Multi-scale Network Inference Framework Using Single Cell Transcriptomic Data

by

YİĞİT ŞİBAL

A Dissertation Submitted to the
Graduate School of Sciences and Engineering
in Partial Fulfillment of the Requirements for
the Degree of

Master of Science

in

Computational Sciences and Engineering

KOÇ
ÜNİVERSİTESİ

August 13, 2024

Multi-scale Network Inference Framework Using Single Cell Transcriptomic Data

Koç University

Graduate School of Sciences and Engineering

This is to certify that I have examined this copy of a master's thesis by

YİĞİT ŞİBAL

and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Committee Members:

_____

Assoc. Prof. Nurcan Tunçbağ (Advisor)

_____

Assist. Prof. Murat Kuşçu

_____

Assoc. Prof. Beste Turanlı

Date: 13.08.2024

To the City of Istanbul,

ABSTRACT

Multi-scale Network Inference Framework Using Single Cell Transcriptomic Data

Yiğit Şibal

Master of Science in Computational Sciences and Engineering

August 13, 2024

Cancer, characterized by uncontrolled cell proliferation, tumor invasion and aberrant signaling, involves complex intercellular and intracellular interactions within tumors. To address this complexity, we developed a framework to integrate single cell transcriptomic data with interactomes and model tumors as multi-scale networks. Our approach integrates reconstructed gene regulatory networks (GRN), signaling networks and receptor-ligand interactions to model intracellular and intercellular communication. We applied this framework to a publicly available single-cell transcriptomic data from LUAD patient tumors and constructed a multi-scale network containing GRNs and signaling networks specific to the cell subtypes and cell-cell interactions across subtypes within the primary tumor. Our analysis identified differentially expressed genes between normal and cancer cells, and between primary tumors and brain metastases, highlighting the critical role of receptor-ligand crosstalk in oncogenic signaling. Our approach revealed the role of several receptor-ligand crosstalk between different cell subtypes that can mediate oncogenic signaling. Additionally, our results suggest that the abnormal expression of transcription factors in cancer cells crucially influences oncogenesis and tumor suppression. Our framework is easily adaptable to various single cell transcriptomic data as well as other omic data types. Overall, our framework uses multi-scale network inference from single-cell transcriptomics to enhance our understanding of intracellular pathways and cell-type interactions, integrating multiple biological network layers.

ÖZETÇE

Tek Hücre Transkriptomik Verisinden Çok Ölçekli Ağ Çıkarım Yöntemi

Yiğit Şibal

Hesaplamalı Bilimler ve Mühendislik, Yüksek Lisans

13 Ağustos 2024

Kontrolsüz hücre bölünmesi, tümör işgali ve anormal işaret iletimleri ile tanımlanan kanser, tümörlerin bünyesindeki karmaşık hücre içi ve hücrelerarası etkileşimleri de içermektedir. Bu karmaşıklığı ele almak amacıyla, tek hücre transkriptomik verilerini interaktomlarla entegre etmek ve tümörleri çok ölçekli ağlar olarak modellemek için bir çerçeve geliştirdik. Yaklaşımımız, hücre içi ve hücrelerarası iletişimi modellemek için yeniden yapılandırılmış gen düzenleyici ağları (GRN), sinyal ağlarını ve reseptör-ligand etkileşimlerini entegre etmektedir. Bu çerçeveyi akciğer adenokarsinoma (LUAD) hasta tümörlerinden elde edilen herkese açık tek hücreli transkriptomik verilere uyguladık ve hücre alt tiplerine özgü gen düzenleyici ağları, sinyal ağlarını ve birincil tümör içindeki alt tipler arasında hücre-hücre etkileşimlerini içeren çok ölçekli bir ağ oluşturduk. Analizimiz, normal hücreler ve kanser hücreleri arasında ve birincil tümörler ile beyin metastazları arasında farklı şekilde ifade edilen genleri tanımlayarak, onkojenik işaret iletimlerinde reseptör-ligand çapraz girişiminin kritik rolünü vurgulamıştır. Yaklaşımımız, onkojenik işaret iletimine aracılık edebilen farklı hücre alt tipleri arasında çeşitli reseptör-ligand çapraz girişiminin rolünü ortaya koymuştur. Genel olarak çerçevemiz, çoklu biyolojik ağ katmanlarını entegre ederek hücre içi yolaklar ve hücre tipi etkileşimleri hakkındaki anlayışımızı geliştirmek için tek hücreli transkriptomiklerden çok ölçekli ağ çıkarımından yararlanır.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to those who have supported me in my journey in completing this master's thesis.

First and foremost, I extend my sincere thanks to my advisor, Assoc. Prof. Nurcan Tunçbağ for her invaluable guidance, mentorship, and unwavering support. Her expertise, encouragement and patience have been instrumental in this research and in my academic development.

I am also grateful to my thesis committee members Assist. Prof. Murat Kuşçu and Assoc. Prof. Beste Turanlı for their insightful feedback and constructive criticism. Your expertise and perspectives have greatly enriched my work and contributed to its development.

I would like to thank the TÜBİTAK ARDEB 1001 program for their funding of project 121E245 and TÜBİTAK BİDEB 2247 program for the research support of project 121C292.

To my colleagues, İdil İlayda Duran, Aslı Dansık, Sertan Ali Dikli, Ülkem Kasapoğlu, Ekin Su Erdem, Cansu Demirel, Berkay Günay, Serra Doğanata, Can Dedeköy and Esra Başaran, I offer my heartfelt thanks. Your collaboration and moral support have made the research process both enjoyable and rewarding. I have learned much from our discussions and shared experiences.

A special thank you to my close friends Ahmet Aydemir, Melis Tanıl, Zeynep Başak Canbay, Zülal Çelik, Ceren Uzun and Enes Sefa Ayar for your constant encouragement, understanding, and companionship. In addition, I would like to thank my dearest friends Oğuzkağan Er, Mahmut Caner Deniz, Kerem Öncel and Emre Hacıyunus for supporting me under any circumstances since my high-school years. Your belief in me has been a source of strength throughout this journey.

Finally, I wish to express my deepest appreciation to my parents, Esra and Bora Şibal. Your unconditional love, support, and sacrifices have been the foundation of my achievements. I am forever grateful for your strong belief in my potential and your constant encouragement.

Thank you all for being an integral part of this journey. This thesis would not have been possible without your support and contributions.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

ANGPT1	Angiopoietin 1
API	Application Programming Interface
APLP2	Amyloid Precursor-like-Protein 2
AQP5	Aquaporin 5
ATAC-seq	Assay for Transposase-Accessible Chromatin sequencing
AXL	AXL Receptor Tyrosine Kinase
CD36	CD36 Molecule
CRE	Cis-regulatory element
DPT	Diffusion Pseudotime
EC	Endothelial cell
EDNRB	Endothelin Receptor Type B
ELF3	E74-like ETS Transcription Factor 3
FDR	False Discovery Rate
FOXP1	Forkhead Box P1
GAS6	Growth Arrest-Specific 6
GGN	Graphlet-Guided Network
GRN	Gene Regulatory Network
GSEA	Gene Set Enrichment Analysis
HMGB3	High Mobility Group Protein B-3
HOPX	HOP Homeobox
HOXA10	Homeobox A10
ID2	Inhibitor of DNA Binding 2
IGFBP3	Insulin-like Growth Factor Binding Protein-3
IGFBP3	Insulin-Like Growth Factor Binding Protein 3
kNN	k-nearest neighbors
LCN2	Lipocalin 2
LUAD	Lung Adenocarcinoma
MAPK	Mitogen-Activated Protein Kinase
MER	MER Proto-Oncogene, Tyrosine Kinase
MYC	MYC Proto-Oncogene, bHLH Transcription Factor
NFKBIA	NFKB Inhibitor Alpha

NTA	Network Topology-Based Analysis
ORA	Over-Representation Analysis
PCA	Principal Component Analysis
PCSF	Prize-Collecting Steiner Forest
PPR	Personalized PageRank
RNA-seq	RNA sequencing
RWR	Random Walk with Restart
S100A4	S100 Calcium-Binding Protein A4
scATAC-seq	Single-cell Assay for Transposase-Accessible Chromatin sequencing
scRNA-seq	Single-cell RNA sequencing
SOX18	SRY-Box Transcription Factor 18
TAM	TYRO3, AXL, MER receptor tyrosin kinases
TCF4	Transcription Factor 4
TF	Transcription Factor
TGF-beta	Transforming growth factor beta
tLung	Lung Adenocarcinoma Tumor
TME	Tumor microenvironment
TMEM176A	Transmembrane Protein 176A
TMEM176B	Transmembrane Protein 176B
TMEM219	Transmembrane Protein 219
TMSB10	Thymosin β10
TNFRSF21	Tumor Necrosis Factor Receptor Superfamily Member 21
TOB1	Transducer of ERBB2,1
TPPP3	Tubulin Polymerization Promoting Protein Family Member 3
tS1	Tumor Cell State 1
tS2	Tumor Cell State 2
tS3	Tumor Cell State 3
tSNE	t-Distributed Stochastic Neighbor Embedding
TYRO3	TYRO3 Protein Tyrosine Kinase
ZFP36	ZFP36 Ring Finger Protein

Chapter 1:
INTRODUCTION

Cancer represents a complex systems disorder, characterized by numerous factors including genomic mutations, changes in transcription patterns, alterations in single-cell behaviors, and disruptions in cell-cell communication. Within this intricate framework, these elements contribute to the progression and development of the disease. It underscores the dynamic interplay among various biological processes, highlighting the need for comprehensive understanding and targeted interventions to combat cancer effectively. Tumors are the main habitats of cancer cells that are present with non-malignant cells and all these cells are defined as tumor microenvironment (TME) (Mao et al., 2021). A key characteristic of cancer is the change in cellular metabolism. Traditionally, cancer research has concentrated on cancer cells themselves, often overlooking the interactions between cancer cells and the surrounding cells in the tumor microenvironment (Martinez-Outschoorn et al., 2014).

Recently, single-cell RNA sequencing (scRNA-seq) technology has significantly advanced our understanding of biological systems. It has enabled the exploration of the cellular diversity in various organisms and uncovered previously unnoticed cellular populations. The significant potential of this technology has empowered computational biologists to develop a variety of analysis tools. The lack of standardization due to the field's relative youth poses a barrier for beginners to single-cell data analysis. (Luecken & Theis, 2019) In the framework, which is the backbone of this thesis, includes scRNA-seq analysis and uniquely combines current optimal practices to establish a groundwork for future standardization of analysis.

In this thesis, we aim to construct a framework that reconstructs intracellular and intercellular networks that consist of transcription factors, genes, proteins and receptor and ligand pairs among cancer and other cells in the primary tumor samples. To achieve this, we used publicly available lung adenocarcinoma samples from various patients and reconstructed multi-scale networks using our framework to have a better understanding of underlying mechanisms of communication of different cell types within the tumor samples. We used the Prize Collecting Steiner Forest (PCSF) Algorithm implemented in

Omics Integrator software to reconstruct intracellular signaling network (Tuncbag et al., 2013).

In Chapter 2, we provide a comprehensive literature review on single-cell RNA sequencing data analysis, cell type identification, gene regulatory network construction, network modeling algorithms and the integration of these components with receptor-ligand interactions among different cell subtypes in the tumor samples. In addition, we reviewed the recently released tools and their methodologies for the network inference using single cell transcriptomic data.

In Chapter 3, we included the methodology of our framework for multi-scale network inference from single cell transcriptomics. In addition, we presented the details of our pipeline involving single-cell RNA sequencing data analysis, gene regulatory network construction and integration of GRNs with the receptor-ligand interaction information to reveal intracellular signaling pathways and crosstalk between different cell subtypes.

In Chapter 4, we demonstrated how our results would contribute to integration of intracellular network inference to receptor-ligand interaction information and eventually cell-cell interactions. In that manner, we constructed intracellular networks of different cell types in our dataset and connected them via corresponding receptor-ligand complexes they possess. Furthermore, we assessed the biological relevance of these receptor-ligand interactions among different cell types through marker gene identification and over-representation analysis. Furthermore, we reflected on the differences in terms of protein coverage in healthy and cancer cells. In addition, we presented our findings on the differentially expressed genes and transcription factors in our resulting networks and indicated their biological relevance to oncogenic processes.

In Chapter 5, we revisited our research question toward developing a novel framework to construct a multi-scale tumor network and summarized our key results by comparing them with the previous studies. Finally, we discussed how our approach towards modeling a tumor from the single-cell transcriptomic data can be improved to better reflect biological processes including increasing the number of layers, patient samples and single-cells, as well as the depth of the data. We also presented the limitations of our final framework.

Chapter 2:
LITERATURE REVIEW

2.1 Single-cell Omics Technologies

Bulk RNA sequencing (RNA-seq) has been extensively utilized over the past decade to examine gene expression patterns at the population level (Thind et al., 2021). The introduction of single-cell RNA sequencing (scRNA-seq) technology provides significant opportunities to investigate gene expression at the individual cell level. Presently, scRNA-seq is preferred for addressing key biological questions including but not limited to cell heterogeneity, drug treatment and cancer research, which involves only a small number of cells (Schäfer et al., 2024). In contrast, bulk RNA-seq predominantly captures the averaged gene expression across thousands of cells, thereby obscuring individual cellular differences. Recently, scRNA-seq has been employed in numerous species, with a particular focus on various human tissues such as normal and cancer tissues, whereas this focus has led to a variability in gene expression among different types of cells (G. Chen et al., 2019).

The advancement of single-cell omics technologies facilitates the creation of techniques for characterizing cell identity. An instance of this is the combination of single-cell RNA sequencing (scRNA-seq) with pooled patient samples, which holds significant potential for studying the genetic regulation of cell identity (N. Kim et al., 2020). Due to the increase in number of single cells in the pooled samples, discriminating the cell type identities becomes relatively more accurate and easier as the computational complexity increases, on the other hand. However, its applicability is limited in various biological contexts. Recently developed computational methods aim to perform single-cell behavior simulations, yet in many of these approaches experimental perturbation data to train models is a requirement, thereby restricting their scalability and utility (Luecken & Theis, 2019).

In addition, single-cell omics datasets can reveal cellular diversity, enabling the reconstruction of intricate differentiation pathways through computational approaches (Raj et al., 2018). Although trajectory inference techniques have unveiled many

biological findings, they usually target snapshot single-cell RNA sequencing data and lack the capacity to integrate other information that is crucial for comprehending cell-state dynamics (Weiler et al., 2024).

Eukaryotic genomes are organized into chromatin in a hierarchical manner, and this packaging is crucial for gene regulation (Mellor, 2005). High throughput approaches that analyze chromatin accessibility, nucleosome positioning, transcription factor occupancy throughout the genome have enhanced valuable information about the epigenetics of the chromatin structure. ATAC-seq enables the simultaneous analysis of TF occupancy, positioning of nucleosome at regulatory locations, and genome-wide chromatin accessibility. To achieve this, it examines the locations of insertions and their lengths, which occurs simultaneously with the transposition (Grandi et al., 2022). ATAC-seq might apply to investigate specific cell subgroups at various stages of aging and development, as well as during disease progression such as cancer, autoimmune conditions, and neuropsychiatric disorders (Buenrostro et al., 2013).

In parallel to the accumulation of single-cell transcriptomic data, new tools and frameworks have been developed for the analysis and processing of this data. Single-cell RNA sequencing analysis has become essential in understanding cellular heterogeneity and dynamics. Several state-of-the-art tools are instrumental in this field, including SCANPY (Wolf et al., 2018), Seurat (Hao et al., 2023), and Monocle3 (Cao et al., 2019). SCANPY is a framework for large-scale single-cell RNA sequencing data analysis for preprocessing, visualization, clustering, and differential gene expression. In addition, Seurat, a widely used R package, operates in multi-scale single-cell data integration, which combines scRNA-seq with other data types such as spatial transcriptomics, chromatin accessibility and protein expression, and possesses robust methods for clustering, trajectory inference, and differential expression analysis, as well. Furthermore, Monocle3 focuses on single-cell trajectory reconstruction, the time-dependent and condition-dependent progression and differentiation of cellular states, empowering our understanding in developmental processes and lineage tracing. These tools enhance the ability to comprehend complex biological systems at single-cell resolution (Cao et al., 2019; Hao et al., 2023; Wolf et al., 2018).

2.2 Gene Regulatory Network Inference

Gene-regulatory network (GRN) modeling strategies show promise by reconstructing systematic transcription factor-gene associations from single-cell omics data. Nevertheless, existing methods for GRN analysis primarily concentrate on the static network structure, posing a challenge in understanding how a static GRN orchestrates cell identity during dynamic biological processes. There is a need for scalable and interpretable approaches to demonstrate the relationship between gene-regulatory mechanisms and single-cell phenotypes observed (Kamimoto et al., 2023).

To obtain information about the cell identity regulation, exclusive in silico frameworks are developed. These frameworks mainly aim to model the gene regulatory networks of transcription factors and their target genes regulated by them. These gene regulatory networks are thereafter transformed into a vector map depicting transitions in cell identity, allowing for an intuitive visualization of simulated changes in cell identity within a lower-dimensional space, such as SCENIC (Aibar et al., 2017), and GENIE3 (Huynh-Thu et al., 2010) and CellOracle (Kamimoto et al., 2023).

SCENIC is a motif-based gene regulatory network inference algorithm that can be run on different complex datasets, and it mainly focuses on in silico knock-out experiments on mouse datasets while utilizing GENIE3 in the background. GENIE3 is also a gene regulatory network inference workflow which focuses on gene expression data, but not specialized on single-cell RNA sequencing datasets.

Another tool to analyze gene regulatory networks based on trajectory inference concept is CellRank2 (Weiler et al., 2024). It is designed to infer trajectories and fate decisions from single-cell omics data robustly, modularly and scalably. CellRank2 integrates various data modalities such as pseudotime, developmental potential, RNA velocity, and experimental time points, to estimate transition probabilities between cells. Its inference results in identification of cell types and cell states such as initial, intermediate and terminal state on basis of cellular differentiation based on pseudotime and diffusion map calculations. Moreover, CellRank2 allows to visualize and cluster of gene expression trends among different cell subtypes and differentiation states, which provides deeper understanding of cellular behavior and lineage decision.

On the other hand, CellOracle can combine multimodal data, such as single-cell RNA sequencing and single-cell ATAC sequencing data, meaning that while constructing gene

regulatory networks, both gene expression data and chromatin accessibility data is considered at single-cell level. Initially, CellOracle builds a base gene regulatory network with directional edges of no weight between transcription factors and their target genes. It uses the genomic DNA sequence of regulatory regions and TF-binding motifs to identify potential regulatory interactions. This involves defining the promoter and enhancer genomic regions, which are inferred from ATAC-seq data. Standard preprocessing of scRNA-seq data is required before GRN construction and simulation. For GRN inference, CellOracle uses a machine learning approach for target gene expression prediction based on regulatory gene expression levels determined in the base GRN fine tuning step. CellOracle infers the edges between the genes quantitatively, as it fits these models to sample data of gene expression. GRN model built by CellOracle must fulfill two criteria for signal propagation: representing transcriptional connections as directed network edges and using linear regression for GRN edges. CellOracle uses genomic sequences and TF-binding motif information to determine the base GRN structure and directionality, eliminating the need to infer causality or directionality from gene expression data. This allows CellOracle to employ a linear machine-learning model. The model predicts the expression of a target gene based on the expression of regulatory candidate genes (Kamimoto et al., 2023).

Gene regulatory network inference methods vary based on the types of data for the cells or samples under study. Initially, transcriptomics data is preprocessed to create a gene expression matrix, giving insights about the level of gene transcript across cells or samples. A set of transcription factor – target gene information from a curated database is sourced to identify genes with regulatory functions. Models are then developed to predict gene expression based on TF transcript presence, resulting in TF-target gene associations. These interactions form a GRN (Pratapa et al., 2020) .

For chromatin accessibility data, preprocessing involves calling peaks to form an accessibility matrix of peaks, which has a binary structure. It involves information on cis-regulatory elements (CREs) in the cells or samples, whether they are open or closed. CREs and the corresponding genes are linked depending on their genomic proximity. TF-CRE binding prediction is conducted by using motif matcher algorithms and databases, which results in TF-CRN-gene triplet formation. (Wittkopp & Kalay, 2012). These triplets are then simplified to TF- target gene complexes and combined into a GRN. When samples are profiled with both transcriptomics and chromatin accessibility data, in other

words, multi-omics data, each data type is preprocessed, and unpaired data modalities are integrated if necessary. This allows methods to use both data types simultaneously to build TF-CRE-gene triplets, which eventually yields a GRN (Badia-i-Mompel et al., 2023).

2.3 Cell-Cell Interactions

From a holistic perspective, interactions between cells of different types, or identities, are facilitated by extracellular signals, where a ligand binds to its corresponding receptor. These ligand-receptor interactions between different cell types can be modeled by using databases possessing context and cell type specific ligand-receptor information. Grasping cellular behavior and responses to neighboring cells is to model the ligand-receptor interactions is fundamental. With advances in single-cell RNA sequencing technology, the capacity to perform the measurement of ligand and receptor expression across different cell types emerges systematically. This allows for the deciphering of intercellular crosstalk networks, contributing to an insight of tissue function in both homeostasis and disease-related alterations. The identification of ligand-receptor interactions from single cell RNA sequencing data entails annotating complex relationships from the literature and employing a statistical framework that brings this knowledge with scRNA-seq data together, picking relevant interactions, namely CellPhoneDB (Efremova et al., 2020). CellPhoneDB is a tool especially designed to analyze cell-cell interactions using single-cell transcriptomics data. It determines ligand-receptor interactions from scRNA-seq data by combining a comprehensive literature-based database of ligand-receptor relationships with a statistical approach that integrates this database and picks relevant interactions from the data. The CellPhoneDB database is a publicly accessible repository that contains curated information on receptors, ligands, and their interactions through literature mining and well-known databases UniProt (Bateman et al., 2023) and PDB (Berman, 2000). It includes detailed subunit structure information for both ligands and receptors, accurately representing heteromeric complexes. Since cell-cell communication often involves multi subunit protein complexes, which are not adequately represented by the binary interactions used in many other databases and studies, providing subunit information is essential. The database

incorporates new, manually reviewed interactions with established roles in cell-cell communication, along with existing datasets related to cellular communication.

