CAPITAL UNIVERSITY OF SCIENCE AND TECHNOLOGY, ISLAMABAD

Prediction and Analysis of Genes Associated with Lung Cancer Using Bioinformatics Tools

by

Afia Anum

A thesis submitted in partial fulfillment for the degree of Master of Science

in the

Faculty of Health and Life Sciences
Department of Bioinformatics and Biosciences

2025

I would like to dedicate this thesis to Allah Almighty, my parents and teachers

CERTIFICATE OF APPROVAL

Prediction and Analysis of Genes Associated with Lung Cancer Using Bioinformatics Tools

by

Afia Anum

(MBS231005)

THESIS EXAMINING COMMITTEE

S. No.	Examiner	Name	Organization
(a)	External Examiner	Dr. Muhammad Muddassar	CUI, Islamabad
(b)	Internal Examiner	Dr. Sania Riaz	CUST, Islamabad
(c)	Supervisor	Dr. Shaukat Iqbal Malik	CUST, Islamabad

Dr. Shaukat Iqbal Malik
Thesis Supervisor
May, 2025

Dr. Syeda Marriam Bakhtiar
Head
Dept. of BI and BS
May, 2025

Dr. Sahar Fazal
Dean
Faculty of Health and Life Sciences
May, 2025

Author's Declaration

I, **Afia Anum** hereby state that my MS thesis titled "**Prediction and Analysis of Genes Associated with Lung Cancer Using Bioinformatics Tools**" is my own work and has not been submitted previously by me for taking any degree from Capital University of Science and Technology, Islamabad or anywhere else in the country/abroad.

At any time if my statement is found to be incorrect even after my graduation, the University has the right to withdraw my MS Degree.

(Afia Anum)

Registration No: MBS231005

Plagiarism Undertaking

I solemnly declare that research work presented in this thesis titled "**Prediction and Analysis of Genes Associated with Lung Cancer Using Bioinformatics Tools**" is solely my research work with no significant contribution from any other person. Small contribution/help wherever taken has been duly acknowledged and that complete thesis has been written by me.

I understand the zero tolerance policy of the HEC and Capital University of Science and Technology towards plagiarism. Therefore, I as an author of the above titled thesis declare that no portion of my thesis has been plagiarized and any material used as reference is properly referred/cited.

I undertake that if I am found guilty of any formal plagiarism in the above titled thesis even after award of MS Degree, the University reserves the right to withdraw/revoke my MS degree and that HEC and the University have the right to publish my name on the HEC/University website on which names of students are placed who submitted plagiarized work.

(Afia Anum)

Registration No: MBS231005

Acknowledgement

Alhamdulillah! I'm extremely thankful to Almighty Allah the most merciful and most magnificent. I offer my gratitude to the Holy Prophet (PBUH) who preached us to seek knowledge for the betterment of mankind. I would like to express my gratitude to my family including my father Ali Akbar, and my mother Farzana Jabeen. I am grateful to my Dean Prof Dr. Sahar Fazal for her continuous support during my work. I'm thankful to my supervisor Prof. Dr.Shokat Iqbal Malik. A lot of people assisted and helped in this work. I want to give special thanks to my teacher Dr Shumaila whose continuous support assisted me in completing my work. I would like to acknowledge all of them, and wish good luck ahead. My thanks and appreciation also go to my friend Arifa Saddiqa.

(Afia Anum)

Abstract

Lung cancer is one of the most aggressive malignancies worldwide, driven by complicated genetic and molecular alterations. A number of the key genetic members, tp53, RB1, and myc play pivotal roles in tumor development. Tp53, a crucial tumor suppressor, is regularly mutated in lung cancer, leading to impaired DNA repair and out-of-control cellular proliferation. Similarly, RB1 inactivation disrupts mobile cycle regulation, selling tumorigenesis, in particular in small-cellular lung cancers (SCLC)and non-small-cell lung cancer(NSCLC). In contrast, the MYC oncogene is often overexpressed, using most cancer's mobile growth, metabolic reprogramming, and resistance to apoptosis. The interplay among these genes highlights crucial pathways concerned with lung cancer initiation and development, making them ability objectives for therapeutic intervention. Advanced bioinformatics gear is increasingly more used to expect and examine gene interactions, offering novel insights for precision medicinal drug and centered therapy improvement in lung cancer remedy.

Contents

List of Figures

List of Tables

Abbreviations

BP	Biological process
CC	Cellular component
ceRNA	Competing endogenous RNA
DAVID	The Database for Annotation, Visualization and Integrated Discovery
DEGCs	Differentially expressed lung cancer genes
DEGs	Differentially expressed genes
GEO	Gene expression omnibus
KEGG	Kyoto encyclopedia of genes and genomes
LC	Lung cancer
MF	Molecular function
NSCLC	Non-Small Cell Lung Cancer
PPI	Protein-protein interaction
SCLC	Small Cell Lung Cancer
STRING	Search tool to retrieve interacting genes and proteins

Chapter 1

Introduction

Lung cancer is a leading cause of cancer-related deaths worldwide, arising when abnormal cells in the lungs grow uncontrollably. It is primarily linked to smoking, but environmental factors and genetic predisposition also play a role. It often remains undetected in its early stages, leading to late diagnoses and reduced survival rates. Depending on the type and stage, treatment options include surgery, chemotherapy, radiation, targeted therapies, and immunotherapy [1].

Non-Small Cell Lung Cancer (NSCLC) accounts for nearly 85% of all lung cancer cases and encompasses several subtypes, including adenocarcinoma, squamous cell carcinoma, and large cell carcinoma. It generally progresses more slowly than small cell lung cancer but is often diagnosed at an advanced stage. Treatment strategies depend on factors such as tumor size, spread, and genetic mutations, with surgery, chemotherapy, radiation, targeted drugs, and immunotherapy being common approaches [1, 2].

Lung cancer is the most frequently diagnosed type of cancer and the primary cause of cancer-related deaths globally, representing 11.6% of all cases and 18.4% of total cancer fatalities. Lung cancer involves the uncontrolled proliferation of abnormal cells, typically originating in one or both lungs. It most commonly affects the cells lining the airways. There are two primary types of lung cancer: small-cell lung cancer (SCLC) and non-small-cell lung cancer (NSCLC). Secondary lung cancer, in contrast, originates in other parts of the body and metastasizes to the lungs.[4]

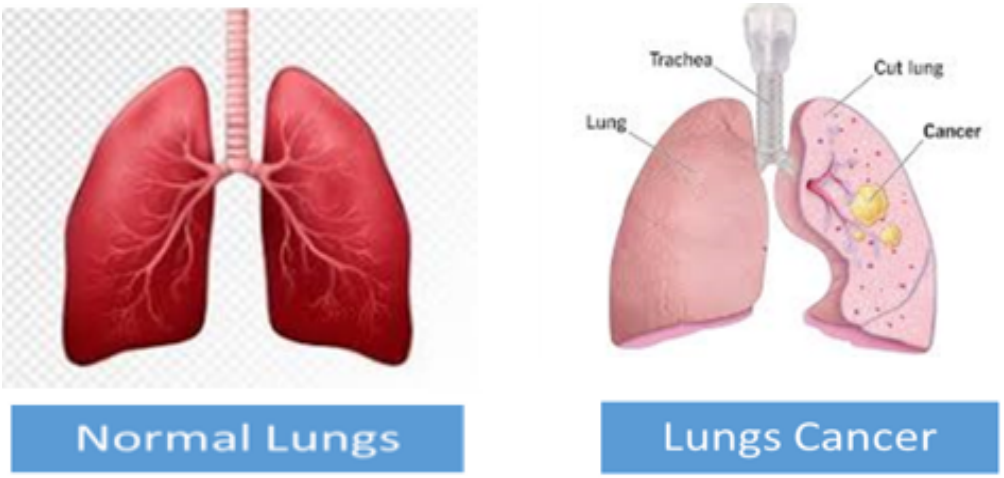


FIGURE 1.1: Description of healthy and Cancerous lung.

Common symptoms of lung cancer include a persistent cough that worsens over time, coughing up blood or rust-colored sputum, and chest pain that becomes more noticeable with deep breathing, coughing, or laughter. Other signs may involve voice changes like hoarseness, reduced appetite, unexplained weight loss, shortness of breath, and persistent fatigue or weakness. Recurrent infections such as bronchitis and pneumonia that fail to resolve or frequently return can also indicate lung cancer. If the cancer spreads to the brain, neurological symptoms may arise, including headaches, limb weakness or numbness, dizziness, balance issues, or seizures. Additionally, swollen lymph nodes, particularly in the neck or above the collarbone, can be a sign. Some types of lung cancer may also cause syndromes, which present as distinct clusters of symptoms [3, 4].

Sputum analysis is used to check for the presence of cancer cells in the mucus that is coughed up. In a laboratory, an aberrant lung tissue sample is taken and examined under a microscope. The appearance of the cells under a microscope can be used to determine the type of cancer if it is discovered in the tissue. Although surgery may be necessary to access the area of concern, a biopsy method is typically used to harvest this tissue during a bronchoscopy.

Sputum analysis is used to check for the presence of cancer cells in the mucus that is coughed up. In a laboratory, an aberrant lung tissue sample is taken and examined under a microscope. The appearance of the cells under a microscope can be used to determine the type of cancer if it is discovered in the tissue. Although surgery

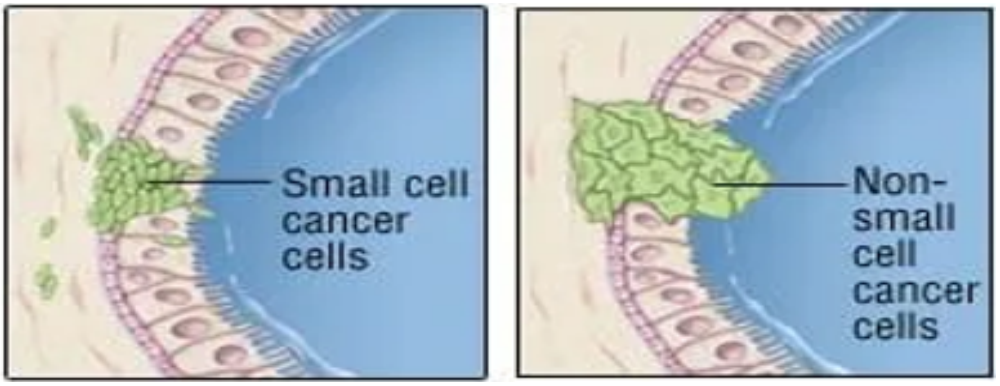


FIGURE 1.2: Display of SCLC and NSCLC.

may be necessary to access the area of concern, a biopsy method is typically used to harvest this tissue during a bronchoscopy [5, 6]

The fine-needle aspiration procedure involves using a CT scan to identify a questionable location, then inserting a thin needle into that area of the lung or pleura. A tiny sample of tissue is taken by the needle for analysis in a lab, which can identify the type of malignancy. procedure for thoracentesis [7]. A sterile needle can be used to drain any fluid that has accumulated in the chest. After that, the fluid is examined for cancerous cells. VATS stands for video-assisted thoracoscopic surgery.Through a tiny incision, a surgeon performs this surgery by inserting a flexible tube equipped with a video camera into the chest. This enables the surgeon to check for malignancy around the edge of the lung and between the lung and the chest wall. It is also possible to harvest abnormal lung tissue [8–10]

Bone scans and CT imaging:These diagnostic tests can reveal lung cancer that has metastasized to the bones, brain, or other regions of the body. Once cancer is diagnosed, it is categorized into a stage. The staging varies for non-small cell lung cancer and small cell lung cancer [11]

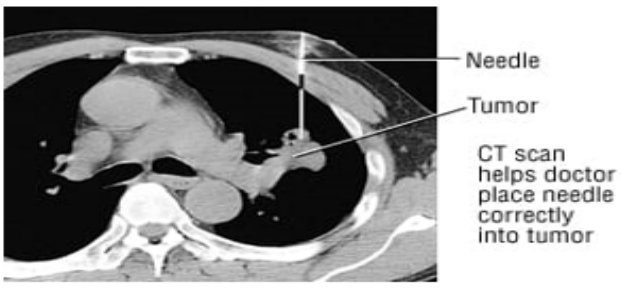


FIGURE 1.3: CT Scan of Lung Cancer.

1.1 Treatment of Lung Cancer

After lung cancer has been diagnosed, the type of treatment depends on the type of cancer and how much the tumor has spread The type of treatment for lung cancer is determined by the cancer's type and stage, or how far the tumor has spread.(its stage) [12].

1.1.1 Non-small Cell Lung Cancer

For non-small cell lung tumors that have not moved beyond of the chest, surgery is the primary treatment. The extent of the malignancy will determine the type of surgery. Additionally, it will depend on the presence of other lung disorders like emphysema.Three categories of surgery exist [13]. Only a small portion of the lung is removed during wedge resection. One lung lobe is removed during a lobectomy. An complete lung is removed during a pneumonectomy. Lymph nodes are also removed and examined to see if the cancer has spread.For non-small cell lung tumors that have not moved beyond of the chest, surgery is the primary treatment. The extent of the malignancy will determine the type of surgery. Additionally, it will depend on the presence of other lung disorders like emphysema.Three categories of surgery exist. Only a small portion of the lung is removed during wedge resection. One lung lobe is removed during a lobectomy. An complete lung is removed during a pneumonectomy. In order to determine whether the cancer has spread, lymph nodes are also removed and analyzed [14, 15]

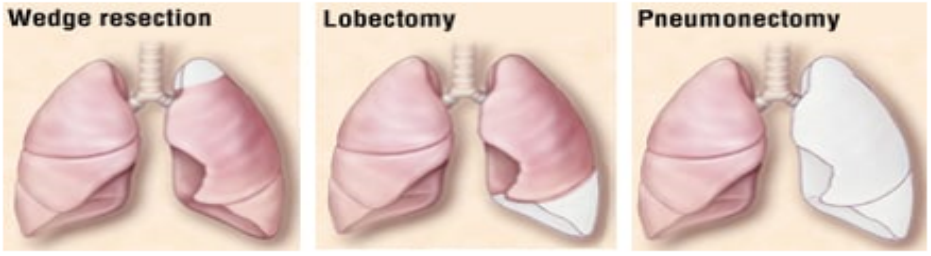


FIGURE 1.4: Lung Cancer treatment options.

The figures presented illustrate the variations in the distribution of radiotherapy doses to the lungs when employing IMRT or VMAT techniques. These visuals

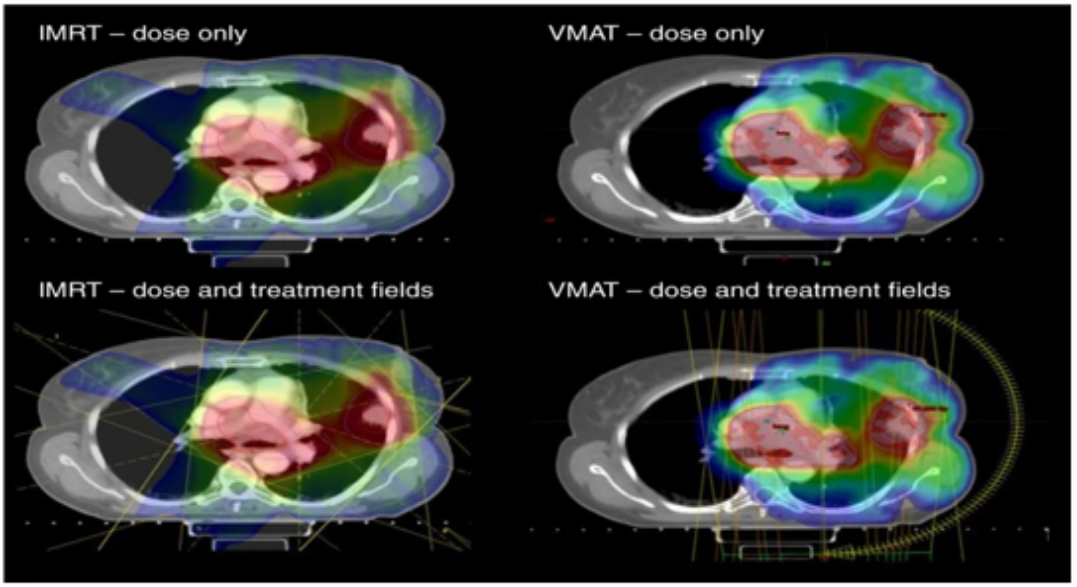


FIGURE 1.5: Intensity-modulated radiotherapy (IMRT) and volumetric arc radiotherapy (VMAT) plans are shown for the same patient with stage III non-small-cell lung cancer (NSCLC).

indicate that the application of advanced IMRT might reduce the V20 dosimetric value for the non-target lung, but it may also lead to a higher percentage of lungs receiving lower radiotherapy doses compared to when VMAT is used [16].

1.1.2 Small Cell Lung Cancer

Small cell lung cancer (SCLC) is a fast-growing and aggressive type of lung cancer. It is strongly associated with smoking and tends to spread quickly to other organs. Treatment usually includes chemotherapy and radiation therapy. Although it initially responds well to treatment, recurrence is common. Due to its rapid progression, early detection is crucial for better outcomes. [17–19].

1.2 Epidemiology

Except for skin cancer, lung cancer, including both small cell and non-small cell types, is the second most prevalent cancers in the United States for both genders. Breast cancer is more frequent among women, while prostate cancer is more common in men.

On a global scale, lung cancer remains the leading cause of death related to cancer. The American Cancer Society anticipates that there will be over 125,070 deaths associated with lung cancer in the US in 2024 (with 65,790 for men and 59,280 for women) and approximately 234,580 new cases of lung cancer (116,310 for men and 118,270 for women) [20, 21]

Lung cancer primarily impacts older individuals. Very few cases are diagnosed in individuals under the age of 45, with most diagnoses occurring in those 65 and older. The average age at diagnosis is around 70 years. Lung cancer condoned for about one in five cancer-related deaths in the US, making it the main cause of mortality linked to cancer. The number of deaths caused by lung cancer each year surpasses the combined fatalities from colon, breast, and prostate cancers. Fortunately, an increasing number of people are either quitting smoking or never starting, which contributes to the ongoing decrease in new lung cancer cases [22].

There is decline in smoking rates and advancements in early detection and treatment, of No. deaths from lung cancer is also decreasing. The lifetime risk of developing lung cancer during one's life is significant. For men, the lifetime risk is about 1 in 16, while for women, it's approximately 1 in 17. This data includes both smokers and non-smokers. Although non-smokers face a lower risk, it is markedly higher for smokers. Compared to white men, black men experience a 12% greater risk of lung cancer, whereas black women have a rate that is roughly 16% lower than that of white women. Lung cancer rates are lower among black and white women compared to men, the gap is closing. Over recent decades, particularly in the last ten years, the cases of lung cancer in men has been on the decline [23].

1.3 Etiology

Tobacco has been an integral part of this nation's cultural and economic landscape since Columbus's time. Initially consumed by chewing or smoking in pipes, tobacco became largely accessible in cigarette form after the invention of machinery for wrapping cigarettes in the mid-1800s. Before World War I, cigarette consumption

in the United States was relatively low. Wynder and Graham estimated that in 1900, the average adult smoked fewer than 100 cigarettes annually. By fifty years later, this figure increased to about 3,500 cigarettes per person each year, peaking at roughly 4,400 cigarettes per person per year in the mid-1960s. In 1964, the US Public Health Service released a significant report by the Surgeon General addressing smoking and its health implications. This pivotal report outlined several key findings [24]. Cigarette smoking was linked to a 70% rise in age-specific mortality rates for men, with a smaller increase for women [25].

The rate of never-smoking lung cancer patients was more than double in women compared to men. The reasons behind this observation remain unclear, but there has been speculation about women's potentially higher susceptibility to non-tobacco environmental carcinogens, increased exposure to environmental tobacco smoke (ETS), or differences related to gender in the metabolic action of non-tobacco environmental carcinogens [26, 27]

1.4 Gene Structure: RB1

Sequencing the **RB1** gene poses unique challenges due to its considerable size (**$\tilde{1}$78 kb**) and an unusually high number of introns (**26**), which make up approximately **99%** of its total length. Genomic alterations in **RB1** are often found in intronic splice-site mutations and structural variants, many of which may go undetected by targeted sequencing techniques commonly used in clinical practice, as these methods primarily focus on exonic regions [28].

