Inaugural-Dissertation

zur

Erlangung der Doktorwürde

der

Gesamtfakultät

für Mathematik, Ingenieur- und Naturwissenschaften

der

Ruprecht–Karls–Universität

Heidelberg

vorgelegt von

Julian Kreis, M.Sc.

aus:	Marburg
Tag der mündlichen Prüfung:	18.6.2024

Enhancing Cancer Diagnostics Through Quality-Focused Gene Expression Signature Analyses

Gutachter:	Prof. Dr. Benedikt Brors
	Prof. Dr. Claudia Scholl

Summary

Analyses of expression profiles between different phenotypes of cancer patient populations or models led to identifying gene modules, or gene expression signatures, that describe specific biological mechanisms. A signature is often used to describe the same mechanism in different contexts, for example, in large-scale cancer gene expression studies. Over the past, gene expression signatures have become readily available and collected in large signature databases like the MsigDB. Although gene expression signatures represent a critical knowledge base for the analysis of cancer populations, they are often carelessly applied or contaminated by confounding processes, such as proliferation. This might be why only a few gene expression signatures have reached clinical relevance.

Therefore, in the first chapter of this thesis, I aim to define a workflow that assesses the quality and translatability of gene expression signatures, thereby enabling the interpretation of their associated biological mechanisms in the context of another cancer study. Additionally, I provide an analytical methodology, to analyze a collection of gene expression signatures that describe a broad range of cancer mechanisms. Additionally, I describe a webapplication named RosettaSX that uses this methodology and allows users to analyze public molecular data of more than 11,000 cancer patients or cancer models. Finally, I will show the applicability of my approach by recapitulating of the intrinsic breast cancer subtypes and other molecular characteristics. This first project will lay the methodological foundation for two subsequent analyses.

In the second project, I use my approach to evaluate a set of gene expression signatures postulated to describe epithelial-to-mesenchymal, mesenchymal, or stemness phenotypes of cancer populations. Previous studies suggested that cells in the tumor microenvironment contribute to individual signatures in specific cancer types. In this study, I applied my methodological approach to analyze multiple levels of data granularity, including cancer cell line and single cell data, in addition to clinical tumor data. The goal was to highlight the impact of contamination on the largest combined set of mesenchymal signatures investigated to date in this context. This project emphasized the significance of thoroughly evaluating cancer cell content when utilizing these signatures. It also demonstrated how incorrect conclusions about cancer characteristics can be drawn when quality control is not rigorously applied in signature analyses.

The final chapter, will apply the methodological approach to evaluate the underdiagnosis of large cell neuroendocrine carcinomas (LCNEC) in a real-world data non-small cell carcinoma (NSCLC) cohort. Although LCNEC was not initially classified as a separate subtype of NSCLC, it has been classified as a separate group in the most recent WHO classification recommendation. The increased recognition of LCNEC over the past has shown an increase in the prevalence of this rare disease, accounting for 1%-3% of all lung cancers. However, today, practical limitations, similarities, and overlap with other NSCLCs are still complicating the diagnosis of LCNEC. Based on a RosettaSX analysis that revealed neuroendocrine differentiation in many patients with NSCLC, I will demonstrate how a machine learning model was used to assess the degree of LCNEC underdiagnosis.

In summary, this work presents an integrated approach for evaluating gene expression signatures in depth. The signature analysis framework was applied in two ways: for the in-depth assessment of gene expression signatures and the comprehensive characterization of a patient population. The approach described herein can provide robust findings for gene expression signatures that are easily interpretable and can reveal previously unknown associations between biomarkers and expression phenotypes.

Zusammenfassung

Durch die Analyse von transkriptionellen Unterschieden zwischen verschiedenen Phänotypen von Krebspatienten oder Krebsmodellen wurde eine Vielzahl von Genmodulen, auch Genexpressionssignaturen genannt, identifiziert, die charakteristische Expressionsprofile für einen biologischen Mechanismus aufweisen. Diese Signaturen werden häufig genutzt, um die Relevanz eines Mechanismus in einem unabhängigen Datensatz zu evaluieren. In den vergangenen Jahren sind große Datenbanken entstanden, wie z.B. MsigDB, die eine große Sammlung von Signaturen bereitstellen. Obwohl solche Signaturen eine wichtige Ressource für die Charakterisierung von Krebspatienten darstellen, werden sie häufig falsch angewendet oder sind von Genen kontaminiert, die in andere Prozesse, wie zum Beispiel Zellproliferation, involviert sind. Unter anderem aus diesen Gründen finden Signaturen nur selten ihren Weg in die klinische Anwendung.

Deshalb ist das Ziel der ersten Studie dieser Thesis die Entwicklung eines methodischen Ansatzes, der es ermöglicht, Genexpressionssignaturen in einem neuen Datensatz zu evaluieren. Dadurch soll der biologische Mechanismus, den die Signatur darstellt, beschrieben werden können. Zusätzlich stelle ich eine Plattform namens RosettaSX vor, welche es ermöglicht, mithilfe einer Sammlung von Signaturen, molekulare Phänotypen von 11.000 öffentlich verfügbaren Krebspatienten und Krebsmodellen zu analysieren. Um den Nutzen meines Analyseansatzes aufzuzeigen, beschreibe ich molekulare Merkmale von etablierten Brustkrebsuntertypen anhand eines bekannten Brustkrebs-Expressionsdatensatzes. Basierend auf den Methoden in diesem Teil der Arbeit führe ich anschließend zwei weitere Studien durch.

Im zweiten Kapitel beschreibe ich, wie ich meinen Ansatz für die Evaluierung einer Gruppe von Genexpressionssignaturen nutze. Diese Signaturen wurden in verschiedenen Krebsarten identifiziert und es wurde postuliert, dass sie epithelial-mesenchymale Transitionen, mesenchymale oder Stamzell-ähnliche Eigenschaften abbilden. Obwohl für einige dieser Signaturen eine starker Einfluss der Komposition der Tumormikroumgebung aufgezeigt wurde, fehlte bislang eine umfangreiche Analyse dieser Signaturen. Ich stelle eine detailierte Analyse dieser Signaturen in unterschiedlichen Genexpressiondaten -aus Einzelzellen, Zellinien oder klinischen Tumoren- dar, die Evidenz für bisherige Fehlinterpretationen dieser Signaturen und Hinweise für die Nutzung solcher Signaturen in Geneexpressionsdaten komplexer Tumorgewebe liefert.

Im letzten Kapitel nutze ich meinen Ansatz zur Untersuchung von Genexpressionssignaturen für die Verbesserung der Diagnose einer aggressiven Form von Lungenkrebs, dem großzellig-neuroendokrinen Lungenkarzinom (LCNEC). Während diese Unterform von der Weltgesundheitsorganisation (WHO) zunächst nicht als eine separate Gruppe beschrieben wurde, stiegen dessen Fallzahlen zuletzt. Heute wird geschätzt, dass 1%-3% aller Lungenkrebse LCNEC sind. Obwohl die Diagnose zuletzt weiter optimiert wurde, erschweren technische Limitierungen und die Ähnlichkeit zu anderen Krebsarten weiterhin die Diagnose. Im dritten Kapitel beschreibe ich, wie ich mit Hilfe einer RosettaSX Analyse zunächst eine häufige neuroendokrine Differenzierung von ursprünglich als NSCLC diagnostizierten Lungentumoren feststelle. Anschließend nutze ich einen Ansatz des maschinellen Lernens, um eine präzisere molekulare Diagnose von LCNEC zu ermöglichen und das Ausmaß der Unterdiagnostik von LCNEC zu untersuchen.

Zusammengefasst zeigt diese Arbeit einen analytischen Ansatz auf, der für die umfassende Evaluierung von Genexpressionsignaturen genutzt werden kann. Die Analyse von mehreren Krebsdatensätzen und verschiedenen Anwendungen zeigen die breite Nutzbarkeit meines Ansatzes. Er gewährt einen sicheren Nutzen von Genexpressionsignaturen für eine breite Menge von onkologischen Genexpressionsdaten und ermöglicht eine differenzierte Interpretation des Neuheitswerts von Signaturen unter Nutzung des Wissens um bereits publizierte Signaturen.

Acknowledgements

I would like to thank my supservisor Benedikt Brors to give me the opportunity to pursue my doctoral studies. Thank you for your supervision. Additionally, I want to thank my Thesis Advisory Commitee (TAC) member, Prof. Dr. Claudia Scholl and Prof. Dr. Julio Saez-Rodriguez, for their valunarable input throughout the last years.

I would like to express my gratitude to the Data Science team of the healthcare business of Merck KGaA, Darmstadt, Germany. Their unwavering passion and patience have played a pivotal role in my personal and professional development. I am particularly obliged to Eike Staub, who has become a mentor to me, shaping me into the person I am today. Additionally, I thank the healthcare business of Merck KGaA (CrossRef Funder ID: 10.13039/100009945) for their grant to pursue my doctoral studies.

The analysis results in this study make use of data that was generated by TCGA Research Network: https://www.cancer.gov/tcga. Additionally I thank Sven-Eric Schelhorn and Jan Feifel for efforts to provide an readily access to the Tempus database and sharing details on its usage.

Lastly I want to thank my family. I want to thank my parents who made it possible for me to pursue this academic path. I want to thank my wife for her continuous backup and support throughout this intensive time of our lives. Last but not least I thank my two children for hanging in there despite spending many weekends without time with your dad.

Table of contents

List of Figures

List of Tables

List of Abbreviations

AIC	Akaike Information Criterion
ADC	Adenocarcinoma
BRCA	Breast Invasive Carcinoma
CCLE	Cancer Cell Line Encyclopedia
CMS	Consensus Molecular Subtypes
CNV	Copy Number Variation
CPE	Consensus Purity Estimation
CPM	Counts Per Million
CRC	Colorectal Cancer
CS	Coherence Score
CTx	Chemotherapy
DLBCL	Diffuse Large B cell Lymphoma
ECM	extracellular matrix
TME	tumor microenvironment
EMT	epithelial-to-mesenchymal transition
GBM	Glioblastoma Multiforme
HNSC	Head and Neck Squamous Cell Carcinoma
HR	Hazard Ratio
ICI	Immune Checkpoint Inhibitor
IHC	Immunohistochemical
MAD	Median Absolute Deviation
LCNEC	Large cell Neuroendocrine Carcinoma
LoT	Line of Therapy
mLNCEC	Molecular LCNEC
$\mathbf{mLNCEC}_{\mathbf{notp}}$	Molecular LCNEC, but not pLCNEC
$\mathbf{mLNCEC}_{\mathbf{p}}$	Molecular LCNEC and pLCNEC
mNSCLC	NSCLC passing RNA-Seq quality checks
NAT	Normal Adjacent to the Tumor
NE	Neuroendocrine
NSCLC	Non-Small Cell Lung Carcinoma
NPV	Negative Predictive Value

OS	Overall Survival
PAAD	Pancreatic Adenocarcinoma
PFS	Progression Free Survival
pLNCEC	pathologically classified LCNEC
PH	Proportional Hazard
PPV	Positive Predictive Value
SCLC	Small Cell Carcinoma
SNV	Single Nucleotide Variation
SqCC	Squamous Cell Carcinoma
TCGA	The Cancer Genome Atlas
TKI	Tyrosine Kinase Inhibitor
TPM	Transcripts Per Million

Chapter 1

Introduction

Understanding the molecular abnormalities that empower cancer cells to maintain abnormal cell growth is crucial in cancer research. By unraveling the riddle of cancer development and progression, researchers can improve the diagnosis of patients and gain insights into novel therapeutic options. The hallmarks of cancer, introduced in 2000, describe fundamental mechanisms for cancer development and have shown their relevance over the past decades (Hanahan 2022). They comprise the ability of cancers to progress and survive (activation of cell growth and deactivation of cell inhibition signals, angiogenesis, and metastasizing) or to block pathways that aim to arrest cell aging and apoptosis (Hanahan and Weinberg 2000). Later, the originally postulated hallmarks were extended by additional mechanisms, such as adapting energy support and immune escape, phenotypic plasticity, and microbiome interplay (Hanahan and Weinberg 2011; Hanahan 2022).

During cancer progression, the deregulation of biological pathways, which initially evolved to maintain the integrity of normal cell growth, leads to the acquisition of hallmark cancer phenotypes. On a transcriptional level, these deregulations are often traceable via molecular footprints that can be described by analyzing unique expression patterns, also called gene expression signatures. Through studies of these signatures, specific phenotypes can be characterized by analyzing cancer populations (e.g., Guinney et al. 2015) or perturbed cancer models (e.g., Bild et al. 2006). Such analyses resulted in a tremendous knowledge base that pictures the association between gene modules and biological processes called MsigDB (Subramanian et al. 2005). It provides an exceptional resource for the characterization of biological processes and can guide the characterization of cancer populations (Guo et al. 2022; McClure et al. 2023; Aran et al. 2017). Functionally, gene expression signatures can be assigned to separate phenomena that they aim to describe: the complex cell type mixtures in cancer samples (e.g., B-cells, Macrophages), the activity of biological signaling pathways (e.g., YAP), larger biological processes (e.g., proliferation, immune response) or the cellular origin of cancer (e.g., germinal center B-cell-like, BRCA-ness). When applied in the context of a cancer cohort, gene expression signature profiles can be used to differentiate patient populations or characterize the relationship between other biomarkers and previously described phenotypes described by signatures.

Gene expression signatures are sometimes limited in reproducibility, as they were identified in specific experimental settings and not evaluated in independent datasets; thus, their applicability can be limited to the original experimental context (Chibon 2013). Similarly, gene expression signatures are often confounded by other processes that deviate from the process initially intended to be described. Consequently, a study by Venet, Dumont and Detours in breast cancer indicated that many prognostic gene expression signatures are strongly associated with the proliferation status (Venet, Dumont, and Detours 2011). Interestingly, they found that random gene expression signatures are often more prognostic than published gene expression signatures for breast cancer. They showed that the random gene expression signatures contained proliferation genes, resulting in the signatures' prognostic value. This indicates that genes involved in other processes can contain gene expression signatures, leading to conclusions not warranted due to the contamination. The difficulty in providing functionally clean signatures is probably one of the main reasons why only a few signatures have been translated into clinical practice, such as the Oncotype DX test for breast cancer (Paik et al. 2004) or coloPrint for colorectal cancer (Salazar et al. 2011) stratification. Therefore, there is a need for an analysis workflow to benefit from well-defined published gene expression signatures that cover most cancer phenomena.

It is essential to evaluate their translatability to derive meaningful conclusions from a gene expression signature. Cancer characterization studies often rely on previously published gene expression signatures Subramanian et al. (2005) and derive conclusions from their signature scores. However, the association of a set of genes does not guarantee the same association between the set of genes in an independent, separate dataset. Therefore, before generalizing signatures to describe a specific phenotype, they must be evaluated for their translatability in independent datasets (Dhawan et al. 2019).

For these reasons, the first objective of my work was to identify methods and approaches to evaluate gene expression signatures across many datasets and sample types (e.g., cancer patient tumors and cancer models like cell lines). The upcoming sections will outline the main methodological aspects of the following chapters. The final section of this chapter explains this thesis's aim and structure.

1.1 Oncogenic Principles in the Light of Gene Expression Signatures

Many cancer hallmarks are driven by the deregulation of pathways or gene modules, which often can be traced by the analyses of gene expression profiles. One of the largest approaches for identifying gene expression signatures describing a wide range of biological phenotypes is the hallmark 50 collection (Liberzon et al. 2015). Liberzon et al. combined unsupervised clustering with expert knowledge to identify a subset of signatures from an enormous collection of gene expression signatures (MsigDB). They identified 50 gene expression signatures (out of 8,000) that describe larger biological mechanisms (e.g., proliferation, angiogenesis, epithelial-to-mesenchymal transition [EMT]) or signaling pathways (e.g., NOTCH, estrogen receptor). This collection has become popular and is frequently used as a reference for characterizing cancer populations (Hu et al. 2022). Besides larger collections, many other signatures were described in the literature. The following sections introduce several hallmarks of cancer and highlight individual gene expression signatures associated with phenomena that contribute to a specific hallmark.

Regulation of Growth Promotion and Resistance

One of the original hallmarks of cancer is the autonomous regulation of cell growth signaling. Multiple checkpoints that regulate a cell's fate tightly control the cell cycle, resulting in its differentiation or apoptosis (p.177-199, Wagener and Müller 2010). Tumor cells adapt this process, allowing them to proliferate and expand continuously. On a transcriptional level, the individual cell cycle phases between the checkpoints have been associated with the upregulation of specific sets of genes. Across many cancer indications, researchers identified gene expression signatures that describe genes related to cell cycle regulation. Dai et al. introduced a 50-gene expression signature, which was shown to be related to genes activated during the G_1 and G_2 phase. In their study of breast cancer patients, the signature profiles were significantly associated with poor outcomes (Dai et al. 2005). Similarly, a clustering approach has been used to identify correlated ‘meta genes’ in more than 1,000 colorectal cancer samples that differentiate a set of colorectal subtypes (Budinska et al. 2013). One of the meta genes was associated with genes involved in the cell cycle, mitosis, and other proliferation-related processes. Using microarray gene expression profiles of 76 glioma samples, Phillips et al. characterized glioma subtypes (Phillips et al. 2006). One of the three identified clusters involved cell cycle regulation-specific markers. Oncogenes and tumor suppressor genes, guardians of balanced cellular functions, are additional fundamental concepts in cancer research. For example, the proto-oncogene MYC influences multiple cancer hallmarks, including proliferation, angiogenesis, and immune surveillance (Dhanasekaran et al. 2022). Through cell perturbation experiments (overexpression of oncogenic signaling pathways), Bild et al. identified a set of gene expression signatures for signaling pathways, including MYC pathway activity (Bild et al. 2006). The signature was significantly associated with patient outcomes and verified across multiple datasets and cancer types. A well-known tumor suppressor pathway is the Hippo pathway, which regulates cell differentiation and organ size across organisms. The pathway gains importance in cancer due to its regulatory mechanisms in tumor initiation, metastasis, and drug resistance (Fu, De Angelis, and Schiff 2021). It can be categorized as the hallmark of avoiding growth suppression and cell death. Thus, it is essential to characterize the activity of Hippo in cancer indications. Wang et al. characterized somatic alterations, as well as transcriptomic differences across many cancer indications (Y. Wang et al. 2018). They indicated that core Hippo pathway genes are only mutated in a minority of cancer indications, primarily in squamous cell carcinomas. However, transcriptionally, they found a set of profoundly dysregulated Hippo downstream genes. This gene set was derived from literature and validated on three ChipSeq studies and cancer cell line protein expression data sets. The derived signature score was significantly associated with patient survival in most analyzed cancer indications.

The Tumor’s Interplay with the Microenvironment

Cell types in the tumor microenvironment (TME), stroma, and extracellular matrix (ECM) can influence processes described by other cancer hallmarks, such as inflammation, induction of angiogenesis, immune destruction, tumor invasion, or genome instability. The stroma and its inherent ECM are key tissue components that can profoundly influence tumor development and progression.

While the ECM is a complex structure of molecules that can affect the interaction between cells or regulate immune cell function (Sutherland, Dyer, and Allen 2023), the stroma is an umbrella term that, besides ECM, additionally comprises cells that can vastly influence tumor development and progression (Valkenburg, Groot, and Pienta 2018). In breast cancer, Farmer et al. identified a 50-gene gene expression signature (metagene) that was proven predictive for chemotherapy resistance (Farmer et al. 2009). Through the analysis of genes associated with the luminal-basal, apocrine, stroma, T Cell, B Cell, adipocytes, proliferation, interferon, and hypoxia, they found that only the stroma metagene was significantly associated with survival across multiple datasets. Subsequently, by analyzing histologic sections, they showed significant enrichment of reactive stroma in samples with high stroma scores. Similarly, in glioblastoma, Liang et al. identified a gene expression signature with genes significantly predictive for poor prognosis (Liang et al. 2005). Analyzing samples from different brain diseases and modeling approaches, they identified a set of genes significantly associated with survival across datasets. Genes involved in this signature were primarily related to neural cell migration capabilities or ECM proteins.

The tumor microenvironment comprises the entirety of the ECM, stroma, immune cells, and endothelial cells (Anderson and Simon 2020). In colorectal cancer (CRC), Bindea et al. performed an in-depth analysis of expression datasets to identify cell type-specific gene expression signatures (Bindea et al. 2013). Interestingly, they showed the high variability of cell type compositions dependent on the tumor stage and a strong association of individual cell types on patient outcome. Their work resulted in a set of gene expression signatures that allows the characterization of cell type compositions in complex tumor samples.

Cellular Plasticity and Metastasis

Cellular plasticity significantly impacts cancer cells' invasiveness and metastatic capability. A process that is involved in at least three cancer hallmarks (invasion and metastasis, cell death resistance, and circumvent immune destruction) is an epithelial-mesenchymal transition (Dongre and Weinberg 2019)). Through this process, cells can lose their epithelial characteristics and gain mesenchymal features, allowing them to switch from stationary to mobile status. This process is well-known in normal tissue for wound healing but has also been extensively studied in the context of cancer research. Several research groups have described cancer subtypes that are characteristic for EMT, mesenchymal or stemness in glioblastoma multiforme (GBM) (Phillips et al. 2006; Verhaak et al. 2010), breast (Taube et al. 2010; Lien et al. 2007; Lehmann et al. 2011), neck cancer squamous cell carcinoma (HNSC) (Walter et al. 2013), colorectal (Guinney et al. 2015) and other indications (Liberzon et al. 2015). In colorectal cancer, multiple studies analyzed gene expression data using unsupervised clustering approaches and indicated a set of subtypes that describe patient populations with distinctive characteristics (Sadanandam et al. 2013; Budinska et al. 2013; Marisa et al. 2013; Roepman et al. 2014; Schlicker et al. 2012; Melo et al. 2013). Each of these studies described partly overlapping subtypes: Most of the approaches identified a subtype that indicated characteristics of immune infiltration, and another subtype was associated with stemness, mesenchymality, or EMT. A subsequent large-scale analysis by Guinney et al. combined the previously described subtyping approaches in four consensus molecular subtypes (CMS). The four subtypes described samples with 1) immune infiltration, 2) canonical, 3) Metabolic deregulation, and 4) mesenchymal and stem cell characteristics. Subsequent studies indicated a strong dependency on the tumor microenvironment of individual subtypes and proposed a separate algorithm for pre-clinical models (Eide et al. 2017).

While the importance of EMT and cancer cell plasticity is generally widely accepted, detecting EMT or a mesenchymal state by published gene expression-based subtyping approaches has been criticized recently, especially for the CMS subtype scheme for CRC (Dunne et al. 2016). Resolving this scientific debate is paramount since subtyping schemes must deliver robust calls for individual tumors and robust prevalence estimates on the population level when used for drug development and treatment decisions.

The Cell of Origin Concept

Besides the hallmarks of cancer, the concept of cell of origin concept has evolved, which describes the specific cell type from which cancer originated (Hoadley et al. 2018). Although the cell-of-origin and hallmarks of cancer are distinct concepts, they are strongly interconnected. In breast cancer, four subtypes were associated with specific characteristics used for patient characterization and stratification: luminal A/B, normal-like, and basal-like (Perou et al. 2000). While the luminal subtypes originate from luminal epithelial cells, the basal-like subtype originates from basal epithelial cells. As such, these subtypes indicate critical signaling cascades that, among others, provide insights into cancer hallmarks, such as proliferation. While the luminal subtype is primarily driven by hormone-driven signaling via the estrogen receptor and progesterone receptors, the basal-like subtype is driven by other growth factors. Multiple gene expression signatures evolved to differentiate these subtypes (Farmer et al. 2009; Calza et al. 2006). Similarly, several studies in diffuse large B-cell lymphoma (DLBCL) established gene expression signatures for the characterization of two main types of glioblastomas: activated B cells (ABC)-like and germinal-center B-cell (GCB)-like, for which there are also established gene expression signatures (Masqué-Soler et al. 2013). As in breast cancer, the ABC-DLBCL subtype showed characteristic pathway aberrations that are associated with proliferation (Compagno et al. 2009). Neuroendocrine cells represent another cell-of-origin group, which, for example, comprises small-cell lung carcinoma (SCLC) and large-cell neuroendocrine carcinoma (LCNEC). SCLC accounts for approximately 20% of all primary lung tumors, and LCNEC for 1% to 3%. Patients with these tumors usually have inferior prognosis, and for LCNEC, there is no standard therapeutic recommendation. Pathological classification is strongly determined by morphological characteristics and immunohistochemical neuroendocrine marker (*NCAM1*, *CHGA*, *SYP*, or *INSM1*) expression. Molecular analysis indicated specific expression patterns that describe the neuroendocrine origin of these carcinomas in the lung and other tissues. In SCLC, by analyzing neuroendocrine (NE) and non-neuroendocrine samples, Zhang et al. could identify a set of highly NE tissue-specific markers. Using these markers, they identified cancer cell lines with high and low NE marker expression to subsequently identify NE and non-NE genes in a differential gene expression analysis (W. Zhang et al. 2018). This and other signatures on neuroendocrine gene expression suggest that gene expression could be the precise diagnosis of LCNEC within NSCLC. Better LCNEC diagnosis could help patients be treated according to the cell of origin of their aggressive cancer.

1.2 Methods for Gene Expression Signatures

Since the introduction of the first gene expression signatures, many resources have evolved. The number of signatures steadily grows, resulting in large databases with thousands of gene expression signatures (PubMed search https://pubmed.ncbi.nlm.nih.gov/, Figure 1.1). Additionally, multiple measures have evolved to characterize a gene expression signature's profile or activity.

This section introduces resources for gene expression signatures and provides an overview of methods for scoring their activity and translatability.

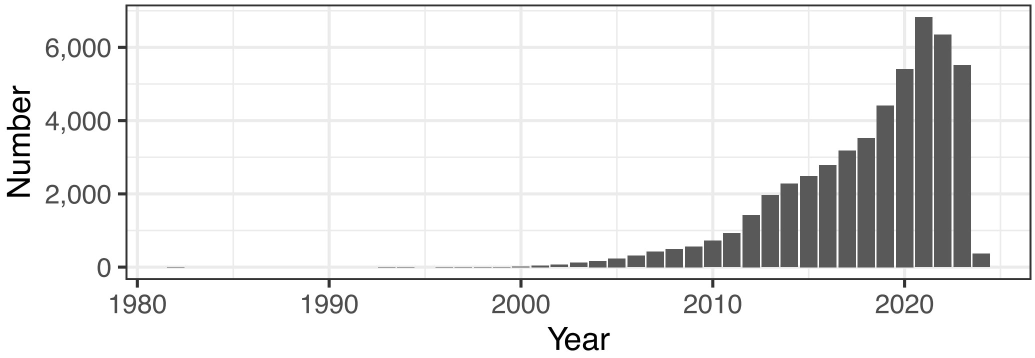

Figure 1.1: Number of publications for the search term 'Gene expression signature' and 'cancer' or 'oncology' by year on PubMed (accessed 2024-01-19).

Signature Collections

Over the past, thousands of gene expression signatures have evolved. The molecular signature database (MsigDB) is one of the largest collections of gene signatures (Subramanian et al. 2005; Liberzon et al. 2015). It comprises signatures for mouse and human organisms. It is split into broad collections (H, C1-8), describing hallmark signatures especially relevant for cancer research and computationally derived or ontology-derived signatures. Each collection comprises thousands of signatures. Although the database is an excellent resource for signatures, its size and generality introduce several limitations. Even though these resources are highly valuable and provide a resource, they must be used with caution. Often, the signatures in these databases are not evaluated on multiple independent datasets and thus might not describe the phenotype to which they were assigned. Venet et al. showed that prognostic gene expression signatures are often not more prognostic than randomly sampled gene sets (Venet, Dumont, and Detours 2011). The reason for that was that many of the signatures were contaminated by proliferation markers that indirectly led to the prognostic association of the signatures.

Similarly, other factors related to different mechanisms might confound gene expression signatures. To comprehensively characterize a new gene expression signature, it is essential to compare it to a suitable set of signatures covering a broad range of cancer mechanisms. Additionally, the new signature should be evaluated in the context of other signatures.

Evaluating the Quality and Activity of Gene Expression Signatures

The evaluation of gene expression signatures requires two steps: a) verifying the integrity of the signature in a new dataset and b) evaluating the activity of the signature. As pointed out, gene expression signatures can originate from data-driven approaches, experiments, or other procedures. Additionally, they might be derived from data generated with different technologies; sometimes, they were validated on multiple datasets and sometimes not in a single dataset. The hypothesis of using a gene expression signature on a dataset is that the expression of the same set of genes in another dataset can robustly explain the same phenomenon that was initially proposed.

To assess the relevance of a gene expression signature in a dataset that was not used for the definition of a signature, it is essential to test the robustness of the signature. One approach is the coherence score (CS), which has been independently developed in the context of pathway analyses by others (Zien et al. 2000; Rahnenführer et al. 2004; Fan et al. 2016; Staub 2012). The score evaluates the correlation of the expression of gene pairs in a gene expression signature. To calculate the coherence score for a signature of size k with X_1, X_2, ..., X_k gene expression variables, for each pair (X, Y) the correlation coefficient (e.g., Pearson correlation coefficient Kirch 2008) is calculated:

$$\mathrm{r}_{XY} = \frac{\sum_i^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_i^n (X_i - \bar{X})^2 \sum_i^n (Y_i - \bar{Y})^2}} \tag{1.1}$$

The average correlation coefficient of the upper triangular matrix of this pairwise correlations matrix with i rows and j columns defines the CS (adapted from Rahnenführer et al. 2004):

$$CS := \frac{1}{\binom{k}{2}} \sum_{1 \leq i \leq j \leq k} r_{ij} \tag{1.2}$$

This formula calculates the average of the individual pairwise correlations (r_{ij}) to determine a score indicating the strength of the association between the genes in a signature. A high absolute value indicates a coordinated expression of the genes in a signature, suggesting a persistent functional relationship. Although similar approaches were implemented earlier (Zien et al. 2000), Staub et al. introduced the term CS (Staub 2012).

For the description of the activity of a gene expression signature, multiple approaches have been described (Foroutan et al. 2018; Yi et al. 2020; Tomfohr, Lu, and Kepler 2005; Hänzelmann, Castelo, and Guinney 2013; E. Lee et al. 2008; Ebi et al. 2009). Early approaches use simple transformations to aggregate gene expression profiles on a gene expression signature level. For example, the 'deregulation index' used by Ebi et al. describes the average of the normalized gene expression values of a gene expression signature (Ebi et al. 2009). A slightly modified version was proposed by Lee et al., who first z-transformed the expression data for each gene and then averaged the signature score by summing the z values and dividing them by the square root of the number of genes within the signature (E. Lee et al. 2008). In an algorithm called PLAGE, Tomfohr et al. proposed to measure the activity of a gene expression signature or pathway by using singular value decomposition (SVD). They termed the first eigenvector as metagene (Tomfohr, Lu, and Kepler 2005). The gene set variation analysis (GSVA) method uses enrichment scores to define gene expression signature scores (Hänzelmann, Castelo, and Guinney 2013). The authors use the dataset's complete gene expression signature matrix and rank the genes for each sample. For each gene set, the distribution of ranks allows them to differentiate whether the genes in a specific signature are coordinately higher or lower expressed than other genes in the respective sample.

While the previous methods are tailored toward large-scale studies, individual measures have evolved to define a single sample's gene expression signature activity. Among them, the single sample gene set variation enrichment analysis (ssGSVEA) that was introduced by Barbie et al. (Barbie et al. 2009) and the singscore, which was introduced for single sample analyses and differentiates bidirectional and unidirectional gene expression signatures (Foroutan et al. 2018). The authors rank genes for the bidirectional signatures based on the expression levels in a single sample (descending for upregulated gene sets and ascending otherwise). The average rank is normalized by the theoretical maximum or minimum value. For signatures with unknown direction, the authors use the absolute median-centered rank of the sample's gene expression abundance ranks.

Platforms

Over the past, there have been a few platforms or software solutions for evaluating gene expression signatures. One of the most evolved software solutions is the SigQC R package (Dhawan et al. 2019), which systematically evaluates individual gene expression signatures based on multiple measures. It also allows computing and comparing different scores for a signature or assessing the contribution of individual signature genes. The overarching goal of sigQC is evaluating a single signature or a small set of signatures and optimizing the signature genes.

1.3 Resources for Molecular Cancer Characterization

To characterize and decode the complex biological phenomena in human tumor tissues, large databases of molecular cancer profiles are required to derive knowledge. Over the past years, many projects have evolved that aim to characterize multiple modalities of single patients.

The Cancer Genome Atlas (TCGA) project is a large consortium that provides access to data from more than 11,000 cancer patients (Weinstein et al. 2013). Over the past decade, the resource offered a tremendous opportunity to characterize cancer indications, develop a bioinformatics approach, or use it as a validation dataset. It, for example, provides access to gene copy number variation, small nucleotide variation, mRNA expression, protein abundance, methylation, and protein expression data. The dataset comprises 32 cancer indications with up to 1,000 samples each.

Tempus is a commercial database for real-world data that is comprised of clinical, response, and molecular data. The healthcare business Merck KGaA, Darmstadt, in-licensed data from multiple cancer indications. In this thesis, I had access to more than 6,000 NSCLC patients from the Tempus database across various modalities (single nucleotide variation [SNV], copy number variation [CNV], gene expression, gene fusions) and clinical data from electronic health records. Tempus developed a DNA-Seq assay, which is used as the basis for copy number and single nucleotide calling, called xT targeted DNA-Seq assays covering 595 (v. 1-2) or 648 (v. 3-4) genes (Beaubier et al. 2019). Similarly, mRNA quantification is done by the RS-targeted RNA-Seq panel assay and spans 39 Mb target region (19,396 genes).

The clinical data provides access to tumor characteristics, medications, outcomes, clinical assessments, and adverse events. In the last chapter, outcome data were not directly available from the Tempus database due to the raw form of real-world evidence data. I will describe additional details of the implementation later used to infer progression-free survival and overall survival in an indication of interest.

The cancer cell line encyclopedia is a project that aims to provide a large-scale dataset of cancer cell line models (Ghandi et al. 2019). Over the past decade, the number of available data modalities for each model has been extended to gene expression, copy number, single nucleotide variation, CRISPR-Cas, shRNA, protein expression (RPPA, Mass Spec), promotor methylation, and more resources.

1.4 Machine Learning Concepts

Throughout this thesis, I will use several types of machine-learning approaches. The overarching goal of machine learning models (i.e., a function to predict the outcome from a set of features) is to learn properties from a set of input variables (or features) to predict an outcome of interest for each sample of input variables. The outcome might be a categorical (class prediction) or a continuous variable (regression). Furthermore, machine learning approaches can be grouped into supervised, unsupervised, and semi-supervised approaches. While in the former approach, positive and negative samples are available for training, and unlabeled samples are classified by the model, in the unsupervised approach, no labeled samples are available. In addition to these two main types of approaches, hybrid or special types of approaches exist. One approach deals with a data situation in which only a few positive training samples are available, but most are unlabeled. Additionally, the unlabeled data is expected to comprise negative samples predomintly, and the model aims to predict the expected additional positive labels among the unlabeled samples. This case is called the Positive-Unlabeled (PU) problem, which requires specific designs of machine learning approaches. I will apply one of these approaches in the last part of this work to predict the cell of origin of a small set of NSCLCs.

The goal of a model is usually to predict the labels of a set of samples given a set of features. To do so, evaluating a model's generalizability (or performance) is pivotal. The following steps are usually performed to train and evaluate a model.

1. Split the data into a training and testing set

2. Preprocess the training data

3. Train the model on the training data

4. Evaluate the model on the testing data

The first step allows the model to be trained on a set of data and evaluate its performance on, for the model, unseen data. In the second step, the data is preprocessed, which is a crucial step and might involve feature engineering, such as a numerical transformation of a feature or the conversion of a continuous to a categorical value. In the third step, the model is trained. It is essential to use strategies to reduce the possibility of the model only being able to predict the labels of the underlying data used to train (over-fitting) or not (under-fitting). A common approach is to further split the data multiple times into subsets called validation sets, in which the model is trained independently, which is a strategy to obtain a more generalizable model. Lastly, the model is evaluated on the test data split from the training data before model training. This last step allows evaluation of the performance of unseen data and provides essential insights into the model's generalizability.

In the following sections, I will describe each step in more detail.

Feature Selection and Hyperparameter Optimization

As indicated above, it is crucial to implement a procedure that allows the model to generalize beyond the data it has seen for training.

Machine learning models require a set of parameters that influence the learning procedure and how well a model applies to a classification or regression task. In the case of a random forest, key hyperparameters are mtry and ntree, representing the number of features used within a random forest's splits and the number of trees used. These parameters depend highly on the dataset under study, and many approaches have evolved to avoid over-fitting. In the last chapter, I use nested cross-validation, which splits the training data into multiple subsets, named outer sets, for selecting features. As a second step, each outer set is divided into additional subsets, which are used to identify the optimal hyperparameters for the selected features based on the outer set. Different methods can be used to split the inner and outer sets. In the third chapter, I combine bootstrapping and k-fold cross-validation (Figure 1.2).

For bootstrapping, a new set of data is generated from the training data of the same size. However, samples are sampled with replacement, resulting in redundancies of individual samples. For k-fold cross-validation, a dataset is split into k parts of the same size. A model is trained on k-1 subsets and evaluated on the left-out subset. This process is repeated for each combination of subsets. This procedure allows the model to assess model parameters or features on many different subsets of the original data and evaluate small subsets of unseen data (validation).

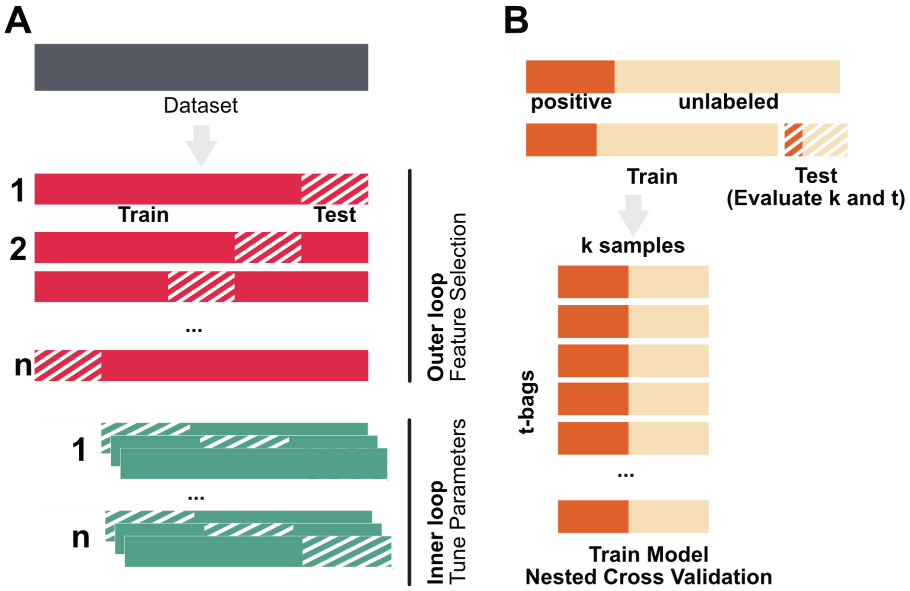

Figure 1.2: Schematic representation of k-fold cross-validation and nested cross-validation. A: There are two nested loops of k-fold cross-validation in the nested cross-validation. The training data is split into k training and validation sets in the outer loop. It is usually used to select features. Each training split is again divided into k parts for hyperparameter optimization. B: In k-fold cross-validation, there is only one loop of splitting, which is used for hyperparameter selection (e.g., when the features are already known)

Learning from Positive and Unlabeled Samples

Positive-unlabeled (PU) learning represents a subclass of machine learning problems in which the data has a small subset of positive samples and a large set of unknown samples. That means that it is known that there is a subset of samples that are positive but were not labeled as such. PU learning aims to identify positive cases in unlabeled samples.

One approach for identifying positive samples in the unlabeled data combines a machine-learning model with a binning approach (Mordelet and Vert 2014) (Figure 1.3). This approach splits the data into k bins of size m. Each bin comprises all known positive cases and a subset of unlabeled samples. A separate model is trained for each bin, including hyperparameter optimization and feature selection. As a last step, the class for all samples not used for training (excluding positive samples) is predicted, and the average prediction for a sample is used as a final prediction.

The controlled binning of samples results in different combinations of positive and unlabeled samples. In some subsets, the unlabeled samples might be true negative; in others, there might be positive cases without a label and a mixture of both. Thus, the trained models will sometimes differentiate true positives from negatives or a mixture of true and false negatives.

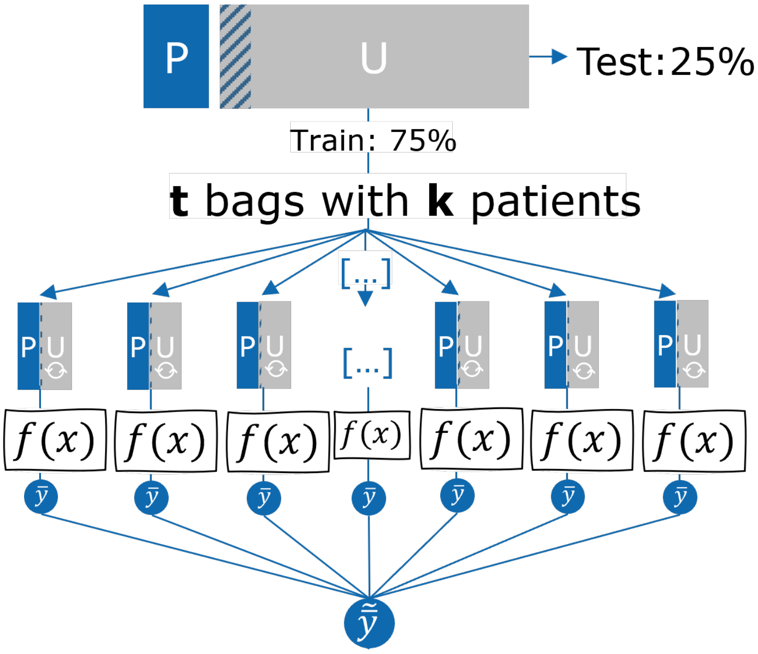

Figure 1.3: Overview of a bagging approach for positive unlabeled learning. The positive and unlabeled data are split into two sets (e.g., 75% and 25%). Subsequently, the training data is divided into t bags with k unlabeled samples (with replacement). Each bag also includes all positive samples and, separate models are trained on each bag. The final prediction combines the classification of the individual models.

Performance Evaluation

Multiple measures have evolved to evaluate how well a model fulfills its prediction task, one frequently applied approach is the F_1 Measure (Christen, Hand, and Kirielle 2023). A contingency table is vital to many measures and counts the number of true and false positives or negatives (e.g., correctly or wrongly classified samples).

	Truth	
Predicted	**true**	**false**
true	true positive (TP)	false positive (FP)
false	false negative (FN)	true negative (TN)

The contingency table can then define the precision (r, i.e., true positive rate) and precision (p, i.e., true negative rate).

$$r = \frac{TP}{TP + FN}$$

$$p = \frac{TN}{TN + FP}$$

In binary classifications, the aim is often to increase TPR and TNR. The F_1-score is the harmonic mean of both.

$$F_1 = 2\frac{p * r}{p + r} \tag{1.3}$$

in the case of PU-learning, there is however a lack of true negatives and thus it is not possible to derive p. If the unlabeled samples are used as true negatives, it often results in a biased performance estimation (Jain, White, and Radivojac 2017). Consequently, Lee et al. described a method that relies on precision only, and was shown to work well (W. S. Lee and Liu 2003). They introduced the following measure as a proxy for the F_1 measure for PU-learning issues:

$$F_1^{PU} = \frac{r^2}{Pr[f(x) = 1]}$$

$Pr[f(x) = 1]$ describing the probability of a sample being classified positive, and r representing recall.

1.5 Aims and Structure of this Work

This study aims to identify signatures and algorithms to establish a well-defined set of high-quality signatures that can be applied to a novel expression dataset. The goal is to provide critical insights into transcriptional phenotypes. Furthermore, I aim to demonstrate how this knowledge can be used to comprehensively characterize cancer populations and molecular readouts by in-depth analysis of signatures. Finally, I strive to illustrate how cancer subtypes and associated gene expression signatures can be identified by machine learning, demonstrate their high quality by applying my gene expression signature analysis framework and provide evidence that it can improve the diagnosis of cancer patients.

This thesis is organized into four chapters. The first chapter, the introductory chapter, provides important aspects for the subsequent chapters. This section provides a background on gene expression signatures and their association with the hallmarks of cancer. It also offers an overview of essential resources and methodologies used in subsequent chapters.

The second chapter will describe a methodological approach to analyzing gene expression signatures. This approach utilizes gene expression signatures to represent their associated phenotype within other contexts. This allows me to characterize gene expression data of cancers with the following goals: analyze gene expression signatures compared to other signatures or characterize sets of cancer samples (e.g., whole patient cohorts or cancer indications) comprehensively with sets of signatures. Using this notion, I will describe a web platform I released to the public that allows users to analyze gene expression signatures across over 10,000 patient cancers and more than 1,000 cancer cell line models.

After describing the methodological approach, the second and third chapters focus on its application. In Chapter 3, the capabilities of the approach described in Chapter 2 will be highlighted. This approach enables the analysis of various cancer datasets and gene expression signatures to derive insights into the associations of signatures suggested to describe cancer EMT, mesenchymal, or stemness characteristics. Finally, in Chapter 4, I apply my workflow to identify large cell neuroendocrine carcinomas (LCNEC), a rare lung cancer subtype, in a real-world evidence dataset. While Chapter 3 demonstrated the applicability of my approach for characterizing gene expression signatures, this final chapter depicts how it can be used to describe biological mechanisms in cancer patients.

Chapter 2

RosettaSX - An Approach for Gene Expression Signature Evaluation and Scoring

The analysis of gene expression signatures is often vital to cancer characterization studies. In this chapter, I introduce a workflow that evaluates the applicability of a collection of gene expression signatures in the study that differs from their discovery study. Using this approach, I implement a public web service that provides access to more than 11,000 pan-cancer patient tumor and cancer cell line data across more than 33 cancer indications from TCGA and DepMap. However, the workflow certainly applies to any dataset, which will be shown in the later chapters. The work described in this chapter was published in an article published in Neoplasia (Kreis et al. 2021).