Several methods have been developed for ligand-receptor pair inference using single-cell transcriptomics data such as CytoTalk (Hu et al., 2021) and CellChat (Jin et al., 2021). These methods essentially rely on identifying significantly expressed genes in two distinct cell types: one for receptors and the other for ligands, and vice versa. In addition, identification of ligand-receptor pairs respective to those distinct cell types to identify signaling pathways is also possible by using these methods. Modeling cell-cell interactions solely based on ligand-receptor interaction information might be limited. However, it becomes feasible to analyze downstream genes by integrating this information with intracellular gene regulatory networks.

On the other hand, NicheNet (Browaeys et al., 2020) is another method designed to model intercellular communication by binding ligands to target genes, revealing how cells influence each other through signaling molecules. This tool integrates multiple types of biological data, including receptor-ligand interactions, gene regulatory networks and signaling pathways, to create a comprehensive communication network. By analyzing gene expression data from interacting cell types, NicheNet predicts which ligands are active and how they affect target gene expression in receiving cells, aiding in identifying key signaling pathways and potential therapeutic targets. Main functionalities of NicheNet are assessing ligand effects on gene expression changes, prioritizing ligands based on their influence, and inferring active ligand-target links. It supports data from single-cell or bulk RNA sequencing and provides visualizations to illustrate the results. Additionally, NicheNet allows users to build and validate their own ligand-target models, offering flexibility for various research contexts.

Taking it to another level, the integration of transcriptomic data with a cell-cell interaction layer derived from scRNA-seq data is essential for a better understanding of tumor samples. Constructing a hypergraph to fulfill a holistic perspective in intracellular and intercellular modeling plays a crucial role in this process as it provides a better and deeper understanding of the cell-cell interactions which originates from their downstream signaling pathways, beginning from the interactions of transcription factors with corresponding genes.

2.4 Network Reconstruction Methods

Network reconstruction methods are techniques used to infer and identify important sub-networks or pathways within large biological networks. These methods are essential for understanding the complex interactions within biological systems, such as protein-protein interaction networks, GRNs and metabolic pathways (Guney & Oliva, 2012). To have a better understanding of the changes in different types of molecules across biological systems and their interconnections, biological datasets can be represented as a network featuring known or predicted interactions. There are state-of-the-art network reconstruction methods such as PathLinker (Gil et al., 2017), DOMINO (Levi et al., 2021), HotNet2 (Leiserson et al., 2015), Random Walk with Restart (RWR) (Köhler et al., 2008) and Prize-Collecting Steiner Forest (PCSF) (Tuncbag et al., 2013).

PathLinker is a network biology-based tool that reveals signaling pathways through protein-protein interaction networks. It utilizes the Steiner Forest algorithm to find the most relevant subnetwork that connects a set of source and target proteins, which identifies potential signaling pathways. In addition, it integrates various types of biological data and generates pathways that might reflect the underlying biological processes more precisely. It has a wide range of application in terms of signaling pathway identification in cancer and cellular response to external stimuli concepts (Gil et al., 2017).

DOMINO (Discovery of active Modules in Networks using Omics) is a framework that aims to determine disease-associated modules in the protein-protein interaction networks. DOMINO dynamically integrates different omic data types such as gene expression data and protein-protein interaction data. By utilizing this combination, it uncovers modules which are differentially regulated in healthy and disease conditions. The tool also aims to identify active network modules with high accuracy of GO term calls and to increase the specificity of GO term enrichments (Levi et al., 2021).

HotNet2 is an algorithm that aims to identify mutated subnetworks among pan-cancer networks. It reveals mutations that are mutually present in low frequency in the cancer signaling pathways to discover new therapeutic agents in many cancer types. The algorithm uses a heat diffusion approach, in which gene expression scores and the topology of the revealed subnetworks are vectorized in combination (Leiserson et al., 2015).

The Random Walk with Restart (RWR) algorithm is a network propagation method used to rank nodes in a network based on their relevance to a set of seed nodes. It simulates a random walker that starts from the seed nodes and iteratively moves to neighboring nodes with a certain probability or returns to the seed nodes with a restart probability. The random walker has a probability γ of returning to the seed node and a probability of 1- γ of moving to a neighboring node, where it continues to walk until the probability distribution over the nodes converges. RWR is effective in ranking candidate genes or proteins in disease-related studies such as cancer studies (Köhler et al., 2008).

Another method that is used for network reconstruction is pyPARAGON (PAgeRAnk-flux on Graphlet-guided network for multi-Omic data integratioN) (Arici & Tuncbag, 2023). It is based on inference of a context-specific subnetwork from an input reference network, using a seed node set based on multi-omic data, where significantly present modules in these subnetworks are determined. It operates in three main steps: constructing a Graphlet-guided network (GGN), propagating and scoring edges using the Personalized PageRank (PPR) algorithm and flux calculation, and reconstructing the subnetwork with highly scored edges. By identifying significantly abundant graphlets, pyPARAGON reduces network complexity and focuses on relevant interactions. Additionally, it performs community analysis by using Louvain community detection method to divide the network into functional modules. This allows pyPARAGON to do biological interpretations like patient stratification and survival analysis (Arici & Tuncbag, 2023).

The Prize-Collecting Steiner Forest (PCSF) algorithm is a network optimization method used to identify sub-networks that maximize the total prize of included nodes while minimizing the total cost of included edges. It is particularly useful for integrating different types of biological data and reconstructing condition-specific networks. In PCSF algorithm, each edge in the network has a cost which represents the confidence of the interaction. And each terminal node in the network has a prize based on its importance. Terminal nodes are the ends of the subnetworks to be constructed. These subnetworks are the networks that collect the highest possible total prize and the lowest possible total cost. Prize-Collecting Steiner Forest Algorithm can be used in biological context as in this thesis study. For a specific disease or condition, pathways as subnetworks can be reconstructed by integrating multi-omics data to identify important regulatory networks.

Moreover, it can be an important part of potential biomarker and therapeutic target discovery (Tuncbag et al., 2013, 2016).

The advances in single-cell omics technologies have dramatically increased our understanding of gene expression and cellular heterogeneity, allowing for a more precise characterization of individual cell identities and their regulatory mechanisms. The integration of scRNA-seq with pooled patient samples and other omics data has enabled more accurate discrimination of cell types and the reconstruction of intricate differentiation pathways, despite the increased computational complexity. The development of gene regulatory network inference methods, such as SCENIC, GENIE3, and CellOracle, has provided powerful tools for modeling transcription factor-gene associations and understanding dynamic biological processes. Furthermore, the exploration of cell-cell interactions through ligand-receptor pair identification and the incorporation of transcriptomic data has deepened our insights into cellular communication, particularly on the tumor microenvironment. Network reconstruction methods like RWR and PCSF have proven essential for deciphering complex biological networks, which supports the identification of key regulatory pathways, therapeutic agents and potential biomarkers. Together, these methodologies represent a comprehensive framework for advancing our knowledge of cellular networks and their implications for diseases like cancer.

Chapter 3:
MATERIALS AND METHOD

3.1 Summary of the method

This thesis study mainly aimed to construct a framework involving processing and network representation of single-cell multi-omic data. We used single-cell RNA seq data from 44 patients with lung adenocarcinoma (LUAD), corresponding to 208,506 single-cells published by Kim et al., Nat Comm. (2020). We first tissue-specifically downsampled and clustered the cells, resulting in 46 clusters and then constructed a gene regulatory network for each cell cluster. In addition, we found significant receptor-ligand pairs that are potential initial nodes for upstream signaling network reconstruction and form cell-cell interaction networks. For each cluster, we integrated regulatory network components with R-L pairs and applied Omics Integrator tool for intracellular network modeling. We elaborated on the cancer cell clusters and their connections to other cell subtypes, the differentially expressed genes among healthy and cancer cells in primary tumor and metastases.

Table 3.1. Composition of lung adenocarcinoma single-cell transcriptomic data by Kim et al. tLung represents the cells from primary tumor site. mBrain represents the brain metastases.

Total number of single cells	208,506
Total number of tLung single cells	45,149
Total number of mBrain single cells	29,060
Total number of main cell types	8
Total number of cell subtypes	46

3.2 Softwares, tools and databases for sc-RNA seq data analysis and integration

3.2.1 SCANPY

SCANPY (Wolf et al., 2018) is a widely utilized library designed to analyze the single-cell RNA sequencing data. It offers a comprehensive tool for preprocessing, visualization, and analysis, including dimensionality reduction, clustering, and trajectory inference. SCANPY is built on top of the AnnData data structure and integrates it with various other single-cell analysis tools. It is efficient in handling large datasets and is flexible and modular, which makes it suitable to study high-throughput scRNA-seq data. In this thesis study, we used SCANPY for preprocessing, dimensionality reduction, clustering and visualization of scRNA-seq data of our interest.

3.2.2 AnnData

AnnData (Virshup et al., 2021) is a data structure designed for efficient storage and manipulation of annotated data matrices, particularly in the context of single-cell genomics. It provides a convenient way to store data along with annotations for observations (cells) and variables (genes), as well as unstructured metadata. AnnData is compatible to SCANPY and other single-cell analysis workflows, facilitating the integration and analysis of complex datasets. It supports sparse and dense matrix formats, making it adaptable to various data sizes and types. This study includes loading and analysis of scRNA-seq data as an AnnData object.

3.2.3 3.2.3 CellOracle

CellOracle (Kamimoto et al., 2023) is a computational tool for constructing and analyzing gene regulatory networks from single-cell RNA-seq data. By inferring gene regulatory interactions and simulating gene expression changes, CellOracle elaborates on the regulatory mechanisms underlying cellular processes and transitions. It combines single-cell gene expression and chromatin accessibility data with prior information from databases and literature to create more accurate and biologically relevant network models, assisting in the exploration of cell fate decisions and gene function.

CellOracle builds cell-subtype-specific gene regulatory networks from scRNA-seq data with respect to the Equation 3.1:

$$x_j = \sum_{i=0}^{n} b_{i,j} x_i + c_j$$

$$(Equation\ 3.1)$$

where x_j is the expression of the target gene, x_i is the expression value of the candidate regulatory gene that regulates gene x_j, b_{ij} is the coefficient of the linear model (with b_{ij} = 0 if i = j), and c is the intercept for this model. After calculating every edge, this process yields a gene regulatory network consisting of TF-gene interactions as a list (Kamimoto et al., 2023).

We utilized CellOracle for gene regulatory network inference for every cell subtype in our scRNA-seq data of interest, where we used the list of transcription factors for further steps.

3.2.4 TRRUST v2

TRRUST v2 (Transcriptional Regulatory Relationships Unraveled by Sentence-based Text mining, version 2) (Han et al., 2018) is a comprehensive database of human and mouse transcriptional regulatory networks. It possesses information through literature-mining on transcription factor - target gene interactions. TRRUST v2 provides insights into the direct regulatory relationships and functional roles of TFs in various biological processes and diseases. In this thesis, we used TRRUST v2 database to involve general transcription factor-gene information in CellOracle gene regulatory network inference part.

3.2.5 CellPhoneDB

CellPhoneDB (Efremova et al., 2020) is an online database and computational tool used to infer cell-cell communication networks from single-cell transcriptomics data. It contains curated information on ligand-receptor interactions, enabling the identification of potential signaling pathways between different cell types. By analyzing the expression levels of ligands and receptors, CellPhoneDB reveals the complex intercellular communication that underlies tissue function and disease progression. We used CellPhoneDB in our study to receive receptor-ligand interaction information for every single cell subtype in our scRNA-seq data.

3.2.6 OmicsIntegrator

OmicsIntegrator (Tuncbag et al., 2016) is a software tool designed to integrate diverse types of omics data such as genomics, transcriptomics, proteomics with biological networks to identify key regulators and pathways involved in specific biological processes or diseases. By combining experimental data with known biological interactions, OmicsIntegrator reconstructs signaling pathways and suggests on the molecular mechanisms that drive observed phenotypes and prioritize targets for further investigation. OmicsIntegrator was one of the most important tools of our framework, where we reconstructed cluster-specific intracellular network models and connected those networks to each other through receptor-ligand interaction information we provide.

3.2.7 HIPPIE

HIPPIE (Human Integrated Protein-Protein Interaction rEference) (Alanis-Lobato et al., 2017) is a comprehensive database of human protein-protein interactions (PPIs). It integrates data from multiple sources and provides confidence scores for each interaction based on the supporting evidence. HIPPIE is a valuable resource for studying the complex protein interaction networks that govern cellular processes, offering insights into the molecular basis of health and disease. HIPPIE was used in our study to provide a reference interactome with confidence scores for each interaction to initiate OmicsIntegrator.

3.2.8 NetworkX

NetworkX (Hagberg et al., 2008) is a Python library designed for the creation, manipulation, and study of complex networks and graphs. It provides a wide range of tools to work with both simple and complex graph structures, including directed, undirected, and multigraphs. NetworkX supports numerous algorithms for network analysis, such as shortest path, clustering, and centrality measures, suitable for biological network analysis. The library integrates well with other scientific Python libraries, facilitating efficient data manipulation and visualization. In this thesis, NetworkX is primarily used for network visualization and manipulation of obtained graphs.

3.2.9 WebGestalt

WebGestalt (WEB-based GEne SeT AnaLysis Toolkit) (Liao et al., 2019) is a widely used bioinformatics tool for functional enrichment analysis. It is designed to interpret large-scale -omics data by translating gene lists into meaningful biological insights. WebGestalt supports multiple methods for enrichment analysis, including Over-Representation Analysis (ORA), Gene Set Enrichment Analysis (GSEA), and Network Topology-based Analysis (NTA). WebGestalt enables programmatic access via an API (Application Programming Interface). Recent updates have included support for phosphoproteomics data, improved visualizations, and methods to reduce redundancy in enriched gene sets, making it a comprehensive and flexible tool for gene set enrichment analysis.

3.2.10 Cytoscape

Cytoscape is a software that is primarily used for visualizing complex networks (Smoot et al., 2011). It supports various biological domains, providing tools to map, analyze, and visualize networks representing molecular pathways, gene regulation, and protein-protein interactions. It extends through plugins, which allow users to add new functionalities and customize the platform according to their specific needs, which enable a wide range of analyses, from network inference to dynamic simulations and integration with large-scale omics data. Cytoscape is often used in combination with other bioinformatics tools to have a better understanding of complex biological systems. In this thesis, we used Cytoscape to visualize our resulting networks.

3.3 scRNA-seq data preprocessing

The single cell RNA sequencing data used to construct a multi-scale network-based representation of tumors was retrieved from Kim et al. (2020). The whole data consists of pooled samples from 44 patients with various progressions of cancer and includes cells from normal tissues, as well. We performed scRNA-seq preprocessing part using scanpy v1.9.3 and anndata 0.9.2. Specifically, single cells that were profiled were from cancer tissue-derived whole cells from primary sites, pleural fluids, lymph nodes, brain metastases, tumor and normal tissues from lungs and lymph nodes. With a total number

of 208,506 cells, the overall network reconstruction pipeline was initiated. The dataset was loaded as an Anndata object. Then, the cell annotation data, which consists of the index of every single cell, the name of the barcode, the sample number, origin of sample, cell type and cell subtype, was loaded. At first, the whole dataset was filtered to the samples of the patients from the primary lung adenocarcinoma tumor, namely tLung. Then, cells without a cell subtype were assigned their cell type as the cell subtype to avoid errors. Thereafter, highly expressed genes throughout the dataset were determined, which was followed by filtering cells that have less than 500 genes expressed and filtering genes that are detected in less than 10 cells.

After tagging the group of mitochondrial genes as mt, those genes were also removed from the dataset, which is a step for noise reduction. Also, the cells having mitochondrial genes with higher than 5% were removed. Then, the whole gene expression matrix was normalized with the parameter of target sum equal to 1e4 and log normalization was performed, since the dataset was provided accordingly for log2TPM normalization, by using scanpy.pp.normalize_total() and scanpy.pp.log1p() functions, respectively. The principal component analysis (PCA) was performed by automatically computed number of principle components, which is equal to 50, with scanpy.tl.pca() function. Relying on this dimensionality reduction method, scanpy.pp.neighbors() function determined the neighbors of each single cell as well as distances for each pair of neighbors and connectivities, in other words weighted adjacency matrix. After this step, Louvain clustering was applied to the data to construct the tSNE visualization with a resolution value of 2.3, yielding 46 clusters in total. By ranking the genes for each cluster, the highest expressed genes which correspond to the marker genes for those clusters were extracted, so that each cluster was assigned to a cell type, based on the marker gene dataset. The assignment of the cell types was validated according to the previously loaded cell annotation file provided by Kim et al. In addition, ranking plot for each cell subtype against the rest of the cell subtypes was constructed by applying t-test, Wilcoxon test and logistic regression. And this gene ranking process yielded a dataframe for each cluster, where genes and their log-fold change values are present, for determining their upregulation or downregulation information.

3.4 Construction of Gene Regulatory Networks

In this step of the framework gene regulatory networks, which consist of transcription factors and genes that are regulated by them, were reconstructed by CellOracle tool. To initiate CellOracle, creating a diffusion map and diffusion pseudotime (DPT) metadata were required by setting a root cell from the diffusion map of the single-cell Anndata object. This step identifying a root cell in trajectory inference or further analyses where a starting point is required. In order to select the same root cell every time, a seed was used.

Firstly, a pre-existing base gene regulatory network data set, automatically fetched by CellOracle v0.18.0s from GimmeMotifs v5—an analysis framework for transcription factor motif analysis that includes chromatin accessibility information (Bruse & Heeringen, 2018) — was integrated by using the function celloracle.data.load_human_promoter_base_GRN(). Then, a dictionary that contains transcription factor-gene interactions was fetched and adapted to the workflow from TRRUST v2 human transcription factor-gene database. Thereafter, by using CellOracle, principal component analysis was performed for network inference this time, which was followed by k-nearest neighbors (kNN) imputation method for further reconstruction with oracle.perform_PCA() and oracle.knn_imputation() functions. Then the gene regulatory network inference was performed for each cell subtype through oracle.get_links() function, where uncertain network edges based on the p-value were removed. If the p-value was greater than 0.05, the edge was removed. After this step, network scores such as degree centrality, betweenness centrality and eigenvector centrality were calculated for each transcription factor present in the cell subtypes. At the end of this step, gene regulatory networks were obtained for each cluster where the columns consisted of transcription factors and genes list. The transcription factors list for every cell subtype cluster was collected for further analysis.

3.5 Receptor-ligand network construction

Receptors and ligands that are significantly expressed in the selected clusters were determined in this step. To achieve this, receptor-ligand interaction data was retrieved from CellPhoneDB and adapted for the pipeline. CellPhoneDB consists of three main dataframes, that are protein input, gene input and interaction input. These three main dataframes of CellPhoneDB were merged by using inner join operation. Then the merged

dataframe was filtered to contain only the receptor-ligand interactions that are related to the cell membrane surface itself. In that manner, either transmembrane, peripheral or secreted proteins were taken, and others were left out. Then, for specific for every cell subtype, ligand and receptor sets were filtered and saved as separate files. This is done by matching the gene expression information coming from the scRNA-seq data with CellPhoneDB dataframe. Thereafter, a dataframe including every receptor-ligand interaction for every combination of cell subtypes was created for further steps.

3.6 Multi-layer network reconstruction

After collecting the transcription factors for each cluster in Section 3.3 and the ligand-receptor information in Section 3.4, they were merged for creating terminal files for each cell subtype. Terminal files are required for Omics Integrator to initiate, as they represent the start nodes and end nodes of the networks to be constructed. A representative for this concept can be viewed in Figure 3.1.

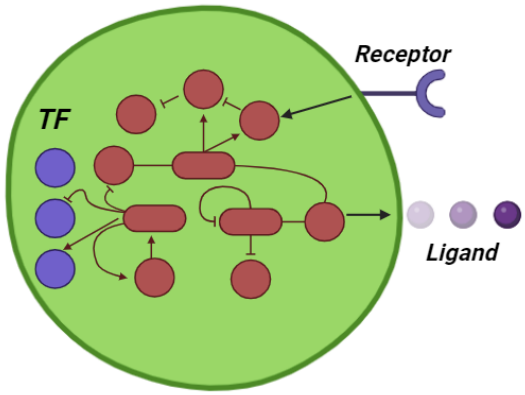


Figure 3.1. A representative of terminals in a cell-type specific network. Purple nodes, that are at both ends of the network, represent the terminals, namely transcription factors (TF), receptors and ligands. Red nodes represent the intermediate proteins that are present in the reference interactome and found in the solution of the Prize-Collecting Steiner Forest algorithm embedded in Omics Integrator.

Omics Integrator tool was used for 'merging' those counterparts of the whole network reconstruction model. As the reference interactome for human, HIPPIE was utilized, in order to assign costs for every protein that is present in this reference interactome. For every single interaction, a cost was assigned with the Equation 3.2.

$$c(e) = 1.5 - p(e)$$

$$(Equation\ 3.2)$$

In Equation 3.1, the c(e) represents the cost of any edge and p(e) represents the confidence value of each edge in the interactome. After obtaining the cost for every interaction for every cell subtype/cluster, terminal sets for all these cell subtypes were determined, and they were introduced as the dummy node set. To include all terminal nodes in the final solution of PCSF for every cell subtype/cluster, especially receptors and ligands, uniform prizes were assigned to the terminal nodes. Then the terminal node file was filtered to exclude the transcription factors that target less than 25 genes, to improve visual clarity of the graphs.

The aim of the PCSF algorithm that is embedded in Omics Integrator in this part is to find a forest which minimizes the objective function in Equation 3.3.

$$f'(F) = \sum_{v \notin Vf} p'(v) + \sum_{e \in E_F} c(e) + \omega * \kappa$$

(Equation 3)

In Equation 3.3, an artificial node, referred to as a 'dummy' node (v_0), is added to the original network and linked to a subset of nodes (N). These nodes in N are chosen from all nodes in the interactome (V) and are given a uniform edge cost (ω) to the dummy node. The forest constraint (F) ensures that F forms a tree, which is a connected graph with no loops, with v_0 serving as the root. The goal is to solve an optimization problem to identify a tree subnetwork. As the optimization problem is solved, the root node (v_0) and its associated edges are eliminated, resulting in a forest network comprising one or more subtrees. These subtrees represent parallel biological pathways, as the solution obtained depends on the specific values of β, ω, and μ (Tuncbag et al., 2016). The reconstruction part was initiated with the parameter search for the Prize Collecting Steiner Forest algorithm that is embedded in the tool. ω, β and μ are the required parameters for the algorithm to be solved. In order to determine the most optimal parameters, a membership dataframe for each cell subtype/cluster was constructed, where the number of total nodes in the output graphs, number of the covering nodes and the number of components of the graphs. Covering nodes indicate the number of terminal nodes that are present in the output graph. The most optimized scenario for those three parameters would be maximizing the number of the covering nodes while minimizing the total nodes. In addition, we desired the number of components to be as minimum as possible, less than

3, in this context. Every cluster/cell subtype yielded an augmented forest posterior to the Omics Integrator initiation. Graphs for each cell subtype/cluster were collected and saved as GraphML (Brandes et al., 2013) files for further steps.

