UNIVERSITÀ
DI SIENA
1240

University of Siena – Department of Medical Biotechnologies

Doctorate in Genetics, Oncology and Clinical Medicine

(GenOMeC)

XXXVI Cycle (2020-2023)

Coordinator: Prof. Ilaria Meloni

Characterization of diagnostic and therapeutic biomarkers using differential expression analysis technique

Scientific disciplinary sector: BIO/10 – Biochemistry; MED/08 – Pathology

Tutor:

Antonio Giordano

PhD Candidate:

Diletta Rosati

Academic Year: 2022/2023

UNIVERSITÀ LUM GIUSEPPE DEGENNARO

Dipartimento di Medicina e Chirurgia

Casamassima, October 8th, 2023

Al Coordinatore del Corso di Dottorato
Genetics, Oncology and Clinical Medicine
XXXVI ciclo
Università di Siena
Chiarissima Prof.ssa Francesca Ariani

Subject: comments on the thesis of the PhD candidate Dr Diletta Rosati

Dear Professor Ariani,

I hope my letter finds you well. First, I would like to thank you for your trust and assignment of Dr Rosati thesis entitled "*Characterization of diagnostic and therapeutic biomarkers using differential expression analysis technique*".

In her thesis work, Dr Rosati set out to identify, through RNAsequencing, genes that are differentially expressed between early lesions in mesothelioma turmorigenesis, such as benign mesothelial hyperplasias, and mesothelioma specimens. As a second task, she also studied the differential gene expression between lung adenocarcinoma and normal tissue, focusing specifically on the tumor microenvironment subpopulations.

Dr Rosati thesis is logically organized, flows well and the English language is excellent. Dr Rosati explained carefully all the methodologies used to carry out her study, demonstrating high knowledge of the technical features of most state-of-the-art bioinformatic techniques, of their advantages and limits. She also demonstrated to master these methods skilfully being able to retrieve list of genes that have the potential to serve as diagnostic/prognostic markers for mesothelioma and lung cancer and, likely, important therapeutic strategies.
For several of the identified genes indeed, data are already present in the literature supporting Dr Rosati findings.
Overall, the PhD candidate study addressed two relevant issues and her original study provided data that hold potential for translation to the clinical setting.

The study present different strengths:

- RNA sequencing is a powerful high-throughput technique to investigate gene expression;
- There are currently few markers able to discriminate malignant pleural mesothelioma and mesothelial hyperplasia; the newly identified genes might serve also as future targets for secondary preventive strategies;
- The bioinformatic ability to dissect the important role of the tumor microenvironment is a crucial aspect of diagnostic/prognostic/therapeutic factors.

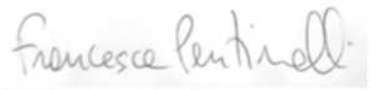
The study also presents some weaknesses:

- The set of clinical samples is quite restricted and might fail in identifying the huge complexity of different patients' specimens (as recognized by the candidate herself);

- At present wet lab functional experiments which validate the candidate's finding are lacking; however, the candidate retrieved many evidence from the literature which further support her data.

So, despite these limitations, the candidate managed to produce promising results which might have a clinical impact.
I support the candidate defence of this thesis work to pursue the PhD title successfully.

Sincerely



Francesca Pentimalli
Professore Associato di Patologia Generale
Dipartimento di Medicina e Chirurgia
Università LUM Giuseppe Degennaro

REGIONE ABRUZZO
Azienda Sanitaria Locale n. 1
Avezzano, Sulmona, L'Aquila
Presidio Ospedaliero San Salvatore
U.O.C. Oncologia Medica
Direttore: Prof. Luciano Mutti

L'Aquila, 18 Oct 2023

To the Director of PhD Programme
Genetics, Oncology and Clinical Medicine
University of Siena

Subject: PhD candidate Dr Diletta Rosati.

Dear Professor Ariani,

it has been my pleasure to review this dissertation

Dr. Rosati has conducted a project aimed at comparing RNA sequencing genes in paired mesothelioma tissue vs. benign mesothelial hyperplasia,

In her side study she also studied gene expression between lung adenocarcinoma vs normal tissue

Dr. Rosati has mastered the techniques and properly applied them in the right direction to achieve the goal of the study

Her study addressed relevant issues which provide potential translational implications whereas the dissertation is well designed and written

The results generated are intriguing and certainly pave the avenue for further studies i.e. the role of fatty acids in Mesothelioma vs hyperplastic mesothelium

If any, the number of samples is small although this limitation is almost unavoidable due to the difficulty of gathering paired in tissue in patients with Mesothelioma

Am definitively in favor to endorse the successful defence of this thesis

The Director
Luciano Mutti

INDEX

Summary

The search for effective diagnostic and therapeutic markers in oncology has emerged as an essential tool for guiding medicine to new frontiers in recent years.

This thesis explores the crucial search for those markers toward more precise diagnoses and more targeted treatments, with the goal of increasing therapy success in the near future. Among the many challenges in oncology is that some diseases are difficult to diagnose and treat. Pleural mesothelioma and lung adenocarcinoma are two emblematic examples of these challenging diseases.

Pleural mesothelioma is a rare tumor located in the lung pleura, and, although there are also rare asbestos-independent cases associated with mutations in tumor suppressors like BAP1 (1), it is mainly related to asbestos exposure. Its early diagnosis is complicated not only by vague and nonspecific symptoms but also by the challenge of distinguishing it from benign, precancerous lesions, such as mesothelial hyperplasia. This similarity can delay proper diagnosis. The search for specific diagnostic markers for pleural mesothelioma is therefore essential, thus offering hope for early diagnosis to patients (2).

Lung adenocarcinoma represents the most common form of lung cancer. Like many others, this tumor is not a monolithic entity but rather a complex ecosystem in which tumor cells and the surrounding tumor microenvironment can interact. A deeper understanding of this microenvironment is essential for developing targeted therapeutic strategies against the microenvironment to promote an anti-tumor response. In light of this complexity, the search for specific therapeutic markers to guide these targeted treatments gains greater relevance, with the aspiration to offer patients more effective therapies (3).

The first part of this thesis, the Main Project, aims to identify variations in genetic transcripts through differential expression analysis between malignant pleural mesothelioma and mesothelial hyperplasia to identify potential effective diagnostic and therapeutic biomarkers.

The first chapter of this section explores bioinformatics as a key tool for identifying molecular differences between two conditions. The essential steps in the planned bioinformatics pipeline will be presented, from data collection and preprocessing to differential expression analysis and biological interpretation of the results.

Chapter 2 outlines the key objectives of our research, while Chapter 3 outlines the bioinformatics methodologies and tools used to guide the study toward identifying differentially expressed genes. In Chapter 4, we present the results that emerged from the analysis. The fifth Chapter, meanwhile, reflects on the conclusions drawn, followed by Chapter 6, which looks ahead, exploring future prospects and suggesting possible directions for further research and applications based on the results of this study.

The second part of this thesis, the Side project, aims to identify variations in genetic transcripts through differential expression analysis within the lung adenocarcinoma immune population compared to control samples to identify potential therapeutic markers for innovative and targeted future therapies.

The chapter 1 of this section introduces lung adenocarcinoma and provides insight into the importance of the relationship between the neoplasm and its microenvironmental context. Chapter 2 defines the objective, followed by the third chapter describing the bioinformatics methods and approaches used. Chapter 4 lays out the results, and Chapter 5 reports the techniques and future prospects of this study.

Main project

1. Introduction

1.1. Malignant pleural mesothelioma

Malignant pleural mesothelioma (MPM) is a rare, heterogeneous, aggressive tumor stemming from the mesothelial cells that cover the pleura, the protective membrane surrounding the lungs (4). It includes three main histological subtypes with different frequencies. The epithelioid subtype is the most common, accounting for 70-85% of all MPM cases. Epithelioid mesothelioma has a better prognosis than other subtypes because they tend to grow more slowly and respond relatively better to treatment. The median survival of epithelioid MPM is about 14 months. The sarcomatoid subtype constitutes about 10% of MPM cases and it is associated with worse prognosis. Its cells have a spindle-shaped appearance and are more aggressive, making this subtype less responsive to treatment. The median survival in this case is only about 3 months from diagnosis. Finally, the biphasic histotype comprises about 10-25% of cases and contains a combination of epithelioid and sarcomatoid cells in different percentages. The prognosis of patients with biphasic MPM is influenced by the specific ratio of each cell type and its characteristics (5).

Malignant pleural mesothelioma is mainly (~80%) caused by exposure to asbestos fibers (6). The first case of asbestos-related MPM was recorded in the United States in 1967, following a pandemic of MPM among miners, which helped establish the link between asbestos exposure and the development of the disease (7). After the second half of the 20th century, due to the indiscriminate use of asbestos in the last century, there was an increase in cases of this rare disease worldwide (7). Although the frequency is currently fairly constant in some countries, such as the United States (3,200 cases per

year) (8), it is increasing in Europe, where a peak of cases has been estimated between 2020 and 2025 (9). Specifically, its overall incidence in recent years in Italy is about 1500-1600 new cases/year (10).

However, globally, the incidence of MPM is still considered low, with an estimate of less than 1 percent. In 2020, the World Health Organization (WHO) detected 30,870 new cases globally, encompassing all ages and both sexes. Of these, 44% were recorded in Europe, 31.5% in Asia, and 13.3% in North America (11).

It should be noted, however, that most likely, the percentages may be even higher considering the paucity of data provided by some countries (12). While European nations such as England, the Netherlands, and Denmark have a long tradition of incidence/mortality monitoring since as early as the 1960s, and Italy has a national registry dedicated to monitoring mesothelioma cases within the country (RENAM) (13) countries in Eastern Europe, Asia, South America, and Africa contribute far less to case detection (14). Note that many of these countries are the same ones that still persist in the use of asbestos today. In fact, even though the latter has been banned in more than 55 countries, it is still widely used: suffice it to mention that more than 2 million tons of asbestos are produced worldwide each year (15).

The aggressive symptomatology of mesothelioma is usually characterized by increasing dyspnea caused by pleural effusion, which may or not be accompanied by non-pleural chest discomfort produced by chest wall invasion. Nonproductive cough, fever, asthenia, hypoxia, weight loss, or night sweats can also be present (16). Symptoms usually appear gradually and over a period of time ranging from 3 to 6 months between the actual onset of tumor growth and diagnosis (17).

The difference between the latency period and growth time in the context of MPM is crucial. While the latency period, which is the time between first exposure to

carcinogenic materials such as asbestos and diagnosis, is very long (30-40 years), the time between the actual onset of tumor growth and diagnosis is much shorter (18). In contrast to slow-growing tumors, which can be diagnosed and treated in early stages, MPM is aggressive and develops rapidly, showing clinical symptoms within a time frame of months or years, not decades (4).

The delayed onset of symptoms is the main reason why this rare disease is often detected at an advanced stage. The lack of a reliable screening approach to detect early-stage disease complicates its diagnosis (4).

In addition, treatment options need a multimodal strategy in an attempt to maximize success (19). In cases of early-stage disease, surgery is an option to reduce tumor mass and improve the outcome of subsequent chemo-radiotherapy treatments. Two types of surgical procedures are common: pleurectomy-decortication (PD), which entails the surgical excision of both the parietal and visceral pleura, along with the pericardium and diaphragm; and pleuro-pneumonectomy, which also involves the removal of the lung from the affected side. Chemotherapy is another treatment approach and it is usually a combination of cisplatin and pemetrexed. Recently, immunotherapy, based on the combination of nivolumab and ipilimumab immune checkpoint inhibitors, has been approved in the first-line setting, although its use is currently highly debated (20,21). Finally, radiation therapy can be used for both pain management in advanced disease and to prevent tumor recurrence at sites where drainage or thoracoscopic procedures have been performed (19).

The overall prognosis is generally dismal due to age, performance status, or comorbidities at the time of diagnosis, with a median survival ranging from 3 to 14 months, depending also on the histologic subtype (9,22).

1.1.1. Mesothelial hyperplasia and differential diagnosis with Mesothelioma

During the latency period of asbestos exposure, which can last several decades, various pre-cancerous lesions can occur in the mesothelium, including mesothelial hyperplasia. Mesothelial hyperplasia is a benign proliferation of mesothelial cells occurring in response to asbestos exposure (23).

It can be challenging to detect, as it often causes no symptoms and sometimes is not even visible on imaging tests. However, mesothelial hyperplasia is a risk factor because can progress into malignant mesothelioma (24).

Discriminating between malignant pleural mesothelioma (MPM) and mesothelial hyperplasia (MH) can be difficult as, in fact, they both present with cytologic atypia, increased cellularity, and mitosis (25). The current differential diagnosis between MPM and mesothelial hyperplasia requires tissue biopsy and pathologic examination. After biopsy, pathological analysis is based on the search for tissue invasion (26). Specific diagnostic markers such as calretinin, *WT-1* (Wilms' tumor 1), cytokeratin ⅝, but also *BAP1* (BRCA1-associated protein-1) and *CDKN2A* (p16) are commonly analyzed for the diagnosis of malignant pleural mesothelioma (25).

BAP1 functions as a ubiquitin hydrolase and plays a multifunctional role in various cellular processes, including DNA repair and cell cycle regulation. In contrast, *CDKN2A*, also known as p16, is a cyclin-dependent kinase inhibitor that plays a direct and critical role in cell cycle regulation. Specifically, it inhibits the action of cyclin-dependent kinase 4 (CDK4), thereby halting the progression from G1 phase to S phase within the cell cycle (27).

These genes show marked functional convergence in the context of the cell cycle, and this functional similarity underscores the importance of cell cycle control in the pathogenesis of malignant pleural mesothelioma (28).

However, the sensitivity of these markers in diagnosing malignant pleural mesothelioma is not always accurate, even when tested together. Thus, there is currently disagreement within the scientific community regarding the ability of these markers to differentiate pleural tumors from benign lesions (29) accurately.

It is necessary, therefore, to identify more sensitive biomarkers for the differential diagnosis of MPM (29).

To untangle this complex topic, we employed RNA sequencing (RNA-Seq) to understand the variation in gene expression occurring during the transition from mesothelial hyperplasia to mesothelioma and identify, therefore, biomarkers useful for differential diagnosis.

The following section outlines the strengths and features of this technology.

1.2. RNA Sequencing

The unveiling of the double-helix configuration of DNA in 1953 revealed the fundamental process of gene interaction (30). This laid the foundation for molecular biology, where genes within DNA are transcribed into RNA, which, in turn, guides the production of proteins. This process underlies several biological and pathological mechanisms (31). As a result, RNA analysis is crucial to comprehending tumor gene expression regulation (31). The complex genomic structures comprising coding and non-coding regions that constitute the transcriptome serve as bridges between genes and proteins. Thus, a thorough transcriptome study is necessary to comprehend the genome's role, determine the cells' molecular makeup, and aid in comprehending the

genesis and progression of diseases (32). Although messenger RNAs (mRNAs) represent the majority of the coding transcriptome, they only make up a small fraction of the total transcriptome, accounting for about 2% of the total RNA in eukaryotic cells (32).

In the past few decades, researchers have developed several techniques for more precise analysis of RNAs and more accurate knowledge of gene expression. Low-throughput methodologies like quantitative polymerase chain reaction (qPCR) are powerful tools for this goal but were unsuitable for studying many different RNA molecules simultaneously (33). In addition, despite the introduction of hybridization-based microarrays, which offers a better method for analyzing gene expression, technical limitations such as cross-hybridization between similar sequences and a lack of precision in quantifying low and highly expressed genes prompted researchers to develop sequence-based techniques to minimize errors in the analysis of transcriptome technologies, using complementary DNA (cDNA) (34). With the introduction of high-throughput next-generation sequencing (NGS), researchers have increased their ability to analyze a wide range of biological data faster, cheaper, and more accurately (35).

This technology, known as RNA sequencing (RNA-Seq), has transformed our understanding of the complex and dynamic nature of the transcriptome by enabling a more accessible 'snapshot' of gene expression at a given time or disease condition in a more comprehensive, sensitive, and quantitative manner (36).

There are several platforms for RNA sequencing. One of the most popular high-throughput sequencing systems is Illumina, which allows the simultaneous determination of sequences of more than 10 million cDNA fragments, resulting in greater sequencing throughput than other sequencing technologies. The technology is based on the use of fluorescently labelled dNTPs for clonal DNA amplification (37). Fluorescence imaging identifies the new base after each cycle of base expansion.

Illumina is well-known for its quick turnaround times, high throughput, and low error rates (1% across its sequencing platforms) (37). Reverse transcription of RNA to cDNA and adaptor ligation of pre-specified Illumina adaptors for recognition by the sequencer are steps necessary for library preparation for Illumina sequencing (37). The RNA-Seq analysis workflow is reported in Figure 1.

Although RNA-Seq was established over a decade ago, it has profoundly transformed how we interpret gene expression, thanks mainly to differential gene expression (DGE) analysis (38). We are now able to identify variations in gene expression in relation to the different physiological or pathological conditions under investigation (39) by following a standardized pathway, which starts with raw data from sequencing and proceeds through in-depth computational analyses (39).

The following chapters and subsections illustrate the crucial steps leading to the identification of differentially expressed genes.

FIGURE 1. RNA Sequencing workflow.

1.3. Computational Analysis of RNA-Seq

The introduction of computational methods associated with sequencing has made it possible to decode the biological complexity of transcriptomes, from quality control needed to filter and assess the quality of raw reads to quantitative analysis of gene expression (40).

Here, we present an overview of the conventional computational pipeline for the differential expression analysis of RNA-Seq.

1.3.1. Quality control (QC)

The sequencing process can introduce errors in the reads, affecting the results of downstream analyses (41).

It is critical, therefore, to use QC methods and correct such inaccuracies before proceeding with subsequent analyses to ensure reliable results and accuracy of interpretations (42). QC includes several metrics to be evaluated, but the key ones are per-base sequence quality, adapter content, sequence duplication levels, and GC content. "Per Base Sequence Quality" refers to the accuracy of each individual base sequenced and is one of the most common metrics in QC of sequencing data (43). This metric is expressed through the Phred score, indicative of the likelihood that a specific base has been sequenced correctly. A higher Phred score indicates higher confidence that the base has been sequenced correctly, while a lower score indicates a higher probability of error at that particular position (43). "Adapter Content" is an essential metric to identify and possibly eliminate adapter sequences before proceeding with subsequent analysis. In fact, during preparation for sequencing, "adapters" are added at the ends of RNA sequences, but if these sequences are short or not the right length, the adapter itself can be incorrectly sequenced. If not removed, these sequences can cause errors in subsequent analyses, such as genome alignment (43). "Sequence Duplication Levels" represent the percentage of duplicated sequences in the sequencing dataset and are therefore present more than once. Understanding this metric and its implications can help ensure reliable and accurate results. In the case of RNA sequencing, expecting duplications is quite normal. Genes are not uniformly expressed within the cell, as, in

fact, some genes are very actively transcribed, producing large amounts of RNA, while others are less expressed (43). This variation in gene expression is reflected in the frequency of any given RNA sequence sequenced. In an RNA-Seq dataset, an RNA sequence derived from highly expressed genes will appear many times, giving the impression of duplications. Thus, in the context of RNA-Seq, these duplications directly reflect gene expression levels in the cell (43). However, it could also be due to certain methodologies, such as PCR, in which a DNA molecule may have been over-amplified, leading to an inflated representation of a particular sequence in the final data. Lastly, the "Sequence GC Content" module in FastQC compares the distribution of GC content in sequences against that expected from a random sample with the same number of sequences. This allows the identification of skewed GC content distributions. A uniform distribution of GC content is what would be expected from a sample (43). However, for RNA-Seq, deviations from the norm are expected. This is because transcripts (RNAs) are not evenly distributed in cells; some genes (with specific GC contents) can be overexpressed relative to others. Extreme deviations can indicate problems, such as contamination or amplification bias during sample preparation (43).

Several widely used QC tools, including FASTQC and MultiQC, summarize the results of a myriad of QC programs in a very practical way (42,44).

FastQC is an invaluable tool that provides an initial glance at the raw sequencing data. By analyzing the FASTQ files, FastQC offers a variety of quality metrics. These metrics include base quality score distributions, sequence content across all bases, duplication levels, and overrepresented sequences. The visual outputs, such as graphs and charts provided by FastQC, give researchers an immediate understanding of whether within a sequence, there are regions of concern that might need further attention or correction (41,42).

While FastQC is very good for analyzing individual samples, it does not perform well with multiple samples, as in fact is not easy to assess each FastQC report separately. In this case, MultiQC is the best option. It aggregates data from multiple FastQC reports into a single comprehensive document, allowing for easier comparison between samples. This comprehensive view enables researchers to identify batch effects, anomalies, or systemic issues that might affect the entire sequencing run (44).

Many steps of the analytical process should include quality control. QC should be proactive and thorough so that you can conduct downstream analysis with suitable assumptions and parameters. Equally essential are the mapping and counting steps outlined in the next section (42–44).

1.3.2. Mapping and counting

Identifying the genetic origin of the sequenced cDNA fragments is necessary to identify the transcripts present in a particular sample. One of the essential steps in all high-throughput sequencing investigations is mapping, which involves assigning sequenced reads to the locus of origin that is most likely their source (45). One approach to solving the string-matching challenge is to find the position along a long sequence known as the transcriptome, where a short series of nucleotides provides the best match. These methods are referred to as read alignment because they need individual letters of two provided strings to line up. Alignment is performed until the greatest number of bases in the read can be aligned, using a portion of the read as the "seed" to guide the tool's search for the ideal match in a reference genome index (46).

Each matching 'seed' is extended on both sides within specific parameters (such as the maximum number of permissible mismatches). At this point, the actual alignment phase is looking for the best possible local alignment of a particular read at a location in the

reference genome anchored around the match of the seed. The Smith-Waterman method (47), one of the pillars of computational biology, is typically used to carry out the local alignment phase (46,48).

Unlike transcriptome alignment programs, which are limited to known transcriptomes and heavily dependent on annotation, the most well-known RNA-Seq alignment tools (e.g., STAR) use the complete genome as a reference and gene annotation as a map (46).

