Integrative scRNA-seq Analysis Illuminates LUAD Progression: Contrasting Primary and Metastatic Landscapes with TCGA Insights

By

Zia Ullah

(Registration No: 00000363387)

Department of Sciences

School of Interdisciplinary Engineering and Sciences

National University of Sciences & Technology (NUST)

Islamabad, Pakistan

(2024)

Integrative scRNA-seq Analysis Illuminates LUAD Progression: Contrasting Primary and Metastatic Landscapes with TCGA Insights

By

Zia Ullah

(Registration No: 00000363387)

A thesis submitted to the National University of Sciences and Technology, Islamabad,

in partial fulfillment of the requirements for the degree of

Master of Science in
Bioinformatics

Supervisor: Dr. Mehak Rafiq

.

School of Interdisciplinary Engineering and Sciences

National University of Sciences & Technology (NUST)

Islamabad, Pakistan

(2024)

Annex A to NUST Letter No.
0972/102/Exams/Thesis-Cert
dated 23 Dec 16.

THESIS ACCEPTANCE CERTIFICATE

Certified that final copy of MS/MPhil thesis written by Mr/Ms Zia Ullah Registration No. 00000363387 of SINES has been vetted by undersigned, found complete in all aspects as per NUST Statutes/Regulations, is free of plagiarism, errors, and mistakes and is accepted as partial fulfillment for award of MS/MPhil degree. It is further certified that necessary amendments as pointed out by GEC members of the scholar have also been incorporated in the said thesis.

Signature with stamp:

DR. MEHAK RAFIQ
Assistant Professor
National University
of Science & Technology
H-12 Islamabad

Name of Supervisor: Dr Mehak Rafiq

Date: 03 July, 2024

Signature of HoD with stamp: Fouzia

Dr Fouzia Malik
HoD Sciences
Professor
SINES NUST Sector H-12
Islamabad

Date: 10-7-2024

Countersign by

Signature (Dean/Principal):

Dr. SYED IRTIZA ALI SHAH
Principal & Dean
SINES • NUST, Sector H-12
Islamabad

Date: 10 JUL 2024

Certificate for Plagiarism

It is certified that MS Thesis Titled Integrative scRNA-seq Analysis Illuminates LUAD Proqression: Contrasting Primary and Metastatic Landscapes with TCGA Insight by Zia Ullah has been examined by us. We undertake the follows:

a. Thesis has significant new work/knowledge as compared already published or are under consideration to be published elsewhere. No sentence, equation, diagram, table, paragraph or section has been copied verbatim from previous work unless it is placed under quotation marks and duly referenced.

b. The work presented is original and own work of the author (i.e. there is no plagiarism). No ideas, processes, results or words of others have been presented as Author own work.

c. There is no fabrication of data or results which have been compiled / analyzed.

d. There is no falsification by manipulating research materials, equipment or processes, or changing or omitting data or results such that the research is not accurately represented in the research record.

e. The thesis has been checked using TURNITIN (copy of originality report attached) and found within limits as per HEC plagiarism Policy and instructions issued from time to time.

Name of Supervisor: Dr Mehak Rafiq

Signature & Stamp of Supervisor:

DR. MEHAK RAFIQ
Assistant Professor
SINES, National University
of Science & Technology
H-12 Islamabad

AUTHOR'S DECLARATION

I, Zia Ullah, hereby state that my MS thesis titled "Integrative scRNA-seq Analysis Illuminates LUAD Progression: Contrasting Primary and Metastatic Landscapes with TCGA Insights" is my own work and has not been submitted previously by me for taking any degree from National University of Sciences and Technology, Islamabad or anywhere else in the country/ world.

At any time, if my statement is found to be incorrect even after I graduate, the university has the right to withdraw my MS degree.

Name of Student: Zia Ullah

Date: ____________________

DEDICATION

Foremost to Almighty Allah for giving me the willpower and strength to complete my dissertation and to my family for their endless love, support, and encouragement throughout my pursuit of education. I hope this achievement will fulfil the dream they envisioned for me.

ACKNOWLEDGEMENT

First, I want to thank and praise Almighty Allah, who has enabled me to understand, think, and complete this case report work in His great mercy and benevolence. My heartiest praise goes to the holiest man in the whole galaxy, Almighty's beloved, Hazrat Muhammad (Peace Be upon Him), humanity's reformer.

I want to express my sincere gratitude and thanks to my supervisor, Dr Mehak Rafiq, for showing me support and guidance for my Post Grad study; she helped me through all the time of study and writing of this report; without Dr Mehak's constant support, and mentorship this would not be possible. I also want to express gratitude towards my GEC members, Dr Zamir Hussain and Dr Nosheen Fatima Rana, for their utmost support and guidance throughout the project.

I want to extend a special express gratitude to the Data Analytics lab members Madhaiha Bint e Masood, Syed M. Ameer Ghaznavi, Zunaira, Faraz Sheikh, and Amman Safeer for their support, discussions, compassion, and respect. Last but not least, I highly acknowledge Rizwan Ahmad for being with me throughout all the thick and thin of my research. The time I spent with them will always be cherished.

TABLE OF CONTENTS

LIST OF FIGURES

Page No.

LIST OF SYMBOLS, ABBREVIATIONS AND ACRONYMS

LUAD:	Lung Adenocarcinoma
mRNA:	messenger RNA
cDNA:	complementary DNA
NCBI:	National Center for Biotechnology Information
UMI:	Unique molecular identifier
NGS:	next-generation sequencing
scRNA-seq:	Single-cell RNA sequencing
PCA:	Principal component analysis
t-SNE:	t-Distributed Stochastic Neighbor Embedding
UMAP:	Uniform Manifold Approximation and Projection
FDR:	False Discovery Rate
DEG:	Differentially Expressed Gene
TSG:	Tumor suppressor gene
GO:	Gene Ontology
GSEA:	Gene Set Enrichment Analysis
QC:	Quality control
GEO:	Gene Expression Omnibus
VST:	Variance Stabilising Transformation
SNN:	Shared Nearest Neighbor
TCGA	The Cancer Genome Atlas
ADGRL2	ADhesion G-protein-coupled Receptor L2
ERO1α	Endoplasmic Reticulum Oxidoreductase 1 alpha
EMT	Epithelial-Mesenchymal Transition
HCC	Hepatocellular Carcinoma
CTLs	Cytotoxic T Lymphocytes
CTCs	Circulating Tumour Cells

ABSTRACT

Lung adenocarcinoma (LUAD) is the most prevalent form of lung cancer and is a significant contributor to cancer-related deaths globally. This cancer type displays extensive diversity and complex characteristics at molecular and cellular levels. Furthermore, LUAD frequently spreads to remote organs, including lymph nodes and the brain, complicating diagnosis, prognosis, and treatment processes. Its progression is a complex and multifaceted process that involves dynamic changes in gene expression patterns, cell identity, and the activation of critical pathways. This comprehensive study utilised single-cell RNA sequencing (scRNA-seq) technology to unveil the complex cellular heterogeneity within primary lung adenocarcinomas and their metastatic sites. This meticulous analysis led to several key findings. The differential gene expression analysis results revealed a striking diversity in gene expression patterns among various cell types within the primary tumour microenvironment. This heterogeneity underscores the specialised roles of distinct cell types in supporting or contributing to tumour growth and maintenance. Conversely, a consistent stemness signature emerged in metastatic samples, suggesting a potential activation of the epithelial-to-mesenchymal transition (EMT), a critical step in cancer metastasis. Our findings shed light on the dynamic changes in gene expression profiles during cancer progression. Gene set enrichment analysis highlighted the significance of several biological pathways in cancer initiation and progression. In primary tumours, the Epithelial-Mesenchymal Transition (EMT) pathway emerged as a central player, emphasising its role in cancer initiation. Other pathways, such as Androgen Response, Adipogenesis, and Coagulation, were also identified, potentially contributing to later-stage tumour growth. Pathways like UV Response Dn and Apical Junction appeared to act as safeguards against uncontrolled cell growth. The p53 Pathway has somehow complex role depending on the specific environment. In metastatic samples, pathways associated with EMT, TNF-alpha Signalling, and IL-2/STAT5 Signalling exhibited high significance, reflecting their importance in promoting tumour growth, metastasis, and immune evasion. Additionally, pathways linked to hypoxia and apoptosis were also found to be actively involved in metastatic processes. By comparing our results with The Cancer Genome Atlas (TCGA) data, we identified two previously unreported genes, ADGRL2 and

ERO1A, with potential roles in cancer progression. ADGRL2 emerged as a stem-cell-specific negative regulator, while ERO1A was associated with poor prognosis in various cancer types and linked to metastasis and the epithelial-to-mesenchymal transition (EMT). Our study offers valuable insights into the intricate processes of cancer progression, emphasising the role of EMT and identifying potential therapeutic targets. It provides a broader context for understanding cancer biology and paves the way for personalised cancer treatments. While acknowledging the study's limitations, including sample size and the need for experimental validation, this research sets the stage for future investigations to explore the identified genes and pathways further, potentially revolutionising cancer diagnosis and treatment.

CHAPTER 1: INTRODUCTION

1.1 Cancer

Cancer is a diverse and complex disease characterised by uncontrolled cell growth of abnormal cells. It can originate in any part of the body where abnormal cells multiply, forming tumours and invading surrounding tissues[1]. This disease can occur in various body parts and may spread to other organs through metastasis. It constitutes a substantial global public health concern, occupying second in global mortality statistics, closely behind cardiovascular diseases[2]. According to the statistics provided by the World Health Organisation (WHO), Cancer is responsible for nearly 10 million deaths globally, with lung cancer being the most common in men and second most common in women[3]. Cancer's lethality depends on its type and stage, making early detection vital for survival. It arises from a mix of genetic, environmental, and lifestyle factors. Common risks include tobacco and alcohol use, an unhealthy diet, lack of exercise, radiation exposure, chemicals, and genetic predisposition. Comprehensive prevention involves reducing these risks, adopting healthy habits, and having regular screenings to catch cancer early and improve survival chances[4, 5].

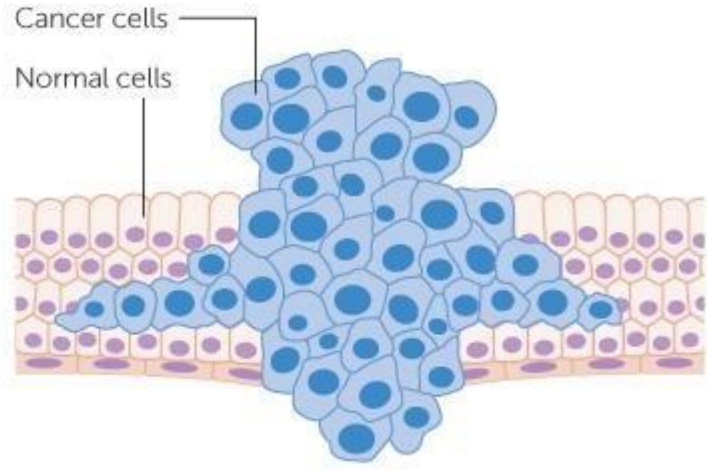

Figure 1.1: Comparison between normal cells and cancer cells within a tissue showing invasive nature of cancer cells.

1.1.1 Lung Cancer

Lung cancer is a highly lethal disease. It ranks as the leading cause of cancer-related deaths. According to data from the World Health Organisation (WHO), in the year 2020, there were approximately 2.21 million newly reported lung cancer cases, along with 1.80 million fatalities[6]. From a histological perspective, lung carcinoma can be divided into non-small cell lung carcinoma (NSCLC), small cell lung carcinoma (SCLC), mesothelioma, sarcoma, and carcinoid. SCLC and NSCLC are the most prevalent, accounting for approximately 90% of all lung cancer cases, while other types are relatively infrequent[7, 8]. Non-small cell lung carcinoma (NSCLC) can be categorised into four distinct types: Adenocarcinoma, squamous cell carcinoma, large-cell undifferentiated carcinoma, and Pancoast tumour [8, 9]. The development of lung cancer is primarily attributed to smoking habits, genetic predisposition, urban living, and environmental influences, including arsenic, toxins, and asbestos exposure[10]. Within the Non-small cell lung carcinoma (NSCLC), Adenocarcinoma has now emerged as the most prevalent histological subtype of primary lung cancer, constituting over 40% of cases, and its frequency continues to grow[11].

1.1.1.1 Lung Adenocarcinoma

Lung Adenocarcinoma (LUAD), a subtype of non-small cell lung cancer, originates from glandular cells in the smaller airways on the outer periphery of the lungs. It is the most common histological type, encompassing roughly 40% of all lung cancer incidences. LUAD is frequently identified at an advanced stage, often accompanied by distant metastasis, which presents considerable hurdles in clinical treatment. Although this form of lung cancer is more common among individuals who currently smoke or have smoked in the past, it is also the predominant type of lung cancer seen in non-smokers[12, 13]. Hence, gaining insight into the molecular and cellular mechanisms that drive the progression and spread of LUAD is of utmost importance in creating innovative approaches for diagnosis and treatment. The prognosis for lung cancer is heavily influenced by the stage at diagnosis. On average, the current 5-year survival rate is around 18%. Early detection can substantially enhance the outcome, leading to a 5-year survival rate of 54% for cases detected at the localised stage[12]. Unfortunately, only 15% of all incidences are detected early, whereas the overwhelming majority (57%) are diagnosed in advanced stages. Consequently, it is crucial to conduct screenings for lung cancer in individuals at high risk[14]. The various stages of primary tumours in lung adenocarcinoma are categorised according to the tumour's size, its extent of infiltration into neighbouring tissues, and regional or distant metastases[15].

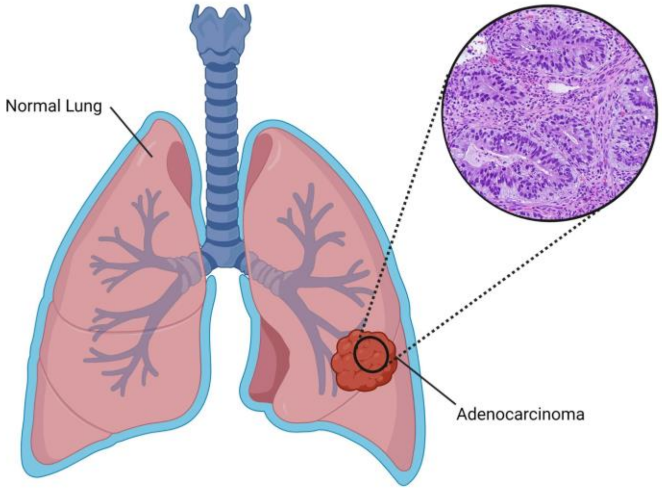

Figure 1.2: Histology of human lungs highlighting the growth of adenocarcinoma.

The different stages of primary tumours in lung adenocarcinoma are typically defined by the TNM staging system, which takes into account the size of the tumour (T), the extent of spread to nearby lymph nodes (N), and the presence of distant metastases (M)[16].

- Stage I: At this stage of lung adenocarcinoma, the tumour is small, with no invasion of nearby tissues or lymph nodes. Typically, its diameter is less than 3 cm and has yet to extend to distant locations.
- Stage II: In stage II lung adenocarcinoma, the tumour is more prominent and may have reached nearby lymph nodes, but it has not metastasized to distant sites. The tumour's diameter can range from 3 to 7 cm.
- Stage III: At stage III of lung adenocarcinoma, the tumour has typically invaded nearby tissues and lymph nodes but has not yet spread to distant sites. This stage has two subcategories: IIIA, where the tumour has extended to nearby lymph nodes and the chest wall, and IIIB, where the tumour has reached the heart's lining or other organs within the chest cavity.

- Stage IV: In stage IV lung adenocarcinoma, the tumour has metastasized to distant sites, including other organs or bones. It may also be present in multiple locations within the lung.

It is crucial to emphasise that various stages of lung adenocarcinoma exhibit unique molecular and cellular attributes, which significantly influence treatment choices and patient results. For instance, early-stage tumours are frequently more amenable to surgical or radiation treatments, whereas advanced-stage tumours might necessitate chemotherapy or targeted interventions. Hence, precise staging of lung adenocarcinoma plays a vital role in determining the most suitable treatment strategy tailored to each patient's needs[13]. Primary tumour sites frequently exhibit multiple metastatic markers in lung adenocarcinoma. These markers refer to molecules or genes present in the primary tumour tissue and are linked to the probability of the tumour spreading to distant areas within the body[17]. Numerous metastatic markers have been discovered and examined in lung adenocarcinoma to assess their potential in predicting disease advancement and gauging treatment effectiveness.

1.2 Single-cell RNA-seq Analysis

Bulk RNA sequencing (RNA-seq) has become a valuable tool for understanding genetic phenomena and biological processes across various animal and plant species. Bulk RNA sequencing (bulk RNA-seq) has played a pivotal role in cancer research by providing insights into gene expression patterns and genomic alterations within tumour samples. Nonetheless, it comes with inherent limitations. One significant constraint is its inability to resolve heterogeneity within cell populations, as it averages gene expression profiles across a mix of cells. It can obscure rare cell subpopulations that may be vital in understanding treatment resistance or specific disease mechanisms. Moreover, bulk RNA-seq struggles to discriminate between distinct cell types within a sample, making it challenging to attribute specific gene expression patterns to individual cell populations. Additionally, it may not efficiently capture alternative splicing, point mutations, novel transcripts, long non-coding RNAs, and gene fusions, which are vital components of cancer biology. Lastly, bulk RNA-seq does not provide information at the single-cell level, limiting its ability to dissect the nuanced heterogeneity of gene expression within tumours, an essential aspect of understanding tumorigenesis and treatment response. These limitations have driven the development of single-cell RNA sequencing (scRNA-seq) and other advanced technologies to address these challenges and provide a more comprehensive view of cancer biology[18]. Unlike traditional bulk RNA sequencing, which averages gene expression across all cells in a tissue sample, scRNA-seq offers a revolutionary approach. It enables the detection of rare cell populations, reveals cell-to-cell differences in gene expression, and uncovers novel cell types and molecular pathways that often remain hidden in bulk sequencing data[19]. scRNA-seq has emerged as a powerful tool for unravelling the complexities of cell biology[20]. It allows researchers to delve into the transcriptomes of individual cells, enabling a comprehensive exploration of gene expression patterns. This innovation has paved the way for a deeper understanding of the masked diversity within cell populations, shedding new light on cellular heterogeneity[21]. By applying scRNA-seq to the study of different cancers, researchers have unveiled previously obscured layers of cellular diversity within tumours, offering unprecedented insights into the complexity of these cancers, including lung adenocarcinoma[22]. This technique provides unprecedented resolution in understanding tumour heterogeneity and dynamics within the microenvironment. In summary, scRNA-seq is a potent tool for exploring the intricate cellular diversity within tumour samples at a deeper level. In turn, it yields fresh insights into the molecular and cellular mechanisms driving cancer and informs the development of more efficacious treatment strategies. This thesis used scRNA-seq to examine LUAD cells from primary and metastatic sites, revealing distinct cellular and transcriptional modules associated with lung cancer survival.

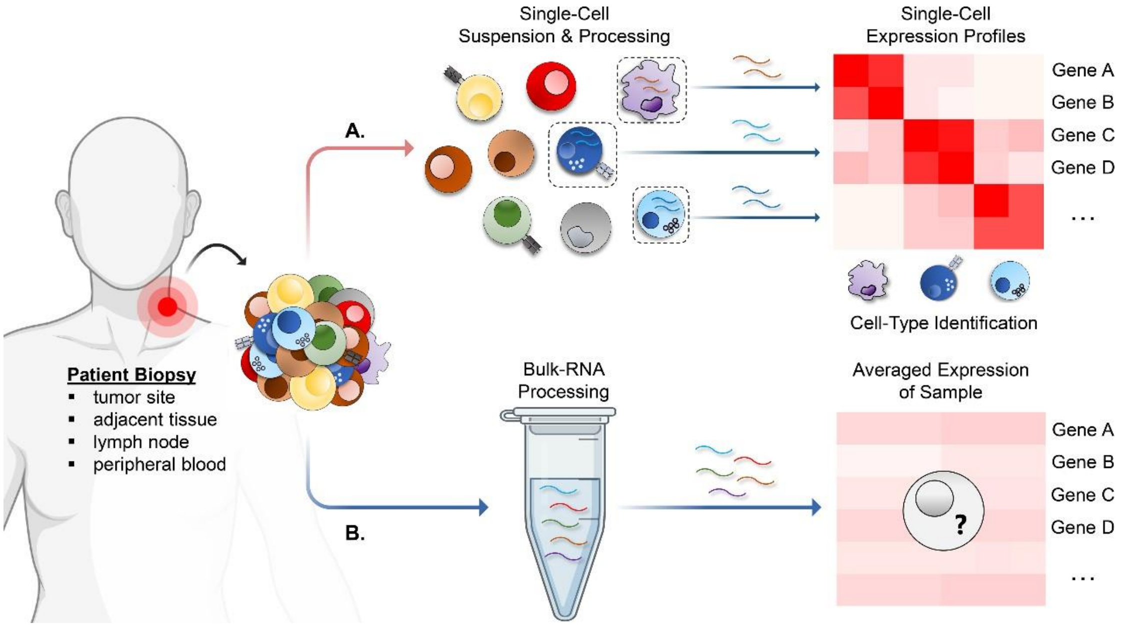

Figure 1.3: Single-Cell RNA-seq vs Bulk RNA-seq

1.3 Problem Statement

Despite substantial advancements in cancer research, lung cancer continues to exhibit a low 5-year survival rate, primarily attributed to its complex and poorly understood heterogeneity at the molecular level. The variability in genetic mutations, cellular phenotypes, and microenvironmental interactions within lung tumors presents a challenging obstacle to effective diagnosis and treatment. Thus, there is a critical need to comprehensively explore and delineate this heterogeneity using advanced methodologies, such as single-cell RNA analysis, to uncover novel molecular signatures that could inform targeted therapeutic strategies and improve patient outcomes.

1.4 Thesis Objectives

The main objectives and contributions of this thesis are:

- The primary aim of this study is to conduct a comparative analysis of Differentially Expressed Genes (DEGs) identified through single-cell RNA sequencing (scRNA-seq) between primary cancer stages and their corresponding metastatic stages. By contrasting the gene expression profiles in these conditions, we seek to elucidate the molecular distinctions underlying cancer progression and metastasis.

- Another key objective is to understand better the functional relevance of the DEGs identified through scRNA-seq. It will be achieved by employing Gene Set Enrichment Analysis (GSEA) to uncover the enrichment of these DEGs in specific biological pathways and molecular processes. Through GSEA, we aim to elucidate the broader biological context in which these genes operate and their potential roles in driving cancer progression.

- The final objective of this study is to establish a meaningful connection between our findings and existing knowledge in the field. The identified DEGs will be compared with those documented in the extensive dataset from The Cancer Genome Atlas (TCGA) to achieve this. This comparative analysis will help validate our results and provide insights into the consistency and relevance of our scRNA-seq findings in the context of a larger-scale genomic database.

1.5 Thesis Outline

The outline of this thesis is as follows: Chapter 2 provides a literature review of the current state-of-the-art LUAD research, focusing on the molecular and cellular mechanisms of LUAD progression and metastasis and the applications of scRNA-seq in lung cancer. Chapter 3 describes the materials and methods used in this study. Chapter 4 presents the results and main findings of scRNA-seq analysis of LUAD cells from primary and metastatic sites. Chapter 5 consists of a discussion, and Chapter 6 concludes the thesis and provides suggestions for future work.

CHAPTER 2: REVIEW OF LITERATURE

This chapter aims to provide an overview of recent research employing RNA sequencing studies, both bulk and single cell, to visualise the genetic makeup of LUAD cells and their surrounding environment at various stages and sites of metastasis. scRNA-seq is a potent tool that can unveil the diverse characteristics and dynamics of cell populations with exceptional detail, leading to the identification of new cell subtypes and potential biomarkers. Below is an overview of a synopsis of the fundamental discoveries and their significance in advancing LUAD research and treatment. Additionally, we will underscore the existing gaps in knowledge and the challenges that warrant further investigation in this domain.

