NATIONAL TECHNICAL UNIVERSITY OF ATHENS
SCHOOL OF ELECTRICAL AND COMPUTER ENGINEERING
SCHOOL OF MECHANICAL ENGINEERING

INTERDISCIPLINARY POSTGRADUATE PROGRAMME
"Translational Engineering in Health and Medicine"

Differential Gene Expression Analysis of RNA-Seq Data: Comparing Lung Squamous Cell Carcinoma and Healthy Lung Cells in Homo sapiens

POSTGRADUATE DIPLOMA THESIS

Kotsa Christina

Supervisor

Dr. George Matsopoulos
Professor in School of Electrical and Computer Engineering
National Technical University of Athens

Co-Supervisor

Dr. Ioannis Makris
Scientific Associate in School of Electrical and Computer Engineering
National Technical University of Athens

Athens, June 2024

NATIONAL TECHNICAL UNIVERSITY OF ATHENS
SCHOOL OF ELECTRICAL AND COMPUTER ENGINEERING
SCHOOL OF MECHANICAL ENGINEERING

INTERDISCIPLINARY POSTGRADUATE PROGRAMME
“Translational Engineering in Health and Medicine”

Differential Gene Expression Analysis of RNA-Seq Data: Comparing Lung Squamous Cell Carcinoma and Healthy Lung Cells in Homo sapiens

POSTGRADUATE DIPLOMA THESIS

Kotsa Christina

Supervisor

Dr. George Matsopoulos
Professor in School of Electrical and Computer Engineering
National Technical University of Athens

Co-Supervisor

Dr. Ioannis Makris
Scientific Associate in School of Electrical and Computer Engineering
National Technical University of Athens

The postgraduate diploma thesis has been approved by the examination committee on 20th June 2024

1st member	2nd member	3rd member
Dr. George Matsopoulos	Dr. Athanasios Panagopoulos	Dr. Panayiotis Tsanakas
Prof. in NTUA	Prof. in NTUA	Prof. in NTUA

Athens, June 2024

....

Kotsa Christina
Graduate of the Interdisciplinary Postgraduate Programme, “Translational Engineering in Health and Medicine”,
Master of Science,
School of Electrical and Computer Engineering,
National Technical University of Athens

The ideas and conclusions presented in this paper are the author’s and do not necessarily reflect the official views of the National Technical University of Athens.

Abstract

Differential gene expression (DGE) analysis of RNA sequencing (RNA-seq) data has become a powerful tool in understanding the molecular mechanisms underlying various diseases. This thesis presents an in-depth examination of RNA-seq methodologies and their applications in the context of human disease, focusing on the DE analysis of RNA-Seq data derived from human lung squamous cell carcinoma and normal lung tissues.

The thesis begins with a brief but comprehensive introduction to the biology of RNA, describing the different types of RNAs and their functions in cellular processes. This is followed by a discussion of the basic concepts of human disease, particularly cancer, and how they relate to genetic and epigenetic factors, laying the foundation for understanding the importance of RNA analysis in disease research.

Subsequently, the revolutionary technology of Next-Generation Sequencing (NGS) is presented, focusing on RNA-Seq. All the critical steps of an RNA-seq experimental procedure are discussed, including RNA extraction, quality control, mRNA enrichment, fragmentation, library preparation, and sequencing. The introductory part of the thesis then focuses on explaining in detail the analysis of the acquired RNA-Seq data; which is the procedure we followed in this thesis.

In this RNA-seq analysis project, the initial steps involved selecting and downloading the dataset of interest from an online database. This process was followed by quality control checks to ensure the reliability of the dataset for the subsequent analyses. Based on the quality control, the data were then trimmed to improve the overall quality, which was checked again. The trimmed reads were then aligned to the human genome in order to gain insights into the level of gene expression in cancer and normal cells.

After counting aligned reads, the next essential step is to normalize them.This process accounts for factors that prevent direct comparison of expression reads, thereby making the data biologically meaningful. The normalized data underwent differential expression analysis to identify the up- and down- regulated genes in cancer cells. The results of the differentially expressed genes were then plotted using various graphs, revealing a variety of information about our dataset. The final step of this project was the enrichment analysis, which revealed significantly enriched molecular functions, biological processes, and pathways in this type of non-small cell lung cancer.

Our analysis revealed important differences in gene expression in tumor and normal cells. Through enrichment analysis we were able to identify several enriched functions and pathways commonly found in cancer according to the literature. Our findings contribute to a better understanding of squamous cell lung carcinoma and highlight potential biomarkers and therapeutic targets for this disease.

Keywords:
RNA Sequencing (RNA-Seq), Next Generation Sequencing (NGS), Lung Cancer, Differential Gene Expression Analysis, Cancer Genomics, Transcriptomics, High-Throughput Sequencing

Summary

Innovative reasearch in cancer prevention, diagnosis, and treatment is crucial, with genetics playing a pivotal role. Bioinformatics, combining molecular biology and computer science, is essential for analyzing genetic information, particularly through Next-Generation Sequencing (NGS) technologies. NGS revolutionized bioinformatics by enabling large-scale genomic and transcriptomic sequencing at lower costs compared to older sequencing methods. NGS workflow and analysis involve multiple steps including library generation, sequencing, quality checks, genome alignment, variant identification, annotation, and data interpretation. Comparing genomes and analyzing transcriptomes of cancerous and healthy cells are key applications of NGS in cancer research. One such application is RNA Sequencing (RNA-Seq), a technique that utilizes NGS, commonly to identify differentially expressed genes between these two conditions.

RNA-Seq, introduced around 2010, allows comprehensive profiling of transcriptomes, capturing both coding and non-coding RNAs. It is currently preferred over microarrays, which have been used since 2000 for mRNA quantification, as it offers greater dynamic range, the ability to detect unknown transcripts and provides insights on alternative splicing. RNA-Seq requires careful sampling, sophisticated preparation methods, and advanced analysis techniques for accurate gene expression profiling. Various tools and statistical frameworks are used to analyze RNA-Seq data, and several databases provide free access to genomic and transcriptomic information for research purposes.

This thesis focuses on uncovering genes that are differentially expressed between cancerous and healthy cells, as well as identifying enriched functions and pathways associated with lung squamous cell carcinoma (LUSC). LUSC is a type of non-small-cell lung cancer (NSCLC) that commonly occurs in the central part of the lungs or the main airways. It is the second most prevalent form of NSCLC, particularly in women. The primary risk factor for LUSC is smoking, though other factors like age, heredity, exposure to second-hand smoke, and occupational hazards also contribute. The subset of data we obtained from the GEO database comprised paired-end RNA-Seq data from both cancerous and normal samples collected from 8 patients during surgery. The RNA sequencing was conducted using Illumina technology.

After the data were downloaded, we performed basic quality control checks using FastQC and MultiQC, which generates a summary report of all FastQC files. The results revealed high levels of duplication in the RNA-Seq data, which is typical for such datasets. The "per-base sequence quality" was generally good, with a slight decrease at the ends of the sequence, a common observation in Illumina sequencers. GC content is roughly normally distributed for most samples, with some variation possibly due to contamination. Sequence length was consistent at 101 base pairs in all samples. In a significant number of samples there were warnings of over-represented sequences and adapter content, particularly the Illumina Universal adapter, suggesting the importance of careful handling. Despite these issues, the overall quality of the data was sufficient for subsequent analysis.

Once the quality control results were examined, we trimmed the reads to eliminate adapter sequences and low-quality bases. For this purpose we used Trimmomatic and adjusted parameters to suit our dataset's requirements. Following the trimming process of the paired-end data, the vast majority of paired-end reads survived, with adapter contamination reduced to less than 0.1% across all samples.

The next step in our analysis involved read alignment using the STAR Aligner. STAR was chosen for its ability to efficiently handle RNA-seq data, particularly in mapping reads that contain splice junctions. We utilized the Homo sapiens GRCh38.80 genome and corresponding GTF annotation file from ENSEMBL to build a comprehensive genome index optimized for alignment. Aligning reads to the genome produced BAM files, which were subjected to quality control checks using SAMtools. Visual analysis of alignment metrics revealed high percentages of uniquely mapped reads across tumor and normal tissues, indicating robust data quality suitable for downstream analyses. We then used featureCounts, a tool used post-read alignment to quantify mapped reads across genomic features. The BAM and genome annotation files were used as inputs, producing outputs that include detailed read counts per gene. There are several factors that prevent direct comparison of expression reads, including the length of the transcript, the sequencing depth, the expression levels of other genes in the sample and the expression levels of a particular gene. Thus, it is essential to normalize the read counts to reflect biological differences rather than technical artifacts. We employed DESeq2's default

method to normalize read counts, as this is the tool we used for the differential gene expression (DGE) analysis.

Before the differential gene expression analysis (DGE), we visually interpreted the normalized read counts. The majority of statistical methods used for multivariate data analysis do not work well with heteroscedastic data, such as RNA-Seq. Therefore, we transformed the normalized data using log2, regularized-algorithm (rlog) and variance-stabilizing (VST) transformations to stabilize variance and reduce outliers. Box plots reveal that rlog and VST transformations provide more uniform distributions compared to raw and log2-transformed data. Moreover, pairwise scatter plots show that rlog and VST reduce variance more effectively than log2. Standard deviation versus mean plots indicate that VST achieves the most consistent variance across gene expression levels. PCA on VST-transformed data reveals a relatively clear separation between tumor and normal samples, with normal samples clustering tightly and tumor samples showing greater variability. Hierarchical clustering supports these results, showing distinct groups for cancer and normal samples, with greater similarity within groups. These analyses confirm the expected gene expression patterns and highlight the effectiveness of VST transformation in managing the variance in heteroscedastic data.

For the subsequent DGE analysis, as we previously mentioned, we used DESeq2, a popular statistical tool for RNA-seq data analysis which is specifically designed to identify genes that are differentially expressed when analyzing different conditions. The results of DGE analysis show that there are several statistically significantly differently expressed genes in tumor samples. In general, gene expression between similar conditions is consistent, but differs significantly between different conditions, allowing us to clearly identify patterns of differential expression.

Finally, enrichment analysis was performed, using clusterProfiler, to further elucidate the functions of differentially expressed (DE) genes and identify patterns between them. Through enrichment analysis using GO terms, DE genes are categorized into various biological processes and molecular functions, providing a detailed picture of their roles in cancer. Additionally, by identifying significantly enriched pathways using KEGG MEDICUS, the analysis uncovered key biological mechanisms that may drive the observed changes in this type of cancer cells.

The enriched biological processes in cancer include rapid DNA replication and chromosome segregation, which are indicative of aggressive growth and proliferation. Cancer cells also exhibit significant changes in their cytoskeleton, enhancing motility and invasiveness. Processes similar to wound healing, such as ECM remodeling, are also enriched, facilitating tumor growth and spread.

In LUSC cells, the enriched molecular functions include DNA helicases and ATP-dependent activities, as well as enhanced initiation of DNA replication, which highlights the increased need for replication and repair in proliferation states. Alterations in actin binding, immune function and ECM remodeling further emphasize the complex interplay of structural, immune and signaling pathways that allow for tumor growth, metastasis and immune evasion.

The enriched pathways in lung cancer cells align with previous findings associated with molecular functions and biological processes. Some of the key pathways are DNA replication and repair mechanisms, such as pre-RC formation and the DNA replication licensing system. TRAIP-dependent replisome disassembly and kinetochore organization are crucial for maintaining genome integrity and proper chromosome segregation during cell division. Enrichment of the Fanconi anemia and homologous recombination pathways underlines the importance of DNA damage repair in cancer progression. In addition, pathways involving viral interactions, such as EBNA3C from Epstein-Barr virus and Tax from HTLV-1, appear to be enriched, underlining the role of viral proteins in carcinogenesis.

In summary, our analysis offers a comprehensive understanding of gene expression dynamics in LUSC compared to normal cells. These findings, consistent with existing literature, may aid in identifying potential biomarkers and developing targeted therapeutic strategies for this type of cancer.

Contents

1 Introduction

Lung cancer is the most common malignant neoplasm in the majority of countries and the main cause of cancer-related mortality globally in both sexes combined [97]. In 2012, around 1.8 million new cases of lung cancer were diagnosed, accounting for 12.9% of the world's total cancer incidence. The same year lung cancer mortality reached 1.59 million deaths accounting for 19.4% of all cancer deaths [47]. As of 2016, it was the leading cause of cancer death in men and the second leading cause of cancer death in women, trailing only breast cancer in frequency [81]. While tobacco smoking constitutes the most significant risk factor, accounting for 80 to 90% of lung cancer diagnoses, a myriad of other factors contribute to its etiology. Genetics, poor dietary habits, alcohol consumption, chronic inflammation due to infections, and occupational exposure to chemicals have emerged as influential contributors to lung cancer development. [140, 97]

Generally, lung cancers are classified into two main types: small-cell lung cancer (SCLC) and non small-cell lung cancer (NSCLC). SCLC metastasizes faster than SCLC and thus it is more dangerous. NSCLC accounts for about 85% of all lung cancers and is one of the leading causes of death worldwide [144]. The majority of patients with lung cancer are smokers; almost all of them in the case of SCLC and about 90% in the case of NSCLC [131]. For this analysis, we utilized an RNA-Seq dataset from patients diagnosed with squamous cell carcinoma of the lung. Squamous cell carcinoma of the lung (sometimes refereed to as LUSC) is the second of the most common NSCLC, especially in women. LUSC often occur in the central part of the lung or in the main airway [136]. The leading cause of LUSC is smoking. Other risk factors include age, hereditary, exposure to second-hand smoke, and occupational exposure to minerals, and metal particles [136].

In the last decade, research endeavors have aimed to explore and unveil differentially expressed genes (DEGs) between normal and cancerous lung tissues [91]. Recent RNA Sequencing (RNA-Seq) analyses of the transcriptome have shown promising results for the discovery of new biomarkers in various types of cancer [91, 185, 180, 61]. RNA-Seq is a powerful genomic tool that can describe the disease mechanism according to the gene expression process at different stages and levels of the disease [24, 96] and elucidate the relationship between clinical features and their biological changes for novel cancer treatment strategies. Through comparative analysis, RNA-Seq allows for the identification of genes that are either up-regulated or down-regulated under different conditions, shedding light on their functional significance in the genesis and progression of cancer.

1.1 Introduction to the biology of RNA and types of RNA

Ribonucleic acid or RNA, is a vital molecule found in most living organisms and viruses [165]. In 1958, Francis Crick formulated the fundamental theory of the central dogma of molecular biology, according to which, the genetic code, DNA, is transcribed to RNA, which is then translated into proteins (Figure 1). In the following years, it became clear that RNA exhibits remarkable diversity, changing the belief that it is simply a transcription of DNA that serves as an intermediate step in protein synthesis [[120]. RNA is crucial in guiding several intricate processes essential for cell function. Findings from human genome sequencing projects indicate that over 80% of the human genome is actively transcribed into RNA, however, less than 3% is responsible for encoding translated proteins [36]. This introductory chapter briefly discusses the basics of RNA biology, the different types of RNA, and their contribution to biological activities. The dynamic processes of transcription and translation are also reported, along with other regulatory mechanisms governing gene expression.

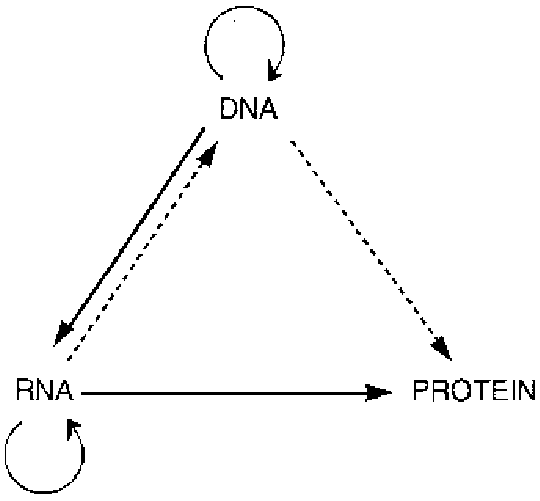


Figure 1: Crick's diagram indicating how he thought things stood in 1958 (reproduced in 1970). The arrows represent transfers of information. Solid arrows indicate "probable" transfers and broken arrows indicate "possible" transfers.Image retrieved from [57]

RNA is a type of nucleic acid, similar to deoxyribonucleic acid (DNA) in that it is composed of nucleotides. A nucleotide consists of a sugar molecule, ribose in the case of RNA, a phosphate group, and a nitrogenous base. The nitrogenous bases in RNA are adenine (A), cytosine (C), guanine (G), and uracil (U), while DNA uses thymine (T) instead of uracil. RNA molecules exhibit structural complexities, ranging from relatively unstructured forms to complex arrangements. These structures include secondary features, such as bulges and loops, and more complex tertiary configurations, which define their spatial arrangements in three dimensions. Secondary and tertiary structures are crucial for ensuring proper functionality and interaction.

A diverse array of RNA types has been identified in humans . An overarching classification (shown in Figure2) based on the functions of RNAs includes two main categories: coding and non-coding RNA. Coding RNA, named messenger RNA (mRNA), carries protein information from the genetic code to the protein-making machinery. RNA polymerase binds to the DNA, causing the double helix to unravel, facilitating the transcription of the genome. The result is a precursor mRNA (pre-mRNA), characterized by the presence of both introns, non-coding regions, and exons, coding sequences. Pre-mRNA undergoes a post-transcriptional refinement, involving the removal of introns and the precise arrangement of exons. Following a series of regulatory checks to ensure the precision of the genetic information, the mature mRNA is ready to be translated into proteins.

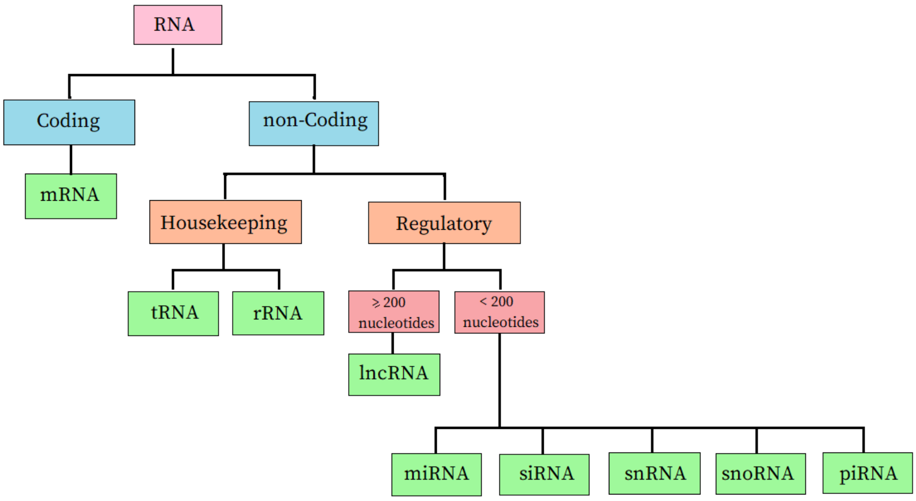


Figure 2: RNA categories classified according to reference [71] (other studies propose different ways of classification)

Non-coding RNAs (ncRNAs) can be further classified into 'housekeeping' and 'regulatory' ncRNAs based on their roles within the cell [71, 32]. Housekeeping ncRNAs are involved in fundamental cellular processes required for the basic maintenance and function of the cell. This category includes transfer RNA (tRNA) and ribosomal RNA (rRNA). tRNAs translate mRNAs into proteins and are the most abundant RNAs (in moles) among all cellular RNAs [118]. They have a distinctive cloverleaf pattern with various double-helical structures, encompassing a 3' acceptor site, 5' terminal phosphate, D arm, T arm, and anticodon arm. The anticodon arm of the tRNA holds the anticodon, which complements the mRNA codon—a triplet sequence of bases that codes for a specific amino acid. The 3' acceptor site transports the amino acid determined by the mRNA codon to a ribosome complex. rRNA molecules account for approximately 80-85% of the total RNA within a cell [120] and act mainly as structural and functional elements of ribosomes, the organelles where protein synthesis occurs. Regulatory ncRNAs are important in regulating gene expression and controlling various cellular processes and pathways. Regulatory ncRNAs are generally transcribed in a location- and time-dependent manner [171]. They can be further divided into two groups based on the number of nucleotides: lncRNAs (200 or more) and small ncRNAs (less than 200). lncRNAs have been found to play diverse and important roles in the regulation of various cellular processes, exerting significant influence on mechanisms such as DNA methylation, histone modification, and transcription [173]. Additionally, they contribute to altering the chromatin structure, affecting the accessibility of genes to the transcriptional machinery, and therefore regulating gene expression patterns [167]. The molecular mechanisms through which lncRNAs execute their functions can be summarized into four modes: signal, decoy, guide, and scaffold. Aberrant expression of lncRNAs is associated with several diseases, such as cancer, neurodegenerative disorders, rheumatoid arthritis, and blood-associated diseases [71, 32].

Small ncRNAs include micro RNAs (miRNAs), small interfering RNAs (siRNAs), small nucleolar RNAs (snoRNAs), small nuclear RNAs (snRNAs), and PIWI-interacting RNAs (piRNAs) [71]. These small molecules are profoundly involved in widespread and essential regulatory mechanisms in almost all eukaryotic cells. miRNA and siRNA both play crucial roles in the regulation of gene expression through a mechanism known as RNA interference (RNAi) [68]. miRNAs are one of the most abundant classes of gene regulatory molecules, which means that they have a major impact on the expression of numerous protein-coding genes [10]. miRNAs act as trans transcriptional regulators, i.e. they are encoded by one gene but act on other genes [171]. They are created by two cleavage reactions. First, the transcription of miRNA genes produces long hairpin-shaped primary transcripts known as pri-miRNAs, which are then fragmented by the Drosha complex to form miRNA precursors (pre-miRNAs). Dicer then further processes the pre-miRNAs to produce mature, functional miRNAs. Mature miRNAs are single-stranded non-coding RNAs, typically consisting of 19 to 22 nucleotides [71], which are loaded onto Argonaute proteins [108], forming RNA-induced silencing complexes

(RISC). This complex usually leads to translational repression or mRNA degradation if the degree of complementarity is higher [68, 171]. miRNAs can also induce transcriptional silencing of genes by inducing chromatin modifications [171]. Notably, miRNAs show promise as potential biomarkers and therapeutic tools for various cancer subtypes [179, 95].

siRNAs and miRNAs operate in comparable manners, share a similar production process, and bind with Argonaute proteins to form RISC complexes [68, 171]. However, the key distinction lies in their origin. siRNAs are mainly derived from the transcription products of the regions they regulate, and are therefore called cis regulators. [171]. This suggests that siRNAs are nearly perfect or perfectly complementary to their target mRNA sequences, inducing mRNA cleavage and rapid degradation. Due to their higher level of complementarity, siRNAs are generally more efficient at silencing specific genes than miRNAs. Thus, they have promising prospects in drug development as they can target a wide range of disease-related genes[68].

snRNAs are essential components of the spliceosome, a large and dynamic complex responsible for the precise removal of introns and joining of exons in pre-mRNA during gene expression. These snRNAs, including U1, U2, U4, U5, and U6, have specific roles at different stages of the splicing process [77]. The spliceosome is not only composed of snRNAs but also proteins (more than 70 [120]), forming small nuclear ribonucleoproteins (snRNPs). The spliceosome's dynamic nature involves structural changes throughout the splicing process. SnRNA is also responsible for maintaining the integrity of the telomeres [120].

snoRNAs are 60–300 nucleotides long that are found in the nucleolus in large quantities [44]. In humans, snoRNAs typically originate from intronic regions [12]. Following the splicing process, introns are removed as lariats, which are then degraded. However, snoRNAs avoid this degradation by forming protein complexes [12]. Their function is to modify rRNA, tRNA, and snRNA through some chemical modifications, such as methylation and pseudouridylation [44]. These modifications are essential for the proper functioning of ribosomes and spliceosomes.

piRNAs have a length of 24-30 nucleotides and are associated with PIWI Argonaute proteins forming RISC complexes [139, 186]. They are primarily expressed in the germline and germline-bordering somatic cells of animals, where they have a critical role in protecting the genomic integrity of the germ cells. In particular, piRNAs act as guides for the PIWI family of proteins to target and silence transposable elements, which are mobile genetic elements that can cause mutations and genomic instability. The piRNA pathway regulates gene expression both transcriptionally and post-transcriptionally, involving PIWI proteins that can cleave transposon transcripts [186]. Some PIWI proteins act as epigenetic regulators, influencing chromatin modifications and DNA methylation [6, 44]. Additionally, the piRNA pathway is important in antiviral defense, incorporating viral sequences into piRNA clusters to suppress their gene expression [186]. Beyond germ cells, however, PIWI proteins and piRNAs are found in other cell types. Their dysregulation is linked to diseases such as various types of cancer, making the piRNA pathway a potential biomarker and therapeutic target [178].

1.2 Human diseases: Basic concepts and effects on genetic material

Human diseases represent a complex range of conditions that lead to the impairment of physiological functions of the affected organism. These conditions can present themselves at different levels, from cellular to systemic dysfunction and damage. Understanding the underlying mechanisms of human disease is essential to developing medical research, diagnosis, and treatment.

Diseases can arise from various factors, including genetic, environmental, and lifestyle influences [119]. Individual susceptibility to almost all human diseases is influenced, to some extent, by genetic variation [26]. Some diseases are caused by the presence of abnormal genes inherited from parents, such as Cystic Fibrosis [98], Hemophilia A [50], Sickle Cell disease, β-Thalassemias [148], and Huntington's Disease [110]. However, the genome is under constant attack from both environmental factors and endogenous processes, often leading to DNA damage, including the introduction of new mutations or cleavages [119]. These alterations in the DNA sequence can have profound consequences for the cell and its progeny. Since the genetic information in our genes acts as a blueprint for protein synthesis and the regulation of biological processes, errors in these instructions can lead to either the absence of a protein or the production of a dysfunctional protein variant. The absence or malfunction of a specific protein can potentially culminate in the onset of diseases within the organism. Several studies have identified genetic variants in genes associated with diseases such as age-related macular degeneration [40], diabetes [143], and obesity [64]. Other conditions known to have a genetic component are neurodegenerative diseases such as Parkinson's and Alzheimer's [146, 156], as well as autoimmune diseases [72] and cancer.

The complex relationship between genetic factors and cancer development is evidenced by various mechanisms and observations. Genetic alterations including mutations, loss of heterozygosity, deletions, insertions, and aneuploidy have been associated with carcinogenesis [87]. It has been observed that in some families, certain types of cancer appear with great frequency and are known as hereditary or familial cancers. Most frequently, however, cancers arise without a clear hereditary link and are referred to as non-hereditary or sporadic cancers. The incidence of cancer increases with exposure to mutagenic agents, such as chemicals or ionizing radiation [134]. Additionally, some viruses have been identified as cancer-causing, suggesting that the expression of viral genes introduced by the host can disrupt normal cell cycle control [134]. Further confirming cancer as a genetic disease is the observation that all progeny of cancer cells also exhibit cancerous characteristics, leading to the formation of tumors [134].

Of the many hundreds of miRNAs identified in humans, over half are situated in genomic regions frequently disrupted in cancers. Consequently, these genes often undergo deletion or amplification, depending on the nature of the chromosomal rearrangement. Similar to the protein-coding genes associated with cancer, miRNAs are classified as either tumor suppressors or oncogenes, depending on their absence or presence in cancer respectively. For instance, the miR-17-92 oncogenic miRNA gene located in chromosome 13 experiences increased expression in various types of cancer, including lung cancer, compared to normal tissues [171].

Apart from genetics, epigenetic modifications modulate gene expression as well, impacting the function of the cell and contributing to disease pathogenesis [76]. The term 'epigenetic landscape' was coined by Conrad Waddington in the early 1940s [129] to describe the molecular mechanisms that convert genetic information into observable characteristics and phenotypes [42]. Monozygotic twins, whose genes are an exact match, can develop different phenotypes due to their experiences and exposures. Factors like diet, lifestyle choices, and encounters with environmental toxins leave distinct marks on their genetic instructions. Consequently, the once-identical twins develop different epigenetic profiles, affecting how their shared genes are activated or silenced. These subtle but crucial differences can impact their individual characteristics, from personality traits to health susceptibilities, highlighting the complex interplay between genetics and environment in shaping uniqueness [122]. In numerous cases, epigenetic patterns and their associated phenotypes persist through the processes of mitosis and meiosis [76]. This suggests that epigenetic gene expression can also be inherited in addition to being acquired.

Epigenetic marks such as DNA methylation, histone modifications, and nucleosome positioning are critical factors for regulating gene and non-coding RNA expression, as well as influencing chromatin states [76, 122]. DNA methylation involves adding a methyl group to the cytosine residue of a CpG dinucleotide (cytosine-phosphate-guanine sequence), which is often found in the promoter region of genes. This mechanism influences the accessibility of DNA to various regulatory proteins and transcription factors, leading to either gene transcriptional activation or repression [129]. Histone modifications, such as acetylation and methylation, can also affect gene expression by changing chromatin structure and accessibility [76]. Posttranslational modifications, such as phosphorylation and ubiquitination, can impact protein activity and stability, thereby inducing changes in cellular signaling pathways [76]. These epigenetic factors engage in intricate interactions, and the observed effects are the cumulative result of their interplay [122].

Lesions in the mechanisms described above can affect the phenotype by interfering with normal gene expression and leading to disease [45]. Epigenetic modifications have been implicated in a wide range of diseases, including cancer, neurological and autoimmune diseases, cardiovascular diseases, metabolic diseases, and myopathies [122]. Particularly in cancers, dysregulation of epigenetic mechanisms may lead to improper activation of oncogenes or, conversely, improper inactivation of tumor suppressors. In cancerous cells, DNA methylation patterns can be disrupted, leading to hypo- or hypermethylation of specific regions. Hypomethylation is a common feature of cancer, typically causing genomic instability and activation of oncogenes. In contrast, hypermethylation of the promoter region of tumor suppressor genes can lead to their silencing, which in turn can induce uncontrolled cell growth and cancer development [122]. Deregulations involving histone-modifying complexes and histone marks may be also a significant mechanism underlying the onset and progression of cancer. One of the most prominent alterations in histone modifications found in cancer cells includes the loss of monoacetylation and trimethylation of histone H4 [49]. The absence of these specific histone modifications in cancer cells can lead to abnormal gene expression patterns and disruption of chromatin organization.

Gene expression, influenced by both genetics and epigenetics, is imprinted in the transcriptome of each cell. The variations observed relate to the amount and the specific genes that are activated. Focusing on cancer, where aberrant gene expression plays a crucial role, it is imperative to exploit methodologies that can decipher the differential transcriptional landscape. The following chapters explore bioinformatics, revealing the sophisticated processes that enable the analysis of RNA Sequencing data. Studying how gene activity changes in healthy versus diseased cells is crucial, offering insights into what makes each disease unique and potential targets for treatment.

1.3 Introduction to NGS and its use for RNA-seq analysis

The prevalence of human diseases is a matter of global concern, affecting individuals, communities, and entire populations. Diseases not only impose a heavy burden on individual health but also have an impact on broader national socio-economic challenges. Their impact on global health extends to factors such as quality of life, life expectancy, and the sustainability of healthcare systems. Therefore, it is crucial to explore innovative approaches for disease prevention, diagnosis, and treatment.

Genetics, being the basis for many diseases, is a crucial area to investigate. While exploring genetic information is significantly challenging as the relevant data is concealed within vast volumes of biological sequences, bioinformatics serves as a fundamental tool in understanding the intricacies encoded in our genes. This interdisciplinary field combines molecular biology and computer science, and its applications lie in employing algorithms and databases for the analysis of genes, proteins, and even the whole genome.

Between 2004 and 2006, Next-Generation Sequencing (NGS) technologies emerged and revolutionized bioinformatics. NGS has since been implemented for the analysis of large-scale genomic and transcriptomic sequencing data at a lower cost compared to Sanger, the traditional first-generation sequencing method introduced in 1977 [100, 138]. A modern NGS platform can read hundreds of base pairs in each sequencing reaction and the number of reads produced by each platform ranges from millions to billions. This massive parallel reading is key to the success of this technique. Although NGS has been very time-consuming, the process has been significantly optimized, reducing the time required from hours to days [120]. Some of the most widely used NGS platforms are Illumina, SOLiD, Ion Torrent, Pacific Biosciences and Roche 454 [120].

The process followed by an NGS technology could be divided into the following steps [117]. Initially, libraries are typically generated from the genetic material of interest (e.g., genomic DNA, cDNA from RNA) and undergo various preparation steps to enable high-throughput sequencing. The generated library is then sequenced using some of the available NGS platforms mentioned above. The acquired data undergo quality checks, with filtering applied to remove low-quality areas. Thereafter, alignment to the reference genome is performed and variants such as point polymorphisms, small insertions or deletions, and structural variants are identified. Following this step, an annotation, i.e. a comparison with existing databases, is carried out. Finally, the data are visualized, evaluated, and interpreted.

NGS offers unparalleled insight into genetic landscapes. Firstly, it allows the sequencing of DNA in the genomes of organisms across the entire life spectrum. For each species of organism, a reference genome can be defined, i.e., a prototype representative of each organism. Re-sequencing the genome of individuals is simplified using NGS compared to the previously used techniques. For instance, it is possible to compare the genome sequence of certain individuals suffering from a disease with the reference genome sequence. This process makes it possible to identify genetic variations across the whole genome, among individuals of the same species. An additional application of NGS is the comparison of genetic alterations between different cell types in the same individual. By sequencing the genome of each cell it is feasible to identify somatic mutations in the case of cancer. When NGS technology is applied to RNA, through a technique called RNA Sequencing (RNA-Seq), it enables the identification and quantification of RNA transcripts. There are many other specialized applications of NGS technology, one of which is Chromatin Immunoprecipitation Sequencing (ChIP-Seq). This technique is implied to identify DNA sequences associated with transcriptional regulation and can also be used to identify methylated regions of DNA. In this project, we will focus our study on transcriptomic data. Over the years, various techniques have been employed to assess mRNA steady-state levels in biological samples. These techniques enable the observation of differential gene expression when comparing normal and diseased tissues. It should be noted that differential gene expression is also a result of the analysis of samples coming from different body tissues, developmental stages, or responses to environmental stimuli (such as drug exposure) [120].

