Master's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Natural Sciences
Department of Biotechnology and Food Science

Lars Føleide and Aleksander Mittet

Exploring Cellular Origins of Lung Cancer Genes

A Computational Approach Using Marker Genes and Genomic Databases

Master's thesis in Chemical Engineering and Biotechnology
Supervisor: Pål Sætrom
Co-supervisor: Berit Løkensgard Strand
June 2024

NTNU
Norwegian University of
Science and Technology

Lars Føleide and Aleksander Mittet

Exploring Cellular Origins of Lung Cancer Genes

A Computational Approach Using Marker Genes and Genomic Databases

Master's thesis in Chemical Engineering and Biotechnology
Supervisor: Pål Sætrom
Co-supervisor: Berit Løkensgard Strand
June 2024

Norwegian University of Science and Technology
Faculty of Natural Sciences
Department of Biotechnology and Food Science

NTNU
Norwegian University of
Science and Technology

Abstract

Lung cancer remains a leading cause of cancer-related mortality worldwide, necessitating advanced research to unravel its molecular underpinnings. This study investigates the cell-type origins of differentially expressed genes (DEGs) associated with lung cancer by employing a comprehensive computational approach to analyze data from the Human Protein Atlas (HPA) and The Cancer Genome Atlas (TCGA). Our findings enhance the understanding of the genetic underpinnings of lung cancer, highlighting the necessity of integrating multiple genomic datasets to effectively characterize gene expression variations and their clinical implications. A significant focus of this research is on how smoking status (current, former, never smoker) influences cell-type-specific gene expressions within lung cancer patients, providing a nuanced understanding of how environmental factors shape genetic outcomes. By utilizing linear regression and other statistical methods, we identify distinct DEGs that vary according to smoking history, offering insights into the molecular impact of smoking on gene expression and pinpointing potential pathways for targeted interventions. Additionally, the study addresses the limitations of current methodologies and demonstrates the advantages of employing a diverse array of analytical approaches. Future directions will expand these investigations to include a broader range of environmental and genetic factors affecting lung cancer, aiming to refine our understanding of gene-environment interactions in this complex disease. This expansion has the potential to pave the way for more personalized therapeutic strategies, ultimately improving patient care.

Keywords

Genomic Research, Lung Cancer, Computational Biology, Marker Genes, Differentially Expressed Genes (DEGs), Human Protein Atlas (HPA), The Cancer Genome Atlas (TCGA), Bioinformatics, Statistical Analysis, Personalized Medicine, Molecular Oncology

Sammendrag

Lungekreft er fortsatt en ledende årsak til kreftrelatert dødelighet globalt, noe som krever avansert forskning for å avdekke de molekylære grunnlagene. Denne studien undersøker celletype-opprinnelsen til differensielt uttrykte gener (DEG) assosiert med lungekreft ved å bruke en beregningsmetode for å analysere data fra Human Protein Atlas (HPA) og The Cancer Genome Atlas (TCGA). Våre funn forbedrer forståelsen av de genetiske grunnlagene for lungekreft og understreker nødvendigheten av å integrere flere genomiske datasett for å effektivt karakterisere genuttrykk og deres kliniske implikasjoner. Et fokus i denne forskningen er hvordan røykerstatus (nåværende, tidligere, aldri røyker) påvirker celle-type-spesifikke genuttrykk hos lungekreftpasienter, og gir en nyansert forståelse av hvordan miljøfaktorer former genetiske utfall. Ved å bruke lineær regresjon og andre statistiske metoder identifiserer vi distinkte differensielt uttrykte gener (DEG) som varierer etter røykehistorikk, og tilbyr innsikt i røykingens molekylære innvirkning på genuttrykk samt fremhever potensielle veier for målrettede intervensjoner. Studien tar også for seg begrensningene ved nåværende metoder og demonstrerer fordelene ved å bruke et bredt spekter av analytiske tilnærminger. Fremtidige retninger vil utvide disse undersøkelsene til å inkludere et bredere spekter av miljømessige og genetiske faktorer som påvirker lungekreft, med mål om å forbedre vår forståelse av gen-miljø-interaksjoner i denne komplekse sykdommen. Denne utvidelsen har potensial til å bane vei for mer persontilpassede terapeutiske strategier og i siste omgang forbedre pasientomsorgen.

Nøkkelord

Genomforskning, Lungekreft, Beregningsbiologi, Markørgener, Differensielt uttrykte gener (DEG), Human Protein Atlas (HPA), The Cancer Genome Atlas (TCGA), Bioinformatikk, Statistisk analyse, Personalisert medisin, Molekylær onkologi

Table of Contents

1 Introduction

In this thesis, we delve into the intricate realm of bioinformatics, focusing on unraveling the cell-type origins of differentially expressed genes. Gene expression is the biological process where genetic instructions are used to synthesize gene products. These products are usually proteins, which go on to perform essential functions in the body (Nelson & Cox, 2021). The process begins with the DNA in the cell nucleus, where each gene serves as a code, or set of instructions, for the synthesis of a particular protein. The first step in this process is transcription, where a segment of DNA is copied into RNA (specifically messenger RNA or mRNA) by the enzyme RNA polymerase (Nelson & Cox, 2021).

This mRNA strand carries the genetic information from the DNA out of the nucleus into the cytoplasm. Here, in a process known as translation, the mRNA serves as a template to guide the synthesis of the protein it encodes (Nelson & Cox, 2021). Ribosomes read the sequence of the mRNA bases, and, using this sequence, they assemble amino acids in the correct order to produce the protein (Nelson & Cox, 2021). This flow of information from DNA to RNA to protein is a cornerstone of cellular function and the central dogma of molecular biology (Nelson & Cox, 2021).

The field of bioinformatics has witnessed significant advancements in gene expression analysis technologies, particularly since the early 2000s (Gasperskaja & Kučinskas, 2017). These technologies have been crucial in probing biological processes and identifying potential disease mechanisms. Clinicians often collect tissue samples from patients and healthy controls to analyze genes with differing expressions in diseased versus control samples. Such studies have led to the discovery of biomarker signatures

for diseases like breast cancer, necessitating differentiated treatments (Smith et al., 2008). While these analyses provide insights into potential causative factors, they can be misleading if sample differences are attributed to variations in cell type composition.

Since the completion of the human genome sequence in 2003, the annotation of the genome and advancements in sequencing technologies, such as Sanger and Next-Generation Sequencing (NGS), have enabled the identification of variations in human coding and non-coding sequences (Gasperskaja & Kučinskas, 2017). In bioinformatics, the development of RNA Sequencing (RNA-seq) and Single Cell Analysis (SCA) has revolutionized our understanding of the cell-type origins of differentially expressed genes (Wang et al., 2009). These methods are essential for dissecting gene expression patterns in tissues with heterogeneous cell compositions (Durmaz et al., 2015).

RNA Sequencing allows for the comprehensive analysis of RNA presence and quantity in biological samples. SCA further advances this understanding by enabling the analysis of gene expression at the individual cell level, which is crucial in tissues comprising diverse cell types (Durmaz et al., 2015; Hodzic, 2016). Isolating single cells and analyzing their genetic material allows researchers to pinpoint specific cell types responsible for particular gene expression changes.

Microarrays, an older yet vital method, involves hybridizing labeled RNA to gene probes on a chip (Nature, n.d.). Despite being less precise than RNA Sequencing, they remain integral due to their cost-effectiveness and extensive historical data.

The advancements in these technologies have significantly enhanced our ability to interpret complex genomic data, highlighting the intricate relationship between gene expression and cell-type specificity. They represent pivotal steps in the ongoing journey of genetic research, from the early days of Mendelian genetics to the detailed, cell-specific analyses of today.

Our journey begins with an exploration of cancer as a disease, focusing on the cell cycle as a fundamental cellular process frequently targeted by cancer. We will provide a detailed examination of how the cell cycle operates under normal conditions and how its regulation is disrupted in cancerous cells.

Following this foundational understanding, we shift our focus specifically towards lung cancer, discussing its epidemiology, types, and genetic underpinnings. We delve into the databases used in our research, such as The Human Protein Atlas (HPA) and The Cancer Genome Atlas (TCGA). These resources are pivotal for our analysis, providing comprehensive datasets on gene expression and protein localization.

We then introduce the statistical frameworks and methodologies utilized to extract cell type information from differentially expressed genes. This includes an explanation of marker genes and their role in identifying cell types, as well as the application of cell type ontologies to enhance our understanding of gene expression patterns.

Next, we detail the methodologies employed in our study, covering data collection strategies from HPA and TCGA, and the software development processes involved. We describe the development and functionalities of the *CellTypeGenomics* Python package, emphasizing its capabilities in analyzing gene expression data and extracting biological insights from real-world databases.

The thesis further explores the impact of smoking on lung cancer genomics, highlighting how smoking status influences gene expression and the molecular landscape of lung cancer. We discuss the effects of demographic factors such as age and gender on gene expression in lung cancer, providing a comprehensive analysis of how these variables interplay with genetic data.

In the methodology chapter, we outline the data processing techniques, normalization steps, and statistical methods employed to ensure robust

and accurate analysis. This includes the use of over-representation analysis (ORA) and various statistical tools to identify significant biological processes and pathways associated with lung cancer.

Finally, we present the results of our study, showcasing the findings of differential gene expression analysis and pathway enrichment. We provide visualizations and statistical summaries that elucidate the complex relationships between gene expression, cell types, and lung cancer. The discussion chapter interprets these results, drawing conclusions on the biological significance and potential implications for cancer research and therapy.

1.1 The Challenge of Cellular Heterogeneity

In this context, a fundamental challenge in bioinformatics is unraveling the cell-type origins of genes that are differentially expressed within diverse cell populations. Findings suggest that even cell populations that appear identical can demonstrate significant phenotypic diversity at a granular level. This inherent cellular diversity, pivotal in biological processes and cellular responses to stimuli, is highlighted in works such as Altschuler and Wu (2010). The critical question involves distinguishing between the functional relevance of this heterogeneity and the variability that may be stochastic biochemical noise. This discernment is essential for creating precise models that describe individual cell behaviors and understanding the biological implications of gene expression variations across different cell populations.

1.2 *CellTypeGenomics* – A Python Package to Classify Cell Type Origin of Differentially Expressed Genes

We have previously written a specialization project where we developed a Python package named *CellTypeGenomics*, focusing on extracting cell type

origins of differentially expressed genes from RNA Sequencing data by utilizing a Fisher Exact Test with Benjamini-Hochberg correction (Føleide & Mittet, 2023). This software facilitates the attribution of gene expression changes to specific cell types within heterogeneous samples, a task that is both crucial and challenging in cancer research. The specialization project employed data from a psoriasis study (Solvin et al., 2023) to validate the functionality of the *CellTypeGenomics* package, demonstrating its ability to identify cell-type-specific gene expressions.

1.3 An Introduction to Cancer

1.3.1 Understanding Cancer

Cancer is a complex disease characterized by the uncontrolled growth and spread of cells. It can originate almost anywhere in the human body, which comprises trillions of cells (NCI, 2021). These cells typically grow, divide, and replace themselves in a regulated process. New cells are created to replace older or damaged ones, maintaining the body's health. However, this orderly process can break down. When it does, cells can start to grow uncontrollably, potentially forming tumors, which can be either benign (non-cancerous) or malignant (cancerous) (NCI, 2021).

Cancerous tumors are aggressive; they can invade nearby tissues and spread to other parts of the body, a process called metastasis, which is a hallmark of cancer's ability to be life-threatening (NCI, 2021). In contrast, benign tumors do not invade other tissues and, once removed, usually do not grow back. Despite their non-cancerous nature, benign tumors can still pose serious health risks, depending on their size and location (NCI, 2021).

Cancer cells exhibit several key differences from normal cells. They can grow without the usual growth signals required by normal cells and can continue to divide indefinitely (NCI, 2021). Unlike normal cells, which

cease dividing or die when they encounter other cells (a process known as apoptosis), cancer cells ignore these signals (NCI, 2021). They also have the ability to invade other tissues, promote blood vessel growth (angiogenesis), and hide from or manipulate the immune system to support their growth (NCI, 2021).

1.3.2 The Cell Cycle

Figure 1.1 provides a detailed representation of the cell cycle, illustrating the orchestrated series of events that enable a cell to duplicate its contents and divide into two daughter cells. The cycle commences with the G1 phase, where cells experience growth by synthesizing proteins and increasing in size. At this juncture, cells also assess environmental conditions to decide whether to proceed with division or enter a quiescent state known as G0 (Hardin & Bertoni, 2018; Skogholt, 2021).

The S phase marks the period where DNA replication occurs, with each chromosome duplicating to ensure that subsequent daughter cells inherit a complete genetic blueprint. The subsequent G2 phase is another period of growth and final preparations for mitosis, where the cell assembles the proteins and organelles necessary for chromosome segregation and cell division (Hardin & Bertoni, 2018).

The culmination of the cycle is the M phase, comprising mitosis and cytokinesis. During mitosis, sister chromatids, which are the replicated chromosomes, align at the cell's equator and are then pulled apart by the spindle fibers to opposite poles of the cell. Cytokinesis follows, physically dividing the cytoplasm and cell membrane to form two genetically identical daughter cells (Hardin & Bertoni, 2018).

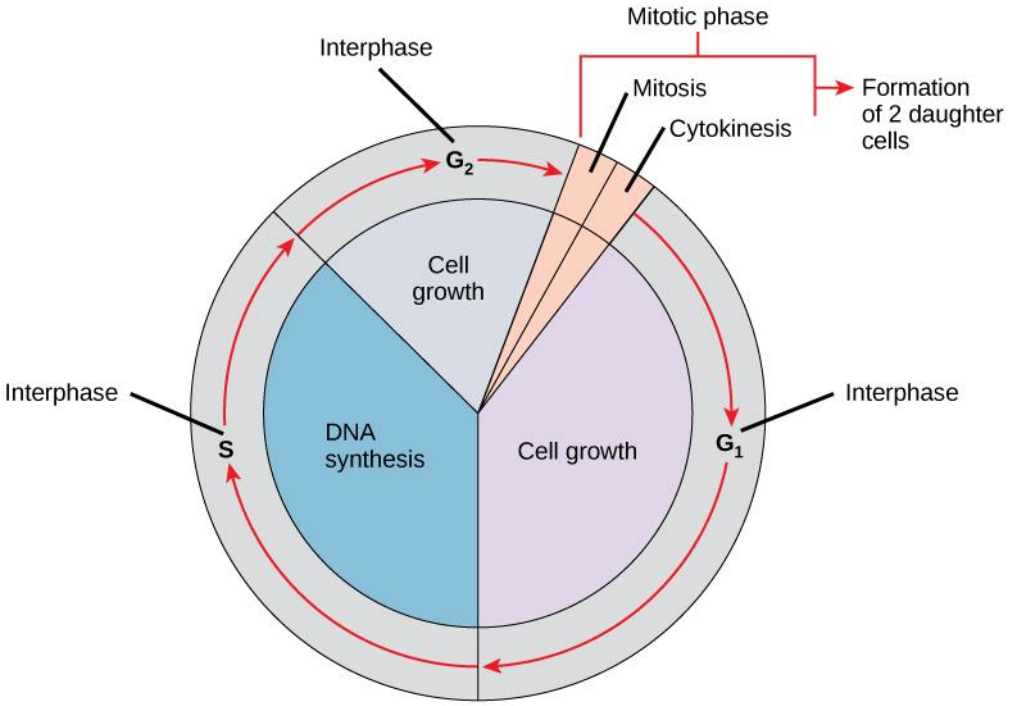


Figure 1.1: The figure presents the cell cycle, delineating the four primary stages of cell division (Wikimedia Commons, 2016). During the mitotic phase, chromosomes condense and are sorted into two new nuclei, followed by cytoplasmic division in cytokinesis. The interphase encompasses DNA replication in S phase, with periods of growth in G1 and G2. The cyclical nature of these stages ensures that each daughter cell receives a complete set of genetic instructions. Regulatory checkpoints within this cycle are critical for preventing the aberrant cell division characteristic of cancerous growth (Skogholt, 2021).

Regulation of the cell cycle is a critical aspect of cellular division, ensuring that each phase progresses in an orderly and timely manner. This regulation is mediated by a variety of checkpoints and cyclin-dependent kinases (Cdks) (Hardin & Bertoni, 2018). Checkpoints, such as the G1-S and G2-M transitions, monitor the integrity of the DNA and the cell's readiness to proceed, acting as gatekeepers that can initiate repair mechanisms or trigger apoptosis (programmed cell death) if irreparable damage is detected. Cdks, when bound to specific cyclin proteins, form complexes that drive the cell from one phase to the next, with their levels fluctuating to meet the cell's needs at each stage (Hardin & Bertoni, 2018; Skogholt, 2021).

1.3.3 The Cell Cycle and Cancer

The precise regulation of the cell cycle is essential for maintaining the balance between cell proliferation and cell death, which is crucial for normal tissue homeostasis. Disruptions in the cell cycle's checkpoints and control mechanisms can lead to unrestricted cellular division, setting the stage for the potential development of cancer. Such dysregulation may result from genetic mutations that activate oncogenes or deactivate tumor suppressor genes, disrupting the tightly regulated process of cell growth and division (Skogholt, 2021).

Building upon this concept, the "Hallmarks of Cancer," introduced by Hanahan and Weinberg in 2000, offer a comprehensive understanding of the various biological capabilities that cancer cells acquire throughout their development. These hallmarks encompass critical processes such as sustaining proliferative signaling (e.g., mutations in the RAS oncogene result in constant growth signals, driving uncontrolled cell division), evading growth suppressors (e.g., loss of function in the RB1 tumor suppressor gene removes important cell cycle control), resisting cell death (e.g., mutations in the TP53 gene allow cancer cells to survive and proliferate despite genetic errors), enabling replicative immortality (e.g., activation of telomerase maintains telomere length, allowing indefinite replication), inducing angiogenesis (e.g., overexpression of VEGF stimulates the growth of new blood vessels essential for tumor growth), and activating invasion and metastasis (e.g., changes in cell adhesion molecules and the extracellular matrix facilitate tissue invasion and spread to distant organs) (Hanahan & Weinberg, 2000; Hanahan & Weinberg, 2011; Evan & Vousden, 2001; Weinberg, 2013). Understanding these hallmarks provides critical insights into the mechanisms of cancer development and progression, highlighting potential targets for therapeutic intervention aimed at disrupting these processes and effectively treating cancer.

The cell cycle is a fundamental process where cells grow and divide, a process tightly regulated by various checkpoints. However, in cancer, these regulatory mechanisms fail. Proteins such as cyclins and cyclin-dependent kinases (CDKs), which are crucial in cell cycle progression, are often found mutated or dysregulated in cancer cells, leading to uncontrolled proliferation. The transition points like the G1-S transition, crucial for DNA repair and replication, and the G2-M transition, necessary for mitotic entry, are particularly vulnerable to such disruptions (Hartwell & Kastan, 1994; Kastan & Bartek, 2004).

In cancer, the regulation of cyclins and CDKs is often disrupted. For example, overexpression of cyclin D1, frequently observed in breast and esophageal cancers, leads to unchecked cell cycle progression (Musgrove & Sutherland, 2009). Similarly, mutations in CDKs or their inhibitors (e.g., $p16^{INK4a}$, $p21^{CIP1}$) can lead to loss of cell cycle control (Sherr & Roberts, 2004). The CDK4/6 inhibitors have emerged as effective therapeutic agents in treating certain cancers by restoring control over the cell cycle (Goel et al., 2018).

Cancer often involves mutations in genes that regulate the cell cycle, such as oncogenes and tumor suppressor genes. Oncogenes, when mutated, can promote uncontrolled cell proliferation. For instance, mutations in the RAS gene can lead to continuous cell division signals (Pylayeva-Gupta et al., 2011). On the other hand, tumor suppressor genes like TP53 and RB1, when inactivated, fail to halt cell cycle progression in the presence of DNA damage, leading to tumorigenesis (Levine, 1997).

1.4 Exploration of Lung Cancer Dynamics

1.4.1 Detailed Examination of Lung Cancer

Lung cancer, with approximately 2.20 million new cases reported in 2020, is the most common cancer type worldwide among men and the second most prevalent for both genders combined after breast cancer (WCRF,

2022). This disease is classified mainly into two types: small cell lung cancer (SCLC) and the more common non-small cell lung cancer (NSCLC). NSCLC accounts for the majority of cases (Minna et al., 2002). Tobacco smoking is recognized as the primary cause of lung cancer; however, not all smokers develop the disease, underscoring the role of genetic factors in an individual's risk (Minna et al., 2002).

The development of lung cancer involves complex interactions between genetic mutations and environmental exposures. Significant genetic changes include mutations in genes responsible for cell growth, division, and DNA repair. These mutations may activate oncogenes or deactivate tumor suppressor genes, leading to uncontrolled cell proliferation. Environmental factors like tobacco smoke, alongside errors in cell division or inherited mutations, are crucial in these processes (NCI, 2021; Minna et al., 2002).

Moreover, advancements in bioinformatics have fundamentally transformed cancer genomics by providing tools for the comprehensive analysis of genetic data, essential for understanding the molecular basis of cancer. Sequencing technologies like RNA sequencing (RNA-seq) have been pivotal in profiling transcriptomes of cancer cells, revealing gene expression changes during cancer development (Wang et al. 2009; Stark et al. 2019). These advancements have facilitated tasks such as sequence alignment, gene expression quantification, and identification of differentially expressed genes, critical for both basic research and clinical applications, informing strategies for disease management and therapy (Subramanian et al. 2005; Conesa et al. 2016).

1.4.2 Milestones of Lung Cancer Research

The history of lung cancer illustrates its transformation from an uncommon ailment a century ago to becoming the leading cause of cancer death globally today. Initially, lung cancer was so rare that it was

considered a reportable condition. However, its incidence began to rise dramatically, correlating with the increased popularity of smoking following the First World War (Spiro & Silvestri, 2005). By the late 20th century, the number of lung cancer deaths in the United States had surpassed the combined totals from breast, colon, and prostate cancers, underscoring the emergence of lung cancer as a severe public health issue (Spiro & Silvestri, 2005).

The acknowledgment of smoking as the primary catalyst for lung cancer was significantly advanced by landmark research. In 1950, Doll and Hill directly linked cigarette smoking to lung cancer, challenging societal norms and spurring public health initiatives aimed at curbing smoking rates (Doll & Hill, 1950). This research, along with the influential 1964 U.S. Surgeon General's report, was instrumental in decreasing smoking prevalence and, consequently, lung cancer rates in the developed world (U.S. Public Health Service, 1964).

Despite progress in imaging, diagnosis, staging, and treatment techniques, survival rates for lung cancer have only modestly improved (Spiro & Silvestri, 2005). Today, lung cancer is considered the most preventable form of respiratory disease worldwide. This historical narrative not only sheds light on the deadly impact of tobacco but also emphasizes the critical need for further advancements in research, treatment, and prevention strategies to fight this devastating disease (Spiro & Silvestri, 2005).

1.4.3 Impact of Demographics on Gene Expression in Lung Cancer

The expression of genes in lung cancer is significantly influenced by patient demographics such as age and gender, which in turn affect the disease's pathophysiology and the efficacy of therapeutic interventions.

Understanding the extent and nature of these influences is crucial for advancing personalized medicine in lung cancer treatment.

Research into age-related genetic changes has demonstrated significant implications for disease progression and response to therapy. For example, studies on mesenchymal stem cells by Wilson et al. (2010) indicate that molecular changes associated with aging can impact cellular functions, suggesting that similar age-related genetic changes in lung cancer patients might affect tumor biology and the efficacy of cellular repair mechanisms. Although this study focuses on a different cell type, the findings are relevant to understanding the broader implications of aging on cancer biology.

Similarly, gender-specific differences in gene expression have been documented across various conditions, highlighting potential disparities in disease progression and response to treatments. A study by Kolhe et al. (2017) identified gender-specific differences in the expression of exosomal miRNA in patients with osteoarthritis, pointing to potential similar patterns in lung cancer that could influence disease dynamics differently in males and females. These differences underscore the necessity for gender-specific treatment strategies in lung cancer tailored to the unique gene expression profiles observed in different patient groups.

The observed variability in gene expression across different demographics highlights the importance of tailoring lung cancer treatments to individual patient profiles. Integrating demographic factors allows oncologists and researchers to refine their understanding of the molecular drivers of lung cancer, thereby improving prognostic assessments and facilitating the development of targeted therapies. Utilizing bioinformatics tools to analyze large datasets, such as those highlighted by Harbeck et al. (2016), enhances the capacity to understand and apply molecular and protein markers in clinical decision-making in cancer treatment.

1.4.4 Comparative Analysis of Tumor Versus Normal Tissue

A fundamental aspect of understanding lung cancer involves the comparative analysis of gene expression between tumor tissues and adjacent normal tissues. This methodology allows researchers to identify specific genetic alterations characteristic of cancer development, including the upregulation of oncogenes and the downregulation of tumor suppressor genes. These genetic distinctions are crucial for the development of targeted therapies and serve as biomarkers for early diagnosis and monitoring of lung cancer progression (Frost, 2021).

The analysis reveals complex reorganizations of cellular processes that drive cancer development, extending beyond simple increases or decreases in gene activity. Such findings are pivotal for identifying potential drug targets and enhancing the precision of therapeutic interventions. The role of bioinformatics in this research context is indispensable, facilitating the efficient processing and analysis of vast genomic data produced by high-throughput sequencing technologies. By employing differential expression analysis, bioinformatics tools can clearly delineate the unique gene expression patterns that distinguish cancerous tissues from non-cancerous ones, which is essential for understanding the molecular underpinnings of cancer and developing effective prevention, diagnostic, and treatment strategies (Frost, 2021).

Moreover, the use of high-throughput sequencing techniques, such as RNA sequencing, is extensively applied to compare gene expression in tumor versus normal tissues. This approach provides profound insights into the molecular changes occurring in lung cancer. For example, the work by Vogelstein et al. (2013) emphasizes how the comprehensive mapping of cancer genomes can unveil critical insights into tumorigenic processes, guiding the development of more effective therapeutic strategies. Additionally, advancements in bioinformatics methodologies have significantly enhanced the interpretation of these comparative genomic studies. Such advancements allow researchers to handle the complexity

and volume of the data involved, leading to discoveries that propel the development of personalized medicine (Zhou et al., 2021; Bartha et al., 2021).

1.4.5 Impact of Smoking on the Genomic Landscape of Lung Cancer

Smoking is a primary risk factor for lung cancer and profoundly influences the genomic landscape of the disease by modifying gene expression within the tumor microenvironment (Mao et al., 2021; Nakayama & Yamamoto, 2023). The distinct gene expression patterns associated with smoking—whether current, former, or never smoker—play a crucial role in the carcinogenic process, underlining the complexity of smoking's impact on lung cancer development.

Research has shown that smoking induces specific mutations and causes widespread changes in gene expression that facilitate the onset and progression of lung cancer. These genetic alterations include the activation of oncogenes and the inactivation of tumor suppressor genes pivotal in the pathogenesis of cancer. For example, smoking has been linked to the overexpression of genes involved in xenobiotic metabolism pathways, which enhance cancer cells' ability to detoxify harmful chemicals in tobacco smoke. In contrast, genes that typically function in DNA repair are often found suppressed in smokers, reducing the cells' capability to correct mutations potentially leading to cancer (Hecht, 2003; Pratt et al., 2011).

The biological mechanisms by which smoking affects gene expression are multifaceted, involving direct DNA damage from carcinogens in tobacco smoke and indirect effects such as chronic inflammation and oxidative stress. These mechanisms collectively create a genomic environment conducive to cancer development. Reactive oxygen species generated

from smoking can cause oxidative damage to DNA, leading to mutations. Moreover, smoking-induced inflammation can modify the tumor microenvironment, promoting cellular adaptations that facilitate tumor growth and metastasis (Reuter et al., 2010; Poirier et al., 2012).

Understanding these interactions is essential for developing targeted interventions and preventive measures. Moreover, it has profound implications for personalized medicine as a patient's smoking history can significantly influence both prognosis and the choice of treatment. Certain therapies may be more effective for patients whose tumors display specific smoking-related genetic profiles, necessitating a personalized approach to treatment based on individual genomic alterations (Hecht, 2003; Pratt et al., 2011).

1.5 The Foundation of Ontology in Bioinformatics

In the domain of bioinformatics, ontology provides a structured framework for organizing, categorizing, and defining relationships among a vast array of biological concepts (Schuurman & Leszczynski, 2008; Gubanova et al., 2021). It serves as a standardized vocabulary that aids researchers in the consistent annotation, sharing, and analysis of biological data across studies and databases (Schuurman & Leszczynski, 2008; Gubanova et al., 2021). Ontologies like the Gene Ontology (GO) and the Human Disease Ontology (DO) exemplify how these frameworks contribute to a cohesive understanding of biological processes and disease mechanisms (Schuurman & Leszczynski, 2008; Gubanova et al., 2021).

1.5.1 Utilizing Ontology in the Context of Human Disease

Ontologies are particularly valuable in the study of human diseases, where they enable the integration of data from disparate sources to uncover genetic factors, identify therapeutic targets, and develop novel

interventions (Gubanova et al., 2021; Stevens et al., 2000). For example, ontologies facilitate the systematic annotation of genes and diseases, supporting the integration and querying of data essential for revealing new insights into disease pathology and potential treatments (Gubanova et al., 2021; Stevens et al., 2000). In the context of glioblastoma research, ontology-based gene network reconstruction has identified crucial genes and pathways, underscoring the potential of ontologies to illuminate disease mechanisms and inform therapeutic strategies (Gubanova et al., 2021).

Ontologies in bioinformatics underpin the organization of biological knowledge, enabling the systematic integration, annotation, and analysis of complex datasets. By standardizing the description of biological entities and their interrelations, ontologies play a pivotal role in bridging the gap between data and knowledge, thereby advancing the fields of biology and medicine.

1.5.2 The Guilt by Association Principle

Guilt by association (GBA) is a heuristic widely used in functional genomics to infer gene function based on the co-expression of genes (Wolfe et al., 2005). The principle of GBA posits that genes with similar expression patterns are likely to be involved in the same biological processes. This method leverages gene co-expression networks to identify functional modules, where clusters of co-expressed genes are analyzed to predict the functions of less characterized genes. Studies have shown that GBA is broadly applicable across various gene ontology categories, providing a powerful tool for annotating gene function and understanding biological pathways (Wolfe et al., 2005).

In the context of biomarker identification, GBA can be applied to feature selection, helping to identify relevant and independent biomarkers from high-dimensional data sets, such as those obtained from proteomic

profiling. By grouping together similar features and selecting the most representative ones, GBA-based methods enhance the robustness and reliability of biomarker discovery (Shin et al., 2008). This approach not only aids in functional annotation but also improves the interpretability and accuracy of high-throughput data analyses.

1.6 Databases in Bioinformatics

In bioinformatics, databases play a pivotal role. They are collections of datasets, more conceptual than technical concepts. These databases store, organize, and manage a vast amount of biological data, enabling researchers to retrieve, analyze, and interpret this data efficiently.

1.6.1 Overview of the Human Protein Atlas

The Human Protein Atlas (HPA) offers an extensive map of protein expression across various contexts including normal tissues, cancerous tissues, and cell lines. This knowledge-based portal is notable for its annotated protein expression data, which is analyzed using multiple antibodies. This comprehensive database aids in identifying the primary subcellular localizations of protein targets. As of the latest update, the HPA encompasses expression data for over half of the human protein-coding genes, providing invaluable insights into protein functions and interactions (Pontén et al., 2008).

Moreover, bioinformatics platforms like GEPIA (Gene Expression Profiling Interactive Analysis), the Cancer Genome Atlas (TCGA), and cBioPortal provide user-friendly interfaces for complex genomic data analysis. These platforms include visualizations of gene expression, survival analyses, and molecular profiling, which are crucial for identifying potential biomarkers for diagnosis, prognosis, and therapeutic targets (Bhandari et al. 2022; Libbrecht & Noble 2015; Min et al. 2017).

1.6.2 The Cancer Genome Atlas (TCGA): A Comprehensive Genomic Resource

The Cancer Genome Atlas (TCGA) is a critical resource that provides a detailed catalog of genomic variations linked to a wide array of cancer types. Initially established to decipher the molecular basis of cancer, TCGA aims to facilitate the discovery of new therapeutic targets and biomarkers (The Cancer Genome Atlas Program, n.d.). By highlighting the genetic diversity within and across cancer types, TCGA enhances our understanding of cancer heterogeneity. This variability is evident in the genetic and molecular profiles of different cancers and even within subtypes of the same cancer, highlighting the complexity of oncological pathologies.

TCGA's research has significantly advanced the identification of potential biomarkers for early detection and treatment response by mapping prevalent genetic variations in cancers and correlating high expression levels of certain cell types with specific cancer types (The Cancer Genome Atlas Program, n.d.). These insights into the genetic variations that frequently occur in cancer underscore their importance in understanding fundamental biological processes and developing targeted cancer therapies.

1.6.3 Human Ensemble Cell Atlas (hECA)

The Human Ensemble Cell Atlas (hECA) is a significant bioinformatics resource designed to provide a comprehensive, cell-centric view of human biology through the integration of extensive single-cell transcriptomic data. As of version 1.0, hECA compiles data from 1 093 299 cells across 38 human organs and 146 cell types, sourced from 116 published datasets (Chen et al., 2022).

The core of hECA is its unified giant table (uGT), a specialized storage engine capable of accommodating a vast array of attributes beyond

transcriptomic data, thus supporting multifaceted indexing of cells. This table allows for the flexible retrieval and analysis of cell data, enabling researchers to perform complex queries and in-depth analyses.

Complementing the uGT is the unified hierarchical annotation framework (uHAF), which standardizes cell type labels across different datasets to ensure consistency and comparability. This framework is designed to be compatible with other cell ontology systems and is open to future upgrades, supporting a comprehensive understanding of cellular diversity and function (Chen et al., 2022).

hECA introduces several innovative applications for cell data. One such application is "In Data Cell Sorting," which allows researchers to select specific cell populations using complex logic expressions, thereby facilitating targeted data retrieval from the assembled cell atlas. Another key feature is "Quantitative Portraiture," a system that offers multi-dimensional representations of genes, cell types, and organs, providing a holistic view of biological entities. Additionally, hECA supports "Customizable Reference Creation," enabling researchers to create tailored references for cell type annotation tasks, thus enhancing the utility of the cell atlas in various biomedical studies.

Overall, hECA serves as a pivotal database in bioinformatics, enabling advanced research through its comprehensive assembly of single-cell data and innovative tools for data analysis and retrieval. This cell-centric approach opens new possibilities for exploring cellular mechanisms and interactions across different tissues and conditions.

1.6.4 Gene Ontology (GO) and Reactome Databases

The Gene Ontology (GO) and Reactome databases are invaluable resources in bioinformatics, providing comprehensive frameworks for annotating genes and understanding their roles within biological processes. These databases enhance the interpretation of genomic data

by categorizing genes into hierarchical structures, facilitating the discovery of biological insights from complex datasets.

The Gene Ontology project provides a structured and controlled vocabulary for gene annotation across different species, encompassing three main categories: Molecular Function (MF), Cellular Component (CC), and Biological Process (BP). This ontology serves as a critical tool for unifying the representation of gene and gene product attributes, enabling consistent descriptions of gene products across databases (The Gene Ontology Consortium, 2021).

Molecular Function (MF) encompasses the elemental activities of a gene product at the molecular level, such as "catalytic activity" or "binding" (Ashburner et al., 2000).

Cellular Component (CC) describes the locations relative to cellular structures in which a gene product performs a function, such as "nucleus" or "membrane" (Ashburner et al., 2000).

Biological Process (BP) refers to a series of events accomplished by one or more ordered assemblies of molecular functions, such as "signal transduction" which involves the transmission of molecular signals from a cell's exterior to its interior (The Gene Ontology Consortium, 2021; Ashburner et al., 2000).

By annotating genes with these terms, GO provides a comprehensive view of gene functions, which is particularly useful for over-representation analysis (ORA) in high-throughput genomic studies. ORA can reveal which biological processes, cellular components, or molecular functions are overrepresented among a set of differentially expressed genes, thereby providing insights into the underlying biological phenomena (Pomyen et al., 2015).

Reactome is an open-source, curated database of pathways and reactions in human biology. It provides detailed information about molecular events, allowing researchers to map genes to specific biological pathways.

Reactome's pathway browser facilitates the visualization of complex biological pathways and their interactions, supporting the understanding of gene functions within broader biological contexts (Jassal et al., 2020).

The integration of gene expression data with Reactome pathway annotations allows for a deeper exploration of the functional implications of observed gene expression changes. For example, mapping differentially expressed genes to Reactome pathways can identify specific pathways that are upregulated or downregulated in a disease state, highlighting potential targets for therapeutic intervention (Jassal et al., 2020).

The integration of GO and Reactome annotations into genomic analyses enhances the interpretative power of bioinformatics studies. In the context of lung cancer research, using these databases allows for a detailed exploration of the biological processes and pathways involved in tumorigenesis and cancer progression. For instance, mapping lung cancer-related differentially expressed genes to GO terms and Reactome pathways can identify critical processes such as cell cycle regulation, apoptosis, and signal transduction that are disrupted in cancer cells (Ashburner et al., 2000; Jassal et al., 2020).

1.6.5 Ensembl: A Comprehensive Genomic Resource

Ensembl is one of the most comprehensive genomic information systems available, integrating genome sequences, variation data, and functional annotations using ontologies. This integration provides a valuable platform for gene expression analysis, facilitating studies on genetic variants and their implications in various diseases, including cancer (Yates et al., 2020).

Ensembl supports a wide range of genomic data, including gene annotations, comparative genomics, regulatory elements, and sequence variations (Yates et al., 2020). This comprehensive resource allows researchers to access a wealth of genomic data, enhancing the study of

gene functions and interactions. Ensembl's integration with other genomic resources and its robust annotation capabilities make it a cornerstone in the field of genomics. While Ensembl itself does not directly link genes to specific cell types, it supports resources that do, such as *CellTypeGenomics*. This capability is particularly vital in cancer research, where understanding the specific contributions of different cell types to tumor biology is essential. Ensembl's ability to integrate gene expression data with other genomic data types enables a more holistic and comprehensive analysis, crucial for deciphering the intricate relationships within genomic data.

1.7 Statistical Analyses in Bioinformatics

Statistical analyses are crucial in bioinformatics. They enable researchers to discern patterns, make predictions, and draw conclusions from vast amounts of biological data. These analyses ensure that the findings are scientifically valid and reproducible.

1.7.1 Over-Representation Analysis

Over-Representation Analysis (ORA) is a statistical method widely used in genomic studies to determine if a predefined set of genes (such as those belonging to specific pathways, functions, or diseases) is represented more than expected within a larger set of genes under study (Yu, 2022). This method is particularly useful in the context of high-throughput experiments, like microarray or RNA Sequencing, where researchers aim to identify biological processes or pathways significantly associated with a specific condition or disease.

The central hypothesis in ORA is that genes involved in a particular biological function or process are not randomly distributed but are often functionally related. For instance, in a gene expression study comparing

diseased vs. healthy states, if a specific pathway is significantly altered or implicated in the disease, genes associated with that pathway should be over-represented among the differentially expressed genes (Pomyen et al., 2015).

To perform ORA, a list of genes of interest (e.g., differentially expressed genes) is compared against a background list (usually the entire genome or a larger set of genes from which the gene list was derived). Statistical methods, such as the hypergeometric test or Fisher's exact test (Fisher, 1922), are employed to calculate the probability that the number of genes from the list of interest falling into a specific category (like a pathway) is higher than expected by chance (Pomyen et al., 2015). The p-values obtained from these tests are adjusted for multiple testing, often using methods like the Benjamini-Hochberg procedure, to control the false discovery rate (Benjamini & Hochberg, 1995; Pomyen et al., 2015).

ORA allows researchers to move beyond the analysis of individual genes to understand the broader biological implications of their data. It helps in identifying key pathways or processes potentially disrupted or altered in the condition under study, thereby providing insights into disease mechanisms or potential therapeutic targets.

While ORA is a powerful tool, it comes with certain limitations. It assumes that genes act independently, which is not always the case in complex biological systems (Pomyen et al., 2015). Additionally, the results of ORA can be influenced by the size of the gene set categories and the choice of background list. Researchers must carefully select their gene lists and categories to avoid biases.

Visualization tools like Venn diagrams and confusion matrices clarify the interpretations of ORA, with the Venn diagram illustrating gene set overlaps and the confusion matrix showing true and false positives and negatives. Examples of a Venn diagram and confusion matrix are given in Figure 1.2 and Figure 1.3.

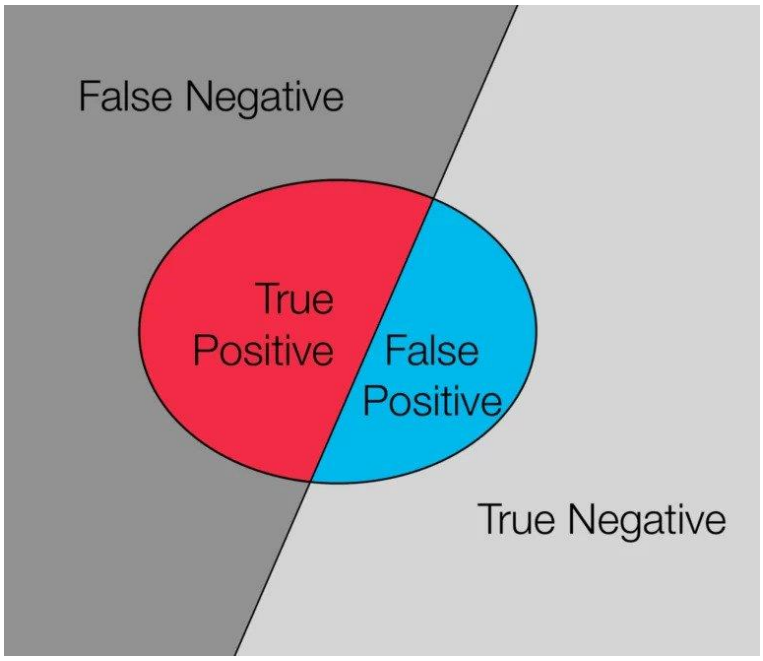


Figure 1.2: Example illustration of a Venn diagram for statistical analysis, with all possible solution spaces marked (Marzell, 2019).

		Predicted	
		Negative (**N**) **-**	Positive (**P**) **+**
Actual	Negative **-**	True Negatives (T**N**)	False Positives (F**P**) **Type I error**
	Positive **+**	False Negatives (F**N**) **Type II error**	True Positives (T**P**)

Figure 1.3: Example illustration of a confusion matrix with all possible solutions spaces marked (Gupta, 2023).

The odds ratio is a statistical measure that quantifies the strength of association between a specific condition (such as the presence of a disease) and a particular gene or set of genes being studied (Szumilas, 2010). It compares the odds of a gene being over-represented in a list of interest (like differentially expressed genes in diseased tissue) to the odds of its representation in a background or reference list (such as the entire genome or a control tissue). An odds ratio greater than one implies that the gene or genes are more likely to be associated with the condition under study than not. An odds ratio of less than one suggests a negative association, whereas an odds ratio of exactly one indicates no association. This measure is particularly insightful when determining if the presence of certain genes is non-randomly associated with the condition being investigated in the bioinformatics study.

1.7.2 Advanced Statistical Techniques for Genomic Data Analysis

Genomic data analysis has dramatically evolved with the introduction of high-throughput biological techniques, such as gene expression microarray and high-throughput sequencing. These advancements allow for the simultaneous measurement of thousands of biomolecules, necessitating sophisticated statistical methods for data analysis and interpretation (Pomyen et al., 2015).

In the realm of 'big data' genomics, multiple hypotheses testing is standard practice. This leads to challenges such as the Familywise Error Rate (FWER), and the probability of making one or more type I errors (false positives) across multiple tests. The Bonferroni correction, a straightforward method to control FWER, involves adjusting the significance level by the number of hypotheses being tested (Watkins, 2023).

In genomic studies, Fisher's Exact Test serves as a cornerstone for analyzing categorical data when sample sizes are small. This test hinges on the hypergeometric distribution to determine the exact probability of observing a specific combination of outcomes (Hoffman, 2015). To illustrate the situation, we consider a population of size N that has c_1 objects with A and c_2 with not-A ($\bar{A}$). Then we draw a sample of r_1 objects and find *a* with A. This is visualized in Table 1.1. The equation for Fisher's Exact Test is given in Equation 1.1 (Hoffman, 2015). It is particularly useful when evaluating the significance of associations within 2x2 contingency tables (confusion matrices), such as the presence or absence of a particular gene variant in disease vs. healthy states.

Table 1.1: Overview of the various groups in the hypergeometric population of size N. The population has c_1 objects with A and c_2 with not-A ($\bar{A}$). The scenario is drawing a sample of r_1 objects and finding *a* with A (Hoffman, 2015).

	A	$\bar{A}$	Total
In sample	a	b	r_1
Not in sample	c	d	r_2
	c_1	c_2	N

$$P(\textit{choosing a number of } \bar{A}\textit{s}) = \frac{\binom{c_1}{a}\binom{c_2}{b}}{\binom{N}{r_1}} = \frac{\frac{c_1!}{a!\,c!} \times \frac{c_2!}{b!\,d!}}{\frac{N!}{r_1!\,r_2!}} = \frac{c_1!\,c_2!\,r_1!\,r_2!}{N!\,a!\,b!\,c!\,d!} \qquad (1.1)$$

Where $\binom{n}{k}$ denotes the binomial coefficient, quantifying the number of ways of picking *k* unordered outcomes from *n* possibilities. $\binom{N}{r_1}$ are the number of possible samples, $\binom{c_1}{a}$ is the number of ways of choosing A in a sample of size c_1, $\binom{c_2}{b}$ is the number of ways of choosing $\bar{A}$ in a sample size N-c_1=c_2. Since these events are independent, there are $\binom{c_1}{a}\binom{c_2}{b}$ ways of choosing *a* number of As and *b* number of $\bar{A}$s, given by the chain law in probability. The exclamation mark here means factorial. For example,

consider a scenario where we're examining a gene's association with a disease. Fisher's Exact Test allows us to calculate precisely whether the observed number of disease cases with the gene variant is higher than what we would expect by chance, considering the total number of cases and controls, and the overall frequency of the variant. This precision is crucial when the dataset is too small to rely on approximations provided by other tests like chi-square, ensuring that the conclusions drawn are as accurate as possible (Hoffman, 2015).

When dealing with larger tables or multiple groups, Fisher's Exact Test can be adapted, although the calculations become more complex. In these instances, specialized software or online tools can assist in computing the probabilities and making sense of the data. The test's reliance on the hypergeometric distribution also means that each selection affects subsequent selections, a crucial consideration when sampling without replacement – a common feature in biological data where the population size is not infinitely large (Hoffman, 2015). The application of Fisher's Exact Test in genomic data analysis is emblematic of the field's reliance on precise, robust statistical tools to draw meaningful conclusions from complex biological data. Its use is essential in the face of large-scale data analyses, where traditional methods may fall short.

However, as the number of tests increases, the focus often shifts from ensuring all hypotheses are true to a more practical criterion: the False Discovery Rate (FDR). FDR is the expected proportion of false discoveries among the rejected hypotheses, a concept important in genomics with its typically large number of tests (Watkins, 2023).

To control FDR, the Benjamini-Hochberg procedure is widely used. This procedure involves ranking the p-values, setting a threshold, and rejecting hypotheses up to the point where the p-value is less than or equal to the threshold set by the FDR level (Watkins, 2023). Benjamini-Hochberg correction reduces type I errors by adjusting p-values for multiple

comparisons, ensuring more reliable results in large-scale genomic studies.

Genomics has dramatically evolved with the introduction of high-throughput biological techniques such as gene expression microarrays and high-throughput sequencing. These advancements allow for the simultaneous measurement of thousands of biomolecules, necessitating sophisticated statistical methods for data analysis and interpretation (Pomyen et al. 2015).

Furthermore, the development of computational tools for data analysis has been essential for advancing our understanding of gene expression patterns and their biological implications. Machine learning and computational methods enhance the precision of genomic research and its applications in oncology (Bhandari et al. 2022).

1.7.3 Comprehensive Gene Expression Analysis with the limma Package

The *limma* package, short for Linear Models for Microarray Data, is a cornerstone in the field of bioinformatics, widely used for gene expression analysis across both microarray and RNA sequencing technologies. Initially developed to address the challenges of high-dimensional genomic datasets, *limma* has evolved into a robust statistical framework capable of addressing the complexities associated with modern genomic research (Ritchie et al., 2015).

At its core, *limma* utilizes linear modeling techniques, allowing for the assessment of differential gene expression across various experimental conditions. This approach is especially beneficial for multifaceted experiments involving multiple factors or covariates. The *lmFit* function exemplifies this, fitting a linear model to each gene, which effectively delineates differential expression across diverse groups defined by clinical or phenotypic traits (Smyth, 2004).

Subsequent to model fitting, the eBayes function calculates empirical Bayes statistics, crucial for enhancing the reliability of differential expression results by stabilizing variance estimates across genes. This process is particularly vital in genomic studies where sample sizes are typically small, thus requiring robust statistical methods to bolster the confidence in the findings (Smyth, 2004).

The voom function is integral in transforming count data from RNA sequencing into log-counts per million, effectively adjusting for the inherent mean-variance relationship in count data. This transformation is critical for preparing the data for subsequent analysis using linear models that are more statistically appropriate (Law et al., 2014).

For hypothesis testing, *limma* offers the makeContrasts and contrasts.fit functions. These functions facilitate the construction of contrast matrices that align with specific experimental hypotheses, enabling comparisons such as between tumor and normal tissues or among different patient subgroups. The topTable function then identifies the most statistically significant genes, prioritizing them based on their evidence of differential expression (Ritchie et al., 2015).

A practical application of *limma* can be seen in lung cancer research, where it is used to explore complex genomic interactions, identifying pivotal genes that may influence disease progression, response to therapy, or disease onset.

1.8 Methods for Gene Classification

The classification of genes based on their expression in various cell types is a fundamental aspect of genomics, crucial for deciphering complex biological processes and disease mechanisms. Understanding the expression patterns of genes across different cell types enables researchers to uncover the roles of these genes in health and disease, identify potential biomarkers, and develop targeted therapies. This

chapter explores significant methodologies for gene classification, focusing on the Human Protein Atlas (HPA) approach, cellular deconvolution techniques, and the theoretical foundations and applications of cell type-specific marker genes. The integration of these methodologies provides a comprehensive framework for understanding cellular diversity and function in both healthy and diseased states.

1.8.1 The Human Protein Atlas Gene Classification Approach

The HPA conducted a systematic classification of all protein-coding genes according to their expression patterns across various cell types, building upon a methodology described earlier (Karlsson et al., 2021). Specifically, 2005 genes were identified as "Cell Type–Enriched", where the expression of a gene, measured as adjusted transcripts per million (TPM), was found to be at least fourfold higher in one specific cell type compared to all other analyzed cell types. Additionally, 2893 genes were classified as "Group-Enriched", denoting genes that were enriched in a set of up to 10 cell types. Moreover, there are 9062 "Cell Type–Enhanced" genes, where the expression of such genes was at least fourfold higher in one cell type compared to the mean expression across all other cell types (Karlsson et al., 2021).

Interestingly, 4257 genes exhibited a low cell type specificity, showing roughly equivalent levels across all examined cell types (Uhlén, 2015; Karlsson et al., 2021). A mere 11% of the genes were detected in all the cell types analyzed, reinforcing prior estimates about the count of universal "housekeeping" genes that are indispensable in every cell (Karlsson et al., 2021). The classification results further highlighted the testis as having the most cell type elevated genes, aligning with previous findings. Numerous elevated genes were also pinpointed in the eye, especially in photoreceptor cells, bipolar cells, and horizontal cells, as well as in the ciliated cells of the lung (Karlsson et al., 2021).

An integral aspect of HPA's methodology is the fusion of multiple analysis platforms, facilitating the validation of single-cell data via antibody-based image profiling in tissues (Karlsson et al., 2021). This approach, using immunohistochemistry, provides a detailed view not only at the single-cell level but also gives insights into the exact spatial pattern, variations between cells, and subcellular localization. The study by Karlsson et al. provided examples of this validation, emphasizing proteins that were specifically expressed in unique structures, including renal collecting ducts, retinal photoreceptor cells, early spermatids, intercalated discs in cardiomyocytes, and hepatic Kupffer cells (Pomyen et al., 2015; Karlsson et al., 2021).

1.8.2 Cellular Deconvolution

Cellular deconvolution, a pivotal technique in bioinformatics, focuses on estimating the proportions of various cell types in mixed tissue samples. Its significance lies in unraveling the complexities within tissues composed of different cell populations, like those found in tumors. Traditional gene expression analyses often fail to capture the nuances of these mixed tissues, leading to overlooked signals from less prevalent cell types (Avila Cobos et al., 2018). Cellular deconvolution addresses this by imputing both cell type abundances and their specific expression profiles, enhancing our understanding of gene expression in mixed cell populations.