2.1 Overview of LUAD

LUAD is the most prevalent histological subtype of non-small cell lung cancer (NSCLC), arising from the epithelial cells that line the bronchioles and alveoli in the lung. LUAD constitutes approximately 40% of all lung cancer cases, making it the most common type within the NSCLC category[16]. It is frequently linked to smoking but can also develop in non-smokers, particularly in women and younger individuals. LUAD exhibits significant heterogeneity within and between tumours, reflecting its diverse cellular origins, molecular changes, and influences from the surrounding microenvironment. LUAD can be categorised into various subtypes based on its histological, molecular, and clinical characteristics, each with important implications for its diagnosis and treatment[13].

In most cases, spontaneous tumours begin with a single cell. However, when clinically diagnosed, many human tumours exhibit significant heterogeneity in various aspects, including cell surface receptor expression, growth potential, and angiogenesis. This diversity can be partly explained by morphological and epigenetic adaptability. Nevertheless, there is compelling evidence to suggest the presence of genetically distinct tumour cell populations coexisting within these tumours[23]. Due to this property of cancer, diagnosing lung cancer is challenging. This heterogeneity exists at cellular, histological, and molecular levels. It affects diagnosis, understanding of pre-neoplastic lesions, and identifying cell origins. Molecular differences within and between tumours and temporal changes further complicate the picture. Tumor heterogeneity has significant implications for understanding pathogenesis, making accurate diagnoses, selecting tissues for molecular testing, and deciding on treatments. Recognising and addressing this heterogeneity is essential for future advancements in lung cancer management[24].

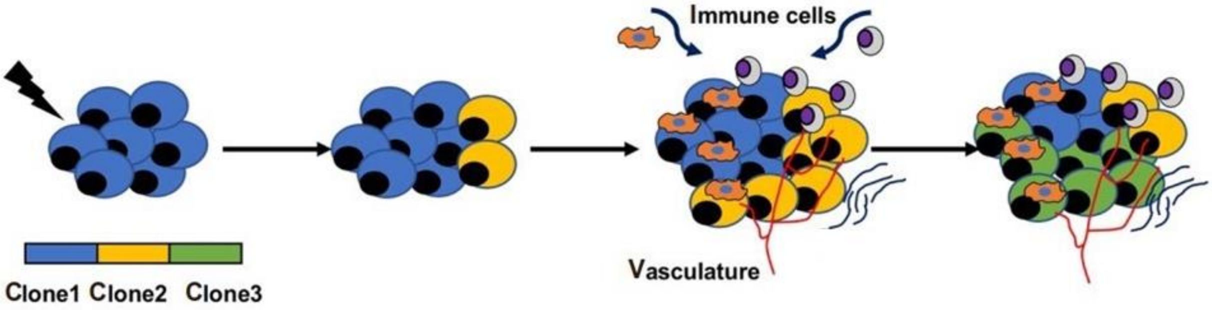

Figure 2.1: Progression of tumour leading to tumour heterogeneity.

Over the past few decades, the transcriptome study has become a standard practice in unravelling the biological processes underlying normal physiology and pathological mechanisms. This interest has driven the rapid advancement of sequencing technologies, progressing from whole-tissue analysis to cell population studies and, most recently, to the single-cell level. While these advancements have significantly expanded our understanding of molecular signalling, intercellular communication networks, and the identification of rare cell sub-populations, it is crucial to recognise that there is no one-size-fits-all approach to RNA sequencing[25]. Researchers must meticulously design their studies, considering sample availability, preparation methods, platform advantages and limitations, and specific sequencing attributes.

Lung adenocarcinoma (LUAD) is the prevailing lung cancer type and frequently exhibits distant metastasis, notably to the brain and lymph nodes. Metastasis is the primary driver of mortality in individuals diagnosed with LUAD, presenting significant clinical complexities and challenges[26]. Understanding the intricate molecular and cellular processes driving the progression and spread of LUAD (Lung Adenocarcinoma) is pivotal for developing innovative diagnostic and therapeutic approaches.

2.2 Bulk RNA-seq Studies on LUAD

The scientists utilised RNA-seq data for analysis to present an RNA-seq prognostic signature for lung adenocarcinoma, identifying a four-gene signature, including a lncRNA gene (LINC00941), with prognostic solid association. The signature was validated in multiple patient cohorts, demonstrating its potential for individualising therapy decisions in early-stage lung adenocarcinoma and risk stratification in advanced cases, particularly for EGFR mutant patients. These findings offered a promising avenue for improving clinical management and tailoring treatment strategies for this challenging cancer[27].

2.2.1 Impact of Smoking on Gene Expression

Bulk RNA sequencing was used again to investigate gene expression variations in lung adenocarcinoma patients, with a specific focus on distinguishing between smokers and nonsmokers. The findings revealed that tobacco smoking substantially influenced the RNA expression profiles in tumour tissues, exacerbating the dysregulation of gene expression, particularly the downregulation of genes. Notably, specific genes, such as Secreted Phosphoprotein 1(SPP1) and Family with Sequence Similarity 83 Member D (FAM83D), consistently exhibited overexpression in smokers and nonsmokers, suggesting their potential utility as biomarkers for early lung cancer detection and therapeutic targeting. The study shows the significance of RNA sequencing studies on lung adenocarcinoma. It proposes potential RNA expression signatures associated with smokers and nonsmokers while recognising the need for further validation and exploration in diverse patient populations[28].

2.2.2 Stage-Specific Biomarkers

Another study performed on LUAD aimed to identify stage-specific biomarkers for distinguishing different stages of progression using bulk RNA-seq data, which provides an average expression profile of tissue samples. The research revealed a set of genes significantly correlated with LUAD stages, including ANGPTL5, C7orf16, EDN3, LOC150622, HOXA11AS, IL1F5, and USH1G, which were identified as potential biomarkers for distinguishing stage III from stages I and II. Additionally, GJB6 was implicated in the gap junction trafficking pathway, while C7orf16 and EDN3 were associated with the Wnt signalling pathway, cell cycle, and G protein-coupled receptor (GPCR) signalling. These findings offer insights into the molecular mechanisms underlying LUAD progression based on bulk RNA-seq data, providing potential targets for further investigation and therapy[29].

2.3 scRNA Seq Studies

Single-cell RNA sequencing (scRNA-seq) is a potent method capable of unveiling the intricacies and fluctuations within cell populations with exceptional precision. It identifies novel cell subtypes and potential biomarkers and offers profound insights into research and treatment for lung adenocarcinoma (LUAD). However, it is essential to emphasise that unexplored areas and unresolved challenges within this field still demand attention. In a particular study, researchers devised a seven-gene signature linked to metastasis, aiming to forecast the prognosis of patients with lung adenocarcinoma (LUAD) using scRNA-seq data from LUAD tissues. They pinpointed 14 metastasis-associated genes (MAGs) that displayed differential expression patterns between primary and metastatic LUAD cells. The validity of this signature was then confirmed across various independent datasets. Additionally, the study demonstrated a correlation between this signature and immune cell infiltration in LUAD[30].

A study conducted scRNA-seq on a substantial dataset comprising 208,506 individual cells extracted from 44 patients diagnosed with either primary or metastatic Lung Adenocarcinoma (LUAD). The investigation unveiled the existence of a distinct cancer cell subtype that exhibited a marked deviation from the standard differentiation trajectory, and notably, this subtype predominated during the metastatic phase of the disease. Furthermore, the study shed light on significant ontological and functional changes occurring within the stromal and immune cell populations, collectively creating a microenvironment conducive to tumour growth and immune suppression in metastatic LUAD[31]. Researchers also utilised scRNA-seq to generate comprehensive molecular profiles from five patients diagnosed with ground glass nodule (GGN) adenocarcinoma, an early-stage variant of Lung Adenocarcinoma (LUAD). They compared them with profiles from five patients with solid adenocarcinoma (SADC), a more advanced form of the disease. Their investigation revealed distinct molecular characteristics between these two subtypes. Specifically, the cancer cells in GGN-ADC exhibited downregulation of signalling pathways associated with cell proliferation, suggesting a potentially less aggressive phenotype than SADC. Meanwhile, the stromal cells displayed differential effects on processes such as angiogenesis, fibrosis, and immunity in GGN-ADC compared to SADC. These findings highlight the intricate interplay between cancer cells and the tumour microenvironment in different LUAD subtypes and may provide valuable insights into their progression and potential treatment strategies[32].

2.3.1 Integrating scRNA-seq with Genomic Data

scRNA-seq was conducted on 1,368 LUAD cells sourced from 10 patients. They subsequently integrated this data with bulk whole-genome sequencing and RNA sequencing to understand the molecular landscape better. Through this integrated approach, the team successfully delineated the functional repercussions of co-occurring genomic alterations in LUAD cells. Moreover, their findings illuminated the remarkable heterogeneity in the expression of driver genes and the activity of pertinent pathways within individual tumours and across the cohort of patients. This newfound insight into the intricate genomic and functional diversity within LUAD tumours holds significant promise for advancing our comprehension of disease progression and tailoring more effective therapeutic strategies[33].

2.3.2 Reference Component Analysis

Researchers utilised reference component analysis to examine scRNA-seq data obtained from 23 colorectal tumours. This analysis revealed the presence of six major cell types and 11 distinct subtypes within the tumour samples. Furthermore, comparable cell types and subtypes were identified when applying this method to scRNA-seq data from LUAD. This study effectively showcased the utility of reference component analysis in unravelling the intricate cellular diversity and tumour microenvironment within human malignancies[32].

2.3.3 Therapy-Induced Evolution

In a different study, scientists used 17,648 cells from 20 treatment-naive and 10 post-treatment LUAD tumours to perform scRNA-seq. They demonstrated how LUAD cells and their microenvironment underwent therapy-induced evolution following treatment with tyrosine kinase inhibitors or immune checkpoint inhibitors. In LUAD, they discovered unexpected cell types linked to immune evasion and treatment resistance[34].

2.4 scRNA-seq in Metastatic LUAD

Metastatic LUAD has a poor prognosis and few treatment options. Several studies have used single-cell RNA sequencing (scRNA-seq) to study the transcriptomic diversity, dynamics of LUAD cells, and their microenvironment throughout different stages and locales of metastasis to obtain insights into the molecular and cellular causes of LUAD metastasis. scRNA-seq effectively captures individual cellular gene expression levels and identifies novel cell subtypes and biomarkers[16]. This article offers an overview of several studies, delving into their findings and significance in research and treatment for Lung Adenocarcinoma (LUAD). Below, are the some of these related studies in detail.

scRNA-seq analysis was performed on 60,459 cells from 10 patients with ground glass nodule (GGN) adenocarcinoma, an early variant of LUAD, and 10 patients with solid adenocarcinoma (SADC), a more advanced variety, in one of the investigations. They discovered that GGN-ADC cancer cells showed downregulated cell growth signalling pathways, but stromal cells had different effects on angiogenesis, fibrosis, and immunity in GGN-ADC and SADC. They also discovered a new subtype of cancer-associated fibroblasts (CAFs) in GGN-ADC that expressed high levels of C-X-C motif chemokine 12 (CXCL12) and platelet derived growth factor receptor beta (PDGFRB)[32]. scRNA-seq was also conducted on 1,368 LUAD cells obtained from 10 different patients. They combined this scRNA-seq data with bulk whole-genome sequencing and RNA sequencing data to explore the functional consequences of co-occurring genomic alterations in LUAD cells. Their analysis uncovered notable variations in the expression of driver genes and pathway activities within individuals and across different tumours. Specifically, they observed that KRAS mutations were linked to increased activity in glycolysis and oxidative phosphorylation pathways, whereas EGFR mutations were associated with reduced cell cycle pathway activity. Additionally, the researchers identified a previously unrecognised subgroup of LUAD cells characterized by high expression levels of AXL receptor tyrosine kinase (AXL) and Growth Arrest Specific 6 (GAS6)[26]. scRNA-seq on 17,648 cells from 20 treatment-naive and 10 post-treatment LUAD tumours was utilised to demonstrate the evolution of LUAD cells and their microenvironment following treatment with tyrosine kinase inhibitors or immune checkpoint inhibitors. They discovered new cell states in LUAD that are linked to medication resistance and immune evasion. They discovered that after tyrosine kinase inhibitor treatment, EGFR-mutant LUAD cells gained mesenchymal epithelial transition factor (MET) amplification or human epidermal growth factor receptor 2 (HER2) activation, whereas PD-L1-positive LUAD cells acquired TGF-beta or WNT signalling [35]. Spatial distribution and functional states of immune cells in LUAD tissues was observed using scRNA-seq on 20,000 cells from 20 LUAD patients. They discovered that immune cells showed distinct spatial patterns depending on their cell type and activation state. They also discovered a new subtype of cytotoxic T cells with impressive anti-tumor activity and high granzyme B (GZMB), Perforin-1 (PRF1), interferon gamma (IFNG), and tumour necrosis factor (TNF) levels[36]. A study found that IFN-signaling pathway genes are heterogeneously expressed and correlated with other genes in single cancer cells, including MHC class II (MHCII). Downregulation of genes in IFN-signaling pathways in cell lines corresponds to an acquired resistance phenotype. This study was performed through scRNA-seq analysis[22]. scRNA-seq also identified a new population of tumour-associated neutrophils (TANs) that aid in the spread of lung cancer. The researchers discovered that TANs increase lung cancer spread by secreting IL-8. In the study, galectin-1 was also identified as a possible therapeutic target for non-small cell lung cancer. Galectin-1 is a protein in many malignancies linked to a bad prognosis[37]. Another study, published in PubMed, discovered that single-cell RNA profiling of lung adenocarcinoma gives new prognostic information based on the microenvironment and may assist in predicting therapy response and disclose potential target cell types for future therapeutic approaches[27].

2.5 scRNA-seq in Breast Cancer

Apart from studies on Luad, scRNA-seq analysis was also used for many other cancer types, resulting in some novel findings. Such a type of study was done on breast cancer. The study used single-cell transcriptomics to examine HER2+ breast cancer cells and their response to trastuzumab treatment. They successfully distinguished treated and untreated cells, attributing the separation to trastuzumab. Known and new expression patterns were identified. Some genes upregulated under trastuzumab treatment, like Clusterin (CLU) and SELENOP selenoprotein P (SEPP1), may be markers for cardiotoxicity. Trastuzumab's impact on tumor-promoting genes was noted. Upregulated MGP may indicate better outcomes. Additionally, a 48-gene signature related to cardiomyocyte cell death was identified, shedding light on cardiotoxicity mechanisms. The researchers also identified new genes previously not shown to be upregulated in previous studies, including C-X-C Motif Chemokine Ligand 1 (CXCL1), C-X-C Motif Chemokine Ligand 8 (CXCL8), Calcium and Integrin Binding 1 (CIB1), Interferon Induced Transmembrane Protein 1 (IFITM1), and Secreted Phosphoprotein 1 (SPP1). This research contributes to a better understanding of trastuzumab therapy's molecular mechanisms and the usage of scRNA-seq technology[38].

The key findings presented in this thorough literature review illustrate the growing importance of single-cell RNA sequencing in figuring out the complex cellular landscape of primary lung adenocarcinomas and their metastatic counterparts. This refined technique has enabled researchers to identify distinct cellular subpopulations inside lung adenocarcinomas, identify the underlying molecular pathways driving tumour initiation, and gain critical insights into metastatic processes. Furthermore, single-cell RNA sequencing has revealed attractive treatment targets for lung cancer and set the groundwork for pursuing personalised medicine approaches. However, despite these remarkable strides, substantial knowledge gaps remain within this domain that necessitate further exploration. For instance, a deeper comprehension of the heterogeneity inherent in the tumour microenvironment is imperative, as is the identification of novel biomarkers that can serve as predictive indicators for enhanced treatment responses in lung adenocarcinomas. Another compelling avenue for future research could be the development of more precise and effective therapies for lung adenocarcinoma, specifically targeting distinct cellular subsets within the tumour microenvironment. Efforts to bridge these knowledge gaps are pivotal in advancing our understanding of lung adenocarcinoma and enhancing patient outcomes. The findings summarised in this literature review underscore the immense potential of single-cell RNA sequencing in cancer research, hinting at its transformative capacity to redefine the diagnosis and treatment of cancer.

CHAPTER 3: METHODS

This chapter details the methodology used to investigate the cellular atlas of primary lung adenocarcinomas and metastatic lesions using single-cell RNA sequencing (scRNA-seq). It includes data processing steps such as quality control, gene expression normalisation, PCA, and cell clustering using Seurat v4.0. Additionally, differential gene expression analysis and functional enrichment were performed. Furthermore, a comparison between DEGs identified through scRNA-seq and TCGA data was carried out to identify unique genes not present in the TCGA database.

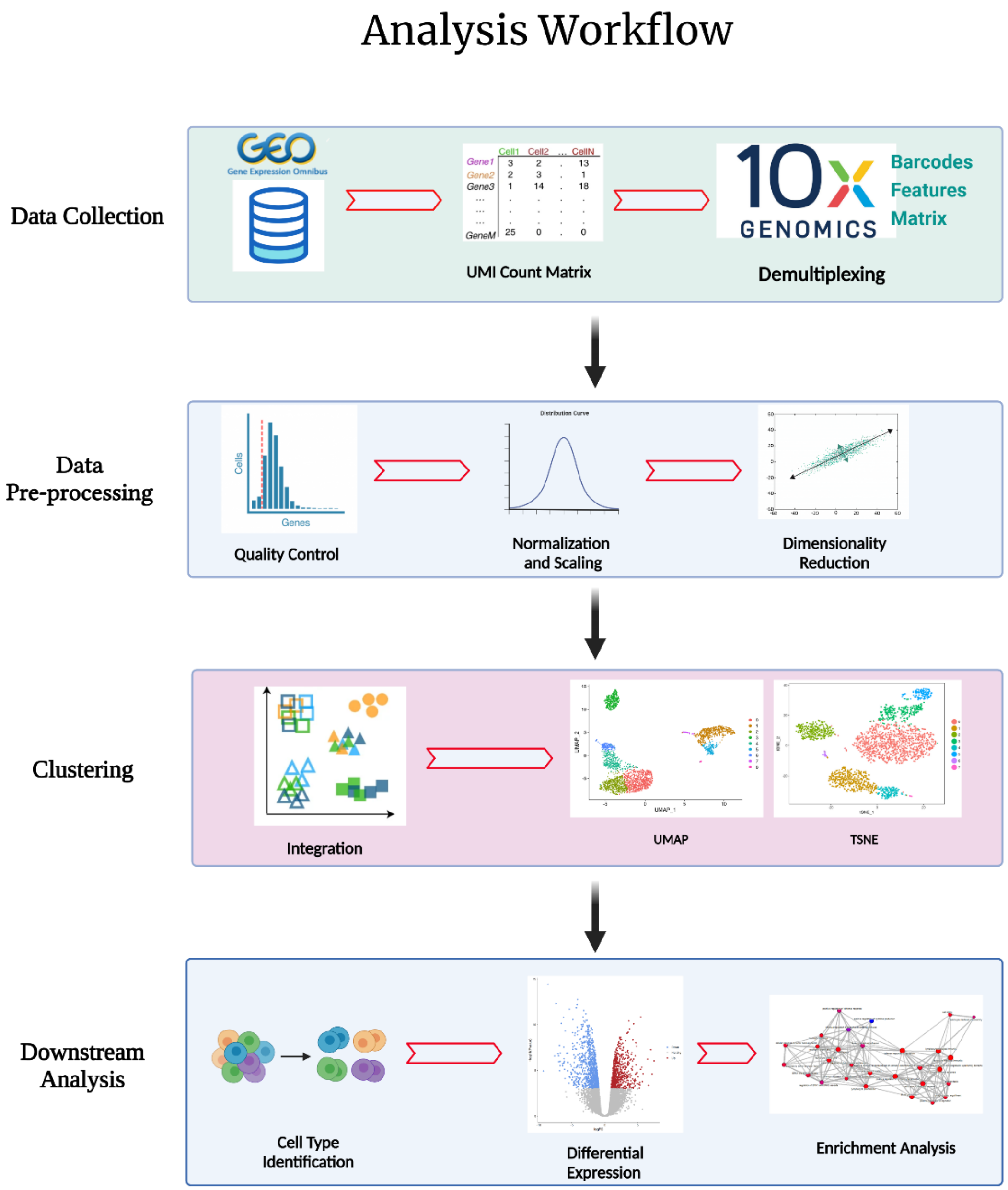

Figure 3.1: An overview of the methodology implemented in this study

3.1 Data Collection

The scRNA-seq dataset investigated in this study was sourced from the NCBI Gene Expression Omnibus (GEO) database under the accession number GSE131907. This dataset was generated by high-throughput sequencing of 44 lung adenocarcinoma patients, producing 58 samples. It encompasses scRNA-seq data from 208,506 cells originating from 15 primary lung adenocarcinoma samples, seven lymph node metastases, and 10 brain metastases. Furthermore, the dataset includes five pleural effusion samples, ten normal lymph node samples, and 11 distant standard samples. For this research, a subset of 11 primary tumour samples and 11 metastatic tumour samples were selected, with the metastatic samples consisting of 5 lymph node samples and 6 brain samples. The original dataset was provided as a .rds.gz count matrix, as raw sequencing data was withheld due to privacy concerns for the patients involved[31].

3.2 Tools

To begin the analysis of the scRNA-seq data using the Seurat pipeline in R, the initial step involved converting the .rds file into the 10X format, which includes Barcodes, Features, and Count Matrices. This analysis was conducted in R, a programming language, with the assistance of R-Studio, an integrated development environment (IDE) for R. The Seurat pipeline, designed explicitly for single-cell data analysis, was then employed in addition to harmony to perform subsequent data processing and analysis tasks. The harmony package was used for the integration step in single-cell RNA-seq analysis. Enrich R, a web tool, was used for the pathway analysis, and the TCGA data was accessed through The University of Alabama at Birmingham Cancer Data Analysis Portal (UALCAN)[39, 40].

3.2.1 R Language

R is a widely accepted programming language and software environment known for its statistical computing and graphics proficiency. It encompasses extensive statistical techniques, ranging from linear and nonlinear modelling to time-series analysis, classification, and clustering. This open-source, freely available software is compatible with Windows, Mac OS X, and Linux operating systems and boasts an active community of dedicated users and developers[41, 42].

3.2.1.1 R studio

R Studio is a robust and user-friendly integrated development environment (IDE) tailored explicitly for the R programming language. It boasts many tools and features that simplify data management, code writing, and visualization creation. Its intuitive code editor, with features like syntax highlighting and code completion, enhances code efficiency and accuracy[41, 42].

3.2.2 Seurat v4.0 and Harmony

Seurat is a prominent and widely used analysis pipeline for single-cell RNA sequencing (scRNA-seq) data, developed by Rahul Satija and his team at the New York Genome Center[43]. The Seurat pipeline encompasses several critical steps, including quality control, normalisation, clustering, and cell type identification, making it a standard tool for scRNA-seq data analysis.

3.2.3 Enrich R

A web tool called Enrich R was used for Gene set enrichment Analysis. Enrich R is a powerful and user-friendly web-based tool widely used in bioinformatics for pathway analysis. Its primary purpose is to help researchers uncover biological pathways and processes significantly associated with a list of genes of interest. By identifying which pathways or processes are overrepresented in the gene list, Enrich R aids researchers in gaining critical insights into the molecular mechanisms and functions of their genes[44-46].

3.3 Demultiplexing

The original dataset was in the form of a compressed ***.rds.gz*** count matrix, as privacy concerns prevented sharing raw sequencing data related to patients. The initial dataset underwent a demultiplexing process, creating 10X file formats, including Barcodes, Features, and Count Matrices. This transformation was carried out for each of the 11 selected samples. Following the data conversion into the 10X file format, the next step involved importing this processed data into R Studio, serving as the starting point for our analytical investigations.