2.1 Project Outline

As outlined in the first chapter, gene expression signatures have a long history and have contributed significantly to the characterization of biological phenomena in cancer research (Section 1.1). Their analysis allows the identification of subpopulations in a cancer indication, the evaluation of active processes, and the presence of cell types in a complex sample. Consequently, in cancer research, studies often utilize gene expression signatures to describe the characteristics of a population or to characterize a biomarker of interest (Qian et al. 2021). Such studies often use previously defined collections of gene expression signatures (e.g., hallmark 50) that were postulated to be of pivotal relevance in other datasets and use gene expression signature scoring approaches that seek to represent the activity of a signature (Section 1.2). However, the composition of a population of samples might strongly vary between different studies. Thus, it is essential to reevaluate the relevance of a gene expression signature in the context of a new dataset. Otherwise, the gene expression signature might not represent the phenomenon under study but the noise in the analyzed dataset (Staub 2012). Additionally, collections, such as the hallmark 50 collection, lack established signatures for individual cancer subtypes, such as cell of origin signatures (Section 1.1), that are of pivotal importance for the characterization of such cancers. When not using such signatures to analyze a cancer indication, important characteristics might be overlooked.

Therefore, in this project, I implement a platform that combines the evaluation of gene expression signatures to identify a robust set of signatures and subsequently analyze these only in the context of a new dataset. The combined analysis can help to identify relevant phenomena that drive the phenotypes of the sample population. Once a set of relevant signatures is selected, they can be used to characterize patterns of biomarkers, such as gene expression markers. The embedded information can inform us about phenomena previously associated with a set of genes. Similarly, known subgroups of samples can be further characterized by the signatures.

2.2 Methods

Here, I implemented a systematic approach that uses this property to verify the translatability of a signature, selects relevant GES in a dataset, and finally provides a web interface to analyze relevant signatures. This method section as adapted from (Kreis et al. 2021)

Preprocessing Gene Expression Data

For the normalization of the gene expression data, I applied the trimmed mean of M-values (TMM) method of the edgeR package (version 3.34.0, Robinson, McCarthy, and Smyth 2009) to factor out differences in RNA composition between the samples of a cohort (i.e., TCGA cohort, and Experimental Factor Ontology [EFO] anatomical entities for cancer cell lines). For the normalization, I removed genes with less than ten reads in 70% of the samples in a cohort or with less than 15 reads overall.

Scoring Gene Expression Signatures

The platform selects relevant gene expression signatures using the CS (Staub 2012; Kreis et al. 2021). The CS evaluates whether the expression profiles across a set of genes are synchronized. The score is defined by the average pairwise correlation of all combinations of gene pairs, and a value approaching one indicates a coherent signature (see Equation 1.2). I calculate the pairwise Pearson correlation for all gene expression values (TMM normalized counts per million [CPM]) and average the upper off-diagonal correlation coefficients in a pairwise correlation matrix for each gene expression signature. A positive score indicates that the genes are synchronously upregulated in at least a subset of the analyzed sample population. Such a pattern might suggest that the gene set as a group is coordinately regulated and possibly involved in the same mechanisms. The significance of the CS is primarily dependent on the signature size. Empirical p-values indicated that CSs larger than 0.2 are highly significant (p-value $<$ .0001) for a gene set of size 10 (see Section 1.2, Figure 2.2).

For the platform, I pre-calculate gene expression signature scores. For each signature, I first log2 transformed and added one to the expression values (transcripts per million [TPM] to account for gene length) and then z-scale the values for each gene. Subsequently, I calculated the mean expression for each sample. Others have referred to this score as a z-score or deregulation index (Hänzelmann, Castelo, and Guinney 2013; Ebi et al. 2009).

Software

Shiny is an R package for implementing interactive user interfaces (version 1.5.0, Chang et al. 2022). The package and its extensions (version 1.1, Attali 2021) enable building scalable and highly customized user interfaces. I combined shiny with a lightweight data structure (fst 0.9.2 Klik 2022) to quickly access molecular, clinical, and phenotype data. The shiny app's central part is a heatmap showing gene expression signatures. The user interface consists of interactive inputs that allow the user to select heatmap annotations, such as molecular readouts (e.g., gene expression, gene copy number alterations) or clinical/phenotype information (e.g., gender).

The app runs on a public server (www.rosettasx.com) with 4 CPUs and 16 GB of RAM with R 3.6.3. For further session information and R-package versions, see Section B.1.

Data availability

For this work, I used molecular and clinical data that was downloaded from the Xena (Doldman et al. 2020) and genomic data commons (GDC) (Grossman et al. 2016) database (accessed: 2020-03-15 and 2020-04-14), harmonized and shared by Sven-Eric Schelhorn. For the cancer cell lines, I accessed molecular and gene dependency data (SNP6 SNV: 19Q1, methylation: 19Q1, dependency: 20q2, RPPA: 19Q1, WES CNV: 20Q2) from the DepMap data portal (Meyers et al. 2017; Barretina, Caponigro, and Stransky 2012; Dempster et al. 2019) and used cell line annotations (provided by Sven-Eric Schelhorn) from Cellosaurus (Bairoch 2018) to map cell lines to EFO. Gene expression signatures listed in Table C.1 have been curated by Johanna Mazur, Miriam Urban, Sven-Eric Schelhorn, Thomas Grombacher, Eike Staub, and me.

Statistical Tests

All analyses were performed in R (version 3.6.3, R Core Team 2021). Statistical tests were implemented wiht base R functions. For the comparison of gene expression levels and the IHC level of a biomarker I used Wilcoxon rank sum test (Hollander, A. Wolfe, and Chicken 2015). For the evaluation of a significant difference in the frequencies of a categorical variable (e.g., copy number change) between two groups (e.g., intrinsic subtypes) I used the Pearson chi-square test (Agresti 2006).

2.3 Results

The above-described obstacles guided me to implement an analytical workflow and web service that allows users to analyze public cancer datasets. In the following, I will first evaluate a collection of gene expression signatures gathered by my research group. Subsequently, I will introduce a methodological approach on which I build a web service named RosettaSX (Kreis et al. 2021). Using this web service, I will analyze molecular data of TCGA Breast Invasive Carcinoma (BRCA) patients and characterize the PAM 50 subtypes.

Evaluating the Significance of the Coherence Score

To demonstrate the effect of the CS, I will analyze a gene expression signature proposed to describe the presence of T-cells (Bindea et al. 2013) in cancer patients and cancer cell line models (Figure 2.1). To empirically evaluate the significance of the CSs for this signature, I generated 10,000 random gene expression signatures and determinated the number of signatures that reach an equal or larger CS than the T-cell signature (i.e., an empirical p-value). While the T-cell gene expression signature had a CS of 0.64 (empirical p-value $< .001$) in the cancer patient data, the CS was lower in the TME naive samples (0.11, empirical p-value $= .895$). This highlights that this signature reaches a significantly higher CS than random signatures, but only in the context where the signature is expected to be relevant (i.e., biopsies comprising T-cells).

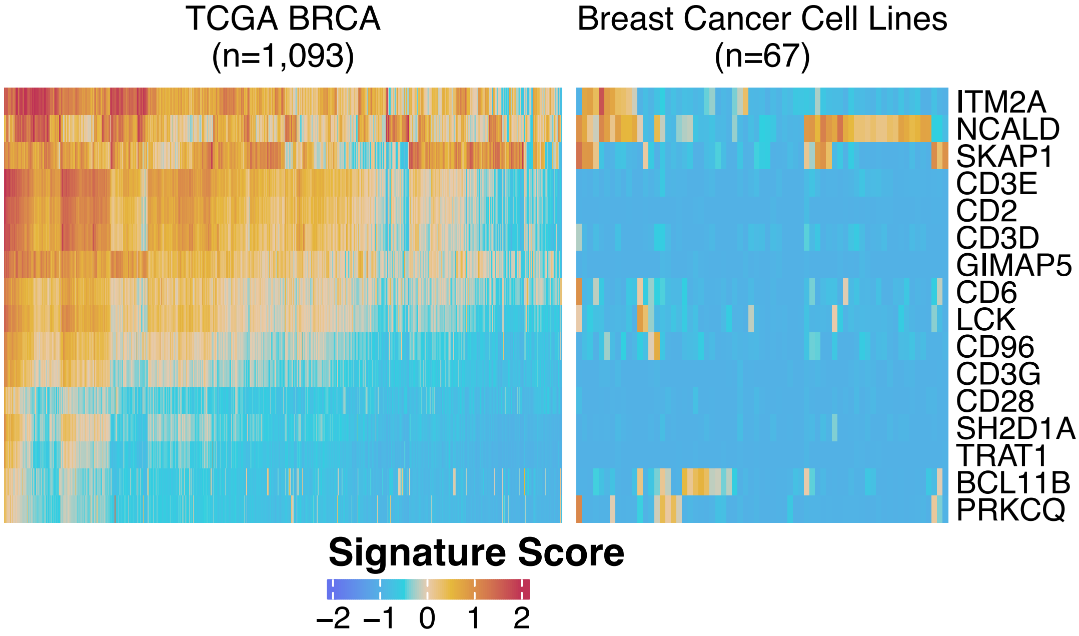

Figure 2.1: Gene expression levels of genes described as a gene set for T-cells in bulk sequencing data. While the left heatmap shows the expression of the signature genes in cancer patient data (TCGA BRCA), the right heatmap shows the same set of genes in breast cancer cell lines. Red indicated a high activity of the signature and blue a low activity. In patients with BRCA, almost all genes have a coordinated expression. While patients on the left side have high scores in most genes, patients on the right have almost exclusively low scores. The cancer cell lines, however have an irregular expression of all genes, indicating no functional association of the genes in this context.

The above example showed the importance of the CS on the interpretability of a gene expression signature.

A more thorough evaluation of the empirical distribution of the CS allows me to derive significant CS levels for gene expression signatures. Thus, for a set of gene expression signature sizes (i.e., 3-100), I next created 10,000 random gene expression signatures to derive empirical p-values (i.e., the number of random signatures that reach an equal or larger value than a signature under study), that later can be used as a measure for the evaluation of the significance of a CS for a gene expression signatures of interest (Figure 2.2). This analysis indicated that the CS strongly depends on the size of the signature and that a cutoff of 0.2 filters signatures of size three, which are significant at an empirical *p*-value of .05 (i.e., only 50 out of 10,000 random signatures of size three had a higher CS than 0.2). Similarly, signatures of size 10 have an empirical *p*-value < .0001 (i.e., at most, one out of 10,000 random signatures of size 10 had a higher CS than 0.2).

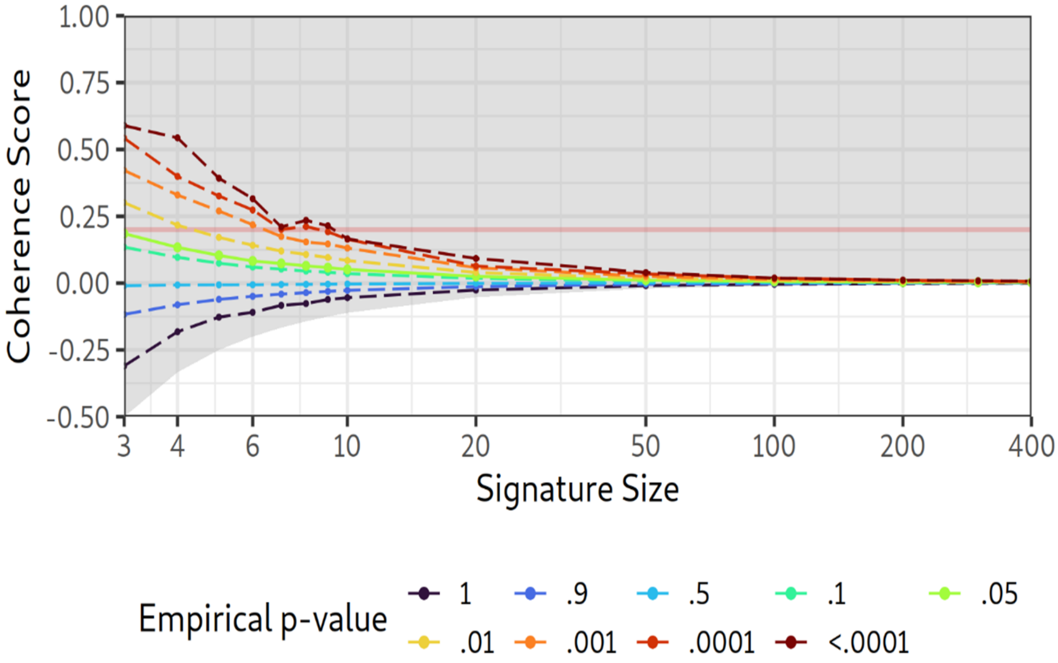

Figure 2.2: Distribution of empirical *p*-values for the CSs of signatures a size of 3 to 400. For each entry on the x-axis (i.e., the size of the signature), I sampled 10,000 random gene expression signatures in TCGA BRCA. The colored lines indicate the empirical *p*-values for the individual size of the signature. (Reprinted from Kreis et al. 2021)

I next evaluated the significance of the gene expression signature collection in BRCA and compared it with a random gene expression signature collection of comparable size (Figure 2.3). While the collection of random gene expression signatures reached a maximal CS of 0.1, 156 signatures in the collection of RosettaSX reached a score larger than 0.1. This indicates that the signatures have a strong expression footprint that is not observable for random signatures. However, 150 signatures did not reach a CS larger than 0.1, indicating that these signatures do not describe a robust (i.e., coherently expressed) gene expression module in this context. Still, these signatures might be relevant in another context.

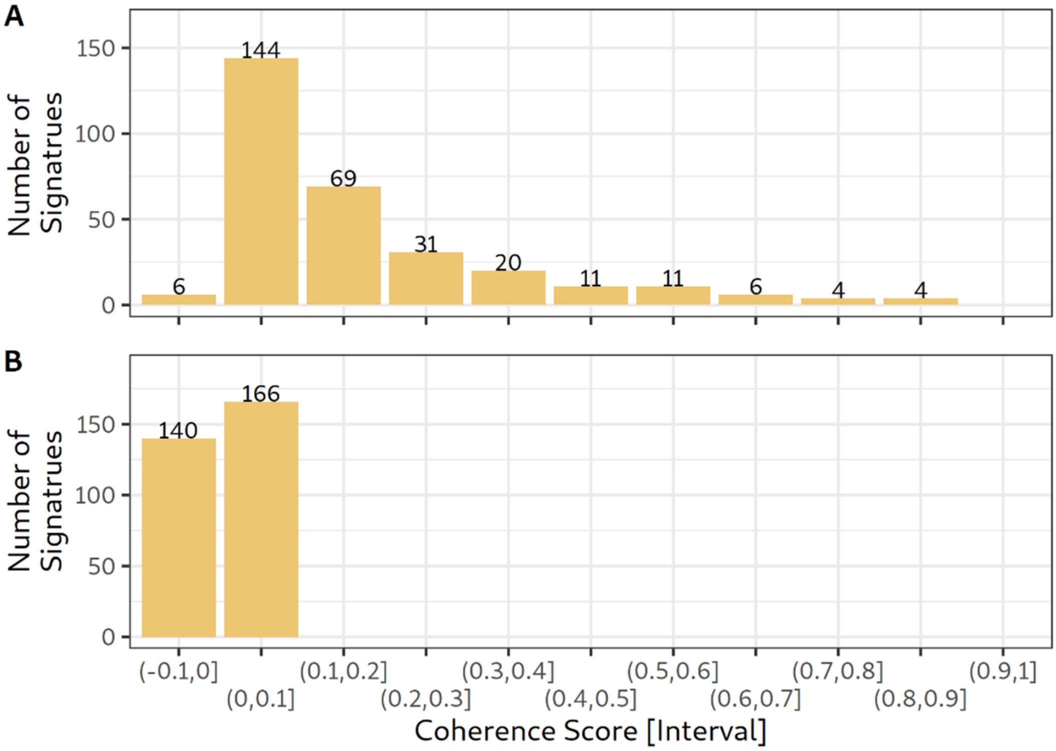

Figure 2.3: Distribution of CSs for gene expression signatures in the RosettaSX gene expression signature collection and in a random collection of equal size. Bar plots for the distribution of CSs grouped into intervals of size 0.1. The top plot shows the CSs for the signatures in the RosettaSX collection, and the bottom plot for a randomly generated collection of equal size. (Reprinted from Kreis et al. 2021)

Comparison of Gene Expression Signature Scoring Methods

In the introduction section, I introduced multiple scores proposed for evaluating the activity of a gene expression signature (Section 1.2). In my work, I use the z-score, which, for each sample, represents the average z-scaled expression value of a set of genes. To evaluate the differences between the different scoring methods, I compared the alignment of the scores to the CSs of the individual gene expression signatures (Figure 2.4). Interestingly, the gene expression signature scores deviate a lot when the CS of the signature was low (e.g., below 0.2). However, the agreement between the different scoring systems was high if the CS increased. These results indicate that the choice of scoring approach has a minor influence on the analysis of the signatures that are filtered for their robustness using the CS filter.

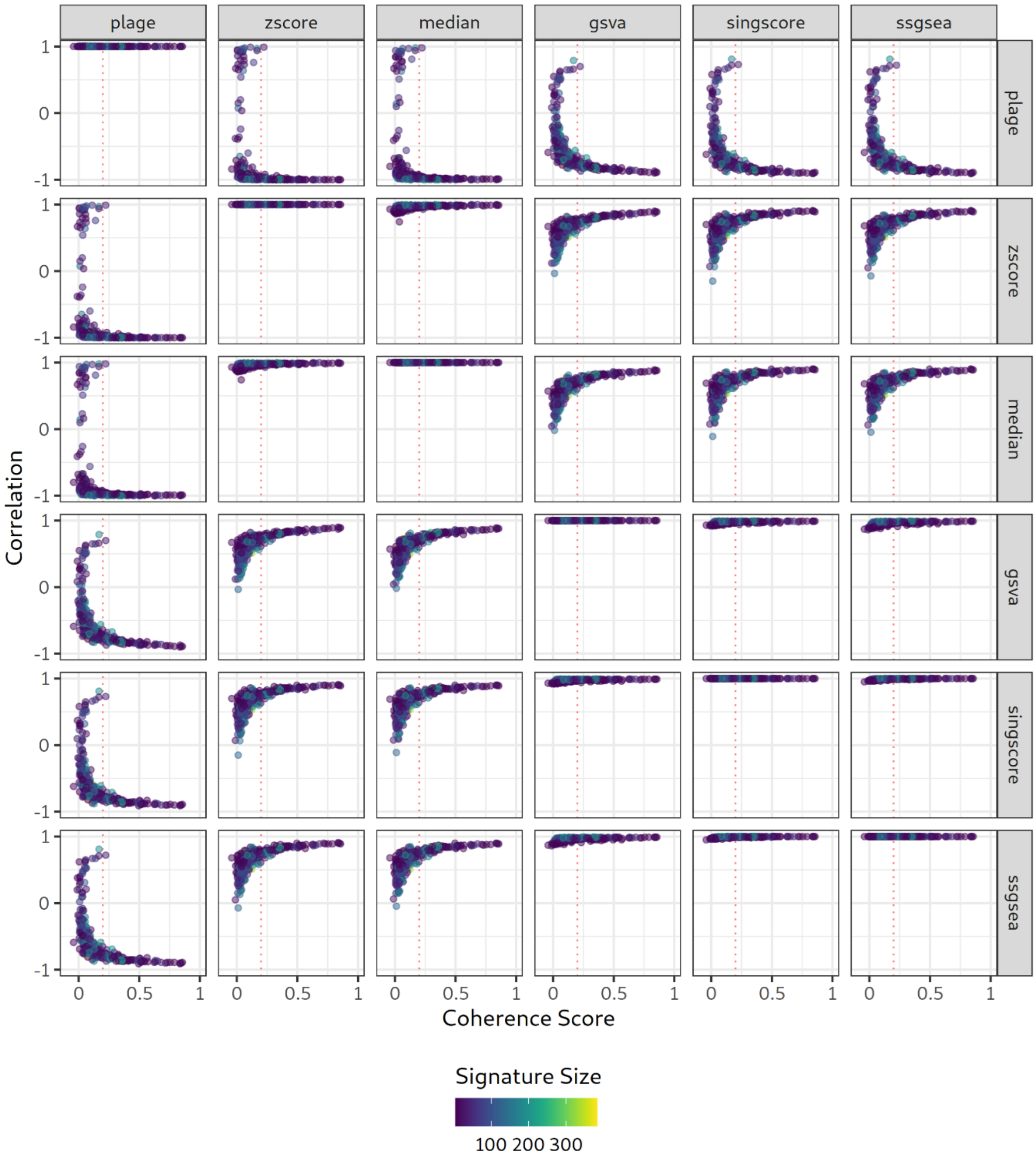

Figure 2.4: Evaluation of gene expression signature scoring approaches compared to the CS. Each point in the plot represents a gene expression signature of the RosettaSX collection. The x-axis represents the CS of a signature, and the y-axis represents the Pearson correlation between the gene expression signature scores calculated by the different scoring approaches. Each panel compares the methods indicated at the top and right sides of the plot. Colors describe the size of the signature. (Reprinted from Kreis et al. 2021)

Approach for Identifying Robust Gene Expression Signatures

The general problem I seek to solve with this methodological approach is the identification of relevant gene expression signatures in the context (e.g., TCGA BRCA) of interest. Thus, the first step in this approach is selecting an indication and dataset of interest (Figure 3.1). Subsequently, a set of filtering criteria allows the choice of relevant signatures (e.g., using the coherence filter). In the next step, relevant annotations can be added to the analysis (e.g., already defined cancer subtypes or relevant biomarkers). Finally, the filtered gene expression signatures are visualized in a heatmap that can either be clustered using hierarchical clustering or ordered by an annotation of interest. In the latter case, the gene expression signatures are correlated with the annotation and ordered by their association. This representation highlights gene expression signatures that are associated with the respective annotation. Similarly, the clustered representation highlights clusters of samples with mutual molecular characteristics.

RosettaSX Analysis of Breast Cancer Intrinsic Subtypes

BRCA continues to be one of the leading diseases in women worldwide and is expected to increase in the future (Xu et al. 2023). Previous studies described luminal A and B, basal-like, human epidermal growth factor 2 (HER2)-enriched, normal-like (Perou et al. 2000), and triple-negative (Brenton et al. 2005) as intrinsic molecular breast cancer subtypes. While luminal cancers usually express one or two hormone receptors (i.e., estrogen receptor [ER], progesterone receptor [PR]), the basal subtype usually does not express either receptor (Harbeck et al. 2019). These subtypes and prognostic markers like Ki-67 have shown clinical significance and are important prognostic markers in the clinics (Allison 2021; Loibl et al. 2024). To showcase a workflow of RosettaSX, I analyzed the TCGA BRCA cohort to recapitulate major findings in breast cancer. Figure 2.6 shows the final configuration of a RosettaSX analysis. The heatmap was first filtered for coherent gene expression signatures and then supplemented with important clinical markers, as provided by TCGA (ER, PR, and HER2 Immunohistochemistry [IHC] status) and pre-computed PAM50 subtypes (Fougner et al. 2020). Additionally, I supplemented the heatmap with important gene expression (*ESR1*: ER, and *MKI67*: Ki-67 proliferation marker) and genomic (*ERBB2*: HER2, and *TP53*) markers.

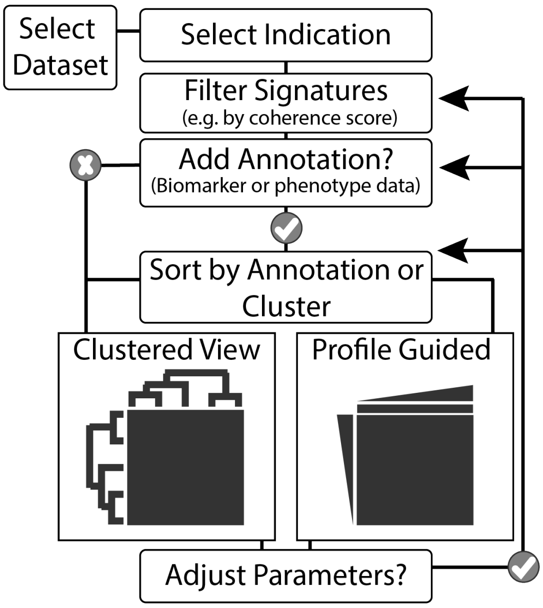

Figure 2.5: Overview of a typical workflow analyzing of gene expression signatures on RosettaSX. The user can choose from either cancer patients of the TCGA dataset or cancer models of the CCLE and the cancer indication of interest. Subsequently, the pre-computed gene expression signatures can be filtered for a specific set of signatures or by the CS (default filter CS > 0.2). Next, the user can analyze the gene expression signatures in a hierarchically clustered heatmap or select additional molecular (e.g., SNV, CNV, gene expression) readouts relevant to the planned analysis. Additional readouts will be added at the top of the heatmap as column annotations. If the user selects an annotation, there is the option to correlate the filtered gene expression signatures with the annotation and order the samples in descending order, automatically highlighting associated signatures. The heatmaps show signatures in the rows and samples in the columns. (Reprinted from Kreis et al. 2021)

First, I compared the annotated markers with the annotated intrinsic subtypes. In line with the description of intrinsic subtypes, the patients with luminal cancer have a higher expression of ESR1 and stained positive for ER (two-sided Wilcoxon rank sum test comparing ER status with *ESR1* expression, *p*-value < .0001, Figure 2.7 A.I) and PR, or both. Accordingly, most patients with basal-like BRCA are negative for the hormone receptors and thus comprise a large fraction of triple-negative cancers. Lastly, the HER2-like subtype is enriched for *ERBB2* copy number gains (chi-square, *p*-value < .0001, Figure 2.7 A.II).

Next, I compared the gene expression signature scores with the annotated intrinsic subtypes. A set of gene expression signatures by Calza et al. and Farmer et al. describe the PAM50 subtypes (bold row labels). Increased levels of the signatures descriptive for basal (Figure 2.6 clustered in the middle, Figure 2.7 B fifth plot), luminal A (Figure 2.6 clustered at the bottom) and *ERBB2*/HER2 (Figure 2.6 clustered below the luminal and basal signatures, Figure 2.7 B third

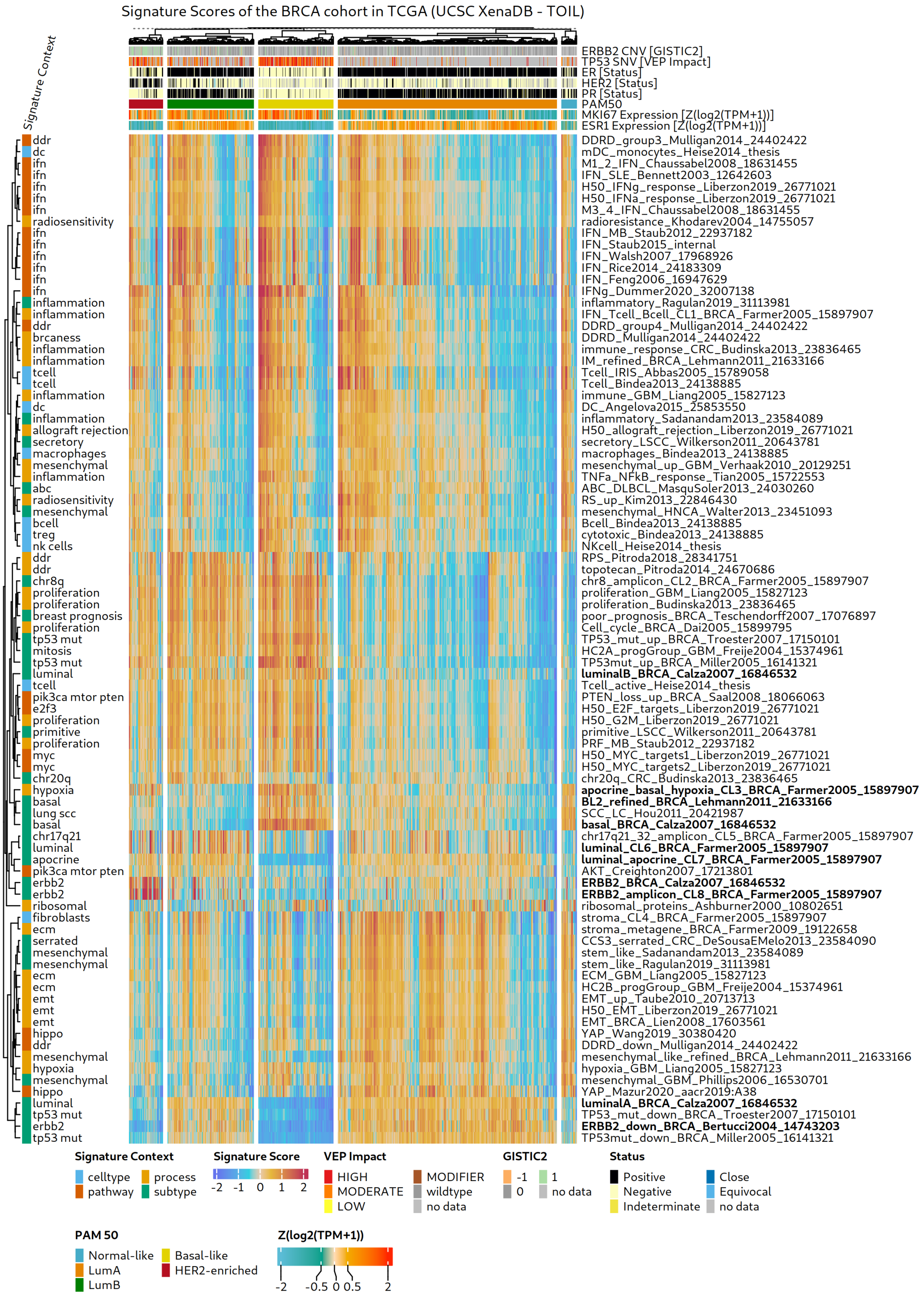

Figure 2.6: RosettaSX analysis of intrinsic molecular subtypes in TCGA BRCA. The heatmap is split into the intrinsic molecular breast cancer subtypes and shows gene expression signatures in the rows and patients in the columns. Red and yellow colors indicate increased gene expression signature scores and blue low scores. Additionally, the heatmap is supplemented with the receptor status (ER, PR, and HER2), gene expression markers (*MKI67* and *ESR1*), and genomic aberrations (*ERBB2* and *TP53*). The annotation on the left side categorizes the signatures into different biological phenomena. I filtered for gene expression signatures with CS ≥ 0.2, more than 60% of the signature genes are available and the signature size is between 3 and 300. (Adapted from Kreis et al. 2021)

plot) showed a good agreement with the annotated PAM50 subtypes. However, my analysis showed a substantial difference in the proliferation status between TCGA luminal B and luminal A classified subtypes. While the luminal A signature described the luminal intrinsic subtypes well, the luminal B signature (Calza et al. 2006) was also upregulated in the basal-like and HER2 enriched subtypes (Figure 2.7 B first plot). Additionally, the luminal B signature was strongly associated with proliferation signatures (e.g., Budinska et al. 2013; Phillips et al. 2006) and the expression of *MKI67* (Figure 2.7 A III). This pattern aligned with previous studies that indicated high proliferation scores in luminal B cancers (Feeley et al. 2014; Ades et al. 2014). Overall, this might indicate, that the luminal B signature rather describes proliferation signals.

Previous studies showed that TP53 mutations are most prevalent in basal-like breast cancer tumors (Mitri et al. 2022). In my analysis, the basal subtype was associated with mutations in *TP53* and elevated levels of signatures descriptive for *TP53* mutations (Miller et al. 2005; Troester et al. 2006). Additionally, the proliferation marker *MKI67* and proliferation scores were among the highest in this subtype (Budinska et al. 2013; Liberzon et al. 2015).

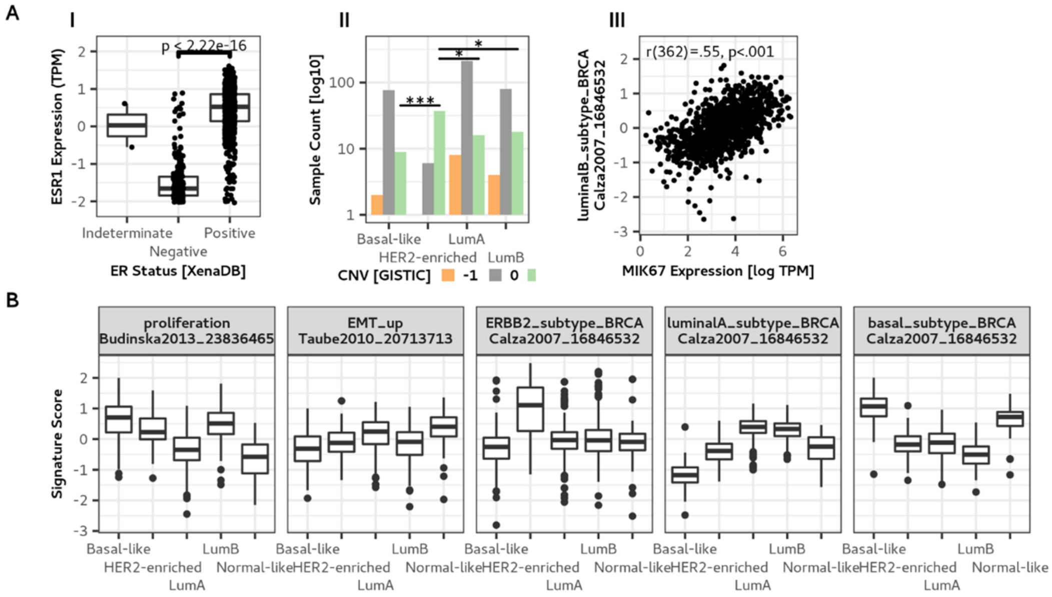

Figure 2.7: Detailed representations of TCGA BRCA RosettaSX analysis. **A.I**: Association between *ESR1* mRNA expression and ER IHC status. **A.II**: Number of patients with an *ERBB2* copy number alterations. Gain: 1 (green), loss: -1 (orange), diploid: 0 (grey), **A.III**: Comparison of the luminal B gene expression score (Calza et al. 2006) and the mRNA expression of the proliferation marker MKI67. **B:** Gene expression signature scores for different biological phenotypes and PAM50 subtypes in the intrinsic molecular BRCA subtypes. (Reprinted from Kreis et al. 2021)

At the upper part of the heatmap, there was a large cluster of signatures related to inflammation (Budinska et al. 2013; Ragulan et al. 2019; Lehmann et al. 2011), interferon signaling (Dummer et al. 2020; Liberzon et al. 2015) and the presence of cell types in the TME (Bindea et al. 2013; Angelova et al. 2015). Although the pivotal importance of hormone receptors is primal for patient stratification, more recent studies have started to evaluate the importance of immune response (Klopfenstein et al. 2021). For example, the presence of tumor-infiltrating lymphocytes (TIL) has gained importance and was found to be an important prognostic marker, especially in patients with triple-negative breast cancer, where TIL is associated with a better prognosis (Denkert, Wienert, and Klauschen 2018). Similarly, using cancer models, Lan et al. showed that type I interferon might be a predictive marker for chemotherapy-treated ER-negative patients (Lan et al. 2019).

Another cluster of signatures in the lower part of the heatmap indicated high gene expression signature scores for stroma, EMT, stemness, and mesenchymal gene expression signatures, especially in the normal-like and luminal A subtype (Sadanandam et al. 2013; Liberzon et al. 2015; Phillips et al. 2006; Farmer et al. 2009, Figure 2.7 B second plot). Previous studies in breast cancer hypothesized a substantial contribution of stromal cells to the expression profiles of normal-like breast cancers (Prat and Perou 2011). Accordingly, high levels of these signatures have been associated with fibroblast content in breast, colorectal, ovarian, and pan-cancer studies (Kreis et al. 2024; Puram et al. 2017; Izar et al. 2020; Tyler and Tirosh 2021), possibly indicating a strong influence if stromal cells in these profiles. However, concurrently, in-vivo studies showed the presence of cancer cell-specific epithelial to mesenchymal plasticity in breast cancer (Lüönd et al. 2021).

Overall, this analysis could recapitulate major characteristics of the inherent breast cancer subtypes and highlight their complexity.

2.4 Discussion

In this chapter, I introduced RosettaSX, an analytical workflow for selecting robust gene expression signatures, and a web server for analyzing cancer datasets. The server provides access to more than 11,000 cancer patients and cancer models. It guides the user through selecting a cancer indication, filtering for relevant gene expression signatures, and annotating relevant molecular of published information. Although individual components of this work have been used in the past, to my knowledge, no comparable service is available.

To demonstrate the capabilities of RosettaSX, I exemplarily applied my methodological approach to the characterization of intrinsic subtypes in patients of the TCGA BRCA cohort. In this analysis, I could recapitulate critical molecular characteristics of the subtypes, such as the high proliferative features of luminal B and basal-like subtypes (Feeley et al. 2014; Ades et al. 2014). Additionally, I was also able to point out mechanisms that are still under active research Lüönd et al. (2021), independent from the evolved intrinsic subtypes.

A key component of my approach is the gathered list of gene expression signatures that build the foundation of the analyses. This collection warranted robustness in broader applications across multiple cancer indications (unpublished work). The collection aims to comprise a set of sparsely overlapping gene expression signatures that describe fundamental biological phenomena in a controlled, redundant manner. Combined with the step-wise filtering of the collection, using the established CS (Staub 2012; Rahnenführer et al. 2004), I am confident that many cancer properties can be characterized. I evaluated the significance of the collection by thoroughly evaluating empirical CS distributions (using random gene expression signatures), and a comparison with a random gene expression signature collection or shuffled gene expression values highlighted the significance of these genes.

The main criterion for assessing gene expression signatures I use is the CS. Using empirical *p*-value distributions, I showed that the CS filters non-randomly generated signatures and, by that, identifies robust signatures. Additionally, the empirical *p*-values indicated that a CS larger than 0.2 filters for robust signatures of size three at a significance level of p-value $< .05$ for signatures. Furthermore, in the exemplary analysis of the TCGA BRCA cohort, multiple lowly overlapping gene expression signatures that describe the same biological mechanism passed the CS filter and were subsequently co-clustered.

In this chapter, I introduced a publicly available web service that enables users to utilize gene expression signatures to characterize their biomarker of interest. This platform combines well-established and frequently applied methods for gene expression analysis to exploit important knowledge hidden in gene expression signatures.

Chapter 3

Investigating the True Source of Increased EMT-Related Signature Scores

In this chapter, I will apply the introduced RosettaSX framework to evaluate a set of gene expression signatures. While the previous exemplary RosettaSX analysis primarily focused on the characterization of known subtypes in a larger cancer indication, this project focuses on comparing multiple gene expression signatures with the signature collection of RosettaSX. This analysis allows me to utilize gene expression signatures to correlate their expression profiles with the signatures of interest and explore associated biological phenomena. The subsequent analysis was published as part of my PhD studies in Clinical Research Communications (Kreis et al. 2024).

3.1 Project Outline

The ability of cells to change from a static epithelial state to a mobile mesenchymal state, called EMT, is involved in normal tissue development, wound healing, and cancer migration (J. Yang et al. 2020). Additionally, it was shown that EMT also enables cells to enter a stem cell-like state (Mani et al. 2008), which, in combination with the plasticity of EMT, is hypothesized to be an essential requirement for cancer metastasis (Xu et al. 2023). While the precise differentiation of these processes is still under active research (Wilson et al. 2020), several studies in BRCA, CRC, GBM, or HNSC have postulated gene expression signatures that describe cancer subtypes with EMT (Lien et al. 2007; Taube et al. 2010; Guinney et al. 2015; Eide et al. 2017; Liberzon et al. 2015), stemness (Ragulan et al. 2019; Sadanandam et al. 2013) or mesenchymal

(Walter et al. 2013; Verhaak et al. 2010; Phillips et al. 2006; Lehmann et al. 2011) characteristics (hereafter referred to as EMT-related).

Most of these studies analyzed gene expression data of bulk RNA sequencing (RNA-Seq) biopsies, data comprising cancer cells and cells from the tumor microenvironment (TME). Thus, expression profiles in these biopsies are based on complex mixed biological pathways describing the state of cancer cells and cells in the TME. Therefore, signals from non-cancerous cells may confound the derived gene expression signatures. Hence, low cancer cell content biopsies describe expression profiles of cells in the TME, and those with high content describe cancer-specific profiles. Individual studies in HNSC (Puram et al. 2017), ovarian (Izar et al. 2020), CRC cancer (Calon et al. 2015; Isella et al. 2015; H. O. Lee et al. 2020; Chowdhury et al. 2021) and pan-cancer (Tyler and Tirosh 2021) have shown that there is a high contribution of non-cancer cells for individual EMT-related signatures. Puram et al. showed that patients initially classified as mesenchymal HNSC could be assigned to other HNSC subtypes when correcting for the influence of stromal signals. Similarly, in CRC and breast cancer, studies of individual EMT-related signatures were shown to be influenced by signals of fibroblasts. Consequently, these signatures were indicated to be unsuitable for cancer cell lines models (Eide et al. 2017) and are strongly dependent on the TME (Puram et al. 2017; Lehmann et al. 2016; Chowdhury et al. 2021), resulting in discordant assignments of subtypes (Piskol et al. 2019). Although previous studies highlighted these issues in individual indications and signatures, a comprehensive analysis of EMT-related signatures in bulk sequencing and single-cell RNA-Seq (scRNA-Seq) is lacking.

Therefore, in this project, I investigated a set of gene expression signatures frequently used across different indications to describe cancer-specific EMT-related characteristics (Figure 3.1 left side). I will evaluate nine questions that systematically narrow down the signals that might drive EMT-related gene expression signature scores. Starting from a high-level perspective of bulk sequencing data in cancer patient tumors, I will evaluate their association with gene expression signatures of the RosettaSX collection (Figure 3.1 green cells). As this analysis primarily indicates the dependence of the signatures on the TME, I will subsequently analyze bulk sequencing RNA-Seq data of cancer models that lack influences of the TME (Figure 3.1 yellow cells). Finally, I will analyze the signatures in scRNA-Seq data to differentiate gene expression signatures on a single cell level (Figure 3.1 red cells).

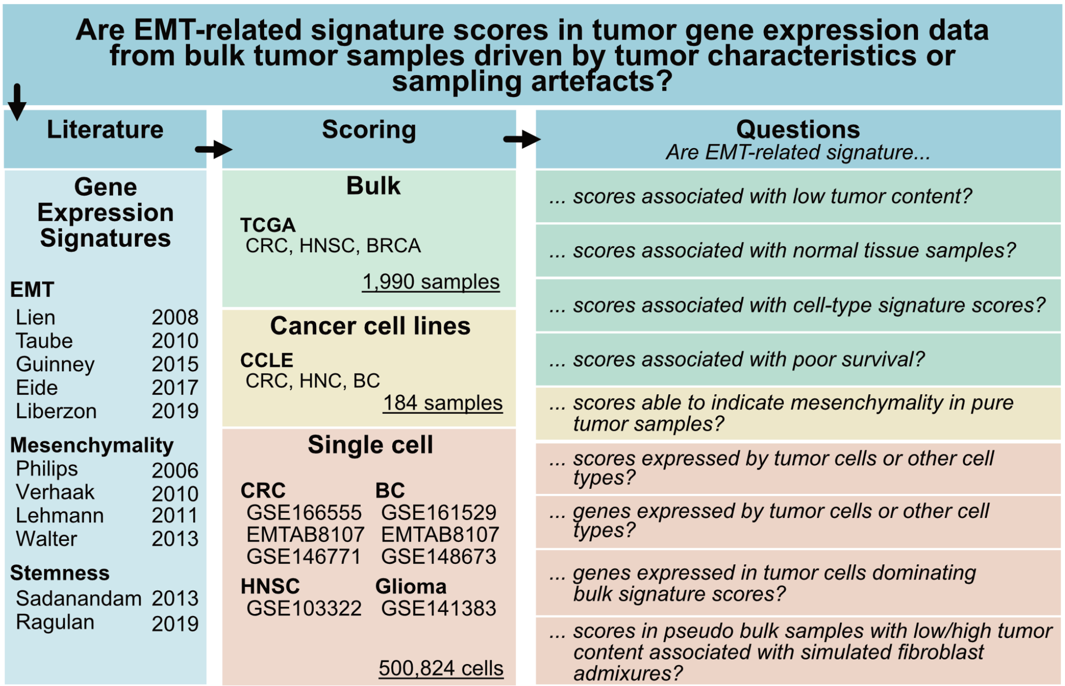
Are EMT-related signature scores in tumor gene expression data from bulk tumor samples driven by tumor characteristics or sampling artefacts?
Literature
Scoring
Questions
Are EMT-related signature...
Gene Expression Signatures
EMT
Lien 2008
Taube 2010
Guinney 2015
Eide 2017
Liberzon 2019
Mesenchymality
Philips 2006
Verhaak 2010
Lehmann 2011
Walter 2013
Stemness
Sadanandam 2013
Ragulan 2019
Bulk
TCGA
CRC, HNSC, BRCA
1,990 samples
Cancer cell lines
CCLE
CRC, HNC, BC
184 samples
Single cell
CRC
GSE166555
EMTAB8107
GSE146771
BC
GSE161529
EMTAB8107
GSE148673
HNSC
GSE103322
Glioma
GSE141383
500,824 cells
... scores associated with low tumor content?
... scores associated with normal tissue samples?
... scores associated with cell-type signature scores?
... scores associated with poor survival?
... scores able to indicate mesenchymality in pure tumor samples?
... scores expressed by tumor cells or other cell types?
... genes expressed by tumor cells or other cell types?
... genes expressed in tumor cells dominating bulk signature scores?
... scores in pseudo bulk samples with low/high tumor content associated with simulated fibroblast admixures?

Figure 3.1: Overview of investigated gene expression signatures, datasets, and questions analyzed in this chapter's context. I first selected 11 gene expression signatures used across many cancer indications. I subsequently investigated these signatures in bulk RNA-Seq data from cancer patient tumors and cancer cell line models to examine the questions listed on the right side in single-cell RNA-Seq data. (Reprinted from Kreis et al. 2024)

3.2 Methods

The following methods section is partly adopted from a paper written and published about this chapter in Cancer Research Communications (p. 517, Kreis et al. 2024).