FIGURE 1.6: Gene Structure

Immunohistochemistry (IHC) is a well-established approach in pathology for evaluating **RB1** expression. However, a large-scale comparison between targeted sequencing and IHC for **RB1** assessment in **SCLC** has yet to be conducted [29].

1.5 Gene Function: RB1

The sequencing of the RB1 gene poses unique challenges due to its considerable size (approximately 178 kb) and the higher-than-average number of introns it contains (26 total), with intronic regions making up about 99% of the overall gene length [30].

Changes in the RB1 gene are particularly common in splice-site mutations located within introns and in structural variants (4, 16); these alterations may escape detection by conventional targeted sequencing methods used in clinical settings, which mainly focus on exonic areas. Immunohistochemistry (IHC) techniques for evaluating Rb expression are well-established in clinical pathology. However, large-scale comparisons of targeted sequencing against IHC for assessing Rb in small cell lung cancer (SCLC) have yet to be carried out [31].

Given the potential intricacies involved in determining Rb status, this study aimed to investigate not only the genomic alterations in RB1 and their expression but also the functional status of Rb by examining key regulators of the G1–S cell cycle checkpoint within the Rb pathway. Research on both cells and clinical samples that encompass a wide range of tumor types has identified a closely linked pattern of expression between Rb and its upstream regulators—cyclin-dependent kinase (CDK) inhibitor p16INK4A (p16), which is produced by the CDKN2A gene, and cyclin D1, produced by the CCND1 gene [32].

1.6 Causes

Causes and risk factors of lung cancer can be noted in Fig below.

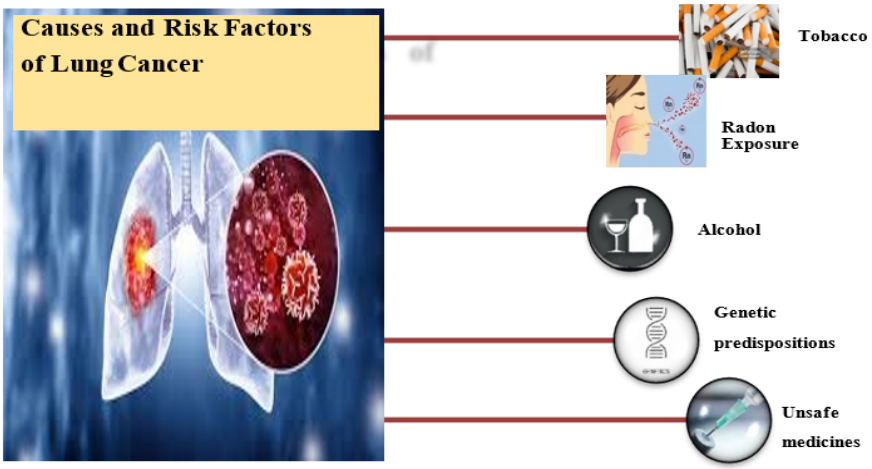


FIGURE 1.7: Causes and risk factors of Lung Cancer

1.7 Mutations

Numerous studies have focused on candidate vulnerability genes characterized by low penetrance and high frequency.

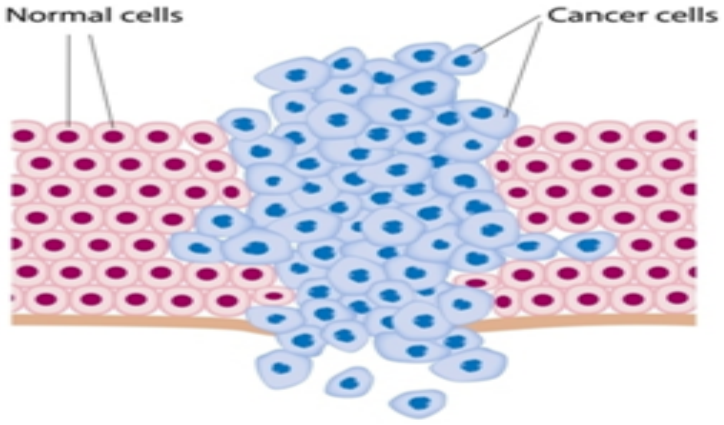


FIGURE 1.8: Mutations in Cell.

1.7.1 Point Mutations

1.7.1.1 Missense Mutations

Mutations known as missense mutations are characterized by the presence of a single nucleotide change that transforms the sequence of amino acids that are present in the resulting protein. The lung cancer protein may be dysfunctional as a consequence of a missense mutation that is associated with the RB1 gene [33].

1.7.1.2 Nonsense Mutations

In this particular instance, a point mutation causes the introduction of an early stop codon, which in turn truncates the synthesis of the protein. In most cases, the protein that is produced is either completely dysfunctional or severely impaired [34].

1.7.1.3 Insertions and Deletions (INDELS)

These mutations involve the insertion or deletion of nucleotides, which results in a frameshift, which is the process of reading frames during reading frame during translation. As a consequence of this mutation, the protein that is produced is frequently either shortened or rendered,irrelevant.Splice acceptor or donor site mutations: Changes in the regions that determine the splicing of introns during the processing of mRNA can result in aberrant splicing patterns. These mutations can occur at either the splice acceptor or the donor site [35]. As a result, this could result in the inclusion of incorrect exons or the exclusion of essential exons, which would ultimately have an impact on the operational capabilities of the protein. In this particular instance, a particular kind of mutation known as a point mutation causes the insertion of an early stop codon, which ultimately results in the production of the protein being shortened. In the majority of cases, the protein that is produced is either completely devoid of any functional properties or significantly impaired [36, 37].

1.7.2 Large Deletions or Duplications

1.7.2.1 Deletions from the Chromosomes

RB1 gene deletions on a large scale have the potential to cause the gene's function to be reduced or eliminated entirely. It is possible for this kind of mutation to have an effect on multiple exons and to cause structural disruptions in the lung cancer protein [38].

1.7.2.2 Chromosome Duplications

It is possible for there to be instances in which chromosomal segments, including the RB1gene, are duplicated simultaneously. Consequently, this could result in an over expression of the lung cancer protein, which could potentially have an effect on the regulation of the cell cycle [39].

1.8 Problem Statement

Lung Cancer is one of the leading causes of cancer-related deaths Worldwide. Lung cancer is often diagnosed with advanced stages then the opportunity of treatment is limited. Identification of genes associated with lung cancer is essential for understanding the development and cure of disease. The main goal is to diagnose early-stage LC cancer accurately and timely by using advanced bioinformatics tools because of late diagnosis, Only a minority of patients can have curative treatment.

1.9 Aim and Objectives

This research aims to find and analyze the genes associated with Lung Cancer.

Objectives

1. To use text mining to forecast genes linked to Lung Cancer

2. To carry out functional annotation of putative genes linked to Lung Cancer

3. To investigate the role of anticipated genes in pathways relevant to Lung Cancer

The purpose of this study is to identify the significant genes involved in LC through the Bioinformatics Method and to reveal the Potential underlying mechanisms for the detection and then cure of lung Cancer

Chapter 2

Literature Review

Lung cancer arises when lung cells start dividing excessively, leading to the formation of tumors. While anyone can be diagnosed with lung cancer, the risk is heightened by cigarette smoking and exposure to smoke, harmful chemicals, or other toxic substances. Although lung cancer can be deadly, advancements in diagnoses and treatments such as chemotherapy and surgery are enhancing prognosis [40].

2.1 Epidemiology

According to the latest GLOBOCAN data, 2,094,000 new lung cancer cases were reported worldwide in 2018, making it the most frequently diagnosed cancer. It ranks as the second most common cancer in men (1,369,000 cases) after prostate cancer and in women (725,000 cases) after breast cancer. The age-standardized lifetime risk of developing lung cancer is 3.8% for men and 1.77% for women [41, 42].

In Pakistan, in 2020, lung cancer was responsible for approximately 9.6% of all cancer-related deaths, with an age-standardized mortality rate of 6.2 per 100,000 people across both genders. A nationwide survey in Pakistan from 1998 revealed that 21.6% of individuals aged 15 and above were smokers, with a higher prevalence among men (36%) compared to women (9%). [43].

2.2 Diagnosis

In accordance with the guidelines, treatment and care should be provided to young children and families who have been affected by lung cancer [44]. It is essential to ascertain the extent of the tumor or the degree of its severity before beginning the process of formulating a treatment plan. Following the determination of the size of the tumor and the likelihood that it will spread, a particular stage of cancer is assigned to the patient. In the diagnosis of lung cancer, MRI and tomography are the method of choice.

To elucidate the genetic risk that is associated with a pathogenic variant of germline RB1, it was suggested that the following staging be utilized [45].

- **TP53:** This gene is involved in the production of tumor protein 40 to 50 % of all cases of NSCLC, Acquired in smokers and Non-Smokers
- **KRAS:** 30% of all lung cancer, More common in the smoker
- **EGFR**: Epidermal Growth Factor is a protein on the surface of the cell involved in growth and division. NSCLS have too much of this protein. Around 10 to 15% of all lung cancer
- **ALK**: Anaplastic Lymphoma Kinase Kinase About 5% (More common in Young and Nonsmokers) [46]
- **MET gene**: Mesenchymal Epithelial Transition is changed up to 5%
- **PIK3CA** (Phosphatidylinositol -4,5 Bisphosphate 3-Kinase Catalytic subunit: presence is 6%, More common in Squamous Cell Carcinoma
- **BRAF & BRAF V600E**:Upto the 3.5 to 4% of NSCLS test positive BRAF mutation ,More common in females than Males [47]
- **HER2:** Between 1% to 4% of NSCLC involve human Epidermal growth receptor genes.It more common in Adenocarcinoma

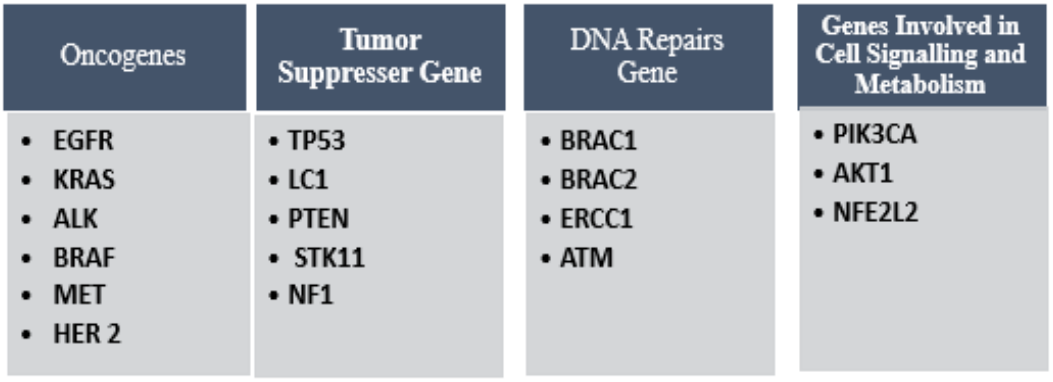


FIGURE 2.1: Associated Genes involved in Lung Cancer.

2.3 Signs and Symptoms

These include a persistent cough that worsens over time, coughing up blood or rust-colored sputum, and chest pain that becomes more intense with deep breathing, coughing, or laughing. Voice changes, such as hoarseness, along with a loss of appetite and unexplained weight loss, may also occur. Many patients experience shortness of breath, fatigue, or weakness. Recurrent infections like bronchitis or pneumonia that fail to resolve or frequently return can also be a warning sign. If the cancer spreads to the brain, neurological symptoms such as headaches, dizziness, balance issues, numbness or weakness in the limbs, and seizures may develop. Additionally, swelling of lymph nodes, particularly in the neck or above the collarbone, can indicate lung cancer [48, 49].

2.4 Nature of Disease

Lung cancer predominantly begins in the lung tissues, yet it can have effects that reach beyond this initial location. As the illness advances, cancerous cells may infiltrate nearby lung regions, resulting in disrupted respiratory function. For example, tumors could block air passages, leading to symptoms like chronic coughing, difficulty breathing, and chest discomfort. Moreover, lung cancer may lead to the formation of fluid in the chest a referred to as pleural effusion, which further hinders breathing.

Additionally, lung cancer can metastasise to various body parts, such as lymph nodes, the bones, the brain, the liver, and the adrenal glands. The degree of lung involvement and the presence of metastases significantly affect treatment choices and prognosis. Identifying the disease early is essential, as cancers that are localized and limited to a small portion of the lung are more likely to be surgically removed and generally have a more favorable outlook compared to later stages where cancer has widely spread. Grasping the patterns of lung cancer progression and dissemination is vital for formulating effective treatment plans and enhancing patient outcomes [49].

2.4.1 Types of Lung Cancer

Hereditary: The inheritance of hereditary characteristics can occur within families.

Sporadic: Occurring in a sporadic manner and not being passed down through extended families.

2.4.1.1 Non-small Cell Lung Cancer (NSCLC)

The initial assessment of the extent of the disease is crucial for patients newly diagnosed with NSCLC, as it helps in choosing the most suitable treatment options and provides prognostic insights. It is particularly important to accurately distinguish patients with potentially curable (early-stage) disease, who might gain from radical surgery, from those deemed unresectable, who will then be directed towards chemotherapy, radiotherapy, or a combination of both.

1. Adenocarcinoma: A type of NSCLC frequently located in the lung's outer regions. It originates in the cells of epithelial tissues, which line body cavities and surfaces while forming glands.

2. Squamous cell carcinoma: This is a form of non-small cell lung cancer (NSCLC) usually located in the central area of the lung, near a bronchus.

3. Large cell carcinoma: This type of NSCLC can arise in any area of the lung and typically has a faster growth rate and spreads more quickly than both adenocarcinoma and squamous cell carcinoma.

2.4.1.2 Small Cell Lung Cancer

Small cell lung carcinoma (SCLC) constitutes about 10–15% of all lung cancer cases. Clinically, SCLC is more aggressive compared to non-small cell lung cancer (NSCLC), showing a rapid doubling time and a higher likelihood of widespread metastatic spread. The overall outlook is grim, as although initial responses to chemotherapy can be favorable, the majority of SCLC patients experience relapse and succumb to recurrent illness. SCLC has traditionally been categorized into two stages: limited disease (LD) or extensive disease.

2.4.2 Groups of Classification

Patients who have been diagnosed with lung cancer are classified by the which can be found below in table 2.1 [50].

TABLE 2.1: Classification of people diagnosed with lung cancer.

Groups	Descriptor
Stage I	The tumor is confined to the lung and is relatively small (4 centimeters or less). It has not metastasized to adjacent lymph nodes or beyond the chest area. Typically, surgery is the primary treatment for most stage 1 tumors. For patients who are not candidates for surgery, we provide radiation therapy as an alternative. Most patients with stage 1 tumors do not require chemotherapy, targeted therapy, or immunotherapy.
Stage II	There are bigger tumors of more than 4 centimeters or signs that cancer has spread to adjacent lymph nodes but not outside the

Continue on next page

Groups	Descriptor
	lung. We typically treat stage 2 lung cancer with surgery followed by chemotherapy and targeted or immunotherapy. For patients who cannot undergo surgery, chemotherapy and radiation are offered instead.
Stage III	In stage 3 lung cancer, there is cancer in the lymph nodes of the chest away from the lung, or there may be large tumors that have spread to adjacent lymph nodes. Most individuals with stage 3 cancer receive multiple treatments, including various combinations of chemotherapy, surgery, radiation, targeted therapies, and immunotherapy.
Stage IV	The cancer has spread to distant organs. The most common metastasis sites include the other lung, bones, brain, and adrenal glands. Treatment depends on the tumor type and may include chemotherapy, targeted therapy, immunotherapy, or a combination of these approaches.

2.5 Genetics

2.5.1 MYCN

MYCN is a gene that belongs to the MYC family of proto-oncogenes. The transcription factor known as MYCN plays a function in the regulation of significant processes that occur during the embryonic development process. When it comes to signaling pathways, the MYCN protein is located downstream of many of them. Pathways are responsible for facilitating the growth, proliferation, and metabolism of progenitor cells in various organs and tissues that are in the process of developing. Unregulated MYCN signaling, on the other hand, is responsible for the proliferation of certain types of cancers, particularly those that occur in early childhood. These cancers include neuroblastoma and lung cancer.

2.5.2 RB1

The gene known as RB1 gene is the first gene to be identified as a tumor suppressor. The size of the gene is quite substantial since it is 190 kilobases in length and is composed of 27 exons. This gene is accountable for the encoding of a messenger RNA (mRNA) molecule that is 4.7 kilobases in length. This mRNA molecule, in turn, plays a role in the translation of proteins that are composed of 928 amino acids [32]. As a tumor suppressor, the LC1 gene, which may be found on chromosome 13q14, is responsible for its function. This particular gene is accountable for the production of a nuclear phosphoprotein that is referred to as LC. During the course of the cell cycle, the process that is known as checkpoints is responsible for regulating the transition of cells from the G-phase to the S-phase [51].

2.5.3 Mutation in RB1 and Amplification of MYCN

Alterations such as point mutations, promoter methylation, and indel mutations are examples of the kinds of changes that can have an effect on the functioning of LC. The majority of the time, LC is classified as a controller of the cell cycle, specifically in terms of its ability to regulate the process of cell division. After some time has passed, it makes a connection with E2F transcription factors, which inhibit the production of genes that are important in the proliferation of cells. Cyclin-dependent kinases (CDKs) are responsible for the hyperphosphorylation of LC, which occurs when mitogenic signals are received. Repression is frequently alleviated as a result of this process, which also makes the transition from the G1 phase to the S phase easier. As a result of the lack of LC, the cell cycle is able to continue even when mitogenic stimuli are not present, which prevents the crackdown from occurring. It is possible to postulate that LC plays a significant part in suppressing E2F transcription factors and that the failure of this function is the fundamental mechanism that leads to the development of lung cancer [43]. In addition to the amplification of MYCN oncogenes, there are a number of other causes that have been discovered for the non-inactivation of LC1. Recently, it has been discovered that patients who have been diagnosed with LC have a lesion [52].

There have been two separate instances in which mutations have been found in exon 20 of the RB1 gene. When a C-T transformation takes place at the 661st codon, which results in the conversion of an arginine (CGG) to tryptophan (TGG), this is an example of a specific point mutation that can be noticed [53]. Almost all patients are observed to have limited penetration and a mild phenotypic manifestation, which is a common observation. A proposition that suggests that changes in the shape and function of proteins can be brought about by alterations in individual amino acids. Codon 675 has been shown to contain a mutation in another instance of the disease. The glutamine (GAA) is converted into a stop codon (TAA) as a result of the transition that takes place through the G-T.

2.6 Protein Association with Lung Cancer

Several proteins are associated with lung cancer, playing crucial roles in its development, progression, and treatment response. **TP53**, a tumor suppressor protein, is frequently mutated in lung cancer, leading to uncontrolled cell growth. **EGFR (Epidermal Growth Factor Receptor)** mutations and overexpression are common in non-small cell lung cancer (NSCLC), making it a key target for targeted therapies like tyrosine kinase inhibitors (TKIs).

KRAS, another oncogenic protein, is involved in cell signaling pathways that drive lung cancer proliferation. Additionally, proteins like **ALK (Anaplastic Lymphoma Kinase and ROS1** are implicated in specific lung cancer subtypes and are targeted by drugs such as crizotinib.

Other significant proteins include **BCL2**, which regulates apoptosis, **PI3K/AKT/mTOR** pathway proteins that influence cell survival and metabolism, and **VEGF (Vascular Endothelial Growth Factor)**, which promotes angiogenesis in tumors.

Understanding the molecular roles of these proteins has led to advancements in personalized medicine and targeted lung cancer therapies. Table 2.2 shows key interacting proteins associated with lung cancer.

TABLE 2.2: Interacting Proteins for LC.

