

T.R.
GEBZE TECHNICAL UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

**IDENTIFICATION AND INVESTIGATION
OF CELL TYPES AND CELL-TO-CELL COMMUNICATION
IN NON-SMALL CELL LUNG CANCER**

CEMAL YILDIZ
A THESIS SUBMITTED FOR THE DEGREE OF
MASTER OF SCIENCE
DEPARTMENT OF BIOENGINEERING

GEBZE

2020

T.R.
GEBZE TECHNICAL UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

**IDENTIFICATION AND INVESTIGATION
OF CELL TYPES AND CELL-TO-CELL
COMMUNICATION IN NON-SMALL CELL
LUNG CANCER**

CEMAL YILDIZ
A THESIS SUBMITTED FOR THE DEGREE OF
MASTER OF SCIENCE
DEPARTMENT OF BIOENGINEERING

THESIS SUPERVISOR
ASSIST. PROF. PINAR PİR

GEBZE
2020

T.C.
GEBZE TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

KÜÇÜK HÜCRE DIŞI AKCİĞER
KANSERİNDE HÜCRE TİPLERİNİN VE
HÜCRE-HÜCRE ETKİLEŞİMLERİNİN
BELİRLENMESİ VE ARAŞTIRILMASI

CEMAL YILDIZ
YÜKSEK LİSANS TEZİ
BİYOMÜHENDİSLİK ANABİLİM DALI

DANIŞMANI
DR. ÖĞR. ÜYESİ PINAR PİR

GEBZE
2020



YÜKSEK LİSANS JÜRİ ONAY FORMU

GTÜ Fen Bilimleri Enstitüsü Yönetim Kurulu'nun 17/07/2020 tarih ve 2020/34 sayılı kararıyla oluşturulan jüri tarafından 10/08/2020 tarihinde tez savunma sınavı yapılan Cemal YILDIZ'ın tez çalışması Biyomühendislik Anabilim Dalında Biyoinformatik ve Sistem Biyolojisi Programında YÜKSEK LİSANS tezi olarak kabul edilmiştir.

JÜRİ

ÜYE

(TEZ DANIŞMANI) : Dr. Öğr. Üyesi Pınar PİR

ÜYE

: Doç. Dr. Tunahan ÇAKIR

ÜYE

: Prof. Dr. Devrim GÖZÜAÇIK

ONAY

Gebze Teknik Üniversitesi Fen Bilimleri Enstitüsü Yönetim Kurulu'nun

...../...../..... tarih ve/..... sayılı kararı.

İMZA/MÜHÜR

SUMMARY

Cancer is a complex and heterogeneous disease caused by generally somatic mutations. These mutations cause abnormality in cell growth, proliferation, migration that leads to tumor formation and eventually secondary tumors (metastasis). The most predominant cancer outcomes are metastasis and relapse that make cancer cells more aggressive. Metastasis is the spread of cancer cells from the primary tumor to surrounding tissues or to distant organs and is the main cause of high cancer morbidity and mortality. In relapse situation, cancer cells stay in a non-growth state which is called “dormancy” and start proliferating after few or more years. These dormant cells can escape immune surveillance and chemotherapy via growth arrest. A sample taken from a patient includes more than one cell type and these cells communicate each other by ligand-receptor interactions. Identifying interactions between cell types could lead to the right treatment of an individual patient. Therefore, molecular interactions should be better understood to decrease or abolish the effects of the disease via targeting the key interaction partners. In this thesis study, single-cell RNA sequencing datasets were analyzed, cell-to-cell communications were identified, interactions were associated with signaling pathways, personalized drug targets were identified and approved candidate drugs were proposed for drug repositioning in lung cancer treatment. The scientific contribution of the thesis will be the determination of potential signaling pathways targets to improve the prognosis of non-small cell lung cancer via an effective treatment strategy, which targets the dormant cells as well as the proliferating cancer cells. The results provide a proof-of-principle in use of scRNA-seq data in personalized medicine practices.

Keywords: Non-Small Cell Lung Cancer (NSCLC), Single-Cell RNA Sequencing, Cell-Cell Communication, Ligand-Receptor Interaction, Cell Type Identification, Dormancy.

ÖZET

Kanser, genellikle somatik mutasyonların sebep olduğu karmaşık ve heterojen bir hastalıktır. Bu mutasyonlar hücre büyümesinde düzensizliklere, çoğalmaya, tümör oluşumuna yol açan göçlere ve sonunda ikincil tümörlere (metastaz) sebep olur. Ortaya çıkabilecek en yaygın durumlar, kanser hücrelerini daha çok agresif yapan metastaz oluşumu ve hastalığın nüksetmesidir. Metastaz kanser hücrelerinin birincil tümörden etraftaki dokulara veya uzak organlara yayılmasıdır ve yüksek kanser hastalık ve ölüm oranının ana sebebidir. Nüks durumunda, kanser hücreleri uyku hali denilen büyümeme durumunda beklerler ve birkaç veya daha fazla yıl sonra tekrar çoğalmaya başlarlar. Bu uyku halindeki hücreler zorlu yaşam koşullarında immün takipten ve kemoterapinin tedavisinden büyümelerini durdurarak kaçarlar. Bir hastadan alınan bir örnek birden fazla hücre tipi içerir ve bu hücreler kendi aralarında reseptör-ligand etkileşimiyle haberleşirler. Hücre tipleri arasındaki etkileşimlerin belirlenmesi bir hasta için doğru bireysel tedavinin uygulanmasını sağlayabilir. Bu sebeple, önemli etkileşim çiftlerini hedef alarak hastalığın etkisini azaltmak veya yok etmek için moleküler etkileşimler daha iyi anlaşılmalıdır. Bu tez çalışmasında, tek hücre transkriptom verileri analiz edildi ve hücre-hücre etkileşimleri belirlendi, sonuçlar sinyal yollarıyla ilişkilendirildi, kişiselleştirilmiş ilaç hedefleri belirlendi ve akciğer kanseri tedavisinde ilaç yeniden konumlandırma için aday ilaçlar önerildi. Tezin bilimsel katkısı, uyku halinde olan ve çoğalan kanser hücrelerini hedefleyen etkili bir tedavi stratejisi ile küçük hücre dışı akciğer kanserinin prognozunu iyileştirmek için potansiyel sinyal yolları hedeflerinin belirlenmesi olacaktır. Sonuçların kişiselleştirilmiş tıp uygulamalarında tek hücre RNA dizileme verisinin kullanımında kavram kanıtı sağlaması hedeflenmiştir.

Anahtar Kelimeler: Küçük Hücre Dışı Akciğer Kanseri (KHDAK), Tek-Hücre RNA Dizileme, Hücre-Hücre Haberleşmesi, Ligand-Reseptör Etkileşimi, Hücre Tipi Tanımlama, Uyku Hali.

ACKNOWLEDGEMENTS

Firstly, I would like to thank my supervisor Asisst. Prof. Pınar PİR for her patience, contribution, support and motivation during my Master's study. I would like to thank the members of my thesis committee, Assoc. Prof. Tunahan ÇAKIR, and Prof. Dr. Devrim GÖZÜAÇIK for their comments and suggestions on my thesis. I thank for the opportunity of being part of a large collaborative project lead by Prof. Gözüaık. I thank Prof. Martin Hemberg and his team for their guidance and assistance during my internship in Wellcome Trust Sanger Institute in Summer 2019.

I gratefully thank the whole project team for the fruitful and friendly collaboration environment, and acknowledge TUBITAK for the financial support (Project Code: 216S489). I thank Erasmus+ Scheme and Sanger Institute for the financial support during my internship.

I would like to thank my teammates Merve KUTAY, Dilara UZUNER and iğdem Hazal TİMUÇİN in this project for their collaboration. I would like to thank each member of PIRISMLAB for their collaboration, always being helpful and creating a pleasant working environment. I am grateful to all members of Computational Systems Biology Group at Gebze Technical University for their collaboration and for the fun times we spent together.

I would like to thank to all my housemates, Ali KAYNAR, Bayram KAŞ, Ömer İLTAŞ and Semih YEŞİLYURT for their supportive attitude and also thank to Halil YILMAZ, Ali KAVALL and Masum TEKİN for their friendship and fun times in Chicken Party on every Friday.

Finally, I have to express very profound gratitude to my mother Figen YILDIZ, my father Mehmet Şeref YILDIZ and my brother Cenk YILDIZ for being a loving family and supporting me at all times.

TABLE of CONTENTS

	<u>Page</u>
SUMMARY	v
ÖZET	vi
ACKNOWLEDGEMENTS	vii
LIST of ABBREVIATIONS and ACRONYMS	x
LIST of FIGURES	xi
LIST of TABLES	xiii
1. INTRODUCTION	1
2. BACKGROUND ASPECTS	3
2.1. Cancer	3
2.2. Lung Cancer	4
2.3. Heterogeneity	5
2.4. Metastasis	6
2.5. Tumor Dormancy	7
2.6. Extracellular Matrix (ECM)	9
2.7. Cancer Treatment	11
2.8. RNA-sequencing	12
3. METHODS	14
3.1. Construction of ECM Component Lists	14
3.2. Single Cell RNAseq Analysis	16
4. RESULTS	20
4.1. ECM Analysis	20
4.1.1. Protein Expression Analysis	21
4.1.2. RNA Expression Analysis	25
4.2. Single Cell Analysis	28
4.2.1. Identification of Cell Types	29
4.2.2. Ligand-Receptor Interactions	40
4.2.3. Drug Repositioning	51

5. DISCUSSION	53
6. CONCLUSIONS	55
REFERENCES	57
BIOGRAPHY	62
APPENDICES	63

LIST of ABBREVIATIONS and ACRONYMS

<u>Abbreviations and Acronyms</u>	<u>Explanations</u>
AC	: Adenocarcinoma
CAF	: Cancer Associated Fibroblast
CTC	: Circulating Tumor Cell
DTC	: Disseminated Tumor Cell
ECM	: Extracellular Matrix
EMT	: Epithelial to Mesenchymal Transition
GEO	: Gene Expression Omnibus
LCC	: Large Cell Carcinoma
NSCLC	: Non-small Cell Lung Cancer
NGS	: Next Generation Sequencing
RNA	: Ribonucleic Acid
SCC	: Squamous Cell Carcinoma
SCLC	: Small Cell Lung Cancer
scRNAseq	: Single Cell Ribonucleic Acid Sequencing
t-SNE	: T-distributed Stochastic Neighbor Embedding

LIST of FIGURES

<u>Figure No:</u>	<u>Page</u>
2.1: The hallmarks of cancer.	4
2.2: Types of tumor dormancy.	8
2.3: Mechanisms of ECM function.	10
4.1: Protein expression comparison of lung, metastasis and non-metastasis tissues as a result of protein expression analysis.	22
4.2: Protein expression comparison of lung, metastasis and gender specific included non-metastasis tissues.	24
4.3: RNA expression comparison between lung, at least one metastasis tissue, and non-metastasis tissues.	26
4.4: RNA expression comparison between lung, at least one metastasis tissue, and gender specific included non-metastasis tissues.	27
4.5: Overview of method.	29
4.6: t-SNE plot of P1 normal tissue data.	30
4.7: Known markers for T cell and B cell in P1 tumor data.	32
4.8: GMM of CD83 marker.	34
4.9: t-SNE plot of P4 data.	34
4.10: Known markers for T cell in P4 normal data.	35
4.11: SingleCellSignalR t-SNE results of P4 data.	38
4.12: Summarized view of interactions between cell types and interaction numbers.	40
4.13: Variable interactions between cell types in tumor data.	41
4.14: Ligand receptor interactions between tumor B, T cells and CAFs and Cancer and Dormant cells.	42
4.15: GO biological process terms enriched by MET and SEMA4D genes.	43
4.16: Most variable interactions between cell types in P4 normal tissue data.	44
4.17: Ligand receptor interactions in normal data between B cell and Dormant cell and T cell and Dormant cell.	45

4.18:	Most variable interactions between cell types in metastasis data.	46
4.19:	Comparison of cancer and CAFs interactions.	46
4.20:	Genes that take part in CAFs and NSCLC cancer cell interactions (but not in metastasis)	47
4.21:	Genes that take part in CAFs and metastatic cancer cell interactions (but not in primary NSCLC tumor)	48
4.22:	The significant interactions between cells in cancer tissue.	49
4.23:	g:Profiler enrichment result of genes taking part in significant interactions in tumor tissue.	50
4.24:	Autocrine interactions of P4 and P1 tumor data.	51

LIST of TABLES

<u>Table No:</u>	Page
4.1: Classification of the tissues included in the ECM list.	20
4.2: Common genes between lung and at least one metastasis tissue as a result of protein expression analysis.	22
4.3: Common genes between lung and metastasis tissue when gender specific tissues are included as a result of protein expression analysis.	24
4.4: Common genes between lung and at least one metastasis tissue as a result of RNA expression analysis.	26
4.5: Markers for each clusters in P1 tumor data.	30
4.6: Markers for each clusters in P1 normal data.	31
4.7: Cell type marker list for P1.	32
4.8: Markers for each clusters in P4 tumor data.	35
4.9: Markers for each clusters in P4 normal data.	36
4.10: Cell type marker list for P4.	37
4.11: Cell type numbers in P4 data SingleCellSignalR analysis.	38
4.12: Cell types and number of assigned cells for all datasets.	39
4.13: Drug targets for NSCLC according to ligand-receptor interactions found in analysis.	52

1. INTRODUCTION

Cancer is a complex disease that is the second common reason of deaths and lung cancer is one of the most common cancer type globally. Cancer is basically caused by mutations in somatic cells. Viral, epigenetic and environmental factors can also lead mediate emergence of cancer. These factors and mutations lead to alterations in cell behavior. Cancer cells become more aggressive after gaining new capabilities such as sustaining their own proliferation, resisting immune response by avoiding immune cells, activation of invasion and metastasis mechanisms to survive in body. Metastasis refers to spread of cancer cells from primary tissue to secondary tissue and responsible for a large part of deaths. After a successful treatment, cancer can reappear in the same or in a different tissue, this type of recurrence is often caused by dormancy. Dormancy is inactive state of cancer cells. Cancer cells stop their growth and become undetectable in dormant state. It is crucial to detect these dormant cells and target them in cancer therapy to avoid the recurrence of the disease

Metastasis and dormancy mechanisms do not only depend on intracellular changes but also alterations in tumor microenvironment. Extracellular matrix is a three-dimensional network of noncellular macromolecules that support tissue structurally and biochemically. Collagens, proteoglycans, integrins are some of these macromolecules. These molecules support tissue architecture and regulate signaling pathways that are important for cancer progression. Cancer cells may metastasize to a distant organ due to suitability of the extracellular matrix of the new organ for survival of the metastatic cells.

Tumors are also heterogeneous due to gradual accumulation of mutations and the types of cells they contain. Tumors have different cell types such as cancer associated fibroblasts, immune cells, epithelial cells, and they communicate with each other via ligand-receptor interactions. For this reason, it is of great interest to identify the composition of extracellular matrix and cellular interactions in tumors for choosing the right target to eliminate both cancer cells and the dormant cells.

Rapid advances in the development of next-generation sequencing (NGS) technologies in the last 10 years have contributed many valuable perspectives into complex diseases such as cancer. Transcriptomics is one of the omics sciences and study

transcriptome that refers to the complete set of RNA molecules in a cell or cell population under specific conditions by using high throughput methods such as microarray or NGS. Recently, NGS based technologies for transcriptomics are progressively focused on the characterization of individual cells. These single cell analyses allow researchers to understand the diseases better. Trajectory analysis, cell type analysis, and cell cycle stage analysis are some of these single cell analysis studies that are often used to investigate disease mechanisms in cellular resolution. There are many software tools developed to analyze single cell data. Most popular one is Seurat which provides a large range of analysis tools, Monocle is a popular tool for trajectory analysis, numerous tools are available and being developed for cell type identification and discovery of cell to cell interactions. Single cell omics data are available on databases such as Gene Expression Omnibus (GEO), Single Cell Expression Atlas from EBI, Single Cell Portal from Broad Institute, Human Cell Atlas and Sequence Read Archive (SRA) form NCBI.

The aim of this study is construction of extracellular matrix components of organs and tissue especially lung and its metastatic organs such as liver, bone marrow, adrenal gland, and to understand the communication between cancer and dormant cells and their interactions with immune cells better to be able to propose efficient therapy strategies. Single cell transcriptomics data containing normal and cancer samples from 4 non-small cell lung cancer patients and a secondary dataset from adenocarcinoma brain metastasis samples were used for investigation of cell to cell interactions in samples. Data were downloaded from GEO, clustering analysis and determination of marker genes that are associated with clusters in samples were performed with Seurat in R programming language. Classification of cell types in samples was performed with algorithm that was developed in Matlab programming language by Kumar and colleagues. Ligand-receptor interactions between cell types were identified by using SingleCellSignalR tool in R. After cell to cell interaction scores were determined, interactions between cancer and dormant cells and between immune cells analyzed and high scored interactions were determined. Finally, drugs that target genes involved in these high scored interactions were identified by using the database DrugBank, and candidate drugs were proposed for drug repositioning towards more effective treatment of lung cancer via using validated drugs to block or increase these interactions to eliminate cancer and dormant cells.

Chapter 2 presents literature information about cancer, dormancy, extracellular matrix and single cell methodologies. Chapter 3 presents Methods that were used in these analyses. Chapter 4 presents the Results. Chapter 5 provides a Discussion of the results and Chapter 6 provides Conclusions of the study.

2. BACKGROUND ASPECTS

2.1. Cancer

Cancer is the disease that causes second largest number of deaths globally after cardiovascular diseases. Cancer is mainly caused by mutations in a single somatic cell. The mutations cause abnormality in cell growth, proliferation, migration that leads to tumor formation and finally secondary tumors (metastasis) which consist of cells with mutations different from their origin. There are about 100 cancer types and all of them follow common progression profiles such as genome instability and mutation, sustaining proliferative signaling, evading growth suppressors, resisting cell death, enabling replicative immortality, inducing angiogenesis, activating invasion and metastasis, avoiding immune destruction, tumor-promoting inflammation, deregulating cellular energetics and different mechanisms mediate the acquisition of these capabilities (Figure 2.1) (Hanahan & Weinberg, 2011).