Techniques utilized for the study of genome expression include Northern blotting, quantitative Reverse Transcription-Polymerase Chain Reaction (qRT-PCR), cDNA library analysis, microarrays, and RNA-Seq [120]. qRT-PCR and Northern blotting are low-throughput techniques. The procedures followed in these techniques are laborious and the information they give is scarce, meaning that few genes or one gene at a time are studied. Nevertheless, they are very reliable methods and can be used to confirm the results obtained from high throughput techniques [120]. The high-throughput techniques, that are now widely used, allow the study of gene expression on a large scale. The complete analysis of gene expression allows the identification of genes whose expression differs from one biological state to another.

In high-throughput techniques such as microarray and RNA-seq, total RNA or mRNA is first isolated from samples, usually corresponding to two biological conditions under comparison. Subsequently, RNA is converted to cDNA, using reverse transcription, because cDNA is inherently less sensitive to chemical and enzymatic degradation than RNA [120]. cDNA can be labeled and serve as a valuable tool for quantifying gene expression in microarray experiments. In addition, cDNA clone libraries can be generated for traditional Sanger sequencing, or more advanced NGS libraries can be constructed for analysis using the most powerful high-throughput technique, RNA-Seq. Figure 3 provides an overview of both microarray and RNA-Seq analyses.

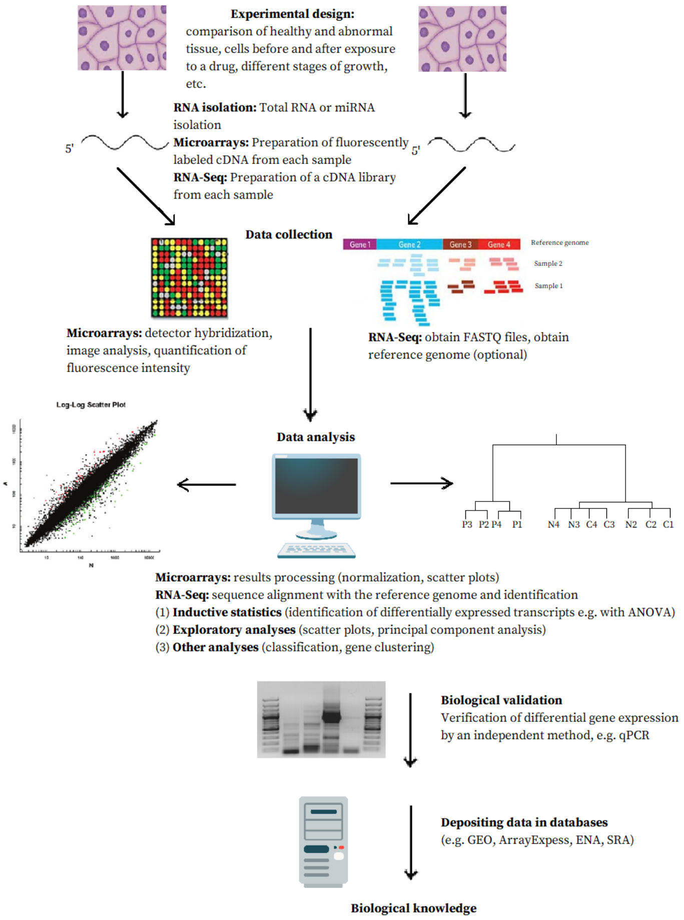

Experimental design:
comparison of healthy and abnormal tissue, cells before and after exposure to a drug, different stages of growth, etc.
RNA isolation: Total RNA or miRNA isolation
5'
Microarrays: Preparation of fluorescently labeled cDNA from each sample
5'
RNA-Seq: Preparation of a cDNA library from each sample
Data collection
Gene 1
Gene 2
Gene 3
Gene 4
Reference genome
Sample 2
Sample 1
Microarrays: detector hybridization, image analysis, quantification of fluorescence intensity
RNA-Seq: obtain FASTQ files, obtain reference genome (optional)
Log-Log Scatter Plot
Data analysis
P3 P2 P4 P1 N4 N3 C4 C3 N2 C2 C1
Microarrays: results processing (normalization, scatter plots)
RNA-Seq: sequence alignment with the reference genome and identification
(1) Inductive statistics (identification of differentially expressed transcripts e.g. with ANOVA)
(2) Exploratory analyses (scatter plots, principal component analysis)
(3) Other analyses (classification, gene clustering)
Biological validation
Verification of differential gene expression by an independent method, e.g. qPCR
Depositing data in databases
(e.g. GEO, ArrayExpress, ENA, SRA)
Biological knowledge

Figure 3: Overview of gene expression analysis experiments with microarrays and RNA-Seq. Image based on reference [120]

Since around 2000, DNA microarrays have been a powerful tool for mRNA quantification and have been used more than any other technique to compare gene expression between different samples [120]. a microarray consists of a solid substrate on which single-stranded DNA molecules, which are typically cDNAs, of known sequence are pinned to stacked microscopic positions. A microarray may contain a few hundred thousand DNA microscopic spots, called features, each of which may have a different sequence attached, in such a way that all the genes of the genome of the organism under study are represented. At the beginning of the experimental process, the extracted RNA from the samples is usually converted into fluorescently labeled cDNA. The microarray is then incubated and the cDNA is selectively hybridized with its complementary nucleic acids on the surface of the microarray. After removing the excess of the labeled molecules that were not hybridized, the microarray is scanned with a laser, and the fluorescent signal emitted from each position is recorded. This technique enables the simultaneous measurement of the expression levels of thousands of genes by analyzing the acquired image and quantifying the fluorescence intensity.

At the beginning of 2010, RNA-Seq was introduced, enabling the unbiased exploration of the abundance and the complexity of the gene expression [101, 109]. This innovative technique allows for the comprehensive profiling of the transcriptome, capturing the entirety of RNA molecules present in a biological sample, ranging from protein-coding mRNAs to non-coding RNAs [120]. The experimental design of RNA-Seq is based on that of NGS presented briefly earlier and will be discussed in detail in the next chapter, as it forms the basis for our analytical approach.

RNA-Seq is currently more widely used than the microarray technique, as the former has several advantages over the latter. While microarray allows measurements to be made only for transcripts whose sequences have been predetermined to be pinned to it, RNA-Seq is in principle able to measure the expression of any transcript, including transcripts that are not known in advance. RNA-Seq generally offers a greater dynamic range compared to microarrays [30], as it does not have an upper limit for quantification and can accurately detect a wide range of expression levels in the sample [56, 170]. In addition, in RNA-Seq experiments, the flexibility to adjust the depth of analysis is significant, as it enhances the ability to detect variants and rare transcripts. RNA-Seq also facilitates the characterization of alternative splicing patterns by identifying new isoforms and providing quantitative insights by measuring the expression of each specific exon [120, 170]. In addition to the study of mRNAs, this technique can, through appropriate manipulations, focus on the analysis of small, ncRNAs [120].

Nevertheless, RNA-Seq also presents certain disadvantages. The cost of this technique is not negligible, leading many scientists to limit the number of iterations and, consequently, the accuracy of the experiment. RNA-Seq is often accompanied by experimental errors that decrease its reliability. If, for example, the control and experimental samples are extracted on different days, differences may be observed due to varying RNA isolation conditions (this situation is described as a 'complete confound') [120]. Another concern is that in the case of protein-coding genes, the end product of expression is the proteins, so changes in mRNA levels may have little biological significance.

2 RNA Sequencing

In the field of transcriptomics, accurate imaging of gene expression profiling depends on meticulous sampling, sophisticated sample preparation methodologies, and advanced RNA-seq analysis techniques. This chapter focuses on the fundamental principles and intricate practices underlying these processes. Sampling and sample preparation techniques require very careful handling as they directly affect the quality of RNA-seq data. The chapter then reviews the diverse landscape of RNA-seq analysis methods, outlining some of the tools and statistical frameworks commonly used to decipher gene expression patterns. Finally, we list some databases that provide free access to genomic and transcriptomic data, which can be utilized for RNA-Seq analysis.

2.1 Sampling and sample preparation for RNA-seq

2.1.1 RNA extraction

The primary step of the procedure is the extraction of RNA from the biological samples of interest. To ensure a successful experiment, the quality of the isolated RNA should be of sufficient quality, and degradation and contamination should be minimal to produce a sequencing library. RNA-Seq can analyze the entire transcriptome, providing a comprehensive overview of all transcribed RNA molecules in a biological sample. Alternatively, it can be targeted to specific subsets, such as mRNA fractions for specific examination of gene expression levels or focusing on non-coding RNA to study the regulatory aspects of the transcriptome. Several methods and commercial kits are accessible for RNA-Seq applications, each tailored to specific research questions. For extracting total RNA from cells and tissues, the most commonly employed methods are the phenol-chloroform-based (e.g. TRIZol) and the silica-gel-based column procedures (e.g. Qiagen) [151].

2.1.2 Quality control

The input quality of the RNAs is the most important factor in RNA-Seq. The quality of the isolated RNA is typically calculated using an Agilent Bioanalyzer, which produces an RNA Integrity Number (RIN) [82]. The RIN is usually reported on a scale from 1 to 10, with higher values indicating better RNA integrity. These values are obtained through gel electrophoresis and analysis of rRNA peaks. The presence of distinct and sharp 18S and 28S bands is indicative of high-quality RNA. The ratio of the 28S to 18S peaks is also used as an additional measure of RNA integrity [82]. Low-quality RNA, with a RIN number lower than 6 [82], can affect the sequencing results leading to uneven gene coverage, and varying degrees of 3'–5' transcript bias. In cases where the sample is of low quality, for example when it is acquired from human autopsy samples, the effect of degraded RNA on sequencing results should be carefully considered as it potentially leads to erroneous biological conclusions [133].

2.1.3 mRNA Enrichment

One important aspect of the experimental design is the RNA-extraction protocols used to remove the rRNA and tRNA molecules, which are highly abundant in the cell [27, 164]. These RNAs should be removed before library construction, otherwise, they will consume most of the sequencing reads, reducing the overall depth of sequence coverage and therefore limiting the detection of less abundant RNAs that contain useful information [82]. There are specialized kits available designed to selectively remove tRNAs and small RNAs from the sample [164]. Many protocols focus on enriching mRNA molecules by selecting polyadenylated (poly-A) RNAs, as in most eukaryotic organisms, most mRNAs and many lncRNAs contain a poly-A tail. This is achieved by targeting the 3' poly-A tails of the RNA molecules using substrates (e.g. magnetic beads or cellulose beads) covered with covalently attached poly-T oligo molecules [82]. Another treatment that can be used to isolate mRNAs, is called rRNA depletion. Such methods selectively remove the rRNA sequences from the total RNA population. Both the above options can be used for eukaryotic samples, whereas prokaryotic samples must undergo rRNA depletion as their mRNAs are not polyadenylated. Additionally to the library preparation protocols mentioned above, protocols have been developed to selectively target small RNAs, which are key regulators of gene expression. Due to the low concentration of small RNAs, short length, and lack of poly-A tails, more complex methods are preferred to sequence these RNA species [107]. Specifically, small RNAs can be extracted from the RNA pool with procedures such as by size separation [82, 11] or by using commercially available kits specifically designed for small RNA extraction, such as miRNeasy or MagnaZol [11]. It is also essential to remove contaminating genomic DNA from RNA samples after RNA enrichment, using for instance DNase I treatment [164].

2.1.4 Fragmentation

Fragmentation is a necessary step during library preparation because of the size limitation of most currently used sequencing platforms (e.g. <600 bp on Illumina sequencers) [67]. The fragmentation step can happen at different stages depending on the library preparation protocol and the sequencing platform. Usually, after poly-A priming or rRNA depletion, RNA molecules are subject to RNA fragmentation to a specific size range before reverse transcription. Uniformity in size ensures that all the fragments have an equal likelihood of being sequenced. RNAs can be fragmented with alkaline solutions, solutions with divalent cations, or enzymes, such as RNase III. If the fragmentation is not uniform, certain regions of the transcriptome may be overrepresented or underrepresented in the final sequencing data. Alternatively, intact RNA can be reverse transcribed and then, the resulting full-length cDNA can be fragmented, both mechanically and chemically. A common method to fragment cDNA utilizes acoustic shearing. Furthermore, full-length double-stranded cDNAs can be fragmented using the enzyme DNase. The development in using a transposon-based, so-called tagmentation method, facilitated the cDNA fragmentation while adding adapter sequences at the same time [107]. However, it is notable that these enzyme-based cDNA fragmentation methods require a precise enzyme:DNA ratio, making method optimization more complicated than RNA fragmentation. Thus, fragmenting RNA is most often used in RNA-Seq library preparation [67].

2.1.5 Library Preparation and Sequencing

Most RNA-Seq experiments are performed on instruments that sequence DNA molecules due to the technical maturity of commercial instruments designed for DNA-based sequencing. Therefore, although direct RNA sequencing is possible [116], which aids in mitigating certain biases [52], usually cDNA libraries are built. The preparation of a cDNA library varies based on the specific RNA species under examination, which can differ in terms of size, sequence, structural features, and abundance [67]. Furthermore, the subsequent steps in cDNA library preparation can differ among protocols, however, all common RNA-Seq protocols involve RT [162]. One notable characteristic of reverse transcriptases is their tendency to produce spurious second-strand cDNA due to their DNA-dependent DNA polymerase activity [150]. This can introduce complications in distinguishing sense from antisense transcripts, posing challenges when strand-specificity is crucial.

In the standard Illumina protocol, double-stranded cDNA synthesis initiation relies on random primers [162]. The use of random primers is a good approach, ensuring a more uniform sequence representation across transcripts [174]. Contrary to its name, random primer hybridization is not truly random but instead, exhibits a preference for specific nucleotide compositions. While simple, this approach does not retain the information of the clones, meaning that the information about which DNA strand corresponds to the sense strand of RNA. This information is important for studies that attempt to identify antisense and novel RNA species [67]. Thus, different methods have been developed for strand-specific RNA seq analysis, which fall into two main classes [14].

The first class involves attaching adapters in a known orientation relative to the 5' and 3' ends of the original mRNA. The ligation products are subsequently reverse-transcribed and amplified by PCR. This method begins with the removal of the 3' phosphate group from fragmented RNA and the addition of a 5' phosphate group. Subsequently, there are consecutive ligations involving a 5' adenylated 3' adapter through a truncated RNA ligase II and a 5' adapter ligation using RNA ligase I [67]. The distinction in sequence between the 5' and 3' adapters maintains the information about the RNA strand. A notable example within this category is the Illumina RNA ligation method, which was originally designed for small RNA-Seq [58]. Although this approach is straightforward, it has significant biases due to the influence of both 5' and 3' end sequences in the ligation steps. This issue, however, can be mitigated by the use of random nucleotides at the binding end of each adapter [74].

The second class, the dUTP protocol is similar to the original Illumina protocol, however, during the second strand cDNA synthesis, dUTP is incorporated. The incorporation of dUTP in the second cDNA strand serves as a chemical mark and facilitates the degradation of this second strand with DNA uracil glycosylase (UDG) before PCR amplification [162]. Furthermore, the U-containing chain is a very poor matrix for thermostable polymerases making it unamplified. Therefore, only the first chain of cDNA is amplified with defined adapter sequences, imparting directional information to the sequencing reads. A systematic evaluation of various protocols for strand-specific RNA-Seq revealed that the dUTP-based method exhibited the highest effectiveness in achieving evenness in coverage [88]. Currently, this method is widely employed in commercial directional RNA-Seq library preparation protocols. However, this method is generally unsuitable for low-input samples due to its demand for additional enzymatic and purification steps which can result in material loss. Efforts persist in the ongoing search for solutions to tackle this challenge [85, 104, 65].

Subsequent steps of most protocols after the preparation of cDNA are usually end-repair, A-tailing, and adapter ligation. The repair-end step addresses the repair of the 3' and 5' overhangs in the cDNA, ultimately yielding fragments with uniform blunt ends. A-tailing is a reaction that adds an 'A' base to the 3' end of the blunt-ended phosphorylated cDNA, thereby facilitating subsequent ligation adapters that carry a 'T' base overhang at their 3' end [84]. Ligation of adapters to the ends of the cDNA fragments prepares them for attachment to the flow cells used in cluster generation for sequencing and the subsequent amplification [174].

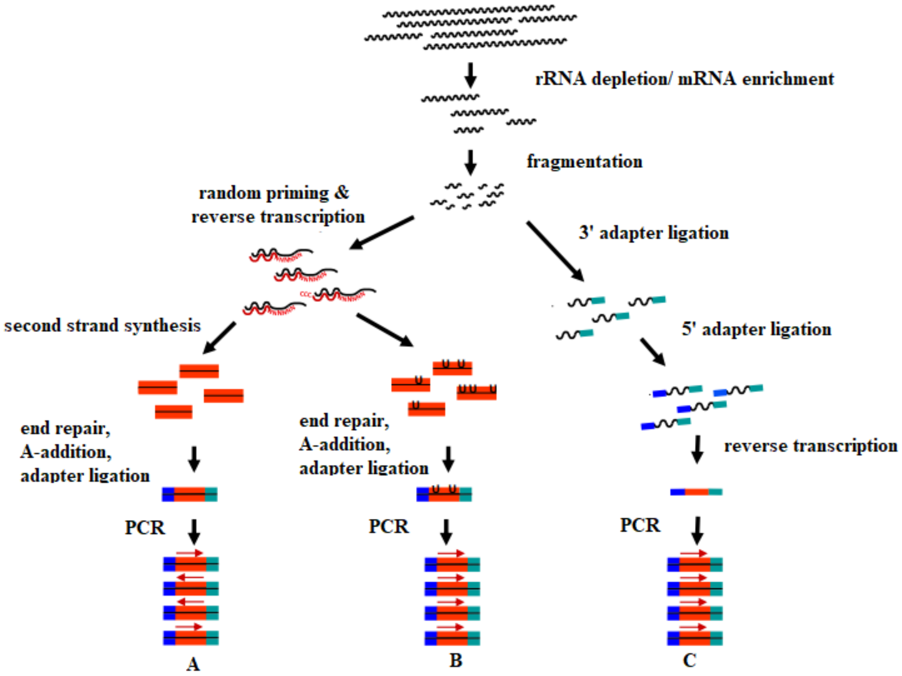


Figure 4: Three most common RNA-Seq protocols. (A) Classical Illumina protocol: Random-primed double-stranded cDNA synthesis, followed by adapter ligation and PCR. (B) First class of strand-specific method: relies on marking one strand by chemical modification. dUTP is incorporated during the second cDNA strand synthesis, preventing amplification of the second strand by PCR. (C) Second class of strand-specific method: relies on attaching different adapters to the 5' and 3' ends of the RNA transcript. Image retrieved from reference [162]

cDNA libraries are often amplified by PCR before sequencing due to the detection limit of the sequencers [67]. In the Illumina platforms, specifically, the amplification process commonly involves cluster generation chemistry [84]. During this process, the fragments from the cDNA library are loaded onto a flow cell, which consists of a glass slide with physically separated lanes coated with a lawn of surface-bound oligos complementary to the library adapters. Each single-stranded adapter ligator fragment is then hybridized to the flow cell surface and acts as a template for synthesizing a complementary strand. This process involves repeated denaturation and extension cycles, leading to the formation of clonal clusters through bridge amplification.

Sequencing-by-synthesis (SBS), a widely utilized NGS technology, employs four fluorescently labeled nucleotides for the parallel sequencing of extensive clusters [181]. In each sequencing cycle, a single labeled deoxynucleotide triphosphate (dNTP) is added to the nucleic acid chain, which acts as a "reversible terminator" for polymerization. Following the incorporation of dNTP, laser excitation identifies the fluorescent dye, revealing a distinct color corresponding to the incorporated nucleotide. After imaging, the fluorophore is cleaved, and the terminator is reversed, enabling the incorporation of the next base in the sequence. This iterative process continues until the determined read length is achieved. Each cluster on the flow cell yields a singular sequencing read [21]. This process is repeated as many times as necessary until the predefined reading length is reached. Reads range from 30 to 400 bp depending on the NGS technology employed [166].

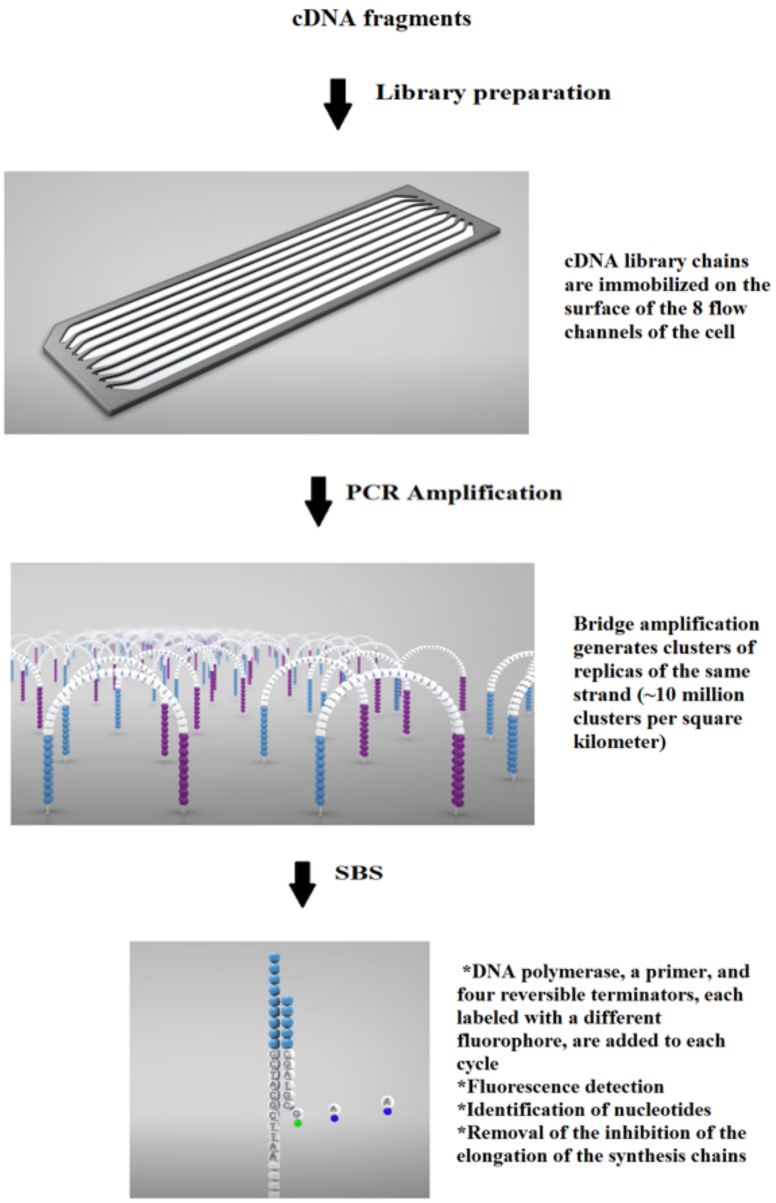

cDNA fragments
Library preparation
cDNA library chains are immobilized on the surface of the 8 flow channels of the cell
PCR Amplification
Bridge amplification generates clusters of replicas of the same strand (~10 million clusters per square kilometer)
SBS
*DNA polymerase, a primer, and four reversible terminators, each labeled with a different fluorophore, are added to each cycle
*Fluorescence detection
*Identification of nucleotides
*Removal of the inhibition of the elongation of the synthesis chains

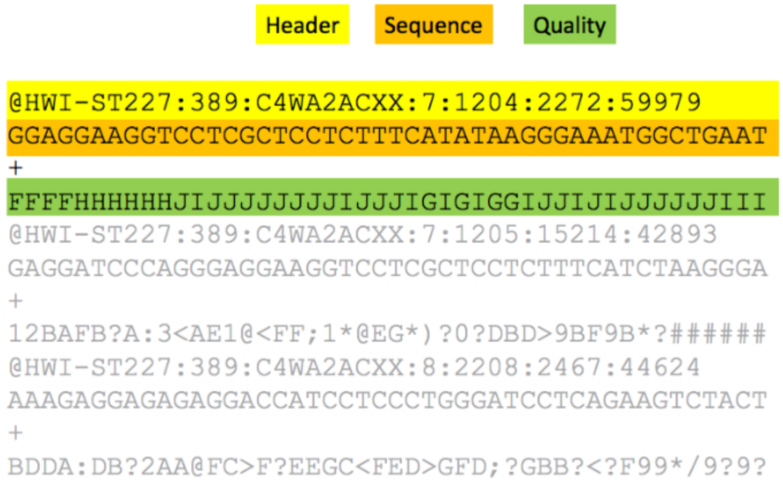
Figure 5: Sequencing with Illumina technology. Based on reference [120]. The three images are retrieved from:[70]

2.2 RNA-seq data analysis

Every cDNA fragment, regardless of whether it has undergone PCR amplification, is sequenced in a high-throughput manner, generating millions of short sequences. Depending on the project's objectives, the sequencing approach can be either single-end, involving the synthesis of only one strand, or pair-end, where both strands of the fragments are read. If the study aims to investigate changes in gene expression, single-end sequencing is sufficient, while pair-end sequencing is more useful for whole-genome sequencing, alternative splicing, and de novo transcriptome studies. After sequencing, the produced reads are aligned to a reference genome and subsequently are assembled based on annotated reference transcripts. Additionally, de novo assembly approaches allow the generated reads to construct a transcriptome map on a genome scale, even without relying on the genomic sequence. RNA-Seq downstream analyses encompass identifying differential gene expression across distinct samples, controlling allele-specific expression, as well as pinpointing genes exhibiting alternative splicing or novel transcripts [166]. In this work, we will particularly focus on differential gene expression between healthy and cancerous lung cells.

2.2.1 FASTQ files

When cDNA is sequenced, primary image files are created, which are used to determine which nucleotide is identified at each position read. The sequencing machine itself then uses these image files to derive the sequences of the different cDNA fragments and calculate the probability of error. These two pieces of information, the sequence, and the error probability, are stored in FASTQ-format files. Typically, the raw data used for further analysis are the FASTQ files rather than the original images on which they are based. FASTQ files are uncompressed and therefore quite large, as they contain the following four rows of information for each sequence read (Figure 6) [120].


Header
Sequence
Quality

```
@HWI-ST227:389:C4WA2ACXX:7:1204:2272:59979
GGAGGAAGGTCCTCGCTCCTCTTTCATATAAGGGAAATGGCTGAAT
+
FFFFHHHHHJIJJJJJJJIJJJIGIGIGGIJJIJIJJJJJJIII
@HWI-ST227:389:C4WA2ACXX:7:1205:15214:42893
GAGGATCCCAGGGAGGAAGGTCCTCGCTCCTCTTTCATCTAAGGGA
+
12BAFB?A:3<AE1@<FF;1*@EG*)?0?DBD>9BF9B*?######
@HWI-ST227:389:C4WA2ACXX:8:2208:2467:44624
AAAGAGGAGAGAGGACCATCCTCCCTGGGATCCTCAGAAGTCTACT
+
BDDA:DB?2AA@FC>F?EEGC<FED>GFD;?GBB?<?F99*/9?9?
```

Figure 6: Example of FASTQ file format. The differently highlighted rows contain the following information. (1) The '@' sign precedes the read ID and possibly information about the sequencing run. (2) Sequenced bases appear (3) The '+' sign (in other cases, the '+' sign is followed by the read ID again or some other description). (4) Quality scores for each base of the sequence (encoded in ASCII). Image retrieved from [5]

The reason why FASTQ files store the cDNA sequence of each read along with a position-specific quality score that represents the error probability, is due to the inevitable uncertainty that is inherited from the imperfect nature of the sequencing process and the limitations of the instruments used. The quality score is called Phred Q and is proportional to the probability p that a base call is incorrect. The Phred score is given by the following equation: is $Q = 10 \log_{10}(p)$

There are different assignments for base quality, and it is essential to identify the specific version of the Phred score being used in each analysis. Various types of bias can be introduced during sequencing. For example, the number of biases increases with the length of the reads. In particular, with Illumina technology, as the cycles increase, distinguishing between signal and noise during nucleotide integration becomes increasingly difficult. Furthermore, it has been observed that the number of errors is associated with GC nucleotide content [120]. Many applications are employed for evaluating the quality of sequencing data, such as FASTQC and ShortRead. These tools facilitate data filtering, i.e. the removal of regions with low quality. Quality checks should be carried out at each step of the analysis, to assist in making the subsequent analyses with appropriate assumptions and parameters.

2.2.2 Trimming reads

Trimming of RNA-seq data is a crucial pre-processing step which aims to improve the overall quality and reliability of subsequent analyses. It usually includes the deletion of sequencing adaptors, low-quality bases and short reads, if any, that could alter the results or interfere with accurate interpretation. Before truncation, it is mandatory to thoroughly examine the quality of the raw data to identify problems such as adapter contamination, poor read quality and different length. Understanding the problems in the dataset facilitates the selection of modules and parameters of the trimming process to result in useful information.

Several tools can be used for this process, such as Trimmomatic [13], Cutadapt [99], or Fastp [22]. Trimmomatic is a robust pre-processing tool with pair awareness, optimized specifically for Illumina NGS data. It includes a variety of processing steps for trimming and filtering reads. However, its main algorithmic innovations are related to adapter sequence detection and quality filtering. The result produced by Trimmomatic is competitive and potentially superior to alternative tools in the field. Cutadapt is a tool that focuses on identifying and removing adapters, with fault tolerance, from each read. It offers a user-friendly interface that supports 454, Illumina and SOLiD (color space) data and provides two adapter cutoff algorithms. astp is an extremely fast FASTQ preprocessor that provides useful quality control and data filtering. It can perform quality control, adapter trimming, quality filtering, per-read quality trimming, and several other operations with a single scan of the FASTQ data. Based on reviews, fastp is 2-5 times faster than other FASTQ preprocessing tools, including the two mentioned above.

Once trimming is complete, it is important to reassess the quality of the data to ensure that the trimming process has effectively improved their overall quality.

2.2.3 Mapping reads

In order to identify the transcript from each sample, the genomic origin of the sequenced cDNA fragments has to be established. Sequencing the reads to the most likely locus, called mapping, is a crucial step in the majority of hight-throughput sequencing experiments. The procedure of mapping involves finding where a short nucleotide sequence has the best match along the very long sequence of the transcribed genome. Given that this involves aligning individual letters of usually two string data, these approaches are commonly known as read alignment. Numerous factors come into play when aligning the huge number of readings acquired. These considerations include dealing with mismatches and gaps, determining the necessity of global alignment, and determining the threshold where two strings cannot be reasonably aligned. Thus, mapping millions of reads accurately and within a reasonable time is challenging for short-read alignment. Different alignment programs use various strategies to accelerate the process and achieve a good balance between mapping accuracy and fault tolerance. The majority of such programs contain algorithms that follow the "seed-and-extend" approach [39]. In this approach, a subset of the total reads is selected as a "seed" for which the tool finds the best match in an index made from the reference genome. Then, every matched seed is extended to both directions until as much of the read is aligned as possible. This procedure is performed under certain constraints, such as a predefined maximum number of mismatches specifically set for each experiment. Moreover, for the local alignment step, the programs usually employ the Smith-Waterman algorithm [147].

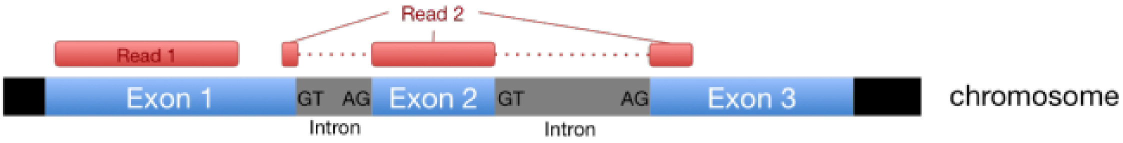


Figure 7: The genome contains both introns and exons within its gene sequences. Introns are removed during RNA processing, while exons are retained in the mature mRNA. Consequently, the major computational challenge in aligning RNA-Seq reads to the reference geneome is handling of spliced junctions, where one part of the read maps to the end of one exon and the rest of the read maps to another exon (which can possibly be located thousands of base pairs away from the first exon). Image retrieved from: [39]

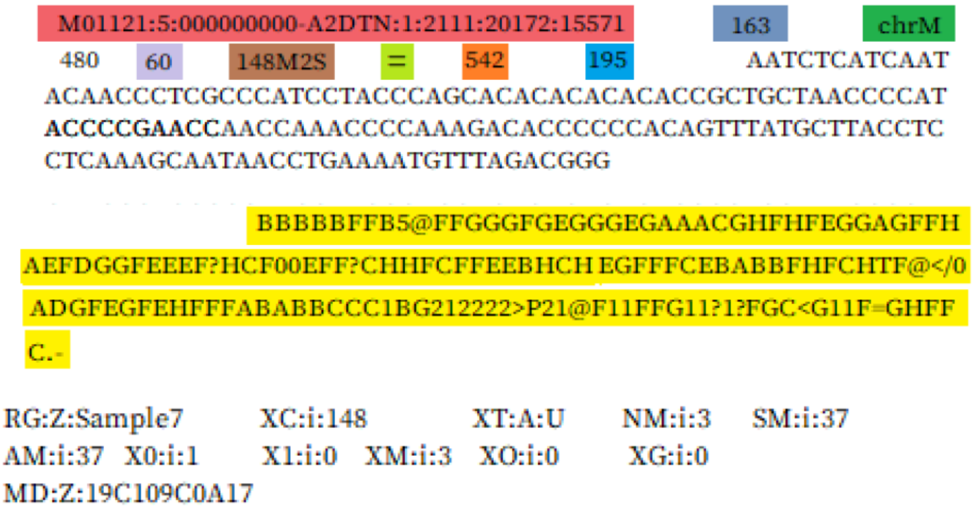
In contrast to genome sequencing, the most challenging aspects of RNA-Seq are the spliced alignment of exon-exon-spanning reads and the presence of different isoforms of the same gene. These RNA-seq reads span across two or more exons, reflecting the original arrangement of exons in the gene before splicing (i.e. intron sequences are included). Popular RNA-seq alignment programs, such as STAR [37], TopHat [160], and GSNAP [177], follow the process of aligning RNA-Seq reads to the entire reference genome and exploit existing gene annotation as a guide to where large gaps should be expected. In case, the genes cannot be aligned despite acknowledging the known introns, most tools will try to identify alternative splicing events. However, considering that the tools search for the most parsimonious combinations of exons, the assumptions made for the splicing may not reflect the actual biological driving force behind the creation of isoforms. Another issue arises from the fact that lowly expressed isoforms may have limited reads and that splice junctions with very few supporting reads are more likely to be unreliable. Thus, novel splice junctions are more likely to be detected in strongly expressed genes because there is a higher chance of obtaining reads that span these junctions, resulting in a bias towards those genes. As technology evolves and larger reads become more commonplace, it is expected that these challenges will be mitigated, improving the accuracy and reliability of alignment of spliced reads in RNA-seq data analysis [39].