Sometimes, trimming techniques are performed before mapping to remove adaptor sequences and low-quality sequencing bases (assessed by prior quality control). However, these procedures can strongly influence the estimation of gene expression levels, thereby altering downstream results (49). On the other hand, it is known that modern readout aligners can perform "soft-clipping" (50). This way, soft-clipped bases are still included in the read mapping results but are marked as 'soft-clipped'. Soft-clipping is a pre-processing technique for RNA-Seq data used to improve the accuracy of alignment results without losing useful gene expression information needed for subsequent steps in the bioinformatic analysis of RNA-Seq data (50).

After mapping, the aligned reads are used to calculate the counts and estimate expression levels. Tools like HT-Seq-count can be used to estimate the expression levels of transcripts by counting the amount of reads that match to certain reference transcripts (51).

The htseq-count script is a tool for analyzing RNA-Seq data. It counts for each gene how many aligned reads overlap with its exons, given a SAM/BAM file and a GTF or GFF file with gene models (52).

Since the script is designed specifically for differential expression analysis, only reads that unambiguously map to a single gene are counted, while reads aligned to multiple

positions or overlapping with more than one gene are discarded for not compromising the accuracy of the analysis (52).

In fact, it is problematic when a reading is counted for both genes. Additional reads from the actual differentially expressed gene could contribute to misclassifying of the other gene as differentially expressed as well (53).

Furthermore, HT-Seq-count counts fragments instead of individual reads in the case of bidirectional (forward and reverse) data to provide more consistent and meaningful results in differential expression analysis (51). This approach avoids overestimation in counts, increases consistency between samples, and simplifies the interpretation of results, thereby making the analysis more biologically meaningful and reliable (52).

The accuracy of RNA-Seq quantification certainly affects the performance of subsequent analysis and is often required by many statistical methods designed to discover genes with significant expression changes under various physiological or pathological conditions, such as differential gene expression analysis (39).

1.4. Differential gene expression analysis

Differential gene expression analysis (DGE) is a statistical method used to detect variations in gene expression levels between different experimental conditions or sample groups (54). Consequently, since it allows to study of how variation in the genome translates into phenotypic changes, it is possible to obtain information on genes involved in a particular biological process, disease, or response to treatment (55). Therefore, this analysis is often used to study diseases and identify genes involved in pathological processes such as tumor development and can help in identifying biomarkers for diagnosis and/or prognosis (56) (57).

The initial phase of this analysis is the normalization and preliminary processing of the data. Given the wide range of instruments and methods used in collecting biological data, it is crucial to maintain consistency between different data sets. 'Noise' in the analysis is minimized by reducing distortions due to instruments or other interferences, thereby ensuring that conclusions are primarily based on biological processes (58).

After optimizing and standardizing the data, the next step is choosing the most appropriate model for them. This model is useful for identifying relevant trends from the data, and its it may depend on the specificity and complexity of the dataset. It is vital therefore to opt for the right model, as it determines the accuracy of subsequent analyses (59).

In the third step, the processed data are examined in depth to recognize genes with high variations in expression levels. With different expressions, these genes can reveal essential information on disease origin and therapeutic targets, or even serve as diagnostic indicators (60).

The last step focuses on the visual representation of the results. Once genes with variable expressions have been identified and their implications understood, it is essential to illustrate these data in a clear and intuitive graphic format (61) (62).

The complete workflow of the differential gene expression analysis is shown in Figure 2.

FIGURE 2. Differential gene expression analysis workflow. It presents the main steps described in the previous chapters (RNA sequencing, Quality Control and Mapping & Counting) and those that we will describe in the following chapters (Normalisation, Differential gene expression and Pathway enrichment).

1.4.1. Normalization

In biomedical research, the accuracy of the result is closely linked to the quality of the data examined particularly when examining gene expression, as in fact inconsistencies can arise being biological systems complex and with numerous techniques for collecting data. This highlights the crucial role of normalization (63). Normalization is an essential step in preliminary processing of data making

measurements under different conditions consistent and able to be compared directly. Converting the original data minimizes technical fluctuations caused by different sets of experiments, variations in sample quality, discrepancies in equipment, differences in the amount of RNA obtained, and differences in library creation and sequencing efficiency (58).

Normalization by geometric averaging is widely recognized among the myriad of normalization methods available (63).

In the geometric averaging-based normalization method, the geometric mean of its expression levels across all samples is determined for each gene, providing a measure that reflects the variation between samples. After calculating this average, the expression level of a gene in a particular sample is divided by the geometric mean. This procedure adjusts and aligns the expression levels of each gene around a reference point, usually 1, assuming that there is no relevant biological variation. After normalization, a gene with uniform expression in all samples will have a value approaching 1. This facilitates the comparison of data between different samples (63).

1.4.2. DGE model selection

In recent times, numerous methods have been designed to analyze differential gene expression from microarray and RNA-Seq data. Techniques such as DESeq rely on parametric models, assuming that the expression of each gene follows a specific distribution. These methods calculate the parameters of this distribution to compare gene expression between different conditions. Statistical distributions like Poisson or negative binomial are frequently adopted for RNA-Seq data, where gene expression is quantified by counting reads (59). The main difference between

various parametric approaches based on the same distribution is the algorithm used to calculate these parameters. In contrast, methods such as NOIseq (59) or SAMseq (60) are based on non-parametric models to represent gene expression distribution, suggesting that certain distributional assumptions or inaccurate parameter estimates can generate unreliable results. Although such methods provide a more detailed picture of gene expression distribution, they often require large sample sizes to provide meaningful information (59).

Although this field of study is rapidly developing, there is no agreement on the best methodology or the most reliable approach. Having said that, DESeq2 is one of the most widely used and well-validated methods for analyzing sequence count data and one of the most popular, for both reliability of the results produced and for easy use and reproducibility (63).

1.4.3. Deseq2

Deseq2 (Differential Expression analysis for Sequence Count data) is available in R and relies on the negative binomial distribution to model count data in RNA sequencing experiments (38). The negative binomial distribution is solely determined by the parameters μ and φ. The approach assumes that the number of times a particular gene is read in an RNA sequencing experiment follows a certain pattern, known as the negative binomial distribution. This pattern is defined by two factors: the average number of reads for a given gene (μij) and a factor called the dispersion parameter (φi), which accounts for variability in the data. Both are specific for each gene (i) and sample (j) studied. For the purpose of identifying DEG isoforms, the accurate measurement of the dispersion parameter φ i for each isoform is essential (64)(63). It describes the probability of obtaining a certain number of counts under a given

experimental condition, given the mean of counts and a dispersion parameter considering the variance overlap. The key feature different from other tools is the use of data-based contraction estimators for dispersion and log2FC. In particular, it uses a specific approach to estimate the dispersion parameter based on a model assuming a similar level of variability for genes with similar average expression levels. This approach is most appropriate when samples have relatively low biological variability (65). Furthermore, in DESeq2, the isoform-specific dispersion is restricted to an adjusted smooth curve using an empirical Bayesian technique, assuming that isoforms with comparable average expression levels have similar dispersion (63). When the expression level is low, DESeq2 reduces log2FC estimates towards zero, overcoming the challenge of estimating log2FC for weakly expressed isoforms (66).

Furthermore, DESeq2 performs an internal normalization in which the geometric mean of each gene across samples is determined (63). Then, this average is divided by the gene count in each sample. The sample size factor is the median of these ratios (63). This process corrects for biases related to library size and RNA composition, which can occur when only a small subset of genes are substantially expressed in one experimental context but not in another (64). Due to internal normalization, DESeq2 analysis is more flexible than other differential expression analysis tools (58,63).

1.4.4. Identification of Differential Expressed Genes

Specific measures of statistical significance are used in differential gene expression (DGE) analysis using DESeq2. These measures include p-values (67), p-values (68) adjusted by different techniques (padj) such as the false discovery rate (FDR) (69) and the magnitude of the difference, typically represented by the log2 fold change

(log2FC) (60). It should be emphasized that in order to avoid a high number of false positives in differential expression analysis, it is necessary to correct for multiple tests and select different expressed genes by setting a cutoff on the adjusted p-value. The log2FC can therefore be used to classify genes mostly representative of the difference between two experimental conditions.

Correct cut-off thresholds are crucial for identifying relevant genes in DGE analysis. Inadequate selection of these thresholds could lead to incorrect or insignificant conclusions (70).

1.4.5. Graphical representation

After conducting a differential expression analysis and identifying genes based on an adjusted p-value below a certain threshold, it is crucial to graphically visualize the results and correctly interpret the data (54). These representations help highlight statistical significance (67), mean expression levels (71), and the magnitude of observed differences (72). Common techniques include MA diagrams, which compare fold-change with normalized mean counts (61), Volcano diagrams, which relate p-value and fold-change (62), and PCA (Principal Component Analysis) useful for analysing and visualizing variations in large data sets (73).

MA diagrams are often used to visualize the differential expression between two conditions, relating log2FC to the mean expression (61). The y-axis represents log2FC in the diagram, while the x-axis shows normalized mean expression. Genes with significant expression differences are visible as points away from the horizontal axis. Interestingly, genes with low average expression values tend to have greater variability in log2FC than those with high values, causing a fan effect in the graph. Dashed lines usually indicate the limits for log2FC in the MA diagram. However, a

critical point of these graphs is that they do not provide information on statistical significance, as the p-value is not considered (74,75). MA graphs are valuable because they combine details on the magnitude of change (y-axis) with the average intensity of gene expression (x-axis). Thus, it is possible to observe both genes with considerable variation in expression and those with consistently high or low expression levels in both contexts.

The analysis of alterations in gene expression between two conditions is often visualized by means of a 'volcano diagram', which compares the statistical significance and magnitude of change (62). In this graph, the y-axis shows the negative logarithm to base 10 of the p-values, representing significance, while the x-axis highlights log2FC, indicating how marked the difference in expression is for each gene. Thus, points high on the y-axis identify genes with very significant differences. In parallel, extreme positions on the x-axis represent significant changes in expression (62). Genes very far from the origin in either direction are those with both statistically and biologically significant changes.

Principal Components Analysis (PCA) is a technique in multivariate statistics employed for the reduction of dimensionality of a dataset while retaining most of the original variance (73). It operates through a linear transformation of the initial variables, generating a new set of variables called 'principal components' (PCs). These components are orthogonal to each other and are ordered so that the first principal component (PC1) represents the direction along which there is the greatest variability within the dataset, the second principal component (PC2) represents the next direction of maximum variation, and so on (73). This reduces the data's complexity by focusing on the most important patterns of variation for analysis. PCA relies on the singular value decomposition (SVD) (76)of the data matrix, which

divides the original matrix into three distinct matrices, providing a set of vectors representing the principal components (73).

A PCA plot, commonly referred to as a scatter plot or biplot, represents each sample as a point in a two-dimensional (or sometimes three-dimensional) space defined by the first two (or three) principal components. The distance between points in this space reflects differences in gene expression between samples. Samples that have comparable patterns of gene expression will be close together, while those with different expression profiles will be far apart. A clear separation between groups of samples in a PCA graph may suggest substantial differences in gene expression between these groups (73).

1.5. Pathway enrichment

Pathway enrichment is a computational analysis that identifies biological pathways significantly enriched in differentially expressed genes, proteins or associated with a particular condition/disease (77). The construction of pathway enrichment networks is critical in understanding the biology of differentially expressed genes between two different conditions. It allows the identification of biological pathways involved in specific biological functions, such as the regulation of apoptosis or cell signaling. In addition, it can help in understanding how genes or proteins interact within these pathways and how they might be involved in pathology (78). This type of analysis offers a more comprehensive and integrated view of biological functions and pathological mechanisms, providing insights for developing new therapies or diagnostic strategies. These techniques, which facilitate the concentration of a large amount of high-dimensional information into a few biological processes underlying

specific phenotypes, have emerged as one of the most critical techniques for interpreting biological data (79).

1.5.1. Pathway Enrichment Analysis Tools and Databases

Various sophisticated tools are available for researchers to decipher the complexity of genetic and proteomic data. ClusterProfiler is rapidly emerging as a sophisticated tool for pathway enrichment analysis. Created as part of the Bioconductor project, this tool not only provides methods for comparative studies and visualizations of biological profiles, but being developed for R, but can also harmonize and integrate with many other genetic analysis platforms (80,81).

In parallel, the well-known DAVID, an acronym for Database for Annotation, Visualization, and Integrated Discovery, is established as a milestone in the field. With its ability to synthesize and integrate multiple sources of gene annotation, DAVID has made identifying biological pathways a more intuitive and informative practice, further enriching the analysis by introducing clustering based on expression data (82).

Finally, while DAVID and ClusterProfiler focus more on genomic aspects and pathways, STRING comes into play when understanding interactions at the molecular level. This tool, which stands for Search Tool for the Retrieval of Interacting Genes/Proteins, sheds light on protein-protein interaction networks, providing valuable details from direct experiments and predictions and pathways (81).

In summary, while each tool has its peculiarities and strong points, combining several can provide a more complete and detailed view, helping researchers navigate the complexity of bioinformatics research (83).

2. Aim

In order to identify possible new specific diagnostic and therapeutic biomarkers we obtained the transcriptome of paired MPM and hyperplastic tissue to identify variations of gene expression level associated with disease progression.

3. Methods

This chapter outlines the procedures and techniques utilized for the transcriptome analysis and identifies genes that show differential expression between mesothelial hyperplasia and malignant pleural mesothelioma samples. The pipeline includes RNA extraction, library preparation, sequencing, quality control, alignment of reads, quantification of reads, normalization, and statistical analysis using Deseq2. Furthermore, the two-sample Welch's t-test, enrichment analysis, and comparison of our results with those obtained from the GEO database are described. Finally, the graphical representations performed in this study are described for a better understanding of the transcriptome analysis.

3.1. Study design and patient samples

Five patients, four males and one female aged between 65 and 78, with epithelioid malignant pleural mesothelioma diagnosed between 2018 and 2021 and paired mesothelial hyperplasia identified by histopathological analysis, were included in this study. All samples used in this study were collected at the Azienda Ospedaliero-Universitaria Senese in Siena, Italy.

One patient and his cancer and hyperplasia samples were excluded from our analysis due to limitations noted during the library preparation process. These samples, collected back in 2018 at the Azienda Ospedaliero-Universitaria Senese in Siena,

presented potential problems with RNA stability and integrity due to the time elapsed. Furthermore, the small sample size could have compromised the quality of the data obtained during sequencing. In this regard, a standard RNA storage protocol was used, but older samples naturally have a higher probability of degradation.

Consequently, for the analysis, we considered the four remaining patients with hyperplasia and cancer between 2020 and 2021. In this study, it is, therefore, necessary to recognize the limitation of the size of the sample set examined. Selecting paired samples allowed us to focus on a more homogeneous group regarding diagnosis and timing, minimizing the potential impact of variations. This gave us an opportunity to identify and characterize changes in gene expression specifically associated with hyperplasia-to-tumor progression.

3.2. RNA extraction

Total RNA from Formalin-Fixed, Paraffin-Embedded (FFPE) biopsy specimens at diagnosis and from surgical specimens was extracted using an automated extractor and the FFPET RNA Isolation Kit (Roche diagnostic GmbH, Germany) following the manufacturer's instructions. This automated methodology ensures high reproducibility and efficiency in extracting RNA from FFPE samples, which are often complex to process due to the chemical changes introduced by formalin.

After extraction, the RNA's concentration and purity was performed using Nanodrop. For more accurate RNA quantification, we additionally used the Qubit RNA HS Assay kit (Thermo Fisher Scientific Cat. No. Q32852) on Qubit equipment. This specific system provides a sensitive measurement of RNA. This step is crucial to ensure that the amount of RNA used in subsequent steps is optimal and that results are as accurate as possible.

3.3. Library preparation and RNA sequencing

Eight tissue samples, derived from primary tumors and mesothelial hyperplasia samples, were subjected to sequencing to acquire a detailed profile of gene expression in both samples.

100 ng of mRNA reverse-transcribed into cDNA. Subsequently, DNA libraries were prepared using the Illumina Stranded Total RNA Prep Ligation kit with Ribo-Zero Plus (Illumina, San Diego CA). This methodology enriches RNA fractions extracted from paraffin-embedded samples and removes ribosomal RNA, which could interfere with transcriptome analysis.

After preparing the libraries, we checked the quality and quantity of the material produced using the Agilent™ High Sensitivity DNA kit (Agilent Technologies, Palo Alto, CA), run on an Agilent2100 Bioanalyzer (Agilent Technologies). This instrument provides a detailed analysis of libraries' size, quality, and concentration.

For a complementary quantitative evaluation, we used the Qubit™ dsDNA BR Assay Kits on a Qubit 2.0 fluorometer (Invitrogen, Carlsbad, CA, USA), following the instructions provided in the user guide. This step allowed an accurate estimate of the DNA concentration in our libraries, ensuring that the samples were suitable for the next sequencing steps.

We then proceeded to the sequencing phase. We chose the NextSeq 550 platform (Illumina, San Diego, CA) for sequencing. The NextSeq 550 has high throughput and the ability to produce large amounts of data in a single run while ensuring precise sequencing resolution, which is necessary for this analysis.

Sampling was performed using a paired-end read configuration, which allows both ends of the cDNA fragments to be sequenced. This mode greatly increases the accuracy of variant identification and provides a more uniform transcriptome coverage.

3.4. BCL to fastq

The sequencing data were obtained in "BCL" format. This format contains the raw data from the base call and sequencer quality score. Although the BCL format includes information, it is unsuitable for most downstream bioinformatic analysis.

We used the bcl2fastq2 version 2.20 conversion software to facilitate this conversion. This software, by Illumina, is designed to convert BCL files into the widely used FASTQ format. The FASTQ format effectively represents both nucleotide sequences and corresponding quality scores, making it particularly suitable for subsequent bioinformatic analysis.

In addition to the bcl2fastq2 software, we also used the Sample Sheet (v2). This sheet plays a crucial role in guiding the demultiplexing process. Demultiplexing is essential, especially when multiple samples are sequenced in a single run. The Sample Sheet contains metadata about the barcodes used for each sample, ensuring that the reads are segregated according to their origin during conversion.

Once the conversion process was completed, we were provided with the FASTQ files. These files contained the nucleotide sequences that directly matched the RNA fragments from our tissue samples. The transformation of raw sequencing data into interpretable nucleotide sequences was a significant step in our research process, and the foundation for subsequent data analysis.

3.5. Quality control and Read alignment

Once the sequencing was completed and the FASTQ files obtained, they were subjected to further quality control (QC) to ensure that the reads obtained met the expected quality standards. Indeed, sequencing can sometimes introduce inaccuracies due to instrument errors, sample preparation inconsistencies, or contaminants. We employed two widely used software tools to address these concerns: FastQC (version 0.12.1) and MultiQC (version 1.14).

Once we verified the quality of the sequenced samples, we aligned the reads. We used STAR software (2.7.10b) for sequence alignment on reference genomes. To perform the mapping, the files Homo_sapiens.GRCh38.dna_sm.primary_assembly.fa and Homo_sapiens.GRCh38.109.gtf were downloaded from Ensembl (ensembl.org), respectively, fasta files containing the nucleotide sequences of the reference genome and gtf files containing the annotations of the genomic regions. These files are needed to generate an optimized STAR index for aligning RNA-Seq reads to the reference genome.

Two types of alignment were performed on the same samples, with or without soft-clipping, to determine which sample group had better sequencing and mapping quality, verified with MultiQC (version 1.14). This step is crucial to make informed decisions regarding data interpretation and subsequent analysis.

3.6. Reads quantification

To quantify gene expression levels from alignment data in BAM format, we used the HTSeq tool (version 2.0.3). In this counting step, the sequences associated with each gene were counted according to their genomic position. The tool's output was a tabular file containing the gene name and the associated sequence count.

3.7. Normalization and Deseq2 analysis

The count data obtained were used for analysing variations in gene expression using the DESeq2 package (version 1.40.1).

To ensure that gene counts were comparable between samples, we used the normalization method proposed by DESeq2. This approach is based on average normalization factors, calculated for each sample to reduce systematic differences between samples due to variations in sequencing depth or other factors. We adopted the standard normalization method provided directly by the DESeq2 package without changing the default parameters.

In the statistical model adopted, we considered pleural mesothelioma as the reference condition and hyperplasia as the comparison condition. Consequently, the coefficients derived from the model reflect the differences in the logarithmic expression between the two conditions. A positive coefficient value (Log2FC > 0) suggests that a gene is more highly expressed in hyperplasia than in mesothelioma. In contrast, a negative coefficient (Log2FC < 0) indicates lower expression in hyperplasia than in pleural mesothelioma. Genes were considered statistically significant in their difference in expression if they had an adjusted p-value and p-value of less than 0.05.

3.8. Graphical representation

To assess the variation and distribution of the samples in our dataset based on their gene expression, we performed a PCA analysis. PCA was performed on the normalized counts derived from DESeq2. The outcomes of the PCA were depicted graphically through the ggplot2 package in R. The resulting graphs represent each sample as a point in the two-dimensional space determined by the first two principal components, which are the ones that explain the largest variance in our data. The coloring and labels of the points in the

PCA graph were used to distinguish between the different experimental conditions and allow easy interpretation of the main trend and grouping of samples.

As part of the data analysis, we assessed the correlation between samples to understand the similarity or dissimilarity in their gene expression. To do this, we calculated a correlation matrix based on the normalized gene counts of each sample. The correlation matrix was calculated using Pearson's method. Once the matrix was calculated, it was visualized as a heatmap through the pheatmap package in R. Each row and column in this chart represent a sample, and cells are colored according to the degree of correlation between samples. Warmer colors (e.g., red) represent high positive correlation, while cooler colors (e.g., blue) represent high negative correlation. Intermediate colors represent instead moderate correlation. This graphical representation allows rapid identification of samples with similar or dissimilar expression profiles. Rows and columns of the heatmap were ordered using a hierarchical clustering algorithm, which groups similar samples close to each other, facilitating correlations.