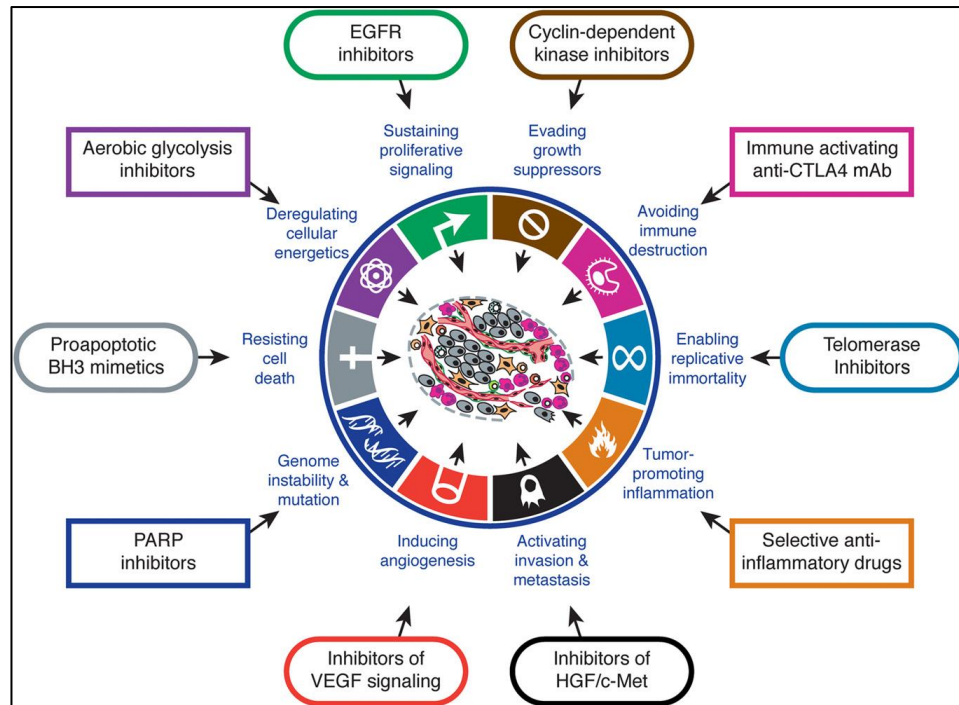


Figure 2.1: The hallmarks of cancer

In a healthy individual, tissues tightly control the signals related to cell growth and division by considering homeostasis in cell number and maintenance of normal tissue architecture and function. However, cancer cells deregulate these signals and control their own growth and division mechanisms via various mechanisms. These deregulation mechanisms can involve somatic mutations that activate growth factor receptors upstream to proliferation pathways such as MAPK, PI3K, TGFB and oncogenes such as RAS, RAF, PTEN, mTOR kinase, or inhibit tumor suppressor genes such as TP53, NF2, and LKB1. As a result of these activating or suppressing mutations, cancer cells gain ability for continuous proliferation, escape from growth suppression and death signals, induction of angiogenesis, and finally tissue invasion and metastasis (Dagogo-Jack & Shaw, 2018).

2.2. Lung Cancer

The most common cancer types are breast cancer for women and prostate cancer for men, whereas lung cancer is the most common cancer type irrespective of gender. However, the most lethal among these three types is lung cancer. Lung cancer is

heterogeneous, which contains several subtypes pathologically and clinically. Lung cancer is divided into two subtypes. The first one is small cell lung cancer (SCLC) which is responsible for ~15% of the cases. The second one is non-small cell lung cancer (NSCLC) responsible for ~85% of the lung cancer cases. Adenocarcinoma (AC) (~40% of cases), squamous cell carcinoma (SCC) (~25–30% of cases) and large cell carcinoma (LCC) (~10–15% of cases) constitute the NSCLC subtypes (Bidkhor, Moeini, & Masoudi-Nejad, 2012; Wood, Pernemalm, Crosbie, & Whetton, 2014).

Adenocarcinoma (AC) is originated from epithelial cells which shows glandular differentiation. Generally, AC takes place at periphery of lung and can be detected by positive marker expression such as napsin A or thyroid transcription factor 1 (TTF1) (Rodriguez-Canales, Parra-Cuentas, & Wistuba, 2016). Squamous cell carcinoma (SCC) is also originated from epithelial cells however shows keratinization. SCC is generally located centrally in the lungs, arising in a main or lobar bronchus and can be identified by positive markers such as p40, p63, and cytokeratins 5/6, while TTF1 is a negative marker (Rodriguez-Canales et al., 2016). Large cell carcinoma (LCC) is also originated from epithelial cells, however it is an undifferentiated NSCLC carcinoma type which shows no differentiation such as glandular differentiation (Rodriguez-Canales et al., 2016). In this study, samples from two AC patients and an AC patient with brain metastasis are analyzed.

2.3. Heterogeneity

Cancer is a complex, dynamic and heterogeneous disease. Heterogeneity in tumor can be mainly divided into intertumoral and intratumoral heterogeneity (Dagogo-Jack & Shaw, 2018). Intertumoral heterogeneity refers to heterogeneity between patients stemming from patient specific factors such as genetic variations, difference in somatic mutations, and environmental changes. Intratumoral heterogeneity refers to heterogeneity between cancer cells in an individual patient in different tissues or in the same tissue but bearing different genotypes (Dagogo-Jack & Shaw, 2018).

Identification of tumor heterogeneity holds a significant promise not only on tumor classification but also on determining the prognosis and choosing the right treatment. So, it is essential to understand the indication of tumor heterogeneity in a regular diagnosis

procedure. Tumors consist of many different cancer cells that have different mutation profiles and also different cell types such as immune cells and non-immunologic cell types (Kumar et al., 2018).

2.4. Metastasis

The main cause of cancer morbidity and mortality are the secondary tumors, also known as metastasis, which are responsible for approximately 90 % of the cancer related deaths. Metastasis is a general term in cancer that refers to spread of cancer cells from the primary tumor to different tissues or organs in the body (Seyfried & Huysentruyt, 2013).

As mentioned above, cancer cells gain some important abilities for out-growth and survival. Another important ability is metastasis caused by changing of tumor microenvironment and dissemination of cancer cells to other regions. Once cancer cells are exposed to a stress, which could be lack of oxygen or nutrition, these malignant cells became invasive cancer cells and settle in new tissues or organs after being transported through blood vessels and lymphatic vessels. This mechanism occurs respectively and termed ‘invasion-metastasis cascade’.

Initially, as a result of epithelial to mesenchymal transition (EMT), invasive cancer cells detach from tissue of origin and invade the tissue layers which is local invasion, second, these cells get into blood flow or lymphatic vessels via a process called intravasation, third, they have to survive in the circulation, fourth, they arrive at a distant organ site, fifth, they get out from the flow and attach to a new organ layer and get in the tissue, this process is called extravasation, sixth, they must survive in a new microenvironment and initiate the micrometastasis formation, and finally primary cancer cells colonize in the new organ which are secondary tumors, thereafter known as metastatic colonization or macrometastasis which may be caused by two types of cells which are circulating tumor cells (CTCs) if they are detected in blood or disseminated tumor cells (DTCs) if they are detected in a secondary tissue such as bone marrow or lymph node (Valastyan & Weinberg, 2011).

2.5. Tumor Dormancy

After successful treatment with surgery, radiotherapy, and chemotherapy, the disease still can relapse many years later. As mentioned above, abilities acquired by cancer cells are important for disease prognosis. Another important ability is tumor dormancy. Tumor dormancy can be explained as a state of sleep during which the cancer cells stop their cell cycle process, and they are not detectable during treatment. Cancer cells can stay and survive in the dormant state for several years, they are undetectable even under stresses such as immunologic attack or lack of oxygen. During this period, CTCs or DTCs with inhibited cell proliferation and activated cell survival pathways stay in the dormant state to be able adapt the new microenvironment (Yadav et al., 2018). Still, cancer metastasis is the most important reason for death of patients however, relapse of dormant cells is responsible for the majority of these deaths. This relapse is caused by generally DTCs that spread from the primary tumor to distant organs however, some primary cancer cells also become dormant.

There are two models of dormancy (Figure 2.2). One of them is cellular dormancy or tumor cell dormancy which refers to G0/G1 arrest in cell cycle process. The other one is tumor mass dormancy or tumor population dormancy which refers to equilibrium between growth rate and cell death, this type of dormancy is further classified as angiogenic and immunologic dormancy (Enderling, Hahnfeldt, Hlatky, & Almog, 2012).

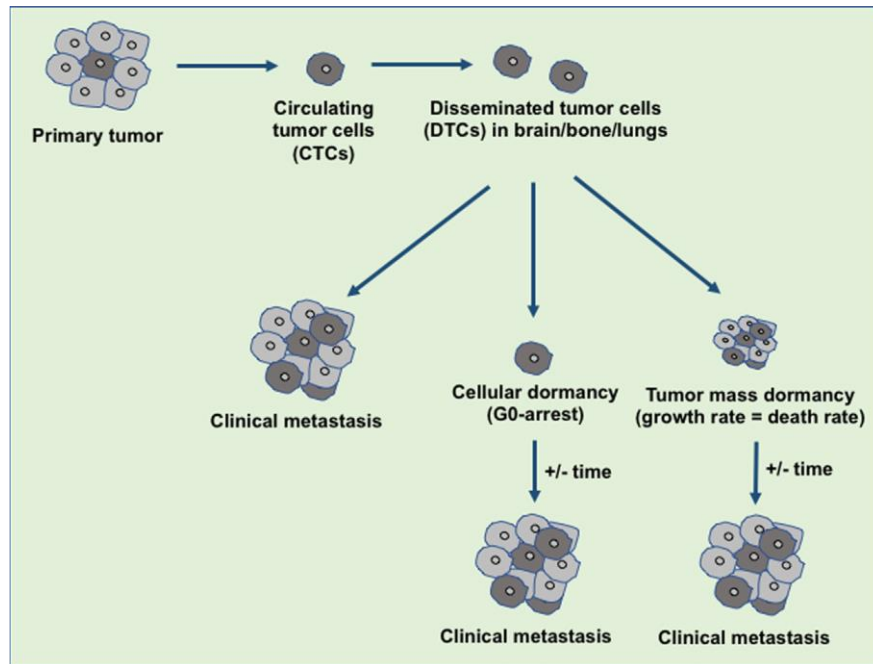


Figure 2.2: Types of tumor dormancy.

Cellular dormancy, also known as solitary cell dormancy, is one of the two models of tumor dormancy. Cellular dormancy can be characterized by minimization of proliferation, minimization of cell death and reversibility (Endo & Inoue, 2019). These characteristic properties emerge when cells are exposed to a stress in their microenvironment. In cellular dormancy, signaling pathways are responsible for the decision whether if the cell will undergo dormancy or not. The inhibition of PI3K-Akt, Ras-MEK-ERK/MAPK, JNK and Wnt pathways drive cancer cells to dormancy (Yadav et al., 2018). It has been shown in many studies that low ERK:p38 ratio induces dormancy and while high ratio induces proliferation. Downregulation of proliferation marker Ki67 is one of the most useful markers to identify of dormant cells. Also interaction between mitogen activated and cyclin dependent kinases such as CDK4, MKK4, MKK6, MKK7, MAPKK4 and stress induced signaling pathways regulate dormancy (Gomis & Gawrzak, 2017; Yadav et al., 2018). Secreted factors in the tumor microenvironment, such as bone morphogenic proteins (BMPs), especially BMP7 and growth arrest-specific 6 (GAS6) inhibit cancer cell proliferation and induce dormancy (Gomis & Gawrzak, 2017). Hypoxia is another stress mechanism in cancer, which refers to lack of oxygen. NR2F1, DEC2,

HIF1 α , p27 genes regulate hypoxia and induce dormancy (Yadav et al., 2018). Also, autophagy mechanisms can activate cellular dormancy.

Another model of dormancy is tumor mass dormancy, which refers to balance between cell proliferation and cell death. There are two mechanisms of tumor mass dormancy. One of them is angiogenic dormancy caused by lack of vascularization. Cancer cells cannot access to enough amount of nutrients and oxygen, therefore they cannot survive under these conditions. Angiogenesis is controlled by pro- and anti-angiogenic factors such as vascular endothelial growth factor (VEGF) and angiostatin. Angiostatin, thrombospondin-1 (TSP1), TGF β 2 molecules induce the angiogenic dormancy. On the contrary, VEGF, TGF β 1, HSP27 inhibits angiogenic dormancy and induce pro-angiogenic factors (Gomis & Gawrzak, 2017). Another model is immunologic dormancy which refers to clearance of cancer cells by immune cells to balance rate of cell death and proliferation. It has three steps; elimination of cancer cells, establishment of equilibrium and escape from immune system (Yeh & Ramaswamy, 2015). Mostly NK cells, macrophages, and T cells take place in this mechanism (Yadav et al., 2018). MHC class I, IL3, IFN γ are inducers of immunologic dormancy. On the other hand, TNF α , IL β , EGF are inhibitors of immunologic dormancy (Gomis & Gawrzak, 2017).

2.6. Extracellular Matrix (ECM)

Extracellular matrix (ECM) has important roles both in metastasis and dormancy. ECM is a three dimensional network of noncellular components such as of water, proteins and polysaccharides that support the cell architecture and environment (Frantz, Stewart, & Weaver, 2010). ECM also regulates tissue behavior, development, and homeostasis. Dysregulation of ECM causes neoplasia. Many macromolecules such as collagens, integrins, and proteoglycans are components of the ECM. These macromolecules regulate most of the cancer hallmarks via receptors such as integrins or growth factor receptors that are located in the ECM and have important roles in signaling pathways (Pickup, Mouw, & Weaver, 2014). Collagens are the most abundant fibrous proteins in ECM and have various functions such as tissue scaffolding, cell adhesion, cell migration, cancer, angiogenesis, tissue morphogenesis and tissue repair. There are 28 types of collagens. Collagens constitute the primary structural component of the ECM, provide tensile

strength, regulate cell adhesion, support chemotaxis and migration, and direct tissue development, so they are important for both metastasis and dormancy mechanisms. Most of the collagens are transcribed and secreted by fibroblasts, therefore fibroblasts are important in cancer disease (Frantz et al., 2010). Various types of collagens, including collagen I, II, III, V, and IX, increase during tumor formation. Also, matrix metalloproteinases (MMPs), which are ECM degrading enzymes, get more active and deregulate ECM formation during aging. Many other ECM components and their receptors such as CD44 and heparan sulfate proteoglycans that enable growth signaling are overexpressed in cancer. Tissue stiffness is an important factor and should be in balance in normal conditions. In tumor tissues, stroma of tumor is 10 times stiffer than normal stroma. In this condition, ECM stiffness help the tumor cell invasion and progression. Lysyl oxidase (LOX), which cross-links collagen fibers and other ECM components facilitates increased tissue stiffness and overexpression of LOX has been observed in many cancers. Overproduction of LOX also induce ERK and PI3K signaling. As a result, all of these ECM components directly or indirectly regulate cancer formation (Figure 2.3)(Lu, Weaver, & Werb, 2012)

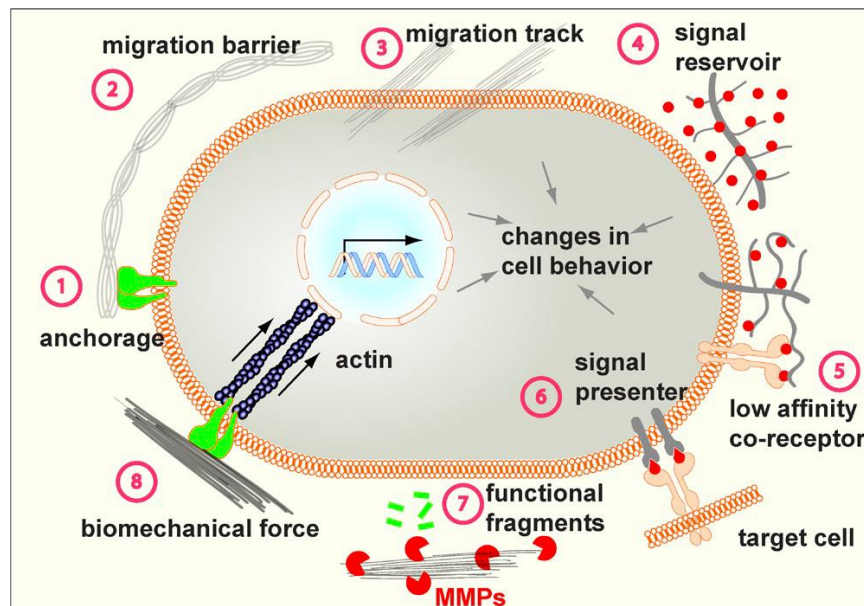


Figure 2.3: Mechanisms of ECM function.

Some ECM components can be unique to specific organs, hence metastasis in some cancer types occur only at specific distant organs, for example, ECM of bone marrow can be different from pancreas, and the differences in ECM may be the reason that cancer cells originated from lung are more metastatic to bone marrow than to pancreas (Peinado et al., 2017).

2.7. Cancer Treatment

There are different types of cancer treatments such as surgery, chemotherapy, radiotherapy which are conventional methods being currently in used in clinic. The type of therapy to be used depends on cancer type and stage of the disease. Surgery aims removal of the tumor and it is recommended for early stage cancers. Surgery is also utilized for diagnosis and staging of cancer as samples can be examined by biopsy (Hirsch et al., 2017; Yildizhan et al., 2018). Another conventional treatment strategy is radiation therapy or radiotherapy which uses high doses of radiation to damage DNA of cancer cells in order to kill or decrease their proliferation. Radiotherapy can be either ‘neoadjuvant’ which means before surgery to reduce tumor size or ‘adjuvant’ to eliminate remaining cancer cells after surgery. Chemotherapy is another strategy for cancer treatment that uses chemical drugs for killing cancer cells by damaging their DNA or stop their growth by blocking the cell cycle (Dasari & Bernard Tchounwou, 2014). Combinations of chemotherapy drugs can be used to reduce the side effects or achieve more effective treatment. Chemotherapy also can be neoadjuvant or adjuvant. These surgery, radiotherapy and chemotherapy are currently standard treatment strategies for cancer. While surgery and radiotherapy are locally applicable, chemotherapy can target cancer cells in different parts of body.

There are also other promising strategies to fight against advanced cancer. Targeted therapies and immunotherapies are examples of these kind of treatments. In targeted therapy, cancer cells are eliminated with drugs that target specific properties of the cancer tissue such as angiogenesis or specific genes which mediate for tumor growth or migration. Erlotinib, Gefitinib, Afatinib are most known drugs that block EGFR protein and inhibit cell growth (Web1, 2020). Bevacizumab, Ramucirumab are monoclonal antibodies used in targeted therapy strategies that target VEGF and its receptor to block

generation of new blood vessels (Beljanski, 2007). Another strategy is immunotherapy which aims to support patient's immune system to fight against the cancer more efficiently. There are several immunotherapy types such as immune checkpoint inhibitors, T-cell transfer therapy, monoclonal antibodies, treatment vaccines, and immune system modulators. Immune checkpoint inhibitors prevent T cell inactivation by blocking PD1/PD-L1 interaction and CTLA4 to keep T cells active and fight against cancer. Monoclonal antibodies can be also used for immunotherapy by marking cancer cells to make them detectable by immune cells. Another immunotherapy type is T-cell transfer also known as adoptive immunotherapy that includes Tumor Infiltrating Lymphocytes (TILs) and chimeric antigen receptor (CAR) T cell therapy. For CAR-T cell therapy, T cells are taken from the patient and CAR protein is induced in T cells and then given back to the patient to recognize and kill cancer cells more effectively. On the other hand, for TIL therapy, immune cells in tumor are removed from patient and grown in laboratory and then given back to patient to fight against cancer cells which are already recognized (Web2, 2020). It is shown that these immunotherapies are effective against cancer (Reynders & De Ruyscher, 2016; Robertson, Salm, & Dangl, 2019), hence many new immunotherapy strategies are under development.