Protein	**Function**	**Interaction in Lung Cancer**
TP53	Tumor suppressor	Frequently mutated, leading to loss of cell cycle control
EGFR	Growth factor receptor	Overexpressed/mutated in NSCLC, drives tumor growth
KRAS	Oncogene	Activates signaling pathways (MAPK, PI3K) for tumor progression
ALK	Tyrosine kinase receptor	Fusion proteins drive uncontrolled proliferation
ROS1	Tyrosine kinase receptor	Rearrangements activate oncogenic signaling
BCL2	Apoptosis regulator	Inhibits cell death, promoting cancer cell survival
PIK3CA	PI3K pathway regulator	Enhances cell survival and proliferation
AKT1	Kinase in PI3K/AKT pathway	Promotes tumor growth and resistance to therapy
VEGF	Angiogenesis regulator	Stimulates blood vessel formation for tumor growth
STAT3	Transcription factor	Regulates immune evasion and inflammation in cancer
NF-κB	Transcription factor	Promotes inflammation and cancer cell survival
CDKN2A	Cell cycle regulator	Frequently inactivated, leading to unchecked proliferation
MYC	Oncogene	Drives cell growth and metabolic reprogramming
HIF-1α	Hypoxia response regulator	Supports tumor adaptation in low-oxygen conditions

2.7 Genetic Counselling

Genetic counseling is a process that assists families in comprehending the characteristics of genetic anomalies, including their inheritance status and the implications of these abnormalities. It offers information that can be used to assist in making judgments on personal choices and individual medical treatments.

The autosomal dominant inheritance pattern is the one that is responsible for heritable lung cancer. The dangers that members of the family face are as follows:

2.7.1 Parents of Affected Individual

- Lung cancer is primarily associated with environmental risk factors, particularly smoking. However, genetic predispositions also play a significant role in its development. Studies have shown that 8–15% of individuals diagnosed with lung cancer carry hereditary mutations that increase their risk.

- These inherited genetic factors can contribute to lung cancer susceptibility, even in non-smokers.

- For parents of individuals affected by lung cancer, it's important to understand that while the majority of lung cancer cases are linked to environmental exposures, there is a hereditary component that may influence risk. Family members, including parents, may have an elevated risk due to shared genetic factors. Genetic counseling and testing can provide valuable insights into individual and familial risk profiles, guiding decisions on screening and preventive measures.

- In summary, while environmental factors remain the predominant cause of lung cancer, hereditary elements can also contribute to its occurrence. Parents of affected individuals should consider genetic evaluation to better understand and manage potential risks within the family [54].

2.7.2 Risk to Siblings of Affected Individual

Both the genetic and phenotypic characteristics of the probe and the parents will determine the level of risk that is posed to the siblings.

2.7.2.1 Parents Phenotype

- The risk of developing lung cancer is fifty percent for siblings of a proband who have bilateral lung cancer along with their parents for developing the disease themselves.
- The risk to siblings is lowered to between one and two percent in the event of low penetrance, which occurs when parents frequently do not exhibit clinical symptoms but possibility of lung cancer is still there.

2.7.2.2 Parents Genetic Status

- In the event that both parents have heterozygous pathogenic variants of RB1, then both the probe and the parent are affected, and there is a fifty percent chance that the risk of cancer will be passed on to their siblings.
- This means that the possibility of the sibling being harmed is less than five percent in the case that the penetrance is diminished. Even so, it is still recommended to conduct a test on the sibling to determine whether or not the probe has a de novo inherited mutation. In the event that this particular mutation is not discovered in the siblings of this individual, such siblings are regarded as being unaffected [55–57].
- When a somatic LC1 mutation is present in the probe, the likelihood that a sibling may be affected by the mutation is greatly reduced.

2.7.2.3 Risk to Proband Offspring

The risk to a proband's offspring regarding lung cancer, based on genetic factors,

depends on several variables. If a genetic predisposition to lung cancer exists within the family, particularly due to mutations in certain genes (like EGFR, KRAS, TP53, BRCA2, or STK11), offspring may inherit those genetic variants, potentially increasing their risk. Lung cancer is influenced not only by genetics but also by environmental factors, such as smoking, exposure to second-hand smoke, and environmental pollutants. The inheritance pattern of lung cancer susceptibility can vary based on the specific gene involved, and many cases of lung cancer are complex, involving both genetic and environmental factors. For instance:

- If a genetic mutation is autosomal dominant (like some in the TP53 gene), there could be a higher risk for offspring.
- A family history of lung cancer may increase the likelihood of a genetic predisposition, but environmental factors can still play a major role.

2.7.2.4 Genotype/Phenotype Correlation

Lung cancer can arise from alterations in specific genes, and these alterations affect the biological characteristics of the cancer (phenotype). Some significant genotype-phenotype associations in lung cancer include:

- EGFR Mutations (Epidermal Growth Factor Receptor): Genotype: Alterations in the EGFR gene, particularly exon 19 deletions and L858R point mutations, are frequently observed in non-smokers, especially among individuals of East Asian descent.Phenotype Such as mutations are linked to an improved response to targeted treatments, like tyrosine kinase inhibitors (TKIs), and are often associated with the adenocarcinoma subtype of non-small cell lung cancer (NSCLC).
- KRAS Mutations Genotype: KRAS mutations are frequently found in smokers and are often present in adenocarcinoma cases.Phenotype: These mutations are connected to a more aggressive type of lung cancer and a worse prognosis. They are generally resistant to targeted therapies, complicating treatment options.

- TP53 Mutations Genotype: Changes in the TP53 tumor suppressor gene are present in various cancer types, including lung cancer.phenotype of TP53 mutations correlate with a poor prognosis, resistance to therapies, and a heightened chance of developing small cell lung cancer (SCLC), which is typically more aggressive.
- ALK Rearrangements (Anaplastic Lymphoma Kinase) Genotype: Alterations in the ALK gene create fusion proteins that promote cancer progression.Phenotype of ALK,mutations are often observed in younger, non-smoking patients with adenocarcinoma and are linked to a favorable response to targeted therapies such as ALK inhibitors (e.g., crizotinib).
- BRCA2 Mutations Genotype: Changes in BRCA2 are related to a higher risk of lung cancer, especially in individuals with a family history of other cancers (e.g., breast or ovarian).Its Phenotype is these mutations might increase the risk of developing lung cancer at a younger age and can also influence treatment approaches, particularly with PARP inhibitors.

2.7.3 Environmental and Lifestyle Factors Influencing Phenotype

While genetic factors (genotype) significantly contribute, environmental and lifestyle elements, such as smoking, air pollution, and workplace exposures, can alter the phenotype of lung cancer. For instance, smokers with KRAS mutations tend to develop lung cancer with a unique set of molecular features compared to non-smokers with EGFR mutations.

2.7.4 Impact on Treatment

Genotype plays a crucial role in predicting the response of lung cancer to treatment. For example:EGFR mutations show a good response to targeted therapies (e.g., erlotinib, gefitinib). ALK rearrangements are responsive to ALK inhibitors

(e.g., crizotinib, alectinib). KRAS mutations tend to be more resistant to existing targeted therapies, often treated with chemotherapy or immunotherapy.

2.7.5 Hereditary Lung Cancer Syndromes

Certain inherited mutations (e.g., LKB1, TP53) make individuals more susceptible to lung cancer. In families with these mutations, the genotype-phenotype relationship might indicate an earlier onset of lung cancer and a higher occurrence among multiple family members, which can inform clinical monitoring and early intervention strategies.

In summary, genotype and phenotype are closely linked in lung cancer. The genetic composition of the tumor, including mutations and rearrangements, directly affects the clinical features, progression, and treatment response. By investigating these associations, researchers can create more accurate diagnostic tools and tailored therapies for lung cancer patients.

2.8 Management

Approach to managing lung cancer requires a team-oriented strategy that combines diagnostic, therapeutic, and supportive care methods customized to the patient's cancer stage, tissue type, genetic characteristics, and overall health status.

2.8.1 Diagnosis and Staging

Achieving a precise diagnosis is critical for effective treatment. Methods such as CT scans, PET scans, bronchoscopy, and biopsies assist in assessing the tumor's location, size, and extent of spread. Employing molecular and genetic profiling through bioinformatics resources like Coremine, STRING, KEGG Pathway, and DAVID helps to identify mutations (e.g., EGFR, ALK, KRAS) that inform targeted therapies.

2.8.2 Treatment Strategies

Management approaches differ based on whether the cancer is non-small cell lung cancer (NSCLC) or small cell lung cancer (SCLC) and the stage of the disease.

- **Surgery:** This is the preferred option for early-stage NSCLC (lobectomy, pneumonectomy) but is rarely utilized in SCLC due to the early occurrence of metastasis.
- **Radiotherapy:** This modality is applied for localized tumors or as palliative treatment in advanced stages.
- **Chemotherapy:** This is the standard treatment for advanced and metastatic lung cancer, particularly in SCLC, with platinum-based agents (cisplatin, carboplatin) being commonly used.
- **Targeted Therapy**: Patients with certain genetic mutations can receive targeted treatments (e.g., EGFR inhibitors like osimertinib, ALK inhibitors like crizotinib).

2.8.3 Supportive and Palliative Care

Patients facing advanced lung cancer gain from symptom alleviation strategies, including pain management, oxygen therapy, and dietary support. Psychosocial interventions play a vital role in enhancing the overall quality of life.

2.8.4 Future Perspectives and Research

Innovations in bioinformatics, artificial intelligence, and precision medicine are transforming the management of lung cancer. Predictive modeling that leverages gene expression data aids in crafting personalized treatment strategies, leading to improved survival rates and minimized side effects. Current research is investigating new biomarkers, liquid biopsy methods, and combination therapies to optimize lung cancer treatment results.

Chapter 3

Research Methodology

Through its unique text mining algorithms and comprehensive biomedical data integration and analysis pipeline, the Norwegian bioinformatics firm PubGene has created a range of tools referred to as COREMINETM. These products were developed by the company.

Building upon the COREMINE Platform, Coremine MedicalTM is the first informational community specifically designed for a distinct domain. This free online service enables users to search, update, and share medical information, functioning as a search engine and a social network [37]

In the realm of text mining, it is crucial to have access to relevant materials on the subject [12]. The primary source, the PubMed database, offers this information through a published corpus about lung cancer within the biomedical literature [13]. To find genes associated with lung cancr a search was conducted within the MeSH database for the main topic "lung cancer." A collection of texts published over the last ten years, from 2014 to 2024, was compiled.

3.1 Retrieval of Genes

Retrieval of genes related to lung cancer by using COREMINE Significance of the Step

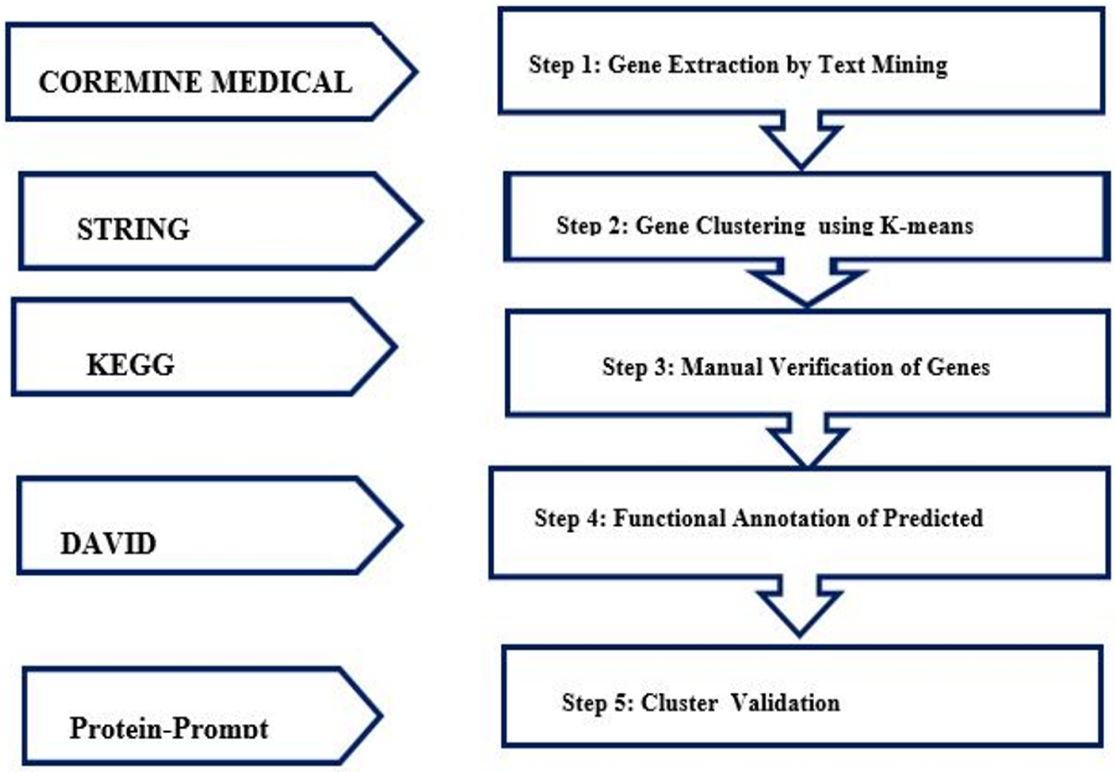


FIGURE 3.1: Mapping the path: Five bioinformatics tools unraveling liver cancer genes [40–44, 55–62].

1. Current and Relevant Information: Retrieving literature published within the last 10 years ensures access to up-to-date information, which is crucial for informing and advancing scientific understanding in field of lung cancer

2. Targeted Search: By searching for both MeSH major topics and specific terms, the retrieved literature is highly relevant to the topic of interest, minimizing irrelevant results and optimizing the effectiveness of text mining efforts

3.2 Exploring Connections from Biomedical Literature

When it comes to analyzing text in the biomedical literature, there is a variety of tools available. A biomedical text mining technology called COREMINE medical was used in order to extract a variety of information from published scholarly works. This information included genes, proteins, MeSH terms, processes, ailmen-

ts, and drugs. Downloaded in a text document were the results that were discovered.

3.3 Occurrence of Genes by Text Mining

Text mining is a fascinating and developing area of thesis that tries to find a solution to the problem of having an excessive amount of information present. The field of data mining encompasses a wider range of approaches, which are included into text mining [47, 63, 64]. However, because it focuses on unorganized data, a large percentage of the text mining approach revolves around the key pre-processing stage of arranging the document collections. This is because the pre-processing stage is particularly important. Techniques such as text classification [48, 65], word extraction [49], and information extraction (IE) [50, 51] are utilized in order to do this. Pre-processing of the document collection is one of the phases that are involved in the text mining process, which consists of multiple steps: [66]

1. It involves storing the intermediate representations
2. Assessing these representations through the use of methods such as distribution analysis
3. Formation of Clustre [52]
4. Trend analysis [53]
5. Association rules [63, 67]
6. Visualizing the results [21, 65]

A typical text mining system begins with a collection of documents that have not been processed, which implies that these documents do not have any labels or tags related to them. When the papers are first tagged, the tags are automatically generated based on the extraction of 'categories,' 'entities,' or 'relationships' directly from the documents. After that, the categories, entities, or connections

that have been retrieved are utilised in order to facilitate a variety of data mining actions that are performed on the documents. A subfield of information retrieval [63, 66, 68] known as text categorization [69–72] involves the process of breaking a large collection of documents into subgroups according to certain criteria that have been established beforehand. As of right now, users of PubMed have access to curated collections of documents that they can search through. These collections include parts that are specifically devoted to AIDS literature and the History of Medicine. The huge collection contains a large number of documents, each of which is labelled with words that represent characteristics of particular groups. Labelling in this manner makes it possible to establish a connection between the document (or website) in question and an appropriate category. Streamlining subsequent tasks for mining tools and increasing the likelihood that these tools will extract the most relevant information from the text can be accomplished by limiting the range of documents that are subject to analysis to certain subcategories that are pertinent to the study. In most cases, the identification of factual information included within the text is performed through the utilisation of Information Extraction (IE) techniques.

3.4 Clustering in Associated Genes in Lung Cancer and Verification by KEGG

Clustering is a technique used in data mining and machine learning to group together data points that are similar to each other based on specific properties. Creating a condition where the data points within each cluster are more comparable to each other than the data points within other clusters is one of the objectives when dividing a dataset into subsets or clusters. Several different clustering techniques, including K-means, hierarchical clustering, and DBSCAN, are used to automatically find these groupings [73, 74].

The Kegg pathway contained 165 pathways of lung cancer. We observed clusters ranging from 3 to 20 from the STRING database in all 165 pathways. We observed

matching genes in following pathways:

Pathway 1 naming (Pathway of Non-small cell lung cancer) and Cluster no 3 i.e.Tp53

Pathway 2 naming (Small Cell Lung Cancer) and Cluster no 4 i.e. Tp53

Pathway 7 naming(Pathway of cancer) and Cluster no 15 i.e. Tp53

Remaining genes RAR1 RALC RXRA RXLC RXRG TP53 CDKNA GADD45G GADD45A GADD45B BAX BCL2 BAK1 DDB2 POLK CYCS APAF1 CASP3 MYC MAX ZBTB17 CDKN2B CK4 CDK6 CCND1 CKS1B CKS2 SKP2 CDKNB CDK2 CCNE1 CCNE2 LC1 E2F1 E2F2 E2F3 COL4A1 COL4A2 COL4A3 COL4A4 COL4A5 COL4A6 LAMA1 LAMA2 LAMA3 LAMA5 LAMA4 LAMB1 LAMB2 LAMB3 LAMC3 LAMC1 LAMC2 FN1 ITGA2 ITGAB ITGA3 ITGA6 ITGAV ITGB1 PTK2 PIK3CA PIK3CB PIK3CD P3B3URE PIK3R3 PIK3R1 PIK3R2 PIK3R3 PTEN AKT3 AKT1 AKT2 CHUK IKBKG RELA XIAP TRAF4 TRAF4 PTGS2 NFKBTA BIRC2 BIRC7 TRAF2 TRAF5 NOS2 NFKB1 BIRC3 BCL2L1 TRAF3 TRAF6 Non-small cell lung cancer genes. FHIT RALC RXRA RXLC RARβ RXLC RXRA CDKN2A CDKU CDK6 CCND1 LC1 E2F1 E2F2 E2F3 KRAS RASSF1 RASSF5 STK4 PIK3CA PIK3CB PIK3CD PIK3R1 PIK3R2 PIK3R3 PDPK1 AKT3 AKT1 HGF BAD CASP9 FOXO3 SOS2 TGFA EGFR ELCB2 RAF1 HRAS ARAF BRAF MAPK1 MAP2K1 MAP2K2 PLCG1 PLCG2 PRKCA PRKCB PRKCG TP53 CDKNA GADD45G GADD45A GADD45B BAX BAK1 DDB2 POLK KIF5A KIF5B KIF5C RET EML4 ALK STAT5A STAT5B JAK3 STAT3 were all predicted genes

Table 3.1: Observing matching genes in LC pathways and Clusters.

