ECM-Guided Integrative Network Modeling for Patient Stratification

by

Aslı Dansık

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

Master of Science

in

Chemical and Biological Engineering

KOÇ ÜNİVERSİTESİ

August 8, 2024

ECM-Guided Integrative Network Modeling for Patient Stratification

Koç University

Graduate School of Sciences and Engineering

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

Aslı Dansık

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

Committee Members:

_____

Assoc. Prof. Nurcan Tunçbağ (Advisor)

_____

Assist. Prof. Ece Öztürk [Co-Advisor]

_____

Prof. Z. Özlem Keskin Özkaya

_____

Prof. Türkan Haliloğlu

_____

Assoc. Prof. Özlen Konu Karakayalı

Date: _____

To my country,

ABSTRACT

ECM-Guided Integrative Network Modeling for Patient Stratification
Aslı Dansık
Master of Science in Chemical and Biological Engineering
August 8, 2024

The extracellular matrix (ECM) plays a crucial role in tumor initiation, progression, and drug response. Consequently, the gene/protein expression signatures of ECM in tumors could serve as significant prognostic factors. This study aims to stratify lung adenocarcinoma (LUAD) patients using an ECM-guided multi-omic approach and further define ECM profiles of these groups through network modeling, providing deeper insights into prognosis and facilitating the selection of patient-specific therapies. Publicly available multi-omic datasets from 101 LUAD patients with paired tumor-normal adjacent tissue samples were used in this study. Tumors were labeled based on the multi-omic expression scores of ECM categories, followed by consensus clustering, resulting in four distinct patient clusters. For each patient, important ECM genes and transcriptional regulators were identified and utilized in network modeling. Various enrichment analyses were performed on the resulting networks, and scores were calculated for all enriched pathways, hallmarks, and transcriptional regulators for each patient. For each cluster, a consensus network was constructed using the patient networks, and a proximity analysis was performed to identify effective drug molecules. This study revealed patient groups representing different ECM grades, and these ECM grades showed distinct clinical features, mutation profiles, and cellular heterogeneity. It was shown that pathways closely relevant to metastasis like epithelial-to-mesenchymal transition and angiogenesis, cancer stemness like Wnt and hedgehog signaling, and cancer-related inflammation like TNF-α and NF-κB were also highly enriched in ECM grades of high severity. Drug screening revealed drugs that could be more effective on patients of a particular ECM grade. Overall, with this thesis, we were able to capture the heterogeneous nature of the ECM observed in tumors by partitioning patients with discrete ECM characteristics into groups that can be targeted with unique therapeutic approaches.

ÖZETÇE

ECM Yönlendirmeli Entegre Ağ Modellemesiyle Hasta Sınıflandırması
Aslı Dansık
Kimya ve Biyoloji Mühendisliği, Yüksek Lisans
8 Ağustos 2024

Hücre dışı matris (ECM), tümor başlangıcı, ilerlemesi ve ilaç yanıtı üzerinde önemli rol oynar. Dolayısıyla, ECM gen ve protein imza ifadeleri hastalık seyri için önemli faktörler olabilirler. Bu çalışma, akciğer adenokarsinoma hastalarını hücre dışı ortamlarına göre çoklu-omik yaklaşım kullanarak sınıflandırmayı ve ağ modellemesiyle bu sınıfların hücre dışı profillerini detaylandırmayı hedeflemektedir. Benimsenen yaklaşım, hastalık seyriyle ilgili fikir elde etmeyi ve hastaya özgü tedavi yöntemlerinin seçimini kolaylaştırmayı sağlayacaktır. Bu çalışmada, 101 akciğer adenokarsinoma hastasının, eşli tümor ve bitişik normal doku örneklerinin çoklu-omik verilerini barındıran halka açık bir veri seti kullanılmıştır. Tümörler, ECM kategorilerinin çoklu-omik ifadelerine dayanan skorlarla etiketlendirilip, oydaşım kümlenedirme yöntemiyle dört farklı gruba ayrılmıştır. Her hasta için önemli olan ECM genleri ve onların transkripsiyonel düzenleyicileri belirlenip, hastaya özgü ağ modellemesinde kullanılmıştır. Bu ağlar üzerinde çeşitli zenginleştirme analizleri yapılıp bulunan her yolak, ana işaret ve transkripsiyon faktörü için bir skor hesaplanmıştır. Her hasta kümesi için, kümeye ait hasta ağlarının bileştirilmesiyle bir oydaşım ağı oluşturulmuş ve etkin ilaçları belirlemek üzere yakınlık analizinde kullanılmıştır. Bu çalışma, farklı ECM kademelerini temsil eden hasta gruplarını açığa çıkarmıştır ve bu kademelerin farklı klinik özellikler, mutasyon profilleri ve hücresel heterojenite taşıdığını göstermiştir. Epitel mezenkimal geçiş ve anjiyogenez gibi metastazla ilgili yolakların, Wnt ve Hedgehog gibi kanser kök hücreleriyle ilgili sinyal yolaklarının ve TNF-α ve NF-κB gibi kansere bağlı enflamasyonla ilgili sinyal yolaklarının ECM kademesi arttıkça zenginleştiği gösterilmiştir. İlaç taraması, belirli ECM kademelerine daha etkili olabilecek ilaç moleküllerini ortaya çıkarmıştır. Özet olarak, bu tezle, hücre dışı matrisin tümörlerde gözlemlenen heterojen doğasını, hastaları farklı ECM niteliklerine göre gruplayarak yakalanmış ve bu grupların özgün ilaçlarla hedeflenebildiği gösterilmiştir.

ACKNOWLEDGMENTS

First and foremost, I would like to express my deepest gratitude to my advisors, Assoc. Prof. Dr. Nurcan Tunçbağ and Assist. Prof. Ece Öztürk for the immense guidance they have provided for me to search for and find knowledge, but also for the wisdom they shared in their fields. They have not only taught me how to do research but also how to overcome the stress we often face in academia by openly sharing their experiences. I would like to thank them for the trust they have in me and for their sincerity.

Furthermore, I would like to thank Prof. Z. Özlem Keskin Özkaya, Prof. Türkan Haliloğlu, and Assoc. Prof. Özlen Konu Karakayalı for accepting to be on my thesis committee and sparing their time to criticize my work. I appreciate the valuable feedback that you have provided.

I would like to thank the TÜBİTAK ARDEB 1001 program for their funding of project 121E245.

I would also like to thank my colleagues, Ceren Uzun, Sefa Ayar, Sina Dadmand, Ülkem Kasapoğlu, Cansu Demirel, Esra Başaran, Yiğit Şibal, Berkay Günay, Sertan Dikli, Can Dedeköy, Serra Doğanata, Ekin Erdem, Sevgi Sarıca, Sena Özkan, Kardelen Yangın, Gökçe Serttaş, Alican Kuşoğlu, and Emir Tekoğlu for the great scientific and social discussions we had, especially during lunch and coffee breaks. I especially thank İdil İlayda Duran for always being so supportive and helpful. I would also like to thank my dear friend Berfin Çobanoğlu for her joyful companionship.

Finally, I would like to thank my parents, Tülin and Fevzi Dansık, and my brother Cem Dansık for always supporting me in every way. I deeply appreciate their understanding and the comfort they have provided me in following my path. I also thank my beloved, Ali Berkay Kan, for being the person that I can trust, no matter what, and always finding a way to make me smile.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

AACR	American Association for Cancer Research
ABC	ATP-Binding Cassette
ADAMs	A Disintegrin and Metalloproteases
APOLLO	Applied Proteogenomics Organizational Learning and Outcomes
BM	Basement Membrane
CAFs	Cancer-Associated Fibroblasts
CIMP	CpG Island Methylator Phenotype
CPTAC	Clinical Proteomic Tumor Analysis Consortium
CSC	Cancer Stem Cell
DEGs	Differentially Expressed Genes
DEPOD	The human Dephosphorylation Database
DOROTHEA	Discriminant Regulon Expression Analysis
DOX	Doxorubicin
ECM	Extracellular Matrix
EMT	Epithelial-Mesenchymal Transition
EPIC	Estimating the Proportions of Immune and Cancer Cells
ESTIMATE	Estimation of STromal and Immune cells in MAlignant Tumor tissues using Expression
FAK	Focal Adhesion Kinase
FDA	U.S. Food and Drug Administration
GAGs	Glycosaminoglycans
GDC	Genomic Data Commons

GENIE	Genomics Evidence Neoplasia Information Exchange
GO	Gene Ontology
HA	Hyaluronic Acid
HIPPIE	Human Integrated Protein-Protein Interaction Reference
HR	Hazard Ratio
KEGG	Kyoto Encyclopedia of Genes and Genomes
LOX	Lysyl Oxidases
LUAD	Lung Adenocarcinoma
MCP-Counter	Microenvironment Cell Populations-Counter
MET	c-Met receptor tyrosine kinase
MMPs	Matrix Metalloproteinases
MSigDB	Molecular Signatures Database
M3C	Monte Carlo Reference-Based Consensus Clustering
NAT	Normal Adjacent Tissue
NCI	The National Cancer Institute
NFκB	Nuclear Factor κB
NK Cells	Natural Killer Cells
NSCLC	Non-Small Cell Lung Cancer
PAC	Proportion of Ambiguous Clustering
PARP	Poly (ADP-Ribose) Polymerase
PCA	Principal Component Analysis
PCSF	Prize Collecting Steiner Forest
PCST	Prize Collecting Steiner Tree
PDC	Proteomic Data Commons
PI	Proximal Inflammatory
PP	Proximal Proliferative
PPI	Protein-Protein Interaction
PTMs	Post-Translational Modifications

RPKM	Reads Per Kilobase Million
RCSI	Relative Cluster Stability Index
ssGSEA	Single Sample Gene Set Enrichment Analysis
TARGET	Therapeutically Applicable Research to Generate Effective Treatments
TCGA	The Cancer Genome Atlas
TFs	Transcription Factors
TIMER	Tumor IMmune Estimation Resource
TME	Tumor Microenvironment
TMT	Tandem Mass Tag
TRNs	Transcriptional Regulatory Networks
TRU	Terminal Respiratory Unit
TRRUST	Transcriptional Regulatory Relationships Unraveled by Sentence-based Text Mining
VEGF	Vascular Endothelial Growth Factor
VIPER	Virtual Inference of Protein Activity by Enriched Regulon Analysis

Chapter 1

SCOPE OF THE THESIS

The extracellular matrix (ECM) is the non-cellular component within the tissues and acts as structural support for the tissues via collagen fibers and adhesive proteins like fibronectins. The ECM also plays an important role in the adhesion of the cells to their environment, mainly via integrins. ECM components like proteoglycans and glycosaminoglycans (GAGs) can bind to many soluble or unsoluble factors, which enables the ECM to act as reservoirs for growth factors and other bioactive molecules and play an important role in cell signaling [Yue, 2014, Barkovskaya et al., 2020]. ECM not only induces biochemical signaling but is also a major regulator of mechanosensing and mechanotransduction in cells [Mierke, 2024]. The network of ECM proteins, their interactions with each other and the cells, and their capacity to present or enable diffusion of various biomolecules result in the regulation of many biological processes, including differentiation, migration, immune cell infiltration, proliferation, wound healing, etc. ECM is also not a static mesh of proteins but is dynamic; it constantly undergoes remodeling via regulators, crosslinkers, or degradation via proteases. Therefore, its composition, mechanical properties, and abundance can be altered in specific situations [Lu et al., 2011].

The disruption of the ECM is a notable aspect of cancer within the tumor microenvironment (TME). Cancer cells play a role in increasing ECM stiffness, which in turn affects the behavior of the cells within the TME [Huang et al., 2021]. The rigid and dense ECM can act as a barrier to diffusion, making it difficult for drugs to access tumor cells and shielding the tumor from receiving therapeutically effective doses of drugs. Similarly, reduced diffusion through the ECM can also limit the supply of nutrients and oxygen to the cells. In return, pathological responses to

metabolic stress and hypoxia, such as increased expression of drug efflux pumps and impaired apoptosis and senescence, can make the drugs less effective even if they reach the undersupplied cells. In addition to the physical aspects, direct contact of the cells with the ECM can affect pathways involved in the cellular response to cytotoxic stress. Activation of integrin and FAK can lead to increased prosurvival signaling, reduced apoptotic response, and avoidance of cell cycle arrest induced by chemotherapy damage. It can also trigger epithelial-to-mesenchymal transition (EMT) [Henke et al., 2020]. Altered ECM characteristics are not only influential on the tumor cells but can also induce abnormal vasculature and lymphatic networks within the TME, interfere with cell migration into and out of the TME, or even aid in the recruitment of healthy stromal cells into malignant cells [Quail and Joyce, 2013, Kessenbrock et al., 2010]. Thus, gaining a thorough understanding of ECM dysregulation in the TME could lead to the identification of potential targets for cancer therapy.

Although it is shown that the ECM is altered in the TME, the means of these alterations are not always the same in every tumor. Puttock et al. could show five different ECM composition groups by performing hierarchical clustering on the proteomic data of ovarian cancer samples [Puttock et al., 2023]. Similarly, Bergamaschi et al. showed four ECM subgroups in breast carcinomas by performing differential expression analysis on a list of ECM genes, further showing the clinical relevance of these groups [Bergamaschi et al., 2008]. Parker et al. also showed two matreotypes in the squamous cell carcinoma subtype of non-small cell lung cancer (NSCLC) based on the expression levels of the ECM genes, with one matreotype showing low expression and the other showing higher expression, and further investigated the differences in these matreotypes [Parker et al., 2022]. Therefore, it can be concluded that the alterations in the ECM of tumors are heterogeneous among tumor samples and can show different clinical outcomes.

Lung adenocarcinoma (LUAD) is the predominant form of NSCLC in the United States. Despite a decrease in its occurrence and fatalities, it still ranks as the foremost cause of cancer-related deaths [Kim and Reckamp, 2024]. Given the impact of ECM alterations on tumor establishment, progression, invasiveness, and drug

response, it becomes important to demonstrate various ECM profiles within patients.

While the aforementioned studies could show that the matrisome differs among groups of patients, they were not sufficient to shed light on the underlying regulatory mechanisms that resulted in such differences. Moreover, they were limited to single omic layers. Therefore, this study aims to group LUAD patients based on their ECM profiles and investigate the reasons behind these profiles using multi-omic data integration with network modeling. This approach aims to understand the regulation of ECM deposition, investigate the distinctive characteristics of different ECM profiles, and ultimately identify drug targets or guide the decision-making of treatment strategies.

In this thesis, we first investigate the different ECM expression tumor and normal samples have at multi-omic scale and calculate a difference value to show ECM dysregulation. With this calculation, we barcode the patients, and with these barcodes, we perform reference-based consensus clustering to reveal patient groups of different ECM characteristics. Simultaneously, we construct patient networks using multi-omic data, which gives information on triggered pathways and signaling processes that occur to ultimately give rise to altered ECM profiles within patients. We then further investigate the differences these ECM groups have on aspects like cellularity, tumor purity, enriched pathways, and transcriptional regulators. Finally, we perform network proximity analysis to reveal possible drugs that may have significant potential to be efficient on patients of a particular ECM group.

In Chapter 2, we provide a thorough description of the literature regarding ECM and network biology. We briefly explain the complex nature of ECM and how it is altered by the TME. We inform on how these alterations influence many aspects of tumors, like their progression, establishment, metastasis, cellularity, and drug response. We mention recent therapeutic strategies that target the ECM directly or indirectly. We also report the benefits of multi-omic data integration, how this can be done via network modeling, and how these models can be used in medicine.

In Chapter 3, we describe in detail the methodology of the thesis and the principles of the tools used. We first introduce the cohort and the datasets. Then we present our calculations on the barcoding of the patients and their clustering based

on their ECM profiles. We show our method for the investigation of patient groups from various perspectives, like clinical features, cellular heterogeneity, and correlation with the expression of important gene sets. We share the network modeling strategy performed on each patient and the similarity and enrichment analysis done on these networks. And, finally, we describe the consensus network construction process followed by network proximity analysis.

In Chapter 4, we present and discuss our results in detail. We first show the ECM characteristics in the cohort and their difference in tumor and normal samples. Then, we show our results on identified ECM groups and describe the different features these groups have. We describe how networks represent these patients and the enriched pathways and transcriptional regulators related to the patient networks. And, finally, we give our results on networks that best represent these patient groups and possible drugs that may target these ECM group-specific networks.

In Chapter 5, we share the highlights and major conclusions we have arrived at from this thesis and discuss how this work contributes to the field. We also discuss the drawbacks of the thesis and some future work.

Chapter 2

INTRODUCTION

2.1 The Extracellular Matrix

2.1.1 The ECM of the Tumor Microenvironment

The ECM is the non-cellular component of the tissues. It consists of multi-domain macromolecules that bind to each other or other biomolecules and form a network in which the cells reside. This formation of a network enables the tissues to have a stable structure and provides mechanical support. Major structural proteins called collagens play a great part in the physical structure, and elastin contributes to the elasticity of the ECM. One of the important roles of the ECM is to provide a surface for the cells to attach to since many cells within our tissues are adherent and if not attached to a surface undergo a form of programmed cell death called anoikis. Fibronectins, which are referred to as "biological glues" are a very important component of the ECM that aids the attachment of the cells. In addition to that, cells bind to collagens of the ECM via integrins. The basement membrane (BM), which is another form of ECM that is in a sheet-like structure, is comprised primarily of laminin and collagen type IV and is responsible for compartmentalizing the tissues [Morrissey and Sherwood, 2015]. The ECM also acts as a reservoir for bioactive molecules like growth factors and cytokines and regulates their availability. Proteoglycans, in particular heparan sulfate proteoglycans, strongly bind to growth factors like fibroblast growth factor family and vascular endothelial growth factors (VEGF) and interact with receptors of these growth factors [Kim et al., 2011]. The ECM shows a dynamic nature. Since it is in constant communication with cells, it can undergo remodeling. Depending on stimuli, it can be secreted by cells or degraded by various proteases like matrix metalloproteinases (MMPs) and a disintegrin and metalloproteases (ADAMs). It can also be crosslinked by lysyl oxidases

(LOX) [Bonnans et al., 2014]. Given such properties, the ECM can conduct physical and chemical cues. Since it is the component of the tissues that fill in between the cells, cell-to-cell communication is also transmitted through the ECM. Similarly, the reserving of the bioactive molecules exemplifies chemical message passing and the cell-matrix interactions via integrin play a crucial role in mechanotransduction. These all in turn show that the ECM has control over cellular characteristics and processes including differentiation, adhesion, migration, polarity, proliferation, and apoptosis [Yue, 2014].

The dynamic reciprocity of the ECM makes it unique to tissues and organs. Not surprisingly, different disease and cancer types also show altered ECM characteristics. ECM synthesis, crosslinking, and degradation in tumor tissues are highly dysregulated. This dysregulation may sometimes facilitate tumor progression or act as an inhibitor [Cox, 2021]. In general, ECM components are at elevated levels in tumors. This increased deposition of ECM and thus the formation of a dense fibrous tissue is defined as desmoplasia and it is reported to cause resistance to treatment and poor prognosis. The main source of desmoplastic reactions is cancer-associated fibroblasts (CAFs) along with tumor cells. In addition to elevated ECM amounts, incorrect organization of the ECM is another characteristic feature of tumors [Sainio, 2013].

Collagen which is one of the major components of the ECM and its crosslinker enzyme LOX are both expressed at high levels in many cancers and this can cause excessive crosslinking of collagens. This in return results in increased rigidity of tumors. Increased rigidity then would exert mechanical force on the cells triggering focal adhesion kinase (FAK) mediated signaling [Chuang et al., 2022]. Another phenomenon involving collagens is their alignment. Some tumor cells can organize fibrillar collagens in such a way that they align and serve as highways for tumor cells to migrate [Song et al., 2022].

Proteoglycans, which are made up of GAGs linked to a core protein also show altered expression in different cancer types. For example, glypican-3 is shown to be overexpressed in lung cancer and hepatocarcinoma and takes part in hedgehog signaling [Zhou et al., 2017]. Syndecan-1 and syndecan-4 expression are shown to

point at breast carcinomas [Lendorf et al., 2011]. Glycan-1 is considered a biomarker for colorectal and pancreatic cancers [Li et al., 2017]. Proteoglycan expression level itself can serve as a biomarker yet sulfation levels are also important factors in patient survival. For example, reduced sulfation presents higher overall survival in patients with lung cancer[Rangel et al., 2018]. Hyaluronic acid (HA), which is a type of GAG, is not linked to a core protein and similarly with collagens and proteoglycans is shown to be expressed at high levels in many cancers. It is of particular importance due to its role as a ligand for CD44 which is a marker of stem cells and excessive HA-CD44 interaction may increase metastasis and invasiveness in cancer [Heldin et al., 2014].

The TME holds a very rich cellular composition. It not only consists of proliferative cancer cells, or the ECM but also houses many other types of stromal and immune cells. It is known that tumor cells also present various subclones with distinct mutations, and this may influence the faith of the tumor tissue, especially in the presence of cancer stem cells (CSC), which are shown to play a great part in recurrence due to differentiation and self-renewal abilities [Alizadeh et al., 2015, Cui et al., 2021]. The TME also includes stromal, immune, and vascular cells. Native and healthy cells may be recruited or attracted to the tumor site to serve for its growth. For example, hypoxic conditions created within the tumor may result in the recruitment of endothelial cells and the formation of new blood vessels that would provide nutrient and oxygen supply to the tumor [Liu et al., 2023]. Stromal cells like fibroblasts can be transformed into CAFs and take part in tumorigenesis and metastasis [Zhao et al., 2023]. Innate immune cells like macrophages, neutrophils, and dendritic cells also infiltrate the TME. The content of these cells is important in tumors because it is shown that the M1 subtype of macrophages shows an anti-tumorigenic effect while M2 is pro-tumorigenic [Lu et al., 2023]. A similar context is also present for T cells. A type of CD4+ T cells called regulatory T cells is shown to be very abundant in the TME and show immunosuppressive effects on tumors [Togashi et al., 2019, Tay et al., 2023]. Considering the rich cell population within the TME and how different cell types and states may influence tumor progression, there exist methods that can interpret cellular content from bulk RNA-Seq

data. Some of the computational tools include CIBERSORT, which uses v-support vector regression; Tumor IMmune Estimation Resource (TIMER), which uses linear least squares regression; xCell, which uses single-sample Gene Set Enrichment Analysis (ssGSEA); Microenvironment Cell Populations-Counter (MCP-Counter), which takes the average expression of cell markers; Estimating the Proportions of Immune and Cancer Cells (EPIC); and quanTIseq, which both use constrained least squares regression [Newman et al., 2015, Li et al., 2016, Aran et al., 2017, Becht et al., 2016, Racle et al., 2017, Finotello et al., 2019]. These tools mainly focus on deconvolving immune cells, but some of them also include the quantification of CAFs and endothelial cells.

2.1.2 Influence of the ECM Alterations on Tumor Progression

Abnormal ECM composition, remodeling, organization, and abundance are influential on many hallmarks of cancer. This specialized form of ECM in the TME can affect the proliferation of cancer cells and their genome stability. It was shown that tumor cells proliferate more rapidly in stiffer matrices than soft. In addition, this stiffened matrix also may take part in the rupture of the nuclear envelope which could lead to DNA damage [Alonso-Nocelo et al., 2018, Tamiello et al., 2013]. The ability of the ECM to pool growth factors is another aspect that supports tumor growth and progression. Other biological processes influenced by the ECM properties include abnormal angiogenesis, metastasis, resistance to therapeutics, and immunosuppression [Huang et al., 2021].

Role of the ECM in Each Step of Cancer Metastasis

The concept of metastasis briefly involves tumor cells acquiring motility and invasive capability followed by modulation to a secondary site and colonize [Fares et al., 2020]. The ECM takes a central role in these features of metastasis.

The ECM has been shown to aid the invasive behavior of the tumor cells in many ways. As mentioned earlier, tumor cells can reorganize collagen fibers and use them for migration or secrete enzymes for matrix degradation at the tumor periphery

and start invasion [Feinberg et al., 2018]. Increased collagen I content in the TME or stiffer ECM can result in reorganization of the cytoskeleton, forming an actin-rich protrusion which is an indication of cellular movement [Coelho and McCulloch, 2018].

EMT is a process in which cells switch from a polarized, immotile, epithelial phenotype to a motile, stem-cell-like mesenchymal phenotype [Ribatti et al., 2020]. EMT is closely related to metastasis since cells acquire invasive characteristics. The ECM is shown to induce EMT mainly via stiffening. Nuclear translocation of transcription factors (TFs) that are known to drive EMT like TWIST1, YAP, and TAZ was observed with increased stiffness. This was followed by loss of cell-cell junctions, downregulation of epithelial markers like E-cadherin, upregulation of mesenchymal markers like vimentin, and activated cytoskeletal reorganization [Jiang et al., 2018, Wei et al., 2015].

The ECM characteristics are also influential on vascularization, intravasation, and extravasation. Highly dense ECM is often a promoter of hypoxic conditions within the TME. This can cause the release of VEGFs that call for the endothelial cells. These cells then form new blood vessels which contribute to tumor growth and easier access for the tumor cells to the circulation [Liu and Agarwal, 2010]. Tumor cells need to pass through the BM by digestion with MMPs during intravasation and extravasation and some of the BM components like laminin were shown to be important in endothelial transmigration [Chen et al., 2016].

The ECM also takes part in the establishment of a hospitable site for newly arrived cancer cells by forming the premetastatic niche. One of the main players in ECM remodeling at this stage is the Bone marrow-derived cells. They release MMP inhibitors which in return causes accumulation of collagens. Once the cancer cells are seeded to the prematastatic niche a dormant state is observed. During this state, cancer cells are not proliferative and can avoid immune surveillance by MHC-1 downregulation. The switch from dormant to growth can be driven by integrin and DDR1 signaling which are direct targets of the ECM [Amos and Choi, 2021].

Role of the ECM in Cancer Therapy

The aberrant form of the ECM in the TME influences response to therapy. Excessive accumulation of the ECM molecules impedes the diffusion of drugs, acting as a barrier. Similarly, nutrients and oxygen are also not permitted through the thick ECM as efficiently. This metabolic stress and hypoxia cause increased expression of drug efflux pumps like ATP-binding cassette (ABC) transporters and impaired apoptosis. ECM-cell interactions via integrins and activation of FAK signaling are also shown to support the survival of the cancer cells and help avoid cell cycle arrest. EMT is not only relevant to metastasis but is also linked to chemoresistance. Transitioning cells express ABC transporters at higher levels and increase some cellular mechanisms for detoxification [Henke et al., 2020].

The ECM not only influences the efficiency of chemotherapy but also radiotherapy. Studies have shown that some ECM molecules like fibronectins and ECM receptors like integrin $\beta 1$ were important in developing resistance. Blocking their function and downstream PI3K/Akt signaling presented a better outcome of radiotherapy [Cordes and Beinke, 2004]. Oxygenation of the tumor is also important for the ionizing radiation effect on cells and hypoxic condition is known to reduce the efficiency. Pretreatment of murine tumors with anti-angiogenic drugs showed improved oxygenation and resulted in better response to radiotherapy [Matsumoto et al., 2011].

Immunotherapy refers to a range of treatments designed to activate the patient's immune system to combat cancer. The ECM can intervene with the immunotherapeutic concepts in a few ways. For a functioning immunotherapy, immune cells should be able to infiltrate the TME, yet, the dense form of ECM diverts the immune cells along the surface of the TME rather than following the cytokine gradient. Similar to chemotherapeutic agents, immune checkpoint inhibitory drugs are also not permitted efficiently through the TME. An increase in metabolic stress and hypoxia causes elevated expression in immunosuppressive factors like TGFβ, CCL18, IL-10, and CCL22 [Schaaf et al., 2018]. These factors cause an increased population of regulatory T cells and M2 macrophages while decreasing natural killer (NK)

cells and CD8+ cytotoxic lymphocytes that fight off cancer. This hypoxia was also shown to result in reduced expression of VCAM1, ICAM1, and ICAM2 which are cell surface glycoproteins expressed in endothelial cells that aid attachment and extravasation of the cytotoxic T cells. Downregulation of these molecules impedes with infiltration of these cells to the tumor site [Henke et al., 2020].

2.1.3 Therapeutic Strategies for Modulating the ECM in Cancer

The ECM shows distinct alterations within the TME and its central role in many cellular processes makes it a subject of prognostic value. Therefore there exists some therapeutic approaches used in combination with cancer drugs that target the ECM directly or indirectly.

Targeting ECM Components

Due to their high prevalence and tumor-promoting properties, fibrillar collagens and HA are appealing targets for therapy. Collagenases and hyaluronidases degrade these molecules and manipulate matrix stiffness, enabling a more efficient drug delivery process. Yet, this could also cause the release of cytokines and growth factors embedded in the collagens, stimulating tumor growth and inflammation. Secondly, it may facilitate the migration of the cancer cells [Fang et al., 2013, Parks et al., 2004].