Volcano diagrams were represented with the R package EnhancedVolcano and MA diagrams were performed with the R package plotMA using the results of differential expression analysis as input data, containing information about the statistical significance and fold changes of the tested genes.

3.9. Welch's two-sample t-test

To assess differences between the means of two distinct groups, we turned to Welch's two-sample t-test. Tailored specifically for scenarios where the two groups may not have equal variances, this test offers a significant advantage over the traditional student t-test. The key distinction lies in the assumptions: the student t-test assumes equal variances between groups, while Welch's two-sample t-test does not (84). This lack of

assumption about equal variances makes Welch's approach especially valuable when the groups being compared might be unequal. Furthermore, when faced with differing variances across groups, Welch's two-sample t-test tends to generate p-value estimates that are more accurate than those derived from the student t-test (84). Within the context of our research, we employed Welch's two-sample t-test to dissect the differences in the means from data produced by alignments with or without soft clipping. Furthermore, it assessed the difference between the averages of sequenced tumor and hyperplastic samples from patients. Normalized and logarithmically transformed expression data were used as input.

3.10. Pathway enrichment

Genes identified as differentially expressed in our analysis were further investigated to understand the potential biological pathways involved and their interrelationships. For this purpose, we used pathway enrichment tools and databases. Using ClusterProfiler, we performed enrichment analysis to identify the main GO (Gene Ontology) categories and KEGG (Kyoto Encyclopedia of Genes and Genomes) pathways associated with our genes of interest. Furthermore, we used DAVID to explore the functional associations of genes and enhance the understanding of the biological interactions involved. Finally, using STRING, we constructed interaction networks between differentially expressed genes to identify potential key hubs and interaction modules within our data.

3.11. GEO database data detection

Given our limited number of samples, we supplemented and corroborated our results by interrogating publicly available gene expression data. For this purpose, we turned to GEO (Gene Expression Omnibus) database, a public resource that has various gene expression datasets.

After a deep dive into this database, we identified the dataset with the identifier GSE51024, which contains 41 paired samples of Malignant Pleural Mesothelioma (MPM) and related normal tissues. This selection provided an opportunity to further strengthen and validate our initial results through comparison with a larger number of samples.

Sample selection and data download were performed following the guidelines and tools of the GEO website. As the data were already normalized and analyzed with the Affymetrix platform, we continued with a differential expression analysis through the limma package in R (85).

3.12. Differentially expressed genes comparison

Although the analysis of the GSE51024 dataset provided inherently different evaluation parameters from our analysis with Deseq2 due to the different platforms and normalization methodologies used, our main focus remained the comparison of differential gene expression profiles between our initial and publicly available sets. In particular, the focus was on differentially expressed genes, rather than on statistical details or set-specific parameters, to identify shared coherences and trends between the two data collections and ensure greater robustness and reliability of the overall results.

3.12.1. Graphical representation for genes comparison

We made Venn diagrams for a graphical representation highlighting the overlaps and intersections between the differentially expressed genes in the experimental conditions of our initial data set and the external set. These diagrams allowed us to visualize the number of unique and shared genes between the various conditions and to provide an overall view of the interrelationships between the different sets of genes. The creation and visualization of the Venn diagrams were carried out using the matplotlib-venn library in Python. The diagrams obtained provided valuable information on the overlaps

between the analyses of the two datasets and contributed to a better interpretation of our findings.

To further investigate the common genes identified in the Venn diagrams between our initial dataset and the external set, we decided to visualize their expression through heatmaps. These heatmaps provide a graphical representation of gene expression values, allowing us to identify shared or distinct expression patterns between samples. The heatmap was performed using R's heatmap function.

Furthermore, we also crafted pie charts to represent the expression of common genes in the two distinct experiments. These charts provided an intuitive and visual breakdown of the gene expression proportions, allowing for a direct comparison of their presence and prominence in each experiment. The pie charts facilitated a clearer understanding of how these shared genes behaved and were expressed in the context of both datasets.

4. Results

4.1. Quality control results

Several key parameters related to sequencing quality and alignment efficiency were observed during our transcriptome analysis. First, using the STAR alignment tool, we observed that between 80% and 90% of reads were mapped to the reference genome. This is an effective mapping rate indicating the quality of the reads and the relevance of the reference genome used.

Next, we examined the percentage of duplicated reads generated during the library amplification step. Such duplicates can affect the amount of useful information obtained. Fortunately, the percentage of these duplicated reads was between 20% and 30%, a

range considered typical for an RNA sequencing experiment. This ensures that the data are representative and the library amplification was performed correctly.

Regarding base composition, the average GC content in the sequencing dataset was around 50%. This percentage is in line with what is expected from a human RNA sequencing dataset, further confirming the quality and representativeness of our sample.

We also explored the effectiveness of soft clipping alignment on a group of 4 paired samples. Following this alignment with STAR, we noted an improvement in data quality of 0.1% without losing any useful information for differential expression analysis. Welch's two-sample t-test further confirmed the validity of this method. The results showed a higher significance of the alignment with soft clipping (p.value < 2.2e-16) than the alignment without soft clipping (p.value = 2.098e-05).

4.2. Differential expression analysis

During our research, we adopted several analytical techniques to better understand the differences between hyperplastic and tumor samples. Our main methodology was principal component analysis (PCA), illustrated in the graph in Figure 3. This technique allowed us to reduce the dimensionality of our dataset by highlighting the variance between observations. We noticed that the first two principal components, PC1 and PC2 captured 52% of total variance in the dataset. However, it was interesting that not all samples classified “a priori” as hyperplastic were located far from the tumor samples. In particular, only samples N15151 and N729 diverged from clustering with the tumor samples, whereas samples N26191 and N16949 tended to cluster with the tumor samples.

FIGURE 3. PCA graph: represents the structure of our multivariate data set in a two-dimensional space. Each point on the graph represents an analyzed sample. The coordinates are determined by the first two principal components, which represent the linear combinations of the original variables with the highest variance. Neighbouring points are similar in their properties, while distant points are different. This visualization allows observation patterns to be identified and outliers to be detected.

To consolidate these results, we further validated the analysis with a two-sample Welch's t-test and a correlation matrix (Figure 4).

Two-sample Welch's t-test allowed us to compare the averages of the two paired samples, assessing the magnitude of their mutual deviation in terms of standard deviations and statistical significance of this comparison. We observed that the hyperplastic samples N15151 and N729, compared with their corresponding tumor samples, showed significant t-values with p-values of 0.001528 and 0.01151, respectively. On the other hand, samples N16949 and N26191, which were “a priori” identified as hyperplastic, did not show statistical significance when compared to the

paired tumor samples. Furthermore, the correlation matrix further revealed that samples N15151 and N729 showed a strong negative correlation with most of the other samples.

.

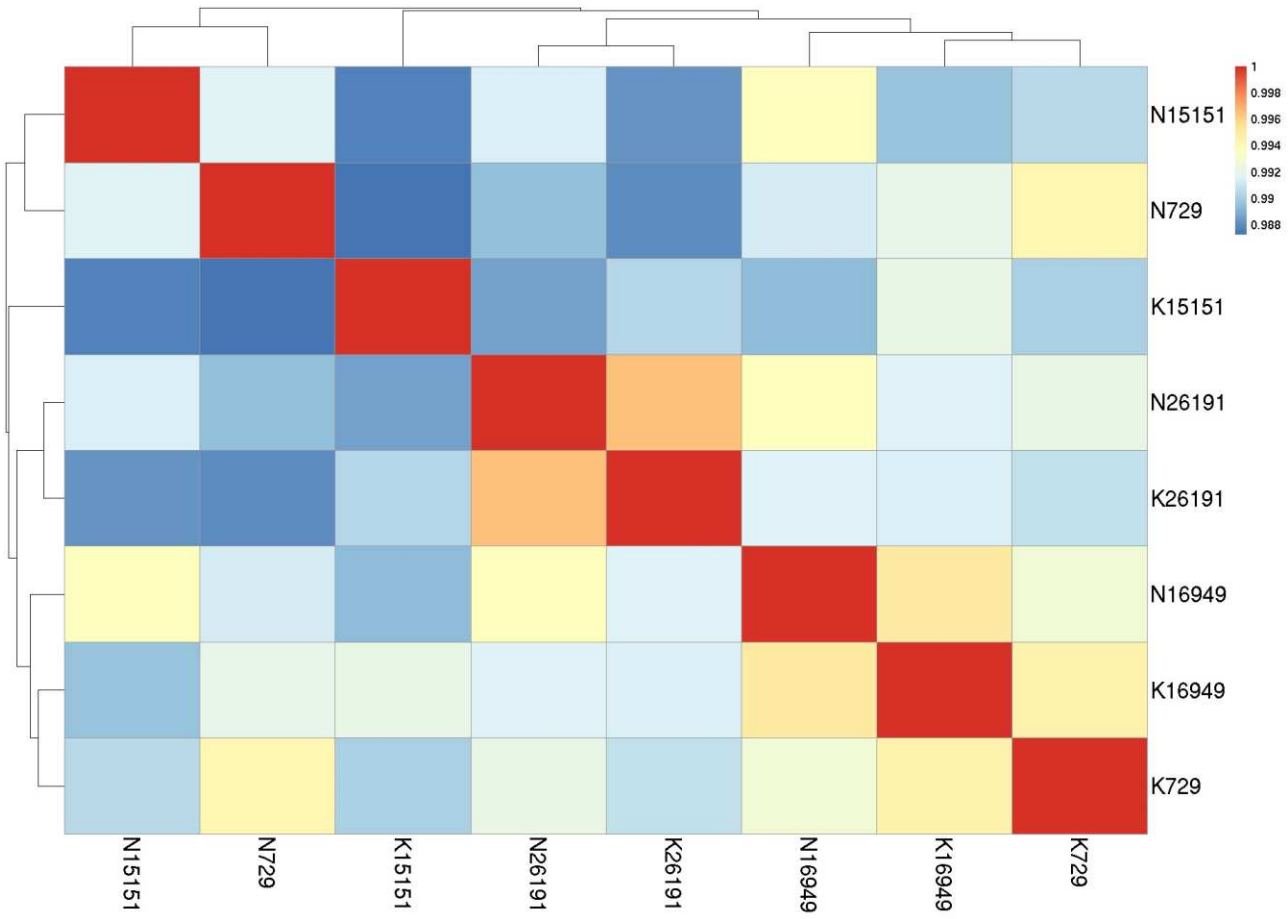


FIGURE 4. Correlation matrix: In this heatmap, the intense colors represent the correlations between the samples. Red indicates high positive correlations, and blue indicates high negative correlations. The color scale makes it easy to identify correlation patterns within the data.

After confirming a statistically significant difference in gene expression only between samples N15151 and N729 and the respective tumor samples, we proceeded with the differential expression analysis between these two samples and the other six samples in the cluster, all of which were identified as tumor samples. We started by examining the MA plot (Figure 5) and the volcano plot (Figure 6). In the MA plot, the x-axis represents the average of gene expression levels between the two conditions or groups of samples,

while the y-axis depicts the logarithm of the ratio of gene expression levels. The uniform distribution centred on the y-axis indicates a lack of systematic differences in gene expression between conditions. However, our analysis showed a skewed distribution of points, indicating a variation in gene expression between conditions. This further confirms the results of previous analyses.

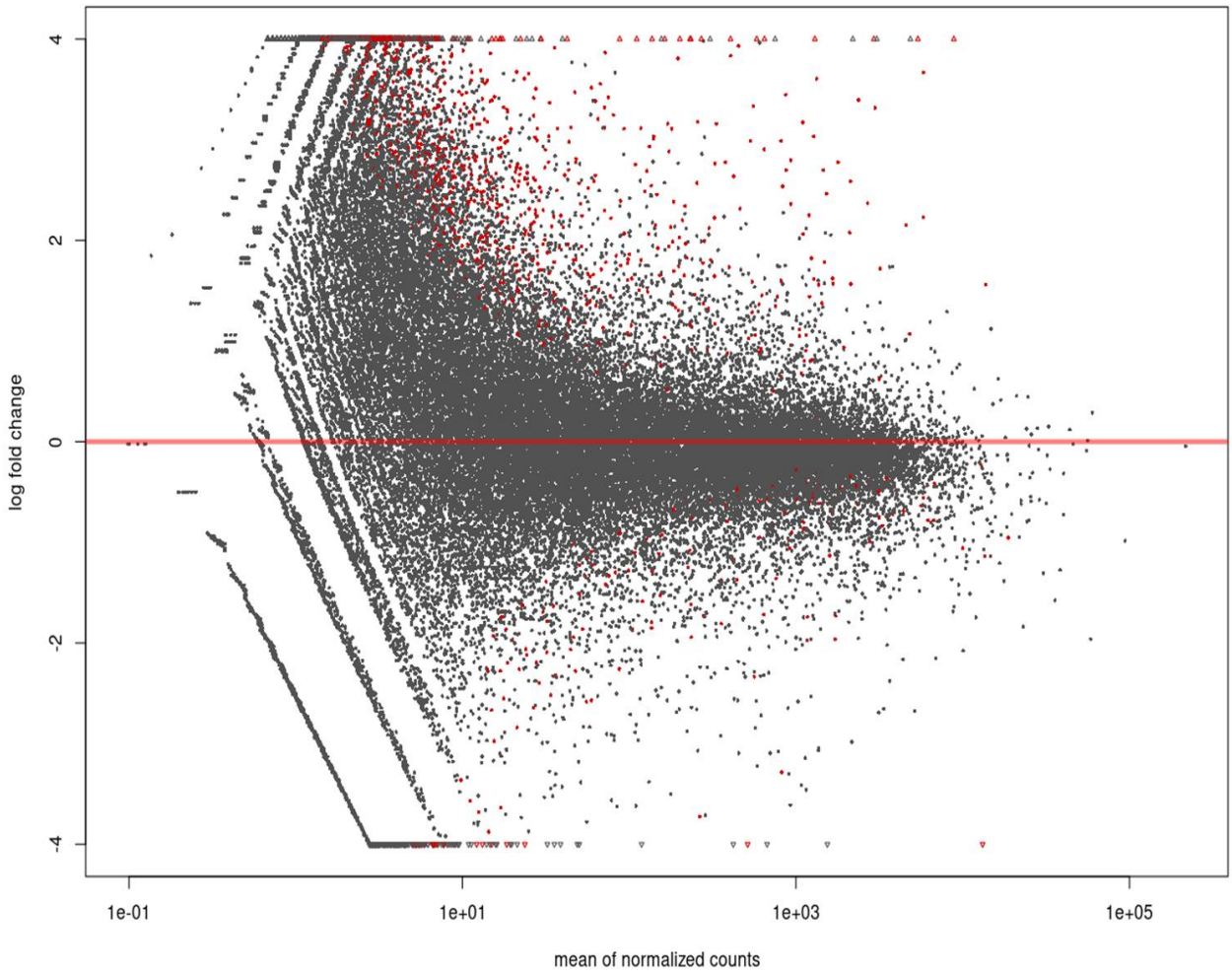


FIGURE 5. MA Plot: The M-axis represents the logarithm of the expression ratio between the two conditions, while the A-axis represents the mean of the logarithm of the expression. Points above zero indicate overexpression in the second condition relative to the first, while points below zero indicate downregulation. The MA plot makes it possible to identify differentially expressed genes and understand gene expression variations between the conditions analyzed.

The volcano plot showed that out of a total of 62,710 variables, 681 genes were overexpressed in hyperplasia and downregulated in tumor, while 98 were overexpressed in tumor and downregulated in hyperplasia.

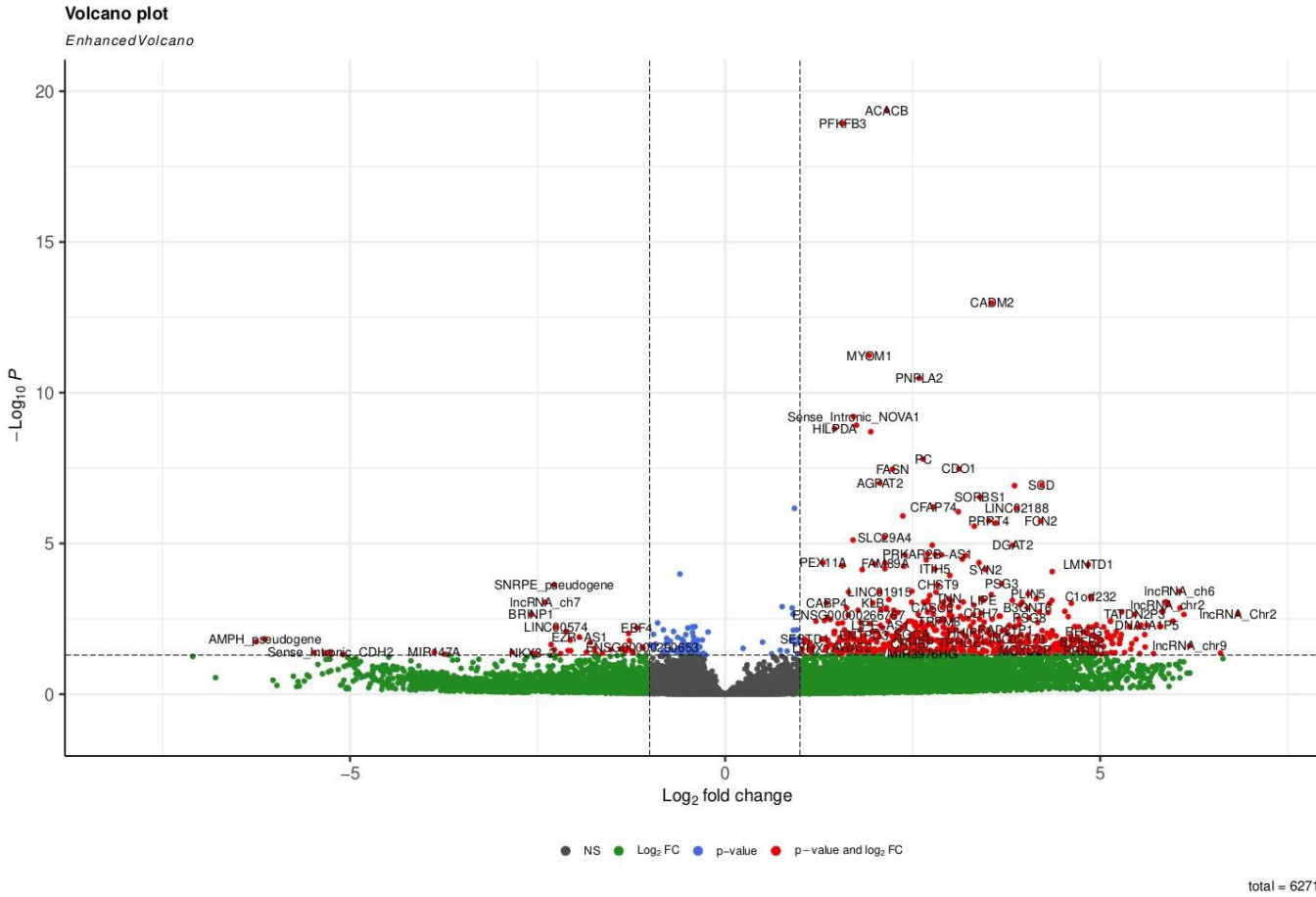


FIGURE 6. Volcano plot: shows the analysis of statistical significance and differential expression of a set of genes. The axes represent the significance of the differential expression (log2 fold change) and the test's statistical significance (-log10 p-value). Dots in red indicate genes that are significantly different between the two groups. The vertical lines indicate the cutoff for fold change, while the horizontal line indicates the cutoff for p-value. This graph makes it possible to identify significantly and differentially expressed genes.

For the next step of pathway enrichment, we included only coding genes, focusing on regions of the genome directly involved in gene expression. We, therefore, chose 469 coding genes from those significantly downregulated in mesothelioma and upregulated in hyperplasia and 80 from those significantly overexpressed in mesothelioma and downregulated in hyperplasia. Both sets of genes were subjected to pathway enrichment analysis.

4.3. Pathway enrichment analysis

We used two-stage enrichment pathway analysis for a comprehensive understanding of the functional relevance of differentially expressed genes in hyperplasia and tumors. Our primary objective was to ascertain the underlying biological processes and signaling pathways potentially impacted by these gene expressions.

Initially, our analysis utilized the ClusterProfiler package in R to enrich the 469 up-regulated genes in hyperplasia e down-regulated in MPM (Figure 7 and 8). Through the Gene Ontology database, this analysis showed a strong connection with fatty acid metabolism pathways, which is an important regulatory signaling in cells. Alterations in lipid metabolic processes can result in changes to the membrane's composition and permeability, two factors important for cell's integrity and function (86). Furthermore, lipid metabolism produces an array of biological intermediates. Many of these intermediates function as signaling molecules pivotal in orchestrating pathways modulating cell growth, proliferation, differentiation, apoptosis, motility, inflammation, survival, and membrane homeostasis (87).