Processing of bulk RNA-Seq and scRNA-Seq Data

I used the trimmed mean of the M-Values (TMM) method with default parameters to normalize bulk RNA-Seq TCGA and CCLE mRNA gene expression data (edgeR, version 3.36.0 Robinson, McCarthy, and Smyth 2009). I used the downloaded TISCH2 scRNA-seq data without further evaluating cell type clusters.

I applied a previously described method to integrate pseudobulk and TCGA bulk gene expression data (Barrett et al. 2022). In short, counts of bulk RNA-Seq gene expression were length-normalized (gencode Release 23, GRCh38.p3, Frankish et al. 2019), and subsequently, I applied TMM normalization on both pseudobulk and bulk RNA-Seq samples.

Simulation of Pseudobulk Gene Expression Data

To differentiate the contribution of fibroblast content in bulk sequencing RNA-Seq data, I simulated pseudobulk RNA-Seq data with different ratios of fibroblast and malignant cell RNA content. I used the SimBu R package (version 1.1.5, Dietrich 2023) and CRC scRNA-Seq data (GSE146771, see Table 3.1) to simulate 20 pseudobulk samples that were generated with 20%-80% of fibroblast cell RNA and 80%-20% malignant cell RNA.

Definition and Scoring of Gene Expression Signatures

For scoring and evaluating gene expression signatures, I used the previously described gene expression signature collection (Section 2.2) and gene expression signature scoring methods (Section 1.2). Similarly, I applied the previously described coherence filtering to select robust gene expression signatures (Equation 1.2).

However, for each cell type cluster in scRNA-seq data, I used the *addModule* to score gene expression signatures (Seurat, version 4.1.0, Hao et al. 2023, 2021; Stuart et al. 2019; Butler et al. 2018; Satija et al. 2015). For the pseudobulk data, I used the same scoring approaches described in Section 1.2 and calculated CSs on the integrated bulk RNA-Seq and pseudobulk data.

Besides the described signatures, I added gene expression signatures for the CMS4 subtype (EMT signatures). Guinney et al. and Eide et al. described two separate machine learning models for classifying CMS subtypes in patient data and preclinical models (Guinney et al. 2015; Eide et al. 2017). Thus, no gene expression signature is available, and I derived a signature as a proxy for the Guinney et al. CMS4 signature, I derived high-importance genes (using mean accuracy decrease) from the trained random forest model (CMSClassifier R package, version 1.0.0, Guinney et al. 2015). Accordingly, for the template genes provided by the CMSCaller package (version 0.99.2) by Eide et al., I extracted the CMS4 template genes. Subsequently, using a one-sided Wilcoxon rank sum test, I identified genes that had a significantly higher expression in the CMS4 compared to the CMS1-3 TCGA CRC cohort (with a $\log_2$ fold change larger two and a Bonferroni corrected p-value $< .05$, Bonferroni 1936). This procedure resulted in signatures of size 139, which I derived from the CMSclassified model, and 33, which I derived from the CMSCaller model. This procedure only approximates the original models but uses the information of the most important genes associated with CMS4.

Identifification of Cell Type Specifc Markers in scRNA-seq Data

I used the *FindMarkers* function of the Seurat package (version 4.1.0, Hao et al. 2023, 2021; Stuart et al. 2019; Butler et al. 2018; Satija et al. 2015) to find genes of a gene expression signature that are differentially expressed between cell type clusters (e.g., between fibroblasts and malignant cells). I filtered for genes with a fold-change larger than 2 and a p-value $< 1x10^{-10}$.

Survival analysis

I downloaded the clinical outcome endpoints disease-free interval (DFI) and overall survival (OS) from the supplementary data of (Liu et al. 2018). I fitted univariate Cox PH models to the EMT-related gene expression signature scores in BRCA, CRC, HNSC, and PAAD. Similarly, I fitted multivariate Cox PH models, which accounted for the American Joint Committee on Cancer staging (excluding patients with missing data) and adjusted p-values using the Holm method (Holm 1979).

Methods for the Visualization of Results

All results were implemented with R (version 4.1.1 R Core Team 2021). RosettaSX analysis results are visualized with the ComplexHeatmap R package (version 2.10.0), and heatmaps are clustered using Euclidean distance and complete hierarchical clustering. I used Pearson correlation to correlate heatmap annotations (e.g., the tumor cell content) with gene expression signature scores. Other plots are generated using ggplot2 (version 3.4.0). Tables are created with (version 0.8.0, Iannone et al. 2024) and gtsummary (version 1.7.0, Sjoberg et al. 2021). Statistical tests (i.e., Wilcoxon Rank sum test, Pearson correlation) were implemented with rstatix (version 0.7.1, Kassambara 2023) and survival analyses were implemented with the survival (version 3.3-1, Therneau and Grambsch 2000) package. For additional session information for this chapter see Section B.1.

Data Availability

TCGA and CCLE data were accessed as noted in Chapter 2 (Section 2.2). Additionally, I downloaded cancer cell content measures (Consensus Purity Estimation [CPE] and ABSOLUTE) from TCGAbiolinks (Mounir et al. 2019) and from the supplementary material of (Raphael et al. 2017). The TCGA deconvolution analysis uses precomputed data access from (Luca et al. 2021) and https://ecotyper.stanford.edu/carcinoma/.

Michail Yekelchyk downloaded the scRNA-seq data from the TISCH2 database (D. Sun et al. 2021; Han et al. 2023), and I used the data in Table 3.1.

Table 3.1: Listing of scRNA-seq datasets accessed from the TISCH2 database. The tumor type column indicates which tumor types of the respective dataset were used and how many cells were extracted (D. Sun et al. 2021; Han et al. 2023). (Downloaded by Michail Yekelchyk).

Accession	Reference	Tumor Type	Technology	Cells
EMTAB8107	Qian et al. (2021)	CRC, BRCA	10x Genomics	56,219
GSE148673	Gao et al. (2021)	BRCA	10x Genomics	10,359
GSE161529	Pal et al. (2021)	BRCA	10x Genomics	332,168
GSE146771	L. Zhang et al. (2020)	CRC	10x Genomics	43,817
GSE166555	Uhlitz et al. (2021)	CRC	10x Genomics	66,050
GSE141383	A. X. Chen et al. (2021)	Glioma	Microwell-seq	10,502
GSE103322	Puram et al. (2017)	HNSC	Smart-seq2	5,902

3.3 Results

I will analyze gene expression signatures in bulk sequencing data in the first section. The overarching goal of the study of bulk sequencing data is to differentiate signals originating from tumor cells, the TME, or the tumor macroenvironment (i.e., paired NAT tissue). An in-depth comparison of biological processes associated with the EMT-related signatures and the analysis of several types of bulk sequencing data will allow me to differentiate these signals, subsequently guiding me to identifying cell type-specific expression patterns.

TME processes are Associated with EMT-related Signatures

One way to utilize RosettaSX is to analyze the association of a signature with other previously published signatures and evaluate which other phenomena a signature is associated with. Here, I analyzed a set of gene expression signatures that describe immune-related (e.g., Finotello et al. 2019; Bindea et al. 2013; Samoszuk, Tan, and Chorn 2005), stroma (e.g., Farmer et al. 2009; Liang et al. 2005), cell of origin (Sadanandam et al. 2013; Budinska et al. 2013), oncogenic mechanisms (Staub 2012; Liberzon et al. 2015; Melo et al. 2013) and biological signaling pathways (Creighton 2007; Liberzon et al. 2015) and compared them with EMT-related signatures. First, I started with the analysis of CRC. I filtered the 285 gene expression signature collection of RosettaSX for a CS larger than 0.18 and selected signatures associated explicitly with the TME or oncogenic signaling pathways. I reduced the CS by 0.02 because individual subtype-specific gene expression signatures did not reach a CS of 0.2. However, as indicated in Figure 2.2, a signature larger than four is significant at a significance level of .05. I next grouped the signatures, which described similar phenomena (e.g., proliferation, cell types, pathways) and reduced the number of signatures within a group to at most three signatures using the Jaccard index (at most 0.25, excluding EMT-related signatures). The filtered signatures are displayed in Figure 3.3 in the left heatmap. Interestingly, cancer cell content was negatively correlated with EMT-related (Figure 3.3 left heatmap, Figure 3.2 A). Additionally, their profiles aligned with each other and also with different gene expression signatures describing cell types (e.g., B cell, T cell, macrophages Bindea et al. 2013; Thorsson et al. 2018), stroma, ECM or fibroblasts (Farmer et al. 2009; Samoszuk, Tan, and Chorn 2005; Liang et al. 2005) (Figure 3.3, Figure 3.4). While the epithelial markers *CDH1* and *EPCAM* are expressed at a lower level in EMT-high samples (Figure 3.2 B), the mesenchymal *VIM* marker is comparably higher expressed in these samples.

This analysis indicated that EMT-related signatures are significantly associated with low cancer cell content biopsies and signatures that describe the TME.

EMT-related Signatures are Less Coherent in TME-naïve Cancer Models

The above analyses indicated that EMT-related signatures are strongly linked to processes in the TME. By analyzing cancer cell lines, I can evaluate gene expression profiles originating from cancer cells, not TME cells. Thus, I next evaluated the cancer cell specificity of EMT-related signature scores. Figure 3.3, on the left side, the second heatmap shows the RosettaSX analysis of colorectal cancer cell lines. Interestingly, multiple gene expression signatures that describe the presence of cell types did not reach a CS larger than 0.2, indicating that the signatures should not be evaluated in these samples (which is expected, as they lack TME). However, individual stroma (Farmer et al. 2005) gene expression signatures reached a sufficient CS. These signatures had higher scores in a small subpopulation of cell lines with high mesenchymal gene marker expression (*VIM*) and low epithelial expression values (*EPCAM* and *CDH1*). A thorough

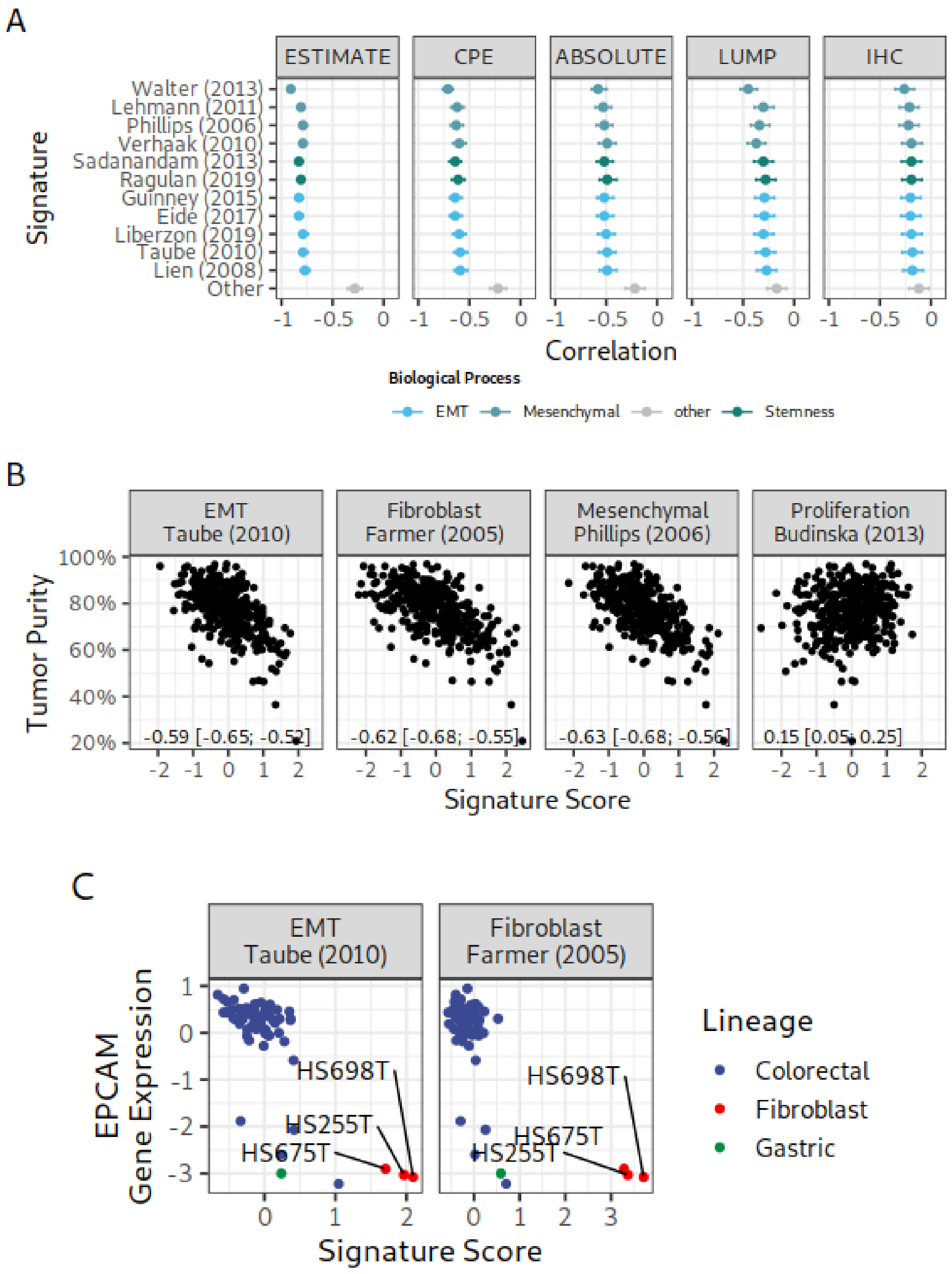

Figure 3.2: Gene expression signatures and their association with tumor content and the epithelial markers in cancer patient tumors and cancer cell lines. **A:** Analysis of the association of cancer cell content readouts with EMT-related signatures and other signatures of the RosettaSX collection (excluding TME-related signatures). **B:** Pearson Correlation of EMT (Taube et al. 2010), fibroblast/stroma (Farmer et al. 2009), mesenchymal (Phillips et al. 2006), and a proliferation signature (Budinska et al. 2013) with cancer cell content (CPE) (Mounir et al. 2019), bottom left annotation: Pearson correlation [95% CI]. **C:** Comparison of EMT (Taube et al. 2010) and fibroblast/stroma (Farmer et al. 2009) gene expression signature scores in relation to an epithelial gene expression marker (*EPCAM*). (Reprinted from Kreis et al. 2024)

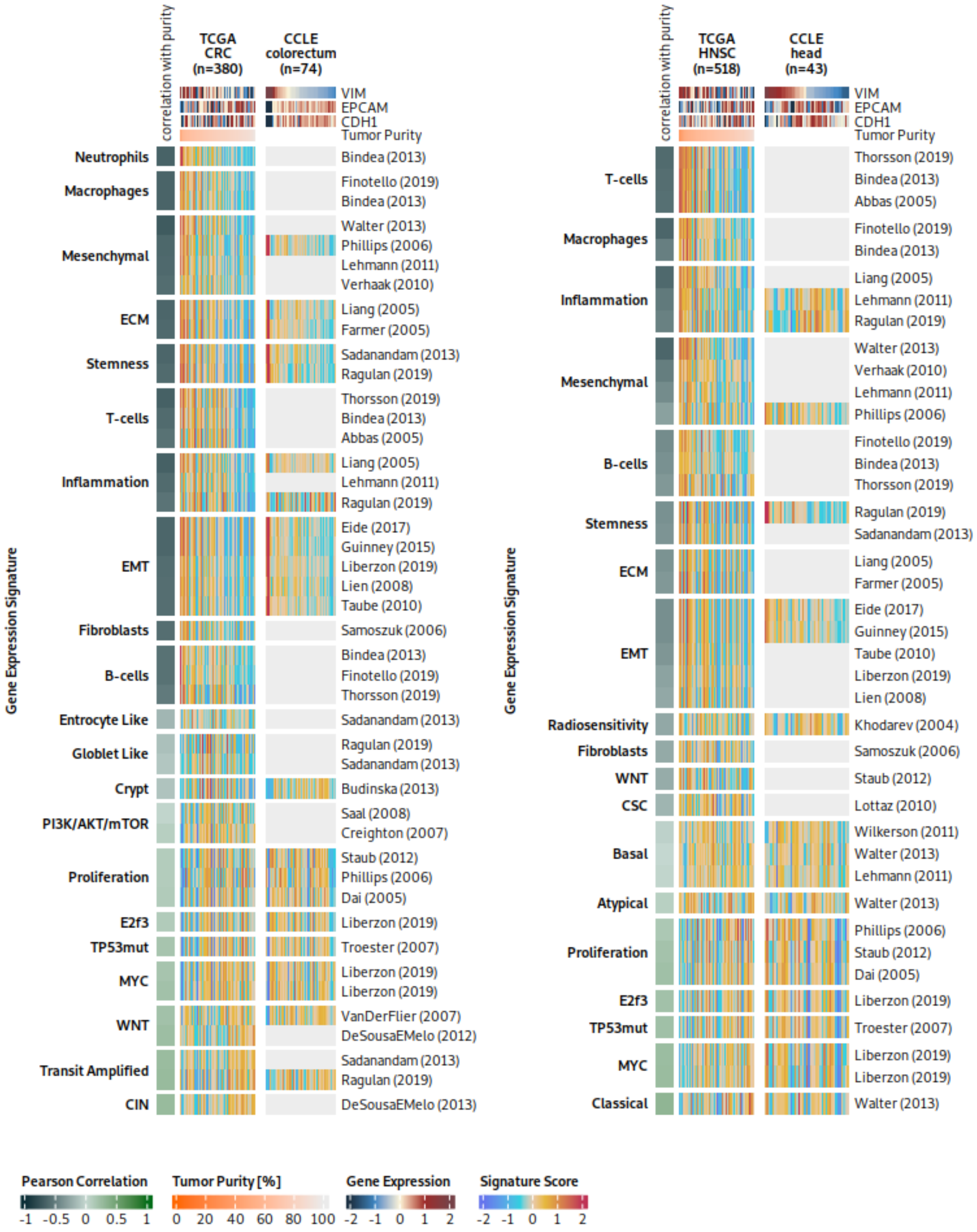

Figure 3.3: RosettaSX analyses of CRC (left) and HNSC (right) cancer patient tumors and cancer cell lines. The two pairs of heatmaps show gene expression signatures for filtered gene expression signatures (rows) for either TCGA tumors (columns) or cancer cell lines. Dark and light green indicates the correlation coefficient with tumor purity on the left side of the heatmaps. Additionally, two epithelial markers (*CDH1* and *EPCAM*) and a mesenchymal marker (*VIM*) are annotated above the heatmap. The annotated tumor purity orders cancer patient samples. Cancer cell lines are ordered by *VIM* expression. Gene expression signatures are ordered by their correlation with the sample's cancer cell content (primary tumors) or *VIM* expression levels (cancer cell lines). At the top of the heatmaps are gene expression signatures negatively correlated with cancer cell content. (Reprinted from Kreis et al. 2024)

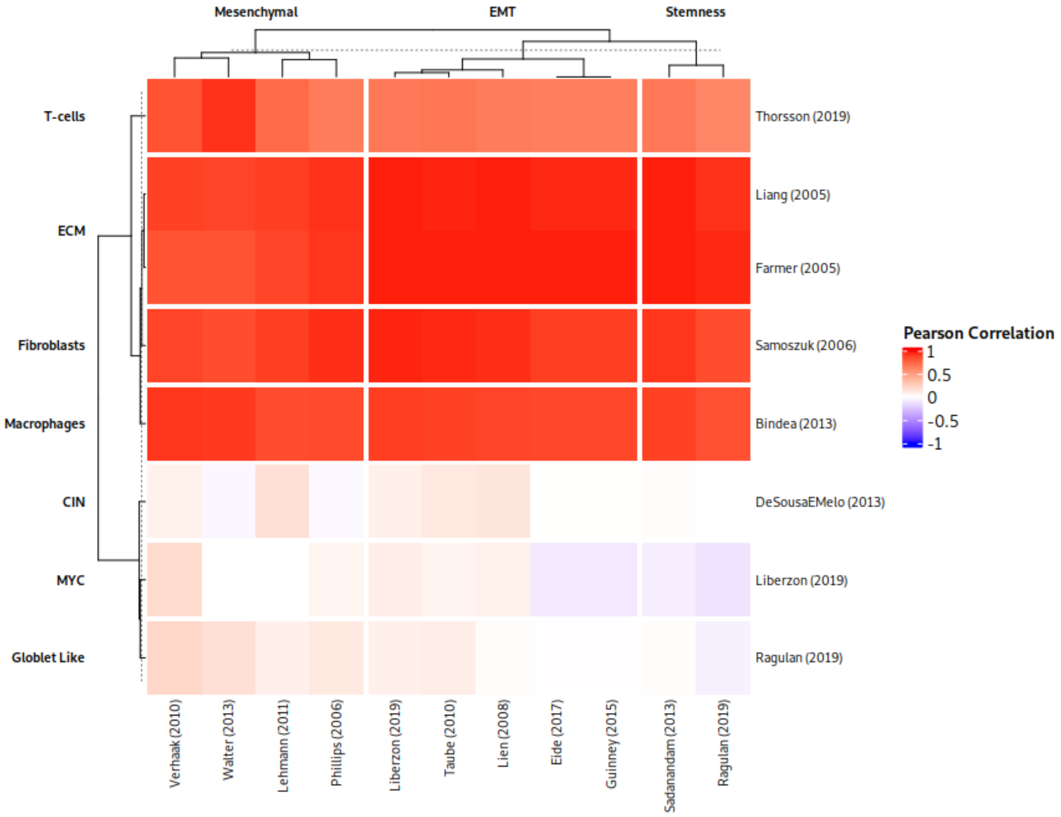

Figure 3.4: Pearson correlation coefficients between EMT-related gene expression signatures and signatures associated with low cancer cell content (T-cell, ECM, fibroblasts, or macrophages (Thorsson et al. 2018; Liang et al. 2005; Farmer et al. 2009; Samoszuk, Tan, and Chorn 2005; Bindea et al. 2013)) and high cancer cell content (CIN, MYC, Goblet-like (Liberzon et al. 2015; Ragulan et al. 2019; Melo et al. 2013)). Red: positive association, white: no association, blue: negative association. (Reprinted from Kreis et al. 2024)

investigation of cell lines with high signature scores indicated that they were lineages of fibroblasts (i.e., HS698T, HS675T, and HS255T; Figure 3.2 C). To evaluate the influence of these cell lines on the CS of the signatures, I removed the cell lines with a fibroblast lineage. I reevaluated the CS of the EMT-related signatures (Figure 3.5). Interestingly, the CS of all gene expression signatures decreased, highlighting the large effect of the cell lines population derived from a fibroblast lineage.

I then repeated the above-outlined RosettaSX analyses in HNSC, BRCA, GBM, and PAAD (Figure 3.3 third and fourth heatmap, Figure A.1, Figure A.2, Figure A.3). Congruent with the analyses in CRC, the signatures showed that low cancer cell content was associated with high signature scores in the analyses of cancer patients and correlated with the mesenchymal markers *VIM* in cancer cell lines.

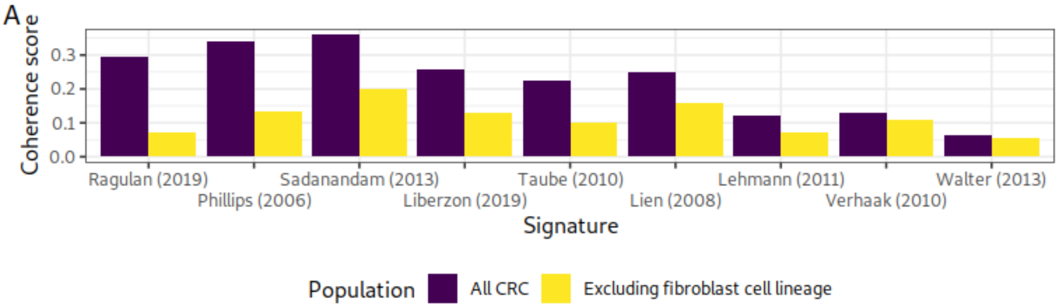

Figure 3.5: CS for the 11 EMT-related gene expression signatures in colorectal cancer cell lines. Purple: using all cell lines, yellow: excluding fibroblast lineages. (Reprinted from Kreis et al. 2024)

These results indicated that in TME-naïve samples, EMT-related signatures have uncoordinated expression profiles and are not cancer cell-specific.

The Environment of a Tumor Influences Scores of EMT-related Signatures

After I outlined the associations of EMT-related signature scores in cancer patient biopsies and TME-naive cancer models, I evaluated the expression of the signatures in the broader range of a tumor sample - normal tissue adjacent to the normal (NAT). TCGA provides paired tumor samples with NAT tissue for a subset of their patients. To evaluate the influence of the TME and the tumor macroenvironment (i.e., NAT used as a proxy), I divided the patient biopsies based on the cancer cell content into low and high. Subsequently, I compared the signature score levels in these samples with those of paired NAT samples (Figure 3.6). I noted that the distribution of cancer cell content across the indication is different, with the highest tumor content in BRCA (96.01%), CRC (95.69%), and lowest in HNSC (85.95%).

The analysis indicated that in CRC and BRCA, all gene expression signature scores were significantly lower in samples with low tumor content (Figure 3.6 lower comparisons). However, in HNSC, only the Walter et al., Verhaak et al., and Lehmann et al. gene expression signatures were significantly higher expressed in low compared to high tumor content samples (Lehmann et al. 2011; Walter et al. 2013; Verhaak et al. 2010). Interestingly, in BRCA (the cohort with the highest tumor content samples), there was a significantly higher expression of all signatures in the NAT samples than in the high cancer cell content samples (Figure 3.6 lower comparisons). While in HNSC, none of the signatures had higher scores in the NAT samples, in CRC, only the Walter et al. and Lehmann et al. signatures had significantly higher scores in NAT samples (Walter et al. 2013; Lehmann et al. 2011).

These results indicate a profound influence of the TME and the macroenvironment (i.e., NAT), especially in BRCA. This signal was less stringent in CRC and HNSC, possibly attributable to the lower cancer cell content.

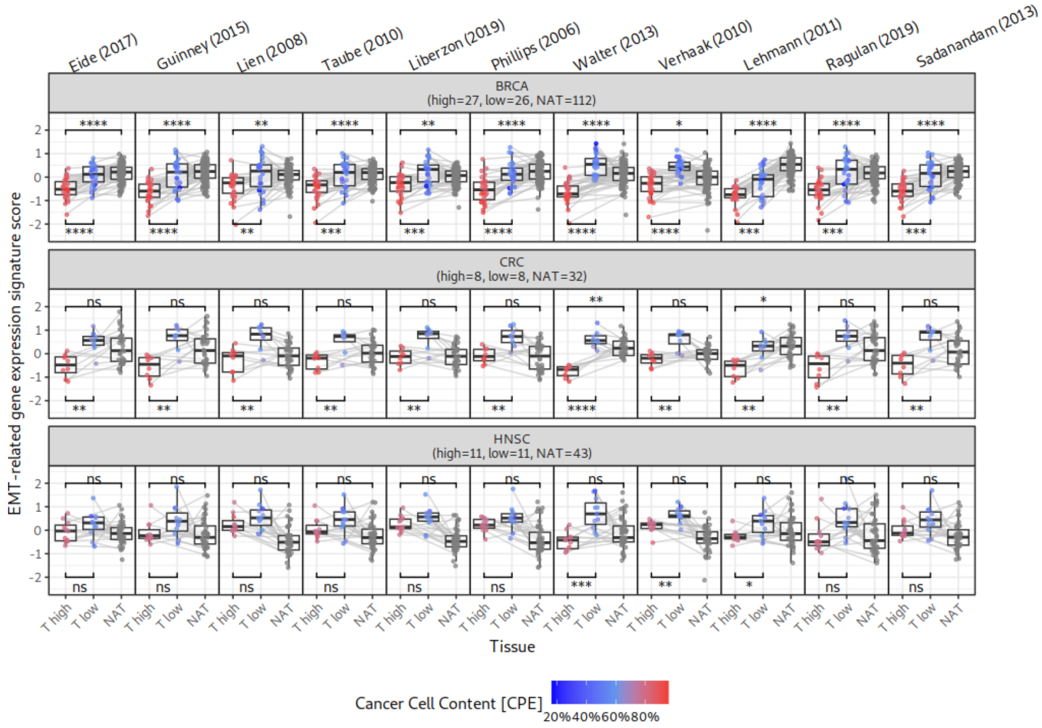

Figure 3.6: Gene expression signature scores in paired patient tumor (T) and NAT biopsies. For TCGA BRCA, CRC, and HNSC, the tumor samples are categorized into low or high cancer cell content samples using the upper and lower quartiles. The bars indicate my comparisons using a Wilcoxon rank sum test. ns: not significant, ***:** $p < .05$, ****:** $p < .01$, *****:** $p < .001$, ******:** $p < .0001$. (Reprinted from Kreis et al. 2024)

Signatures are Associated with Cell Types Enriched in Normal Tissue

Next, I evaluated the association of the gene expression signature scores with different cell types and cell type states using precomputed deconvolution cell type abundances (Figure 3.7). The data provides information on the cell type and state in tumor and NAT tissue. Interestingly, most signatures were strongly correlated with cell states specifically enriched in NAT tissue (e.g., epithelial cells, CAF2, M2-like monocytes, mast cells). Still, there was also a positive correlation with tumor-associated cell states; the strongest association was with pro-angiogenic epithelial cells. However, a gene that was indicated to be strongly associated with the pro-angiogenic epithelial cell state was *COL1A1*. Luca et al. listed the gene as one of the ten most associated with the pro-angiogenic epithelial cell state (Luca et al. 2021). However, simultaneously, the gene was also significantly related to fibroblasts by them and others (M. Li and Lu 2020; Mingyue Li et al. 2020; Y. Chen et al. 2023).

I repeated the analysis in the CRC, HNSC, and PAAD TCGA datasets (Figure A.4; Figure A.5; Figure A.6). Overall, the correlation between the cell state abundances and the EMT-related gene expression signatures was similar to that in BRCA. The signature scores correlated with cell states reported to be enriched in normal tissue. The only exceptions in CRC were Walter et al. and Verhaak et al. (Walter et al. 2013; Verhaak et al. 2010).

Overall these results indicated, that overall the signatures are often strongly correlated with cell types and states that are enriched in NAT tissue and also cell states in tumor tissue, but that the deconvolution of these cell states might be insufficient.

Malignant Cells Only Lowly Express EMT-related Signatures

The above analyses showed that signals from the TME primarily drive EMT-related signatures and that there is little evidence that they emerge from cancer cell-intrinsic signals. Therefore, I next analyzed multiple scRNA-Seq datasets to deconvolute the contribution of individual cell types to gene expression signature scores (Figure 3.8, Figure 3.9). While cell types with the highest expression of the gene expression signatures comprised myofibroblasts, fibroblasts, and endothelial, malignant cells had lower scores of the signatures, and only a tiny fraction of malignant cells expressed the signatures. This pattern was recognizable across cancer indications (BRCA, CRC, HNSC, and Glioma). The only exception of gene expression signatures not primarily associated with fibroblasts were two mesenchymal signatures by Walter et al. and Verhaak et al. (Walter et al. 2013; Verhaak et al. 2010). Although many fibroblasts, myofibroblasts, and endothelial cells expressed the signatures at a lower level, their expression was higher in monocytes and macrophages (Figure 3.9).

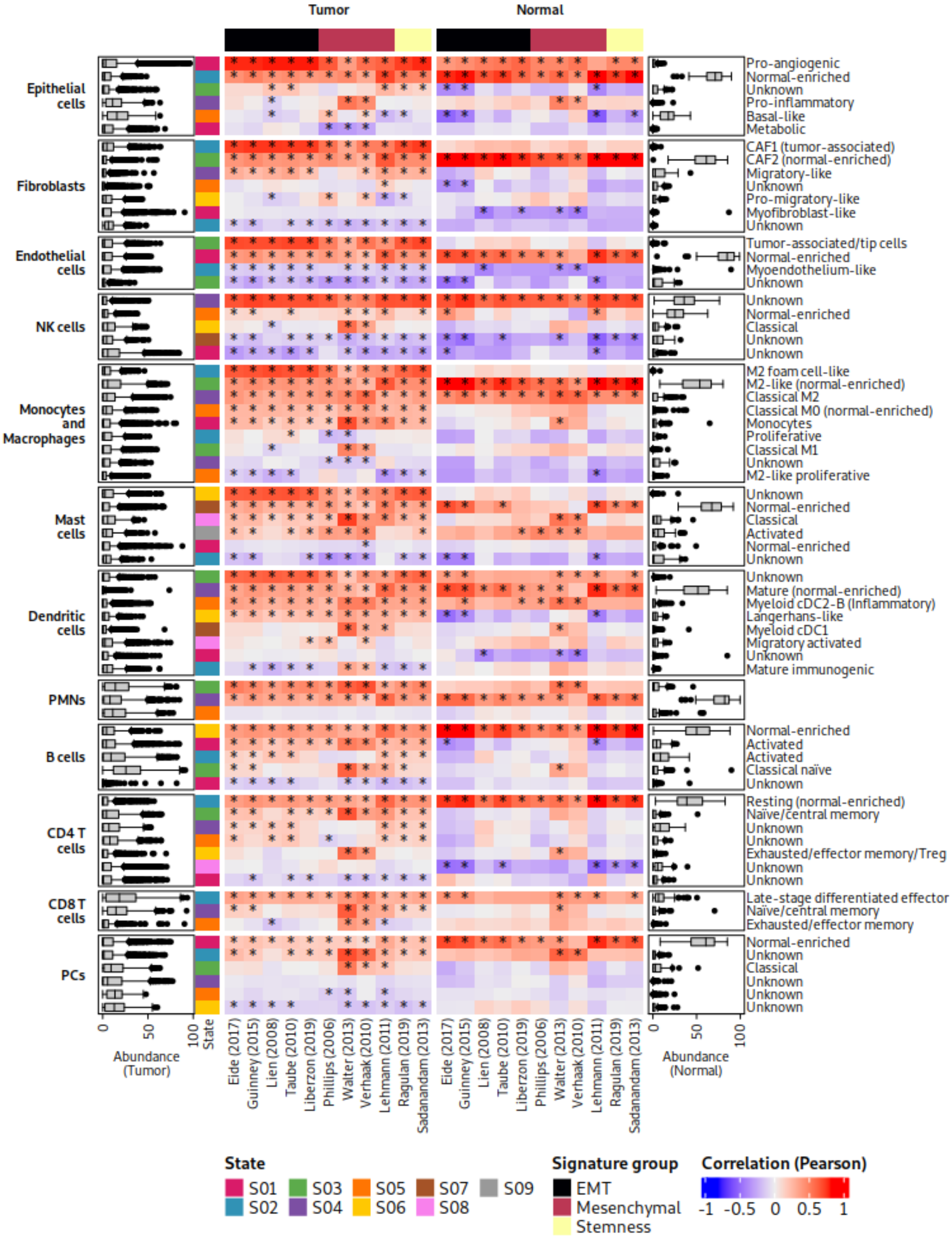

Figure 3.7: Deconvolution analysis of tumor and NAT samples in TCGA BRCA. The EMT-related gene expression signatures (columns) levels correlate with the abundance of different cell types and type states. Red indicates a positive correlation, and blue ia a negative correlation. ***:** significant association. (Reprinted from Kreis et al. 2024)

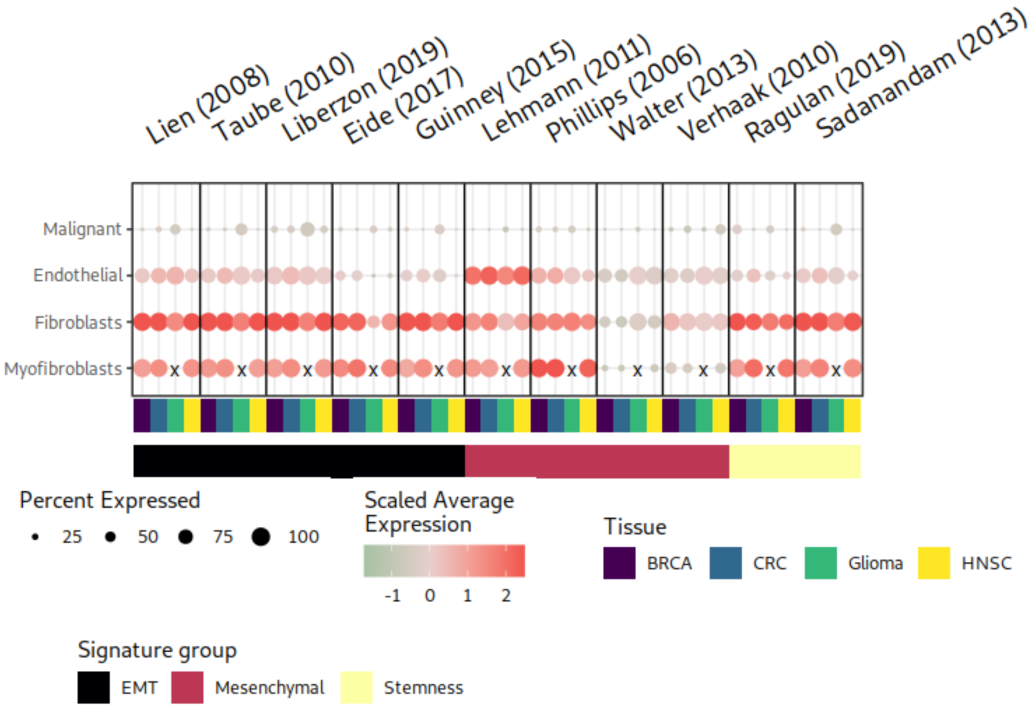

Figure 3.8: Single Cell RNA-seq gene expression signature levels for the studied EMT-related signatures in BRCA, CRC, HNSC, and Glioma. If multiple datasets for a cancer indication were available, I averaged the percent expressed and scaled average expression. Displayed are only malignant cells and cell types with a high percentage of expressing cells. Red: high expression, green low expression, small dot: few cells express the signature, large dot: many cells expressing the signature, x: cell type unavailable. (Reprinted from Kreis et al. 2024)

EMT-related Signature Genes are Sparsely Expressed in Malignant Cells

The previous analyses indicated a minor contribution of tumor cells to the expression levels of the EMT-related gene expression signatures. Thus, I evaluated if individual signature genes are significantly higher expressed in malignant cells than in fibroblasts (the major contributors from the above section). Table 3.2 shows the percentage of differentially expressed genes in malignant cells or fibroblasts across all analyzed EMT-related gene expression signatures. The maximal percentage of differentially expressed genes in cancer cells was 3.70% for the stemness signature by Ragulan et al. (Ragulan et al. 2019). Besides this, signature genes were frequently expressed in malignant cells (with up to 64% of the genes Phillips et al. 2006). This effect was most prominent for the stemness signatures (Ragulan et al. 2019; Sadanandam et al. 2013), a mesenchymal signature (Phillips et al. 2006), and individual EMT signatures (Taube et al. 2010; Lien et al. 2007; Liberzon et al. 2015) in BRCA and CRC. This analysis indicates that only a small fraction of genes are significantly more expressed in malignant cells than fibroblasts.

Table 3.2: Results of a differential gene expression analysis, comparing the expression of genes of gene expression signatures between malignant cells and fibroblasts. Signatures are grouped into EMT, stemness, and mesenchymal. Each column shows the percentage of differentially expressed genes of gene expression signatures in the malignant and fibroblasts across the different indications. (Reprinted from Kreis et al. 2024)

Geneset	BRCA[1]	CRC[1]	Glioma[1]	HNSC[1]
EMT				
Eide (2017)	2.1 (0.0)	2.6 (0.0)	0.0 (0.0)	5.6 (0.0)
Guinney (2015)	6.7 (0.0)	8.3 (0.0)	0.0 (0.0)	10.5 (0.0)
Liberzon (2019)	25.5 (0.2)	29.0 (0.0)	13.2 (0.5)	18.4 (1.0)
Lien (2008)	38.3 (0.0)	45.7 (0.0)	14.8 (0.0)	33.3 (0.0)
Taube (2010)	25.2 (0.4)	26.6 (0.0)	17.3 (0.0)	18.2 (1.1)
Mesenchymal				
Lehmann (2011)	3.4 (0.0)	5.7 (0.0)	2.9 (0.0)	3.5 (0.0)
Phillips (2006)	56.4 (0.0)	64.1 (0.0)	38.5 (0.0)	38.5 (0.0)
Verhaak (2010)	3.7 (0.0)	4.1 (0.0)	2.6 (0.0)	2.6 (0.7)
Walter (2013)	5.6 (0.3)	6.3 (0.0)	2.8 (0.0)	5.2 (0.4)
Stemness				
Ragulan (2019)	44.4 (3.7)	53.7 (0.0)	25.0 (0.0)	44.4 (0.0)
Sadanandam (2013)	27.4 (0.5)	35.7 (0.0)	8.2 (0.5)	25.9 (0.0)

[1] DEG fibroblast % (DEG malignant cells %)

Fibroblast-enriched Pseudobulk Samples Resemble Low Cancer Content Samples

To further evaluate the association of fibroblasts on bulk sequencing samples, I used a CRC single-cell dataset (H. O. Lee et al. 2020) to simulate pseudobulk (Dietrich 2023) samples with varying fibroblast cell content. I compared these pseudobulk samples with TCGA CRC samples in an integrated RosettaSX analysis (Figure 3.10). The study indicated a cluster of gene expression signatures that describe CRC subtypes: transit amplified (Ragulan et al. 2019), crypt (Budinska et al. 2013), or goblet-like (Ragulan et al. 2019). Additionally, a cluster with oncogenic processes like proliferation (Staub 2012) had higher scores in samples with intermediate and high cancer cell content (top and bottom clusters). Besides that, a large cluster of signatures is associated with cell types, processes describing phenomena in the TME (Samoszuk, Tan, and Chorn 2005; Liang et al. 2005; Bindea et al. 2013), and EMT-related gene expression signatures. In this cluster of signatures, the pseudobulk samples with high fibroblast content had higher expression levels and co-occurred with low cancer cell content.

This analysis highlighted the pivotal contribution of high fibroblast content to the EMT-related signature scores. Additionally, it became apparent that fibroblast-enriched and low cancer-cell content samples have highly congruent expression profiles.

Low Cancer Content Associated Genes Contribute Most to Signature Scores

In the previous sections, I analyzed the gene expression signatures either on bulk sequencing or scRNA-Seq data level. A combined analysis of the previously described readouts can differentiate which genes of a gene expression signature contribute most to the signature scores in bulk RNA-Seq data and show in which cell types these genes were primarily expressed. For this, I evaluated three measures: a) the correlation of a signature gene's expression values with the respective gene expression signature score, b) the correlation of a signature gene's expression value with tumor purity, and c) the log fold change of cancer cell expression vs. fibroblast expression from the single cell analysis. Figure 3.11. While genes with the highest contribution to the signature scores were lower expressed in high cancer cell content samples, those with a lower influence on the signature scores were primarily associated with low cancer cell content. Additionally, high-influence genes often had higher scores in fibroblasts compared to malignant cells, especially for the Liberzon et al., Taube et al., and Sadanandam et al. signatures (Taube et al. 2010; Liberzon et al. 2015; Sadanandam et al. 2013). The Genes most often present in these signatures and highly expressed in fibroblasts were *COL3A1*, *COL1A2,* and *COL1A1*.

Thus, for all signatures, I observed a high contribution of genes associated with low tumor content and often highly expressed by fibroblasts. This analysis indicated that all signatures strongly depend on signals from the TME, in most cases, on the fibroblast content.

EMT-related Gene Expression Signatures are Not Associated with Prognosis

Individual signatures of the herein analyzed EMT-related gene expression signatures have been described as prognostic (Sadanandam et al. 2013; Calon et al. 2015; Isella et al. 2015), but others could not recapitulate such findings (Mak et al. 2016; Tan et al. 2014). Thus, I next reevaluated their prognostic value in untreated TCGA patients. I evaluated their survival in univariate (Figure 3.12) and multivariate proportional hazard (PH) models that account for tumor stage, gender, and age of the patients Figure 3.13. Overall, there was no significant association between decreased disease-free survival and overall survival across any signature in any cohort.

3.4 Discussion

In this chapter, I decoded the contribution of signals stemming from tumor cells and cells in the TME to EMT-related gene expression signature scores. The in-depth analysis of intermingled gene expression profiles from bulk sequencing data to single cell level using scRNA-Seq data clearly showed that none of the 11 analyzed gene expression signatures was expressed by malignant cells, but cells in the TME, primarily fibroblasts. Although the herein analyzed set of signatures is, to my knowledge, the largest set of analyzed EMT-related signatures, the highlighted issues with these signatures are possibly also present in other indications.

The mesenchymal, stemness, or EMT characteristics have been characterized in gene expression signatures across multiple cancer indications. My analysis revealed that neither in the indications from which the signatures originated nor in other indications did these signatures describe cancer-specific mesenchymal characteristics (Figure 3.3, Figure A.1, Figure A.2, Figure A.3). Although previous studies showed separate analyses for individual gene expression signatures in ovarian, CRC, or HNSC (Calon et al. 2015; Isella et al. 2015; Puram et al. 2017), a comprehensive analysis as in this work was lacking. My studies showed that these signatures describe a strong signal observable across cancer indications (as indicated by the CS). However, phenomena in the TME and macroenvironment most often drive this association. Additionally, the EMT-related signatures had low CSs in cancer models (lacking signals from cells in the TME), reinforcing the high dependency on cancer cell extrinsic signals.