2.8. RNA-sequencing

Transcriptome refers to the complete set of RNA molecules that could be all RNAs or mRNAs in a cell or cell population such as a tissue or specific cells, and their quantity, for a specific developmental stage or condition (Wang, Gerstein, & Snyder, 2009). RNA sequencing is a technique that provides a measure of the transcriptome profile of cells. Bulk RNA sequencing is a milestone for transcriptomic research area. Average gene expression profiles of thousands of cells found in homogenized tissue are measured with this technology. Using bulk RNA sequencing technique, generally differences between two conditions are studied. For example, differentially expressed genes are found between normal and disease states and then genes are identified as drug targets to treat the disease.

The shortcomings of bulk RNA sequencing technology have led scientists to single cell RNA technology. Single cell RNA sequencing was firstly applied by Tang and colleagues in 2009 (Tang et al., 2009). Using this technology, gene expression profile of

every single cell can be measured individually. The arrival of single cell RNA sequencing provides new opportunities for exploring gene expression profile at the single-cell level. Single cell RNA sequencing technique has become favorable for studying the heterogeneous diseases or cells such as cancer or embryos (Chen, Ning, & Shi, 2019). In this technique, first step is isolation of every single cell in the sample by approaches such as flow-activated cell sorting (FACS), laser capture microdissection (LCM), and microfluidics (Hwang, Lee, & Bang, 2018). Reverse transcription, and cDNA amplification follows the cell isolation. After these steps, samples are ready for library preparation and sequencing. After sequencing, sequence reads are analyzed by bioinformatics approaches. Firstly, low quality cells, bases and barcodes are removed from the data via quality control tools. Read alignment is the second step to assign transcripts as exonic, intronic, or intergenic. Last step is normalization to eliminate the cell specific biases in data. After construction of the count matrix, results are ready for further analysis. Many types of downstream analyses can be applied to data such as differential expression analysis, cell population identification and trajectory inference (Chen et al., 2019). There are numerous tools developed for analysis single cell data. Most popular one is Seurat for clustering and find differentially expressed genes, Monocle is another popular method for trajectory analysis, tools for assigning cell types to single cells and finding cell-to-cell interactions have been developed recently and many others are under development (Butler, Hoffman, Smibert, Papalexi, & Satija, 2018; Qiu et al., 2017). The single cell data are stored and freely downloadable from databases such as Gene Expression Omnibus (GEO), Single Cell Expression Atlas from EBI, Single Cell Portal from Broad Institute, Human Cell Atlas and Sequence Read Archive (SRA) from NCBI (Barrett et al., 2013; Papatheodorou et al., 2020).

Single cell RNA sequencing reflects gene expression profile of an individual cell. Single cell sequencing provides powerful tools to scientists in a wide range of research topics such as in developmental biology or neuroscience to understand cell differentiation or identify and characterize neuronal subtypes better in single cell resolution. In cancer, single cell sequencing is also used frequently to understand cancer cells' behavior and mutation profiles of cancer cells (Olsen & Baryawno, 2018). Because of heterogeneity in cancer, single cell sequencing become more important to understand and characterize

cancer development or mutation profile changes between cancer cells. For example, in one study, single cell RNA-sequencing was applied to investigate the cellular composition and transcriptional dynamics of tumor and adjacent normal tissues from 4 early-stage non-small cell lung cancer and a differentiation path from CD14⁺ monocytes to M2 macrophages was identified by using trajectory analysis (GSE117570) (Song et al., 2019). Another study, a unique alternative activation (M2) gene expression pattern of the tumor-associated macrophage in the brain metastasis and a lack of known T cell co-stimulator expression was identified by using single cell RNA-sequencing in brain metastasis of non-small lung cancer (GSE143423) (L. Wang et al., 2019).

Recently, cancer studies have focused on cell type heterogeneity by identifying cell types and uncover their communications with ligand-receptor interaction. In this study, GSE117570 and GSE143423 datasets were analyzed and cell types were assigned to each cell and ligand-receptor interactions were determined to find difference between cancer cells, dormant cells by comparison of normal, tumor and metastasis conditions. Potential drug targets were identified and candidate drugs were proposed.

3. METHODS

3.1. Construction of ECM Component Lists

The Matrisome Project includes an ECM atlas developed in 2016 by Hynes Lab at MIT and Naba Lab at the University of Illinois at Chicago, compiling the proteomic data on ECM components of 14 different tissues and tumor types (Naba et al., 2016).

UniProt, The Universal Protein Resource, was established in 2002 as a result of a collaborative work between the European Bioinformatics Institute (EMBL-EBI), the Swiss Bioinformatics Institute (SIB) and the Protein Information Resource (PIR), and it is run by Alex Bateman (PI), Cathy Wu and Alan Bridge. It is a comprehensive set of protein sequence and annotation data and contains UniProt Knowledgebase (UniProtKB), UniProt Reference Clusters (UniRef) and UniProt Archive (UniParc) databases (Bateman, 2019).

The Human Protein Atlas (ProteinAtlas) is produced by a Swedish-based research program launched in 2003 that uses the integrative omics approach such as antibody-based

imaging, mass spectrometry-based proteomics, transcriptomics and systems biology to map all human proteins in cells, tissues and organs (Uhlén et al., 2015).

While creating the lists used in this project, Human Matrisome data was downloaded from The Matrisome Project website. Then, the generic names of the proteins contained in the extracellular matrix were searched in the UniProt database and those associated with human were downloaded. Finally, “normal_tissue.tsv.zip”, “pathology.tsv.zip” and “rna_tissue.tsv.zip” data were downloaded from the ProteinAtlas database.

“Category”, “Gene Symbol” and “UniProt_IDs” columns of data downloaded from The Matrisome Project website, “Entry” and “Gene names” columns of data downloaded from Uniprot database, “Gene name”, “Tissue”, “Cell type” and “Level” columns of “normal_tissue.tsv.zip” file, “Gene name”, “Cancer”, “prognostic – favorable”, “unprognostic – favorable”, “prognostic – unfavourable” and “unprognostic – unfavourable” columns of “pathology.tsv.zip” file, “Gene name”, “Sample”, “Value” and “Unit” columns of “rna_tissue.tsv.zip” file downloaded from ProteinAtlas database are used for merging the data from these heterogeneous resources.

While downloading data from Uniprot, the terms chemokine, collagen, cytokine ECM, extracellular matrix, glycoprotein, growth factor receptor, growth factor, integrin, lectin, and proteoglycan were searched separately, and the results were downloaded and merged. Data downloaded from Uniprot was filtered in Excel based on the term ‘extracellular’ being in the Gene Ontology (GO) column.

To construct the list, the data downloaded from The Matrisome Project and Uniprot databases were compared using the UniProt IDs via a script written in Matlab R2018a programming language (The Mathworks Inc., 2018). As a result of comparison, the common and database-specific proteins were determined in the data downloaded from the Uniprot database and The Matrisome Project website, and a list was created by merging the proteins from the two databases.

After the list was constructed, the list was reorganized by combining the proteins that were controlled by genes encoding more than one protein. Expression and Pathology information at the Protein and RNA levels for each gene were added to the list using data downloaded from the ProteinAtlas database. The tissues lung, liver, adrenal gland, bone marrow, brain, spleen, heart muscle, and kidney were chosen for protein localization

analysis as these 8 tissues are most common secondary sites for lung cancer metastasis. Protein expression levels and mRNA levels measured in these tissues were added to the list for further investigation

3.2. Single Cell RNAseq Analysis

Two datasets that report single cell mRNA measurement from primary tumors / adjacent normal tissue and brain metastatic tissue / primary adenocarcinoma were used in this project.

Single cell RNAseq data, GSE117570, was downloaded from GEO website. The data includes 4 patients' tumor and adjacent normal samples. Patient 1 (P1), a Caucasian male, has 1832 adenocarcinoma single cell samples and 2495 adjacent normal single cell samples. Patient 2 (P2), a Caucasian female, has 1304 squamous cell carcinoma single cell samples and 1409 adjacent normal single cell samples. Patient 3 (P3), a Caucasian female, has 328 adenocarcinoma single cell samples and 667 adjacent normal single cell samples. Patient 4 (P4), a Caucasian female has 1423 adenocarcinoma single cell samples and 2027 adjacent normal single cell samples. Data from P1 and P4 were chosen for further analysis as they have largest number single cells.

Seurat is the one of most popular R package for single cell RNAseq analysis developed by Satijalab (Butler et al., 2018). Seurat was designed for quality control, analysis and exploration of single cell RNAseq data. It is basically used for quality control and pre-processing of data, dimension reduction, clustering, and differential expression analyses. It also allows integration of multiple datasets, analysis of spatial datasets, analysis of CITE-seq data, integration of single cell ATAC-seq and single cell RNAseq data (Butler et al., 2018).

A Gaussian Mixture Model (GMM) based algorithm for cell type identification and analysis was developed in Matlab by Kumar and colleagues. This algorithm has modules for preprocessing of single cell data by removing cells with low number of genes detected and also normalizes data on log two base by dividing count matrix to the average expression level of house-keeping genes. Cell types are assigned by using a decision tree classifier according to specified marker genes that define each cell type. After cell type

identification, cell-to-cell communication is computed based on level of ligand and receptor expression in each cell type (Kumar et al., 2018).

Another tool for cell-to-cell signaling developed in R is SingleCellSignalR (Cabello-Aguilar et al., 2020). In this tool, raw data is normalized according to 99th percentile which finds 99th quantile of each sample and creates quantile matrix. Then, count matrix is divided to related quantile value/median of quantile matrix (Appendix A). Finally, data is log transformed. Clustering can be done via two methods, single-cell interpretation via multikernel learning (SIMLR) model which learns a similarity measure from single-cell RNA-seq data in order to perform dimension reduction, clustering and visualization or K-means (B. Wang, Zhu, Pierson, Ramazzotti, & Batzoglou, 2017; Zheng et al., 2017). Cell types are assigned based on a given list of gene signatures that represent cell types of interest. Algorithm has also a default marker list that is retrieved from PanglaoDB (Franzén, Gan, & Björkegren, 2019). Algorithm computes the average gene expression of markers across all cells and generates a matrix of gene signature expression. After the matrix is generated, a threshold is determined to maximize the number of cells assigned to a single cell type (as opposed to being assigned to no cell types or multiple cell types). After cell assignment is completed, ligand-receptor interactions are scored between pairs of cell types by considering differentially changed genes that encode ligands and receptors. Most tools including Kumar's algorithm use average expression of ligand and receptor pairs in their analyses (Appendix A). However, in SingleCellSignalR authors introduced a regularized score which is computed by multiplying the square roots of ligand and receptor expressions and dividing the product by sum of mean of expression matrix and square roots of ligand and receptor expressions to control false positives due to high expression genes being favored by non-normalized scores while computing ligand-receptor score (Appendix A) (Cabello-Aguilar et al., 2020). After interactions are found, algorithm generates plots of interactions of interest and interaction subtypes such as paracrine, which means ligand from one cell type and receptor from another cell type are in interaction, or autocrine, which means self-signaling mechanism of a cell type is active.

In this project, normal and tumor data were initially analyzed with Seurat package in R to find cell clusters in the samples and cluster-specific marker genes were identified. Markers were further filtered to find immune and cancer cell associated genes by

comparing the information from databases such as CellMarker and Uniprot (Bateman, 2019; Zhang et al., 2019). After marker genes were determined, a table of marker genes that represent each cell type in data was generated. By using the algorithm developed in Matlab by Kumar and colleagues, cells were assigned cell type labels (such as cancer cell or T-cell) (Kumar et al., 2018). Finally, SingleCellSignalR was used to determine ligand-receptor interactions between the cell types. Further details of each step in the analysis is provided below.

Initially, expression matrix, gene names and cell names of samples were read with “read.csv” function in R. Seurat object was created with “CreateSeuratObject” function. Percentage of mitochondrial genes was determined with “PercentageFeatureSet” function. Data was filtered according to number of detected genes and percentage of mitochondrial genes to exclude samples without an intact cell (number of detected genes has to be greater than 200 and percentage of counts from mitochondrial genes has to be less than 5). Data was normalized with “NormalizeData” function. Variable genes were determined by using “FindVariableFeatures” function. “ScaleData” function is used for scaling. PCA is performed via “RunPCA” function. Neighbor cells were found with “FindNeighbors” function. Clusters of neighbor cells are generated by using “FindClusters” function. T-distributed Stochastic Neighbor Embedding (t-SNE), which is one of the non-linear dimension reduction techniques used for visualization and exploration of data, is performed via “RunTSNE” function (Van Der Maaten & Hinton, 2008). Finally, marker genes were determined by using “FindAllMarkers” function.

After markers for clusters were determined, marker genes that can be used for identification of cell types such as T cells, B cells, Macrophages, Endothelial cells, Cancer Associated Fibroblast (CAF), Cancer, Natural Killer cells (NK), Dormant cells (DRM), and Dendritic cells (DC) were manually retrieved from from databases. For P1 and P4 data, a different marker set was used because sparse scRNAseq data provides a different set of genes for each cell.

The input data was reorganized into the format required by Kumar’s algorithm. P1 and P4 data were preprocessed by using housekeeping gene information and log2-normalization. For cell type classification, firstly, marker gene expressions were fit a Gaussian Mixture Model (GMM) to identify if there are multiple subpopulations in the

data with respect to relevant marker expression, genes with bimodal expression patterns were chosen as markers as expression levels of these genes have the power to discriminate subpopulations of cells from each other, for example a gene can be expressed very highly in one cell type but very lowly in the others, leading to a bimodal distribution across all cells. Cells with high expression of the gene then can be assigned to the corresponding cell type. After selection of such markers, the marker matrix is constructed based on literature information, cell types are given in rows and genes are given in the columns, marker genes are labeled with “AND” if they are expected to be highly expressed in the cell type given in the column or with “NOT” if they should not be expressed in the cell type. Then, training dataset was created by using marker expression information and cells with distinct expression these genes, the parameters obtained from the training set is applied to all other cell, and finally cell types are assigned to all cells.

After cell types were identified using GMM and decision tree algorithm, SingleCellSignalR tool was used to determine the ligand-receptor interactions between cells. For this purpose, cell types identities were assigned as column of count matrix then algorithm was run and interaction scores were generated.

To investigate the metastasis mechanisms driven by ligand-receptor interactions, GSE143423 dataset, a scRNAseq dataset collected from brain metastasis of lung adenocarcinoma samples, was downloaded from GEO. This dataset includes samples from two male and one female patients. 4,610 single cells from LBM1 patient, 1,696 single cells from LBM2 patient and 5,890 single cells from LBM3 patient are included in dataset. For analyzing this dataset Kumar’s method and SingleCellSignalR were used as described above. Marker list was built as specified in the original data article, however markers for dormancy, dendritic cells, and CAF had to be added to the list for this study because these cell types were not interested in article and no markers for them were reported.

To analyze the datasets, Seurat tool was used to determine clusters and marker genes. For metastasis data, marker genes which reported in the original article were used. Kumar’s algorithm was used to identify cell types by using specified markers that represent the cell types of interest. SingleCellSignalR package was used to identify ligand-receptor interactions between cell types.

Enrichment analysis is performed for significant interactions by using g:Profiler and NetworkAnalyst (Raudvere et al., 2019; G. Zhou et al., 2019). For drug repositioning study, the proteins encoded by the genes that are involved in significant interactions were searched in DrugBank (Wishart et al., 2018) and known drugs targeting those genes were identified.

4. RESULTS

4.1. ECM Analysis

Metastasis is considered to have a non-random distribution mechanism known as organotropism, therefore probability of metastasis from a given source tissue is not equal for all target tissues. Cancer cells succeed to metastasize or not, based on the microenvironment of the distant organ (Gao et al., 2019). Hence, RNA and protein levels in tissues may be an indication of outcome of the metastasis mechanisms in each specific source and target tissue. We analyzed ECM genes to investigate whether there are any genes affecting the metastasis of NSCLC cells from lungs to secondary sites. The constructed ECM list contains 3187 genes and 5445 proteins encoded by these genes, together with their expression levels in tissues and pathology status.

After the list was completed as described in Methods Section, metastasis and non-metastasis tissues for lung cancer were determined and genes that are involved in lung and metastasis cases were compared, common and different genes were identified (Table 4.1). Comparison was done either between lung and metastasis tissues, lung and non-metastasis tissues, or lung and non-metastasis tissues including gender specific tissues. “Lymph node” and “tonsil” were excluded from the analysis because all cancer types may metastasize to these two tissues.

Table 4.1: Classification of the tissues included in the ECM list.

Metastasis by lung cancer	No (or rare) metastasis by lung cancer	Gender specific tissues	Excluded tissues
adrenal gland	adipose tissue	breast	lymph node
bone marrow	appendix	cervix, uterine	tonsil
cerebral cortex (brain)	colon	endometrium	

Table 4.1: Continued.

Metastasis by lung cancer	No (or rare) metastasis by lung cancer	Gender specific tissues	Excluded tissues
heart muscle	duodenum	epididymis	
kidney	esophagus	fallopian tube	
liver	gallbladder	ovary	
spleen	pancreas	placenta	
	parathyroid gland	prostate	
	rectum	seminal vesicle	
	salivary gland	testis	
	skeletal muscle		
	skin		
	small intestine		
	smooth muscle		
	stomach		
	thyroid gland		
	urinary bladder		

4.1.1. Protein Expression Analysis

In protein expression analysis, we find that 7 genes are specific for lung, 21 genes are shared between the lung and at least one metastasis tissue, and 161 genes are specific for at least one metastasis tissue. Similarly, 7 genes are shared between the lung and all non-metastasis tissues, 161 genes are specific for all non-metastasis tissues, 432 genes are shared between all non-metastasis and at least one metastasis tissue, and 1334 genes are commonly expressed in all tissues (Figure 4.1).

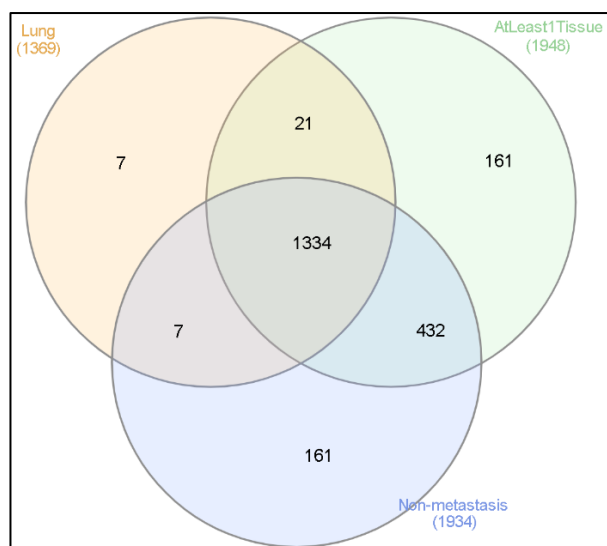


Figure 4.1: Protein expression comparison of lung, metastasis and non-metastasis tissues as a result of protein expression analysis.