Genome sequences and annotation are often produced by consortia such as (mod)ENCODE [39]. Someone can access the data generated by (mod)ENCODE through dedicated websites or pan-species databases, such as one hosted by the University of California, Santa Cruz (UCSC) [16], or the European genome resource, ENSEML [41]. UCSC and ENSEMBL are platforms that organize and provide access to genome and annotation data. Reference sequences, which are essentially extensive strings, are usually stored in FASTA files. Annotation files, containing information about the positions of transcription sites, introns, exons, etc., are typically formatted as GFF (General Feature Format) and GTF (Gene Transfer Format). Annotating Notably, GTF is an extension of GFF but is more strictly defined. In the course of the entire analysis, obtaining a properly formatted GFF/GTF file is a pivotal undertaking. Neglecting this step poses the risk of introducing potentially impactful issues that might go unnoticed and lead to unusual results solely due to formatting issues [39].

The output of alignment programs, specifically from the STAR read alignment, is a file format called SAM (Sequence Alignment/Map) and BAM (Binary Alignment/Map). SAM format files contain a general form of nucleotide alignment that describes the alignment of sequence reads (query sequences) with a reference. In a SAM file for each of the millions of reads, there are 12 mandatory fields [120] containing information about the alignment, including the sequence followed by the corresponding quality scores for each base, as well as information about the alignment of the read to the reference genome. The format of a SAM file is shown in Figure 8 and the information provided is analytically listed in Table 1. The readable SAM files can be compressed into the binary BAM format. BAM files with positional classification can be indexed so that all reads aligned to a position can be efficiently retrieved without having to load the entire file into memory.

Field Name	**Example Entry**	**Description**
QNAME	Read1	Query template (=read) name (PE: read pair name)
FLAG	83	Information about the read's mapping properties encoded as bit-wise flags
RNAME	chrI	Reference sequence name where the alignment is mapped. This should match a @SQ line in the header
POS	15364	1-based left-most mapping position of the first matching base. Set as 0 for an unmapped read without coordinates
MAPQ	30	Mapping quality of the alignment (Phred-scaled)
CIGAR	51M	Extended CIGAR string. Describes the position of insertions/deletions/matches in the alignment, and encodes splice junctions, for example
RNEXT	=	PE reads: reference sequence name of the next read. Set to "=" if both mates are mapped to the same chromosome
PNEXT	15535	PE reads: leftmost mapping position of the next read
TLEN	232	PE reads: inferred template length (fragment size)
SEQ	CCA..GGC	The sequence of the aligned read on the forward strand
QUAL	BBH. . . 1+B	Base quality (same as the quality string in the FASTQ format, but always in Sanger format [ASCII+33])
OPT	NM:i:0	Optional fields (format: <TAG>:<TYPE>:<VALUE>)

Table 1: Mandatory fields in the SAM format (there may be other optional fields)



```
M01121:5:000000000-A2DTN:1:2111:20172:15571        163        chrM
480    60    148M2S    =    542    195        AATCTCATCAAT
ACAACCCTCGCCCATCCTACCCAGCACACACACACACCGCTGCTAACCCCAT
ACCCCGAACCAACCAAACCCCAAAGACACCCCCCACAGTTTATGCTTACCTC
CTCAAAGCAATAACCTGAAAATGTTTAGACGGG

                BBBBBFFB5@FFGGGFGEGGGEGAAACGHFHFEGGAGFFH
AEFDGGFEEEF?HCF00EFF?CHHFCFFEEBHCH EGFFFCEBABBFHFCHTF@</0
ADGFEGFEHFFFABABBCCC1BG212222>P21@F11FFG11?1?FGC<G11F=GHFF
C.-

RG:Z:Sample7       XC:i:148          XT:A:U      NM:i:3     SM:i:37
AM:i:37   X0:i:1   X1:i:0   XM:i:3   XO:i:0      XG:i:0
MD:Z:19C109C0A17
```

Figure 8: Example of SAM file format. (1) The name of the read (M0112...) (2) The value of FLAG is 163 (3) The name of the reference sequence is chrM (it is a mitochondrial genome) (4) Position 480 is the leftmost position of this read (5) The mapping quality on the Phred scale is 60 (error frequency 10^{-6}) (6) The CIGAR line (148M2S) shows that 148 bases are aligned, while 2 bases at the ends of the sequence are not aligned (7) The '=' symbol indicates that the name of the reference sequence is the same for both paired-end reads (8) Position 524 is the leftmost paired reading position (9) Insert size is 195 bases (10) The sequence begins with AATCT and ends with ACGG (length 150 bases) (11) Each base has a quality rating (12) This reading has additional optional fields that accompany the MiSeq analysis. Image based on reference [120]

2.2.4 Quality Control of Aligned Reads

The properties of aligned reads should be assessed before downstream analyses. An alignment of RNA-seq reads is typically considered to be successful if the mapping rate is greater than 70% [39]. Basic alignment assessments, such as the total number of mapped reads and the number of uniquely mapped reads, are included in the logs that are produced by the mapping programs. However, there are useful tools (SAMtools, RSeQC, QoRTs) that can provide more information about the number and types of reads stored in a BAM file. The visual representation of aligned reads provides valuable insights into the accuracy of the alignment process, i.e. the amount of reads aligned to the expected regions and the amount of mismatches.

Based on statistical information of the quality control, one can draw conclusions regarding certain common biases observed in RNA-Seq [39]. In mRNA-Seq, the majority of the aligned reads are expected to overlap with exons. However, if a substantial number of reads align to introns, this may suggest incomplete poly(A) enrichment. Furthermore, if there is an abundance of reads aligned outside of annotated gene sequences, this may be indicative of either genomic DNA contamination or the presence of abundant non-coding transcripts. Additionally, biases at the 3' or 5' ends, signifying over-representation of these specific transcript portions, may suggest RNA degradation.

2.2.5 Quantification

The next phase of the RNA-Seq workflow involves the quantification of transcripts, a step of particular interest, especially for studying differential gene expression. Overall, there are two different ways to quantify the expression level of individual transcripts [161]. The first way is to directly count reads overlapping with gene loci, after the alignment step. Counting reads overlapping with the genome is, in principle, a straightforward task. However, the complexity lies in considering specific details in the quantification tools (such as HTSeq-count and featureCounts) relevant to each case [39]. To begin with, the overlap size, which determines the distinction between full reads and partial overlap with genomic features, differs between tools. Secondly, there is variability in how the tools handle multiple-match reads, i.e., reads that align to multiple positions in the genomic sequence. Thirdly, when reads overlap with multiple instances of the same genomic feature, such as different isoforms of a gene, programs can either allocate reads to all features or apply specific criteria for assignment. Finally, it is important to consider how the program handles the reads that overlap introns, to ensure efficient analysis of gene expression and splicing patterns. The second way of quantifying the obtained reads is by attempting to determine the number of individual transcripts, avoiding the tedious alignment step. This approach yields a probabilistic measure of how many reads indicate the presence of a transcript. Despite the significant increase in the speed of analysis, this method is less efficient than the former, as it is prone to spurious alignment, especially for the limited and small transcripts, as well as transcripts where the splice variants are quite similar to each other.

2.2.6 Normalization and Exploring global read count patterns

For the RNA-Seq it is important to remember that, even if completely uniform sampling of a diverse pool of transcripts were achieved (which is not possible due to biases), it is not enough to interpret the numbers of reads overlapping with a given gene as absolute indicators of individual gene expression levels. The number of sequences mapped to a gene naturally depends, in addition to its own expression level, on read length (longer transcripts give rise to more short fragments), the depth of sequencing, and the overall expression of all other genes in a sample. To compare gene expression between two conditions, the essential task involves determining the proportion of reads assigned to each gene concerning both the total number of reads and the comprehensive RNA repertoire, which may differ significantly between samples. There are various normalization methods for comparing gene read counts within the same sample (e.g. RPKM, FPKM, TPM) and gene read counts between different conditions (e.g. Total Count, Counts Per Million, DESeq's size factor, TMM) [39]. Normalization plays a crucial role in RNA-Seq, aiming to eliminate systematic effects that are not indicative of the biological differences of interest. This process is essential to ensure the integrity of exploratory analyses and to maintain statistical control for differential gene expression accurately.

Replicates, whether they are technical (multiple measurements of the same sample) or biological (independent samples from the same condition), should exhibit similar gene expression patterns. If either biological or technical replicates show dissimilar expression patterns, it could be indicative of errors in the experimental procedure. Furthermore, if gene expression patterns between different experimental conditions show little variation, it becomes challenging to discern which genes are truly responsive to the experiment. Thus, before proceeding with the identification of differentially expressed genes in RNA-seq data analysis it is important to conduct initial assessments of gene expression patterns. To assess these patterns, three of the most commonly used methods for RNA-seq data are pairwise correlation, hierarchical clustering, and Principal Component Analysis (PCA) [39].

Principal Component Analysis (PCA) is a linear dimensionality reduction technique, which aims to capture the maximum amount of variation in a data set by projecting it into a lower dimensional space while maintaining the essential structure of the given data. PCA is commonly used in bioinformatics to analyze high-dimensional data, such as gene expression profiles acquired from RNA-Seq experiments. This technique has several objectives, including extracting the most important information from the data, finding relationships between observations, identifying and removing outliers, and reducing the dimension of the data by retaining only the important information [157].

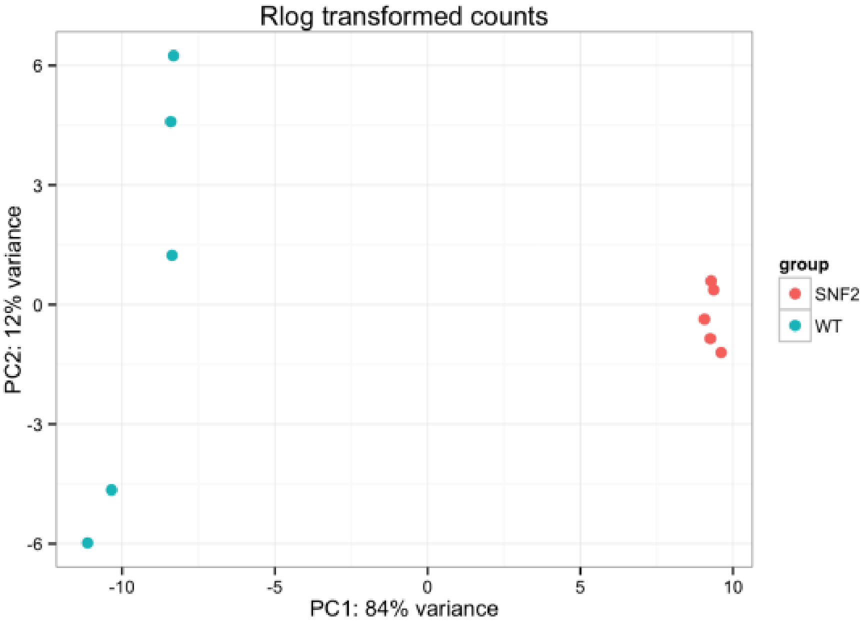


Figure 9: The figure showcases the results of Principal Component Analysis (PCA) on log transformed counts. It demonstrates the power of PCA in reducing dimensionality while preserving significant patterns in the data. The clear separation along PC1 suggests that the differences between WT and SNF2 samples are substantial and are the primary drivers of variability in the dataset. PC2 captures less variance and represents additional, smaller differences within the data. Image retrieved from [120]

2.2.7 Differential Gene Expression

To detect differential gene expression (DGE), various statistical methods have been developed specifically for RNA-Seq data. Cuffdiff, part of the Tuxedo suite tools, is a popular differential expression tool designed to analyze RNA-Seq data [159]. Other packages that support examining the differential expression, are baySeq [62] , DEGseq [168], DESeq [4], DESeq2 [94], edgeR [128], and limma-voom [127]. The last four are considered the best-performing tools of DGE [142]. Each model makes specific assumptions that may not be correct for the acquired data in a particular experiment. Therefore, it is mandatory to understand the model parameters and their limitations so that the resulting conclusions are biologically accurate [18].

These tools follow a similar approach, which consists of the following two basic steps. Firstly, the gene expression difference for a given gene is estimated, while taking the factors of the previously executed classification and clustering into consideration [39]. The procedure is followed by a statistical test based on the null hypothesis that the difference is close to zero, which means that there is no significant difference in the gene expression values that could be explained by the experimental conditions.

In order to assess whether the observed differences in read counts between different conditions for a particular gene are statistically significant, DGE tools often employ the linear regression model. The goal is to estimate the difference in expression levels between conditions for each gene while considering the inherent variability within each condition. Instead of using a linear model, other models are also employed to fit the observed read counts and arrive at the estimate for the difference. These models are based on statistical distributions such as the negative binomial (baySeq, Cuffdiff, DESeq, DESeq2, EBSeq, and edgeR), beta-binomial (BBSeq), binomial (DEGSeq), Poisson (PoissonSeq), and log-normal (limma) [31]. The statistical significance calculated by those DGE tools typically relies on being able to accurately measure the mean count and variance for each gene [31]. In experiments where there is a significant number of replicates per condition (more or equal to 12 replicates), it is feasible to accurately calculate mean and variance estimators directly from the data. However, many RNA-Seq DGE studies have a limited number of replicates per condition (less than or equal to 3 replicates). To address the challenge of low replicas, various DGE analysis tools such as DESeq, DESeq2, edgeR, and limma use strategies that model the mean-variance relationship by borrowing information across genes with similar expression values [39, 34]. The stabilized variance helps to avoid excessive fluctuations that might lead to spurious false positives and negatives but is strongly dependent on an assumed read count distribution and on the assumption that the majority of the gene counts are not truly differentially expressed [31].

Instead of using the raw estimates, these calculated values for the mean and variation are employed to assess differential gene expression. The test used to assign the p-values depends on the tool used. Among the statistical tests used in differential gene expression analysis, t-test, and ANOVA are two of the most well-known ones [39]. The null hypothesis in the t-test is that the given gene under two conditions has the same mean count, whereas in ANOVA is that there are no significant differences among the group means. Regardless of the test statistic, the resulting p-values represent the probability of observing the data if there were no true differences in expression between conditions. It is important to adjust the p-values when analyzing a large number of genes to avoid an inflated number of false positives and to ensure more reliable results for downstream analyses. One form of correction that many tools offer is the Benjamini- Hochberg correction [39].

2.2.8 Data visualization

It is essential to visualize the results obtained from the DEG analysis in order to assess the quality of the analysis and identify potential problems. Also, visualization of the RNA-Seq data can aid in their interpretation by highlighting patterns, relationships, and trends within the data set. Tools often use diagnostic charts, such as istograms (Figure 10 left), heatmaps (Figure 12), volcano plots (Figure 11), MA charts(Figure 10 right), and pathway enrichment plots (Figure 14, Figure 15)[39, 81]

Histograms are a simple way of getting an idea of how frequently certain values are present in a data set. For instance, they are particularly useful, in the case of depicting p-values for all genes tested in the null hypothesis. Heatmaps are used to visualize expression values in individual samples across different conditions or samples, offering insights into patterns and trends in the data. Volcano plots are a type of scatter plot that facilitates the detection of changes in large data sets consisting of repeated data and are commonly used in high-throughput genomics and omics studies. In a volcano plot, each data point represents a gene or transcript, with the x-axis typically showing the fold change in gene expression between conditions (e.g., diseased versus healthy) and the y-axis representing the statistical significance of the difference in expression, often expressed as the negative logarithm (base 10) of the p-value obtained from statistical tests (e.g., t-tests, ANOVA). MA plots are similar to volcano plots in that they display the fold change against the log10 p-value. They show the relationship between the change in expression among the different conditions (M), the average gene expression power (A), and the ability of the algorithm to identify genes with small p-values (highlighted in a different color from the others).

The data and the acquired results can be visualized at the different stages of the RNA-Seq workflow. For instance, visualization can be done at the level of reads (e.g. using ReadXplorer [66]) or at the level of processed coverage (read pileup), not normalized (e.g. total count) or normalized, using genome browsers such as the UCSC browser [79], Integrative Genomics Viewer (IGV) [158], Genome Maps [102], or Savant [48]. Certain visualization tools are specifically designed for visualizing multiple RNA-seq samples, like RNAseqViewer [130], which provides flexible ways of displaying read abundances in exons, transcripts, and crossovers. Introns can be hidden to better reveal signals in exons, and heatmaps can aid in the visual comparison of signals across different samples. Some of the software packages for differential gene expression analysis (such as DESeq2 or DEXseq in Bioconductor) have functions for visualizing gene expression [27]. Enrichr and g:Profiler provide visualizations of enriched biological pathways. EnrichmentMap is another application that visualizes the results of pathway enrichment analysis and facilitates interpretation, displaying pathways as a network in which overlapping pathways are clustered to identify important biological features in the results [126].

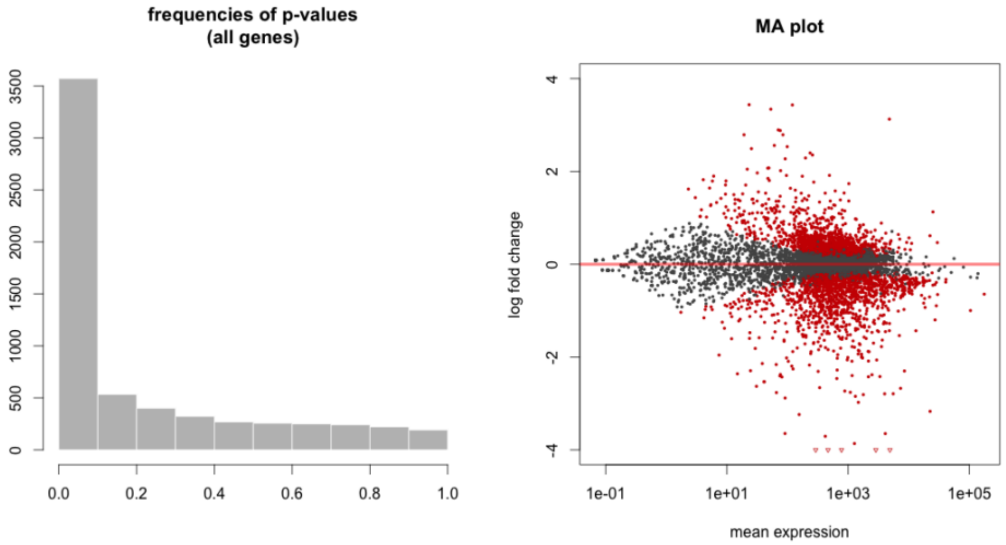


Figure 10: Plot examples. Left: Histogram of p-values for all genes tested for no differential expression between the two conditions, (healthy and pathological). Right: The MA plot shows the relationship between the expression change (M) and average expression strength (A); genes with adjusted p-values < 0.05 are marked in red. Plots retrieved from [120]

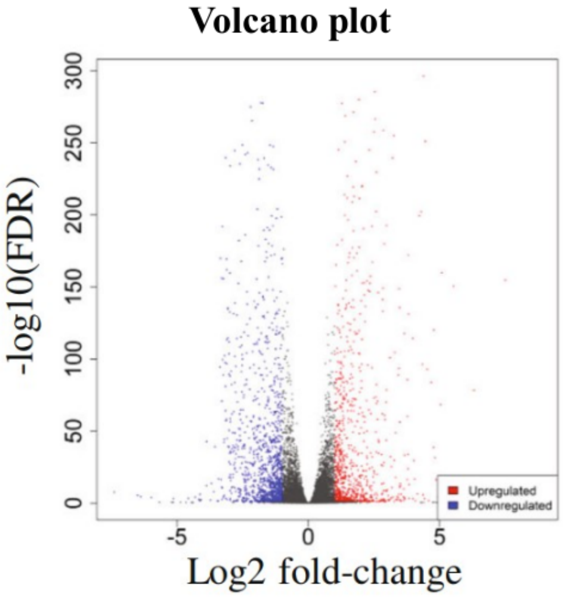


Figure 11: Volcano plot of differential expression analysis. Dots in red represent the up-regulated genes, whereas dots in blue represent down-regulated genes. Genes in gray are not significantly differentially expressed. Plot retrieved from [55]

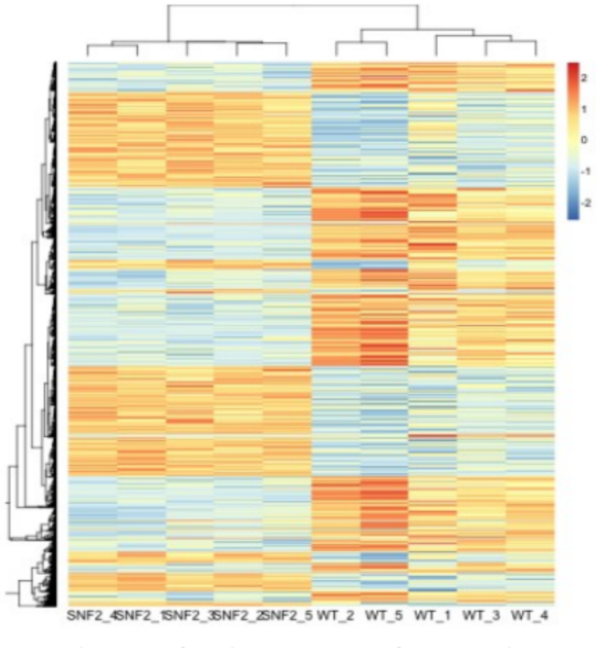


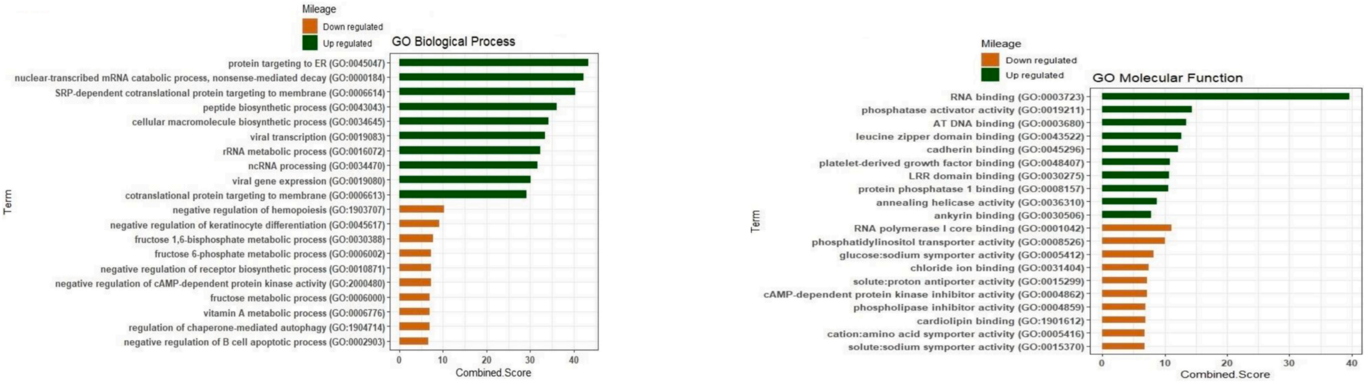
Figure 12: Heatmap example of rlog-transformed read counts for genes with adjusted p-values$<$ 0.05 in the DGE analysis. Genes sorted according to hierarchical clustering with read count values are scaled per row so that the colors actually represent z-scores rather than the underlying read counts. (WT and SNF2 are the two different examined conditions in this particular example). Plot retrieved from [120]

2.2.9 Annotation of differentially expressed genes and Enrichment analysis

To gain a better understanding of the biological significance of the identified differentially expressed genes, additional information is assigned through the annotation process. This process involves annotating gene lists obtained from DE analysis with the associated functional terms, such as Gene Ontology (GO) terms, signaling pathways (relying on databases like Panther, KEGG, Biocarta), or protein domain annotations (for instance based on InterPro) [176]. GO terms are categorized into biological process (BP), molecular function (MF), and cellular component (CC) [145], ranging from general (e.g., activation of the innate immune response) to specific (e.g., antigen processing and presentation of endogenous peptide antigen through the major histocompatibility complex class II [GO:0002491]) [7, 28]. Additionally, functional annotation clustering analysis can be performed, which groups these terms into related functional categories, giving a first insight into the biological impact of the discovered differences [69]. KEGG (Kyoto Encyclopedia of Genes and Genomes) is a comprehensive database resource that integrates genomic, chemical, and systemic functional information.

Enrichment analysis is a statistical method that determines whether a particular set of DEGs are significantly up-or down-regulated for specific biological conditions or annotations compared to a reference database, such as GO terms and KEGG [145]. Statistical tests, such as Fisher's exact test [126] or hypergeometric test [81], are commonly used to determine the significance of enrichment. An example of the results of a GO enrichment analysis of biological processes and molecular functions for genes with up- and down-regulation between healthy and diseased tissues is shown in Figure 13. Exploring pathways enriched in gene lists derived from DE analysis is also essential, unraveling the mechanistic implications of gene expression changes. This method identifies biological pathways (i.e. genes that cooperate to carry out a biological process) that are enriched in a gene list more than would be expected by chance. GO annotations of biological processes serve as valuable resources for pathway enrichment analysis [126]. Tools commonly employed for enrichment analysis include DAVID (Database for Annotation, Visualization, and Integrated Discovery) [35], Enrichr [83], g:Profiler [125] or clusterProfiler [183].

Enrichments are usually evaluated using one of two methods: i) over-representation analysis (ORA) or ii) gene set enrichment analyses (GSEA) [80, 3]. All types of enrichment analyses are based on gene-to-gene comparisons, which means that gene-specific biases, such as gene length, can cause false findings. It has been observed that long transcripts and genes are much more likely to be detected as differentially expressed and also tend to be overrepresented in most of the widely used databases [182]. This is a result of their length making them more likely to be detected in many different experiments.



(a)

(b)

Figure 13: Results of Gene Ontology enrichment analysis of (a) biological processes and (b) molecular functions for up-and down-regulated genes between normal vs tumor ovarian samples. Retrieved from [145]

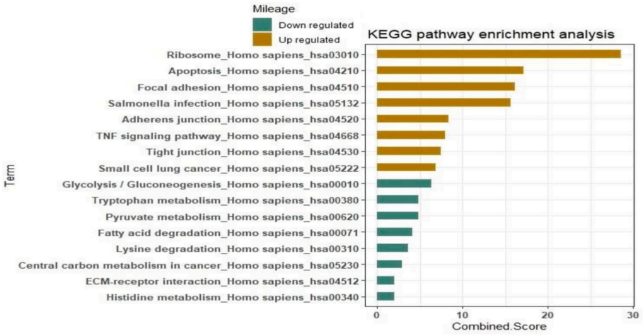


Figure 14: KEGG pathway enrichment analysis for up- and down-regulated genes between normal vs tumor ovarian samples. Retrieved from [145]

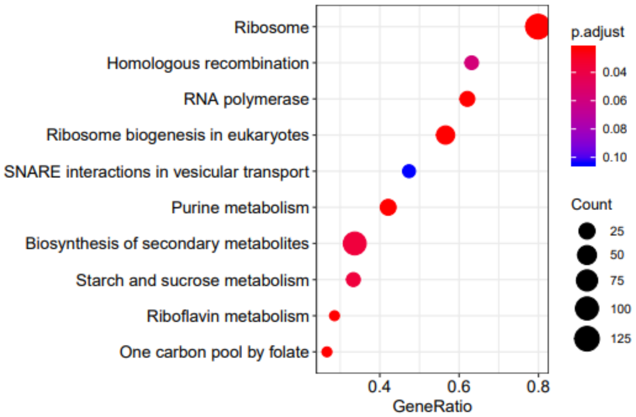


Figure 15: Scatter plot of KEGG pathway enrichment analysis provided by clusterProfiler. Shown here are the most significantly enriched pathways. Retrieved from [39]

2.2.10 Validation of the results

RNA-Seq experiments allow quantitative measurements of thousands of RNAs. Analysis of the data collected from such experiments (depending on the specific experiment and the statistical methodology used) usually reveals that tens or hundreds of thousands of genes are differentially expressed. However, for many of these genes, the result may be a false positive. For this reason, the detected differences in gene expression levels must be confirmed by a different method [120]. Validation methods include in silico analyses, using online databases or simulated datasets, and technical validation with the same RNA samples. However, validation using independent biological replicates is preferred over the aforementioned to confirm true positive DEGs between two or more biological conditions [25].

Differential gene expression is preferably validated by randomly selecting several genes from the list of DEGs identified by RNA-Seq, using independent biological replicates and a gene expression analysis technique, like Western blotting or qPCR [124]. Quantitative high-throughput reverse transcription PCR (qPCR) is a well-established technique with standardized protocols and serves as a reliable tool to confirm the differential gene expression observed in high-throughput sequencing [120]. It is the most common validation method due to its high sensitivity, specificity, and dynamic range, allowing the detection of even subtle differences in gene expression between different conditions. In addition, qPCR instruments offer high throughput and automation, simplifying the validation process across multiple samples or conditions. Transcription factors that are up- or down-regulated may have accompanying epigenetic modifications. Thus, other high-throughput sequencing methods such as ChIP-seq and ATAC-seq (assay for chromatin sequencing with transposase access) can be employed to further unravel their role [17, 103]. Chromatin profiling can contribute to uncovering both temporal and spatial expression changes [175]. Furthermore, causality claims should be supported by functional studies or genetic ablation, optimally limited to the cell type or cell line of interest, to mitigate confounding effects from neighboring cells and the microenvironment [81].

2.2.11 Interpretation of the RNA-Seq results

The final step of the RNA-Seq workflow is interpreting the acquired data and the downstream analysis. After a brief recapitulation of the biological question or hypothesis that guided the RNA-seq study, the main findings of the differential expression analysis are summarized, including the number and identity of DEGs and the enriched pathways.

The biological functions and roles of the identified DEGs in the context of the biological question need to be discussed, highlighting specific genes that are significantly up-or down-regulated, as well as, their known or predicted functions, as identified through functional annotation analysis. It is essential to unravel the biological significance of enriched pathways in the context of the disease phenotype under study and to interpret how their dysregulation contributes to disease onset or progression. The possible interactions and cross-talk between different pathways, and how these may affect overall cellular processes, should also be considered. If the results are validated using a different experimental method, it is important to report whether there is biological relevance to the results of the key findings obtained from RNA-Seq.

The results of the RNA-Seq analysis should be interpreted critically, considering the study's limitations, potential error sources, and alternative explanations for the observations. For instance, in case the RNA-seq results suggest changes in expression that differ dramatically from what is already known from previous studies, interpretation of the results should be very cautious. It is important to avoid over-interpretation of the data and to present the results clearly and transparently, as over-interpretation can lead to inaccurate conclusions and misrepresentations of findings, potentially influencing subsequent research directions and decision-making processes [81].

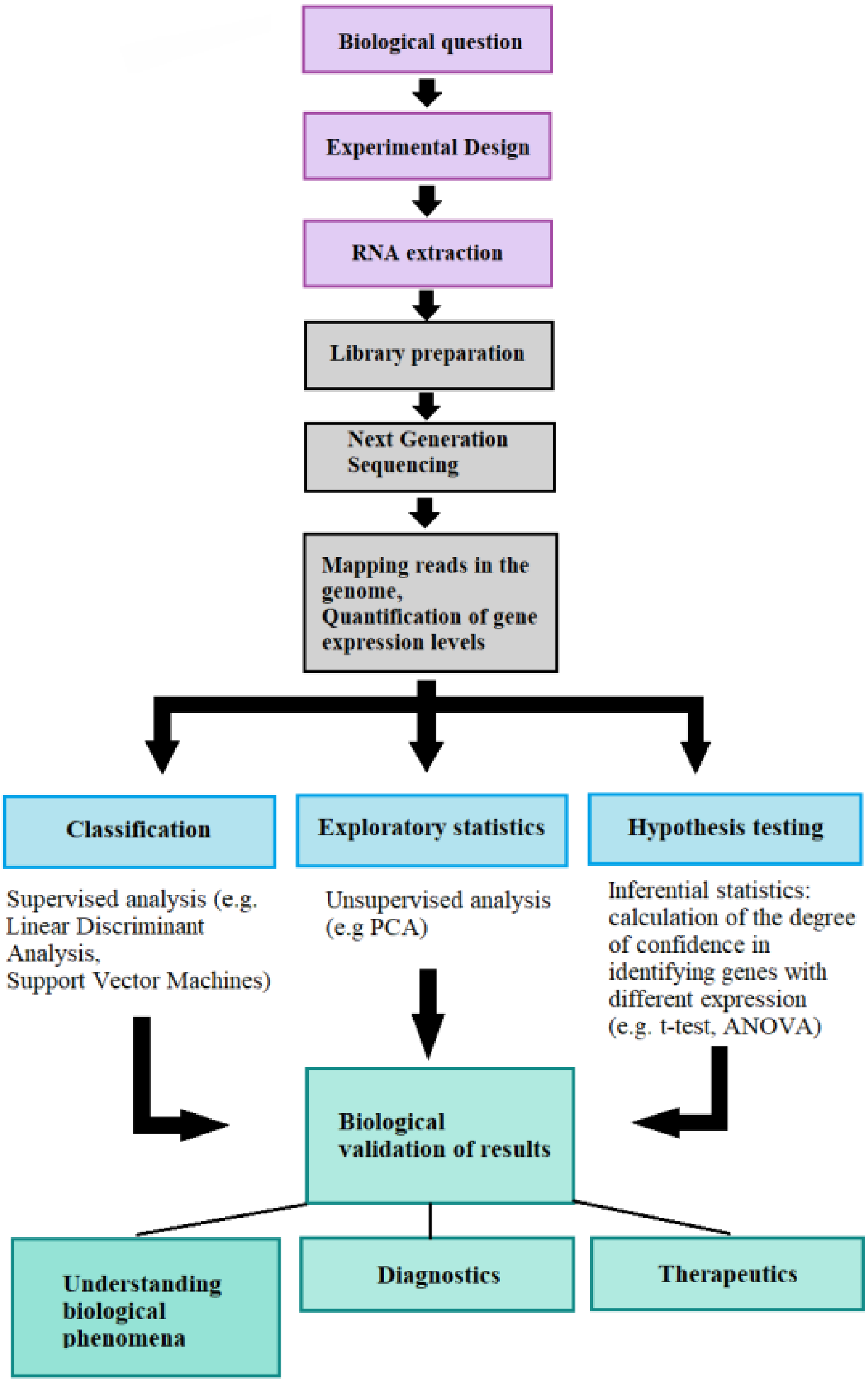

Biological question
Experimental Design
RNA extraction
Library preparation
Next Generation Sequencing
Mapping reads in the genome, Quantification of gene expression levels
Classification
Exploratory statistics
Hypothesis testing
Supervised analysis (e.g. Linear Discriminant Analysis, Support Vector Machines)
Unsupervised analysis (e.g PCA)
Inferential statistics: calculation of the degree of confidence in identifying genes with different expression (e.g. t-test, ANOVA)
Biological validation of results
Understanding biological phenomena
Diagnostics
Therapeutics

Figure 16: Overview of gene expression assessment workflow with RNA-Seq. Purple frames: First a biological question is formulated, then the experimental design is devised and the RNA is isolated. Grey frames: The RNA is converted into complementary DNA (cDNA) and a library is built. The sequence of each molecule is identified and the sequence reads are mapped to a reference genome. Gene expression levels are then quantified. Blue frames: In RNA-Seq, hypothesis testing is performed using t-tests, ANOVA, and other statistical procedures to determine which transcripts are overexpressed or underexpressed in an experiment. It is possible to apply exploratory statistics such as the clustering of genes or samples. In supervised learning approaches, genes or samples are associated with specific categories according to already available classification information (e.g. normal versus pathological tissue) and gene expression measurements can be used to predict which unknown samples can be classified as healthy or pathological. Green frames: Finally, after the RNA-Seq analysis, the experimental validation is carried out These approaches can potentially unravel complex biological processes and provide important insights into diseases. They can facilitate the discovery of diagnostic markers and the development of therapeutic intervention strategies. Image based on [120]

2.3 RNA-Seq Databases

RNA-Seq databases play a critical role in the storage, organization, and accessibility of RNA-Seq data generated from various experiments. These databases are valuable resources for researchers studying gene expression, transcriptomics, and biologically related processes. Some of the most commonly used platforms are listed below.