Sr. No.	**Pathway Name**	**Cluster No.**	**Genes in Cluster**
1	Non-small cell lung cancer	Cluster No 3	TP53
2	Small Cell Lung Cancer	Cluster No 4	TP53
3	Chemical carcinogenesis - DNA adducts		

Continue on next page

Sr. No.	Pathway Name	Cluster No.	Genes in Cluster
4	Chemical carcinogenesis - receptor activation		
5	Chemical carcinogenesis - reactive oxygen species		
6	Purine metabolism		
7	Cancer Pathway		
8	Metabolic pathways		
9	EGFR tyrosine kinase inhibitor resistance		
10	Endocrine resistance		
11	Platinum drug resistance		
12	mRNA surveillance pathway		
13	Virion - Hepatitis viruses		
14	MAPK signaling pathway		
15	ELCB signaling pathway		
16	Ras signaling pathway		
17	Rap1 signaling pathway		
18	Calcium signaling pathway		
19	cAMP signaling pathway		
20	Chemokine signaling pathway		
21	NF-kappa B signaling pathway		
22	HIF-1 signaling pathway		
23	FoxO signaling pathway		
24	Phosphatidylinositol signaling system		
25	Sphingolipid signaling pathway		
26	Phospholipase D signaling pathway		
27	Hormone signaling		
28	Cell cycle		
29	Oocyte meiosis		
30	p53 signaling pathway		

Continue on next page

Sr. No.	Pathway Name	Cluster No.	Genes in Cluster
31	Mitophagy - animal		
32	Autophagy - animal		
33	Protein processing in endoplasmic reticulum		
34	Endocytosis		
35	mTOR signaling pathway		
36	PI3K-Akt signaling pathway		
37	AMPK signaling pathway		
38	Apoptosis		
39	p53 signaling pathway		
40	Parkinson disease		
41	Huntington disease		
42	Bacterial invasion of epithelial cells		
43	Epithelial cell signaling in Helicobacter pylori infection		
44	Coronavirus disease - COVID-19		
45	Basal cell carcinoma		
46	Calcium signaling pathway		
47	Hematopoietic cell lineage		
48	Hypertrophic cardiomyopathy		
49	Malaria		
50	Influenza A		
51	Tuberculosis		

3.5 Functional Annotation of Lung Cancer Genes by DAVID

The predicted genes were functionally annotated using DAVID. An established web server and web service for functional annotation of gene lists are part of the

bioinformatics resource system DAVID. It comes with a sizable knowledge base and a number of functional analysis tools [89]. DAVID was performed after a list of predicted genes was uploaded. Since these genes are found in humans, Homosapiens species were chosen. A collection of settings is required to operate the DAVID gene ontology. The most stringent categorization criterion was chosen in order to efficiently screen the lung cancer genes' functions [57]

3.6 Cluster Validation through Protein-Protein Interaction

The physical contacts formed between two or more proteins are known as protein-protein interactions (PPIs), and they are essential for a number of biological activities that occur within cells. In addition to numerous other cellular processes, these interactions are essential for signal transduction, enzyme activity regulation, and cellular transport [66]. In this step, we have validated some proteins from cluster no 3 and cluster no 4 through Protein-Protein interactions by server (https://proteinformatics.uni-leipzig.de/ProteinPrompt/). Version 2.0 of the Protein Prompt server is made available by the Protein Bioinformatics Group at the University of Leipzig [62].

Chapter 4

Results

4.1 Retrieval of Genes by COREMINE

In the first step, we extracted the genes by using the COREMINE tool. This was done by opening the search bar of the COREMINE tool and searching the disease i.e.Lung Cancer Then after clicking on extracted associations in the COREMINE interface we had to extract the associations from the year 2014 to 2024. Using the query lung cancer (Genes/Proteins), COREMINE provided a list of 2000 genes implicated lung cancer. An Excel list including the names of the associated genes, diseases, and their descriptions was obtained.

Table 4.1: List of genes of lung cancer.

Genes	**Genes**	**Genes**	**Genes**
RASSF5	PPOX	CCDC6	GPX7
ALKBH5	BRD4	HELLS	BCL2L1
CASP9	FOSL1	PSMB5	BAD
SSTR5	SPA17	EIF3B	CCDC88A
TNS1	FOSL2	MIR17HG	CNTNAP2
MSH2	TRIB2	MTHFD2	LGI1
KEAP1	RALB	NOTCH3	ESRP1

Continue on next page

Genes	**Genes**	**Genes**	**Genes**
HSP90AA1	SST	DVL3	PBK
IGF2BP3	HCP5	MIRLET7D	SMARCAL1
NQO1	DTX1	H3F3AP6	HTR1D
TCEAL1	POU4F3	CDK4	FAT4
YTHDF3	MAP1A	XRS	HDAC10
PVT1	KLHL1	PTPRC	QPCT
FOXM1	LIMK2	PLXNA1	POMC
THM	CDH6	GRIK	SNORD116
EPAS1	NUSAP1	PF4V1	ANXA1
E2F7	MGMT	CEACAM1	CALCB
TRIP4	CD276	SNCA	DACH1
AMBRA1	CXCL12	EPHA7	PAX5
PTPRU	TSPAN8	OPN3	IMPDH1
LTB4R2	EP300	PSMB4	CDS1
PPP1R13L	DPEP3	IDH3A	SYN2
SPRY4	WNT6	MBD3	KDM2A
TNFSF18	METAP1	MAGEA4	MAGI2
HRNR	MELK	CADM1	MB21D1
CSNK1A1	SHOX2	SEMA3C	CDC6
TPD52	DEGS1	TOP2B	CCKBR
CD81	ASS1	IMPDH2	LCM3
TSPAN7	HNRNPK	TRDMT1	RIMBP2
SLC	CARD8	GAST	GAD2
PLCG2	AKAP4	MAPKAPK2	MSI2
DHODH	TNFRSF9	ABCA2	KLK8
MUC1	ATR	DLST	USO1
KIF11	NTNG1	NFKBIB	CGA
TREM1	ZFPM2	SMYD3	LTBP1
WNT2B	AGPAT9	KIAA1524	NHLH1

Continue on next page

Genes	**Genes**	**Genes**	**Genes**
RNPEP	DNMT3A	WLS	TSPAN12
ARHGAP4	CASC5	TAC1	CTAG1B
BUB1B	STK38	COL18A1	CALCA
BMP8B	TXNIP	RUVBL1	SEMA3F
GRM8	ADM2	NCAPG	SULT1A3
MIR184	PCNA	TACC3	CIITA
RXFP2	TP63	MAD2L2	PRKDC
HUWE1	ERCC4	ANK2	NRAS
NANOGP8	MEG3	MAP2K5	CHRNA4
PAX2	NANOG	IFI27	ELCB2IP
NEURL1	CKS1B	KLF15	SRSF1
ATXN8	U2AF2	IGFBP5	SEMA6A
CIB1	AJUBA	TAGLN	TAS2R64P
LTBP4	ASCC3	SYCE1L	DMPK
E2F1	SNHG6	AMER1	RALGAPB
SLC35A1	TRAPPC10	C19orf48	TYMS
ANGPT4	RRM2	AKR1C1	TREX1
KDR	TMEM132D	TOP1	ROBO1
HEPACAM2	IGF1R	RTN1	ANKRD28
EGR4	TGFBR2	G3BP1	LILRB1
DCTN6	HEPACAM	PPP1R11	CDC25A
PID1	APEH	BCYRN1	CACNA2D4
C20orf85	RPS6KB2	NR1D2	IMPDH1P11
CCDC54	KSR1	ABCF1	RFC4
TCP1	PCLO	DDB2	MAML3
ITPKB	RASSF2	MEN1	POLDIP2
NTS	CHMP3	ABCF2	ASPM
BAZ1A	WFDC2	IGHD1-7	STXBP5
CENPF	PAX9	MTRNR2L5	SOX1

Continue on next page

Genes	**Genes**	**Genes**	**Genes**
AFF3	HDAC9	FAIM2	CD44
SSTR3	TRHDE	DHX9	CDK6
MST1	ABCC2	RASGRF2	PARP4
ALDH1A1	KMT2B	NTNG2	RAB3B
MCM4	BMX	S100A16	PIP4K2A
MPP3	TOP2A	CDKN1A	KRT18
AASDHPPT	ORC4	TERC	GTSE1
PSMA1	MAPK15	XRCC1	CDKN2A
CXCR4	CRIM1	SBF2	CACNG3
LRSAM1	TPTE	MAX	ARID2
SIVA1	COTL1	VEGFA	WWTR1
HAVCR2	SCUBE1	SLC35C1	CCNO
ZMYND10	SMARCA4	BIRC5	ETV4
FAT1	PPIE	TBC1D9	ASH1L
ASCL1	POU2F3	DLL3	RB1
MIR513C	MYCL	GPR137C	NEUROD1
BRWD1-AS1	YAP1	GRP	FAM135B
SLFN11	TTF1	LINC00173	CD274
FAM66A	INSM1	CHGA	TP53
HIST1H2AL	NKX2-1	ENO2	NCAM1
SNORA71A	SSTR2	SPICE1	KIAA1211
MYC	EGFR	SYP	MYCN
FAM171B	C11orf53	ZNF429	KDM1A
HIST1H4L	LRP1B	PDCD1	RPS6P2
SEZ6L	COL11A2	PARP1	NFIB
TBC1D2B	ASXL3	MTUS2	SEZ6
SAMD5	ITGA10	PCDHA1	TCEAL2
ELAVL4	MTUS1	METTL5	CEMIP
PI4K2B	MS4A8	SRRM4	NKX1-2

Continue on next page

Genes	**Genes**	**Genes**	**Genes**
RASSF1	TAS2R6P	KMT2D	RPL17
TAS2R18P	MET	BTBD3	MIR486-1
PLEKHA4	REPS1	ZCCHC4	ASTN1
BCL2	SST	REST	ABCC1
TCPT	RALBP1	PIK3CA	TXK
ROS1	ARHGEF19	TSC1	PTPN5
SPTA1	EPHA3	POU3F2	SEMA3B
MSH3	PARVG	MVP	KRAS
ZIC4	FAT3	CSMD3	PARP6
TDGF1P7	KIT	KRT19	ABCB1
MDM4	PARP16	KCTD16	TMSB15A
OTP	CTLA4	RLF	ALK
EPCAM	SSTR1	CCT5	FGFR1
RCVRN	MBD2	NOTCH1	PTEN
GAGE2A	CSMD2	MCTP2	NOL4
ZFHX2	SELV	NKX1-1	RNF113A
RBL2	CEP350	MS4A7	RAB15
BRD9	SLC35G4	SLC35E4	FZD8
SLC35G6	KIF1BP	SLC35E3	ADAMTS19
SCGN	RBM5	SLC35G3	SLC35E2B
SLC35E2	CDR2L	SLC35F6	MIR200B
CCT2	SLC35G1	EZH2	FOXI3
ERCC1	SLC35E1	SLC35F4	HSPB3
TSPAN9	SLC35G2	SLC35A5	SLC35C2
KCNH4	PFN2	MYCNOS	MIR95
LINC00968	ALDH1L2	PTENP1	CASC8
CDK7	RBL1	SLC35F1	RICTOR
DPYSL5	C21orf33	SLC35B1	LAG3
HES1	MTOR	SOX2	SLC35F3

Continue on next page

Genes	**Genes**	**Genes**	**Genes**
PARD6B	TMPRSS13	ZFPM2-AS1	SLC35D3
MYCT1	SPATA2	MCL1	WRNIP1
SLC35D1	PDCD11	SLC35B4	SLC35B3
NOTCH2	ZEB2-AS1	AURKA	CD9
CDH10	HOPX	UBE2C	RET
TIGIT	BUB1	SFMBT2	SBF2-AS1
NBR1	CNN1	PDCD1LG2	NAPSA
ABCE1	CSF3	FBN3	SLC35F2
PPCDC	CEACAM5	PAGE5	CDS2
SKP2	GRPR	RBM15B	GPC2
AURKB	RBM6	SLC35D2	ACTBP2
DNER	RNMT	ACTBP8	NMBR
KDM6A	HGF	SLC35B2	ADGRB3
CREBBP	TERT	CDKN1B	KMT2C
SLC35F5	STK33	CHRNA6	HIF1A-AS2
BMPER	CD151	EPB41L3	PSMA4
PSMA5	STK11	ESPL1	CDH1
MYH8	KRT4	TNKS	ZMAT3
POU2F2	PROM1	YES1	MAGEA5
RRAD	ERBB2	ZNF521	CIDEB
CHEK1	MSMB	DERL2	HOXD8
NPTXR	KITLG	SLC35G5	ASH1L
PIP5K1C	MUC16	TBC1D9	FAT1
AVP	PPIE	BIRC5	ZMYND10
CCNO	ETV4	SLC35C1	HAVCR2
VEGFA	SMARCA4	COTL1	SCUBE1
WWTR1	SIVA1	ARID2	LRSAM1
CACNG3	SBF2	MAX	TPTE
CDKN2A	XRCC1	MAPK15	CRIM1

Continue on next page

Genes	**Genes**	**Genes**	**Genes**
CXCR4	ORC4	PSMA1	KRT18
CDKN1A	GTSE1	TOP2A	MPP3
TERC	AASDHPPT	S100A16	PIP4K2A
RAB3B	NTNG2	KMT2B	BMX
PARP4	RASGRF2	MCM4	ALDH1A1
CDK6	ABCC2	MST1	TRHDE
CD44	FAIM2	DHX9	AFF3
SOX1	SSTR3	MTRNR2L5	CENPF
HDAC9	PAX9	WFDC2	BAZ1A
STXBP5	ASPM	IGHD1-7	CHMP3
NTS	RASSF2	ABCF2	POLDIP2
ITPKB	MAML3	MEN1	PCLO
TCP1	RFC4	DDB2	CCDC54
NR1D2	RPS6KB2	ABCF1	KSR1
CACNA2D4	IMPDH1P11	PID1	CDC25A
C20orf85	BCYRN1	HEPACAM	APEH
LILRB1	TGFBR2	PPP1R11	DCTN6
G3BP1	ANKRD28	RTN1	HEPACAM2
EGR4	TOP1	IGF1R	TMEM132D
ROBO1	RRM2	ANGPT4	TYMS
KDR	TREX1	AKR1C1	TRAPPC10
SLC35A1	RALGAPB	C19orf48	E2F1
DMPK	AMER1	ASCC3	SNHG6
TAS2R64P	SYCE1L	AJUBA	CIB1
SOX2-OT	KRT20	ZNF423	MSI1
TAGLN	LTBP4	GAGE7	PTK2
SNHG3	NCR3LG1	RAF1	PDIA2
CASP3	HOTAIR	SLC35A4	LHX6
HNRNPA2B1	RALA	BRAF	CD47

Continue on next page

Genes	**Genes**	**Genes**	**Genes**
ABCC3	CCNA2	RPS6KA2	GPT2
CADM2	M6PR	FOXI1	GBP5
DLX5	RPA3	RHBDF1	PAK3
LSAMP	PVRL3	NOVA1	GFI1B
HOXB2	STAB2	KRT7	DYRK1B
WEE1	CD24	CYCSP5	CHL1
ITGB6	E2F3	RASSF5	PPOX
CCDC6	GPX7	ALKBH5	BRD4
HELLS	BCL2L1	CASP9	FOSL1
PSMB5	BAD	SSTR5	SPA17
EIF3B	CCDC88A	TNS1	FOSL2
MIR17HG	CNTNAP2	MSH2	TRIB2
MTHFD2	LGI1	KEAP1	RALB
NOTCH3	ESRP1	HSP90AA1	DVL3
PBK	IGF2BP3	HCP5	MIRLET7D
SMARCAL1	NQO1	DTX1	H3F3AP6
HTR1D	TCEAL1	POU4F3	CDK4
FAT4	YTHDF3	MAP1A	XRS
HDAC10	PVT1	KLHL1	PTPRC
QPCT	FOXM1	LIMK2	PLXNA1
POMC	THM	CDH6	GRIK3
SNORD116@	EPAS1	NUSAP1	PF4V1
ANXA1	E2F7	MGMT	CEACAM1
CALCB	TRIP4	CD276	SNCA
DACH1	AMBRA1	CXCL12	EPHA7
PAX5	PTPRU	TSPAN8	OPN3
IMPDH1	LTB4R2	EP300	PSMB4
CDS1	PPP1R13L	DPEP3	IDH3A
SYN2	SPRY4	WNT6	MBD3

Continue on next page

Genes	**Genes**	**Genes**	**Genes**
KDM2A	TNFSF18	METAP1	MAGEA4
MAGI2	HRNR	MELK	CADM1
MB21D1	CSNK1A1	SHOX2	SEMA3C
CDC6	TPD52	DEGS1	TOP2B
CCKBR	CD81	ASS1	IMPDH2
RBM3	TSPAN7	HNRNPK	TRDMT1
RIMBP2	SLC4A3	CARD8	GAST
GAD2	PLCG2	AKAP4	MAPKAPK2
MSI2	DHODH	TNFRSF9	ABCA2
KLK8	MUC1	ATR	DLST
USO1	KIF11	NTNG1	NFKBIB
CGA	TREM1	ZFPM2	SMYD3
LTBP1	WNT2B	AGPAT9	KIAA1524
NHLH1	RNPEP	DNMT3A	WLS
TSPAN12	ARHGAP4	CASC5	TAC1
CTAG1B	BUB1B	STK38	COL18A1
CALCA	BMP8B	TXNIP	RUVBL1
SEMA3F	GRM8	ADM2	NCAPG
SULT1A3	MIR184	PCNA	TACC3
CIITA	RXFP2	TP63	MAD2L2
PRKDC	HUWE1	ERCC4	ANK2
NRAS	NANOGP8	MEG3	MAP2K5
CHRNA4	PAX2	NANOG	IFI27
ERBB2IP	NEURL1	CKS1B	KLF15
SRSF1	ATXN8	U2AF2	IGFBP5

4.2 Network Building

For this step we had to go to the STRING database. Here we inserted a list of

782 genes of SCLC After that we did the selection of organisms i.e. in this case Homo sapiens. After that a network was generated in which we had 61 nodes (Fig 4.1).

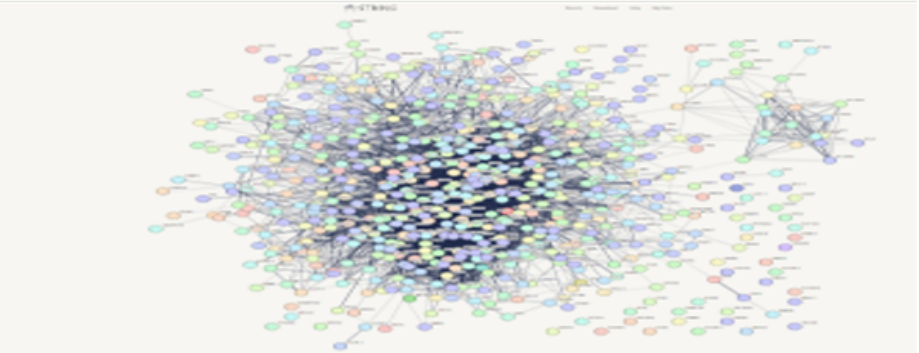


FIGURE 4.1: Network of 782 Genes.

4.3 Cluster Generation

The next step was cluster generation. For this purpose we clicked on clusters and then go to K-means clustering and set numbers of clusters gradually from 3,4,5,6 and so on and then apply (Fig 4.2) [75].

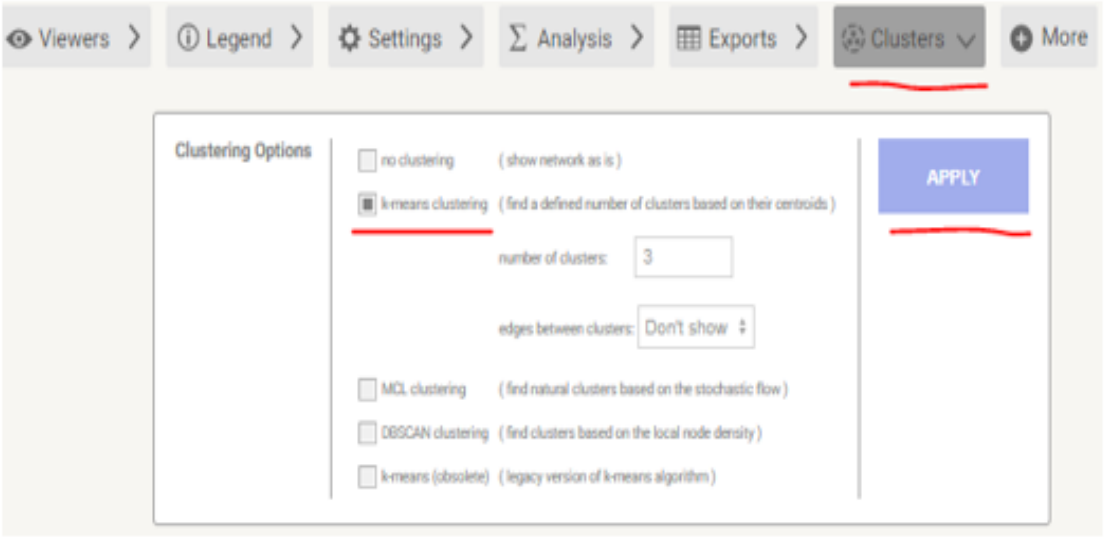


FIGURE 4.2: Executing K-means Clustering.