CIBERSORTx emerges as a novel method in this realm, coined as "in silico flow cytometry." Developed by Newman et al., it leverages gene expression data to approximate the abundances of distinct cell types within a mixed cell population (Newman et al., 2019). What sets CIBERSORTx apart is its capability to process bulk gene expression data along with a signature matrix file, which outlines the expression profile for each cell type. Users have the flexibility to employ existing signature matrices or generate custom ones by supplying pure cell population

expression profiles. With the advent of single-cell RNA Sequencing, CIBERSORTx also allows for the derivation of signature matrices from such data.

CIBERSORTx consists of two key analysis modules. The "Cell Fractions" module measures the proportions of different cell subpopulations within bulk tissue expression profiles. It's distinctive for its capability to deconvolve bulk RNA Sequencing data using signature genes derived from either single-cell transcriptomes or sorted cell populations (Newman et al., 2019). The "Gene Expression" module, on the other hand, infers cell type-specific expression profiles from bulk tissue transcriptomes, without the need for physical cell sorting (Newman et al., 2019). This module functions in two modes: "Group-Level," which generates representative transcriptome profiles for each cell type, aiding the understanding of context-dependent changes in expression, and "High-Resolution," aimed at deducing sample-level expression variations across distinct cell types, suitable for exploring variations in cell type expression without relying on pre-defined biological groupings.

CIBERSORTx's implementation on a web platform, backed by R and PHP (Hypertext Preprocessor), ensures accessibility and ease of use, further enhanced by a user-friendly interface and comprehensive guides, including step-by-step tutorials (Newman et al., 2019).

Each cell type, with its distinct gene expression profile, contributes to the overall molecular information in bulk samples. Consequently, analyses like differential gene expression can be affected by variations in cell type proportions. Addressing this, cellular deconvolution algorithms have found applications in a variety of samples, improving interpretability, and reducing the confounding effects of cellular heterogeneity (Avila Cobos et al., 2018).

1.8.3 Theoretical Foundations and Applications of Cell Type-Specific Marker Genes

Cell type-specific (CTS) genes, also known as marker genes, play a pivotal role in the analysis of RNA transcriptional data by defining cellular identity. These genes are typically highly expressed in one cell type but are lowly expressed in others, which allows them to provide essential insights into the core set of genes that characterize a particular cell type (Qiu et al., 2021). Understanding marker genes is crucial for filling gaps in our knowledge of cell biology and could elucidate the cellular origins of various pathologies.

Marker genes are extensively used to annotate cell clusters, analyze the cellular composition of bulk tissues, and estimate cell type fractions via deconvolution techniques. They also enable the estimation of cell type-specific expression directly from bulk tissue samples (Qiu et al., 2021). This wide array of applications underscores the importance of marker genes in enhancing our understanding of complex biological systems and the intricacies of cellular function.

A common approach to identifying marker genes involves conducting statistical tests on cell type-specific transcriptome data, typically derived from single-cell RNA sequencing (scRNA-seq). Genes that demonstrate significant expression differences between a specific cell type and all others are regarded as marker genes for that cell type. However, despite the appeal of this approach, challenges such as the high cost of single-cell sequencing, difficulties in obtaining viable cells from certain tissues like the human brain, and the inherent noisiness of scRNA-seq data can complicate the direct acquisition and analysis of CTS data. Furthermore, using scRNA-seq data from other species to infer marker genes poses additional challenges due to potential disparities in gene expression profiles across species.

Qualitative Marker Genes in HPA and hECA

The Human Protein Atlas (HPA) utilizes an extensive selection of qualitative marker genes to classify cell types. These markers include both those from original publications and additional markers used in pathology diagnostics, which are chosen based on their well-established specificity to certain cell types. This selection is guided by a strong correlation between the gene's cluster-specific expression and expected expression patterns (Karlsson et al., 2021). These markers are critical for accurately annotating the vast array of cell types identified in the HPA, aiding significantly in enhancing the precision of cell type classifications.

Similarly, the Human Ensemble Cell Atlas (hECA) employs a combinatory approach that incorporates both canonical, knowledge-based marker genes and data-derived differentially expressed genes (DEGs) to construct a comprehensive marker reference. For cell types whose marker genes were not given in the original studies, the hECA team surveyed markers from multiple sources, including Gene Ontology, PanglaoDB, the Human Protein Atlas, and CellMarker to replenish the marker references. In most cases, the top 10 DEGs for each cluster in each dataset were considered, ensuring the robustness of cell type classifications (Chen et al., 2022).

In practical applications, these marker genes are indispensable for the precise identification and characterization of cell types within tissues, especially in pathologies such as cancer where cellular heterogeneity is pronounced. By providing a means to classify cells based on definitive expression profiles, marker genes facilitate a deeper understanding of cellular diversity and function in health and disease. Both the HPA and hECA databases serve as crucial resources for the scientific community, providing access to detailed gene expression profiles and cellular localization data which are instrumental in various scientific investigations.

1.9 Software Development Process

The development of the *CellTypeGenomics* package is delineated through a structured sequence of phases, each integral to the realization of a robust and functional software tool. This sequence comprises requirement analysis, design and prototyping, coding, and comprehensive testing, each of which plays a crucial role in the software's lifecycle.

The process begins with an extensive requirement analysis phase, during which detailed consultations with bioinformatics experts are conducted. This interaction is essential to capture the precise needs and identify potential deficiencies within existing software tools. Such engagements are fundamental in establishing the functional requirements of the *CellTypeGenomics* package, ensuring its utility and relevance to the target user base.

Subsequent to requirement analysis is the design and prototyping phase. In this stage, initial models of the software are constructed and are subject to iterative refinements based on user feedback. This iterative design process is indicative of agile development methodologies, which prioritize flexibility and user feedback over rigid planning and development schedules (Patton, 2014). Agile practices are particularly adept at accommodating changes in user requirements and emerging technologies, thereby enhancing the adaptability and longevity of the software.

The coding phase is executed using Python, a programming language well-regarded for its extensive library support and readability—attributes that are particularly advantageous in the domain of bioinformatics (Van Rossum & Drake, 2009). The choice of Python is strategic, facilitating the integration of complex data structures and algorithms while maintaining clarity and ease of maintenance.

The culmination of the development process is marked by a comprehensive testing phase. This phase employs a combination of unit and integration testing to ensure the software's reliability and accuracy

(Myers et al., 2011). Unit tests evaluate the functionality of individual software components, whereas integration tests assess the cohesive operation of these components within the full software system. Such rigorous testing is essential to ascertain that the software adheres to the specified requirements and performs effectively under varied operational scenarios.

Through these structured phases, the development of the *CellTypeGenomics* package is meticulously orchestrated to meet the specified design and functional criteria. This approach not only ensures compliance with initial specifications but also imbues the software with the necessary flexibility to adapt to future advancements in the field of bioinformatics.

1.10 Research Questions

Investigating the molecular landscape of lung cancer requires a detailed examination of gene expressions and their interactions with cellular and environmental factors. Utilizing comprehensive genomic data from The Cancer Genome Atlas (TCGA), this study employs the *CellTypeGenomics* tool, which leverages data from The Human Protein Atlas (HPA) to enhance its analyses and conducts gene ontology analysis to address several pivotal research questions aimed at elucidating the complex dynamics within cancer cells.

The primary focus of this research is to identify genes that are differentially expressed between tumor and normal lung tissues, specifically determining which genes are upregulated or downregulated in tumors compared to normal tissues. This inquiry is essential for understanding the cellular context of these gene expression changes and gaining insights into their roles in lung cancer. The *CellTypeGenomics* tool uses data from HPA to identify cell types based on Ensembl gene identifiers and associates these cell types with differential gene

expressions observed in TCGA data, thereby enhancing our understanding of the cellular dynamics contributing to tumor development and progression.

The broader biological implications of these differentially expressed genes are further explored through Reactome pathway analysis. By mapping these genes onto specific biological pathways, the research seeks to delineate the functional pathways that are altered in lung cancer. These comprehensive analyses provide a deep view of gene functions, their interactions, and the biochemical pathways disrupted in the disease.

Furthermore, the study leverages TCGA data to perform gene ontology (GO) analysis, which categorizes differentially expressed genes into biological processes (BP), cellular components (CC), and molecular functions (MF). The results of the GO Biological Processes are visualized using Directed Acyclic Graphs (DAGs), providing a structured view of the biological processes involved and their hierarchical relationships. This visualization is crucial for interpreting the complex interactions and expression levels of genes across different samples and conditions, offering insights into the molecular landscape of lung cancer.

Additionally, this thesis examines how smoking status modifies gene expression profiles in lung cancer and identifies the key molecular pathways predominantly affected by these changes. By analyzing the differential gene expression linked to smoking, the research will shed light on the molecular mechanisms altered by this significant environmental factor.

Finally, the robustness and applicability of the *CellTypeGenomics* package are critically evaluated. This evaluation aims to determine to what extent the tool can analyze TCGA lung cancer data, using HPA data to identify cell-type origins across various datasets and experimental conditions. This assessment will help validate the utility of the *CellTypeGenomics* package in broader genomic research applications.

2 Materials and Methods

This chapter provides a detailed account of the methodologies and tools utilized in this study, focusing on the application and development of the *CellTypeGenomics* Python package. The primary aim is to explore the cell-type origins of differentially expressed genes in lung cancer, leveraging the capabilities of this package to interpret and analyze complex genomic data. Two primary data sources, the Human Protein Atlas (HPA) and The Cancer Genome Atlas (TCGA), are employed in this research. Each serves a distinct purpose within the research framework. HPA data provides detailed annotations and normalized expression levels of genes across various tissues and cell types, crucial for associating specific genes with their corresponding cell types. Conversely, TCGA data, serving primarily as a test data source, offers extensive gene expression profiles from lung cancer samples. Additionally, marker genes from the Human Ensemble Cell Atlas (hECA) are integrated to enhance the analysis.

The chapter begins with an overview of data acquisition methods for HPA, TCGA, and hECA, followed by detailed processes for data preparation and integration. Subsequently, it outlines the statistical methodologies employed, including differential expression analysis and pathway analysis, and describes the visualization techniques used to represent the data. The development and application of the *CellTypeGenomics* package are then discussed, along with the quality control and validation measures implemented to ensure the robustness and reliability of the findings.

2.1 Data Sources

This section outlines the acquisition of data from the Human Protein Atlas (HPA) and The Cancer Genome Atlas (TCGA), which are central to the analyses performed in this study. Additionally, it includes the integration of marker genes from the Human Ensemble Cell Atlas (hECA) to enhance the analysis.

2.1.1 Human Protein Atlas (HPA) Data Acquisition

The Human Protein Atlas dataset was sourced from its official website, specifically from the downloadable data section. The dataset, labeled "proteinatlas.tsv.zip," is part of Human Protein Atlas version 23.0 (Human Protein Atlas, 2023). This dataset includes extensive gene annotations covering gene synonyms, Ensembl gene IDs, descriptions, and detailed RNA and protein expression data across various tissues and cell types. The comprehensive nature of the HPA dataset, with its detailed data on gene expression across different cell types, is invaluable for associating specific genes with particular cell types, which is crucial for this study.

2.1.2 The Cancer Genome Atlas (TCGA) Data Acquisition

For this project, comprehensive genomic data was acquired from The Cancer Genome Atlas (TCGA), a resource aggregating genomic information across various cancer types. Data retrieval was facilitated through the Genomic Data Commons (GDC) Data Portal, enabling access to harmonized cancer datasets tailored to lung cancer research (The Cancer Genome Atlas Program, 2024). The downloaded files contained lung cancer tissue data from adenomas, adenocarcinomas and squamous cell neoplasms.

The specific datasets procured included bulk RNA sequencing data and associated clinical metadata critical for identifying gene expression patterns in cancerous tissues of the lung. The selection criteria applied in the GDC Data Portal for downloading the data were specified to ensure reproducibility and to cater to the specific requirements of our research:

- Access Level: Open, ensuring all retrieved data is publicly accessible.
- Primary Site: Bronchus and lung, focusing our analysis specifically on lung-related cancers.
- Data Category: Transcriptome Profiling, selected to provide a comprehensive view of gene expression within the samples.
- Data Type: Gene Expression Quantification, which provides precise measurement of gene expression levels.
- Experimental Strategy: RNA-Seq, a method chosen for its high throughput and accuracy in quantifying transcripts (Wang et al., 2009).

A total of 2001 files were downloaded, comprising approximately 8.48 GB of data, reflecting a robust dataset focused on lung cancer. These files are in TSV format, facilitating easy integration and analysis within bioinformatics tools. This approach ensures that our data selection is tailored to maximize the relevance of our findings on lung cancer genomics.

2.1.3 Marker Genes from HPA and hECA

Marker genes, which are indicative of specific cellular or tissue states, played a pivotal role in our analysis. These genes were identified from both the Human Protein Atlas (HPA) and the Human Ensemble Cell Atlas (hECA) datasets. From HPA, our selection focused on genes that demonstrated a high tissue-specific expression. We specifically employed a four-fold numerical threshold for marker gene selection set by HPA,

which facilitated the identification of genes significantly expressed in distinct tissues. This threshold was crucial for the Fisher test in the original cell-type origin analysis, and the relevant data were directly extracted from the TSV file mentioned in chapter 2.1.2.

For the ontology function of the *CellTypeGenomics* package, we utilized the marker genes data from supplementary data S2 in the HPA study (Karlsson et al., 2021). This subset of data was chosen due to its higher quality and manual curation, which provided more reliable and precise markers for various cell types.

In hECA, we extracted marker genes for the ontology function using the *uHAF marker reference.xlsx* file (Chen et al., 2022), which compiled a robust reference of marker genes, including both knowledge-based marker genes and data-derived DEGs.

2.2 Data Preparation and Integration

This section describes the methods used to prepare and integrate data from HPA, TCGA, and hECA. It covers normalization techniques, alignment of Ensembl gene identifiers, and the creation of a unified analysis pipeline to ensure consistent and reliable comparisons across samples.

2.2.1 Human Protein Atlas Data Processing

After downloading, the compressed file was extracted to access the TSV file. The dataset from this TSV file was then structured for analysis using *Pandas*, a Python library renowned for its data-handling capabilities. The focus was on specific columns that provided insights into the normalized expression levels of genes across different tissues and cell types, namely:

- RNA tissue specific nTPM
- RNA single cell type specific nTPM
- RNA blood cell specific nTPM
- RNA blood lineage specific nTPM

These columns provide valuable insights into the normalized expression levels of genes in different tissues and cell types. A Python script aggregated Ensembl gene IDs based on their expression in various cell types, resulting in a comprehensive JSON file that mapped different cell types to their associated genes as per the HPA dataset.

2.2.2 The Cancer Genome Atlas Data Processing

Building on the foundational understanding of *limma* as detailed in chapter 1.7.3, this section translates its theoretical capabilities into practical applications, specifically for our analysis of The Cancer Genome Atlas (TCGA) data. The initial step involves loading the *limma* library, crucial for employing the sophisticated linear modeling techniques that the package is known for. Our analysis workflow commences with the loading of gene expression data and associated metadata from CSV files. These data sets are meticulously linked by aligning sample identifiers, facilitating an integrated approach to subsequent analyses.

Following data importation, we undertake a meticulous preprocessing of the metadata. This involves standardizing the sample types to ensure consistency across the data set, a step critical for the integrity of the analyses that follow. We focus specifically on samples identified as *Primary Tumor* and *Solid Tissue Normal*, filtering out all irrelevant or incomplete data entries. This selective approach not only streamlines subsequent analyses but also enhances the accuracy of our differential expression analysis.

2.2.3 Data Normalization and Filtration

Normalization of the expression data is executed utilizing the voom function of *limma*, transforming RNA-Seq count data into log2-counts per million, a format amenable to the linear modeling techniques that *limma* executes with high precision. This step is essential to adjust for the inherent technical and biological variability in the data, ensuring that the differential expression analyses are robust and reliable.

With normalized data, we proceed to differential expression analysis. Employing *limma*'s lmFit function, we fit linear models to each gene, systematically exploring differences in expression between conditions such as tumor versus normal tissues. This modeling is crucial for elucidating the molecular underpinnings of lung cancer. To further refine our analysis, specific contrasts are defined using the makeContrasts function, and these are fitted to the models using contrasts.fit. These contrasts are specified in Chapter 2.3.2. The application of the eBayes function follows, enhancing the reliability of our results by stabilizing variance estimates—a critical feature when dealing with the typically small sample sizes in genomic studies.

To ensure that only biologically significant changes are highlighted, a multi-step filtering process is applied to the topTable results from *limma*. Initially, a threshold of 1 is set for log(FPKM) (Fragments Per Kilobase of transcript per Million mapped reads, representing average expression). This preliminary filtering ensures that only genes with a minimum expression level are considered for further analysis.

Following this initial filtering, a dynamic logFC threshold is applied. This dynamic threshold is estimated based on the relationship between average expression (AveExpr) and absolute logFC for significant genes. This approach is particularly useful when dealing with a high number of significant genes, as it provides a robust fit for the dynamic threshold.

The dynamic filtering process begins by modeling the relationship between the average expression of genes and their absolute log-fold change values using a loess (locally estimated scatterplot smoothing) regression. This regression establishes a threshold that varies according to the average expression levels of the genes.

Next, the dynamic threshold is adjusted to ensure it meets a minimum required value of abs(logFC) greater than one 1. This adjustment sets a baseline threshold that all genes must meet or exceed, even if their average expression levels are higher. This step ensures that the threshold is both flexible and stringent enough to capture significant biological changes.

Using the computed dynamic threshold function, genes are then filtered based on whether their absolute log-fold change meets or exceeds the threshold for their average expression levels. This filtering process ensures that only genes showing significant changes in expression are retained for further analysis.

Results are then saved in CSV format, ensuring that they are accessible for further analysis.

2.2.4 Marker Genes Processing

The conversion of marker gene symbols to Ensembl Gene IDs was performed through a systematic, multi-stage approach, leveraging various bioinformatics tools to maximize the resolution of gene identifiers. This process was crucial for integrating gene expression data into wider genomic analyses, especially when correlating cell-type specific expressions with genomic datasets.

Initially, the *mygene* Python package was utilized to convert gene symbols from HPA and hECA datasets into Ensembl IDs. This tool provided a direct query interface to genomic databases, facilitating the retrieval of gene

IDs. Despite *mygene*'s utility, several genes were either not found or there were multiple hits per gene when converting to Ensembl Gene IDs.

Due to unresolved symbols from the first iteration, the *biomart* package was employed. It connected directly to the Ensembl BioMart database offering a more robust search for Ensembl IDs through another layer of validation. This iteration managed to obtain more unique mappings between gene symbols and Ensembl IDs that were faulty in the first iteration.

The remaining unresolved gene symbols were then queried using the REST API provided by Ensembl, which is capable of accessing up-to-date and comprehensive genetic data. This final automated step helped identify several Ensembl IDs, although some genes remained unmatched due to various reasons including possible obscurity or recent reclassifications in genomic databases.

The few persistently unmatched genes were subjected to manual searches using The Human Gene Database and other literature sources to assign the most plausible Ensembl IDs. This step was crucial to ensure completeness of the dataset.

2.3 Statistical Analysis and Tools

This section elaborates on the statistical methodologies and computational tools used to analyze the integrated data from the Human Protein Atlas (HPA) and The Cancer Genome Atlas (TCGA). Focusing on statistical approaches, the analysis utilizes the *limma* package, a comprehensive bioinformatics tool designed for the analysis of gene expression data through linear models.

The process begins with the application of the *limma* package to perform differential expression analysis. This analysis is critical as it identifies genes that are significantly upregulated or downregulated in lung cancer

samples compared to normal tissue samples. The ability of *limma* to handle complex experimental designs and large datasets makes it particularly suitable for the genomic data used in this study.

Data normalization is a preliminary step before differential expression analysis, where RNA-seq data from TCGA are transformed into log2-counts per million using *limma*'s voom function. This normalization method adjusts for library size differences and other technical variabilities, facilitating a more accurate comparison of gene expression levels across samples.

Following the identification of differentially expressed genes, the results are integrated with pathway analysis to ascertain which biological pathways are enriched with these genes. This step is performed using bioinformatics tools that map genes to known pathways, highlighting potential mechanisms underlying lung cancer and suggesting targets for therapeutic intervention.

Additionally, intersection analysis is employed to group genes with similar expression patterns, indicating shared regulatory mechanisms or functional roles in lung cancer. This intersection analysis helps in understanding the complex biological relationships and can guide further experimental design or hypothesis generation.

Quality control measures are stringently applied throughout the analysis to ensure the reliability and reproducibility of the results. These include multiple hypothesis testing adjustments using the Benjamini-Hochberg procedure to control the false discovery rate, ensuring that the findings are statistically significant.

Building on these analytical foundations, the following chapters delve deeper into the specific *limma* designs and contrasts used in our study, examining both the full dataset and a critical subset defined by smoking categories. Chapter 2.3.1 discusses the configuration of experimental setups tailored to our research questions, illustrating how the design

choices enhance the robustness of our statistical conclusions across both the complete dataset and its subsets. Chapter 2.3.2 explores the specific comparisons made between different conditions, including those based on smoking status, to discern the subtle nuances in gene expression that are pivotal for understanding lung cancer progression. Together, these chapters extend our discussion on the statistical framework and bioinformatics tools employed, providing a thorough exploration of the methodologies that drive our research in genomic data interpretation.

2.3.1 Constructing the Design Matrix

In this section, we discuss the construction of the design matrix for both the full dataset and the subset focusing on smoking data. The design matrix is a crucial component in linear modeling as it defines the structure of the covariates and their interactions used in the analysis.

Full Dataset

For the full dataset, the design matrix incorporates the interaction between sample type and gender, along with age as a continuous variable.

Design Matrix for Full Dataset

First, we create an interaction term between `Sample.Type` and `gender`. This interaction term allows us to consider the combined effect of these two variables on gene expression.

```
metadata_filtered$SampleSex <- with(metadata_filtered,
interaction(Sample.Type, gender, sep="_"))
```

Next, we construct the design matrix. The term `normalized_age` is included as an independent variable. It is treated as a continuous variable, representing the normalized age of the patients. This inclusion allows us to assess the effect of age on gene expression directly.

```
design <- model.matrix(~ 0 + SampleSex + normalized_age,
data = metadata_filtered)
```

In this matrix, `SampleSex` represents the interaction of sample type and gender, while `normalized_age` captures the continuous age variable. This design matrix enables the analysis of gene expression variations considering both the interaction of categorical variables (sample type and gender) and the continuous effect of age.

Smoking Dataset

For the smoking dataset, we extend the interaction term to include smoking status, thereby incorporating the effect of smoking on gene expression.

Design Matrix for Smoking Dataset:

First, we create an interaction term that includes `Sample.Type`, `gender`, and `smoking_category`. This comprehensive interaction term accounts for the combined effect of these variables.

```
metadata_filtered$interaction_term <-
with(metadata_filtered, interaction(Sample.Type, gender,
smoking_category))
```

Next, we construct the design matrix for the smoking dataset. Similar to the full dataset, `normalized_age` is included as an independent variable.

```
design <- model.matrix(~ 0 + interaction_term +
normalized_age, data = metadata_filtered)
```

In this matrix, `interaction_term` represents the combined interaction of sample type, gender, and smoking category, while `normalized_age` continues to capture the continuous age variable. This design matrix allows for the analysis of gene expression variations considering the interaction of categorical variables (sample type, gender, smoking status) and the continuous effect of age.

By constructing these design matrices, we ensure that our linear models can effectively incorporate and analyze the influence of multiple covariates and their interactions on gene expression.

2.3.2 Regression Contrasts for both Datasets

The contrasts were defined to investigate specific biological hypotheses and differences between sample groups in our dataset.

Full Dataset

The following contrasts were used for the full dataset.

Tumor vs. Normal: This contrast compares gene expression levels between primary tumor samples and solid tissue normal samples. It is formulated in the code block below.

```
TumorVsNormal = (SampleSexPrimaryTumor_male +
SampleSexPrimaryTumor_female)/2 -
(SampleSexSolidTissueNormal_male +
SampleSexSolidTissueNormal_female)/2
```

This approach averages the expression levels for male and female samples in both the tumor and normal tissue groups to provide a robust comparison.

Tissue x Sex Interaction: This contrast examines the interaction effect between tissue type (tumor or normal) and sex (male or female). It is expressed below.

```
TissueXSex = (SampleSexPrimaryTumor_male -
SampleSexPrimaryTumor_female) -
(SampleSexSolidTissueNormal_male -
SampleSexSolidTissueNormal_female)
```

This formulation captures the differential expression due to the interaction between sex and tissue type.

Male vs. Female: This contrast explores differences in gene expression between male and female samples, irrespective of tissue type. It is formulated below.

```
MaleVsFemale = (SampleSexPrimaryTumor_male +
SampleSexSolidTissueNormal_male)/2 -
(SampleSexPrimaryTumor_female +
SampleSexSolidTissueNormal_female)/2
```

By averaging the expression levels of male and female samples across both tissue types, this contrast isolates the effect of sex on gene expression.

Smoking Dataset

The following contrasts were used for the smoking dataset.

Tumor vs. Normal: This contrast compares gene expression levels between primary tumor samples and solid tissue normal samples. It is formulated in the code block below.

```
    TumorVsNormal =
(interaction_termPrimaryTumor.female.0 +
interaction_termPrimaryTumor.male.0 +
interaction_termPrimaryTumor.female.1 +
interaction_termPrimaryTumor.male.1 +
interaction_termPrimaryTumor.female.2 +
interaction_termPrimaryTumor.male.2) / 6 -

(interaction_termSolidTissueNormal.female.0 +
interaction_termSolidTissueNormal.male.0 +
interaction_termSolidTissueNormal.female.1 +
interaction_termSolidTissueNormal.male.1 +
interaction_termSolidTissueNormal.female.2 +
interaction_termSolidTissueNormal.male.2) / 6
```

This approach averages the expression levels for male and female samples in both the tumor and normal tissue groups to provide a robust comparison.

Tissue x Sex Interaction: This contrast examines the interaction effect between tissue type (tumor or normal) and sex (male or female). It is expressed below.

```
TissueXSex = (
(interaction_termPrimaryTumor.male.0 -
interaction_termPrimaryTumor.female.0) +
(interaction_termPrimaryTumor.male.1 -
interaction_termPrimaryTumor.female.1) +
(interaction_termPrimaryTumor.male.2 -
interaction_termPrimaryTumor.female.2)) / 3 -

((interaction_termSolidTissueNormal.male.0 -
interaction_termSolidTissueNormal.female.0) +
(interaction_termSolidTissueNormal.male.1 -
interaction_termSolidTissueNormal.female.1) +
(interaction_termSolidTissueNormal.male.2 -
interaction_termSolidTissueNormal.female.2)) / 3
```

This formulation captures the differential expression due to the interaction between sex and tissue type.

Male vs. Female: This contrast explores differences in gene expression between male and female samples, irrespective of tissue type. It is formulated below.

```
    MaleVsFemale = (
interaction_termPrimaryTumor.male.0 +
interaction_termPrimaryTumor.male.1 +
interaction_termPrimaryTumor.male.2 +
interaction_termSolidTissueNormal.male.0 +
interaction_termSolidTissueNormal.male.1 +
interaction_termSolidTissueNormal.male.2) / 6 -

(interaction_termPrimaryTumor.female.0 +
interaction_termPrimaryTumor.female.1 +
```

```
interaction_termPrimaryTumor.female.2 +
interaction_termSolidTissueNormal.female.0 +
interaction_termSolidTissueNormal.female.1 +
interaction_termSolidTissueNormal.female.2) / 6
```

By averaging the expression levels of male and female samples across both tissue types, this contrast isolates the effect of sex on gene expression.

Current vs. Former (Tumor): This contrast explores differences in gene expression between current smokers and former smokers, on tumor tissue. It is formulated below.

```
    CurrentVsFormerTumor =
(interaction_termPrimaryTumor.female.2 +
interaction_termPrimaryTumor.male.2) / 2 -

(interaction_termPrimaryTumor.female.1 +
interaction_termPrimaryTumor.male.1) / 2
```

By contrasting the expression levels of current smokers and former smokers across tumor tissue, this contrast isolates the effect of not smoking anymore on gene expression.

Current vs. Never (Tumor): This contrast explores differences in gene expression between current smokers and never smokers, on tumor tissue. It is formulated below.

```
    CurrentVsNeverTumor =
(interaction_termPrimaryTumor.female.2 +
interaction_termPrimaryTumor.male.2) / 2  -
```

```
(interaction_termPrimaryTumor.female.0 +
interaction_termPrimaryTumor.male.0) / 2,

    FormerVsNeverTumor =
(interaction_termPrimaryTumor.female.1 +
interaction_termPrimaryTumor.male.1) / 2 -
(interaction_termPrimaryTumor.female.0 +
interaction_termPrimaryTumor.male.0) / 2
```

By contrasting the expression levels of current smokers and never smokers across tumor tissue, this contrast isolates the effect of never having smoked on gene expression.

Current vs. Former (Normal): This contrast explores differences in gene expression between current smokers and former smokers, on normal tissue. It is formulated below.

```
    CurrentVsFormerNormal =
(interaction_termSolidTissueNormal.female.2 +
interaction_termSolidTissueNormal.male.2) / 2 -

(interaction_termSolidTissueNormal.female.1 +
interaction_termSolidTissueNormal.male.1) / 2
```

By contrasting the expression levels of current smokers and former smokers across normal tissue, this contrast isolates the effect of not smoking anymore on gene expression.

Current vs. Never (Normal): This contrast explores differences in gene expression between current smokers and never smokers, on normal tissue. It is formulated below.

```
    CurrentVsNeverNormal =
(interaction_termSolidTissueNormal.female.2 +
interaction_termSolidTissueNormal.male.2) / 2  -
(interaction_termSolidTissueNormal.female.0 +
interaction_termSolidTissueNormal.male.0) / 2,

    FormerVsNeverNormal =
(interaction_termSolidTissueNormal.female.1 +
interaction_termSolidTissueNormal.male.1) / 2-
(interaction_termSolidTissueNormal.female.0 +
interaction_termSolidTissueNormal.male.0) / 2
```

By contrasting the expression levels of current smokers and never smokers across normal tissue, this contrast isolates the effect of never having smoked on gene expression.

2.4 Development of the *CellTypeGenomics* Package

The development of the *CellTypeGenomics* package is a key component of this thesis, designed to provide robust tools for exploring the cell-type origins of differentially expressed genes in lung cancer by leveraging complex genomic data. This package utilizes detailed data from the Human Protein Atlas (HPA) to accurately map Ensembl gene identifiers to specific cell types, enabling a nuanced understanding of gene expression patterns within various cellular contexts.

2.4.1 Rationale for Development

The *CellTypeGenomics* package was developed to address the need for specialized tools capable of utilizing HPA data for detailed cell type identification. This functionality is crucial for studies on complex diseases such as lung cancer and psoriasis, where understanding cellular dynamics is key to uncovering disease mechanisms and identifying potential therapeutic targets.

2.4.2 Core Functionality

At the heart of the *CellTypeGenomics* package is its capability to map Ensembl gene identifiers to cell types using the extensive numerical data available from HPA. This fundamental feature allows for the detailed analysis of the cellular origins of gene expressions, which is critical for investigating the pathogenesis of diseases. The output from this analysis includes structured data frames that detail the associations between genes and cell types, complete with statistical analyses such as p-values and odds ratios. These outputs provide researchers with precise and actionable data on gene expression patterns.

2.4.3 Additional Features

The *CellTypeGenomics* package includes advanced functionalities that enhance its analytical capacity. It supports both numerical and qualitative marker genes from the Human Protein Atlas (HPA), enhancing the precision of cell-type specificity analysis. Additionally, the package supports qualitative marker genes from the Human Ensemble Cell Atlas (hECA), allowing for a broader exploration of cell types in various datasets. The capability to analyze tissue origins further broadens the genetic analysis, including both cell-type and tissue-specific gene

expression patterns. These features are invaluable for studies aimed at understanding the broader biological contexts of gene expressions.

2.4.4 Handling Data from Multiple Sources

The *CellTypeGenomics* package is adept at processing input lists of Ensembl gene identifiers from diverse genomic datasets, such as The Cancer Genome Atlas and psoriasis-specific studies. By analyzing these lists and returning structured DataFrames that detail gene-cell type associations, the package demonstrates its flexibility and broad applicability in genomic research, enabling comprehensive analyses that incorporate a wide variety of biological and medical contexts.

2.4.5 Optimization and Validation

Extensive efforts have been made to optimize the *CellTypeGenomics* package for handling large genomic datasets efficiently. It has been thoroughly tested with both synthetic benchmarks and real-world data to ensure its accuracy and reliability, making it a dependable tool for scientific research.

2.4.6 Documentation and Community Engagement

To aid users and foster an open-source community, comprehensive documentation is provided alongside the *CellTypeGenomics* package. Available on GitHub and PyPI, the documentation offers detailed instructions on installation, usage, and troubleshooting, encouraging collaboration and ongoing development by researchers worldwide (Føleide, 2024).

2.5 Visualization Techniques

Effective visualization is crucial in genomic research as it aids in the interpretation and communication of complex data sets. This section describes the visualization techniques employed in this study to represent data obtained from the Human Protein Atlas (HPA) and The Cancer Genome Atlas (TCGA). These techniques were integral in elucidating the relationships between differentially expressed genes, their associated cell types, and the biological pathways involved in lung cancer.

2.5.1 Dot Plots

Dot plots were extensively used to display the expression levels of genes across different samples and conditions. This method was particularly useful in showcasing the variability and distribution of gene expression within and between groups defined by clinical or biological parameters. In this study, dot plots helped visualize the expression profiles of genes identified as differentially expressed in lung cancer tissues compared to normal tissues, as well as differences influenced by factors such as patient sex and smoking status. The dot plots were created using the *Matplotlib* library in Python, which provides extensive customization options and robust functionality for creating detailed and informative plots.

2.5.2 Upset Plots

Upset plots, a modern alternative to Venn diagrams, were utilized to visualize the intersections of multiple gene sets and their relationships. These plots were particularly useful in this study for analyzing the overlap between differentially expressed genes across various conditions and subsets within the data. By using the *UpSetPlot* library in Python, the study was able to generate clear and concise representations of complex relationships in the data, highlighting significant overlaps and unique

expressions within subgroups. This visualization technique was crucial for identifying patterns that are not immediately obvious from raw data alone.

2.5.3 Directed Acyclic Graphs (DAGs)

Directed Acyclic Graphs (DAGs) were employed to visualize the hierarchical relationships among Gene Ontology (GO) terms associated with differentially expressed genes. These graphs played a crucial role in elucidating the biological pathways and processes impacted by changes in gene expression observed in lung cancer. By mapping the enriched GO terms in a DAG, the study highlighted the interconnected nature of biological processes and identified key pathways.

First, the Gene Ontology OBO (Open Biological and Biomedical Ontologies) file was streamed from an online URL and temporarily saved to disk using the requests library. This step facilitated the local processing of the GO terms. Next, GO terms with associated p-values and parent-child relationships were loaded from JSON files. These terms were then sorted by p-value to prioritize the most significant terms for analysis. To organize the GO terms, they were structured into hierarchical layers based on their parent-child relationships. Each term was assigned to a specific layer, which represents its level in the hierarchy (top significant, second most significant, etc.). For visual distinction in the DAG plots, each layer of GO terms was assigned a distinct color inspired by the rainbow spectrum. This coloring scheme helps to easily differentiate between various levels of the hierarchy. To ensure the accuracy of the hierarchical relationships, GO terms were loaded again from the OBO file. This step verified that the terms and their relationships were correctly represented. Finally, DAG plots were created using the GOATOOLS library. Each term in the DAG was color-coded according to its assigned layer. The resulting plots were saved as PNG files for further analysis and presentation. The creation of DAGs involved the API of gProfiler, which was used to fetch information

from the Gene Ontology (GO) database, providing comprehensive functional annotations for the gene lists, essential for over-representation analysis (ORA) (Kolberg et al., 2023). The use of these tools made it possible to organize and display the relationships among GO terms in a structured and visually appealing manner.

2.6 Quality Control and Validation

In the rigorous framework of this study, ensuring the reliability and accuracy of data analysis is paramount. Quality control and validation measures are extensively implemented across all stages of the research to maintain data integrity and validate the analytical methods used.

Quality control begins at the data acquisition stage, where initial checks on data completeness and correctness are performed. For the RNA-seq data from The Cancer Genome Atlas and expression data from the Human Protein Atlas, quality control measures include the verification of gene identifiers and inspection of expression level distributions to identify potential outliers or anomalies.

The normalization of data, a crucial step in quality control, is performed using the voom function of the *limma* package, which normalizes RNA-seq data to log2-counts per million to adjust for library size variations and other technical biases. This step is critical for ensuring that subsequent analyses such as differential expression are based on reliable and comparable data.

Validation of the analytical results involves several steps to confirm the biological plausibility and accuracy of the findings. Differential expression results, particularly those related to sex differences in gene expression, are subjected to rigorous validation. For example, the differential expression of genes between male and female samples in the full dataset and in the smoking dataset underwent additional verification. This

included matching the differentially expressed genes against established lists of identifiers known to differentiate male from female gene expressions. Genes such as DDX3Y and ZFY, known for their roles in sex determination and differentiation, showed expected patterns of higher expression in male samples. Conversely, genes like XIST, involved in X-chromosome inactivation, exhibited underexpression in male samples, aligning with known biological functions.

Furthermore, validation checks included the statistical analysis of log fold changes and adjusted p-values to confirm significant differences in gene expression. This comprehensive validation not only confirms the robustness of the dataset and the analytical procedures employed but also enhances the foundational knowledge necessary for further investigations into the genetic determinants of sex-based differences in lung cancer.

Additionally, the validation process extends to other sex-specific genes, particularly those located on the Y chromosome in male samples. The expression patterns observed were consistent with their chromosomal location and biological roles, providing further evidence of the accuracy of the differential expression analysis.

These quality control and validation processes ensure that the results presented are not only statistically significant but also robust and biologically meaningful. This rigorous approach enhances the credibility of the findings and supports the integrity of the research methodology employed in this study.

3 Results

This chapter presents a comprehensive analysis of cell-type specific gene expression in lung cancer, emphasizing the differential effects of demographic variables and smoking status. Utilizing the *CellTypeGenomics* package, the study meticulously examines data from The Human Protein Atlas (HPA) and The Cancer Genome Atlas (TCGA), enabling detailed exploration of how differentially expressed genes (DEGs) correlate with specific cell types and demographic factors.

The structure of the analysis is designed to directly address specific research questions posed in Chapter 1.10, detailing the relationship between gene expression and factors such as age, gender, smoking status, and cell type origins.

Chapter 3.1 introduces the *CellTypeGenomics* package and outlines its fundamental role in mapping Ensembl gene identifiers to cell types using data from the Human Protein Atlas. This chapter directly addresses the second research question regarding the cell type origins of DEGs, demonstrating how this package can link DEGs to specific cell types, which is essential for understanding the cellular dynamics of gene expression in lung cancer.

Chapter 3.2 presents the validation of the *CellTypeGenomics* package using gene lists from psoriasis studies. This chapter focuses on confirming the package's utility in linking gene expression to cell type data within the context of psoriasis, showcasing the versatility of the package.

Chapter 3.3 delves into the gene expression analysis related to tumor versus normal tissue using the full dataset from TCGA. This chapter is pivotal in addressing the first research question about identifying DEGs in lung cancer. It explores fundamental differences in gene expression between tumor tissues and normal counterparts, employing statistical

methods to identify DEGs and using precise cell-type mapping to relate specific cell types to cancerous behavior. The findings from this chapter contribute significantly to our understanding of the molecular alterations associated with tumor development.

Chapter 3.4 focuses on the gene expression analysis related to smoking using a subset of the TCGA dataset. This chapter investigates how smoking status influences gene expression in lung cancer, addressing the fourth research question regarding the impact of smoking on gene expression. By analyzing DEGs associated with smoking status, this chapter highlights key molecular pathways differentially affected by smoking, thus contributing to a more nuanced understanding of environmental influences on lung cancer pathology.

Chapters 3.3 and 3.4 collectively address the fifth research question about the validation and applicability of the *CellTypeGenomics* package in a cancer-specific context by utilizing the package's capabilities to attribute DEGs to specific cell types and pathways, thus validating its efficacy in handling large, complex genomic datasets. Throughout these chapters, both Gene Ontology (GO) and Reactome pathway analyses are extensively used to investigate the biological processes, cellular components, and molecular functions associated with the DEGs. This approach addresses the third research question by providing deeper insights into the molecular mechanisms underpinning lung cancer and elucidating how identified DEGs participate in critical biological pathways such as cell cycle regulation, DNA repair, immune responses, and cellular signaling.

Each chapter employs advanced bioinformatics tools and comprehensive genomic data to systematically address the intertwined dynamics of genetic, demographic, and environmental factors in lung cancer. Through this structured analysis, the study enhances understanding of lung cancer at the molecular level, potentially informing future research, diagnosis, and treatment strategies. This approach not only elucidates the role of specific cell types in lung cancer but also highlights the potential influence

of lifestyle factors such as smoking on the disease's genetic expression landscape.

3.1 The *CellTypeGenomics* Package

In this chapter, we present the results obtained using the *CellTypeGenomics* package, a Python tool developed for analyzing gene expressions concerning specific cell types. The *CellTypeGenomics* package was developed during the fall of 2023, as part of our specialization project with psoriasis genes from a study by Solvin et al. from 2023, used for validation (Føleide & Mittet, 2023). The package was further developed in 2024 to include more data sources. Initially, we outline the rationale behind the package's creation, emphasizing its need and utility in the realm of genomic research. The main results of our initial psoriasis study are presented in chapter 3.2.

The development of the *CellTypeGenomics* package was catalyzed by the necessity for an efficient method to pinpoint the cell-type origins of differentially expressed genes, leveraging numerical data from the Human Protein Atlas. Existing tools, while comprehensive, either lacked the specific functionality we required or did not adequately address our research needs. For instance, methods for automated cell type annotation on scRNA-seq data, such as those discussed by Pasquini et al. (2020), did not use Human Protein Atlas (HPA) for identifying cell type origins, a crucial aspect of our analysis. Similarly, traditional RNA sequencing and microarray techniques often fail to detect differentially expressed genes that are identifiable through single-cell RNA sequencing, as demonstrated by Chen et al. (2020). Furthermore, tools like the Single-cell Mapper (scMappR), which infer cell-type specificities of differentially expressed genes (Sokolowski et al., 2021), do not fully meet the analytical demands of our project. These gaps in existing methodologies underscore the

necessity for developing a specialized package tailored specifically to our research objectives.

3.1.1 Usage of the *CellTypeGenomics* Package

CellTypeGenomics is an open-source Python package, designed to assist researchers in exploring the cell-type origins of differentially expressed genes. The first version of the *CellTypeGenomics* package from 2023 utilized numerical data (proteinatlas.tsv) from the Human Protein Atlas (HPA) to generate a prioritized list of genes, potentially underscoring over-represented genes in the dataset. This data consisted of 16 742 unique genes, with a range of 50 to 3053 genes associated with each cell type across the data. The average amount of genes per cell type accounted to 571.7. Building upon this, the second version of the package from 2024 incorporated qualitative marker genes from both the Human Ensemble Cell Atlas (hECA) and HPA. As described in chapter 1.8.3, these marker genes are based on existing literature. The marker genes from HPA consisted of 148 unique genes, with a range of 1 to 7 genes per cell type and an average of 2.9 genes per cell type. The hECA marker genes consisted of 479 unique genes, a range of 1 to 34 genes per cell type and an average of 5.0 genes per cell type. In addition, there is an option to return tissue origins of genes using HPA data. The package is easily accessible on the Python Package Index (PyPI; https://pypi.org/) and can be installed with a simple command: *pip install celltypegenomics*. Its core functionality, the *celltypefishertest* function, processes a list of Ensembl IDs containing differentially expressed genes and returns a prioritized DataFrame, highlighting genes that are potentially over- or under-represented in certain cell types based on the overlap with the HPA or hECA data.

An example of how to specify qualitative markers from hECA in the *CellTypeGenomics* package is shown below in code.

```
result = celltypefishertest(list_of_ensembl_ids, alpha=0.05,
heca=True)
```

The default data source is numerical HPA marker genes, so then only `list_of_ensembl_ids` is needed as specified input. For qualitative markers from HPA, set `hpa_marker_genes=True`. To analyze tissue origins, set `tissue=True`. The alpha parameter can be adjusted from its default value of 0.05.

3.1.2 *CellTypeGenomics* Package Example

In this example, Genelist1 is read by the *CellTypeGenomics* package, returning a Pandas Dataframe. The following code allows for a list of Ensembl codes to be converted into a list of cell types. The top five most significant results are returned. These results are shown in Table 3.1.

```
with open('Genelist1.txt', 'r') as f:
genelist_content = f.read().splitlines()

import CellTypeGenomics

CellTypeGenomics.celltypefishertest(genelist_content).head(5)
```

Table 3.1: Dataframe returned by the *CellTypeGenomics* package for the example code

Cell Type	P-value	Odds ratio	Count in both	Count in genelist not cell type	Count in cell type not genelist	Count in neither	Adjusted p-value
Suprabasal keratinocytes	4.52e-70	26.86	78	122	464	19498	4.47e-68
Basal keratinocytes	7.36e-38	21.50	43	157	251	19711	3.64e-36
Squamous epithelial cells	1.28e-23	11.34	36	164	379	19583	4.25e-22
Serous glandular cells	1.44e-18	13.18	25	175	214	19748	3.58e-17
Basal respiratory cells	2.15e-18	11.38	27	173	270	19692	4.25e-17

3.1.3 *CellTypeGenomics* Package Overview

Figure 3.1 presents the workflow used by the *CellTypeGenomics* package, tracing the path from data acquisition at the Human Protein Atlas to the analytical results. It visually articulates the sequence of operations, clarifying complex methodologies for the audience. The diagram underscores stages such as data handling, computational analysis through Python, core functionalities of the package, and the availability of the tool in PyPI. Each aspect serves to deepen understanding of the research process and the execution of the study's methods.

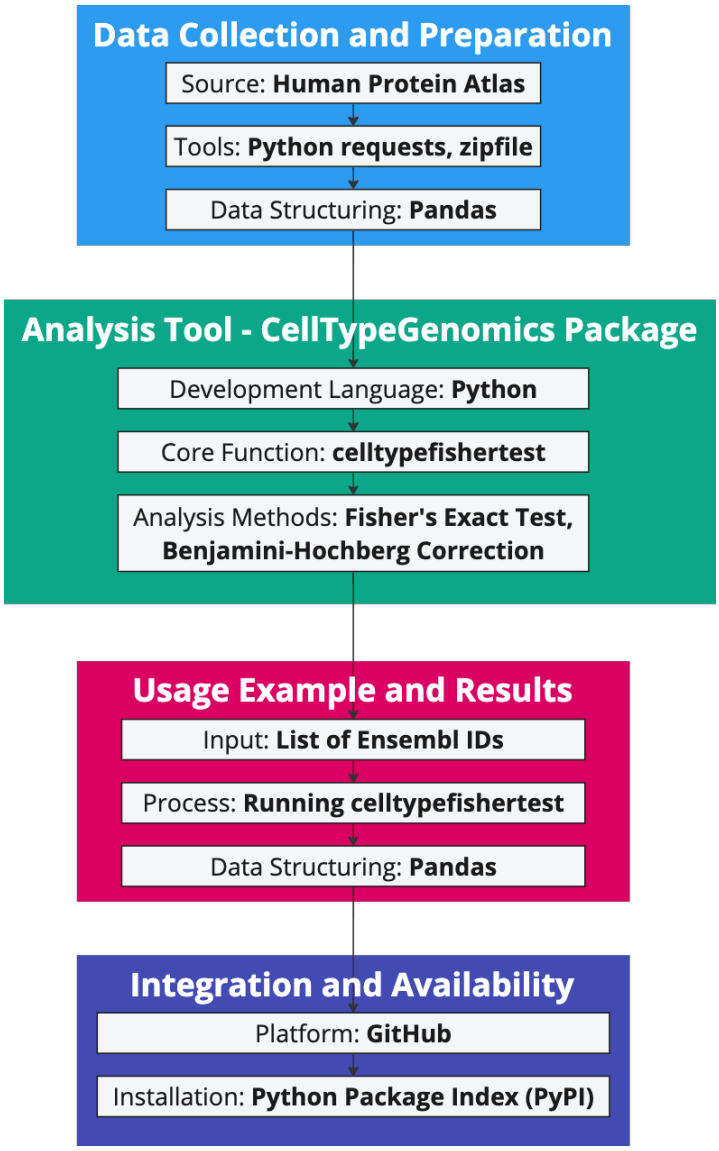


Figure 3.1: The streamlined process from data acquisition in Human Protein Atlas to the analytical output of the *CellTypeGenomics* package is depicted, demonstrating the stages of data handling, computational analysis, and result generation (Føleide & Mittet, 2023).

3.2 Validation of the *CellTypeGenomics* Package with Gene Lists from Psoriasis Data

This section examines the genetic basis of psoriasis using the *CellTypeGenomics* package to analyze two gene lists: Genelist1, which includes overexpressed genes comparing lesional psoriatic skin to healthy control skin, and Genelist2, which includes underexpressed genes comparing lesional psoriatic skin to healthy control skin. This approach aimed to explore the cellular dynamics of psoriasis.

The analysis revealed several key findings, visualized in Figure 3.2. Suprabasal keratinocytes, basal keratinocytes, and squamous epithelial cells showed strong associations with psoriasis in Genelist1. Notably, serous glandular cells were the only cell type found in both Genelist1 and Genelist2 with significant odds ratios, underscoring their central role in the disease's mechanisms.

Further analysis of Genelist2 revealed significant associations for cell types such as Leydig cells, fibroblasts, astrocytes, peritubular cells, and oligodendrocyte precursor cells. These distinct expression patterns between Genelist1 and Genelist2 underscore the complex regulatory mechanisms involved in psoriasis.

Additionally, the same analysis was conducted using qualitative markers from the Human Protein Atlas (HPA) and the Human Ensemble Cell Atlas (hECA). This provided further significant results. For hECA qualitative markers, monocytes (adjusted p-value = 0.0145, odds ratio = 37.9) and neutrophilic granulocytes (adjusted p-value = 0.0145, odds ratio = 15.6) showed significant associations with psoriasis. Intriguingly, Kupffer cells—typically resident in the liver—demonstrated a notably significant association (adjusted p-value = 0.0158, odds ratio = 202). The presence of liver-associated Kupffer cells among significantly associated cell types raises questions about systemic involvement and cross-talk between distant organ systems in psoriasis, suggesting a potentially broader

systemic component to the disease pathophysiology (Gelfand et al., 2006; Takeshita et al., 2017).

These findings align with existing literature on the role of keratinocytes in psoriasis. Keratinocytes are known to be central in the pathogenesis of psoriasis, interacting with immune cells and contributing to inflammation and abnormal skin cell proliferation. Studies highlight the role of cytokines such as IL-17 in inducing keratinocyte proliferation and differentiation abnormalities, which are hallmarks of psoriasis (Nestle, Kaplan, & Barker, 2009; Lowes, Suárez-Fariñas, & Krueger, 2014). Moreover, recent insights into the pathophysiology of psoriasis emphasize the intricate network of immune cells, including monocytes and neutrophils, and their contribution to the disease's chronic inflammatory state (Krueger & Bowcock, 2005). This broader understanding suggests that psoriasis may not only be a localized skin disorder but also involve multiple organ systems, potentially mediated by systemic immune responses.

Both the current analysis and the study by Solvin et al. (2023) acknowledge the significant role of keratinocytes in psoriasis. This analysis shows a notable presence of differentiated keratinocytes in lesional versus non-lesional and control skin, reinforcing the link between keratinocyte activity and psoriatic lesions. Similarly, the Solvin study, through cellular deconvolution, identified differentiated keratinocytes as the most prominent cell type among the DEGs in lesional psoriatic versus healthy control skin. This alignment underscores the pivotal role of keratinocytes in the pathophysiology of psoriasis.

Comparing the results of these two studies is challenging due to differences in analytical methods. The current study employs the *CellTypeGenomics* package, designed specifically to analyze cell type origins of differentially expressed genes, whereas the Solvin et al. study (2023) used CIBERSORTx for cellular deconvolution. CIBERSORTx, known for estimating cell proportions in mixed tissue samples, likely offers differing sensitivity and specificity in detecting cell type fractions

compared to the current method. These methodological differences can influence the detection and interpretation of minor cell populations or subtle expression changes.