My results highlight that the TME strongly influences elevated levels of EMT-related gene expression. It became evident that especially in BRCA, the TME and the macroenvironment (NAT tissue) had higher signature scores than the samples with high cancer cell content (Figure 3.6). In CRC and HNSC, this effect was less pronounced, possibly attributable to the overall lower cancer cell content of all samples. This analysis indicated that sampling errors during a biopsy can significantly impact the scoring of these signatures. Thus, it is crucial to thoroughly control for a high cancer cell content when the goal is to derive such molecular phenotypes from the samples under study. To further differentiate the influence of signals of individual cell states stemming from cells in the tumor and NAT tissue, I performed a gene expression deconvolution analysis (Figure 3.7, Figure A.4, Figure A.5, Figure A.6). It became apparent, that high cell abundances of cell states, that were associated with NAT, strongly correlated with most of the EMT-related signatures. One of the cell states that had the highest correlation with EMT-related signature scores (pro-angiogenic epithelial cells). However, Luca et al., indicated that this cell state was strongly associated with the expression of *COL1A1*, which is also highly expressed by fibroblasts (M. Li and Lu 2020; Mingyue Li et al. 2020; Y. Chen et al. 2023). This possibly indicates a impure deconvolution of this cell state and strengthens the influence of fibroblasts on an increased expression of EMT-related signatures.

In my analyses of scRNA-Seq data, I found low or no contribution of malignant cells to EMT-related gene expression signature scores (Figure 3.8, Figure 3.9, Table 3.2). However, my analyses might lack granularity and the differentiation of small cancer cell populations. More recent concepts describe cells in the TME that guide small clusters of cancer cells via the concept of leader cells and tumor budding (Williams et al. 2019; Vilchez Mercedes et al. 2021). My analysis does not invalidate such concepts, as it lacks the granularity to detect such cell clusters. The main point of my analysis is that the analyzed gene expression signatures cannot differentiate such small populations in complex bulk sequencing cancer samples. Additionally, cells in the TME or macroenvironment often express the genes in these gene expression signatures at higher levels. Thus, these signatures are not sufficient to describe cancer-specific EMT-related processes.

While individual EMT-related gene expression signatures had an association with patient outcome proposed (Sadanandam et al. 2013; Calon et al. 2015; Isella et al. 2015), others could not recapitulate such findings. Except for TCGA PAAD, none of my univariate Cox PH models indicated a significant association between high signature scores and OS or DFI (Figure 3.12, Figure 3.13). Similarly, when correcting for confounding factors like age and gender in multivariate models, only in PAAD, there was a significant association. The discrepancy between these results can be manifold; previous studies separated patients into high and low gene expression signature scores, did not provide sufficient information on the source of outcome data (Sehgal et al. 2024), or subset the gene expression signatures (Calon et al. 2015). Consequently, these results, which do not describe characteristics of the complete set of gene expression signatures, are biased due to cutoff selection (Bennette and Vickers 2012; Busch 2021) or rely on low-quality data (Liu et al. 2018).

This chapter showed RosettaSX's capabilities in comparing a gene expression signature with other gene expression signatures and using the associated phenomena to characterize their profiles. Similar analysis of other biomarkers could be used to describe related phenotypes.

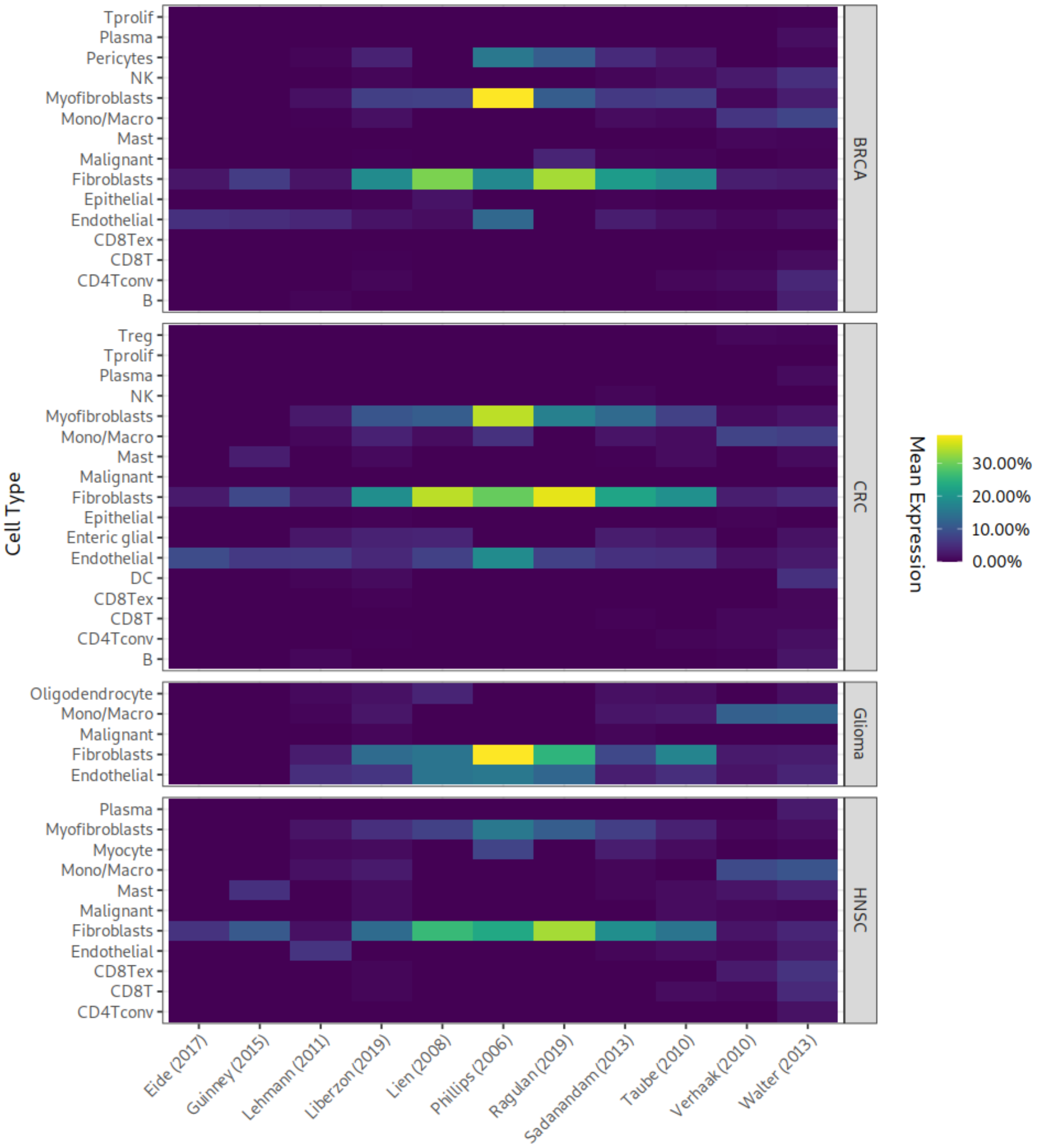

Figure 3.9: Scores for EMT-related signatures across different cancer cell types in scRNA-seq data from BRCA, CRC, HNSC, and Glioma. (Reprinted from Kreis et al. 2024)

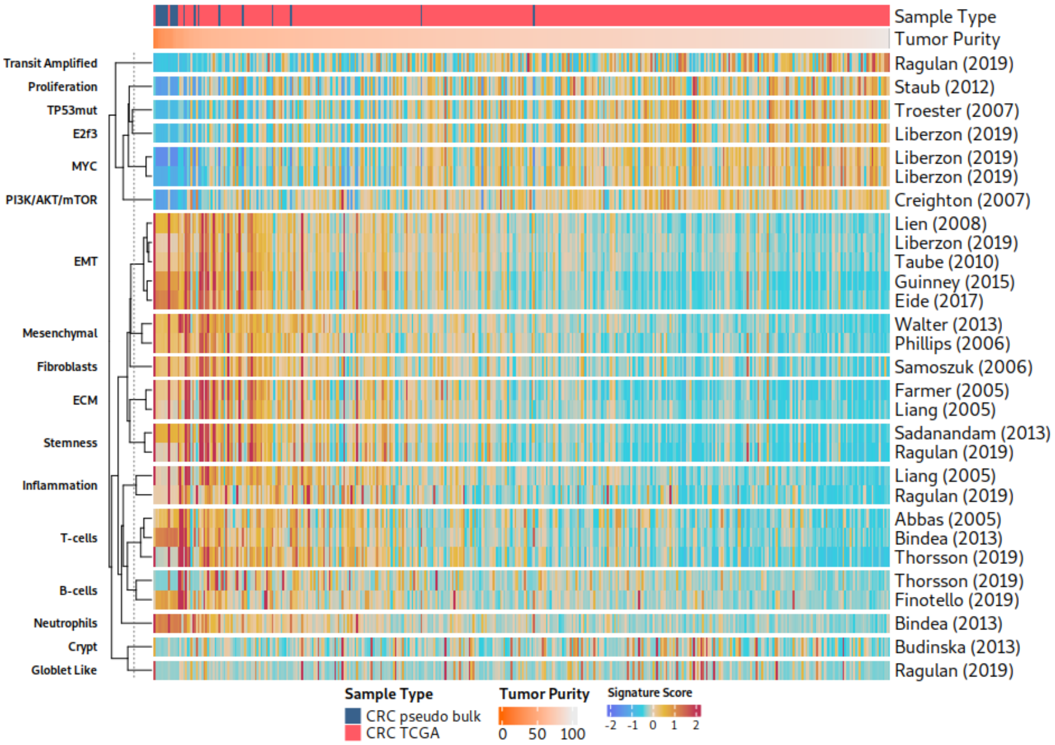

Figure 3.10: RosettaSX analysis of integrated TCGA CRC and pseudobulk samples was generated using CRC scRNA-Seq data. The pseudobulk samples are simulated from malignant and fibroblast cells only. Rows show the coherent gene expression signatures, and columns show the individual samples. High scores are indicated with red/orange colors and low scores with blue. The top annotation indicates the tumor purity (i.e., CPE values for TCGA cancer patient tumors and the fraction of sampled malignant cell RNA for pseudobulk samples) and the sample's origin above it (red: TCGA, blue: pseudobulk sample). Pseudobulk samples contain a fraction of 20% to 80% of RNA sampled from fibroblasts. (Reprinted from Kreis et al. 2024)

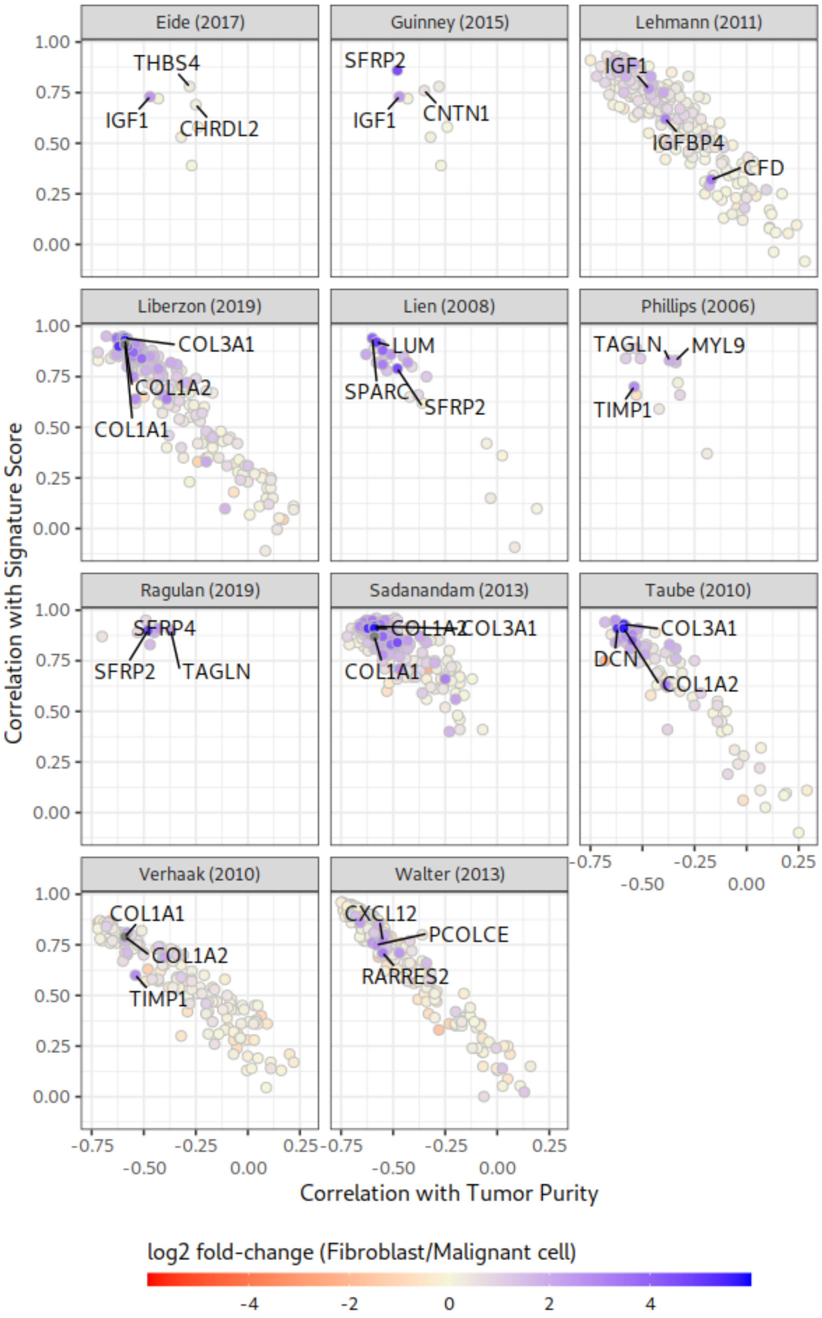

Figure 3.11: Analysis of multiple measures to evaluate the contribution and cell type specificity of individual genes in a gene expression signature. Each panel shows the results for EMT-related signature, and the axes show the correlation of the mRNA gene expression levels with the sample's cancer cell content (x-axis) and the respective gene expression signature (y-axis). Each dot in the panels represents a gene of the gene expression signature. The color indicates the fold-change between a comparison of malignant and fibroblasts (blue: high expression in fibroblasts). (Reprinted from Kreis et al. 2024)

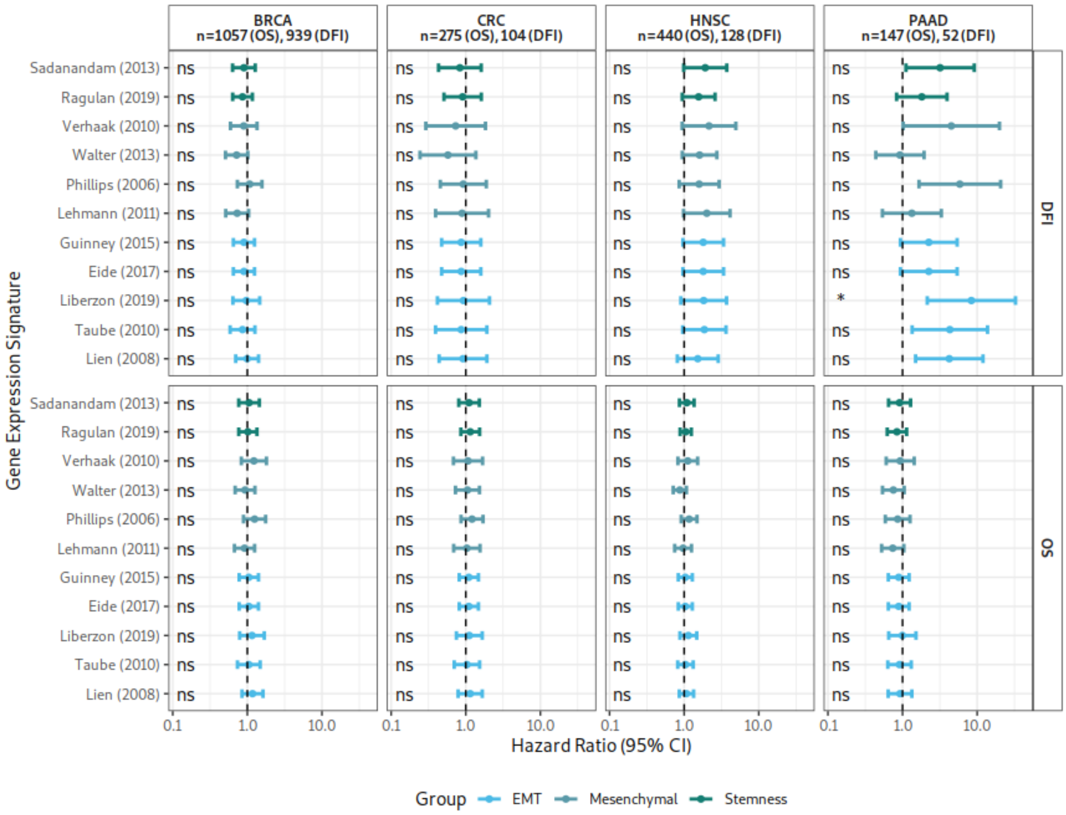

Figure 3.12: Overview of results from univariate Cox PH models in BRCA, CRC, HNSC, and PAAD (column panels), analyzing DFI and OS (row panels). The colors indicate the process described by the gene expression signatures, and each y-axis entry is one gene expression signature whose signature score was used in the Cox PH model. (Reprinted from Kreis et al. 2024)

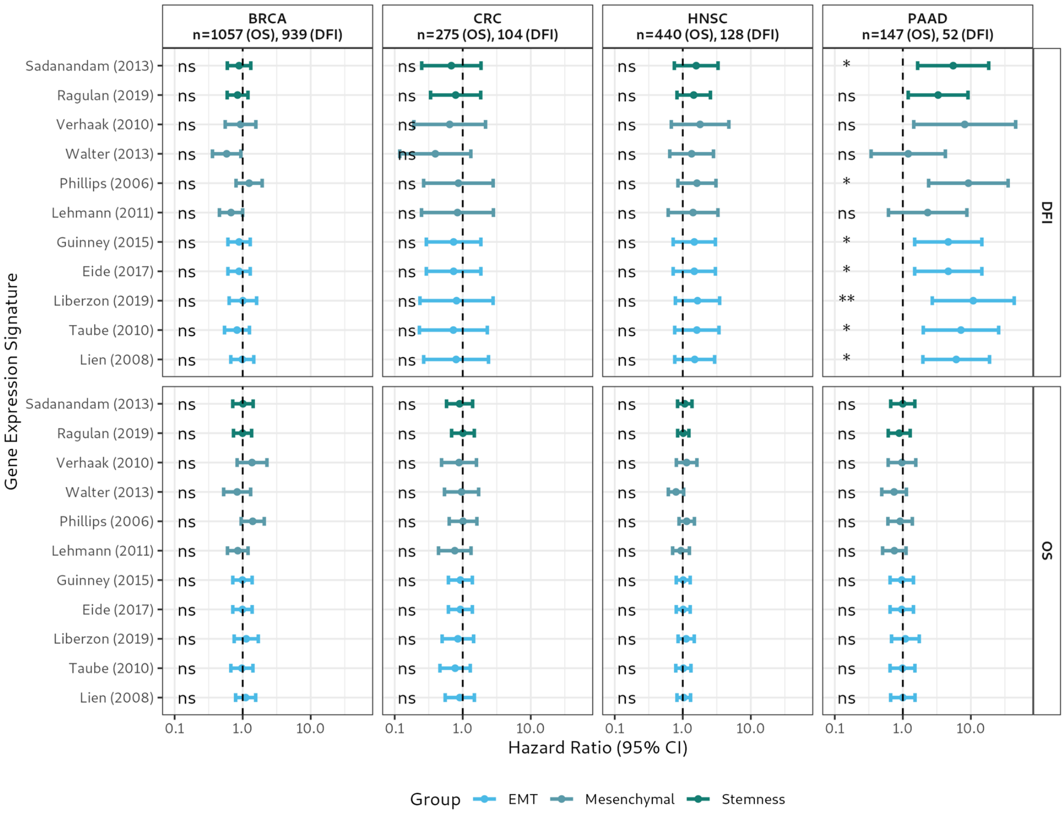

Figure 3.13: Overview of results from multivariate Cox PH models in BRCA, CRC, HNSC, and PAAD (column panels), analyzing DFI and OS (row panels). Each model was corrected for tumor stage, gender, and age of the patients. The colors indicate the process described by the gene expression signatures, and each y-axis entry is one gene expression signature whose signature score was used in the Cox PH model. (Reprinted from Kreis et al. 2024)

Chapter 4

Exposing the Underdiagnosis of Pulmonary LCNEC by Using RosettaSX

I applied my RosettaSX platform to characterize gene expression signatures in the previous chapters. In this chapter, I will apply the framework to identify and describe a rare lung cancer subtype - large cell neuroendocrine carcinoma (LCNEC). Firstly, I will use RosettaSX to identify a subpopulation of patients with NSCLC with high neuroendocrine gene expression signature scores. Secondly, I will train a machine learning model using neuroendocrine gene expression markers to subsequently investigate the underdiagnosis of LCNEC.

4.1 Project Outline

Neuroendocrine (NE) tumors are rare diseases that can occur among others in the lung. NE tumors in the lung comprise 15% of small cell lung cancer (SCLC), 3% (LCNEC), and 2% (carcinoids) of all lung tumors, respectively (Rekhtman 2022). While there are multiple transcriptomic and genomic subtypes for SCLC, LCNEC is less frequently characterized (George et al. 2018; W. Zhang et al. 2018). The main reason for this is its low prevalence and difficulties in pathological classification (Kinslow et al. 2020; Lantuejoul et al. 2020; L. Yang, Fan, and Lu 2022). The WHO guidelines recommend the presence of at least one positive NE IHC marker (*SYP*, *CHGA*, *NCAM1*), along with neuroendocrine morphology, high mitotic count, and visible necrotic tissue (Lindsay et al. 2021; Meihui Li, Yang, and Lu 2022). However, the classification is often difficult due to small biopsy sizes, limited testing for NE differentiation, and similarities with other lung subtypes (Derks et al. 2019; Rekhtman 2022; L. Yang, Fan, and Lu

2022). Additionally, the WHO's 2021 extension to differentiate pure LCNEC from those co-occurring with other NSCLC (called LCNEC combined) further highlights the complexity of PCNEC classification and possible reasons for an underdiagnosis of LCNEC (Lindsay et al. 2021; Kinslow et al. 2020).

Molecular studies can differentiate important markers, which can guide the stratification of patients for therapeutic options. In LCNEC, early genomic studies with up to 90 samples proposed two subtypes that indicate an NSCLC-like or SCLC-like characteristic (Yoshimura et al. 2021). While the SCLC-like subtype has characteristic co-alterations in *TP53* and *RB1*, the NSCLC-like subtype was proposed to have alterations in *TP53* and *STK11* or *KEAP1* (George et al. 2018). A large-scale analysis of genomic data from 1,429 patients with LCNEC recapitulated these genomic subtypes and extended the list of NSCLC-like alterations by *TP53*, *STK11*, *KRAS*, *CDKN2A*, *CDKN2B*, *MTAP*, *SMARCA4*, *CDN11*, *FRG3*, *FGF9*, *FGF13* and *CCND1* (Burns et al. 2024). Transcriptionally, Heijboer et al. indicated that ASCL1-positive patients (transcriptomic subtype in SCLC) are associated with poor prognosis (Heijboer et al. 2023). Clinically, patients suffering from LCNEC are primarily elderly males with poor survival (Kinslow et al. 2020). Compared to patients with NSCLC, those with LCNEC have worse prognosis and no standard therapy. The lack of sufficient patients hinders the evaluation of therapeutic outcomes between therapies. Patients are often treated either by NSCLC-like or SCLC-like therapies. Past studies indicated contradictory results regarding the benefit of NSCLC-like chemotherapy (CTx, e.g., pemetrexed, platinum combinations) or immune checkpoint inhibitors (ICI, e.g., atezolizumab) (Rossi et al. 2005; Naidoo et al. 2016; Sarkaria et al. 2011; J. M. Sun et al. 2012; Dudnik et al. 2021; V. E. Wang et al. 2017; Sherman et al. 2020; Igawa et al. 2010). A study that combined the classification of LCNEC based on genomic subtypes indicated a more prolonged survival for patients with NSCLC-like LCNEC when treated with medication frequently used for patients with SCLC (gemcitabine/taxane, platinum combinations) (Zhuo et al. 2020). Although these efforts provided essential insights into molecular variation and indicated the clinical significance of LCNEC subtypes, limited data availability hindered an in-depth association of these markers with clinical outcomes (Yoshimura et al. 2021; Zhuo et al. 2020).

Therefore, in the following, I evaluate the underdiagnosis of patients with LCNEC. To do so, I analyze molecular data from 5,329 patients with NSCLC to identify those with similar molecular features to pathologically classified LCNEC (pLCNEC) samples. This allows me to subsequently provide an in-depth characterization of the molecular LCNEC (mLCNEC), highlighting clinical, genomic, and transcriptomic differences from patients suffering from NSCLC.

4.2 Methods

This methods section, in part, uses text from a manuscript currently under review for submission.

Quality control of RNA-Seq data

To identify highly comparable gene expression profiles, I evaluated the distribution of RNA-seq read counts across all available NSCLC samples. A high deviation of read count distributions or a low or high library size (number of reads for a sample) potentially reduces the comparability of a sample. Consequently, I removed samples with less than 10 million, more than 60 million reads, or a median absolute deviation (MAD) count value that exceeded the interquartile range of all samples (MAD is within 25th, 75th quartile +/- 1.5 interquartile range). To further improve the comparability of this filtered set of samples, I normalized the data using the trimmed mean of M-values (TMM) (edgeR, version 3.36.0 Robinson, McCarthy, and Smyth 2009). Before applying the method, I filtered out genes with less than one count per million counts in 20% of the samples. Lastly, as two versions of the xR assay were used for the Tempus mRNA expression data generation, I applied the *removeBatchEffect* function of the limma package to remove batch effects (Ritchie et al. 2015).

The batch-corrected CPM values were used to compare gene expression markers and pairwise correlation analyses of mRNA expression data. I used the batch-corrected TPM values provided by Tempus to calculate gene expression signature scores.

Definition of Clinical Annotations

For each patient, I used different time points of the patient's clinical history as a reference for retrieving clinical annotations (this and the following section were elaborated with Jan Feifel). I analyzed the annotations for the survival analyses relative to the time of primary diagnosis or the first line of therapy administration. Similarly, I used the day of the biopsy to define the patient's

NE status for the machine learning model. On each of these reference days, I extracted patient demographics (gender, age, TNM stage, smoking status, and pathologist morphology) from the real-world Tempus database. If there was missing stage information for individual patients, I used Tempus TNM tumor staging information to impute the tumor stage information. For this, relative to the reference day, I accessed TNM staging information (T stage, N stage, and M stage) that preceded the reference day at most one year or occurred no later than 14 days after the reference day. Additionally, if TNM annotations appear within a 30-day window on different days, I combined them to imputate the tumor stage using the AJCC Cancer staging manual (Amin et al. 2016). Finally, I still lacked clinical annotations for individual patients after this procedure. Consequently, using predictive mean matching, I applied multiple chained equations (MICE) to impute stage and patient age with 20 imputations and 20 iterations.

That way, the pathologist's morphology annotations used by my model were accessed relative to the day of the biopsy, and the start date of the first line of therapy was used as a reference for extracting demographics that are used as covariates in the Cox proportional hazard models.

Patient Outcome Analyses

This study analyzes two types of patient outcomes: OS and progression-free survival (PFS). The analyses start on the day of the biopsy or the day of the first line of therapy administration.

For OS, patients were censored at the last known follow-up, and the day of the patient's death was used as a progression event. Similarly, PFS patients were censored on the day of the last known follow-up. The first recurrence, metastasis, or progressive disease outcome events defined the end day for PFS. To account for delays in data acquisition, progression events within a 15-day window after the reference date were ignored.

In all analyses, I accounted for immortal time bias due to the date of biopsy or shifted administration of therapies (start is when the last medication was given if multiple medications are part of a line of therapy) using the landmark approach (Gleiss, Oberbauer, and Heinze 2018). In brief, if the biopsy or the last medication of a line of therapy with medication combinations fell into the period between start and end (immortal day), the number of days between the start date and the immortal day was used as a starting point for subsequent analyses.

All analyses were terminated if less than 10% or less than eight patients were available for a stratum to ensure statistically reliable and unbiased results due to small sample sizes.

The left-truncated time-to-event data is visualized by Kaplan-Meier plots (survminer, version 0.4.9, Kassambara, Kosinski, and Biecek 2021) and supplemented with p-values derived from univariate Cox PH models. For multivariate Cox PH, I evaluated the influence of covariates (gender, age, or TNM stage) using the Akaike Information Criterion (AIC). AIC provides a measure to assess model complexity and the model fit.

Apply Machine Learning to Classify Molecular LCNEC

For this classification task, I hypothesized that a small fraction of the patients were falsely classified as non-NE differentiated (1-2%) and thus expected a class imbalance between NE and non-NE patients. Additionally, I needed concrete examples of non-NE patients, which I lacked, because I hypothesized that not all patients classified as non-NE were indeed non-NE. A group of methods that do not rely on a complete set of defined positive and negative classes is positive unlabeled (PU) learning (see 1.4.2). One of these methods uses bagged sets of data in combination with a support vector machine model (SVM) to predict the class of unlabeled samples (Mordelet and Vert 2014). Here I adopted this approach and used the gene expression values of 151 genes that were published in neuroendocrine gene expression signatures or indicated in association with a neuroendocrine phenotype, SCLC, LCNEC, prostate, or pancreatic cancer (W. Zhang et al. 2018; Ostano et al. 2020; Crona and Skogseid 2016; Tsai et al. 2017; Beltran et al. 2011; Simbolo et al. 2019; Balanis et al. 2019) as features for the PU-learning approach.

For this approach, samples were either categorized into positive (i.e., the class to be predicted) or unlabeled (i.e., the remaining patients) and split into t bags (subsets of data) of size k. Samples are resampled with replacements from a training data set. Therefore, each bag is comprised of all positive samples and k unlabeled samples. Subsequently, a model is trained on each bag separately using the unlabeled samples as negative class. In my case, samples annotated with a neuroendocrine morphology by a pathologist represented positive, and all other samples were unlabeled. The aim was to detect positive samples among the unlabeled ones. Expression features are scaled and centered. Features showing zero variance were removed from the training data.

For the model training, the complete data is divided into 90% training and 10% testing data along the patient dimension. Using the training data, t bags are sampled, and model is trained independently for each bag. For each bag, three bootstraps of the bag data are used as an external set. A random number of features is selected from the one with the strongest association with the positive class (Wilcoxon rank sum test). Next, hyperparameters are selected using two repeats of three-fold cross-validation on the internal set. After choosing the best hyperparameters (using the F_1^{PU} metric, Section 1.4) from the internal cross-validation validation sets, the final model with the best features is selected from the external bootstrap validation sets. Finally, the bag configuration (i.e., t and k) resulting in the best performance on the test set is selected as the final model (i.e., the data that the model never saw).

Finally, each model predicts the probability of a sample belonging to the NE class. Each model only predicts the probability for all positive and unlabeled samples that were not part of the bag on which the model was trained. Finally, a patient's LCNEC status (from here on called mLCNEC) is determined by the average prediction probability across all models that predicted a status for that specific sample. Samples were classified as mNSCLC if their LCNEC prediction probability fell between 0% and 25%, as mLCNEC if it ranged from 75% to 100 %, and as ambiguous if it ranged from 25% to 75%. Ambiguous samples were excluded from downstream analyses.

The analyzed data is highly imbalanced on the expected numbers of examples for positive and negative training cases. For the PU-learning approach, the training data is expected to comprise false negatives in the unlabeled data (patients with mLCNEC). Thus, standard metrics for model performance evaluation are strongly affected and not used here. Instead, a previously described performance measure for comparing models (W. S. Lee and Liu 2003) was used for model evaluation (see Section 1.4, Equation 1.3).

The model reached a F_1^{PU} of 15.13 for the training data and 12.04 for the test data. High values indicate better model performance, but the upper limit of the measure is unbound and lacks a straightforward interpretation. The model's positive predictive value (PPV, train: 0.31, test: 0.29) and negative predictive value (NPV, train: 1.00, test: 1.00) are highly biased due to the nature of PU-learning problems. While the high NPV is explained by the high portion of negative samples expected in the unlabeled data, the PPV turns out lower due to the low fraction of positive cases in the unlabeled data, which is larger than the number of available positive samples.

Statistical Analyses

All statistical tests were implemented with functions of the rstatix package (version 0.7.2, Kassambara 2023) and base R functions (R version 4.1.1 R Core Team 2021), for additional session information see Section B.1. To compare gene expression values or gene expression signatures between groups, I used Wilcoxon rank sum tests and corrected for multiple testing using Holm correction (Holm 1979). When comparing counts across a 2x2 table (e.g., mutation status within mLCNEC cohorts), I applied Fisher's exact test and Chi-squared tests for larger count tables (e.g., number of line of therapies [LoT] between cohorts). Lastly, Pearson correlation coefficients were used to compare gene expression values or signature scores.

Data Availability

The data analyzed in this study were part of the real-world multi-omics cancer database assembled by Tempus AI, Inc. The data is subject to controlled access for privacy and proprietary reasons. Tempus will make access to de-identified data available pending a signed data use agreement. Requests for access should be sent to publication.inquiry@tempus.com and will be responded to promptly, starting the release date of this study.

I accessed de-identified targeted sequencing data from lung cancer tumor samples via the Tempus database. Depending on the assay version, the tumor samples were profiled using the Tempus xT assay, a DNA-Seq panel capturing 598 or 648 genes. The tumors were also profiled using Tempus xR, an RNA whole-transcriptome assay (19,396 genes).

4.3 Results

Analysis of Neuroendocrine Differentiation in an NSCLC Cohort

I applied my RosettaSX framework on the complete NSCLC cohort to analyze samples with NE differentiation. I first filtered for RNA-Seq samples with sufficient quality (i.e., comparability of expression distributions) and analyzed identified gene expression signatures with coherent gene expression signatures (CS $>$ 0.2, Figure 4.1). The analysis indicated a population of NSCLCs that had a high expression of NE gene expression signatures (W. Zhang et al. 2018;

Ostano et al. 2020). A subset of these carcinomas was classified as LCNEC (i.e., LCNEC or Adenocarcinoma [ADC] with NE differentiation), but most of these cases were classified as other NSCLC subtypes (i.e., ADC, SqCC). This indicated that a substantial portion of samples was falsely classified as other NSCLCs, even though, they showed signs of NE differentiation in my analysis.

A Machine Learning Method to Identify Molecular LCNEC

Overall, in the Tempus NSCLC cohort, the original pathological annotations classified 1.77% of the patients as LCNEC. Based on previous reports, I expected an increase of at least 1% (Rekhtman 2022). However, as I retrospectively lacked thorough IHC testing for the positivity of NE markers across all patients with NSCLC, I implemented a machine-learning model that, for each patient, predicts the probability of LCNEC-likeness. For this, I gathered a list of 151 NE markers that have been associated with a neuroendocrine phenotype in neuroendocrine tumors (NET) (Balanis et al. 2019; Crona and Skogseid 2016), prostate cancer (Tsai et al. 2017; Beltran et al. 2011), pancreatic cancer (Ostano et al. 2020), LCNEC (Simbolo et al. 2019), or SCLC (W. Zhang et al. 2018). For my hypothesis, which assumes that there are patients with molecular LCNEC (with LCNEC-like gene expression profiles, subsequently termed mLCNEC), I used the expression of these genes to differentiate patients with LCNEC from those with NSCLC. By that, the identified patient population will share molecular characteristics with pathologically classified LCNEC (pLCNEC). However, unsupervised machine learning approaches require true positive (i.e., NE) and true negative (i.e., non-NE) patients, which I lacked because I could not label patients as non-NE for sure. Therefore, I applied a PU-learning approach to identify mLCNEC patients (Mordelet and Vert 2014) (see methods). Before using my model, I removed 671 (11.18%) patients, which indicated a deviated count distribution across all genes (see methods). For the final patient classification, I averaged the prediction probabilities of all models. Each model only predicted the status of a sample if it was not part of the model training data (pLCNECs are an exception, as they are part of all parts, Figure 4.2 A). Finally, I classified patients with an average prediction probability larger than 75% as mLCNEC, ambiguous with a prediction probability between 25% and 75%, and below 25% as mNSCLC (molecular NSCLC), resulting in cohorts of size 201 (mLCNEC), 4,795 (mNSCLC) and 333 (ambiguous) (Figure 4.2 B). Interestingly, 24 (17 ambiguous and 7 mNSCLC) patients with pLCNEC had a low or

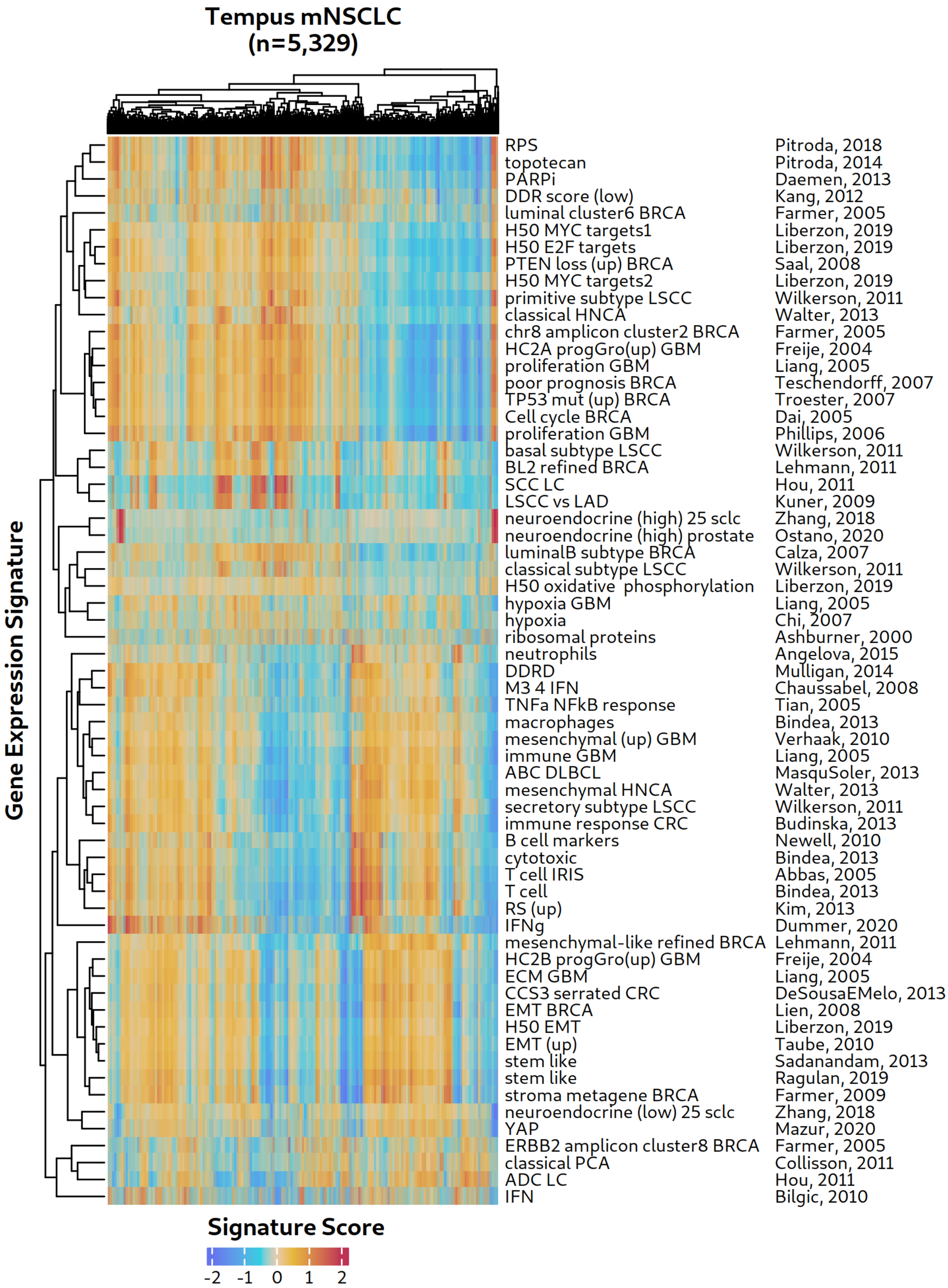

Figure 4.1: RosettaSX analysis of patients with NSCLC in the Tempus database. Rows represent coherent gene expression signatures and columns for patients with NSCLC (CS > 0.2). Red indicates high signature scores, and blue indicates low signature scores. The labels on the right side describe the mechanism, defined by the author's signature and year of publication. The heatmap indicated a population of NSCLC patients with increased gene expression signature scores that are descriptive of NE differentiation.

intermediate prediction probability for a NE phenotype (25th and 75th quantile prediction probability: [18%; 72%]). This indicated that compared to other patients, they only showed reduced signs of NE differentiation, so I removed them from subsequent analyses. In the following sections, I will differentiate multiple subgroups of the mLCNEC cohort. Patients are labeled mLCNEC$_{\mathrm{p}}$ if they were classified as mLCNEC and belonged to the pLCNEC cohort, and patients that only belong to the mLCNEC cohort are labeled mLCNEC$_{\mathrm{notp}}$.

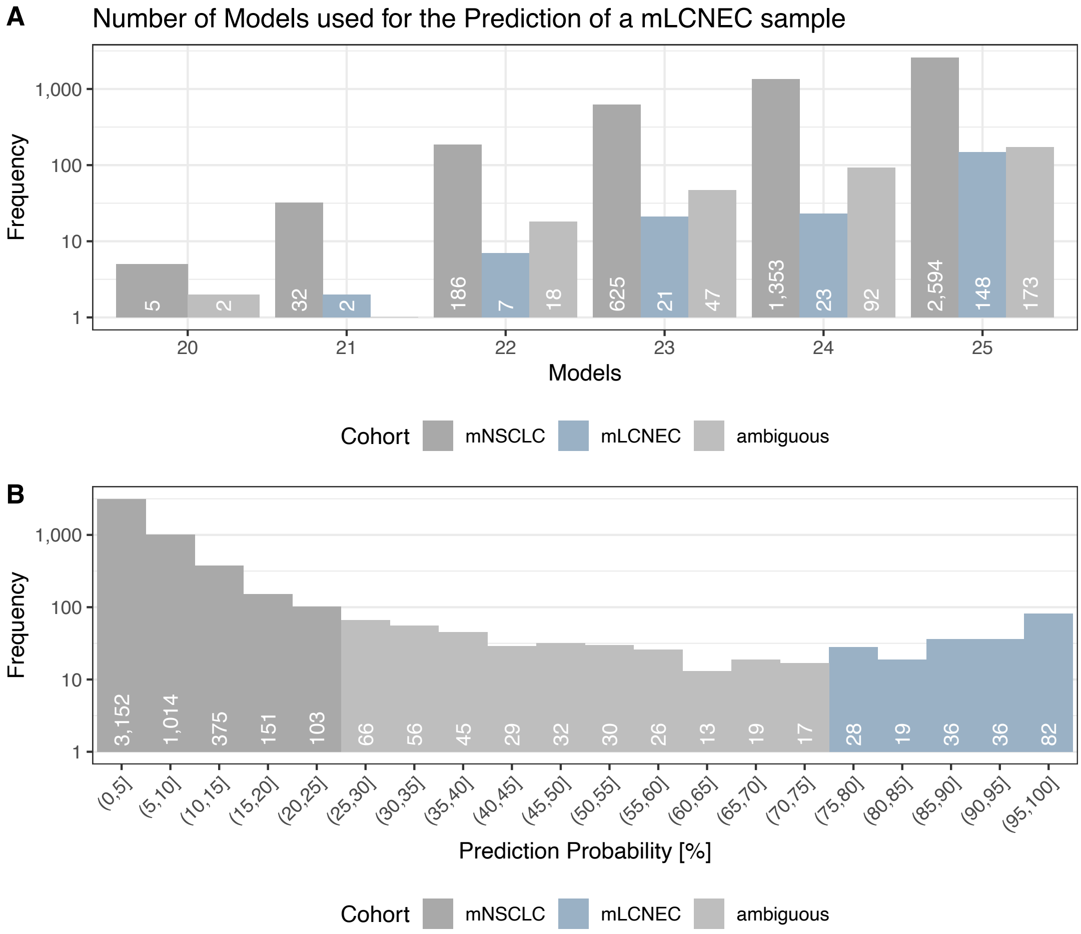

Figure 4.2: Overview of models that predicted the probability for LCNEC-likeness and distribution of final prediction probability in relation to their final classification. **A:** For each patient, multiple models predict the probability of NE differentiation, but only if the patient was not part of its training dataset for the respective model. Each bar represents the number of patients that received a prediction probability by the number of models indicated at the bottom. **B:** Distribution of prediction probability across all patients with NSCLC. The final prediction of a sample is based on the averaged prediction probabilities across all models that provided a prediction for the respective sample. Lastly, the classification of a sample is based on the aggregated probability: ambiguous (25-75%), mNSCLC (0-25%), or mLCNEC (75-100%). Blue: patients with mLCNEC, dark-grey: patients with mNSCLC, ambiguous: patients with unknown LCNEC-likeness

Model Illustrates the Underdiagnosis of Neuroendocrine Carcinomas

To evaluate the model results, I evaluated three NE markers used for LCNEC classification (as I lacked protein expression data, I used mRNA gene expression data of *SYP*, *CHGA*, and *NCAM*). I also evaluated an NE gene expression signature describing NE differentiation in SCLC (W. Zhang et al. 2018). Compared to the mNSCLC cohort, the expression of these markers was significantly higher than in the mLCNEC (and mLCNEC$_{\mathrm{p}}$ and mLCNEC$_{\mathrm{notp}}$) cohorts (Figure 4.3, right side). Lower NE markers were significantly more frequent in the mLCNEC$_{\mathrm{notp}}$ compared to the mLCNEC$_{\mathrm{p}}$ cohort. In a second step, I compared the NE levels among the pathological annotated morphologies between the patients classified as mLCNEC or mNSCLC (Figure 4.3 C). Across all indications, there was a significantly higher expression of at least two neuroendocrine markers in the mLCNEC$_{\mathrm{notp}}$ cohort than in other NSCLCs. The difference was more significant in patients originally annotated as ADC than those with SqCC (Figure 4.3 C right).

Overall, these results show that my model identified patients with solid signs of neuroendocrine differentiation.