The Genotype-Tissue Expression (GTEx) [112] project was launched in 2010. As of 2024, it is still an ongoing project whose main goal is to create a molecular and data analysis resource. It aims to create a tissue bank to study the gene expression in healthy tissues that were acquired from subjects in different developmental stages (neonatal, pediatric, and adolescent tissues). Specifically, according to the latest update, the samples were collected across about 1000 individuals from 54 non-diseased tissues primarily for molecular analyses, mainly for WGS, WES, and RNA-Seq. The GTEx Portal allows open access to gene expression data, Quantitative Trait Loci (QTLs), and histological images.

The Cancer Genome Atlas (TCGA) [115] is an effort between NCI and the National Human Genome Research Institute which began in 2006 and has produced a rich dataset in the field of cancer 'omics'. Over the course of 12 years and with input from more than 11,000 patients, TCGA has produced over 2.5 petabytes of genomic, epigenomic, transcriptomic, and proteomic data. The characteristics of more than 20,000 primary cancer samples have been systematically analyzed along with corresponding normal tissue samples. The project included a wide range of cancer types, a total of 33 different cancer types, gaining comprehensive knowledge of the genetic and molecular alterations associated with cancer. The data are publicly available to the research community and have already been exploited, improving cancer diagnosis, treatment, and prevention.

Gene Expression Omnibus (GEO) [113] is a public repository hosted by NCBI that archives and provides free access to raw and processed microarray, NGS, and other forms of high-throughput functional genomics data submitted by the public research community. GEO Profiles is a huge collection of gene-level data, including cancer-related data, in which users can search for gene expression profiles related to their interests.

3 Selecting a dataset

In our project, we exploited transcriptome data sourced from the GEO. Specifically, we retrieved a targeted subset of data from the GEO database, utilizing the Sequence Read Archive, or Short Read Archive originally (SRA). The SRA is a great resource for acquiring sequencing data that underlie numerous publications and studies. It functions as the primary repository for unassembled reads, encompassing the raw sequencing data generated directly from instruments.

The selected dataset includes data from both cancerous and healthy tissues of LUSC patients, facilitating direct comparisons between pathological and non-pathological states. The cohort study of 37 LUSC patients was established by recruiting patients from the Yonsei Cancer Center (YCC) in Seoul, Korea. Biopsies of tumors and normal tissues were performed on each patient during surgery. Samples were obtained from the Severance Hospital records. Typical procedures of RNA sequencing were performed for the pairs of tissue specimens. Firstly, each sample underwent sequencing library preparation, which was performed using the TruSeq RNA Access Library Prep Guide Part # 15049525 Rev. B with the TruSeq RNA Access Library Prep Kit (Illumina). RNA sequencing was conducted using the HiSeq 2500 system (Illumina) and the sequencing data were processed according to the manufacturer's protocol.

It should be noted that the full dataset (Accession number: GSE158433) encompasses not only the RNA-Seq data of interest (GSE158420) but also DNA methylation data (GSE158422). The primary objective outlined by the investigators who uploaded the data to NCBI was to conduct a comprehensive genome-wide analysis of DNA methylation in both tumor and normal tissues to unravel the epigenetic regulatory mechanisms underlying carcinogenesis [23], rather than focusing on the differential gene expression.

In this work, we used the Linux operating system, which is commonly employed to manage the files resulting from RNA-Seq data analysis.

3.1 Downloading the data

Due to limitations related to computational resources, we chose to selectively analyze a subset of the RNA-Seq data from the 74 samples. Specifically, we focused our analysis on the first 8 patients (both tumor and normal samples for both) listed on the SRA Run Selector page. This decision was necessitated by the significant memory requirements associated with processing the entire dataset. We selected the first 16 runs from the study, totaling about 54.60 Gb of the gzipped data and downloaded them via the SRA Run Selector page.

3.2 Quality Control

After obtaining paired sequencing data we conducted essential quality control checks on the raw data to gain insights into the dataset's integrity and identify potential issues. Utilizing FastQC, we examined the individual files for quality metrics. The tool produces a basic text and an HTML output file that contains all of the valuable results and plots. To generate a comprehensive overview, we then employed MultiQC, creating a summarizing report encompassing quality control outputs of the entire dataset [43]. The MultiQC report results are presented below.

General Statistics

Sample Name	% Dups	% GC	Length	% Failed	M Seqs
SRR12697412_forward	82.3%	50%	101 bp	18%	35.4
SRR12697412_reverse	75.9%	50%	101 bp	9%	35.4
SRR12697413_forward	79.8%	50%	101 bp	18%	33.6
SRR12697413_reverse	74.8%	50%	101 bp	9%	33.6
SRR12697414_forward	80.7%	49%	101 bp	18%	34.9
SRR12697414_reverse	75.4%	49%	101 bp	9%	34.9
SRR12697415_forward	82.9%	49%	101 bp	18%	35.0
SRR12697415_reverse	77.2%	50%	101 bp	9%	35.0
SRR12697416_forward	80.0%	48%	101 bp	18%	36.7
SRR12697416_reverse	74.6%	49%	101 bp	18%	36.7
SRR12697417_forward	85.7%	50%	101 bp	18%	34.2
SRR12697417_reverse	81.7%	50%	101 bp	9%	34.2
SRR12697418_forward	80.6%	49%	101 bp	18%	35.1
SRR12697418_reverse	75.7%	49%	101 bp	18%	35.1
SRR12697419_forward	81.6%	48%	101 bp	18%	34.3
SRR12697419_reverse	75.4%	48%	101 bp	9%	34.3
SRR12697420_forward	83.3%	49%	101 bp	18%	34.0
SRR12697420_reverse	77.2%	50%	101 bp	18%	34.0
SRR12697421_forward	81.4%	49%	101 bp	18%	31.5
SRR12697421_reverse	75.4%	50%	101 bp	18%	31.5
SRR12697422_forward	81.7%	49%	101 bp	18%	34.7
SRR12697422_reverse	75.6%	50%	101 bp	18%	34.7
SRR12697423_forward	81.8%	50%	101 bp	18%	33.9
SRR12697423_reverse	76.5%	50%	101 bp	9%	33.9
SRR12697424_forward	84.1%	48%	101 bp	18%	33.4
SRR12697424_reverse	78.2%	49%	101 bp	9%	33.4
SRR12697425_forward	88.2%	49%	101 bp	27%	33.4
SRR12697425_reverse	84.3%	50%	101 bp	9%	33.4
SRR12697426_forward	81.5%	49%	101 bp	18%	33.9
SRR12697426_reverse	75.9%	50%	101 bp	18%	33.9
SRR12697427_forward	86.7%	50%	101 bp	18%	32.3
SRR12697427_reverse	82.1%	50%	101 bp	9%	32.3

Table 2: The table presents various sequencing metrics for each sample, including the percentage of duplicate reads, GC content, read length, percentage of failed reads, and the number of million sequences.

Sequence Counts

This plot shows the total number of reads, broken down into unique and duplicate if possible.

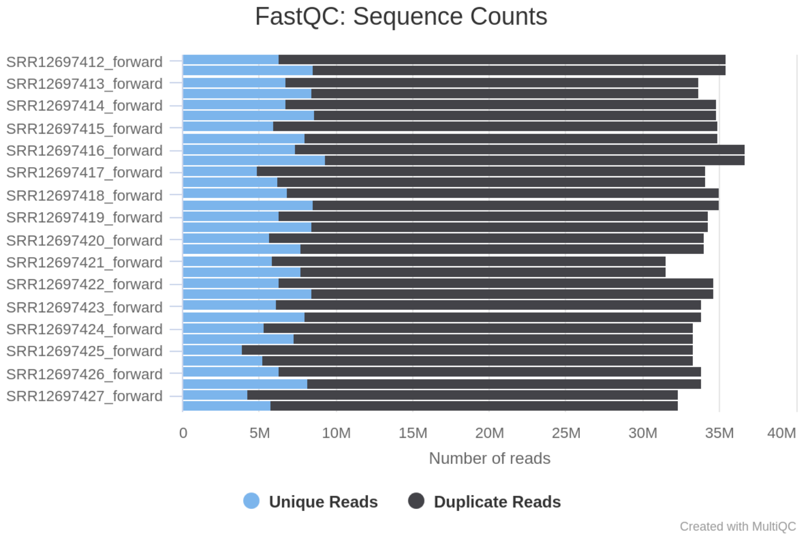


Figure 17: Sequence Counts

Our samples have high proportion of duplicated reads, as it is expected in RNA-Seq data.

Per Base Sequence Quality

This graph shows the mean quality value across each base position in the read. To allow multiple samples to be plotted on the same graph, only the average quality scores are plotted (as opposed to the box plots shown in FastQC reports).

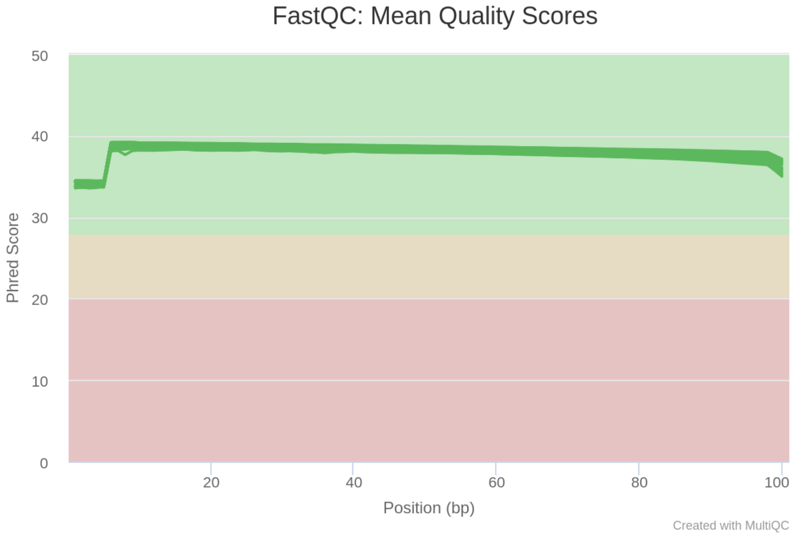


Figure 18: Per Base Sequence Quality

The ’per base sequence quality’ is globally good in this dataset with a slight decrease at the end of the sequences. Moreover, with all Illumina sequencers, it is common for the median quality score to start lower in the first 5-7 bases and then increase, as it is observed in the graph above.

Per Sequence Quality scores

This graph reports the quality score per sequence indicating whether a subset of sequences has universally low quality scores.

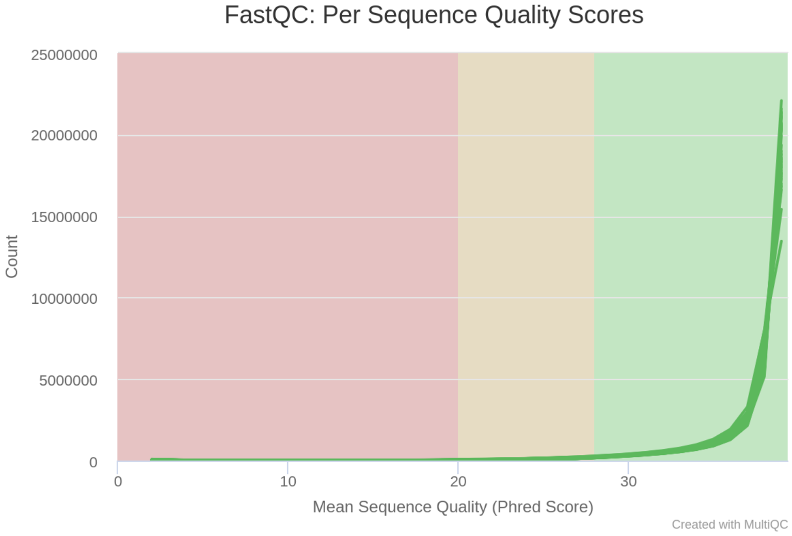


Figure 19: Per Sequence Quality Scores

Sometimes, a subset of sequences has universally low quality, however, these should only represent a small percentage of the total sequences. In this case, the mean quality score has a similar distribution and is quite high for all the reads.

Per Base Sequence Content

In order to view the base composition data for all the samples, the following heatmap is constructed by MultQC. Each color represents the balance between the four bases. Adenine, Thymine, Guanine and Cytosine are represented with blue, red, green and gray respectively. An even distribution should give an even muddy brown color, as all colors contribute in equal amounts.In the htlm MultiQC report, it is possible to see the percentage of the four bases under the cursor, while hovering over the plot. The original FastQC line plots can be shown when clicking on a simple track.

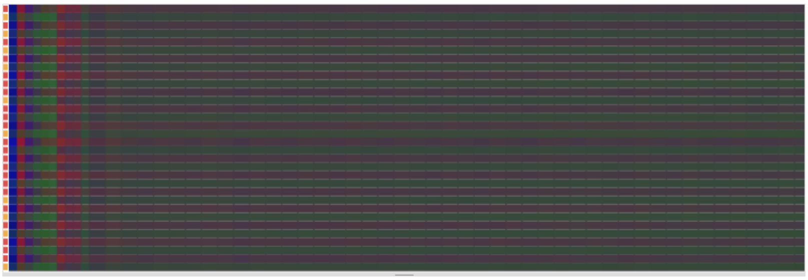


Figure 20: Heatmap of the per Base Sequence Content

In our dataset, 10 of the samples have a warning and the rest have a fail warning. As we can observe in the above heatmap the beginning of each line is dominated by bold colors representing specific bases. This indicates that at the beginning of the read (first 10-12 bp) the proportion of the bases is highly imbalanced, whereas in a random library, the ratio of each of the four bases is expected to be maintained constant throughout the length of the read. It is a true technical bias, commonly encountered in RNA-Seq libraries, and although it cannot be eliminated by trimming, it does not seem to affect the analysis whatsoever.

Per Sequence GC Content

This plot shows the GC content across the whole length of each sequence in all the dataset files and compares it to a modeled normal distribution of GC content.

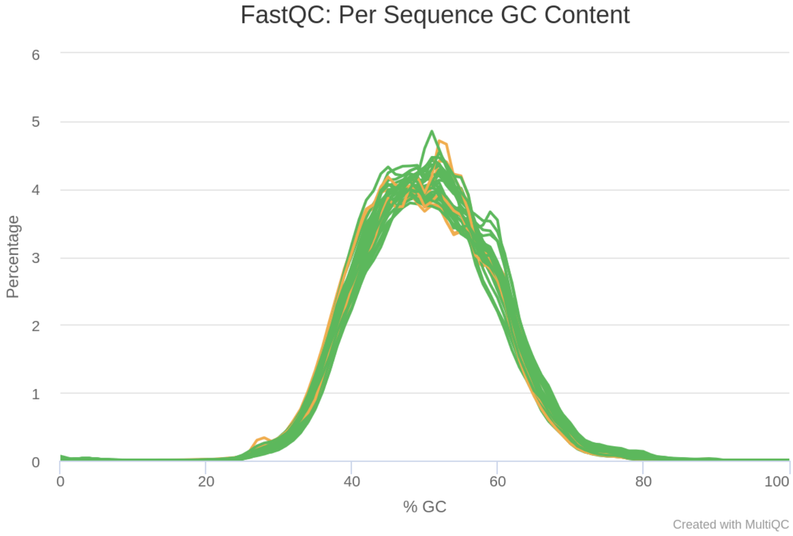


Figure 21: Per Sequence GC Content

The 27 out of the 32 samples follow a roughly normal distribution of GC content. The remaining 5, assigned with orange color, fit the normal distribution worse. There are some sharp peaks observed that are possibly a result of a specific contaminant, such as adapter dimers.

Per Base N Content

In the diagram below is a summary for all the samples, presenting the percentage of N calls at each position for which the sequencer did not have sufficient confidence to call a base.

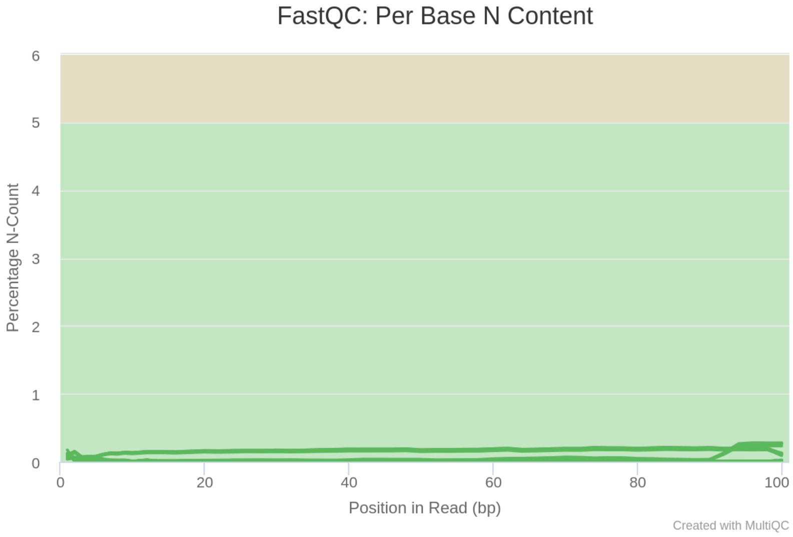


Figure 22: Per Base N Content

We can observe that the sequencers have efficiently identified almost all the bases in the reads as the percentage of N calls is very low across the reads.

Sequence Length Distribution

In the chosen dataset, all 24 paired-end reads have sequences of the same length, 101 base pairs.

Sequence Duplication Levels

The following graph shows the relative number of sequences with varying degrees of duplication for all the samples.

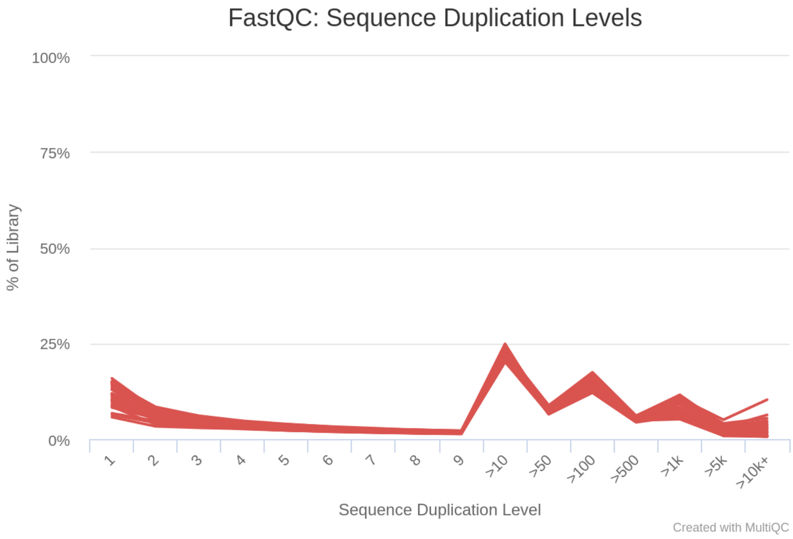


Figure 23: Sequence Duplication Levels

We observe that there are 3 distinct peaks towards the end of the end of the graph, meaning that there is a large number of sequences which is highly duplicated. All the samples have been issued a failure warning, meaning that the amount of non-unique sequences make up more than 50% of the total. In RNA-Seq libraries, however, this duplication profile is common. This difference in the levels of the starting population is done on purpose, to be able to observe lowly expressed transcript. This results in high overall duplication which comes from physically linked regions. An examination of the distribution of duplicates in a specific genomic region allows the distinction between over-sequencing and general technical duplication.

Overrepresented sequences

MultiQC creates a summary report for all the samples by creating a bar plot that shows the total amount of overrepresented sequences found in each library, instead of listing the counts for each overrepresented sequence, as FastQC does. In blue color is denoted the percentage of the top overrepresented sequences and in gray is the percentage of the sum of the remaining over-represented sequences.

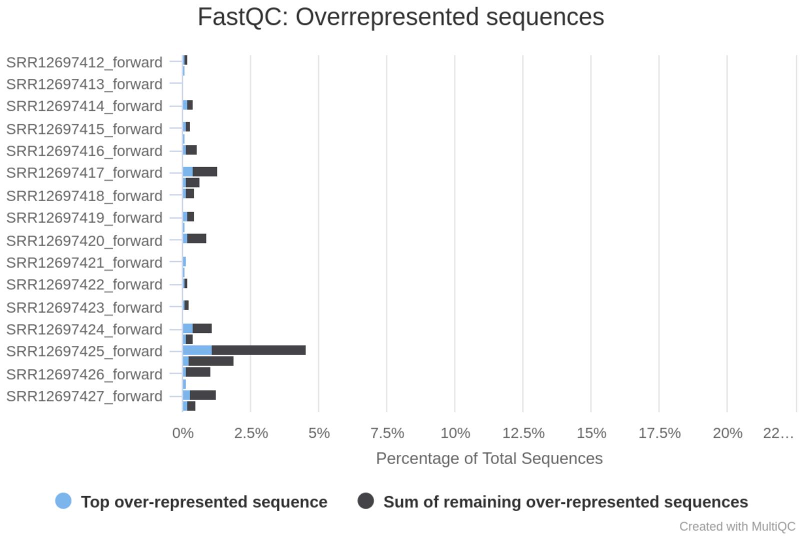


Figure 24: Overrepresented sequences

In our dataset, 23 fastq files have issued a warning, meaning that there is at least one sequence that is found to represent more than 0.1% of the total. 1 of them has failed, indicating that at least one sequence is found to represent more than 1% of the total. In RNA sequencing commonly the data may have certain transcripts listed as an over-represented sequence.

Adapter Content

The line plot below shows the cumulative percentage of the proportion of the libraries that have seen each of the adapter sequences at each position, including Illumina Universal Adapter, Illumina Small RNA 3' Adapter, Illumina Small RNA 5' Adapter, Nextera Transposase Sequence, PolyA, PolyG. Each colored line represents the percentage of sequences where a specific adapter is found across the read at each fastq file. This diagram provides detailed insights into the presence and abundance of these adapters throughout the sequencing data. By hovering the cursor over each line in the html file, we can access more information about the adapters identified in the libraries.

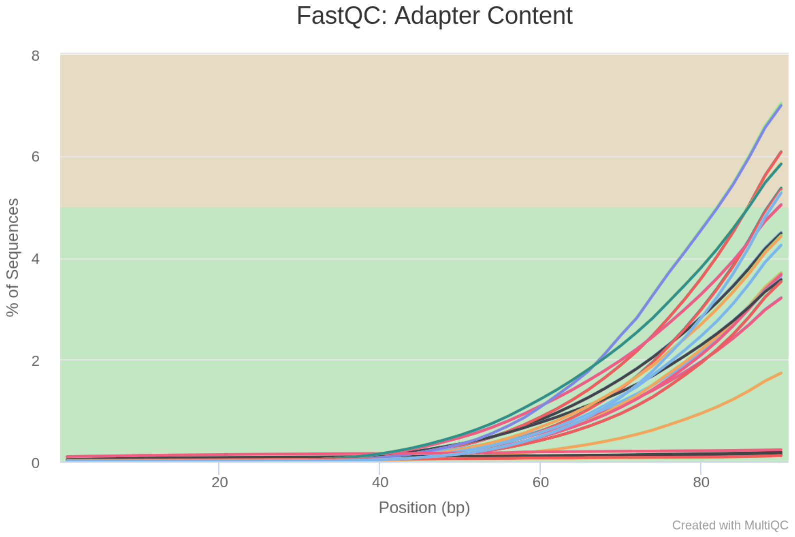


Figure 25: Adapter Content

Of the 32 reads analyzed, 12 issued a warning, indicating the presence of more than one sequence in over 5% of all reads. More precisely, the detected sequence in this scenario was identified as the Illumina Universal adapter. Moreover, the lower sections of the graph disclose the presence of the polyA adapter, although in smaller proportions.

Status Checks The following plot summarizes the status for all the FastQC reports showing whether results seem entirely normal (green), slightly abnormal (yellow) or very unusual (red).

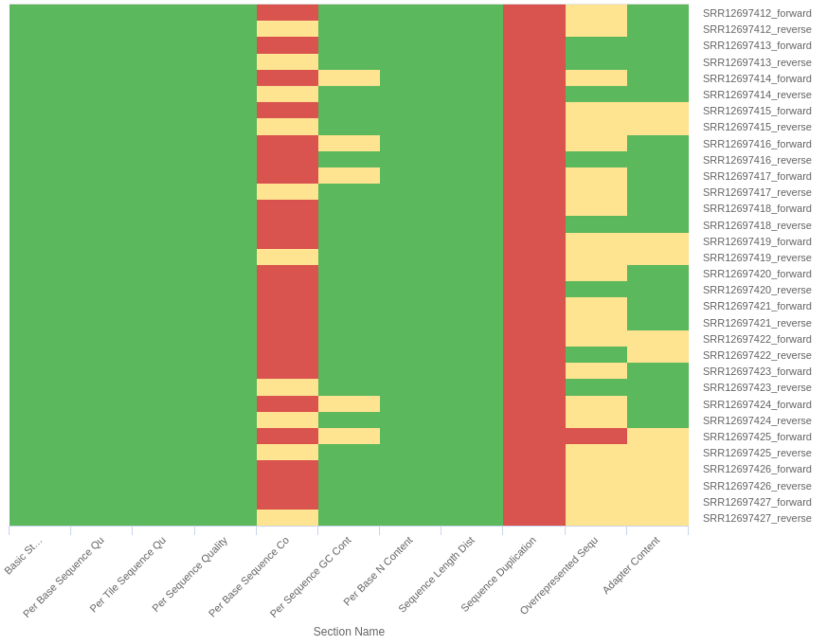


Figure 26: Status Check Heatmap

The majority of samples show good quality across most metrics, but there are consistent issues with Per Base Sequence Content and Sequence Duplication Levels, reflected in the red and yellow bands. Attention should also be given to Overrepresented Sequences and in particular to the Adapter Content to ensure high quality sequencing data.

4 Read Trimming

Subsequently, we trimmed the data to remove low-quality bases and adapter sequences from the raw reads. To achieve this, we chose Trimmomatic, which is a fast, multi-threaded command line tool that performs various useful trimming operations on both paired and single-ended Illumina data.

In particular, the paired-end mode, which was utilized here, preserves the correspondence of paired reads and leverages the additional information contained in the paired reads to more efficiently find adapter fragments or PCR primers introduced by the library preparation process. For this mode, two input files (forward and reverse reads) are specified. Four output files are obtained, two for the "paired" output where both reads survived processing (forward paired and reverse paired reads), and two for the corresponding "unpaired" output where one read survived, but the partner read did not survive (forward unpaired and reverse unpaired reads).

Trimmomatic allows the steps of the trimming process, briefly presented in the table below, to be performed. The parameter selection for each step depends on the information of the dataset, which we obtained from the quality control.

Module	Description
ILLUMINACLIP	Cut adapter and other Illumina-specific sequences from the read
SLIDINGWINDOW	Performs sliding window trimming, cutting once the average quality within the window falls below a predefined threshold
MINLEN	Drops the read if its length is below a specified threshold
LEADING	Cuts bases off the start of a read, if they are below a threshold quality
TRAILING	Cuts bases off the end of a read, if they are below a threshold quality
CROP	Cuts the read to a specified length
HEADCROP	Cuts the specified number of bases from the start of the read
AVGQUAL	Drops the read if the average quality is below a specified value
MAXINFO	Cuts the read adaptively, balancing the reading length and error rate to maximize the value of each read
TOPHRED33	Converts quality scores to Phred-33
TOPHRED64	Converts quality scores to Phred-64

Table 3: Trimmomimg processes avaliable with Trimmomatic

The order of execution of the trimming steps is of particular importance, as the various steps are executed in the order in which they are specified on the command line. As recommended the adapter cut-off is performed first, since correct identification of the adapter using partial matches is more challenging. This process requires the maximum amount of information possible to achieve a delicate trade-off between sensitivity (accurately detecting and eliminating all contaminant sequences) and specificity (precisely identifying contaminant sequences and genuine biological data).

SRA did not clarify which adapters were used in this experiment but mentioned the machine that was employed to sequence the samples. On the same machine, different adapters can be used. To perform the trimming of the adapters in this case, we arbitary chose the TruSeq-PE-2 adapters and they seem to work well for our dataset. For palindrome trimming, a matched pair or multiple matched pairs (Prefix pairs) of adapter sequences are provided. Additionally, during the trimming process, efforts are made to identify and remove other sequences (Long Clipping Sequences). In table below the TruSeq-PE-2 adapters utilized here are presented.

Type	Sequences
PrefixPair	TACACTCTTTCCCTACACGACGCTCTTCCGATCT GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
Long Clipping Sequences	AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTA AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT TACACTCTTTCCCTACACGACGCTCTTCCGATCT

Table 4: Adapter and Illumina-specific Sequences

The remaining parameters were set as follows:

• Maximum mismatch count which still allows a full match to be performed: 2

• Accuracy of the match between two 'adapter-ligated' reads that is required for PE palindrome read alignment: 30

• The accuracy of the match between any adapter etc. sequence that is required to be against a read: 10

We also set the 'LEADING' module to 3, cutting the first 3 bases which are usually of poor quality, and the 'MINLEN' module to 75, keeping sequences of length equal to or greater than 75. We did not use the "SLIDINGWINDOW" module, as the quality scores were quite high (above 30) for all reads in our dataset. After performing trimming on all paired-end data using uniform parameters, we created a bar plot showing the distribution of the input paired reads after trimming. This bar plot depicts the percentage of read pairs that fall into the four categories: those in which both reads survived the trimming process, reads in which only the forward read survived, reads in which only the reverse read survived, and reads that were dropped entirely.

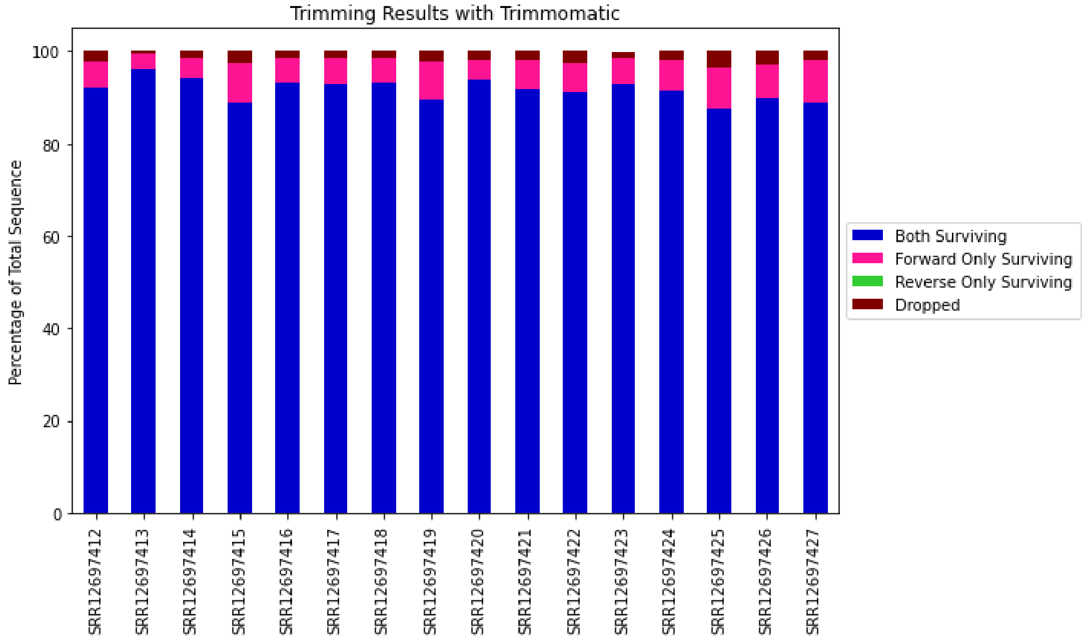


Figure 27: Bar plot showing the percentage of read pairs that: both reads survived the trimming process(blue), reads in which only the forward read survived (fuchsia), reads in which only the reverse read survived (light green), and reads that were dropped entirely (dark red)

As we can observe, the number of paired reads that both survived is large. This is the data we will use for the downstream analysis. The forward only surviving are a smaller percentage and the reads that dropped from both forward and reverse are even less. The reverse only reads that survive constitute only a small percentage and therefore do not appear in the histogram.