Some molecules play a key role in ECM deposition, making them good targets for therapeutics. TGFβ is one such molecule and the use of its inhibitor along with the cancer drug doxorubicin (DOX) in breast carcinomas showed better distribution of the drug [Liu et al., 2012a]. One other such molecule is HIF-1α and silencing of it with siRNAs along with DOX in prostate xenografts improved delivery of the drug [Liu et al., 2012b].

ECM remodeling enzymes like proteases and crosslinkers are also therapeutic targets. MMP inhibitors showed reduced tumor burden and prevented metastasis in preclinical models yet tuning between efficacy and toxicity of the inhibitors makes it difficult to be approved for clinical use [Heath et al., 2006]. LOX inhibitors also

show enhanced drug delivery yet their use in already matured ECM could limit their application [Huang et al., 2021].

Targeting Integrins

Integrin is one of the central cell surface receptors that interact with the ECM. It's involved in many signaling pathways like PI3K/AKT, FAK, SRC kinases, MAP kinase signaling, and nuclear factor κB (NFκB) signaling. These pathways are important in processes like cell survival, tumorigenesis, proliferation, angiogenesis, and inflammation. In addition, they are expressed in fibroblasts, endothelial cells, tumor-associated macrophages, and more importantly, tumor cells. This makes integrins a promising target for drugs aiming for the ECM. Issues faced in integrin-targeting drugs revolve around the efficacy, pharmacodynamics, and specificity of the drugs. Integrins are expressed quite heterogeneously and have the dual role of intracellular pathways and cell adhesion. Therefore, rather than inhibition, labeling and destroying cells that highly express integrins can be beneficial. CEND-1, ProAgio, and 7HP349 are a few drug molecules that target integrins and their trials are ongoing [Sleeboom et al., 2024].

Targeting CAFs

CAFs, as mentioned earlier are the main source of ECM deposition in the TME. They can originate from bone-marrow-derived mesenchymal stem cells, metabolically inactive fibroblasts, and mesenchymal stem cells of the adipose tissue. CAFs are highly heterogeneous yet preserve some common characteristics like high mobility, invasiveness, and promotion of immune response and proliferation. The presence and infiltration of CAFs into the TME further deteriorate the ECM properties and cause poor response to drugs. Initial CAF depletion strategies included targeting the CAF markers, fibroblast activation protein, and α-SMA yet this did not contribute to the poor progression of the disease [Chung et al., 2020]. So, reprogramming of CAFs to a healthier state rather than depleting them was considered. Calcipotriol and paricalcitol are vitamin D analogs that follow this strategy[Sherman et al., 2014].

2.2 Network Biology

2.2.1 Multi-Omic Data Integration with Network Modeling

A network or a graph is a collection of interconnected entities. These entities are referred to as nodes or vertices and the connection between these nodes is referred to as edges or links. Real-life concepts like social connections, transportation, and financial systems can all be represented by networks. In the context of biological systems, nodes can represent, genes, RNA molecules, amino acids, proteins, metabolites, drugs, pathways, cells, and tissues while the edges can show functional, physical, or chemical interactions between them. With such representations, one can, for example, model how a group of cells communicate with each other how a gene is expressed or suppressed how a protein is modified or not, and how this alters the downstream processes or shows the synergy between different pathways. To better represent the complex nature of biological systems network biology combines biological sciences with computational concepts like network science, data mining, graph theory, and machine learning [Barabási and Pósfai, 2016].

The major type of data used in computational biology is omics data. Omics data is the data generated from high-throughput technologies of various omics fields that represent different layers of the biological system. Each of these layers carries particular intel on the state of the biological system. Genomics revolves around the DNA and the entirety of an organism's genes. Whole genome sequencing (WGS) and analysis enable one to identify genetic variations or evolutionary patterns for a given organism. Epigenomics investigates the reversible modifications of the DNA and histone proteins that alter DNA accessibility and expression levels of genes. Transcriptomics is the study involved with the RNA transcripts and quantification of the mRNA helps to understand gene regulation patterns. Proteomics includes the identification and quantification of the proteins within the cell as well as their post-translational modifications (PTMs). This can help researchers to understand which proteins are more abundant or have altered states in specific conditions. Finally, metabolomics is the omic data that involves determining the amounts of small molecules produced within the cells like amino acids, sugars, lipids, and nucleotides.

This data can hint at the metabolic processes the cell undergoes in different conditions [Misra et al., 2019, Ning and Li, 2023, Santiago-Rodriguez and Hollister, 2021]. Other data types that can also be implemented to support these molecular data include screening methods like imaging data, fluorescence-activated cell sorting/cytometry, or clinical data [Tarazona et al., 2021].

Omic data provide detailed and comprehensive information at various molecular levels, but they often appear to be independent of each other. However, in biological networks, different omic layers coexist within cells and constantly interact. For instance, a TF, identifiable from proteomic data, can become activated, which is observable through phosphoproteomic data. This activated TF can then bind to DNA, and if the chromatin is accessible, a state inferred from epigenomic data, it can regulate the expression of a target gene. The resulting gene regulation can subsequently be detected through transcriptomic data yet translation of the transcript can be shown again with proteomic data. Therefore, it can be said that these omic data are complementary to each other, and act together to form functioning phenotypes. To reveal such causality between omic layers and model the reality within biological systems in a holistic manner, integrative methodologies are required and for this, network-based methods are extensively used [Ivanisevic and Sewduth, 2023].

Modeling realistic biological networks involves utilizing omics hits related to a specific context, such as a particular disease or tissue, and mapping it to a reference. These omic hits can be differentially expressed genes (DEGs), enriched TFs, highly mutated genes, and abundant proteins or phosphoproteins. The reference interactome may have protein-protein interaction (PPI) data like Human Integrated Protein-Protein Interaction Reference (HIPPIE) and STRING, gene regulatory interaction data like Transcriptional Regulatory Relationships Unraveled by Sentence-based Text Mining (TRRUST), Discriminant Regulon Expression Analysis (DOROTHEA) and Virtual Inference of Protein Activity by Enriched Regulon Analysis (VIPER), and PTM data like PhosphoSitePlus and The human Dephosphorylation Database (DEPOD) [Ideker et al., 2002, Alanis-Lobato et al., 2016, Szklarczyk et al., 2022, Han et al., 2017, Alvarez et al., 2016, Garcia-Alonso et al., 2019, Hornbeck et al., 2014, Damle and Köhn, 2019]. The interactions from these references

may be derived from high-throughput experiments, conserved co-expression data, predictions, automated text mining, and previous knowledge in databases and almost all have an effort to assign a confidence value to the interactions to manage false negatives and positives. These values may be determined by how many references the interactions have or how reliable the detection method is [Kamburov et al., 2012].

Network modeling algorithms need to evaluate several factors during optimization. Directly mapping omics hits to the interactome is not ideal, as it would result in a crowded, uninterpretable hairball-like structure. Similarly, connecting only the first neighbors of the hits is insufficient, as it would yield an incomplete network. The algorithm should propagate from the omics hits by selecting the most reliable interactions while accounting for bias toward hub proteins that have many confident edges within the interactome because they are extensively studied. Some tools manage this bias by penalizing these hub proteins [Demirel et al., 2022].

Various network modeling tools work with omic data and they have different propagation methods. MEXCOwalk utilizes a random walk that involves moving from one node to another neighboring node in a random manner with transition probabilities [Ahmed et al., 2019]. RWRNET which uses random walk with restart that is similar to random walk yet with an additional probability of returning to the starting node [Liu et al., 2020]. HotNet works with heat diffusion where propagation from the seed nodes with some "heat" value indicating its importance is carried out until there is no significant change in the distribution of the heat value and an equilibrium is reached [Leiserson et al., 2014]. Pathlinker works by calculating multiple short paths from receptors to TFs to give signaling pathways [Ritz et al., 2016]. DOMINO performs an empirical evaluation of Gene Ontology (GO) terms called by active module identification algorithms with the integration of permutation-based methods to get subnetworks with significant activity that represents biological processes [Levi et al., 2021]. Omics Integrator which uses the prize-collecting Steiner forest (PCSF) algorithm [Tuncbag et al., 2016].

2.2.2 Types of Biological Networks

Biological systems are not of a single layer but are multi-layered and each layer has its sort of data that implies a different aspect of the system. Therefore there are different types of network representations. The main types include PPIs, transcriptional regulatory networks (TRNs), cell signaling, and metabolic networks. Other networks include correlation networks which are undirected networks that show if the expression or abundance of a gene or a protein pair is correlated across samples and genetic interaction networks that help to investigate whether simultaneous mutations in multiple genes have a particular phenotypic or functional outcome [Boucher and Jenna, 2013].

Protein Protein Interaction Networks

In PPIs, nodes represent the proteins and the edges represent the interactions between these proteins. These interactions can be direct physical contact between the proteins and this data can be collected with methods like yeast two-hybrid screening, mass-spectrometry, and co-immunoprecipitation but also can be modeled with docking algorithms. These physical links are specific to the protein pair have a biological meaning and need not be permanent. They can be stable bounds or transient interactions generally depictive of a modification or transport of a protein. The link between the proteins can also be representative of the functional role of interacting proteins as a collective. With this, proteins that for example take part in a specific pathway but don't show direct contact, can be connected. The entirety of PPIs in a biological system is called the interactome and although the advanced high-throughput screening technologies provide a large amount of PPI data, the interactome is still noisy and incomplete and contains true negatives and positives [Rual et al., 2005].

Transcriptional Regulatory Networks

TRNs help us to investigate the gene expression patterns in a particular condition. TFs, regulatory RNAs like miRNAs, target genes, and regulatory elements like pro-

moters, and enhancers can be represented by nodes. The relationship between the regulators and targets can be depicted by directional edges which indicate suppression or activation of a gene. TRNs can describe the causality between biomolecules and are generally the last step of a cell signaling network. Cell signaling networks are representative of the cellular communication and transmittance of cellular messages from outside the cell toward the inside. This information flow within the cell can be shown as directed edges where the nodes can depict any element from a particular pathway like a protein or a metabolite. Cells receive signals from other cells or the ECM as ligands via cell surface receptors. These receptors then transfer the message with a cascade of events toward the nucleus, meaning that the message may trigger the activation or deactivation of particular TFs. These TFs would then reach to their targets and ultimately switch the transcriptional program of the cells. This flow of the message can be modeled with signaling networks and the response generated from the message can be modeled with TRNs [Costa et al., 2008].

Metabolic Networks

Metabolic networks represent the metabolic activities within the cells that keep the cell functioning in a particular way. These metabolic activities revolve around cellular processes like reproduction, growth, migration, and response to external stimuli. Nodes depict the metabolites and enzymes that take part in biochemical reactions and edges are generally directed although in principle all of the reactions are reversible. These networks model how particular biomolecules like sugars, amino acids, and lipids come together to form cellular structures and how energy is generated from the breakdown of these molecules. Substrates are converted to products which can also be the substrates of another reaction. So, these networks contain some motifs like chains or cycles. The main players in this conversion are the enzymes since they can accelerate the reactions and influence the rate of the metabolic flow. Regulation of a metabolic network is also controlled by the concentration of substrates and products. By modeling the metabolism, one can investigate the systemic behavior of cells to altered conditions [Dräger and Planatscher, 2013].

2.2.3 Applications of Network Modeling in Medicine

Network modeling enables researchers to gain a holistic understanding of the processes occurring in biological systems. One of the main applications of network biology involves identifying disease-associated subnetworks or disease modules. Alterations in omic layers upon a diseased condition provide hits to researchers to start investigating the disease yet in singularity a gene or a protein is not enough for an explanation of the disease. Disease modules are a collection of genes, proteins, or metabolites that have a role in a particular cellular function and an abnormality in this module results in a diseased phenotype. These modules are important in understanding the mechanism of the disease and can be used to address various biological questions, such as identifying biomarkers, discovering drug targets, or stratifying patients [Barabási et al., 2010].

Biomarker Identification

Biomarkers are important indicators of disease outcomes. They aid in the prediction of the overall survival of patients, the progression of the disease, or whether the patient may be responsive or resistant to a particular drug. Network biology helps detect novel biomarkers. With multi-omic data integration and rewiring of omic hit or already known biomarkers, missing pieces of an essential disease pathway can be revealed. One can identify key motifs or nodes by performing a topological analysis of the disease module or comparing them to healthy control. Machine learning approaches applied to these modules can also reveal patterns that can show potential biomarkers [Ben-Hamo and Efroni, 2013].

Drug Discovery and Repurposing

Drug discovery processes aim to identify targets and medication for the target yet traditional ways of drug discovery are costly and time-consuming. Better means of identifying potential drug targets can significantly reduce time and cost. In addition to discovery steps, trials of drug molecules also take a lot of time. Therefore, drug repurposing becomes a beneficial strategy that involves finding different uses for

already approved drugs. Similarly to biomarker identification, network modeling can be very useful in this field. In addition to investigating the topology of the networks as a static entity, simulating changes in the networks with in-silico knock-out or overexpression of particular genes can be informative of drug targets [Xie et al., 2023]. Or, calculating how closely oriented the disease module is to known drug targets can be useful for drug repurposing [Guney et al., 2016].

Patient Stratification

There are diseases like cancer that do not show the exact same features and outcomes in every individual. Patients themselves are already significantly different from each other and constructing patient-specific networks from patient-driven data is one of the best ways to represent a patient and the patient's disease state.

There are publicly available patient data cohorts from various institutions. The National Cancer Institute (NCI) is the main cancer research institute of the United States of America and one of its objectives is to provide patient-driven data to the cancer research community. It contains two main branches of data repository called, Proteomic Data Commons (PDC) and Genomic Data Commons (GDC) but it also has a cancer medical image archive that contains magnetic resonance images, computed tomography scans, and, histopathological images. These repositories are fed with data collected from several programs. PDC mainly houses data from programs like the Clinical Proteomic Tumor Analysis Consortium (CPTAC) and Applied Proteogenomics Organizational Learning and Outcomes (APOLLO) which mainly focus on proteogenic data. GDC mainly holds data from cancer genome programs like The Cancer Genome Atlas (TCGA) and Therapeutically Applicable Research to Generate Effective Treatments (TARGET). These repositories collectively include various types of data such as gene expression, protein abundance and modification, metabolomic, DNA methylation, and mutation data of tissue samples taken from patients. These samples can be from the primary tumors, metastatic sites, and normal adjacent tissues. Patients in the studies are not of a single cancer type. Currently, there are 12 major primary sites in PDC with the ovary hav-

ing the most cases and 69 in GDC with bone marrow and blood having the most cases. In addition to omic data, these programs collect clinical data of the patients making it possible to correlate molecular changes in disease condition to clinical data [Thangudu et al., 2020, Grossman et al., 2016]. There are other institutes like the American Association for Cancer Research (AACR) with a clinical-genomic data program called Genomics Evidence Neoplasia Information Exchange (GENIE) and UK Biobank which provides large-scale biomedical data from UK participants [André et al., 2017, Sudlow et al., 2015].

In addition to differences between individual patients, tumor tissues also show high heterogeneity and it becomes more and more difficult to generalize a singular diagnostic conclusion and more importantly treatment method for patients with a particular cancer. Network modeling using patient-derived multi-omic data creates a digital twin of the patients. These digital representations then can be used to understand different cellular states and molecular patterns one may observe in different patient groups and guide researchers in developing personalized medicine strategies [Zitnik et al., 2023].

Chapter 3

MATERIALS AND METHODS

3.1 Overview of the Method

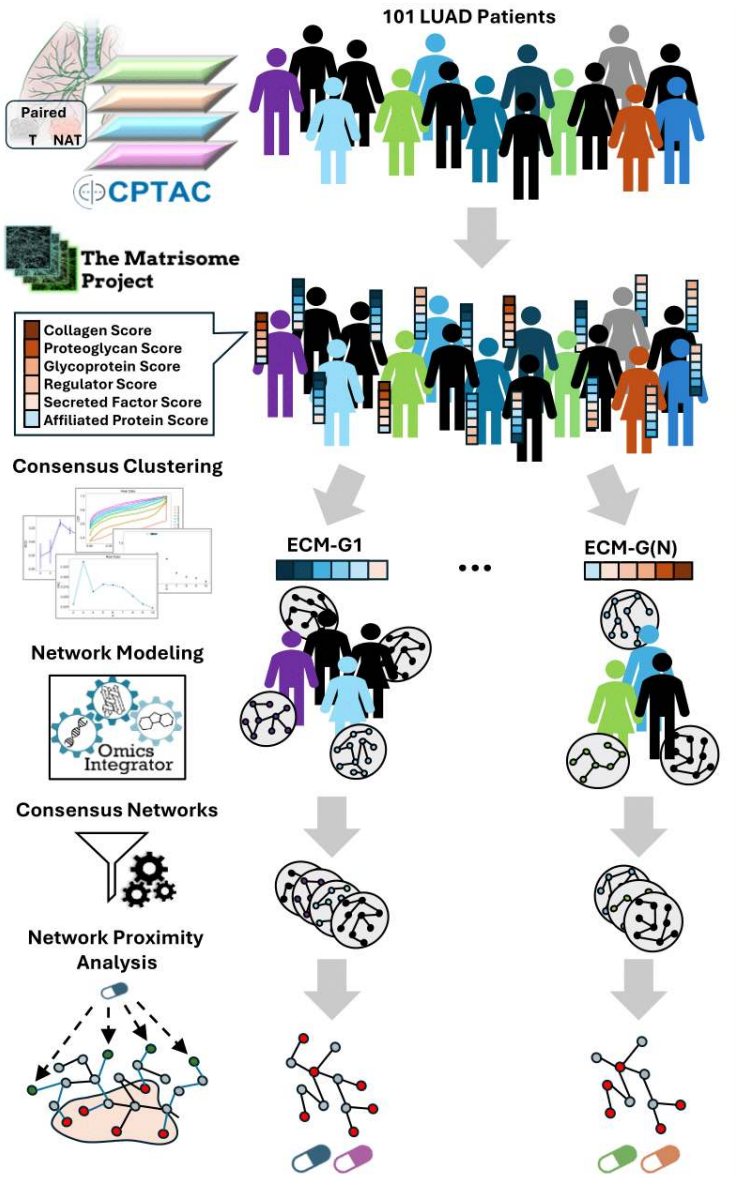


Figure 3.1: Overview of the Method

This study aims to stratify LUAD patients based on their ECM and define the features of these different ECM profiles. This was done by first calculating multi-omic ECM barcodes for every patient using a patient's tumor and normal tissues. Then these barcodes were used for consensus clustering of patients. Once the clusters were formed, they were subjected to comparative analyses to understand the differences and similarities these groups may have. These analyses include cellular heterogeneity, correlation, and clinical data analysis. This was followed by network modeling of every patient for multi-omic data integration, revealing the underlying molecular mechanisms for altered ECM expression. Enrichment analyses were done on these networks to show enrichment patterns in pathways and hallmarks. Then, these networks were utilized for constructing ECM-specific consensus networks, and these networks were subjected to network proximity analysis to identify potential drugs that could be more effective on particular ECM groups. The overview of the method can be seen in Figure 3.1. Some of the icons in this figure were obtained from BioRender.

3.2 Patient Cohort and Datasets

3.2.1 CPTAC

The dataset used in this study was retrieved from the PDC CPTAC-3 project. This dataset includes multi-omic data from LUAD patients. Genomic data is generated with whole genome sequencing and whole exome sequencing and provides germline mutation, somatic mutation, and copy number alteration data. Epigenomic data is generated with a methylation array and provides DNA methylation data. Transcriptomic data is generated with RNA-seq and miRNA-seq and provides mRNA and miRNA quantification data. Proteomic data is generated with mass spectrometry and provides protein abundance, phosphorylation, and acetylation data. The sample size of the dataset is 211 and 111 of them are tumor samples while 101 are matched normal adjacent tissue (NAT). Mainly, 101 patients with tumors and NAT samples were used throughout this study. Clinical data is also present for each patient such as age, sex, ethnicity, smoking status, stage, and body mass index.

The main omic data used in this study are mutation, gene expression, protein abundance, and phosphorylation data. The mutation data includes the mutation status of cancer driver genes like TP53, KRAS, and EGFR and fusion data on genes like ALK, ROS1, and RET for all of the patients. Gene expression data is present in two forms. The first contains un-centered normalized data for tumor and NAT samples where counts of protein-coding genes are first upper-quartile normalized and normalized based on library depth and gene length to get reads per kilobase million (RPKM) values. These values are then log2 transformed. The second contains gene-centered counts where the median of each gene across the entire data is subtracted from the log2(RPKM) values. Both transcription data contain 18,099 features. Protein abundance and phosphorylation data contains log2 transformed tandem mass tag (TMT) ratios to a comparative reference sample which is then centered by subtracting the mean protein abundance from the log2(TMT) values and scaled by dividing by the standard deviation. The protein abundance data contains 10,699 features and the phosphorylation data contains 41,188 features [Gillette et al., 2020].

A tumor z-score data was generated from the un-centered gene expression data to capture the expression trends among tumor samples by calculating gene-wise z-scores specifically for the tumor samples using Equation 3.1. For gene-wise centering, the mean expression value μ of each gene across the tumor samples was subtracted from each expression value x. This was followed by dividing the resulting values by the standard deviation σ for scaling. This processed data was later used in various analyses throughout the study.

$$z = \frac{x - \mu}{\sigma} \tag{3.1}$$

3.2.2 The Matrisome Project

The Matrisome Project is a resource that provides ECM-related data. By nature, the ECM proteins are complex, large, and from insoluble matrices. In addition, ECM proteins contain multiple domains that are shared with other ECM and non-ECM proteins. This makes it difficult to come up with a complete list of ECM proteins and already existing information on ECM proteins does not represent the

true categorization of these molecules and is inconsistent. The Matrisome Project aims to identify genes that encode ECM proteins and factors associated with them and the entirety of these components is referred to as the "matrisome". The matrisome consists of 1027 proteins from two divisions which are the core matrisome and associated proteins called matrisome-associated proteins. The core matrisome includes the most well-known and accepted ECM proteins along with some other proteins shown to exhibit the same structural features of ECM proteins. Core matrisome holds 274 proteins of three main categories which are ECM glycoproteins with 195 proteins, collagens with 44 proteins, and proteoglycans with 35 proteins. Matrisome-associated proteins include proteins that take part in the remodeling of the ECM, frequently bind to ECM, or are shown to be correlated or coexisting with other ECM proteins. Matrisome-associated proteins hold 753 proteins of three main categories which are ECM Regulators, such as crosslinking enzymes or proteases, with 238 proteins, Secreted Factors, such as cytokines or growth factors, with 344 proteins, and ECM-affiliated proteins, such as mucins or galectins, with 171 proteins [Naba et al., 2012a, Naba et al., 2012b].

3.2.3 TRRUST

TRRUST is a TRN database that contains information on TF-target genes of humans and mice. The data is achieved by finding relevant sentences to TRNs from 20 million articles from Pubmed with text mining algorithms followed by manual curation where the hits are reviewed and verified by experts. Currently, the database holds 8444 regulatory relationship data for 800 TFs in humans and 6552 regulatory relationship data for 828 in mice, derived from 11,237 Pubmed articles. Around 59.8% of the regulatory interactions also contain information on the regulation mode that is activation or regression. It is possible to perform queries on the web interface with two options. One can perform a single gene search to identify its regulators, mode of regulation, and references. If the query gene is a TF it also identifies its target genes, other TFs that share the same targets as the query, diseases, pathways, and GO Terms, associated with the TF along with simple subnetworks from the PPI

network and functional network references. The second enables one to query a set of genes and returns a list of candidate key regulators using Fisher's exact test. The output gives the number and list of overlapping genes between the query set and the targets of the TF along with P and Q values that indicate how significant this overlap is [Han et al., 2017].

3.2.4 DRUGBANK

DrugBank is a database that contains drug and drug target information. This data can be collected from the manual curation of scientific literature by experts, and other databases like PubChem and Uniprot. It also collects data from organizations like Genome Canada and The Metabolomics Innovation Centre and regulatory agents like the U.S. Food and Drug Administration (FDA). Total number of drugs is 16,911. There are 2785 approved small molecules, 135 nutraceutical substances, and 1628 approved biotech drugs which include vaccines, allergenics, proteins, and peptides. There are also 6722 experimental drug molecules as well as withdrawn and illicit drugs. The database provides 5293 proteins which include drug targets and enzymes, transporters, and carriers associated with the drug. Protein sequences in FASTA format and drug structures in SDF format can be found in the database and external links to enable cross-referencing. This database can be used for discovering drug targets and predicting drug interactions and metabolism. It can also be used for virtual screening by molecular docking to get hits for drug discovery, repurposing, and design [Wishart et al., 2017, DrugBank, 2024a].

3.2.5 iRefIndex

iRefIndex is an interactome that provides physical PPI data where the proteins are in physical contact with each other. Proteins involved in the interaction, the type of experiment or method that identified the physical interaction, and the publication that reports the interaction are all shared in this database and it holds interaction data collected from public databases like BioGrid, CORUM, IntAct, and InnateDB [Oughtred et al., 2020, Ruepp et al., 2007, Orchard et al., 2013, Breuer et al., 2012].

The main goal of this database is to construct a unifying index for PPIs since there are multiple interaction data available in different PPI databases for a particular protein. With indexing, redundancy in interactions between protein pairs can be eliminated. The interactions between the proteins are scored with MI score which considers the number of publications that call the interaction, reproducibility of the score, and the reliability of the experimental detection method for the interaction and is an indicator of the confidence of the physical interaction between the proteins. The version used in this study is v14 and it includes 182,002 interactions of 15,759 proteins [Razick et al., 2008, VIB, 2023].

3.3 Barcoding and Clustering of Patients

3.3.1 ECM Barcoding

Patients were barcoded based on their ECM categories: collagens, proteoglycans, glycoproteins, regulators, secreted factors, and affiliated genes. For each ECM category, a transcriptomic, proteomic, and phosphoproteomic score is calculated and integrated to obtain a final multi-omic ECM score. This score calculation is referred to as the "barcoding" of the patients.

First, gene-centered transcriptomic data is filtered to include only the matrisome genes. Then, the expression level of NAT samples is subtracted from the expression level of tumor samples for each patient. This calculation provides a value representing the upregulation or downregulation of an ECM gene in the patient's tumor sample with respect to the patient's NAT sample. Then genes having a difference value below -2 or above 2 are retained to get highly dysregulated ECM genes while removing those that do not show considerable change in their regulation. These highly dysregulated genes are then grouped by their ECM categories. Afterward, a single transcriptomic score for each ECM category is calculated by taking the average of the difference values of the highly dysregulated genes within that category. This score represents the overall expression level change for the genes of the ECM category in a patient's tumor.

Proteomic data is processed with the same procedure. Protein-centered abundance data is filtered to include only matrisome proteins and the difference between tumor and NAT sample protein abundance is calculated for each patient. Using the same thresholds, highly dysregulated ECM proteins are identified and grouped based on their ECM categories. By taking the average of these protein subsets, a proteomic score is obtained for each category. This score gives the overall change in protein abundance of an ECM category in the patient's tumor.

Phosphoproteomic score calculation is performed similarly to the proteomic and transcriptomic ECM scores. The difference in phosphoprotein abundance of the ECM proteins in tumor and NAT samples is calculated, filtered using the same thresholds, and then grouped by ECM categories. The average is taken to get a phosphoproteomic score for each ECM category, indicating how the proteins of a particular ECM category show elevated or reduced abundance in their phosphorylated form in a patient's tumor.

Finally, transcriptomic, proteomic, and phosphoproteomic ECM scores are combined by taking the average to get a multi-omic ECM score. This score indicates the increase or decrease in gene expression, protein abundance, and phosphorylated protein abundance of the ECM categories in a patient-specific manner. These barcodes are unique for each patient and provide a holistic view of the regulation profile of the ECM components.

3.3.2 Consensus Clustering

Patients were clustered based on their ECM profiles using the barcodes calculated. For this, a package called Monte Carlo reference-based consensus clustering (M3C) is used by setting the objective function to the proportion of ambiguous clustering (PAC), the clustering algorithm to k-means clustering, and the rest of the arguments to default.

This tool makes use of Monte Carlo simulations to construct several references to be used to calculate some metrics to ultimately give an optimal number of clusters K. M3C starts by generating multiple reference datasets that preserve the covariation

in features within the actual dataset via principal component analysis (PCA) that only form one Gaussian cluster. This gives the basis for the null hypothesis that "the dataset does not form any clusters". Then, the Monti et.al consensus clustering is performed on these references, and a PAC score is calculated for a range of K values. PAC score gives a measure of the proportion of data pairs that show uncertain clustering behaviour over all data pairs and the smaller it is the more confident the clustering is for the particular K value. The same consensus clustering and PAC score calculation is also done on the original dataset for a range of K values. Then, for each K value, a relative cluster stability index (RSCI) is calculated using the PAC values of the generated references and the PAC value of the real dataset. This RSCI evaluates the robustness and reproducibility of the clusters and the higher it is the more stable the clusters are for the particular K value. Finally, a Monte Carlo p-value is calculated simply by measuring the fraction of reference PAC scores that are less than or equal to the PAC score of the real dataset which is again calculated for a range of K values. A Monte-Carlo p-value smaller than 0.05 for a K value means that using that K value for clustering results in the rejection of the null hypothesis and states that the dataset does not form a single Gaussian cluster. It suggests that the data shows significantly less ambiguity with K number of clusters than with a singular cluster [John et al., 2020, Şenbabaoğlu et al., 2014, Monti, 2003].