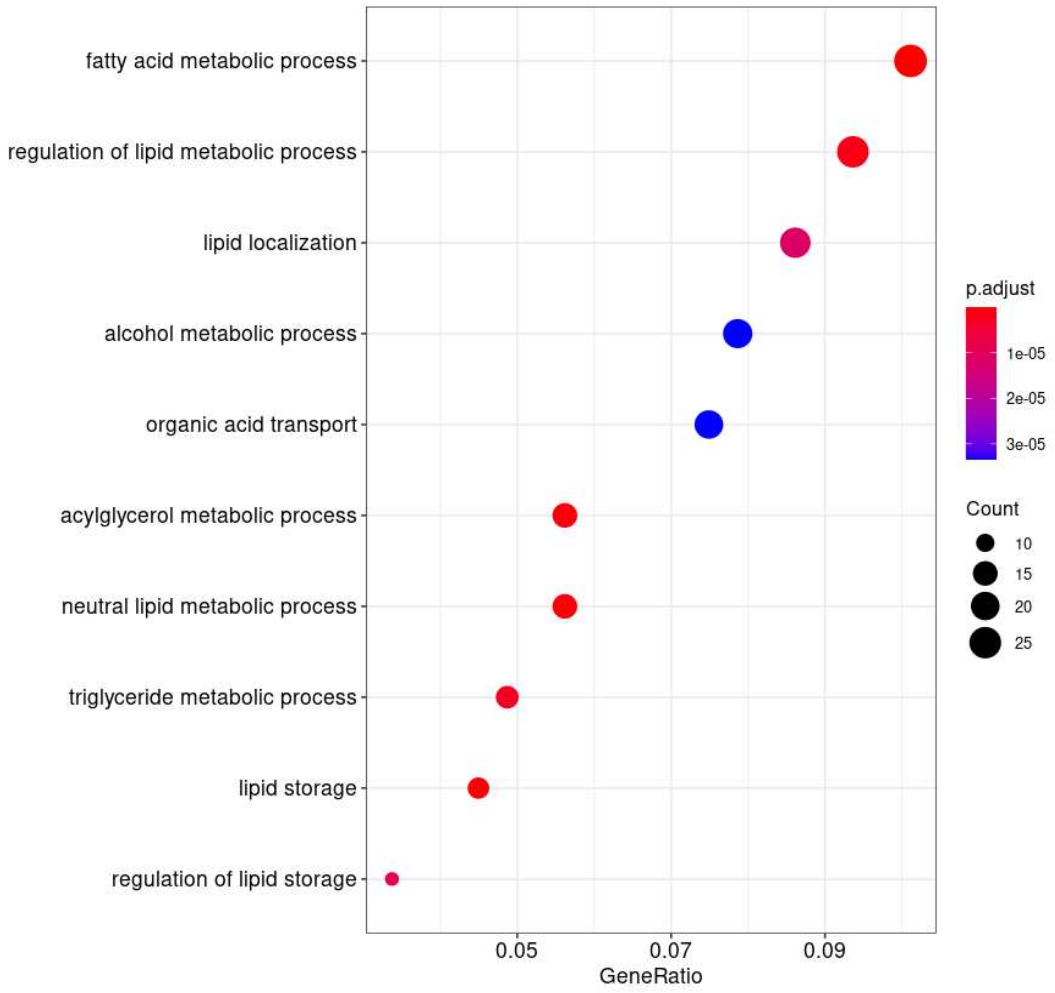


FIGURE 7. Dotplot up regulated gene in Hyperplasia. This dotplot visualizes the genes identified as over-regulated during the hyperplasia study. The dots represent the set of genes expressed for each specific biological process, and the diameter of the dots reflects the quantity of genes associated with that particular process. On the x-axis is the GeneRatio, which represents the fraction of genes associated with a specific process compared to the total number of genes expressed or analyzed in the study. This metric is useful for evaluating the relative importance of a biological process compared to others regarding gene over-regulation.

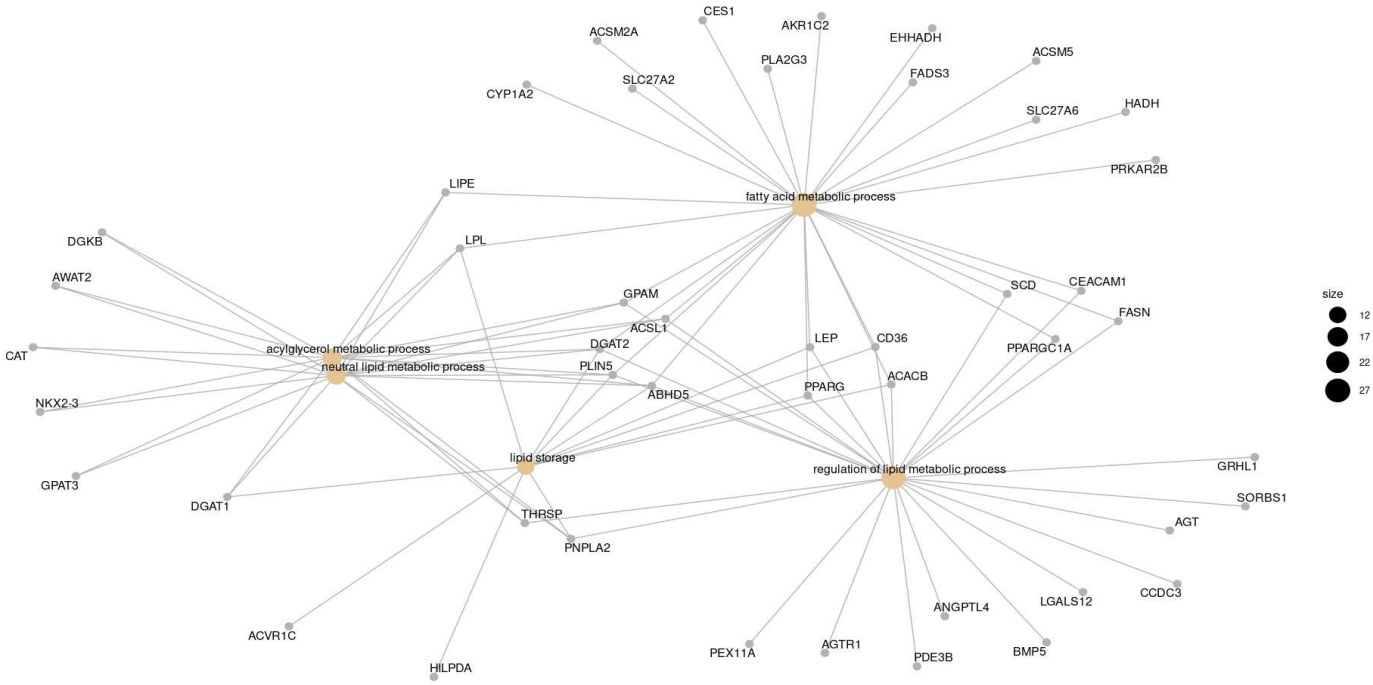


FIGURE 8. Cneplot up regulated genes in Hyperplasia. In this figure, a Cneplot is used to visualise the genes that were identified as over-regulated during the hyperplasia study. The Cneplot provides a detailed analysis of the interactions between these genes and can help identify clusters or groups of genes with coordinated expression. The connections between genes are represented by lines. This visual representation provides insight into the network of genes involved in the hyperplasia response and is valuable for understanding the molecular pathways involved in this condition.

Building on the information gained from ClusterProfiler, our subsequent enrichment analysis utilized the DAVID database. This subsequent analysis showed the importance of the PPAR signaling pathway among the set of overexpressed genes in hyperplasia and dowregulated in mesothelioma. This pathway encompasses transcription factors that bind specific response elements within the DNA, consequently modulating gene expression pivotal to lipid metabolism, inflammation, and cell proliferation (88). The anti-inflammatory attributes of PPARs are particularly noteworthy, as they diminish the synthesis of inflammatory cytokines and invigorate immune cells (88). The PPAR signaling pathway also influences cell proliferation and differentiation across many tissue types, suggesting its elevated expression within hyperplastic cells (89).

Shifting our focus to the tumor environment, enrichment pathway analysis of the 80 up-regulated genes in mesothelioma and downregulated genes in hyperplasia, conducted via the ClusterProfiler package, highlighted different biological processes. While these processes were of lower statistical significancy, they were mostly associated with regulating protein stability and DNA biosynthetic processes (Figure 9 and 10).

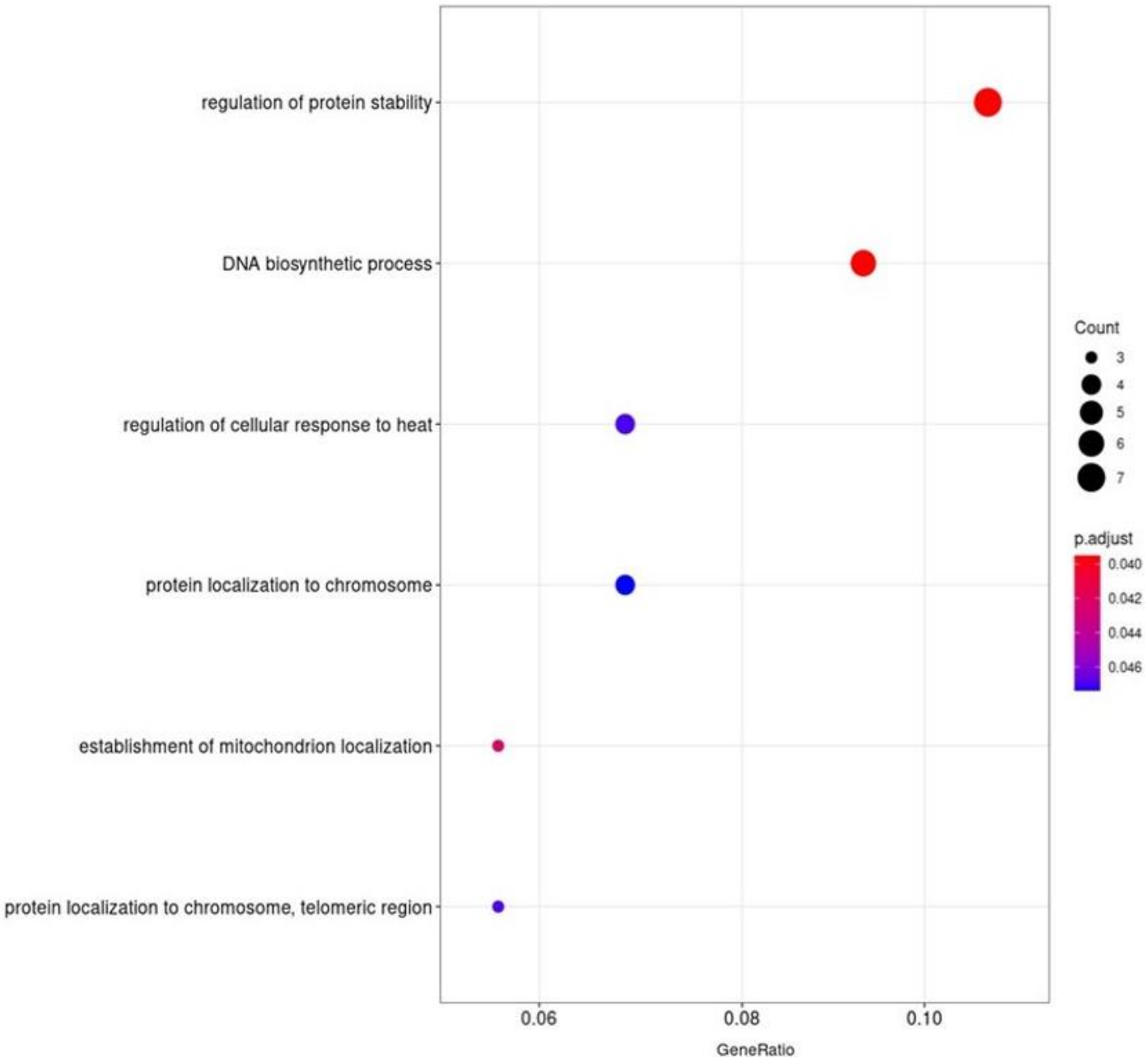


FIGURE 9. Dotplot up regulated genes in Mesothelioma. This dotplot visualizes the genes identified as over-regulated during the mesothelioma study. The dots represent the set of genes expressed for each specific biological process, and the diameter of the dots reflects the quantity of genes associated with that particular process. On the x-axis is the GeneRatio, which represents the fraction of genes associated with

a specific process compared to the total number of genes expressed or analyzed in the study. This metric is useful for evaluating the relative importance of a biological process compared to others regarding gene over-regulation.

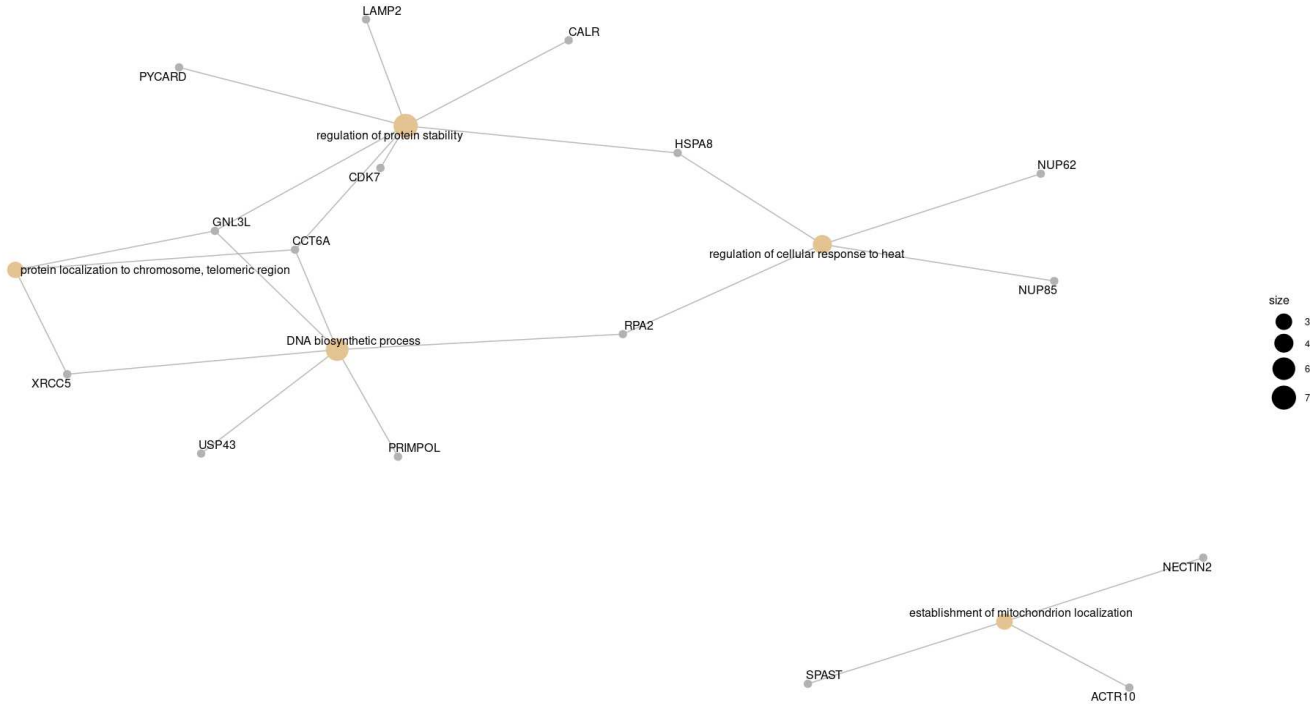


FIGURE 10. Cneplot up regulated genes in Mesothelioma. In this figure, a Cneplot is used to visualize the genes that were identified as up-regulated during the mesothelioma study. The Cneplot provides a detailed analysis of the interactions between these genes and can help identify clusters or groups of genes with coordinated expression. The connections between genes are represented by lines. This visual representation provides insight into the network of genes involved in the mesothelioma response and is valuable for understanding the molecular pathways involved in this condition.

Further analysis with the DAVID database pinpointed a pathway correlated with the progression of the cell cycle into mitosis (p-value = 7,3E-3), further emphasizing its potential importance in MPM. In this thesis, we therefore decided to focus on genes involved in the cell cycle because they are highly critical in tumorigenesis by regulating cell growth and division.

4.3.1. Significant up-regulated genes in the MPM and cell cycle pathway

Notably, eight genes up-regulated in the mesothelioma samples and dowregulated in hyperplastic tissues were related to cell cycle progression. These genes include *MCM3*,

CDK7, *NUP62*, *SPAST*, *NUP85*, *PSMC4*, *RPA2*, and *TUBA1B*, analyzed using STRING (Figure 11). These genes have roles in various cellular functions and they are part of a complex, tightly regulated network that governs many processes from DNA replication to the structural organization of cells.

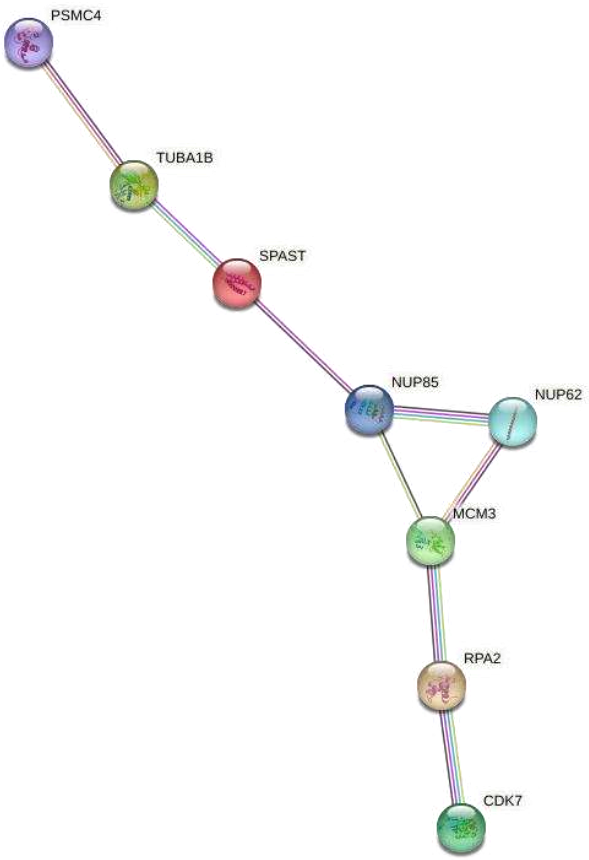


FIGURE 11. STRING analysis of 8 up regulated genes in Mesothelioma, enriched by DAVID database in cell cycle progression pathway.

For instance, RPA2 protein is a critical subunit of replication protein A (RPA), and the main single-stranded DNA binding protein complex in eukaryotes. This protein has fundamental functions such as stabilizing single-stranded DNA strands during replication, repair and recombination. This step is crucial to avoid the formation of secondary DNA structures that could interfere with replication and repair. And is crucial for maintaining genomic integrity. Dysregulation or mutations in RPA2 can have catastrophic consequences, leading to genomic instability. This is particularly

problematic during DNA replication, when the genome is most susceptible to errors that can lead to somatic mutations, a common prerequisite for carcinogenesis (90).

NUP62 and NUP85 are components of the nuclear pore complex, a multi-protein structure that lines nuclear pores and regulates the flow of molecules between the nucleus and cytoplasm. While they have primary functions in nucleus-cytoplasmic transport, they also play important roles in cell cycle, especially in its mitotic phase. During mitosis, the cell nucleus dissolves and reforms, a process that requires precise regulation of the nuclear pore complex. Abnormalities in this regulation can lead to errors in chromosome segregation, one of the key mechanisms in carcinogenesis (91).

These genes contribute to the fine organization that ensures proper cell cycle progression and their dysregulation can lead to MPM carcinogenesis (90,91).

Other genes such as *SPAST*, *PSMC4,* and *TUBA1B*, although there are no data in the literature on MPM, play equally indispensable roles in cellular function.

The *SPAST* gene encodes for spastin, a protein that is crucial for microtubule dynamics in eukaryotic cells. Spastin has the ability to cut microtubules, an essential component of the cellular cytoskeleton. This activity is particularly important during mitosis, the process of cell division, where the precise segregation of chromosomes is essential for genomic stability. Microtubules are dynamic filaments that provide 'scaffolding' for the cell, and their dynamics are finely regulated by a series of proteins, including spastin. During mitosis, the microtubules form the mitotic spindle, a structure that helps separate the chromosomes equally between the two daughter cells. Spastin acts here to ensure that the microtubules are cut and remodeled appropriately, allowing precise chromosome segregation (92,93).

Disregulation of SPAST expression or function can impair microtubule dynamics and, consequently, chromosome segregation during mitosis. This can lead to genomic

instability, one of the risk factors for tumorigenesis. If chromosomes are not properly segregated, daughter cells may inherit an abnormal number of chromosomes or suffer other chromosome abnormalities, creating a favorable environment for tumor formation and progression (93).

The 26S proteasome, a multi-protein complex, plays a key role in maintaining protein homeostasis within the cell. It is responsible for the degradation of proteins that have been labeled for destruction through ubiquitination. PSMC4 (26S Proteasome Regulatory Subunit) is one of the subunits that make up the 26S proteasome, and it is essential in facilitating the degradation of target proteins (94). Proper protein degradation is crucial for several cellular functions, including cell cycle regulation, DNA damage response, and activation and inactivation of signaling pathways. PSMC4, as part of the 26S proteasome, contributes to the control of these crucial functions. Its activity is therefore tightly regulated to ensure that only damaged or excess proteins are degraded, leaving intact proteins necessary for normal cell function (94).

Deregulation of the 26S proteasome can have dramatic consequences, contributing to malignant cell transformation and tumor progression (94).

The *TUBA1B* gene encodes for tubulin alpha-1B, one of the protein subunits of microtubules (95). Tubulin alpha-1B is a key component in the formation of microtubules, which are essential for the mitotic spindle, the structure that separates chromosomes during mitosis. When this process is efficient and precise, it ensures that each daughter cell inherits an exact copy of the genome. TUBA1B is therefore central in maintaining genomic integrity during cell division (95).

Altered funtion of TUBA1B can lead to a defective mitotic spindle and, consequently, inaccurate chromosome segregation and genome instability (96).

Instead, altered gene expression patterns of *MCM3* and *CDK7* have emerged as focal points in various studies, also focusing on MPM (97,98).

4.3.1.1. MCM3 and its Implications in Mesothelioma

The role of the MCM family of proteins, particularly MCM3, in cellular processes, particularly DNA replication, has been a focal point of numerous studies. These proteins are important in cell cycle dynamics and are therefore subjects of intensive research (99).

MCM3, or Mini-Chromosome Maintenance Complex Component 3, is recognized as a potential replicative helicase, imperative for both initiating and elongating DNA replication processes. This protein is a member of the MCM family, which includes proteins MCM2-7. All members of this family play pivotal roles during the DNA replication phases of the cell cycle (99).