Additionally, we performed quality control with FASTQC and then MultiQC to visualize the changes in the dataset after trimming. The two modules that changed the most are the length of the reads and the adapter content. The sequence length distribution is now varying but since we chose to trim the reads, it is not something to consider. The adapter contamination was found less than 0.1% for all the samples.

5 Read alignment

5.1 Indexing the genome and aligning reads using STAR

The next step following the data trimming is read alignment in order to determine the locus of the genome where the reads originate from. The alignment process consists of choosing an appropriate reference genome to map our reads against, annotating the reference genome (defining which parts of the reference sequence correspond to genes) and performing the read alignment employing one of several alignment tools. In this case we used STAR Aligner (Spliced Transcripts Alignment to a Reference), which is designed to specifically deal with many of the challenges of RNA-seq data mapping, utilizing a strategy that takes into account alignments with splicing. The STAR algorithm achieves this particularly efficient matching by performing basically a two-step process, which includes: 1) seed searching and 2) clustering, stitching, and scoring [37].

The basic idea of seed searching focuses on identifying the Maximally Matched Prefix (MMP). In particular, STAR searches for the longest sequence that matches exactly one or more locations on the reference genome. The longer matching sequences are called MMPs. In the initial step, the algorithm locates the MMP starting from the first base of the read. Commonly reads comprise at least one splice junction and cannot be contiguously mapped to the reference genome. Thus, the first seed (seeds are called the parts of the reads that are mapped separately) will be mapped to a donor splice site. Subsequently, the search is repeated for the unmapped portion of the read to find the next MMP. This sequential searching of only the unmapped portions of reads makes the STAR algorithm extremely efficient. STAR implements an uncompressed suffix array (SA) to efficiently search for MMPs, compared to other slower aligners that use algorithms that often search the entire read sequence before splitting reads and performing iterative rounds of mapping. The use of SA allows fast searching even against larger reference genomes. In case there is a failure in finding an exact sequence match for each part of the read due to introns or mismatches, the previous MMPs are expanded. If the expansion does not give good alignment, then the poor quality, adapter sequence, or other contaminating sequence will be soft clipped. The MMP search is performed in both forward and reverse directions of the read sequence.

In the second step of the algorithm, the individual seeds are stitched to form a complete read. This is carried out by clustering the seeds based on their proximity to a chosen set of 'anchor' seeds (seeds that align well to a certain region in the genome) and is guided by the local alignment scoring scheme, with user-degined scores for mismatches, indels, gaps, etc. Additionally, this method leverages the nature of paired-end reads to improve alignment sensitivity. This is achieved by treating the pair-end reads as a single continuous sequence and making better use of the information that the reads come from the same DNA fragment. If STAR finds a correct alignment anchor from one end of the mates it can accurately align the entire read.

The STAR workflow consists of two main steps: the generation of genome markers and the mapping of reads to the genome, which is accomplished as described above. To begin with, we downloaded the Homo sapiens GRCh38.80 genome from ENSEMBL in FASTA file format. We chose the uncovered primary assembly, i.e. the reference genome containing all top-level sequence regions, excluding haplotypes and patches, as and all repeats and low-complexity regions without changes. Also, the corresponding annotation file provided from ENSEMBL was downloaded for indexing the genome; we selected the GTF file in particular.The annotation of the genome was provided through a GTF file, which includes detailed information about exons and genes.

After downloading the necessary files, we utilized them as inputs to STAR to generate a comprehensive genome index in order to facilitate the alignment of sequencing reads. This process requires a large amount of memory usage and the parameters had to be chosen carefully. The output index files contain all the information from the reference genome in a compressed format that is optimized for efficient access and comparison with the query read sequences.

Subsequently, after indexing the genome, STAR is used to match each read to the reference sequence, using the additional information about splicing junctions etc. The alignment step must be performed for each individual FASTQ file. The output of STAR is stored in a SAM or BAM file, the format of which is described in greater detail above. In this case we chose BAM file format as the output.

5.2 Quality control of aligned reads

Once the reads have been aligned, quality control of the output files should be performed before downstream analyses. The basic alignment assessment of aligned reads should be to generally check the aligner's output. With the use of STAR and SAMtools [90] we generated the following files:

File extension	**File contents**

*Aligned.sortedByCoord.out.bam	information about the genomic loci of each read, including the sequence
*Log.final.out	alignment statistics (number of mapped reads)
*Log.out	files, parameters and commands used during the alignment procedure
*Log.progress.out	elapsed time
*SJ.out.mate1	genomic loci where the splicing junctions were detected and the the amount of reads overlapping with them
*Unmapped.out.mate1	unmapped reads (similar to original fastq file)

Table 5: Files generated from STAR and SAMtools

The number of uniquely mapped reads is usually the most important number.

We constructed the following bar plots to visualize the read alignment results.

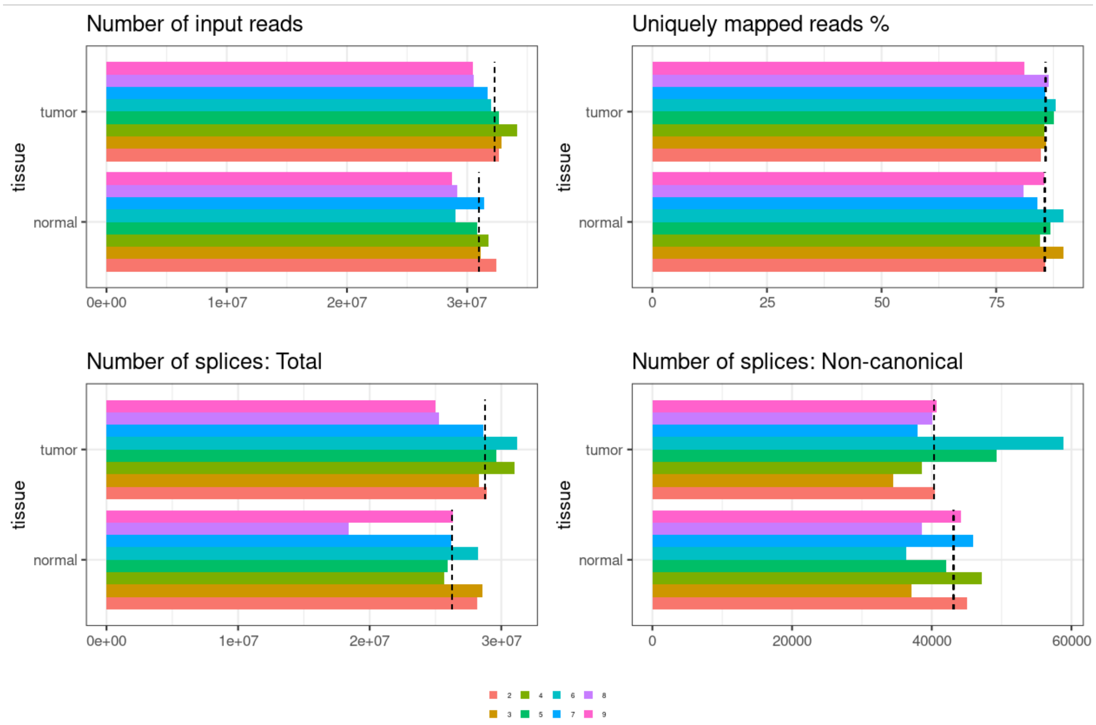


Figure 28: Graphical summary of log files for the 16 samples. The different colors represent the distinct samples and the dashed lines indicate the median values across all samples of the same condition (tumor and normal). The bar plots were created with the help of `https://github.com/friedue/course_RNA-seq2015`.

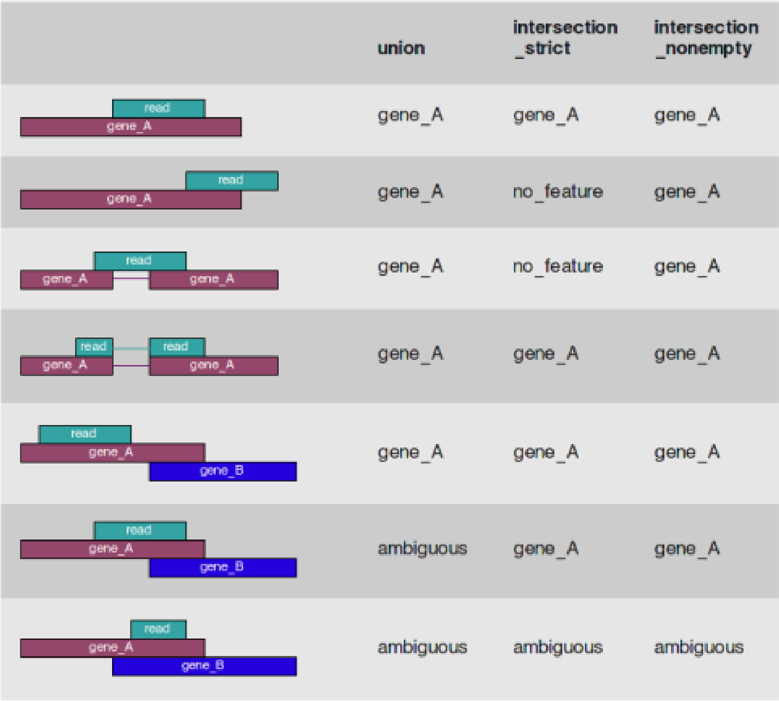
As we can observe from the top left plot, the number of input reads for both tumor and normal tissues appears to be fairly consistent across different samples, which suggests that the sequencing depth was uniform. In the top right plot we can see that the percentage of uniquely mapped reads is relatively high (around 79%) and consistent for both tumor and normal tissues, which is a good indicator of high-quality data. Bottom left bar plot shows the total number of splices for each read, which is fairly consistent across the samples, both for tumor and normal tissues. This consistency suggests that the transcriptomic complexity is comparable across the samples, which is expected for similar tissue types. Small variations may reflect biological differences between tumor and normal tissues, such as alternative splicing events. Finally, in the bottom right plot we can see the amount of non-canonical splices. The presence of non-canonical splices in normal tissue is notable but lower than in tumor samples. These differences could be important, potentially linking splicing variations to the pathology of the tumor samples.

6 Read Quantification

Continuously, we should count the number of reads which correspond to a specific gene. In principle, counting the reads overlapping with genomic features is simple but some details need to be determined. We employed featureCounts to carry out this procedure.

FeatureCounts is a high-performance general-purpose read summary program that counts mapped reads for genomic features such as genes, exons, promoters, gene bodies etc. It accepts SAM/BAM files as input and an annotation file containing the chromosomal coordinates of the features. The output contains read counts mapped to features, as well as statistical information on the overall summary results, including the number of successfully matched reads and the number of those that did not match due to various reasons.

For counting reads, featureCounts gives two options. The first counts reads that match individual features, such as exons, providing a detailed view of the alignment. The second option, counts reads that overlap with entire meta-features, such as genes. This provides a broader perspective by aggregating the reads that align with all the features belonging to the same meta-feature. The default featureCounts option is "union", which is schematically illustrated below. The output of featureCounts consists of two files: a .txt file containing the actual read counts per gene (with the gene ID, the genomic coordinates of the gene, including strand and length) and a .summary file which provides an overview of how many reads could be assigned to genes and the reasons why some of them failed to be assigned.



	union	intersection _strict	intersection _nonempty
	gene_A	gene_A	gene_A
	gene_A	no_feature	gene_A
	gene_A	no_feature	gene_A
	gene_A	gene_A	gene_A
	gene_A	gene_A	gene_A
	ambiguous	gene_A	gene_A
	ambiguous	ambiguous	ambiguous

read
gene_A
gene_B

Figure 29: Schematic illustration of the different ways of handling the reads when counting features. The default option for featureCounts is union. Image retrieved from `http://www-huber.embl.de/users/anders/HTSeq/doc/count.html`.

7 Visual interpretation of normalized and transformed read counts

Since we will be using DESeq2 for our analysis, we will focus on explaining the objects and the methods this particular tool employs. To begin with, DESeq2 commonly stores almost all the experimental information in a specific R object of the DESeqDataSet class, a modified version of the SummarizedExperiment class. These classes were developed to allow storage of both numerical matrices (e.g. raw read counts) and metadata (e.g. condition of each sample), which are a standard requirement of biological studies. In particular, DESeqDataSet classes are organized as shown in the following figure.

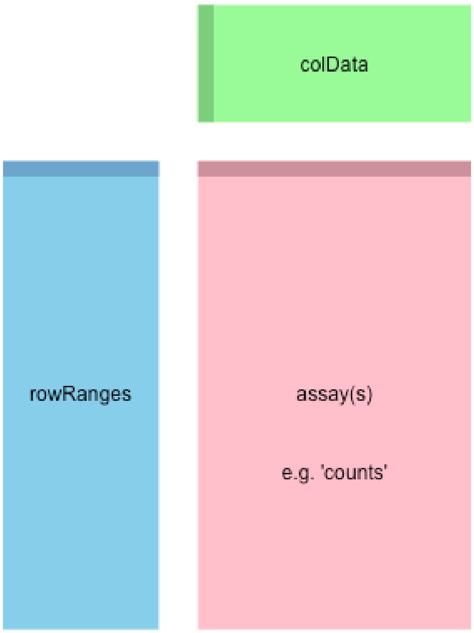


Figure 30: DESeqDataSet class. Image retrieved from: `https://www.bioconductor.org/help/course-materials/2015/CSAMA2015/lab/rnaseqCSAMA.html`

• colData: a data.frame that can contain all the variables known about the samples, like the experimental condition and the type and date of sequencing Its row.names should correspond to the unique sample names.

• RowData: keeps the information about the genes, e.g. gene ID's, their genomic ranges etc.

• assay: contains a matrix(or matrices) of the values associated with the genes and samples (e.g. raw counts).

There are different ways to construct the DESeqDataSet. In our analysis, we used the DESeqDataSetFromMatrix function with the read count matrix we prepared earlier using the featureCounts function on the alignment files. After constructing the DESeqDataSet we removed the genes without any counts.

The read counts overlapping with a particular gene cannot be directly interpreted as absolute values of the gene's expressions levels. The value obtained for each gene in a given sample is based on the number of reads corresponding to that gene. However, the efficiency of amplification and sequencing of cDNA fragments is affected by several factors that must be taken into account for the correct interpretation of the biological information hidden in the reads. In general, the number of sequenced reads corresponding to a gene depends on the length of the transcript (longer transcripts yield more fragments), the sequencing depth, the expression levels of other genes in the sample and the expression levels of the specific gene, which is the metric of interest. Therefore, in order to compare gene expression between the two conditions, the fraction of reads corresponding to each gene relative to the total number of reads and also relative to the entire RNA library must be calculated.

After normalizing the raw read counts by dividing by the size factors, computed by the estimateSizeFactors function (detailed explanation of the DESeq2 normalization procedure will follow), we visually explore the read counts through clustering and classification of the samples before performing differential testing. The majority of statistical methods employed for multidimensional data analysis, (e.g. PCA), do not work well with heteroskedastic data like those obtained from RNA-Seq experiments. For example, if PCA is directly performed on a matrix of normalized read counts, the results usually depend mainly on the few strongly expressed genes. A simple method used to overcome this issue is to take the logarithm of the normalized counts plus a small pseudo-count. Log2 transformation is usually preferred over log10 because it simplifies interpretation by focusing on doubled values. However, this method makes the lower read counts dominate the results.

DESeq2 offers the regularized-logarithm transformation (rlog) and the variance stabilizing transformation (VST) as solution to this problem. These methods are used for data visualization and other analyses that are not directly related to differential expression testing. They help reduce the amount of heteroskedasticity by stabilizing the variance across the range of mean values, making it easier to compare expression levels visually and perform other downstream analyses that benefit from a stabilized variance.

The variance-stabilizing transformation (VST) is calculated based on the fitted variance-mean relationship and then applied to the count data. This generates a table of values that are essentially homoscedastic, meaning they exhibit constant variance across the range of means. The transformation also considers the library size. In our analysis, we used the vst() function, which is a wrapper for the varianceStabilizingTransformation() function. The vst() first identifies 1000 variables that represent the variance trend of the dataset and uses this information to perform the transformation.

DESeq2's rlog() function returns values that are both normalized for sequencing depth and transformed to the log2 scale. The values are also adjusted to fit the experiment-wide trend of the variance-mean relationship in order to stabilize the variance across the range of mean values, making the data more homoscedastic. The following plots show collectively the impact of different normalization and transformation methods on the gene expression data.

By generating box plots of raw and transformed read counts across all the samples, we can see how the transformation makes even this simple graph more easily interpretable.

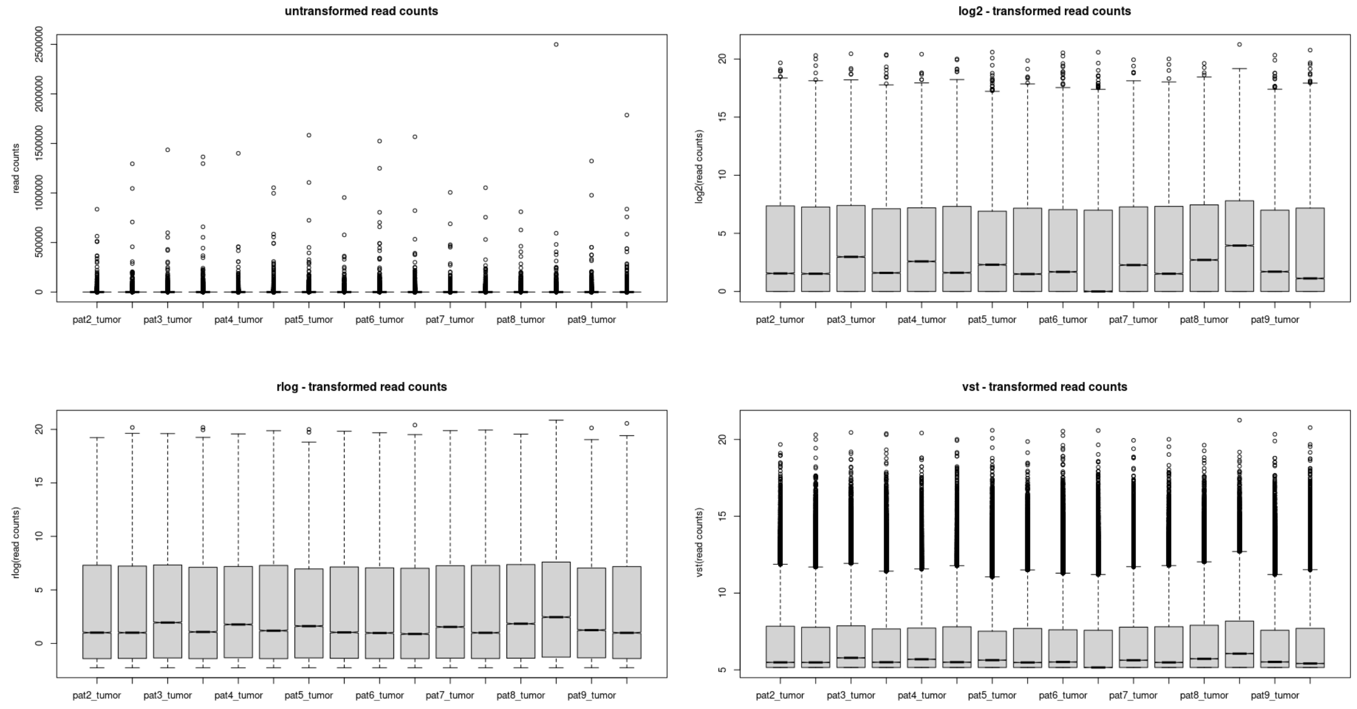
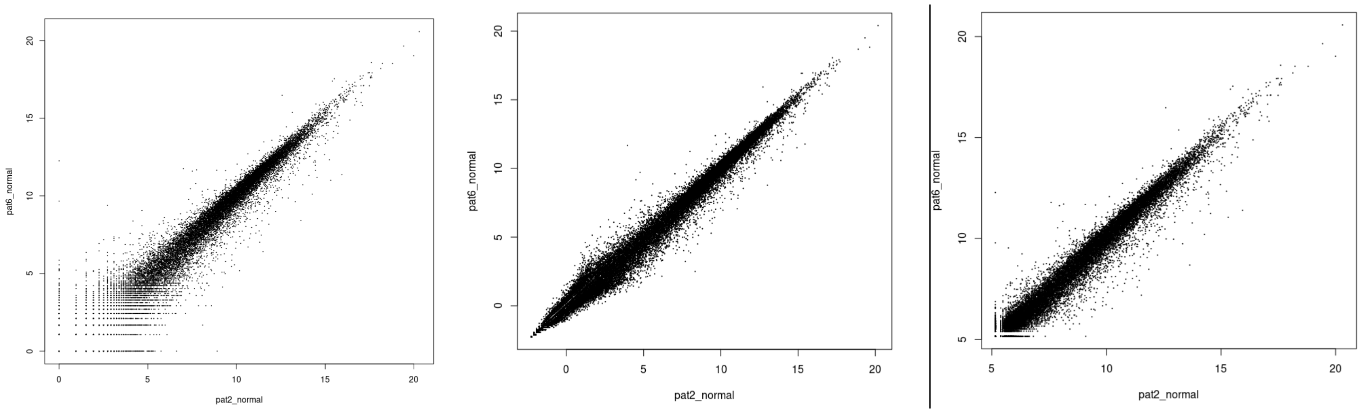


Figure 31: Boxplots comparing different transformations of RNA-Seq read counts across all the samples of our dataset. (Due to lack of space, only the labels of the tumor samples are displayed, but to the right of each of them is the corresponding normal sample of each patient).

The raw read counts exhibit a wide range of values with several extreme outliers, indicating significant variability. The log2 transformation reduces the skewness of the data, bringing the outliers closer to the center. However, there are still considerable differences between samples and some outliers remain, although less pronounced than in the raw data. The rlog transformation aims to stabilize the variance across the range of counts. The resulting distribution is more uniform across samples, with significantly fewer outliers. Similar to rlog, VST also stabilizes the variance between samples. The distribution is more consistent and normalized, but there appear to be many more outliers. The presence of outliers is not inherently bad, as it preserves some biological variability that could be important in subsequent analyses.

To get an impression of how similar read counts are between replicates, we plotted the counts in a pairwise manner. In particular we randomly chose to create the scatter plots of the normalized read counts of two different samples of the same condition (pat2_normal and pat6_normal). Each point represents a gene, with the x-axis showing the normalized read count for the patient 2 normal sample and the y-axis showing the normalized read count for the patient 6 normal sample. The plots show the normalized read counts transformed using log2, rlog, and vst, respectively, from left to right.



(a) log2 transformation

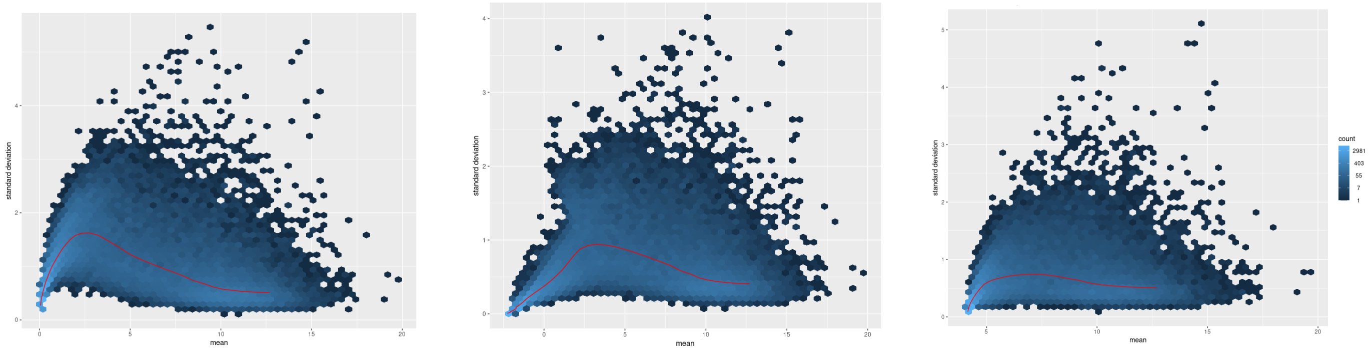
(b) rlog transformation

(c) VST

Figure 32: Pairwise comparison of the same condition samples of log2-, rlog- and vs -transformed read counts.

Points on the y = x suggest that the genes have similar expression levels in both samples whereas points deviating from the diagonal indicate genes with differential expression between the two samples. The left plot (log2 transformed data) shows greater variability for lower counts, suggesting that gene expression is less consistent for genes with lower read counts. This is due to the log transformation compressing higher counts. The middle plot (rlog) and the right plot (vst) show reduced variance and provide a more uniform distribution of points across the range of counts.

The effect of heteroscedasticity is clearly visible when we plot the per-gene standard deviation against the rank of the average expression. These scatter plot shows the relationship between the mean expression levels (x-axis) and the standard deviation (y-axis) of the normalized (and transformed here) read counts. Each point corresponds to a gene's mean expression value and its associated standard deviation. Lighter blue regions correspond to higher counts of genes while darker blue regions indicate lower counts of genes. The red line shows the general trend of the relationship between mean expression and standard deviation. The plots below are generated using the meanSdPlot function from the vsn package.



(a) log2 transformation

(b) rlog transformation

(c) VST

Figure 33: Plot of the standard deviation against the mean of log2 -,rlog- and vs- transformed read counts based on all samples of the dataset.

We do not expect all points to fall on a horizontal line. The red line shows that there is a systematic trend of variances on the mean. After log2 transformation, unequal variances persist, particularly for lower counts, which exhibit elevated standard deviation. Although the elevated standard deviation decreases after rlog transformation, variances are still not entirely uniform for all genes. The standard deviation becomes more consistent across the dynamic range after the VST. The last two transformations and especially VST ensure that the variance is more consistent across expression levels, thus enhancing the accuracy of the differential expression results.

An important step before going deeper into identifying differentially expressed genes is to test whether expectations about key global patterns are met. In this case, samples of the same condition should show similar expression patterns, whereas the expression patterns of two different pathological conditions should be more dissimilar. There are several ways to assess the similarity of expression patterns, including hierarchical clustering and principal component analysis (PCA), which we will focus on here.

In RNA-Seq, each sample is represented as a vector of read counts for thousands of genes, resulting in a high-dimensional dataset. We performed PCA which reduces the dimensionality of this dataset while retaining the most important patterns of variation. This classification method allows the high-dimensional gene expression data to be plotted in two or three dimensions according to the principal components. This visualization aids in understanding the overall structure of the dataset, by identifying groups of samples that have similar gene expression profiles and detecting outliers or abnormal samples that behave differently from the rest. Below, we present the PCA plot generated using variance-stabilizing transformed read counts. We chose to focus on this transformation because, through our analysis, we concluded that manages variance across the dataset in the most effective way compared to the other two transformations.

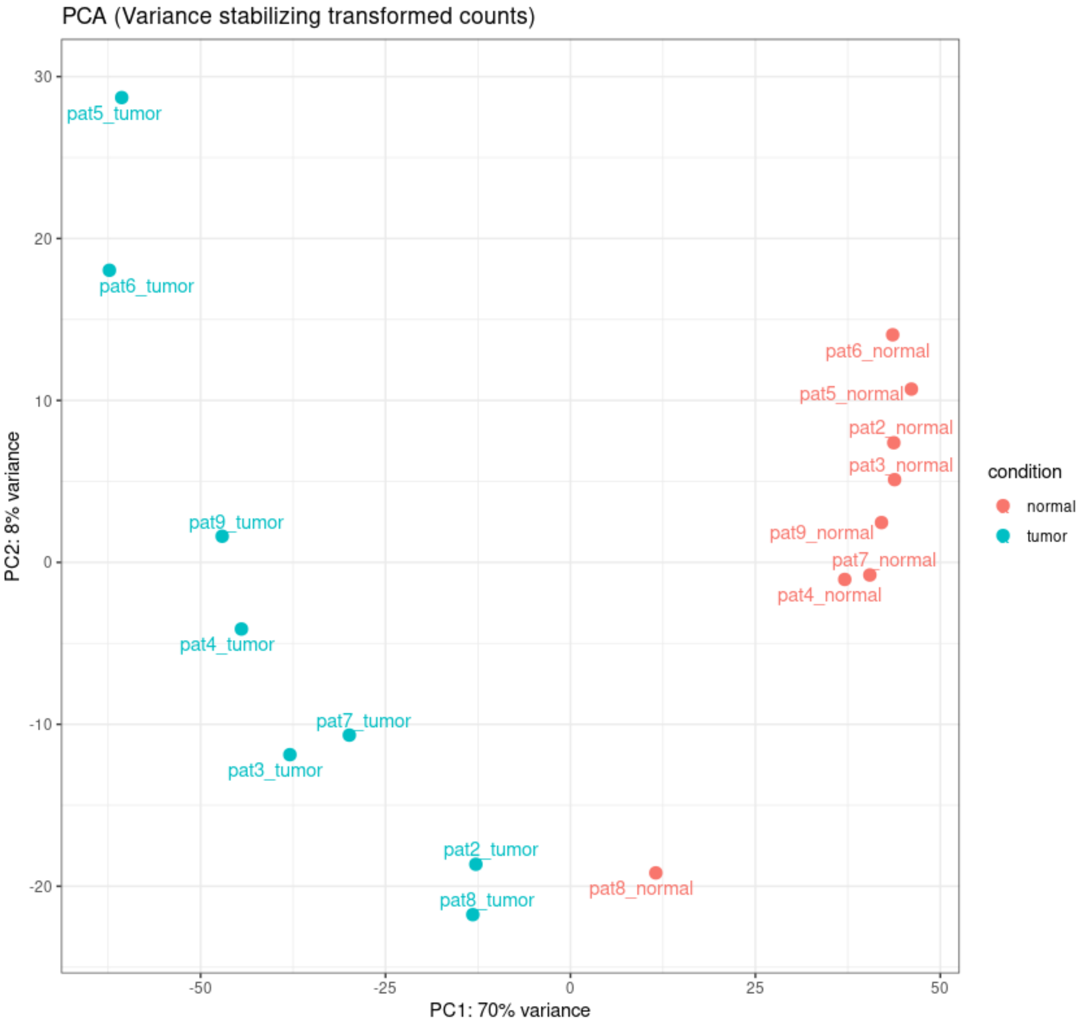


Figure 34: PCA on VST read counts with the DESEq2 function plotPCA().

This PCA plot shows the first two principal components (PC1 and PC2) of the dataset, which account for the majority of the variation in the data. We also checked different PC correlations but they did not offer additional insights into the variance structure of the data; PC1 vs PC2 classify the data best. There is a clear separation between the normal and tumor samples along PC1. This suggests that the primary source of variation (70% variance explained by PC1) in the dataset is the difference between normal and tumor samples. Normal samples, colored in red, cluster together tightly on the right side of the plot, indicating low variability in gene expression profiles. Tumor samples, colored in cyan, are more spread out but generally cluster on the left side of the plot. The normal sample for patient 8 appears somewhat isolated from the normal sample cluster, suggesting it might have distinct gene expression patterns. For the tumor samples of patients 4 and 6 they also seem to be placed further away from the corresponding cluster, but we could not draw the same conclusion as the tumor cluster is more dispersed. The greater dispersion of tumor samples highlights the inherent heterogeneity of tumor biology.

A complementary method for determining whether samples show greater variability between different conditions than between the same conditions is hierarchical clustering. Hierarchical clustering is typically based on pairwise comparisons of individual samples according to a matrix of similarity metrics, which is actually different from the actual gene expression values. We applied the dist() function to the transpose of the variance-stabilizing transformed count matrix to get sample-to-sample distances. Subsequently, we generated a heatmap of sample-to-sample distance matrix with clustering which represents to illustrate the similarity or dissimilarity between RNA-Seq samples. Each cell in the table represents the distance between two samples. The colors of the cells range from dark blue, indicating shorter distance (i.e. high similarity), to light blue, indicating high distance (i.e low similarity).

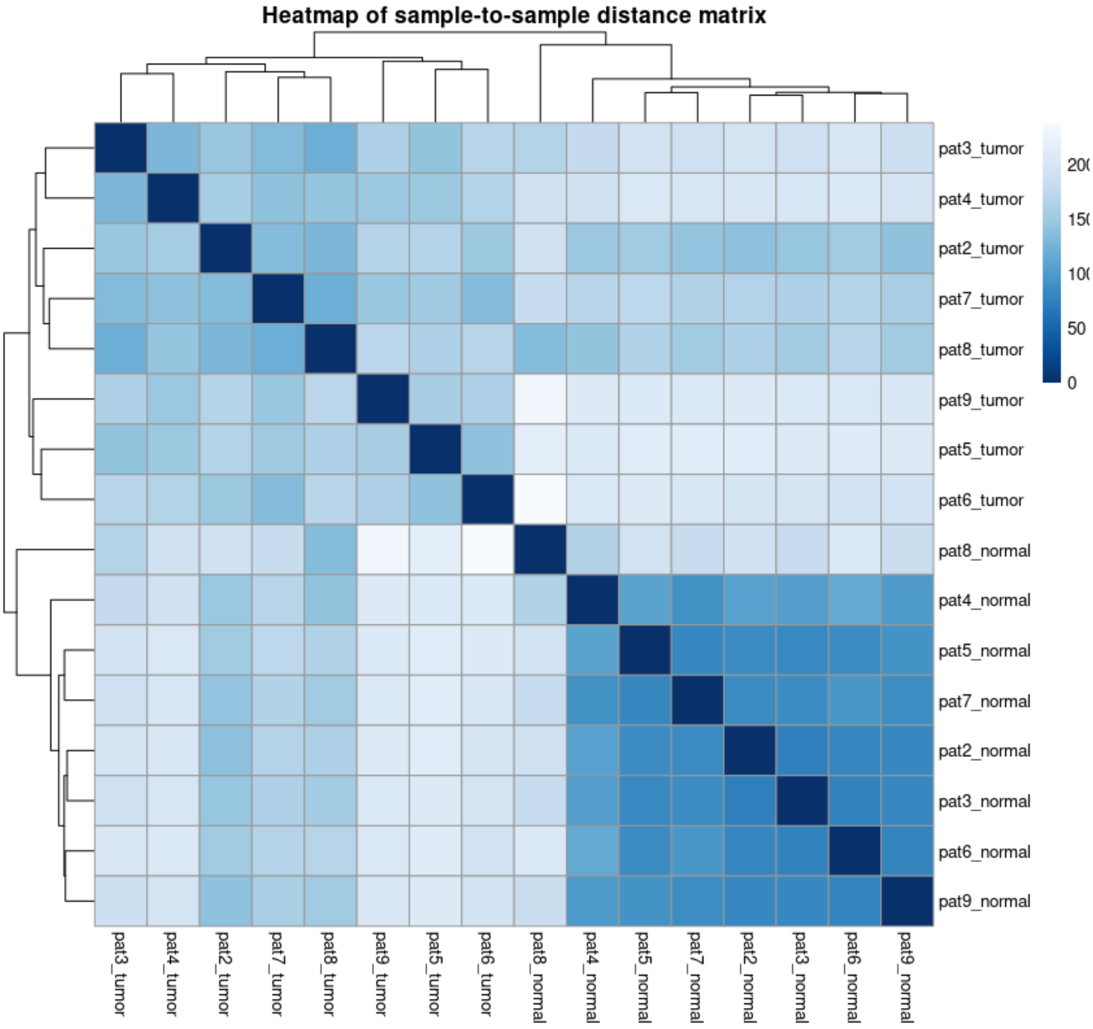


Figure 35: Heatmap of sample-to-sample distance matrix with hierarchical clustering. (The dentrogram is made with Euclidean distance).