3.4 Data Loading

Out of the total of 58 samples in the dataset, there are 15 primary tumour samples, seven from metastatic lymph nodes, 10 from metastatic brain lesions, five from pleural effusion, 10 from normal lymph nodes, and 11 from distant normal tissues. The primary focus for analysis was on the 11 primary tumour samples and their associated metastatic lesions, which included five samples from lymph nodes and six samples from brain metastases[31]. A 10X sparse matrix from the loaded data to effectively handle the complex nature of single-cell data. This matrix creation process was carried out using the R function ***Seurat::Read10X***().

3.5 Quality Control

After creating the Seurat object in the scRNA-seq analysis pipeline, a crucial step involves quality control to ensure the reliability of the dataset. This quality control process involves the application of criteria to retain only high-quality cells and genes. Specifically, cells with at least 100 expressed genes and genes expressed in at least three cells are retained. It was achieved through the ***Seurat::CreateSeuratObject()*** function[47].

Another crucial step involves calculating the percentages of mitochondrial and ribosomal genes in each cell using ***Seurat::PercentageFeatureSet()*** to ensure data integrity. High levels of mitochondrial or ribosomal gene content can distort analysis results and obscure the expression of nuclear genes. A high number of mitochondrial genes can be present due to damaged cells because when a cell's membrane is damaged, its genetic material can leak out. So, cells with more than 8% mitochondrial or ribosomal gene content are removed, enhancing the dataset's suitability for subsequent analyses accomplished through the subset() function[48].

In the final quality control step, cells with over 5000 features per cell are removed using ***subset(subset = nFeature_RNA < 5000)***. It indicates potential doublets that could distort the expression data, so it is necessary to remove them for accurate analysis.

3.6 Normalisation and Scaling

Normalisation is crucial in analyzing single-cell RNA sequencing (scRNA-seq) data to assess gene expression at the individual cell level. Since scRNA-seq experiments involve processing cells in batches, variations in sequencing depth and library size across cells and batches can introduce technical noise that hinders the accurate detection of biologically significant differences in gene expression[49]. Normalisation eliminates these technical variations and effectively aligns gene expression values to compare them across cells and batches [50].

In this analysis, a global-scaling normalisation method called ***"LogNormalise"*** was utilised. This method involves normalising the gene expression measurements for each cell based on the total expression, multiplying the result by a scale factor (typically 10,000 by default), and applying a logarithmic transformation. The resulting normalised values are then stored in the "RNA" assay within the ***'@assay'*** slot of the ***“seu”*** object. This normalisation step was executed using the R function ***"Seurat::NormaliseData()"***[51].

The next stage in our scRNA-seq data analysis involves identifying variable features. A method commonly employed for detecting variable features in scRNA-seq data analysis is the variance stabilizing transformation (VST) method. VST is a data normalisation technique designed to correct technical noise and variability in scRNA-seq data. This technique transforms read counts to approximate a normal distribution, enabling more precise statistical analysis and gene expression comparisons between individual cells[50].

Genes are initially ranked based on their mean expression levels to identify variable features using the VST method. Subsequently, the variance of each gene is computed, and genes are filtered according to their Coefficient of Variation (CV), which measures the relative variability in gene expression across cells. Genes exhibiting high CV values are considered variable and are retained for further analysis[47].

Subsequently, scaling, a standard preprocessing step, was performed before dimensional reduction techniques like Principal Component Analysis (PCA). It was achieved using the “***ScaleData()”*** function, which adjusts the expression of each gene so that the mean expression across cells becomes 0, and the gene expression variance across cells becomes 1. This transformation ensures that downstream analyses assign equal weight to each gene, preventing highly expressed genes from dominating the results. The scaled data is stored in ***“seu$RNA@scale.data”***, where the R function ***“Seurat::ScaleData()”*** was employed for this purpose.

Scaling in scRNA-seq data analysis offers several advantages. Firstly, it facilitates the detection of differentially expressed genes across individual cells or cell types by accounting for variations in sequencing depth and gene expression variability. Secondly, it aids in identifying rare cell types or subpopulations that might exhibit lower sequencing depth than more abundant cell types. Lastly, it allows for comparing gene expression levels across different scRNA-seq datasets or platforms, as scaling ensures that the data is consistently normalised.

3.7 Dimensionality Reduction

In single-cell RNA sequencing (scRNA-seq) data analysis, dimensionality reduction is a crucial step that simplifies the high-dimensional gene expression data into a lower-dimensional representation, making it more manageable and revealing key patterns. In the dimensionality reduction phase, Principal Component Analysis (PCA) was employed as the initial step, followed by t-distributed Stochastic Neighbor Embedding (t-SNE).

PCA identifies patterns in the data and generates orthogonal variables known as principal components (PCs), which capture the most significant data variation sources. PCA is essential for several reasons in scRNA-seq analysis. Firstly, scRNA-seq data is inherently high-dimensional, making visualization and analysis challenging. PCA reduces data dimensionality, facilitating visualization and downstream analysis. Secondly, it helps remove technical noise and batch effects, clarifying biological variability. Lastly, PCA can uncover biologically meaningful subpopulations with coordinated gene expression patterns[52]. To perform PCA, gene expression data is scaled and centred to account for differences in sequencing depth and gene expression variability across cells. The covariance matrix of the scaled data is then computed, and its eigenvectors and eigenvalues determine the principal components. The analysis uses ***"Seurat::RunPCA()"*** for this purpose[51].

Determining the optimal number of principal components to retain is crucial. The elbow plot technique, executed via ***Seurat::ElbowPlot()***, aids in this decision by plotting the variance explained by each PC against the number retained. It visually identifies the point where the increase in variance explained levels off, resembling an "elbow" shape[53]. The goal of the elbow plot is to strike a balance between capturing most of the data's variation and avoiding overfitting or including irrelevant components. Only some components may lose vital data variation, while too many can lead to overfitting, reducing result interpretability.

After PCA analysis, t-distributed Stochastic Neighbor Embedding (t-SNE) was introduced. Unlike PCA, t-SNE is a method that focuses on keeping nearby cells together, preserving their local interactions. After identifying dimensions with PCA, t-SNE captured detailed connections within the data and maintained local structures. The analysis was conducted using the R function

"Seurat::RunTSNE()" and selecting dimensions based on the elbow plot. This two-step approach offered a comprehensive understanding of overall and detailed patterns in the scRNA-seq data, contributing to a more detailed and interpretable analysis. t-SNE's non-linear dimensionality reduction is particularly beneficial in uncovering complex patterns between cells, making it a valuable tool in single-cell transcriptomics for exploring hidden patterns and facilitating the identification of distinct cell populations[54].

3.8 Integration

Integration in single-cell RNA sequencing (scRNA-seq) data analysis is a critical step that involves merging and aligning multiple scRNA-seq datasets from distinct samples or experiments into a unified dataset. This process is essential for enhancing statistical power, enabling cross-sample comparisons, and uncovering shared biological signals[50].

The harmony package integrates scRNA-seq data by identifying shared "anchors" between datasets based on expression similarities. It then employs non-linear dimensionality reduction to map data from each dataset into a common low-dimensional space. Batch correction is applied to minimise technical variations, and the aligned, integrated data can be used for downstream analyses while preserving biological signals and reducing noise[55, 56]. This step was carried out using the R function ***"RunHarmony()"***. The choice of integration method depends on specific research objectives and dataset characteristics.

3.9 Clustering

Clustering is essential for comprehending functional heterogeneity within cell populations, uncovering rare cell types, and gaining insights into complex biological systems. Seurat's ***"FindClusters"*** function is central to identifying cell types and subpopulations based on gene expression patterns. A clustering tree is constructed to determine the optimal resolution for clustering, generating various clustering distributions based on different resolutions[57]. The resolution parameter influences the number of clusters, and the choice is guided by examining the clustering tree. The ***"clustree"*** package's "clustree()" function facilitates this evaluation[51].

Before clustering, neighbouring points need to be identified. Seurat's ***"FindNeighbors"*** function utilises shared gene expression profiles to identify the K-nearest neighbours for each cell, creating a K-nearest neighbour graph using a shared nearest neighbour (SNN) algorithm. The number of dimensions retained is determined by examining the elbow plot[47]. Subsequently, ***"FindClusters"*** performs unsupervised clustering analysis based on the K-nearest neighbour graph, grouping cells through modularity optimisation[57]. Together, these functions enable the characterisation of cell populations and subpopulations, providing valuable insights into gene expression patterns and functional diversity within scRNA-seq data.

3.10 Cell Type Annotation

Cell type annotation can be done manually, relying on expert knowledge and marker gene expression. Although it is a valuable approach for verifying computational results or identifying rare cell types, it is subjective and time-consuming[50]. Alternatively, computational tools like SingleR streamline cell-type annotation. Singler leverages a reference dataset of known cell types, and their gene expression profiles to predict the cell type of individual cells in a new dataset. It assesses the similarity between each cell's expression profile and the reference dataset, assigning cells to the closest matching cell type[58].

This analysis used SingleR with the ***"FindClusters"*** function to annotate cell types within each cluster. The reference dataset was accessed via the R function ***celldex::HumanPrimaryCellAtlas()***, providing a rich resource of over 350,000 single-cell transcriptomes from diverse human tissues. This dataset empowers researchers to explore gene expression profiles, delineate cell types and subtypes, and investigate molecular mechanisms underlying cell function and disease[59].

The Human Primary Cell Atlas, hosted on the CellDex platform, is a comprehensive single-cell data repository encompassing a wide range of human tissues and organs[60]. For cell type annotation using SingleR, the downloaded reference set is compared to the normalised count data, determining the most likely cell type. This automated approach expedites the annotation process, enhances objectivity, and facilitates the discovery of novel or rare cell types in scRNA-seq datasets.

Similarly, the cell cycling state annotation can also be applied to the data. It is achieved by utilising a function called "***CellCycleScore***", which, in turn, employs "***AddModuleScore***" to calculate a score for two critical phases of the cell cycle, G2/M and S (both of which are associated with mitosis). Furthermore, the "***CellCycleScore***" function categorises each cell into three phases: G2/M, S, or G1. The initial step in this process involves extracting the predefined genes associated with cell cycling, which presumably play pivotal roles in regulating and signalling the various phases of the cell cycle. These genes are essential for accurately characterising and categorising cells based on their cycling states. Plots were created to illustrate the distribution of cells across these phases, offering insights into cell behaviour and cell cycle dynamics. These annotations and visualisations are invaluable for understanding the cell cycle's regulation within primary and metastatic stages.

3.11 Differential Gene Expression

Differential expression analysis conducted via the ***"FindAllMarkers"*** function in the Seurat package allows for identifying genes that exhibit significant differences in expression between specific clusters of cells, conditions, or cell types within their scRNA-seq datasets[47].

This analysis leverages two key statistical measures: log2 fold change (log2FC) and p-values. Log2 fold change (log2FC) quantifies the magnitude of gene expression differences between clusters or conditions. In the context of single-cell data, log2FC values are typically positive because scRNA-seq data predominantly captures upregulated genes. A positive log2FC suggests that a gene is upregulated in one cluster relative to another, reflecting the extent of this upregulation. Conversely, P-values are statistical measures that assess the likelihood of observing gene expression differences by random chance. A low p-value indicates that the observed differences are unlikely to be due to chance, signifying significant differential expression[61]. A commonly used threshold for statistical significance is an adjusted p-value of less than 0.05, which indicates that there is less than a 5% chance that the observed difference in expression is due to random chance. However, the threshold for statistical significance can vary depending on the specific research question and the level of stringency desired[57].

3.12 Comparison with TCGA data

After identifying differentially expressed genes, the data is compared with the TCGA dataset for further analysis and insights. The Cancer Genome Atlas (TCGA) is a collaborative project between the National Cancer Institute and the National Human Genome Research Institute that began in 2005. It aims to analyse various cancer types' genetic and molecular characteristics comprehensively. TCGA has generated a wealth of data that improves understanding of cancer and leads to potential treatment targets and personalised medicine approaches[62]. The Cancer Genome Atlas (TCGA) dataset primarily comprises bulk RNA sequencing (RNA-seq) and microarray data from various cancer types. Unique genes that may not be present in typical reference databases are expected to be encountered when analysing data. These unique genes hold significant importance in cancer research as they can play pivotal roles in cancer progression and provide insights into novel pathways within the cell. The presence of these unique genes in cancer suggests many possibilities. For example, these genes may indicate undocumented mutations, serving as potential biomarkers for specific cancer subtypes, or these unique genes might be involved in undiscovered cellular processes, providing insights into cancer development and suggesting new therapeutic targets. It can also lead to the discovery of rare cell types actively involved in cancer. Investigating their functions could reveal unconventional mechanisms driving malignancy. The "**U**niversity of **A**labama at Birmingham **CAN**cer data analysis (UALCAN) Portal" will be utilised to identify these unique genes in the data. It allows the comparison of the differentially expressed genes with the TCGA data. By doing this, those genes not present in the TCGA database can be identified. This targeted approach helps to focus on the most promising candidates for further investigation, maximizing the potential impact of the research.

3.13 Gene Set Enrichment Analysis

Enrichment analysis in single-cell RNA sequencing (scRNA-seq) data plays a pivotal role in identifying dysregulated pathways and biological processes within cancer cells. The analysis offers a deeper understanding of the molecular complexities driving cancer development and progression, revealing potential therapeutic targets[63]. The pathway analysis was exclusively conducted using the EnrichR tool, focusing on the MSigDB hallmark pathways[44-46]. The Molecular Signatures Database (MSigDB), developed by the Broad Institute, is a vital resource for genomics and cancer researchers. Its "Hallmark Gene Sets," a collection of 50 well-defined biological processes and pathways, are commonly used in gene set enrichment analysis (GSEA) to analyze gene expression data from cancer studies to identify key biological pathways and processes that are dysregulated in different types of cancer. It helps to understand the molecular mechanisms underlying the cancer and identify potential therapeutic targets[64]. This step involved the selection of gene lists for primary and metastatic samples based on specific thresholds, which were subjected to pathway analysis using the EnrichR web portal. The focus primarily remained on the MSigDB hallmark pathways. The top ten were preferred among these pathways based on their respective p-values. This stringent selection criterion enabled the identification of the most statistically significant pathways associated with cancer development and metastasis, thereby highlighting critical biological processes highly relevant to cancer biology. The study aimed to show critical biological processes and signalling pathways highly pertinent to cancer biology by selecting this curated collection of predefined gene sets.

CHAPTER 4: RESULTS

This study employed single-cell RNA sequencing (scRNA-seq) technology to delve into the cellular heterogeneity of primary lung adenocarcinomas and their metastatic sites. A robust scRNA-seq data analysis pipeline was meticulously executed, encompassing quality control, normalisation, scaling, dimensionality reduction, cell clustering, cell type identification, differential expression analysis, and enrichment analysis. A comparison with the TCGA dataset was conducted to enhance the analysis. The study culminated in the identification of distinct cell clusters representing known lung cell types, as well as the discovery of previously uncharacterized cell types specific to lung cancer. Furthermore, two unique genes not present in the TCGA Database were identified. These findings also significantly contribute to our understanding of the complex cellular landscape in lung adenocarcinomas, shedding light on the molecular mechanisms driving tumour progression and metastasis. The results of each step are described in detail in the following paragraphs:

4.1 Data Collection

In the initial phase of the methodology, a publicly available scRNA-seq dataset (Accession No. GSE131907) from the Gene Expression Omnibus was accessed. This dataset contained raw UMI count matrices in the formed of compressed .rds.gz format for two distinct samples: Primary Tumour and metastatic sites in the brain and lymph nodes. Subsequently, a demultiplexing process was used to generate 10X genomics files for each sample. This demultiplexing step was essential in identifying and grouping reads belonging to individual cells based on their unique barcode sequences. High-quality transcriptome data was obtained for each cell, which was utilised for subsequent analysis.

The resulting 10X genomics files underwent preprocessing using the Seurat package in R. This preprocessing step involved filtering out low-quality cells that expressed less than 100 genes and genes that are expressed in less than 3 cells per sample. After the preprocessing, 45,149 cells were retained in primary samples, as shown in Figure 4-1 and 31,497 cells in metastatic samples, as shown in Figure 4-2. This rigorous filtration process ensured that only high-quality cells and genes were retained for downstream analysis.

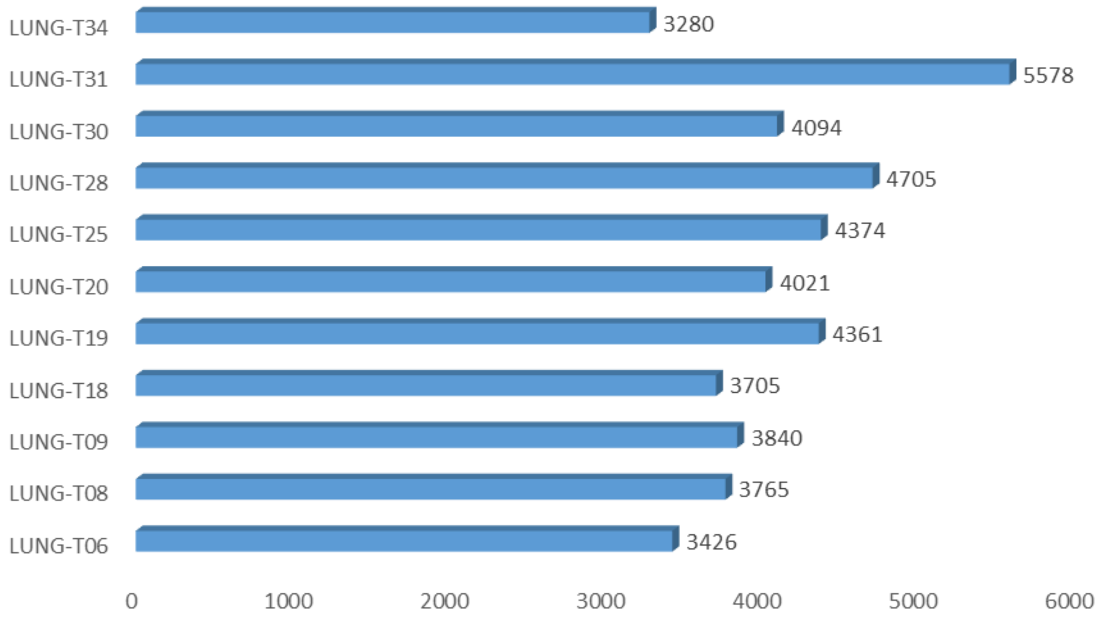

Figure 4.1: Frequency plot of cell abundance in each Primary sample

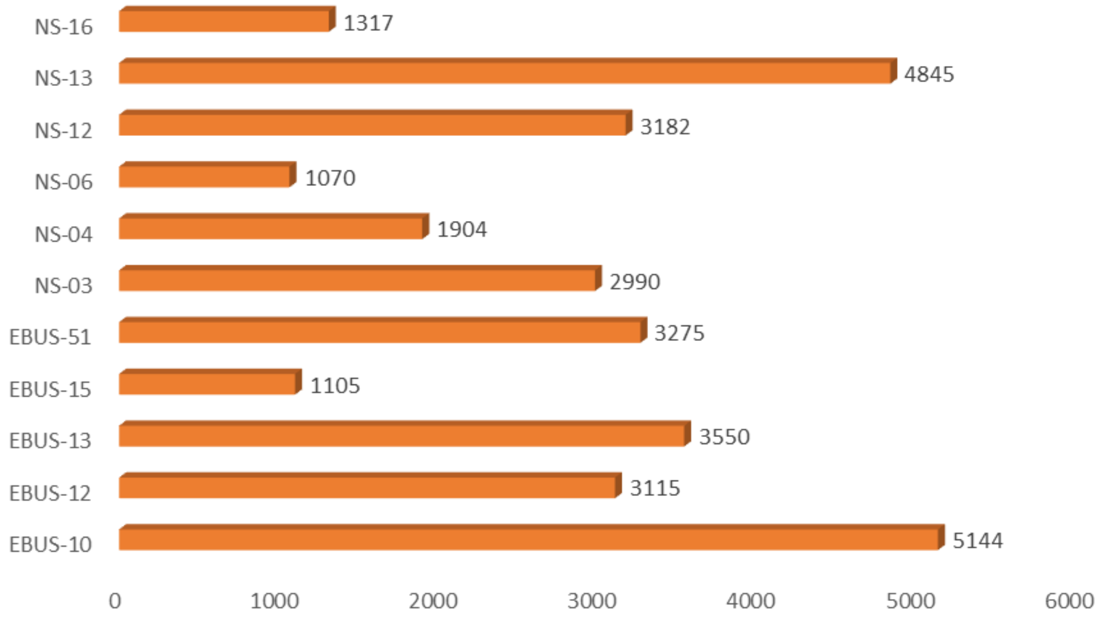

Figure 4.2: Frequency plot of cell abundance in each Metastatic sample

4.2 Quality Control

In the Quality Control (QC) phase, the two assays were treated separately: the primary tumour and metastatic samples. Since the dataset had already undergone some preprocessing, we ended up with 45,149 clean-read cells for the primary tumour and 31,497 for the metastatic samples, following filtering and quality control measures.

During this QC process, various metrics were assessed, including the percentage of mapped reads to the genome, the count of unique molecular identifiers (UMIs), the number of detected genes, and the percentage of mitochondrial and ribosomal genes, among others. In this step, two key filtering steps were applied. Firstly, cells with over 8% expression of mitochondrial genes were removed to prevent potential interference with expression analysis as they depict damaged cells. Figures 4-5 and 4-6 show the mitochondrial, ribosomal and haemoglobin gene percentages. These figures depict removing the cells with mitochondrial genes greater than 8%. Secondly, cells displaying more than 5000 features were identified as potential doublets and excluded from further analysis to maintain data accuracy, as shown in Figures 4-3 and 4-4, respectively. These quality control measures were implemented to ensure the reliability of subsequent data analysis and interpretation. The outcomes of our QC analysis confirmed that all the samples met the predefined QC criteria, and all of them were retained for subsequent analysis.

4.3 Normalisation and Scaling

Normalisation and scaling procedures were carried out to address variations in sequencing depth and technical noise across the samples. After normalisation, an average of around 50,000 unique molecular identifiers (UMIs) per cell across all samples were achieved, ensuring consistency. Subsequently, the ***"ScaleData"*** function from Seurat was utilised to standardise the data, accounting for differences in sequencing depth among individual cells. It shifted gene expression to a mean of 0 across cells and scaled it to have a variance of 1. Furthermore, comparison of the distribution of UMIs per cell was carried out across samples and found no significant differences, suggesting that the normalisation and scaling methods effectively eliminated batch effects[65].

4.4 Dimensionality Reduction

After completing quality control, normalisation, and scaling, Principal Component Analysis (PCA) was applied to reduce the dimensionality of the scRNA-seq data for both Primary and Metastatic samples. Elbow plots were constructed based on the PCA results to determine the most suitable number of principal components to retain. The elbow plots guided the selection of the top 30 principal components in both datasets for further analysis. Subsequently, t-distributed Stochastic Neighbor Embedding (t-SNE) was utilised for data visualisation, which helped observe patterns in the data and identified clusters of cells exhibiting similar expression profiles.

As a preliminary step in dimensionality reduction, a subset of features that display significant cell-to-cell variation within the datasets was identified. In both assays, 2,000 genes (features) per dataset were retained by default for downstream analysis, such as PCA. Focusing on these genes in subsequent analyses helps us emphasise the biological signals within the single-cell datasets. The 2,000 highly variable genes across both datasets are shown in Figures 4-9 and 4-10, respectively. Selecting 2000 variable features helps cover diverse biological information without using enormous computational resources. Additionally, it ensures the inclusion of crucial genes and pathways relevant to the biological question while facilitating effective dimensionality reduction techniques like PCA or t-SNE.