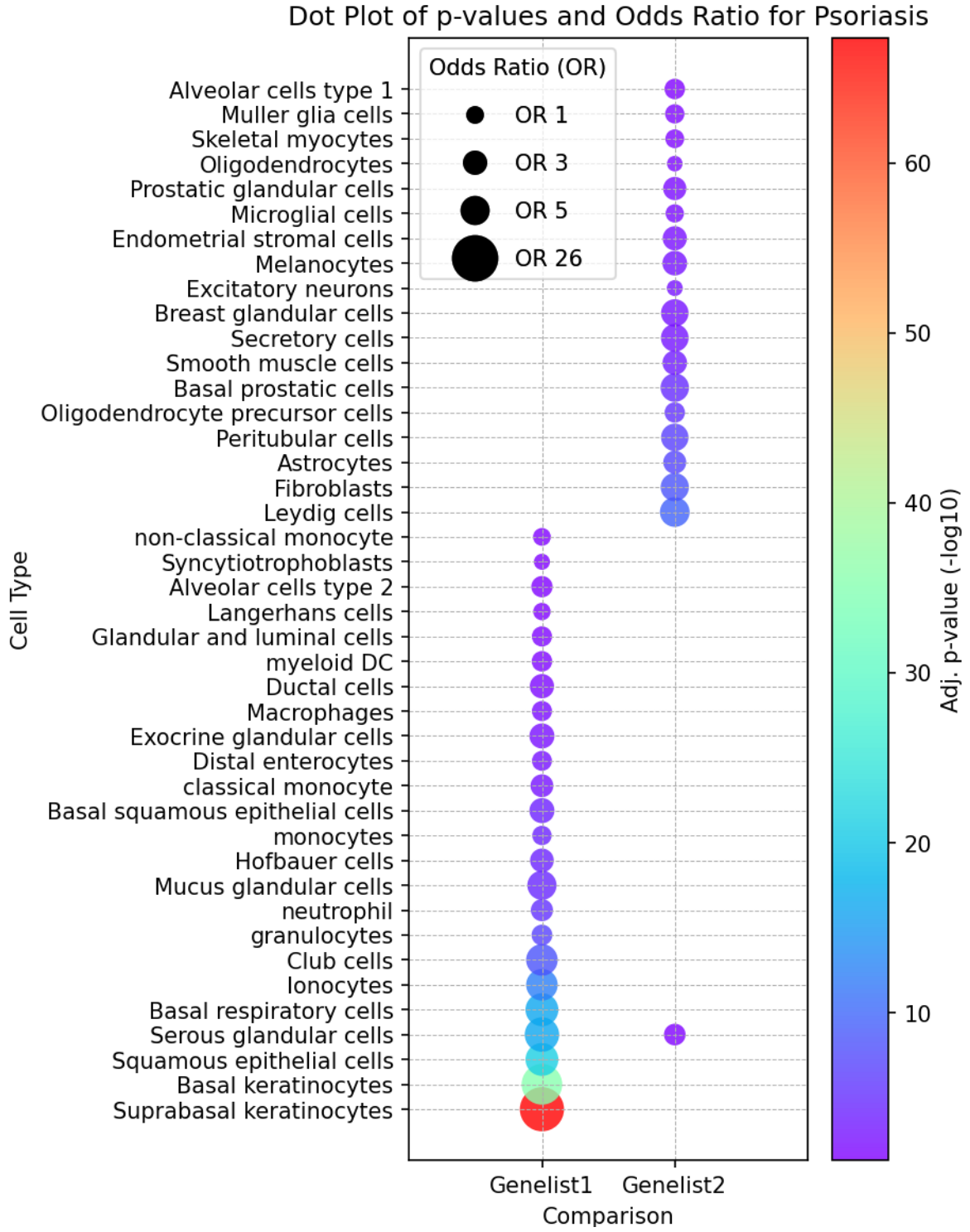


Figure 3.2: **Dot Plot of p-values and odds ratios for psoriasis Genelist1 and Genelist2**. Symbol size vary with odds ratios (OR). The color gradient from red to purple represents the associated metric (e.g., p-value, -log10 scale), highlighting the significance of each observation. Red dots indicate the most significant observations, with purple representing the least significant within the set thresholds.

3.3 Gene Expression Analysis Related to Tumor vs Normal Tissue

In this section, we examine the genetic underpinnings of lung cancer using data from The Cancer Genome Atlas (TCGA). The primary objective is to analyze cell-type specific gene expressions within lung cancer tissues to elucidate the complex interactions between genetic factors and the disease.

The statistical overview of the dataset includes an UpSet plot to illustrate the distribution and interconnectivity of significant gene clusters. This plot highlights the balance between overexpressed and underexpressed genes within the dataset, offering a clear visualization of differential gene expression patterns.

Further, we investigate marker genes from the Human Protein Atlas (HPA) and the Human Ensemble Cell Atlas (hECA). These marker genes are pivotal in identifying cell types exhibiting significant differential expression. The analysis of these genes provides insights into the cellular composition of lung cancer tissues, enhancing our understanding of the disease's molecular landscape.

Reactome pathway analysis is conducted to map the significant gene clusters to biological pathways, providing a deeper understanding of the functional implications of the observed gene expression changes. This analysis is complemented by Gene Ontology (GO) classifications, which further categorize the significant genes into biological processes, cellular components, and molecular functions.

A Directed Acyclic Graph (DAG) is employed to visualize the hierarchical relationships and biological pathways, emphasizing the interconnectivity and shared functions relevant to lung cancer. This graphical representation aids in identifying key pathways and their roles in disease progression.

To ensure the robustness of our findings, a sanity check is performed, validating the consistency and reliability of the differential gene expression results. This step is crucial for confirming the accuracy of our computational analyses.

Additionally, dot plots of differential gene expression in both tissues and cell types are generated. These plots display the odds ratios and adjusted p-values, highlighting the statistical significance of gene expression across various conditions.

Overall, this section leverages advanced bioinformatics tools and comprehensive datasets to dissect the molecular mechanisms of lung cancer, providing a solid foundation for future research and potential therapeutic interventions.

3.3.1 Lung Cancer Full Dataset Statistics

This section provides a comprehensive overview of the demographic and clinical characteristics of the lung cancer dataset obtained from The Cancer Genome Atlas (TCGA). This dataset includes detailed information on gender distribution, age statistics, and other relevant clinical variables, forming a robust foundation for subsequent genomic analyses.

The dataset comprises 1879 lung cancer cases, with 1165 males and 714 females. The mean age for males is 63.31 years (SD = 15.58), while for females it is 61.69 years (SD = 18.11). The age at diagnosis for the overall cohort shows a mean of approximately 62.7 years and a median of approximately 65.9 years. Additionally, the metadata includes comprehensive records for a total of 1997 samples, providing a broad base for in-depth analysis of lung cancer. This extensive dataset supports a detailed exploration of the molecular mechanisms underlying the disease, considering both demographic and clinical variables.

A detailed examination of the dataset reveals that male patients tend to be older than female patients across most cancer stages. The age

difference were most notable in Stage IV, where males have a mean age of 63.6 years compared to 60.1 years for females, with 29 male and 16 female patients.

These demographic statistics highlight the age distribution and gender composition within the lung cancer cohort, which are important for understanding the patient population and ensuring the robustness of the genomic analyses that follow. The age and stage-specific trends underscore the importance of considering demographic factors in lung cancer research and may inform targeted strategies for early detection and treatment. This analysis provides essential context for the genomic studies conducted in subsequent sections, emphasizing the need to account for demographic and clinical variability.

3.3.2 Differentially Expressed Genes of the Full Dataset

The UpSet plot in Figure 3.3 provides a comprehensive visualization of the distribution and interconnectivity of significant gene clusters based on their differential expression in the lung cancer dataset. This plot is essential for elucidating the complex relationships among gene clusters, particularly concerning overexpression and underexpression across different conditions.

The UpSet plot employs color coding to distinguish between underexpressed and overexpressed genes: red indicates underexpressed genes, while green signifies overexpressed genes. The left histogram categorizes gene clusters by size, displaying the number of elements in each cluster. The accompanying bar chart at the top quantifies the elements per cluster, emphasizing the balance between overexpressed and underexpressed genes.

A detailed examination of the dataset reveals several key findings. In the Tumor vs. Normal expression patterns, there are 893 genes significantly overexpressed in tumor tissues compared to normal tissues

(TumorVsNormal_up), with 853 of these genes being exclusively overexpressed in tumor tissues. In contrast, TumorVsNormal_down includes 3069 genes exclusively underexpressed in tumor tissues compared to normal tissues from a total of 3122 significantly underexpressed genes.

Regarding age-related expression patterns, the analysis identified 28 genes significantly overexpressed in relation to age (Age_up) and 43 genes significantly underexpressed (Age_down).

The UpSet plot further highlights the interconnectivity among different gene clusters, illustrating shared pathways and mechanisms between TumorVsNormal and Age-related expression patterns. This interconnectivity offers insights into potential therapeutic targets and emphasizes the importance of considering both tumor-specific and age-related factors in lung cancer research.

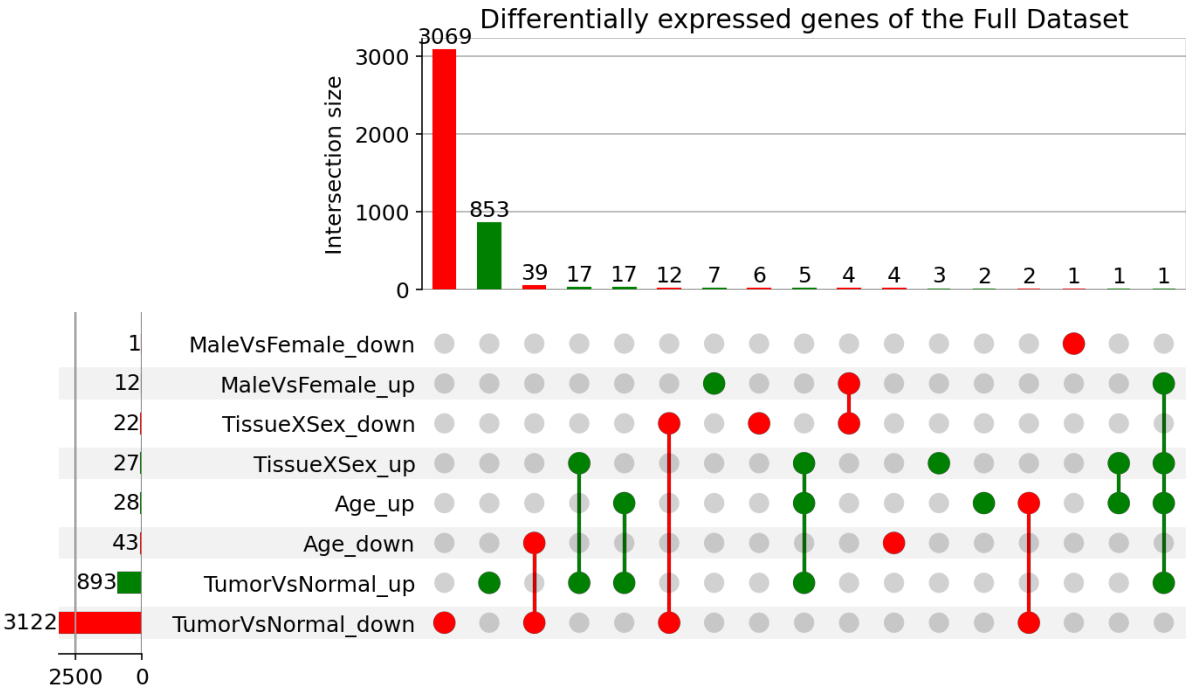


Figure 3.3: **UpSet Plot of Significant Gene Interactions for Full Dataset**. This plot illustrates the distribution and interconnectivity of significant gene interactions based on differential expression—red indicating underexpressed and green signifying overexpressed genes. The left histogram categorizes clusters by size, while the network diagram displays their relationships. The top bar chart quantifies the elements per cluster, highlighting the balance between overexpressed and underexpressed genes. The plot emphasizes distinct expression patterns in tumor versus normal tissues and age-related changes, showcasing intricate relationships among gene interactions.

3.3.3 Differential Expression Analysis of Marker Genes for Full Dataset

The analysis of qualitative marker genes from the Human Protein Atlas (HPA) and the Human Ensemble Cell Atlas (hECA) identifies significant differential gene expressions, categorized as either over- or underexpressed. Figure 3.4 illustrates a comparative analysis of marker genes across various conditions, with larger symbols indicating higher odds ratios, suggesting stronger associations between gene expression and specific cell types. Upward-pointing triangles denote overexpressed genes, while downward-pointing triangles indicate underexpressed genes. A gradient bar reflects statistical significance.

The study spans age-related changes, tissue-specific expressions influenced by sex, and comparisons between tumor and normal tissues. Notably, Alveolar cells type II consistently show underexpression in the TissueXSex condition for both hECA and HPA. This cell type also exhibits significant underexpression in the TumorVsNormal condition for hECA, highlighting their sensitivity to various biological influences, including sex-specific factors and tissue-specific changes. This contrasts with findings by Chaudhary et al. (2023), where alveolar type II cells, upon $KRAS^{G12D}$ activation, show enhanced plasticity and tumor-initiating capabilities, suggesting a differential expression profile under oncogenic stress compared to non-cancerous conditions.

Bronchial epithelium basal cells display highly significant p-values in tumor versus normal tissue comparisons, suggesting their important role in tumor biology and potential as markers for cancer progression. These cells are overexpressed in TumorVsNormal, TissueXSex, and Age conditions for HPA.

Alveolar cells type I are significantly underexpressed in the TissueXSex and TumorVsNormal conditions for hECA, pointing to crucial regulatory mechanisms affecting their expression. Macrophages and endothelial cells show significant underexpression in TumorVsNormal conditions for HPA

and hECA respectively, indicating their roles in tumor immune evasion and the vascular changes associated with tumor growth.

This comprehensive analysis underscores the complex nature of gene regulation across various biological contexts and lays a strong foundation for future research aimed at uncovering the underlying mechanisms. Integrating these findings with additional omics data such as proteomics and metabolomics will enhance our understanding of the regulatory networks involved.

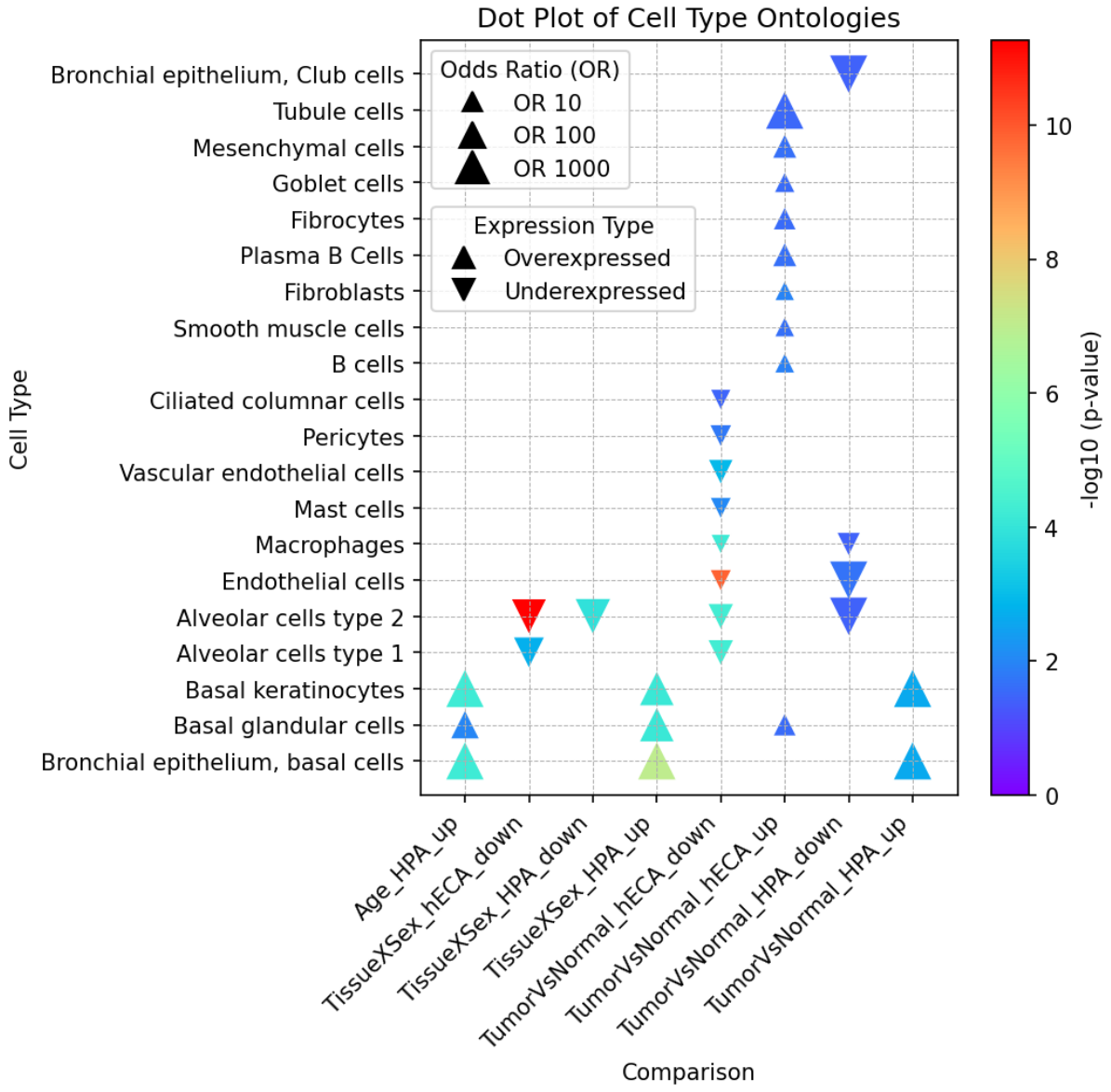


Figure 3.4: **Dot Plot for full dataset displaying differential expression of qualitative marker genes in cell types from the Human Protein Atlas (HPA) and Human Ensemble Cell Atlas (hECA)**. Symbol sizes indicate odds ratios (ORs), with direction denoting overexpression (upward triangles) and underexpression (downward triangles). The color gradient bar shows the statistical significance (-log10 p-value).

3.3.4 Reactome Pathway Analysis for Full Dataset

A comprehensive analysis using the Reactome Pathway database was conducted on the full dataset to identify significant pathways associated with differentially expressed genes in lung cancer. This analysis focused particularly on variations resulting from tumor versus normal tissue comparisons and differences influenced by age and sex. Figure 3.5 presents a detailed comparative analysis, visually demonstrating how these conditions affect gene expression.

The analysis revealed overexpressed pathways in tumor conditions associated with the cell cycle that are highly significant, underlining their crucial role in the progression of lung cancer. The genes involved in these pathways have potential as biomarkers for detecting and monitoring the disease.

Underexpressed age-related changes were also notable, particularly for the terms DNA Damage/Telomere Stress Induced Senescence, DNA methylation and RNA Polymerase I Promoter Opening. These findings suggest a potential dysregulation in the cellular aging processes that could influence tumorigenesis in lung tissue. The underexpression of genes associated with DNA damage response and telomere maintenance could imply a reduced capacity for senescence induction, potentially allowing cells with damaged DNA to proliferate instead of entering a senescent state. This pathway is crucial as it acts as a natural barrier against cancer by stopping the proliferation of cells that have acquired hazardous levels of DNA damage (Blackburn, 2005). Alterations in DNA methylation patterns are a hallmark of aging and cancer, affecting gene expression without altering the DNA sequence. The observed underexpression related to DNA methylation processes might indicate an aberrant epigenetic landscape, which is critical in the regulation of gene expression and maintenance of genomic stability (Jones & Baylin, 2007).
RNA Polymerase I Promoter Opening is essential for the transcription of ribosomal RNA (rRNA), fundamental for ribosome biogenesis and overall

protein synthesis. Underexpression in this pathway might suggest a compromised capacity for cellular protein synthesis, impacting cell growth and proliferation, critical aspects of cancer development and progression. Grummt (2003) highlights the complex regulation of RNA Polymerase I, underscoring its pivotal role in cellular growth mechanisms, which could be disrupted in cancerous tissues (Grummt, 2003).

The pathways Formation of the cornified envelope and Keratinization were significantly overexpressed in age, sex-related comparisons and tissue-sex comparisons. The formation of the cornified envelope involves the creation of a protective barrier in the outer layer of the skin and other tissues. Dysregulation in differentiation processes like cornification can indicate broader epithelial changes relevant to cancer biology, including lung cancer (Carregaro et al., 2013). Keratinization, the process by which keratin proteins form protective layers in epithelial cells, can be a marker of epithelial cell dysregulation, a characteristic of many carcinomas, including lung cancer (Heryanto & Imoto, 2023).

Pathways such as Surfactant Metabolism, Diseases Associated with Surfactant Metabolism and Defective CSF2RA causes SMDP4 were significantly underexpressed for tissue-sex comparisons. Surfactant Metabolism is crucial for the proper functioning of lung tissues (Lopez-Rodriguez et al., 2016). It is a complex mixture of lipids and proteins that lines the alveolar epithelium. At the air-liquid interface, the surfactant lowers surface tension, avoiding alveolar collapse and reducing the work of breathing (Lopez-Rodriguez et al., 2016). Surfactant deficiency can result in diseases such as pulmonary alveolar proteinosis (Lopez-Rodriguez et al., 2016). CSF2RA is a gene that encodes a critical protein involved in immune and inflammatory responses. Defects in the CSF2RA gene can cause Pulmonary Surfactant Metabolism Dysfunction 4 (SMDP4), also known as congenital pulmonary alveolar proteinosis (Whitsett et al., 2015). This is a rare lung disorder due to impaired surfactant homeostasis characterized by alveoli filling with floccular material (Whitsett et al.,

2015). The connection between defects in CSF2RA and lung cancer is unknown.

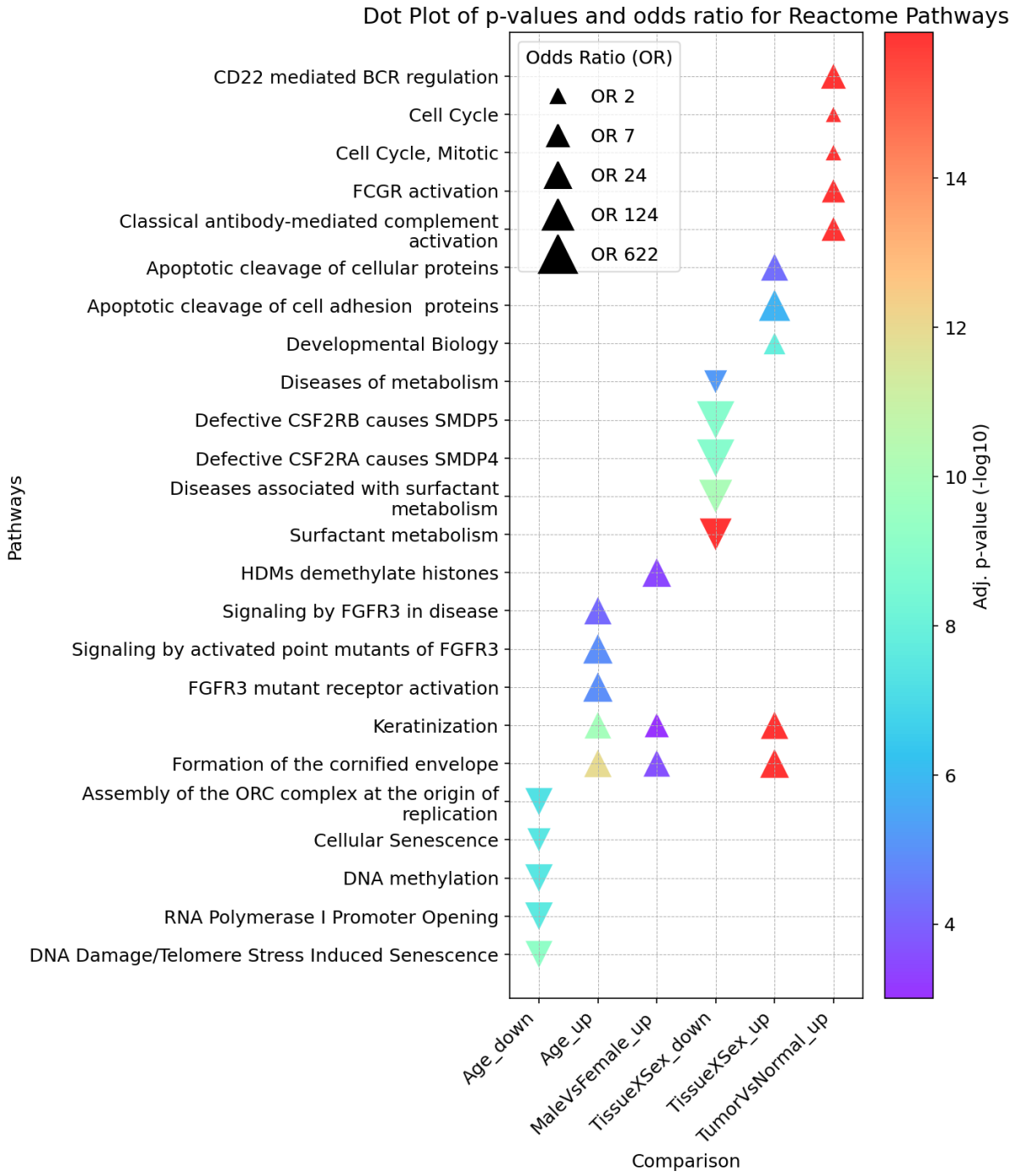


Figure 3.5: **Dot Plot of Reactome Pathways for Full Dataset**. Displays the top 5 significant pathways for various conditions. Symbol sizes indicate odds ratios (ORs), with upward triangles for overexpression and downward triangles for underexpression. The color gradient bar represents the statistical significance (-log10 p-value).

3.3.5 Biological Processes (BP) from Gene Ontology (GO)

An enrichment analysis of Gene Ontology (GO) Biological Processes (BP) was undertaken to investigate the broader biological implications of differentially expressed genes identified in the study. Figure 3.6 visually maps these associations across various conditions including age, sex, and tumor presence.

The analysis underscored several key biological processes exhibiting significant expression patterns. Notably, mitotic processes such as the Mitotic Cell Cycle and Mitotic Nuclear Division were markedly overexpressed in comparisons between tumor and normal tissues. These processes demonstrated exceptionally significant adjusted p-values and high odds ratios, suggesting a pronounced role in tumor biology (Long et al., 2019).

Processes like Cell Migration, Cell Motility, and Locomotion were significantly underexpressed in comparisons between tumor and normal tissues. In the context of lung cancer, it has been observed in mice that intrinsic Interleukin (IL)-15 in cancer cells promotes cell motility and migration, but exogenous IL-15 inhibits these processes (Hu et al., 2024). Overexpression of Regulator of Chromosome Condensation 2 (RCC2) has also been linked to enhanced cell motility in lung adenocarcinoma (Pang et al., 2017). However, the specific underexpression of these processes in lung cancer requires further investigation. It’s important to note that the underexpression of these processes could potentially impact the metastatic capabilities of the cancer cells, as these processes are crucial for tumor progression and spread (Hu et al., 2024; Pang et al., 2017). Further research in this area could provide valuable insights into the development and progression of lung cancer.

For a more comprehensive understanding, Appendix A.6 presents a detailed table listing enriched biological processes. This appendix includes significant findings such as the overexpression of the Immunoglobulin Mediated Immune Response and B Cell Mediated Immunity in tumor

versus normal comparisons, highlighting these processes as potential targets for immune-based therapies. Additionally, Vasculature Development was notably underexpressed in tumor versus normal tissue comparisons, which might signify compromised vascular processes within tumor environments, suggesting its potential as a biomarker for cancer progression (Yang et al., 2021).

Additionally, the study identified significant overexpression of developmental processes including Epidermis Development and Intermediate Filament Organization in both tissue-sex and age categories. This repeated overexpression emphasizes their critical roles in physiological adaptations related to aging and sex-specific biological differences, highlighting significant regulatory interactions (Sharma et al., 2019).

Terms related to the packaging of DNA, such as Nucleosome Assembly, Nucleosome Organization and Protein Localization to Chromatin were shown to be significantly underexpressed associated with aging. These alterations in chromatin structure and function can disrupt the normal regulation of gene expression, contributing to cancer development. Histone modifications, which play a pivotal role in nucleosome assembly and chromatin dynamics, are particularly implicated in this process. Changes in these modifications can affect DNA replication, repair, and overall genomic stability, which are critical factors in cancer biology (Zhang et al., 2023; Prado & Maya, 2017). Furthermore, the aging process itself influences these epigenetic modifications, potentially increasing the vulnerability to cancer as these regulatory mechanisms become less effective. The decoupling of DNA synthesis from nucleosome assembly, a phenomenon more frequently observed in aging cells, contributes to genomic instability—a key feature in cancer progression (Prado & Maya, 2017).

Terms related to keratinization and skin, such as Keratinization, Keratinocyte Differentiation and Epidermis development were shown to be

significantly overexpressed associated with aging. These processes are essential for the maintenance of skin integrity and are intricately linked to the pathophysiological changes observed in Lung Squamous Cell Carcinoma (LUSC). For instance, keratinization is a key histopathological feature of LUSC, where epithelial cells produce keratin as a protective response to external harmful substances. This response is particularly critical in lung tissues exposed to carcinogens like tobacco smoke, which is a common risk factor for LUSC (Heryanto & Imoto, 2023). Research also indicates that proteins like Receptor-Interacting Protein Kinase 4 (RIPK4), which are involved in keratinocyte differentiation, play significant roles in the carcinogenesis process, particularly in Squamous Cell Carcinomas (SCCs), including those of the lung. RIPK4 is implicated in various signaling pathways that regulate epidermal homeostasis and differentiation, and mutations in this protein have been associated with different forms of SCC (Xu et al., 2020).

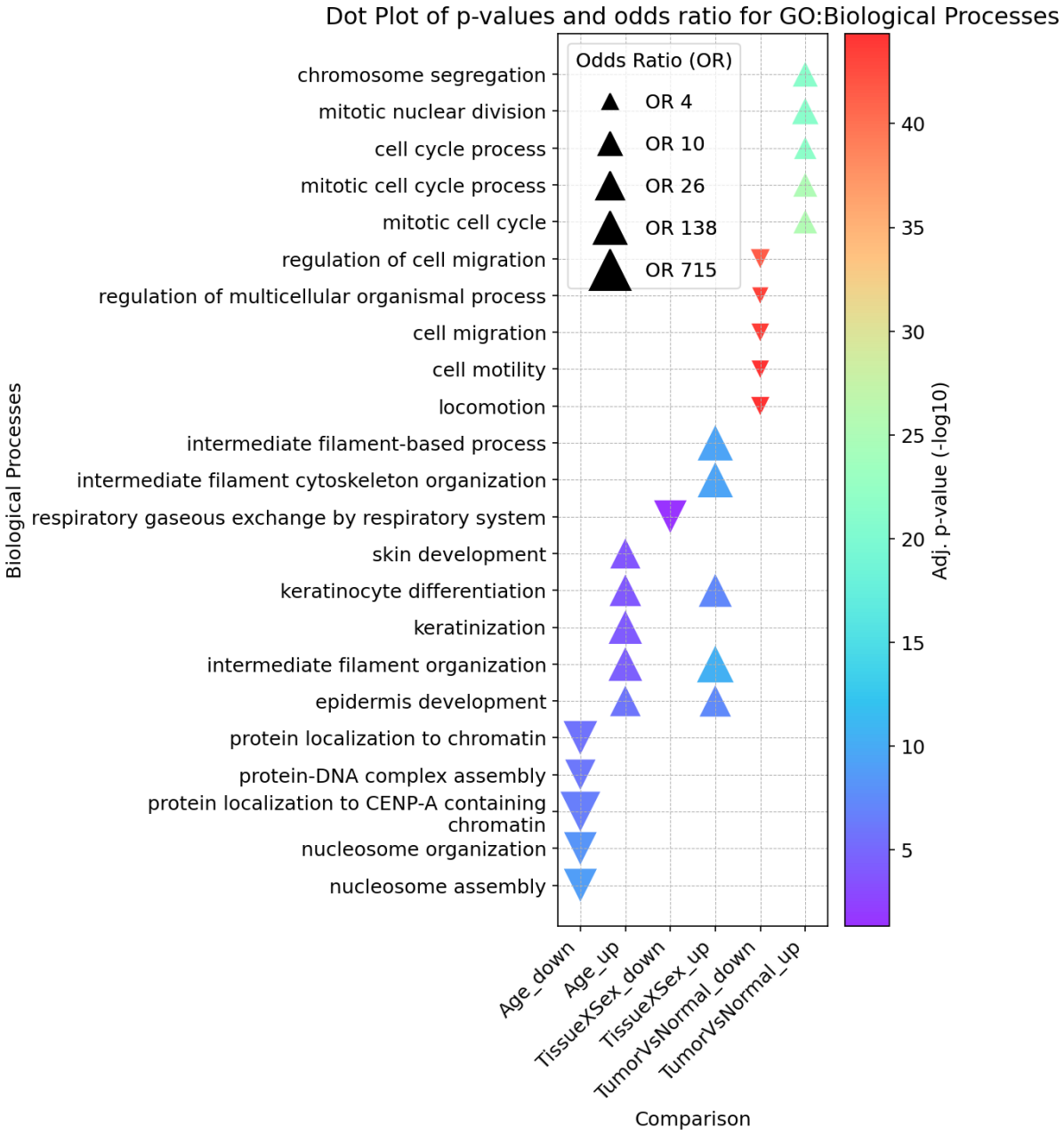


Figure 3.6: **Dot Plot of GO Biological Processes for Full Dataset**. This plot organizes Gene Ontology (GO) Biological Processes along the y-axis, each linked to specific biological conditions such as age, sex, and tumor presence. Vertical stacks of symbols illustrate the involvement within each condition, with the size of each symbol indicating the odds ratio, reflecting the strength of association. The color gradient from purple to red represents the adjusted p-values (-log10), highlighting the statistical significance of gene involvement in each condition.

To explore the relationships between the overexpressed and underexpressed biological processes across multiple contrasts, Directed Acyclic Graphs (DAGs) were constructed. A DAG is a graphical tool that illustrates the hierarchical relationships between different biological processes identified through Gene Ontology (GO) analysis. While DAGs were generated for various contrasts, Figure 3.7 shows a representative DAG for the overexpressed TumorVsNormal contrast. This DAG provides a structured visualization of how the dysregulated processes in tumor samples, compared to normal tissue, are interconnected within a network of biological processes. Each node in the DAG represents a GO term, with arrows indicating parent-child relationships that move from more specific to more general terms. Key details of the nodes include the GO Term ID, level (L), depth (D), and descendant count (d).

The DAG contains three main branches. The left branch contains nodes related to organization within the cell and more specifically nuclear division. The middle branch is the shortest, only containing the cell cycle and more specifically the mitotic cell cycle. The right branch also deals with the cell cycle, but rather with the cell cycle process and the mitotic cell cycle process. The DAG demonstrates that many significant GO Biological Process terms are related, forming a network of interconnected processes. An example of the connectivity within the DAG can be seen with the term Cell Cycle Process (GO:0022402), which connects to the more specific processes Mitotic Cell Cycle Process (GO:1903047) and Mitotic Nuclear Division (GO:140014). This indicates a functional progression from broad to specific cell cycle processes, underscoring the interrelatedness of these biological processes.

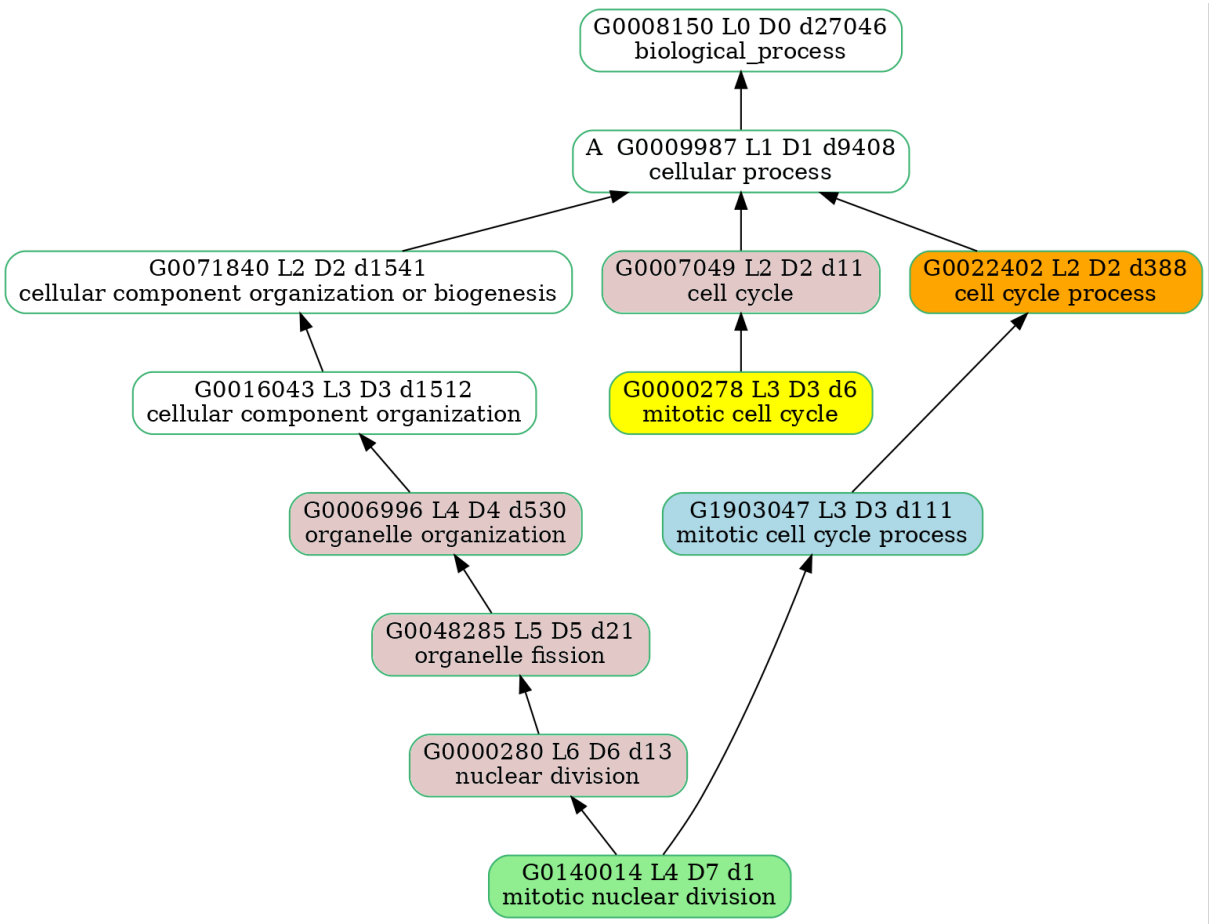


Figure 3.7: **Directed Acyclic Graph (DAG) illustrating the top four most significant Biological Process Gene Ontology (GO:Biological Process) results for the overexpressed TumorVsNormal contrast.** The arrows in the DAG point from child to parent, denoting a progression from more specific to more general terms. This visualization highlights the hierarchical relationships and biological pathways involved, emphasizing the interconnectivity and shared biological functions relevant to the overexpressed TumorVsNormal contrast. The yellow node is the most significant result, the light blue second most significant, the orange third most significant and the light green fourth most significant. The light red nodes represent remaining nodes in the DAG that are found to be significant GO:Biological Process results for the overexpressed TumorVsNormal contrast, but are not among the top four most significant. The white nodes are GO:BP results that are not found to be significant for the overexpressed TumorVsNormal contrast, but are part of the hierarchical structure of the DAG. Each of the nodes contain a unique GO ID, level (*L*) indicating the minimum path from the top root, depth (*D*) indicating the maximum path from the top root term and descendant count (*d*) indicating the total number of GO terms below the given node from the GO hierarchy structure (not shown in this DAG, but a part of the underlying Open Biological and Biomedical Ontologies file) (Klopfenstein et al., 2018). The letters A, B and C at the second most top nodes represent aliases for depth-01 GO terms, used to provide the general location in the GO DAG of any GO term. They stand for cellular process, biological regulation and metabolic process, respectively (Klopfenstein et al., 2018).

A G0009987 L1 D1 d9408
cellular process

3.3.6 Cellular Components (CC) from Gene Ontology (GO)

The Gene Ontology (GO) Cellular Components (CC) enrichment analysis was conducted to assess the broader biological implications of differentially expressed genes within this study. Figure 3.8 visually illustrates these associations across varying conditions such as age, sex-specific tissues, and the presence of tumors.

The analysis demonstrated significant differences in the expression of cellular components when comparing tumor tissues to normal tissues. Notably, components such as the immunoglobulin complex, extracellular region, extracellular space, extracellular exosome, and extracellular vesicle were markedly overexpressed in tumors. The immunoglobulin complex, which exhibited the highest adjusted p-value and an odds ratio of 103, was identified as particularly significant, indicating its crucial role in tumor biology and its potential as a therapeutic target.

In contrast, components including the cell periphery, plasma membrane, cell surface, extracellular region, and collagen-containing extracellular matrix were found to be underexpressed in tumors compared to normal tissues. These components showed lower odds ratios, suggesting a diminished presence in tumor tissues, which may provide insights into the structural and functional alterations occurring within the tumor microenvironment.

In the category of sex-specific tissues, components such as the cornified envelope, keratin filament, intermediate filament, intermediate filament cytoskeleton, and desmosome showed significant overexpression. The cornified envelope, in particular, displayed an exceptionally high odds ratio of 706, underscoring its critical role in gene expression related to sex differences.

Conversely, the underexpressed sex-specific tissue category included components such as the lamellar body, multivesicular body, alveolar lamellar body, multivesicular body lumen, and vesicle. Notably, the

lamellar body and alveolar lamellar body exhibited very high odds ratios of 1308 and 1413, respectively, highlighting the distinct associations of these components with sex-specific tissue differences.

Regarding age-related changes, overexpression was observed in extracellular components such as the extracellular space and extracellular region, alongside the immunoglobulin complex and circulating immunoglobulin complex, and the cornified envelope. These components' enhanced expression suggests their involvement in physiological processes associated with aging.

In the age category with underexpressed components, the nucleosome, nucleolus, CENP-A containing chromatin, CENP-A containing nucleosome, and chromosome centromeric core domain were notably underexpressed. These components, which displayed high odds ratios, are indicative of their reduced presence, emphasizing their potential roles in age-related gene expression changes.

This comprehensive analysis of cellular components and their expression across various conditions offers valuable insights into the complex regulatory mechanisms associated with different cellular components. By elucidating significant associations and expression patterns, the study contributes to a deeper understanding of the biological processes involved in aging, sex-specific differences, and tumor biology.

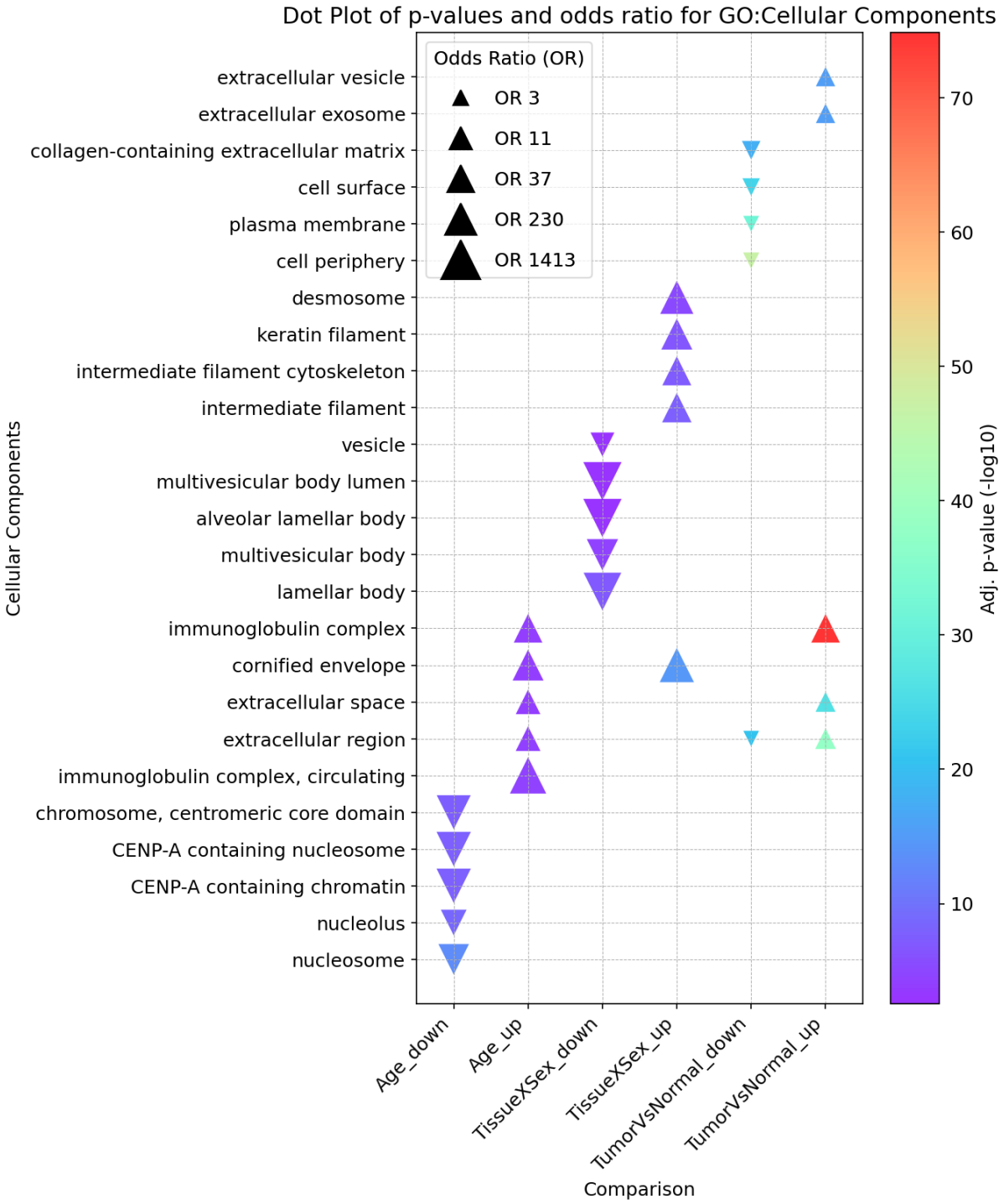


Figure 3.8: **Dot Plot of GO Cellular Components for Full Dataset**. This plot show Gene Ontology (GO) Cellular Components associated with age, sex-specific tissues, and tumor presence. Symbols represent odds ratios (OR), with size indicating the magnitude of the OR. Upward-pointing triangles denote overexpressed genes, while downward-pointing triangles indicate underexpressed genes. The color gradient from purple to red represents adjusted p-values (-log10), with red marking the most statistically significant findings.

3.3.7 Molecular Functions (MF) from Gene Ontology (GO)

The Gene Ontology (GO) Molecular Functions (MF) enrichment analysis elucidated the roles of differentially expressed genes in lung cancer, focusing on the impact of variables such as age, sex, and tumor presence. Visual representation of these associations is provided in Figure 3.9, which highlights the molecular functions implicated in various clinical scenarios.

The analysis comparing tumor tissues to normal tissues revealed significant overexpression of molecular functions, including antigen binding, cell adhesion molecule binding, and protein binding. The molecular function of antigen binding, notable for having the most significant adjusted p-value and an odds ratio of 56, is essential for immune recognition. The prominent role of this function in tumor biology highlights its potential as a therapeutic target, especially for strategies aimed at enhancing the immune response in cancer immunotherapy.

Additionally, the overexpression of cell adhesion molecule binding is significant as these molecules facilitate not only cell-cell and cell-matrix interactions but also modulate the signaling pathways that drive tumor progression. Understanding these functions is crucial for comprehending the invasive capacity of cancer cells, offering potential targets to inhibit metastasis (Harjunpää et al., 2019; Mrozik et al., 2018; Neophytou, 2021).

Conversely, molecular functions such as signaling receptor binding and integrin binding were found to be underexpressed in tumors, suggesting a suppression of specific cellular signaling and regulatory mechanisms within the tumor microenvironment. This suppression may contribute to the pathological state of cancer cells.

We found significant overexpression in male versus female samples, such as histone H3 demethylase activity, histone demethylase activity, protein demethylase activity, general demethylase activity, and 2-oxoglutarate-dependent dioxygenase activity. These findings suggest that

demethylation processes, especially those involving histone modifications, are more pronounced in males, potentially influencing gene expression regulation and contributing to sex differences in disease susceptibility and progression.

Analysis of sex-specific tissues revealed high overexpression of molecular functions like indanol dehydrogenase activity and phenanthrene 9,10-monooxygenase activity, indicating their significant roles in sex-related biological processes. These findings imply that sex differences might influence specific metabolic pathways, which are differentially activated in lung cancer.

The study also highlighted functions associated with age-related changes. Functions such as structural molecule activity and immunoglobulin receptor binding were predominantly overexpressed in older individuals, likely involved in physiological or pathological processes associated with aging. In contrast, functions like structural constituent of chromatin and nucleic acid binding were underexpressed, suggesting a decrease in genomic stability and transcriptional activity with age.

The upregulation of structural constituents of skin epidermis, such as keratins, in tumors relative to normal tissues suggests an epithelial-to-mesenchymal transition (EMT). This process is pivotal for cancer progression, providing epithelial cells with mesenchymal features that enhance their motility and invasiveness, a process elaborated on by Neophytou (2021). Interestingly, this upregulation also correlates with age and is discernible in tissues specific to different sexes, underscoring the complex interplay of various factors in cancer development and progression.

Our study observed both upregulation and downregulation of structural molecule activity, reflecting the complex nature of cancer and aging processes. The adjusted p-values indicate that downregulation is much more pronounced than upregulation. This duality might reflect a balance

between protective adaptations and detrimental changes in cellular structures, highlighting the intricate interplay of molecular functions in lung cancer's pathology.

This enriched understanding of molecular functions through GO analysis underscores the complex interplay of genetic expressions influenced by demographic and pathological factors, offering pathways for targeted therapeutic interventions and deepening the comprehension of lung cancer biology.

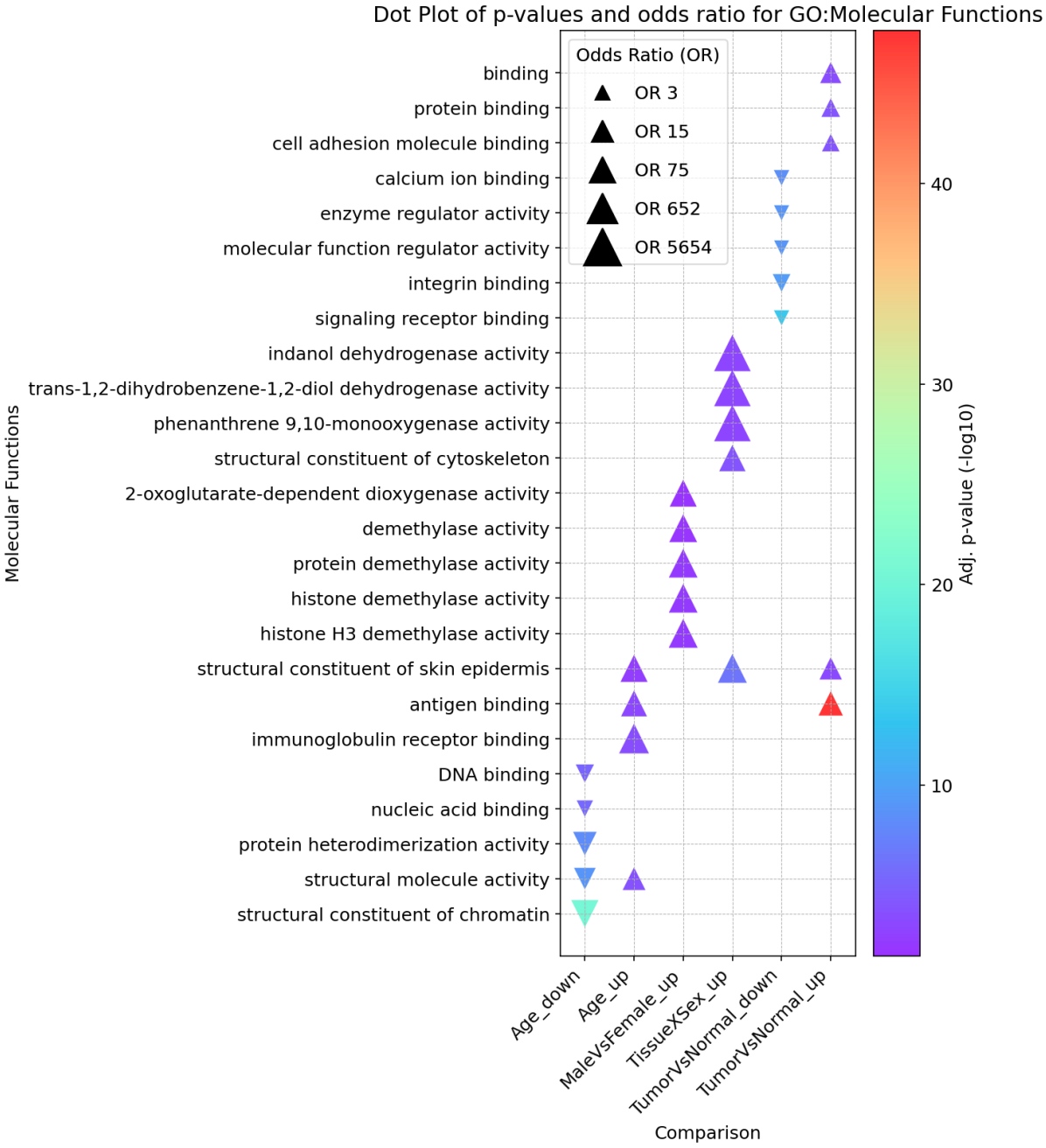


Figure 3.9: **Dot Plot of GO Molecular Functions for Full Dataset**. The dot plot shows Gene Ontology (GO) Molecular Functions associated with age, gender, sex-specific tissues, and tumor presence. Upward-pointing triangles denote overexpressed genes, while downward-pointing triangles indicate underexpressed genes. Larger symbols indicate higher odds ratios (OR), while a color gradient from purple to red represents the significance of p-values (-log10 scale).