“*ACP5, LAMC3, SFTP1, SFTP2, SFTP3, SFTP4 and SFTD*” proteins were specific to lung and not expressed in any other tissue. “*ADAMTS1, AGER, DMBT1, HAVCR2, P4HA3, PI3, and UPK3B*” were common between lung and non-metastasis tissues. 21 genes were common between lung and at least one metastasis tissue, those 21 genes and tissues of expression are listed in (Table 4.2).

Table 4.2: Common genes between lung and at least one metastasis tissue as a result of protein expression analysis.

Gene Name	Tissue1	Tissue2	Tissue3
C1QA	spleen		
C4BPA	liver	bone marrow	
C6	brain	liver	kidney
C8B	liver		
CD209	bone marrow		
CD5L	spleen		
CD84	spleen		
CLEC12A	spleen	bone marrow	
CLEC4A	spleen	bone marrow	
CLEC4D	spleen	bone marrow	
COLEC12	brain		
CTSW	spleen	bone marrow	
FGR	spleen	bone marrow	

Table 4.2: Continued.

Gene Name	Tissue1	Tissue2	Tissue3
GZMM	spleen	bone marrow	
LINGO1	brain		
NAPSA	kidney		
PGLYRP1	spleen	bone marrow	
PLXNA4	brain		
RNASE6	bone marrow		
SIGLEC1	bone marrow		
THSD7A	brain	kidney	

Proteins encoded by genes listed in Table 4.2 may take roles in the metastasis of lung cancer to distant organs as presence of these proteins both in lung and the distant organ may mediate survival of cancer cells in the distant tissue by providing a similar environment for the arriving cells with respect to their source tissue. Focusing on the protein specific to lung and brain, PLXND1, we show that tissue specificity of proteins may indeed mediate organotropism in metastasis. Plexin receptors participate in axon growth guidance, cancer progression, tumor angiogenesis and immune system (Gu & Giraudo, 2013). Roodink and colleagues showed that plexin D1 (PLXND1) is abundantly expressed both in the vasculature and in malignant cells in the majority of clinical tumors and also has important role in brain metastasis of melanoma and neuro-endocrine lung tumor (Roodink, Verrijp, Raats, & Leenders, 2009). Although in our analysis PLXNA4 gene was found only in lung and brain but not in the skin, available data indicates that this gene may take part in mediation of brain metastasis following lung cancer.

Above analysis was repeated between lung, at least one metastasis tissue and non-metastasis tissues including the gender specific tissues for protein expression. 6 genes are specific for lung, 17 genes are shared between the lung and at least one metastasis tissue, 125 genes are specific for at least one metastasis tissue, 8 genes are shared between the lung and non-metastasis tissues, 265 genes are specific for non-metastasis tissues, 468 genes are shared between non-metastasis and at least one metastasis tissue, and 1338 genes are common to all tissues (Figure 4.2).

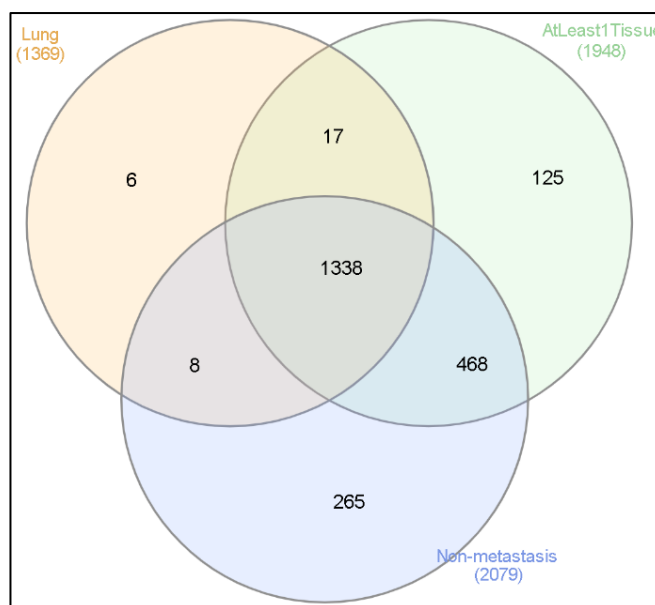


Figure 4.2: Protein expression comparison of lung, metastasis and gender specific included non-metastasis tissues.

“*ACP5, SFTP1, SFTP2, SFTP, SFTPC and SFTD*” genes are expressed only in the lung. “*ADAMTS1, AGER, DMBT1, HAVCR2, LAMC3, P4HA3, PI3, and UPK3B*” genes are common only between lung and non-metastasis tissues. 17 genes are common only between lung and at least one metastasis tissue and the organs where these are expressed are shown in Table 4.3.

Table 4.3: Common genes between lung and metastasis tissue when gender specific tissues are included as a result of protein expression analysis.

Gene Name	Tissue1	Tissue2
C1QA	spleen	
C8B	liver	
CD209	bone marrow	
CD5L	spleen	
CD84	spleen	
CLEC12A	spleen	bone marrow
CLEC4A	spleen	bone marrow
CLEC4D	spleen	bone marrow
CTSW	spleen	bone marrow
FGR	spleen	bone marrow
GZMM	spleen	bone marrow

Table 4.3: Continued.

Gene Name	Tissue1	Tissue2
LINGO1	brain	
NAPSA	kidney	
PGLYRP1	spleen	bone marrow
PLXNA4	brain	
RNASE6	bone marrow	
SIGLEC1	bone marrow	

PLXNA4 is also found in this analysis to be specific to lung and brain. LINGO1 (Leucine Rich Repeat and Immunoglobulin Domain Containing 1) is another protein that is common in lung and brain in both analyses. LINGO1 protein induces downregulation of the activity of EGFR–PI3K–Akt signaling (Z. Zhou, Sathiyamoorthy, & Tan, 2012).

4.1.2. RNA Expression Analysis

The same analysis was done for RNA expression. According to RNA expression comparison of lung, at least one metastasis tissue and non-metastasis tissues, 6 genes are specific for lung, 7 genes are shared between the lung and at least one metastasis tissue, 155 genes are specific for at least one metastasis tissue, 37 genes are shared between the lung and non-metastasis tissues, 139 genes are specific for non-metastasis tissues, 353 genes are shared between non-metastasis and at least one metastasis tissue, and 2136 genes are common to all tissues (Figure 4.3).

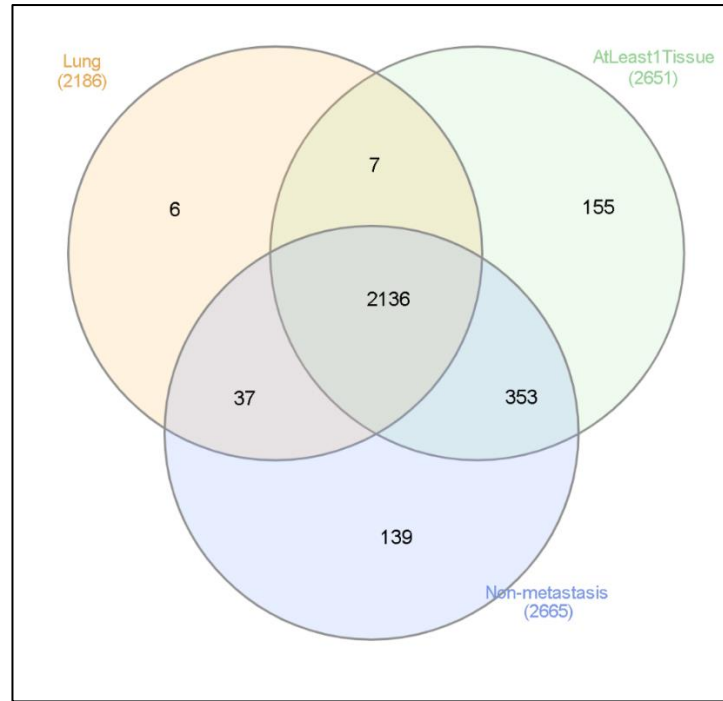


Figure 4.3: RNA expression comparison between lung, at least one metastasis tissue, and non-metastasis tissues.

“*CHIA*, *COL6A5*, *CSF2*, *VWA3A*, *VWA3B* and *WNT3A*” are specific for lung. “*C8B*, *CEACAM8*, *ICAM5*, *LRRN4*, *PCDH15*, *PZP*, and *RSPO4*” genes are shared between lung and at least one metastasis tissue and tissues that express these genes are shown in Table 4.4.

Table 4.4: Common genes between lung and at least one metastasis tissue as a result of RNA expression analysis.

Gene Name	Tissue1	Tissue2	Tissue3
C8B	liver		
CEACAM8	spleen	bone marrow	
ICAM5	brain		
LRRN4	kidney		
PCDH15	adrenal gland	brain	kidney
PZP	liver		
RSPO4	brain		

Also, lung, at least one metastasis tissue and gender specific tissues included non-metastasis tissues were compared and common and different genes were identified for RNA expression. According to results of RNA expression, 1 gene is specific for lung, 1 gene is shared between lung and at least one metastasis tissue, 65 genes are specific for at least one metastasis tissue, 42 genes are shared between lung and non-metastasis tissues, 252 genes specific for non-metastasis tissues, 443 genes are shared between non-metastasis and at least one metastasis tissue, and 2142 genes are expressed in all tissues (Figure 4.4).

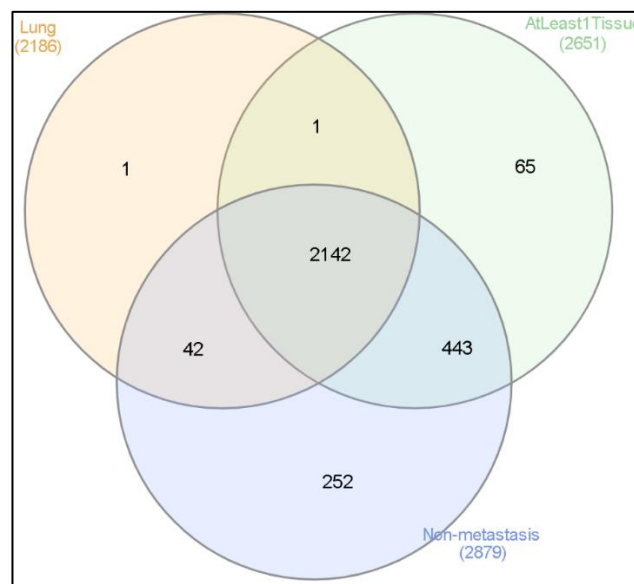


Figure 4.4: RNA expression comparison between lung, at least one metastasis tissue, and gender specific included non-metastasis tissues.

“*CSF2*” gene is specific for lung and “*CEACAM8*” gene is shared between lung and at least one metastasis tissue which are spleen and bone marrow.

Carcinoembryonic antigen-related cell adhesion molecules are cell surface proteins that take part in ECM structure, angiogenesis, apoptosis, tumor suppression, cell invasion and metastasis. Rizeq and colleagues showed that CEACAM6 is expressed in spleen, bone marrow and lung but not expressed in small intestine (Rizeq, Zakaria, & Ouhtit, 2018). In above analysis, CEACAM8 is shared between spleen, bone marrow, and lung so, it could be a mediator in metastasis of lung cancer to these organs.

There is no strong evidence for predicting that genes are related with metastasis to distant organs because, the curated information for ECM in this study are from normal tissues. But this does not exclude the possibility that these genes lead cancer cells to form metastasis in specific organs as cancer studies in literature mentioned above support our results.

4.2. Single Cell Analysis

Tumors are composed of heterogeneous cell types and interactions between the cells in tumors play an important role in cancer progression. In the rest of this study, single cell RNAseq data was analyzed to reveal the interactions between cancer cells, immune cells and dormant cells in NSCLC.

The analysis of the scRNA-seq data towards identification of cell-to-cell interactions requires a multi-step and multi-tool analysis. The first step is finding clusters and choosing marker genes that represent these clusters in the data. It is assumed that data from cells of the same type cluster together and cluster specific marker genes represent each of them. However, different types of cells may also cluster together because they could share common differentially expressed genes. After determination of the marker gene list, canonical marker genes that are reported in the literature to be markers for known cell types are chosen from the list, and combinations of these genes are used as signatures of the cell types to identify the type of the cells that are sampled from the tumor. The last step of analysis is to uncover and score ligand-receptor interactions between cell types according to expression values of both ligand and receptor. For this purpose, Seurat tool was used for determining the marker genes to assign a cell type to each cell in data. Kumar's GMM and decision tree-based algorithm was used for assigning cell types to data by using the marker list generated after Seurat analysis. SingleCellSignalR was used to identify and score the ligand-receptor interactions between cell types. Overview of methods is illustrated in Figure 4.5. Further details of the analysis and results are provided below.

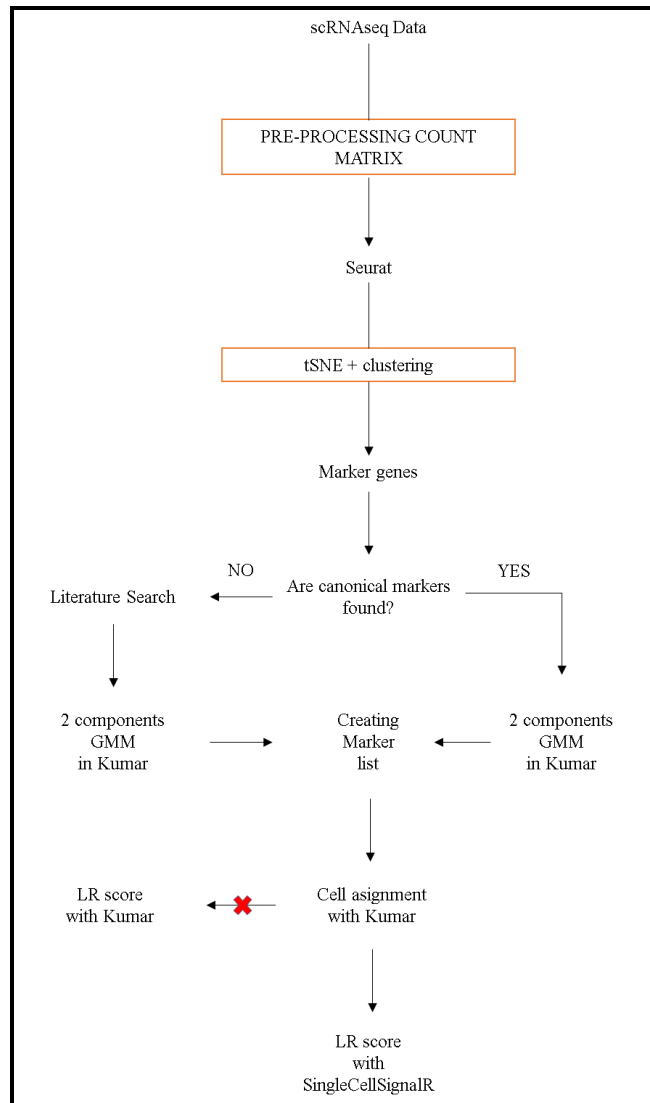


Figure 4.5: Overview of the data analysis pipeline.

4.2.1. Identification of Cell Types

Seurat package was used for pre-analysis of single cell data downloaded from GEO. Data from normal lung and tumor tissues of Patient 1 (P1) were imported to R and Seurat objects were created. Data were normalized and variable genes were found. Data were scaled and principal component analysis was performed to reduce dimension of data. Neighbor cells were found and based on the proximity of the neighboring cells to each other, clusters of cells were generated. t-SNE analysis was performed for visualization by using selected principle components that were determined with PCA (Figure 4.6).

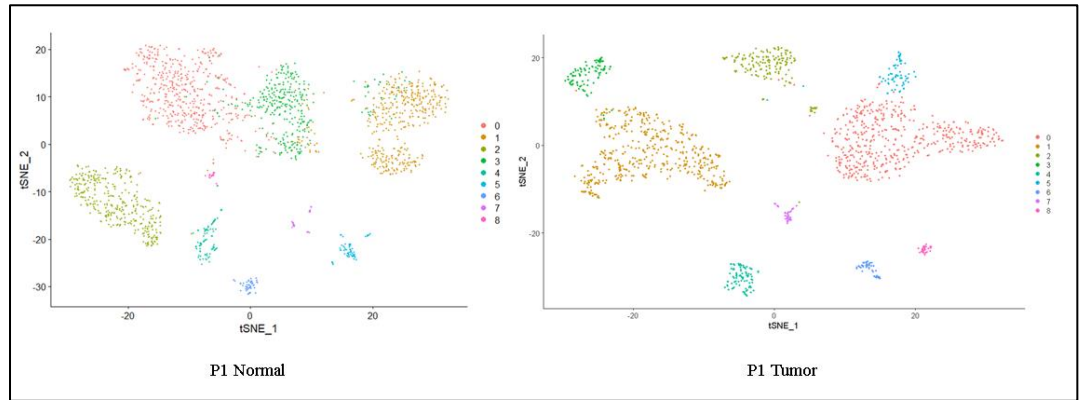


Figure 4.6: t-SNE plot of P1 normal tissue data.

After clusters were generated, marker genes were found by analyzing one cluster versus all others and ordered by only positive (upregulated) markers considering their average fold change value for each cluster for both normal and tumor data. (Table 4.5 and Table 4.6)

Table 4.5: Markers for each cluster in P1 tumor data.

cluster	gene	avg_logFC	p_val	p_val_adj
0	IL7R	1.76	1.13E-151	7.45E-148
0	BATF	1.41	3.46E-55	2.28E-51
1	G0S2	2.80	1.04E-116	6.85E-113
1	CXCL8	2.31	2.17E-77	1.44E-73
2	LY9	1.42	2.77E-79	1.83E-75
2	CD83	1.21	1.47E-41	9.70E-38
3	FABP4	3.77	1.14E-143	7.52E-140
3	C1QB	2.71	1.98E-82	1.31E-78
4	IGKC	6.47	8.04E-57	5.32E-53
4	IGLC2	5.66	4.63E-32	3.06E-28
5	NKG7	2.82	4.76E-111	3.14E-107
5	CCL5	2.54	3.38E-77	2.24E-73
6	TAGLN	3.74	5.16E-109	3.41E-105
6	SPARC	3.42	1.22E-79	8.09E-76
7	SPINK1	4.45	9.41E-36	6.22E-32
7	SFTPC	5.07	2.08E-07	1.37E-03
8	TM4SF1	3.02	4.35E-113	2.87E-109
8	MGP	2.74	1.36E-84	9.00E-81

Table 4.6: Markers for each cluster in P1 normal data.

cluster	gene	avg_logFC	p_val	p_val_adj
0	IL7R	1.94	3.03E-192	2.01E-188
0	DUSP4	1.49	3.80E-98	2.51E-94
1	PTGDS	2.29	6.39E-91	4.23E-87
1	CCL3	2.60	9.43E-78	6.23E-74
2	G0S2	2.99	2.53E-209	1.67E-205
2	IL1B	2.80	1.32E-190	8.70E-187
3	CD8A	1.17	1.80E-91	1.19E-87
3	TRGC2	1.13	3.56E-60	2.36E-56
4	HLA-DQA1	2.36	4.14E-79	2.74E-75
4	HLA-DPB1	2.47	3.85E-60	2.55E-56
5	SFTPA1	4.67	1.08E-90	7.12E-87
5	SFTPC	5.93	6.34E-29	4.19E-25
6	APOC1	3.87	1.05E-109	6.94E-106
6	FABP4	3.84	9.32E-93	6.16E-89
7	CLU	3.17	2.51E-137	1.66E-133
7	CFD	3.99	1.61E-04	1.00E+00
8	IGHM	2.04	3.34E-194	2.21E-190
8	CCR7	2.05	1.60E-40	1.06E-36

In many studies reported in the literature, cell types are assigned to clusters according to marker expression in t-SNE. For example, expression levels of well-known cell type markers were plotted on clusters in t-SNE plot to check whether markers found by clustering represent distinct cell types in the sample (Figure 4.7). But as seen in Figure 4.7, *CD79A* and *CD79B* are expressed in more than one cluster. These markers are B cell markers but expressed in clusters where also *CD2*, *CD3D*, *CD3E* are expressed, which are T cell markers. Thus, using Seurat clusters and their markers for cell type assignment without a thorough investigation of the chosen markers may lead to incorrect cell type assignments. To avoid incorrect assignments, we used more sophisticated algorithms that use a given marker list to assign the cell types correctly.