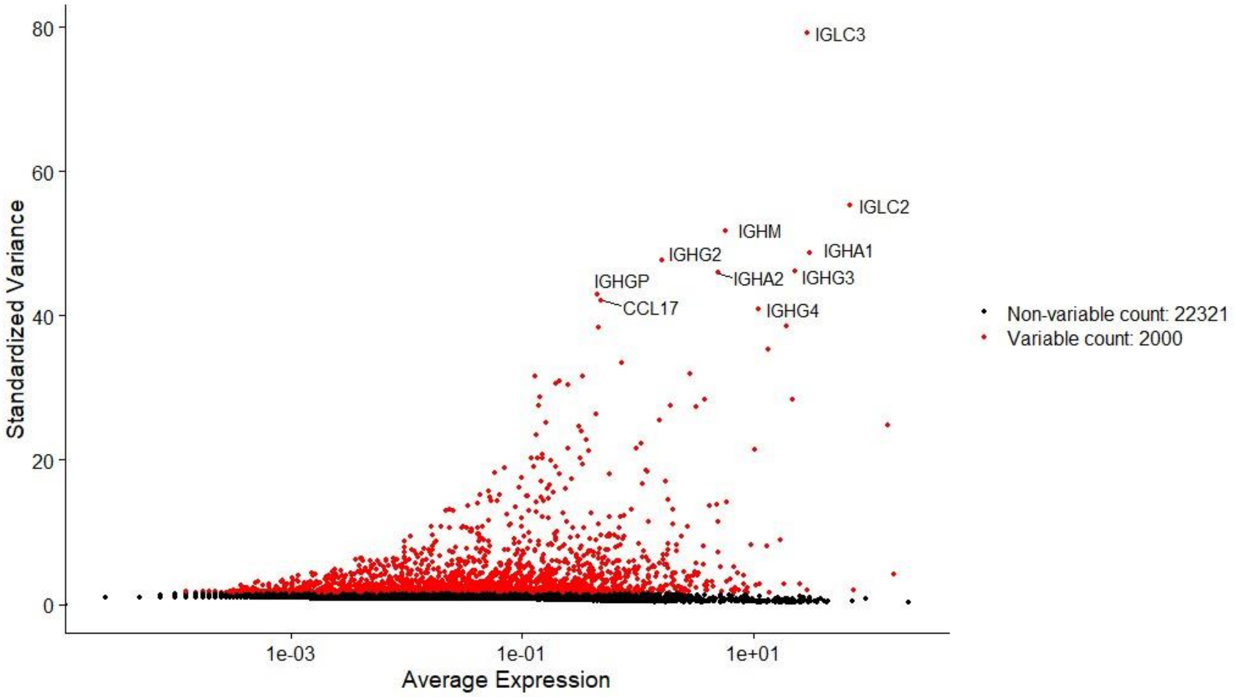

Figure 4.3: Plot showing two thousand variably expressed genes (Primary samples)

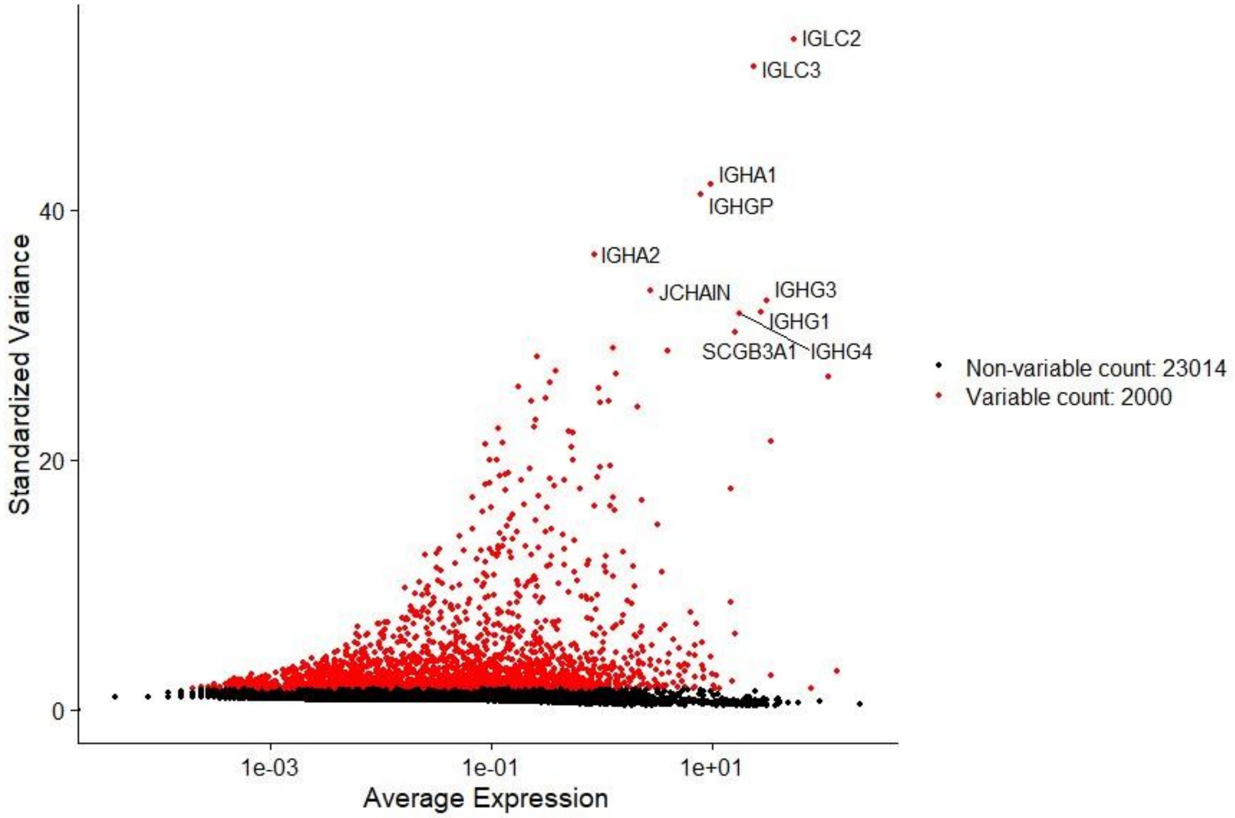

Figure 4.4: Plot showing two thousand variably expressed genes (Metastatic samples)

The elbow plots are visualised to see how many principal components are needed to retain for the downstream analysis. The elbow plot is a powerful tool in scRNA-seq data analysis that helps determine the optimal number of principal components for downstream analysis.

Elbow plots for choosing the number of components for both primary and metastatic samples are given below, as shown in Figures 4-13 and 4-14. There is a sharp decrease in standard deviation in the initial number of components added, which signifies that as components are added, variations are better explained, but after a while, adding more components does not affect the standard deviation and a plateau is reached. The number of components needed to reduce is picked after the plateau is achieved, which looks like the bent elbow of a human, signifying its name. In both datasets, the elbow point, indicating the optimal number of principal components, was observed around the 25th component. However, 30 principal components were selected to enhance data compatibility and minimise potential errors, which may have led to overfitting.

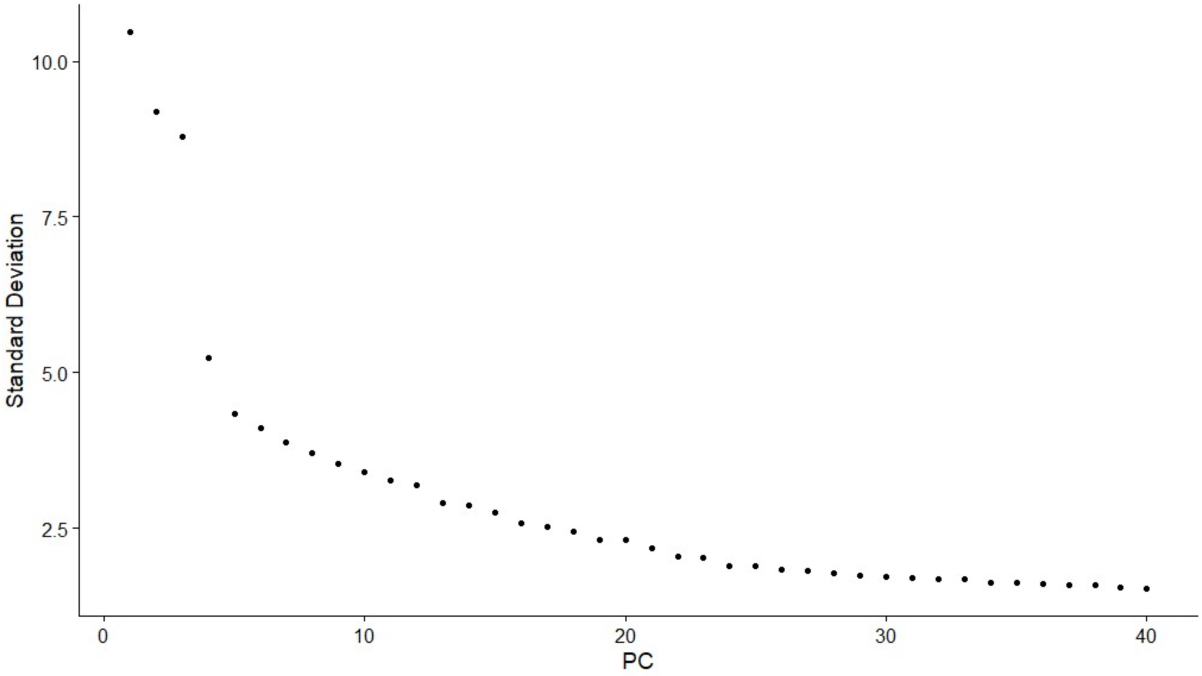

Figure 4.5: Elbow plot for selecting appropriate no. of PCs (Primary samples)

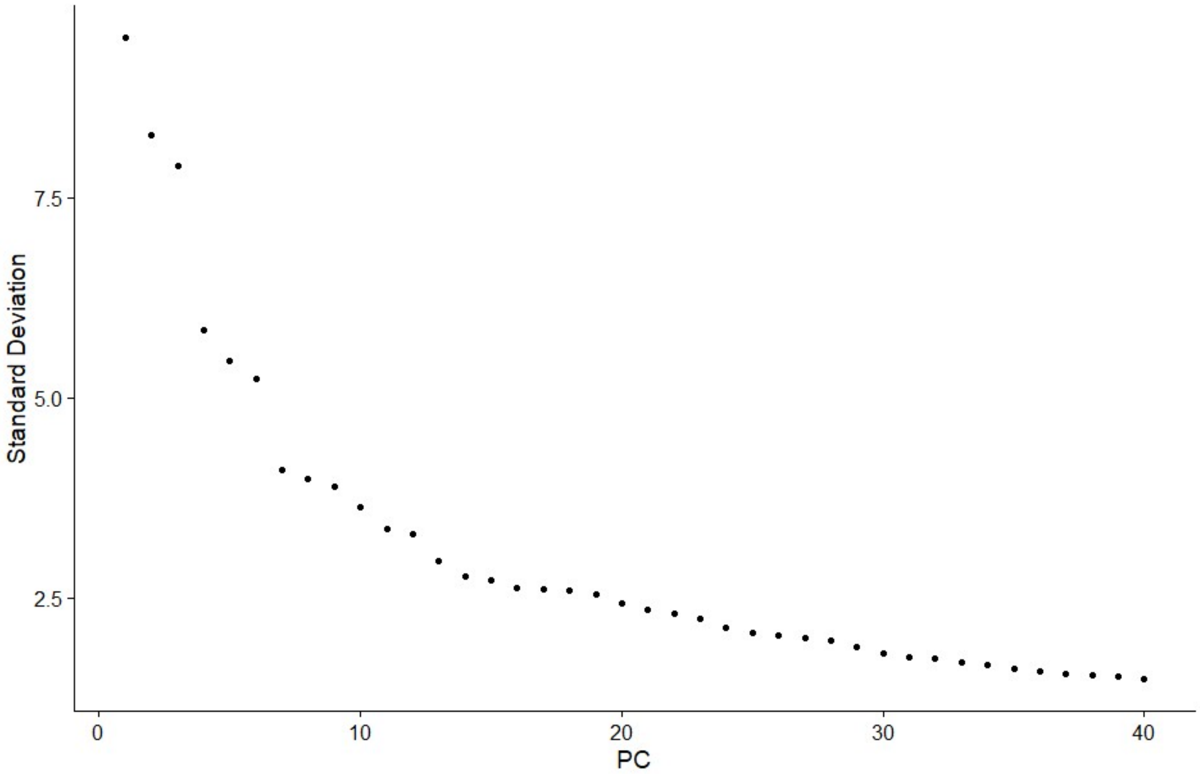

Figure 4.6: Elbow plot for selecting appropriate no. of PCs (Metastatic samples)

The t-SNE plots, based on the ideal number of principal components chosen from the elbow plot, offer a visual representation of the data in two or three dimensions. These plots help to identify distinct clusters within the data. In Figures 4-15 and 4-16, numerous clusters for both primary and metastatic samples indicate different cell types. In Figure 4-14, focusing on the primary data, it is noticed that samples T30 and T34 form separate clusters from the rest of the samples. Further analysis reveals that these patients have never smoked, unlike the others who are either current smokers or ex-smokers. It suggests that specific unique characteristics in these patients might contribute to cancer development.

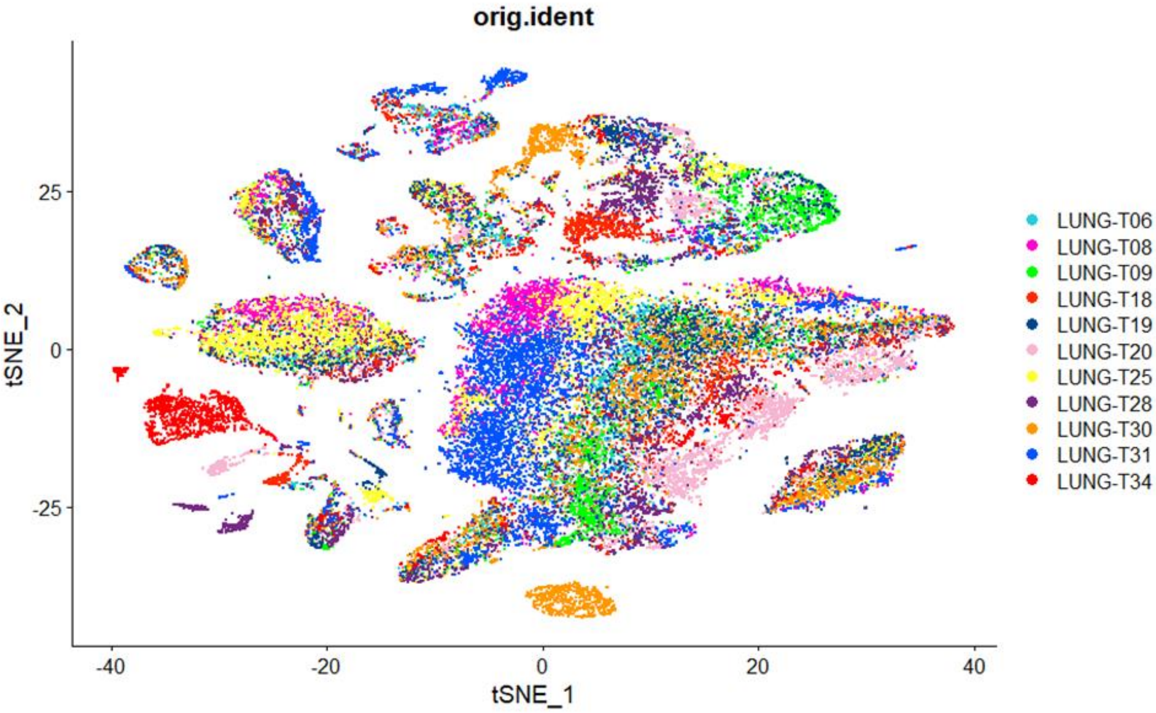

Figure 4.7: Raw TSNE after dimensionality reduction (Primary samples)

Moving to Figure 4-16, which shows metastatic samples, the clusters are clearer and more diverse. It could be attributed to different metastatic regions and varied patient histories. Some patients are in stage 3, while others are in stage 4, and there is a mix of smokers, non-smokers, and ex-smokers among them. Overlaying these clusters with cell type annotations or gene expression data helps extract insights into the underlying biology. By reducing the dimensionality of the data, it becomes easier to identify specific cell subpopulations and explore relationships between different cell types.

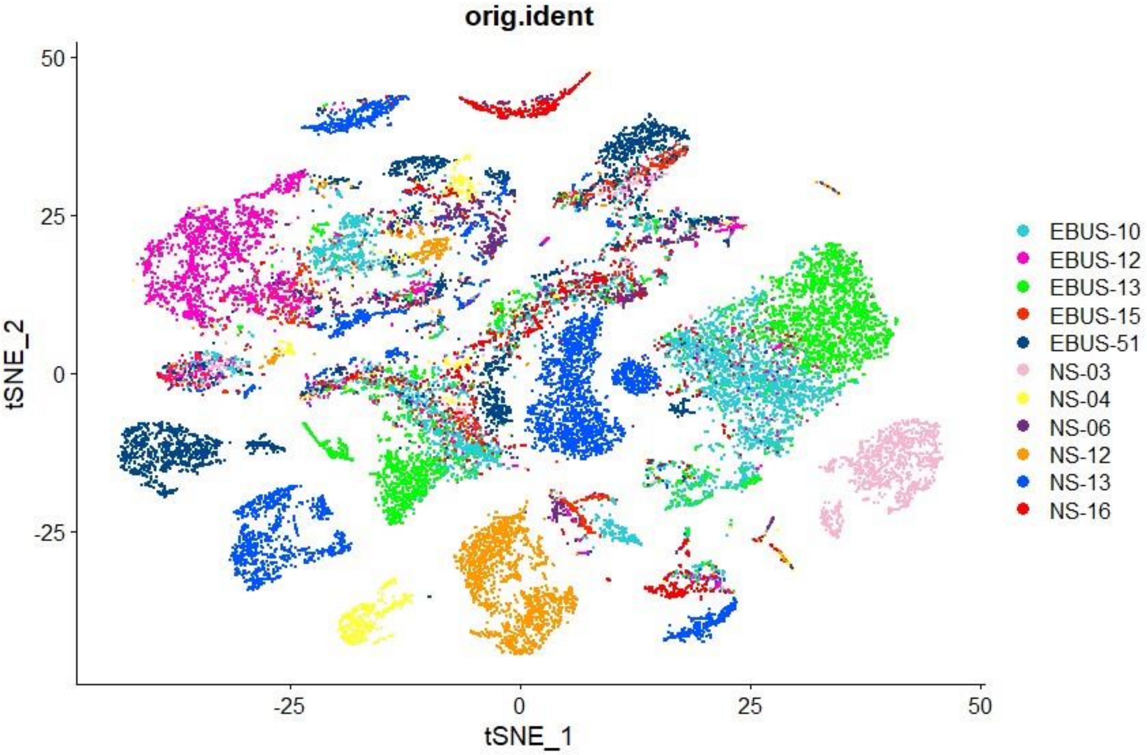

Figure 4.8: Raw TSNE after dimensionality reduction (Metastatic samples)

4.5 Integration

The Harmony package in R was utilised to integrate the data. It constructs shared nearest-neighbour graphs for individual datasets and then employs an iterative optimisation process to align and harmonise cells across batches. The algorithm adjusts cell positions to minimise batch-specific differences while preserving biological signals.

Following integration through Harmony, the t-SNE analysis was carried out again. Before integration, Figures 4-15 and 4-16 show clusters in both primary and metastatic samples in their raw form, scattered and less organized. Cells are primarily clustered by their respective samples, representing an initial clustering stage with no refinement. However, after integration, as shown in Figures 4-17 and 4-18, clusters in both primary and metastatic samples become more distinct and separated from each other compared to the raw t-SNE. Cells are more evenly distributed based on similarities and differences. In the primary data, sample T34 is confined to a single separate cluster, which can be attributed to the presence of distinct features or cells.

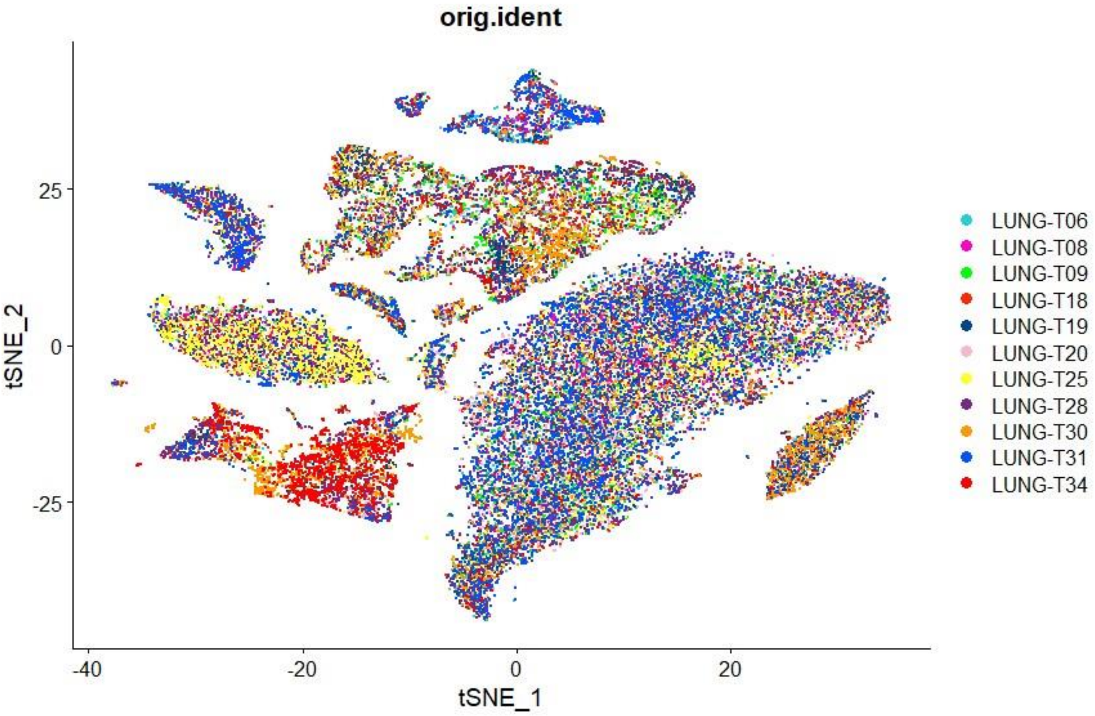

Figure 4.9: Integrated TSNE after dimensionality reduction (Primary samples)

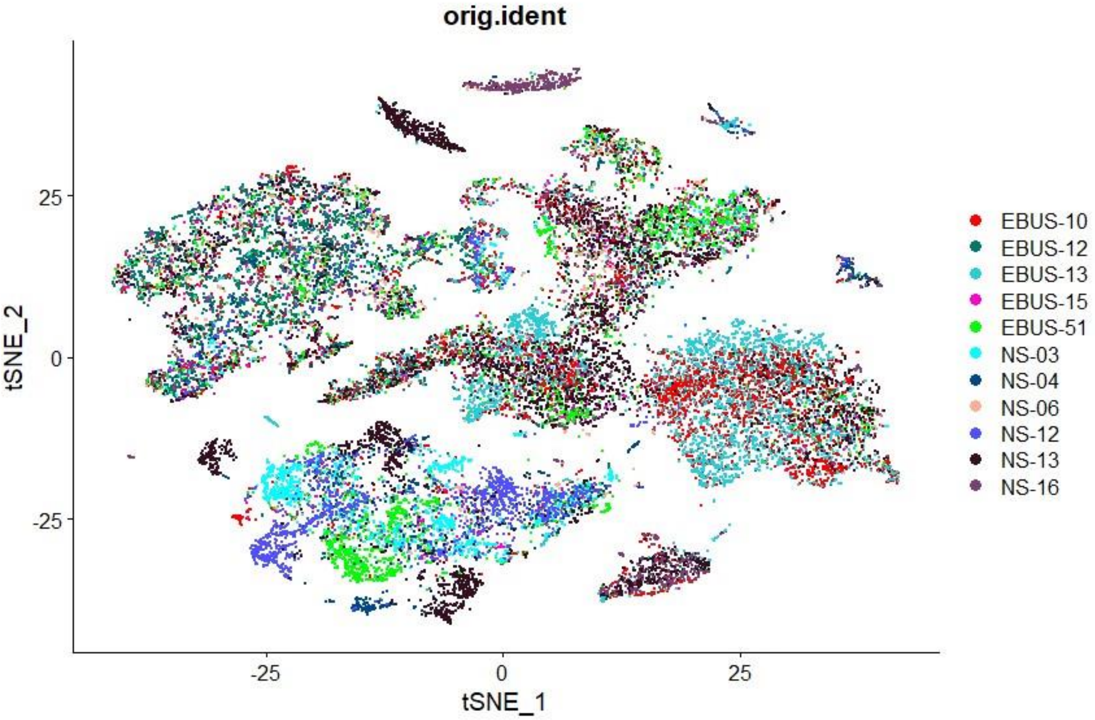

Figure 4.10: Integrated TSNE after dimensionality reduction (Metastatic samples)

The integration process yielded unified datasets that promoted more robust and informative downstream analyses, providing valuable insights into the biology being studied. These harmonized datasets are utilised for subsequent tasks like clustering, cell type annotation, and differential expression analysis.