3.3.8 Validation of Differential Gene Expression Between Male and Female Samples in Full Dataset

In the course of ensuring the reliability of our regression results, a sanity check was performed by verifying the differential expression of genes between male and female samples for the full dataset. The dataset underwent a filtration process to include only those genes previously analyzed in the MaleVsFemale regression study. A meticulous search was conducted to match these genes against established lists of identifiers known to differentiate male from female gene expressions.

The differential expression of selected genes between male and female samples was documented as follows. Genes such as DDX3Y, and ZFY, among others, showed higher expression levels in male samples, indicative of overexpression, whereas genes like XIST exhibited underexpression in the same group, as can be seen in Table 3.2. The observed expression levels for each gene are supported by a comprehensive statistical analysis noting significant differences in expression levels, as evidenced by log fold change values and adjusted p-values, some of which reached levels described as infinite due to their extremity.

The roles of these genes in sex-specific biological processes are well-documented, thereby making them reliable markers for gender-specific expression patterns. For instance, the gene DDX3Y, which is known for its role in RNA helicase activity critical for RNA processing, exhibited a log fold change of 4.90, indicating significant overexpression in males—a finding that is consistent with its essential role in spermatogenesis (Lardone et al., 2007). Similarly, ZFY, a gene involved in sex determination and differentiation, also displayed significant overexpression in male samples with a log fold change of 3.03, underscoring the robustness of these findings (Page et al., 1987).

Conversely, the gene XIST, a long non-coding RNA responsible for X-chromosome inactivation, is typically underexpressed in males who possess only one X chromosome and thereby do not undergo X-chromosome inactivation (Brown et al., 1991). This underexpression was quantified with a log fold change of -3.88, reinforcing the reliability of the observed expression patterns due to its biological functions.

Further analysis extended to other sex-specific genes located on the Y chromosome, such as RPS4Y1, USP9Y, UTY, TXLNG2P, and PRKY, all of which were overexpressed in males. This overexpression aligns with their chromosomal location and specific roles in male biological processes such as protein synthesis, spermatogenesis, and epigenetic regulation (Bellott et al., 2014).

Table 3.2: Differential expression of genes between male and female samples for the full dataset

Ensembl ID	Gene Name	Expr	logFC	Ave Expr	t	adj.P.Val (-log10)
ENSG00000129824	RPS4Y1	Over	6.70	4.83	69.17	infinite
ENSG00000067048	DDX3Y	Over	4.90	3.41	56.91	infinite
ENSG00000067646	ZFY	Over	3.03	2.29	46.63	311.08
ENSG00000229807	XIST	Under	-3.88	2.02	-43.01	277.37
ENSG00000114374	USP9Y	Over	3.04	2.27	39.26	242.25
ENSG00000183878	UTY	Over	3.52	2.48	39.03	240.13
ENSG00000131002	TXLNG2P	Over	3.16	2.38	37.63	227.13
ENSG00000099725	PRKY	Over	2.61	2.13	35.57	208.15

This comprehensive validation of differential expression across gender-specific genes confirms the robustness of our dataset and the statistical analyses employed. The high level of statistical significance associated with these findings supports their validity and confirms the anticipated

biological roles of these genes within the context of sex-specific genetic research. These results not only substantiate the integrity of our analytical procedures but also enhance the foundational knowledge necessary for further investigations into the genetic determinants of sex-based differences in biological traits and diseases.

3.3.9 Comprehensive Analysis of Gene Expression in Human Protein Atlas (HPA) Tissues using Full Dataset

The Human Protein Atlas (HPA) provides an invaluable resource for examining the distribution and expression levels of proteins across various human tissues. Understanding these variations is critical for uncovering the underlying biological mechanisms and their implications for health and disease. This section presents a comprehensive analysis of gene expression variations across different tissue types using HPA data. The primary aim is to explore how gene expression is influenced by age, sex-specific differences, and tumor presence, thereby identifying significant patterns and associations, as shown in Figure 3.10.

The analysis reveals several key findings. Comparisons between tumor and normal tissues provide critical insights into tumor biology. Overexpressed genes in tumor tissues are prominently observed in the esophagus, reflecting active roles in tumor development and progression. On the other hand, significant underexpression in tissues like the lung suggests a loss of function during tumorigenesis, contributing to the altered cellular environment in tumors. The two most significant tissue in the dot plot is esophagus and lung, which might be explained by a significant overlap in the gene expression profiles of developing esophageal and lung tissues (Morrisey & Rustgi, 2018).

Age-related gene expression changes are evident in various tissues. For example, the esophagus exhibits significant age-related overexpression with high odds ratios, suggesting enhanced metabolic activity or stress

responses in this tissue as age advances. In contrast, the lymphoid tissue and bone marrow display marked underexpression, indicative of possible degenerative changes or reduced cellular functions typical of aging tissues. Studies have shown a decline in the functionality and gene expression in bone marrow with age, linked with a reduction in hematopoietic activity and an increase in adiposity, reflecting a shift from a regenerative to a more degenerative state in the tissue (Liu et al., 2011). Similarly, the production of B cells in bone marrow is significantly decreased in aged organisms, attributed to changes in the microenvironment that unfavorably affect survival signals and cellular dynamics necessary for effective hematopoiesis (de Mol et al., 2021).

Sex-specific differences in gene expression are also observed. Overexpression is significant in sex-specific tissues such as the esophagus, urinary bladder and skin. Conversely, underexpression is only noted in the lung, highlighting differential regulatory mechanisms that may be at play between males and females.

This analysis illuminates the complex regulatory mechanisms underlying tissue-specific gene expression and highlights potential targets for therapeutic intervention, particularly in age-related diseases and cancer. Overexpressed genes in aging tissues may reflect compensatory mechanisms or increased demand for specific functions, while underexpressed genes could indicate declines in critical pathways or cellular functions. Understanding sex-specific differences in gene expression is crucial for developing gender-specific treatments and interventions.

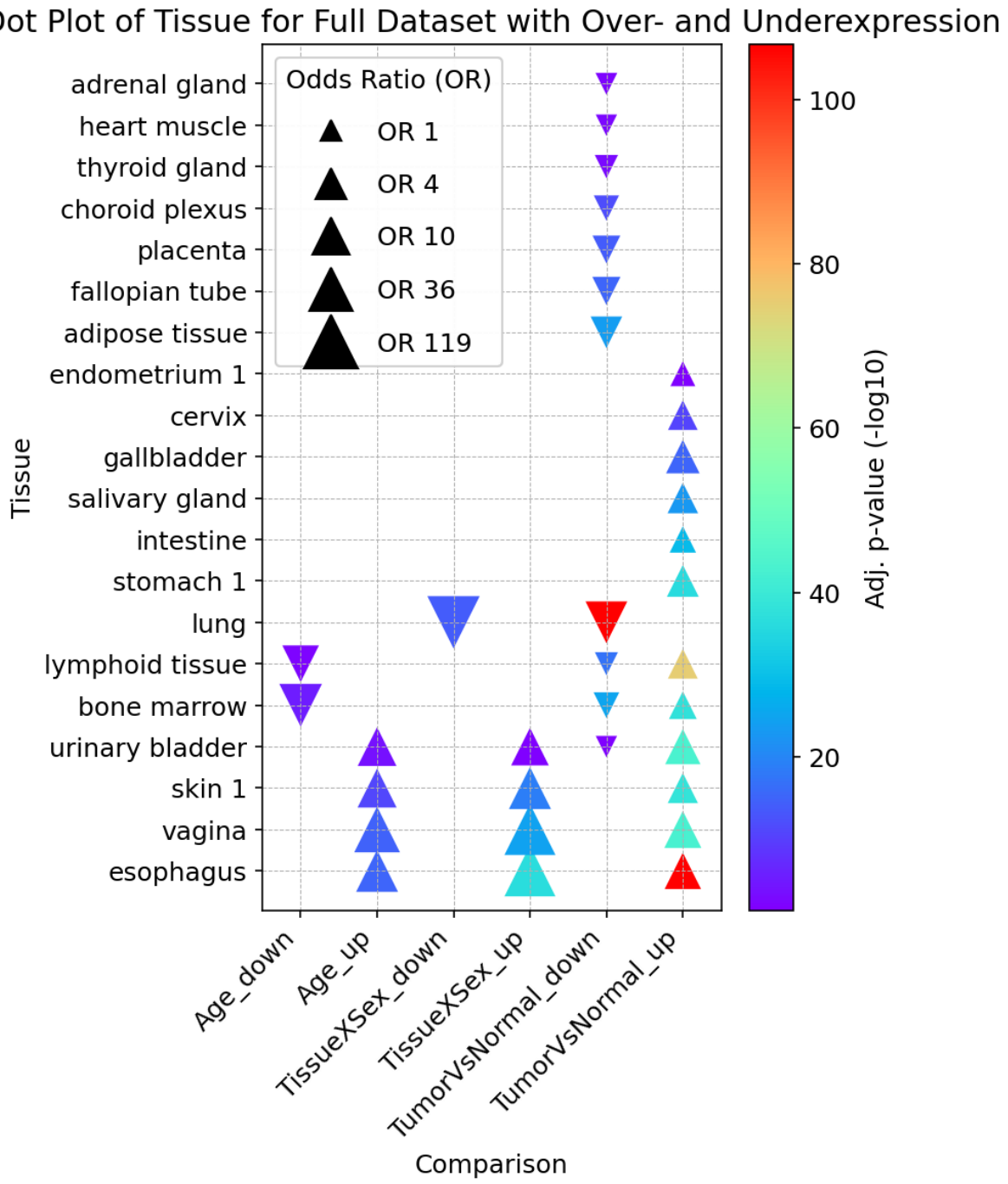


Figure 3.10: **Dot Plot of Differential Gene Expression Analysis in Tissues for Full Dataset.** This visualization illustrates the odds ratios (OR) for gene expression, where upward-pointing triangles indicate overexpression and downward-pointing triangles represent underexpression. The size of each symbol correlates with the odds ratio. The accompanying color gradient denotes the adjusted p-value (-log10), highlighting the statistical significance of each gene's differential expression across various tissues.

3.3.10 Comprehensive Analysis of Gene Expression in Numerical HPA Cell Types using Full Dataset

The Human Protein Atlas (HPA) is a crucial resource that allows for the examination of the distribution and expression levels of proteins across various human cell types. Understanding these variations is fundamental to uncovering the underlying biological mechanisms and their implications for health and disease. This section presents a detailed analysis of gene expression variations across different cell types using numerical HPA data, focusing on how gene expression is influenced by age, sex-specific differences, and tumor presence, as shown in Figure 3.11.

The analysis reveals several key findings across different conditions. Comparisons between tumor and normal cell types provide critical insights into tumor biology. Overexpressed genes in tumor cell types, such as extravillous trophoblasts, plasma cells, and suprabasal keratinocytes, reflect active roles in tumor development and progression. These genes may contribute to the uncontrolled proliferation, invasion, and metastasis characteristic of cancer cells. Conversely, underexpressed genes in adipocytes, endothelial cells, and monocytes suggest a loss of function during tumorigenesis, highlighting the complex interplay between oncogenic signals and the cellular environment.

Significant age-related overexpression is observed in various cell types, such as basal keratinocytes, suprabasal keratinocytes and basal squamous epithelial cells, with high odds ratios indicating enhanced metabolic activity or stress responses in these cells as age advances. This suggests that aging may lead to increased metabolic activity or stress responses in these cells, reflecting an attempt to counteract age-related declines in function or increased exposure to damaging agents over time. Conversely, age-related underexpression in cell types such as plasma cells and erythroid cells highlights potential degenerative changes or reduced cellular functions typical of aging tissues.

Sex-specific differences in gene expression are also observed. For example, the overexpressed genes in basal keratinocytes, suprabasal keratinocytes and basal squamous epithelial cells (the same cell types as for age-related overexpression).

Alveolar cells, type 1 and type 2, are crucial for lung function by facilitating gas exchange and producing surfactant. Analysis of numerical HPA gene expression data reveals significant underexpression of these cells in tissue and sex interactions and in tumor versus normal tissue comparisons. This underexpression suggests reduced functional capacity, likely due to degenerative changes or impaired repair mechanisms associated with sex differences and cancer. Studies on alveolar cell differentiation and function underscore the importance of these cells in maintaining lung integrity and their role in surfactant production, which is crucial for lung function and defense mechanisms. The observed underexpression in conditions like cancer significantly impacts these roles, leading to compromised lung function by disrupting gas exchange and surfactant production capabilities (Zhou et al., 2021; Zhang et al., 2022).

This analysis elucidates the complex regulatory mechanisms underlying cell type-specific gene expression and identifies potential targets for therapeutic intervention, particularly in age-related diseases and cancer. Overexpressed genes in aging cell types may reflect compensatory mechanisms or increased demand for specific functions, while underexpressed genes could indicate declines in critical pathways or cellular functions. Understanding sex-specific differences in gene expression is crucial for developing gender-specific treatments and interventions.

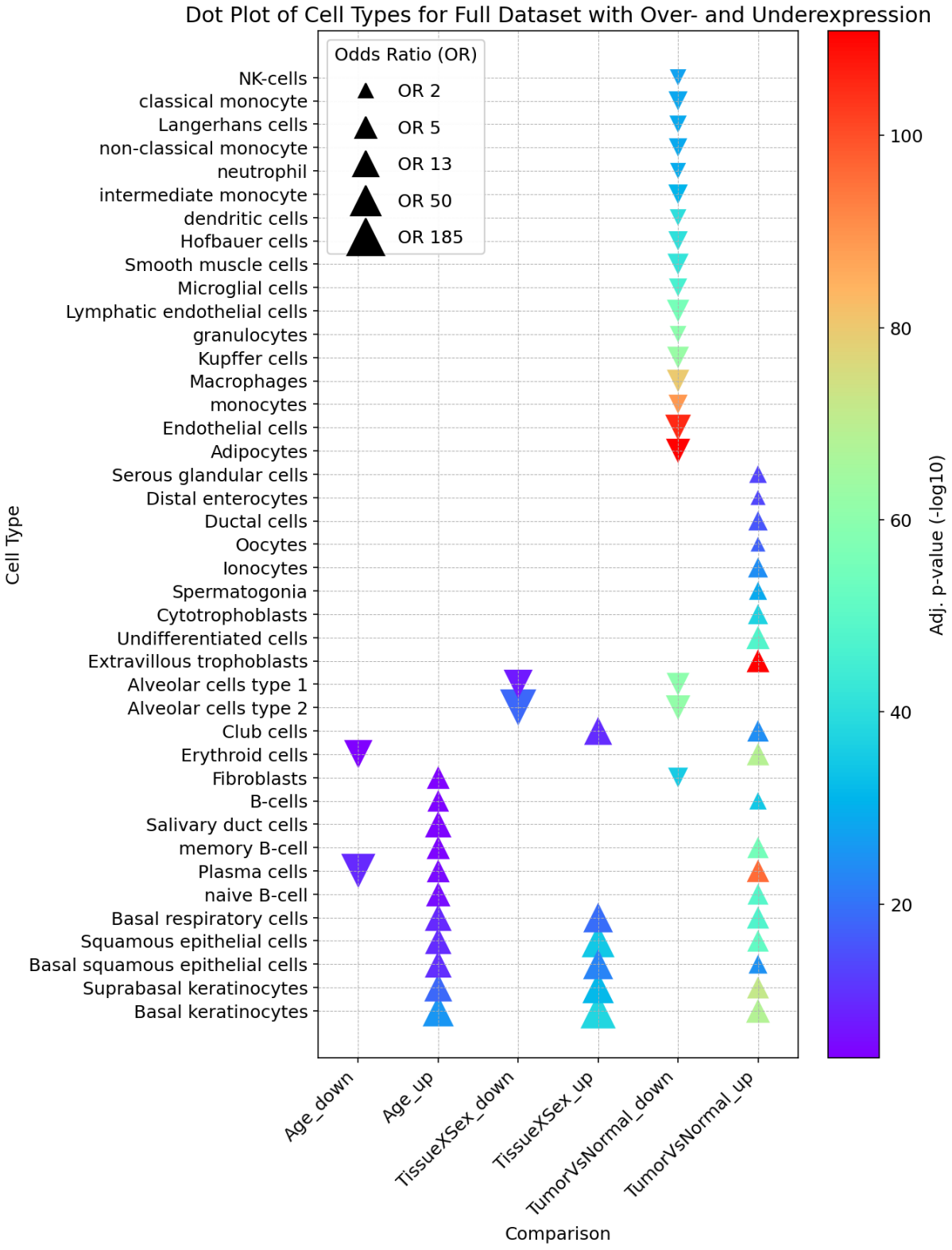


Figure 3.11: **Dot Plot of Top 20 differential gene expression analysis in cell types for the full dataset with over- and underexpression**. The triangles indicate over- or underexpression. Triangles pointing upward represent overexpression, while those pointing downward represent underexpression. Each triangle is color-coded according to a scale that represents -log10(p-value), indicating the statistical significance of the over- or underexpression.

3.4 Gene Expression Analysis Related to Smoking

In this section, we investigate the genetic influences of smoking on lung cancer using data from The Cancer Genome Atlas (TCGA). The primary focus is to analyze cell-type specific gene expressions within lung cancer tissues to elucidate the complex interactions between smoking-related genetic factors and the disease.

Mirroring the structure of chapter 3.3, we begin with a statistical overview of the dataset, employing an UpSet plot to depict the distribution and interconnectivity of significant gene clusters. This visualization highlights the differential gene expression patterns influenced by smoking, showcasing the balance between overexpressed and underexpressed genes.

Subsequently, we explore marker genes from the Human Protein Atlas (HPA) and Human Ensemble Cell Atlas (hECA). These marker genes are critical in identifying cell types that exhibit significant differential expression related to smoking. Analyzing these genes provides insights into the cellular composition of lung cancer tissues, enhancing our understanding of the molecular impacts of smoking.

Reactome pathway analysis is conducted to map the significant gene clusters to biological pathways, providing a deeper understanding of the functional implications of the observed gene expression changes due to smoking. This analysis is complemented by Gene Ontology (GO) classifications, which further categorize significant genes into biological processes, cellular components, and molecular functions.

A Directed Acyclic Graph (DAG) is utilized to visualize the hierarchical relationships and biological pathways, emphasizing the interconnectivity and shared functions relevant to smoking-related lung cancer. This graphical representation aids in identifying key pathways and their roles in disease progression.

To ensure the robustness of our findings, a sanity check is performed, validating the consistency and reliability of the differential gene expression results. This step is crucial for confirming the accuracy of our computational analyses.

Additionally, we generate dot plots of differential gene expression in both tissues and cell types. These plots display the odds ratios and adjusted p-values, highlighting the statistical significance of gene expression across various conditions influenced by smoking.

Finally, we look specifically at qualitative marker genes from hECA and HPA that are shared with numerical marker genes from HPA, highlighting marker genes that are found across multiple sources.

3.4.1 Lung Cancer Smoking Dataset Statistics

This section provides a comprehensive overview of the demographic and clinical characteristics of the lung cancer dataset obtained from The Cancer Genome Atlas (TCGA). This dataset includes detailed information on gender distribution, age statistics, and other relevant clinical variables, forming a robust foundation for subsequent genomic analyses.

The dataset comprises 442 lung cancer cases, with 290 males and 152 females. The mean age for males is 64.57 years (SD = 9.19), while for females it is 63.97 years (SD = 9.80). The age at diagnosis for the overall cohort shows a mean of approximately 64.37 years and a median of approximately 65.27 years. Additionally, the metadata includes comprehensive records for a total of 741 samples, providing a broad base for in-depth analysis of lung cancer. This extensive dataset supports a detailed exploration of the molecular mechanisms underlying the disease, considering both demographic and clinical variables.

The demographic statistics is further broken down for each smoking category in Table 3.3 showing the total number of cases, along with the

gender distribution and the mean age of males and females for each smoking category. It also includes the standard deviation (SD) of age for each group. Figure 3.12 show a bar chart of age distribution for each smoking category.

Table 3.3: This table presents the total number of cases, along with the gender distribution and the mean age of males and females for each smoking category. It also includes the standard deviation (SD) of age for each group.

Smoking Category	Total Cases	Males	Females	Mean Age Males (years)	SD Age Males	Mean Age Females (years)	SD Age Females
Never	126	64	62	61.97	10.64	61.31	10.31
Reformed	179	122	57	68.63	8.11	67.66	8.06
Current	137	104	33	62.23	7.84	62.43	9.66

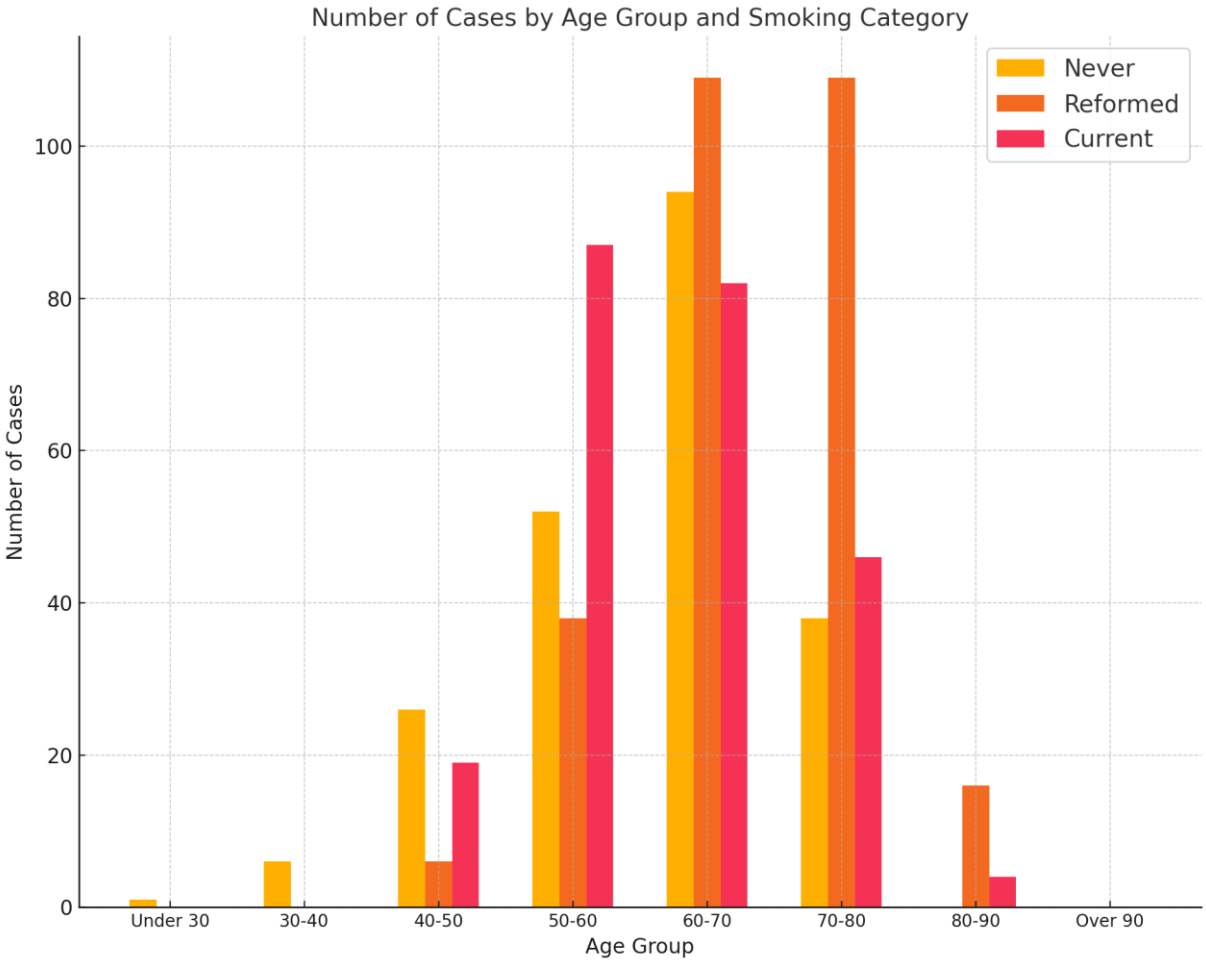


Figure 3.12: **The bar chart illustrates the number of cases categorized by age group and smoking status.** Age groups are represented on the x-axis, ranging from 'Under 30' to 'Over 90', while the number of cases is depicted on the y-axis. The chart differentiates between three smoking categories: Never (yellow), Reformed (orange), and Current (red).

3.4.2 Differentially expressed genes of the Smoking Dataset

The UpSet plot in Figure 3.13 provides a comprehensive visualization of the distribution and interconnectivity of significant gene clusters based on their differential expression in the lung cancer dataset. This plot is essential for elucidating the complex relationships among gene clusters, particularly concerning overexpression and underexpression across different conditions.

The UpSet plot employs color coding to distinguish between underexpressed and overexpressed genes: red indicates underexpressed genes, while green signifies overexpressed genes. The left histogram categorizes gene clusters by size, displaying the number of elements in each cluster. The accompanying bar chart at the top quantifies the elements per cluster, emphasizing the balance between overexpressed and underexpressed genes.

A detailed examination of the dataset reveals several key findings. In the Tumor vs. Normal expression patterns, there are 671 genes significantly overexpressed in tumor tissues compared to normal tissues (TumorVsNormal_up), contributing significantly to the dataset's interconnectivity. In contrast, TumorVsNormal_down includes 2161 genes significantly underexpressed in tumor tissues compared to normal tissues.

Regarding age-related expression patterns, the analysis identified 15 genes significantly overexpressed in relation to age (Age_up) and one gene significantly underexpressed (Age_down).

The FormerVsNever_Tumor_up condition has six significant genes and in former smokers compared to never smokers in tumor tissues (FormerVsNever_Tumor_down) there are 10 significantly underexpressed genes, suggesting potential long-term effects of smoking cessation. The CurrentVsNever_Tumor_down condition, with 19 genes significantly underexpressed, reflects the impact of smoking on gene expression in tumor tissues. This condition intersects with the TumorVsNormal_down

condition, indicating shared pathways of gene underexpression in these contexts.

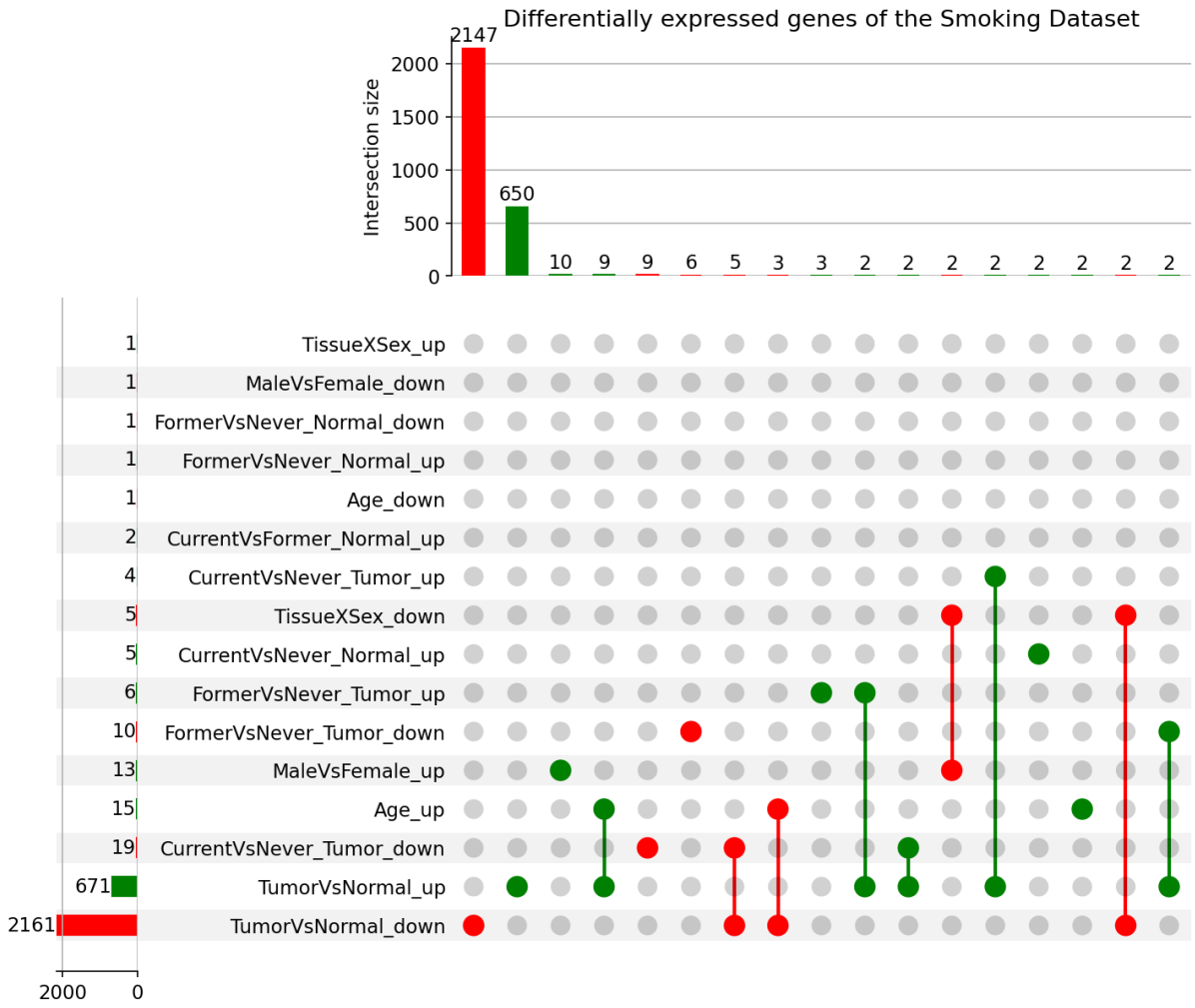


Figure 3.13: **UpSet Plot of Significant Gene Clusters for Smoking Dataset**. This plot illustrates the distribution and interconnectivity of significant gene clusters based on differential expression—red indicating underexpressed and green signifying overexpressed genes. The left histogram categorizes clusters by size, while the network diagram displays their relationships. The top bar chart quantifies the elements per cluster, highlighting the balance between overexpressed and underexpressed genes. The plot emphasizes distinct expression patterns in tumor versus normal tissues and age-related changes, showcasing intricate relationships among gene clusters.

3.4.3 Differential Expression Analysis of Marker Genes for Smoking Dataset

The differential expression analysis of qualitative marker genes from the Human Protein Atlas (HPA) and the Human Ensembl Cell Atlas (hECA) identifies significant variations in gene expressions, classified as either overexpressed or underexpressed. Figure 3.14 presents a comparative analysis of marker genes across various conditions.

The study covers age-related changes, tissue-specific expressions influenced by sex, and comparisons between tumor and normal tissues. Bronchial epithelium basal cells exhibit highly significant p-values in tumor versus normal tissue comparisons, highlighting their crucial role in tumor biology and potential as markers for cancer progression. These cells are overexpressed in TumorVsNormal conditions for HPA, underscoring their importance in lung cancer development and progression. Research has shown that the bronchial epithelium's response to factors like TGF-β1, which is involved in epithelial-mesenchymal transition, further emphasizes their dynamic role in cancerous transformations and their potential utility as therapeutic targets in oncology (Paw et al., 2021). Similarly, fibroblasts display significant p-values in TumorVsNormal_hECA_down and TumorVsNormal_hECA_up conditions, indicating their involvement in both tumor suppression and promotion. This underscores the importance of fibroblasts in the tumor microenvironment.

Alveolar cells type 2 show significant underexpression in the TumorVsNormal condition for hECA, underscoring their sensitivity to tumor-related changes and their potential impact on lung cancer progression. Among the top cell types by p-value, goblet cells demonstrate the highest p-value in the TumorVsNormal_hECA_up condition, suggesting substantial changes in gene expression related to tumor presence. Endothelial cells represent the top cell type by p-value for the TumorVsNormal_hECA_down contrast, as well as for the entire dot plot as a whole, indicating their significant downregulation in the smoking

dataset and highlighting their potential role in the differential expression patterns observed between tumor and normal samples. Studies have demonstrated that endothelial cells are crucial in the tumor microenvironment, influencing both tumor growth and the immune response. They regulate blood flow, control the permeability of blood vessels, and interact with immune cells, making them central to the dynamics of cancer progression and the response to therapy (Leone et al., 2024). The influence of smoking on endothelial cells, particularly in how it affects their function and viability, further highlights the complexities of their role in cancer. Smoking has been shown to induce apoptosis in pulmonary vascular endothelial cells, contributing to diseases such as chronic obstructive pulmonary disease (COPD), which shares some pathological features with lung cancer, indicating a broader impact of smoking on endothelial dysfunction (Song et al., 2021).

Vascular endothelial cells exhibit significant changes in the TumorVsNormal_hECA_down condition, indicating their involvement in the vascular alterations associated with tumors. Neutrophilic granulocytes are significant in the TumorVsNormal_hECA_down condition, pointing to their role in the immune cell response to tumors. Basal keratinocytes and bronchial epithelium basal cells are significant in the TumorVsNormal_HPA_up condition, suggesting their potential as markers for tumor progression. Pericytes, significant in the TumorVsNormal_hECA_down condition, highlight their role in the tumor microenvironment and vascular changes.

In terms of age-related changes, mast cells are significantly overexpressed in the Age_hECA_up condition, indicating their role in immune response modulation and their potential impact on aging processes. Research shows that mast cells, which are integral to immune and allergic responses, can influence age-related conditions like macular degeneration by affecting inflammatory and oxidative stress pathways (Malih et al., 2024). B cells, significant in the Age_hECA_up condition,

emphasize their involvement in the adaptive immune response and their potential impact on aging.

Alveolar cells type 2 consistently show underexpression in the TissueXSex condition for both hECA and HPA, highlighting their sensitivity to various biological influences, including sex-specific factors and tissue-specific changes. Fibroblasts display significant p-values in multiple conditions, underscoring their influence on cancer progression and their role in the tumor microenvironment.

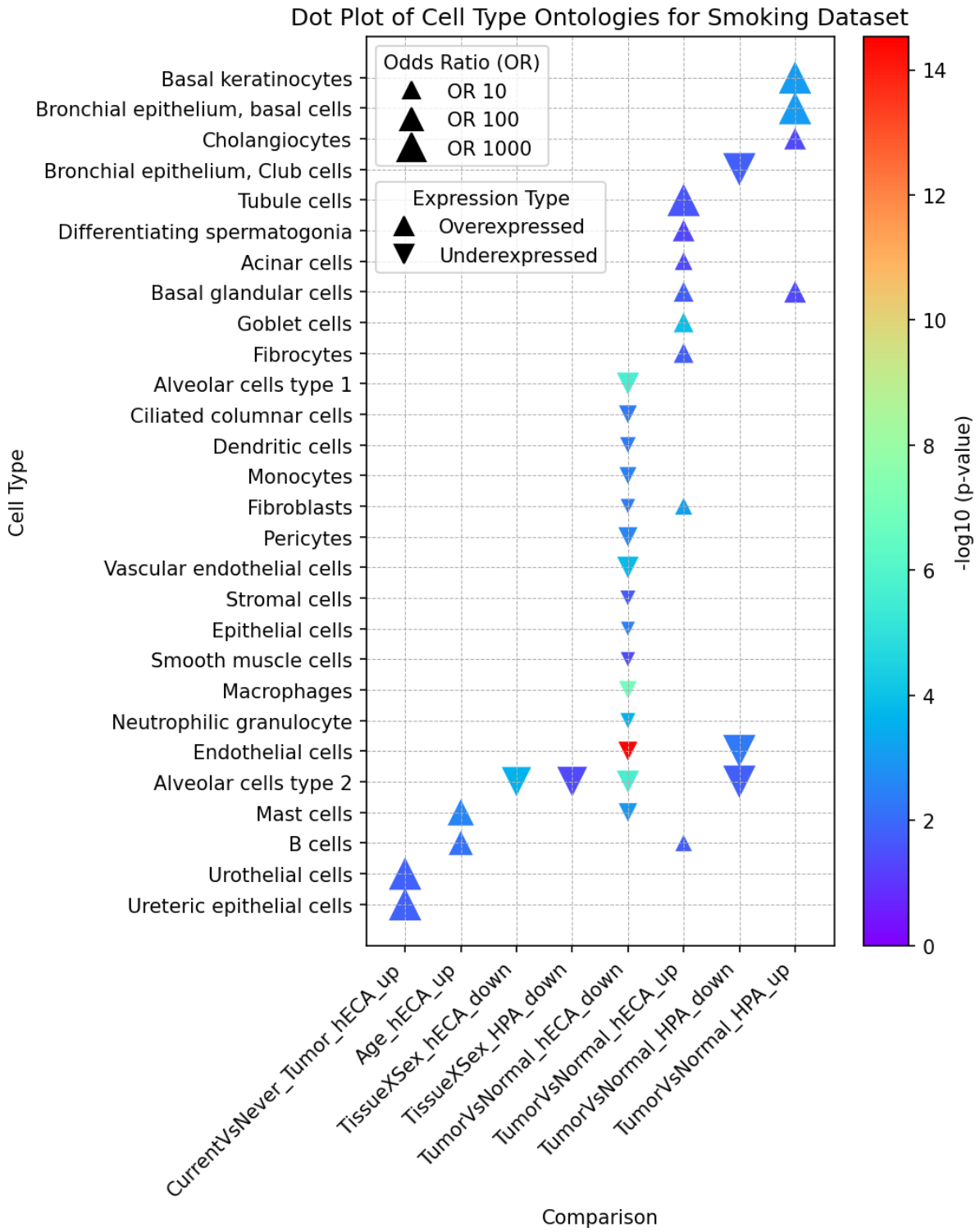


Figure 3.14: **Dot Plot for smoking dataset displaying differential expression of marker genes in cell types from the Human Protein Atlas (HPA) and Human Ensemble Cell Atlas (hECA).** Symbol sizes indicate odds ratios (ORs), with direction denoting overexpression (upward triangles) and underexpression (downward triangles). The color gradient bar shows the statistical significance (-log10 p-value).

3.4.4 Reactome Pathway Analysis using Smoking Dataset

A comprehensive analysis using the Reactome Pathway database was conducted to identify significant pathways associated with differentially expressed genes in lung cancer. This analysis focused particularly on variations resulting from tumor versus normal tissue comparisons and differences influenced by smoking status, age, and sex. Figure 3.15 presents a detailed comparative analysis, visually demonstrating how these conditions affect gene expression.

The analysis revealed several key findings. In tumor versus normal tissue comparisons, pathways associated with the cell cycle, including Cell Cycle and Cell Cycle, Mitotic, are highly significant. These pathways are predominantly overexpressed in tumor conditions, suggesting their crucial role in the progression of lung cancer. The genes involved in these pathways have potential as biomarkers for detecting and monitoring the disease. Pathways related to the immune system, such as Classical antibody-mediated complement activation and FCGR activation, also showed significant overexpression. This suggests an active role in tumor immunity and inflammation processes within the tumor microenvironment. The classical antibody-mediated complement activation pathway involves the activation of the complement system via antibodies, crucial for immune responses. Overactivation can lead to chronic inflammation, promoting a microenvironment conducive to tumor development and progression. Research underscores that the complement system, through components like C5a, can stimulate various cellular responses that enhance tumor progression. Activation of C5a, for instance, can lead to increased inflammation and promote tumor growth by affecting cellular functions such as migration and metastasis formation. Additionally, the dysregulation of this pathway is linked to adverse effects in the tumor microenvironment, potentially contributing to carcinogenesis by modifying cellular behavior and immune cell interactions (Netti et al., 2021; Zhang et al., 2019). FCGRs play a role in antibody-mediated immune responses,

influencing immune responses to cancer cells with potential roles in immune evasion and tumor progression. CD22 mediated BCR regulation, a pathway that regulates BCR signaling essential for B-cell function, is also noteworthy. Dysregulated BCR signaling can support abnormal B-cell activation and immune evasion mechanisms in lung cancer (Zhang et al., 2023).

In age-related comparisons, the top pathways identified as overexpressed are CD22 Mediated BCR Regulation, Regulation of Complement Cascade, Complement Cascade, Classical Antibody-Mediated Complement Activation, and FCGR Activation. The CD22 mediated BCR regulation pathway is crucial for B-cell function, and its dysregulation can lead to abnormal B-cell activation, contributing to cancer cell proliferation and immune evasion in lung cancer. The Regulation of Complement Cascade and Complement Cascade pathways play essential roles in immune response, promoting inflammation and cell lysis. Abnormal complement activity can lead to chronic inflammation, a condition linked to cancer development and progression, including in lung cancer. Classical antibody-mediated complement activation enhances immune defense mechanisms, and its overactivation can contribute to chronic inflammation and immune modulation, factors associated with lung cancer progression (Kharghan, 2017). FCGR activation modulates immune responses against tumors, and its abnormal activation can lead to immune evasion by cancer cells and support tumor growth.

For pathways that are significantly underexpressed in older age groups, the top pathways identified are Scavenging of Heme from Plasma and Binding and Uptake of Ligands by Scavenger Receptors. The Scavenging of Heme from Plasma pathway involves the clearance of free heme from the blood, preventing oxidative damage and maintaining iron homeostasis. Heme metabolism is linked to oxidative stress, which can promote cancer development (Chiang et al., 2021). Abnormal heme scavenging might contribute to the oxidative environment that supports

lung cancer progression. The Binding and Uptake of Ligands by Scavenger Receptors pathway involves scavenger receptors on immune cells binding and internalizing various ligands, including modified lipoproteins and apoptotic cells (Zani et al., 2015). Dysregulation in scavenger receptor pathways can affect inflammation and immune responses, potentially promoting a tumor-supportive microenvironment in the lungs.

There are notable overlaps between pathways identified in TumorVsNormal_up and Age_up conditions. Both conditions highlight the significance of pathways such as CD22 mediated BCR regulation, Regulation of Complement cascade, Complement cascade, Classical antibody-mediated complement activation, and FCGR activation. The CD22 mediated BCR regulation pathway is crucial for B-cell function, and its dysregulation can lead to abnormal B-cell activation, contributing to cancer cell proliferation and immune evasion in lung cancer. The Regulation of Complement cascade and Complement cascade pathways play essential roles in immune response, promoting inflammation and cell lysis. Abnormal complement activity can lead to chronic inflammation, a condition linked to cancer development and progression, including in lung cancer (Dominguez et al., 2021). Classical antibody-mediated complement activation enhances immune defense mechanisms, and its overactivation can contribute to chronic inflammation and immune modulation, factors associated with lung cancer progression. FCGR activation modulates immune responses against tumors, and its abnormal activation can lead to immune evasion by cancer cells and support tumor growth.

Sex-related comparisons revealed substantial differential expression in pathways such as Formation of the anterior neural plate and Formation of the posterior neural plate, indicating a sensitivity to biological variables like sex which could impact tissue architecture and influence the dynamics of the tumor microenvironment. The pathway HDMs demethylate histones was significantly overexpressed in male versus female comparisons,

suggesting sex-specific epigenetic modifications that might contribute to lung cancer susceptibility and progression.

In comparisons involving smoking status, particularly current versus former smokers, pathways like PPARA activates gene expression and Regulation of lipid metabolism by PPARalpha were significantly upregulated in current smokers. This points to ongoing inflammatory and metabolic dysregulation due to smoking, which could exacerbate lung cancer risk. In current versus never smokers, pathways such as Keratinization and Formation of the cornified envelope were prominent, suggesting changes in epithelial cell differentiation and barrier function that may facilitate cancer development. The formation of the cornified envelope involves the creation of a protective barrier in the outer layer of the skin and other tissues. Dysregulation in differentiation processes like cornification can indicate broader epithelial changes relevant to cancer biology, including lung cancer (Carregaro et al., 2013). Keratinization, the process by which keratin proteins form protective layers in epithelial cells, can be a marker of epithelial cell dysregulation, a characteristic of many carcinomas, including lung cancer (Heryanto & Imoto, 2023).

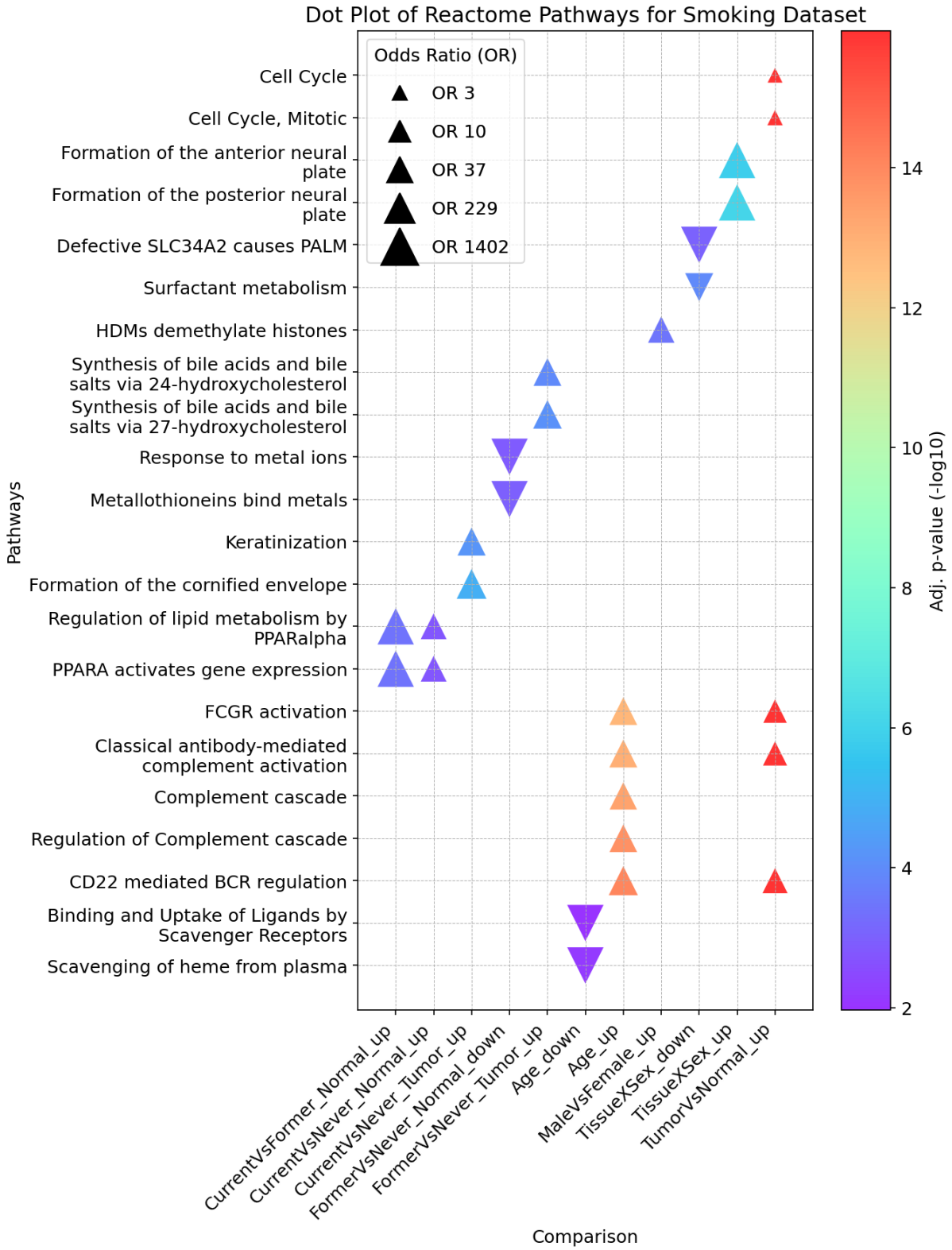


Figure 3.15: **Dot Plot of Reactome Pathways for Smoking Dataset**. Displays the top 5 for Age_up and TumorVsNormal_up conditions and top 2 significant pathways for the remaining conditions. Symbol sizes indicate odds ratios (ORs), with upward triangles for overexpression and downward triangles for underexpression. The color gradient bar represents the statistical significance (-log10 p-value).

3.4.5 Biological Processes (BP) from Gene Ontology (GO)

An enrichment analysis of Gene Ontology (GO) Biological Processes (BP) was undertaken to investigate the broader biological implications of differentially expressed genes identified in the study. Figure 3.16 visually maps these associations across various conditions, including comparisons between current and former smokers (Normal tissue), current and never smokers (Tumor tissue), former and never smokers (Tumor tissue), age-related differences, and tumor versus normal tissue comparisons.

In the comparison between current and former smokers (Normal) where genes are overexpressed, the metabolism of insecticides, often mediated by cytochrome P450 enzymes, was significant. These pathways can influence lung cancer risk by affecting the body's ability to process carcinogens found in tobacco smoke and environmental pollutants. Research detailed in Bernauer et al. (2006) emphasizes the diversity and functionality of cytochrome P450 enzymes in human lung tissue, which play a crucial role in the metabolic activation of chemicals inhaled via tobacco smoke, potentially contributing to carcinogenic effects. Additionally, Stipp and Acco (2021) review how these enzymes, through their interaction with proinflammatory cytokines in the tumor microenvironment, can influence carcinogenesis and modify the efficacy and toxicity of chemotherapy in lung cancer. Furthermore, the dibenzo-p-dioxin catabolic process was identified. Dioxins are environmental pollutants known to be carcinogenic, and the body's ability to break them down can impact lung cancer risk, as these compounds can cause DNA damage and promote carcinogenesis (Valavanidis et al., 2013).

In the comparison between current and never smokers (Tumor), where genes are underexpressed, several processes related to sensory perception were identified. Taste receptors, including those for bitter taste, are expressed in the respiratory system and play roles in detecting harmful substances and triggering protective responses. Underexpression of these genes in tumors might reflect alterations in cell signaling

pathways critical for recognizing and responding to carcinogenic stimuli. This includes the detection of chemical stimuli involved in the sensory perception of bitter taste. Studies like those by Risso et al. (2016) have shown that variations in the TAS2R38 bitter taste receptor influence smoking behavior, potentially due to differential sensitivity to bitter compounds in tobacco smoke. For the comparison between former and never smokers (Tumor) where genes are overexpressed, the cellular response to jasmonic acid stimulus and the response to jasmonic acid were highlighted. Although jasmonic acid is a plant hormone, the cellular response mechanisms it triggers, such as stress and defense responses (Rehman et al., 2023), have parallels in human biology. In the context of lung cancer, overexpression of genes related to these pathways may reflect an increased cellular effort to counteract the stress and damage caused by tumor growth and therapeutic interventions. The metabolism of daunorubicin and doxorubicin, both chemotherapy drugs used to treat various cancers including lung cancer, was also significant. The metabolic processes involved in handling these drugs are crucial for understanding their therapeutic effects and side effects in lung cancer treatment.

For the comparison between tumor and normal tissues where genes are overexpressed, the adaptive immune response was again highlighted, underscoring its importance in targeting and destroying cancer cells. The immunoglobulin-mediated immune response and B cell-mediated immunity were also significant. These responses are part of the body's defense mechanism against cancer, and therapies that utilize or enhance immunoglobulins are being developed for cancer treatment. Additionally, the mitotic cell cycle process, crucial for cell division, was highlighted. In cancer, dysregulation of this process leads to uncontrolled cell proliferation, and targeting the cell cycle is a common strategy in cancer therapy.

In the comparison between tumor and normal tissues where genes are underexpressed, several processes crucial for maintaining normal tissue architecture and function were identified. This includes anatomical structure development, regulation of multicellular organismal processes, cell motility, and multicellular organismal processes. Underexpression of genes involved in these processes in tumors suggests a disruption in normal tissue organization, which is a hallmark of cancer, facilitating tumor invasion and metastasis. Proper segregation of chromosomes during cell division, critical for genomic stability, was also significant. Errors in this process can lead to mutations and cancer progression. Understanding these mechanisms can help in developing treatments that target cancer cell division.