3.3.3 Highly Variable Gene Identification

To observe genes distinctive of the ECM clusters and show that these clusters indeed represent ECM characteristics of clusters, genes with high variability were identified. Tumor z-scores were used to identify highly variable genes among ECM clusters. For each cluster, the number of genes with a z-score above 1 or below -1 was counted separately to account for the higher and lower expression levels, respectively. Genes with count values greater than or equal to 13, except for one cluster where the threshold is 8, in either high or low expression were determined as highly variable genes.

3.4 Cellular Heterogeneity

For detecting cellular heterogeneity, R package immundeconve was used [Sturm et al., 2020]. This package provides a unified platform for detecting various cell populations from bulk RNA-Seq data by combining multiple cellular deconvolution tools. This tool works with transcripts per kilobase million (TPM) normalized data therefore, the expression data was transformed from RPKM to TPM form with Equation 3.2.

$$TPM = 10^6 * \frac{RPKM}{\sum(RPKM)} \tag{3.2}$$

The main cell fraction detectors used in this study were MCP-Counter and EPIC. Both tools detect various immune cells like T-cells, B-cells, macrophages, and NK cells. They also detect endothelial cells and CAF content. MCP-counter simply works by evaluating the expression profiles of marker genes in a particular cell type. MCP-counter takes the average of the expression levels of the well-defined cell markers and calculates an abundance score. This is followed by normalizing these scores so that the cellular fractions are comparable between samples [Becht et al., 2016].

3.5 Correlation Analysis

Cancer-associated gene groups for calculating, EMT, CSC, invasiveness, and integrin scores were retrieved from Kuşoğlu et al. as well as the methodology which is given in Equation 3.3 where i is the gene set, j is the patient, n is the number of genes in the set and, z_k is the z-score of the gene k. In this study, these scores were calculated using tumor z-score data [Jethalia et al., 2023, Kuşoğlu et al., 2024].

$$score(i, j) = \frac{1}{n} \sum_{k=1}^{n} z_k \tag{3.3}$$

Estimation of STromal and Immune cells in MAlignant Tumor tissues using Expression (ESTIMATE) data is a tool that is used to detect tumor purity which is the proportion of tumor cells to stromal and immune cells within the tumors. This tool utilizes gene expression data to detect stromal and immune cell content using ssGSEA. ssGSEA ranks the genes according to their expression and scans

this the ranked list for genes of a specific gene set. While scanning, it calculates a running sum that increases each time the algorithm encounters a gene from the gene set while decreasing every time it doesn't. This gives an enrichment score for the given gene set which is later scaled for comparability. ESTIMATE calculates such a score for capturing stromal cell and immune cell presence within the tumors named stromal score and immune score respectively. The sum of these two scores referred to as the ESTIMATE score represents the enrichment of the portion of the tumor that is not composed of tumor cells and infers tumor purity [Yoshihara et al., 2013]. ESTIMATE score, immune score, and stromal scores were provided in the CPTAC clinical data.

EMT, CSC, invasiveness, integrin, ESTIMATE, immune, and stromal scores were all subjected to correlation analysis with the components of the ECM barcodes, which are proteoglycans, collagens, glycoproteins, secreted factors, regulators, and affiliated proteins scores using spearman rank test.

3.6 Clinical Data Analysis

Survival and recurrence-free analysis were performed on ECM clusters using the survival package, and the log-rank test was done to calculate p-values [Terry M. Therneau and Patricia M. Grambsch, 2000]. Short-term survival and recurrence-free analyses were performed within the first 600 and 300 days, respectively. Patients with events that occurred after these time points were assumed to be survivors or recurrent-free until that time point. In addition to the log-rank test, a multivariate COX proportional hazards model was employed to calculate hazard ratios (HR) by taking into account age, sex, smoking status, and ECM grade.

The status of ECM clusters on clinical features, TCGA mRNA subtype, CpG Island Methylator Phenotype (CIMP), and tumor grade were also investigated. RNA subtype is a categorization system accomplished by analyzing RNA-seq data, which provides information on the expression profiles of genes related to defined cellular processes. There are three RNA subtypes: proximal inflammatory (PI), which correlates with increased inflammatory response and high expression of immune-

related genes; proximal proliferative (PP), which correlates with elevated cell proliferation and expression of genes related to the cell cycle; and terminal respiratory unit (TRU), which is a subtype that resembles the normal tissues the most, and genes expressed in functional tissues are also expressed at the same level for the samples within this subtype. CIMP status provides information about the epigenetic phenotype of tumors. It correlates with the extent of methylation within the CpG islands, which may result in the silencing of tumor suppressors. There are four CIMP subgroups, each representing different levels of methylation. CIMP+ tumors show very high methylation of promoters. The CIMP-1 or CIMP-High subtype indicates high methylation of the DNA but not as much as CIMP+ and is generally correlated with a poor prognosis. CIMP-2 or CIMP-Low tumors show moderate methylation, and the CIMP-subtype shows negligible CpG island methylation. Tumor grade shows the degree of differentiation of healthy tissue into a malignant form using histological evaluation under microscopy. There are four tumor grades. Grade 1 indicates tumor cells that are not very aggressive and show slow growth and spreading. Grade 2 represents tumor cells that have moderate spread and growth. Grade 3 tumors show cells with aggressive behavior, and this grade is highly correlated with a poor prognosis. Grade X represents tumors with an undetermined grade.

3.7 Patient-Specific Network Reconstruction

3.7.1 OmicsIntegrator2

A network representation of the patients was done with the forest module of OmicsIntegrator2. OmicsIntegrator2 requires two inputs. The first is a list of proteins of interest, referred to as the terminals, with some value indicating the importance of the proteins, referred to as prizes. These terminals can be any omics hits, for example, DEGs, proteins, or highly phosphorylated proteins. The prizes can be fold-change values or p-values and must be positive numbers. The second input is the reference interactome with confidence values assigned to edges.

OmicsIntegrator2 maps the terminals to the reference interactome and tries to connect those terminals using a form of the PCSF algorithm, which originates from

the prize-collecting Steiner tree (PCST) problem. The solution for PCST involves collecting prizes by connecting as many terminals as possible while paying the lowest possible cost. This may introduce intermediate nodes, referred to as Steiner nodes, coming from the reference interactome but not the terminal list. This results in a connected subnetwork that doesn't have any cycles, called a tree structure. PCST gives a single tree structure, while PCSF solves for multiple disjoint trees, which gives a forest. This helps to preserve terminals that could not be included in one tree but form their tree or trees within the solution. Since biological systems can show distinct and separate biological processes and signaling pathways upon perturbation, this algorithm enables the simultaneous reconstruction of these concepts.

Let $G(V, E, c(e), p'(v))$ be a directed or undirected network where V is the node-set, E is the edge set, $c(e) > 0$ is the cost for each edge $e \in E$ and $p'(v)$ is the attenuated prize value for each node $v \in V$. The PCSF algorithm of OmicsIntegrator2 aims to find a forest $F(V_F, E_F)$ by minimizing the objective function in Equation 3.4.

$$f'(F) = \sum_{v \notin V_F} p'(v) + \sum_{e \in E_F} c(e) + \omega \cdot \kappa \tag{3.4}$$

The first element calculates the sum of prizes of terminals that were not included in the resulting forest. This $p'(v)$ is an attenuated prize value for the nodes. The scope of this attenuation is the elimination of bias towards hub nodes and each node is penalized for the amount of connections it has. This attenuated prize is calculated with Equation 3.5.

$$p'(v) = \beta \cdot p(v) - \mu \cdot \text{degree}(v) \tag{3.5}$$

Here, $p(v)$ is the prize for node v, degree(v) is the number of first neighbors of node v. β is a parameter for tuning for the amount of terminals to be included in the solution and μ is a parameter that adjusts the influence of the penalization for the node degree.

The second element is the sum of edge costs in the resulting forest. The edge costs $c(e)$, are calculated from the confidence values $p(e)$ with Equation 3.6.

$$c(e) = 1 - p(e) \tag{3.6}$$

The third element is the major player in the function that enables multiple trees to form and result in a forest. This algorithm introduces an artificial node connected to terminals with an edge cost ω. Then, the PCST algorithm is run on the input and a tree rooted from the artificial node is constructed. By removing all of the edges between the nodes within the solution and the artificial node, a forest is formed. κ here is the number of trees in the solution and by changing the parameter ω, it is possible to tune for the number of trees. OmicsIntegrator2 enables users to tune for six parameters in total but there are three that must be configured to get the optimal networks which are β, ω, and D. As mentioned earlier β tunes for the number of terminals to be included in the solution, and ω tunes for how many components there should be. D parameter tunes for the maximum path length between terminals and the artificial node. This influences the number of nodes in total that give the final forest. One should perform parameter tuning to choose the optimal combination of β, ω, and D so that the solution connects almost all the terminals provided with reliable edges without introducing too many Steiner nodes [Tuncbag et al., 2013, Tuncbag et al., 2016].

3.7.2 Network Modeling

Patient-specific signaling networks were constructed with terminal lists that included possible TFs that target differentially expressed ECM genes and dysregulated ECM phosphoproteins. As for the reference interactome, iRefIndex was chosen and the hub proteins determined by Beyge et al. “UBC”, “APP”, “ELAVL1”, “SUMO2“, and “CUL3” were excluded [Unsal-Beyge and Tuncbag, 2022]. The confidence values were converted to edge costs using Equation 3.6.

To determine differentially expressed ECM genes, un-centered gene expression data was used where expression of an ECM gene from a patient’s NAT sample was subtracted from the same patient’s tumor sample to calculate a log2 foldchange value. ECM genes with a log2 foldchange value higher than 2 or smaller than -2 were considered differentially expressed. This gene list was then used for TF enrichment using the TRRUST database. Among the key regulators identified, those with a

q-value below 0.1 were included in the terminal list. The prizes of the TFs were calculated by taking the average of the absolute foldchange values of the differentially expressed ECM genes they regulate. Highly dysregulated ECM phosphoproteins were obtained from the same list produced for calculating phosphoproteomic ECM scores and the highest absolute difference values were assigned as the prizes.

To obtain optimal networks, a set of values was provided for β, ω, and D while the remaining were set to default. For β, ω, and D, values [2,3,4,5] were used in combination, and 64 networks in total were constructed. Networks were then ranked based on the number of components, nodes, and terminal coverage. Those that had the minimal number of components and nodes and maximal coverage were ranked at the top and the parameter values of the best two ranking networks were selected as the optimal values. After the parameter tuning, optimal networks based on these two parameter sets were constructed and combined to give a final consensus network for each patient.

3.7.3 Network Comparison

To show that the networks of patients are representative of their ECM cluster two similarity analyses were performed. For both analyses, the Jaccard index $J(A, B)$ was calculated by dividing the intersection of the node list of patients A and B by their union as shown in Equation 3.7. A Jaccard index of 1 means identical node sets while 0 means no common nodes are present between the two node sets. This was calculated for all patient pairs to construct a 101×101 similarity matrix.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \qquad 0 \leq J(A,B) \leq 1 \tag{3.7}$$

The first analysis evaluates the similarity of patient networks within the same ECM cluster compared to random network collections. For this, a similarity score was calculated by taking the average of the Jaccard index between the patient and the members of the patient group as shown in Equation 3.8 where $Sim(P)$ is the similarity score of patient P against a group of patients G with N members.

$$Sim_P = \frac{1}{N} \sum_{n=1}^{N} J(P, G_n) \tag{3.8}$$

Two sets of similarity scores named, within-cluster, and random were calculated for the ECM clusters individually. The within-cluster similarity score set was built by calculating the similarity scores for all of the patients of an ECM cluster against the same ECM cluster the patients belong to. The random set was again built from all of the patients of an ECM cluster but this time against a randomly selected ECM group other than, but of the same size as, the patients' ECM cluster.

The second analysis involves the comparison of similarities within and between clusters. The within-cluster similarity score set was compared with the between-cluster score sets. These sets were constructed by calculating similarity scores for all of the patients of a particular ECM cluster against the other ECM clusters.

3.8 Cluster-Specific Enrichment Analyses

3.8.1 Pathway Enrichment Analysis

Cluster-specific pathway analysis was performed using the EnrichR [Chen et al., 2013]. Gene sets used in functional enrichment analysis were pathways from the Kyoto Encyclopedia of Genes and Genomes (KEGG) and hallmarks from the Molecular Signatures Database (MSigDB) [Kanehisa et al., 2022, Liberzon et al., 2015]. KEGG Pathways represent a collection of various reactions and incidences in the cells for example, the metabolism of a substance, processing of environmental or genetic information, diseases, and basic cellular processes. MSigDB Hallmarks is a collection constructed from overlaps of different gene set collections within MSigDB and contains 50 Hallmarks including biological processes or states like EMT. Functional enrichment analysis was performed using the nodes of patient-specific networks and pathways with an adjusted p-value below or equal to 0.05 and overlapping gene counts above 2 were considered enriched. These were then subjected to cluster-wise trimming where pathways enriched in more than 25% of the cluster population for at least one of the clusters were retained. Then, for each patient, a pathway enrichment score was calculated using tumor z-score data by taking the average of the z-scores of genes related to the pathways. A similar score was calculated for each cluster by averaging the pathway enrichment scores of patients of the same cluster.

3.8.2 TF Enrichment Analysis

TF enrichment analysis was performed by providing the differentially expressed ECM genes to TRRUST and statistically filtering key regulators, which were included in the terminal list as explained in the section on Network Modeling. The resulting enriched TFs were then subjected to the same cluster-wise trimming performed for enriched pathways. A TF enrichment score was calculated for each patient and each cluster, using a similar calculation done for the pathway enrichment score. In this case, the median was taken instead of the mean. Finally, the resulting list of TFs was filtered to include those that present a continuous trend along the ECM clusters.

3.9 Network to Drug Targets Proximity Calculation

To identify drugs that may be associated with the ECM clusters, network proximity analysis was performed. This analysis takes into consideration the distances between drug target proteins and disease proteins, in our case, ECM proteins. The algorithm maps the set of disease proteins S and drug target proteins T onto a connected interactome and calculates the shortest distance between targets and proteins. For each pair of drug targets, t, and disease protein, s, the shortest path length $d(s,t)$ is calculated, and the minimum length is picked for the pair. This is done on all of the drug targets, and to quantify the distance of the drug targets as a collective, the average of these picked values is taken, and the result is referred to as the closest distance $d_c(S,T)$ as shown in Equation 3.9.

$$d_c(S,T) = \frac{1}{||T||}\sum_{t \in T} min_{s \in S} d(s,t) \tag{3.9}$$

Then, the significance of this distance is determined by comparing this value to randomly generated distance values, and a relative proximity z_c is calculated. A reference distribution of distances is calculated by selecting two sets with the same size and degree as the original disease and drug target proteins and calculating the closest distance. This process is repeated 1000 times, and then the mean $\mu_{\mathrm{d(S,\ T)}}$ and standard deviation $\sigma_{\mathrm{d(S,\ T)}}$ of the randomly generated distance values are calculated.

These values are used to calculate the relative proximity $z(S,T)$ with Equation 3.10. Drugs were considered proximal to the disease proteins when $z_c \leq -0.15$ and significantly proximal when $z_c \leq -2$.

$$z(S,T) = \frac{d_c(S,T) - \mu_{\mathrm{d(S,\ T)}}}{\sigma_{\mathrm{d(S,\ T)}}} \tag{3.10}$$

For each cluster, a consensus network was constructed from the patient-specific networks. To do this, the frequency of each edge within the patient population of an ECM cluster was determined, and those observed in 30% of the patients were retained. Final consensus networks were obtained by selecting the largest connected component from the edge-cleared data. Removal of the edges was done in such a way that cluster-specific node counts were rather equal among clusters and edges present in at least four patients were retained, accounting for the clusters that have small patient sizes. For network proximity analysis, these four cluster-specific networks were used. The reference interactome used in patient-specific network construction was used in this analysis as well. Drug target protein data was retrieved from DrugBank. The calculation of relative proximity was done using the default settings for the "calculate_proximity" function and setting the nodes_to to cluster networks and the nodes_from to drug targets.

3.10 Statistical Analysis

Spearman rank correlation was performed for all of the correlation analyses. A two-sided Mann-Whitney U test was employed to calculate the p-values when comparing the distributions of ECM clusters for all of the scores mentioned except for the similarity score comparison. For that, a paired, two-sided, Wilcoxon signed-rank test was employed. P-values smaller than or equal to 0.05 were considered significant throughout the study.

Chapter 4

RESULTS AND DISCUSSION

4.1 Multi-Omic Investigation of ECM Profiles of Patient Samples

Multi-omic data analysis was performed to understand the ECM gene and protein expression profiles in tumor and NAT samples. Transcriptomic data was refined to show ECM genes, revealing 951 genes out of the 1028 genes listed in the Matrisome Project, giving an ECM gene coverage of 92%. 264 (28%) of the ECM genes were of the core matrisome, and 687 (72%) were matrisome-associated, as shown in Figure 4.1a. The majority of core matrisome content was ECM glycoproteins with 187 (71%) followed by collagens with 44 (17%) and then proteoglycans with 33 (12%) as shown in Figure 4.1b. The majority of matrisome-associated content was secreted factors with 308 (45%) followed by regulators with 220 (32%) and then affiliated with 159 (23%) as shown in Figure 4.1c. These observations overlap with the portion of ECM divisions and categories in the Matrisome Project, showing that the cohort gene expression data can represent the ECM profiles with high coverage.

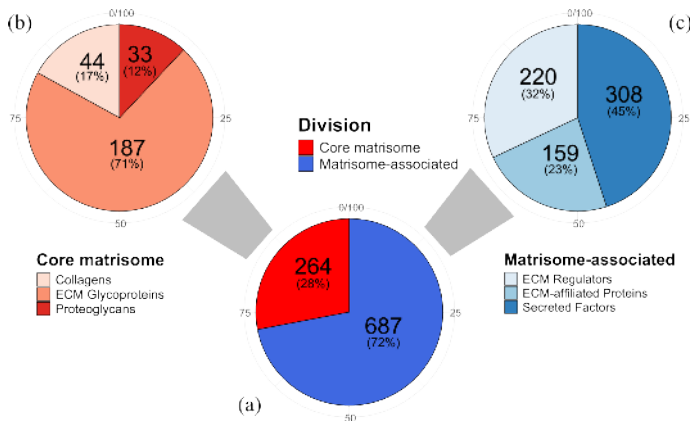
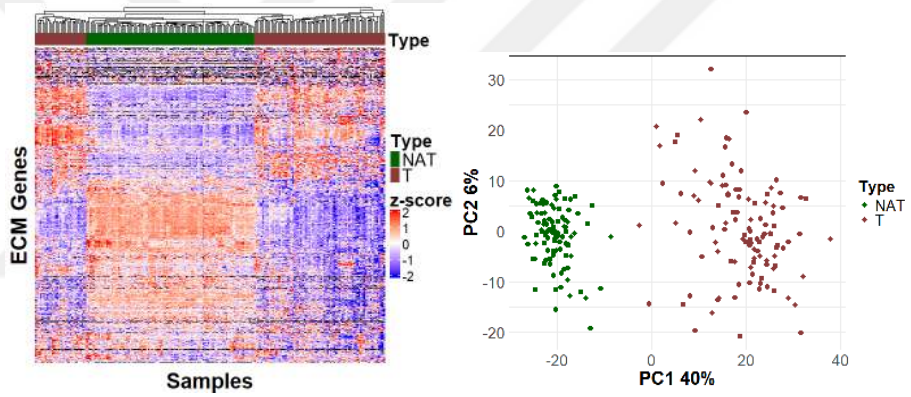


Figure 4.1: ECM Gene Composition: (a) matrisome divisions, (b) core matrisome categories, (c) matrisome-associated categories.

ECM gene expression levels between tumor and NAT samples were represented with z-scores, and samples were subsequently clustered hierarchically, as shown in Figure 4.2a. This showed an apparent separation between tumor samples and NAT samples. NAT samples were all observed under a single branch, while tumor samples were separated into multiples. Similarly, PCA showed separated tumor and NAT clusters, proving that there is indeed an altered expression of ECM genes with tumor development, as shown in Figure 4.2b. PCA also showed a more dispersed cluster for the tumor samples, suggesting more heterogeneous ECM expression profiles for the tumors, while a denser cluster for NAT samples showed a rather uniform ECM gene expression.



(a) (b)

Figure 4.2: ECM Gene Expression Profiles: (a) Heatmap of ECM gene expression, (b) PCA plot of ECM gene expression.

The proteome of the tumor and NAT samples were also analyzed to get ECM protein distributions. The proteome of the samples was filtered to include only ECM proteins. This gave a total of 563 ECM proteins. 210 (37%) of the ECM proteins were of the core matrisome, and 353 (63%) were matrisome-associated, as shown in Figure 4.3a. The majority of core matrisome content was again ECM glycoproteins with 146 (70%) followed by collagens with 41 (20%) and then proteoglycans with 23 (11%), as shown in Figure 4.3b. The majority of matrisome-associated content was ECM regulators with 158 (45%) followed by secreted factors with 101 (29%) and then affiliated with 94 (27%), as shown in Figure 4.3c. Although the ECM

division and category percentages within the cohort proteomic data were very close to the portions in the Matrisome project, small nuances were observed. This can be attributed to the fact that ECM proteins have many isoforms, and the cohort can represent multiple isoforms of a particular protein with RefSeq IDs. Therefore, it can be said that the proteomic data given in the cohort can represent ECM proteins while also capturing their isoforms.

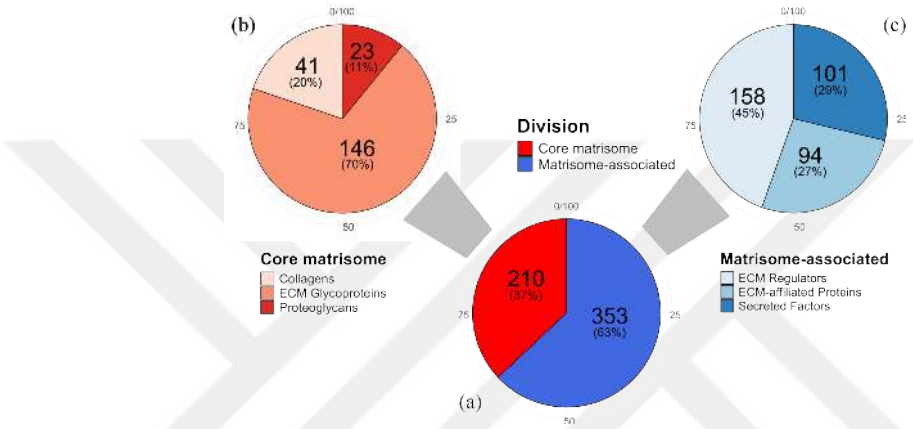
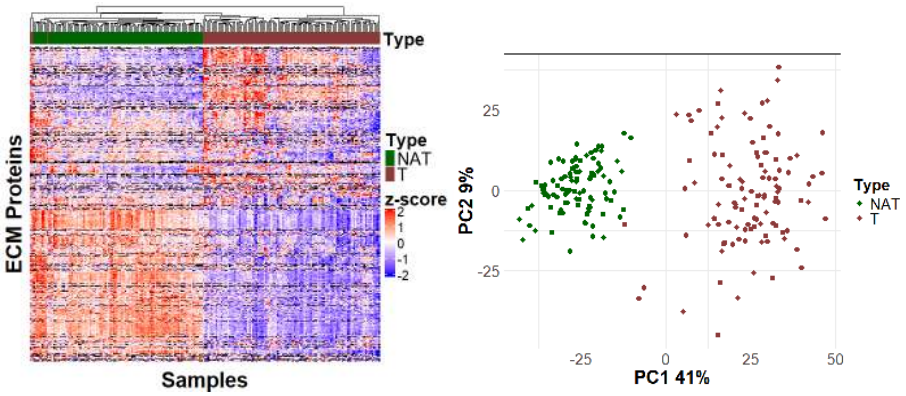


Figure 4.3: ECM Protein Composition: (a) matrisome divisions, (b) core matrisome categories, (c) matrisome-associated categories.

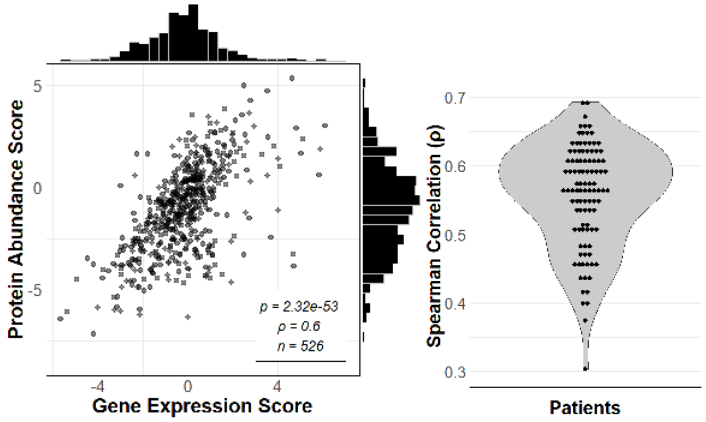
The representation done on the ECM gene expression data was also performed on the protein abundance data. ECM protein levels between tumor and NAT samples were given with z-scores and subjected to hierarchical clustering, as shown in Figure 4.4a. Similarly to the observation for the transcriptomic data, a clear separation between tumor samples and NAT samples was present. Again, NAT samples were all observed in one branch while tumor samples were separated into multiples. Similarly, PCA showed a major distance between tumor and NAT clusters, indicating that the ECM protein abundance profiles for tumor and NAT samples are different, as shown in Figure 4.4b. The same high dispersion observed in tumor samples and compacted distribution of NAT samples in transcriptomic data were also present in this analysis, which again shows the heterogeneous nature of the tumors regarding their ECM abundance and a more stable ECM profile for NAT samples. So, ECM protein abundance changes in tumor samples and shows higher heterogeneity.



(a) (b)

Figure 4.4: ECM Protein Abundance Profiles: (a) Heatmap of ECM protein abundance, (b) PCA plot of ECM protein abundance.

ECM gene and protein distribution and expression profiles among tumor and NAT samples were shown to agree with each other. To further show the relationship between the transcriptomic and proteomic data, correlation analysis was performed. The Spearman rank correlation between the average difference in ECM gene expression and protein abundance within the patient population was calculated, and a significant positive relationship was observed between the two omic layers with a ρ_{spearman} value of 0.6, as can be seen in Figure 4.5a. A similar test was performed on each patient. The difference values of the ECM genes from transcriptomic and proteomic data were correlated for each patient, and the distribution of the ρ_{spearman} value was visualized in Figure 4.5b. The distribution showed that the median ρ_{spearman} value was equal to 0.56. Both of these analyses show a strong positive correlation between the transcriptomic and proteomic data, suggesting that the transcription of an ECM gene is likely to be followed by the translation of its protein at both global and patient-specific scales.



(a) (b)

Figure 4.5: Transcriptomic and Proteomic Data Correlation: (a) global scale, (b) patient scale.

4.2 Identification of ECM Grades with Multi-Omic Barcoding and Reference-Based Clustering of Patients

Multi-omic barcodes of patients were used for reference-based consensus clustering. A Monte Carlo p-value was calculated for each K value up to 10, and a K value of 4 was found to be significant, as shown in Figure 4.6. Forming four clusters was the only option for rejecting the null hypothesis, which is that the population does not form any clusters. Therefore, a K value of four was used for clustering the patients based on their ECM barcodes and the patient IDs of these four ECM clusters can be found in Appendix Table A.1.