3.7 Multi-cell-subtype graph construction

Based on the dataframe that is obtained in the end of Section 3.4, edges representing the receptor-ligand information were added after loading the GraphML files of every cell subtype/cluster. Then, the corresponding receptors and ligands of corresponding cell subtypes were trimmed to create edges between those graphs by a manually created function, namely add_edges_manually. This function collects the information of presence of receptors and ligands in two distinct cell subtypes, and if there is an interaction between them, adds and edge with a cost of 0.1 and includes it within the PCSF solution, so that these manually added receptor-ligand information can be visualized. Applying this function for every combination of cell subtypes present in our dataset yielded a hypergraph with n number of nodes where every node is also a graph.

3.8 Condition-specific Differentially Expressed Gene Analysis

In this part, we identified differentially expressed genes among different conditions. First, we filtered the whole dataset only to the samples taken from the primary lung tumor (tLung), then applied the same preprocessing procedure in Part 3.2. After cell type annotation, we filtered the cells again only to healthy Epithelial cells and cancer cells with three different transcriptional states (tS1, tS2 and tS3) and applied t-test by using scanpy.tl.rank_genes_groups() function to yield a dataframe with genes, their log2 fold changes and adjusted p-values. We only included those with an adjusted p-value smaller than 0.05 and an absolute log2 fold change value greater than 1. We gave the four filtered gene lists to WebGestalt Over-Representation Analysis (ORA) by setting the parameter of False Discovery Rate (FDR) to 0.05 and leaving the other parameters as default. We selected “pathway” and “KEGG” as the functional database. We applied the same procedure to obtain the differentially expressed genes in metastatic cancer cells in brain against primary tumor cancer cells and healthy epithelial cells and obtained biological processes that are significantly present in those cell subtypes.

Chapter 4:
RESULTS

4.1 Overview of the framework

This thesis study mainly aimed to integrate scRNAseq data with known molecular interactions and release a new framework to represent a given dataset from a context as a multi-scale network. It is not limited to only transcriptomic data but also highly adaptable to any single-cell multi-omic data. To demonstrate it, we conducted our pipeline on a scRNA-seq data consisting of pooled samples of 44 patients from lung adenocarcinoma patients by Kim et al. (2020). We used this data to perform tissue-specific downsampling only to primary tumor sample single-cells where we obtained 46 Louvain clusters of distinct cell subtypes, consisting of healthy and cancer cells. Then, by utilizing CellOracle, we constructed gene regulatory networks consisting of transcription factors and their target genes for each cell cluster, where we used TRRUST v2 as our reference transcriptional regulatory network. Thereafter, we collected the transcription factors for each cell type, determined the differentially expressed transcription factors, and identified them as potential initial nodes for our intracellular network models. In addition, we determined cell subtype-specific receptors and ligands through CellPhoneDB and added these proteins as potential terminal nodes for our intercellular network reconstruction, too. For each cluster, we combined the transcription factors with receptors and ligands and generated the terminal files with uniform prizes. Following, we applied Omics Integrator tool for network modeling, where we used HIPPIE as our reference protein-protein interactome and obtained intracellular networks. Additionally, we connected different cell subtypes such as healthy epithelial cells and cancer cells by using the receptor-ligand interaction information we obtained via CellPhoneDB. We also performed differential gene expression analysis between healthy cells and cancer cells in the primary tumor samples, and between primary tumor cells and brain metastases. We also performed over-representation analysis through WebGestalt to determine significantly over-represented or under-represented biological processes in cancer cells in comparison to healthy cells. We conducted another additional over-representation

analysis between healthy epithelial cells and cancer cells in the primary tumor samples, where we included the intermediate proteins determined by Omics Integrator to reveal possible contributions of these proteins to the abnormalities in cancer cells. In summary, we elaborated on the cancer cell clusters and their connections to other cell subtypes, the differentially expressed genes and transcriptional programs among healthy and cancer cells in primary tumor samples and metastases.

Overall workflow of the thesis consists of six main steps, the scheme of which is indicated in Figure 4.1 below:

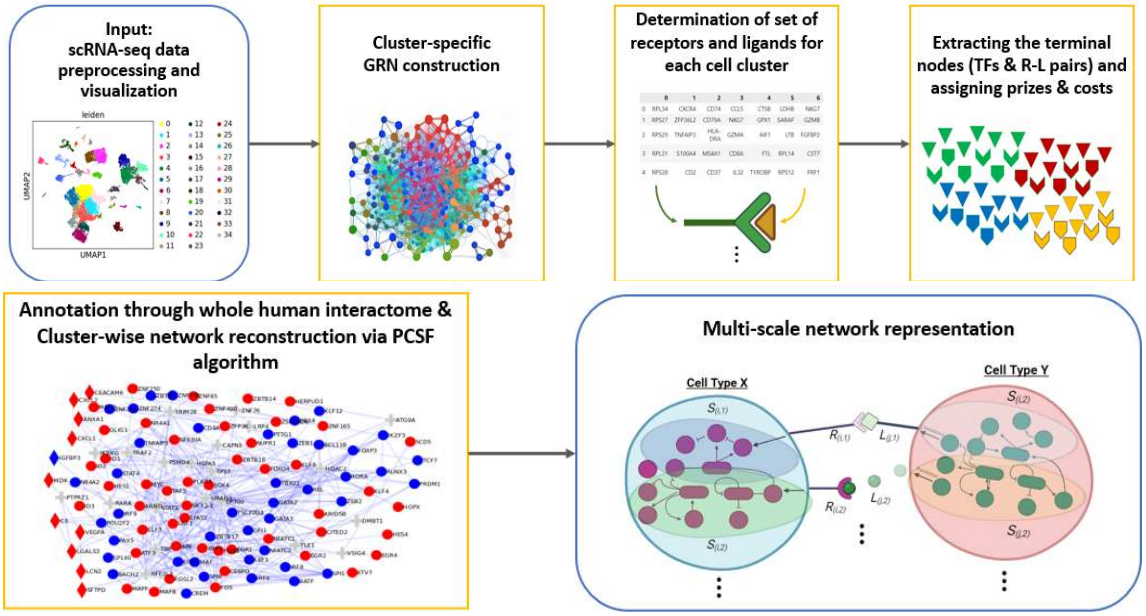


Figure 4.1. Overall framework of multi-scale network representation of tumors. Input begins with the introduction of the single-cell RNA sequencing data of lung adenocarcinoma primary tumor samples by Kim et al. (2020) as an AnnData (Virshup et al., 2021) object to SCANPY (Wolf et al., 2018), its preprocessing and tSNE visualization. It is followed by the construction of cluster specific gene regulatory networks via CellOracle (Kamimoto et al., 2023). Set of ligands and receptors with respect to identified cell clusters were fetched from CellPhoneDB (Efremova et al., 2020). Then prizes for the terminals and the cost for the edges throughout the reconstructed network were given to the Prize Collecting Steiner Forest algorithm (Tuncbag et al., 2013), where cluster-specific networks and their receptor-ligand interactions were included via OmicsIntegrator (Tuncbag et al., 2016) in the last two steps.

4.2 Single-cell RNA sequencing data preprocessing and visualization

We first preprocessed the single-cell RNA sequencing data of metastatic lung adenocarcinoma primary tumor samples with the following main steps; prefiltering based on gene and cell counts, data normalization, cell type annotation. At the beginning, we had 208,506 cells with 29,634 gene counts. Then we filtered out the cells that have expressed less than 500 cells and the genes that have been expressed in less than 10 cells. In addition, we removed the cells that have expressed the mitochondrial genes more than 5% for sake of noise reduction. We also removed the cells according to a gene expression threshold of 0.0125. After the filtering process, we obtained 36,082 cells with 3802 gene counts from the primary tumor samples, consisting of 2,709 cancer cells and 554 healthy epithelial cells. In addition, there were 17,965 cells with 3366 gene counts from the brain metastases, 6,694 of which are metastatic cells and 25 are healthy epithelial cells. Based on the filtering we performed, we obtained two tSNE graphs, the first of which shows the 8 main cell types that are present in the primary tumor samples, where the second tSNE graph yielded 46 clusters, in other words, cell subtypes, in the same part of the data. Both plots are shown in Figure 4.2 and 4.3, respectively.

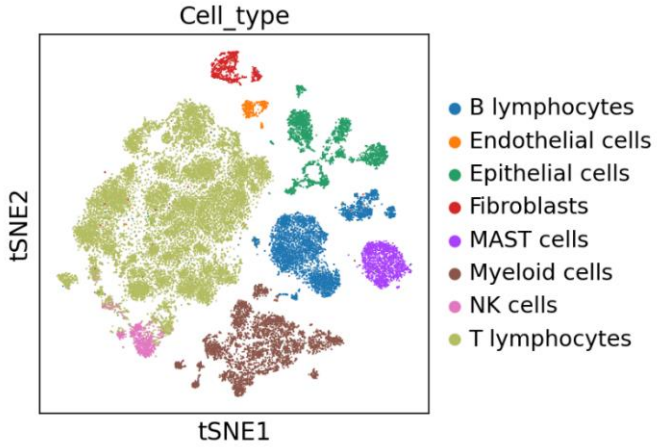


Figure 4.2. tSNE plot of the main cell types in the dataset. Each cluster represents the main cell types in the dataset, which consists of the analysis from the samples taken from the lung adenocarcinoma tumor area. Clusters were determined via scanpy, using Louvain clustering method after downsampling, principal component analysis, normalization and cell type annotation steps.

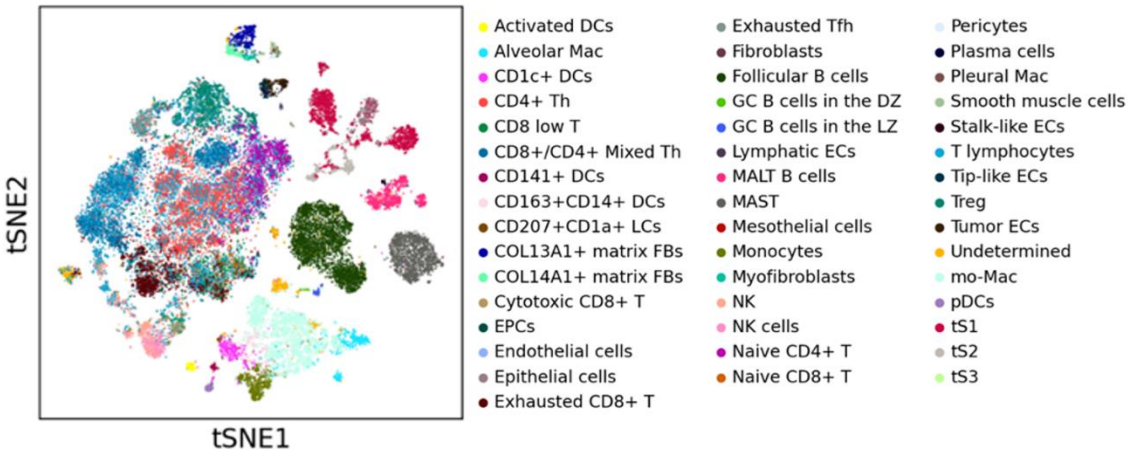


Figure 4.3. tSNE plot of the cell subtypes in the dataset. Each cluster represents the cell subtypes in the dataset, which consists of the analysis from the samples taken from the lung adenocarcinoma primary tumor. Clusters were determined via scanpy, using Louvain clustering method after downsampling, principal component analysis, normalization and cell subtype annotation steps. tS1, tS2 and tS3 are the clusters of cancer cells with three different transcriptional states, which are determined by Kim et al. by differential expression analysis.

As seen in Figure 4.2 and 4.3, there were 46 cell subtypes in total, consisting of 36,082 single-cells, consisting of immune cell groups, fibroblasts, endothelial cells and epithelial cells. We identified the main cell types in Figure 4.2 and Figure 4.3 through the marker gene identification by gene ranking with t-test and the annotation file provided by Kim et al.

Following, we determined the clusters that might be closely associated with the cancer cells, such as fibroblasts with their subtypes, endothelial cells with their subtypes and epithelial cells and their subtypes. We built the gene regulatory networks in the corresponding cell subtypes. Then, we performed differential expression analysis by using t-test to determine the differentially expressed transcription factors among those from CellOracle, from which we extracted the transcription factor lists for each cell subtype as the initial nodes for our reconstructed networks. A gene regulatory network that is constructed by CellOracle for one of our cell types is filtered to the differentially expressed transcription factors. A representation of this gene regulatory network of the top three differentially expressed transcription factors in this cell type is indicated in Figure 4.4.

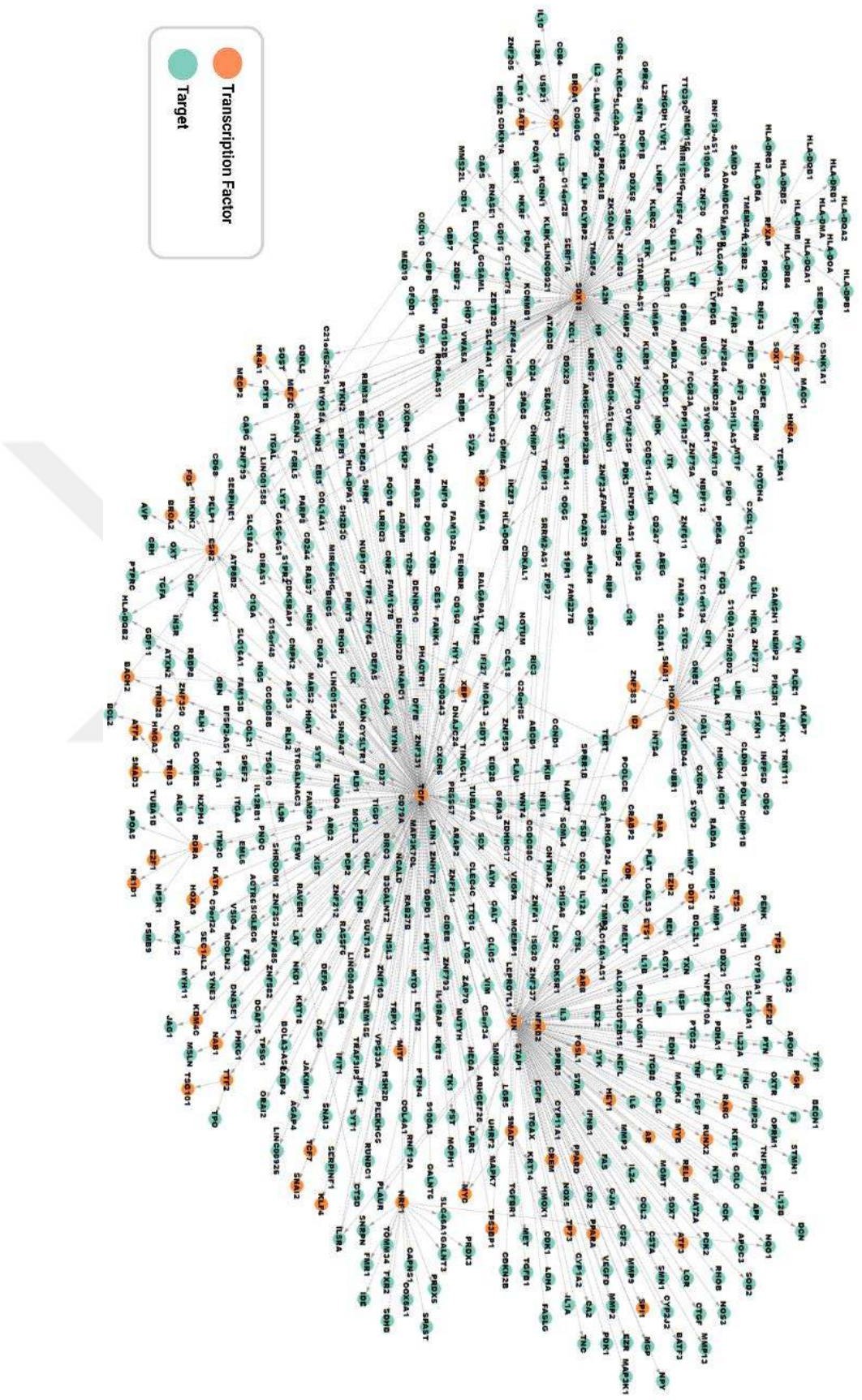

Transcription Factor
Target

Figure 4.4 Network modeling of top 3 highly expressed transcription factors. These are TCF4 (Transcription factor 4), SOX18 (SRY-box transcription factor 18) and HOXA10 (Homeobox A10). and their targets in transcriptional state 1 (tS1) in primary tumor samples. We also indicated the targets of the first neighbors of these transcription factors, which are also transcription factors. There are 71 transcription factors in total. Orange nodes indicate the transcription factors, turquoise nodes indicate their targets. There are 710 nodes and 771 edges in this network.

Moreover, we performed an over-representation analysis in WebGestalt with TCF4 and its targets in the tS1 gene regulatory network, which yielded Figure 4.5.

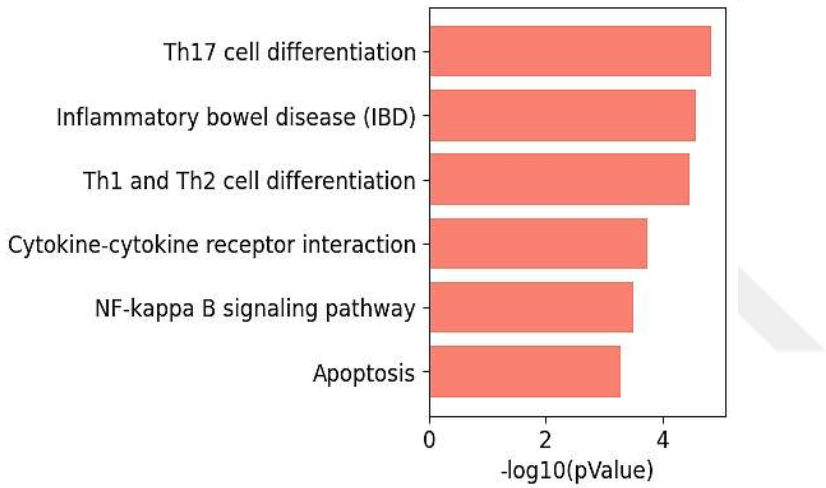


Figure 4.5. Over-representation analysis of the targets of TCF4 transcription factor in transcriptional state 1 (tS1) in primary tumor samples. The bar chart indicates the –log10(pValue) of the biological processes that are significantly over-represented in tS1. The results are from Webgestalt Over-representation Analysis with a false discovery rate (FDR) equal smaller than 0.05.

TCF4 and its targets have been over-represented in cytokine-cytokine receptor interaction process in tS1. In lung cancer types, it is reported that cytokine proteins are active to trigger the immune system and are involved in pro-tumorigenic processes (Abolfathi et al., 2021). In addition, TCF4 expression is reported to inhibit apoptosis, as we found TCF4 and its targets to be over-represented in apoptotic processes in some cancer types (Xie et al., 2012).

Then, we enriched our terminal node list by adding the significantly expressed ligands and receptors. The properties of the terminal node sets as the input of Omics Integrator, and the properties of the resulting networks are indicated in Table 4.1.

Table 4.1. Properties of intracellular networks of cell types that might be associated with cancer cells. # represents the number. The number of covered nodes indicate the number of nodes which are present in the terminal set and the resulting network. Number of total nodes indicates the number of nodes in the resulting network including the intermediate proteins determined by Omics Integrator.

Cell Types	# of TFs	# of Ligands	# of Receptors	# of Terminals	# of Covered Nodes	# of Total Nodes	Coverage (%)	# of Edges in Solution
Endothelial cells	85	25	15	125	123	179	98.4	178
Epithelial cells	13	22	10	45	45	65	100.00	64
Fibroblasts	61	14	10	85	83	120	97.65	119
Mesothelial cells	139	15	12	166	155	222	93.37	221
Tumor ECs	43	22	25	91	86	119	94.51	118
tS1	24	14	14	52	52	75	100.00	74
tS2	19	21	12	52	51	65	98.08	64
tS3	84	8	4	96	95	141	98.96	140

In Table 4.1, there are 8 cell types that we determined to construct the intracellular networks of. All of them resulted in a forest with a node number of n and an edge number of (n-1), where the node coverage of our networks is not less than 93.37%.

In addition, we constructed the heatmaps of the ligands and the receptors that are present in these eight cell types, and we showed the interactions between the ligands and receptors with the heatmaps combined in Figure 4.6.