In the age-related comparison where genes are overexpressed, the adaptive immune response is critical in targeting and destroying cancer cells. Age-related changes in the immune system can impact the effectiveness of this response against lung cancer. The immune response in general, including the generation of diverse immune receptors capable of recognizing a wide range of antigens, is essential for effective immune surveillance. This includes processes such as the immunoglobulin-mediated immune response and B cell-mediated immunity. Antibodies can recognize and bind to tumor antigens, marking cancer cells for destruction by the immune system.

The analysis also identified several biological processes that are shared across different comparisons, underscoring common pathways potentially involved in lung cancer pathogenesis. Specifically, adaptive immune response, immunoglobulin-mediated immune response, and B cell-mediated immunity were found to be significantly enriched in both the age-related and tumor versus normal comparisons. These shared processes highlight the pivotal role of the immune system in recognizing and responding to tumor cells. The adaptive immune response involves T cells that can target and destroy cancer cells, while immunoglobulin-

mediated responses involve antibodies that mark cancer cells for destruction. B cells, which produce these antibodies, are critical for mounting an effective immune response against tumors. The presence of these shared processes across different comparisons suggests that enhancing these immune pathways could be a key strategy in developing effective treatments for lung cancer. Understanding the commonalities in these biological processes provides valuable insights into potential therapeutic targets and biomarkers for lung cancer.

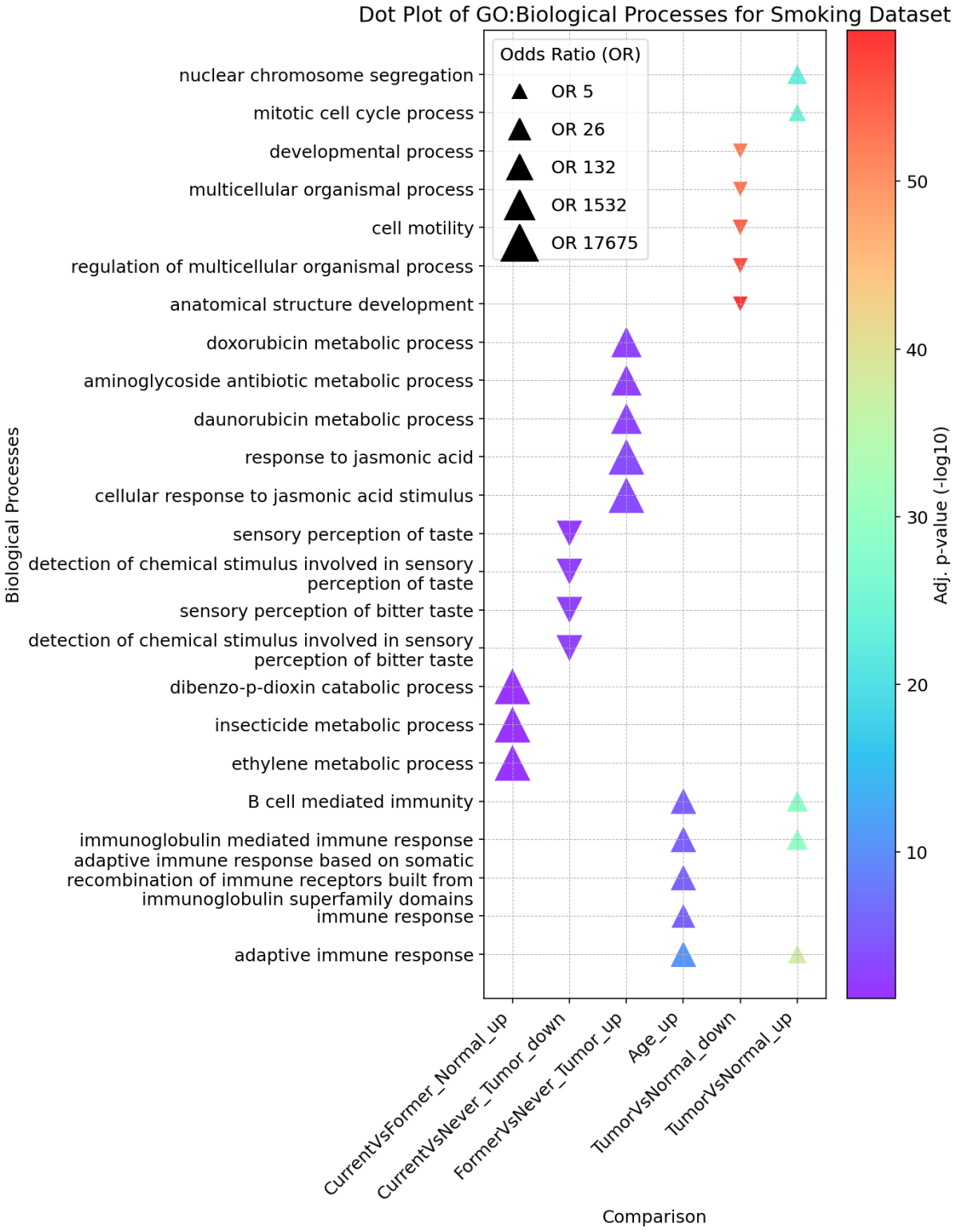


Figure 3.16: **Dot Plot of GO Biological Processes for Smoking Dataset**. This plot organizes Gene Ontology (GO) Biological Processes along the y-axis, each linked to specific biological conditions such as smoking category, age, sex, and tumor presence. Vertical stacks of symbols illustrate the involvement within each condition, with the size of each symbol indicating the odds ratio, reflecting the strength of association. The color gradient from purple to red represents the adjusted p-values (-log10), highlighting the statistical significance of gene involvement in each condition.

As in chapter 3.3.5, a Directed Acyclic Graph (DAG) was generated to explore the relationships between the overexpressed biological processes for the TumorVsNormal contrast. This is shown in Figure 3.17. This DAG shows several similarities with the DAG constructed from the full dataset (without specific smoking categories), emphasizing core biological pathways involved in lung cancer. Furthermore, letters such as A, B, and C in front of the node names are aliases for depth-01 GO terms, providing a general location within the DAG. For example, "C" is the alias for metabolic process, so terms descended from metabolic processes will have a "C" associated with them, such as immune system response (Klopfenstein et al., 2018).

The DAG reveals that many significant GO terms are related, forming a network of interconnected processes. It consists of three branches, where the left branch is related to the immune system process, the middle branch is related to immune response and the right branch is related to the cell cycle process. There are some differences between the DAGs in Figure 3.7 and Figure 3.17. The two branches related to the immune system in Figure 3.17 are unique for this DAG. In Figure 3.7, the left branch related to organization within the cell is unique for that DAG. There are a few similarities between the DAGs. Particularly, the node Cell Cycle Process (GO:0022402) connects to the more specific node Mitotic Cell Cycle Process (GO:1903047).

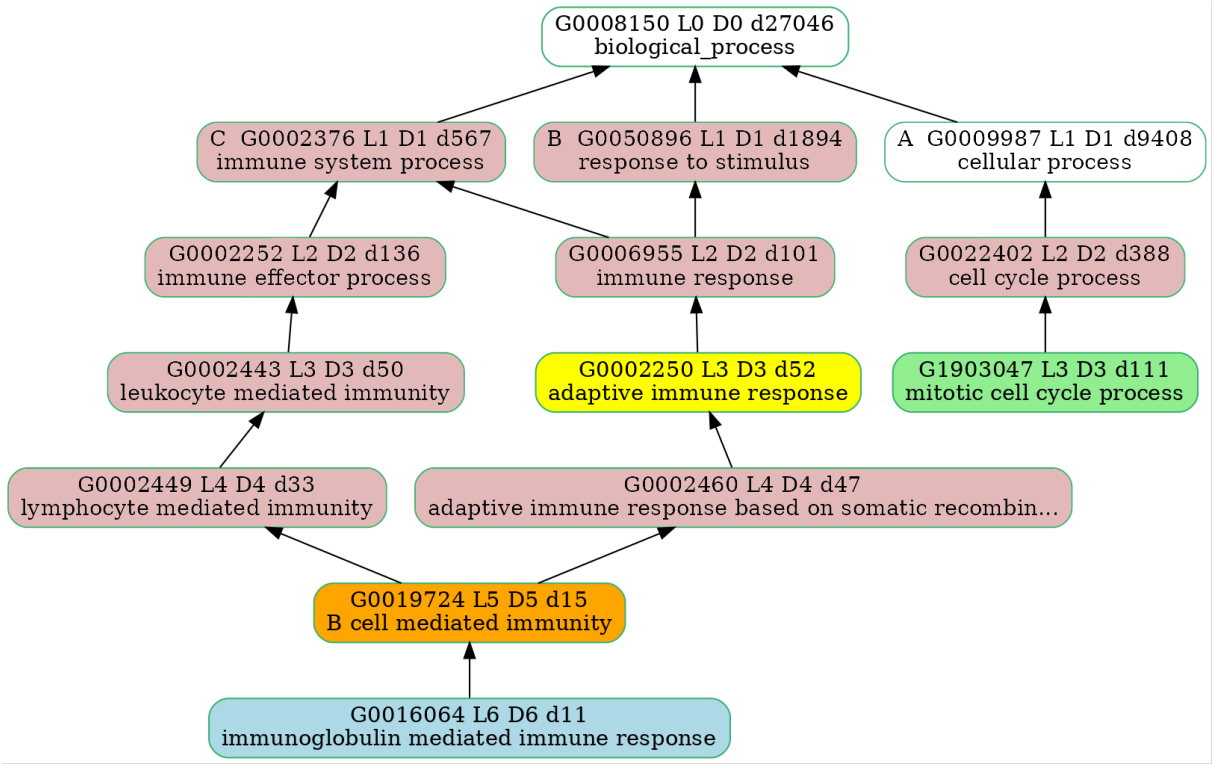


Figure 3.17: **Directed Acyclic Graph (DAG) illustrating the top four most significant Biological Process Gene Ontology (GO:Biological Process) results for the overexpressed TumorVsNormal contrast.** The arrows in the DAG point from child to parent, denoting a progression from more specific to more general terms. This visualization highlights the hierarchical relationships and biological pathways involved, emphasizing the interconnectivity and shared biological functions relevant to the overexpressed TumorVsNormal contrast. The yellow node is the most significant result, the light blue second most significant, the orange third most significant and the light green fourth most significant. The light red nodes represent remaining nodes in the DAG that are found to be significant GO:Biological Process results for the overexpressed TumorVsNormal contrast, but are not among the top four most significant. The white nodes are GO:BP results that are not found to be significant for the overexpressed TumorVsNormal contrast, but are part of the hierarchical structure of the DAG. Each of the nodes contain a unique GO ID, level (*L*) indicating the minimum path from the top root, depth (*D*) indicating the maximum path from the top root term and descendant count (*d*) indicating the total number of GO terms below the given node from the GO hierarchy structure (not shown in this DAG, but a part of the underlying Open Biological and Biomedical Ontologies file) (Klopfenstein et al., 2018). The letters A, B and C at the second most top nodes represent aliases for depth-01 GO terms, used to provide the general location in the GO DAG of any GO term. They stand for cellular process, biological regulation and metabolic process, respectively (Klopfenstein et al., 2018). The full name of the node with GO ID GO002460 is “adaptive immune response based on somatic recombination of immune receptors built from immunoglobulin superfamily domains”.

3.4.6 Cellular Components (CC) from Gene Ontology (GO)

An enrichment analysis of Gene Ontology (GO) Cellular Components (CC) was conducted to explore the broader cellular implications of differentially expressed genes identified in the study. Figure 3.18 visually maps these associations across various conditions, including comparisons between current and former smokers (Normal), current and never smokers (Tumor), age-related differences, and tumor versus normal tissue comparisons.

In the comparison between current and former smokers (Normal) where genes are overexpressed, the plasma membrane proton-transporting V-type ATPase complex was highlighted. This complex plays a crucial role in regulating the acidification of intracellular compartments, which is essential for various cellular processes including protein degradation and receptor-mediated endocytosis (Pamarthy et al., 2018). The overexpression of this complex may indicate enhanced cellular activity and metabolic processes associated with cancer progression.

In the comparison between current and never smokers (Tumor) where genes are underexpressed, several cellular components related to vesicle formation and membrane structures were identified. These include vesicles, the side of the membrane, the external side of the plasma membrane, extracellular exosomes, and extracellular vesicles. These components are integral to processes such as intracellular transport, cell communication, and the immune response. Underexpression in tumors may suggest a reduction in these critical cellular functions, potentially contributing to tumor development and immune evasion.

In the comparison between tumor and normal tissues where genes are underexpressed, several components critical for maintaining cellular structure and function were identified. These include the cell periphery, plasma membrane, extracellular region, cell surface, and the external encapsulating structure. The cell periphery and plasma membrane are vital for cell integrity and communication. The extracellular region and cell

surface are involved in interactions with the extracellular matrix and other cells, which are crucial for tissue organization and function. The external encapsulating structure provides structural support and protection to cells. Underexpression of these components in tumors suggests a breakdown in these critical functions, facilitating tumor invasion and metastasis.

For the comparison between tumor and normal tissues where genes are overexpressed, components such as the immunoglobulin complex, extracellular region, extracellular space, nucleosome, and cell periphery were highlighted. The nucleosome is involved in the organization and regulation of DNA, playing a critical role in gene expression and cellular function. Overexpression of these components in tumors indicates an increase in cellular activities related to immune response and gene regulation, which could be a response to the presence of cancer cells.

In the age-related comparison where genes are overexpressed, components such as the immunoglobulin complex, extracellular region, extracellular space, cell periphery, and blood microparticles were highlighted. The immunoglobulin complex is vital for the immune response, with antibodies playing a key role in identifying and neutralizing pathogens and cancer cells. The extracellular region and space, along with blood microparticles, are crucial for intercellular communication and the immune response. The cell periphery, including structures involved in cell signaling and interaction with the extracellular environment, is essential for maintaining cellular integrity and function. Overexpression of these components in older individuals may reflect an enhanced immune surveillance mechanism, potentially impacting the body's ability to recognize and respond to cancer cells.

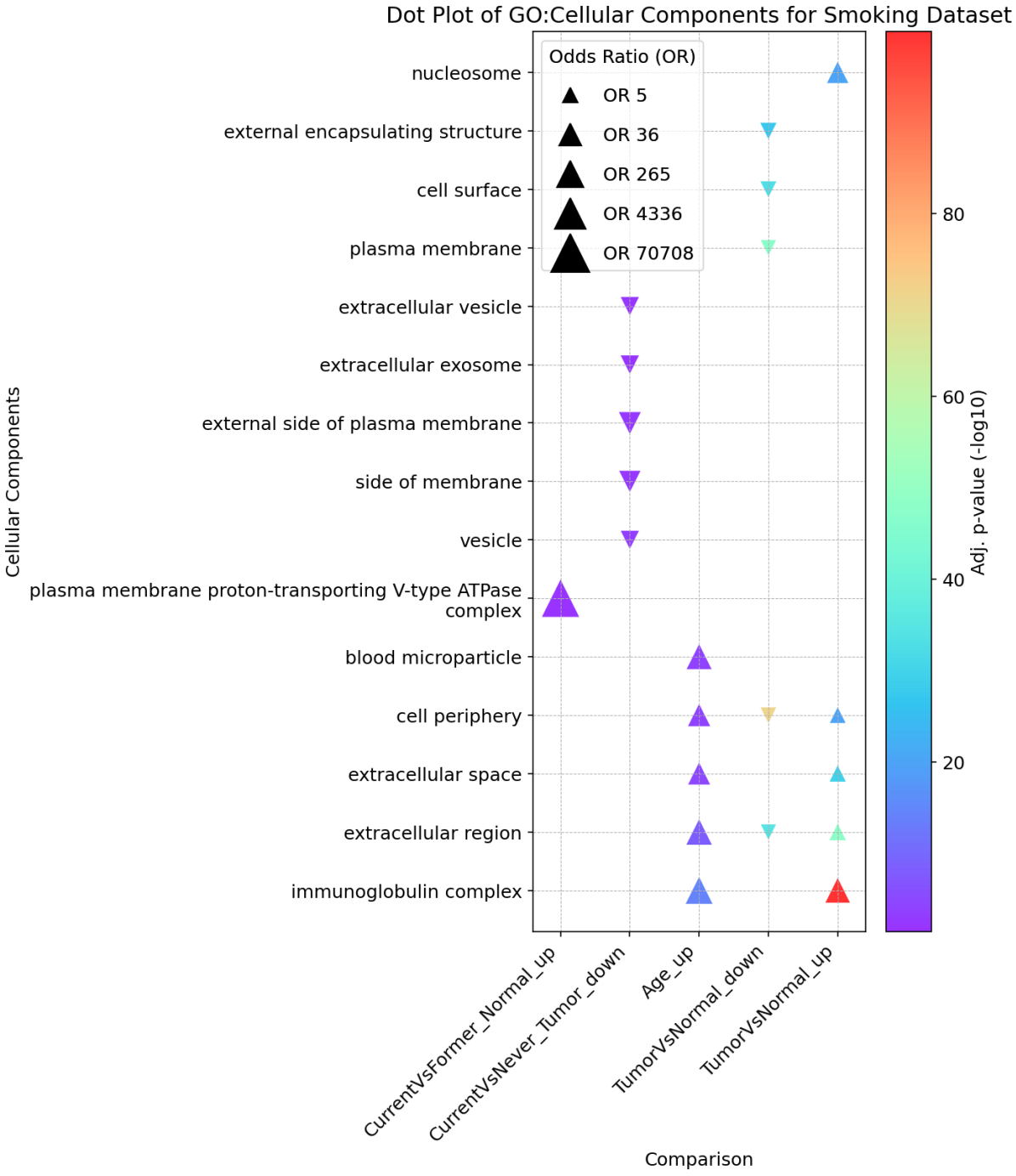


Figure 3.18: **Dot Plot of GO Cellular Components for Smoking Dataset**. This plot show Gene Ontology (GO) Cellular Components associated with smoking category, age, sex-specific tissues, and tumor presence. Symbols represent odds ratios (OR), with size indicating the magnitude of the OR. Upward-pointing triangles denote overexpressed genes, while downward-pointing triangles indicate underexpressed genes. The color gradient from purple to red represents adjusted p-values (-log10), with red marking the most statistically significant findings.

3.4.7 Molecular Functions (MF) from Gene Ontology (GO)

An enrichment analysis of Gene Ontology (GO) Molecular Functions (MF) was conducted to explore the broader molecular implications of differentially expressed genes identified in the study. Figure 3.19 visually maps these associations across various conditions, including comparisons between current and former smokers (Normal), current and never smokers (Normal), former and never smokers (Tumor), age-related differences, comparisons between males and females, and tumor versus normal tissue comparisons.

In the comparison between current and former smokers (Normal), where genes are overexpressed, flavonoid 3'-monooxygenase activity was highlighted. However, it is important to note that flavonoid 3'-monooxygenase does not exist in humans. This enzyme is specific to plants, where it plays a role in the metabolism of flavonoids—compounds known for their antioxidant and potential anti-cancer properties. The identification of this activity in human gene expression data presents an interpretive challenge and suggests possible issues with annotation or cross-species comparisons.

Similarly, in the comparison between current and never smokers (Normal), flavonoid 3'-monooxygenase activity was again highlighted. The repeated identification of this plant-specific enzyme in human data underscores the complexity of interpreting such results and raises questions about the accuracy of the annotations used.

In the comparison between former and never smokers (Tumor) where genes are overexpressed, several dehydrogenase activities were identified, including alcohol dehydrogenase (NADP+) activity, aldo-keto reductase (NADP) activity, phenanthrene 9,10-monooxygenase activity, indanol dehydrogenase activity, and trans-1,2-dihydrobenzene-1,2-diol dehydrogenase activity. These enzymes are involved in the oxidation-reduction processes essential for detoxifying carcinogens and other harmful substances (Orywal et al., 2020). Overexpression of these

activities may indicate enhanced detoxification capacity in response to the carcinogenic environment associated with smoking.

In the comparison between tumor and normal tissues where genes are underexpressed, several molecular functions crucial for cellular communication and signaling were highlighted. These include protein binding, signaling receptor binding, integrin binding, molecular function regulator activity, and calcium ion binding. Underexpression of these functions in tumors suggests a disruption in normal signaling pathways, which can contribute to uncontrolled cell growth and metastasis.

For the comparison between tumor and normal tissues where genes are overexpressed, antigen binding was again significant, along with structural constituent of chromatin, structural molecule activity, protein heterodimerization activity, and cell adhesion molecule binding. Overexpression of these functions indicates enhanced cellular activities related to immune response, structural integrity, and cell-cell interactions, potentially reflecting the body's attempt to counteract tumor growth and spread.

In the comparison between males and females where genes are overexpressed, various demethylase activities were highlighted, including histone H3 demethylase activity, histone demethylase activity, protein demethylase activity, demethylase activity, and 2-oxoglutarate-dependent dioxygenase activity. These enzymes play crucial roles in the regulation of gene expression through epigenetic modifications, which can impact cancer development and progression.

In the age-related comparison where genes are overexpressed, antigen binding was prominently identified. This function is critical for immune surveillance and the identification of pathogens and cancer cells. The overexpression of antigen binding in older individuals may suggest an increased reliance on immune mechanisms to combat age-related changes and the emergence of cancer cells.

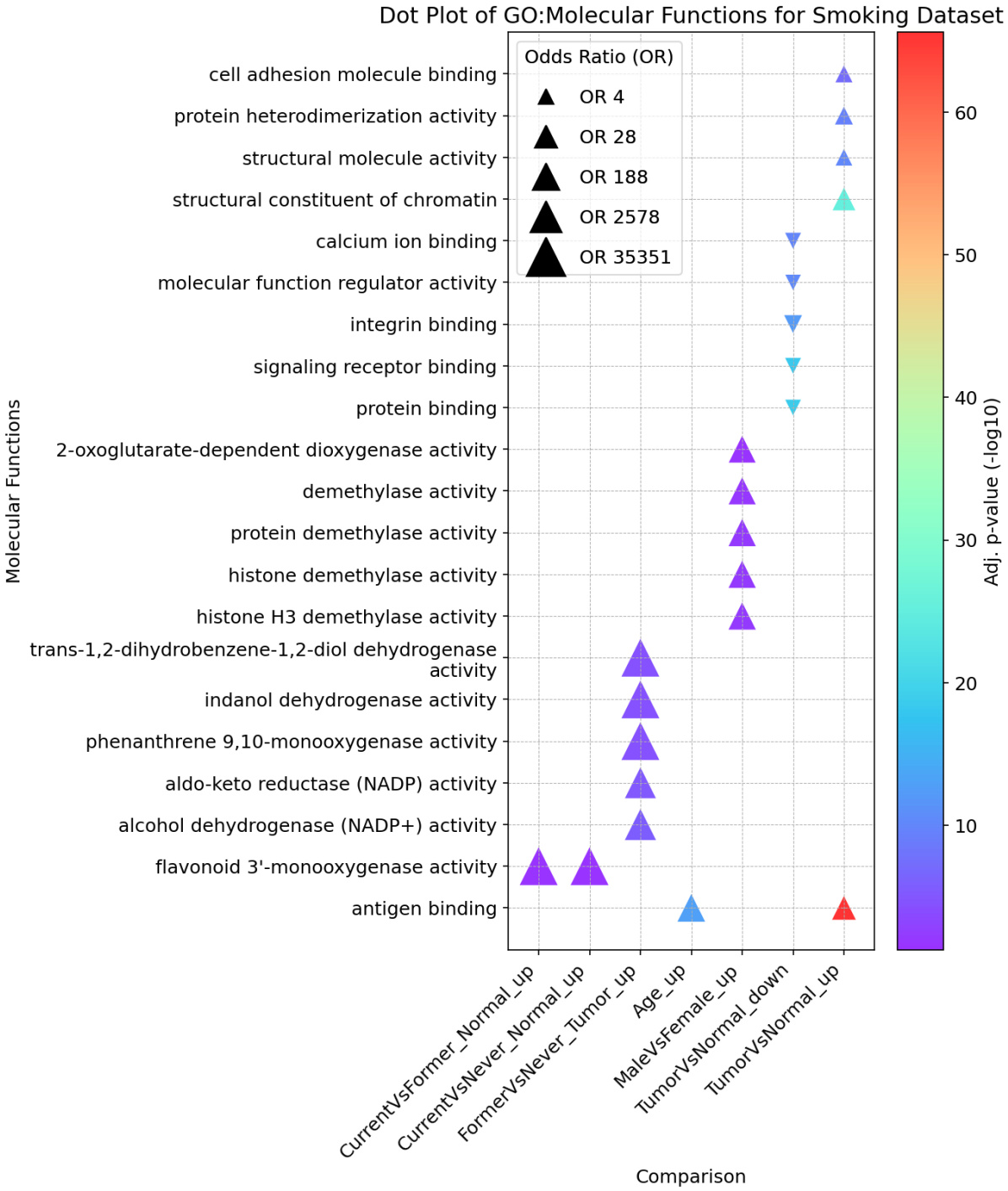


Figure 3.19: **Dot Plot of GO Molecular Functions for Smoking Dataset**. The dot plot shows Gene Ontology (GO) Molecular Functions associated with smoking category, age, gender, sex-specific tissues, and tumor presence. Upward-pointing triangles denote overexpressed genes, while downward-pointing triangles indicate underexpressed genes. Larger symbols indicate higher odds ratios (OR), while a color gradient from purple to red represents the significance of p-values (-log10 scale).

3.4.8 Validation of Differential Gene Expression Between Male and Female Samples in Smoking Dataset

To ensure the robustness of the regression results, a validation check was conducted to confirm the differential expression of all genes between male and female samples. This followed the same structure as in Chapter 3.3.8. The dataset underwent a filtration process to include only those genes previously analyzed in the MaleVsFemale regression study. A comprehensive search was performed to match these genes against established lists of identifiers known to distinguish male from female gene expressions. This step was crucial in verifying that the identified genes were accurately reflecting sex-specific differences in expression, thereby bolstering the credibility of the regression findings.

The selection of genes was based on their established roles in differentiating male and female gene expressions, specifically focusing on genes located on sex chromosomes (X and Y) and those known to be influenced by sex-specific factors. The dataset was filtered to include only those genes that were part of the MaleVsFemale regression analysis, ensuring consistency and relevance in the validation process. Each gene was cross-referenced with established databases and literature to confirm its differential expression between male and female samples. Key sources included scientific articles, genetic databases, and specialized studies on sex-linked gene expression.

The verification process confirmed that certain genes located on the Y chromosome, such as RPS4Y1, DDX3Y, ZFY, USP9Y, UTY, PSMA6P1, LINC00278, TXLNG2P, PRKY, KDM5D, EIF1AY, and CD24P4, were overexpressed in male samples as shown in Table 3.4. These genes are known for their roles in male-specific functions and are typically not present or not expressed in female samples due to the absence of the Y chromosome.

Similarly like in Chapter 3.3.8, the gene XIST, located on the X chromosome, was verified to be underexpressed in male samples. XIST is involved in X-chromosome inactivation, a process essential in female samples to balance the dosage of X-linked genes (Brown et al., 1991).

An interesting case was the IGHG4 gene, located on chromosome 14, which is not directly linked to sex chromosomes. However, its expression levels can be influenced by sex-specific factors such as hormonal differences and immune system variations. IgG4-related disease (IgG4-RD), associated with the overexpression of IGHG4, shows a higher prevalence in middle-aged and elderly males. This disease involves fibroinflammatory infiltration of various organs and highlights how immune responses can differ between sexes (Guinee, 2018).

Table 3.4: Differential expression of genes between male and female samples in smoking dataset

Ensembl ID	Gene Name	Expr	logFC	AveExpr	t	adj.P.Val (-log10)
ENSG00000229807	XIST	Under	-5.38	2.51	-73.9	infinite
ENSG00000129824	RPS4Y1	Over	6.54	5.03	65.33	299.9
ENSG00000067048	DDX3Y	Over	5.33	4.15	62.53	288.43
ENSG00000012817	KDM5D	Over	4.49	3.58	56.21	260.91
ENSG00000067646	ZFY	Over	3.51	2.92	52.6	244.34
ENSG00000183878	UTY	Over	4.46	3.56	52.38	243.40
ENSG00000114374	USP9Y	Over	3.77	3.1	47.15	218.04
ENSG00000215414	PSMA6P1	Over	3.05	3.12	43.51	199.52
ENSG00000231535	LINC00278	Over	2.79	2.45	42.34	193.41
ENSG00000131002	TXLNG2P	Over	3.89	3.2	42.01	191.73
ENSG00000198692	EIF1AY	Over	3.03	2.61	41.38	188.44
ENSG00000099725	PRKY	Over	3.02	2.64	39.93	180.68
ENSG00000185275	CD24P4	Over	1.84	2.36	23.92	89.59
ENSG00000211892	IGHG4	Over	1.06	6.2	5.74	5.46

The validation process confirmed that the genes exhibit differential expression between male and female samples consistent with their known biological roles and chromosomal locations. Genes on the Y chromosome were consistently overexpressed in males, while XIST was underexpressed in males, aligning with its function in X-chromosome inactivation in females. The expression of IGHG4, although not sex-linked, was noted to vary due to immune response differences and its association with IgG4-RD, more common in males. These findings support the reliability of the regression results and underscore the importance of considering both genetic and epigenetic factors in sex-specific gene expression studies.

3.4.9 Comprehensive Analysis of Gene Expression in Human Protein Atlas (HPA) Tissues using Smoking Dataset

The Human Protein Atlas (HPA) provides an invaluable resource for examining the distribution and expression levels of proteins across various human tissues. Understanding these variations is critical for uncovering the underlying biological mechanisms and their implications for health and disease. This section presents a comprehensive analysis of gene expression variations across different tissue types using HPA data. The primary aim is to explore how gene expression is influenced by smoking status, age, sex-specific differences, and tumor presence, thereby identifying significant patterns and associations, as shown in Figure 3.20.

In the comparison of current smokers versus never smokers in tumor tissues, salivary gland tissue exhibited significant underexpression, indicating possible direct metastatic involvement or systemic effects on gland function. Intestinal tissues also showed underexpression, suggesting metabolic or inflammatory responses linked to lung cancer progression. Similarly, comparisons between former smokers and never smokers in tumor tissues revealed significant underexpression in the salivary gland, reinforcing the possibility of persistent systemic effects or direct

metastatic involvement. Salivary gland-type tumors of the lung are rare and typically originate from the submucosal exocrine glands of the large airways (Horio et al., 2024). These tumors are often misdiagnosed due to their rarity and the need for differential diagnosis to distinguish between primary and metastatic diseases. The management of these tumors requires comprehensive knowledge of diagnostics, including molecular characteristics, and treatment modalities like surgery, radiotherapy, and chemotherapy. Persistent underexpression of salivary gland tissue in smokers could reflect the complex interaction of systemic effects and direct metastatic involvement associated with lung cancer progression.

The Tumor vs. Normal dataset was analyzed to understand the expression patterns in cancerous versus normal tissues. Overexpressed genes were prominently observed in lymphoid tissue, bone marrow, esophagus, stomach, and intestine. These tissues exhibited significant overexpression, reflecting active roles in tumor development and progression. For instance, overexpression in lymphoid tissue may indicate an immune response or lymphoid metastasis in lung cancer patients. Further analysis revealed that significant underexpression was noted in the lung, fallopian tube, adipose tissue, choroid plexus, and bone marrow. The lung tissue showed marked underexpression, emphasizing its central involvement in lung cancer.

Sex-specific differences in gene expression were also observed. Underexpression was noted with a high odds ratio in tissues such as the lung, highlighting differential regulatory mechanisms that may be at play between males and females.

The analysis also identified overlapping comparisons for various tissues and expression types. For example, intestine, lymphoid tissue, and stomach 1 showed significant gene expression changes across multiple comparisons including Age and Tumor vs. Normal. The lung was notable for its significant underexpression in both the Tissue vs. Sex and Tumor vs. Normal comparisons, reflecting its critical role in lung cancer

pathology. The salivary gland was another tissue with significant overlaps, showing underexpression in both Current vs. Never Tumor and Former vs. Never Tumor comparisons, highlighting persistent systemic effects or direct metastatic involvement.

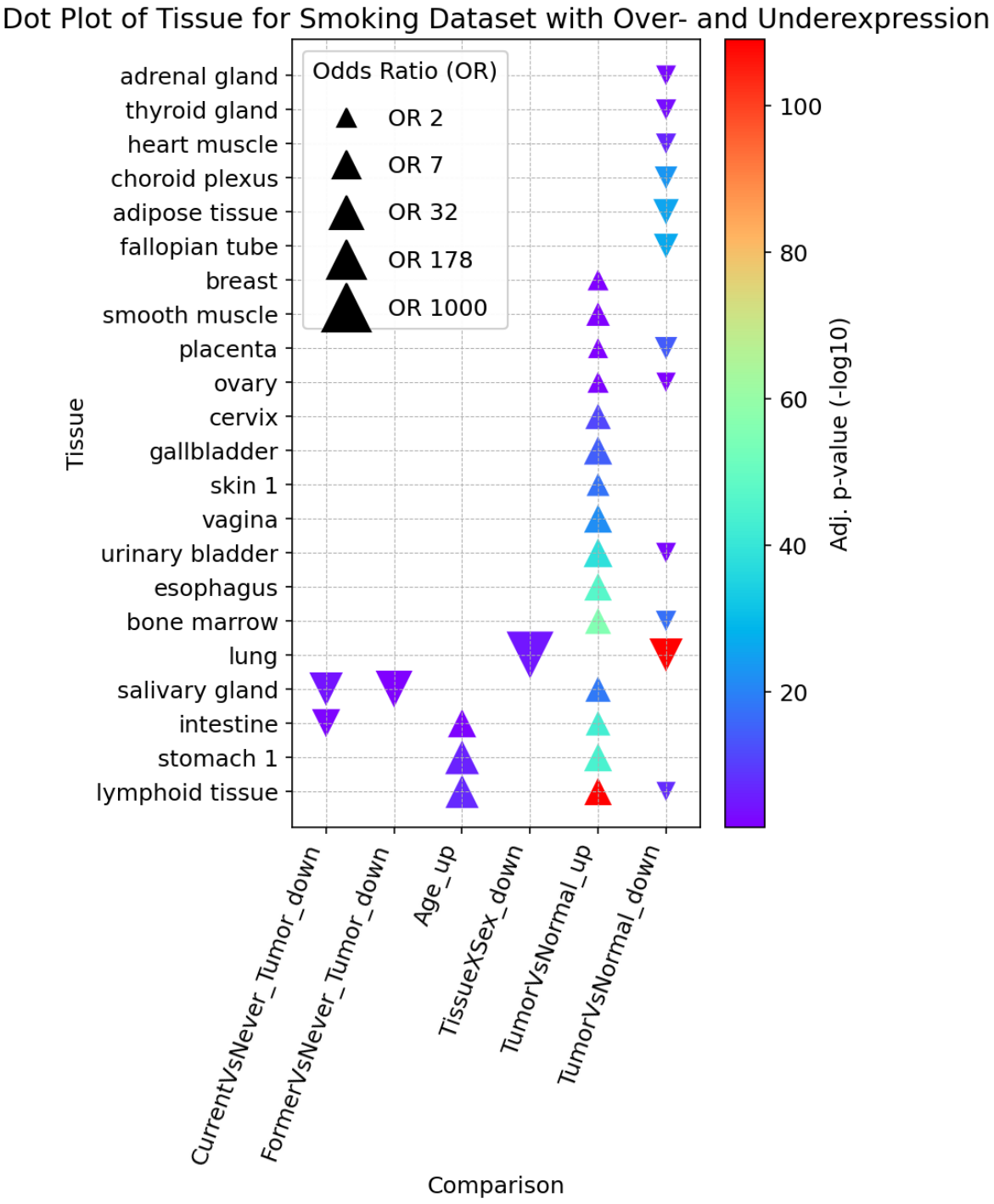


Figure 3.20: **Dot Plot of Differential Gene Expression Analysis in Tissues using Smoking Dataset**. This visualization illustrates the odds ratios (OR) for gene expression, where upward-pointing triangles indicate overexpression and downward-pointing triangles represent underexpression. The size of each symbol correlates with the odds ratio. The accompanying color gradient denotes the adjusted p-value (-log10), highlighting the statistical significance of each gene's differential expression across various tissues.

3.4.10 Comprehensive Analysis of Gene Expression in Numerical HPA Cell Types using Smoking Dataset

The Human Protein Atlas (HPA) provides an invaluable resource for examining the distribution and expression levels of proteins across various human cell types. Understanding these variations is essential for uncovering the underlying biological mechanisms and their implications for health and disease. This section presents a detailed analysis of gene expression variations across different cell types using numerical HPA data, focusing on how gene expression is influenced by smoking status, age, sex-specific differences, and tumor presence, as shown in Figure 3.21.

In the comparison between current and never smokers with tumors, several cell types exhibit significant underexpression. Distal enterocytes show a loss of normal cellular function, possibly due to smoking-induced damage or cancer-related changes. Similarly, proximal enterocytes demonstrate underexpression, which may indicate compromised gut-lung axis interactions and overall cellular health (Haldar et al., 2023). Cholangiocytes exhibit reduced expression, highlighting systemic effects of smoking and its indirect impact on lung tissue. Serous glandular cells' underexpression points to potential disruptions in glandular secretions, which could influence lung mucosal environments and cancer risk. Myeloid dendritic cells show decreased expression, suggesting impaired antigen presentation and immune surveillance in the lung microenvironment (Hato et al., 2024).

Comparing former smokers to never smokers with tumors reveals significant overexpression in specific cell types. Basal respiratory cells have high odds ratios indicating substantial overexpression, reflecting their role in maintaining respiratory epithelium and potential involvement in tumorigenesis. Exocrine glandular cells show increased secretory activity in response to past smoking, contributing to a pro-tumorigenic environment.

In the Tumor vs. Normal comparisons, there are notable findings for both overexpressed and underexpressed genes. Overexpressed genes in plasma cells reflect increased antibody production and immune response, potentially aiding in tumor progression. Memory B-cells exhibit heightened expression, suggesting an active immune role and possibly contributing to inflammation and tumor microenvironment modulation. Erythroid cells show increased expression, likely related to compensatory responses to hypoxia within tumors, promoting angiogenesis and tumor survival (Shevchenko et al., 2023). Naive B-cells' overexpression signifies enhanced immune activation, influencing tumor-immune interactions. B-cells generally show robust overexpression, indicating a significant immune response that may affect tumor development and progression.

Conversely, underexpressed genes in adipocytes suggest metabolic alterations and loss of adipose-related signaling in the tumor environment. Endothelial cells show underexpression, reflecting compromised vascular function critical for tumor growth and metastasis. Monocytes exhibit decreased expression, indicating impaired immune responses and reduced phagocytic activity within the tumor microenvironment. Lymphatic endothelial cells' underexpression suggests disrupted lymphatic function, impacting immune surveillance and fluid balance in the lungs. Alveolar cells type 1 show reduced expression, affecting gas exchange and indicating significant functional loss in lung cancer.

Significant age-related overexpression is observed in various cell types, such as memory B-cells, naive B-cells, plasma cells, and B-cells. These high odds ratios indicate enhanced metabolic activity or stress responses in these cells as age advances, suggesting that aging leads to increased metabolic activity or stress responses. This reflects an attempt to counteract age-related declines in function or increased exposure to damaging agents over time.

In the context of tissue and sex interactions, significant underexpression is observed in alveolar cells type 2 and type 1. This underexpression suggests a loss of surfactant production, critical for lung function, potentially exacerbating cancer-related lung dysfunction. Reduced expression affects gas exchange, indicating significant lung impairment in the context of cancer and sex differences.

Several cell types appear in more than one comparison, highlighting their critical roles. Memory B-cells are involved in both Age_up and TumorVsNormal_up comparisons, indicating their importance in aging and cancer. Naive B-cells appear in Age_up and TumorVsNormal_up, suggesting their role in immune responses across different conditions. Plasma cells are found in Age_up and TumorVsNormal_up, reflecting their significant role in antibody production and cancer progression. B-cells are present in Age_up and TumorVsNormal_up, highlighting their importance in immune surveillance and tumor interaction. Alveolar cells type 1 appear in both TissueXSex_down and TumorVsNormal_down, indicating their crucial function in lung health and disease.

This comprehensive analysis elucidates the complex regulatory mechanisms underlying cell type-specific gene expression and identifies potential targets for therapeutic intervention, particularly in smoking-related diseases and cancer. Overexpressed genes in smoking-affected cell types may reflect compensatory mechanisms or increased demand for specific functions, while underexpressed genes could indicate declines in critical pathways or cellular functions. Understanding sex-specific and age-related differences in gene expression is crucial for developing targeted treatments and interventions. Identifying overexpressed genes in tumor cell types suggests potential therapeutic targets, whereas underexpressed genes may represent lost tumor suppressor functions.

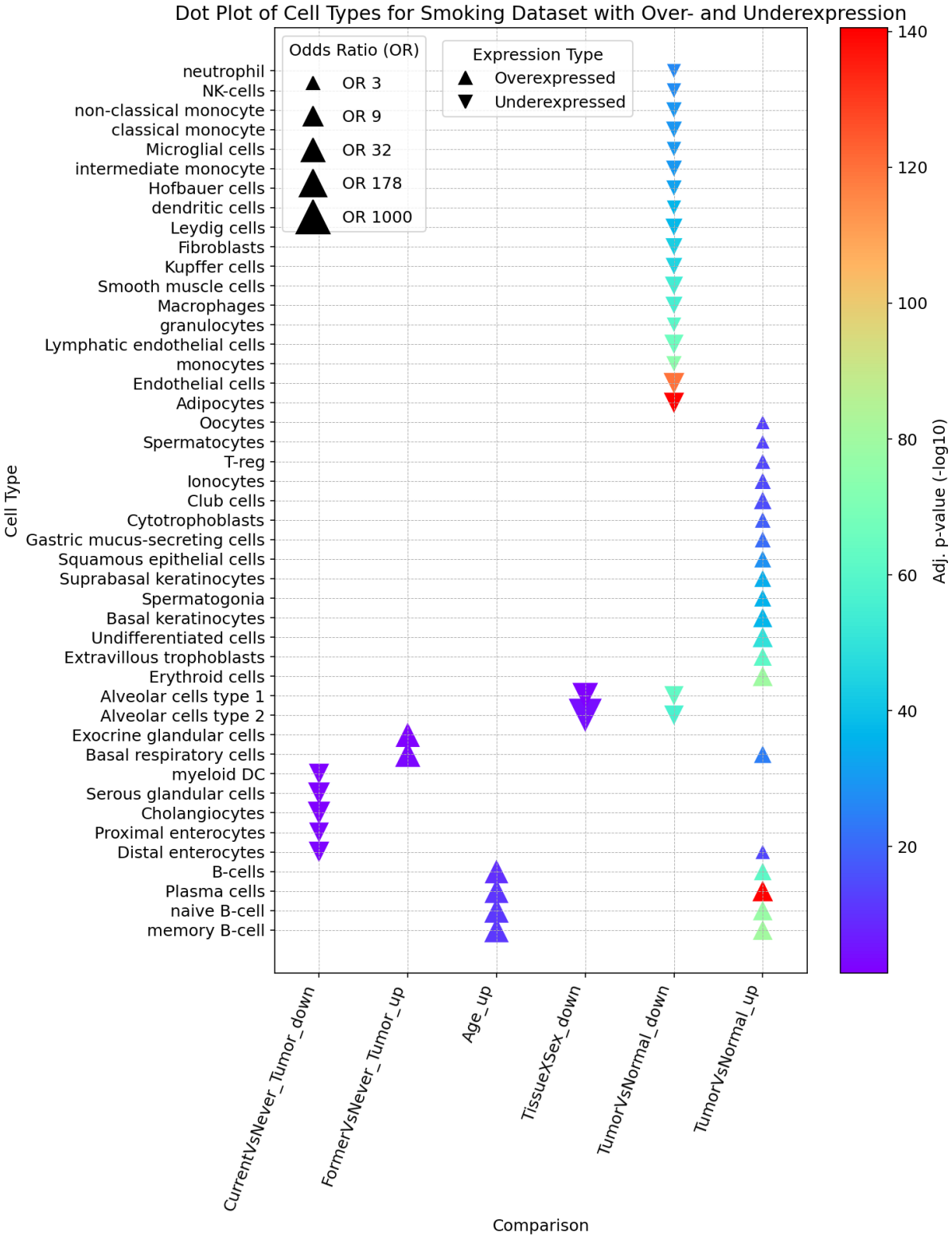


Figure 3.21: **Dot Plot of Top 20 differential gene expression analysis in cell types for the smoking dataset with over- and underexpression**. The triangles indicate over- or underexpression. Triangles pointing upward represent overexpression, while those pointing downward represent underexpression. Each triangle is color-coded according to a scale that represents -log10(p-value), indicating the statistical significance of the over- or underexpression.

3.4.11 Analysis of Shared Qualitative and Numerical Marker Genes

The study of cell type ontologies associated with lung cancer through both qualitative and numerical data from the Human Protein Atlas (HPA), as well as qualitative data from the Human Ensemble Cell Atlas (hECA) enriches our understanding of the molecular landscape within the lung cancer microenvironment. By visualizing the odds ratios and adjusted p-values across different cell types, as illustrated in Figure 3.22, variations in gene expression under diverse experimental conditions can be examined.

Significant findings from this analysis include the following observations. Alveolar cells type 2 are markedly downregulated in TissueXSex interactions and when comparing tumor versus normal tissues, which may reflect the impact of sex-specific factors and the aggressive nature of tumor growth on alveolar function. Endothelial cells exhibit significant downregulation in tumor versus normal tissue comparisons, indicating possible alterations in vascular structures within tumors or changes in angiogenic signaling. Macrophages and smooth muscle cells demonstrate notable changes in TumorVsNormal comparisons, possibly linked to their roles in tumor-stroma interactions and structural integrity of lung tissue (Cao et al., 2024; Ramamonjisoa & Ackerstaff, 2017).

B-cells show enhanced expression levels in both age-related conditions and tumor versus normal comparisons, particularly noted in the hECA and numerical HPA marker genes. This suggests a potential role of B-cells in age-associated immune responses and their adaptation within the tumor microenvironment. In an intriguing anomaly observed in our empirical findings, B-cells displayed both upregulation and downregulation in the numerical tumor versus normal comparisons conducted using the Human Protein Atlas (HPA). Specifically, the comparison "Num_TumorVsNormal_up" shows B-cells being significantly upregulated with an odds ratio of approximately 6.97 and a high adjusted p-value (-

log10) of 49.52, suggesting a robust overexpression in certain tumor environments. Conversely, the same cell type in the comparison "Num_TumorVsNormal_down" exhibits downregulation with a much lower odds ratio of about 1.55 and an adjusted p-value (-log10) of 3.20, indicating a relatively subdued expression. Appendix A.1 has a table of all adjusted p-values in scientific notation. This division in B-cell expression, characterized by distinct upregulation and downregulation within the same experimental framework, hints at complex, context-dependent roles of B-cells in the tumor microenvironment. Notably, the underexpressed B-cells demonstrate significantly lower odds ratios and p-values compared to their overexpressed counterparts, which may reflect variations in immune responses or cellular adaptation mechanisms triggered by different tumor microenvironments. Research indicates that B cells can contribute significantly to tumor immunity through various mechanisms, including the production of antibodies and cytokines, and modulation of T-cell responses (Zhang et al., 2023). These functionalities underscore the dual nature of B-cell activities, where they can both support and inhibit tumor growth depending on their state and the surrounding microenvironmental conditions.

The behavior of fibroblasts within the tumor versus normal comparisons reveals significant insights into their role in lung cancer. In the numerical comparisons from HPA, fibroblasts were observed both upregulated and downregulated, suggesting variable interactions with the tumor microenvironment. Specifically, in the comparison "Num_TumorVsNormal_down," fibroblasts demonstrated a substantial upregulation with an odds ratio of approximately 6.09 and a high adjusted p-value (expressed as -log10) of 43.75, indicating a strong association under these conditions. Conversely, the "Num_TumorVsNormal_up" comparison showed a lesser degree of upregulation with an odds ratio of about 2.17 and a lower adjusted p-value of 2.72.

Further analysis using qualitative data from hECA also showed fibroblasts experiencing both upregulation and downregulation. The "TumorVsNormal_hECA_down" comparison displayed a moderate upregulation with an odds ratio of 5.26 and an adjusted p-value of 2.37, while the "TumorVsNormal_hECA_up" comparison exhibited a more pronounced upregulation with an odds ratio of 11.79 and an adjusted p-value of 3.16.

These observations underline the dual nature of fibroblast behavior in lung cancer, potentially contributing to both tumor support through the construction of tumor microenvironments and resistance against tumor progression. Fibroblasts, particularly cancer-associated fibroblasts (CAFs), have been shown to contribute both to tumor progression and resistance mechanisms. They are implicated in various oncogenic processes such as angiogenesis, invasion, metastasis, and modulation of therapy resistance. The identification of specific fibroblast subtypes and their signaling pathways offers potential targets for therapeutic intervention, aimed at manipulating their tumor-promoting and -resisting roles (Fiori et al., 2019; Joshi et al., 2021). The variability in their expression and the significance of their regulatory effects highlight the need for further investigation into the specific signals and pathways that govern fibroblast activity in different tumor conditions. This nuanced understanding could lead to targeted therapies that manipulate fibroblast functions to hinder tumor growth and progression.

These observations underscore the complex interplay between different cell types in the lung and their responses to both intrinsic factors like sex and age and extrinsic pressures such as tumor presence. The implications of these findings are profound, suggesting that targeted therapies need to consider not only the heterogeneous nature of lung tumors but also the diversity of the cellular landscape in which these tumors exist.

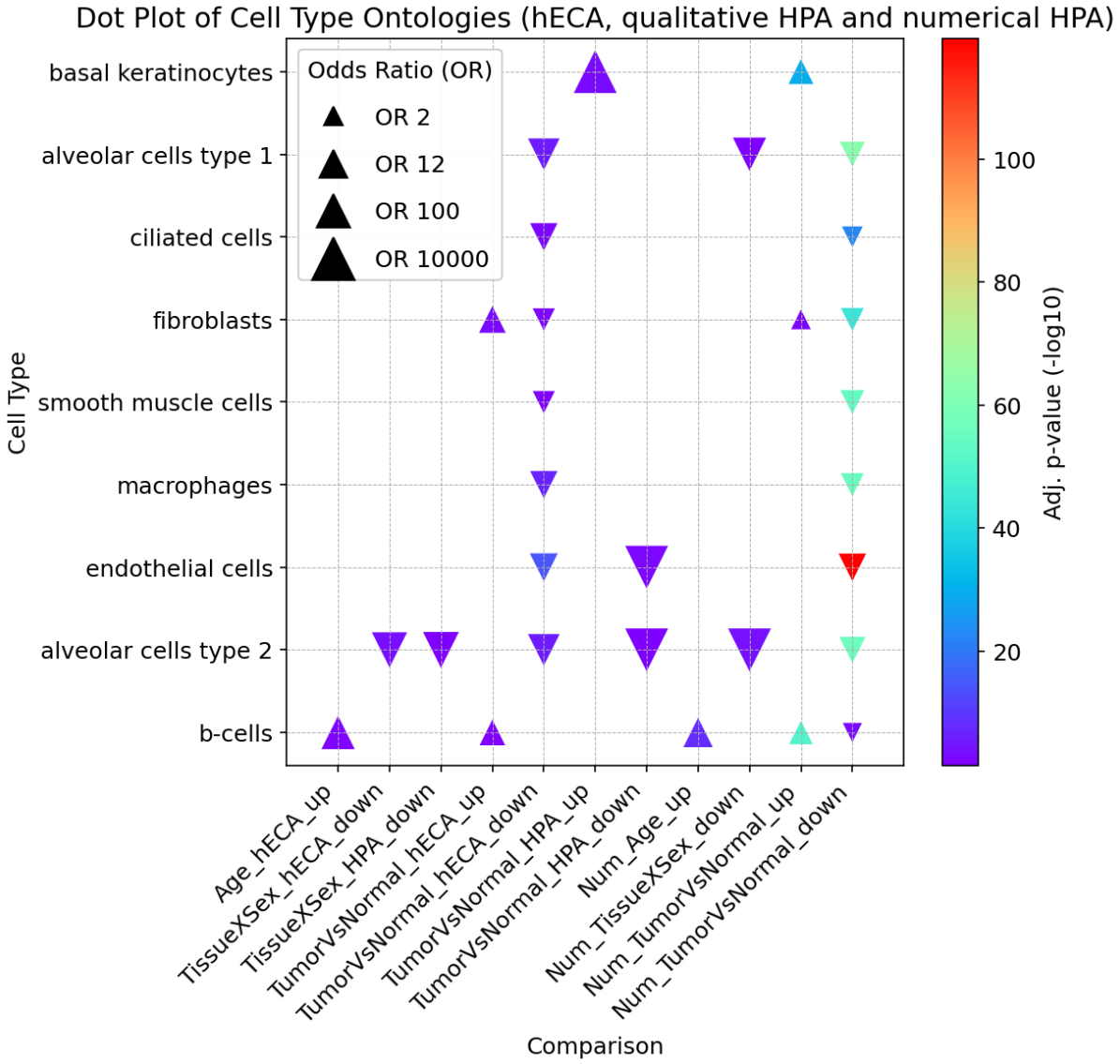


Figure 3.22: **Dot plot of cell type ontologies illustrating gene expression variations in lung cancer, using data from the Human Ensemble Cell Atlas (hECA) and the Human Protein Atlas (HPA)**. This plot arranges cell types vertically and experimental conditions horizontally, distinguishing "Num_" prefixed measurements from numerical HPA marker genes and others from qualitative hECA and HPA markers. Only cell types shared between numerical and qualitative marker genes are shown. Dot sizes indicate the Odds Ratio, illustrating expression strength, while the color gradient shows statistical significance (-log10 adj. p-value scale).