Hierarchical clustering of the samples reveals two distinct groups, corresponding to cancer and normal samples, confirming the different expression profiles expected between these two conditions. The analysis within the same group reveals that the normal samples have a high degree of similarity to each other, as indicated by the dark blue colors in the off-diagonal elements within the group of normal samples. Correspondingly, tumor samples are also showing higher similarity within the tumor sample group. The normal sample for patient number 8 does not clearly belong to a group, as the respective cell colours are not dark enough for either. However, the distances (colors) appear more uniform within the normal patient group and we can see from the dentogram that the sample has been correctly clustered with the normal patient group. The lighter colors in the off-diagonal data that compare tumor samples to normal samples suggest greater distance, i.e. significant differences in gene expression profiles between the two conditions. This heatmap is in line with the PCA results.

8 Differential Expression Analysis

8.1 DESeq2

After analytically studying the dataset based on normalized and transformed measures of expression levels, we are interested in statistically testing our dataset, determining whether the expression of a gene varies between tumor and normal samples. The DGE tool we employed to perform this statistical analysis is DESeq2, which is one of the most popular DGE tools due to its exceptional performance [94].It is an R based tool developed by Michael Love, Wolfgang Huber, and Simon Anders and is part of the Bioconductor project. It that uses advanced statistical methods, such as shrinkage estimation for fold changes and dispersions, to improve the accuracy and reliability of differential expression analysis.

DESeq2 models RNA-Seq read counts using a negative binomial distribution. Selecting the binomial distribution, instead of normal or Poisson distributions for instance, helps account for overdispersion, which is often observed in RNA-seq data where the variance is higher than the mean. Specifically, the read counts (K_{ij}) for gene i in sample j are modeled with mean (μ_{ij}) and dispersion (α_i) parameters. The mean (μ_{ij}) of the negative binomial distribution is expressed as $\mu_{ij} = s_{ij}q_{ij}$. Here, q_{ij} represents a quantity proportional to the concentration of cDNA fragments from gene i in sample j, while s_{ij} is a normalization factor (which commonly remains constant across all genes within a sample, denoted as s_j). This last parameter is estimated using the median-of-ratios method, where $s_j = \text{median}\left(\frac{g_i}{x_{ij}}\right)$, x_{ij} is the count for gene i in sample j, and g_i is the geometric mean of counts for gene i across all samples. Normalization is crucial to ensure that observed differences in read counts reflect biological differences rather than technical artifacts. DESeq2 does not use normalization methods like RPKM, FPKM, and TPM, which are commonly utilized for adjusting differences in overall read counts among libraries. It is designed to be used in all kinds of datasets and thus the normalization can handle differences in library sizes (varying sequencing depth) and in library composition (e.g., comparing normal vs diseased).

DESeq2 employs generalized linear models (GLMs) with a logarithmic link to model the relationship between read counts and experimental conditions. The logarithmic link function is given by the formula $\log_2 q_{ij} = \sum_r x_{jr}\beta_{ir}$, where x_{jr} are the design matrix elements (in a simple two-group comparison they indicate whether a sample j belongs to one group or another) and β_{ir} are the coefficients. The GLM fit returns coefficients that indicate the overall expression level of each gene and the $\log_2$ fold change between the conditions studied.

An important feature of DESeq2 is the application of empirical Bayes shrinkage techniques to both dispersion and log fold change (LFC) estimates estimates. Shrinkage is a technique used to make these estimates more reliable by borrowing information across genes. For dispersion estimates, it is assumed that genes of similar average expression strength have similar dispersion. For log fold changes, this technique stabilizes estimates, particularly for genes with low counts or high variability, by shrinking raw log fold change estimates towards zero.

After a GLM is fit for each gene, DESeq2 tests whether the coefficients of these models (which represent the effects of different conditions on gene expression) are significantly different from zero. A coefficient considerably different from zero implies that the condition has a significant effect on gene expression. To determine the significance of the shrunken LFC estimates, DESeq2 uses the Wald test, which provides a straightforward approach for individual coefficient testing. The shrunken LFC estimate is divided by its standard error to produce a z-statistic. The Wald test compares the produced z-statistic to a standard normal distribution to obtain a p-value, indicating the likelihood that the observed effect is due to random chance.

In RNA-seq, the multiple testing problem arises when numerous statistical tests increase the risk of false positives. To control the False Discovery Rate (FDR), several methods that adjust for multiple testing are employed. However, these methods often reduce statistical power, i.e. the ability to correctly detect true differentially expressed genes. This loss of power occurs because stricter significance thresholds are applied, potentially leaving true positives undetected. DESeq2 addresses this by using the Benjamini-Hochberg procedure. This procedure filters out genes with low average expression, which are more prone to noise, and focuses on genes more likely to be differentially expressed. It selects a filtering threshold that maximizes the number of detected genes at a specified FDR, balancing the need to control false positives with the goal of detecting true positives.

8.2 Differential Expression Analysis with DESeq2

Running the DGE analysis with DESeq2 is very simple. Only by applying the DESeq() function we get the results we want. DESeq() is basically a wrapper of the functions estimateSizeFactors(), which computes the size factors according to the sequencing depth, estimateDispersions (), which estimates dispersion across all samples and nbinomWaldTest () which fits a negative binomial GLM and applies Wald statistics to each gene. DESeq2 offers a range of powerful visualization tools to help interpret and present the results (including MA charts and heatmaps).

To begin with, it is important to observe the distribution of p-values to understand the significance of the differential expression results.

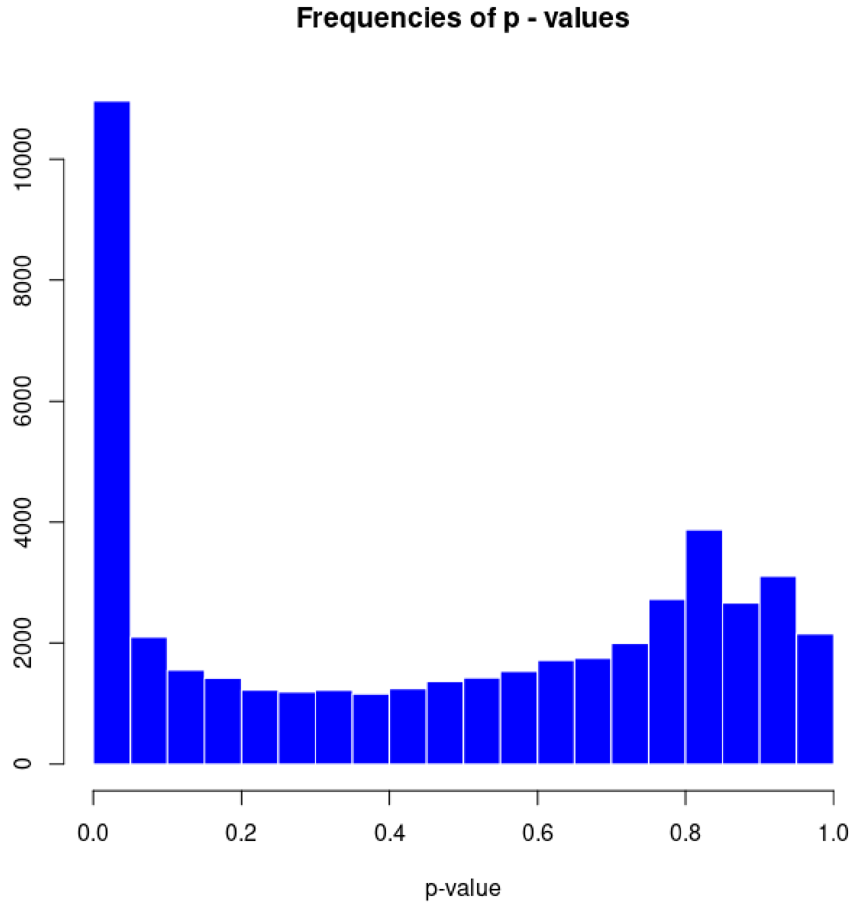


Figure 36: Histogram of p-values for all genes examined for non-differential expression between the two conditions, normal vs. tumor

The histogram of p-values reveals that the majority of p-values are very close to zero. The concentration of p-values close to zero suggests that these genes show significant changes in expression, strongly indicating their potential role in carcinogenesis. The presence of p-values around 0.8 suggests that a subset of genes do not exhibit statistically significant differences in expression between normal and cancer samples. This distribution of p-values for higher values may be due to random variation rather than a true biological effect.

The MA plot of our RNA-Seq data comparing normal and tumor samples shows a typical distribution of gene expression changes. The x-axis represents the mean expression level of each gene, and the y-axis shows the log fold change between the two conditions.

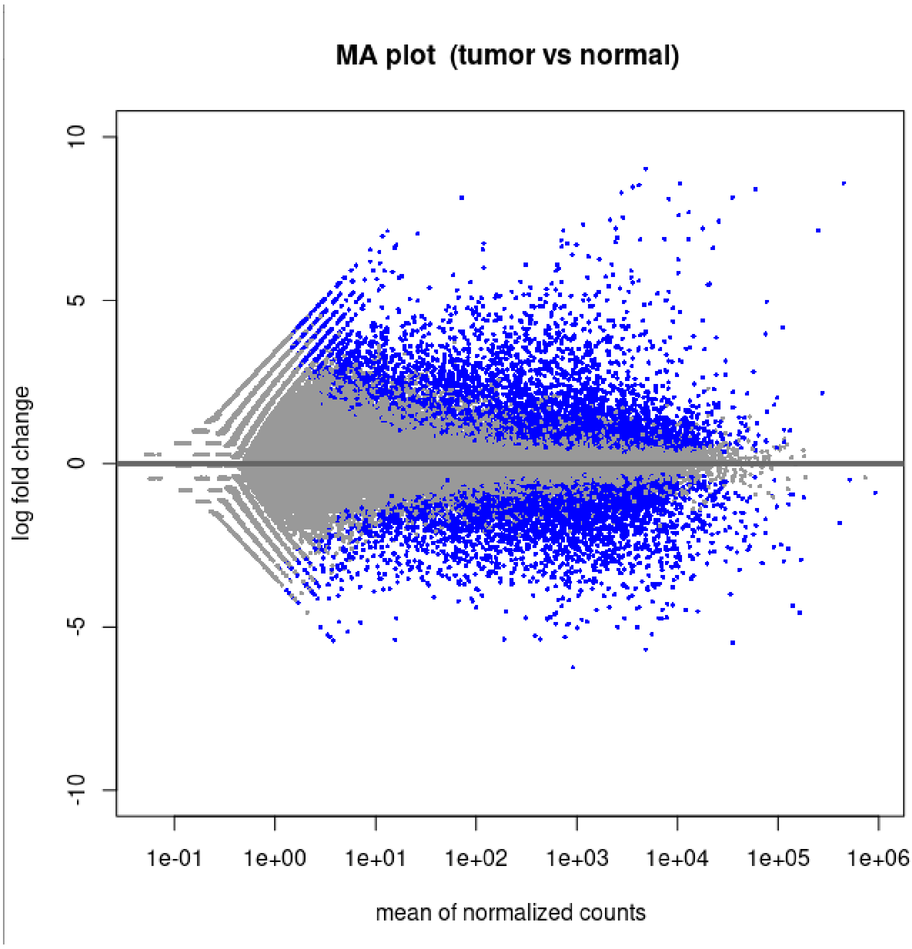


Figure 37: MA plot shows the relationship between the expression level (mean of normalized counts, A) and the fold change (log fold change, M) of genes between two conditions. Each point on the plot represents a gene. Genes with adjusted p-values <0.05 are marked in blue.

As expected, the dispersion relative to y-axis (fold change = 1, i.e. no difference in expression) is fairly even, indicating that the upregulated and downregulated genes are balanced. In a well-controlled biological experiment, a systematic bias favoring either up- or down-regulation would not be expected, unless there is a specific biological reason. The presence of many significantly differentially expressed genes (the dots highlighted in blue) between the two conditions suggests extensive transcriptional reprogramming associated with the tumor state. This diagram also suggests that mainly genes with a high average normalized number have sufficient information to generate a statistically significant result.

It is also important to consider both statistical significance (p-value) and biological significance (fold change) when comparing different conditions. This dual criterion helps in filtering out noise and focusing on biologically meaningful changes. Thus, we construct a volcano plot which is a useful tool for identifying the most significantly differentially expressed genes in RNA-Seq data.The x-axis represents the log2 transformed fold change in gene expression between the two conditions. Positive values indicate upregulation in tumor samples, while negative values are indicative of downregulation.

The y-axis represents the negative log10 transformed p- values from statistical tests comparing gene expression between conditions. Higher values have higher statistical significance.

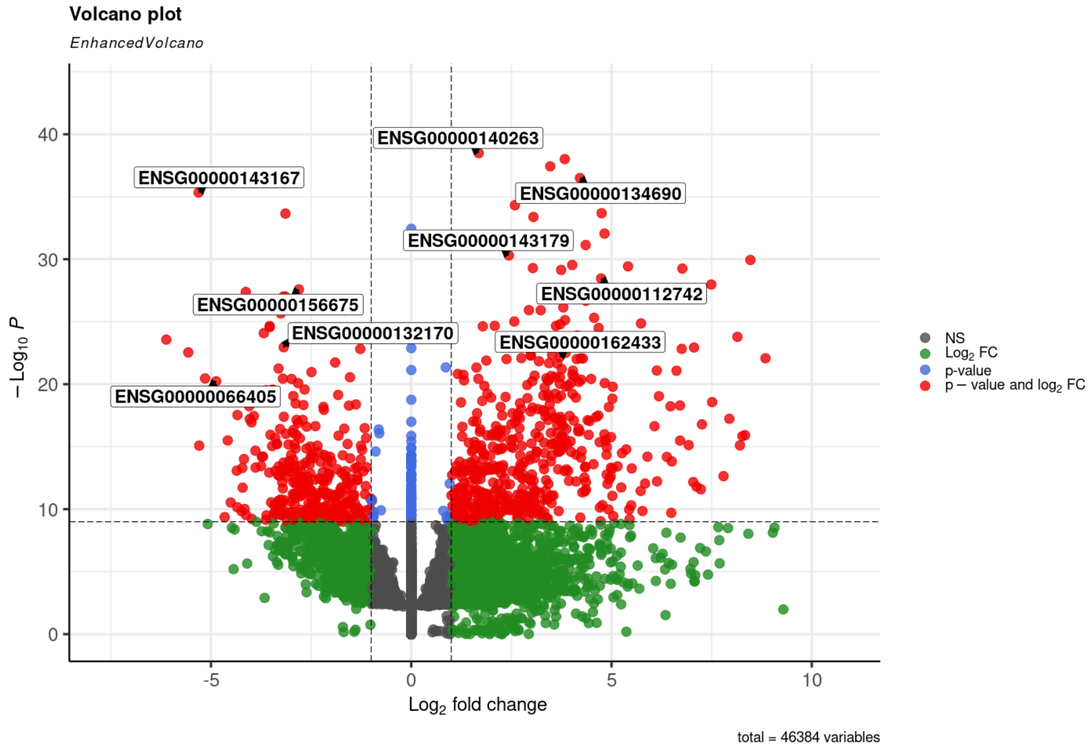


Figure 38: Volcano plot generated by the EnrichedVolcano tool from BioManager package

In this particular volcano plot red dots represent genes that are significantly differentially expressed with both high fold change and low p-value (passing the significance thresholds for both). Blue dots are genes with significant p-values but fold changes not meeting the threshold. Green dots represent genes with significant fold changes but p-values not meeting the threshold. Finally, gray dots represent genes that are not significantly differentially expressed (NS meaning Non-Significant). The dashed vertical lines represent the thresholds for log2 fold change set at -1 and 1. Thedashed horizontal line represents the p-value threshold, which were set at 10e-10. We also chose to print certain gene labels which are of particular interest on the plot.

In our volcano plot, we observe a large amount of genes with log2 fold changes greater than 1 or less than -1 (red and green dots), which can be considered biologically significant because they represent at least a twofold change in expression.. Although the $\log_1 0p$ threshold is strictly set at 9, there is a considerable number of red dots that are statistically significantly differentially expressed.

Continuously, we created heatmaps, which are a comprehensive method for examining gene expression data, and identifying significant patterns. There are numerous functions for generating heatmaps in R, including the NMF::aheatmap() which we used here. The following heatmaps for differently transformed read counts show the importance of scaling and clustering for this kind of visualization technique. The rows represent genes and the columns the RNA-Seq samples. The color scale ranges from red to blue, with warmer colors being indicative of higher expression of a particular gene in a sample, while cooler tones are indicative of lower expression levels. The genes in the heatmaps we made have an adjusted p-value<0.05, suggesting that these genes are significantly differentiated.

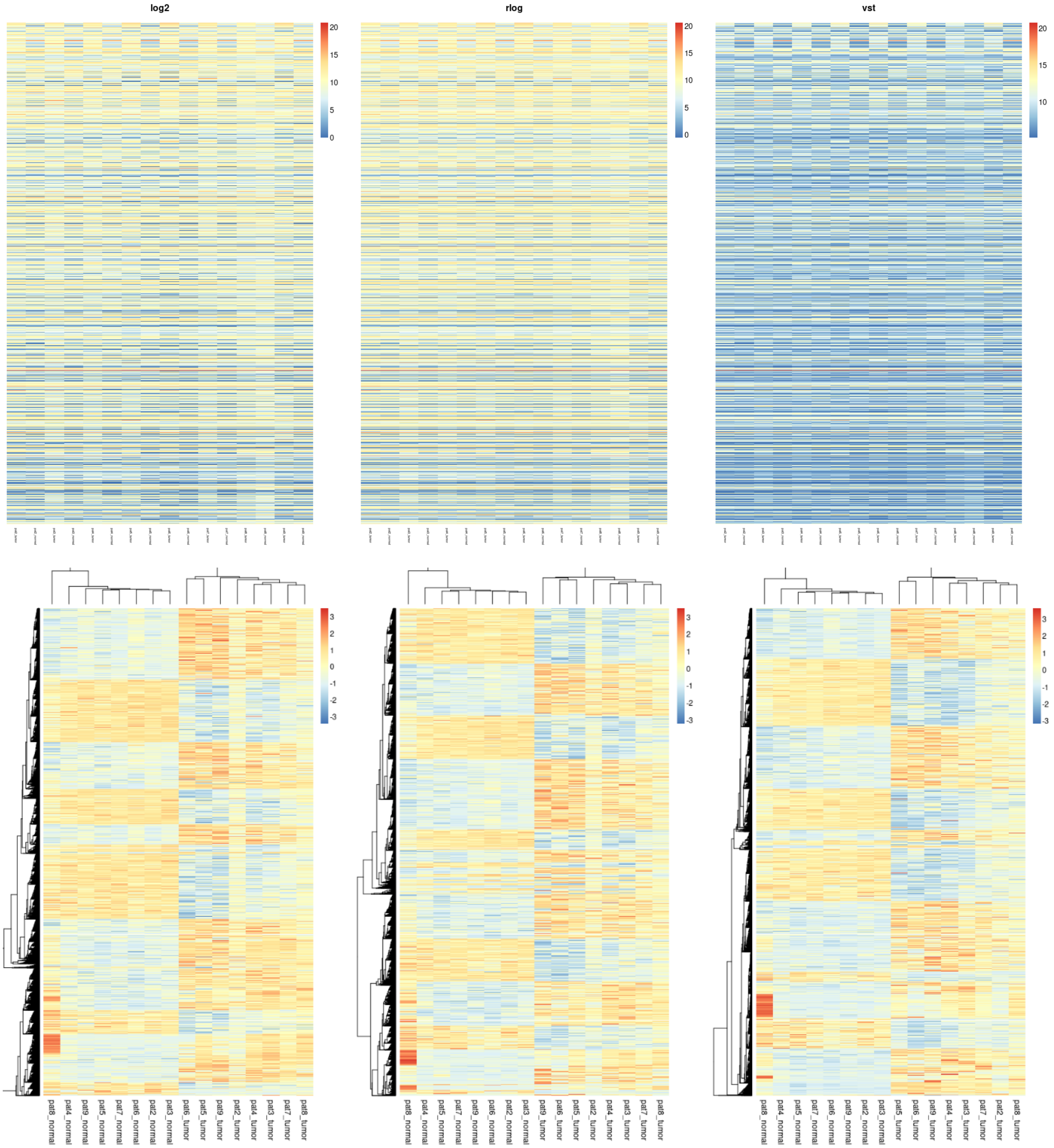


Figure 39: Heatmaps displaying log2-, rlog-, and VS-transformed read counts (from left to right) for genes with adjusted p-values less than 0.05 from the DGE analysis. In the first row, genes are sorted based on their adjusted p-values. In the final row, genes are sorted according to hierarchical clustering, with read count values scaled per row so that the colors represent z-scores instead of the raw read counts.

The first row of heatmaps show the normalized and transformed DE gene read counts sorted by adjusted p-value. As we can observe, examining gene expression solely on the basis of sorted adjusted p-values makes it impossible to interpret the plots and draw meaningful conclusions.

For the final row of heatmaps, the relative abundances have been scaled. In this case, the scaling was performed by row, i.e., on a per-gene basis. Data is basically converted into z-scores, calculated by the formula:

$$\text{z-score} = \frac{\text{actual value} - \text{mean of the group}}{\text{standard deviation}}$$

This makes it easy to identify whether a particular gene is down- or up-regulated among samples. It should be noted that with this scaling, we cannot compare genes within the same sample.

Another modification to the data involved grouping rows (genes) based on similarity. This clustering approach helped in identifying genes that exhibit the highest levels of transcription in samples from different conditions. The methods we used for determining the similarities in order to cluster the data are the Euclidean distance metric and the average as the clustering method.

Due to these modifications the resulting heatmaps, are easily interpretable. The samples are clearly clustered into two groups corresponding to cancerous and normal conditions. This clear clustering suggests that the expression profiles of the selected genes can effectively distinguish between tumor and normal tissues. Hierarchical clustering of genes reveals distinct groups with similar expression patterns. Several genes are expressed at high levels in tumor samples and at lower levels in normal samples and vice versa, suggesting their possible implication in carcinogenesis.

9 Enrichment Analysis

To elucidate the functions of the DE genes and identify patterns among them, we performed enrichment analysis. Enrichment analysis allows us to categorize the DE genes into various biological processes and molecular functions, providing a comprehensive overview of their roles in tumor cells. Moreover, by identifying significantly enriched pathways, we are able to highlight key biological themes and mechanisms potentially underlying the observed changes in gene expression in cancerous samples. The enrichment analysis also enables the discovery of regulatory networks and interaction pathways.

Our analysis includes enrichments of GO terms, molecular functions and biological processes in particular, and enrichments of certain pathways identified by KEGG. For this analysis we employed clusterProfiler. ClusterProfiler is an R package that automates the process of classification of biological terms and enrichment analysis, also offering modules for visualizing the results of the analysis, thus facilitating their interpretation. The clusterProfiler package provides two functions that follow different analytical procedures for enrichment analysis. The first one is the enricher() function, for hypergeometric testing, and the second one is the GSEA() function, for gene set enrichment analysis, both designed to accept user-defined annotation. In this project we utilized enricher() as our analyzer. It is designed to work with lists of genes categorized as significantly differentially expressed, here, based on a fold change cutoff (> 1) and a p-value threshold (<0.05). The hypergeometric testing is used to determine whether the overlap between the gene list of interest and the known pathway is greater than expected by chance. Despite the limitations, which mainly concern ignoring the magnitude of the change, we chose this approach as it is simple and at the same time yields satisfactory results.

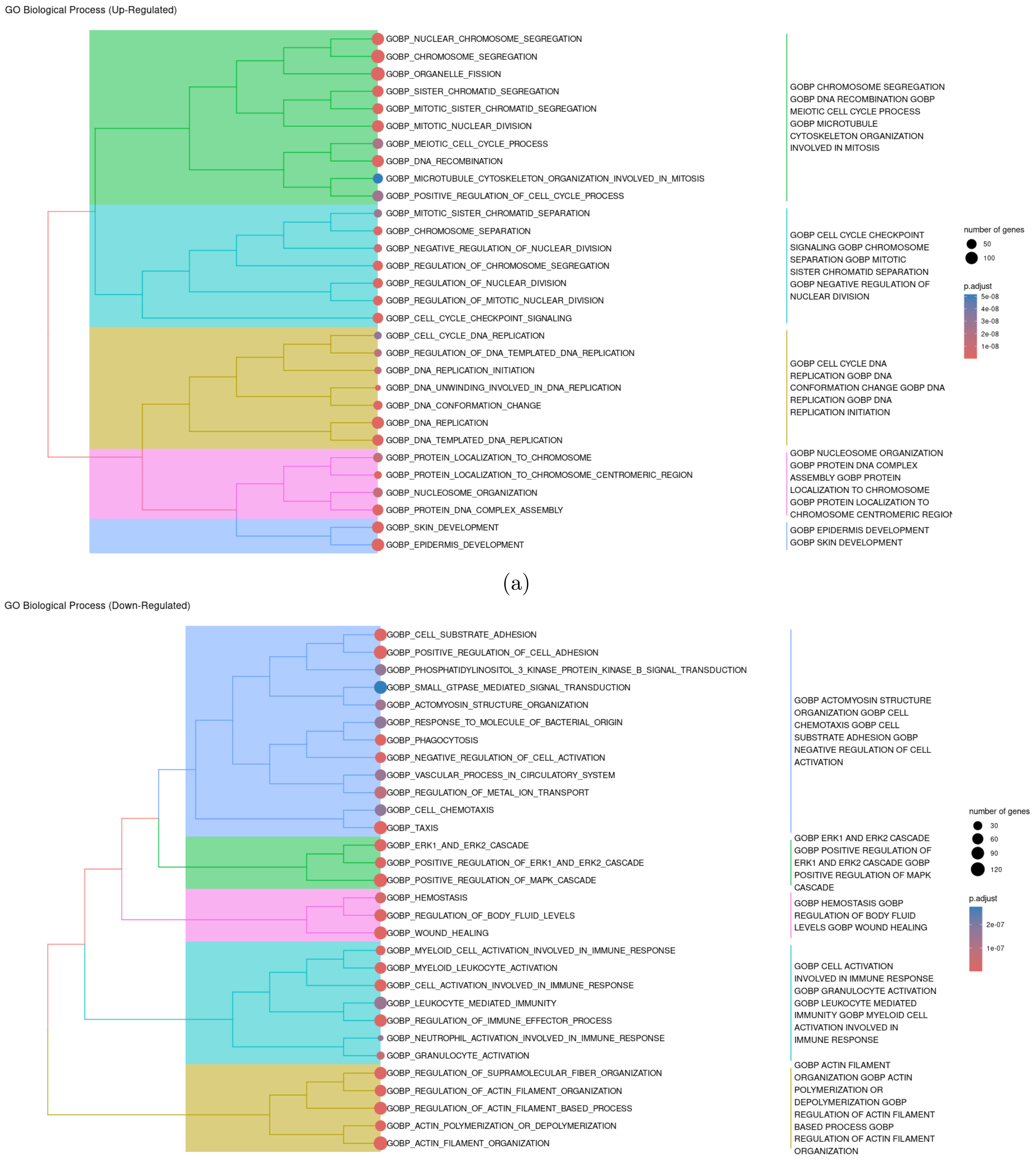
Firstly, we downloaded gmt (Gene Matrix Transposed) files of GO Biological Process, GO Molecular Function, and KEGG Medicus. This file format is commonly used to store gene sets and pathways. Since our dataset was annotated with Ensembl IDs, it was necessary to convert these to gene symbols to ensure consistency with the annotations in the GMT files. We employed both the mapIds function from the AnnotationDbi package and the biomaRt package, which are commonly used tools in bioinformatics for gene annotation. By combining the results from both tools, we successfully converted most Ensembl IDs to gene symbols. However, 293 out of 8233 genes did not have corresponding gene symbols in the databases. We proceeded with the enrichment analysis, excluding these unmatched genes.

The analysis using clusterProfiler yielded the following results for each category (GO Biological Process, GO Molecular Function and KEGG Medicus), for up- and down- regulated genes. It is crucial to clarify that, for instance, pathways identified through up-regulated gene enrichment analysis may not reflect actual up-regulation of the pathways themselves. Rather, these pathways are enriched among the up-regulated genes identified in our dataset. This enrichment suggests the presence of genes that may encode repressors or inhibitors of these pathways. Consequently, activities of these pathways and processes could potentially be down-regulated in cancerous cells compared to healthy cells. Similar considerations apply to the list of down-regulated genes.

We chose to present our results with tree plots, as they are more inclusive. They include the statistical significance of our enrichment analysis, i.e. the p-adjusted value (warmer colors are indicative of higher significance), the number of expressed genes (dot size is proportional to the expression number), as well as, the hierarchical clustering of enriched processes and pathways, making it easier to see related biological themes and groupings.

9.1 GO Biological Process

The enrichment analysis of GO Biological Processes provides insights into the underlying biological mechanisms that are enriched in tumor cells compared to normal cells. We discuss some of the prominent enriched processes, to better understand the broader biological processes and systems impacted by the DE genes (up- and down- regulated), offering valuable insights into their roles in cancer development.



(a)

(b)

Figure 40: Treeplots of enriched GO Biological Processes for (a) up- and (b) down- regulated genes

To begin with, in treeplot 40a we observe that there is significant enrichment of DNA replication pathways, which emphasizes the high proliferative capacity of cancer cells, which is a hallmark of malignancy. This indicates that cancer cells undergo rapid and uncontrolled cell division.

The proper segregation of chromosomes -the partitioning of genetic material into two daughter cells- is critical during cell division for maintaining genomic stability. Tumor cells frequently exhibit genomic instability, which can result from errors in chromosome segregation. Chromosome missegregation is related to cell death, but also to chromosome instability, and aneuploidy (an alteration in the number of whole chromosomes), which lead to cancer [154]. The enrichment of these processes suggests that tumor cells are making attempts to maintain genomic integrity despite the rapid division. This is consistent with observations that cancer cells exhibit non-random segregation of chromosomes [92], and highlights the importance of chromosomal stability and the impact of proteins that modulate it in certain types of cancer [54].

Organelle fission and mitotic nuclear division are essential processes for cell division and the equitable distribution of cellular components to daughter cells. The enrichment of these processes further supports the idea of increased cell proliferation in cancer cells. Regarding organelle fission, it is worth mentioning that increased mitochondrial fission is commonly observed in cancer cells. It is mainly mediated by DRP1 and FIS1 proteins, which are associated with increased aerobic glycolysis, a peculiar metabolic advantage of cancer cells to support their metabolic requirements and survival [8].

Furthermore, the results of our enrichment analysis highlight significant alterations in pathways associated with mitosis and cell-cycle regulation in tumor cells compared to normal cells. Mitosis is a highly regulated process crucial for maintaining genomic stability during cell division. The enrichment of mitotic pathways highlights the aggressive proliferative nature of cancer cells. While these processes are essential for cell proliferation, erroneous mitotic events have been associated with aneuploidy and polyploidy, conditions frequently observed in cancer cells [29]. Additionally, the enrichment of cell cycle control pathways highlights the necessity to manage high rates of DNA damage during cell division. The cell cycle checkpoint term, refers to the mechanisms by which the cell actively halts progression through the cell cycle until it ensures that a process, like DNA replication or mitosis, is complete. Frequent mutations in checkpoint and DNA repair genes in tumors can lead to a failure in such mechanisms, allowing cells with damaged DNA to continue dividing. This failure can significantly contribute to the development of cancer [78].

In plot, 40bb we observe that several biological processes enriched, are differing from those associated with the up-regulated genes.

The invasive nature of cancer cells is linked to increased force of attraction, high deformability and altered cell adhesion. These alterations are directly associated with the remodelling of the cell cytoskeleton mainly in the actin structure [153]. Thus, biological processes such as Actin Filament Organization and Regulation of Actin Filament Based Process are of high relevance to cancers.

In cancer cells, the enrichment of immune response-related pathways is a notable observation, often reflecting the complex interactions within the tumor microenvironment (TME). In the TME, tumor cells orchestrate a highly immunosuppressive microenvironment by secreting immunosuppressive mediators, expressing immune checkpoint ligands, and downregulating human leukocyte antigen expression [105]. Thus although tumour-infiltrating lymphocytes and natural killer cells are trying to detect and destroy malignant cells, they eventually fail. [93].

Accelerated malignant growth coincides with the massive accumulation of immature myeloid leukocytes in the TME, which explains the presence of this enriched pathway in our analysis. Ultimately, this accumulation disrupts the dynamic balance between protective T cell responses and proliferating tumor cells. Although tumors also lose recognizable antigens during their progression, they appear to remain significantly immunogenic [135].