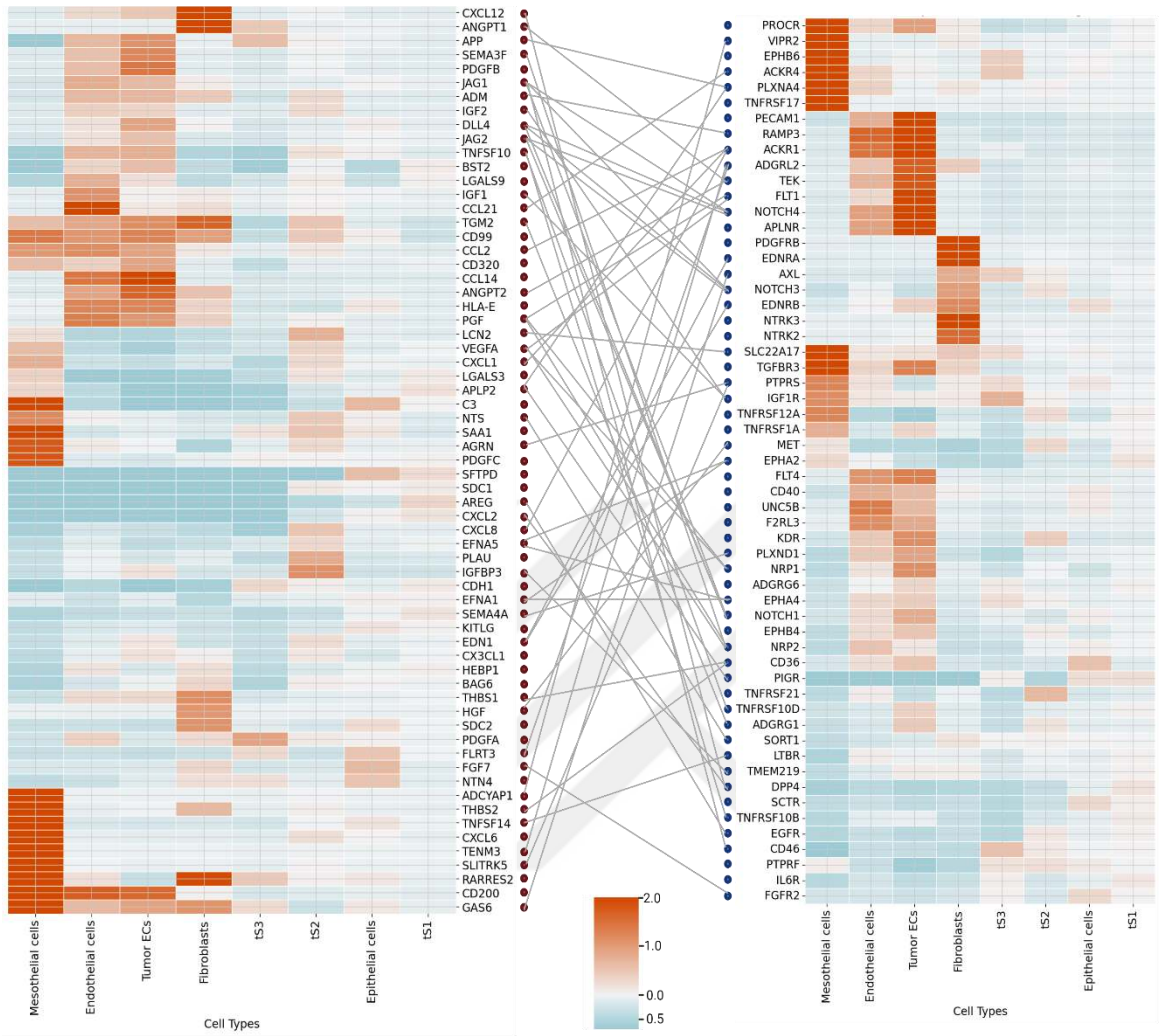


Figure 4.6. Heatmap of the expression levels of ligands and receptors in eight cell types. The cell types are the ones that are relevant to the cancer cell association. We took the union of the ligands and th receptors that are differentially expressed in these cell types. We connected the receptor-ligand pairs by using the receptor-ligand complex information in CellPhoneDB. The expression values are from scRNA-seq data by Kim et al., log2-transformed and normalized.

According to Figure 4.6, we determined that the expression level of IGFBP3 (Insulin like growth factor binding protein 3) in tS2 cancer cells is higher than the healthy epithelial cells, fibroblasts, endothelial cells and mesothelial cells. This aberrant expression of IGFBP3 is stated to trigger lung cancer cells through activation of IGF1 (Insulin like growth factor 1) signaling activity in lung carcinogenesis (Y. A. Wang et al.,

2017). In addition we revealed that in tS2, the expression of PLAU (Plasminogen activator, urokinase) is relatively higher than the healthy epithelial cells and other cell types, as well. It has been shown that overexpression of PLAU is a driving factor for cell growth in non-small cell lung cancer types (Zheng et al., 2024). However, in our receptor set in Figure 4.6, there were no receptors that PLAU targets. Following, we determined that LCN2 (Lipocalin 2) expression is relatively higher in comparison to other healthy cell types. Indeed, LCN2 is involved in differentiation of epithelial cells, and plays a role in various types of cancers such as breast cancer and might be a therapeutic target (Bao et al., 2024). We may link LCN2 to lung adenocarcinoma, as well. Additionally, ANGPT1 (Angiopoietin 1) is higher expressed in tS3 and fibroblasts. ANGPT1 is a protein that acts as a tumor suppressor, and it has been revealed that higher expression of ANGPT1 can increase the survival rate of patients suffering from lung cancer (Yao et al., 2020). On the other side of Figure 4.6, we examined the receptors present across eight cell types. There is a relatively higher expression of TNFRSF21 in tS2 than the other healthy cell types. TNFRSF21 is reported to be highly expressed in lung adenocarcinoma, where it promotes tumor invasion and abnormal cell growth, which might make TNFRSF21 a potential target for lung cancer therapy (Zhou et al., 2021). Furthermore, we found that EDNRB is relatively less expressed in tS1, tS2 and tS3 than healthy epithelial cells. Indeed, EDNRB expression in low levels in lung adenocarcinoma patients occurs, as it decreases the survival rate, and might be able to inhibit the migration and proliferation of lung adenocarcinoma cells (Wei et al., 2020). Another transcription factor that contrasts in cancer cells is CD36 (CD36 molecule). It has a lower expression in tS1, tS2 and tS3, as its expression is relatively higher in healthy epithelial cells. It possesses a dual role in cancer, as it may act as a tumor suppressor or an oncogene (Jiang et al., 2024). In our case, CD36 shows lower expression values in cancer cells relative to the healthy cells, which might indicate the tumor suppressor role of CD36. Its lower expression might have driven the cells to become cancer cells.

After extracting the terminal nodes that include the transcription factors obtained via CellOracle, receptors and ligands obtained via CellPhoneDB in Sections 3.3 and 3.4, we reconstructed our multi-cell-subtype graphs via Omics Integrator. One of the graphs that represents the combination of intracellular networks of two cell types and their intercellular networks, as the solution to the PCSF problem found by Omics Integrator, is demonstrated in Figure 4.7.

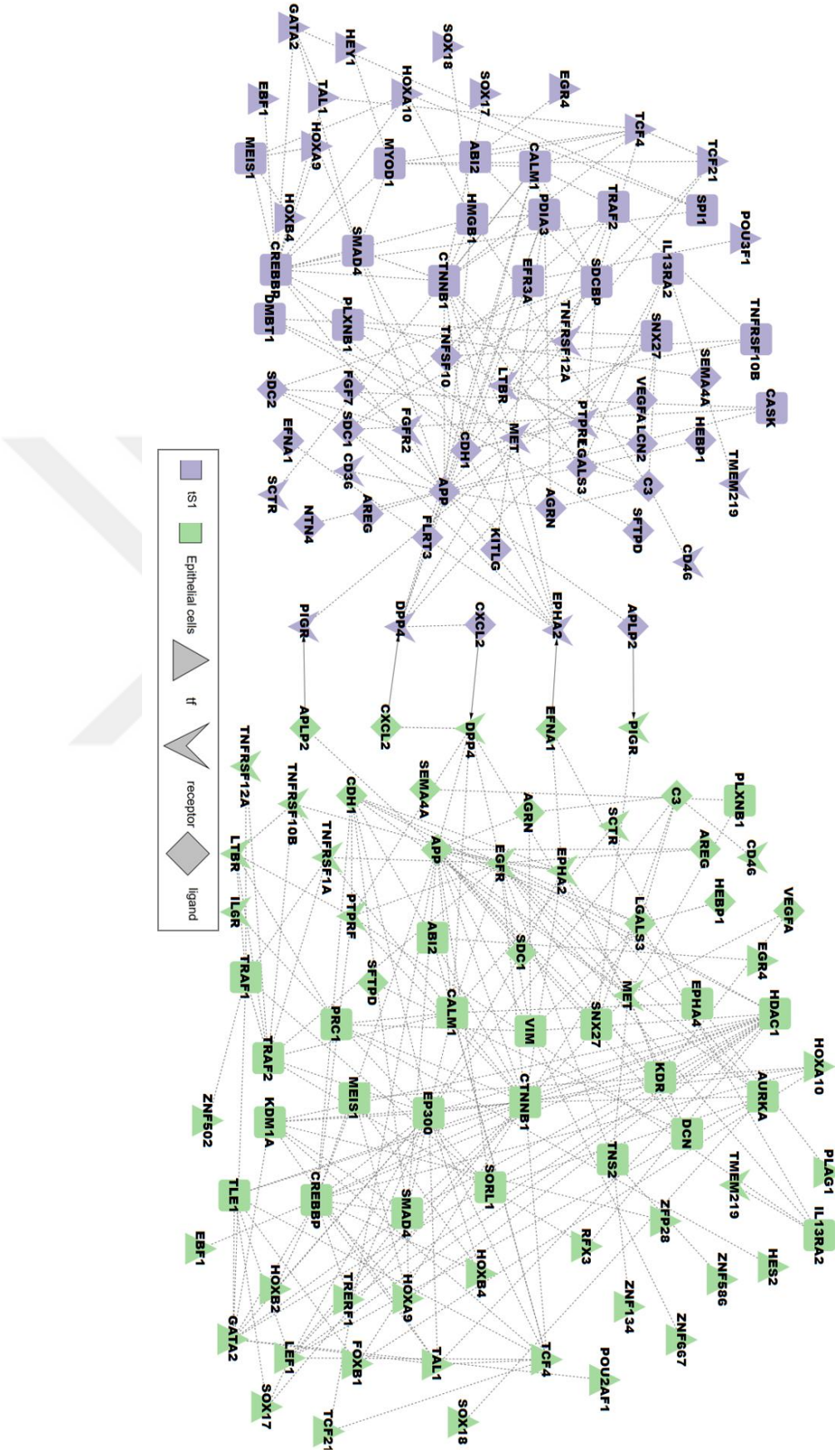

tS1
Epithelial cells
tf
receptor
ligand
GATA2
HEY1
SOX18
SOX17
EGR4
TCF4
TCF21
EBF1
TAL1
HOXA10
MEIS1
HOXA9
MYOD1
ABI2
CALM1
SPI1
POU3F1
HOXB4
HMGB1
PDIA3
TRAF2
CREBBP
SMAD4
IL13RA2
MBT1
CTNNB1
EFR3A
SDCBP
TNFRSF10B
PLXNB1
TNFSF10
TNFRSF12A
SNX27
CASK
SDC2
FGF7
LTBR
SEMA4A
EFNA1
SDC1
FGFR2
MET
VEGFA
HEBP1
CD36
CDH1
LCN2
TMEM219
SCTR
AREG
APP
PTPRE
GALS3
C3
NTN4
FLRT3
AGRN
SFTPD
KITLG
CD46
PIGR
DPP4
CXCL2
EPHA2
APLP2
CXCL2
DPP4
EFNA1
PIGR
APLP2
TNFRSF12A
TNFRSF10B
CDH1
SEMA4A
AGRN
SCTR
C3
PLXNB1
LTBR
TNFRSF1A
APP
EGFR
EPHA2
CD46
AREG
IL6R
PTPRF
ABI2
SDC1
LGALS3
HEBP1
VEGFA
TRAF1
SFTPD
CALM1
MET
EPHA4
EGR4
HDAC1
PRC1
VIM
SNX27
HOXA10
TRAF2
MEIS1
EP300
CTNNB1
KDR
AURKA
ZNF502
KDM1A
SORL1
TNS2
DCN
PLAG1
TLE1
CREBBP
SMAD4
TMEM219
IL13RA2
EBF1
ZFP28
HOXB4
RFX3
HES2
HOXB2
TRERF1
HOXA9
ZNF586
ZNF134
GATA2
LEF1
FOXB1
TAL1
TCF4
ZNF667
POU2AF1
SOX17
SOX18
TCF21

Figure 4.7. Multi-layer network representation of two cell subtypes. The network includes the intracellular regulatory and intercellular receptor-ligand interaction networks of two cell subtypes in lung adenocarcinoma primary tumor samples (tLung) dataset by Kim et al. Purple nodes represent the transcriptional state of cancer 1 (tS1) in lung adenocarcinoma identified by Kim et al. Green nodes represent the healthy epithelial cells in the primary tumor samples. Triangle-shaped nodes represent the transcription factors (tf), V-shaped nodes represent the receptors, diamond-shaped nodes represent the ligands. Dashed lines represent the intracellular edges. Solid arrows represent the intercellular edges, pointing from ligand towards receptor.

In the intracellular and intercellular network representation that connects the Epithelial cells and tS1, there are 7 receptor-ligand pairs. Epithelial cells intracellular network consists of 157 nodes and 431 edges, where tS1 intracellular network consists of 149 nodes and 431 edges, respectively. In epithelial cells there are 39 intermediate proteins in addition to the terminal nodes, which are found to be in the solution node set of PCSF algorithm solved by Omics Integrator. In tS1, there are 35 intermediate proteins that are in solution of PCSF.

As the final part of this section, we constructed a network in which the nodes represent receptors and ligands and the edges represent the interactions between them, by using the information from CellPhoneDB. We showed the abundance of receptors and ligands in different cell types in the lung adenocarcinoma primary tumor samples depending on the node size, as demonstrated in Figure 4.8.

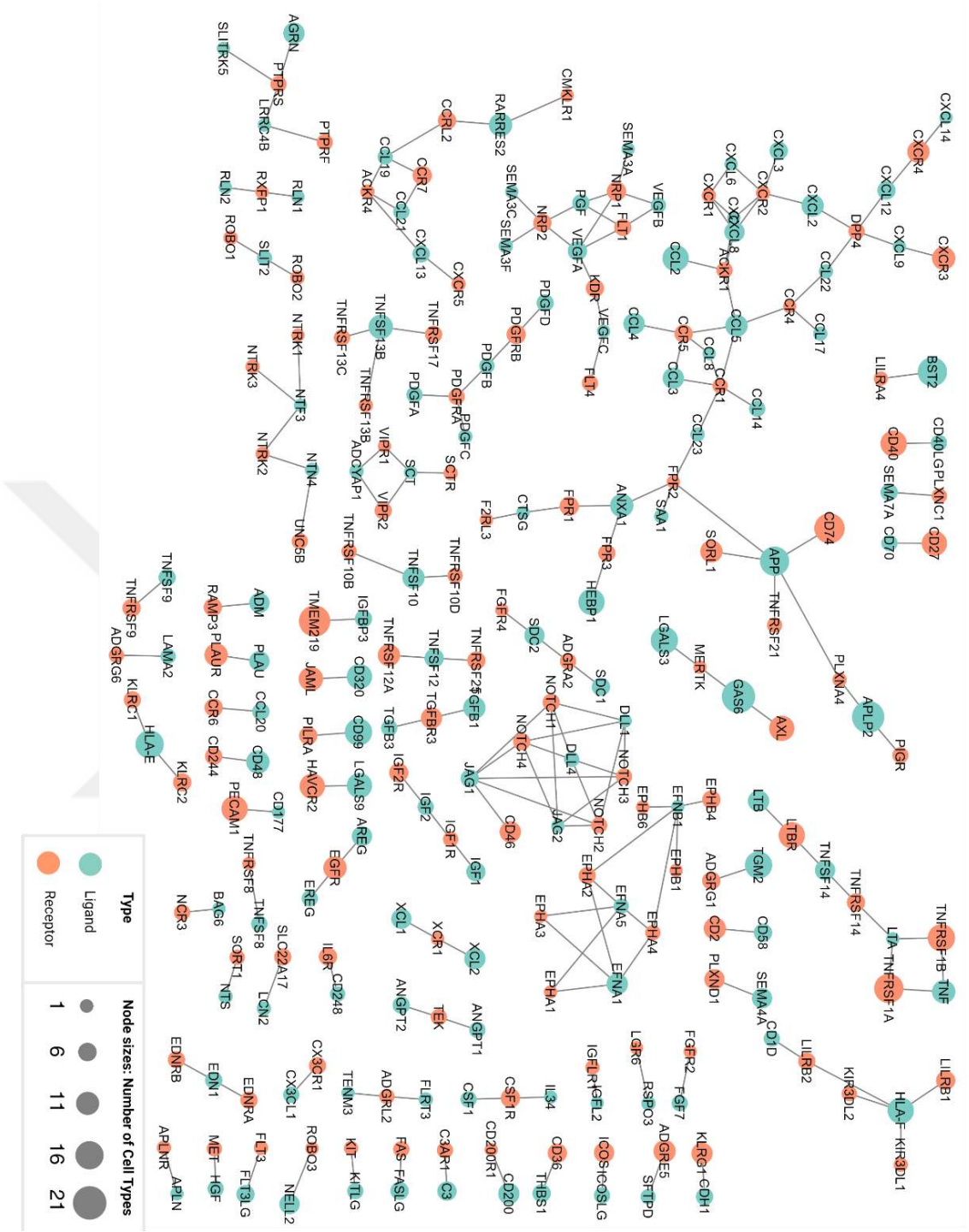


Figure 4.8. Receptor-ligand network according to the cell types. Visualization of receptors and ligands present in the cell subtypes of lung adenocarcinoma primary tumor microenvironment (tLung) from the samples in the dataset by Kim et al. Node sizes represent the number of cell types in which that protein is present. The turquoise color represents the ligands, the orange color represents the receptors.

According to Figure 4.8, the ligand that is abundant in the most cell types is GAS6, with 21 cell types. GAS6, Growth Arrest-Specific 6, is a which builds a complex with TYRO3, AXL, and MER receptor tyrosine kinases, abbreviated as TAM (G. Wu et al., 2018). GAS6/TAM complexes are found to be abundant in tumor microenvironments of many types of cancer, which causes inhibition of cell death, giving rise to the tumor cell proliferation. In addition, GAS6/TAM complex can result in tumor cell migration and metastasis (Mikuteit et al., 2022). Indeed, we also found that AXL, a member of TAM, is present in 13 cell types in the primary tumor samples of lung adenocarcinoma data that we analyzed, ranking one of the highly represented receptors in the dataset.

Secondly, APLP2, Amyloid precursor-like-protein 2, abundant in 20 cell types, is known to be aberrantly expressed in pancreatic cancer, as well as lung metastases (Pandey et al., 2015). Thus, it might be relevant to observe APLP2 expression in lung adenocarcinoma tumor samples, as well.

Thirdly, we detected the presence of IGFBP3/TMEM219 (Insulin-like growth factor binding protein-3/Transmembrane protein 219) pathway with a representation rate with 5 cell types and 19 cell types, respectively. It is suggested that IGFBP3 can be a diagnostic biomarker for many types of tumors, and IGFBP3/TMEM219 signaling pathway might be a potential therapeutic target for cancer (Cai et al., 2020).

4.3 Comparative and condition-dependent analysis of cell subtypes

4.3.1 Healthy cells and cancer cells in primary tumor samples

In this part, we firstly filtered the single cells of samples from lung adenocarcinoma tumor samples only to healthy epithelial cells and cancer cells, which are categorized into three different transcriptional states, which are abbreviated as tS1, tS2 and tS3. These states are determined by Kim et. al in 2020, with differential expression analysis. We plotted these cells in four different clusters by using tSNE as indicated in Figure 4.9.

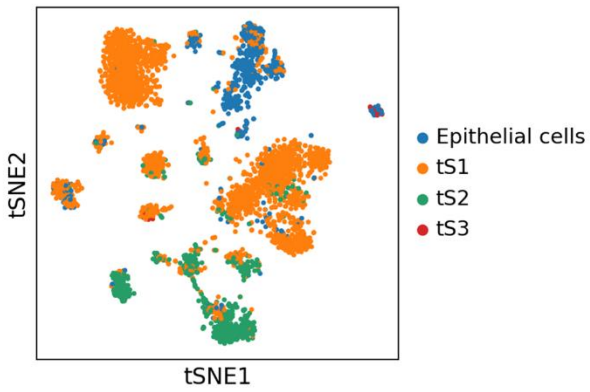


Figure 4.9. tSNE plot of the healthy epithelial and cancer cells. Total number of 3263 cells taken from the lung adenocarcinoma tumor area. Clusters were determined via scanpy, using Louvain clustering method after filtering out the cells that are out of interest in this comparative analysis. tS1, tS2 and tS3 are the clusters of cancer cells with three different transcriptional states, which are determined by Kim et al. by differential expression analysis.

In addition, we ranked the genes with t-test that are differentially expressed in these four cell types according to their fraction of cells in group and mean expression in group, and we plotted the first five genes with the highest ranks among these cell types as shown in Figure 4.10.

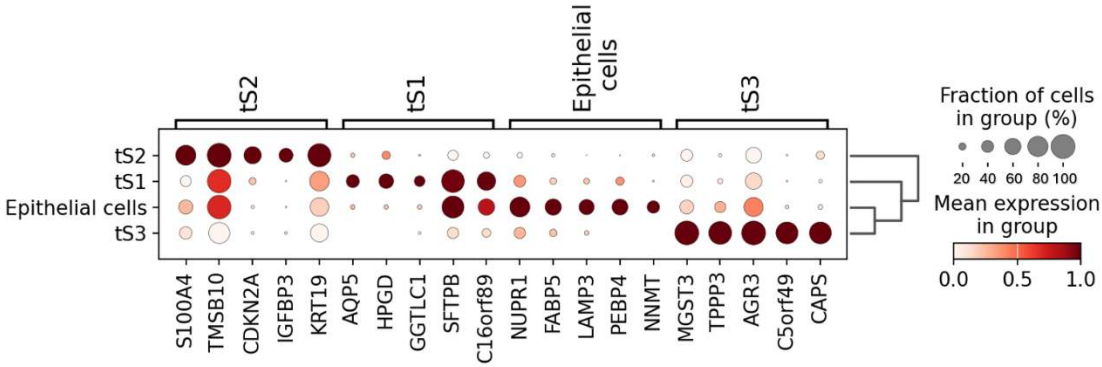


Figure 4.10. Dot plot of the top five highly expressed genes among four cell subtypes. tS1, tS2 and tS3 represent the three different transcriptional states determined by Kim et al. by performing differential expression analysis. Epithelial cells are healthy cells. The color of the dots represents the mean expression in each cell subtype. The size of the dots represents the fraction of cells in the group.

Among the highest ranked genes according to Figure 4.7, S100A4 (S100 calcium-binding protein A4) is a gene, the accumulation of which is used for prognosis for lung cancer (Kagimoto et al., 2023). TMSB10 (Thymosin β10) is known to contribute to tumor progression in most of the solid cancer types (Zeng et al., 2020). These are two of the "marker" genes for tS2 epithelial cancer cell types.