4 Discussion

This chapter evaluates the significant findings of the study, contextualizing them within the broader scientific literature on lung cancer genomics and assessing their implications for future research and clinical practice. Through an in-depth analysis of differentially expressed genes between tumor and normal lung tissues, and the elucidation of the molecular impacts of smoking on gene expression, this chapter seeks to bridge empirical data with established theoretical frameworks. It explores the interplay between genetic factors and environmental influences, enhancing our understanding of lung cancer pathophysiology.

The findings from the use of the *CellTypeGenomics* package, which facilitated the analysis of complex genomic data from The Human Protein Atlas (HPA) and The Cancer Genome Atlas (TCGA), are evaluated. This evaluation assesses the tool's efficacy in identifying and analyzing cell-type origins of differentially expressed genes, providing a cornerstone for understanding the cellular dynamics at play in tumor environments versus normal tissues. Furthermore, the impact of smoking on gene expression patterns offers insight into how environmental factors modify genomic landscapes, which is essential for comprehending the variability in tumor biology and patient responses to treatments.

Comparative analyses between over-representation analysis and cellular deconvolution are presented to highlight the methodological strengths and potential areas for further enhancement. The synthesis of findings from these comparisons establishes a perspective on the current research landscape and the contributions of this study.

Moreover, this chapter discusses the clinical ramifications of the research, considering the potential for the findings to inform early detection strategies, prognosis, and personalized therapy approaches. It critically

evaluates the strengths and limitations of the study, offering a balanced view that underscores both the scientific advancements achieved, and the challenges encountered. This chapter provides an integrated analysis of the study's findings with a broad discussion on the cell-type origins of differentially expressed genes in lung cancer. It connects empirical data with the theoretical constructs that have traditionally guided lung cancer research, illuminating the interplay between genetic dynamics and environmental influences such as smoking. This examination not only deepens our understanding of the molecular underpinnings of lung cancer but also evaluates the implications for future research and clinical practice.

By evaluating the significance of the identified genes and their cellular origins, this discussion contextualizes the results within the broader field of lung cancer genomics. It reflects on how these insights enhance our comprehension of tumor biology and patient variability, supported by meticulous comparisons with existing literature and the integration of bioinformatics tools. This approach validates the research methodology and results, explores the clinical implications of the findings, and critically assesses the strengths and limitations of the study. This chapter underscores the contributions of this research to the field and outlines the path forward for subsequent investigations.

4.1 Interpretation of Results

The overarching aim of this thesis was to explore the cell-type origins of differentially expressed genes associated with lung cancer, with a particular focus on how smoking impacts gene expression. Leveraging data from The Human Protein Atlas (HPA) and The Cancer Genome Atlas (TCGA), the study has provided significant insights into the molecular mechanisms underpinning lung cancer.

4.1.1 Differential Expression in Tumor vs. Normal Tissue

This thesis utilizes data from The Cancer Genome Atlas (TCGA) to elucidate significant differences in gene expression between tumor and normal lung tissues. Incorporating analyses on the impact of smoking enriches our understanding of environmental contributions to these molecular disparities. A comprehensive examination reveals distinct patterns of gene expression alterations, particularly emphasizing the consistent upregulation of genes fundamental to cancer progression, such as those governing cell cycle control, DNA repair, and apoptosis.

A key finding from this study is the significant overexpression of bronchial epithelium basal cells in tumor tissues, suggesting their pivotal role in cancer progression and their potential utility as biomarkers for detecting malignant transformations within lung tissue. In contrast, type II alveolar cells are notably underexpressed, signifying a loss of their normal physiological roles under oncogenic stress, which could lead to impaired lung function and alterations in the tumor microenvironment.

The Reactome Pathway Analysis further enriches these observations by highlighting the predominant overexpression of cell cycle pathways, elucidating their role in driving the uncontrolled cellular proliferation characteristic of malignancies. This analysis also reveals a marked underexpression of immune-related pathways, indicating sophisticated mechanisms by which tumors may evade immune detection.

Further insights from Biological Processes (BP) derived from Gene Ontology (GO) underscore the significant overexpression of mitotic cell cycle processes and immune-related functions, such as immunoglobulin-mediated responses and B cell-mediated immunity. These findings not only depict the aggressive nature of tumor cells but also reveal potential therapeutic targets that could harness these immune interactions to combat cancer more effectively.

Analyses of Cellular Components (CC) and Molecular Functions (MF) highlight the crucial roles of the immunoglobulin complex and antigen-binding functions, which are significantly active within tumor biology. These components offer unique opportunities for developing targeted therapies that could disrupt these interactions to mitigate tumor growth and spread.

The analysis of Differential Gene Expression (DEG) between tumor and normal tissues provides insights into the consistent upregulation of a diverse array of cell types. Notably, extravillous trophoblasts and erythroid cells demonstrate substantial overexpression. The presence of extravillous trophoblasts, typically associated with placental biology, may suggest mechanisms of invasive behavior akin to tumor cells, illuminating aspects of metastatic processes. Similarly, the upregulation of erythroid cells might reflect changes in oxygenation within the tumor microenvironment, potentially affecting tumor growth and treatment responses. Furthermore, the significant upregulation of cell types such as plasma cells, memory B-cells, naive B-cells, B-cells, suprabasal keratinocytes, basal keratinocytes, and squamous epithelial cells in lung cancer tissue provides essential insights into the disease's pathophysiology. The pronounced activity of various B-cell types, particularly plasma cells, suggests a robust immune response to tumor antigens, critical for developing immunotherapy strategies. The pronounced expression of keratinocytes and squamous epithelial cells points to disruptions in epithelial cell differentiation and proliferation, common features of lung carcinogenesis.

Moreover, the analysis reveals notable underexpression of genes typically expressed in adipocytes, endothelial cells, and monocytes during tumorigenesis, highlighting a loss of normal physiological functions. This underexpression underscores the complex interplay between oncogenic signals and the cellular environment, suggesting that tumor progression

involves not only the activation of oncogenic pathways but also the suppression of normal cellular functions.

This integrated analysis, propelled by the analytical capabilities of the *CellTypeGenomics* package, offers a nuanced understanding of the molecular differences between tumor and normal lung tissues, including the specific exacerbating effects of smoking. By detailing how key genes and pathways are altered in lung cancer, this research not only deepens our molecular understanding of the disease but also highlights critical targets for enhancing diagnostic and therapeutic strategies. These insights emphasize the clinical relevance of the molecular differences identified, suggesting their significant potential to impact lung cancer management and treatment outcomes effectively. This comprehensive examination sets a precedent for considering both genetic and environmental factors in cancer research, paving the way for more personalized and precise oncological interventions.

4.1.2 Differential Expression in Tissue X Sex

The investigation into differential gene expression across tissue types and sexes, particularly in the context of lung cancer, employs a nuanced analytical approach termed Tissue X Sex. This contrast analysis is essential as it offers more detailed insights than straightforward comparisons such as Male vs. Female or Tumor vs. Normal. Such contrasts are crucial for revealing complex dynamics in gene expression that are influenced by both sex and the disease state, providing a deeper understanding of the underlying molecular mechanisms.

The Tissue X Sex contrast is precisely defined by the following mathematical expression:

$$\text{TissueXSex} = (T_M - T_F) - (N_M - N_F)$$

In this formula, T_M and T_F denote gene expression levels in tumor tissues from males and females, respectively, while N_M and N_F represent expression levels in normal tissues from males and females, respectively. This setup allows for an assessment of how sex differences influence gene expression in tumor tissues as compared to normal tissues.

To further elucidate the effects of these sex-based differences, four distinct scenarios have been identified, each reflecting a unique pattern of differential expression:

1. **Greater Negative Deviation in Tumor than in Normal** ($T_M<T_F$ and $|T_M-T_F|>|N_M-N_F|$): This scenario suggests a more pronounced decrease in gene expression in male tumors than in female tumors, and to a greater extent than observed in normal tissues, pointing to enhanced gene repression in male tumors.
2. **Less Pronounced Positive Deviation in Tumor than in Normal** ($T_M<T_F$ and $|T_M-T_F|<|N_M-N_F|$): Here, the increase in gene expression in tumors is less pronounced compared to that in normal tissues, indicating a subdued activation of expression in the tumor environment.
3. **Greater Positive Deviation in Tumor than in Normal** ($T_M>T_F$ and $|T_M-T_F|>|N_M-N_F|$): This pattern demonstrates an increase in gene expression in male tumors compared to female tumors, which is greater than the difference observed in normal tissues. This suggests an active sex-specific regulatory mechanism modifying gene expression in the tumor environment.
4. **Lesser Positive Deviation in Tumor than in Normal** ($T_M>T_F$ and $|T_M-T_F|<|N_M-N_F|$): This final scenario shows increased gene expression in male tumors compared to female tumors, but with a less severe difference than in normal tissues, implying a decrease in positive regulatory influences within the tumor setting.

To further illustrate the process of determining which of the identified scenarios applies when a cell type is underexpressed, consider the

example of Alveolar cells type 2 as presented in Chapter 3.3.10. This cell type is underexpressed in the Tissue X Sex contrast. Utilizing the *CellTypeGenomics* package, a comprehensive Fisher test analysis was conducted, resulting in the identification of 12 genes that align with the first scenario. This scenario highlights that Alveolar cells type 2 are underexpressed in male tumors compared to female tumors, a pattern not observed in normal tissues. This consistent underexpression of Alveolar cells type 2 in male tumors, but not in normal tissues, suggests the existence of a sex-specific molecular pathway potentially significant in the pathogenesis of lung cancer.

The detailed analysis of the significant 43 Ensembl IDs derived from the Tissue X Sex regression contrast reveals important insights into gene expression patterns influenced by sex in lung cancer. Among these significant IDs, 15 fall under Scenario 1, indicating a greater negative deviation in tumors compared to normal tissues. Additionally, 3 Ensembl IDs correspond to Scenario 4, signifying a lesser negative deviation in tumors than in normal tissues.

For overexpressed genes, 25 Ensembl IDs align with Scenario 3. This scenario represents a greater positive deviation in tumors than in normal tissues. These findings suggest a complex regulation of gene expression influenced by both sex and the disease state in lung cancer. The increased expression of these genes in male tumors compared to female tumors, coupled with their lower expression in normal tissues, implies an active role in sex-specific tumorigenesis driven by regulatory mechanisms unique to the tumor environment.

Understanding such patterns is crucial for developing targeted therapeutic strategies that effectively account for sex-specific variations in gene expression. By identifying specific genes that follow distinct patterns of differential expression, researchers can pinpoint potential targets for therapeutic intervention. This strategy enhances the precision of

treatments and underscores the necessity of integrating molecular diagnostics into clinical practices to optimize patient outcomes.

4.1.3 Impact of Smoking on Gene Expression

Chapter 3.4 provides a comprehensive analysis of how smoking status profoundly influences gene expression within lung cancer tissues, utilizing extensive data from the Human Protein Atlas (HPA). This analysis underscores the complexity of gene expression changes across various cell types and elucidates the intricate relationships between smoking, tumor development, and tissue-specific variations. The findings indicate that smoking not only triggers oncogenic processes but also leaves a lasting molecular imprint that significantly shapes the disease's progression and response to treatment.

The study identified 671 genes significantly overexpressed in tumor tissues compared to normal tissues, indicating their potential roles in cancer progression and their linkage to smoking. Conversely, 2161 genes were significantly underexpressed in tumor tissues, reflecting a loss of normal cellular functions and possible tumor suppressive properties being overridden by oncogenic processes.

In particular, genes involved in xenobiotic metabolism pathways were notably upregulated among smokers. This overexpression reflects a biological adaptation aimed at detoxifying the myriad of harmful compounds present in tobacco smoke. The upregulation of these pathways highlights the body’s attempt to counteract the carcinogenic effects of smoking-related compounds, which can directly contribute to DNA damage and subsequent cancer initiation.

Reactome Pathway Analysis further highlighted pathways related to cell cycle control, such as Cell Cycle and Mitotic Cell Cycle, which were predominantly overexpressed. This signifies their vital role in the rapid proliferation characteristic of cancer cells. Immune-related pathways also

showed increased activity, suggesting a complex interplay within the tumor microenvironment that could facilitate immune evasion and tumor growth.

Biological Processes from Gene Ontology (GO) analysis revealed that processes involved in the metabolism of xenobiotics by cytochrome P450 were notably overexpressed. This links smoking directly to increased lung cancer risk through the metabolic activation of carcinogens. Sensory perception pathways, particularly those related to chemical stimulus detection, were underexpressed, potentially reducing the lung's ability to detect and respond to carcinogenic threats.

The study also conducted a comprehensive analysis of gene expression in HPA cell types, comparing smokers and non-smokers. Significant underexpression in current smokers was observed in distal enterocytes and myeloid dendritic cells, indicative of smoking-induced damage and compromised immune surveillance. Conversely, notable overexpression in former smokers was observed in basal respiratory cells and exocrine glandular cells, suggesting alterations in respiratory epithelium and secretory processes potentially contributing to a pro-tumorigenic environment. Enhanced expression in plasma cells and memory B-cells across different smoker categories points to an adaptive immune response, while alterations in metabolic pathways underline the physiological impact of smoking on tumor and normal tissues alike.

A focused analysis of qualitative and numerical marker genes provided substantial insights into how cellular responses to smoking influence tumor progression. This detailed evaluation revealed that alveolar cells type 2, crucial for gas exchange, were markedly downregulated in interactions influenced by sex and in comparisons between tumor and normal tissues. This downregulation underscores the impact of sex-specific factors and the aggressive nature of tumor growth on alveolar function, potentially compromising lung function and responsiveness to treatment. Similarly, alveolar cells type 1 also showed significant

downregulation under these conditions, both in the Human Ensemble Cell Atlas (hECA) for qualitative marker genes and in the Human Protein Atlas (HPA) for numerical marker genes. The consistent downregulation across these two cell types raises concerns about the overall integrity of the lung's alveolar structure in the face of tumorigenic stress, possibly leading to diminished lung capacity and impaired respiratory function.

Furthermore, endothelial cells, which form the linings of blood vessels, exhibited significant downregulation in tumor versus normal tissue comparisons, indicating potential alterations in vascular structures or changes in angiogenic signaling within tumors. Such vascular changes are crucial as they could affect tumor blood supply, influencing tumor growth and metastasis potential, thereby highlighting the critical role of the vascular component in cancer progression.

In addition, B-cells, known for their role in the immune response, showed enhanced expression levels in both age-related conditions and in comparisons between tumor and normal tissues. This increase suggests their involvement in age-associated immune responses and their adaptation within the tumor microenvironment. The numerical HPA also found a significant downregulation of B-cells, but at a fraction of the significant value of the upregulated B-cells, indicating a predominant upregulation. The enhanced activity of B-cells could indicate a compensatory mechanism to counteract the immunosuppressive environment created by tumors, or it might reflect an age-related increase in inflammatory responses that inadvertently support tumor progression.

4.1.4 Utilization of Bioinformatics Tools and Integration with Genomic Databases

The integration of The Human Protein Atlas (HPA) and The Cancer Genome Atlas (TCGA) databases facilitated a comprehensive exploration of gene expression variations across different cell types and conditions.

The *CellTypeGenomics* package was pivotal in managing these complex datasets, showcasing its utility in genomic research. The iterative refinement and application of this package underscored its adaptability and robustness, enhancing the statistical integrity and comprehensiveness of the analysis.

Utilizing the TCGA database allowed for a thorough examination of gene expression profiles in lung cancer, while the integration with HPA data provided functional context to the genetic information. This dual-database approach enriched the study by combining genomic and proteomic data, offering a holistic view of the molecular mechanisms underpinning lung cancer.

The *CellTypeGenomics* package enabled precise mapping of differentially expressed genes (DEGs) to specific cell types, facilitating a deeper understanding of cellular dynamics in lung cancer. Its flexibility in handling large datasets and capability for iterative refinement were crucial for the analysis.

The effective use of bioinformatics tools and databases not only enhanced the accuracy and depth of the findings but also demonstrated the power of computational approaches in genomic research. This approach reinforced the reliability of the findings and highlighted the potential of integrated bioinformatics tools in advancing genomic research.

4.2 Contextualization and Synthesis of Findings

This thesis significantly advances our understanding of lung cancer genetics, with a particular focus on how smoking modulates gene expression. The contextualization of these findings draws upon and is compared with significant prior studies, including the influential work by Alexandrov et al. (2016). This synthesis not only confirms previous

observations but also provides new insights that enhance our comprehension of lung cancer's molecular basis.

4.2.1 Comparative Analysis with Other Studies

The observed differential expression aligns with findings from pivotal studies that mapped genomic changes in lung cancers, corroborating the role of specific genes in tumorigenesis. The impact of smoking on gene expression echoes the mutational signatures identified by Alexandrov et al. (2016), which linked smoking with increased mutation burdens and specific mutational signatures in lung cancers. This alignment provides a functional context to the mutational changes induced by smoking, enhancing the understanding of their role in cancer progression. For instance, the upregulation of genes involved in xenobiotic metabolism pathways observed in this study reflects the body's response to detoxify harmful compounds found in tobacco smoke, a process highlighted in the mutational signatures described by Alexandrov et al.

Additionally, studies by Govindan et al. (2012) and Cancer Genome Atlas Research Network (2014) have shown similar patterns of gene expression changes associated with smoking in lung cancer, supporting the findings of this thesis. These studies reinforce the notion that smoking induces widespread genetic alterations that contribute to lung carcinogenesis.

4.2.2 Bridging Molecular Insights with Clinical Observations

The study's molecular insights correlate well with clinical observations regarding differential treatment responses in smokers versus non-smokers. This correlation suggests that gene expression profiles influenced by smoking could affect the tumor microenvironment and response to therapy. For example, the alterations in immune-related genes suggest modifications in the tumor microenvironment that could

affect tumor immunity and potentially offer new targets for immunotherapy in smokers. Understanding these molecular differences can guide the development of targeted therapies, providing a more personalized approach to lung cancer care and improving treatment outcomes.

Studies by Schiller et al. (2002) and Herbst et al. (2008) have documented differences in treatment efficacy between smokers and non-smokers, which may be attributed to the molecular changes identified in this thesis. These clinical observations highlight the potential for gene expression profiles to serve as biomarkers for tailoring treatment strategies.

By contextualizing these findings within existing literature, this thesis underscores the critical role of smoking in shaping the genetic landscape of lung cancer and highlights the importance of integrating genomic and proteomic data to fully understand the impact of environmental factors on cancer development. The alignment of this study's results with those of Alexandrov et al. (2016), Govindan et al. (2012), and Cancer Genome Atlas Research Network (2014) reinforces the validity of the findings and contributes to a deeper, more nuanced understanding of lung cancer genetics in the context of smoking.

4.2.3 Methodology Comparison and Integration

This chapter presents an in-depth comparison between the methodologies employed in this study, specifically Over-Representation Analysis (ORA) using the *CellTypeGenomics* package, and cellular deconvolution methods, exemplified by CIBERSORTx. It further discusses the potential advantages of integrating these methodologies to refine cell type-specific annotations in lung cancer research.

Over-Representation Analysis (ORA) is a statistical technique designed to ascertain if a predefined group of genes, such as those linked to specific cell types, is more frequently represented within a larger gene set than would be expected by chance. This process typically involves comparing a list of differentially expressed genes (DEGs) against a background gene set using statistical tests like the hypergeometric test or Fisher's exact test to determine the probability that the observed gene overlap occurs by chance.

In our methodology, the *CellTypeGenomics* package maps DEGs to cell types using marker genes cataloged in the HPA. This process involves mapping Ensembl gene IDs to their respective cell types based on HPA data, followed by statistical analysis to verify the presence of these cell type-specific genes among the DEGs using Fisher's exact test, complemented by the Benjamini-Hochberg procedure to control the false discovery rate (FDR).

Unlike cellular deconvolution methods like CIBERSORTx, which estimate cell type proportions, ORA as implemented in our study focuses on identifying statistically significant links between DEGs and specific cell types, thereby offering advantages such as precision in annotation through comprehensive databases like the HPA and enhanced statistical rigor from combining Fisher’s exact test with FDR correction.

Cellular deconvolution, particularly through the application of CIBERSORTx, is a sophisticated approach utilized to annotate genes with cell type identities based on bulk RNA sequencing data. CIBERSORTx applies a gene expression matrix based on bulk RNA sequencing data and a signature expression matrix from known cell types (often derived from single-cell RNA sequencing data) to decompose experimental RNA sequencing data into proportions of cell types present, summarized in a cell type matrix. Cellular deconvolution applies the signature matrix and the cell type matrix to explain the gene expression matrix, presenting a prediction problem. This methodology is especially beneficial in analyzing

heterogeneous tissues, where direct measurement of individual cell types is impractical. A notable issue with CIBERSORTx is its limitation to only identify cell types present in the signature matrix, which may not encompass all cell types within a sample. Thus, CIBERSORTx's ability to identify and quantify cell types is confined by the data provided in the signature matrix. Furthermore, CIBERSORTx often exhibits discrepancies in the relative percentages of cell types it predicts, although the relative difference across similar cell types often aligns with other methodologies like Fluorescence-Activated Cell Sorting (FACS).

Potential Integration of Methods

Future research could benefit from integrating cellular deconvolution and over-representation analysis methodologies. Cellular deconvolution estimates cell type proportions within a sample, providing quantitative insights into tissue composition. When combined with over-representation analysis, which identifies significant associations between differentially expressed genes and specific cell types, this approach offers a comprehensive view of the cellular landscape in complex tissues.

Using cellular deconvolution to estimate cell type proportions, followed by over-representation analysis to map gene expression changes to specific cell types, enhances precision in cell type-specific annotations. The *CellTypeGenomics* package, with its detailed gene annotation capabilities using data from the Human Protein Atlas, complements this approach by providing high-resolution insights into the cellular origins of gene expression changes. Furthermore, cellular deconvolution includes quantization per sample, allowing for more precise measurements of gene expression at the cellular level.

In lung cancer research, this combined methodology can elucidate how smoking influences gene expression at the cellular level, leading to refined models of disease progression and treatment response. Such integrated

models are invaluable for developing personalized therapeutic strategies, accounting for tumor heterogeneity and environmental factors.

Beyond lung cancer, this integrated approach can be applied to other cancers and complex diseases, broadening the impact of bioinformatics tools in genomic research. Leveraging detailed gene annotation capabilities and proportional insights enables a nuanced understanding of gene expression dynamics.

4.3 Implications of the Findings

This section explores the practical implications of our findings, discussing how they contribute to lung cancer research, influence clinical applications, and impact broader oncological practices.

4.3.1 Clinical Implications and Therapeutic Opportunities

Our findings provide potential targets for therapeutic intervention through the identification of genes differentially expressed in tumor versus normal tissues. These genes, particularly those involved in cell proliferation and survival pathways, could serve as focal points for the development of targeted drug therapies. Inhibiting these upregulated genes in tumors might effectively reduce tumor growth or improve the response to existing treatments. Additionally, understanding the modulation of gene expression by smoking offers a clear path toward personalized medicine. Given the distinct gene expression profiles associated with different smoking histories, treatment plans could be tailored more precisely to enhance efficacy and minimize side effects, based on a patient's smoking status. This approach would not only personalize treatment but also optimize resource allocation in clinical settings.

4.3.2 Potential for Early Detection and Prognosis

The identification of specific biomarkers from differentially expressed genes in early-stage tumors presents opportunities for early detection. Developing diagnostic tests based on these biomarkers could significantly improve early detection rates, leading to earlier intervention and potentially better patient outcomes. Furthermore, the gene expression profiles linked to smoking status might also provide prognostic tools, helping predict the aggressiveness of the disease and guiding treatment decisions. Early detection biomarkers can also facilitate routine screening, allowing for timely treatment and improved survival rates.

4.3.3 Enhancing the Understanding of Lung Cancer Pathophysiology

The comprehensive analysis enabled by the *CellTypeGenomics* package enhances our understanding of lung cancer's pathophysiology. By elucidating the cellular origins of differentially expressed genes and their association with environmental factors like smoking, this study contributes to a deeper understanding of how external factors can influence cellular behavior and disease progression. These insights are crucial for developing new models of lung cancer that reflect its biological complexity more accurately, potentially influencing both research and clinical approaches to the disease. Additionally, integrating genomic and proteomic data facilitates a more holistic understanding of the interactions between genetic and environmental factors in lung cancer development.

4.3.4 Policy and Public Health Implications

The findings underscore the importance of smoking cessation programs and policies aimed at reducing tobacco use to mitigate lung cancer risk. Public health strategies can be informed by the molecular evidence linking

smoking to specific genetic alterations, reinforcing the need for preventive measures and education. This research supports the development of targeted public health campaigns that address the molecular impact of smoking, potentially reducing the incidence of lung cancer and improving population health outcomes.

4.4 Challenges and Considerations

While the study provided valuable insights into the molecular dynamics of lung cancer, it faced several challenges inherent in the integration and interpretation of complex genomic data. The reliance on public genomic databases such as TCGA and HPA introduced potential biases due to variations in sample collection, data processing, and demographic diversity. These factors could limit the generalizability of the findings, as variations in sample handling and population representation might influence the observed gene expression patterns.

Integrating diverse data sources posed additional challenges, particularly in harmonizing data across different platforms and ensuring consistency in annotations and gene identifiers. These issues underscore the need for robust bioinformatics pipelines capable of managing and standardizing heterogeneous datasets effectively. Iterative refinement and validation of these pipelines are essential to improve the reliability of genomic analyses.

Moreover, interpreting differential gene expression and pathway analysis results requires careful consideration of biological context and experimental conditions. The dynamic nature of gene expression, influenced by both intrinsic and extrinsic factors, necessitates robust validation of computational predictions through experimental methods such as quantitative PCR (qPCR) and functional assays. This validation is crucial to ascertain the biological relevance of identified gene expression changes and their associations with specific cell types and pathways.

To overcome these challenges, continuous efforts are needed to refine data collection and analysis techniques in genomic research. Enhancing the accuracy and completeness of public genomic databases through improved sample collection protocols, standardized data processing methods, and comprehensive metadata annotation is imperative. Additionally, fostering collaborations between computational and experimental biologists can help bridge the gap between theoretical predictions and empirical validation, thus strengthening the overall robustness of genomic studies.

In summary, while this study has provided valuable insights into the molecular dynamics of lung cancer, the challenges encountered highlight the importance of continuous methodological improvements and interdisciplinary approaches in genomic research. Addressing these challenges will ensure that future studies can build on these findings with greater accuracy and generalizability, further advancing our understanding of lung cancer and its underlying mechanisms.

4.5 Strengths and Limitations

4.5.1 Strengths

One of the primary strengths of this study is the utilization of advanced bioinformatics tools, particularly the *CellTypeGenomics* package. This package is not only tailored for cell-type origin studies to ensure relevancy and precision but is also optimized for efficiency, facilitating rapid analysis of extensive gene lists. Being open source, it encourages collaboration and further development within the scientific community, enhancing the potential for innovative approaches. The integration of large-scale genomic data from The Human Protein Atlas (HPA) and The Cancer Genome Atlas (TCGA) provided a robust framework for a comprehensive analysis of gene expression variations. This approach enabled a detailed

exploration of complex genetic interactions and their implications for lung cancer, offering novel insights into how smoking modulates gene expression at the cellular level. Additionally, the statistical rigor applied through methods like Fisher's exact test and the Benjamini-Hochberg procedure ensured the reliability and accuracy of the results.

4.5.2 Limitations

Despite these strengths, the study has certain limitations. The reliance on secondary genomic data sources, such as The Human Protein Atlas (HPA) and The Cancer Genome Atlas (TCGA), introduces potential biases related to data collection and sample heterogeneity. For example, the findings indicate that males are generally diagnosed at older ages, particularly in later stages of lung cancer. This trend may be influenced by several factors, including higher smoking rates among males and potential delays in seeking medical care, leading to later-stage diagnoses. These factors may affect the generalizability of the findings and restrict the ability to draw definitive causal inferences.

Additionally, our approach considers a data-driven gene selection method where the Human Protein Atlas determines a numeric threshold for signature genes. This can be contrasted with methods that employ strong signature genes specific to particular cell types, adding a more qualitative dimension to the analysis. While this method aids in the identification of potential signature genes, it may also introduce biases by potentially overlooking strong, type-specific signature genes that do not meet the numeric threshold. Furthermore, the observational nature of the study limits the ability to establish direct cause-and-effect relationships between smoking and gene expression changes. Future research should aim to incorporate primary data collection and experimental validation to confirm these findings and address the identified biases. Expanding the study to

include diverse populations and additional environmental factors could also enhance the comprehensiveness and applicability of the results.

5 Conclusion

This chapter synthesizes the findings from the investigation into the molecular dynamics of lung cancer, reflecting on the implications of these results for future research and clinical practice. It draws together the key outcomes of the analyses conducted using The Human Protein Atlas (HPA) and The Cancer Genome Atlas (TCGA), facilitated by the capabilities of the *CellTypeGenomics* package. The conclusions drawn not only highlight advances in understanding lung cancer biology but also underscore the potential for these insights to inform more effective and personalized treatment strategies. The following sections detail the principal findings, their broader implications, the inherent challenges encountered, and the recommended directions for future research.

5.1 Summary of Key Findings

The comprehensive investigation conducted in this thesis has yielded significant insights into the differential gene expression between tumor and normal lung tissues, particularly highlighting the impact of smoking on these genetic alterations. Utilizing advanced bioinformatics tools, including the *CellTypeGenomics* package, and genomic data from HPA and TCGA, this study identified key genes that are significantly upregulated or downregulated in tumors, offering valuable insights into the cellular dynamics driving lung cancer progression and revealing potential targets for therapeutic intervention and biomarkers for early detection.

A pivotal aspect of this research was associating these differentially expressed genes with specific cell types, made possible through the utilization of HPA data within the *CellTypeGenomics* package. This analysis has significantly advanced our understanding of the cellular context of

these gene expression changes and their roles in lung cancer. For example, bronchial epithelium basal cells were significantly overexpressed in tumor tissues, suggesting their vital role in cancer progression and their potential as biomarkers for identifying malignant transformations. Conversely, alveolar cells type II exhibited notable underexpression, indicating a loss of their normal physiological roles under oncogenic stress.

Further, the integration of gene ontology (GO) and Reactome pathway analyses categorized these genes according to their roles in biological processes, cellular components, and molecular functions, mapping the disrupted functional pathways in lung cancer. The Reactome Pathway Analysis underscored the predominant overexpression of cell cycle pathways, elucidating their role in fostering uncontrolled cellular proliferation, a hallmark of malignancy. Immune-related pathways showed a marked underexpression, suggesting mechanisms by which tumors evade immune surveillance.

The analysis of Differential Gene Expression (DEG) between tumor and normal lung tissues reveals significant upregulation of various cell types, such as erythroid cells, suggesting altered oxygenation in the tumor microenvironment. Enhanced expression of plasma cells and different B-cell types, alongside keratinocytes, indicates robust immune responses and disruptions in epithelial cell functioning, which are critical for understanding lung cancer pathophysiology and identifying potential therapeutic targets. Conversely, the notable underexpression of genes in adipocytes, endothelial cells, and monocytes suggests a loss of normal functions, highlighting the dual nature of lung cancer progression through oncogenic activation and suppression of regular cellular activities.

The study also revealed sex-specific differences in gene expression, emphasizing the necessity for personalized therapeutic strategies. For instance, the Tissue X Sex contrast analysis identified genes that exhibited different patterns of expression between male and female tumors

compared to normal tissues. This analysis provided insights into sex-specific regulatory mechanisms that influence tumorigenesis, underscoring the importance of considering sex as a biological variable in lung cancer research and treatment.

Overall, the integration of TCGA and HPA data through the *CellTypeGenomics* package has facilitated a nuanced understanding of the molecular differences between tumor and normal lung tissues. This research not only deepens our comprehension of lung cancer's molecular basis but also highlights potential targets for diagnostic and therapeutic strategies, paving the way for more personalized and effective interventions. These insights have significant potential to impact lung cancer management and treatment outcomes, emphasizing the importance of integrating genetic and environmental factors in cancer research.

5.2 Implications and Significance

The findings from this study have profound implications for the scientific understanding and clinical management of lung cancer. By elucidating the differential gene expression between tumor and normal lung tissues, especially in the context of smoking, this research advances our knowledge of the molecular underpinnings of lung cancer. This enhanced understanding supports the development of more precise diagnostic tools and targeted therapeutic strategies.

A key outcome of this study is the demonstration of the *CellTypeGenomics* package as a powerful tool for researchers globally. While this study specifically applies the package to lung cancer and the autoimmune condition psoriasis, there is significant potential for its application across a wide range of diseases. By identifying key genes and pathways that are differentially expressed and associating these changes with specific cell types, the *CellTypeGenomics* package enables the tailoring of treatments

to individual genetic profiles and environmental factors, such as smoking habits. This personalization of therapy is crucial for improving treatment efficacy and patient outcomes.

The detailed insights provided by the *CellTypeGenomics* package into the roles of specific cell types and molecular pathways in disease progression are invaluable for developing new therapeutic strategies. For instance, the identification of overexpressed bronchial epithelium basal cells and underexpressed alveolar cells type II in tumor tissues suggests specific cellular targets for intervention. Targeting these cell types and the pathways they influence could lead to more effective treatments that address the underlying mechanisms of lung cancer.

Moreover, the study's findings on the impact of smoking on gene expression profiles highlight the need for considering smoking history in the molecular profiling of lung cancer patients. Understanding how smoking-induced alterations affect tumor biology can guide the development of therapies that exploit these molecular vulnerabilities, potentially improving the efficacy of treatments for smokers with lung cancer.

Overall, the integration of genomic data from TCGA and HPA, facilitated by the *CellTypeGenomics* package, has provided a comprehensive and nuanced understanding of lung cancer biology. These insights have significant potential to impact clinical practice by informing the development of more personalized and targeted treatment strategies, ultimately improving patient care and outcomes. The *CellTypeGenomics* package stands as a promising resource for researchers worldwide, offering the potential to make better and more informed decisions across a wide range of diseases. This study underscores the importance of integrating genomic and environmental data to refine therapeutic approaches, paving the way for advancements in the precision and effectiveness of disease treatments.

5.3 Future Work

To broaden the impact and enhance the scientific rigor of this research, future studies should prioritize the development and utilization of experimental models to validate the causal relationships suggested in this study. Longitudinal studies are particularly crucial for examining the reversibility of smoking-related gene expression changes. These studies will offer vital insights into the temporal dynamics of gene expression and assess whether the alterations induced by smoking can be mitigated following cessation.

Additionally, integrating survival analyses will be pivotal. Such analyses are designed to explore the relationship between specific gene expression levels and patient survival rates, providing essential insights into how individual genes may influence the progression of cancer. This understanding is particularly significant in oncology, where gene expression data can directly inform therapeutic strategies and prognostication.

The *CellTypeGenomics* package has demonstrated considerable potential in this study, and future work should aim to validate its broader applicability. This package could become a powerful tool for researchers globally, applicable to a wide range of diseases, including various types of cancer, autoimmune conditions, genetic disorders, physiological diseases, degenerative diseases, and pathological infections. Future studies should focus on validating this potential across diverse disease contexts to establish the package's utility in different areas of biomedical research. By doing so, the *CellTypeGenomics* package could significantly advance our understanding of complex diseases and contribute to the development of more effective and personalized therapeutic strategies.

Experimental validation of computational predictions is equally crucial. Techniques such as quantitative PCR (qPCR), Western blotting, and functional assays should be employed to confirm the pathways and

mechanisms identified in this thesis. This step will not only strengthen the reliability of the findings but also deepen our understanding of the molecular impacts of smoking on lung cancer.

Furthermore, the utility of the *CellTypeGenomics* package should be expanded to assess its performance across various cancer types. This exploration will help validate the effectiveness of the developed tools and may lead to notable advancements in cancer diagnostics and treatment.

Incorporating multi-omics approaches—including proteomics, metabolomics, and transcriptomics—will provide a more comprehensive view of the molecular mechanisms at play. This integrative strategy will offer a holistic view of how smoking and other environmental factors influence cellular processes, thereby elucidating the intricate interactions between genetics and the environment in the development of cancer.

The application of machine learning and artificial intelligence (AI) offers a transformative potential for advancing lung cancer genomic research. These technologies can significantly enhance the ability to identify complex patterns and predictive markers that are not apparent through traditional methods. Machine learning models can be trained to predict outcomes such as disease progression, response to treatment, and patient survival rates from gene expression profiles. Moreover, AI can facilitate the integration of multi-omics data, accelerating the analysis process and improving the identification of potential therapeutic targets. Embracing these advanced computational tools will likely advance personalized medicine by enabling treatment plans tailored to the genetic profiles of individual tumors, optimizing therapeutic efficacy while minimizing side effects.

Collaborative efforts with clinical researchers will be essential for translating these computational and experimental findings into clinical practice. Linking predictive models and validations with real patient data and outcomes will help in crafting personalized treatment strategies and

enhancing prognostic tools. Such collaborative endeavors will ultimately improve patient care in oncology, effectively bridging the gap between research findings and their practical application in clinical settings.

References

Afshar-Kharghan, V. (2017). The role of the complement system in cancer. *The Journal of Clinical Investigation*, 127(3), 780-789.

Alexandrov, L. B., Ju, Y. S., Haase, K., Van Loo, P., Martincorena, I., Nik-Zainal, S., ... & Stratton, M. R. (2016). Mutational signatures associated with tobacco smoking in human cancer. *Science*, 354(6312), 618-622.

Altschuler, S. J., & Wu, L. F. (2010). Cellular heterogeneity: do differences make a difference? *Cell*, 141(4), 559–563. https://doi.org/10.1016/j.cell.2010.04.033

Avila Cobos, F., Vandesompele, J., Mestdagh, P., & De Preter, K. (2018). Computational deconvolution of transcriptomics data from mixed cell populations. *Bioinformatics*, 34(11), 1969-1979. https://doi.org/10.1093/bioinformatics/bty019

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*. Series B (Methodological), 57(1), 289-300.

Bellott, D. W., Hughes, J. F., Skaletsky, H., Brown, L. G., Pyntikova, T., Cho, T. J., ... & Page, D. C. (2014). Mammalian Y chromosomes retain widely expressed dosage-sensitive regulators. *Nature*, *508*(7497), 494-499.

Bernauer, U., Heinrich-Hirsch, B., Tönnies, M., Peter-Matthias, W., & Gundert-Remy, U. (2006). Characterisation of the xenobiotic-metabolizing Cytochrome P450 expression pattern in human lung tissue by immunochemical and activity determination. *Toxicology letters*, 164(3), 278-288.

Bhandari, N., Walambe, R., Kotecha, K., & Khare, S. P. (2022). A comprehensive survey on computational learning methods for analysis of gene expression data. *Frontiers in Molecular Biosciences*, 9, 907150.

Blackburn, E. H. (2005). Telomeres and telomerase: their mechanisms of action and the effects of altering their functions. *FEBS letters*, 579(4), 859-862.

Brown, C. J., Ballabio, A., Rupert, J. L., Lafreniere, R. G., Grompe, M., Tonlorenzi, R., & Willard, H. F. (1991). A gene from the region of the human X inactivation centre is expressed exclusively from the inactive X chromosome. *Nature*, 349(6304), 38-44.

Cancer Genome Atlas Research Network. (2014). Comprehensive molecular profiling of lung adenocarcinoma. *Nature*, 511(7511), 543-550.

Cao, L., Meng, X., Zhang, Z., Liu, Z., & He, Y. (2024). Macrophage heterogeneity and its interactions with stromal cells in tumour microenvironment. *Cell & Bioscience*, 14(1), 16.

Carregaro, F., Stefanini, A. C. B., Henrique, T., & Tajara, E. H. (2013). Study of small proline-rich proteins (SPRRs) in health and disease: a review of the literature. *Archives of Dermatological Research*, 305, 857-866.

Chaudhary, P., Xu, X., Wang, G., Hoj, J. P., Rampersad, R. R., Asselin-Labat, M. L., ... & Onaitis, M. W. (2023). Activation of KrasG12D in subset of alveolar Type II cells enhances cellular plasticity in lung adenocarcinoma. *Cancer Research Communications*, 3(11), 2400-2411.

Chen, S., Luo, J., Gao, H., Li, F., Chen, Y., Li, J.,…& Zhang, X. (2022). hECA: The cell-centric assembly of a cell atlas. *iScience*, 25(1). https://doi.org/10.1016/j.isci.2022.104318

Chen, Z., Zhao, M., Li, M., Sui, Q., Bian, Y., Liang, J., Hu, Z., Zheng, Y., Lu, T., Huang, Y., Zhan, C., Jiang, W., Wang, Q., & Tan, L. (2020). Identification of differentially expressed genes in lung adenocarcinoma

cells using single-cell RNA sequencing not detected using traditional RNA sequencing and microarray. *Laboratory Investigation*, 100(10), 1318-1329. https://doi.org/10.1038/s41374-020-0428-1

Chiang, S. K., Chen, S. E., & Chang, L. C. (2021). The role of HO-1 and its crosstalk with oxidative stress in cancer cell survival. *Cells*, 10(9), 2401.

Conesa, A., et al. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biology*, 17(1), 13.

de Mol, J., Kuiper, J., Tsiantoulas, D., & Foks, A. C. (2021). The dynamics of B cell aging in health and disease. *Frontiers in Immunology*, 12, 733566.

Doll, R. & Hill, AB. (1950). Smoking and carcinoma of the lung. *British Medical Journal*.

Chavez-Dominguez, R., Perez-Medina, M., Aguilar-Cazares, D., Galicia-Velasco, M., Meneses-Flores, M., Islas-Vazquez, L., ... & Lopez-Gonzalez, J. S. (2021). Old and new players of inflammation and their relationship with cancer development. *Frontiers in Oncology*, 11, 722999.

Dransfield, B. & Brightwell, B. (n.d.). Fisher's exact test: Use & misuse. (2x2 contingency table, fixed factors, test of association). https://influentialpoints.com/Training/Fishers_exact_test_use_and_misuse.htm

Durmaz, A. A., Karaca, E., Demkow, U., Toruner, G., Schoumans, J., & Cogulu, O. (2015). Evolution of genetic techniques: past, present, and beyond. *BioMed research international*, 2015, 461524. https://doi.org/10.1155/2015/461524

Evan, G. I., & Vousden, K. H. (2001). Proliferation, cell cycle and apoptosis in cancer. *Nature*, 411(6835), 342-348.

Fiori, M. E., Di Franco, S., Villanova, L., Bianca, P., Stassi, G., & De Maria, R. (2019). Cancer-associated fibroblasts as abettors of tumor progression

at the crossroads of EMT and therapy resistance. *Molecular Cancer*, 18, 1-16.

Fisher, R. A. (1922). On the interpretation of χ2 from contingency tables, and the calculation of P. *Journal of the Royal Statistical Society*, 85(1), 87-94.

Frost, H. R. (2021). Analyzing cancer gene expression data through the lens of normal tissue-specificity. PLoS Computational Biology, 17(6), e1009085.

Føleide, L. & Mittet, A. (2023). Cell Type Origin of Differentially Expressed Genes (specialization project). Unpublished.

Føleide, L. (2024). *CellTypeGenomics*: Source Code and Documentation [Software and documentation]. Retrieved from https://github.com/Zyron/*CellTypeGenomics*

Gasperskaja, E. & Kučinskas, V. (2017). The most common technologies and tools for functional genome analysis. *Acta medica Lituanica*, *24*(1), 1–11. https://doi.org/10.6001/actamedica.v24i1.3457

Gelfand, J. M., Neimann, A. L., Shin, D. B., Wang, X., Margolis, D. J., & Troxel, A. B. (2006). Risk of myocardial infarction in patients with psoriasis. *Jama*, 296(14), 1735-1741.

Gene Ontology Consortium. (2021). The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Research*, 49(D1), D325-D334. https://doi.org/10.1093/nar/gkaa1113

Govindan, R., Ding, L., Griffith, M., Subramanian, J., Dees, N. D., Kanchi, K. L., ... & Wilson, R. K. (2012). Genomic landscape of non-small cell lung cancer in smokers and never-smokers. *Cell*, 150(6), 1121-1134.

Goel, S., DeCristo, M. J., McAllister, S. S., & Zhao, J. J. (2018). CDK4/6 inhibition in cancer: beyond cell cycle arrest. *Trends in cell biology*, 28(11), 911-925.

Grummt, I. (2003). Life on a planet of its own: regulation of RNA polymerase I transcription in the nucleolus. *Genes & development*, 17(14), 1691-1702.

Gubanova, N. V., Orlova, N.G., Dergilev A. I., Oparina, N. Y. & Orlov, Y. O. (2021). Glioblastoma gene network reconstruction and ontology analysis by online bioinformatics tools. *Journal of Integrative Bioinformatics*. https://doi.org/10.1515/jib-2021-0031

Guinee, D. G. (2018). Lymphoid Lesions of the Lung. In D. S. Zander & C. F. Farver (Eds.), *Pulmonary Pathology (Second Edition)* (pp. 445-485.e1). Elsevier. https://doi.org/10.1016/B978-0-323-39308-9.00022-4

Gupta, R. (2023). Accuracy, Precision, Recall, F-1 Score, Confusion Matrix, and AUC-ROC. *Medium*. https://medium.com/@riteshgupta.ai/accuracy-precision-recall-f-1-score-confusion-matrix-and-auc-roc-1471e9269b7d

Haldar, S., Jadhav, S. R., Gulati, V., Beale, D. J., Balkrishna, A., Varshney, A., ... & Shah, R. M. (2023). Unravelling the gut-lung axis: insights into microbiome interactions and Traditional Indian Medicine's perspective on optimal health. *FEMS Microbiology Ecology*, 99(10), fiad103.

Hanahan, D. & Weinberg, R.A. (2000). The Hallmarks of Cancer. *Cell*, 100(1), 57-70. https://doi.org/10.1016/S0092-8674(00)81683-9

Hanahan, D. & Weinberg, R.A. (2011). Hallmarks of Cancer: The Next Generation. *Cell*, 144(5), 646-674. https://doi.org/10.1016/j.cell.2011.02.013

Harbeck, N., Sotlar, K., Wuerstlein, R., & Doisneau-Sixou, S. (2014). Molecular and protein markers for clinical decision making in breast cancer: today and tomorrow. *Cancer treatment reviews*, 40(3), 434-444.

Hardin, J. & Bertoni, G. (2018). Becker’s World of the Cell. Pearson.

Harjunpää, H., Llort Asens, M., Guenther, C., & Fagerholm, S. C. (2019). Cell adhesion molecules and their roles and regulation in the immune and tumor microenvironment. *Frontiers in immunology*, 10, 448153.

Hartwell, L. H., & Kastan, M. B. (1994). Cell cycle control and cancer. *Science*, 266(5192), 1821-1828.

Hato, L., Vizcay, A., Eguren, I., Pérez-Gracia, J. L., Rodríguez, J., Gállego Pérez-Larraya, J., ... & Santisteban, M. (2024). Dendritic Cells in Cancer Immunology and Immunotherapy. *Cancers*, 16(5), 981.

Hecht, S. S. (2003). Tobacco carcinogens, their biomarkers and tobacco-induced cancer. *Nature Reviews Cancer*, 3(10), 733-744.

Herbst, R. S., Heymach, J. V., & Lippman, S. M. (2008). Lung cancer. *New England Journal of Medicine*, 359(13), 1367-1380.

Heryanto, Y. D., & Imoto, S. (2023). Identifying key regulators of keratinization in lung squamous cell cancer using integrated TCGA analysis. *Cancers*, 15(7), 2066.

Hodzic, E. (2016). Single-cell analysis: Advances and future perspectives. *Bosnian journal of basic medical sciences*, 16(4), 313-314. https://doi.org/10.17305/bjbms.2016.1371

Hoffman, J. I. (2015). Hypergeometric distribution. Biostatistics for medical and biomedical practitioners. Academic Press, Cambridge, MA, 179-182. https://doi.org/10.1016/B978-0-12-802387-7.00013-5

Horio, Y., Kuroda, H., Masago, K., Matsushita, H., Sasaki, E., & Fujiwara, Y. (2024). Current diagnosis and treatment of salivary gland-type tumors of the lung. *Japanese Journal of Clinical Oncology*, 54(3), 229-247.

Hu, S., Meng, K., Wang, T., Qu, R., Wang, B., Xi, Y., ... & Li, L. (2024). Lung cancer cell-intrinsic IL-15 promotes cell migration and sensitizes murine lung tumors to anti-PD-L1 therapy. *Biomarker Research*, 12(1), 40.

Human Protein Atlas. (2023). Human Protein Atlas version 23.0: Data in Tab-Separated Format [Data file]. Retrieved from https://www.proteinatlas.org/download/proteinatlas.tsv.zip

Jassal, B., Matthews, L., Viteri, G., Gong, C., Lorente, P., Fabregat, A., ... & D'Eustachio, P. (2020). The Reactome pathway knowledgebase. *Nucleic Acids Research*, 48(D1), D498-D503. https://doi.org/10.1093/nar/gkz1031

Jones, P. A., & Baylin, S. B. (2007). The epigenomics of cancer. *Cell*, 128(4), 683-692.

Joshi, R. S., Kanugula, S. S., Sudhir, S., Pereira, M. P., Jain, S., & Aghi, M. K. (2021). The role of cancer-associated fibroblasts in tumor progression. *Cancers*, 13(6), 1399.

Karlsson, M., Zhang, C., Méar, L., Zhong, W., Digre, A., Katona, B., ... & Lindskog, C. (2021). A single–cell type transcriptomics map of human tissues. *Science advances*, 7(31), eabh2169. https://doi.org/10.1126/sciadv.abh2169

Lopez-Rodriguez, E., Gay-Jordi, G., Mucci, A., Lachmann, N., & Serrano-Mollar, A. (2017). Lung surfactant metabolism: early in life, early in disease and target in cell therapy. *Cell and Tissue Research*, 367, 721-735.

Kastan, M. B., & Bartek, J. (2004). Cell-cycle checkpoints and cancer. *Nature*, 432(7015), 316-323.

Klopfenstein, D. V., Zhang, L., Pedersen, B. S., Ramírez, F., Warwick Vesztrocy, A., Naldi, A., ... & Tang, H. (2018). GOATOOLS: A Python library for Gene Ontology analyses. *Scientific reports*, 8(1), 1-17.

Kolberg, L., Raudvere, U., Kuzmin, I., Adler, P., Vilo, J., & Peterson, H. (2023). g: Profiler—interoperable web service for functional enrichment analysis and gene identifier mapping (2023 update). *Nucleic acids research*, 51(W1), W207-W212.

Kolhe, R., Hunter, M., Liu, S., Jadeja, R. N., Pundkar, C., Mondal, A. K., ... & Fulzele, S. (2017). Gender-specific differential expression of exosomal miRNA in synovial fluid of patients with osteoarthritis. *Scientific reports*, 7(1), 2029.

Krueger, J. G., & Bowcock, A. (2005). Psoriasis pathophysiology: current concepts of pathogenesis. *Annals of the rheumatic diseases*, 64(suppl 2), ii30-ii36.

Lardone, M. C., Parodi, D. A., Valdevenito, R., Ebensperger, M., Piottante, A., Madariaga, M., ... & Castro, A. (2007). Quantification of DDX3Y, RBMY1, DAZ and TSPY mRNAs in testes of patients with severe impairment of spermatogenesis. *Molecular human reproduction*, 13(10), 705-712.

Law, C. W., Chen, Y., Shi, W., & Smyth, G. K. (2014). voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology*, 15(2), R29.

Leone, P., Malerba, E., Susca, N., Favoino, E., Perosa, F., Brunori, G., ... & Racanelli, V. (2024). Endothelial cells in tumor microenvironment: insights and perspectives. *Frontiers in Immunology*, 15, 1367875.

Levine, A. J. (1997). p53, the cellular gatekeeper for growth and division. *Cell*, 88(3), 323-331.

Libbrecht, M. W., & Noble, W. S. (2015). Machine learning applications in genetics and genomics. *Nature Reviews Genetics*, 16(6), 321-332.