Phagocytosis is a critical process in the immune system where phagocytes, such as macrophages and dendritic cells, engulf and destroy cancer cells. This process is important for the direct clearance of cancer cells and also plays a significant role in antigen presentation, thereby promoting adaptive immune responses against the tumor. However, cancer cells can also phagocytose other cancer cells or host cells (cell cannibalism)[19], and has also been found to phagocytose natural killer cells, eventually killing these immune cells [169]. The enrichment of the phagocytosis pathway further underscores the alterations in the complex immune processes within the TEM.

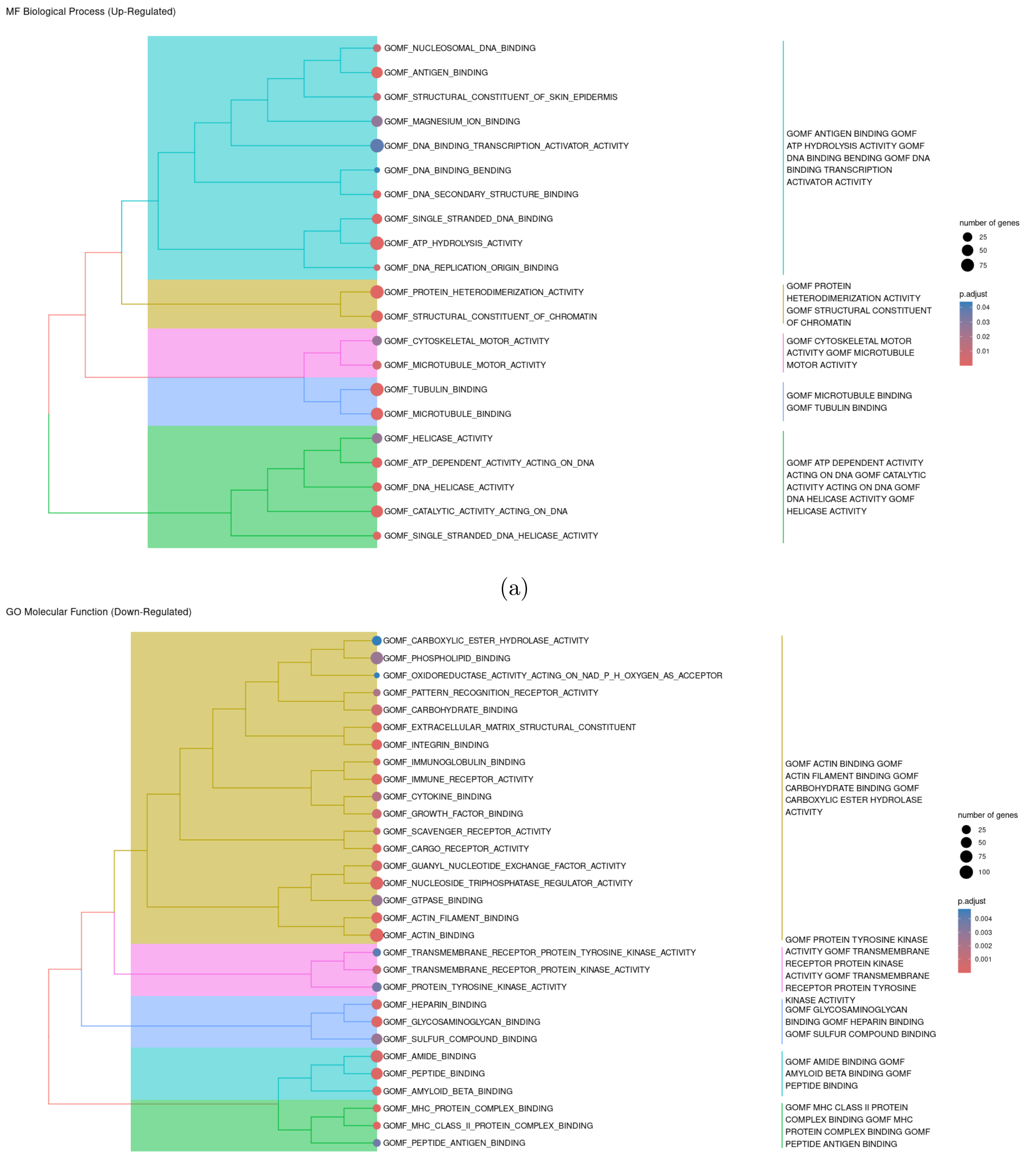
The biological process of wound healing is expected to be present in cancer cells. This is due to the fact that both wound healing and cancer are characterized by cellular proliferation, remodelling of the extracellular matrix, cell invasion and migration, formation of new blood vessels and modification of blood coagulation [33]

Among all the signalling networks, the MAPK signal transmission pathway plays an important role in controlling various physiological processes in cells, such as cell growth, development, division and death. ERK is a member of the MAPK family, and the ERK/MAPK signalling pathway is the core of the signalling network involved in regulating cell growth, development and division. The

ERK/MAPK signalling pathway is also related to tumor formation [132]. In particular, elevated ERK expression has been detected in various human tumors, including lung cancer [155].

9.2 GO Molecular Function

In this section, we highlight the results of the GO molecular functions enrichment analysis. We are focusing on some of the key enriched functions to gain a better understanding of the specific biochemical activities and their roles in squamous cell carcinoma of the lungs.



(a)

(b)

Figure 41: Treeplots of enriched GO Molecular Functions for (a) up- and (b) down- regulated genes

We begin by exploring the enriched molecular functions of up-regulated genes, as illustrated Figure 41 a.

Helicases are enzymes crucial for DNA replication, repair, recombination, and transcription. Helicase-dependent mechanisms aid cells to deal with endogenous or exogenous stress to maintain cellular homeostasis and prevent chromosomal instability. The expression of DNA helicases is up-regulated in transformed or neoplastic cells and tissues [51] and is mandatory for cancer cell proliferation. The continuous up-regulation of helicase gene expression in cancer is crucial for addressing increased DNA damage and managing replication lesions that occur in highly proliferative states. Thus, these DNA Helicase Activity functions are prominently represented in our enrichment analysis. It is worth mentioning that in contrast to the above, chromosomal instability from loss of helicase functions in inherited helicase disorders promotes carcinogenesis [15].

Catalytic activity acting on DNA includes enzymes that use ATP to catalyze reactions involving DNA, such as replication and repair. This enrichment highlights the increased need for DNA replication and repair mechanisms in rapidly dividing cancer cells. Moreover, the specific enrichment of ATP-dependent activity acting on DNA indicates a high demand for ATP in cancer cells, not only for basic survival but also for proliferation and metastasis.This is related to the fact that one of the primary causes of cancer is the shift from oxygen respiration in normal body cells to the fermentation of sugar [152].

Microtubules are vital for cell proliferation, trafficking, signalling, and migration in eukaryotic cells. The enrichment in these categories could be indicative of the increased proliferation rate of cancer cells. Notably, due to the critical role of microtubules, several microtubule-binding agents have been developed, including those aimed at cancer treatment [38].

Nucleosome binding involves the interaction of specific proteins or molecules with nucleosomes, playing a crucial role in gene regulation, epigenetic modifications, and chromatin organization. According to the literature, nucleosome-binding protein 1 has been implicated as oncogene in various types of human cancers, particularly in non-small cell lung cancer, which is the focus of our investigation in this project [89, 172]

The enrichment of DNA replication origin binding function indicates the critical role of initiating DNA replication. This reinforces the concept of enhanced DNA replication activity in cancer, as these cells require increased DNA synthesis to support their rapid proliferation.

In the GO Biological Processes enrichment analysis for down-regulated genes, we observe significant enrichment in categories related to actin and actin filament binding, as well as immune-related activities. These findings are also in agreement with the results of GO Molecular Functions analysis.

Extracellular vesicles (EVs), which are a heterogeneous collection of membrane-bound structures, are important in intercellular communication and are crucial regulators of tumorigenesis. These vesicles function largely by transporting cargo molecules (such as DNA, RNA and proteins) between tumor cells and cells in the tumor stroma [73]. The enrichment of Cargo Receptor Activity highlights the importance of cargo receptors in this process, as they are integral to the selective and efficient loading of specific molecules into vesicles. This facilitates the precise delivery of regulatory substances, emphasizing the crucial role of cargo receptors in modulating the content and function of EVs.

Integrins are fundamental cell adhesion receptors associated with signaling, mechanotransduction, and cell migration, playing a critical role in cancer progression from the development of the primary tumor to metastasis. Tumors often exhibit altered integrin expression, supporting oncogenic growth factor receptor signaling and promoting cancer cell migration [2]. Up-regulation of integrin binding partners particularly, has been associated with metastatic cancer [9], underscoring the role of integrins in enhancing metastatic potential.

Many tumous down-regulate the expression of MHC molecules (molecules that present protein fragments to CD8+ and CD4+ T cells), suggesting a role for the immune system in controlling the progression of cancer [123]. In case of impairments in the presentation or recognition of tumor-associated antigens in the context of both MHC Class I and Class II molecules (which bind the antigenic peptides in a groove in their membrane distal part), tumor cells are aided to avoid detection and destruction by the immune system. However, anti- tumor responses may unintentionally promote the growth of more competent cancer cells through a process known as immunostimulation [141]. This complex interaction between immune recognition and tumor adaptation underlines the importance of the MHC class II protein complex binding function observed in our enrichment analysis, as it reflects the complex balance between immune surveillance and tumor avoidance mechanisms.

Amyloid beta binding refers to the interaction between amyloid beta peptides and various molecules or receptors in the brain. Amyloid beta (Aβ) is a peptide that has a variety of forms and its interactions are more commonly discussed in the context of neurodegenerative diseases; mostly Alzheimer's [59]. However, Aβ have been positively associated with all cancers, as they appear particularly increased in tumor cells compared to normal cells [75].

The enrichment of the structural components of the extracellular matrix (ECM) suggests the critical role of the ECM in tumor biology. The ECM is a complex network of proteins and polysaccharides that dynamically remodels based on the function of each tissue. It provides the structural basis for tissue function and regulates various cellular processes. Cancer evolves within a dynamically changing ECM that shapes almost every aspect of cancer cell and cancer-associated stromal cell behaviour [121]. The hallmarks of cancer are heavily affected by biophysical and biochemical cues from the tumor-associated ECM [60].

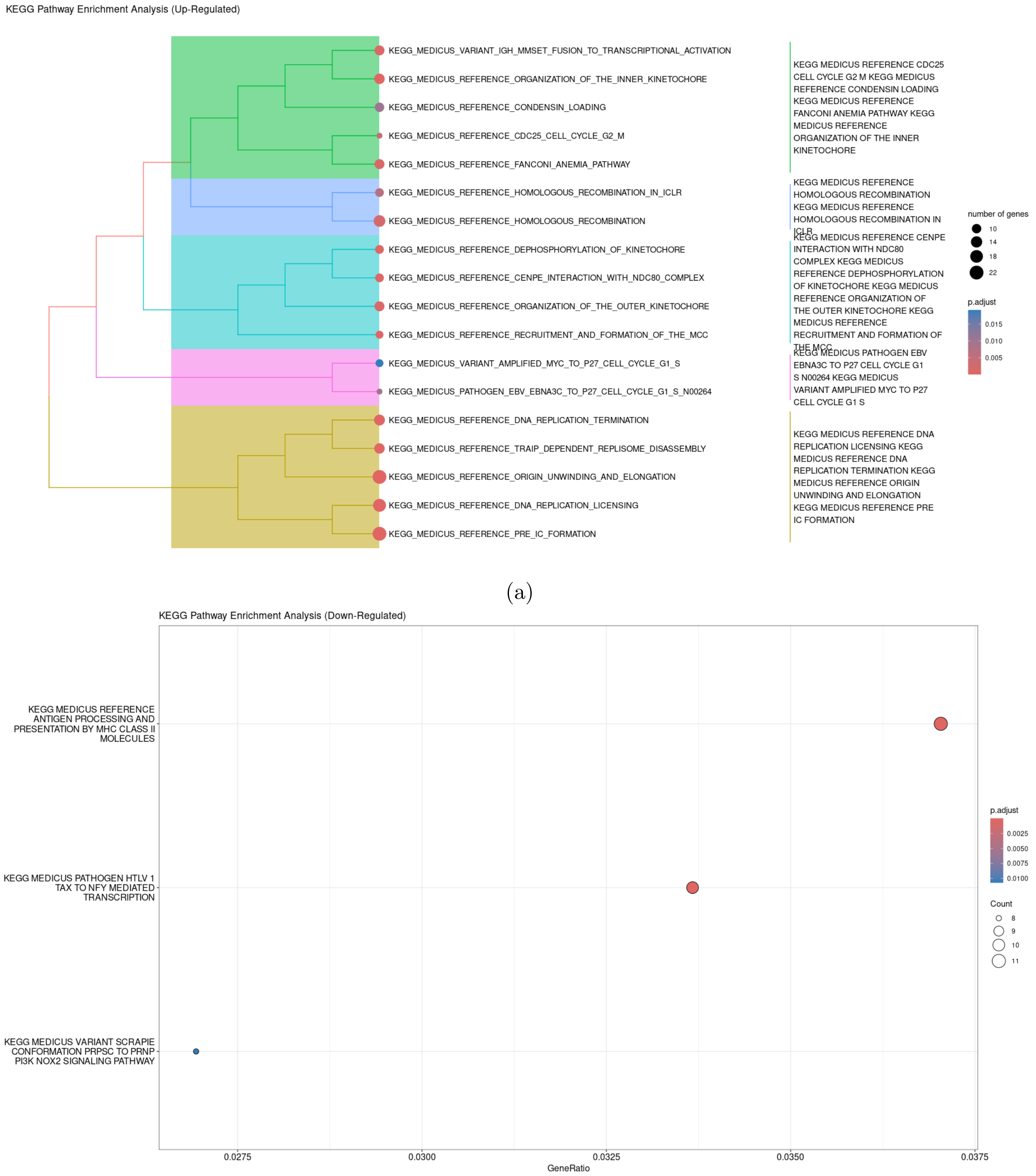
Scavenger receptors are a large family of proteins expressed mainly in macrophages and other cells of the immune system that are essential in removing pathogens, modified lipoproteins as well as dead cells. In addition to their critical role in maintaining host homeostasis, scavenger receptors have been implicated in the pathogenesis of various diseases, including atherosclerosis, neurodegeneration and metabolic disorders. They have also been found to be important regulators of tumor behavior and host immune responses to cancer [184], which explains their enriched profile in our analysis.

Transmembrane receptor protein kinase activity is a critical molecular function that involves combining with a signal molecule on one side of the cell membrane and transmitting that signal to the other side to initiate a change in cellular activity. This process is catalyzed by the reaction: a protein + ATP = a phosphoprotein + ADP. Transmembrane receptor protein kinase activity, is enriched in our data for down-regulated genes, highlighting its crucial role in cancer development. These receptors regulate cell growth and survival, and their dysregulation can lead to uncontrolled proliferation and tumorigenesis [20].

9.3 KEGG MEDICUS Pathway

Finally, KEGG MEDICUS was used for pathway enrichment analysis. KEGG MEDICUS is a specialized tool in the KEGG suite intended for medical and clinical research.Through this analysis we can identify biological pathways that are enriched in cancer cells due to differential gene expression.

Due to the small number of enriched pathways identified for down-regulated genes, it was not possible to construct a tree diagram, therefore we have included the dotplot.



(a)

(b)

Figure 42: (a)Treeplot of GO Molecular Functions for up-regulated genes (b) Dotplot of GO Molecular Functions for down-regulated genes

Initially, we examine the KEGG Medicus enriched pathways for the up-regulated genes (Figure

42 a).

Pre-replicative Complex Formation and Origin Unwinding and Elongation pathways are essential steps in DNA replication and repair. The presence of these enriched pathways is in line with the aggressive growth characteristic of tumors. The enrichment of the pre-RC formation pathway is indicative of the observed overexpression of several pre-RC proteins in cancer, which consequently makes them good tumour markers [86].

The DNA replication licensing system ensures the precise duplication of chromosomal DNA prior to cell division. When inappropriate replication origin licensing occurs in the same cell cycle, the nuclear genome is amplified. This process is known as re-replication and is usually accompanied by the appearance of DNA damage, genomic stress, or instability, which are related to cell cycle arrest, senescence, and apoptosis. Thus, although the function of replication licensing remains unclear at a certain level, is highly associated with multiple clinical pathogenesis and tumorigenesis [149].

TRAIP Dependent Replisome Disassembly is a mitotic pathway, named after the ubiquitin ligase of the same name. It is an important pathway for genome integrity in human cells [46]. More specifically, mechanisms during which the replisome is unloaded include DNA replication termination, mitosis to overcome under-replicated DNA, and ICL repair. These mechanisms are strictly regulated throughout the cell cycle, ensuring that Replisome Disassembly occurs at the proper place and time. Thus, deregulation of these mechanisms could be damaging to the maintenance of genome integrity and contribute to tumorogenesis [106].

The Organization of the Inner and Outer Kinetochore are crucial pathways for chromosome segregation and their enrichment could be associated with the high rate of cancer cell division. Kinetochore is a macromolecular complex that consists of CEN DNA and associated proteins and is primarily responsible for mediating the attachment of sister chromatids to the microtubules of the spindle, as well as for guiding the movement of chromosomes during mitosis and meiosis [111]. Kinetochore proteins are not usually mutated in cancer, but instead are dysregulated, contributing to mitotic dysfunction and chromosomal instability (CIN) [163].

Fanconi anemia (FA) pathway, is a complex DNA repair mechanism. Large-scale genomic data have revealed somatic monoallelic activation of FA genes in sporadic cancers, indicating a role in tumorigenesis [114]. Consistent with these findings, individuals with FA are predisposed to various types of cancer [53]. Additionally, recent studies suggest that the Fanconi Anemia (FA) pathway operates within a tumor-suppressor network to maintain genomic integrity by stabilizing replication forks, reducing replication stress, and regulating cytokinesis [114].

Homologous recombination (HR) is a major pathway which is also involved in the repair of DNA double-strand breaks in mammalian cells. Many cancers have mutations or epigenetic silencing of HR genes, resulting in the genetic instability that leads to cancer progression [63]. Given the significant impact of homologous recombination (HR) on cancer at multiple levels [63], it is anticipated that this pathway would be prominently enriched in our analysis.

The enriched pathway of EBNA3C, one of the Epstein-Barr virus (EBV)-encoded latent antigens, is essential for primary B-cell transformation. EBV is a ubiquitous human herpesvirus, which is associated with the development of multiple cancers [137].

The most significantly enriched pathway in the dotplot for down-regulated genes, is Antigen Processing and Presentation by MHC Class II Molecules. This molecule's function, was discussed in the GO Biological Processes subsection, as it was also found to be enriched for the down-regulated genes.

Human T-cell lymphotropic virus type 1 (HTLV-1) is the etiological agent of adult T-cell leukaemia/lympho (ATLL), which is a neoplasm of CD4+CD25+ T-cells. Tax is one of the HTLV-1 encoded proteins that interfere with viral persistence and latency. It functions as a transcription co-factor and binds to transcription factors, including NFY (Nuclear Transcription Factor Y). Tax translocates to various subcellular compartments to activate anti-apoptotic genes and deregulates the cell cycle causing genomic instability, which usually leads to cell immortality and malignant transformation.

Prion protein has two isoforms including cellular prion protein (PrP$^{\mathrm{C}}$) and scrapie prion protein (PrP$^{\mathrm{Sc}}$). Recent studies have revealed the association of PrP$^{\mathrm{C}}$ in various aspects of cancer biology. PrP$^{\mathrm{C}}$ is involved in processes related to proliferation, cell survival, invasion/metastasis and resistance to chemotherapy [1]. However, $\mathrm{PrP^{Sc}}$ is the abnormal accumulated form of the prion protein and is known to play an important role in neurodegenerative diseases, but does not appear to be associated with cancer according to the literature. Consequently, the enrichment of the Scrapie Conformation $\mathrm{PrP^{Sc}}$ to Prnp Pi3k Nox2 Signaling Pathway pathway is not easily interpretable in the case of our study.

10 Conclusions

In our project, we used RNA Sequencing (RNA-Seq) data to examine the differences between lung squamous cell carcinoma and healthy lung cells. RNA-Seq is a technique that uses the Next-Generation Sequencing (NGS) method to quantify and sequence RNA in the studied samples. NGS has enabled the recording of genetic variations in extraordinary detail, previously unimaginable. It has been used to sequence genomes in a wide variety of organisms. By comparing the genomes of all these different organisms, we have gained a much better understanding of the principles and phenomena of biology.

RNA-Seq has been widely used for more than a decade and is considered revolutionary in the field of genomics and transcriptomics. This is due to its ability to detect and capture basically all RNA transcripts, offering a broader dynamic counting range compared to similar methods. In addition, it is capable of identifying new transcripts and their isoforms and allows the quantification of alternative splicing products. Its ability to provide a comprehensive, delicate and accurate insight into gene expression makes it an invaluable tool in biological and medical research.

However, this new technology presents exceptional challenges in data analysis. RNA-Seq analysis is often considered more difficult than other methods (such as microarrays). Different analytical approaches to analyze RNA-Seq data can yield significantly different results on gene expression. These variations are largely influenced by shifts in read coverage in transcribed regions of the genome [120], an issue that RNA-Seq should overcome in the coming years. Currently, it is crucial to mention the tools, methods, and parameters selected for each analysis.

This thesis provides a detailed analysis of RNA-seq data acquired from Lung Squamous Cell Carcinoma (LUSC) and normal cells, focusing on the critical steps involved in data processing and interpretation. It details methods for mapping reads to human genome, quantifying gene expression levels, normalizing read counts to account for differences between samples, and identifying genes that are differentially expressed between tumor and normal cells.

Differential Gene Expression (DGE) analysis is a key technique in genomics that allows the comparison of gene expression levels in different conditions (e.g. as diseased versus healthy or treated versus untreated). For this step, we employed DESeq2, a widely used statistical tool for analyzing RNA-seq data, specifically designed to identify differentially expressed genes when analyzing different conditions. The results of the DGE analysis indicate that there are several statistically significant differentially expressed genes in tumor samples. It is easy to infer this, for example, by looking at the volcano diagram 38. In general, gene expression between similar conditions (normal versus normal and tumor versus tumor) is consistent, but differs significantly between different conditions. This allows us to clearly identify differential expression patterns, as shown in plots like Figure 39.

It is important to carefully execute the RNA-Seq analysis and select the correct parameters for each method, tailored to the specific dataset, so that the results of the downstream analysis can be biologically meaningful and reflect the nature of the tissues being studied. To achieve this, it is also crucial to perform rigorous quality control at each stage; first to understand the quality of the data acquired from the sequencer and then to gain insight into how each process affects the dataset.

The final step of the downstream RNA-Seq analysis is performing enrichment analysis in the differentially expressed genes to gain insight into the biological significance of the alterations in gene expression levels. We chose the databases GO Molecular Function, GO Biological Process and KEGG MEDICUS; the first two contain functional terms and the last one contains signalling pathways. The enriched biological processes in malignancies are indicative of aggressive growth and proliferation of cancer cells. They include rapid DNA replication, increased mitosis and aberrant chromosome segregation. Cancer cells also show significant changes in their cytoskeleton, which aids their motility, as well as their invasiveness. Processes similar to wound healing, such as ECM remodelling, are also enriched and facilitate tumor growth and spread.

Molecular functions that appear to be enriched in cancer cells include DNA helicases and ATP-dependent activities, as well as enhanced initiation of DNA replication, which emphasizes the elevated need for replication and repair in proliferative states. Alterations in actin binding, immune functions, extracellular vesicles, integrins and ECM remodeling highlight further the complex interplay of structural, immune and signaling pathways that enable tumor growth, metastasis and immune evasion.

The enriched pathways in lung cancer cells are also consistent with previous findings of molecular functions and biological processes. Key pathways include DNA replication and repair mechanisms, such as pre-RC formation and the DNA replication licensing system. TRAIP-dependent replisome disassembly and kinetochore organization, which are crucial for maintaining genome integrity and proper chromosome segregation during cell division are also enriched. The Fanconi anemia and homologous recombination pathways highlight the importance of DNA damage repair in cancer progression. In addition, pathways that involve viral interactions, such as EBNA3C from Epstein-Barr virus and Tax from HTLV-1, highlight the role of viral proteins in carcinogenesis.

Overall, the findings from the analysis of the RNA-Seq data are consistent with the literature and highlight the multifaceted nature of cancer biology, enhancing our understanding of LUSC, and cancer in general. They also help to identify potential biomarkers, which can serve as valuable indicators for early disease detection. These discoveries, along with the potential therapeutic directions they suggest, highlight the remarkable power and importance of RNA-Seq in advancing the understanding and treatment of diseases.

References

[1] Roland Abi Nahed, Hasan Safwan-Zaiter, Kevin Gemy, Camille Lyko, Mélanie Boudaud, Morgane Desseux, Christel Marquette, Tiphaine Barjat, Nadia Alfaidy, and Mohamed Benharouga. The multifaceted functions of prion protein (prpc) in cancer. *Cancers*, 15(20):4982, 2023.

[2] Mohammad Mohawsh Al Zeyadi. Molecular and genetic aspects of lung cancer. *Biotechnology & Biotechnological Equipment*, 27(5):4051–4060, 2013.

[3] Monther Alhamdoosh, Milica Ng, Nicholas J Wilson, Julie M Sheridan, Huy Huynh, Michael J Wilson, and Matthew E Ritchie. Combining multiple tools outperforms individual methods in gene set enrichment analyses. *Bioinformatics*, 33(3):414–424, 2017.

[4] Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Nature Precedings*, pages 1–1, 2010.

[5] Anonymous. Fastq format for sequencing reads. `https://biocorecrg.github.io/RNAseq_course_2019/fastq.html`. Accessed: 2024-06-05.

[6] Alexei A Aravin, Ravi Sachidanandam, Deborah Bourc'his, Christopher Schaefer, Dubravka Pezic, Katalin Fejes Toth, Timothy Bestor, and Gregory J Hannon. A pirna pathway primed by individual transposons is linked to de novo dna methylation in mice. *Molecular cell*, 31(6):785–799, 2008.

[7] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000.

[8] Matteo Audano, Silvia Pedretti, Simona Ligorio, Maurizio Crestani, Donatella Caruso, Emma De Fabiani, and Nico Mitro. "the loss of golden touch": Mitochondria-organelle interactions, metabolism, and cancer. *Cells*, 9(11):2519, 2020.

[9] Jorge Barbazan, Lorena Alonso-Alconada, Laura Muinelo-Romay, Maria Vieito, Alicia Abalo, Marta Alonso-Nocelo, Sonia Candamio, Elena Gallardo, Beatriz Fernandez, Ihab Abdulkader, et al. Molecular characterization of circulating tumor cells in human metastatic colorectal cancer. *PloS one*, 7(7):e40476, 2012.

[10] David P Bartel. Micrornas: genomics, biogenesis, mechanism, and function. *cell*, 116(2):281–297, 2004.

[11] Sarka Benesova, Mikael Kubista, and Lukas Valihrach. Small rna-sequencing: approaches and considerations for mirna analysis. *Diagnostics*, 11(6):964, 2021.

[12] Danny Bergeron, Laurence Faucher-Giguère, Ann-Kathrin Emmerichs, Karine Choquet, Kristina Sungeun Song, Gabrielle Deschamps-Francoeur, Étienne Fafard-Couture, Andrea Rivera, Sonia Couture, L Stirling Churchman, et al. Intronic small nucleolar rnas regulate host gene splicing through base pairing with their adjacent intronic sequences. *Genome Biology*, 24(1):160, 2023.

[13] Anthony M Bolger, Marc Lohse, and Bjoern Usadel. Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics*, 30(15):2114–2120, 2014.

[14] Tatiana Borodina, James Adjaye, and Marc Sultan. A strand-specific library preparation protocol for rna sequencing. In *Methods in enzymology*, volume 500, pages 79–98. Elsevier, 2011.

[15] Robert M Brosh Jr. Dna helicases involved in dna repair and their roles in cancer. *Nature Reviews Cancer*, 13(8):542–558, 2013.

[16] UCSC Genome Browser. Ucsc genome browser. `https://genome.ucsc.edu`. Accessed: 2024-06-05.

[17] Jason D Buenrostro, Paul G Giresi, Lisa C Zaba, Howard Y Chang, and William J Greenleaf. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, dna-binding proteins and nucleosome position. *Nature methods*, 10(12):1213–1218, 2013.

[18] James H Bullard, Elizabeth Purdom, Kasper D Hansen, and Sandrine Dudoit. Evaluation of statistical methods for normalization and differential expression in mrna-seq experiments. *BMC bioinformatics*, 11(1):1–13, 2010.

[19] Carla E Cano, María José Sandí, Tewfik Hamidi, Ezequiel L Calvo, Olivier Turrini, Laurent Bartholin, Céline Loncle, Véronique Secq, Stéphane Garcia, Gwen Lomberk, et al. Homotypic cell cannibalism, a cell-death process regulated by the nuclear protein 1, opposes to metastasis in pancreatic cancer. *EMBO molecular medicine*, 4(9):964–979, 2012.

[20] Antonio Caretta and Carla Mucignat-Caretta. Protein kinase a in cancer. *Cancers*, 3(1):913–926, 2011.

[21] Vijender Chaitankar, Gökhan Karakülah, Rinki Ratnapriya, Felipe O Giuste, Matthew J Brooks, and Anand Swaroop. Next generation sequencing technology and genomewide data analysis: Perspectives for retinal research. *Progress in retinal and eye research*, 55:1–31, 2016.

[22] Shifu Chen, Yanqing Zhou, Yaru Chen, and Jia Gu. fastp: an ultra-fast all-in-one fastq preprocessor. *Bioinformatics*, 34(17):i884–i890, 2018.

[23] Jae-Won Cho, Hyo Sup Shim, Chang Young Lee, Seong Yong Park, Min Hee Hong, Insuk Lee, and Hye Ryun Kim. The importance of enhancer methylation for epigenetic regulation of tumorigenesis in squamous lung cancer. *Experimental & molecular medicine*, 54(1):12–22, 2022.

[24] Yongjun Chu and David R Corey. Rna sequencing: platform selection, experimental design, and data interpretation. *Nucleic acid therapeutics*, 22(4):271–274, 2012.

[25] Rodrigo F Chuaqui, Robert F Bonner, Carolyn JM Best, John W Gillespie, Michael J Flaig, Stephen M Hewitt, John L Phillips, David B Krizman, Michael A Tangrea, Mamoun Ahram, et al. Post-analysis follow-up and validation of microarray experiments. *Nature genetics*, 32(4):509–514, 2002.

[26] Melina Claussnitzer, Judy H Cho, Rory Collins, Nancy J Cox, Emmanouil T Dermitzakis, Matthew E Hurles, Sekar Kathiresan, Eimear E Kenny, Cecilia M Lindgren, Daniel G MacArthur, et al. A brief history of human disease genetics. *Nature*, 577(7789):179–189, 2020.

[27] Ana Conesa, Pedro Madrigal, Sonia Tarazona, David Gomez-Cabrero, Alejandra Cervera, Andrew McPherson, Michał Wojciech Szcześniak, Daniel J Gaffney, Laura L Elo, Xuegong Zhang, et al. A survey of best practices for rna-seq data analysis. *Genome biology*, 17(1):1–19, 2016.

[28] Gene Ontology Consortium. Expansion of the gene ontology knowledgebase and resources. *Nucleic acids research*, 45(D1):D331–D338, 2017.

[29] Fergus J Couch, Xianshu Wang, William R Bamlet, Mariza De Andrade, Gloria M Petersen, and Robert R McWilliams. Association of mitotic regulation pathway polymorphisms with pancreatic cancer risk and outcome. *Cancer epidemiology, biomarkers & prevention*, 19(1):251–257, 2010.

[30] Nicholas J Croucher and Nicholas R Thomson. Studying bacterial transcriptomes using rna-seq. *Current opinion in microbiology*, 13(5):619–624, 2010.

[31] Bethina da Rocha Camargo, Vanessa das Graças Pereira de Souza, Rainer Marco López Lapa, Patricia Pintor dos Reis, and Rogério Antonio Oliveira. A decision tree-based classifier compares three data analysis methods for the identification of mirnas associated with early-stage lung cancer. *REVISTA FOCO*, 16(5):e2031–e2031, 2023.

[32] Swati Dahariya, Indira Paddibhatla, Santosh Kumar, Sanjeev Raghuwanshi, Adithya Pallepati, and Ravi Kumar Gutti. Long non-coding rna: Classification, biogenesis and functions in blood cells. *Molecular immunology*, 112:82–92, 2019.

[33] Daniel J Dauer, Bernadette Ferraro, Lanxi Song, Bin Yu, Linda Mora, Ralf Buettner, Steve Enkemann, Richard Jove, and Eric B Haura. Stat3 regulates genes common to both wound healing and cancer. *Oncogene*, 24(21):3397–3408, 2005.

[34] Benoît De Hertogh, Bertrand De Meulder, Fabrice Berger, Michael Pierre, Eric Bareke, Anthoula Gaigneaux, and Eric Depiereux. A benchmark for statistical microarray data analysis that preserves actual biological and technical variance. *BMC bioinformatics*, 11:1–14, 2010.

[35] Glynn Dennis, Brad T Sherman, Douglas A Hosack, Jun Yang, Wei Gao, H Clifford Lane, and Richard A Lempicki. David: database for annotation, visualization, and integrated discovery. *Genome biology*, 4:1–11, 2003.

[36] Sarah Djebali, Carrie A Davis, Angelika Merkel, Alex Dobin, Timo Lassmann, Ali Mortazavi, Andrea Tanzer, Julien Lagarde, Wei Lin, Felix Schlesinger, et al. Landscape of transcription in human cells. *Nature*, 489(7414):101–108, 2012.

[37] Alexander Dobin, Carrie A Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R Gingeras. Star: ultrafast universal rna-seq aligner. *Bioinformatics*, 29(1):15–21, 2013.

[38] Charles Dumontet and Mary Ann Jordan. Microtubule-binding agents: a dynamic field of cancer therapeutics. *Nature reviews Drug discovery*, 9(10):790–803, 2010.