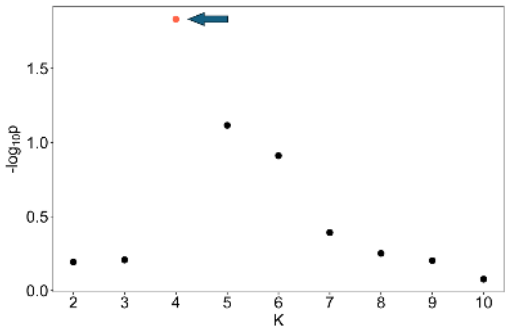


Figure 4.6: Monte Carlo p-values of K values

A distinctive gradient-like pattern was observed between these four clusters that go from low grade to high grade of ECM enrichment, which can be seen in Figure 4.7. The group with the lowest grade of ECM richness, with 21 patients, was named ECM Grade-1 or ECM-G1. This group presented almost no enrichment in ECM categories but only a very mild expression in collagens compared to other groups. The largest and second-least ECM-enriched group with 35 patients was named ECM Grade-2, or ECM-G2, and was more enriched in collagen and glycoproteins when compared with ECM-G1. The third group with 32 patients showed higher ECM enrichment than ECM-G2 and was named ECM Grade-3 or ECM-G3. This group was again more enriched in collagens and glycoproteins, but additionally in proteoglycans, when compared with ECM-G2. The final and least populated group of 13 patients that showed the highest enrichment in all ECM categories was named ECM Grade-4 or ECM-G4. It can be said that for all the clusters, the collagen content was the common ECM category that was enriched, although at different levels. And, for almost each ECM grade, a new ECM category became enriched in a rather step-wise manner, especially the categories belonging to the core matrisome, which indicates the ECM heterogeneity in LUAD patients.

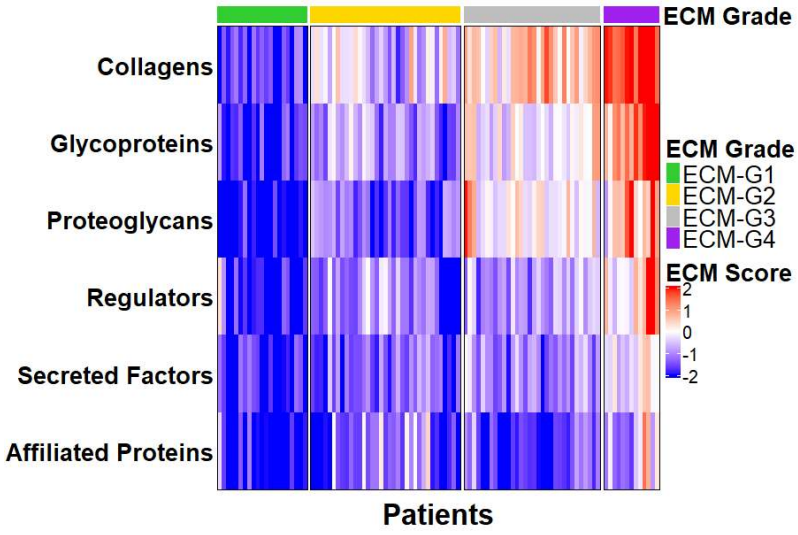


Figure 4.7: ECM Grades

To understand the relative distances between these ECM clusters, PCA was performed, as shown in Figure 4.8. It was observed that the clusters ECM-G1 and ECM-G4 were the furthest from each other on PC1, suggesting a great difference between these groups. This is expected since these groups, ECM-G1 and ECM-G4, represent the two ends of the ECM gradient, which is from poor to rich ECM content, respectively. Also, the ECM-G4 group showed higher dispersion suggesting that the individuals within this group may be more unique. The other two groups, ECM-G2 and ECM-G3, resided between ECM-G1 and ECM-G4 along PC1, suggesting that these groups may be intermediate groups, yet ECM-G2 was closer to ECM-G1 and ECM-G3 was closer to ECM-G4, which could suggest that ECM-G2 and ECM-G3 may be more likely to transition into groups ECM-G1 and ECM-G4, respectively.

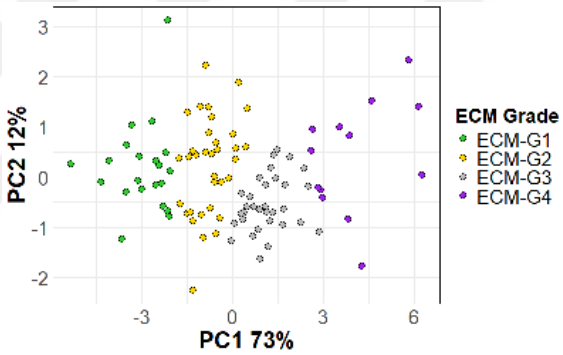


Figure 4.8: PCA Plot of ECM Grades

To investigate whether these clusters show significantly changed ECM profiles, the ECM scores of the patient groups were compared, as shown in Figure 4.9. A statistically significant increase was observed between all ECM grades for collagen score (Figure 4.9a), glycoprotein score (Figure 4.9b), proteoglycan score (Figure 4.9c), regulator score (Figure 4.9d), and secreted factor score (Figure 4.9e). Only for affiliated protein scores, ECM-G2 and ECM-G3 did not show a significant change (Figure 4.9f). Considering the proximity observed for these two clusters in the PCA, not observing significance can be expected. All in all, it can be said that these four clusters show distinct changes in all dimensions of the ECM and a gradual increase in the enrichment of ECM components. This also shows that the ECM scores

calculated in this study perform well at showing the enrichment of the ECM since they can result in confident stratification of patients.

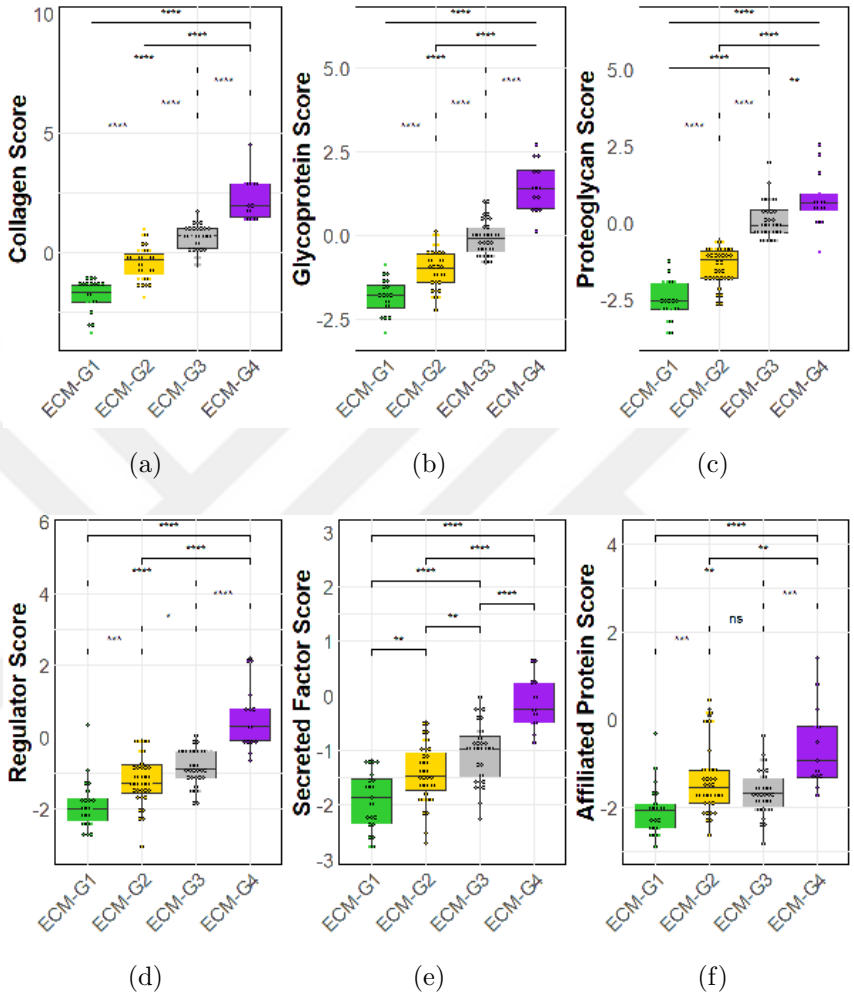


Figure 4.9: ECM Scores of Patient Clusters: (a) collagen score, (b) glycoprotein score, (c) proteoglycan score, (d) regulator score, (e) secreted factor score, (f) affiliated protein score.

To investigate the altered expression levels of other genes within these clusters, highly variable genes were identified. With very stringent and cluster-specific thresholding, 39 genes were marked as highly variable, and 28% (11) of these genes were of the matrisome genes, as shown in Figure 4.10. Among these genes, some genes showed elevated expression as the ECM grade was increased, such as ITGA11. This protein is a known collagen receptor that is not normally expressed in alveolar ep-

ithelia, but its overexpression in lung cancer cell lines was highly correlated with invasiveness and migration [Ando et al., 2019]. One of the genes called PIK3R2 was especially expressed at higher levels for ECM-G3. This gene is known to be a downstream mediator of FAK signaling and can activate AKT, which results in the inhibition of apoptosis [Aboubakar Nana et al., 2019].

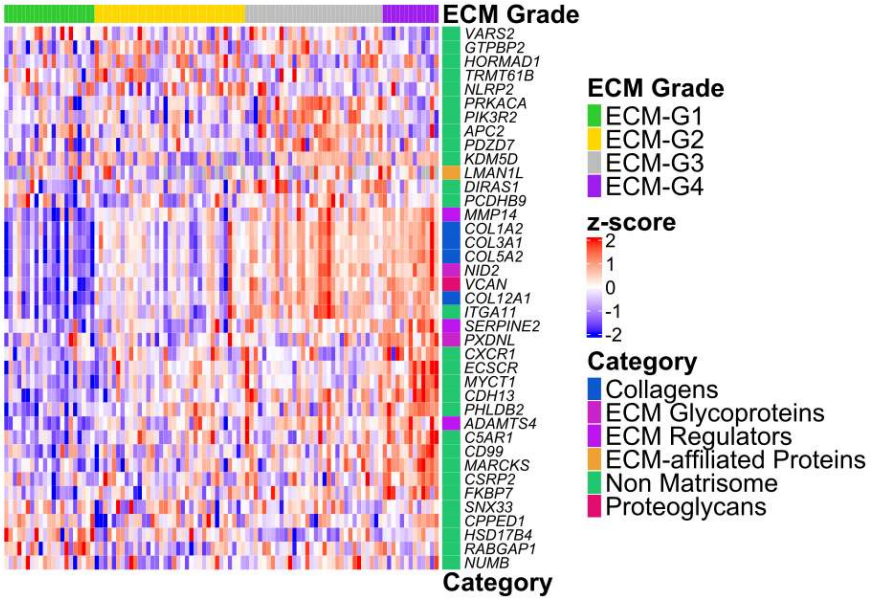


Figure 4.10: Highly Variable Genes

Functional enrichment analysis was done on these highly variable genes. The most enriched REACTOME pathway was ECM organization, followed by pathways related to the synthesis, modification, formation, and degradation of ECM or ECM components, as shown in Figure 4.11a. Similarly, the most enriched GO biological process was ECM organization, followed by processes related to external structure and fibril organization, as shown in Figure 4.11b. All in all, both the percentage of ECM genes within the highly variable gene set and the ECM-related pathway and biological process enrichments from the enrichment analysis suggest that these clusters can capture ECM features of the population and give groups that show different ECM profiles.

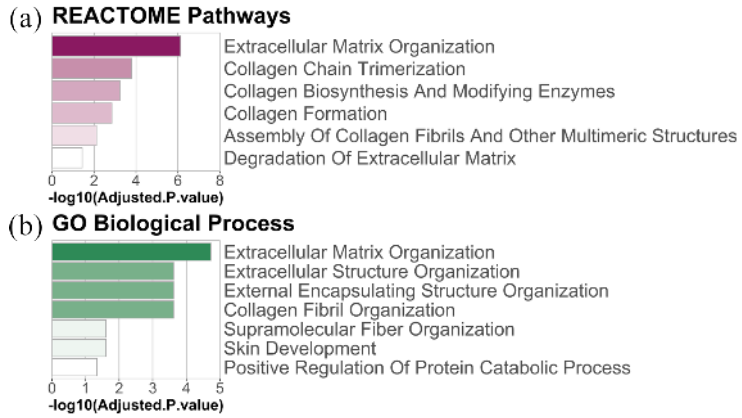


Figure 4.11: Functional Enrichment Analysis on Highly Variable Genes: (a) enriched REACTOME pathways, (b) enriched GO biological processes.

4.3 Cellular Deconvolution of Patient TMEs and Correlation Analyses

Cellular heterogeneity of patient tumor and NAT samples was determined using MCP-Counter, and the change in cellular enrichment between paired samples was calculated to understand the composition of cells within the TME, as shown in Figure 4.12. For almost all of the ECM clusters, no change in the enrichment of immune-related cells was observed. This may be attributed to the potential loss of immune cell content during sample acquisition, preservation, and preparation before high-throughput analyses. Trends clearly show that the loss of endothelial cell content in ECM-G1 tumor samples is much greater when compared with other grades, and this change decreases as the ECM grade becomes more severe. A similar but more apparent pattern is also observed for CAFs. For ECM-G1, although the variance is high, the average change in CAF enrichment change seems to be close to zero, suggesting that there might not be a considerable change in CAF content in the tumor samples of patients within this group. Yet, a clear and gradual increase in CAF enrichment is observed as the severity of the ECM grade increases again.

A statistical comparison of the cell enrichment changes in tumor samples of ECM grades shows significant enrichment in CAF abundance as the ECM grade is elevated, as shown in Figure 4.13a. CAF enrichment is almost the same for ECM-G1, while for

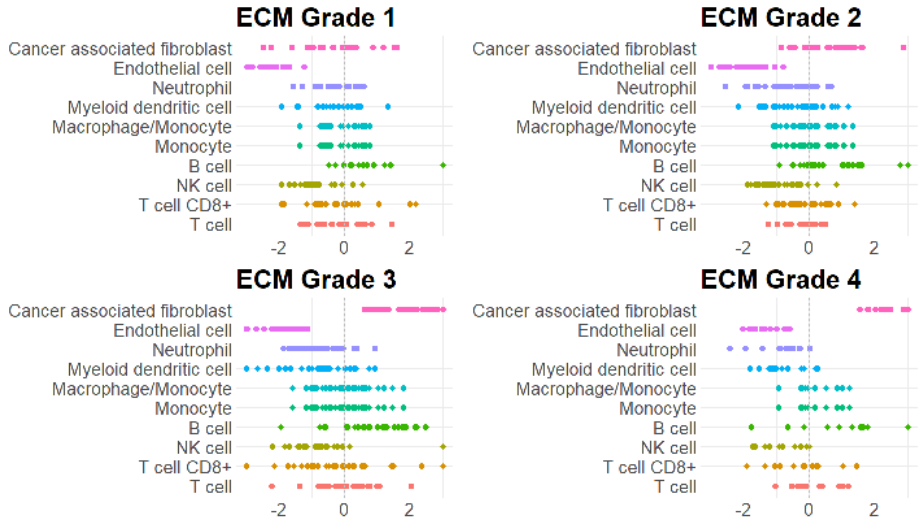
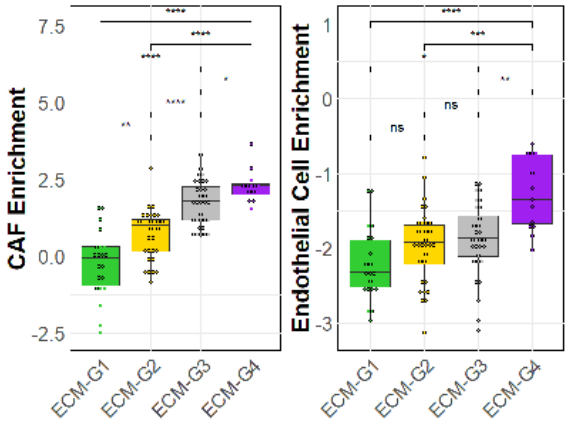


Figure 4.12: Cellular Heterogeneity of ECM Grades

ECM-G4, the tumor samples show 4.89 times the CAF enrichment observed for their NAT match. A significant increase in CAF enrichment between each ECM grade indicates that each component of the matrisome, along with their abundance, has an influence on CAF recruitment and differentiation and proves that the abundance of ECM content is positively correlated with CAF maintenance. Regarding the endothelial cell content, tumor samples from all groups show decreased enrichment compared to their NAT match, as can be seen in Figure 4.13b. Yet, the decrease observed for ECM-G4 is significantly less than all the other groups, suggesting that the TME in this group of patients is more likely to maintain endothelial cell growth. It is known that elevated ECM deposition may lead to hypoxic conditions and mediate the recruitment of endothelial cells to build vasculature. Considering that this group is the most ECM-enriched group, it is more likely that this group can call for and support endothelial cells. Similar results were also obtained using another deconvolution tool called EPIC (Appendix Figure A.1).

A correlation analysis was performed to further reveal the relevance of stromal content and tumor purity to ECM profiles. ESTIMATE stomal scores showed positive correlations with the ECM scores, while tumor purity was negatively correlated as expected, as shown in Figure 4.14a. All of the ECM scores showed a significantly

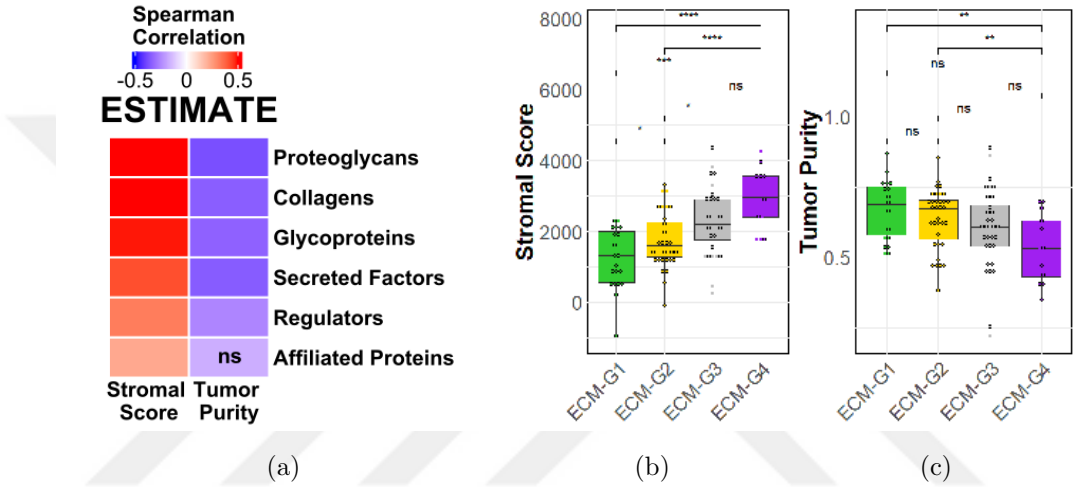


(a) (b)

Figure 4.13: CAF and Endothelial Cell Enrichment of ECM Grades: (a) CAF enrichment, (b) endothelial cell enrichment.

positive correlation with the stromal score, with the lowest being the affiliated proteins with a ρ_{spearman} equal to 0.22. Especially core matrisome genes were highly correlated, with the highest being the proteoglycans, with ρ_{spearman} equal to 0.58. In a similar sense, tumor purity showed a significant negative correlation with the ECM scores, except for the affiliated protein scores, which were not significant. Again, the proteoglycan score showed a high negative correlation with a ρ_{spearman} value of -0.38. Comparison of the stromal scores of ECM grades also showed a significant increase as the grade severity increased, except for the groups ECM-G3 and ECM-G4, as shown in Figure 4.14b. The tumor purity of ECM-G4 was significantly less than that of ECM-G2 and ECM-G1, with an average purity of 0.53, and ECM-G1 showed the highest purity with 0.67, as shown in Figure 4.14c. Tumor purity calculation is done on ESTIMATE scores, which take into account both the stromal and immune cell content. A highly significant increase in stromal scores was present between ECM groups, but this was not the case for the immune cells and a loss of immune cell content in tumor samples was observed in the cellular deconvolution analysis as well (Appendix Figure A.2a). This depletion may be the reason why differences between ECM groups were not apparent in both the ESTIMATE score and tumor purity (Appendix Figure A.2b). Overall, these analyses suggest that the cellular het-

erogeneity of the TME, which includes stromal cells like CAFs and endothelial cells, is increased as the complexity and abundance of the ECM are elevated, whereas a more tumor-pure TME is present for lower ECM grades. Another conclusion that can be drawn is that ECM grades, although separated well from each other, may show some patterns where they show characteristics of another, especially those that are between the two poles of the ECM grades, which are ECM-G1 and ECM-G4.



(a) (b) (c)

Figure 4.14: ESTIMATE Correlations with ECM Scores and Grades: (a) correlation analysis, (b) ESTIMATE stromal score, (c) tumor purity.

Scoring was done on cancer-associated gene groups like EMT, CSC, invasiveness, and integrin richness, and their correlation with the ECM scores was calculated, as shown in Figure 4.15. A robust positive correlation with the core matrisome scores was observed for the EMT and INV scores. Especially the proteoglycan score was the most positively correlated for both with a ρ_{spearman} correlation of 0.64 for EMT and 0.77 for invasiveness, which is consistent with the statements given by Kuşoğlu et al. Not a major correlation was observed between ECM scores and CSC and INT scores. Only the collagen score was somewhat correlated with the CSC score, given the maintenance of the self-renewal of CSCs by collagen [Suh and Han, 2011].

In accordance with the correlation analysis, the EMT scores and INV scores of patients between ECM grades showed a significant increase as the grade increased. ECM-G1 was the significantly lowest group for both scores, and each ECM group

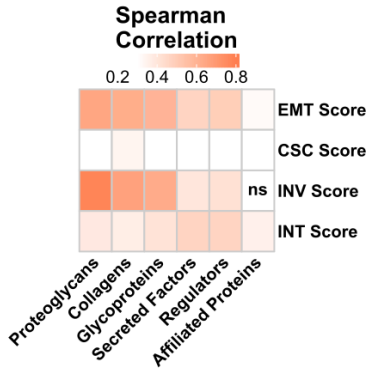
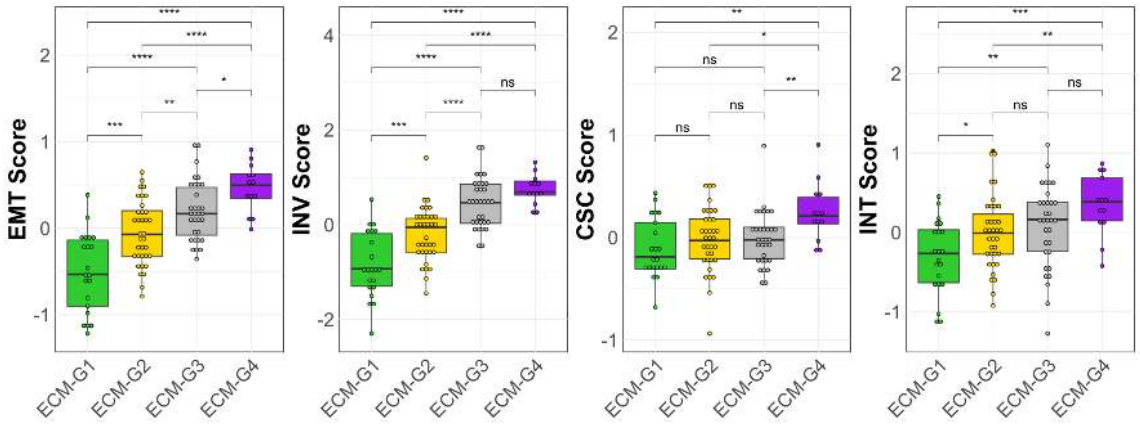


Figure 4.15: Correlation between Cancer-Associated Gene Group and ECM Scores

showed significantly increased EMT scores, as shown in Figure 4.16a. Only for invasiveness, no significant increase was observed between ECM-G3 and ECM-G4, yet given their proximity and the tendency to transition that ECM-G3 has on ECM-G4, this was understandable, as shown in Figure 4.16b. For the CSC score, ECM-G1 to ECM-G3 did not show a significant change in CSC scores, and ECM-G4 was significantly higher than all of the other groups, as shown in Figure 4.16c. This suggests that the maintenance of CSCs could require very aggressive ECM profiles. ECM-G1 showed significantly reduced INT scores when compared with other groups except for ECM-G2, yet the remaining groups showed rather similar INT scores. This may suggest that some level of ECM enrichment is necessary for ECM-integrating interaction, and when ECM profiles are poor, no such interaction could be facilitated, as shown in Figure 4.16d.

4.4 Clinical Interpretation of ECM Grades

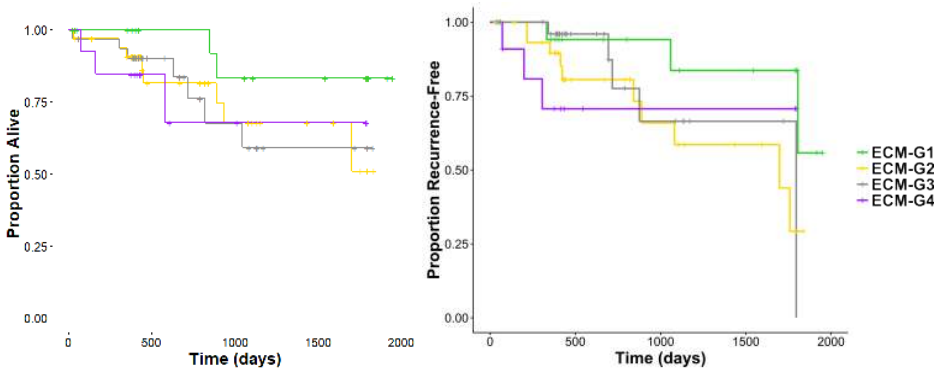
Survival analysis was conducted on ECM grades to review overall survival patterns, and ECM-G4 presented a rapid decline in survival probability and was one of the poorest groups to survive, as shown in Figure 4.17a. ECM-G1, on the other hand, had the best survival probability among the groups. ECM-G2 and ECM-G3 showed moderate survival for most of the time period yet decreased even more than ECM-G1 within the last three years. A similar trend was also observed for recurrence-free



(a) (b) (c) (d)

Figure 4.16: Cancer-Associated Gene Group Scores in ECM Grades: (a) EMT score, (b) INV score, (c) CSC score, (d) INT score.

patterns, as shown in Figure 4.17b. Again, ECM-G4 showed a rapid decline in recurrence-free proportion more than the other groups, and ECM-G1 presented a lower recurrence probability. Cox proportional regression analysis also showed that patients with ECM-G4 have six times the risk of dying than patients with ECM-G1, with a p-value of 0.058 (Appendix Figure A.3).



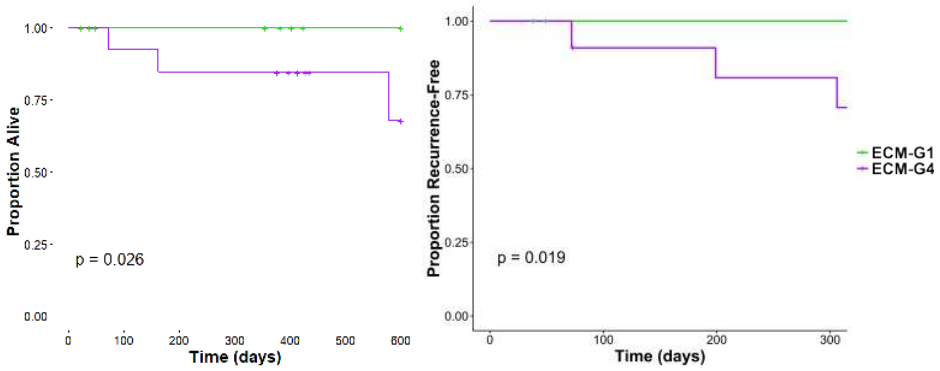
(a) (b)

Figure 4.17: Survival and Recurrence-Free Analysis: (a) survival, (b) recurrence-free.

To investigate the rapid decline in both curves, a short-term analysis was performed. The analysis revealed, according to the log-rank test, a significant decrease in survival and recurrence-free probabilities between ECM-G1 and ECM-G4 within the first 600 and 300-day period, respectively, as shown in Figure 4.18a,b. Here we can speculate that patients that have more enriched ECM present poorer survival and a higher probability of recurrence than those that have less enriched ECM. The high rate of recurrence within the ECM-rich group can be attributed to the fact that this group was the only group that showed higher CSC enrichment among others, and CSCs are a well-known source of recurrence due to their self-renewal and differentiation capabilities and give rise to new tumors [Aramini et al., 2022]. It can also be speculated that the recurrence observed within the first 300 days may have resulted in the poor survival observed within the first 600 days. All in all, it can be said that the ECM profiles of patients indeed have an effect on poor prognosis at some level. Yet, it is important to know that the survival and recurrence data of this cohort contains much right-censored data, meaning that the event has not happened within or even after the time interval of the given study, and the overall survival time is unknown. In this dataset, 78 patients out of 101 are censored, and it should be stated that the amount of censored data is inversely proportional to the reliability of the survival analysis. Therefore, survival and recurrence analysis done on this may be indicative of prognosis, yet more follow-up is necessary for proper evaluation [Rich et al., 2010].

Other clinical data related to the ECM groups were investigated to show possible patterns, and a trend in percentages of clinical factors like sex, mRNA subtypes, CIMP status, and tumor grade was observed, as shown in Figure 4.19.