Liu, L. F., Shen, W. J., Ueno, M., Patel, S., & Kraemer, F. B. (2011). Characterization of age-related gene expression profiling in bone marrow and epididymal adipocytes. *BMC genomics*, 12, 1-18.

Long, T., Liu, Z., Zhou, X., Yu, S., Tian, H., & Bao, Y. (2019). Identification of differentially expressed genes and enriched pathways in lung cancer using bioinformatics analysis. *Molecular medicine reports*, 19(3), 2029-2040.

Lowes, M. A., Suárez-Fariñas, M., & Krueger, J. G. (2014). Immunology of psoriasis. *Annual Review of Immunology*, 32, 227-255. https://doi.org/10.1146/annurev-immunol-032713-120225

Mao, Y., Huang, P., Wang, Y., Wang, M., Li, M. D., & Yang, Z. (2021). Genome-wide methylation and expression analyses reveal the epigenetic landscape of immune-related diseases for tobacco smoking. *Clinical Epigenetics*, 13, 1-14.

Marzell, T. (2019). Machine Learning Performance Indicators. RocketLoop. https://rocketloop.de/en/blog/machine-learning-performance-indicators/

Min, S., Lee, B., & Yoon, S. (2017). Deep learning in bioinformatics. *Briefings in Bioinformatics*, 18(5), 851-869.

Minna, J. D., Roth, J. A., Gazdar, A. F. (2002). Focus on lung cancer. *Cancer Cell*. https://doi.org/10.1016/S1535-6108(02)00027-2

Morrisey, E. E., & Rustgi, A. K. (2018). The lung and esophagus: developmental and regenerative overlap. *Trends in cell biology*, 28(9), 738-748.

Mrozik, K. M., Blaschuk, O. W., Cheong, C. M., Zannettino, A. C. W., & Vandyke, K. (2018). N-cadherin in cancer metastasis, its emerging role in haematological malignancies and potential as a therapeutic target in cancer. *BMC cancer*, 18(1), 939.

Musgrove, E. A., & Sutherland, R. L. (2009). Biological determinants of endocrine resistance in breast cancer. *Nature Reviews Cancer*, 9(9), 631-643.

Myers, G. J., Sandler, C., & Badgett, T. (2011). The Art of Software Testing. Wiley.

Nakayama, J., & Yamamoto, Y. (2023). Cancer-prone phenotypes and gene expression heterogeneity at single-cell resolution in cigarette-smoking lungs. *Cancer Research Communications*, 3(11), 2280-2291.

National Cancer Institute (NCI). (2021). What Is Cancer? https://www.cancer.gov/about-cancer/understanding/what-is-cancer

National Institute of Standards and Technology (NIST). (n.d.). Kolmogorov-Smirnov Goodness-of-Fit Test. https://www.itl.nist.gov/div898/handbook/eda/section3/eda35g.htm

Nature. (n.d.) Microarray. https://www.nature.com/scitable/definition/microarray-202/

Nelson, D. L., & Cox, M. M. (2021). Lehninger principles of biochemistry. Macmillan Learning.

Nestle, F. O., Kaplan, D. H., & Barker, J. (2009). Psoriasis. *New England Journal of Medicine*, 361(5), 496-509. https://doi.org/10.1056/NEJMra0804595

Netti, G. S., Franzin, R., Stasi, A., Spadaccino, F., Dello Strologo, A., Infante, B., ... & Stallone, G. (2021). Role of complement in regulating inflammation processes in renal and prostate cancers. *Cells*, 10(9), 2426.

Newman, A. M., Steen, C. B., Liu, C. L., Gentles, A. J., Chaudhuri, A. A., Scherer, F., ... & Alizadeh, A. A. (2019). Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nature biotechnology*, 37(7), 773-782. https://doi.org/10.1038/s41587-019-0114-2

Orywal, K., Jelski, W., Kozłowski, M. D., & Mroczko, B. (2020). Activity of alcohol dehydrogenase and aldehyde dehydrogenase in lung cancer cells. *Anticancer Research*, 40(7), 3857-3863.

Page, D. C., Mosher, R., Simpson, E. M., Fisher, E. M., Mardon, G., Pollack, J., ... & Brown, L. G. (1987). The sex-determining region of the human Y chromosome encodes a finger protein. *Cell*, 51(6), 1091-1104.

Pang, B., Wu, N., Guan, R., Pang, L., Li, X., Li, S., ... & Jin, Y. (2017). Overexpression of RCC2 enhances cell motility and promotes tumor metastasis in lung adenocarcinoma by inducing epithelial–mesenchymal transition. *Clinical Cancer Research*, 23(18), 5598-5610.

Pasquini, G., Rojo Arias, J. E., Schäfer, P., & Busskamp, V. (2020). Automated methods for cell type annotation on scRNA-seq data. *Computational and Structural Biotechnology Journal*, 19, 961-969. https://doi.org/10.1016/j.csbj.2021.01.015

Pamarthy, S., Kulshrestha, A., Katara, G. K., & Beaman, K. D. (2018). The curious case of vacuolar ATPase: regulation of signaling pathways. *Molecular Cancer*, 17, 1-9.

Neophytou, C. M., Panagi, M., Stylianopoulos, T., & Papageorgis, P. (2021). The role of tumor microenvironment in cancer metastasis: Molecular mechanisms and therapeutic opportunities. *Cancers*, 13(9), 2053.

Patton, M. Q. (2014). Qualitative Research & Evaluation Methods: Integrating Theory and Practice. Sage publications.

Paw, M., Wnuk, D., Jakieła, B., Bochenek, G., Sładek, K., Madeja, Z., & Michalik, M. (2021). Responsiveness of human bronchial fibroblasts and epithelial cells from asthmatic and non-asthmatic donors to the transforming growth factor-β 1 in epithelial-mesenchymal trophic unit model. *BMC Molecular and Cell Biology*, 22, 1-14.

Peto, R., Darby, S., Deo, H., Silcocks, P., Whitley, E., & Doll, R. (2000). Smoking, smoking cessation, and lung cancer in the UK since 1950: combination of national statistics with two case-control studies. Bmj, 321(7257), 323-329.

Poirier, M. C. (2004). Chemical-induced DNA damage and human cancer risk. *Nature Reviews Cancer*, 4(8), 630-637.

Pomyen, Y., Segura, M., Ebbels, T. M., & Keun, H. C. (2015). Over-representation of correlation analysis (ORCA): a method for identifying associations between variable sets. *Bioinformatics*, 31(1), 102-108. https://doi.org/10.1093/bioinformatics/btu589

Pontén, F., Jirström, K., & Uhlén, M. (2008). The Human Protein Atlas—a tool for pathology. *Pathological Society*, pp. 387-393.

Prado, F., & Maya, D. (2017). Regulation of replication fork advance and stability by nucleosome assembly. *Genes*, 8(2), 49.

Pratt, M. M., John, K., MacLean, A. B., Afework, S., Phillips, D. H., & Poirier, M. C. (2011). Polycyclic aromatic hydrocarbon (PAH) exposure and DNA adduct semi-quantitation in archived human tissues. *International journal of environmental research and public health*, 8(7), 2675-2691.

Pylayeva-Gupta, Y., Grabocka, E., & Bar-Sagi, D. (2011). RAS oncogenes: weaving a tumorigenic web. *Nature Reviews Cancer*, 11(11), 761-774.

Qiu, Y., Wang, J. & R., K. (2021). Identification of cell-type-specific marker genes from co-expression patterns in tissue samples. *Bioinformatics*, 37(19), 3228-3234. https://doi.org/10.1093/bioinformatics/btab257

Ramamonjisoa, N., & Ackerstaff, E. (2017). Characterization of the tumor microenvironment and tumor–stroma interaction by non-invasive preclinical imaging. *Frontiers in Oncology*, 7, 3.

Rehman, M., Saeed, M. S., Fan, X., Salam, A., Munir, R., Yasin, M. U., ... & Gan, Y. (2023). The multifaceted role of jasmonic acid in plant stress mitigation: An overview. *Plants*, 12(23), 3982.

Reuter, S., Gupta, S. C., Chaturvedi, M. M., & Aggarwal, B. B. (2010). Oxidative stress, inflammation, and cancer: how are they linked?. *Free radical biology and medicine*, 49(11), 1603-1616.

Risso, D. S., Kozlitina, J., Sainz, E., Gutierrez, J., Wooding, S., Getachew, B., ... & Drayna, D. (2016). Genetic variation in the TAS2R38 bitter taste receptor and smoking behaviors. *PLoS One*, 11(10), e0164157.

Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., & Smyth, G.K. (2015). limma powers differential expression analyses for RNA Sequencing and microarray studies. *Nucleic Acids Research*, 43(7), e47.

Schiller, J. H., Harrington, D., Belani, C. P., Langer, C., Sandler, A., Krook, J., ... & Johnson, D. H. (2002). Comparison of four chemotherapy regimens for advanced non-small-cell lung cancer. *New England Journal of Medicine*, 346(2), 92-98.

Schuurman, N., Leszczynski, A. (2008). Ontologies for Bioinformatics. *Bioinformatics and Biology Insights*. https://doi.org/10.4137/BBI.S451

Sharma, P., Alsharif, S., Fallatah, A., & Chung, B. M. (2019). Intermediate filaments as effectors of cancer development and metastasis: a focus on keratins, vimentin, and nestin. *Cells*, 8(5), 497.

Sherr, C. J., & Roberts, J. M. (1999). CDK inhibitors: positive and negative regulators of G1-phase progression. *Genes & development*, 13(12), 1501-1512.

Shevchenko, J. A., Nazarov, K. V., Alshevskaya, A. A., & Sennikov, S. V. (2023). Erythroid Cells as Full Participants in the Tumor Microenvironment. *International Journal of Molecular Sciences*, 24(20), 15141.

Shin, H., Sheu, B., Joseph, M., & Markey, M. K. (2008). Guilt-by-association feature selection: Identifying biomarkers from proteomic profiles. *Biomedical Informatics*, 41(2), 124-136. https://doi.org/10.1016/j.jbi.2007.04.003

Skogholt, A.H. (2021). Identifying robust blood-based messenger RNA (mRNA) markers for potential detection of lung cancer (Unpublished doctoral dissertation). Norwegian University of Science and Technology.

Smith, D. D., Sætrom, P., Snøve, O., Lundberg, C., Rivas, G. E., Glackin, C., & Larson, G. P. (2008). Meta-analysis of breast cancer microarray studies in conjunction with conserved cis-elements suggest patterns for coordinate regulation. *BMC Bioinformatics*, 9, 1-15. https://doi.org/10.1186/1471-2105-9-63

Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 3(1).

Sokolowski, D. J., Faykoo-Martinez, M., Erdman, L., Hou, H., Chan, C., Zhu, H., ... & Wilson, M. D. (2021). Single-cell mapper (scMappR): using scRNA-seq to infer the cell-type specificities of differentially expressed genes. *NAR genomics and bioinformatics*, 3(1), lqab011.

Solvin, Å. Ø., Chawla, K., Jenssen, M., Olsen, Lene C., Furberg, AS., … & Løset, M. (2023). Meta-analysis of RNA-seq data from 534 skin samples shows substantial IL-17 effects in non-lesional psoriatic skin. *medRxiv*. Preprint. https://www.medrxiv.org/content/10.1101/2023.11.03.23298021v1

Song, Q., Chen, P., & Liu, X. M. (2021). The role of cigarette smoke-induced pulmonary vascular endothelial cell apoptosis in COPD. *Respiratory Research*, 22(1), 39.

Spiro, S. G. & Silvestri, G. A. (2005). One Hundred Years of Lung Cancer. *American Journal of Respiratory and Critical Care Medicine*. 172(5).

Stark, R., Grzelak, M., & Hadfield, J. (2019). RNA Sequencing: the teenage years. *Nature Reviews Genetics*, 20(11), 631-656.

Stevens, R., Goble, C & Bechhofer, S. (2000). Ontology-based knowledge representation for bioinformatics. *Briefings in Bioinformatics*, 1(4), 398-414. https://academic.oup.com/bib/article/1/4/398/2530008

Stipp, M. C., & Acco, A. (2021). Involvement of cytochrome P450 enzymes in inflammation and cancer: a review. *Cancer chemotherapy and pharmacology*, 87(3), 295-309.

Subramanian, A., et al. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43), 15545-15550.

Szumilas, M. (2010). Explaining odds ratios. *Journal of the Canadian Academy of Child and Adolescent Psychiatry* = Journal de l'Academie canadienne de psychiatrie de l'enfant et de l'adolescent, 19(3), 227–229. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2938757/

Takeshita, J., Grewal, S., Langan, S. M., Mehta, N. N., Ogdie, A., Van Voorhees, A. S., & Gelfand, J. M. (2017). Psoriasis and comorbid diseases: epidemiology. *Journal of the American Academy of Dermatology*, 76(3), 377-390.

Tata, P. R., & Rajagopal, J. (2017). Plasticity in the lung: making and breaking cell identity. *Development*, 144(5), 755-766.

The Cancer Genome Atlas Program. (n.d.) The Cancer Genome Atlas Program (TCGA). https://www.cancer.gov/ccg/research/genome-sequencing/tcga

The Cancer Genome Atlas Program. (2024). Genomic Data Commons Data Portal. https://www.portal.gdc.cancer.gov/

Thul, P. J., & Lindskog, C. (2018). The Human Protein Atlas: A spatial map of the human proteome. *Protein Science*, 27(1), 233-244.

Uhlén, M., Fagerberg, L., Hallström, B. M., et al. (2015). Tissue-based map of the human proteome. *Science*, 347(6220), 1260419.

U.S. Public Health Service. (1964). Surgeon General's advisory committee on smoking and health. Washington, DC: U.S. Government Printing Office. Publication No. 1103.

Valavanidis, A., Vlachogianni, T., Fiotakis, K., & Loridas, S. (2013). Pulmonary oxidative stress, inflammation and cancer: respirable particulate matter, fibrous dusts and ozone as major causes of lung carcinogenesis through reactive oxygen species mechanisms. *International Journal of Environmental Research and Public Health*, 10(9), 3886-3907.

Van Rossum, G., & Drake, F. L. (2009). Python 3 Reference Manual. *CreateSpace*.

Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz Jr, L. A., & Kinzler, K. W. (2013). Cancer Genome Landscapes. *Science*, 339(6127), 1546-1558.

Wang, Z., et al. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1), 57-63.

Watkins, J.C. (n.d.). Hypothesis Testing - Multiple Testing. The University of Arizona. https://www.math.arizona.edu/~jwatkins/H6_multiple.pdf

Weinberg, R. A. (2013). The biology of cancer 2nd edition. *Garland Science*. Cambridge, MA, 658-691.

Whitsett, J. A., Wert, S. E., & Weaver, T. E. (2015). Diseases of pulmonary surfactant homeostasis. *Annual Review of Pathology: Mechanisms of Disease*, 10, 371-393.

Wilson, A., Shehadeh, L. A., Yu, H., & Webster, K. A. (2010). Age-related molecular genetic changes of murine bone marrow mesenchymal stem cells. *BMC genomics*, 11, 1-14.

Wikimedia Commons. (2016). Figure 10 02 01.jpg. https://www.commons.wikimedia.org/wiki/File:Figure_10_02_01.jpg

Wolfe, C. J., Kohane, I. S. & Butte, A. K. (2005). Systematic survey reveals general applicability of guilt-by-association within gene coexpression networks. *BMC Bioinformatics*, 6(227). https://www.doi.org/10.1186/1471-2105-6-227

World Cancer Research Fund International (WCRF). (2022). Worldwide cancer data. https://www.wcrf.org/cancer-trends/worldwide-cancer-data/

Xu, J., Wei, Q., & He, Z. (2020). Insight into the function of RIPK4 in keratinocyte differentiation and carcinogenesis. *Frontiers in oncology*, 10, 1562.

Yang, T., Xiao, H., Liu, X., Wang, Z., Zhang, Q., Wei, N., & Guo, X. (2021). Vascular normalization: a new window opened for cancer therapies. *Frontiers in Oncology*, 11, 719836.

Yates, A. D., Achuthan, P., Akanni, W., Allen, J., Allen, J., Alvarez-Jarreta, J., ... & Flicek, P. (2020). Ensembl 2020. *Nucleic acids research*, 48(D1), D682-D688.

Zani, I. A., Stephen, S. L., Mughal, N. A., Russell, D., Homer-Vanniasinkam, S., Wheatcroft, S. B., & Ponnambalam, S. (2015). Scavenger receptor structure and function in health and disease. *Cells* 4: 178–201.

Zhou, B., Stueve, T. R., Mihalakakos, E. A., Miao, L., Mullen, D., Wang, Y., ... & Marconett, C. N. (2021). Comprehensive epigenomic profiling of human alveolar epithelial differentiation identifies key epigenetic states and transcription factor co-regulatory networks for maintenance of distal lung identity. *BMC Genomics*, 22, 1-25.

Zhou, Y., Xu, B., Zhou, Y., Liu, J., Zheng, X., Liu, Y., ... & Jiang, J. (2021). Identification of key genes with differential correlations in lung adenocarcinoma. *Frontiers in Cell and Developmental Biology*, 9, 675438.

Zhang, E., Ding, C., Li, S., Zhou, X., Aikemu, B., Fan, X., ... & Yang, X. (2023). Roles and mechanisms of tumour-infiltrating B cells in human cancer: a new force in immunotherapy. *Biomarker Research*, 11(1), 28.

Zhang, L., Luo, W., Liu, J., Xu, M., Peng, Q., Zou, W., ... & Fu, Z. (2022). Modeling lung diseases using reversibly immortalized mouse pulmonary alveolar type 2 cells (imPAC2). *Cell & Bioscience*, 12(1), 159.

Zhang, R., Liu, Q., Li, T., Liao, Q., & Zhao, Y. (2019). Role of the complement system in the tumor microenvironment. *Cancer Cell International*, 19(1), 300.

Zhang, Y., Zhang, Q., Zhang, Y., & Han, J. (2023). The role of histone modification in DNA replication-coupled Nucleosome Assembly and Cancer. *International Journal of Molecular Sciences*, 24(5), 4939.

A. Appendices

This chapter provides a compilation of supplementary materials that enhance the insights discussed throughout this thesis. It includes detailed tables and additional data that complement the visual representations and analyses within the main text.

Appendix A.1 provides a table summarizing cell type ontologies from qualitative hECA, qualitative HPA, and numerical HPA data, focusing on shared cell types. It details adjusted p-values and odds ratios for various cell types under different conditions, such as age, sex-specific interactions, and tumor versus normal tissue comparisons.

Appendix A.2 presents the cell types of differentially expressed genes for the full dataset. It offers a detailed view of cellular changes associated with lung cancer.

Appendix A.3 delves into the impact of smoking on gene expression, presenting data from the smoking-related dataset. It highlights how smoking alters gene expression across different cell types, providing insights into the molecular adjustments that occur in lung cancer due to smoking.

Appendix A.4 lists the top 20 Reactome pathways derived from the full dataset, illustrating the significant biological processes and pathways that are perturbed in lung cancer. This section provides a deeper understanding of the pathway dynamics involved in the disease.

Appendix A.5 outlines the top 20 Reactome pathways from the smoking-related dataset, emphasizing the pathways that are predominantly influenced by smoking. This appendix helps to pinpoint specific biological processes that smoking impacts, aiding in the understanding of its role in lung cancer progression.

Appendix A.6 through A.11 provide detailed Gene Ontology (GO) analyses for different aspects of lung cancer:

Appendix A.6 details the biological processes affected in lung cancer as revealed by Gene Ontology (GO) analysis of the full dataset.

Appendix A.7 outlines the biological processes influenced by smoking, based on GO analysis of the smoking dataset.

Appendix A.8 identifies the top 20 cellular components affected in lung cancer, providing insights from GO analysis of the full dataset.

Appendix A.9 specifies the top 20 cellular components influenced by smoking, as detailed in the GO analysis of the smoking dataset.

Appendix A.10 displays the molecular functions affected in lung cancer, highlighting findings from GO analysis of the full dataset.

Appendix A.11 reports on the molecular functions influenced by smoking, as uncovered in the GO analysis of the smoking dataset.

Each appendix is designed to extend the data representation and analysis presented in the thesis, providing a richer and more comprehensive understanding of the genetic and molecular landscape of lung cancer.

A.1 Cell Type Ontologies (hECA, Qualitative HPA and Numerical HPA)

Comparison	Cell Type	Adj. p-value	Odds Ratio
Age hECA up	b-cells	6.71e-03	172.18
TumorVsNormal hECA up	b-cells	1.63e-02	10.08
TissueXSex hECA down	alveolar cells type 2	2.44e-04	1919.33
TissueXSex HPA down	alveolar cells type 2	4.02e-02	2519.88
TumorVsNormal hECA down	alveolar cells type 2	2.17e-06	74.88
TumorVsNormal HPA down	alveolar cells type 2	1.63e-02	inf
TumorVsNormal hECA down	endothelial cells	2.89e-15	22.61
TumorVsNormal HPA down	endothelial cells	4.73e-03	inf
TumorVsNormal hECA down	macrophages	1.49e-07	14.59
TumorVsNormal hECA down	smooth muscle cells	3.64e-02	5.1
TumorVsNormal hECA down	fibroblasts	4.25e-03	5.26
TumorVsNormal hECA up	fibroblasts	6.98e-04	11.79
TumorVsNormal hECA down	ciliated cells	4.25e-03	15.58
TumorVsNormal hECA down	alveolar cells type 1	2.17e-06	74.88
TumorVsNormal HPA up	basal keratinocytes	8.99e-04	inf
Num Age up	b-cells	1.30e-08	42.54
Num TissueXSex down	alveolar cells type 2	1.75e-04	inf
Num TissueXSex down	alveolar cells type 1	4.04e-02	119.08
Num TumorVsNormal up	b-cells	3.01e-50	6.97
Num TumorVsNormal up	basal keratinocytes	1.50e-30	9.02
Num TumorVsNormal up	fibroblasts	1.90e-03	2.17
Num TumorVsNormal down	endothelial cells	2.79e-120	17.14
Num TumorVsNormal down	alveolar cells type 1	1.75e-63	9.24
Num TumorVsNormal down	alveolar cells type 2	5.25e-57	10.89
Num TumorVsNormal down	macrophages	5.05e-56	6.24
Num TumorVsNormal down	smooth muscle cells	5.25e-55	7.51
Num TumorVsNormal down	fibroblasts	1.79e-44	6.09
Num TumorVsNormal down	ciliated cells	3.96e-23	3.49
Num TumorVsNormal down	b-cells	6.32e-04	1.55

A.2 Cell Types of Differentially Expressed Genes for the Full Dataset

Comparison	Cell Types	Adj. p-value	Odds Ratio
Age down	Plasma cells	1.51e-10	140.12
Age down	Erythroid cells	9.28e-05	38.31
Age up	Basal keratinocytes	2.42e-17	76.37
Age up	Suprabasal keratinocytes	2.34e-12	34.42
Age up	Basal squamous epithelial cells	4.67e-07	24.01
Age up	Squamous epithelial cells	1.17e-06	20.41
Age up	Basal respiratory cells	1.92e-06	23.95
Age up	naive B-cell	7.16e-04	11.34
Age up	Plasma cells	2.29e-03	8.85
Age up	memory B-cell	5.09e-03	9.31
Age up	Salivary duct cells	6.54e-03	20.34
Age up	B-cells	6.54e-03	6.72
Age up	Fibroblasts	1.73e-02	8.6
TissueXSex down	Alveolar cells type 2	2.30e-19	185.95
TissueXSex down	Alveolar cells type 1	1.71e-08	44.07
TissueXSex up	Basal keratinocytes	7.33e-26	161.9
TissueXSex up	Squamous epithelial cells	2.03e-23	111.87
TissueXSex up	Suprabasal keratinocytes	1.72e-21	84.21
TissueXSex up	Basal squamous epithelial cells	2.63e-15	57.88
TissueXSex up	Basal respiratory cells	7.41e-13	50.9
TissueXSex up	Club cells	1.38e-06	35.03
TumorVsNormal down	Adipocytes	1.31e-111	11.04
TumorVsNormal down	Endothelial cells	4.10e-106	14.3
TumorVsNormal down	monocytes	9.47e-90	4.59
TumorVsNormal down	Macrophages	2.07e-80	7.95
TumorVsNormal down	Kupffer cells	3.32e-63	6.94
TumorVsNormal down	Alveolar cells type 2	2.40e-62	11.49
TumorVsNormal down	Alveolar cells type 1	8.20e-61	8.55
TumorVsNormal down	granulocytes	1.06e-60	3.25
TumorVsNormal down	Lymphatic endothelial cells	2.47e-56	7.92
TumorVsNormal down	Microglial cells	1.23e-46	4.09
TumorVsNormal down	Smooth muscle cells	1.38e-42	5.96
TumorVsNormal down	Hofbauer cells	2.75e-41	4.76
TumorVsNormal down	dendritic cells	7.02e-41	3.27
TumorVsNormal down	Fibroblasts	3.30e-36	5.1
TumorVsNormal down	intermediate monocyte	9.92e-32	4.84
TumorVsNormal down	neutrophil	4.08e-30	2.92
TumorVsNormal down	non-classical monocyte	9.50e-30	4.22
TumorVsNormal down	Langerhans cells	1.99e-29	3.61
TumorVsNormal down	classical monocyte	4.34e-29	4.61
TumorVsNormal down	NK-cells	1.57e-28	3.21

TumorVsNormal down	Leydig cells	4.51e-27	4.73
TumorVsNormal down	myeloid DC	4.33e-26	4.32
TumorVsNormal down	T-cells	7.56e-21	2.38
TumorVsNormal down	Peritubular cells	8.71e-18	3.42
TumorVsNormal down	Schwann cells	4.99e-17	3.45
TumorVsNormal down	eosinophil	1.64e-15	2.72
TumorVsNormal down	Mesothelial cells	3.15e-14	3.74
TumorVsNormal down	Oligodendrocyte precursor cells	3.88e-14	1.9
TumorVsNormal down	Astrocytes	1.78e-13	2
TumorVsNormal down	basophil	2.26e-13	2.17
TumorVsNormal down	Ciliated cells	9.69e-13	2.56
TumorVsNormal down	Oligodendrocytes	1.03e-10	1.75
TumorVsNormal down	Cardiomyocytes	1.86e-09	2.06
TumorVsNormal down	plasmacytoid DC	2.09e-09	2.22
TumorVsNormal down	Glandular and luminal cells	3.40e-07	2.34
TumorVsNormal down	Excitatory neurons	9.09e-07	1.48
TumorVsNormal down	Sertoli cells	1.48e-06	2.36
TumorVsNormal down	Skeletal myocytes	2.27e-06	1.98
TumorVsNormal down	Basal prostatic cells	5.33e-06	2.62
TumorVsNormal down	Inhibitory neurons	8.91e-06	1.44
TumorVsNormal down	Ovarian stromal cells	1.45e-05	2.36
TumorVsNormal down	Endometrial stromal cells	2.29e-05	2.23
TumorVsNormal down	B-cells	3.24e-05	1.65
TumorVsNormal down	MAIT T-cell	3.97e-05	2.26
TumorVsNormal down	Muller glia cells	3.41e-04	1.77
TumorVsNormal down	Hepatocytes	6.09e-04	1.53
TumorVsNormal down	NK-cell	1.01e-03	1.85
TumorVsNormal down	gdT-cell	4.76e-03	1.86
TumorVsNormal down	Ionocytes	5.03e-03	1.77
TumorVsNormal down	Secretory cells	8.86e-03	2.03
TumorVsNormal down	naive CD4 T-cell	9.05e-03	1.68
TumorVsNormal down	Mucus glandular cells	9.05e-03	2
TumorVsNormal down	Cholangiocytes	9.74e-03	1.82
TumorVsNormal down	Prostatic glandular cells	2.45e-02	1.61
TumorVsNormal down	naive CD8 T-cell	2.71e-02	1.59
TumorVsNormal up	Extravillous trophoblasts	7.14e-76	8.66
TumorVsNormal up	Plasma cells	1.10e-65	8.05
TumorVsNormal up	Suprabasal keratinocytes	5.52e-49	7.15
TumorVsNormal up	Erythroid cells	6.07e-47	7.95
TumorVsNormal up	Basal keratinocytes	1.03e-46	10.35
TumorVsNormal up	memory B-cell	1.84e-37	6.32
TumorVsNormal up	Squamous epithelial cells	4.13e-35	6.62
TumorVsNormal up	naive B-cell	2.95e-33	5.72
TumorVsNormal up	Undifferentiated cells	1.12e-32	9.36
TumorVsNormal up	Basal respiratory cells	2.34e-32	7.69

TumorVsNormal up	Cytotrophoblasts	3.38e-25	5.23
TumorVsNormal up	B-cells	1.20e-23	3.66
TumorVsNormal up	Spermatogonia	1.48e-19	4.04
TumorVsNormal up	Basal squamous epithelial cells	1.34e-16	4.47
TumorVsNormal up	Club cells	2.08e-16	6.57
TumorVsNormal up	Ionocytes	2.19e-16	5.13
TumorVsNormal up	Oocytes	7.75e-12	2.43
TumorVsNormal up	Ductal cells	1.57e-10	4.76
TumorVsNormal up	Distal enterocytes	2.85e-09	2.54
TumorVsNormal up	Serous glandular cells	5.26e-09	3.81
TumorVsNormal up	T-reg	5.94e-09	2.94
TumorVsNormal up	Salivary duct cells	3.47e-08	5
TumorVsNormal up	Syncytiotrophoblasts	4.13e-07	2.17
TumorVsNormal up	Spermatocytes	3.37e-06	1.79
TumorVsNormal up	Pancreatic endocrine cells	1.06e-05	3.18
TumorVsNormal up	Intestinal goblet cells	4.43e-05	2.46
TumorVsNormal up	plasmacytoid DC	6.13e-05	2.07
TumorVsNormal up	Glandular and luminal cells	3.78e-04	2.27
TumorVsNormal up	Proximal enterocytes	4.37e-04	1.74
TumorVsNormal up	Breast myoepithelial cells	1.59e-03	2.6
TumorVsNormal up	Paneth cells	1.59e-03	2.02
TumorVsNormal up	Exocrine glandular cells	1.84e-03	2.47
TumorVsNormal up	Fibroblasts	1.00e-02	1.82
TumorVsNormal up	Gastric mucus-secreting cells	1.09e-02	1.74
TumorVsNormal up	Mucus glandular cells	1.15e-02	2.37
TumorVsNormal up	Breast glandular cells	1.25e-02	2.32
TumorVsNormal up	T-cells	2.90e-02	1.39
TumorVsNormal up	Cholangiocytes	3.75e-02	1.92
TumorVsNormal up	Endometrial stromal cells	4.93e-02	1.75

A.3 Cell Types of Differentially Expressed Genes for the Smoking Dataset

Comparison	Cell Types	Adj. p-value	Odds Ratio
Age up	memory B-cell	2.99e-10	74.39
Age up	naive B-cell	2.99e-10	71.91
Age up	Plasma cells	1.68e-09	56.04
Age up	B-cells	1.30e-08	42.54
CurrentVsNever Tumor down	Distal enterocytes	6.34e-03	16.18
CurrentVsNever Tumor down	Proximal enterocytes	1.14e-02	12.18
CurrentVsNever Tumor down	Cholangiocytes	1.23e-02	26.27
CurrentVsNever Tumor down	Serous glandular cells	1.35e-02	23.01
CurrentVsNever Tumor down	myeloid DC	3.89e-02	14.58
FormerVsNever Tumor up	Basal respiratory cells	3.06e-03	101.34
FormerVsNever Tumor up	Exocrine glandular cells	4.71e-02	67.55
TissueXSex down	Alveolar cells type 2	1.75e-04	inf
TissueXSex down	Alveolar cells type 1	4.04e-02	119.08
TumorVsNormal down	Adipocytes	3.18e-141	14.65
TumorVsNormal down	Endothelial cells	2.79e-120	17.14
TumorVsNormal down	monocytes	3.49e-75	4.29
TumorVsNormal down	Lymphatic endothelial cells	2.56e-66	9.50
TumorVsNormal down	Alveolar cells type 1	1.75e-63	9.24
TumorVsNormal down	granulocytes	6.43e-61	3.38
TumorVsNormal down	Alveolar cells type 2	5.25e-57	10.89
TumorVsNormal down	Macrophages	5.05e-56	6.24
TumorVsNormal down	Smooth muscle cells	5.25e-55	7.51
TumorVsNormal down	Kupffer cells	4.70e-46	5.71
TumorVsNormal down	Fibroblasts	1.79e-44	6.09
TumorVsNormal down	Leydig cells	4.83e-38	6.19
TumorVsNormal down	dendritic cells	7.76e-38	3.24
TumorVsNormal down	Hofbauer cells	5.48e-33	4.30
TumorVsNormal down	intermediate monocyte	7.89e-31	4.92
TumorVsNormal down	Microglial cells	9.74e-31	3.37
TumorVsNormal down	classical monocyte	4.09e-30	4.89
TumorVsNormal down	non-classical monocyte	5.50e-30	4.39
TumorVsNormal down	NK-cells	4.95e-28	3.29
TumorVsNormal down	neutrophil	1.62e-27	2.89
TumorVsNormal down	myeloid DC	2.05e-24	4.30
TumorVsNormal down	Peritubular cells	9.37e-24	4.15

TumorVsNormal down	Langerhans cells	2.31e-23	3.31
TumorVsNormal down	Ciliated cells	3.96e-23	3.49
TumorVsNormal down	eosinophil	1.13e-19	3.12
TumorVsNormal down	T-cells	1.25e-18	2.34
TumorVsNormal down	Schwann cells	2.08e-15	3.37
TumorVsNormal down	Mesothelial cells	2.62e-15	4.03
TumorVsNormal down	basophil	2.25e-11	2.10
TumorVsNormal down	Cardiomyocytes	2.77e-11	2.25
TumorVsNormal down	Sertoli cells	7.51e-11	3.07
TumorVsNormal down	Astrocytes	1.67e-10	1.89
TumorVsNormal down	Oligodendrocyte precursor cells	4.25e-09	1.71
TumorVsNormal down	Ovarian stromal cells	2.09e-08	2.92
TumorVsNormal down	Skeletal myocytes	3.93e-08	2.22
TumorVsNormal down	Muller glia cells	3.93e-08	2.32
TumorVsNormal down	Basal prostatic cells	4.87e-08	3.08
TumorVsNormal down	plasmacytoid DC	1.11e-07	2.11
TumorVsNormal down	NK-cell	9.21e-07	2.37
TumorVsNormal down	gdT-cell	7.41e-06	2.54
TumorVsNormal down	Glandular and luminal cells	2.33e-05	2.14
TumorVsNormal down	Endometrial stromal cells	3.18e-05	2.26
TumorVsNormal down	Oligodendrocytes	6.42e-05	1.46
TumorVsNormal down	Hepatocytes	4.52e-04	1.56
TumorVsNormal down	Breast myoepithelial cells	4.79e-04	2.35
TumorVsNormal down	B-cells	6.32e-04	1.55
TumorVsNormal down	Inhibitory neurons	7.13e-04	1.34
TumorVsNormal down	MAIT T-cell	7.60e-04	2.04
TumorVsNormal down	Secretory cells	2.25e-03	2.23
TumorVsNormal down	Excitatory neurons	2.74e-03	1.30
TumorVsNormal down	Cholangiocytes	2.93e-03	2.00
TumorVsNormal down	Melanocytes	8.62e-03	1.76
TumorVsNormal down	Ionocytes	9.43e-03	1.76
TumorVsNormal down	Granulosa cells	2.21e-02	1.67
TumorVsNormal down	Serous glandular cells	4.88e-02	1.58
TumorVsNormal up	Plasma cells	2.35e-113	16.28
TumorVsNormal up	memory B-cell	2.86e-65	11.86
TumorVsNormal up	Erythroid cells	3.03e-64	12.61
TumorVsNormal up	naive B-cell	6.22e-63	11.21
TumorVsNormal up	B-cells	3.01e-50	6.97
TumorVsNormal up	Extravillous trophoblasts	5.92e-50	7.74
TumorVsNormal up	Undifferentiated cells	1.41e-40	13.70
TumorVsNormal up	Basal keratinocytes	1.50e-30	9.02
TumorVsNormal up	Spermatogonia	1.01e-29	6.21
TumorVsNormal up	Suprabasal keratinocytes	5.25e-29	6.02
TumorVsNormal up	Squamous epithelial cells	2.25e-23	6.06
TumorVsNormal up	Basal respiratory cells	1.26e-20	6.75

TumorVsNormal up	Gastric mucus-secreting cells	4.70e-18	4.94
TumorVsNormal up	Cytotrophoblasts	1.24e-15	4.66
TumorVsNormal up	Club cells	2.44e-13	6.85
TumorVsNormal up	Ionocytes	6.66e-13	5.26
TumorVsNormal up	Distal enterocytes	2.48e-12	3.32
TumorVsNormal up	T-reg	2.48e-12	4.03
TumorVsNormal up	Spermatocytes	7.70e-12	2.50
TumorVsNormal up	Oocytes	2.19e-11	2.71
TumorVsNormal up	Basal squamous epithelial cells	3.65e-10	3.96
TumorVsNormal up	Serous glandular cells	1.75e-09	4.61
TumorVsNormal up	Ductal cells	2.20e-08	4.87
TumorVsNormal up	Intestinal goblet cells	5.55e-07	3.24
TumorVsNormal up	Salivary duct cells	2.62e-06	4.98
TumorVsNormal up	Pancreatic endocrine cells	2.31e-05	3.56
TumorVsNormal up	Cholangiocytes	3.90e-05	3.41
TumorVsNormal up	Exocrine glandular cells	6.20e-05	3.40
TumorVsNormal up	Paneth cells	2.69e-04	2.44
TumorVsNormal up	Glandular and luminal cells	2.92e-04	2.59
TumorVsNormal up	Proximal enterocytes	3.36e-04	1.91
TumorVsNormal up	Breast glandular cells	4.43e-04	3.38
TumorVsNormal up	Breast myoepithelial cells	1.07e-03	3.05
TumorVsNormal up	plasmacytoid DC	1.72e-03	2.01
TumorVsNormal up	Fibroblasts	1.90e-03	2.17
TumorVsNormal up	T-cells	3.79e-03	1.62
TumorVsNormal up	Endometrial stromal cells	6.02e-03	2.27
TumorVsNormal up	Enteroendocrine cells	1.25e-02	2.15
TumorVsNormal up	Syncytiotrophoblasts	2.29e-02	1.59
TumorVsNormal up	Mucus glandular cells	2.41e-02	2.38
TumorVsNormal up	Collecting duct cells	4.31e-02	2.07

A.4 Reactome Pathways for Full Dataset (Top 20)

Comparison	Reactome:id	Pathway	Adj. p-value	Odds ratio
TissueXSex up	R-HSA-6805567	Keratinization	1.11e-16	47.95
TissueXSex up	R-HSA-5619043	Defective SLC2A1 causes GLUT1 deficiency syndrome 1 (GLUT1DS1)	3.85e-03	430.73
TissueXSex up	R-HSA-193775	Synthesis of bile acids and bile salts via 24-hydroxycholesterol	3.15e-03	21.78
TissueXSex up	R-HSA-193807	Synthesis of bile acids and bile salts via 27-hydroxycholesterol	1.99e-03	27.92
TissueXSex up	R-HSA-5357801	Programmed Cell Death	1.09e-03	8.15
TissueXSex up	R-HSA-9754189	Germ layer formation at gastrulation	1.00e-03	40.65
TissueXSex up	R-HSA-9823739	Formation of the anterior neural plate	6.33e-04	52.63
TissueXSex up	R-HSA-109581	Apoptosis	5.00e-04	10.13
TissueXSex up	R-HSA-9832991	Formation of the posterior neural plate	3.46e-04	74.59
TissueXSex up	R-HSA-452723	Transcriptional regulation of pluripotent stem cells	3.44e-03	20.76
TissueXSex up	R-HSA-75153	Apoptotic execution phase	1.58e-04	27.33
TissueXSex up	R-HSA-9725554	Differentiation of keratinocytes in interfollicular epidermis in mammalian skin	9.86e-05	32.44
TissueXSex up	R-HSA-9734767	Developmental Cell Lineages	9.86e-05	32.44
TissueXSex up	R-HSA-111465	Apoptotic cleavage of cellular proteins	5.62e-05	39.88
TissueXSex up	R-HSA-351906	Apoptotic cleavage of cell adhesion proteins	1.41e-06	174.91
TissueXSex up	R-HSA-1266738	Developmental Biology	1.49e-08	8.14
TissueXSex up	R-HSA-6809371	Formation of the cornified envelope	1.11e-16	82.36
TissueXSex up	R-HSA-446107	Type I hemidesmosome assembly	2.14e-04	99.48
TissueXSex down	R-HSA-913709	O-linked glycosylation of mucins	7.14e-03	15.68
TissueXSex down	R-HSA-3906995	Diseases associated with O-glycosylation of proteins	8.10e-03	14.64
TissueXSex down	R-HSA-427589	Type II Na+/Pi cotransporters	8.63e-03	133.35
TissueXSex down	R-HSA-427652	Sodium-coupled phosphate cotransporters	1.21e-02	88.88

TissueXSex down	R-HSA-5173105	O-linked glycosylation	2.22e-02	8.45
TissueXSex down	R-HSA-168179	Toll Like Receptor TLR1:TLR2 Cascade	2.32e-02	8.26
TissueXSex down	R-HSA-2142770	Synthesis of 15-eicosatetraenoic acid derivatives	3.75e-02	25.36
TissueXSex down	R-HSA-1500931	Cell-Cell communication	3.31e-02	6.77
TissueXSex down	R-HSA-166016	Toll Like Receptor 4 (TLR4) Cascade	3.34e-02	6.73
TissueXSex down	R-HSA-3781865	Diseases of glycosylation	4.77e-02	5.5
TissueXSex down	R-HSA-5621481	C-type lectin receptors (CLRs)	4.81e-02	5.48
TissueXSex down	R-HSA-168898	Toll-like Receptor Cascades	4.98e-02	5.37
TissueXSex down	R-HSA-168249	Innate Immune System	6.76e-03	3.44
TissueXSex down	R-HSA-181438	Toll Like Receptor 2 (TLR2) Cascade	2.32e-02	8.26
TissueXSex down	R-HSA-5621480	Dectin-2 family	5.71e-03	17.69
TissueXSex down	R-HSA-1566977	Fibronectin matrix formation	1.21e-02	88.88
TissueXSex down	R-HSA-5619045	Defective SLC34A2 causes pulmonary alveolar microlithiasis (PALM)	5.19e-03	266.74
TissueXSex down	R-HSA-5687583	Defective SLC34A2 causes PALM	5.19e-03	266.74
TissueXSex down	R-HSA-5683826	Surfactant metabolism	1.11e-16	221.67
TissueXSex down	R-HSA-5688890	Defective CSF2RA causes SMDP4	1.21e-09	622.44
Age up	R-HSA-6791312	TP53 Regulates Transcription of Cell Cycle Genes	1.11e-02	13.6
Age up	R-HSA-6799198	Complex I biogenesis	9.22e-03	15.04
Age up	R-HSA-109704	PI3K Cascade	8.92e-03	15.31
Age up	R-HSA-380108	Chemokine receptors bind chemokines	8.63e-03	15.59
Age up	R-HSA-2173782	Binding and Uptake of Ligands by Scavenger Receptors	7.99e-03	8.03
Age up	R-HSA-2033515	t(4;14) translocations of FGFR3	7.30e-03	207.35
Age up	R-HSA-5654741	Signaling by FGFR3	7.51e-03	16.82
Age up	R-HSA-8853334	Signaling by FGFR3 fusions in cancer	7.30e-03	207.35

Age up	R-HSA-1266738	Developmental Biology	6.80e-03	3.09
Age up	R-HSA-112399	IRS-mediated signalling	1.11e-02	13.6
Age up	R-HSA-2029482	Regulation of actin dynamics for phagocytic cup formation	6.76e-03	8.55
Age up	R-HSA-1500931	Cell-Cell communication	7.61e-03	8.18
Age up	R-HSA-2029480	Fcgamma receptor (FCGR) dependent phagocytosis	1.16e-02	6.95
Age up	R-HSA-6783783	Interleukin-10 signaling	1.88e-02	10.18
Age up	R-HSA-9664433	Leishmania parasite growth and survival	1.16e-02	6.95
Age up	R-HSA-2428928	IRS-related events triggered by IGF1R	1.24e-02	12.78
Age up	R-HSA-5690714	CD22 mediated BCR regulation	1.35e-02	12.23
Age up	R-HSA-2428924	IGF1R signaling cascade	1.35e-02	12.23
Age up	R-HSA-74751	Insulin receptor signalling cascade	1.35e-02	12.23
Age up	R-HSA-2404192	Signaling by Type 1 Insulin-like Growth Factor 1 Receptor (IGF1R)	1.38e-02	12.06
Age down	R-HSA-73886	Chromosome Maintenance	4.02e-04	8.66
Age down	R-HSA-4839726	Chromatin organization	4.72e-04	5.78
Age down	R-HSA-3247509	Chromatin modifying enzymes	4.72e-04	5.78
Age down	R-HSA-8939211	ESR-mediated signaling	4.89e-04	5.73
Age down	R-HSA-8878171	Transcriptional regulation by RUNX1	5.16e-04	5.66
Age down	R-HSA-9609646	HCMV Infection	6.22e-04	5.42
Age down	R-HSA-195258	RHO GTPase Effectors	1.42e-03	4.47
Age down	R-HSA-75153	Apoptotic execution phase	3.74e-04	16.38
Age down	R-HSA-195721	Signaling by WNT	1.54e-03	4.38
Age down	R-HSA-157118	Signaling by NOTCH	5.07e-04	5.69
Age down	R-HSA-201681	TCF dependent signaling in response to WNT	2.21e-04	6.88
Age down	R-HSA-69306	DNA Replication	7.02e-05	8.9
Age down	R-HSA-157579	Telomere Maintenance	1.73e-04	10.93
Age down	R-HSA-73884	Base Excision Repair	1.24e-04	11.98
Age down	R-HSA-212165	Epigenetic regulation of gene expression	9.20e-05	8.38

Age down	R-HSA-9610379	HCMV Late Events	8.50e-05	8.53
Age down	R-HSA-9018519	Estrogen-dependent gene expression	4.53e-05	9.82
Age down	R-HSA-1474165	Reproduction	3.40e-05	10.46
Age down	R-HSA-9816359	Maternal to zygotic transition (MZT)	2.68e-05	11.02
Age down	R-HSA-68875	Mitotic Prophase	2.42e-05	11.28
TumorVsNormal up	R-HSA-9725554	Differentiation of keratinocytes in interfollicular epidermis in mammalian skin	6.98e-05	4.13
TumorVsNormal up	R-HSA-69273	Cyclin A/B1/B2 associated events during G2/M transition	3.19e-04	4.57
TumorVsNormal up	R-HSA-176417	Phosphorylation of Emi1	2.09e-04	19.4
TumorVsNormal up	R-HSA-69306	DNA Replication	1.74e-04	2.06
TumorVsNormal up	R-HSA-1592389	Activation of Matrix Metalloproteinases	1.34e-04	4.67
TumorVsNormal up	R-HSA-6805567	Keratinization	1.32e-04	1.88
TumorVsNormal up	R-HSA-6811434	COPI-dependent Golgi-to-ER retrograde traffic	1.19e-04	2.54
TumorVsNormal up	R-HSA-69481	G2/M Checkpoints	1.08e-04	2.19
TumorVsNormal up	R-HSA-69275	G2/M Transition	9.92e-05	1.94
TumorVsNormal up	R-HSA-9734767	Developmental Cell Lineages	6.98e-05	4.13
TumorVsNormal up	R-HSA-1474228	Degradation of the extracellular matrix	6.59e-05	2.28
TumorVsNormal up	R-HSA-983705	Signaling by the B Cell Receptor (BCR)	3.83e-13	3.77
TumorVsNormal up	R-HSA-983189	Kinesins	6.06e-05	3.32
TumorVsNormal up	R-HSA-69190	DNA strand elongation	5.60e-05	4.77
TumorVsNormal up	R-HSA-453274	Mitotic G2-G2/M phases	5.05e-05	2
TumorVsNormal up	R-HSA-68962	Activation of the pre-replicative complex	3.47e-05	5.15
TumorVsNormal up	R-HSA-69239	Synthesis of DNA	3.21e-05	2.49
TumorVsNormal up	R-HSA-5653656	Vesicle-mediated transport	2.21e-05	1.39
TumorVsNormal up	R-HSA-68886	M Phase	2.16e-05	1.66
TumorVsNormal up	R-HSA-176974	Unwinding of DNA	1.79e-05	16.33
MaleVsFemale up	R-HSA-3214842	HDMs demethylate histones	3.55e-04	77.15

MaleVsFemale up	R-HSA-6809371	Formation of the cornified envelope	2.36e-04	27.36
MaleVsFemale up	R-HSA-6805567	Keratinization	9.87e-04	16.43

A.5 Reactome Pathways for Smoking Dataset (Top 20)

Comparison	Reactome:id	Pathway	Adj. p-value	Odds ratio
Age down	R-HSA-2168880	Scavenging of heme from plasma	6.81e-03	inf
Age down	R-HSA-2173782	Binding and Uptake of Ligands by Scavenger Receptors	1.08e-02	inf
Age up	R-HSA-5690714	CD22 mediated BCR regulation	8.77e-15	199.20
Age up	R-HSA-977606	Regulation of Complement cascade	1.63e-14	127.96
Age up	R-HSA-166658	Complement cascade	4.55e-14	112.99
Age up	R-HSA-2871837	FCERI mediated NF-kB activation	1.01e-11	75.63
Age up	R-HSA-2029481	FCGR activation	1.51e-13	133.82
Age up	R-HSA-2168880	Scavenging of heme from plasma	1.90e-13	129.69
Age up	R-HSA-2730905	Role of LAT2/NTAL/LAB on calcium mobilization	2.05e-13	128.37
Age up	R-HSA-166786	Creation of C4 and C2 activators	2.74e-13	123.34
Age up	R-HSA-166663	Initial triggering of complement	5.08e-13	113.34
Age up	R-HSA-2871796	FCERI mediated MAPK activation	6.59e-13	109.39
Age up	R-HSA-173623	Classical antibody-mediated complement activation	9.39e-14	142.92
Age up	R-HSA-2871809	FCERI mediated Ca+2 mobilization	9.02e-13	104.82
Age up	R-HSA-9664323	FCGR3A-mediated IL10 synthesis	1.82e-12	95.26
Age up	R-HSA-198933	Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell	2.92e-12	68.62
Age up	R-HSA-202733	Cell surface interactions at the vascular wall	3.86e-12	66.36
Age up	R-HSA-9664407	Parasite infection	4.27e-12	84.91
Age up	R-HSA-9664422	FCGR3A-mediated phagocytosis	4.27e-12	84.91
Age up	R-HSA-9664417	Leishmania phagocytosis	4.27e-12	84.91
Age up	R-HSA-2029482	Regulation of actin dynamics for phagocytic cup formation	4.49e-12	84.34
Age up	R-HSA-2173782	Binding and Uptake of Ligands by Scavenger Receptors	7.30e-12	78.99
MaleVsFemale up	R-HSA-3214842	HDMs demethylate histones	3.55e-04	70.13

TissueXSex down	R-HSA-5619045	Defective SLC34A2 causes pulmonary alveolar microlithiasis (PALM)	9.63e-04	1402.50
TissueXSex down	R-HSA-5683826	Surfactant metabolism	1.11e-04	148.97
TissueXSex down	R-HSA-5687583	Defective SLC34A2 causes PALM	9.63e-04	1402.50
TissueXSex down	R-HSA-427652	Sodium-coupled phosphate cotransporters	2.25e-03	467.33
TissueXSex down	R-HSA-5687613	Diseases associated with surfactant metabolism	5.13e-03	186.78
TissueXSex down	R-HSA-427589	Type II Na+/Pi cotransporters	1.60e-03	701.12
TissueXSex up	R-HSA-9832991	Formation of the posterior neural plate	8.08e-07	1869.17
TissueXSex up	R-HSA-9758941	Gastrulation	1.29e-04	126.31
TissueXSex up	R-HSA-449147	Signaling by Interleukins	1.79e-03	32.23
TissueXSex up	R-HSA-2892245	POU5F1 (OCT4), SOX2, NANOG repress genes related to differentiation	1.28e-03	inf
TissueXSex up	R-HSA-9834899	Specification of the neural plate border	3.08e-03	inf
TissueXSex up	R-HSA-452723	Transcriptional regulation of pluripotent stem cells	8.35e-06	520.19
TissueXSex up	R-HSA-9754189	Germ layer formation at gastrulation	2.38e-06	1018.64
TissueXSex up	R-HSA-9823739	Formation of the anterior neural plate	1.49e-06	1318.82
TissueXSex up	R-HSA-2892247	POU5F1 (OCT4), SOX2, NANOG activate genes related to proliferation	2.70e-03	inf
TissueXSex up	R-HSA-8986944	Transcriptional Regulation by MECP2	1.28e-02	inf
TissueXSex up	R-HSA-1280215	Cytokine Signaling in Immune system	4.98e-03	18.47
TissueXSex up	R-HSA-3769402	Deactivation of the beta-catenin transactivating complex	5.64e-03	inf
TissueXSex up	R-HSA-195721	Signaling by WNT	4.21e-02	inf
TissueXSex up	R-HSA-168256	Immune System	2.92e-02	6.44
TissueXSex up	R-HSA-201681	TCF dependent signaling in response to WNT	2.74e-02	inf
TissueXSex up	R-HSA-6785807	Interleukin-4 and Interleukin-13 signaling	1.84e-04	105.44
TissueXSex up	R-HSA-1266738	Developmental Biology	1.21e-02	11.10
TissueXSex up	R-HSA-9856649	Transcriptional and post-translational regulation of MITF-M expression and activity	5.77e-03	inf