Model Selected Genes with High Specificity

Next, I evaluated which genes were most frequently selected by my implemented bagging approach (i.e., genes with the highest classification importance). All 25 models selected *BSN*, *KIF1A*, *RUNDC3A*, *SCG3*, and *SYP*, and 24 out of 25 models selected *INSM1*, *KIF5C*, and *MAST1* as the most essential genes for mLCNEC classification. Figure 4.4 highlights the gene expression values of these genes in mNSCLC, mLCNEC, and pLCNEC patients, with a high alignment of the scores in the mLCNEC cohort. These genes are a subset of the NE differentiation SCLC gene expression signature (W. Zhang et al. 2018). However, they have not been mentioned in the context of LCNEC. Many of these genes are well-known neuronal markers (Tsai et al. 2017; Swarts, Ramaekers, and Speel 2015) or have been associated with other NE tumors (Lázaro et al. 2019; W. Zhang et al. 2018).

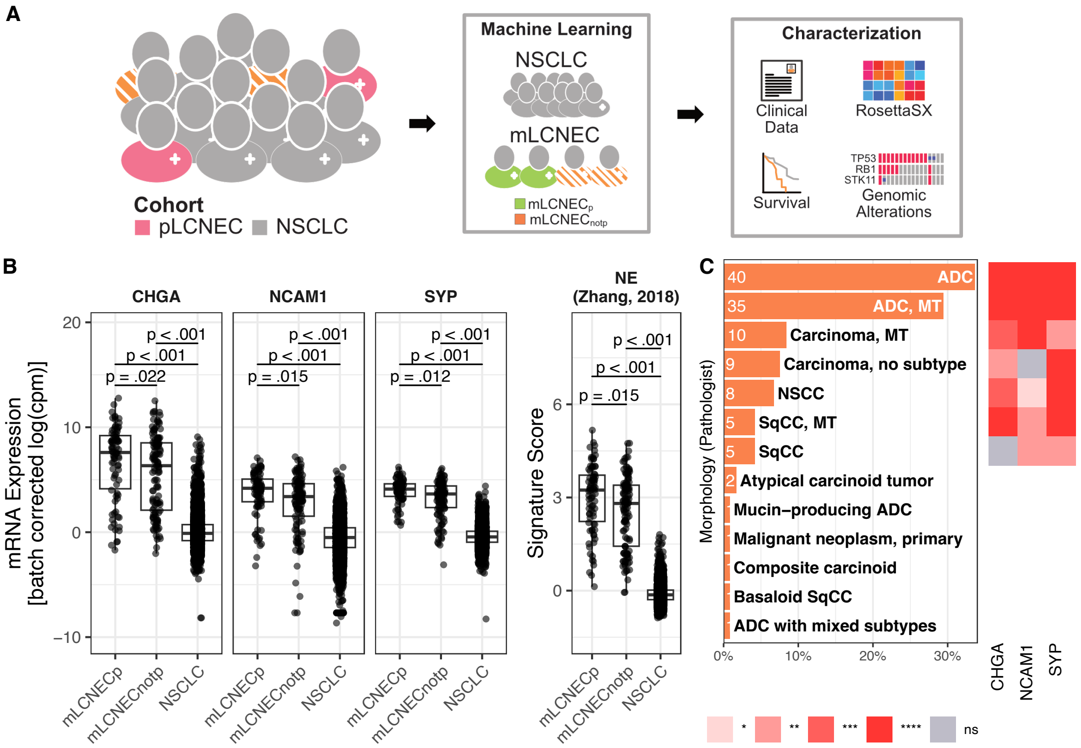

Figure 4.3: Graphical abstract of this chapter and evaluation of the model. **A:** In this section, I analyze a population of patients with NSCLC to identify patients that share molecular characteristics (mLCNEC, orange) with pathologically classified LCNECs (pLCNEC, pink). For this, I will implement a machine-learning model that allows me to identify patients that were either classified as pLCNEC and mLCNEC (mLCNEC$_p$, green) or only molecular LCNEC (mLCNEC$_{notp}$, orange). Finally, I will use this enriched mLCNEC cohort to characterize genomic, transcriptomic and clinical properties. **B:** Gene expression of NE markers (*CHGA*, *NCAM1* and *SYP*) currently used in the clinics and gene expression signature scores of a NE signature (W. Zhang et al. 2018) between patients with NSCLC, mLCNEC$_p$ and mLCNEC$_{notp}$. **C:** Pathological annotation of the identified patients with mLCNEC (left) and the difference of NE marker expression of the identical morphologies in the remaining NSCLC patients.

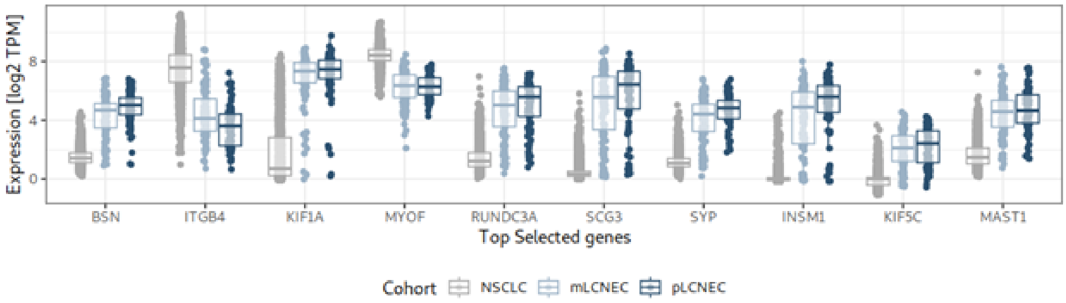

Figure 4.4: Expression of gene expression markers most frequently selected by the bagging approach. For each gene, the expression in patients with mNSCLC (grey), mLCNEC (light blue), or pLCNEC (dark blue) is shown. Except for *ITGB4* and *MYOF*, the median expression is higher in patients with mLCNE or pLCNEC for all genes.

Core Molecular LCNEC Genes Are Coherent Across Studies

The above analyses indicated a significantly higher expression of NE markers in the identified sample populations. However, a validation of the genes in independent cohorts will further strengthen the capability of the genes to describe NE differentiation. Therefore, I next used the most critical, positively associated genes as an 8-gene expression signature and evaluated their coherent expression in independent cohorts (Figure 4.5). One cohort comprised 66 LCNEC and another 81 SCLC samples (George et al. 2018, 2015). I compared the CS of my 8-gene signature with 10,000 random signatures of equal size sampled from the 151 NE markers I used for the model training. Additionally, for comparison, I applied the same procedure to the NE SCLC (25-gene) Zhang et al. signature (W. Zhang et al. 2018). The empirical *p*-values (i.e., a fraction of random signatures of equal size with the same or more extreme p-value) were determined for my signature and the Zhang et al. signature and was significantly higher than random signatures in both LCNEC and SCLC cohorts (*p*-value $<$ .0001, and *p*-value $<$ .0001).

One of the studies also comprised the IHC status for synaptophysin (SYP), chromogranin (*CHGA*), and neural cell adhesion molecule 1 (*NSCM1*) for 66 patients (George et al. 2018). The comparison of gene expression signature scores of my 8-gene signature showed a significantly higher expression in IHC-positive patients for SYP (one-sided Wilcoxon rank sum test, *p*-value $<$ .001) and *CHGA* (one-sided Wilcoxon rank sum test, *p*-value $<$ .001) but not for *NCAM1* (one-sided Wilcoxon rank sum test, *p*-value = .56)

These results indicated that my signature (i.e., the top eight mLCNEC positively associated markers selected by my model) represents a coordinately expressed gene module across different cohorts. Additionally, the signature aligned well with the protein expression status of two out of three NE markers currently used in the clinical classification of LCNEC, indicating that the top features describe an NE phenotype well.

Molecular LCNECs Resemble Clinical LCNEC Characteristics

I next evaluated the clinical characteristics of my identified mLCNEC cohort to determine similarities with the characteristics of patients with mNSCLC or those with pLCNEC (Table 4.1). Overall, I classified 201 (3.77%) patients of the

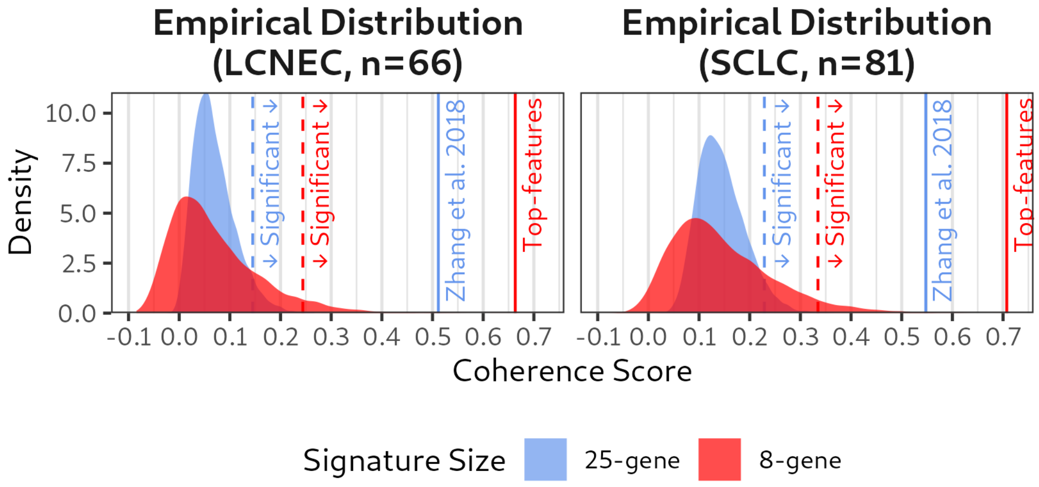

Figure 4.5: Empirical p-value distribution of 10,000 randomly sampled 8-gene (red) and 25-gene (blue) expression signatures compared to the 25-gene Zhang et al. NE and my 8-gene expression signature. Solid vertical lines indicate the CS of the annotated gene expression signature, and dashed lines indicate an empirical p-value smaller than .05.

Tempus NSCLC cohort (that had sufficient RNA-Seq quality) as mLCNEC. 77 of these patients were initially diagnosed with ADC, 27 with carcinoma, 11 with SqCC, 3 carcinoid, and 1 with malignant neoplasm by a pathologist (Figure 4.3). Patients with mLCNEC had a median age of 64 and were more frequently female with a history of smoking. While patients with mNSCLC often were never smokers (NSCLC: 15%, mLCNEC: 9%), patients with mLCNEC were frequently current smokers or past smokers at the time of biopsy (NSCLC: 85%, mLCNEC: 91%). Additionally, patients with mLCNEC are usually diagnosed with stage IV cancers (NSCLC: 67%, mLCNEC: 81%), while stage I-III tumors were more frequent in the mNSCLC cohort. I did not observe substantial differences in ethnicity or the number of lines of therapy (LoT, Fisher's exact test, $p = .21$) that a patient received between the cohorts.

Table 4.1: Clinical characteristics of the NSCLC, pLCNEC (all pathologically classified LCNEC), mNSCLC (QC passed NSCLC only), mLCNEC (including mLCNEC$_{p}$ and mLCNEC$_{notp}$), mLCNEC$_{p}$ and mLCNEC$_{notp}$.

Characteristic	**NSCLC**, N = 5,329 (100%)	**pLCNEC**, N = 106 (2.0%)	**mNSCLC**, N = 4,795 (90%)	**mLCNEC**, N = 201 (3.8%)	**mLCNEC$_{\mathbf{p}}$**, N = 82 (1.5%)	**mLCNEC$_{\mathbf{notp}}$**, N = 119 (2.2%)
Gender, n (%)						
Female	2,674 (50%)	53 (50%)	2,400 (50%)	107 (53%)	39 (48%)	68 (57%)
Male	2,655 (50%)	53 (50%)	2,395 (50%)	94 (47%)	43 (52%)	51 (43%)
Age, Median (IQR)	67 (61, 73)	66 (58, 71)	67 (61, 74)	65 (59, 71)	65 (57, 71)	64 (61, 71)
Missing	19	0	17	1	0	1
Smoking Status, n (%)						
Current smoker	412 (18%)	10 (23%)	364 (18%)	21 (23%)	7 (21%)	14 (24%)
Never smoker	321 (14%)	5 (12%)	305 (15%)	9 (9.7%)	5 (15%)	4 (6.8%)
Past smoker	1,571 (68%)	28 (65%)	1,392 (68%)	63 (68%)	22 (65%)	41 (69%)
Missing	3,025	63	2,734	108	48	60
Ethnicity, n (%)						
American Indian or Alaska Native	10 (0.3%)	0 (0%)	7 (0.2%)	0 (0%)	0 (0%)	0 (0%)
Asian	185 (4.7%)	0 (0%)	175 (4.9%)	5 (3.6%)	0 (0%)	5 (6.0%)
Black or African American	468 (12%)	8 (11%)	414 (12%)	17 (12%)	7 (13%)	10 (12%)
Native Hawaiian or Other Pacific Islander	3 (<0.1%)	0 (0%)	3 (<0.1%)	0 (0%)	0 (0%)	0 (0%)
White	3,275 (83%)	68 (89%)	2,959 (83%)	117 (84%)	49 (88%)	68 (82%)
Missing	1,388	30	1,237	62	26	36
Stage, n (%)						
I-III	1,584 (33%)	19 (20%)	1,479 (35%)	35 (19%)	13 (18%)	22 (21%)

IV	3,156 (67%)	75 (80%)	2,781 (65%)	145 (81%)	61 (82%)	84 (79%)
Missing	589	12	535	21	8	13
Line of Therapies, n (%)						
1	2,730 (51%)	56 (53%)	2,453 (51%)	100 (50%)	45 (55%)	55 (46%)
2+	1,662 (31%)	33 (31%)	1,480 (31%)	75 (37%)	26 (32%)	49 (41%)
none	937 (18%)	17 (16%)	862 (18%)	26 (13%)	11 (13%)	15 (13%)

Genomic Aberrations of Molecular LCNECs Align with LCNEC Aberrations

Genomic information was available for all 201 patients of the identified mLCNEC cohort. Figure 4.6 displays the top mutated genes (SNV, amplification, or deletion). 30.85% and 59.20% of the patients with mLCNEC belonged to the SCLC-like (*TP53* and *RB1* mutated) and NSCLC-like (*TP53*, *CDKN2B*, *CDKN2A*, *MTAP*, *STK11*, *KRAS*, *SMARCA4*, *FGF3*, *FGF4* or *CCND1*) genomic LCNEC subtypes (Burns et al. 2024; W. Zhang et al. 2018). My identified mLCNEC$_{\mathrm{notp}}$ cohort primarily belonged to the NSCLC-like subtype (Fisher's exact test, $p < 0.001$). Additionally, in the NSCLC-like subtype, I observed a significant enrichment of never-smokers with ch9p21 amplifications (one-sided Wilcoxon rank Sum test, $p < 0.001$). Most of the highly mutated genes have previously been either identified as enriched in SCLC-like (*PTEN*) or NSCLC-like (*KEAP1*) or as similarly prevalent in both subtypes (*NF1*, *NOTCH1*) (Burns et al. 2024). Previous LCNEC studies have not identified *FOXA1* amplification, which I found in 9% of the patients with mLCNEC. It frequently co-occurred with previously mentioned amplifications of *NKX2-1* (comparing increased alterations in either gene between mLCNEC and mNSCLC, Fisher's exact test, $p = .01$) (Burns et al. 2024; George et al. 2018).

Besides these known subtype alterations, I also observed gene alterations related to proliferation (*APC*, *NF1*), RAS/RAF/MAPK pathway (*KRAS*), PI3K/AKT/mTOR pathway (*PI3CA*), DNA Damage Response (DDR) (*FAT1*, *SMARCA4*, *KMT2C*, *KMT2D*), Notch signaling (*NOTCH1*, *NOTCH3*). Additionally, I found alterations in *MYC* (6%) and *MYCL* (8%) that have been mentioned in the context of SCLC, LCNEC, and NSCLC (Burns et al. 2024; Mollaoglu et al. 2017; Eftekhari Kenzerki et al. 2023).

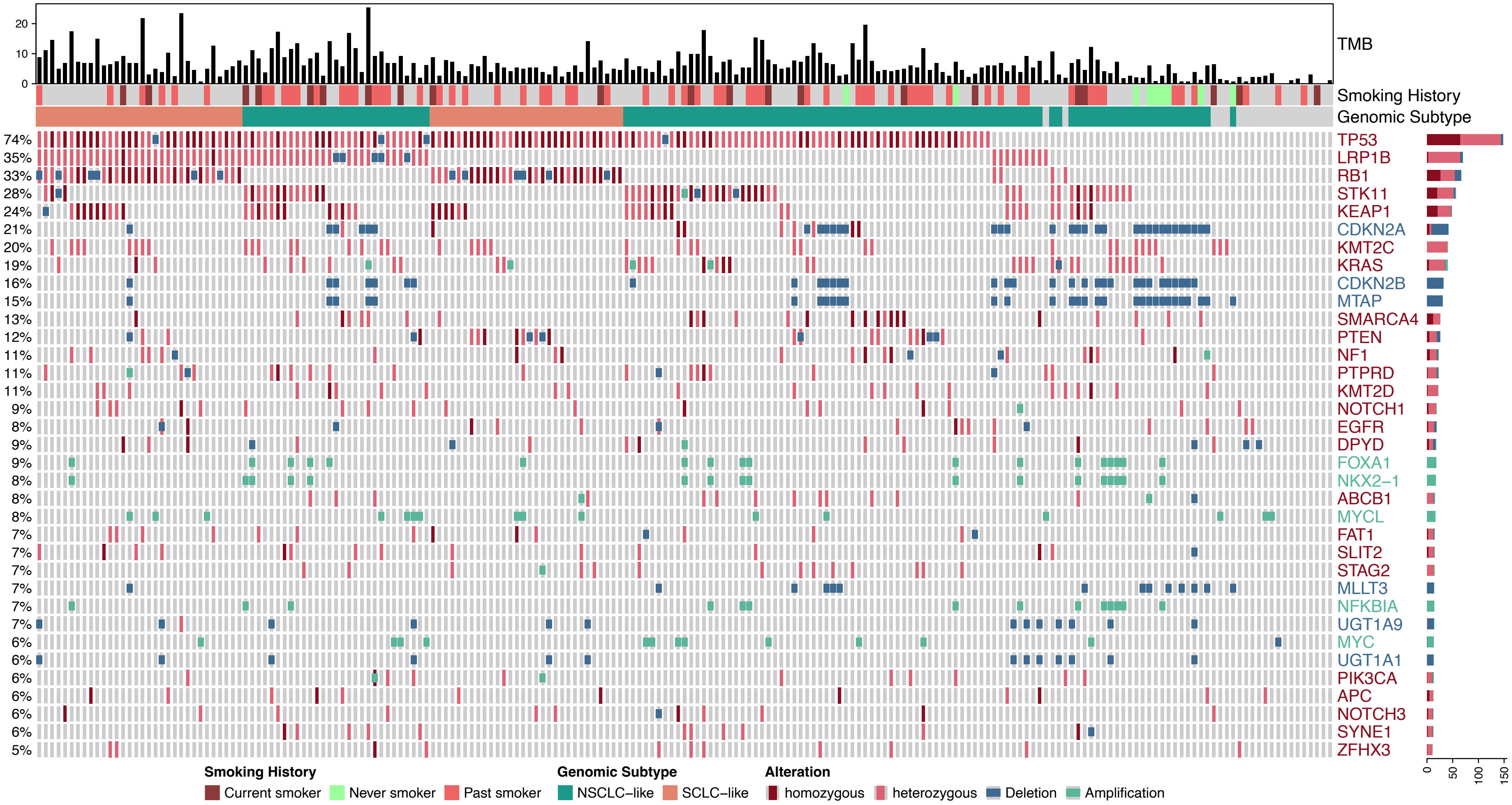

Figure 4.6: Overview of most frequent alterations in the patients with mLCNEC. At the top of the heatmap, the tumor mutational burden (TMB), smoking history, and genomic subtype (i.e., NSCLC-like or SCLC-like) are annotated. Genes are listed if gene copy number deletions, amplifications, or SNVs exist in more than 4% of the cohort. The color of the gene names indicated if the gene was selected due to a high frequency of amplifications (green), deletions (blue), or a high impact loss (either SNV, deletion, or both). The bar plot indicates the overall alteration frequencies.

RosettaSX Reveals Main Transcription Programs in mLCNEC

The most extensive transcriptional analysis of LCNEC samples was provided by George et al. and only comprised 66 patients (George et al. 2018). I analyzed the 201 mLCNEC to identify transcription programs that indicate transcriptional subtypes within the mLCNEC cohort using RosettaSX. For the analysis of 306 signatures, I first narrowed the signatures down to 85 signatures that had a CS below 0.2. Finally, I filtered the signatures using the Jaccard index (maximally 0.25), selecting a maximum of 3 signatures for an individual process (e.g., proliferation, EMT). The remaining 63 signatures are shown in Figure 4.7. I used unsupervised clustering for the patients (columns) and signatures (rows). Additionally, I supplemented the heatmap with markers for SCLC-like LCNEC stratification (*TP53*, *RB1* alterations), the genomic LCNEC subtypes (NSCLC-like and SCLC-like), and mLCNEC ($\mathrm{mLCNEC_p}$ or $\mathrm{mLCNEC_{notp}}$) subtypes.

Patients clustered into two mutually inclusive clusters, with predominantly high proliferation scores in the first cluster (Dai et al. 2005; Phillips et al. 2006) and high immune and stroma signatures in the second cluster (Farmer et al. 2009; Bindea et al. 2013). *TP53*, *RB1* co-alterations were significantly enriched in the proliferation cluster (Wilcoxon rank sum test, $p < .001$), and homozygous mutations in *TP53* were significantly associated with lower scores of the Farmer et al. stroma gene expression signature (Wilcoxon rank sum test, $p < .001$). A group of patients had increased EMT, mesenchymal, or stemness gene expression signatures but low proliferation scores. As indicated in the previous section, such a pattern possibly indicates a high contribution of stromal cells, not cancer cells, that results in this expression pattern (Kreis et al. 2024). Biopsies in this cluster were significantly associated with lower cancer cell content (one-sided Wilcoxon rank sum test, $p < .001$).

Accordingly, the gene expression signature scores clustered into two larger clusters (at the bottom of the heatmap), describing proliferation and immune signatures, but also a third cluster describing a variety of mechanisms, like neuroendocrine differentiation (W. Zhang et al. 2018; Ostano et al. 2020), *PTEN* loss (Saal et al. 2008), SqCC, or ADC (Hou et al. 2010). Interestingly, high scores of signatures characteristic for ADC or SqCC were significantly associated with the original pathological annotations for ADC or SqCC (one-sided Wilcoxon Rank Sum test, other mLCNEC vs. ADC: $p < .001$, other mLCNEC vs. SqCC: $p < .001$).

Finally, when comparing the gene expression signatures between the patients with $\mathrm{mLCNEC_p}$ or $\mathrm{mLCNEC_{notp}}$, there were no significant differences, except for a NE signature, which was significantly higher in the $\mathrm{mLCNEC_p}$, compared to the $\mathrm{mLCNEC_{notp}}$ cohort (one-sided Wilcoxon rank sum test, $p < .001$).

Clinical and Molecular LCNEC Receive Different Treatments

As described in the project outline, patients with LCNEC lack specific recommendations for therapies and instead are often treated with NSCLC-like (e.g., pemetrexed or gefitinib) or SCLC-like (e.g., etoposide and platinum chemotherapy) regimens (Dingemans et al. 2021; Hendriks et al. 2023). Therefore, I next evaluated which therapies were administered to patients with mLCNEC as the first line of therapies (LoT). 175 (87%) patients with mLCNEC received their first LoT between 2013 and 2023 and overall received a median of 1.6 LoT (ranging from zero to ten LoTs). The patients received combinations of CTx, Immune checkpoint inhibitors (ICI), ICI in combination with CTx, Tyrosine kinase inhibitors (TKI), and other targeted therapies (Table 4.2). The most frequently administered drug types administered to patients with mLCNEC were combinations of CTx (79) or CTx + ICI (75). Although such combinations were also most frequently administered to mNSCLC patients, the ratio of administrated therapies differed significantly between mNSCLC and mLCNEC (Chi-square test, $p < .001$). However, when comparing the ratio of administered therapies between patients with $\mathrm{mLCNEC_p}$ and $\mathrm{mLCNEC_{notp}}$, there was no significant difference (CTx or CTx+ICI, Fisher's exact test, $p = .255$).

When comparing the frequencies of administered drug types between $\mathrm{mLCNEC_p}$ and $\mathrm{mLCNEC_{notp}}$ patients, the latter received therapies comparable with mNSCLC (carboplatin, pembrolizumab, and pemetrexed (22) and carboplatin, durvalumab, and paclitaxel (15)). In contrast, patients with $\mathrm{mLCNEC_p}$ received therapies that were comparable with the pLCNEC cohort (cisplatin and etoposide (18), atezolizumab, carboplatin, and etoposide (15), and carboplatin and etoposide (15)). It is worth highlighting that the most frequently administered drugs for both $\mathrm{mLCNEC_{notp}}$ and $\mathrm{mLCNEC_p}$ are the standard therapies for either NSCLC (carboplatin, pembrolizumab, and pemetrexed, (Velcheti et al. 2021; Hendriks et al. 2023)), or SCLC (cisplatin, etoposide with and without atezolizumab, (Dingemans et al. 2021)).

Interestingly, only patients with $\mathrm{mLCNEC_{notp}}$ also received targeted therapies,

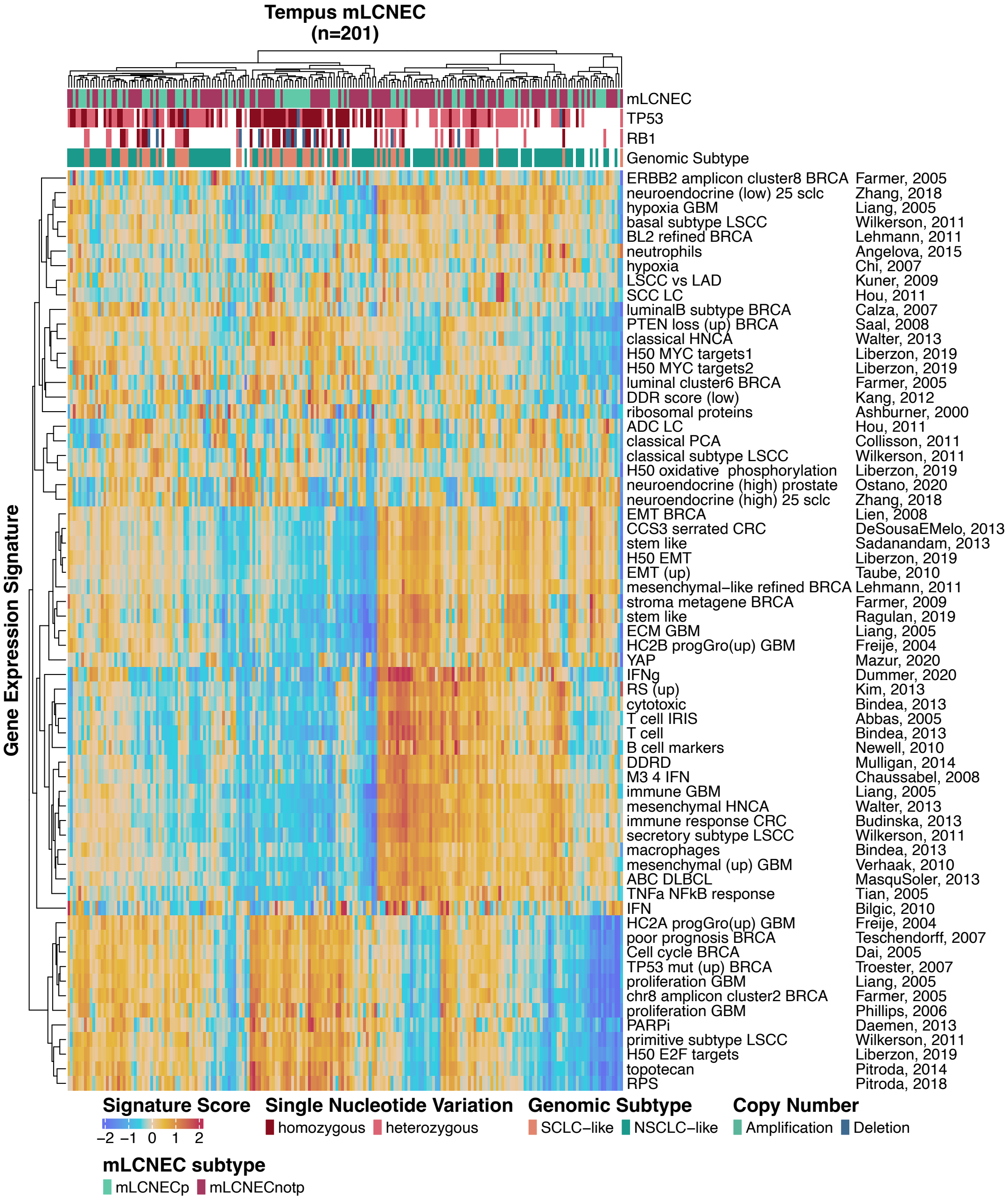

Figure 4.7: RosettaSX analysis of the mLCNEC cohort. The rows of the heatmap list gene expression signatures that passed the filtering criteria, and the columns are the patients with mLCNEC (red high signature scores, blue low signature scores). At the top of the heatmap, the Genomic subtype, the alterations of *RB1* and *TP53*, and the mLCNEC subtype (mLCNEC$_{\mathrm{p}}$ or mLCNEC$_{\mathrm{notp}}$) are visualized. The low labels on the right side briefly describe the signature, the author, and the year of publication from which the signature was derived.

the standard first-line therapy for ADCs with an EGFR mutation, EGFR tyrosine kinase inhibitors (TKIs). Studies have shown that patients indeed initially benefit from this therapy but that these patients inevitably develop resistance, for example, by histologic transformation of ADCs to LCNECs or SCLCs (Baglivo et al. 2017; Lim et al. 2014; M. Lee et al. 2022). Six patients with $\text{mLCNEC}_{\text{notp}}$ had a biopsy after the treatment with osimertinib (a TKI inhibitor). Consequently, the therapy selection might have resulted in a transformation of ADCs to LCNECs.

These results show that patients with $\text{mLCNEC}_{\text{notp}}$ (the patients my model identified) are treated differently than mLCNEC_{p} patients, in the worst case, to the patient's disadvantage.

Table 4.2: Listing of therapies that were administered to at least four patients with mLCNEC in comparison to patients with mNSCLC, mLCNEC_p and $\text{mLCNEC}_\text{notp}$. Medications are grouped into drug types (CTx, CTx+ICI or TKI) and percentages are relative to all administered therapies (including the medications not shown here).

1st LoT[1]	**mNSCLC**,N=3,893	**mLCNEC**,N=173	$\textbf{mLCNEC}_\textbf{notp}$,N=104	$\textbf{mLCNEC}_\textbf{p}$,N=69	**pLCNEC**,N=74
CTx, ICI, n (%)					
Carboplatin, Pembrolizumab, Pemetrexed	705 (18.1)	27 (15.6)	22 (21.2)	5 (7.2)	6 (8.1)
Carboplatin, Durvalumab, Paclitaxel	378 (9.7)	17 (9.8)	15 (14.4)	2 (2.9)	2 (2.7)
Carboplatin, Paclitaxel, Pembrolizumab	184 (4.7)	5 (2.9)	3 (2.9)	2 (2.9)	2 (2.7)
Atezolizumab, Carboplatin, Etoposide	1 (0.0)	19 (11.0)	4 (3.8)	15 (21.7)	16 (21.6)
CTx, n (%)					
Carboplatin, Paclitaxel	465 (11.9)	15 (8.7)	13 (12.5)	2 (2.9)	3 (4.1)
Carboplatin, Pemetrexed	347 (8.9)	8 (4.6)	8 (7.7)	- (-)	- (-)
Cisplatin, Pemetrexed	260 (6.7)	7 (4.0)	7 (6.7)	- (-)	- (-)
Cisplatin, Etoposide	77 (2.0)	22 (12.7)	4 (3.8)	18 (26.1)	20 (27.0)
Carboplatin, Etoposide	17 (0.4)	20 (11.6)	5 (4.8)	15 (21.7)	15 (20.3)
Tyrosine Kinase Inhibitors (TKI), n (%)					
Osimertinib	282 (7.2)	6 (3.5)	6 (5.8)	- (-)	- (-)

[1] Only listing therapies with at least four administrations.

Molecular LCNEC is Associated with Poor Prognosis

Patients with LCNEC were reported to have a comparably shorter OS of 9 (CI 8.2, 9.8) months compared to 11 (CI 10.9, 11.1) months for those with NSCLC (Kinslow et al. 2020). Thus, I next evaluated the prognosis of my identified mLCNEC cohort in relation to the mNSCLC cohort (Figure 4.8). Due to the low number of patients with stage I-III mLCNEC, I either analyzed patients with stage IV tumors only or patients with stage I-IV tumors. I accessed clinical records with a median follow-up time of 17.5 (mNSCLC) and 15.8 (mLCNEC) months. Using a univariate Cox PH model to evaluate differences between the OS of patients with mLCNEC and those with mNSCLC (stage I-IV), I found that patients with mLCNEC had a significantly shorter OS (hazard ratio [HR] = 1.65 (95% CI 1.34, 2.03; $p < .001$)). Similarly, patients with mNSCLC had a significantly longer PFS (HR = 0.64 (95% CI 0.54, 0.77; $p < .001$)). To reduce the effect of the enrichment of stage I-III patients in the mLCNEC$_{\text{notp}}$ cohort, I only evaluated the survival of stage IV tumors. The results of these analyses aligned with the previous results, highlighting a shorter survival for mLCNEC patients compared to those with mNSCLC (OS, HR = 1.47 (95% CI 1.17, 1.84; $p < .001$), PFS, HR = 1.62 (95% CI 1.28, 2.06; $p < .001$)).

To evaluate the influence of confounding factors, I next used multivariate Cox PH models to account for confounding factors. A stage-stratified Cox PH model, correcting for gender and age at primary diagnosis, showed a significantly shorter PFS with an HR of 1.80 (95% CI 1.46, 2.22; $p < .001$). Similarly, a gender- and age-stratified CoxPH model, correcting for stage, also indicated a shorter OS with an HR of 1.47 (95% CI 1.17, 1.83; $p < .001$)) (Figure 4.8 A top panel). This indicated a significantly worse prognosis for patients with mLCNEC compared to those with mNSCLC, even when I accounted for commonly described confounding factors like gender (Q. Yang et al. 2019).

Finally, I evaluated prognostic differences between mNSCLC, mLCNEC$_{\text{p}}$, and mLCNEC$_{\text{notp}}$. A multivariate Cox PH model, stratified by gender and tumor stage, adjusted for age, indicated reduced OS for both mLCNEC$_{\text{p}}$ and mLCNEC$_{\text{notp}}$ with an HR of 1.95 (95% CI 1.42, 2.68; $p < .001$) and 1.32 (95% CI 1.01, 1.74; $p = .045$), respectively.

Overall, these results highlight that patients with mLCNEC have a poor prognosis compared to those with mNSCLC. My analyses showed that this is true not only for the mLCNEC$_{\text{p}}$ cohort but also for the mLCNEC$_{\text{notp}}$ cohort.

First Line of Therapy Influences Prognosis of Molecular LCNEC

I next compared PFS and OS between the major drug types (CTx, CTx+PD1, and CTx+PDL1) within the mLCNEC cohort. Due to the small number of patients with stage I-III mLCNEC that received a first line of therapy, I restricted my analysis to stage IV cancers. The median OS was not reached (CTx, 95% CI: [11.14, -]), 7.6 (CTx + PD1, 95% CI: [6.02, -]), and 11.2 (CTx + PDL1, 95% CI: [6.61, -]), respectively and for PFS, none of the therapy types reached a median survival. Finally, I used a Cox PH model, stratified by gender, to account for confounding effects and evaluate the influence of the first LoT drug type on OS. Compared to a combination of CTx, a therapy with CTx and ICI + anti-PDL1 had a non-significant increased HR of 1.86 (95% CI 0.75, 4.62; p = .2). Similarly, a combination of CTx and ICI + anti-PD1 therapy has a significantly increased HR of 2.58 (95% CI 1.12, 5.90; p = .025). These results highlight the poor prognosis of patients with mLCNEC$_\mathrm{p}$ and those with mLCNEC. Additionally, it highlights that the most advantageous current therapeutic option for mLCNEC patients is a combination of CTx.

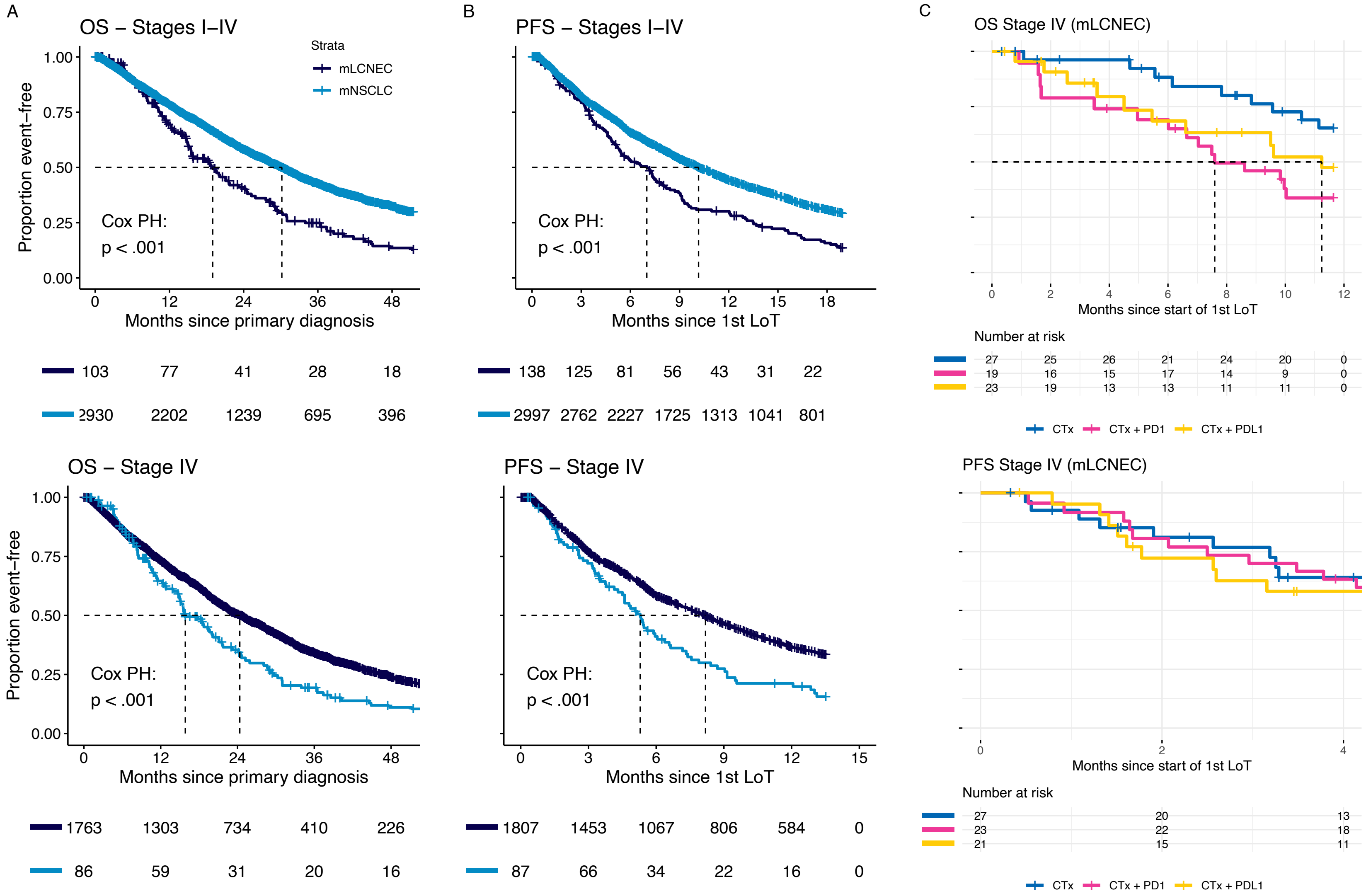

Figure 4.8: OS and PFS Kaplan-Meier curves are separated by the patient's disease (mLCNEC or mNSCLC) or the type of therapy administered to patients with mLCNEC. **A, B:** OS and PFS in patients with mLCNEC (dark blue) or those with mNSCLC (light blue). The top plots show the Kaplan-Meier curves of all patients, and the bottom plots show Kaplan-Meier curves for patients with stage IV mLCNEC.

Investigation of the Association between Molecular Alterations and Outcome

I identified multiple molecular characteristics that differentiate molecular subtypes in my mLCNEC cohort in the previous analyses. As a final analysis, I set these alterations in relation to PFS and OS. I selected genomic aberrations and transcription programs with enough patients and analyzed their association with OS using univariate Cox PH models. From the transcriptomic programs (either immune infiltration or proliferation), only the proliferation gene expression signature (Dai et al. 2005) was significantly associated with poor prognosis (HR = 1.75, 95% CI 1.12, 2.73; p = .014) and immune infiltration (Budinska et al. 2013) indicated a non-significant trend towards increased survival (HR = 0.78, 95% CI 0.57, 1.07; p = .121). Except for deletions of *TP53* (HR = 2.07, 95% CI 1.22, 3.50; p = .007) or deletions in either gene of the top 5 mutated genes (*TP53*, *RB1*, *LRBP1B*, *KEAP1*, or *STK11*, HR = 2.45, 95% CI 1.23, 4.88; p = .011), none of the alterations were significantly associated with PFS. However, none of the markers was significantly associated with OS in a multivariate Cox PH model (stratified by gender and corrected for age and stage).

Consequently, besides a proliferation signature transcriptionally, no biomarker was associated with patient outcome.

4.4 Discussion

This project focused on applying RosettaSX to characterize a population of NSCLC with signs of NE differentiation. Using my RosettaSX approach, I first identified the prevalent NE differentiation status of the samples and then applied RosettaSX for the transcriptional characterization of the cohort. This study reinforced previously reported misclassification of patients with LCNEC in an NSCLC real-world evidence dataset (Lindsay et al. 2021; Kinslow et al. 2020; Zhuo et al. 2020). More stringent testing for the NE differentiation status of patients with NSCLC might further increase the prevalence of LCNEC.

The WHO recommendation for LCNEC classification obligates an NE morphology and positive IHC staining for at least one NE marker (*CHGA*, *SYP*, *NCAM1*, Andrini et al. 2022). As I lacked NE marker IHC readouts, I analyzed mRNA expression data. All single-gene NE markers were significantly more highly expressed in patients with mLCNEC than mNSCLC (Figure 4.3). However, previous studies have shown these markers may have low LCNEC specificity (Andrini et al. 2022). Therefore, I evaluated a well-established SCLC NE gene expression signature, which provides a more robust NE differentiation status (W. Zhang et al. 2018), which was also significantly higher in patients with mLCNEC (Figure 4.3). Although IHC data might not recapitulate NE mRNA marker expression perfectly, my analysis indicates that the actual prevalence of LCNEC is almost 4%, not 1-3%, as previously stated (Rekhtman 2022). Lastly, in a validation cohort, using the high-importance markers of my model for two out of three NE markers, I recapitulated high scores in IHC-positive patients. Thus, in line with Zhuo et al., my analyses also revealed an underdiagnosis of LCNEC (Zhuo et al. 2020). A more comprehensive NE differentiation in current clinical practice might mitigate this problem (i.e., using a more robust gene expression panel).

Another factor that can influence the variable NE marker expression might be the introduction of LCNEC combined in the 2021 WHO recommendation (Nicholson et al. 2022). The pLCNEC cohort comprised 14 LCNECs with evidence of NE differentiation combined with other NSCLCs. Although I could not retrospectively differentiate LCNEC combined, my gene expression signature analysis indicated a subset of patients with increased NE and ADC or SqCC characteristic gene expression signatures. This suggests that my mLCNEC cohort also comprised a fair number of patients with LCNEC combined SqCC or ADC.

Besides the molecular characteristics, the herein identified mLCNEC cohort showed a good agreement with the previously described characteristics of LCNEC. The identified mLCNEC cohort is enriched for late-stage tumors with a comparably poor prognosis. The set of genomic mutations and aberrations that I found recapitulated findings from previous studies (Burns et al. 2024; George et al. 2018) with the two previously described NSCLC-like and SCLC-like subtypes. In addition, I observed 333 patients who had ambiguous prediction probabilities for their NE status. A subset of these ambiguous samples might be a sample with an NE differentiation status, which would further increase the frequency of LCNEC. My analysis indicated that patients treated with CTx + ICI-anti PD1 in their first LoT had significantly worse survival at an HR = 2.58 (95% CI 1.12, 5.90; p=0.025). Combined with the shorter PFS and OS (starting from primary diagnosis), these results depict insufficient therapies for patients with mLCNEC. My analysis describes a sufficiently sized patient population that, despite its poor prognosis, lacks targeted therapies.

This project highlighted the drastic underdiagnosis of LCNEC and its still lacking therapy options. I highlighted the beneficial effect of a more thorough NE marker evaluation or gene panel screening, which might improve the diagnosis of LCNEC. Although my analysis was based on one of the largest NSCLC cohorts for which transcriptomic data is available today, future studies are required to investigate the advantage of individual therapy regimens further.

Chapter 5

Discussion

The objectives of this study were the description and application of a workflow and the implementation of a platform for the evaluation of gene expression signatures. I first established the RosettaSX framework for in-depth signature analyses throughout the two subsequent chapters. I then highlighted the framework's capabilities to recapitulate breast cancer's molecular features. In the subsequent chapters, I demonstrated the platform's utility to unravel the actual phenotype of EMT-related gene expression signatures and improve the molecular diagnosis of the phenotype of lung cancer patients. This study describes a methodology easily transferable to various cancer datasets and enables in-depth evaluations of gene expression signatures.

5.1 The Unique Value of My RosettaSX Approach for Gene Expression Signature Analyses

In the second chapter, I outlined a method that can generically be used to assess phenotypic patterns in gene expression data of cancers using gene expression signatures, the RosettaSX collection. Through the utilization of gene expression signatures, my approach makes use of the prior knowledge of published gene expression signatures. In new experimental contexts, it provides access to functional links proposed for the RosettaSX signatures in their original studies. Capturing this knowledge is valuable since it has often been verified experimentally or through comprehensive analyses in the literature and therefore comes with a high likelihood that such signatures are of good quality.