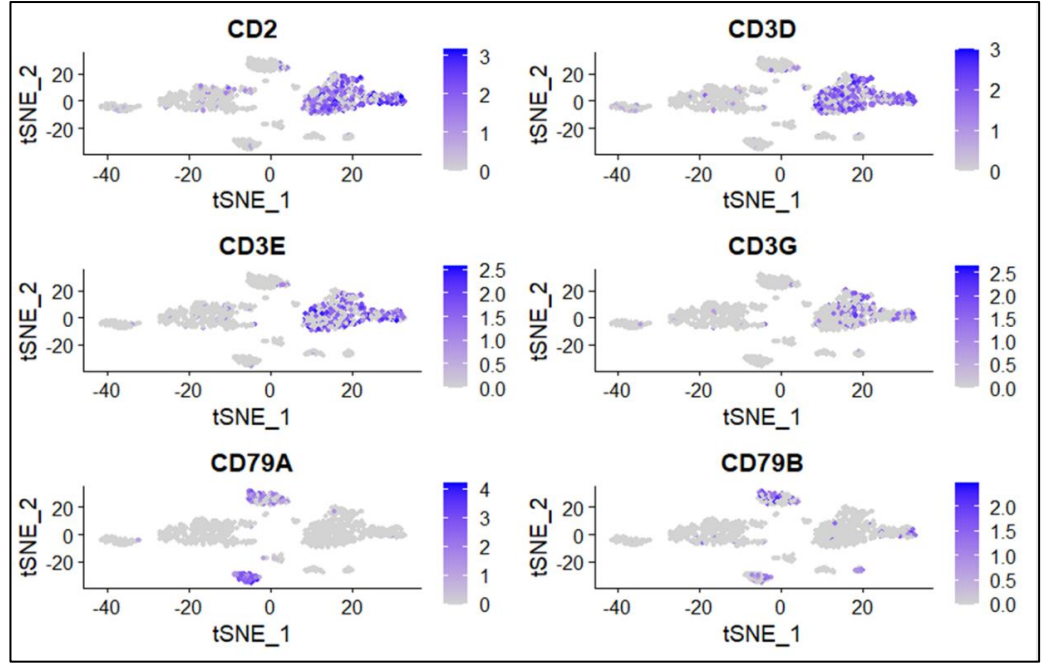


Figure 4.7: Known markers for T cell and B cell in P1 tumor data.

To be able to construct a reliable marker list, we identified the Seurat cluster markers, and among those, we chose the truly cluster-specific markers and checked their associated cell type from literature and databases. This procedure allowed us to produce dataset-specific and literature-supported reliable marker lists (Table 4.7). Although CD83 has low fold change it was used as a marker because it has two components (Figure 4.8). So it can be used as marker in Kumar algorithm.

Table 4.7: Cell type marker list for P1.

	Tcell	Bcell	Mac	Endo	CAF	Cancer	NK	DRM	DC
CD2	AND								
CD3D	AND								
CD3E	AND								
CD79A		AND							
CD79B		AND							
CD163			AND						
CD14			AND						
CSF1R			AND						
PECAM1				AND					
NRP1				AND					

Table 4.7: Continued.

CD58					AND				
FGFR1					AND				
EGFR						AND			
KRAS						AND			
TP53						NOT			
NAPSA						AND			
KLRD1							AND		
CD226							AND		
MAPK14						NOT		AND	
CDKN1A								AND	
CDKN1B								AND	
ITGAX									AND
CD83									AND

Once marker lists were created, data were analyzed in Matlab by using Kumar's algorithm. Markers were cross-checked also by using Gaussian mixture model whether subpopulations exist or not in the data with respect to the marker, to verify the discriminative power of their expression levels of the chosen markers (Figure 4.8). If there are two peaks in left panel of Figure 4.8, this means that there are two subpopulations which are colored red (mean =0) and blue (mean~2) in the data with respect to the marker of interest and right panel shows the number of mixture components that are found by the GMM based on the expression levels of the marker at the minimum Bayesian information criterion value which is used for model selection. For all chosen markers, it is cross checked that there are strictly two components in the data, so these markers could be reliably used to identify cell types. The results of this analysis will be given in the next section.

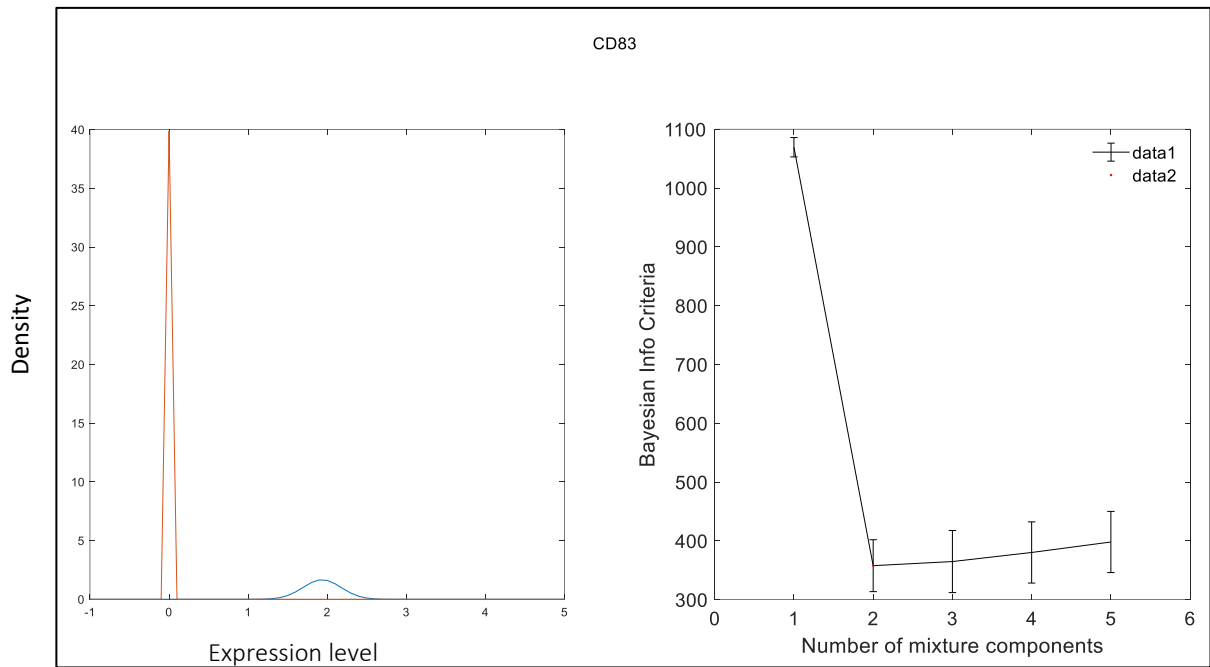


Figure 4.8: GMM of CD83 marker.

The same analysis was applied for P4 normal and tumor data. Firstly, data were analyzed in R with Seurat tool to obtain clusters (Figure 4.9) and marker genes (Table 4.8 and Table 4.9).

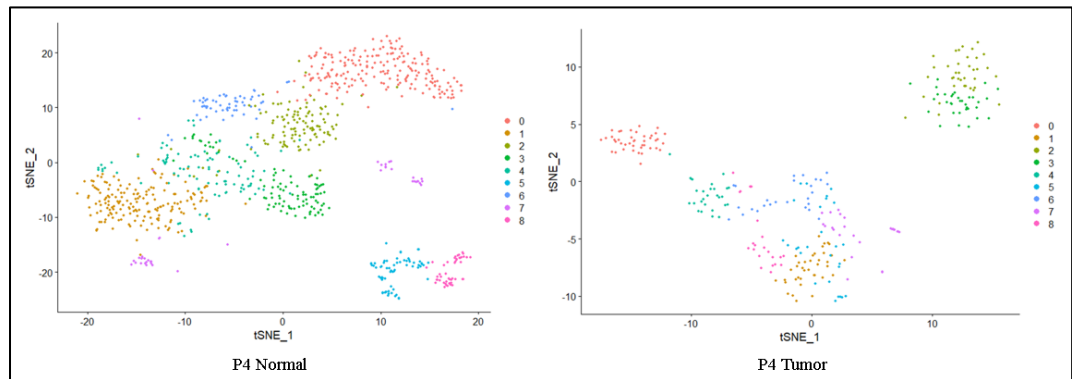


Figure 4.9: t-SNE plot of P4 data.

In Figure 4.9, it can be seen that the clusters were not separated well, and unsurprisingly, computationally it is difficult to distinguish which cluster represents which cell types in this dataset. Particularly, cells in P4 normal data are not separated well, unlike

the clearer separation in the tumor data. In this case, alternative methods were explored to identify cell types because known marker genes were not expressed in only one cluster (Figure 4.10). The cluster markers reported by Seurat are given in Table 4.8 and 4.9, most of these markers cannot be verified by the marker lists in the literature, therefore the cluster markers were supplemented by additional markers chosen from the literature, finally markers shown in Table 4.10 were used for further analysis. Once the marker genes were determined, tumor and normal data were analyzed in Matlab by using Kumar's algorithm.

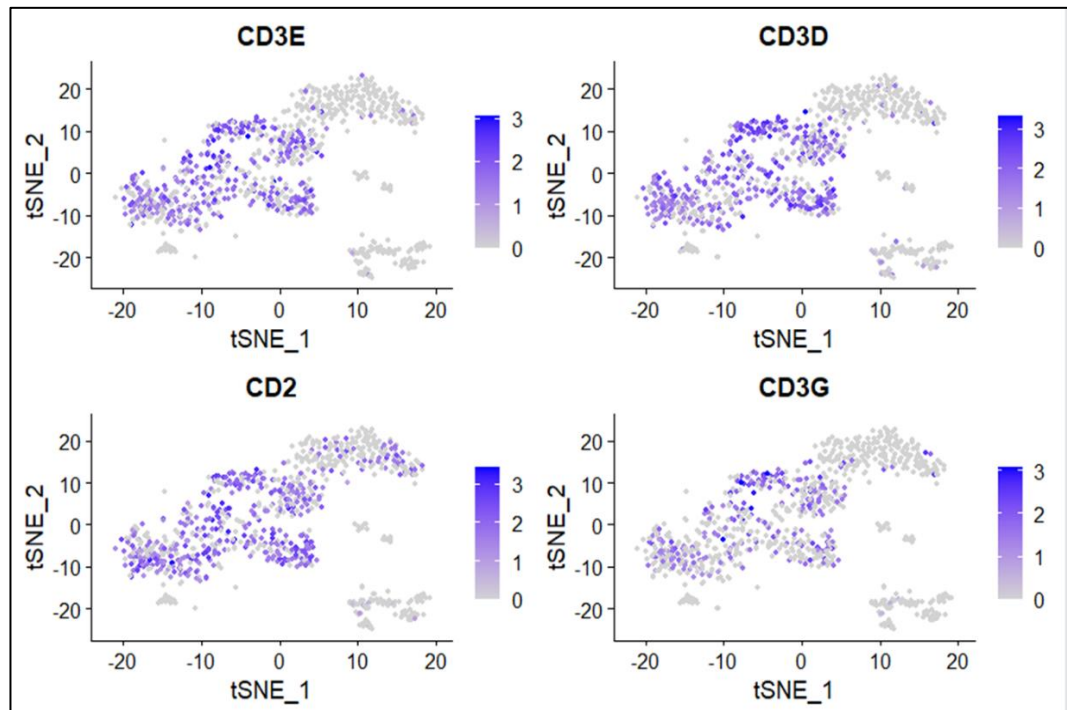


Figure 4.10: Known markers for T cell in P4 normal data.

Table 4.8: Markers for each clusters in P4 tumor data.

cluster	gene	avg_logFC	p_val	p_val_adj
0	RP11-1143G9.4	2.72	3.37E-29	2.53E-25
0	LYZ	2.99	1.07E-24	8.00E-21
1	CD83	1.19	2.41E-13	1.81E-09
1	GPR183	1.22	9.39E-13	7.05E-09
2	CEACAM6	2.16	7.47E-26	5.60E-22

Table 4.8: Continued

cluster	gene	avg_logFC	p_val	p_val_adj
2	SFTPB	2.32	4.87E-15	3.66E-11
3	WFDC2	2.29	4.11E-23	3.09E-19
3	SPP1	2.37	3.89E-11	2.92E-07
4	APOC1	2.97	4.60E-18	3.45E-14
4	C1QB	1.91	1.74E-10	1.30E-06
5	FCER1A	1.32	4.38E-12	3.28E-08
5	IL1B	1.25	4.11E-08	3.08E-04
6	HLA-DRB1	1.15	1.35E-10	1.01E-06
6	RGS1	1.21	1.97E-08	1.48E-04
7	LTB	1.33	3.18E-07	2.38E-03
7	BIRC3	1.46	4.90E-07	3.67E-03
8	EREG	1.57	7.61E-11	5.71E-07
8	CXCL3	2.30	1.50E-05	1.13E-01

Table 4.9: Markers for each clusters in P4 normal data.

cluster	gene	avg_logFC	p_val	p_val_adj
0	KLRF1	1.91	1.72E-109	1.29E-105
0	SPON2	1.80	4.26E-85	3.19E-81
1	LTB	1.17	5.42E-56	4.07E-52
1	BIRC3	1.25	1.01E-44	7.55E-41
2	TRGC2	0.82	1.42E-21	1.07E-17
2	HSPA6	0.86	1.08E-14	8.12E-11
3	RBPJ	1.38	2.60E-32	1.95E-28
3	DUSP4	1.38	5.02E-25	3.77E-21
4	IL7R	1.22	6.65E-28	4.99E-24
4	RGS1	0.97	4.53E-04	1.00E+00
5	LYZ	3.00	1.92E-72	1.44E-68
5	HLA-DRA	3.09	1.76E-59	1.32E-55
6	TRGC2	1.46	4.32E-25	3.24E-21
6	GZMH	1.34	2.03E-22	1.52E-18
7	SCGB3A1	4.29	1.78E-06	1.33E-02
7	SCGB1A1	4.57	2.04E-04	1.00E+00
8	S100A9	3.59	1.11E-81	8.34E-78
8	S100A8	3.42	2.20E-66	1.65E-62

Table 4.10: Cell type marker list for P4.

	Tcell	Bcell	Mac	Endo	CAF	Cancer	NK	DRM	DC
CD2	AND								
CD3D	AND								
CD3E	AND								
CD37		AND							
TNFSF13B		AND							
ITGAM			AND						
CD14			AND						
CSF1R			AND						
PECAM1				AND					
NRP1				AND					
CD58					AND				
FGFR1					AND				
TP53						NOT			
EGFR						AND			
KRAS						AND			
NAPSA						AND			
KLRB1							NOT		
KLRD1							AND		
MAPK14						NOT		AND	
CDKN1A								AND	
CDKN1B								AND	
ITGAX									AND
CD83									AND

SingleCellSignalR (Cabello-Aguilar et al., 2020) is another tool to analyze single cell RNAseq data and uncover ligand-receptor interactions between cell types, the tool is capable of doing t-SNE plots and clustering similar to Seurat, but also capable of doing cell type assignment and ligand-receptor analysis similar to Kumar's algorithm. P4 tumor and normal data are analyzed in R by using SingleCellSignalR tool in addition to Seurat and Kumar's algorithm. Using SingleCellSignalR tool, clusters in both tumor and normal data were distinguished better in t-SNE when compared to Seurat results (Figure 4.11). Both tools run on R, however it is likely that they use different implementations of t-SNE algorithm and generate different results, in addition t-SNE uses stochastic parameters in its computations, which may lead to slightly different results in each run. Further, the two tools use different clustering algorithms, therefore it is expected to find different t-SNE and clustering results by Seurat and SingleCellSignalR. Nevertheless, t-SNE and

clustering results of these tools were only used for visualizing the data and aiding the marker selection, but not used in cell type assignments.

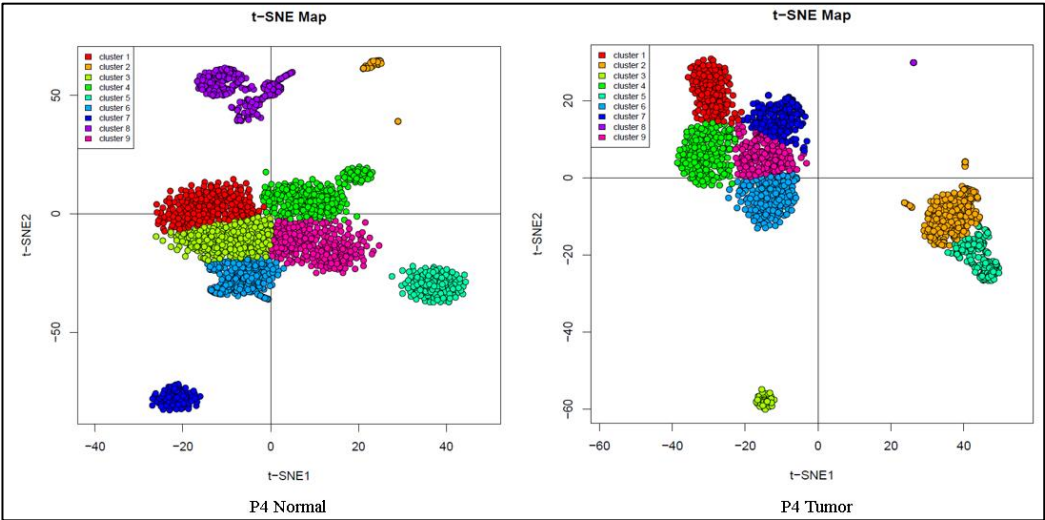


Figure 4.11: SingleCellSignalR t-SNE results of P4 data.

Cell type assignment algorithm was run by using marker genes given in tool by adding cancer and dormancy markers and cell types were assigned (Table 4.11). For P4 tumor data, 2 T cells, 42 B cells, 184 Macrophages, 4 Endothelial cells, 4 CAFs, 33 Dendritic cells, 1 NK cell, 45 Dormant cells, 856 Cancer cells were assigned and 252 cells were unassigned. For P4 normal data, 568 T cells, 65 B cells, 195 Macrophages, 3 Endothelial cells, 3 CAFs, 9 Dendritic cells, 43 NK cells, 351 Dormant cells, 31 Cancer cells were assigned and 759 cells were unassigned.