Of particular interest is the crucial role MCM3 plays in safeguarding the integrity and completeness of DNA replication. A change in its expression levels or function can hinder DNA replication and lead to DNA anomalies, laying the foundation for oncogenic transformations (100). Numerous studies have demonstrated aberrations in MCM3 expression across diverse tumor types. A study by Oluf Dimitri Røe et al. 2009 (97) demonstrated that MCM2, 3, and 6 are overexpressed in mesothelioma, and are linked to a poor prognosis. This link is particularly evident for MCM3, which is, overexpressed in a wide spectrum of neoplasms (97). This study demonstrated a 90% of mice exhibiting epithelial tumors within six weeks post-injection with cells transfected with MCM3 (97). The role of MCM3 in DNA replication and its altered expression in tumorigenesis underscores its potential as a biomarker and therapeutic target (97).

4.3.1.2. CDK7 and its Implications in Mesothelioma

CDK7 (Cyclin-Dependent Kinase 7) is a serine/threonine kinase with a dual role in the regulation of cell cycle and RNA transcription by RNA polymerase II. This kinase is the catalytic subunit of CAK (Cyclin-dependent Activating Kinase), which is essential for several crucial transitions during the cell cycle. Notably, CAK itself is a trimeric complex composed of CDK7, cyclin-H and Mat1. This complex plays a crucial role in cellular processes, being essential for cell cycle progression, which includes the transition from the G2 phase to the M phase, as well as from the G1 phase to the S phase (101).

Indeed, the role of CDK7 in the G1-S transition involves the activation of CDK2 through the formation of the CDK2/cyclin E complex. The role of the CAK complex, in this case, is to catalyze the phosphorylation of CDK2, which allows the binding to cyclin E. This is an crucial step for DNA replication and subsequent progression into S phase, where DNA synthesis occurs (101).

Similarly, the activation and formation of the CDK1/cyclin-B complex during the G2-M transition is highly dependent on CDK7. When CDK7 phosphorylates CDK1, it activates CDK1, thereby allowing binding to cyclin-B. This CDK1/cyclin-B complex, is critical for driving cells from the G2 phase to mitosis (M phase). The initiation of the G2-M transition is a critical checkpoint in cell cycle regulation, ensuring proper cell transition into mitosis (102).

Recent studies have shed light on CDK7 as a potential therapeutic target for MPM. A study by Miao J. et al., 2020 has shown that inhibition of CDK7 led to a reduction in the level of YAP protein, thereby suppressing migration and invasion of MPM cells (98).

This suggests the important potential of CDK7 not only as a molecular marker for understanding pathological mechanisms but also as a target for pharmaceutical interventions in mesothelioma (98).

Furthermore, during the G2-M transition, it is important to point out that CDK7 also indirectly influences other regulatory proteins such as Greatwall kinase (Gwl) and ARPP19, which in turn inhibit protein phosphatase 2A (PP2A). PP2A acts as a regulator of cell signaling, inhibiting cell cycle progression. CDK7, through its indirect effect on PP2A, adds an additional layer of complexity to cell cycle regulation (103). Activation of Gwl kinase and ARPP19 through CDK7 essentially modulates PP2A activity, thereby affecting a number of processes that contribute to cell cycle progression (104).

4.4. GEO database gene comparison

Once the results of our analysis were obtained, given the limited number of samples used, which could make our results less relevant, we wanted to validate and strengthen their validity by comparing them with the results of a larger number of samples. For this reason, we downloaded 41 paired samples of MPM and control normal tissue from the GEO database (GSE51024 dataset) and performed differential expression analysis using R's limma package (85).

We subsequently compared the differential gene expression profiles between our initial data set and the limma-analyzed GSE51024 set to validate our results. Despite methodological and normalization differences between the two sets, the focus was specifically on differentially expressed genes. This choice was made to identify shared coherences and trends between the two sets in order to give greater robustness and reliability to the results obtained.

Using Venn diagrams, depicted in Figure 12, we identified 22 common upregulated mesothelioma genes consistent across both analytical methods of the two analyzed datasets.

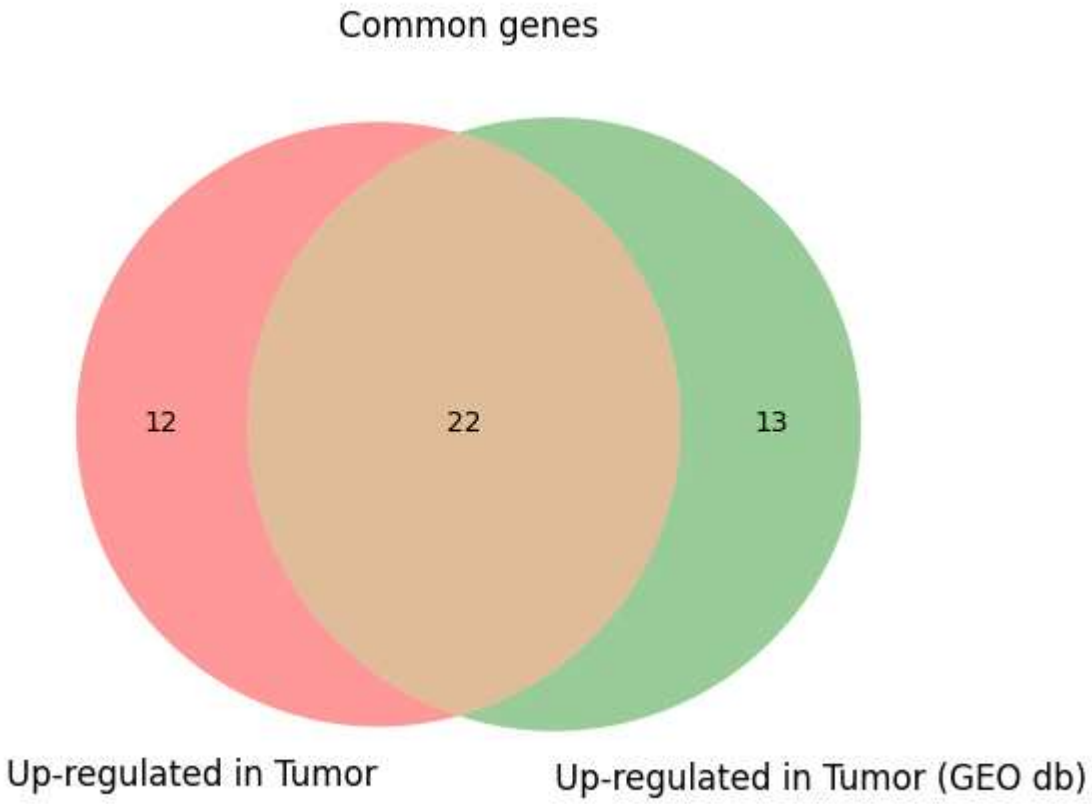


FIGURE 12. Venn diagram of genes in MPM: this Venn diagram represents the intersection of differentially expressed upregulated genes in MPM. The red circle represents upregulated genes in mesothelioma obtained by differential expression analysis with Deseq2, while the green circle represents up-regulated genes in mesothelioma obtained by limma analysis of Geo data. The overlapping area between them indicates the upregulated genes in mesothelioma shared between the two datasets.

Common genes between the two experiments were represented by heatmaps, to visualize common genes within our experiment (Figure 13).

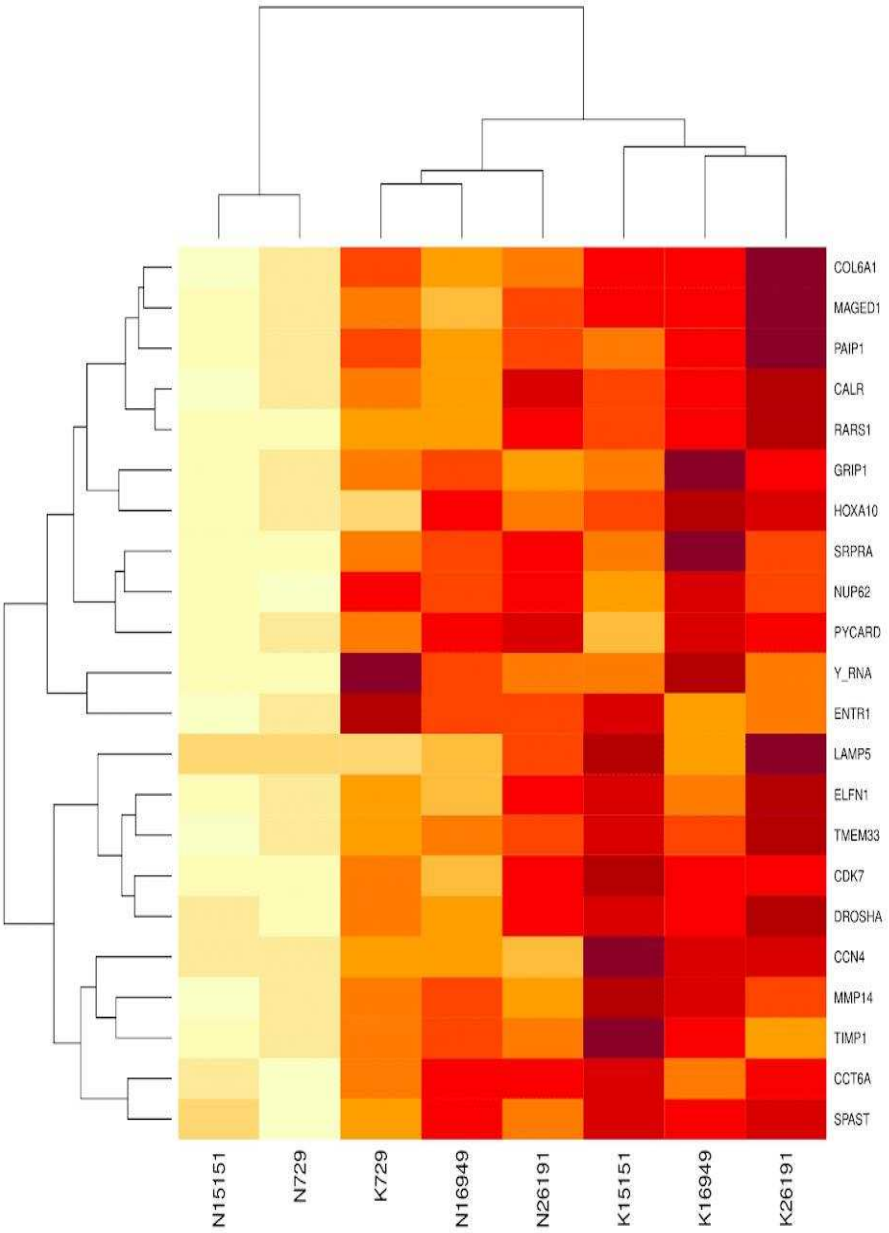


FIGURE 13. Heatmap representing common up-regulated genes in mesothelioma between our dataset and GEO dataset.

Notably, among these 22 genes, the upregulation of the *CDK7* gene and *NUP62* were confirmed. The elevated expression levels of *NUP62* may be influenced and driven by the altered activity of *CDK7*. Instead, the *CDK7* gene is closely associated with malignancy, making it a potential biomarker for the progression or severity of mesothelioma.

Subsequently, another analysis with Venn diagrams was conducted to compare gene expression between normal tissue of the GSE51024 dataset and hyperplasia of our study. This was done to identify up-regulated genes in the hyperplasia still associated with tissue benignity. It is illustrated in Figure 14, where 165 common upregulated genes were identified.

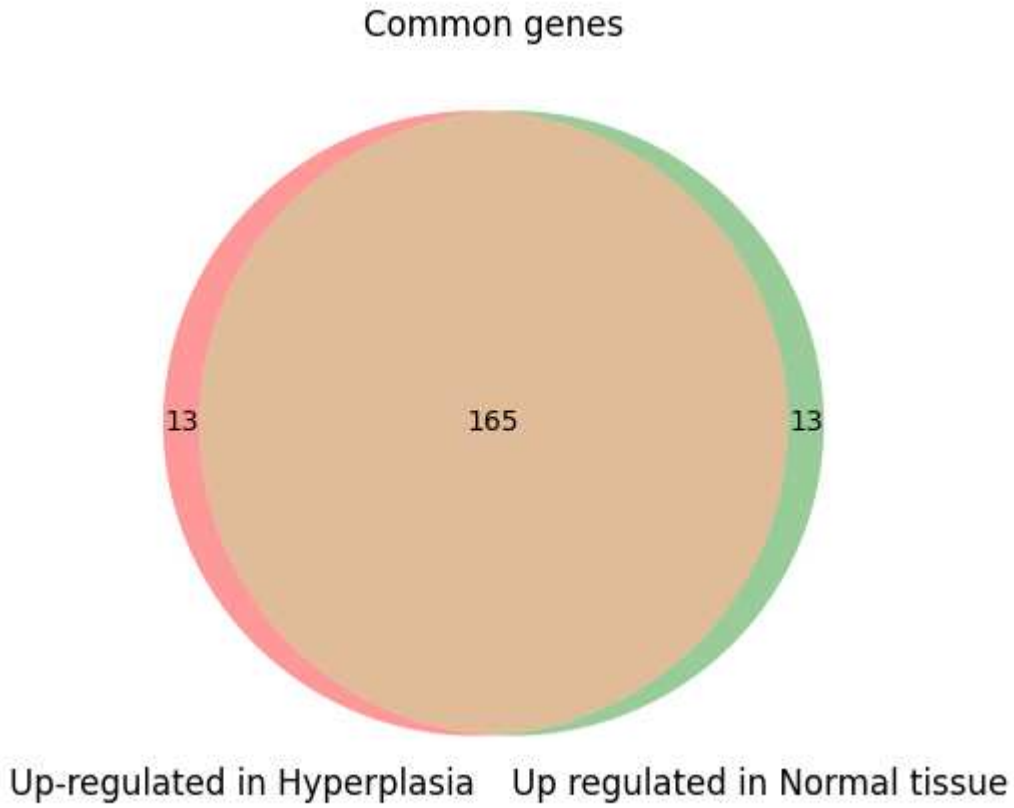


FIGURE 14. Venn diagram genes associated with benignity: this Venn diagram represents the intersection of up-regulated genes in hyperplasia of our study and up-regulated genes in normal samples from the dataset GSE51024. The red circle represents up-regulated genes in hyperplasia obtained by differential expression analysis with Deseq2, while the green circle represents up-regulated genes in normal tissue (GSE51024) obtained by limma analysis. The overlap area between the two indicates the up-regulated genes shared between hyperplastic tissue and normal tissue.

The common genes between the two experiments were represented by heatmap (Figure 15).

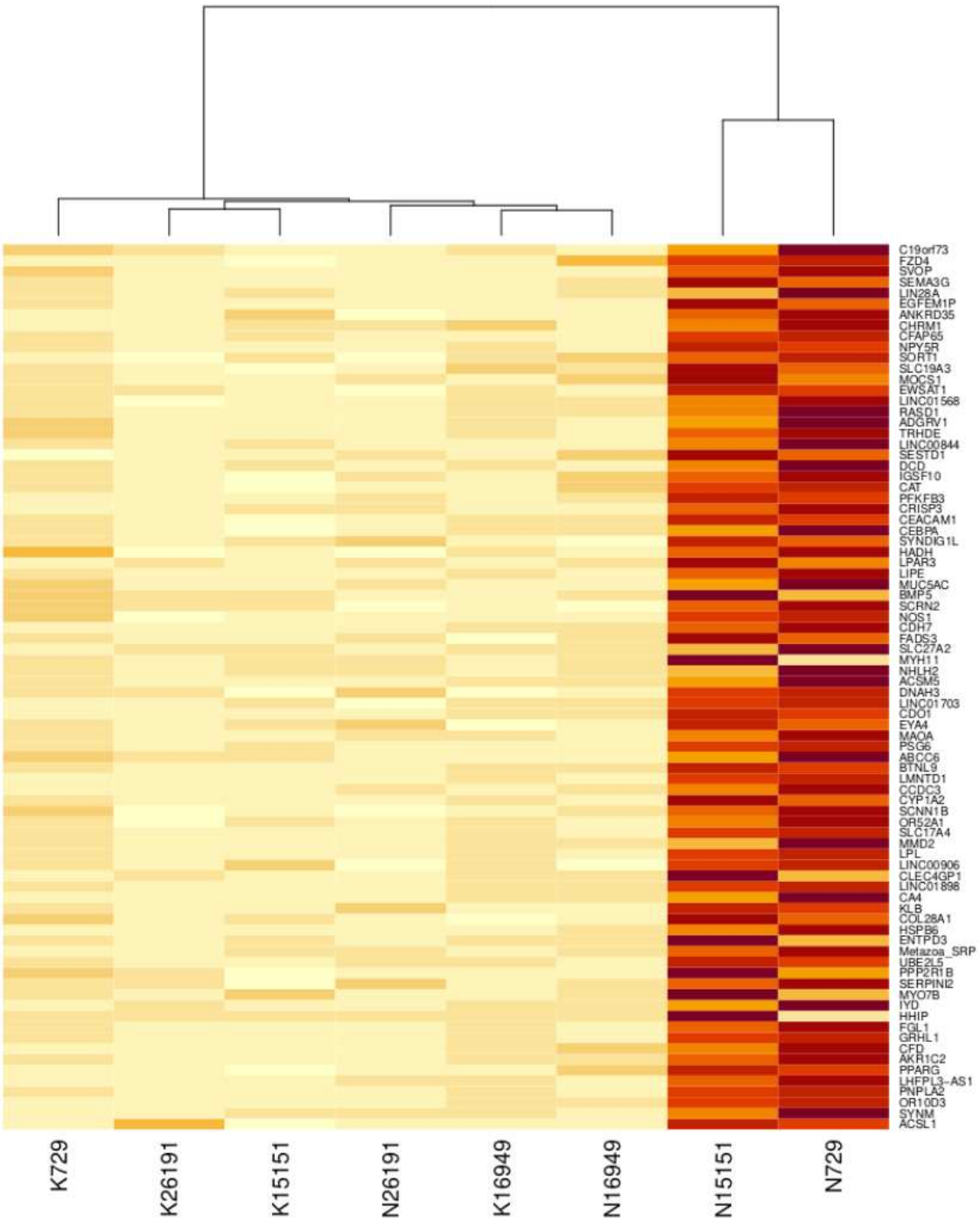


FIGURE 15. Heatmap representing common up-regulated genes in hyperplasia and normal tissue between our dataset and GEO dataset.

Among the genes identified, the common upregulation of the PP2R1B gene in both our dataset among the hyperplastic samples and in the GSE51024 dataset among the normal tissue samples stands out interestingly in contrast to the upregulation of the malignancy-associated CDK7 gene identified in the malignant pleural mesothelioma of both datasets compared.

4.4.1. PP2R1B and its Implications in Tumorigenesis

The PPP2R1B gene (Protein Phosphatase 2 Regulatory Subunit A Beta) encodes the scaffolding (or A) subunit beta of protein phosphatase 2 (PP2A). PPP2R1B is responsible for the structural integrity of the PP2A complex and plays a crucial role in its assembly. The PP2A enzyme complex is composed of three components: the catalytic subunit C, the structural scaffolding subunit A, and a variable regulatory subunit B (103).

PPP2R1B can influence several cellular processes, including cell cycle exit.

Several studies have highlighted the importance of the PPP2R1B gene in tumor suppression. Wang et al. (1998) were pioneers in recognizing it as a putative human tumor suppressor gene (105). Their study found somatic alterations in the PPP2R1B gene in 15 percent of primary lung cancers, 6 percent of lung cancer-derived cell lines, and 15 percent of primary colon cancers (105). Importantly, these mutations generated a truncated PP2A-Abeta protein unable to bind to the PP2AC subunit, suggesting a crucial role of the PPP2R1B gene product in cell cycle regulation and, consequently, tumor suppression (105).

Several other studies, such as those conducted by Schonthal et al., 2001 (98), Esplin E. et al., 2006 (106), Sablina et al., 2007(107), and Ruvolo et al., 2016 (108), have further corroborated the crucial role of the PPP2R1B gene in tumor suppression. In particular, Schonthal et al., 2001 examined several PP2A complexes and their implication in cancer development, pointing out that mutations in the PP2A Aβ structural subunit occur in many human cancers (109,110).

Sablina et al., 2007 (107) further elaborated on this point, showing that introducing an Aβ WT allele into lung cancer cell lines with Aβ mutations partially reverses the tumorigenic phenotype. Finally, the work revealed that suppression of A-beta

expression of PP2A allowed immortalized human cell lines to acquire a tumorigenic state.

We then wanted to show via pie charts (Figures 16 and 17) changes in the expression of different PP2A subunits between the normal tissues of the GSE51024 dataset and the hyperplastic tissues of our study. In particular, as can be seen from the graph, the PPP2R1B subunit appears significant and upregulated in normal tissues from the GSE51024 dataset along with other subunits of the PP2A protein. Subsequently, of all the PP2A subunits, only PPP2R1B appears to be significant and upregulated in the hyperplastic tissue of our study, and then disappears completely in malignant pleural mesothelioma. It is interesting to note this progressive silencing of this important PP2A subunit in the transition from normal, hyperplastic to tumor tissue.

