RNA-seq data analysis.” *Genome Biology*, vol. 17, 2016, p. 13.

42. Li, Heng, et al. “The Sequence Alignment/Map format and SAMtools.” *Bioinformatics*, vol. 25, no. 16, 2009, pp. 2078-2079, DOI: 10.1093/bioinformatics/btp352.

43. Liao, Yang, et al. “featureCounts: an efficient general purpose program for assigning sequence reads to genomic features.” *Bioinformatics*, vol. 30, no. 7, 2014, pp. 923-930, DOI: 10.1093/bioinformatics/btt656.

44. Love, Michael I., et al. “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.” *Genome Biology*, vol. 15, no. 12, 2014, p. 550, DOI: 10.1186/s13059-014-0550-8.

45. Yaqoob, Abrar, et al. “Applications and Techniques of Machine Learning in Cancer Classification: A Systematic Review.” *Human-Centric Intelligent Systems*, 2023, DOI: 10.1007/s44230-023-00041-3.

46. Aziz, Rabia Musheer, et al. “Computer vision model with novel cuckoo search based deep learning approach for classification of fish image.” *Multimedia Tools and Applications*, vol. 82, 2023, pp. 3677-3696, DOI: 10.1007/s11042-022-13437-3.

47. Yaqoob, Abrar, et al. “A Review on Nature-Inspired Algorithms for Cancer Disease Prediction and Classification.” *Mathematics*, vol. 11, no. 5, p. 1081, DOI: 10.3390/math11051081.

48. Jakhar, Amit Kumar, et al. “SELF: a stacked-based ensemble learning framework for breast cancer classification.” *Evolutionary Intelligence*, 2023, DOI: 10.1007/s12065-023-00824-4.

49. Yaqoob, Abrar, et al. “Applications and Techniques of Machine Learning in Cancer Classification: A Systematic Review.” *Human-Centric Intelligent Systems*, 2023, DOI: 10.1007/s44230-023-00041-3.