ECM-G4 consisted of 84.61% males, with the highest percentage among the other grades. It has been shown that there are sex-specific trends in lung cancer, and males show higher risk and mortality compared to females [May et al., 2023]. In our case, ECM-G4 showed one of the lowest survival and short survival trends, which is consistent with the ratio of men within this group. In the survival analysis, the HR for men, although not significant, was 59% higher than for females.



(a) (b)

Figure 4.18: Short-Term Survival and Recurrence-Free Analysis: (a) survival, (b) recurrence-free.

Regarding the mRNA subtype, ECM-G4 had the highest percentage of members having the PP subtype with 69.23% and the lowest percentage of TRU subtype, while ECM-G1 had the contrary with 71.42% TRU and 9.52% PP. An apparent increase in the percentage of patients with the PP subtype was present as the ECM grade was elevated. This suggests that the transcriptional profile is shifting from a normal lung tissue profile, that is, TRU, to a profile with a higher cell cycle and proliferation-related gene expression profile, that is, PP. The PI percentage remained rather constant among all grades except ECM-G3, which has 40.62% of its members in this subtype, suggesting that this group may have an altered immune response.

For the CIMP status, again, ECM-G1 had the highest percentage for the CIMP-2 status with lower methylation than CIMP-A and CIMP+, with 47.61%, while this percentage showed a decrease as the ECM aggressiveness increased, and the lowest percentage of CIMP-2 status was observed for ECM-G4 with 7.69%. Although a higher percentage of CIMP+ patients was observed in ECM-G1, the overall percentage of CIMP-1 and CIMP+ was much higher in ECM-G4 and ECM-G3. This suggests that overall hypermethylation of CpG islands is much more common in ECM-G3 and ECM-G4, making these groups more susceptible to the silencing of tumor suppressors.

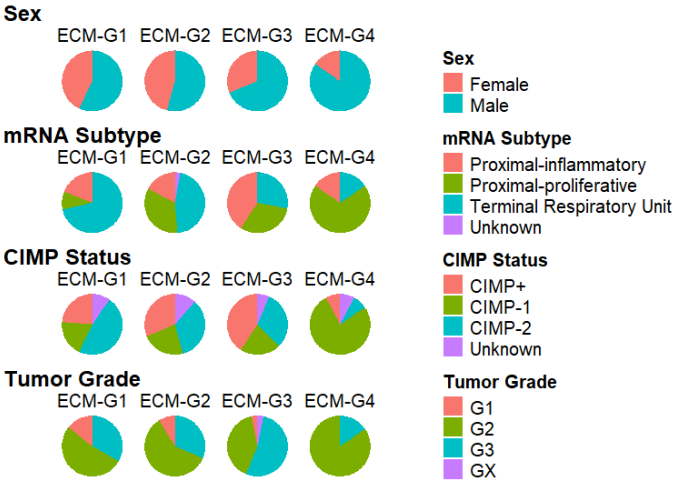


Figure 4.19: Clinical Data on ECM Grades

Finally, tumor grade showed an altered pattern as the ECM grades shifted. ECM-G1 was the group that showed the highest percentage for the mildest tumor grade with 14%, while this was 8.57% for ECM-G2, 3.12% for ECM-G3, and 0% for ECM-G4. Similarly to CIMP status, tumor G1 was observed more in ECM-G1 than ECM-G4, yet the entire population of ECM-G4 showed more aggressive tumors. Especially, ECM-G3 held the highest percentage for the most aggressive tumor grade G3 with a percentage of 53.12%. Overall, it can be said that tumor grade is also influenced by the ECM grade, where a slight decrease in mild tumor grade is observed and replaced by more aggressive tumors that grow and spread much more quickly.

Mutation profiles of ECM clusters were investigated to see relevant patterns, as shown in Figure 4.20. The number of somatic mutations was evaluated, and the highest median mutation count was observed for ECM-G4 with 216, followed by ECM-G3 with 188.5, followed by ECM-G2 with a drastic decline to 55.5, and finally with the smallest load of 42 for ECM-G1. It can be said that somatic mutations seem to accumulate more in ECM-rich groups than in poor ones. High mutation counts are known to give rise to more aggressive cancer phenotypes and higher tumor heterogeneity, which was shown to be the case for ECM grades as they went

higher [Ding et al., 2008, Pan et al., 2005]. These results may suggest that there could be a mechanism led by the ECM that triggers genomic instability. Indeed, ECM can directly affect DNA repair mechanisms and influence the probability of somatic mutations by, for instance, ECM-induced stress or by stimulating mechanical signaling pathways by exerting mechanical forces on the cell [Ingber, 2008, Sokolova et al., 2020].

TP53 tumor suppressor, which is one of the most mutated genes in LUAD, was also highly mutated in this cohort, mainly with a missense mutation, which is shown to be enough to disrupt its function [Database, 2024]. Around 61% of the patients within ECM-G2, ECM-G4, and 46% for ECM-G3 had a TP53 mutation, while ECM-G1 had the lowest percentage with 38%. And a great portion of the TP53 mutations were missense mutations. So, the overall mutation on TP53 was quite high for all of the groups except for ECM-G1, which correlates with a better prognosis observed for this group. STK11 mutation was also heavily accumulated on the ECM-G4 side of the spectrum and mutation of this gene is shown to cause poor immunotherapy response [Papillon-Cavanagh et al., 2020].

KRAS and EGFR are well-known genes that activate pathways related to survival and cell growth. EGFR is a receptor on the cell surface, and KRAS is involved with signal transduction mechanisms. These genes show mutual exclusivity and are not observed to be mutated at the same time, considering that they activate similar pathways like MAPK/ERK [Suda et al., 2010]. Therefore, the existence of a mutation in one of them is sufficient to sustain tumor growth. In ECM grades, we observed a higher KRAS mutation accumulating on ECM-G4 and a higher EGFR mutation on ECM-G1. This presents not only the clear nature of mutual exclusiveness between KRAS and EGFR but also how ECM grade favors mutation of KRAS since the increase in abundance of ECM can compensate for EGFR mutation and ECM poor groups like ECM-G1 and ECM-G2 require mutation of EGFR to stimulate growth signaling pathways.

To observe similar ECM clusters within the TCGA LUAD cohort, we barcoded TCGA patients using the transcriptomic data and assigned them to ECM groups. The result of this cohort also represented a gradient of ECM abundance, as shown

in Figure 4.21. KEAP1 mutations were accumulated in the ECM-G1 group, which is also known for its negative effect on immunotherapy [Hassanein et al., 2021].

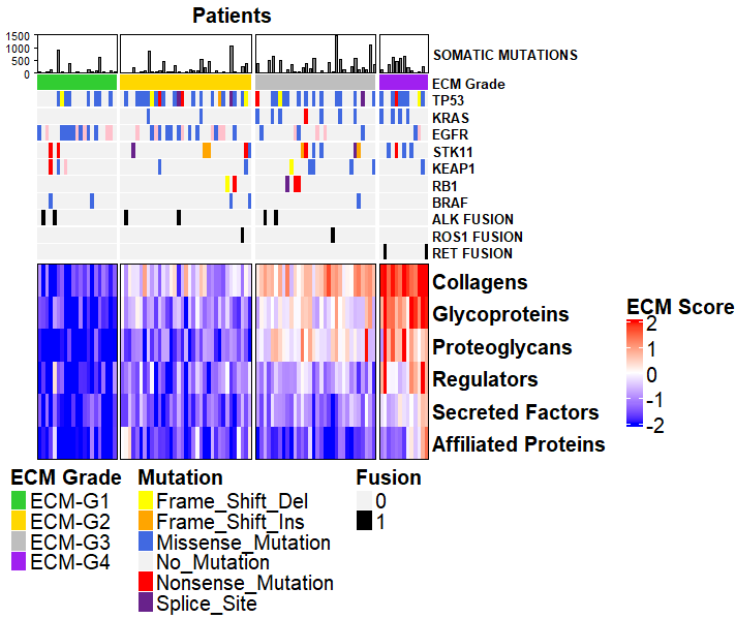


Figure 4.20: Mutation Profiles of Patients

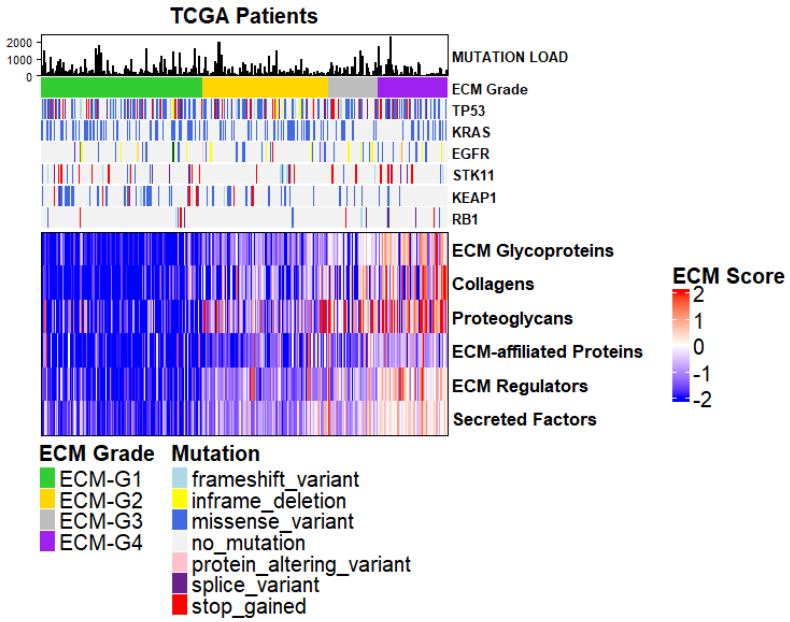


Figure 4.21: Mutation Profiles of TCGA Patients

4.5 Network Modeling of Patient Tumors within ECM Grades

Network modeling was performed using terminal lists prepared for each patient, and patient-specific networks were constructed. In total, 101 ECM-relevant signaling networks were established, and as an example, the network of patient C3N.00738 from patient group ECM-G3 is shown in Figure 4.22. This network consists of 101 nodes and 126 edges. It shows the category of each node, data on whether the node is a terminal or a Steiner, and to which louvain cluster the nodes are members. The edges are colored by the level of cost, and a smaller cost means better confidence for the edge.

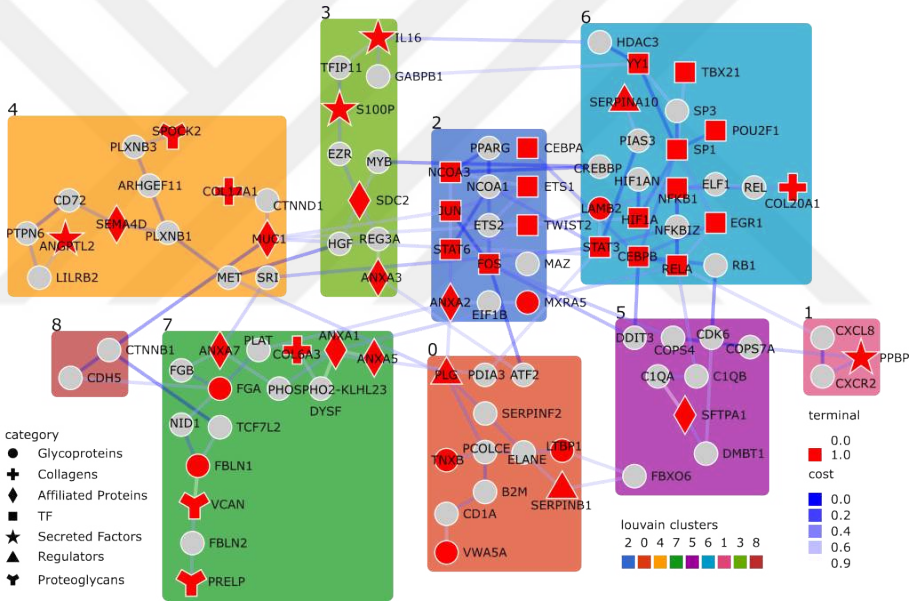


Figure 4.22: Network of Patient C3N.00738 of ECM-G3

Data on terminal, node, edge count, and edge confidence distribution is given in Figure 4.23. On average, the terminals determined for all the patients had 130 nodes with a minimum terminal count of 50 and a maximum of 180; the average node count after network modeling was 205 with a minimum of 50 and a maximum of 270. The average edge count of the patient population was 264, with a minimum of 125 and a maximum of 356. Here we can say that the initial data provided for network modeling was already quite heterogeneous, that is, each patient was

represented with a different number of terminal nodes acquired from multi-omic data. This shows that patient heterogeneity was somewhat expressed by the list constructed. Similarly, the node and edge count of patients after modeling were also different from each other, which shows that the network modeling strategy could preserve patient heterogeneity. For each network, the average of the edge confidence was taken, and it was observed that on average, the confidence in patient networks was 0.62 with a minimum of 0.58 and a maximum of 0.65. This shows that overall the edges given in the solution for patient networks were reliable.

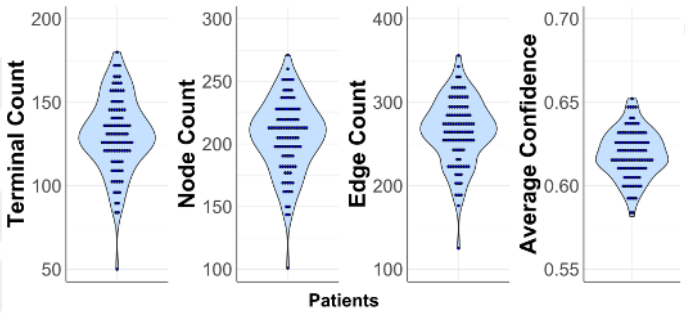


Figure 4.23: Basic Network Metrics

To understand the patient specificity of the nodes within patient networks, the percentage of occurrence of a node within the patient population was counted, as shown in Figure 4.24. If the occurrence of a node is close to 100%, this means that the node is shared in most of the patients, and if it is close to 0%, it is shared in just several patients. It can be seen in the figure that the distribution accumulates at the left side, close to 0%, and a decreasing trend is present as the occurrence is increased. There were 11 nodes shared by all of the patients, while 133 were observed in only one patient. The number of nodes observed in less than 25% of the population was 550, while more than or equal to 25% was 319, and the node count of occurrences above 75% was 58. This suggests that there are some nodes preserved within the entire population, but the majority has higher patient specificity.

The number of nodes observed within the ECM grades was obtained by taking the union of the nodes of patient networks of a particular ECM grade. Nodes that were only observed in 3 or fewer patients were removed to obtain more representative

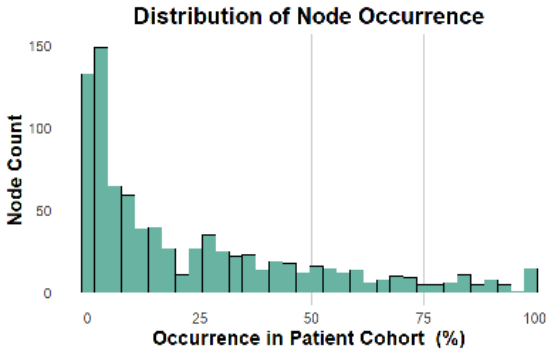


Figure 4.24: Distribution of Node Occurence

node sets for the ECM groups. Node sizes of ECM grades and shared nodes between these groups are shown in Figure 4.25. The largest node sizes were for ECM-G2 and ECM-G3, with a set size of 465 for both, followed by ECM-G1 with a node size of 382 and ECM-G4 with 259. There were 232 nodes shared by all of the groups. The most common nodes shared only by a pair of ECM groups were between ECM-G2 and ECM-G3 with 41 nodes, and the least was between ECM-G1 and ECM-G4 with no shared nodes. Node sizes of the ECM groups can be related to the number of patients within each group. ECM-G2 and ECM-G3 had similar and higher patient sizes than ECM-G1 and ECM-G4. So, it is understandable to have more nodes for these groups. Also, observing higher and lower commonality between ECM-G2 and ECM-G3 and ECM-G1 and ECM-G4, respectively, is consistent with the observations in the PCA of ECM grades (see Figure 4.8). Here we can also say that there is a great number of nodes shared among all of the ECM groups, and these genes may be essential for the overall maintenance of the ECM.

To further show the capability of these networks to represent these ECM clusters, we integrated tumor z-score data onto the nodes within the patient networks. Patients and ECM groups may be presenting the same nodes in their networks, yet the expression of these nodes may be different among patients. It can be seen that a slight gradient of node expression from low to high was present as the ECM grade was elevated, as shown in Figure 4.26a. To show this concept more in detail, a network score was calculated for each patient by taking the average of tumor z-scores

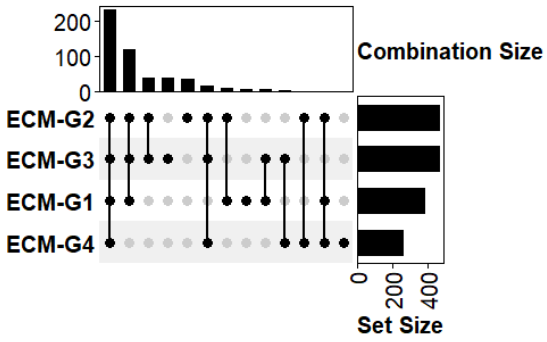
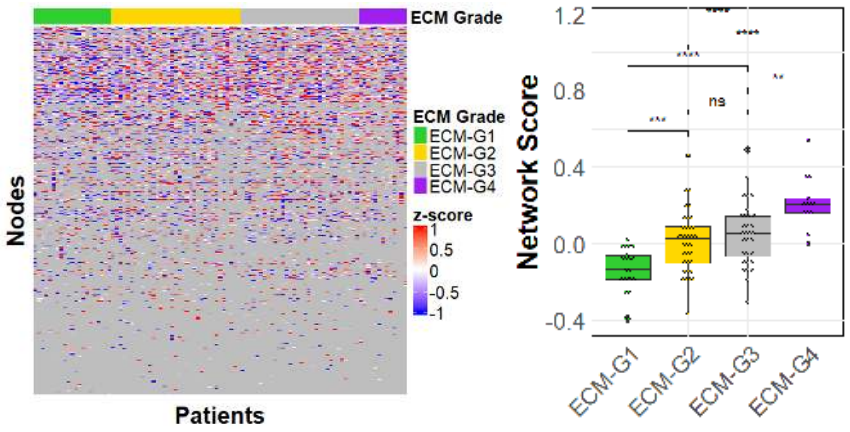


Figure 4.25: Combination and Set Sizes of ECM Grades

of the nodes within a patient network. It can be seen that the network score was significantly increased as the ECM grade was increased, as shown in Figure 4.26b. This significance was not present between ECM-G2 and ECM-G3, but we have already observed the similarity of these two clusters to each other in many aspects. Here, it can be said that these networks could represent patients and ECM groups significantly while preserving their heterogeneous nature and common ECM characteristics by considering both the node composition and expression levels of the nodes.



(a) (b)

Figure 4.26: Patient Node Expression: (a) heatmap, (b) network score.

4.6 Similarity Analysis Between ECM Grades Using Patient Networks

To further reveal that the networks can represent patients and their ECM groups robustly, a similarity analysis was performed, as shown in Figure 4.27. Similarity scores calculated for ECM-G1 patients against the remaining portion of the ECM-G1 showed significantly higher scores than the scores calculated against a randomly selected patient list. The same concept was also observed for ECM-G2, ECM-G3, and ECM-G4. Therefore, it can be said that the networks of the patients within the same ECM group are significantly more similar to each other than to a randomly selected group of patients, and again, these networks could capture similar ECM profiles of patients of the same ECM group. It is also observed that ECM-G3 shows on average the most within-cluster similarity when compared with other groups, which is in accordance with the PCA of ECM grades, where ECM-G3 showed the least dispersion among others and was more clustered together (Figure 4.8).

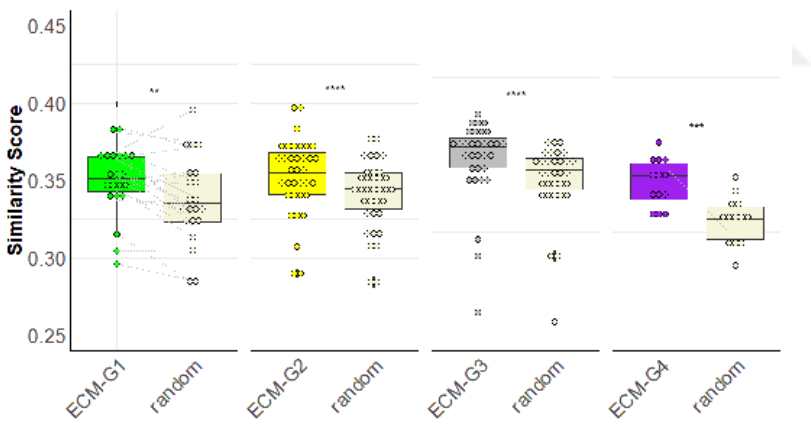
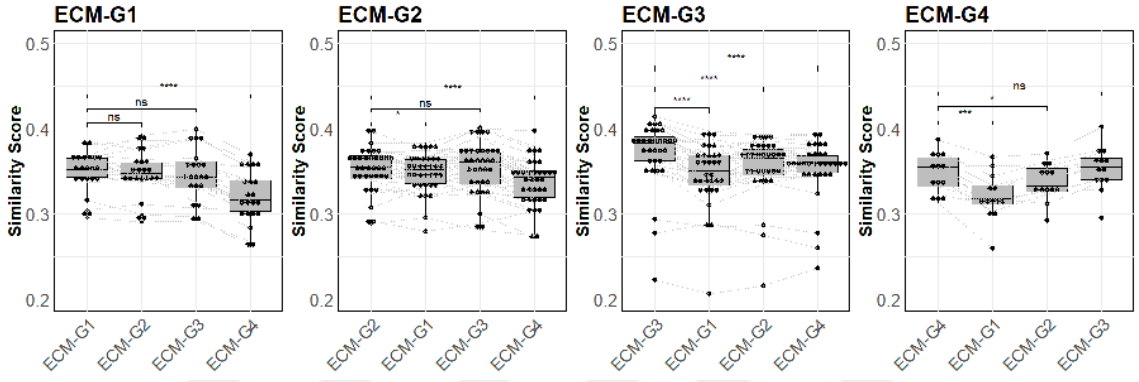


Figure 4.27: Similarity Score Comparison with Random Groups

Another similarity analysis was performed to detect which ECM groups are most similar or different from each other regarding the patient networks they hold, as shown in Figure 4.28. ECM-G1 within-cluster similarity scores were significantly higher than scores calculated against ECM-G4, while no significant difference was observed for ECM-G2 and ECM-G3, as shown in Figure 4.28a. This suggests that the network representations of individuals within ECM-G1 are very similar to the

patients in ECM-G2 and ECM-G3 but significantly different from the patients in ECM-G4. ECM-G2 within-cluster similarity was significantly higher than ECM-G1 and ECM-G4 but showed no significant difference with ECM-G3, as shown in Figure 4.28b. This suggests that patient networks in ECM-G2 are very similar to those in ECM-G3 which is consistent with the results shown in the PCA (Figure 4.8), node commonality (Figure 4.25) and network score comparison (Figure 4.26b).



(a) (b) (c) (d)

Figure 4.28: Similarity Score Comparison Between ECM Grades: (a) ECM-G1, (b) ECM-G2, (c) ECM-G3, (d) ECM-G4.

ECM-G3 within-cluster similarity was significantly different than all of the other ECM groups, which is rather surprising considering the resemblance it shows to ECM-G2 regarding some of the other analyses. Yet among the other two groups, ECM-G2 still showed the closest average similarity scores to ECM-G3, followed by ECM-G4, as shown in Figure 4.28c. Finally, ECM-G4 within-cluster similarity was significantly higher than ECM-G1 and ECM-G2, and the difference was not significant for ECM-G3. This suggests that the patients in this cluster are most similar to patients in ECM-G3 and are very different from the other groups, as shown in Figure 4.28d.

Overall, it can be said that ECM groups show greater similarity within themselves when compared with randomly selected groups, and some of the groups are very similar to each other while some are significantly different.

4.7 Enrichment Analyses of ECM Grades

Enrichment analysis was performed on the patient networks to make biological sense of the nodes, as shown in Figure 4.29.

Figure 4.29: Cluster-Specific Enriched KEGG Pathways: "*" shows pathways enriched in less than 25% of the patients in an ECM grade.

As the ECM grades increased, we observed many key KEGG pathways showing higher enrichment. These include pathways related to, of course, ECM and cell adhesion, proliferation, immunity, EMT, metastasis, apoptosis, cell senescence, signal transduction, metabolism, and hormonal signaling.

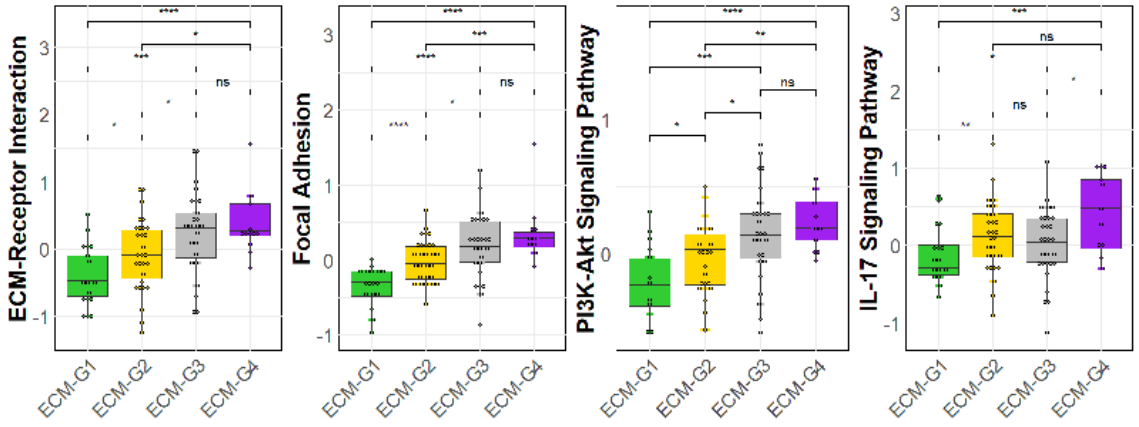
Many ECM-related pathways like ECM-receptor interaction, focal adhesion, regulation of actin cytoskeleton, and proteoglycans in cancer were enriched at higher levels for higher-grade ECM groups, as expected. Focal adhesions connect the actin cytoskeleton to the ECM, providing the cell-ECM interaction, and are important players in cell migration. Abnormalities observed for these pathways were shown to be highly correlated with the promotion of invasion [Carragher and Frame, 2004]. Cell adhesion molecules were also observed to be enriched highly in higher ECM grades. These molecules are found on the cell surface and are involved in cell-cell and cell-ECM interaction. Although their enrichment was elevated, known cell-cell junctions like gap, adherens, and tight junctions do not show the same pattern of enrichment. In particular, tight junctions were enriched below 25% of the ECM-G4 group with lower enrichment scores, while both the percentage and enrichment score were higher for the ECM-G1. This suggests that ECM grade elevation may be positively correlated with EMT since tight junctions are epithelial markers and their presence is known to suppress EMT [Kyuno et al., 2021]. The TGF-β signaling pathway, which is another pathway closely related to EMT, was also highly enriched for higher ECM grades [Kim et al., 2020].

Among the apoptosis-related pathways, a very clear enrichment was present for the p53 signaling pathway. It has been shown that ECM can send out pro-survival messages, mainly via focal adhesion and cytoskeletal rearrangement, and this could in turn decrease p53-caused apoptosis [Li et al., 2003]. Ras and MAPK signaling pathways were enriched with ECM grades as well. These pathways are well-known pathways that conduct signals from the ECM to the nucleus and trigger activation of many genes related to a plethora of cellular processes [Molina and Adjei, 2006]. It can be speculated that as the ECM content is enriched within the TME, signaling across the cell becomes more stimulated and messages are passed through the nucleus constantly. This may also explain the elevated enrichment in transcriptional regula-

tion in the cancer pathway. Another pathway that shows a clear enrichment trend is the phospholipase D signaling pathway, which is mainly involved in exocytosis and vesicular trafficking. Enrichment of this pathway could be another contributor to the ECM-rich groups that show higher deposition of the synthesized ECM content to the cell exterior. Some immune-related pathways like Th1, Th2, Th17 cell differentiation, toll-like receptor, T and B cell receptor signaling, chemokine, and cytokine signaling pathways were also enriched, which indicates that the immune response patterns are changed as the ECM features are altered. Especially, the switch between Th1 and Th2 cell differentiation is important in tumor faith since Th1, in general, shows an anti-tumorigenic effect while Th2 shows a pro-tumorigenic effect [Basu et al., 2021]. Not surprisingly, pathways in cancer were also enriched for ECM-rich grades, suggesting that the ECM content for these groups is more capable for the maintenance of tumors.