I had verified clusters from 3 to 24 from the KEGG pathway. In the KEGG pathway there were 165 pathways of lung cancer and clusters from 3 to 20 were verified from all 165 pathways.

4.4 Functional Annotation by DAVID Tool

Functional annotation was carried out using the David tool. Clusters were the output reseult of the functional annotation process. We were able to form thirty distinct clusters for the lung cancer genes based on the output of the David tool. The outcomes are displayed below:

4.4.1 Functional Annotation Clustering

4.4.1.1 Annotation Cluster 1

Annotation Cluster 1 (Fig 4.3) has a 4.58 enrichment score in lung cancer gene functional annotation clustering. With an enrichment score of 6.93, genes related to foreign substance transportation (p-value: 6.4E-8), and the lower side of plasma membranes (p-value: 9.5E-8) increased. This suggests certain biological processes and cellular components may cause lung cancer.

Annotation Cluster 1	Enrichment Score: 6.93	G		Count	P_Value	Benjamini
GOTERM_BP_DIRECT	vascular endothelial growth factor receptor-1 signaling pathway	RT		11	6.4E-8	1.7E-5
GOTERM_MF_DIRECT	protein tyrosine kinase collagen receptor activity	RT		11	6.4E-8	3.1E-6
GOTERM_MF_DIRECT	brain-derived neurotrophic factor receptor activity	RT		11	6.4E-8	3.1E-6
GOTERM_MF_DIRECT	GPI-linked ephrin receptor activity	RT		11	6.4E-8	3.1E-6
GOTERM_MF_DIRECT	insulin receptor activity	RT		11	6.4E-8	3.1E-6
GOTERM_MF_DIRECT	macrophage colony-stimulating factor receptor activity	RT		11	6.4E-8	3.1E-6
GOTERM_MF_DIRECT	boss receptor activity	RT		11	6.4E-8	3.1E-6
GOTERM_MF_DIRECT	platelet-derived growth factor alpha-receptor activity	RT		11	6.4E-8	3.1E-6
GOTERM_MF_DIRECT	stem cell factor receptor activity	RT		11	6.4E-8	3.1E-6
GOTERM_MF_DIRECT	placental growth factor receptor activity	RT		11	6.4E-8	3.1E-6
GOTERM_BP_DIRECT	platelet-derived growth factor receptor-alpha signaling pathway	RT		11	7.8E-8	1.8E-5
GOTERM_MF_DIRECT	epidermal growth factor receptor activity	RT		11	7.9E-8	3.1E-6
GOTERM_MF_DIRECT	fibroblast growth factor receptor activity	RT		11	7.9E-8	3.1E-6
GOTERM_MF_DIRECT	hepatocyte growth factor receptor activity	RT		11	7.9E-8	3.1E-6
GOTERM_MF_DIRECT	platelet-derived growth factor beta-receptor activity	RT		11	7.9E-8	3.1E-6
GOTERM_BP_DIRECT	Kit signaling pathway	RT		11	9.5E-8	1.9E-5

FIGURE 4.3: Functional annotation cluster 1, of lung cancer.

4.4.1.2 Annotation Cluster 2

Functional annotation clustering study of lung cancer genes showed 3.87 enrich-

ment in Annotation Cluster 2 (Fig 4.4). Many measures related to amino acid transport and transmembrane transporter activity increased in the study. Significant enrichment was seen for "amino acid transmembrane ABC transporter activity" (p-value: 4.2E-5) and "Amino-acid transporter 2" (5.9E-4). This suggests amino acid transport mechanisms may cause lung cancer.

Annotation Cluster 2	Enrichment Score: 3.87			Count	P_Value	Benjamini
UP_SEQ_FEATURE	DOMAIN:ABC transporter 1	RT		7	4.2E-5	8.2E-3
UP_SEQ_FEATURE	DOMAIN:ABC transporter 2	RT		7	4.2E-5	8.2E-3
INTERPRO	ABC_transporter-like_CS	RT		7	3.3E-4	2.2E-2
INTERPRO	ABC_transporter-like_ATP-bd	RT		7	5.9E-4	3.3E-2

FIGURE 4.4: Functional annotation cluster 2, of lung cancer.

4.4.1.3 Annotation Cluster 3

Functional annotation cluster 3 (Fig 4.5) of lung cancer genes has a 3.71 enrichment score. The study found significant increases in transmembrane transporter and antiporter activities. The "EGF-like 13 calcium binding (p-value: 1.5E-4) and EGF-like 14 calcium binding" (p-value: 1.5E-4) activities were considerably enriched. Additionally, EGF like 11 (p-value: 3.4E-4) was enriched. These data suggest that transmembrane transporter activity and EGF like activities may cause lung cancer.

Annotation Cluster 3	Enrichment Score: 3.71			Count	P_Value	Benjamini
UP_SEQ_FEATURE	DOMAIN:EGF-like 13; calcium-binding	RT		5	1.5E-4	2.2E-2
UP_SEQ_FEATURE	DOMAIN:EGF-like 14; calcium-binding	RT		5	1.5E-4	2.2E-2
UP_SEQ_FEATURE	DOMAIN:EGF-like 11; calcium-binding	RT		5	3.4E-4	4.5E-2

FIGURE 4.5: Functional annotation cluster 3, of lung cancer.

4.4.1.4 Annotation Cluster 4

Functional annotation cluster 4 (Fig 4.6) of lung cancer genes has an enrichment value of 2.96. The study found significant increases in transmembrane transporter and antiporter activities. The "EGF-like 28 calcium binding (p-value: 4.5E-4) and EGF-like 23 calcium binding" (p-value: 6.6E-4) activities were considerably enriched. Additionally, EGF like 21 (p-value: 6.6E-4) was enriched. These data suggest that transmembrane transporter activity and EGFlike activities may cause lung cancer.

Annotation Cluster 4	Enrichment Score: 2.96			Count	P_Value	Benjamini
UP_SEQ_FEATURE	DOMAIN:EGF-like 28	RT		4	4.5E-4	5.5E-2
UP_SEQ_FEATURE	DOMAIN:EGF-like 23; calcium-binding	RT		4	6.6E-4	6.9E-2
UP_SEQ_FEATURE	DOMAIN:EGF-like 21; calcium-binding	RT		4	6.6E-4	6.9E-2
UP_SEQ_FEATURE	DOMAIN:EGF-like 15; calcium-binding	RT		4	2.6E-3	1.9E-1
UP_SEQ_FEATURE	DOMAIN:EGF-like 16; calcium-binding	RT		4	3.3E-3	2.0E-1

FIGURE 4.6: Functional annotation cluster 4, of lung cancer.

4.4.1.5 Annotation Cluster 5

Annotation Cluster 5 (Fig 4.7) with an enrichment score of 2.82, highlights the presence of Retinoblastoma (LC)-associated domains, including LCC, LCB, LCA, and LCN, along with LC family-associated features. These domains are critical for the function of the LC tumor suppressor protein, which regulates cell cycle progression by inhibiting E2F transcription factors. The presence of the Retinoblastoma-associated protein A-box further supports its role in tumor suppression. Additionally, the cluster includes a biological process term related to lipid kinase activity regulation, which is linked to cell signaling, proliferation, and survival pathways—all of which are frequently dysregulated in cancer. The identification of DUF3452, a conserved protein family of unknown function, suggests potential novel regulatory interactions. Given the well-established role of LC1 in lung cancer, particularly in small-cell lung carcinoma (SCLC), this cluster underscores a possible involvement of LC-associated pathways in lung tumorigenesis, reinforcing the need for further investigation into LC.

Annotation Cluster 5	Enrichment Score: 2.82			Count	P_Value	Benjamini
INTERPRO	RB_C	RT		3	1.2E-3	5.1E-2
INTERPRO	RB_B	RT		3	1.2E-3	5.1E-2
INTERPRO	RB_A	RT		3	1.2E-3	5.1E-2
INTERPRO	RB_fam	RT		3	1.2E-3	5.1E-2
INTERPRO	RB_N	RT		3	1.2E-3	5.1E-2
UP_SEQ_FEATURE	REGION:Domain A	RT		3	1.2E-3	1.1E-1
UP_SEQ_FEATURE	DOMAIN:Retinoblastoma-associated protein A-box	RT		3	1.2E-3	1.1E-1
GOTERM_BP_DIRECT	regulation of lipid kinase activity	RT		3	1.3E-3	4.4E-2
SMART	DUF3452	RT		3	2.2E-3	8.8E-2
SMART	RB_A	RT		3	2.2E-3	8.8E-2
SMART	Rb_C	RT		3	2.2E-3	8.8E-2
UP_SEQ_FEATURE	REGION:Domain B	RT		3	2.5E-3	1.8E-1

FIGURE 4.7: Functional annotation cluster 5, of lung cancer.

4.4.1.6 Annotation Cluster 6

Annotation Cluster 6 (Fig 4.8), with an enrichment score of 2.66, highlights the presence of bromodomains, which are critical for recognizing acetylated lysine resi-

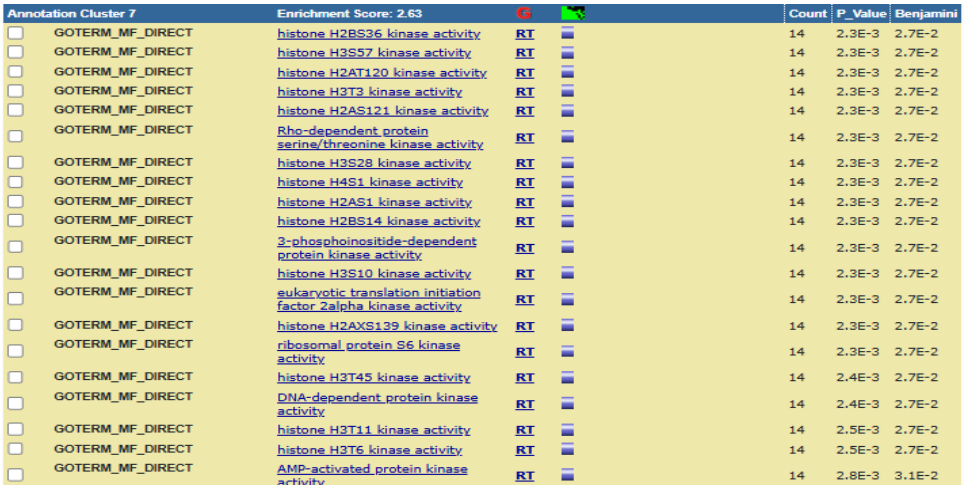
dues in histone proteins and play a key role in epigenetic regulation of gene expression. Bromodomain-containing proteins, such as BRD4, are frequently implicated in cancer progression, including lung cancer, by modulating transcriptional programs that drive tumor cell proliferation and survival. The enrichment of bromodomains in this dataset suggests a potential involvement of epigenetic dysregulation in lung cancer, making bromodomain inhibitors (BET inhibitors) a promising therapeutic avenue for targeting oncogenic transcription in lung tumors.

Annotation Cluster 6	Enrichment Score: 2.66			Count	P_Value	Benjamini
UP_SEQ_FEATURE	DOMAIN:Bromo	RT		6	1.5E-3	1.3E-1
UP_KW_DOMAIN	Bromodomain	RT		6	1.6E-3	1.4E-2
INTERPRO	Bromodomain	RT		6	1.8E-3	7.2E-2
SMART	BROMO	RT		6	5.6E-3	1.5E-1

FIGURE 4.8: Functional annotation cluster 6, of lung cancer.

4.4.1.7 Annotation Cluster 7

Annotation Cluster 7 (Fig 4.9) (enrichment score: 2.63) is enriched with histone kinase activities, including phosphorylation of histone H2B, H3, and H4, along with other kinases like ribosomal protein S6 kinase and AMP-activated protein kinase. Histone phosphorylation is crucial for chromatin remodeling, gene expression, and cell cycle regulation, processes often dysregulated in lung cancer.



Annotation Cluster 7	Enrichment Score: 2.63			Count	P_Value	Benjamini
GOTERM_MF_DIRECT	histone H2BS36 kinase activity	RT		14	2.3E-3	2.7E-2
GOTERM_MF_DIRECT	histone H3S57 kinase activity	RT		14	2.3E-3	2.7E-2
GOTERM_MF_DIRECT	histone H2AT120 kinase activity	RT		14	2.3E-3	2.7E-2
GOTERM_MF_DIRECT	histone H3T3 kinase activity	RT		14	2.3E-3	2.7E-2
GOTERM_MF_DIRECT	histone H2AS121 kinase activity	RT		14	2.3E-3	2.7E-2
GOTERM_MF_DIRECT	Rho-dependent protein serine/threonine kinase activity	RT		14	2.3E-3	2.7E-2
GOTERM_MF_DIRECT	histone H3S28 kinase activity	RT		14	2.3E-3	2.7E-2
GOTERM_MF_DIRECT	histone H4S1 kinase activity	RT		14	2.3E-3	2.7E-2
GOTERM_MF_DIRECT	histone H2AS1 kinase activity	RT		14	2.3E-3	2.7E-2
GOTERM_MF_DIRECT	histone H2BS14 kinase activity	RT		14	2.3E-3	2.7E-2
GOTERM_MF_DIRECT	3-phosphoinositide-dependent protein kinase activity	RT		14	2.3E-3	2.7E-2
GOTERM_MF_DIRECT	histone H3S10 kinase activity	RT		14	2.3E-3	2.7E-2
GOTERM_MF_DIRECT	eukaryotic translation initiation factor 2alpha kinase activity	RT		14	2.3E-3	2.7E-2
GOTERM_MF_DIRECT	histone H2AXS139 kinase activity	RT		14	2.3E-3	2.7E-2
GOTERM_MF_DIRECT	ribosomal protein S6 kinase activity	RT		14	2.3E-3	2.7E-2
GOTERM_MF_DIRECT	histone H3T45 kinase activity	RT		14	2.4E-3	2.7E-2
GOTERM_MF_DIRECT	DNA-dependent protein kinase activity	RT		14	2.4E-3	2.7E-2
GOTERM_MF_DIRECT	histone H3T11 kinase activity	RT		14	2.5E-3	2.7E-2
GOTERM_MF_DIRECT	histone H3T6 kinase activity	RT		14	2.5E-3	2.7E-2
GOTERM_MF_DIRECT	AMP-activated protein kinase activity	RT		14	2.8E-3	3.1E-2

FIGURE 4.9: Functional annotation cluster 7, of lung cancer.

Abnormal histone modifications contribute to oncogenesis, therapy resistance, and uncontrolled proliferation. Specifically, histone H3 phosphorylation (e.g., H3T3, H3T6, H3T11) is linked to mitotic progression and DNA damage response, key factors in lung cancer development. The enrichment of these kinase activities suggests their potential role as biomarkers or therapeutic targets in lung cancer.

4.4.1.8 Annotation Cluster 8

Annotation Cluster 8 (Fig 4.10) (enrichment score: 2.61) is associated with NAD+-dependent ADP-ribosyltransferase activity, specifically targeting histone H2BE glutamate residues. ADP-ribosylation is a key post-translational modification involved in DNA repair, transcriptional regulation, and chromatin remodeling, all of which play critical roles in lung cancer progression. Dysregulation of ADP-ribosylation can lead to genomic instability, altered gene expression, and resistance to DNA-damaging therapies, contributing to lung cancer development. The enrichment of these activities suggests a potential link between aberrant histone ADP-ribosylation and oncogenic pathways, highlighting its relevance as a possible biomarker or therapeutic target in lung cancer.

Annotation Cluster 8	Enrichment Score: 2.61			Count	P_Value	Benjamini
GOTERM_MF_DIRECT	NAD+-histone H2BE2 glutamate ADP-ribosyltransferase activity	RT		4	1.8E-3	2.7E-2
GOTERM_MF_DIRECT	NAD+-histone H2BE18 glutamate ADP-ribosyltransferase activity	RT		4	1.8E-3	2.7E-2
GOTERM_MF_DIRECT	NAD+-histone H2BE35 glutamate ADP-ribosyltransferase activity	RT		4	1.8E-3	2.7E-2
GOTERM_MF_DIRECT	NAD+-protein ADP-ribosyltransferase activity	RT		4	6.1E-3	5.7E-2

FIGURE 4.10: Functional annotation cluster 8, of lung cancer.

4.4.1.9 Annotation Cluster 9

Annotation Cluster 9 (Fig 4.11) (enrichment score: 2.59) is enriched with NAD+-dependent ADP-ribosyltransferase activity, including poly(ADP-ribose) polymerase (PARP) catalytic domains, nucleotidyltransferase, and glycosyltransferase activities.

Annotation Cluster 9	Enrichment Score: 2.59			Count	P_Value	Benjamini
GOTERM_MF_DIRECT	NAD+-protein-aspartate ADP-ribosyltransferase activity	RT		5	1.2E-4	3.0E-3
INTERPRO	Poly(ADP-ribose)pol_cat_dom	RT		5	3.3E-4	2.2E-2
UP_SEQ_FEATURE	DOMAIN:PARP catalytic	RT		5	3.4E-4	4.5E-2
GOTERM_MF_DIRECT	NAD+-protein poly-ADP-ribosyltransferase activity	RT		5	1.5E-3	2.7E-2
GOTERM_MF_DIRECT	nucleotidyltransferase activity	RT		5	1.7E-2	1.4E-1
UP_KW_MOLECULAR_FUNCTION	Glycosyltransferase	RT		5	8.3E-1	1.0E0

FIGURE 4.11: Functional annotation cluster 9, of lung cancer.

PARP enzymes play a crucial role in DNA repair, genomic stability, and cellular stress responses, processes frequently dysregulated in lung cancer. Increased PARP activity is often linked to tumor survival, therapy resistance, and altered metabolism in lung cancer cells. Inhibitors targeting PARP have shown promise

in enhancing the efficacy of DNA-damaging treatments, making these enriched activities potential biomarkers or therapeutic targets in lung cancer.

4.4.1.10 Annotation Cluster 10

Annotation Cluster 10 (Fig 4.12) (enrichment score: 2.49) is enriched with transcriptional regulation factors associated with the Myc family, including Myc transcription factors and Myc protein. Myc proteins are critical regulators of cell growth, differentiation, and apoptosis, with their overexpression frequently observed in various cancers, including lung cancer. Dysregulation of MYC activity contributes to uncontrolled cell proliferation, metabolic reprogramming, and resistance to therapy in lung cancer. These transcription factors are key drivers of tumorigenesis, making MYC proteins potential biomarkers and therapeutic targets for lung cancer treatment strategies.

Annotation Cluster 10	Enrichment Score: 2.49			Count	P_Value	Benjamini
INTERPRO	Tscrpt_reg_Myc	RT		3	2.4E-3	8.8E-2
INTERPRO	Tscrpt_reg_Myc_N	RT		3	2.4E-3	8.8E-2
INTERPRO	Myc_transcription_factors	RT		3	2.4E-3	8.8E-2
PIR_SUPERFAMILY	Myc_protein	RT		3	7.3E-3	4.6E-1

FIGURE 4.12: Functional annotation cluster 10, of lung cancer.

4.4.1.11 Annotation Cluster 11

The comprehensive annotation cluster statement reveals a notable association with lung cancer through significant enrichment of gene functions related to xenobiotic transport activities. Specifically, the cluster includes "ABC-type xenobiotic transporter activity," "xenobiotic transmembrane transporter activity," and "ATPase-coupled transmembrane transporter activity," each exhibiting strong statistical significance with p-values of 1.7E-4, 1.86E-3, and 8.3E-3, respectively. These transporter activities are crucial in mediating the cellular response to environmental toxins and drugs, which are critical factors in lung cancer pathogenesis and progression. Additionally, the presence of translocase functions suggests involvement in membrane transport processes that may influence cellular detoxification pathways, further highlighting the relevance of these molecular functions in the

context of lung cancer biology. Collectively, these annotations underscore the importance of transporter mechanisms in the modulation of lung cancer susceptibility and treatment responses (Fig 4.13).