[39] Friederike Dündar, Luce Skrabanek, and Paul Zumbo. Introduction to differential gene expression analysis using rna-seq. *Appl. Bioinformatics*, pages 1–67, 2015.

[40] Albert O Edwards, Robert Ritter III, Kenneth J Abel, Alisa Manning, Carolien Panhuysen, and Lindsay A Farrer. Complement factor h polymorphism and age-related macular degeneration. *Science*, 308(5720):421–424, 2005.

[41] Ensembl. Ensembl genome browser. `http://www.ensembl.org`. Accessed: 2024-06-05.

[42] Manel Esteller. Epigenetics in evolution and disease. *The Lancet*, 372:S90–S96, 2008.

[43] Philip Ewels, Måns Magnusson, Sverker Lundin, and Max Käller. Multiqc: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, 32(19):3047–3048, 2016.

[44] Marina Falaleeva and Stefan Stamm. Processing of snornas as a new source of regulatory non-coding rnas: snorna fragments form a new class of functional rnas. *Bioessays*, 35(1):46–54, 2013.

[45] Andrew P Feinberg and Benjamin Tycko. The history of cancer epigenetics. *Nature Reviews Cancer*, 4(2):143–153, 2004.

[46] Wanjuan Feng, Yingying Guo, Jun Huang, Yiqun Deng, Jianye Zang, and Michael Shing-Yan Huen. Traip regulates replication fork recovery and progression via pcna. *Cell Discovery*, 2(1):1–14, 2016.

[47] Jacques Ferlay, Isabelle Soerjomataram, Morten Ervik, Rajesh Dikshit, Sultan Eser, Colin Mathers, Marise Rebelo, Donald M Parkin, D Forman, and F Bray. Globocan 2012 v1. 0, cancer incidence and mortality worldwide. *Iarc Cancerbase*, 11, 2013.

[48] Marc Fiume, Vanessa Williams, Andrew Brook, and Michael Brudno. Savant: genome browser for high-throughput sequencing data. *Bioinformatics*, 26(16):1938–1944, 2010.

[49] Mario F Fraga, Esteban Ballestar, Ana Villar-Garea, Manuel Boix-Chornet, Jesus Espada, Gunnar Schotta, Tiziana Bonaldi, Claire Haydon, Santiago Ropero, Kevin Petrie, et al. Loss of acetylation at lys16 and trimethylation at lys20 of histone h4 is a common hallmark of human cancer. *Nature genetics*, 37(4):391–400, 2005.

[50] Massimo Franchini and Pier Mannuccio Mannucci. Hemophilia a in the third millennium. *Blood reviews*, 27(4):179–184, 2013.

[51] Kazunobu Futami, Sachiko Ogasawara, Hideyuki Goto, Hirohisa Yano, and Yasuhiro Furuichi. Recql1 dna repair helicase: A potential tumor marker and therapeutic target against hepatocellular carcinoma. *International journal of molecular medicine*, 25(4):537–545, 2010.

[52] Daniel R Garalde, Elizabeth A Snell, Daniel Jachimowicz, Botond Sipos, Joseph H Lloyd, Mark Bruce, Nadia Pantic, Tigist Admassu, Phillip James, Anthony Warland, et al. Highly parallel direct rna sequencing on an array of nanopores. *Nature methods*, 15(3):201–206, 2018.

[53] Juan I Garaycoechea and KJ Patel. Why does the bone marrow fail in fanconi anemia? *Blood, The Journal of the American Society of Hematology*, 123(1):26–34, 2014.

[54] Jeison Garcia and Fernando Lizcano. Kdm4c activity modulates cell proliferation and chromosome segregation in triple-negative breast cancer. *Breast cancer: basic and clinical research*, 10:BCBCR–S40182, 2016.

[55] Steven Xijin Ge, Eun Wo Son, and Runan Yao. idep: an integrated web application for differential expression and pathway analysis of rna-seq data. *BMC bioinformatics*, 19:1–24, 2018.

[56] Eugenia G Giannopoulou, Olivier Elemento, and Lionel B Ivashkiv. Use of rna sequencing to evaluate rheumatic disease patients. *Arthritis research & therapy*, 17(1):1–10, 2015.

[57] James R Griesemer. The informational gene and the substantial body: On the generalization of evolutionary theory by abstraction. In *Idealization XII: Correcting the model*, pages 59–115. Brill, 2005.

[58] Markus Hafner, Pablo Landgraf, Janos Ludwig, Amanda Rice, Tolulope Ojo, Carolina Lin, Daniel Holoch, Cindy Lim, and Thomas Tuschl. Identification of micrornas and other small regulatory rnas using cdna library sequencing. *Methods*, 44(1):3–12, 2008.

[59] Ian W Hamley. The amyloid beta peptide: a chemist's perspective. role in alzheimer's and fibrillization. *Chemical reviews*, 112(10):5147–5192, 2012.

[60] Douglas Hanahan and Robert A Weinberg. Hallmarks of cancer: the next generation. *cell*, 144(5):646–674, 2011.

[61] Jun-Mei Hao, Juan-Zhi Chen, Hong-Mei Sui, Xue-Qing Si-Ma, Guang-Qiu Li, Chao Liu, Ji-Liang Li, Yan-Qing Ding, and Jian-Ming Li. A five-gene signature as a potential predictor of metastasis and survival in colorectal cancer. *The Journal of Pathology: A Journal of the Pathological Society of Great Britain and Ireland*, 220(4):475–489, 2010.

[62] Thomas J Hardcastle and Krystyna A Kelly. bayseq: empirical bayesian methods for identifying differential expression in sequence count data. *BMC bioinformatics*, 11:1–14, 2010.

[63] Thomas Helleday. Homologous recombination in cancer development, treatment and development of drug resistance. *Carcinogenesis*, 31(6):955–960, 2010.

[64] Alan Herbert, Norman P Gerry, Matthew B McQueen, Iris M Heid, Arne Pfeufer, Thomas Illig, H-Erich Wichmann, Thomas Meitinger, David Hunter, Frank B Hu, et al. A common genetic variant is associated with adult and childhood obesity. *Science*, 312(5771):279–283, 2006.

[65] Erin E Heyer, Hakan Ozadam, Emiliano P Ricci, Can Cenik, and Melissa J Moore. An optimized kit-free method for making strand-specific deep sequencing libraries from rna fragments. *Nucleic acids research*, 43(1):e2–e2, 2015.

[66] Rolf Hilker, Kai Bernd Stadermann, Daniel Doppmeier, Jörn Kalinowski, Jens Stoye, Jasmin Straube, Jörn Winnebald, and Alexander Goesmann. Readxplorer—visualization and analysis of mapped sequences. *Bioinformatics*, 30(16):2247–2254, 2014.

[67] Radmila Hrdlickova, Masoud Toloue, and Bin Tian. Rna-seq methods for transcriptome analysis. *Wiley Interdisciplinary Reviews: RNA*, 8(1):e1364, 2017.

[68] Bo Hu, Liping Zhong, Yuhua Weng, Ling Peng, Yuanyu Huang, Yongxiang Zhao, and Xing-Jie Liang. Therapeutic sirna: state of the art. *Signal transduction and targeted therapy*, 5(1):101, 2020.

[69] Da Wei Huang, Brad T Sherman, Qina Tan, Jack R Collins, W Gregory Alvord, Jean Roayaei, Robert Stephens, Michael W Baseler, H Clifford Lane, and Richard A Lempicki. The david gene functional classification tool: a novel biological module-centric algorithm to functionally analyze large gene lists. *Genome biology*, 8:1–16, 2007.

[70] Illumina, Inc. Explore illumina sequencing technology. `https://www.illumina.com/science/technology/next-generation-sequencing/sequencing-technology.html`. Accessed: 2024-06-05.

[71] Kentaro Inamura. Major tumor suppressor and oncogenic non-coding rnas: clinical relevance in lung cancer. *Cells*, 6(2):12, 2017.

[72] Pietro Invernizzi and M Eric Gershwin. The genetics of human autoimmune disease. *Journal of autoimmunity*, 33(3-4):290–299, 2009.

[73] James Jabalee, Rebecca Towle, and Cathie Garnis. The role of extracellular vesicles in cancer: cargo, function, and therapeutic implications. *Cells*, 7(8):93, 2018.

[74] Anitha D Jayaprakash, Omar Jabado, Brian D Brown, and Ravi Sachidanandam. Identification and remediation of biases in the activity of rna ligases in small-rna deep sequencing. *Nucleic acids research*, 39(21):e141–e141, 2011.

[75] Wang-Sheng Jin, Xian-Le Bu, Yu-Hui Liu, Lin-Lin Shen, Zhen-Qian Zhuang, Shu-Sheng Jiao, Chi Zhu, Qing-Hua Wang, Hua-Dong Zhou, Tao Zhang, et al. Plasma amyloid-beta levels in patients with different types of cancer. *Neurotoxicity research*, 31:283–288, 2017.

[76] Luciane T Kagohara, Genevieve L Stein-O'Brien, Dylan Kelley, Emily Flam, Heather C Wick, Ludmila V Danilova, Hariharan Easwaran, Alexander V Favorov, Jiang Qian, Daria A Gaykalova, et al. Epigenetic regulation of gene expression in cancer: techniques, resources and analysis. *Briefings in functional genomics*, 17(1):49–63, 2018.

[77] John Karijolich and Yi-Tao Yu. Spliceosomal snrna modifications and their function. *RNA biology*, 7(2):192–204, 2010.

[78] Michael B Kastan and Jiri Bartek. Cell-cycle checkpoints and cancer. *Nature*, 432(7015):316–323, 2004.

[79] W James Kent, Charles W Sugnet, Terrence S Furey, Krishna M Roskin, Tom H Pringle, Alan M Zahler, and David Haussler. The human genome browser at ucsc. *Genome research*, 12(6):996–1006, 2002.

[80] Purvesh Khatri, Marina Sirota, and Atul J Butte. Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS computational biology*, 8(2):e1002375, 2012.

[81] Clarissa M Koch, Stephen F Chiu, Mahzad Akbarpour, Ankit Bharat, Karen M Ridge, Elizabeth T Bartom, and Deborah R Winter. A beginner's guide to analysis of rna sequencing data. *American journal of respiratory cell and molecular biology*, 59(2):145–157, 2018.

[82] Kimberly R Kukurba and Stephen B Montgomery. Rna sequencing and analysis. *Cold Spring Harbor Protocols*, 2015(11):pdb–top084970, 2015.

[83] Maxim V Kuleshov, Matthew R Jones, Andrew D Rouillard, Nicolas F Fernandez, Qiaonan Duan, Zichen Wang, Simon Koplev, Sherry L Jenkins, Kathleen M Jagodnik, Alexander Lachmann, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic acids research*, 44(W1):W90–W97, 2016.

[84] Ravi Kumar, Yasunori Ichihashi, Seisuke Kimura, Daniel H Chitwood, Lauren R Headland, Jie Peng, Julin N Maloof, and Neelima R Sinha. A high-throughput method for illumina rna-seq library preparation. *Frontiers in plant science*, 3:202, 2012.

[85] Stanley A Langevin, Zachary W Bent, Owen D Solberg, Deanna J Curtis, Pamela D Lane, Kelly P Williams, Joseph S Schoeniger, Anupama Sinha, Todd W Lane, and Steven S Branda. Peregrine: a rapid and unbiased method to produce strand-specific rna-seq libraries from small quantities of starting material. *RNA biology*, 10(4):502–515, 2013.

[86] Eric Lau, Toshiya Tsuji, Liping Guo, Shih-Hsin Lu, and Wei Jiang. The role of pre-replicative complex (pre-rc) components in oncogenesis. *The FASEB Journal*, 21(14):3786–3794, 2007.

[87] Christoph Lengauer, Kenneth W Kinzler, and Bert Vogelstein. Genetic instabilities in human cancers. *Nature*, 396(6712):643–649, 1998.

[88] Joshua Z Levin, Moran Yassour, Xian Adiconis, Chad Nusbaum, Dawn Anne Thompson, Nir Friedman, Andreas Gnirke, and Aviv Regev. Comprehensive comparative analysis of strand-specific rna sequencing methods. *Nature methods*, 7(9):709–715, 2010.

[89] Dongfan Li, Xusheng Du, An Liu, and Peng Li. Suppression of nucleosome-binding protein 1 by mir-326 impedes cell proliferation and invasion in non-small cell lung cancer cells. *Oncology Reports*, 35(2):1117–1124, 2016.

[90] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. The sequence alignment/map format and samtools. *bioinformatics*, 25(16):2078–2079, 2009.

[91] Yafang Li, Xiangjun Xiao, Xuemei Ji, Bin Liu, and Christopher I Amos. Rna-seq analysis of lung adenocarcinomas reveals different gene expression profiles between smoking and nonsmoking patients. *Tumor Biology*, 36:8993–9003, 2015.

[92] Wenyu Liu, Gajan Jeganathan, Sohrab Amiri, Katherine M Morgan, Bríd M Ryan, and Sharon R Pine. Asymmetric segregation of template dna strands in basal-like human breast cancer cell lines. *Molecular Cancer*, 12:1–10, 2013.

[93] David Loose and Christophe Van de Wiele. The immune system and cancer. *Cancer Biotherapy and Radiopharmaceuticals*, 24(3):369–376, 2009.

[94] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, 15:1–21, 2014.

[95] Lauren MacDonagh, Steven G Gray, Stephen P Finn, Sinead Cuffe, Kenneth J O'Byrne, and Martin P Barr. The emerging role of micrornas in resistance to lung cancer treatments. *Cancer treatment reviews*, 41(2):160–169, 2015.

[96] Christopher A Maher, Chandan Kumar-Sinha, Xuhong Cao, Shanker Kalyana-Sundaram, Bo Han, Xiaojun Jing, Lee Sam, Terrence Barrette, Nallasivam Palanisamy, and Arul M Chinnaiyan. Transcriptome sequencing to detect gene fusions in cancer. *Nature*, 458(7234):97–101, 2009.

[97] Jyoti Malhotra, Matteo Malvezzi, Eva Negri, Carlo La Vecchia, and Paolo Boffetta. Risk factors for lung cancer worldwide. *European Respiratory Journal*, 48(3):889–902, 2016.

[98] Marcus A Mall and Dominik Hartl. Cftr: cystic fibrosis and beyond, 2014.

[99] Marcel Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal*, 17(1):10–12, 2011.

[100] W Richard McCombie, John D McPherson, and Elaine R Mardis. Next-generation sequencing technologies. *Cold Spring Harbor perspectives in medicine*, 9(11), 2019.

[101] Paul A McGettigan. Transcriptomics in the rna-seq era. *Current opinion in chemical biology*, 17(1):4–11, 2013.

[102] Ignacio Medina, Francisco Salavert, Ruben Sanchez, Alejandro de Maria, Roberto Alonso, Pablo Escobar, Marta Bleda, and Joaquín Dopazo. Genome maps, a new generation genome browser. *Nucleic acids research*, 41(W1):W41–W46, 2013.

[103] Clifford A Meyer and X Shirley Liu. Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. *Nature Reviews Genetics*, 15(11):709–721, 2014.

[104] David FB Miller, Pearlly S Yan, Aaron Buechlein, Benjamin A Rodriguez, Ayse S Yilmaz, Shokhi Goel, Hai Lin, Bridgette Collins-Burow, Lyndsay V Rhodes, Chris Braun, et al. A new method for stranded whole transcriptome rna-seq. *Methods*, 63(2):126–134, 2013.

[105] Ikko Mito, Hideyuki Takahashi, Reika Kawabata-Iwakawa, Shota Ida, Hiroe Tada, and Kazuaki Chikamatsu. Comprehensive analysis of immune cell enrichment in the tumor microenvironment of head and neck squamous cell carcinoma. *Scientific reports*, 11(1):16134, 2021.

[106] Sara Priego Moreno and Agnieszka Gambus. Mechanisms of eukaryotic replisome disassembly. *Biochemical Society Transactions*, 48(3):823–836, 2020.

[107] Ryan D Morin, Yongjun Zhao, Anna-Liisa Prabhu, Noreen Dhalla, Helen McDonald, Pawan Pandoh, Angela Tam, Thomas Zeng, Martin Hirst, and Marco Marra. Preparation and analysis of microrna libraries using the illumina massively parallel sequencing technology. *RNAi and microRNA-Mediated Gene Regulation in Stem Cells: Methods, Protocols, and Applications*, pages 173–199, 2010.

[108] Kevin V Morris and John S Mattick. The rise of regulatory rna. *Nature Reviews Genetics*, 15(6):423–437, 2014.

[109] Kai-Oliver Mutz, Alexandra Heilkenbrinker, Maren Lönne, Johanna-Gabriela Walter, and Frank Stahl. Transcriptome analysis using next-generation sequencing. *Current opinion in biotechnology*, 24(1):22–30, 2013.

[110] Richard H Myers. Huntington's disease genetics. *NeuroRx*, 1:255–262, 2004.

[111] Kim Nasmyth. Segregating sister genomes: the molecular biology of chromosome separation. *Science*, 297(5581):559–565, 2002.

[112] Bethesda Maryland 20892 U.S. Department of Health National Institutes of Health, 9000 Rockville Pike and Human Services. Genotype-tissue expression program. `https://commonfund.nih.gov/GTEx`. Accessed: 2024-06-05.

[113] NIH. Gene expression omnibus. `https://www.ncbi.nlm.nih.gov/geo/`. Accessed: 2024-06-05.

[114] Joshi Niraj, Anniina Färkkilä, and Alan D D'Andrea. The fanconi anemia pathway in cancer. *Annual review of cancer biology*, 3:457–478, 2019.

[115] U.S. Department of Health and National Cancer Institute USA.gov Human Services, National Institutes of Health. The cancer genome atlas program. `https://www.cancer.gov/ccg/research/genome-sequencing/tcga`. Accessed: 2024-06-05.

[116] Fatih Ozsolak, Adam R Platt, Dan R Jones, Jeffrey G Reifenberger, Lauryn E Sass, Peter McInerney, John F Thompson, Jayson Bowers, Mirna Jarosz, and Patrice M Milos. Direct rna sequencing. *Nature*, 461(7265):814–818, 2009.

[117] Stephan Pabinger, Andreas Dander, Maria Fischer, Rene Snajder, Michael Sperk, Mirjana Efremova, Birgit Krabichler, Michael R Speicher, Johannes Zschocke, and Zlatko Trajanoski. A survey of tools for variant analysis of next-generation genome sequencing data. *Briefings in bioinformatics*, 15(2):256–278, 2014.

[118] Tao Pan. Modifications and functional genomics of human transfer rna. *Cell research*, 28(4):395–404, 2018.

[119] Lauren Pecorino. *Molecular biology of cancer: mechanisms, targets, and therapeutics*. Oxford university press, 2021.

[120] Jonathan Pevsner. *Bioinformatics and functional genomics*. John Wiley & Sons, 2015.

[121] Michael W Pickup, Janna K Mouw, and Valerie M Weaver. The extracellular matrix modulates the hallmarks of cancer. *EMBO reports*, 15(12):1243–1253, 2014.

[122] Anna Portela and Manel Esteller. Epigenetic modifications and human disease. *Nature biotechnology*, 28(10):1057–1068, 2010.

[123] Isabel Poschke, Dimitrios Mougiakakos, and Rolf Kiessling. Camouflage and sabotage: tumor escape from the immune system. *Cancer Immunology, Immunotherapy*, 60:1161–1171, 2011.

[124] Anto P Rajkumar, Per Qvist, Ross Lazarus, Francesco Lescai, Jia Ju, Mette Nyegaard, Ole Mors, Anders D Børglum, Qibin Li, and Jane H Christensen. Experimental validation of methods for differential gene expression analysis and sample pooling in rna-seq. *BMC genomics*, 16:1–8, 2015.

[125] Uku Raudvere, Liis Kolberg, Ivan Kuzmin, Tambet Arak, Priit Adler, Hedi Peterson, and Jaak Vilo. g: Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic acids research*, 47(W1):W191–W198, 2019.

[126] Jüri Reimand, Ruth Isserlin, Veronique Voisin, Mike Kucera, Christian Tannus-Lopes, Asha Rostamianfar, Lina Wadi, Mona Meyer, Jeff Wong, Changjiang Xu, et al. Pathway enrichment analysis and visualization of omics data using g: Profiler, gsea, cytoscape and enrichmentmap. *Nature protocols*, 14(2):482–517, 2019.

[127] Matthew E Ritchie, Belinda Phipson, DI Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth. limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic acids research*, 43(7):e47–e47, 2015.

[128] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. edger: a bioconductor package for differential expression analysis of digital gene expression data. *bioinformatics*, 26(1):139–140, 2010.

[129] Manuel Rodríguez-Paredes and Manel Esteller. Cancer epigenetics reaches mainstream oncology. *Nature medicine*, 17(3):330–339, 2011.

[130] Xavier Rogé and Xuegong Zhang. Rnaseqviewer: visualization tool for rna-seq data. *Bioinformatics*, 30(6):891–892, 2014.

[131] William N Rom, John G Hay, Theodore C Lee, Yixing Jiang, and Kam-Meng Tchou-Wong. Molecular and genetic aspects of lung cancer. *American journal of respiratory and critical care medicine*, 161(4):1355–1367, 2000.

[132] Hadara Rubinfeld and Rony Seger. The erk cascade: a prototype of mapk signaling. *Molecular biotechnology*, 31:151–174, 2005.

[133] Udo Rudloff, Umesh Bhanot, William Gerald, David S Klimstra, William R Jarnagin, Murray F Brennan, and Peter J Allen. Biobanking of human pancreas cancer tissue: impact of ex-vivo procurement times on rna quality. *Annals of surgical oncology*, 17:2229–2236, 2010.

[134] Peter J Russell and Keith Gordey. *IGenetics*. Number QH430 R87. Benjamin Cummings San Francisco, 2002.

[135] Melanie R Rutkowski, Tom L Stephen, and Jose R Conejo-Garcia. Anti-tumor immunity: myeloid leukocytes control the immune landscape. *Cellular immunology*, 278(1-2):21–26, 2012.

[136] Bhanusivakumar R Sabbula, David P Gasalberti, and Fatima Anjum. Squamous cell lung cancer. In *StatPearls [Internet]*. StatPearls Publishing, 2023.

[137] Abhik Saha, Sabyasachi Halder, Santosh K Upadhyay, Jie Lu, Pankaj Kumar, Masanao Murakami, Qiliang Cai, and Erle S Robertson. Epstein-barr virus nuclear antigen 3c facilitates g1-s transition by stabilizing and enhancing the function of cyclin d1. *PLoS pathogens*, 7(2):e1001275, 2011.

[138] Frederick Sanger, Steven Nicklen, and Alan R Coulson. Dna sequencing with chain-terminating inhibitors. *Proceedings of the national academy of sciences*, 74(12):5463–5467, 1977.

[139] Baby Santosh, Akhil Varshney, and Pramod Kumar Yadava. Non-coding rnas: biological functions and applications. *Cell biochemistry and function*, 33(1):14–22, 2015.

[140] Matthew B Schabath and Michele L Cote. Cancer progress and priorities: lung cancer. *Cancer epidemiology, biomarkers & prevention*, 28(10):1563–1579, 2019.

[141] Robert D Schreiber, Lloyd J Old, and Mark J Smyth. Cancer immunoediting: integrating immunity's roles in cancer suppression and promotion. *Science*, 331(6024):1565–1570, 2011.

[142] Nicholas J Schurch, Pieta Schofield, Marek Gierliński, Christian Cole, Alexander Sherstnev, Vijender Singh, Nicola Wrobel, Karim Gharbi, Gordon G Simpson, Tom Owen-Hughes, et al. Evaluation of tools for differential gene expression analysis by rna-seq on a 48 biological replicate experiment. *arXiv preprint arXiv:1505.02017*, 2015.

[143] GRANT SF. Variant of transcription factor 7-like 2 (tcf7l2) gene confers risk of type 2 diabetes. *Nat Genet*, 38:320–323, 2006.

[144] Frances A Shepherd, Paul A Bunn Jr, and Luis Paz-Ares. Lung cancer in 2013: state of the art therapy for metastatic disease. *American Society of Clinical Oncology Educational Book*, 33(1):339–346, 2013.

[145] A Siavoshi, M Taghizadeh, E Dookhe, and M Piran. Gene expression profiles and pathway enrichment analysis to identification of differentially expressed gene and signaling pathways in epithelial ovarian cancer based on high-throughput rna-seq data. *Genomics*, 114(1):161–170, 2022.

[146] Andrew B Singleton, Matthew J Farrer, and Vincenzo Bonifati. The genetics of parkinson's disease: Progress and therapeutic implications. *Movement Disorders*, 28(1):14–23, 2013.

[147] Temple F Smith and Michael S Waterman. Comparison of biosequences. *Advances in applied mathematics*, 2(4):482–489, 1981.

[148] Maria de Fátima Sonati and Fernando Ferreira Costa. The genetics of blood disorders: hereditary hemoglobinopathies. *Jornal de Pediatria*, 84:S40–S51, 2008.

[149] Shaoran Song, Yaochun Wang, and Peijun Liu. Dna replication licensing factors: Novel targets for cancer therapy via inhibiting the stemness of cancer cells. *International Journal of Biological Sciences*, 18(3):1211, 2022.

[150] S Spiegelman, Arsène Burny, MR Das, J Keydar, J Schlom, M Travnicek, and K Watson. Dna-directed dna polymerase activity in oncogenic rna viruses. *Nature*, 227(5262):1029–1031, 1970.

[151] Marc Sultan, Vyacheslav Amstislavskiy, Thomas Risch, Moritz Schuette, Simon Dökel, Meryem Ralser, Daniela Balzereit, Hans Lehrach, and Marie-Laure Yaspo. Influence of rna extraction methods and library selection schemes on rna-seq data. *BMC genomics*, 15:1–13, 2014.

[152] Bryan T Oronsky, Neil Oronsky, Gary R Fanger, Christopher W Parker, Scott Z Caroen, Michelle Lybeck, and Jan J Scicinski. Follow the atp: tumor energy production: a perspective. *Anti-Cancer Agents in Medicinal Chemistry (Formerly Current Medicinal Chemistry-Anti-Cancer Agents)*, 14(9):1187–1198, 2014.

[153] Mohammad Tafazzoli-Shadpour, Ehsan Mohammadi, and Elham Torkashvand. Mechanics of actin filaments in cancer onset and progress. In *International Review of Cell and Molecular Biology*, volume 355, pages 205–243. Elsevier, 2020.

[154] Kozo Tanaka and Toru Hirota. Chromosome segregation machinery and cancer. *Cancer science*, 100(7):1158–1165, 2009.

[155] Qing Tang, JingJing Wu, Fang Zheng, Swei Sunny Hann, and YuQing Chen. Emodin increases expression of insulin-like growth factor binding protein 1 through activation of mek/erk/ampkα and interaction of pparγ and sp1 in lung cancer. *Cellular Physiology and Biochemistry*, 41(1):339–357, 2017.

[156] Rudolph E Tanzi. The genetics of alzheimer disease. *Cold Spring Harbor perspectives in medicine*, 2(10), 2012.

[157] Alaa Tharwat. Principal component analysis-a tutorial. *International Journal of Applied Pattern Recognition*, 3(3):197–240, 2016.

[158] Helga Thorvaldsdóttir, James T Robinson, and Jill P Mesirov. Integrative genomics viewer (igv): high-performance genomics data visualization and exploration. *Briefings in bioinformatics*, 14(2):178–192, 2013.

[159] Cole Trapnell, David G Hendrickson, Martin Sauvageau, Loyal Goff, John L Rinn, and Lior Pachter. Differential analysis of gene regulation at transcript resolution with rna-seq. *Nature biotechnology*, 31(1):46–53, 2013.

[160] Cole Trapnell, Lior Pachter, and Steven L Salzberg. Tophat: discovering splice junctions with rna-seq. *Bioinformatics*, 25(9):1105–1111, 2009.

[161] Koen Van den Berge, Katharina M Hembach, Charlotte Soneson, Simone Tiberi, Lieven Clement, Michael I Love, Rob Patro, and Mark D Robinson. Rna sequencing data: hitchhiker's guide to expression analysis. *Annual Review of Biomedical Data Science*, 2:139–173, 2019.

[162] Erwin L Van Dijk, Yan Jaszczyszyn, and Claude Thermes. Library preparation methods for next-generation sequencing: tone down the bias. *Experimental cell research*, 322(1):12–20, 2014.

[163] Ioannis A Voutsadakis. Clinical implications of chromosomal instability (cin) and kinetochore abnormalities in breast cancers. *Molecular Diagnosis & Therapy*, 23(6):707–721, 2019.

[164] Irina O Vvedenskaya, Seth R Goldman, and Bryce E Nickels. Preparation of cdna libraries for high-throughput rna sequencing analysis of rna 5 ends. *Bacterial Transcriptional Control: Methods and Protocols*, pages 211–228, 2015.

[165] D Wang and A Farhana. Biochemistry, rna structure-statpearls-ncbi bookshelf. *Biochemistry, RNA Structure-StatPearls-NCBI Bookshelf. https://www. ncbi. nlm. nih. gov/books/NBK55*, 8999, 2022.

[166] Jinglu Wang, Dylan C Dean, Francis J Hornicek, Huirong Shi, and Zhenfeng Duan. Rna sequencing (rna-seq) and its application in ovarian cancer. *Gynecologic oncology*, 152(1):194–201, 2019.

[167] Kevin C Wang, Yul W Yang, Bo Liu, Amartya Sanyal, Ryan Corces-Zimmerman, Yong Chen, Bryan R Lajoie, Angeline Protacio, Ryan A Flynn, Rajnish A Gupta, et al. A long noncoding rna maintains active chromatin to coordinate homeotic gene expression. *Nature*, 472(7341):120–124, 2011.

[168] Likun Wang, Zhixing Feng, Xi Wang, Xiaowo Wang, and Xuegong Zhang. Degseq: an r package for identifying differentially expressed genes from rna-seq data. *Bioinformatics*, 26(1):136–138, 2010.

[169] Shan Wang, Zhen Guo, Peng Xia, Tingting Liu, Jufang Wang, Shan Li, Lihua Sun, Jianxin Lu, Qian Wen, Mingqian Zhou, et al. Internalization of nk cells into tumor cells requires ezrin and leads to programmed cell-in-cell death. *Cell research*, 19(12):1350–1362, 2009.

[170] Zhong Wang, Mark Gerstein, and Michael Snyder. Rna-seq: a revolutionary tool for transcriptomics. *Nature reviews genetics*, 10(1):57–63, 2009.

[171] J.D. Watson. *Molecular Biology of the Gene*. Always learning. Pearson, 2014.

[172] Feng Wei, Fan Yang, Xiangli Jiang, Wenwen Yu, and Xiubao Ren. High-mobility group nucleosome-binding protein 1 is a novel clinical biomarker in non-small cell lung cancer. *Tumor Biology*, 36:9405–9410, 2015.

[173] Jian-Wei Wei, Kai Huang, Chao Yang, and Chun-Sheng Kang. Non-coding rnas as regulators in epigenetics. *Oncology reports*, 37(1):3–9, 2017.

[174] Brian T Wilhelm, Samuel Marguerat, Ian Goodhead, and Jürg Bähler. Defining transcribed regions using rna-seq. *Nature protocols*, 5(2):255–266, 2010.

[175] Deborah R Winter, Steffen Jung, and Ido Amit. Making the case for chromatin profiling: a new tool to investigate the immune-regulatory landscape. *Nature Reviews Immunology*, 15(9):585–594, 2015.

[176] Markus Wolfien, Christian Rimmbach, Ulf Schmitz, Julia Jeannine Jung, Stefan Krebs, Gustav Steinhoff, Robert David, and Olaf Wolkenhauer. Trapline: a standardized and automated pipeline for rna sequencing data analysis, evaluation and annotation. *BMC bioinformatics*, 17:1–11, 2016.

[177] Thomas D Wu and Serban Nacu. Fast and snp-tolerant detection of complex variants and splicing in short reads. *Bioinformatics*, 26(7):873–881, 2010.

[178] Xi Wu, Yutian Pan, Yuan Fang, Jingxin Zhang, Mengyan Xie, Fengming Yang, Tao Yu, Pei Ma, Wei Li, and Yongqian Shu. The biogenesis and functions of pirnas in human diseases. *Molecular Therapy-Nucleic Acids*, 21:108–120, 2020.

[179] Nozomu Yanaihara, Natasha Caplen, Elise Bowman, Masahiro Seike, Kensuke Kumamoto, Ming Yi, Robert M Stephens, Aikou Okamoto, Jun Yokota, Tadao Tanaka, et al. Unique microrna molecular profiles in lung cancer diagnosis and prognosis. *Cancer cell*, 9(3):189–198, 2006.

[180] Wataru Yasui, Naohide Oue, Reiko Ito, Kazuya Kuraoka, and Hirofumi Nakayama. Search for new biomarkers of gastric cancer through serial analysis of gene expression and its clinical implications. *Cancer science*, 95(5):385–392, 2004.

[181] Yuko Yoshinaga, Christopher Daum, Guifen He, and Ronan O'Malley. Genome sequencing. *Fungal Genomics: Methods and Protocols*, pages 37–52, 2018.

[182] Matthew D Young, Matthew J Wakefield, Gordon K Smyth, and Alicia Oshlack. Gene ontology analysis for rna-seq: accounting for selection bias. *Genome biology*, 11:1–12, 2010.

[183] Guangchuang Yu, Li-Gen Wang, Yanyan Han, and Qing-Yu He. clusterprofiler: an r package for comparing biological themes among gene clusters. *Omics: a journal of integrative biology*, 16(5):284–287, 2012.

[184] Xiaofei Yu, Chunqing Guo, Paul B Fisher, John R Subjeck, and Xiang-Yang Wang. Scavenger receptors: emerging roles in cancer biology and immunology. *Advances in cancer research*, 128:309–364, 2015.

[185] W Zhai, X-D Yao, Y-F Xu, B Peng, H-M Zhang, M Liu, J-H Huang, G-C Wang, and J-H Zheng. Transcriptome profiling of prostate tumor and matched normal samples by rna-seq. *European Review for Medical & Pharmacological Sciences*, 18(9), 2014.

[186] Jin Zhang, Sizhuo Chen, and Ke Liu. Structural insights into pirna biogenesis. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, 1865(2):194799, 2022.