The hippo signaling pathway also showed enrichment with elevated ECM grade. This pathway is shown to be related to the regulation of tissue-specific stem cells and takes a role in organ size control and tissue regeneration. Downstream mediators of this pathway are TFs, YAP, and TAZ, which are known to induce tumor development and recurrence [Mo et al., 2014, Luo et al., 2023, Liang et al., 2024]. The cellular senescence pathway was also enriched with the ECM grade. Cellular senescence, mainly induced by therapy, is shown to derive tumor dormancy. This dormancy enables cancer cells to evade treatment strategies, and these cells may later recover their proliferative nature and mediate the relapse of the disease [Pluquet et al., 2019, Saleh and Gewirtz, 2022]. It has been discussed that the composition, modification, and remodeling of the ECM, as well as their interaction and the secreted factors they hold, have an influence on cellular senescence mainly via β-catenin [Parker and Cox, 2020]. This transcription factor is a downstream effector of the Wnt signaling pathway, which also showed increased enrichment with the ECM grades, and this pathway is also known to be associated with CSCs [Ruan et al., 2020]. The overall reduced proportion of recurrence-free trend and significantly higher risk of short-term recurrence probability of ECM-G4 may be attributed to such enriched pathways (Figures 4.17b, 4.18b).

In addition to enriched pathways, those that were significantly altered between ECM groups were also identified, as shown in Figure 4.30. It can be seen that ECM-receptor interaction increases significantly between all ECM grades except for ECM-G3 and ECM-G4, as shown in Figure 4.30a. It has already been shown with the ECM scores that the ECM content and abundance are significantly different from each other, and with this enrichment result, it can be said that the altered ECM profiles are complemented by their receptors. Another significantly differential pathway was the focal adhesion pathway among all ECM grades again except for ECM-G3 and ECM-G4, as shown in Figure 4.30b. As mentioned earlier, focal adhesions are the structures that maintain cell-ECM interaction and transduce mechanical cues to the nucleus via actin filaments. The focal adhesion complex consists of proteins like vinculin, α-actinin, talin, FAK, and integrins [Rikitake and Takai, 2011]. Especially, FAK is a very important player in cell migration and is shown to protect cells from anoikis [McLean et al., 2005]. Therefore, enrichment of this pathway indicates that the invasiveness capability of tumor cells is higher for increased ECM grade, which is consistent with our findings using the invasiveness score (Figure 4.16b). PI3K-Akt signaling was another enriched pathway significantly different between ECM groups except for again ECM-G3 and ECM-G4, as shown in Figure 4.30c. PI3K is known to be a hub for mechanotransduction. High ECM deposition, especially HA, which swells, observed in the TME exerts mechanical stress onto tumor cells and triggers the canonical PI3K/Akt pathway, which in turn promotes cell migration and tumor growth [Di-Luoffo et al., 2021]. Therefore, observing more enrichment in higher ECM grades is understandable since elevated ECM content would provide more stimuli for this pathway. Enrichment of IL17 signaling was significantly less in the ECM-G1 group compared to others, as shown in Figure 4.30d. IL17 is a signature cytokine of specialized CD4+ T cells called Th17 cells, which have a dual role as a tumor promoter and suppressor, and it has also been reported that they can exhibit immunosuppressive features [Chang, 2019, Basu et al., 2021]. Th17 cell differentiation was also less enriched in ECM-G1 (Figure 4.29). These suggest that a switch in the Th17 cellular program may be observed with altered ECM and higher grade groups may be more capable of escaping immune surveillance.

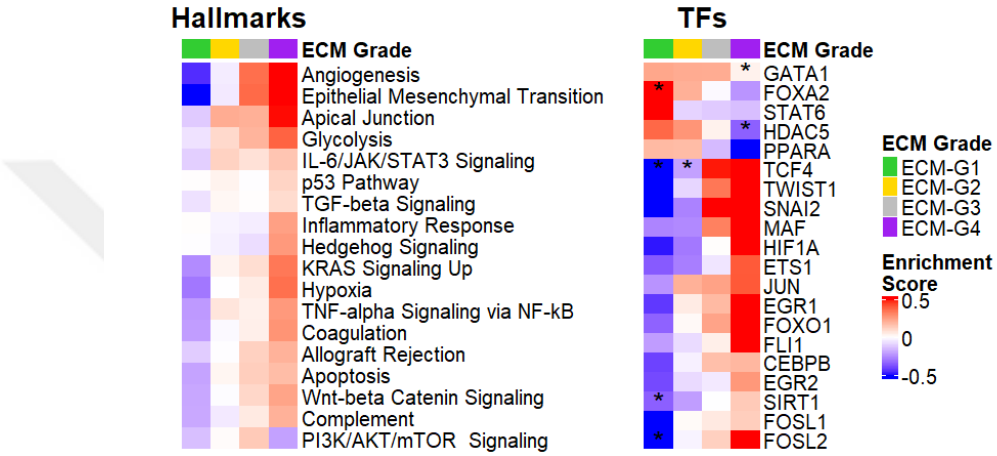


(a) (b) (c) (d)

Figure 4.30: Differentially Enriched KEGG Pathways: (a) ECM-receptor interaction, (b) focal adhesion, (c) PI3K-Akt signaling pathway, (d) IL-17 signaling pathway.

In addition to KEGG pathways, MSigDB hallmarks were also used for enrichment analysis, as shown in Figure 4.31a. Some of the enriched KEGG pathways (Figure 4.30) were also present in this enrichment analysis, like p53, TGF-β, TNF-α, NFκB, Wnt, and PI3K signaling. The inflammatory response hallmark enriched in these results was also implied by the immune-related pathways revealed from the previous enrichment analysis. Both the inflammatory response hallmark and related KEGG pathways were observed to have an increased enrichment trend with ECM grade, and it has been shown that tumor-related inflammation can promote immunosuppression and support tumor development [Ma et al., 2013]. These inflammation patterns may also be attributed to the significantly higher abundance of CAFs observed for higher ECM grades (Figure 4.13a) since CAFs are also known to produce many chemokines and cytokines like CXCL1, CCL-5, and IL6, as well as TGF-β [Özdemir et al., 2014, Yang et al., 2010]. This may also be one of the reasons why the IL16/JAK/STAT3 signaling hallmark was also observed to be highly enriched for ECM grades of higher levels, and this pathway is a very well-known signal transduction pathway that is involved in tumor development [Huang et al., 2022]. Hedgehog signaling hallmark showed the highest enrichment in ECM-G4,

and its enrichment in other groups was rather low. This pathway is a well-known pathway that is profoundly associated with CSCs and cancer therapy resistance [Giroux-Leprieur et al., 2018]. Here it can be understood that ECM-G4 seems to provide the best environment for the maintenance of CSC, which is also consistent with our findings using the CSC score (Figure 4.16c).



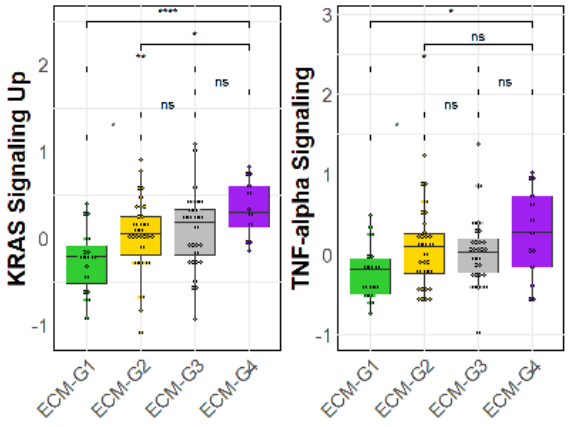
(a) (b)

Figure 4.31: Hallmark and TF Enrichment Analysis: : "*" shows TFs enriched in less than 25% of the patients in an ECM grade.

Enriched TFs were also determined to understand the key regulators that give rise to the ECM grades, shown in Figure 4.31b. A strikingly decreased TF for ECM-rich groups was HDAC5, which is a histone deacetylase and is known to modulate the binding of TFs to DNA molecules [Yang et al., 2021]. This may be one of the reasons transcriptional misregulation in the cancer pathway was more enriched as the ECM grade was elevated (Figure 4.29). Another such TF was PPARA, which is a nuclear receptor that regulates inflammation and lipid metabolism, and its activation has been shown to inhibit tumorigenesis [Lakshmi et al., 2017]. So, enrichment of this TF in lower ECM grades may be one of the reasons milder tumor aggression was observed in these groups. ETS1 enrichment was increased for ECM high groups, and higher expression of this TF is correlated with poor prognosis and EMT by regulating TWIST1 expression [Yamaguchi et al., 2007, Millien et al., 2018].

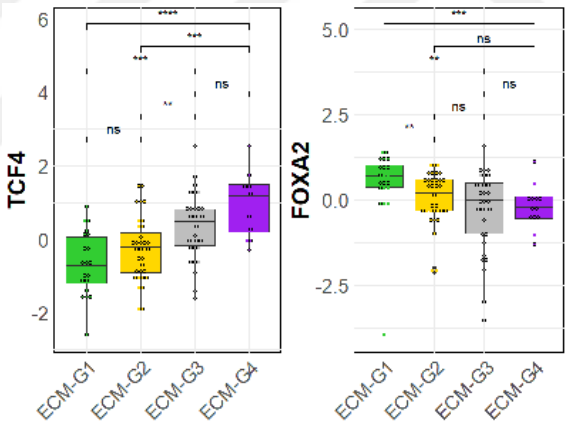
To emphasize hallmarks that are significantly different among ECM groups, enrichment scores were compared. Hallmarks, KRAS signaling, TNF-α signaling via NFκB, EMT, angiogenesis, and hypoxia were differentially enriched. KRAS signaling enrichment was significantly enriched in ECM-G4, while ECM-G2 and ECM-G3 showed similar enrichments, as shown in Figure 4.32a. It was also shown from the mutation profiles that KRAS mutations were highly accumulated on ECM-rich grades (Figure 4.20). It has been shown that KRAS mutant cancer cells can play a great part in the modulation of ECM structure and composition, mainly by remodeling it via MMPs [Dias Carvalho et al., 2018]. It can be speculated that mutated KRAS expression may be one of the mediators for richer ECM content. TNF-α signaling via NFκB was also significantly less enriched in ECM low grades, as shown in Figure 4.32b. This was in accordance with the trends observed for both TNF and NFκB signaling pathways given in enrichment results for KEGG pathways (Figure 4.29). TNF-α is a cytokine involved in many cellular processes, and the binding of this cytokine to its receptor, mainly TNFR2, triggers a cascade of events that ultimately activate the TF NFκB. This pathway is mainly involved in inflammation but also cell survival, proliferation, and death [Wang and Lin, 2008, Wu and Zhou, 2010, Alim et al., 2024]. Therefore, it can be said that ECM-rich groups are more prone to cancer-related inflammation, cell survival, and proliferation.

As for the TFs, TCF4, FOXA2, TWIST1, SNAI2, and HIF1A were differentially enriched, as were other TFs, FOSL1, FOSL2, FLI1, and MAF, which are given in Appendix Figure A.4. The TCF4 TF showed a significant increase in enrichment across all ECM grades, with the exception of ECM-G1 and ECM-G2, and ECM-G3 and ECM-G4, as shown in Figure 4.33a. This TF is known to be involved with ECM remodeling via controlling the regulation of various MMPs as well as Wnt signaling, and its high expression was shown to lead to poor prognosis in LUAD patients [Bleckmann et al., 2012, Greb-Markiewicz et al., 2019, Marchenko et al., 2002]. So, it can be said that enrichment of this TF may contribute to the higher ECM score calculated for ECM regulators in ECM-rich groups, and it explains the elevated invasive capability observed for these groups.



(a) (b)

Figure 4.32: Hallmarks KRAS and TNF-α Signaling: (a) KRAS signaling up, (b) TNF-α signaling via NFκB.



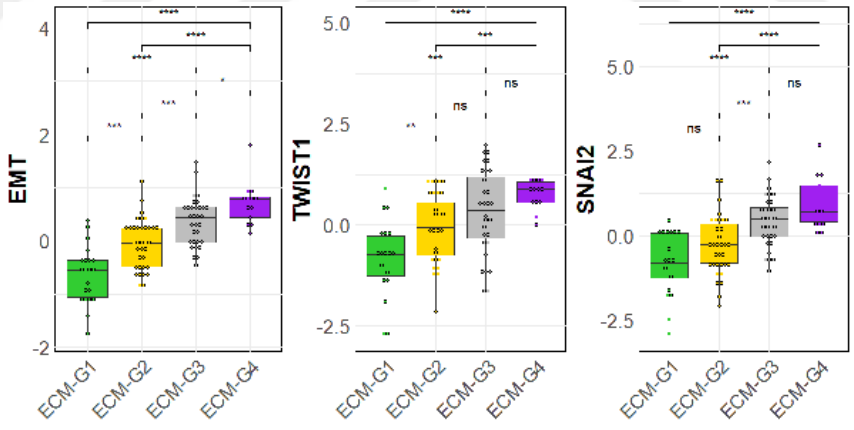
(a) (b)

Figure 4.33: TFs TCF4 and FOXA2: (a) TCF4, (b) FOXA2

FOXA2 was one of the enriched TFs that showed highly reduced enrichment in ECM-rich groups and was significantly enriched in ECM-G1, as shown in Figure 4.33b. High expression of this TF was shown to be correlated with suppression of invasion and TGF-β-induced EMT, and its loss has been shown to promote EMT and invasiveness in lung cancer cell lines [Tang et al., 2010, Li et al., 2015]. It has also been shown that the expression of this TF in EGFR mutant lung tumors resulted in suppression of tumor development, while in KRAS mutant cases it gave more

invasive tumors, suggesting that its tumor suppression and invasiveness-supporting nature can be dependent on the mutation profile of the lung tumor [Tomoshige et al., 2023]. We have shown that ECM grades of higher extent had more KRAS mutations while poor ECM groups held more EGFR mutations. Considering the duality of FOXA2 TF with these two mutation profiles, our calculation on invasiveness directly correlates with the literature (Figures 4.16b, 4.20).

It is known that ECM can be highly influential on EMT, and in that regard, we have observed that enrichment of the EMT hallmark was significantly and discretely increased as the ECM grades were elevated, as shown in Figure 4.34a. In agreement with these findings, the two most well-known EMT markers, TFs TWIST1 and SNAI2, were also significantly enriched, as shown in Figure 4.34b for TWIST1 and Figure 4.34c for SNAI2. This is consistent with our calculations using the EMT score (Figure 4.16a). So, it can be said that the EMT potential of patients within the ECM group of higher ECM grade is greater than those in poor ECM groups.



(a) (b) (c)

Figure 4.34: Hallmarks and TFs Related with EMT: (a) EMT hallmark, (b) TWIST TF, (c) SNAI2 TF.

It was discussed that elevated ECM content may mediate hypoxia and activate HIF1A which is a known TF that regulates VEGF expression [Faloppi et al., 2016]. This may then induce angiogenesis. In this context, the enrichment of angiogenesis and hypoxia hallmarks was observed to be significantly and almost discretely

increased with increasing ECM grade, as shown in Figure 4.35a and Figure 4.35b respectively. HIF1A enrichment was also significantly enriched in ECM-G4 compared to others, as shown in Figure 4.35c. Here we can say that the abundant ECM composition in ECM high grades indeed may result in hypoxic conditions, and this may be followed by the activation of pathways that result in angiogenesis, making these groups more prone to metastasis.

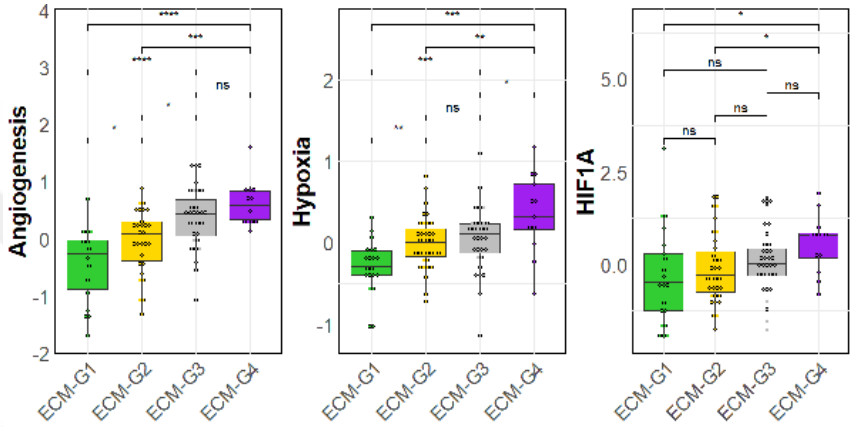
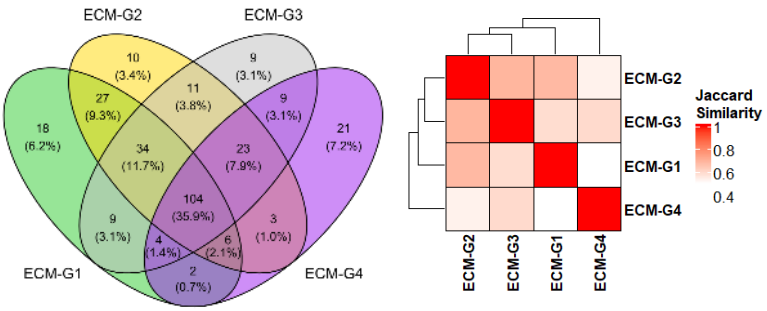


Figure 4.35: Hallmarks and TFs Related with Angiogenesis and Hypoxia: (a) angiogenesis hallmark, (b) hypoxia hallmark, (c) HIF1A TF.

4.8 Drug Targets to Consensus Network Proximity Analysis

To obtain networks that best represent the ECM grades, consensus networks were created by merging the patient networks of ECM grades. While doing so, edges that show frequency below 30% of the ECM grade population were filtered out to obtain the most representative consensus networks. Node sizes of consensus networks can be seen in Figure 4.36a. The consensus network of ECM-G2 held the highest node count of 218, followed by ECM-G1 with 204, ECM-G3 with 203, and ECM-G4 with 172. All of the groups shared 104 nodes. These consensus networks also revealed ECM grade-specific nodes that were not shared by any other ECM group.



(a) (b)

Figure 4.36: Comparison of Consensus Networks: (a) Venn diagram, (b) pair-wise Jaccard index.

Consensus networks of ECM grades were then compared with each other using Jaccard similarity, as shown in Figure 4.36b. It was observed that the most similar groups were ECM-G2 and ECM-G3, followed by ECM-G1 and ECM-G2, followed by ECM-G4 and ECM-G3. ECM-G1 and ECM-G2 were least similar to ECM-G4, and ECM-G4 and ECM-G3 were least similar to ECM-G4, which is consistent with our findings presented in Figure 4.28. Here we can conclude that consensus networks can preserve group-specific features while also capturing the similarities ECM groups have with each other.

The consensus network of ECM-G4 with 172 nodes and 203 edges can be seen in Figure 4.37. The shape of the nodes represents the category of the nodes; the red color represents the terminal, and grey represents the Steiner nodes. The thickness and transparency of the edges represent decreasing frequency and confidence, respectively. Some of the pathways and hallmarks enriched in ECM-G4 were shown to be represented within the consensus network by some relevant nodes. For example, NF-κB signaling by NFKBIZ, NFKB1, and NFKBIA, TGF-β signaling by TGFB1, the p53 pathway with TP53, Wnt signaling by CTNNB1, and hypoxia with HIF1A. Many EMT-related proteins can also be observed within the network, like TWIST1, TWIST2, SNAI2, MMP2, and FN1, as well as invasiveness-related proteins like COL1A1, VCAN, and FBN1. Stemness-related proteins can also be observed, like KLF4, SOX2, FOXA2, and POU5F1. It should be emphasized that

some of the proteins mentioned that represent the enriched pathways are intermediate nodes that come from the interactome by the network modeling algorithm. This shows that the modeling strategy could fill in the blanks and reveal relevant players that could not be identified by high-throughput technologies. It is also important to mention that the construction of the consensus network could represent the solution of the algorithm for patients without distortion and did not result in the loss of valuable and consistent information gathered from patient networks.

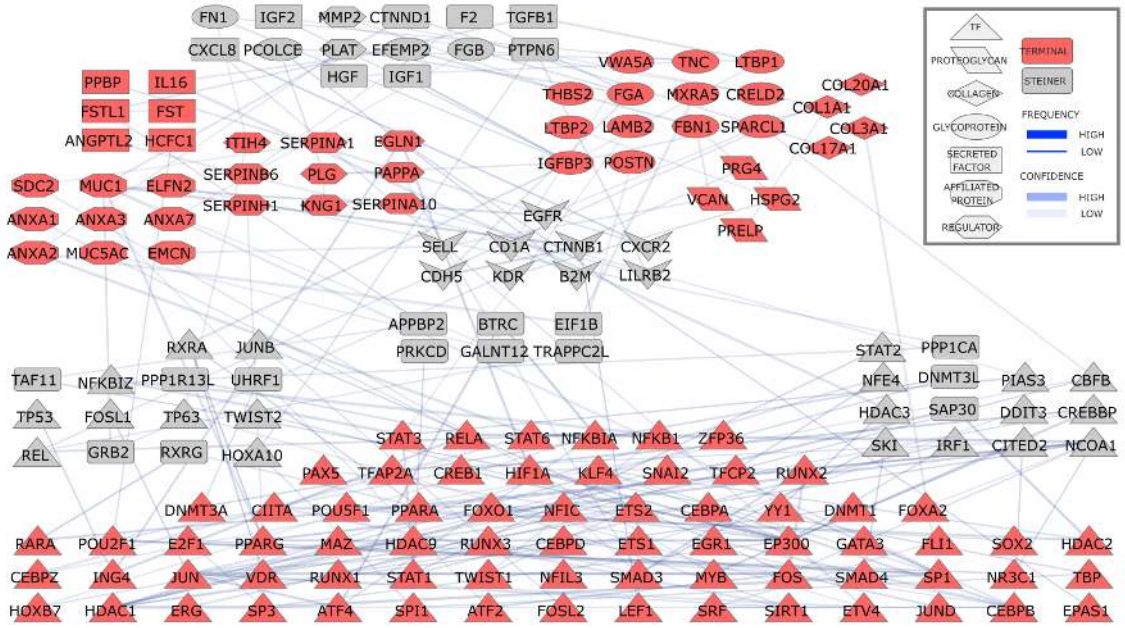


Figure 4.37: Consensus Network of ECM-G4

Consensus networks of ECM grades were then used for network proximity analysis to identify potential drugs that could target these groups. The resulting list of drug targets-to-network proximities was refined to show drugs within high-affinity inhibitor drugs curated from OncoTarget analysis [Mundi et al., 2023]. This subset was further filtered to show drugs that had a z-score below -2 for at most 2 groups and those that showed a continuous trend of proximity values between ECM grades. This filtering revealed cluster-specific drugs as well as large-spectrum drugs with possibly higher efficiencies on specific ECM groups, as shown in Figure 4.38.

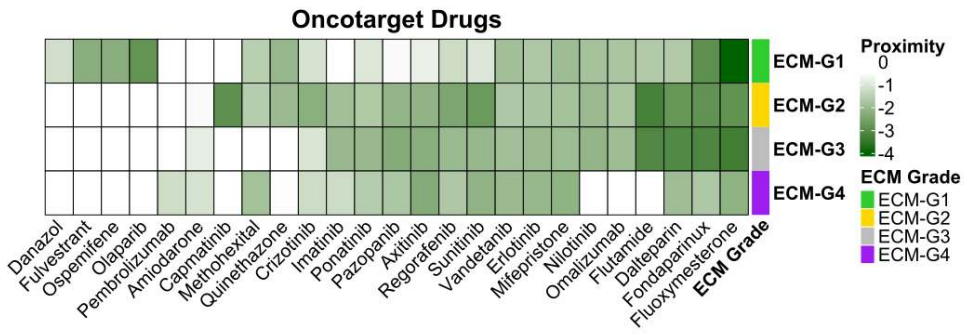


Figure 4.38: Proximate OncoTarget Drugs to ECM Grades

Danazol, Fluvestrant, Ospemifene, and Olaparib were in close proximity with ECM-G1 while showing no proximity to others. In particular, Olaparid was significantly proximate to this group with a z-score of -2.85. This drug is an approved inhibitor of poly (ADP-ribose) polymerase (PARP), which is known to be used in several cancers and mainly acts on PARP1 and PARP2 [DrugBank, 2024d]. Both of these proteins were present in the consensus network of ECM-G1, while only PARP1 was present in ECM-G2, and none of them were present in the remaining groups. This could explain the significantly higher proximity of this drug to ECM-G1. Capmatinib was significantly proximate to ECM-G2 with a z-score of -2.92, while it showed no proximity to others. This drug is an approved inhibitor of c-Met receptor tyrosine kinase (MET) that is known to be used in NSCLC [DrugBank, 2024b]. This kinase was only present in ECM-G2, which explains the proximity of this drug to this group. ECM-G3 did have a unique drug yet, Amiodarone, an anti-arrhythmic drug, was proximate to ECM-G3 and ECM-G4 with a relative proximity value of -0.87 and -1.08, respectively. Ponatinib, which is a multiple tyrosine kinase inhibitor, showed significant proximity to ECM-G3 with a z-score of -2.02, while z-scores for ECM-G1, ECM-G2, and ECM-G4 were -0.97, -1.67, and -1.58, respectively. So, it can be suggested that the use of this drug on patients of this group may be more effective than on patients of other groups. Flutamide, which is a nonsteroidal antiandrogen used mainly in metastatic prostate cancer, was significantly proximate to ECM-G2 and ECM-G3 with z-scores -3.26 and -3.12, respectively, since both

contained the well-known targets of this drug, AR and AHR [DrugBank, 2024c]. This suggests that Flutamide may also be more effective in these groups than the remaining. Finally, for ECM-G4, Pembrolizumab was found to be the only drug with a high relative proximity of -1.21. This approved biotech drug is an antibody that blocks PD-1 and is known to be used for the treatment of various cancers like metastatic melanoma and NSCLC [DrugBank, 2024e]. In addition to cluster-specific drugs, drugs that show high proximity to all of the groups were observed, like Erlotinib and Vandetanib, which are both tyrosine kinase inhibitors mainly targeting EGFR and VEGFR, respectively [Gao et al., 2012, Middleton et al., 2017]. All in all, this analysis shows that ECM grades may be targeted by specific drug molecules, or some of the drugs, although shared among the groups, may be more effective on some than others.

Chapter 5

CONCLUSION

The ECM of the TME is highly altered when compared with normal tissues and is highly influential on cancer proliferation, angiogenesis, metastasis, and immunosuppression [Yuan et al., 2023]. Therefore, expression profiles of the ECM may be highly important in predicting patient prognosis and developing strategies for treatment. These alterations are not always the same for every tumor, and different profiles may present distinct outcomes in tumor development [Malandrino et al., 2018]. So, categorizing patients based on their ECM characteristics can be beneficial for further understanding the impact of ECM on the TME as well as discovering personalized treatment strategies rather than high-spectrum. LUAD being one of the most common types of NSCLC, revealing ECM profiles for LUAD patients can be very important for predicting their survival.

Analysis of the multi-omic data in this study showed that indeed, expression and abundance of ECM components were highly altered and more heterogeneous in tumor samples compared to their paired NAT. Barcoding and clustering of patients gave four groups named ECM-G1, ECM-G2, ECM-G3, and ECM-G4, showing a discretely and significantly increasing pattern of ECM richness. Independently identifying highly variable genes for these groups further proved that these groups were profoundly capable of capturing the alterations in ECM components. Investigation of cellular heterogeneity and tumor purity revealed increased stromal cell content, including CAFs and endothelial cells, in richer ECM grades. A significantly higher CSC score of ECM-G4 from correlation analysis also suggests that this group might maintain CSCs. This analysis also showed that EMT and invasiveness of tumors in ECM-rich groups were significantly higher, and attachment of ECM groups of low enrichment to the ECM via integrins was significantly less than others. Clinical interpretation of these clusters showed that groups with higher ECM richness

had lower survival and higher recurrence probability, especially in the short term. Significantly higher short-term recurrence probability can be attributed to the maintenance of CSCs in the ECM-G4 group. Other clinical features showed that expression of cell-proliferative genes, hypermethylation of the DNA, and aggressiveness of tumors were much more observed in high-grade ECM groups. The mutation load of higher ECM grades was also much more than lower grades, and EGFR-KRAS mutual exclusiveness was at the ECM grade level, where lower ECM grades had a higher mutation count for EGFR while higher ECM grades had higher KRAS mutations. Network representation of the patients revealed hidden players that could not be detected from the multi-omic data and were central to many cellular processes. This enabled the completion of signaling cascades that resulted in the creation of different ECM profiles observed for patients. Subsequent enrichment analyses were done on these complete frameworks, and some of the pathways enriched in higher ECM grades were consistent with the previous findings. For instance, ECM-receptor interactions, focal adhesion pathways with the integrin score comparison or hedgehog, and Wnt signaling pathways with the CSC score comparison and recurrence probability. In addition to this, TF enrichment analysis also revealed enriched TF, which was very in synergy with the results of pathway and hallmark enrichment analysis. For instance, TCF4 TF was known to be an active player in Wnt signaling, and FOXA2 TF was known to suppress TGF-β-induced EMT. Yet, the most consistent and apparent pathways enriched with increasing ECM severity were EMT and angiogenesis via hypoxia, which was also supported by the enrichment of relevant TFs, TWIST1, SNAI2, and HIF1A. Drug screening done using proximity analysis of drug targets and consensus networks identified possible hits specific for ECM grades, such as olaparib for ECM-G1, Capmatinib for ECM-G2, Ponatinib for ECM-G3, and Pembrolizumab for ECM-G4, along with some other high-spectrum drugs that could be more effective on specific ECM grades.