Many studies use an alternative approach to analyze new cancer gene expression data sets. It is based on de-novo clustering of gene expression data followed by annotation of cluster-associated molecular or clinical phenotypes (e.g., George et al. 2018). Although this approach might sometimes lead to the identification of new patterns in a dataset under study, it hardly provides links to already existing knowledge about gene expression programs. It disregards already identified gene expression signatures that have been demonstrated to be associated with a specific phenotype (e.g., W. Zhang et al. 2018; Masqué-Soler et al. 2013; Perou et al. 2000). It frequently leads to the re-discovery of signatures for phenomena for which other (and often better) signatures have been described already.

My signature set in RosettaSX has been carefully selected for limited functional redundancy. This means there is a limited number of signatures that describe one specific phenomenon, like immune cell infiltration (Bindea et al. 2013), cancer cell type (W. Zhang et al. 2018; Perou et al. 2000), proliferation activity (Dai et al. 2005), or interferon signaling (Dummer et al. 2020). I try to cover all hallmarks that can be recurrently detected in gene expression studies of cancer cohorts by such small sets of signatures. The gene sets of signatures that stand for a particular phenomenon or hallmark hardly overlap. For a new data set that will be investigated with my RosettaSX analysis framework, this limited redundancy of signatures can provide useful information about gene expression programs. The rediscovery of relevance (by the CS) and co-clustering (by hierarchical clustering of signature profiles) of a set of functionally related signatures -that have been detected independently in different published studies- is a strong hint at the relevance of their function in the data set that is under investigation. No other signature analysis framework gives access to such comprehensive information in a similar way. Published workflows based on Gene Set Enrichment Analysis (Subramanian et al. 2005) variants also allow conclusions about large sets of signatures. Still, associations between signatures are not detected, the redundancy of signatures for specific functions is not controlled, and the translatability of signatures into a new data context is not assessed.

My approach builds on a gene expression signature collection that can differentiate several processes involved in the hallmarks of cancer (Liberzon et al. 2015) and, also beyond that, comprises signatures that describe important cancer-specific properties (e.g., cell of origin). Therefore, the signature collection is not only limited to a specific set of signatures (e.g., biological pathways, hallmark signatures) but can also be extended to involve signatures relevant to the context under investigation. This is different from other approaches, which rely on smaller sets of signatures that try to describe the activity of specific signaling pathways (Schubert et al. 2018). These methods are most often tailored to predicting the activity of particular pathways but require additional data that is usually unavailable for large-scale datasets or many hallmark phenomena. Instead, I rigorously apply the CS concept to be able to focus on signatures that show coordinated regulation of their genes as a measure of relevance in a new data context (Staub 2012). The CS concept needs minimal information; it does not even need annotation of sample groups in the data under investigation which is a requirement for canonical Gene Set Enrichment Analyses that are frequently performed on the extensive collection of MsigDB signatures (Subramanian et al. 2005). My RosettaSX approach fills a gap in the landscape of analytical procedures for gene expression signatures.

5.2 Context Matters – The Influence of Confounding Factors

While studies frequently apply gene expression signatures to characterize a biomarker, they often skip an essential step - the context under study. Without further evaluation, many studies use gene expression signature collections, such as the hallmark 50 dataset. Although these signatures were shown to describe a particular phenomenon in discovery data, assuming the set of genes explains the same phenomenon in another data context (i.e., another gene expression dataset) is not valid. Regardless of whether a signature is applicable in a new context, the signature profiles can be calculated and will indicate populations with low and high signature scores. However, it is questionable if these profiles describe the originally anticipated phenomenon if the genes are not coordinately up- and down-regulated across samples of the new data set. This became especially apparent in the second chapter of my work in which I evaluated gene expression signatures in different contexts. My analyses in section Section 3.3 highlighted that signatures describing phenomena in the TME are only coherent in the context of data that comprises TME.

However, Chapter 3 highlighted that, even though a gene expression signature might be highly relevant in another context, there are signs of contamination. While the EMT-related signatures indicated a highly reproducible expression footprint across various datasets, they did not describe the originally anticipated phenomena which frequently have been attributed to the mesenchymal status of cancer. At this point, my approach again highlighted its advantageous capabilities. The analysis of multiple gene expression signatures that cover a wide range of cancer-specific phenomena, associating the signatures with each other and evaluating the integrity of signatures. The observer can determine the biological context with which the signature is associated.

The current RosettaSX implementation is limited to analyzing cancer patients and cancer models. However, a future extension of the platform might be archived by the addition of further datasets, such as pseudobulk data from scRNA-seq studies, as analyzed in section Section 3.3. Through the analyses of pseudobulk data, tailored towards samples that describe cell type-specific expression phenotypes, I was able to evaluate gene expression signatures and their association with the cell composition of the TME.

5.3 The Importance of a Comprehensive Gene Expression Signature Collection

In the third study on the underdiagnosis of LCNEC within NSCLC, I highlighted the capability of my analytical framework, RosettaSX, to quickly identify biological meaningful cancer subpopulations in a large cancer cohort. The analysis revealed high signature scores for well-established NE signatures in a subset of patients. Subsequent analyses highlighted that these patients recapitulated molecular features of previously described LCNEC characteristics. Thus, while Chapter 2 highlighted the application of my analytical framework in the gene expression signature dimension, in this chapter I highlighted the utility of my analytical framework on the patient dimension. As indicated in the previous section, the analysis in section Section 4.3 showed that the NE expression profiles represent a unique cell of origin phenotype that cannot be detected with other gene expression signature collections (e.g., hallmark 50 Liberzon et al. 2015). Only the composition of the gene expression signature collection of RosettaSX, which combines hallmark signatures with specific cancer-specific signatures, allows a comprehensive analysis of cancer indications. Indeed, the evolution of a signature collection is a continuous process. While this thesis is finalized, new signatures for other phenomena included in RosettaSX manifest themselves in cancer expression data. Comprehensive analyses of large signature databases for the identification of additional signatures, as by Cantini et al., or the evaluation of signatures using single-cell data will be of primal importance for the future development of RosettaSX (Cantini et al. 2018).

For the field of LCNEC, my analysis indicates that there is a population of patients that share molecular similarities with pLCNEC but are most often classified as ADCs or SqCC. Overall, these patients had a poorer prognosis than the patient population pathologists classified them. The threshold for the prediction of a mLCNEC phenotype was chosen rather conservatively. Further analyses in this direction and a more thorough investigation of borderline negative cases might even result in a higher prevalence of this aggressive form of lung cancer and ultimately, in the exploration of better therapeutic options for these patients.

Chapter 6

Conclusion

In this thesis, I describe an integrated approach to gene expression signature analysis that can be applied in multiple ways to better characterize molecular phenomena of importance to tumors based on expression data. My analysis approach facilitates the discovery of associations with other biomarkers. Through analyzing of multiple data types and cancer indications, I demonstrated the power of the approach for the characterization of known and the discovery of novel gene expression signatures. As examples, I describe the comprehensive analysis of breast cancer gene expression data, the recovery of limited applicability and interpretation flaws of signatures for EMT and mesenchymality in the clinical context, and the discovery and in-depth characterization of a gene expression signature for molecularly defined LCNEC.

The consistent evaluation of the quality and applicability of a gene expression signature, which is often ignored, can, at worst result in misapplication of gene expression signatures over a long period and ultimately lead to false conclusions. In contrast, the rigorous application of quality principles in gene expression signature studies helps to discover new ways of molecular diagnoses, as shown here for LCNEC.

References

Ades, Felipe, Dimitrios Zardavas, Ivana Bozovic-Spasojevic, Lina Pugliano, Debora Fumagalli, Evandro De Azambuja, Giuseppe Viale, Christos Sotiriou, and Martine Piccart. 2014. "Luminal b Breast Cancer: Molecular Characterization, Clinical Management, and Future Perspectives." *Journal of Clinical Oncology* 32: 2794–2803. https://doi.org/10.1200/JCO.2013.54.1870.

Agresti, Alan. 2006. *An Introduction to Categorical Data Analysis*. *Wiley Series in Probability and Statistics*. Wiley. https://doi.org/10.1002/0470114754.

Allison, Kimberly H. 2021. "Prognostic and Predictive Parameters in Breast Pathology: A Pathologist's Primer." *Modern Pathology* 34 (January): 94–106. https://doi.org/10.1038/s41379-020-00704-7.

Amin, Mahul B, Stephen Edge, Frederick L Greene, Richard L Schilsky, David R Byrd, Lauri E Gaspar, Mary Kay Washington, Jeffrey Evan Gershenwald, Carolyn C Compton, and Kenneth R Hess. 2016. *AJCC Cancer Staging Manual*. 8th ed. Cham, Switzerland: Springer International Publishing.

Anderson, Nicole M, and M Celeste Simon. 2020. "The Tumor Microenvironment." *Current Biology*.

Andrini, Elisa, Paola Valeria Marchese, Dario De Biase, Cristina Mosconi, Giambattista Siepe, Francesco Panzuto, Andrea Ardizzoni, Davide Campana, and Giuseppe Lamberti. 2022. "Large Cell Neuroendocrine Carcinoma of the Lung: Current Understanding and Challenges." *Journal of Clinical Medicine* 11 (5): 1461. https://doi.org/10.3390/jcm11051461.

Angelova, Mihaela, Pornpimol Charoentong, Hubert Hackl, Maria L. Fischer, Rene Snajder, Anne M. Krogsdam, Maximilian J. Waldner, et al. 2015. "Characterization of the Immunophenotypes and Antigenomes of Colorectal Cancers Reveals Distinct Tumor Escape Mechanisms and Novel Targets for Immunotherapy." *Genome Biology* 16: 1–17. https://doi.org/10.1186/s13059-015-0620-6.

Aran, Dvir, Roman Camarda, Justin Odegaard, Hyojung Paik, Boris Oskotsky, Gregor Krings, Andrei Goga, Marina Sirota, and Atul J. Butte. 2017. "Comprehensive Analysis of Normal Adjacent to Tumor Transcriptomes." *Nature Communications* 8 (December): 1077. https://doi.org/10.1038/s41467-017-

01027-z.

Attali, Dean. 2021. *Shinyjs: Easily Improve the User Experience of Your Shiny Apps in Seconds*. https://deanattali.com/shinyjs/.

Baglivo, Sara, Vienna Ludovini, Angelo Sidoni, Giulio Metro, Biagio Ricciuti, Annamaria Siggillino, Alberto Rebonato, Salvatore Messina, Lucio Crinò, and Rita Chiari. 2017. "Large Cell Neuroendocrine Carcinoma Transformation and EGFR-T790M Mutation as Coexisting Mechanisms of Acquired Resistance to EGFR-TKIs in Lung Cancer." *Mayo Clinic Proceedings* 92 (8): 1304–11. https://doi.org/https://doi.org/10.1016/j.mayocp.2017.03.022.

Bairoch, Amos. 2018. "The Cellosaurus, a Cell-Line Knowledge Resource." *Journal of Biomolecular Techniques* 29 (2): 25–38. https://doi.org/10.7171/jbt.18-2902-002.

Balanis, Nikolas G., Katherine M. Sheu, Favour N. Esedebe, Saahil J. Patel, Bryan A. Smith, Jung Wook Park, Salwan Alhani, et al. 2019. "Pan-cancer Convergence to a Small-Cell Neuroendocrine Phenotype that Shares Susceptibilities with Hematological Malignancies." *Cancer Cell* 36 (1): 17–34.e7. https://doi.org/10.1016/j.ccell.2019.06.005.

Barbie, David A., Pablo Tamayo, Jesse S. Boehm, So Young Kim, Susan E. Moody, Ian F. Dunn, Anna C. Schinzel, et al. 2009. "Systematic RNA Interference Reveals That Oncogenic KRAS-Driven Cancers Require TBK1." *Nature* 462: 108–12. https://doi.org/10.1038/nature08460.

Barretina, Jordi, Giordano Caponigro, and Nicolas Stransky. 2012. "The Cancer Cell Line Encyclopedia Enables Predictive Modeling of Anticancer Drug Sensitivity." *Nature* 483: 603–7. https://doi.org/10.1038/nature11003.The.

Barrett, Alec, Erdem Varol, Alexis Weinreb, Seth R. Taylor, Rebecca D McWhirter, Cyril Cros, Manasa Basavaraju, et al. 2022. "Integrating Bulk and Single Cell RNA-Seq Refines Transcriptomic Profiles of Specific c. Elegans Neurons." *bioRxiv*, 1–45.

Beaubier, Nike, Martin Bontrager, Robert Huether, Catherine Igartua, Denise Lau, Robert Tell, Alexandria M. Bobe, et al. 2019. "Integrated genomic profiling expands clinical options for patients with cancer." *Nature Biotechnology* 37 (11): 1351–60. https://doi.org/10.1038/s41587-019-0259-z.

Beltran, Himisha, David S. Rickman, Kyung Park, Sung Suk Chae, Andrea Sboner, Theresa Y. MacDonald, Yuwei Wang, et al. 2011. "Molecular characterization of neuroendocrine prostate cancer and identification of new drug targets." *Cancer Discovery* 1 (6): 487–95. https://doi.org/10.1158/2159-8290.CD-11-0130.

Bennette, Caroline, and Andrew Vickers. 2012. "Against Quantiles: Categorization of Continuous Variables in Epidemiologic Research, and Its Discontents." *BMC Medical Research Methodology* 12 (1): 21. https://doi.org/10.1186/1471-2288-12-21.

Bild, Andrea H., Guang Yao, Jeffrey T. Chang, Quanli Wang, Anil Potti, Dawn Chasse, Mary Beth Joshi, et al. 2006. "Oncogenic Pathway Signatures in Human Cancers as a Guide to Targeted Therapies." *Nature* 439: 353–57. https://doi.org/10.1038/nature04296.

Bindea, Gabriela, Bernhard Mlecnik, Marie Tosolini, Amos Kirilovsky, Maximilian Waldner, Anna C. Obenauf, Helen Angell, et al. 2013. "Spatiotemporal dynamics of intratumoral immune cells reveal the immune landscape in human cancer." *Immunity* 39 (4): 782–95. https://doi.org/10.1016/j.immuni.2013.10.003.

Bonferroni, Carlo E. 1936. "Teoria Statistica Delle Classi e Calcolo Delle Probabilita." *Pubblicazioni Del R Istituto Superiore Di Scienze Economiche e Commericiali Di Firenze* 8: 3–62. https://cir.nii.ac.jp/crid/1570009749360424576.

Brenton, James D., Lisa A. Carey, Ashour Ahmed, and Carlos Caldas. 2005. "Molecular Classification and Molecular Forecasting of Breast Cancer: Ready for Clinical Application?" *Journal of Clinical Oncology*. https://doi.org/10.1200/JCO.2005.03.3845.

Budinska, Eva, Vlad Popovici, Sabine Tejpar, Giovanni D'Ario, Nicolas Lapique, Katarzyna Otylia Sikora, Antonio Fabio Di Narzo, et al. 2013. "Gene expression patterns unveil a new level of molecular heterogeneity in colorectal cancer." *Journal of Pathology* 231 (1): 63–76. https://doi.org/10.1002/path.4212.

Burns, Laura, Hanna Tukachinsky, Kira Raskina, Richard S. P. Huang, Alexa B. Schrock, Jacob Sands, Matthew H. Kulke, Geoffrey R. Oxnard, and Umit Tapan. 2024. "Real-World Comprehensive Genomic Profiling Data for Diagnostic Clarity in Pulmonary Large-Cell Neuroendocrine Carcinoma." *Lung Cancer* 188 (February): 107454. https://doi.org/10.1016/j.lungcan.2023.107454.

Busch, Evan L. 2021. "Cut Points and Contexts." *Cancer*. John Wiley; Sons Inc. https://doi.org/10.1002/cncr.33838.

Butler, Andrew, Paul Hoffman, Peter Smibert, Efthymia Papalexi, and Rahul Satija. 2018. "Integrating Single-Cell Transcriptomic Data Across Different Conditions, Technologies, and Species." *Nature Biotechnology* 36: 411–20. https://doi.org/10.1038/nbt.4096.

Calon, Alexandre, Enza Lonardo, Antonio Berenguer-Llergo, Elisa Espinet, Xavier Hernando-Momblona, Mar Iglesias, Marta Sevillano, et al. 2015. "Stromal Gene Expression Defines Poor-Prognosis Subtypes in Colorectal Cancer." *Nature Genetics* 47 (4): 320–29. https://doi.org/10.1038/ng.3225.

Calza, Stefano, Per Hall, Gert Auer, Judith Bjöhle, Sigrid Klaar, Ulrike Kronenwett, Edison T. Liu, et al. 2006. "Intrinsic Molecular Signature of Breast Cancer in a Population-Based Cohort of 412 Patients." *Breast Cancer Research* 8 (July): R34. https://doi.org/10.1186/bcr1517.

Cantini, Laura, Laurence Calzone, Loredana Martignetti, Mattias Rydenfelt, Nils Blüthgen, Emmanuel Barillot, and Andrei Zinovyev. 2018. "Classification of Gene Signatures for Their Information Value and Functional Redundancy." *Npj Systems Biology and Applications* 4: 2. https://doi.org/10.1038/s41540-017-0038-8.

Chang, Winston, Joe Cheng, JJ Allaire, Carson Sievert, Barret Schloerke, Yihui Xie, Jeff Allen, Jonathan McPherson, Alan Dipert, and Barbara Borges. 2022. *Shiny: Web Application Framework for r*. https://shiny.rstudio.com/.

Chen, Andrew X., Robyn D. Gartrell, Junfei Zhao, Pavan S. Upadhyayula, Wenting Zhao, Jinzhou Yuan, Hanna E. Minns, et al. 2021. "Single-Cell Characterization of Macrophages in Glioblastoma Reveals MARCO as a Mesenchymal Pro-Tumor Marker." *Genome Medicine* 13 (December): 88. https://doi.org/10.1186/s13073-021-00906-x.

Chen, Yajun, Qican Deng, Hui Chen, Jianguo Yang, Zhenzhou Chen, Juncai Li, and Zhongxue Fu. 2023. "Cancer-Associated Fibroblast-Related Prognostic Signature Predicts Prognosis and Immunotherapy Response in Pancreatic Adenocarcinoma Based on Single-Cell and Bulk RNA-Sequencing." *Scientific Reports* 13 (December): 16408. https://doi.org/10.1038/s41598-023-43495-y.

Chibon, Frederic. 2013. "Cancer Gene Expression Signatures-the Rise and Fall?" *European Journal of Cancer* 49: 2000–2009. https://doi.org/10.1016/j.ejca.2013.02.021.

Chowdhury, Saikat, Matan Hofree, Kangyu Lin, Dipen Maru, Scott Kopetz, and John Paul Shen. 2021. "Implications of Intratumor Heterogeneity on Consensus Molecular Subtype (CMS) in Colorectal Cancer." *Cancers* 13 (19): 4923. https://doi.org/10.3390/cancers13194923.

Christen, Peter, David J. Hand, and Nishadi Kirielle. 2023. "A Review of the f-Measure: Its History, Properties, Criticism, and Alternatives." *ACM Computing Surveys* 56 (October): 1–24. https://doi.org/10.1145/3606367.

Compagno, Mara, Wei Keat Lim, Adina Grunn, Subhadra V. Nandula, Manisha Brahmachary, Qiong Shen, Francesco Bertoni, et al. 2009. "Mutations of Multiple Genes Cause Deregulation of NF-b in Diffuse Large b-Cell Lymphoma." *Nature* 459 (June): 717–21. https://doi.org/10.1038/nature07968.

Creighton, C. J. 2007. "A Gene Transcription Signature of the Akt/mTOR Pathway in Clinical Breast Tumors." *Oncogene* 26 (July): 4648–55. https://doi.org/10.1038/sj.onc.1210245.

Crona, Joakim, and Britt Skogseid. 2016. "Genetics of neuroendocrine tumors." *European Journal of Endocrinology* 174 (6): R275–90. https://doi.org/10.1530/EJE-15-0972.

Dai, Hongyue, Laura van't Veer, John Lamb, Yudong D. He, Mao Mao, Bernard M. Fine, Rene Bernards, et al. 2005. "A Cell Proliferation Signature Is a Marker of Extremely Poor Outcome in a Subpopulation of Breast Cancer Patients." *Cancer Research* 65 (10): 4059–66. https://doi.org/10.1158/0008-5472.can-04-3953.

Dempster, Joshua M., Jordan Rossen, Mariya Kazachkova, Joshua Pan, Guillaume Kugener, David E. Root, and Aviad Tsherniak. 2019. "Extracting Biological Insights from the Project Achilles Genome-Scale CRISPR Screens in Cancer Cell Lines." *bioRxiv*, 720243. https://doi.org/10.1101/720243.

Denkert, Carsten, Stephan Wienert, and Frederick Klauschen. 2018. "Analyzing the Immunological Landscape of a Tumor—Heterogeneity of Immune Infiltrates in Breast Cancer as a New Prognostic Indicator." *Journal of the National Cancer Institute*. Oxford University Press. https://doi.org/10.1093/jnci/djx188.

Derks, Jules L., Anne Marie C. Dingemans, Robert Jan van Suylen, Michael A. den Bakker, Ronald A. M. Damhuis, Esther C. van den Broek, Ernst Jan Speel, and Erik Thunnissen. 2019. "Is the sum of positive neuroendocrine immunohistochemical stains useful for diagnosis of large cell neuroendocrine carcinoma (LCNEC) on biopsy specimens?" *Histopathology* 74 (4): 555–66. https://doi.org/10.1111/his.13800.

Dhanasekaran, Renumathy, Anja Deutzmann, Wadie D. Mahauad-Fernandez, Aida S. Hansen, Arvin M. Gouw, and Dean W. Felsher. 2022. "The MYC Oncogene — the Grand Orchestrator of Cancer Growth and Immune Evasion." *Nature Reviews Clinical Oncology*. Nature Research. https://doi.org/10.1038/s41571-021-00549-2.

Dhawan, Andrew, Alessandro Barberis, Wei Chen Cheng, Enric Domingo, Catharine West, Tim Maughan, Jacob G. Scott, Adrian L. Harris, and

Francesca M. Buffa. 2019. "Guidelines for Using sigQC for Systematic Evaluation of Gene Signatures." *Nature Protocols* 14: 1377–1400. https://doi.org/10.1038/s41596-019-0136-8.

Dietrich, Alexander. 2023. *SimBu: Simulate Bulk RNA-Seq Datasets from Single-Cell Datasets*. https://github.com/omnideconv/SimBu.

Dingemans, A. M. C., M. Früh, A. Ardizzoni, B. Besse, C. Faivre-Finn, L. E. Hendriks, S. Lantuejoul, et al. 2021. "Small-Cell Lung Cancer: ESMO Clinical Practice Guidelines for Diagnosis, Treatment and Follow-up ." *Annals of Oncology* 32 (July): 839–53. https://doi.org/10.1016/j.annonc.2021.03.207.

Doldman, Mary J, Brain Craft, Mim Hastie, Kristupas Repečka, Fran McDade, Akhil Kamath, Ayan Banerjee, et al. 2020. "Visualizing and Interpreting Cancer Genomics Data via the Xena Platform." *Nature Biotechnology* 38 (6): 669–73. https://doi.org/10.1038/s41587-020-0550-z.

Dongre, Anushka, and Robert A. Weinberg. 2019. "New Insights into the Mechanisms of Epithelial–Mesenchymal Transition and Implications for Cancer." *Nature Reviews Molecular Cell Biology.* Nature Publishing Group. https://doi.org/10.1038/s41580-018-0080-4.

Dudnik, Elizabeth, Samuel Kareff, Mor Moskovitz, Chul Kim, Stephen V Liu, Anastasiya Lobachov, Teodor Gottfried, et al. 2021. "Real-World Survival Outcomes with Immune Checkpoint Inhibitors in Large-Cell Neuroendocrine Tumors of Lung." *Journal for ImmunoTherapy of Cancer* 9 (2): e001999. https://doi.org/10.1136/jitc-2020-001999.

Dummer, Reinhard, Jan C. Brase, James Garrett, Catarina D. Campbell, Eduard Gasal, Matthew Squires, Daniel Gusenleitner, et al. 2020. "Adjuvant Dabrafenib Plus Trametinib Versus Placebo in Patients with Resected, BRAFV600-Mutant, Stage III Melanoma (COMBI-AD): Exploratory Biomarker Analyses from a Randomised, Phase 3 Trial." *The Lancet Oncology* 21: 358–72. https://doi.org/10.1016/S1470-2045(20)30062-0.

Dunne, Philip D., Darragh G. McArt, Conor A. Bradley, Paul G. O'Reilly, Helen L. Barrett, Robert Cummins, Tony O'Grady, et al. 2016. "Challenging the Cancer Molecular Stratification Dogma: Intratumoral Heterogeneity Undermines Consensus Molecular Subtypes and Potential Diagnostic Value in Colorectal Cancer." *Clinical Cancer Research* 22 (August): 4095–4104. https://doi.org/10.1158/1078-0432.CCR-16-0032.

Ebi, Hiromichi, Shuta Tomida, Toshiyuki Takeuchi, Chinatsu Arima, Takahiko Sato, Tetsuya Mitsudomi, Yasushi Yatabe, Hirotaka Osada, and Takashi Takahashi. 2009. "Relationship of Deregulated Signaling Converging onto mTOR with Prognosis and Classification of Lung Adenocarcinoma Shown by Two Independent in Silico Analyses." *Cancer Research* 69: 4027–35. https://doi.org/10.1158/0008-5472.CAN-08-3403.

Eftekhari Kenzerki, Maryam, Mohsen Ahmadi, Pegah Mousavi, and Soudeh Ghafouri-Fard. 2023. "MYC and Non-Small Cell Lung Cancer: A Comprehensive Review." *Human Gene*. Elsevier BV. https://doi.org/10.1016/j.humgen.2023.201185.

Eide, Peter W., Jarle Bruun, Ragnhild A. Lothe, and Anita Sveen. 2017. "CMScaller: An r Package for Consensus Molecular Subtyping of Colorectal Cancer Pre-Clinical Models." *Scientific Reports* 7: 1–8. https://doi.org/10.1038/s41598-017-16747-x.

Fan, Jean, Neeraj Salathia, Rui Liu, Gwendolyn E. Kaeser, Yun C. Yung, Joseph L. Herman, Fiona Kaper, et al. 2016. "Characterizing Transcriptional Heterogeneity Through Pathway and Gene Set Overdispersion Analysis." *Nature Methods* 13: 241–44. https://doi.org/10.1038/nmeth.3734.

Farmer, Pierre, Hervé Bonnefoi, Pascale Anderle, David Cameron, Pratyakasha Wirapati, Véronique Becette, Sylvie André, et al. 2009. "A stroma-related gene signature predicts resistance to neoadjuvant chemotherapy in breast cancer." *Nature Medicine* 15 (1): 68–74. https://doi.org/10.1038/nm.1908.

Farmer, Pierre, Herve Bonnefoi, Veronique Becette, Michele Tubiana-Hulin, Pierre Fumoleau, Denis Larsimont, Gaetan MacGrogan, et al. 2005. "Identification of Molecular Apocrine Breast Tumours by Microarray Analysis." *Oncogene* 24 (29): 4660–71. https://doi.org/10.1038/sj.onc.1208561.

Feeley, Linda P., Anna M. Mulligan, Dushanthi Pinnaduwage, Shelley B. Bull, and Irene L. Andrulis. 2014. "Distinguishing Luminal Breast Cancer Subtypes by Ki67, Progesterone Receptor or TP53 Status Provides Prognostic Information." *Modern Pathology* 27: 554–61. https://doi.org/10.1038/modpathol.2013.153.

Finotello, Francesca, Clemens Mayer, Christina Plattner, Gerhard Laschober, DIetmar Rieder, Hubert Hackl, Anne Krogsdam, et al. 2019. "Molecular and Pharmacological Modulators of the Tumor Immune Contexture Revealed by Deconvolution of RNA-Seq Data." *Genome Medicine* 11 (May): 34. https://doi.org/10.1186/s13073-019-0638-6.

Foroutan, Momeneh, Dharmesh D. Bhuva, Ruqian Lyu, Kristy Horan, Joseph Cursons, and Melissa J. Davis. 2018. "Single Sample Scoring of Molecular Phenotypes." *BMC Bioinformatics* 19 (November): 1–10. https://doi.org/10.1186/s12859-018-2435-4.

Fougner, Christian, Helga Bergholtz, Jens Henrik Norum, and Therese Sørlie. 2020. "Re-Definition of Claudin-Low as a Breast Cancer Phenotype." *Nature*

Communications 11 (December): 1787. https://doi.org/10.1038/s41467-020-15574-5.

Frankish, Adam, Mark Diekhans, Anne Maud Ferreira, Rory Johnson, Irwin Jungreis, Jane Loveland, Jonathan M. Mudge, et al. 2019. "GENCODE Reference Annotation for the Human and Mouse Genomes." *Nucleic Acids Research* 47 (January): D766–73. https://doi.org/10.1093/nar/gky955.

Fu, Xiaoyong, Carmine De Angelis, and Rachel Schiff. 2021. "Interferon Signaling in Estrogen Receptor–positive Breast Cancer: A Revitalized Topic." *Endocrinology* 163 (1): bqab235. https://doi.org/10.1210/endocr/bqab235.

Gao, Ruli, Shanshan Bai, Ying C. Henderson, Yiyun Lin, Aislyn Schalck, Yun Yan, Tapsi Kumar, et al. 2021. "Delineating Copy Number and Clonal Substructure in Human Tumors from Single-Cell Transcriptomes." *Nature Biotechnology* 39 (May): 599–608. https://doi.org/10.1038/s41587-020-00795-2.

George, Julie, Jing Shan Lim, Se Jin Jang, Yupeng Cun, Luka Ozretia, Gu Kong, Frauke Leenders, et al. 2015. "Comprehensive Genomic Profiles of Small Cell Lung Cancer." *Nature* 524 (August): 47–53. https://doi.org/10.1038/nature14664.

George, Julie, Vonn Walter, Martin Peifer, Ludmil B Alexandrov, Danila Seidel, Frauke Leenders, Lukas Maas, et al. 2018. "Integrative genomic profiling of large-cell neuroendocrine carcinomas reveals distinct subtypes of high-grade neuroendocrine lung tumors." *Nature Communications* 9 (1): 1048. https://doi.org/10.1038/s41467-018-03099-x.

Ghandi, Mahmoud, Franklin W. Huang, Judit Jané-Valbuena, V. Gregory Kryukov, Christopher C. Lo, E. Robert McDonald, Jordi Barretina, et al. 2019. "Next-Generation Characterization of the Cancer Cell Line Encyclopedia." *Nature* 569 (7757): 503–8. https://doi.org/10.1038/s41586-019-1186-3.

Gleiss, Andreas, Rainer Oberbauer, and Georg Heinze. 2018. "An Unjustified Benefit: Immortal Time Bias in the Analysis of Time-Dependent Events." *Transplant International* 31 (2): 125–30. https://doi.org/https://doi.org/10.1111/tri.13081.

Grossman, Robert L., Allison Heath P., Vincent; Ferretti, Harold E. Varmus, Douglas R. Lowy, Warren A. Kibbe, and Stuadt. Louis M. 2016. "Toward a Shared Vision for Cancer Genomic Data." *The New England Journal of Medicine*, 1109–12.

Guinney, Justin, Rodrigo Dienstmann, Xin Wang, Aurélien De Reyniès, Andreas Schlicker, Charlotte Soneson, Laetitia Marisa, et al. 2015. "The

Consensus Molecular Subtypes of Colorectal Cancer." *Nature Medicine* 21: 1350–56. https://doi.org/10.1038/nm.3967.

Guo, Chengbin, Yuqin Tang, Zhao Yang, Gen Li, and Yongqiang Zhang. 2022. "Hallmark-Guided Subtypes of Hepatocellular Carcinoma for the Identification of Immune-Related Gene Classifiers in the Prediction of Prognosis, Treatment Efficacy, and Drug Candidates." *Frontiers in Immunology* 13 (August). https://doi.org/10.3389/fimmu.2022.958161.

Han, Ya, Yuting Wang, Xin Dong, Dongqing Sun, Zhaoyang Liu, Jiali Yue, Haiyun Wang, Taiwen Li, and Chenfei Wang. 2023. "TISCH2: Expanded Datasets and New Tools for Single-Cell Transcriptome Analyses of the Tumor Microenvironment." *Nucleic Acids Research* 51 (January): D1425–31. https://doi.org/10.1093/nar/gkac959.

Hanahan, Douglas. 2022. "Hallmarks of Cancer: New Dimensions." *Cancer Discovery* 12 (1): 31–46. https://doi.org/10.1158/2159-8290.CD-21-1059.

Hanahan, Douglas, and Robert A. Weinberg. 2000. "The Hallmarks of Cancer Review." *Cell* 100: 57–70. https://doi.org/10.1016/S0092-8674(00)81683-9.

———. 2011. "Hallmarks of cancer: The next generation." *Cell* 144 (5): 646–74. https://doi.org/10.1016/j.cell.2011.02.013.

Hänzelmann, Sonja, Robert Castelo, and Justin Guinney. 2013. "GSVA: Gene Set Variation Analysis for Microarray and RNA-Seq Data." *BMC Bioinformatics*. http://www.biomedcentral.com/1471-2105/14/7http://www.bioconductor.org.Background.

Hao, Yuhan, Stephanie Hao, Erica Andersen-Nissen, William M. Mauck, Shiwei Zheng, Andrew Butler, Maddie J. Lee, et al. 2021. "Integrated Analysis of Multimodal Single-Cell Data." *Cell* 184 (13): 3573–3587.e29. https://doi.org/10.1016/j.cell.2021.04.048.

Hao, Yuhan, Tim Stuart, Madeline H. Kowalski, Saket Choudhary, Paul Hoffman, Austin Hartman, Avi Srivastava, et al. 2023. "Dictionary Learning for Integrative, Multimodal and Scalable Single-Cell Analysis." *Nature Biotechnology* 42 (2): 293–304. https://doi.org/10.1038/s41587-023-01767-y.

Harbeck, Nadia, Frédérique Penault-Llorca, Javier Cortes, Michael Gnant, Nehmat Houssami, Philip Poortmans, Kathryn Ruddy, Janice Tsang, and Fatima Cardoso. 2019. *Breast Cancer*. *Nature Reviews Disease Primers*. Vol. 5. https://doi.org/10.1038/s41572-019-0111-2.

Heijboer, F., J. Derks, N. Rijnsburger, B. Hermans, L. Hillen, E. van den Broek, E-J. Speel, A-M.C. Dingemans, and J. von der Thüsen. 2023. "2202P Large Cell Neuroendocrine Carcinoma (LCNEC) Subtyping Based on NEUROD1, ASCL1, POU2F3 and YAP1 Expression." *Annals of Oncology* 34 (October):

S1138. https://doi.org/10.1016/j.annonc.2023.09.984.

Hendriks, L. E., K. M. Kerr, J. Menis, T. S. Mok, U. Nestle, A. Passaro, S. Peters, et al. 2023. "Non-Oncogene-Addicted Metastatic Non-Small-Cell Lung Cancer: ESMO Clinical Practice Guideline for Diagnosis, Treatment and Follow-up ." *Annals of Oncology* 34 (April): 358–76. https://doi.org/10.1016/j.annonc.2022.12.013.

Hoadley, Katherine A., Christina Yau, Toshinori Hinoue, Denise M. Wolf, Alexander J. Lazar, Esther Drill, Ronglai Shen, et al. 2018. "Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer." *Cell* 173 (April): 291–304.e6. https://doi.org/10.1016/j.cell.2018.03.022.

Hollander, Myles, Douglas A. Wolfe, and Eric Chicken. 2015. *Nonparametric Statistical Methods. Wiley Series in Probability and Statistics.* Wiley. https://doi.org/10.1002/9781119196037.

Holm, Sture. 1979. "A Simple Sequentially Rejective Multiple Test Procedure a Simple Sequentially Rejective Multiple Test Procedure." *Source: Scandinavian Journal of Statistics.*

Hou, Jun, Joachim Aerts, Bianca den Hamer, Wilfred van IJcken, Michael den Bakker, Peter Riegman, Cor van der Leest, et al. 2010. "Gene Expression-Based Classification of Non-Small Cell Lung Carcinomas and Survival Prediction." *PLoS ONE* 5 (April): e10312. https://doi.org/10.1371/journal.pone.0010312.

Hu, Wei, Yangjun Wu, Qili Shi, Jingni Wu, Deping Kong, Xiaohua Wu, Xianghuo He, Teng Liu, and Shengli Li. 2022. "Systematic Characterization of Cancer Transcriptome at Transcript Resolution." *Nature Communications* 13 (December): 6803. https://doi.org/10.1038/s41467-022-34568-z.

Iannone, Richard, Joe Cheng, Barret Schloerke, Ellis Hughes, Alexandra Lauer, and JooYoung Seo. 2024. *Gt: Easily Create Presentation-Ready Display Tables.*

Igawa, Satoshi, Reiko Watanabe, Ichiro Ito, Haruyasu Murakami, Toshiaki Takahashi, Yukiko Nakamura, Asuka Tsuya, et al. 2010. "Comparison of Chemotherapy for Unresectable Pulmonary High-Grade Non-Small Cell Neuroendocrine Carcinoma and Small-Cell Lung Cancer." *Lung Cancer* 68 (June): 438–45. https://doi.org/10.1016/j.lungcan.2009.07.003.

Isella, Claudio, Andrea Terrasi, Sara Erika Bellomo, Consalvo Petti, Giovanni Galatola, Andrea Muratore, Alfredo Mellano, et al. 2015. "Stromal Contribution to the Colorectal Cancer Transcriptome." *Nature Genetics* 47 (4): 312–19. https://doi.org/10.1038/ng.3224.

Izar, Benjamin, Itay Tirosh, Elizabeth H. Stover, Isaac Wakiro, Michael S. Cuoco, Idan Alter, Christopher Rodman, et al. 2020. "A Single-Cell Landscape of High-Grade Serous Ovarian Cancer." *Nature Medicine* 26 (8): 1271–79. https://doi.org/10.1038/s41591-020-0926-0.

Jain, Shantanu, Martha White, and Predrag Radivojac. 2017. "Recovering True Classifier Performance in Positive-Unlabeled Learning." arXiv. https://doi.org/10.48550/ARXIV.1702.00518.

Kassambara, Alboukadel. 2023. *Rstatix: Pipe-Friendly Framework for Basic Statistical Tests*. https://rpkgs.datanovia.com/rstatix/.

Kassambara, Alboukadel, Marcin Kosinski, and Przemyslaw Biecek. 2021. *Survminer: Drawing Survival Curves Using 'Ggplot2'*. https://rpkgs.datanovia.com/survminer/index.html.

Kinslow, Connor J., Michael S. May, Anjali Saqi, Catherine A. Shu, Kunal R. Chaudhary, Tony J. C. Wang, and Simon K. Cheng. 2020. "Large-Cell Neuroendocrine Carcinoma of the Lung: A Population-Based Study." *Clinical Lung Cancer* 21 (2): e99–113. https://doi.org/10.1016/j.cllc.2019.07.011.

Kirch, Wilhelm, ed. 2008. "Pearson's Correlation Coefficient." In *Encyclopedia of Public Health*, 1090–91. Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-1-4020-5614-7_2569.

Klik, Mark. 2022. *Fst: Lightning Fast Serialization of Data Frames*. http://www.fstpackage.org.

Klopfenstein, Quentin, Valentin Derangère, Laurent Arnould, Marion Thibaudin, Emeric Limagne, Francois Ghiringhelli, Caroline Truntzer, and Sylvain Ladoire. 2021. "Evaluation of Tumor Immune Contexture Among Intrinsic Molecular Subtypes Helps to Predict Outcome in Early Breast Cancer." *Journal for ImmunoTherapy of Cancer* 9 (6): e002036. https://doi.org/10.1136/jitc-2020-002036.

Kreis, Julian, Bogac Aybey, Felix Geist, Benedikt Brors, and Eike Staub. 2024. "Stromal Signals Dominate Gene Expression Signature Scores That Aim to Describe Cancer Cell–Intrinsic Stemness or Mesenchymality Characteristics." *Cancer Research Communications* 4 (2): 516–29. https://doi.org/10.1158/2767-9764.crc-23-0383.

Kreis, Julian, Boro Nedić, Johanna Mazur, Miriam Urban, Sven Eric Schelhorn, Thomas Grombacher, Felix Geist, Benedikt Brors, Michael Zühlsdorf, and Eike Staub. 2021. "RosettaSX: Reliable Gene Expression Signature Scoring of Cancer Models and Patients." *Neoplasia (United States)* 23: 1069–77. https://doi.org/10.1016/j.neo.2021.08.005.

Lan, Qiang, Sanam Peyvandi, Nathalie Duffey, Yu Ting Huang, David Barras, Werner Held, François Richard, et al. 2019. "Type i Interferon/IRF7 Axis Instigates Chemotherapy-Induced Immunological Dormancy in Breast Cancer." *Oncogene* 38 (April): 2814–29. https://doi.org/10.1038/s41388-018-0624-2.

Lantuejoul, Sylvie, Lynnette Fernandez-Cuesta, Francesca Damiola, Nicolas Girard, and Anne McLeer. 2020. "New molecular classification of large cell neuroendocrine carcinoma and small cell lung carcinoma with potential therapeutic impacts." *Translational Lung Cancer Research* 9 (5): 2233–44. https://doi.org/10.21037/TLCR-20-269.

Lázaro, Sara, Miriam Pérez-Crespo, Corina Lorz, Alejandra Bernardini, Marta Oteo, Ana Belén Enguita, Eduardo Romero, et al. 2019. "Differential Development of Large-Cell Neuroendocrine or Small-Cell Lung Carcinoma Upon Inactivation of 4 Tumor Suppressor Genes." *Proceedings of the National Academy of Sciences of the United States of America* 116 (October): 22300–22306. https://doi.org/10.1073/pnas.1821745116.

Lee, Eunjung, Han Yu Chuang, Jong Won Kim, Trey Ideker, and Doheon Lee. 2008. "Inferring Pathway Activity Toward Precise Disease Classification." *PLoS Computational Biology* 4: e1000217. https://doi.org/10.1371/journal.pcbi.1000217.

Lee, Hae Ock, Yourae Hong, Hakki Emre Etlioglu, Yong Beom Cho, Valentina Pomella, Ben Van den Bosch, Jasper Vanhecke, et al. 2020. "Lineage-Dependent Gene Expression Programs Influence the Immune Landscape of Colorectal Cancer." *Nature Genetics* 52 (6): 594–603. https://doi.org/10.1038/s41588-020-0636-z.

Lee, Matthew, Dhruv Patel, Sebastian Jofre, Shabnam Fidvi, Mark Suhrland, Perry Cohen, and Haiying Cheng. 2022. "Large Cell Neuroendocrine Carcinoma Transformation as a Mechanism of Acquired Resistance to Osimertinib in Non-Small Cell Lung Cancer: Case Report and Literature Review." *Clinical Lung Cancer* 23 (May): e276–82. https://doi.org/10.1016/j.cllc.2021.08.002.

Lee, Wee Sun, and Bing Liu. 2003. "Learning with Positive and Unlabeled Examples Using Weighted Logistic Regression." In *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*, 448–55. ICML'03. Washington, DC, USA: AAAI Press.

Lehmann, Brian D., Joshua A. Bauer, Xi Chen, Melinda E. Sanders, A. Bapsi Chakravarthy, Yu Shyr, and Jennifer A. Pietenpol. 2011. "Identification of Human Triple-Negative Breast Cancer Subtypes and Preclinical Models for Selection of Targeted Therapies." *Journal of Clinical Investigation* 121: 2750–67. https://doi.org/10.1172/JCI45014.

Lehmann, Brian D., Bojana Jovanović, Xi Chen, V. Monica Estrada, Kimberly N. Johnson, Yu Shyr, Harold L. Moses, Melinda E. Sanders, and Jennifer A. Pietenpol. 2016. "Refinement of Triple-Negative Breast Cancer Molecular Subtypes: Implications for Neoadjuvant Chemotherapy Selection." *PLoS ONE* 11 (6): 1–22. https://doi.org/10.1371/journal.pone.0157368.

Li, Meihui, Lan Yang, and Hongyang Lu. 2022. "Pulmonary Combined Large Cell Neuroendocrine Carcinoma." *Pathology and Oncology Research* 28 (November): 1–6. https://doi.org/10.3389/pore.2022.1610747.

Li, Mingyue, Jiaying Wang, Conghui Wang, Lili Xia, Junfen Xu, Xing Xie, and Weiguo Lu. 2020. "Microenvironment Remodeled by Tumor and Stromal Cells Elevates Fibroblast-Derived COL1A1 and Facilitates Ovarian Cancer Metastasis." *Experimental Cell Research* 394 (1): 112153. https://doi.org/10.1016/j.yexcr.2020.112153.

Li, M., and W. Lu. 2020. "Fibroblasts-Secreted Collagen Type i Alpha 1 Drives a Metastasis-Promoting Microenvironment in Ovarian Cancer." *Gynecologic Oncology* 159 (October): 348. https://doi.org/10.1016/j.ygyno.2020.05.648.

Liang, Yu, Maximilian Diehn, Nathan Watson, Andrew W. Bollen, Ken D. Aldape, M. Kelly Nicholas, Kathleen R. Lamborn, et al. 2005. "Gene Expression Profiling Reveals Molecularly and Clinically Distinct Subtypes of Glioblastoma Multiforme." *Proceedings of the National Academy of Sciences of the United States of America* 102: 5814–19. https://doi.org/10.1073/pnas.0402870102.

Liberzon, Arthur, Chet Birger, Helga Thorvaldsdóttir, Mahmoud Ghandi, Jill P. Mesirov, and Pablo Tamayo. 2015. "The Molecular Signatures Database Hallmark Gene Set Collection." *Cell Systems* 1 (6): 417–25. https://doi.org/10.1016/j.cels.2015.12.004.

Lien, H. C., Y. H. Hsiao, Y. S. Lin, Y. T. Yao, H. F. Juan, W. H. Kuo, Mien Chie Hung, K. J. Chang, and F. J. Hsieh. 2007. "Molecular Signatures of Metaplastic Carcinoma of the Breast by Large-Scale Transcriptional Profiling: Identification of Genes Potentially Related to Epithelial-Mesenchymal Transition." *Oncogene* 26: 7859–71. https://doi.org/10.1038/sj.onc.1210593.