TissueXSex up	R-HSA-9730414	MITF-M-regulated melanocyte development	2.13e-02	inf
TumorVsNormal up	R-HSA-2029481	FCGR activation	1.11e-16	24.14
TumorVsNormal up	R-HSA-9664417	Leishmania phagocytosis	1.11e-16	10.65
TumorVsNormal up	R-HSA-5690714	CD22 mediated BCR regulation	1.11e-16	33.96
TumorVsNormal up	R-HSA-69278	Cell Cycle, Mitotic	1.11e-16	3.64
TumorVsNormal up	R-HSA-1640170	Cell Cycle	1.11e-16	3.08
TumorVsNormal up	R-HSA-69620	Cell Cycle Checkpoints	1.11e-16	4.24
TumorVsNormal up	R-HSA-9664323	FCGR3A-mediated IL10 synthesis	1.11e-16	12.77
TumorVsNormal up	R-HSA-9664422	FCGR3A-mediated phagocytosis	1.11e-16	10.65
TumorVsNormal up	R-HSA-9664407	Parasite infection	1.11e-16	10.65
TumorVsNormal up	R-HSA-212300	PRC2 methylates histones and DNA	5.81e-11	10.54
TumorVsNormal up	R-HSA-2730905	Role of LAT2/NTAL/LAB on calcium mobilization	1.11e-16	18.85
TumorVsNormal up	R-HSA-2029480	Fcgamma receptor (FCGR) dependent phagocytosis	1.11e-16	7.74
TumorVsNormal up	R-HSA-983705	Signaling by the B Cell Receptor (BCR)	1.11e-16	5.72
TumorVsNormal up	R-HSA-173623	Classical antibody-mediated complement activation	1.11e-16	28.07
TumorVsNormal up	R-HSA-166786	Creation of C4 and C2 activators	1.11e-16	20.34
TumorVsNormal up	R-HSA-2029482	Regulation of actin dynamics for phagocytic cup formation	1.11e-16	10.54
TumorVsNormal up	R-HSA-9662851	Anti-inflammatory response favouring Leishmania parasite infection	1.11e-16	7.74
TumorVsNormal up	R-HSA-9664433	Leishmania parasite growth and survival	1.11e-16	7.74
TumorVsNormal up	R-HSA-2454202	Fc epsilon receptor (FCERI) signaling	1.11e-16	5.31
TumorVsNormal up	R-HSA-166663	Initial triggering of complement	1.11e-16	17.27

A.6 Gene Ontology Biological Processes (GO:BP) for Full Dataset

Comparison	GO:id	Biological Process	Adj. p-value	Odds ratio
TissueXSex up	GO:0045109	intermediate filament organization	3.40e-11	457.45
TissueXSex up	GO:0045104	intermediate filament cytoskeleton organization	2.77e-10	345.65
TissueXSex up	GO:0045103	intermediate filament-based process	3.02e-10	341.67
TissueXSex up	GO:0008544	epidermis development	3.04e-08	107.97
TissueXSex up	GO:0030216	keratinocyte differentiation	4.54e-08	176.73
TissueXSex up	GO:0030855	epithelial cell differentiation	7.31e-07	65.97
TissueXSex up	GO:0009913	epidermal cell differentiation	7.47e-07	122.56
TissueXSex up	GO:0009888	tissue development	2.17e-06	43.04
TissueXSex up	GO:0043588	skin development	5.36e-06	94.66
TissueXSex up	GO:0060429	epithelium development	1.05e-05	45.82
TissueXSex up	GO:0031424	keratinization	4.85e-03	157.43
TissueXSex up	GO:0097435	supramolecular fiber organization	8.46e-03	35.26
TissueXSex up	GO:0018149	peptide cross-linking	8.68e-03	353.3
TissueXSex up	GO:0009753	response to jasmonic acid	1.42e-02	2827.28
TissueXSex up	GO:0071395	cellular response to jasmonic acid stimulus	1.42e-02	2827.28
TissueXSex up	GO:0048513	animal organ development	3.07e-02	20.76
TissueXSex down	GO:0007585	respiratory gaseous exchange by respiratory system	4.73e-02	148.66
Age up	GO:0008544	epidermis development	9.65e-07	86.72
Age up	GO:0045109	intermediate filament organization	4.39e-05	225.75
Age up	GO:0031424	keratinization	7.92e-05	199.34
Age up	GO:0030216	keratinocyte differentiation	1.02e-04	113.12
Age up	GO:0043588	skin development	1.50e-04	74.7
Age up	GO:0045104	intermediate filament cytoskeleton organization	1.58e-04	172.43
Age up	GO:0045103	intermediate filament-based process	1.66e-04	170.51
Age up	GO:0009913	epidermal cell differentiation	8.08e-04	78.73
Age up	GO:0019730	antimicrobial humoral response	8.29e-04	121.73
Age up	GO:0030855	epithelial cell differentiation	3.25e-03	38.22
Age up	GO:0009888	tissue development	1.89e-02	22.23
Age up	GO:0006959	humoral immune response	2.13e-02	61.49
Age up	GO:0006958	complement activation, classical pathway	2.80e-02	223.09
Age down	GO:0006334	nucleosome assembly	1.01e-09	176.21
Age down	GO:0034728	nucleosome organization	4.38e-09	148.2
Age down	GO:0061644	protein localization to CENP-A containing chromatin	2.44e-07	715.13
Age down	GO:0065004	protein-DNA complex assembly	7.12e-07	81.42

Age down	GO:0071168	protein localization to chromatin	1.77e-06	220.22
Age down	GO:0006396	RNA processing	2.39e-06	23.04
Age down	GO:0071459	protein localization to chromosome, centromeric region	1.82e-05	265.54
Age down	GO:0045653	negative regulation of megakaryocyte differentiation	6.61e-05	483.1
Age down	GO:0034502	protein localization to chromosome	1.54e-04	99.49
Age down	GO:0045652	regulation of megakaryocyte differentiation	1.10e-03	219.53
Age down	GO:0006325	chromatin organization	2.88e-03	26.27
Age down	GO:0006338	chromatin remodeling	4.62e-03	28.29
Age down	GO:0016070	RNA metabolic process	6.92e-03	15.55
Age down	GO:0071824	protein-DNA complex organization	8.02e-03	23.36
Age down	GO:0030219	megakaryocyte differentiation	1.09e-02	118.72
Age down	GO:0051276	chromosome organization	3.06e-02	25.53
Age down	GO:0090304	nucleic acid metabolic process	3.50e-02	14.27
Age down	GO:0032200	telomere organization	4.53e-02	49.86
TumorVsNormal up	GO:0000278	mitotic cell cycle	3.20e-26	12.5
TumorVsNormal up	GO:1903047	mitotic cell cycle process	4.64e-26	13.52
TumorVsNormal up	GO:0022402	cell cycle process	2.25e-22	10.11
TumorVsNormal up	GO:0140014	mitotic nuclear division	4.87e-22	20.88
TumorVsNormal up	GO:0007059	chromosome segregation	6.11e-22	16.34
TumorVsNormal up	GO:0016064	immunoglobulin mediated immune response	9.18e-22	25.37
TumorVsNormal up	GO:0019724	B cell mediated immunity	1.83e-21	24.89
TumorVsNormal up	GO:0002250	adaptive immune response	2.66e-21	12.28
TumorVsNormal up	GO:0051301	cell division	1.21e-20	12.76
TumorVsNormal up	GO:0050896	response to stimulus	1.49e-20	8.12
TumorVsNormal up	GO:0098813	nuclear chromosome segregation	6.84e-20	18.17
TumorVsNormal up	GO:0000819	sister chromatid segregation	2.43e-19	21.73
TumorVsNormal up	GO:0051276	chromosome organization	4.22e-19	12.47
TumorVsNormal up	GO:0000070	mitotic sister chromatid segregation	1.17e-18	24.14
TumorVsNormal up	GO:0000280	nuclear division	1.54e-18	14.54
TumorVsNormal up	GO:0007049	cell cycle	9.12e-17	8.1
TumorVsNormal up	GO:0048285	organelle fission	9.31e-17	13.15

TumorVsNormal up	GO:0044770	cell cycle phase transition	1.09e-16	12.63
TumorVsNormal up	GO:0051983	regulation of chromosome segregation	1.63e-15	27.87
TumorVsNormal up	GO:0010564	regulation of cell cycle process	5.52e-15	10.57
TumorVsNormal down	GO:0040011	locomotion	4.59e-45	6.18
TumorVsNormal down	GO:0048870	cell motility	4.71e-45	5.42
TumorVsNormal down	GO:0016477	cell migration	3.08e-44	5.66
TumorVsNormal down	GO:0051239	regulation of multicellular organismal process	5.93e-44	4.48
TumorVsNormal down	GO:0030334	regulation of cell migration	3.84e-42	6.79
TumorVsNormal down	GO:2000145	regulation of cell motility	2.85e-41	6.53
TumorVsNormal down	GO:0040012	regulation of locomotion	6.90e-41	6.38
TumorVsNormal down	GO:0001944	vasculature development	3.28e-39	7.28
TumorVsNormal down	GO:0072359	circulatory system development	3.76e-37	5.89
TumorVsNormal down	GO:0001568	blood vessel development	3.79e-37	7.23
TumorVsNormal down	GO:0048856	anatomical structure development	7.86e-36	3.63
TumorVsNormal down	GO:0007155	cell adhesion	1.69e-34	5.11
TumorVsNormal down	GO:0007166	cell surface receptor signaling pathway	5.01e-33	4.15
TumorVsNormal down	GO:0009653	anatomical structure morphogenesis	9.95e-33	4.19
TumorVsNormal down	GO:0032501	multicellular organismal process	4.03e-32	3.44
TumorVsNormal down	GO:0007275	multicellular organism development	5.09e-32	3.68
TumorVsNormal down	GO:0030335	positive regulation of cell migration	1.20e-31	7.73
TumorVsNormal down	GO:2000147	positive regulation of cell motility	1.62e-31	7.51
TumorVsNormal down	GO:0040017	positive regulation of locomotion	2.25e-31	7.4
TumorVsNormal down	GO:0048646	anatomical structure formation involved in morphogenesis	5.15e-31	5.37

A.7 Gene Ontology Biological Processes (GO:BP) for Smoking Dataset

Comparison	GO:id	Biological Process	Adj. p-value	Odds ratio
Age up	GO:0002250	adaptive immune response	4.08e-11	264.30
Age up	GO:0006955	immune response	1.92e-06	95.59
Age up	GO:0002460	adaptive immune response based on somatic recombination of immune receptors built from immunoglobulin superfamily domains	2.20e-06	167.22
Age up	GO:0016064	immunoglobulin mediated immune response	2.87e-06	237.37
Age up	GO:0019724	B cell mediated immunity	3.13e-06	233.81
Age up	GO:0002376	immune system process	6.79e-05	67.56
Age up	GO:0002449	lymphocyte mediated immunity	8.63e-05	132.10
Age up	GO:0002252	immune effector process	1.59e-04	88.65
Age up	GO:0002443	leukocyte mediated immunity	3.61e-04	103.15
FormerVsNever Tumor up	GO:0071395	cellular response to jasmonic acid stimulus	2.15e-04	17675.75
FormerVsNever Tumor up	GO:0009753	response to jasmonic acid	2.15e-04	17675.75
FormerVsNever Tumor up	GO:0044597	daunorubicin metabolic process	1.29e-03	5049.86
FormerVsNever Tumor up	GO:0030647	aminoglycoside antibiotic metabolic process	1.61e-03	4418.56
FormerVsNever Tumor up	GO:0044598	doxorubicin metabolic process	1.61e-03	4418.56
FormerVsNever Tumor up	GO:0030638	polyketide metabolic process	1.61e-03	4418.56
FormerVsNever Tumor up	GO:0042448	progesterone metabolic process	4.87e-03	2356.33
FormerVsNever Tumor up	GO:0016137	glycoside metabolic process	6.81e-03	1963.53
FormerVsNever Tumor up	GO:1902644	tertiary alcohol metabolic process	8.27e-03	1767.12
FormerVsNever Tumor up	GO:0071398	cellular response to fatty acid	2.13e-02	1070.79
FormerVsNever Tumor up	GO:0008207	C21-steroid hormone metabolic process	2.51e-02	981.51
FormerVsNever Tumor up	GO:1901661	quinone metabolic process	3.08e-02	883.31
FormerVsNever Tumor up	GO:0006693	prostaglandin metabolic process	4.03e-02	768.03
FormerVsNever Tumor up	GO:0006692	prostanoid metabolic process	4.20e-02	751.68

CurrentVsFormer Normal up	GO:0009692	ethylene metabolic process	4.95e-02	inf
CurrentVsFormer Normal up	GO:0017143	insecticide metabolic process	4.95e-02	inf
CurrentVsFormer Normal up	GO:0019341	dibenzo-p-dioxin catabolic process	4.95e-02	inf
CurrentVsNever Tumor down	GO:0001580	detection of chemical stimulus involved in sensory perception of bitter taste	1.51e-03	389.66
CurrentVsNever Tumor down	GO:0050913	sensory perception of bitter taste	2.23e-03	339.68
CurrentVsNever Tumor down	GO:0050912	detection of chemical stimulus involved in sensory perception of taste	2.40e-03	331.18
CurrentVsNever Tumor down	GO:0050909	sensory perception of taste	9.25e-03	206.92
TumorVsNormal down	GO:0048856	anatomical structure development	1.02e-59	5.62
TumorVsNormal down	GO:0051239	regulation of multicellular organismal process	4.28e-57	6.47
TumorVsNormal down	GO:0048870	cell motility	1.23e-55	7.75
TumorVsNormal down	GO:0032501	multicellular organismal process	2.00e-53	5.28
TumorVsNormal down	GO:0032502	developmental process	3.59e-53	5.35
TumorVsNormal down	GO:0007275	multicellular organism development	1.42e-52	5.61
TumorVsNormal down	GO:0016477	cell migration	6.95e-52	7.92
TumorVsNormal down	GO:0009653	anatomical structure morphogenesis	4.51e-48	6.22
TumorVsNormal down	GO:0001944	vasculature development	2.05e-47	10.46
TumorVsNormal down	GO:0048731	system development	2.31e-47	5.59
TumorVsNormal down	GO:0072359	circulatory system development	5.09e-47	8.53
TumorVsNormal down	GO:0040011	locomotion	1.85e-46	8.16
TumorVsNormal down	GO:0001568	blood vessel development	3.99e-46	10.54
TumorVsNormal down	GO:0050896	response to stimulus	1.34e-45	5.03
TumorVsNormal down	GO:0007155	cell adhesion	5.88e-43	7.24
TumorVsNormal down	GO:0007166	cell surface receptor signaling pathway	3.71e-42	5.85
TumorVsNormal down	GO:0030334	regulation of cell migration	3.16e-41	8.72
TumorVsNormal down	GO:0040012	regulation of locomotion	4.50e-41	8.30

TumorVsNormal down	GO:2000145	regulation of cell motility	5.08e-41	8.45
TumorVsNormal down	GO:0051716	cellular response to stimulus	2.01e-40	4.89
TumorVsNormal up	GO:0002250	adaptive immune response	2.34e-38	18.52
TumorVsNormal up	GO:0016064	immunoglobulin mediated immune response	8.24e-30	35.52
TumorVsNormal up	GO:0019724	B cell mediated immunity	1.73e-29	34.84
TumorVsNormal up	GO:1903047	mitotic cell cycle process	7.02e-25	14.34
TumorVsNormal up	GO:0098813	nuclear chromosome segregation	4.37e-24	22.58
TumorVsNormal up	GO:0007059	chromosome segregation	1.33e-23	18.75
TumorVsNormal up	GO:0140014	mitotic nuclear division	3.89e-23	23.91
TumorVsNormal up	GO:0000278	mitotic cell cycle	8.39e-23	12.58
TumorVsNormal up	GO:0000070	mitotic sister chromatid segregation	1.13e-22	30.73
TumorVsNormal up	GO:0051276	chromosome organization	3.90e-22	14.58
TumorVsNormal up	GO:0000819	sister chromatid segregation	4.13e-22	26.36
TumorVsNormal up	GO:0050896	response to stimulus	1.72e-20	7.72
TumorVsNormal up	GO:0022402	cell cycle process	3.44e-20	10.26
TumorVsNormal up	GO:0000280	nuclear division	5.02e-20	16.58
TumorVsNormal up	GO:0002460	adaptive immune response based on somatic recombination of immune receptors built from immunoglobulin superfamily domains	9.91e-20	18.13
TumorVsNormal up	GO:0006955	immune response	3.57e-19	8.64
TumorVsNormal up	GO:0044770	cell cycle phase transition	5.81e-19	14.61
TumorVsNormal up	GO:0002449	lymphocyte mediated immunity	7.52e-19	18.05
TumorVsNormal up	GO:0051301	cell division	1.65e-18	13.08
TumorVsNormal up	GO:0048285	organelle fission	7.89e-18	14.75

A.8 Gene Ontology Cellular Components (GO:CC) for Full Dataset (Top 20)

Comparison	GO:id	Cellular Components	Adj. p-value	Odds ratio
TissueXSex up	GO:0001533	cornified envelope	2.19e-15	706.34
TissueXSex up	GO:0005882	intermediate filament	2.12e-08	141.3
TissueXSex up	GO:0045111	intermediate filament cytoskeleton	7.13e-08	120.56
TissueXSex up	GO:0045095	keratin filament	2.79e-07	219.23
TissueXSex up	GO:0030057	desmosome	5.19e-06	534.3
TissueXSex up	GO:0005911	cell-cell junction	3.38e-04	47.97
TissueXSex up	GO:0099512	supramolecular fiber	3.70e-04	33.32
TissueXSex up	GO:0099081	supramolecular polymer	3.94e-04	33.06
TissueXSex up	GO:0005576	extracellular region	4.44e-04	23.04
TissueXSex up	GO:0099513	polymeric cytoskeletal fiber	5.32e-04	36.83
TissueXSex up	GO:0070161	anchoring junction	1.24e-03	32.76
TissueXSex up	GO:0005829	cytosol	2.92e-03	20.27
TissueXSex up	GO:0099080	supramolecular complex	4.33e-03	24.44
TissueXSex up	GO:0005615	extracellular space	5.53e-03	19.02
TissueXSex up	GO:0005737	cytoplasm	1.19e-02	37.89
TissueXSex down	GO:0042599	lamellar body	1.10e-07	1308.83
TissueXSex down	GO:0005771	multivesicular body	5.38e-05	237.79
TissueXSex down	GO:0097208	alveolar lamellar body	2.29e-03	1413.68
TissueXSex down	GO:0097486	multivesicular body lumen	2.29e-03	1413.68
TissueXSex down	GO:0031982	vesicle	2.72e-03	20.05
TissueXSex down	GO:0031906	late endosome lumen	3.91e-03	1009.74
TissueXSex down	GO:0005576	extracellular region	4.65e-03	18.99
TissueXSex down	GO:0045334	clathrin-coated endocytic vesicle	8.80e-03	131.15
TissueXSex down	GO:0030141	secretory granule	1.17e-02	29.41
TissueXSex down	GO:0005770	late endosome	2.04e-02	50.45
TissueXSex down	GO:0005615	extracellular space	2.15e-02	17.06
TissueXSex down	GO:0099503	secretory vesicle	3.10e-02	24.56
Age up	GO:0042571	immunoglobulin complex, circulating	4.74e-05	1060.12
Age up	GO:0005576	extracellular region	6.00e-05	24.49
Age up	GO:0005615	extracellular space	1.20e-04	23.65
Age up	GO:0001533	cornified envelope	1.31e-04	214.02
Age up	GO:0019814	immunoglobulin complex	1.69e-04	104.31
Age up	GO:0005882	intermediate filament	9.92e-04	71.92

Age up	GO:0045111	intermediate filament cytoskeleton	2.08e-03	61.49
Age up	GO:0071735	IgG immunoglobulin complex	7.42e-03	906.12
Age up	GO:0045095	keratin filament	4.08e-02	89.16
Age down	GO:0000786	nucleosome	6.88e-14	195.56
Age down	GO:0005730	nucleolus	1.91e-09	30.15
Age down	GO:0061638	CENP-A containing chromatin	2.17e-08	715.13
Age down	GO:0043505	CENP-A containing nucleosome	2.17e-08	715.13
Age down	GO:0034506	chromosome, centromeric core domain	2.94e-08	664.04
Age down	GO:0043232	intracellular non-membrane-bounded organelle	1.21e-07	22.55
Age down	GO:0043228	non-membrane-bounded organelle	1.22e-07	22.54
Age down	GO:0005634	nucleus	3.67e-07	26.42
Age down	GO:0031981	nuclear lumen	6.94e-07	20.02
Age down	GO:0043233	organelle lumen	5.24e-05	16.3
Age down	GO:0031974	membrane-enclosed lumen	5.24e-05	16.3
Age down	GO:0070013	intracellular organelle lumen	5.24e-05	16.3
Age down	GO:0000785	chromatin	2.82e-03	17.56
Age down	GO:0032993	protein-DNA complex	4.44e-03	16.7
Age down	GO:0000775	chromosome, centromeric region	1.75e-02	36.48
Age down	GO:0043231	intracellular membrane-bounded organelle	4.40e-02	16.84
Age down	GO:0000781	chromosome, telomeric region	4.69e-02	43.3
TumorVsNormal up	GO:0019814	immunoglobulin complex	1.41e-75	103.58
TumorVsNormal up	GO:0005576	extracellular region	2.15e-38	9.31
TumorVsNormal up	GO:0005615	extracellular space	7.50e-27	8.3
TumorVsNormal up	GO:0070062	extracellular exosome	3.40e-16	7.54
TumorVsNormal up	GO:1903561	extracellular vesicle	4.16e-16	7.51
TumorVsNormal up	GO:0043230	extracellular organelle	4.35e-16	7.5
TumorVsNormal up	GO:0065010	extracellular membrane-bounded organelle	4.35e-16	7.5
TumorVsNormal up	GO:0098687	chromosomal region	3.60e-15	13.43
TumorVsNormal up	GO:0001533	cornified envelope	2.00e-14	47.64
TumorVsNormal up	GO:0000793	condensed chromosome	3.55e-14	15.58
TumorVsNormal up	GO:0000775	chromosome, centromeric region	2.11e-13	15.83

TumorVsNormal up	GO:0000779	condensed chromosome, centromeric region	2.71e-13	19.33
TumorVsNormal up	GO:0005819	spindle	9.00e-12	11.44
TumorVsNormal up	GO:0005694	chromosome	1.65e-11	6.86
TumorVsNormal up	GO:0000776	kinetochore	6.53e-11	17.9
TumorVsNormal up	GO:0071944	cell periphery	8.14e-11	6.2
TumorVsNormal up	GO:0072686	mitotic spindle	1.03e-10	16.92
TumorVsNormal up	GO:0000940	outer kinetochore	4.50e-10	105.65
TumorVsNormal up	GO:0005737	cytoplasm	4.67e-10	8
TumorVsNormal up	GO:0099080	supramolecular complex	2.84e-09	6.91
TumorVsNormal down	GO:0071944	cell periphery	7.93e-48	3.87
TumorVsNormal down	GO:0005886	plasma membrane	3.95e-32	3.54
TumorVsNormal down	GO:0009986	cell surface	3.31e-24	5.33
TumorVsNormal down	GO:0005576	extracellular region	2.03e-21	3.38
TumorVsNormal down	GO:0062023	collagen-containing extracellular matrix	1.55e-18	6.58
TumorVsNormal down	GO:0030312	external encapsulating structure	1.59e-18	5.82
TumorVsNormal down	GO:0031012	extracellular matrix	4.22e-18	5.77
TumorVsNormal down	GO:0098552	side of membrane	6.62e-18	5.17
TumorVsNormal down	GO:0031982	vesicle	1.06e-17	3.27
TumorVsNormal down	GO:0009897	external side of plasma membrane	2.31e-17	6.69
TumorVsNormal down	GO:0070161	anchoring junction	7.00e-17	4.66
TumorVsNormal down	GO:0005615	extracellular space	1.94e-13	3.18
TumorVsNormal down	GO:0031410	cytoplasmic vesicle	2.44e-12	3.27
TumorVsNormal down	GO:0097708	intracellular vesicle	3.18e-12	3.26
TumorVsNormal down	GO:0030141	secretory granule	4.47e-12	4.19
TumorVsNormal down	GO:0015629	actin cytoskeleton	9.18e-12	5.02
TumorVsNormal down	GO:0005911	cell-cell junction	2.61e-10	4.78
TumorVsNormal down	GO:0042995	cell projection	4.54e-10	3.17
TumorVsNormal down	GO:0099503	secretory vesicle	8.81e-10	3.75
TumorVsNormal down	GO:0030054	cell junction	9.91e-10	3.18

A.9 Gene Ontology Cellular Components (GO:CC) for Smoking Dataset (Top 20)

Comparison	GO:id	Cellular Components	Adj. p-value	Odds ratio
Age up	GO:0019814	immunoglobulin complex	6.64e-15	740.07
Age up	GO:0005576	extracellular region	1.04e-08	221.71
Age up	GO:0005615	extracellular space	2.42e-05	56.31
Age up	GO:0071944	cell periphery	9.14e-05	67.44
Age up	GO:0072562	blood microparticle	1.65e-04	194.39
Age up	GO:0005886	plasma membrane	7.37e-03	31.17
CurrentVsFormer Normal up	GO:0033181	plasma membrane proton-transporting V-type ATPase complex	3.69e-02	70708
CurrentVsNever Tumor down	GO:0031982	vesicle	6.79e-04	18.56
CurrentVsNever Tumor down	GO:0098552	side of membrane	4.06e-03	34.9
CurrentVsNever Tumor down	GO:0009897	external side of plasma membrane	6.10e-03	49.34
CurrentVsNever Tumor down	GO:0070062	extracellular exosome	7.85e-03	19.03
CurrentVsNever Tumor down	GO:1903561	extracellular vesicle	8.44e-03	18.81
CurrentVsNever Tumor down	GO:0043230	extracellular organelle	8.47e-03	18.8
CurrentVsNever Tumor down	GO:0065010	extracellular membrane-bounded organelle	8.47e-03	18.8
CurrentVsNever Tumor down	GO:0005615	extracellular space	1.69e-02	14.88
CurrentVsNever Tumor down	GO:0005886	plasma membrane	1.92e-02	12.59
CurrentVsNever Tumor down	GO:0071944	cell periphery	4.00e-02	11.52
CurrentVsNever Tumor down	GO:0030659	cytoplasmic vesicle membrane	4.96e-02	20.32
TumorVsNormal down	GO:0071944	cell periphery	7.22e-72	6.01
TumorVsNormal down	GO:0005886	plasma membrane	6.95e-48	5.28
TumorVsNormal down	GO:0005576	extracellular region	1.30e-35	5.05
TumorVsNormal down	GO:0009986	cell surface	1.50e-33	7.96
TumorVsNormal down	GO:0030312	external encapsulating structure	2.81e-28	9.19
TumorVsNormal down	GO:0031012	extracellular matrix	9.54e-28	9.1
TumorVsNormal down	GO:0062023	collagen-containing extracellular matrix	4.43e-27	10.32

TumorVsNormal down	GO:0031982	vesicle	1.18e-23	4.53
TumorVsNormal down	GO:0005615	extracellular space	2.46e-22	4.61
TumorVsNormal down	GO:0098552	side of membrane	7.83e-22	7.22
TumorVsNormal down	GO:0016020	membrane	1.92e-20	4.27
TumorVsNormal down	GO:0070161	anchoring junction	2.04e-20	6.42
TumorVsNormal down	GO:0009897	external side of plasma membrane	1.05e-19	9.2
TumorVsNormal down	GO:0031410	cytoplasmic vesicle	1.36e-16	4.48
TumorVsNormal down	GO:0097708	intracellular vesicle	1.80e-16	4.47
TumorVsNormal down	GO:0042995	cell projection	4.96e-15	4.41
TumorVsNormal down	GO:0030141	secretory granule	7.81e-15	5.72
TumorVsNormal down	GO:0005737	cytoplasm	7.81e-15	4.24
TumorVsNormal down	GO:0120025	plasma membrane bounded cell projection	4.75e-14	4.38
TumorVsNormal down	GO:0005911	cell-cell junction	9.94e-14	6.8
TumorVsNormal up	GO:0019814	immunoglobulin complex	9.08e-102	174.84
TumorVsNormal up	GO:0005576	extracellular region	4.28e-49	10.86
TumorVsNormal up	GO:0005615	extracellular space	3.21e-30	8.96
TumorVsNormal up	GO:0000786	nucleosome	1.08e-20	35.47
TumorVsNormal up	GO:0071944	cell periphery	3.00e-20	7.15
TumorVsNormal up	GO:0005694	chromosome	6.70e-19	8.38
TumorVsNormal up	GO:0098687	chromosomal region	3.33e-17	15.55
TumorVsNormal up	GO:0000775	chromosome, centromeric region	1.27e-16	19.58
TumorVsNormal up	GO:0005886	plasma membrane	4.28e-14	6.36
TumorVsNormal up	GO:0000793	condensed chromosome	1.07e-13	16.63
TumorVsNormal up	GO:0000779	condensed chromosome, centromeric region	1.35e-13	21.52
TumorVsNormal up	GO:0070062	extracellular exosome	1.58e-13	7.21
TumorVsNormal up	GO:1903561	extracellular vesicle	3.61e-13	7.13

TumorVsNormal up	GO:0065010	extracellular membrane-bounded organelle	3.74e-13	7.12
TumorVsNormal up	GO:0043230	extracellular organelle	3.74e-13	7.12
TumorVsNormal up	GO:0000776	kinetochore	1.56e-12	21.28
TumorVsNormal up	GO:0000940	outer kinetochore	5.36e-12	141.69
TumorVsNormal up	GO:0072562	blood microparticle	2.57e-11	23.16
TumorVsNormal up	GO:0001533	cornified envelope	3.49e-10	39.76
TumorVsNormal up	GO:0005819	spindle	1.09e-09	11.15

A.10 Gene Ontology Molecular Functions (GO:MF) for Full Dataset

Comparison	GO:id	Molecular Functions	Adj. p-value	Odds ratio
TissueXSex up	GO:0030280	structural constituent of skin epidermis	2.98e-07	517.98
TissueXSex up	GO:0005200	structural constituent of cytoskeleton	1.04e-04	148.52
TissueXSex up	GO:0018636	phenanthrene 9,10-monooxygenase activity	1.66e-03	5654.64
TissueXSex up	GO:0047115	trans-1,2-dihydrobenzene-1,2-diol dehydrogenase activity	1.66e-03	5654.64
TissueXSex up	GO:0047718	indanol dehydrogenase activity	1.66e-03	5654.64
TissueXSex up	GO:0005198	structural molecule activity	1.90e-03	31.89
TissueXSex up	GO:0047086	ketosteroid monooxygenase activity	5.51e-03	1884.83
TissueXSex up	GO:0047023	androsterone dehydrogenase activity	1.54e-02	942.37
TissueXSex up	GO:0047044	androstan-3-alpha,17-beta-diol dehydrogenase activity	1.98e-02	807.74
TissueXSex up	GO:0032052	bile acid binding	2.47e-02	706.76
TissueXSex up	GO:0019215	intermediate filament binding	3.62e-02	565.39
TissueXSex up	GO:0004032	alditol:NADP+ 1-oxidoreductase activity	3.62e-02	565.39
Age down	GO:0030527	structural constituent of chromatin	2.17e-21	272.6
Age down	GO:0005198	structural molecule activity	9.07e-10	21.96
Age down	GO:0046982	protein heterodimerization activity	6.72e-09	45.27
Age down	GO:0003676	nucleic acid binding	2.89e-06	5.75
Age down	GO:0003677	DNA binding	7.25e-06	9.29
Age down	GO:0046983	protein dimerization activity	4.51e-05	14.39
Age down	GO:0097159	organic cyclic compound binding	2.49e-04	4.05
Age down	GO:0031492	nucleosomal DNA binding	6.31e-04	143.17
Age down	GO:0003723	RNA binding	1.03e-03	6.58
Age down	GO:0031491	nucleosome binding	3.45e-03	79.03
Age down	GO:0031490	chromatin DNA binding	1.82e-02	44.46
Age up	GO:0005198	structural molecule activity	1.81e-04	35.47
Age up	GO:0034987	immunoglobulin receptor binding	2.95e-04	706.71
Age up	GO:0003823	antigen binding	8.81e-04	90.7
Age up	GO:0030280	structural constituent of skin epidermis	4.55e-03	256.91
TumorVsNormal down	GO:0005102	signaling receptor binding	2.61e-14	3.77

TumorVsNormal down	GO:0005178	integrin binding	3.33e-10	8.19
TumorVsNormal down	GO:0098772	molecular function regulator activity	9.48e-10	3.2
TumorVsNormal down	GO:0030234	enzyme regulator activity	1.02e-09	3.55
TumorVsNormal down	GO:0005509	calcium ion binding	3.44e-09	4.1
TumorVsNormal down	GO:0005515	protein binding	4.36e-09	2.88
TumorVsNormal down	GO:0140375	immune receptor activity	5.88e-09	8.05
TumorVsNormal down	GO:0005201	extracellular matrix structural constituent	4.06e-08	7.13
TumorVsNormal down	GO:0003779	actin binding	5.06e-08	4.64
TumorVsNormal down	GO:0019838	growth factor binding	1.61e-07	7.73
TumorVsNormal down	GO:0005488	binding	2.40e-07	2.95
TumorVsNormal down	GO:0030695	GTPase regulator activity	2.51e-07	4.35
TumorVsNormal down	GO:0060589	nucleoside-triphosphatase regulator activity	2.51e-07	4.35
TumorVsNormal down	GO:0019955	cytokine binding	4.15e-07	7.22
TumorVsNormal down	GO:0008092	cytoskeletal protein binding	4.37e-07	3.51
TumorVsNormal down	GO:0060089	molecular transducer activity	3.56e-05	3.03
TumorVsNormal down	GO:0038023	signaling receptor activity	3.56e-05	3.03
TumorVsNormal down	GO:0008047	enzyme activator activity	3.88e-05	3.76
TumorVsNormal down	GO:0038024	cargo receptor activity	1.77e-04	7.71
TumorVsNormal down	GO:0019199	transmembrane receptor protein kinase activity	1.94e-04	7.99
TumorVsNormal down	GO:0030246	carbohydrate binding	2.33e-04	4.51
TumorVsNormal down	GO:0005085	guanyl-nucleotide exchange factor activity	2.78e-04	4.86
TumorVsNormal down	GO:0005096	GTPase activator activity	6.18e-04	4.48
TumorVsNormal down	GO:0005539	glycosaminoglycan binding	6.89e-04	4.61
TumorVsNormal down	GO:0008201	heparin binding	8.15e-04	5.23
TumorVsNormal down	GO:0140678	molecular function inhibitor activity	8.78e-04	3.67
TumorVsNormal down	GO:0004713	protein tyrosine kinase activity	1.19e-03	5.55

TumorVsNormal down	GO:0001540	amyloid-beta binding	1.26e-03	7.26
TumorVsNormal down	GO:0140677	molecular function activator activity	1.29e-03	3.02
TumorVsNormal down	GO:0050839	cell adhesion molecule binding	1.52e-03	3.51
TumorVsNormal down	GO:0004888	transmembrane signaling receptor activity	2.05e-03	2.93
TumorVsNormal down	GO:0004896	cytokine receptor activity	2.36e-03	6.42
TumorVsNormal down	GO:0071813	lipoprotein particle binding	3.07e-03	14.05
TumorVsNormal down	GO:0071814	protein-lipid complex binding	3.07e-03	14.05
TumorVsNormal down	GO:0019899	enzyme binding	4.01e-03	2.69
TumorVsNormal down	GO:0038187	pattern recognition receptor activity	4.24e-03	11.85
TumorVsNormal down	GO:0044877	protein-containing complex binding	5.44e-03	2.74
TumorVsNormal down	GO:0008289	lipid binding	6.42e-03	3.09
TumorVsNormal down	GO:0030169	low-density lipoprotein particle binding	7.20e-03	21.7
TumorVsNormal down	GO:0042277	peptide binding	1.01e-02	3.86
TumorVsNormal down	GO:0051015	actin filament binding	1.39e-02	4.42
TumorVsNormal down	GO:0004714	transmembrane receptor protein tyrosine kinase activity	1.62e-02	7.4
TumorVsNormal down	GO:0033218	amide binding	1.69e-02	3.59
TumorVsNormal down	GO:0019956	chemokine binding	1.90e-02	10.86
TumorVsNormal down	GO:1901681	sulfur compound binding	2.95e-02	3.95
TumorVsNormal down	GO:0050431	transforming growth factor beta binding	3.46e-02	13.02
TumorVsNormal down	GO:0019865	immunoglobulin binding	3.46e-02	13.02
TumorVsNormal down	GO:0050840	extracellular matrix binding	3.48e-02	7.25
TumorVsNormal down	GO:0005021	vascular endothelial growth factor receptor activity	3.57e-02	54.21
TumorVsNormal down	GO:0004672	protein kinase activity	4.15e-02	3.18
TumorVsNormal up	GO:0003823	antigen binding	2.40e-48	55.6
TumorVsNormal up	GO:0050839	cell adhesion molecule binding	1.01e-04	7.86
TumorVsNormal up	GO:0005515	protein binding	1.21e-04	10.02

TumorVsNormal up	GO:0005488	binding	3.42e-04	16.35
TumorVsNormal up	GO:0030280	structural constituent of skin epidermis	5.56e-04	30.4
TumorVsNormal up	GO:0045296	cadherin binding	2.34e-03	8.53
TumorVsNormal up	GO:0008017	microtubule binding	3.48e-03	9.01
TumorVsNormal up	GO:0005200	structural constituent of cytoskeleton	5.79e-03	13.11
TumorVsNormal up	GO:0034987	immunoglobulin receptor binding	8.99e-03	52.47
TumorVsNormal up	GO:0016538	cyclin-dependent protein serine/threonine kinase regulator activity	9.34e-03	20.8
TumorVsNormal up	GO:0017116	single-stranded DNA helicase activity	1.34e-02	34.47
TumorVsNormal up	GO:0005198	structural molecule activity	1.66e-02	5.77
TumorVsNormal up	GO:0015631	tubulin binding	1.98e-02	7.49
TumorVsNormal up	GO:0030020	extracellular matrix structural constituent conferring tensile strength	3.55e-02	19.73
TumorVsNormal up	GO:0005046	KDEL sequence binding	4.11e-02	inf
TumorVsNormal up	GO:0003777	microtubule motor activity	4.26e-02	15.26
MaleVsFemale up	GO:0141052	histone H3 demethylase activity	7.17e-03	523.5
MaleVsFemale up	GO:0032452	histone demethylase activity	8.21e-03	487.38
MaleVsFemale up	GO:0140457	protein demethylase activity	8.21e-03	487.38
MaleVsFemale up	GO:0032451	demethylase activity	1.44e-02	362.36
MaleVsFemale up	GO:0016706	2-oxoglutarate-dependent dioxygenase activity	3.21e-02	239.46

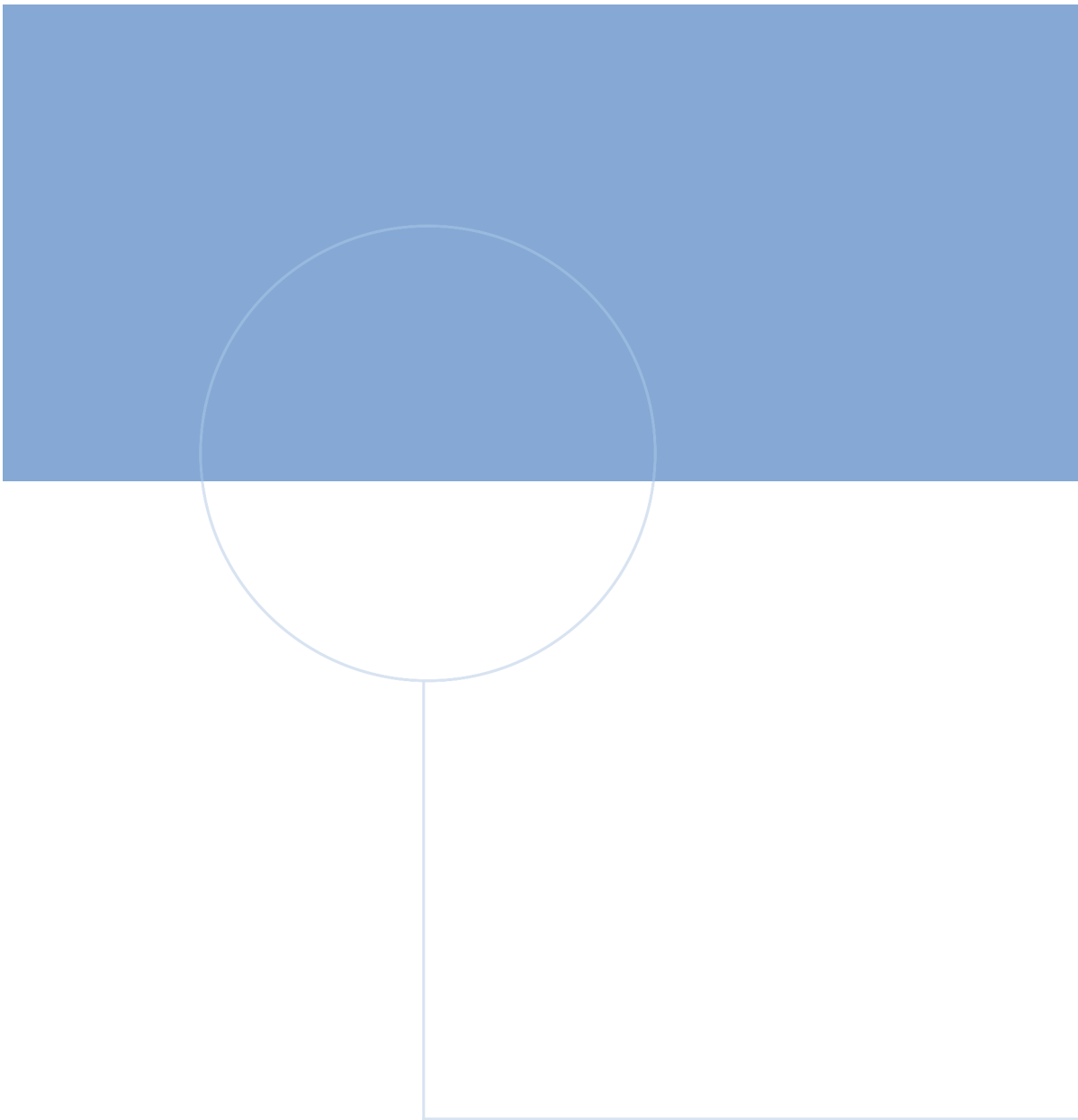
A.11 Gene Ontology Molecular Functions (GO:MF) for Smoking Dataset

Comparison	GO:id	Molecular Functions	Adj. p-value	Odds ratio
Age up	GO:0003823	antigen binding	9.97e-14	641.19
FormerVsNever Tumor up	GO:0008106	alcohol dehydrogenase (NADP+) activity	2.59e-06	3534.25
FormerVsNever Tumor up	GO:0004033	aldo-keto reductase (NADP) activity	5.94e-06	2617.7
FormerVsNever Tumor up	GO:0018636	phenanthrene 9,10-monooxygenase activity	2.96e-05	35352
FormerVsNever Tumor up	GO:0047718	indanol dehydrogenase activity	2.96e-05	35352
FormerVsNever Tumor up	GO:0047115	trans-1,2-dihydrobenzene-1,2-diol dehydrogenase activity	2.96e-05	35352
FormerVsNever Tumor up	GO:0047086	ketosteroid monooxygenase activity	9.87e-05	11783.67
FormerVsNever Tumor up	GO:0047023	androsterone dehydrogenase activity	2.76e-04	5891.58
FormerVsNever Tumor up	GO:0047044	androstan-3-alpha,17-beta-diol dehydrogenase activity	3.55e-04	5049.86
FormerVsNever Tumor up	GO:0032052	bile acid binding	4.44e-04	4418.56
FormerVsNever Tumor up	GO:0016616	oxidoreductase activity, acting on the CH-OH group of donors, NAD or NADP as acceptor	4.96e-04	564.64
FormerVsNever Tumor up	GO:0016614	oxidoreductase activity, acting on CH-OH group of donors	6.22e-04	522.74
FormerVsNever Tumor up	GO:0004032	alditol:NADP+ 1-oxidoreductase activity	6.51e-04	3534.75
FormerVsNever Tumor up	GO:0004303	estradiol 17-beta-dehydrogenase [NAD(P)] activity	1.87e-03	1963.53
FormerVsNever Tumor up	GO:0033764	steroid dehydrogenase activity, acting on the CH-OH group of donors, NAD or NADP as acceptor	5.20e-03	1139.9
FormerVsNever Tumor up	GO:0016628	oxidoreductase activity, acting on the CH-CH group of donors, NAD or NADP as acceptor	6.20e-03	1039.28
FormerVsNever Tumor up	GO:0016229	steroid dehydrogenase activity	6.55e-03	1009.57
FormerVsNever Tumor up	GO:0016709	oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular	6.91e-03	981.51

		oxygen, NAD(P)H as one donor, and incorporation of one atom of oxygen		
FormerVsNever Tumor up	GO:0016655	oxidoreductase activity, acting on NAD(P)H, quinone or similar compound as acceptor	1.57e-02	642.27
FormerVsNever Tumor up	GO:0016627	oxidoreductase activity, acting on the CH-CH group of donors	2.10e-02	551.88
FormerVsNever Tumor up	GO:0033293	monocarboxylic acid binding	3.10e-02	452.74
FormerVsNever Tumor up	GO:0016651	oxidoreductase activity, acting on NAD(P)H	3.75e-02	410.58
FormerVsNever Tumor up	GO:0047042	androsterone dehydrogenase (B-specific) activity	4.99e-02	inf
MaleVsFemale up	GO:0141052	histone H3 demethylase activity	7.17e-03	475.9
MaleVsFemale up	GO:0032452	histone demethylase activity	8.21e-03	443.07
MaleVsFemale up	GO:0140457	protein demethylase activity	8.21e-03	443.07
MaleVsFemale up	GO:0032451	demethylase activity	1.44e-02	329.41
MaleVsFemale up	GO:0016706	2-oxoglutarate-dependent dioxygenase activity	3.21e-02	217.69
CurrentVsFormer Normal up	GO:0016711	flavonoid 3'-monooxygenase activity	5.00e-02	inf
CurrentVsNever Normal up	GO:0016711	flavonoid 3'-monooxygenase activity	5.00e-02	inf
TumorVsNormal down	GO:0005515	protein binding	1.73e-19	5.02
TumorVsNormal down	GO:0005102	signaling receptor binding	7.31e-19	5.51
TumorVsNormal down	GO:0005178	integrin binding	7.52e-13	12.48
TumorVsNormal down	GO:0098772	molecular function regulator activity	3.51e-11	4.43
TumorVsNormal down	GO:0005509	calcium ion binding	1.40e-10	5.72
TumorVsNormal down	GO:0140375	immune receptor activity	1.14e-08	10.64
TumorVsNormal down	GO:0005201	extracellular matrix structural constituent	1.28e-08	9.82
TumorVsNormal down	GO:0030234	enzyme regulator activity	1.53e-07	4.46
TumorVsNormal down	GO:0019955	cytokine binding	2.15e-07	9.86
TumorVsNormal down	GO:0019838	growth factor binding	4.29e-07	10.09

TumorVsNormal down	GO:0008092	cytoskeletal protein binding	5.82e-07	4.65
TumorVsNormal down	GO:0005488	binding	8.37e-07	5.13
TumorVsNormal down	GO:0003779	actin binding	3.84e-06	5.7
TumorVsNormal down	GO:0030246	carbohydrate binding	4.33e-06	6.66
TumorVsNormal down	GO:0050839	cell adhesion molecule binding	8.77e-06	5.2
TumorVsNormal down	GO:0140678	molecular function inhibitor activity	1.40e-05	5.34
TumorVsNormal down	GO:0038024	cargo receptor activity	1.68e-05	11.33
TumorVsNormal down	GO:0038023	signaling receptor activity	2.76e-05	4.04
TumorVsNormal down	GO:0060089	molecular transducer activity	2.76e-05	4.04
TumorVsNormal down	GO:0001540	amyloid-beta binding	3.22e-05	11.36
TumorVsNormal down	GO:0019199	transmembrane receptor protein kinase activity	9.73e-05	11.02
TumorVsNormal down	GO:0008201	heparin binding	1.30e-04	7.42
TumorVsNormal down	GO:0005539	glycosaminoglycan binding	1.56e-04	6.42
TumorVsNormal down	GO:0008289	lipid binding	6.68e-04	4.28
TumorVsNormal down	GO:0004888	transmembrane signaling receptor activity	8.93e-04	3.93
TumorVsNormal down	GO:0004857	enzyme inhibitor activity	9.40e-04	5.2
TumorVsNormal down	GO:0004896	cytokine receptor activity	1.00e-03	8.91
TumorVsNormal down	GO:1901681	sulfur compound binding	1.38e-03	5.82
TumorVsNormal down	GO:0140677	molecular function activator activity	1.90e-03	3.96
TumorVsNormal down	GO:0042277	peptide binding	3.56e-03	5.28
TumorVsNormal down	GO:0033218	amide binding	3.73e-03	4.97
TumorVsNormal down	GO:0044877	protein-containing complex binding	4.14e-03	3.64
TumorVsNormal down	GO:0050840	extracellular matrix binding	7.59e-03	10.64
TumorVsNormal down	GO:0004955	prostaglandin receptor activity	1.16e-02	47.71
TumorVsNormal down	GO:0050431	transforming growth factor beta binding	1.22e-02	19.11
TumorVsNormal down	GO:0030545	signaling receptor regulator activity	1.32e-02	4.38

TumorVsNormal down	GO:0005044	scavenger receptor activity	1.81e-02	11.21
TumorVsNormal down	GO:0004954	prostanoid receptor activity	2.40e-02	38.17
TumorVsNormal down	GO:0030169	low-density lipoprotein particle binding	3.23e-02	24.75
TumorVsNormal down	GO:0008047	enzyme activator activity	3.72e-02	4.2
TumorVsNormal down	GO:0005126	cytokine receptor binding	4.14e-02	5.09
TumorVsNormal down	GO:0005518	collagen binding	4.43e-02	8.87
TumorVsNormal down	GO:0071814	protein-lipid complex binding	4.98e-02	15.08
TumorVsNormal down	GO:0071813	lipoprotein particle binding	4.98e-02	15.08
TumorVsNormal up	GO:0003823	antigen binding	2.55e-66	84.53
TumorVsNormal up	GO:0030527	structural constituent of chromatin	4.69e-26	56.59
TumorVsNormal up	GO:0005198	structural molecule activity	5.06e-11	8.34
TumorVsNormal up	GO:0046982	protein heterodimerization activity	2.70e-10	12.88
TumorVsNormal up	GO:0050839	cell adhesion molecule binding	6.97e-08	9.44
TumorVsNormal up	GO:0008017	microtubule binding	6.10e-05	10.67
TumorVsNormal up	GO:0046983	protein dimerization activity	6.27e-05	6.43
TumorVsNormal up	GO:0045296	cadherin binding	5.98e-04	9.1
TumorVsNormal up	GO:0003777	microtubule motor activity	9.18e-04	20.46
TumorVsNormal up	GO:0005201	extracellular matrix structural constituent	1.28e-03	12.11
TumorVsNormal up	GO:0030020	extracellular matrix structural constituent conferring tensile strength	1.33e-03	26.44
TumorVsNormal up	GO:0004222	metalloendopeptidase activity	3.59e-03	14.39
TumorVsNormal up	GO:0015631	tubulin binding	2.51e-02	7.41
TumorVsNormal up	GO:0010997	anaphase-promoting complex binding	3.34e-02	84
TumorVsNormal up	GO:0005488	binding	4.79e-02	8.96



NTNU

Norwegian University of
Science and Technology