Annotation Cluster 11	Enrichment Score: 2.3			Count	P_Value	Benjamini
GOTERM_MF_DIRECT	ABC-type xenobiotic transporter activity	RT		5	1.7E-4	4.0E-3
GOTERM_MF_DIRECT	xenobiotic transmembrane transporter activity	RT		5	1.8E-3	2.7E-2
GOTERM_MF_DIRECT	ATPase-coupled transmembrane transporter activity	RT		5	8.3E-3	7.3E-2
UP_KW_MOLECULAR_FUNCTION	Translocase	RT		5	2.5E-1	1.0E0

FIGURE 4.13: Functional annotation cluster 11, of lung cancer.

4.4.1.12 Annotation Cluster 12

The comprehensive annotation cluster statement highlights a significant association with lung cancer through the enrichment of homeobox-related genes, specifically in Annotation Cluster 12. This cluster, characterized by an enrichment score of 2.28, reveals key functional elements including "homeobox" and "HD" domains, both of which display notable p-values of 2.3E-3 and 2.8E-3, respectively. The presence of HOX genes further underscores their critical role in regulating developmental processes and cellular differentiation, which are often disrupted in lung cancer pathology. With a total count of 14 genes linked to these annotations, the data suggests a vital link between homeobox gene expression and lung cancer progression, emphasizing the potential for these factors as targets for therapeutic intervention and biomarker identification in lung cancer research (Fig 4.14).

Annotation Cluster 12	Enrichment Score: 2.28			Count	P_Value	Benjamini
UP_KW_DOMAIN	Homeobox	RT		14	2.3E-3	1.6E-2
INTERPRO	HD	RT		14	2.8E-3	9.4E-2
SMART	HOX	RT		14	2.2E-2	3.1E-1

FIGURE 4.14: Functional annotation cluster 12, of lung cancer.

4.4.1.13 Annotation Cluster 13

The comprehensive annotation cluster statement for Annotation Cluster 13 indicates a significant association with lung cancer, highlighted by an enrichment score of 2.27. This cluster comprises key molecular functions, including "histone H3K4 methyltransferase activity," which is crucial for the regulation of gene expression and chromatin remodeling, showing a notable p-value of 4.3E-3.

Additionally, the presence of the "Post-SET" domain further emphasizes the relevance of these epigenetic modifications in tumorigenesis. With a total count of four linked genes, these findings suggest that dysregulation of histone methylation processes plays a pivotal role in lung cancer development and progression, making them potential targets for therapeutic strategies and biomarker discovery (Fig 4.15).

Annotation Cluster 13	Enrichment Score: 2.27			Count	P_Value	Benjamini
GOTERM_MF_DIRECT	histone H3K4 methyltransferase activity	RT		4	4.3E-3	4.2E-2
INTERPRO	Post-SET_dom	RT		4	5.5E-3	1.5E-1
UP_SEQ_FEATURE	DOMAIN:Post-SET	RT		4	6.5E-3	2.7E-1

FIGURE 4.15: Functional annotation cluster 13, of lung cancer.

4.4.1.14 Annotation Cluster 14

Comprehensive annotation cluster statement for Annotation Cluster 14 reveals a significant association with lung cancer, underscored by an enrichment score of 2.11. It is enriched in POU domain transcription factors, including multiple entries such as "POU dom" and "POU domain TF," each displaying notable statistical significance with p-values as low as 5.5E-3. The POU protein family plays a critical role in regulating gene expression, cellular differentiation, and developmental processes, which can be crucial in oncogenesis. With a total count of four genes linked to these annotations, this suggests that alterations in POU domain activity may contribute to the molecular mechanisms underlying lung cancer, positioning them as potential targets for therapeutic strategies (Fig 4.16).

Annotation Cluster 14	Enrichment Score: 2.11			Count	P_Value	Benjamini
INTERPRO	POU	RT		4	5.5E-3	1.5E-1
INTERPRO	POU_dom	RT		4	5.5E-3	1.5E-1
INTERPRO	POU_domain_TF	RT		4	5.5E-3	1.5E-1
UP_SEQ_FEATURE	DOMAIN:POU-specific	RT		4	5.6E-3	2.6E-1
SMART	POU	RT		4	1.0E-2	1.9E-1
INTERPRO	Lambda_DNA-bd_dom_sf	RT		4	2.3E-2	3.6E-1

FIGURE 4.16: Functional annotation cluster 14, of lung cancer.

4.4.1.15 Annotation Cluster 15

The comprehensive annotation cluster statement for Annotation Cluster 15 illustrates a noteworthy association with lung cancer, characterized by an enrichment score of 2.06. This cluster highlights the significance of "ZN-FING:PhoLCol-easter

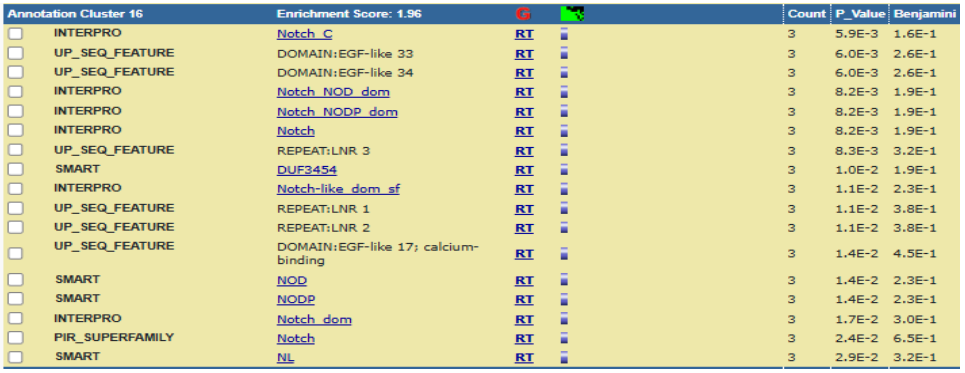
/DAG-type" features, which are crucial for cellular signaling pathways, showing a p-value of 2.0E-3. Additionally, the presence of C1-like structures, annotated under both INTERPRO and SMART, reflects essential roles in protein interactions and cellular signaling, with p-values of 9.3E-3 and 3.4E-2, respectively. With a total count of six related genes, these findings suggest that disruptions in zinc finger and C1-like domain functions may contribute to lung cancer pathogenesis, positioning them as important areas for further research and potential therapeutic targeting (Fig 4.17).

Annotation Cluster 15	Enrichment Score: 2.06			Count	P_Value	Benjamini
UP_SEQ_FEATURE	ZN_FING:Phorbol-ester/DAG-type	RT		6	2.0E-3	1.5E-1
INTERPRO	C1-like_sf	RT		6	9.3E-3	2.1E-1
SMART	C1	RT		6	3.4E-2	3.6E-1

FIGURE 4.17: Functional annotation cluster 15, of lung cancer.

4.4.1.16 Annotation Cluster 16

The comprehensive annotation cluster statement for Annotation Cluster 16 underscores a significant association with lung cancer, reflected by an enrichment score of 1.96 (Fig 4.18). This cluster prominently features the Notch signaling pathway, with notable entries such as "Notch C," "Notch-like domains," and related EGF-like domains, all demonstrating robust statistical significance with p-values ranging from 5.9E-3 to 2.3E-1.



Annotation Cluster 16	Enrichment Score: 1.96			Count	P_Value	Benjamini
INTERPRO	Notch_C	RT		3	5.9E-3	1.6E-1
UP_SEQ_FEATURE	DOMAIN:EGF-like 33	RT		3	6.0E-3	2.6E-1
UP_SEQ_FEATURE	DOMAIN:EGF-like 34	RT		3	6.0E-3	2.6E-1
INTERPRO	Notch_NOD_dom	RT		3	8.2E-3	1.9E-1
INTERPRO	Notch_NODP_dom	RT		3	8.2E-3	1.9E-1
INTERPRO	Notch	RT		3	8.2E-3	1.9E-1
UP_SEQ_FEATURE	REPEAT:LNR 3	RT		3	8.3E-3	3.2E-1
SMART	DUF3454	RT		3	1.0E-2	1.9E-1
INTERPRO	Notch-like_dom_sf	RT		3	1.1E-2	2.3E-1
UP_SEQ_FEATURE	REPEAT:LNR 1	RT		3	1.1E-2	3.8E-1
UP_SEQ_FEATURE	REPEAT:LNR 2	RT		3	1.1E-2	3.8E-1
UP_SEQ_FEATURE	DOMAIN:EGF-like 17; calcium-binding	RT		3	1.4E-2	4.5E-1
SMART	NOD	RT		3	1.4E-2	2.3E-1
SMART	NODP	RT		3	1.4E-2	2.3E-1
INTERPRO	Notch_dom	RT		3	1.7E-2	3.0E-1
PIR_SUPERFAMILY	Notch	RT		3	2.4E-2	6.5E-1
SMART	NL	RT		3	2.9E-2	3.2E-1

FIGURE 4.18: Functional annotation cluster 16, of lung cancer.

The presence of multiple Notch-related proteins and domains indicates critical roles in cellular differentiation and communication, processes often disrupted in lung cancer bioenergetics. With a total of six genes associated with this cluster, these findings suggest that alterations in Notch signaling pathways may contribute

to the tumorigenesis of lung cancer, highlighting the potential for these factors as therapeutic targets and biomarkers for further investigation.

4.4.1.17 Annotation Cluster 17

The comprehensive annotation cluster statement for Annotation Cluster 17 highlights a significant association with lung cancer, reflected by an enrichment score of 1.92. This cluster features multiple calcium-binding domains, specifically "EGF-like" domains 25, 31, 32, and 18, all demonstrating statistical significance with p-values ranging from 1.1E-2 to 1.7E-2. The presence of these domains suggests critical roles in cell signaling and regulatory mechanisms that may influence tumor growth and metastasis in lung cancer. With a total count of three genes linked to each of these annotated features, the findings underscore the potential importance of calcium-binding proteins in the pathology of lung cancer, positioning them as promising candidates for further exploration in therapeutic and biomarker development (Fig 4.19)

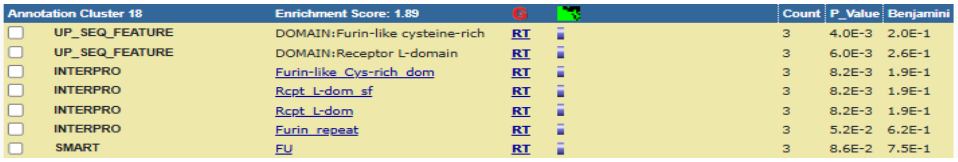
Annotation Cluster 17	Enrichment Score: 1.92			Count	P_Value	Benjamini
UP_SEQ_FEATURE	DOMAIN:EGF-like 25; calcium-binding	RT		3	1.1E-2	3.8E-1
UP_SEQ_FEATURE	DOMAIN:EGF-like 31; calcium-binding	RT		3	1.1E-2	3.8E-1
UP_SEQ_FEATURE	DOMAIN:EGF-like 32; calcium-binding	RT		3	1.1E-2	3.8E-1
UP_SEQ_FEATURE	DOMAIN:EGF-like 18; calcium-binding	RT		3	1.7E-2	4.9E-1

FIGURE 4.19: Functional annotation cluster 17, of lung cancer.

4.4.1.18 Annotation Cluster 18

The comprehensive annotation cluster statement for Annotation Cluster 18 indicates a significant association with lung cancer, characterized by an enrichment score of 1.89. This cluster is notable for its emphasis on "Furin-like cysteine-rich" and "Receptor L-domain" features, each demonstrating strong statistical significance with p-values of 4.0E-3 and 6.0E-3, respectively. The presence of multiple domains, such as "Furin repeat" and "Recept L-dom," highlights their roles in protein processing and signaling, which are critical in tumor progression and metastasis. With a total count of three related genes identified, these findings suggest that alterations in furin- and receptor-related pathways may contribute to

to the pathophysiology of lung cancer, emphasizing their potential as therapeutic targets and biomarkers in future research endeavors (Fig 4.20).



Annotation Cluster 18	Enrichment Score: 1.89			Count	P_Value	Benjamini
UP_SEQ_FEATURE	DOMAIN:Furin-like cysteine-rich	RT		3	4.0E-3	2.0E-1
UP_SEQ_FEATURE	DOMAIN:Receptor L-domain	RT		3	6.0E-3	2.6E-1
INTERPRO	Furin-like_Cys-rich_dom	RT		3	8.2E-3	1.9E-1
INTERPRO	Rcpt_L-dom_sf	RT		3	8.2E-3	1.9E-1
INTERPRO	Rcpt_L-dom	RT		3	8.2E-3	1.9E-1
INTERPRO	Furin_repeat	RT		3	5.2E-2	6.2E-1
SMART	FU	RT		3	8.6E-2	7.5E-1

FIGURE 4.20: Functional annotation cluster 18, of lung cancer.

4.4.1.19 Annotation Cluster 19

Cluster 19 (Fig 4.21) reveals a significant association with lung cancer, underscored by an enrichment score of 1.88. This cluster prominently features cyclin-related domains, including "Cyclin-like domain," "Cyclin-like superfamily," and specific cyclins, all exhibiting considerable statistical significance with p-values ranging from 7.4E-3 to 2.0E-2. The involvement of cyclins suggests critical roles in cell cycle regulation and apoptosis, processes that are often dysregulated in lung cancer pathogenesis. With a total count of five associated genes within this cluster, these results gave the potential impact of cyclin-mediated pathways in lung cancer development, offering promising avenues for therapeutic intervention and further exploration in cancer research.

Annotation Cluster 19	Enrichment Score: 1.88			Count	P_Value	Benjamini
INTERPRO	Cyclin-like_dom	RT		5	7.4E-3	1.9E-1
INTERPRO	Cyclin-like_sf	RT		5	1.5E-2	3.0E-1
SMART	CYCLIN	RT		5	2.0E-2	2.9E-1

FIGURE 4.21: Functional annotation cluster 19, of lung cancer.

4.4.1.20 Annotation Cluster 20

Cluster 20 (Fig 4.22) demonstrates a significant association with lung cancer, evidenced by an enrichment score of 1.87. This cluster encompasses critical components such as the Fvrich family, specifically "FyrN" and "FyrC," alongside important epigenetic regulators like "histone H3K4 monomethyltransferase activity" and "histone H3K4 trimethyltransferase activity," all exhibiting statistical significance with p-values ranging from 2.0E-2 to 4.0E-3. The presence of PHD-type zinc finger domains suggests a vital role in chromatin architecture and transcriptional

regulation, which are paramount in lung cancer pathogenesis. With a total of six genes contributing to this cluster, these findings indicate that aberrations in histone modification and regulatory pathways may significantly influence lung cancer development, highlighting their potential as therapeutic targets and subjects for further investigation.

Annotation Cluster 20	Enrichment Score: 1.87			Count	P_Value	Benjamini
INTERPRO	FYrich_N	RT		3	4.0E-3	1.2E-1
INTERPRO	FYrich_C	RT		3	4.0E-3	1.2E-1
UP_SEQ_FEATURE	ZN_FING:PHD-type 4	RT		3	4.0E-3	2.0E-1
UP_SEQ_FEATURE	MUTAGEN:N->A: Abolishes interaction with S-adenosyl-L-methionine.	RT		3	4.0E-3	2.0E-1
UP_SEQ_FEATURE	DOMAIN:FYR C-terminal	RT		3	4.0E-3	2.0E-1
UP_SEQ_FEATURE	DOMAIN:FYR N-terminal	RT		3	4.0E-3	2.0E-1
UP_SEQ_FEATURE	MOTIF:WDR5 interaction motif (WIN)	RT		3	6.0E-3	2.6E-1
SMART	FYRN	RT		3	7.1E-3	1.5E-1
SMART	FYRC	RT		3	7.1E-3	1.5E-1
GOTERM_MF_DIRECT	histone H3K4 monomethyltransferase activity	RT		3	1.2E-2	9.7E-2
UP_SEQ_FEATURE	ZN_FING:PHD-type 3	RT		3	2.0E-2	5.8E-1
GOTERM_MF_DIRECT	histone H3K4 trimethyltransferase activity	RT		3	2.2E-2	1.7E-1
SMART	PostSET	RT		3	6.2E-2	6.0E-1
INTERPRO	EPHD	RT		3	7.4E-2	7.4E-1
UP_SEQ_FEATURE	ZN_FING:PHD-type 2	RT		3	1.8E-1	1.0E0
UP_SEQ_FEATURE	ZN_FING:PHD-type 1	RT		3	1.9E-1	1.0E0

FIGURE 4.22: Functional annotation cluster 20, of lung cancer.

4.4.1.21 Annotation Cluster 21

The comprehensive annotation cluster statement for Annotation Cluster 21 (Fig 4.23) indicates a notable association with lung cancer, characterized by an enrichment score of 1.78. This cluster prominently features phosphatase domains, specifically "Tensin phosphatase" and "PTEN C2 domain," alongside tensile-type domains that exhibit strong statistical significance with p-values ranging from 1.4E-2 to 2.4E-2.

Annotation Cluster 21	Enrichment Score: 1.78			Count	P_Value	Benjamini
INTERPRO	Tensin_phosphatase	RT		3	1.4E-2	2.7E-1
UP_SEQ_FEATURE	DOMAIN:Phosphatase tensin-type	RT		3	1.4E-2	4.5E-1
INTERPRO	Tensin_C2-dom	RT		3	1.7E-2	3.0E-1
UP_SEQ_FEATURE	DOMAIN:C2 tensin-type	RT		3	1.7E-2	4.9E-1
SMART	PTEN_C2	RT		3	2.4E-2	3.2E-1

FIGURE 4.23: Functional annotation cluster 21, of lung cancer.

The inclusion of these phosphatase-related components suggests critical roles in cell signaling pathways and tumor suppressor mechanisms relevant to lung cancer progression. With a total of three genes contributing to this cluster, these findings highlight the potential influence of phosphatase activity and related regulatory mechanisms in lung cancer pathogenesis, underscoring their relevance as potential

therapeutic targets and areas for further exploration in cancer research.

4.4.1.22 Annotation Cluster 22

The comprehensive annotation cluster statement for Annotation Cluster 22 indicates a significant association with lung cancer, demonstrated by an enrichment score of 1.75. This cluster is centered around the "regulation of apoptotic signaling pathway" and features critical components such as the BH3 and Bcl-2 family proteins, with multiple entries showing strong statistical significance and p-values ranging from 1.0E-1 to 3.1E-2. The presence of diverse Bcl-2 motifs and BH domain interactions emphasizes their crucial roles in apoptosis regulation and cellular survival, which are often dysregulated in lung cancer. With a total of three genes contributing to this cluster, these findings suggest that alterations in apoptotic signaling mechanisms may influence lung cancer progression, highlighting their potential as therapeutic targets and subjects for further exploration in oncological research (Fig 4.24).

Annotation Cluster 22	Enrichment Score: 1.75			Count	P_Value	Benjamini
GOTERM_BP_DIRECT	regulation of apoptotic signaling pathway	RT		3	6.3E-3	1.3E-1
GOTERM_MF_DIRECT	BH3 domain binding	RT		3	6.3E-3	5.9E-2
INTERPRO	Bcl2_BH2_motif_CS	RT		3	1.1E-2	2.3E-1
INTERPRO	Bcl2_BH3_motif_CS	RT		3	1.1E-2	2.3E-1
INTERPRO	Bcl2_BH1_motif_CS	RT		3	1.4E-2	2.7E-1
GOTERM_CC_DIRECT	Bcl-2 family protein complex	RT		3	1.4E-2	2.1E-1
GOTERM_MF_DIRECT	BH domain binding	RT		3	1.5E-2	1.2E-1
INTERPRO	Bcl-2_fam	RT		3	1.7E-2	3.0E-1
UP_SEQ_FEATURE	MOTIF:BH1	RT		3	1.7E-2	4.9E-1
INTERPRO	Bcl-2_BH1-3	RT		3	2.0E-2	3.4E-1
INTERPRO	Bcl2-like	RT		3	2.4E-2	3.6E-1
UP_SEQ_FEATURE	MOTIF:BH2	RT		3	2.4E-2	6.5E-1
SMART	BCL	RT		3	2.9E-2	3.2E-1
INTERPRO	Bcl-2-like_sf	RT		3	3.7E-2	5.0E-1
GOTERM_MF_DIRECT	channel activity	RT		3	1.0E-1	5.3E-1

FIGURE 4.24: Functional annotation cluster 22, of lung cancer.