Additionally, TPPP3 (Tubulin polymerization promoting protein family member 3) has been reported to promote cell proliferation and tumor cell migration in lung carcinoma (Li et al., 2018), which is one of the highest ranked genes in tS3.

Aquaporins (AQPs) have been shown to play a significant role in cancer development and progression. AQP5 (Aquaporin 5) is thought to be a potential player in lung adenocarcinoma (Jaskiewicz et al., 2023). Indeed, AQP5 is the highest ranked gene in tS1.

To look from a more general point of view in terms of biological processes, we filtered our differentially expressed genes in these four cell types and performed Over-Representation Analysis in WebGestalt. We compared the resulting enriched biological processes with a False Discovery Rate smaller than or equal to 0.05 in these cell types in a heatmap according to their -log10(p-value), as seen in Figure 4.11.

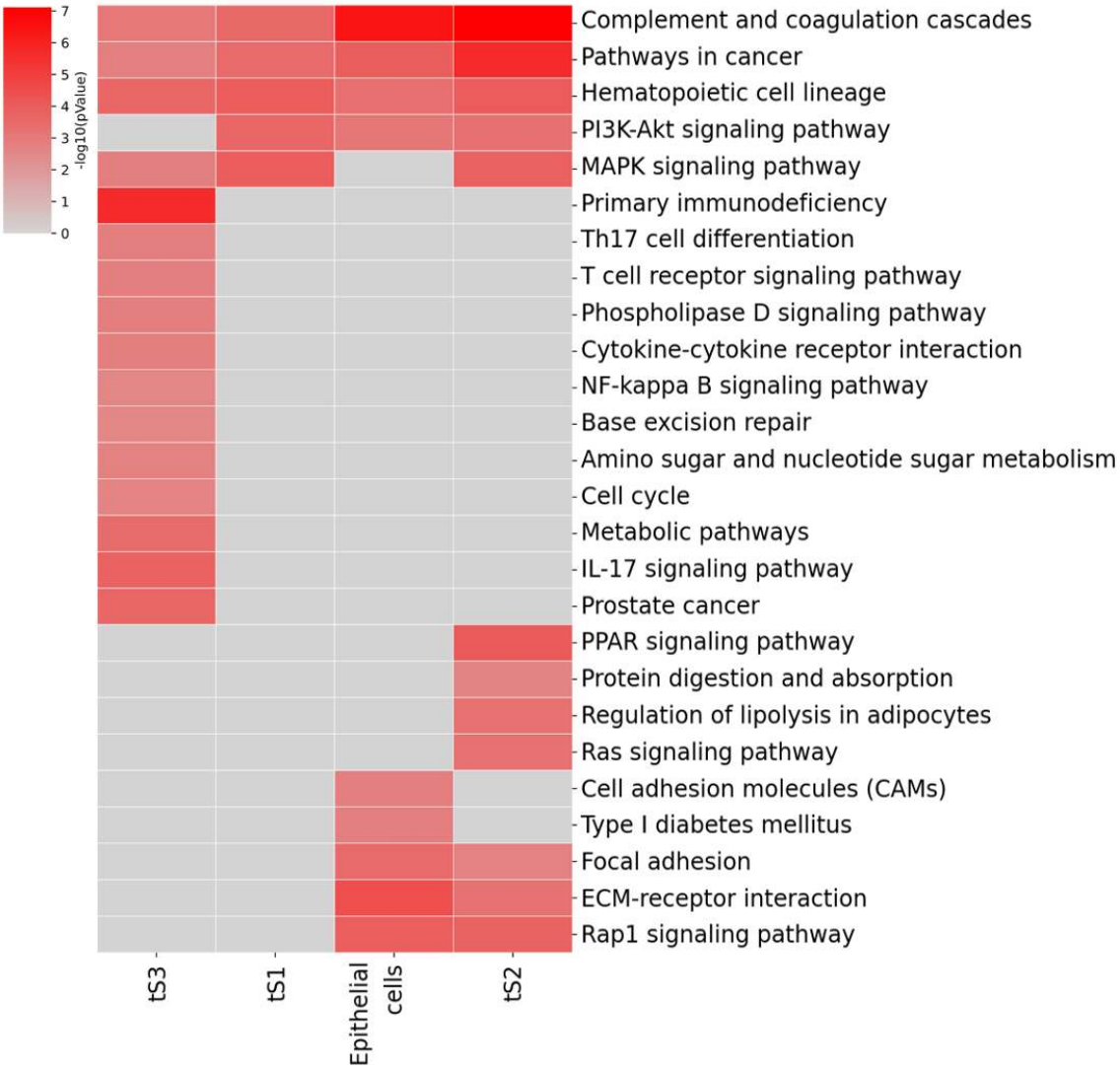


Figure 4.11. Comparative heatmap of biological processes that are significantly present in four different cell types. The cell types are healthy Epithelial cells, and cancer cells with three different transcriptional states, tS1, tS2 and tS3. The heatmap includes WebGestalt Over-representation Analysis results of these four cell subtypes with a false discovery rate (FDR) smaller than or equal to than 0.05. The coloring of heatmap is done by indicating -log10(p-value) of every biological process in four cell subtypes.

In Figure 4.11, we observed that MAPK signaling pathway is over-represented in all three transcriptional states of cancer, tS1, tS2, and tS3, but not in healthy Epithelial cells. Indeed, activation of MAPK signaling pathway is an observed phenomenon in lung adenocarcinoma. Therefore, choosing MAPK pathway as a therapeutic target may yield promising results for lung adenocarcinoma (Stutvoet et al., 2019).

Primary immunodeficiency is also highly over-represented in tS1, which is reported to give rise to autoimmune disease, serious infections, and cancer (Salavoura et al., 2008). Furthermore, RAS signaling pathway is over-represented in tS3. It is known that RAS signaling pathway activation is oncogenic and RAS activation has been observed in 84% of lung adenocarcinoma (East et al., 2022). T-cell receptor signaling pathway is also over-represented in tS1, differently than Epithelial cells. Indeed, the increased crosstalk between lung adenocarcinoma cancer cells and T-cells is reported to play a vital role in promotion of tumor development (Z. Chen et al., 2020). Lastly, the over-representation of IL-17 signaling pathway validates the information that in non-small cell lung cancer, the activation of IL-17 pathway may promote the growth of tumors by increasing the anti-apoptotic capacity of cells in the microenvironment and enhancing cell proliferation (Z. Wu et al., 2019).

Additionally, we constructed network representations of two cell types separately, healthy Epithelial cells and tS1, respectively. The nodes of these networks consist of the transcription factors, receptors, ligands and intermediate proteins that are found to be in the solution of the PCSF algorithm of Omics Integrator. The intracellular network representations are shown in Figure 4.12 and 4.13.

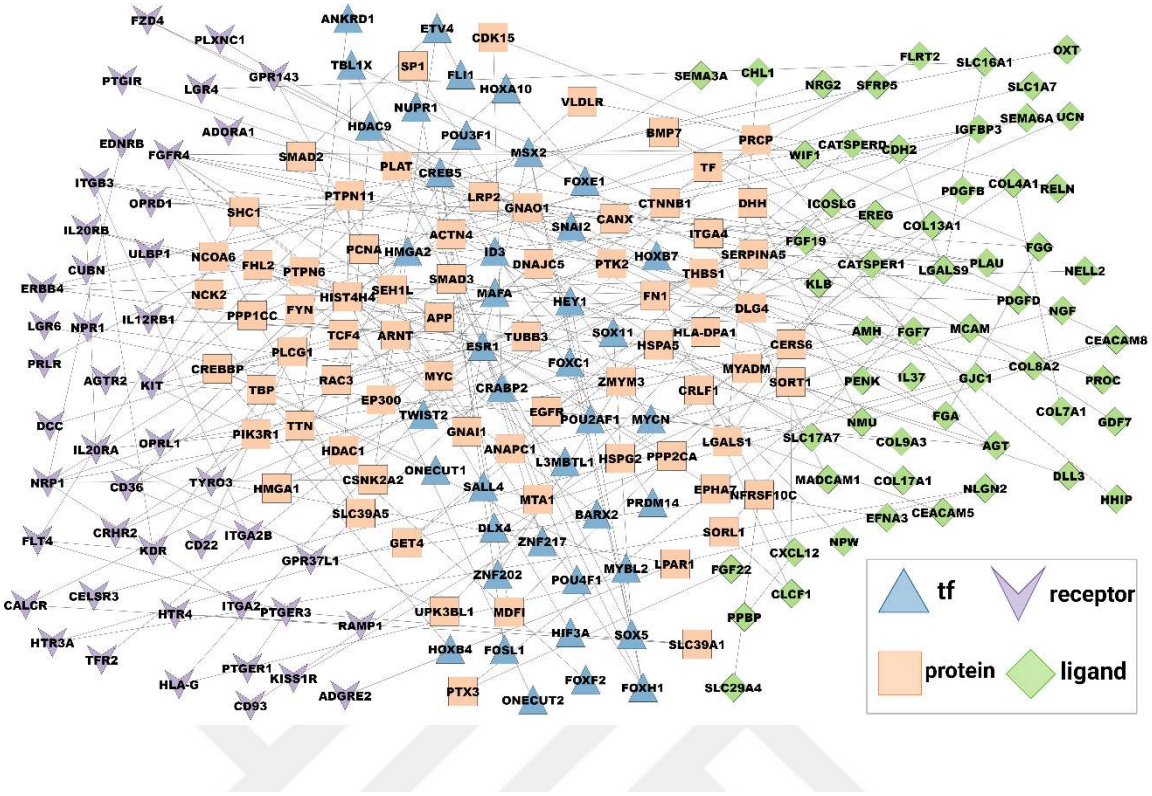


Figure 4.12 Intracellular network of healthy epithelial cells. We visualized the Omics Integrator output as differentially expressed genes of transcription factors, receptors and ligands in healthy epithelial cells against three different transcriptional states of cancer cells in lung adenocarcinoma (tS1, tS2 and tS3) determined by Kim et al. In this network, a differentially expressed gene is defined by an adjusted p-value smaller than 0.05 and an absolute log2 fold change value greater than 2. Blue triangles represent the transcription factors (tf), purple V shapes represent the receptors, orange rectangles represent the intermediate proteins that are in the solution of PCSF algorithm solved by Omics Integrator 2, green diamonds represent the ligands.

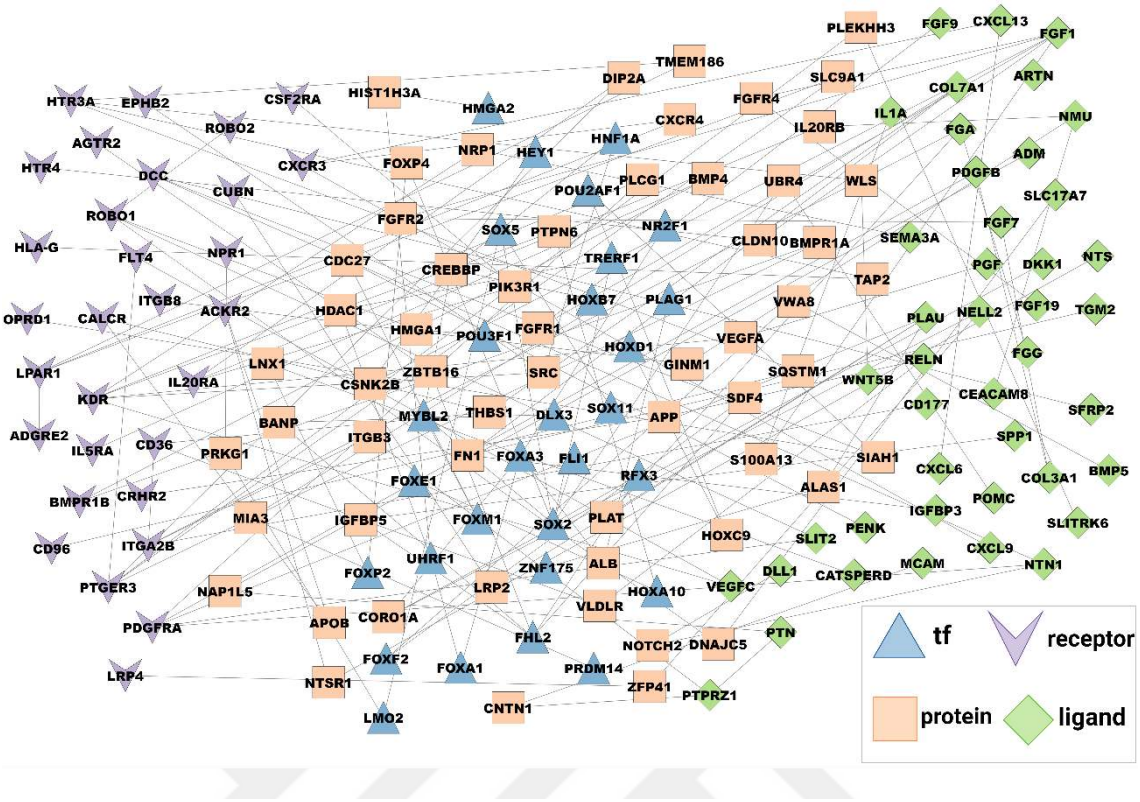


Figure 4.13. Intracellular network of cancer cells. We visualized the Omics Integrator output as differentially expressed genes of transcription factors, receptors and ligands in transcriptional state of cancer 1 (tS1) against healthy epithelial cells and two different transcriptional states of cancer cells in lung adenocarcinoma (tS2 and tS3) determined by Kim et al. In this network, a differentially expressed gene is defined by an adjusted p-value smaller than 0.05 and an absolute log2 fold change value greater than 2. Blue triangles represent the transcription factors (tf), purple V shapes represent the receptors, orange rectangles represent the intermediate proteins that are in the solution of PCSF algorithm solved by Omics Integrator, green diamonds represent the ligands.

We analyzed the two networks in Figures 4.12 and 4.13, respectively. The intracellular network of healthy Epithelial cells in lung adenocarcinoma primary tumor samples yielded 216 nodes, which consists of transcription factors, intermediate proteins, receptors and ligands. tS1 has 160 nodes with the same composition. There are 66 nodes in common in these two intracellular networks, which gives a similarity percentage of 30.56%. We calculated the similarity percentage of tS1 to healthy Epithelial cells by dividing the total number of common nodes to the total number of Epithelial cell nodes. Additionally, we revealed that the difference percentage in Epithelial cells is %69.44 and

the difference percentage in tS1 is 58.75%. We calculated these percentages by dividing the number of unique nodes in both networks to the total number of the nodes in the corresponding network. In Table 4.1, we provide a breakdown of the comparative analysis of these two networks.

Table 4.2. Node coverage analysis of healthy Epithelial cell intracellular network and the transcriptional state of cancer 1 (tS1) in lung adenocarcinoma primary tumor samples.

Total number of nodes in Epithelial cells	216
Total number of nodes in tS1	160
Common nodes	66
Similarity (%)	30.56
Difference (%) of Epithelial cells	69.44
Difference (%) of tS1	58.75

After constructing these networks, we performed another Over-Representation Analysis in WebGestalt where we included the intermediate proteins revealed by Omics Integrator in the gene list for these two cell types. We plotted a heatmap to visualize the significantly enriched biological processes as shown in Figure 4.15.

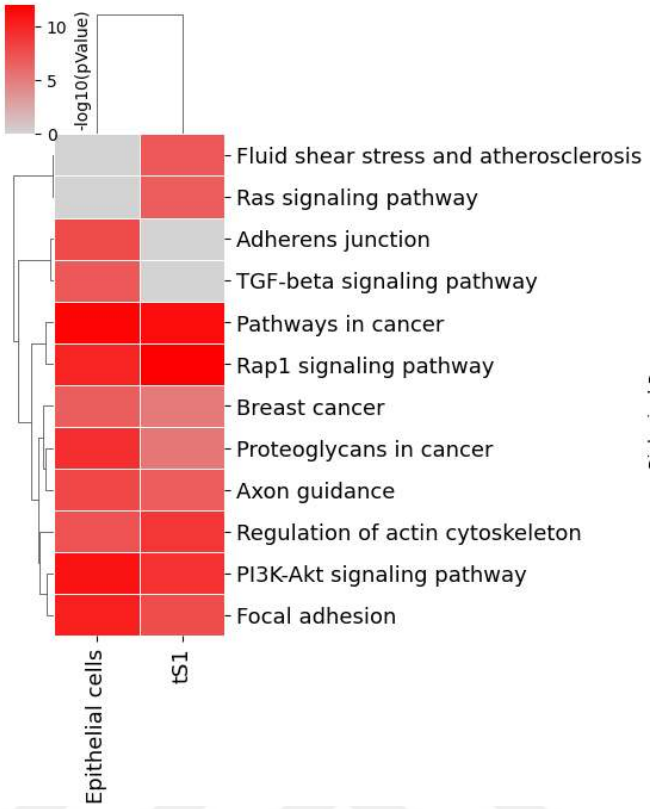


Figure 4.14. Comparative heatmap of biological processes that are significantly present in healthy Epithelial cells and cancer cells (tS1). These cell types are located in lung adenocarcinoma primary tumor samples (tLung). The gene list introduced to WebGestalt includes the intermediate proteins (Steiner nodes) that are present in the network representations of Epithelial cells and tS1 in Figure X and Y, respectively. The heatmap includes WebGestalt Over-representation Analysis results of these five cell subtypes with a false discovery rate (FDR) less than or equal to 0.05. The coloring of heatmap is done by indicating -log10(p-value) of biological processes in both cell subtypes.

In Figure 4.15, TGF-beta signaling pathway is over-represented in healthy Epithelial cells. However, it is not even represented in transcriptional state of cancer in lung adenocarcinoma (tS1). Our findings in this context matches with the literature, where it is stated that TGF-beta signaling pathway repression can increase the invasiveness of cancer and TGF-beta type II receptor, which is a member of the pathway, acts as a tumor suppressor (Toonkel et al., 2010). A key factor in cancer cell migration and metastasis is the suppression of cell adhesion molecules, causing a reduction in cell-to-cell adhesion. Lack of expression of proteins associated to adherens junction is a common issue in lung adenocarcinoma (Kim et al., 2013). Indeed, we found that adherens junction is over-

represented in healthy Epithelial cells but not in tS1, which may have resulted in being one of the driver biological processes. Additionally, it has been reported that increased fluid shear stressed is associated with the apoptotic processes, tumor invasion, cell proliferation and metastasis of cancer cells (Huang et al., 2018). Our findings support this information in a way that fluid shear stressed is over-represented in tS1, but not in healthy Epithelial cells. Lastly, the difference in over-representation in RAS signaling pathway appears the same as in Figure 4.8.

4.3.2 Cancer cells in tumor and metastatic cancer cells

Moreover, we elaborated on the significant differences between cancer cells in lung adenocarcinoma primary tumor samples and the brain metastases of the disease. In that sense, we mapped these cells onto a tSNE plot in Figure 4.16. The cell types are categorized into five, namely healthy epithelial cells, malignant cells of the brain metastases, and tS1, tS2 and tS3, which are the three transcriptional states of cancer in the primary tumor samples.

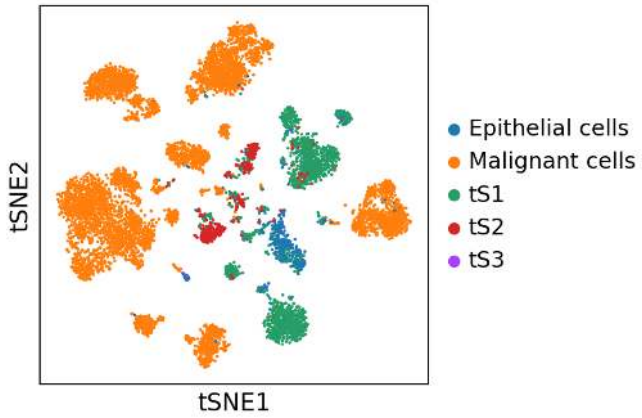


Figure 4.15. tSNE plot of healthy epithelial cells and cancer cells from primary tumor site, and from brain metastases. Total number of 9982 cells with 3620 gene counts taken from the lung adenocarcinoma tumor and brain metastases. Clusters were determined via SCANPY, using Louvain clustering method after filtering out the cells that are out of interest in this comparative analysis. tS1, tS2 and tS3 are the clusters of cancer cells with three different transcriptional states; malignant cells are the cancer cells from brain metastases which are determined by Kim et al. by differential expression analysis.

Additionally, we performed the t-test gene ranking to reveal differentially expressed genes in these four cell types according to their fraction of cells in group and mean expression in group, and we visualized the first five highest-ranked genes in these cell types in a dot plot as shown in Figure 4.3.2.2.

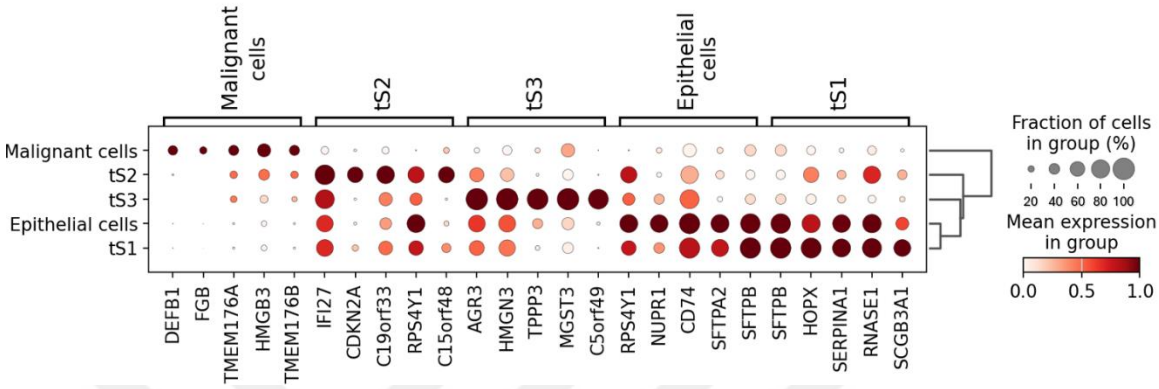


Figure 4.16. Dot plot of the top five highly expressed genes in five cell subtypes. tS1, tS2 and tS3 represent the three different transcriptional states determined by Kim et al. by performing differential expression analysis. Malignant cells are cancer cells in brain metastases. Epithelial cells are healthy cells in both brain and lungs. The color of the dots represents the mean expression in each cell subtype. The size of the dots represents the fraction of cells in the group.