4.6 Clustering

Before conducting clustering, a hierarchical clustering tree was created to depict the relationships between cell clusters at various resolutions. A range from 0.1 to 0.8 was explored to determine the most suitable clustering resolution. The tree illustrated different levels of cell clusters at each resolution, with broader clusters at lower resolutions and more refined, specific clusters at higher resolutions.

In primary samples, a resolution of 0.3 emerged as the most suitable value for clustering, yielding 17 distinct clusters, as shown in Figure 4-19. This choice balanced interpretability and computational efficiency. Notably, this resolution facilitated clear clustering, with only one additional division observed at cluster 7. Upon closer examination, it became evident that cluster 0 harboured the highest number of transcripts, followed by clusters 1 and 2. The colour intensity of edges within cluster 0 suggested elevated expression levels compared to other clusters. These clusters likely represent the dominant cell types in the data, such as immune cells or epithelial cells. Further observation of edge transparency at the 0.3 resolution revealed stable clustering patterns within the data.

For metastatic samples, a resolution of 0.5 emerged as the most reasonable value, as shown in Figure 4-20, resulting in 22 distinct clusters. Although clusters 0 and 2 show further projection towards clusters 6 and 12, respectively, they exhibited less effectiveness due to their transparency, as depicted in Figure 4-20. This indicates that the cells within Cluster 2 share some similarities with Cluster 12, while those in Cluster 0 are akin to Cluster 6. Cluster 0 contained the highest number of cells, followed by clusters 1 and 2. The colour of the edges also suggests that cluster 1 exhibits the highest expression level, followed by clusters 0, 3, 2, and 4. Furthermore, the edges indicate that the clustering at this resolution is stable. The more considerable number of clusters in metastatic samples suggests that they contain cells from other regions and lung cells.

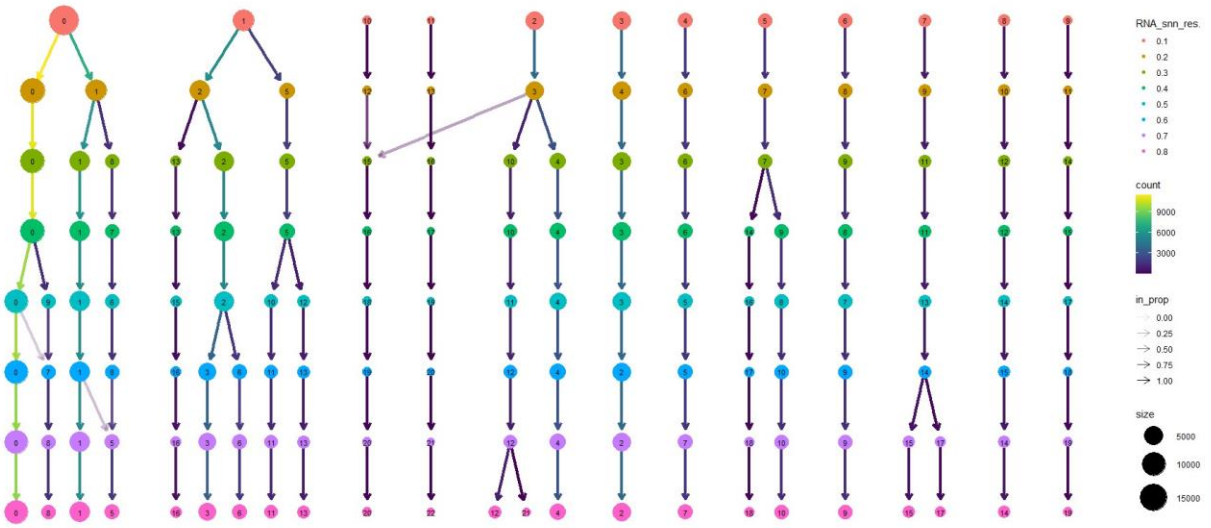

Figure 4.11: Clustering tree of resolution ranging between 0.1 - 0.8 (Primary samples)

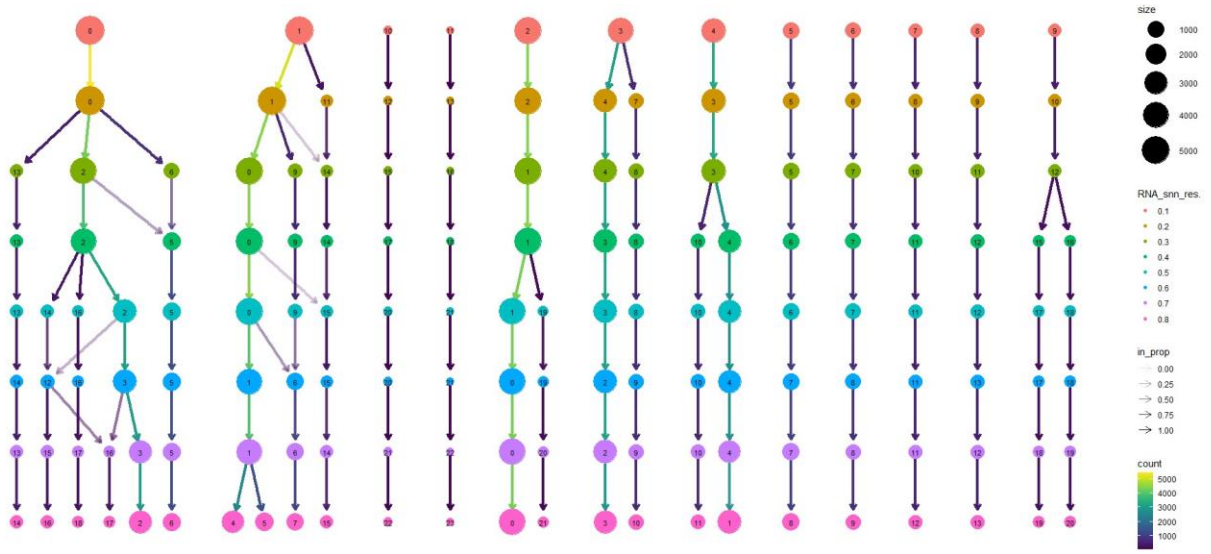

Figure 4.12: Clustering tree of resolution ranging between 0.1 - 0.8 (Metastatic samples)

Based on the hierarchical clustering tree, 0.3 for primary samples and 0.5 for metastatic samples were selected to cluster cells. Once clustering was completed, we applied the dimensionality reduction technique known as t-SNE to visualise the clustering outcomes. Separate t-SNE plots were created for each of our samples, and the t-SNE plots for both states can be found below:

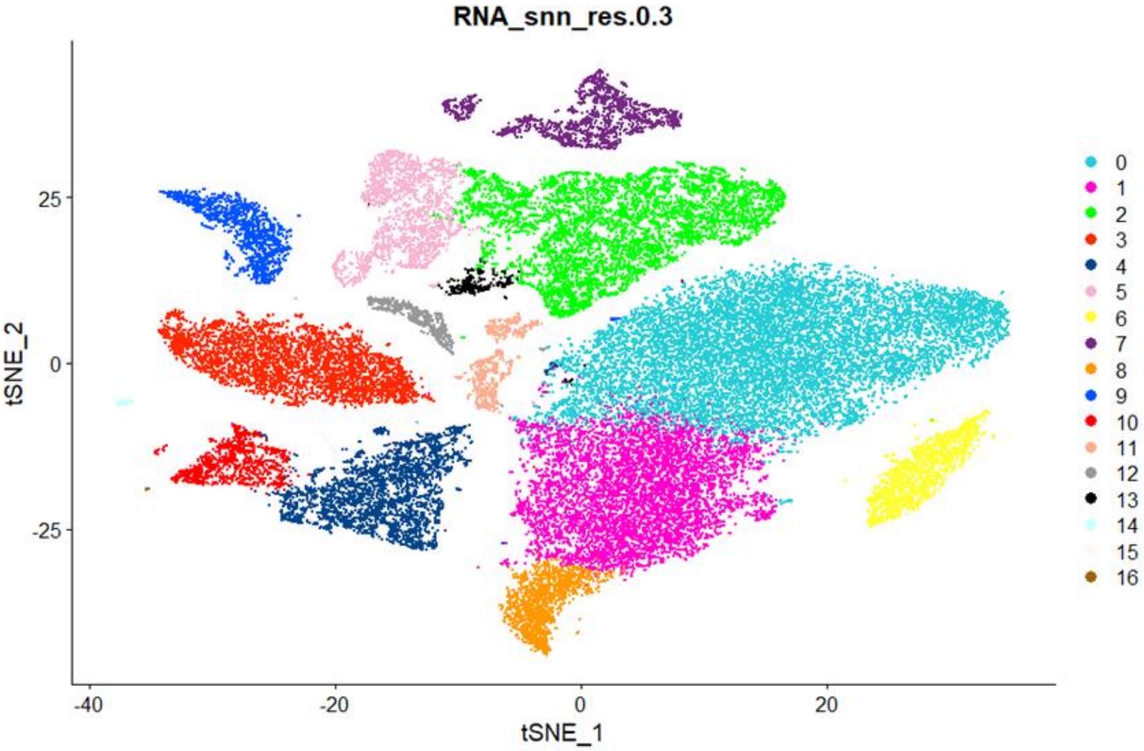

Figure 4.13: TSNE with 0.3 resolution (Primary samples)

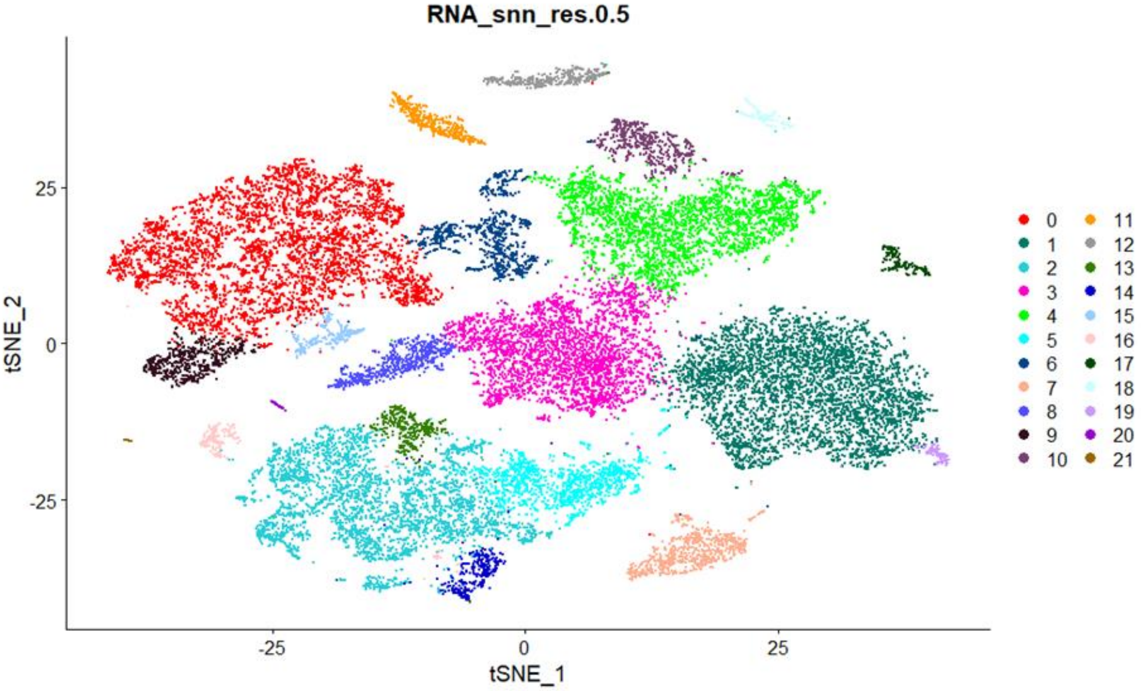

Figure 4.14: TSNE with 0.5 resolution (Metastatic samples)

4.7 Cell Type Annotation

The cell type annotation step entails assigning a biological identity to each cluster identified during the prior clustering phase. This study used The SingleR package with the Human Primary Cell Atlas as a reference dataset for cell annotation. 19 distinct cell type clusters were readily shown in primary samples and 22 in metastatic samples, encompassing various cell types like tumour, endothelial, T, B, and macrophages. A predominant T cell population, constituting approximately 40-45% of all cells, was notable in the primary samples depicted in Figure 4-23. These T cells are a crucial part of the immune system and actively fight against the tumour. Other prominent cell types included macrophages, epithelial cells, and B lymphocytes. Smaller populations of endothelial cells, monocytes, and natural killer cells, each making up less than 10% of the total cell count, were also identified.

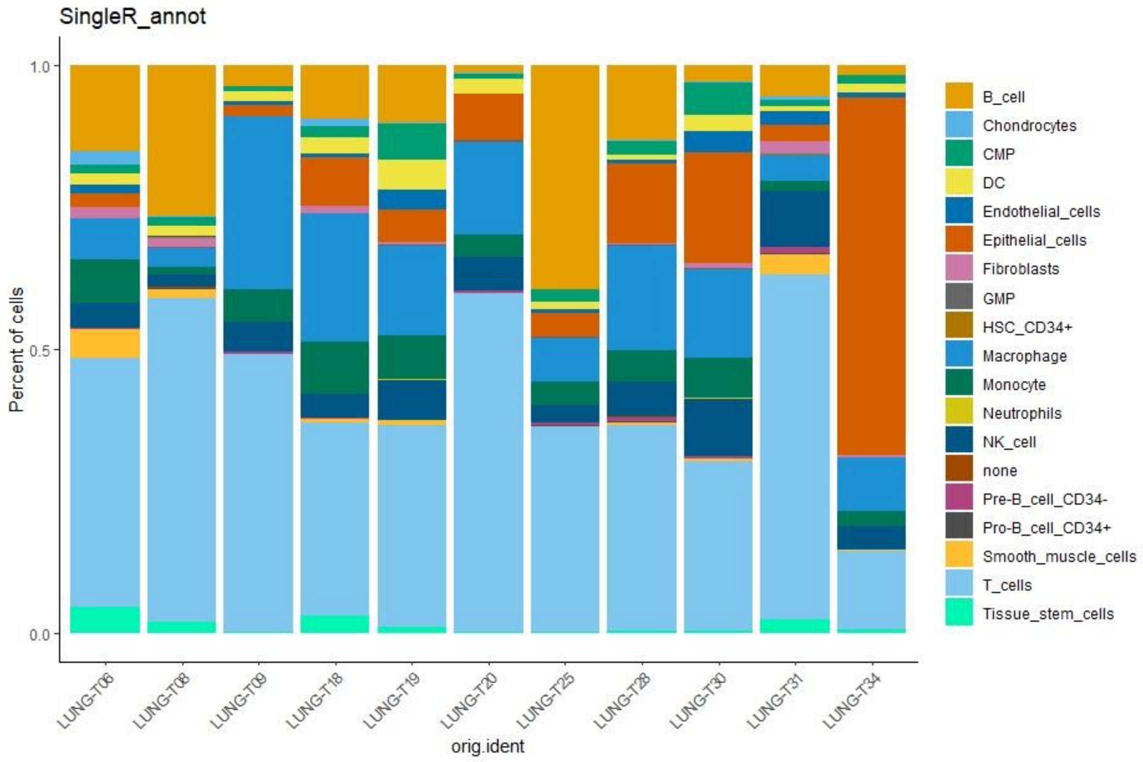

Figure 4.15: Stacked bar plot showing cell type abundance in each Sample (Primary samples)

In contrast, metastatic tumours exhibited a more diverse range of cell types, as shown in Figure 4-24. It shows that they metastasised to other body parts. Along with the abundance of immune cells, metastatic tumours also contained substantial numbers of other types of cells, including epithelial cells, fibroblasts, astrocytes and monocytes, as shown in Figure 4-24. The variation in cell composition between primary and metastatic tumours highlights the dynamic nature of cancer progression. As tumours evolve and disseminate to distant locations, they undergo genetic, epigenetic, and microenvironmental changes that shape the cellular landscape. These changes and factors, such as treatment effects and site-specific influences, contribute to the observed differences in cell composition. Understanding these variations is crucial for developing targeted therapies personalised to the specific cell types in metastatic tumours, ultimately improving outcomes for patients with cancer.

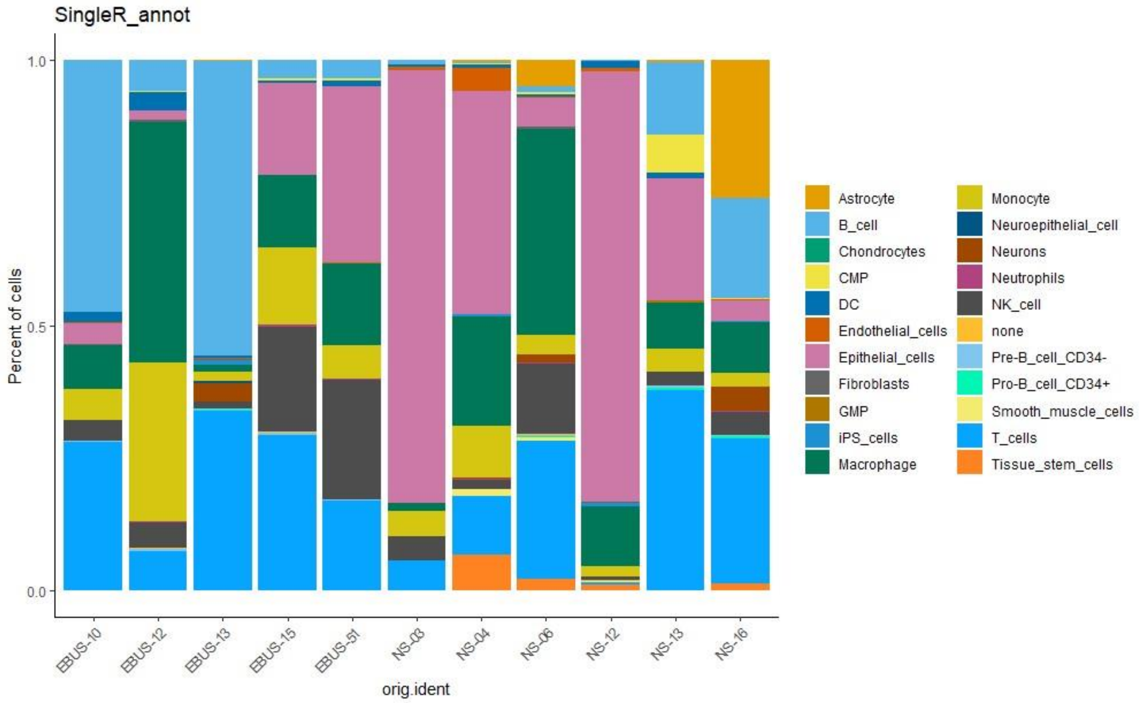

Figure 4.1: Stacked bar plot showing cell type abundance in each Sample (Metastatic samples)

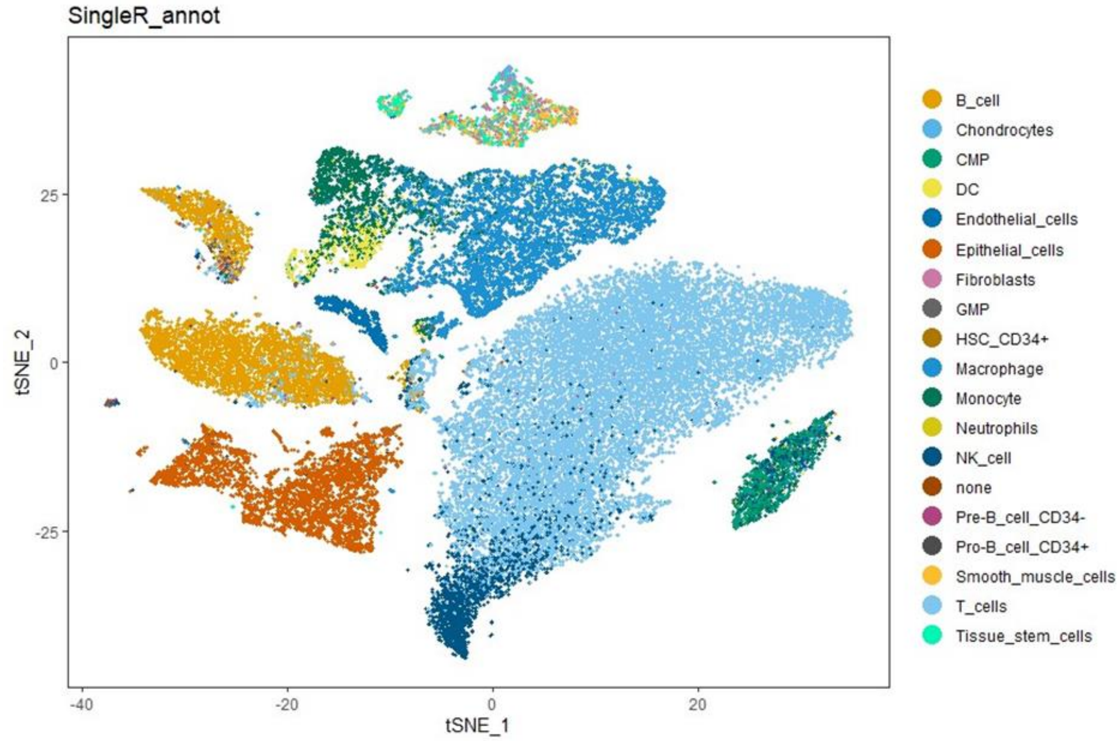

Figure 4.17: TSNE labelled with cell types (Primary samples)

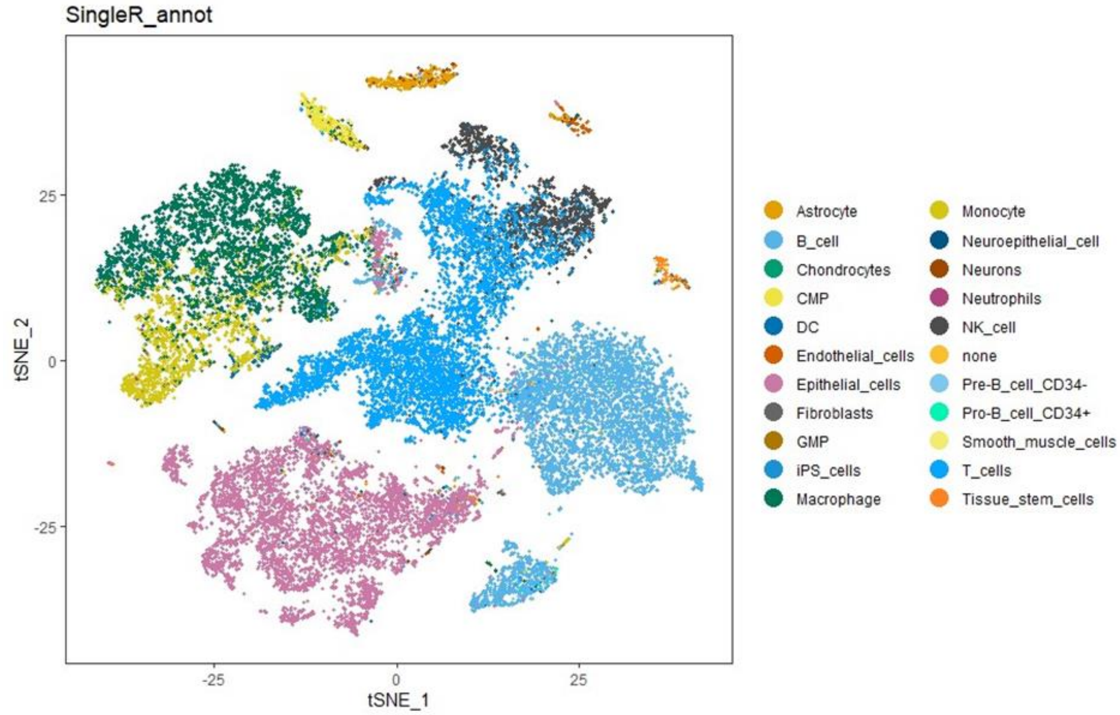

Figure 4.18: TSNE labelled with cell types (Metastatic samples)

The cell cycle dynamics observed in primary and metastatic LUAD samples depict distinct stages of tumour progression and microenvironmental adaptations. In primary tumour stages characterized by localized growth, a prevalence of cells in the G2/M and S phases signifies active proliferation and DNA replication, typical of rapidly dividing cancer cells, as shown in Figure 4-27. Additionally, T and B cells in similar phases suggest a concurrent immune response, possibly driven by the tumour's antigens. However, as the disease advances to the metastatic stage, where cancer cells spread to distant sites, a notable shift toward G1 phase dominance emerges, as shown in Figure 4-28. This alteration hints at a potential slowdown in cell cycle progression, likely influenced by changes in the microenvironment. Metastatic cells may enter an inert state to adapt to new environmental conditions, evade immune detection, or initiate the formation of secondary tumours. The observed differences indicate the interactions between cancer cells and their surroundings, shaping tumour development and invasion to distant sites throughout the disease progression.