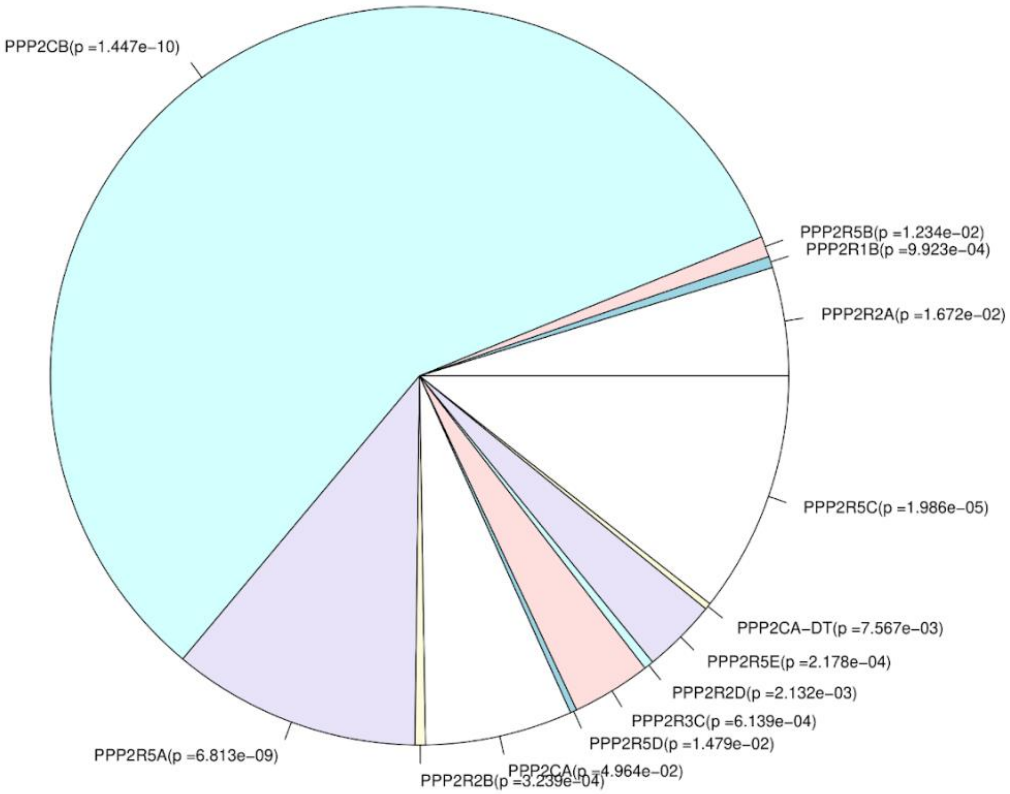


FIGURE 16. Pie chart representing all subunits of PP2A up regulated in normal tissue of the GSE51024 dataset.

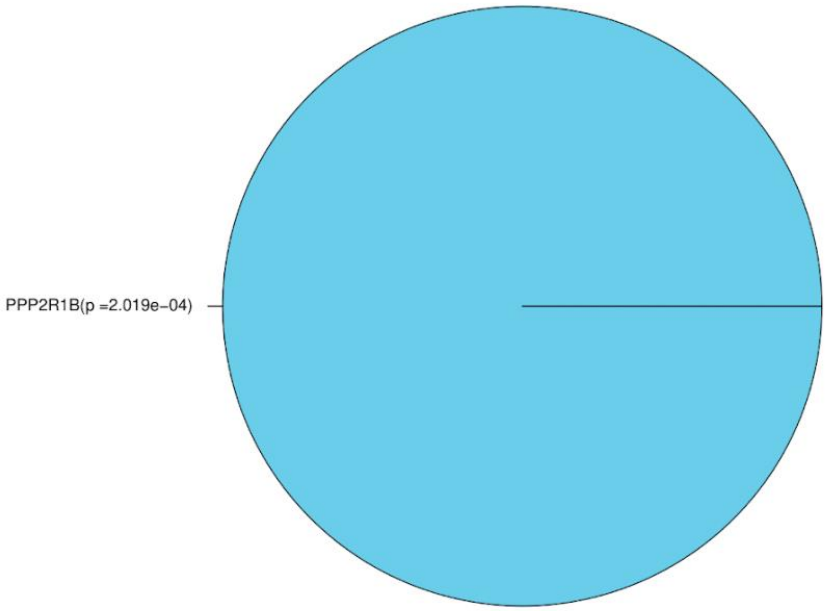


FIGURE 17. Pie chart representing the only significant and upregulated PP2A protein subunit (PPP2R1B) in hyperplastic tissue of our study.

5. Conclusion

In conclusion, this thesis has employed a multi-faceted approach, utilizing RNA sequencing, bioinformatic differential expression, and statistical analysis to detect differentially expressed genes between mesothelial hyperplasia and MPM. These approaches have provided a reliable means of analyzing changes in gene expression, adding robustness to our results. Relevant is the identification of key pathways through enrichment analyses. Specifically, we discovered a pronounced enrichment of the lipid metabolism pathway in hyperplastic tissues, suggesting a potential functional role for lipid metabolic processes in hyperplasia.

Equally important was the identification of a subset of genes that were upregulated in tumorous tissues and involved in cell cycle progression. The detection of specific genes involved in cell cycle progression deepens our molecular insight into mesothelioma development and opens doors for the exploration of new therapeutic avenues. Among these, the upregulation of CDK7 in mesothelioma samples stands out. Given its pivotal role in the formation of the CDK1/cyclin-B complex and subsequent cell cycle progression, CDK7 emerges as a likely candidate for further research into mesothelioma therapies.

Another significant revelation was the identification of PP2R1B as an upregulated gene in hyperplastic tissues. This gene, involved in the formation of the PP2A holoenzyme complex, is implicated in the negative regulation of cell growth and division. The increased expression of PP2R1B in hyperplastic tissue, juxtaposed with its inhibited state in tumor samples, implicates it as a potential tumor suppressor.

Collectively, these findings not only broaden our understanding of the genetic landscape that differentiates hyperplastic and tumorous tissues but also highlight promising avenues for future research. The upregulated genes CDK7 and PP2R1B, in particular, offer valuable insights for the development of targeted therapies for mesothelioma and other related conditions. Thus, the results of this thesis provide a substantive foundation for the advancement of more effective diagnostic tools and treatments.

6. Future perspectives

This thesis has laid a foundational groundwork that warrants further exploration in several directions. First and foremost, the markers identified, particularly CDK7 and PP2R1B, necessitate more comprehensive *in vivo* and *in vitro* studies to confirm their roles and potential as therapeutic targets. As this research was limited by the number of samples, especially in differentiating between normal, hyperplastic, and tumorous

tissues, future work should involve differential expression analysis utilizing a larger dataset. This will provide a more robust statistical framework to confirm the markers and pathways identified.

Finally, the enrichment of the lipid metabolism pathway in hyperplastic tissues represents an intriguing avenue of investigation. Future studies will delve deeper into the role of lipid metabolism in mesothelial hyperplasia, aiming to provide further mechanistic insights into how hyperplastic and MPM tissues differ at the gene expression level.

Our future goal will be to use the results of this study as a basis to try to understand in greater depth and detail the differences at the molecular level between mesothelial hyperplasia and malignant pleural mesothelioma so as to enable more immediate diagnosis and effective therapies against this cancer.

Side project

Differential expression analysis using Deseq2 is a technique widely used by researchers to identify differentially expressed genes between different physiological and pathological conditions. In this thesis, we have used this tool to detect variations in the transcriptome pathological of a rare tumor, such as malignant pleural mesothelioma and one of the most frequent lung neoplasm histotypes: adenocarcinoma. It is precisely on lung adenocarcinoma that we now focus on.

1. Introduction

1.1. Lung adenocarcinoma

Lung cancer (LC) represents the most prevalent form of cancer and the primary reason of cancer-related deaths worldwide (110). There are an estimated 2.20 million new cases of LC each year globally, accounting for about 12 percent of all cancer cases. In 2020, LC was estimated to cause about 1.8 million deaths worldwide, accounting for approximately 19% of total cancer-related deaths (111). It is classifiable into two main categories: small cell lung cancer, which is less common and affects about 10-15% of patients, and non-small cell lung cancer (NSCLC) comprising 85% of cases (111). Among NSCLCs, lung adenocarcinoma (LUAD) is the most prevalent subtype and is detected in about 50% of patients with NSCLC (112). The incidence of lung adenocarcinoma has increased significantly over the past two decades and is now the most common form of lung cancer. It represents the most frequent form in females and men younger than 50 years of age. This lesion is usually peripheral and at the level of the smaller caliber bronchi and often arises near areas of parenchymal scarring. This tumor is a multifactorial disease andthe result of a combination of genetic predisposition and exogenous factors (3).

Genetic factors are responsible for the interindividual variability in susceptibility to carcinogens. In fact, with equal exposure to exogenous factors, such as tobacco smoke and occupational carcinogens, only some individuals develop lung cancer. In addition, a family history of lung cancer heightens the likelihood of being diagnosed with the disease, especially among non-smokers (113).

Exogenous factors that can play a role in the onset of lung cancer include: cigarette smoking, which is the major risk factor, environmental and occupational exposures to asbestos, heavy metals, ionizing radiation, smoke, and dust. Notably, cigarette smoking is now considered the most important risk factor as in fact 80-90% of all patients diagnosed with lung cancer are smokers or former smokers and the incidence of lung cancer has been increasing in parallel with the spread of tobacco use since the late 1800s (113). This risk correlates with how long the individual has smoked and the daily average cigarette consumption. In fact, a chronic smoker of 20 cigarettes per day has 1 in 8 chance of dying of lung cancer, with a risk ratio between smokers and nonsmokers of about 25 to 1, increasing to 60 to 1 for consumers of an average of 40 cigarettes per day and started smoking early in life (113). Conversely, smoking cessation may lead to a reduction in cancer risk, which still remains relevant for 10 to 15 years after smoking cessation and still remains higher than non-smokers even after 40 years (113). Regarding occupational exposure to carcinogens, on average, it is estimated that work-related lung neoplasms constitute 10-15% of the total (112).

Symptomatically, patients with lung adenocarcinoma may present with persistent cough, dyspnea, chest pain, and sometimes hemoptysis (3).

After analyzing the symptoms and risk factors of lung adenocarcinoma, it is critical to consider the environment in which the tumor grows and evolves (114). The microenvironment of the tumor has a crucial role in tumor progression, response to

therapy, and metastasis. It consists not only of tumor cells but also of a variety of other cells, such as cells of the immune system, stromal cells, and blood vessels, which interact with each other and tumor cells in complex ways. These interactions can influence tumor growth, its ability to spread to other organs, and its resistance to treatment (115).

1.2. Tumor microenvironment

In lung adenocarcinoma, the tumor microenvironment (TME) is composed of immune cells that can influence tumor fate and play contrasting roles (116,117). Some of these cells, such as M2-like macrophages (118), regulatory T lymphocytes (119), and myeloid-derived suppressor cells (120), tend to promote tumor and are associated with poor prognosis. Others, such as CD8+ T lymphocytes (121) and M1 polarized macrophages, are associated with a favorable prognosis (122).

For instance, M2-type macrophages are a subpopulation of macrophages that play a role in promoting tumor growth and suppressing anti-tumor immune responses. They can produce immunosuppressive molecules and promote angiogenesis in the TME, thereby contributing to tumor survival and growth (118). Regulatory T lymphocytes are a subpopulation of T lymphocytes that play a role in suppressing anti-tumor immune responses. These cells help maintain a suppressed immune environment in the TME, allowing the tumor to escape immune destruction (119). Finally, myeloid derived suppressor cells (MDSCs) are a heterogeneous population of cells that can suppress antitumor immune responses. They act partly by suppressing T-cell activities and by promoting immunosuppression in the TME (120).

On the other hand, CD8+ T lymphocytes are cells of the immune system that hold a pivotal role in the destruction of tumor cells. Their presence in the TME is often associated with an effective antitumor response and a favorable prognosis (121).

Finally, M1-type polarized macrophages are a sub-population of macrophages that play a role in activating anti-tumor immune responses. They can produce proinflammatory molecules and contribute to the destruction of tumor cells in the TME (122).

In lung adenocarcinoma, the tumor's TME comprises several other components, including tertiary lymphoid structures, which organize T- and B-lymphocytes. A higher proportion of B lymphocytes within tertiary lymphoid structures often correlates with a better prognosis. However, the exact role of tumor-infiltrating B lymphocytes is still debated (123). B lymphocytes within the TME may have conflicting roles. On the one hand, some tumor-infiltrating B lymphocytes produce specific antibodies directed against tumor cells or their antigens, thus eliminating tumor cells and enhancing the anti-tumor immune response. On the other hand, some tumor-infiltrating B lymphocytes may promote tumor growth by producing cytokines and growth factors that stimulate survival and proliferation of tumor cells. In addition, they may contribute to forming a suppressed immune microenvironment that protects the tumor from immune responses (124).

Finally, the extracellular matrix, an acellular component of the TME, holds a vital role in governing the interactions between tumor cells and the surrounding tissue. This influences tumor growth, migration and the possibility of metastasis (125).

Understanding this complexity within the TME is essential for identifying new therapeutic targets and improving our knowledge of tumor biology. This approach is crucial for developing more effective targeted therapies and therapeutic strategies for adenocarcinoma lung cancer, thus contributing to improved treatment options for patients with this form of cancer.

2. Aim

Our goal is to conduct a transcriptomic analysis of 22 patients with lung adenocarcinoma, focusing on 730 genes associated with the immune system, in order to identify variations compared to normal control in genetic transcripts within the immune population as a basis for future treatments.

3. Methods

3.1. Sampling, RNA extraction and transcriptome profiling

Surgical samples from the Azienda Ospedaliero-Universitaria Senese in Siena were processed using the QIAGEN® RNeasy Mini Kit, Germany following the manufacturer's instructions. Tissue samples, from normal lung parenchyma and tumor neoplasm, were prepared by homogenizing them in 1.4 mL of the RLT lysis buffer of the QIAGEN® RNeasy Mini Kit (Catalog Number 74106). The lysates obtained from this procedure were collected in a 15 mL tube in preparation for RNA extraction. The RNA was extracted, using the QIAGEN® RNeasy Mini Kit, and then stored at a temperature of -80 °C. The nCounter® PanCancer Immune Profiling Panel (NanoString Technologies, Inc., Seattle WA, USA), contains a selection of 770 genes pertaining to the profiling of infiltrating immune cells, analysis of both adaptive and innate immune function, assessment of response to immunotherapy, and recognition of tumor-specific antigens. The process included targeted hybridization of the panel probes with RNA extracted from the samples, followed by digital counting of the hybridized molecules for direct quantification of gene expression. In addition, a quality control was implemented that included the use of housekeeping genes, present in the panel, for a first level of data standardisation. Subsequently, the normalization of the raw counts was carried out by applying advanced statistical methods within the DESeq2 software, in order to

guarantee maximum precision and reliability in the analysis of transcriptomic profiles (Figure 18). This panel was used to characterize the transcriptomic profile of the 22 patients with lung adenocarcinoma (LUAD) who were included in the study. Among these patients, 7 are active smokers, 13 are ex-smokers, and 2 are nonsmokers. The cohort is composed of 8 women and 14 men, with an average age of 72 years.

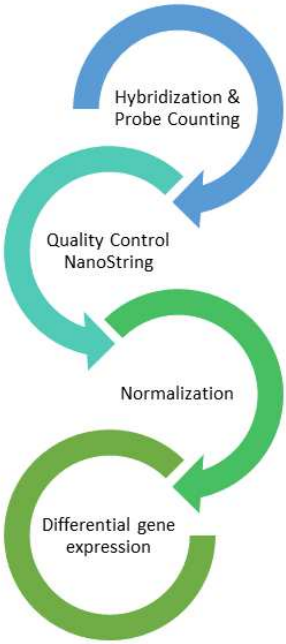


FIGURE 18. Differential gene expression analysis workflow after Nanostring nCounter analysis.

3.2. Differential gene expression analysis

The data obtained from the transcriptomic analysis were analyzed using Deseq2 in RStudio (R-Studio version 3.6). The Deseq2 package was then used to perform static analysis and discover quantitative variations in gene expression levels between the two experimental groups.

To ensure that the gene counts were homogeneous between samples, we adopted the normalization methodology suggested by DESeq2. The normalization process was

carried out following the standard directions of the DESeq2 package, keeping the preset settings unchanged.

In our analytical framework, normal tissue was taken as the benchmark, while tumor tissue was used for comparisons. The coefficients emerging from this framework show the discrepancies in logarithmic gene expression between the two conditions. A coefficient with a value above zero (Log2FC > 0) signals higher gene expression in tumor tissue than in normal tissue, while a value below zero (Log2FC < 0) suggests the opposite. We considered a difference in expression as statistically significant if the gene had a corrected p-value of less than 0.05.

3.3. Graphical representation

We used the Volcano plot to obtain a visual representation of the difference in gene expression (Log2FC), in relation to its statistical significance (-log10 of the p-value), for each gene under consideration. The creation of the Volcano graph was done by taking advantage of the R package called EnhancedVolcano. In addition, to get a clear idea of the distribution of gene expression counts, we used the PlotCount graph. The latter allows to visualize the frequency of counts for a specific gene in all the considered samples, giving a detailed overview of the expression activity of that specific gene in the different samples.

While the approach offered by DESeq2 is primarily focused on identifying and quantifying differences in gene expression between different samples, we also used an additional tool, known as the Microenvironment Cell Populations-counter (MCP-counter). The latter was specifically developed with the goal of accurately quantifying the absolute abundance of as many as seven distinct types of cell populations present in the tumor microenvironment, commonly known as TME (Tumor MicroEnvironment).

Using MCP-counter, we are able to obtain a detailed and clear view of the cellular composition within the TME. This, in turn, provides valuable information about the nature of the interaction between tumor cells and the different cell populations that coexist in their surrounding environment. Using these tools in combination allows not only a deeper understanding of the gene activity within the tumor, but also of how the tumor interacts and responds to the complex environment in which it is located.

4. Results and Conclusions

4.1. Differenial gene expression analysis

Our data, in order to obtain a clear and intuitive visualization of gene expression differences between LUAD samples and normal control samples, were illustrated through a Volcano Plot, as shown in Figure 19. In this graphical representation, each individual dot corresponds to a distinct gene. Variations in gene expression are highlighted through the arrangement and color of each dot: genes that demonstrate significant overexpression in tumor samples compared to non-tumor samples are outlined with blue coloring, contrasting with underexpressed genes, which are instead highlighted with shades of red.

This graph confirms the relevant difference in gene expression between tumor and normal samples and allows for easy identification of genes with significant differential expression.

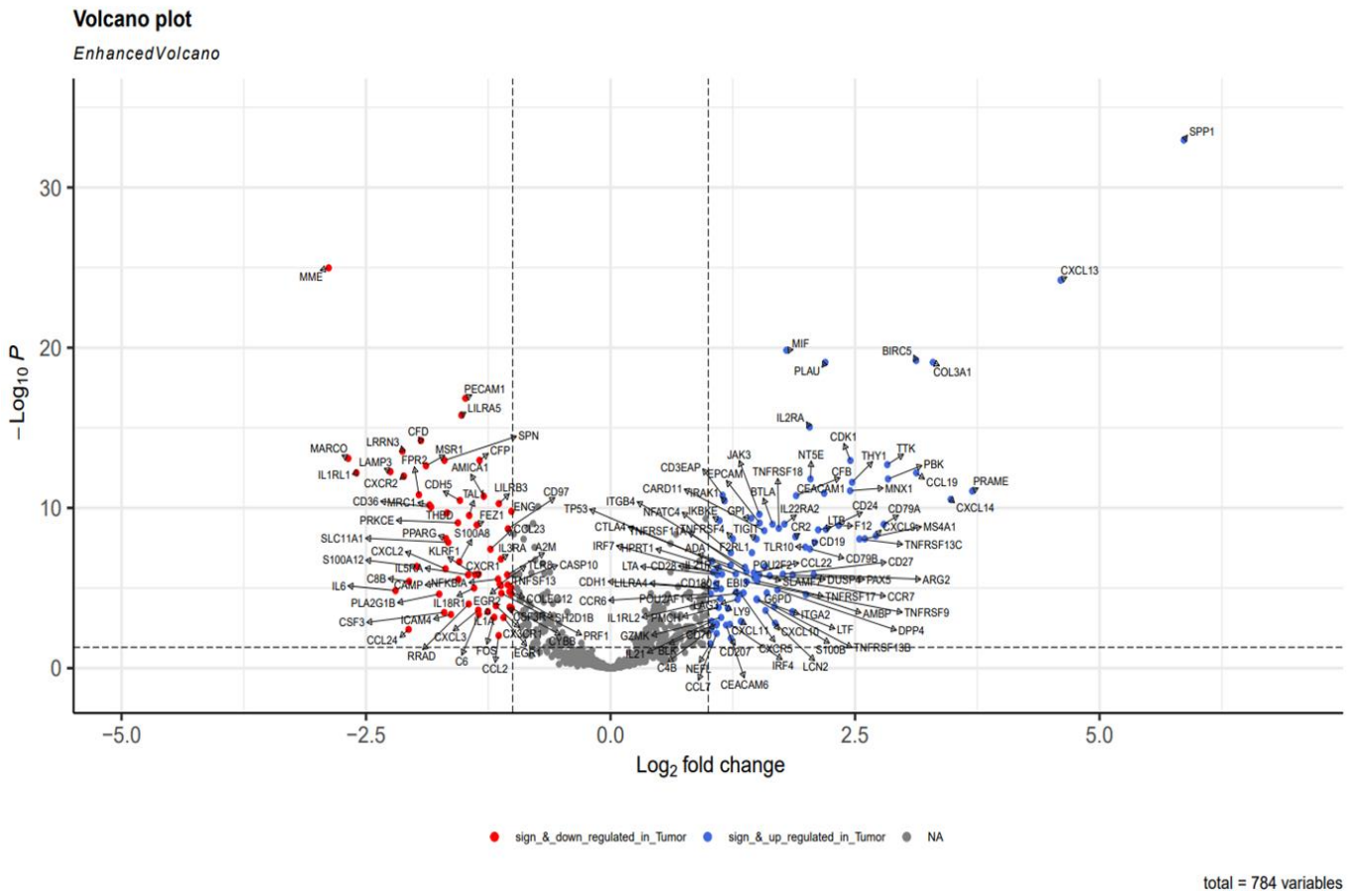


FIGURE 19. Volcano plot showing the relevant difference in gene expression between tumor and normal samples. In the graph, logarithms of fold change (log2FC) are represented on the x-axis, while p-values (-log10) are represented on the y-axis. Genes that exhibit a significant difference in expression between the two groups (both in terms of fold change and statistical significance) are highlighted as points at the top of the graph. These results provide an overview of the alterations in gene expression that characterize the tumor compared to normal samples.