Lim, Jeong Uk, In Sook Woo, Yun Hwa Jung, Jae Ho Byeon, Chan Kwon Park, Tae Jung Kim, and Hyo Rim Kim. 2014. "Transformation into Large-Cell Neuroendocrine Carcinoma Associated with Acquired Resistance to Erlotinib in Nonsmall Cell Lung Cancer." *Korean Journal of Internal Medicine*. Korean Association of Internal Medicine. https://doi.org/10.

3904/kjim.2014.29.6.830.

Lindsay, Colin R., Emily C. Shaw, David A. Moore, Doris Rassl, Mariam Jamal-Hanjani, Nicola Steele, Salma Naheed, et al. 2021. “Large cell neuroendocrine lung carcinoma: consensus statement from The British Thoracic Oncology Group and the Association of Pulmonary Pathologists.” Springer Nature. https://doi.org/10.1038/s41416-021-01407-9.

Liu, Jianfang, Tara Lichtenberg, Katherine A. Hoadley, Laila M. Poisson, Alexander J. Lazar, Andrew D. Cherniack, Albert J. Kovatich, et al. 2018. “An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics.” *Cell* 173 (2): 400–416.e11. https://doi.org/10.1016/j.cell.2018.02.052.

Loibl, S, F André, T Bachelot, C H Barrios, J Bergh, H J Burstein, M J Cardoso, et al. 2024. “Early Breast Cancer: ESMO Clinical Practice Guideline for Diagnosis, Treatment and Follow-up 5 Behalf of the ESMO Guidelines Committee.” *Annals of Oncology* 35: 159–82. https://doi.org/10.1016/j.

Luca, Bogdan A., Chloé B. Steen, Magdalena Matusiak, Armon Azizi, Sushama Varma, Chunfang Zhu, Joanna Przybyl, et al. 2021. “Atlas of Clinically Distinct Cell States and Ecosystems Across Human Solid Tumors.” *Cell* 184 (October): 5482–5496.e28. https://doi.org/10.1016/j.cell.2021.09.014.

Lüönd, Fabiana, Nami Sugiyama, Ruben Bill, Laura Bornes, Carolina Hager, Fengyuan Tang, Natascha Santacroce, et al. 2021. “Distinct Contributions of Partial and Full EMT to Breast Cancer Malignancy.” *Developmental Cell* 56 (December): 3203–3221.e11. https://doi.org/10.1016/j.devcel.2021.11.006.

Mak, Milena P., Pan Tong, Lixia Diao, Robert J. Cardnell, Don L. Gibbons, William N. William, Ferdinandos Skoulidis, et al. 2016. “A Patient-Derived, Pan-Cancer EMT Signature Identifies Global Molecular Alterations and Immune Target Enrichment Following Epithelial-to-Mesenchymal Transition.” *Clinical Cancer Research* 22 (February): 609–20. https://doi.org/10.1158/1078-0432.CCR-15-0876.

Mani, Sendurai A, Wenjun Guo, Mai Jing Liao, Elinor Ng Eaton, Ayyakkannu Ayyanan, Alicia Y Zhou, Mary Brooks, et al. 2008. “The Epithelial-Mesenchymal Transition Generates Cells with Properties of Stem Cells.” *Cell* 133: 704–15. https://doi.org/10.1016/j.cell.2008.03.027.

Marisa, Laetitia, Aurélien de Reyniès, Alex Duval, Janick Selves, Marie Pierre Gaub, Laure Vescovo, Marie Christine Etienne-Grimaldi, et al. 2013. “Gene Expression Classification of Colon Cancer into Molecular Subtypes: Characterization, Validation, and Prognostic Value.” *PLoS Medicine* 10: e1001453.

https://doi.org/10.1371/journal.pmed.1001453.

Masqué-Soler, Neus, Monika Szczepanowski, Christian W Kohler, Rainer Spang, and Wolfram Klapper. 2013. "Molecular Classification of Mature Aggressive b-Cell Lymphoma Using Digital Multiplexed Gene Expression on Formalin-Fixed Paraffin-Embedded Biopsy Specimens." *Blood* 122: 1985–86. https://doi.org/https://doi.org/10.1182/blood-2013-06-508937.

McClure, Marni B., Yasunori Kogure, Naser Ansari-Pour, Yuki Saito, Hann-Hsiang Chao, Jonathan Shepherd, Mariko Tabata, et al. 2023. "Landscape of Genetic Alterations Underlying Hallmark Signature Changes in Cancer Reveals TP53 Aneuploidy–Driven Metabolic Reprogramming." *Cancer Research Communications* 3 (February): 281–96. https://doi.org/10.1158/2767-9764.crc-22-0073.

Melo, Felipe De Sousa E, Xin Wang, Marnix Jansen, Evelyn Fessler, Anne Trinh, Laura P. M. H. De Rooij, Joan H. De Jong, et al. 2013. "Poor-Prognosis Colon Cancer Is Defined by a Molecularly Distinct Subtype and Develops from Serrated Precursor Lesions." *Nature Medicine* 19: 614–18. https://doi.org/10.1038/nm.3174.

Meyers, Robin M., Jordan G. Bryan, James M. McFarland, Barbara A. Weir, Ann E. Sizemore, Han Xu, Neekesh V. Dharia, et al. 2017. "Computational Correction of Copy Number Effect Improves Specificity of CRISPR-Cas9 Essentiality Screens in Cancer Cells." *Nature Genetics* 49: 1779–84. https://doi.org/10.1038/ng.3984.

Miller, Lance D., Johanna Smeds, Joshy George, Vinsensius B. Vega, Liza Vergara, Alexander Ploner, Yudi Pawitan, et al. 2005. "An Expression Signature for P53 Status in Human Breast Cancer Predicts Mutation Status, Transcriptional Effects, and Patient Survival." *Proceedings of the National Academy of Sciences of the United States of America* 102: 13550–55. https://doi.org/10.1073/pnas.0506230102.

Mitri, Zahi I., Nour Abuhadra, Shaun M. Goodyear, Evthokia A. Hobbs, Andy Kaempf, Alastair M. Thompson, and Stacy L. Moulder. 2022. "Impact of TP53 Mutations in Triple Negative Breast Cancer." *Npj Precision Oncology* 6 (December). https://doi.org/10.1038/s41698-022-00303-6.

Mollaoglu, Gurkan, Matthew R. Guthrie, Stefanie Böhm, Johannes Brägelmann, Ismail Can, Paul M. Ballieu, Annika Marx, et al. 2017. "MYC Drives Progression of Small Cell Lung Cancer to a Variant Neuroendocrine Subtype with Vulnerability to Aurora Kinase Inhibition." *Cancer Cell* 31 (February): 270–85. https://doi.org/10.1016/j.ccell.2016.12.005.

Mordelet, F., and J. P. Vert. 2014. "A bagging SVM to learn from positive and unlabeled examples." *Pattern Recognition Letters* 37 (1): 201–9. https://doi.org/10.1016/j.patrec.2013.06.010.

Mounir, Mohamed, Marta Lucchetta, Tiago C. Silva, Catharina Olsen, Gianluca Bontempi, Xi Chen, Houtan Noushmehr, Antonio Colaprico, and Elena Papaleo. 2019. "New Functionalities in the TCGAbiolinks Package for the Study and Integration of Cancer Data from GDC and GTEx." Edited by Edwin Wang. *PLOS Computational Biology* 15 (3): e1006701. https://doi.org/10.1371/journal.pcbi.1006701.

Naidoo, Jarushka, Maria L. Santos-Zabala, Tunc Iyriboz, Kaitlin M. Woo, Camelia S. Sima, John J. Fiore, Mark G. Kris, et al. 2016. "Large Cell Neuroendocrine Carcinoma of the Lung: Clinico-Pathologic Features, Treatment, and Outcomes." *Clinical Lung Cancer* 17 (September): e121–29. https://doi.org/10.1016/j.cllc.2016.01.003.

Nicholson, Andrew G., Ming S. Tsao, Mary Beth Beasley, Alain C. Borczuk, Elisabeth Brambilla, Wendy A. Cooper, Sanja Dacic, et al. 2022. "The 2021 WHO Classification of Lung Tumors: Impact of Advances Since 2015." *Journal of Thoracic Oncology* 17 (3): 362–87. https://doi.org/10.1016/j.jtho.2021.11.003.

Ostano, Paola, Maurizia Mello-grand, Debora Sesia, Ilaria Gregnanin, Caterina Peraldo-neia, Francesca Guana, Elena Jachetti, Antonella Farsetti, and Giovanna Chiorino. 2020. "Gene Expression Signature Predictive of Neuroendocrine Transformation in Prostate Adenocarcinoma." *International Journal of Molecular Sciences* 21: 1078. https://doi.org/10.3390/ijms21031078.

Paik, Soonmyung, Steven Shak, Chungyeul Kim, Joffre Baker, Maureen Cronin, Frederick L Baehner, Michael G Walker, et al. 2004. "A Multigene Assay to Predict Recurrence of Tamoxifen-Treated, Node-Negative Breast Cancer." *The New England Journal of Medicine* 351: 2817–43. www.nejm.org.

Pal, Bhupinder, Yunshun Chen, François Vaillant, Bianca D Capaldo, Rachel Joyce, Xiaoyu Song, Vanessa L Bryant, et al. 2021. "A Single-cell RNA Expression Atlas of Normal, Preneoplastic and Tumorigenic States in the Human Breast." *The EMBO Journal* 40 (11): e107333. https://doi.org/https://doi.org/10.15252/embj.2020107333.

Perou, Charles M., Therese Sørile, Michael B. Eisen, Matt Van De Rijn, Stefanie S. Jeffrey, Christian A. Ress, Jonathan R. Pollack, et al. 2000. "Molecular portraits of human breast tumours." *Nature* 406 (6797): 747–52. https://doi.org/10.1038/35021093.

Phillips, Heidi S., Samir Kharbanda, Ruihuan Chen, William F. Forrest, Robert H. Soriano, Thomas D. Wu, Anjan Misra, et al. 2006. "Molecular Subclasses of High-Grade Glioma Predict Prognosis, Delineate a Pattern of Disease Progression, and Resemble Stages in Neurogenesis." *Cancer Cell* 9: 157–73. https://doi.org/10.1016/j.ccr.2006.02.019.

Piskol, Robert, Ling Huw, Ismail Sergin, Christiaan Kljin, Zora Modrusan, Doris Kim, Noelyn Kljavin, et al. 2019. "A Clinically Applicable Gene-Expression Classifier Reveals Intrinsic and Extrinsic Contributions to Consensus Molecular Subtypes in Primary and Metastatic Colon Cancer." *Clinical Cancer Research* 25: 4431–42. https://doi.org/10.1158/1078-0432.CCR-18-3032.

Prat, Aleix, and Charles M. Perou. 2011. "Deconstructing the Molecular Portraits of Breast Cancer." *Molecular Oncology*. John Wiley; Sons Ltd. https://doi.org/10.1016/j.molonc.2010.11.003.

Puram, V. Sidharth, Itay Tirosh, Anuraag S. Parikh, Anoop P. Patel, Keren Yizhak, Shawn Gillespie, Christopher Rodman, et al. 2017. "Single-Cell Transcriptomic Analysis of Primary and Metastatic Tumor Ecosystems in Head and Neck Cancer." *Cell* 171 (7): 1611–1624.e24. https://doi.org/10.1016/j.cell.2017.10.044.

Qian, Yuquan, Jimmy Daza, Timo Itzel, Johannes Betge, Tianzuo Zhan, Frederik Marmé, and Andreas Teufel. 2021. "Prognostic Cancer Gene Expression Signatures: Current Status and Challenges." *Cells*. MDPI. https://doi.org/10.3390/cells10030648.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Ragulan, Chanthirika, Katherine Eason, Elisa Fontana, Gift Nyamundanda, Noelia Tarazona, Yatish Patil, Pawan Poudel, et al. 2019. "Analytical Validation of Multiplex Biomarker Assay to Stratify Colorectal Cancer into Molecular Subtypes." *Scientific Reports* 9 (1): 1–12. https://doi.org/10.1038/s41598-019-43492-0.

Rahnenführer, Jörg, Francisco S. Domingues, Jochen Maydt, and Thomas Lengauer. 2004. "Calculating the Statistical Significance of Changes in Pathway Activity from Gene Expression Data." *Statistical Applications in Genetics and Molecular Biology* 3: 1–29. https://doi.org/10.2202/1544-6115.1055.

Raphael, Benjamin J., Ralph H. Hruban, Andrew J. Aguirre, Richard A. Moffitt, Jen Jen Yeh, Chip Stewart, A. Gordon Robertson, et al. 2017. "Integrated Genomic Characterization of Pancreatic Ductal Adenocarcinoma." *Cancer Cell* 32 (August): 185–203.e13. https://doi.org/10.1016/j.ccell.2017.07.007.

Rekhtman, Natasha. 2022. “Lung neuroendocrine neoplasms: recent progress and persistent challenges.” *Modern Pathology* 35 (July 2021): 36–50. https://doi.org/10.1038/s41379-021-00943-2.

Ritchie, Matthew E., Belinda Phipson, Di Wu, Yifang Hu, Charity W. Law, Wei Shi, and Gordon K. Smyth. 2015. “Limma Powers Differential Expression Analyses for RNA-Sequencing and Microarray Studies.” *Nucleic Acids Research* 43 (January): e47. https://doi.org/10.1093/nar/gkv007.

Robinson, Mark D., Davis J. McCarthy, and Gordon K. Smyth. 2009. “edgeR: A Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data.” *Bioinformatics* 26 (1): 139–40. https://doi.org/10.1093/bioinformatics/btp616.

Roepman, Paul, Andreas Schlicker, Josep Tabernero, Ian Majewski, Sun Tian, Victor Moreno, Mireille H. Snel, et al. 2014. “Colorectal Cancer Intrinsic Subtypes Predict Chemotherapy Benefit, Deficient Mismatch Repair and Epithelial-to-Mesenchymal Transition.” *International Journal of Cancer* 134: 552–62. https://doi.org/10.1002/ijc.28387.

Rossi, Giulio, Alberto Cavazza, Alessandro Marchioni, Lucia Longo, Mario Migaldi, Giuliana Sartori, Nazzarena Bigiani, et al. 2005. “Role of Chemotherapy and the Receptor Tyrosine Kinases KIT, PDGFR , PDGFR , and Met in Large-Cell Neuroendocrine Carcinoma of the Lung.” *Journal of Clinical Oncology* 23: 8774–85. https://doi.org/10.1200/JCO.2005.02.8233.

Saal, Lao H., Sofia K. Gruvberger-Saal, Camilla Persson, Kristina Lövgren, Mervi Jumppanen, Johan Staaf, Göran Jönsson, et al. 2008. “Recurrent Gross Mutations of the PTEN Tumor Suppressor Gene in Breast Cancers with Deficient DSB Repair.” *Nature Genetics* 40 (January): 102–7. https://doi.org/10.1038/ng.2007.39.

Sadanandam, Anguraj, Costas A. Lyssiotis, Krisztian Homicsko, Eric A. Collisson, William J. Gibb, Stephan Wullschleger, Liliane C.Gonzalez Ostos, et al. 2013. “A colorectal cancer classification system that associates cellular phenotype and responses to therapy.” *Nature Medicine* 19 (5): 619–25. https://doi.org/10.1038/nm.3175.

Salazar, Ramon, Paul Roepman, Gabriel Capella, Victor Moreno, Iris Simon, Christa Dreezen, Adriana Lopez-Doriga, et al. 2011. “Gene Expression Signature to Improve Prognosis Prediction of Stage II and III Colorectal Cancer.” *Journal of Clinical Oncology* 29 (January): 17–24. https://doi.org/10.1200/JCO.2010.30.1077.

Samoszuk, Michael, Jenny Tan, and Guillaume Chorn. 2005. “Clonogenic growth of human breast cancer cells co-cultured in direct contact with serum-activated fibroblasts." *Breast Cancer Research* 7 (3): R274. https://doi.org/10.1186/bcr995.

Sarkaria, Inderpal S, Akira Iyoda, Mee Soo Roh, Gabriel Sica, Deborah Kuk, Camelia S Sima, Maria C Pietanza, Bernard J Park, William D Travis, and Valerie W Rusch. 2011. "Neoadjuvant and Adjuvant Chemotherapy in Resected Pulmonary Large Cell Neuroendocrine Carcinomas: A Single Institution Experience." *The Annals of Thoracic Surgery* 92: 1180–87. https://doi.org/https://doi.org/10.1016/j.athoracsur.2011.05.027.

Satija, Rahul, Jeffrey A Farrell, David Gennert, Alexander F Schier, and Aviv Regev. 2015. "Spatial Reconstruction of Single-Cell Gene Expression Data." *Nature Biotechnology* 33: 495–502. https://doi.org/10.1038/nbt.3192.

Schlicker, Andreas, Garry Beran, Christine M. Chresta, Gael McWalter, Alison Pritchard, Susie Weston, Sarah Runswick, et al. 2012. "Subtypes of Primary Colorectal Tumors Correlate with Response to Targeted Treatment in Colorectal Cell Lines." *BMC Medical Genomics* 5: 1–15. https://doi.org/10.1186/1755-8794-5-66.

Schubert, Michael, Bertram Klinger, Martina Klünemann, Anja Sieber, Florian Uhlitz, Sascha Sauer, Mathew J. Garnett, Nils Blüthgen, and Julio Saez-Rodriguez. 2018. "Perturbation-Response Genes Reveal Signaling Footprints in Cancer Gene Expression." *Nature Communications* 9: 20. https://doi.org/10.1038/s41467-017-02391-6.

Sehgal, Manas, Soundharya Ramu, Joel Markus Vaz, Yogheshwer Raja Ganapathy, Srinath Muralidharan, Sankalpa Venkatraghavan, and Mohit Kumar Jolly. 2024. "Characterizing Heterogeneity Along EMT and Metabolic Axes in Colorectal Cancer Reveals Underlying Consensus Molecular Subtype-Specific Trends." *Translational Oncology* 40 (February): 101845. https://doi.org/10.1016/j.tranon.2023.101845.

Sherman, Shira, Ofer Rotem, Tzippy Shochat, Alona Zer, Assaf Moore, and Elizabeth Dudnik. 2020. "Efficacy of Immune Check-Point Inhibitors (ICPi) in Large Cell Neuroendocrine Tumors of Lung (LCNEC)." *Lung Cancer* 143 (May): 40–46. https://doi.org/10.1016/j.lungcan.2020.03.008.

Simbolo, Michele, Stefano Barbi, Matteo Fassan, Andrea Mafficini, Greta Ali, Caterina Vicentini, Nicola Sperandio, et al. 2019. "Gene Expression Profiling of Lung Atypical Carcinoids and Large Cell Neuroendocrine Carcinomas Identifies Three Transcriptomic Subtypes with Specific Genomic Alterations." *Journal of Thoracic Oncology* 14 (9): 1651–61. https://doi.org/10.1016/j.jtho.2019.05.003.

Sjoberg, Daniel D., Karissa Whiting, Michael Curry, Jessica A. Lavery, and Joseph Larmarange. 2021. “Reproducible Summary Tables with the Gtsummary Package.” *The R Journal* 13: 570–80. https://doi.org/10.32614/RJ-2021-053.

Staub, Eike. 2012. “An interferon response gene expression signature is activated in a subset of medulloblastomas.” *Translational Oncology* 5 (4): 297–304. https://doi.org/10.1593/tlo.12214.

Stuart, Tim, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M Mauck III, Yuhan Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija. 2019. “Comprehensive Integration of Single-Cell Data.” *Cell* 177: 1888–1902. https://doi.org/10.1016/j.cell.2019.05.031.

Subramanian, Aravind, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, et al. 2005. “Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles.” *Proceedings of the National Academy of Sciences of the United States of America* 102: 15545–50. https://doi.org/10.1073/pnas.0506580102.

Sun, Dongqing, Jin Wang, Ya Han, Xin Dong, Jun Ge, Rongbin Zheng, Xiaoying Shi, et al. 2021. “TISCH: A comprehensive web resource enabling interactive single-cell transcriptome visualization of tumor microenvironment.” *Nucleic Acids Research* 49 (D1): D1420–30. https://doi.org/10.1093/nar/gkaa1020.

Sun, Jong Mu, Myung Ju Ahn, Jin Seok Ahn, Sang Won Um, Hojoong Kim, Hong Kwan Kim, Young Soo Choi, et al. 2012. “Chemotherapy for Pulmonary Large Cell Neuroendocrine Carcinoma: Similar to That for Small Cell Lung Cancer or Non-Small Cell Lung Cancer?” *Lung Cancer* 77 (August): 365–70. https://doi.org/10.1016/j.lungcan.2012.04.009.

Sutherland, Tara E., Douglas P. Dyer, and Judith E. Allen. 2023. “The Extracellular Matrix and the Immune System: A Mutually Dependent Relationship.” *Science*. American Association for the Advancement of Science. https://doi.org/10.1126/science.abp8964.

Swarts, Dorian R. A., Frans C. S. Ramaekers, and Ernst J. M. Speel. 2015. “Gene Expression Profiling of Pulmonary Neuroendocrine Neoplasms: A Comprehensive Overview.” *Cancer Treatment Communications* 4: 148–60. https://doi.org/10.1016/j.ctrc.2015.09.002.

Tan, Tuan Zea, Qing Hao Miow, Yoshio Miki, Tetsuo Noda, Seiichi Mori, Ruby Yun-Ju Huang, and Jean Paul Thiery. 2014. “Epithelial-mesenchymal Transition Spectrum Quantification and Its Efficacy in Deciphering Survival and

Drug Responses of Cancer Patients." *EMBO Molecular Medicine* 6 (October): 1279–93. https://doi.org/10.15252/emmm.201404208.

Taube, Joseph H., Jason I. Herschkowitz, Kakajan Komurov, Alicia Y. Zhou, Supriya Gupta, Jing Yang, Kimberly Hartwell, et al. 2010. "Core Epithelial-to-Mesenchymal Transition Interactome Gene-Expression Signature Is Associated with Claudin-Low and Metaplastic Breast Cancer Subtypes." *Proceedings of the National Academy of Sciences of the United States of America* 107: 15449–54. https://doi.org/10.1073/pnas.1004900107.

Therneau, Terry M., and Patricia M. Grambsch. 2000. *Modeling Survival Data: Extending the Cox Model. Statistics for Biology and Health.* Springer New York. https://doi.org/10.1007/978-1-4757-3294-8.

Thorsson, Vésteinn, David L. Gibbs, Scott D. Brown, Denise Wolf, Dante S. Bortone, Tai Hsien Ou Yang, Eduard Porta-Pardo, et al. 2018. "The Immune Landscape of Cancer." *Immunity* 48 (4): 812–830.e14. https://doi.org/10.1016/j.immuni.2018.03.023.

Tomfohr, John, Jun Lu, and Thomas B. Kepler. 2005. "Pathway Level Analysis of Gene Expression Using Singular Value Decomposition." *BMC Bioinformatics* 6: 1–11. https://doi.org/10.1186/1471-2105-6-225.

Troester, Melissa A., Jason I. Herschkowitz, Daniel S. Oh, Xiaping He, Katherine A. Hoadley, Claire S. Barbier, and Charles M. Perou. 2006. "Gene Expression Patterns Associated with P53 Status in Breast Cancer." *BMC Cancer* 6: 1–13. https://doi.org/10.1186/1471-2407-6-276.

Tsai, Harrison K., Jonathan Lehrer, Mohammed Alshalalfa, Nicholas Erho, Elai Davicioni, and Tamara L. Lotan. 2017. "Gene expression signatures of neuroendocrine prostate cancer and primary small cell prostatic carcinoma." *BMC Cancer* 17 (1): 1–21. https://doi.org/10.1186/s12885-017-3729-z.

Tyler, Michael, and Itay Tirosh. 2021. "Decoupling Epithelial-Mesenchymal Transitions from Stromal Profiles by Integrative Expression Analysis." *Nature Communications* 12 (1): 1–13. https://doi.org/10.1038/s41467-021-22800-1.

Uhlitz, Florian, Philip Bischoff, Stefan Peidli, Anja Sieber, Alexandra Trinks, Mareen Lüthen, Benedikt Obermayer, et al. 2021. "Mitogen-activated Protein Kinase Activity Drives Cell Trajectories in Colorectal Cancer." *EMBO Molecular Medicine* 13 (October): e14123. https://doi.org/10.15252/emmm.202114123.

Ushey, Kevin, and Hadley Wickham. 2023. *Renv: Project Environments.*

Valkenburg, Kenneth C., Amber E. De Groot, and Kenneth J. Pienta. 2018.

"Targeting the Tumour Stroma to Improve Cancer Therapy." *Nature Reviews Clinical Oncology.* Nature Publishing Group. https://doi.org/10.1038/s41571-018-0007-1.

Velcheti, Vamsidhar, Xiaohan Hu, Bilal Piperdi, and Thomas Burke. 2021. "Real-World Outcomes of First-Line Pembrolizumab Plus Pemetrexed-Carboplatin for Metastatic Nonsquamous NSCLC at US Oncology Practices." *Scientific Reports* 11 (December): 9222. https://doi.org/10.1038/s41598-021-88453-8.

Venet, David, Jacques E. Dumont, and Vincent Detours. 2011. "Most Random Gene Expression Signatures Are Significantly Associated with Breast Cancer Outcome." *PLoS Computational Biology* 7: e1002240. https://doi.org/10.1371/journal.pcbi.1002240.

Verhaak, Roel G. W., Katherine A. Hoadley, Elizabeth Purdom, Victoria Wang, Yuan Qi, Matthew D. Wilkerson, C. Ryan Miller, et al. 2010. "Integrated Genomic Analysis Identifies Clinically Relevant Subtypes of Glioblastoma Characterized by Abnormalities in PDGFRA, IDH1, EGFR, and NF1." *Cancer Cell* 17: 98–110. https://doi.org/10.1016/j.ccr.2009.12.020.

Vilchez Mercedes, Samuel A., Federico Bocci, Herbert Levine, José N. Onuchic, Mohit Kumar Jolly, and Pak Kin Wong. 2021. "Decoding leader cells in collective cancer invasion." *Nature Reviews Cancer* 21 (9): 592–604. https://doi.org/10.1038/s41568-021-00376-8.

Wagener, Christoph, and Oliver Müller. 2010. *Molekulare Onkologie - Entstehung, Progression, Klinische Aspekte ; 95 Tabellen.* Stuttgart: Georg Thieme Verlag.

Walter, Vonn, Xiaoying Yin, Matthew D. Wilkerson, Christopher R. Cabanski, Ni Zhao, Ying Du, Mei Kim Ang, et al. 2013. "Molecular Subtypes in Head and Neck Cancer Exhibit Distinct Patterns of Chromosomal Gain and Loss of Canonical Cancer Genes." Edited by Muy-Teck Teh. *PLoS ONE* 8 (2): e56823. https://doi.org/10.1371/journal.pone.0056823.

Wang, Victoria E., Anatoly Urisman, Lee Albacker, Siraj Ali, Vincent Miller, Rahul Aggarwal, and David Jablons. 2017. "Checkpoint Inhibitor Is Active Against Large Cell Neuroendocrine Carcinoma with High Tumor Mutation Burden." *Journal for ImmunoTherapy of Cancer* 5 (September): 75. https://doi.org/10.1186/s40425-017-0281-y.

Wang, Yumeng, Xiaoyan Xu, Dejan Maglic, Michael T. Dill, Kamalika Mojumdar, Patrick Kwok Shing Ng, Kang Jin Jeong, et al. 2018. "Comprehensive Molecular Characterization of the Hippo Signaling Pathway in Cancer." *Cell Reports* 25: 1304–1317.e5. https://doi.org/10.1016/j.celrep.2018.10.001.

Weinstein, John N., Eric A. Collisson, Gordon B. Mills, Kenna R.Mills Shaw, Brad A. Ozenberger, Kyle Ellrott, Chris Sander, et al. 2013. "The Cancer Genome Atlas Pan-Cancer Analysis Project." *Nature Genetics* 45 (October): 1113–20. https://doi.org/10.1038/ng.2764.

Williams, Elizabeth D., Dingcheng Gao, Andrew Redfern, and Erik W. Thompson. 2019. "Controversies around epithelial–mesenchymal plasticity in cancer metastasis." *Nature Reviews Cancer* 19 (12): 716–32. https://doi.org/10.1038/s41568-019-0213-x.

Wilson, Molly M., Robert A. Weinberg, Jacqueline A. Lees, and Vincent J. Guen. 2020. "Emerging Mechanisms by Which EMT Programs Control Stemness." *Trends in Cancer* 6 (9): 775–80. https://doi.org/10.1016/j.trecan.2020.03.011.

Xu, Yuyan, Maoyuan Gong, Yue Wang, Yang Yang, Shu Liu, and Qibing Zeng. 2023. "Global Trends and Forecasts of Breast Cancer Incidence and Deaths." *Scientific Data* 10 (December): 334. https://doi.org/10.1038/s41597-023-02253-5.

Yang, Jing, Parker Antin, Geert Berx, Cédric Blanpain, Thomas Brabletz, Marianne Bronner, Kyra Campbell, et al. 2020. "Guidelines and Definitions for Research on Epithelial–Mesenchymal Transition." *Nature Reviews Molecular Cell Biology*. https://doi.org/10.1038/s41580-020-0237-9.

Yang, Lan, Ying Fan, and Hongyang Lu. 2022. "Pulmonary Large Cell Neuroendocrine Carcinoma." *Pathology and Oncology Research* 28 (October): 123. https://doi.org/10.3389/PORE.2022.1610730/BIBTEX.

Yang, Qiao, Zihan Xu, Xiewan Chen, Linpeng Zheng, Yongxin Yu, Xianlan Zhao, Mingjing Chen, Bangyu Luo, Jianmin Wang, and Jianguo Sun. 2019. "Clinicopathological Characteristics and Prognostic Factors of Pulmonary Large Cell Neuroendocrine Carcinoma: A Large Population-Based Analysis." *Thoracic Cancer* 10 (April): 751–60. https://doi.org/10.1111/1759-7714.12993.

Yi, Ming, Dwight V. Nissley, Frank McCormick, and Robert M. Stephens. 2020. "ssGSEA Score-Based Ras Dependency Indexes Derived from Gene Expression Data Reveal Potential Ras Addiction Mechanisms with Possible Clinical Implications." *Scientific Reports* 10 (December): 10258. https://doi.org/10.1038/s41598-020-66986-8.

Yoshimura, Masayo, Kurumi Seki, Andrey Bychkov, and Junya Fukuoka. 2021. "Molecular Pathology of Pulmonary Large Cell Neuroendocrine Carcinoma: Novel Concepts and Treatments." *Frontiers in Oncology* 11 (April). https://doi.org/10.3389/fonc.2021.671799.

Zhang, Lei, Ziyi Li, Katarzyna M. Skrzypczynska, Qiao Fang, Wei Zhang, Sarah A. O'Brien, Yao He, et al. 2020. "Single-Cell Analyses Inform Mechanisms of Myeloid-Targeted Therapies in Colon Cancer." *Cell* 181 (2): 442–459.e29. https://doi.org/10.1016/j.cell.2020.03.048.

Zhang, Wei, Luc Girard, Yu An Zhang, Tomohiro Haruki, Mahboubeh Papari-Zareei, Victor Stastny, Hans K. Ghayee, et al. 2018. "Small Cell Lung Cancer Tumors and Preclinical Models Display Heterogeneity of Neuroendocrine Phenotypes." *Translational Lung Cancer Research* 7 (February): 32–49. https://doi.org/10.21037/tlcr.2018.02.02.

Zhuo, Minglei, Yanfang Guan, Xue Yang, Lingzhi Hong, Yuqi Wang, Zhongwu Li, Runzhe Chen, et al. 2020. "The Prognostic and Therapeutic Role of Genomic Subtyping by Sequencing Tumor or Cell-Free DNA in Pulmonary Large-Cell Neuroendocrine Carcinoma." *Clinical Cancer Research* 26 (February): 892–901. https://doi.org/10.1158/1078-0432.CCR-19-0556.

Zien, Alexander, Robert Kiiffner, Ralf Zimmer, and Thomas Lengauer. 2000. "Analysis of Gene Expression Data with Pathway Scores." *Proceedings. International Conference on Intelligent Systems for Molecular Biology* 8: 407–17. www.aaai.org.

Author’s Publications

Manuscript that deals with the content of the 4th chapter (Chapter 4):

Julian Kreis, Benedikt Brors, Eike Staub (2024). Expression profiling in a real-world, non-small cell lung cancer cohort reveals severe underdiagnosis of large cell neuroendocrine carcinomas. *Finished manuscript, submission planned for April 2024.*

Journal article for the 3rd chapter (Chapter 3):

Julian Kreis, Bogac Aybey, Felix Geist, Benedikt Brors, Eike Staub (2024). Stromal signals dominate gene expression signature scores that aim to describe cancer cell-intrinsic stemness or mesenchymality characteristics. *Cancer Research Communications, 4*(2), 516-526

Journal article for the 2nd chapter (Chapter 2):

Julian Kreis, Boro Nedic, Johanna Mazuret et al. (2021). RosettaSX: Reliable gene expression signature scoring of cancer models and patients. *Neoplasia (United States), 23*(11), 1069-1077

Journal articles with the author’s contribution

Eike-Benjamin Braune, Felix Geist, Xiaojia Tang, et al. (2024). Identification of a Notch transcriptomic signature for breast cancer. *Breast Cancer Research, 26*(1), 1-22

Hanrui Zhang, Julian Kreis, Sven Eric Schelhorn et al. (2021). Mapping combinatorial drug effects to DNA damage response kinase inhibitors. *Nature Communications, 14*(1), 1-8

Krzysztof Koras, Dilafruz Juraeva, Julian Kreis et al. (2020). Feature selection strategies for drug sensitivity prediction. *Nature Research, 1*(10), 9377

Poster Presentations

Julian Kreis, Dilafruz Juraeva, Eike Staub, Benedikt Brors (2019). Pathway Activation Prediction using Signed Random Walk with Restart *Gernam Conference on Bioinformatics, German Cancer Research Center*

Julian Kreis, Benedikt Brors, Dilafruz Juraeva, Eike Staub (2019). The Use of Genome-Scale Data for the Prediction of Pathway Activity, and Drug Efficacy *ASCONA Workshop-Statistical Challenges in Medical Data Science, Asconam CH*

Appendix A

Epithelial-to-mesenchymal

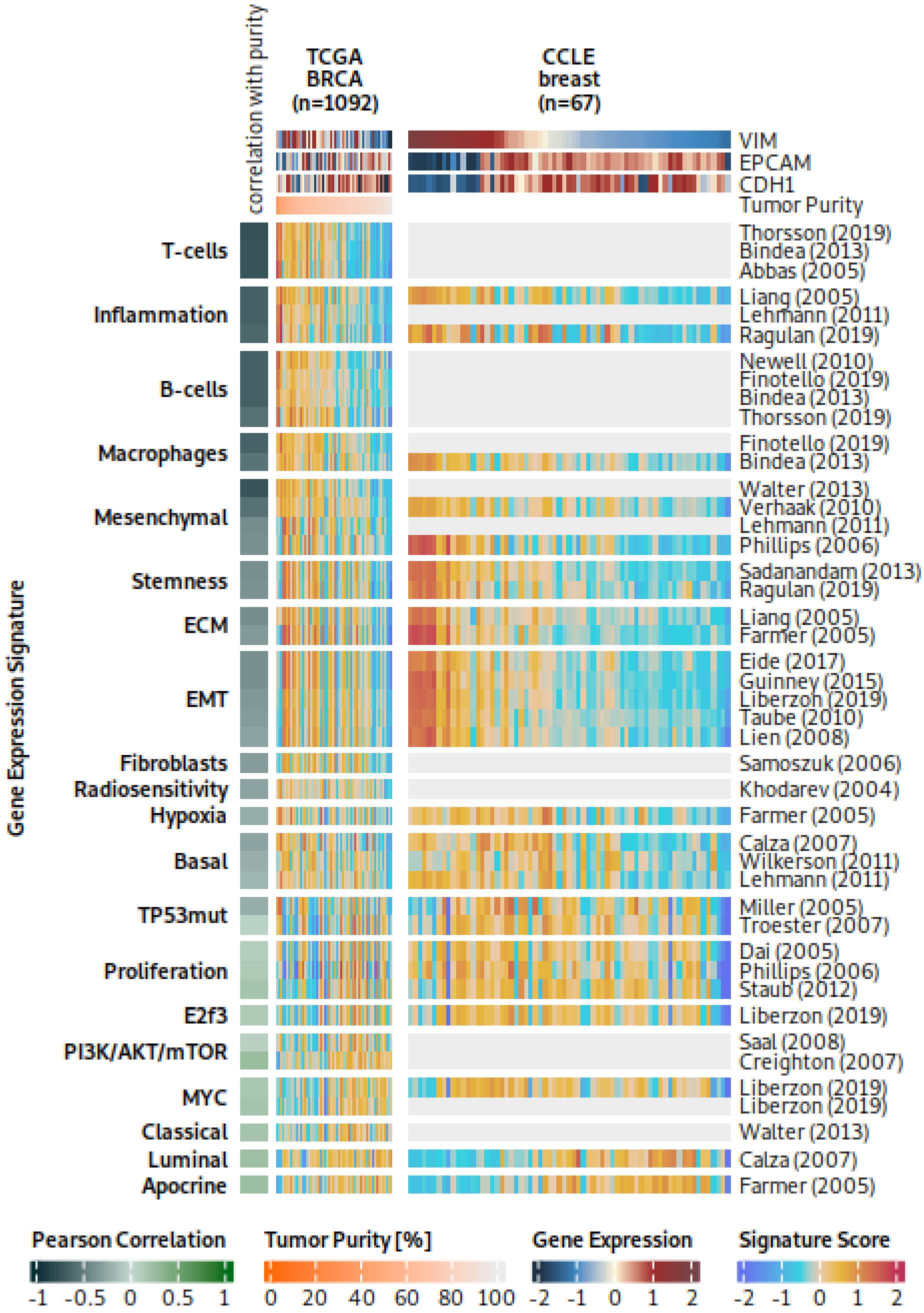

Figure A.1: RosettaSX analysis for TCGA BRCA samples and CCLE breast cancer cell lines. Please see Figure 3.3 for a detailed description of color codes and analysis details.

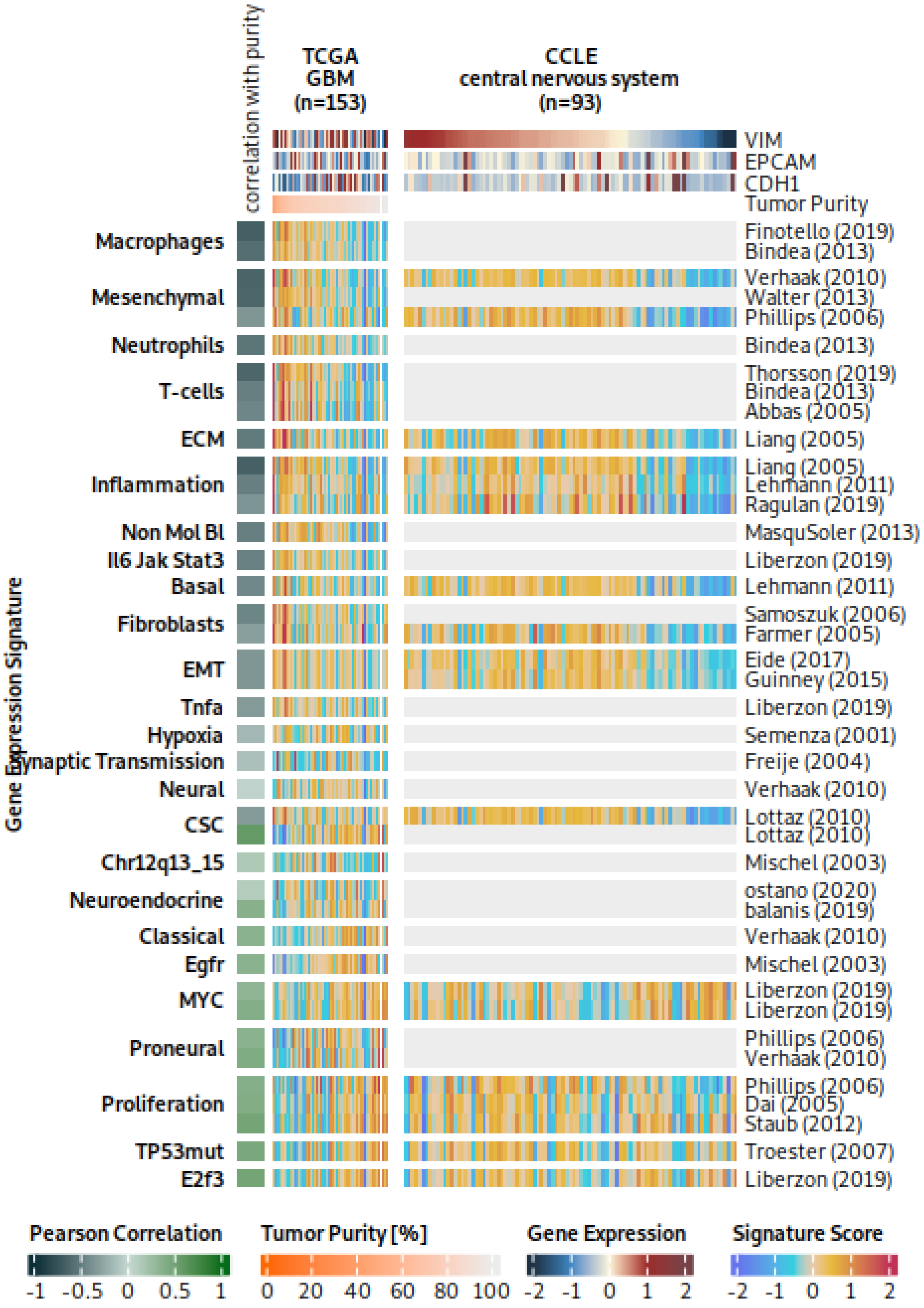

Figure A.2: RosettaSX analysis for TCGA GBM samples and CCLE central nervous system cancer cell lines. Please see Figure 3.3 for a detailed description of color codes and analysis details.

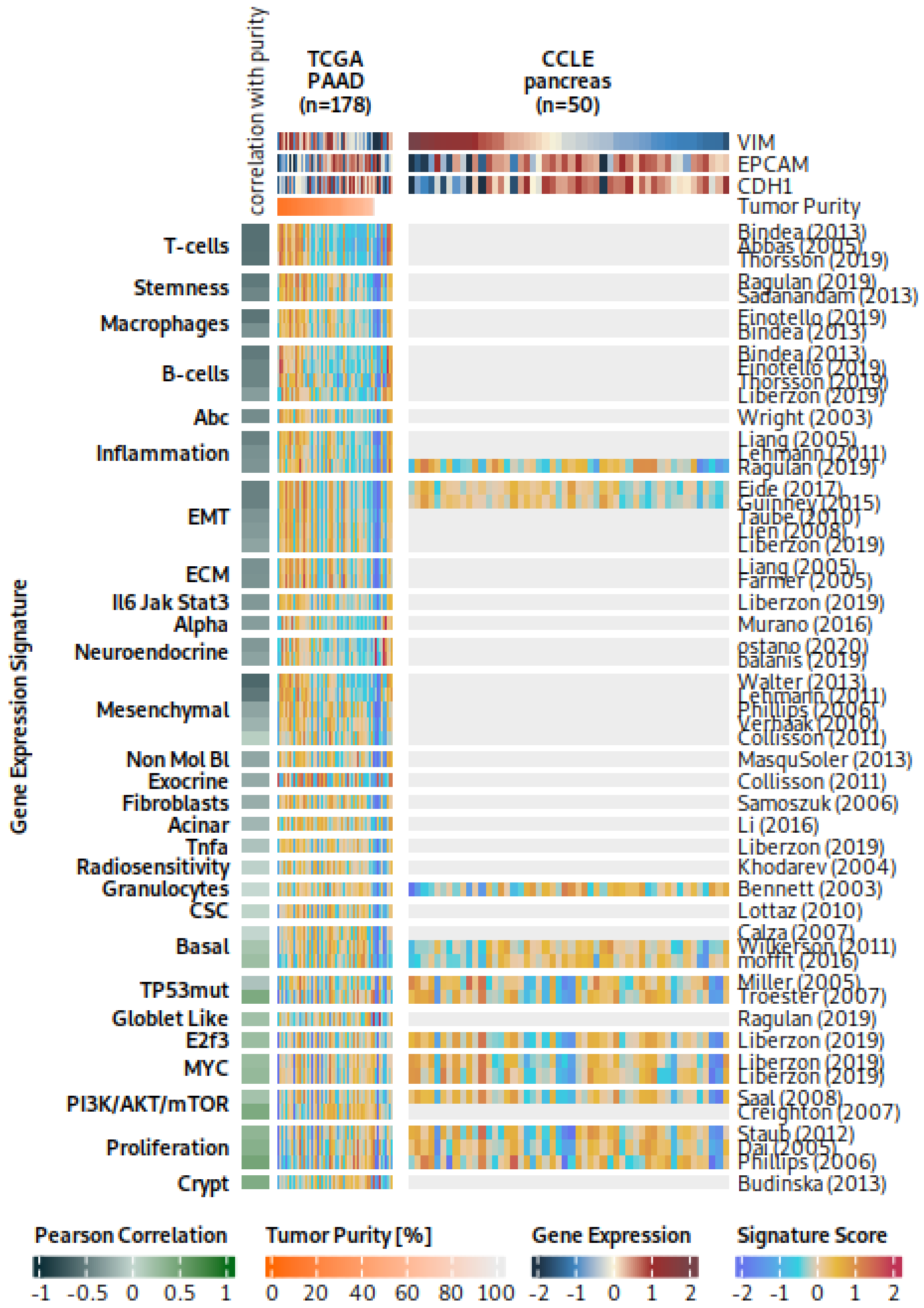

Figure A.3: RosettaSX analysis for TCGA PAAD and CCLE cancer cell lines. Please see Figure 3.3 for a detailed description of color codes and analysis details.