In Figure 4.17, the highest ranked genes from brain metastases of lung adenocarcinoma are indicated. Upregulation of HMGB3 (high mobility group protein B-3) is known to promote tumor development, as well as cell proliferation, cell migration and invasion, as it also contributes to inhibition of apoptosis (Wen et al., 2021). Furthermore, TMEM176A (Transmembrane protein 176A) is also one of the top genes in Malignant cells of brain metastases. It has revealed that TMEM176A plays a crucial role in tumor progression in glioblastoma cells, and it is clinically utilized for tumor prognosis, as a biomarker (Z. Liu et al., 2018). TMEM176B (Transmembrane protein 176B) is also found to be differentially expressed in brain metastasis patients with small-cell lung carcinoma (Mikaeili Namini et al., 2022). Indeed, our results match these findings, as we revealed TMEM176B to be differentially expressed in brain Malignant cells.

In this context, we extracted the differentially expressed genes as the output of t-test based gene ranking in these five cell types and conducted Over-Representation Analysis in WebGestalt. We compared the enriched biological processes with a False Discovery

Rate smaller than or equal to 0.05 in these cell types in a heatmap based on their -log10(p-value), as demonstrated in Figure 4.18.

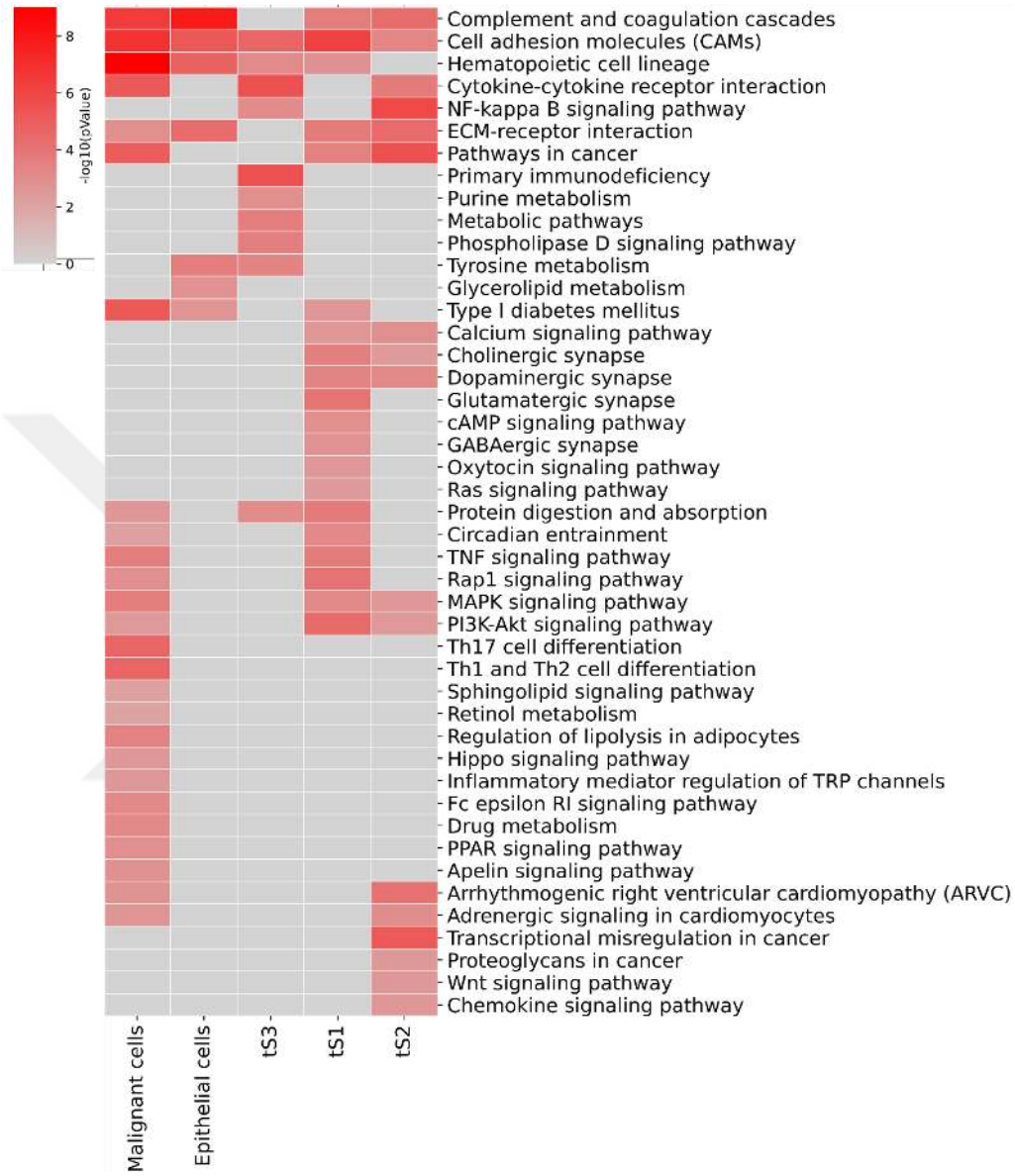


Figure 4.17. Comparative heatmap of biological processes that are significantly present in five different cell types. tS1, tS2 and tS3 represent the three different transcriptional states determined by Kim et al. by performing differential expression analysis. Malignant cells are cancer cells in brain metastases. Epithelial cells are healthy cells in both brain and lungs. The heatmap includes WebGestalt Over-representation Analysis results of these five cell subtypes with a false discovery rate (FDR) less than or equal to 0.05. The coloring of heatmap is done by indicating -log10(p-value) of every biological process in four cell subtypes.

In Figure 4.18, we demonstrated the differing biological processes in brain metastases, namely Malignant cells, to the healthy Epithelial cells and cancer cells in the primary tumor samples. First, pathways in cancer are over-represented in brain metastases and in tS1 and tS2, but not in healthy Epithelial cells. Moreover, we focused on the relevant biological processes that are over-represented in brain metastases, but not in healthy epithelial cells and cancer cells in the primary tumor samples. We found that apelin signaling pathway is over-represented in brain metastases, but not in others. Indeed, it is reported that apelin signaling pathway is involved in metastasis of lung adenocarcinoma and might be considered a therapeutic target to prevent metastasis in lung adenocarcinoma patients (L. Liu et al., 2021).

4.4 Network modeling of tumor

Looking from another point of view, we visualized the cell subtypes in the tumor samples that might be closely associated with the three types of cancer cells. These cell types are healthy epithelial cells, endothelial cells with their subtypes and fibroblasts with their subtypes. We visualized them in another tSNE graph as shown in Figure 4.15.

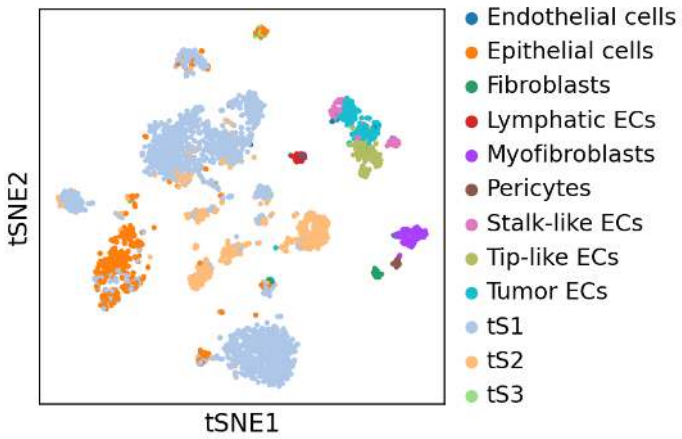


Figure 4.18. tSNE plot of primary tumor site associated cell types. There is a total number of 3505 cells taken from the lung adenocarcinoma tumor samples. Clusters were determined via SCANPY, using Louvain clustering method after filtering out the cells that are out of interest. tS1, tS2 and tS3 are the clusters of epithelial cancer cells with three different transcriptional states, ECs are subtypes of Endothelial cells. Pericytes and myofibroblasts are subtypes of fibroblasts.

Additionally, we studied the number of receptor-ligand interactions between the cell types mentioned in Figure 4.4.1, based on their expression levels greater than 0.01 in the scRNA-seq matrix, as well as their presence in CellPhoneDB database. A representative for this network modeling for two cell types in the dataset is in Figure 4.20.

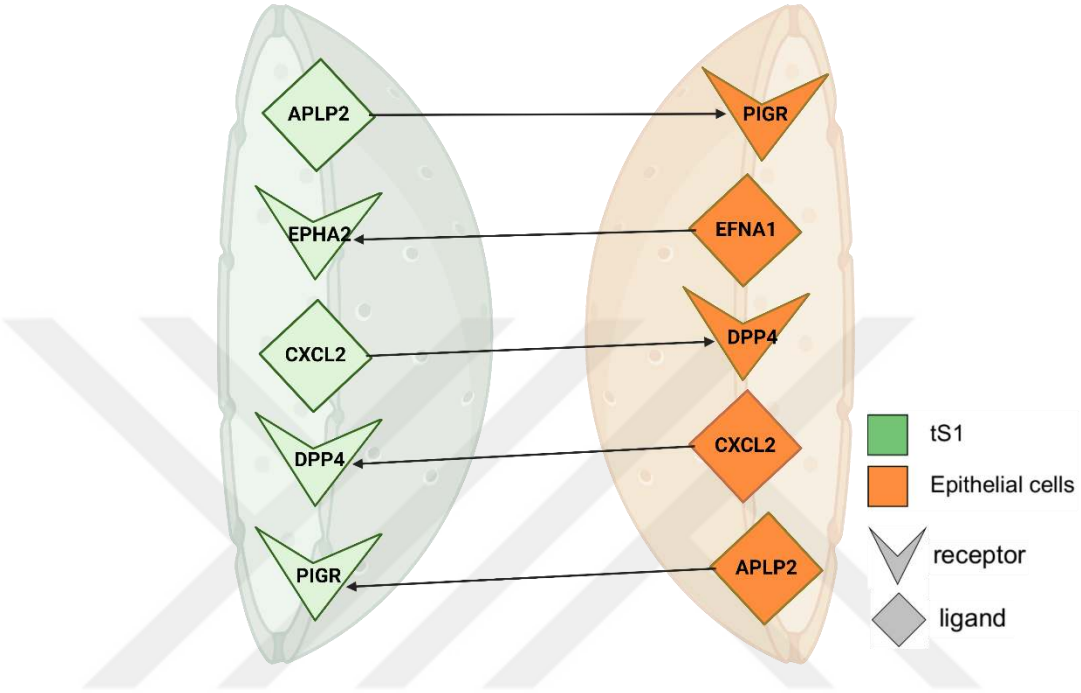


Figure 4.19. Network modeling of crosstalk between healthy and cancer cells. It represents the network visualization of the receptors and ligands that are compatible to each other in two cell subtypes in lung adenocarcinoma primary tumor samples (tLung) dataset by Kim et al. Green cellular model and the green shapes represent the transcriptional state of cancer 1 (tS1) in lung adenocarcinoma identified by Kim et al. Orange cellular model and the orange shapes represent the healthy epithelial cells in the primary tumor samples. V-shaped nodes represent the receptors, and the diamond-shaped nodes represent the ligands.

In Figure 4.20, we found 5 ligand-receptor pairs to be significantly present, to provide a better insight into the intercellular layer that we added to our multi-scale network modeling of lung adenocarcinoma tumor framework. In addition to this, we performed this part for every cell type that might be closely associated with healthy and cancer epithelial cells with every combination possible. We transferred our results to a network

of the corresponding cell types and their subtypes in the lung adenocarcinoma primary tumor samples based on the number of receptor-ligand interactions proportional to the edge thickness in Figure 4.21.

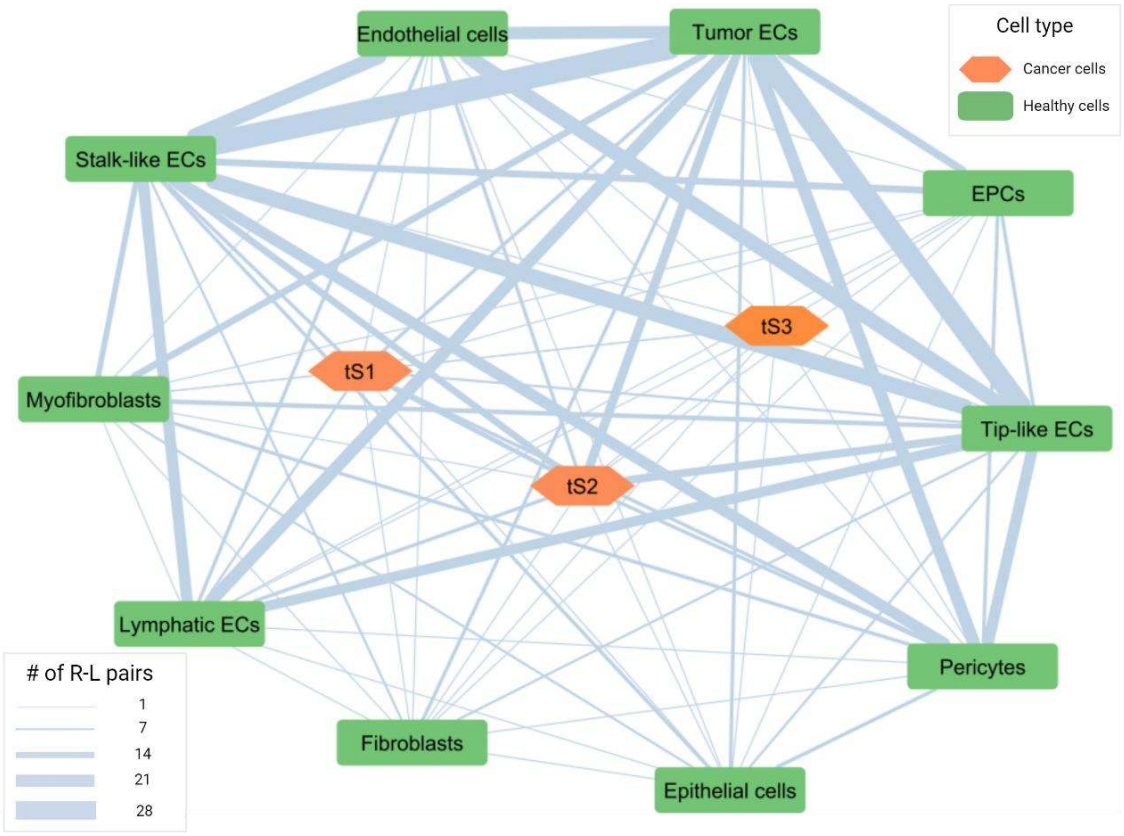


Figure 4.20. Network of cell subtypes in primary tumor site. We visualized of the connectedness regarding the number of receptor-ligand pairs among cell subtypes of lung adenocarcinoma primary tumor samples (tLung) from the samples in the dataset by Kim et al. tS1, tS2 and tS3 are the clusters of epithelial cancer cells with three different transcriptional states, ECs are subtypes of Endothelial cells. EPCs are endothelial progenitor cells. Pericytes and myofibroblasts are subtypes of fibroblasts. Edge thickness represents the number of receptor-ligand (R-L) pairs between two nodes. Orange hexagons represent the cancer cells, green rounded rectangles represent the healthy cells.

In Figure 4.21, the tightest connection, in other words the greatest number of receptor-ligand interactions, of a cancer cell type and a healthy cell type is between the Tumor ECs and tS2, with 15 receptor ligand pairs. Findings in the literature indicate that tumor stromal cells, which develop unique properties within the tumor

microenvironment, enhance tumor invasion. The new blood vessel formation plays a profound role in tumor progression. These blood vessels carry nutrients and oxygen and facilitate metastasis. Tumor endothelial cells, which the tumor blood vessels are composed of, show numerous altered phenotypes compared to normal endothelial cells, promoting tumor progression (Maishi & Hida, 2017). Additionally, the other subtypes of endothelial cells have tight connections to all the cancer cells, tS1, tS2 and tS3, as well. In the literature, it is stated that there are endothelial cell subtypes promoting tumor progression via cytokine secretion, which activates receptors on tumor cells and/or inhibits the anti-tumor immune response by reducing the cytotoxic activity of immune cells (Yang et al., 2021). Furthermore, we included fibroblasts and some of its subtypes present in the dataset to our network, which we found to possess receptor-ligand interactions with cancer cells. Indeed, fibroblasts, myofibroblasts and pericytes in the tumor microenvironment may include cancer associated fibroblasts, which have pathobiological roles. They are reported to increase tumor cell growth in lung adenocarcinoma (Min et al., 2021).

Transcription factors, the initial nodes of the intracellular networks in Omics Integrator solutions, were also analyzed in this part. There were 188 transcription factors in total in these 12 cell types which are significantly expressed. We performed Wilcoxon test to obtain z-score for every transcription factor in every cell type, as shown in Figure 4.22.

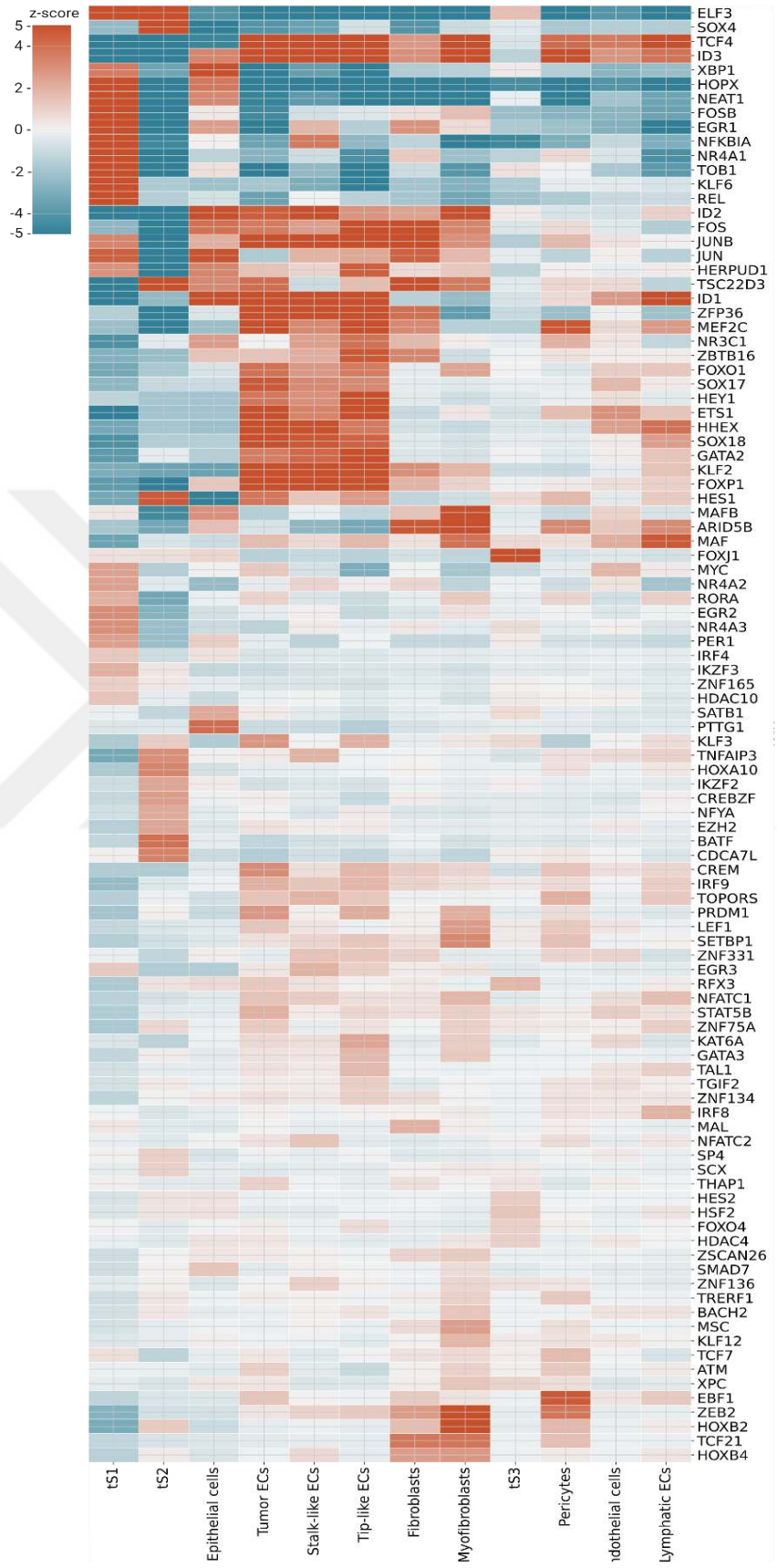

z-score
5
4
2
0
-2
-4
-5
ELF3
SOX4
TCF4
ID3
XBP1
HOPX
NEAT1
FOSB
EGR1
NFKBIA
NR4A1
TOB1
KLF6
REL
ID2
FOS
JUNB
JUN
HERPUD1
TSC22D3
ID1
ZFP36
MEF2C
NR3C1
ZBTB16
FOXO1
SOX17
HEY1
ETS1
HHEX
SOX18
GATA2
KLF2
FOXP1
HES1
MAFB
ARID5B
MAF
FOXJ1
MYC
NR4A2
RORA
EGR2
NR4A3
PER1
IRF4
IKZF3
ZNF165
HDAC10
SATB1
PTTG1
KLF3
TNFAIP3
HOXA10
IKZF2
CREBZF
NFYA
EZH2
BATF
CDCA7L
CREM
IRF9
TOPORS
PRDM1
LEF1
SETBP1
ZNF331
EGR3
RFX3
NFATC1
STAT5B
ZNF75A
KAT6A
GATA3
TAL1
TGIF2
ZNF134
IRF8
MAL
NFATC2
SP4
SCX
THAP1
HES2
HSF2
FOXO4
HDAC4
ZSCAN26
SMAD7
ZNF136
TRERF1
BACH2
MSC
KLF12
TCF7
ATM
XPC
EBF1
ZEB2
HOXB2
TCF21
HOXB4
tS1
tS2
Epithelial cells
Tumor ECs
Stalk-like ECs
Tip-like ECs
Fibroblasts
Myofibroblasts
tS3
Pericytes
ıdothelial cells
Lymphatic ECs

Figure 4.21. Comparative heatmap of transcription factor expression in primary tumor site. We constructed the heatmap of the 188 transcription factors that are significantly expressed in Epithelial cells and their subtypes including cancer cells with three different transcription profiles (tS1, tS2 and tS3), Endothelial cells and their subtypes, Fibroblasts and their subtypes. The Wilcoxon test was applied to the data that yields z-scores for every transcription factor. The coloring of the heatmap ranges from red (the highest) and blue (the lowest) from 5 to -5. The values beyond this interval are colored the same as the interval boundaries. We performed row-wise and column-wise hierarchical clustering for transcription factors and cell subtypes, respectively, for a clearer visualization.