We have investigated the similarities and differences of ECM grades regarding their clinical features, expression levels of cancer-related genesets, cellular heterogeneity, and tumor purity. We have also conducted both patient-network level and consensus network level similarity analyses on ECM grades. First, ECM scores,

which indicate the richness of ECM components, were all significantly different between every patient group. These differences were in an increasing step-wise trend, suggesting that these groups represent ECM grades rather than configurations, so it might be possible to observe patients within the transient of serial ECM grades. Consistently, the distance of ECM-G1 to ECM-G2, ECM-G2 to ECM-G3, and ECM-G3 to ECM-G4 was quite short in PCA, suggesting that although ECM groups are distinct, some groups may be more similar to each other than others, especially intermediate groups ECM-G2 and ECM-G3. Indeed, such differences and similarities were observed throughout the analyses; for instance, the invasiveness scores of ECM-G3 and ECM-G4 did not show a significant difference, suggesting that ECM-G3 is more similar to ECM-G4 regarding invasiveness. Although some similarities were observed between some of the groups, even ECM-G2 and ECM-G4, almost no similarity or shared features were observed for ECM-G1 and ECM-G4, suggesting that there are two poles of the ECM gradient, an ECM poor and an ECM rich, respectively. We could capture the uniqueness of ECM grades while preserving the similarities they might have with each other with our patient network modeling strategy. The network score comparison clearly showed overall the ECM groups that resembled each other while those that were most dissimilar. We also showed similarities between ECM grades more in-depth with the similarity scores, and a clear and significant difference was observed between ECM-G1 and ECM-G4. These relative similarities, to some extent, were preserved for the consensus networks. All in all, barcoding, consensus clustering, patient-network modeling, and consensus network construction strategies we have applied could present distinct ECM grades while preserving the similarities they may have with each other.

There are various studies that aim to find ECM groups among different cancers or identify ECM signatures that could be used as prognostic biomarkers [Zeng et al., 2022, Yu et al., 2021]. Yet, these studies utilize only one omic layer losing valuable information preserved in other omic layers and they mostly limit their scope to several ECM-related proteins to categorize patients as high and low-risk groups. To our knowledge, this study is the first to integrate multi-omic data for the investigation of ECM groups and reveal the underlying signaling pathways using network biology.

Although we could group patients successfully, we faced some limitations. The goal of this study was to use multi-omic data, but available cohorts on LUAD like TCGA only consisted of transcriptomic data, although the size of the cohort was around 510 samples. Most of these samples were also not paired with healthy tissue; therefore, we went with the CPTAC data. Yet, this cohort only had 101 patients with paired NAT samples, giving us a rather small set of data to work with. Also, survival data of this cohort was released recently and had too much censored data, making it difficult to speculate on patient prognosis.

In conclusion, we could group LUAD patients based on their ECM characteristics and represent these patients with network modeling using multi-omic data. This approach enabled us to show biological processes and pathways that give rise to different ECM grades, further define the similarities and differences ECM grades have, and finally identify key molecular mechanisms altered with ECM grades that can be targeted with different drugs, which can be potential candidates for personalized treatment strategies.

BIBLIOGRAPHY

[Aboubakar Nana et al., 2019] Aboubakar Nana, F., Vanderputten, M., and Ocak, S. (2019). Role of focal adhesion kinase in small-cell lung cancer and its potential as a therapeutic target. *Cancers*, 11(11):1683.

[Ahmed et al., 2019] Ahmed, R., Baali, I., Erten, C., Hoxha, E., and Kazan, H. (2019). Mexcowalk: mutual exclusion and coverage based random walk to identify cancer modules. *Bioinformatics*, 36(3):872–879.

[Alanis-Lobato et al., 2016] Alanis-Lobato, G., Andrade-Navarro, M. A., and Schaefer, M. H. (2016). Hippie v2.0: enhancing meaningfulness and reliability of protein–protein interaction networks. *Nucleic Acids Research*, 45(D1):D408–D414.

[Alim et al., 2024] Alim, L. F., Keane, C., and Souza-Fonseca-Guimaraes, F. (2024). Molecular mechanisms of tumour necrosis factor signalling via tnf receptor 1 and tnf receptor 2 in the tumour microenvironment. *Current Opinion in Immunology*, 86:102409.

[Alizadeh et al., 2015] Alizadeh, A. A., Aranda, V., Bardelli, A., Blanpain, C., Bock, C., and et al., B. (2015). Toward understanding and exploiting tumor heterogeneity. *Nature Medicine*, 21(8):846–853.

[Alonso-Nocelo et al., 2018] Alonso-Nocelo, M., Raimondo, T. M., Vining, K. H., López-López, R., de la Fuente, M., and Mooney, D. J. (2018). Matrix stiffness and tumor-associated macrophages modulate epithelial to mesenchymal transition of human adenocarcinoma cells. *Biofabrication*, 10(3):035004.

[Alvarez et al., 2016] Alvarez, M. J., Shen, Y., Giorgi, F. M., Lachmann, A., Ding, B. B., Ye, B. H., and Califano, A. (2016). Functional characterization of somatic

mutations in cancer using network-based inference of protein activity. *Nature Genetics*, 48(8):838–847.

[Amos and Choi, 2021] Amos, S. E. and Choi, Y. S. (2021). The cancer microenvironment: Mechanical challenges of the metastatic cascade. *Frontiers in Bioengineering and Biotechnology*, 9.

[Ando et al., 2019] Ando, T., Kage, H., Matsumoto, Y., Zokumasu, K., Yotsumoto, T., and et al., M. (2019). Integrin (
alpha
)11 in non–small cell lung cancer is associated with tumor progression and postoperative recurrence. *Cancer Science*, 111(1):200–208.

[André et al., 2017] André, F., Arnedos, M., Baras, A. S., Baselga, J., Bedard, P. L., Berger, M. F., and et al., B. (2017). Aacr project genie: Powering precision medicine through an international consortium. *Cancer Discovery*, 7(8):818–831.

[Aramini et al., 2022] Aramini, B., Masciale, V., Grisendi, G., Bertolini, F., and et al., M. (2022). Dissecting tumor growth: The role of cancer stem cells in drug resistance and recurrence. *Cancers*, 14(4):976.

[Aran et al., 2017] Aran, D., Hu, Z., and Butte, A. J. (2017). xcell: digitally portraying the tissue cellular heterogeneity landscape. *Genome Biology*, 18(1).

[Barabási and Pósfai, 2016] Barabási, A. and Pósfai, M. (2016). *Network Science*. Cambridge University Press.

[Barabási et al., 2010] Barabási, A.-L., Gulbahce, N., and Loscalzo, J. (2010). Network medicine: a network-based approach to human disease. *Nature Reviews Genetics*, 12(1):56–68.

[Barkovskaya et al., 2020] Barkovskaya, A., Buffone, A., Zidek, M., and Weaver, V. M. (2020). Proteoglycans as mediators of cancer tissue mechanics. *Frontiers in Cell and Developmental Biology*, 8.

[Basu et al., 2021] Basu, A., Ramamoorthi, G., Albert, G., Gallen, C., Beyer, A., Snyder, C., Koski, G., Disis, M. L., Czerniecki, B. J., and Kodumudi, K. (2021). Differentiation and regulation of th cells: A balancing act for cancer immunotherapy. *Frontiers in Immunology*, 12.

[Becht et al., 2016] Becht, E., Giraldo, N. A., Lacroix, L., Buttard, B., Elarouci, N., and et al., P. (2016). Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. *Genome Biology*, 17(1).

[Ben-Hamo and Efroni, 2013] Ben-Hamo, R. and Efroni, S. (2013). Network as biomarker: Quantifying transcriptional co-expression to stratify cancer clinical phenotypes. *Systems Biomedicine*, 1(1):35–41.

[Bergamaschi et al., 2008] Bergamaschi, A., Tagliabue, E., Sørlie, T., Naume, B., Triulzi, T., and et al., O. (2008). Extracellular matrix signature identifies breast cancer subgroups with different clinical outcome. *The Journal of Pathology*, 214(3):357–367.

[Bleckmann et al., 2012] Bleckmann, A., Siam, L., Klemm, F., Rietkötter, E., Wegner, C., Kramer, F., Beissbarth, T., Binder, C., Stadelmann, C., and Pukrop, T. (2012). Nuclear lef1/tcf4 correlate with poor prognosis but not with nuclear β-catenin in cerebral metastasis of lung adenocarcinomas. *Clinical & Experimental Metastasis*, 30(4):471–482.

[Bonnans et al., 2014] Bonnans, C., Chou, J., and Werb, Z. (2014). Remodelling the extracellular matrix in development and disease. *Nature Reviews Molecular Cell Biology*, 15(12):786–801.

[Boucher and Jenna, 2013] Boucher, B. and Jenna, S. (2013). Genetic interaction networks: better understand to better predict. *Frontiers in Genetics*, 4.

[Breuer et al., 2012] Breuer, K., Foroushani, A. K., Laird, M. R., Chen, C., Sribnaia, A., and et al., L. (2012). Innatedb: systems biology of innate immunity

and beyond—recent updates and continuing curation. *Nucleic Acids Research*, 41(D1):D1228–D1233.

[Carragher and Frame, 2004] Carragher, N. O. and Frame, M. C. (2004). Focal adhesion and actin dynamics: a place where kinases and proteases meet to promote invasion. *Trends in Cell Biology*, 14(5):241–249.

[Chang, 2019] Chang, S. H. (2019). T helper 17 (th17) cells and interleukin-17 (il-17) in cancer. *Archives of Pharmacal Research*, 42(7):549–559.

[Chen et al., 2013] Chen, E. Y., Tan, C. M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G. V., Clark, N. R., and Ma'ayan, A. (2013). Enrichr: interactive and collaborative html5 gene list enrichment analysis tool. *BMC Bioinformatics*, 14(1).

[Chen et al., 2016] Chen, M. B., Lamar, J. M., Li, R., Hynes, R. O., and Kamm, R. D. (2016). Elucidation of the roles of tumor integrin $\beta 1$ in the extravasation stage of the metastasis cascade. *Cancer Research*, 76(9):2513–2524.

[Chuang et al., 2022] Chuang, H.-H., Zhen, Y.-Y., Tsai, Y.-C., Chuang, C.-H., Hsiao, M., Huang, M.-S., and Yang, C.-J. (2022). Fak in cancer: From mechanisms to therapeutic strategies. *International Journal of Molecular Sciences*, 23(3):1726.

[Chung et al., 2020] Chung, V., Alistar, A., George, B., Kim, K., Kindler, H., Oh, D., Allen, S., and et al., B. (2020). So-4 phase ib/ii, open-label, randomised evaluation of atezolizumab plus ro6874281 vs control in morpheus–pancreatic ductal adenocarcinoma. *Annals of Oncology*, 31:S218.

[Coelho and McCulloch, 2018] Coelho, N. M. and McCulloch, C. A. (2018). Mechanical signaling through the discoidin domain receptor 1 plays a central role in tissue fibrosis. *Cell Adhesion & Migration*, page 1–15.

[Cordes and Beinke, 2004] Cordes, N. and Beinke, C. (2004). Fibronectin alters cell

survival and intracellular signaling of confluent a549 cultures after irradiation. *Cancer Biology & Therapy*, 3(1):47–53.

[Costa et al., 2008] Costa, L. d. F., Rodrigues, F. A., and Cristino, A. S. (2008). Complex networks: the key to systems biology. *Genetics and Molecular Biology*, 31(3):591–601.

[Cox, 2021] Cox, T. R. (2021). The matrix in cancer. *Nature Reviews Cancer*, 21(4):217–238.

[Cui et al., 2021] Cui, X., Liu, R., Duan, L., Cao, D., Zhang, Q., and Zhang, A. (2021). Car-t therapy: Prospects in targeting cancer stem cells. *Journal of Cellular and Molecular Medicine*, 25(21):9891–9904.

[Damle and Köhn, 2019] Damle, N. P. and Köhn, M. (2019). The human dephosphorylation database depod: 2019 update. *Database*, 2019.

[Database, 2024] Database, C. (2024). Feature 45: Summary. Accessed: 2024-07-22.

[Demirel et al., 2022] Demirel, H. C., Arici, M. K., and Tuncbag, N. (2022). Computational approaches leveraging integrated connections of multi-omic data toward clinical applications. *Molecular Omics*, 18(1):7–18.

[Di-Luoffo et al., 2021] Di-Luoffo, M., Ben-Meriem, Z., Lefebvre, P., Delarue, M., and Guillermet-Guibert, J. (2021). Pi3k functions as a hub in mechanotransduction. *Trends in Biochemical Sciences*, 46(11):878–888.

[Dias Carvalho et al., 2018] Dias Carvalho, P., Guimarães, C. F., Cardoso, A. P., Mendonça, S., Costa, u. M., Oliveira, M. J., and Velho, S. (2018). Kras oncogenic signaling extends beyond cancer cells to orchestrate the microenvironment. *Cancer Research*, 78(1):7–14.

[Ding et al., 2008] Ding, L., Getz, G., Wheeler, D. A., Mardis, E. R., McLellan, M. D., and et al., C. (2008). Somatic mutations affect key pathways in lung adenocarcinoma. *Nature*, 455(7216):1069–1075.

[Dräger and Planatscher, 2013] Dräger, A. and Planatscher, H. (2013). *Metabolic Networks*, page 1249–1251. Springer New York.

[DrugBank, 2024a] DrugBank (2024a). About drugbank. Accessed: 2024-07-03.

[DrugBank, 2024b] DrugBank (2024b). Capmatinib. `https://go.drugbank.com/drugs/DB11791`. Accessed: 2024-07-27.

[DrugBank, 2024c] DrugBank (2024c). Flutamide. DrugBank Online. Accessed: 2024-07-27.

[DrugBank, 2024d] DrugBank (2024d). Olaparib. `https://go.drugbank.com/drugs/DB09074`. Accessed: 2024-07-27.

[DrugBank, 2024e] DrugBank (2024e). Pembrolizumab. DrugBank Online. Accessed: 2024-07-27.

[Faloppi et al., 2016] Faloppi, L., Casadei Gardini, A., Masi, G., Silvestris, N., Loretelli, C., Ulivi, P., Vivaldi, C., and et al., B. (2016). Angiogenesis polymorphisms profile in the prediction of clinical outcome of advanced hcc patients receiving sorafenib: Combined analysis of vegf and hif-1α—final results of the alice-2 study. *Journal of Clinical Oncology*, 34(4_suppl):280–280.

[Fang et al., 2013] Fang, M., Yuan, J., Peng, C., and Li, Y. (2013). Collagen as a double-edged sword in tumor progression. *Tumor Biology*, 35(4):2871–2882.

[Fares et al., 2020] Fares, J., Fares, M. Y., Khachfe, H. H., Salhab, H. A., and Fares, Y. (2020). Molecular principles of metastasis: a hallmark of cancer revisited. *Signal Transduction and Targeted Therapy*, 5(1).

[Feinberg et al., 2018] Feinberg, T. Y., Zheng, H., Liu, R., Wicha, M. S., Yu, S. M., and Weiss, S. J. (2018). Divergent matrix-remodeling strategies distinguish developmental from neoplastic mammary epithelial cell invasion programs. *Developmental Cell*, 47(2):145–160.e6.

[Finotello et al., 2019] Finotello, F., Mayer, C., Plattner, C., Laschober, G., Rieder, D., and et al., H. (2019). Molecular and pharmacological modulators of the tumor immune contexture revealed by deconvolution of rna-seq data. *Genome Medicine*, 11(1).

[Gao et al., 2012] Gao, H., Ding, X., Wei, D., Cheng, P., Su, X., Liu, H., Aziz, F., Wang, D., and Zhang, T. (2012). Erlotinib in patients with advanced non-small-cell lung cancer: A meta-analysis. *Translational Lung Cancer Research*, 1(2).

[Garcia-Alonso et al., 2019] Garcia-Alonso, L., Holland, C. H., Ibrahim, M. M., Turei, D., and Saez-Rodriguez, J. (2019). Benchmark and integration of resources for the estimation of human transcription factor activities. *Genome Research*, 29(8):1363–1375.

[Gillette et al., 2020] Gillette, M. A., Satpathy, S., Cao, S., Dhanasekaran, S. M., Vasaikar, S. V., and et al., K. (2020). Proteogenomic characterization reveals therapeutic vulnerabilities in lung adenocarcinoma. *Cell*, 182(1):200–225.e35.

[Giroux-Leprieur et al., 2018] Giroux-Leprieur, E., Costantini, A., Ding, V. W., and He, B. (2018). Hedgehog signaling in lung cancer: From oncogenesis to cancer treatment resistance. *International Journal of Molecular Sciences*, 19(9):2835.

[Greb-Markiewicz et al., 2019] Greb-Markiewicz, B., Kazana, W., Zarebski, M., and Ozyhar, A. (2019). The subcellular localization of bhlh transcription factor tcf4 is mediated by multiple nuclear localization and nuclear export signals. *Scientific Reports*, 9(1).

[Grossman et al., 2016] Grossman, R. L., Heath, A. P., Ferretti, V., Varmus, H. E., Lowy, D. R., Kibbe, W. A., and Staudt, L. M. (2016). Toward a shared vision for cancer genomic data. *New England Journal of Medicine*, 375(12):1109–1112.

[Guney et al., 2016] Guney, E., Menche, J., Vidal, M., and Barábasi, A.-L. (2016). Network-based in silico drug efficacy screening. *Nature Communications*, 7(1).

[Han et al., 2017] Han, H., Cho, J.-W., Lee, S., Yun, A., Kim, H., and et al., B. (2017). Trrust v2: an expanded reference database of human and mouse transcriptional regulatory interactions. *Nucleic Acids Research*, 46(D1):D380–D386.

[Hassanein et al., 2021] Hassanein, S. S., Abdel-Mawgood, A. L., and Ibrahim, S. A. (2021). Egfr-dependent extracellular matrix protein interactions might light a candle in cell behavior of non-small cell lung cancer. *Frontiers in Oncology*, 11.

[Heath et al., 2006] Heath, E. I., Burtness, B. A., Kleinberg, L., Salem, R. R., Yang, S. C., and et al., H. (2006). Phase ii, parallel-design study of preoperative combined modality therapy and the matrix metalloprotease (mmp) inhibitor prinomastat in patients with esophageal adenocarcinoma. *Investigational New Drugs*, 24(2):135–140.

[Heldin et al., 2014] Heldin, P., Basu, K., Kozlova, I., and Porsch, H. (2014). *HAS2 and CD44 in Breast Tumorigenesis*, page 211–229. Elsevier.

[Henke et al., 2020] Henke, E., Nandigama, R., and Ergün, S. (2020). Extracellular matrix in the tumor microenvironment and its impact on cancer therapy. *Frontiers in Molecular Biosciences*, 6.

[Hornbeck et al., 2014] Hornbeck, P. V., Zhang, B., Murray, B., Kornhauser, J. M., Latham, V., and Skrzypek, E. (2014). Phosphositeplus, 2014: mutations, ptms and recalibrations. *Nucleic Acids Research*, 43(D1):D512–D520.

[Huang et al., 2022] Huang, B., Lang, X., and Li, X. (2022). The role of il-6/jak2/stat3 signaling pathway in cancers. *Frontiers in Oncology*, 12.

[Huang et al., 2021] Huang, J., Zhang, L., Wan, D., Zhou, L., Zheng, S., Lin, S., and Qiao, Y. (2021). Extracellular matrix and its therapeutic potential for cancer treatment. *Signal Transduction and Targeted Therapy*, 6(1).

[Ideker et al., 2002] Ideker, T., Ozier, O., Schwikowski, B., and Siegel, A. F. (2002).

Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics*, 18(suppl_1):S233–S240.

[Ingber, 2008] Ingber, D. E. (2008). Can cancer be reversed by engineering the tumor microenvironment? *Seminars in Cancer Biology*, 18(5):356–364.

[Ivanisevic and Sewduth, 2023] Ivanisevic, T. and Sewduth, R. N. (2023). Multi-omics integration for the design of novel therapies and the identification of novel biomarkers. *Proteomes*, 11(4):34.

[Jethalia et al., 2023] Jethalia, M., Jani, S. P., Ceccarelli, M., and Mall, R. (2023). Pancancer network analysis reveals key master regulators for cancer invasiveness. *Journal of Translational Medicine*, 21(1).

[Jiang et al., 2018] Jiang, Z., Zhou, C., Cheng, L., Yan, B., Chen, K., and et al., C. (2018). Inhibiting yap expression suppresses pancreatic cancer progression by disrupting tumor-stromal interactions. *Journal of Experimental & Clinical Cancer Research*, 37(1).

[John et al., 2020] John, C. R., Watson, D., Russ, D., Goldmann, K., Ehrenstein, M., Pitzalis, C., Lewis, M., and Barnes, M. (2020). M3c: Monte carlo reference-based consensus clustering. *Scientific Reports*, 10(1).

[Kamburov et al., 2012] Kamburov, A., Stelzl, U., and Herwig, R. (2012). Intscore: a web tool for confidence scoring of biological interactions. *Nucleic Acids Research*, 40(W1):W140–W146.

[Kanehisa et al., 2022] Kanehisa, M., Furumichi, M., Sato, Y., Kawashima, M., and Ishiguro-Watanabe, M. (2022). Kegg for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Research*, 51(D1):D587–D592.

[Kessenbrock et al., 2010] Kessenbrock, K., Plaks, V., and Werb, Z. (2010). Matrix metalloproteinases: Regulators of the tumor microenvironment. *Cell*, 141(1):52–67.

[Kim et al., 2020] Kim, B. N., Ahn, D. H., Kang, N., Yeo, C. D., Kim, Y. K., Lee, K. Y., Kim, T.-J., and et al., L. (2020). Tgf-β induced emt and stemness characteristics are associated with epigenetic regulation in lung cancer. *Scientific Reports*, 10(1).

[Kim and Reckamp, 2024] Kim, E. S. and Reckamp, K. L. (2024). Lung adenocarcinoma. `https://www.ncbi.nlm.nih.gov/books/NBK519578/`. Accessed: 2024-07-19.

[Kim et al., 2011] Kim, S.-H., Turnbull, J., and Guimond, S. (2011). Extracellular matrix and cell signalling: the dynamic cooperation of integrin, proteoglycan and growth factor receptor. *Journal of Endocrinology*, 209(2):139–151.

[Kuşoğlu et al., 2024] Kuşoğlu, A., Örnek, D., Dansık, A., Uzun, C., Nur Özkan, S., Sarıca, S., Yangın, K., Özdinç, S., Sorhun, D. T., Solcan, N., Doğanalp, Tuncbag, N., and Öztürk, E. e. a. (2024). Extracellular matrix sulfation in the tumor microenvironment stimulates cancer stemness and invasiveness. *Advanced Science*.

[Kyuno et al., 2021] Kyuno, D., Takasawa, A., Kikuchi, S., Takemasa, I., Osanai, M., and Kojima, T. (2021). Role of tight junctions in the epithelial-to-mesenchymal transition of cancer cells. *Biochimica et Biophysica Acta (BBA) - Biomembranes*, 1863(3):183503.

[Lakshmi et al., 2017] Lakshmi, S. P., Reddy, A. T., Banno, A., and Reddy, R. C. (2017). Ppar agonists for the prevention and treatment of lung cancer. *PPAR Research*, 2017:1–8.

[Leiserson et al., 2014] Leiserson, M. D. M., Vandin, F., Wu, H.-T., Dobson, J. R., Eldridge, J. V., and et al., T. (2014). Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nature Genetics*, 47(2):106–114.

[Lendorf et al., 2011] Lendorf, M. E., Manon-Jensen, T., Kronqvist, P., Multhaupt, H. A. B., and Couchman, J. R. (2011). Syndecan-1 and syndecan-4 are indepen-

dent indicators in breast carcinoma. *Journal of Histochemistry & Cytochemistry*, 59(6):615–629.

[Levi et al., 2021] Levi, H., Elkon, R., and Shamir, R. (2021). Domino: a network-based active module identification algorithm with reduced rate of false calls. *Molecular Systems Biology*, 17(1).

[Li et al., 2016] Li, B., Severson, E., Pignon, J.-C., Zhao, H., Li, T., and et al., N. (2016). Comprehensive analyses of tumor immunity: implications for cancer immunotherapy. *Genome Biology*, 17(1).

[Li et al., 2015] Li, C. M.-C., Gocheva, V., Oudin, M. J., Bhutkar, A., Wang, S. Y., Date, S. R., Ng, S. R., Whittaker, C. A., Bronson, R. T., Snyder, E. L., Gertler, F. B., and Jacks, T. (2015). Foxa2 and cdx2 cooperate with nkx2-1 to inhibit lung adenocarcinoma metastasis. *Genes &; Development*, 29(17):1850–1862.

[Li et al., 2017] Li, J., Chen, Y., Guo, X., Zhou, L., Jia, Z., and et al., P. (2017). ¡scp¿gpc¡/scp¿1 exosome and its regulatory mi¡scp¿rna¡/scp¿s are specific markers for the detection and target therapy of colorectal cancer. *Journal of Cellular and Molecular Medicine*, 21(5):838–847.

[Li et al., 2003] Li, L.-N., Wang, D.-R., Sato, M., Kojima, N., Imai, K., Higashi, N., and Senoo, H. (2003). Extracellular matrix-regulated p53 expression and nuclear localization in cultured detroit 562 cells derived from pharyngeal carcinoma. *Archives of Histology and Cytology*, 66(5):419–428.

[Liang et al., 2024] Liang, H., Xu, Y., Zhao, J., Chen, M., and Wang, M. (2024). Hippo pathway in non-small cell lung cancer: mechanisms, potential targets, and biomarkers. *Cancer Gene Therapy*, 31(5):652–666.

[Liberzon et al., 2015] Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J., and Tamayo, P. (2015). The molecular signatures database hallmark gene set collection. *Cell Systems*, 1(6):417–425.

[Liu and Agarwal, 2010] Liu, J. and Agarwal, S. (2010). Mechanical signals activate vascular endothelial growth factor receptor-2 to upregulate endothelial cell proliferation during inflammation. *The Journal of Immunology*, 185(2):1215–1221.

[Liu et al., 2012a] Liu, J., Liao, S., Diop-Frimpong, B., Chen, W., Goel, S., and et al., N. (2012a). Tgf-β blockade improves the distribution and efficacy of therapeutics in breast carcinoma by normalizing the tumor stroma. *Proceedings of the National Academy of Sciences*, 109(41):16618–16623.

[Liu et al., 2020] Liu, W., Sun, X., Peng, L., Zhou, L., Lin, H., and Jiang, Y. (2020). Rwrnet: A gene regulatory network inference algorithm using random walk with restart. *Frontiers in Genetics*, 11.

[Liu et al., 2012b] Liu, X.-Q., Xiong, M.-H., Shu, X.-T., Tang, R.-Z., and Wang, J. (2012b). Therapeutic delivery of sirna silencing hif-1 alpha with micellar nanoparticles inhibits hypoxic tumor growth. *Molecular Pharmaceutics*, 9(10):2863–2874.

[Liu et al., 2023] Liu, Z.-L., Chen, H.-H., Zheng, L.-L., Sun, L.-P., and Shi, L. (2023). Angiogenic signaling pathways and anti-angiogenic therapy for cancer. *Signal Transduction and Targeted Therapy*, 8(1).

[Lu et al., 2023] Lu, C., Liu, Y., Ali, N. M., Zhang, B., and Cui, X. (2023). The role of innate immune cells in the tumor microenvironment and research progress in anti-tumor therapy. *Frontiers in Immunology*, 13.

[Lu et al., 2011] Lu, P., Takai, K., Weaver, V. M., and Werb, Z. (2011). Extracellular matrix degradation and remodeling in development and disease. *Cold Spring Harbor Perspectives in Biology*, 3(12):a005058–a005058.