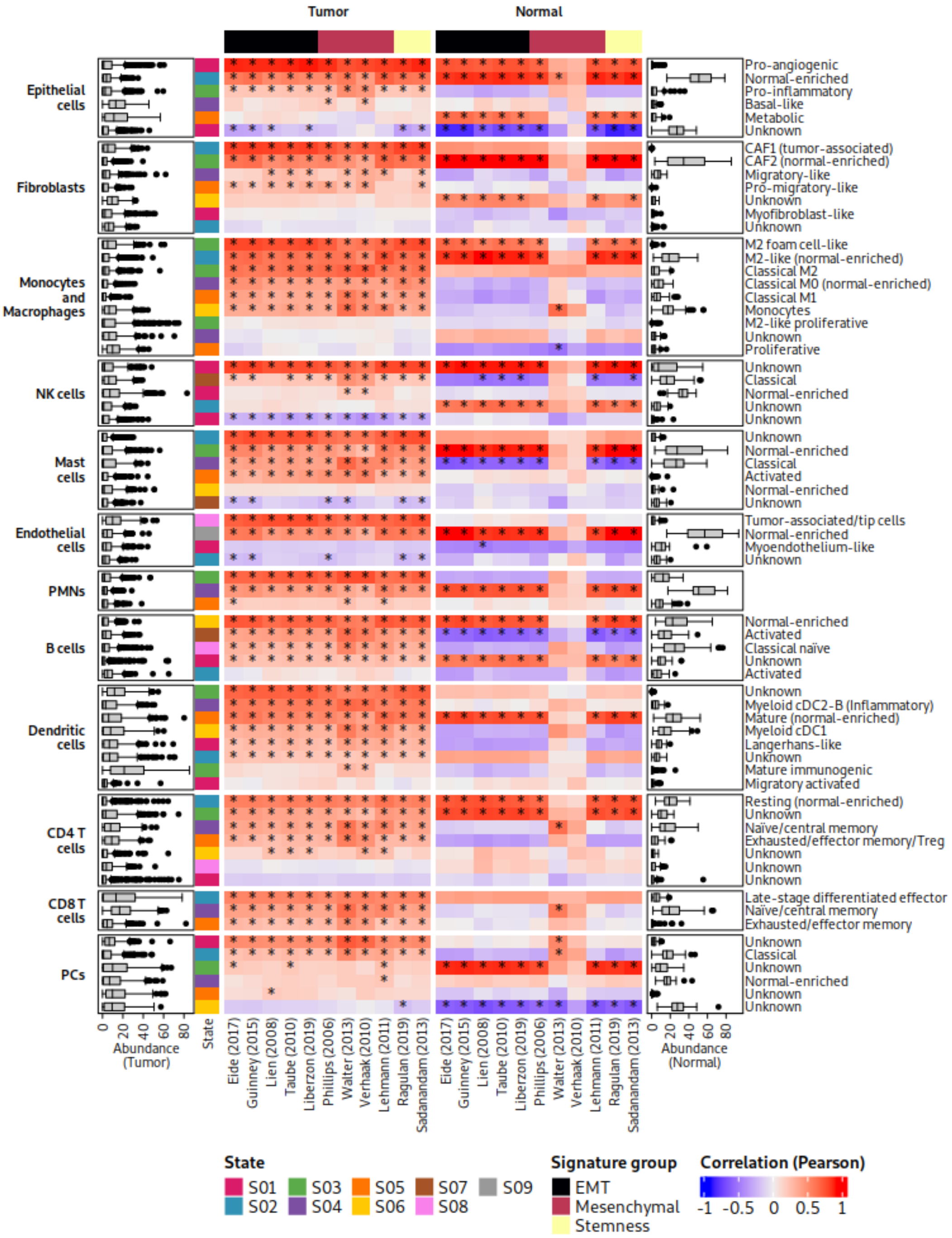

Figure A.4: Deconvolution analysis of TCGA CRC tumor and NAT tissue. For details, see Figure 3.7.

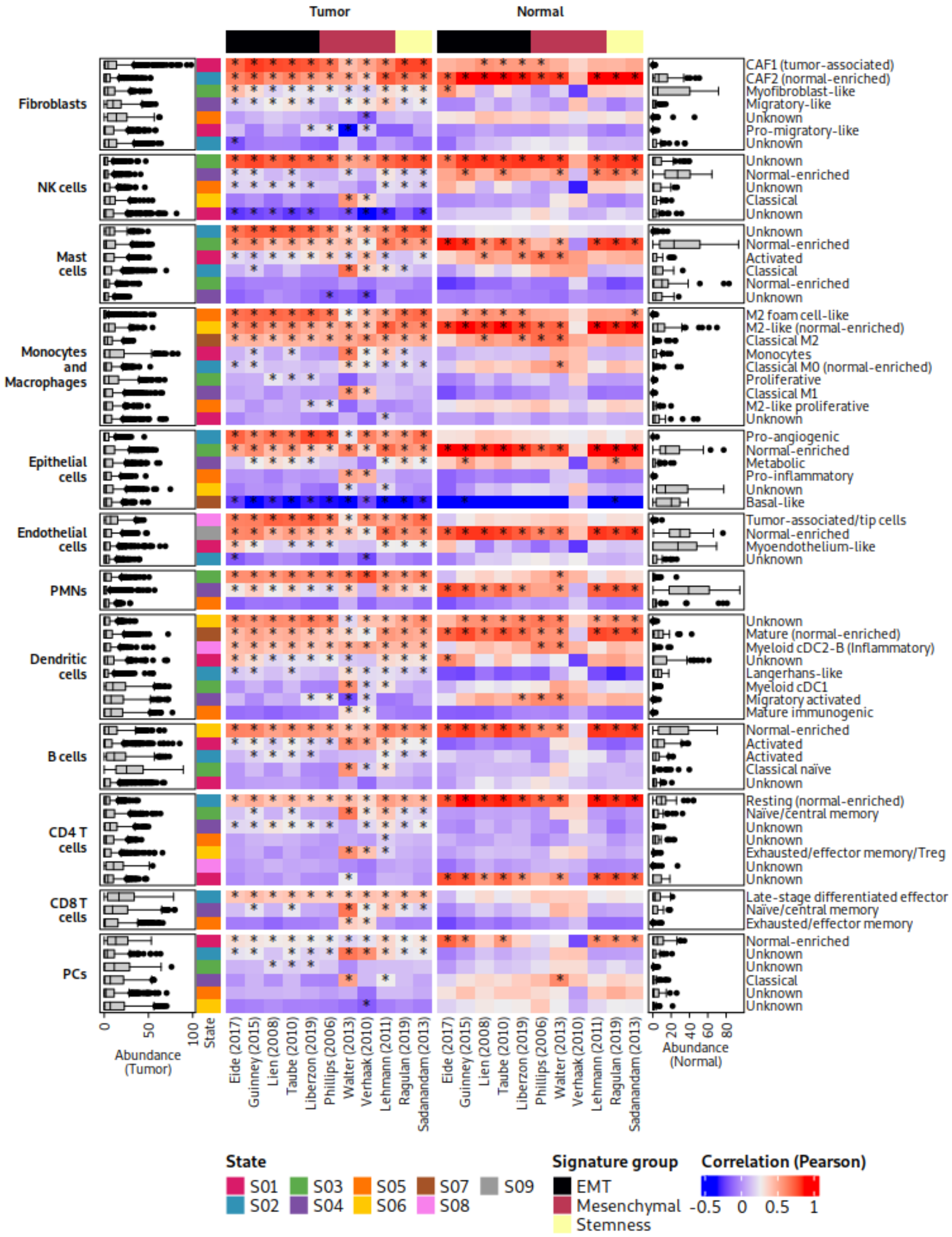

Figure A.5: Deconvolution analysis of TCGA HNSC tumor and NAT tissue. For details, see Figure 3.7.

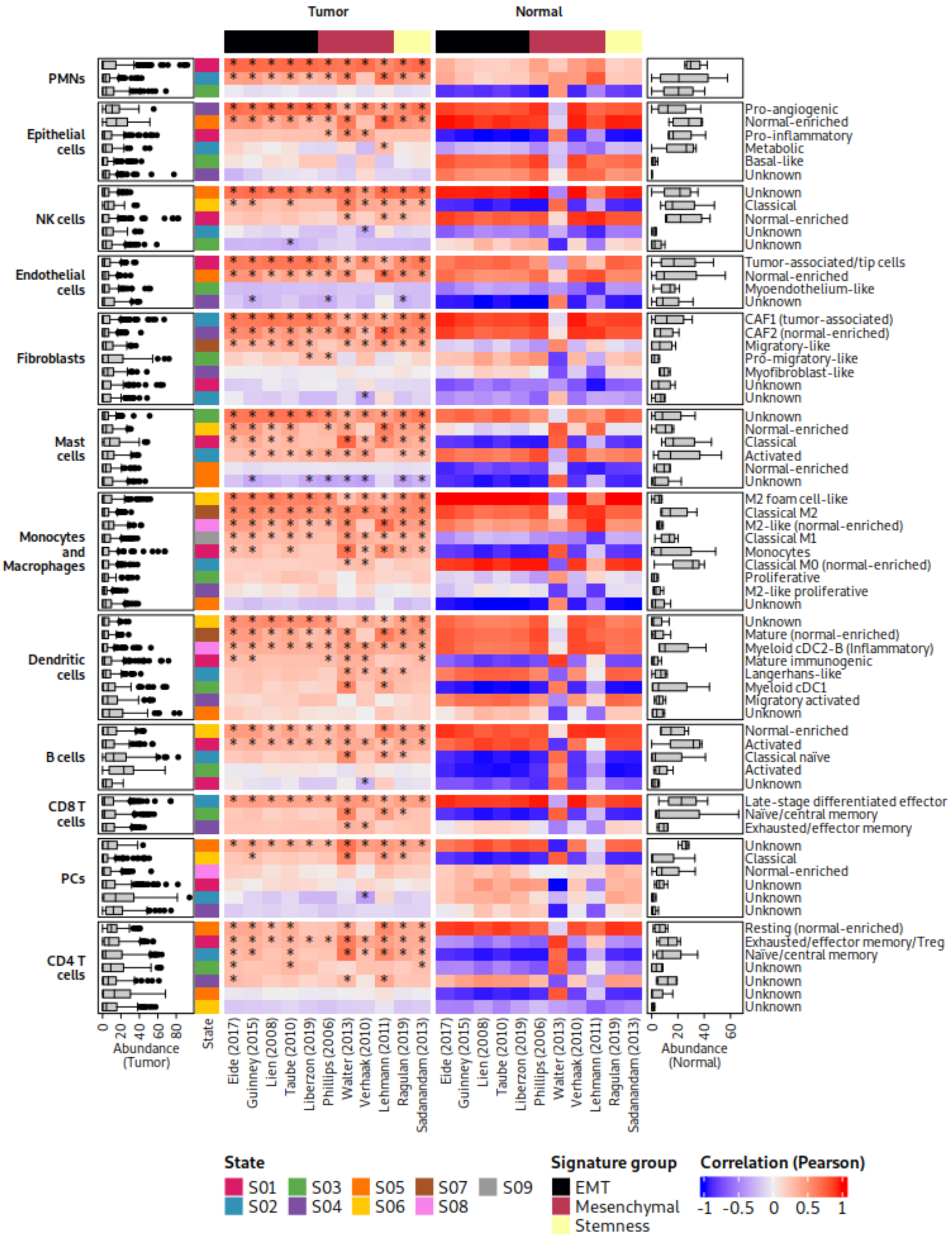

Figure A.6: Deconvolution analysis of TCGA PAAD tumor and NAT tissue. For details, see Figure 3.7.

Appendix B

Reproducibility

I implemented all analyses in R. The RosettaSX platform is accessible via www.rosettasx.com. Additionally, scripts and reproducible environments (using renv, Ushey and Wickham 2023) for the individual projects are available upon request.

B.1 R Session Information

Quarto Project of The Thesis

R version 4.1.1 (2021-08-10)

Platform: x86_64-pc-linux-gnu (64-bit)

locale: *LC_CTYPE=C.UTF-8*, *LC_NUMERIC=C*, *LC_TIME=C.UTF-8*, *LC_COLLATE=C.UTF-8*, *LC_MONETARY=C.UTF-8*, *LC_MESSAGES=C.UTF-8*, *LC_PAPER=C.UTF-8*, *LC_NAME=C*, *LC_ADDRESS=C*, *LC_TELEPHONE=C*, *LC_MEASUREMENT=C.UTF-8* and *LC_IDENTIFICATION=C*

attached base packages: *stats*, *graphics*, *grDevices*, *datasets*, *utils*, *methods* and *base*

loaded via a namespace (and not attached): *Rcpp(v.1.0.12)*, *digest(v.0.6.34)*, *jsonlite(v.1.8.8)*, *evaluate(v.0.23)*, *rlang(v.1.1.2)*, *cli(v.3.6.2)*, *renv(v.1.0.3)*, *rstudioapi(v.0.15.0)*, *rmarkdown(v.2.25)*, *tools(v.4.1.1)*, *pander(v.0.6.5)*, *xfun(v.0.41)*, *yaml(v.2.3.8)*, *fastmap(v.1.1.1)*, *compiler(v.4.1.1)*, *BiocManager(v.1.30.22)*, *htmltools(v.0.5.7)* and *knitr(v.1.45)*

Chapter 2 - RosettaSX Server

R version 3.6.3 (2020-02-29)

Platform: x86_64-pc-linux-gnu (64-bit)

locale: *LC_CTYPE=de_DE.utf8*, *LC_NUMERIC=C*, *LC_TIME=de_DE.utf8*, *LC_COLLATE=de_DE.utf8*, *LC_MONETARY=de_DE.utf8*, *LC_MESSAGES=de_DE.utf8*, *LC_PAPER=de_DE.utf8*, *LC_NAME=de_DE.utf8*, *LC_ADDRESS=de_DE.utf8*, *LC_TELEPHONE=de_DE.utf8*, *LC_MEASUREMENT=de_DE.utf8* and *LC_IDENTIFICATION=de_DE.utf8*

attached base packages: *stats*, *graphics*, *grDevices*, *utils*, *datasets*, *methods* and *base*

other attached packages: shiny(v.1.5.0)

loaded via a namespace (and not attached): *tidyr(v.1.0.0)*, *jsonlite(v.1.6)*, *foreach(v.1.4.7)*, *RhpcBLASctl(v.0.20-137)*, *R.utils(v.2.9.2)*, *bupaR(v.0.4.4)*, *rintrojs(v.0.2.2)*, *pander(v.0.6.5)*, *xlsxjars(v.0.6.1)*, *shinyhelper(v.0.3.2)*, *xopmodules(v.0.6.7)*, *yaml(v.2.2.0)*, *pillar(v.1.8.1)*, *glue(v.1.4.0)*, *digest(v.0.6.23)*, *xopdata(v.1.0.1)*, *RColorBrewer(v.1.1-2)*, *promises(v.1.1.0)*, *colorspace(v.1.4-1)*, *plyr(v.1.8.5)*, *shinycssloaders(v.0.2.0)*, *htmltools(v.0.5.4)*, *httpuv(v.1.5.2)*, *R.oo(v.1.23.0)*, *pkgconfig(v.2.0.3)*, *GetoptLong(v.0.1.7)*, *purrr(v.0.3.3)*, *xtable(v.1.8-4)*, *scales(v.1.1.0)*, *later(v.1.0.0)*, *tibble(v.3.1.8)*, *ggplot2(v.3.3.2)*, *generics(v.0.0.2)*, *DT(v.0.13)*, *withr(v.2.1.2)*, *shinyjs(v.1.1)*, *fst(v.0.9.2)*, *cli(v.3.6.0)*, *magrittr(v.1.5)*, *mime(v.0.8)*, *R.methodsS3(v.1.7.1)*, *fansi(v.0.4.0)*, *xoprosettasxshinyapp(v.0.3.4)*, *doParallel(v.1.0.15)*, *forcats(v.0.4.0)*, *shinycustomloader(v.0.9.0)*, *shinydashboard(v.0.7.1)*, *tools(v.3.6.3)*, *data.table(v.1.12.8)*, *GlobalOptions(v.0.1.1)*, *lifecycle(v.1.0.3)*, *ComplexHeatmap(v.2.3.1)*, *stringr(v.1.4.0)*, *xlsx(v.0.6.3)*, *munsell(v.0.5.0)*, *shinyEventLogger(v.0.1.1)*, *cluster(v.2.1.4)*, *eventdataR(v.0.2.0)*, *compiler(v.3.6.3)*, *rlang(v.1.0.6)*, *shinyjqui(v.0.3.2)*, *grid(v.3.6.3)*, *shinydashboardPlus(v.0.7.0)*, *iterators(v.1.0.12)*, *xopcode(v.0.1.19)*, *rjson(v.0.2.20)*, *htmlwidgets(v.1.5.1)*, *circlize(v.0.4.8)*, *miniUI(v.0.1.1.1)*, *shinyWidgets(v.0.5.1)*, *gtable(v.0.3.0)*, *codetools(v.0.2-19)*, *mongolite(v.2.2.0)*, *R6(v.2.4.1)*, *sparkline(v.2.0)*, *lubridate(v.1.7.9)*, *knitr(v.1.26)*, *dplyr(v.1.1.0)*, *fastmap(v.1.1.0)*, *utf8(v.1.1.4)*, *clue(v.0.3-57)*, *shape(v.1.4.4)*, *rJava(v.0.9-11)*, *stringi(v.1.4.3)*, *parallel(v.3.6.3)*, *Rcpp(v.1.0.3)*, *vctrs(v.0.5.2)*, *png(v.0.1-7)*, *xfun(v.0.11)* and *tidyselect(v.1.2.0)*

Chapter 3

R version 4.1.1 (2021-08-10)

Platform: x86_64-pc-linux-gnu (64-bit)

locale: *LC_CTYPE=C.UTF-8*, *LC_NUMERIC=C*, *LC_TIME=C.UTF-8*, *LC_COLLATE=C.UTF-8*, *LC_MONETARY=C.UTF-8*, *LC_MESSAGES=C.UTF-8*, *LC_PAPER=C.UTF-8*, *LC_NAME=C*, *LC_ADDRESS=C*, *LC_TELEPHONE=C*, *LC_MEASUREMENT=C.UTF-8* and *LC_IDENTIFICATION=C*

attached base packages: *stats*, *graphics*, *grDevices*, *datasets*, *utils*, *methods* and *base*

loaded via a namespace (and not attached): *Rcpp(v.1.0.10)*, *ps(v.1.6.0)*, *digest(v.0.6.29)*, *later(v.1.3.0)*, *R6(v.2.5.1)*, *jsonlite(v.1.8.4)*, *evaluate(v.0.15)*, *rlang(v.1.0.6)*, *cli(v.3.6.0)*, *renv(v.1.0.1)*, *rstudioapi(v.0.13)*, *rmarkdown(v.2.20)*, *tools(v.4.1.1)*, *pander(v.0.6.5)*, *xfun(v.0.36)*, *fastmap(v.1.1.0)*, *yaml(v.2.3.5)*, *compiler(v.4.1.1)*, *processx(v.3.5.3)*, *BiocManager(v.1.30.19)*, *htmltools(v.0.5.2)*, *knitr(v.1.38)* and *quarto(v.1.2)*

Chapter 4

R version 4.1.1 (2021-08-10)

Platform: x86_64-pc-linux-gnu (64-bit)

locale: *LC_CTYPE=C.UTF-8*, *LC_NUMERIC=C*, *LC_TIME=C.UTF-8*, *LC_COLLATE=C.UTF-8*, *LC_MONETARY=C.UTF-8*, *LC_MESSAGES=C.UTF-8*, *LC_PAPER=C.UTF-8*, *LC_NAME=C*, *LC_ADDRESS=C*, *LC_TELEPHONE=C*, *LC_MEASUREMENT=C.UTF-8* and *LC_IDENTIFICATION=C*

attached base packages: *stats*, *graphics*, *grDevices*, *datasets*, *utils*, *methods* and *base*

loaded via a namespace (and not attached): *backports(v.1.4.1)*, *circlize(v.0.4.15)*, *xopdata(v.1.2.2)*, *workflows(v.1.1.3)*, *igraph(v.1.6.0)*, *splines(v.4.1.1)*, *RApiSerialize(v.0.1.2)*, *listenv(v.0.9.0)*, *ggplot2(v.3.4.4)*, *digest(v.0.6.33)*, *yardstick(v.1.2.0)*, *foreach(v.1.5.2)*, *htmltools(v.0.5.7)*, *targets(v.1.4.1)*, *parsnip(v.1.1.1)*, *fansi(v.1.0.6)*, *magrittr(v.2.0.3)*, *tune(v.1.1.2)*, *gtsummary(v.1.7.2)*, *base64url(v.1.4)*, *cluster(v.2.1.6)*, *doParallel(v.1.0.17)*, *tzdb(v.0.4.0)*, *readr(v.2.1.4)*, *recipes(v.1.0.9)*, *ComplexHeatmap(v.2.10.0)*, *globals(v.0.16.2)*, *gower(v.1.0.1)*, *RcppParallel(v.5.1.7)*, *matrixStats(v.1.2.0)*, *R.utils(v.2.12.3)*, *hardhat(v.1.3.0)*, *timechange(v.0.2.0)*, *rsample(v.1.2.0)*, *dials(v.1.2.0)*, *colorspace(v.2.1-0)*, *xfun(v.0.41)*, *dplyr(v.1.1.4)*, *callr(v.3.7.3)*, *crayon(v.1.5.2)*, *probably(v.1.0.2)*, *stringfish(v.0.16.0)*, *survival(v.3.5-7)*, *zoo(v.1.8-12)*, *iterators(v.1.0.14)*, *glue(v.1.6.2)*, *survminer(v.0.4.9)*, *gtable(v.0.3.4)*, *ipred(v.0.9-14)*, *GetoptLong(v.1.0.5)*, *car(v.3.1-2)*, *future.apply(v.1.11.0)*, *shape(v.1.4.6)*, *BiocGenerics(v.0.40.0)*, *abind(v.1.4-5)*, *scales(v.1.3.0)*, *rstatix(v.0.7.2)*, *Rcpp(v.1.0.11)*, *xtable(v.1.8-4)*, *clue(v.0.3-65)*, *GPfit(v.1.0-8)*, *bit(v.4.0.5)*, *km.ci(v.0.5-6)*, *stats4(v.4.1.1)*, *lava(v.1.7.3)*, *prodlim(v.2023.08.28)*, *fstcore(v.0.9.18)*, *RColorBrewer(v.1.1-3)*, *pkgconfig(v.2.0.3)*, *R.methodsS3(v.1.8.2)*, *nnet(v.7.3-19)*, *utf8(v.1.2.4)*, *here(v.1.0.1)*, *tidyselect(v.1.2.0)*, *rlang(v.1.1.2)*, *DiceDesign(v.1.10)*, *munsell(v.0.5.0)*, *tools(v.4.1.1)*, *cli(v.3.6.2)*, *generics(v.0.1.3)*, *broom(v.1.0.5)*, *ggdendro(v.0.1.23)*, *stringr(v.1.5.1)*, *fastmap(v.1.1.1)*, *yaml(v.2.3.8)*, *processx(v.3.8.3)*, *RhpcBLASctl(v.0.23-42)*, *knitr(v.1.45)*, *bit64(v.4.0.5)*, *pander(v.0.6.5)*, *survMisc(v.0.5.6)*, *purrr(v.1.0.2)*, *future(v.1.33.0)*, *R.oo(v.1.25.0)*, *arrow(v.14.0.0.1)*, *xml2(v.1.3.6)*, *brio(v.1.1.4)*, *compiler(v.4.1.1)*, *rstudioapi(v.0.15.0)*, *png(v.0.1-8)*, *testthat(v.3.2.1)*, *ggsignif(v.0.6.4)*, *gt(v.0.10.0)*, *chisq.posthoc.test(v.0.1.2)*, *tibble(v.3.2.1)*, *lhs(v.1.1.6)*, *broom.helpers(v.1.14.0)*, *stringi(v.1.7.6)*, *ps(v.1.7.5)*, *forcats(v.1.0.0)*, *lattice(v.0.22-5)*, *Matrix(v.1.6-2)*, *KMsurv(v.0.1-5)*, *vctrs(v.0.6.5)*, *furrr(v.0.3.1)*, *pillar(v.1.9.0)*, *lifecycle(v.1.0.4)*, *BiocManager(v.1.30.22)*, *GlobalOptions(v.0.1.2)*, *data.table(v.1.14.10)*, *R6(v.2.5.1)*, *qs(v.0.26.1)*, *renv(v.1.0.3)*, *gridExtra(v.2.3)*, *IRanges(v.2.28.0)*, *parallelly(v.1.36.0)*, *codetools(v.0.2-19)*, *MASS(v.7.3-60)*, *assertthat(v.0.2.1)*, *rprojroot(v.2.0.4)*, *rjson(v.0.2.21)*, *withr(v.2.5.2)*, *S4Vectors(v.0.32.4)*, *hms(v.1.1.3)*, *parallel(v.4.1.1)*, *fst(v.0.9.8)*, *grid(v.4.1.1)*, *rpart(v.4.1.23)*, *timeDate(v.4032.109)*, *tidyr(v.1.3.0)*, *class(v.7.3-22)*, *carData(v.3.0-5)*, *ggpubr(v.0.6.0)* and *lubridate(v.1.9.3)*

Appendix C

RosettaSX Platform

C.1 Gene expression signature collection

Table C.1: Listing of gene expression signatures that are used by the RosettaSX approach

Geneset Name	No. Genes	PubMed ID
classical_PCA_Collisson2011_21460848	22	21460848
claudin_low_down_BRCA_Prat2013_20813035	356	20813035
FA_down_Gene2013_24036430	179	24036430
classical_HNCA_Walter2013_23451093	62	23451093
EMTspheres_CRC_up_Hwang2011_21640118	50	21640118
bcell_Bindea2013_24138885	35	24138885
RS_down_Kim2013_22846430	21	22846430
hallmark50_apical_junction_Liberzon2019_26771021	200	26771021
neutrophils_Bindea2013_24138885	31	24138885
crypt_markers_CRC_Budinska2013_23836465	16	23836465
stemness_RamalhoSantos2002_12228720	202	12228720
TP53mut_up_BRCA_Miller2005_16141321	5	16141321
HC1B_progGroup_GBM_Freije2004_15374961	9	15374961
basal_subtype_LSCC_Wilkerson2011_20643781	17	20643781
hallmark50_inflammation_response_Liberzon2019_26771021	200	26771021
ERBB2_subtype_BRCA_Calza2007_16846532	7	16846532
M5_12_IFN_Chaussabel2008_18631455	58	18631455
polypeptide_PDAC_normal_Enge2017_28965763	1	28965763
CCS2_MSI_CRC_DeSousaEMelo2013_23584090	45	23584090
EGFR_Mischel2003_12700671	16	12700671
chr17q21_32_amplicon_cluster5_BRCA_Farmer2005_15897907	17	15897907
serrated_adenoma_CRC_ConesaZamora2013_22696308	9	22696308
TA_Sadanandam2013_23584089	183	23584089
CellCycle_BRCA_Dai2005_15899795	33	15899795
immune_responce_CRC_Budinska2013_23836465	102	23836465
luminal_cluster6_BRCA_Farmer2005_15897907	16	15897907
hallmark50_xenobiotic_metabolism_Liberzon2019_26771021	200	26771021

MSI_up_CRC_Jorissen2009_19088021	192	19088021
clusterC_MB_Kool2008_18769486	136	18769486
AKT_Creighton2007_17213801	57	17213801
BRAFlikeness_up_Popovici2012_22393095	31	22393095
HIF1A_targets_Semenza2001_11248550	36	11248550
IFNg_Dummer2020_32007138	5	32007138
hippo_up_YAP_transfection_up_Zhang2008_18413746	40	18413746
LAR_refined_BRCA_Lehmann2011_21633166	233	21633166
chr12q13_15_Mischel2003_12700671	17	12700671
TA_Ragulan2019_31113981	7	31113981
MSI_down_CRC_Jorissen2009_19088021	182	19088021
WNT_NanoStr_MB_Northcott2013_22057785	5	22057785
MTOR_PI3K_S6K_inhib_Heinonen2008_18652687	16	18652687
hypoxia_GBM_Liang2005_15827123	21	15827123
MSI_down_CRC_Watanabe2006_17047040	57	17047040
hallmark50_MTORC1_Liberzon2019_26771021	200	26771021
hallmark50_hypoxia_Liberzon2019_26771021	200	26771021
LSCC_vs_LAD_Kuner2009_18486272	10	18486272
tcell_Bindea2013_24138885	19	24138885
hallmark50_faty_acid_metabolism_Liberzon2019_26771021	158	26771021
ERBB2_down_BRCA_Bertucci2004_14743203	5	14743203
EMT_down_Taube2010_20713713	152	20713713
MSI_CRC_Watanabe2006_17047040	29	17047040
NKcell_monocytes_Heise2014_thesis	18	thesis
IFN_Staub2015_internal	7	internal
hallmark50_MYC_targets1_Liberzon2019_26771021	200	26771021
beta_PDAC_normal_Enge2017_28965763	1	28965763
polypeptide_PDAC_normal_Murano2016_27693023	18	27693023
hallmark50_G2M_Liberzon2019_26771021	200	26771021
HC2B_progGroup_GBM_Freije2004_15374961	14	15374961
mesenchymal_GBM_Phillips2006_16530701	15	16530701
HRD_Peng2015_24553445	223	24553445
WNT_DeSousaEMelo2012_22056143	55	22056143
metastasis_BRCA_lung_Minn2005_16049480	54	16049480
hallmark50_complement_Liberzon2019_26771021	200	26771021
mesenchymal_up_GBM_Verhaak2010_20129251	164	20129251
stroma_metagene_BRCA_Farmer2005_19122658	49	19122658
EMTspheres_CRC_down_Hwang2011_21640118	57	21640118
BL1_refined_BRCA_Lehmann2011_21633166	28	21633166
gem_pdac_organoid_tiriac2018_29853643	130	29853643
NE_top50_weight_pan_balanis2019_31287989	50	31287989
5fu_pdac_organoid_tiriac2018_29853643	62	29853643
BRCAness_high_Konstantinopoulos2010_20547991	32	20547991
mesenchymal_down_GBM_Verhaak2010_20129251	45	20129251
ABC_DLBCL_Scott2014_24398326	8	24398326
basal_HNCA_Walter2013_23451093	199	23451093

PRF_MB_Staub2012_22937182	21	22937182
chr8_amplicon_cluster2_BRCA_Farmer2005_15897907	33	15897907
RAS_Bild2006_16273092	237	16273092
MSI_down_CRC_Kim2009_19034969	22	19034969
SCC_LC_Hou2011_20421987	47	20421987
hallmark50_androgen_response_Liberzon2019_26771021	100	26771021
Bcell_Heise2014_thesis	48	thesis
BRAFlikeness_down_Popovici2012_22393095	31	22393095
Tcell_rest_Heise2014_thesis	48	thesis
IGF_down_Creighton2008_18757322	455	18757322
basal_subtype_BRCA_Calza2007_16846532	17	16846532
YAP_Mazur2020_aacr2019:A38	4	aacr2019:A38
hallmark50_EMT_Liberzon2019_26771021	400	26771021
hallmark50_IFNg_response_Liberzon2019_26771021	200	26771021
ER_pos_BRCA_Abba2006_15762987	9	15762987
IFN_Feng2006_16947629	5	16947629
BL2_refined_BRCA_Lehmann2011_21633166	23	21633166
bronchoid_markers_LC_Hayes2006_17075127	17	17075127
pDC_Angelova2015_25853550	18	25853550
Bcell_IRIS_Abbas2005_15789058	86	15789058
classical_up_GBM_Verhaak2010_20129251	152	20129251
FOLFOX_resistance_CRC_Tsuji2012_22095227	18	22095227
proneural_down_GBM_Verhaak2010_20129251	73	20129251
Tcell_IRIS_Abbas2005_15789058	14	15789058
classical_down_GBM_Verhaak2010_20129251	56	20129251
HC2A_progGroup_GBM_Freije2004_15374961	10	15374961
NKcell_Heise2014_thesis	49	thesis
hallmark50_KRAS_signaling_down_Liberzon2019_26771021	200	26771021
VGFA_HUVEC_Abe2002_12197474	12	12197474
DC_Angelova2015_25853550	34	25853550
MSI_up_CRC_Kim2009_19034969	36	19034969
poor_prognosis_BRCA_Teschendorff2007_17076897	15	17076897
MSI_Kruhffer2005_15956967	9	15956967
hallmark50_mitotic_spindle_Liberzon2019_26771021	199	26771021
tcpeI_CSC_Lottaz2010_20145155	8	20145155
human_SC_Conrad2008_18849962	39	18849962
poor_prognosis_CRC_Laiho2007_16819509	32	16819509
DDRD_Mulligan2014_24402422	24	24402422
beta_PDAC_normal_Li2016_26691212	100	26691212
inflammatory_Sadanandam2013_23584089	175	23584089
TNFa_NFkB_response_Tian2005_15722553	20	15722553
hallmark50_coagulation_Liberzon2019_26771021	138	26771021
WNT_MB_Staub2012_22937182	11	22937182
clusterB_MB_Kool2008_18769486	197	18769486
hallmark50_protein_secretion_Liberzon2019_26771021	96	26771021
radioresistance_up_BRCA_Speers2016_25904749	23	25904749

IFN_Walsh2007_17968926	6	17968926
GCB_DLBCL_Scott2014_24398326	7	24398326
hcc_tumorinItiatingEpCAMpos_Yamashita2009_19150350	45	19150350
radioresistance_Khodarev2004_14755057	51	14755057
neutrophils_Angelova2015_25853550	16	25853550
bcell_markers_Newell2010_20501946	15	20501946
CIP2_knockdown_up_Niemel2013_22809314	42	22809314
QMPDA_PCA_Collisson2011_21460848	20	21460848
hallmark50_myogenesis_Liberzon2019_26771021	200	26771021
hallmark50_TNFa_via_NFKb_Liberzon2019_26771021	200	26771021
secretory_subtype_LSCC_Wilkerson2011_20643781	24	20643781
ERBB2_up_BRCA_Bertucci2004_14743203	23	14743203
hallmark50_unfloald_protein_response_Liberzon2019_26771021	113	26771021
metastasis_BRCA_bone_Kang2004_12842083	11	12842083
non_serrated_conventional_CRC_Laiho2007_16819509	68	16819509
dendriticCell_Heise2014_thesis	17	thesis
KRAS2i_up_SweetCordero2005_15608639	23	15608639
KRAS2i_down_SweetCordero2005_15608639	26	15608639
ABC_MethExp_DLBCL_Shaknovich2010_20610814	10	20610814
neural_up_GBM_Verhaak2010_20129251	80	20129251
LPS_NFkB_targets_Sharif2007_17222336	69	17222336
cytotoxic_Bindea2013_24138885	17	24138885
MTOR_PI3K_S6K_siRNA_Heinonen2008_18652687	45	18652687
mesenchymal_HNCA_Walter2013_23451093	245	23451093
CCS3_serrated_CRC_DeSousaEMelo2013_23584090	46	23584090
pac_pdac_organoid_tiriac2018_29853643	111	29853643
oxa_pdac_organoid_tiriac2018_29853643	98	29853643
hallmark50_PI3K_AKT_MTOR_Liberzon2019_26771021	105	26771021
RS_up_Kim2013_22846430	10	22846430
delta_PDAC_normal_Murano2016_27693023	17	27693023
IFN_tcell_bcell_cluster1_BRCA_Farmer2005_15897907	43	15897907
acinar_PDAC_normal_Li2016_26691212	100	26691212
alpha_PDAC_normal_Enge2017_28965763	1	28965763
GCB_DLBCL_Wright2003_12900505	22	12900505
hallmark50_Bell_pca_Liberzon2019_26771021	80	26771021
molBL_DLBCL_MasquSoler2013_24030260	6	24030260
hallmark50_ocidative_phosphorylation_Liberzon2019_26771021	200	26771021
TP53mut_down_BRCA_Miller2005_16141321	16	16141321
DDRD_group3_Mulligan2014_24402422	7	24402422
macrophages_Bindea2013_24138885	33	24138885
alpha_PDAC_normal_Li2016_26691212	100	26691212
BRAF_high_up_Wong2011_20802181	80	20802181
stroma_cluster4_BRCA_Farmer2005_15897907	19	15897907
EMT_up_Taube2010_20713713	93	20713713
mesenchymal_PDAC_normal_Enge2017_28965763	1	28965763
BRAF_Kannengiesser2009_19383316	24	19383316

goblet_like_Sadanandam2013_23584089	94	23584089
SHH_MB_Staub2012_22937182	21	22937182
exocrine_PCA_Collisson2011_21460848	20	21460848
beta_PDAC_normal_Murano2016_27693023	14	27693023
alpha_PDAC_normal_Murano2016_27693023	17	27693023
IFN_Rice2014_24183309	4	24183309
proliferation_GBM_Phillips2006_16530701	5	16530701
IFN_MB_Staub2012_22937182	10	22937182
hippo_up_YAP_transfection_down_Zhang2008_18413746	42	18413746
hallmark50_NOTCH_Liberzon2019_26771021	32	26771021
M3_4_IFN_Chaussabel2008_18631455	59	18631455
CCS1_CIN_CRC_DeSousaEMelo2013_23584090	51	23584090
DDR_score_low_Kang2012_22505474	4	22505474
EMT_BRCA_Lien2008_17603561	32	17603561
proliferation_GBM_Liang2005_15827123	22	15827123
clusterE_MB_Kool2008_18769486	75	18769486
hallmark50_IL5_JAK_STAT3_signaling_Liberzon2019_26771021	87	26771021
PTEN_loss_down_BRCA_Saal2008_18066063	69	18066063
hallmark50_WNT_bCatenin_Liberzon2019_26771021	42	26771021
typeII_CSC_Lottaz2010_20145155	21	20145155
hallmark50_reactive_oxygen_species_Liberzon2019_26771021	49	26771021
SRC_Bild2006_16273092	55	16273092
hallmark50_adipogenesis_Liberzon2019_26771021	200	26771021
MSL_refined_BRCA_Lehmann2011_21633166	275	21633166
immune_GBM_Liang2005_15827123	35	15827123
hallmark50_spermatogenesis_Liberzon2019_26771021	135	26771021
BRCAness_low_Konstantinopoulos2010_20547991	27	20547991
MYC_Bild2006_16273092	187	16273092
EMTspheres_up_OV_Wang2012_22160925	33	22160925
BRCAness_Severson2018_28851423	77	28851423
MED12_KD_down_Huang2012_23178117	18	23178117
stem_like_Ragulan2019_31113981	9	31113981
WNT_DwAftDomNegTCFexpr_VanDerFlier2007_17320548	15	17320548
hallmark50_bile_acis_metabolism_Liberzon2019_26771021	112	26771021
proliferation_Budinska2013_23836465	83	23836465
proliferation_BRCA_Cardoso2016_27557300	47	27557300
FA_up_Gene2013_24036430	52	24036430
classical_subtype_LSCC_Wilkerson2011_20643781	18	20643781
EMTspheres_down_OV_Wang2012_22160925	28	22160925
IM_refined_BRCA_Lehmann2011_21633166	174	21633166
granulopoiesis_SLE_Bennett2003_12642603	20	12642603
chr20q_CRC_Budinska2013_23836465	33	23836465
hallmark50_IFNa_response_Liberzon2019_26771021	97	26771021
CIP2_knockdown_down_Niemel2013_22809314	15	22809314
YAP_Wang2019_30380420	21	30380420
ECM_GBM_Liang2005_15827123	19	15827123

hallmark50_P53_Liberzon2019_26771021	200	26771021
treg_Bindea2013_24138885	1	24138885
Tcell_active_Heise2014_thesis	50	thesis
chr1p_loss_down_GBM_Ngo2007_17440165	7	17440165
atypical_HNCA_Walter2013_23451093	132	23451093
luminal_apocrine_cluster7_BRCA_Farmer2005_15897907	20	15897907
hallmark50_KRAS_signaling_up_Liberzon2019_26771021	200	26771021
NE_high_25_sclc_zhang2018_29535911	25	29535911
classical_pdac_moffit2016_26343385	62	26343385
NE_low_25_sclc_zhang2018_29535911	25	29535911
GCB_MethExp_DLBCL_Shaknovich2010_20610814	6	20610814
eosinophil_Angelova2015_25853550	15	25853550
MED12_KD_MEKi_resistant_Huang2012_23178117	54	23178117
DDR_score_high_Kang2012_22505474	19	22505474
luminalB_subtype_BRCA_Calza2007_16846532	9	16846532
RSI_Eschrich2009_19735873	9	19735873
hallmark50_ER_late_Liberzon2019_26771021	200	26771021
M_refined_BRCA_Lehmann2011_21633166	25	21633166
topotecan_Pitroda2014_24670686	12	24670686
enterocyte_Ragulan2019_31113981	9	31113981
PTENi_down_Vivanco2007_17560336	45	17560336
enterocyte_Sadanandam2013_23584089	127	23584089
deathFromCancer_Glinsky2005_15931389	11	15931389
MSI_up_CRC_Staubna_internal	63	internal
hypoxia_Chi2007_16417408	17	16417408
acinar_PDAC_normal_Enge2017_28965763	1	28965763
ERBB2_amplicon_cluster8_BRCA_Farmer2005_15897907	7	15897907
claudin_low_up_BRCA_Prat2013_20813035	424	20813035
RPS_Pitroda2018_28341751	4	28341751
hallmark50_glycolysis_Liberzon2019_26771021	200	26771021
hallmark50_uv_response_up_Liberzon2019_26771021	158	26771021
HRD_score_Lu2016_25062964	114	25062964
hypoxia_Harris2002_11902584	81	11902584
ductal_PDAC_normal_Enge2017_28965763	1	28965763
serrated_CRC_Laiho2007_16819509	82	16819509
hallmark50_apical_surface_Liberzon2019_26771021	44	26771021
ABC_DLBCL_MasquSoler2013_24030260	11	24030260
IFN_SLE_Bennett2003_12642603	26	12642603
high_risk_LC_Chen2007_17202451	5	17202451
primitive_subtype_LSCC_Wilkerson2011_20643781	30	20643781
hallmark50_heme_metabolism_Liberzon2019_26771021	200	26771021
stem_like_Sadanandam2013_23584089	207	23584089
ABC_DLBCL_Wright2003_12900505	21	12900505
IGF_up_Creighton2008_18757322	381	18757322
GCB_DLBCL_MasquSoler2013_24030260	9	24030260
MSI_up_CRC_Watanabe2006_17047040	68	17047040

DDRD_down_Mulligan2014_24402422	18	24402422
WNT_cluster_MB_Kool2008_18769486	46	18769486
SCC_markers_LC_Hayes2006_17075127	15	17075127
luminalA_subtype_BRCA_Calza2007_16846532	12	16846532
ADC_LC_Hou2011_20421987	5	20421987
SHH_NanoStr_MB_Northcott2013_22057785	5	22057785
neural_down_GBM_Verhaak2010_20129251	129	20129251
hallmark50_apoptosis_Liberzon2019_26771021	161	26771021
apocrine_basal_hypoxia_cluster3_BRCA_Farmer2005_15897907	15	15897907
TGFb_early_Verrecchia2001_11279127	50	11279127
LAD_vs_LSCC_Kuner2009_18486272	9	18486272
IFN_Bilgic2010_19877033	3	19877033
hallmark50_E2F_targets_Liberzon2019_26771021	400	26771021
clusterD_MB_Kool2008_18769486	67	18769486
DDRD_group4_Mulligan2014_24402422	10	24402422
hallmark50_cholesterol_homeostasis_Liberzon2019_26771021	74	26771021
HC1A_progGroup_GBM_Freije2004_15374961	10	15374961
PTEN_loss_up_BRCA_Saal2008_18066063	110	18066063
proneural_up_GBM_Verhaak2010_20129251	135	20129251
hallmark50_peroxisome_Liberzon2019_26771021	104	26771021
PARPi_Daemen2013_22875744	7	22875744
TP53_mut_down_BRCA_Troester2007_17150101	16	17150101
WNT_canonical_KRASdep_Singh2012_22341439	32	22341439
recurr_score_oncotypeDB_BRCA_Paik2005_15591335	16	15591335
MYC_Chandriani2010_19690609	101	19690609
neutrophils_Heise2014_thesis	50	thesis
hallmark50_uv_response_down_Liberzon2019_26771021	144	26771021
hallmark50_TGFb_Liberzon2019_26771021	54	26771021
E2F3_Bild2006_16273092	224	16273092
EGFRmut_LC_Shibata2007_17459062	26	17459062
M1_2_IFN_Chaussabel2008_18631455	32	18631455
hallmark50_angiogenesis_Liberzon2019_26771021	36	26771021
BRAF_high_down_Wong2011_20802181	27	20802181
TP53_mut_up_BRCA_Troester2007_17150101	32	17150101
delta_PDAC_normal_Enge2017_28965763	1	28965763
platinum_sensitivity_Kang2012_22505474	23	22505474
radioresistance_down_BRCA_Speers2016_25904749	26	25904749
hallmark50_IL2_STAT5_signaling_Liberzon2019_26771021	200	26771021
hallmark50_MYC_targets2_Liberzon2019_26771021	58	26771021
hallmark50_allograft_rejection_Liberzon2019_26771021	200	26771021
nonMolBL_DLBCL_MasquSoler2013_24030260	4	24030260
rapamycin_Akcakanat2010_19778445	27	19778445
hallmark50_DDR_Liberzon2019_26771021	150	26771021
hallmark50_ER_early_Liberzon2019_26771021	200	26771021
goblet_like_Ragulan2019_31113981	5	31113981
mDC_monocytes_Heise2014_thesis	14	thesis

inflammatory_Ragulan2019_31113981	8	31113981
BetaCatenin_Bild2006_16273092	76	16273092
hallmark50_hedgehog_Liberzon2019_26771021	36	26771021
PTENi_up_Vivanco2007_17560336	17	17560336
ribosomal_proteins_Ashburner2000_10802651	68	10802651
proNeural_GBM_Phillips2006_16530701	14	16530701
NE_high_prostate_ostano2020_32041153	8	32041153
sn38_pdac_organoid_tiriac2018_29853643	137	29853643
basal_like_pdac_moffit2016_26343385	26	26343385