Table 4.11: Cell type numbers in P4 data SingleCellSignalR analysis.

Data	T cell	B cell	Mac	Endo	CAF	Cancer	NK	DRM	DC	Unassigned
P4 Tumor	2	42	184	4	4	856	1	45	33	252
P4 Normal	568	65	195	3	3	31	43	351	9	759

For SingleCellSignalR results, clustering results were better than Seurat but when cell types were assigned, unexpectedly a number of cancer cells in normal tissue data were detected and therefore cell type assignments of the tool were ignored in further analysis as mentioned above. Kumar’s algorithm was used for cell type assignment and SingleCellSignalR was used only to uncover ligand-receptor interactions.

The cell type identification analysis by Kumar algorithm can be summarized as follows: For P1 normal tissue data, 905 T cells, 9 B cells, 6 Macrophages, 28 Endothelial cells, 19 CAFs, 86 Cancer cells, 416 NK cells, 36 Dormant cells, 797 Dendritic cells were identified and 193 cells unassigned. For P1 tumor tissue data, 586 T cells, 174 B cells, 51 Macrophages, 121 Endothelial cells, 2 CAFs, 86 Cancer cells, 4 NK cells, 4 Dormant cells, 558 Dendritic cells were assigned, and 246 cells were unassigned.

For P4 tumor tissue data, 51 T cells, 262 B cells, 1 Macrophage, 5 Endothelial cells, 3 CAFs, 603 Cancer cells, 196 Dormant cells, 57 Dendritic cells were identified and 245 cells were unassigned and no NK cells were found. For P4 normal tissue data, 631 T cells, 175 B cells, 4 Endothelial cells, 10 CAFs, 861 NK cells, 6 Dormant cells, 154 Dendritic cells were identified, and 186 cells were unassigned and no Macrophage or cancer cells were found.

In addition, adenocarcinoma brain metastasis data was analyzed by using marker genes reported in the original data article and cell types were assigned with Kumar's algorithm. 1 T cell, 1278 B cells, 4274 Endothelial cells, 140 CAFs, 3278 Cancer cells, 348 NK cells, 14 Dormant cells, 502 Dendritic cells were identified, 2361 cells were unassigned and no Macrophages were found.

All cell types and number of assigned cells are given in Table 4.12. P1 data from the normal tissue was excluded from further analysis because cancer cells were identified in normal tissue data, which may be an indication of sample contamination or another technical issue with the dataset.

Table 4.12: Cell types and number of assigned cells for all datasets.

Data	T cell	B cell	Mac	Endo	CAF	Cancer	NK	DRM	DC	Unassigned
P1_tumor	905	9	6	28	19	36	416	36	797	193
P1_normal	586	174	51	121	2	86	4	4	558	246
P4_tumor	51	262	1	5	3	603	0	196	57	245
P4_normal	631	175	0	4	10	0	861	6	154	186
Metastasis	1	1278	0	4274	140	3278	348	14	502	2361

4.2.2. Ligand-Receptor Interactions

After evaluation of the results by using different tools, it is revealed that each tool has its own strengths and weaknesses. Based on the evaluation of the results, a hybrid pipeline was used for all analysis (Figure 4.5): Seurat tool was used for finding marker genes and these marker genes were filtered and verified by using literature information, Kumar's algorithm was used for assignment of cell types in P4 analysis as no cancer cells in normal tissue data was found, hence results from this algorithm are more reliable in terms of cell type assignment. And finally, SingleCellSignalR is used to score ligand-receptor interactions as the regularized scoring of this tool produces fewer false positives.

P4 and metastasis data were analyzed by using the hybrid pipeline described above. Cell types were identified by using marker genes and ligand-receptor interactions were scored between cells and generated chord plot by algorithm (Figure 4.12). Chord plot summarizes the number of interactions between cell types via a color code where darker color indicates higher number of interactions.

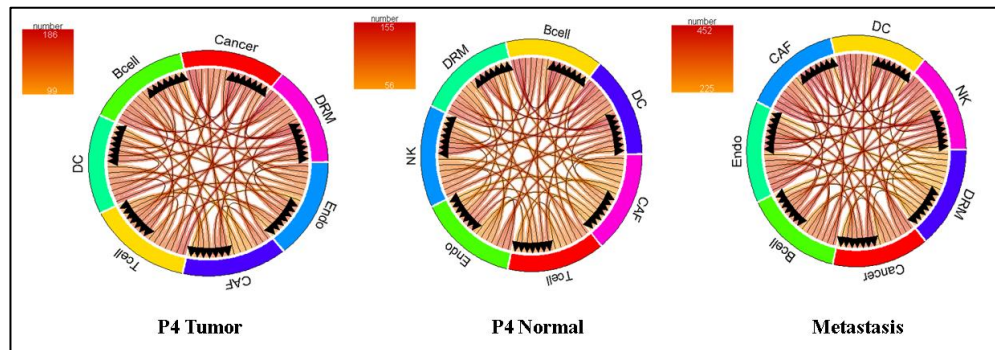


Figure 4.12: Summarized view of interactions between cell types and interaction numbers.

In P4 tumor tissue data, most variable interactions between cell types were shown in Figure 4.13, the interactions were clustered by two dimensional hierarchical clustering, hence pair of cell types with similar interactions appear together on the horizontal axis and ligand-receptor pairs that interact in similar cell type pairs appear together on the vertical axis. Interactions that take place between all cell types at similar strength were

filtered out by choosing the most variable interactions. Therefore, Figure 4.13 focuses only on interactions that are relatively strong between specific cell-types, but not in others.

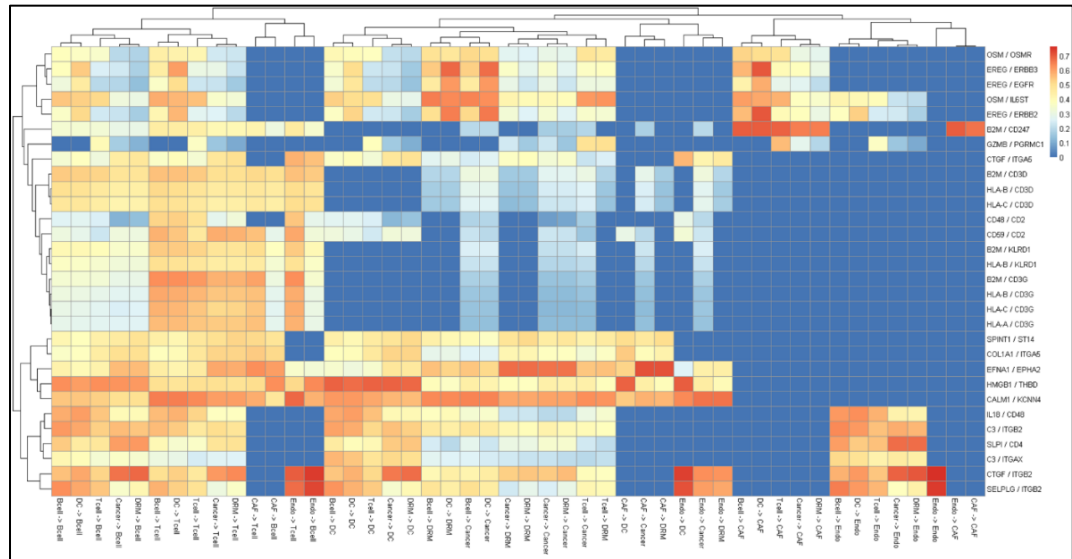


Figure 4.13. Variable interactions between cell types in tumor data.

To identify the differences in interactions of cancer cells and dormancy cells with other cell types in P4 tumor samples, interactions between B cell, T cell, or CAFs and Cancer or Dormant cells were compared (Figure 4.14).

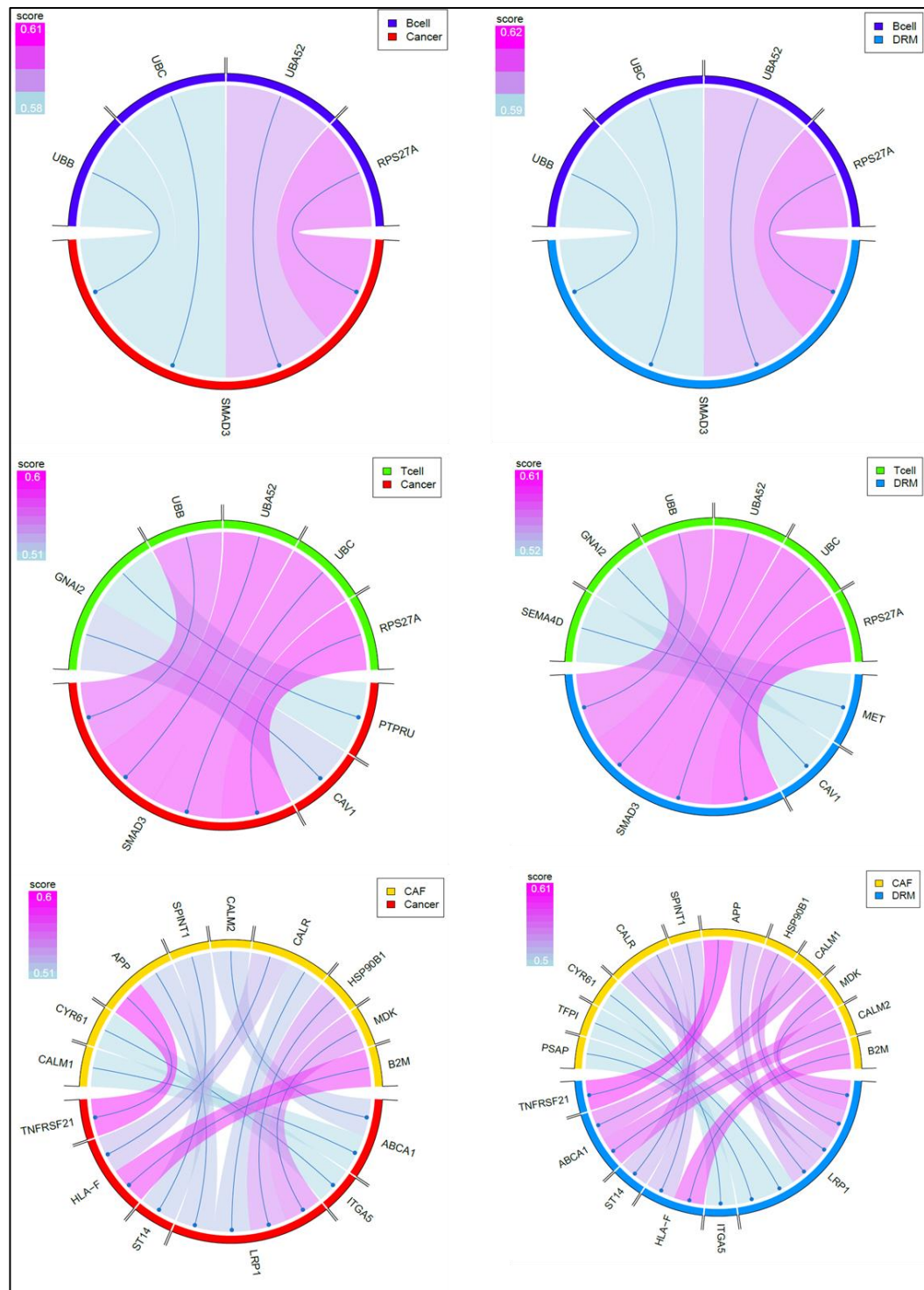


Figure 4.14. Ligand receptor interactions between tumor B, T cells and CAFs and Cancer and Dormant cells.

As shown in Figure 4.14, B cell interactions of cancer and dormant cells are the same but have slightly different scores. On the other hand, when T cell interactions are compared, there is one different interaction in dormant cells which is *SEMA4D_MET*. Soong and Scott showed that SEMA4D suppresses activation of the MET hepatocyte growth factor receptor (Soong & Scott, 2013), therefore this interaction targets activity of MET, which is promotion of tumor growth, angiogenesis and metastasis. The interaction found in our analysis indicates that SEMA4D secreted by T cells may inhibit MET activity in dormant cells and mediate initiation or maintenance of dormancy, and cancer cells escape this inhibition as they do not express MET and maintain their growth via alternative pathways. On the other hand, both genes are annotated by the GO term ‘negative regulation of cell death’, therefore it is possible that dormant cells regulate their activity and proliferation with this interaction and escape programmed cell death to survive (Figure 4.15). Other differential interactions between cancer or dormant cells and CAFs are, *PSAP_LRP1* and *TFPI_LRP1*. It was shown that PSAP inhibits metastasis by stimulating the expression of TSP1 which induces angiogenic dormancy (Kang et al., 2009) and also it was shown that TFPI-2 suppress tumor cancer cell proliferation and invasion in breast cancer (G. Wang et al., 2018). Hence, it can be concluded that dormant cells express various receptors to receive growth inhibition signals from CAFs as well as T cells.

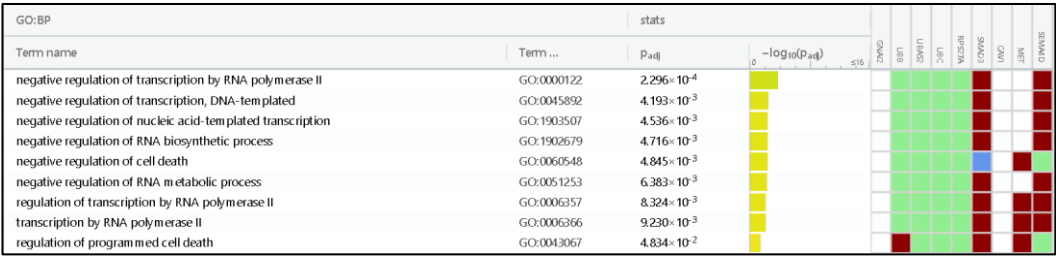


Figure 4.15. GO biological process terms enriched by MET and SEMA4D genes.

In P4 normal tissue data, most variable interactions between cell types were shown in Figure 4.16. The most significant interaction between CAF and other cells is HSP90AA1-EGFR. HSP90AA1 encodes the heat shock protein 90-alpha (HSP90A) acts as a molecular chaperone to modify the structures and interactions of other proteins.

C5AR1 decreases tumor growth and delays the development of tumors in mammary tissues (Markiewski et al., 2017).

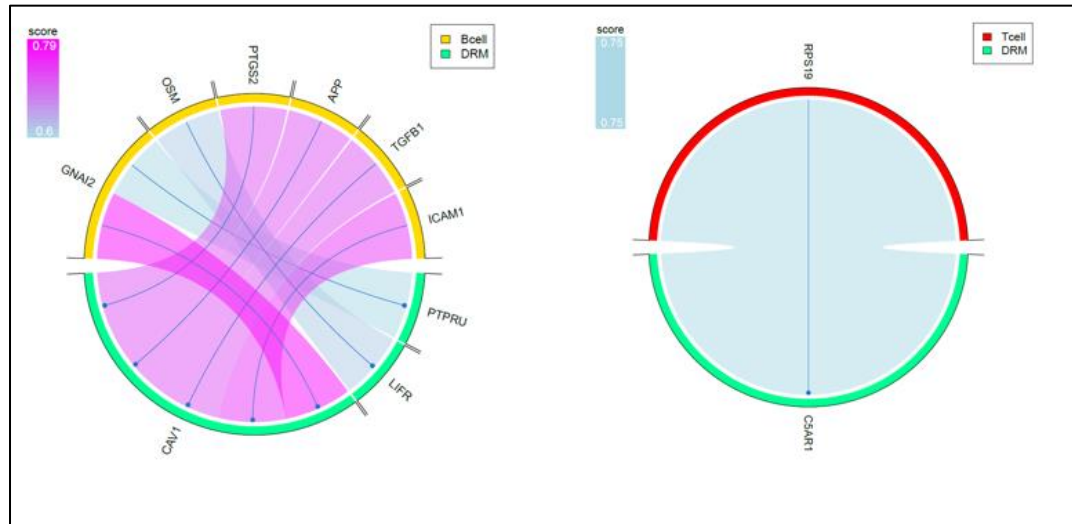


Figure 4.17. Ligand receptor interactions in P4 normal tissue data between B cell - dormant cell and T cell -dormant cell

Most variable interactions between cell types in metastasis data were shown in Figure 4.18. RTN4-LINGO1 interaction, placed in the first row, is found to be active in both cancer and dormant cells, dormant cells have a stronger interaction as a result of higher expression of LINGO1 receptor. RTN4 (Reticulon 4) encodes neurite outgrowth inhibitor (NOGO), which has important roles in cancer such as regulating cell motility and growth (Schwab, 2010). LINGO1 (Leucine Rich Repeat and Immunoglobulin Domain Containing 1) is one of the NOGO receptor complexes and inhibits activity of PI3K/Akt signaling pathways (Fernandez-Enright, Andrews, Newell, Pantelis, & Huang, 2014). In ECM analysis of protein expression given in the Section 4.1.1, LINGO1 was found to be exclusive to lung and brain. This interaction may induce cells to enter dormancy either in lung or brain, increasing the risk of a future metastasis in either of the organs.

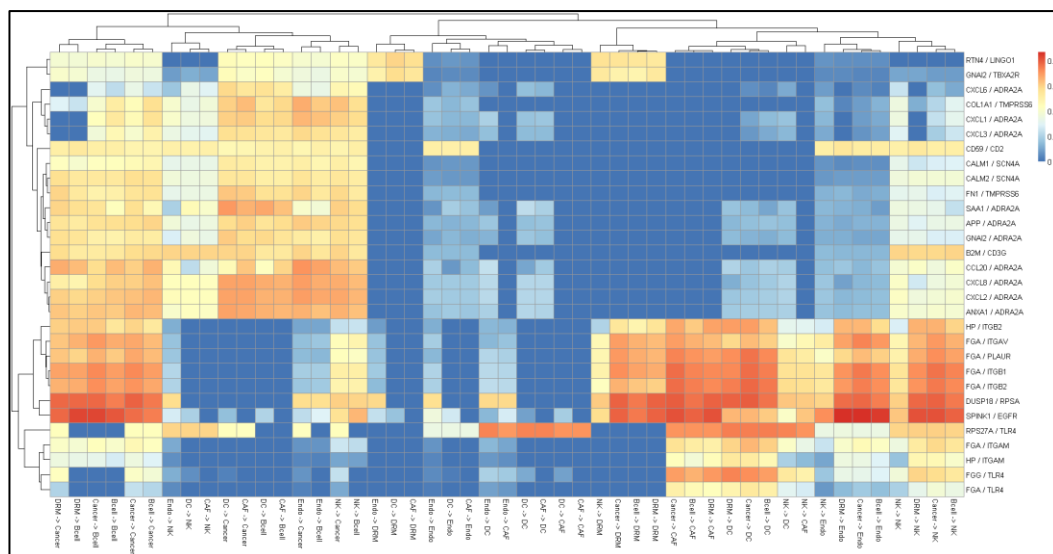


Figure 4.18. Most variable interactions between cell types in metastasis data.