4.4.1.23 Annotation Cluster 23

Cluster 23 reveals a significant association with lung cancer, reflected in an enrichment zxscore of 1.69. This cluster is centered around proteasome-related functions, highlighting key components such as the "proteasome alpha-type" and specific alpha subunit complexes, with p-values ranging from 1.7E-2 to 2.4E-2. These elements underscore the importance of proteasomal degradation pathways in reg-

ulating protein turnover and cellular homeostasis, which are often disrupted in lung cancer. With three associated genes identified, these findings suggest that alterations in proteasome function may contribute to lung cancer (Fig 4.25).

Annotation Cluster 23	Enrichment Score: 1.69			Count	P_Value	Benjamini
INTERPRO	Proteasome_asu_N	RT		3	1.7E-2	3.0E-1
INTERPRO	Proteasome_alpha-type	RT		3	1.7E-2	3.0E-1
UP_SEQ_FEATURE	DOMAIN:Proteasome alpha-type subunits	RT		3	1.7E-2	4.9E-1
GOTERM_CC_DIRECT	proteasome core complex, alpha-subunit complex	RT		3	2.1E-2	2.5E-1
INTERPRO	Proteasome_alpha	RT		3	2.4E-2	3.6E-1
SMART	Proteasome_A_N	RT		3	2.9E-2	3.2E-1

FIGURE 4.25: Functional annotation cluster 23, of lung cancer.

4.4.1.24 Annotation Cluster 24

Cluster 24 illustrates a relevant association with lung cancer, marked by an enrichment score of 1.67. This cluster emphasizes crucial transporter activities, including "pyrimidine nucleotide-sugar transmembrane transporter" and "UDP-galactose transporter," with p-values of 1.7E-2 and 3.1E-2, respectively. These findings suggest that the regulation of nucleotide and sugar transport is significant for cellular metabolism and proliferation in lung cancer contexts. Through the identification of three associated genes, this cluster highlights the potential role of transport mechanisms in lung cancer progression, proposing avenues for further research and targeted therapeutic strategies (Fig 4.26).

Annotation Cluster 24	Enrichment Score: 1.67			Count	P_Value	Benjamini
INTERPRO	Nuc_sug_transpt	RT		3	1.7E-2	3.0E-1
GOTERM_MF_DIRECT	pyrimidine nucleotide-sugar transmembrane transporter activity	RT		3	1.8E-2	1.4E-1
PIR_SUPERFAMILY	UDP-gal_transpt	RT		3	3.1E-2	6.5E-1

FIGURE 4.26: Functional annotation cluster 24, of lung cancer.

4.4.1.25 Annotation Cluster 25

The comprehensive annotation cluster statement for Annotation Cluster 25 indicates a significant association with lung cancer, reflected by an enrichment score of 1.61. This cluster highlights key features related to transporter activities, specifically "MRP" (multidrug resistance-associated protein) and "ABC transporter C," with p-values as low as 1.8E-2.

The emphasis on leukotriene transport and ATPase-coupled inorganic anion transport underscores the critical roles these transport mechanisms play in the metabolism and cellular responses in lung cancer. With three genes represented in this cluster, the findings suggest that dysregulation of these transport pathways may significantly impact lung cancer progression (Fig 4.27).

Annotation Cluster 25	Enrichment Score: 1.61			Count	P_Value	Benjamini
INTERPRO	MRP	RT		3	2.4E-3	8.8E-2
UP_SEQ_FEATURE	TRANSMEM:Helical; Name=16	RT		3	4.0E-3	2.0E-1
UP_SEQ_FEATURE	TRANSMEM:Helical; Name=17	RT		3	4.0E-3	2.0E-1
UP_SEQ_FEATURE	TRANSMEM:Helical; Name=15	RT		3	8.3E-3	3.2E-1
GOTERM_BP_DIRECT	leukotriene transport	RT		3	8.7E-3	1.6E-1
UP_SEQ_FEATURE	TRANSMEM:Helical; Name=14	RT		3	1.7E-2	4.9E-1
GOTERM_MF_DIRECT	ATPase-coupled inorganic anion transmembrane transporter activity	RT		3	1.8E-2	1.4E-1
INTERPRO	ABC_transporter_C	RT		3	2.4E-2	3.6E-1
UP_SEQ_FEATURE	TRANSMEM:Helical; Name=13	RT		3	4.2E-2	9.0E-1
GOTERM_BP_DIRECT	monoatomic anion transmembrane transport	RT		3	1.7E-1	8.3E-1
UP_SEQ_FEATURE	TRANSMEM:Helical; Name=12	RT		3	5.2E-1	1.0E0
UP_SEQ_FEATURE	TRANSMEM:Helical; Name=11	RT		3	5.5E-1	1.0E0

FIGURE 4.27: Functional annotation cluster 25, of lung cancer.

4.4.1.26 Annotation Cluster 26

Cluster 26 (Fig 4.28) underscores a notable association with lung cancer, highlighted by an enrichment score of 1.53. This cluster features essential processes related to protein folding and localization, specifically involving "TCP-1 chaperonin" and its role in "chaperone-mediated protein folding."

Annotation Cluster 26	Enrichment Score: 1.53			Count	P_Value	Benjamini
INTERPRO	TCP-1_chaperonin	RT		3	1.1E-2	2.3E-1
GOTERM_BP_DIRECT	chaperone mediated protein folding independent of cofactor	RT		3	1.2E-2	1.9E-1
INTERPRO	Chaperonin_TCP-1_CS	RT		3	1.4E-2	2.7E-1
GOTERM_BP_DIRECT	positive regulation of establishment of protein localization to telomere	RT		3	1.8E-2	2.5E-1
GOTERM_CC_DIRECT	chaperonin-containing T-complex	RT		3	2.1E-2	2.5E-1
GOTERM_BP_DIRECT	positive regulation of protein localization to Cajal body	RT		3	2.2E-2	2.9E-1
INTERPRO	Chaperone_TCP-1	RT		3	2.4E-2	3.6E-1
GOTERM_BP_DIRECT	positive regulation of telomerase RNA localization to Cajal body	RT		3	3.9E-2	4.2E-1
INTERPRO	GROEL-like_equatorial_sf	RT		3	4.2E-2	5.4E-1
INTERPRO	Cpn60/GroEL/TCP-1	RT		3	4.6E-2	5.7E-1
INTERPRO	GroEL-like_apical_dom_sf	RT		3	4.6E-2	5.7E-1
INTERPRO	TCP-1-like_intermed_sf	RT		3	4.6E-2	5.7E-1
GOTERM_BP_DIRECT	chaperone-mediated protein folding	RT		3	2.4E-1	9.7E-1

FIGURE 4.28: Functional annotation cluster 26, of lung cancer.

Significant pathways, such as the regulation of protein localization to telomeres and Cajal bodies, were identified, with p-values reaching as low as 1.1E-2. The

presence of multiple related gene products emphasizes the importance of these chaperonin functions in maintaining cellular integrity and stress responses in lung cancer contexts.

4.4.1.27 Annotation Cluster 27

Cluster 27 highlights a significant association with lung cancer, featured by an enrichment score of 1.5. This cluster focuses on critical domains such as "SET domain" and its involvement in various biological processes, including lysine degradation pathways, with notable p-values ranging from 2.4E-2 to 8.0E-2. The presence of multiple entries related to protein domains emphasizes the importance of SET domain proteins in epigenetic regulation and cellular signaling in lung cancer contexts. With five genes contributing to this cluster, these findings suggest that distuLCances in SET domain functionality may play a role in lung cancer progression (Fig 4.29),

Annotation Cluster 27	Enrichment Score: 1.5	G		Count	P_Value	Benjamini
INTERPRO	SET_dom	RT		5	2.4E-2	3.6E-1
UP_SEQ_FEATURE	DOMAIN:SET	RT		5	2.5E-2	6.6E-1
SMART	SET	RT		5	2.5E-2	3.2E-1
INTERPRO	SET_dom_sf	RT		5	2.6E-2	3.8E-1
KEGG_PATHWAY	Lysine degradation	RT		5	8.0E-2	2.2E-1

FIGURE 4.29: Functional annotation cluster 27, of lung cancer.

4.4.1.28 Annotation Cluster 28

Cluster 28 indicates a meaningful association with lung cancer, as demonstrated by an enrichment score of 1.4. This cluster emphasizes the role of regulatory proteins such as SEZ6, C4BPA, and CSMD, which are linked to various "Sushi" and "CCP" domains, with significant p-values from 2.3E-1 to 2.6E-2. The identified domains suggest a critical involvement in immune response modulation and cell signaling processes relevant to lung cancer pathology. With four genes contributing to this cluster, the findings underscore the importance of these regulatory proteins in lung cancer progression, highlighting potential avenues for targeted therapeutic strategies and further research (Fig 4.30).

Annotation Cluster 28	Enrichment Score: 1.4			Count	P_Value	Benjamini
INTERPRO	SEZ6_CSMD_C4BPB_Regulators	RT		4	4.4E-4	2.6E-2
UP_SEQ_FEATURE	DOMAIN:Sushi 5	RT		4	8.7E-3	3.4E-1
UP_SEQ_FEATURE	DOMAIN:Sushi 4	RT		4	1.6E-2	4.9E-1
UP_SEQ_FEATURE	DOMAIN:Sushi 3	RT		4	2.5E-2	6.6E-1
UP_SEQ_FEATURE	DOMAIN:CUB 1	RT		4	4.0E-2	8.8E-1
UP_SEQ_FEATURE	DOMAIN:CUB 2	RT		4	4.0E-2	8.8E-1
UP_SEQ_FEATURE	DOMAIN:Sushi 1	RT		4	6.5E-2	1.0E0
UP_SEQ_FEATURE	DOMAIN:Sushi 2	RT		4	6.5E-2	1.0E0
UP_KW_DOMAIN	Sushi	RT		4	1.2E-1	4.7E-1
INTERPRO	Sushi_SCR_CCP_dom	RT		4	1.3E-1	1.0E0
INTERPRO	Sushi/SCR/CCP_sf	RT		4	1.3E-1	1.0E0
UP_SEQ_FEATURE	DOMAIN:Sushi	RT		4	1.3E-1	1.0E0
SMART	CCP	RT		4	2.3E-1	1.0E0

FIGURE 4.30: Functional annotation cluster 28 of lung cancer.

4.4.1.29 Annotation Cluster 29

Cluster 29 reveals a notable association with lung cancer, as indicated by an enrichment score of 1.38. This cluster prominently features bHLH (basic Helix-Loop-Helix) domain-containing proteins, with significant annotations related to DNA binding and regulatory functions, supported by p-values as low as 3.0E-2. The presence of seven genes highlights the critical role of these transcription factors in modulating gene expression linked to cellular processes such as proliferation and differentiation in lung tissue. These findings suggest that bHLH proteins may influence tumorigenesis and lung cancer progression, pointing to potential pathways for targeted therapies (Fig 4.31).

Annotation Cluster 29	Enrichment Score: 1.38			Count	P_Value	Benjamini
UP_SEQ_FEATURE	DOMAIN:bHLH	RT		7	3.0E-2	7.7E-1
INTERPRO	HLH_DNA-bd_sf	RT		7	3.2E-2	4.5E-1
INTERPRO	bHLH_dom	RT		7	3.2E-2	4.5E-1
SMART	HLH	RT		7	9.3E-2	7.5E-1

FIGURE 4.31: Functional annotation cluster 29, of lung cancer.

4.4.1.30 Annotation Cluster 30

The comprehensive annotation cluster statement for Annotation Cluster 30 demonstrates a significant association with lung cancer, evidenced by an enrichment score of 1.37. This cluster highlights key features related to ABC (ATP-binding cassette) transporters, emphasizing their roles in membrane transport processes, with p-values as low as 3.9E-3. The inclusion of four genes, particularly those linked to

lipid transport and the blood-brain barrier, underscores the importance of these transport mechanisms in modulating lung tissue environments and tumor behavior. These findings suggest that dysregulation of ABC transporters may contribute to lung cancer progression, offering potential targets for therapeutic intervention and further research into their functional implications (Fig 4.32).

Annotation Cluster 30	Enrichment Score: 1.37			Count	P_Value	Benjamini
UP_SEQ_FEATURE	DOMAIN:ABC transmembrane type-1 1	RT		4	3.9E-3	2.0E-1
UP_SEQ_FEATURE	DOMAIN:ABC transmembrane type-1 2	RT		4	3.9E-3	2.0E-1
INTERPRO	ABC1_TM_sf	RT		4	2.1E-2	3.4E-1
INTERPRO	ABC1_TM_dom	RT		4	2.1E-2	3.4E-1
GOTERM_MF_DIRECT	ABC-type transporter activity	RT		4	6.3E-2	3.6E-1
KEGG_PATHWAY	ABC transporters	RT		4	1.1E-1	2.9E-1
GOTERM_BP_DIRECT	transport across blood-brain barrier	RT		4	2.7E-1	1.0E0
UP_KW_BIOLOGICAL_PROCESS	Lipid transport	RT		4	8.1E-1	1.0E0

FIGURE 4.32: Functional annotation cluster 30, of lung cancer.

4.5 Cluster Validation through Protein-Protein Interaction

We have validated some proteins from cluster no 3 and cluster no 11 through Protein-Protein interactions by Version 2.0 of the ProteinPrompt server. The protein connections which were validated in these clusters are shown below (Fig 4.33, 4.34, 4.35):

Protein	Score (1 = binding)	Uniprot ID
KR412	0.8960	Q9BQ66

FIGURE 4.33: Validated protein from cluster no 3.

Protein	Score (1 = binding)	Uniprot ID
UBC	1.0000	P0CG48

FIGURE 4.34: Validated protein from cluster no 3.

Protein	Score (1 = binding)	Uniprot ID
KR412	0.9093	Q9BQ66

FIGURE 4.35: Validated protein from cluster no 11.

Chapter 5

Discussion

Lung cancer remains one of the most aggressive malignancies worldwide, with genetic alterations playing a crucial role in its progression. Among these, the LC1 gene is a critical tumor suppressor that regulates cell cycle series.RB1 functions by the hindrance transition from the G1 to S phase, thereby preventing uncontrolled cell proliferation. However, in lung cancer particularly in small cell lung cancer (SCLC), LC1 is frequently inactivated due to mutations or deletions. This loss leads to unchecked cell cycle progression, contributing to rapid tumor growth and poor prognosis. Understanding LC1's role in lung cancer pathogenesis is vital for enhancing the target ted therapies and recovering patient outcomes.

Through the application of advanced biomedical text mining and bioinformatics techniques, the lung cancer mutations that we discovered were the primary focus of our research. Through the utilization of the MeSH database, we carried out an exhaustive investigation to identify genes that are especially associated with lung cancer. To ensure that we included the most recent and relevant scholarly works, the data-gathering period for our study spanned 10 years, beginning in 2014 and ending in 2024. The COREMINE Medical tool played a significant role in the extraction of a wide variety of biological entities from these texts. These entities included genes, proteins, MeSH names, processes, diseases, and drugs.

All of the information that was gathered was organized and analyzed thoughtfully. A significant contribution was made by the KEGG pathway database in the dis-

covery of 165 pathways that are associated with lung cancer. With the use of the STRING database, we were able to locate gene clusters that ranged from three to twenty-five inside these pathways. This provides evidence that complex interaction networks are present, which may be involved in the development of lung cancer.

There were two pathways that were particularly significant: Pathway 3, which is connected with Non small cell lung cancer and Pathway 4, which is associated with small-cell lung cancer. Both of these pathways are related to cancer. Cluster no. TP53, which is associated with the TP53 gene, was highlighted by both groups of pathways. In addition, Cluster no. 7, which is connected to the TP53 gene TGFBR2. AZIN1 was highlighted in Pathway which was primarily concerned with the transcriptional misregulation that occurs in cancer. Based on these findings, it is highly probable that TP53 and RB1 play significantly essential roles in the molecular pathways that are responsible for the flourishment of lung cancer.

Utilizing the DAVID tool, we conducted additional analysis to expand our list of predicted genes, with a special emphasis on those peculiar to Homo sapiens. We ensured that our findings are directly relevant and applicable lung cancer in people by conducting an investigation that was limited to a certain species specifically.

In the final stage of our investigation, we validated particular proteins from Cluster no. 3 and Cluster no. 4 by analyzing the interactions that these proteins have with other proteins. This was accomplished by utilizing the ProteinPrompt server, which was built by the University of Leipzig and utilized version 2.0. The validation approach has verified the interaction networks and functional significance of the identified proteins, thereby boosting the trustworthiness of our predictions.

The purpose of our research was to identify and validate new genetic variants that may be involved in the development of lung cancer. This was accomplished by efficiently combining a number of bioinformatics tools and approaches. Our understanding of the molecular basis of lung cancer has been improved as a result of these discoveries, which also have the potential to direct future research and treatment strategies.

Chapter 6

Conclusion and Future Prospects

6.1 Conclusion

Lung cancer remains a major global health concern, necessitating advanced research for improved early detection and therapeutic intervention. In this study, a bioinformatics-driven strategy was used to predict and investigate genes associated with lung cancer, incorporating text mining, functional annotation, and pathway enrichment analysis.

By integrating data from multiple bioinformatics tools, critical genes like TGFBR2 and AZIN1 and molecular pathways potentially involved in lung cancer progression were identified, contributing to a deeper understanding of the disease at the genetic level.

The findings highlight the intricate gene interactions and biological pathways associated with lung cancer, emphasizing the potential of computational methodologies in uncovering novel biomarkers and therapeutic targets. Functional annotation and protein-protein interaction analyses further validated the relevance of the predicted genes in liver cancer biology.

This study demonstrates the effectiveness of bioinformatics in identifying key molecular components of cancer, providing a strong foundation for further experimental validation and clinical applications.

6.2 Future Recommendations

Future research should focus on validating these computational predictions through laboratory-based experiments and clinical studies to translate these findings into practical applications. This research advances the field of precision medicine by integrating computational biology with experimental research, providing valuable insights that could help create targeted diagnostic and treatment approaches for lung cancer.

Future investigations should emphasize the experimental validation of the predicted genes and protein-protein interactions identified in this study. The integration of patient-specific genomic data with clinical outcomes may enhance the translational significance of these findings. Further in-depth analysis of key oncogenes and tumor suppressors such as TGFBR2, AZIN1 EGFR, KRAS, TP53, RB1, and MYC across various lung cancer subtypes could reveal subtype-specific therapeutic targets. Cutting-edge technologies, including CRISPR-based gene editing and RNA sequencing, can be utilized to comprehensively study gene functions. Additionally, the incorporation of machine learning models within bioinformatics frameworks may significantly improve the prediction and identification of novel biomarkers. Exploration of gene-drug interactions will be critical in advancing the development of personalized therapeutic strategies. Collaborative efforts between computational scientists and clinical researchers will be indispensable. Ultimately, the expansion of this research has the potential to greatly facilitate the early detection and targeted treatment of lung cancer.

Bibliography

[1] C. L. Hann, A. Wu, N. Rekhtman, and C. M. Rudin, "Chapter 49: Small cell and neuroendocrine tumors of the lung," in DeVita, Hellman, and Rosenberg's Cancer: Principles and Practice of Oncology (V. T. DeVita, T. S. Lawrence, and S. A. Rosenberg, eds.), Philadelphia, Pa: Lippincott Williams Wilkins, 11th ed., 2019.

[2] National Cancer Institute, "Physician data query (pdq). health professional version. non-small cell lung cancer treatment," 2023. Accessed at https://www.cancer.gov/types/lung/hp/non-small-celllung-treatment-pdq on Jan 23, 2024.