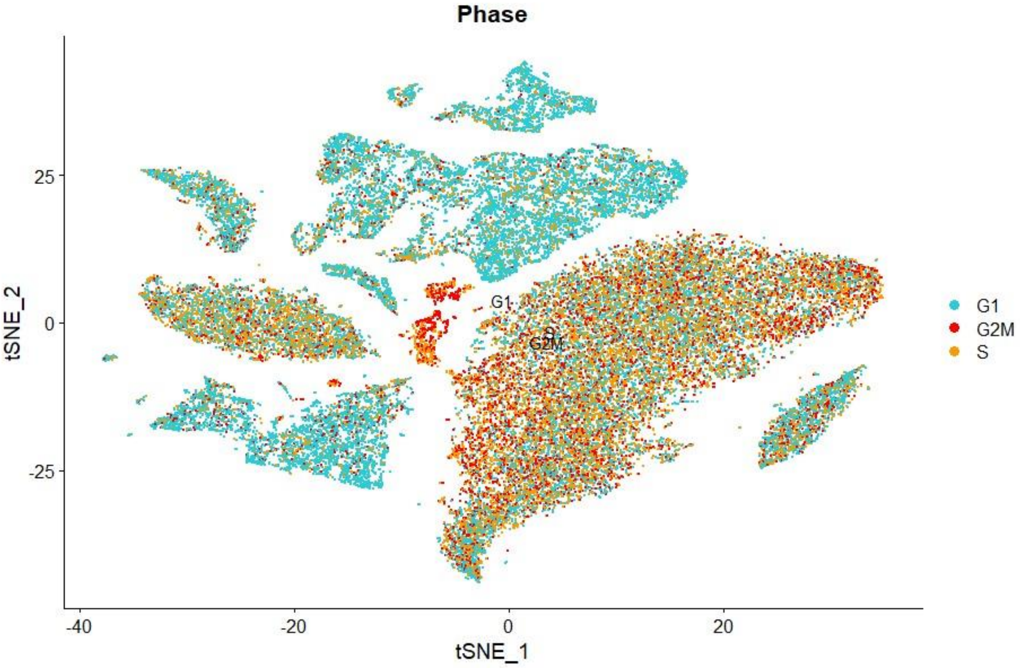

Figure 4.19: TSNE labelled with cell Cycle Phases (Primary samples)

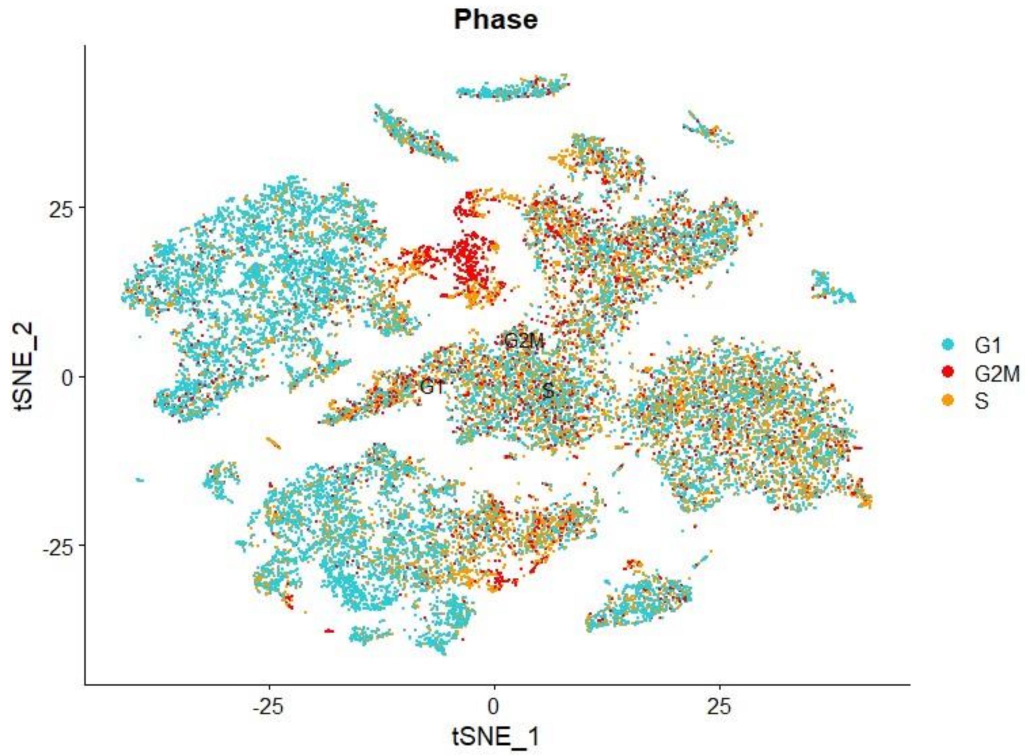

Figure 4.20: TSNE labelled with cell Cycle Phases (Metastatic samples)

4.8 Differential Gene Expression

Differential expression analysis revealed distinct gene expression patterns among different cell types within primary tumour samples, as indicated by the "findallmarkers" analysis conducted separately for primary and metastatic samples. Several genes showed a unique expression profile in primary tumour samples, as shown in Figure 4-29, with only a few being expressed across multiple cell clusters: Ferritin Light Chain (FTL), Cystatin B (CSTB), Surfactant Protein C (SFTPC), Immunoglobulin Lambda Constant 2 (IGLC2) and Immunoglobulin Kappa Constant (IGKC). FTL and CSTB are associated with tumour-promoting activities. FTL is known for its involvement in gliomas. The hypoxic microenvironment characteristic of these tumours influences it. It is crucial in regulating the Epithelial Mesenchymal Transition (EMT) process, contributing to tumour invasion and metastasis. It induces EMT through the Protein Kinase B/Glycogen Synthase Kinase 3 Beta/Beta-catenin (AKT/GSK3β/β-catenin) signalling pathway, enhancing tumour aggressiveness[66]. Cystatin B (CSTB) is implicated in hepatocellular carcinoma (HCC), where its overexpression correlates with aggressive tumorigenesis and poor prognosis.

Functionally, CSTB acts as an oncogene, driving tumour growth and invasion by promoting cell proliferation and inducing G2 phase cell cycle arrest. CSTB also positively regulates EMT and influences vital signalling pathways such as MAPK, mTOR, and AKT, thereby contributing to cancer progression[67]. In contrast, SFTPC exhibits tumour-suppressing activity in LUAD by inhibiting the PI3K/AKT/mTOR pathway, which controls cell proliferation. Low levels of SFTPC are associated with poor prognosis in LUAD patients and increased sensitivity to Programmed Cell Death Protein 1 (PD-1) and Cytotoxic T-Lymphocyte-Associated Protein 4 (CTLA-4) antibodies, suggesting its potential as a predictive marker for immunotherapy response[68]. Targeting SFTPC or its downstream pathways may offer new therapeutic avenues for LUAD treatment. IGLC2 and IGKC, as components of the immune system, contribute to antibody production and play a role in identifying and neutralizing abnormal cells, thus aiding the body's defence against cancer[69, 70]. The dot plot analysis of primary samples depicted distinct gene expression patterns among different cell types within the primary tumour microenvironment, indicating that cells maintain their unique characteristics and polarity within the primary stage.

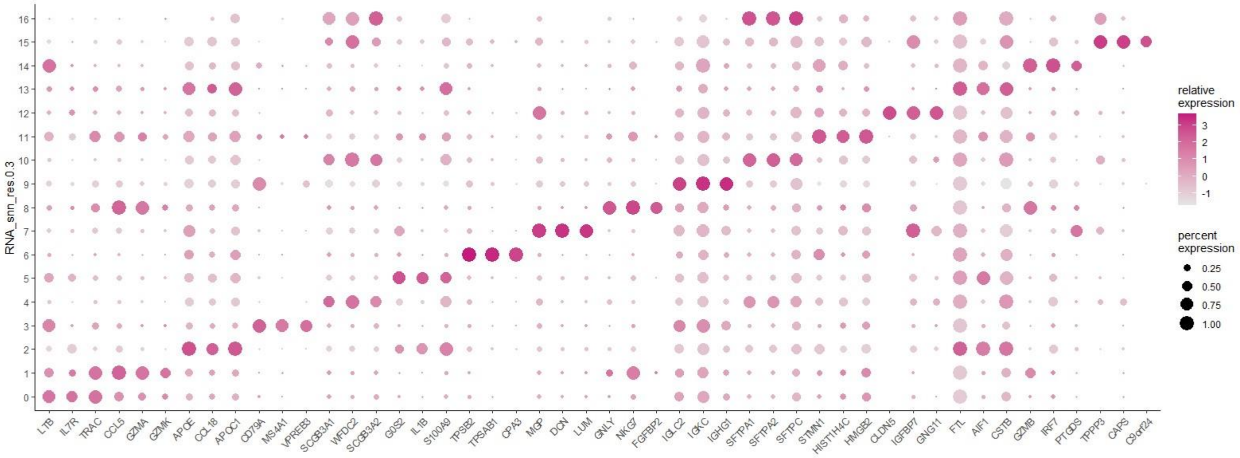

Figure 4.21: Dot plot showing top three markers for each cluster (Primary samples)

In metastatic samples, a consistent expression pattern of genes, including Microsomal Glutathione S-Transferase 1 (MGST1), Keratin 19 (KRT19), CD9 Molecule (CD9), Keratin 18 (KRT18), Growth Hormone 3 (GHG3), Immunoglobulin Lambda Constant 2 (IGLC2), Adenomatous Polyposis Coli (APC), Secreted Phosphoprotein 1 (SPP1), Aldo-Keto Reductase Family 1 Member B1 (AKR1B1), Surfactant Protein A1 (SFTPA1), Phospholipase C Gamma 2 (PLCG2), Surfactant Protein C (SFTPC), X Inactive Specific Transcript (XIST), BCL2 Interacting Protein 3 (BNIP3) and Secretory Leukocyte Peptidase Inhibitor (SLPI) was observed as shown in Figure: 4-30, suggesting a stemness signature across various cell types. This stemness indicates heightened expression of multiple genes associated with stem cell properties and genes related to tumour suppression and immune response. This uniform stemness signature implies a shift in cancer cell biology at the metastatic stage, potentially indicating a more aggressive phenotype characterized by epithelial-to-mesenchymal transition (EMT). EMT is a crucial process wherein cancer cells acquire traits enhancing their invasive potential, facilitating detachment from the primary tumour and initiation of metastasis. The activation of EMT in later cancer stages is pivotal for metastatic spread, as observed in the differential expression results. CD9, KRT19, KRT18, and SPP1 are some of the genes contributing to this stemness. Research indicates that CD9 expression in pancreatic ductal adenocarcinoma promotes annexin A6-induced cell migration and EMT via the p38 MAPK pathway[71]. KRT19 and KRT18 play pivotal roles in circulating tumour cells' survival and metastatic potential (CTCs) by enhancing structural integrity, enabling clustered CTCs to withstand shear forces and evade anoikis[72]. Additionally, they contribute to immune evasion mechanisms, potentially hindering interactions with Cytotoxic T Lymphocytes (CTLs) via Major Histocompatibility Complex-1 (MHCI) inhibition. These genes are implicated in various aspects of cancer biology, including EMT regulation, cell adhesion, and migration. Their expression in metastatic samples further supports the shift in gene expression patterns during cancer progression, elucidating the molecular mechanisms driving tumour growth and metastasis. Furthermore, osteopontin/secreted phosphoprotein 1 (SPP1) has been highlighted for its pivotal role in promoting stemness characteristics in pancreatic cancer (PC) cells within the tumour microenvironment[73]. Elevated SPP1 levels also correlate with poor prognosis in LUAD patients, indicating its potential as a prognostic biomarker. SPP1 promotes LUAD progression and metastasis by driving the EMT pathway, enhancing cancer cell mobility and invasiveness. Additionally, SPP1 regulates the tumour microenvironment by modulating Collagen Type XI Alpha 1 Chain (COL11A1) expression, further fuelling EMT and metastasis in LUAD cells[74]. Targeting the SPP1-COL11A1 axis presents a promising therapeutic strategy to inhibit LUAD progression and metastasis. Additionally, genes such as APC, BNIP3, IGHG3, and IGLC2 were found to be expressed in multiple cell types independent of their cellular function, and they are implicated in tumour suppression and inhibition of metastasis[69, 70, 75, 76].

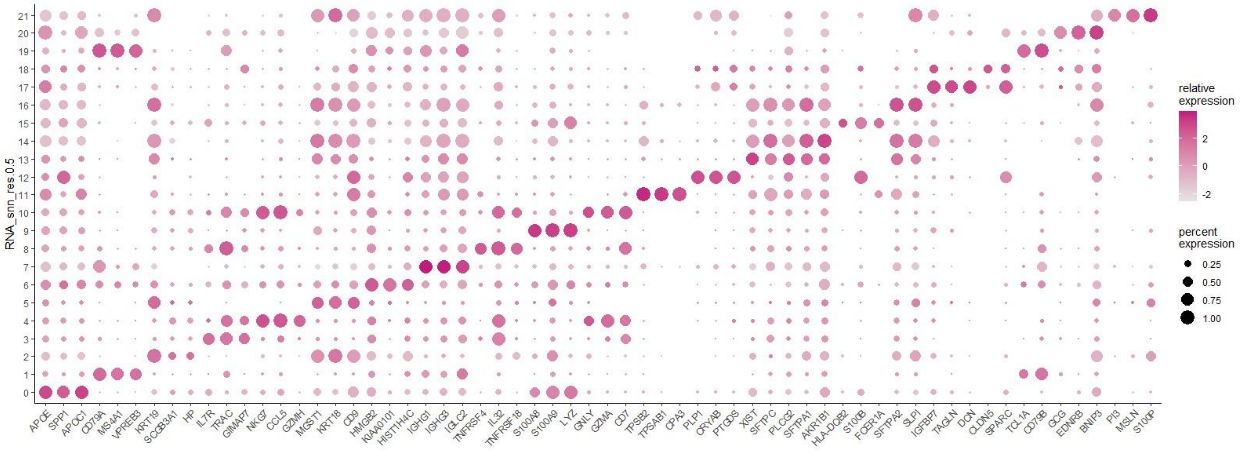

Figure 4.22: Dot plot showing top three markers for each cluster (Metastatic samples)

The differential expression highlights how cancer cells change as they spread. Initially, different genes are active in different cell types within the tumour. However, as cancer progresses to the metastatic stage, a common set of genes linked to stem cell properties becomes active across various cell types. This shift marks a crucial step where cancer cells become more aggressive and able to spread to other parts of the body. Understanding these changes can help develop better treatments for various stages of cancer.

4.9 Comparison with TCGA data

The Cancer Genome Atlas (TCGA) is a pioneering project by the National Cancer Institute and the National Human Genome Research Institute. It has been running since 2005. It aims to analyze various cancers' genetic and molecular features to improve understanding of cancer and personalised treatments. TCGA data primarily includes bulk RNA sequencing and microarray data from diverse cancer types. By comparing the differential analysis results with TCGA data, the scRNA-seq also identified two more genes not reported in TCGA. One is Adhesion G Protein-Coupled Receptor L2 (ADGRL2), downregulated in the metastatic stage, and the second is Endoplasmic Reticulum Oxidoreductase 1 Alpha (ERO1A), which is overexpressed in the metastatic stage. In primary samples, looking at Figure 4-13, ADGRL2 shows a higher expression level with a log2FC of 3 and is only expressed in endothelial cells. However, in metastatic samples, the expression of ADGRL2 decreases with a log2FC value below 2, as shown in Figure 4-32, while it is expressed in endothelial cells along with smooth muscle cells. This shift in expression pattern suggests a potential loss of specificity towards endothelial lineage and a broader distribution within the tumour microenvironment.

On the other hand, ERO1A demonstrates a significant upregulation in metastatic samples, as shown in Figure 4-34, characterized by a log2FC value of 5. In primary samples, ERO1A exhibits a lower log2FC value of around 3, as shown in Figure 4-3, and is confined to epithelial cells and macrophages. The metastatic environment showed an expansion of ERO1A expression across multiple cell types, such as epithelial cells, dendritic cells (DC), natural killer (NK) cells, monocytes, and macrophages. This alteration in expression profiles of ADGRL2 and ERO1A suggests a transition toward stemness in the metastatic stage. The acquisition of stemness characteristics is often associated with cancer as it progresses into the metastatic stage. Moreover, the presence of ERO1A in epithelial cells, macrophages, and other immune cells in the metastatic microenvironment hints at a potential role in promoting tumour progression and immune evasion, contributing to the acquisition of stem-like properties by the tumour cells.

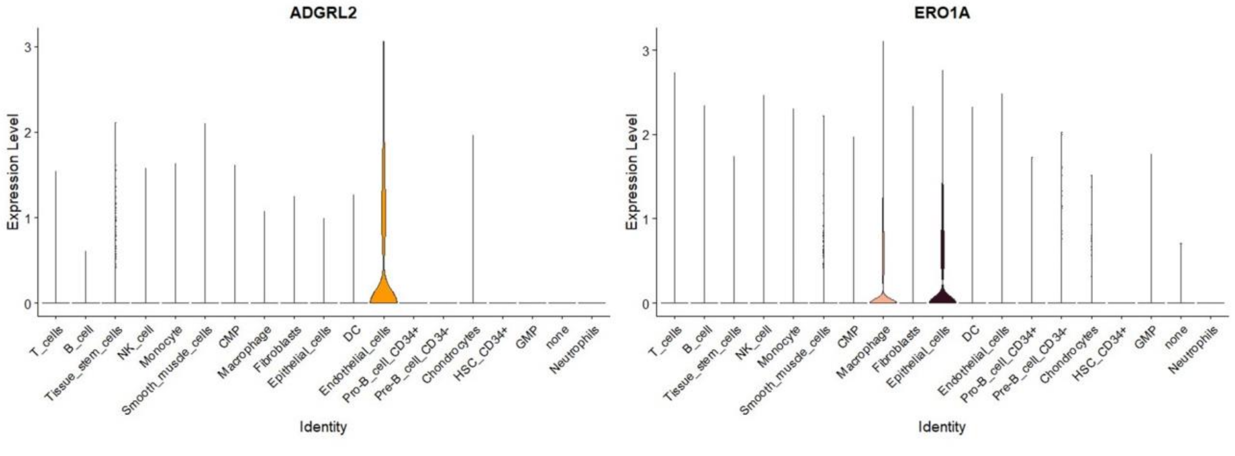

Figure 4.23: Expression of ADGRL2 & ERO1A in Primary Samples

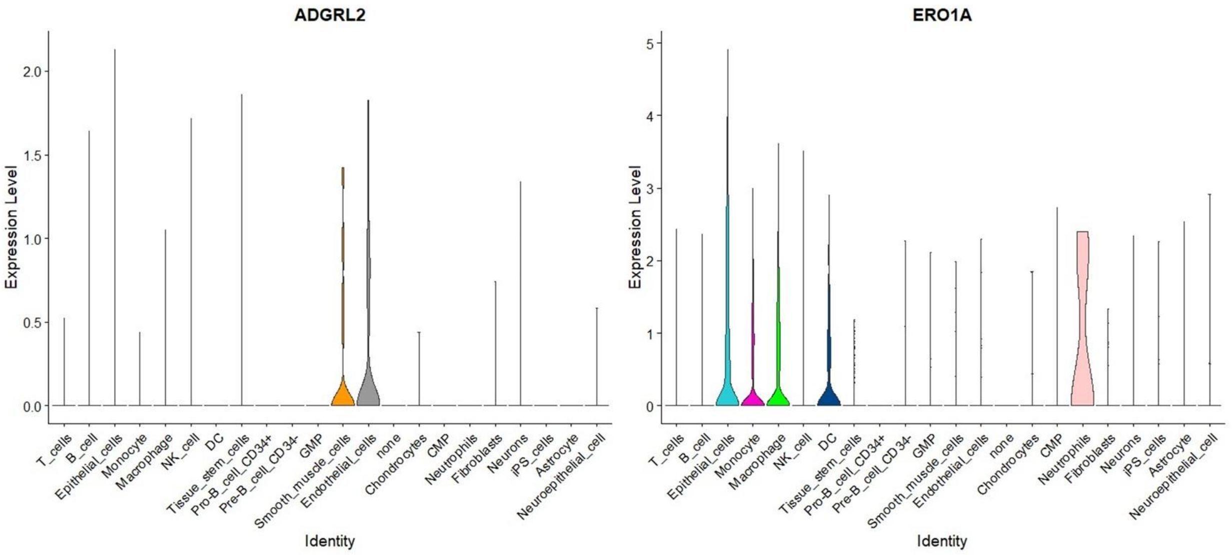

Figure 4.24: Expression of ADGRL2 & ERO1A in Metastatic Samples

4.10 Gene Set Enrichment Analysis

Enrichment analysis is crucial in analyzing single-cell RNA sequencing (scRNA-seq) data, as it helps uncover dysregulated pathways and biological processes within cancer cells. This analysis provides valuable insights into the molecular intricacies underlying cancer development and progression, offering potential therapeutic targets. In this study, we exclusively employed the EnrichR tool, explicitly focusing on the MSigDB hallmark pathways.