CAFs are important in regulation of ECM in tumor tissues. Secreted factors in ECM and their receptors take part in cancer development, metastasis, and dormancy, placing CAFs in the center of cancer research. To understand the mechanisms of CAF regulation in cancer, interactions of CAFs and cancer cells in NSCLC (P4 tumor tissue) and metastasis were compared (Figure 4.19), and found that 41 interactions are specific for NSCLC, 266 interactions are specific for metastasis and 109 interactions are common.

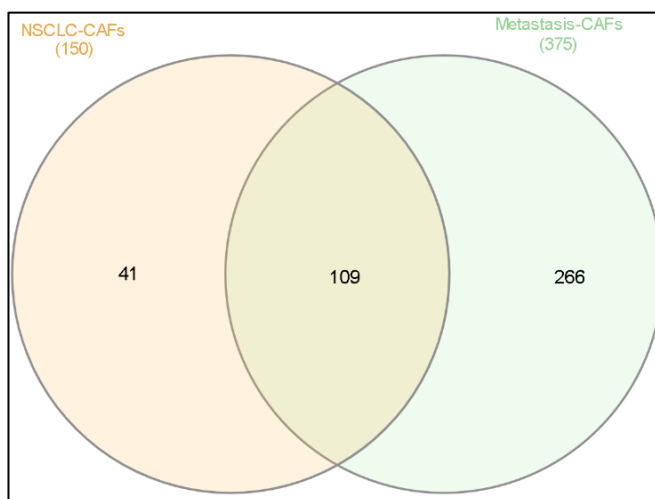


Figure 4.19: Comparison of cancer and CAFs interactions.

CAF-cancer cell interactions exclusive to NSCLC are mostly related to cell adhesion as shown in Figure 4.20 and blue colored genes have role in this cell adhesion. CAF-cancer cell interactions exclusive to metastasis are mostly related to cell migration as shown in Figure 4.21 and red colored genes have role in this cell migration. NetworkAnalyst tool was used to visualize the ligand-receptor interactions specific for both NSCLC and metastasis given in Figure 4.19, to find the enriched GO terms and KEGG pathways in the set of genes that encode the interacting proteins and to color the genes which are members of a selected GO term or KEGG Pathway as shown in the Figure 4.20 and Figure 4.21.

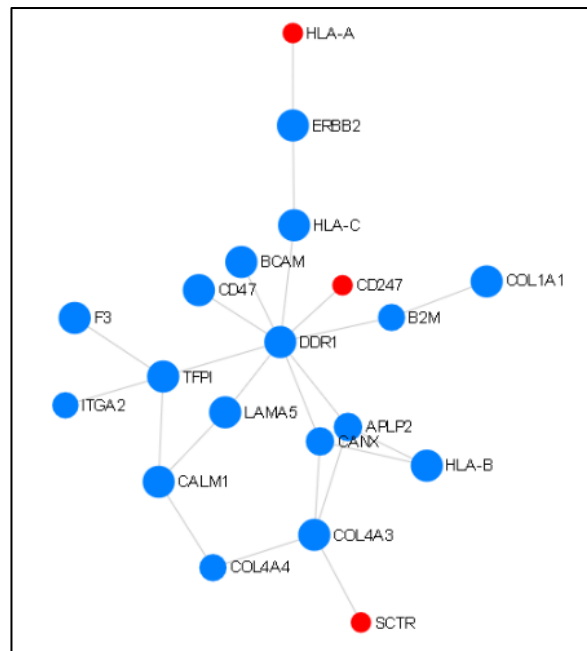


Figure 4.20: Genes that take part in CAFs and NSCLC cancer cell interactions (but not in metastasis)

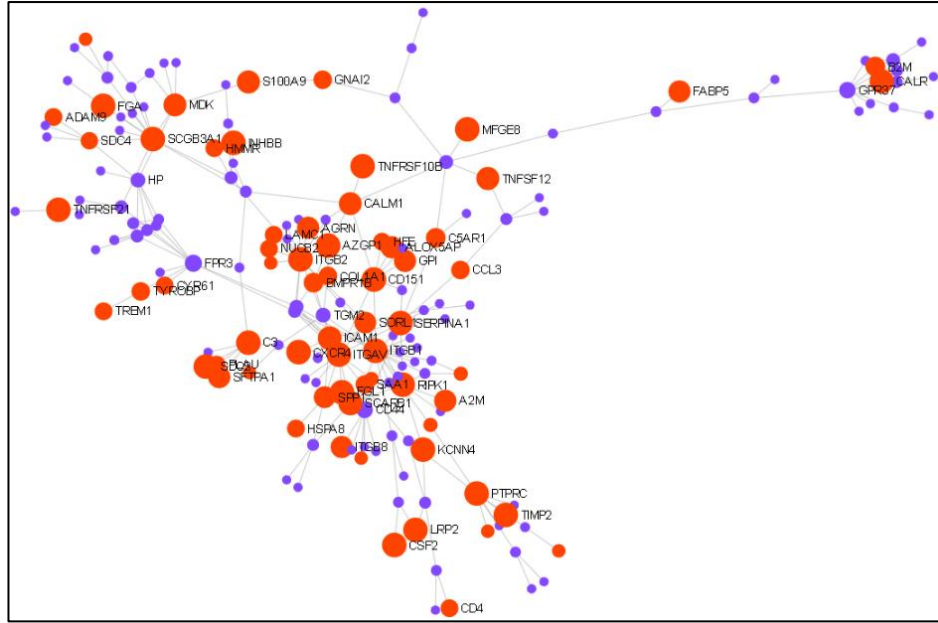


Figure 4.21: Genes that take part in CAFs and metastatic cancer cell interactions (but not in primary NSCLC tumor)

The results above demonstrated that nature of the interactions of CAF cells with cancer cells are different from each other in primary NSCLC tumor and metastatic brain tumor. CAFs lead to signals which promote cell adhesion in primary tumor whereas they lead to signals which promote cell migration in the metastatic tissue. Therefore, more aggressive nature of metastatic tumors maybe triggered by such signals received and secreted by cancer cells in the metastatic tissues.

Finally, interactions between all cell type pairs in the tumor tissues were investigated to reveal the general ECM profile and interaction network in the tumor environment. Tumor samples of P4 data were analyzed for all cell-to-cell interactions and common interactions between cancer cells and all the other cells were found.

Most of high scored significant interactions between all cell types common between cancer cells and all the other cells included *B2M_HLA-F*, *HLA-A_APLP2*, *TIMP1_CD63*, *APP_CD74*, *HLA-B_CANX* (Figure 4.22). HLA-A, HLA-B and HLA-F are human leukocyte antigens and responsible for regulation of immune system. It is expected to find these interactions between these cells because immune cells interact with cancer cells to eliminate them and these results demonstrate that signaling towards immune response in cancer tissue is highly active.

Based on pathway enrichment analysis of the interactions between all cell types in tumor tissues, results are mostly related with ECM-receptor interaction, pathways in cancer such as Non-small cell lung cancer, PI3K-Akt, MAPK signaling pathway, EGFR, antigen processing and antigen presentation, which demonstrate that strong interactions between the cells in tumors are related to proliferation pathways such as PI3K-Akt and MAPK as well as immune response pathways.

g:Profiler was used as an alternative tool to visualize the enrichment analysis of common interactions in tumor tissue (Figure 4.23). Most enriched terms are ligand-receptor interactions such as integrin binding, protein-containing complex binding, cell surface (because surface proteins are often receptors), ECM-receptor interactions, PI3K-Akt signaling pathway (which is important for cancer development), immune system (because of interactions between immune cells and cancer cells) and abnormality of the immune system.

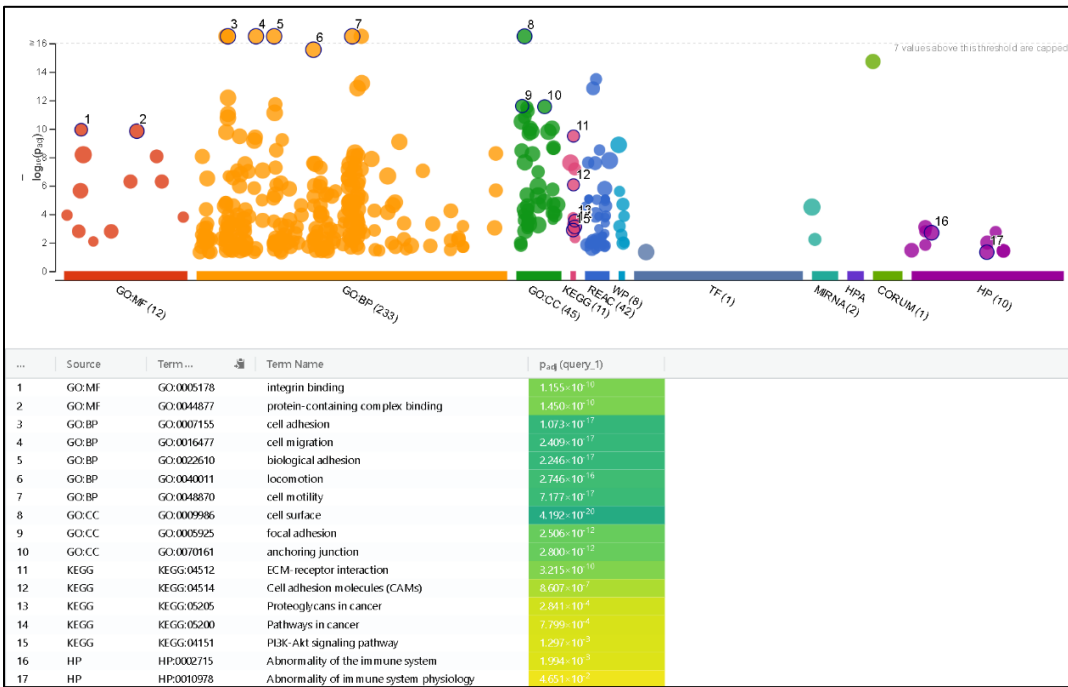


Figure 4.23: g:Profiler enrichment result of genes taking part in significant interactions in tumor tissue.

In this section, a general view of the cell-to-cell interaction in normal, tumor and metastatic tissues was given. Relevance of the results for better understanding the disease mechanisms was demonstrated by focusing on few exemplary interactions and also by providing a functional analysis of all interactions in the tumor microenvironment. The analyses exemplified for few cases in this section were carried out for all datasets and all combinations of cell types. Complete lists and images of these all-by-all analyses are given in Supplementary Files and a list of Supplementary Files are provided in the Appendix section.

4.2.3. Drug Repositioning

Cancer is caused by mutations and via the mutations, cancer cells gain some capabilities such as sustaining their own proliferation and avoiding immune attack. To understand the self-signaling between the cancer cells, cancer autocrine interactions were analyzed for both P1 tumor and P4 tumor data. 177 interactions were identified for P4 tumor data, 90 interactions were found for P1 tumor data, and 69 interactions were shared between them (Figure 4.24).

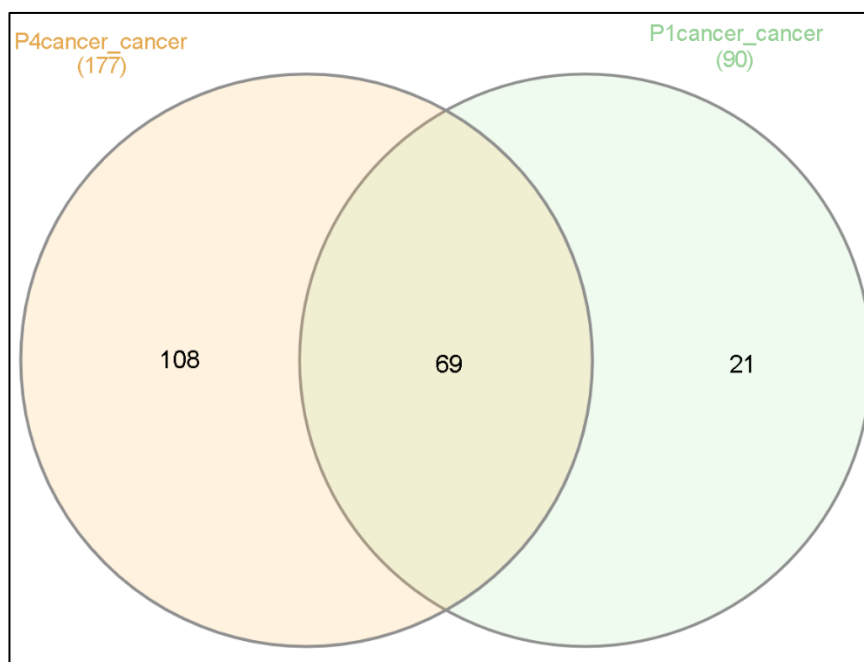


Figure 4.24: Autocrine interactions of P4 and P1 tumor data.

In Figure 4.24, there are 108 interactions found only in P4 tumor data, 21 interactions in P1 tumor data and 69 interactions are common to both. To uncover approved drugs that can be used for NSCLC treatment, drug-gene list, constructed by Dilara Uzuner was analyzed and used to score the drugs by the number of their targets (Uzuner, 2020). Patient specific and common drugs to target the interactions between cancer cells are listed in Table 4.13.

Table 4.13: Drug targets for NSCLC according to ligand-receptor interactions found in analysis.

For both patients		
drug	gene	score
COPPER	RPSA, APOE, CALR, HSPA8, PKM, B2M, APP, HSP90B1, C3	9
ARTENIMOL	PKM, VIM, HSPA8, LGALS1, RPS19	5
CALCIUM CITRATE	CALR, CALM3, CALM1, CANX, CALM2	5
CALCIUM PHOSPHATE	CALM1, CANX, CALR, CALM3, CALM2	5
CALCIUM PHOSPHATE DIHYDRATE	CANX, CALM2, CALR, CALM1, CALM3	5
ZINC	APOE, APP, C3, APLP2	4
ZINC ACETATE	APP, C3, APOE, APLP2	4
ZINC CHLORIDE	APLP2, APP, APOE, C3	4
COCCIDIOIDES IMMITIS SPHERULE	HLA-A, HLA-B, HLA-A	3
FORESKIN KERATINOCYTE (NEONATAL)	VEGFA, TGFB1, TGFB2	3
For P4		
LANOTEPLASE	CALR, LAMC1, LRP1, LAMA5, FN1, PLAUR	6
COPPER	CALR, HSP90B1, RPSA, HSP90AA1, APP	5
FOSTAMATINIB	ACVR1, EPHA2, ERBB2, EGFR, DDR1	5
CALCIUM CITRATE	CALR, CALM3, CALM1, CALM2	4
CALCIUM PHOSPHATE	CALM1, CALR, CALM2, CALM3	4
CALCIUM PHOSPHATE DIHYDRATE	CALM1, CALM3, CALR, CALM2	4
TENECTEPLASE	CALR, LRP1, PLAUR	3
PALIFERMIN	HSPG2, NRP1	2
TRASTUZUMAB	ERBB2, EGFR	2
BRIGATINIB	EGFR, ERBB2	2
For P1		
MYRISTIC ACID	CALM1, LY96, TLR4	3
(2S)-2-(3-BROMOPHENYL)-3-(5-CHLORO-2-HYDROXYPHENYL)-1,3-THIAZOLIDIN-4-ONE	ARPC5, ACTR2	2
3-HYDROXY-MYRISTIC ACID	TLR4, LY96	2
ADAPALENE	PTGS2, RXRA	2
COPPER	B2M, C3	2
ETODOLAC	RXRA, PTGS2	2
FORESKIN FIBROBLAST (NEONATAL)	TGFB2, TGFB1	2
FORESKIN KERATINOCYTE (NEONATAL)	TGFB2, TGFB1	2
GOLOTIMOD	TLR4, TLR2	2
LAURIC ACID	TLR4, LY96	2

Some drugs can target more than one gene as shown in Table 4.13. As a result of drug repositioning analysis, some drugs can be used in both patients, such as Copper. Also some unique drugs can be suggested for each patient such as Trastuzumab for P4 and Lauric Acid for P1. To have a more effective treatment, these drugs can be administered as a combination.

5. DISCUSSION

Lung cancer is one of the most common cancer types worldwide. Non-small cell lung cancer is responsible for ~85 % of the lung cancer cases and adenocarcinoma is responsible for ~40 % of the non-small cell lung cancer cases. Tumor, metastasis and normal tissues from adenocarcinoma cases were investigated in this study to better understand the disease mechanisms. Tumors are heterogeneous and dynamic tissues, the heterogeneity stems from both accumulating mutations and also from variety of cell types that co-exist in the tumor. Epithelial cells, immune cells, cancer cells and some rare cell types such as dormant cells both inhabit and transform the microenvironment. For this reason, it is important to identify the cell types in tumors and determine ligand-receptor interactions between these cell types to better understand the mechanisms of cancer progression and develop new strategies for treatment. While surgery aims removal of tumor tissue and radiotherapy aims to destroy cancer cells by damaging their DNA, chemotherapy and immunotherapy changes the intracellular and extracellular mechanisms to deal with cancer. Immunotherapy is a type of cancer treatment that makes the body's natural defenses stronger to fight cancer. It uses substances produced by the body or in the lab to improve or restore immune system function. Immunotherapy includes treatments that work in different ways. Some increase the body's immune system in a very general way. Others help to train the immune system specifically to attack cancer cells. It helps the immune system cells to interact with cancer cells and aims to destroy cancer cells, stop their division, or prevent metastases.

In this study, ECM components list was created by using literature information. Gene expression in both protein and RNA level in lung, metastasis and non-metastasis tissues were compared and candidate genes that mediate metastasis from lung to distant tissues were identified. Also, single cell RNA-seq data of samples from non-small cell lung cancer, adjacent normal tissue and lung adenocarcinoma metastasized to brain were analyzed to reveal the interactions between cell types. To analyze the datasets, two packages developed in R and one algorithm developed in Matlab were used. Firstly, GSE117570 and GSE143423 were downloaded from GEO. GSE117570 was analyzed with Seurat and clusters and marker genes were found. These marker genes were

compared with known markers and marker list was generated to identify cell types in data for both P1 and P4 data. Different marker lists were used for assigning cell types because some markers were not applicable to data from both patients. Kumar algorithm was used for identification of cell types in data by considering the marker gene expression in cells in each data. For P1, cancer cells were assigned in the normal tissue data, which may be a result of contamination or other technical problems, hence P1 normal tissue data was excluded from further analysis. There were no cancer cells in P4 data as expected, hence this set was further analyzed for cell-to-cell interactions. For metastasis data, marker genes for related cell types were already provided by the team who produced the data, and by using these markers cell types were assigned. Additionally, P4 analyzed with SingleCellSignalR package to identify cell types and cell types and numbers were shown in Table 4.11. Because cancer cells were found in P4 normal data, SingleCellSignalR was not used for cell type assignment. After assignment of cell types were completed via Kumar's algorithm, SingleCellSignalR was used only to find ligand-receptor interactions between cell types and to score the interaction. This algorithm uses regularized product score to compute interaction score, hence superior to other tools. The tests with Seurat, SingleCellSignalR and Kumar's algorithm demonstrated that alternative tools must be tested before choosing the most appropriate tool and a hybrid pipeline may be the best option to complete a series of tasks.