In this thesis, we focused on up-regulated genes in the tumor, consistent with our goal of analyzing variations in gene expression within the immune population of LUAD to identify possible new therapeutic biomarkers against this tumor.

As shown in this graph, the two genes found most upregulated (Log2FC > 3) and statistically significant in the cancerous tissue relative to the normal control are the SPP1 gene (p-value = 1.5e-36) and CXCL13 (p-value = 2.4e-27), represented by the PlotCount graph (Figure 20 and Figure 21).

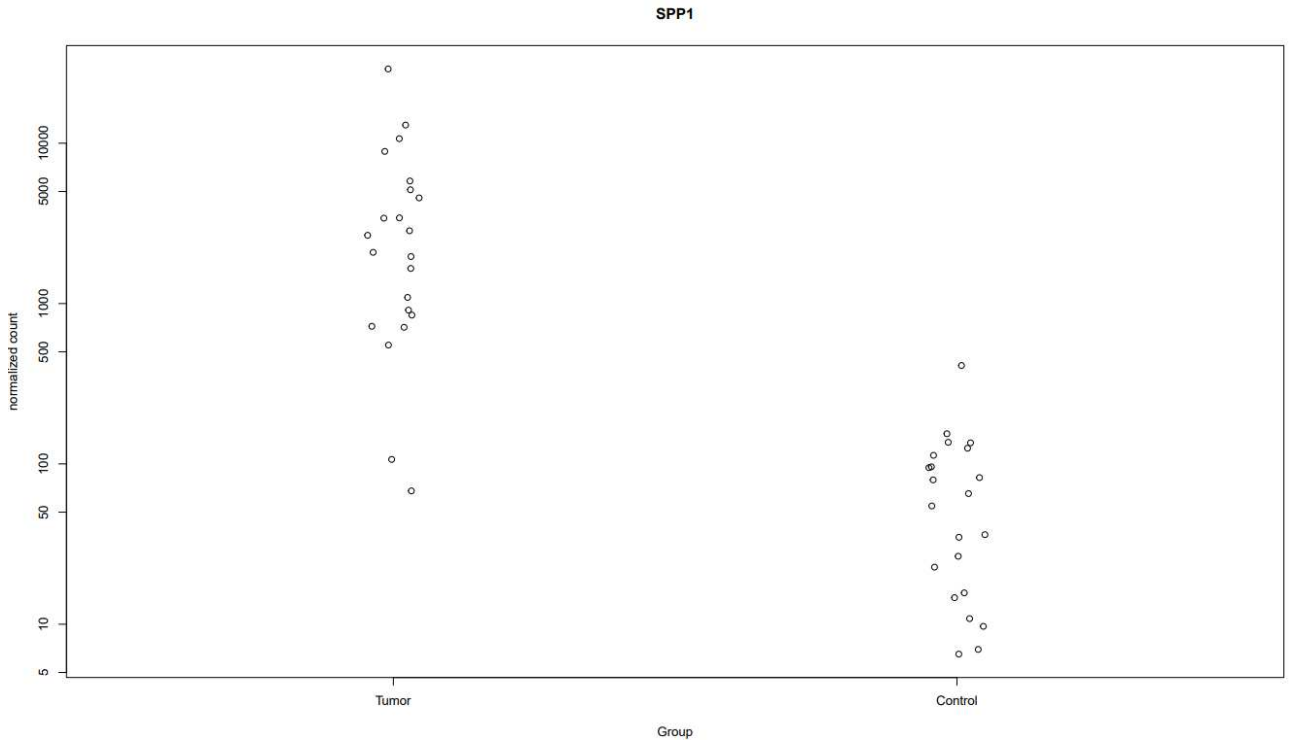


FIGURE 20. PlotCount graph of SPP1 gene. In the graph, the gene expression of normalized counts of the SPP1 gene in tumor samples and normal control samples are shown.

SPP1, also known as osteopontin, is a protein expressed in multiple tissues and organs. It has attracted particular attention in oncology because its expression has been linked to unfavorable outcomes in several malignancies, including lung adenocarcinoma (126).

In lung adenocarcinoma, osteopontin is not only produced by tumor cells, but also by tumor-associated macrophages (TAMs). These macrophages play a key role in tumor progression, contributing to tumor growth, invasion, and metastasis. Expression of osteopontin, in TAMs may therefore offer profound insights into tumor behavior and the interaction between the tumor and its microenvironment (127).

Several research investigations have explored the prognostic significance of SPP1 expression in TAMs in patients with lung adenocarcinoma. They have found that increased expression of this protein in TAMs is closely related to a less favorable clinical course. This link is particularly evident when we consider specific clinical metrics such as cancer-specific survival (CSS) and progression-free survival (PFS) (127,128). Notably,

the high SPP1 expression is associated with poor prognosis, particularly in patients with EGFR-wild-type lung adenocarcinoma, suggesting that in this subgroup of patients osteopontin may play a particularly critical role. These results highlight the potential of targeting SPP1 as a therapeutic strategy in lung adenocarcinoma (128,129).

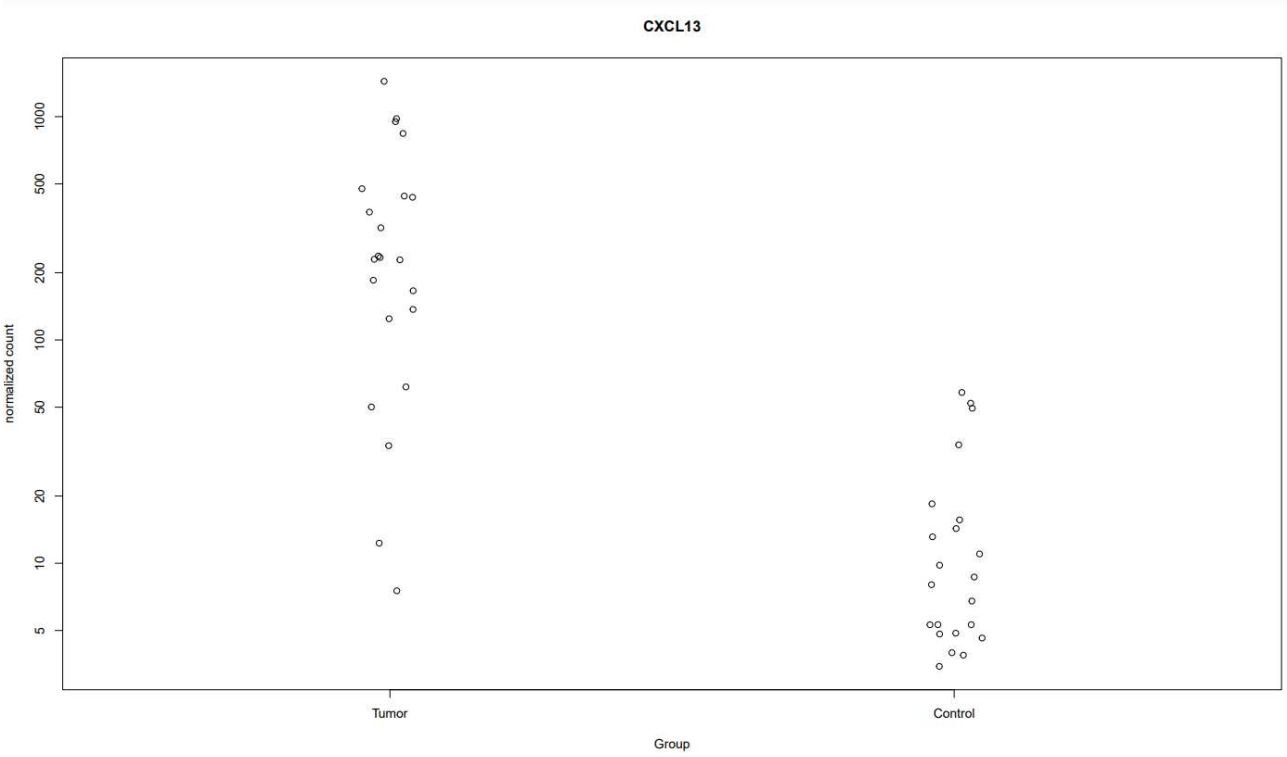


FIGURE 21. PlotCount graph of CXCL13. In the graph, the gene expression of normalized counts of the CXCL13 gene in tumor samples (yes) and normal samples (no) are shown.

In contrast, the CXCL13 gene, known as C-X-C motif chemokine ligand 13, is an essential signaling molecule with roles ranging from orchestrating immune responses to the complexities of tumorigenesis. Traditionally recognized for its central role in the immune system, CXCL13 is critical for guiding B cells to the follicles within secondary lymphoid organs, such as lymph nodes. It is essential for the formation of tertiary lymphoid structures (TLS), ectopic lymphoid aggregates that can develop in nonlymphoid tissues during events of chronic inflammation or specific immune responses. The appearance of these structures, facilitated by CXCL13, has been

considered both a protective mechanism and a contributing factor to autoimmune diseases (130).

However, the depth of CXCL13's biological influence does not end with immune regulation. Recent research paints a picture of a molecule with a dual nature, in which its overexpression in tumor microenvironment can direct signaling in a pro-tumorigenic direction. This transformation becomes evident when CXCL13 binds to its receptor, CXCR5, present in tumor cells. This binding event activates a cascade of intracellular pathways, with the CXCL13-CXCR5 axis identified as a key factor in oncogenesis (131). Activation of this axis can stimulate pathways such as PI3K-Akt, known to safeguard cell survival and counteract apoptosis, and the MAPK/ERK pathway, a central player in promoting cell proliferation (132,133). Taken together, these pathways promote the growth and survival of cancer cells, giving them greater capacity for migration and invasion and thus potentially promoting metastatic spread (134).

CXCL13's dual role in both immune function and tumor biology indicates its potential role as a therapeutic marker. Disruption of CXCL13-CXCR5 signaling could offer a strategy to limit tumor progression, invasion, and therapeutic resistance. Moreover, its influence on B-cell movement provides valuable insights into the mechanisms by which tumors canevade or modify host immune defenses (135).

Its deep involvement in the complex dance of cancer makes it a promising candidate for future research, both as a diagnostic biomarker and as a target for innovative therapeutic interventions.

4.2. Microenvironment Cell Populations-Counter

The next step in our analysis aimed at obtaining information regarding the cellular composition of the tumor microenvironment (TME) of our samples by Microenvironment Cell Populations-counter (MCP-counter) (Figure 22).

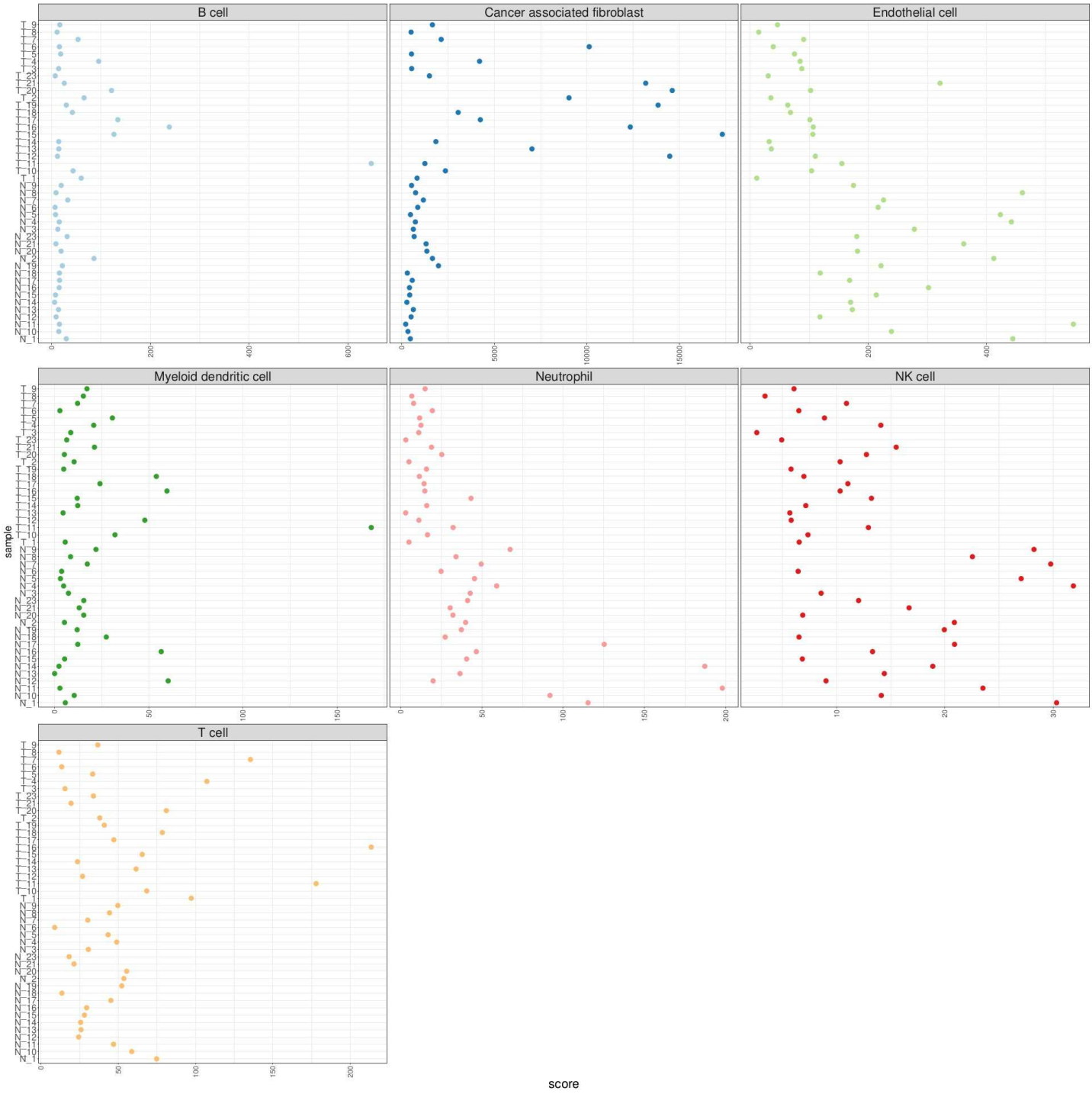


FIGURE 22. MCP-counter graph. This MCP-counter graph shows the quantitative analysis of the different populations of cells within the tumor microenvironment (TME) of LUAD. The different lines represent

estimates of the abundance of cell populations, encompassing immune cells such as T lymphocytes, B cells, NK cells, and myeloid dendritic cells, along with stromal cells such as cancer-associated fibroblasts (CAFs) and endothelial cells. Represented with different colorings. This study offers a comprehensive overview of the cellular composition of the TME, making it possible to assess the dynamics of the immune system and stromal tissue within the tumor.

This tool is designed to quantify the absolute abundance of specific cell populations in heterogeneous samples using transcriptomic gene expression data. Specifically, it targets five cell populations immune cell populations: B cells, myeloid dendritic cells, neutrophils, natural killer cells, and T lymphocytes as well as two stromal cell populations: cancer-associated fibroblasts (CAFs) and endothelial cells (136).

Gene expression data from tumor (T) and noncancer (N) samples were then compared with reference profiles to estimate the proportions of each cell type by deconvolution analysis. This analysis thus provided us with a comprehensive overview of the cellular composition of the tumor microenvironment (137).

The results obtained from our preliminary analysis showed a predominant role of cancer-associated fibroblasts (CAFs) within the tumor microenvironment (TME) of the LUAD samples compared to control samples (138). These fibroblasts, when activated, assume a crucial function in shaping the tumor ecosystem, influencing both tumor growth and immune response dynamics. Various metabolic and molecular pathways are involved in fibroblast activation in the TME (139). For example, transforming growth factor-beta (TGF-β) is responsible for the induction of αSMA marker expression typical of fibroblasts in an active state. Similarly, interleukin-1 (IL-1) modulates the transcriptional factor NF-κB, while interleukin-6 (IL-6) impacts the STAT transcription factors. These biological mechanisms emphasize the crucial role that CAFs play in regulating the tumor environment and tumor evolutionary dynamics (140).

These details underscore how essential it is to gain a deeper understanding of the different cell populations present in the TME of lung adenocarcinoma and how they

correlate with specific patterns of gene expression, with the goal of delineating an increasingly clear picture of tumor biology and identifying potential new therapeutic targets.

5. Future perspectives

In light of the results and conclusions obtained from this analysis, several interesting opportunities for future research emerge. The key to a deeper understanding of lung adenocarcinoma lies in the use of further advanced approaches and tools.

We intend to explore clustering analysis as part of the next steps in our research. This methodology will allow us to further subdivide samples into subgroups based on common gene expression patterns. This approach could reveal distinct tumor subtypes within lung adenocarcinoma, offering valuable information on biological variability and potential differences in treatment response.

Another key aspect will be the in-depth analysis of gene pathways. This strategy will allow us to gain a more detailed insight into the biological processes that drive tumor progression. Identifying specific molecular pathways involved could open the way to new therapeutic targets and more targeted treatment strategies.

Finally, for a more in-depth view of cell populations within the tumor microenvironment, we will also turn to advanced tools such as CIBERSORT. This tool will help us more accurately estimate the composition of cell populations in the tumor microenvironment, providing detailed information on tumor immunity and the role of stromal cells.

We aim to gain a deeper and more detailed understanding of lung adenocarcinoma in order to develop increasingly targeted and personalized therapies, thereby improving the prognosis and well-being of patients with this form of cancer.

Conclusive remarks

At a time when medicine is pushing beyond traditional frontiers, our research has positioned itself at the intersection of molecular biology, bioinformatics and oncology, attempting to delve into some of the many complex mechanisms of cancer.

The two projects, presented in this thesis, although distinct in some respects, converge toward a common goal: use differential expression analysis to search for possible diagnostic and therapeutic biomarkers that could improve patient lives. The main project tackled one of the most difficult challenges: identifying differentially expressed genes between malignant pleural mesothelioma and precancerous lesions to identify possible new diagnostic biomarkers, emphasizing the crucial importance of early and accurate diagnosis. On the other hand, the side project probed the complex field of interactions between lung adenocarcinoma and its microenvironment, laying the foundation for potentially innovative therapies.

The findings presented in this thesis may be considered the springboard for further future efforts to turn discoveries into tangible solutions for patients.

Bibliography

1. Attanoos RL, Churg A, Galateau-Salle F, Gibbs AR, Roggli VL. Malignant mesothelioma and its non-asbestos causes. Vol. 142, Archives of Pathology and Laboratory Medicine. 2018.
2. Moro J, Sobrero S, Cartia CF, Ceraolo S, Rapanà R, Vaisitti F, et al. Diagnostic and Therapeutic Challenges of Malignant Pleural Mesothelioma. Vol. 12, Diagnostics. 2022.
3. Butnor KJ. Controversies and challenges in the histologic subtyping of lung adenocarcinoma. Vol. 9, Translational Lung Cancer Research. 2020.
4. Schumann SO, Kocher G, Minervini F. Epidemiology, diagnosis and treatment of the malignant pleural mesothelioma, a narrative review of literature. Vol. 13, Journal of Thoracic Disease. 2021.
5. Vorster T, Mthombeni J, teWaterNaude J, Phillips JI. The Association between the Histological Subtypes of Mesothelioma and Asbestos Exposure Characteristics. Int J Environ Res Public Health. 2022;19(21).
6. Janosikova M, Nakladalova M, Stepanek L. Current causes of mesothelioma: how has the asbestos ban changed the perspective? Vol. 167, Biomedical Papers. 2023.
7. Hajj GNM, Cavarson CH, Pinto CAL, Venturi G, Navarro JR, de Lima VCC. Malignant pleural mesothelioma: an update. Vol. 47, Jornal Brasileiro de Pneumologia. 2021.
8. Jane Henley S, Larson TC, Wu M, Antao VCS, Lewis M, Pinheiro GA, et al. Mesothelioma incidence in 50 states and the District of Columbia, United States, 2003- 2008. Int J Occup Environ Health. 2013;19(1).
9. Carbone M, Adusumilli PS, Alexander HR, Baas P, Bardelli F, Bononi A, et al. Mesothelioma: Scientific clues for prevention, diagnosis, and therapy. CA Cancer J Clin. 2019;69(5).
10. Bianchi C, Bianchi T. Malignant mesothelioma: Global incidence and relationship with asbestos. Vol. 45, Industrial Health. 2007.
11. World Health Organization (WHO). PRESS RELEASE N° 292. Press Release N° 292. 2020;
12. Odgerel CO, Takahashi K, Sorahan T, Driscoll T, Fitzmaurice C, Yoko-O M, et al. Estimation of the global burden of mesothelioma deaths from incomplete national mortality data. Occup Environ Med. 2017;74(12).
13. Binazzi A, Di Marzio D, Verardo M, Migliore E, Benfatto L, Malacarne D, et al. Asbestos exposure and malignant mesothelioma in construction workers—epidemiological remarks by the italian national mesothelioma registry (Renam). Int J Environ Res Public Health. 2022;19(1).
14. Bianchi C, Bianchi T. Global mesothelioma epidemic: Trend and features. Indian J Occup Environ Med. 2014;18(2).
15. Scherpereel A, Scherpereel A, Opitz I, Berghmans T, Psallidas I, Glatzer M, et al. ERS/ESTS/EACTS/ESTRO guidelines for the management of malignant pleural mesothelioma. Vol. 55, European Respiratory Journal. 2020.
16. Mercadante S, Degiovanni D, Casuccio A. Symptom burden in mesothelioma patients admitted to home palliative care. Curr Med Res Opin. 2016;32(12).
17. Walker SL, Dickson VV, Cacchione PZ. A Pilot Mixed-Methods Study of Malignant Pleural Mesothelioma Symptoms. Oncol Nurs Forum. 2022;49(6).
18. Bianchi C, Giarelli L, Grandi G, Brollo A, Ramani L, Zuch C. Latency periods in asbestos-related mesothelioma of the pleura. European Journal of Cancer Prevention. 1997;6(2).
19. Hiddinga BI, Rolfo C, van Meerbeeck JP. Mesothelioma treatment: Are we on target? A review. Vol. 6, Journal of Advanced Research. 2015.
20. Somekawa K, Horita N, Kaneko A, Tagami Y, Fukuda N, Matsumoto H, et al. Adverse events induced by nivolumab and ipilimumab combination regimens. Ther Adv Med Oncol. 2021;13.