In Figure 4.22, we identified which transcription factors are highly expressed (red) or lowly expressed (blue) across different cell types and subtypes, depending on their z-scores. We determined a group of transcription factors in tS1 that have a very high z-score in comparison to healthy epithelial cells. These transcription factors are ELF3 and HOPX. Indeed, increased expression of HOPX, ELF3, is observed in some lung cancer types (Liang et al., 2019; Z. Zhang et al., 2022). On the contrary, low z-score of NFKBIA, TOB1, ID2, ZFP36, FOXP1 in tS2 validates the findings in the literature that these transcription factors are tumor suppressors, thus, their lack of expression, may give rise to lung adenocarcinoma (Jiao et al., 2012; Peng et al., 2022; Sheng et al., 2019; T. Zhang et al., 2023; Zhu et al., 2023). Another singleton that has a positive z-score unlike the healthy epithelial cells is MYC, which has indeed an oncogene role in lung cancer and poses a potential therapeutic target (Massó-Vallés et al., 2020).

Chapter 5:
DISCUSSION AND CONCLUSION

Cancer is a disease that originates from uncontrolled cell proliferation, invasion of tumors and aberrant malfunction of signaling pathways. Since the tumor microenvironment is composed of cancer cells, as well as immune cells, healthy cells and numerous other cell types, utilizing single-cell omics technologies as a more comprehensive approach may yield more accurate results (Cheng et al., 2020). In order to elaborate on the primary tumor, we built a framework to analyze single cell transcriptomics data. This framework allows us to reconstruct intracellular signaling networks and the crosstalk between different cell types and subtypes. We applied this framework to lung adenocarcinoma (LUAD) data to gain insights into its microenvironment. In that manner, we built gene regulatory networks in every cell subtype in the dataset, which is filtered to those differentially expressed in primary tumor samples. Then, we combined our results with identification of cell subtype specific significant receptor-ligand pairs, which are potential initial nodes for upstream signaling network reconstruction and the formation of a cell-cell interaction network. Additionally, for each cell subtype, we integrated regulatory network components with receptor-ligand pairs, and we applied Omics Integrator tool for intracellular network modeling. Thereafter, we built heatmaps of ligands and receptors across these cell types in which we also indicated the receptor ligand complexes. As a result, we found that IGFBP3 is highly expressed in tS2 cancer cells compared to healthy epithelial cells, fibroblasts, endothelial cells, and mesothelial cells, linked to lung cancer progression through IGF1 signaling. Additionally, PLAU expression is elevated in tS2 cancer cells, a factor in non-small cell lung cancer growth, though no receptors for PLAU were found in our dataset. LCN2, significantly higher in tS2 cancer cells, plays a role in epithelial cell differentiation and various cancers, suggesting a relation to lung adenocarcinoma. ANGPT1, expressed higher in tS3 and fibroblasts, acts as a tumor suppressor, potentially increasing lung cancer patient survival. TNFRSF21, with higher expression in tS2, promotes tumor invasion in lung adenocarcinoma, making it a potential therapeutic target. Conversely, EDNRB is less expressed in tS1, tS2, and tS3, with low levels correlating to decreased

survival and inhibited cell migration in lung adenocarcinoma. Lastly, CD36 shows lower expression in cancer cells compared to healthy cells, indicating a tumor suppressor role.

Then, we linked every intracellular network through receptor-ligand complex information that we obtained via CellPhoneDB. We further performed differential gene expression analysis across different conditions, by comparing the healthy and cancer epithelial cells in the primary tumor samples, and by contrasting the malignant cells in the brain metastases and the cancer cells in the primary tumor.

First, we demonstrated a multi-scale network representation of healthy epithelial cells and cancer cells in lung adenocarcinoma primary tumor samples, which is followed by a network of presence of receptors and ligands among the cell subtypes in the same dataset. We found that GAS6/TAM complex abundance is common in most of the cancer types, resulting in proliferative cell behavior and metastasis. In addition, a high representation of APLP2 among the cell subtypes supported the finding that in some lung metastases, its overexpression is observed. Moreover, we identified the IGFBP3/TMEM219 pathway in numerous cell types, proposing that IGFBP3 is a biomarker for various tumors, and the IGFBP3/TMEM219 pathway can be potentially targeted for cancer therapy.

We determined that S100A4, TMSB10, TPPP3, and AQP5 are highly expressed in three transcriptional states of cancer cells (tS1, tS2, and tS3), all of which are known to be involved in oncogenic processes. Furthermore, we discovered an over-representation of the MAPK and RAS signaling pathways, as well as the T-cell receptor signaling and IL-17 pathways in cancer cells, which have been reported to show increased activation in lung cancers.

Next, we constructed intracellular networks for tS1 and healthy Epithelial cells in primary tumor samples and compared the similarities and differences in terms of nodes. We detected 66 proteins to be common in both networks, however, healthy epithelial cells showed a difference percentage of 69.44 as tS1 has 58.75%. In addition, we compared the over-representation of biological processes among these two networks, by including the intermediate proteins. The TGF-beta signaling pathway is over-represented in healthy epithelial cells but absent in the tS1, which aligns with literature stating that TGF-beta pathway repression can increase invasiveness and that the TGF-beta type II receptor acts as a tumor suppressor (Toonkel et al., 2010). A key factor in cancer metastasis is the suppression of cell adhesion molecules, reducing cell-to-cell adhesion. We found that

adherens junctions are over-represented in healthy epithelial cells but not in tS1, contributing to cancer progression. Additionally, fluid shear stress, linked to apoptosis, tumor invasion, and metastasis is over-represented in tS1 but not in healthy epithelial cells.

In the following step, we performed a comparative analysis between the brain metastatic cells, primary tumor cells and healthy epithelial cells. HMGB3, TMEM176A and TMEM176B can be considered oncogenic in the brain tissue according to the literature, which supports our findings. Moreover, we elaborated on the differently over-represented biological processes in brain metastases and found that Pathways in cancer are over-represented in brain metastases and in tS1 and tS2, but not in healthy epithelial cells. We identified the apelin signaling pathway as over-represented in brain metastases but not in the other cell types as it is reported to be involved in lung adenocarcinoma metastasis and may be a therapeutic target to prevent metastasis.

In further analyses, we constructed another network of the fibroblasts, endothelial cells and epithelial cells of the primary tumor samples and represented the strength of their connections with the number of receptor-ligand pairs they possess. We showed that endothelial cells, especially tumor endothelial cells are in a dense crosstalk with cancer cells, which is result of tumor progression and anti-tumor immune response inhibition, as per literature. We also analyzed fibroblasts and their subtypes in our network, revealing the receptor-ligand interactions with cancer cells. Fibroblasts, including myofibroblasts and pericytes, may contain cancer-associated fibroblasts that enhance tumor growth in lung adenocarcinoma.

Lastly, we plotted transcription factors with high and low expression across various cell types based on their z-scores in a heatmap. In tS1, ELF3 and HOPX showed significantly high z-scores compared to healthy epithelial cells, reflecting their increased expression in some lung cancers. On the other hand, low z-scores for NFKBIA, TOB1, ID2, ZPF36, and FOXP1 in tS2 confirmed their roles as tumor suppressors, with their reduced expression potentially contributing to lung adenocarcinoma. Additionally, MYC displayed a positive z-score in tS1, indicating its oncogenic role in lung cancer and potential as a therapeutic target.

By combining intracellular and intercellular networks, we have gone beyond only modeling receptor-ligand interactions or only intracellular networks. For instance, unlike NicheNet (Browaeys et al., 2020), which focuses solely on receptor-ligand interactions

within the tumor microenvironment, and SCENIC (Aibar et al., 2017) which targets solely on intracellular gene regulatory network inference, we aimed to combine both types of networks to create a hypergraph of the tumor. This hypergraph includes every cell subtype in the tumor as a node, and the intercellular crosstalk would compose the edges between them. Every node would be another intracellular network itself.

Furthermore, instead of utilizing previously acknowledged pathway annotations as in NicheNet or SoptSC (S. Wang et al., 2019), we reconstruct those networks and identify those pathways and biological processes at the cell subtype level using our own framework. By extracting terminal node sets specific to cell subtypes, we have become able to create condition-specific or cell-subtype-specific intracellular networks and link them through receptor-ligand interactions.

We are focusing on the tumor region of lung adenocarcinoma in this thesis study; however, to validate our method, we can also apply it to different cancer types and datasets or increase the number of single cells in the datasets we are working on. Additionally, tools such as NicheNet and CytoTalk (Hu et al., 2021), which reveal receptor-ligand interactions across cell types, also include a perturbation module that allows for in silico knock-out or overexpression experiments. The absence of this module in our method is a limitation. Another point that may limit our framework is that the lack of assignment of an importance score for ligands specifically for every cell subtype of interest, which is a module of NicheNet. It provides a prediction of ligand activity, as it might fluctuate but not be stable across different cell types. Lastly, we utilized the annotation information of cell types and subtypes that were readily conducted by Kim et al. in the dataset. However, cell identity annotation is a very well-known challenge for single-cell omics datasets, as only determining marker genes to reveal cell identity might not be sufficient (Luecken & Theis, 2019). In that manner, well-trained machine learning models such as scGPT (Cui et al., 2024), which performs cell type annotations in the light of its training set with over a number of 33 million human single-cells, might be integrated to improve accuracy of cell type annotation.

Additionally, state-of-the-art tools in the literature such as CellRank2 (Weiler et al., 2024), which include an RNA velocity module, can classify cell types into differentiation states such as initial, intermediate, and terminal states based on intercellular interactions, providing a more detailed analysis. In contrast, our framework involves a more static

analysis of intracellular networks and pathways of cell subtypes and the intercellular crosstalk.

This framework can be considered innovative in several ways, especially in terms of the uniqueness of the dataset that was utilized, which is a single cell transcriptomics data. Single cell transcriptomics data supports the aim of this pipeline such that it provides the opportunity to investigate the datasets at cell subtype resolution. In addition, this framework might help extend the knowledge of intercellular networks, by modelling their downstream pathways and visualizing them in a compact way. Another point that makes this pipeline advantageous is that it brings transcriptomics data in multiple layers together. These layers consist of transcription factor-gene interactions and ligand-receptor interactions, where there are additional intermediate (Steiner) nodes might exist through the reconstruction of optimal subnetworks by Prize Collecting Steiner Forest algorithm. The Steiner nodes that are not present in the pathways of the reference interactome might be or not be biologically relevant for the reconstructed network models, however, they should be further investigated through literature, which possesses a high potential for contribution to the discovery of new driver or passenger genes and proteins associated with cancer. Moreover, those models also play a vital role in modelling the tumor microenvironment, as the specificity and significance of the cell-cell interactions through ligand-receptor information should be considered. This pipeline provides this feature as intracellular and intercellular interactions are modelled due to their specificity to the cell subtypes.

Overall, we created a flexible framework for multi-scale network-based analysis of single-cell transcriptomics data. This flexibility means that this pipeline can be enriched by providing more information such as chromatin accessibility data to address the emerging problems in spatial transcriptomics in terms of locating a biological process in a tissue sample by measuring the significant gene activity in different cell types, as well as modeling and discovering the patterns in those reconstructed models.

BIBLIOGRAPHY

Abolfathi, H., Sheikhpour, M., Shahraeini, S. S., Khatami, S., & Nojoumi, S. A. (2021). Studies in lung cancer cytokine proteomics: a review. Expert Review of Proteomics, 18(1), 49–64. https://doi.org/10.1080/14789450.2021.1892491

Aibar, S., González-Blas, C. B., Moerman, T., Huynh-Thu, V. A., Imrichova, H., Hulselmans, G., Rambow, F., Marine, J.-C., Geurts, P., Aerts, J., van den Oord, J., Atak, Z. K., Wouters, J., & Aerts, S. (2017). SCENIC: single-cell regulatory network inference and clustering. Nature Methods, 14(11), 1083–1086. https://doi.org/10.1038/nmeth.4463

Alanis-Lobato, G., Andrade-Navarro, M. A., & Schaefer, M. H. (2017). HIPPIE v2.0: enhancing meaningfulness and reliability of protein–protein interaction networks. Nucleic Acids Research, 45(D1), D408–D414. https://doi.org/10.1093/nar/gkw985

Arici, M. K., & Tuncbag, N. (2023). Unveiling Hidden Connections in Omics Data via pyPARAGON: an Integrative Hybrid Approach for Disease Network Construction. BioRxiv, 2023.07.13.547583. https://doi.org/10.1101/2023.07.13.547583

Badia-i-Mompel, P., Wessels, L., Müller-Dott, S., Trimbour, R., Ramirez Flores, R. O., Argelaguet, R., & Saez-Rodriguez, J. (2023). Gene regulatory network inference in the era of single-cell multi-omics. Nature Reviews Genetics, 24(11), 739–754. https://doi.org/10.1038/s41576-023-00618-5

Bao, Y., Yan, Z., Shi, N., Tian, X., Li, J., Li, T., Cheng, X., & Lv, J. (2024). LCN2: Versatile players in breast cancer. Biomedicine & Pharmacotherapy, 171, 116091. https://doi.org/10.1016/j.biopha.2023.116091

Bateman, A., Martin, M.-J., Orchard, S., Magrane, M., Ahmad, S., Alpi, E., Bowler-Barnett, E. H., Britto, R., Bye-A-Jee, H., Cukura, A., Denny, P., Dogan, T., Ebenezer, T., Fan, J., Garmiri, P., da Costa Gonzales, L. J., Hatton-Ellis, E., Hussein, A., Ignatchenko, A., … Zhang, J. (2023). UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids Research, 51(D1), D523–D531. https://doi.org/10.1093/nar/gkac1052

Berman, H. M. (2000). The Protein Data Bank. Nucleic Acids Research, 28(1), 235–242. https://doi.org/10.1093/nar/28.1.235

Brandes, U., Eiglsperger, M., Lerner, J., & Pich, C. (2013). Graph markup language (GraphML).

Browaeys, R., Saelens, W., & Saeys, Y. (2020). NicheNet: modeling intercellular communication by linking ligands to target genes. Nature Methods, 17(2), 159–162. https://doi.org/10.1038/s41592-019-0667-5

Bruse, N., & Heeringen, S. J. van. (2018). GimmeMotifs: an analysis framework for transcription factor motif analysis. BioRxiv, 474403. https://doi.org/10.1101/474403

Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., & Greenleaf, W. J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nature Methods, 10(12), 1213–1218. https://doi.org/10.1038/nmeth.2688

Cai, Q., Dozmorov, M., & Oh, Y. (2020). IGFBP-3/IGFBP-3 Receptor System as an Anti-Tumor and Anti-Metastatic Signaling in Cancer. Cells, 9(5), 1261. https://doi.org/10.3390/cells9051261

Cao, J., Spielmann, M., Qiu, X., Huang, X., Ibrahim, D. M., Hill, A. J., Zhang, F., Mundlos, S., Christiansen, L., Steemers, F. J., Trapnell, C., & Shendure, J. (2019). The single-cell transcriptional landscape of mammalian organogenesis. Nature, 566(7745), 496–502. https://doi.org/10.1038/s41586-019-0969-x

Chen, G., Ning, B., & Shi, T. (2019). Single-Cell RNA-Seq Technologies and Related Computational Data Analysis. Frontiers in Genetics, 10. https://doi.org/10.3389/fgene.2019.00317

Chen, Z., Yang, X., Bi, G., Liang, J., Hu, Z., Zhao, M., Li, M., Lu, T., Zheng, Y., Sui, Q., Yang, Y., Zhan, C., Jiang, W., Wang, Q., & Tan, L. (2020). Ligand-receptor interaction atlas within and between tumor cells and T cells in lung adenocarcinoma. International Journal of Biological Sciences, 16(12), 2205–2219. https://doi.org/10.7150/ijbs.42080

Cheng, H., Fan, R., & Wei, W. (2020). Cancer Systems Biology in the Era of Single-Cell Multi-Omics. PROTEOMICS, 20(13). https://doi.org/10.1002/pmic.201900106

Cui, H., Wang, C., Maan, H., Pang, K., Luo, F., Duan, N., & Wang, B. (2024). scGPT: toward building a foundation model for single-cell multi-omics using generative AI. Nature Methods. https://doi.org/10.1038/s41592-024-02201-0

East, P., Kelly, G. P., Biswas, D., Marani, M., Hancock, D. C., Creasy, T., Sachsenmeier, K., Swanton, C., Downward, J., & de Carné Trécesson, S. (2022). RAS oncogenic activity predicts response to chemotherapy and outcome in lung adenocarcinoma. Nature Communications, 13(1), 5632. https://doi.org/10.1038/s41467-022-33290-0

Efremova, M., Vento-Tormo, M., Teichmann, S. A., & Vento-Tormo, R. (2020). CellPhoneDB: inferring cell–cell communication from combined expression of multi-subunit ligand–receptor complexes. Nature Protocols, 15(4), 1484–1506. https://doi.org/10.1038/s41596-020-0292-x

Gil, D. P., Law, J. N., & Murali, T. M. (2017). The PathLinker app: Connect the dots in protein interaction networks. F1000Research, 6, 58. https://doi.org/10.12688/f1000research.9909.1

Grandi, F. C., Modi, H., Kampman, L., & Corces, M. R. (2022). Chromatin accessibility profiling by ATAC-seq. Nature Protocols, 17(6), 1518–1552. https://doi.org/10.1038/s41596-022-00692-9

Guney, E., & Oliva, B. (2012). Exploiting Protein-Protein Interaction Networks for Genome-Wide Disease-Gene Prioritization. PLoS ONE, 7(9), e43557. https://doi.org/10.1371/journal.pone.0043557

Hagberg, A., Swart, P., & Schult, D. (2008). Proceedings of the 7th Python in Science Conference (SciPy2008). 11–15.

Han, H., Cho, J.-W., Lee, S., Yun, A., Kim, H., Bae, D., Yang, S., Kim, C. Y., Lee, M., Kim, E., Lee, S., Kang, B., Jeong, D., Kim, Y., Jeon, H.-N., Jung, H., Nam, S., Chung, M., Kim, J.-H., & Lee, I. (2018). TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. Nucleic Acids Research, 46(D1), D380–D386. https://doi.org/10.1093/nar/gkx1013

Hao, Y., Stuart, T., Kowalski, M. H., Choudhary, S., Hoffman, P., Hartman, A., Srivastava, A., Molla, G., Madad, S., Fernandez-Granda, C., & Satija, R. (2023). Dictionary learning for integrative, multimodal and scalable single-cell analysis. Nature Biotechnology. https://doi.org/10.1038/s41587-023-01767-y

Hu, Y., Peng, T., Gao, L., & Tan, K. (2021). CytoTalk: De novo construction of signal transduction networks using single-cell transcriptomic data. Science Advances, 7(16). https://doi.org/10.1126/sciadv.abf1356

Huang, Q., Hu, X., He, W., Zhao, Y., Hao, S., Wu, Q., Li, S., Zhang, S., & Shi, M. (2018). Fluid shear stress and tumor metastasis. American Journal of Cancer Research, 8(5), 763–777.