We initially selected 438 highly upregulated genes from the primary assay to identify critical pathways encompassing unique and shared genes. Similarly, the metastatic assay yielded 628 genes, meeting the same selection criteria. Subsequently, both sets of genes were subjected to the pathway analysis mentioned earlier, resulting in the identification of enriched pathways for both groups. We chose the top 10 from these enriched pathways based on their statistical significance, as indicated by the p-value. Figures below show these top 10 enriched pathways for the primary and metastatic assays.

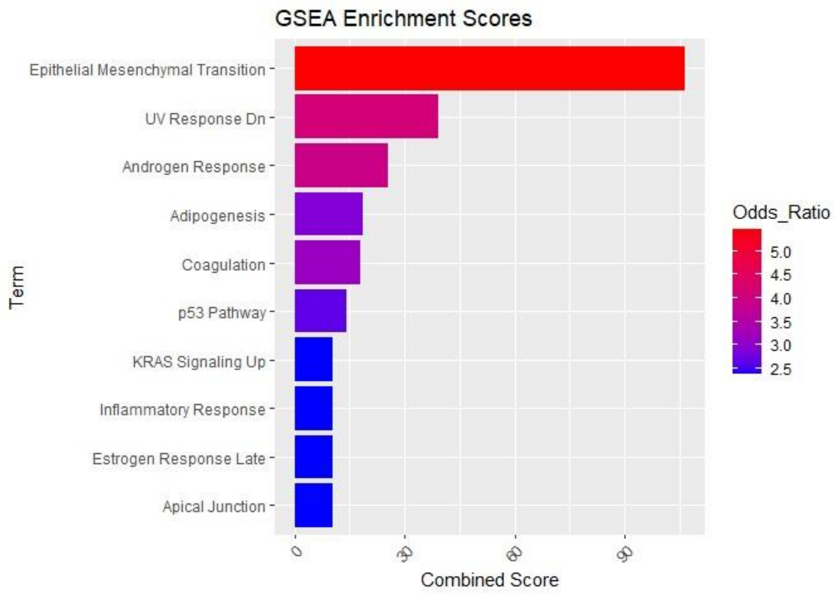

Figure 4.25: Bar plot of enriched terms (Primary samples)

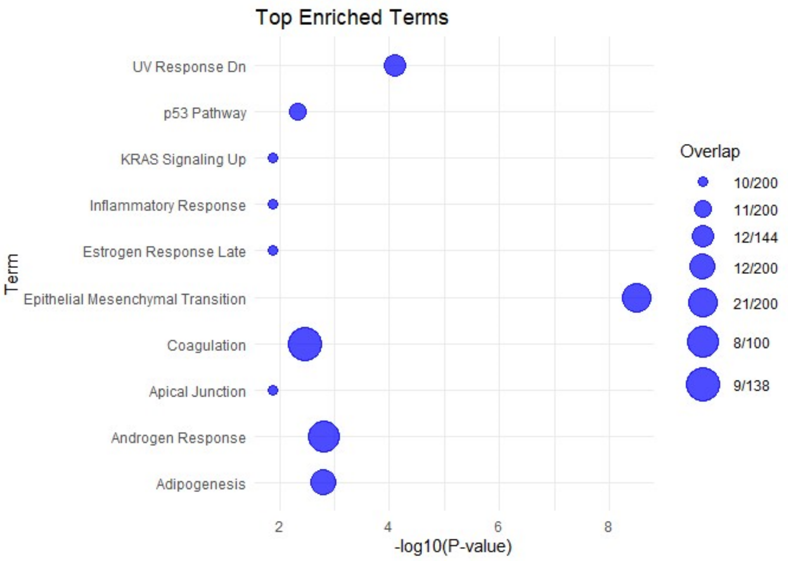

Figure 4.26: Dot plot of enriched terms (Primary samples)

The analysis of the primary and metastatic samples reveals intriguing insights into the roles of various biological pathways in cancer. In the primary samples, as shown in Figures 4-33 and 4-34, the Epithelial-Mesenchymal Transition (EMT) pathway is critically involved in cancer initiation, metastasis, and treatment resistance. It enables cells to transition from stationary to mobile, which is crucial in tumour development[77]. Induced by EMT-convincing transcription factors (EMT-TFs) and regulated by microRNAs, EMT induces epithelial cells to acquire mesenchymal traits, promoting their migration and invasion into surrounding tissues. This process disrupts cell-cell adhesion and alters marker expression, enhancing cancer cell motility and invasiveness, which are vital steps in metastatic spread. Moreover, EMT grants cancer cells stem-like properties, making them resistant to therapy and prone to tumour recurrence. Targeting EMT and its regulatory networks holds promise for impeding cancer progression and improving patient outcomes. The heightened activity of the EMT pathway in primary cancer underscores its significant role in driving tumour aggressiveness and progression, laying the groundwork for metastasis[78]. Conversely, pathways like Androgen Response, Adipogenesis, and Coagulation, while less statistically significant, may contribute to later-stage tumour growth by promoting cellular processes like proliferation, survival, and angiogenesis[79-81]. Interestingly, pathways like ultraviolet (UV) Response Downn, p53 Pathway, and Apical Junction inhibit or prevent tumour growth as potential safeguards against uncontrolled cell growth[82, 83]. 'UV-response-DN' are genes that are downregulated in response to UV light. This downregulation may indicate an inhibitory function against cancer, particularly during the early stages, as they were downregulated in the metastatic stage[84].

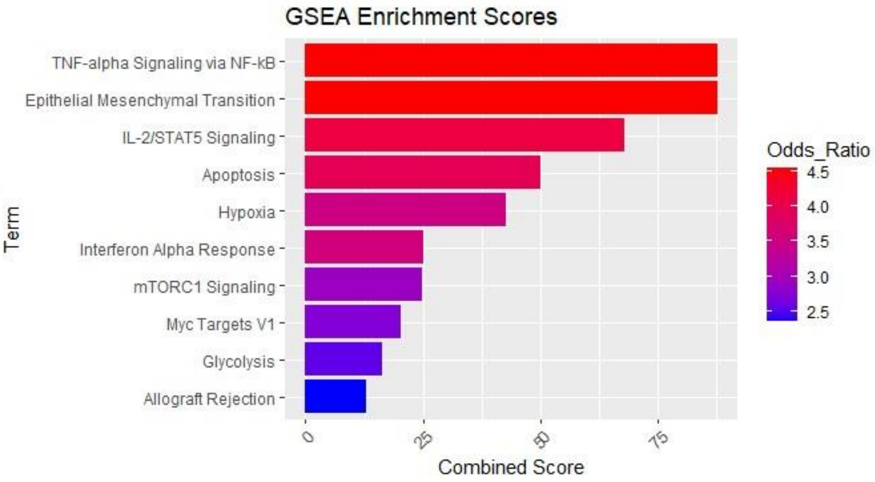

Figure 4.27: Bar plot of enriched terms (Metastatic samples)

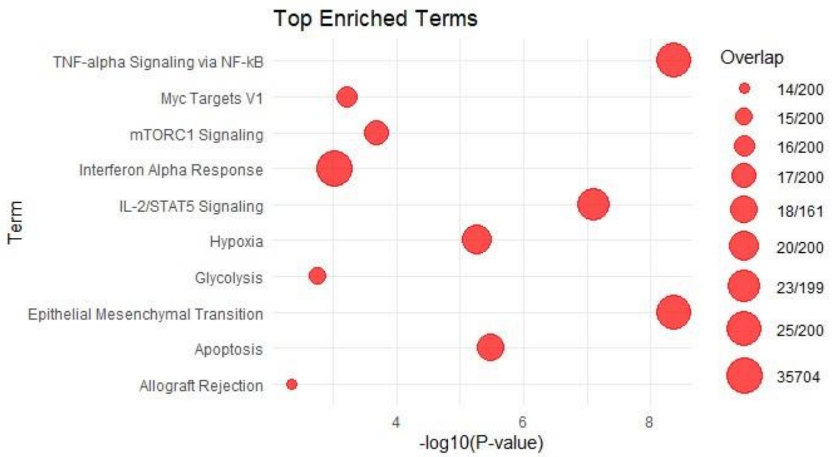

Figure 4.28: Dot plot of enriched terms (Metastatic samples)

In the metastatic samples, pathways associated with EMT, TNF-alpha Signaling, and IL-2/STAT5 Signaling exhibit high combined scores and negative log10 p-values, as shown in Figures 4-35 and 4-36, reflecting their importance in cancer metastasis. EMT drives tumour metastasis by triggering molecular and cellular changes in primary tumour cells, allowing them to spread to distant organs. Various signals like Transforming Growth Factor Beta (TGF-β), cytokines, hypoxia, and Extracellular Matrix (ECM) stiffness initiate EMT, leading to decreased epithelial junction proteins and increased mesenchymal markers. This facilitates the loss of cell-cell adhesion, promotes motility, and enhances invasion capabilities. Key transcription factors like Snail Family Transcriptional Repressor (SNAIL), Twist Family BHLH Transcription Factor (TWIST), and Zinc Finger E-Box Binding Homeobox (ZEB) families drive these changes by suppressing epithelial markers and activating mesenchymal gene expression. Although three tumour-suppressing genes, TIMP Metallopeptidase Inhibitor 3 (TIMP3), Growth Arrest and DNA Damage Inducible Alpha (GADD45A) and Growth Arrest and DNA Damage Inducible Beta (GADD45B) were shown to be upregulated in the metastatic stage, their expression levels were lower than other genes involved in EMT. EMT's reversible nature enables transitions between epithelial and mesenchymal states, with partial EMT aiding metastatic progression. Evidence from cell studies, mouse models, and patient samples highlights EMT's crucial role in metastasis and its potential as a therapeutic target in combating metastatic cancer.[85] Additionally, pathways linked to hypoxia and apoptosis show moderate values, aligning with their recognised roles in metastatic processes. Hypoxia is associated with metastasis by influencing cellular behaviours, while apoptosis regulates cell survival and tissue remodelling in metastatic contexts[86]. Categorising pathways based on their impact on tumour progression reveals tumour-promoting and tumour-suppressing pathways. EMT, IL-2/STAT5 Signaling, and other pathways like hypoxia and glycolysis promote tumour growth, metastasis, and immune evasion[86-89]. On the other hand, apoptotic pathway genes have a complex nature. Most of the genes were involved in tumour progression, and few were involved in tumour suppression[90-93]. Typically, TNF-alpha signalling through Nuclear Factor Kappa-Light-Chain-Enhancer of Activated B Cells (NF-κB) is involved in the immune response. However, in cancer, it is found to cause problems by promoting inflammation and helping cancer cells to go into a metastatic state.[94] Dual Specificity Phosphatase 1 (DUSP1), a gene shown to be upregulated, can be a good guy in this situation. It acts as a stop sign for NF-kB, reducing inflammation and slowing cancer growth[95]. Another gene called G0/G1 Switch 2 (GOS2) is involved in how the body manages fats, but when it comes to cancer, its role is very complex, depending on the type of cell. Generally, it promotes effects in tumorigenesis, but overexpression in estrogen receptor cells decreases cell proliferation[96].

CHAPTER 5: DISCUSSION

Lung cancer is a deadly disease responsible for a large number of cancer-related deaths worldwide.[6] It consists of different types, with small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC) being the most common types with around 90% share.[7, 8] Among NSCLC, lung adenocarcinoma (LUAD) is the most prevalent, with 40% of all the cases not only occurring in smokers but also in non-smokers, particularly women and younger individuals.[11, 12, 16] LUAD is known for its diverse cellular origins and tumour heterogeneity, making it challenging to diagnose and treat effectively.[24] Recent technological advancements, such as single-cell RNA sequencing (scRNA-seq), have allowed scientists to uncover the complex cellular heterogeneity within cancer tumours and their metastatic sites[38]. In. this comprehensive study, single-cell RNA sequencing (scRNA-seq) technology was utilised to elucidate the complex cellular heterogeneity within primary lung adenocarcinomas and their metastatic sites. The study encompassed various data processing steps, including quality control, gene expression normalization, principal component analysis (PCA), and cell clustering using Seurat v4.0. Integration of datasets was conducted utilizing Harmony. Additionally, the analysis included differential gene expression analysis and functional enrichment analysis. Furthermore, a comparative analysis between differentially expressed genes (DEGs) identified through scRNA-seq and data from The Cancer Genome Atlas (TCGA) was undertaken to pinpoint unique genes absent in the TCGA database.

Differential expression analysis uncovered distinct gene expression patterns among different cell types within primary tumour samples. In primary tumour samples, most genes displayed a unique expression profile, with only a few being expressed across multiple cell clusters, including FTL, CSTB, SFTPC, IGLC2, and IGKC. FTL and CSTB are associated with tumor-promoting activities. FTL is found to be involved in gliomas. It is influenced by the hypoxic tumour microenvironment, regulating the epithelial-to-mesenchymal transition (EMT) process and thereby contributing to tumour invasion into metastasis through the AKT/GSK3β/β-catenin signalling pathway[66]. CSTB, known for its implication in hepatocellular carcinoma (HCC), acts as an oncogene. It is involved in tumour progression, invasion, and induction of EMT by modulating key signalling pathways such as MAPK, mTOR, and AKT[67]. Conversely, SFTPC exhibits tumor-suppressing activity in lung adenocarcinoma (LUAD) by inhibiting the PI3K/AKT/mTOR pathway, thus controlling cell proliferation. Low levels of SFTPC are linked to poor prognosis in LUAD patients and increased sensitivity to PD-1 and CTLA-4 antibodies, suggesting its potential as a predictive marker for immunotherapy response[68]. Targeting SFTPC or its downstream pathways may offer new therapeutic avenues for LUAD treatment. IGLC2 and IGKC are the components of the immune system that contribute to antibody production and aid in identifying and neutralizing abnormal cells, boosting the body's defense against cancer[69, 70]. Dot plot analysis of primary samples as shown in figure() explained distinct gene expression patterns among different cell types within the primary tumor microenvironment, indicating the preservation of distinct cellular characteristics and polarity during the primary stage.

In metastatic samples, a consistent expression pattern of most genes, including MGST1, KRT19, CD9, KRT18, GHG3, IGLC2, APCE, SPP1, AKR1B1, SFTPA1, PLCG2, SFTPC, XIST, BNIP3, and SLPI was observed that suggests the occurrence of stemness across various cell types. It implies the upregulated expression of multiple genes in multiple cell types, including genes associated with stem cell properties and a few genes related to tumour suppression and immune response. It indicates a shift in cancer cell biology at the metastatic stage, potentially reflecting a more aggressive phenotype characterized by EMT. CD9, KRT19, KRT18, and SPP1 are a few genes that contribute to this stemness. CD9 expression in pancreatic ductal adenocarcinoma is found to promote cell migration and EMT via the p38 MAPK pathway[71]. KRT19 and KRT18 enhance the survival and metastatic potential of circulating tumour cells (CTCs) by promoting structural integrity and immune evasion mechanisms[72]. Additionally, SPP1 has been highlighted for its critical role in promoting stemness characteristics in pancreatic cancer (PC) cells within the tumour microenvironment[9]. Increased SPP1 levels also correlate with adverse outcomes in LUAD patients, suggesting its potential role as a prognostic indicator. SPP1 fuels LUAD advancement and metastasis by stimulating the EMT pathway, enhancing cancer cell mobility and invasiveness. Moreover, SPP1 modulates the tumour microenvironment by regulating COL11A1 expression, further promoting EMT and metastasis in LUAD cells[74]. Targeting the SPP1-COL11A1 axis can be a promising therapeutic approach to impede LUAD progression and metastasis. Additionally, genes such as APC, BNIP3, IGHG3, and IGLC2 were expressed in multiple cell types independent of their cellular function and are implicated in tumour suppression and inhibition of metastasis[69, 70, 75, 76]. The change in gene expression pattern highlights the evolution of cancer cells during cancer progression toward metastasis. Initially, diverse genes are engaged across various cell types within the tumour. However, as cancer advances to the metastatic phase, a shared group of genes associated with stem cell traits becomes activated across diverse cell types. This transition signifies a pivotal phase wherein cancer cells heighten their aggressiveness and capacity for dissemination to distant sites. Grasping these alterations can potentially refine therapeutic strategies designed for distinct cancer stages.

The gene set enrichment analysis further elucidates the role of various biological pathways in cancer progression. In the primary samples, the Epithelial-Mesenchymal Transition (EMT) pathway stands out with the highest enrichment score and significant negative log10 p-value, emphasizing its critical role in cancer initiation. EMT enables cells to transition from a stationary to a mobile state, an essential step in tumour development[77]. EMT is induced by EMT-transcription factors (EMT-TFs) and regulated by microRNAs that enable epithelial cells to acquire mesenchymal traits, increasing their motility and invasiveness, which are the critical steps in metastasis. This process disrupts cell-to-cell adhesion, altering their marker expression and providing cancer cells with stem-like properties that cause resistance to therapy and recurrence of the cancer.[78] Targeting EMT and its regulatory networks offers the potential for impeding cancer progression and improving outcomes. The heightened activity of EMT in primary cancers underscores its crucial role in driving tumour aggressiveness and metastasis. Conversely, pathways like Androgen Response, Adipogenesis, and Coagulation, while less statistically significant, may contribute to later-stage tumour growth by promoting cellular processes like proliferation, survival, and angiogenesis[79-81]. On the other hand, pathways like UV Response Dn, p53 Pathway, and Apical Junction inhibit or prevent tumour growth as potential safeguards against uncontrolled cell growth[82, 83]. 'UV-response-DN' are genes that are downregulated in response to UV light. This downregulation may indicate an inhibitory function against cancer, particularly during the early stages, as they were downregulated in the metastatic stage[84].

In the metastatic samples, EMT, TNF-alpha Signaling, and IL-2/STAT5 Signaling exhibit high combined scores and negative log10 p-values, as shown in figures(), reflecting their importance in cancer metastasis. EMT drives tumour metastasis by triggering molecular and cellular changes in primary tumour cells, allowing them to spread to distant organs. Various signals like TGF-β, cytokines, hypoxia, and ECM stiffness initiate EMT, leading to decreased epithelial junction proteins and increased mesenchymal markers. This facilitates the loss of cell-cell adhesion, promotes motility, and enhances invasion capabilities. Key transcription factors like SNAIL, TWIST, and ZEB families drive these changes by suppressing epithelial markers and activating mesenchymal gene expression. Although three tumor-suppressing genes, named TIMP3, GADD45A, and GADD45B, were shown to be upregulated in the metastatic stage, their expression levels were lower than other genes involved in EMT. EMT's reversible nature enables transitions between epithelial and mesenchymal states, with partial EMT aiding metastatic progression. Evidence from cell studies, mouse models, and patient samples highlights EMT's crucial role in metastasis and its potential as a therapeutic target in combating metastatic cancer[85]. Additionally, pathways linked to hypoxia and apoptosis show moderate values, aligning with their recognised roles in metastatic processes. Hypoxia is associated with metastasis by influencing cellular behaviours, while apoptosis regulates cell survival and tissue remodelling in metastatic contexts[86]. Categorizing pathways based on their impact on tumor progression reveals tumor-promoting and tumor-suppressing pathways. EMT, IL-2/STAT5 Signaling, hypoxia, glycolysis and others are associated with promoting tumour growth, metastasis, and immune evasion[86-89]. On the other hand, apoptotic pathway genes have a complex nature. Most of the genes were involved in tumour progression, and a few were involved in tumour suppression[90-93]. Usually, TNF-alpha signalling through NF-kB is involved in immune response. However, in cancer, it is found to cause problems by promoting inflammation and helping cancer cells to enter a metastatic state[94]. DUSP1, a gene shown upregulated, can be a positive factor in this situation. It acts as a stop sign for NF-kB, reducing inflammation and slowing down cancer growth[95]. Another gene called GOS2 is involved in how the body manages fats, but when it comes to cancer, its role is very complex, depending on the type of cell. Generally, it promotes tumorigenesis, but overexpression in estrogen receptor cells decreases cell proliferation[96].

By comparing the findings with The Cancer Genome Atlas (TCGA) data, two genes not reported in TCGA were identified: ADGRL2 and ERO1A. ADGRL2's role in restricting stem cell expansion and its potential as a stem-cell-specific negative regulator in colorectal cancer are intriguing findings. Similarly, ERO1A's association with poor prognosis in various cancer types, including gastric, breast, and pancreatic cancer, suggests its significance in tumour metastasis and the epithelial-to-mesenchymal transition. Further investigation into the Akt/mTOR pathway and the potential upstream role of ERO1A adds complexity to its involvement in cancer progression.

The adhesion G-protein-coupled receptor L2 (ADGRL2), also known as Latrophilin-2, has yet to be sufficiently explored in human health despite its known importance in heart development and neural synaptogenesis. One previous study was carried out to investigate the physiological function of ADGRL2 in the colon. Previous findings indicated that ADGRL2 is lost in colorectal cancer (CRC) and downregulated during hyperproliferative recovery from colitis. It is enriched in stem and progenitor cells, and cell growth is arrested in stem cells, but stemness features are a hallmark of cancer. In this study, intestinal-epithelial specific Adgrl2 knockout mice (Adgrl2 IE KO) were used to assess the effect of Adgrl2 on cell proliferation in the colon. These findings support the hypothesis that ADGRL2 modulates epithelial regeneration in the colon, potentially functioning as a stem-cell-specific negative regulator. This research may provide insights into dysregulated growth and repair mechanisms in the colon. It could potentially reveal targets to prevent colitis-associated cancer in patients with inflammatory bowel disease (IBD)[97].

ERO1A was upregulated in the metastatic stage. In various cancer types like gastric, breast, and pancreatic cancer, high expression of ERO1A is linked to poor prognosis. Similar findings were observed in cholangiocarcinoma (CCA) in another study and were shown to be associated with shortened patient survival and pathological and clinical stages. ERO1A's involvement in proliferation and migration was supported by experiments showing that depleting ERO1A inhibited these processes in CCA cells while overexpression enhanced them. It can be proposed that ERO1A might be linked to tumour metastasis and epithelial-mesenchymal transition (EMT) because it is involved in hypoxia, mTORC1 signalling pathway, and glycolysis, both significant factors in cancer progression. It can be hypothesized that ERO1A can be a successful therapeutic target, as said in the previous study above, that depletion of ERO1A suppressed EMT progression, and its overexpression accelerated it in CCA cells[98]. Further investigation into the Akt/mTOR pathway was conducted due to its relevance in cellular processes like proliferation and survival. Depleting ERO1A reduced activation of this pathway, while overexpression had the opposite effect. Though the exact mechanism remains unclear, our study suggested that ERO1A might be positioned upstream of the Akt signalling pathway. In conclusion, ERO1A's role in promoting proliferation and migration, influencing EMT, and interacting with the Akt/mTOR pathway makes it an appealing target for diagnosis and potential therapeutic interventions in LUAD.