21. Wright JJ, Powers AC, Johnson DB. Endocrine toxicities of immune checkpoint inhibitors. Vol. 17, Nature Reviews Endocrinology. 2021.

22. Andujar P, Lacourt A, Brochard P, Pairon JC, Jaurand MC, Jean D. Five years update on relationships between malignant pleural mesothelioma and exposure to asbestos and other elongated mineral particles. J Toxicol Environ Health B Crit Rev. 2016;19(5–6).

23. Herbert RA, Janardhan KS, Pandiri AR, Cesta MF, Chen V, Miller RA. Lung, Pleura, and Mediastinum. In: Boorman's Pathology of the Rat. 2018.

24. Fulciniti F, Barizzi J, Cappellesso R, Nicolè L, Fassina A, Takinishi Y, et al. Diagnostic Pathology of Malignant Pleural Mesothelioma: Cytological, Histological, Immunohistochemical Features. In: Malignant Pleural Mesothelioma: A Guide for Clinicians. 2019.

25. Husain AN, Colby T, Ordonez N, Krausz T, Attanoos R, Beasley MB, et al. Guidelines for pathologic diagnosis of malignant mesothelioma: 2012 update of the consensus statement from the International Mesothelioma Interest Group. Vol. 137, Archives of Pathology and Laboratory Medicine. 2013.

26. Chapel DB, Vivero M, Sholl LM. Malignant and borderline mesothelial tumors of the pleura. In: Practical Pulmonary Pathology: A Diagnostic Approach. 2022.

27. Bruno R, Alì G, Poma AM, Proietti A, Libener R, Mariani N, et al. Differential Diagnosis of Malignant Pleural Mesothelioma on Cytology. The Journal of Molecular Diagnostics. 2020;22(4).

28. Pippa R, Boffo S, Odero MD, Giordano A. Data mining analysis of the PP2A cell cycle axis in mesothelioma patients. J Cell Physiol. 2020;235(6).

29. Husain AN, Colby T V., Ordóñez NG, Allen TC, Attanoos RL, Beasley MB, et al. Guidelines for pathologic diagnosis of Malignant Mesothelioma: 2017 Update of the consensus statement from the International Mesothelioma Interest Group. Vol. 142, Archives of Pathology and Laboratory Medicine. 2018.

30. Portin P. The birth and development of the DNA theory of inheritance: Sixty years since the discovery of the structure of DNA. J Genet. 2014;93(1).

31. Ergin S, Kherad N, Alagoz M. RNA sequencing and its applications in cancer and rare diseases. Vol. 49, Molecular Biology Reports. 2022.

32. Wang Z, Gerstein M, Snyder M. RNA-Seq: A revolutionary tool for transcriptomics. Vol. 10, Nature Reviews Genetics. 2009.

33. Fassbinder-Orth CA. Methods for quantifying gene expression in ecoimmunology: From qPCR to RNA-Seq. In: Integrative and Comparative Biology. 2014.

34. Negi A, Shukla A, Jaiswar A, Shrinet J. Applications and challenges of microarray and RNA-sequencing. In: Bioinformatics: Methods and Applications. 2021.

35. Metzker ML. Sequencing technologies the next generation. Vol. 11, Nature Reviews Genetics. 2010.

36. Ozsolak F, Milos PM. RNA sequencing: Advances, challenges and opportunities. Vol. 12, Nature Reviews Genetics. 2011.

37. Emiyu K, Lelisa K. Review on Illumina sequencing technology. Austin J Vet Sci & Anim Husb. 2022;9(1).

38. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. 2010;11(10).

39. Deshpande D, Chhugani K, Chang Y, Karlsberg A, Loeffler C, Zhang J, et al. RNA-seq data science: From raw data to effective interpretation. Vol. 14, Frontiers in Genetics. 2023.

40. Oshlack A, Robinson MD, Young MD. From RNA-seq reads to differential expression results. Vol. 11, Genome Biology. 2010.

41. Andrews S, others. FastQC: a quality control tool for high throughput sequence data. 2010. Https://WwwBioinformaticsBabrahamAcUk/Projects/Fastqc/. 2019;

42. Andrews S. FastQC - A quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/. Babraham Bioinformatics. 2010;

43. Babraham. Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data. Vol. 5, Soil. 2020.

44. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: Summarize analysis results for multiple tools and samples in a single report. Bioinformatics. 2016;32(19).

45. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14).

46. Dobin A, Gingeras TR. Mapping RNA-seq Reads with STAR. Curr Protoc Bioinformatics. 2015;51(1).

47. Bayat A, Gaëta B, Ignjatovic A, Parameswaran S. Pairwise alignment of nucleotide sequences using maximal exact matches. BMC Bioinformatics. 2019;20(1).

48. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: Ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1).

49. Del Fabbro C, Scalabrin S, Morgante M, Giorgi FM. An extensive evaluation of read trimming effects on illumina NGS data analysis. PLoS One. 2013;8(12).

50. Liao Y, Shi W. Read trimming is not required for mapping and quantification of RNA-seq reads at the gene level. NAR Genom Bioinform. 2020;2(3).

51. NCI Genomic Data Commons. The National Cancer Institute (NCI). 2017. mRNA Analysis Pipeline.

52. Anders S, Pyl PT, Huber W. HTSeq-A Python framework to work with high-throughput sequencing data. Bioinformatics. 2015;31(2).

53. Päll T, Luidalepp H, Tenson T, Maiväli Ü. A field-wide assessment of differential expression profiling by high-throughput sequencing reveals widespread bias. PLoS Biol. 2023;21(3).

54. McDermaid A, Monier B, Zhao J, Liu B, Ma Q. Interpretation of differential gene expression results of RNA-seq data: Review and integration. Vol. 20, Briefings in Bioinformatics. 2019.

55. Kebschull M, Juliafittler M, Demmer RT, Papapanou PN. Differential expression and functional analysis of high- throughput - Omics data using open source tools. In: Methods in Molecular Biology. 2017.

56. Kakati T, Bhattacharyya DK, Barah P, Kalita JK. Comparison of Methods for Differential Co-expression Analysis for Disease Biomarker Prediction. Vol. 113, Computers in Biology and Medicine. 2019.

57. Wen H, Gallo RA, Huang X, Cai J, Mei S, Farooqi AA, et al. Incorporating Differential Gene Expression Analysis with Predictive Biomarkers to Identify Novel Therapeutic Drugs for Fuchs Endothelial Corneal Dystrophy. J Ophthalmol. 2021;2021.

58. Liu S, Wang Z, Zhu R, Wang F, Cheng Y, Liu Y. Three differential expression analysis methods for rna sequencing: Limma, edger, deseq2. Journal of Visualized Experiments. 2021;2021(175).

59. Costa-Silva J, Domingues D, Lopes FM. RNA-Seq differential expression analysis: An extended review and a software tool. Vol. 12, PLoS ONE. 2017.

60. Liang W, Sun F, Zhao Y, Shan L, Lou H. Identification of Susceptibility Modules and Genes for Cardiovascular Disease in Diabetic Patients Using WGCNA Analysis. J Diabetes Res. 2020;2020.

61. Zhao T, Wang Z. GraphBio: A shiny web app to easily perform popular visualization analysis for omics data. Front Genet. 2022;13.

62. Wodrich MD, Sawatlon B, Busch M, Corminboeuf C. The Genesis of Molecular Volcano Plots. Acc Chem Res. 2021;54(5).

63. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12).

64. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. 2010;11(3).

65. Schurch NJ, Schofield P, Gierliński M, Cole C, Sherstnev A, Singh V, et al. How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? RNA. 2016;22(6).

66. Stupnikov A, McInerney CE, Savage KI, McIntosh SA, Emmert-Streib F, Kennedy R, et al. Robustness of differential gene expression analysis of RNA-seq. Comput Struct Biotechnol J. 2021;19.

67. Andrade C. The P value and statistical significance: Misunderstandings, explanations, challenges, and alternatives. Vol. 41, Indian Journal of Psychological Medicine. 2019.

68. Carlström EL, Niazi A, Etemadikhah M, Halvardson J, Enroth S, Stockmeier CA, et al. Transcriptome analysis of post-mortem brain tissue reveals up-regulation of the complement cascade in a subgroup of schizophrenia patients. Genes (Basel). 2021;12(8).

69. Chumbley JR, Friston KJ. False discovery rate revisited: FDR and topological inference using Gaussian random fields. Neuroimage. 2009;44(1).

70. Tarazona S, García-Alcalde F, Dopazo J, Ferrer A, Conesa A. Differential expression in RNA-seq: A matter of depth. Genome Res. 2011;21(12).

71. Farahbod M, Pavlidis P. Differential coexpression in human tissues and the confounding effect of mean expression levels. Bioinformatics. 2019;35(1).

72. Evans C, Hardin J, Stoebel DM. Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions. Brief Bioinform. 2018;19(5).

73. Elhaik E. Principal Component Analyses (PCA)-based findings in population genetic studies are highly biased and must be reevaluated. Sci Rep. 2022;12(1).

74. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, et al. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. 2004;5(10).

75. Soneson C, Delorenzi M. A comparison of methods for differential expression analysis of RNA-seq data. BMC Bioinformatics. 2013;14.

76. Garcia E. Singular value decomposition (svd) a fast track tutorial. Dr E Garcia. 2006;

77. Mubeen S, Tom Kodamullil AT, Hofmann-Apitius M, Domingo-Fernández D. On the influence of several factors on pathway enrichment analysis. Vol. 23, Briefings in Bioinformatics. 2022.

78. Merico D, Isserlin R, Stueker O, Emili A, Bader GD. Enrichment map: A network-based method for gene-set enrichment visualization and interpretation. PLoS One. 2010;5(11).

79. Reimand J, Isserlin R, Voisin V, Kucera M, Tannus-Lopes C, Rostamianfar A, et al. Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap. Nat Protoc. 2019;14(2).

80. Yu G, Wang LG, Han Y, He QY. ClusterProfiler: An R package for comparing biological themes among gene clusters. OMICS. 2012;16(5).

81. Szklarczyk D, Morris JH, Cook H, Kuhn M, Wyder S, Simonovic M, et al. The STRING database in 2017: Quality-controlled protein-protein association networks, made broadly accessible. Nucleic Acids Res. 2017;45(D1).

82. Jiao X, Sherman BT, Huang DW, Stephens R, Baseler MW, Lane HC, et al. DAVID-WS: A stateful web service to facilitate gene/protein list analysis. Bioinformatics. 2012;28(13).

83. Zhao K, Rhee SY. Interpreting omics data with pathway enrichment analysis. Vol. 39, Trends in Genetics. 2023.

84. Kelter R. Simulation data for the analysis of Bayesian posterior significance and effect size indices for the two-sample t-test to support reproducible medical research. BMC Res Notes. 2020;13(1).

85. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43(7).

86. Casares D, Escribá P V., Rosselló CA. Membrane lipid composition: Effect on membrane and organelle structure, function and compartmentalization and therapeutic avenues. Int J Mol Sci. 2019;20(9).

87. Pan M, Qin C, Han X. Lipid Metabolism and Lipidomics Applications in Cancer Research. In: Advances in Experimental Medicine and Biology. 2021.

88. Bougarne N, Weyers B, Desmet SJ, Deckers J, Ray DW, Staels B, et al. Molecular actions of PPARα in lipid metabolism and inflammation. Vol. 39, Endocrine Reviews. 2018.

89. Orozco Morales ML, Rinaldi CA, de Jong E, Lansley SM, Gummer JPA, Olasz B, et al. PPARα and PPARγ activation is associated with pleural mesothelioma invasion but therapeutic inhibition is ineffective. iScience. 2022;25(1).

90. Chen CC, Juan CW, Chen KY, Chang YC, Lee JC, Chang MC. Upregulation of RPA2 promotes NF-κB activation in breast cancer by relieving the antagonistic function of menin on NF-κB-regulated transcription. Carcinogenesis. 2017;38(2).

91. Xu S, Powers MA. Nuclear pore proteins and cancer. Vol. 20, Seminars in Cell and Developmental Biology. 2009.

92. Yu W, Jin H, Deng J, Nan D, Huang Y. A novel SPAST gene mutation identified in a Chinese family with hereditary spastic paraplegia. BMC Med Genet. 2020;21(1).

93. Verriello L, Lonigro IR, Pessa ME, Betto E, Pauletto G, Fogolari F, et al. Amplifying the spectrum of SPAST gene mutations. Acta Biomedica. 2021;92.

94. Frankland-Searby S, Bhaumik SR. The 26S proteasome complex: An attractive target for cancer therapy. Vol. 1825, Biochimica et Biophysica Acta - Reviews on Cancer. 2012.

95. TUBA1B Gene. In: Definitions. 2020.

96. Xu QQ, Qin LT, Liang SW, Chen P, Gu JH, Huang ZG, et al. The Expression and Potential Role of Tubulin Alpha 1b in Wilms' Tumor. Biomed Res Int. 2020;2020.

97. Røe OD, Anderssen E, Helge E, Pettersen CH, Olsen KS, Sandeck H, et al. Genome-wide profile of pleural mesothelioma versus parietal and visceral pleura: The emerging gene portrait of the mesothelioma phenotype. PLoS One. 2009;4(8).

98. Miao J, Kyoyama H, Liu L, Chan G, Wang Y, Urisman A, et al. Inhibition of cyclin-dependent kinase 7 down-regulates yes-associated protein expression in mesothelioma cells. J Cell Mol Med. 2020;24(1).

99. Evrin C, Clarke P, Zech J, Lurz R, Sun J, Uhle S, et al. A double-hexameric MCM2-7 complex is loaded onto origin DNA during licensing of eukaryotic DNA replication. Proc Natl Acad Sci U S A. 2009;106(48).

100. Burnham DR, Kose HB, Hoyle RB, Yardimci H. The mechanism of DNA unwinding by the eukaryotic replicative helicase. Nat Commun. 2019;10(1).

101. Fisher RP. Secrets of a double agent: CDK7 in cell-cycle control and transcription. J Cell Sci. 2005;118(22).

102. Malumbres M, Barbacid M. Cell cycle, CDKs and cancer: A changing paradigm. Vol. 9, Nature Reviews Cancer. 2009.

103. Mochida S, Hunt T. Protein phosphatases and their regulation in the control of mitosis. Vol. 13, EMBO Reports. 2012.

104. Nilsson J. Protein phosphatases in the regulation of mitosis. Vol. 218, Journal of Cell Biology. 2019.

105. Wang SS, Esplin ED, Li JL, Huang L, Gazdar A, Minna J, et al. Alterations of the PPP2R1B gene in human lung and colon cancer. Science (1979). 1998;282(5387).

106. Esplin ED, Ramos P, Martinez B, Tomlinson GE, Mumby MC, Evans GA. The glycine 90 to aspartate alteration in the Aβ subunit of PP2A (PPP2R1B) associates with breast cancer and causes a deficit in protein function. Genes Chromosomes Cancer. 2006;45(2).

107. Sablina AA, Chen W, Arroyo JD, Corral L, Hector M, Bulmer SE, et al. The Tumor Suppressor PP2A Aβ Regulates the RalA GTPase. Cell. 2007;129(5).

108. Ruvolo PP. The broken "Off" switch in cancer signaling: PP2A as a regulator of tumorigenesis, drug resistance, and immune surveillance. Vol. 6, BBA Clinical. 2016.

109. Schönthal AH. Role of serine/threonine protein phosphatase 2A in cancer. Vol. 170, Cancer Letters. 2001.

110. The Global Cancer Observatory. Estimated number of new cases in 2020,wordwide, both sexes, all ages. Word Health Organization. 2020;291.

111. Mithoowani H, Febbraro M. Non-Small-Cell Lung Cancer in 2022: A Review for General Practitioners in Oncology. Vol. 29, Current Oncology. 2022.

112. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2018;68(6).

113. Jemal A, Thun MJ, Ries LAG, Howe HL, Weir HK, Center MM, et al. Annual report to the nation on the status of cancer, 1975-2005, featuring trends in lung cancer, tobacco use, and tobacco control. J Natl Cancer Inst. 2008;100(23).

114. Quail DF, Joyce JA. Microenvironmental regulation of tumor progression and metastasis. Vol. 19, Nature Medicine. 2013.

115. Hanahan D, Coussens LM. Accessories to the Crime: Functions of Cells Recruited to the Tumor Microenvironment. Vol. 21, Cancer Cell. 2012.

116. Farc O, Cristea V. An overview of the tumor microenvironment, from cells to complex networks (Review). Exp Ther Med. 2020;21(1).

117. Tan Z, Xue H, Sun Y, Zhang C, Song Y, Qi Y. The Role of Tumor Inflammatory Microenvironment in Lung Cancer. Vol. 12, Frontiers in Pharmacology. 2021.

118. Roszer T. Understanding the mysterious M2 macrophage through activation markers and effector mechanisms. Vol. 2015, Mediators of Inflammation. 2015.

119. Liang J, Bi G, Shan G, Jin X, Bian Y, Wang Q. Tumor-Associated Regulatory T Cells in Non-Small-Cell Lung Cancer: Current Advances and Future Perspectives. Vol. 2022, Journal of Immunology Research. 2022.

120. Sheida F, Razi S, Keshavarz-Fathi M, Rezaei N. The role of myeloid-derived suppressor cells in lung cancer and targeted immunotherapies. Vol. 22, Expert Review of Anticancer Therapy. 2022.

121. Xie Q, Ding J, Chen Y. Role of CD8+ T lymphocyte cells: Interplay with stromal cells in tumor microenvironment. Vol. 11, Acta Pharmaceutica Sinica B. 2021.

122. Yuan A, Hsiao YJ, Chen HY, Chen HW, Ho CC, Chen YY, et al. Opposite Effects of M1 and M2 Macrophage Subtypes on Lung Cancer Progression. Sci Rep. 2015;5.

123. Genova C, Dellepiane C, Carrega P, Sommariva S, Ferlazzo G, Pronzato P, et al. Therapeutic Implications of Tumor Microenvironment in Lung Cancer: Focus on Immune Checkpoint Blockade. Vol. 12, Frontiers in Immunology. 2022.

124. Wang F, Yang M, Luo W, Zhou Q. Characteristics of tumor microenvironment and novel immunotherapeutic strategies for non-small cell lung cancer. Vol. 2, Journal of the National Cancer Center. 2022.

125. Hassanein SS, Abdel-Mawgood AL, Ibrahim SA. EGFR-Dependent Extracellular Matrix Protein Interactions Might Light a Candle in Cell Behavior of Non-Small Cell Lung Cancer. Vol. 11, Frontiers in Oncology. 2021.

126. Zhang Y, Du W, Chen Z, Xiang C. Upregulation of PD-L1 by SPP1 mediates macrophage polarization and facilitates immune escape in lung adenocarcinoma. Exp Cell Res. 2017;359(2).

127. Li Y, Liu H, Zhao Y, Yue D, Chen C, Li C, et al. Tumor-associated macrophages (TAMs)-derived osteopontin (OPN) upregulates PD-L1 expression and predicts poor prognosis in non-small cell lung cancer (NSCLC). Thorac Cancer. 2021;12(20).

128. Matsubara E, Komohara Y, Esumi S, Shinchi Y, Ishizuka S, Mito R, et al. SPP1 Derived from Macrophages Is Associated with a Worse Clinical Course and Chemo-Resistance in Lung Adenocarcinoma. Cancers (Basel). 2022;14(18).

129. Si Z, Hu K. Identification of CXCL13 as an immune-related biomarker associated with tumorigenesis and prognosis in cutaneous melanoma patients. Medical Science Monitor. 2021;27.

130. Kazanietz MG, Durando M, Cooke M. CXCL13 and its receptor CXCR5 in cancer: Inflammation, immune response, and beyond. Vol. 10, Frontiers in Endocrinology. 2019.

131. Wang B, Wang M, Ao D, Wei X. CXCL13-CXCR5 axis: Regulation in inflammatory diseases and cancer. Vol. 1877, Biochimica et Biophysica Acta - Reviews on Cancer. 2022.

132. Jiang N, Dai Q, Su X, Fu J, Feng X, Peng J. Role of PI3K/AKT pathway in cancer: the framework of malignant behavior. Vol. 47, Molecular Biology Reports. 2020.

133. Wen X, Jiao L, Tan H. MAPK/ERK Pathway as a Central Regulator in Vertebrate Organ Regeneration. Vol. 23, International Journal of Molecular Sciences. 2022.

134. Singh R, Gupta P, Kloecker GH, Singh S, Lillard JW. Expression and clinical significance of CXCR5/CXCL13 in human non-small cell lung carcinoma. Int J Oncol. 2014;45(6).

135. Hsieh CH, Jian CZ, Lin LI, Low GS, Ou PY, Hsu C, et al. Potential Role of CXCL13/CXCR5 Signaling in Immune Checkpoint Inhibitor Treatment in Cancer. Cancers (Basel). 2022;14(2).

136. Becht E, Giraldo NA, Lacroix L, Buttard B, Elarouci N, Petitprez F, et al. Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. Genome Biol. 2016;17(1).

137. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, et al. Robust enumeration of cell subsets from tissue expression profiles. Nat Methods. 2015;12(5).

138. Özdemir BC, Pentcheva-Hoang T, Carstens JL, Zheng X, Wu CC, Simpson TR, et al. Depletion of carcinoma-associated fibroblasts and fibrosis induces immunosuppression and accelerates pancreas cancer with reduced survival. Cancer Cell. 2014;25(6).

139. Sahai E, Astsaturov I, Cukierman E, DeNardo DG, Egeblad M, Evans RM, et al. A framework for advancing our understanding of cancer-associated fibroblasts. Vol. 20, Nature Reviews Cancer. 2020.

140. Kalluri R. The biology and function of fibroblasts in cancer. Vol. 16, Nature Reviews Cancer. 2016.