Huynh-Thu, V. A., Irrthum, A., Wehenkel, L., & Geurts, P. (2010). Inferring Regulatory Networks from Expression Data Using Tree-Based Methods. PLoS ONE, 5(9), e12776. https://doi.org/10.1371/journal.pone.0012776

Jaskiewicz, L., Romaszko-Wojtowicz, A., Doboszynska, A., & Skowronska, A. (2023). The Role of Aquaporin 5 (AQP5) in Lung Adenocarcinoma: A Review Article. Cells, 12(3), 468. https://doi.org/10.3390/cells12030468

Jiang, M., Karsenberg, R., Bianchi, F., & van den Bogaart, G. (2024). CD36 as a double-edged sword in cancer. Immunology Letters, 265, 7–15. https://doi.org/10.1016/j.imlet.2023.12.002

Jiao, Y., Sun, K., Zhao, L., Xu, J., Wang, L., & Fan, S. (2012). Suppression of human lung cancer cell proliferation and metastasis in vitro by the transducer of ErbB-2.1 (TOB1). Acta Pharmacologica Sinica, 33(2), 250–260. https://doi.org/10.1038/aps.2011.163

Jin, S., Guerrero-Juarez, C. F., Zhang, L., Chang, I., Ramos, R., Kuan, C.-H., Myung, P., Plikus, M. V., & Nie, Q. (2021). Inference and analysis of cell-cell communication using CellChat. Nature Communications, 12(1), 1088. https://doi.org/10.1038/s41467-021-21246-9

Kagimoto, A., Tsutani, Y., Kushitani, K., Kambara, T., Mimae, T., Miyata, Y., Takeshima, Y., & Okada, M. (2023). Usefulness of serum <scp>S100A4</scp> and positron-emission tomography on lung cancer accompanied by interstitial pneumonia. Thoracic Cancer, 14(4), 381–388. https://doi.org/10.1111/1759-7714.14757

Kamimoto, K., Stringa, B., Hoffmann, C. M., Jindal, K., Solnica-Krezel, L., & Morris, S. A. (2023). Dissecting cell identity via network inference and in silico gene perturbation. Nature, 614(7949), 742–751. https://doi.org/10.1038/s41586-022-05688-9

Kim, A., KIM, E. Y., CHO, E. N., KIM, H. J., KIM, S. K., CHANG, J., AHN, C. M., & CHANG, Y. S. (2013). Notch1 destabilizes the adherens junction complex through upregulation of the Snail family of E-cadherin repressors in non-small cell lung cancer. Oncology Reports, 30(3), 1423–1429. https://doi.org/10.3892/or.2013.2565

Kim, N., Kim, H. K., Lee, K., Hong, Y., Cho, J. H., Choi, J. W., Lee, J.-I., Suh, Y.-L., Ku, B. M., Eum, H. H., Choi, S., Choi, Y.-L., Joung, J.-G., Park, W.-Y., Jung, H. A., Sun, J.-M., Lee, S.-H., Ahn, J. S., Park, K., … Lee, H.-O. (2020). Single-cell RNA sequencing demonstrates the molecular and cellular reprogramming of metastatic lung adenocarcinoma. Nature Communications, 11(1), 2285. https://doi.org/10.1038/s41467-020-16164-1

Köhler, S., Bauer, S., Horn, D., & Robinson, P. N. (2008). Walking the Interactome for Prioritization of Candidate Disease Genes. The American Journal of Human Genetics, 82(4), 949–958. https://doi.org/10.1016/j.ajhg.2008.02.013

Leiserson, M. D. M., Vandin, F., Wu, H.-T., Dobson, J. R., Eldridge, J. V, Thomas, J. L., Papoutsaki, A., Kim, Y., Niu, B., McLellan, M., Lawrence, M. S., Gonzalez-Perez, A., Tamborero, D., Cheng, Y., Ryslik, G. A., Lopez-Bigas, N., Getz, G., Ding, L., & Raphael, B. J. (2015). Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. Nature Genetics, 47(2), 106–114. https://doi.org/10.1038/ng.3168

Levi, H., Elkon, R., & Shamir, R. (2021). DOMINO: a network-based active module identification algorithm with reduced rate of false calls. Molecular Systems Biology, 17(1). https://doi.org/10.15252/msb.20209593

Li, Y., Bai, M., Xu, Y., Zhao, W., Liu, N., & Yu, J. (2018). TPPP3 Promotes Cell Proliferation, Invasion and Tumor Metastasis via STAT3/ Twist1 Pathway in Non-Small-Cell Lung Carcinoma. Cellular Physiology and Biochemistry, 50(5), 2004–2016. https://doi.org/10.1159/000494892

Liang, H., Wang, C., Gao, K., Li, J., & Jia, R. (2019). microRNA-421 promotes the progression of non-small cell lung cancer by targeting HOPX and regulating the Wnt/β-catenin signaling pathway. Molecular Medicine Reports. https://doi.org/10.3892/mmr.2019.10226

Liao, Y., Wang, J., Jaehnig, E. J., Shi, Z., & Zhang, B. (2019). WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. Nucleic Acids Research, 47(W1), W199–W205. https://doi.org/10.1093/nar/gkz401

Liu, L., Yi, X., Lu, C., Wang, Y., Xiao, Q., Zhang, L., Pang, Y., & Guan, X. (2021). Study Progression of Apelin/APJ Signaling and Apela in Different Types of Cancer. Frontiers in Oncology, 11. https://doi.org/10.3389/fonc.2021.658253

Liu, Z., An, H., Song, P., Wang, D., Li, S., Chen, K., & Pang, Q. (2018). Potential targets of TMEM176A in the growth of glioblastoma cells. OncoTargets and Therapy, Volume 11, 7763–7775. https://doi.org/10.2147/OTT.S179725

Luecken, M. D., & Theis, F. J. (2019). Current best practices in single-cell RNA-seq analysis: a tutorial. Molecular Systems Biology, 15(6). https://doi.org/10.15252/msb.20188746

Maishi, N., & Hida, K. (2017). Tumor endothelial cells accelerate tumor metastasis. Cancer Science, 108(10), 1921–1926. https://doi.org/10.1111/cas.13336

Mao, X., Xu, J., Wang, W., Liang, C., Hua, J., Liu, J., Zhang, B., Meng, Q., Yu, X., & Shi, S. (2021). Crosstalk between cancer-associated fibroblasts and immune cells in the tumor microenvironment: new findings and future perspectives. Molecular Cancer, 20(1), 131. https://doi.org/10.1186/s12943-021-01428-1

Martinez-Outschoorn, U. E., Lisanti, M. P., & Sotgia, F. (2014). Catabolic cancer-associated fibroblasts transfer energy and biomass to anabolic cancer cells, fueling tumor growth. Seminars in Cancer Biology, 25, 47–60. https://doi.org/10.1016/j.semcancer.2014.01.005

Massó-Vallés, D., Beaulieu, M.-E., & Soucek, L. (2020). MYC, MYCL, and MYCN as therapeutic targets in lung cancer. Expert Opinion on Therapeutic Targets, 24(2), 101–114. https://doi.org/10.1080/14728222.2020.1723548

Mellor, J. (2005). The Dynamics of Chromatin Remodeling at Promoters. Molecular Cell, 19(2), 147–157. https://doi.org/10.1016/j.molcel.2005.06.023

Mikaeili Namini, A., Jahangir, M., Mohseni, M., Kolahi, A. A., Hassanian-Moghaddam, H., Mazloumi, Z., Motallebi, M., Sheikhpour, M., & Movafagh, A. (2022). An in silico comparative transcriptome analysis identifying hub lncRNAs and mRNAs in brain metastatic small cell lung cancer (SCLC). Scientific Reports, 12(1), 18063. https://doi.org/10.1038/s41598-022-22252-7

Mikuteit, M., Zschäbitz, S., Erlmeier, M., Autenrieth, M., Weichert, W., Hartmann, A., Steffens, S., & Erlmeier, F. (2022). Growth Arrest-Specific 6 in Chromophobe Renal Cell Carcinoma. Oncology, 100(10), 536–541. https://doi.org/10.1159/000525601

Min, K.-W., Kim, D.-H., Noh, Y.-K., Son, B. K., Kwon, M. J., & Moon, J.-Y. (2021). Cancer-associated fibroblasts are associated with poor prognosis in solid type of lung adenocarcinoma in a machine learning analysis. Scientific Reports, 11(1), 16779. https://doi.org/10.1038/s41598-021-96344-1

Pandey, P., Rachagani, S., Das, S., Seshacharyulu, P., Sheinin, Y., Naslavsky, N., Pan, Z., Smith, B. L., Peters, H. L., Radhakrishnan, P., McKenna, N. R., Giridharan, S. S. P., Haridas, D., Kaur, S., Hollingsworth, M. A., MacDonald, R. G., Meza, J. L., Caplan, S., Batra, S. K., & Solheim, J. C. (2015). Amyloid precursor-like protein 2 (APLP2) affects the actin cytoskeleton and increases pancreatic cancer growth and metastasis. Oncotarget, 6(4), 2064–2075. https://doi.org/10.18632/oncotarget.2990

Peng, W., Chen, J., He, R., Tang, Y., Jiang, J., & Li, Y. (2022). ID2 inhibits lung adenocarcinoma cell malignant behaviors by inhibiting the activation of the PI3K/AKT/mTOR signaling pathway. Tissue and Cell, 79, 101950. https://doi.org/10.1016/j.tice.2022.101950

Pratapa, A., Jalihal, A. P., Law, J. N., Bharadwaj, A., & Murali, T. M. (2020). Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. Nature Methods, 17(2), 147–154. https://doi.org/10.1038/s41592-019-0690-6

Raj, B., Wagner, D. E., McKenna, A., Pandey, S., Klein, A. M., Shendure, J., Gagnon, J. A., & Schier, A. F. (2018). Simultaneous single-cell profiling of lineages and cell types in the vertebrate brain. Nature Biotechnology, 36(5), 442–450. https://doi.org/10.1038/nbt.4103

SALAVOURA, K., KOLIALEXI, A., TSANGARIS, G., & MAVROU, A. (2008). Development of Cancer in Patients with Primary Immunodeficiencies. Anticancer Research, 28(2B), 1263. http://ar.iiarjournals.org/content/28/2B/1263.abstract

Schäfer, S., Smelik, M., Sysoev, O., Zhao, Y., Eklund, D., Lilja, S., Gustafsson, M., Heyn, H., Julia, A., Kovács, I. A., Loscalzo, J., Marsal, S., Zhang, H., Li, X., Gawel, D., Wang, H., & Benson, M. (2024). scDrugPrio: a framework for the analysis of single-cell transcriptomics to address multiple problems in precision medicine in immune-mediated inflammatory diseases. Genome Medicine, 16(1), 42. https://doi.org/10.1186/s13073-024-01314-7

Sheng, H., Li, X., & Xu, Y. (2019). Knockdown of FOXP1 promotes the development of lung adenocarcinoma. Cancer Biology & Therapy, 20(4), 537–545. https://doi.org/10.1080/15384047.2018.1537999

Smoot, M. E., Ono, K., Ruscheinski, J., Wang, P.-L., & Ideker, T. (2011). Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics, 27(3), 431–432. https://doi.org/10.1093/bioinformatics/btq675

Stutvoet, T. S., Kol, A., de Vries, E. G., de Bruyn, M., Fehrmann, R. S., Terwisscha van Scheltinga, A. G., & de Jong, S. (2019). MAPK pathway activity plays a key role in PD-L1 expression of lung adenocarcinoma cells. The Journal of Pathology, 249(1), 52–64. https://doi.org/10.1002/path.5280

Thind, A. S., Monga, I., Thakur, P. K., Kumari, P., Dindhoria, K., Krzak, M., Ranson, M., & Ashford, B. (2021). Demystifying emerging bulk RNA-Seq applications: the application and utility of bioinformatic methodology. Briefings in Bioinformatics, 22(6). https://doi.org/10.1093/bib/bbab259

Toonkel, R. L., Borczuk, A. C., & Powell, C. A. (2010). TGF-2 Signaling Pathway in Lung Adenocarcinoma Invasion. Journal of Thoracic Oncology, 5(2), 153–157. https://doi.org/10.1097/JTO.0b013e3181c8cc0c

Tuncbag, N., Braunstein, A., Pagnani, A., Huang, S.-S. C., Chayes, J., Borgs, C., Zecchina, R., & Fraenkel, E. (2013). Simultaneous Reconstruction of Multiple Signaling Pathways via the Prize-Collecting Steiner Forest Problem. Journal of Computational Biology, 20(2), 124–136. https://doi.org/10.1089/cmb.2012.0092

Tuncbag, N., Gosline, S. J. C., Kedaigle, A., Soltis, A. R., Gitter, A., & Fraenkel, E. (2016). Network-Based Interpretation of Diverse High-Throughput Datasets through the Omics Integrator Software Package. PLOS Computational Biology, 12(4), e1004879. https://doi.org/10.1371/journal.pcbi.1004879

Virshup, I., Rybakov, S., Theis, F. J., Angerer, P., & Wolf, F. A. (2021). anndata: Annotated data. BioRxiv.

Wang, S., Karikomi, M., MacLean, A. L., & Nie, Q. (2019). Cell lineage and communication network inference via optimization for single-cell transcriptomics. Nucleic Acids Research, 47(11), e66–e66. https://doi.org/10.1093/nar/gkz204

Wang, Y. A., Sun, Y., Palmer, J., Solomides, C., Huang, L.-C., Shyr, Y., Dicker, A. P., & Lu, B. (2017). IGFBP3 Modulates Lung Tumorigenesis and Cell Growth

through IGF1 Signaling. Molecular Cancer Research, 15(7), 896–904. https://doi.org/10.1158/1541-7786.MCR-16-0390

Wei, F., Ge, Y., Li, W., Wang, X., & Chen, B. (2020). Role of endothelin receptor type <scp>B</scp> (<scp>EDNRB)</scp> in lung adenocarcinoma. Thoracic Cancer, 11(7), 1885–1890. https://doi.org/10.1111/1759-7714.13474

Weiler, P., Lange, M., Klein, M., Pe'er, D., & Theis, F. (2024). CellRank 2: unified fate mapping in multiview single-cell data. Nature Methods. https://doi.org/10.1038/s41592-024-02303-9

Wen, B., Wei, Y., & Zhao, K. (2021). The role of high mobility group protein B3 (HMGB3) in tumor proliferation and drug resistance. Molecular and Cellular Biochemistry, 476(4), 1729–1739. https://doi.org/10.1007/s11010-020-04015-y

Wittkopp, P. J., & Kalay, G. (2012). Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. Nature Reviews Genetics, 13(1), 59–69. https://doi.org/10.1038/nrg3095

Wolf, F. A., Angerer, P., & Theis, F. J. (2018). SCANPY: large-scale single-cell gene expression data analysis. Genome Biology, 19(1), 15. https://doi.org/10.1186/s13059-017-1382-0

Wu, G., Ma, Z., Cheng, Y., Hu, W., Deng, C., Jiang, S., Li, T., Chen, F., & Yang, Y. (2018). Targeting Gas6/TAM in cancer cells and tumor microenvironment. Molecular Cancer, 17(1), 20. https://doi.org/10.1186/s12943-018-0769-1

Wu, Z., He, D., Zhao, S., & Wang, H. (2019). IL-17A/IL-17RA promotes invasion and activates MMP-2 and MMP-9 expression via p38 MAPK signaling pathway in non-small cell lung cancer. Molecular and Cellular Biochemistry, 455(1–2), 195–206. https://doi.org/10.1007/s11010-018-3483-9

Xie, J., Xiang, D.-B., Wang, H., Zhao, C., Chen, J., Xiong, F., Li, T.-Y., & Wang, X.-L. (2012). Inhibition of Tcf-4 Induces Apoptosis and Enhances Chemosensitivity of Colon Cancer Cells. PLoS ONE, 7(9), e45617. https://doi.org/10.1371/journal.pone.0045617

Yang, D., Guo, P., He, T., & Powell, C. A. (2021). Role of endothelial cells in tumor microenvironment. Clinical and Translational Medicine, 11(6). https://doi.org/10.1002/ctm2.450

Yao, S., Dong, S., Ding, J., Rong, Y., Zhang, Y., Chen, H., Chen, J., Chen, Y., Yan, H., Dai, Z., & Guo, Y. (2020). Sex-specific SNP-SNP interaction analyses within

topologically associated domains reveals ANGPT1 as a novel tumor suppressor gene for lung cancer. Genes, Chromosomes and Cancer, 59(1), 13–22. https://doi.org/10.1002/gcc.22793

Zeng, J., Yang, X., Yang, L., Li, W., & Zheng, Y. (2020). Thymosin β10 promotes tumor-associated macrophages M2 conversion and proliferation via the PI3K/Akt pathway in lung adenocarcinoma. Respiratory Research, 21(1), 328. https://doi.org/10.1186/s12931-020-01587-7

Zhang, T., Qiu, L., Cao, J., Li, Q., Zhang, L., An, G., Ni, J., Jia, H., Li, S., & Li, K. (2023). ZFP36 loss-mediated BARX1 stabilization promotes malignant phenotypes by transactivating master oncogenes in NSCLC. Cell Death & Disease, 14(8), 527. https://doi.org/10.1038/s41419-023-06044-z

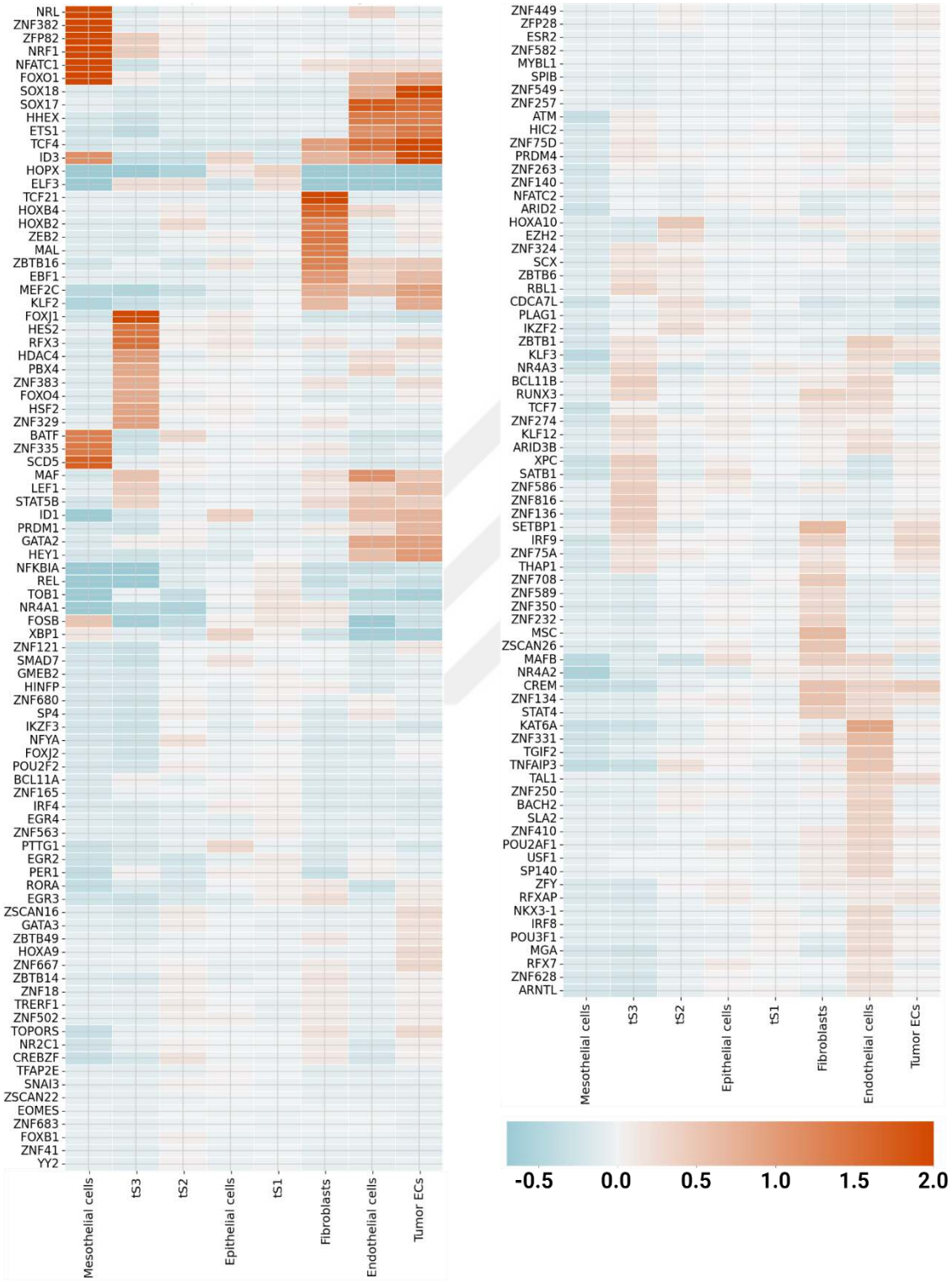
Zhang, Z., Nong, L., Chen, M.-L., Gu, X.-L., Zhao, W.-W., Liu, M.-H., & Cheng, W.-W. (2022). LncRNA ELF3-AS1 Promotes Non-Small Cell Lung Cancer Cell Invasion and Migration by Downregulating miR-212. Cancer Biotherapy and Radiopharmaceuticals, 37(2), 119–124. https://doi.org/10.1089/cbr.2019.3506

Zheng, Y., Zhang, L., Zhang, K., Wu, S., Wang, C., Huang, R., & Liao, H. (2024). PLAU promotes growth and attenuates cisplatin chemosensitivity in ARID1A-depleted non-small cell lung cancer through interaction with TM4SF1. Biology Direct, 19(1), 7. https://doi.org/10.1186/s13062-024-00452-7

Zhou, C., Chen, Z., Liu, J., & Fang, S. (2021). Aberrant upregulation of TNFRSF21 enhances tumor aggressiveness in lung cancer via activation of the ERK/FOXM1 signaling cascade.

Zhu, W., Yu, Y., Ye, Y., Tu, X., Zhang, Y., Wu, T., Ni, L., Huang, X., Wang, Y., & Cui, R. (2023). MiR-196b-5p activates NF-κB signaling in non-small cell lung cancer by directly targeting NFKBIA. Translational Oncology, 37, 101755. https://doi.org/10.1016/j.tranon.2023.101755

Appendix A:


NRL
ZNF382
ZFP82
NRF1
NFATC1
FOXO1
SOX18
SOX17
HHEX
ETS1
TCF4
ID3
HOPX
ELF3
TCF21
HOXB4
HOXB2
ZEB2
MAL
ZBTB16
EBF1
MEF2C
KLF2
FOXJ1
HES2
RFX3
HDAC4
PBX4
ZNF383
FOXO4
HSF2
ZNF329
BATF
ZNF335
SCD5
MAF
LEF1
STAT5B
ID1
PRDM1
GATA2
HEY1
NFKBIA
REL
TOB1
NR4A1
FOSB
XBP1
ZNF121
SMAD7
GMEB2
HINFP
ZNF680
SP4
IKZF3
NFYA
FOXJ2
POU2F2
BCL11A
ZNF165
IRF4
EGR4
ZNF563
PTTG1
EGR2
PER1
RORA
EGR3
ZSCAN16
GATA3
ZBTB49
HOXA9
ZNF667
ZBTB14
ZNF18
TRERF1
ZNF502
TOPORS
NR2C1
CREBZF
TFAP2E
SNAI3
ZSCAN22
EOMES
ZNF683
FOXB1
ZNF41
YY2
Mesothelial cells
tS3
tS2
Epithelial cells
tS1
Fibroblasts
Endothelial cells
Tumor ECs
ZNF449
ZFP28
ESR2
ZNF582
MYBL1
SPIB
ZNF549
ZNF257
ATM
HIC2
ZNF75D
PRDM4
ZNF263
ZNF140
NFATC2
ARID2
HOXA10
EZH2
ZNF324
SCX
ZBTB6
RBL1
CDCA7L
PLAG1
IKZF2
ZBTB1
KLF3
NR4A3
BCL11B
RUNX3
TCF7
ZNF274
KLF12
ARID3B
XPC
SATB1
ZNF586
ZNF816
ZNF136
SETBP1
IRF9
ZNF75A
THAP1
ZNF708
ZNF589
ZNF350
ZNF232
MSC
ZSCAN26
MAFB
NR4A2
CREM
ZNF134
STAT4
KAT6A
ZNF331
TGIF2
TNFAIP3
TAL1
ZNF250
BACH2
SLA2
ZNF410
POU2AF1
USF1
SP140
ZFY
RFXAP
NKX3-1
IRF8
POU3F1
MGA
RFX7
ZNF628
ARNTL
Mesothelial cells
tS3
tS2
Epithelial cells
tS1
Fibroblasts
Endothelial cells
Tumor ECs
-0.5
0.0
0.5
1.0
1.5
2.0

Appendix 1. Heatmap of the expression levels of transcription factors in eight cell types in Table 4.1. The cell types are the ones that are relevant with the cancer cell association. We took the union of the transcription factors that are differentially expressed in these cell types. The expression values are from scRNA-seq data by Kim et al., log2-transformed and normalized.