[Luo et al., 2023] Luo, J., Deng, L., Zou, H., Guo, Y., Tong, T., Huang, M., Ling, G., and Li, P. (2023). New insights into the ambivalent role of yap/taz in human cancers. *Journal of Experimental & Clinical Cancer Research*, 42(1).

[Ma et al., 2013] Ma, Y., Adjemian, S., Mattarollo, S., Yamazaki, T., Aymeric, L., Yang, H., Portela Catani, J., Hannani, D., Duret, H., Steegh, K., Martins, I., Schlemmer, F., and et al., M. (2013). Anticancer chemotherapy-induced intratumoral recruitment and differentiation of antigen-presenting cells. *Immunity*, 38(4):729–741.

[Malandrino et al., 2018] Malandrino, A., Mak, M., Kamm, R. D., and Moeendarbary, E. (2018). Complex mechanics of the heterogeneous extracellular matrix in cancer. *Extreme Mechanics Letters*, 21:25–34.

[Marchenko et al., 2002] Marchenko, G. N., MARCHENKO, N. D., LENG, J., and STRONGIN, A. Y. (2002). Promoter characterization of the novel human matrix metalloproteinase-26 gene: regulation by the t-cell factor-4 implies specific expression of the gene in cancer cells of epithelial origin. *Biochemical Journal*, 363(2):253–262.

[Matsumoto et al., 2011] Matsumoto, S., Batra, S., Saito, K., Yasui, H., Choudhuri, R., and et al., G. (2011). Antiangiogenic agent sunitinib transiently increases tumor oxygenation and suppresses cycling hypoxia. *Cancer Research*, 71(20):6350–6359.

[May et al., 2023] May, L., Shows, K., Nana-Sinkam, P., Li, H., and Landry, J. W. (2023). Sex differences in lung cancer. *Cancers (Basel)*, 15(12):3111.

[McLean et al., 2005] McLean, G. W., Carragher, N. O., Avizienyte, E., Evans, J., Brunton, V. G., and Frame, M. C. (2005). The role of focal-adhesion kinase in cancer — a new therapeutic opportunity. *Nature Reviews Cancer*, 5(7):505–515.

[Middleton et al., 2017] Middleton, G., Palmer, D. H., Greenhalf, W., Ghaneh, P., Jackson, R., Cox, T., Evans, A., Shaw, V. E., and et al., W. (2017). Vandetanib plus gemcitabine versus placebo plus gemcitabine in locally advanced or metastatic pancreatic carcinoma (vip): a prospective, randomised, double-blind, multicentre phase 2 trial. *The Lancet Oncology*, 18(4):486–499.

[Mierke, 2024] Mierke, C. T. (2024). Extracellular matrix cues regulate mechanosensing and mechanotransduction of cancer cells. *Cells*, 13(1):96.

[Millien et al., 2018] Millien, G., Cao, Y., O'Hara, C. J., Tagne, J.-B., Hinds, A., Williams, M. C., Ramirez, M. I., and Kathuria, H. (2018). Ets1 regulates twist1 transcription in a $kras^{G12D}/lkb1^{-/-}$ metastatic lung tumor model of non-small cell lung cancer. *Clinical & Experimental Metastasis*, 35(3):149–165.

[Misra et al., 2019] Misra, B. B., Langefeld, C., Olivier, M., and Cox, L. A. (2019). Integrated omics: tools, advances and future approaches. *Journal of Molecular Endocrinology*, 62(1):R21–R45.

[Mo et al., 2014] Mo, J., Park, H. W., and Guan, K. (2014). The hippo signaling pathway in stem cell biology and cancer. *EMBO reports*, 15(6):642–656.

[Molina and Adjei, 2006] Molina, J. R. and Adjei, A. A. (2006). The ras/raf/mapk pathway. *Journal of Thoracic Oncology*, 1(1):7–9.

[Monti, 2003] Monti, S. (2003). Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data. *Machine Learning*, 52(1/2):91–118.

[Morrissey and Sherwood, 2015] Morrissey, M. A. and Sherwood, D. R. (2015). An active role for basement membrane assembly and modification in tissue sculpting. *Journal of Cell Science*.

[Mundi et al., 2023] Mundi, P. S., Dela Cruz, F. S., Grunn, A., Diolaiti, D., Mauguen, A., Rainey, A. R., Guillan, K., Siddiquee, A., and et al., Y. (2023). A transcriptome-based precision oncology platform for patient–therapy alignment in a diverse set of treatment-resistant malignancies. *Cancer Discovery*, 13(6):1386–1407.

[Naba et al., 2012a] Naba, A., Clauser, K. R., Hoersch, S., Liu, H., Carr, S. A., and Hynes, R. O. (2012a). The matrisome: In silico definition and in vivo character-

ization by proteomics of normal and tumor extracellular matrices. *Molecular & Cellular Proteomics*, 11(4):M111.014647.

[Naba et al., 2012b] Naba, A., Hoersch, S., and Hynes, R. O. (2012b). Towards definition of an ecm parts list: An advance on go categories. *Matrix Biology*, 31(7–8):371–372.

[Newman et al., 2015] Newman, A. M., Liu, C. L., Green, M. R., Gentles, A. J., Feng, W., Xu, Y., Hoang, C. D., Diehn, M., and Alizadeh, A. A. (2015). Robust enumeration of cell subsets from tissue expression profiles. *Nature Methods*, 12(5):453–457.

[Ning and Li, 2023] Ning, K. and Li, Y. (2023). *Introduction to Multi-Omics*, page 1–10. Springer Nature Singapore.

[Orchard et al., 2013] Orchard, S., Ammari, M., Aranda, B., Breuza, L., Briganti, L., and et al., B.-C. (2013). The mintact project—intact as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Research*, 42(D1):D358–D363.

[Oughtred et al., 2020] Oughtred, R., Rust, J., Chang, C., Breitkreutz, B., Stark, C., and et al., W. (2020). The ¡scp¿biogrid¡/scp¿ database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Science*, 30(1):187–200.

[Özdemir et al., 2014] Özdemir, B., Pentcheva-Hoang, T., Carstens, J., Zheng, X., Wu, C.-C., Simpson, T., and et al., L. (2014). Depletion of carcinoma-associated fibroblasts and fibrosis induces immunosuppression and accelerates pancreas cancer with reduced survival. *Cancer Cell*, 25(6):719–734.

[Pan et al., 2005] Pan, Q., Pao, W., and Ladanyi, M. (2005). Rapid polymerase chain reaction-based detection of epidermal growth factor receptor gene mutations in lung adenocarcinomas. *The Journal of Molecular Diagnostics*, 7(3):396–403.

[Papillon-Cavanagh et al., 2020] Papillon-Cavanagh, S., Doshi, P., Dobrin, R., Szustakowski, J., and Walsh, A. M. (2020). Stk11 and keap1 mutations as prognostic biomarkers in an observational real-world lung adenocarcinoma cohort. *ESMO Open*, 5(2):e000706.

[Parker et al., 2022] Parker, A. L., Bowman, E., Zingone, A., Ryan, B. M., Cooper, W. A., Kohonen-Corish, M., Harris, C. C., and Cox, T. R. (2022). Extracellular matrix profiles determine risk and prognosis of the squamous cell carcinoma subtype of non-small cell lung carcinoma. *Genome Medicine*, 14(1).

[Parker and Cox, 2020] Parker, A. L. and Cox, T. R. (2020). The role of the ecm in lung cancer dormancy and outgrowth. *Frontiers in Oncology*, 10.

[Parks et al., 2004] Parks, W. C., Wilson, C. L., and López-Boado, Y. S. (2004). Matrix metalloproteinases as modulators of inflammation and innate immunity. *Nature Reviews Immunology*, 4(8):617–629.

[Pluquet et al., 2019] Pluquet, O., Abbadie, C., and Coqueret, O. (2019). Connecting cancer relapse with senescence. *Cancer Letters*, 463:50–58.

[Puttock et al., 2023] Puttock, E. H., Tyler, E. J., Manni, M., Maniati, E., Butterworth, C., and et al., B. R. (2023). Extracellular matrix educates an immunoregulatory tumor macrophage phenotype found in ovarian cancer metastasis. *Nature Communications*, 14(1).

[Quail and Joyce, 2013] Quail, D. F. and Joyce, J. A. (2013). Microenvironmental regulation of tumor progression and metastasis. *Nature Medicine*, 19(11):1423–1437.

[Racle et al., 2017] Racle, J., de Jonge, K., Baumgaertner, P., Speiser, D. E., and Gfeller, D. (2017). Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data. *eLife*, 6.

[Rangel et al., 2018] Rangel, M. P., de Sá, V. K., Prieto, T., Martins, J. R. M., Olivieri, E. R., and et al., C. (2018). Biomolecular analysis of matrix proteoglycans as biomarkers in non small cell lung cancer. *Glycoconjugate Journal*, 35(2):233–242.

[Razick et al., 2008] Razick, S., Magklaras, G., and Donaldson, I. M. (2008). irefindex: A consolidated protein interaction database with provenance. *BMC Bioinformatics*, 9(1).

[Ribatti et al., 2020] Ribatti, D., Tamma, R., and Annese, T. (2020). Epithelial-mesenchymal transition in cancer: A historical overview. *Translational Oncology*, 13(6):100773.

[Rich et al., 2010] Rich, J. T., Neely, J. G., Paniello, R. C., Voelker, C. C. J., Nussenbaum, B., and Wang, E. W. (2010). A practical guide to understanding kaplan-meier curves. *Otolaryngology–Head and Neck Surgery*, 143(3):331–336.

[Rikitake and Takai, 2011] Rikitake, Y. and Takai, Y. (2011). *Directional Cell Migration*, page 97–143. Elsevier.

[Ritz et al., 2016] Ritz, A., Poirel, C. L., Tegge, A. N., Sharp, N., Simmons, K., Powell, A., Kale, S. D., and Murali, T. (2016). Pathways on demand: automated reconstruction of human signaling networks. *npj Systems Biology and Applications*, 2(1).

[Rual et al., 2005] Rual, J.-F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., and et al., L. (2005). Towards a proteome-scale map of the human protein–protein interaction network. *Nature*, 437(7062):1173–1178.

[Ruan et al., 2020] Ruan, Y., Ogana, H., Gang, E., Kim, H. N., and Kim, Y.-M. (2020). *Wnt Signaling in the Tumor Microenvironment*, page 107–121. Springer International Publishing.

[Ruepp et al., 2007] Ruepp, A., Brauner, B., Dunger-Kaltenbach, I., Frishman, G., Montrone, C., and et al., S. (2007). Corum: the comprehensive resource of mammalian protein complexes. *Nucleic Acids Research*, 36(Database):D646–D650.

[Sainio, 2013] Sainio, A. (2013). Extracellular matrix macromolecules in tumour microenvironment with special reference to desmoplastic reaction and the role of matrix proteoglycans and hyaluronan. *Journal of Carcinogenesis &a Mutagenesis*, S13.

[Saleh and Gewirtz, 2022] Saleh, T. and Gewirtz, D. A. (2022). Considering therapy-induced senescence as a mechanism of tumour dormancy contributing to disease recurrence. *British Journal of Cancer*, 126(10):1363–1365.

[Santiago-Rodriguez and Hollister, 2021] Santiago-Rodriguez, T. M. and Hollister, E. B. (2021). Multi 'omic data integration: A review of concepts, considerations, and approaches. *Seminars in Perinatology*, 45(6):151456.

[Schaaf et al., 2018] Schaaf, M. B., Garg, A. D., and Agostinis, P. (2018). Defining the role of the tumor vasculature in antitumor immunity and immunotherapy. *Cell Death & Disease*, 9(2).

[Sherman et al., 2014] Sherman, M., Yu, R., Engle, D., Ding, N., Atkins, A., and et.al, T. (2014). Vitamin d receptor-mediated stromal reprogramming suppresses pancreatitis and enhances pancreatic cancer therapy. *Cell*, 159(1):80–93.

[Sleeboom et al., 2024] Sleeboom, J. J. F., van Tienderen, G. S., Schenke-Layland, K., van der Laan, L. J. W., Khalil, A. A., and Verstegen, M. M. A. (2024). The extracellular matrix as hallmark of cancer and metastasis: From biomechanics to therapeutic targets. *Science Translational Medicine*, 16(728).

[Sokolova et al., 2020] Sokolova, O. A., Mikhaleva, E. A., Kharitonov, S. L., Abramov, Y. A., Gvozdev, V. A., and Klenov, M. S. (2020). Special vulnerability of somatic niche cells to transposable element activation in drosophila larval ovaries. *Scientific Reports*, 10(1).

[Song et al., 2022] Song, K., Yu, Z., Zu, X., Li, G., Hu, Z., and Xue, Y. (2022). Collagen remodeling along cancer progression providing a novel opportunity for cancer diagnosis and treatment. *International Journal of Molecular Sciences*, 23(18):10509.

[Sturm et al., 2020] Sturm, G., Finotello, F., and List, M. (2020). Immunedeconv: An r package for unified access to computational methods for estimating immune cell fractions from bulk rna-sequencing data. In Boegel, S., editor, *Bioinformatics for Cancer Immunotherapy*, volume 2120 of *Methods in Molecular Biology*, pages 223–232. Humana.

[Suda et al., 2010] Suda, K., Tomizawa, K., and Mitsudomi, T. (2010). Biological and clinical significance of kras mutations in lung cancer: an oncogenic driver that contrasts with egfr mutation. *Cancer and Metastasis Reviews*, 29(1):49–60.

[Sudlow et al., 2015] Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., and et al., D. (2015). Uk biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLOS Medicine*, 12(3):e1001779.

[Suh and Han, 2011] Suh, H. N. and Han, H. J. (2011). Collagen i regulates the self-renewal of mouse embryonic stem cells through $\alpha 2\beta 1$ integrin- and ddr1-dependent bmi-1. *Journal of Cellular Physiology*, 226(12):3422–3432.

[Szklarczyk et al., 2022] Szklarczyk, D., Kirsch, R., Koutrouli, M., Nastou, K., Mehryary, F., and et al., H. (2022). The string database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Research*, 51(D1):D638–D646.

[Tamiello et al., 2013] Tamiello, C., Kamps, M. A., van den Wijngaard, A., Verstraeten, V. L. R. M., Baaijens, F. P., Broers, J. L., and Bouten, C. C. (2013). Soft substrates normalize nuclear morphology and prevent nuclear rupture in fibrob-

lasts from a laminopathy patient with compound heterozygous lmna mutations. *Nucleus*, 4(1):61–73.

[Tang et al., 2010] Tang, Y., Shu, G., Yuan, X., Jing, N., and Song, J. (2010). Foxa2 functions as a suppressor of tumor metastasis by inhibition of epithelial-to-mesenchymal transition in human lung cancers. *Cell Research*, 21(2):316–326.

[Tarazona et al., 2021] Tarazona, S., Arzalluz-Luque, A., and Conesa, A. (2021). Undisclosed, unmet and neglected challenges in multi-omics studies. *Nature Computational Science*, 1(6):395–402.

[Tay et al., 2023] Tay, C., Tanaka, A., and Sakaguchi, S. (2023). Tumor-infiltrating regulatory t cells as targets of cancer immunotherapy. *Cancer Cell*, 41(3):450–465.

[Terry M. Therneau and Patricia M. Grambsch, 2000] Terry M. Therneau and Patricia M. Grambsch (2000). *Modeling Survival Data: Extending the Cox Model*. Springer, New York.

[Thangudu et al., 2020] Thangudu, R. R., Rudnick, P. A., Holck, M., Singhal, D., MacCoss, M. J., and et al., E. (2020). Abstract lb-242: Proteomic data commons: A resource for proteogenomic analysis. *Cancer Research*, 80(16_Supplement):LB–242–LB–242.

[Togashi et al., 2019] Togashi, Y., Shitara, K., and Nishikawa, H. (2019). Regulatory t cells in cancer immunosuppression—implications for anticancer therapy. *Nature Reviews Clinical Oncology*, 16(6):356–371.

[Tomoshige et al., 2023] Tomoshige, K., Stuart, W. D., Fink-Baldauf, I. M., Ito, M., Tsuchiya, T., Nagayasu, T., and et al., Y. (2023). Foxa2 cooperates with mutant kras to drive invasive mucinous adenocarcinoma of the lung. *Cancer Research*, 83(9):1443–1458.

[Tuncbag et al., 2013] Tuncbag, N., Braunstein, A., Pagnani, A., Huang, S.-S. C., Chayes, J., Borgs, C., Zecchina, R., and Fraenkel, E. (2013). Simultaneous re-

construction of multiple signaling pathways via the prize-collecting steiner forest problem. *Journal of Computational Biology*, 20(2):124–136.

[Tuncbag et al., 2016] Tuncbag, N., Gosline, S. J. C., Kedaigle, A., Soltis, A. R., Gitter, A., and Fraenkel, E. (2016). Network-based interpretation of diverse high-throughput datasets through the omics integrator software package. *PLOS Computational Biology*, 12(4):e1004879.

[Unsal-Beyge and Tuncbag, 2022] Unsal-Beyge, S. and Tuncbag, N. (2022). Functional stratification of cancer drugs through integrated network similarity. *npj Systems Biology and Applications*, 8(1).

[VIB, 2023] VIB (2023). irefindex: A comprehensive resource for protein-protein interaction data. `https://irefindex.vib.be/`. Accessed: 2024-07-03.

[Wang and Lin, 2008] Wang, X. and Lin, Y. (2008). Tumor necrosis factor and cancer, buddies or foes? *Acta Pharmacologica Sinica*, 29(11):1275–1288.

[Wei et al., 2015] Wei, S. C., Fattet, L., Tsai, J. H., Guo, Y., Pai, V. H., and et al., M. (2015). Matrix stiffness drives epithelial–mesenchymal transition and tumour metastasis through a twist1–g3bp2 mechanotransduction pathway. *Nature Cell Biology*, 17(5):678–688.

[Wishart et al., 2017] Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., and et al., G. (2017). Drugbank 5.0: a major update to the drugbank database for 2018. *Nucleic Acids Research*, 46(D1):D1074–D1082.

[Wu and Zhou, 2010] Wu, Y. and Zhou, B. P. (2010). Tnf-α/nf-κb/snail pathway in cancer cell migration and invasion. *British Journal of Cancer*, 102(4):639–644.

[Xie et al., 2023] Xie, H., Crawford, L., and Conard, A. M. (2023). Multioviz: an interactive platform forin silicoperturbation and interrogation of gene regulatory networks. *bioRxiv*.

[Yamaguchi et al., 2007] Yamaguchi, E., Nakayama, T., Nanashima, A., Matsumoto, K., Yasutake, T., Sekine, I., and Nagayasu, T. (2007). Ets-1 proto-oncogene as a potential predictor for poor prognosis of lung adenocarcinoma. *The Tohoku Journal of Experimental Medicine*, 213(1):41–50.

[Yang et al., 2021] Yang, J., Gong, C., Ke, Q., Fang, Z., Chen, X., Ye, M., and Xu, X. (2021). Insights into the function and clinical application of hdac5 in cancer management. *Frontiers in Oncology*, 11.

[Yang et al., 2010] Yang, L., Pang, Y., and Moses, H. L. (2010). Tgf-β and immune cells: an important regulatory axis in the tumor microenvironment and progression. *Trends in Immunology*, 31(6):220–227.

[Yoshihara et al., 2013] Yoshihara, K., Shahmoradgoli, M., Martínez, E., Vegesna, R., Kim, H., and et al., T.-G. (2013). Inferring tumour purity and stromal and immune cell admixture from expression data. *Nature Communications*, 4(1).

[Yu et al., 2021] Yu, C., You, M., Zhang, P., Zhang, S., Yin, Y., and Zhang, X. (2021). A five-gene signature is a prognostic biomarker in pan-cancer and related with immunologically associated extracellular matrix. *Cancer Medicine*, 10(13):4629–4643.

[Yuan et al., 2023] Yuan, Z., Li, Y., Zhang, S., Wang, X., Dou, H., Yu, X., Zhang, Z., Yang, S., and Xiao, M. (2023). Extracellular matrix remodeling in tumor progression and immune escape: from mechanisms to treatments. *Molecular Cancer*, 22(1).

[Yue, 2014] Yue, B. (2014). Biology of the extracellular matrix: An overview. *Journal of Glaucoma*, 23:S20–S23.

[Zeng et al., 2022] Zeng, Z., Zuo, Y., Jin, Y., Peng, Y., and Zhu, X. (2022). Identification of extracellular matrix signatures as novel potential prognostic biomarkers in lung adenocarcinoma. *Frontiers in Genetics*, 13.

[Zhao et al., 2023] Zhao, Y., Shen, M., Wu, L., Yang, H., Yao, Y., Yang, Q., Du, J., Liu, L., Li, Y., and Bai, Y. (2023). Stromal cells in the tumor microenvironment: accomplices of tumor progression? *Cell Death & Disease*, 14(9).

[Zhou et al., 2017] Zhou, F., Shang, W., Yu, X., and Tian, J. (2017). Glypican-3: A promising biomarker for hepatocellular carcinoma diagnosis and treatment. *Medicinal Research Reviews*, 38(2):741–767.

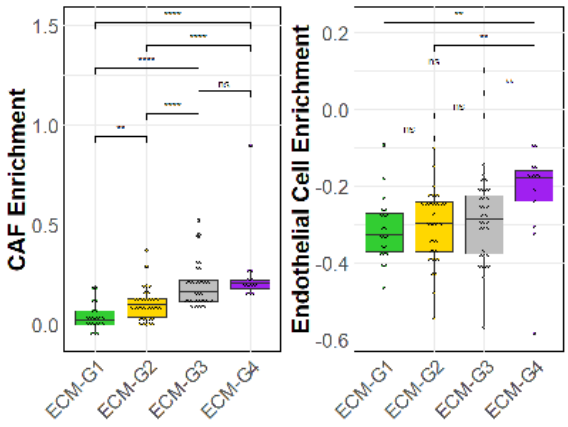
[Zitnik et al., 2023] Zitnik, M., Li, M. M., Wells, A., Glass, K., Gysi, D. M., and et al., K. (2023). Current and future directions in network biology.

[Șenbabaoğlu et al., 2014] Șenbabaoğlu, Y., Michailidis, G., and Li, J. Z. (2014). Critical limitations of consensus clustering in class discovery. *Scientific Reports*, 4(1).

Appendix A

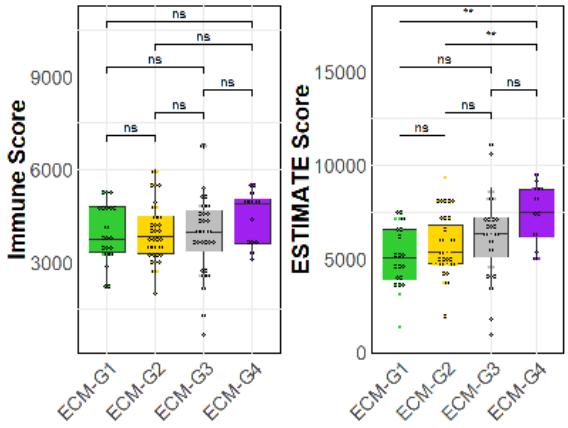
Table A.1: Patient IDs with ECM Grades.

ECM-Grade	Patient ID
ECM-G1	C3L.00422, C3L.02350, C3N.02582, C3L.00604, C3L.02348, C3L.00080, C3N.02529, C3L.00140, C3N.00546, C3N.02380, C3N.00223, C3N.02158, C3N.02002, C3N.00203, C3L.00510, C3N.02424, C3N.02089, C3L.01889, C3L.00893, C3N.01799, C3N.02587
ECM-G2	C3N.00580, C3L.00094, C3N.00560, C3N.00579, C3N.02588, C3L.01890, C3N.00704, C3N.01415, C3N.01410, C3N.01413, C3N.01071, C3N.01823, C3L.02219, C3L.00973, C3N.02067, C3N.02586, C3L.02549, C3N.02379, C3N.01024, C3N.00167, C3N.00550, C3L.00093, C3N.00572, C3L.01682, C3N.00959, C3N.02003, C3N.02087, C3L.00263, C3N.02155, C3L.00368, C3L.01330, C3L.00001, C3N.02433, C3N.00559, C3N.01414
ECM-G3	C3N.00738, C3L.02345, C3L.01632, C3L.02365, C3N.00293, C3N.02000, C3N.02145, C3N.01488, C3N.01405, C3N.02423, C3N.00217, C3N.00549, C3N.01030, C3N.00551, C3N.00199, C3N.01023, C3N.00547, C3L.01683, C3N.01072, C3N.00556, C3N.02421, C3N.00180, C3N.02572, C3N.00737, C3N.00552, C3N.02729, C3N.00578, C3N.00433, C3N.01416, C3L.01924, C3L.00144, C3N.00175
ECM-G4	C3N.00169, C3L.00913, C3L.00279, C3N.01016, C3N.02149, C3N.01489, C3L.00095, C3N.01021, C3N.00574, C3L.02508, C3L.00009, C3L.00083, C3L.00412



(a) (b)

Figure A.1: CAF and Endothelial Cell Enrichment of ECM Grades using EPIC: (a) CAF enrichment, (b) endothelial cell enrichment



(a) (b)

Figure A.2: ESTIMATE Scores of ECM Grades: (a) immune score, (b) ESTIMATE score.

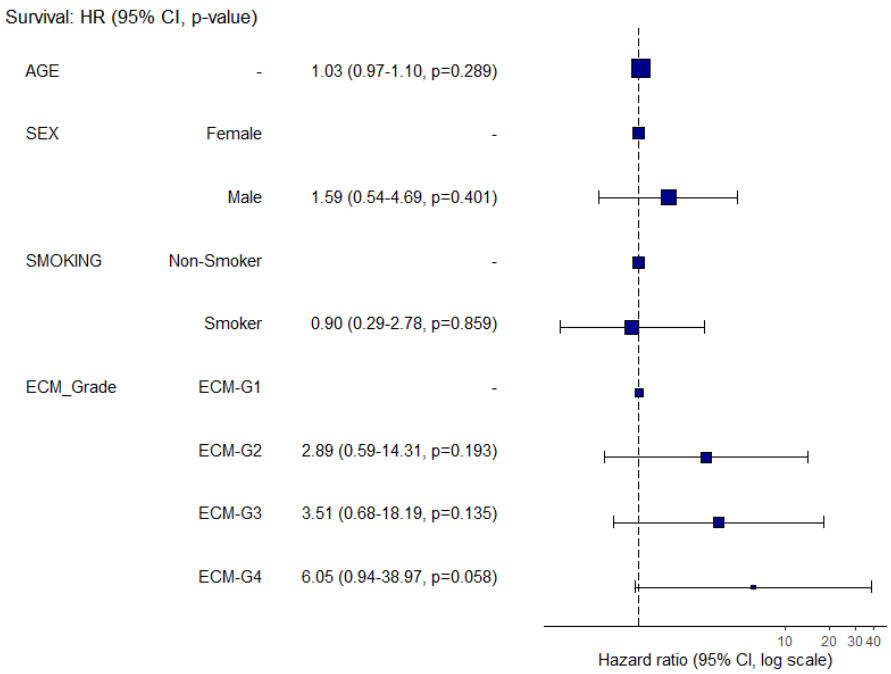
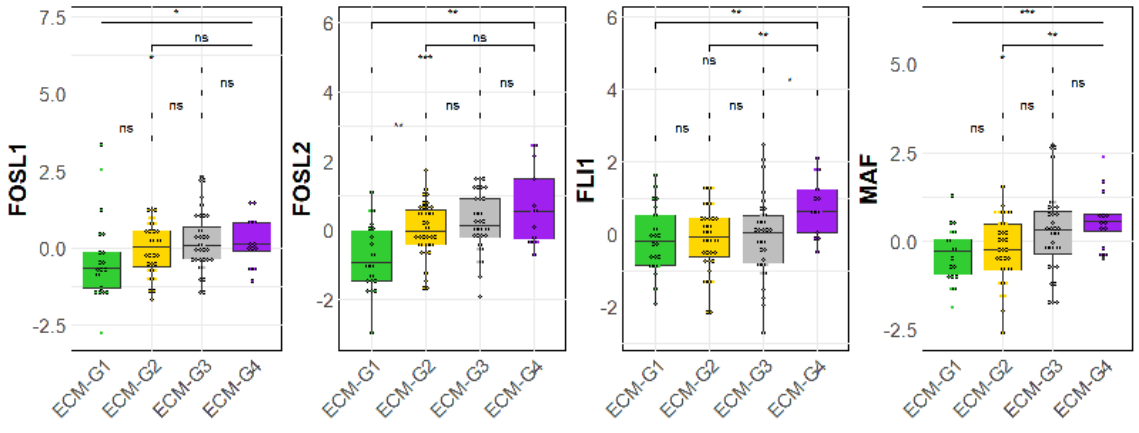


Figure A.3: Hazard Ratios of Multvariates



(a) (b) (c) (d)

Figure A.4: Other Enriched TFs