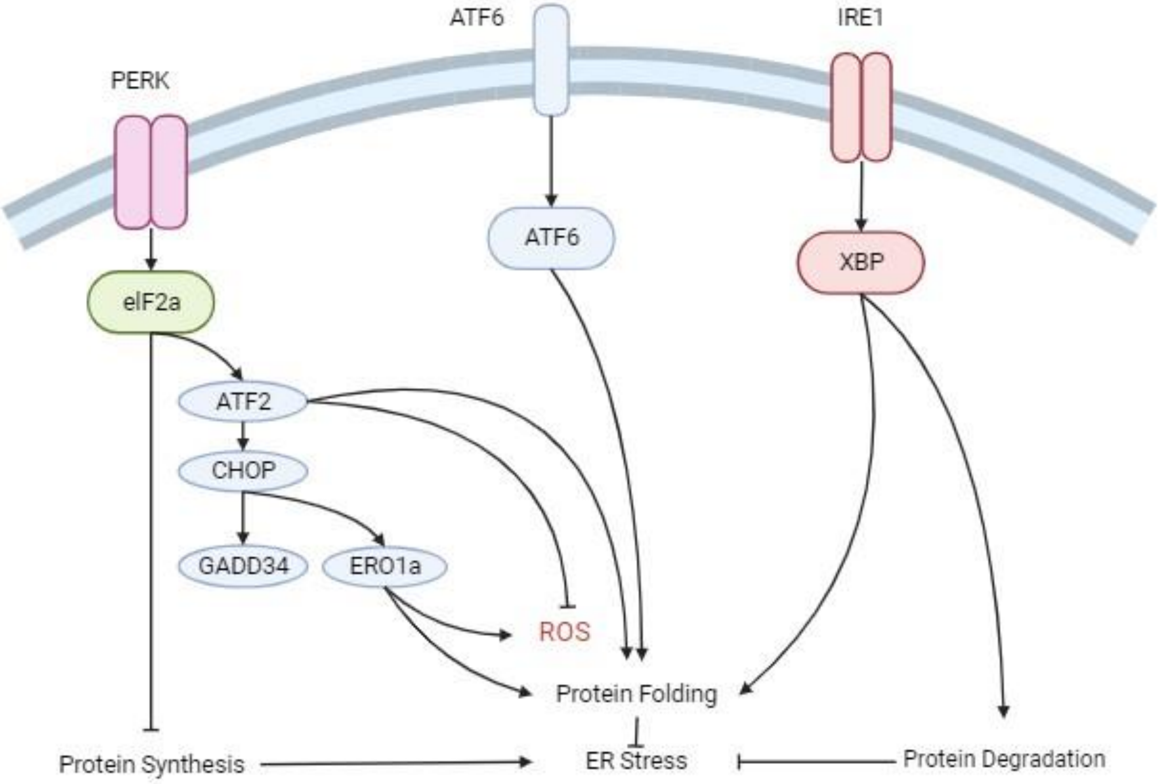

Figure 5.1: Molecular Mechanisms of ERO1a in Angiogenesis

In conclusion, this research provides a comprehensive analysis of differential gene expression and pathway enrichment in cancer progression, highlighting the central role of EMT and identifying potential therapeutic targets. However, it is essential to acknowledge the study's limitations, such as sample size and the need for experimental validation. Future research should address these limitations and further explore the identified genes and pathways, ultimately contributing to cancer diagnosis and treatment advancements.

CHAPTER 6: CONCLUSIONS

In conclusion, our study has delved into the intricate landscape of lung adenocarcinoma (LUAD), shedding light on its diverse and complex nature at both molecular and cellular levels. This prevalent form of lung cancer poses a significant global health challenge, contributing to many cancer-related deaths. Through the application of cutting-edge single-cell RNA sequencing (scRNA-seq) technology, we have uncovered valuable insights into the progression of this aggressive disease. Our analysis revealed remarkable heterogeneity in gene expression patterns among various cell types within the primary tumour microenvironment. This diversity underscores the specialized roles distinct cell types play in supporting or driving tumour growth and maintenance. The gene set enrichment analysis illuminated the importance of various biological pathways in cancer initiation and progression. Comparing our results with The Cancer Genome Atlas (TCGA) data, we uncovered two previously unreported genes, ADGRL2 and ERO1A, with potential roles in cancer progression. ADGRL2 emerged as a stem-cell-specific negative regulator, while ERO1A was associated with poor prognosis in various cancer types and linked to metastasis and EMT. Our findings provide a broader context for understanding the dynamic processes involved in cancer progression. They emphasise the pivotal role of EMT and pinpoint potential therapeutic targets for intervention. Our findings concluded that scRNA-seq is a powerful tool for investigating the molecular mechanisms of cancer development and progression. The ability to characterize the transcriptomic profiles of individual cells provides a more accurate and detailed understanding of the complex biology of tumours. Our study also highlights the importance of studying tumour microenvironments and their interactions with cancer cells, which could lead to identifying novel therapeutic targets for cancer treatment.

In conclusion, this study contributes to understanding the cellular and molecular mechanisms underlying lung adenocarcinomas and their metastatic potential. Our findings provide insights into tumours' complex biology and microenvironments and offer potential avenues for developing new therapeutic strategies.

Future Aspects

To conclude this study, looking at the promising future aspects arising from these findings is essential. The insights gained from our research open several avenues for further exploration and potential breakthroughs in the understanding and treatment of lung adenocarcinoma (LUAD) and cancer in general. First and foremost, the identified genes and pathways hold great promise as targets for novel therapeutic interventions. Developing strategies to modulate EMT and the associated pathways could lead to more effective treatments to inhibit metastasis, a significant challenge in cancer management. Additionally, the potential roles of ADGRL2 and ERO1A in cancer progression provide exciting prospects for targeted therapies and further research into their mechanisms of action. Expanding our understanding of the intricate cellular heterogeneity within primary tumours and their metastatic sites could pave the way for more personalised treatment approaches. By tailoring therapies to the specific characteristics of individual tumours, we may improve treatment outcomes and reduce the side effects associated with conventional cancer treatments. The field of single-cell RNA sequencing (scRNA-seq) continues to advance rapidly. As this technology becomes more accessible and cost-effective, it holds the potential to become a standard tool in cancer research and clinical practice. Future studies with larger sample sizes and more diverse datasets can further refine our knowledge of cancer progression and provide a more comprehensive picture of the disease. Furthermore, comparing our results with data from The Cancer Genome Atlas (TCGA) highlights the importance of collaborative efforts in cancer research. Continued integration of data from various sources and the development of comprehensive databases can accelerate the discovery of novel genes and pathways associated with cancer, ultimately leading to more effective diagnosis and treatment strategies.

In conclusion, our study serves as a crucial steppingstone in the ongoing battle against lung adenocarcinoma and cancer in general. The future holds tremendous potential for translating these findings into improved patient outcomes, and we anticipate further breakthroughs in understanding and managing this challenging disease. As researchers and clinicians continue to work together, building upon these insights, we move closer to realizing more effective, personalised, and targeted approaches to cancer diagnosis and treatment.

REFERENCES

1. Compton, C. and C. Compton, *The Nature and Origins of Cancer.* Cancer: The Enemy from Within: A Comprehensive Textbook of Cancer's Causes, Complexities and Consequences, 2020: p. 1-23.
2. Roser, M. and H. Ritchie, *Cancer. Our World in Data, 2015.* Reference Source.
3. Cruz, C.S.D., L.T. Tanoue, and R.A. Matthay, *Lung cancer: epidemiology, etiology, and prevention.* Clinics in chest medicine, 2011. **32**(4): p. 605-644.
4. Danaei, G., et al., *Causes of cancer in the world: comparative risk assessment of nine behavioural and environmental risk factors.* The Lancet, 2005. **366**(9499): p. 1784-1793.
5. Calvayrac, O., et al., *Molecular biomarkers for lung adenocarcinoma.* European Respiratory Journal, 2017. **49**(4).
6. *Cancer*, in *World Health Organization*. 3 Feb 2022.
7. Litzky, L.A., *Pulmonary sarcomatous tumors.* Archives of pathology & laboratory medicine, 2008. **132**(7): p. 1104-1117.
8. Zorzetto, M., et al., *MET genetic lesions in non-small-cell lung cancer: pharmacological and clinical implications.* Translational Lung Cancer Research, 2012. **1**(3): p. 194.
9. Panagopoulos, N., et al., *Pancoast tumors: characteristics and preoperative assessment.* Journal of thoracic disease, 2014. **6**(Suppl 1): p. S108.
10. Sharma, A., et al., *Advances in lung cancer treatment using nanomedicines.* ACS omega, 2022. **8**(1): p. 10-41.
11. Hutchinson, B.D., et al., *Spectrum of Lung Adenocarcinoma.* Seminars in Ultrasound, CT and MRI, 2019. **40**(3): p. 255-264.
12. Siegel, R.L., K.D. Miller, and A. Jemal, *Cancer statistics, 2018.* CA: a cancer journal for clinicians, 2018. **68**(1): p. 7-30.
13. Li, Q., et al., *Molecular profiling of human non-small cell lung cancer by single-cell RNA-seq.* Genome medicine, 2022. **14**(1): p. 1-18.
14. Torre, L.A., R.L. Siegel, and A. Jemal, *Lung cancer statistics.* Lung cancer and personalized medicine: current knowledge and therapies, 2016: p. 1-19.
15. Hanahan, D., *Hallmarks of cancer: new dimensions.* Cancer discovery, 2022. **12**(1): p. 31-46.
16. Araujo, L.H., et al., *69 - Cancer of the Lung: Non–Small Cell Lung Cancer and Small Cell Lung Cancer*, in *Abeloff's Clinical Oncology (Sixth Edition)*, J.E. Niederhuber, et al., Editors. 2020, Elsevier: Philadelphia. p. 1108-1158.e16.
17. Pillai, R.N., et al., *HER2 mutations in lung adenocarcinomas: A report from the Lung Cancer Mutation Consortium.* Cancer, 2017. **123**(21): p. 4099-4105.
18. Li, X. and C.-Y. Wang, *From bulk, single-cell to spatial RNA sequencing.* International Journal of Oral Science, 2021. **13**(1): p. 36.
19. Luecken, M.D. and F.J. Theis, *Current best practices in single-cell RNA-seq analysis: a tutorial.* Molecular systems biology, 2019. **15**(6): p. e8746.
20. Chen, Z., et al., *Identification of differentially expressed genes in lung adenocarcinoma cells using single-cell RNA sequencing not detected using traditional RNA sequencing and microarray.* Laboratory investigation, 2020. **100**(10): p. 1318-1329.
21. Liang, L., et al., *Integration of scRNA-Seq and bulk RNA-Seq to analyse the heterogeneity of ovarian cancer immune cells and establish a molecular risk model.* Frontiers in oncology, 2021. **11**: p. 711020.

22. Zhang, Y., et al., *Single-cell RNA sequencing in cancer research.* Journal of Experimental & Clinical Cancer Research, 2021. **40**: p. 1-17.
23. Marusyk, A. and K. Polyak, *Tumor heterogeneity: Causes and consequences.* Biochimica et Biophysica Acta (BBA) - Reviews on Cancer, 2010. **1805**(1): p. 105-117.
24. Diaz-Cano, S.J., *Tumor heterogeneity: mechanisms and bases for a reliable application of molecular marker design.* International journal of molecular sciences, 2012. **13**(2): p. 1951-2011.
25. Hegenbarth, J.-C., et al., *Perspectives on bulk-tissue RNA sequencing and single-cell RNA sequencing for cardiac transcriptomics.* Frontiers in Molecular Medicine, 2022. **2**: p. 839338.
26. He, J., et al., *Development of metastasis-associated seven gene signature for predicting lung adenocarcinoma prognosis using single-cell RNA sequencing data.* Mathematical Biosciences and Engineering, 2021. **18**(5): p. 5959-5977.
27. Shukla, S., et al., *Development of a RNA-Seq based prognostic signature in lung adenocarcinoma.* JNCI: Journal of the National Cancer Institute, 2017. **109**(1): p. djw200.
28. Li, Y., et al., *RNA-seq analysis of lung adenocarcinomas reveals different gene expression profiles between smoking and nonsmoking patients.* Tumor Biology, 2015. **36**: p. 8993-9003.
29. Liang, J., J. Lv, and Z. Liu, *Identification of stage-specific biomarkers in lung adenocarcinoma based on RNA-seq data.* Tumor Biology, 2015. **36**: p. 6391-6399.
30. Chan, J.M., et al., *Single cell profiling reveals novel tumor and myeloid subpopulations in small cell lung cancer.* bioRxiv, 2020: p. 2020.12. 01.406363.
31. Kim, N., et al., *Single-cell RNA sequencing demonstrates the molecular and cellular reprogramming of metastatic lung adenocarcinoma.* Nature communications, 2020. **11**(1): p. 2285.
32. Lu, T., et al., *Single-cell transcriptome atlas of lung adenocarcinoma featured with ground glass nodules.* Cell discovery, 2020. **6**(1): p. 69.
33. Saito, M., et al., *Gene aberrations for precision medicine against lung adenocarcinoma.* Cancer science, 2016. **107**(6): p. 713-720.
34. Xue, Q., et al., *Promising immunotherapeutic targets in lung cancer based on single-cell RNA sequencing.* Frontiers in Immunology, 2023. **14**: p. 1148061.
35. Maynard, A., et al., *Therapy-induced evolution of human lung cancer revealed by single-cell RNA sequencing.* Cell, 2020. **182**(5): p. 1232-1251. e22.
36. Guo, X., et al., *Global characterization of T cells in non-small-cell lung cancer by single-cell sequencing.* Nature medicine, 2018. **24**(7): p. 978-985.
37. Yang, Q., et al., *Single-cell RNA sequencing reveals the heterogeneity of tumor-associated macrophage in non-small cell lung cancer and differences between sexes.* Frontiers in Immunology, 2021. **12**: p. 756722.
38. Wang, J., et al., *Single-cell RNA sequencing reveals novel gene expression signatures of trastuzumab treatment in HER2+ breast cancer: a pilot study.* Medicine, 2019. **98**(26).
39. Chandrashekar, D.S., et al., *UALCAN: An update to the integrated cancer data analysis platform.* Neoplasia, 2022. **25**: p. 18-27.
40. Chandrashekar, D.S., et al., *UALCAN: a portal for facilitating tumor subgroup gene expression and survival analyses.* Neoplasia, 2017. **19**(8): p. 649-658.
41. Giorgi, F.M., C. Ceraolo, and D. Mercatelli, *The R language: an engine for bioinformatics and data science.* Life, 2022. **12**(5): p. 648.
42. Racine, J.S., *RStudio: a platform-independent IDE for R and Sweave*. 2012, JSTOR.

43. Hafemeister, C. and R. Satija, *Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression.* Genome biology, 2019. **20**(1): p. 296.

44. Chen, E.Y., et al., *Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool.* BMC Bioinformatics, 2013. **14**: p. 128.

45. Kuleshov, M.V., et al., *Enrichr: a comprehensive gene set enrichment analysis web server 2016 update.* Nucleic Acids Res, 2016. **44**(W1): p. W90-7.

46. Xie, Z., et al., *Gene Set Knowledge Discovery with Enrichr.* Current Protocols, 2021. **1**(3): p. e90.

47. Lu, J., et al., *scRNA-seq data analysis method to improve analysis performance.* IET nanobiotechnology, 2023.

48. Caron, M., et al., *Single-cell analysis of childhood leukemia reveals a link between developmental states and ribosomal protein expression as a source of intra-individual heterogeneity.* Scientific Reports, 2020. **10**(1): p. 8079.

49. Choudhary, S. and R. Satija, *Comparison and evaluation of statistical error models for scRNA-seq.* Genome biology, 2022. **23**(1): p. 27.

50. Babcock, B.R., et al., *Data matrix normalization and merging strategies minimize batch-specific systemic variation in scRNA-seq data.* bioRxiv, 2021: p. 2021.08. 18.456898.

51. Townes, F.W., et al., *Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model.* Genome biology, 2019. **20**: p. 1-16.

52. Becht, E., et al., *Dimensionality reduction for visualizing single-cell data using UMAP.* Nature biotechnology, 2019. **37**(1): p. 38-44.

53. Peng, L., et al., *Single-cell RNA-seq clustering: datasets, models, and algorithms.* RNA biology, 2020. **17**(6): p. 765-783.

54. Kobak, D. and P. Berens, *The art of using t-SNE for single-cell transcriptomics.* Nature Communications, 2019. **10**(1): p. 5416.

55. Sun, S., et al., *Accuracy, robustness and scalability of dimensionality reduction methods for single-cell RNA-seq analysis.* Genome biology, 2019. **20**(1): p. 1-21.

56. Korsunsky, I., et al., *Fast, sensitive and accurate integration of single-cell data with Harmony.* Nature Methods, 2019. **16**(12): p. 1289-1296.

57. Chen, G., B. Ning, and T. Shi, *Single-cell RNA-seq technologies and related computational data analysis.* Frontiers in genetics, 2019. **10**: p. 317.

58. Sun, Y. and P. Qiu, *Domain adaptation for supervised integration of scRNA-seq data.* Communications Biology, 2023. **6**(1): p. 274.

59. Prazanowska, K.H. and S.B. Lim, *An integrated single-cell transcriptomic dataset for non-small cell lung cancer.* Scientific Data, 2023. **10**(1): p. 167.

60. Chowdhury, H.A. *Effective clustering of scRNA-seq data to identify biomarkers without user input*. in *Proceedings of the AAAI Conference on Artificial Intelligence*. 2021.

61. Jabato, F.M., et al., *Gene expression analysis method integration and co-expression module detection applied to rare glucide metabolism disorders using ExpHunterSuite.* Sci Rep, 2021. **11**(1): p. 15062.

62. Weinstein, J.N., et al., *The Cancer Genome Atlas Pan-Cancer analysis project.* Nat Genet, 2013. **45**(10): p. 1113-20.

63. Holland, C.H., et al., *Robustness and applicability of transcription factor and pathway analysis tools on single-cell RNA-seq data.* Genome biology, 2020. **21**: p. 1-19.

64. Liberzon, A., et al., *The Molecular Signatures Database (MSigDB) hallmark gene set collection.* Cell Syst, 2015. **1**(6): p. 417-425.

65. Shaath, H., et al., *Single-cell long noncoding RNA (lncRNA) transcriptome implicates MALAT1 in triple-negative breast cancer (TNBC) resistance to neoadjuvant chemotherapy.* Cell Death Discovery, 2021. **7**(1): p. 23.

66. Liu, J., et al., *Hypoxia induced ferritin light chain (FTL) promoted epithelia mesenchymal transition and chemoresistance of glioma.* Journal of Experimental & Clinical Cancer Research, 2020. **39**(1): p. 137.

67. Zhu, W., et al., *CSTB accelerates the progression of hepatocellular carcinoma via the ERK/AKT/mTOR signaling pathway.* Heliyon, 2024. **10**(1): p. e23506.

68. Zuo, B., et al., *Abnormal low expression of SFTPC promotes the proliferation of lung adenocarcinoma by enhancing PI3K/AKT/mTOR signaling transduction.* Aging (Albany NY), 2023. **15**(21): p. 12451-12475.

69. Chang, Y.T., et al., *A Novel IGLC2 Gene Linked With Prognosis of Triple-Negative Breast Cancer.* Front Oncol, 2021. **11**: p. 759952.

70. Schmidt, M., et al., *Immunoglobulin kappa chain as an immunologic biomarker of prognosis and chemotherapy response in solid tumors.* Oncoimmunology, 2012. **1**(7): p. 1156-1158.

71. Shao, S., et al., *The role of Tetraspanins in digestive system tumor development: update and emerging evidence.* Frontiers in Cell and Developmental Biology, 2024. **12**.

72. Takan, I., et al., *"In the light of evolution:" keratins as exceptional tumor biomarkers.* PeerJ, 2023. **11**: p. e15099.

73. Nallasamy, P., et al., *Pancreatic Tumor Microenvironment Factor Promotes Cancer Stemness via SPP1–CD44 Axis.* Gastroenterology, 2021. **161**(6): p. 1998-2013.e7.

74. Yi, X., et al., *SPP1 facilitates cell migration and invasion by targeting COL11A1 in lung adenocarcinoma.* Cancer Cell International, 2022. **22**(1): p. 324.

75. Zhang, L. and J.W. Shay, *Multiple roles of APC and its therapeutic implications in colorectal cancer.* JNCI: Journal of the National Cancer Institute, 2017. **109**(8): p. djw332.

76. Yu, Q., et al., *BNIP3 as a potential biomarker for the identification of prognosis and diagnosis in solid tumours.* Molecular Cancer, 2023. **22**(1): p. 143.

77. Roche, J., *The Epithelial-to-Mesenchymal Transition in Cancer.* Cancers (Basel), 2018. **10**(2).

78. Islam, M.S., et al., *The role of inflammations and EMT in carcinogenesis.* Advances in Cancer Biology - Metastasis, 2022. **5**: p. 100055.

79. Aurilio, G., et al., *Androgen Receptor Signaling Pathway in Prostate Cancer: From Genetics to Clinical Applications.* Cells, 2020. **9**(12).

80. Oshi, M., et al., *Adipogenesis in triple-negative breast cancer is associated with unfavorable tumor immune microenvironment and with worse survival.* Scientific Reports, 2021. **11**(1): p. 12541.

81. Bluff, J.E., et al., *Tissue factor, angiogenesis and tumour progression.* Breast Cancer Res, 2008. **10**(2): p. 204.

82. Vazquez, A., et al., *The genetics of the p53 pathway, apoptosis and cancer therapy.* Nature Reviews Drug Discovery, 2008. **7**(12): p. 979-987.

83. Takahashi, H., et al., *Gastric cancer with enhanced apical junction pathway has increased metastatic potential and worse clinical outcomes.* Am J Cancer Res, 2022. **12**(5): p. 2146-2159.

84. Hiller, T.W., et al., *Solar ultraviolet radiation and breast cancer risk: a systematic review and meta-analysis.* Environmental Health Perspectives, 2020. **128**(1): p. 016002.

85. Yeung, K.T. and J. Yang, *Epithelial–mesenchymal transition in tumor metastasis.* Molecular Oncology, 2017. **11**(1): p. 28-39.

86. Chen, Z., et al., *Hypoxic microenvironment in cancer: molecular mechanisms and therapeutic interventions.* Signal Transduction and Targeted Therapy, 2023. **8**(1): p. 70.
87. Halim, C.E., et al., *Involvement of STAT5 in Oncogenesis.* Biomedicines, 2020. **8**(9): p. 316.
88. Fadaka, A., et al., *Biology of glucose metabolization in cancer cells.* Journal of Oncological Sciences, 2017. **3**(2): p. 45-51.
89. Tian, T., X. Li, and J. Zhang, *mTOR Signaling in Cancer and mTOR Inhibitors in Solid Tumor Targeting Therapy.* Int J Mol Sci, 2019. **20**(3).
90. Pfeffer, C.M. and A.T.K. Singh, *Apoptosis: A Target for Anticancer Therapy.* Int J Mol Sci, 2018. **19**(2).
91. Fan, C., et al., *PRF1 is a prognostic marker and correlated with immune infiltration in head and neck squamous cell carcinoma.* Transl Oncol, 2021. **14**(4): p. 101042.
92. Li, M., et al., *The arginine methyltransferase PRMT5 and PRMT1 distinctly regulate the degradation of anti-apoptotic protein CFLAR L in human lung cancer cells.* Journal of Experimental & Clinical Cancer Research, 2019. **38**: p. 1-13.
93. Cruceriu, D., et al., *The dual role of tumor necrosis factor-alpha (TNF-α) in breast cancer: molecular insights and therapeutic approaches.* Cellular Oncology, 2020. **43**(1): p. 1-18.
94. Wu, Y. and B.P. Zhou, *TNF-α/NF-κB/Snail pathway in cancer cell migration and invasion.* British Journal of Cancer, 2010. **102**(4): p. 639-644.
95. Gil-Araujo, B., et al., *Dual specificity phosphatase 1 expression inversely correlates with NF-κB activity and expression in prostate cancer and promotes apoptosis through a p38 MAPK dependent mechanism.* Mol Oncol, 2014. **8**(1): p. 27-38.
96. Corbet, A.K., et al., *G0S2 promotes antiestrogenic and pro-migratory responses in ER+ and ER- breast cancer cells.* Translational Oncology, 2023. **33**: p. 101676.
97. Bucar, E.B., et al., *Loss of Adhesion G-Protein-Coupled Receptor L2 Expression Impacts Colonic Epithelial Proliferation.* The FASEB Journal, 2022. **36**.
98. Yan, W., et al., *Expression of endoplasmic reticulum oxidoreductase 1-α in cholangiocarcinoma tissues and its effects on the proliferation and migration of cholangiocarcinoma cells.* Cancer Manag Res, 2019. **11**: p. 6727-6739.