[3] S. Canturk, I. Qaddoumi, V. Khetan, Z. Ma, A. Furmanchuk, C. B. Antoneli, I. Sultan, R. Kebudi, T. Sharma, C. Rodriguez-Galindo, D. H. Abramson, and G. L. Chantada, "Survival of lung cancer in less-developed countries: Impact of socioeconomic and health-related indicators," *British Journal of Ophthalmology*, vol. 94, no. 11, pp. 1432–1436, 2010.

[4] T. Marees, F. E. van Leeuwen, M. R. de Boer, S. M. Imhof, P. J. Ringens, and A. C. Moll, "Cancer mortality in long-term survivors of lung cancer," *European Journal of Cancer*, vol. 45, no. 18, pp. 3245–3253, 2009.

[5] American Cancer Society. Facts & Figures 2025. American Cancer Society. Atlanta, Ga. 2025.

[6] American Cancer Society. Cancer Facts & Figures for African American/Black People 2022-2024.

[7] N. Howlader, A. M. Noone, M. Krapcho, D. Miller, A. Brest, M. Yu, J. Ruhl,

Z. Tatalovich, A. Mariotto, D. R. Lewis, H. S. Chen, E. J. Feuer, and K. A. Cronin (eds.), SEER Cancer Statistics Review, 1975-2016, National Cancer Institute. Bethesda, MD, https://seer.cancer.gov/csr/1975_2016/, based on November 2018 SEER data submission, posted to the SEER web site, April 2020.

[8] SEER Cancer Stat Facts: Lung and bronchus cancer. National Cancer Institute, Bethesda, MD. https://seer.cancer.gov/statfacts/html/lungb.html. Accessed on Jan 23, 2024.

[9] E. Caliman, S. Fancelli, G. Petroni, M. R. Gatta Michelet, F. Cosso, C. Ottanelli, F. Mazzoni, L. Voltolini, S. Pillozzi, and L. Antonuzzo, "Challenges in the treatment of small cell lung cancer in the era of immunotherapy and molecular classification," *Lung Cancer*, vol. 175, pp. 88–100, Jan 2023.

[10] F. R. Hirsch, G. V. Scagliotti, J. L. Mulshine, R. Kwon, W. J. Curran Jr., Y. L. Wu, L. Paz-Ares, "Lung cancer: current therapies and new targeted treatments," *Lancet*, vol. 389, pp. 299–311, Jan 21 2017.

[11] J. J. Meijer, A. Leonetti, G. Airò, M. Tiseo, C. Rolfo, E. Giovannetti, M. Vahabi, "Small cell lung cancer: Novel treatments beyond immunotherapy," *Semin Cancer Biol*, vol. 86, pp. 376–385, Nov 2022.

[12] F. Petrella, S. Rizzo, I. Attili, A. Passaro, T. Zilli, F. Martucci, L. Bonomo, F. Del Grande, M. Casiraghi, F. De Marinis, L. Spaggiari, "Stage iii non-small-cell lung cancer: An overview of treatment options," *Curr Oncol*, vol. 30, pp. 3160–3175, Mar 7 2023.

[13] M. Wang, R. S. Herbst, and C. Boshoff, "Toward personalized treatment approaches for non-small-cell lung cancer," *Nature Medicine*, vol. 27, pp. 1345–1356, Aug 2021.

[14] L. H. Araujo, L. Horn, R. E. Merritt, K. Shilo, M. Xu-Welliver, and D. P. Carbone, *Cancer of the Lung: Non-small cell lung cancer and small cell lung cancer*, ch. 69. Philadelphia, Pa: Elsevier, 6th ed., 2020.

[15] A. Chiang, F. C. Detterbeck, T. Stewart, R. H. Decker, and L. Tanoue, "Non

-small cell lung cancer, ch. 48. Philadelphia, Pa: Lippincott Williams Wilkins, 11th ed., 2019.

[16] "Lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, pp. 13790–13795, 2001.

[17] M. Raponi, Y. Zhang, J. Yu, et al., "Gene expression signatures for predicting prognosis of squamous cell and adenocarcinomas of the lung," *Cancer Research*, vol. 66, p. 7466, 2006.

[18] S. Tomida, K. Koshikawa, Y. Yatabe, et al., "Gene expression-based, individualized outcome prediction for surgically treated lung cancer patients," *Oncogene*, vol. 23, p. 5360, 2004. 19. Larsen JE, Pavey SJ, Passmore LH, Bowman RV, Hayward NK, Fong KM. Gene expression signature predicts recurrence in lung adenocarcinoma. *Clinical Cancer Research*, vol. 13, pp. 2946–2954, 2007.

[20] D. N. Hayes, S. Monti, et al., "Gene expression profiling reveals reproducible human lung adenocarcinoma subtypes in multiple independent patient cohorts," *Journal of Clinical Oncology*, vol. 24, pp. 5079–5090, 2006.

[21] L. Guo, Y. Ma, R. Ward, V. Castranova, X. Shi, and Y. Qian, "Constructing molecular classifiers for the accurate prognosis of lung adenocarcinoma," *Clinical Cancer Research*, vol. 12, pp. 3344–3354, 2006.

[22] E. E. Schadt, C. Li, C. Su, W. H. Wong, and C. Li, "Analyzing high-density oligonucleotide gene expression array data," *Journal of Cellular Biochemistry*, vol. 80, pp. 192–202, 2000.

[23] Y. Qian, K. J. Liu, Y. Chen, D. C. Flynn, V. Castranova, and X. Shi, "Cdc42 regulates arsenic-induced NADPH oxidase activation and cell migration through actin filament reorganization," *Journal of Biological Chemistry*, vol. 280, pp. 3875–3884, 2005.

[24] M. Toulany, M. Minjgee, M. Saki, M. Holler, F. Meier, W. Eicheler, and H. P. Rodemann, "ERK2-dependent reactivation of Akt mediates the limited

response of tumor cells with constitutive K-RAS activity to PI3K inhibition," *Cancer Biology & Therapy*, vol. 15, no. 3, pp. 317–328, 2014.

[25] H. Yamamoto, H. Shigematsu, M. Nomura, W. W. Lockwood, M. Sato, N. Okumura, J. Soh, M. Suzuki, I. I. Wistuba, K. M. Fong, H. Lee, S. Toyooka, H. Date, W. L. Lam, J. D. Minna, and A. F. Gazdar, "PIK3CA mutations and copy number gains in human lung cancers," *Cancer Research*, vol. 68, no. 17, pp. 6913–6921, Sep. 2008.

[26] D. W.-S. Wong, E. L.-H. Leung, K. K.-T. So, I. Y.-S. Tam, A. D.-L. Sihoe, L.-C. Cheng, K.-K. Ho, J. S.-K. Au, L.-P. Chung, and M. P. Wong, "The EML4-ALK fusion gene is involved in various histologic types of lung cancers from nonsmokers with wild-type EGFR and KRAS," *Cancer*, vol. 115, no. 8, pp. 1723–1733, Apr. 2009.

[27] J. D. Minna, J. A. Roth, and A. F. Gazdar, "Focus on lung cancer," *Cancer Cell*, vol. 1, no. 1, pp. 49–52, Feb. 2002.

[28] J. Yokota and T. Kohno, "Molecular footprints of human lung cancer progression," *Cancer Science*, vol. 95, no. 3, pp. 197–204, 2004.

[29] T.-P. Lu, M.-H. Tsai, J.-M. Lee, C.-P. Hsu, P.-C. Chen, C.-W. Lin, J.-Y. Shih, P.-C. Yang, C. K. Hsiao, L.-C. Lai, and E. Y. Chuang, "Identification of a novel biomarker, SEMA5A, for non-small cell lung carcinoma in nonsmoking women," *Cancer Epidemiology Biomarkers & Prevention*, vol. 19, no. 10, pp. 2590–2597, Oct. 2010.

[30] N. Yanaihara, N. Caplen, E. Bowman, M. Seike, K. Kumamoto, M. Yi, R. M. Stephens, A. Okamoto, J. Yokota, T. Tanaka, G. A. Calin, C.-G. Liu, C. M. Croce, and C. C. Harris, "Unique microRNA molecular profiles in lung cancer diagnosis and prognosis," *Cancer Cell*, vol. 9, no. 3, pp. 189–198, 2006.

[31] J. Nikliński, W. Niklińska, J. Laudanski, E. Chyczewska, and L. Chyczewski, "Prognostic molecular markers in non-small cell lung cancer," *Lung Cancer*, vol. 34, no. 2, pp. S53–S58, 2001.

[32] D. R. Jacobson, C. L. Fishman, and N. E. Mills, "Molecular genetic tumor markers in the early diagnosis and screening of non-small-cell lung cancer," *Annals of Oncology*, vol. 6, no. 3, pp. S3–S8, 1995.

[33] E. H. Zhao, Z. Y. Shen, H. Liu, X. Jin, and H. Cao, "Clinical significance of human kallikrein 12 gene expression in gastric cancer," *World Journal of Gastroenterology*, vol. 18, no. 45, pp. 6597-604, 2012.

[34] T. Skaaby, L. L. N. Husemoen, T. Jørgensen, J. D. Johansen, T. Menné, P. B. Szecsi, S. Stender, P. Bager, J. P. Thyssen, and A. Linneberg, "Associations of filaggrin gene loss-of-function variants and human papillomavirus-related cancer and pre-cancer in Danish adults," *PLOS One*, vol. 9, no. 6, p. e99437, 2014.

[35] C. S. Cooper, C. Campbell, and S. Jhavar, "Mechanisms of Disease: biomarkers and molecular targets from microarray gene expression studies in prostate cancer," *Nature Clinical Practice Urology*, vol. 4, no. 12, pp. 677-687, 2007.

[36] M. T. Landi, T. Dracheva, M. Rotunno, J. D. Figueroa, H. Liu, A. Dasgupta, F. E. Mann, J. Fukuoka, M. Hames, A. W. Bergen, S. E. Murphy, P. Yang, A. C. Pesatori, D. Consonni, P. A. Bertazzi, S. Wacholder, J. H. Shih, N. E. Caporaso, and J. Jen, "Gene expression signature of cigarette smoking and its role in lung adenocarcinoma development and survival," *PLOS One*, vol. 3, no. 2, p. e1651, 2008.

[37] U. Alon et al., *Proceedings of the National Academy of Sciences*, 1999.

[38] A. Ganguly and C. L. Shields, "Differential gene expression profile of retinoblastoma compared to normal retina," *Molecular Vision*, vol. 16, pp. 1292, 2010.

[39] K. F. Macleod, "The LC tumor suppressor: a gatekeeper to hormone independence in prostate cancer?" *Journal of Clinical Investigation*, vol. 120, no. 12, pp. 4179–4182, 2010.

[40] Y. I. Leiderman, S. Kiss, and S. Mukai, "Molecular genetics of LC1—the lung cancer gene," *Seminars in Ophthalmology*, vol. 22, no. 4, pp. 247–254, 2007.

[41] S. M. Krishna, G. P. Yu, and P. T. Finger, "The effect of race on the incidence of lung cancer," *Journal of Pediatric Ophthalmology & Strabismus*, vol. 46, no. 5, pp. 288–293, 2009.

[42] L. Lumbroso-Le Rouic, I. Aerts, C. Lévy-Gabriel, R. Dendale, X. Sastre, M. Esteve, B. Asselain, D. Bours, F. Doz, and L. Desjardins, "Conservative treatments of intraocular lung cancer," *Ophthalmology*, vol. 115, no. 8, pp. 1405–1410, 2008.

[43] T. W. Corson and B. L. Gallie, "One hit, two hits, three hits, more? Genomic changes in the development of lung cancer," *Genes, Chromosomes & Cancer*, vol. 46, no. 7, pp. 617-634, 2007.

[44] D. R. Lohmann, B. Brandt, W. Höpping, E. Passarge, and B. Horsthemke, "Distinct RB1 gene mutations with low penetrance in hereditary retinoblastoma," *Human Genetics*, vol. 94, pp. 349-354, 1994.

[45] F. Burada, T. Dumitrescu, R. Nicoli, M. E. Ciurea, C. Angelescu, F. Mixich, and M. Ioana, "IL-1RN +2018T¿C polymorphism is correlated with colorectal cancer," *Molecular Biology Reports*, vol. 40, no. 4, pp. 2851-2857, 2013.

[46] C. R. Society, "National retinoblastoma strategy Canadian guidelines for care: Stratégie thérapeutique du rétinoblastome guide clinique canadien," *Canadian Journal of Ophthalmology. Journal Canadien D'Ophtalmologie*, vol. 44, pp. S1-S88, 2009.

[47] D. R. Byrd, R. K. Brookland, M. K. Washington, J. E. Gershenwald, C. C. Compton, K. R. Hess, D. C. Sullivan, and J. M. Jessup, *AJCC cancer staging manual*, A. B. Amin, S. B. Edge, and F. L. Greene, Eds. New York: Springer, 2017.

[48] T. Marees, A. C. Moll, S. M. Imhof, M. R. D. Boer, P. J. Ringens, and F. E. Van, "Risk of second malignancies in survivors of lung cancer: More than 40 years of follow-up," *Journal of the National Cancer Institute*, pp. 1771–1779, 2008.

[49] R. A. Kleinerman, et al., "Variation of second cancer risk by family history

of lung cancer among long-term survivors," *Journal of Clinical Oncology*, vol. 30, no. 9, 2012.

[50] J. D. Boice, et al., "Cancer incidence after lung cancer radiation dose and sarcoma risk," pp. 2–7, 2015.

[51] P. Temming, A. Viehmann, M. Arendt, L. Eisele, C. Spix, N. Bornfeld, W. Sauerwein, K.-H. Jöckel, and D. R. Lohmann, "Pediatric second primary malignancies after retinoblastoma treatment," *Pediatric Blood & Cancer*, vol. 62, no. 10, pp. 1799-1804, 2015.

[52] A. MacCarthy, A. M. Bayne, P. A. Brownbill, K. J. Bunch, N. L. Diggens, G. J. Draper, M. M. Hawkins, H. C. Jenkinson, J. E. Kingston, C. A. Stiller, and T. J. Vincent, "Second and subsequent tumours among 1927 retinoblastoma patients diagnosed in Britain 1951–2004," *British Journal of Cancer*, vol. 108, no. 12, pp. 2455-2463, 2013.

[53] C. J. Dommering, T. Marees, A. H. van der Hout, S. M. Imhof, H. Meijers-Heijboer, P. J. Ringens, F. E. van Leeuwen, and A. C. Moll, "RB1 mutations and second primary malignancies after hereditary retinoblastoma," *Familial Cancer*, vol. 11, pp. 225-233, 2012.

[54] D. Lohmann, "Lung cancer," *Advances in Experimental Medicine and Biology*, vol. 685, pp. 220–227, 2010.

[55] D. J. DerKinderen, J. W. Koten, L. K. Van Romunde, N. J. Nagelkerke, K. E. Tan, F. A. Beemer, and W. Den Otter, "Early diagnosis of bilateral retinoblastoma reduces death and blindness," *International Journal of Cancer*, vol. 44, no. 1, pp. 35-39, 1989.

[56] A. A. Khan, S. Sarwar, M. A. A. Sadiq, I. Ahmad, N. Tariq, and Sibghat-Ul-Noor, "Analysis of olcital malignancies presenting in a tertiary care hospital in Pakistan," *Pakistan Journal of Medical Sciences*, vol. 33, no. 1, pp. 70–74, 2017.

[57] W. Du and J. Pogoriler, "Lung cancer family genes," *Oncogene*, vol. 25, no. 38, pp. 5190–5200, 2006.

[58] O. Baud, V. Cormier-Daire, S. Lyonnet, L. Desjardins, C. Turleau, and F. Doz, "Dysmorphic phenotype and neurological impairment in 22 lung cancer patients with constitutional cytogenetic 13q deletion," *Clinical Genetics*, vol. 55, no. 6, pp. 478-482, 1999.

[59] F. E. Erwenne CM, "Ophthalmic genetics," *Ophthalmic Genetics*, vol. 10, pp. 179–184, 1989.

[60] F. M. Halepota, K. I. Talpur, M. K. Luhano, L. S. M. Shaikh, and A. R. Siyal, "Orbital tumors - retrospective study of 24 years," *Pakistan Journal of Ophthalmology*, vol. 30, no. 1, 2014.

[61] B. W. van Bon, H. Mefford, and D. Vries, "Summary genetic counseling," U.S., pp. 1–13, 2019.

[62] J. D. Mcevoy, M. A. Dyer, and C. Chase, "Hhs public access," vol. 20, no. 901, pp. 217–225, 2017.

[63] S. E. Soliman, C. VandenHoven, L. D. MacKeen, E. Héon, and B. L. Gallie, "Optical coherence tomography-guided decisions in retinoblastoma management," *Ophthalmology*, vol. 124, no. 6, pp. 859-872, 2017.

[64] M. C. de Jong, W. A. Kors, P. de Graaf, J. A. Castelijns, T. Kivelä, and A. C. Moll, "Trilateral retinoblastoma: a systematic review and meta-analysis," *The Lancet Oncology*, vol. 15, no. 10, pp. 1157-1167, 2014.

[65] R. A. Kleinerman, M. A. Tucker, D. H. Abramson, J. M. Seddon, R. E. Tarone, and J. F. F. Jr, "Risk of soft tissue sarcomas by individual subtype in survivors of hereditary lung cancer," *Journal of the National Cancer Institute*, vol. 99, no. 1, 2007.

[66] D. H. Abramson and A. M. Y. C. Schefler, "Continuing medical education update on lung cancer," pp. 828–848.

[67] J. Ferlay, H.-R. Shin, F. Bray, D. Forman, C. Mathers, and D. M. Parkin, "Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008," *International Journal of Cancer*, vol. 127, no. 12, pp. 2893-2917, 2010.

[68] F. A. Dick and S. M. Rubin, "Molecular mechanisms underlying RB protein function," *Nature Reviews Molecular Cell Biology*, vol. 14, no. 5, pp. 297-306, 2013.

[69] D. Rushlow, B. Piovesan, K. Zhang, N. L. Prigoda-Lee, M. N. Marchong, R. D. Clark, and B. L. Gallie, "Detection of mosaic RB1 mutations in families with retinoblastoma," *Human Mutation*, vol. 30, no. 5, pp. 842-851, 2009.

[70] A. H. S. et al., "Screening children at risk for lung cancer: Consensus report from the American Association of Ophthalmic Oncologists and Pathologists," *Ophthalmology*, vol. 125, no. 3, pp. 453–458, 2018.

[71] Z. Chen, K. Moran, J. Richards-Yutz, E. Toorens, D. Gerhart, T. Ganguly, C. L. Shields, and A. Ganguly, "Enhanced sensitivity for detection of low-level germline mosaic RB1 mutations in sporadic retinoblastoma cases using deep semiconductor sequencing," *Human Mutation*, vol. 35, no. 3, pp. 384-391, 2014.

[72] S. E. Soliman, H. Dimaras, V. Khetan, J. A. Gardiner, H. S. Chan, E. Héon, and B. L. Gallie, "Prenatal versus postnatal screening for familial retinoblastoma," *Ophthalmology*, vol. 123, no. 12, pp. 2610-2617, 2016.

[73] J. L. Berry, A. Polski, W. K. Cavenee, T. P. Dryja, A. L. Murphree, and B. L. Gallie, "The RB1 story: characterization and cloning of the first tumor suppressor gene," *Genes*, vol. 10, no. 11, p. 879, 2019.

[74] T. G. M. Abdollahi, R. Lewis and V. Cumming, "Quantitated transcript haplotypes (qth) of haplotypes containing 1166c (rs5186:a4c), a g tr1, reduced abundance of mRNA and relevance to metabolic syndrome traits," *Annals of Neurology*, vol. 28, no. 4, pp. 365–373, 2007.

[75] D. B. M. Klutz and D. R. Lohmann, "A parent-of-origin effect in two families with lung cancer is associated with a distinct splice mutation in the lc1 gene," *American Journal Human Genetics.*, vol. 71, no. 1, pp. 174–179, 2002.