In this study, only P4 tumor and normal, P1 tumor data and lung brain metastasis data were analyzed, because, cell type markers did not identify the cells in P2, sample size of P3 was not enough to identify cells and cancer cells were identified in P1 normal data which was not expected. The most important technical challenge was to choose reliable markers and determine the cell types. In addition, because of nature of scRNAseq data, some genes could be missing or there could be high levels of noise in the data. Another challenge is the dependence on data available in the literature. Usually, the data in the literature is not designed for addressing the biological question asked by a computational biologist and analyses are restricted to the available data.

6. CONCLUSIONS

In recent years single cell RNA sequencing technologies are widely used. These techniques provide researchers with new opportunities for their studies and give more information about individual cell's expression profile. There are many algorithms or tools to analyze single cell RNA sequencing data. These analyses provide better understanding of biological systems such as diseases especially complex diseases like cancer. By using these analyses new and effective therapies could be discovered against aggressive cancer types such as non-small cell lung cancer.

In this study extracellular matrix of tissues were collected from literature and analyzed lung and more metastatic tissues of NSCLC. NSCLC and lung brain metastasis single-cell RNA sequencing data were analyzed and cell types in data were identified by using bioinformatics approaches. Finally, ligand-receptor interactions between cell types were found in cancer, normal and metastasis conditions and drugs were proposed that targets important interactions.

Analysis of the genes expressing the extracellular matrix components indicated that some genes were shared between lung and metastasis tissues and these genes may regulate metastasis from lung to distant metastatic organs. Genes and metastatic sites were given in Table 4.2, Table 4.3 and Table 4.4. In these analyses the curated information for ECM is from measurements in normal tissues, therefore there is no strong evidence for predicting the genes are mediator in metastasis from lung to distant organs. But this does not exclude the possibility that these genes lead to metastasis in specific organs, as cancer studies in literature support our results (Gu & Giraudo, 2013; Rizeq et al., 2018; Roodink et al., 2009)

Analysis for P4 data, SEMA4D_MET interaction is significantly different in dormant cells when compared to cancer cells and this interaction may lead cancer cells to dormant state as SEMA4D ligand inhibits MET hepatocyte growth factor receptor (Soong & Scott, 2013) and also PSAP and TFPI may lead to dormancy by inhibiting metastasis and proliferation (Kang et al., 2009; G. Wang et al., 2018). For normal tissue data, HSP90AA1_EGFR interaction was found in CAFs and this interaction may regulate ECM and cause abnormality in the cells. The potential of HSP90AA1 as a drug target was shown

previously (Coskunpinar et al., 2014; Liu et al., 2019). Also RPS19_C5AR1 interaction plays important role in developing cancer and decrease in tumor growth by targeting this interaction was shown in literature (Markiewski et al., 2017). For metastasis data analysis, RTN4_LINGO1 interaction is different in dormant cells when compared to cancer cells because of receptor expression level. This interaction regulate cell growth and inhibits PI3K/Akt pathway, and this pathway is also important both cancer development and dormancy mechanism (Fernandez-Enright et al., 2014; Yadav et al., 2018). Lastly, results of drug repositioning analysis showed that, although drugs to target genes involved in the same interactions will be effective, only one patient-specific interactions may exist, and combinations of drugs to target them may be more effective.

For future work, single-cell RNA sequencing data of different tissues will be analyzed and metastasis and dormancy mechanisms will be understood better in a broader range of cancer types.

REFERENCES

- Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., ... Soboleva, A. (2013). NCBI GEO: Archive for functional genomics data sets - Update. *Nucleic Acids Research*. <https://doi.org/10.1093/nar/gks1193>
- Bateman, A. (2019). UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Research*. <https://doi.org/10.1093/nar/gky1049>
- Bidkhor, G., Moeini, A., & Masoudi-Nejad, A. (2012). Modeling of Tumor Progression in NSCLC and Intrinsic Resistance to TKI in Loss of PTEN Expression. *PLoS ONE*, 7(10), 1–13. <https://doi.org/10.1371/journal.pone.0048004>
- Butler, A., Hoffman, P., Smibert, P., Papalexi, E., & Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology*. <https://doi.org/10.1038/nbt.4096>
- Cabello-Aguilar, S., Alame, M., Kon-Sun-Tack, F., Fau, C., Lacroix, M., & Colinge, J. (2020). SingleCellSignalR: inference of intercellular networks from single-cell transcriptomics. *Nucleic Acids Research*. <https://doi.org/10.1093/nar/gkaa183>
- Chen, G., Ning, B., & Shi, T. (2019). Single-cell RNA-seq technologies and related computational data analysis. *Frontiers in Genetics*, 10(APR), 1–13. <https://doi.org/10.3389/fgene.2019.00317>
- Coskunpinar, E., Akkaya, N., Yildiz, P., Oltulu, Y. M., Aynaci, E., Isbir, T., & Yaylim, I. (2014). The significance of HSP90AA1, HSP90AB1 and HSP90B1 gene polymorphisms in a Turkish population with non-small cell lung cancer. *Anticancer Research*.
- Dagogo-Jack, I., & Shaw, A. T. (2018, February 1). Tumour heterogeneity and resistance to cancer therapies. *Nature Reviews Clinical Oncology*. Nature Publishing Group. <https://doi.org/10.1038/nrclinonc.2017.166>
- Enderling, H., Hahnfeldt, P., Hlatky, L., & Almog, N. (2012). Systems biology of tumor dormancy: Linking biology and mathematics on multiple scales to improve cancer therapy. In *Cancer Research* (Vol. 72, pp. 2172–2175). NIH Public Access. <https://doi.org/10.1158/0008-5472.CAN-11-3269>
- Endo, H., & Inoue, M. (2019). Dormancy in cancer. *Cancer Science*, 110(2), 474–480. <https://doi.org/10.1111/cas.13917>
- Fernandez-Enright, F., Andrews, J. L., Newell, K. A., Pantelis, C., & Huang, X. F. (2014). Novel implications of Lingo-1 and its signaling partners in schizophrenia. *Translational Psychiatry*. <https://doi.org/10.1038/tp.2013.121>

- Frantz, C., Stewart, K. M., & Weaver, V. M. (2010). The extracellular matrix at a glance. *Journal of Cell Science*, 123(24), 4195–4200. <https://doi.org/10.1242/jcs.023820>
- Franzén, O., Gan, L. M., & Björkegren, J. L. M. (2019). PanglaoDB: A web server for exploration of mouse and human single-cell RNA sequencing data. *Database*. <https://doi.org/10.1093/database/baz046>
- Gao, Y., Bado, I., Wang, H., Zhang, W., Rosen, J. M., & Zhang, X. H. F. (2019). Metastasis Organotropism: Redefining the Congenial Soil. *Developmental Cell*, 49(3), 375–391. <https://doi.org/10.1016/j.devcel.2019.04.012>
- Gomis, R. R., & Gawrzak, S. (2017). Tumor cell dormancy. *Molecular Oncology*, 11(1), 62–78. <https://doi.org/10.1016/j.molonc.2016.09.009>
- Gu, C., & Giraudo, E. (2013). The role of semaphorins and their receptors in vascular development and cancer. *Experimental Cell Research*. <https://doi.org/10.1016/j.yexcr.2013.02.003>
- Hanahan, D., & Weinberg, R. A. (2011). Hallmarks of cancer: The next generation. *Cell*, 144(5), 646–674. <https://doi.org/10.1016/j.cell.2011.02.013>
- Hwang, B., Lee, J. H., & Bang, D. (2018). Single-cell RNA sequencing technologies and bioinformatics pipelines. *Experimental and Molecular Medicine*, 50(8). <https://doi.org/10.1038/s12276-018-0071-8>
- Kang, S. Y., Halvorsen, O. J., Gravdal, K., Bhattacharya, N., Jung, M. L., Liu, N. W., ... Watnick, R. S. (2009). Prosaposin inhibits tumor metastasis via paracrine and endocrine stimulation of stromal p53 and Tsp-1. *Proceedings of the National Academy of Sciences of the United States of America*. <https://doi.org/10.1073/pnas.0903120106>
- Kumar, M. P., Du, J., Lagoudas, G., Jiao, Y., Sawyer, A., Drummond, D. C., ... Raue, A. (2018). Analysis of Single-Cell RNA-Seq Identifies Cell-Cell Communication Associated with Tumor Characteristics. *Cell Reports*, 25(6), 1458–1468.e4. <https://doi.org/10.1016/j.celrep.2018.10.047>
- Liu, K., Kang, M., Li, J., Qin, W., & Wang, R. (2019). Prognostic value of the mRNA expression of members of the HSP90 family in non-small cell lung cancer. *Exp Ther Med*, 17(4), 2657–2665. <https://doi.org/10.3892/etm.2019.7228>
- Lu, P., Weaver, V. M., & Werb, Z. (2012). The extracellular matrix: A dynamic niche in cancer progression. *Journal of Cell Biology*, 196(4), 395–406. <https://doi.org/10.1083/jcb.201102147>
- Markiewski, M. M., Vadrevu, S. K., Sharma, S. K., Chintala, N. K., Ghouse, S., Cho, J.-H., ... Karbowiczek, M. (2017). The Ribosomal Protein S19 Suppresses

Antitumor Immune Responses via the Complement C5a Receptor 1. *The Journal of Immunology*. <https://doi.org/10.4049/jimmunol.1602057>

Naba, A., Clauser, K. R., Ding, H., Whittaker, C. A., Carr, S. A., & Hynes, R. O. (2016). The extracellular matrix: Tools and insights for the “omics” era. *Matrix Biology*, 49, 10–24. <https://doi.org/10.1016/j.matbio.2015.06.003>

Olsen, T. K., & Baryawno, N. (2018). Introduction to Single-Cell RNA Sequencing. *Current Protocols in Molecular Biology*, 122(1), 1–14. <https://doi.org/10.1002/cpmb.57>

Papathodorou, I., Moreno, P., Manning, J., Fuentes, A. M. P., George, N., Fexova, S., ... Brazma, A. (2020). Expression Atlas update: From tissues to single cells. *Nucleic Acids Research*. <https://doi.org/10.1093/nar/gkz947>

Peinado, H., Zhang, H., Matei, I. R., Costa-Silva, B., Hoshino, A., Rodrigues, G., ... Lyden, D. (2017). Pre-metastatic niches: Organ-specific homes for metastases. *Nature Reviews Cancer*, 17(5), 302–317. <https://doi.org/10.1038/nrc.2017.6>

Pickup, M. W., Mouw, J. K., & Weaver, V. M. (2014). The extracellular matrix modulates the hallmarks of cancer. *EMBO Reports*, 15(12), 1243–1253. <https://doi.org/10.15252/embr.201439246>

Qiu, X., Mao, Q., Tang, Y., Wang, L., Chawla, R., Pliner, H. A., & Trapnell, C. (2017). Reversed graph embedding resolves complex single-cell trajectories. *Nature Methods*. <https://doi.org/10.1038/nmeth.4402>

Raudvere, U., Kolberg, L., Kuzmin, I., Arak, T., Adler, P., Peterson, H., & Vilo, J. (2019). G:Profiler: A web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Research*. <https://doi.org/10.1093/nar/gkz369>

Rizeq, B., Zakaria, Z., & Ouhtit, A. (2018). Towards understanding the mechanisms of actions of carcinoembryonic antigen-related cell adhesion molecule 6 in cancer progression. *Cancer Science*. <https://doi.org/10.1111/cas.13437>

Rodriguez-Canales, J., Parra-Cuentas, E., & Wistuba, I. I. (2016). Diagnosis and molecular classification of lung cancer. In *Cancer Treatment and Research* (Vol. 170, pp. 25–46). Kluwer Academic Publishers. https://doi.org/10.1007/978-3-319-40389-2_2

Roodink, I., Verrijp, K., Raats, J., & Leenders, W. P. J. (2009). Plexin D1 is ubiquitously expressed on tumor vessels and tumor cells in solid malignancies. *BMC Cancer*. <https://doi.org/10.1186/1471-2407-9-297>

Schwab, M. E. (2010). Functions of Nogo proteins and their receptors in the nervous system. *Nature Reviews Neuroscience*. <https://doi.org/10.1038/nrn2936>

Seyfried, T. N., & Huysentruyt, L. C. (2013). On the origin of cancer metastasis. *Critical Reviews in Oncogenesis*, 18(1–2), 43–73.

Song, Q., Hawkins, G. A., Wudel, L., Chou, P. C., Forbes, E., Pullikuth, A. K., ... Zhang, W. (2019). Dissecting intratumoral myeloid cell plasticity by single cell RNA-seq. *Cancer Medicine*, 8(6), 3072–3085. <https://doi.org/10.1002/cam4.2113>

Soong, J., & Scott, G. (2013). Plexin B1 inhibits MET through direct association and regulates Shp2 expression in melanocytes. *Journal of Cell Science*. <https://doi.org/10.1242/jcs.119487>

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., ... Surani, M. A. (2009). mRNA-Seq whole-transcriptome analysis of a single cell. *Nature Methods*, 6(5), 377–382. <https://doi.org/10.1038/nmeth.1315>

The Mathworks Inc. (2018). MATLAB R2018a. www.mathworks.com/products/matlab. <https://doi.org/2016-11-26>

Uhlén, M., Fagerberg, L., Hallström, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., ... Pontén, F. (2015). Tissue-based map of the human proteome. *Science*. <https://doi.org/10.1126/science.1260419>

Uzuner, D. (2020). INTEGRATIVE ANALYSIS OF TRANSCRIPTOME DATA AND CELLULAR NETWORKS TO REVEAL MOLECULAR INTERACTIONS OF METASTASIS MECHANISMS IN CANCER. *GebzeTechnical University*.

Valastyan, S., & Weinberg, R. A. (2011). Tumor metastasis: Molecular insights and evolving paradigms. *Cell*, 147(2), 275–292. <https://doi.org/10.1016/j.cell.2011.09.024>

Van Der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*.

Wang, B., Zhu, J., Pierson, E., Ramazzotti, D., & Batzoglou, S. (2017). Visualization and analysis of single-cell rna-seq data by kernel-based similarity learning. *Nature Methods*. <https://doi.org/10.1038/nMeth.4207>

Wang, G., Huang, W., Li, W., Chen, S., Chen, W., Zhou, Y., ... Gu, W. (2018). TFPI-2 suppresses breast cancer cell proliferation and invasion through regulation of ERK signaling and interaction with actinin-4 and myosin-9. *Scientific Reports*. <https://doi.org/10.1038/s41598-018-32698-3>

Wang, L., Dai, J., Han, R.-R., Dong, L., Feng, D., Zhu, G., ... Jin, W.-L. (2019). Single-Cell Map of Diverse Immune Phenotypes in the Metastatic Brain Tumor Microenvironment of Non Small Cell Lung Cancer. *BioRxiv*, 2019.12.30.890517. <https://doi.org/10.1101/2019.12.30.890517>

Wang, Z., Gerstein, M., & Snyder, M. (2010). Nihms229948, 10(1), 57–63.

<https://doi.org/10.1038/nrg2484.RNA-Seq>

Web1. (2020). <https://www.cancer.org/cancer/lung-cancer/treating-non-small-cell/targeted-therapies.html>.

Web2. (2020). <https://www.cancer.gov/about-cancer/treatment/types/immunotherapy/t-cell-transfer-therapy>.

Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., ... Wilson, M. (2018). DrugBank 5.0: A major update to the DrugBank database for 2018. *Nucleic Acids Research*. <https://doi.org/10.1093/nar/gkx1037>

Wood, S. L., Pernemalm, M., Crosbie, P. A., & Whetton, A. D. (2014). The role of the tumor-microenvironment in lung cancer-metastasis and its relationship to potential therapeutic targets. *Cancer Treatment Reviews*, 40(4), 558–566. <https://doi.org/10.1016/j.ctrv.2013.10.001>

Yadav, A. S., Pandey, P. R., Butti, R., Radharani, N. N. V., Roy, S., Bhalara, S. R., ... Kumar, D. (2018). The biology and therapeutic implications of tumor dormancy and reactivation. *Frontiers in Oncology*, 8(MAR), 1–11. <https://doi.org/10.3389/fonc.2018.00072>

Yeh, A. C., & Ramaswamy, S. (2015). Mechanisms of cancer cell dormancy-another hallmark of cancer? *Cancer Research*, 75(23), 5014–5022. <https://doi.org/10.1158/0008-5472.CAN-15-1370>

Zhang, X., Lan, Y., Xu, J., Quan, F., Zhao, E., Deng, C., ... Xiao, Y. (2019). CellMarker: A manually curated resource of cell markers in human and mouse. *Nucleic Acids Research*. <https://doi.org/10.1093/nar/gky900>

Zheng, G. X. Y., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., ... Bielas, J. H. (2017). Massively parallel digital transcriptional profiling of single cells. *Nature Communications*. <https://doi.org/10.1038/ncomms14049>

Zhou, G., Soufan, O., Ewald, J., Hancock, R. E. W., Basu, N., & Xia, J. (2019). NetworkAnalyst 3.0: A visual analytics platform for comprehensive gene expression profiling and meta-analysis. *Nucleic Acids Research*. <https://doi.org/10.1093/nar/gkz240>

Zhou, Z., Sathiyamoorthy, S., & Tan, E.-K. (2012). LINGO-1 and Neurodegeneration: Pathophysiologic Clues for Essential Tremor? *Tremor and Other Hyperkinetic Movements*. <https://doi.org/10.5334/tohm.111>

BIOGRAPHY

Cemal Yıldız was born in İskenderun in October 13, 1993. He graduated from İnönü University Department of Molecular Biology and Genetics in 2016. He is a M.Sc student in Bioinformatics and Systems Biology Program under Bioengineering Department at Graduate School of Natural and Applied Sciences in Gebze Technical University.

APPENDICES

Appendix A: The formulas to compute ligand-receptor score.

Kumar algorithm computation: $LRscore = \frac{l}{n1} * \frac{r}{n2}$

SingleCellSignalR computation: $LRscore = \frac{\sqrt{lr}}{\mu + \sqrt{lr}}$

SingleCellSignalR normalization: $x = \frac{C(i)}{q(i)/median(q)}$

Where l means ligand expression, r means receptor expression, n1 means number of interested cell type, n2 means number of interested cell type, μ means mean of normalized count matrix, x means normalized value, C means count matrix, q means quantile vector, i means sample.

Appendix B: All interaction results.

Analysis results are given as Supplementary File for all datasets used in this study. For each dataset a folder was generated and following files were provided for each dataset:

- “cell_signaling” folder includes all interactions for both autocrine and paracrine between cell types identified in data.
- “images_autocrine” folder contains chord plots for all autocrine interactions.
- “images_paracrine” folder contains chord plots for all paracrine interactions.
- “networks_autocrine” folder contains all autocrine interactions as graphml format
- “networks_paracrine” folder contains all paracrine interactions as graphml format